Exploratory Data Analysis

Introduction:

As the junior data scientist, my responsibilities include preprocessing the data and presenting the modelling team with a dataset suitable for the model they will create. In this use case, we will perform EDA and initial cleaning on the dataset provided. The purpose of this document is to present the cleaned dataset to the modelling team along with the justification for each choice made during the cleaning process.

Data Description:

The dataset contains information about residential properties that were sold in a particular region. The dataset has 16 columns and 5001 rows. The columns include sold price, year built, bedrooms, bathrooms, square footage, garage size, lot size, number of fireplaces, taxes, and zipcode.

Data Cleaning Process:

Dropping Null Values and Duplicates:

To ensure that the data is clean and accurate, we must remove any null values and duplicates. We use the dropna() method to remove any rows with null values. The resulting DataFrame with null values removed is stored in data_clean. We use the drop_duplicates() method to get rid of any duplicate rows from the dataframe. The resulting cleaned data has 4974 rows and 16 columns.

Converting Data Types:

We noticed that some of the columns had the wrong data types. To fix this issue, we converted certain columns to integer or float data types as appropriate. The columns that were converted to integer data types include year_built, bedrooms, bathrooms, sqrt_ft, garage, lot_acres, fireplaces, and zipcode. The columns that were converted to float data types include sold_price and taxes.

Dropping Rows with Missing Data:

To ensure that the data is as accurate as possible, we drop any rows with missing data in numeric columns. This was done using a loop that iterates over the columns of the DataFrame and drops any rows with null/NaN values in numeric columns.

Removing Outliers:

We calculated the z-score for the sold_price column, which measures how many standard deviations away from the mean each value is. We then filtered the dataframe to only include rows where the z-score is between -3 and 3, which corresponds to the 99.7% confidence interval. Finally, we dropped the z-score column from the dataframe.

Conclusion:

In conclusion, we have successfully cleaned the dataset by removing null values and duplicates, converting data types, dropping rows with missing data, and removing outliers. The resulting cleaned dataset has 4974 rows and 16 columns. This cleaned dataset is now ready to be used for modelling.

Attached to this document are plots and charts to visualise the data.

We hope that this cleaned dataset and accompanying visualisations will be useful to the modelling team in creating an accurate and reliable model.