Catalina Chavarría

  Curso: Machine Learning

Caso de análisis 2: Análisis exploratorio (train)

Contenido

[**Introducción** 2](#_Toc149248264)

[1. Preparando datos 2](#_Toc149248265)

[2. Desarrollo 3](#_Toc149248266)

[**Conclusión** 13](#_Toc149248267)

# **Introducción**

Para este primer caso de análisis exploratorio de datos (EDA). Se utilizará el dataset Train en el cual se explorará sus datos y realizando análisis y creando mapas utilizando Python en este proyecto

## Preparando datos

Para este primer apartado importamos las librerías que necesitaremos para realizar este ejercicio.

**%matplotlib inline**: este comando se utiliza para mostrar gráficos o visualizaciones en lugar de abrir ventanas emergentes separadas.

**plt.style.use('bmh')**: este comando se utiliza para cambiar el estilo de las gráficas predefinidos.

Seguido carga el dataset train para realizar el análisis de la base de datos.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

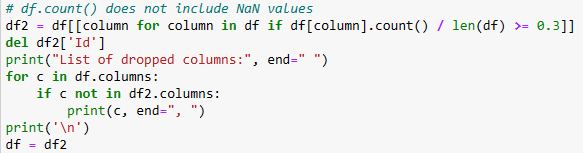
Descripción generada automáticamente

Imagen que contiene Interfaz de usuario gráfica

Descripción generada automáticamente

## Desarrollo

En la salida anterior donde nos muestra 5 filas de este dataset podemos ver varios valores NaN que no nos serán útiles para poder realizar análisis, por lo que en el siguiente código empezamos a eliminar valores de las columnas que contenga valores NaN que estén entre el rango de 30% o menos



Dando de resultado eliminando de las siguientes columnas valores NaN.

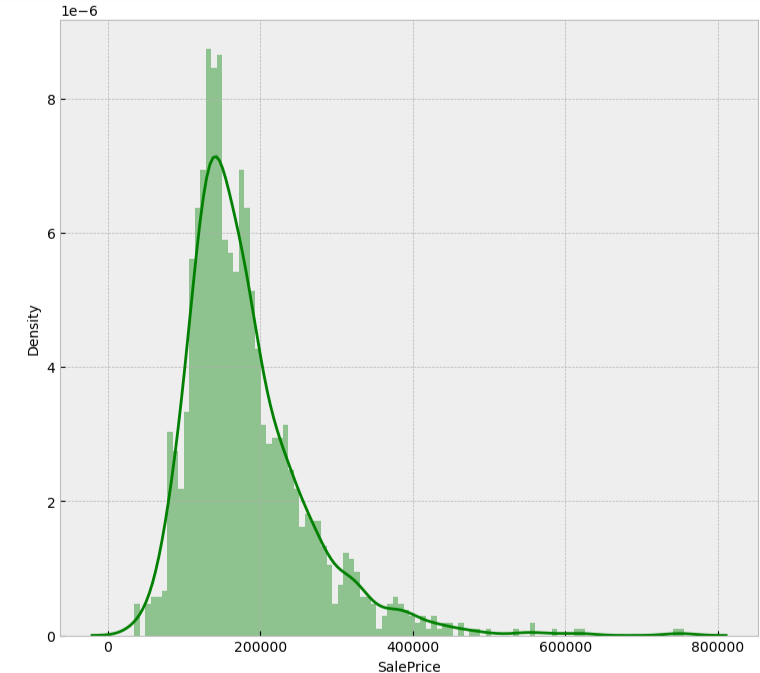


En este comando se usa para generar estadísticas y un histograma de la columna SalePrice.

Interfaz de usuario gráfica, Texto, Aplicación, Chat o mensaje de texto

Descripción generada automáticamente

El histograma muestra cómo se distribuyen los valores de la columna SalePrice en el conjunto de datos y proporciona información sobre su forma y dispersión.



En este apartado miramos la distribución de los datos de nuestro dataset.

Primero se lista todos los tipos de nuestra data mostrando solo las variables numéricas.

Tabla

Descripción generada automáticamente

Seguido de esto se genera un histograma de todas las columnas que enlistamos que son numéricas.

Mucha de su sintaxis es usando la variable **df\_num** que se estableció para las variables numéricas, el método **hist** para crear el histograma y seguidos parámetros para ajustar los gráficos, tanto en tamaño, etiquetas, los ejex, etc.

Diagrama

Descripción generada automáticamente

Como en el caso 1 este código calcula la correlación entre Pearson vs SalPrice.

Seguido identifica las columnas con correlación más fuerte con SalePrice. Esto nos ayuda para identificar las características numéricas más relevantes en relación con 'SalePrice' en un análisis de datos, dando en este caso 10 valores fuertemente correlación a SalePrice.

Interfaz de usuario gráfica, Aplicación, Word

Descripción generada automáticamente

Seguido de esta correlación y teniendo una lista de valores fuertemente correlacionado, se crea un conjunto de gráficos de dispersión para visualizar las relaciones entre la variable 'SalePrice'

Texto

Descripción generada automáticamente

El resultado son los siguientes gráficos de dispersión, esto permite visualizar las relaciones entre 'SalePrice' y grupos de cinco variables numéricas a la vez.Gráfico, Gráfico de dispersión

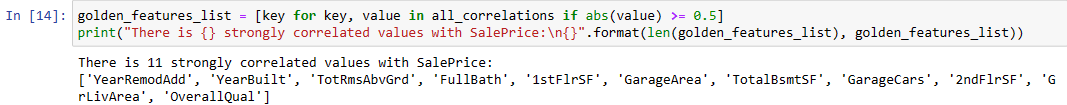
Descripción generada automáticamente

En el siguiente código se elimina los valores 0 y se usa para encontrar valores correlacionados:

Texto

Descripción generada automáticamente

Este código crea una lista llamada **golden\_features\_list** que contiene los nombres de las columnas que son numéricas y tiene correlación mayor o igual a 0.5 con la columna 'SalePrice'. Luego, muestra cuántas de estas características están fuertemente correlacionadas con 'SalePrice' y enumera sus nombres.



En el siguiente código crea un mapa de calor de las correlaciones entre las columnas numéricas, pero no usa SalePrice.

Imagen que contiene Texto

Descripción generada automáticamente

El resultado resalta las correlaciones más fuertes (positivas y negativas) entre las columnas numéricas lo que ayuda a identificar las relaciones significativas.



Este código crea un nuevo dataframe llamado **df\_quantitative\_values** que contiene solo las columnas numéricas enumeradas en la lista **quantitative\_features\_list**.

Imagen que contiene Tabla

Descripción generada automáticamente

Este código crea una lista llamada **features\_to\_analyse** que contiene los nombres de las características numéricas que se encuentran en la lista **quantitative\_features\_list** además de la columna 'SalePrice'.

Interfaz de usuario gráfica, Aplicación, Word

Descripción generada automáticamente

Seguido de esto creamos un conjunto de gráficos de dispersión para visualizar la relación entre cada característica en la lista features\_to\_analyse y la variable 'SalePrice'.

Texto

Descripción generada automáticamente

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Este código tiene como objetivo seleccionar y crear un nuevo DataFrame llamado **df\_categ** que contiene las columnas categóricas del DataFrame original df.

Captura de pantalla de computadora

Descripción generada automáticamente

Este código nos ayuda identificar y mostrar las columnas no numéricas en el DataFrame **df\_categ,** muestra cuántas columnas no numéricas hay y cuáles son sus nombres.

Texto

Descripción generada automáticamente con confianza media

Seguido creamos un gráfico de caja que visualiza la distribución de los precios de la columna **SalePrice** en función de la exposición del sótano **BsmtExposure** del dataframe df\_categ

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Creamos otro grafico de caja para visualizar la distribución entre SalePrice contra SaleCondition del dataframe df\_categ.

Texto

Descripción generada automáticamente

El gráfico de caja proporciona información sobre la dispersión entre SalePrice y SaleCondition, lo que facilita la comparación de las distribuciones entre estas columnas.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Y creamos el ultimo grafico este de barras para visualizar la distribución de las columnas categóricas de df\_not\_num.

Texto

Descripción generada automáticamente con confianza media

Cada gráfico nos muestra una categoría diferente, y la altura de las barras indica la frecuencia de esa categoría en los datos.

Gráfico en cascada

Descripción generada automáticamente con confianza media

# **Conclusión**

El EDA es un proceso crucial a la hora de realizar análisis de datos y poder explorar y comprender un conjunto de datos antes de realizar un análisis más profundo.   
Este enfoque nos ayuda a tomar las mejores decisiones y a obtener información valiosa de los datos.