Capstone project 1: Bank Marketing

Report on Data Storytelling

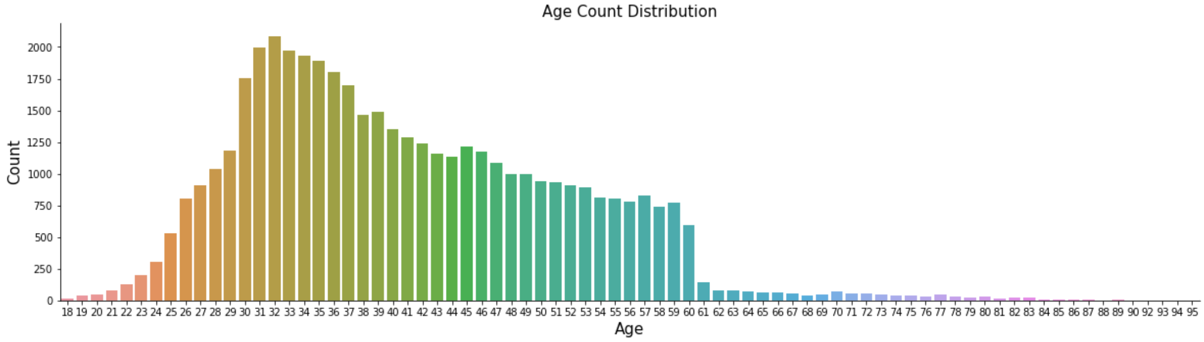
Overview:

This document provides a report on the data storytelling aspects of the Capstone project on Bank Marketing.

Bank Clients Information -

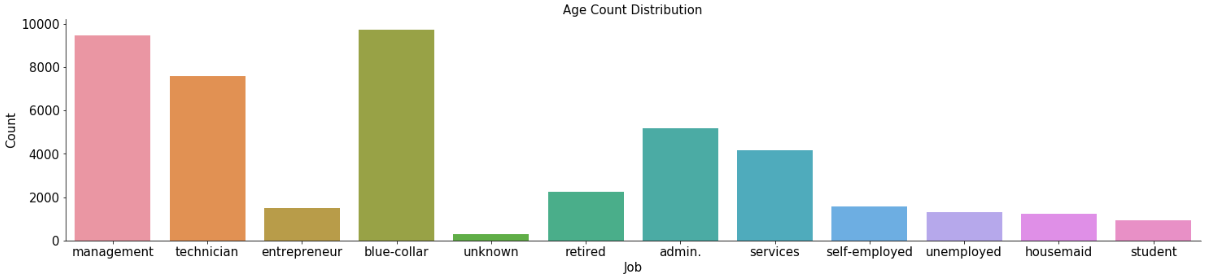
AGE - age of bank clients.

Countplot from seaborn gives a very good explanation of how the age of the clients is distributed from the minimum age of 18 to the maximum age of 95. Apart from countplot, histogram and boxplot were plotted.



Even though the minimum and maximum age are largely at the end of the graphs, most of the age groups are centered in the middle. The mean age is about 40yrs and the standard deviation is 10 years. Most of the clients are middle-aged.

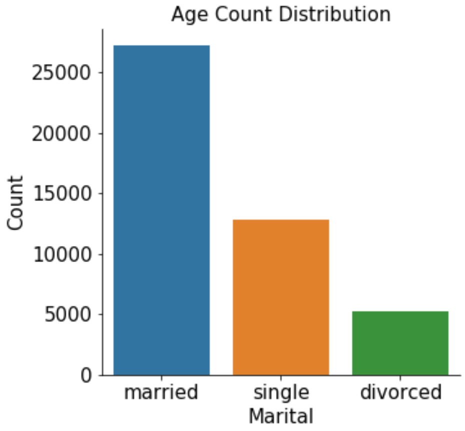
JOB



The job categories - management, technician, and blue-collar have a high count of jobs among all other job categories.
The proportion of bank clients who are getting low or no income (student, housemaid, unemployed, self-employed, unknown, entrepreneur) is very less

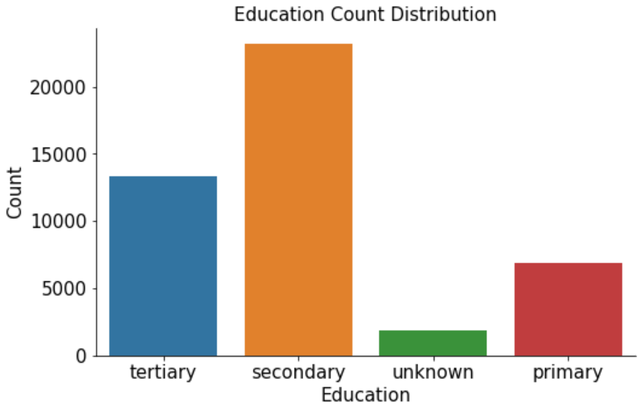
compared to that of clients who are regularly paid. Maybe they are less prone to open a bank account.

MARITAL - marital status



Married clients having a bank account is higher than single and divorced. Since the data is not that skewed, the data may represent the general population.

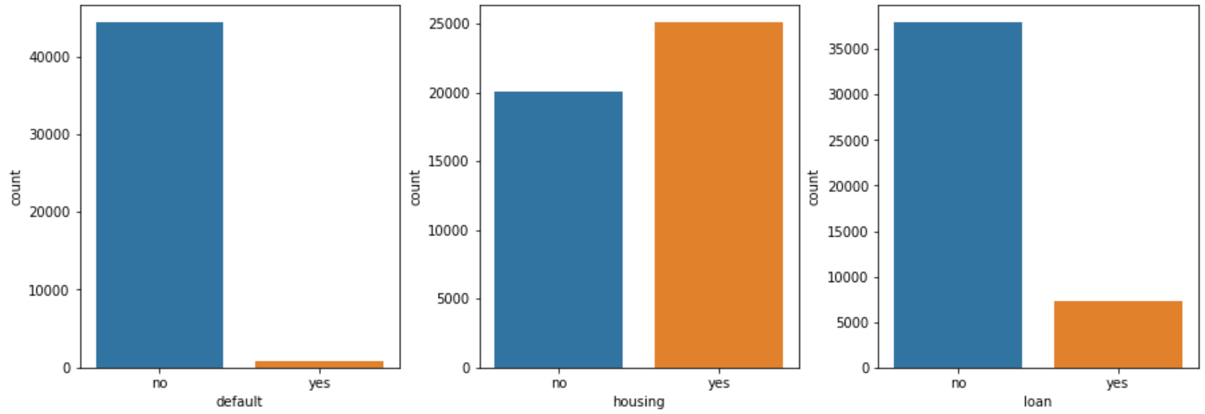
Education



Education is grouped into 4 categories- primary, secondary, tertiary, and unknown. Among these 4 categories, secondary and tertiary categories are on the higher side. The higher count of secondary and tertiary education can be because of being able to better provide for themselves economically and hence the need for a bank account.

Default, Housing, Loan -

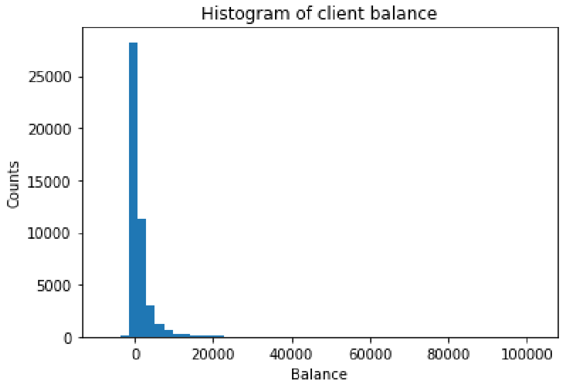
Whether the client has credit in default, whether the client has a housing loan, and whether the client has a personal loan.

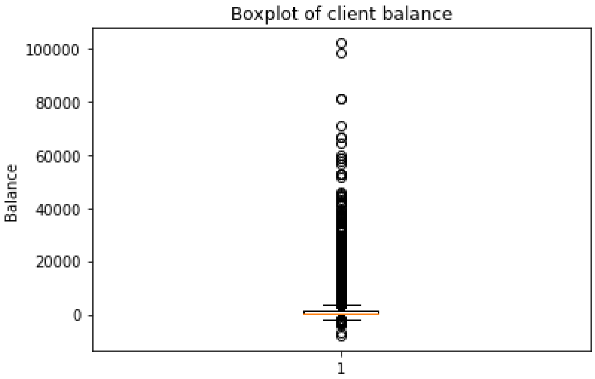


It is understood from data visualization that most of the clients do not have any credit in default and do not have a personal loan.
The clients having housing loans are nearly equal to the clients who do not have housing loans.

Balance

- average yearly balance, in euros. This is one of the numerical columns of the data set.



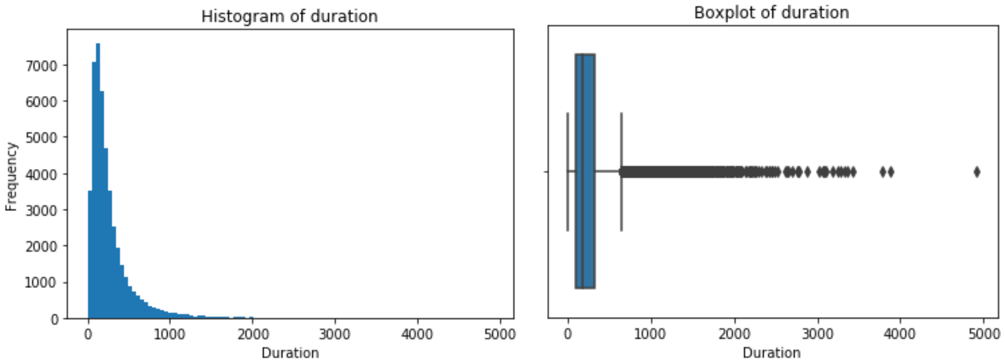


Balance ranges from a minimum of -8019 to a maximum of 102127. But 25th and 75th percentiles range between 72 to 1428 euros. It can be seen from the histogram that many clients have a negative balance. There are a lot of outliers, but they are kept for further analysis.

The clients who had negative balances were further analyzed. The value_counts() method on the term deposit ‘y’ for clients who had negative balances was performed. Out of 7280 clients, 502 clients subscribed for a term deposit. Hence, even clients who had negative balances are important.

Attributes related to the last contact of the current campaign :

Duration - call duration made by the telephone executive regarding the term deposit.

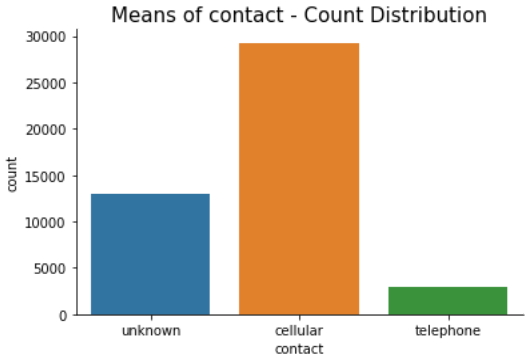


A histogram and boxplot were plotted to see the distribution of the call duration. It is understood that most of the call duration is grouped around the mean duration at around 258s.

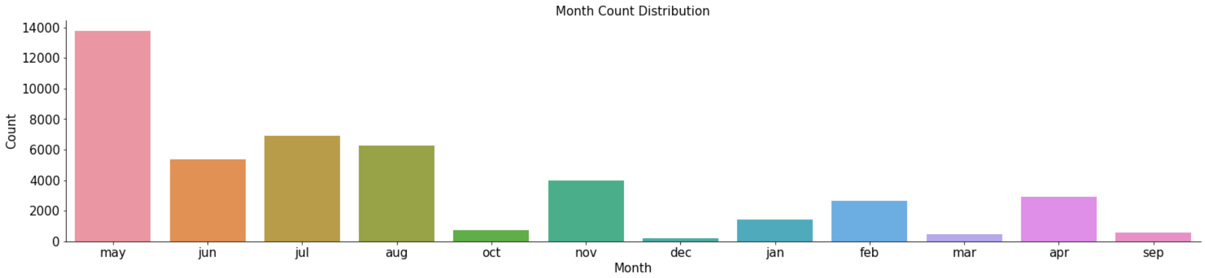
Most of the data is skewed towards the left. With data between the minimum and 75th percentile lies between 0 and 319 seconds. Whereas the maximum duration is 4918 seconds. The data below the 75th percentile for the duration of the call gives a clearer picture. The call duration is usually about 5 minutes. But occasionally it got a little higher. And the maximum value of 4918 s could be an outlier.

Contact, Month, Day

Communication type - unknown, telephone, and cellular.

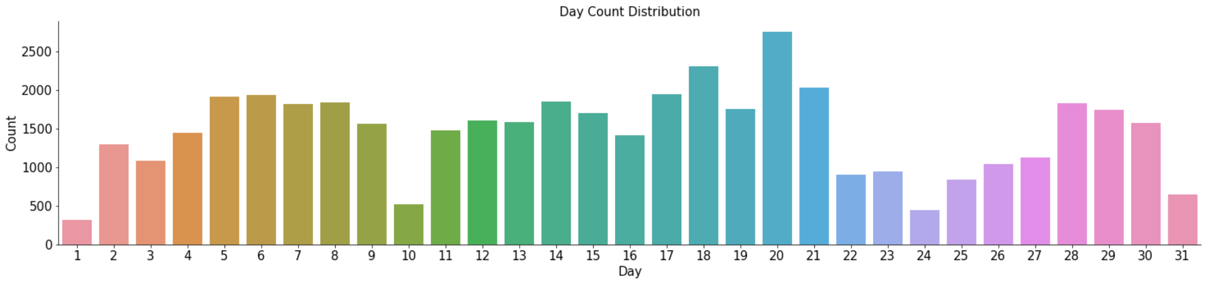


Cellular contact is the highest of the three categories. The higher proportion of cellular contact shows that it is the usual method of contact. And most people have and use cellular telephones rather than a usual (landline) telephone.



- last contact month of the year.

May month has the highest data points. Jun, July, August have the next highest data points. The rest of the months do not have many data points.
Many clients are contacted during these months - May to August. Summer seems to be the usual contact period regarding a term deposit. Maybe because the financial year starts in April and the banks may be under pressure to create a new stimulus. Winter might be a dormant period for contacting clients.

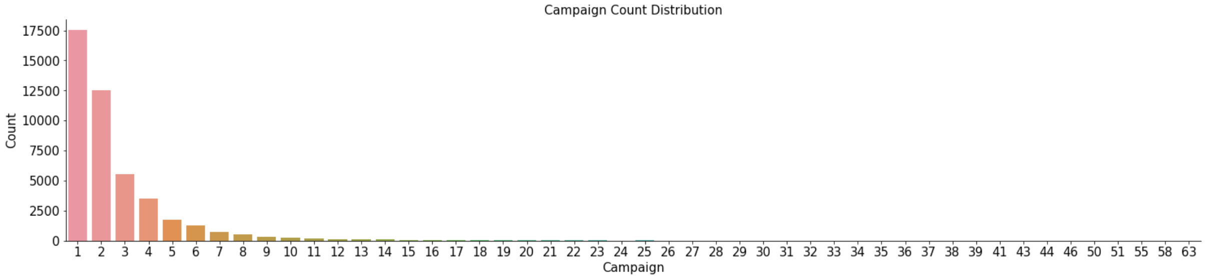


- last contact day of the month.

The last contact day of the month is nearly equally distributed. Among all the days of the month. This means that the clients are equally contacted all the time of the month.

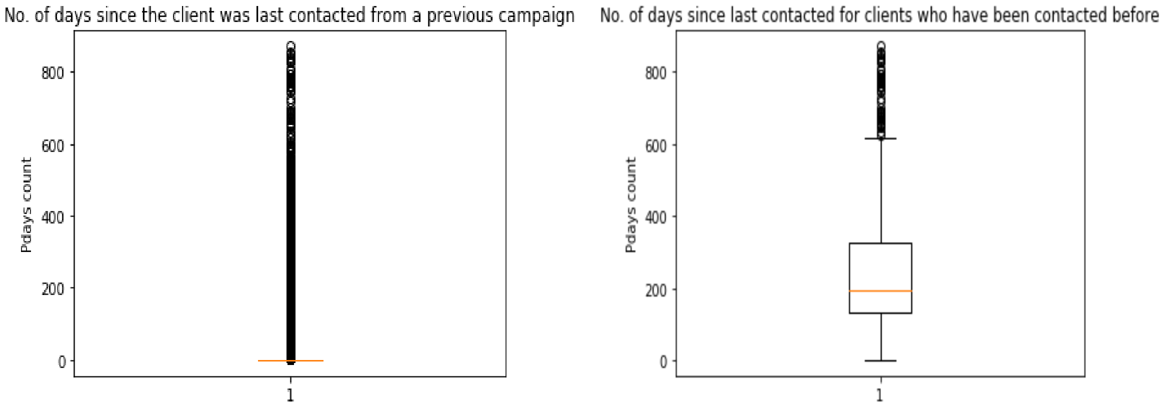
OTHER ATTRIBUTES

Campaign - Number of contacts performed during this campaign and for this client.



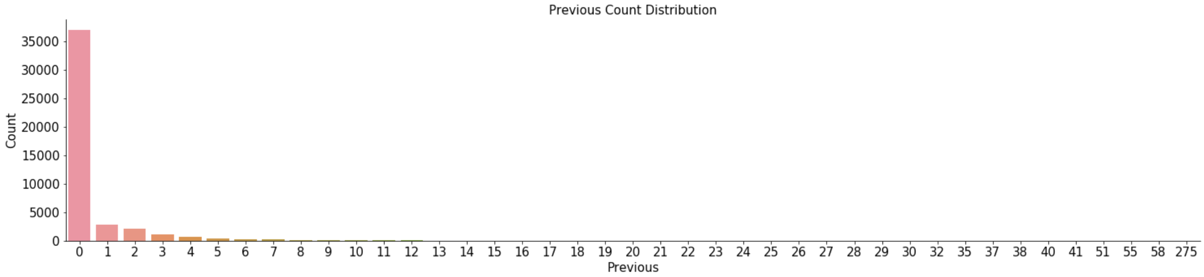
75 percent of the clients were contacted once, twice, or thrice. Making clients who were contacted more than 6 times as outliers. About 3064 outliers are making the total percentage of the outliers to about 6.78%. But the outliers are not removed because the data among the outliers have a significant amount of clients who have subscribed for a term deposit ‘y’.

Pdays - number of days that passed by after the client was last contacted from a previous campaign.



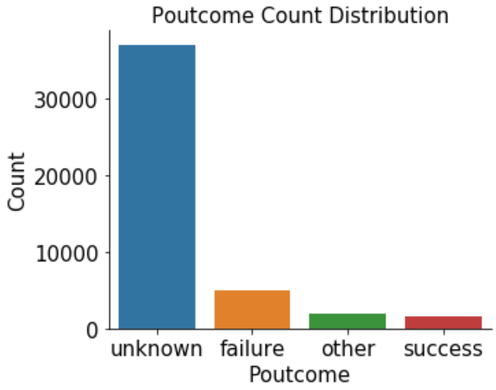
The second figure is a boxplot of the clients who have been contacted before. There are about 36954 clients who have not been contacted by a previous campaign (they are given as pdays = 0). Pdays = 0 data is removed and plotted with a boxplot to get a general sense of the data. Apart from Pdays = 0, the rest of the data is normally distributed.

Previous - number of contacts performed before this campaign and for this client.



Like, the previous attribute, this attribute also has about 36,954 clients who have not been contacted by a previous campaign (denoted by previous = 0). Apart from that, the rest of the data is normally distributed.

<u>Poutcome</u> - the outcome of the previous marketing campaign.
Poutcome is grouped into 4 categories- unknown, other, failure, and success.

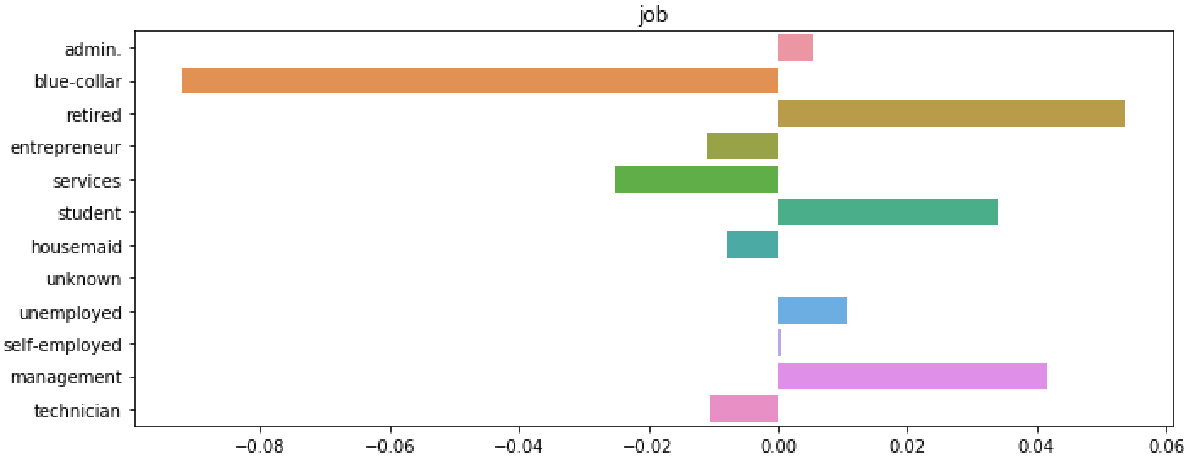


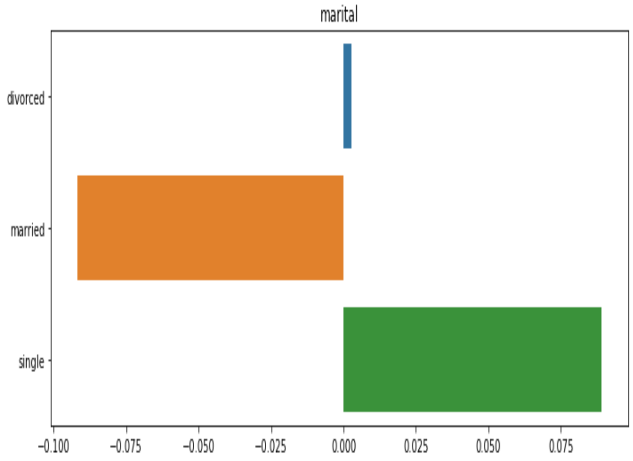
Unknown and other categories are pretty much the same and hence they can be grouped and it has the highest proportion of data points. It is understood from this that the data regarding the previous outcome is not known.

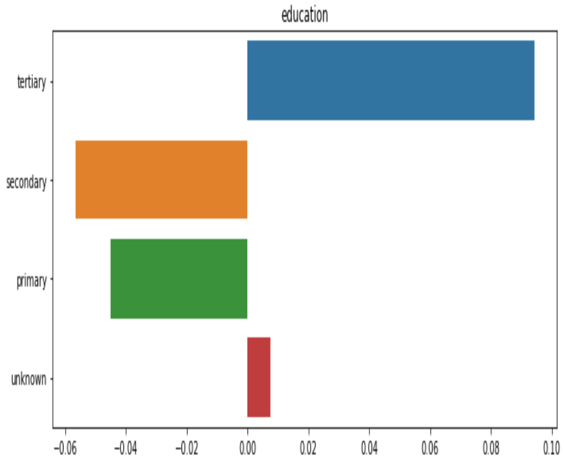
Categorical Columns

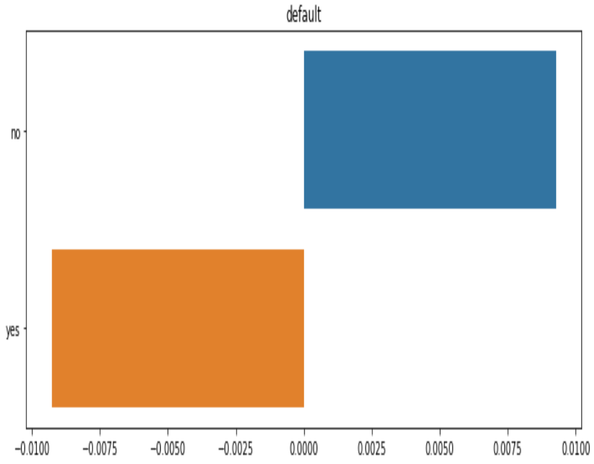
Categorical columns are further treated to analyze every class of each feature. This is done to understand what attributes of the clients or attributes relating the campaign have a higher chance of clients subscribing to a term deposit.

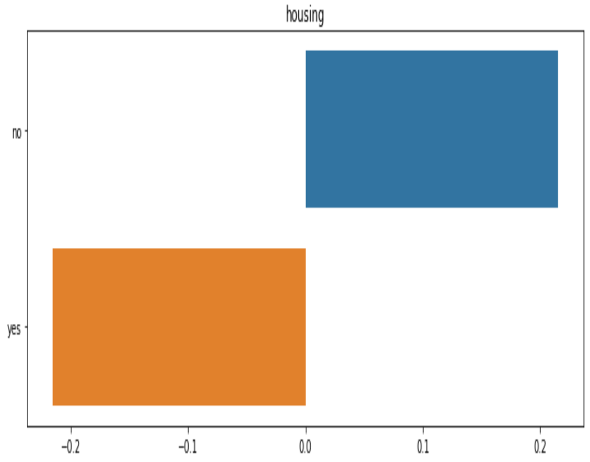
Each class per feature is normalized and the differences are plotted between positive and negative frequencies. Positive values imply this category favours clients that will subscribe and the negative values category that favour not buying the product.

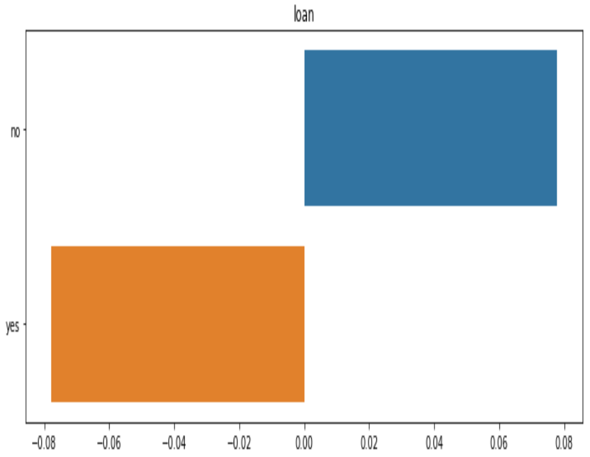


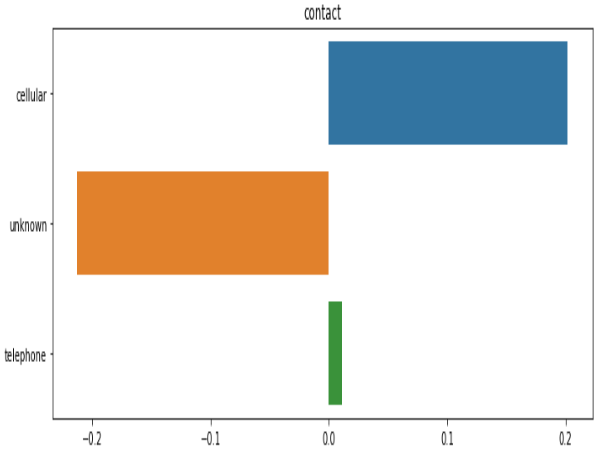

marital
divorced
married
single
−0.100
−0.075
−0.050
−0.025
0.000
0.025
0.050
0.075

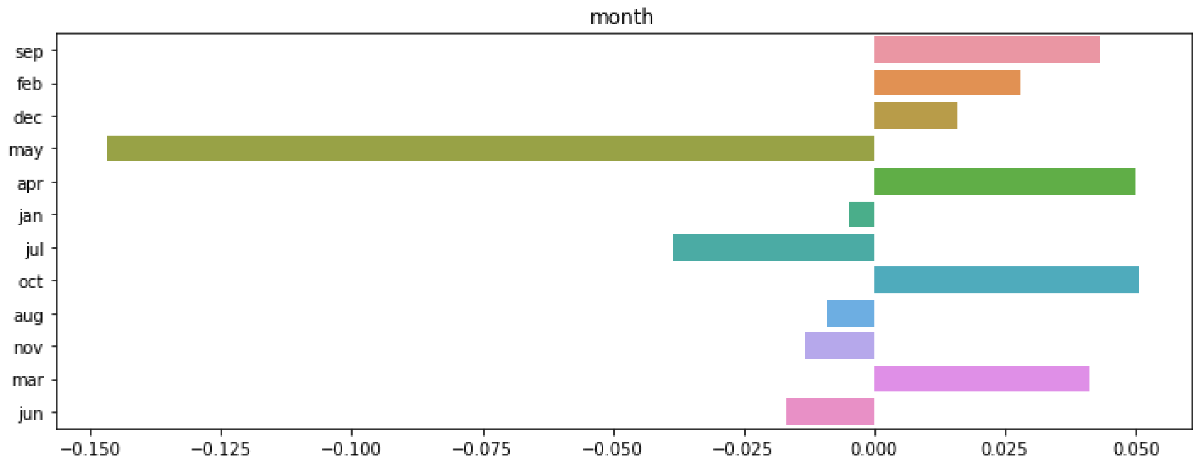

education
tertiary
secondary
primary
unknown
−0.06
−0.04
−0.02
0.00
0.02
0.04
0.06
0.08
0.10

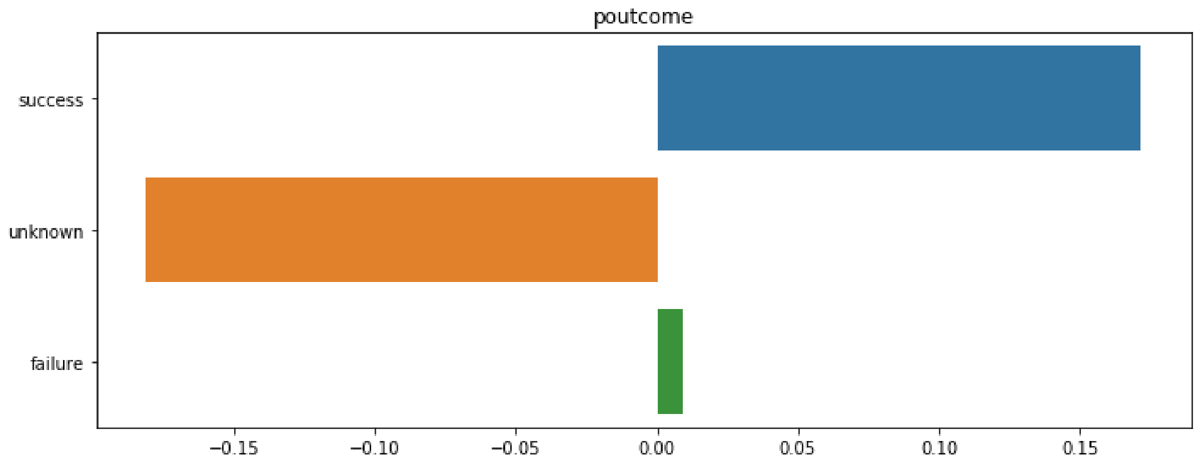

default
no
yes
−0.0100
−0.0075
−0.0050
−0.0025
0.0000
0.0025
0.0050
0.0075
0.0100


housing
no
yes
−0.2
−0.1
0.0
0.1
0.2


loan
no
yes
−0.08
−0.06
−0.04
−0.02
0.00
0.02
0.04
0.06
0.08


contact
cellular
unknown
telephone
−0.2
−0.1
0.0
0.1
0.2





There are many take-away for the marketing team from the above plots:

- Retired clients, students, and management professionals are to be more targeted and blue-collar professionals less.
- Target single clients compared to married clients
- Target clients with higher education
- Target clients who do not have any credits in default, no housing loans, and no personal loan.
- Cellular contact is better.
- Do not contact in May month. It may be because of the summer vacations that families generally take during this month.
- Target the clients for whom the outcome of the previous campaign was successful.

The conclusion from data storytelling

Client information attributes such as age, job, education, marital status gives a general representation of people having a bank account. Balance in the bank account gives an interesting story. There are various clients with balance 0 or less than zero and they still subscribe to a term deposit. The marketing campaign must target retired clients and also clients with higher education first.

Attributes related to the last contact of the current campaign tell us that most clients are contacted in summer and it could be any day of the month with a usual contact method being cellular and the duration mostly under 5 minutes. But the clients contacted in May have the highest no’s in the target variable.

And some attributes talk about the previous campaigns. It is better to target clients who have already been contacted before.

Conclusion :

Client information attributes such as age, job, education, marital status gives a general representation of people having a bank account. Balance in the bank account gives an interesting story. There are various clients with balance 0 or less than zero and they still subscribe to a term deposit.

Attributes related to the last contact of the current campaign tell us that most clients are contacted in summer and it could be any day of the month with a usual contact method being cellular and the duration mostly under 5 minutes.

And some attributes talk about the previous campaigns. It can be understood from that data that most clients were not contacted before. Only a small fraction of them was contacted before.