### Housing: Livable Communities - Public Policy Institute

### \*Hosuing data set \*

### Ragul v

DA/DS

B3MAY(ONLINE)

60524OLE002

# INTRODUCTION

## A housing dataset contains detailed information about real estate properties, typically including features such as property size, location, number of rooms, age, and sale prices. This dataset is crucial for analyzing market trends, predicting property values, and understanding factors influencing real estate prices. It often includes demographic data, historical sales, and economic indicators to provide a comprehensive view of the housing market. By leveraging this data, stakeholders can make informed decisions regarding property investments, pricing strategies, and market analysis, ultimately enhancing their ability to navigate and capitalize on real estate opportunities.

**AIM**

* The aim of a housing dataset is to provide comprehensive, accurate, and actionable information about real estate properties. It seeks to facilitate the analysis of market trends, property valuations, and factors influencing real estate prices. By integrating diverse data points, such as property features, geographic location, and historical sales, the dataset supports predictive modeling and trend analysis. This enables stakeholders—such as real estate agents, investors, and analysts—to make informed decisions, optimize pricing strategies, and identify investment opportunities. Ultimately, the dataset aims to enhance strategic planning and decision-making in the real estate market

**\*Bottom of FormBUSINESS PROBLEM/PROBLEM STATEMENT**

**In the real estate industry, accurate and comprehensive housing datasets are crucial for understanding market trends, pricing strategies, and investment opportunities. A common business problem arises when housing datasets are fragmented, outdated, or incomplete, making it difficult to perform reliable market analysis and forecasts. For example, a real estate firm might struggle with integrating data from various sources, leading to inconsistencies in property valuations, neighbourhood trends, and demographic insights. This issue hampers the ability to provide clients with accurate recommendations and can result in missed investment opportunities. The problem statement, therefore, is: “The current fragmentation and inconsistency in our housing dataset hinder effective market analysis and strategic decision-making, affecting our ability to provide accurate property valuations and investment advice. We need a unified, updated, and comprehensive dataset to enhance data-driven insights and support better business outcomes.”**

**\*Project Workflow**

**A housing dataflow project workflow outlines the systematic approach to managing, processing, and analyzing housing data from collection to actionable insights. Here’s a structured workflow:**

1. **Data Collection: Gather housing data from diverse sources, such as real estate listings, public records, and market reports. Ensure data is relevant, accurate, and comprehensive.**
2. **Data Integration: Combine data from various sources into a unified system or database. This step involves cleaning and harmonizing data formats to ensure consistency.**
3. **Data Cleaning and Transformation: Address missing values, outliers, and inconsistencies. Transform data into a usable format by normalizing, aggregating, and enriching it as needed.**
4. **Data Storage: Store the cleaned and transformed data in a secure and scalable storage solution, such as a relational database or a data warehouse.**
5. **Data Analysis: Perform exploratory data analysis (EDA) to identify trends, patterns, and insights. Use statistic methods and data visualization tools to understand market dynamics.**
6. **Model Building: Develop predictive models or algorithms to forecast housing trends, pricing, and other key metrics. Validate and refine models to improve accuracy.**
7. **Reporting and Visualization: Create dashboards and reports to present findings. Use visualization tools to convey insights clearly to stakeholders.**
8. **Decision Making: Utilize insights from analysis and models to make informed business decisions, such as pricing strategies, investment opportunities, or market expansion.**
9. **Monitoring and Feedback: Continuously monitor the performance of data models and workflows. Gather feedback from users to improve data quality and system efficiency**
10. **Iteration and Improvement: Regularly update data sources, refine models, and enhance processes based on new information and evolving business needs.**

**This workflow ensures a structured approach to handling housing data, leading to actionable insights and better decision-making.**

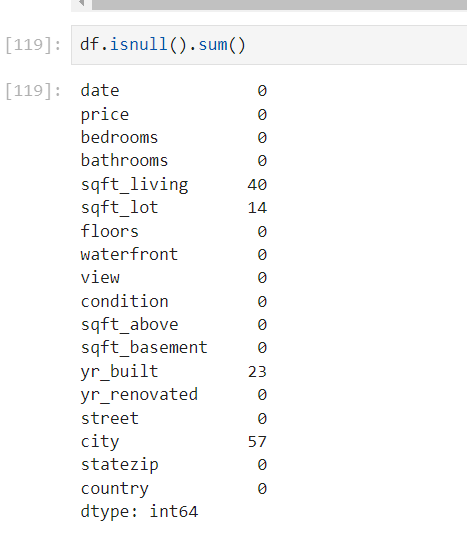
**\*Data Understanding**

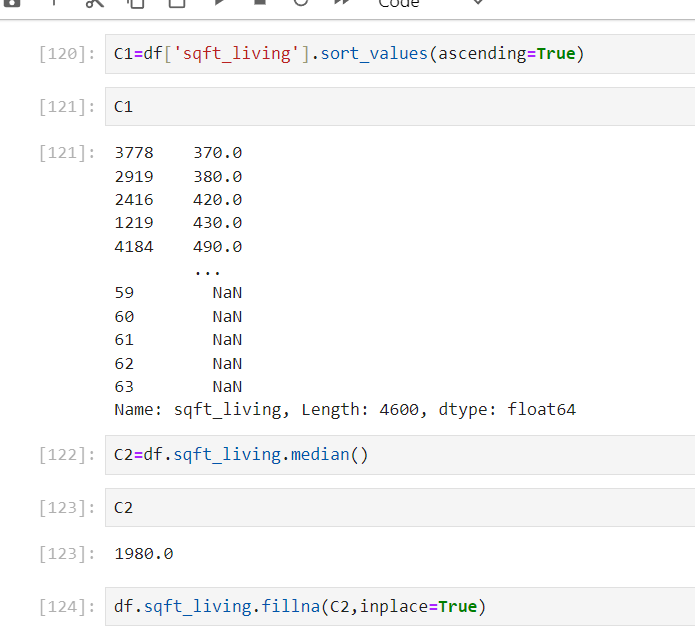
**Data Understanding is a critical phase in working with a housing dataset, focusing on comprehending the data’s nature, quality, and relevance to the project's objectives. This stage involves several key activities:**

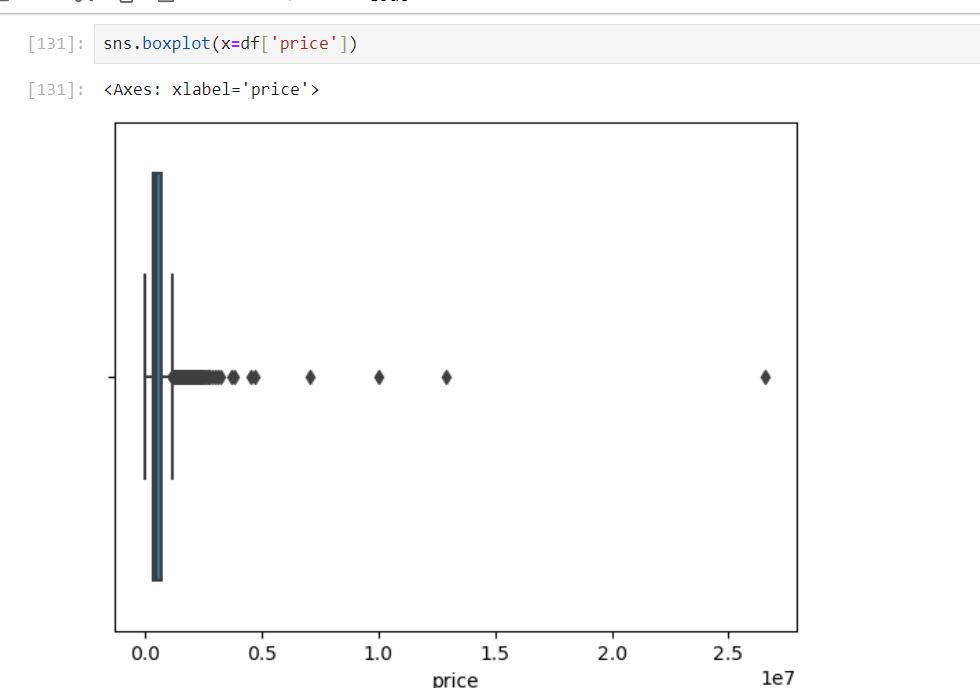
1. **Data Exploration: Begin by examining the dataset's structure, including the types of variables (e.g., numerical, categorical) and their distributions. Look at summary statistics, such as mean, median, and standard deviation, to understand the central tendencies and variability of the data.**
2. **Data Quality Assessment: Identify and assess issues related to data quality, such as missing values, duplicates, inconsistencies, and outliers. Determine the impact of these issues on analysis and plan remediation strategies.**
3. **Contextual Analysis: Understand the context in which the data was collected, including the geographic and temporal scope. For housing data, this might involve understanding local market conditions, property types, and economic factors influencing the data.**
4. **Feature Analysis: Investigate individual features to understand their relevance and relationships with other variables. For instance, examine how factors like location, size, and amenities correlate with property prices.**
5. **Data Distribution: Analyze the distribution of key variables to identify patterns and trends. Visualization tools such as histograms, scatter plots, and box plots can help reveal insights about the data’s distribution and relationships.**
6. **Stakeholder Input: Engage with stakeholders to ensure the data meets their needs and to clarify any additional requirements or objectives.**

**Understanding these aspects of the housing dataset provides a foundation for effective data processing, analysis, and ultimately, making informed business decisions**

**\*Missing value**

****

****



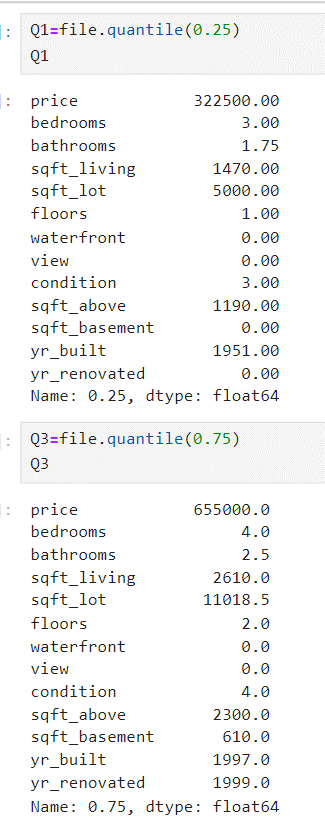
* **Bivariate analysis**

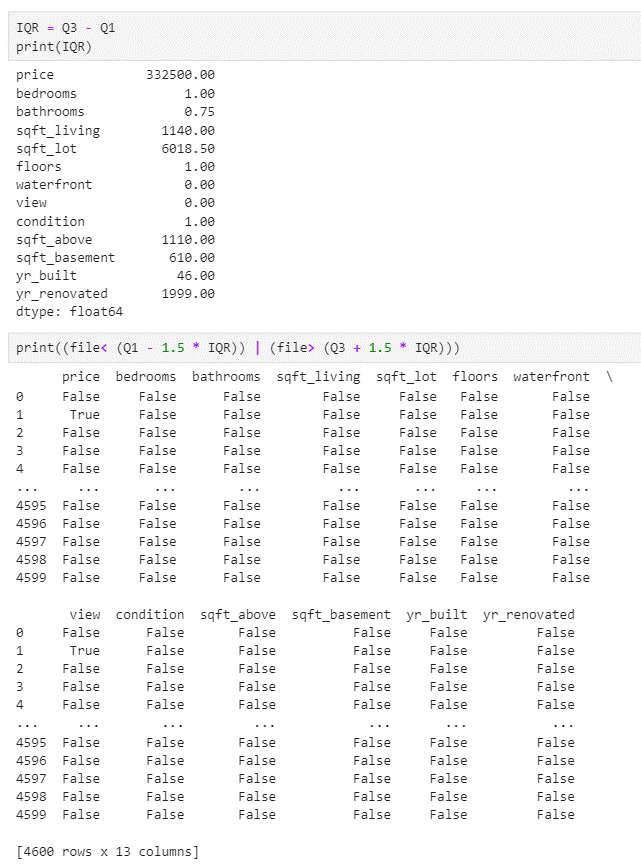
**Bivariate analysis is a statistical method used to analyze the relationship between two variables. It can help to identify and measure the strength and direction of the association between these variables. The most common techniques for bivariate analysis include:**

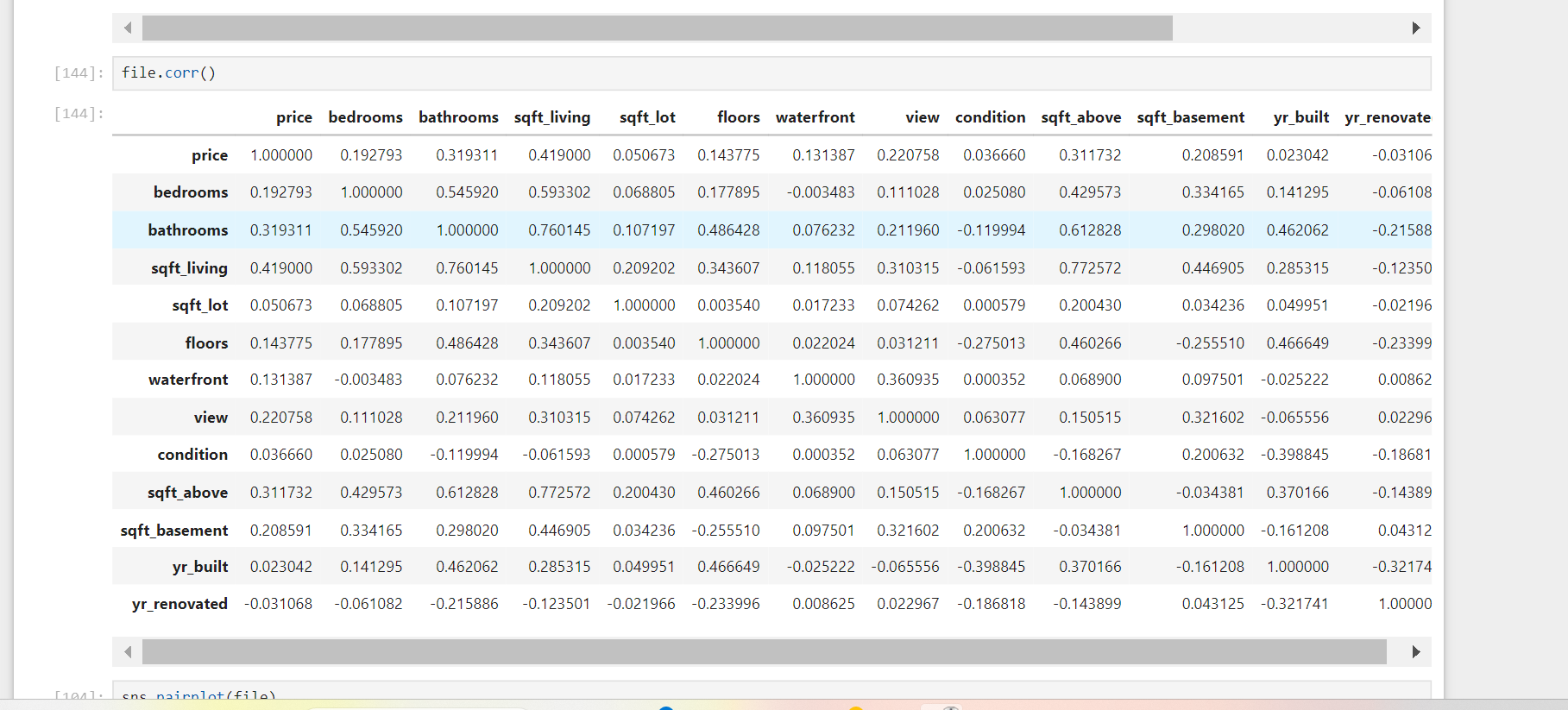
**Scatter Plot:** **A graphical representation where each point represents a pair of values. It helps visualize the relationship between two continuous variables**

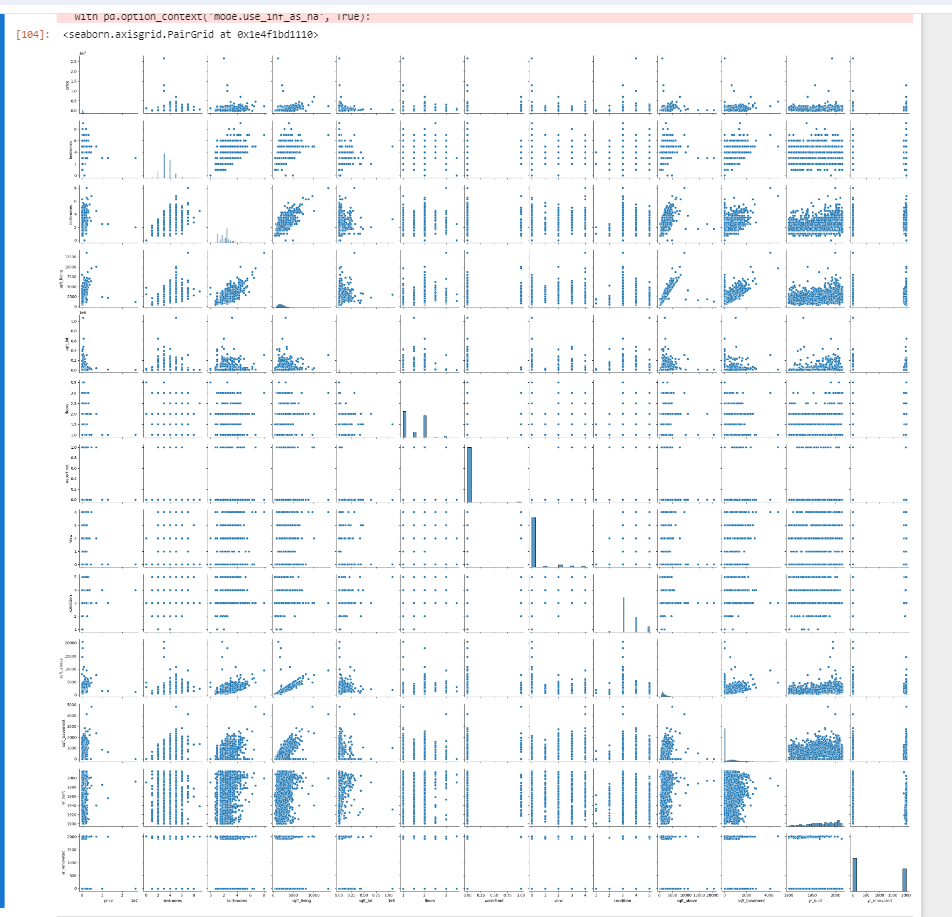
**Correlation analysis: is a statistical method used to evaluate the strength and direction of the relationship between two continuous variables.**

**A box plot : (or box-and-whisker plot) is a standardized way of displaying the distribution of data based on a five-number summary: minimum, first quartile (Q1), median, third quartile (Q3), and maximum. It helps to visualize the spread and skewness in the data, and identify potential outliers.**









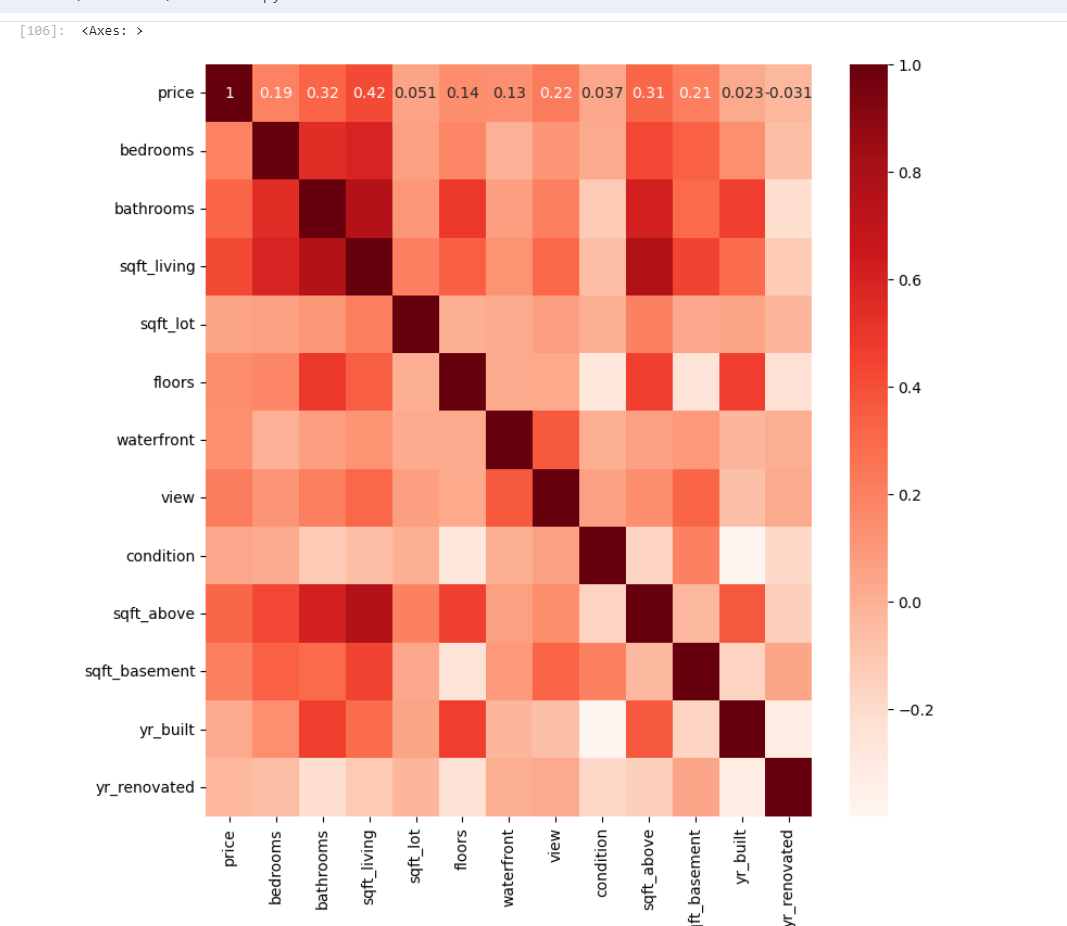
Exploratory Data Analysis (EDA) is a critical process in data analysis that involves summarizing the main characteristics of the data, often with visual methods. Here are several key insights typically obtained

A heatmap : is a graphical representation of data where individual values contained in a matrix are represented as colors. Heatmaps are commonly used to visualize the magnitude of data values across two dimensions, making it easy to see patterns, correlations, and outliers.

**Key Uses of Heatmaps**

1. **Correlation Matrix**: Visualize the correlation between multiple variables.
2. **Confusion Matrix**: In classification tasks, visualize the performance of a model.
3. **Missing Data**: Identify patterns in missing data.
4. **Data Distribution**: Compare the distribution of data across different categories.

\*HEATMAP



* Conclusion

The exploratory data analysis (EDA) of the housing dataset provides valuable insights into the factors affecting house prices. The key determinants include the size of the house, the number of rooms, and the location. Outliers in the data highlight exceptional cases that may require further investigation. Understanding these patterns can inform better decision-making for buyers, sellers, and policymakers in the housing market