Practical Data Science Assignment 1

S3697150

Data Preparation

1. I used:
   import pandas as pd
   bank = pd.read_csv('Bank.csv', sep = ';',decimal = '.', header = 0)
   to load the Bank.csv file, separated columns by a semicolon, header = 0 so the first row will be the names of attributes.

   Reference: lecture notes week3 – data summarisation, slide 10, Yongli Ren.

2. By checking .dtypes I found a few wrong datatypes, further checking into each column by .value_counts() showed the source data has incorrect double quote which caused values squeezed and shifted to the wrong columns.

   I open Bank.csv and read in python as a string, replace the incorrect double quote in string, open a csv file with new name, write the replaced string to this new name and close it.

   Read the new csv file in python as step 1. Then check values for each column using .value_counts(), all incorrect quotes are removed, some values shifted to the proper columns. Some values are still squeezed but now are due to typos.

   Reference:
   - *https://stackoverflow.com/questions/23903094/how-to-import-data-from-a-csv-file-and-store-it-in-a-variable?utm_medium=organic&utm_source=google_rich_qa&utm_campaign=google_rich_qa*
   - *https://stackoverflow.com/questions/12277864/python-clear-csv-file/12277993?utm_medium=organic&utm_source=google_rich_qa&utm_campaign=google_rich_qa*

3. I changed the following typos while all other values in each column had been verified as correct by .value_counts():
   bk.loc[bk.marital == "divorceded", 'marital'] = "divorced"

```
bk.loc[bk.education == "basic.6yes", 'education'] = "basic.6y"
bk.loc[bk.default == "noo", 'default'] = "no"
bk.loc[bk.housing == "yess", 'housing'] = "yes"
bk.loc[bk.duration == "-", 'duration'] = ""
bk['duration'] = bk['duration'].apply(pd.to_numeric)
```

Reference: Tute lab2 materials

4. I used bk['job'].str.strip() to strip extra whitespaces for each column.
   Reference: Lecture notes week 2- Data Curation, slide 36, Yongli Ren.

5. I used bk['job'].str.lower() to change every column's values to lower case.
   Reference: Lecture notes week 2- Data Curation, slide 37, Yongli Ren.

6. For sanity check on numeric variables, I can use .describe() to view max and min values, but I wanted to try sort_values() and printed the first 10 rows and the last 8 rows, from there I can see if there is any impossible values for age. There were two impossible max age values which I masked them to None by: bk.loc[bk.age >= 100, 'age'] = None.
   Then I checked whether the rest of the numeric variables have negative values where it should not be, eg. bk.loc[(bk['duration'] <= 0)]
   For categorical variables, there was not any impossible values.

   Reference: Tute lab2 materials

7. From the earlier steps like .count() or .info() it is already obvious that missing value exists. I used .isnull() to check where are the missing values in the dataframe. Then I filled 13 missing values with column-wise mean:

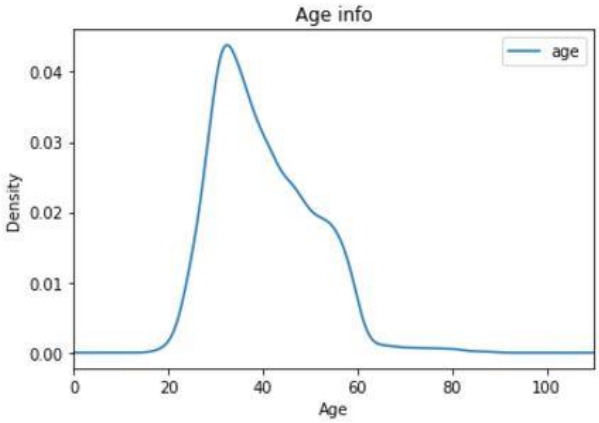
   ```
   bk['age'].fillna(bk['age'].mean(axis = 0),inplace = True)
   bk['duration'].fillna(bk['duration'].mean(axis = 0),inplace = True)
   bk['cons.price.idx'].fillna(bk['cons.price.idx'].mean(axis = 0),inplace = True)
   bk['euribor3m'].fillna(bk['euribor3m'].mean(axis = 0),inplace = True)
   bk['nr.employed'].fillna(bk['nr.employed'].mean(axis = 0),inplace = True)
   ```

   Reference: Lecture notes week 2- Data Curation, slide 50, Yongli Ren.

Data Exploration

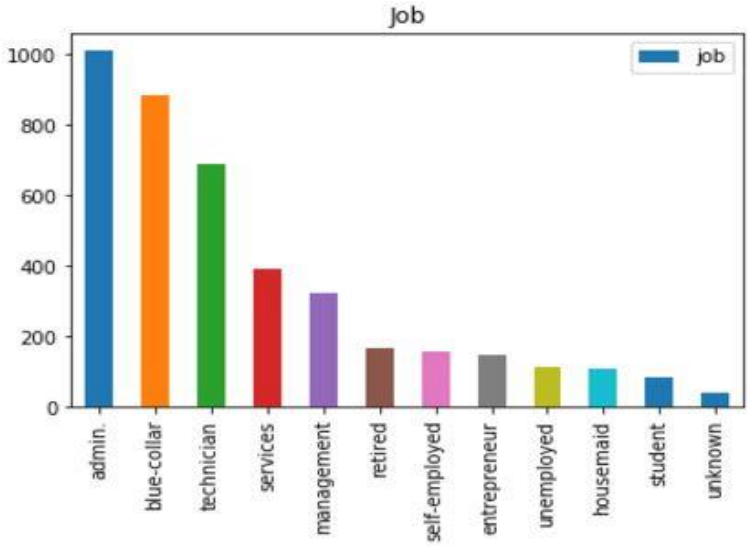
1. Choice of graph type:

    1. Age.



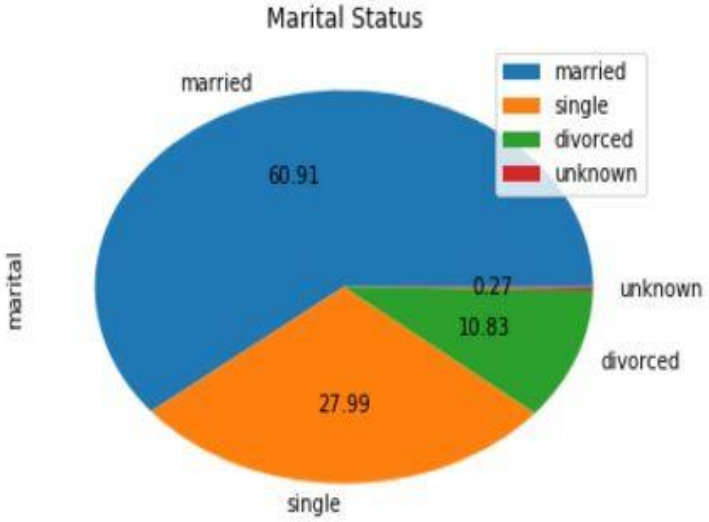
A density graph can obviously show distribution of all clients' age, the sharp peak indicates the majority clients involved in campaign are around 30's.

    2. Job.



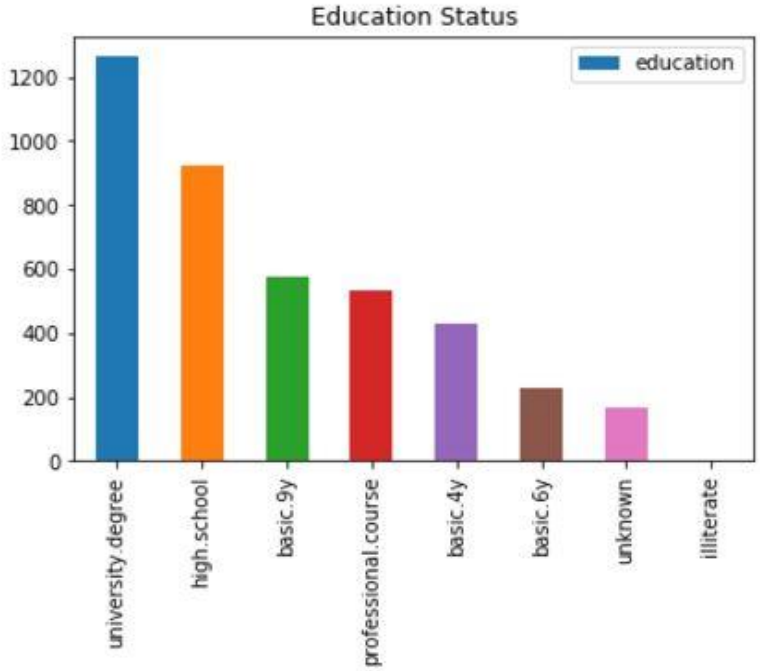
Since there are 12 categories in the Job attribute, a bar chart can clearly label each type of job with distinguished colours, the frequency of each job category is also clearly shown.

3. Marital Status.



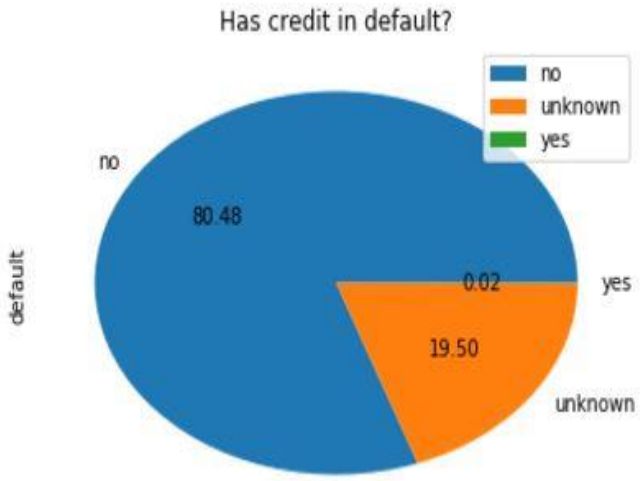
A pie chart represents the proportion of each marital status within the population of clients, from the graph percentage of each portion can be read out. From this plot, correlations with other attributes can be explored easily.

4. Education



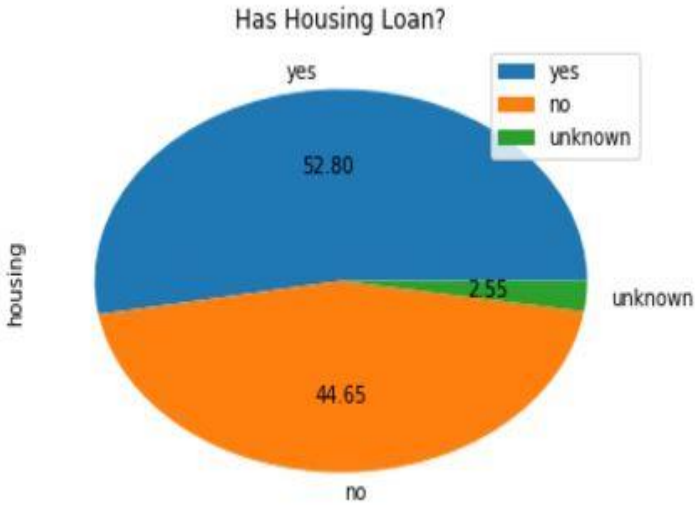
The education status is best interpreted by a bar chart since the education attribute as distinct categories, the height of each bar indicates the number of clients who had undertaken that corresponding level of education. For example the bar chat tells that the most number of clients undertook university degree.

5. Default.



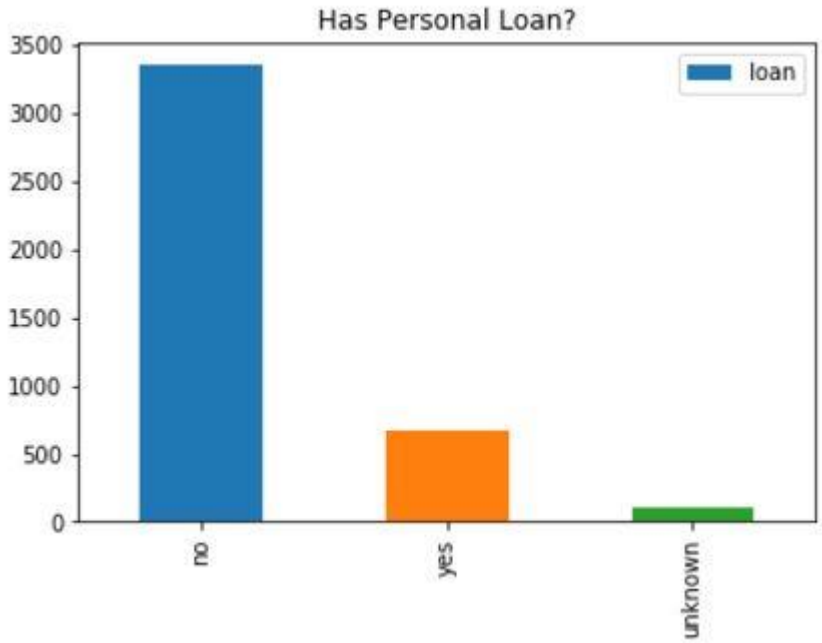
The pie chart indicates how many clients have credit in default, again clearly shows the percentile components of client situations. There is one observation of yes (has credit) calculated by .value_counts() in Task 1, whereas the majority of clients do not have credit and a small proportion of 19.5% of clients don't have this information available.

6. Housing.



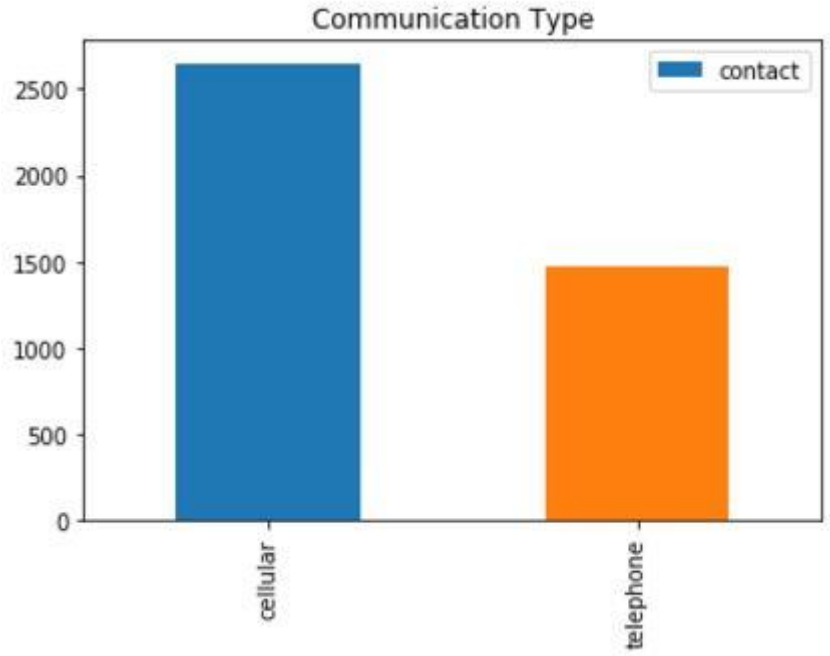
A pie chart can also clearly indicate the rations of clients between who have and who do not have a housing loan, and between those who provided unknown information on whether they has a housing loan.

7. Loan.



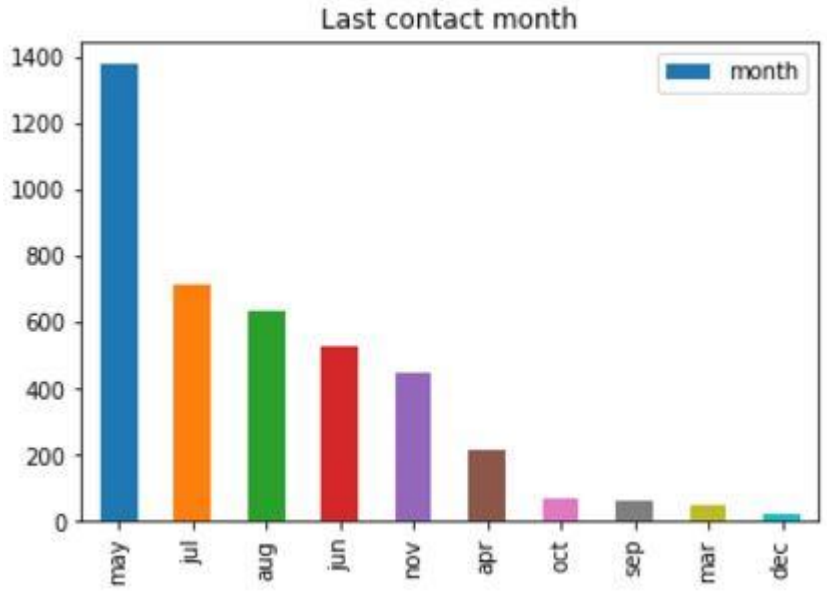
A bar chart can also clearly indicate the proportions of clients who has a personal loan against who does not have, and those unknown.

8. Contact.



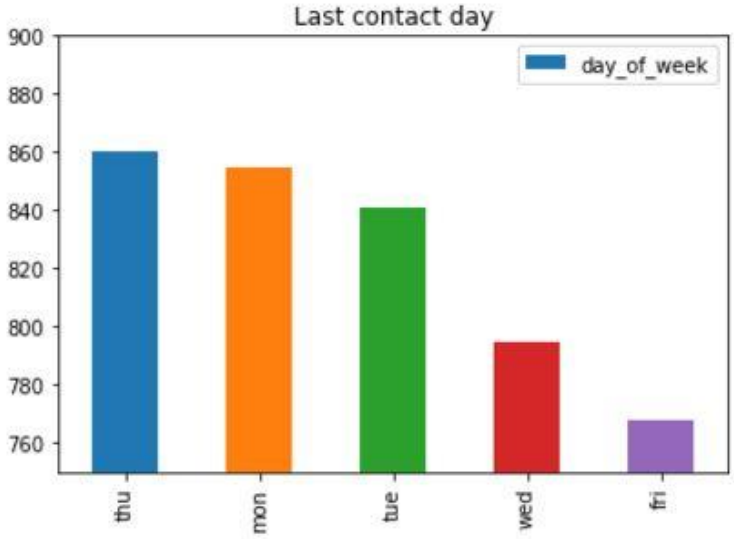
A bar chart can best interprets the ratio of communication type used during a campaign.

9. Month.



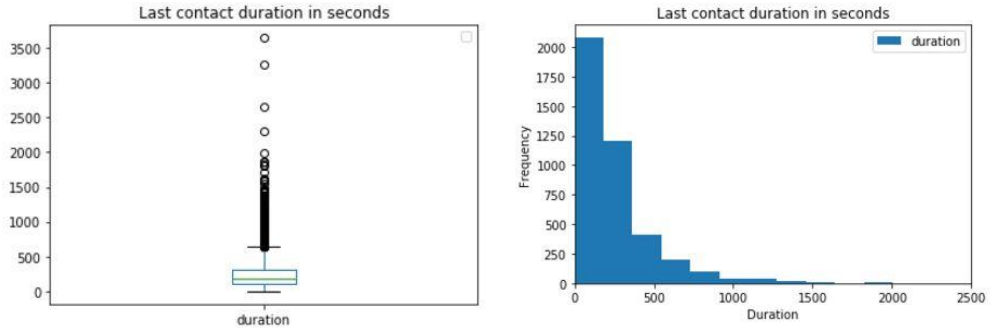
A bar chart can not only indicate the frequency of campaigns took place in the respective months, but also shows during May where the most number of times of campaigns were conducted, and there are two months (January and February) where no campaigns conducted at all, this may deduce some market information.

10. Day_of_week.



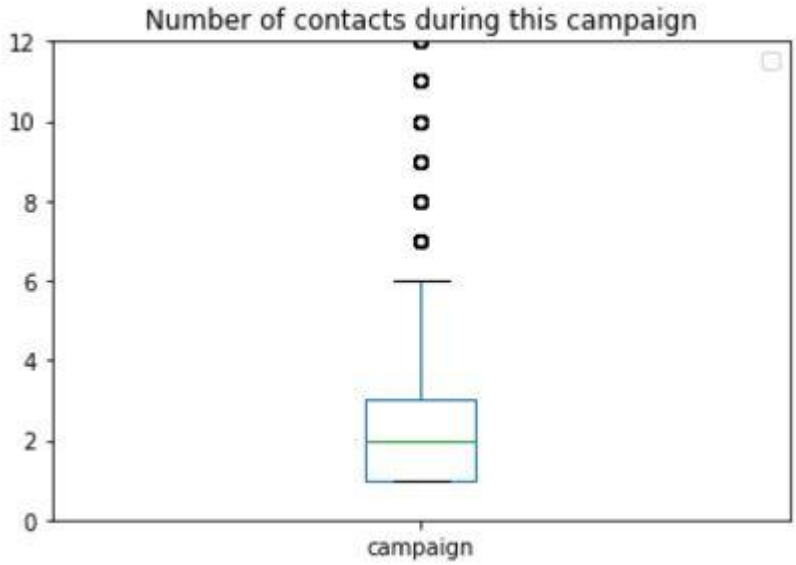
This bar chart apparently shows the time of contact from last campaign was mainly on Thursday, and there are very few opportunities to contact clients on Friday in comparison. This plot can help decide the optimal day to conduct future campaigns to increase success, concluded from past habits of clients.

11. Duration.



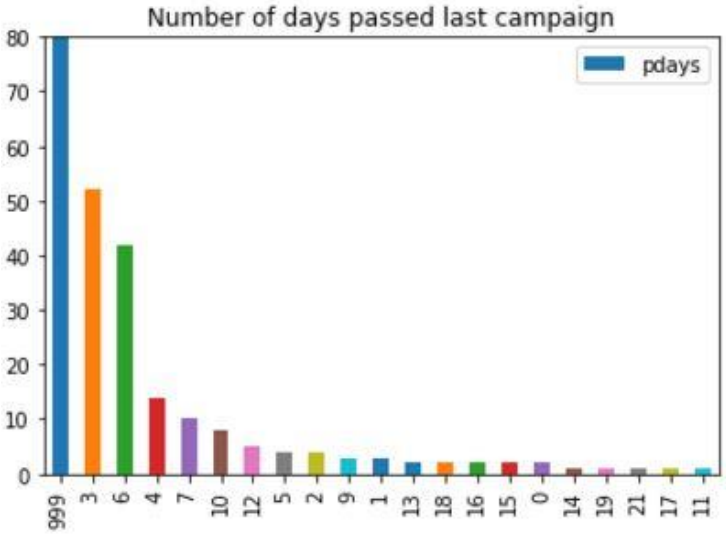
The duration in seconds of the last contact to a client was firstly plotted by a boxplot, which shows a lot of outliers because there are rare occasions that a contact to a particular client was really long (3643 seconds equivalent to an hour). When the duration is illustrated in a histogram, the pattern became clearer. Most of the contact ended within 200 seconds, majority (or 75%) of contacts finished in around 500 seconds.

12. Campaign.



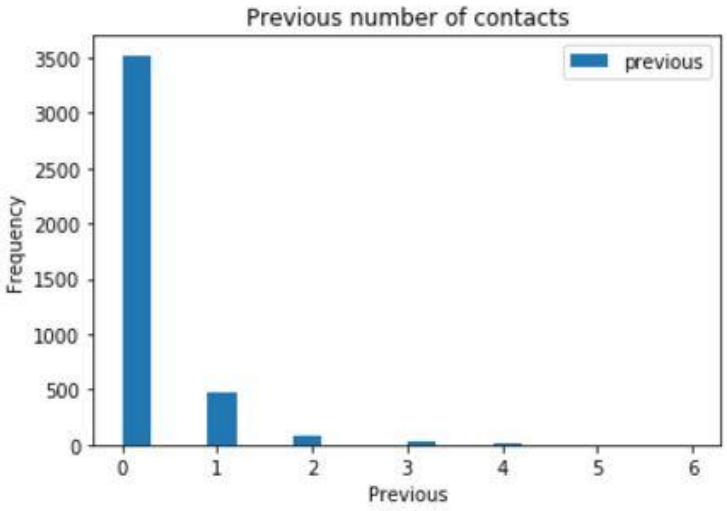
A boxplot best describes the number of contacts made to a client during this campaign. All clients had at least been contacted once, each client had been contacted twice on average, some particular clients have been contacted for over 7 times in one month.

13. Pdays.



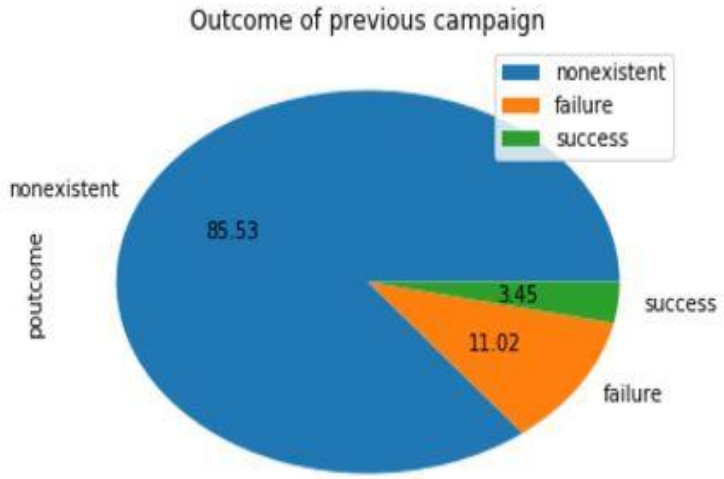
A bar chart can best show the distribution of the number of days passed the last campaign when a client was contacted. 999 days indicates the most representative group in all clients are those who were not previously contacted, probably they are new clients.

14. Previous.



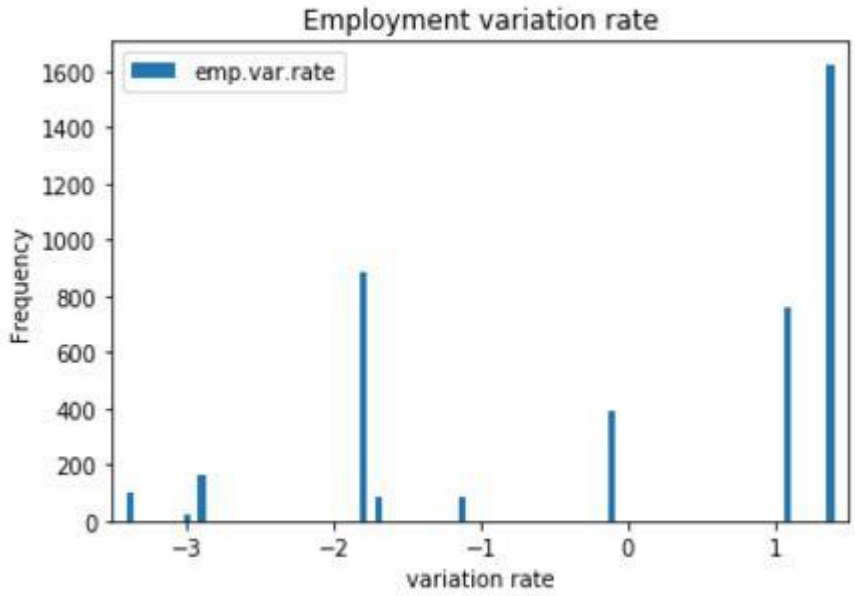
A histogram clearly shows that most clients were not previously contacted before, again indicating these clients are probably the new customers. The histogram also clearly shows the frequency of number of times of contact made before this campaign to the same client, the second biggest group of clients were previously contacted once.

15. Poutcome.



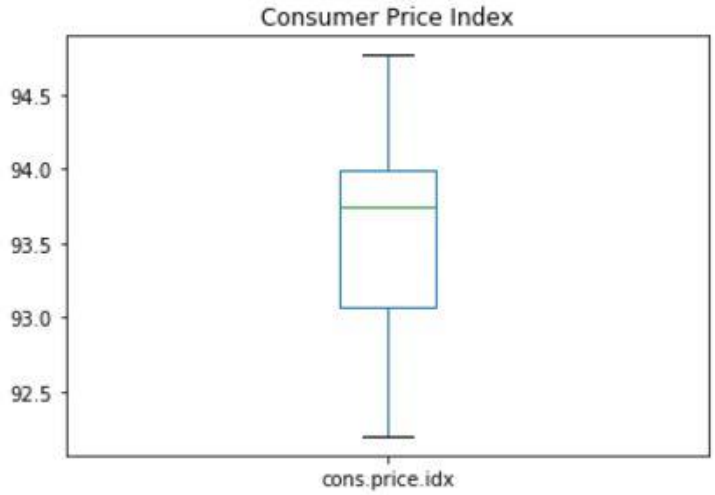
This pie chart makes the results from previous campaign extremely obvious. 85.53% of contact were non-existent which emphases this majority of clients are emerging customers.

16. Emp.var.rate.



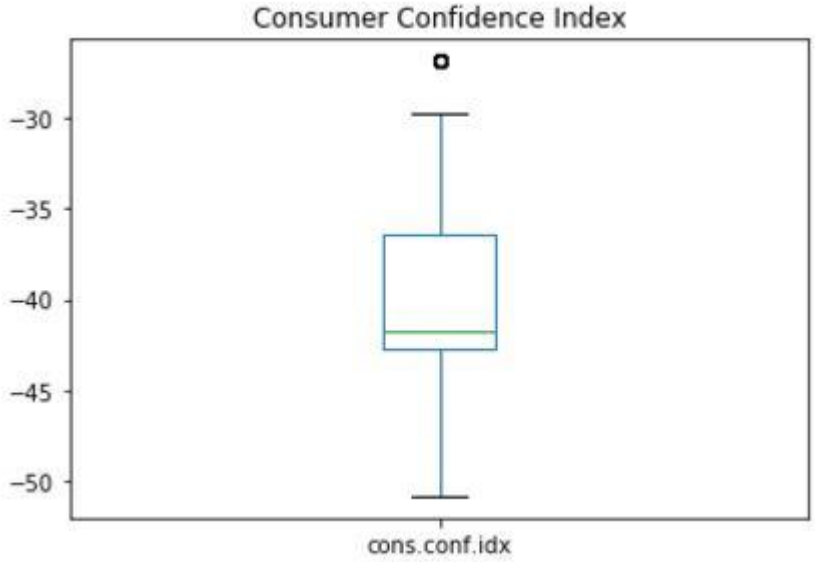
The histogram gives a good presentation of employment variation rate, it clearly indicates the employment situation of each client. The most number of clients have a good employment, whereas there are also a number of clients who are unemployed?

17. Cons.price.idx.



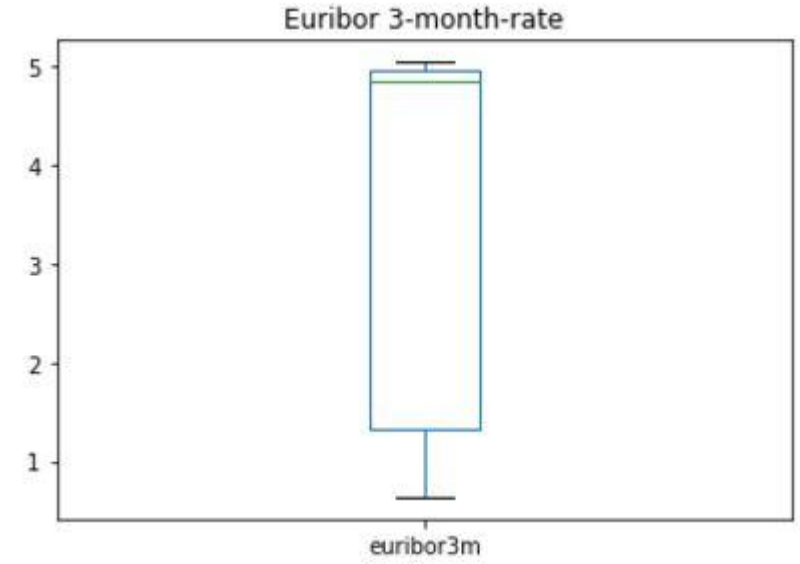
Boxplot is used for illustrating the distribution of the consumer price index variable. Since box plot can best summarises the mean, max, min and other statistics for numerical attributes, the same plot will be used on depicting the rest indexes and rates variables.

18. Cons.conf.idx.



Boxplot can clearly show how strong the clients are confident as consumers, even the most confident outlier client has a negative consumer confidence index, indicating unhappy consumer psychology.

19. Euribor3m.

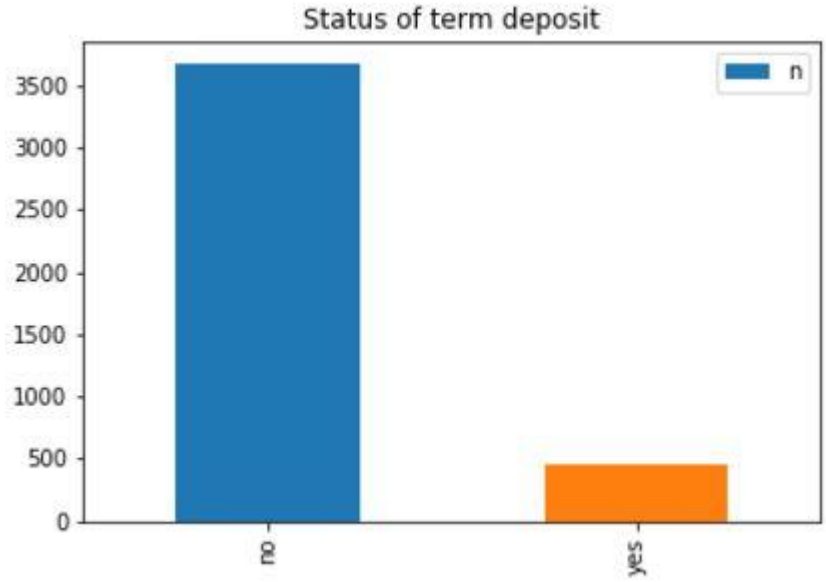


The boxplot clearly shows that the daily European euribor interest rate (3 months) varies between numbers that are below 1 and maximally below 5.

20. Nr.employed.

This boxplot presents the number of employees per quarter at this institution, which varies roughly between 4950 and 5500 people, minimally above 4950 and maximally below 5500 employees.

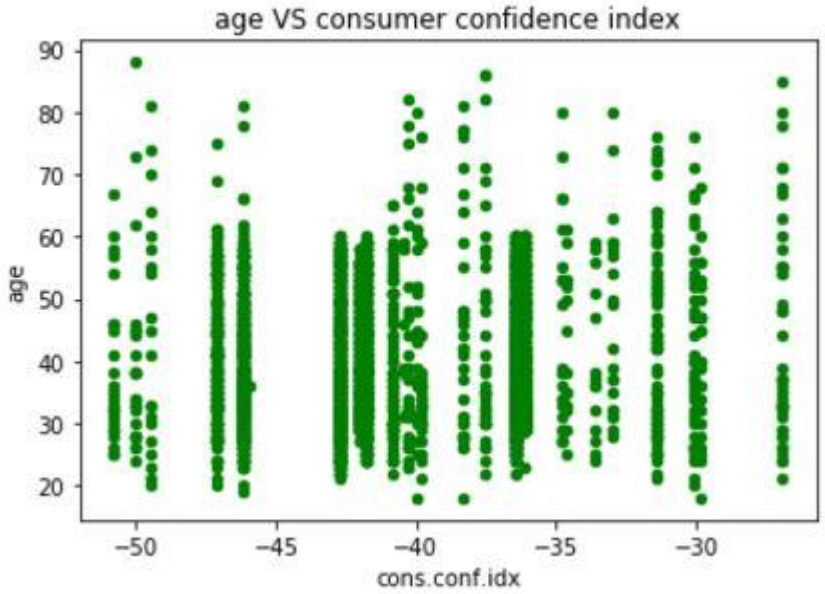
21. Y.



This bar chart represents how well clients subscribed to a term deposit after marketing campaigns. With only two types of results, bar graph can confidently interprets whether the outcome of campaign was successful or failure.

2. Investigating relationships:

    1. Relationship between age and consumer confidence index.



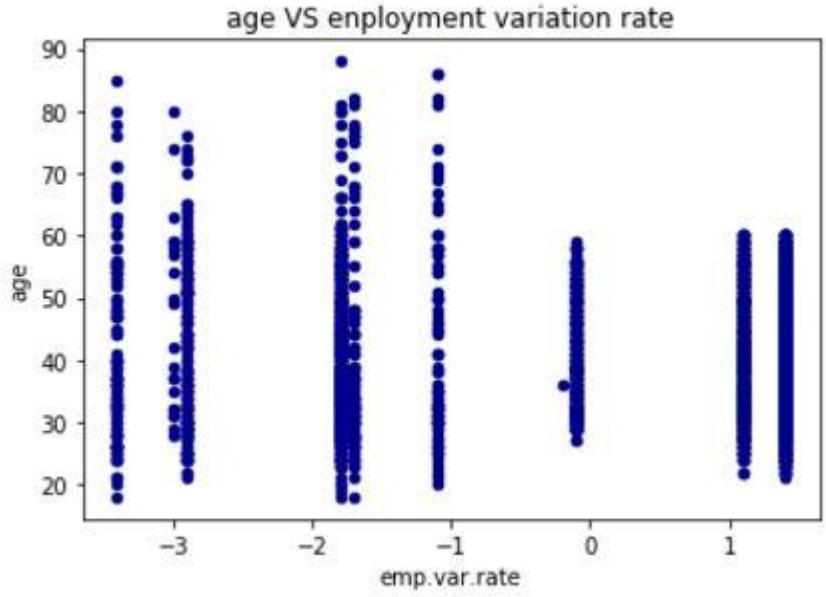
I would investigate if there is any relationship between the clients' age and the degree of their consumer confidence (whether the clients prefer to make a term deposit).

Supposed that for clients whose age is between 25 to 45 are more likely to invest in term deposits, compared to clients at other age.

The scatter plot of age against cons.conf.idx reveals that the most number of clients (highest density in the graph) hold average views for choosing a term deposit and their age ranges from 25 to 58, so they can potentially become a consumer of term deposits which is consistent with my assumptions.

There is another interesting pattern that both the most confident and the lease confident group of clients are aged between 25 to 35, the reason might become clearer if further analysis can be done on the relationship with a third attribute which affects this pattern most significantly.

2. Relationship between age and employment variation rate

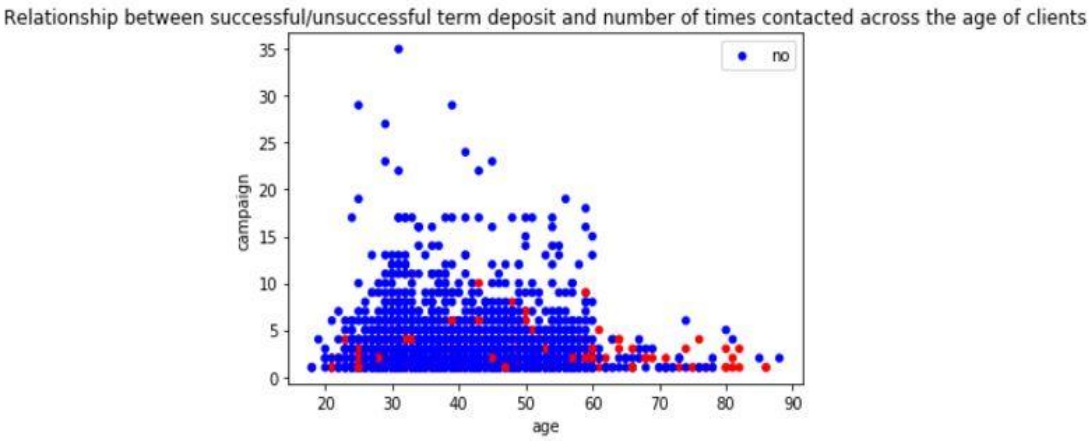


I want to investigate whether the employment variation rate associates with age, assumed those clients who were confident consumers should have positive employment variation rate and this should align with the age group explored in previous graph.

In this scatter plot, clients who are between 25 to 60 years old generally have a steady or positive employment variation rate, so that's why they are also the age group who are more consumer-confident than others.

But this age group also contains a subpopulation having negative employment variation rate, this may explains this subpopulation is more negative or less confident than average in the previous graph.

3. Relationship between age and campaign.

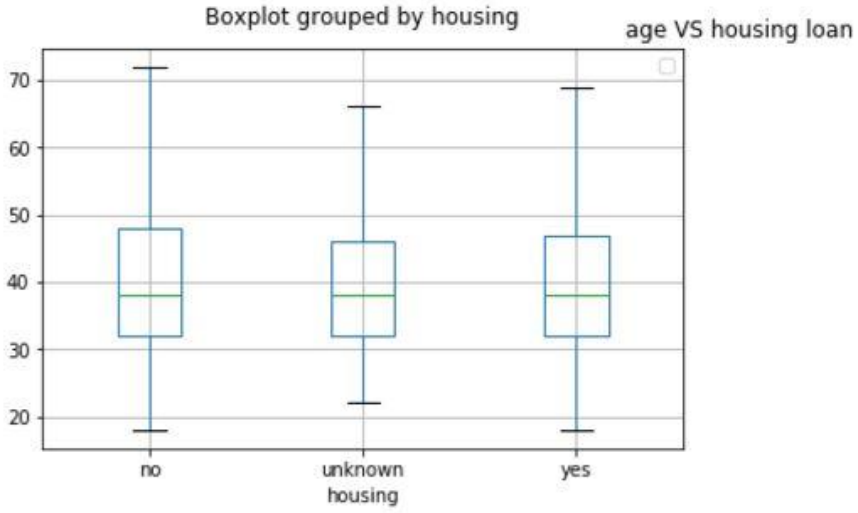


I firstly investigate the relationship between age and campaign (the number of times that a client was contacted), because there may be some age range that are more open to marketing campaigns, hence sales campaigns could be made easier for employees.

From the blue scatter plot it is easy to see that clients in their 30's and 40's are the main recipients of the campaign. But the higher-than-average number of contacts was made to the clients aged 40 to 60, they are happy to make contact, it can be deduced that people in this age range might have the real need to make some term deposits, so they could become targeting customers.

Then I labelled all successful outcomes in red. These clients finally turned into a term deposit subscriber. The successful rate is low but most of the subscribers are aged 60 to 80 even though they cost so few times of contacts. From earlier distribution graphs for single attributes, a conclusion was made that most of the clients this time are new customers since they haven't been contacted previously, so it turns out that the elderly people are easier to accept phone call campaigns and accept the product on offer.

4. Relationship between age and housing loan.



I wanted to see some relationships between age and hosing loan owners, but the mean, median, min, max of age are fairly equal for clients who has or hasn't got a housing loan, so there is not a really obvious relationship between these two attributes.

3. The scatter matrix for all numeric attributes: