**EXECUTIVE SUMMARY**

Finance and banking are one of the most extensive and extremely competitive markets. Any organisation going face-to-face with the big players need to fortify that they understand precisely how buyers like to interact with their sales and marketing processes. Customers today lean on both online as well as offline means to shop before making a decision. Almost on every occasion, an offline phenomenon such as making a phone call or visiting a branch is a positive indicator of a possible conversion. In this report, a bank marketing dataset of a Portuguese bank is selected where the marketing campaigns were based on phone calls. The report is based on the sample of the entire dataset which is about 50% of the original dataset. Initially, the dataset is split into training data, testing data and evaluation data. The outcome of the campaign is predicted using various classification and prediction models. The models used for predicting the result are Decision trees (C5.0, CART, QUEST, CHAID, RANDOM TREES), Neural Networks, Bayesian network and Support Vector Machines (SVM). Each model is presented briefly in this report and the best model is selected for analysis. CHAID produced the best results out

**INTRODUCTION**

Banks and financial institutions exist to offer financial services to people and to make huge profits. Having said that, banks also devote remarkable resources and business intellect to gain capital. One of the most common ways for banks to do this is to engage in direct marketing campaigns like phone calls and face-to-face meetings to promote and provide services. Phone calls, i.e. Telemarketing is a conventional marketing technique that helps to soar profits for any given business. Moreover, it also offers a more interactive and personal medium of sale service which can initiate an instant rapport with the prospective customers. Furthermore, telemarketing can help an organization to reach out more customers than with in-person or by going door-to-door and it can benefit a company to sell a product to both existing and new customers. For banking industry, telemarketing can be useful to communicate with large number of customers and offer them with all the services that they have for them. This may include, information about loans, term deposits, mortgages, Overdraft facility, Credit cards etc.

For this project, a data set of a Portuguese Bank direct marketing campaign is used. This dataset is obtained from the UCI machine learning repository (<https://archive.ics.uci.edu/ml/datasets/Bank+Marketing>). The primary objective of this project is to find the best model to predict whether a customer will subscribe for a term deposit or not using various classification techniques. Our secondary objective is to determine what factors in this data set would contribute the most for the sale of term deposits to the potential customers. The target users for this project are the marketing team of a banking institution who are looking to increase their inflow of cash deposits. The following sections of this report includes the description of the dataset in detail, all the methods (classification techniques) that has been applied on the dataset to get the results and eventually the best one is described thoroughly. Moreover, the results for the best model is presented followed by conclusion which summarises the most important findings and the scope of future research is suggested.

**DESCRIPTION OF THE DATASET**

**Original Data:**

The original data has been extracted from the UCI Machine Learning Repository. The data is a result of a direct marketing campaign executed by a Portuguese banking institution to promote term deposits. The campaign was based on phone calls. It contains a total of 45211 instances. There are 17 attributes in total, out of which 16 are independent variables and 1 is dependent variable (outcome variable). The outcome variable is binary (yes/no), where yes means a customer will subscribe for a term deposit or no otherwise. The description of all the attributes is given in the table below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sr.#** | **ATTRIBUTES** | **TYPE** | **DESCRIPTION** | **VALUES** |
| ***CLIENT DATA*** | | | | |
| 1 | age | Numeric | Age of the Client | Positive Integer |
| 2 | job | Categorical | Type of Job | admin., unknown, unemployed, management, housemaid, entrepreneur, student, blue-collar, self-employed, retired, technician, services |
| 3 | marital | Categorical | Marital status of the client | married, divorced, single |
| 4 | education | Categorical | Education level of the client | unknown, secondary, primary, tertiary |
| 5 | default | Binary | Whether a client has credit in default or not | yes, no |
| 6 | balance | Numeric | Average yearly balance in Euros | Integer |
| 7 | housing | Binary | Whether a client has housing loan or not | yes, no |
| 8 | loan | Binary | Whether a client has a personal loan or not | yes, no |
| ***CONTACT RELATED DATA*** | | | | |
| 9 | contact | Categorical | Contact communication type | unknown, telephone, cellular |
| 10 | day | Numeric | Last contact day of the month | Positive integer (1 – 31) |
| 11 | month | Categorical | Last contact month of the year | jan,feb, mar…,dec |
| 12 | duration | Numeric | Last contact duration in seconds | Positive integer |
| ***CAMPAIGN RELATED DATA*** | | | | |
| 13 | campaign | Numeric | number of contacts performed during this campaign | Positive Integer |
| 14 | pdays | Numeric | number of days that passed by after the client was last contacted from a previous campaign | Integer (-1 means client was not contacted previously) |
| 15 | previous | Numeric | number of contacts performed before this campaign and for this client | Integer |
| 16 | poutcome | Categorical | outcome of the previous marketing campaign | unknown, other, failure, success |
| 17 | y | Binary | Whether a client has subscribed a term deposit or not | yes, no |

Table 1: Description of the original data

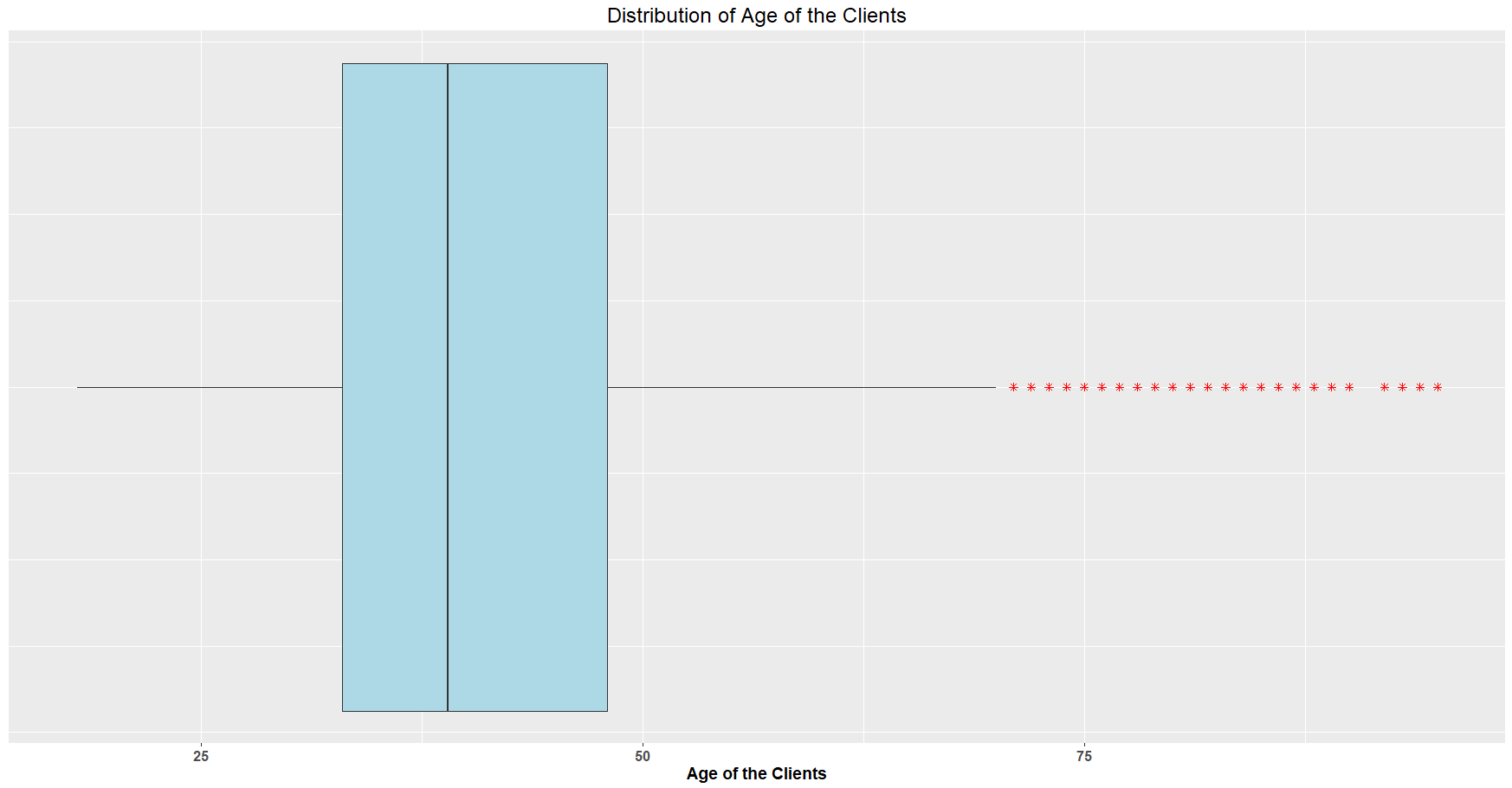
**Data Pre-processing:**

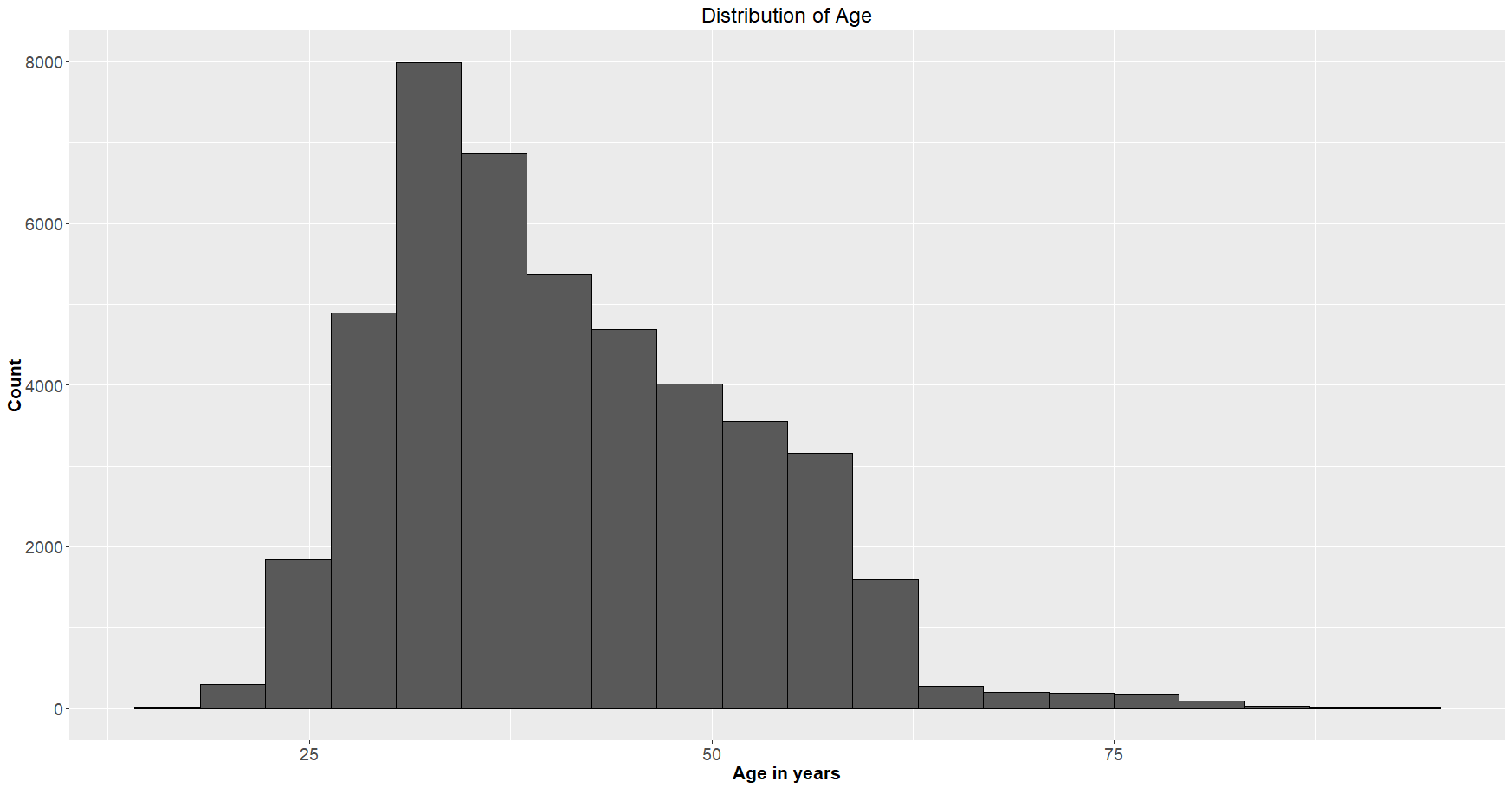
The bank marketing data set that we have used is almost a clean dataset and does not contain any missing values. However, there are some changes such as renaming of variables, identifying the outliers etc that have been made in few attributes as mentioned below:

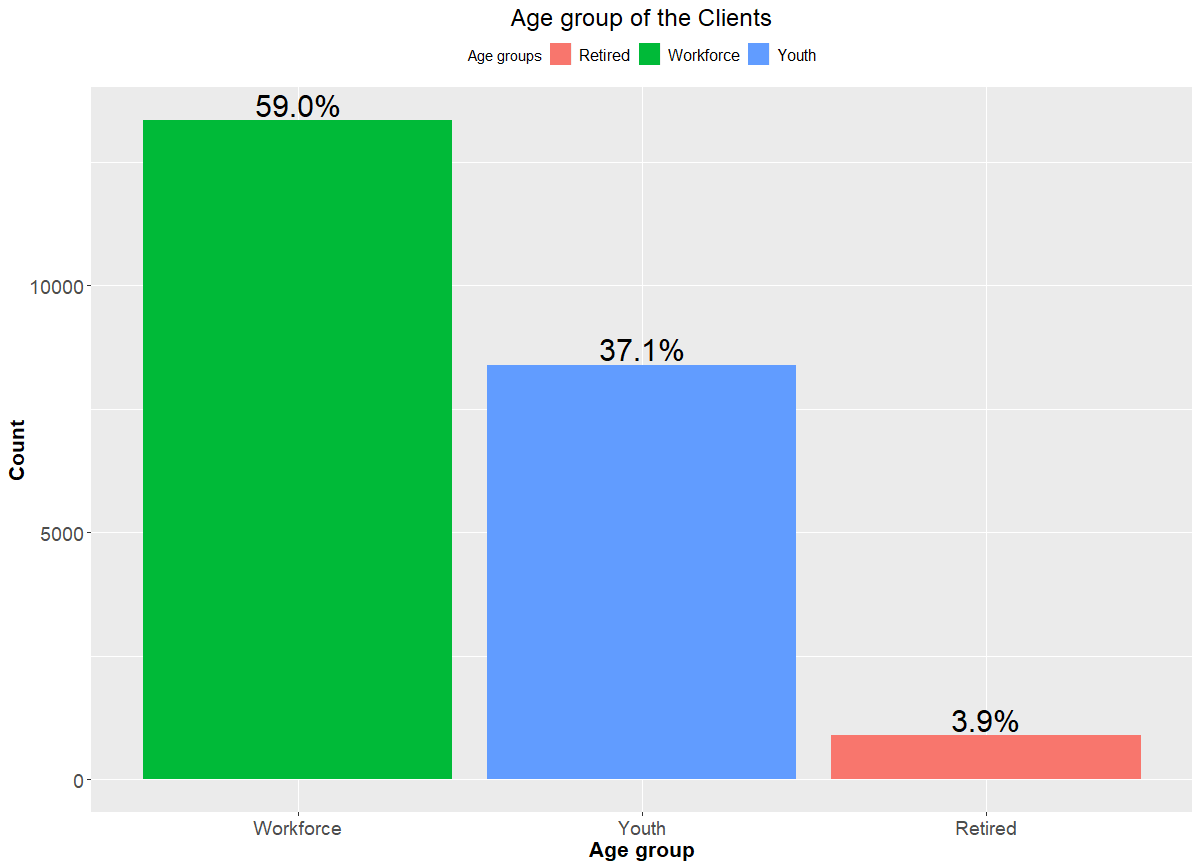
* Firstly, we have selected a sample data from the original data which contains 22606 instances (50% of the original data). This has been done to get accurate models with minimum misclassification.
* Dataset observations – 22606, predictors – 16, outcome variable - 1
* There are no missing values in the dataset.
* The variable "default" has been renamed as "default\_credit" which means whether the client has any credit in default or not. The renaming has been done to make the variable more specific.
* The variable "y" (outcome variable) which indicates whether a client has subscribed a term deposit or not has been renamed as "term\_deposit\_susbscribed". The renaming variable gives a meaning to this variable which makes the outcome easy to understand.
* **Age:** The variable "age" had 180 outliers(Age >= 74) [min = 18, max = 95 ] , so the data was right-skewed. We couldn’t remove the outliers as they add meaning to the data, and they showed nature of the age of people(right-skewed). Just to have better understanding of the distribution of the age, it has been converted to categorical type. Three distinct age groups are created which are: Youth (age between 18 and 35), Workforce (between 36 and 59) and Retired (above 60).
* **Balance:** This variable had 2400 outliers with some outliers(balance < -1884 & balance > 3412) under debt. The maximum account balance was recorded as 98417 euros and maximum debt was 8019 euros, which are plausible amounts therefore no imputation was carried out.
* **Duration:** This variable had 1600 outliers(duration >= 646) with minimum and maximum call duration of 2 and 3881 seconds. As, this was also reliable measure so here also no imputation was done.
* **Campaign:** This variable is right-skewed and has around 1500 outliers(campaign >= 7)[min = 1, max = 58]. In order to maximize the Clients for subscription of term deposit 58 is quite reliable count so no change was made in this variable as well.
* **Pdays:** This variable was also right skewed and had around 4000 outliers(pdays >= 1)[min = -1, max = 854]. Most values were -1 which means client was not previously contacted. Due to very high frequency of -1 the other values were treated as outliers. So, we didn’t remove them.
* **Previous:** This variable was also right skewed and had around 4000 outliers(previous >= 1)[min = 0, max = 275]. Most values were 0 which means client was not contacted before campaign. Due to very high frequency of 0 the other values were treated as outliers. So, we didn’t remove them.
* **Poutcome:** The observation "other" was merged with "unknown" as they belong to the same hierarchy according to the dataset and also there were total 900 records for ‘others’. And also, they both don’t convey any meaning.

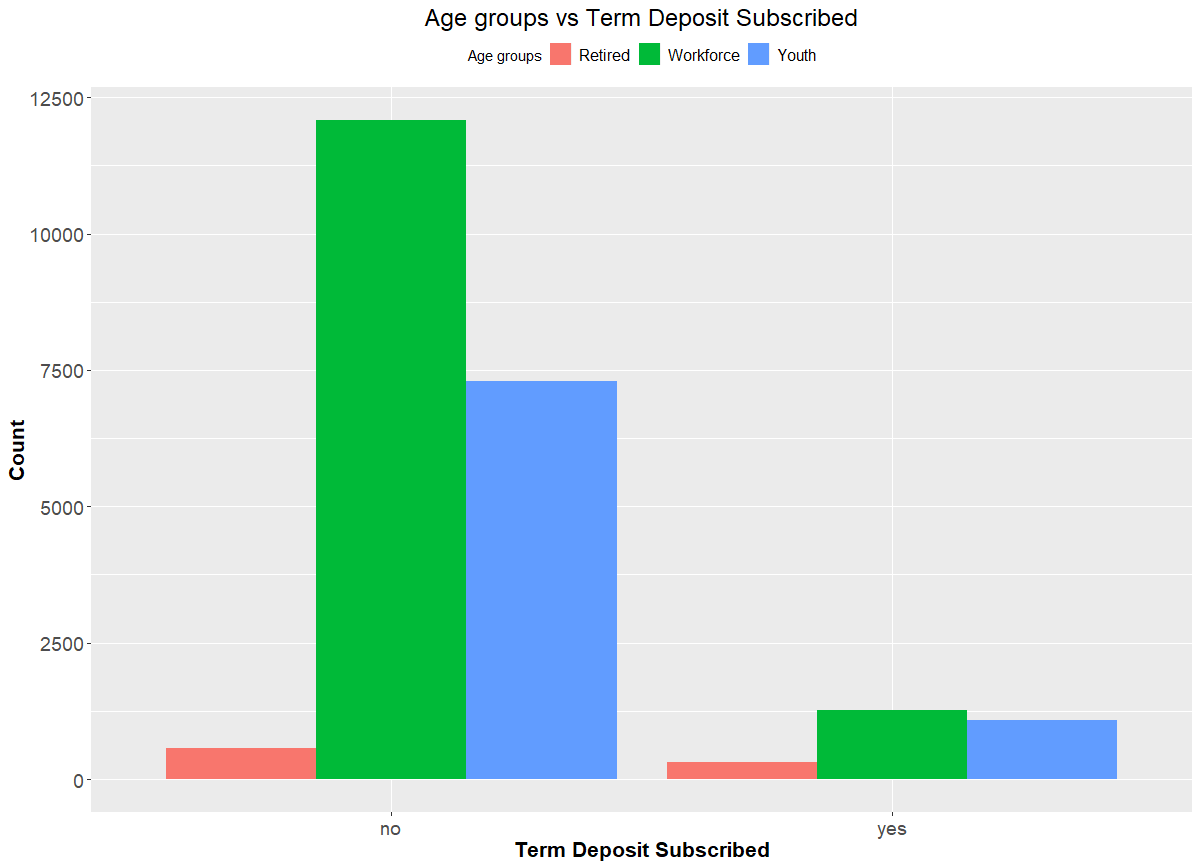
Neither new variables nor any concept hierarchies were created. Neither any observations nor any variables were excluded from the analysis. Dataset after data preprocessing observation – 22606, predictors – 16, outcome variable - 1

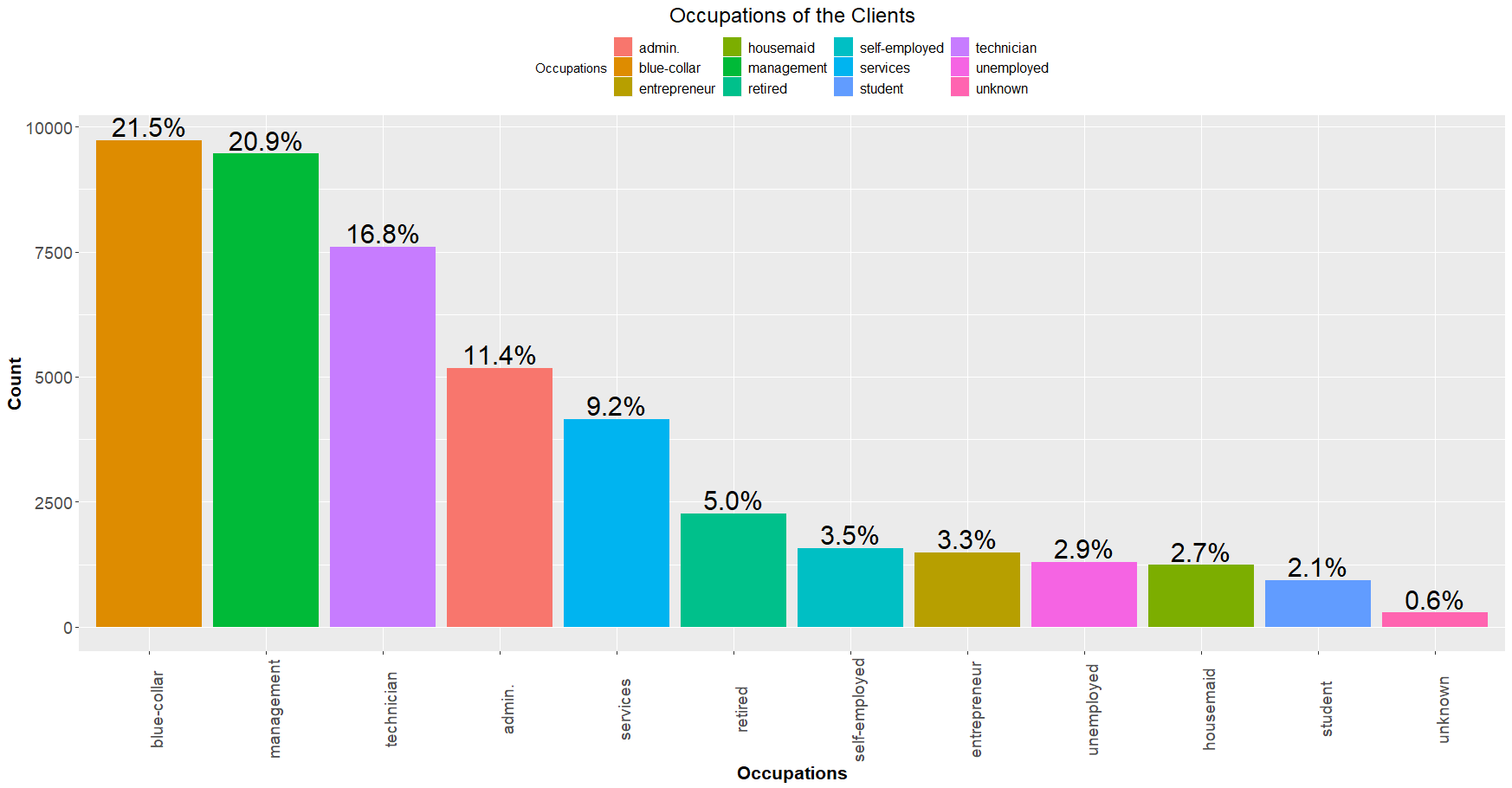
**METHODOLOGY**

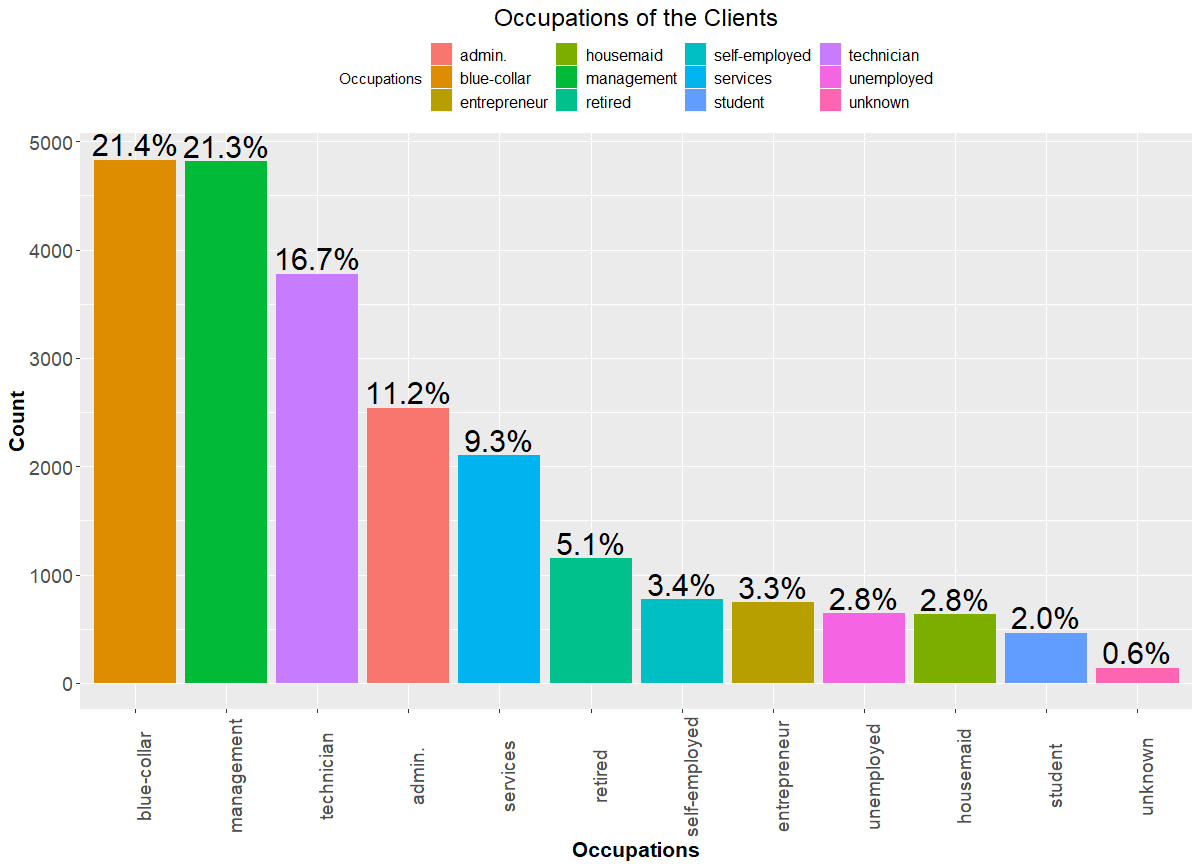


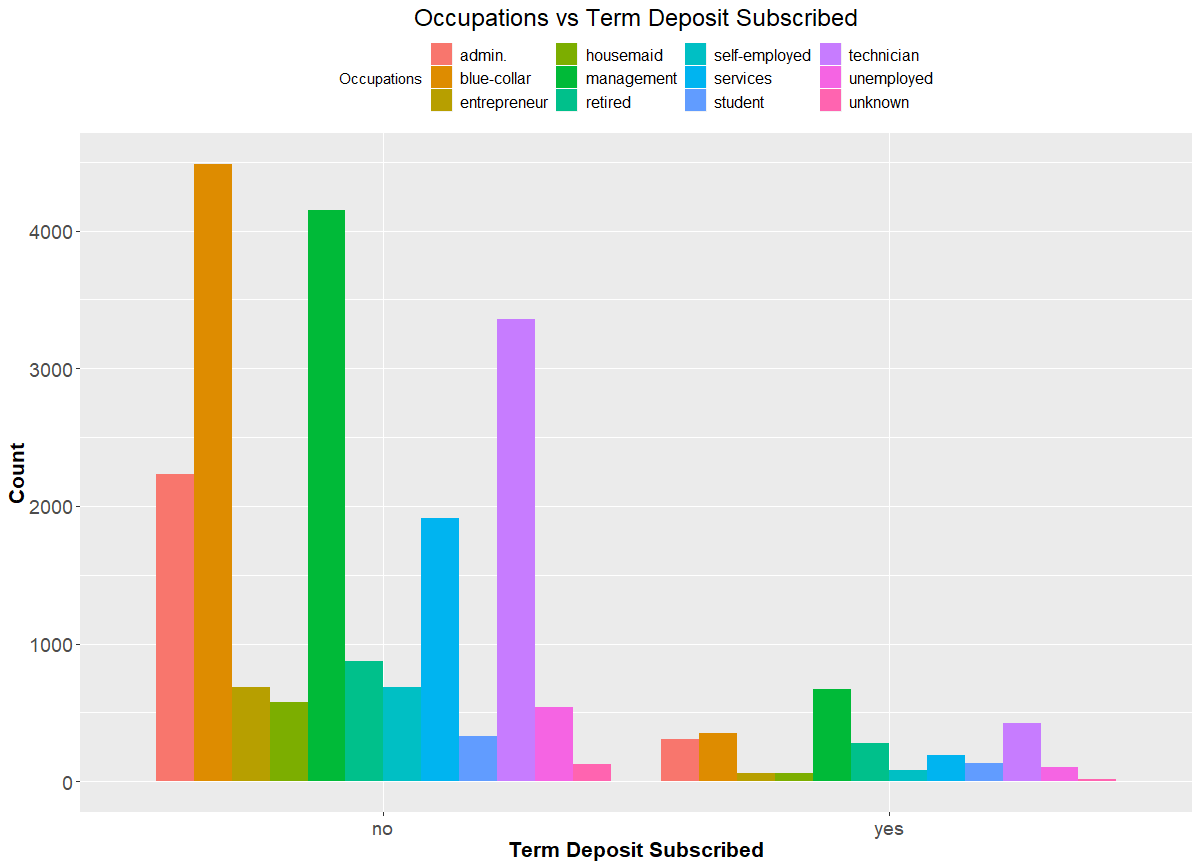


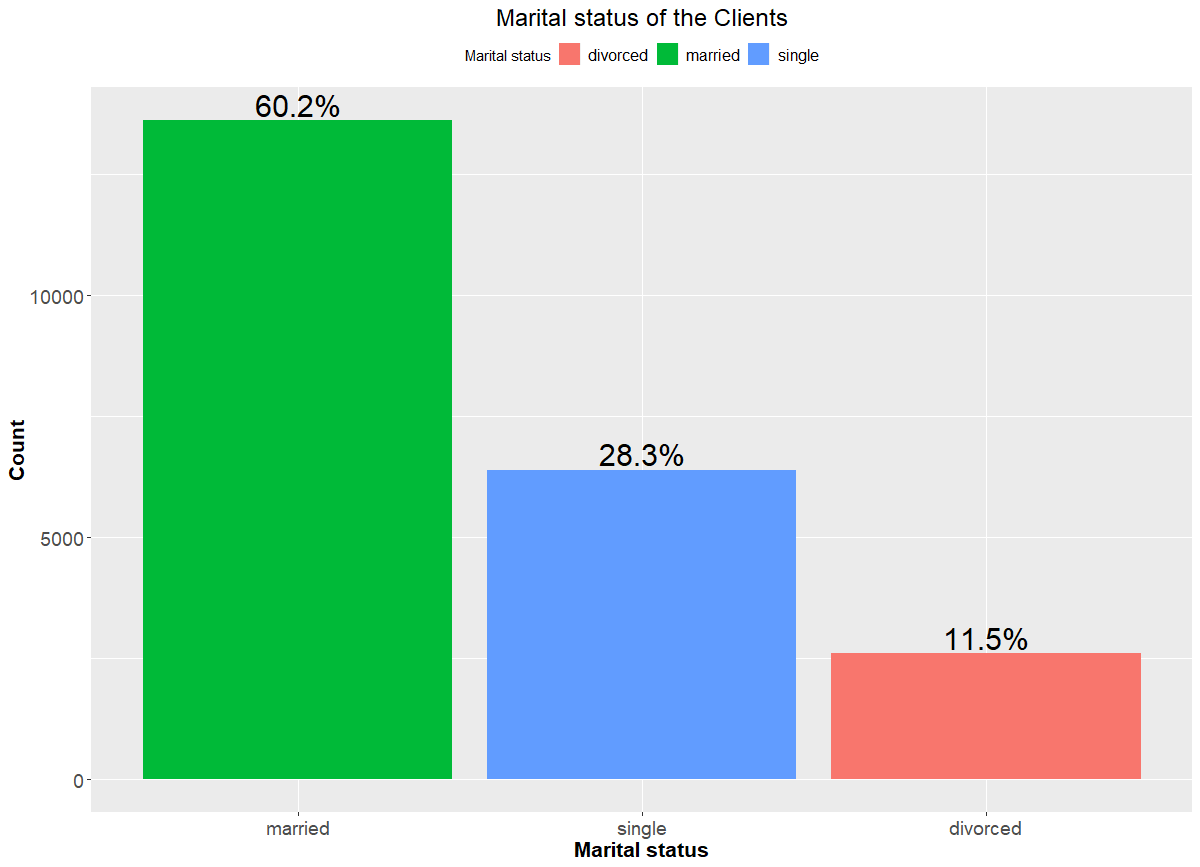


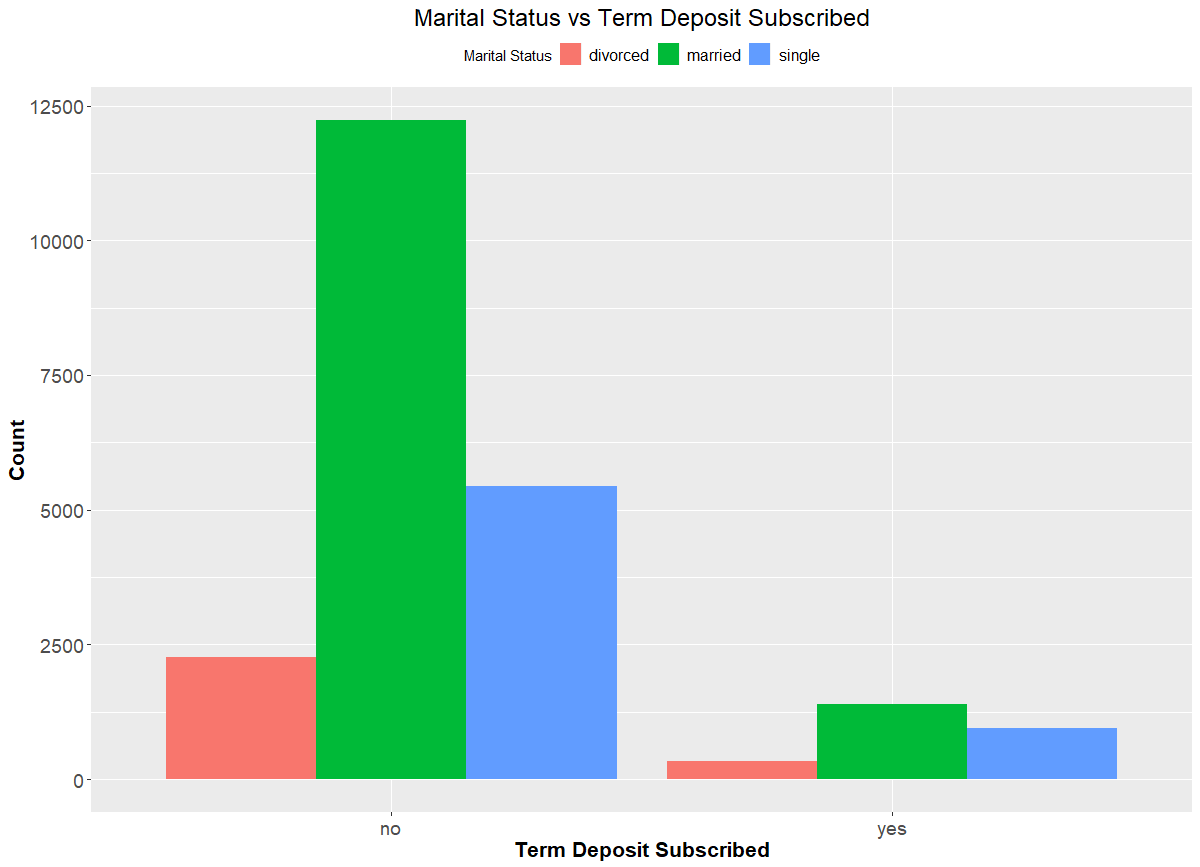


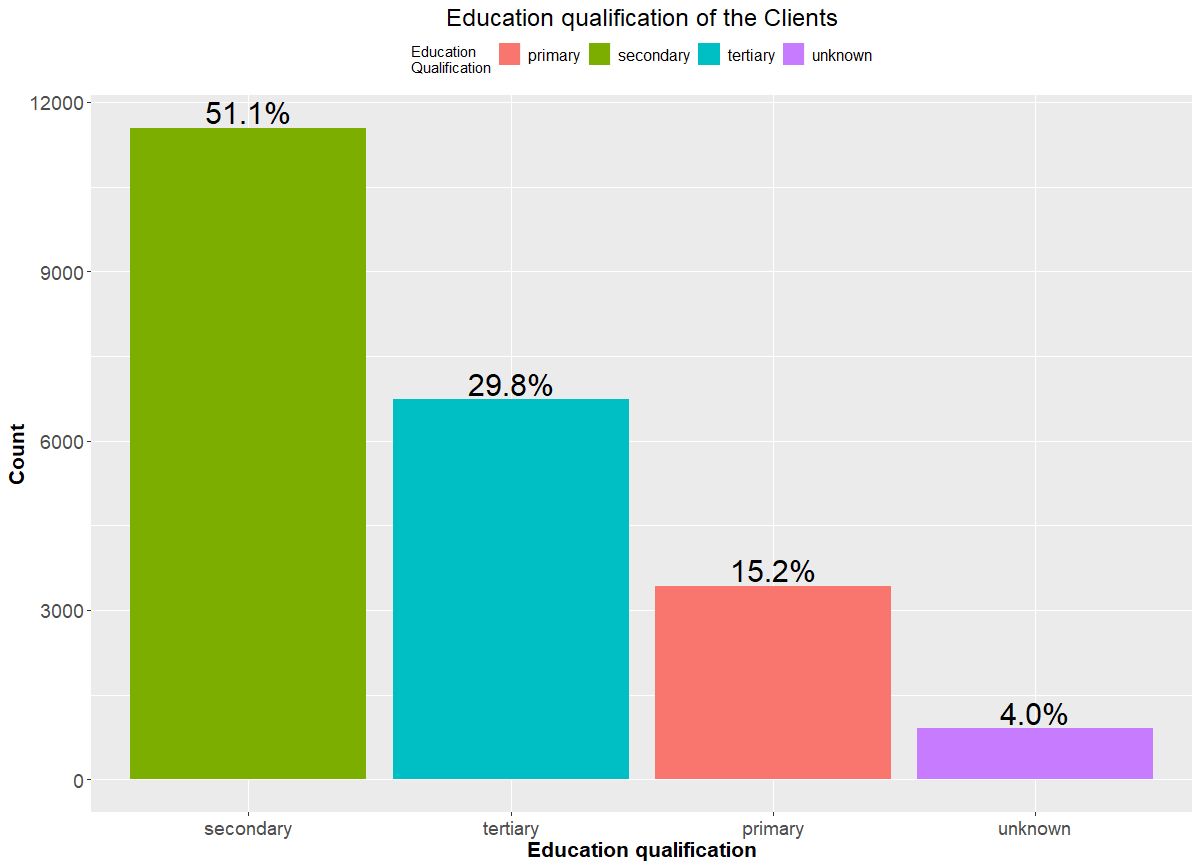


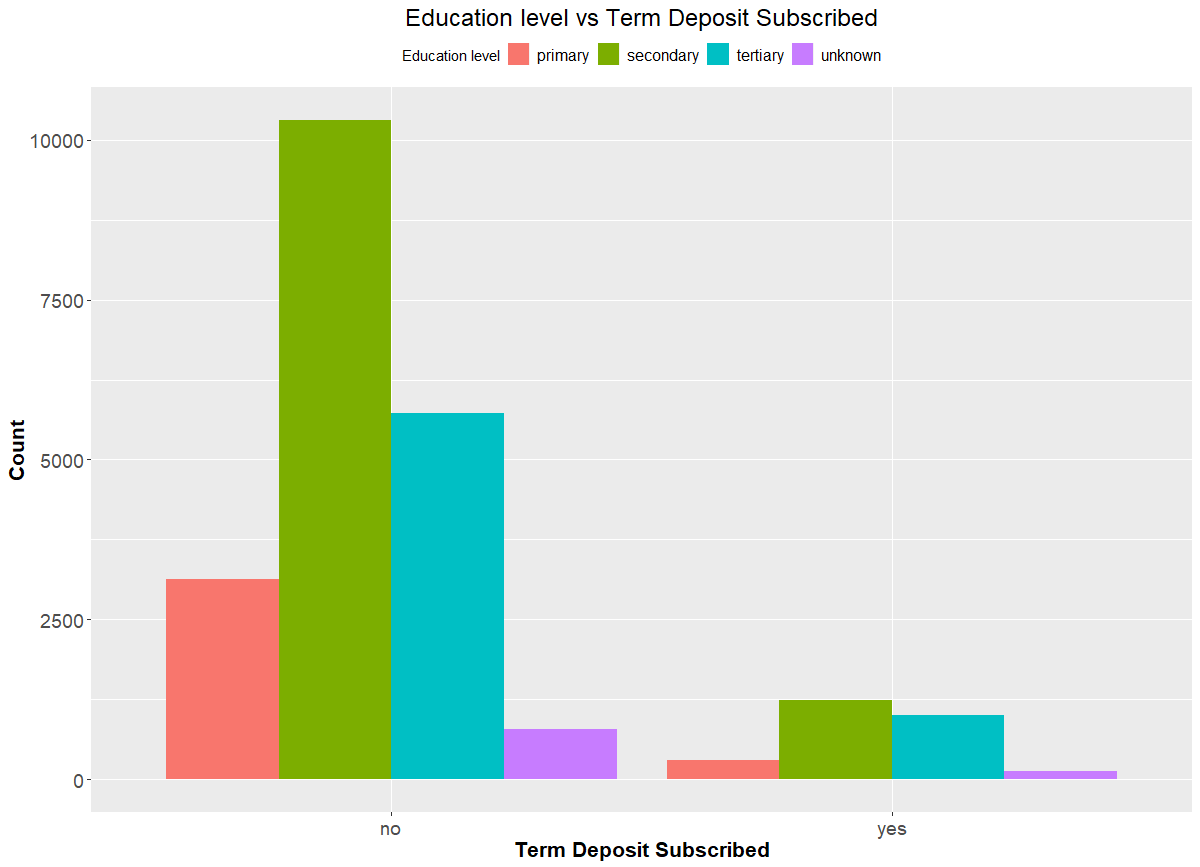


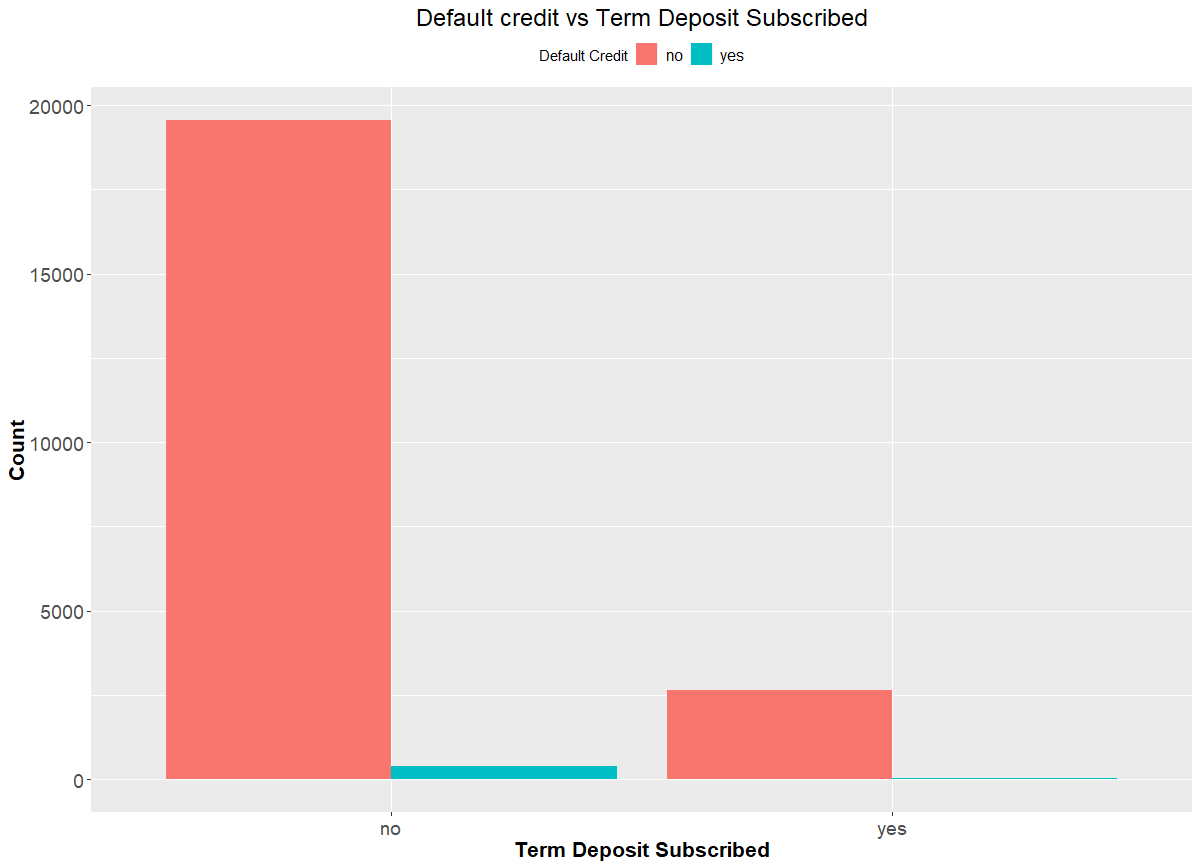
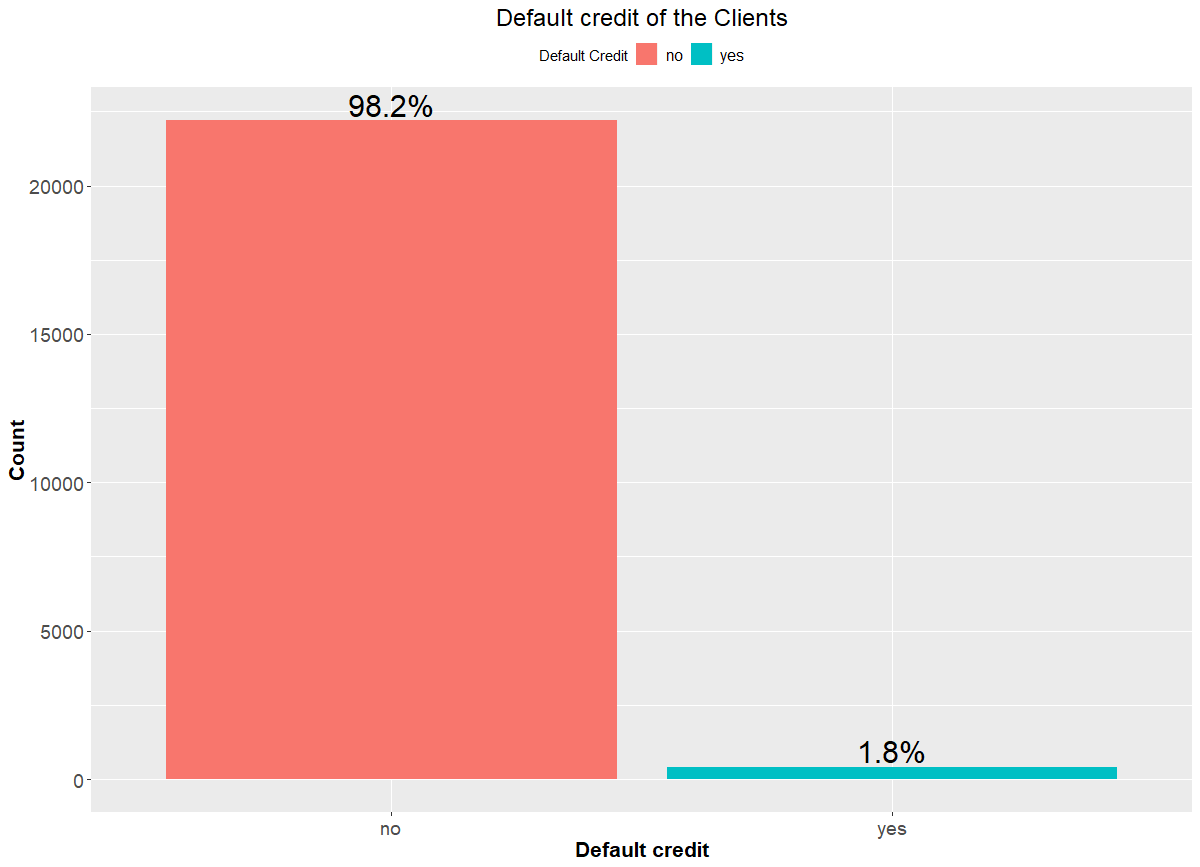


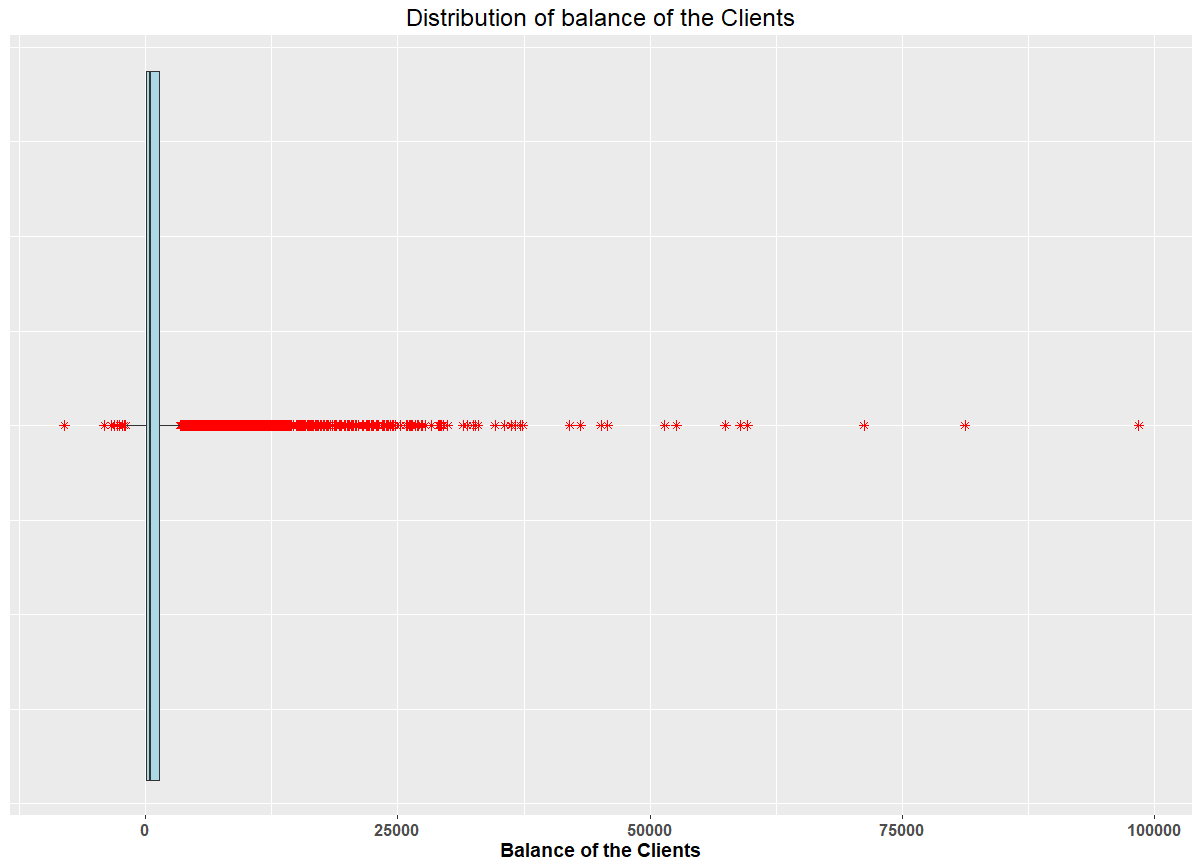


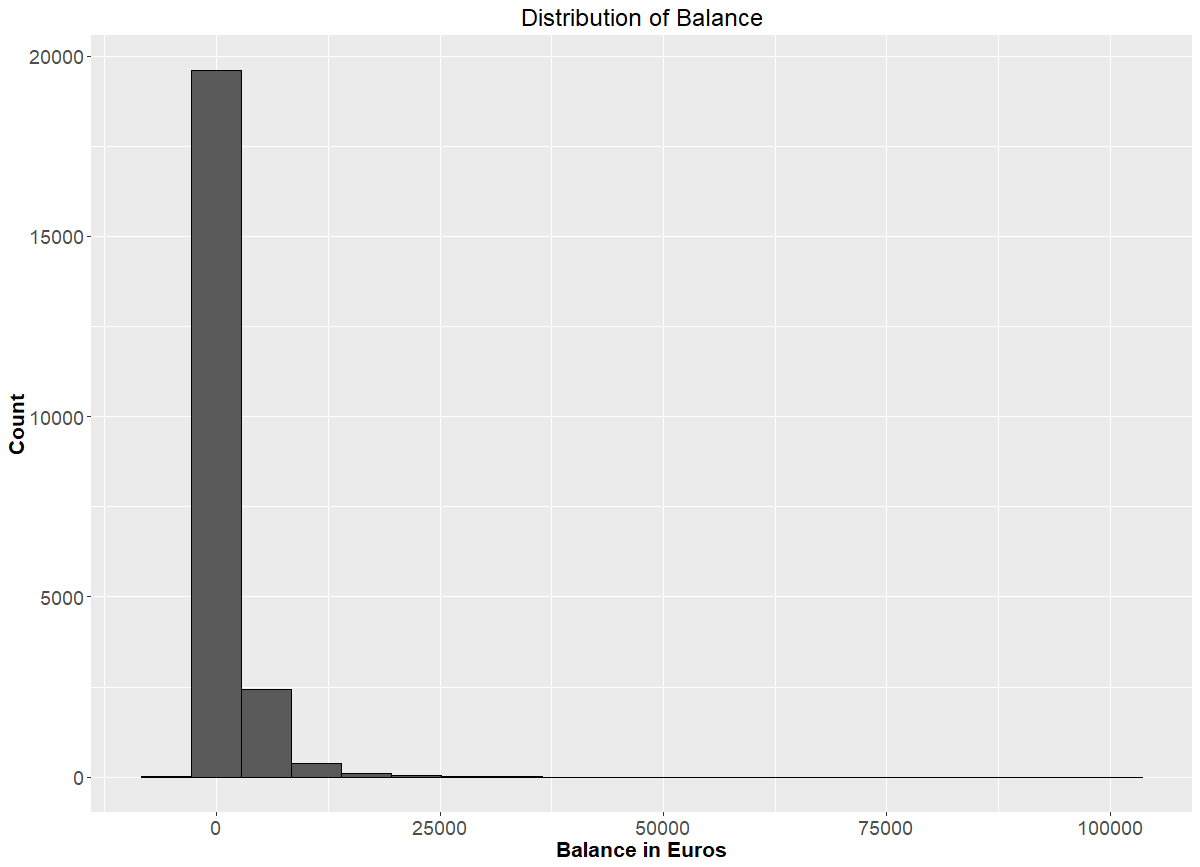


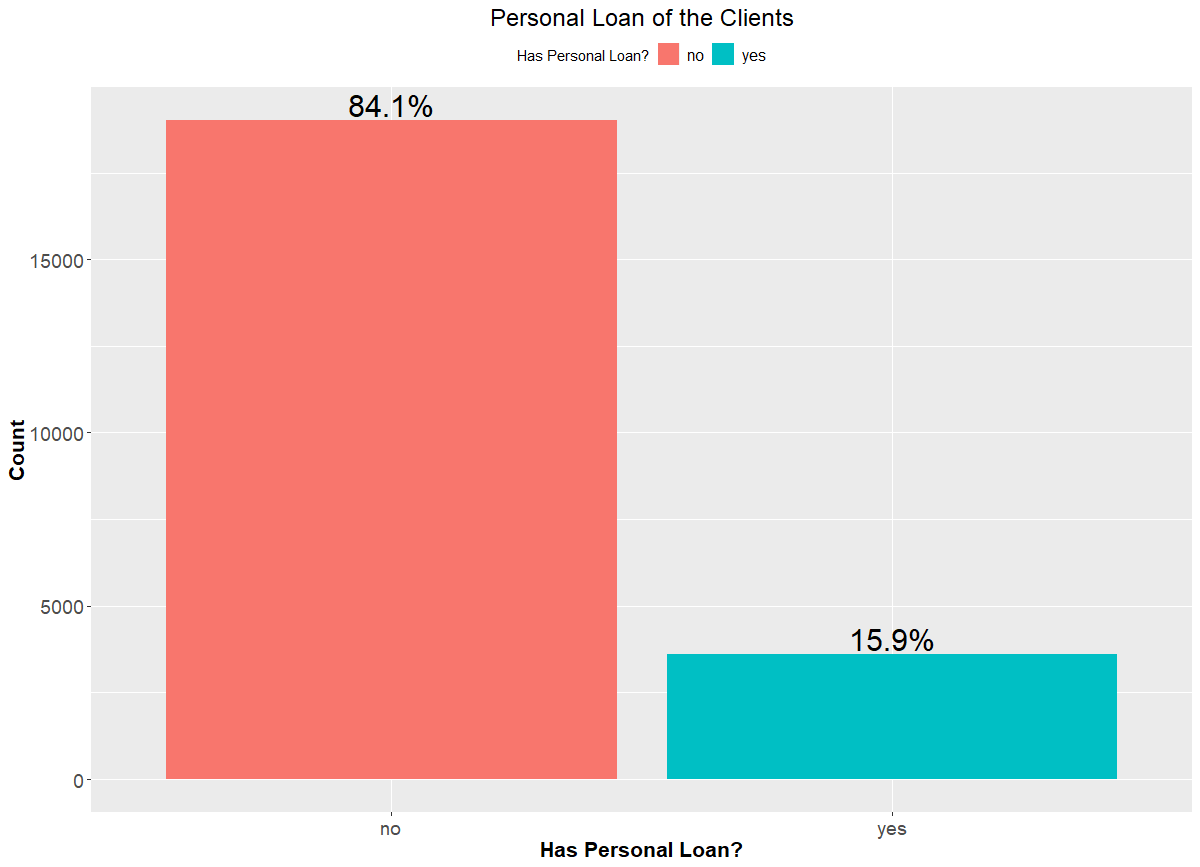


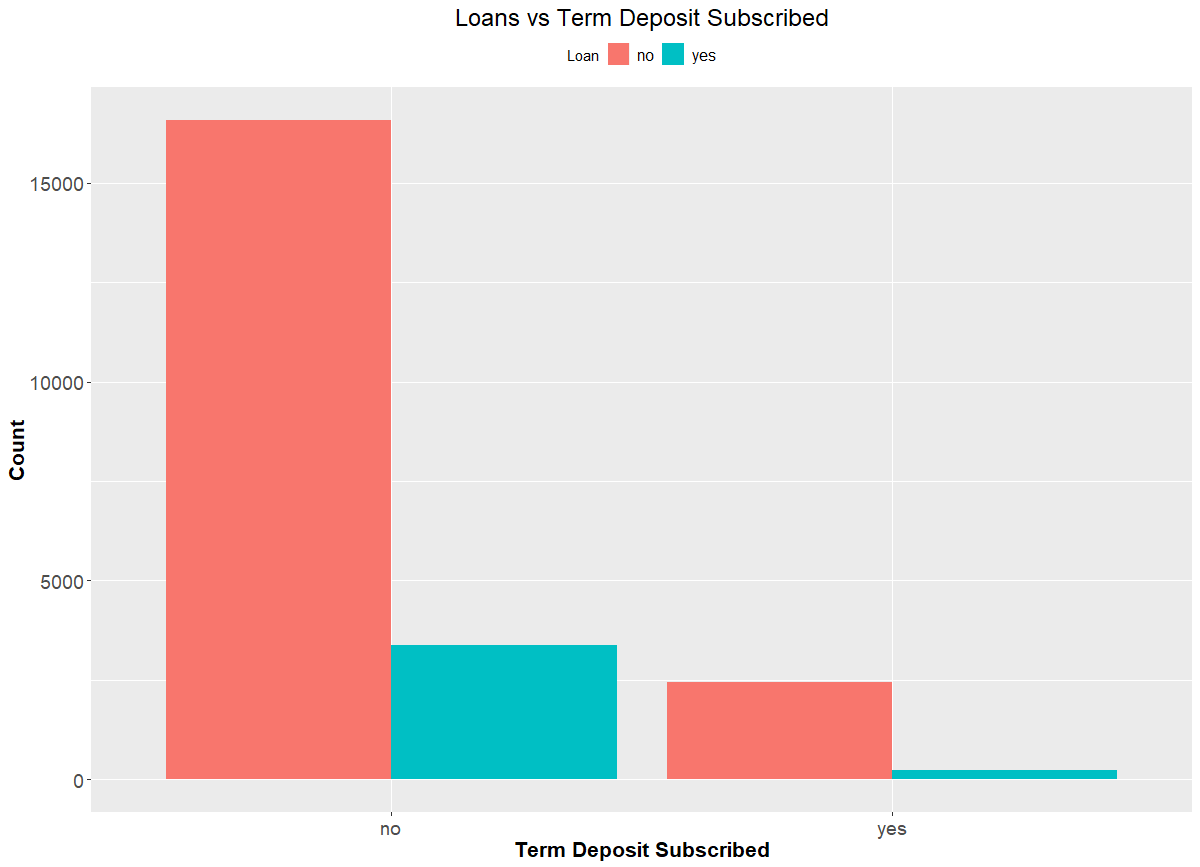


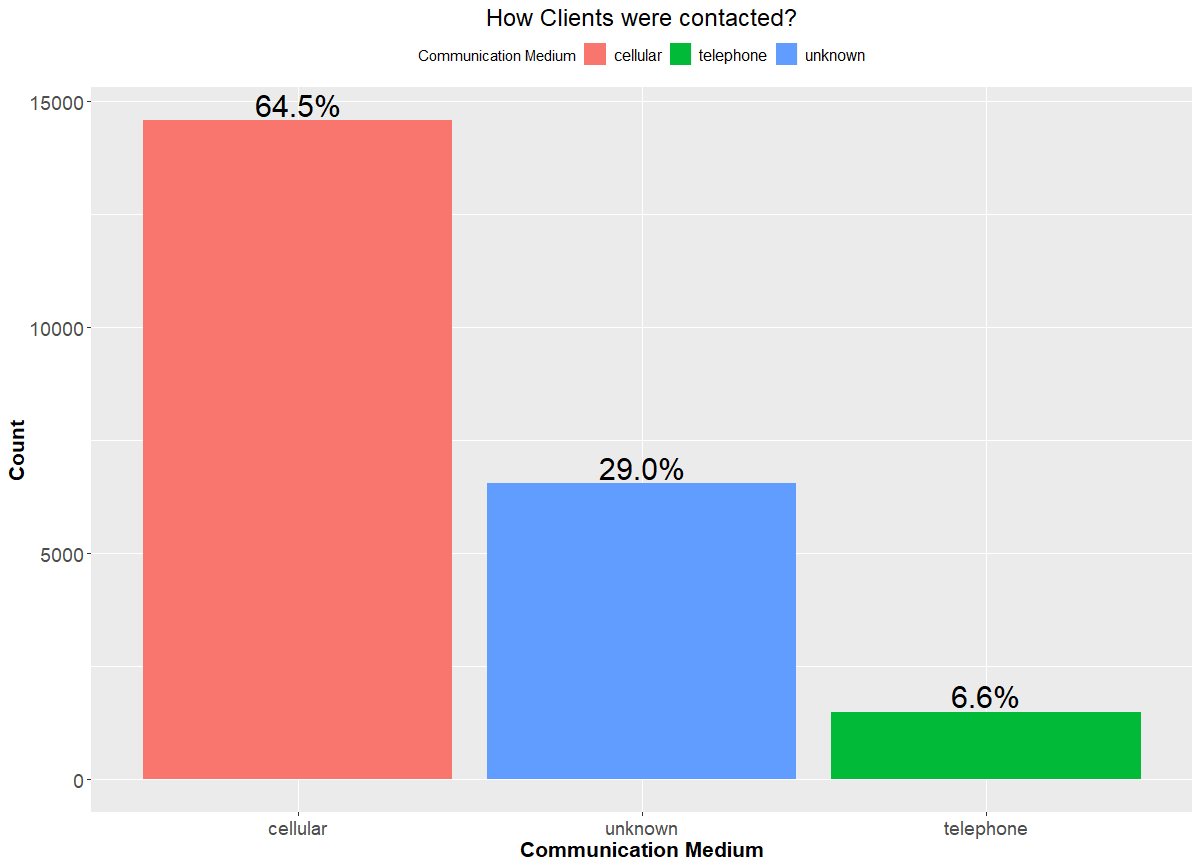


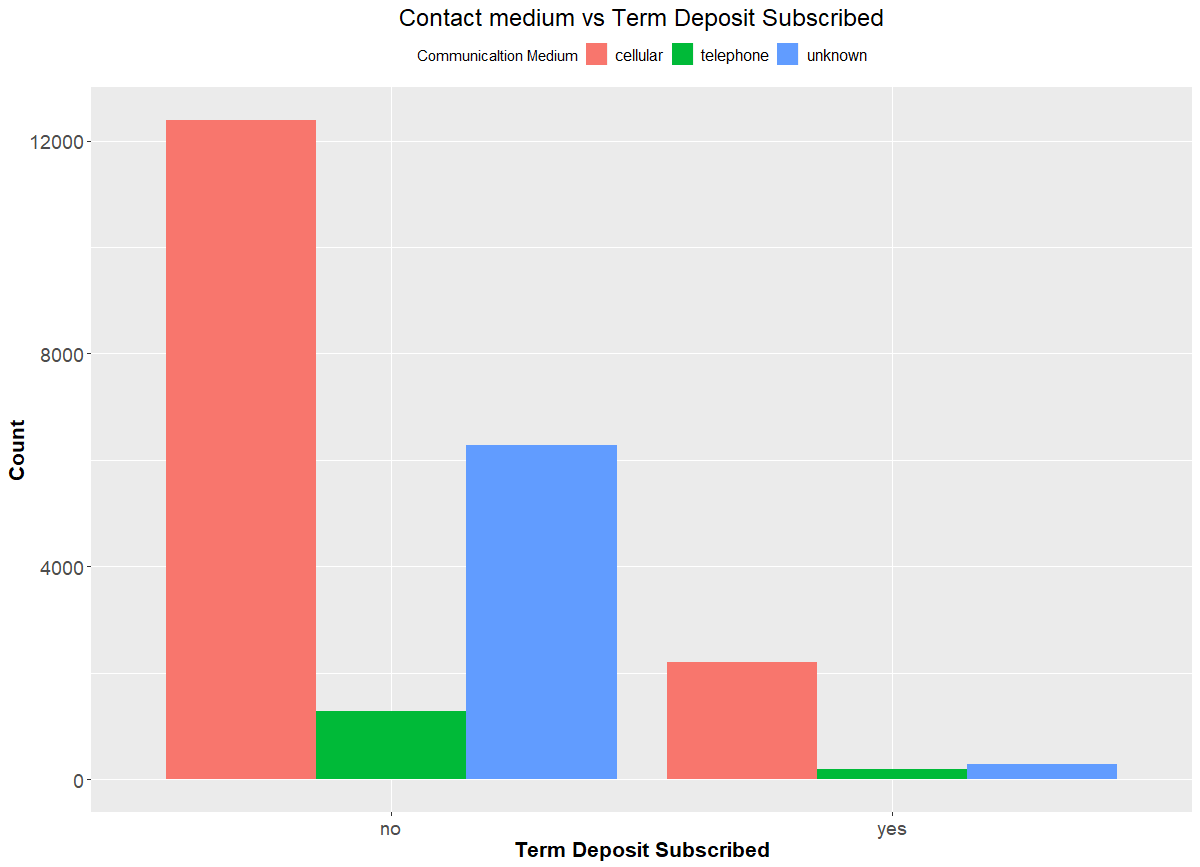


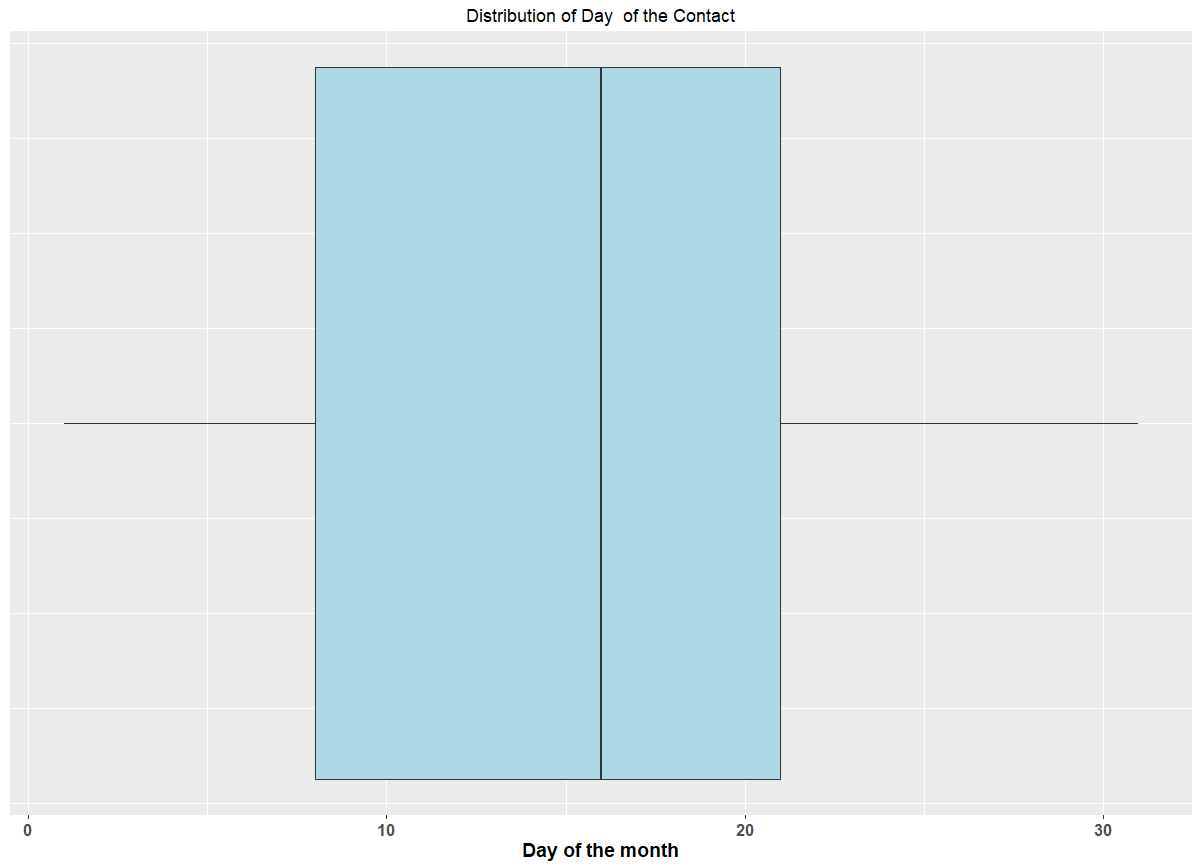


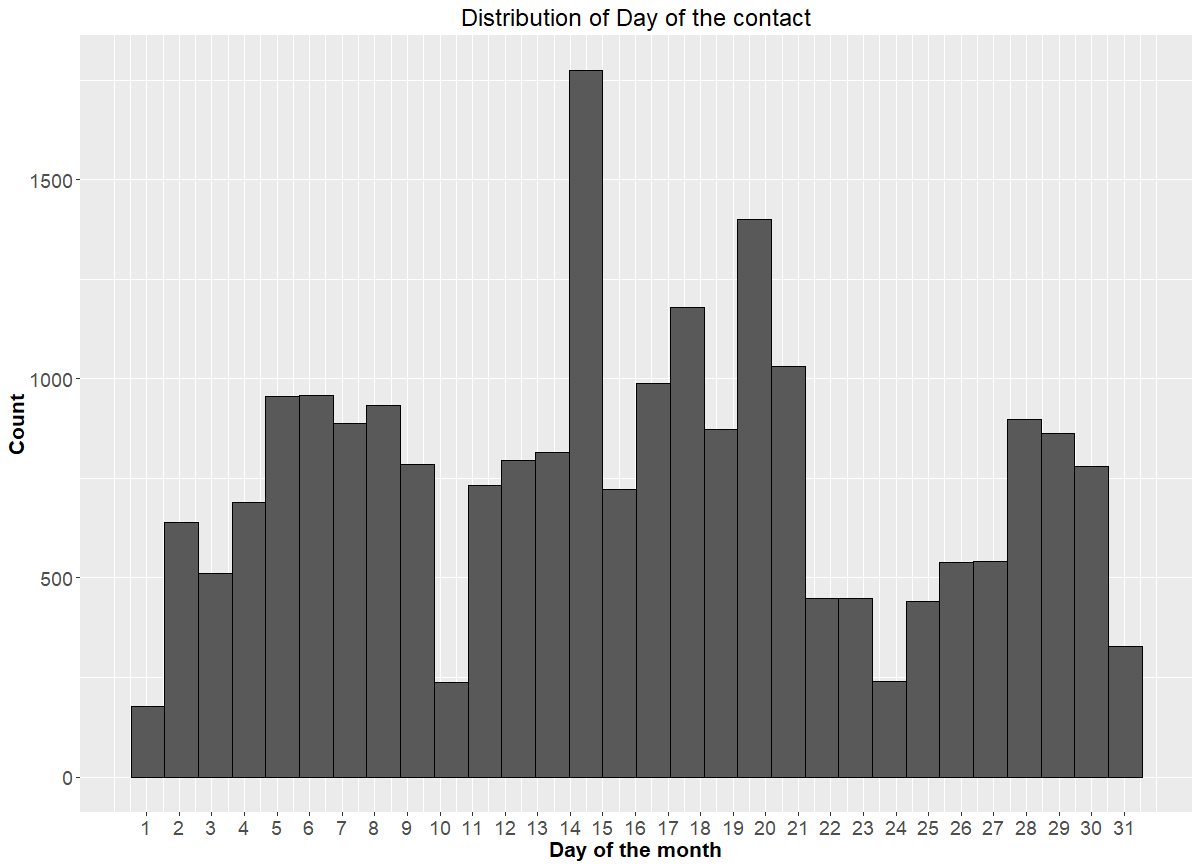


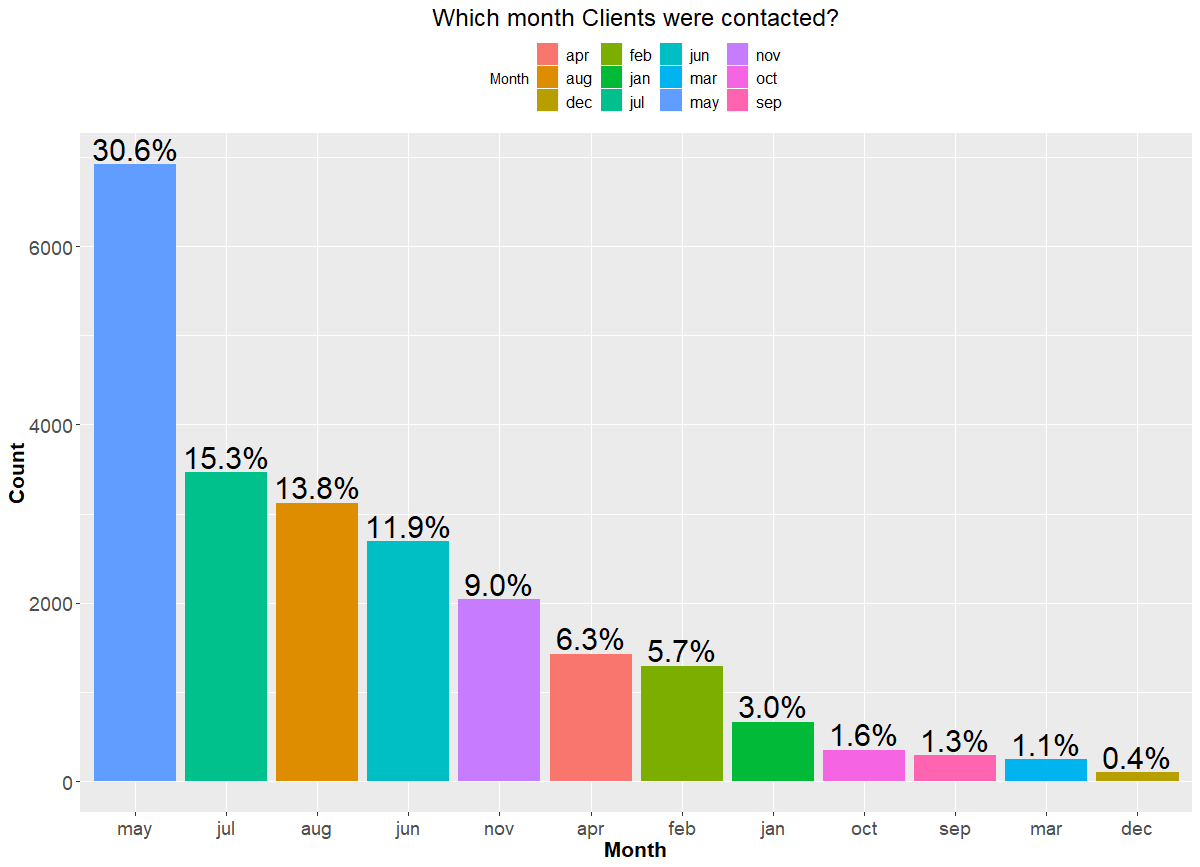


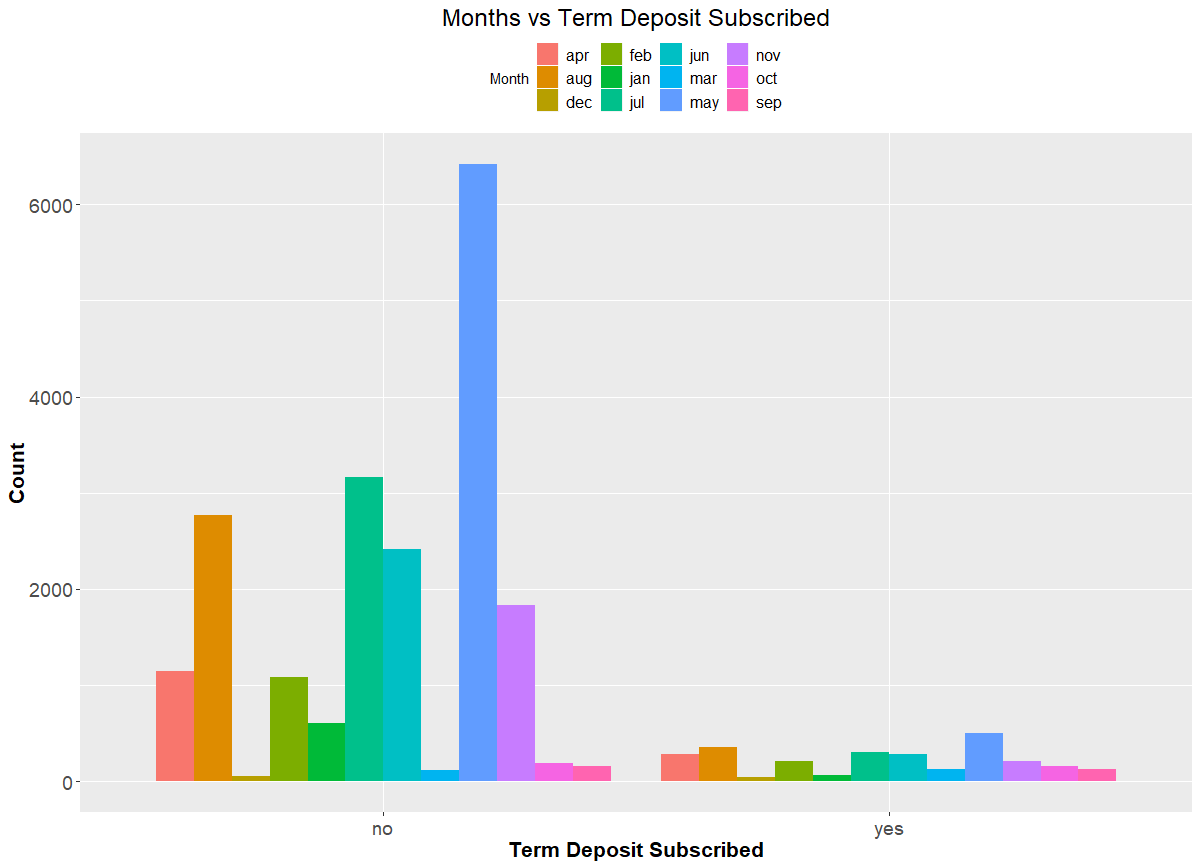


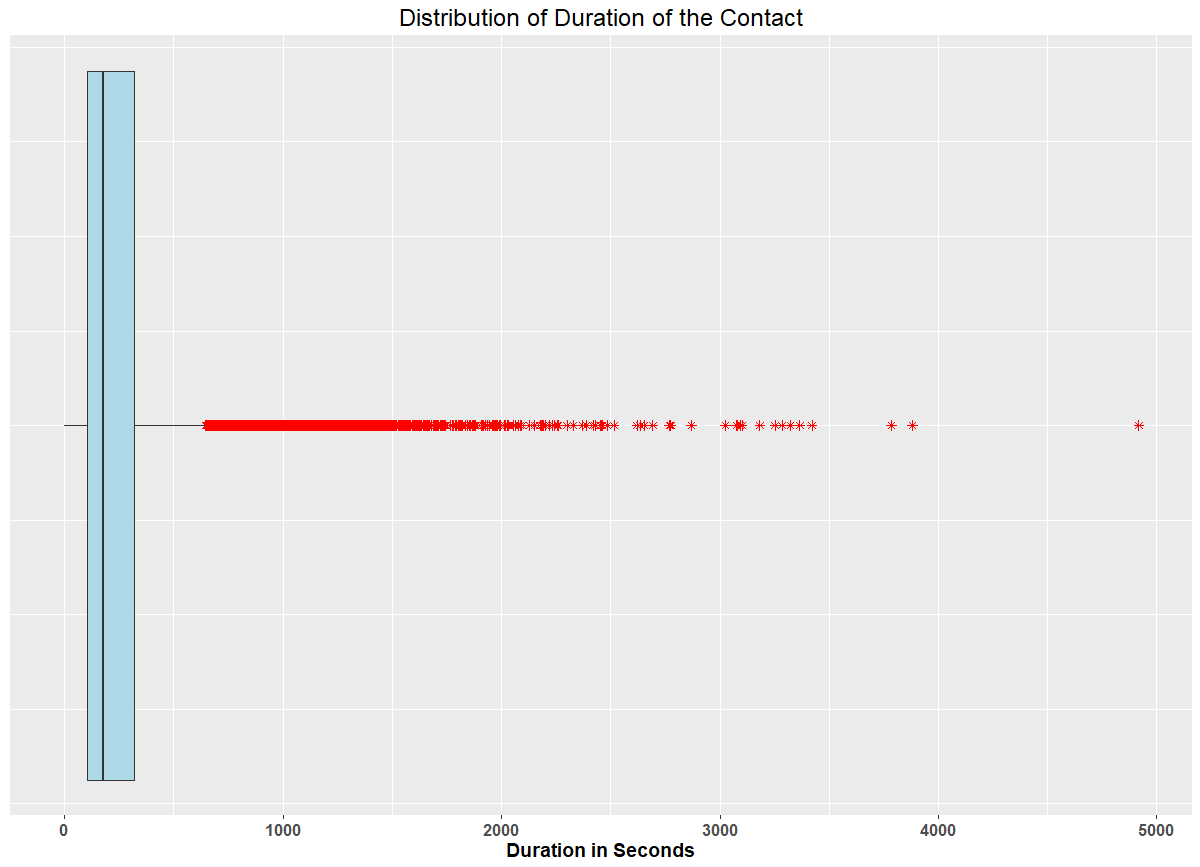


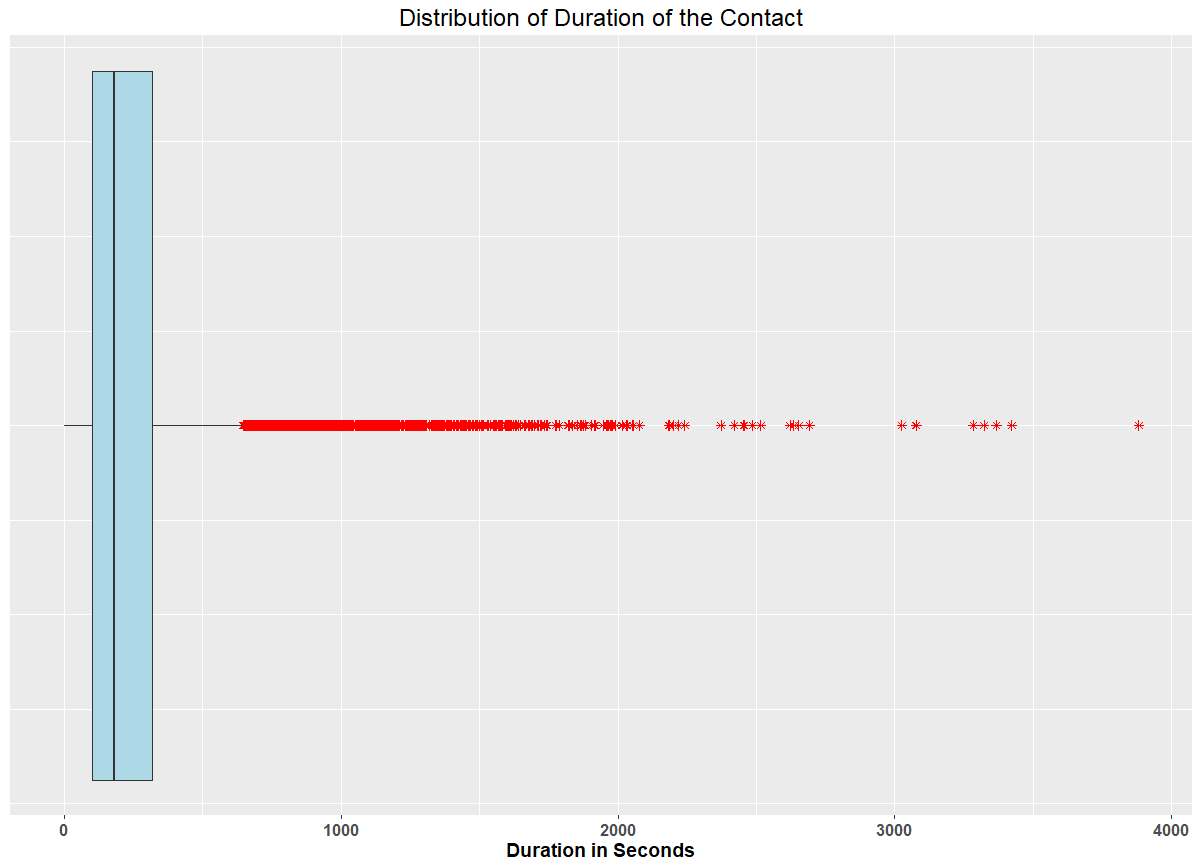


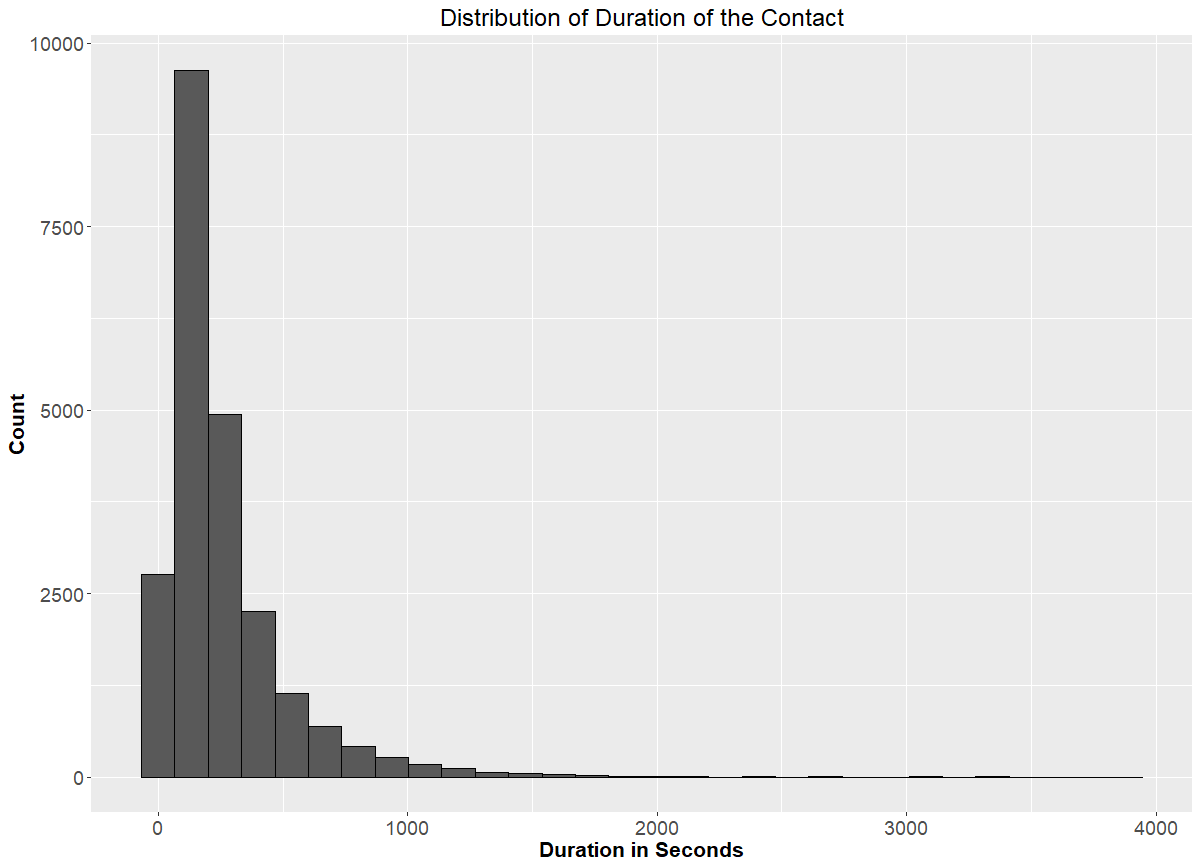


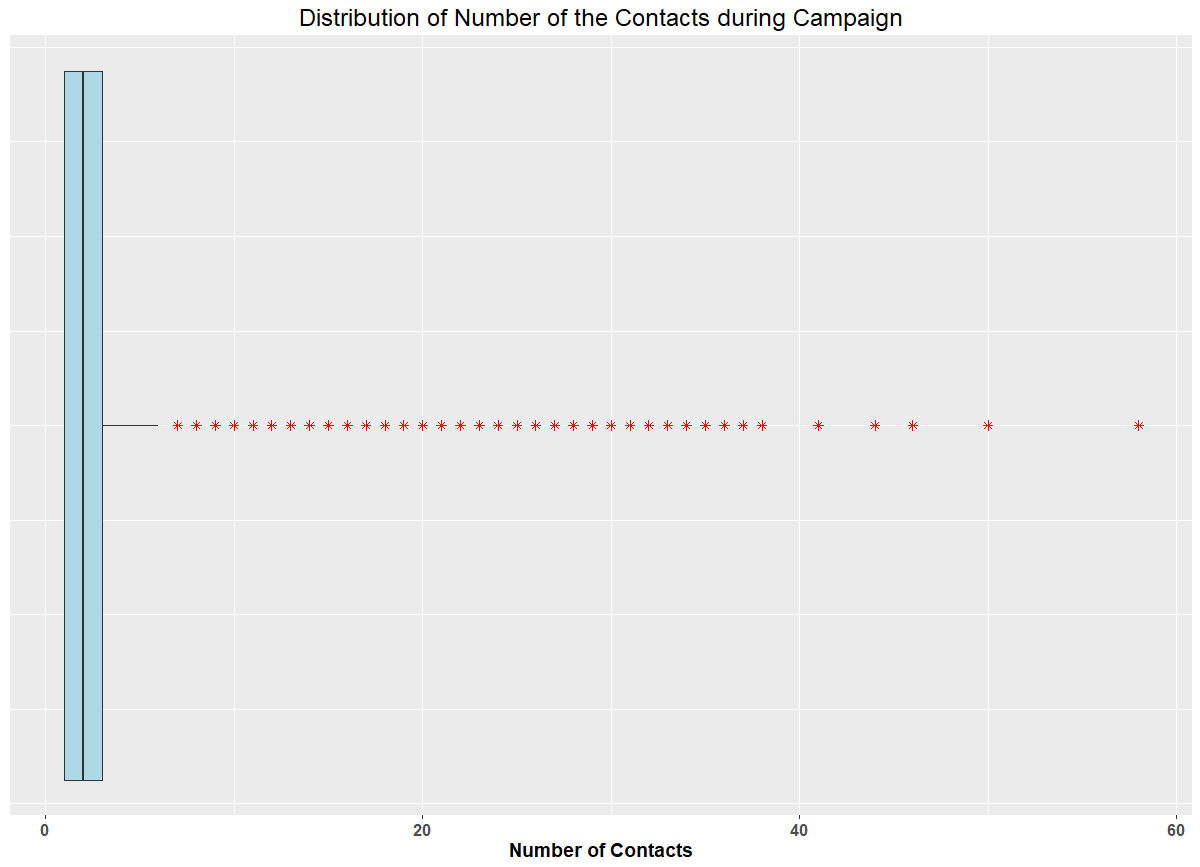


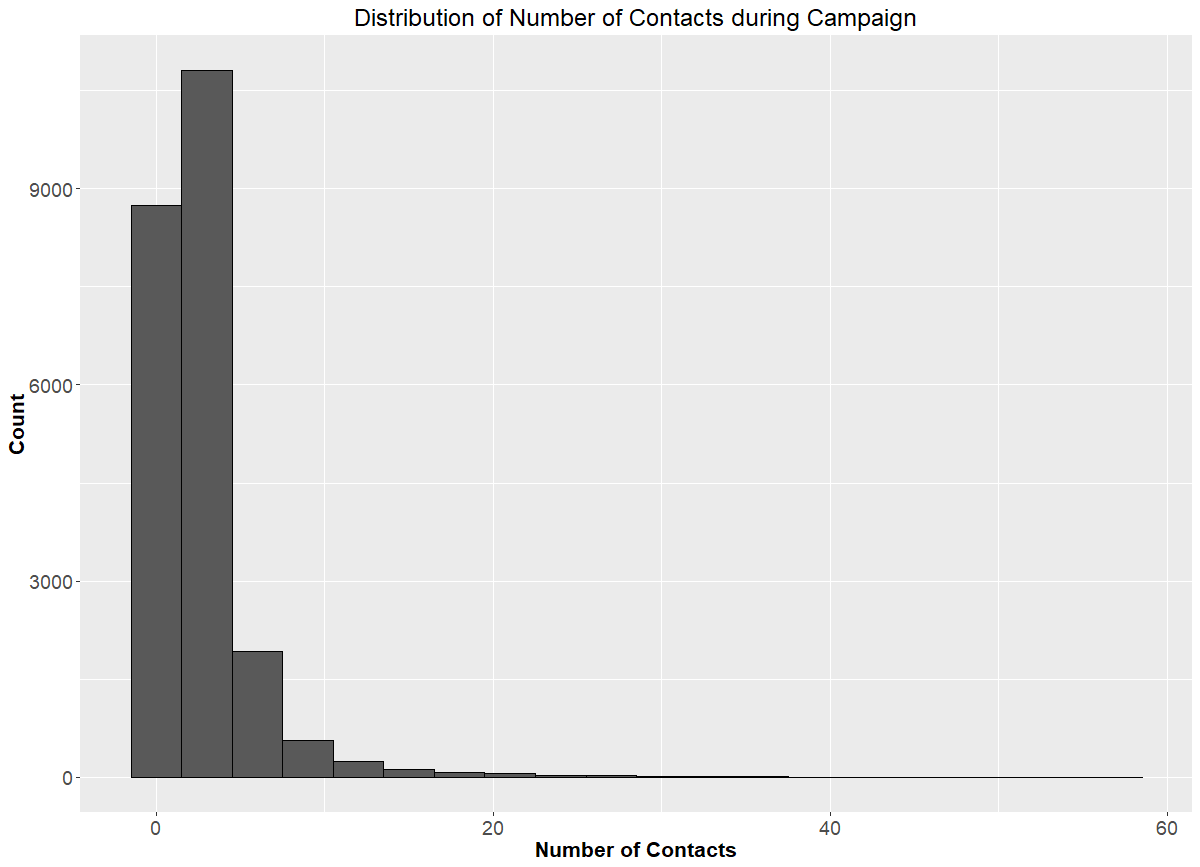


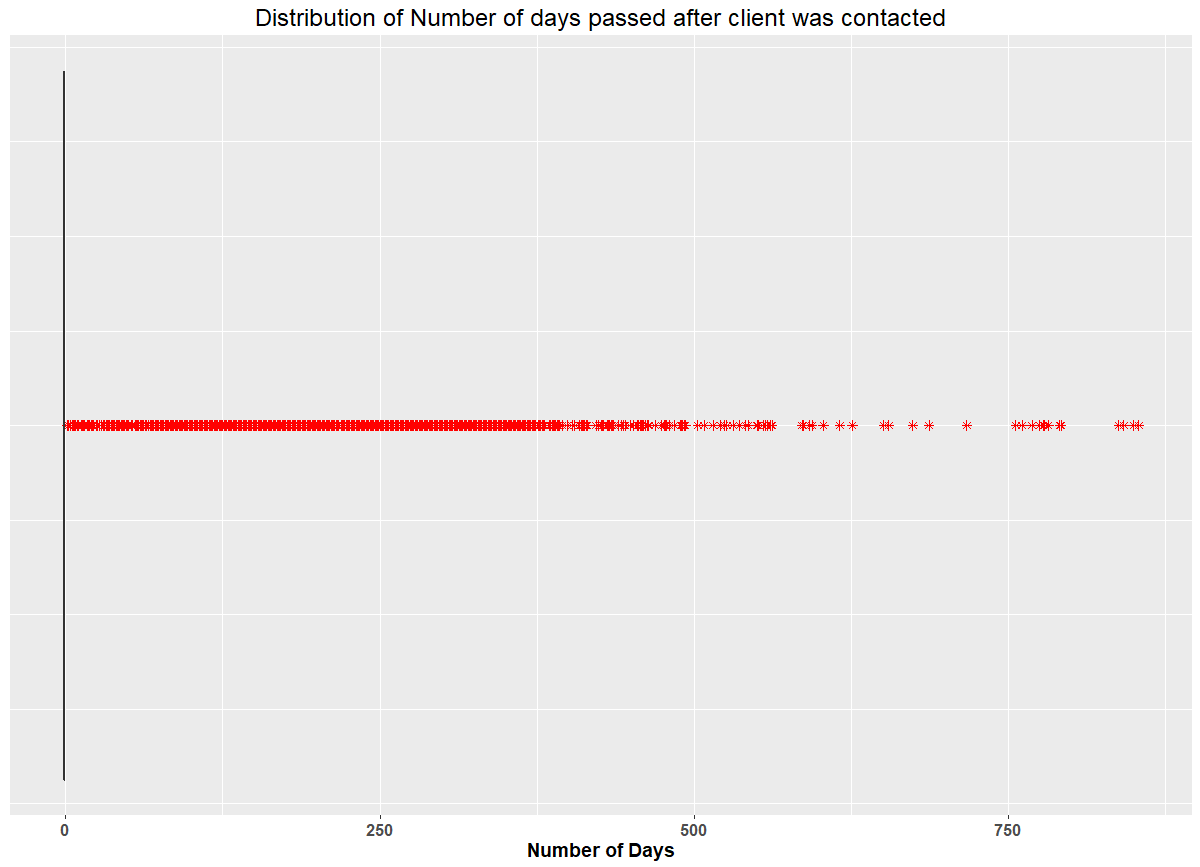


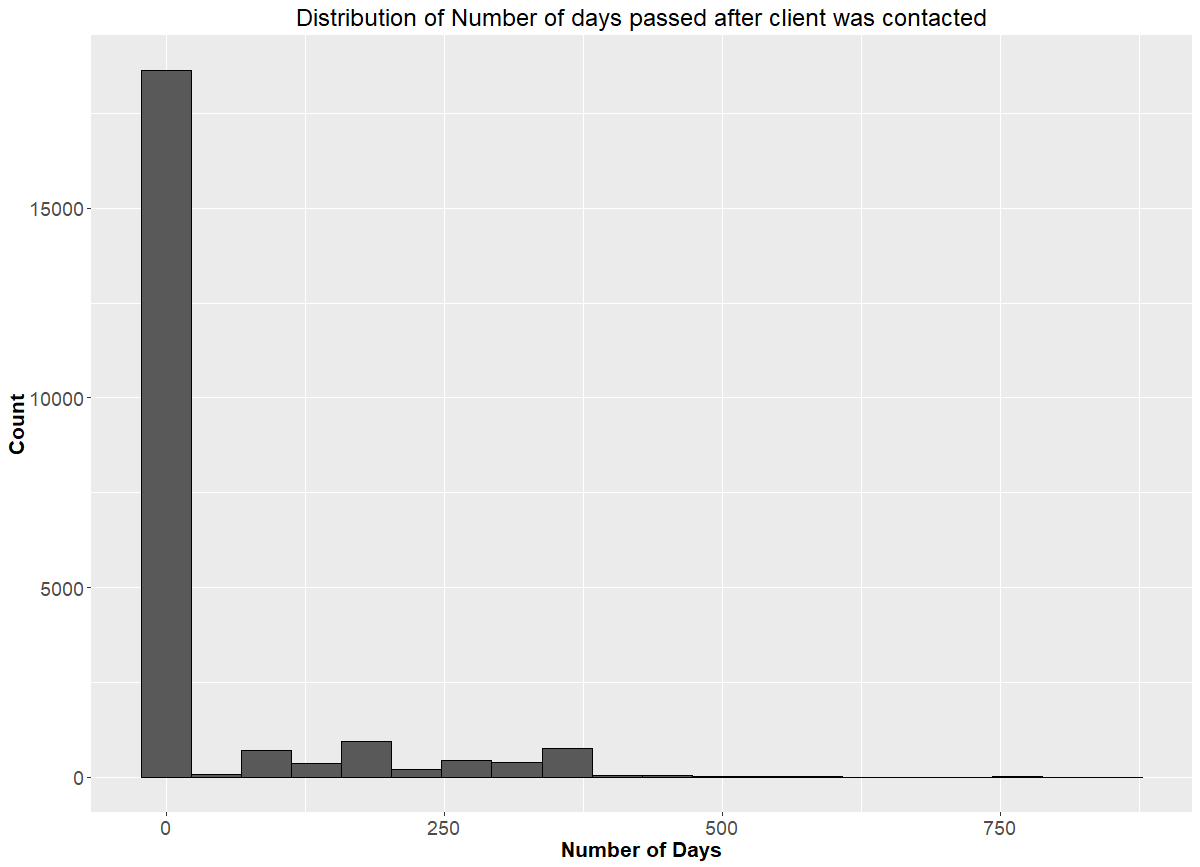


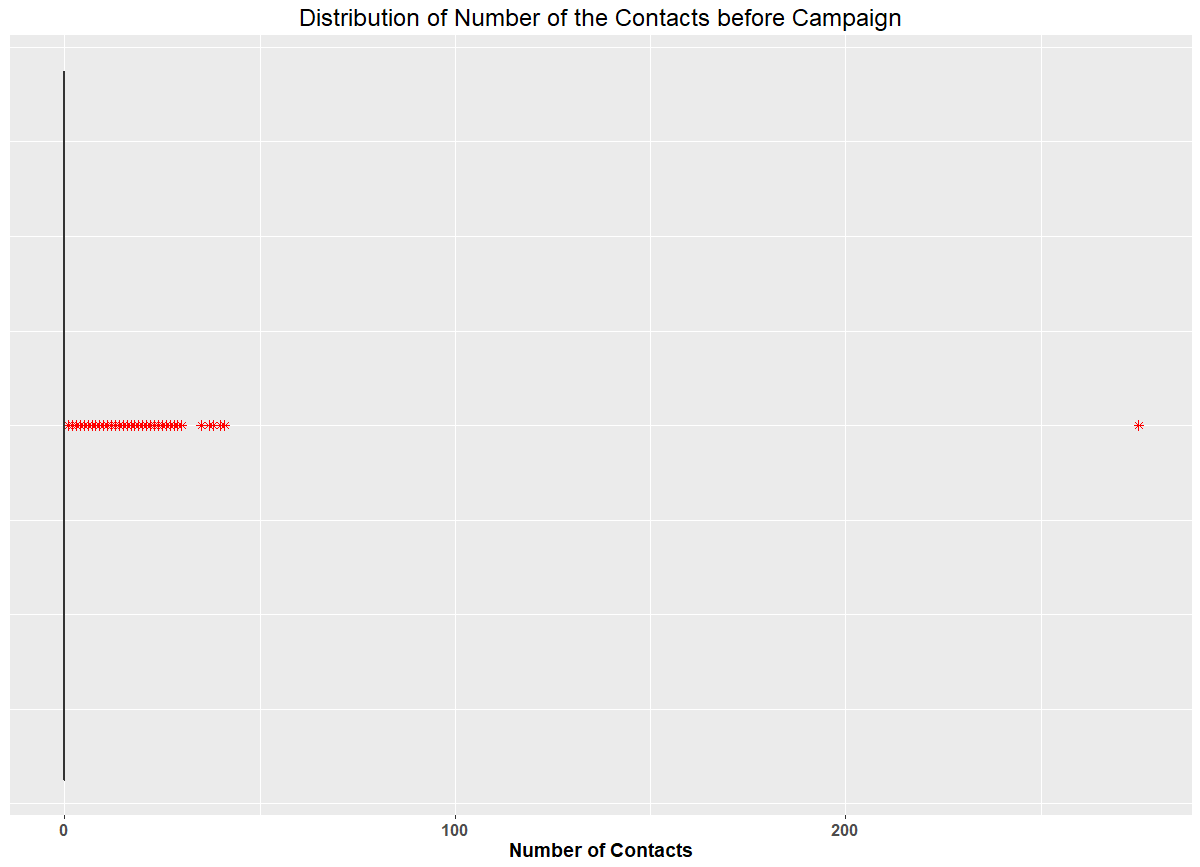


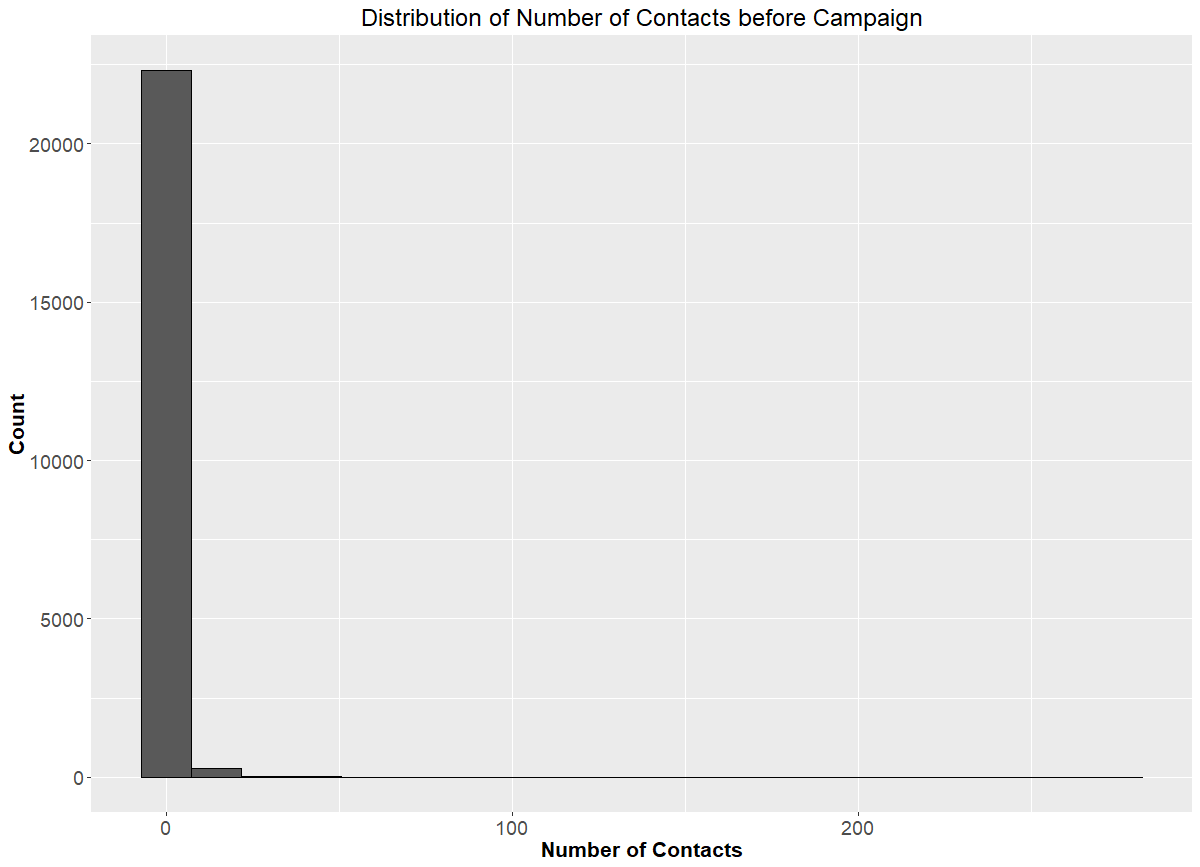


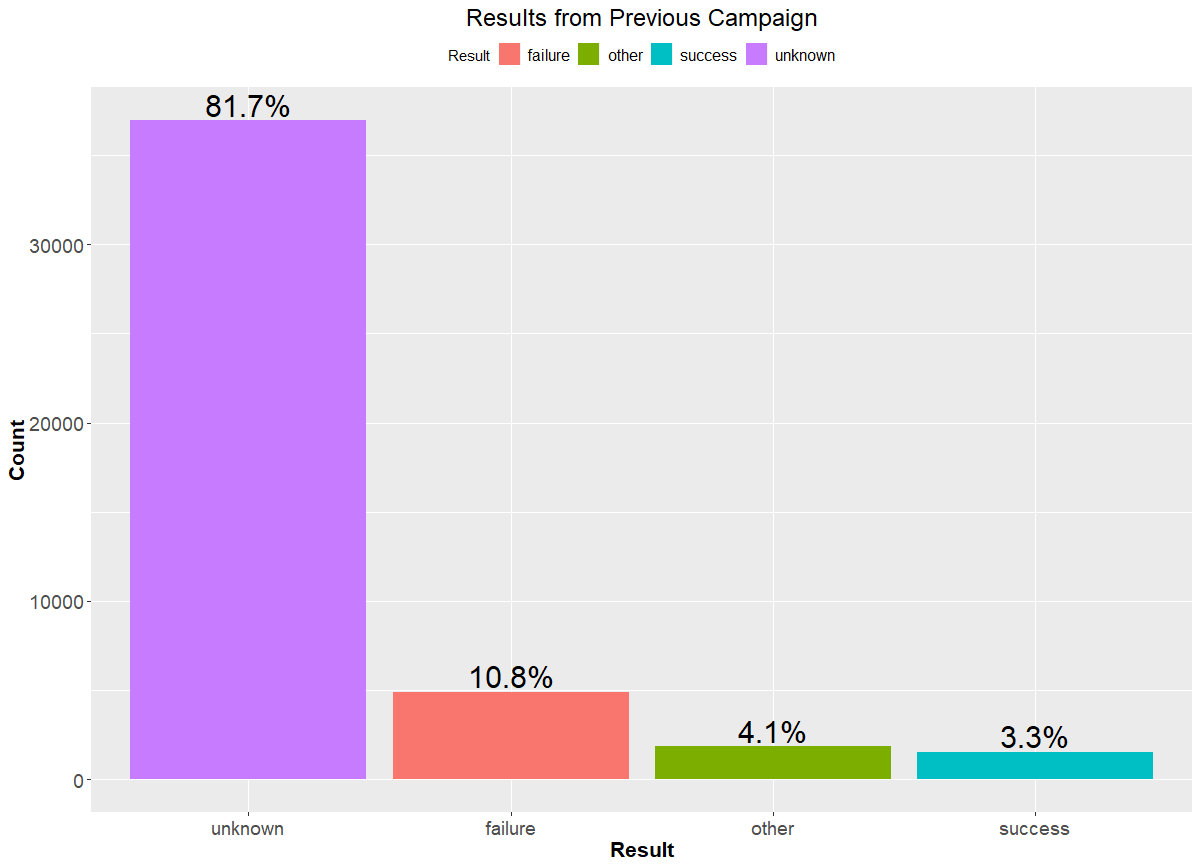


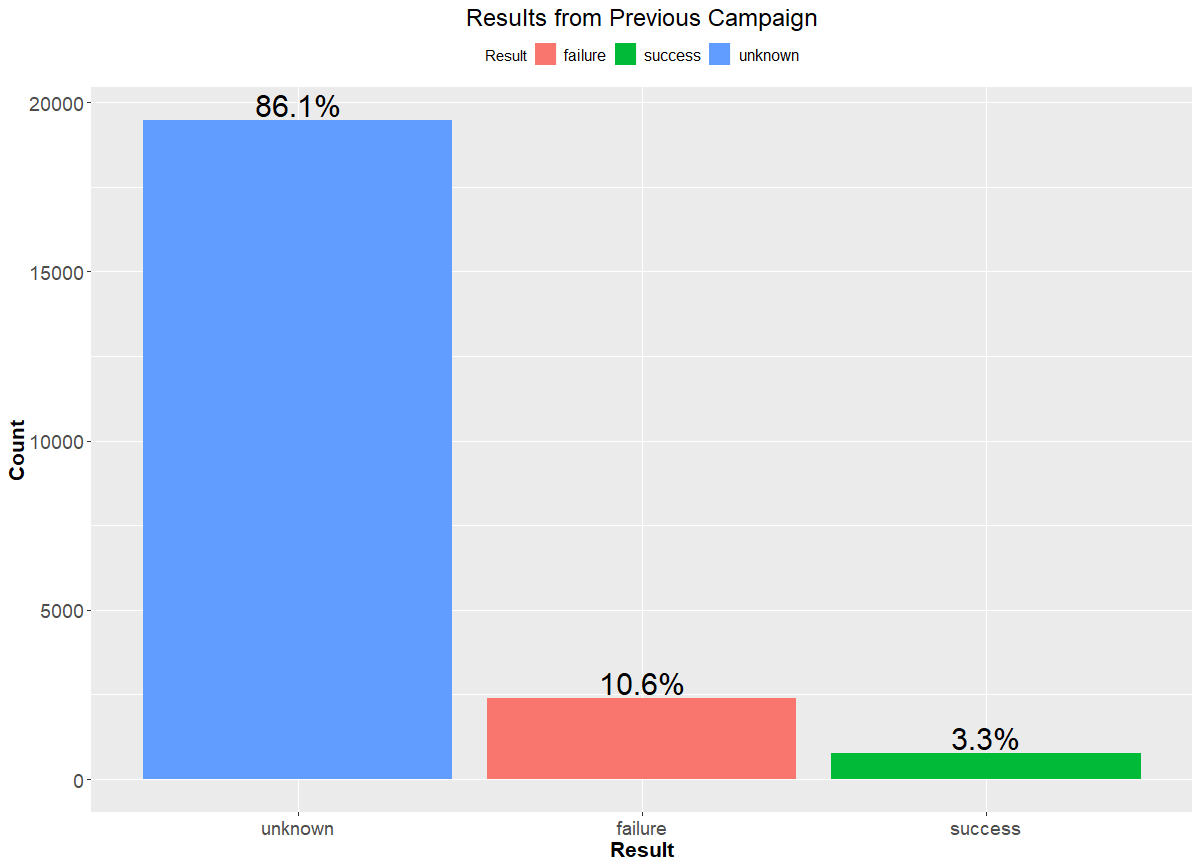


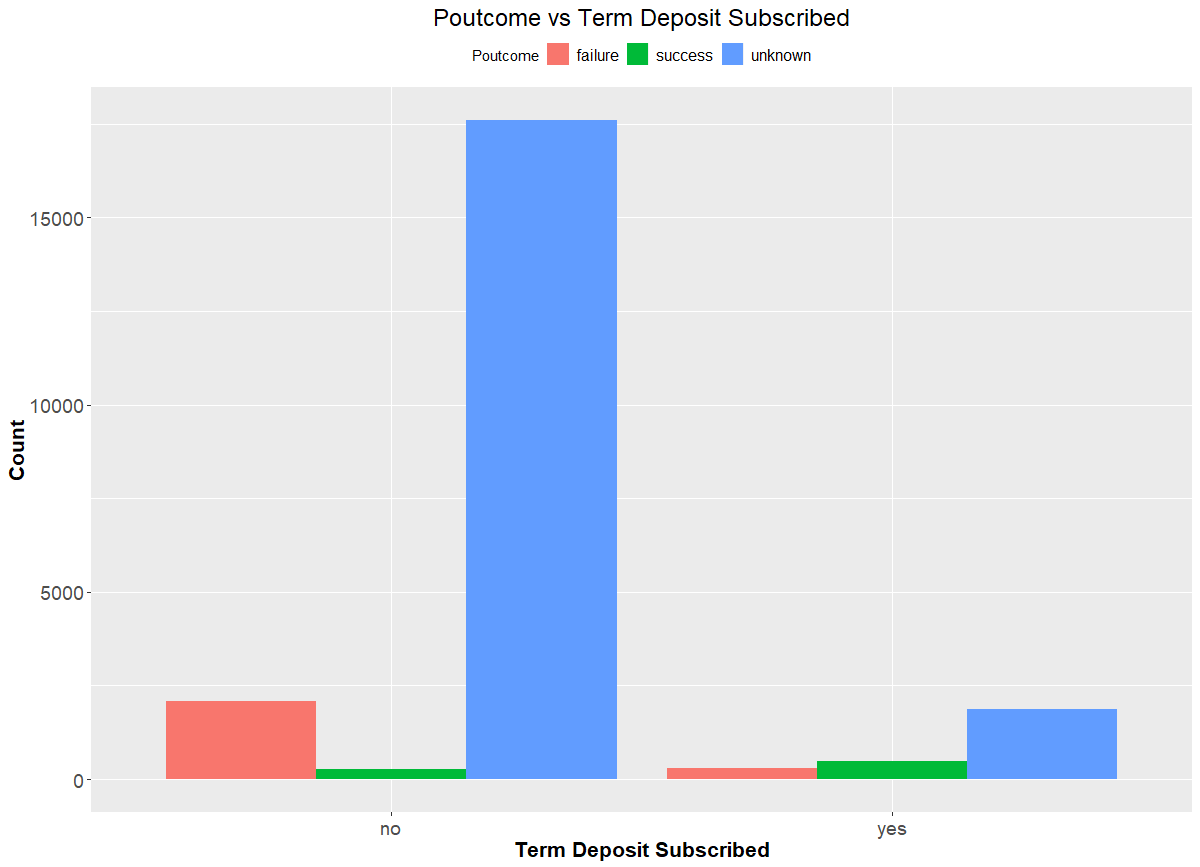












**Results**

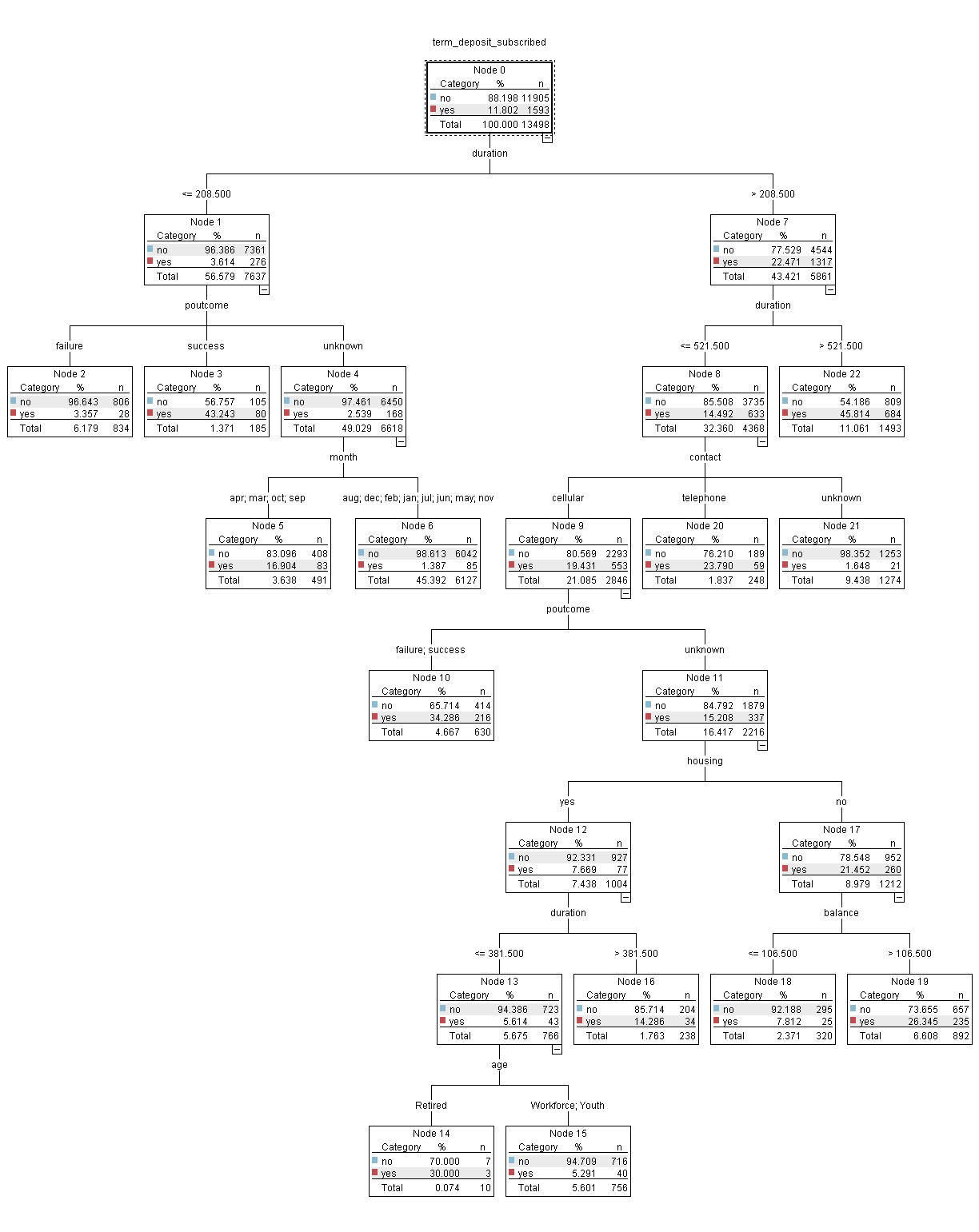
The best model is C5.0 with 30% pruning severity, 500 minimum nodes in child branch, with global pruning, with cost ratio of 9:1, with overall accuracy of 74.5%, with 73% accuracy and lift 1.11 for class ‘no’ and with 89% accuracy and lift 2.6 for class ‘yes’. We wanted to correctly identify ‘yes’ labels so we compromised the overall accuracy and class ’no’ accuracy. The accuracies and lift were consistent throughout all the dataset so we can rely on the model. We can see results in Fig *NUMBER.*

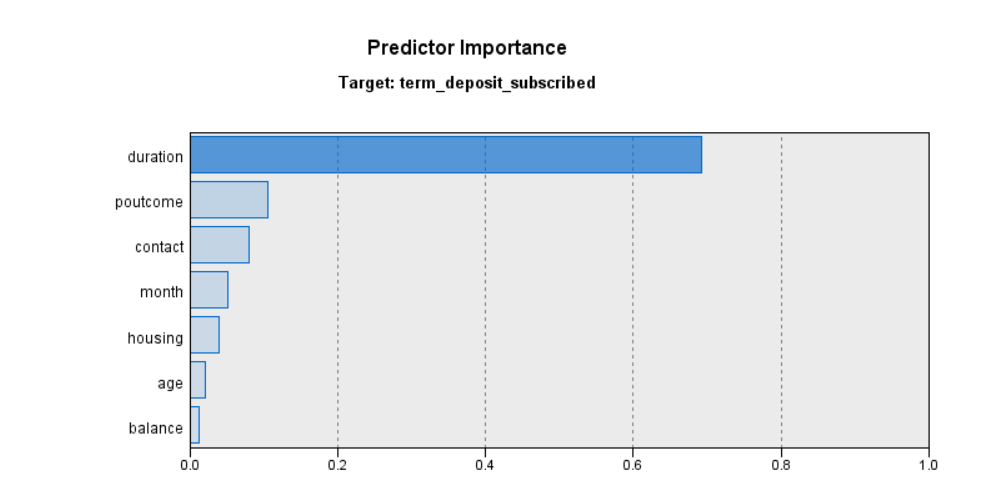
Many predictors were given as an input but we can see that the 3 most important of them are duration, poutcome and contact in Fig *NUMBER*. We can also see most important rules in the table below[For all rules Table *NUMBER*].

|  |  |  |
| --- | --- | --- |
| **#Rule** | **Rule statement** | **Classified as** |
| 1 | duration <= 208.5 and poutcome = ‘failure’ | no |
| 2 | duration <= 208.5 and poutcome = ‘success’ | yes |
| 3 | duration <= 208.5 and duration > 521.5 | yes |
| 4 | duration <= 208.5 and duration <= 521.5 and contact = ‘telephone’ | yes |
| 5 | duration <= 208.5 and duration <= 521.5 and contact = ‘unknown’ | no |
| 6 | duration <= 208.5 and duration <= 521.5 and contact = ‘cellular’ and poutcome = [‘failure’, ‘success’] | yes |

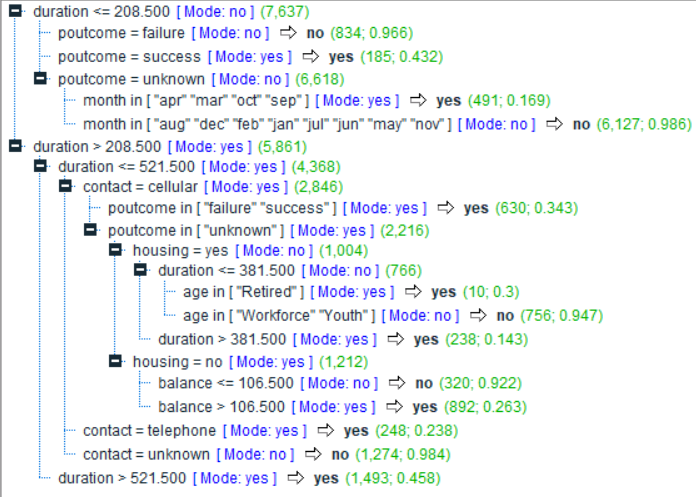
**Table *NUMBER***

First rule is obvious as duration and poutcome were the most important predictors. And also, if duration of the call is shorter, it seems that client is not interested in the campaign details and s/he is not looking forward to subscribing the term deposit. Same goes with the result of previous campaign(poutcome), if a client didn’t invest in the previous campaign s/he is more likely not to subscribe the term deposit. So that’s why those observations are labelled as ‘no’. Second rule is the exact opposite of the former one and we can see that it is classified as ‘yes’. Third rule states the same that higher the duration higher the interest of the client in the campaign and higher the chances of the client subscribing the term deposit. Now the third most important predictor comes into the picture which is contact. Forth rule indicates that if communication medium is telephone and duration is high then those observation can be classified as ‘yes’. The clients are using telephone that suggests that either clients are retired or housewives or someone who stays at home. As these clients have money and easy to convince, so they are labelled as ‘yes’. Fifth rule same as forth rule, different in the terms of only communication medium which is ‘unknown’. Unknown can mean many things that client was not contacted or contacted through mediums like word of mouth by bank employee, other platforms like bank’s mobile application or emails. They are labelled as ‘no’. Sixth rule is the combination of all the above rules except the communication medium is ‘cellular’ and poutcome is either ‘failure’ or ‘success’ and they are labelled as ‘yes’. These rules show that duration, poutcome and the contact is the most associated with the outcome whether client subscribes a term deposit or not.





|  |  |  |
| --- | --- | --- |
| **#Rule** | **Rule statement** | **Classified as** |
| 1 | duration <= 208.5 and poutcome = ‘failure’ | no |
| 2 | duration <= 208.5 and poutcome = ‘success’ | yes |
| 3 | duration <= 208.5 and poutcome = ‘unknown’ and month = [‘apr’, ‘march’, ‘oct’, ‘sep’] | yes |
| 4 | duration <= 208.5 and poutcome = ‘unknown’ and month != [‘apr’, ‘march’, ‘oct’, ‘sep’] | no |
| 5 | duration <= 208.5 and duration > 521.5 | yes |
| 6 | duration <= 208.5 and duration <= 521.5 and contact = ‘telephone’ | yes |
| 7 | duration <= 208.5 and duration <= 521.5 and contact = ‘unknown’ | no |
| 8 | duration <= 208.5 and duration <= 521.5 and contact = ‘cellular’ and poutcome = [‘failure’, ‘success’] | yes |
| 9 | duration <= 208.5 and duration <= 521.5 and contact = ‘cellular’ and poutcome = ‘unknown’ and housing = ‘no’ and balance <= 106.5 | no |
| 10 | duration <= 208.5 and duration <= 521.5 and contact = ‘cellular’ and poutcome = ‘unknown’ and housing = ‘no’ and balance > 106.5 | yes |
| 11 | duration <= 208.5 and duration <= 521.5 and contact = ‘cellular’ and poutcome = ‘unknown’ and housing = ‘yes’ and duration > 381.5 | yes |
| 12 | duration <= 208.5 and duration <= 521.5 and contact = ‘cellular’ and poutcome = ‘unknown’ and housing = ‘yes’ and duration <= 381.5 and age = ‘Retired’ | yes |
| 13 | duration <= 208.5 and duration <= 521.5 and contact = ‘cellular’ and poutcome = ‘unknown’ and housing = ‘yes’ and duration <= 381.5 and age = [‘Workforce’, ‘Youth’] | no |



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Type** | **Min** | **Max** | **Median** | **Mean** | **Std.Dev** |
| age | Factor | Retired: 883 Youth: 8385 Workforce: 13338 | | | | |
| job | Factor | admin: 2538 blue-collar: 4831 entrepreneur: 743 housemaid: 634 student: 461 management: 4815 retired: 1149 self-employed: 770 services: 2102 technician: 3779 unemployed: 643 unknown: 141 | | | | |
| marital | Factor | married: 13615 single: 6390 divorced: 6390 | | | | |
| education | Factor | primary: 3430 secondary: 11541 tertiary: 6732 unknown: 903 | | | | |
| default\_credit | Binary | no: 22198 yes: 408 | | | | |
| balance | Numeric | -8019 | 98417 | 442 | 1349 | 3013.346 |
| housing | Binary | yes: 12588 no: 10018 | | | | |
| loan | Binary | yes: 3598 no: 19008 | | | | |
| contact | Factor | cellular: 14577 telephone: 1482 unknown: 6547 | | | | |
| day | Numeric | 1 | 31 | 16 | 15.85 | 8.29 |
| month | Factor | jan: 668 feb: 1293 mar: 246 apr: 1427 may: 6917 jun: 2691 jul: 3464 aug: 3121 sep: 288 oct: 351 nov: 2040 dec: 100 | | | | |
| duration | Numeric | 2 | 3881 | 181 | 258.6 | 257.1873 |
| campaign | Numeric | 1 | 58 | 2 | 2.756 | 3.0656 |
| pdays | Numeric | -1 | 854 | -1 | 39.4 | 98.998 |
| previous | Numeric | 0 | 275 | 0 | 0.5768 | 2.599 |
| poutcome | Factor | failure: 2389 success: 756 unknown: 19461 | | | | |
| term\_deposit\_subscribed | Factor | no: 19939 yes: 2667 | | | | |