**PROJECT DOCUMENTATION**

EXPLORATORY DATA ANALYSIS USING PYTHON

|  |  |
| --- | --- |
| TITLE | Exploring Patterns in Housing Dataset |
| NAME | MATHESWARAN L |
| COURSE | DADS – Offline |
| BATCH | August- 2025 |

TABLE OF CONTENT

|  |  |
| --- | --- |
| 1 | **Introduction** |
| 2 | **Aim** |
| 3 | **Business Problem / Problem Statement** |
| 4 | **Project Workflow** |
| 5 | **Data Understanding** |
| 6 | **Data Cleaning**  Missing Values Imputation  Outlier Treatment  Handling Inconsistent Values |
| 7 | **Obtaining Derived Metrics** |
| 8 | **Filtering Data for Analysis** |
| 9 | **Statistical Analysis**  Descriptive analysis  Test statistics and hypothesis testing |
| 10 | **Exploratory Data Analysis (EDA) - Univariate Analysis** |
| 11 | **Bivariate Analysis** |
| 12 | **Multivariate Analysis** |
| 13 | **Overall Insights from Analysis** |
| 14 | **Conclusion** |

**INTRODUCTION**

In today’s data-driven world, understanding the factors that influence housing prices is crucial for buyers, sellers, investors, and real estate professionals. This project focuses on performing Exploratory Data Analysis (EDA) on a housing dataset to uncover key patterns, trends, and relationships that affect property prices. By analysing various features such as square footage, number of bedrooms and bathrooms, grade, and location-related attributes, the project aims to extract meaningful insights from the data. Python-based analytical methods and visualizations are used to explore the dataset and prepare it for future predictive modelling or decision-making applications.

This study helps in understanding how different variables affect property value and assists in making data-driven decisions for buyers, sellers, and real-estate analysts. The project also includes correlation analysis, univariate and bivariate analysis, multivariate analysis, feature engineering, and visualization techniques to extract insights from the dataset

**AIM OF THE PROJECT**

The main aim of this project is to analyse the housing dataset using Exploratory Data Analysis techniques to understand the underlying structure and key factors influencing home prices.

The project seeks to identify important variables, discover trends, study relationships between numerical and categorical features, and highlight how various attributes affect price variations.

Another aim of the analysis is to clean, prepare, and transform the dataset so that it becomes suitable for further statistical modelling.

Overall, the project intends to provide insights that can support better decision-making in real estate pricing and investment.

**PROBLEM STATEMENT**

The real estate market is highly dynamic, with property prices influenced by numerous factors such as location, physical characteristics, quality, and amenities. Many buyers struggle to evaluate whether a listed price is reasonable, while sellers lack data-backed insights to price their properties competitively. Real estate agencies also face challenges in identifying market trends and understanding customer preferences.  
 This project addresses the problem of understanding the key determinants of housing prices by conducting an in-depth exploratory analysis of a housing dataset. By identifying meaningful patterns and correlations between property features and their prices, the analysis provides clarity on which attributes significantly affect home value. These insights can help in improving pricing strategies, enhancing investment decisions, and developing future predictive models for automated price estimation.

**PROJECT WORKFLOW**

Data Collection: Loaded housing dataset and examined basic structure. Verified data dimensions and initial data quality

Data Cleaning & Pre-processing: Handled missing values and converted data types. Created simplified categorical variables from complex descriptions

Exploratory Data Analysis (EDA): Conducted univariate analysis of key variables. Performed bivariate analysis to examine relationships. Applied multivariate analysis for deeper insights.

Visualization & Insights: Created charts to identify patterns and trends. Examined distributions and relationships visually.

Feature Engineering: Developed new categorical groupings and binary indicators. Generated time-based variables for analysis.

Hypothesis Testing: Tested statistical relationships between different variables. Validated patterns through appropriate statistical tests.

Conclusion & Recommendations: Summarized key findings and patterns. Provided insights for stakeholders.

**DATA UNDERSTANDING**

The dataset (housing.csv) contains multiple features that describe the physical and qualitative characteristics of residential properties, along with their sale prices.

* **Rows – 4600**
* **Columns – 18**

**Key Variables**

Date floors yr\_built

Price waterfront yr\_renovated

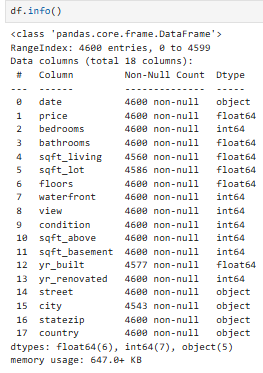
Bedrooms view street

Bathrooms condition city

Sqft\_living sqft\_above statezip

Sqft\_lot sqft\_basement country

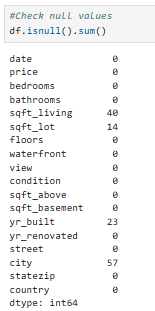
This dataset has 5 object columns, 6 in float and 7 in int columns. Columns like latitude and longitude are neglected because, not relevant here otherwise it can be used for more map based analysis.



**CLEANING AND TRANSFORMATION**

Data cleaning is the next step in EDA, Cleaning is one of the most important step in Data Science and clean data ensure accurate and proper insights. Data Cleaning steps include handling missing values, Changing columns names, Dropping Irrelevant columns, Formatting the data into proper format.

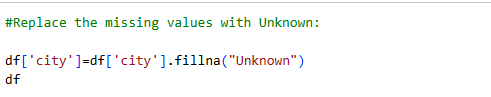
**Checking for null values:**





The dataset have 134 null values.

**Replacing with Unknown:**

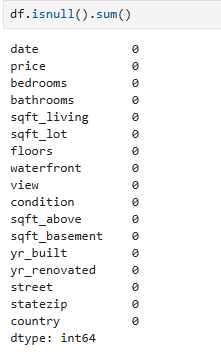


The null values are filled with Unknown.

**Converting to median:**

The null values are filled by using fillna ().

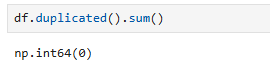
Numeric columns are filled with median value. Because the data are normally distributed in this dataset.

****

After filling null values the dataset does not contain any null values.

****

And I converted Date into Date format and using to\_datetime () function.



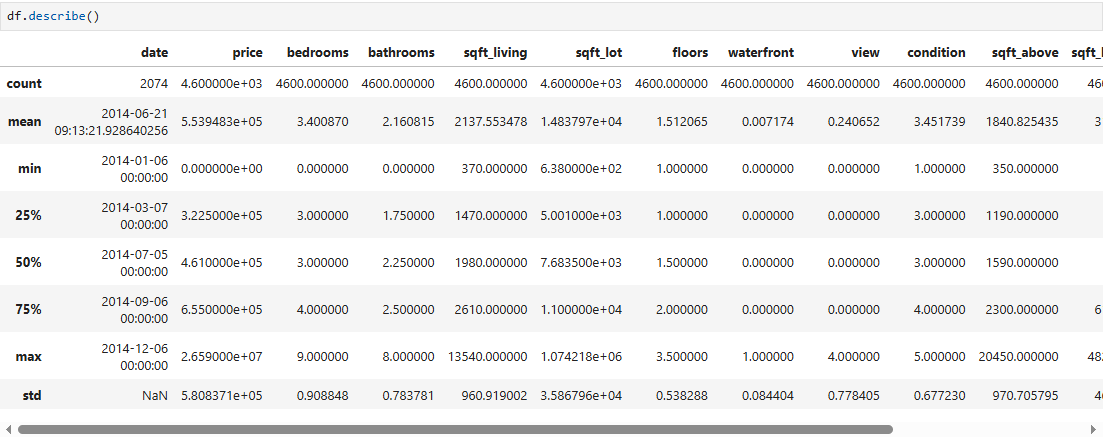
After I checked the dataset have any duplicate values and i find 0 duplicates.

**STATISTICAL ANALYSIS AND TESTING**

**Descriptive Statistics:**

Descriptive statistics summarize and describe the main features of a dataset. They include metrics like mean, median, standard deviation, minimum, maximum, and quartiles, which give a snapshot of the distribution and spread of values.

The descriptive statistics of the score column provide an overview of laptop price analysing performance across ~4600 records.

****

**Hypothesis Testing:**

Hypothesis testing is a statistical method used to validate whether an observed pattern in the data is due to chance or represents a meaningful difference/relationship. It involves setting up a Null Hypothesis (H0), which assumes no effect or difference, and an Alternative Hypothesis (H1), which assumes there is a significant effect or difference. The decision to reject or accept H0 is based on a p-value compared to a significance level (alpha = 0.05 in this case).

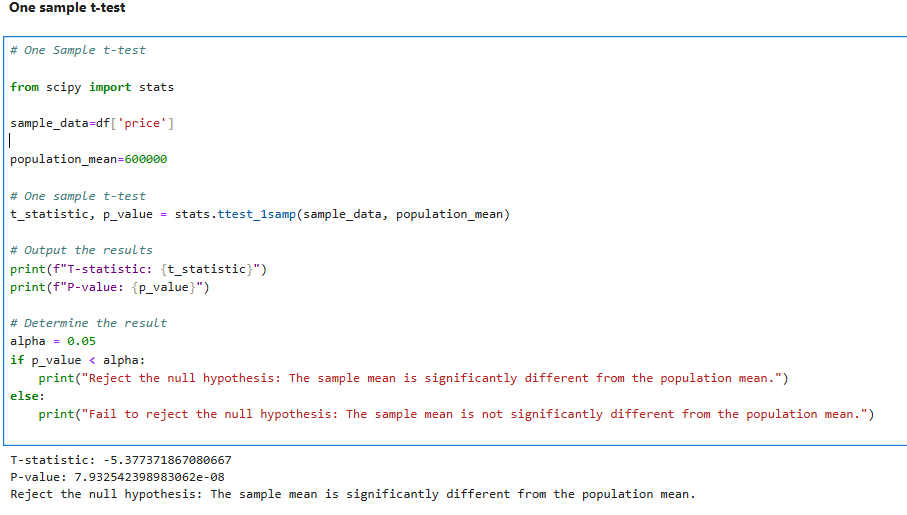
**Tests Performed**

**One sample t-test:**

**H0**: A one-sample t-test was conducted to determine whether the mean housing price differs from the population mean of ₹6, 00,000.

**H1**: A one-sample t-test was conducted to determine whether the mean housing price differs from the population mean is not ₹6, 00,000.

**Result:** Reject the null hypothesis: The sample mean is significantly different from the population mean.



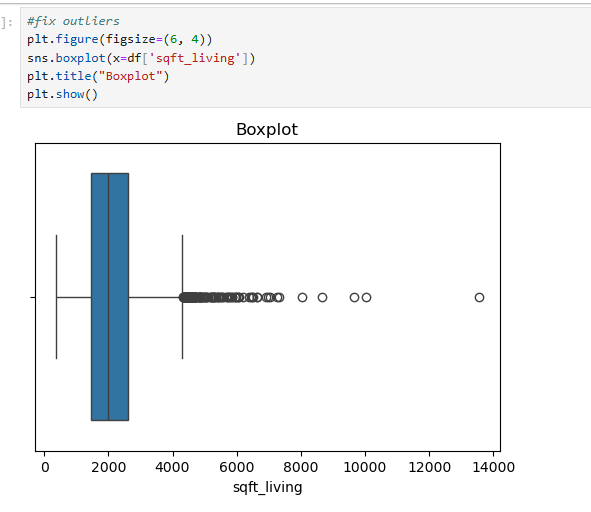
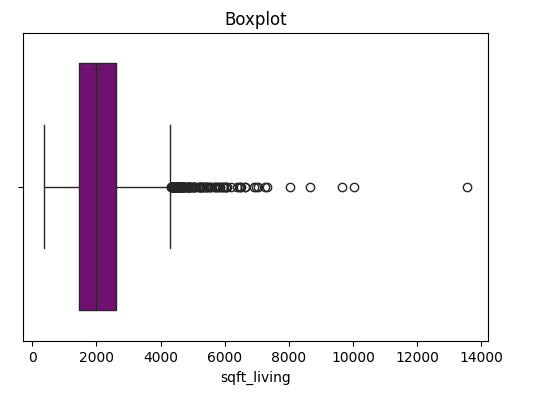
**Chi Square Test: Bedrooms vs Condition**

**H0:** Housing Bedrooms and Housing Condition are independent of each other.

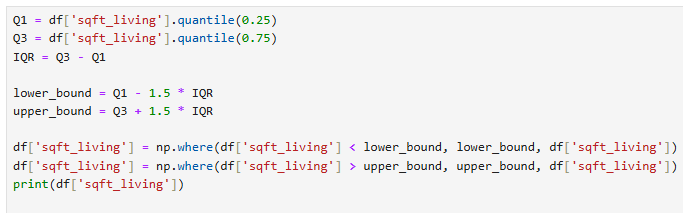
**H1:** Housing Bedrooms and Housing Condition are dependent (associated).

**Result:** Reject H0 there is a significant relationship.

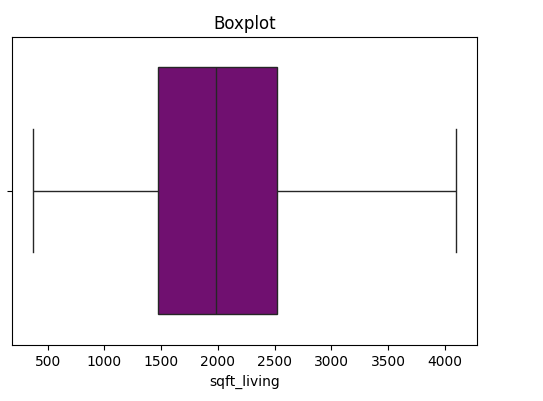


**Outliers **

Outliers are unusual values that don’t fit the general pattern of the data. They can arise due to data entry errors, rare cases, or genuine market differences. Handling them properly ensures meaningful insights.



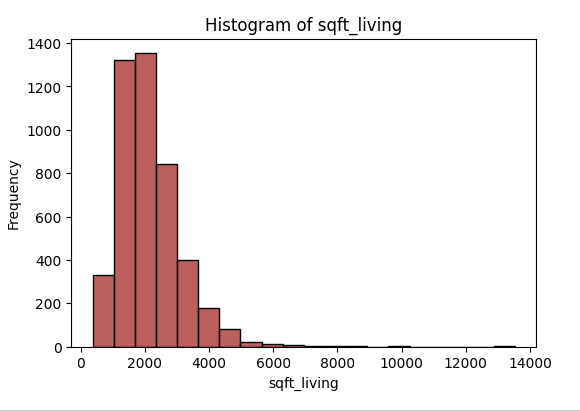
Boxplot is used to find outliers in Exploratory Data Analysis.

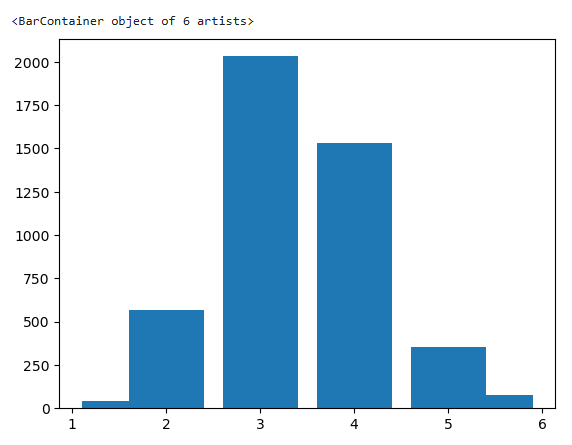


**EXPLORATORY DATA ANALYSIS - UNIVARIATE ANALYSIS**

Univariate analysis focuses on exploring individual variables to understand their distribution, patterns, and anomalies.

|  |  |  |
| --- | --- | --- |
| Variable | Chart Type | Key Insights |
| Sqft\_living | Histplot | Most Housing are highly Frequency between 2000 to 2500. |
| Bedrooms | Count Plot | 3 and 4 is highly Selled. |

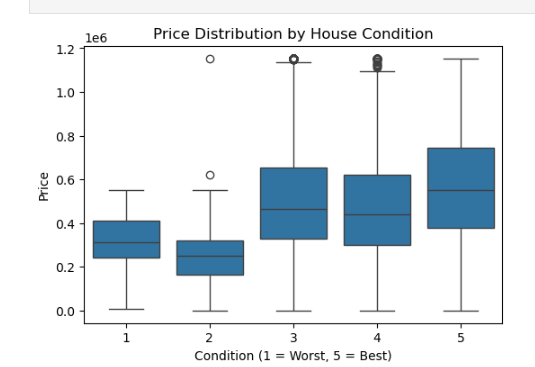




**EXPLORATORY DATA ANALYSIS - BIVARIATE ANALYSIS**

Bivariate analysis helps explore relationships between two variables, uncovering trends, correlations, and dependencies that impact housing prices.

In this bivariate analysis, the relationship between **house price** (numerical) and **condition** (categorical) is examined to determine whether the physical condition of a house influences its market price.



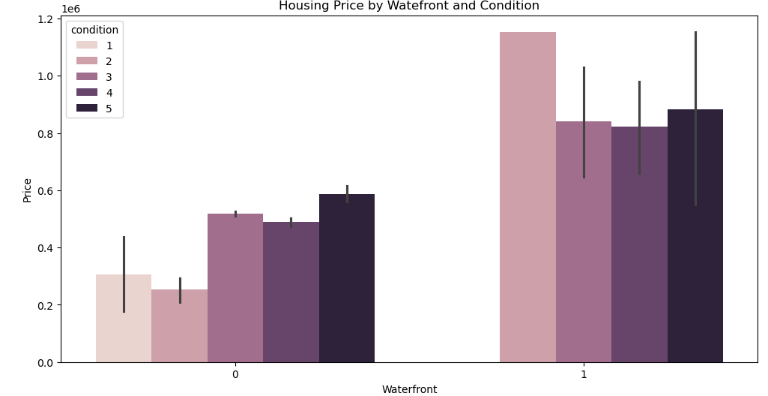
A **boxplot** was used to compare the distribution of prices across the five condition categories (1 = Worst, 5 = Best). The plot clearly shows that houses with **higher condition ratings** generally tend to have **higher median prices** compared to homes in poorer condition. Additionally, lower condition categories (1 and 2) show greater price variability and the presence of outliers, indicating inconsistent pricing for poorly maintained homes.

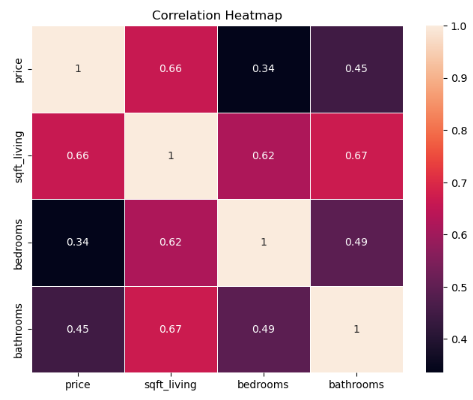
Overall, the distribution suggests that **better condition houses are typically more expensive**, supporting the expectation that property condition is an important factor influencing housing prices.

**EXPLORATORY DATA ANALYSIS - MULTIVARIATE ANALYSIS**

Multi variate analysis is used to find relationship between multiple attributes, Mostly heatmaps and pair plots are used to find the correlation and relationship between numerical columns. This dataset has only one valid numerical column, Instead of correlation heatmap, other charts show multiple relationships is used.

|  |  |  |
| --- | --- | --- |
| Variable | Chart Type | Key Insights |
| Waterfront vs  Price vs  Condition | Bar plot | Well-maintained waterfront homes reach the highest price ranges, while non-waterfront homes in lower conditions are priced much lower. |
| Numeric Columns | Heatmap | 1. Price has a strong positive correlation with sqft\_living, showing that larger homes generally cost more.  2. Bedrooms and bathrooms have moderate correlation with price, meaning they influence price but not as strongly as total living area. |





**OVERALL INSIGHTS FROM ANALYSIS.**

* Price correlates strongly with sqft\_living.
* Waterfront properties command significantly higher prices.
* Bedrooms and bathrooms positively influence price, with common configurations being 3-4 beds/1.5-2.5 baths.
* Floors and condition have a significant relationship.
* Price distribution is right-skewed with outliers, which were treated.
* No clear short-term price trend observed over time.

**CONCLUSION**

Missing Value Handling: Identified and addressed missing values in sqft\_living, sqft\_lot, and yr\_built through median imputation, and dropped the city column.

Outlier Treatment: Outliers in sqft\_lot, bedrooms, sqft\_living, and price were treated using the IQR method to prevent skewed analysis.

Data Type Conversion: The date column was successfully converted to a datetime object, enabling time-series analysis.

Key Price Drivers: Confirmed that sqft\_living, bedrooms, bathrooms, and waterfront access are significant factors influencing housing prices.

Statistical Insights: Statistical tests revealed a significant relationship between bedrooms and condition, and a notable difference in mean prices for houses with/without waterfront.

Enhanced Dataset: New features like age\_of\_property and price\_per\_sqft were engineered, resulting in a cleaned and enriched dataset suitable for predictive modelling.

**Statistical analysis and**

**Transformation**

**Cleaning and Transformation**

**Data cleaning is the next step in EDA, Cleaning is one of**