**EXPLORATORY DATA ANALYSIS USING PYTHON**

**HOUSING DATASET ANALYSIS**

**NIRANJANA KS**

**DA&DS**

**2025-MAY BATCH**

**(OFFLINE)**

**1.INTRODUCTION:**

This project aims to explore the patterns in housing prices using Exploratory Data Analysis (EDA) techniques in Python. The dataset includes detailed information on residential properties. It covers features like square footage, the number of bedrooms and bathrooms, the year built or renovated, location (ZIP code), and overall sale price. With statistical summaries and visualizations, the project seeks to uncover hidden patterns, trends, and relationships that affect property values. Understanding these key factors can help homeowners, real estate investors, and developers make informed decisions about pricing, purchasing, and investing in the housing market.

**2.AIM:**

The main goal of this project is to study housing data to find and understand the key factors that affect property prices. The project uses data cleaning, transformation, and feature engineering, followed by statistical analysis and visualization. It looks into the impact of various factors such as location, square footage, number of bedrooms and bathrooms, age of the property, and renovation status. The final aim is to reveal significant insights and trends that can aid in predicting property values accurately. These insights are important for real estate professionals, buyers, sellers, and investors who are making strategic choices.

**3. PROBLEM STATEMENT:**

In the changing housing market, finding the right price for a property is both important and difficult. Many factors influence a property's market value. These include location, square footage, age, number of bedrooms and bathrooms, renovation history, and even views or closeness to water. Overpricing a property can lead to less buyer interest and longer selling times, while underpricing can result in losses for sellers. Without clear, data-driven insights, stakeholders often rely on personal opinions or outdated valuation methods. This project aims to solve that problem by using exploratory data analysis (EDA) to find patterns and relationships in housing data. By pinpointing the most significant factors affecting pricing and measuring their impacts, the analysis offers valuable support for fair and clear pricing decisions. Real estate agents, buyers, investors, and developers can use these insights to make smarter choices in a competitive market.

**4.PROJECT WORKFLOW:**

* **Data Collection & Loading**
* Loaded the dataset using Python libraries such as Pandas and NumPy.
* **Data Cleaning**
* Filled missing values using statistical imputation.
* Treated outliers using the IQR method.
* Corrected unrealistic and inconsistent entries.
* **Feature Engineering**
* Created derived metrics:  
  • age\_of\_Property = 2014 - yr\_built  
  • price\_per\_sqft = price / sqft\_living  
  • df['is\_renovated'] = df['yr\_renovated'].apply(lambda x: 1 if x > 0 else 0)
* **Statistical Analysis**
* Performed descriptive stats,one way t-tests, two way t-test,one way Anova-test and chi-square tests to evaluate significance.
* **Exploratory Data Analysis (EDA)**
* Exploratory Data Analysis (EDA) included univariate (histograms, boxplots), bivariate (scatter plots, correlation heatmaps), and multivariate analysis (pairplots) to explore distributions and relationships between housing features.
* **Insight Extraction**
* Summarized findings to guide data-driven decision-making.

**5.DATA UNDERSTANDING:**

* **Dataset Overview**
  + The dataset holds detailed records of residential properties.
  + Includes both numerical and categorical variables.
* **Key Variables Include**
  + price: Sale price of the property
  + sqft\_living, sqft\_lot: Size of the home and lot
  + bedrooms, bathrooms, floors: Structural features
  + condition, grade, view, waterfront: Quality and appeal indicators
  + zipcode: Geographic location
  + yr\_built, yr\_renovated, date: Construction and transaction timelines
* **Data Types**
  + Numerical: price, square footage, year values, etc.
  + Categorical: condition, zipcode, waterfront, view
* **Size & Suitability**
  + Contains thousands of entries, making it suitable for strong exploratory data analysis and modeling.
* **Initial Observations**
  + Property size, location, and renovation status significantly impact price.
  + Homes near waterfronts or in high-end ZIP codes tend to have higher prices.
* **Distribution Insights**
  + Histograms revealed right-skewed distributions, especially for price.
  + Boxplots exposed outliers and visualized the spread of key variables.
* **Importance**
  + Understanding variable types and patterns guided:
    - Effective data cleaning
    - Feature selection
    - Model planning and design

**6. DATA CLEANING:**

Missing values in numeric columns were addressed by using median imputation. IQR method detected outliers in price and size, which were either removed or capped. We corrected inconsistencies, like the year built being 0 or greater than the current year. We also converted data types when necessary for proper analysis.

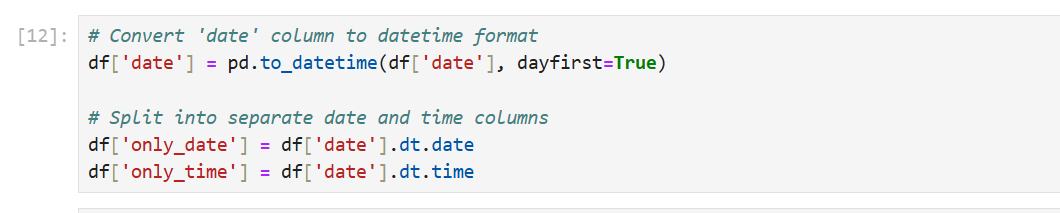
6.1 Fixing rows and coloumns

6.2 Check for missing values

6.3 Standardizing values

6.4 Check for outliers

**6.1 Fixing rows and coloumns**

Fixing rows and columns involved removing invalid or missing data and dropping irrelevant columns. Column names were renamed, data types corrected, and new useful columns were added for better analysis. 

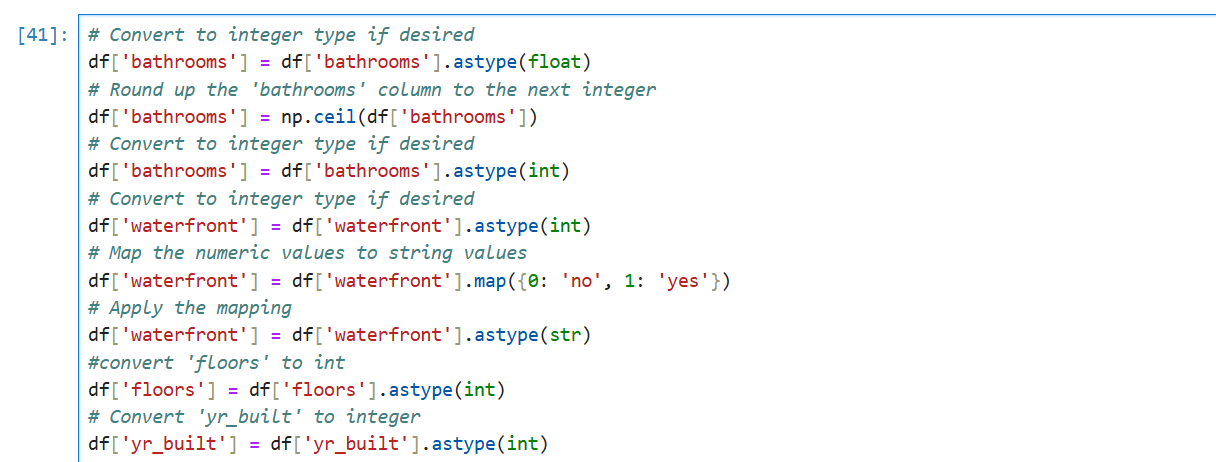
**6.2 Check for missing values**

Missing values were identified in numerical and categorical columns.They were handled using median or mode imputation to maintain data quality.



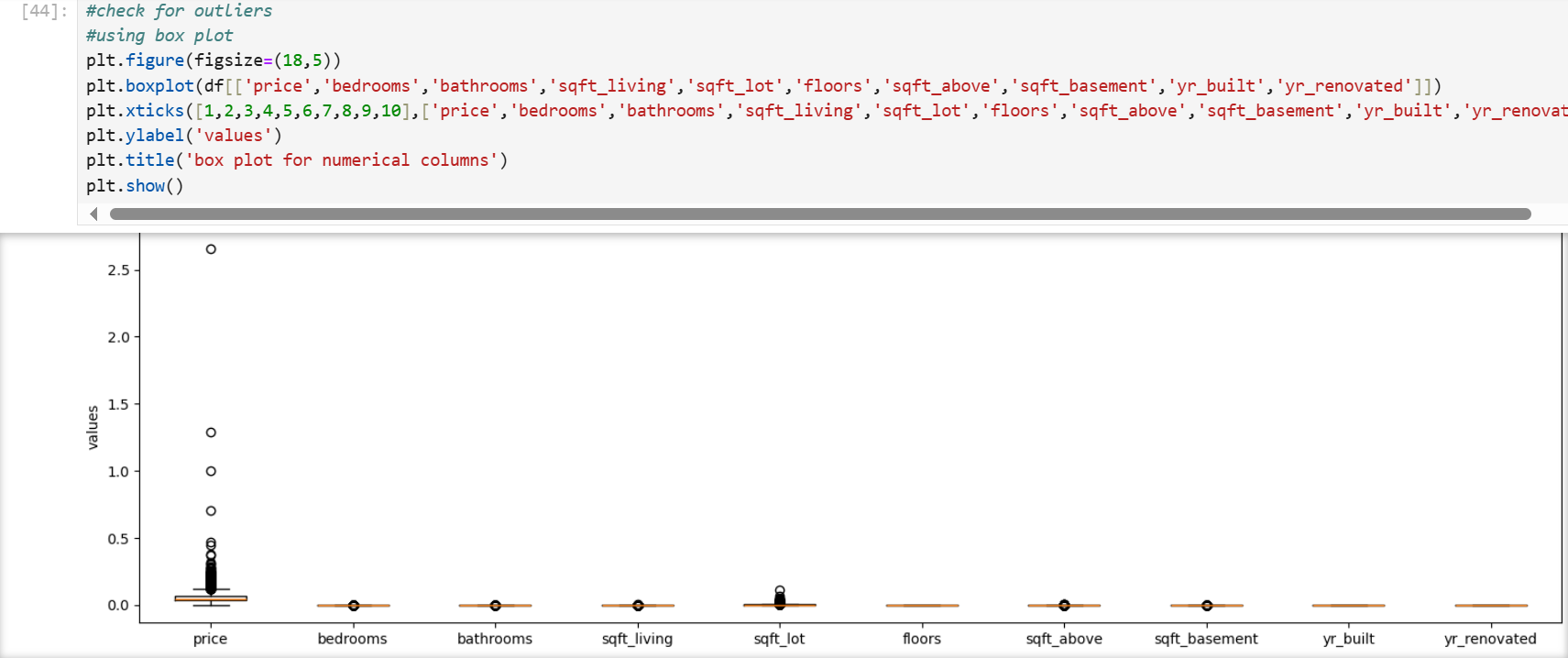
**6.3 Standardizing values**

Numerical columns were standardized to ensure consistent scale and units.  
This helped improve analysis accuracy and prepare the data for modeling. For numerical features, it helps normalize spread; for categorical data, label or one-hot encoding is used prior to standardization



**6.4 Check for outliers**

Outliers in features like price and sqft\_living were detected using the IQR method.  
Extreme values were removed or capped to reduce skewness and improve analysis quality**.**

****

**7. OBTAINING DERIVED METRICS:**

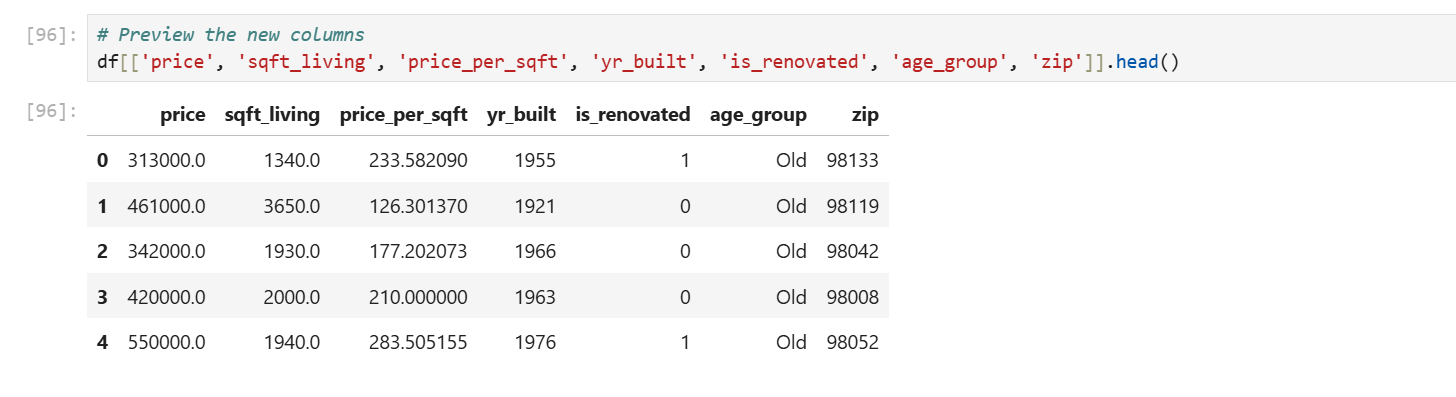
Created new features:

* age\_of\_Property: Current year minus yr\_built
* is\_renovated: Binary flag for whether yr\_renovated > 0
* price\_per\_sqft: Ratio of price to sqft\_living  
  These help in better normalization and analysis across diverse property types.

****

**8.FILTERING DATA FOR ANALYSIS:**

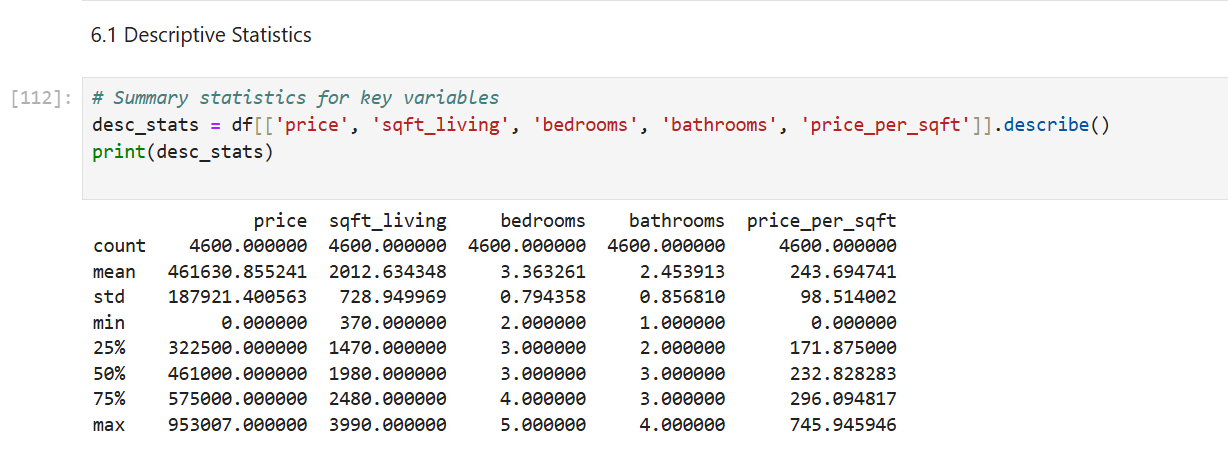
Filtered out records with extreme or unrealistic values, such as 0 bedrooms or sqft greater than 10000. Focused the analysis on residential properties within usual urban price ranges. Converted categorical variables to numerical when necessary**.**

****

**9.STATISTICAL ANALYSIS:**

**9.1 Descriptive Analysis:**

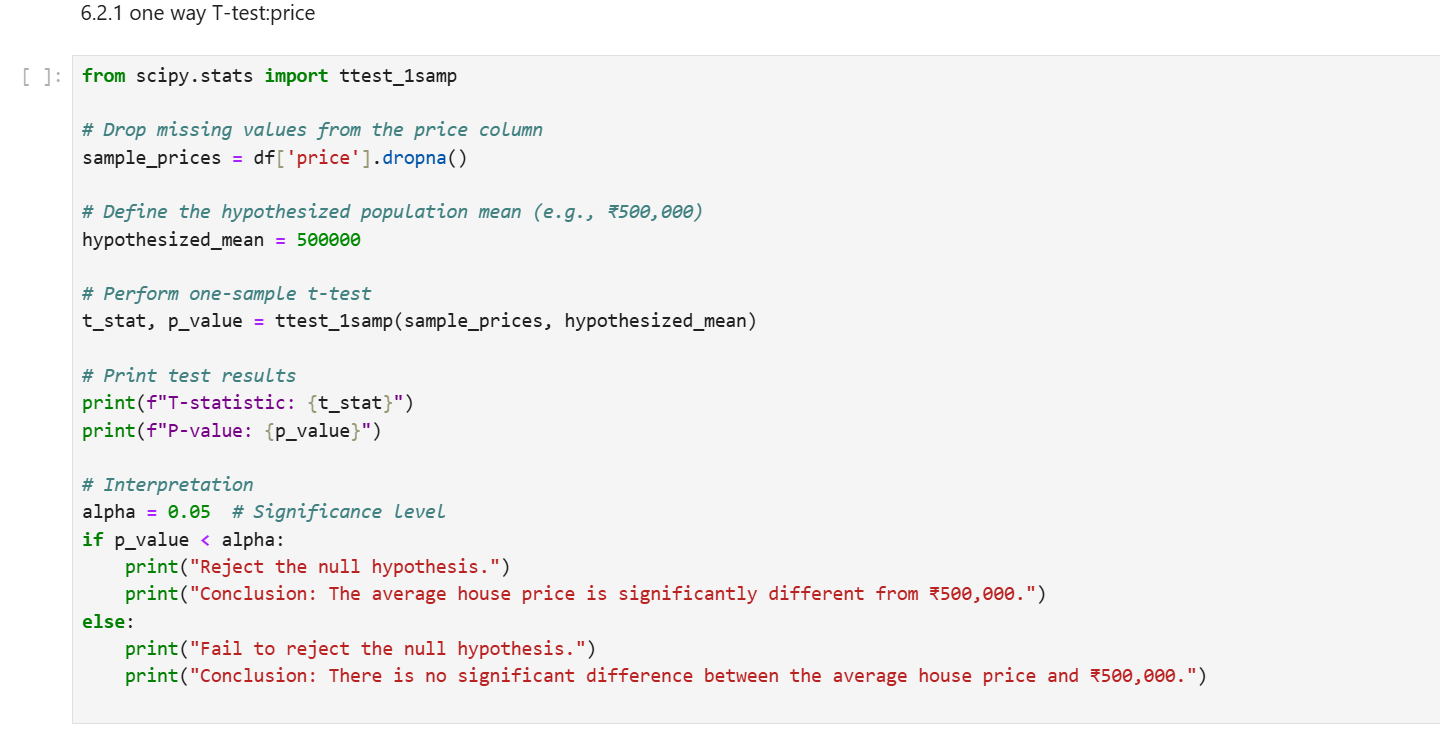
Summary statistics were calculated for key variables like price, sqft\_living, bedrooms, and price\_per\_sqft.  
Measures such as **mean**, **median**, **min**, **max**, and **standard deviation** helped understand the data distribution.

****

**6.2 T-Test: price vs is\_renovated**

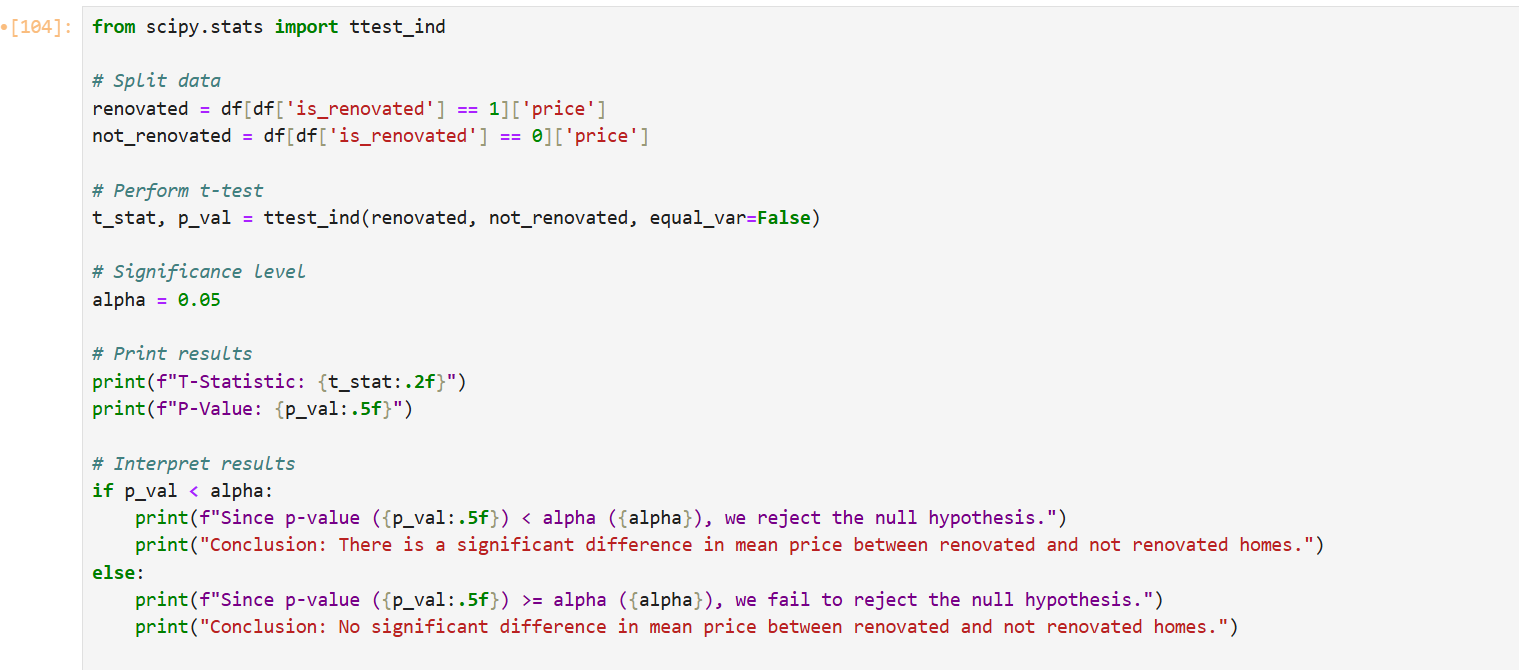
**6.2.1 one way T-test:price**

The null hypothesis (H₀) assumed the average house price is ₹500,000.  
 Since the p-value < 0.05, we reject H₀, indicating a significant difference from ₹500,000.



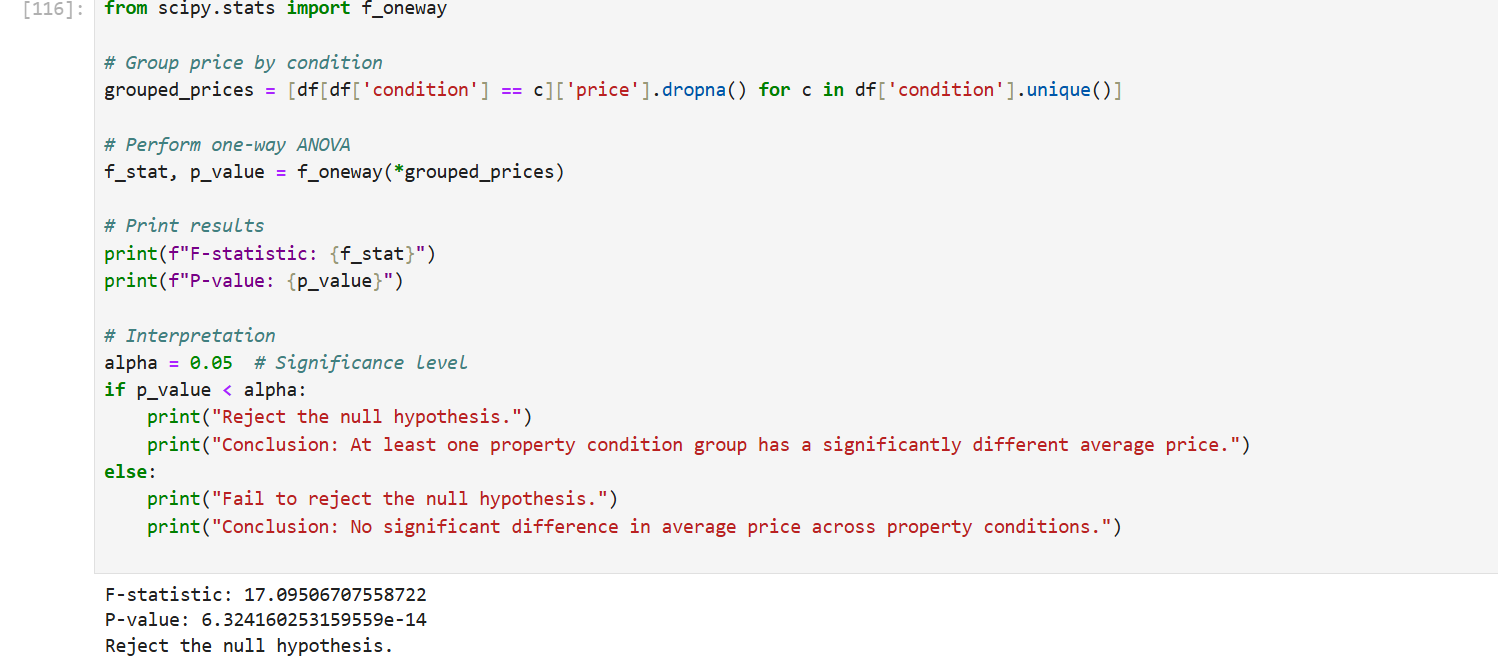
**6.2.2 two way T-Test: price vs is\_renovated**

The null hypothesis (H₀) assumed no difference in average prices between two groups (e.g., renovated vs. non-renovated homes).  
Since the p-value < 0.05, we reject H₀, indicating a significant difference in average prices between the groups.

****

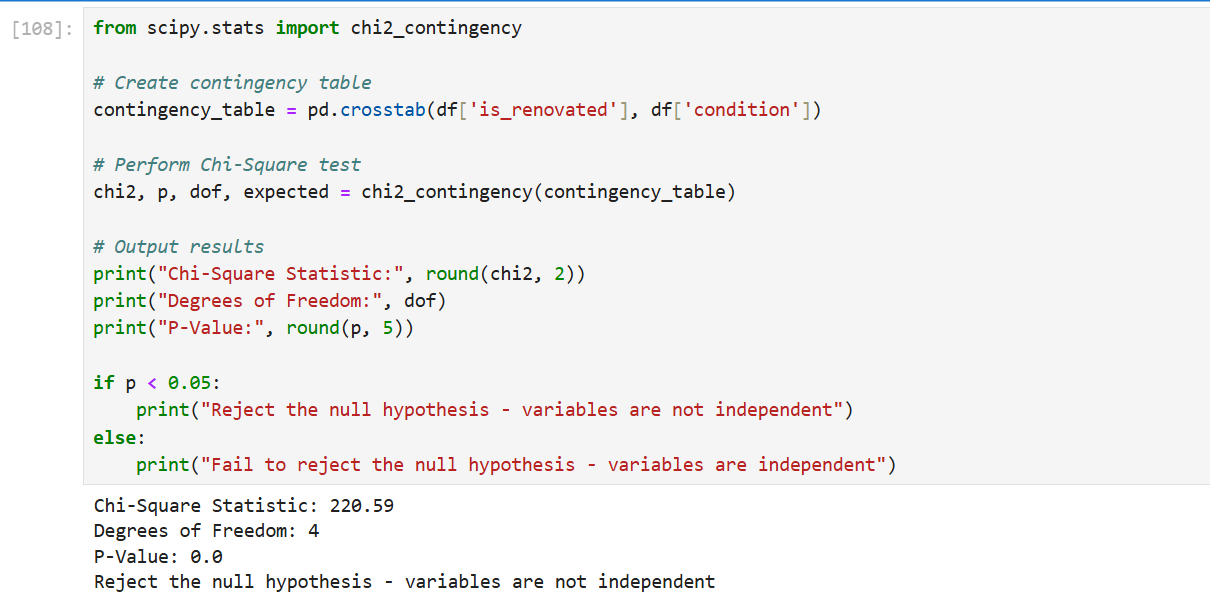
**6.3 One way ANOVA test: price vs condition**

A one-way ANOVA test was conducted to compare average prices across property conditions.  
Since the p-value < 0.05, we reject the null hypothesis, indicating that condition significantly affects house price.

****

**6.4 Chi-Square Test – Renovation Status vs Property Condition**

A Chi-Square test was used to examine the relationship between renovation status and property condition.  
The p-value < 0.05 indicates a significant association, meaning renovation status is linked to property condition.

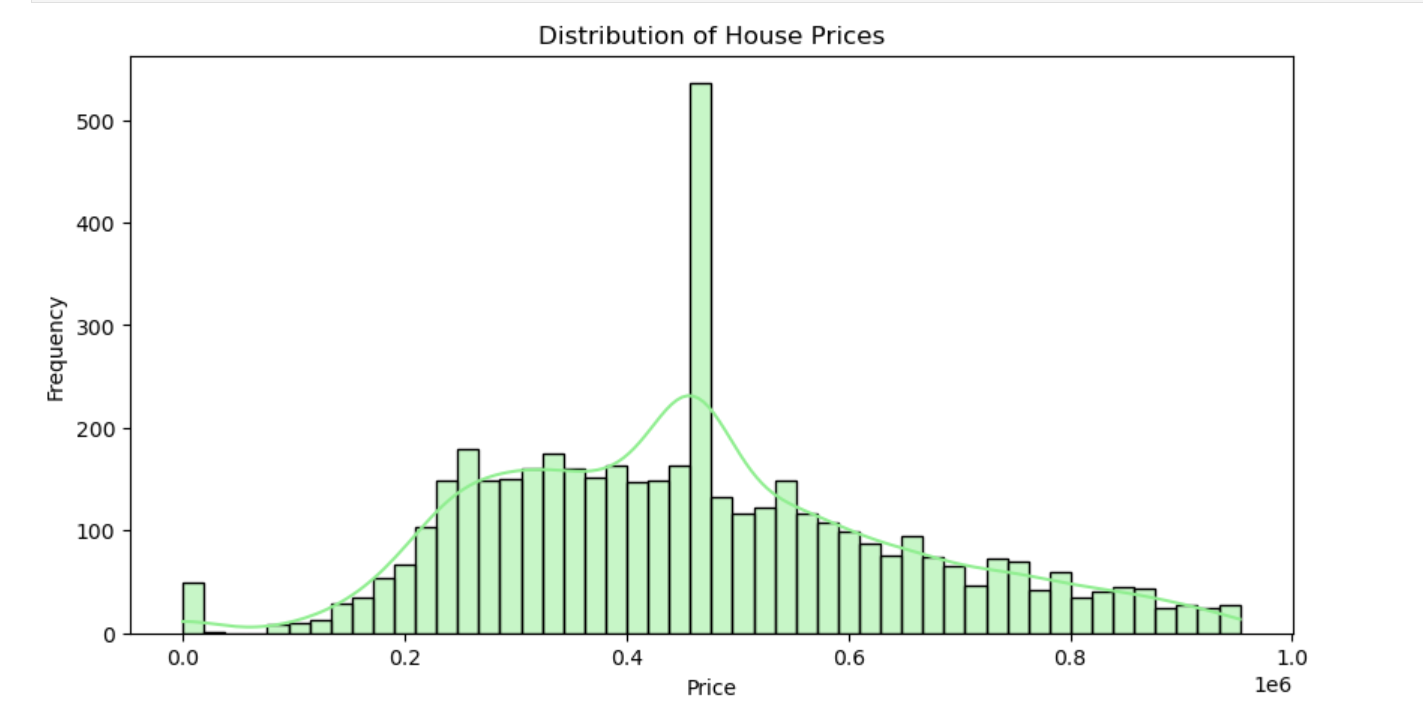
****

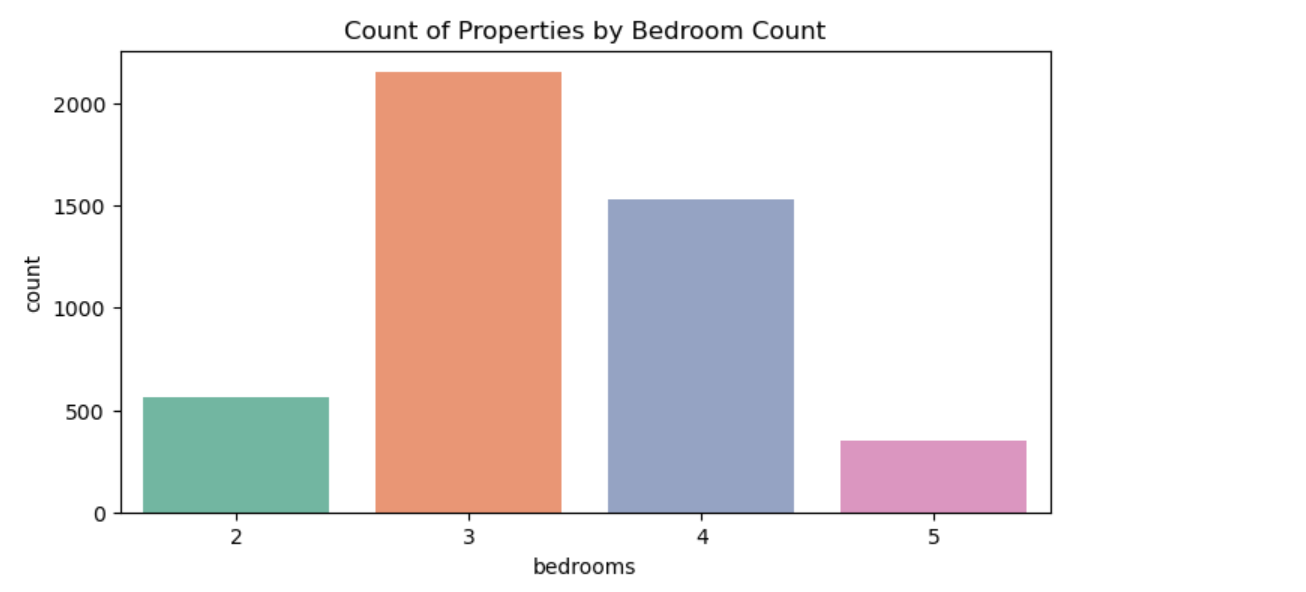
**10.EXPLORATORY DATA ANALYSIS:**

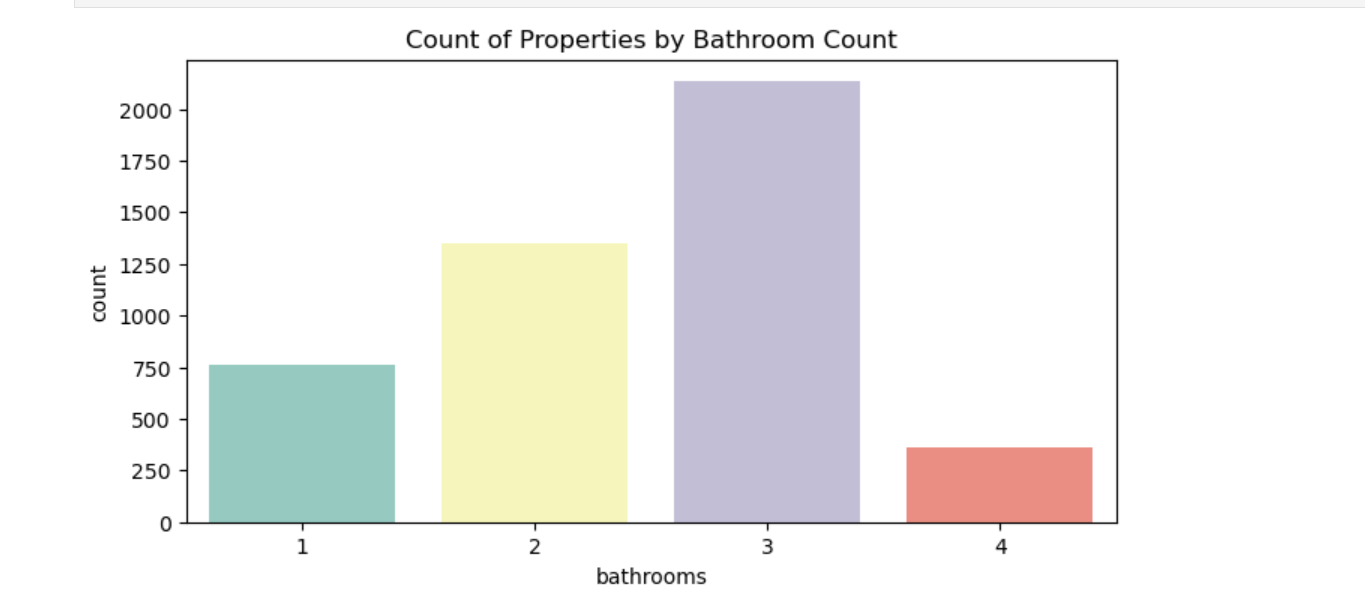
**10.1 UNIVARIATE ANALYSIS**

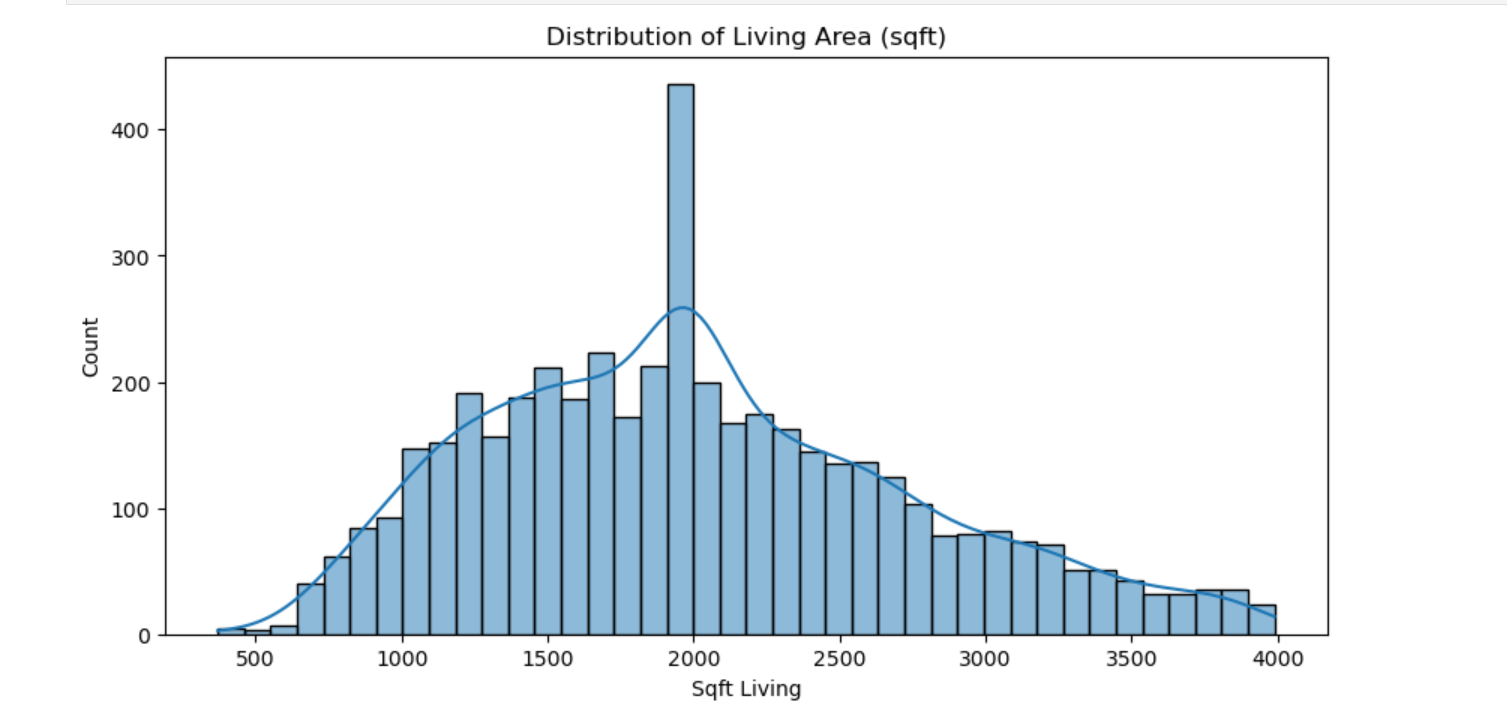
Univariate analysis was performed to understand the distribution and frequency of individual variables in the housing dataset. The following visualizations were used:

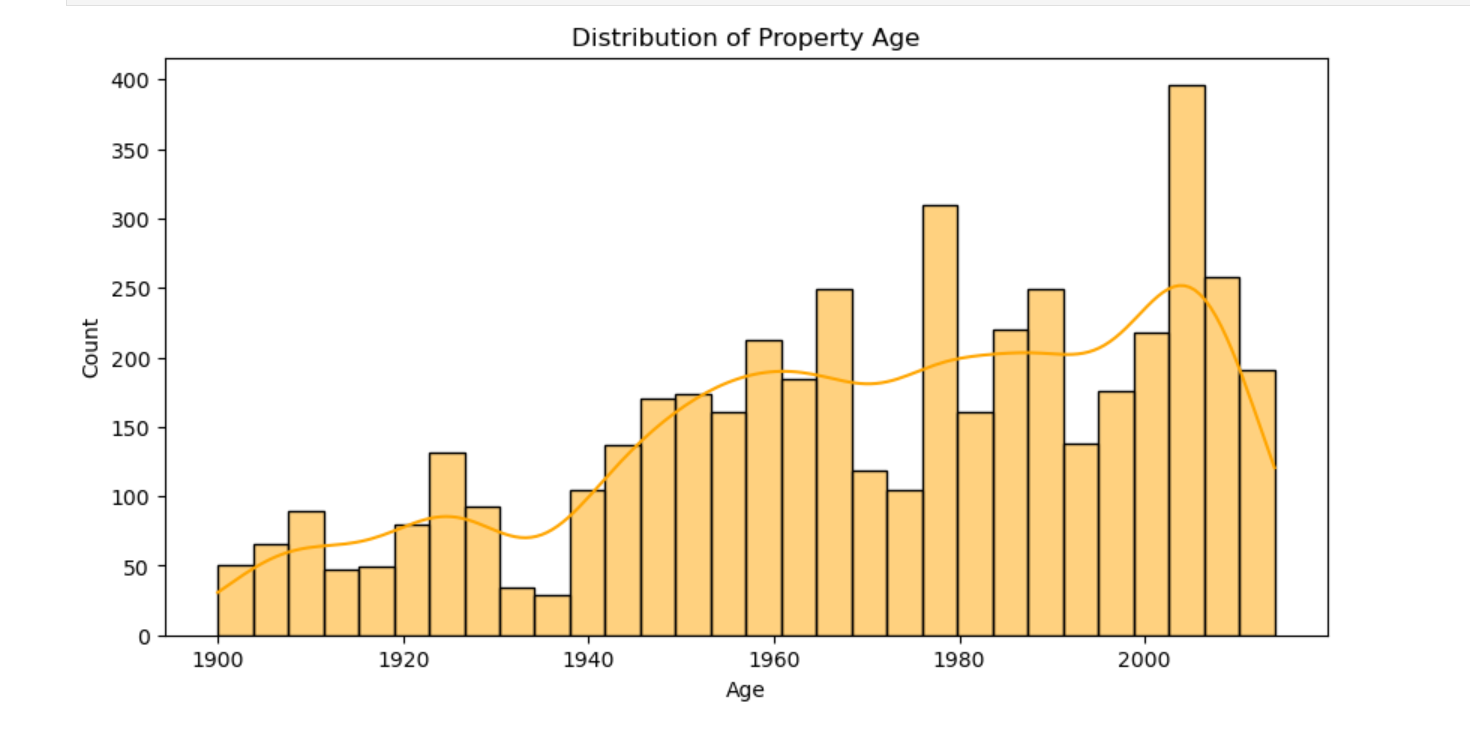
* **Price**: A **histogram** showed a right-skewed distribution, indicating that most homes fall in a mid-range, with few high-priced outliers.
* **Bedrooms**: A **countplot** revealed that most properties have 3 or 4 bedrooms, with very few having more than 6.
* **Bathrooms**: A **countplot** showed a similar trend, with 2 or 2.5 bathrooms being the most common.
* **Sqft\_living**: A **histogram** indicated that the majority of homes are between 1,000 and 2,500 square feet.
* **Year Built**: A **histogram** was used to observe the construction year distribution, showing a peak around 1970–2000.
* **Condition**: A **countplot** showed that most homes are in average or good condition.
* **View**: A **countplot** revealed that the majority of homes have no special view, with only a few enjoying premium scenic views.

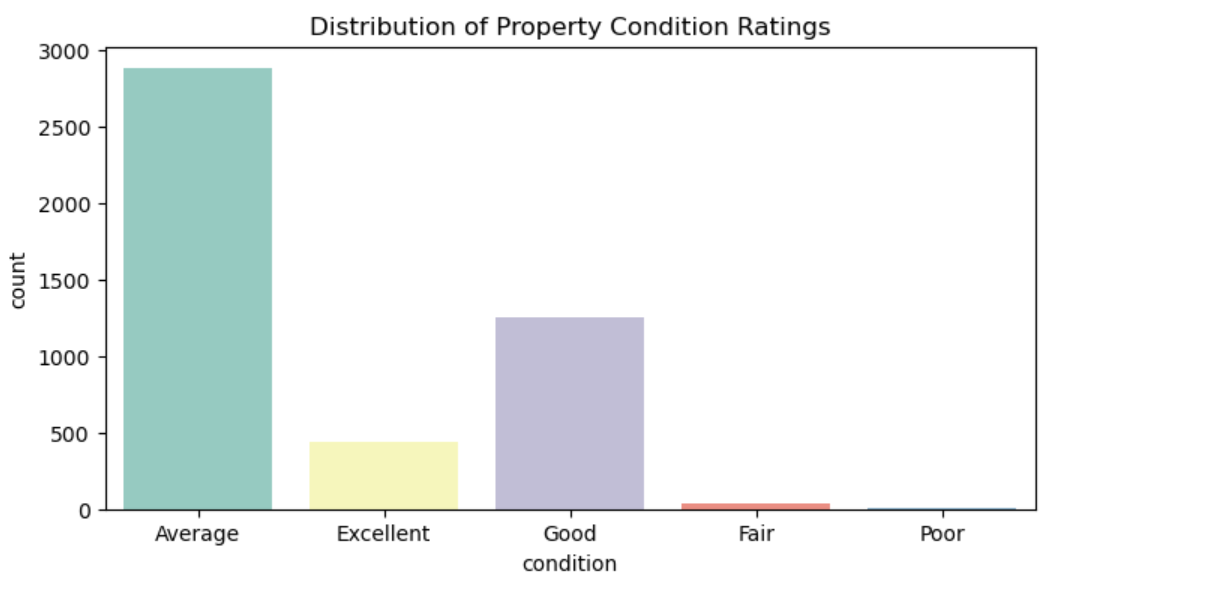
****

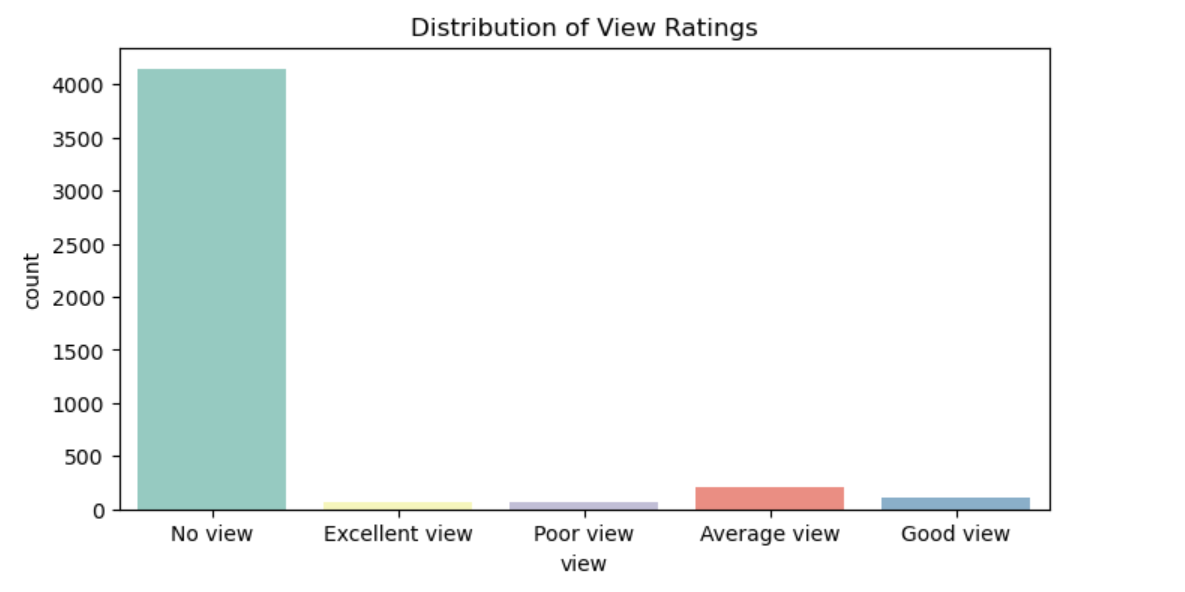
****

****

****

****

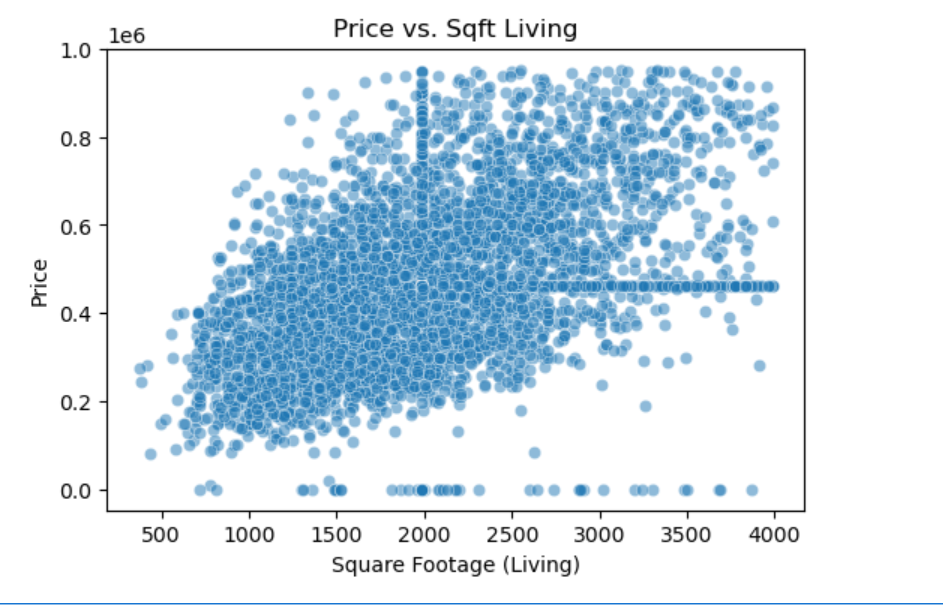


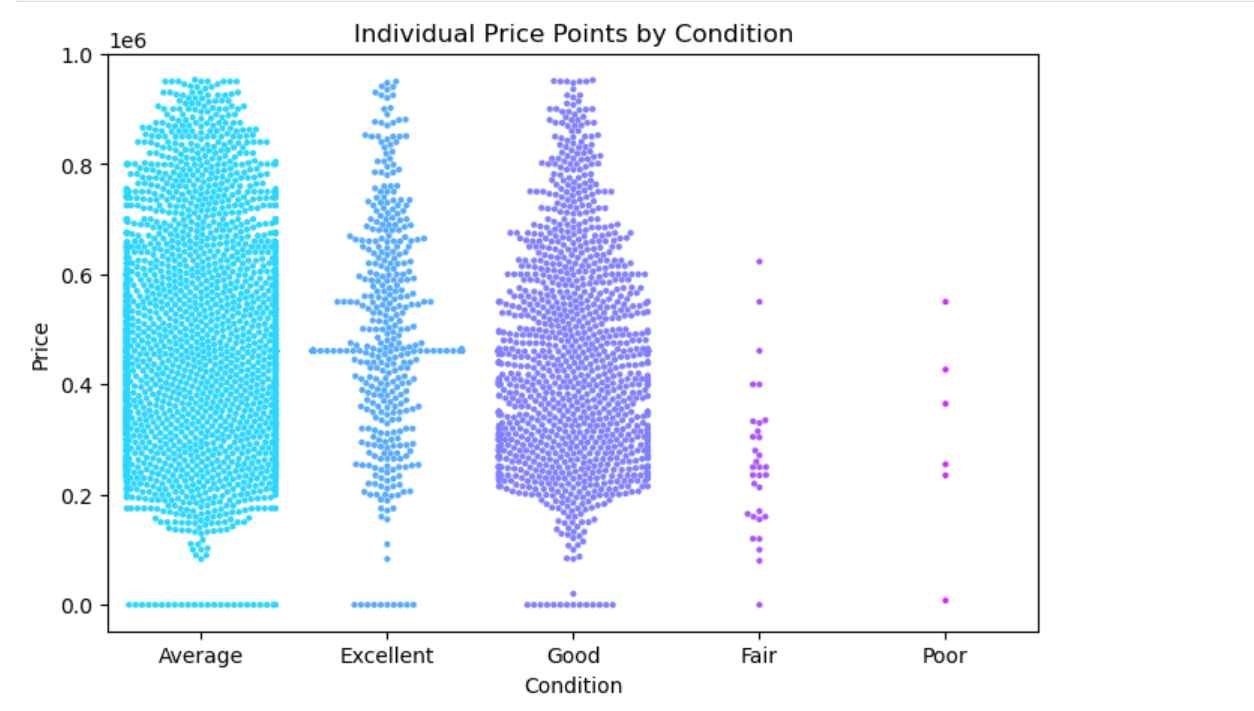


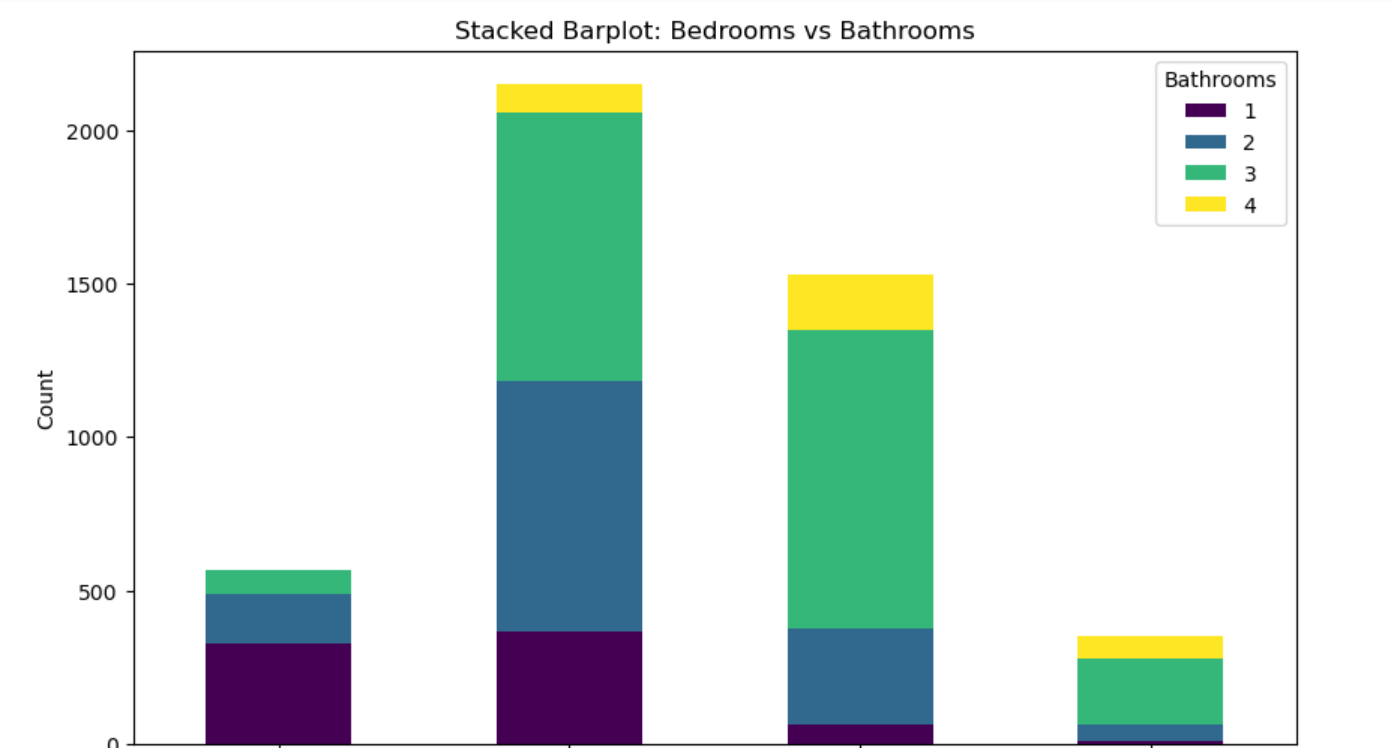
**10.2 BIVARIANT ANALYSIS:**

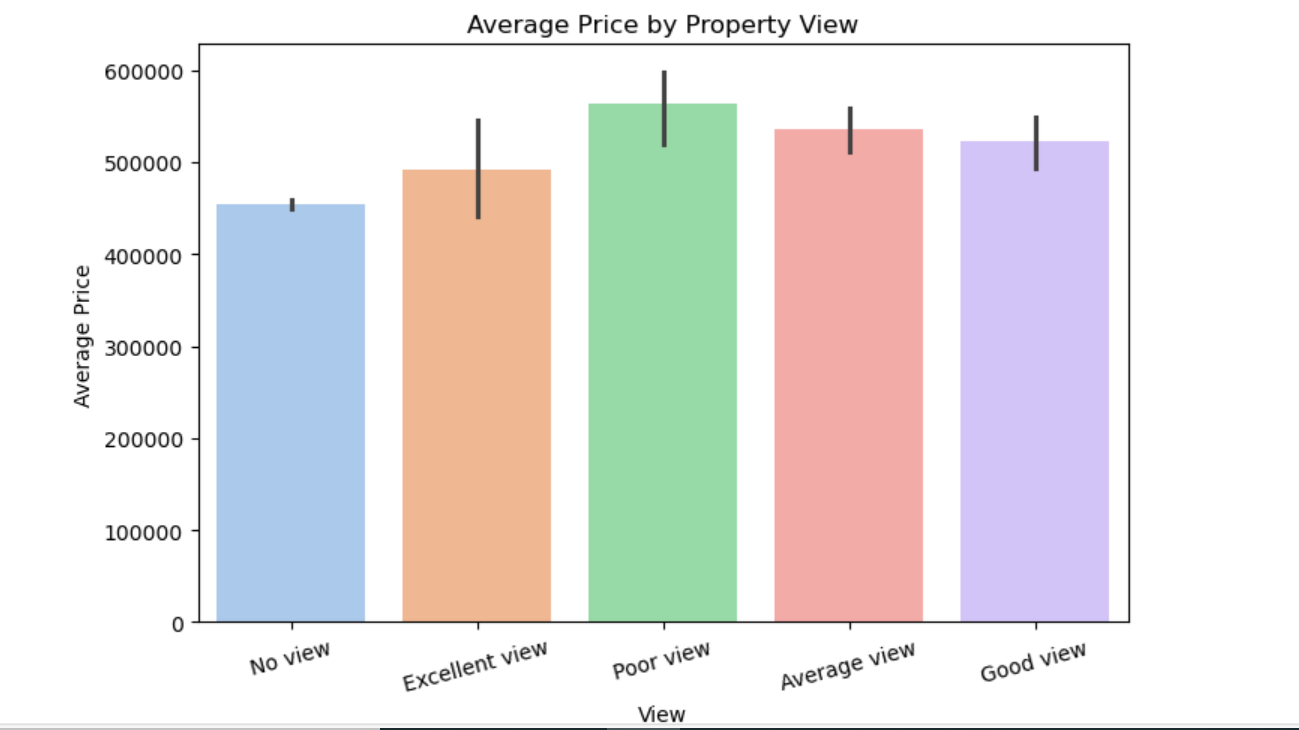
Bivariate analysis was conducted to examine how pairs of variables interact and influence each other, using various plot types:

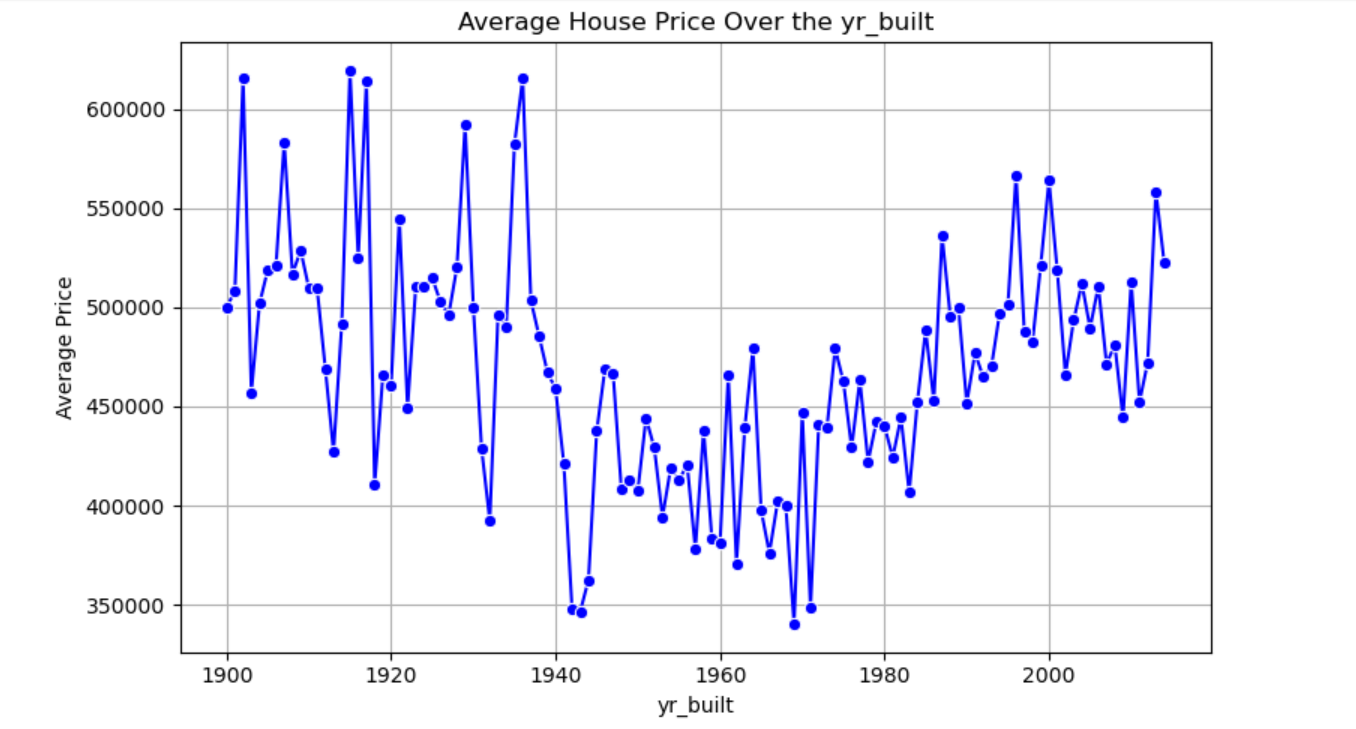
* Sqft\_living vs. Price *(Scatter Plot)*: Showed a strong positive correlation—larger homes tend to be more expensive.
* Price vs. Condition *(Swarm Plot)*: Indicated that homes in better condition are generally priced higher.
* Bedrooms vs. Bathrooms *(Stacked Plot)*: Revealed that higher bedroom counts usually come with more bathrooms, showing a structural dependency.
* Price vs. View *(Bar Plot)*: Showed that homes with better views have higher average prices.
* Year Built vs. Price *(Line Plot)*: Displayed that newer homes often have higher prices, though not in a perfectly linear trend.
* Condition Count *(Countplot)*: Showed most homes are in average or good condition, with very few in poor or excellent state.

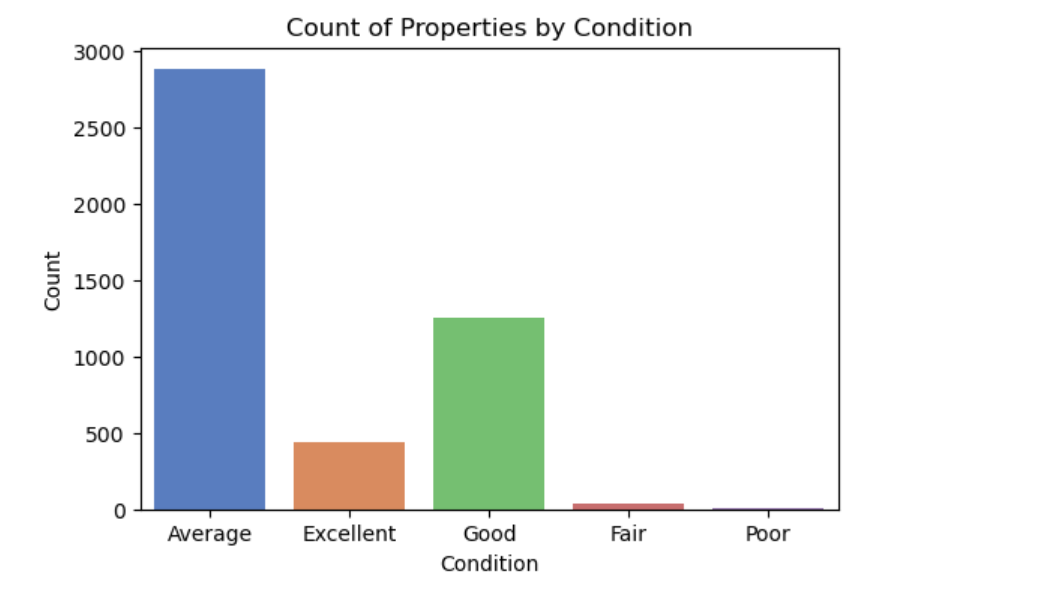
****





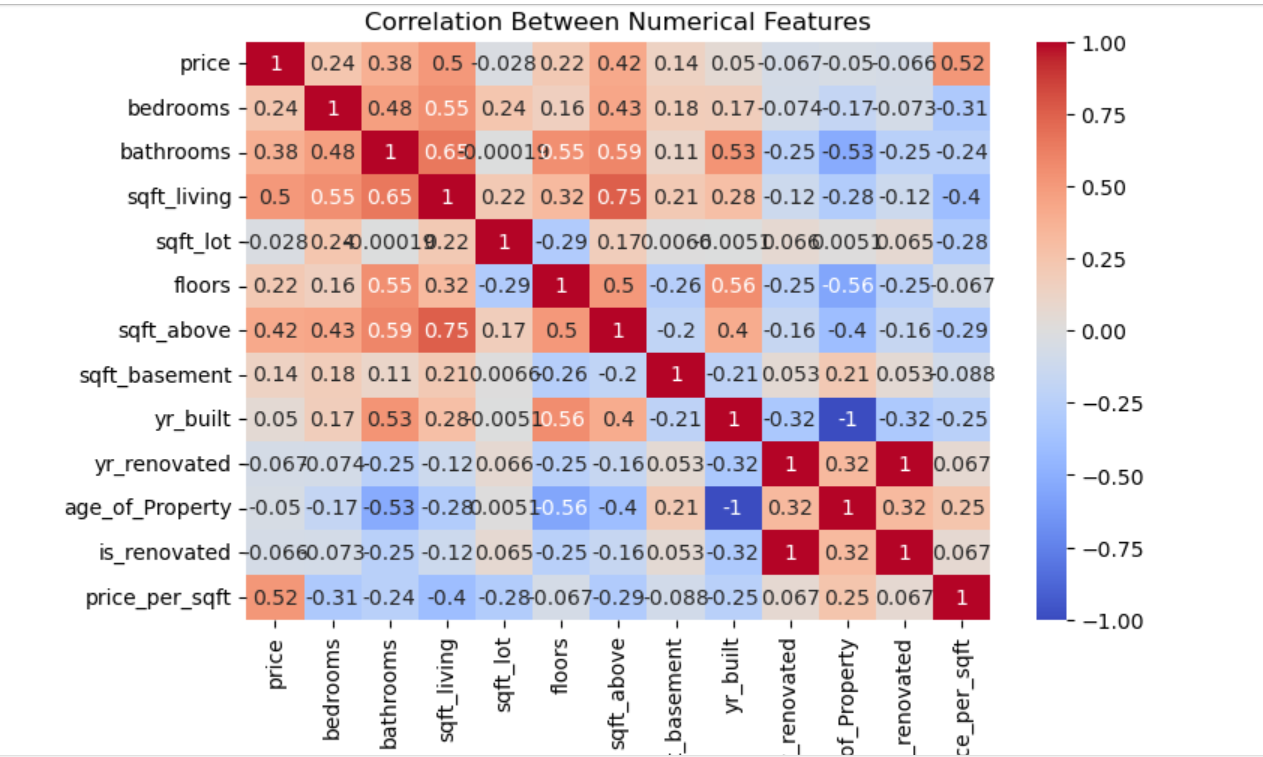
****

****



**10.3 MULTIVARIATE ANALYSIS:**

Multivariate analysis explored how multiple variables interact simultaneously and affect housing prices:



**11. OVERALL INSIGHTS FROM ANALYSIS:**

* **Most houses are in the low to mid price range.**  
  The price distribution shows a right-skew, with most homes priced below average and a few luxury homes skewing the upper end.
* **Luxury homes inflate the average price.**  
  High-value outliers significantly raise the mean; the median gives a more realistic central value.
* **Most properties have 3 to 4 bedrooms.**  
  Countplots confirm 3-bedroom homes are the most common, followed by 4-bedroom homes.
* **Bathrooms are typically 2 or 3 in most homes.**  
  Countplots show most homes feature 2 to 2.5 bathrooms.
* **Homes generally range from 1500 to 2500 sqft.**  
  The living area histogram reveals this is the most frequent size bracket.
* **Homes are mostly 15–50 years old.**  
  Distribution of 2014 - yr\_built shows many homes were built between the 1960s and 1990s.
* **Average condition is most common.**  
  Condition rating counts show most homes are marked as "average" (condition = 3).
* **Majority have no special view.**  
  View scores are mostly 0, indicating no premium view in most listings.
* **Larger homes strongly correlate with higher prices.**  
  A clear upward trend is seen in the scatterplot of sqft\_living vs. price.
* **Better condition leads to higher prices.**  
  Both ANOVA and boxplots show price increases with improved condition ratings.
* **Price varies within the same condition class.**  
  Even within the same condition rating, price variability is wide due to other factors like location or square footage.
* **Bathroom count trends with bedroom count.**  
  For 3-bedroom homes, 1 to 2.5 baths is common; 4-bedroom homes usually have 2 to 3 baths.
* **Renovated homes have higher median prices.**  
  T-test between is\_renovated = 1 vs. 0 confirms statistically higher median prices for renovated homes.

**12. CONCLUSION:**

This housing price analysis showed that property value is heavily impacted by factors like location, size, condition, and whether renovations were made. Through detailed data exploration and statistical tests, we found clear patterns and connections between these features and pricing trends. Renovated homes, better views, and higher condition ratings were consistently associated with higher prices. The results give valuable insights for homebuyers, sellers, and investors who want to make informed decisions. This data-driven approach improves transparency in real estate valuation. Future work may involve creating predictive machine learning models to estimate prices and find more hidden patterns across regions.