Universitat Oberta de Catalunya

uoc.edu

Data Science y KNIME, combinación perfecta para el éxito en la toma de decisiones

Pelayo García Bermejo
Grado en Ingeniería Informática
Inteligencia Artificial

Consultora: Elena Álvarez de la Campa Crespo
Profesor: Carles Ventura Royo

Fecha Entrega: 04/01/2022


Reconocimiento-NoComercial-SinObraDerivada - (CC BY-NC-ND, versión 3.0 SE)
(http://creativecommons.org/licenses/by-nc-nd/3.0/es/deed.ca).

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Data Science y KNIME, combinación perfecta para el éxito en la toma de decisiones*
Nombre del autor:	*Pelayo García Bermejo*
Nombre del consultor/a:	*Elena Álvarez de la Campa Crespo*
Nombre del PRA:	*Carles Ventura Royo*
Fecha de entrega (01/2022):	01/2022
Titulación:	*Grado en Ingeniería Informática*
Área del Trabajo Final:	*Inteligencia Artificial*
Idioma del trabajo:	*Español*
Palabras clave	*KNIME, sector inmobiliario, predicción, Machine Learnig, Data Science, Inteligencia Artificial, Deep Learning, Data Quality, Data Wrangling, Regresión lineal, K-means, Clustering, Redes Neuronales.*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados y conclusiones del trabajo.*

RESUMEN

Este proyecto tiene por objetivo presentar, de una manera sencilla, los pasos a seguir para convertir los datos en información y esta en conocimiento con el objetivo de obtener tiempo gracias a la automatización, predicción y una certera toma de decisiones.

Gracias a la combinación de técnicas BI, algoritmos de ML y un software intuitivo como es knime es posible crear un sistema básico de inteligencia artificial.

A partir de los datos en bruto (dataset), este sistema permite:

- *Limpiar y preparar los datos, perfilando, transformando y enriqueciéndolos*
- *Explorar los datos aplicando técnicas estadísticas que permiten conocer y seleccionar los datos relevantes*
- *Aplicar modelos de aprendizaje automático en función del conocimiento adquirido anteriormente en la exploración estadística.*

Los modelos elegidos para aplicar ML sobre un dataset de ejemplo con datos del sector inmobiliario se basan en el aprendizaje supervisado, donde se parte de un conjunto de datos etiquetado, y en el aprendizaje no supervisado, donde los datos no están etiquetados previamente. Estos son:

- *Regresión lineal, que permitirá modelar el comportamiento de la variable cuantitativa "Price" en función de otras variables predictoras*
- *K-means, que permitirá agrupar los inmuebles según la similitud de sus características*

Mediante distintas estadísticas, métricas y gráficos se podrá valorar la efectividad del modelo en función de las variables o número de cluster's elegid@s.

Abstract (in English, 250 words or less):

This project aims to present, in a simple way, the steps to follow to convert data into information and knowledge in order to obtain time thanks to automation, prediction and accurate decision-making.

Thanks to the combination of BI techniques, ML algorithms and intuitive software such as knime, it is possible to create a basic artificial intelligence system.
From the raw data (dataset), this system allows:

- *Clean and prepare the data, profiling, transforming and enriching it*
- *Explore the data by applying statistical techniques that allow knowing and selecting the relevant data*
- *Apply machine learning models based on previously acquired knowledge in statistical exploration.*

The models chosen to apply ML to a sample dataset with real estate data are based on supervised learning, where you start from a tagged data set, and unsupervised learning, where the data is not pre-tagged. These are:

- *Linear regression, which will allow modeling the behavior of the quantitative variable "Price" in function of other predictor variables.*
- *K-means, which will allow the properties to be grouped according to the similarity of their characteristics*

Through different statistics, metrics and graphs, the effectiveness of the model can be assessed based on the variables or number of clusters chosen.

Índice

Lista de figuras

Nota: No se contemplan como ilustraciones los pantallazos adjuntos que reflejan los resultados de acciones propias de los flujos de trabajo.

1. INTRORUCCIÓN

1.1 Contexto y justificación del Trabajo

En los últimos años se ha producido una explosión en el volumen de los datos disponibles. La tecnología actual permite expresar las actividades comerciales, industriales e individuales en datos digitales que son almacenados y procesados con el fin de extraer valor para las empresas y sus clientes.

En el siglo XXI los datos podrían considerarse como el "nuevo petróleo" [1] con la diferencia de que el petróleo tiene los días contados y los datos justo lo contrario. Cada año que pasa el volumen de producción aumenta desaforadamente y la tendencia es al alza.

Cualquier sector es consciente de que, en la era de la digitalización, la explotación del dato implica beneficios en todos los sentidos, además de ser ya una necesidad para sobrevivir en el mundo empresarial. Por ello nace la ciencia de datos o *Data Science*[2] que es un campo interdisciplinario que involucra métodos científicos, procesos y sistemas para extraer conocimiento o un mejor entendimiento de datos en sus diferentes formas, ya sea estructurados o no estructurados, lo cual es una continuación de algunos campos de análisis de datos como la estadística, la minería de datos, el aprendizaje automático, y la analítica predictiva.

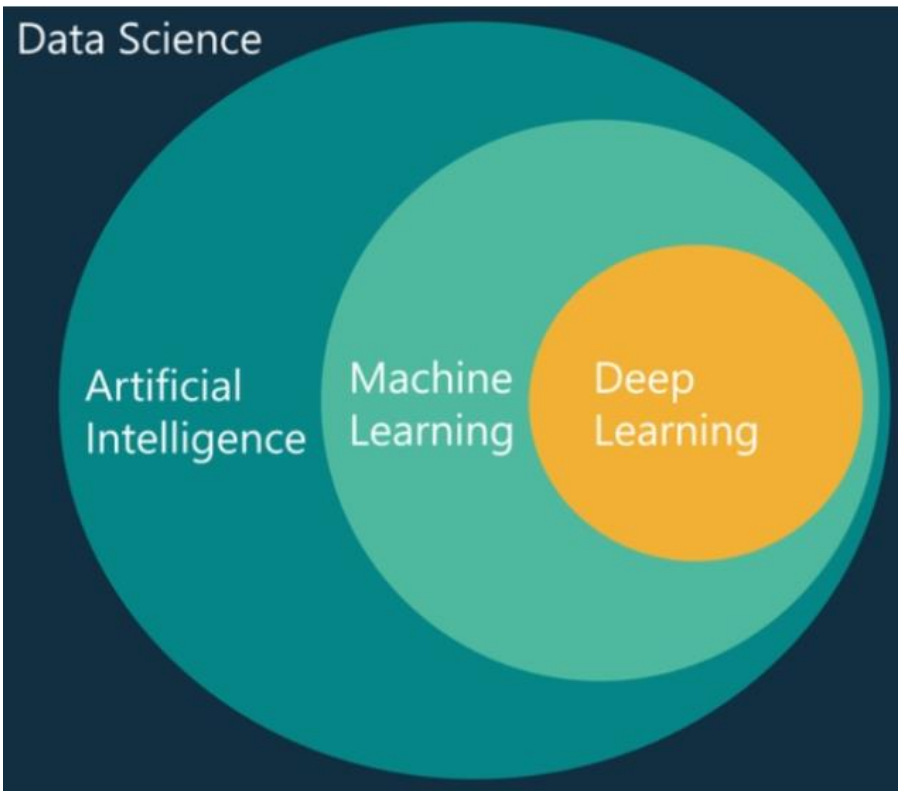


Ilustración 1 Data Science

Para identificar mejor cada campo de actuación, a continuación se expone una breve definición de cada área:

- *Artificial Intelligence* se dedica al desarrollo y aprendizaje de máquinas para atribuirles una funcionalidad humana.
- *Machine Learning*, es, digamos, una fase incluida dentro de DS y AI dónde se testean (con datos) algoritmos y modelos, y estos se emplean en el aprendizaje de la máquina.
- *Deep Learning* es un aspecto más especializado de *Machine Learning* dónde se trabaja con redes neuronales simulando aspectos del cerebro.

Los modelos de aprendizaje aplicados en datos preprocesados y depurados permiten predecir situaciones y afinar la toma de decisiones. Para conseguir este objetivo nacen distintos perfiles profesionales y múltiples herramientas/plataformas y lenguajes de programación.

Este proyecto quiere dar visibilidad a la plataforma de manejo de datos *KNIME* (*Konstanz Information Miner*) que permite el desarrollo de un proyecto analítico completo, siendo posible desarrollar cualquier fase del proyecto, desde ingestas y transformaciones, hasta modelos analíticos, predicciones y visualizaciones. *KNIME* está desarrollado sobre la plataforma *Eclipse* y programado en *Java* siendo de software libre.

Para mostrar su potencial se ha decido resolver una problemática común en el sector inmobiliario, la especulación con los precios de los inmuebles. Más concretamente, se usarán dos *dataset* del sector inmobiliario de la ciudad de New York[3], con datos del 2006 al 2010 incluidos, proporcionado por Telefónica en una formación de 10 horas a la que asistí en abril de 2021. Con estos datos se desarrollarán dos modelos analíticos, uno que prediga el precio de un inmueble (aprendizaje supervisado) y otro que agrupe los inmuebles según la similitud de sus características, es decir, segmentación (aprendizaje no supervisado). Además, se obtendrán distintas métricas y gráficos que convierten los datos en información y conocimiento.

1.2 Objetivos del Trabajo

Por suerte o por desgracia no todos sabemos de todo, y por ese motivo, se cometen errores a la hora de tomar decisiones. Este proyecto tiene como objetivo principal mostrar los beneficios que supone el uso de los datos en dicha tarea y adentrarse en el mundo del *Data Science*[2].

Una correcta toma de decisiones es un éxito en cualquier ámbito de la vida, y los datos combinados con herramientas de análisis que permitan aplicar técnicas de aprendizaje automático y el conocimiento humano son la clave.

Mediante un ejemplo claro y real, el alcance del proyecto se centrará en aplicar técnicas de ML sobre dos *dataset* para demostrar cómo se incrementa el éxito en la toma de decisiones.

Para ello se realizará un proceso analítico completo usando la plataforma de software libre, *KNIME*, mediante los datos de venta del mercado inmobiliario de la ciudad de New York, permitiendo, entre otras cosas, adelantarse y predecir el precio de los inmuebles en la ciudad de New York con el fin de evitar la especulación y orientar al posible comprador del valor real.

Las etapas de un proceso analítico y por tanto las que definen el alcance de este proyecto son las siguientes:

1. Obtención de datos: se utilizará un *dataset* en formato plano, csv, denominado "*housing_price_sin_tratar.csv"*, que contempla las siguientes variables:

    - ID = Identificador
    - LotArea = Código del área
    - LotFrontage = Código dentro del área
    - PoolArea = Área de la piscina
    - 1stFlrSF = Superficie
    - YearBuilt = Año de construcción
    - DtSold = Fecha de venta/compra
    - MSZoning = siglas de la zona
    - Foundation = base de la construcción
    - BldgType = tipo de construcción
    - Street = tipo de suelo
    - CentralAir = Aire central
    - OverallCond = Valoración, del 1 al 10, de las condiciones del inmueble
    - OverallQual = Valoración, del 1 al 10, de las calidades del inmueble
    - SalePrice = Precio de venta

2. Procesamiento y limpieza (Data Quality): mediante procesos ETL se organizan, limpian y dan formato para realizar la ingesta de los datos o en nuestro caso tratarlos.
3. Análisis exploratorio (Data Wrangling): se exploran los datos, se determina la calidad y se seleccionan los atributos y variables adecuadas para resolver el problema.

4. Modelos y algoritmos: en esta etapa se utilizan distintas técnicas de análisis y algoritmos de *machine learning*[5] para ajustar modelos a partir de un conjunto de datos, mediante un proceso de aprendizaje que nos permite descubrir patrones, hacer predicciones, o describir los datos.
5. Evaluación y visualización: se determina si el modelo obtenido ha sido satisfactorio y cumple con los objetivos preestablecidos. Se presentan los resultados mediante técnicas de visualización y representación de la información. Finalmente se extraen conclusiones y el valor para el negocio.

Una imagen que puede resumirlo podría ser la siguiente[6]:

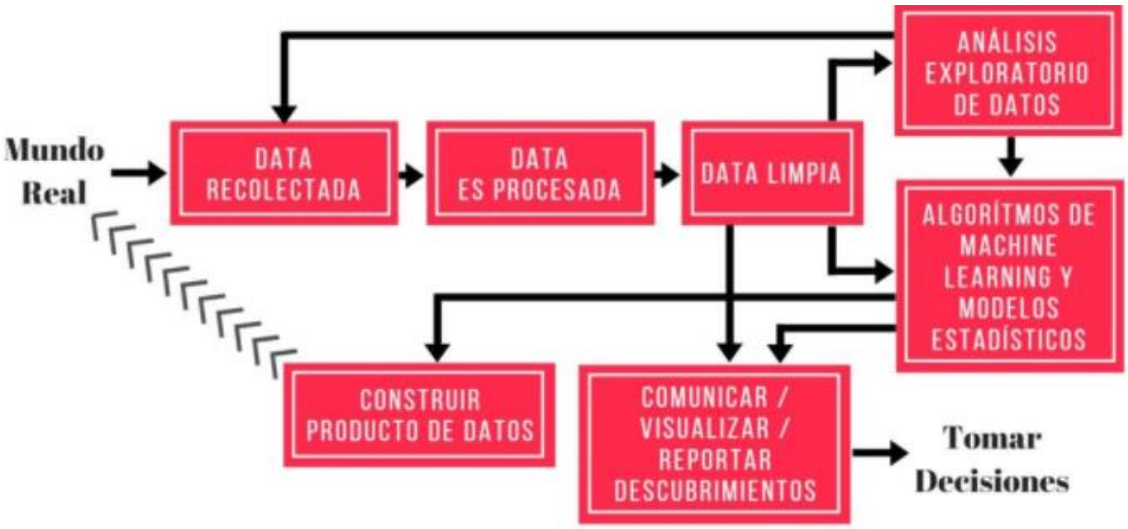


Ilustración 2 Ciclo del proceso analítico de datos

Por tanto, el proyecto se dividirá en 4 entregables que consistirán en lo siguiente:

1. Perfilar, transformar y enriquecer los datos mejorando su calidad y valor (Data Quality[16] y Data Wrangling[17]). Esto permitirá limpiar y preparar el conjunto de datos y detectar las problemáticas típicas de la ingeniería de datos.

   Perfilar y transformar (Data Quality):

   - Realizar un análisis de las variables para conocer el fichero csv y la información que contiene.
   - Evaluar la limpieza de los datos y corregir los errores que detectemos (formateo, manipulación y normalización de variables, detección y tratamiento de nulos, duplicados y *outliers*)

   Enriquecer (Data Wranglig):

   - Enriquecer el fichero recibido, para ello se tiene pensado relacionar con información sobre la descripción de la zona, para ampliar el detalle, y realizar una agrupación por ella para saber el número de ventas y precio medio por zona

2. Realizar un análisis exploratorio (EDA[18]) como etapa previa a seleccionar la técnica de *Machine Learning*[5] a aplicar.

    - Realizar un análisis gráfico y descriptivo de las variables individuales, de las relaciones entre ellas que cuantifique su interrelación, identificar los valores extremos (outliers) y evaluar el impacto de datos ausentes sobre la representatividad de los datos. Se usarán técnicas gráficas y descriptivas simples como histogramas (Histogram), gráficos de barras, probabilidades, diagramas de caja (BoxPlot), relación entre variables con Scatter Plot etc.y algún trazado estadístico sencillo como, gráficos de media y desviación estándar.
    - Seleccionar un modelo adecuado para obtener información acerca de los diferentes perfiles poblacionales, esto es, qué tipo de clientes existen, sus rasgos principales y el número de cada tipo.

      Con esto se podrán responder múltiples cuestiones como, por ejemplo, variables con mayor número de nulos, valor medio de venta de los inmuebles, el mes del año que se producen más operaciones de venta de inmuebles u obtener una lista de los 10 precios que más veces se repiten en el dataset proporcionado.

3. Aplicar modelos de aprendizaje supervisados[7] (por ejemplo, para predecir el precio de inmuebles se podría usar regresión) y no supervisados[9] (como Kmeans[10], agrupando en clusters las propiedades inmobiliarias, seguramente necesite otro dataset).

4. Evaluar y visualizar la efectividad de cada modelo de desarrollo y mostrar los datos.

1.3 Enfoque y método seguido

Se desarrollará un nuevo producto, usando la herramienta de *machine learning KNIME*, que partirá de los datos proporcionados por Telefónica en una formación a la que asistí en 2021. Este origen de datos se divide en dos dataset en formato csv y contienen información del sector inmobiliario de la ciudad de New York entre los años 2006 y 2010. El volumen es limitado, dando por sentado que la capacidad de realizar el tratamiento ágil y masivo es posible gracias al software utilizado y a la rapidez de procesamiento, siendo posible el uso de sistemas distribuidos en función de la demanda.

Siguiendo el material de estudio de la asignatura Gestión de Proyectos", se combinará la guía del PMI, el PMBOK[13], y metodologías ágiles, concretamente scrum[14] permitiendo una estrategia de desarrollo incremental y no planificación y ejecución completa del producto (metodología tradicional), esto, sin duda, garantiza calidad y evita sorpresas en el ciclo de vida completo del proyecto. Esta diferencia se puede ver mejor gráficamente:

Ilustración 3 Comparativa metodología tradicional y agile

Por tanto, se seguirán las siguientes fases:

- Iniciación: donde se realiza el estudio de viabilidad, se definen la necesidad, los requerimientos, el software a utilizar, la metodología de gestión y los recursos.
- Planificación: se define el alcance del proyecto, en tiempo coste y recursos, y los objetivos cuantificadores. Para ello, el proyecto se divide en entregables. Ver punto siguiente 1.4 Planificación del Trabajo.
- Ejecución: en esta fase se ejecuta cada hito planificado, compuestos de tareas y que forman parte de los entregables. Siguiendo la metodología scrum[15], estos hitos no superan los 4 días de ejecución, tienen un seguimiento diario y finalizan con la puesta en común con todos los stakeholder (iteración). Con esto es posible detectar nuevos requerimientos asumibles dentro del proyecto o en próximos, impresiones del cliente final, errores y/o mal funcionamiento por posibles malas interpretaciones o casos de uso no contemplados entre otros problemas, además incrementan el entregable final de una manera muy positiva ya que queda revisada

y prácticamente aprobada esta parte/hito. En el caso de este proyecto, el producto será revisado por mi y por los profesores asignados. A continuación, se ve gráficamente como es esta fase:

Ilustración 4 Fase ejecución en Scrum

- Seguimiento y control: se lleva un control diario de las tareas, permitiendo anteponerse a posibles problemas. En un proyecto normal se debería contar con la comunicación directa con un Director de Proyectos para resolver dudas y problemas y unos potenciales usuarios del producto, en este caso, no es posible, asumiendo yo este rol.
- Finalización: esta última fase permitirá presentar el producto sin fallos. Al haberse aplicado la metodología indicada con entregables incrementales y periódicos, se evitarán sorpresas.

1.4 Planificación del Trabajo

La siguiente planificación da por realizada la fase de iniciación, concretamente es la fase de planificación y se ciñe al alcance funcional del proyecto, comprendiendo el diseño de 4 diagramas de flujo de datos, pudiéndose considerar también como 4 ETL's, que muestran el proceso de análisis de datos completo y la aplicación de modelos de aprendizaje automático (ML).

Para conseguir estos objetivos en el tiempo estipulado, se sigue la siguiente planificación dividida en hitos y entregables:

Entregable 1: Flujo BI Limpiar y preparar los datos

- Hito 1 Data Quality
- Hito 2 Data Wranglig
- Hito 3 Documentar

Entregable 2: Flujo EDA Análisis exploratorio

- Hito 1 Análisis descriptivo
- Hito 2 Análisis Gráfico
- Hito 3 Documentar

Entregable 3: Flujo ML supervisado

- Hito 1 Algoritmo Regresión Lineal
- Hito 2 Evaluación del modelo
- Hito 3 Documentar

Entregable 4: Flujo ML no supervisado

- Hito 1 Algoritmo K-Means
- Hito 2 Evaluación del modelo
- Hito 3 Documentar

Revisión general

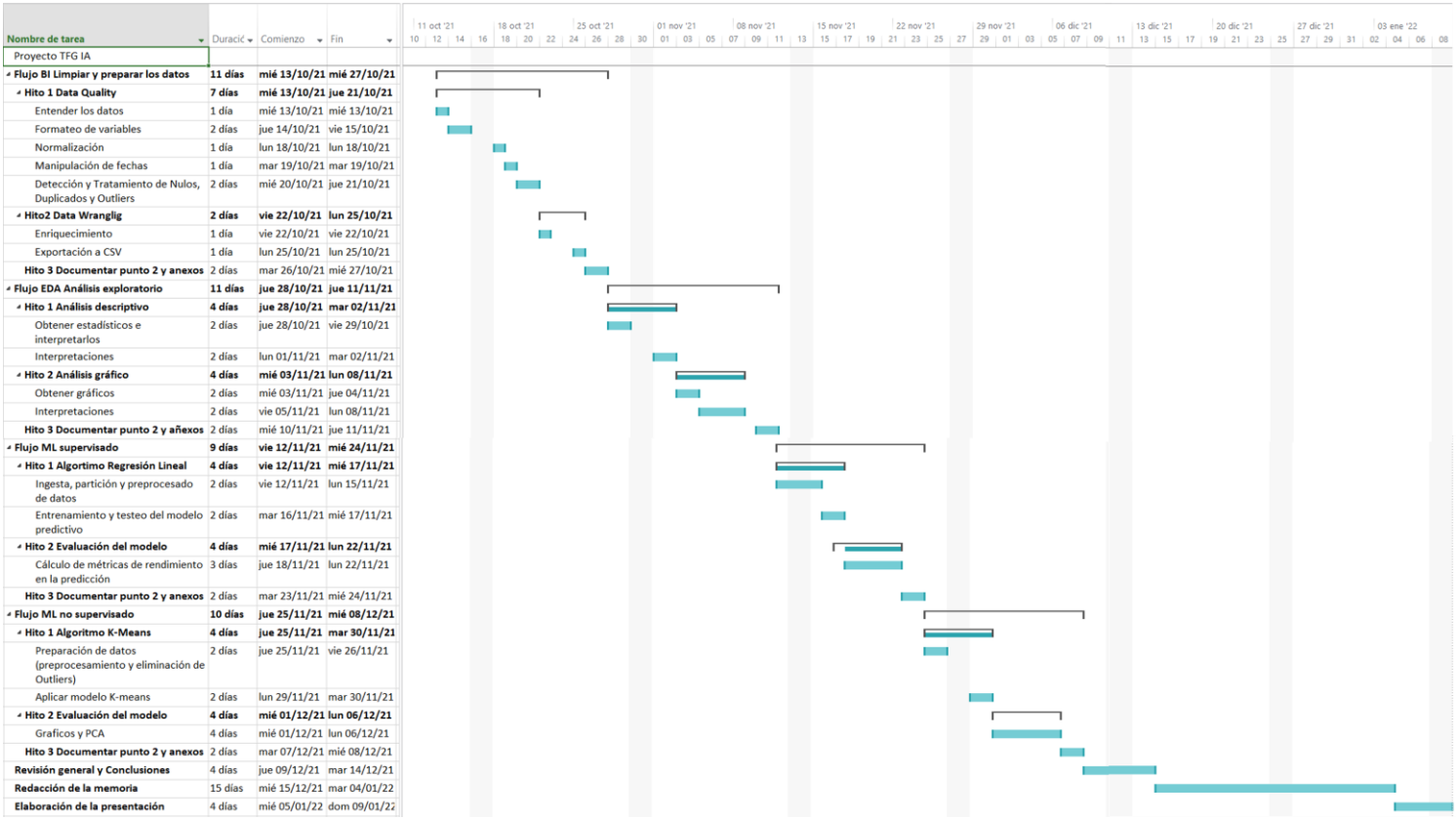
Redactado de la memoria

Elaboración de la presentación final

A continuación, se muestra la tabla desglosada por tareas con sus tiempos y fechas:

Nombre de tarea	Duración	Comienzo	Fin
Proyecto TFG IA	**63 días**	**mié 13/10/21**	**dom 09/01/22**
Flujo BI Limpiar y preparar los datos	**11 días**	**mié 13/10/21**	**mié 27/10/21**
Hito 1 Data Quality	**7 días**	**mié 13/10/21**	**jue 21/10/21**
Entender los datos	1 día	mié 13/10/21	mié 13/10/21
Formateo de variables	2 días	jue 14/10/21	vie 15/10/21
Normalización	1 día	lun 18/10/21	lun 18/10/21
Manipulación de fechas	1 día	mar 19/10/21	mar 19/10/21
Detección y Tratamiento de Nulos, Duplicados y Outliers	2 días	mié 20/10/21	jue 21/10/21
Hito2 Data Wranglig	**2 días**	**vie 22/10/21**	**lun 25/10/21**
Enriquecimiento	1 día	vie 22/10/21	vie 22/10/21
Exportación a CSV	1 día	lun 25/10/21	lun 25/10/21
Hito 3 Documentar punto 2 y anexos	2 días	mar 26/10/21	mié 27/10/21
Flujo EDA Análisis exploratorio	**11 días**	**jue 28/10/21**	**jue 11/11/21**
Hito 1 Análisis descriptivo	**4 días**	**jue 28/10/21**	**mar 02/11/21**
Obtener estadísticos e interpretarlos	2 días	jue 28/10/21	vie 29/10/21
Interpretaciones	2 días	lun 01/11/21	mar 02/11/21
Hito 2 Análisis gráfico	**4 días**	**mié 03/11/21**	**lun 08/11/21**
Obtener gráficos	2 días	mié 03/11/21	jue 04/11/21
Interpretaciones	2 días	vie 05/11/21	lun 08/11/21
Hito 3 Documentar punto 2 y anexos	2 días	mié 10/11/21	jue 11/11/21
Flujo ML supervisado	**9 días**	**vie 12/11/21**	**mié 24/11/21**
Hito 1 Algoritmo Regresión Lineal	**4 días**	**vie 12/11/21**	**mié 17/11/21**
Ingesta, partición y preprocesado de datos	2 días	vie 12/11/21	lun 15/11/21
Entrenamiento y testeo del modelo predictivo	2 días	mar 16/11/21	mié 17/11/21
Hito 2 Evaluación del modelo	**4 días**	**mié 17/11/21**	**lun 22/11/21**
Cálculo de métricas de rendimiento en la predicción	3 días	jue 18/11/21	lun 22/11/21
Hito 3 Documentar punto 2 y anexos	2 días	mar 23/11/21	mié 24/11/21
Flujo ML no supervisado	**10 días**	**jue 25/11/21**	**mié 08/12/21**
Hito 1 Algoritmo K-Means	**4 días**	**jue 25/11/21**	**mar 30/11/21**
Preparación de datos (preprocesamiento y eliminación de Outliers)	2 días	jue 25/11/21	vie 26/11/21
Aplicar modelo K-means	2 días	lun 29/11/21	mar 30/11/21
Hito 2 Evaluación del modelo	**4 días**	**mié 01/12/21**	**lun 06/12/21**
Gráficos y PCA	4 días	mié 01/12/21	lun 06/12/21
Hito 3 Documentar punto 2 y anexos	2 días	mar 07/12/21	mié 08/12/21
Revisión general y Conclusiones	4 días	jue 09/12/21	mar 14/12/21
Redacción de la memoria	15 días	mié 15/12/21	mar 04/01/22
Elaboración de la presentación	4 días	mié 05/01/22	dom 09/01/22

El diagrama de Gantt quedaría de la siguiente manera:



Nombre de tarea	Duració	Comienzo	Fin
Proyecto TFG IA			
Flujo BI Limpiar y preparar los datos	**11 días**	**mié 13/10/21**	**mié 27/10/21**
Hito 1 Data Quality	**7 días**	**mié 13/10/21**	**jue 21/10/21**
Entender los datos	1 día	mié 13/10/21	mié 13/10/21
Formateo de variables	2 días	jue 14/10/21	vie 15/10/21
Normalización	1 día	lun 18/10/21	lun 18/10/21
Manipulación de fechas	1 día	mar 19/10/21	mar 19/10/21
Detección y Tratamiento de Nulos, Duplicados y Outliers	2 días	mié 20/10/21	jue 21/10/21
Hito2 Data Wranglig	**2 días**	**vie 22/10/21**	**lun 25/10/21**
Enriquecimiento	1 día	vie 22/10/21	vie 22/10/21
Exportación a CSV	1 día	lun 25/10/21	lun 25/10/21
Hito 3 Documentar punto 2 y anexos	2 días	mar 26/10/21	mié 27/10/21
Flujo EDA Análisis exploratorio	**11 días**	**jue 28/10/21**	**jue 11/11/21**
Hito 1 Análisis descriptivo	**4 días**	**jue 28/10/21**	**mar 02/11/21**
Obtener estadísticos e interpretarlos	2 días	jue 28/10/21	vie 29/10/21
Interpretaciones	2 días	lun 01/11/21	mar 02/11/21
Hito 2 Análisis gráfico	**4 días**	**mié 03/11/21**	**lun 08/11/21**
Obtener gráficos	2 días	mié 03/11/21	jue 04/11/21
Interpretaciones	2 días	vie 05/11/21	lun 08/11/21
Hito 3 Documentar punto 2 y añexos	2 días	mié 10/11/21	jue 11/11/21
Flujo ML supervisado	**9 días**	**vie 12/11/21**	**mié 24/11/21**
Hito 1 Algortimo Regresión Lineal	**4 días**	**vie 12/11/21**	**mié 17/11/21**
Ingesta, partición y preprocesado de datos	2 días	vie 12/11/21	lun 15/11/21
Entrenamiento y testeo del modelo predictivo	2 días	mar 16/11/21	mié 17/11/21
Hito 2 Evaluación del modelo	**4 días**	**mié 17/11/21**	**lun 22/11/21**
Cálculo de métricas de rendimiento en la predicción	3 días	jue 18/11/21	lun 22/11/21
Hito 3 Documentar punto 2 y anexos	2 días	mar 23/11/21	mié 24/11/21
Flujo ML no supervisado	**10 días**	**jue 25/11/21**	**mié 08/12/21**
Hito 1 Algoritmo K-Means	**4 días**	**jue 25/11/21**	**mar 30/11/21**
Preparación de datos (preprocesamiento y eliminación de Outliers)	2 días	jue 25/11/21	vie 26/11/21
Aplicar modelo K-means	2 días	lun 29/11/21	mar 30/11/21
Hito 2 Evaluación del modelo	**4 días**	**mié 01/12/21**	**lun 06/12/21**
Graficos y PCA	4 días	mié 01/12/21	lun 06/12/21
Hito 3 Documentar punto 2 y anexos	2 días	mar 07/12/21	mié 08/12/21
Revisión general y Conclusiones	4 días	jue 09/12/21	mar 14/12/21
Redacción de la memoria	15 días	mié 15/12/21	mar 04/01/22
Elaboración de la presentación	4 días	mié 05/01/22	dom 09/01/22

Se ha usado el software Microsoft Project 2019 para realizar esta planificación que servirá para hacer un seguimiento del proyecto completo, fijando una línea de base que permitirá comparar el proyecto entre lo planificado y lo realizado realmente, es decir, seguimiento del progreso del proyecto.

1.5 Breve sumario de productos obtenidos

Con este proyecto se quiere mostrar los pasos que se deben seguir para explotar los datos con el objetivo de obtener conocimiento y poder acertar en la toma de decisiones. Para ello se combinarán técnicas de BI y ML con el software *Knime* sobre dos *dataset* con información del mercado inmobiliario (2006-2010) en la ciudad de New York.

El producto final se acota al tiempo especificado por la UOC, en torno a 3 meses. En el punto 1.4 se ve el detalle de la planificación y, por tanto, el alcance del proyecto.

Dentro de este marco, el producto final contendrá los siguientes flujos:

1. Flujo BI: donde se perfilan, transforman y enriquecen los datos.
2. Flujo EDA: con los datos obtenidos en el flujo anterior se realizará un análisis exploratorio mediante un flujo que incorporará estadísticos y gráficos (Histograma, *BoxPlot* y gráficos de barra).
3. Flujo ML supervisado: se aplicará regresión lineal sobre los datos. Este modelo entrenará y testeará los precios de los inmuebles y dará las métricas de rendimiento en la predicción.
4. Flujo ML no supervisado: se aplicará *Kmeans*, agrupando por zonas y viendo la distribución de cada variable individualmente. La evaluación, a diferencia del flujo anterior, no puede hacerse de forma automática, se realizará viendo las características de las agrupaciones, el número de datos de cada *cluster* y la distancia entre los centroides para saber si el número K (número de *clusters*) es el idóneo para el modelo.

1.6 Breve descripción de los otros capítulos de la memoria

En esta memoria, además de lo visto en el punto 1 (antecedentes, objetivos, enfoque, planificación y productos conseguidos) se ve necesario detallar cada punto incluido en la planificación, es decir:

- Flujo BI: Se analizará el origen de los datos que se utilizará para alimentar el sistema. Se definirán las distintas acciones y los pasos/nodos a usar en *Knime* para perfilar y transformar los datos.
- Flujo EDA: se definirán los pasos/nodos a usar en *Knime* para obtener estadísticos y gráficos que permitan realizar un análisis exploratorio.
- Flujo ML supervisado: se explicará cómo se aplicará el algoritmo de regresión lineal, qué pasos/nodos se usarán para obtener la predicción y las métricas usadas para obtener conclusiones sobre la aplicación del modelo.
- Flujo ML no supervisado: se explicará cómo se aplicará el algoritmo de K-*means*, qué pasos/nodos se usarán para obtener la predicción y las métricas usadas para obtener conclusiones sobre la aplicación del modelo.
- Conclusiones: a nivel técnico, funcional y personal.
- Glosario
- Bibliografía
- Anexos: que proporcionan información detallada y técnica de los puntos más relevantes en la definición del producto (puntos previamente expuestos).

2. PRODUCTOS

Mediante 4 flujos o *pipeline* distintos y usando la herramienta de Machine Learning *KNIME*, se quiere reflejar una progresión de cómo se realiza un análisis completo de los datos hasta obtener conocimiento de estos.

Se podrán ver los problemas típicos, metodologías y procesos que se aplican en un proyecto de analítica de datos.

Además, se proporcionan alternativas de aprendizaje y diversas métricas que permitirán evaluar, en función de la disposición de los datos, los resultados obtenidos enfocados a la toma de decisiones.

2.1 Flujo BI, limpiar y preparar los datos. Definición

El primer paso para comenzar con la analítica de datos es limpiar y preparar los *dataset* y detectar las problemáticas de la ingeniería de datos en general[19].

Para ello se ha decidido usar datos del mercado inmobiliario en la ciudad de New York entre los años 2006 y 2010 ("*housing_price_sin_tratar.csv"*).

Después de realizar un análisis de las variables para conocer los datos y la información que contiene el primer *dataset* (ver página 7 de este documento) se marcan los siguientes objetivos para este primer flujo:

- Limpiar y corregir errores detectados.
- Enriquecer el fichero recibido con otra información relacionable con el primer *dataset*.
- Exportar a *csv* un fichero limpio para la realización del siguiente flujo dirigido al análisis exploratorio de estos datos.

La preparación de los datos consume entre un 60% y un 80% del tiempo total de un proyecto analítico, es muy importante determinar la calidad de estos (exactitud, integridad, confiabilidad, relevancia, actualización) y procesarlos, convirtiéndolos o mapeándolos a otro formato que permita acceder a los datos de forma más organizada.

A continuación, se detalla el perfilado y la transformación que sufrirán los datos en el *pipeline* de este apartado. Además, se mostrarán los pasos realizados en KNIME y el flujo completo con los resultados obtenidos.

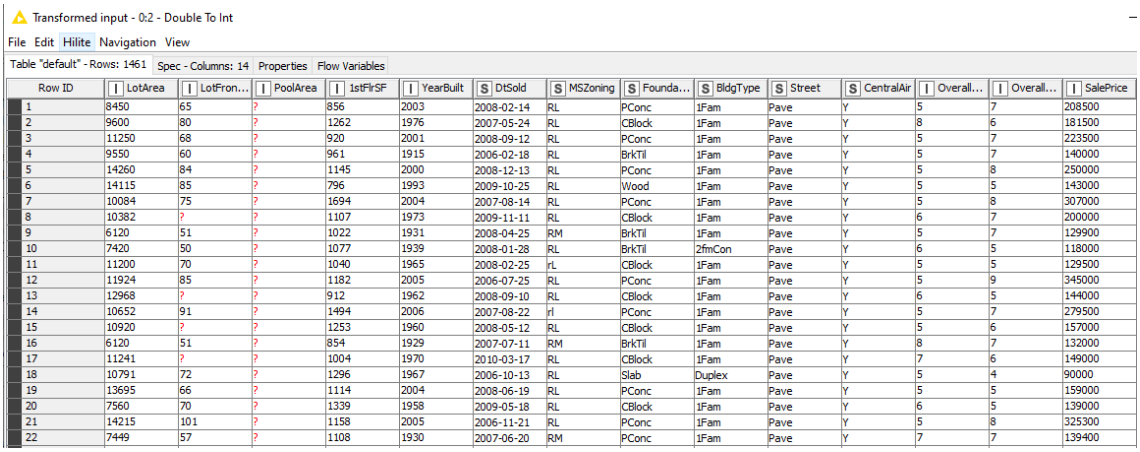
2.1.1 Perfilado y transformación. Definición

Permite verificar sintáctica y semánticamente los datos, el rango en el que se encuentran y la distribución de los valores o relación con otros campos.

En la limpieza del conjunto de datos que ocupa a este proyecto se opta por realizar las siguientes acciones usando distintos nodos/pasos de KNIME:

1-Leer datos: lee los datos en bruto del csv ubicado en una ruta especificada.

2-Formatear variables numéricas: se redondean las variables con decimales convirtiéndolas en enteros. Quedando de la siguiente manera:



Transformed input - 0:2 - Double To Int

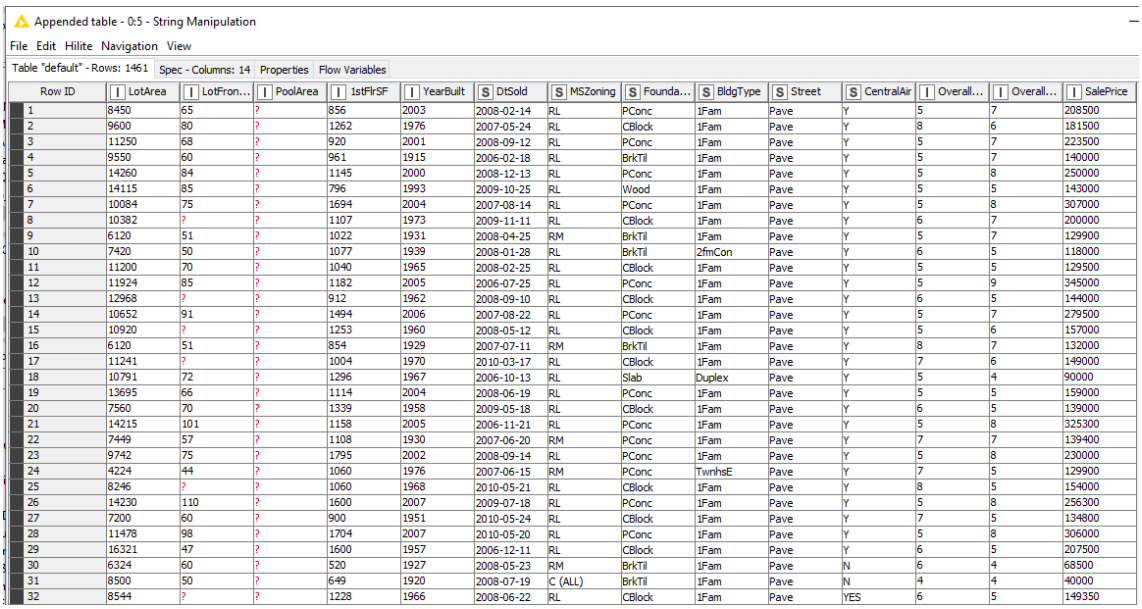
File Edit Hilite Navigation View

Table "default" - Rows: 1461 Spec - Columns: 14 Properties Flow Variables

Row ID	LotArea	LotFron...	PoolArea	1stFlrSF	YearBuilt	DtSold	MSZoning	Founda...	BldgType	Street	CentralAir	Overall...	Overall...	SalePrice
1	8450	65	?	856	2003	2008-02-14	RL	PConc	1Fam	Pave	Y	5	7	208500
2	9600	80	?	1262	1976	2007-05-24	RL	CBlock	1Fam	Pave	Y	8	6	181500
3	11250	68	?	920	2001	2008-09-12	RL	PConc	1Fam	Pave	Y	5	7	223500
4	9550	60	?	961	1915	2006-02-18	RL	BrkTil	1Fam	Pave	Y	5	7	140000
5	14260	84	?	1145	2000	2008-12-13	RL	PConc	1Fam	Pave	Y	5	8	250000
6	14115	85	?	796	1993	2009-10-25	RL	Wood	1Fam	Pave	Y	5	5	143000
7	10084	75	?	1694	2004	2007-08-14	RL	PConc	1Fam	Pave	Y	5	8	307000
8	10382	?	?	1107	1973	2009-11-11	RL	CBlock	1Fam	Pave	Y	6	7	200000
9	6120	51	?	1022	1931	2008-04-25	RM	BrkTil	1Fam	Pave	Y	5	7	129900
10	7420	50	?	1077	1939	2008-01-28	RL	BrkTil	2fmCon	Pave	Y	6	5	118000
11	11200	70	?	1040	1965	2008-02-25	rL	CBlock	1Fam	Pave	Y	5	5	129500
12	11924	85	?	1182	2005	2006-07-25	RL	PConc	1Fam	Pave	Y	5	9	345000
13	12968	?	?	912	1962	2008-09-10	RL	CBlock	1Fam	Pave	Y	6	5	144000
14	10652	91	?	1494	2006	2007-08-22	rl	PConc	1Fam	Pave	Y	5	7	279500
15	10920	?	?	1253	1960	2008-05-12	RL	CBlock	1Fam	Pave	Y	5	6	157000
16	6120	51	?	854	1929	2007-07-11	RM	BrkTil	1Fam	Pave	Y	8	7	132000
17	11241	?	?	1004	1970	2010-03-17	RL	CBlock	1Fam	Pave	Y	7	6	149000
18	10791	72	?	1296	1967	2006-10-13	RL	Slab	Duplex	Pave	Y	5	4	90000
19	13695	66	?	1114	2004	2008-06-19	RL	PConc	1Fam	Pave	Y	5	5	159000
20	7560	70	?	1339	1958	2009-05-18	RL	CBlock	1Fam	Pave	Y	6	5	139000
21	14215	101	?	1158	2005	2006-11-21	RL	PConc	1Fam	Pave	Y	5	8	325300
22	7449	57	?	1108	1930	2007-06-20	RM	PConc	1Fam	Pave	Y	7	7	139400

3-Formatear variables string:

Se detecta que las variables *MSZoning* y *CentrailAir* tienen mayusculas y minúsculas en sus valores, se decide reemplazar en la misma variables los valores formateándolos a mayúsculas. Quedando de la siguiente manera:



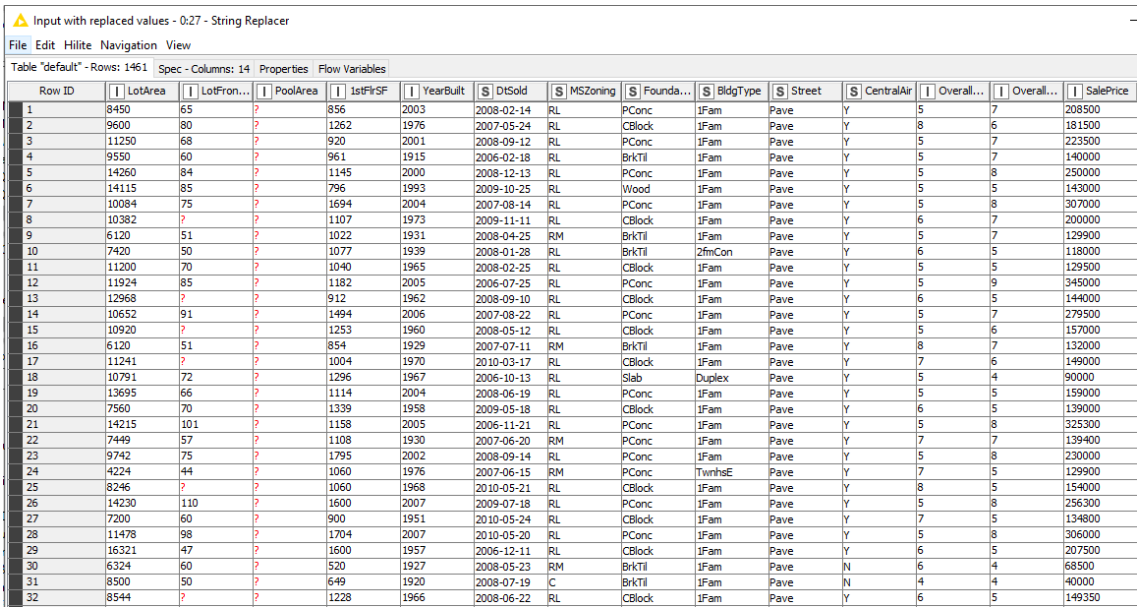
Appended table - 0:5 - String Manipulation

File Edit Hilite Navigation View

Table "default" - Rows: 1461 Spec - Columns: 14 Properties Flow Variables

Row ID	LotArea	LotFron...	PoolArea	1stFlrSF	YearBuilt	DtSold	MSZoning	Founda...	BldgType	Street	CentralAir	Overall...	Overall...	SalePrice
1	8450	65	?	856	2003	2008-02-14	RL	PConc	1Fam	Pave	Y	5	7	208500
2	9600	80	?	1262	1976	2007-05-24	RL	CBlock	1Fam	Pave	Y	8	6	181500
3	11250	68	?	920	2001	2008-09-12	RL	PConc	1Fam	Pave	Y	5	7	223500
4	9550	60	?	961	1915	2006-02-18	RL	BrkTil	1Fam	Pave	Y	5	7	140000
5	14260	84	?	1145	2000	2008-12-13	RL	PConc	1Fam	Pave	Y	5	8	250000
6	14115	85	?	796	1993	2009-10-25	RL	Wood	1Fam	Pave	Y	5	5	143000
7	10084	75	?	1694	2004	2007-08-14	RL	PConc	1Fam	Pave	Y	5	8	307000
8	10382	?	?	1107	1973	2009-11-11	RL	CBlock	1Fam	Pave	Y	6	7	200000
9	6120	51	?	1022	1931	2008-04-25	RM	BrkTil	1Fam	Pave	Y	5	7	129900
10	7420	50	?	1077	1939	2008-01-28	RL	BrkTil	2fmCon	Pave	Y	6	5	118000
11	11200	70	?	1040	1965	2008-02-25	RL	CBlock	1Fam	Pave	Y	5	5	129500
12	11924	85	?	1182	2005	2006-07-25	RL	PConc	1Fam	Pave	Y	5	9	345000
13	12968	?	?	912	1962	2008-09-10	RL	CBlock	1Fam	Pave	Y	6	5	144000
14	10652	91	?	1494	2006	2007-08-22	RL	PConc	1Fam	Pave	Y	5	7	279500
15	10920	?	?	1253	1960	2008-05-12	RL	CBlock	1Fam	Pave	Y	5	6	157000
16	6120	51	?	854	1929	2007-07-11	RM	BrkTil	1Fam	Pave	Y	8	7	132000
17	11241	?	?	1004	1970	2010-03-17	RL	CBlock	1Fam	Pave	Y	7	6	149000
18	10791	72	?	1296	1967	2006-10-13	RL	Slab	Duplex	Pave	Y	5	4	90000
19	13695	66	?	1114	2004	2008-06-19	RL	PConc	1Fam	Pave	Y	5	5	159000
20	7560	70	?	1339	1958	2009-05-18	RL	CBlock	1Fam	Pave	Y	6	5	139000
21	14215	101	?	1158	2005	2006-11-21	RL	PConc	1Fam	Pave	Y	5	8	325300
22	7449	57	?	1108	1930	2007-06-20	RM	PConc	1Fam	Pave	Y	7	7	139400
23	9742	75	?	1795	2002	2008-09-14	RL	PConc	1Fam	Pave	Y	5	8	230000
24	4224	44	?	1060	1976	2007-06-15	RM	PConc	TwnhsE	Pave	Y	7	5	129900
25	8246	?	?	1060	1968	2010-05-21	RL	CBlock	1Fam	Pave	Y	8	5	154000
26	14230	110	?	1600	2007	2009-07-18	RL	PConc	1Fam	Pave	Y	5	8	256300
27	7200	60	?	900	1951	2010-05-24	RL	CBlock	1Fam	Pave	Y	7	5	134800
28	11478	98	?	1704	2007	2010-05-20	RL	PConc	1Fam	Pave	Y	5	8	306000
29	16321	47	?	1600	1957	2006-12-11	RL	CBlock	1Fam	Pave	Y	6	5	207500
30	6324	60	?	520	1927	2008-05-23	RM	BrkTil	1Fam	Pave	N	6	4	68500
31	8500	50	?	649	1920	2008-07-19	C (ALL)	BrkTil	1Fam	Pave	N	4	4	40000
32	8544	?	?	1228	1966	2008-06-22	RL	CBlock	1Fam	Pave	YES	6	5	149350

4-Normalizar Codificaciones: se detectan valores distintos para el mismo significado, en la variable *MSZoning* “C (ALL)” deberá reemplazarse por “C” y en la variable *CentralAir* “YES” por “Y” y “NO” por “N”. Quedando de la siguiente manera:



Input with replaced values - 0:27 - String Replacer

File Edit Hilite Navigation View

Table "default" - Rows: 1461 Spec - Columns: 14 Properties Flow Variables

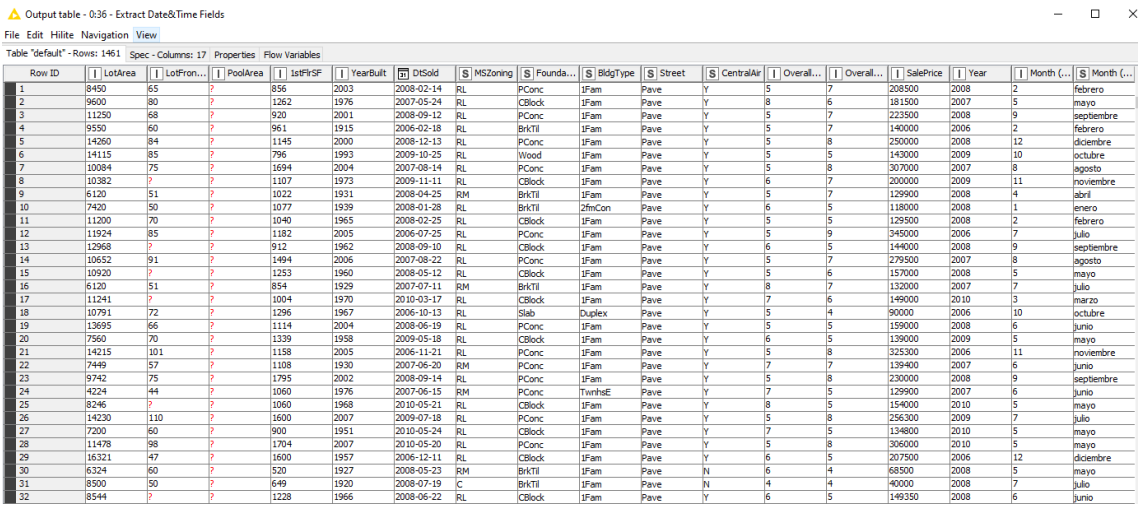
Row ID	LotArea	LotFron...	PoolArea	1stFlrSF	YearBuilt	DtSold	MSZoning	Founda...	BldgType	Street	CentralAir	Overall...	Overall...	SalePrice
1	8450	65	?	856	2003	2008-02-14	RL	PConc	1Fam	Pave	Y	5	7	208500
2	9600	80	?	1262	1976	2007-05-24	RL	CBlock	1Fam	Pave	Y	8	6	181500
3	11250	68	?	920	2001	2008-09-12	RL	PConc	1Fam	Pave	Y	5	7	223500
4	9550	60	?	961	1915	2006-02-18	RL	BrkTil	1Fam	Pave	Y	5	7	140000
5	14260	84	?	1145	2000	2008-12-13	RL	PConc	1Fam	Pave	Y	5	8	250000
6	14115	85	?	796	1993	2009-10-25	RL	Wood	1Fam	Pave	Y	5	5	143000
7	10084	75	?	1694	2004	2007-08-14	RL	PConc	1Fam	Pave	Y	5	8	307000
8	10382	?	?	1107	1973	2009-11-11	RL	CBlock	1Fam	Pave	Y	6	7	200000
9	6120	51	?	1022	1931	2008-04-25	RM	BrkTil	1Fam	Pave	Y	5	7	129900
10	7420	50	?	1077	1939	2008-01-28	RL	BrkTil	2fmCon	Pave	Y	6	5	118000
11	11200	70	?	1040	1965	2008-02-25	RL	CBlock	1Fam	Pave	Y	5	5	129500
12	11924	85	?	1182	2005	2006-07-25	RL	PConc	1Fam	Pave	Y	5	9	345000
13	12968	?	?	912	1962	2008-09-10	RL	CBlock	1Fam	Pave	Y	6	5	144000
14	10652	91	?	1494	2006	2007-08-22	RL	PConc	1Fam	Pave	Y	5	7	279500
15	10920	?	?	1253	1960	2008-05-12	RL	CBlock	1Fam	Pave	Y	5	6	157000
16	6120	51	?	854	1929	2007-07-11	RM	BrkTil	1Fam	Pave	Y	8	7	132000
17	11241	?	?	1004	1970	2010-03-17	RL	CBlock	1Fam	Pave	Y	7	6	149000
18	10791	72	?	1296	1967	2006-10-13	RL	Slab	Duplex	Pave	Y	5	4	90000
19	13695	66	?	1114	2004	2008-06-19	RL	PConc	1Fam	Pave	Y	5	5	159000
20	7560	70	?	1339	1958	2009-05-18	RL	CBlock	1Fam	Pave	Y	6	5	139000
21	14215	101	?	1158	2005	2006-11-21	RL	PConc	1Fam	Pave	Y	5	8	325300
22	7449	57	?	1108	1930	2007-06-20	RM	PConc	1Fam	Pave	Y	7	7	139400
23	9742	75	?	1795	2002	2008-09-14	RL	PConc	1Fam	Pave	Y	5	8	230000
24	4224	44	?	1060	1976	2007-06-15	RM	PConc	TwnhsE	Pave	Y	7	5	129900
25	8246	?	?	1060	1968	2010-05-21	RL	CBlock	1Fam	Pave	Y	8	5	154000
26	14230	110	?	1600	2007	2009-07-18	RL	PConc	1Fam	Pave	Y	5	8	256300
27	7200	60	?	900	1951	2010-05-24	RL	CBlock	1Fam	Pave	Y	7	5	134800
28	11478	98	?	1704	2007	2010-05-20	RL	PConc	1Fam	Pave	Y	5	8	306000
29	16321	47	?	1600	1957	2006-12-11	RL	CBlock	1Fam	Pave	Y	6	5	207500
30	6324	60	?	520	1927	2008-05-23	RM	BrkTil	1Fam	Pave	N	6	4	68500
31	8500	50	?	649	1920	2008-07-19	C	BrkTil	1Fam	Pave	N	4	4	40000
32	8544	?	?	1228	1966	2008-06-22	RL	CBlock	1Fam	Pave	Y	6	5	149350

5-<u>Tratamiento del campo *DtSold* de tipo *Date*</u>: en un proceso analítico puede ser beneficioso tener la fecha desglosada para poder agrupar por años o meses. Para tratar las fechas se puede hacer como tipo fecha o como *string*.

<u>Tratamiento como fecha</u>:

- Se convierte el *string* a tipo de dato *Date*
- De *DtSold* (fecha) se obtiene la parte del año y del mes (como número y como nombre)

Quedando de la siguiente manera:



Output table - 0:36 - Extract Date&Time Fields

File Edit Hilite Navigation View

Table "default" - Rows: 1461 Spec - Columns: 17 Properties Flow Variables

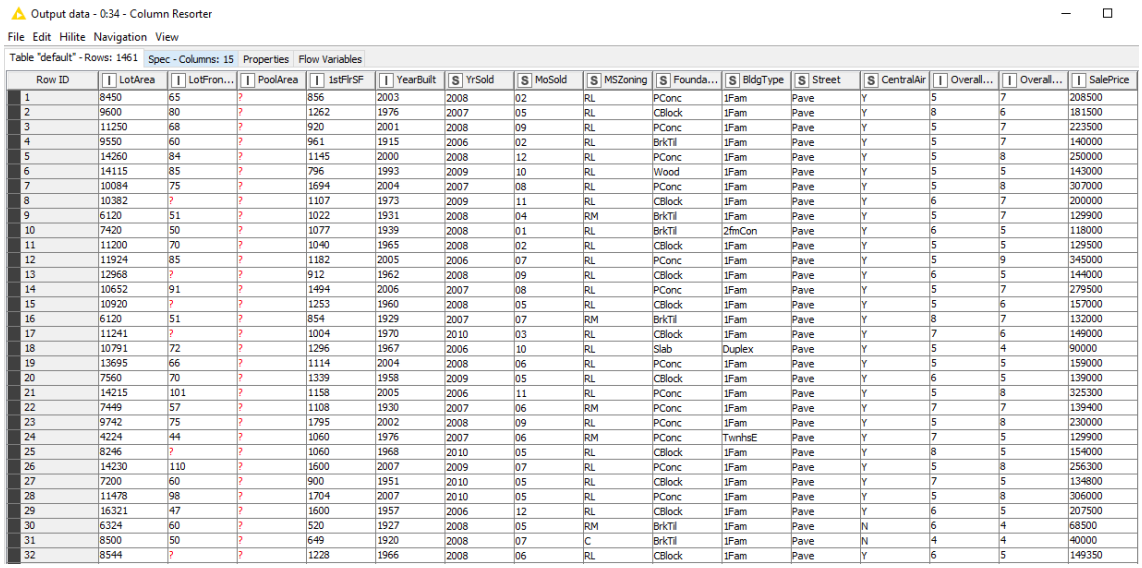
Row ID	LotArea	LotFron...	PoolArea	1stFlrSF	YearBuilt	DtSold	MSZoning	Founda...	BldgType	Street	CentralAir	Overall...	Overall...	SalePrice	Year	Month (...	Month (...
1	8450	65	?	856	2003	2008-02-14	RL	PConc	1Fam	Pave	Y	5	7	208500	2008	2	febrero
2	9600	80	?	1262	1976	2007-05-24	RL	CBlock	1Fam	Pave	Y	8	6	181500	2007	5	mayo
3	11250	68	?	920	2001	2008-09-12	RL	PConc	1Fam	Pave	Y	5	7	223500	2008	9	septiembre
4	9550	60	?	961	1915	2006-02-18	RL	BrkTil	1Fam	Pave	Y	5	7	140000	2006	2	febrero
5	14260	84	?	1145	2000	2008-12-13	RL	PConc	1Fam	Pave	Y	5	8	250000	2008	12	diciembre
6	14115	85	?	796	1993	2009-10-25	RL	Wood	1Fam	Pave	Y	5	5	143000	2009	10	octubre
7	10084	75	?	1694	2004	2007-08-14	RL	PConc	1Fam	Pave	Y	5	8	307000	2007	8	agosto
8	10382	?	?	1107	1973	2009-11-11	RL	CBlock	1Fam	Pave	Y	6	7	200000	2009	11	noviembre
9	6120	51	?	1022	1931	2008-04-25	RM	BrkTil	1Fam	Pave	Y	5	7	129900	2008	4	abril
10	7420	50	?	1077	1939	2008-01-28	RL	BrkTil	2fmCon	Pave	Y	6	5	118000	2008	1	enero
11	11200	70	?	1040	1965	2008-02-25	RL	CBlock	1Fam	Pave	Y	5	5	129500	2008	2	febrero
12	11924	85	?	1182	2005	2006-07-25	RL	PConc	1Fam	Pave	Y	5	9	345000	2006	7	julio
13	12968	?	?	912	1962	2008-09-10	RL	CBlock	1Fam	Pave	Y	6	5	144000	2008	9	septiembre
14	10652	91	?	1494	2006	2007-08-22	RL	PConc	1Fam	Pave	Y	5	7	279500	2007	8	agosto
15	10920	?	?	1253	1960	2008-05-12	RL	CBlock	1Fam	Pave	Y	5	6	157000	2008	5	mayo
16	6120	51	?	854	1929	2007-07-11	RM	BrkTil	1Fam	Pave	Y	8	7	132000	2007	7	julio
17	11241	?	?	1004	1970	2010-03-17	RL	CBlock	1Fam	Pave	Y	7	6	149000	2010	3	marzo
18	10791	72	?	1296	1967	2006-10-13	RL	Slab	Duplex	Pave	Y	5	4	90000	2006	10	octubre
19	13695	66	?	1114	2004	2008-06-19	RL	PConc	1Fam	Pave	Y	5	5	159000	2008	6	junio
20	7560	70	?	1339	1958	2009-05-18	RL	CBlock	1Fam	Pave	Y	6	5	139000	2009	5	mayo
21	14215	101	?	1158	2005	2006-11-21	RL	PConc	1Fam	Pave	Y	5	8	325300	2006	11	noviembre
22	7449	57	?	1108	1930	2007-06-20	RM	PConc	1Fam	Pave	Y	7	7	139400	2007	6	junio
23	9742	75	?	1795	2002	2008-09-14	RL	PConc	1Fam	Pave	Y	5	8	230000	2008	9	septiembre
24	4224	44	?	1060	1976	2007-06-15	RM	PConc	TwnhsE	Pave	Y	7	5	129900	2007	6	junio
25	8246	?	?	1060	1968	2010-05-21	RL	CBlock	1Fam	Pave	Y	8	5	154000	2010	5	mayo
26	14230	110	?	1600	2007	2009-07-18	RL	PConc	1Fam	Pave	Y	5	8	256300	2009	7	julio
27	7200	60	?	900	1951	2010-05-24	RL	CBlock	1Fam	Pave	Y	7	5	134800	2010	5	mayo
28	11478	98	?	1704	2007	2010-05-20	RL	PConc	1Fam	Pave	Y	5	8	306000	2010	5	mayo
29	16321	47	?	1600	1957	2006-12-11	RL	CBlock	1Fam	Pave	Y	6	5	207500	2006	12	diciembre
30	6324	60	?	520	1927	2008-05-23	RM	BrkTil	1Fam	Pave	N	6	4	68500	2008	5	mayo
31	8500	50	?	649	1920	2008-07-19	C	BrkTil	1Fam	Pave	N	4	4	40000	2008	7	julio
32	8544	?	?	1228	1966	2008-06-22	RL	CBlock	1Fam	Pave	Y	6	5	149350	2008	6	junio

<u>Tratamiento como string:</u>

- Se divide la columna *DtSold* en tres partes, usando el separador '-' para dividir, creandose tres columnas tipo string (año, mes y día).
- Se renombran las columnas que contienen el año y el mes.

- Se quita la columna con la parte dia de la fecha de venta, con el objetivo de facilitar la agrupación posterior a nivel mensual.
- Se reordenan las nuevas variables de modo que *YrSold* y *MoSold* estén entre *YearBuilt* y *MSZoning*.

Quedando de la siguiente manera:


Output data - 0:34 - Column Resorter

File Edit Hilite Navigation View

Table "default" - Rows: 1461 | Spec - Columns: 15 | Properties | Flow Variables

Row ID	LotArea	LotFron...	PoolArea	1stFlrSF	YearBuilt	YrSold	MoSold	MSZoning	Founda...	BldgType	Street	CentralAir	Overall...	Overall...	SalePrice
1	8450	65	?	856	2003	2008	02	RL	PConc	1Fam	Pave	Y	5	7	208500
2	9600	80	?	1262	1976	2007	05	RL	CBlock	1Fam	Pave	Y	8	6	181500
3	11250	68	?	920	2001	2008	09	RL	PConc	1Fam	Pave	Y	5	7	223500
4	9550	60	?	961	1915	2006	02	RL	BrkTil	1Fam	Pave	Y	5	7	140000
5	14260	84	?	1145	2000	2008	12	RL	PConc	1Fam	Pave	Y	5	8	250000
6	14115	85	?	796	1993	2009	10	RL	Wood	1Fam	Pave	Y	5	5	143000
7	10084	75	?	1694	2004	2007	08	RL	PConc	1Fam	Pave	Y	5	8	307000
8	10382	?	?	1107	1973	2009	11	RL	CBlock	1Fam	Pave	Y	6	7	200000
9	6120	51	?	1022	1931	2008	04	RM	BrkTil	1Fam	Pave	Y	5	7	129900
10	7420	50	?	1077	1939	2008	01	RL	BrkTil	2fmCon	Pave	Y	6	5	118000
11	11200	70	?	1040	1965	2008	02	RL	CBlock	1Fam	Pave	Y	5	5	129500
12	11924	85	?	1182	2005	2006	07	RL	PConc	1Fam	Pave	Y	5	9	345000
13	12968	?	?	912	1962	2008	09	RL	CBlock	1Fam	Pave	Y	6	5	144000
14	10652	91	?	1494	2006	2007	08	RL	PConc	1Fam	Pave	Y	5	7	279500
15	10920	?	?	1253	1960	2008	05	RL	CBlock	1Fam	Pave	Y	5	6	157000
16	6120	51	?	854	1929	2007	07	RM	BrkTil	1Fam	Pave	Y	8	7	132000
17	11241	?	?	1004	1970	2010	03	RL	CBlock	1Fam	Pave	Y	7	6	149000
18	10791	72	?	1296	1967	2006	10	RL	Slab	Duplex	Pave	Y	5	4	90000
19	13695	66	?	1114	2004	2008	06	RL	PConc	1Fam	Pave	Y	5	5	159000
20	7560	70	?	1339	1958	2009	05	RL	CBlock	1Fam	Pave	Y	6	5	139000
21	14215	101	?	1158	2005	2006	11	RL	PConc	1Fam	Pave	Y	5	8	325300
22	7449	57	?	1108	1930	2007	06	RM	PConc	1Fam	Pave	Y	7	7	139400
23	9742	75	?	1795	2002	2008	09	RL	PConc	1Fam	Pave	Y	5	8	230000
24	4224	44	?	1060	1976	2007	06	RM	PConc	TwnhsE	Pave	Y	7	5	129900
25	8246	?	?	1060	1968	2010	05	RL	CBlock	1Fam	Pave	Y	8	5	154000
26	14230	110	?	1600	2007	2009	07	RL	PConc	1Fam	Pave	Y	5	8	256300
27	7200	60	?	900	1951	2010	05	RL	CBlock	1Fam	Pave	Y	7	5	134800
28	11478	98	?	1704	2007	2010	05	RL	PConc	1Fam	Pave	Y	5	8	306000
29	16321	47	?	1600	1957	2006	12	RL	CBlock	1Fam	Pave	Y	6	5	207500
30	6324	60	?	520	1927	2008	05	RM	BrkTil	1Fam	Pave	N	6	4	68500
31	8500	50	?	649	1920	2008	07	C	BrkTil	1Fam	Pave	N	4	4	40000
32	8544	?	?	1228	1966	2008	06	RL	CBlock	1Fam	Pave	Y	6	5	149350

6-Detección y filtrado de duplicados: se eliminan registros duplicados ya sea para encontrar varios registros exactamente iguales en el dataset, debido a problemas en el diseño y consistencia de los datos, o bien para localizar determinados subconjuntos de datos con condiciones que se repitan dentro de una misma tabla.

7-Tratamiento Nulos y Outliers[20]: para agilizar esta tarea es bueno contar primero con distintos estadísticos resumen sobre nulos y outliers.

Para tratar los **valores nulos** se puede optar por distintas opciones dependiendo del tipo de dato y su significado. Se podría por ejemplo asignar el valor más frecuente, filtrar filas, si fuera numérico la media, moda, mediana, rellenar con un código comodín como el 0 para enteros o N para string, etc. En este dataset se aprecia que las variables *LotFrontage* y *PoolArea*, de tipo *integer* (entero) y informan de una subárea y de la supercifie, son las que contienen valores nulos. Dada la naturaleza de su contenido se decide realizar lo siguiente en cada una:

- *LotFrontage:* rellenar con el valor más frecuente
- *PoolArea:* rellenar con 0

Para tratar los **valores inusuales (*outliers*)** de cada una de las columnas se utiliza el rango intercuartílico (IQR), calculando el primer y tercer cuartil (Q1 y Q3) y marcando como valor atípico el que se encuentra fuera del rango R = [Q1 – k (IQR), Q3 + k (IQR)] donde IQR = Q3-Q1 y k>=0. En KNIME por defecto se establece el valor del multiplicador de rango intercuartílico como k=1.5, siendo el valor más pequeño en R correspondiente al extremo inferior del bigote de un diagrama de caja y el valor más grande a su extremo superior.

Quedando de la siguiente manera:

Treated table - 0:12 - Numeric Outliers

File Edit Hilite Navigation View

Table "default" - Rows: 1460 Spec - Columns: 15 Properties Flow Variables

Row ID	LotArea	LotFron...	PoolArea	1stFlrSF	YearBuilt	YrSold	MoSold	MSZoning	Founda...	BldgType	Street	CentralAir	Overall...	Overall...	SalePrice
1	1450	32	0	561	1980	2009	05	RM	CBlock	TwnhsE	Pave	Y	6	6	124000
2	1477	32	0	630	1970	2009	04	RM	CBlock	TwnhsE	Pave	Y	4	4	80000
3	1491	32	0	630	1972	2010	05	RM	CBlock	TwnhsE	Pave	Y	6	4	75500
4	1526	32	0	630	1970	2009	05	RM	CBlock	Twnhs	Pave	Y	7	4	86000
5	1533	32	0	630	1970	2006	08	RM	CBlock	Twnhs	Pave	Y	7	5	92000
6	1533	32	0	798	1970	2009	05	RM	CBlock	Twnhs	Pave	Y	6	4	97000
7	1596	32	0	526	1973	2009	11	RM	CBlock	Twnhs	Pave	Y	5	4	91000
8	1680	32	0	483	1971	2008	07	RM	CBlock	Twnhs	Pave	Y	5	6	85400
9	1680	32	0	483	1971	2008	08	RM	CBlock	TwnhsE	Pave	Y	4	6	89500
10	1680	32	0	483	1972	2006	06	RM	CBlock	Twnhs	Pave	Y	5	6	94500
11	1680	32	0	483	1972	2009	05	RM	CBlock	Twnhs	Pave	Y	7	6	118000
12	1680	32	0	483	1973	2008	11	RM	CBlock	Twnhs	Pave	Y	5	6	100000
13	1680	32	0	525	1971	2010	03	RM	CBlock	Twnhs	Pave	Y	5	6	88000
14	1680	32	0	630	1971	2009	05	RM	CBlock	Twnhs	Pave	Y	5	5	112000
15	1680	32	0	630	1972	2009	02	RM	CBlock	Twnhs	Pave	Y	7	5	119500
16	1680	32	0	672	1971	2006	04	RM	CBlock	Twnhs	Pave	Y	5	6	91500
17	1680	32	0	672	1972	2006	05	RM	CBlock	Twnhs	Pave	Y	7	6	118000
18	1869	32	0	483	1970	2008	09	RM	CBlock	Twnhs	Pave	Y	6	6	106000
19	1890	32	0	630	1972	2008	06	RM	CBlock	Twnhs	Pave	Y	7	4	81000
20	1890	32	0	672	1973	2007	05	RM	CBlock	Twnhs	Pave	Y	5	6	113000
21	1920	32	0	765	1971	2007	08	RM	CBlock	TwnhsE	Pave	Y	5	5	122500
22	1936	32	0	630	1970	2007	12	RM	CBlock	Twnhs	Pave	Y	6	4	84500
23	1950	32	0	716	1980	2008	07	RM	CBlock	Twnhs	Pave	Y	6	6	151000
24	1953	32	0	483	1973	2006	06	RM	CBlock	Twnhs	Pave	Y	5	6	83000
25	1974	60	0	546	1973	2010	05	RM	CBlock	TwnhsE	Pave	Y	5	4	83500
26	2001	32	0	546	1970	2007	01	RM	CBlock	Twnhs	Pave	Y	5	4	75000
27	2016	32	0	630	1970	2007	04	RM	CBlock	TwnhsE	Pave	Y	5	5	106000
28	2117	60	0	756	2000	2007	06	FV	PConc	TwnhsE	Pave	Y	5	6	168500
29	2117	60	0	769	2000	2007	06	FV	PConc	Twnhs	Pave	Y	5	6	177000
30	2117	60	0	769	2000	2010	06	FV	PConc	Twnhs	Pave	Y	5	6	177500
31	2160	32	0	624	1999	2008	03	FV	PConc	Twnhs	Pave	Y	5	7	160000
32	2217	32	0	546	1970	2009	08	RM	CBlock	TwnhsE	Pave	Y	4	4	88000

8-Exportación a csv: para pasar al siguiente nivel y diferenciar los distintos flujos se deben exportar los datos perfilados a csv y con ellos comenzar con el análisis exploratorio (EDA). El nuevo archivo se denomina "*housing_price.csv*".

2.1.2 Enriquecimiento y transformación. Definición

En este punto se intenta mostrar como después de un buen perfilado y tratamiento de los datos iniciales (antes del analisis exploratorio) ya se puede obtener información e incluso dislumbrar conocimiento.

Esta fase tiene como objetivo permitir acceder a los datos de una forma más organizada y ágil, esto se consigue mediante la conversión y el mapeado a otro formato de los datos, facilitando además su comprensión (los puntos 5, 6 y 7 del apartado 2.1.1 podrían entrar también en el desarrollo de este apartado).

Como ejemplo de esta fase, tomando como partida los datos ya perfilados en el punto anterior, se va a proporcionar un informe con el total de ventas y precio de venta medio agrupado por zonas, campo MSZoning. Se dará una descripción a los códigos de este campo, mediante el fichero "*lookup_MSZoning.csv*", para que se crucen con el dataset original y se muestren en el reporte. Tomando como partida los Los pasos serán los siguientes:

- Leer csv con la relación entre las siglas y la descripción de las zonas (campo *MSZoning*)
- Unir el dataset inicial perfilado con el dataset que contine la descripción de las zonas
- Se agrupa por la descripción de la zona, calculando el número de ventas y el precio medio desde el campo *SalesPrice*
- Se renombran las nuevas columnas con las métricas
- Se redondean los valores de las métricas

Quedando la siguiente tabla:

Row ID	S MSZoning	D NºVentas	D PrecioMedio
Row0	Commercial	10	74,528
Row1	Floating Village Residential	65	213,541
Row2	Residential High Density	16	131,558
Row3	Residential Low Density	1,151	186,647
Row4	Residential Medium Density	218	125,458

El flujo1 ejecutado quedaría de la siguiente manera:

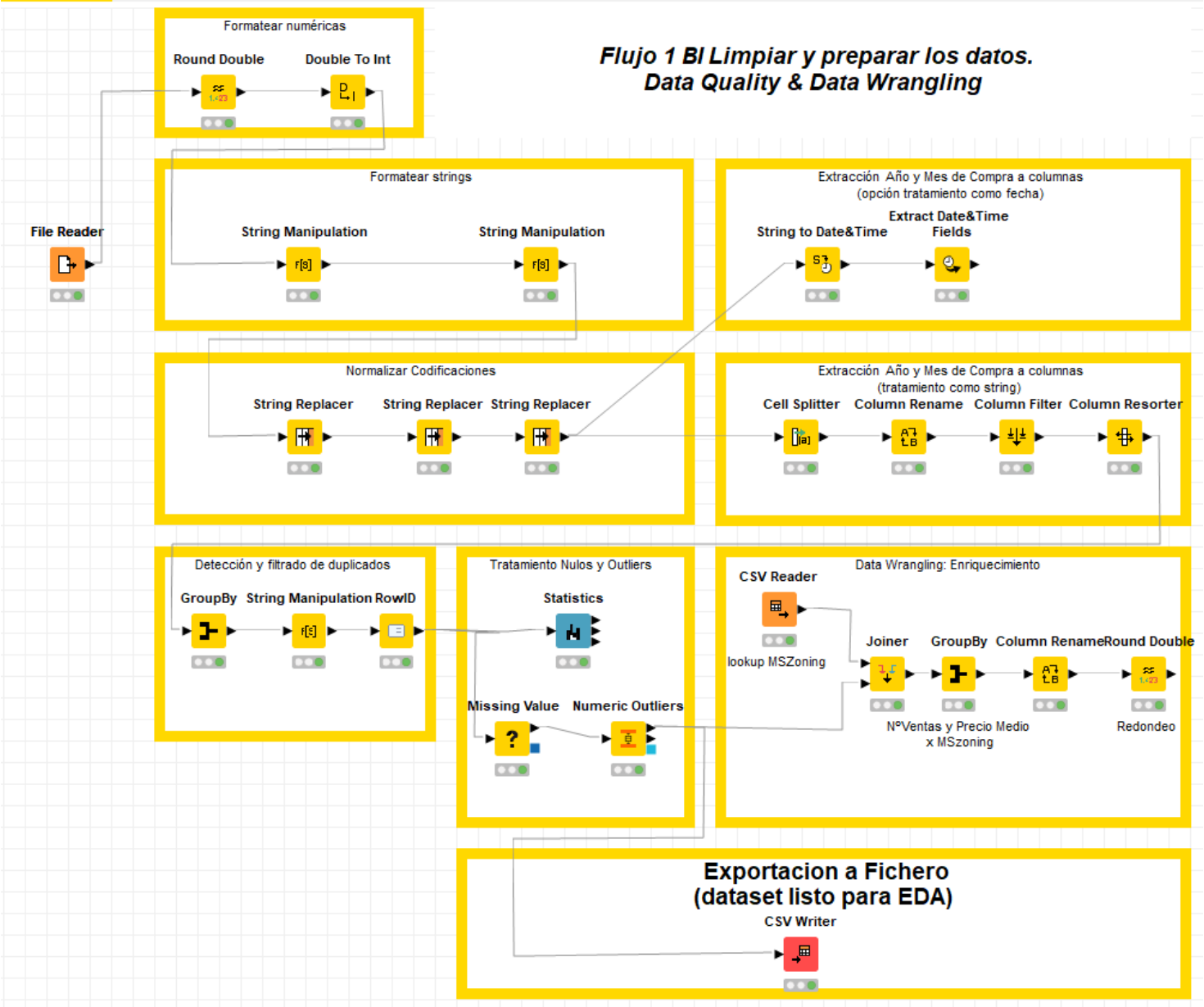


Ilustración 6 Flujo1 Limpiar y preparar los datos. Data Quality & Data Wrangling

2.2 Flujo EDA, análisis exploratorio. Definición

El análisis exploratorio de datos[22] consiste en el tratamiento estadístico al que se someten las muestras recogidas durante un proceso de investigación. El conocido estadístico John W. Tukey[23], en 1977, lo definía como "una actitud, un estado de flexibilidad, una voluntad de buscar aquellas cosas que creemos que no están allí, también aquellas que creemos que sí están"[24].

Para hacer una buena descripción de la matriz de datos se deben considerar dos pasos básicos:

- Medición y descripción, aplicando la estadística descriptiva[25]
- Comparación, aplicando la estadística inferencial[26]

El ciclo iterativo del EDA[27] se puede diferenciar en tres fases:

- Generar preguntas acerca de los datos
- Buscar respuestas y examinar la calidad de los datos visualizando, transformando y modelándolos
- Usar lo aprendido para refinar las preguntas y/o generar nuevos interrogantes

Con esto se consigue ganar intuición sobre los datos, detectar valores atípicos, extraer variables importantes; y en general, descubrir estructuras subyacentes en los datos. Permite, además, organizar los datos, detectar fallos, y evaluar la existencia de datos ausentes. Todo, en gran parte, a técnicas gráficas y descriptivas simples, como:

- Representación de datos sin procesar (datos crudos) mediante histogramas, gráficos de barras, gráficos de probabilidades, diagramas de caja, entre otros.
- Trazado de estadísticos sencillos, como: gráficos de media y desviación estándar, diagramas de cajas, entre otros.

Para realizar el análisis exploratorio se parte de los datos perfilados que se encuentran en el fichero "*housing_price.csv*" proporcionado al aplicar al dataset original el flujo1 de limpieza y preparación de los datos. Se evalúa la estructura estadística de los datos que contiene y se selecciona el modelo adecuado para obtener información acerca de los diferentes perfiles poblacionales, esto es, qué tipo de clientes existen, sus rasgos principales y el número de cada tipo.

Las etapas del EDA, previas a la selección de la técnica de machine learning a aplicar a los datos, que se van a seguir en este flujo son:

- Preparar los datos para el estudio estadístico (Flujo 1 descrito en el punto 2.1)
- Realizar análisis gráfico y descriptivo univariante
- Realizar análisis gráfico y descriptivo multivariante
- Evaluar algunos supuestos o hipótesis como normalidad o linealidad
- Identificar la presencia de valores extremos, *outliers*
- Valorar el impacto de los valores ausentes (*missing values*) sobre la representatividad de los datos

2.2.1 Análisis descriptivo y gráfico

El análisis descriptivo de datos es la disciplina estadística que consiste en la descripción cuantitativa de las principales características de un conjunto de datos. Comprende técnicas gráficas y numéricas (estadísticos) que permiten organizar y resumir convenientemente la información contenida en grandes volúmenes de datos. Hay dos tipos:

- Univariante: corresponde al estudio del comportamiento de una variable característica, o atributo, de forma individual. Las técnicas de análisis univariante se agrupan en tres categorías:
    - Medidas de centralidad (media, mediana, moda, cuartiles)
    - Medidas de dispersión (rango, varianza, desviación típica)
    - Medidas de forma (asimetría, curtosis)
- Multivariante: son métodos que estudian las relaciones que hay entre variables tomadas de dos en dos (bivariante) o más (multivariante), como medidas de correlación, tabulación cruzada etc.

El análisis exploratorio gráfico de datos es el conjunto de herramientas que permiten organizar y representar los datos de forma gráfica. A diferencia del análisis descriptivo mediante estadísticos, el análisis exploratorio gráfico es intuitivo y tiene como finalidad extraer información cualitativa de los datos.

En el análisis exploratorio de datos los métodos descriptivos y gráficos son complementarios y por lo tanto se recomienda el uso de ambos.

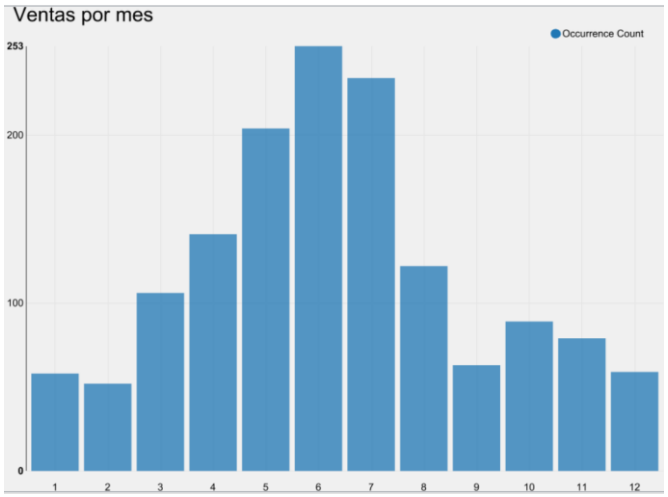
Mientras el objetivo de EDA es ganar intuición sobre los datos, detectar errores, outliers, etc; el objetivo fundamental de la etapa de visualización es comunicar los resultados obtenidos y servir de soporte en la toma de decisiones.

2.2.1.1 Análisis exploratorio gráfico univariante

Para realizar este análisis, siguiendo el ciclo iterativo del EDA definido anteriormente, se proponen las siguientes preguntas y métricas/grafos iniciales:

- ¿Cuántas ventas se realizan por mes? Tabla y gráfico de barras:

Row ID	count
1	58
2	52
3	106
4	141
5	204
6	253
7	234
8	122
9	63
10	89
11	79
12	59



Se aprecia que en verano aumentan las ventas con respecto a los demás periodos del año, siendo junio el mejor mes y febrero el peor.

- Obtener estadísticos básicos de las variables numéricas, como valor máximo, mínimo, media, mediana, varianza, valores nulos (missings), asimetría de la distribución de probabilidad de cada variable (skewness) y la presencia de valores atípicos en la distribución (kurtosis)[29], etc.

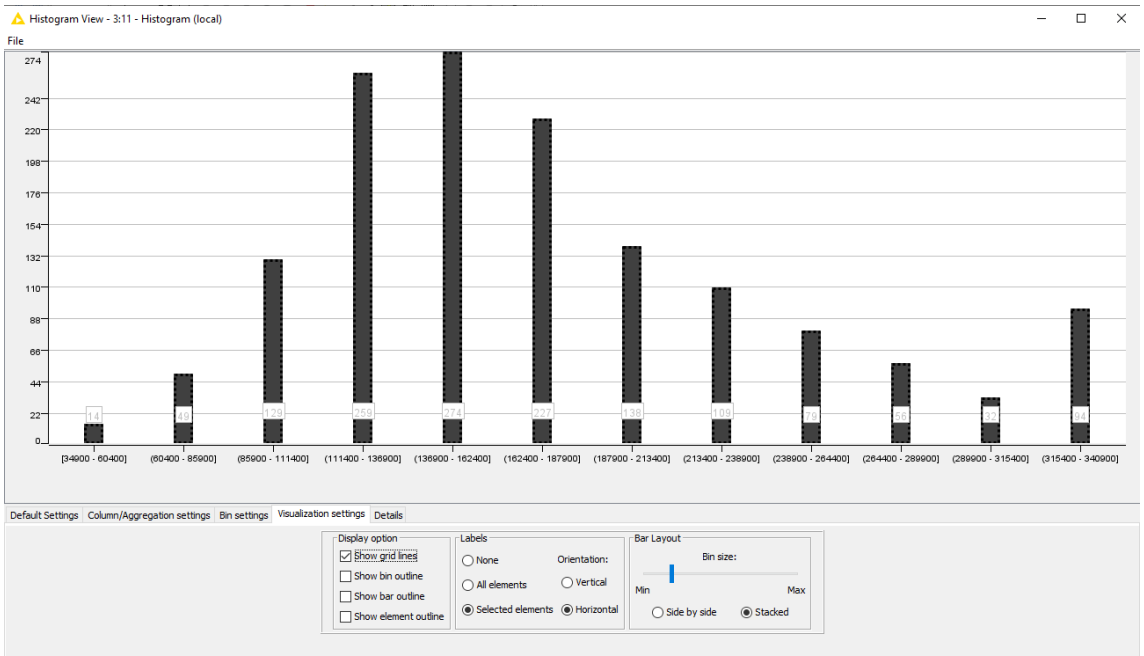
Column	Min	Max	Mean	Std. deviation	Variance	Skewness	Kurtosis	Overall sum	No. missings	No. NaNs	No. +oos	No. -oos	Median	Row count	Histogram
row ID	1	1,460	730.5	421.61	177,755	0	-1.2	1,066,530	0	0	0	0	730.5	1460	
LotArea	1,450	17,690	9,648.121	3,596.106	12,931,97...	0.211	0.087	14,086,256	0	0	0	0	9,478.5	1460	
LotFrontage	32	107	67.49	17.501	306.299	0.337	0.013	98,536	0	0	0	0	63	1460	
PoolArea	0	0	0	0	0	0	0	0	0	0	0	0	0	1460	
1stFlrSF	334	2,154	1,157.003	362.541	131,435.666	0.664	-0.064	1,689,224	0	0	0	0	1,087	1460	
YearBuilt	1,885	2,010	1,971.299	30.108	906.516	-0.594	-0.519	2,878,097	0	0	0	0	1,973	1460	
YrSold	2,006	2,010	2,007.816	1.328	1.764	0.096	-1.191	2,931,411	0	0	0	0	2,008	1460	
MoSold	1	12	6.322	2.704	7.31	0.212	-0.404	9,230	0	0	0	0	6	1460	1 12
OverallCond	4	7	5.522	0.883	0.78	0.564	-0.775	8,062	0	0	0	0	5	1460	4 7
OverallQual	2	10	6.101	1.378	1.9	0.25	-0.014	8,907	0	0	0	0	6	1460	2 10
SalePrice	34,900	340,150	177,336...	67,217.227	4,518,155,...	0.803	0.089	258,910,890	0	0	0	0	163,000	1460	

Si lo enfocamos al precio de venta, *SalePrice*, y al tamaño de la muestra del *dataset*, se destaca lo siguiente:

- No hay datos nulos ni en blanco (campos *No. Missings* y *No NaNs*).
- Existe una asimetría en los valores del precio de venta de tipo positivo sesgado o sesgado a la derecha (valor positivo de la columna *Skewness*), el valor concreto, 0,803 está dentro del rango que indica que los datos están ligeramente sesgados, pero no extremadamente. Además, se deduce y se ve en otras columnas que la media es superior a la mediana y la moda el valor más alto.
- La columna *Kurtosis* contiene el valor 0.089 indicando que la grafica de la distribución de los valores de la muestra es más plana que la distribución normal.

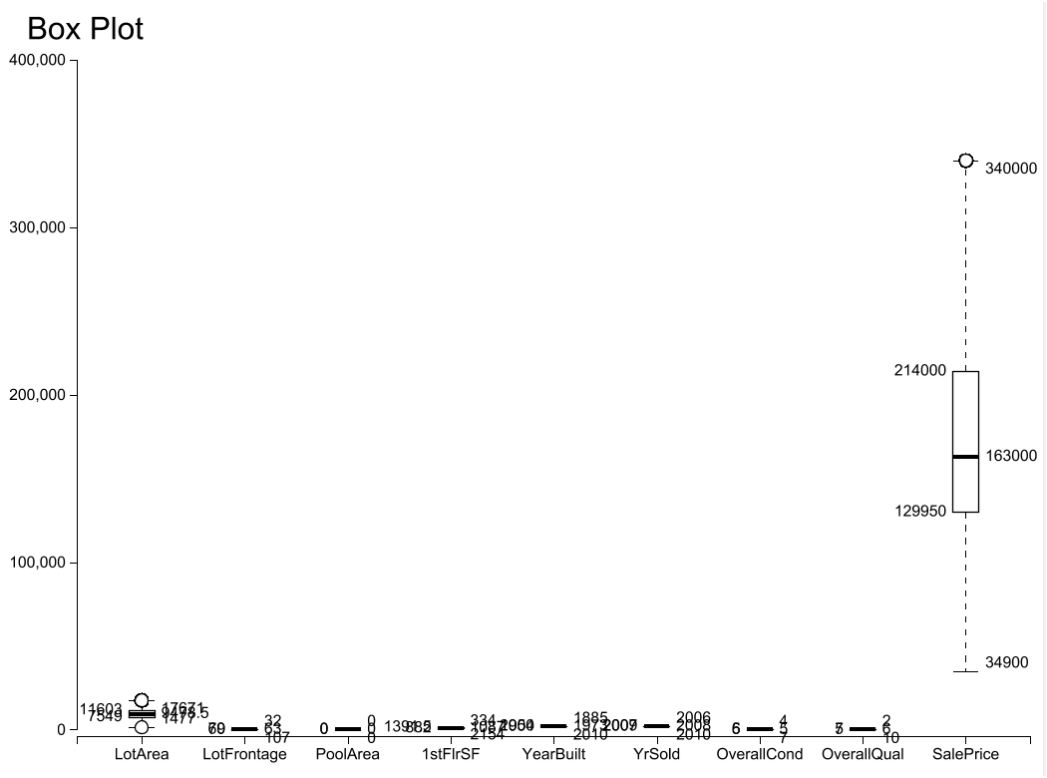
Estos dos últimos valores (*Skewness* y *Kurtosis*) dan pie a pensar que los valores de cola pueden detectarse como atípicos para le modelo estadístico, por ejemplo, para modelos basados en regresión afecta de manera negativa al rendimiento, mientras que para modelos basados en árboles no ya que son más resistentes a este tipo de valores (los valores atípicos). La solución sería transformar los datos sesgados para que se acerquen a una distribución normal.

- Histogramas[30], por ejemplo, siguiendo con la variable *SalePrice* del *dataset* se realizan agrupaciones por rango de precios, quedando de la siguiente manera:



Se puede apreciar que el rango de precios más común es entre 136900 y 162400 donde hay 274 ventas.

- Boxplot[31], permite ver los grupos de datos numéricos a través de los cuartiles de cada variable y los posibles outliers en función del rango intercuartílico (ver punto 7 del apartado 2.1.1 de este documento para conocer como se calcula). Este tipo de diagramas de caja se pueden mostrar en paralelo (en el mismo gráfico) siempre que compartan la misma dimensión de valores, sino es ineficiente ya que ocurre lo siguiente:

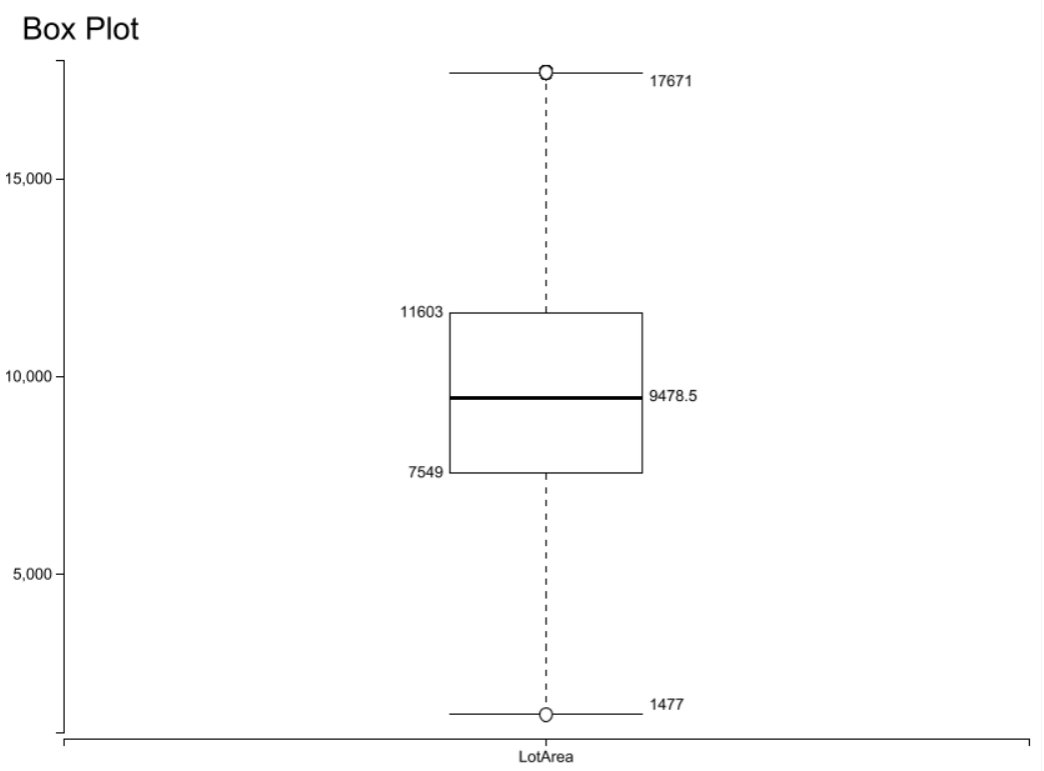


Sin embargo, los datos en tabla si son útiles:

Row ID	LotArea	LotFron...	PoolArea	1stFlrSF	YearBuilt	YrSold	Overall...	Overall...	SalePrice
Minimum	1,450	32	0	334	1,885	2,006	4	2	34,900
Smallest	1,477	32	0	334	1,885	2,006	4	2	34,900
Lower Quartile	7,549	60	0	882	1,954	2,007	5	5	129,950
Median	9,478.5	63	0	1,087	1,973	2,008	5	6	163,000
Upper Quartile	11,603	79	0	1,391.5	2,000	2,009	6	7	214,000
Largest	17,671	107	0	2,154	2,010	2,010	7	10	340,000
Maximum	17,690	107	0	2,154	2,010	2,010	7	10	340,150

Por tanto, si no hay una misma dimensión de valores en las variables lo mejor es hacerlo de forma individual, por ejemplo:

Diagrama de caja de la variable *LotArea*:



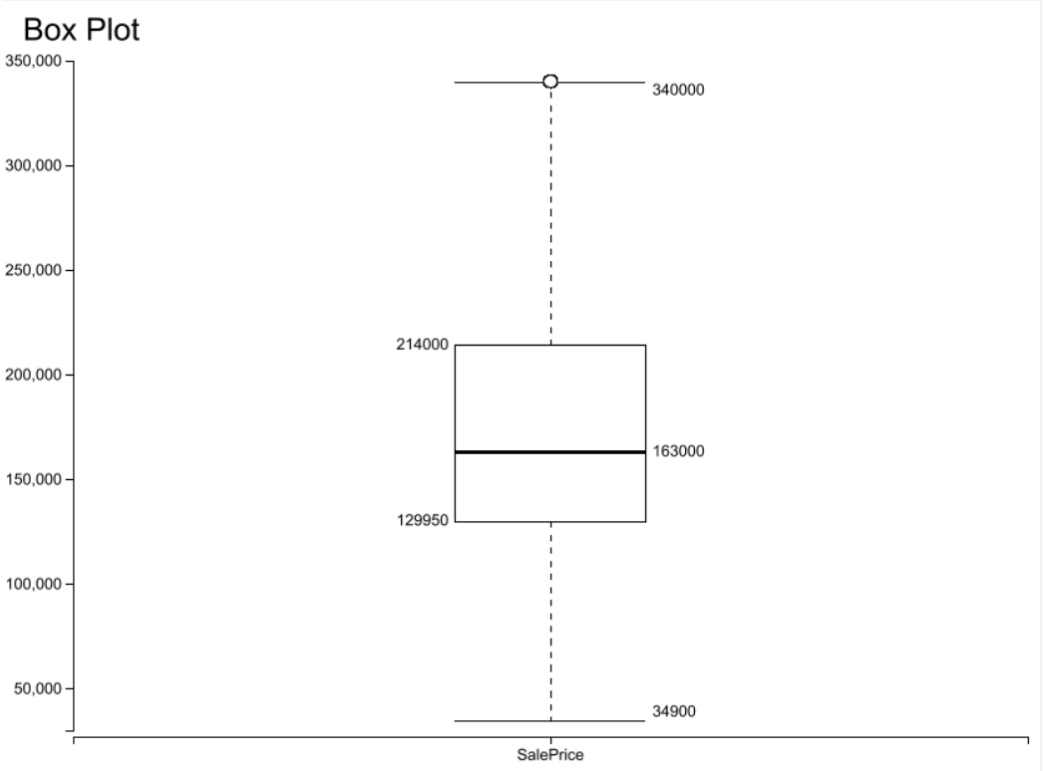
Donde se aprecia lo siguiente:

- No hay valores atípicos (outliers) ya que todos los valores están dentro de los bigotes del gráfico.
- El valor mínimo (Q_0) es 1477
- El valor máximo (Q_4) es 17671
- El valor de la mediana (Q_2) es 9478,5
- El valor del primer cuartil (Q_1) es 7549
- El valor del tercer cuartil es (Q_2) 11603

Se puede deducir que hay una ligera asimetría positiva o segada a la derecha ya que la parte más larga de la caja es la parte superior a la mediana, pero es muy leve.

Para la variable *SalePrice*, además del diagrama de caja, se calculan los ouliers, considerando lo indicado en el punto 7 del apartado 2.1.1 de este documento, quedando lo siguiente:

S Outlier ...	I Membe...	I Outlier count	D Lower bound	D Upper bound
SalePrice	1460	0	3,750	340,150



Donde se aprecia que no hay outlier según la configuración de los parámetros de cálculo del rango intercuartílico y, además, se corrobora lo deducido a la hora de interpretar los datos estadísticos básicos, hay una ligera asimetría positiva o segada a la derecha ya que la parte más larga de la caja es la parte superior a la mediana, pero es muy leve.

2.2.1.2 Análisis exploratorio gráfico multivariante

Para realizar este análisis, siguiendo el ciclo iterativo del EDA definido anteriormente, se proponen las siguientes métricas y grafos:

- Matriz o mapa de correlaciones[32], representa la relación estadística, causal o no, entre dos o más variables aleatorias, representándose de forma lineal.

  Las correlaciones son útiles porque pueden indicar una relación predictiva que puede explotarse en la práctica, aunque estas no implican causalidad, por ejemplo, un productor de energía puede generar menos electricidad en días templados según la correlación entre la demanda eléctrica y el clima.

  Hay varios coeficientes que miden el grado de correlación entre las variables, los dos más comunes son:

  - Coeficiente de correlación de Pearson[33]: sensible solo a una relación lineal entre dos variables cuantitativas.
  - Coeficiente de correlación de rango de Spearman[34]: son más robustos que los de Pearson, siendo más sensibles a las relaciones no lineales.

A continuación, se generarán los mapas de correlación aplicando los dos coeficientes y se explican los resultados para entender sus bondades:

Matriz de correlación de Pearson: mide linealmente el comportamiento de dos variables numéricas. Aplicado a las variables numéricas del dataset se obtienen las siguientes medidas:

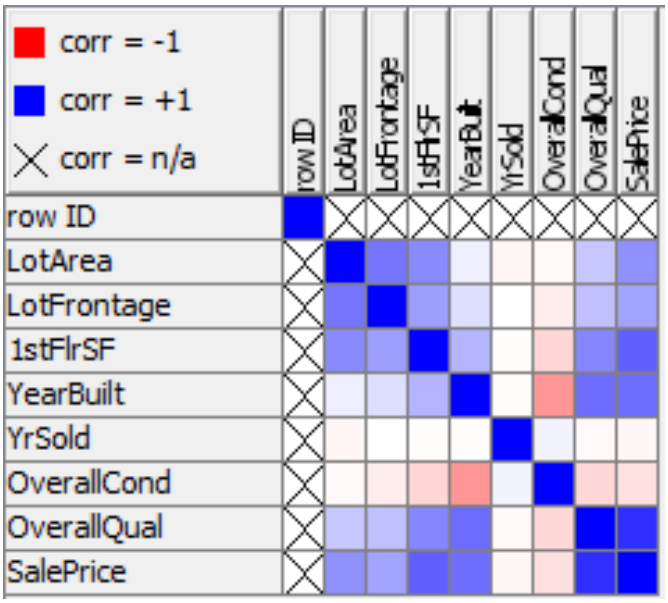
First column name	Second column name	Correlation value	p value	Degrees of freedom
LotArea	LotFrontage	0.5395991935300973	0.0	1458
LotArea	1stFlrSF	0.45808001112462077	0.0	1458
LotArea	YearBuilt	0.06568889158859455	0.01205497...	1458
LotArea	YrSold	-0.03685410136250602	0.15928947...	1458
LotArea	OverallCond	-0.027610569199836708	0.29174753...	1458
LotArea	OverallQual	0.2215210142371321	0.0	1458
LotArea	SalePrice	0.4358058656332907	0.0	1458
LotFrontage	1stFlrSF	0.38047278112752797	0.0	1458
LotFrontage	YearBuilt	0.12835079380084846	8.63496180...	1458
LotFrontage	YrSold	0.0022387302514464565	0.93188885...	1458
LotFrontage	OverallCond	-0.06946325004018311	0.00792798...	1458
LotFrontage	OverallQual	0.24654838166602155	0.0	1458
LotFrontage	SalePrice	0.36406253902536406	0.0	1458
1stFlrSF	YearBuilt	0.28843128745045876	0.0	1458
1stFlrSF	YrSold	-0.013547868859551549	0.60498605...	1458
1stFlrSF	OverallCond	-0.15894789736151094	1.01488226...	1458
1stFlrSF	OverallQual	0.47392419129009655	0.0	1458
1stFlrSF	SalePrice	0.6219046231261974	0.0	1458
YearBuilt	YrSold	-0.013292322687628004	0.61181361...	1458
YearBuilt	OverallCond	-0.41060954676096134	1.76413451...	1458
YearBuilt	OverallQual	0.5751598628971031	0.0	1458
YearBuilt	SalePrice	0.570293084814646	0.0	1458
YrSold	OverallCond	0.04755733882304255	0.06927304...	1458
YrSold	OverallQual	-0.02729935050553934	0.29722276...	1458
YrSold	SalePrice	-0.031357991049684224	0.23112910...	1458
OverallCond	OverallQual	-0.15577696072936445	2.18132412...	1458
OverallCond	SalePrice	-0.12334233865157934	2.27943085...	1458
OverallQual	SalePrice	0.8176765672649933	0.0	1458

Row ID	row ID	LotArea	LotFron...	1stFlrSF	YearBuilt	YrSold	Overall...	Overall...	SalePrice
row ID	1.0	?	?	?	?	?	?	?	?
LotArea	?	1.0	0.53959919...	0.45808001...	0.06568889...	-0.0368541...	-0.0276105...	0.22152101...	0.43580586...
LotFrontage	?	0.53959919...	1.0	0.38047278...	0.12835079...	0.00223873...	-0.0694632...	0.24654838...	0.36406253...
1stFlrSF	?	0.45808001...	0.38047278...	1.0	0.28843128...	-0.0135478...	-0.1589478...	0.47392419...	0.62190462...
YearBuilt	?	0.06568889...	0.12835079...	0.28843128...	1.0	-0.0132923...	-0.4106095...	0.57515986...	0.57029308...
YrSold	?	-0.0368541...	0.00223873...	-0.0135478...	-0.0132923...	1.0	0.04755733...	-0.0272993...	-0.0313579...
OverallCond	?	-0.0276105...	-0.0694632...	-0.1589478...	-0.4106095...	0.04755733...	1.0	-0.1557769...	-0.1233423...
OverallQual	?	0.22152101...	0.24654838...	0.47392419...	0.57515986...	-0.0272993...	-0.1557769...	1.0	0.81767656...
SalePrice	?	0.43580586...	0.36406253...	0.62190462...	0.57029308...	-0.0313579...	-0.1233423...	0.81767656...	1.0

Donde se aprecia que los valores oscilan en el rango [-1, 1]. Su significado es el siguiente (siendo r el grado de correlación de Pearson):

- Si r=1: correlación positiva perfecta. Refleja la dependencia total perfecta entre ambas variables, la que se denomina relación directa (cuando una de las variables aumenta la otra aumenta en proporción constante)
- Si 0<r<1: se da una correlación positiva
- Si r=0: no hay relación lineal. Esto no significa que las variables sean independientes, ya que puede haber relaciones no lineales entre ambas variables
- Si -1<r<0: se da una correlación negativa
- Si r=-1: correlación negativa perfecta. Refleja una dependencia total entre ambas variables conocida como relación inversa, que se da cuando una de las variables aumenta, la otra en cambio disminuye en proporción constante

Siendo la matriz del dataset, correspondiente a las métricas expuestas anteriormente en este punto, de la siguiente manera:



Se representa como una matriz cuadrada, con tantas filas como variables numéricas, representando la correlación de cada par con colores en escala del -1 al 1. Se aprecia que la matriz es simétrica y como tal, la relación entre las mismas variables es una correlación positiva perfecta (1 azul). Se destacan las siguientes correlaciones:

- La valoración de las condiciones del inmueble (*OverallCond*) tiene una correlación negativa (-0,41, color rojizo) con el año de construcción del inmueble (*YearBuilt*), es decir, según es mayor el año de construcción parece que las condiciones son peores
- Al contrario pasa con las calidades de los inmuebles (*OverallQual*) que tienen una correlación positiva (0,58, color azulado) con respecto al año de construcción del inmueble (*YearBuilt*), es decir, cuanto más nueva es la construcción, mejores calidades.
- La valoración de las calidades (*OverallQual*) tiene una correlación casi perfecta (0,82, color azulado) con respecto al precio de venta (*SalePrice*), es decir, cuanto más caro mejores calidades.
- La superficie de los inmuebles (*1stFlrSF*) tiene una correlación positiva (0,62) con respecto al precio del inmueble (*SalePrice*), es decir, cuantos más grande sea el inmueble más vale.
- El código de zona (LotArea) tiene una correlación positiva (0,44) con respecto al precio del inmueble (*SalePrice*), es decir, en el precio influye la zona donde se encuentra.

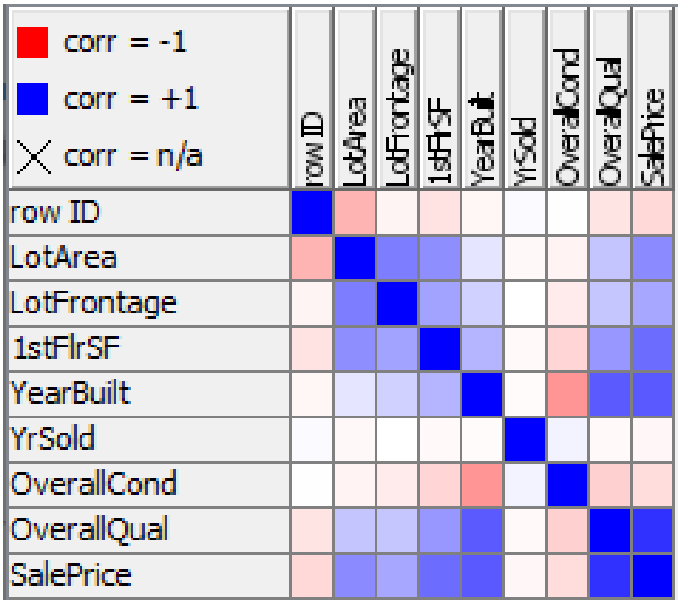
Matriz de correlación de rango de Spearman: mide la dependencia estadística entre las clasificaciones de dos variables (interdependencia de dos variables continuas), siendo igual a la correlación de Pearson entre los valores de rango de dos variables, es decir, si las observaciones de dos variables tienen un rango similar, la correlación será alta. Mientras que la correlación de Pearson evalúa relaciones lineales, la correlación de Spearman evalúa todo tipo de relaciones monotónicas (lineales o no), es decir, va más allá. Aplicado a las variables numéricas del dataset se obtienen las siguientes medidas:

Row ID	First col...	Second...	Correlation value	p value	Degree...
Row0	row ID	LotArea	-0.29393993103373...	0.0	1458
Row1	row ID	LotFrontage	-0.04355074018966...	0.09622531...	1458
Row2	row ID	1stFlrSF	-0.11357400096666...	1.36124310...	1458
Row3	row ID	YearBuilt	-0.03653344124346...	0.16295223...	1458
Row4	row ID	YrSold	0.020640772789832...	0.43064332...	1458
Row5	row ID	OverallCond	0.003555524005254...	0.89202646...	1458
Row6	row ID	OverallQual	-0.10722386434584...	4.03704490...	1458
Row7	row ID	SalePrice	-0.15152403033896...	5.94166427...	1458
Row8	LotArea	LotFrontage	0.5095286495452701	0.0	1458
Row9	LotArea	1stFlrSF	0.44348362929222757	0.0	1458
Row10	LotArea	YearBuilt	0.1031576477190362	7.85403727...	1458
Row11	LotArea	YrSold	-0.02769688397012...	0.29024108...	1458
Row12	LotArea	OverallCond	-0.04645887304532...	0.07595919...	1458
Row13	LotArea	OverallQual	0.23294749355973426	0.0	1458
Row14	LotArea	SalePrice	0.4555814150180786	0.0	1458
Row15	LotFrontage	1stFlrSF	0.3650630110453706	0.0	1458
Row16	LotFrontage	YearBuilt	0.1792891303450814	5.17297316...	1458
Row17	LotFrontage	YrSold	-0.00190306254254...	0.94208169...	1458
Row18	LotFrontage	OverallCond	-0.07699860366010...	0.00324037...	1458
Row19	LotFrontage	OverallQual	0.2218437992561246	0.0	1458
Row20	LotFrontage	SalePrice	0.34397757949890445	0.0	1458
Row21	1stFlrSF	YearBuilt	0.2932885698383333	0.0	1458
Row22	1stFlrSF	YrSold	-0.0223987256632922	0.39242474...	1458
Row23	1stFlrSF	OverallCond	-0.1653159319007171	2.08323136...	1458
Row24	1stFlrSF	OverallQual	0.4086751606379143	0.0	1458
Row25	1stFlrSF	SalePrice	0.5751800326950135	0.0	1458
Row26	YearBuilt	YrSold	-0.01410048162352...	0.59033961...	1458
Row27	YearBuilt	OverallCond	-0.41575452085243...	0.0	1458
Row28	YearBuilt	OverallQual	0.6474104558065921	0.0	1458
Row29	YearBuilt	SalePrice	0.6525635651835422	0.0	1458
Row30	YrSold	OverallCond	0.04860562736024367	0.06335026...	1458
Row31	YrSold	OverallQual	-0.02546444175598...	0.33089175...	1458
Row32	YrSold	SalePrice	-0.03004560501794...	0.25125076...	1458
Row33	OverallCond	OverallQual	-0.18419538201527...	1.31383792...	1458
Row34	OverallCond	SalePrice	-0.1330958846432245	3.32630008...	1458
Row35	OverallQual	SalePrice	0.8097139547441321	0.0	1458

Row ID	row ID	LotArea	LotFrontage	1stFlrSF	YearBuilt	YrSold	OverallCond	OverallQual	SalePrice
row ID	1.0	-0.293939931033733...	-0.043550740189666...	-0.113574000966666...	-0.036533441243466...	0.020640772789832...	0.0035555240052546...	-0.107223864345844...	-0.151524030338960...
LotArea	-0.2939399...	1.0	0.5095286495452701	0.443483629292227 57	0.1031576477190362	-0.027696883970125...	-0.046458873045322 24	0.232947493559734 26	0.4555814150180786
LotFrontage	-0.0435507...	0.5095286495452701	1.0	0.3650630110453706	0.1792891303450814	-0.001903062542545...	-0.076998603660103 55	0.221843799256124 6	0.3439775794989044 5
1stFlrSF	-0.1135740...	0.443483629292227 57	0.3650630110453706	1.0	0.2932885698383333	-0.022398725663292 2	-0.165315931900717 1	0.408675160637914 3	0.5751800326950135
YearBuilt	-0.0365334...	0.1031576477190362	0.1792891303450814	0.2932885698383333	1.0	-0.014100481623527...	-0.415754520852438 75	0.647410455806592 1	0.6525635651835422
YrSold	0.020640772...	-0.027696883970125...	-0.001903062542545...	-0.022398725663292 2	-0.014100481623527...	1.0	0.048605627360243 67	-0.025464441755985 02	-0.030045605017944...
OverallCond	0.003555524...	-0.046458873045322 24	-0.076998603660103 55	-0.165315931900717 1	-0.415754520852438 75	0.048605627360243 67	1.0	-0.184195382015278 38	-0.133095884644322 45
OverallQual	-0.1072238...	0.232947493559734 26	0.221843799256124 6	0.408675160637914 3	0.647410455806592 1	-0.025464441755985 02	-0.184195382015278 38	1.0	0.8097139547441321
SalePrice	-0.1515240...	0.4555814150180786	0.3439775794989044 5	0.5751800326950135	0.6525635651835422	-0.030045605017944...	-0.1330958846432245	0.8097139547441321	1.0

Donde se aprecia que los valores oscilan en el rango [-1, 1] igual que en la correlación de Pearson, pero esta vez aplicado a entre los valores de rango de dos variables. El significado del grado de correlación es el mismo descrito para Pearson.

La matriz del dataset, correspondiente a las métricas expuestas anteriormente en este punto, es de la siguiente manera:



Se pueden obtener unas conclusiones muy similares a las obtenidas con la correlación de Pearson.

- Scatter Plot[35], es un diagrama de dispersión que muestra como dos variables se relacionan entre sí, permitiendo estudiar, los problemas o causas relacionados con la calidad.

  Los valores de este tipo de gráfico, correspondientes a un conjunto de datos, se ubican como puntos de un plano cartesiano mostrando la relación como positiva (los valores aumentan juntos), negativa (un valor disminuye a medida que el otro aumenta), nulo (sin correlación) lineal, exponencial o en forma de U, pudiendo ser la fuerza de correlación fuerte, débil o ninguna[36].

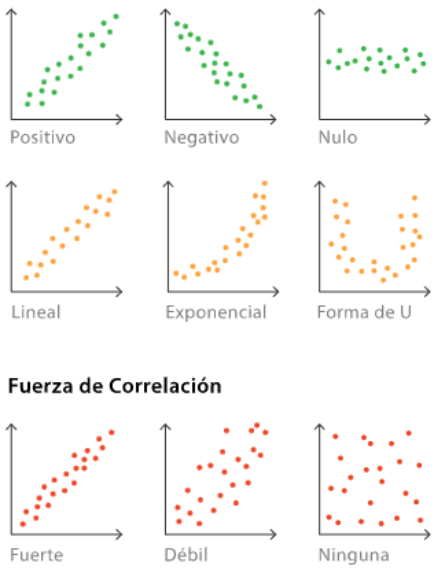
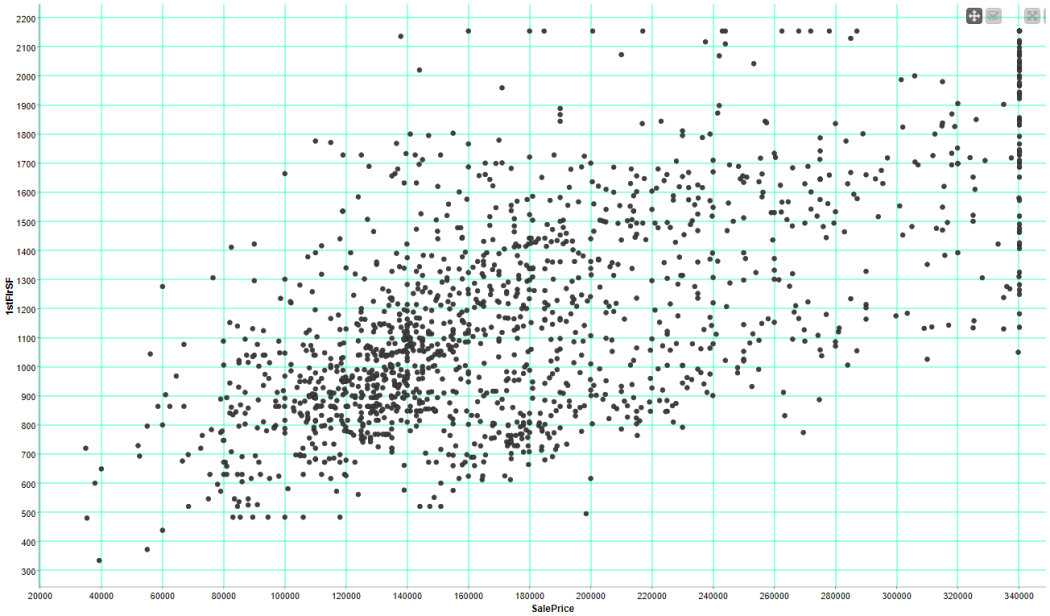


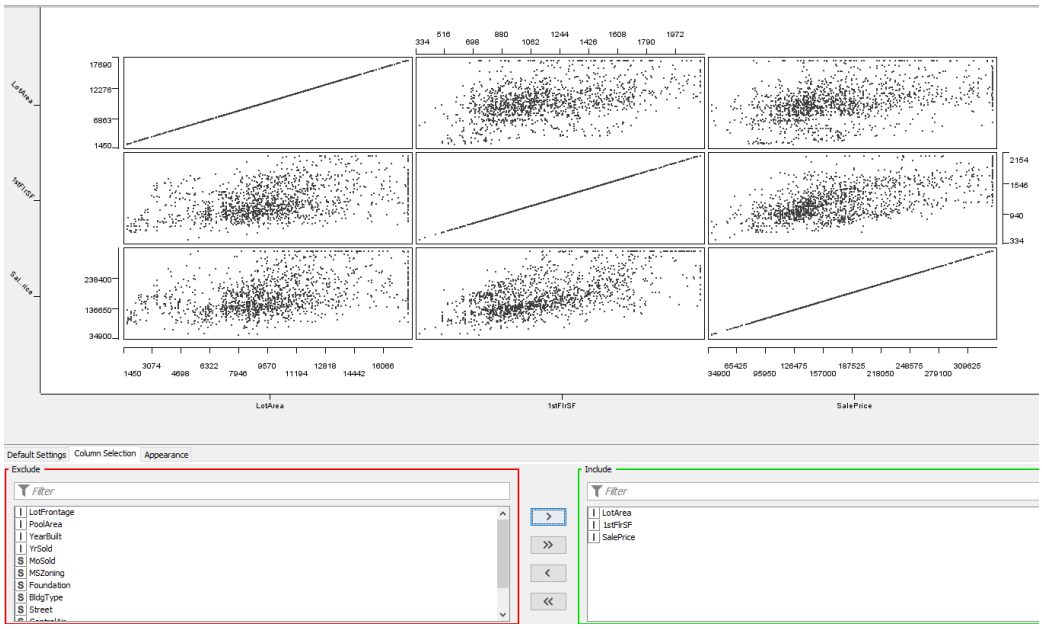
Ilustración 7 Tipos y fuerza de correlación en un diagrama de dispersión

Aplicando *Scatter Plot* al dataset, se pueden obtener, por ejemplo, la relación entre el precio del inmueble (*SalePrice*) y la superficie del mismo (*1stFlrSF*):



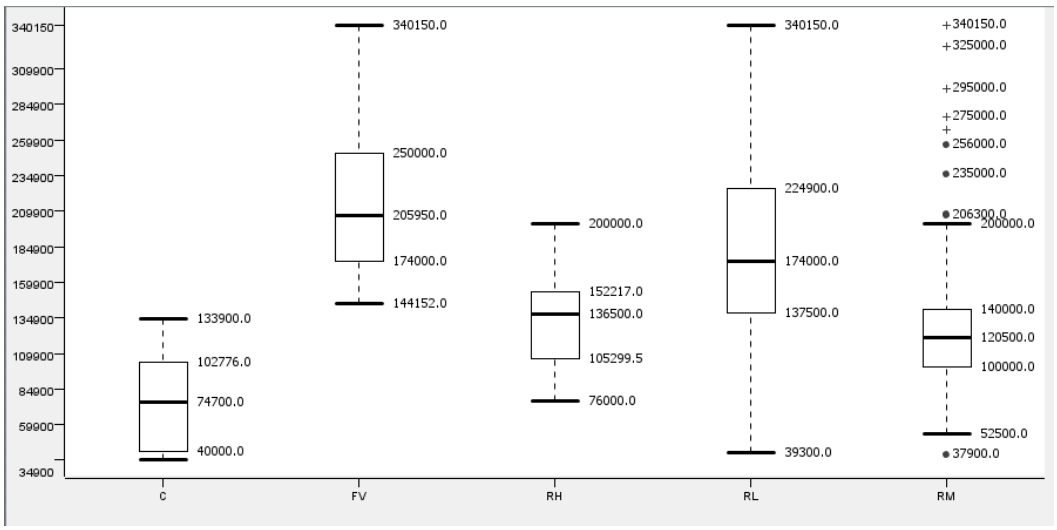
Mostrando una relación positiva débil, es decir, se pude decir que cuanto más cuesta un piso es probable que más superficie tenga.

Se pueden generar los gráficos que se considere, por ejemplo, se puede añadir el codigo de la zona (*LotArea*):



Se observa que la relación entre el código de zona y la superficie o el precio del piso es nula o positiva muy débil.

- Boxplot de dos variables (numérica y categórica), se puede obejeter el valor número por cada categoría, por ejemplo, el precio de venta (*SalesPrice*) en cada uno de los tipos de zona (*MSZoning*):



Observando, por ejemplo, que:

- La zona más cara es FV (*Floating Village Residential*) y la más barata C (*Commercial*)
- En la zona RM (*Residential Medium Density*) se han vendido varios inmuebles muy por encima de su valor de mercado
- La zona RL (*Residential Low Density*) tiene el mayor rango de precios

El flujo2 ejecutado quedaría de la siguiente manera:

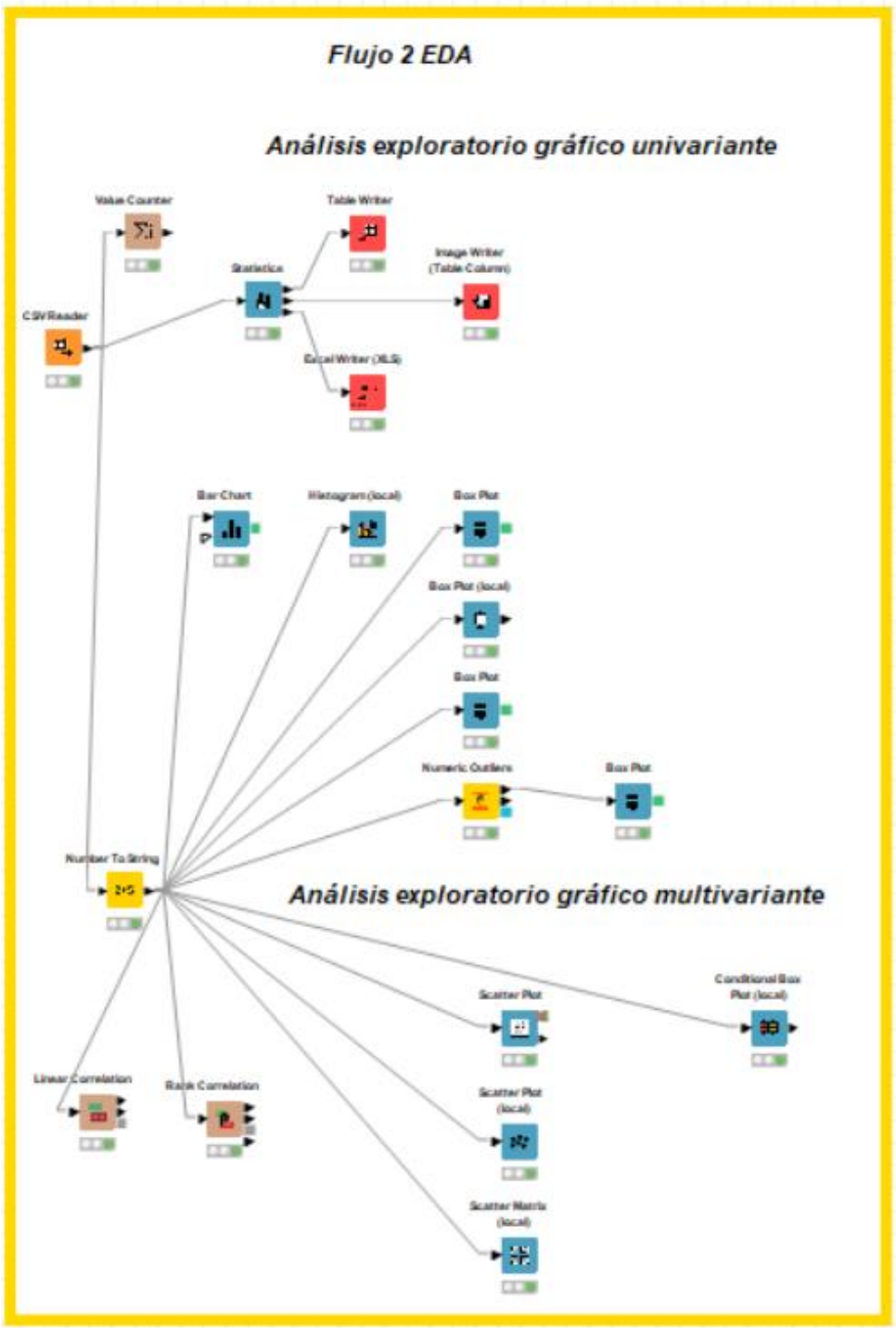


Ilustración 8 Flujo2 EDA

2.3 Flujo ML Supervisado. Definición

El aprendizaje supervisado[37], es una técnica para deducir una función (Test) a partir de datos de entrenamiento (Train). La salida de la función puede ser un valor numérico (regresión[38]) o una etiqueta de clase (clasificación[39]). Esta función debe ser capaz de predecir el valor correspondiente a cualquier objeto de entrada válida después de haber visto una serie de ejemplos, los datos de entrenamiento.

La elección del modelo a utilizar en ML supervisado se basará en si la relación entre los predictores y la respuesta es lineal, pudiendo usar regresión lineal, o no, pudiendo usar árboles de decisión (hay otros tipos de modelos predictivos supervisados, pero se consideran estos como los más relevantes).

A continuación, partiendo del conocimiento del *dataset* de propiedades inmobiliarias visto en los flujos anteriores, se va a desarrollar un modelo analítico, basados en técnicas de aprendizaje supervisado, que predigan el precio de un inmueble. Para ello se usará el modelo basado en regresión lineal y se evaluará su efectividad en función de las variables independientes elegidas.

2.3.1 Acciones previas

Para aplicar cualquier modelo de aprendizaje supervisado es recomendable realizar las siguientes acciones iniciales comunes:

2.3.1.1 Ingesta y partición de los datos

1-Leer datos: lee los datos en bruto del csv ubicado en una ruta especificada.

2-Visualizar los datos: mediante una matriz scatter o matriz de diagramas de dispersión se muestran las relaciones entre dos o más variables, permitiendo estudiar los problemas o causas relacionados con la calidad. Además, permite ver la linealidad existente entre ellas.

3-Particionar datos: se dividen los datos para el entrenamiento del modelo, 85% destinados a los datos de entrenamiento, *Train*, y el 15% para comparar el resultado del entrenamiento, *Test*. Knime utiliza el procedimiento *hold-out* o validación cruzada.

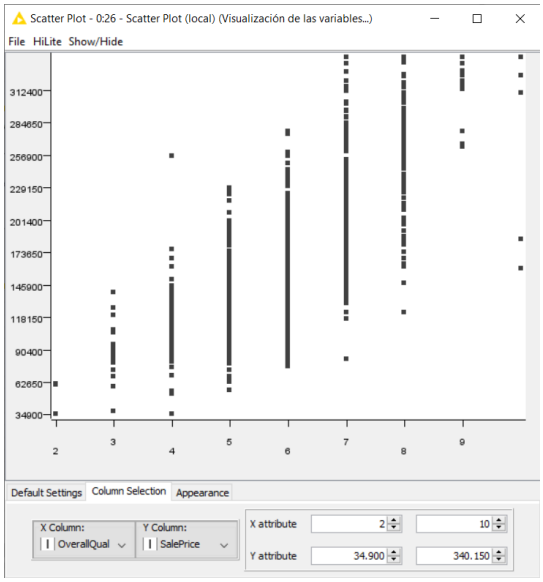
4-Estadísticos básicos: aunque ya se ha visto en los flujos anteriores, se pueden volver a calcular estadísticos básicos que permitan conocer mejor los valores y la relación entre variables, en esta ocasión se puede hacer sobre los datos de entrenamiento que suponen el 85% del conjunto total de datos, viendo que siguen la misma línea que los ya calculados (ver punto 7 del apartado 2.1.1 de este documento). También se podrían calcular sobre el conjunto completo de datos.

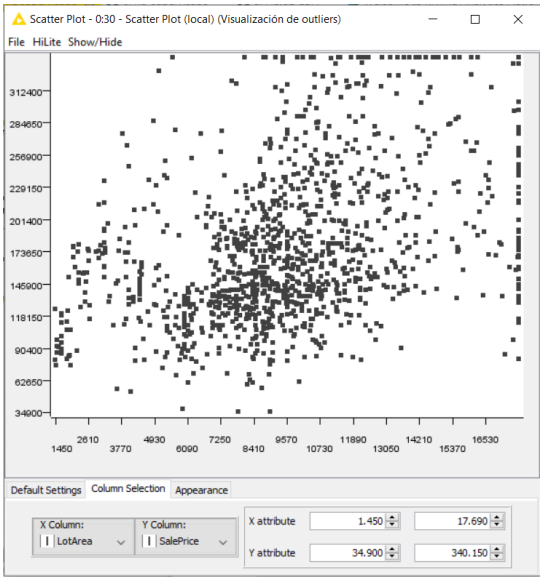
2.3.1.2 Preprocesamiento

Como se ha explicado en este documento (puntos 2.1.1 y 2.2.1) es muy importante contar con unos datos limpios y perfilados, para ello se deben aplicar distintas acciones antes del procesamiento de los datos. Aunque los datos ya vienen preparados de flujos anteriores, se cree interesante ver como en este flujo se aplican los siguientes pasos:

- Imputación de nulos del set de entramiento
- Filtrado de Outliers (sobre LotArea y SalesPrice)
- Visualización del resultado del preprocesado

También se pueden obtener distintos gráficos interesantes que permitan comprender los datos y su relación. Por ejemplo, visualizando cualquier variable predictora, una ya vista anteriormente es *OverallQual* o *LotArea*, frente a la variable objeto, *SalePrince*:





El preprocesado en este flujo sirve simplemente para contrastar puesto que se trabaja con un dataset procedente de los dos flujos anteriores, por tanto, ya está perfilado y transformado.

2.3.2 Regresión lineal Multivariante

La regresión lineal[41] es un modelo sencillo que ofrece ciertas ventajas:

- Interpretabilidad de los resultados
- Facilidad de uso
- Poco coste computacional

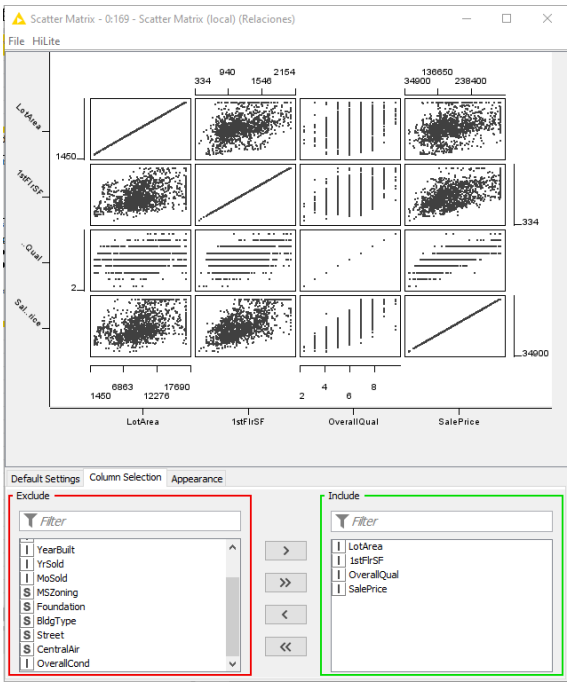
Por otro lado, tiene muchas limitaciones, por ejemplo:

- La relación entre los predictores y la salida debe ser lineal o casi lineal
- Es sensible a los *outliers* y a la colinealidad entre los predictores
- El número de muestras debe ser superior al de variables

Gracias al análisis exploratorio realizado en el flujo anterior, Flujo EDA (punto 2.2.1) se pueden apreciar el tipo de relaciones entre las distintas variables de un *dataset*. Concretamente, las correlaciones obtenidas en el análisis descriptivo y gráfico del flujo 2, punto 2.2.1.2 (matrices de correlación de Pearson y Spearman), muestran las relaciones más lineales entre las variables independientes y la variable dependiente del precio de venta, *SalePrice*, (ver página 33 de este documento) Estas variables independientes, que harán que este método predictivo sea más efectivo, son:

- *1stFlrSF* = Superficie. Cuanto mayor superficie mayor es el precio de venta.
- *LotArea* = Código del área, muy relacionado con la variable *MSZoning* pero la ventaja de esta es que es numérica.
- *MSZoning* = siglas de la zona.
- *OverallQual* = Valoración, del 1 al 10, de las calidades del inmueble. Tiene la mejor relación con el precio de venta, *SalesPrice*.

Se puede apreciar gráficamente en la siguiente matriz:

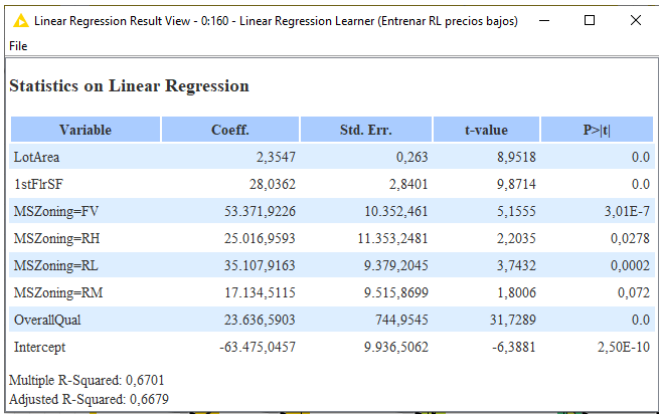


1-Segmentación, como hipótesis de investigación para mejorar el modelo se opta por dividir el *dataset* en precios bajos y altos. Se filtran los datos de entrenamiento y test haciendo dos grupos por cada conjunto:

- Inmuebles con precios bajos, límite 260.000
- Inmuebles con precios altos, supera el límite de 260.000

2-Creación del modelo, se aplica el modelo predictivo a los datos de entrenamiento, es decir, se entrenan los datos de *train*, separados por precios bajos y altos. A parte de los datos entrenados, *regresor*, se obtienen distintos coeficientes y estadísticas de las variables para cada una de las variables del conjunto de datos con respecto a la variable dependiente *SalePrice*:

Precios bajos


Linear Regression Result View - 0:160 - Linear Regression Learner (Entrenar RL precios bajos)

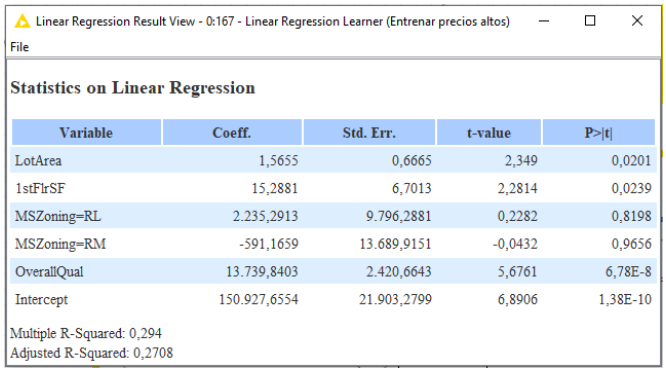
File

Statistics on Linear Regression

Variable	Coeff.	Std. Err.	t-value	P>|t|
LotArea	2,3547	0,263	8,9518	0.0
1stFlrSF	28,0362	2,8401	9,8714	0.0
MSZoning=FV	53.371,9226	10.352,461	5,1555	3,01E-7
MSZoning=RH	25.016,9593	11.353,2481	2,2035	0,0278
MSZoning=RL	35.107,9163	9.379,2045	3,7432	0,0002
MSZoning=RM	17.134,5115	9.515,8699	1,8006	0,072
OverallQual	23.636,5903	744,9545	31,7289	0.0
Intercept	-63.475,0457	9.936,5062	-6,3881	2,50E-10

Multiple R-Squared: 0,6701
Adjusted R-Squared: 0,6679

Precios altos


Linear Regression Result View - 0:167 - Linear Regression Learner (Entrenar precios altos)

File

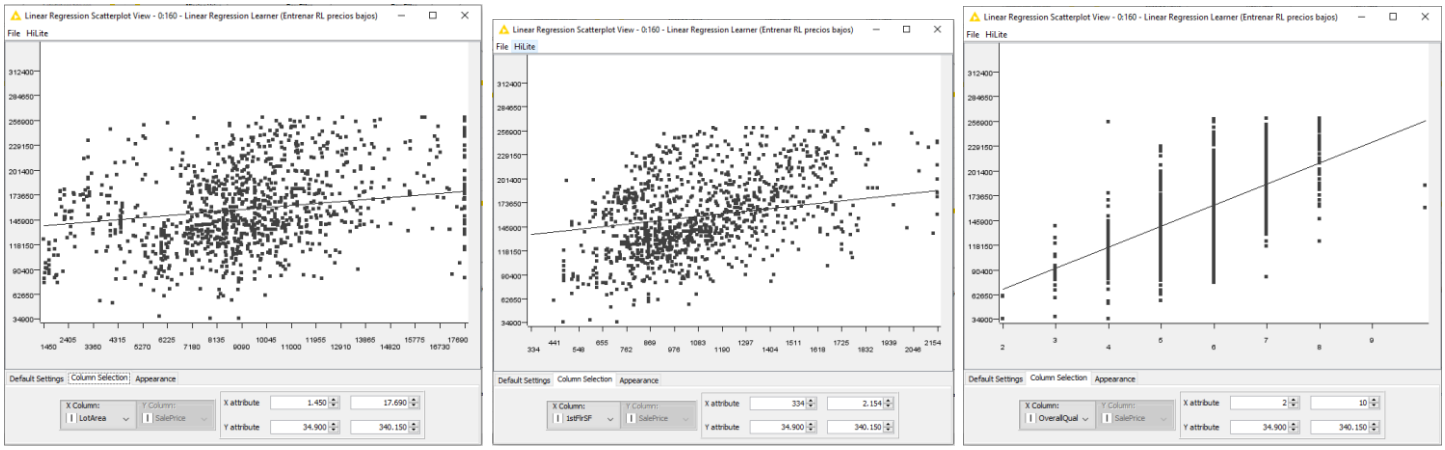
Statistics on Linear Regression

Variable	Coeff.	Std. Err.	t-value	P>|t|
LotArea	1,5655	0,6665	2,349	0,0201
1stFlrSF	15,2881	6,7013	2,2814	0,0239
MSZoning=RL	2.235,2913	9.796,2881	0,2282	0,8198
MSZoning=RM	-591,1659	13.689,9151	-0,0432	0,9656
OverallQual	13.739,8403	2.420,6643	5,6761	6,78E-8
Intercept	150.927,6554	21.903,2799	6,8906	1,38E-10

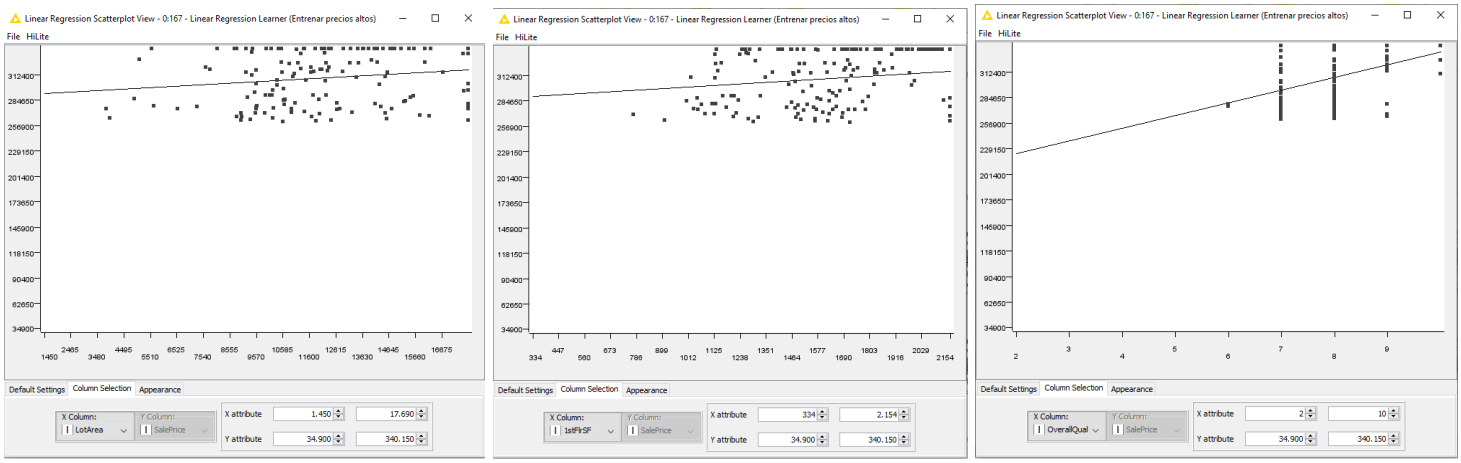
Multiple R-Squared: 0,294
Adjusted R-Squared: 0,2708

Una vez ejecutado este entrenamiento, también se pueden ver los valores estimados por el modelo lineal (línea recta) frente a los valores reales de cada variable, de forma gráfica:

Precios bajos:



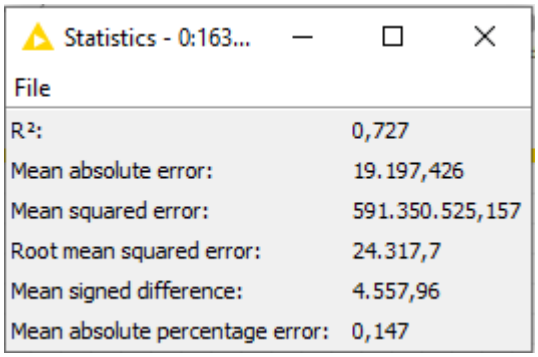
Precios altos:



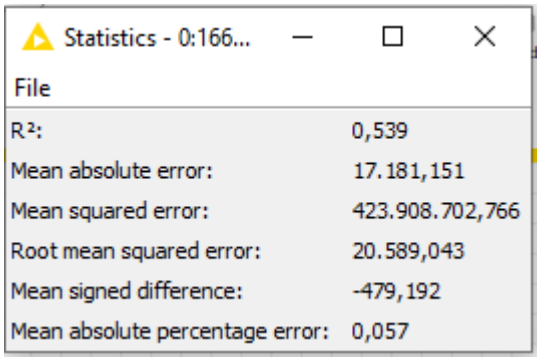
3-Evaluación de la predicción, se aplica el modelo prediciendo la respuesta mediante el modelo de regresión. Se conectan los datos del modelo regresor, datos de la predicción o entrenamiento obtenidos del paso anterior, con los de *test* y se agrega una nueva columna *Prediction (SalePrice)*, que permitirá comparar y valorar la calidad de la predicción y del modelo.

4-Estadísticas y métricas, se obtienen métricas de la calidad de la predicción entre el precio de test y el obtenido mediante la aplicación del modelo predictivo, como por ejemplo, R^2, error absoluto medio, error cuadrático medio, etc.:

Precios bajos



Precios altos



El coeficiente de determinación R^2[42] es un estadístico que mide la bondad del ajuste como la proporción de variación de los resultados que pueden explicarse por el modelo, es decir, prueba la hipótesis de nuestro modelo. Este coeficiente determina la calidad del modelo para replicar los resultados.

Se calcula dividiendo la desviación respecto a la media explicada por el modelo (SSR) entre la suma de esta misma desviación y la desviación respecto a la media no explicada por el modelo (SSE).

Para interpretar este coeficiente en la aplicación de este modelo a este dataset con las variables independientes seleccionadas sobre la variable dependiente se debe tener en cuenta que si el valor es igual a 1, indica que existe un ajuste lineal perfecto ya que la variación total de la variable *SalePrice* es explicada por el modelo de regresión. Mientras que si el valor es 0 indica la no representatividad del modelo lineal, lo que supone que el modelo no explica nada de la variación total de la variable *SalePrice*.

Se aprecia que este valor es, para el grupo de precios bajos 0,727 y para el grupo de precios altos 0,539. Siendo más efectivo el modelo para el conjunto de precios bajos.

Para este modelo se podrían aplicar otros pasos que obtienen distintas métricas como validación cruzada de K iteraciones (K-fold cross-validation), además, se podrían usar otras variables independientes para probar si mejoran el ajuste lineal (se ha probado y la combinación elegida es acertada). También, la regresión lineal se podría resolver con un modelo basado en árboles de decisión e incluso para este mismo caso se podrían usar modelos no lineales basados en clasificación (árboles).

El flujo3 ejecutado quedaría de la siguiente manera:

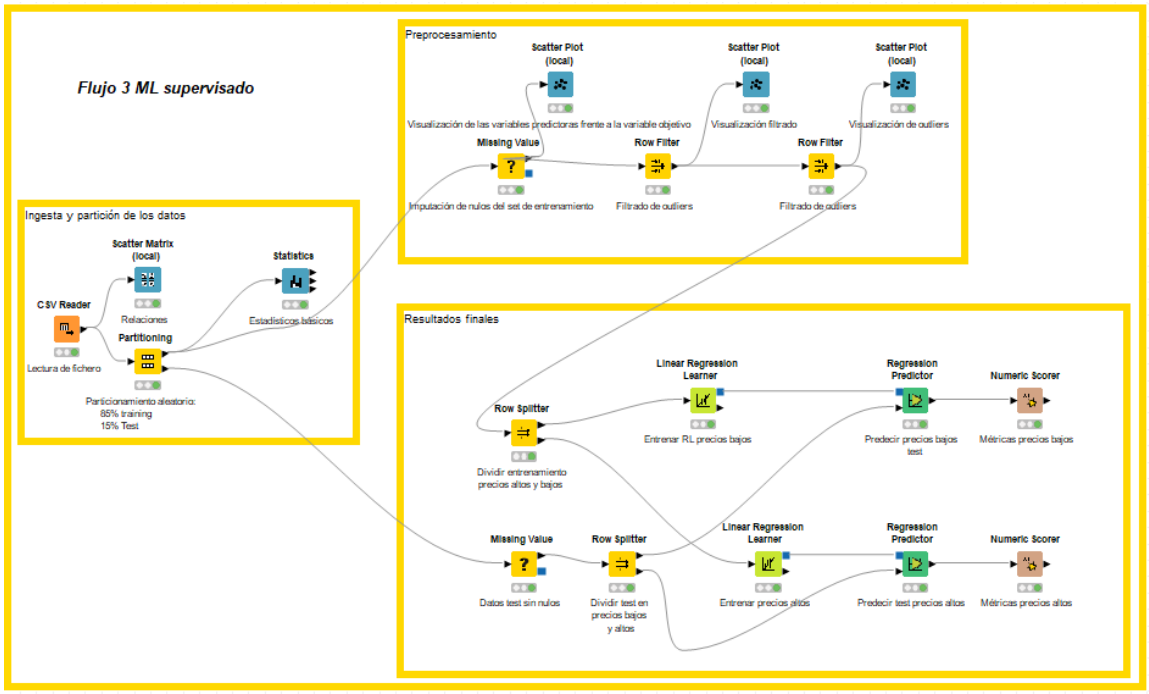


Ilustración 9 Flujo3 ML supervisado

2.4 Flujo ML No Supervisado. Definición

El aprendizaje no supervisado[43], es un método de aprendizaje automático donde un modelo se ajusta a las observaciones. Se distingue del aprendizaje supervisado por el hecho de que no hay un conocimiento a priori, tratando los objetos de entrada como un conjunto de variables aleatorias. Lo más destacado de este tipo de aprendizaje sería lo siguiente:

- Los datos de entrenamiento no están etiquetados con una salida Y.
- A diferencia del aprendizaje supervisado, en el no supervisado no hay forma determinística de verificar el performance del modelo. Sólo se puede evaluar con conocimiento del negocio.
- Algunas aplicaciones típicas del aprendizaje no supervisado son:
    - Segmentación de clientes
    - Detección de fraude o de anomalías
- Otra aplicación importante es la reducción de dimensionalidad.

El aprendizaje no supervisado se puede usar en conjunto con la inferencia bayesiana[44] para producir probabilidades condicionales, es decir, aprendizaje supervisado. Otra forma es la agrupación o *clustering*[45] que a veces no es probabilístico.

A continuación, partiendo del conocimiento del *dataset* de propiedades inmobiliarias visto en los flujos anteriores, se va a desarrollar un modelo analítico que agrupe los inmuebles según la similitud de sus características. Para ello se usará un modelo basado en el algoritmo de *clustering K-means* para agrupar en *cluster* las propiedades inmobiliarias, después se evaluará su efectividad.

2.4.1 Acciones previas

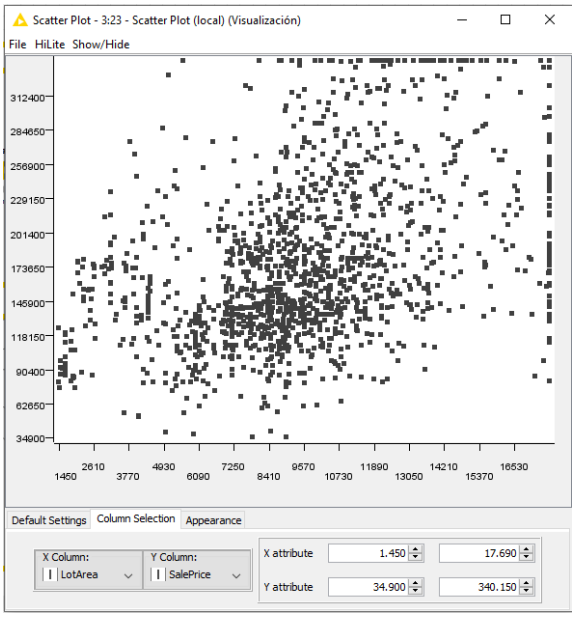
Para aplicar cualquier modelo de aprendizaje supervisado es recomendable realizar las siguientes acciones iniciales comunes:

2.4.1.1 Ingesta y corrección de nulos

1-Leer datos: lee los datos en bruto del csv ubicado en una ruta especificada.

2-Corrección de valores nulos: se asigna valor 0 a los nulos del *dataset*. Esto permitirá un mejor tratamiento del conjunto y una correcta normalización futura.

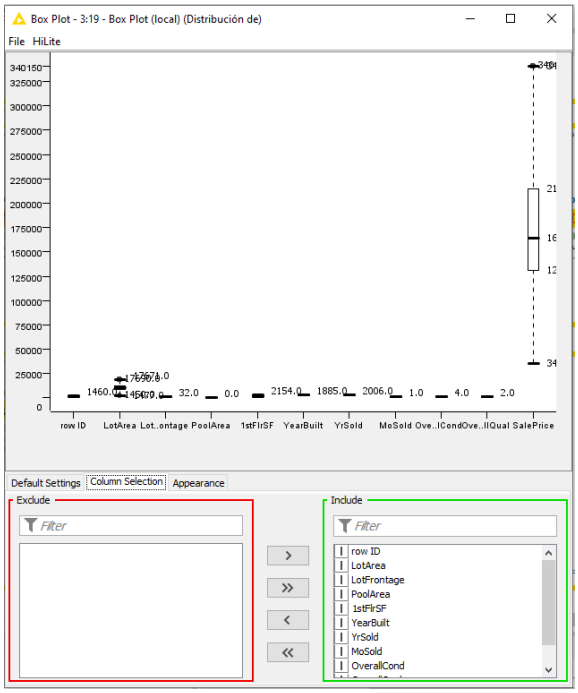
3-Visualizar variables: se analizan visualmente las variables mediante gráficos 2D interactivos, por ejemplo, *LotArea* con *SalePrice*:



4-Distribución de las variables: se observan los grupos de datos numéricos a través de los cuartiles de cada variable y los posibles *outliers* en función del rango intercuartílico (ver página 28, apartado 2.2.1.1), además, se obtienen distintos estadísticos que permitirán conocer el conjunto de datos en grupo y de forma individual:

Row ID	row ID	LotArea	LotFron...	PoolArea	1stFlrSF	YearBuilt	YrSold	MoSold	Overall...	Overall...	SalePrice
Minimum	1	1,450	32	0	334	1,885	2,006	1	4	2	34,900
Smallest	1	1,477	32	0	334	1,885	2,006	1	4	2	34,900
Lower Quartile	365.5	7,549	60	0	882	1,954	2,007	5	5	5	129,950
Median	730.5	9,478.5	63	0	1,087	1,973	2,008	6	5	6	163,000
Upper Quartile	1,095.5	11,603	79	0	1,391.5	2,000	2,009	8	6	7	214,000
Largest	1,460	17,671	107	0	2,154	2,010	2,010	12	7	10	340,000
Maximum	1,460	17,690	107	0	2,154	2,010	2,010	12	7	10	340,150

Hay que recordar que este tipo de diagramas de caja se pueden mostrar en paralelo (en el mismo gráfico) pero siempre que compartan la misma dimensión de valores, sino es ineficiente. En el siguiente gráfico se aprecia que los valores de las variables *LotArea* y *SalePrice* no están en la misma dimensión que el resto, por tanto, en el preprocesado (apartado siguiente) se deberán manipular.
:



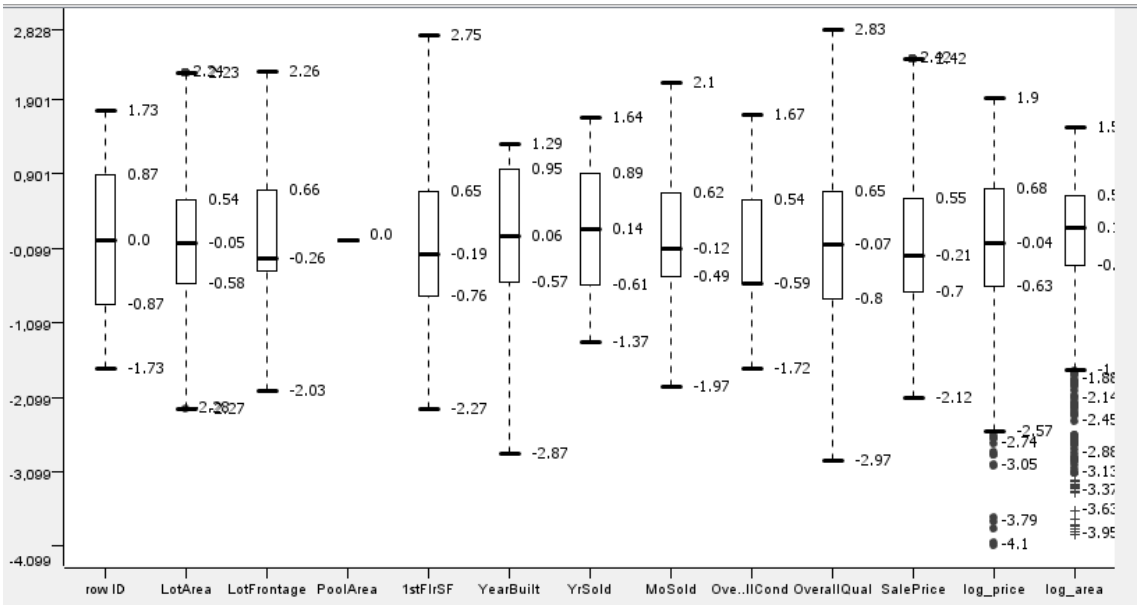
2.4.1.2 Preprocesamiento

KNIME implementa sin normalizar el algoritmo o modelo k-means, elegido para realizar la agrupación en *cluster* por propiedades inmobiliarias del conjunto de datos de ejemplo. Además, se aprecia en el punto anterior que las variables LotArea y SalePrice tienen una dimensión muy distinta que las demás, es por ello que se debe realizar lo siguiente:

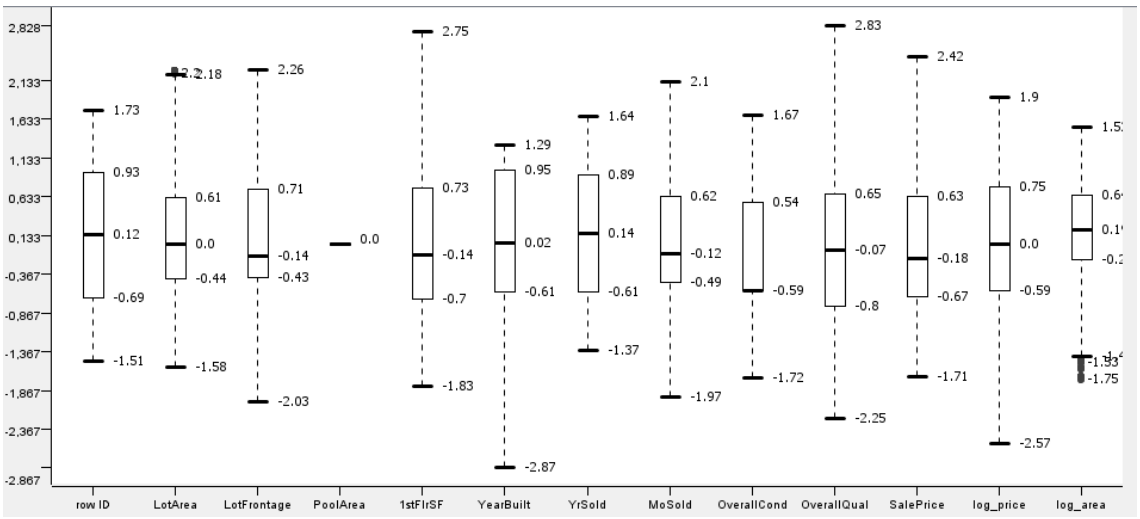
1-Transformar variables[46] *LotArea* y *SalePrice*, una buena forma de entender estas variables que superan las dimensiones de las demás, es ver su crecimiento en una escala estándar inferior, por ello, para reducir las diferencias entre valores grandes y pequeños se considera transformar sus valores y expresarlo en base logarítmica (aplicar logaritmos).

2-Normalización estándar[47], se ajustan los valores de todas las variables, que están a escalas diferentes, para pasar a medirse en una escala común. Se utiliza una normalización Gaussiana, distribución normal, que consiste en restar la media y dividir por la desviación típica. Esto hará que la escala en la que está representada cada variable no afecte la decisión sobre a qué *cluster* pertenece una observación.

3-Visualización distribuciones, se genera un gráfico de cajas en paralelo con todas las variables para observar que ahora, los datos tienen una dimensión común y serán medidos en la misma escala:



4-Eliminación de *outliers*, se eliminan posibles valores extremos que puedan desvirtuar los clusters en la aplicación del modelo k-means. Como los datos ya se han tratado en flujos anteriores, para evitar que se eliminen datos de interés aplicando demasiadas veces la técnica de elimiación de outliers, se acota su aplicación a las variables *LotFrontage*, *1stFlrSF*, *OverallCond*, *OverallQual*, *log_prince* y *log_*area y se vuelven a visualizar todas las variables en paralelo mediante un diagrama de cajas:



.

2.4.2 K-means

2.4.2.1 Introducción

El modelo *K-means clustering*[48] tiene como objetivo dividir n observaciones en k grupos (*clusters*) en los que cada observación pertenece al grupo con la media más cercana (centros de grupo o centroides). Con esto se consiguen grupos donde los miembros de cada uno son muy similares entre si y los miembros de grupos distintos son muy diferentes.

En función de lo estudiado en la asignatura Aprendizaje computacional en el primer semestre de 2021, este algoritmo se puede explicar de la siguiente manera:

- De manera aleatoria se generan k centroides. Cada observación es asignada al centroide más cercano.
- Se itera sobre los siguientes pasos hasta que las asignaciones a los cluster dejen de cambiar:
    - Para cada cluster, se calcula el centroide. El centroide será un vector compuesto por la media de los D predictores de las observaciones del mismo cluster.
    - Reasignar cada observación al cluster cuyo centroide esté más cercano a la observación.

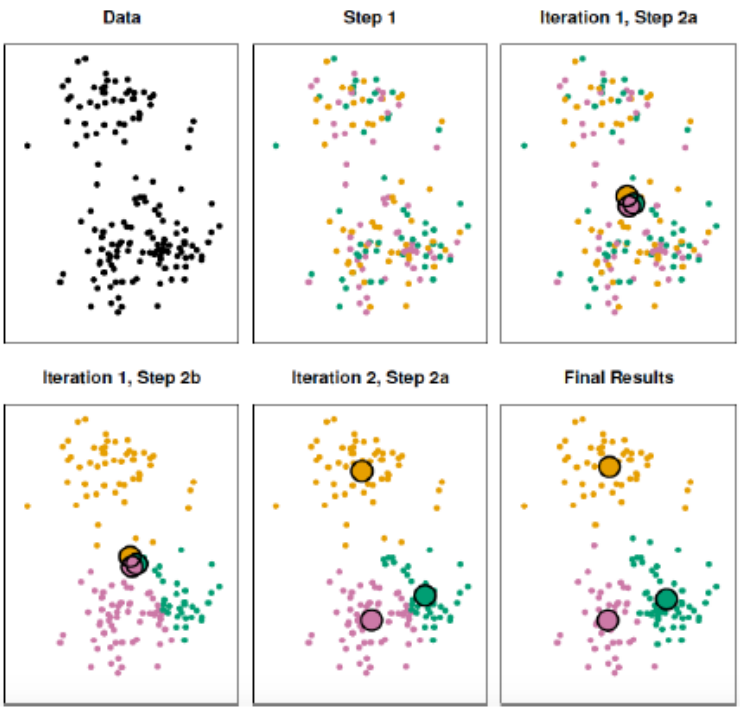


Ilustración 10 Algoritmo k-means

El número k (número de *clusters*) es un hiperparámetro que hay que darle al algoritmo.

2.4.2.1 Aplicación e interpretación

1-*Clustering*, al aplicar el modelo *k-means* sobre los datos ya tratados se deben elegir las variables mediante las cuales se quiere *clusterizar*. En función del diagrama de cajas, obtenido en el apartado 2.4.1.2, se decide utilizar las variables que tenían un gráfico similar, es decir:

- *row ID*
- *LotFrontage*
- *1stFlrSF*
- *OverallCond*
- *OverallQuadl*
- *log_price*
- *log_area*

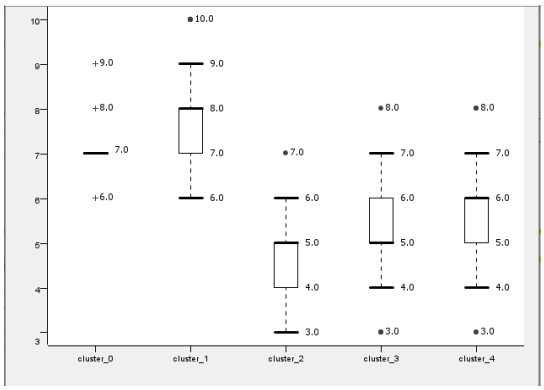
Además, se debe escoger la cantidad de cluster que se quieren obtener, lo normal y lo que se va a aplicar en este caso son 5 (k=5).

En este paso se añade al dataset original (tratado según se indica en el apartado 2.4.1) una columna donde se muestra el cluster o agrupamiento al que pertenece cada observación.

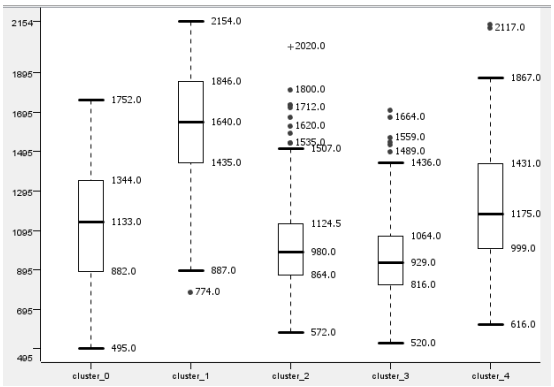
2-Deshacer la normalización, se debe deshacer la normalización realizada antes de realizar el *clustering* para poder interpretar los resultados con datos reales.

3-Visualizar las diferencias por cluster: se puede asignar una variable y ver la diferencia de sus observaciones entre los distintos cluster, por ejemplo:

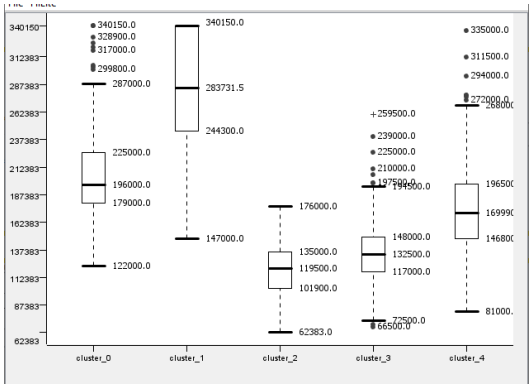
LotFrontage

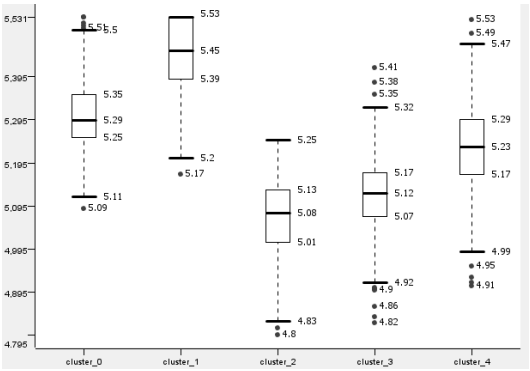


1stFlrSF

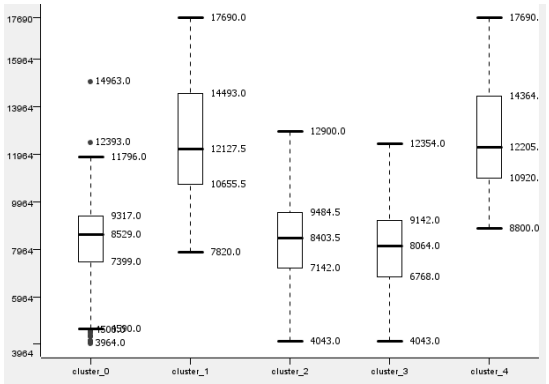


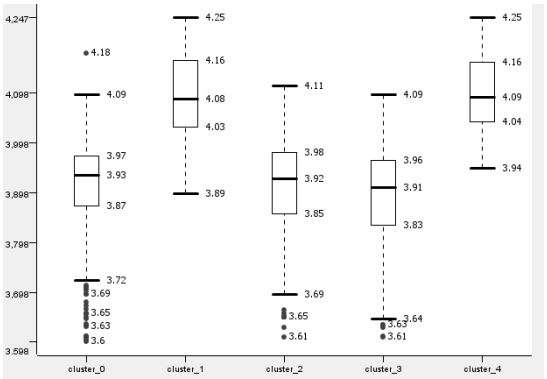
SalePrice *log_price*





LotArea *log_area*



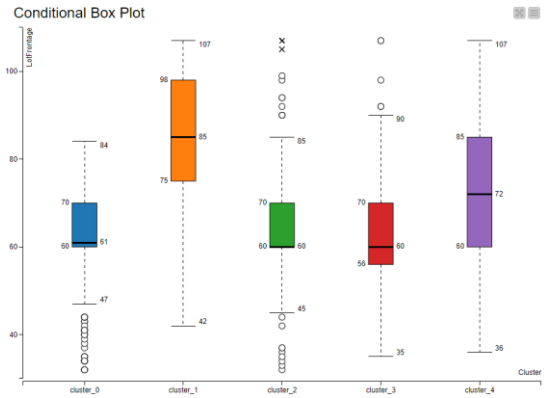


4-<u>Valores medios</u>, se pueden obtener los valores medios de los distintos cluster para las variables con que se ha realizado el *clustering*:

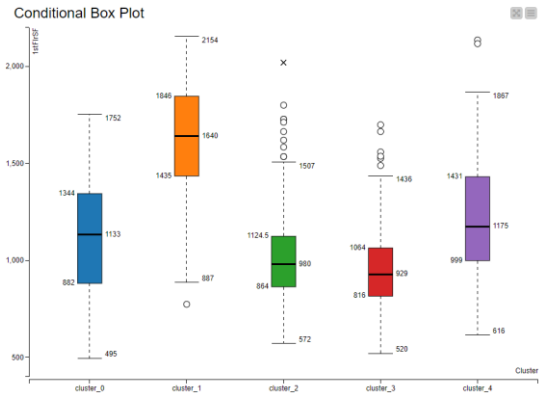
Cluster	Mean(LotArea)	Mean(1stFlrSF)	Mean(OverallCond)	Mean(OverallQual)	Mean(LotFrontage)	Mean(SalePrice)
cluster_0	8,206.601	1,132.366	5.049	6.984	61.272	204,707.733
cluster_1	12,732.693	1,629.713	5.172	7.963	84.869	282,568.664
cluster_2	8,272.557	1,018.646	4.759	4.877	64.108	118,265.472
cluster_3	7,945.755	949.52	6.627	5.254	63.019	133,563.508
cluster_4	12,986.066	1,218.346	5.642	5.797	73.779	175,226.296

5-Asignar color a cada cluster y visualizar en 2D y 3D, para ver las diferencias de cada cluster según sus variables. Otra forma es usar de nuevo el diagrama de cajas para cada cluster en función de una variable concreta:

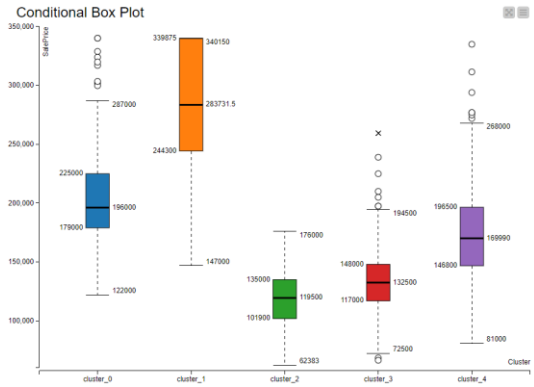
LotFrontage



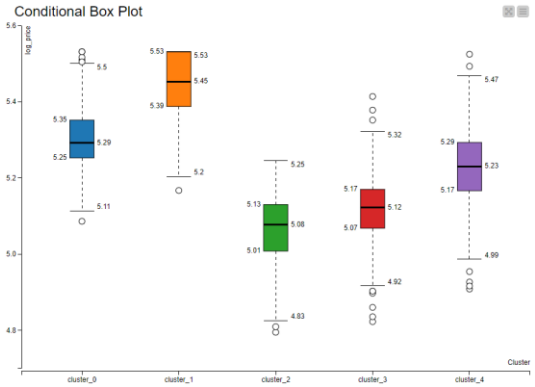
1stFlrSF



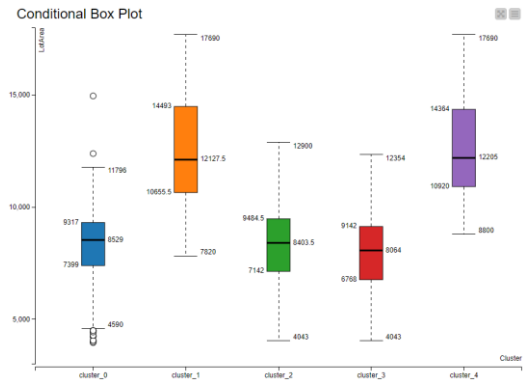
SalePrice



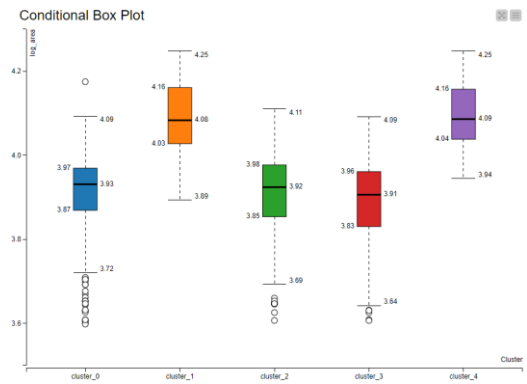
log_price



LotArea



log_area



2.4.2.1 Métricas de evaluación y PCA

La forma de evaluar este modelo es heurística y los puntos que se suelen comprobar son los siguientes:

- Si hay algún *cluster* con muy pocos datos significa que el número de *clusters* es demasiado alto, es necesario disminuir k.
- Si hay 'centroides' que están demasiado cerca entre sí, quiere decir que el número de *clusters* es demasiado alto, por tanto, es necesario disminuir k.
- Se pueden realizar representaciones en dos dimensiones de pares de características de las que se componen los datos, para ver si hay una clara agrupación gráfica de los *clusters*. Esto sólo es útil cuando el número de características es bajo.
- Si no se aprecia agrupación según características de las que se esperaba que existiera agrupación, significa que hay pocos centroides.

Se puede apreciar por el flujo y gráficos generados y explicados en el punto anterior que la elección de 5 cluster es correcta para la aplicación del modelo k-means sobre el conjunto de datos inmobiliarios que se maneja.

La evaluación de los modelos de *clustering* no puede hacerse de forma automatizada ya que, al ser una técnica no supervisada, no existen datos etiquetados y estos se necesitan contrastar con conocimiento de negocio, lo que si se puede es utilizar técnicas de visualización como las vistas en el apartado 2.4.2.1.
Una forma de ayudar a entender los resultados y valorar si el modelo necesita alguna modificación con respecto a las variables etc. y poder aplicar la lógica de negocio sobre los mismos es realizar un análisis de componentes principales o PCA[49].

La aplicación de PCA se realiza posterior a la del modelo k-means y con los datos normalizados. Permitirá limpiar el conjunto de datos describiéndolo en términos de nuevas variables no correlacionadas y reducir la dimensionalidad del *dataset*.

1-Calcular PCA, se realiza el análisis de componentes principales sobre los datos de entrada normalizados después de aplicar el modelo k-means. Como salida se obtiene la matriz de covarianza de las columnas de entrada:

Row ID	D row ID	D LotArea	D LotFron...	D 1stFlrSF	D Overall...	D Overall...	D SalePrice
row ID	0.874	0.797	0.397	0.379	-0.083	0.26	0.415
LotArea	0.797	0.787	0.347	0.355	-0.066	0.218	0.372
LotFrontage	0.397	0.347	0.858	0.288	-0.107	0.246	0.309
1stFlrSF	0.379	0.355	0.288	0.951	-0.189	0.456	0.585
OverallCond	-0.083	-0.066	-0.107	-0.189	1.004	-0.182	-0.156
OverallQual	0.26	0.218	0.246	0.456	-0.182	0.982	0.815
SalePrice	0.415	0.372	0.309	0.585	-0.156	0.815	0.994

2-Aplicar PCA, después de calcular el PCA se proyectan los datos de entrada en un espacio de menor dimensión mientras se conserva un máximo de información. Se decide que la reducción sea a dos dimensiones porque es muy cómodo de visualizar en un eje de coordenadas.

3-Se asigna un color a cada *cluster* y se visualizan las variables principales en 2 dimensiones, por ejemplo:

Gracias a la agrupación de los inmuebles del dataset según la similitud de sus características se podrán agrupar nuevas observaciones y aplicar el conocimiento de cada cluster. Muy interesante como enfoque ágil en la toma certera de decisiones.

El flujo4 ejecutado quedaría de la siguiente manera:

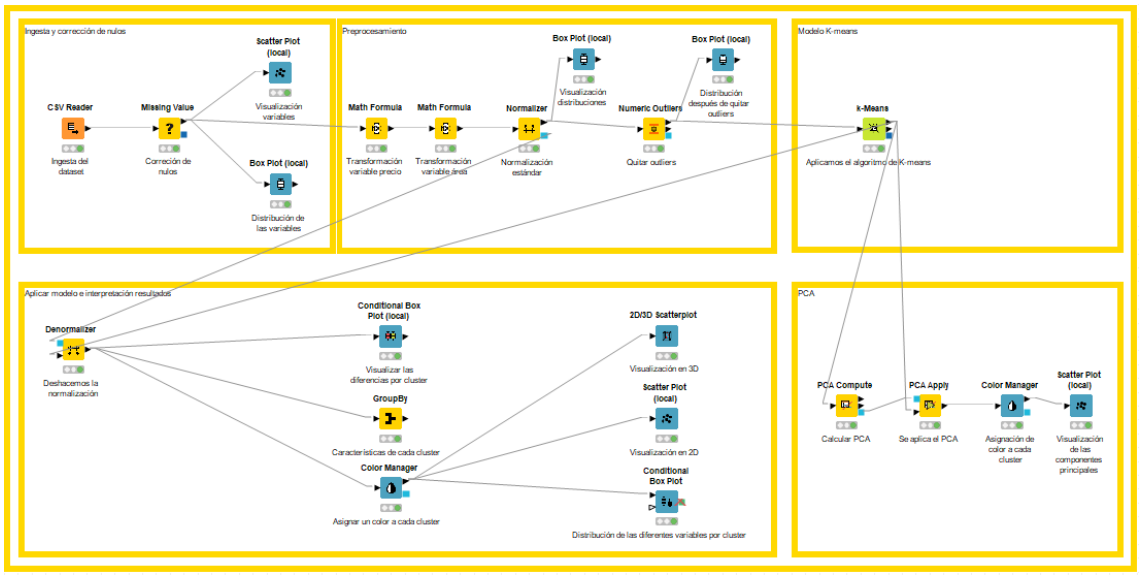


Ilustración 11 Flujo4 ML no supervisado

3. Conclusiones

En la actualidad la cantidad de datos consumidos en todo el mundo es de 79ZB y se prevé que en 2025 alcance los 180ZB[50]. Algunas de las cifras clave de lo que sucede en un minuto en el entorno digital, en internet, son las siguientes:

- Los usuarios de Tiktok ven 167 millones de vídeos
- Facebook Live recibe 44 millones de visitas
- 12 millones de usuarios envían un iMessage
- 6 millones de personas compran online
- Los usuarios de YouTube transmiten 694.000 vídeos
- Se postean 575.000 tweets
- Microsoft Teams conecta a 100.000 usuarios
- En Instagram los usuarios comparten 65.000 fotos

Ilustración 12 Cifras clave de data en un minuto

Al final, el dato bruto en general, ya sea en grandes cantidades (big data) o en cantidades más reducidas y de distintos tipos (estructurados, semiestructurados o no estructurados), necesita ser explotado para obtener conocimiento del mismo con el objetivo de predecir situaciones futuras y tomar decisiones certeras, proporcionando sobre todo tiempo a las personas que, en mi opinión, es lo más valioso.

Realizar una limpieza y análisis de los datos permite conocer bien las fuentes, el contenido y focalizar lo que interesa, si a esto se le añade automatización, percepción, capacidad de razonamiento y decisión según los objetivos marcados por los humanos se puede considerar un sistema de Inteligencia Artificial (IA). Estos sistemas consiguen que las herramientas sean activas y generativas en lugar de pasivas y forman un tándem perfecto con el Big Data que puede aplicarse para mejorar cualquier sector.

En este TFG y dentro de un sistema de estas características, IA, me he enfocado en el aprendizaje sobre los datos, dando una pincelada al potencial que tiene el ML (rama de la IA y de la Ciencia de datos), destacando lo siguiente:

- Los datos son el oro del futuro.
- La tecnología puede ser inteligente pero no sabia, debe ser creada y supervisada por humanos con conocimiento de negocio.
- La clave reside en la aplicación de modelos de interacción óptimos desarrollados por humanos expertos tecnológicos y/o de negocio, en definitiva, equipos multidisciplinares.
- En torno a los datos aparecen nuevos perfiles profesionales, entre ellos el *Data Scientist* que será el encargado de la extracción de valor y conocimiento de los datos, analizando la información, elaborando modelos predictivos y *repoting* estadísticos. Este TFG está muy enfocado a este perfil que debe tener habilidades o conocimientos sólidos de estadística, matemáticas, programación y entorno de desarrollo.
- El uso de patrones agiliza la aplicación de modelos y elimina el preguntar por el por qué.
- Los sistemas basados en ML deben ir acompañados de herramientas BI que asienten sus bases con modelos de datos ajustados al modelo de negocio y escalables, herramientas de extracción, transformación y carga (ETL's) y plataformas de almacenamiento y gestión distribuida.
- En la actualidad hay múltiples herramientas y lenguajes que permiten aplicar modelos de ML, lo importante es asentar los conocimientos para entender las distintas arquitecturas y tener presente la mejora y aprendizaje continuo del individuo ya que este campo evoluciona rápidamente.
- El potencial está aún por dimensionar en este mundo tan apasionante del dato.

Se ha seguido el método y la planificación definida en los puntos 1.3 y 1.4 de este documento que han garantizado el éxito del proyecto. Dividiendo el desarrollo en cuatro hitos/entregables que han permitido actuar sobre pequeños imprevistos de forma rápida y sin que afectara en los tiempos. Estos han sido los siguientes:

- Recursos: Este proyecto consta de un solo recurso personal, yo, y he tenido que compaginar el desarrollo con dos asignaturas más. Una de ellas me ha llevado más tiempo del necesario y en algunas tareas he tenido retrasos, afortunadamente he podido recuperar el tiempo sin que se hayan visto afectados los entregables.
- Se ha tardado más de lo esperado en entender bien la funcionalidad de los nodos de la herramienta *Knime*, tiene muchísimo potencial aún por descubrir.
- En el alcance del proyecto, dentro del flujo 3, no se incluyó aplicar un modelo no lineal y compararlo con la regresión lineal, pero lo tenía en mente y me hubiera gustado mucho poder haber llegado.

Los cuatro flujos resultantes son una muestra de las posibilidades de aplicar modelos de aprendizaje sobre datos preprocesados y depurados. A partir del *dataset* de ejemplo de propiedades inmobiliarias se han desarrollado dos modelos analíticos:

- Técnica de aprendizaje supervisado, mediante la regresión lineal se consigue predecir el precio de un inmueble
- Técnica de aprendizaje no supervisado, mediante la aplicación del modelo *k-means* se agrupan los inmuebles según la similitud de sus características permitiendo segmentar las viviendas y con ello los potenciales clientes

El ML tiene muchísimo más recorrido, algunas líneas de trabajo futuro pasarían por explorar los múltiples algoritmos de aprendizaje que existen (supervisados, no supervisados y de refuerzo), para obtener automatismos, compararlos y aprender de los datos facilitando la toma de decisiones. Además, existe una rama que combina estos algoritmos, el *Deep learning*, que básicamente los estructura en capas (neuronas), para crear una red neuronal artificial, que puede aprender y tomar decisiones por sí misma, es decir, el conjunto de algoritmos busca reproducir los mismos resultados que un cerebro humano.

A pesar de que tengo mucho que aprender y mejorar, a día de hoy ya considero que puedo aportar mi granito de arena y eso me satisface, lo aprendido hasta la fecha me ha llevado a cambiar de trabajo hace un par de meses y adentrarme en un proyecto técnico en la gestión de los datos de ciudades inteligentes para Telefónica.

Cada vez estoy más convencido en que la transformación digital parte de cambiar la cultura empresarial, que debe estar orientada al dato. Desde la directiva de las compañías se debe inculcar a todos los niveles de la organización. Al principio la inversión es elevada, pero merece la pena. Es muy importante contar con un equipo de profesionales unido, con experiencia y con conocimientos técnicos y de negocio.

Este TFG solo ha confirmado que la IA es algo apasionante con muchísimo recorrido de aprendizaje y con lo que estoy seguro que continuaré disfrutando.

4. Glosario

PAML: Predictive Analytics And Machine Learning
BI: Business Intelligence
ML: Machine Learning
ETL: Extract, Transform and Load
PCA: Principal Component Analysis
DS: Data Science
AI: Artificial Intelligence
IA: Inteligencia Artificial
DL: Deep Learning
PMI: Project Management Institute
PMBOK: Project Management Body of Knowledge
CSV: Comma-Separated Values
DQ: Data Quality
DW: Data Wrangling
EDA: Exploratory Data Analysis
IQR: InterQuartile Range
Q1: Quartile 1
Q2: Quartile 2

5. Bibliografía

[1]: El gran mapa del Big Data: ¿de dónde vienen todos nuestros datos? (elperiodico.com) 24/09/2021
[2]: Ciencia de datos - Wikipedia, la enciclopedia libre 24/09/2021
[3]: Reporte del Mercado Inmobiliario en NY - Febrero 2018 | Optimal Spaces 24/09/2021
[4]: Machine learning - Wikipedia 24/09/2021
[5]: Aprendizaje automático - Wikipedia, la enciclopedia principal 24/09/2021
[6]: El Proceso de Analítica de Datos y su Aplicación - Bruno Tafur 24/09/2021
[7]: Aprendizaje supervisado - Wikipedia, la enciclopedia principal 25/09/2021
[8]: Regresión lineal - Wikipedia, la enciclopedia principal 25/09/2021
[9]: Unsupervised learning - Wikipedia 25/09/2021
[10]: k-means clustering - Wikipedia 25/09/2021
[11]: http://www.knime.com 28/09/2021
[12]: https://www.analyticalpost.com/2019/05/the-forrester-wave-analisis-multimodal.html 28/09/2021
[13]: https://es.wikipedia.org/wiki/Project_Management_Institute 28/09/2021
[14]: https://es.wikipedia.org/wiki/Scrum_(desarrollo_de_software) 28/09/2021
[15]: http://sgv.es/nueva/index.php/empresa/metodologia 28/09/2021
[16]: Calidad de datos - Wikipedia, la enciclopedia libre 01/10/2021
[17]: Disputa de datos - Wikipedia, la enciclopedia principal 01/10/2021
[18]: Análisis exploratorio de datos - Wikipedia, la enciclopedia principal 01/10/2021
[19]: https://es.wikipedia.org/wiki/Limpieza_de_datos 10/10/2021
[20]: https://en.wikipedia.org/wiki/Outlier 10/10/2021
[21]: https://en.wikipedia.org/wiki/Quantile#Estimating_quantiles_from_a_sample 10/10/2021
[22]: Análisis exploratorio de datos - Wikipedia, la enciclopedia libre 27/10/2021
[23]: John W. Tukey - Wikipedia, la enciclopedia libre 27/10/2021
[24]: ANALIZANDO DATOS CON PYTHON - Datahack 27/10/2021
[25]: Estadística descriptiva - Wikipedia, la enciclopedia libre 27/10/2021
[26]: Estadística inferencial - Wikipedia, la enciclopedia libre 27/10/2021
[27]: 7 Análisis exploratorio de datos (EDA) | _main.utf8 (hadley.nz) 27/10/2021
[28]: 02-EDA-tutorial-solution – KNIME Hub 30/10/2021
[29]: Skewness and Kurtosis |Shape of data: Skewness and Kurtosis (analyticsvidhya.com) 30/10/2021
[30]: Histograma - Wikipedia, la enciclopedia principal 30/10/2021
[31]: Diagrama de caja - Wikipedia, la enciclopedia secundaria 30/10/2021
[32]: Correlation - Wikipedia 09/11/2021
[33]: Pearson correlation coefficient - Wikipedia 09/11/2021
[34]: Spearman's rank correlation coefficient - Wikipedia 09/11/2021
[35]: Scatter plot - Wikipedia 09/11/2021
[36]: Diagrama de Dispersión (datavizcatalogue.com) 09/11/2021
[37]: Aprendizaje supervisado - Wikipedia, la enciclopedia libre 14/11/2021
[38]: Análisis de la regresión - Wikipedia, la enciclopedia libre 14/11/2021
[39]: Clasificación estadística - Wikipedia, la enciclopedia libre 14/11/2021
[40]: Regresión lineal - Wikipedia, la enciclopedia libre 14/11/2021
[41]: 8.3.pdf (ugr.es) 15/11/2021
[42]: Coeficiente de determinación - Wikipedia, la enciclopedia libre 15/11/2021
[43]: Aprendizaje no supervisado - Wikipedia, la enciclopedia libre 25/11/2021
[44]: Inferencia bayesiana - Wikipedia, la enciclopedia libre 25/11/2021
[45]: Algoritmo de agrupamiento - Wikipedia, la enciclopedia libre 25/11/2021
[46]: tema4.dvi (uc3m.es) 25/11/2021
[47]: Normalización (estadística) - Wikipedia, la enciclopedia libre 25/11/2021
[48]: k-means clustering - Wikipedia 29/11/2021
[49]: Análisis de componentes principales - Wikipedia, la enciclopedia libre 03/12/2021
[50]: ¿Qué pasa en un minuto en Internet en 2021? - Telefónica (telefonica.com) 15/12/2021

6. Anexos

6.1. Conociendo KNIME

6.1 Introducción a KNIME

Knime[11] es una herramienta de soluciones PAML multimodales. Estas soluciones proveen la más amplia gama de entornos de trabajo. Ofrecen múltiples paradigmas de interfaces de usuario, enormes soluciones de entorno, interfaces gráficos, configuraciones con asistentes, automatización y entornos de desarrollo de código.

Según el informe Forrester[12], Knime se encuentra dentro de las 4 herramientas multimodales PALM principales, y de uso gratuito es la primera.

Ilustración 5 Clasificación en el mercado de la herramienta KNIME

El hecho de que sea una herramienta open source provoca que tenga una extensa comunidad de desarrolladores, lo que permite que esté en constante mejora su capacidad de computación, así como el aumento de la versatilidad de la herramienta.

KNIME está desarrollado sobre la plataforma Eclipse y programado en Java. Fue concebida como herramienta gráfica colaborativa y de investigación basada en la modificación de datos con nodos de cómputo independientes.

A continuación se muestra un cuadro con los enlaces más relevantes de la comunidad KNIME:

Comunidad KNIME	
Enlace	**Descripción**
http://www.knime.com	Página principal de KNIME, primera página a consultar en caso de buscar información de producto. KNIME Analytics Platform se puede descargar desde esta página
https://www.knime.com/knime-introductory-course	Página para aprender sobre funcionalidades específicas de KNIME. Engloba el ciclo analítico completo del dato.
http://www.knime.com/learning-hub	Materiales sobre campos específicos de uso de la herramienta (text mining, química, nodos básicos, Spark…)
https://forum.knime.com	Foro en el que plantear dudas sobre el uso de KNIME. Igualmente sirve como repositorio de preguntas
https://www.knime.com/knime-labs	En esta página se pueden encontrar nodos que se encuentran todavía en desarrollo.

Consideraciones sobre el entorno de trabajo:

- Al iniciar KNIME se mostrará una primera ventana en la que se pedirá elegir el *path* del *Workspace* de trabajo. Este directorio es sobre el que trabaja KNIME y se guardarán los *workflows* realizados y su configuración. Es recomendable organizar los distintos proyectos en *workspaces* diferentes.

- KNIME trabaja con flujo gráficos de trabajo o *workflows* que son secuencias analíticas que producen uno o más cambios de estado. Estos cambios de estado se realizan mediante el objeto atómico de procesamiento llamado paso o nodo. Por tanto, un conjunto de nodos forma un *workflow* y serán los que definan los pasos realizados sobre un conjunto de datos.

- Un paso o nodo es la unidad básica de procesamiento de un *workflow* que normalmente tendrá un input, una tarea de procesamiento y una salida. Los nodos poseen cuatro posibles estados, asimilados a los estados de un semáforo:

    - Rojo: Nodo inactivo sin configurar
    - Amarillo: Nodo configurado, pero sin ejecutar
    - Verde: Nodo ejecutado satisfactoriamente
    - Rojo con cruz blanca: Nodo ejecutado no satisfactoriamente

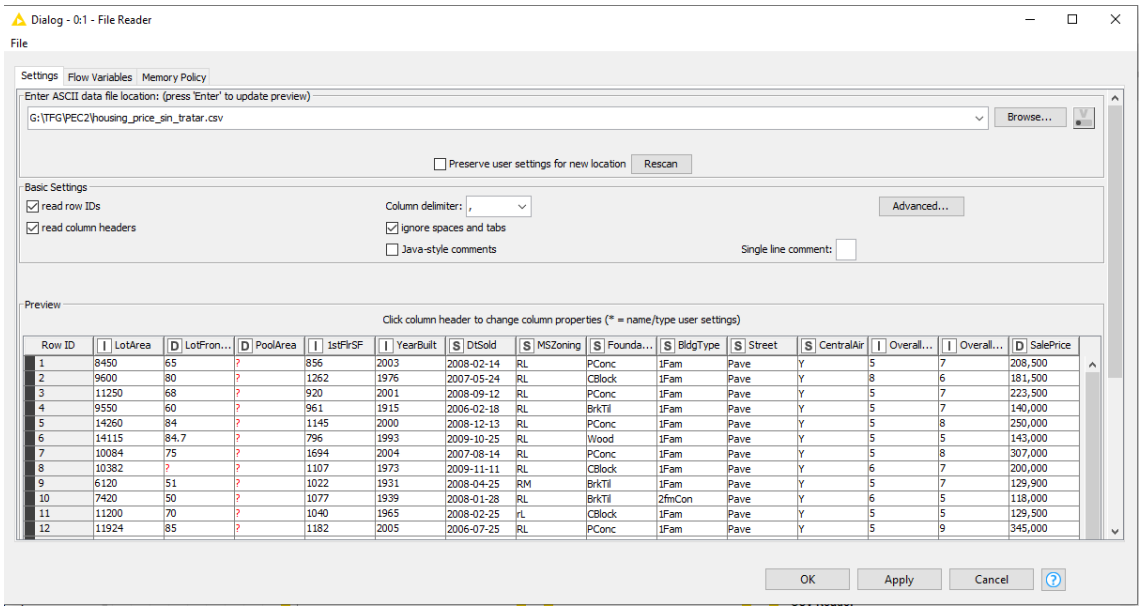
- Los *workflows* de KNIME pueden ser empaquetados y exportados en dos extensiones, .knwf (*workflow* único) y .knar (grupo de *workflows*).

6.2 Flujo BI, limpiar y preparar los datos. Estructura

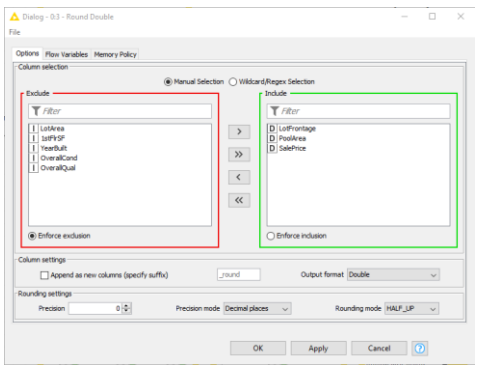
6.2.1 Perfilado y transformación. Estructura

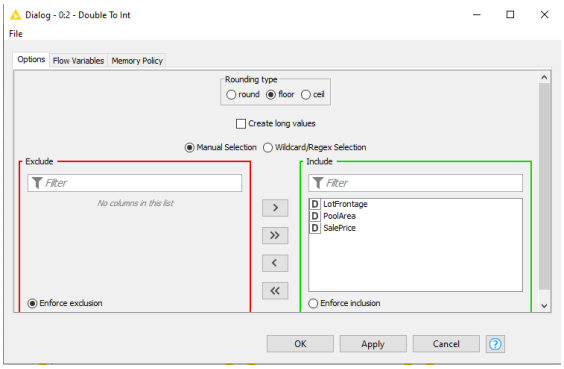
A continuación, se detallan la estructura y configuración de los nodos usados para esta tarea definida en el punto 2.1.1 de este documento:

1-Leer datos: se usa el paso *File Reader* indicando la delimitación por coma y que considere como primera columna.

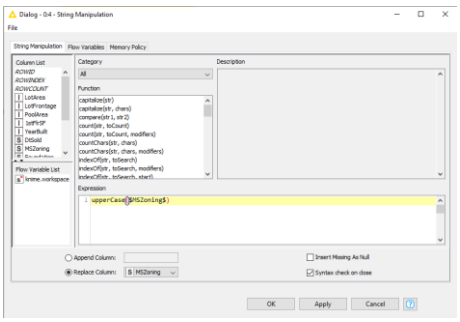


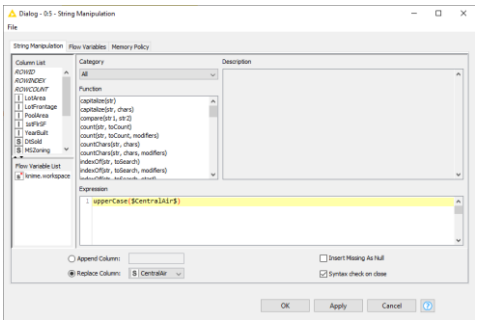
2-Formatear variables numéricas: se usan dos pasos *Round Double y Double To Int*.



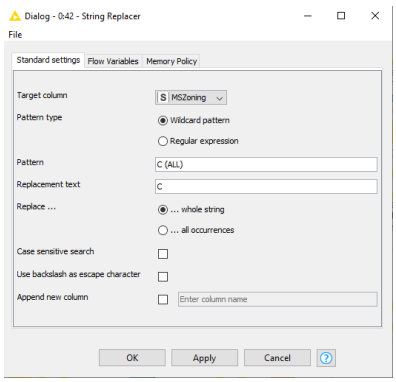


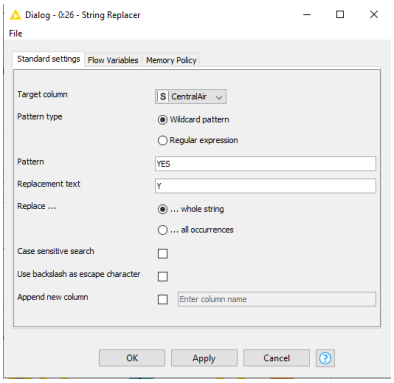
3-Formatear variables string: se utilizará el paso *String Manipulation* para reemplazar en la misma variables los valores formatenadolos a mayúsculas.

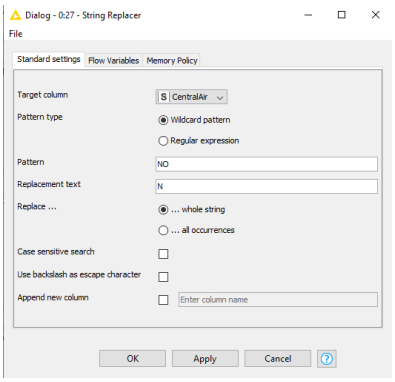




4-Normalizar Codificaciones: se usa el paso String Replacer.



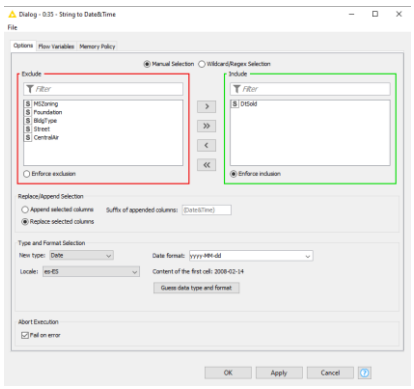




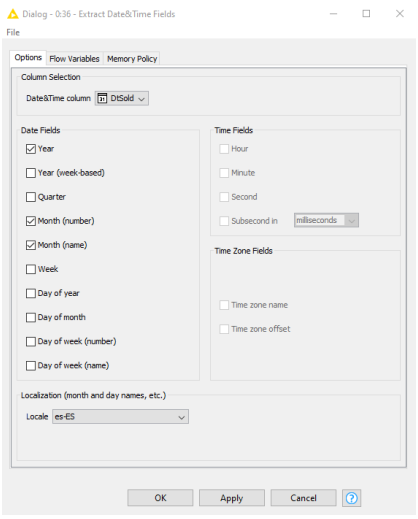
5-Tratamiento del campo *DtSold* de tipo *Date*.

Tratamiento como fecha:

- Se convierte el *string* a tipo de dato *Date* usando el paso *String to Date&Time*:

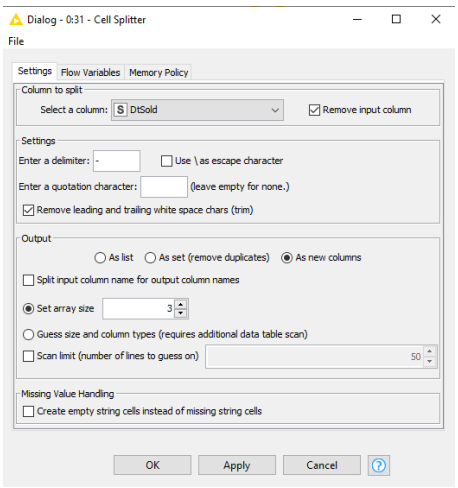


- Extraer la información de *DtSold* (fecha), se usa el paso *Extract Date&Time Fields*:

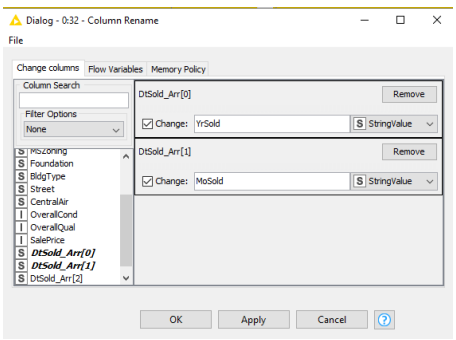


Tratamiento como string:

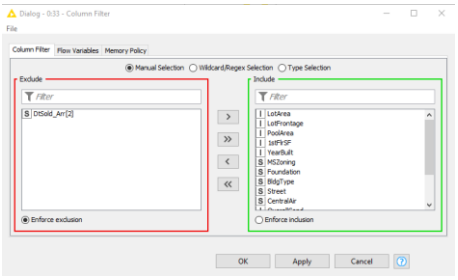
- Se divide la columna *DtSold* usando el paso *Cell Splitter*:



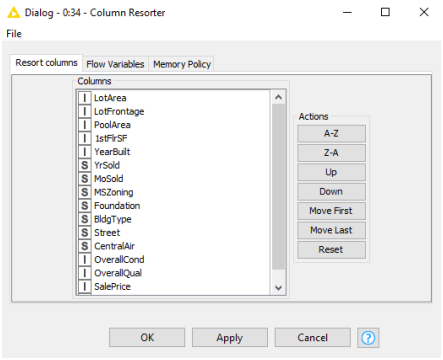
- Se renombran las columnas que contienen el año y el mes. Se usa el paso *Column Rename*:



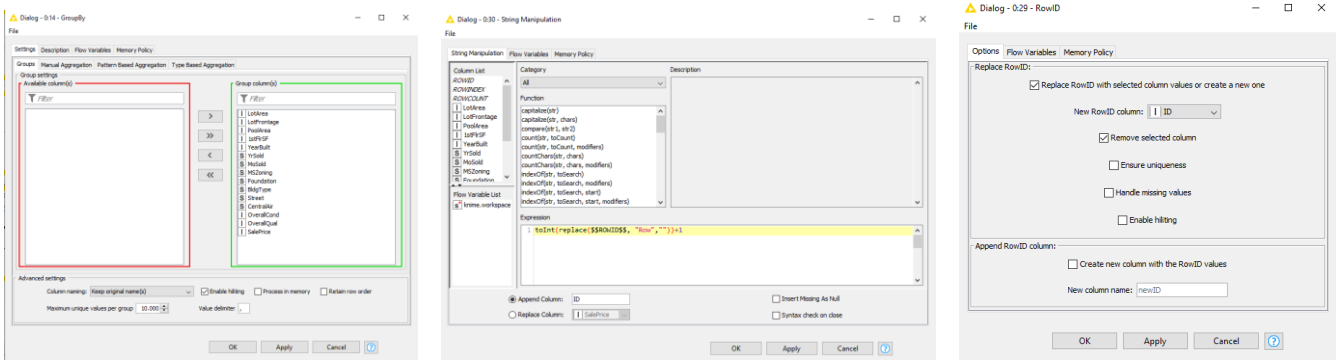
- Se quita la columna con la parte dia de la fecha de venta, con el objetivo de facilitar la agrupación posterior a nivel mensual. Se usa el paso *Column Filter:*



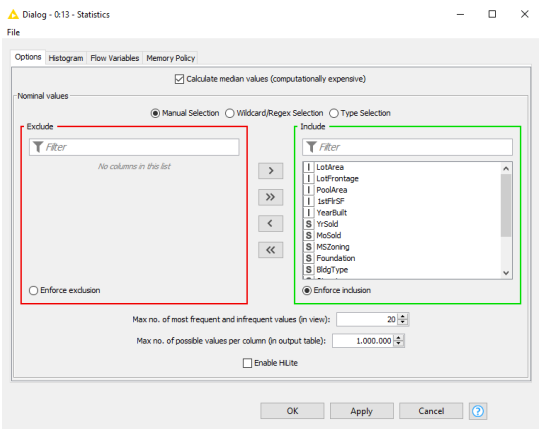
- Se reordenan las nuevas variables de modo que *YrSold* y *MoSold* estén entre *YearBuilt* y *MSZoning*. Se usa el paso *Column Resorter:*



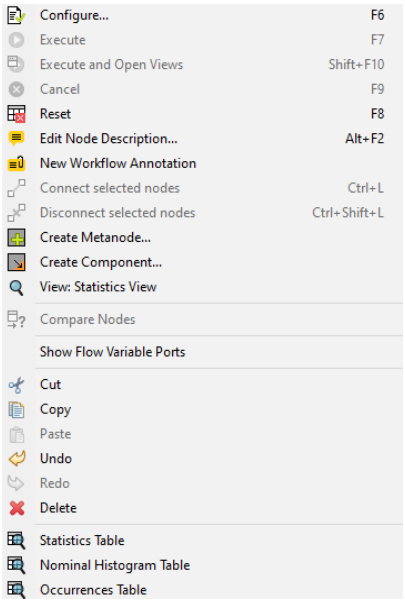
6-<u>Detección y filtrado de duplicados</u>: no se dispone de nodos específicos para esta tarea en KNIME por lo que se realizará la operación agrupando por todos los campos y recalculando el identificador ya que el paso de agrupación cambia el formato del identificador, campo *Row ID*. Se usarán los pasos *GroupBy*, *String Manipulation* y *RowID*:



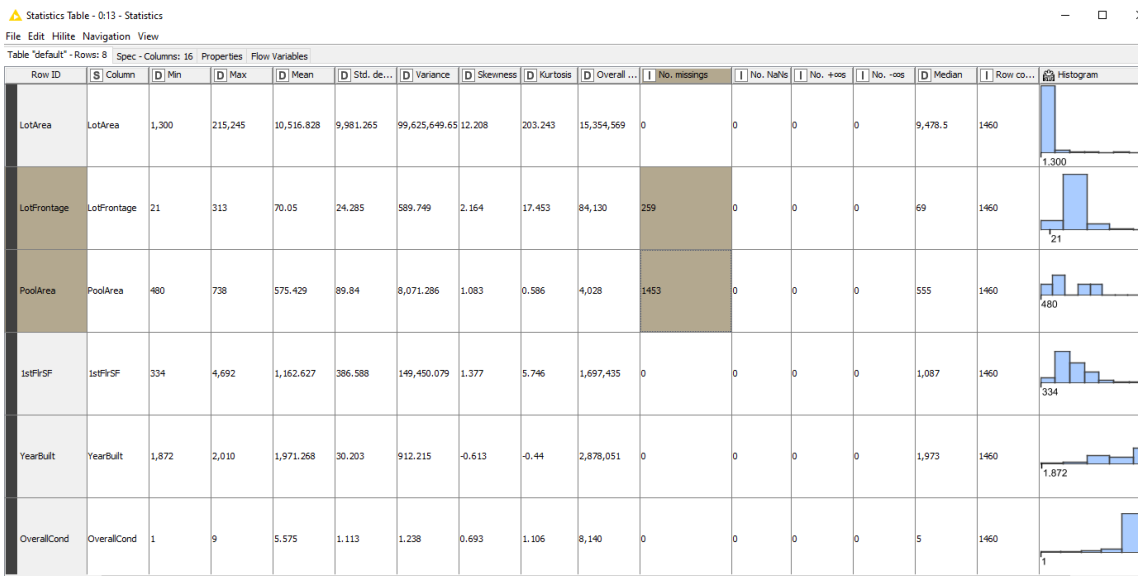
7-Tratamiento Nulos y Outliers[20]: para agilizar esta tarea es bueno contar primero con distintos estadísticos resumen sobre nulos y outliers, para ello, KNIME tiene el paso *Statistics:*



Y en las opciones del paso, una vez configurado como se muestra en la imagen anterior, al final se pueden sacar distintas métricas del dataset:



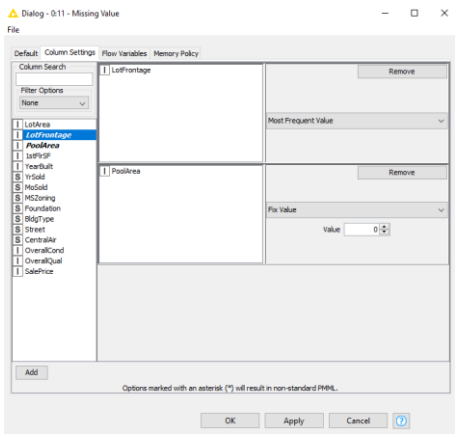
Mediante la opción *Statistics Table* se pueden ver distintos datos como número de registros, valor máximo, mínimo y medio, número de nulos etc. En la siguiente imagen se muestran las variables con nulos, *LotFrontage* y *PoolArea*:



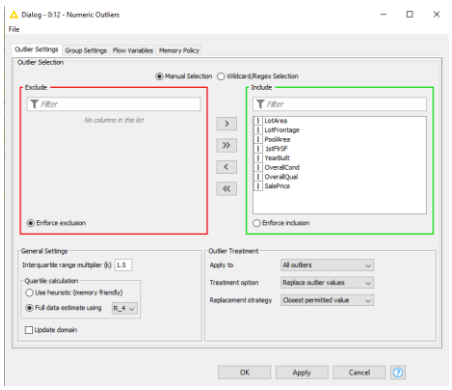
Row ID	Column	Min	Max	Mean	Std. de...	Variance	Skewness	Kurtosis	Overall ...	No. missings	No. NaNs	No. +oos	No. -oos	Median	Row co...	Histogram
LotArea	LotArea	1,300	215,245	10,516.828	9,981.265	99,625,649.65	12.208	203.243	15,354,569	0	0	0	0	9,478.5	1460	1.300
LotFrontage	LotFrontage	21	313	70.05	24.285	589.749	2.164	17.453	84,130	259	0	0	0	69	1460	21
PoolArea	PoolArea	480	738	575.429	89.84	8,071.286	1.083	0.586	4,028	1453	0	0	0	555	1460	480
1stFlrSF	1stFlrSF	334	4,692	1,162.627	386.588	149,450.079	1.377	5.746	1,697,435	0	0	0	0	1,087	1460	334
YearBuilt	YearBuilt	1,872	2,010	1,971.268	30.203	912.215	-0.613	-0.44	2,878,051	0	0	0	0	1,973	1460	1.872
OverallCond	OverallCond	1	9	5.575	1.113	1.238	0.693	1.106	8,140	0	0	0	0	5	1460	1

Para tratar los valores nulos se opta por configurar el paso *Missing Value* de la siguiente manera:

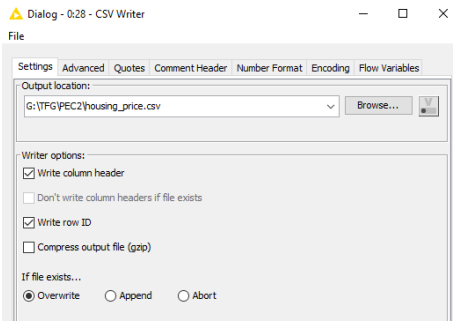
- *LotFrontage:* rellenar con el valor más frecuente
- *PoolArea:* rellenar con 0



Después de definir el tratamiento de los valores inusuales (*outliers*) y el uso del rango intercuartílico (IQR), la estrategia a seguir es aplicar el valor por defecto, K (1.5, y seleccionar el uso de estimación de datos completos, es decir, imputar *outliers* al valor más extremo permitido[21]. Para ello se usa y configura el nodo *Numeric Outliers*:



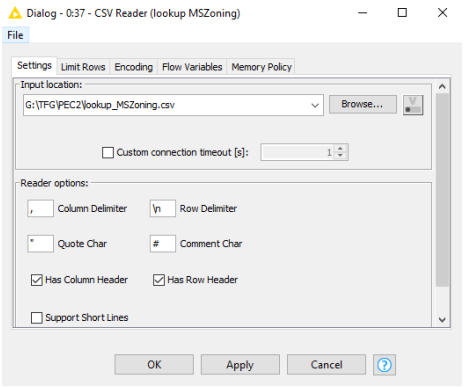
8-Exportación a csv: se usa el paso *CSV Writer*:



6.2.2 Enriquecimiento y transformación. Estructura

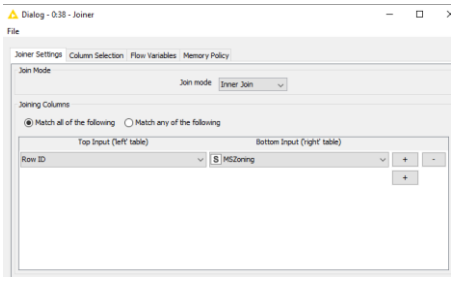
A continuación, se detallan la estructura y configuración de los nodos usados para esta tarea definida en el punto 2.2.1 de este documento:

1-Leer csv con la relación entre las siglas y la descripción de las zonas (campo *MSZoning*). Se usa el paso *CSV Reader:*



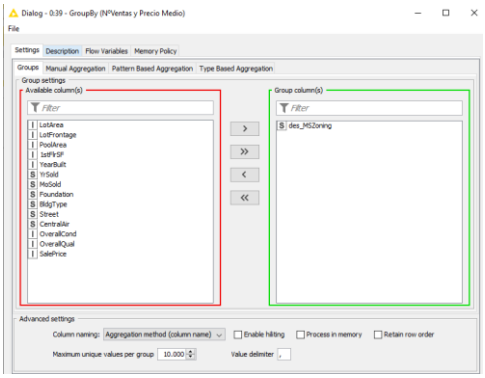
2-Unir el dataset inicial perfilado con el dataset que contine la descripción de las zonas. Se usa el paso Joiner:

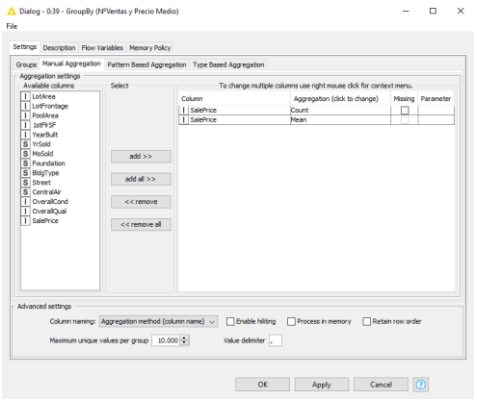
- Se une la columna *Row ID* con *MSZoning*:



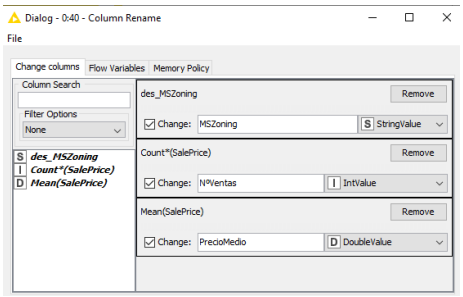
- Se listan las columnas de entrada que se incluirán en la salida, en este caso todas (*dataset* original y *dataset* descriptivo de la zona)

- Se agrupa por la descripción de la zona, calculando el número de ventas y el precio medio desde el campo *SalesPrice* y usando el paso *GroupBy*:

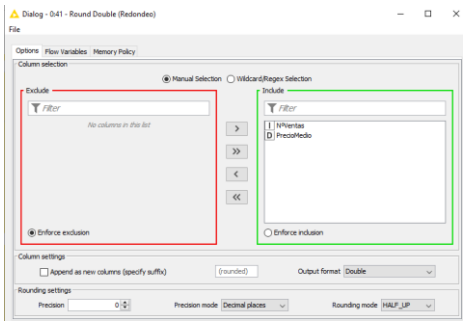




- Se renombran las nuevas columnas con las métricas con el paso *Column Rename*:



- Se redondean los valores de las métricas con el paso *Round Double*:

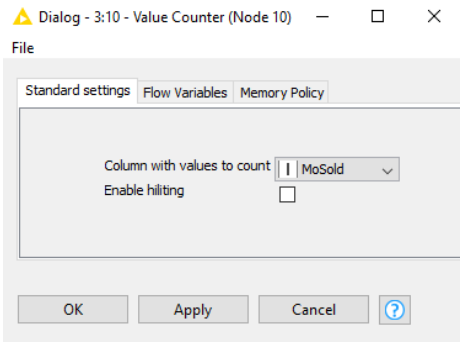


6.3 Flujo EDA, análisis exploratorio. Estructura

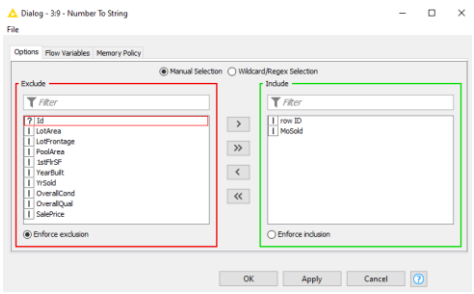
6.3.1 Análisis exploratorio gráfico univariante

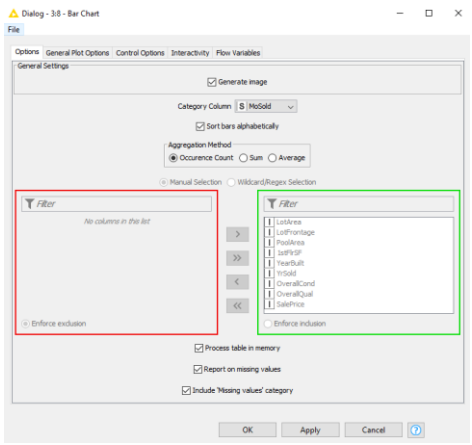
- ¿Cuántas ventas se realizan por mes?

  Para obtener la tabla se usa el nodo *Value Counter* agrupando por la variable entera *MoSold*:

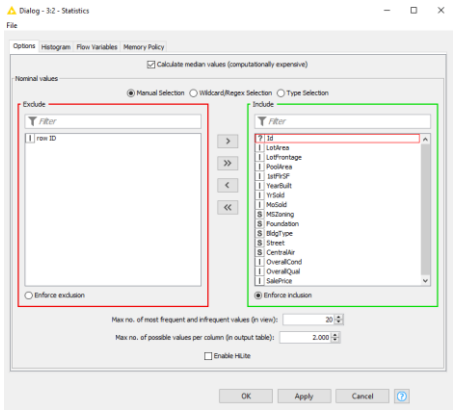


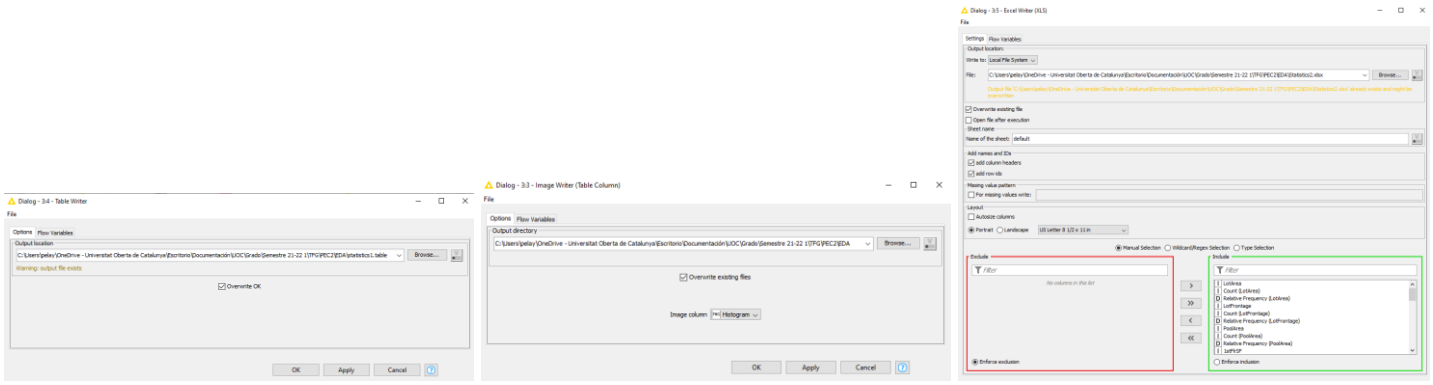
  Para obtener el gráfico de barras primero se pasará a cadena la variable MoSold, permitindo categorizar por ella en este y los demás gráficos del flujo. Se usan los nodos *Number To String* y *Bar Chart*:



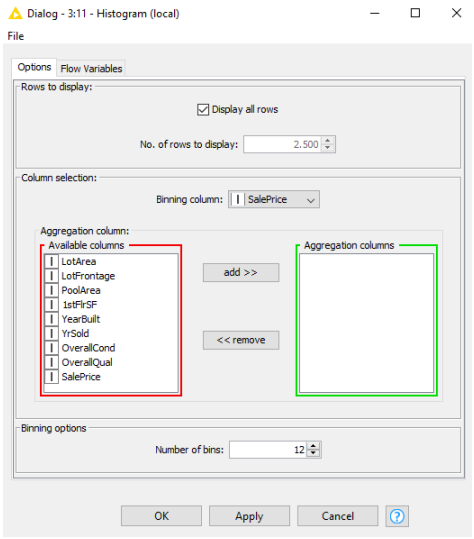


- Obtener estadísticos básicos de las variables numéricas, se usa el nodo Statistics, pudiendo exportar los datos con los nodos *Table Writer*, *Image Writer* y *Excel Writer*:





- Histogramas[30], se utiliza el nodo *Histogram (local)* y se configura para que realice 12 agrupaciones o rangos en función del precio:



- Boxplot[31], se utiliza el nodo *Box Plot* y *Box Plot (Local)*:

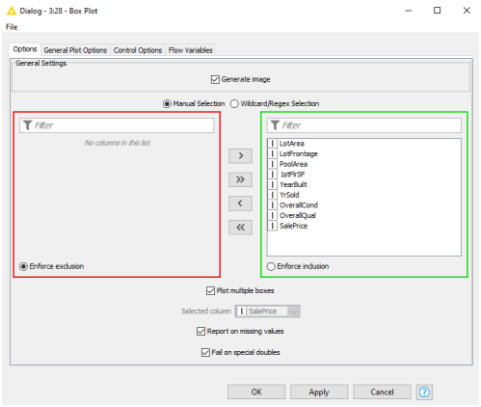
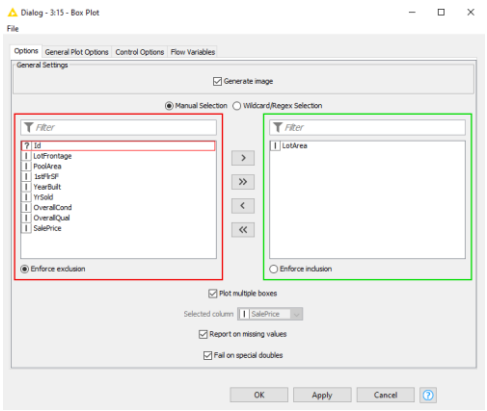
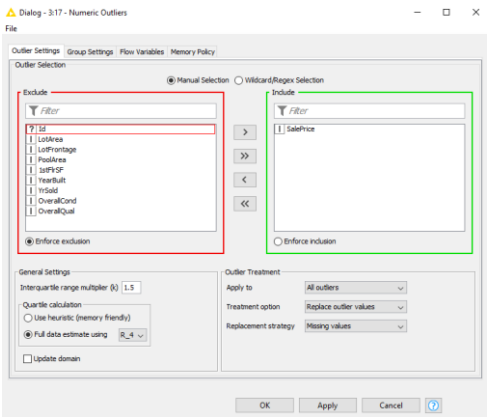
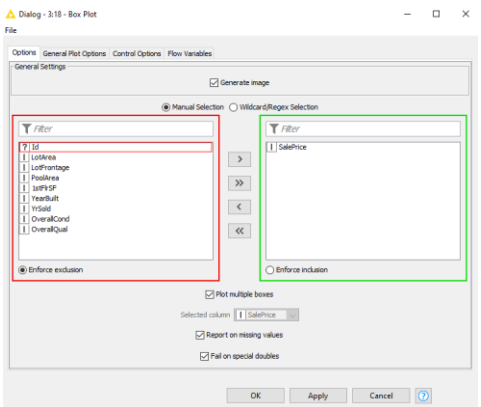


Diagrama de caja de la variable *LotArea*:



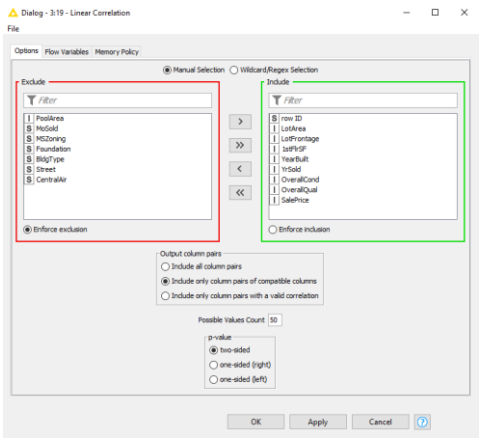
Cálculo de *outlier* (nodo *Numeric Outliers*) y diagrama de caja de la variable *SalePrice*:



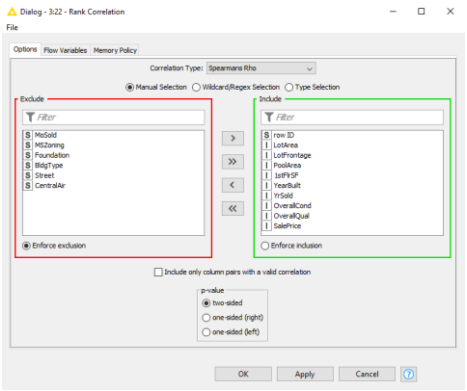


6.3.2 Análisis exploratorio gráfico multivariante

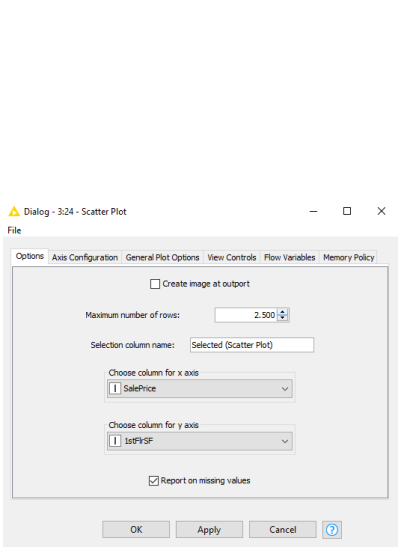
- Matriz de correlación de Pearsont[33], se utiliza el nodo *LinearCorrelation* donde se incluyen las variables numéricas:

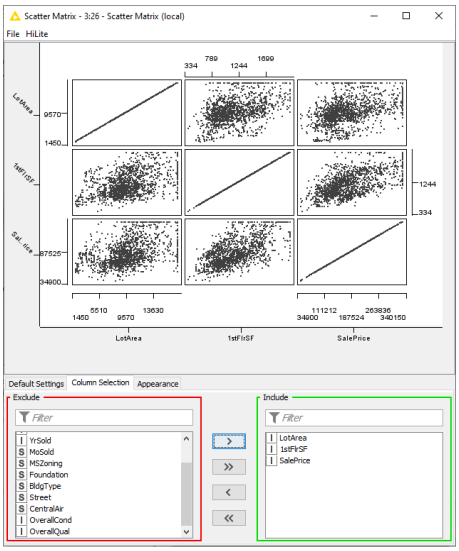


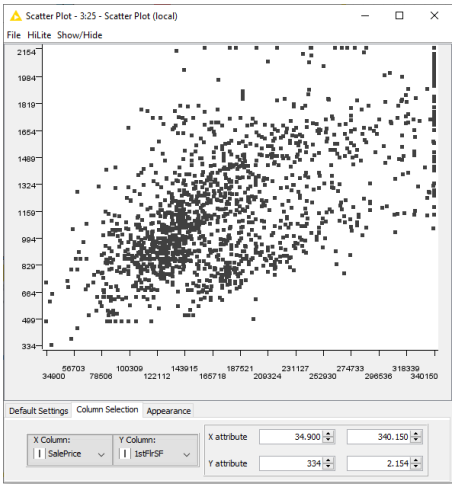
- Matriz de correlación de rango de Spearman[34], se utiliza el nodo *Rank Correlation* donde se incluyen las variables numéricas:



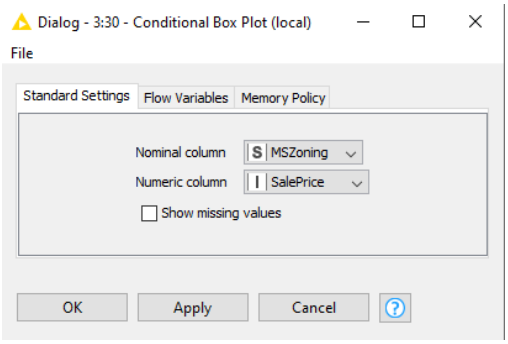
- Scatter Plot[35],se utilizan distintos nodos, *Scatter Plot*, *Scatter Matrix (local)* y *Scatter Plot (local)* con la siguiente configuración:







- Boxplot de dos variables (numérica y categórica), se utiliza el nodo *Conditional Box Plot (local)*, con la siguiente configuración:

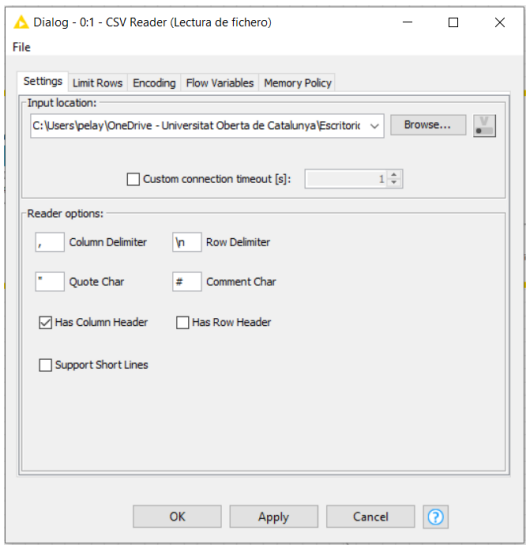


6.4 Flujo ML Supervisado. Estructura

6.4.1 Acciones previas

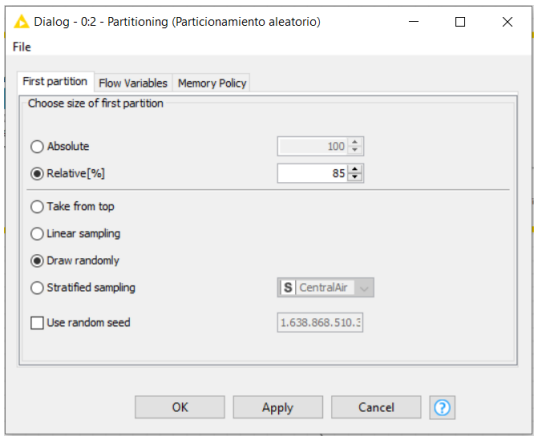
6.4.1.1 Ingesta y partición de los datos

1-Leer datos: se usa el paso *CSV Reader* indicando la delimitación por coma y que considere como primera columna.

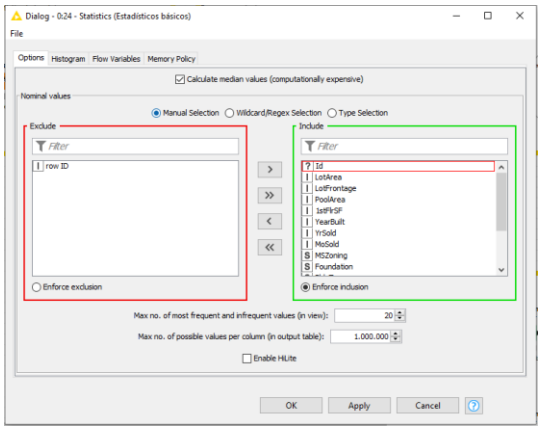


2-Scatter Matriz: paso configurable para mostrar las relaciones entre dos o más variables.

3-Particionar datos: Se usa el paso *Partitioning* para asignar los porcentajes correspondientes a los datos de entrenamiento, *Train*, y datos de comparación con los resultados del entrenamiento, *Test*.



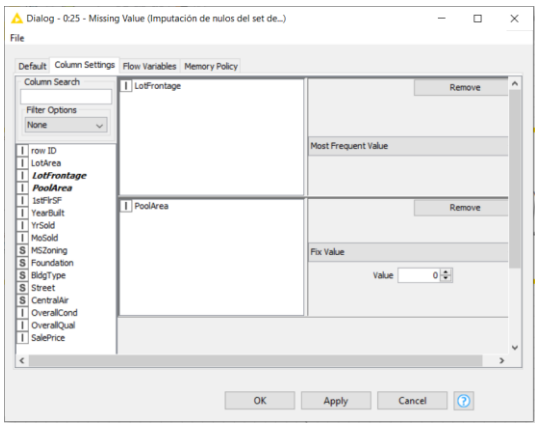
4-Estadísticos básicos: se calculan valores estadísticos, utilizando el paso *Statistics*, sobre los datos de entrenamiento que permitirán compararlos con los obtenidos sobre el conjunto total en los flujos anteriores (ver punto 7 de los apartados 2.1.1 y 6.2.1 de este documento).



6.4.1.2 Preprocesamiento

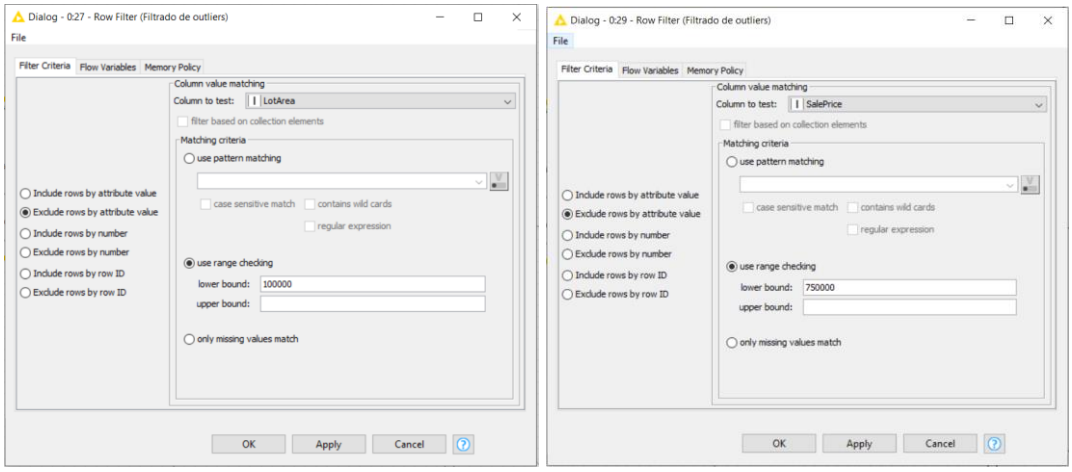
1-Imputación de nulos del set de entramiento, como en el apartado 6.2.1, se opta por configurar el paso *Missing Value* de la siguiente manera:

- *LotFrontage:* rellenar con el valor más frecuente
- *PoolArea:* rellenar con 0



2-Filtrado de Outliers (sobre LotArea y SalesPrice)

Se utilizará el paso Row Filter:

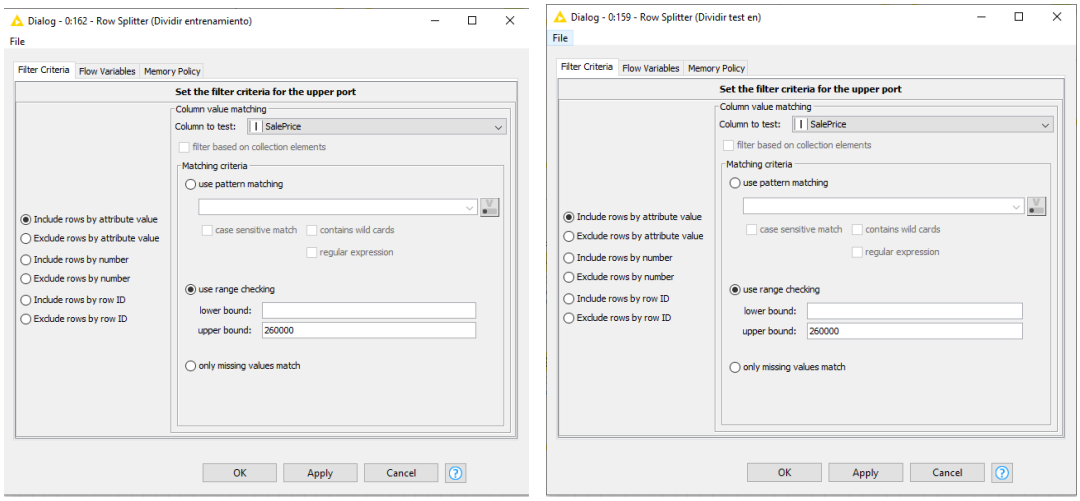


3-Visualización del resultado del preprocesado

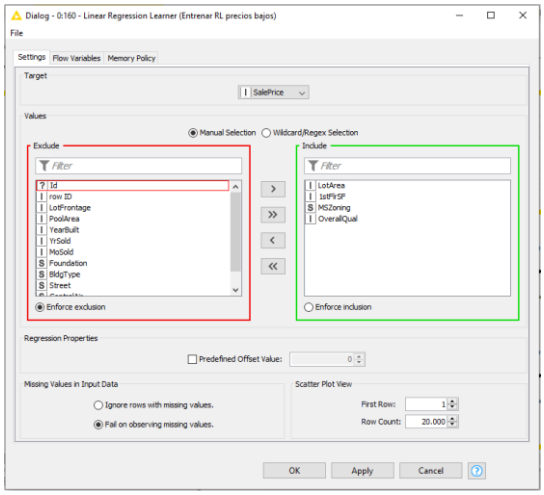
Se utilizan pasos *Scatter Plot (local)* permitiendo configurar el grafico con dos o más variables, ver punto 6.3.2

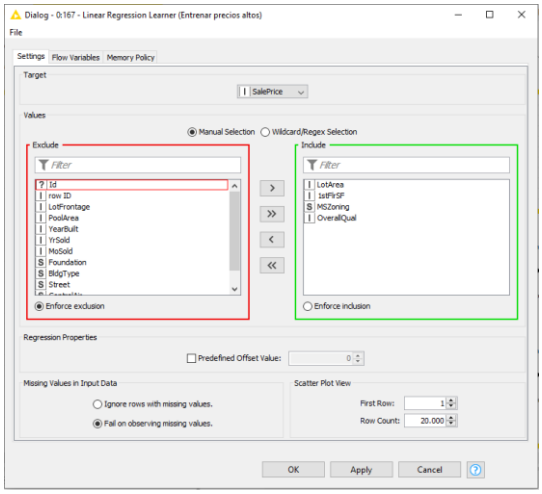
6.4.2 Regresión lineal Multivariante

1-Segmentación: se utiliza el paso *Row Splitter* para los dos conjuntos de datos, *train* y *test*:

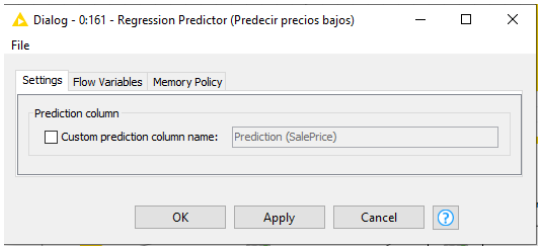


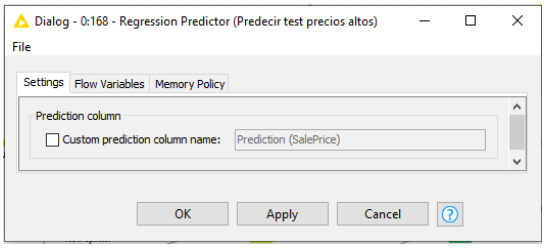
2- Creación del modelo: se utiliza el paso *Linear Regression Learner* para los dos conjuntos de datos de train, precios altos y bajos:



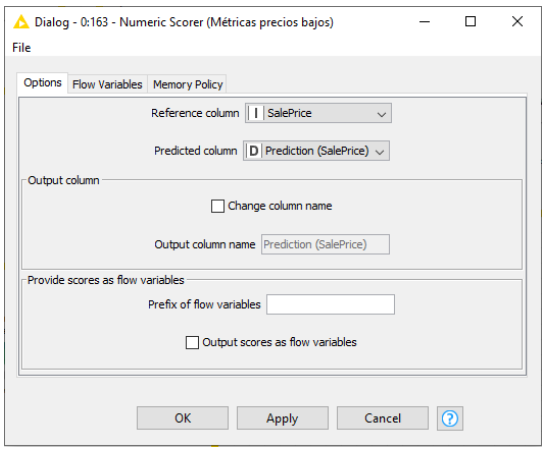


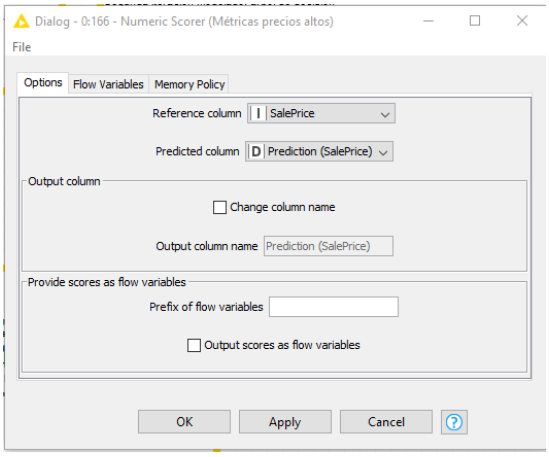
3- Evaluación de la predicción: se utiliza el paso *Regression Predictor* juntando los datos del modelo regresor con los de *test* para ambos conjuntos, precios bajos y altos:





4-Estadísticas y métricas: para obtener las estadísticas entre el precio de test y el obtenido mediante la aplicación del modelo predictivo se utiliza el nodo *Numeric Scorer*



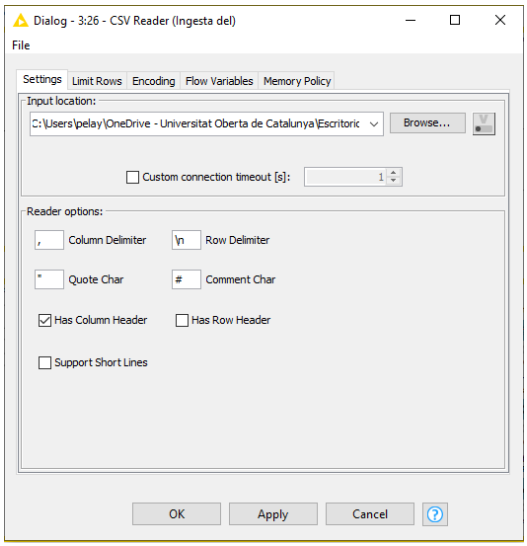


6.5 Flujo ML No Supervisado. Estructura

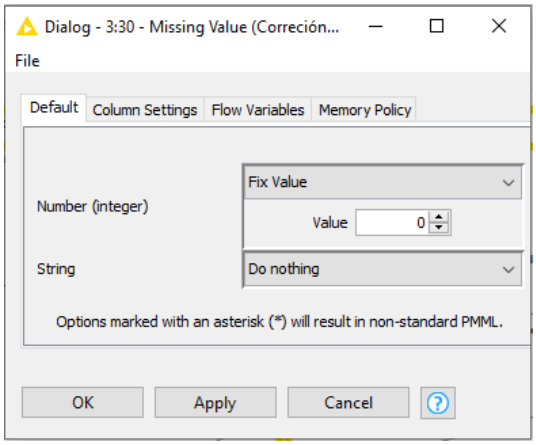
6.5.1 Acciones previas

6.5.1.1 Ingesta y corrección de nulos

1-Leer datos, se usa el paso *CSV Reader* indicando la delimitación por coma y que considere como primera columna:



2-Corrección de valores nulos, se utiliza el paso Missing Value aplicado a todas las variables, asignando un 0 donde se encuentro un null:

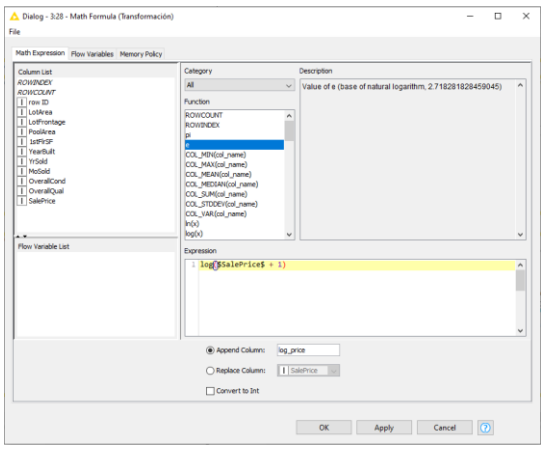
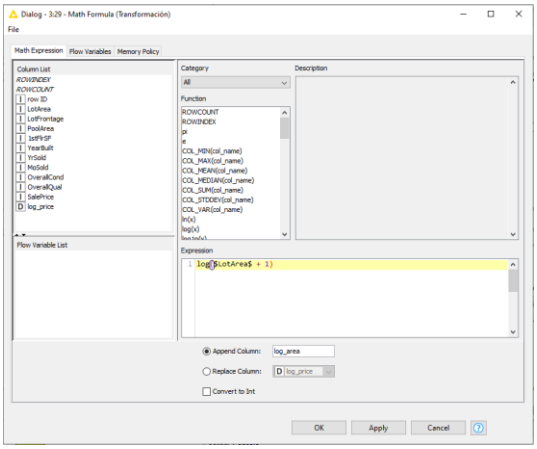


3-Visualizar los datos, se utiliza el paso configurable *Scatter Plot (local)* que muestra las relaciones entre dos variables.

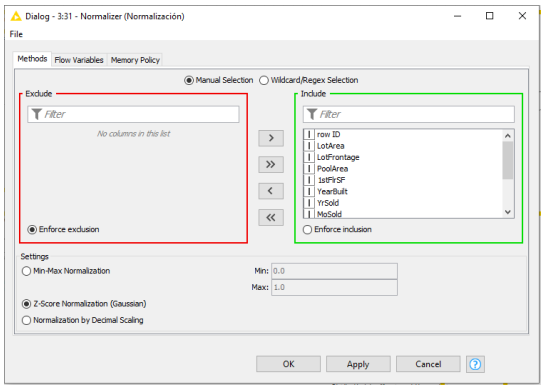
4-Distribución de las variables, se utiliza el paso *Box Plot (local)* que muestra diagramas de cajas en paralelo por cada variable.

6.5.1.2 Preprocesamiento

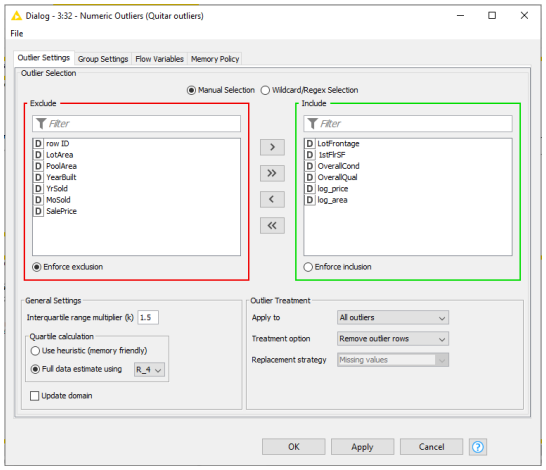
1-Transformar variables[46] *LotArea* y *SalePrice*, se utiliza el paso *Math Formula* con la siguiente configuración:

2-Normalización estándar, se utiliza el paso *Normalizer* aplicado a todas las variables numéricas:



3-Eliminación de *outliers*, se utiliza el paso *Numeric Outliers* configurandolo para que se aplique a las variables numéricas que se consideren, cuidado que si se aplica demasiadas veces a una variable puede eliminar valores relevantes:

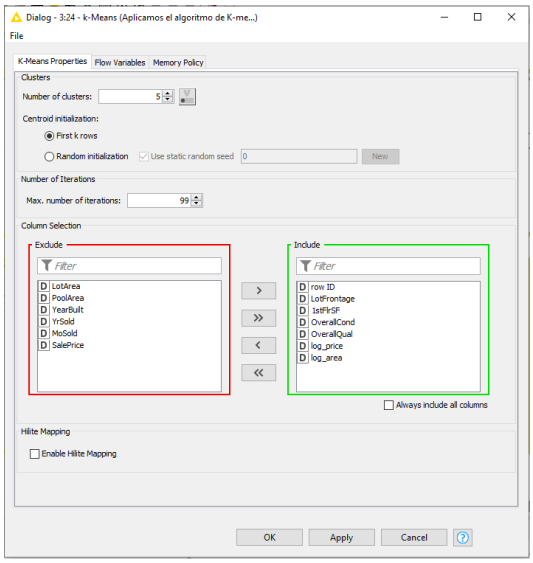


3-Visualización distribuciones, se utiliza el paso *Box Plot (local)*, aplicándolo en dos puntos después de la normalización, antes y después de la eliminación de outliers y así permitirá comparar.

6.5.2 K-means

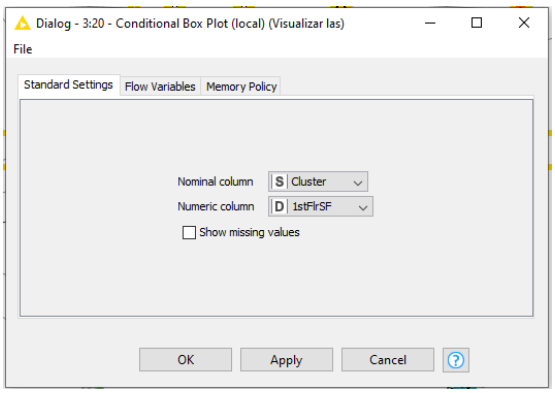
6.5.2.1 Aplicación e interpretación

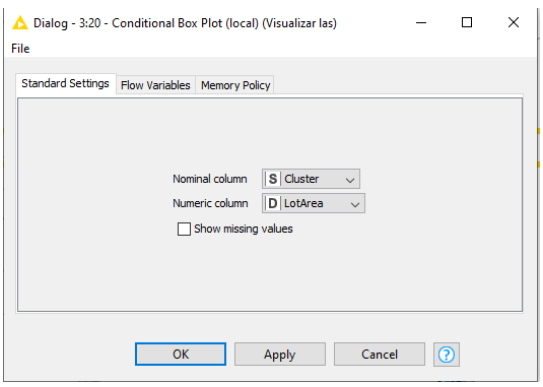
1-*Clustering*, se utiliza el paso *k-means* configurado con las variables elegidas y el número de *clusters*:



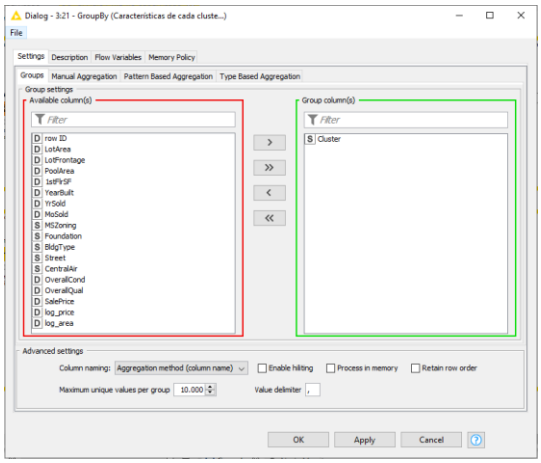
2-Deshacer la normalización, se utiliza el paso *Denormalizer* al que se le pasa como entrada los datos normalizados antes de la aplicación del modelo y los datos obtenidos después.

3-Visualizar las diferencias por *cluster*, se utiliza el paso *Conditional Box Plot (local)* cambiando la variable numérica que se quiere mostrar en los 5 *clusters:*

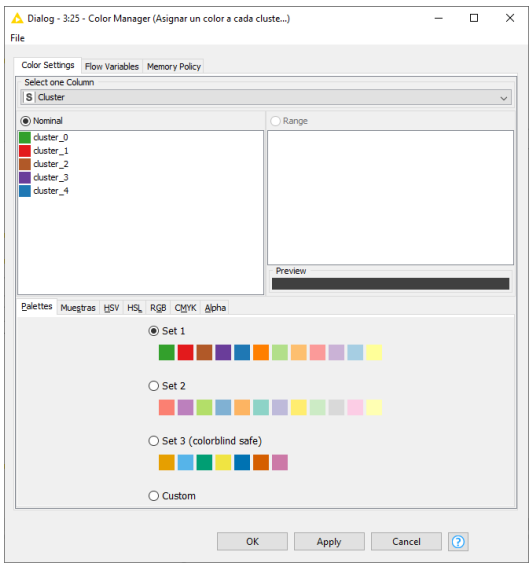




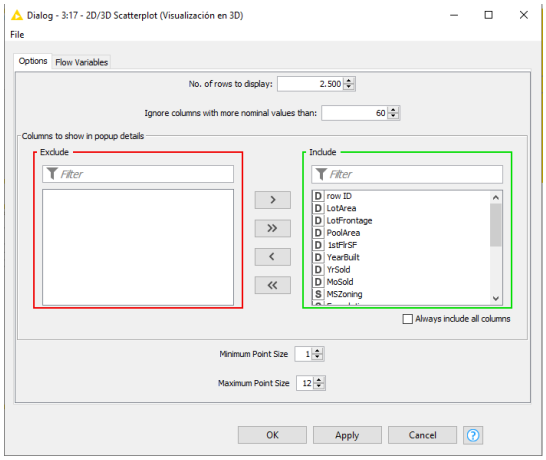
4-Valores medios, se utiliza el paso *GroupBy* configurando la agrupación de las observaciones por la nueva columna cluster:

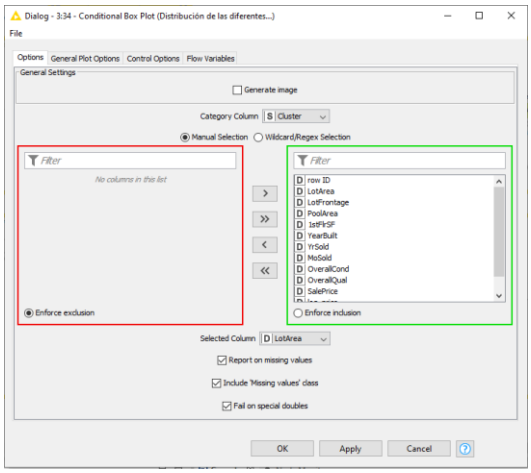


5-Asignar color a cada cluster y visualizar en 2D y 3D, para asignar color se utiliza el paso *Color Manager*:



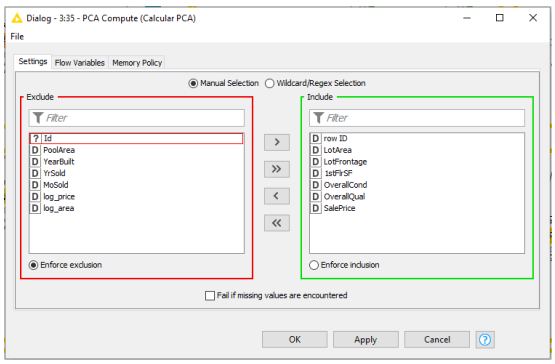
Para visualizar los datos en 2D se utliza el paso *Scatter Plot (local)*, en 3D el paso *2D/3D Scatterplot* y para ver la distribución de las diferentes variables por cluster se usa el paso *Conditional Box Plot* (en la configuración se debe seleccionarla variable número/columna que se quiere mostrar en los 5 *clusters:*



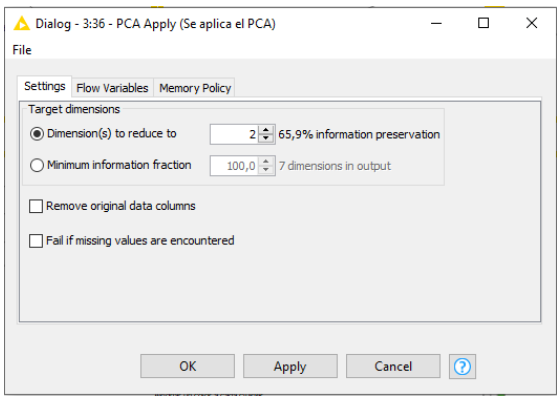


6.5.2.1 Métricas de evaluación y PCA

1-Calcular PCA, se utiliza el paso *PCA Compute* sobre las variables numéricas:



2-Aplicar PCA, se utiliza el paso *PCA Apply* indicando a cuantas dimensiones se quiere reducir (2)



3-Dar color a los cluster y visualizar, se utilizan los pasos *Color Manager* y *Scatter Plot (local)*: