**Project Documentation**

**Exploratory Data Analysis using Python**

**Project Information** - **Title:** Housing Price Analysis Using EDA & Hypothesis Testing

**Name:** Muruganantham.R

**Course:** DA/DS , Offline

**Roll Number:** 16525CBRE21

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** |

### 1. Introduction

This project analyzes a real estate dataset with 4600 entries and 18 features to understand the key variables that impact housing prices. Using Python libraries such as pandas, matplotlib, seaborn, and scipy, we perform descriptive analytics, statistical testing, and data visualization. The project helps in uncovering trends, correlations, and patterns that drive pricing in the housing market.

### 2. Aim

The objective is to conduct an exploratory data analysis and perform hypothesis testing on a housing dataset to discover influential features affecting home prices. Through visualizations and statistics, we aim to validate assumptions and offer data-driven insights for better decision-making in real estate pricing and investment.

### 3. Business Problem / Problem Statement

In real estate, pricing is influenced by numerous variables including size, location, and amenities. Buyers and sellers often make assumptions without validating them with data. This project seeks to analyze real housing market data to answer: Does having more square footage or waterfront access significantly increase price? Are there seasonal price trends? Insights from this study can guide agents, investors, and policy makers.

### 4. Project Workflow

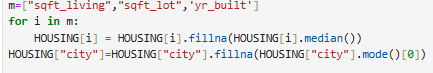
* Load the housing dataset
* Perform data cleaning and handle missing values
* Conduct feature engineering (derived metrics)
* Perform hypothesis testing (t-test, ANOVA, correlation)
* Conduct univariate, bivariate, and multivariate analysis
* Generate insights
* through visualizations

### 5. Data Understanding

The dataset has 4600 rows and 18 columns. Key fields include price, bedrooms, bathrooms, sqft\_living, and waterfront. Price is the target variable. Most houses have 3–4 bedrooms, 1–2 bathrooms, and 1000–2500 sqft living area. Data types are mostly numeric, with a few categorical (e.g., city, statezip).

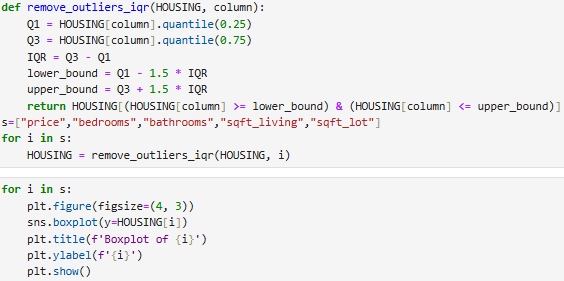
### 6. Data Cleaning

* **Missing Values:** Some missing values in columns like sqft\_living and city were dropped for analysis integrity.





* **Outliers:** Outliers in price and sqft were detected using boxplots. Extreme values were retained since they represent luxury properties.



* **Inconsistencies:** Date and categorical columns were cleaned and transformed for analysis. No severe structural inconsistencies were found.

### 7. Obtaining Derived Metrics

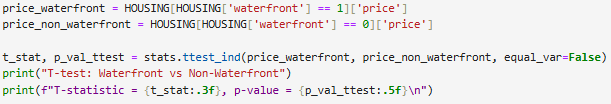
* Created a binary feature is\_renovated from yr\_renovated.
* total\_sqft = sqft\_above + sqft\_basement
* Extracted month from date to analyze seasonality in prices.

### 8. Filtering Data for Analysis

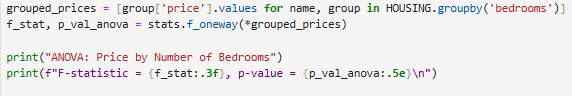
To ensure clean statistical results, rows with null values in key columns (price, sqft\_living, bedrooms, waterfront) were dropped. Data was filtered to include only entries with valid price ranges and living areas.

### 9. Statistical Analysis

* **Descriptive Analysis:**
  + Average price: ~$540,000
  + Most homes: 3 bedrooms, 2 bathrooms
  + sqft\_living typically between 1000–2500
* **Hypothesis Testing:**
  + **T-test:** Waterfront vs non-waterfront price (p = 0.00095) → significant



* + **ANOVA:** Price differs by bedrooms (p < 1.48e-34) → significant



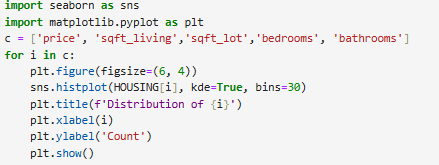
* + **Correlation:** sqft\_living vs price (r = 0.419, p < 2.02e-193) → positive correlation

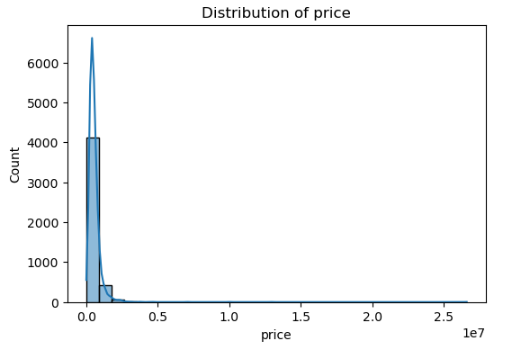
### 10. Exploratory Data Analysis (Univariate)

* Price distribution is right-skewed
* Bedroom and bathroom counts show clear modes (3 bedrooms, 2 bathrooms)
* sqft\_living follows a bell shape with a long right tail
* sqft\_lot has a wide range with few very large properties

### 11. Bivariate Analysis

* Price increases with sqft\_living (confirmed visually)



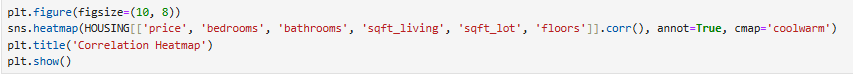


* Bedrooms vs price: not linear, but 3-4 bedrooms are most valuable
* Correlation matrix shows highest relation between price and sqft\_living, bathrooms, and bedrooms

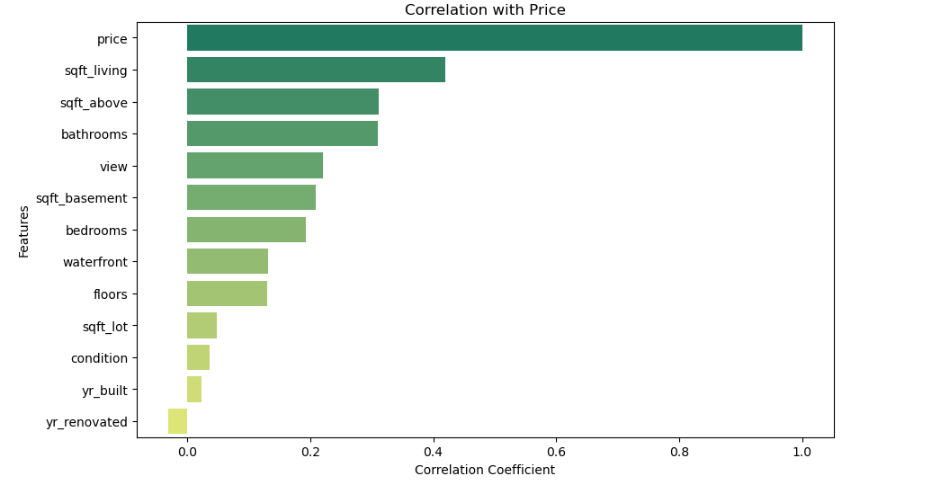
### 

### 12. Multivariate Analysis

* Heatmap confirms high correlation between sqft\_living, bathrooms, and price



* Pairplots show linear trends between price and continuous variables
* Bar plot: Top cities like Mercer Island and Clyde Hill have highest average prices



* Price trend line: Prices dropped mid-year and recovered slowly, showing seasonality

### 13. Overall Insights

* Square footage (sqft\_living) is the most reliable predictor of price
* Homes on waterfronts are significantly more expensive
* Bedroom count affects price, but only up to a point (3–5 bedroom homes most common)
* There are clear outliers in price, especially among luxury homes with large sqft or special locations
* Prices dip mid-year and rise again toward the end, indicating a seasonal pattern

### 14. Conclusion

Through EDA and statistical testing, this project demonstrates the power of data in understanding housing prices. Significant relationships were found between price and features like sqft\_living, waterfront access, and bedroom count. These insights can help in real estate forecasting, pricing strategies, and investment decisions.