EXPLORATORY DATA ANALYSIS USING PYTHON

PROJECT TITLE: HOUSING DATASET ANALYSIS NAME: POOJA K

BATCH: DA & DS MAY -2025 BATCH ONLINE / OFFLINE: OFFLINE

# CONTENTS

|  |  |  |
| --- | --- | --- |
| SNO | CONTENT | PAGE NO |
| 1. | INTRODUCTION | 3 |
| 2. | AIM OF THE PROJECT | 4 |
| 3. | PROBLEM STATEMENT | 4 |
| 4. | PROJECT WORKFLOW   * 1. DATA COLLECTION   2. DATA CLEANING   3. EDA (EXPLORATORY DATA ANALYSIS)   4. FEATURE ENGINEERING   5. VISUALIZATION   6. OBTAINED DERIVED METRICS   7. STATISTICAL ANALYSIS | 4 |
| 5. | OVERALL INSIGHT | 19 |
| 6. | CONCLUSION | 19 |

1. INTRODUCTION:

This housing dataset provides detailed information on residential properties, including price, size, layout, and location. The goal of this project is to thoroughly analyze the data and extract meaningful insights to support informed decision- making in the real estate industry.

The dataset consists of 4,600 rows and 18 columns, covering various property features such as price, size, location, and condition.

* **Date**: Record the date of the property.
* **Price**: Property sale price.
* **Bedrooms/Bathrooms**: Number of bedrooms and bathrooms.
* **Sqft\_living**: Total living space area.
* **Sqft\_lot**: Total lot size.
* **Floors**: Number of floors.
* **Waterfront**: 1 if waterfront, else 0.
* **View**: View quality (0–4).
* **Condition**: Overall condition (1–5).
* **Sqft\_above**: Area above ground level.
* **Sqft\_basement**: Basement area.
* **Yr\_built**: Year built.
* **Yr\_renovated**: Last renovation year.
* **Street**: Property street address.
* **City**: Property city.
* **Statezip**: State and ZIP code.
* **Country**: Property country.

### AIM:

This analysis aims to explore residential housing data to uncover key factors influencing property prices. By examining variables like location, size, condition, and other factors, the project identifies patterns and trends that affect real estate value. It also involves data cleaning, feature engineering, and visualization to enhance insights. The final goal is to support informed decision-making for investors, developers, and buyers through data-driven analysis and geospatial mapping.

### PROBLEM STATEMENT :

In this project, our goal is to determine the appropriate price for residential properties using the housing dataset. By analyzing key factors such as location, renovations, property size, and layout, we will examine how these attributes influence pricing. The dataset provides detailed information, including price, number of bedrooms and bathrooms, square footage, and location specifics. The goal is to conduct a thorough analysis to gain meaningful insights that can help stakeholders in the real estate industry make informed decisions.

### PROJECT WORFLOW :

This project follows a systematic workflow to analyze residential housing data, aiming to extract meaningful statistics and actionable insights to support real estate decision-making.

## Data Collection

* Load the housing dataset (CSV or database).
* Understand the structure and types of features.

## Data Preprocessing

* Handle missing values and duplicates.
* Convert data types (e.g., date columns).
* Encode categorical variables if needed (e.g., city, waterfront).

## Exploratory Data Analysis (EDA)

* Summary statistics of key features.
* Visualizations (histograms, scatter plots, box plots).
* Correlation matrix to identify feature relationships.

## Feature Engineering

* Create new features (e.g., age of house, total area).
* Normalize or scale features if required.

## Data Visualization

* Use charts to display patterns and trends.
* Geospatial maps for location-based price analysis.

## Statistics

* Calculate descriptive statistics (mean price, average bedrooms, etc.).
* Analyze distributions of key variables.

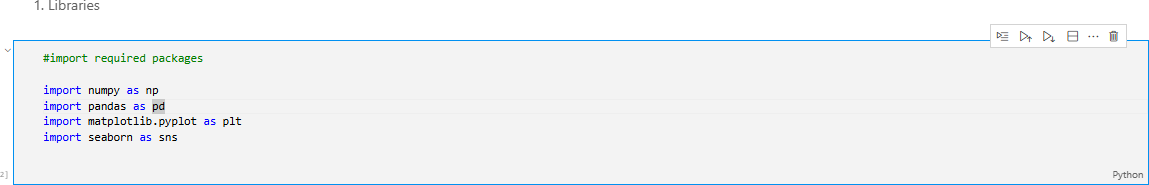
## Insights

* Identify correlations between features and price.
* Analyze geospatial trends and property conditions.
* Detect outliers and their impact on price distribution.
* Summarize market trends and key takeaways.

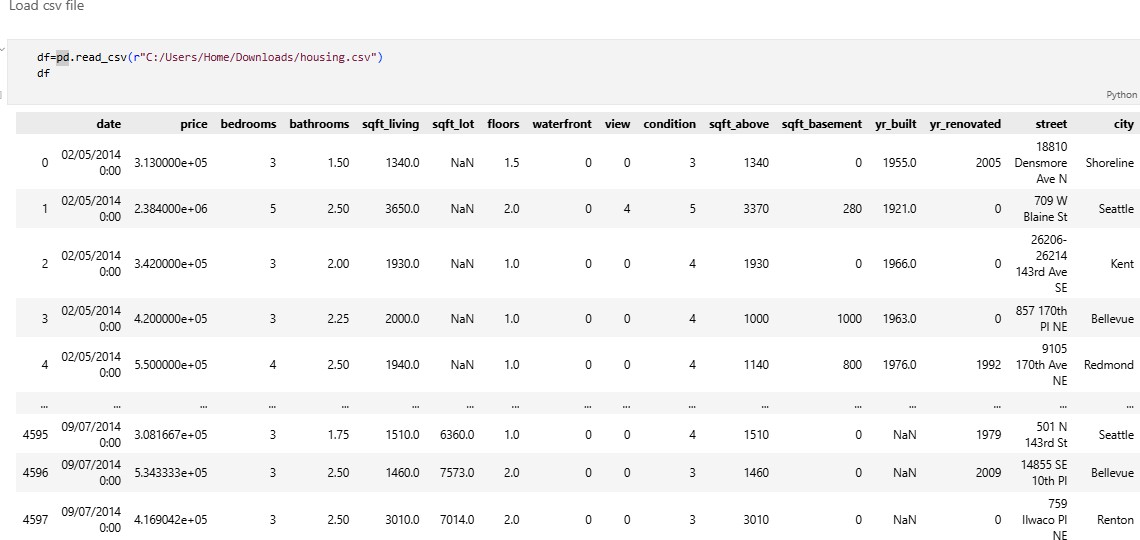
### DATA COLLECTION:

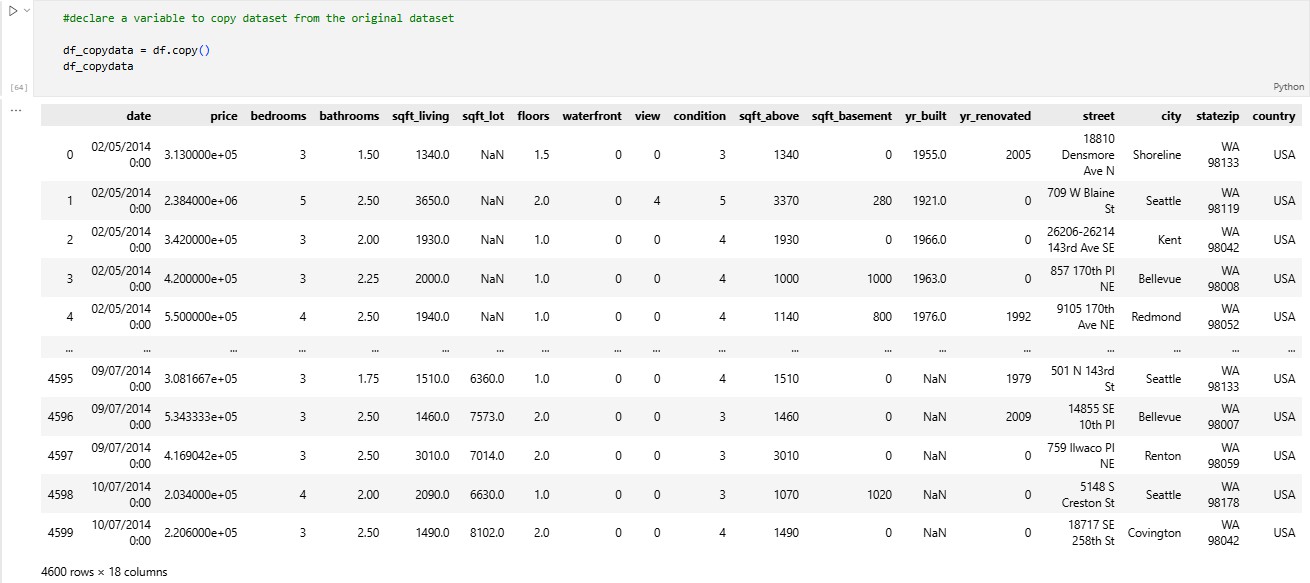
The first step involves gathering the housing dataset, which contains detailed information on various property attributes necessary for analysis. This step focuses on acquiring the raw data that forms the foundation of the project. It ensures that all relevant property details—such as price, size, location, and condition—are collected and ready for further processing and analysis.

In this project, I have stored my dataset in the format of a CSV file in the D/ drive. In my project, I have imported libraries such as Pandas, Numpy, Matplotlib, and Seaborn.



After importing the necessary libraries, declaring a variable name as df, and making the CSV file to read as pd.read\_csv, and taking a copy of the original data, which is already declared using df, and storing it as df\_copydata



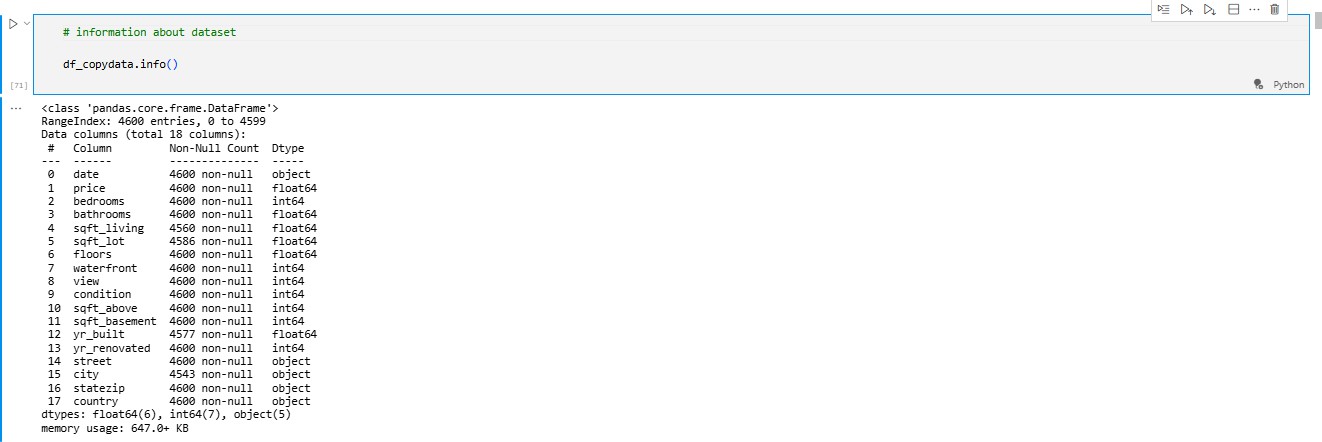


After this, usually we get the first five rows and the last five rows, shape, length, sample, a range of rows from the dataset, information of the data, describe the mathematical values in the dataset, and its datatype.

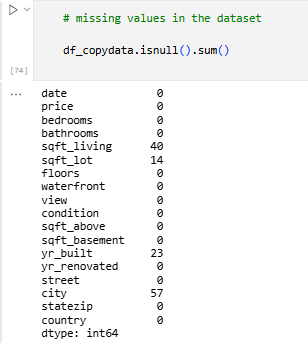
### DATA CLEANING

Data cleaning is essential to prepare the dataset for analysis by addressing missing values, removing duplicates, and correcting inconsistencies. This step ensures the quality and accuracy of the data by fixing errors and handling incomplete information. Clean data helps produce reliable analysis results and prevents misleading insights.

First, we need to check the data type of the dataset by using df\_copydata.info(), which checks the datatypes, non-null count, and columns.



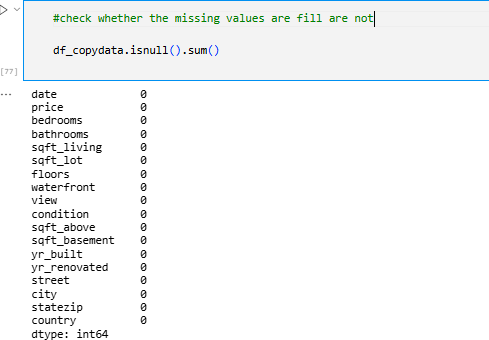
Next, we used to identify the null values present in the columns using df\_copydata.isnull().sum()



Using fillna, it fills the missing values in the dataset. Since the dataset has missing values in both

**numerical** and **categorical** columns, the typical strategy is:

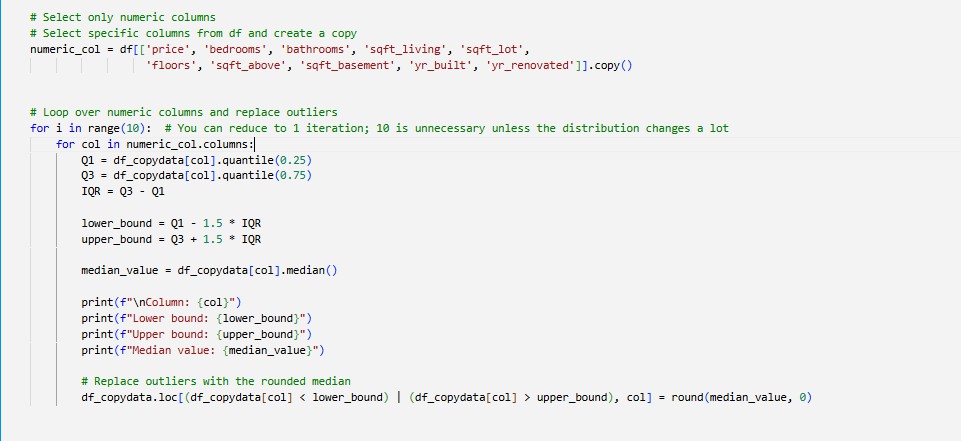
* + - **Numerical columns** → fill with **median**
    - **Categorical columns** → fill with **mode**
    - sqft\_living, sqft\_lot, and yr\_built are numerical — best filled with the **median** to avoid influence from extreme values.
    - City is a categorical variable — best filled with the **mode** to preserve consistency.



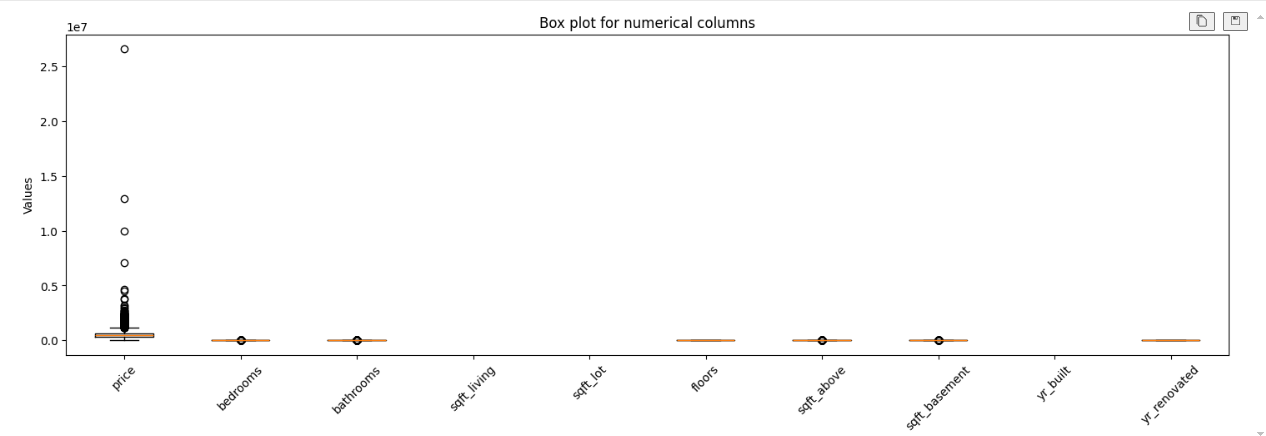
Outliers in the dataset :

Outliers are data points that differ significantly from other observations in the dataset. In the housing data, outliers can appear in features like price or square footage, often due to rare luxury properties or data entry errors. Identifying and handling outliers is important, as they can skew statistical results and impact model accuracy. Common methods to detect outliers include boxplots and the Interquartile Range (IQR) technique. Depending on the context, outliers can be removed, capped, or transformed to improve data quality.

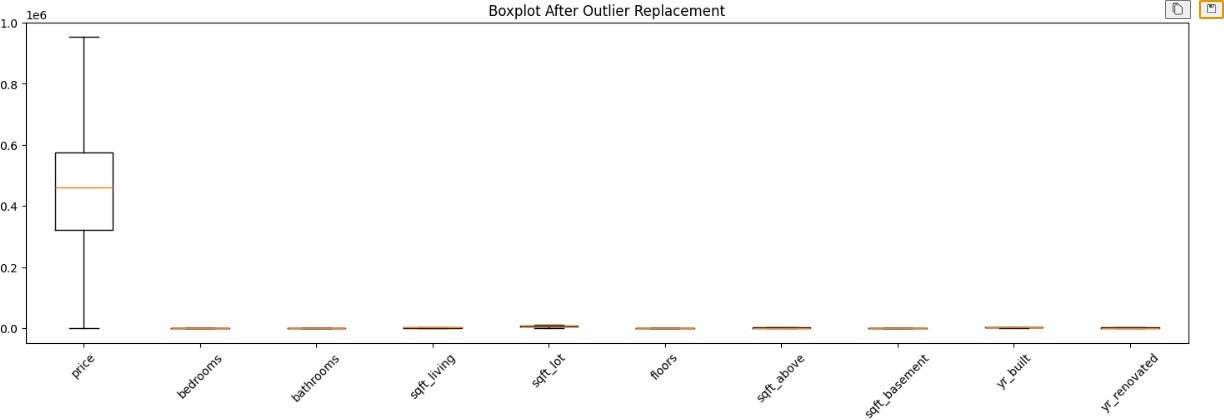
It calculates the median value for each specified column and its upper and lower bounds.



Before outliers



After removing the outliers



* 1. EDA (Exploratory Data Analysis)

EDA is the process of analyzing datasets to summarize their main characteristics using visual and statistical methods. It helps in understanding the data, identifying patterns, and preparing for modeling.

**Types of EDA**

## Univariate Analysis

* + Examines a single variable.
  + Example: Histogram of price, boxplot of sqft\_living.

## Bivariate Analysis

* + Examines relationships between two variables.
  + Example: Scatter plot of price vs sqft\_living, boxplot of price by

condition.

## Multivariate Analysis

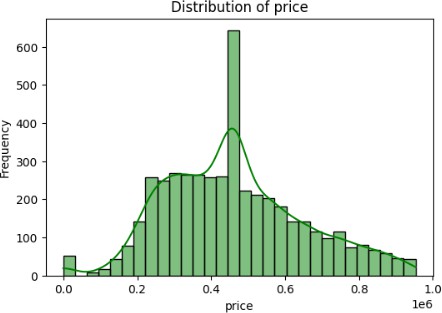
* + Analyzes more than two variables together.
  + Example: Pair plots, grouped bar charts.

Continous variable: price,sqft\_living,sqft\_lot,floors,sqft\_above,sqft\_basement,yr\_built,yr\_renovated

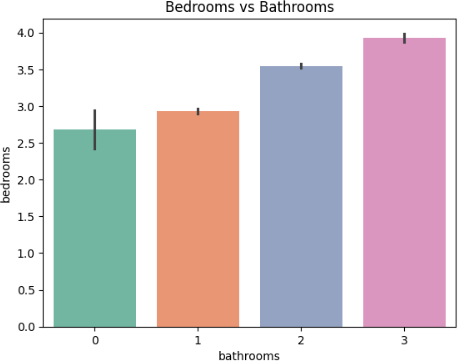
Categorical variable:

Bedrooms, Bathrooms, Floors, Waterfront, Views, Condition, City, State, Zip, Country, Year, Month, Day

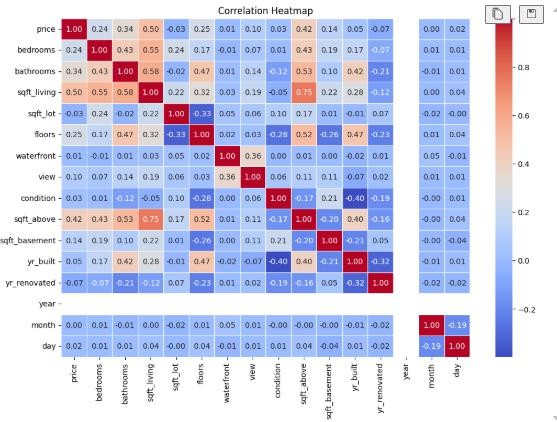
Univariate Analysis:



Bivariate analysis:



Multivariate Analysis



* 1. Feature Engineering

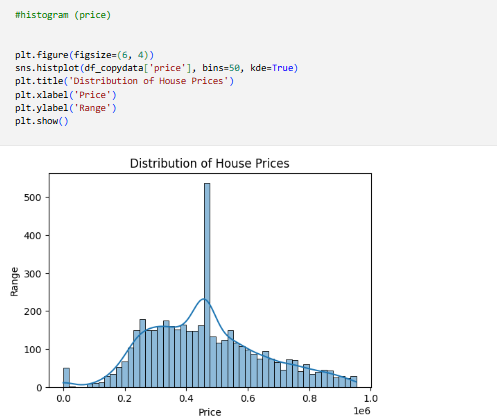
Feature engineering involves creating new features to improve model performance and gain deeper insights. In this project, new variables were derived to enhance the prediction and understanding of housing prices, such as:

* + - **house\_age**: Number of years since the house was built.
    - **renovation\_age**: Years since the last renovation; 0 if never renovated.
    - **price\_per\_sqft**: Price efficiency — how much each square foot costs.
    - **multi\_floor**: Binary indicator (1 or 0) if the house has more than one floor.
    - **basement\_ratio**: Proportion of living space in the basement

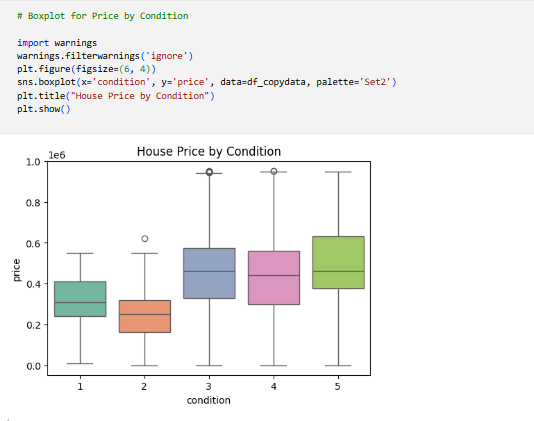


### VISUALIZATION

To better understand the relationships and trends within the housing dataset, several visualizations were created. These plots helped reveal hidden patterns, outliers, and correlations among various features. It shows the distribution of house prices using a histogram.



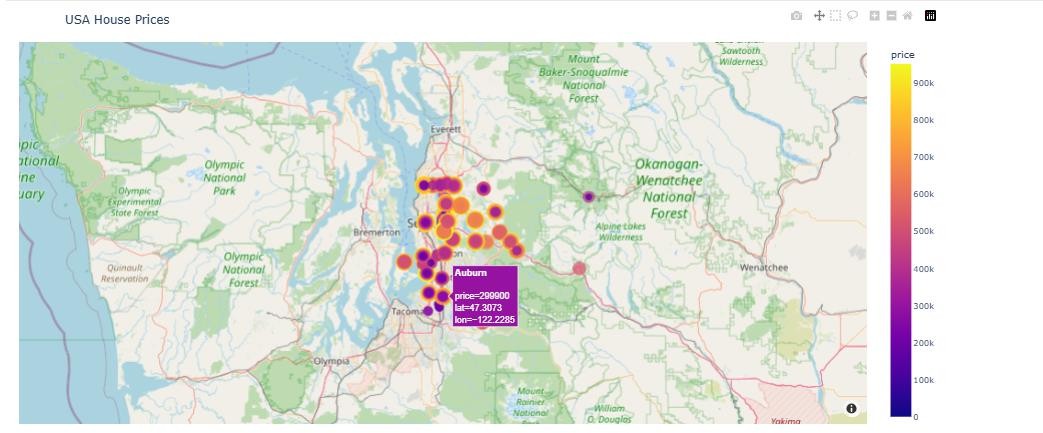
Boxplot for price by condition



A scatter plot is used to visualize the land area vs price.



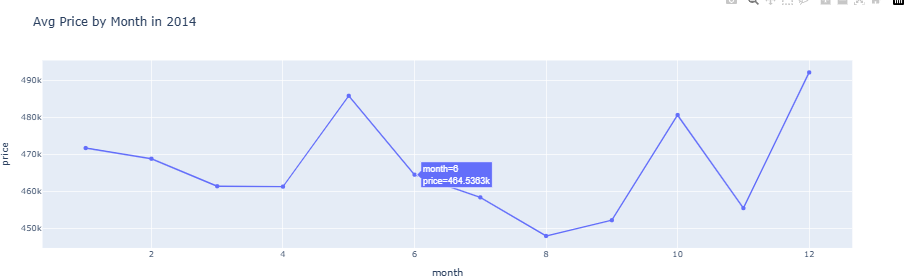
Map plot is used to plot the exact country(USA) based on its city and state, zip code, the price of the house in the dataset, and the latitude and longitude of the cities in the dataset.



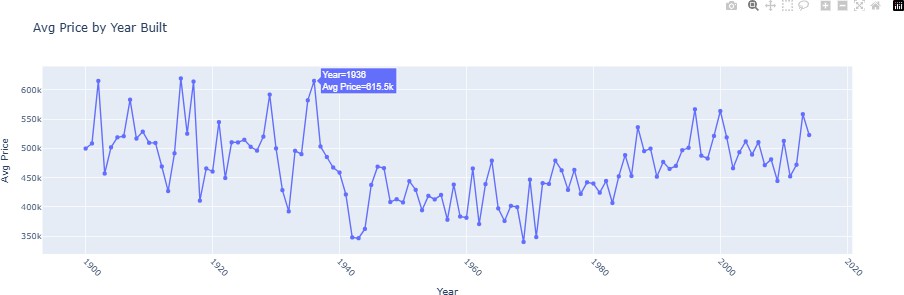
A bar plot is used for the waterfront based on the price in the dataset.



Average price in the months



Average Price by the year built in the dataset



### OBTAINED DERIVED METRICS

The **age of the property** was calculated by subtracting the year\_built from the current year to show how old each property is, allowing better comparison than using raw build years. The **renovation age** was derived by subtracting yr\_renovated from the current year when applicable, indicating how recently the property was updated. Additionally, **price per square foot** was computed by dividing the price by the living area (price / sqft\_living), providing a standardized measure for comparing properties of different sizes.

### FILTERING DATA FOR ANALYSIS

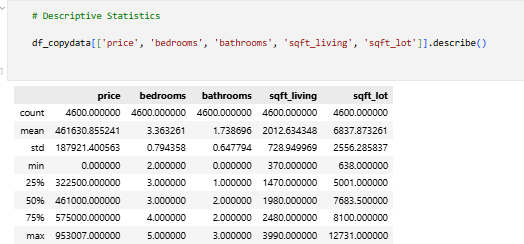
Data filtering involves selecting relevant subsets of the dataset based on specific criteria to improve the quality and focus of the analysis. This process removes irrelevant, duplicate, or erroneous records and narrows down the data to the most meaningful entries. Filtering can be done by applying conditions on columns such as date ranges, location, price limits, or property features, ensuring that the analysis reflects accurate and targeted insights.



### STATISTICAL ANALYSIS

Statistical analysis involves applying mathematical techniques to summarize, describe, and interpret data patterns. It helps identify relationships, trends, and differences within the dataset. Common methods include calculating measures of central tendency (mean, median), dispersion (variance, standard deviation), and performing hypothesis tests to validate assumptions. This analysis supports data- driven decisions by providing quantitative evidence and insights.

Descriptive statistics are used to describe specified data in the dataset.



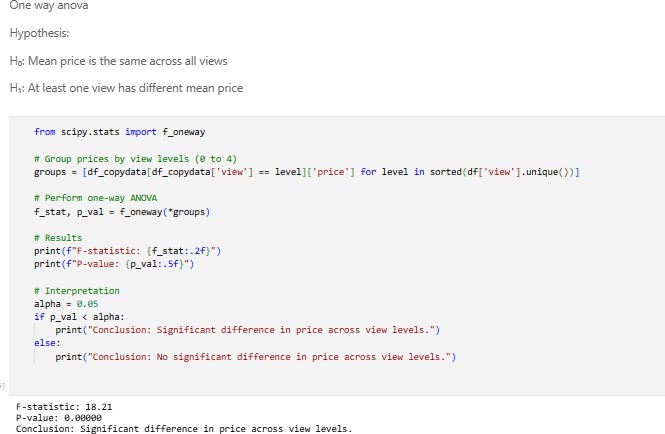
Two-way t-test:

A two-sample t-test was conducted to compare the mean house prices between different view categories. This test helps determine whether the differences in average prices are statistically significant based on the property’s view rating. By evaluating the p-value, we assess if the variation in price is likely due to the view factor or occurred by chance, thus revealing whether view influences pricing trends.



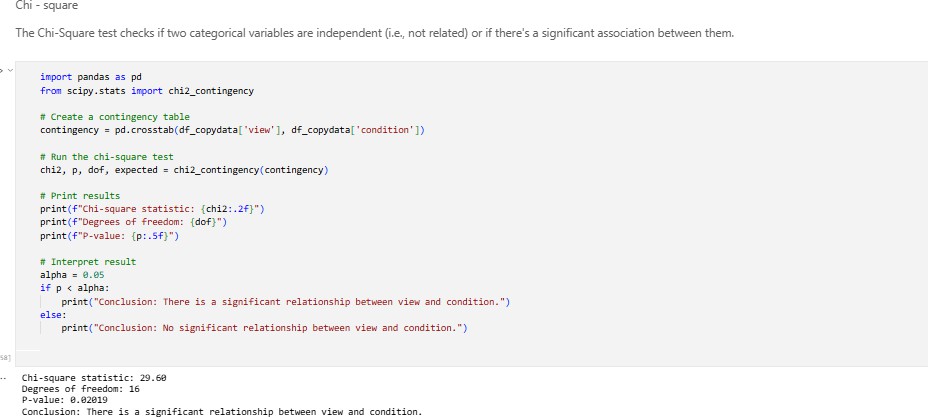
# One-way ANOVA:

A one-way ANOVA test was performed to examine whether there are significant differences in average house prices across multiple view categories. This statistical method helps identify if at least one view group has a mean price significantly different from the others. A low p-value indicates that the view rating has a meaningful impact on property pricing.



Chi-square:

A chi-square test of independence was conducted to examine the relationship between property view and condition. This test assesses whether these two categorical variables are statistically dependent or independent. A significant result (low p-value) indicates that the distribution of property condition varies meaningfully across different view categories, suggesting a potential association between the two.



1. OVERALL INSIGHT:

The housing market data reveals that **property price is significantly influenced by factors such as view quality, condition, square footage, and location**. Homes with better views and in good condition tend to be priced higher. Age and renovation history also impact value— newer or recently renovated homes command more. Statistical tests confirm **significant relationships** between view, price, and condition, indicating that aesthetic and physical quality drive market demand. The addition of derived metrics like **price per sqft** and **property age** enables deeper, more actionable insights for buyers, sellers, and investors.

### CONCLUSION

This housing data analysis demonstrates that **key features such as view quality, condition, square footage, and renovation status have a substantial impact on property prices**. Statistical analyses, including t-tests, ANOVA, and chi-square tests, validate the **significant associations** between these variables. Derived metrics like **house age**, **renovation age**, and **price per square foot** provided enhanced clarity and depth to the insights.

Overall, this study helps understand how both **physical attributes and visual appeal** contribute to housing market trends, aiding **better decision-making** for stakeholders such as buyers, sellers, and real estate investors.