EDA project summary

Manoj Koongahawatte

24/03/2025

The objective of this project was to perform EDA on a USA housing data set and provide a clean dataset which can be used for model development. I checked the features, records and size of the dataset and then identified the columns with null values, missing data and outliers. The dataset did not have any duplicate records. I differentiate the columns where naturally null values are possible to occur (e.g. fireplace) and the columns where it is impossible (e.g. square feet). Then I decided to impute mean or median values to the columns with numerical data types where it is inaccurate to contain missing data and null values. Calculating mean and median from a column where there are null/missing values and outliers is not accurate. Since that I created a data frame filtering records with all null/missing values from the original data set. After that I calculated the interquartile range (IQR), lower bound and upper bound for selected columns. Then I plot the data distribution of those selected columns within the lower bound and upper bound removing outliers. Ploting the histograms, I identified whether the data distribution of each of those columns were skewed or normal. I imputed the columns with mean value for normally distributed data and with median for skewed distributed data columns. Finally, I plotted a correlation graph for each 'imputed dataset including outliers' and 'imputed dataset excluding outliers' to visualize the relationships.

1. I completed this project on Google-Colab and Python 3.
2. I used Pandas Library to import CSV file and to create data frames.
3. I created 3 functions to:
    a. Filter nulls and zeros, calculate the IQR, filter the data and plot histogram of selected columns.
    b. Calculate the mean value, impute the normally distributed columns and update the data frame.
    c. Calculate the median value, impute the columns where distribution is skewed and update the data frame.
4. I utilized 'IQR' approach to identify outliers.
5. I utilized Seaborn Library to plot correlation diagrams, histograms, box plot and pair plot.
6. I utilized 'Plotly express' library to plot box plots with pointer labels.

7. I created a presentation to visualize the EDA process utilizing Microsoft 365 PowerPoint.