PROJECT DOCUMENTATION

Exploratory Data Analysis

Housing Dataset

Name : Mohan Prabhu D

Branch : Data Science

Batch Number : 1

Mode of Study : Offline

Roll Number : 150324CBR003

Table of Contents

1.Introduction

The purpose of this project is to perform an Exploratory Data Analysis (EDA) on housing Dataset to uncover underlying patterns, relationships, and insights. A housing dataset typically contains information about their features such as Location Details that contains Address, City, Zip Code, and various properties, including Bedrooms, Bathrooms, Square Footage, Lot Size, Year Built. Such datasets are widely used in real estate analysis, housing market studies, and machine learning applications for predictive modelling.

2.Aim

The goal of this project is to conduct a comprehensive analysis of the dataset to derive insights and it can be utilized to address various business problems in the real estate industry such as Property Valuation, Market Analysis and Trends, Investment Decision-Making, Customer Segmentation, demand forecasting, Pricing Strategy etc.

3. Business Problem / Problem Statement

Accurate property valuation is crucial for buyers, sellers, real estate agents, and investors. Traditional methods of property valuation can be time-consuming and subjective, while the buyers need to know the Location Details such as Address, City, Zip Code, and various properties of the house, including Bedrooms, Bathrooms, Square Footage, Lot Size, Year Built.

4. Project Workflow

The project workflow includes the following steps:

1. Data Collection
2. Data Cleaning and Preprocessing
3. Exploratory Data Analysis
4. Data Visualization
5. Deriving Insights
6. Reporting

5. Data Understanding

The dataset contains of 4600 rows and 18 columns.

The key variables include

- Price
- Bedrooms
- Bathrooms
- Sqft_living
- Sqft_lot
- Floors
- Condition
- Year Built
- City

Data Types

- Object
- Float
- Int

6.Data Cleaning

Data cleaning involved:

- Dropping null values:

  Since there are more null values rows present. So, dropping it.

- Dropping unwanted columns;

  There is a column time and country which is not necessary for the analysis. So, removing it.

- Imputing missing values using mode method:

  There are a greater number of missing values in columns Sqft_living, Sqft_lot, year built and city, imputing it using the mean and mode method.

- Replacing the inconsistent address:

There is inconsistent address in the column street. So, removing the inconsistent address and regularizing the spellings in the address.

- Fixing the invalid values:

There are invalid float values in the column floor and bathrooms. So, regularizing the values of floors and bathrooms using ceil method.

- Standardizing the values:

The values of waterfront, view and condition are replaced with interactive inputs of scaling for the easy understanding of the client.

- Identifying and treating outliers using quantile method:

Treating the outliers in the column's bedrooms, bathrooms, sqft_living, sqft_lot, sqft_above, sqft_basement and price using the quantile method.

7. Obtaining derived metrics

Extracted three new columns such as "category_of_house" based on sqft_living and "type_of_house" based on bedrooms and finally "age_of_house" from the column year_built to get the overall insight.

bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	sqft_above	sqft_basement	yr_built	city	category_of_house	type_of_house	age_of_house
2	1340.0	14856.81	2	Not Available	Fair View	Very Good Condition	1340	0	1955.0	Shoreline	Compact House	3BHK	69
2	1930.0	14856.81	1	Not Available	Fair View	Excellent Condition	1930	0	1966.0	Kent	Compact House	3BHK	58
3	2000.0	14856.81	1	Not Available	Fair View	Excellent Condition	1000	1000	1963.0	Bellevue	Compact House	3BHK	61
3	1940.0	14856.81	1	Not Available	Fair View	Excellent Condition	1140	800	1976.0	Redmond	Compact House	4BHK	48
1	880.0	14856.81	1	Not Available	Fair View	Very Good Condition	880	0	1938.0	Seattle	Tiny House	2BHK	86
...	...	...	...	...	...	...	...	...	...	...	...	...	...
2	1510.0	6360.00	1	Not Available	Fair View	Excellent Condition	1510	0	1970.0	Seattle	Compact House	3BHK	54

8. Filtering data for analysis

Filtering data for analysis involves selecting a subset of your dataset that meets specific criteria relevant to your analysis goals. Here we are filtering the required columns for the analysis. The most required columns of analysis is "price", "bedrooms", "bathrooms", "sqft_living", "sqft_lot", "floors", "waterfront", "view", "condition", "sqft_above", "sqft_basement", "yr_built", "city".

9. EDA - Univariate Analysis

Univariate Analysis revealed:

1.House Price:

From the skewness value, it is confirmed that it is a right skewed distribution and it is longer on the right side of its peak than on its left side. It is also called as positive skew. It indicates that there are observations at one end of the extreme ends of the distribution, but they are relatively infrequent. A right skewed distribution has a long tail on its right side.

From the above kurtosis value, it is found that we have obtained a leptokurtic distribution. Leptokurtic distributions are more kurtotic than a normal distribution. In Leptokurtic distributions there is a greater tendency for outliers.

2. Bedrooms of house:

Most of the data points (middle 50%) are concentrated between 3 and 4 bedrooms.

The distribution is relatively symmetrical around the median.

The plot suggests that most properties have between 2 and 5 bedrooms, there are some properties with only 1 bedroom and some with up to 6 bedrooms.

3. Bathrooms of houses:

Most of the data points (middle 50%) are concentrated between 1.5 and 3 bathrooms.

The distribution shows a slight positive skew, with a few properties having significantly more bathrooms than the majority.

4. Condition of houses:

This plot shows the conditions of number of houses.

Maximum of the houses are in very good condition.

only few houses are in average and fair condition.

10. Segmented Univariate Analysis

Segmented analysis showed:

1.Housing Price analysis:

After analyzing the housing prices,

Inference:

The skewness value is 1.44 and the kurtosis value is 2.97

Normal distributions have a kurtosis of 3, so any distribution with a kurtosis of approximately 3 is mesokurtic.

The kurtosis value is nearly 3, so it is a mesokurtic distribution and is medium-tailed, so outliers are neither highly frequent, nor highly infrequent.

The shape of the histogram suggests a right-skewed distribution, where most of the data points cluster towards the lower end of the price range, and fewer data points are present at the higher end.

The peak of the histogram is around the 250,000 to 500,000 (0.25e6 to 0.5e6) price range.

11. Bivariate Analysis

Bivariate analysis included:

1.'Floors' vs 'condition':

Comparing the floors and condition of houses.

Inference:

The chi-square test is used to analyse the relation between floors, conditions of the houses in the city.

It says that the floors are more correlated with the city.

It shows that the houses with single floors are well maintained and good in condition.

When the floors of the house increase the condition is affected

2.'year_built' vs 'Price':

Comparing the prices and year built of the houses,

Inference:

The anova test is used to analyse the relation between price,year built and condition of the houses in the city.

It says that the year built is more correlated with the condition.

12. Multivariate Analysis

Multivariate analysis revealed:

1.Analysis to determine the condition of the house:

Comparing the relationship between the features to determine the condition of the house.

Inference:

The test indicates that the features bathrooms and floors determine the condition of the houses.

2.Analysis to determine the view of the house:

Comparing the relationship between the features to determine the view of the house.

Inference:

The test indicates that the price of house determines the view of the houses.

3.Analysis to determine the waterfront of the house:

Comparing the relationship between the features to determine the waterfront of the house.

Inference:

The test indicates that the price of house determines whether the house has waterfront or not.

13. Overall Insights Obtained from Analysis

Based on the inferences drawn from your data exploration and visualization, here's a detailed analysis and interpretation of the findings, along with recommendations for the housing data:

Analysis and Interpretation

Key Factors Influencing House Pricing:

1. Number of bedrooms:

- Houses with bedrooms 2 to 5 are more popular and clients like to have at least 2 or more number of bedrooms in their house.

2. Number of bathrooms:

- Houses with bedrooms 2 to 3 are more popular and clients like to have at least 2 or more number of bathrooms in their house depending on the number of bedrooms.

3. Condition of houses:

- Clients likes the house to be in well maintained good condition. Houses with one floor are well maintained.

4. Distribution of houses in city:

- The data shows that the distribution of houses in the city is based on the number of floors.

Recommendations:

1. Number of bedrooms:

- In future there will be no more requirement of more bedrooms in the houses. 4 or 5 number of bedrooms will be sufficient. If there are a greater number of bedrooms then price will be high so it will not attract the clients.

2. Condition of the house:

- Consider to renovate the house at certain period of time and always ensure that houses are well maintained and they are in ready to occupy condition. Then it will attract more number of clients.

13. Conclusion

In conclusion, a multifaceted analysis of the housing dataset underscores the intricate interplay of various factors driving property prices. While location remains paramount, factors such as property size, condition, market trends, amenities, and economic indicators collectively shape the pricing landscape. By understanding these dynamics, stakeholders can make informed decisions regarding property investments, pricing strategies, and market positioning.