REPORT

ABSTRACT

The aim of the project is to format and clean the data to make it suitable for analysis. Then, draw conclusions and interpretation from the data by visualizing the formatted data. Formatting and cleaning of data involves a process called 'Data Analysis'. Python provides us several libraries such as pandas, numpy and matplotlib to do so. More detailed information is given along with project details.

INTRODUCTION

Data Analysis is the process of evaluating data using analytical and logical reasoning to examine each component of the data provided. Here the objective is to analyse the data and to draw conclusions from the data. It also includes drawing conclusions by seeing through the relationships among different parameters of the data. Project involves the analysis of the data present in the Bank.csv file.

Data visualization describes any effort to help understand the significance of data by placing it in a visual context. Simply stating, data visualization is the presentation of data in a pictorial or graphical format. It includes pie charts, bar graphs, histograms, box plots and many more pictorial formats.

DATA PREPARATION

Firstly, we import pandas. Then, 'Bank.csv' is loaded using the pandas function 'read_csv'. Arguments used are- ; as separators and the first row as header. While loading this data into a data frame, the detail that is taken care of is the extra double quotes present in the csv.

Functions such as 'dtypes', 'head' and 'columns' are used to ensure that the data in the dataframe (bank_df) is same as the data in the csv.

The column 'duration' did not have a numeric data type. On checking the rows, it was found that '-' was in some of the rows. It had to be converted to NaN and then the 'duration' column

got it's float64 datatype. Similarly, it was found that the column 'age' had an empty value of ''. It was replaced by NaN and then converted into float64 datatype. Rest all columns were ok.

There were some typos in the loaded dataframe. For instance, divorceded" instead of divorced. Typos were present in fields - marital, loan, housing, day_of_week, default, education. These were corrected and made consistent with the data.

To ensure that there are no instances of whitespaces, a function was created that would iterate through all columns and rows to remove whitespaces using the 'strip' function.

The function 'lower' was used on all columns in order to cast data to lower-case. This would help maintaining the consistency in data.

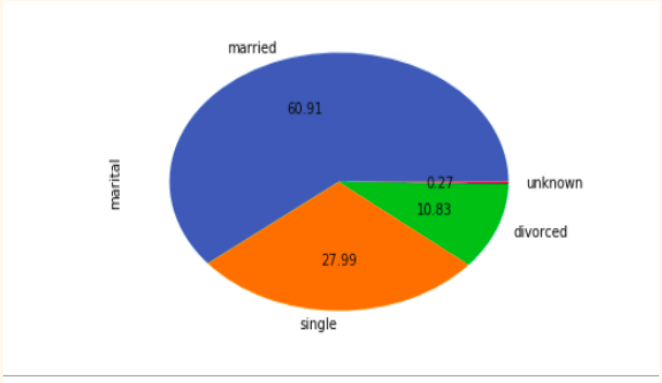
Tests were carried out to ensure the apt data-type for 'duration' and 'age' column as stated above. Rest of the tests included looking for the columns which contain the NaN. These NaN values in columns of - 'age', 'cons.price.idx', 'emp.var.rate', 'nr.employed' and 'euribor3m'. These were replaced by mean values of respective columns.

DATA EXPLORATION

Part 1

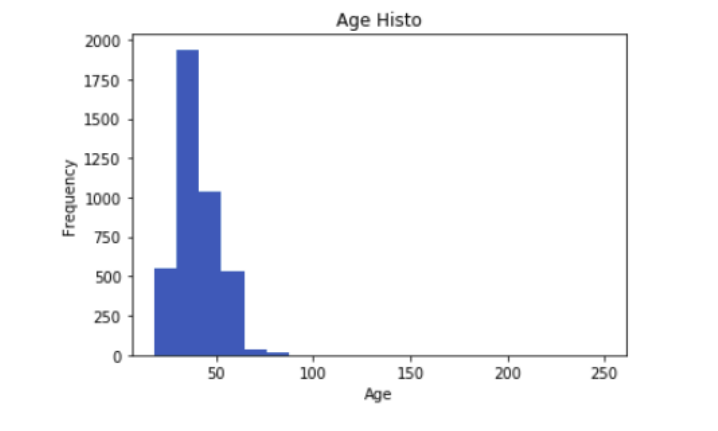
We create a visualization for each column by producing appropriate types of graph. The graphs used for individual representation of the data are Pie Chart, Histogram, Box Plot and Bar Graph.

Pie Chart



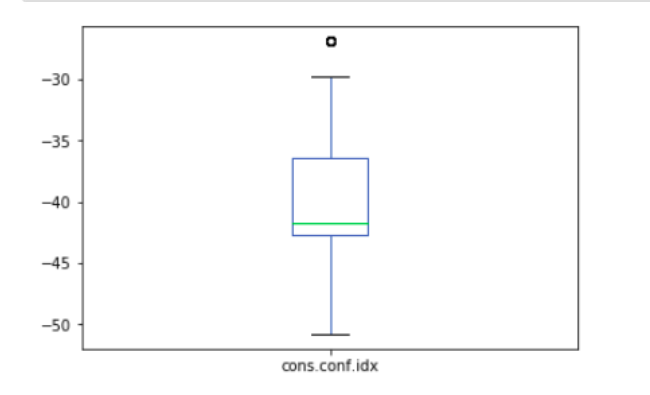
Pie Charts are made to show - marital, default, housing, loan, contact, poutcome, y. Pie Chart is used here because these fields have limited possibilities. e.g . y has only two possibilities yes or no and so pie chart is helpful for visualisation here as the percentage of people can be easliy distinguished without any overcrowding.

Histogram



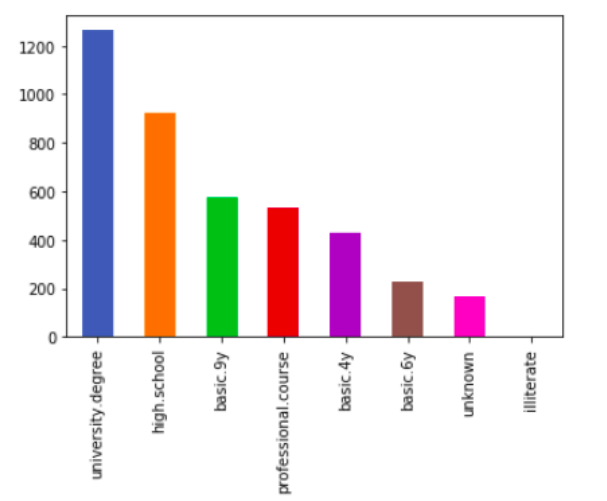
Histogram is used to plot numerical data and show the frequency of occurance of the data. It is made for fields - age, campaign, pdays, duration.

Box Plot



Box Plot is made for numeric values but only for those which are close to each-other. The data which is apart from the box (the outliers) are considered as exceptional cases.

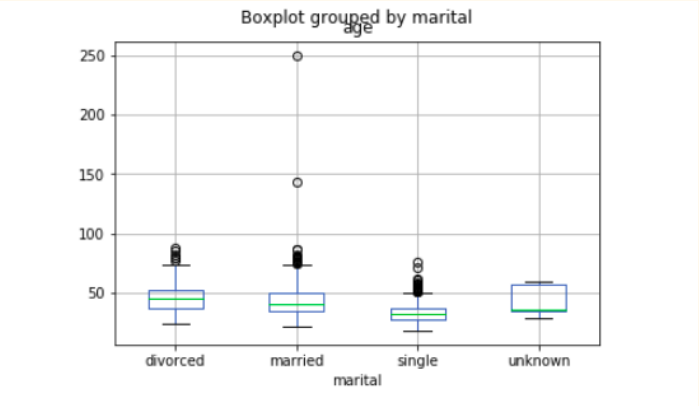
Bar Graph



These are used to visualize the object data in which we have many choices. Data with many choices lead to overcrowding of the Pie Chart. So, we make use of bar graphs. Plus, bar graphs can show up the exact values, rather than percentage as in case of pie chart.

Part 2

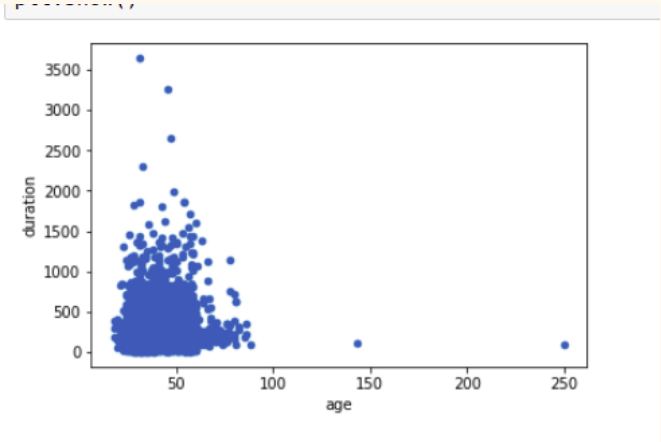
We explore the relationship between different columns here. We use box plots to show relation between numeric and object data. And we use scatter plot to show relations between numeric data.



.

The box plot between age and marital shows that married people and divorced people have the same age whereas the age of singles is obviously less. One interesting fact here is majority of the data is unknown here.

Another box plot is made between age and loan which shows that people who are under debt numbers are almost similar to those who are not under loan. Here also much data is unknown.

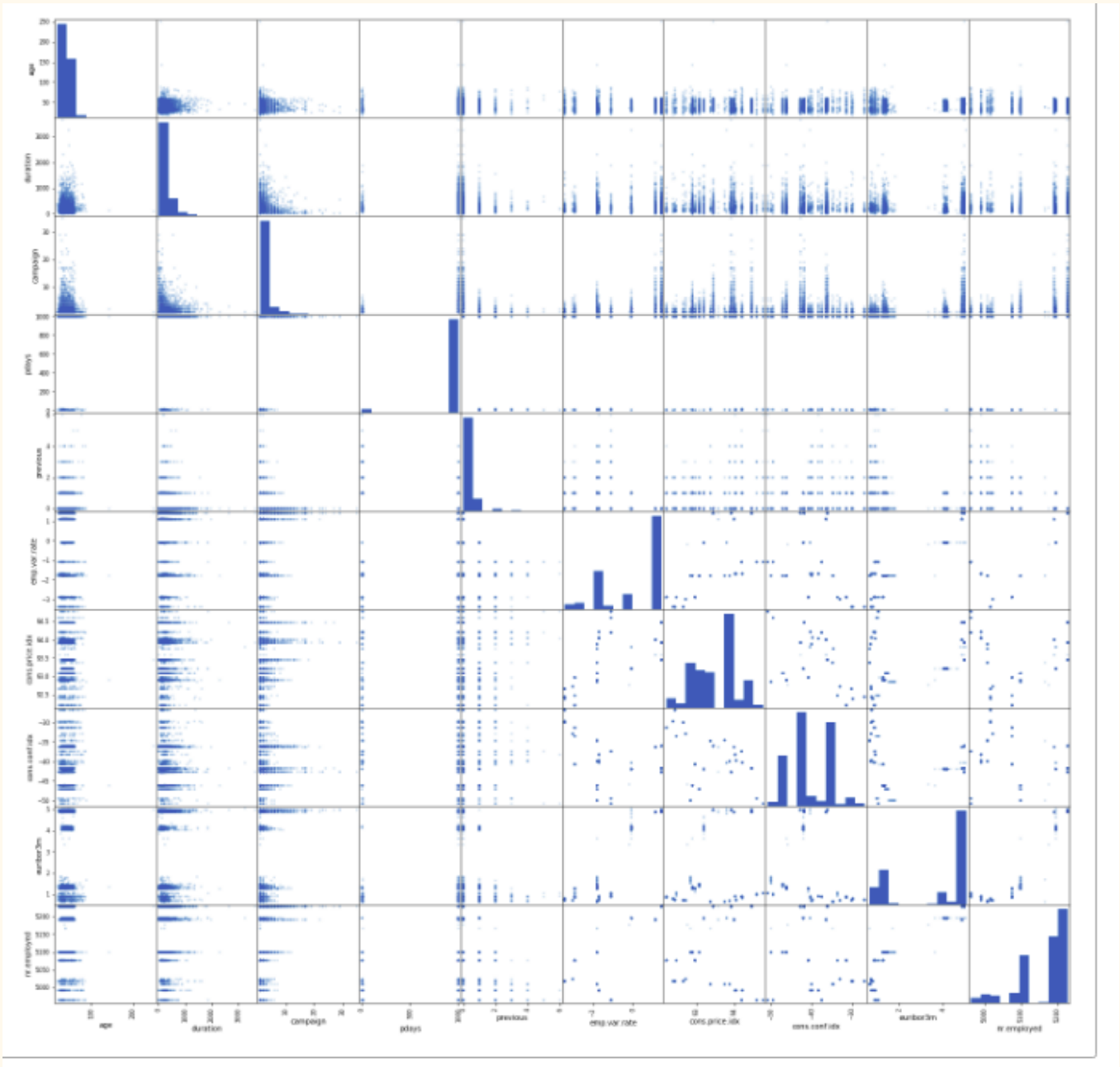


Scatter plot between age and duration shows that most people have a duration of under 2000. Very less have extraordinary durations exceeding 3000.

Part 3

Build a scatter matrix for all numerical columns.

The scatter plot is as shown below-



CONCLUSION

From above it is clear that Data Analysis is a vital tool for making meaning out of data. It is clear how Python automates this complex Data Analysis for us. Libraries such as numpy, matplotlib and pandas automates data analysis and data visualization for us.

Thus, we have learnt and made use of useful Data Analysis tools in analysing and visualising the data in ‘Bank.csv’.