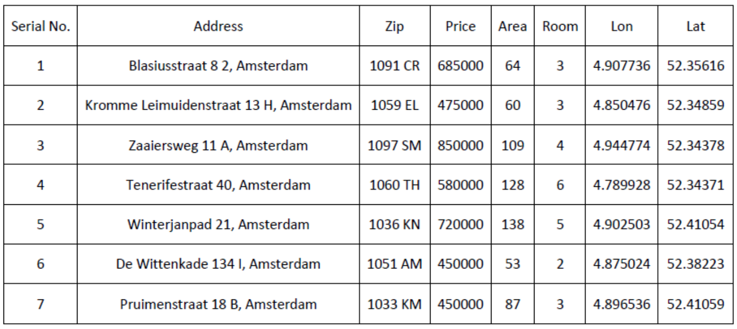
EDA of Amsterdam Housing Prices Dataset

Aniket Santra, MSc Data Science-1st Year, Chennai Mathematical Institute

Abstract

Amsterdam is the capital and most populous city of Netherlands. It is an urban city known for its artistic heritage, narrow houses with gabled facades, legacies of the city's 17th-century Golden Age. We have the dataset of Amsterdam Housing Prices. In this dataset we have unique address, Zip code, Area, Price of 924 houses of Amsterdam. We also have no. of rooms, the latitude and longitude of the locations of each house. In this project it has been attempted to study and analyse the several components of the dataset. We did EDA or exploratory data analysis of this dataset by several statistical graph. We have analysed each component of the dataset and found some statistical characteristics of the components. We have also checked any dependency in between the component by bivariate and multivariate graphical analysis. To visualize and analyse the dataset, the dataset's component, to construct the graphs R programming and R-Shiny Dashboard are used. This project helps to study the several characteristics of the houses in Amsterdam city.

Dataset



Serial No.	Address	Zip	Price	Area	Room	Lon	Lat
1	Blasiusstraat 8 2, Amsterdam	1091 CR	685000	64	3	4.907736	52.35616
2	Kromme Leimuidenstraat 13 H, Amsterdam	1059 EL	475000	60	3	4.850476	52.34859
3	Zaaiersweg 11 A, Amsterdam	1097 SM	850000	109	4	4.944774	52.34378
4	Tenerifestraat 40, Amsterdam	1060 TH	580000	128	6	4.789928	52.34371
5	Winterjanpad 21, Amsterdam	1036 KN	720000	138	5	4.902503	52.41054
6	De Wittenkade 134 I, Amsterdam	1051 AM	450000	53	2	4.875024	52.38223
7	Pruimenstraat 18 B, Amsterdam	1033 KM	450000	87	3	4.896536	52.41059

Format of the dataset

The dataset has total 924 entries i.e. 924 rows and each row has total 8 components i.e. no of column in our dataset is 8. Among the 8 components three components consist of categorical data and other five components consist of quantitative data:-

Serial No. - The first component Serial No. consists of integers starting from 1 which are to mark the all houses serially uniquely. This is a categorical variable of nominal type.

Address - The address component stores the unique addresses of each of the houses and it is also a categorical variable of nominal type.

Zip - This variable or component stores the Zip codes of the respective houses' location. The type of the variable Zip is categorical of nominal type.

Price - This is the price variable stores the prices of each of the houses in Euros. The Price variable is a continuous variable.

Area – This variable stores the areas of each of the houses in the unit square metre. The Area variable is a continuous variable.

Room - The room variable is a discrete variable stores the no of rooms in each respective house.

Lon - This variable stores the numeric part of longitude of the location of each houses and it's a continuous variable.

Lat - Like Lon variable it's also a continuous variable stores the numeric part of latitude of the location of each house.

Methodology

Data Cleaning

In the data cleaning process, it was found that four rows in our dataset have null values. Those four rows were omitted and we continued our further analysis with the remaining 920 rows.

Analysis by Diagrammatic Representation

Diagrams like graphs, charts, maps, pictures etc. are attractive and effective means for presentation of statistical data. It is more effective than tabular representation, being easily intelligible to a layman. Indeed, diagrams are almost essential whenever it is required to convey any statistical information to the general public. Diagrams are readily capable of revealing some features of the exhibited data.

Univariate Plots:-

Five-number Summary & Boxplot

The five-number summary is a set of descriptive statistics that provides information about a dataset. It consists of the five most important sample percentiles:- the sample minimum (smallest observation), the lower quartile or first quartile, the median (the middle value), the upper quartile or third quartile, the sample maximum (largest observation).

A boxplot or box plot is a standardized way of displaying the dataset based on the five-number summary. In descriptive statistics it is a method for graphically demonstrating the locality, spread and skewness groups of numerical data through their quartiles. In addition to the box on a boxplot, there can be lines (which are called whiskers) extending from the box indicating variability outside the upper and lower quartiles, thus, the plot is also termed as the box-and-whisker plot. Outliers that differ significantly from the rest of the dataset may be plotted as individual points beyond the whiskers on the box-plot. Box plots are non-parametric; they display variation in samples of a statistical population without making any assumptions of the underlying statistical distribution. The spacings in each subsection of the box-plot indicate the degree of dispersion (spread) and skewness of the data, which are usually described using the five-number summary. Boxplots can be drawn either horizontally or vertically.

**The summary() function in R provides all the descriptive statistics of the five-number summary and also the sample mean.

Bar Diagram

In representing the frequency distribution of a discrete variable graphically, at the outset, we may take two mutually perpendicular axes of coordinates, the horizontal and vertical axes respectively showing the variate values and the frequencies; scale of each axis has to be appropriately chosen. Next, bars with equal width having heights equal to the frequencies of the variable values are drawn at the corresponding point(indicating variable values) on the horizontal axis. The diagram so formed is called bar diagram or frequency bar diagram. It should be noted that this diagram can also be drawn using relative frequencies instead of absolute frequencies.

Histogram

It is an appropriate diagram for representing the frequency distribution of a continuous variable in the sense that it considers the fact that the frequency of a class is dispersed over the interval. Here two coordinate axes are taken and the class-boundaries are shown on the horizontal axis for locating the class intervals. Next, a rectangle is drawn over each class-interval so that its area indicates the corresponding class frequency. In other words, the height of a rectangle becomes equal to the corresponding frequency density. In this manner, a series of adjoining rectangles are erected so that the area covered by this entire group of rectangles exhibits the total frequency. The diagram so formed is called the histogram of the frequency distribution. It should be noted that the widths of rectangles, which are same as corresponding class widths, are not necessarily equal.

Bivariate Plots:-

Side By Side Boxplot

Side-By-Side boxplots are used to display the distribution of several quantitative variables or a single quantitative variable along with a categorical variable. It helps to compare between dataset features. In this plot several boxplots of a quantitative variable is drawn for each of the category of a categorical variable or for the different values of a discrete data.

Scatter Plot

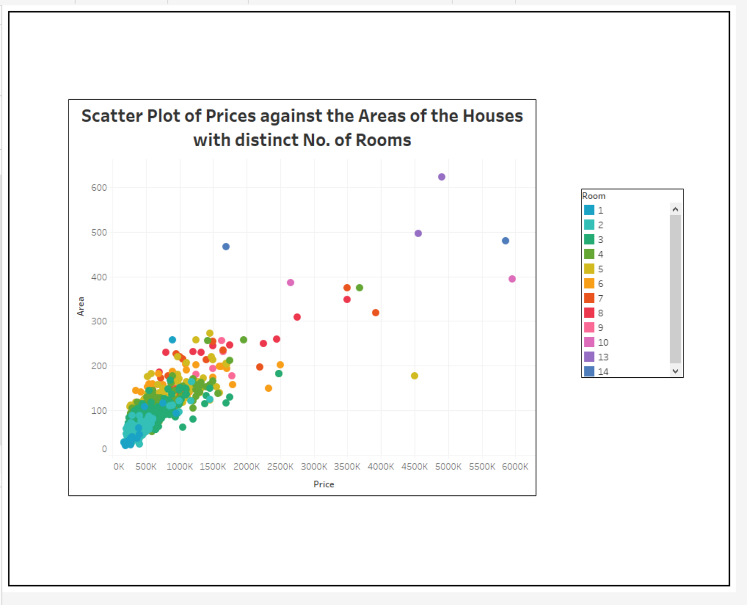
A scatter plot is a type of plot using Cartesian coordinates to display values for typically two variables for a set of data. Scatter plots are used to plot data points on a horizontal and a vertical axis in the attempt to show how much one variable is affected by another i.e. its primary uses are to observe and show relationships between two numeric variables. The dots in a scatter plot not only report the values of individual data points, but also patterns when the data are taken as a whole.

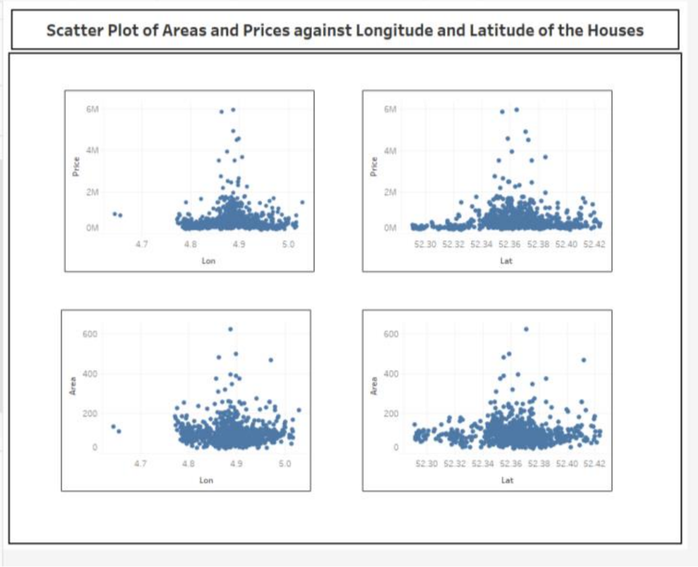
Interactive Mapping

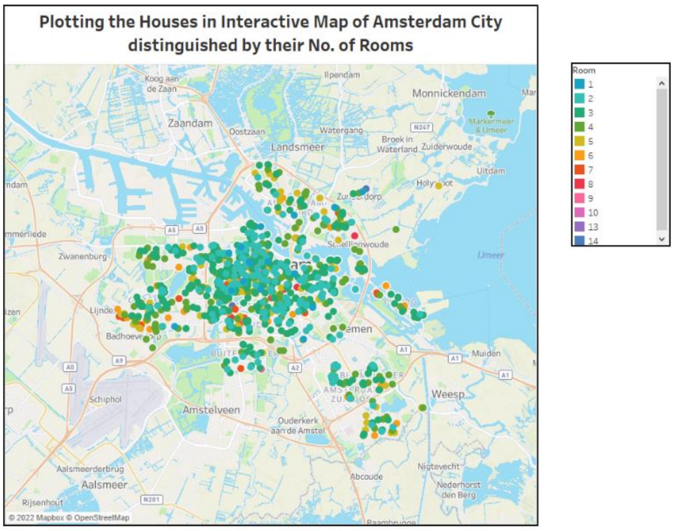
Interactive mapping involves using maps that allow zooming in and out, panning around, identifying specific features, querying underlying data such as by topic or a specific indicator (e.g., socioeconomic status), generating reports and other means of using or visualising select information in the map. Interactive maps are powerful tools for presentation. Interactive maps have several features – It creates layers of information that can be shown at the click of a button; The zoom functions allow users to focus on the details of a particular region and gain a quick overview of a wider area. The nature and distribution of a problem is made clearer by use of different map layers to give new insights and comparisons. It is easy to demonstrate how an issue affects different populations and geographic areas.

We've visualized and analysed the dataset of houses in Amsterdam using these univariate and bivariate plots. For the visualization part we've used R-Shiny dashboard. The dashboard link of this visualization project is given in the reference part. Visualizing those graphs and plots we've also analysed the components of the dataset. The analysis part is explained well in a video which link is also given in reference part.

Some plots which we have been used for Visualization and Analysis


Scatter Plot of Prices against the Areas of the Houses with distinct No. of Rooms


Scatter Plot of Areas and Prices against Longitude and Latitude of the Houses


Plotting the Houses in Interactive Map of Amsterdam City distinguished by their No. of Rooms

Results and Findings

We visualize the whole house price dataset of Amsterdam. We analyse each and every component of the dataset with statistical graphical analysis. From the analysis we found many interesting facts about the dataset.

• First consider the analysis using univariate plot -
Visualizing and analysing the component room i.e. basically the no. of rooms using five-number summary, boxplot and bar plot in the houses we've found that each of the houses in Amsterdam have on an average 2-4 no. of rooms and most of the houses in Amsterdam there are total three rooms.
We've also found after visualizing and analysing the area component using boxplot and histogram that the average area of the houses in Amsterdam is about 80-95 square metres and analysing the price component, we've got to know that the average price of the houses in Amsterdam varies between 500k Euros to 700k Euros. There are some houses which prices and areas are excessively high, in the boxplots those are pointed out as outliers. If we ignore the outliers, from the histograms of these two components we get that the both the distribution of area and price data are positively skewed.

From google we can find that the longitude of Amsterdam is 4.9041°E and latitude is 52.3676°N. We have visualized, analysed the boxplots and histograms of the components longitude, latitude of each house and found that in our dataset the distribution of the longitude and latitude data is almost symmetric. Then check the descriptive statistics mean, median, 1st & 3rd quartile of those two variables and comparing with the longitude and latitude of Amsterdam city we've got that the house dataset covers almost the whole Amsterdam city i.e. the house data are collected from the entire city not a specific part(like only the middle part or the extreme part of the Amsterdam city). The houses in our dataset are equally distributed all around the Amsterdam city.

• Now come to the bivariate and multivariate analysis part -
From the scatter plot of area and price against the room (no. of rooms in the house) we've found that as no. of rooms in house increases price and area also increase and it's obvious because the total size of a house gets higher if more rooms are added to the house, also for this house becomes pretty much expensive but there are some exceptional that no. of rooms in some houses are less but price and area are high. This may be for the houses' position, decoration or the rooms of those houses are large or for some other quality characteristics of those house.
In the scatter plot of price against area we've seen that the prices of the houses are going higher as well as the areas are getting larger which is obvious. There are some houses also which have not so large area but expensive.
From the scatter plots of room, price and area against longitude and latitude we've found that the houses having no. of rooms between two to four, the houses having area less than or equal to 90 square metres and the houses of price not more than

700k euros are equally distributed all over the Amsterdam city. So, in any part of the Amsterdam city we can get houses having those characteristics(no. of rooms, area, price).
We've also noticed from those scatter plots that as well as the no. of rooms, area and price of the houses increase excessively the longitude and latitude values get close to the central value of their range i.e. getting close to their mean and median. Now, we've seen earlier from the longitudes and latitudes of the houses in our dataset that the houses are distributed equally all over the city. So, the longitudes and latitudes are going towards centre of their range means the positions of the houses are going to the central part of the Amsterdam city (as well as the no. of rooms, area and price of the houses become extremely high).
We know the houses which have excessive no. of rooms, large area and are too much expensive those are the luxurious house or villa type house or bungalow. Hence, the scatter plots show that these villa type houses or bungalows are situated in the central part of the Amsterdam city. This is an important fact and it's familiar for any metropolitan city.

• Interactive Maps -
The houses in our dataset are plotted in maps basis on their longitudes and latitudes. Houses are plotted in three interactive maps in which the houses are distinguished based on the no. of rooms, areas and prices.
From the interactive maps it has been found again that most of the houses in Amsterdam city have no. of rooms two to four, the area of most of the houses is between 80-95 square metres and the price range of most of the houses is between 500k to 700k euros. We've also seen in interactive maps that the houses having large no. of rooms, big area and of high cost price are located in the central part of Amsterdam.

Conclusion

We have visualized and analysed the house dataset of Amsterdam city. Each quantitative component of the dataset has been analysed using univariate plots and any relationship between any two variables or components is also analysed with bivariate and multivariate plots.

We can conclude from the results and findings of analysis that on an average the houses of Amsterdam have two-four no. of rooms, 80-95 square metres of area and cost price approximate 500k-700k euros. If we want a house having large no. of rooms and large area then we have to expense much.

We can also conclude that if anyone want to get a house having no. of rooms two-four, area less than equal to 90 square metres and cost price not more than 700k euros he can find such houses in any part of the Amsterdam city but if he want to get a house having excessively large number of rooms, large area and of higher price he has to go to the central part of the city.

This study and analysis represent an overview of information and characteristics of the houses in Amsterdam city. It will help to gain some idea in purchasing houses in Amsterdam.

Acknowledgement

I am indebted to so many people for helping me in the preparation of this project.

I owe a deep debt of gratitude to my supervisor Dr. Sourish Das for necessary guidance, for this presentation of this dissertation, valuable comments and suggestions. I am extremely grateful to him for the necessary stimulus, support and valuable time.

Finally, my earnest thanks go to my friends who were always beside me when I needed them without any excuses.

References

Sources of Data:-
https://www.kaggle.com/thomasnibb/amsterdam-house-price-prediction?select=HousingPrices-Amsterdam-August-2021.csv

Visualization and Analysis:-

R-Shiny Dashboard Link-
https://aniketvisualization.shinyapps.io/aniket_visualizationprojectdashboard/?_ga=2.171247971.23266111.1639583986-455046010.1638821410

Analysis and Presentation Video Link-
https://youtu.be/wR79Q8R_5z0

Other sources which have been used during the project:-

https://en.wikipedia.org/
https://ggplot2.tidyverse.org/
https://shiny.rstudio.com/