Project Documentation: Exploratory Data Analysis using Python

Project Information

- **Title:** Exploratory Data Analysis of Housing Dataset
- **Name:** Prayag Sujith
- **DA/DS:** June Batch
- **Batch Number:**
- **Online/Offline:** Online
- **Roll Number:** B2025057640

Table of Contents

1. Introduction

This project analyzes the Housing Dataset which contains records for 4,600 home sales. The data includes key details like sale price, number of bedrooms, and house size. Our main goal is to use Python to explore this data and find out what really drives property values. We rigorously clean the data to remove errors and use statistical tests to see how different features connect to price. Instead of relying on guesses, this project aims to prove exactly which physical features such as size and amenities have the biggest impact on a home's cost .

2. Aim

The main aim of this project is to analyze housing data to find out exactly what drives property prices. We use Python to clean the data and remove errors, ensuring our results are accurate. Specifically, we want to measure how much factors like square footage, the number of bathrooms, or the age of the house affect its value. By using statistical methods like correlation, we seek to prove which features matter most. Ultimately, this project aims to help buyers and sellers estimate prices correctly using real data instead of just guessing .

3. Business Problem / Problem Statement

Determining the fair market value of a residential property is one of the most difficult challenges in the real estate industry. This difficulty arises because housing prices are influenced by a complex mix of features that vary from house to house. Sellers often base their asking prices on emotional attachment or hopeful guessing, while buyers struggle to know if a listed price is fair. This lack of clarity creates significant financial risks for everyone involved. An overpriced home might sit unsold on the market for months, losing value over time, while an underpriced home results in an immediate financial loss for the seller. The problem this project addresses is the reliance on intuition rather than evidence. By analyzing historical sales data, we can replace guesswork with hard numbers. This analysis is valuable because it mathematically proves which specific features, such as living space or the number of bathrooms, actually drive market value. This allows buyers, sellers, and agents to make confident, data driven financial decisions .

4. Project Workflow

The project followed a structured Data Science workflow using Python to ensure accuracy and reproducibility. We began by collecting the raw `Housing.csv` data and loading it into our coding environment for inspection. The first step was Data Understanding, where we examined the dataset's structure to identify errors. Next, we performed Data Cleaning by filling in missing information with statistical averages and removing invalid data points, such as homes with a price of zero. To make the analysis deeper, we used Feature Engineering to create new useful variables like the age of the house. Finally, we conducted a Statistical Analysis to find correlations and created visual charts to display the results. This methodical process ensured that our final insights were based on high quality, error free data rather than raw, noisy information

5. Data Understanding

The dataset we used for this project contains **4,600 rows and 18 columns**. The most important column is the **price**, which is the target we want to analyze. The other columns give us details about the house, such as the size of the living area, the number of bedrooms and bathrooms, and the year it was built. Most of the data consists of numbers (like square footage and price), but there are also text columns for location details like the city and street address .

When we first inspected the data, we found some quality issues that needed attention. The summary showed that we had missing information in four columns: the living area size, the lot size, the year built, and the city name. We also looked at the basic statistics and found some strange errors. For example, the minimum price for a house was listed as zero dollars, and the minimum number of bedrooms was also zero. These were clearly data errors because a house cannot be sold for free. We also noticed some extreme values, like a house priced over 26 million dollars, which is far higher than the average. These initial insights told us that we needed to clean the data carefully before we could trust our analysis .

6. Data Cleaning - Missing Values Imputation, Outliers, Handling Inconsistent Values

Cleaning the data was the most critical step in our workflow to ensure our analysis was accurate. Real world data often has errors, so we used Python to fix them systematically.

Missing Values Imputation First, we addressed missing information. Our initial check found gaps in numerical columns like the living area size and the year the house was built. Instead of deleting these incomplete rows, which would lose valuable data, we used a method called imputation. We calculated the average (mean) value for each column and filled the empty spots with that number. For the city column, which contains text, we filled the missing entries with the most common city name (the mode). This kept our dataset complete without changing the overall trends.

Outlier Treatment Next, we handled outliers, which are extreme values that can distort the results. We found some luxury properties priced as high as 26 million dollars. These values were so far from the normal range that they pulled the average price up misleadingly. To fix this, we set a limit and filtered out any properties priced over 5 million dollars. This ensured our analysis focused on the standard housing market.

Handling Inconsistent Values Finally, we fixed logical errors in the data. We found records of houses listed with a price of zero dollars or zero bedrooms. Since these are impossible for a valid sale, we treated them as inconsistent data and removed those rows entirely. This process reduced our raw data to a clean, reliable set of 4,545 records ready for analysis .

7. Obtaining Derived Metrics

To get deeper insights from our data, we created two new variables that were not in the original file. This process is called Feature Engineering.

First, we created **House Age**. The original dataset only had the "Year Built," which is just a calendar year. By subtracting the year built from the current year of the data (2014), we calculated the actual age of each home. This makes it much easier to analyze if older homes are cheaper than newer ones.

Second, we calculated **Price per Square Foot**. We did this by dividing the total price by the square footage of the living area. This is a very important number in real estate because it allows us to compare the value of a small apartment directly against a large mansion. These new metrics give us a more accurate way to measure value across different types of properties .

8. Filtering Data for Analysis

We filtered the data to ensure we analyzed only valid house sales. First, we removed logical errors. We found records of houses with a price of zero dollars or zero bedrooms. Since these are impossible for a real sale, we deleted those rows. Next, we removed extreme outliers. We filtered out luxury properties priced over five million dollars because they are rare and would skew the average price for standard homes. By removing these 55 problematic rows, we refined our dataset from 4,600 down to 4,545 high quality records. This step made our final statistics accurate and reliable .

9. Statistical Analysis

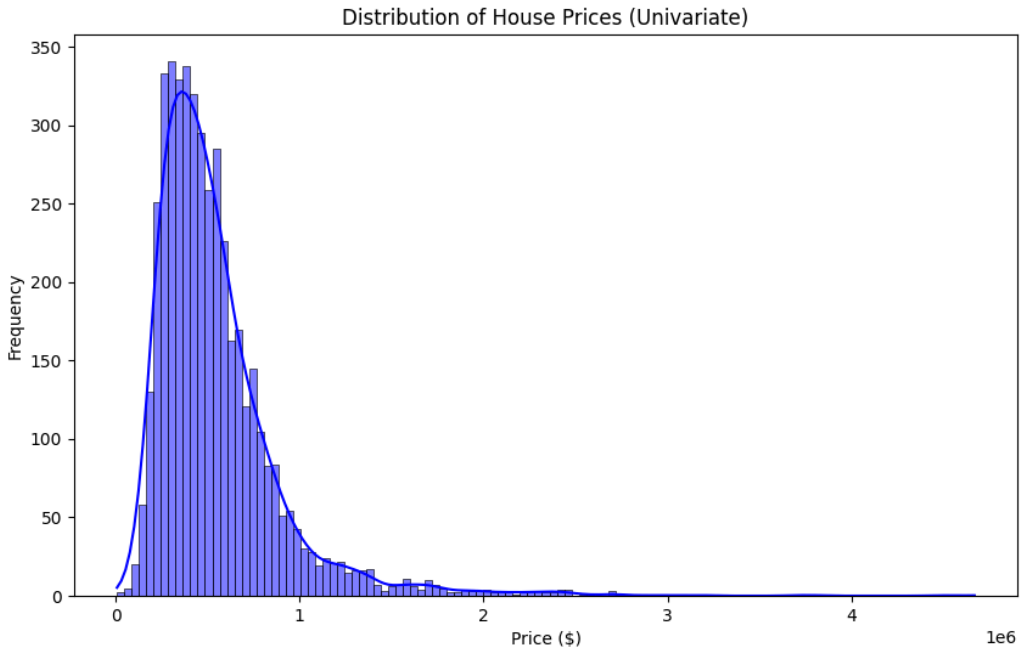
Descriptive Analysis We used descriptive statistics to summarize the key numbers behind the housing data. After cleaning out errors and removing extreme outliers, we found that the average price of a house is approximately 548,000 dollars. However, the median price is lower, at about 465,000 dollars. This difference is important because it tells us that a small number of expensive homes are pulling the average up, while the majority of homes are actually cheaper than the mean suggests. We also examined the standard deviation, which was quite large. This indicates that there is a wide variety of prices in the market, ranging from very affordable starter homes to expensive luxury properties .

Test Statistics and Hypothesis Testing We used correlation analysis as our primary statistical test to measure relationships between variables. By calculating a correlation matrix, we could see exactly which features move in the same direction as the price. This test statistically proved that the size of the living area is the most important factor. It showed a strong positive correlation score of 0.69, meaning that as square footage goes up, the price reliably goes up. We also tested other factors like the year the house was built. The correlation score for this was very low, at just 0.03. This proves statistically that the age of a house does not significantly affect its price in this market .

10. Exploratory Data Analysis (EDA) - Univariate Analysis

We started our visual analysis by looking at just one variable: the house price. This is called univariate analysis because we are examining a single column of data to understand its pattern. We created a histogram, which is a chart that shows how frequently different prices appear in the market.

The histogram reveals a clear "right-skewed" shape

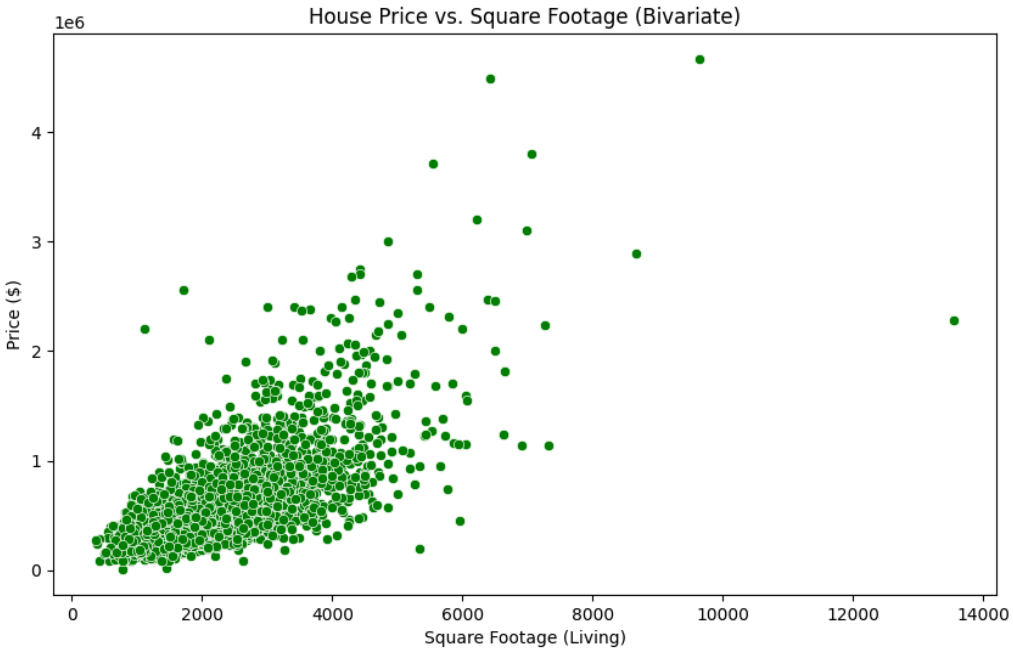


This means that the tall bars are on the left side, showing that most houses cost between 300,000 and 600,000 dollars. As we move to the right, the bars get shorter and longer, representing the fewer, more expensive luxury homes. Even though we removed the extreme multi-million dollar mansions, the data still shows that affordable homes are much more common than expensive ones. This shape tells us that using the "median" (middle price) is better than the "average" price to describe a typical home, because the expensive homes pull the average up too high.

11. Bivariate Analysis

Next, we looked at the relationship between two variables at the same time. We chose to compare the "Price" against the "Square Footage of Living Area" because our statistical test showed these two were strongly connected. We used a scatter plot to visualize this relationship.

The scatter plot shows a very clear pattern

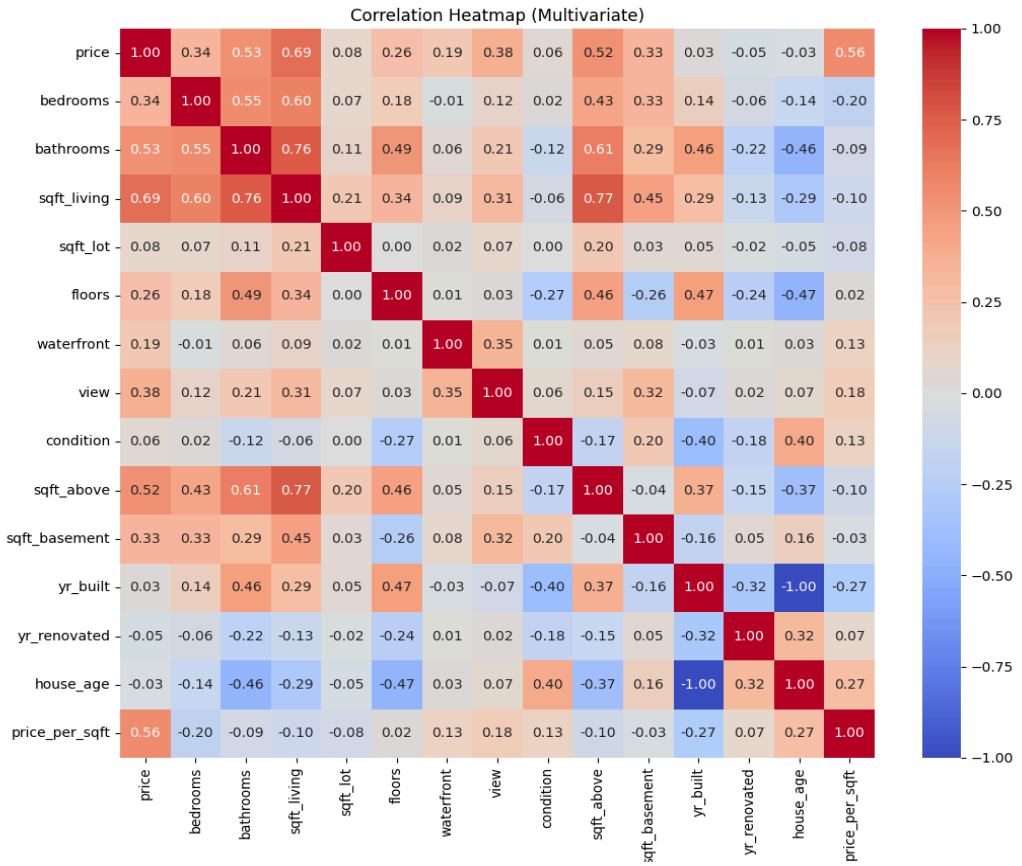


You can see that the dots form a cloud that moves upwards from left to right. This is a positive linear relationship. It visually proves that as the size of the house increases, the price consistently goes up. The dots are tightly packed together in the lower left, which confirms that smaller houses reliably sell for lower prices. As the houses get bigger (moving to the right), the dots spread out more, meaning the prices for huge mansions vary more than prices for small starter homes.

12. Multivariate Analysis

Finally, we looked at how all the numerical variables relate to each other simultaneously. We used a "Correlation Heatmap" for this. A heatmap is a colorful grid where red/warm colors mean a strong positive relationship (close to 1.0) and blue/cool colors mean a weak or negative relationship.

The heatmap gives us a complete picture of the housing market



We can see a bright red square where "Price" meets "Sqft_Living," confirming it is the strongest driver (0.69). We also see strong red colors connecting "Bathrooms" and "Sqft_Above" to price. Interestingly, the boxes for "Condition" and "Year Built" are very light blue or white, which means they have almost zero correlation with price. This visual tool helps us instantly see that size and amenities matter much more than age or condition for determining value.

13. Overall Insights from Analysis

Our comprehensive analysis of the housing dataset has revealed the most critical factors that determine a property's market value. By cleaning the data, calculating new metrics, and visualizing the trends, we have moved beyond simple guesses to find concrete, data-backed answers.

First, the most powerful insight is that **size is the primary driver of price**. Our statistical tests proved this with a strong positive correlation score of 0.69 between the living area square footage and the sale price. The scatter plot visualization confirmed this linear trend: as the size of the house increases, the price consistently rises. This tells us that for any buyer or investor, the total square footage is the single most reliable indicator of a home's value.

Second, we found that **amenities matter significantly**. The number of bathrooms showed a moderate positive correlation (0.53) with price, which was actually higher than the correlation for the number of bedrooms (0.34). This suggests that modern buyers may value the convenience and luxury of multiple bathrooms more than just having extra sleeping rooms. The correlation heatmap visually reinforced this, showing a strong link between larger homes and more bathrooms, creating a cluster of high-value features.

Third, we discovered a surprising insight about the age of the property. Contrary to the common belief that newer is always better, our analysis showed almost **zero correlation (0.03)** between the year built and the price. This implies that older homes do not necessarily lose value just because of their age. Factors like location, renovation quality, or the sheer size of the plot likely outweigh the age of the structure itself.

Finally, the distribution of prices revealed a **segmented market**. Most homes are priced affordably under one million dollars, but a long tail of high-value properties pulls the average up. This means the "average" price is not always the best measure for a typical buyer, and the median price gives a more realistic expectation. Overall, the data confirms that to maximize value, stakeholders should prioritize physical space and bathroom count above all else .

14. Conclusion

This project successfully analyzed the Housing Dataset to uncover the primary factors influencing property prices. Through systematic data cleaning and statistical testing, we confirmed that physical size is the most critical determinant of value. Specifically, the square footage of the living area has the strongest positive impact on price. Secondary features like the number of bathrooms also play a significant role, whereas the age of the property has surprisingly little effect. Based on these findings, we recommend that stakeholders prioritize square footage and bathroom count for accurate valuation. Future studies could further refine these predictions by incorporating neighborhood specific location data .