EM 622 / SYS 622 Spring 2022
Final Project
Bank Marketing Data Visualization

Team Members –

- Yuxin Cao
- Mehul Kumar Jain
- Sonali Pandurang Raut

Introduction –

The bank's revenue has decreased, and they want to know what steps to take next. Following a survey, it was discovered that the fundamental cause is that their consumers are not depositing as regularly as they used to. Term deposits allow banks to keep a deposit for a certain period, allowing them to lend more and profit more. Furthermore, banks have a better chance of persuading term deposit consumers to purchase other products such as funds or insurance in order to enhance their revenue.

The data we have chosen is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed. Our main objective is to find out the categories of the people who subscribed to the term deposit and during which months was the campaign most successful.

There are four datasets:

1. bank-additional-full.csv with all examples (41188) and 20 inputs, ordered by date (from May 2008 to November 2010), very close to the data analyzed in [Moro et al., 2014]
2. bank-additional.csv with 10% of the examples (4119), randomly selected from 1), and 20 inputs.
3. bank-full.csv with all examples and 17 inputs, ordered by date (older version of this dataset with less inputs).
4. bank.csv with 10% of the examples and 17 inputs, randomly selected from 3 (older version of this dataset with less inputs).

We have chosen the “bank-additional-full.csv” file to perform data analysis and visualization.

Dataset link – https://data.world/data-society/bank-marketing-data

Understanding the dataset –

Some of the important variables of the dataset –

1 - age (Numeric)
2 - job: type of job (Categorical: 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')
3 - marital : marital status (Categorical: 'divorced', 'married', 'single', 'unknown'; note: 'divorced' means divorced or widowed)
4 - education (Categorical: 'basic.4y', 'basic.6y', 'basic.9y', 'high.School', 'illiterate', 'professional.course' ,'university.degree', 'unknown')
5 - default: has credit in default? (Categorical: 'no', 'yes', 'unknown')
6 - housing: has housing loan? (Categorical: 'no', 'yes', 'unknown')
7 - loan: has personal loan? (Categorical: 'no', 'yes', 'unknown')
8 - contact: contact communication type (Categorical: 'cellular', 'telephone')
9 - month: last contact month of year (Categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
10 - day_of_week: last contact day of the week (Categorical: 'mon', 'tue', 'wed', 'thu', 'fri')
11 - duration: last contact duration, in seconds (Numeric).
12 - campaign: number of contacts performed during this campaign and for this client (Numeric, includes last contact)
13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (Numeric; 999 means client was not previously contacted)
14 - previous: number of contacts performed before this campaign and for this client (Numeric)
15 - poutcome: outcome of the previous marketing campaign (Categorical: 'failure', 'nonexistent', 'success')

The main goal is to find out if the client will subscribe (yes/no) a term deposit (variable y).

	age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration
1	56	housemaid	married	basic.4y	no	no	no	telephone	may	mon	261
2	57	services	married	high.school	unknown	no	no	telephone	may	mon	149
3	37	services	married	high.school	no	yes	no	telephone	may	mon	226
4	40	admin.	married	basic.6y	no	no	no	telephone	may	mon	151
5	56	services	married	high.school	no	no	yes	telephone	may	mon	307

Fig 1 – First 5 rows of our dataset

Data Preparation –

One of the important steps in data analysis is data preparation. The steps which we have taken for data preparation are –

1) Renaming the column names for few of our columns – col “y” became “Desired_Target”, “poutcome” became “Previous_Outcome”
2) Changing the target column’s values – “yes” became “Subscribed”, “no” became “Not Subscribed”
3) Also converting the “month” and “day_of_week” column values into a proper form –
    a. In month column – “aug” became “August”, “dec” became “December”. etc.
    b. In “Day_of_week” – “mon” became “Monday” and “tue” became “Tuesday”. etc.
4) Lastly changed data type of a couple of categorical variables from "chr" to "factor"

Handling Missing values –

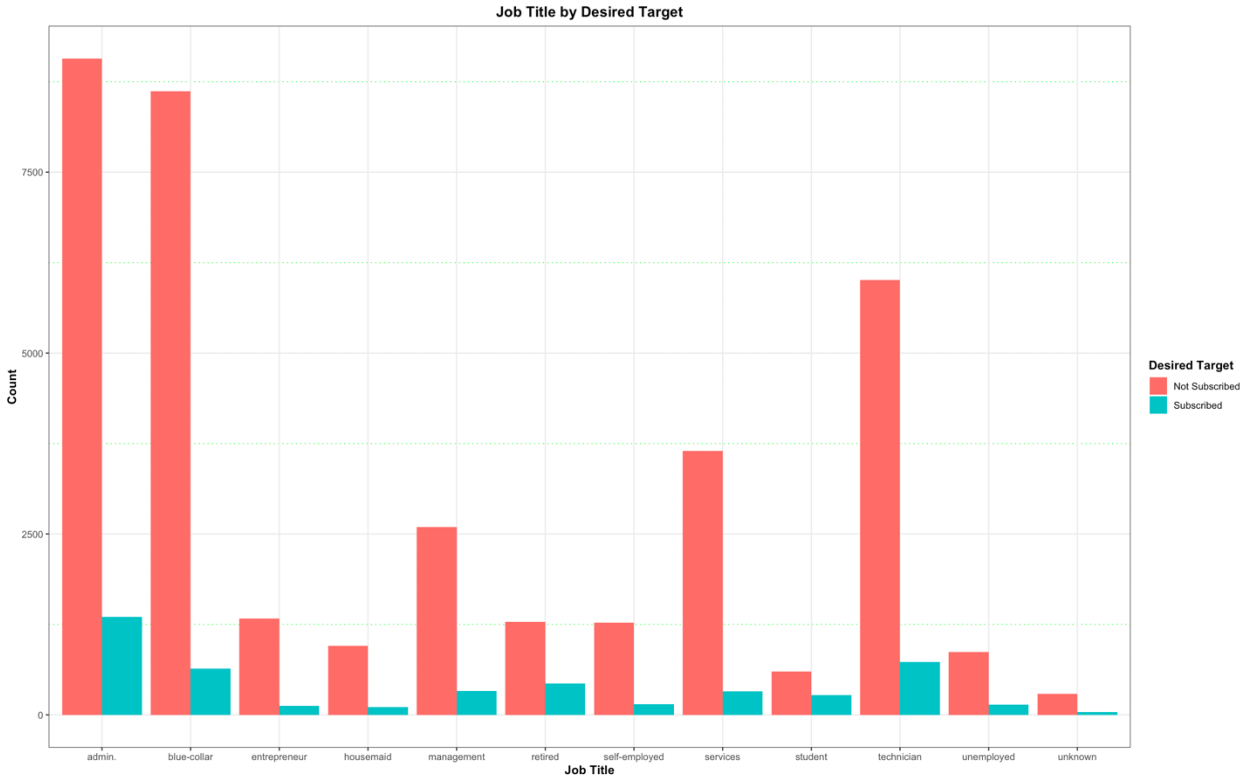
There are several missing values in some categorical attributes, all are coded with the "unknown" label. These missing values are treated as a possible class label.

Proposed Questions –

- Who were the targeted audience of the bank marketing campaign?
- In which categories most of the subscribed audience fall into?
- Is there any relationship between the previous and current campaign outcomes?
- In which months and days of the week, was the campaign more successful?
- How many people in the subscribed audience had loan?
- How successful was the campaign?

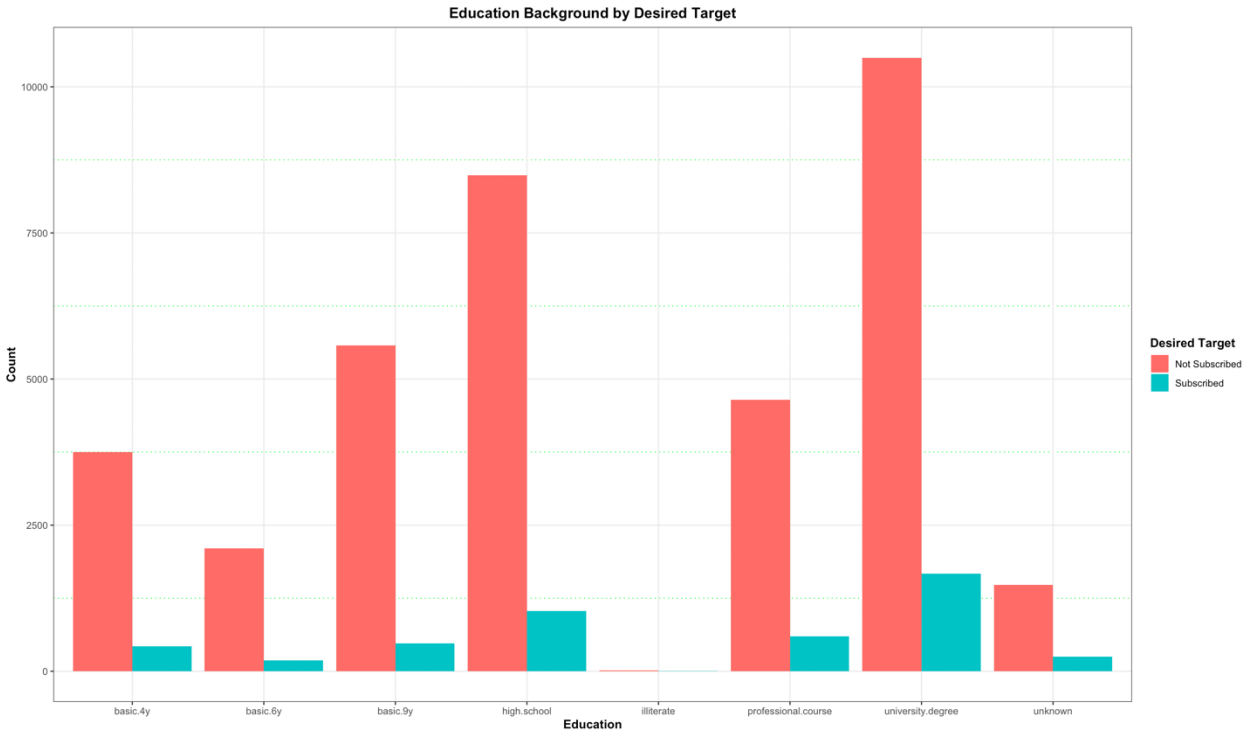
Visualizations –

Plot 1 –



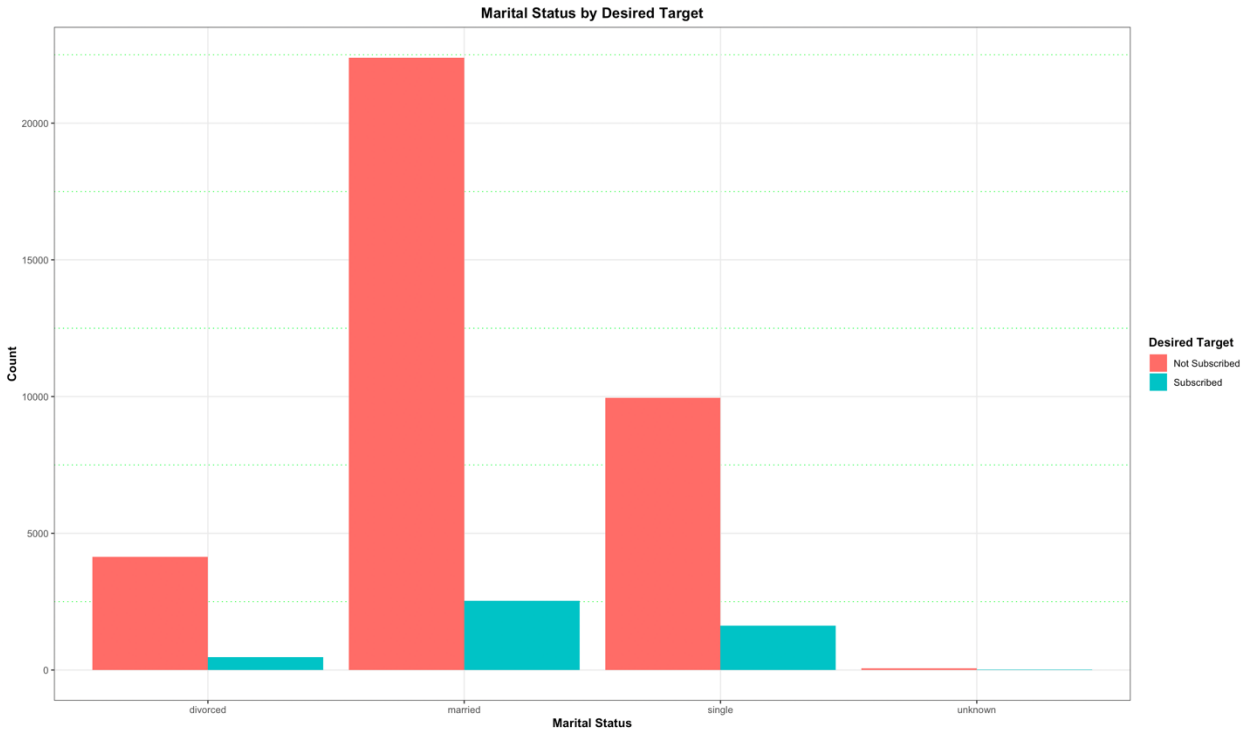
From the above graph, we can see that most of the targeted audience and the people who actually subscribed to the service (term deposit) were **admins by profession**. Around **1250** people who subscribed to service were admins. The campaign also targeted **people with blue-collar jobs** and out of the people with blue-collar jobs, around **800** of them subscribed to the service (term deposit).

Plot 2 –



The above visualization gives us insights about the Educational background of the targeted audience and the subscribers. We can see that, most of the targeted audience and the subscribers had a **university degree**. Around **1500** people with a university degree subscribed to the service. Many people with just **high school education** also subscribed to the service.

Plot 3 –



The above visualization gives us information about the Marital status of the targeted and the subscribed audience. We can see that most of the targeted audience and the subscribers were **married.**

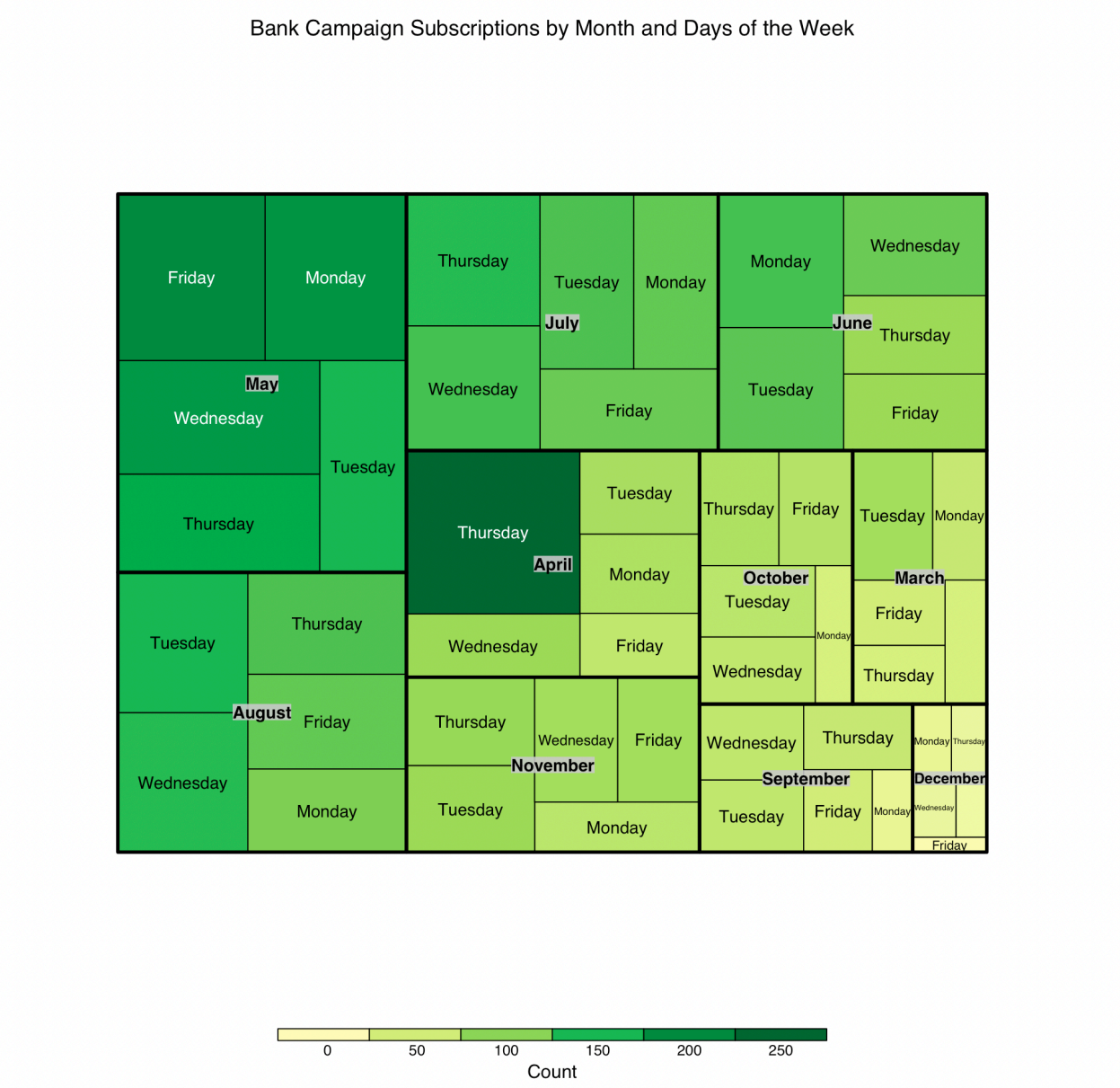
We chose histograms for the above visualizations because, we wanted to analyze the counts of the different categorical variables and histograms are the best type of visualizations to provide us with the required information.

Plot 4 –

The above visualization is scatter plot between the duration of the call and the age of the subscribed audience, split by the outcomes of the previous campaign. We can see that the outcome of most of the subscribed audience was either **successful or nonexistent** and the highest contact type was **cellular**. Also, there are very less subscribers whose previous campaign was failure, which totally makes sense and the most the subscribers belonged to the ages from **20 to 50 years**.

Since there was a mixture of categorical and numerical variables, scatter plot is the best type of visualization to get insights about the required information.

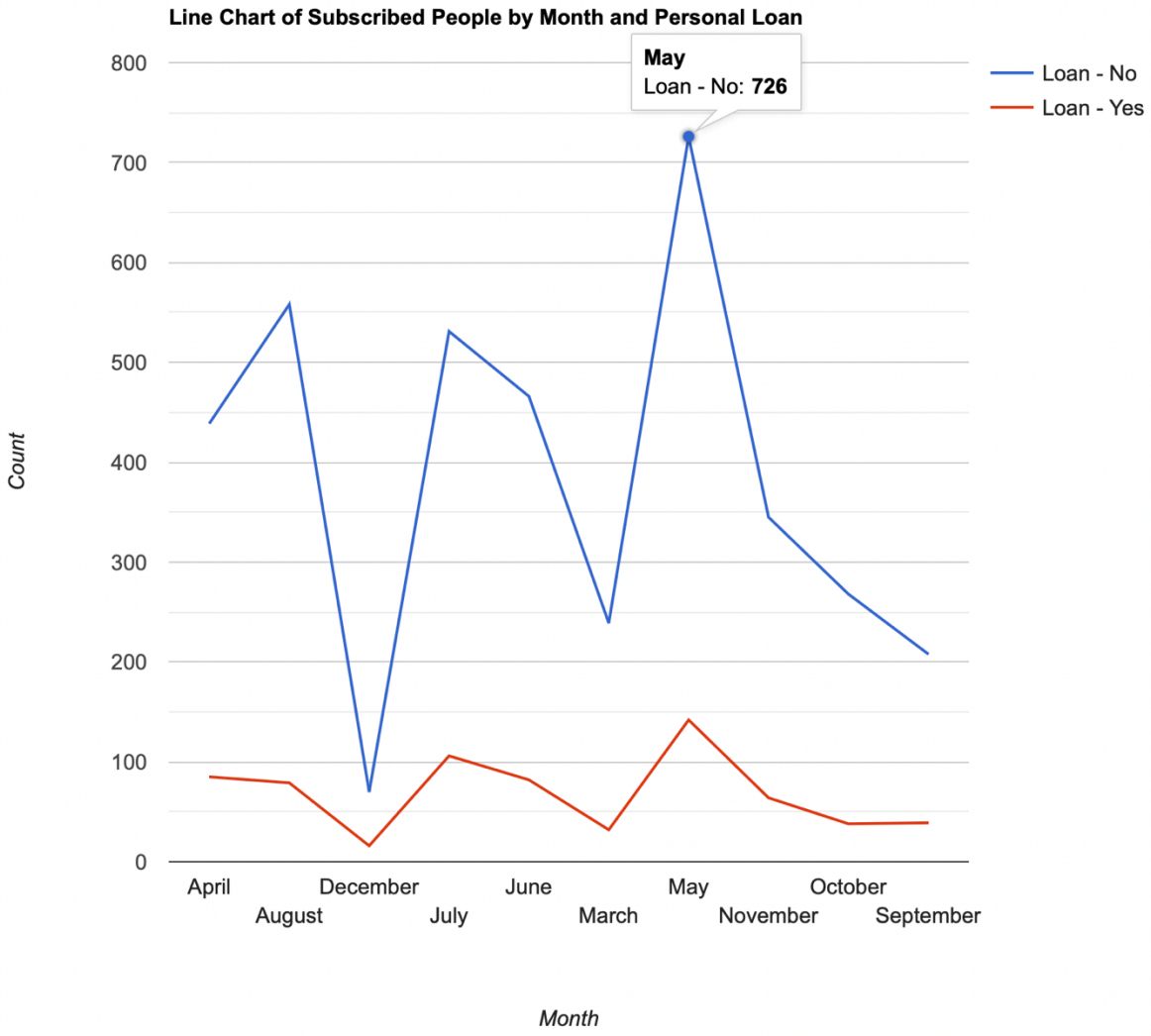
Plot 5 –



The above visualization is a tree map used to get insights about the months and the days of the week in which the campaign was more successful. We can observe that, the campaign was relatively more successful in the months of **May, June, July and August**. If we look carefully, we can notice that the campaign observed highest number of subscriptions on **Thursdays in the month of April**.

Since we wanted to analyze the contributions of the months and days of the week, tree maps are easier to understand the contributions by different categories.

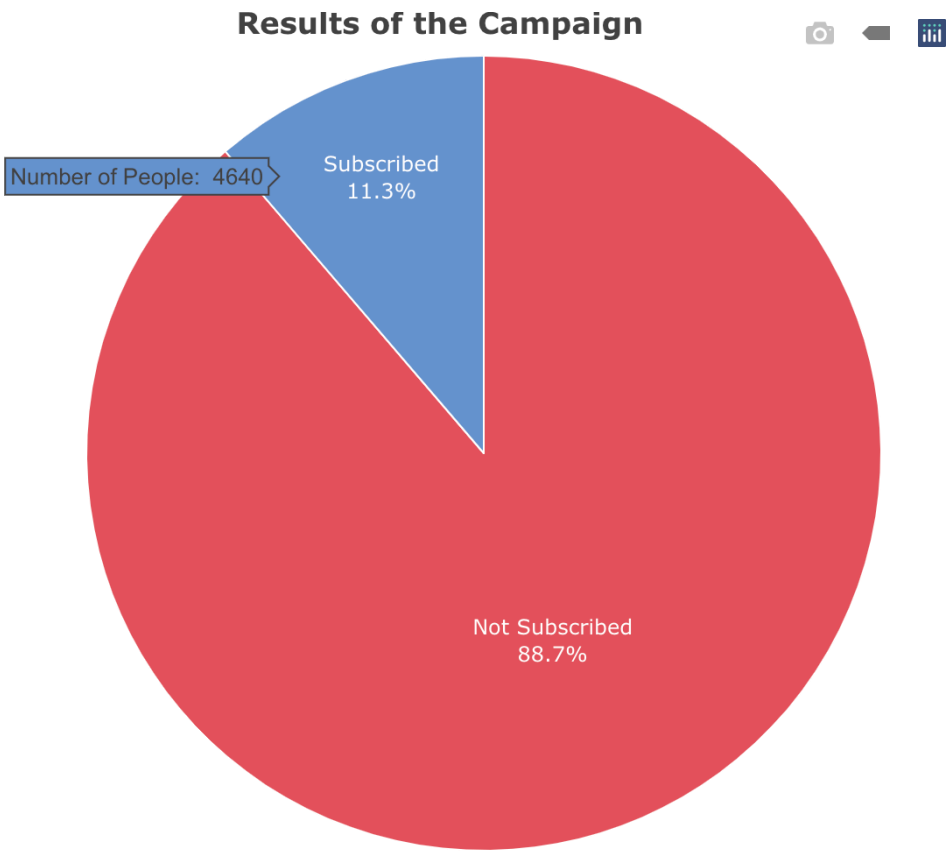
Plot 6 –



The above visualization is an interactive line chart between the month and the count of the subscribed people who have personal loan or not. We can see that most of the people who subscribed to the service (term deposit) **did not have personal loan**. Also, in the month of May during which the campaign was relatively more successful many people did not have loan.

The interactive visualization gives us the exact count of the subscribers who had personal loan or not in every month.

Summary –



To understand how successful the campaign was, the above visualization is used. It gives us details about the percentage and the total number of people who subscribed and did not subscribe to the service (term deposit). We can see that around **10%** of the targeted audience opted for term deposit which seems to be relatively low. But we guess that's how campaigns work, we need to make thousands of phone calls to achieve so little.

Since we wanted to analyze the share of the subscribed and not subscribed people, pie chart is the best type of visualization.

To conclude, we can say that most of the subscribed audience were **admins**, **married** and had a **university degree**. The outcome of the previous campaign had an impact on the current campaign, but it was **not very significant**. Also, most of the people who subscribed to term deposit **did not have loan** and the campaign was most successful in the months of **May**, **June**, **July and August**.