

Exploratory Data Analysis of Housing Dataset

Manjula M | Data Analytics | B1 ( E )|Online Mode

## Table of Contents

* Introduction
* Aim
* Business Problem / Problem Statement
* Project Workflow
* Data Understanding
* Data Cleaning
* Obtaining Derived Metrics
* Filtering Data for Analysis
* EDA - Univariate Analysis
* Segmented Univariate Analysis
* Bivariate Analysis
* Multivariate Analysis
* Overall Insights Obtained from Analysis
* Conclusion

### Introduction

This project introduction to analyse a housing dataset to identify key factors influencing housing prices and to handle data anomalies such as outliers and missing values. The dataset contains various attributes of houses, including price, number of bedrooms, bathrooms, square footage, lot size, and other relevant features. The objectives include detecting and handling outliers, imputing missing values, performing exploratory data analysis (EDA), and deriving insights through statistical tests and visualizations. This analysis will help in understanding the primary drivers of housing prices and provide a basis for further predictive modelling and decision-making in real estate.

### Aim

The aim of this project is to comprehensively analyze a housing dataset to uncover the primary factors that influence housing prices. This involves detecting and handling outliers, imputing missing values, and performing exploratory data analysis (EDA) to identify significant patterns and correlations. Additionally, the project seeks to create new features, such as property age and price per square foot, to enhance the analysis. Ultimately, the goal is to derive actionable insights that can inform real estate decision-making and lay the groundwork for building robust predictive models to forecast housing prices.

### Business Problem / Problem Statement

The real estate market is highly dynamic, with numerous factors influencing property prices, making it challenging for buyers, sellers, and investors to make informed decisions. The primary business problem addressed by this project is the need for a systematic analysis to identify the key determinants of housing prices. This is crucial for various stakeholders, including real estate agents, developers, and financial institutions, to accurately assess property values, forecast market trends, and optimize investment strategies.

This project focuses on a dataset containing multiple features of residential properties, such as price, number of bedrooms, bathrooms, square footage, and lot size. By detecting and handling data anomalies, performing exploratory data analysis (EDA), and conducting statistical tests, the project aims to uncover significant patterns and relationships among these features. The insights derived from this analysis will help stakeholders make data-driven decisions, enhance market strategies, and potentially increase profitability in the real estate sector.

### Project Workflow

The project follows a systematic workflow to achieve its objectives.

1. **Data Collection and Loading:** The first step involves loading the housing dataset using pandas, ensuring the data is accessible for analysis.
2. **Data Exploration:** Initial exploration includes examining the dataset’s structure, summary statistics, and identifying potential issues such as missing values or outliers.
3. **Data Cleaning:** This step involves handling missing values by imputing numerical columns with their means and categorical columns with their modes. Outliers are detected using the Interquartile Range (IQR) method and replaced with the mode.
4. **Exploratory Data Analysis (EDA):** Univariate, bivariate, and multivariate analyses are conducted to visualize distributions, relationships, and correlations among variables using plots and correlation matrices.
5. **Feature Engineering:** New features, such as property age and price per square foot, are created to enhance the dataset.
6. **Statistical Testing:** ANOVA (F-test) is performed to identify significant features impacting housing prices.
7. **Insights and Recommendations:** The final step involves interpreting the results, identifying key drivers of housing prices, and providing actionable insights for stakeholders.

This structured methodology ensures a thorough and insightful analysis of the housing market.

### Data Understanding

**Description of the Dataset:** The dataset comprises various features of residential properties, aiming to facilitate an in-depth analysis of housing prices. Key attributes include:

* **Price:** The sale price of the house.
* **Bedrooms:** Number of bedrooms.
* **Bathrooms:** Number of bathrooms.
* **Sqft\_living:** Square footage of the living space.
* **Sqft\_lot:** Size of the lot.
* **Floors:** Number of floors.
* **Waterfront:** Indicator if the property has a waterfront view.
* **View:** Quality of the view.
* **Condition:** Condition of the house.
* **Sqft\_above:** Square footage of the house apart from basement.
* **Sqft\_basement:** Square footage of the basement.
* **Yr\_built:** Year the house was built.
* **Yr\_renovated:** Year the house was renovated.
* **City:** Location of the property.

The dataset has 21 columns with mixed data types, including numerical and categorical features, and consists of several thousand rows, representing individual housing records.

**Summary Statistics and Initial Insights:** Initial data exploration reveals:

* The price feature shows a wide range, indicating diverse property values.
* Numerical features like sqft\_living, sqft\_lot, and bedrooms exhibit variability, suggesting different property sizes and configurations.
* Categorical features like condition, waterfront, and view show variations, which could impact housing prices.

Summary statistics provide insights into central tendencies and distributions:

* **Mean and Median:** Offer central values for features like price and square footage.
* **Standard Deviation:** Indicates variability, crucial for identifying outliers.
* **Min and Max:** Highlight the range of values, useful for understanding the data spread.

**Initial Insights:**

* Outliers are present in features like price and sqft\_lot, necessitating careful handling.
* Missing values are minimal but require imputation to ensure data completeness.
* Significant correlations are expected between price and features like sqft\_living, bathrooms, and bedrooms.

These insights guide the data cleaning, feature engineering, and further analysis steps, ensuring a robust approach to understanding housing market dynamics.

### Data Cleaning

**Missing Values Imputation:** The dataset contains some missing values, which need to be addressed to ensure robust analysis. The following steps were taken:

* **Numerical Columns:** Missing values in numerical columns were imputed using the mean value of each column. This approach maintains the overall distribution and avoids skewing the data.

**PROGRAM:**

numerical\_cols = data.select\_dtypes(include=['number']).columns

data[numerical\_cols] = data[numerical\_cols].fillna(data[numerical\_cols].mean())

**Categorical Columns:** Missing values in categorical columns were imputed using the mode, which is the most frequently occurring value in each column. This method is effective in preserving the categorical distribution.

**PROGRAM:**

categorical\_cols = data.select\_dtypes(include=['object']).columns

data[categorical\_cols]=

data[categorical\_cols].fillna(data[categorical\_cols].mode().iloc[0])

**Outlier Detection and Handling:** Outliers can distort analysis and models. The Interquartile Range (IQR) method was used to detect and handle outliers:

* **Detection:** For each numerical column, the first quartile (Q1) and third quartile (Q3) were calculated. The IQR was then used to determine the lower and upper bounds beyond which values are considered outliers.

**PROGRAM:**

def detect\_outliers\_iqr(df, column):

Q1 = df[column].quantile(0.25)

Q3 = df[column].quantile(0.75)

IQR = Q3 - Q1

lower\_bound = Q1 - 1.5 \* IQR

upper\_bound = Q3 + 1.5 \* IQR

return df[(df[column] < lower\_bound) | (df[column] > upper\_bound)], lower\_bound, upper\_bound

outliers\_dict = {}

bounds\_dict = {}

for col in numerical\_cols:

outliers, lower\_bound, upper\_bound = detect\_outliers\_iqr(data, col)

outliers\_dict[col] = outliers

bounds\_dict[col] = (lower\_bound, upper\_bound)

**Handling:** Detected outliers were replaced with the mode of the respective column, as this approach retains the column's distribution while mitigating the impact of extreme values.

**PROGRAM:**

def replace\_outliers\_with\_mode(df, column, lower\_bound, upper\_bound):

mode\_value = df[column].mode()[0]

df[column] = df[column].apply(lambda x: mode\_value if x < lower\_bound or x > upper\_bound else x)

for col in numerical\_cols:

lower\_bound, upper\_bound = bounds\_dict[col]

replace\_outliers\_with\_mode(data, col, lower\_bound, upper\_bound)

**Handling Inconsistent Values:** Consistency in data values is crucial for accurate analysis. The following steps were taken:

* **Categorical Data:** Ensured categorical data values were consistent, correcting any typos or variations in categorical entries.
* **Numerical Data:** Checked numerical columns for any unlikely values (e.g., negative square footage) and corrected or removed them.

Through these data cleaning steps, the dataset was prepared for robust and reliable analysis, enhancing the accuracy of subsequent insights and models.

### Obtaining Derived Metrics

To enrich the dataset and provide deeper insights into the factors influencing housing prices, several derived metrics were created:

**1. Age of Property:**

This metric represents the age of the property in years and is calculated by subtracting the year the house was built (`yr\_built`) from the current year (2024). It helps in understanding how the age of a property impacts its price.

**Python Program:**

data['age'] = 2024 - data['yr\_built']

**2. Price per Square Foot:**

This metric provides a measure of the value of the living space by dividing the house price (`price`) by the square footage of the living area (`sqft\_living`). It allows for a normalized comparison of property values regardless of size.

**Python Program:**

data['price\_per\_sqft'] = data['price'] / data['sqft\_living']

These derived metrics enhance the dataset by introducing new dimensions for analysis. The age of the property can reveal trends related to the depreciation or appreciation of older homes. The price per square foot helps in comparing properties of different sizes on a common scale, making it easier to identify overvalued or undervalued properties.

By incorporating these derived metrics, the analysis can more effectively identify key drivers of housing prices and provide actionable insights for real estate decision-making.

### Filtering Data for Analysis

In addition to the data cleaning steps previously outlined, a few additional filtering and preprocessing steps were performed to prepare the data for analysis:

* **Removing Irrelevant Columns:** Any columns that were deemed irrelevant for the analysis, such as unique identifiers or redundant information, were removed to streamline the dataset.
* **Handling Categorical Variables:** Categorical variables were encoded using techniques like one-hot encoding or label encoding to convert them into numerical format, enabling their inclusion in analytical models.
* **Handling Sparse Categories:** Sparse categories in categorical variables were consolidated or removed to improve model performance and interpretability.

These steps ensure that the dataset is optimized for meaningful analysis while minimizing noise and irrelevant information.

### EDA - Univariate Analysis

Univariate analysis focuses on exploring individual variables in the dataset to understand their distributions and characteristics. Here are some insights gained from univariate analysis, along with visualizations of key variables:

1. **Price Distribution:** The distribution of house prices reveals the range and central tendency of property values. A histogram or kernel density plot can visually represent this distribution, showing whether prices are skewed or normally distributed.

2. **Square Footage:** Analyzing the square footage of living spaces (`sqft\_living`) and lots (`sqft\_lot`) provides insights into the size distribution of properties. Histograms or kernel density plots can illustrate the distribution of square footage, highlighting common property sizes and any outliers.

3. **Number of Bedrooms and Bathrooms**: Histograms or bar plots can showcase the distribution of bedrooms and bathrooms in the dataset. Understanding the frequency of different configurations (e.g., 2-bedroom vs. 3-bedroom homes) helps in assessing market demand and preferences.

4. **Property Condition:** Exploring the distribution of property condition ratings (e.g., on a scale of 1 to 5) provides insights into the overall condition of properties in the dataset. A bar plot or countplot can visualize the frequency of each condition rating.

5. **Year Built:** A histogram or line plot can display the distribution of property construction years (`yr\_built`). Understanding the age distribution of properties helps in identifying trends related to older or newer homes.

These visualizations and insights from univariate analysis lay the foundation for further exploration and modeling, providing valuable context for understanding the dataset's characteristics and distributions.

### Segmented Univariate Analysis

Segmented univariate analysis involves analyzing data segments or categories within a single variable to gain deeper insights into their individual characteristics. For example, segmenting the dataset based on geographical regions or property types allows for a more granular examination of each segment's distribution and characteristics. This approach helps in identifying variations and patterns specific to each segment, which may not be apparent in the overall dataset. Visualizations such as histograms, box plots, or bar plots can be used to compare the distributions of different segments, revealing any disparities or trends. Segmenting the data based on relevant factors such as property condition, waterfront status, or city location enables a more nuanced understanding of how these variables influence housing prices and market dynamics.

### Bivariate Analysis

### Bivariate analysis explores relationships between pairs of variables in the dataset to uncover patterns, correlations, and dependencies. Here are some key aspects of bivariate analysis:

**1.** **Price vs. Square Footage:** A scatter plot can illustrate the relationship between house prices and square footage of living space (`sqft\_living`). This analysis helps in understanding how property prices vary with size, revealing potential trends such as larger homes commanding higher prices.

**2. Price vs. Bedrooms/Bathrooms**: Box plots or scatter plots can depict how house prices vary with the number of bedrooms or bathrooms. This analysis provides insights into the impact of these key features on property values, helping in assessing market demand and preferences.

**3. Price vs. Location (City):** Box plots or bar plots can compare house prices across different cities or regions. Understanding how prices vary geographically helps in identifying areas with higher or lower property values, informing investment decisions.

4**. Price vs. Condition:** Box plots or bar plots can show how property prices vary with condition ratings. This analysis reveals the relationship between property condition and market value, guiding renovation or improvement strategies.

Bivariate analysis provides valuable insights into the relationships between variables, guiding further exploration and modeling to understand the factors influencing housing prices.

### Multivariate Analysis

Multivariate analysis delves into the complex relationships involving multiple variables simultaneously, offering deeper insights into the interplay of various factors influencing housing prices. Here are some aspects of multivariate analysis:

1. **Pairwise Relationships:** Investigating how multiple variables interact with each other, such as examining the combined effects of square footage, number of bedrooms, and location on house prices. Techniques like scatter plots or pairplots can visualize these relationships comprehensively.
2. **Correlation Analysis:** Assessing the correlations among multiple variables using techniques like correlation matrices or heatmap visualizations. This analysis helps in identifying which features are strongly correlated with each other and with house prices, aiding feature selection and model building.
3. **Regression Analysis:** Employing regression models to predict house prices based on multiple predictor variables. Techniques like multiple linear regression or polynomial regression can uncover the relative importance of different features in determining property values.
4. **Cluster Analysis:** Identifying groups or clusters of properties with similar characteristics using clustering algorithms like k-means clustering. This analysis helps in segmenting the dataset based on shared attributes, providing insights into market segments and pricing dynamics.

Multivariate analysis enables a comprehensive understanding of the intricate relationships among multiple variables, facilitating informed decision-making and predictive modeling in the real estate domain.

### Overall Insights Obtained from Analysis

The analysis of the housing dataset yielded several key insights and findings that shed light on the factors influencing housing prices and market dynamics:

1. **Primary Drivers of Housing Prices:** Square footage (sqft\_living), number of bedrooms and bathrooms, and property location emerged as significant drivers of housing prices. Larger homes with more bedrooms and bathrooms tend to command higher prices, while properties located in desirable areas exhibit higher values.
2. **Geographical Variations:** Housing prices vary significantly across different cities or regions. Certain locations, such as urban centers or waterfront areas, tend to have higher property values compared to suburban or rural areas. Understanding these geographical variations is crucial for making informed investment decisions.
3. **Impact of Property Condition:** Property condition plays a vital role in determining housing prices. Well-maintained properties with higher condition ratings generally fetch higher prices in the market. Renovation or improvement strategies aimed at enhancing property condition can positively impact resale values.
4. **Age and Renovation Effects:** The age of the property (yr\_built) influences its market value, with newer properties typically commanding higher prices. Additionally, renovated properties tend to have higher resale values, highlighting the importance of updating older homes to attract buyers.
5. **Square Footage and Price per Square Foot:** Square footage of the living space (sqft\_living) significantly impacts housing prices, with larger homes generally priced higher. The analysis also revealed insights into the price per square foot, providing a standardized measure for comparing property values regardless of size.
6. **Market Segmentation:** Segmenting the dataset based on various factors such as property type, location, or condition revealed distinct market segments with unique pricing dynamics. Understanding these segments enables targeted marketing strategies and tailored pricing approaches.

Overall, the analysis provides valuable insights into the complex interplay of factors shaping the housing market. These findings can inform various stakeholders, including buyers, sellers, developers, and investors, in making informed decisions and strategies in the real estate domain.

### Conclusion

In conclusion, the analysis of the housing dataset has provided valuable insights into the factors influencing housing prices and market dynamics. Key drivers such as square footage, property location, condition, and age have been identified, highlighting the importance of these factors in determining property values. Understanding geographical variations and market segmentation can guide investment decisions and marketing strategies.

### Recommendations

Moving forward, further analysis could focus on:

* Advanced predictive modeling to forecast housing prices.
* Incorporating additional data sources such as economic indicators or demographic trends for more comprehensive analysis.
* Implementing geographic information systems (GIS) to analyze spatial patterns and trends in property values.
* Conducting customer segmentation analysis to tailor marketing strategies to specific buyer demographics.

These recommendations will enhance decision-making processes and optimize outcomes in the dynamic real estate market.

**\*-\*-\*-\*-\*-\***