Exploratory Data Analysis of Housing Dataset

DIVYA PRASATH R S

DA&DS-MAY’25

REINFORCEMENT PROJECT-1

23-07-2025

(OFFLINE)

Table of Content

1.Introduction :

The Housing market is so Dynamic ,We need a deep Understanding in this data and we need to find what drives the price of the property to change because this data is essential for anyone looking to buy, sell or invest in the properties.

In this project we are going to dive deep into the detailed Housing Data to discover the trends and Features which influence the home values.

In this dataset we have important columns like:

- 'bathrooms',
- 'bedrooms',
- 'city',
- 'condition',
- 'country',
- 'date',
- 'floors',
- 'price',
- 'sqft_above',
- 'sqft_basement',
- 'sqft_living',
- 'sqft_lot',
- 'statezip',
- 'street',
- 'view',
- 'waterfront',
- 'yr_built',
- 'yr_renovated'.

With these columns we can get an view over the housing market.

We use Python as the programming language and we use Libraries like Pandas, NumPy, Matplotlib, and Seaborn. We can use these libraries for Data Understanding, Data cleaning, EDA and for Data Visualization.

2.AIM:

The Main goal of this project is to build a simple but efficient python code to find the feature that matters the most important while defining prices of the properties.

To achieve this goal we will follow a certain order that will start from:

- Data Understanding
- Data Cleaning
- Exploratory Data Analysis (EDA)
- Data Visualization
- Feature Engineering
- Analysis and Interpretation

By Data analyzing the key factors like Location, Property size, Condition of the property, Number of Bathrooms and Bedrooms, Views and whether the house is Renovated or not .From these features we can analyze the shape of market value.

3.Problem Statement:

In this Housing Dataset we are going the Define the Right price for a property.

By considering the factors ranging from location to the renovations.

The housing dataset provides comprehensive information on various attributes associated with residential properties, including price, number of bedrooms and bathrooms, square footage, location details, and other relevant features. The objective of this project is to conduct an in depth analysis of the dataset to derive valuable insights for stakeholders in the real estate industry.

4.Project Workflow:

To understand the factors which makes changes in home prices, we will follow these simple steps:

Data Collection :

First we will get the data from a Kaggle dataset or from other sources . In this project I got my data from my Teacher and stored it in the .csv file formate in my local drive (C:\Users\divya\Downloads).

Then we will assign the data to a variable and then we will take the assigned variable and we will take a copy of the Housing dataset which is stored in x variable then I stored the copied data in a variable name **house_data** . Because if we don't take a copy from the original dataset we might corrupt or change the data from it so we use copy method here.

Data Cleaning

We will find the null values in the given dataset . By using **house_data.info()** to check the datatypes, Non-null count and columns .

We will use **house_data.isnull().sum()** to identify the null values in the columns

EDA:

We will use columns to give a visual analysis of the data we use EDA it has 3 types:

- ❖ Univariate Analysis
- ❖ Bivariate Analysis
- ❖ Multivariate Analysis
  - ✓ Univariate Analysis
  - ✓ Bivariate Analysis
  - ✓ Multivariate Analysis

Lets discuss these in detail in the upcoming sections .

Adding new Features:

We create new columns based on the given Data like Age of the Property, price per square feet, Does the property has a basement and whether the house is renovated or not.

Insights :

We will analyse the data and give the analysed data to the clients.

5.Data Understanding:

Provide a detailed description of the dataset, including the structure, dimensions (rows and columns), and data types and basic statistics and early insights.

We use **house_data.shape** to find how much Rows and columns present in the dataset .

We use house_data.describe() to do descriptive statistics.

6.Data Cleaning:

We will find the null values in the given dataset.

We will use **house_data.isnull().sum()** to identify the null values in the columns .

Checking Outliers

By using the IQR method we will find the outliers from the columns.

And we will replace the Outliers with the Median of the respective column.

7.Obtaining Derived Metrics:

We create new columns based on the given Data like Age of the Property, price per square feet, Does the property has a basement and whether the house is renovated or not.

We create a new column named **house_age**, **price_per_sqft**, **has_basement and is_renovated.**

- ✓ Property Age: How old each home is, based on the year it was built.
  Price per Square Foot: A way to compare homes of different sizes fairly.
  Renovation Status: A simple indicator showing whether the home had been renovated.
  These new features helped us better understand how age, size, and renovations impact prices.
  They also made it easier to identify trends and compare different homes across the dataset.

8.Filtering Data for Analysis:

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

9.Statistical Analysis:

Descriptive Analysis: Summarize the key descriptive statistics (mean, median, mode, etc.).

One way ANOVA test

- ❖ 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.

Two Way T-test

- ❖ The null hypothesis (H_0) assumed no difference in average prices between two groups (e.g., renovated vs. non-renovated homes).
- ❖ Since the p-value < 0.05, we reject H_0, indicating a significant difference in average prices between the groups.

10.Exploratory Data Analysis (EDA)- Univariate Variables:

Exploratory Data Analysis (EDA) is important for several reasons in the context of data science and statistical modeling. Here are some of the key reasons:

- ✓ It helps to understand the dataset by showing how many features it has, what type of data each feature contains and how the data is distributed.
- ✓ It helps to identify hidden patterns and relationships between different data points which help us in and model building.

- ✓ Allows to identify errors or unusual data points (outliers) that could affect our results.
- ✓ The insights gained from EDA help us to identify most important features for building models and guide us on how to prepare them for better performance.
- ✓ By understanding the data it helps us in choosing best modeling techniques and adjusting them for better results.

UNIVARIATE ANALYSIS:

Univariate analysis focuses on studying one variable to understand its characteristics. It helps to describe data and find patterns within a single feature. Various common methods like histograms are used to show data distribution, box plots to detect outliers and understand data spread and bar charts for categorical data. Summary statistics like **Mean, Median, Mode, Variance and Standard deviation** helps in describing the central tendency and spread of the data

11. Bivariate Analysis:

Bivariate Analysis focuses on identifying relationship between two variables to find connections, correlations and dependencies. It helps to understand how two variables interact with each other. Some key techniques include:

- Scatter plots which visualize the relationship between two continuous variables.
- **C**orrelation coefficient measures how strongly two variables are related which commonly use **Pearson's correlation** for linear relationships.
- Cross-tabulation or contingency tables shows the frequency distribution of two categorical variables and help to understand their relationship.
- **Line graphs** are useful for comparing two variables over time in time series data to identify trends or patterns.

- **Covariance** measures how two variables change together but it is paired with the correlation coefficient for a clearer and more standardized understanding of the relationship.

12.Multivariate Analysis:

Multivariate Analysis identify relationships between two or more variables in the dataset and aims to understand how variables interact with one another which is important for statistical modeling techniques. It include techniques like:

- **Pair plots** which shows the relationships between multiple variables at once and helps in understanding how they interact.
- Another technique is **Principal Component Analysis (PCA)** which reduces the complexity of large datasets by simplifying them while keeping the most important information.
- **Spatial Analysis** is used for geographical data by using maps and spatial plotting to understand the geographical distribution of variables.
- **Time Series Analysis** is used for datasets that involve time-based data and it involves understanding and modeling patterns and trends over time. Common techniques include line plots, autocorrelation analysis, moving averages and **ARIMA** models.

13.Overall Insights from Analysis:

• Living Area is the Most Influential Factor

sqft_living has the strongest link to price. Larger homes demand significantly higher prices.

• More Bathrooms > More Bedrooms

Bathrooms have a stronger relationship with price than bedrooms. This shows that utility matters more than the number of rooms.

• Renovated Homes Are Worth More

A t-test showed that renovated homes have significantly higher average prices than non-renovated ones ($p < 0.05$).

- **Renovation & Condition Are Related**

A Chi-Square test confirmed that renovated homes usually have better condition ratings. Condition also affects price.

- **Condition Drives Value**

ANOVA showed that homes in better condition cost more, even when other factors are held constant.

- **Properties with Premium Views Are Pricier**

View score is positively linked to price. Homes with higher view ratings generally sell for more.

- **Newer Homes Usually Cost More**

The property_age distribution and line plot showed that newer homes often fetch higher prices, although the trend is not perfectly linear.

- **Many Homes Have No Basement, But It Adds Value**

Many homes have sqft_basement = 0, but those with basements generally sell for higher prices, especially if they are finished.

- **Most Homes Are Mid-Aged (20–50 years)**

The age distribution shows a mature market, where many homes could benefit from renovation and modernization.

- **Price is Right-Skewed Due to Luxury Homes**

A small percentage of high-value properties significantly raise the average price. This highlights the need to analyze luxury listings separately.

- **3 Bedrooms and 2–3 Bathrooms is the Market Norm**

Most homes have 3 bedrooms and 2–2.5 bathrooms. This shows a typical family-oriented property layout.

14.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.