

## Assessment task 1: Exploration of data skills and issues



Course Title: Statistical Thinking for Data Science

Student Name: Yasaman Mohammadi  
Student ID: 24612626  
Date of Submission:  
11.59pm Sunday 27 August 2023

# 1. Introduction

---

## 1.1 Problem Statement

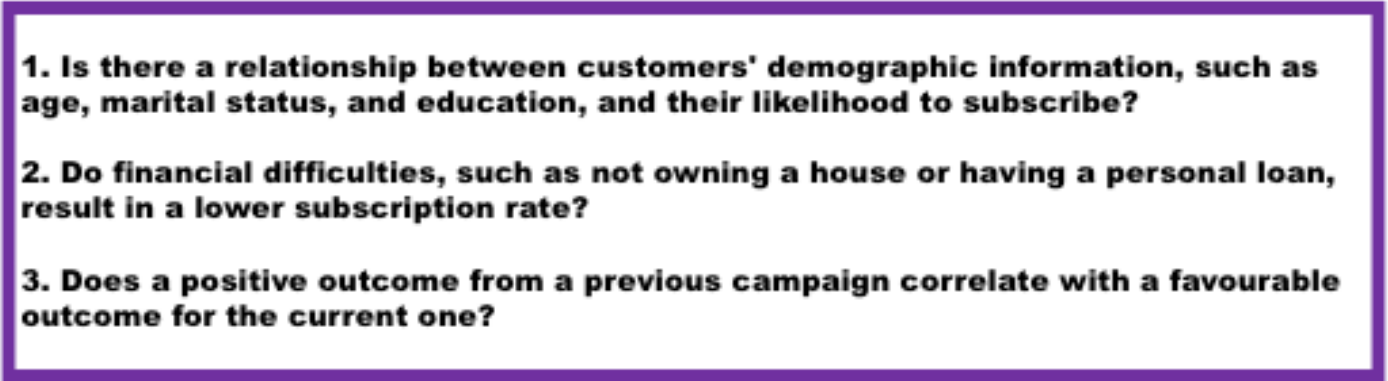
A telecommunication company recently launched a marketing campaign to promote the adoption of their new subscription plan among customers. The company seeks assistance in understanding its customers comprehensively and identifying the customer segments with the highest responsiveness to marketing campaigns.

## 1.2 Rationale

This study explores the campaign's impact, considering its substantial investments and potential outcomes. It delves into customer segmentation to identify the most receptive segments, offering insights for optimizing current and future marketing strategies.

## 1.3 Project aims and Objectives

An objective of this project is to assist the company in gaining an understanding of its customers and identifying the segments of customers with the highest potential for responding to marketing campaigns.

- 
- 1. Is there a relationship between customers' demographic information, such as age, marital status, and education, and their likelihood to subscribe?**
  - 2. Do financial difficulties, such as not owning a house or having a personal loan, result in a lower subscription rate?**
  - 3. Does a positive outcome from a previous campaign correlate with a favourable outcome for the current one?**

*Figure 1-Hypothesis*

We engaged in Exploratory Data Analysis (EDA) to uncover insights from the dataset.

# 2. Methodology

---

## 2.1 Methods Overview

In this project, we collect and analyse customer data to comprehend their behaviour and responses to marketing campaigns. We structured and cleaned the data, explored numerical and categorical variables, and examined their relationships. By calculating summary statistics, creating visualisations, and investigating key variables, we aim to identify customer segments most responsive to marketing efforts. This approach enables us to tailor campaigns effectively, optimising our strategies for better results.

## 2.2 Methods details

	<i>The steps Taken for EDA</i>	<i>Explanation</i>
1	Data Acquisition	UTS Canvas was used to acquire the data.
2	Understand Data Structure	The dataset consists of 41180 records of customer information. It consists of 20 variable predictors, including both categorical and continuous variables.
3	Basic Summary Statistics	Performing basic summary is crucial for spotting data issues, understanding the data's distribution, aiding feature and model selection, handling outliers, guiding preprocessing, setting baselines, facilitating communication, identifying patterns, and recognizing data imbalances, all of which collectively lay the foundation for informed decision-making and effective analysis.
4	Data Cleaning	Pre-processing the data is necessary to eliminate irrelevant, missing, or corrupted data points, and to rectify any discrepancies. The data contained 12 duplicated rows and no missing values.
5	Explore the response variable	Constructing a pie chart illustrates data imbalance, highlighting a larger portion of the data representing non-subscriber customers.
6	Explore numerical predictor variables	We generated various plots to comprehend the numerical variables.
7	Explore categorical predictor variables	We generated various plots to comprehend the categorical variables.
8	Explore relationship between variables	We generated a heatmap to visualize the correlations between variables.
9	Additional visualisations	Plotting different charts in Python and Tableau provides Visual understanding of data.

Table 1-The Steps Taken for EDA

### 3. Results

#### 2.1 Key findings

plotting different charts from categorical and numerical features helped to gain insight:

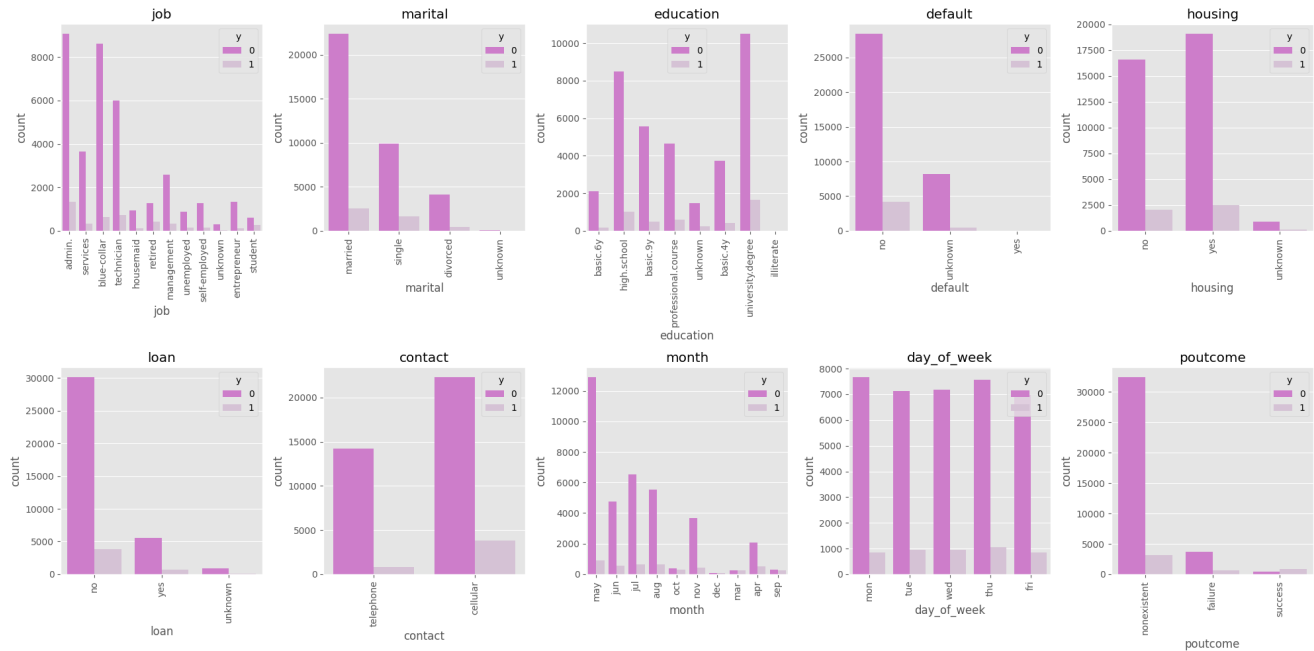


Figure 2-Bar plots of categorical variables

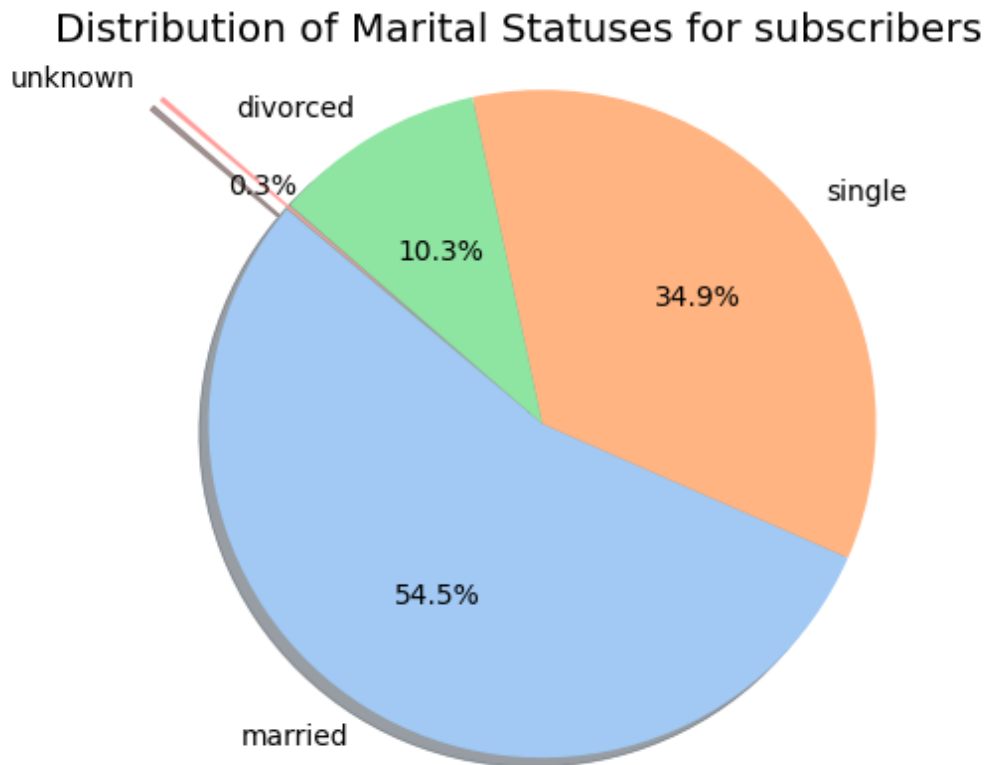
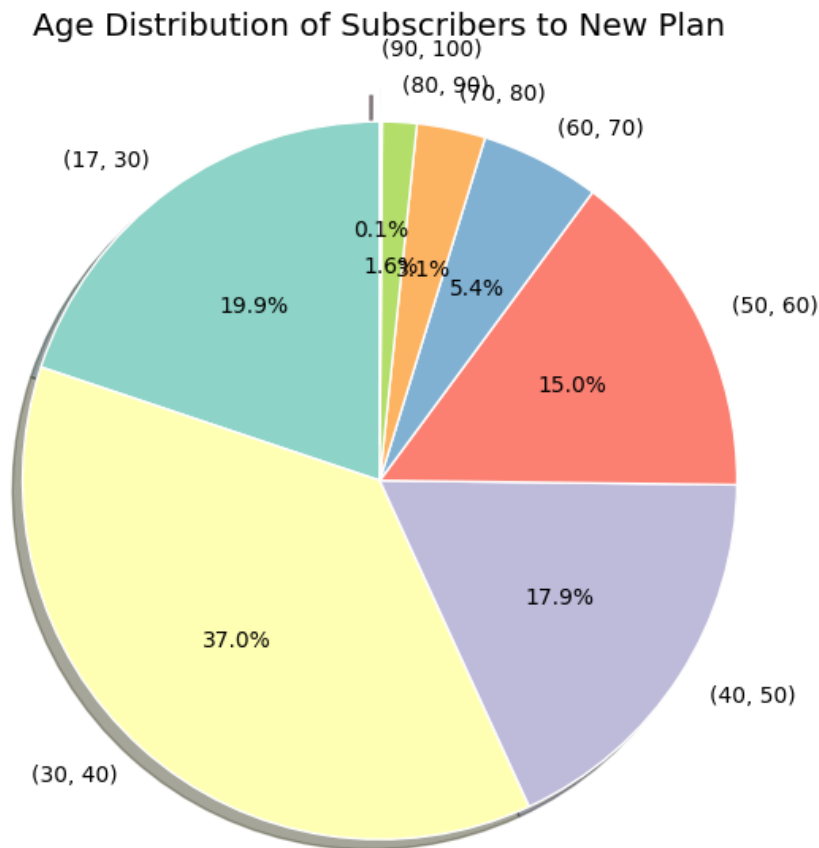


Figure 3- Pie Chart of Subscribers' Marital Statuses



*Figure 4- Pie Chart of Age Distribution Among Subscribers of the New Plan*

The charts present critical findings:

- Blue-collar, technician, and administrative job roles have the highest subscription rates.
- Married individuals subscribe more than divorced or single ones.
- Subscribed customers often have university degrees.
- Financial struggles (no credit history, loans) lead to lower subscriptions.
- Communication method affects subscription rates.
- May shows the least successful outcomes; different months have varying impacts on subscriptions rate.
- The subscription rate remains relatively consistent across all days of the week.
- Previous campaign success aids new campaigns.
- Most subscribers are aged 30 to 40.
- Shorter time since previous campaign contact correlates with better response to new campaign.

However, it's important to acknowledge the data's imbalance and recognize that some elements of the experiment may not be accurate. Employing further experiments and machine learning techniques is advisable to attain more refined insights.

## 4. Conclusion

---

In summary, the analysis reveals connections between customer demographic data and subscription rates, as well as a correlation between financial difficulties and a reduced likelihood of subscribing.

Furthermore, the method of communication, the timing of campaigns, and the positive impact of previous campaign outcomes have been identified as influential factors in subscription rates.

It's essential to note the data imbalance and potential inaccuracies in the experiment. To enhance the precision of our insights, it is advisable to conduct additional experiments and leverage advanced machine-learning techniques.

Moreover, when pursuing these insights, it's imperative to uphold ethical considerations to protect customer privacy. Safeguarding sensitive information and adhering to privacy regulations should remain a paramount concern throughout the analysis and implementation processes.

## 5. References

1. TIBCO. (n.d.). What is a Parallel Coordinate Plot? TIBCO Spotfire Documentation.

Retrieved from:

[https://docs.tibco.com/pub/spotfire/6.5.2/doc/html/para/para\\_what\\_is\\_a\\_parallel\\_coordinate\\_plot.htm#:~:text=A%20parallel%20coordinate%20plot%20maps,a%20plot%20is%20substantially%20different.](https://docs.tibco.com/pub/spotfire/6.5.2/doc/html/para/para_what_is_a_parallel_coordinate_plot.htm#:~:text=A%20parallel%20coordinate%20plot%20maps,a%20plot%20is%20substantially%20different.)

## 6. Appendices

	age	duration	campaign	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
count	41180.000000	41180.000000	41180.000000	41180.000000	41180.000000	41180.000000	41180.000000	41180.000000	41180.000000	41180.000000
mean	40.021710	258.280427	2.567800	962.516707	0.172705	0.081901	93.575508	-40.501999	3.621422	5167.053344
std	10.419593	259.299856	2.770225	186.809028	0.493719	1.571037	0.578762	4.627358	1.734385	72.230334
min	17.000000	0.000000	1.000000	0.000000	0.000000	-3.400000	92.201000	-50.800000	0.634000	4963.600000
25%	32.000000	102.000000	1.000000	999.000000	0.000000	-1.800000	93.075000	-42.700000	1.344000	5099.100000
50%	38.000000	180.000000	2.000000	999.000000	0.000000	1.100000	93.749000	-41.800000	4.857000	5191.000000
75%	47.000000	319.000000	3.000000	999.000000	0.000000	1.400000	93.994000	-36.400000	4.961000	5228.100000
max	98.000000	4918.000000	56.000000	999.000000	7.000000	1.400000	94.767000	-26.900000	5.045000	5228.100000

Figure 5-Basic Summary Statistics

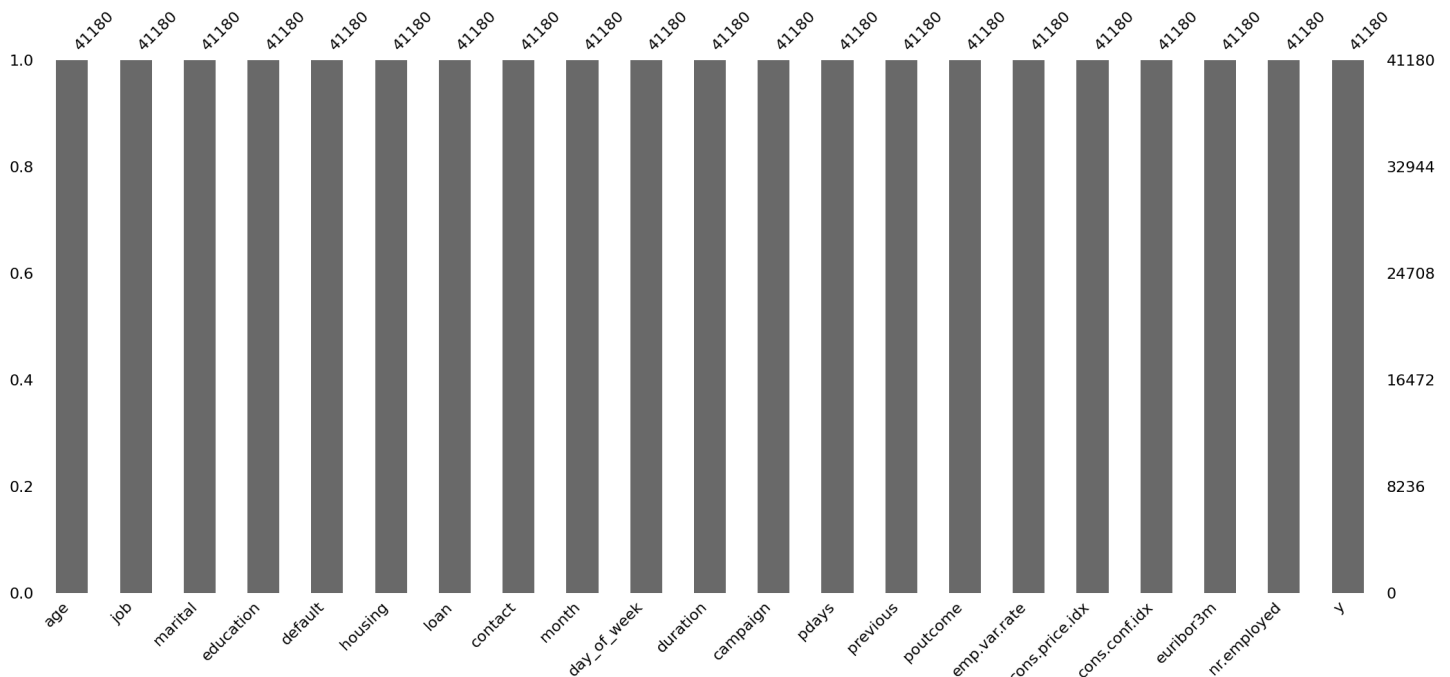


Figure 6-missingno chart for showing missing values

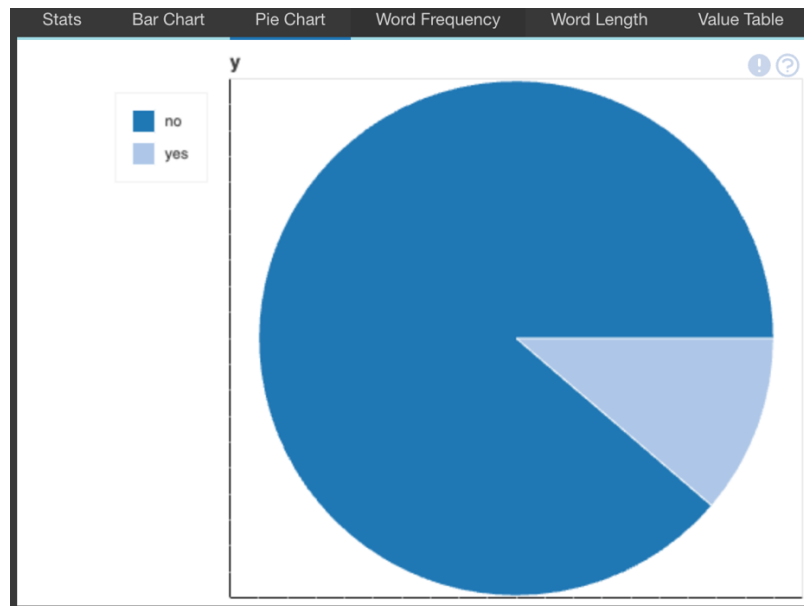


Figure 7-Pie chart of response variable

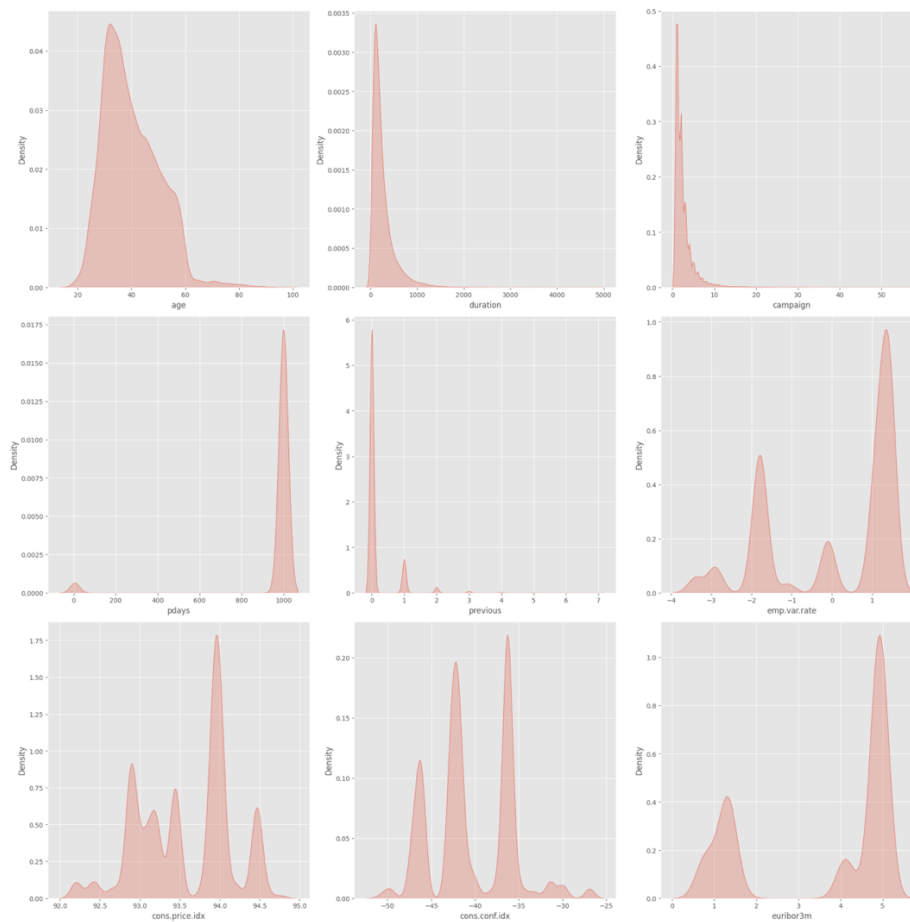


Figure 8-Density plots (Duration, campaign, and age features are left skewed.)



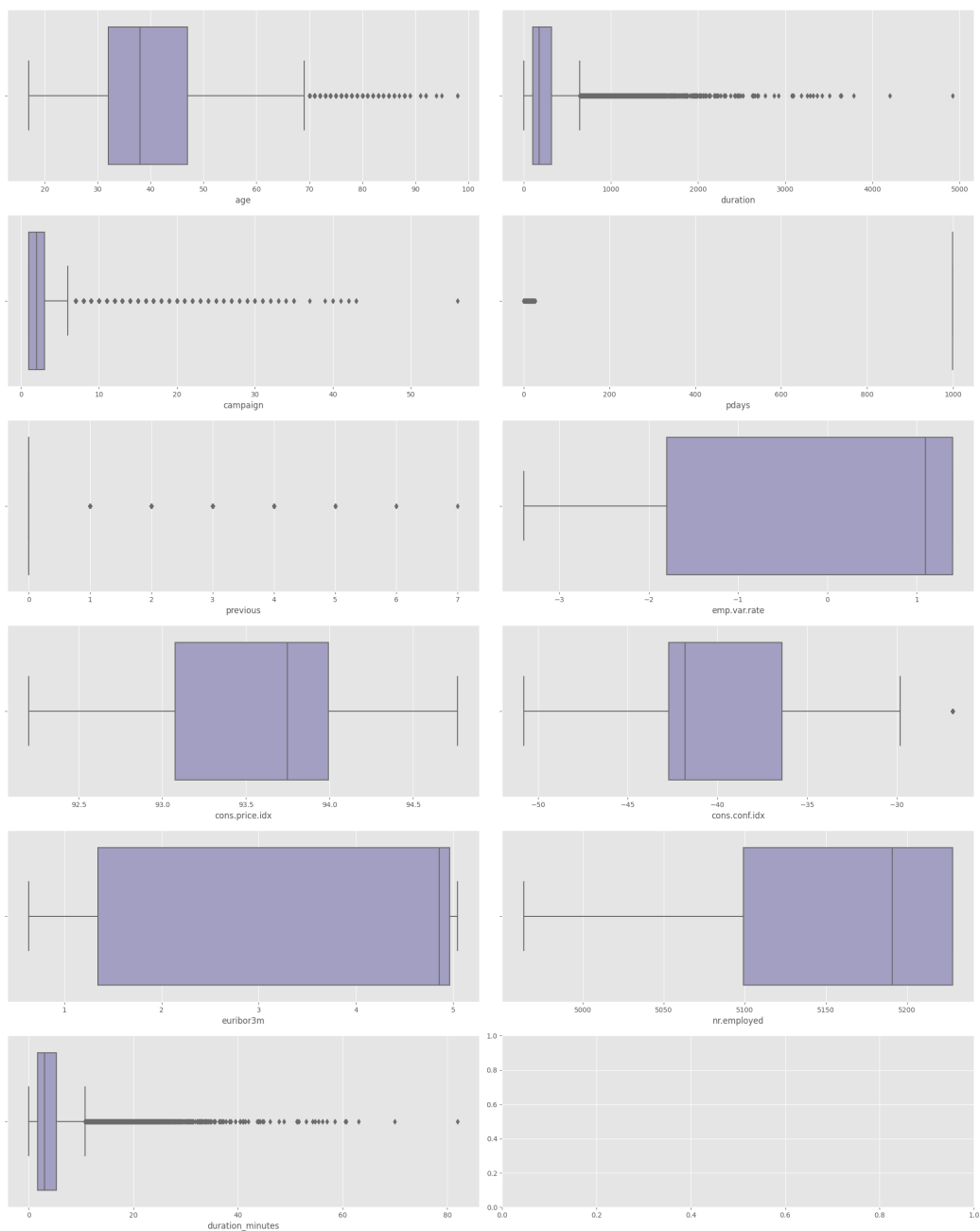


Figure 9-box plots of numerical variables

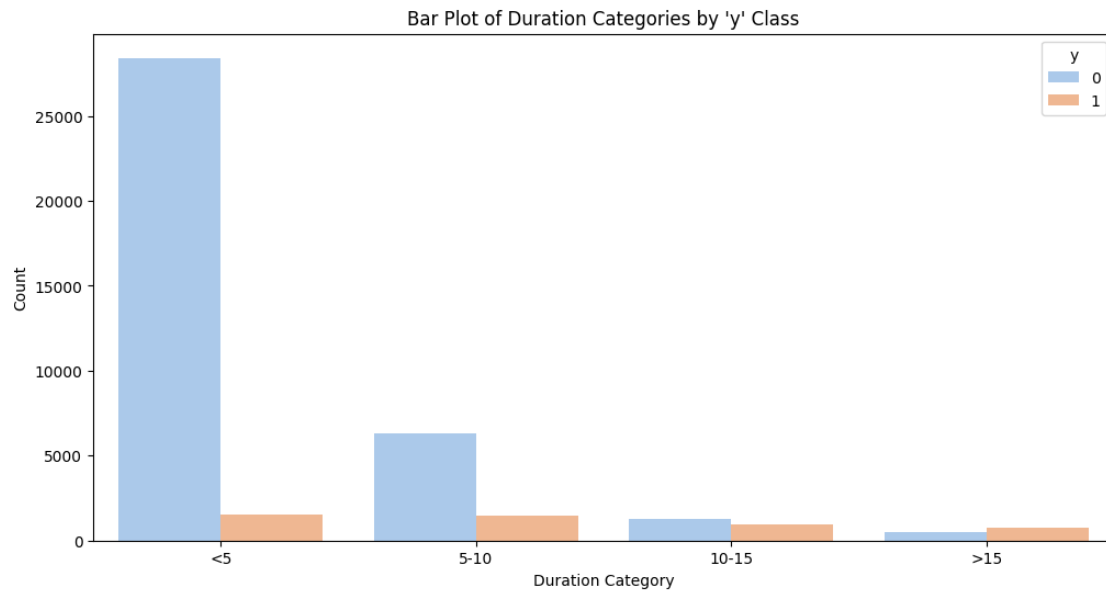


Figure 10-bar plot of Duration Categories

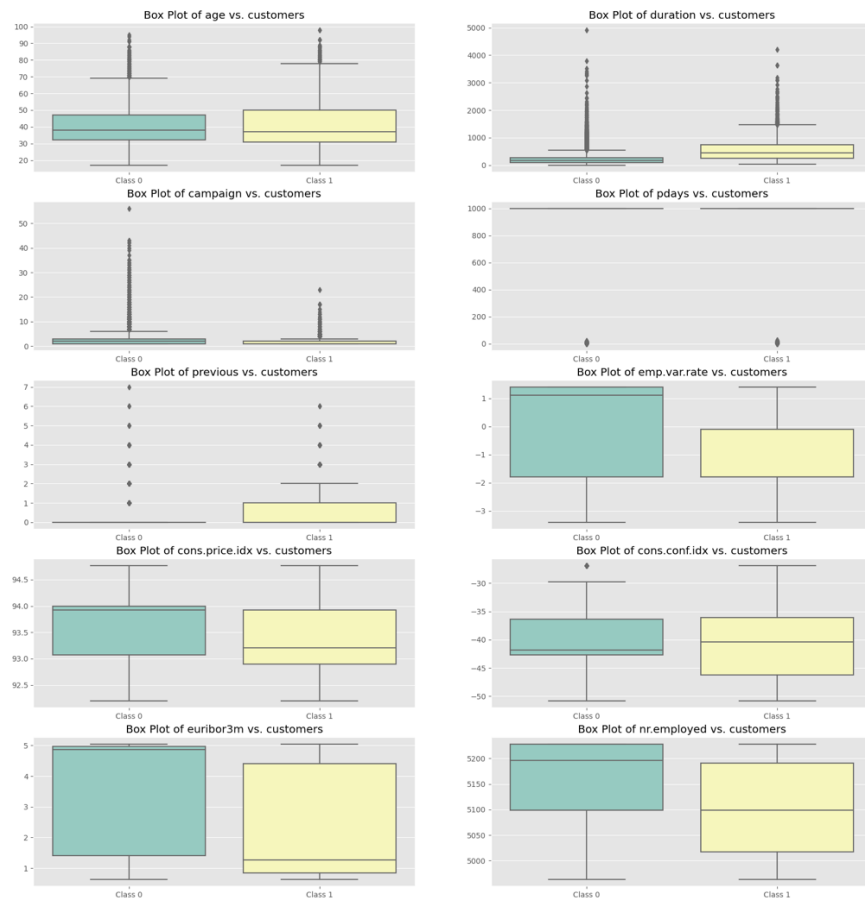


Figure 11-box plot of each class

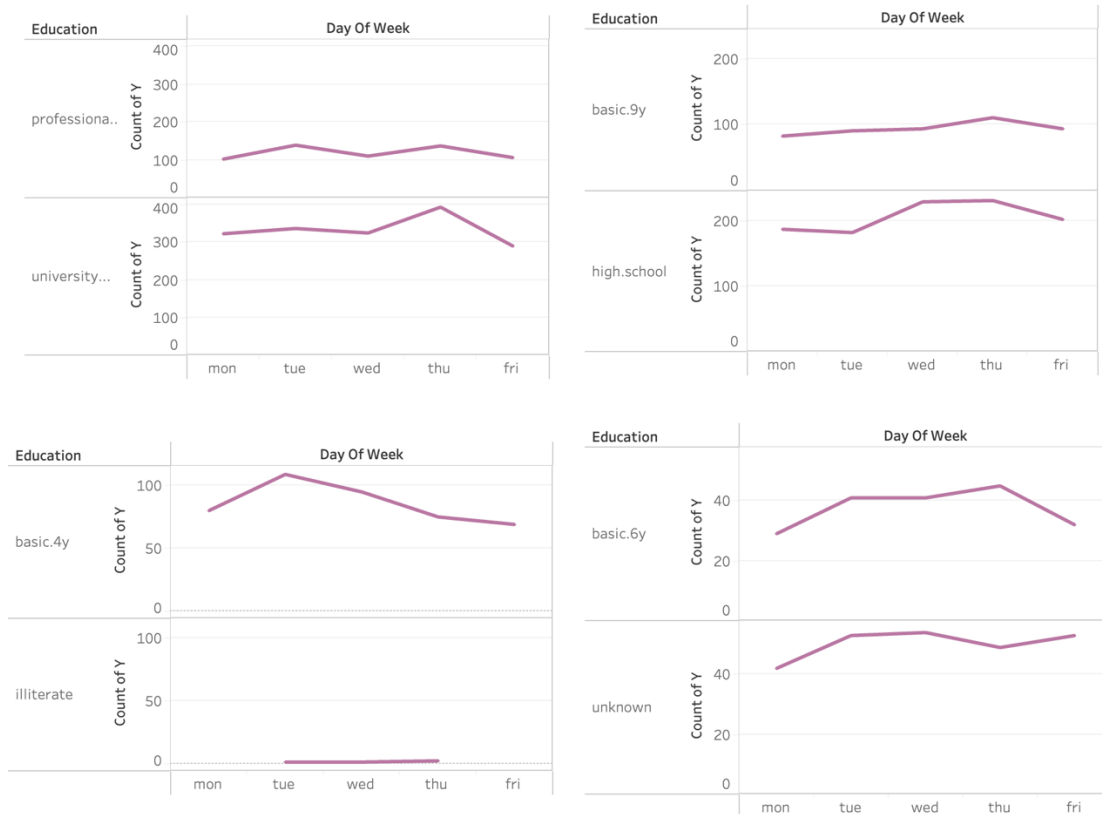


Figure 12- Parallel Coordinates plot by Tableau (finding relationship between days of week and education level vs subscription rate)

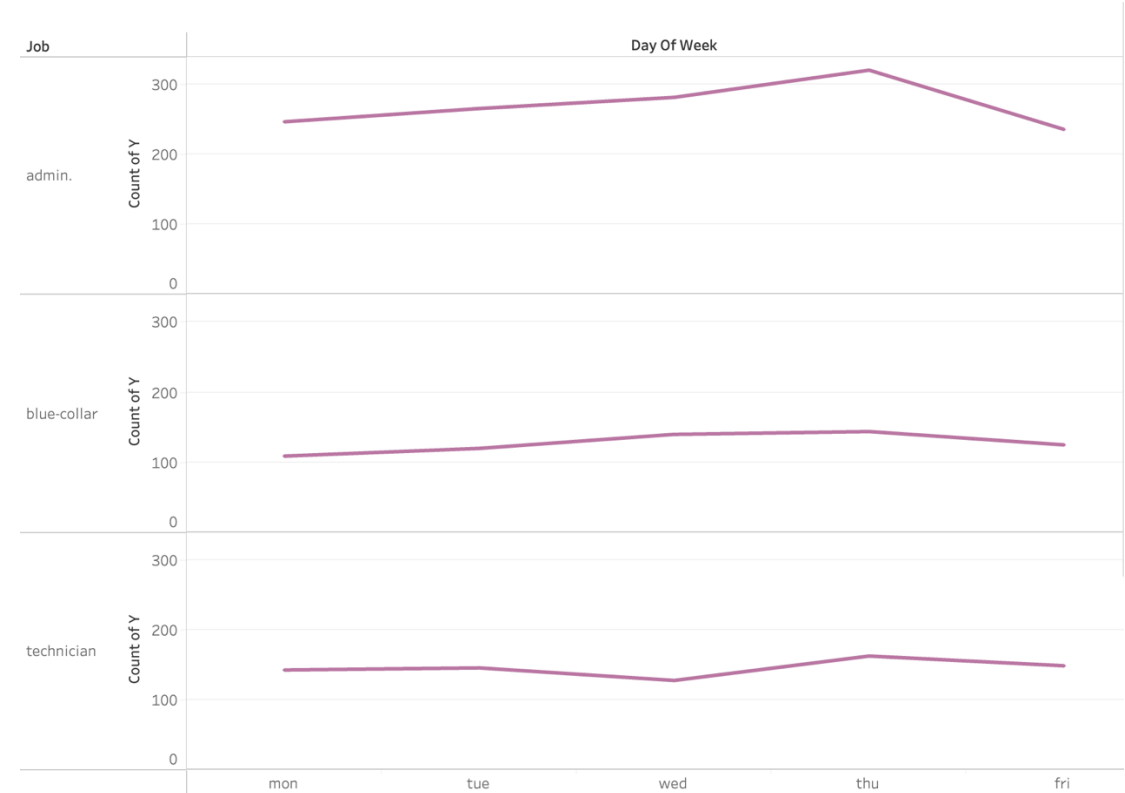


Figure 13- Parallel Coordinates plot by Tableau (finding relationship between days of week and job vs subscription rate)

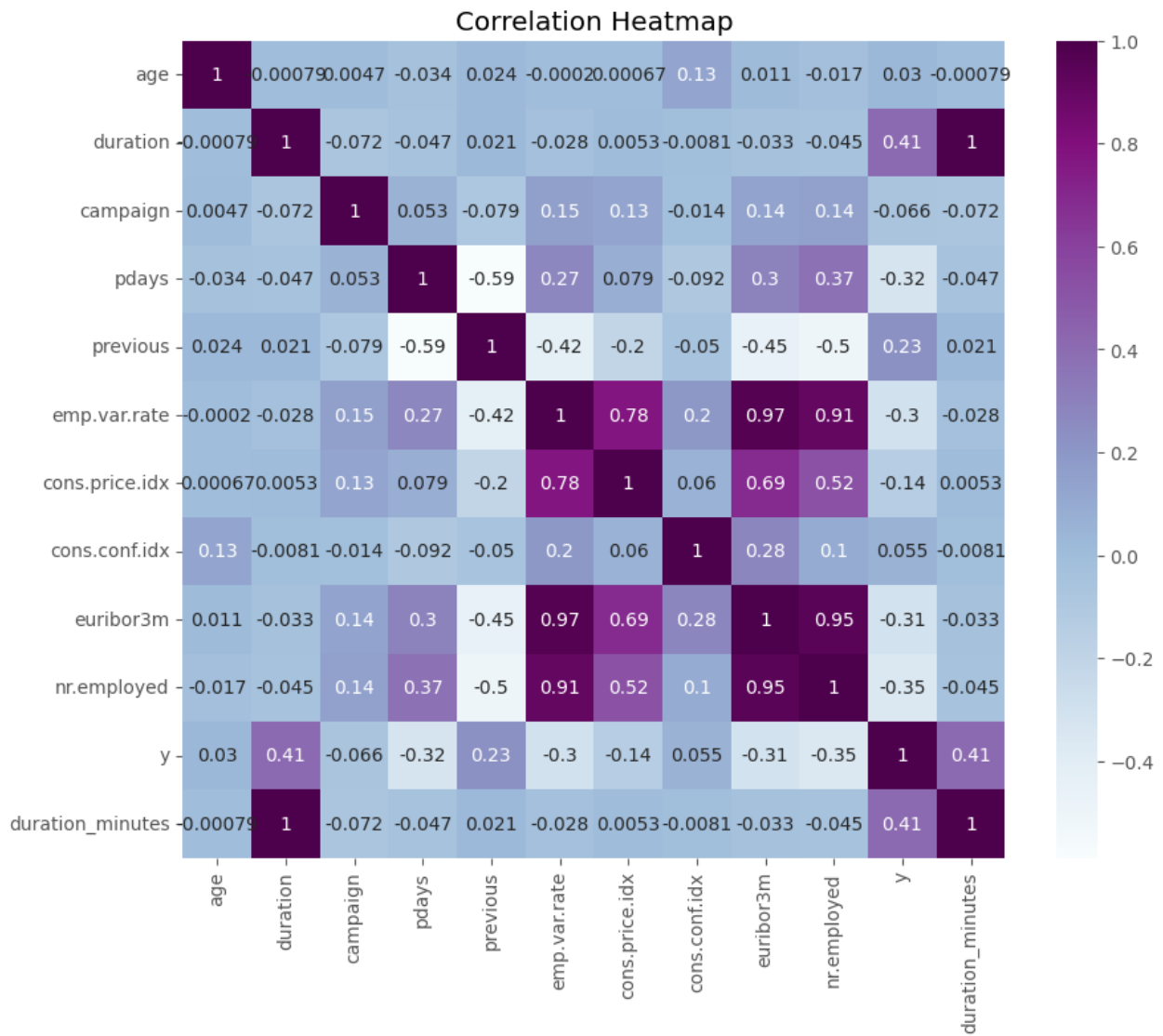


Figure 14-Correlation Heatmap

Variable Name	Description
age	Age
job	Type of job
marital	Marital status
education	Level of education
default	Has credit in default
balance	Average yearly balance
housing	Has a housing loan
loan	Has a personal loan
contact	Contact communication type
day	Day of contact
month	Month of contact
duration	Last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y="no"). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known.
campaign	Number of contacts performed during this campaign and for this client
pdays	Number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)
previous	Number of contacts performed before this campaign and for this client
poutcome	Outcome of the previous marketing campaign
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 employed - quarterly indicator (numeric)
y	Did the client subscribe to a Telecom plan?

Figure 15-Data Dictionary

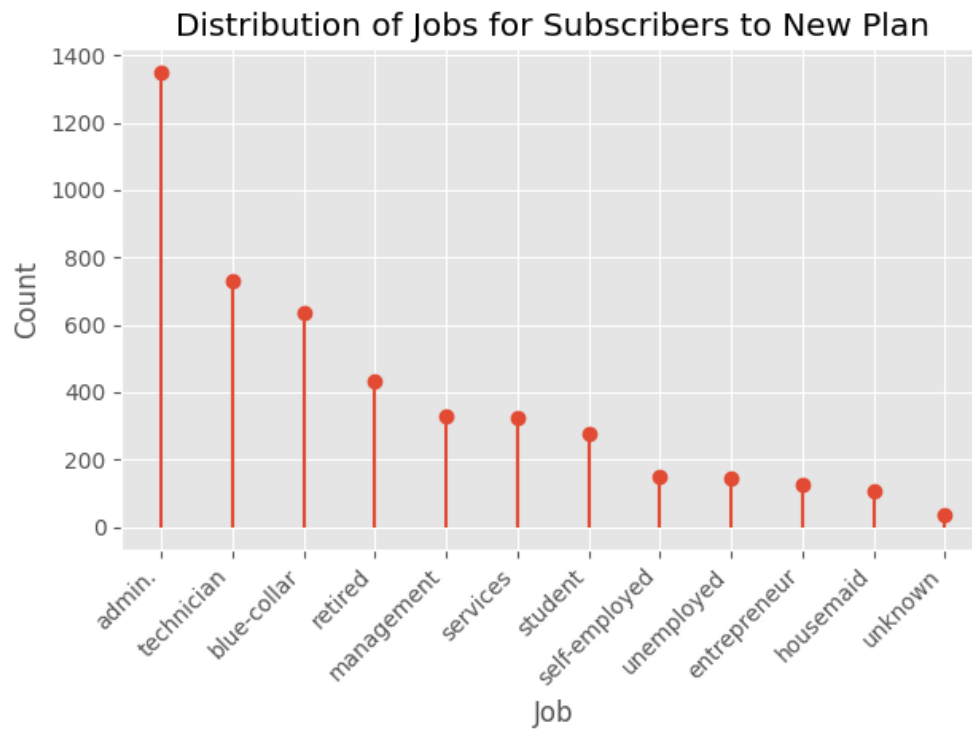


Figure 16-Distribution of Occupations Among Subscribers of the New Plan

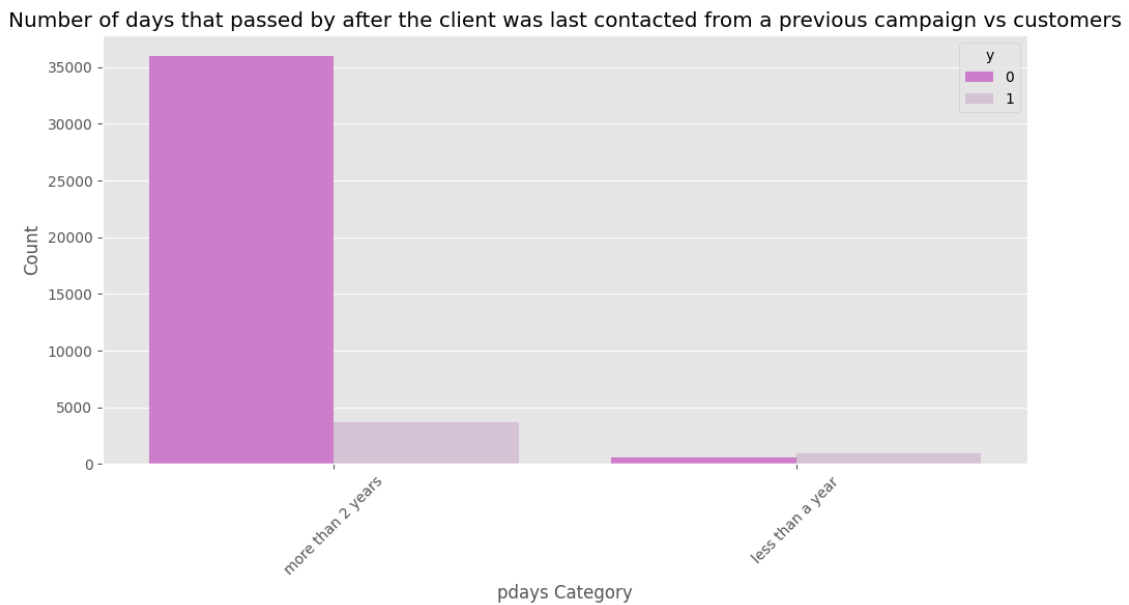


Figure 17-Number of days that passed by after the client was last contacted from a previous campaign vs customers.