**Exploratory Data Analysis using Python**

**HOUSING DATASET**

**JEEVANANTHAM R**

**DA&DS-MAY’25 REINFORCEMENT PROJECT-1**

**23-07-2025 (OFFLINE)**

**Table of Content :**

**1. Introduction**

**2. Aim**

**3. Problem Statement**

**4. Project Workflow**

**5. Data Understanding**

**6. Data Cleaning**

**7. Obtaining Derived Metrics**

**8. Filtering Data for Analysis**

**9. Statistical Analysis**

**10. Exploratory Data Analysis (EDA)**

**11. Uni-variate Analysis**

**12. Bi-variate Analysis**

**13. Multi-variate Analysis**

**14. Data Visualization**

**15. Overall Insights from Analysis**

**16. Conclusion**

**1.Introduction :**

The House marketing is so dynamic, so we need look deep into the data. We need to find what factors affecting the price of the house, it is essential for anyone looking to buy, sell or invest in the property.

We have to look deep into the housing data to discover the price trends over time and features of the house.

In this we have a columns like;

'bathrooms', 'bedrooms', 'city', 'condition','country', 'date', 'floors', 'price', 'sqft\_above', 'sqft\_basement', 'sqft\_living', 'sqft\_lot', 'state\_zip', 'street', 'view', 'waterfront', 'yr\_built', 'yr\_renovated'.

With these columns we can overview of the data set housing.

Here we use python as a programming language.

We use libraries like pandas, numpy, matplotlib, datetime and seaborn.

**2.Aim:**

The main goal of the project is to build the simple code for easy understanding and the factors affecting the price of the house.

To achieve our goal we have follow this steps:

1. Data Understanding

2.Data Cleaning

3.Exploratory Data Analysis (EDA)

4.Data Visualization

5.Feature Engineering

6.Analysis and Interpretation

By data analyzing the factors like property size, location, bedrooms and bathrooms, views and condition of the house, whether the house is renovated or not renovated.

By this we can analyze the value of the house according to the records.

**3.PROBLEM STATEMENT:**

In this data we have find the accurate price for the data.

It helps the person to make a correct decision.

The housing data provides a comprehensive factor that helps in finding the price, location, year built, renovations, square area of the house and other relevant factors.

**4.PROJECT OVERFLOW:**

It shows the factors affecting and how it varies the price of the house.

**DATA COLLECTION:**

First we have to load the data set. The data is in CSV format in my personal drive.

Then we have to assign the variable for our data.

Then we have to make an copy of our data.

If we not take a copy the original file may change or corrupt, so we always have to take the copy of our original data.



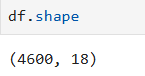


**5.DATA UNDERSTANDING:**

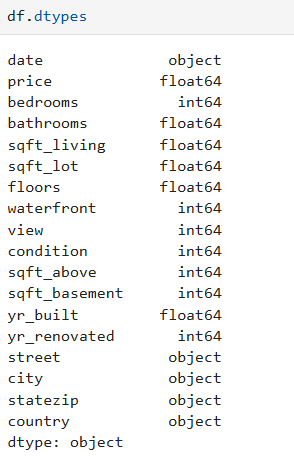
Then we have to understand the basic information from our data set.

Like shape, data types, info of the data set.

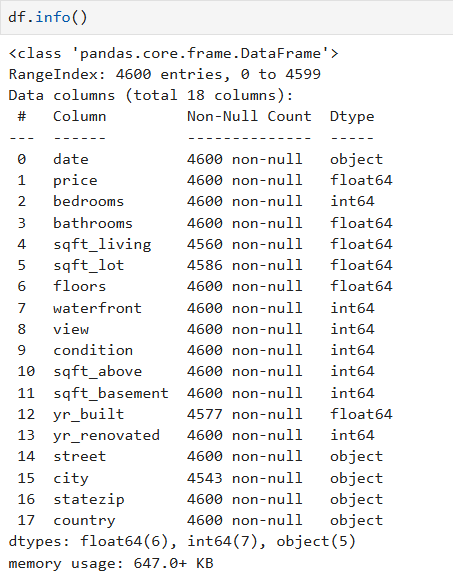
Shape of the data set:



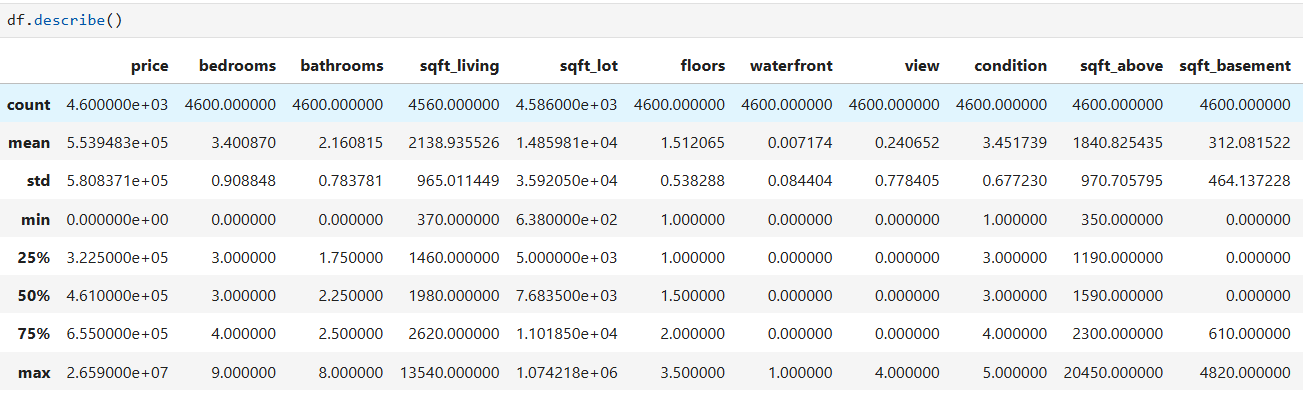
Data types of the data set:



Info of the data set:



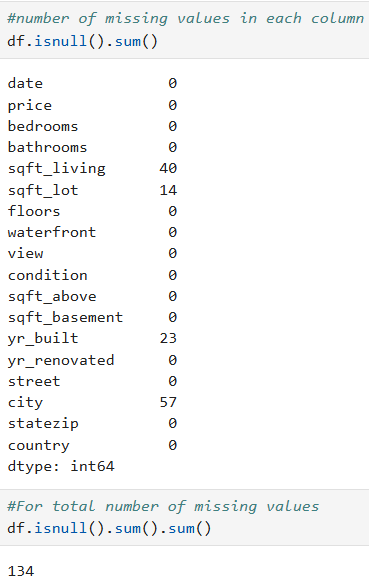
Describe of the data set:



**6.DATA CLEANING:**

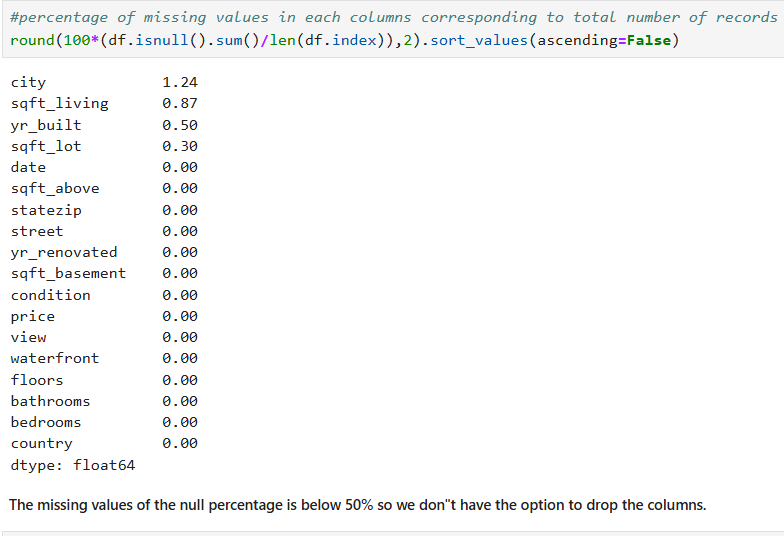
Then we have find the null values in our data set, check outliers and fill the missing value.

checking null values:

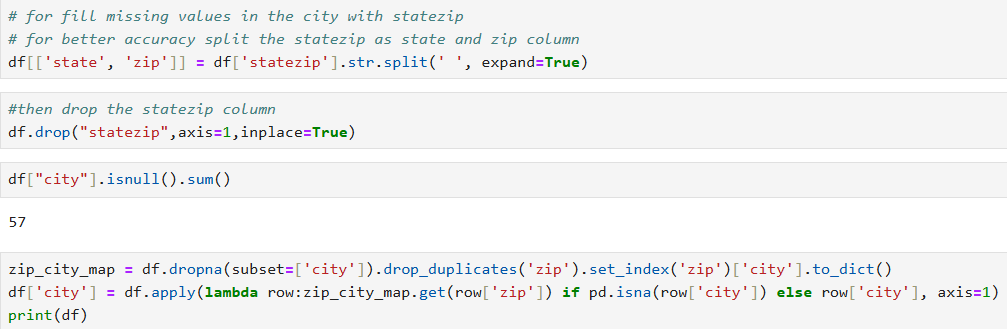


We have null values in our dataset.

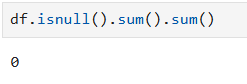
Let’s see how to fill the missing values.



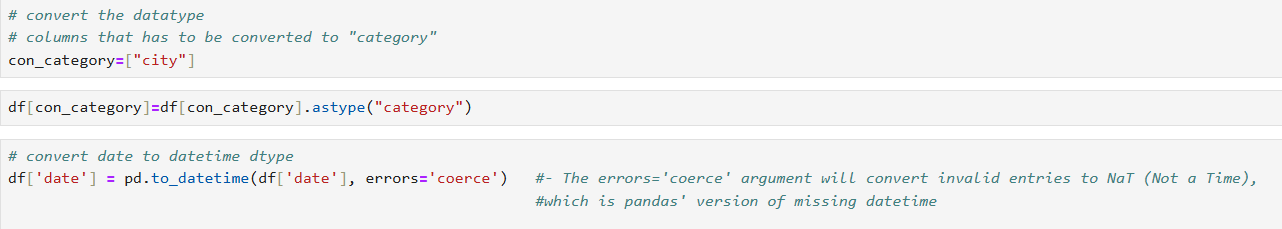




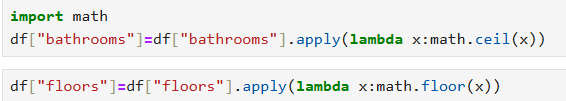
After cleaning:



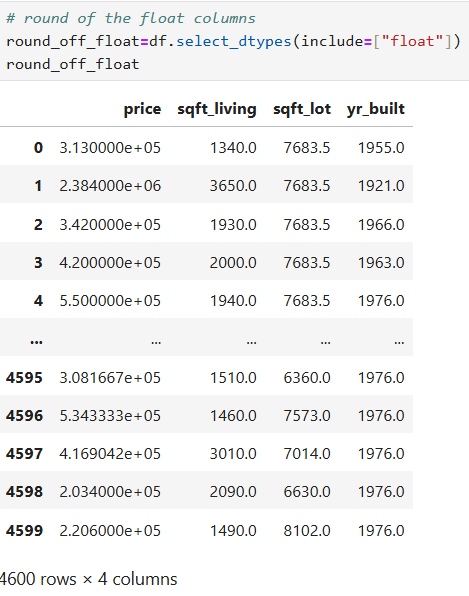
Then change the data types of the columns according to our analysis.

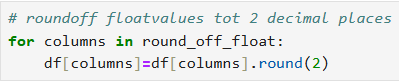


Then the bathrooms and floors are in decimal so we have fill it with ceil and floor function.

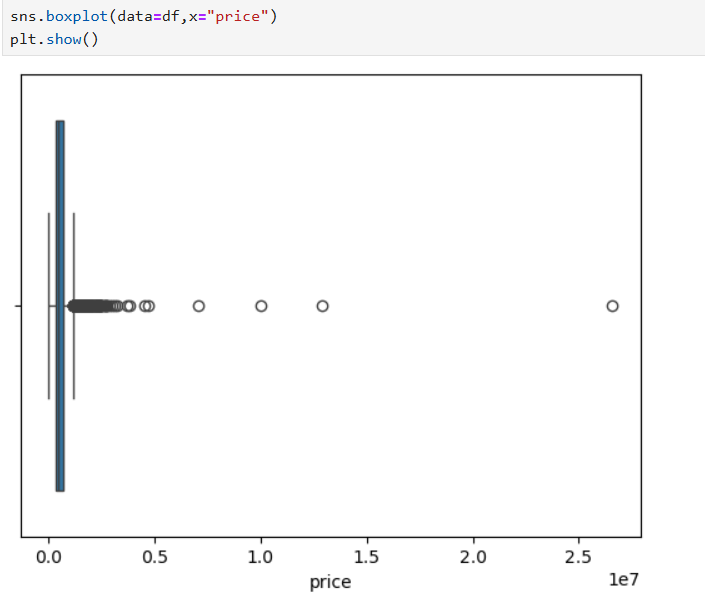


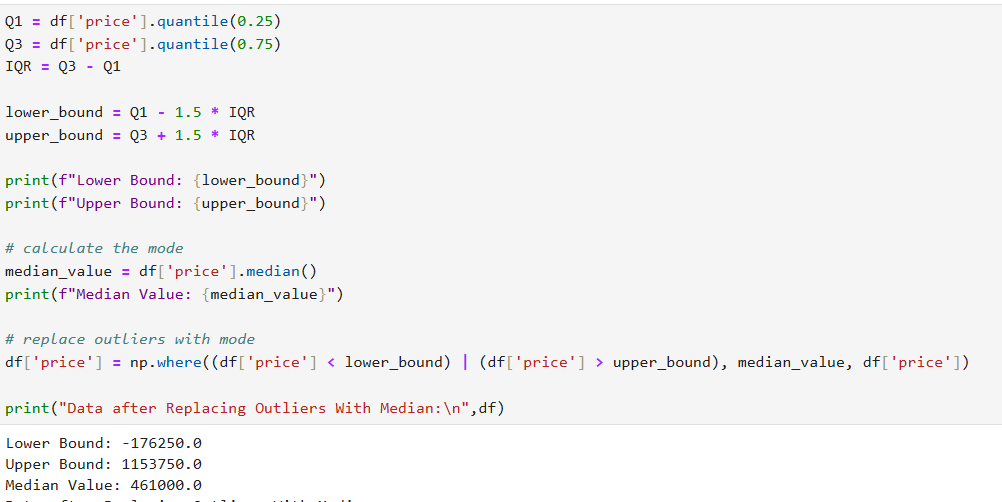
Then round-off the float columns.





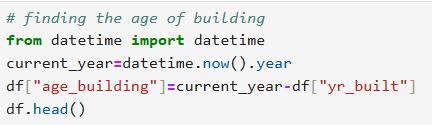
Checking outliers:

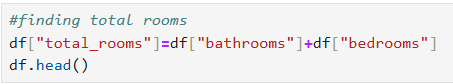




**7.OBTAINING DERIVED METRICS:**

We create new columns like Age of the building, Total rooms of the building, Total square area of the building.







By this we give more relevant information about the building. By this new columns we have make an better comparison with price.

**8.FILTERING DATA FOR ANALYSIS:**

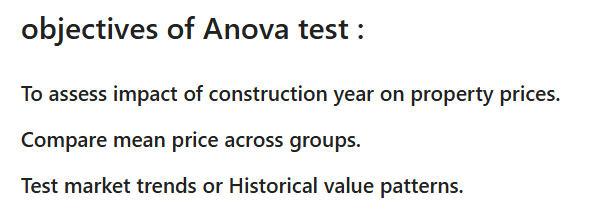
Filtering our records with extreme or unrealistic records. such as the bedrooms with 0 and square area more than 10000.And price range between city and rural areas. convert the data types when it is necessary.

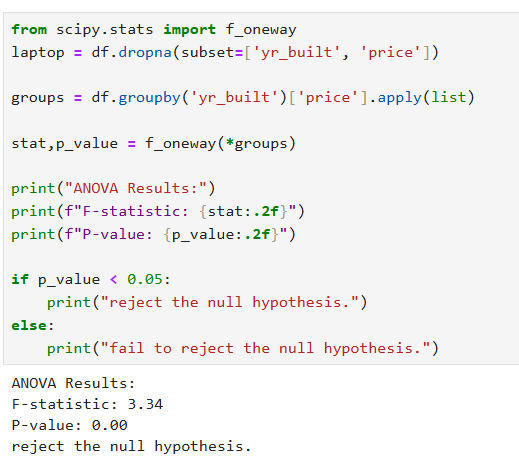
**9.STATISTICAL ANALYSIS:**

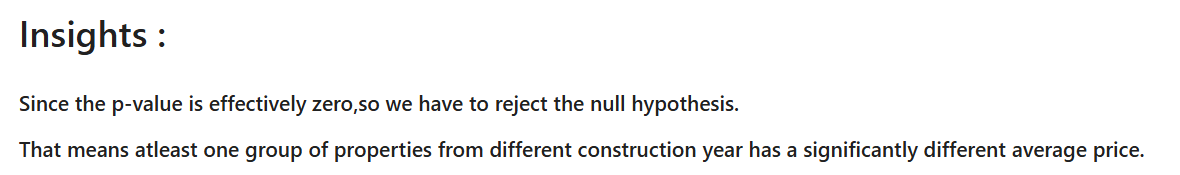
**Descriptive statistics:** Summarize the descriptive statistics(mean, median, mode)



**ANOVA TEST:**







**10.EXPLORATORY DATA ANALYSIS(EDA):**

Exploratory Data Analysis is the process of analyzing datasets and summarizing values. It is a crucial step in data modelling.

**11.UNIVARIATE ANALYSIS:**

It helps to find describe data and find patterns within a single feature.

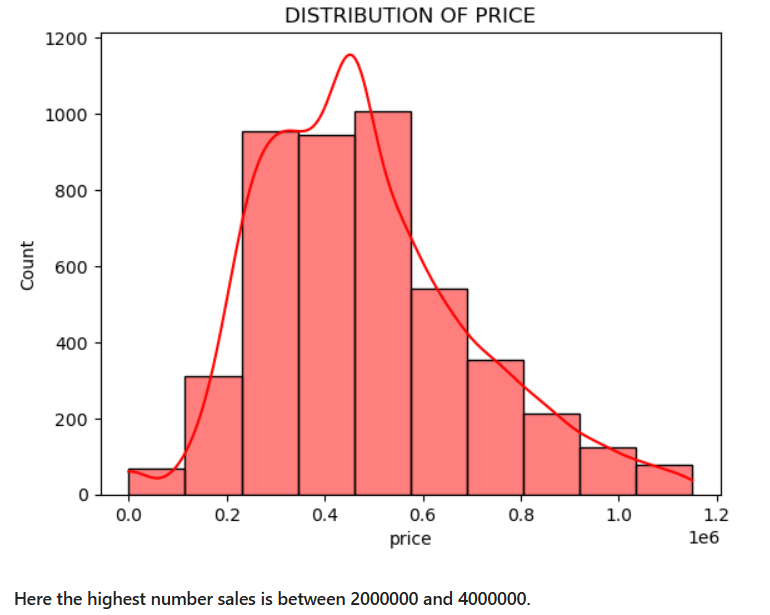
The term "uni" means "one," so this type of analysis doesn’t explore relationships between variables—it’s all about understanding the characteristics of a single data point.

****Common Techniques**:**

* **Summary statistics**: Mean, median, mode, range, variance, standard deviation.
* **Frequency distributions**: How often each value occurs.
* **Visualizations**: Histograms, box plots, pie charts, and bar graphs.

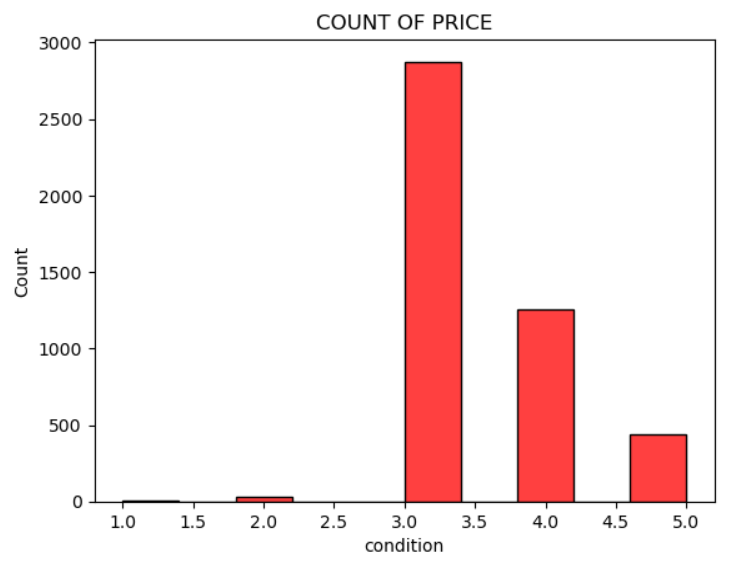
**Plots:**

Distribution of price:



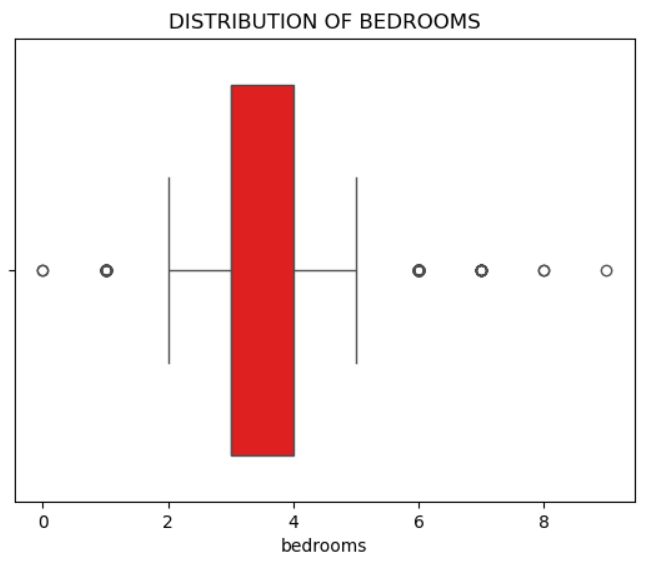
Here the highest number of sales is between 2000000 and 6000000.

Distribution of condition of the house:



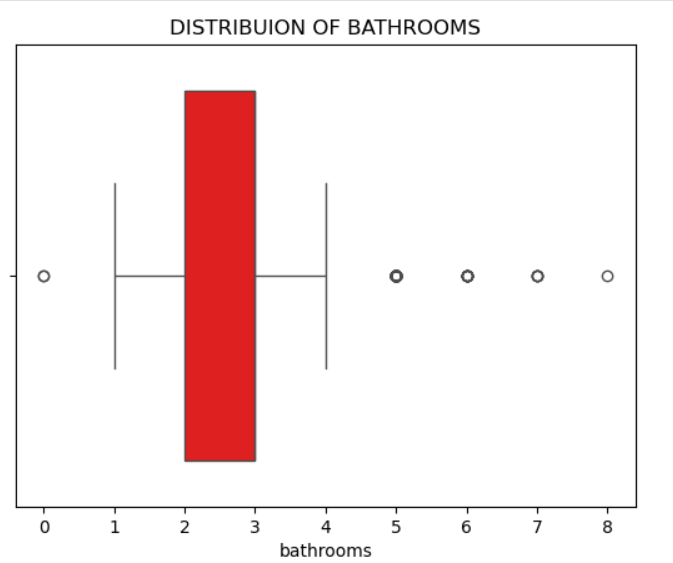
The 3 condition house has the highest number of sales as compared to others.

Distribution of bedrooms:



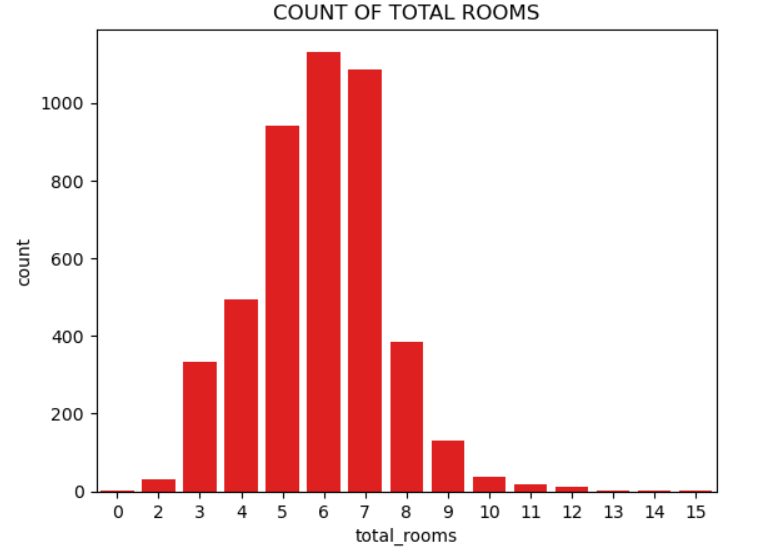
Here the house with bedrooms 3 and 4 has the highest number of sales.

Distribution of bathrooms:



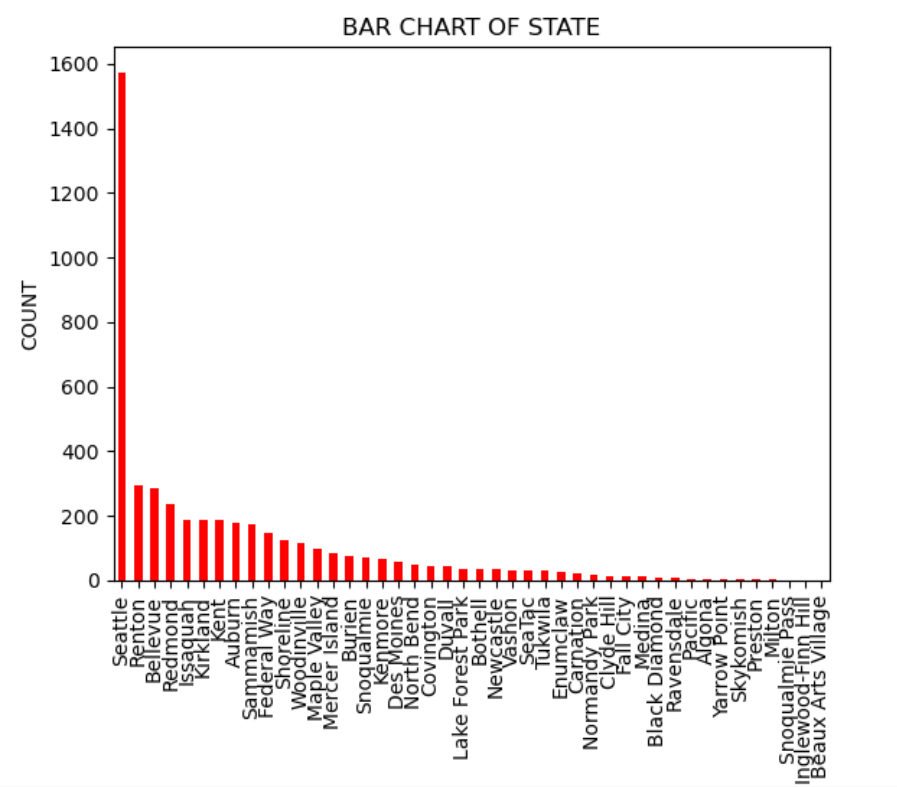
Here the house with bathrooms 2 and 3 has the highest number of sales.

Distribution of total rooms:



Here the house with total rooms 6 has the highest number of sales.

Total number of sales by city:



Here the Seattle city has the highest number of sales.

The Beaux Arts Village has the lowest number of sales.

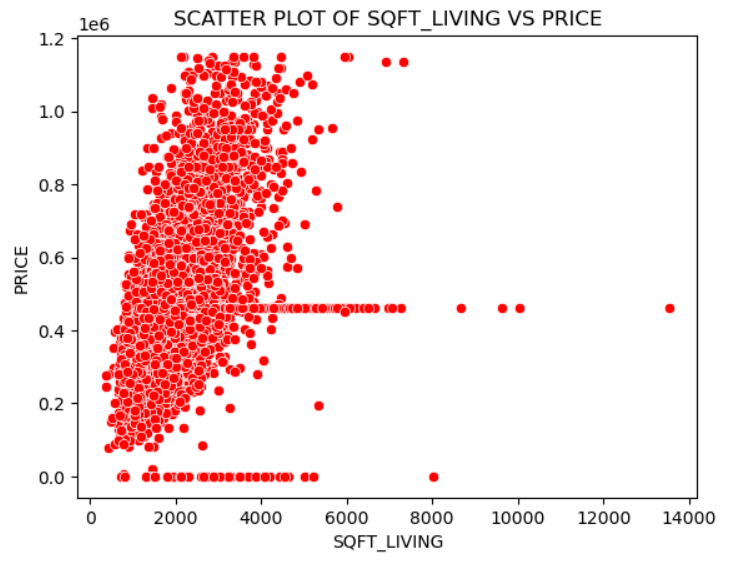
**12.BIVARIATE ANALYSIS:**

It is a statistical method that explores the relationship between two variables-often labeled as X and Y.

**Common Techniques**:

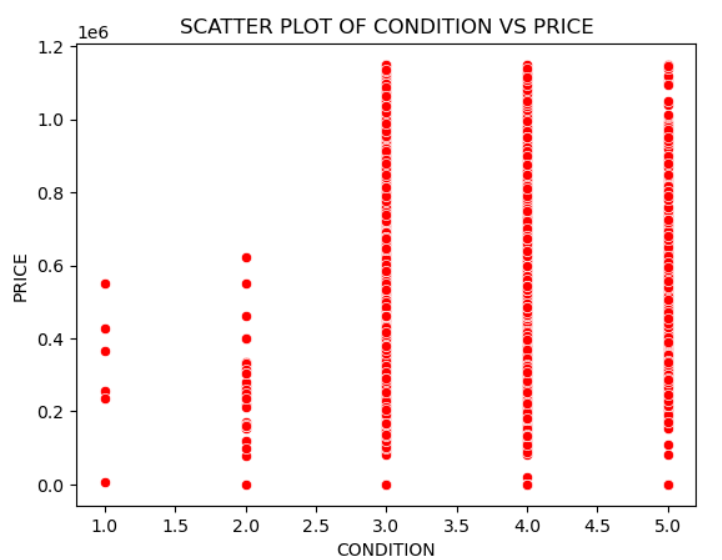
* **Scatter plots**: Visualize the relationship between variables.
* **Correlation analysis**: Measures the strength and direction of association.
* **Regression analysis**: Predicts one variable based on the other.

Scatter plot for sqft\_living vs price:



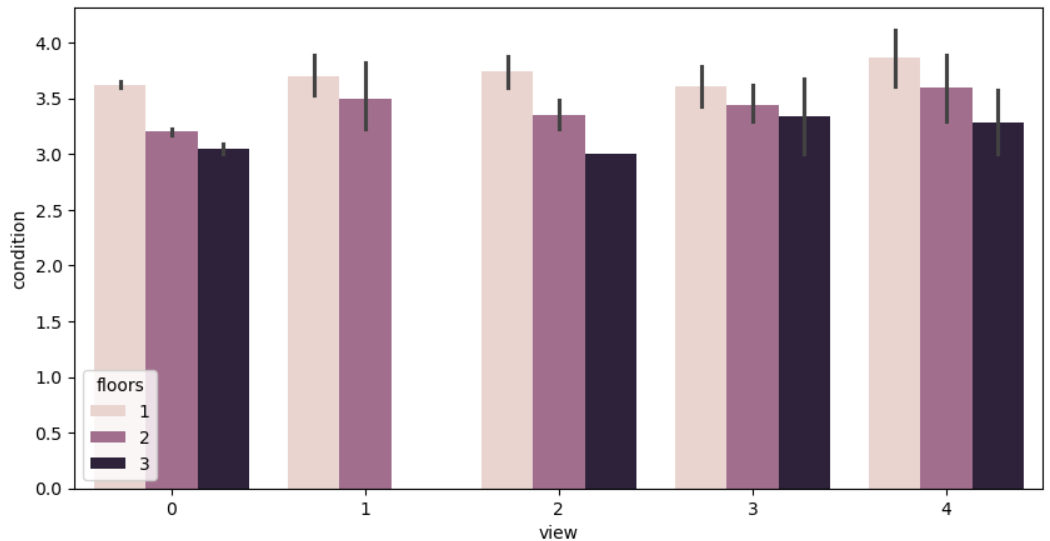
#### Here the sqft\_living between 1000 to 4000 has the highest number of sales with the highest price rate.

Scatter plot for condition vs price:

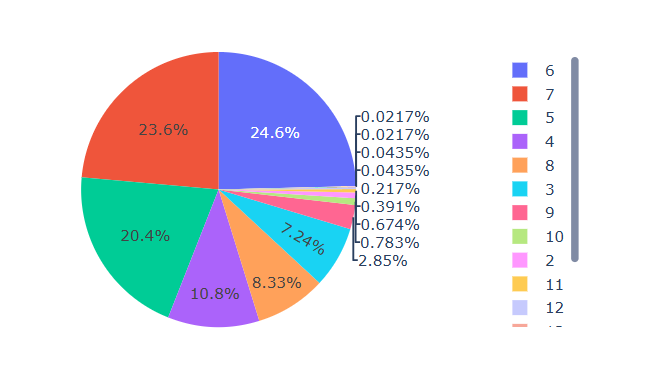


#### Here the condition with 3,4 and 5 has the highest price rate.

Boxplots for condition, views, floors:

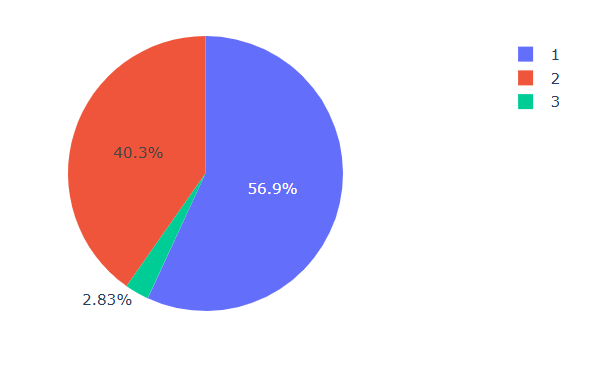


Pie chart:



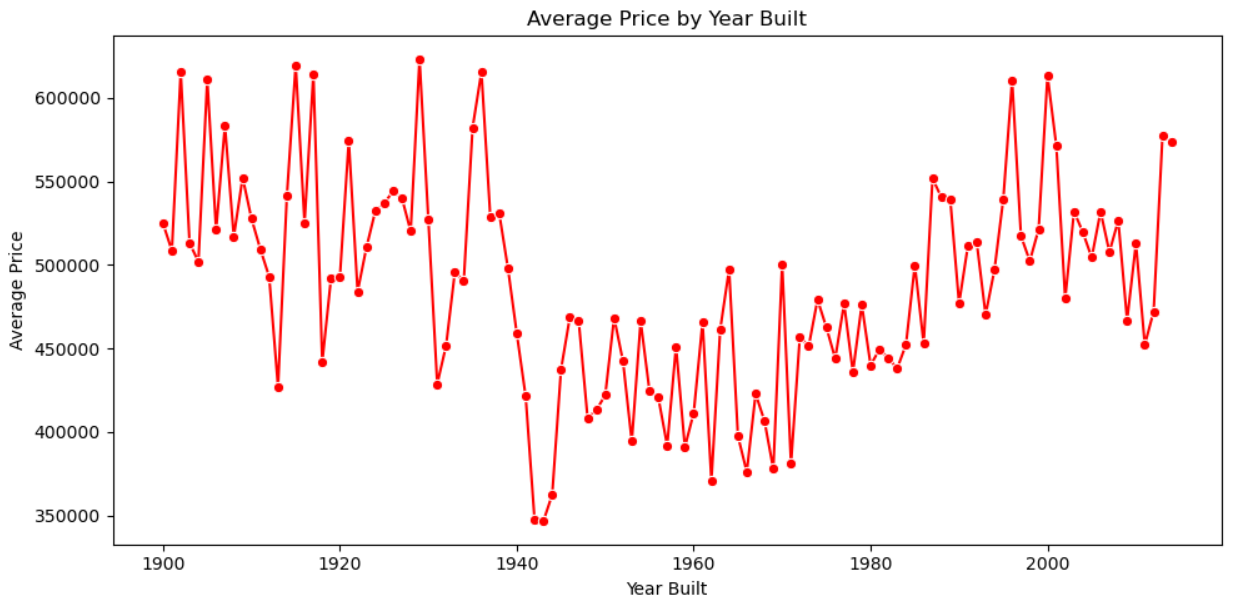
#### Here the total rooms with 6 has the highest number of sales.

#### And the total rooms with 2,10,11,12,13,14,15 has the lowest number of sales.

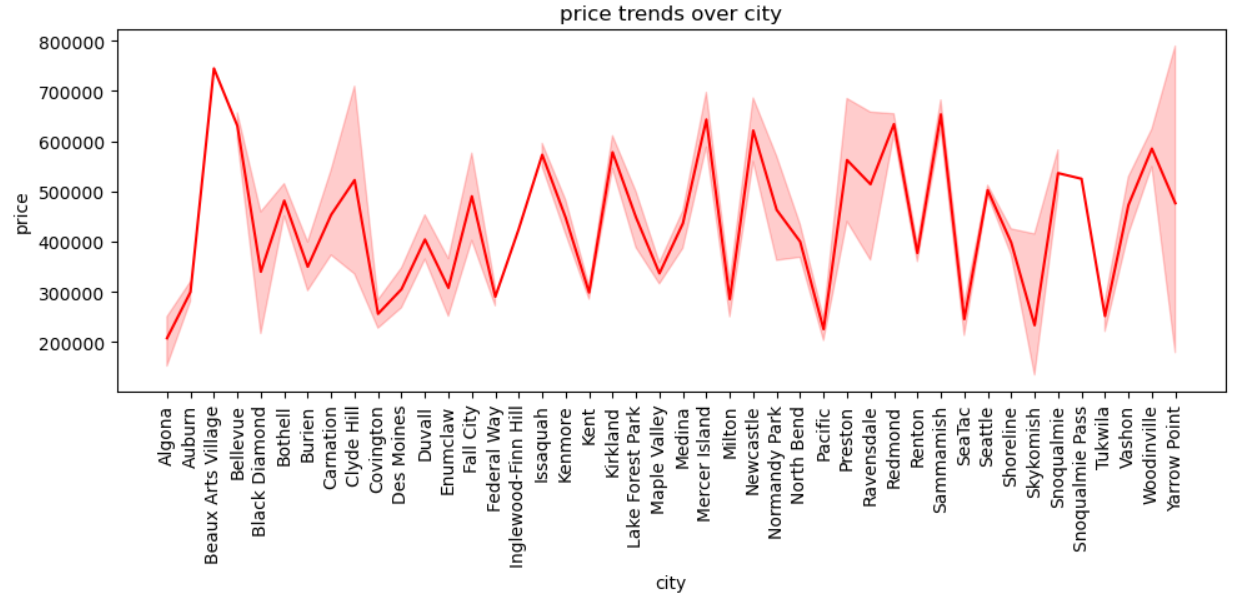


#### Here the floors with 1 has the highest number of sales.

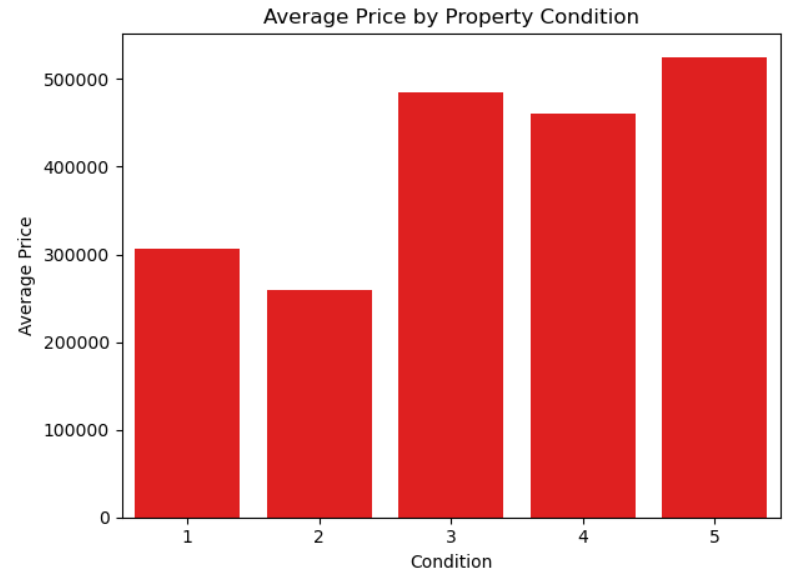
Lineplot:



#### Here the highest sales is between 1920 to 1940.

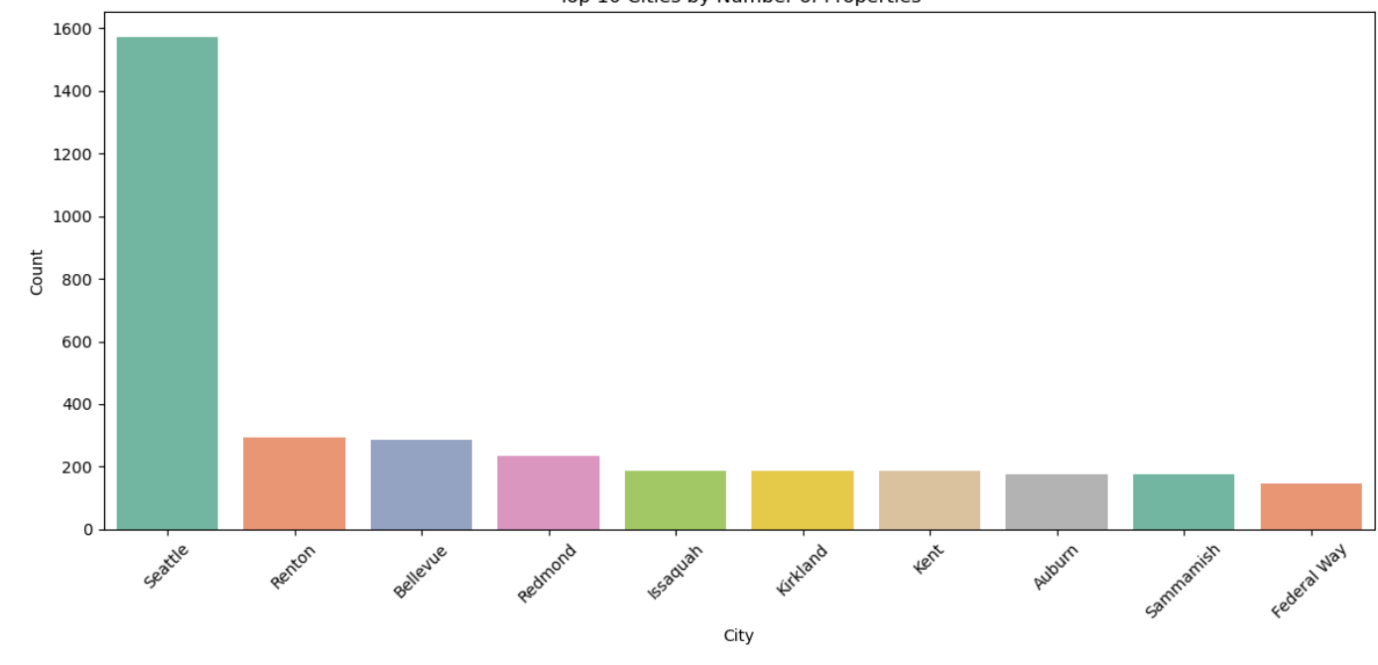


Bar plot:



#### Here the condition with 5 has the highest sales.

Count plot:



#### Here SEATTLE city has the highest count of sales.

**13.MULTIVARIATE ANALYSIS:**

It is a statistical technique used to examine data that involves multiple variables simultaneously

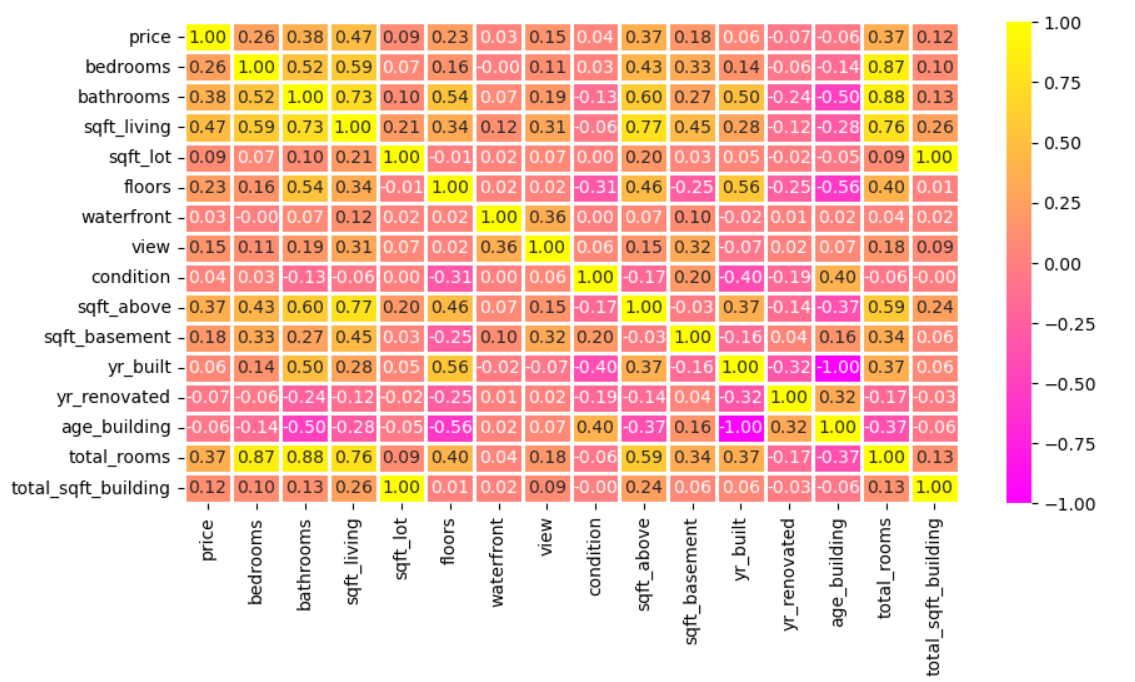
Instead of analyzing one variable at a time (uni-variate) or two variables (bi-variate) , multivariate analysis explores the relationships among three or more variables to uncover patterns, correlations, and influences.

**Common Techniques:**

**Multiple Regression**: Predicts one variable using several others

**Factor Analysis**: Reduces data by grouping correlated variables

**Heat map:**



1. **INSIGHTS AND RECOMMENDATIONS:**

**• Living area has the more influencing factor to the price of the sales.**

Total square area of the building is strongly connected with the price.since,the larger square area has the demand of higher sales price.

**• Total rooms of the building also plays the most significant roles in the price of the sales.**

If the house has more bedrooms and bathrooms sold to the highest price.

The bedrooms with 3 and 4 has the highest rate of sales.

The bathrooms with 2 and 3 has the highest rate of sales.

More customers have the potential to buy the total number of rooms with 6.

So we have to built more number of building with 6 rooms.

So we can increase our sales rate and also the price rate.

**• Condition have the strong link with price of the house**

The condition with 4 and 5 has the highest number of sales.

So we have to built the condition with 4 and 5.

By this we can increase our sales rate.

**• Properties with views strong relationship with price.**

The views with 3 and 4 has the highest number of sales.

So we have to built the views with 3 and 4.

It increase the sales rate.

**• Newly built house has the highest demand rate**

The customers are preferring newly built house.

It increase the price of the building.

**• Renovated building also plays major role**

The customer also preferring the renovated buildings.

So we have renovate the old buildings to increase the price of the building.

**• Most buildings are mid-aged(20-40 years)**

The age of the building shows that market is mature and the buildings benefits from the renovation in a continuous period of time.

**• Price is right-skewed**

The price is right-skewed due the luxury buildings.

The also buying luxury buildings.

So we have built luxury buildings for attracting customers and expand our market value.

**• Average price of the building**

The average price of the building is between 4000000 and 6000000.

So we have to built more buildings in this price range to increase our sales rate.

**• Sales by city wise**

The SEATTLE has the highest number of sales.

So we have to increase the construction rate of the building to increase the sales.

Many of the cities has lowest sales, we have to focus on that cities to increase our sales rate.

1. **conclusion:**

This housing data shows that value of the property is impacted by the location of the building,total square area of the building, condition of the building, renovation of the building,total rooms of the building.

By detailed data exploration and statistical tests we find the clear patterns and trends over the time and the factors and the price trends.

This analysis gives the detailed insights for the buyers,sellers and investors who want to make a decision to purchase or selling the property.

This analysis is helpful for the real estate persons to make a decision.