Table of Contents

ABSTRACT

The point of this report was to analyse and clean the given "Bank" dataset to discover insights and connection between the different columns in them. The data is identified with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns depended on phone calls recorded over a period. In general, the outcomes demonstrate that the marketing effort procedure is shameful. The report presumes that the general marketing effort was a disappointment as a dominant part of the customers did not settle on term deposits. It is suggested that an appropriate file of campaigns like these must be kept up to predict and investigate legitimate patterns in data and for upgrading the business.

INTRODUCTION

The data which is decided for analysis is gotten from a Portuguese banking institution. Since the data gathered was from phone calls, there are potential outcomes of imperfections and errors in the nature of data. In addition, the data gathered was identified with direct marketing campaigns. Because of the reason that there is in excess of one contact to the same client to get to if the item (bank term deposit) would be ('yes') or not ('no') subscribed. The dataset contains 21 attributes, where each quality has its own particular data type. The given dataset is in the crude format and should be cleaned to get legitimate outputs from it. This report will talk about on how the data is prepared, cleaned and investigated and what conclusions and bits of knowledge that has been drawn from the cleaned data. This report will likewise talk about the different methods for how the cleaned data has been interpreted to get proper knowledge from it.

METHODOLOGY

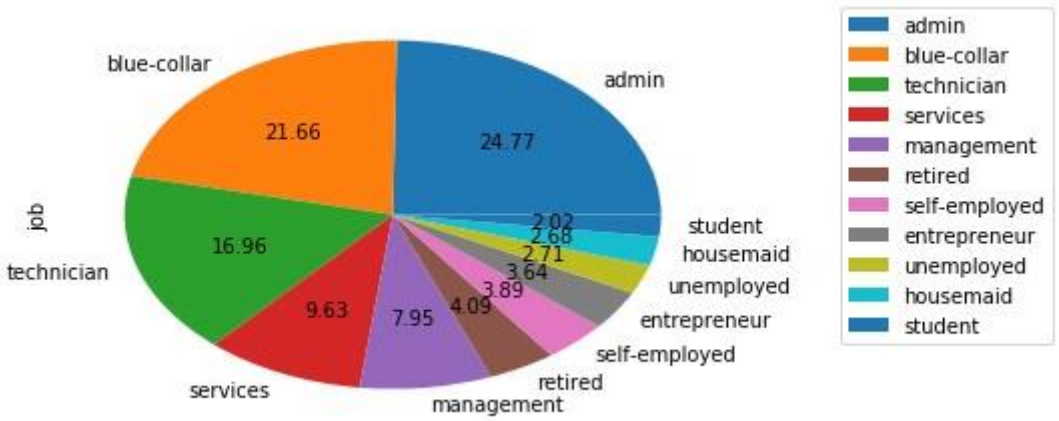
Python's pandas feature is the primary tool which is chosen for analysis of the above raw data. Matplotlib feature of python is used in crucial visualisation. Pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language. The above bank dataset is from a Portuguese financial institution which conducted direct marketing campaigns. The dataset is provided for analysis and has not been collected by me for this analysis.

The story goes this way, the data is first loaded from a csv file into a pandas dataframe, as the csv file cannot be directly used for analysis. The cleaning of raw data is performed on the dataframe and is stored in a separate csv file. The new csv file, which cleaned and free of errors is then plotted and visualised using matplotlib feature of python library. The whole analysis is performed on jupyternotebook. The visualisations are done on various attributes separately and some attributes are compared with other attributes for visualisations.

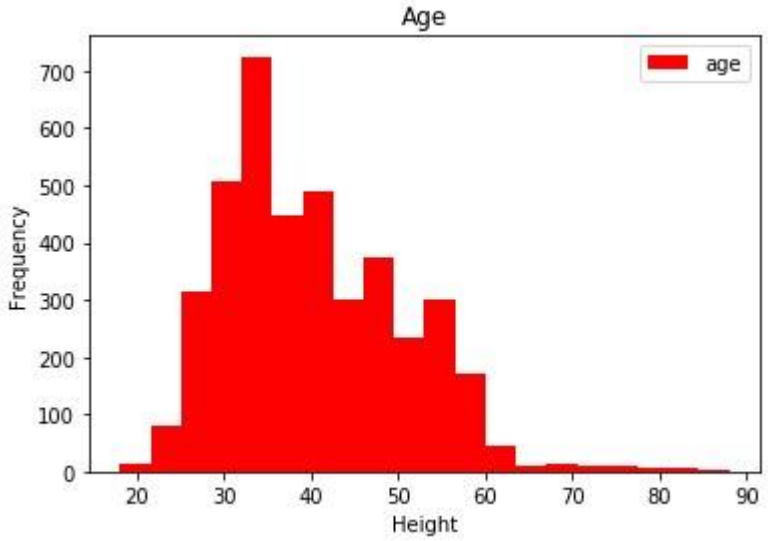
DATA PREPARATION

1. The csv file is loaded into the dataframe “Bank”. The separator used here is “;” and the decimal as “.”.
2. On checking the datatypes, majority of the datatypes are “object”. All required operations are done to change the datatypes of the required attributes are changed.
3. Value_counts is used to check and rectify the appropriate typos in the dataframe. There were a few instances where typos were encountered and replaced manually. In some cases, values in the attributes were shifted suitably to match their set of data.
4. The extra spaces in the data where checked using the str.isspace() function. This function returns a FALSE – if there are no white spaces and TRUE – if there is a white space. str.strip() is made use of in eliminating white spaces.
5. We can change the text data into lowercase str.lower function.
6. Sanitary checks are used to identifying impossible values in data. The Euribor interest rate must be greater than 0. Another check was that the Age must be greater than 18 and less than 99.
7. The fillna function is used to fill-up all the missing values in each attribute by its mean value.

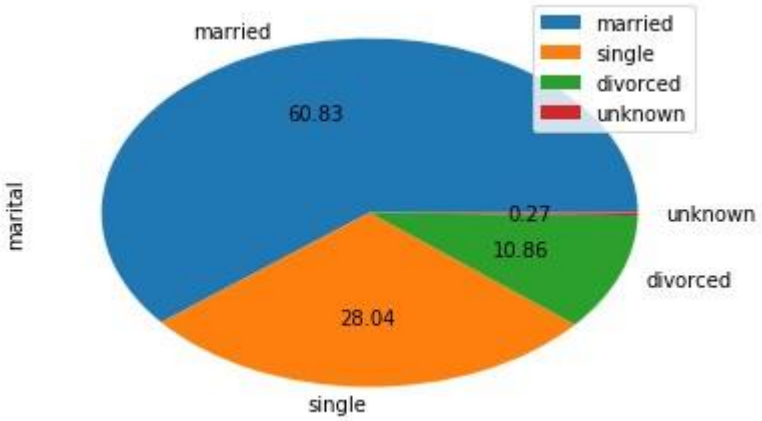
DATA EXPLORATION



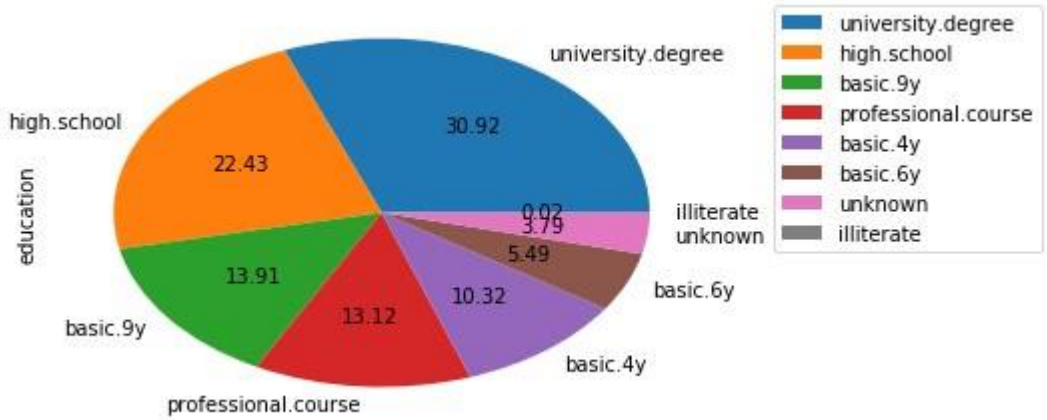
The attribute "job" is plotted in a pie chart to depict the various categories. It is used to show the frequency of the data.



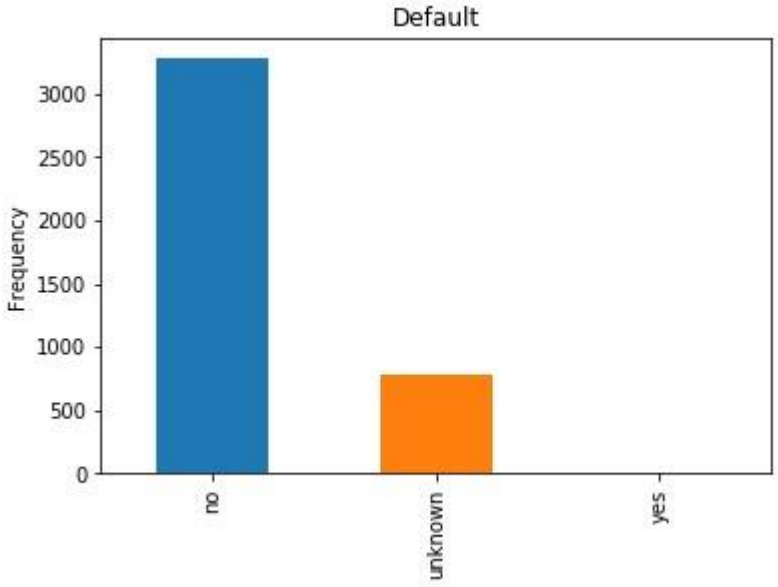
The histogram plot is chosen for depiction of the above data. As histogram shows values in a continuous manner which easy for comparison of peaks and downs.



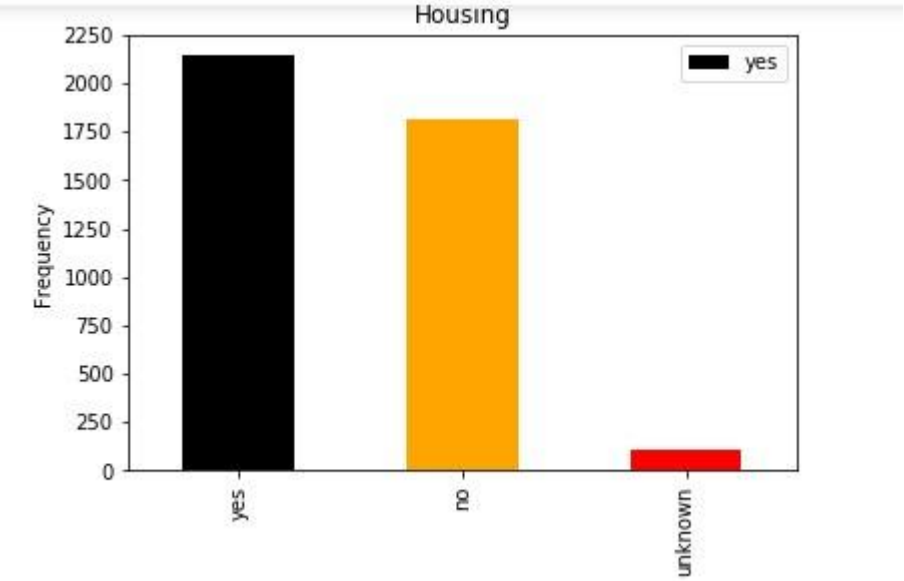
A pie chart is used to show the attributes such as married, single, divorced and unknown. It is used to describe the categorical data.

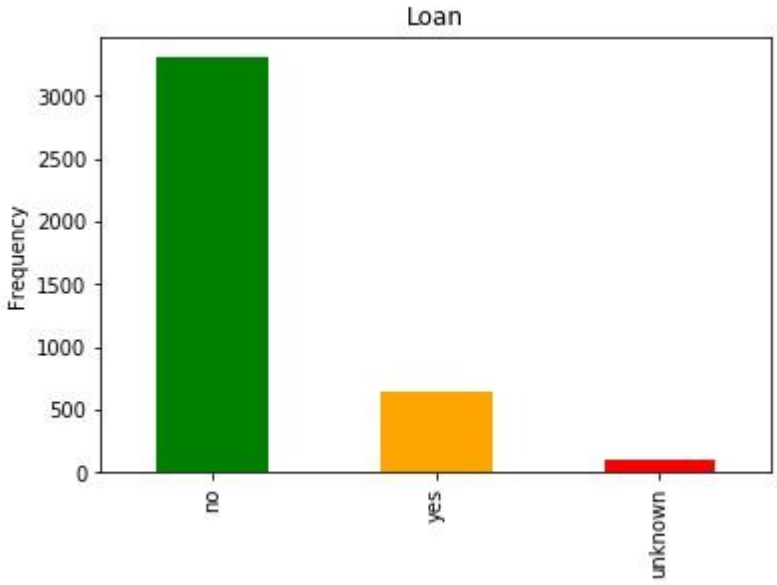


A pie is also used for depicting the previous and this graph as there are many values to compare.

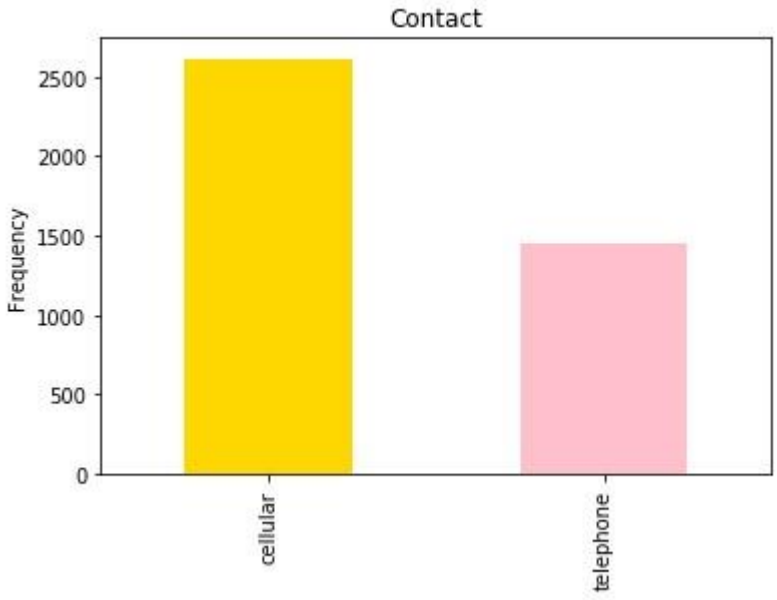


The bar chart is used to represent the Default attribute.

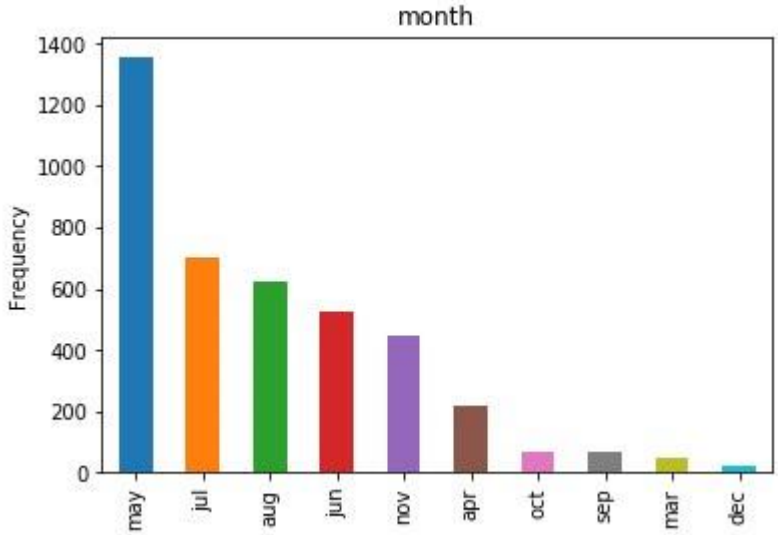




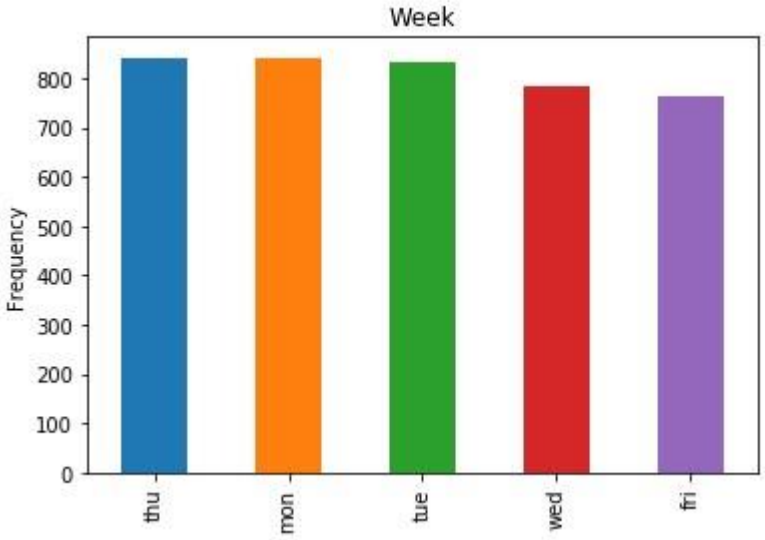
The Bar chart is used to show the Loan column.



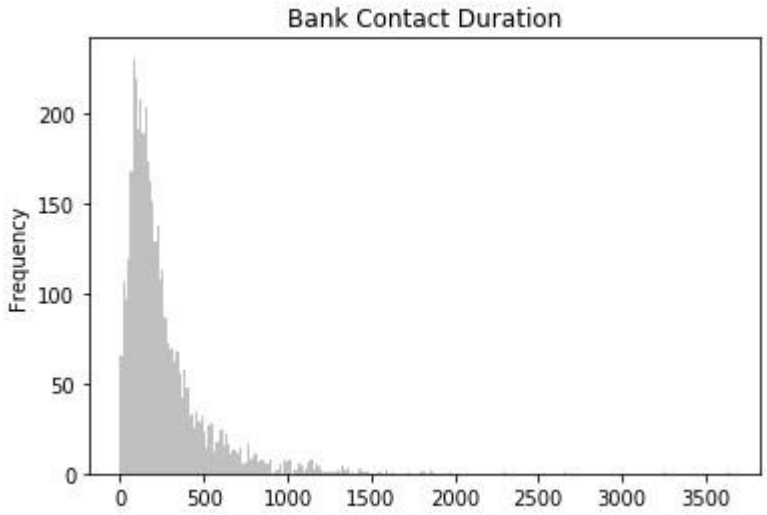
A Bar graph is used here since there are only two parameters to compare. The graph is plotted as Frequency vs Contact.

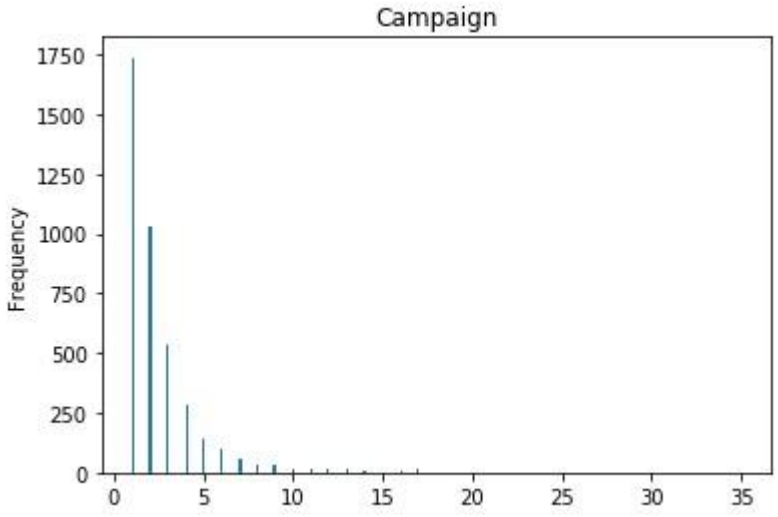


A continuous bar graph is used above so as to show various categories of months of the year. The month is plotted over frequency.

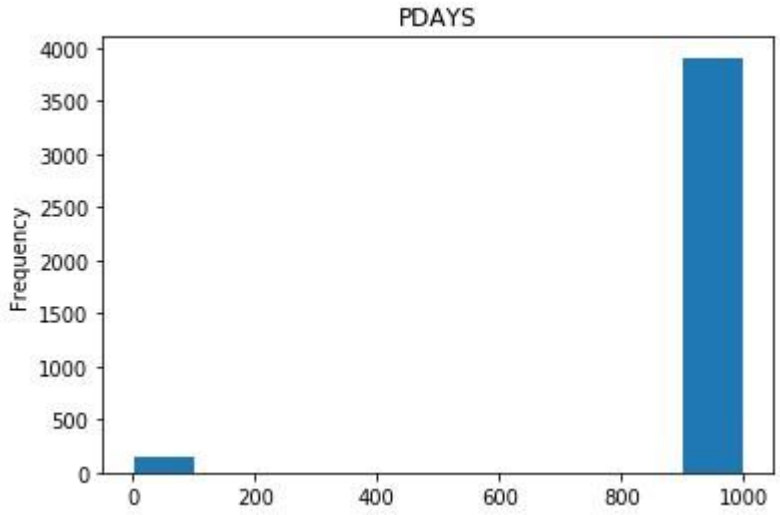


The bar chart is used to show the Week attribute.

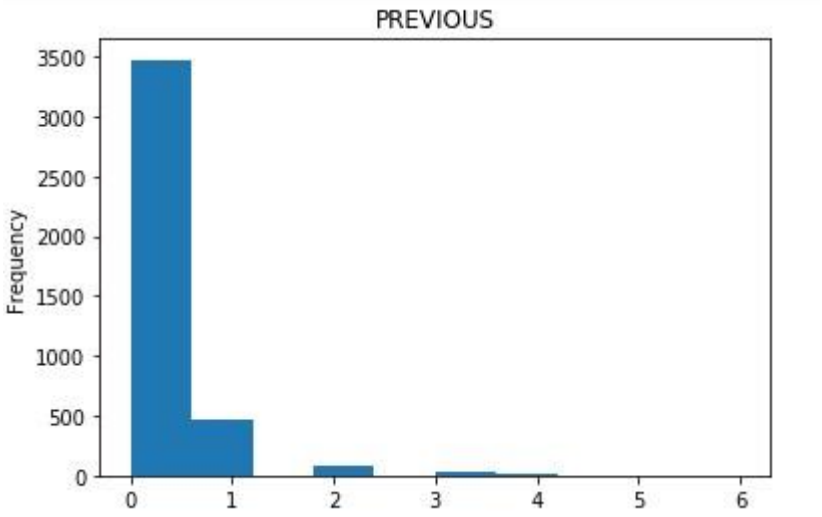




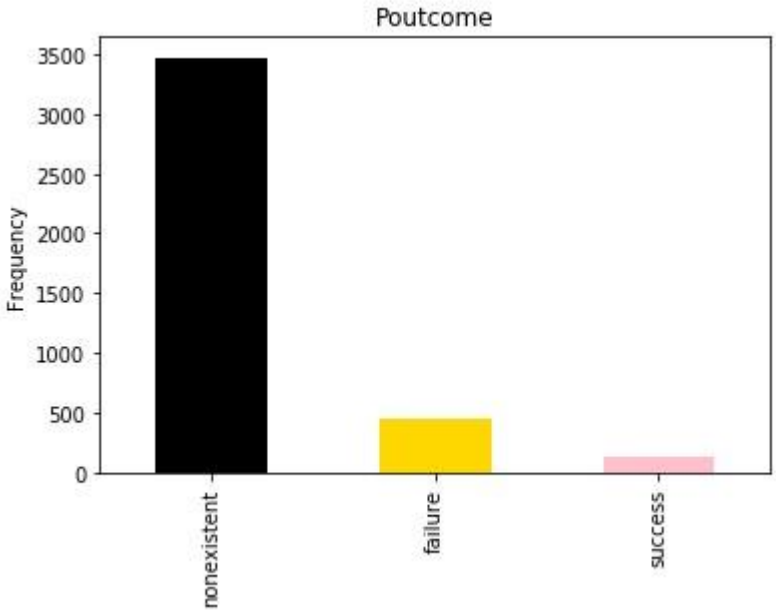
The Bar chart is used to represent the Campaign attribute.



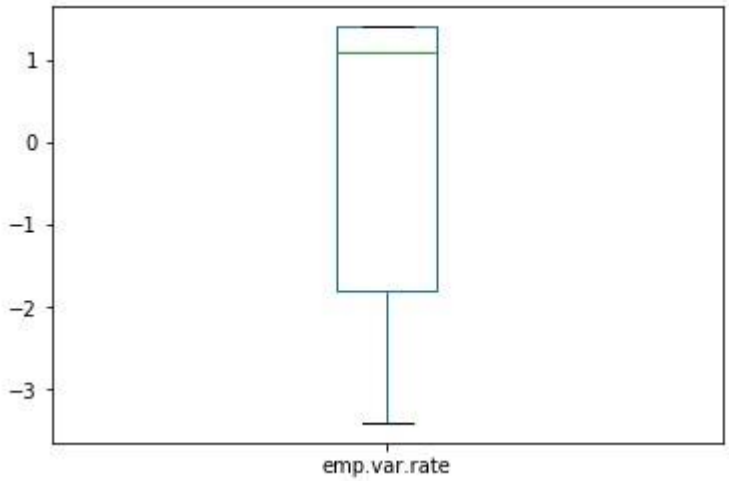
The Bar chart is used to represent the Previous attribute.

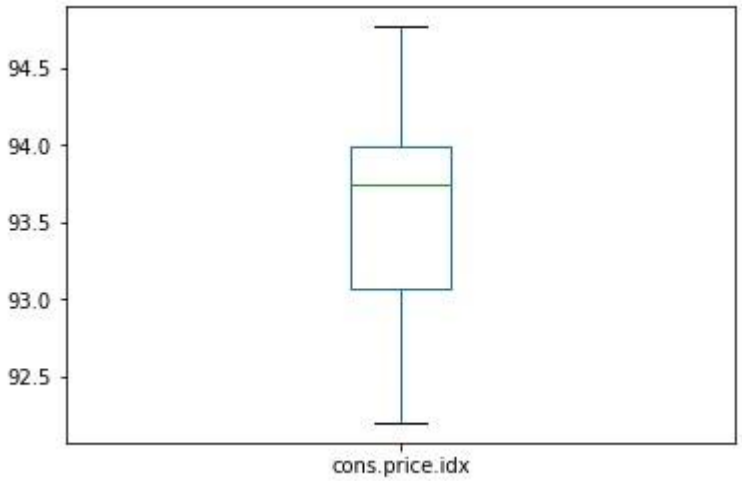


The Bar chart is used to represent the Poutcome attribute.

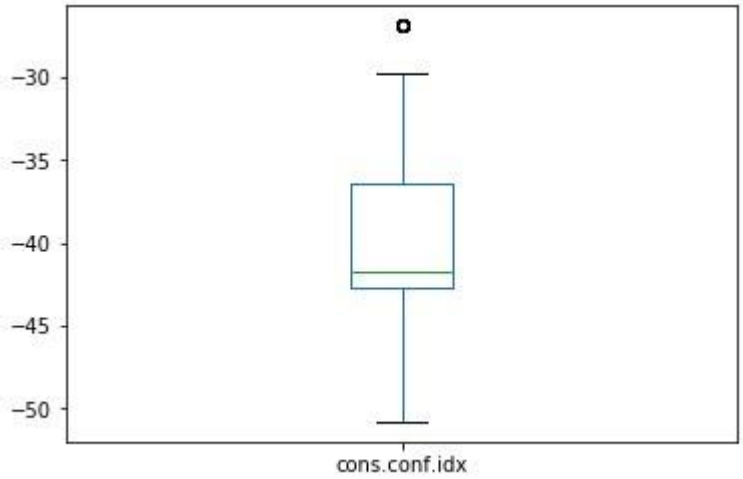


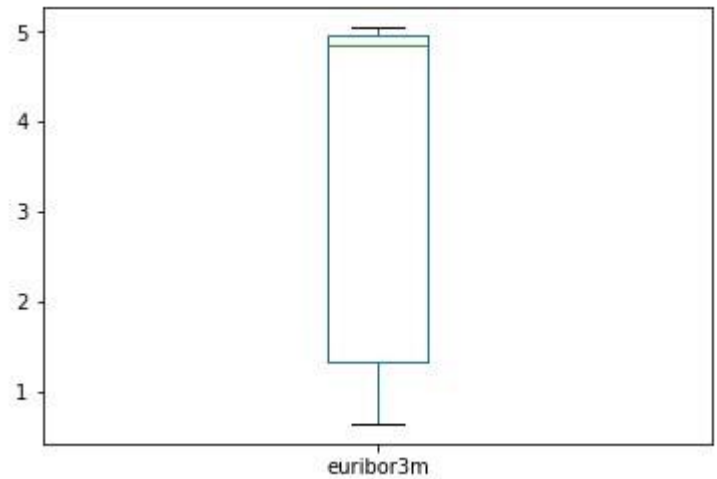
The boxplot chart is used to represent the emp.var.rate attribute.



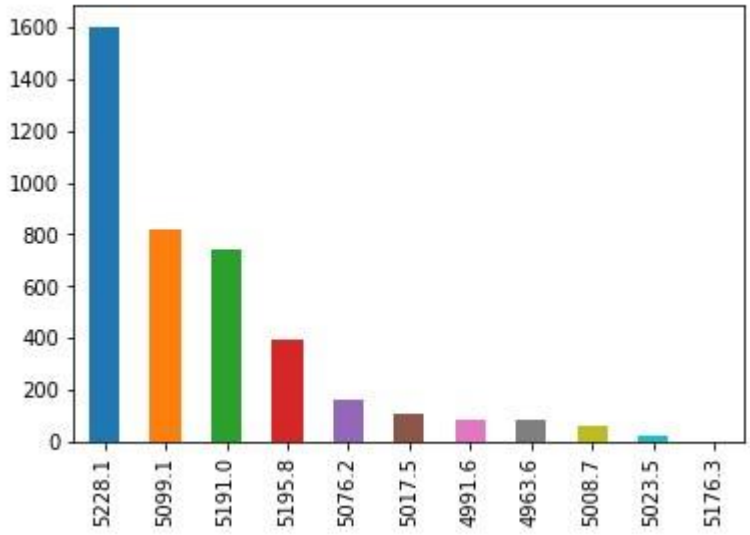


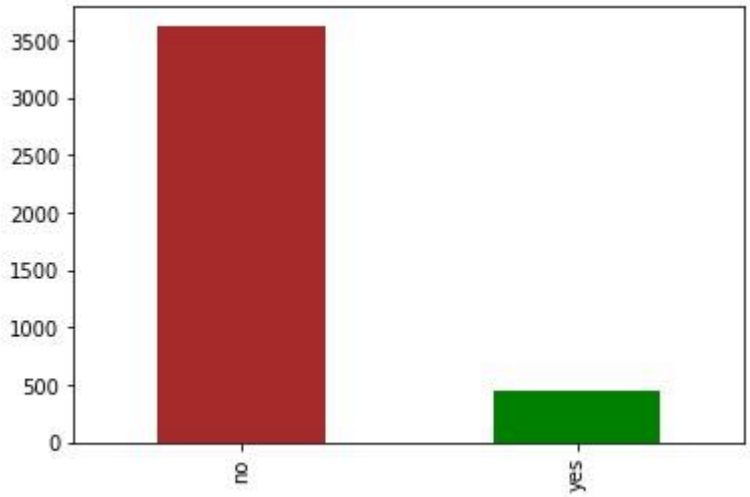
The boxplot is used to represent the cons.price.idx,euribor3m and cons.conf.idx attributes.

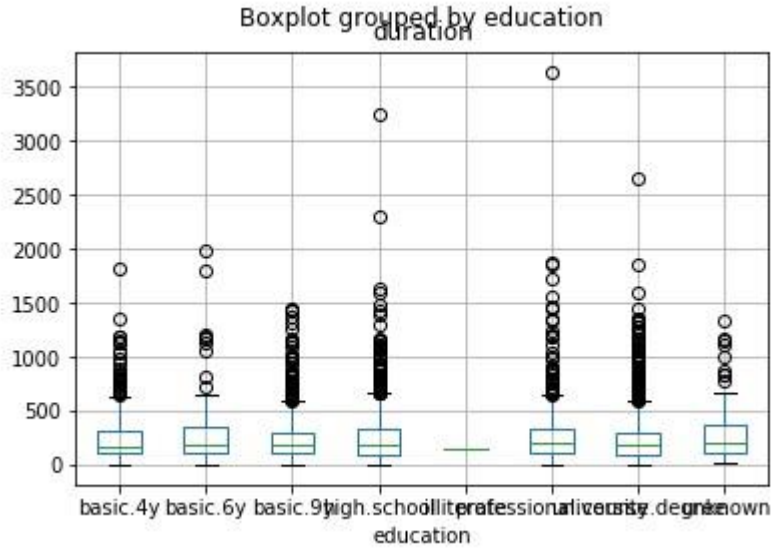




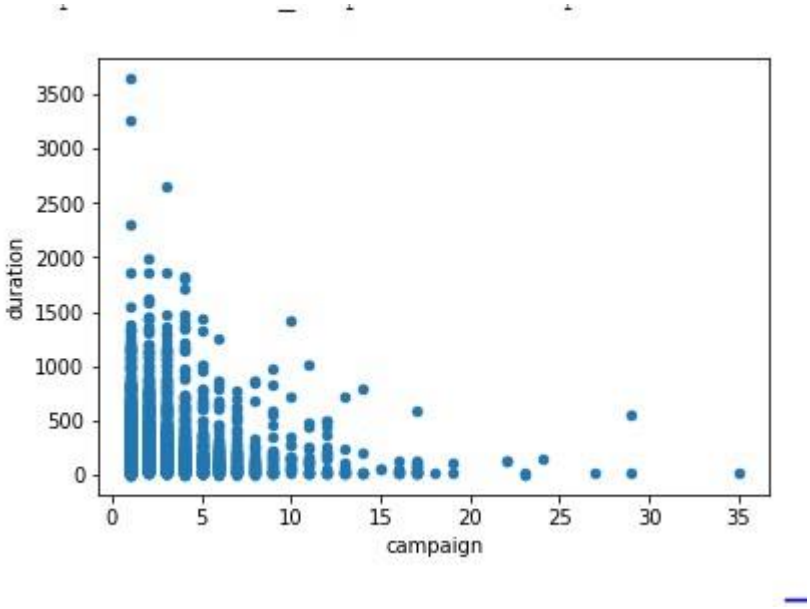
A Bar chart used to depict the below 2 bar attributes.


1600
1400
1200
1000
800
600
400
200
0
5228.1
5099.1
5191.0
5195.8
5076.2
5017.5
4991.6
4963.6
5008.7
5023.5
5176.3

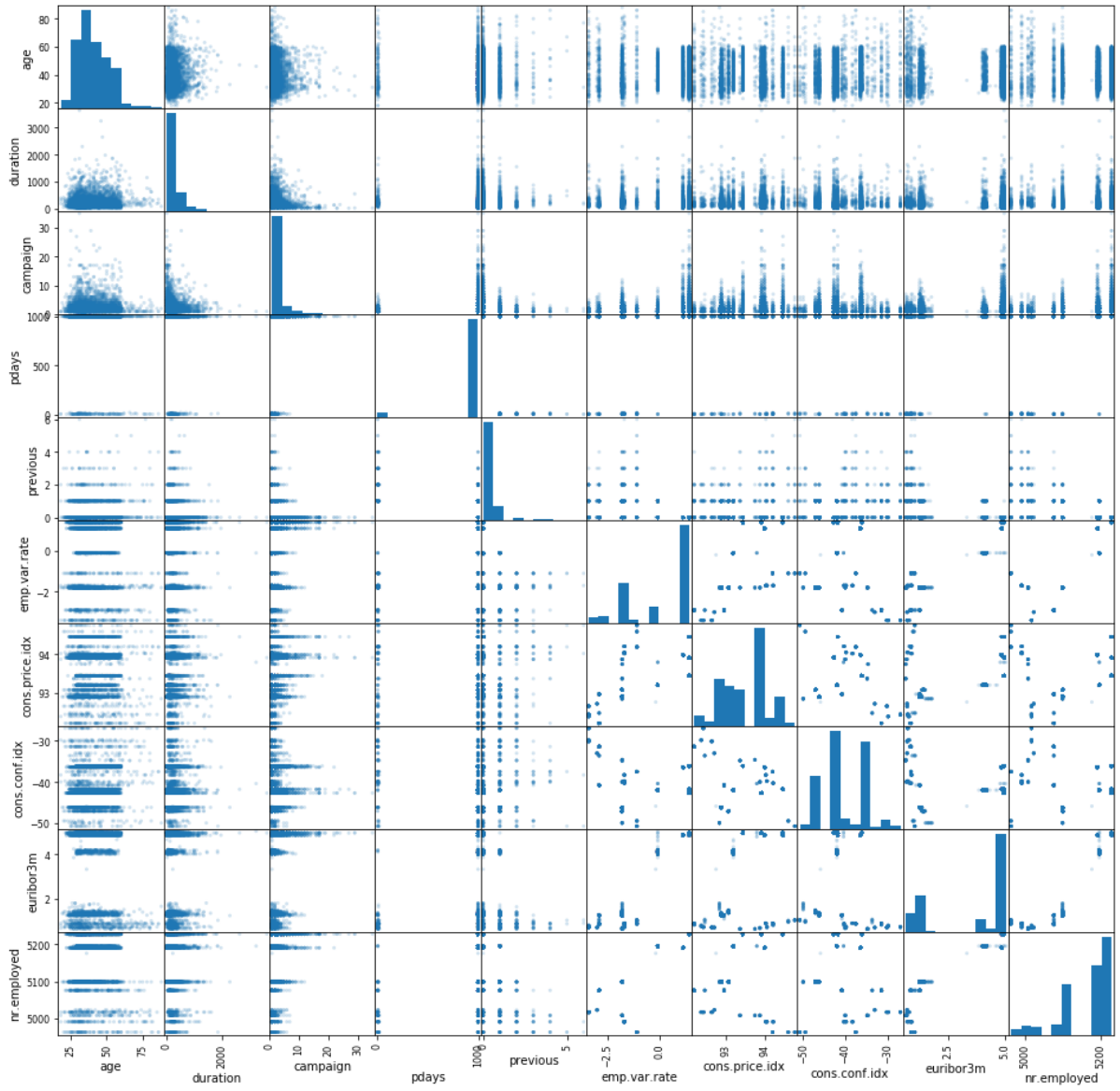

3500
3000
2500
2000
1500
1000
500
0
no
yes



Scatter plot is used to represent the relation between Campaign and Duration



Scatter Matrix


age
duration
campaign
pdays
previous
emp.var.rate
cons.price.idx
cons.conf.idx
euribor3m
nr.employed

DISCUSSION

The reason behind performing this analysis on the data is that the given bank dataset is based on phone calls and may contain errors and wrong information which we must not encounter while performing analysis on the data. So, the part of data cleaning is done to cleanse the dirty dataset so as to convert it to its pure form on which data exploration is done to get valuable insights from it. Data in its raw form is always invaluable unless and until cleaned and analyse.

The advantage of data analysis is that, the relationship between attributes or the relationship between various attributes can fetch useful information upon visualisation which can in-turn be used for bettering business.

Take for instance, the customers who have spent an average time on their call have opted for new term deposits, on the contrary the customers who have spent the maximum and the minimum amount of time in their phone calls have not opted for the new term deposit. This shows that the call duration of the customers needs to be crisp and short in order to convince the customers to open a term deposit. This analysis is shown in a boxplot.

CONCLUSION

The beneath are the focuses that are outlined from the above analysis:

- The customer's confidence index is not a deciding factor in whether a customer will opt for a term deposit or not. The next time the campaign is conducted, it is advised not to consider this factor for analysis.
- On seeing the number of customer contacted, a large amount of customer was contacted in the month of May whereas, extremely few customers were contacted for this campaign in the month of December. The graph showed a decreasing trend. It is suggested that the customers need to be contacted at regular time intervals.
- The customers having a collateral and the customers who don't have collateral are contacted almost equally which shows that the bank does not see the collateral as a criterion for opening bank deposits.
- The level of educational qualification that the customer has does not depend upon the customers confidence index.
- To sum up, the whole marketing campaign is concluded as a failure since a majority of customers did not open term deposits.

Future developments and applications:

A few aspects are not collected properly for data analysis. The previous campaign shows that the data is not collected properly as a majority of it remains unknown. Going forward, this data can be further used for business improvements.

REFERENCES

1. https://en.wikipedia.org/wiki/Euribor
2. https://pandas.pydata.org/