EDA-Project

Team Members

Group Name	INSIGHT

Name	Farzin Valiloo	Maryam Aliakbari

Background & Motivation

Short Problem Statement

Problem Statement

How to increase effectiveness of direct marketing campaign based on the dataset

Industry Knowledge domain

Investopedia

Telemarketing is **the direct marketing of goods or services to potential customers over the telephone or the Internet**.
Four common kinds of telemarketing include outbound calls, inbound calls, lead generation, and sales calls. [1]

Bank is one of the organization use telemarketing method for selling banking products or services.

The Telemarketing Main Problems

As much as we are crazy for telemarketing, we cannot ignore the persistent problems facing it today! [3]

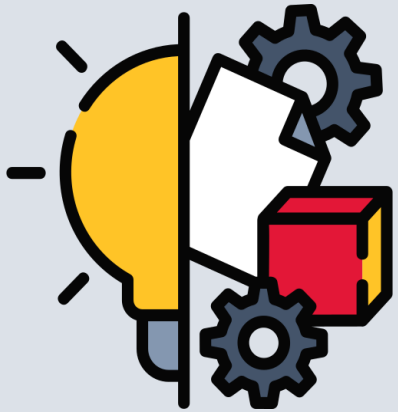
Rejections

B2B marketers are getting rejected as more people have cynical whenever they receive a telemarketer

Increased costs

The Companies cannot initiate a cold calling campaign without churning out a hefty sum of money

Background & Motivation



Background & Motivation

- A good telemarketing campaign plan clearly outlines the details and the scope of the campaign, so everyone is clear about what it involves and how it'll work. [2]
- Telemarketing campaigns can help us reach a group of targeted prospects or customers to communicate a message, gather feedback, and determine a next step for the relationship. [2]

Background & Motivation

- Marketing strategy could evolve over time.
- As the bank learn more about what is and what is not working, it will build a deeper understanding of the marketplace.
- Building an effective campaign strategy can help bank and its marketing department to create the better positioning for understanding the customers requirements.

Benefits of effective Marketing Campaigns [4]

- Attracts more sales
- Improve the reputation of the bank
- Improve Undertesting of the market & customers
- Better long-term marketing campaigns

Problem Statement Project Proposal &

Analysis Questions

Question

1. Who are the top targeted prospects in terms of successful outcome?
2. What is the best time to contact prospect customers?
3. What are the top characteristics of a successful call?

Project Proposal Statement

The product of this project will be a descriptive analytics that examines the characteristics of a successful phone call for the purpose of direct marketing in terms of targeted prospect customers and properties of the call session, based on the data set.

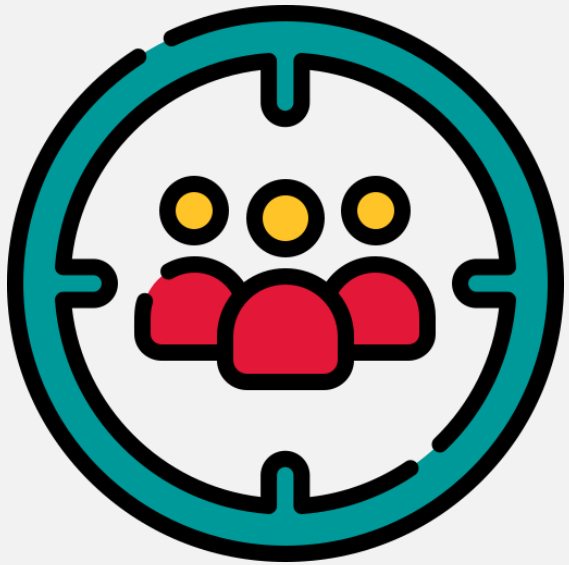
The output of this project may help marketing managers in similar industries (e.g. finance, other service industries) to improve the performance of their direct marketing efforts by increasing the efficiency and effectiveness of the phone calls by exploring some of the aspects of successful direct marketing calls recorded in the dataset.

Proposal Statement

Project Audience

Audiences

- Marketing Department
- Medium & High-level Managers
- Sales Department
- Call Centers



Dataset Description

To describe and analyze the data, we need to understand the nature of data.

The type of statistical analysis is influenced by the type of data and can be performed on it.

About Dataset

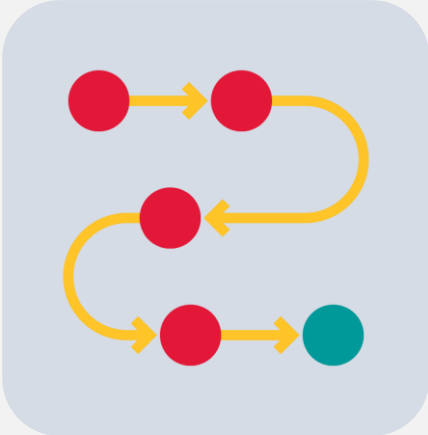
Title	Bank Marketing Data
Source	Data.world
Dataset	[5] & [6]
Data Link	[7]
Description	The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. [6]
Main Article	Moro et al., 2014, S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014 [5]

Dataset Characteristic

- The data 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.

EDA Steps

In this chapter, for the EDA:



- The data types of variables will be explored
- Each variable will be introduced
- A plot and short description of each variable will be discussed.
- A general exploration of correlation among numerical variables will be provided

DataFrame Structure

Primary Dataset Structure

No. Variables	21
No. Rows	41188

New-Structured Dataset

No. Variables	16
No. Rows	41188

Social and Economic context Attributes that have been Eliminated based on Problem statement

Name	Description	Data Type
emp.var.rate	employment variation rate - quarterly indicator	Numeric
cons.price.idx	consumer price index - monthly indicator	Numeric
cons.conf.idx	consumer confidence index - monthly indicator	Numeric
euribor3m	euribor 3 month rate - daily indicator	Numeric
nr.employed	number of employees - quarterly indicator	Numeric

DataFrame Structure

Primary Dataset Structure

No. Variables	21
No. Rows	41188

New-Structured Dataset

No. Variables	16
No. Rows	41188

Social and Economic context Attributes that have been Eliminated based on Problem statement

Name	Description	Data Type
emp.var.rate	employment variation rate - quarterly indicator	Numeric
cons.price.idx	consumer price index - monthly indicator	Numeric
cons.conf.idx	consumer confidence index - monthly indicator	Numeric
euribor3m	euribor 3 month rate - daily indicator	Numeric
nr.employed	number of employees - quarterly indicator	Numeric

Data Type

Categorical

- Job
- Education
- Marital
- Credit_default
- Personal_loan
- Housing_loan
- Last_month
- Last_weekday
- PrevCampaign_result
- Call_type
- Campaign_Success

Numerical

- Age
- LastCall_Dur
- NewCampaign_CallNo
- PrevCampaign_CallNo
- Campaign_Intervals_Day

DataFrame Structure

Primary Dataset Structure

No. Variables	21
No. Rows	41188

New-Structured Dataset

No. Variables	16
No. Rows	41188

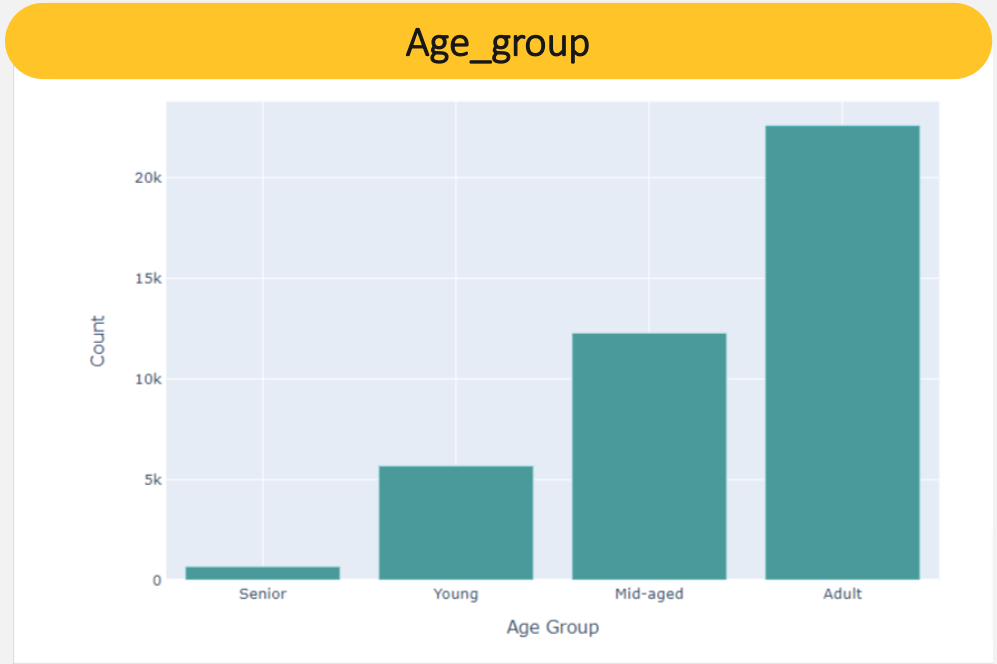
Social and Economic context Attributes that have been Eliminated based on Problem statement

Name	Description	Data Type
emp.var.rate	employment variation rate - quarterly indicator	Numeric
cons.price.idx	consumer price index - monthly indicator	Numeric
cons.conf.idx	consumer confidence index - monthly indicator	Numeric
euribor3m	euribor 3 month rate - daily indicator	Numeric
nr.employed	number of employees - quarterly indicator	Numeric

Define New Variable

- We define new variable “Age_Group” based on “Age” variable. It is a categorical variable.

Age_Group	
Age Range	Value
[15 , 30)	Young
[30 , 45)	Adult
[45 ,65)	Mid_aged
[65, 100)	Senior

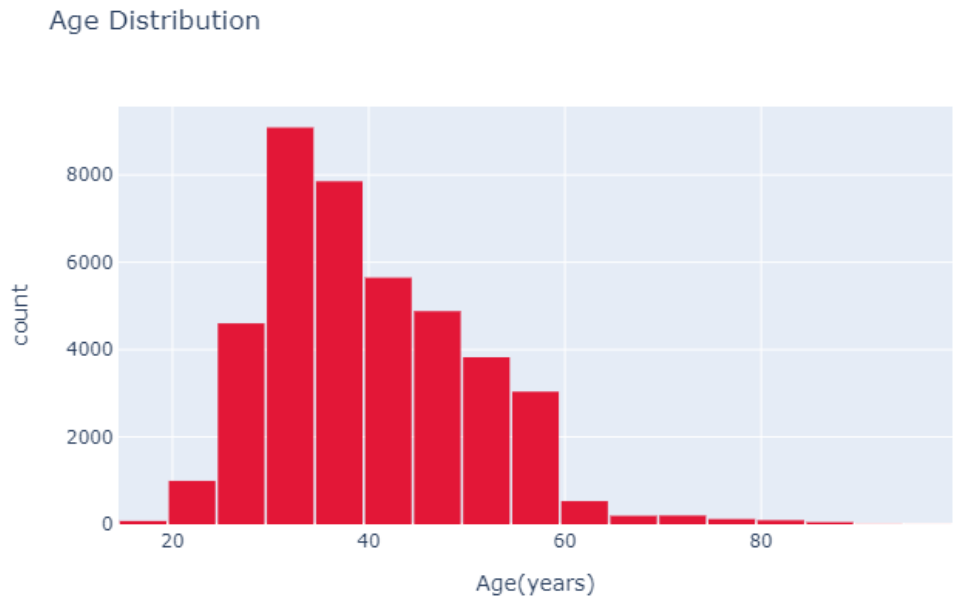


EDA:
Univariant Preliminary Visualization

Numeric Variables

Preliminary Visualization: Numeric Variables

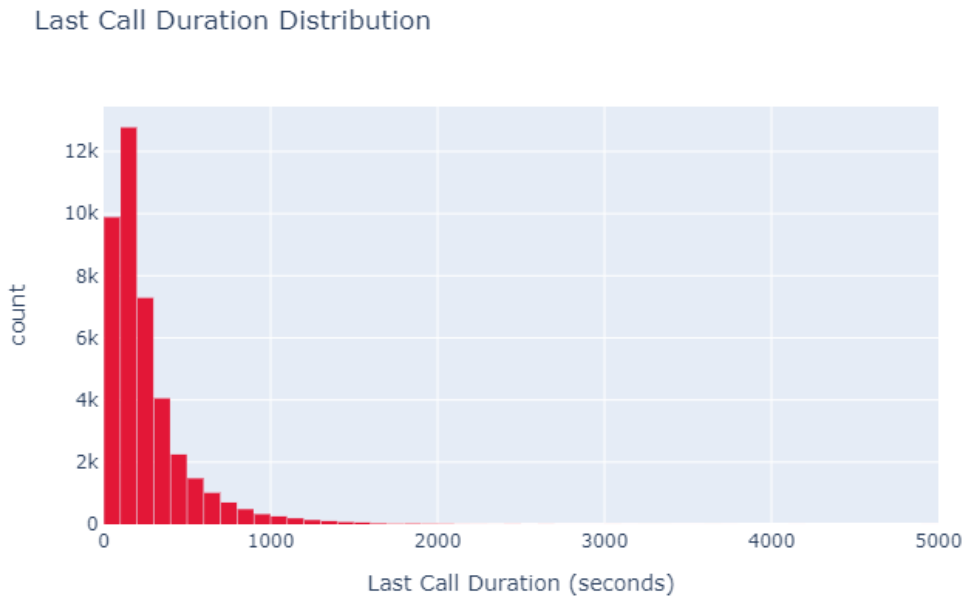
Age



The data for age variable is between 15 and 98 and more than 50% of the clients are younger than 40 years old.

Mean	Median	Var	Distinct No	Skew
40.02	38	108.6	78	Right

LastCall_Dur

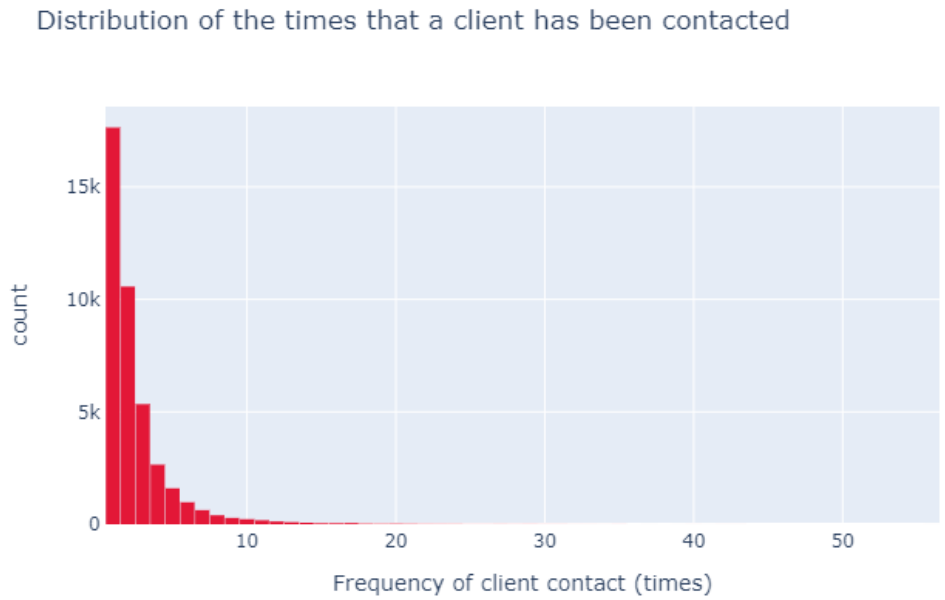


More than 30% of the call durations fall between 100 and 200 seconds. There maybe some outlier data in this variable.

Mean	Median	Var	Distinct No	Skew
258.3	180	259.3	1544	Right

Preliminary Visualization: Numeric Variables

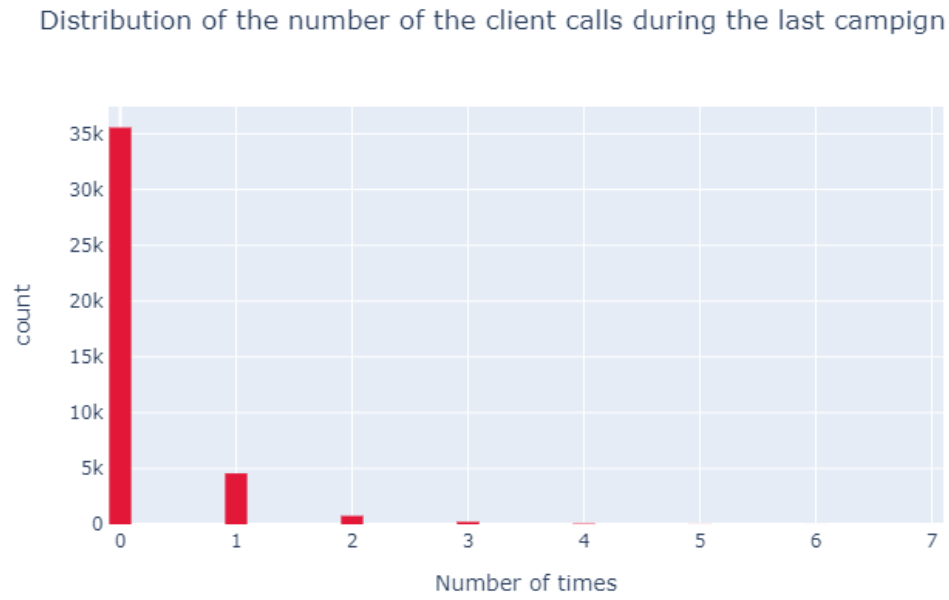
NewCampaign_CallNo



This variable is highly skewed and almost 70% of the observations were under 2. It has a long tail, and it has possible a few outlier data.

Mean	Median	Var	Distinct No	Skew
2.56	2	7.67	42	Right

PrevCampaign_CallNo

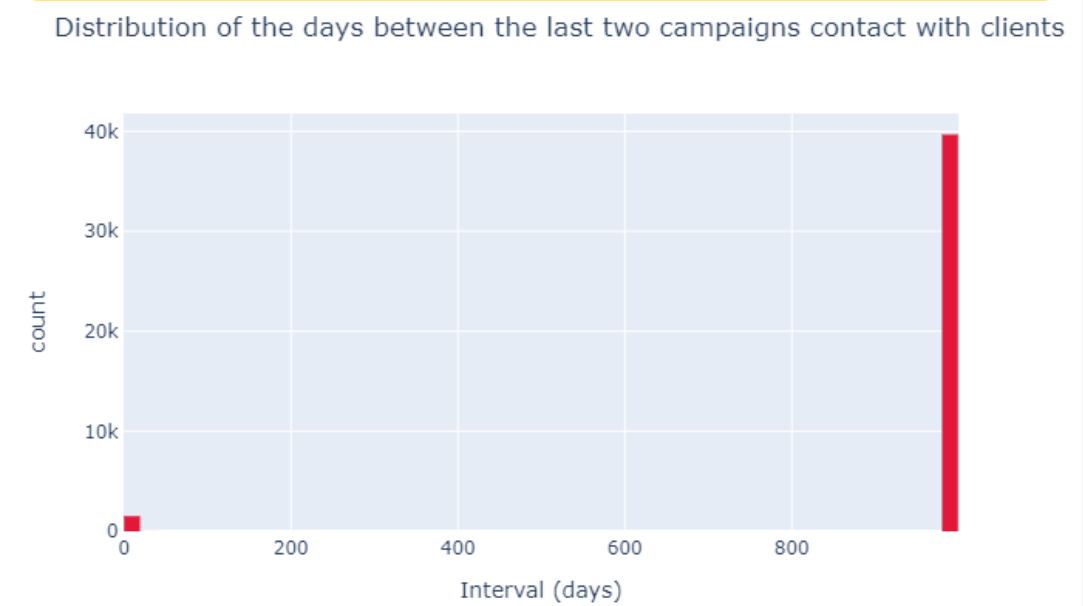


This plot shows that most of the clients has not been included in the previous campaign (about 85%).

Mean	Median	Var	Distinct No	Skew
0.17	0	0.25	8	Right

Preliminary Visualization: Numeric Variables

Campaign_Intervals_Day



This plot is consistent with the previous plot and shows that most of the clients have not been contacted during the previous campaign.

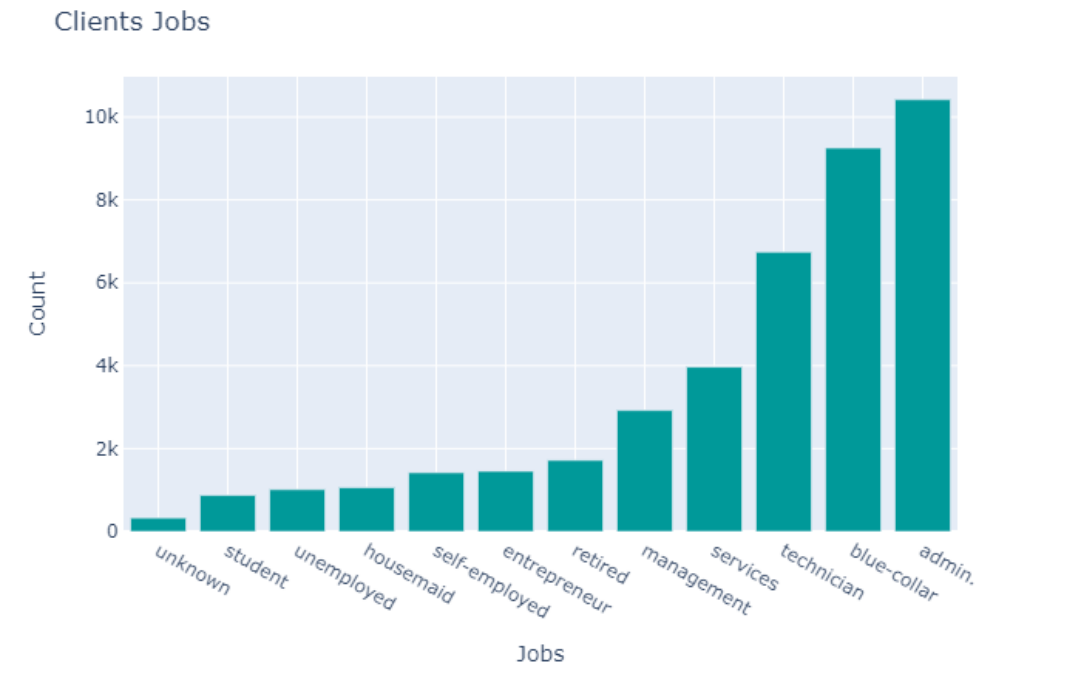
Mean	Median	Var	Distinct No	Skew
962.4	999	186.911	27	Left

EDA:
Univariate Preliminary Visualization

Categorical Variables

Preliminary Visualization: Categorical Variables

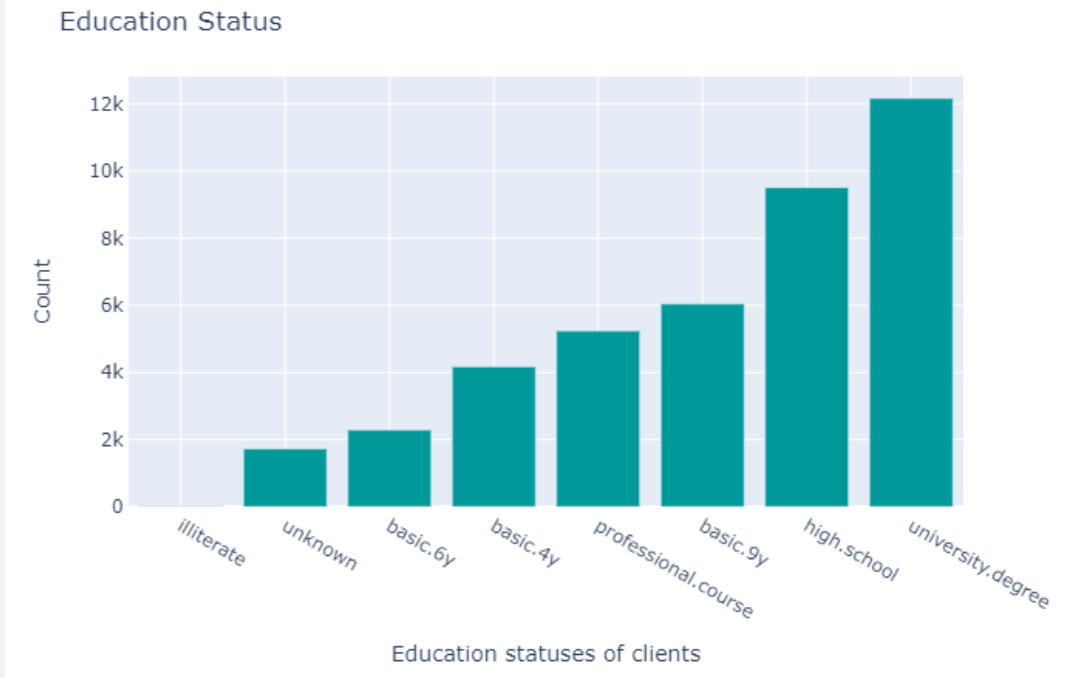
Job



About half of the phone calls are made with clients that either have administrative roles or are blue-collar workers.

Distinct No
12

Education

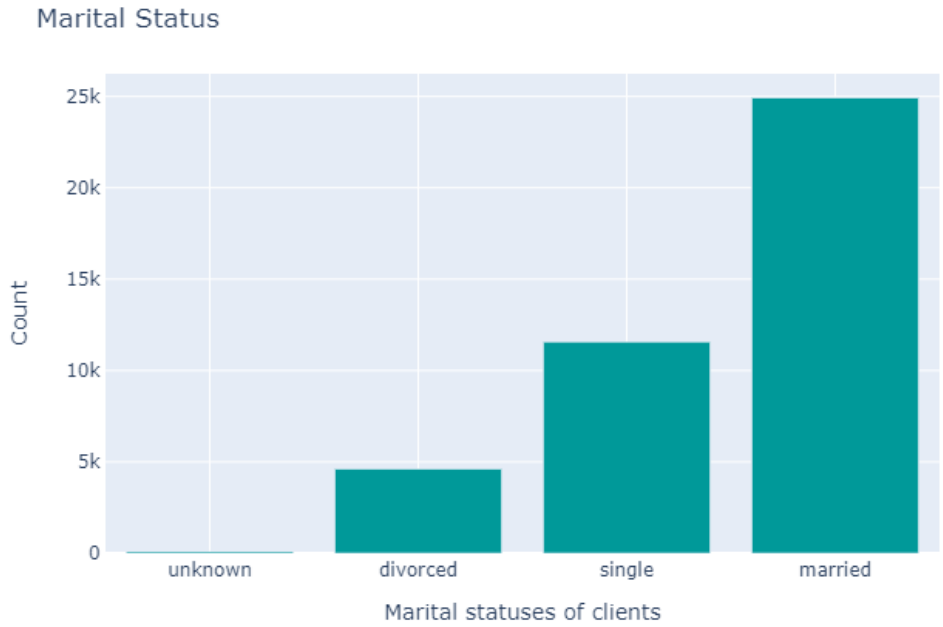


About 50% of the clients in this campaign have at least high-school level education. Education level of about 4% is unknown.

Distinct No
8

Preliminary Visualization: Categorical Variables

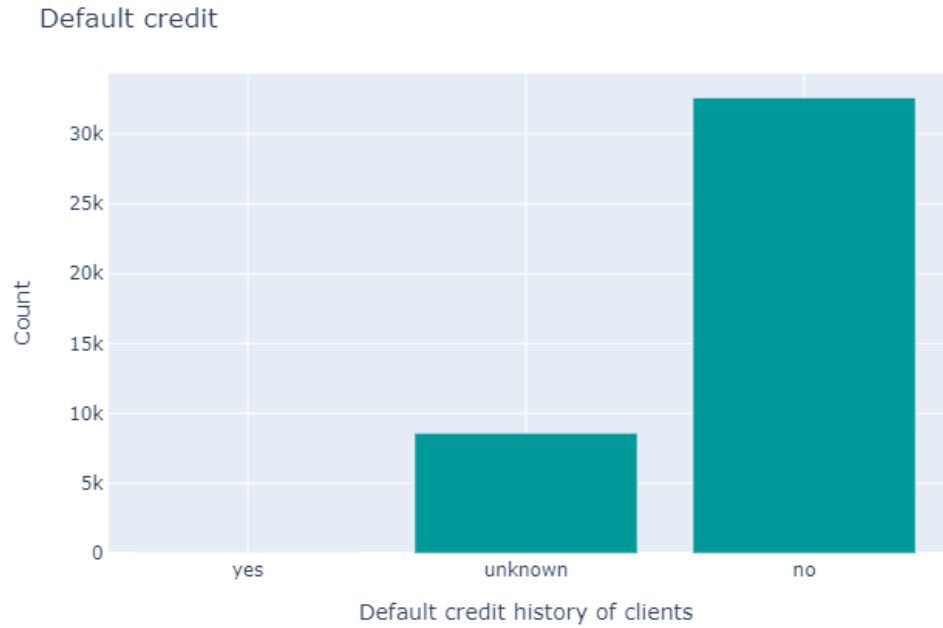
Marital



About 60% of the correspondents are married. Comparison of this data with the age group may have interesting outcomes.

Distinct No
4

Credit_default



Only 3 correspondents have default credits. The status of about 20% of the clients is unknown.

Distinct No
3

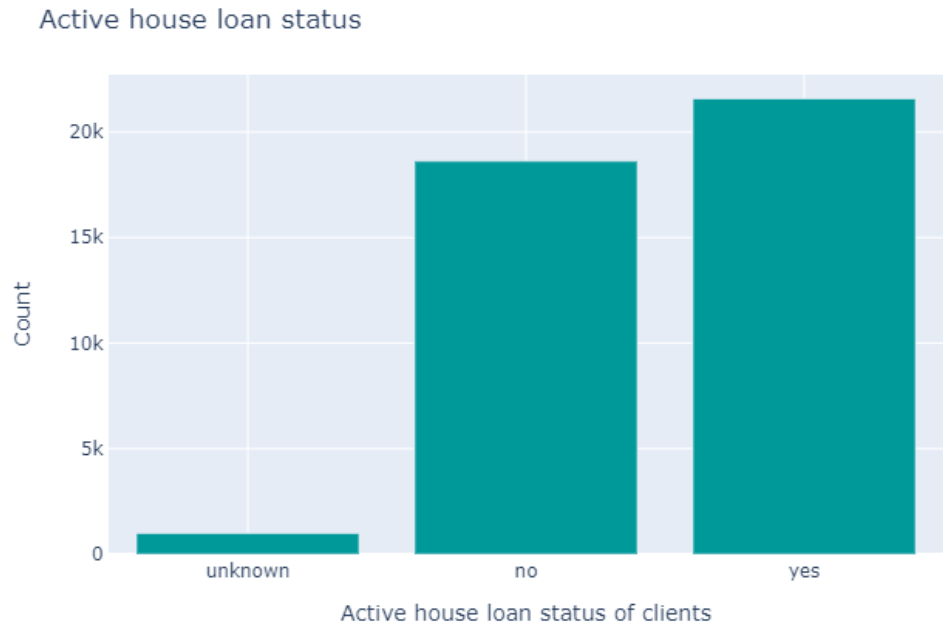
Preliminary Visualization: Categorical Variables

Personal_loan

More than 80% percent of the correspondents have no personal loans.

Distinct No
3

Housing_loan

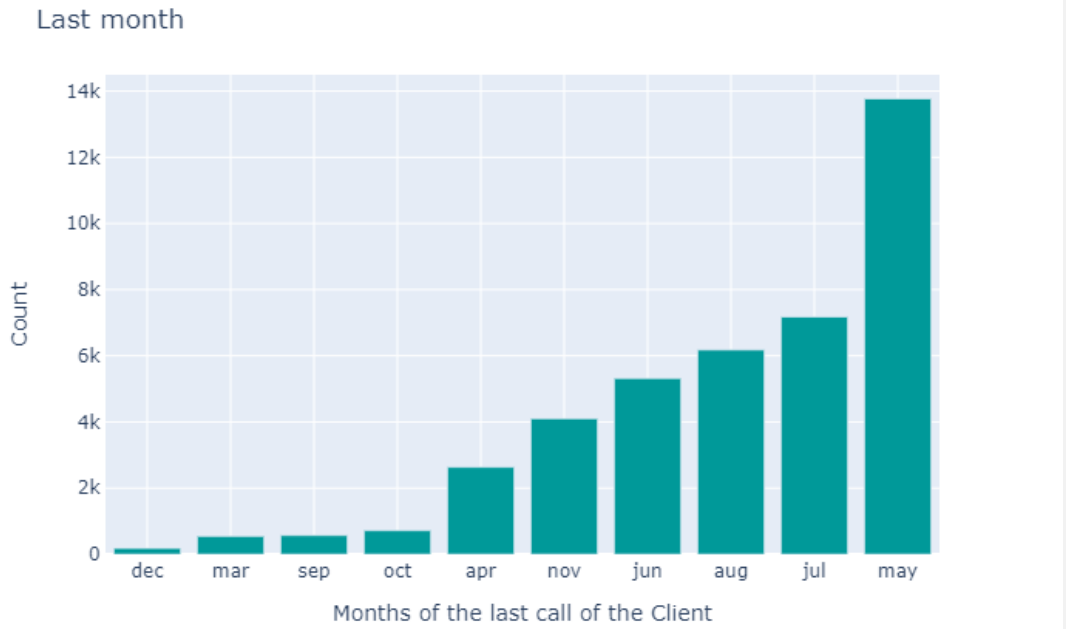


More than 50% of the correspondents have some kind of housing loans.

Distinct No
3

Preliminary Visualization: Categorical Variables

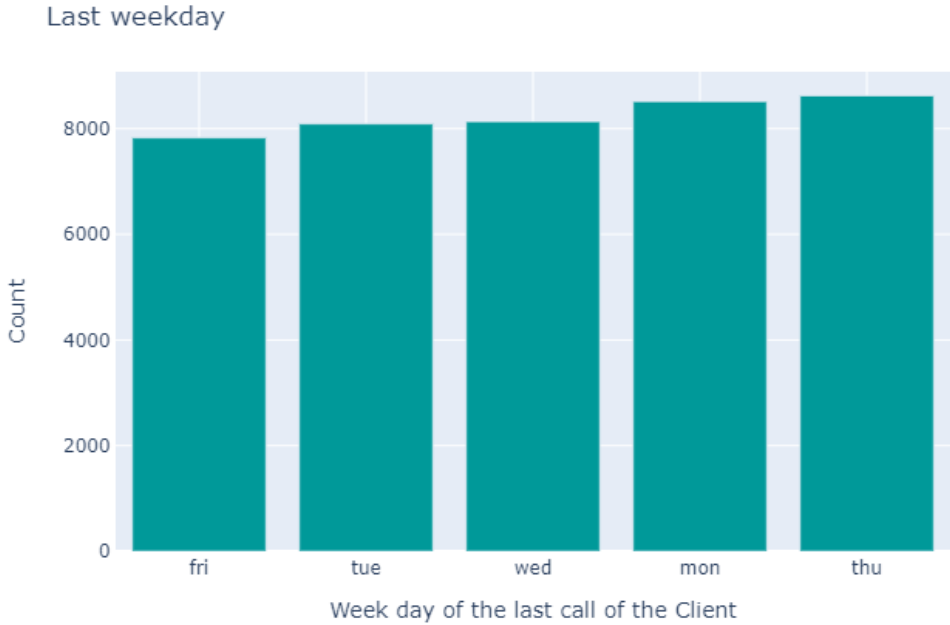
Last_month



No calls were made during January and February. Most of the calls were made during May (approx. 34%).

Distinct No
10

Last_weekday

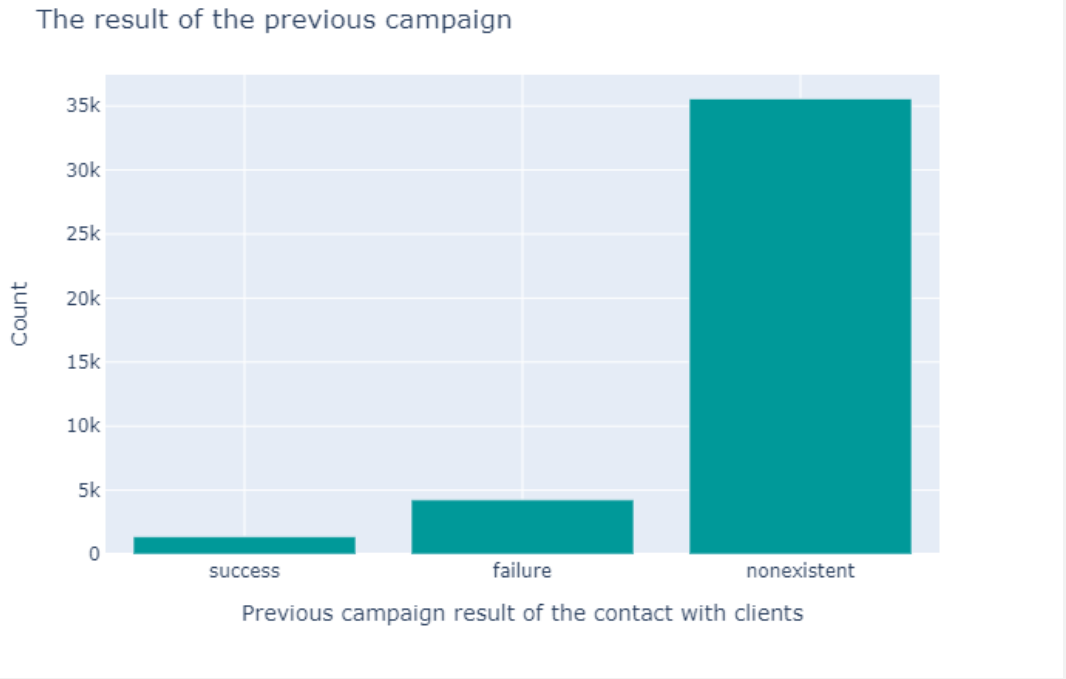


It seems that the calls were made in balance regarding the day of the week that calls were made.

Distinct No
5

Preliminary Visualization: Categorical Variables

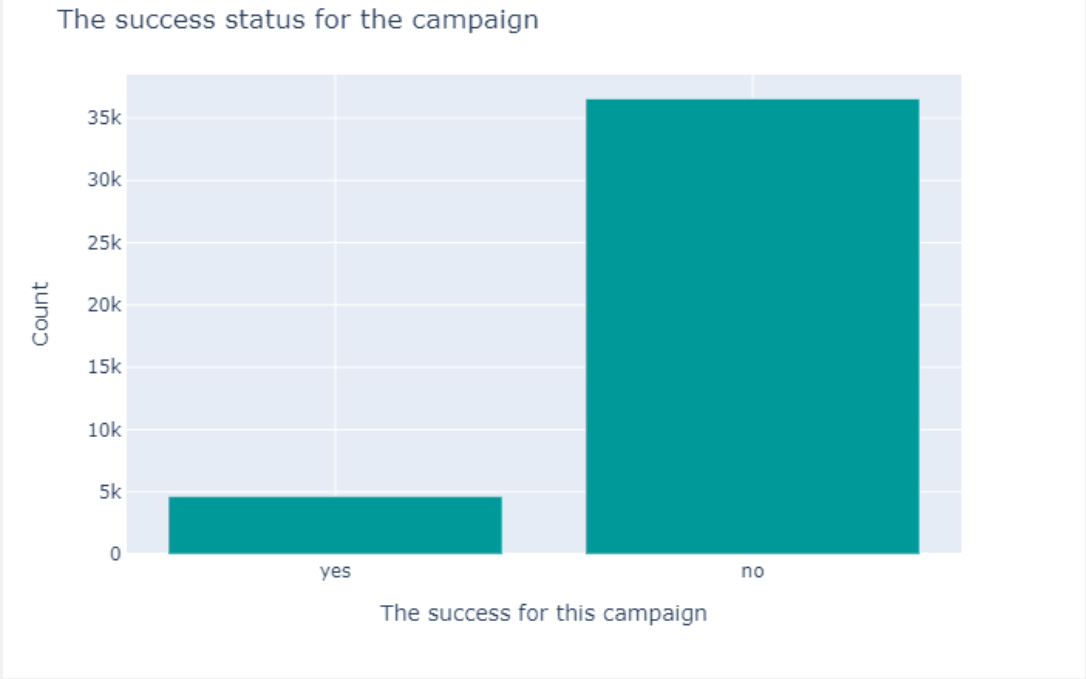
PrevCampaign_Result



It seems majority of the correspondents are selected from clients that have not been targeted during previous campaigns (87%).

Distinct No
3

Campaign_Success

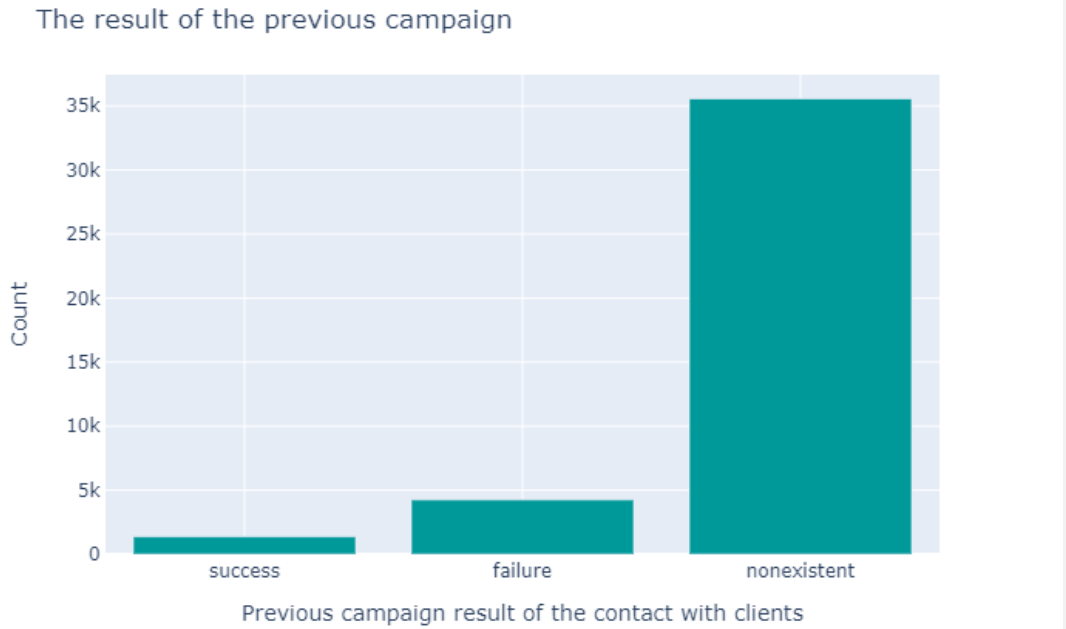


The rate of success (signing up to deposit into a special account) in this campaign was 11.2%

Distinct No
2

Preliminary Visualization: Categorical Variables

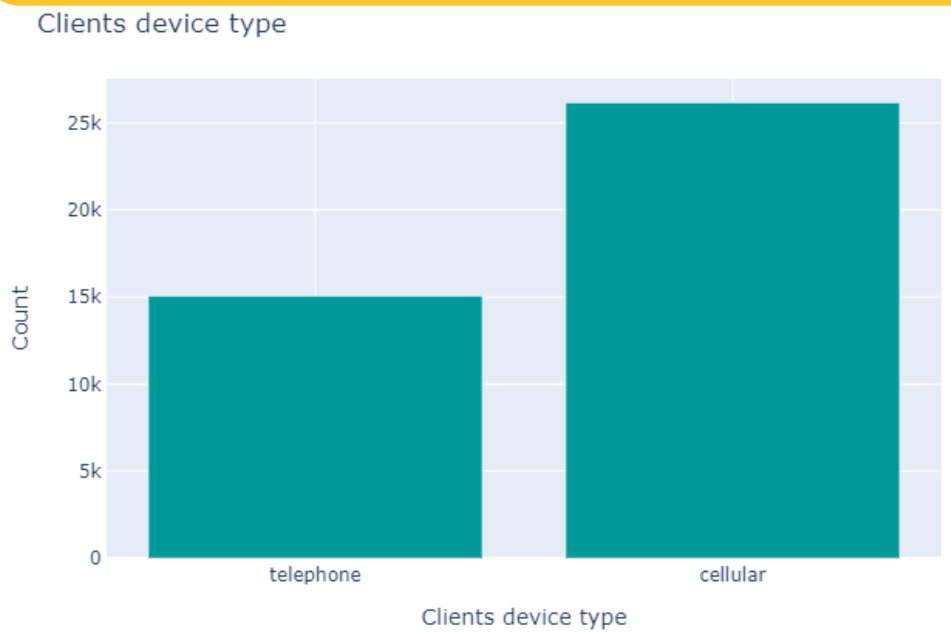
PrevCampaign_Result



It seems majority of the correspondents are selected from clients that have not been targeted during previous campaigns (87%).

Distinct No
3

Call_type

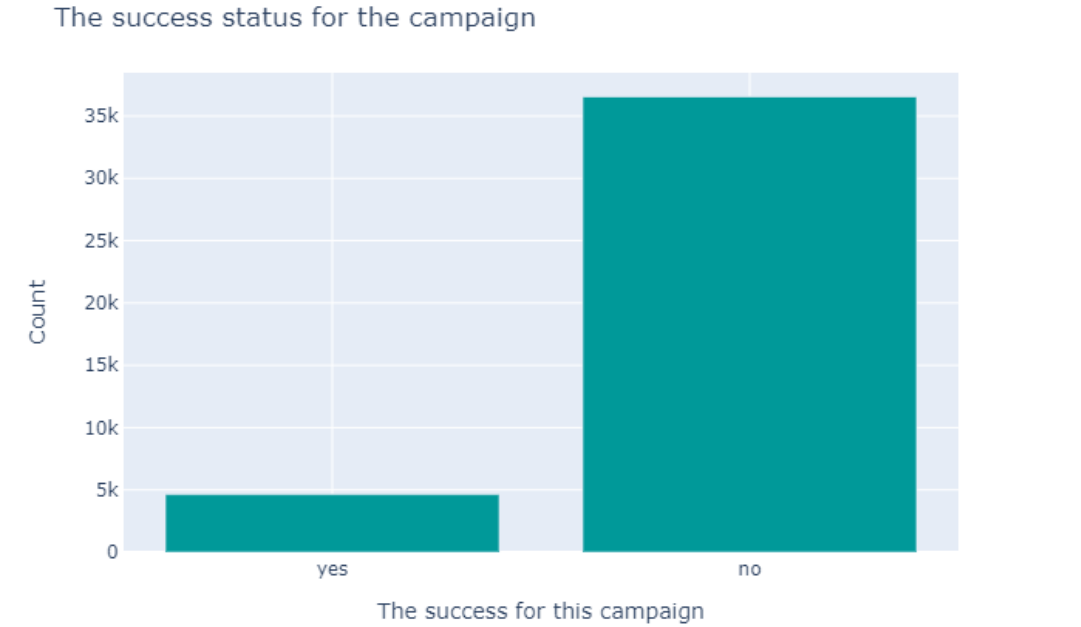


Most of the correspondents were contacted via their cellphones rather than their telephones (63.5% vs. 36.5%).

Distinct No
2

Preliminary Visualization: Categorical Variables

Campaign_Success



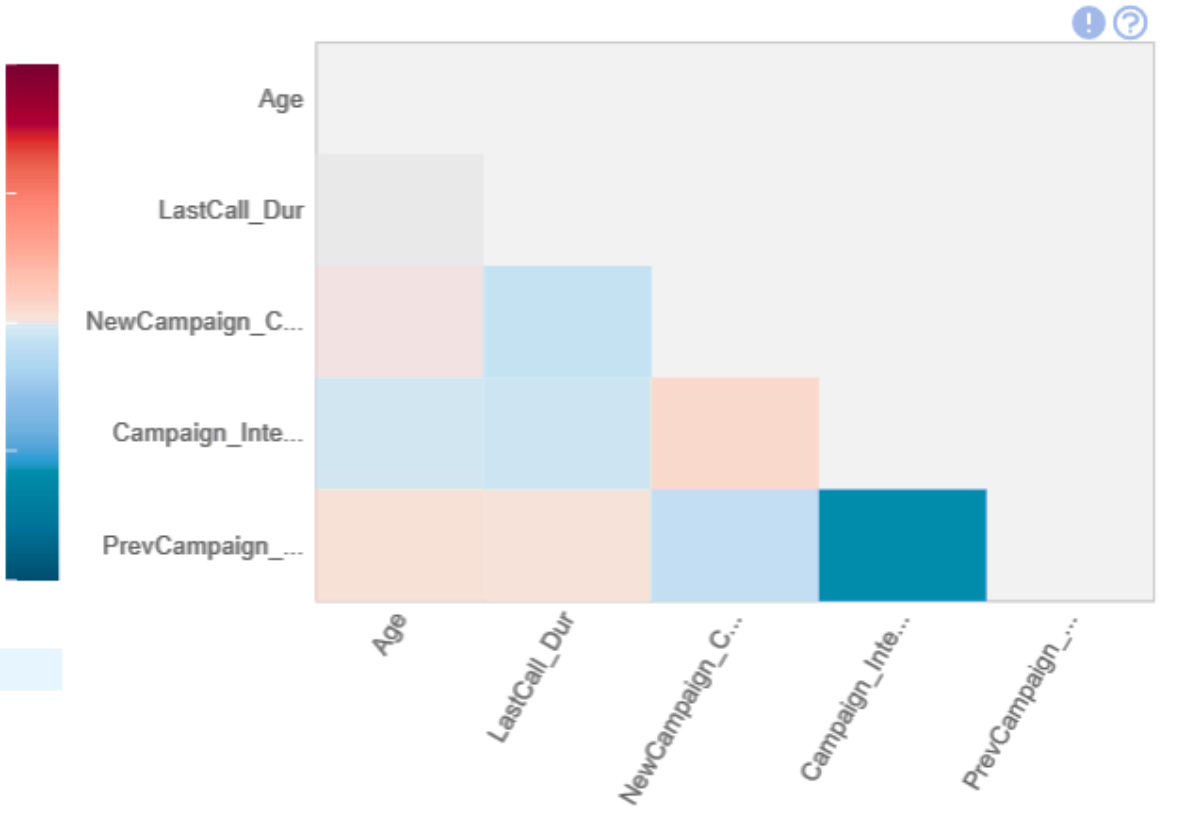
The rate of success (signing up to deposit into a special account) in this campaign was 11.2%

Distinct No
2

Correlations

- Pearson correlation matrix

Pearson correlation matrix (by dataprep.eda)



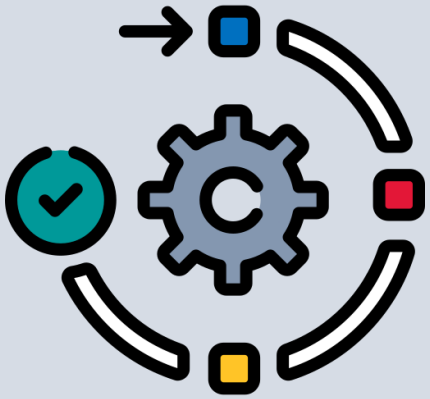
Cor. Method	Pearson	Spearman	KendallTau
Highest Positive Correlation	0.053	0.056	0.05
Highest Negative Correlation	-0.588	-0.51	-0.5
Lowest Correlation	0.001	0.001	0.001
Mean Correlation	-0.058	-0.054	-0.05

In a general view of the correlation matrix of a subset of the dataset (only the numerical variables are included in this subset), it seems only “PrevCampaign_CallNo” and “Campaign_Intervals_Day” have a considerable negative correlation. This correlation is somewhat expected due to the nature of these variables. More thorough analysis is needed to determine other possible correlations.

Data Cleaning & Transformation

The Data Transformation and Cleaning Steps

Data Transformation & Cleaning



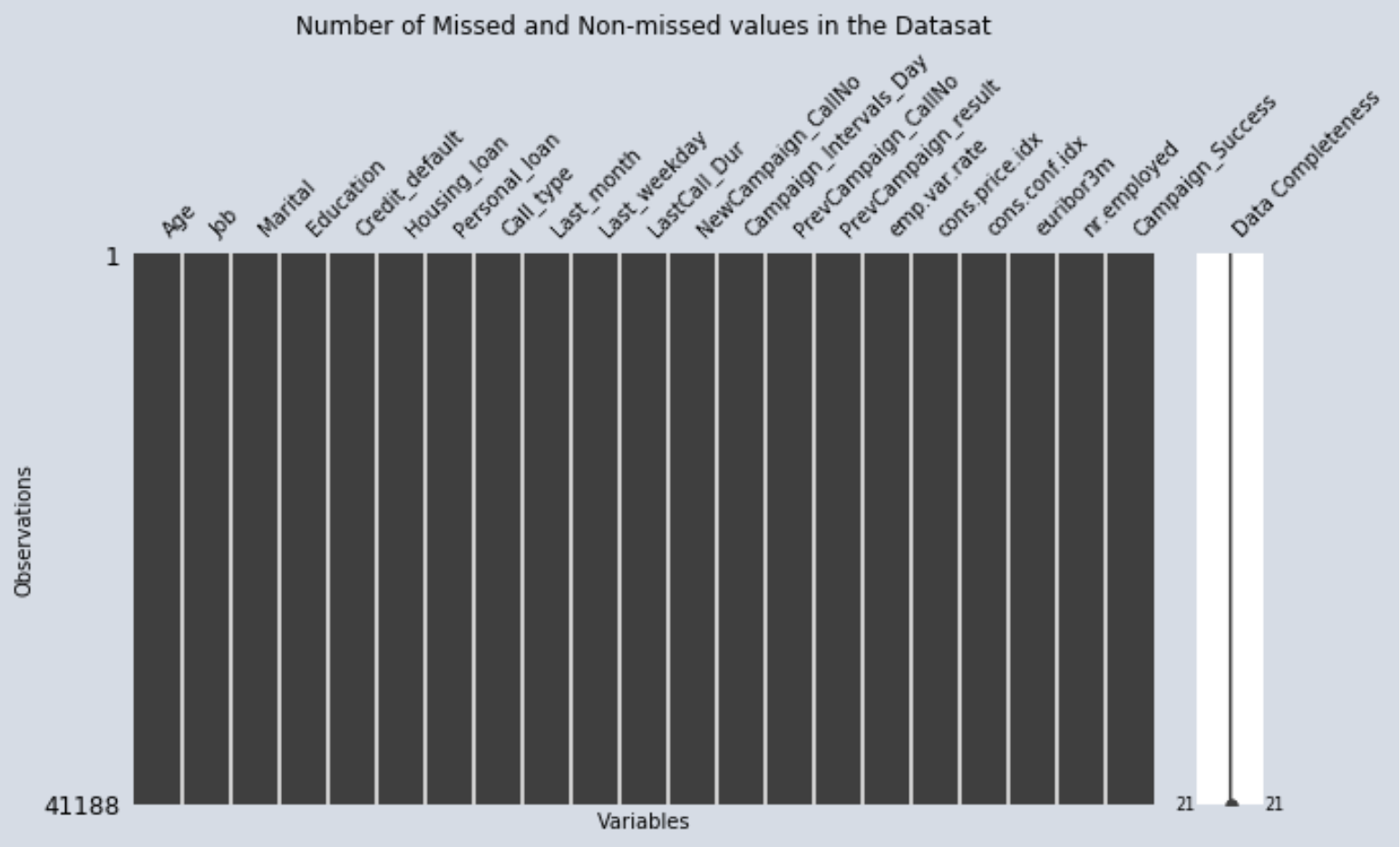
Steps

I. Changing variables' names
II. Identifying Missing values
III. Identifying Unfit Variables
IV. Finding and working with Outliers
V. Define new variable
VI. Creating a series of dataframe that group_by each categorical variables and the result variable which is "campaign_success"
VII. Calculate the result variable values in each values of each categorical variables.

1 Changing the variables' names

Previous_Name		New_Name
age	Numeric	Age
job	Categorical	Job
marital	Categorical	Marital
education	Categorical	Education
default	Categorical	Credit_default
housing	Categorical	Housing_loan
loan	Categorical	Personal_loan
contract	Categorical	Call_type
month	Categorical	Last_month
day_of_week	Categorical	Last_weekday
duration	Numeric	LastCall_Dur
campaign	Numeric	NewCampaign_CallNo
pdays	Numeric	Campaign_Intervals_Day
previous	Numeric	PrevCampaign_CallNo
poutcome	Categorical	PrevCampaign_Result
y	Categorical	Campaign_Success

2 Identifying Missing Values



As illustrated, there is no missing values in this dataset.

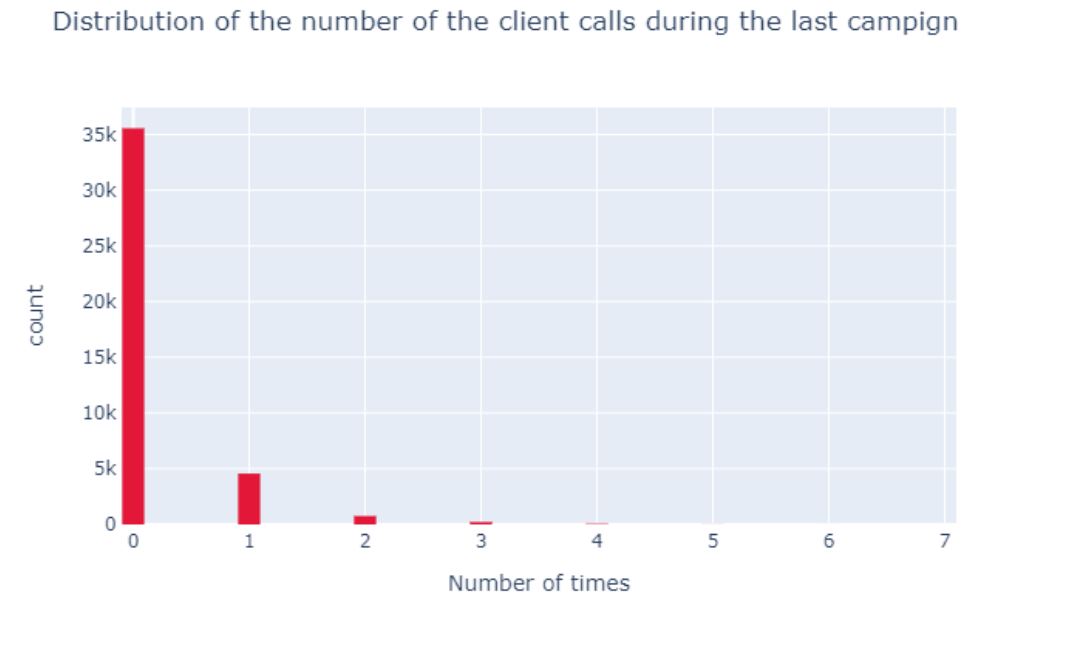
2 Identifying Missing Values

"Unknown" Values		
Job	330	% 0.8
Marital	80	% 0.19
Education	1731	% 4.2
Credit_default	8597	% 20.9
Housing_Loan	990	% 2.4
Personal_Loan	990	% 2.4

Missing Values	0

3 Identifying Unfit Variables

PrevCampaign_CallNo



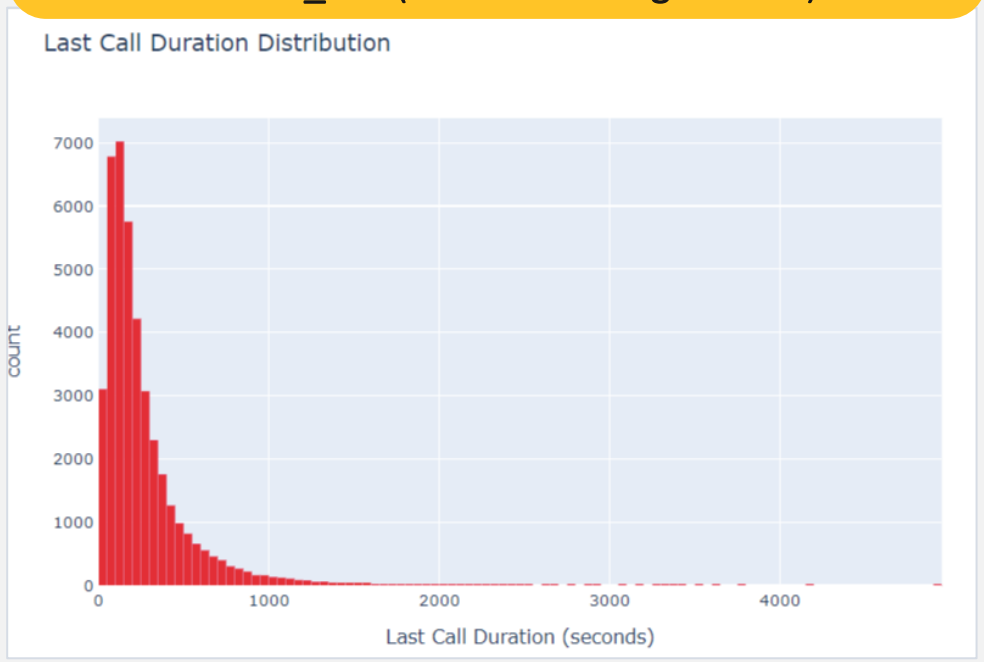
For "PrevCampaign_CallNo" because this variable had no use for the analysis, and we have a categorical variable which indicates whether a client has been contacted during previous campaign, we dropped it

4 Finding and working with Outliers

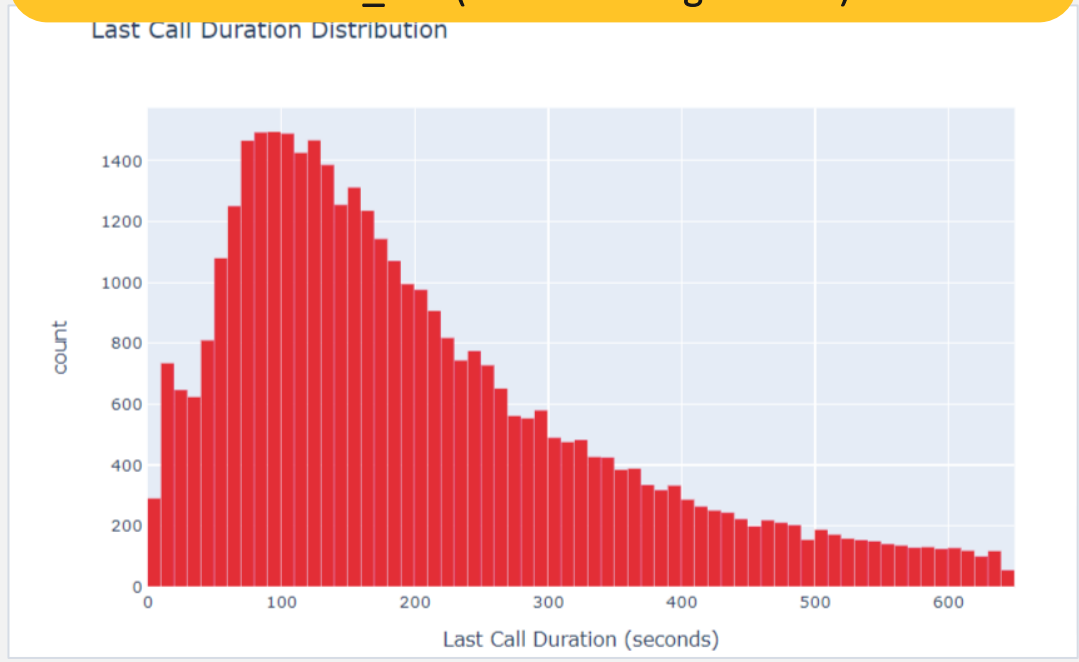
For "LastCall_Dur" variable we used IQR method to detect and omit outliers. We chose to delete outliers.

- It seemed that a very long duration of a call seemed to be unreasonable and unexplainable.

LastCall_Dur (Before omitting outliers)



LastCall_Dur (After omitting outliers)

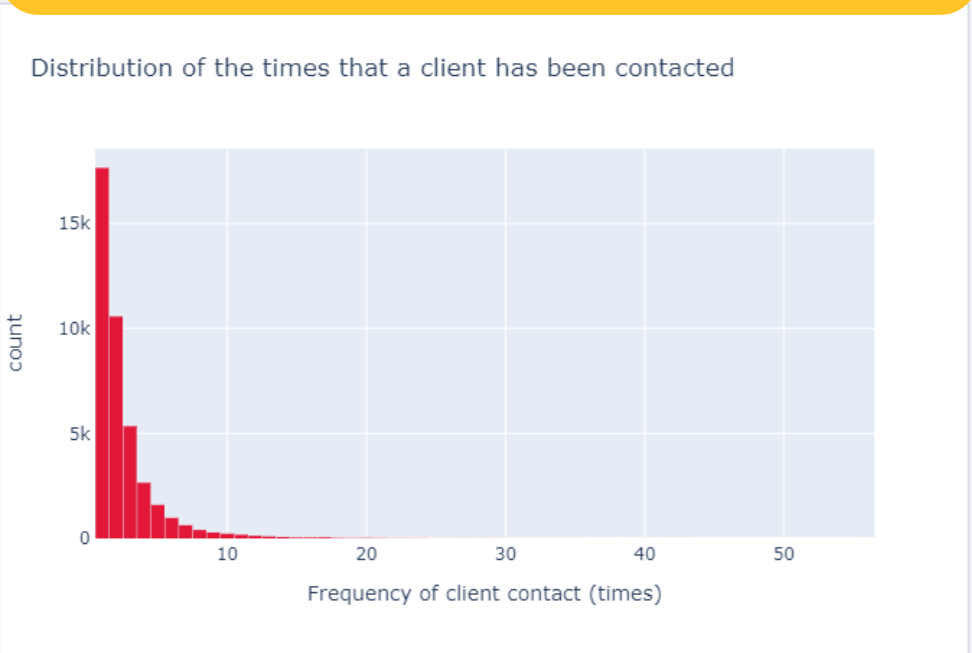


4 Finding and working with Outliers

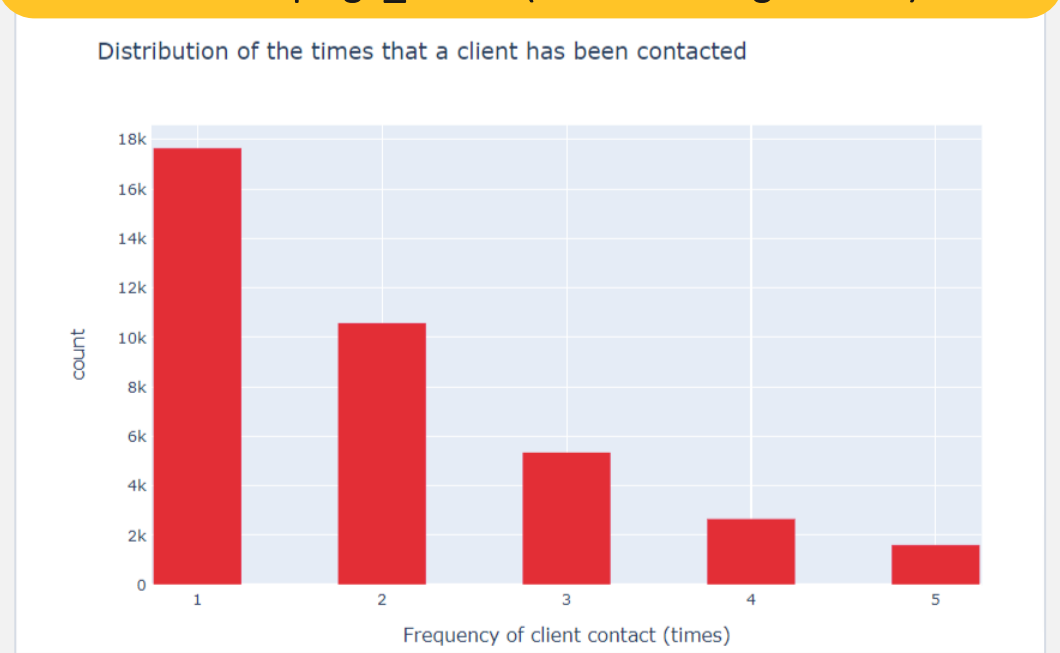
For “NewCampaign_CallNo” variable we also used IQR method to detect and omit outliers.

- For the number of calls made to a client it seemed to be some kind of mistake that some operators kept calling a particular client for over 6 times

NewCampaign_CallNo(Before omitting outliers)



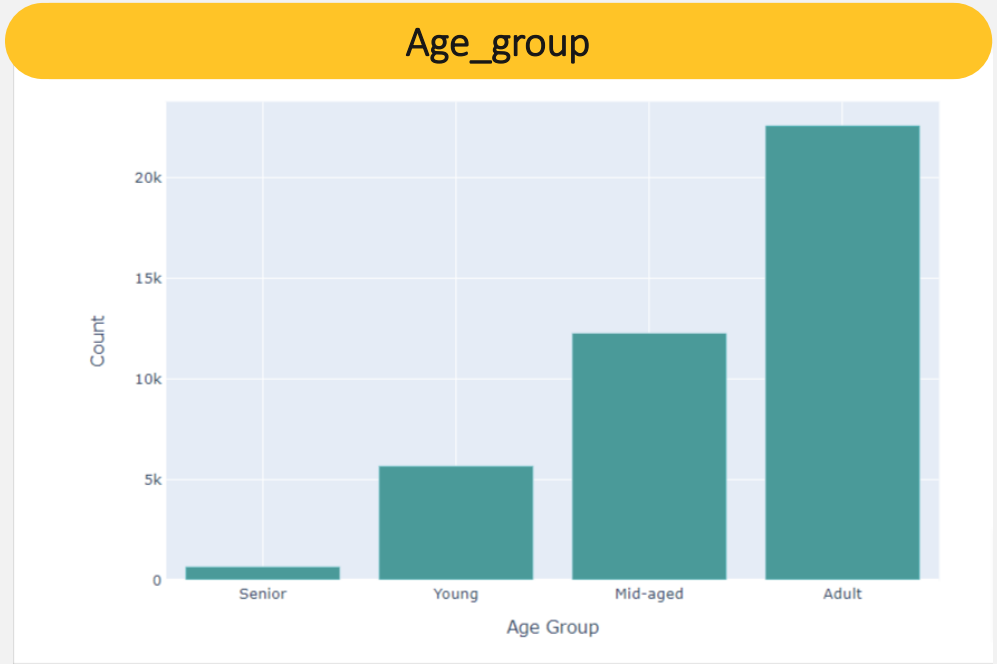
NewCampaign_CallNo (After omitting outliers)



5 Define New Variable

- We define new variable “Age_Group” based on “Age” variable. It is a categorical variable.

Age_Group	
Age Range	Value
[15 , 30)	Young
[30 , 45)	Adult
[45 ,65)	Mid_aged
[65, 100)	Senior



Data Analysis

Analysis Questions: Review

Question

1. Who are the top targeted prospects in terms of successful outcome?
2. What is the best time to contact prospect customers?
3. What are the top characteristics of a successful call?

Analysis Questions: Review

Question

1

Who are the top targeted prospects in terms of successful outcome?

Answer to Question #1

Q1 Who are the top targeted prospects in terms of successful outcome?

Methodology

1

To answer this question, we have calculated the percentage of our result variable (Campaign_Success) in each one of our categorical variables that can describe a customer segment

- Job
- Education
- Marital
- Age_Group
- Personal_loan
- Housing_loan
- PrevCampaign_Result
- Credit_default

2

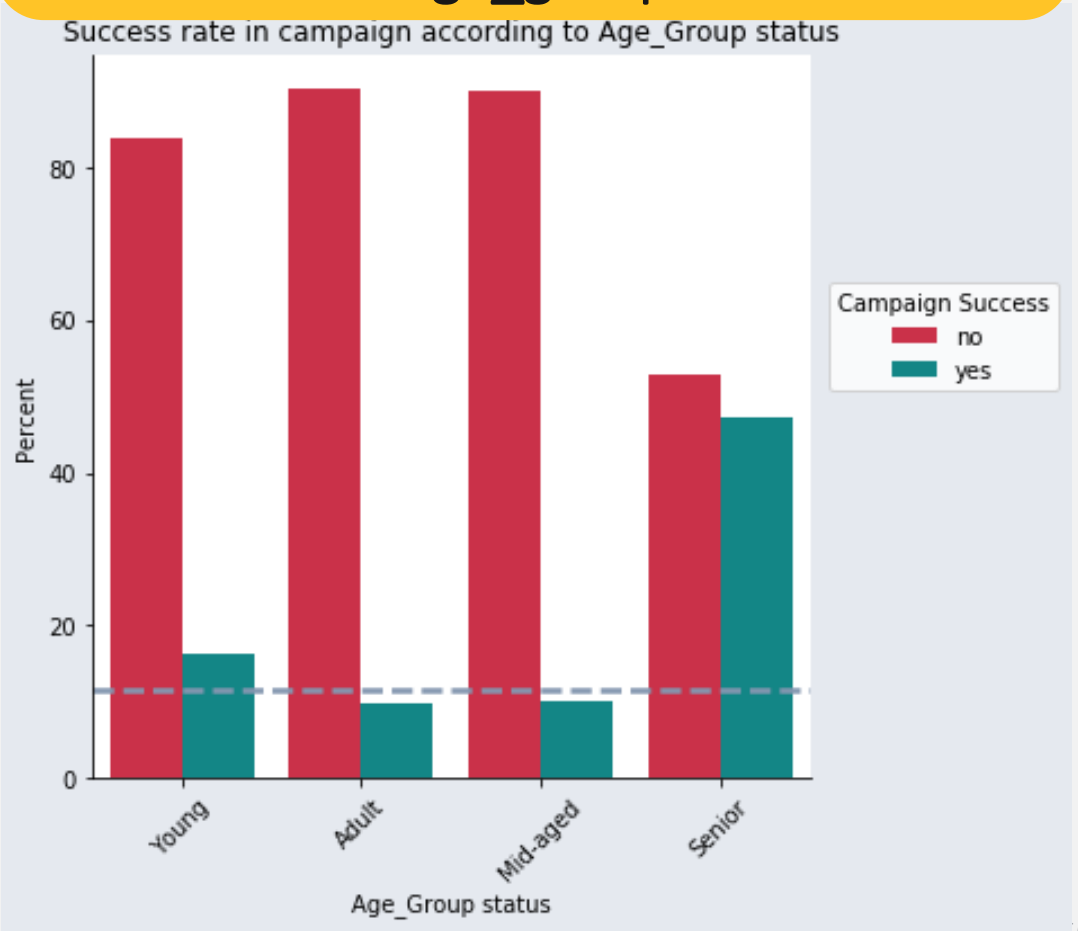
Based on this calculation, when in a segment the success rate is above overall success rate (11.625%), we mark that segment of the variable as a desirable segment to target

Answer to Question #1

Q1 Who are the top targeted prospects in terms of successful outcome?

- The gray line shows the campaign success rate among all clients which is 11.625%
- As the plot shows, “Senior” and “Young” age groups have been more enrolling to the campaign calls as the percentage of “yes” in “Campaign_Success” variable for these groups exceeds the over all success rate.

Age_group

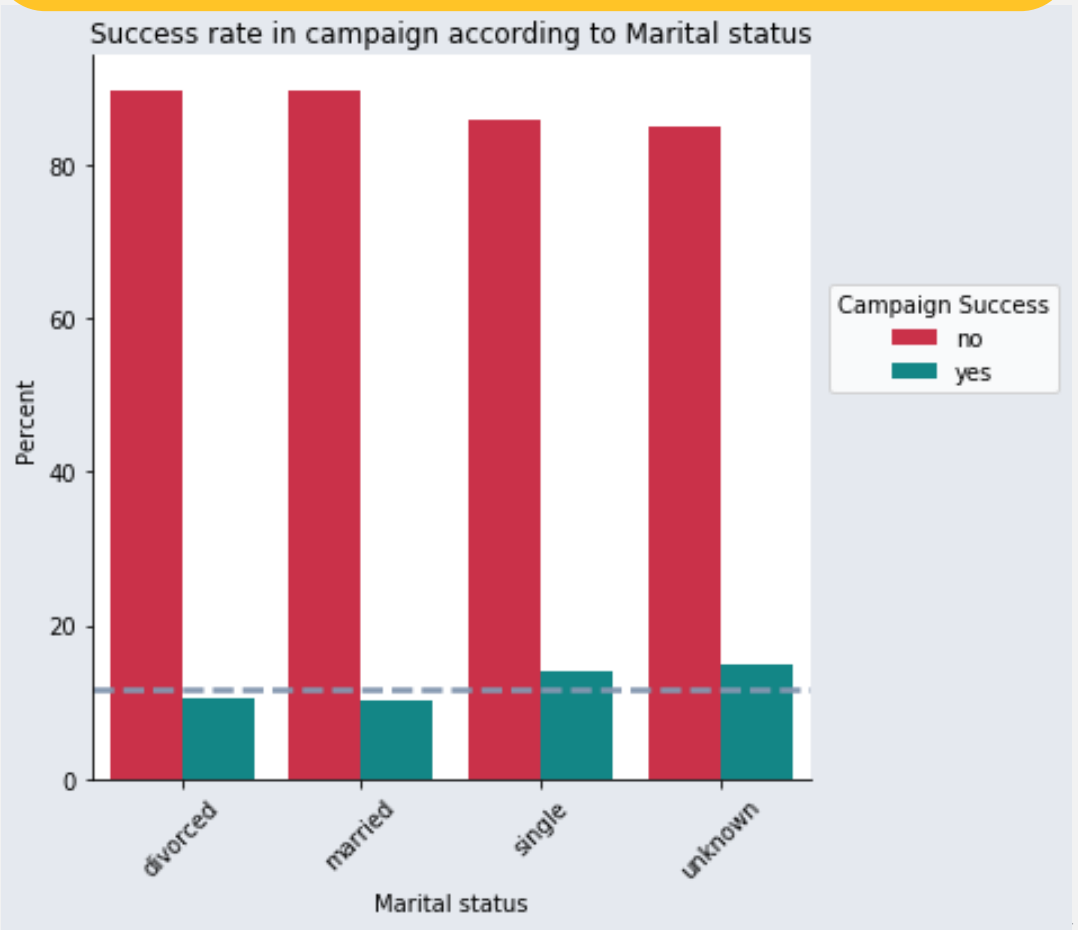


Answer to Question #1

Q1 Who are the top targeted prospects in terms of successful outcome?

- As the plot shows, “Unknown” and “Single” clients have been more enrolling to the campaign calls as the percentage of “yes” in “Campaign_Success” variable for these groups exceeds the over all success rate.
- As “unknown” can not use in target customer detection we can ignore it
- So, “single” clients were slightly more responsive to this campaign.

Marital

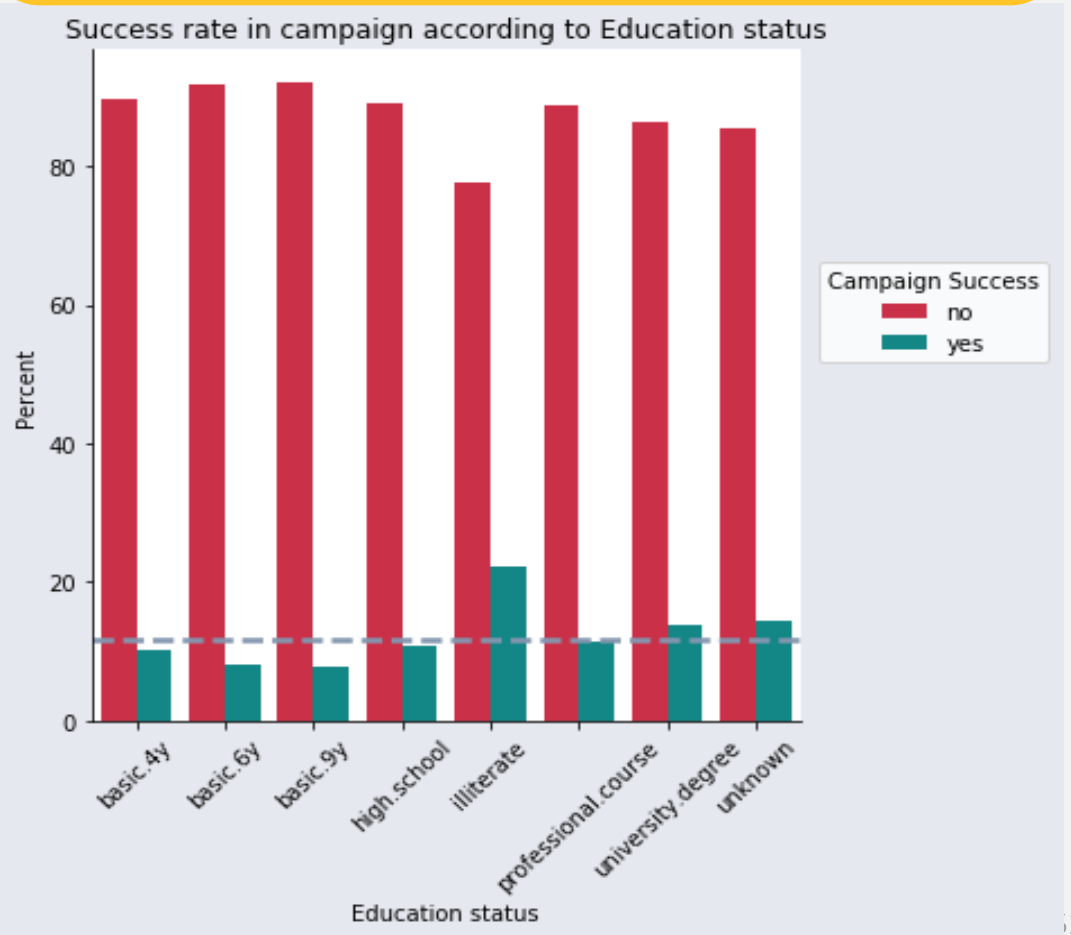


Answer to Question #1

Q1 Who are the top targeted prospects in terms of successful outcome?

- As the plot shows, clients with “unknown” ,“illiterate” and “university degree” level of education, were better targets for this campaign
- Also “unknown” can not be used in target customer detection we can ignore it

Education

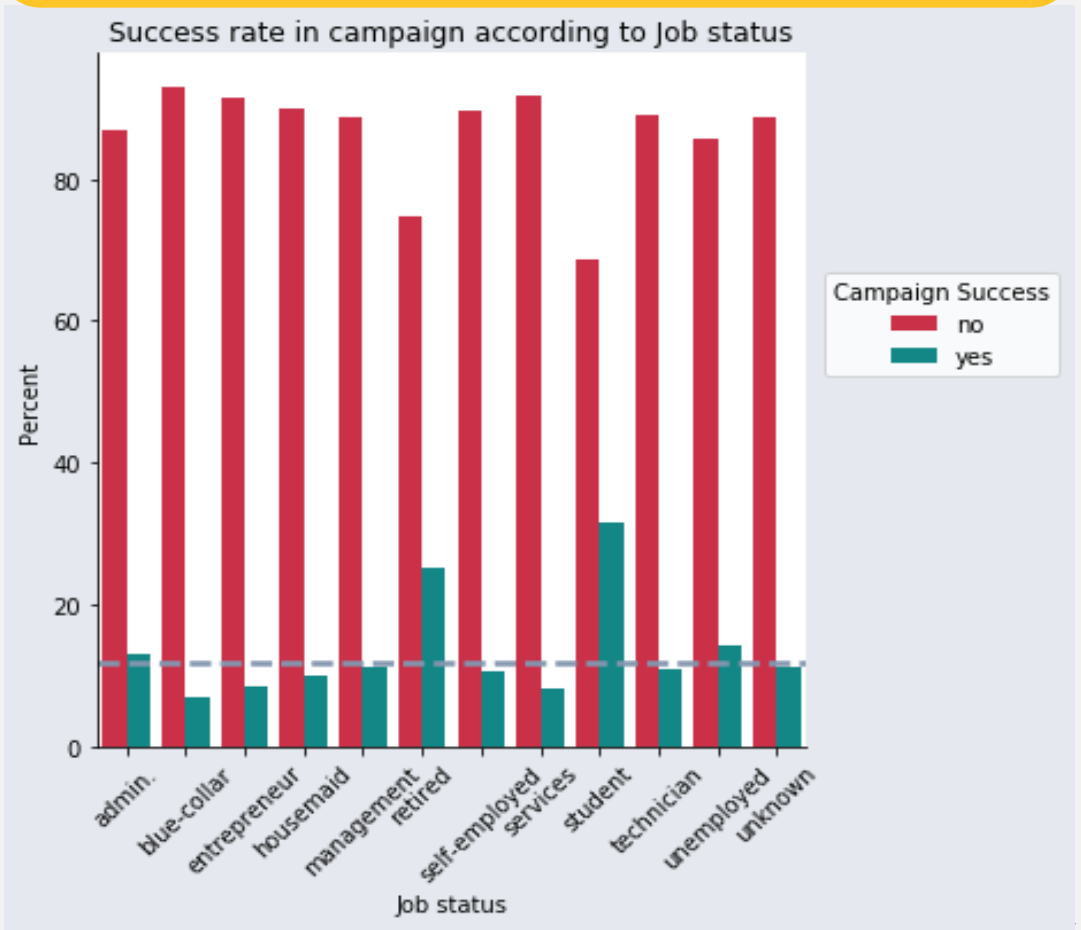


Answer to Question #1

Q1 Who are the top targeted prospects in terms of successful outcome?

- As the plot shows, clients with “administrative” ,“retired”, “student” and “unemployed” job categories were better targets for this campaign

Job

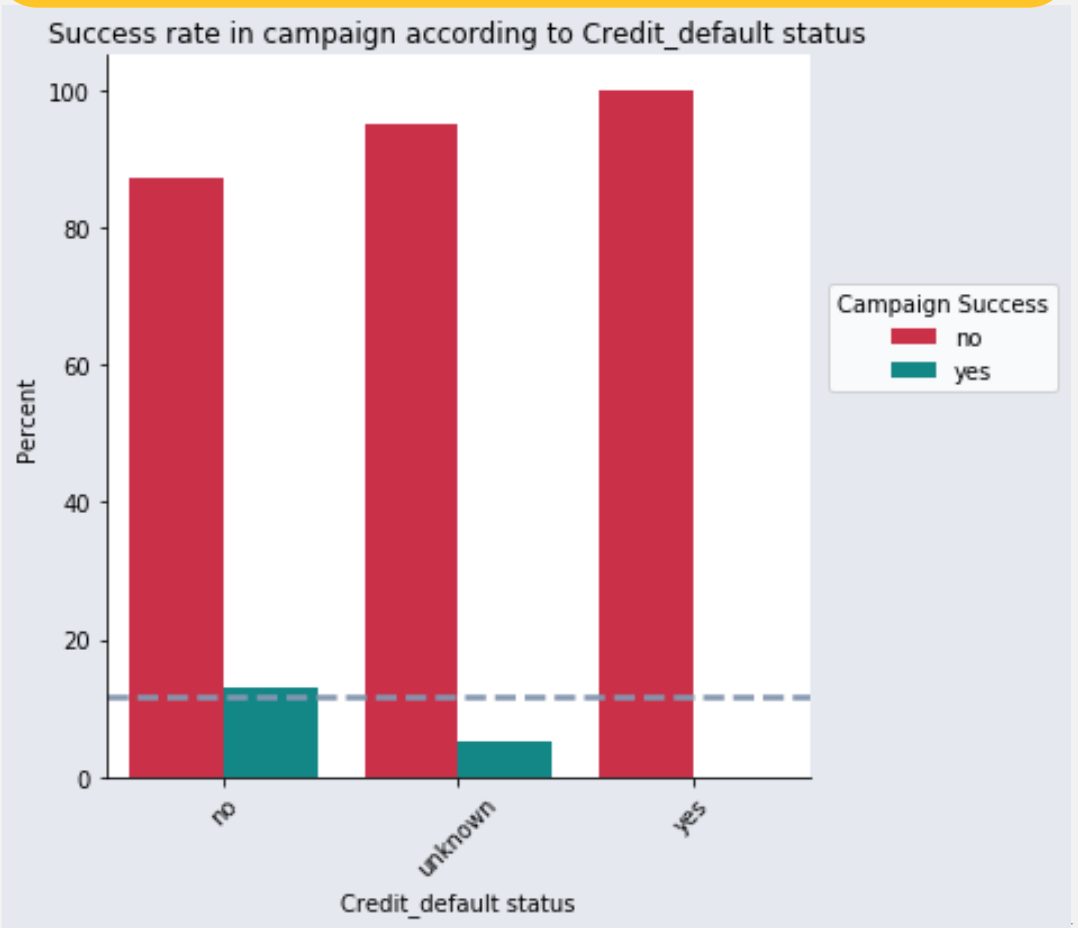


Answer to Question #1

Q1 Who are the top targeted prospects in terms of successful outcome?

- As the plot shows, clients with "no" credit defaults were better targets in this campaign

Credit_default

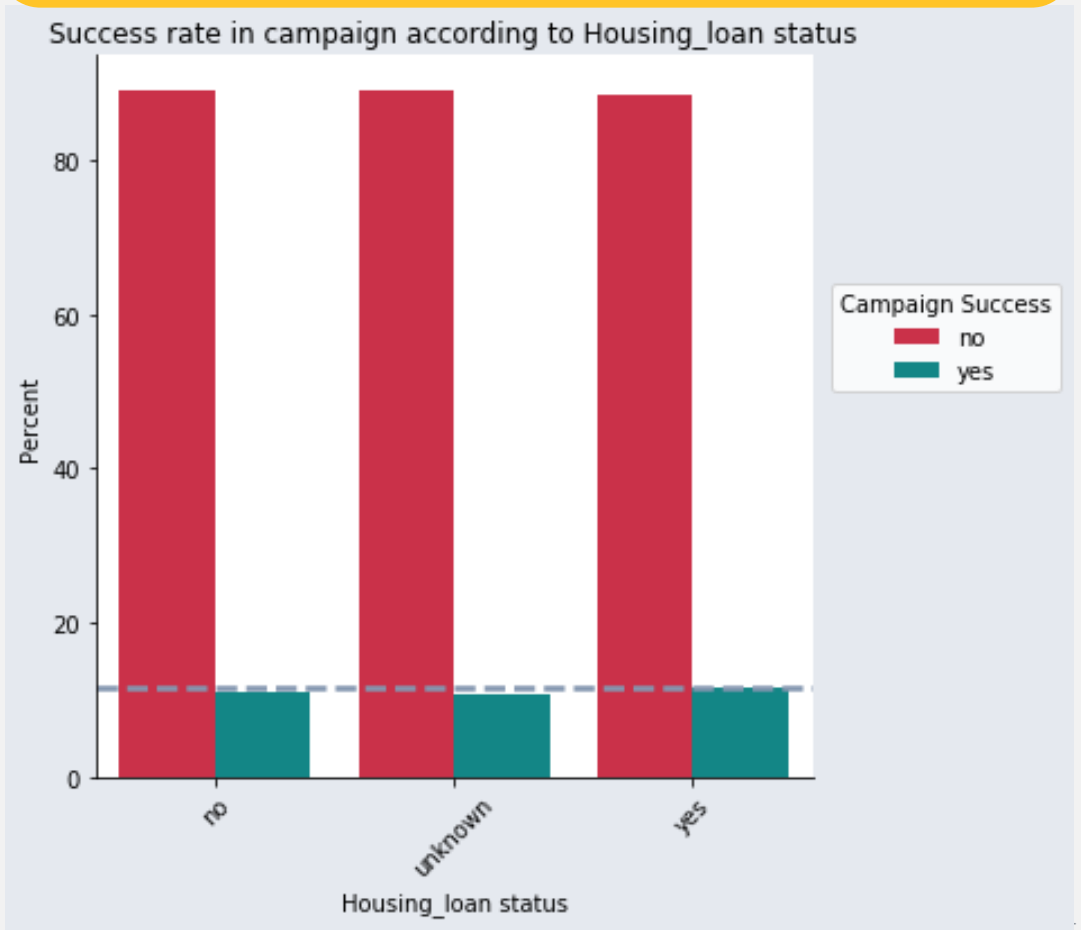


Answer to Question #1

Q1 Who are the top targeted prospects in terms of successful outcome?

- As the plot shows there is no significant difference among these groups in terms of campaign success

Housing_loan

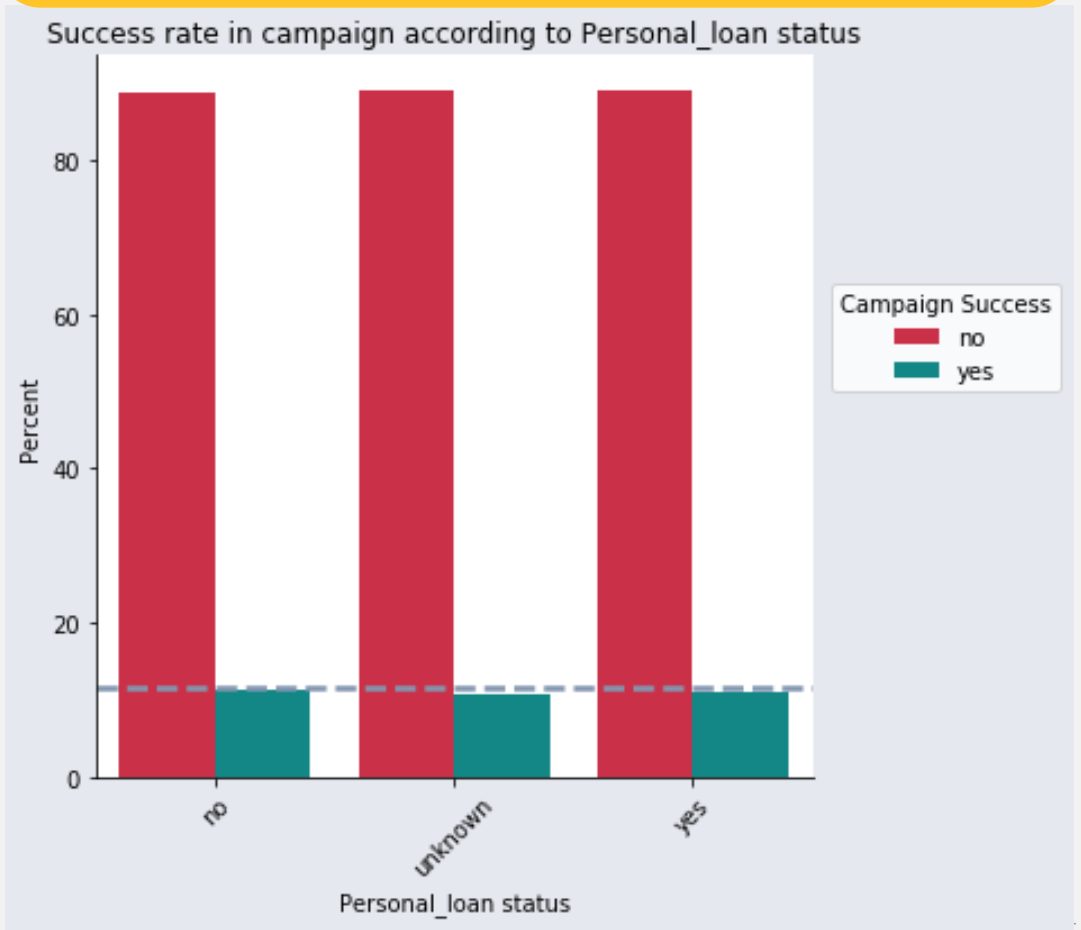


Answer to Question #1

Q1 Who are the top targeted prospects in terms of successful outcome?

- As the plot shows there is no significant difference among these groups in terms of campaign success

Personal_loan

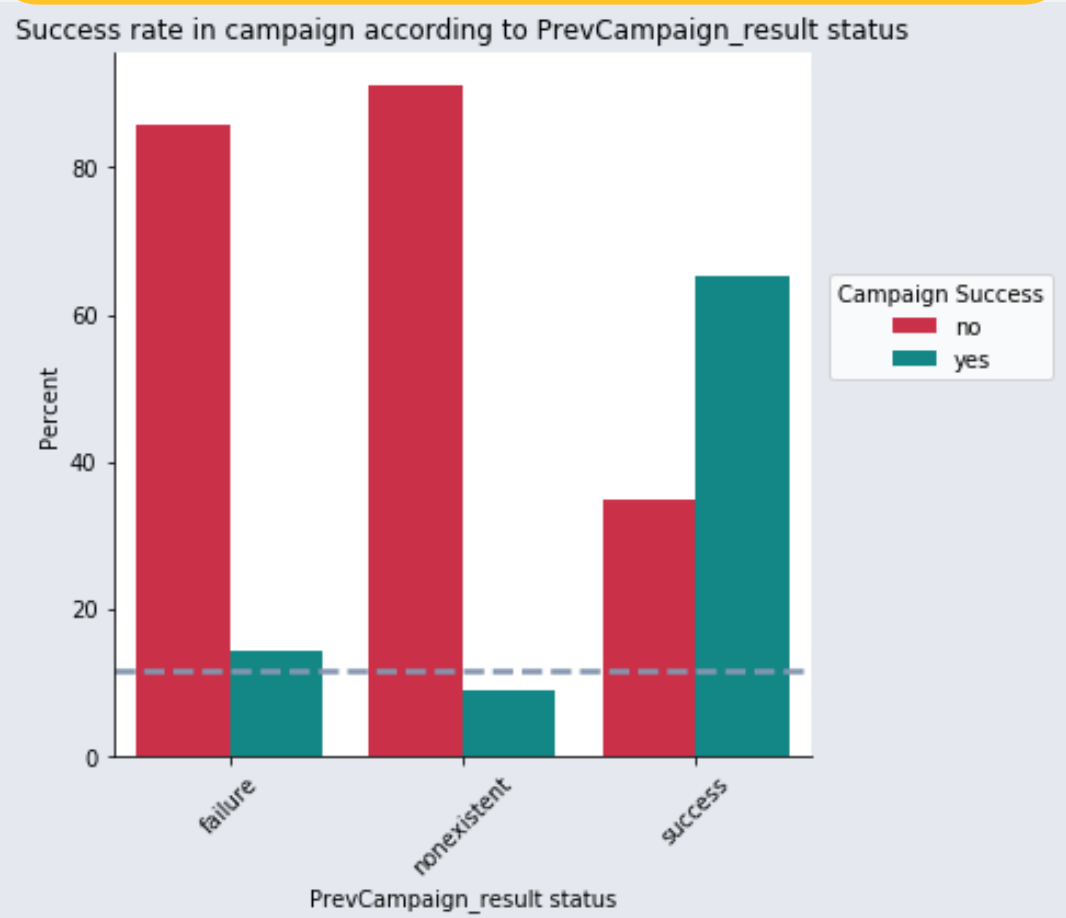


Answer to Question #1

Q1 Who are the top targeted prospects in terms of successful outcome?

- As the plot shows clients that have been contacted during previous campaigns enrolled more in this campaign, especially clients that have enrolled in previous campaign successfully.
- So, we can consider “failure” and ”success” values for our purpose

PrevCampaign_result



Answer to Question #1

Q1 Who are the top targeted prospects in terms of successful outcome?

Demographics

Age Group	Marital Status	Education Level	Job
Senior	Single	Illiterate	Retired
Young		University Degree	Unemployed
			Student
			Administrative

Behavioral

Prev Campaign Result	Credit Default
Failure	No
Success	

The top targeted clients' segment demographic and behavioral characteristics

Analysis Questions: Review

Question

2

What is the best time to contact prospect customers?

Answer to Question #2

Q2 What is the best time to contact prospect customers?

Methodology

1

To answer this question, we have calculated the percentage of the result variable (Campaign_Success) in each one of the variables that can describe time to contact a client

Last_month

Last_weekday

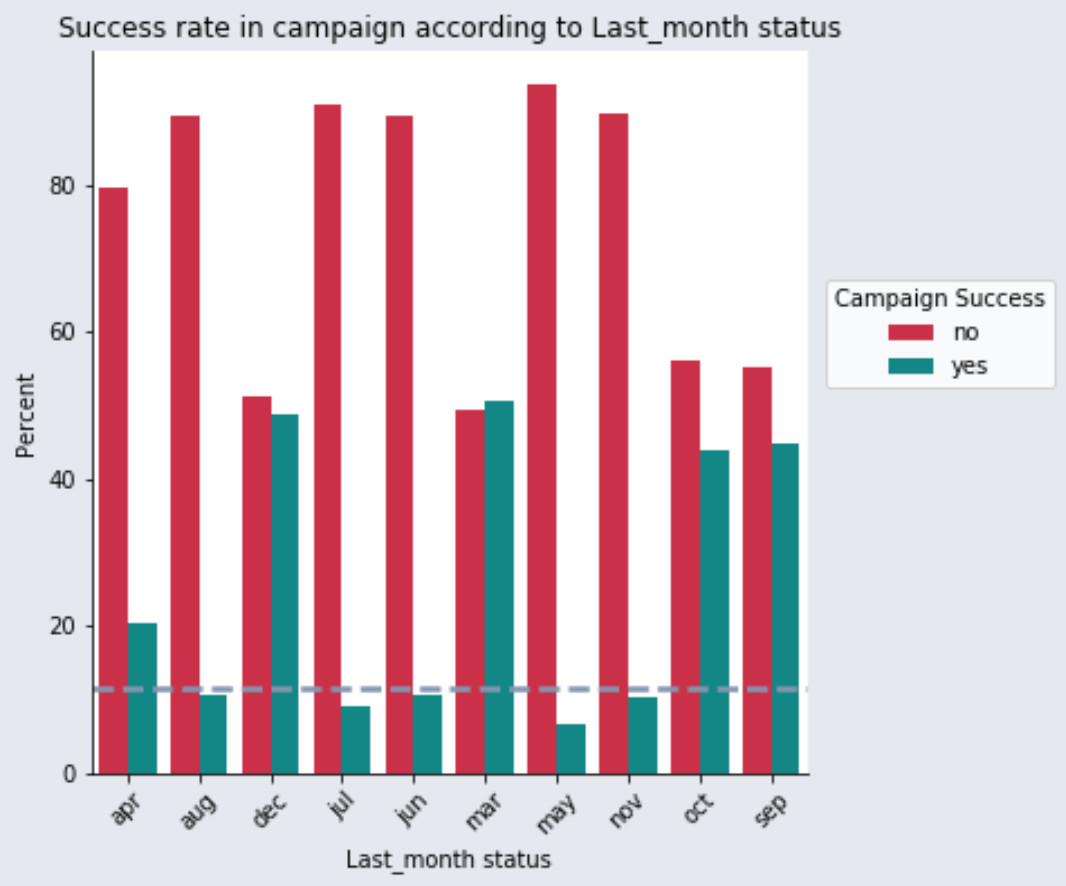
2

Based on this calculation, when in a time slot the success rate is above overall success rate (11.625%), we mark that time slot of the variable as a desirable time slot

Answer to Question #2

Q2 What is the best time to contact prospect customers?

Last_month

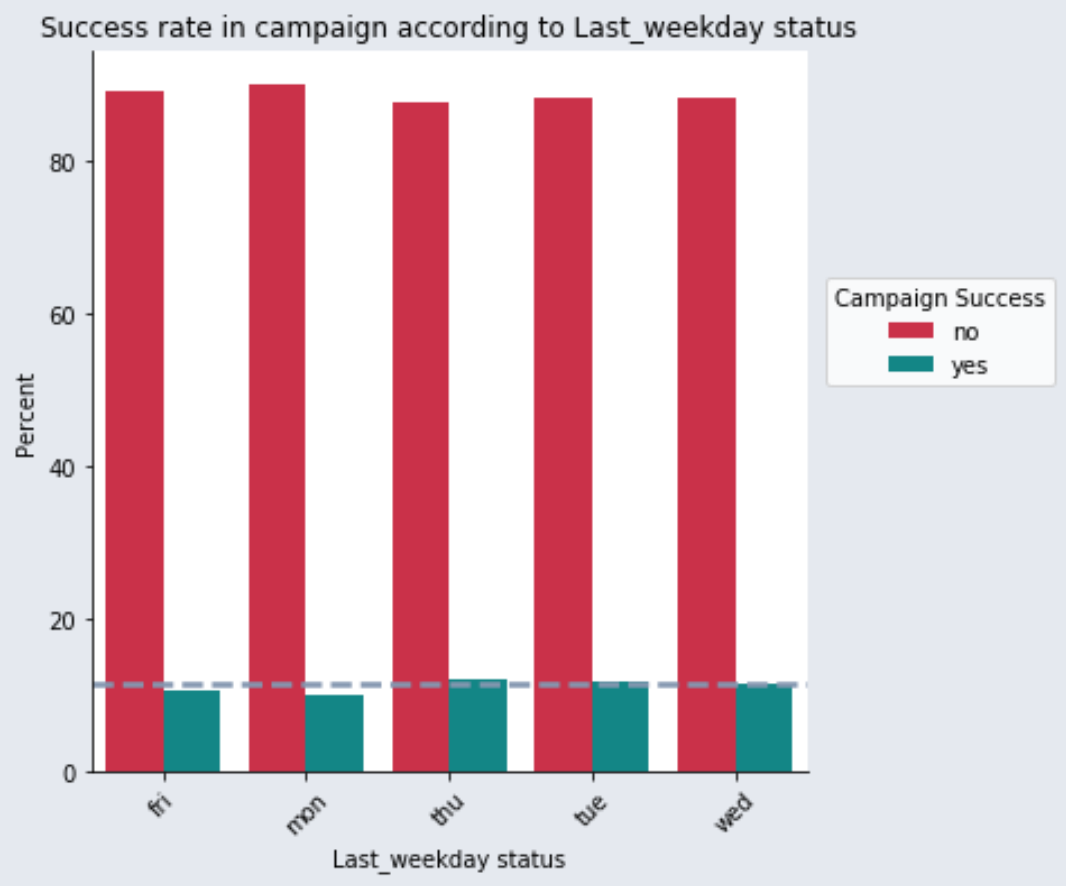


- According to the plot, “March”, “April”, “September”, “October” and “December” are the months that more successful calls are made in.

Answer to Question #2

Q2 What is the best time to contact prospect customers?

Last_weekday



- According to the plot, there is not any significant variations in campaign success rate among different weekdays.

Answer to Question #2

Q2 What is the best time to contact prospect customers?

Month

- March
- April
- September
- October
- December

Weekday

- No difference

The best time to contact clients to have successful outcome

Analysis Questions: Review

Question

3

WHAT are the top characteristics of a successful call?

Answer to Question #3

Q3 What are the top characteristics of a successful call?

Methodology

1

We have calculated the percentage of the result variable (Campaign_Success) in the categorical variable that can describe characteristics of a successful call which includes:
Call_type (indicate if the call made to clients' cellphone or telephone)
segment

Call_type

2

As for numerical variables we tried to plot box-plot for these variables and compare the median of the distribution according to campaign success ("yes" and "no" in "Campaign_Success" variable)

LastCall_Dur

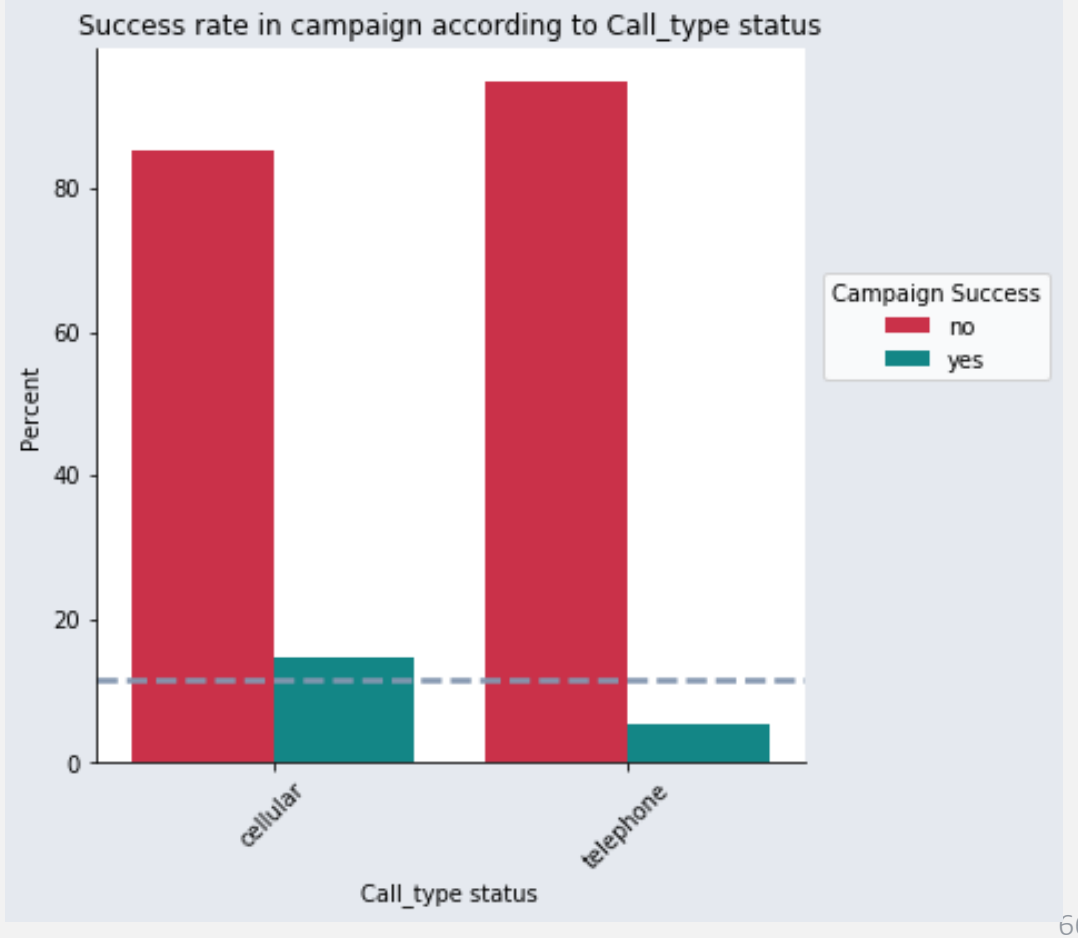
NewCampaign_CallNo

Answer to Question #3

Q3 What are the top characteristics of a successful call?

- According to the plot, clients who have been called through their "cellular" phones, performed slightly better in terms of enrolling in the campaign.

Call_type

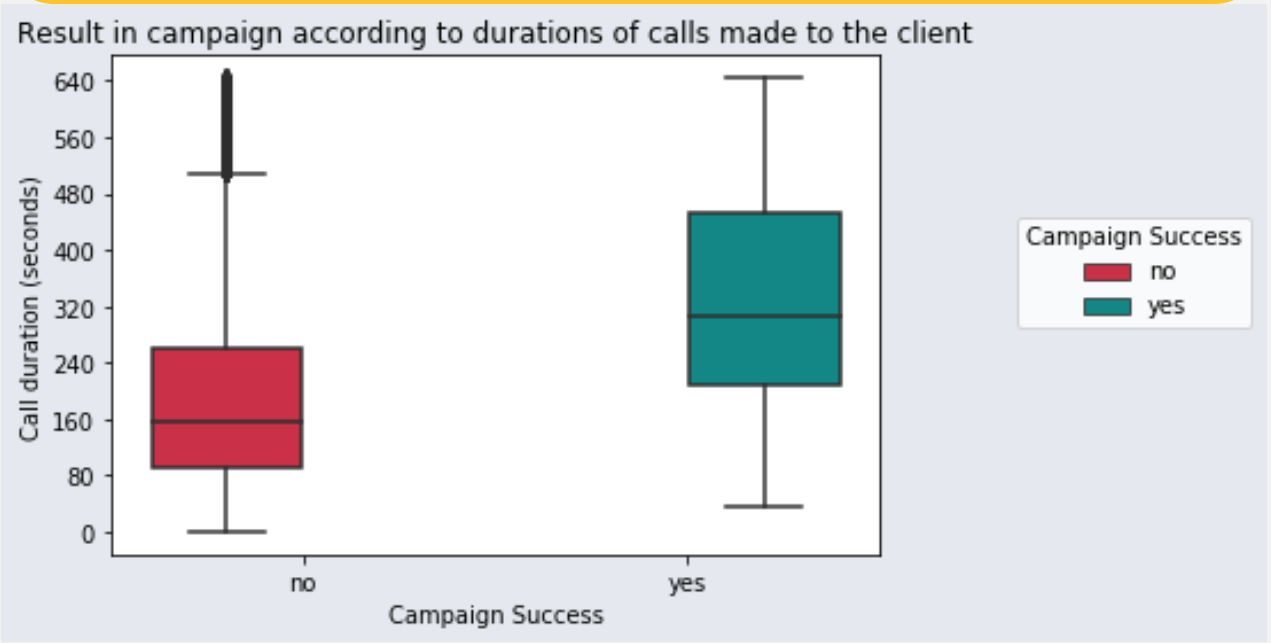


Answer to Question #3

Q3 What are the top characteristics of a successful call?

- As the plot shows, the median of duration of calls with the clients who submitted in the campaign is higher than the median of those who didn't.
- Although not always, but it seems that it is natural to a successful call to a client takes longer than an unsuccessful one. So, there is a need to look more into this matter but the data we have is not enough for more through analysis.

LastCall_Dur

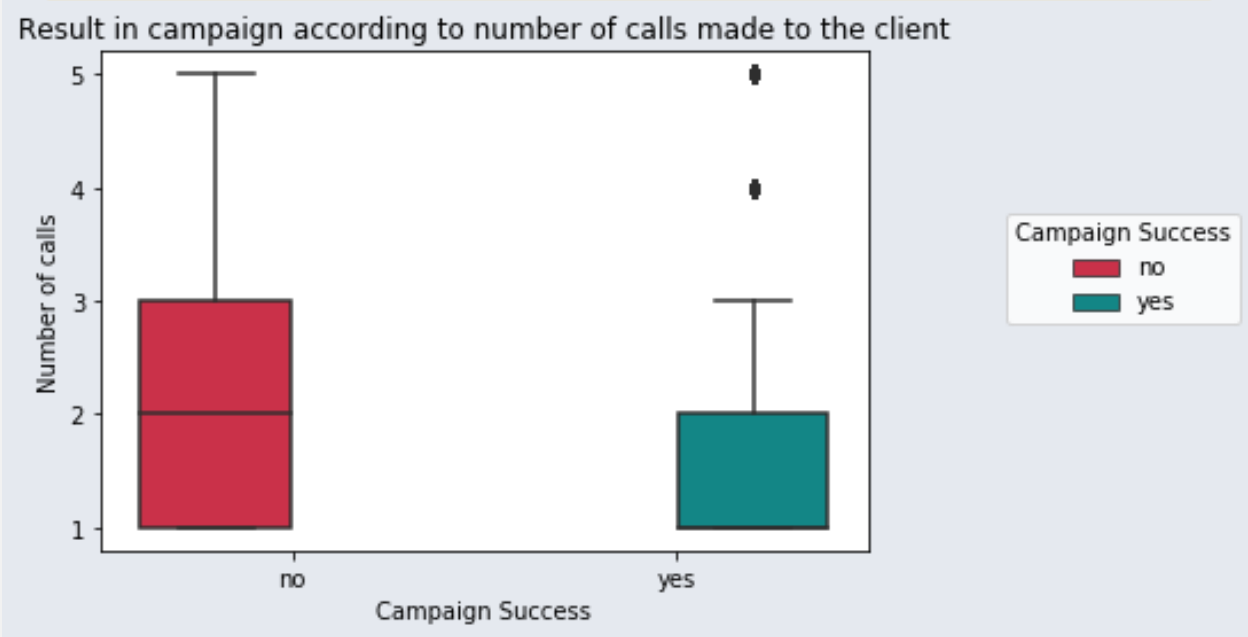


Answer to Question #3

Q3 What are the top characteristics of a successful call?

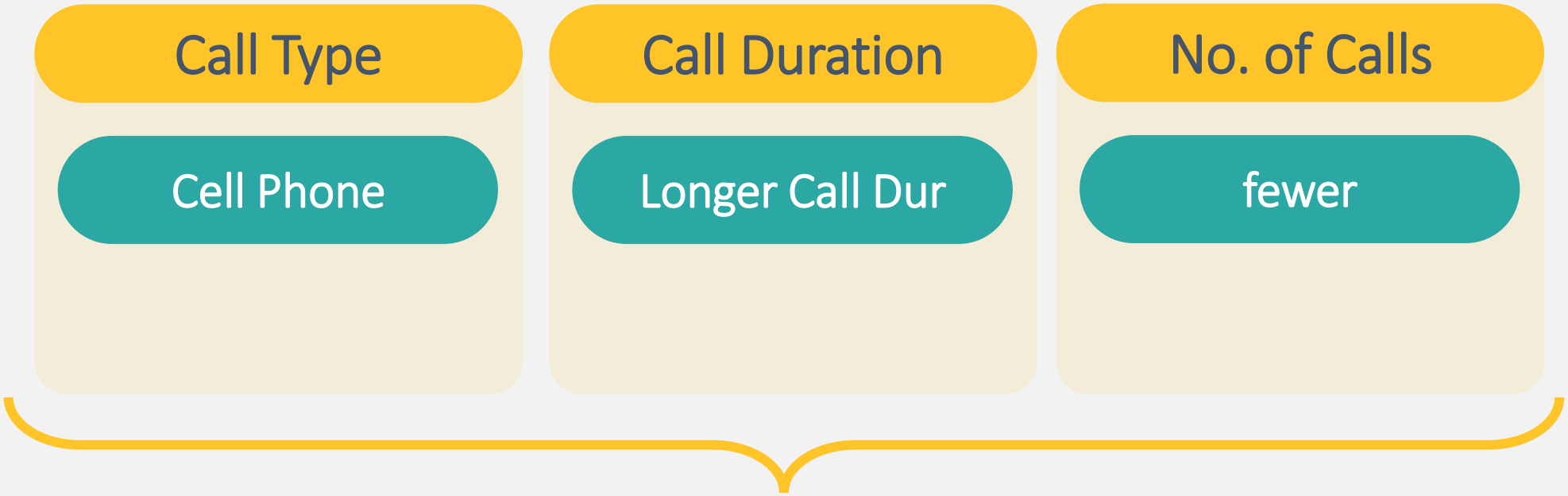
- The plot indicates that the median of the number of call to clients who didn't participate in the campaign is higher than the median of those who did.
- So, we may concur that the more times we call a client to enroll him in the campaign, less likely she will do so.

NewCampaign_CallNo



Answer to Question #3

Q3 What are the top characteristics of a successful call?



Conclusion

Conclusion

WHO?

- The demographic characteristics that can describe the participating clients in this campaign are:
  - Young, or senior
  - Single
  - Illiterate or with a university degree
  - Unemployed, retired, student or administrator
- The behavioral characteristics of these clients are:
  - With no credit defaults
  - Contacted during previous campaign (may have participated or not, although participated clients had better result in this campaign)

WHEN?

- In this campaign, clients who have been contacted during following months:
  - March
  - April
  - September
  - October
  - December
- As for weekdays, there is no significant difference among weekdays in terms of campaign success

HOW?

- In this campaign, clients who have been contacted via their cellphones were more participating in the campaign
- Also, the more times a client have been contacted the less the chances of him participating in the campaign. This insight need more data and exploration.
- In terms of call duration, the longer the operators spoke with the client the more likely the client participated in the campaign. But it could be the other way around: successful calls have longer call duration. This insight need more data and exploration as well.

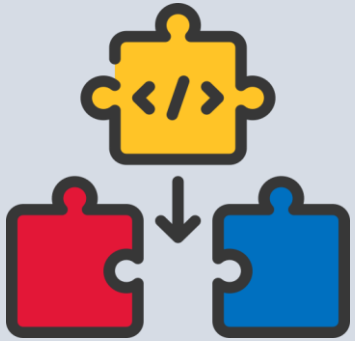
Recommendations

These recommendations are all based on the assumption that the goal is to purely increase the percentage of the participating client in direct marketing campaign. Other concerns and conditions such as goals of the campaign, number of planned participants and more, will greatly affect the course of action that should be taken based on this data and analysis.

Problem Statement

How to increase effectiveness of direct marketing campaign based on the dataset?

Analysis



Recommendations

- Focus on indicated demographics and behavioral segments (the answer to Q1) when starting new direct marketing campaign
- Hold the future direct marketing campaigns in the months that have been specified (the answer for Q2)
- Try to reach out to the clients via their cellphones rather than telephones
- Try to engage the clients more in the conversation and make him feel more invested in it.

Code & Repository Address

Code & Files repository Link

The project Python code could be found in GitHub code repository:

Dataset & Data Description

- Dataset Description
- Data Description

Dataset Description

- About Dataset
- Dataset Characteristic

About Dataset

Title	Bank Marketing Data
Source	Data.world
Dataset	[5] & [6]
Data Link	[7]
Description	The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. [6]
Main Article	Moro et al., 2014, S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014 [5]

About Dataset

There are four datasets extracted from Main Dataset used in the article [5]:

- **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]
- **bank-additional.csv** with 10% of the examples (4119), randomly selected from 1), and 20 inputs.
- **bank-full.csv** with all examples and 17 inputs, ordered by date (older version of this dataset with less inputs).
- **bank.csv** with 10% of the examples and 17 inputs, randomly selected from 3 (older version of this dataset with less inputs).

The smallest datasets are provided to test more computationally demanding machine learning algorithms (e.g., SVM).

The classification goal was to predict if the client will subscribe (yes/no) a term deposit (variable y).

Dataset Characteristic

- The data 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.
- The data is example of data very close to the data analyzed in the main article ordered by date from **May 2008 to November 2010**.

Resources

Resources

1. https://www.investopedia.com/terms/t/telemarketing.asp#:~:text=Key%20Takeaways-,Telemarketing%20is%20the%20direct%20marketing%20of%20goods%20or%20services%20to,lead%20generation%2C%20and%20sales%20calls
2. https://qlutch.com/telemarketing/create-a-plan-for-a-telemarketing-campaign
3. https://www.callboxinc.com/telemarketing/the-problem-with-telemarketing-today/
4. https://mytekrescue.com/benefits-of-effective-marketing-campaigns/
5. https://archive.ics.uci.edu/ml/datasets/Bank+Marketing
6. https://data.world/data-society/bank-marketing-data
7. https://query.data.world/s/lqrsaugj7kwkyazkvydowjuxjrybxx

Sources: Dataset

Dataset	https://data.world/data-society/bank-marketing-data http://archive.ics.uci.edu/ml/datasets/Bank+Marketing
Data Link	https://query.data.world/s/lqrsaugj7kwkyazkvydowjuxjrybxx