**7PAM2000 Applied Data Science 1**

**Visualization Assignment**

**Daniel Fiyinfoluwa Oyeduntan**

**21080788**

**M.Sc. Data Science**

**Table of Content**

1. Background
2. Data source and information about the data
3. Python Environment Set up
4. Importing and Retrieving the Data
5. Definition of Functions
6. Data preparation and preprocessing
7. Data Analysis
   1. Bar Chart
   2. Scatter Plot
   3. Line Chart

8.0 Links and References

1. **BACKGROUND**

Data Visualization otherwise known as Exploratory Data Analysis (E.D.A.) is one of the most important aspect of Data Science – it is the summarization of data using graphical representation to provide meaningful and readable insights. There are so many statistical graphics and method deployed in this aspect of data science some of which are line charts, box plots, scatter plots, line plots etc; all of these have various best use and are more appropriate for different datasets, understanding how and when to deploy them is one of the most important skill to have as a Data Scientist.

In this report, we will be doing an Exploratory Data Analysis on a particular dataset and drawing inference and insights from the data.

**2.0 DATA SOURCE AND INFORMATION**

The data to be used in this analysis was gotten from kaggle.com; one of the world’s biggest platform for data science and house to numerous datasets. The data used is a USA real estate dataset gathered by **AHMED SHAHRIAR SAKIB** 5 months ago – June 2022, it has been updated 9 times since then with the last update happening in August 2022. The dataset consists of 923159 rows and 12 columns (923159 x 12) and is about 120mb in size. The 12 column headers with what they represent are as follows:

|  |  |  |
| --- | --- | --- |
| **S/n** | **Column Header** | **What they represent** |
| 1 | Status | Current market status – for sale or ready to build |
| 2 | Price | Selling price |
| 3 | Bed | Number of bedrooms |
| 4 | Bath | Number of Bathroooms |
| 5 | Acre\_lot | The measurement the land in square metres weighed to acres |
| 6 | Full Address | The full address of the property |
| 7 | Street | The street the property is located |
| 8 | City | The city the property is located |
| 9 | State | The state the property is located |
| 10 | Zip\_code | The zipcode for the property |
| 11 | House\_size | The size in square metres of the actual house |
| 12 | Sold\_date | The date the house was sold |

**Link to website**: <http://kaggle.com/>

**Link to dataset page**: <https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset>

**Direct download link**: [download link](https://storage.googleapis.com/kaggle-data-sets/2218738/4063426/compressed/realtor-data.csv.zip?X-Goog-Algorithm=GOOG4-RSA-SHA256&X-Goog-Credential=gcp-kaggle-com%40kaggle-161607.iam.gserviceaccount.com%2F20221110%2Fauto%2Fstorage%2Fgoog4_request&X-Goog-Date=20221110T113528Z&X-Goog-Expires=259200&X-Goog-SignedHeaders=host&X-Goog-Signature=)

**3.0 PYTHON ENVIRONMENT SET UP**

For this exercise, python 3.9 was used for code work and analysis and the IDE used was Sypder (v 5.3.3). Python is a simple to learn multi-purpose programming language, which means it can be used to web and software development alongside data science. Due to its versality and simplicity, it is arguably the most used programming language right now in the Data Science world.

Python has numerous libraries available to its users to help achieve propose tasks; for data scientist, it has Pandas, Numpy, SkLearn, Matplotlib, Scipy, Seaborn etc. For the purpose of this exercise we will be making use of 3 powerful python libraries:

1. Pandas – For data wrangling and manipulation
2. Numpy – For mathematical computing
3. Matplotlib – For Visualisation (graphical representations)

To retrieve these libraries, a simple ‘import’ keyword is used and also give an alias (optional)

i.e. *import pandas as pd*

**4.0 IMPORTING AND RETRIEVING THE DATA**

After setting up the python environment, the next step is to import the dataset into a pandas data frame to enable pre-processing and cleaning. For this exercise we imported two extra modules (urllib.request and zipfile), this was necessary due to the nature of the datasets and the Kaggle.com. After downloading the dataset we then imported it into a named panda data frame.

**5.0 DEFINITION OF FUNCTIONS**

A function in python in a term used to describe a group of related statements that performs a specific task, they could either be set to take in arguments and are called in the middle of the program to perform certain tasks.

For this analysis, we defined 7 functions to make our program a lot better and concise, details on these functions are listed below:

|  |  |  |  |
| --- | --- | --- | --- |
| **S/n** | **Name of Function** | **What they do** | **Arguments** |
| 1. | nan\_dataframe (dataframe) | Produces a data frame showing the number of missing values in each column of the data frame passed | Takes one argument which is the data frame |
| 2. | drop\_nan (df,column = "") | Produces a new data frame after dropping all missing values from all columns or specified columns | Takes one or two argument - the first one being the data frame and second(optional) is the specific column that we want to drop from |
| 3. | group\_by (dataframe,groupby\_column, grouped\_columns\_list,group\_method) | Groups or summarizes data frame by methods and given criteria | Takes 4 arguments - the data frame, the columns we want to group by, a list of columns we want grouped and the method of grouping |
| 4. | des\_hist (dataframe) | Prints a table of the description of the data frame passed into it and plots a histogram of all numeric columns. | Takes only one argument which is the data frame. |
| 5. | bar\_chart\_subplot (x\_axis,my\_list,labels) | Plots a subplots of bar charts. | Takes 3 arguments - the x axis, a list of y axis and a corresponding list of labels |
| 6. | scatter\_plot (my\_list,label) | Produces scatter plots of variables in a list, each plotted against every other variable | Takes two arguments - a list of arrays to be plotted and the corresponding labels |
| 7. | line\_plot(x\_axis,y\_axis,y\_label,x\_label) | Plot multiple line graphs. | Takes 4 arguments - the common x axis, a list of y axis to be plotted, a corresponding list of labels for the y-axis and the x-axis label. |

After defining functions, we make use of them in between the program by calling them.

**6.0 DATA PREPARATION AND PRE-PROCESSING**

One of the major challenges data scientist face is the issue of messy data, most of the real world dataset are filled missing values or duplicates and as such it is important that the dataset is cleaned and pre-processed before analysis.

For this analysis, our dataset is not so messy; however, it is not completely clean so we still have some pre-processing to do.

First of all, a full description check was done to know initial state of the dataset and I noticed that it had quite a number of missing values; to fix this we took out the all rows with missing values by calling our predefined function. After this, our dataset was ready for the actual EDA.

***Picture of histogram before EDA***

Chart, box and whisker chart

Description automatically generated

***Picture of histogram after EDA***

Chart, box and whisker chart

Description automatically generated

***Please ignore the zip code graph, it is not exactly a numeric column, but because zip codes come in numeric forms, the code assumed it was one***

From both pictures above, we will observe that the distribution of the variables did not change after the pre-processing. Infact, we now have a better zoom in histogram.

**7.0 DATA ANALYSIS**

The process of analysing data is majorly done to gain insights from the data (i.e. harness useful information). A lot of times, before plotting a graph after the data has been cleaned, we still need to summarize them so that it will make sense to be plotted on a graph. In this exercise, we did many summarizations, either by calling the group by function previously defined or by slicing existing data frames to create new ones.

We will be making using three matplotlib charts in this exercise, all of which already have predefined functions to help plot easily:

1. Bar Chart
2. Scatter Plot
3. Line Plots

**7.1 Bar Chart**

A bar chart graphically represents data by displaying vertical bars next to each other, lined up on the horizontal axis. Each bar represents a different category, and the height of the bar correlates with numbers on the values axis. They give you an immediate way to compare values for related data sets side by side, highlighting trends in a swift, visual way.

We will this graphical method to compare the averages of prices of houses, number of bathroom, number of bedroom, size in acre and the size of the house for the different states.

***Bar Chart showing average prices, no of bathrooms and bedrooms, size in acre and house size in all regions***

***Chart, bar chart

Description automatically generated***

From the above graph, we are able to conclude the following:

1. Virgin Islands has the most expensive houses followed by New York and Massachusetts
2. The average number of bedrooms in house in these states are between 3 – 5 except Puerto Rico where we see the average number of bedrooms being 7.
3. We can also see from the third bar chart, the average number of bathrooms correspond with the number of bedrooms across all state which is between 2 – 3 (about one less than the number of bedrooms). We however can see that Virgin Islands averages about 5 bathrooms which is about the same as the average number of bedrooms as seen in the second graph.
4. The fourth graph is a particularly interesting because it shows us how big lands are in various states; we can deduce that lands in Rhode Island, New Jersey and Vermouth are a lot bigger than in certain states like Puerto Rico, Virgin Island and Pennsylvania. What is more interesting is that, when we compare the prices and also the house size of these states, they are confliction with the Land size, meaning if a customer is looking to geta cheap house on a large land without minding the size of the house, the best state to look at will be Rhode island.
5. The fifth graph shows us the average size of houses in these state, one very interesting insight from this graph is that houses in Virgin Island are way bigger than other states, now when we look at other graphs to get a bigger picture of Virgin island, we can see that for its price even if it’s a smaller land, you still get a big house with about 4 bed rooms and 5 bathrooms. One could assume that residents of Virgin island are not so keen on company or lot size.

**7.2 Scatter Plots**

Scatter charts are a rather unique method to visualizing data. It is a kind of plot that rely on Cartesian co-ordinates. They are great and most effective in scenarios where you want to display both distribution and the relationship between two variables.

We will use this plot to check whether or not there is a relationship between certain features of the dataset (prices of houses, number of bathroom, number of bedroom, size in acre and the size of the house for the different states).

***Scatter plots of all numeric features of the data sets plotted against each other (average prices, no of bathrooms and bedrooms, size in acre and house size in all regions)***

*7.2.1 Price against price, bed, bath, acre\_lot, house\_size*

***Chart, scatter chart

Description automatically generated***

*7.2.2 Bed against price, bed, bath, acre\_lot, house\_size*

***Chart, scatter chart

Description automatically generated***

*7.2.3 Bathe against price, bed, bath, acre\_lot, house\_size*

***Chart, line chart, scatter chart

Description automatically generated***

*7.2.4 Acre\_lot against price, bed, bath, acre\_lot, house\_size*

***Chart, scatter chart

Description automatically generated***

*7.2.5 House\_size against price, bed, bath, acre\_lot, house\_size*

***Chart, scatter chart

Description automatically generated***

From the above plots, we are able to conclude the following:

1. There is a weak or negligible correction between price of houses and the number of bedrooms (r = 0.29)
2. There is a stronger correction between price of houses and number of bathrooms though it is still deemed fair statistical (r = 0.54)
3. There is no correlation between the prices of houses and how large the land is
4. There is a medium correlation between the prices of houses and the sizes of the house (r=0.43)
5. The most correlation variables are number of bedrooms & number of bathrooms and size of the house & number of bathrooms.

**7.3 Line Plots**

Line charts plot data points on a graph and then join them up with a single line from each point to the next, they are simple and very popular as they give you an instant picture of a trend over a period of time. You can see peaks and dips, whether the overall values are moving up or down, and when there’s a sharp spike or drop in numbers. You can also compare changes over the same period of time for more than one group or category.

In this exercise, we will plotting three line plots;

1. Average prices of house over the years
2. Average house sizes over the years
3. Multiple line plot of average number of bedrooms, bathrooms and land size over the years.

We had to split this because of the scale of the various features, if they are plotted together we won’t see some features.

***Line plots of average prices, no of bathrooms and bedrooms, size in acre and house size over the years***

*7.3.1 Line plot of average prices of houses over the years*

Chart, line chart

Description automatically generated

*7.3.1 Line plot of average prices of houses over the years*

***Chart, line chart

Description automatically generated***

7.3.2 *Line plot of average house sizes over the years*

*Chart, histogram

Description automatically generated*

*7.3.3 Line plot of average number of bedrooms, bathrooms and land size over the years*

From the above plots, we are able to conclude the following:

1. The prices of houses saw a steady uprise from around 1907 up until around 1953 where we see an all-time crash, it started to increase in subsequent years, however in the 1970s we see a significant spike, dropping again in the 1980s and spiking again to hit an all-time peak in the early 1990s which dipped again towards the end of the 1990s and in the 21st century we see a continuous upward and downward movement (net downwards at the end)
2. The sizes of house started to drop in 1907 just about the same time the prices started to go up until it the sizes dropped significantly in early 1950s. After which builders started making the houses bigger as we can see the trend with the biggest of houses sold in 1970s.
3. If we compare the prices and the size of the house across the entire time period, it might be difficult to see a relationship, but if we break it into 3 time brackets – 1900s to 1950s, 1960s – 1990s and 1900s upward; we can notice that in the first interval, they had a negative relationship, but in the second time bracket. The began to move together.
4. In the third graph, we can observe that all through the years, the number of bed rooms and bathrooms remained consistent, though the size of lands can boast of almost similar trend it did experience a significant spike in the 1970s, just about the same time the size of house got bigger and the prices also increased.

**8.0 LINKS AND REFERENCES**

**Link to website:** [**http://kaggle.com/**](http://kaggle.com/)

**Link to dataset page:** [**https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset**](https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset)

**Direct download link:** [**download link**](https://storage.googleapis.com/kaggle-data-sets/2218738/4063426/compressed/realtor-data.csv.zip?X-Goog-Algorithm=GOOG4-RSA-SHA256&X-Goog-Credential=gcp-kaggle-com%40kaggle-161607.iam.gserviceaccount.com%2F20221110%2Fauto%2Fstorage%2Fgoog4_request&X-Goog-Date=20221110T113528Z&X-Goog-Expires=259200&X-Goog-SignedHeaders=host&X-Goog-Signature=)

**Link to github:** [**https://github.com/odcalculus**](https://github.com/odcalculus)

**Link to github repo:** [**https://github.com/odcalculus/USA-Real-Estate-ADS1**](https://github.com/odcalculus/USA-Real-Estate-ADS1)