

Project Summary

This project involved comprehensive Exploratory Data Analysis (EDA) and data cleaning on a real estate dataset containing 5000 property records with 16 features. The analysis focused on understanding the dataset's structure, identifying and addressing missing values, detecting outliers, analysing feature distributions, and preparing the data for modelling. The cleaned dataset provides a solid foundation for predictive modelling of property prices.

Specific Actions and Technologies:

1. Utilized Python with Pandas, NumPy, Matplotlib, and Seaborn libraries for data analysis and visualization
2. Performed initial data exploration, including examining the first few rows and basic statistics of the dataset
3. Conducted missing value analysis, identifying HOA fees as the feature with the highest percentage of missing values
4. Implemented outlier detection using box plots and Q-Q plots for numeric features
5. Analysed feature distributions using histograms and count plots
6. Executed data cleaning and transformation, including handling missing values
7. Created visualizations of the transformed data to verify the effectiveness of the cleaning process
8. Generated a cleaned dataset saved as a CSV file for further modelling and analysis

This summary captures the key aspects of the work done in the Google Collab notebook, highlighting the technologies used and the main steps in the data analysis and cleaning process.