



# **Data Science Intern at Data Glacier**

**Week 8: Report on Group Project**

**Topic: Bank Marketing (Campaign)**

**Group Name: Campaign Catalysts**

**Specialization: Data Science**

**Batch Code: LISUM19**

**Date: 26<sup>th</sup> April 2023**

**Submitted to: Data Glacier**

## Group Member Details

S.No.	Name	Email	Country	College/ Company
1.	Yash Jayesh Doshi	yashjdoshi99@gmail.com	UAE	Orpheuss LLC
2.	Anuj Singh	dsanuj21@gmail.com	India	Mumbai University
3.	Yash Jadwani	yash.jadwani1998@gmail.com	United Kingdom	Kingston University
4.	Harold Wilson	haroldwilson537@gmail.com	United Kingdom	University of Buckingham

# 1. Problem Description

ABC Bank wants to sell its term deposit product to customers and before launching the product they want to develop a model which help them in understanding whether a particular customer will buy their product or not (based on customer's past interaction with bank or other Financial Institution).

Bank wants to use ML model to shortlist customer whose chances of buying the product is more so that their marketing channel (tele marketing, SMS/email marketing etc.) can focus only to those customers whose chances of buying the product is more.

## 2. Data Understanding

We will be following the CRISP-DM (Cross-Industry Standard Process for Data Mining) approach to understand the data at hand. This approach involves the following steps:

### 2.1 Collecting the data

As described in the data intake report, we have imported four datasets for this project.

The first two of these data sets are:

- A) **bank-full.csv** with all examples, ordered by date (from May 2008 to November 2010). The bank-full data set contains 45211 X 17 observations.
- B) **bank.csv** with 10% of the examples (4521), randomly selected from bank-full.csv. The bank data set contains 4521 X 17 observations and is 10% of the examples (4521), randomly selected from bank-full.csv to test more computationally demanding machine learning algorithms (e.g. SVM).

More meta-data for these two datasets can be found below.

- **Data set location:**

This dataset is publicly available for research. The details are described in [Moro et al., 2011] where it used for Data Mining for Bank Direct Marketing: relating to a direct marketing campaigns of a Portuguese banking institution. [1]

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 subscribed or not.

The data is available and can be located at:

- [pdf] <http://hdl.handle.net/1822/14838>
- [bib] <http://www3.dsi.uminho.pt/pcortez/bib/2011-esm-1.txt>

- **Data Attributes/Variables/Features:**

The main variables /attributes of the data are:

- 1 - age (numeric)
- 2 - job: type of job (categorical, "admin", "unknown", "unemployed", "management", "housemaid", "entrepreneur", "student", "blue-collar", "self-employed", "retired", "technician", "services")
- 3 - marital: marital status (categorical: "married", "divorced", "single"; note: "divorced" means divorced or widowed)
- 4 - education (categorical: "unknown", "secondary", "primary", "tertiary")
- 5 - default: has credit in default? (binary: "yes", "no")
- 6 - balance: average yearly balance, in euros (numeric)
- 7 - housing: has housing loan? (binary: "yes", "no")
- 8 - loan: has personal loan? (binary: "yes", "no")
- 9 - contact: contact communication type (categorical: "unknown", "telephone", "cellular")
- 10 - day: last contact day of the month (numeric)
- 11 - month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")
- 12 - duration: last contact duration, in seconds (numeric)
- 13 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
- 14 - 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)
- 15 - previous: number of contacts performed before this campaign and for this client (numeric)
- 16 - poutcome: outcome of the previous marketing campaign (categorical: "unknown", "other", "failure", "success")

Output variable (desired target):

- 17 - y - has the client subscribed a term deposit? (binary: "yes", "no")

The other two datasets are:

- C) **bank-additional-full.csv** with all examples, ordered by date (from May 2008 to November 2010). The bank-additional-full data set contains 41188 X 21 observations.
- D) **bank-additional.csv** with 10% of the examples (4119), randomly selected from bank-additional-full. The bank-additional data set contains 4119 X 21 observations and is 10% of the examples (4119), randomly selected from bank-full.csv to test more computationally demanding machine learning algorithms (e.g. SVM).

More meta-data for these two datasets can be found below.

- **Data set location:**

This dataset is publicly available for research. The details are described in [Moro et al., 2014]

The data is available and can be located at:

- [pdf] <http://dx.doi.org/10.1016/j.dss.2014.03.001>
- [bib] <http://www3.dsi.uminho.pt/pcortez/bib/2014-dss.txt>

This dataset is based on "Bank Marketing" UCI, and is enriched by the addition of five new social and economic features/attributes collected from a national wide indicators from a 10M population country and published by the Banco de Portugal and publicly available at:

<https://www.bportugal.pt/estatisticasweb> and was found to lead to a successful substantial improvement in the prediction process.

- **Data Attributes/Variables/Features:**

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")

16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)

17 - cons.price.idx: consumer price index - monthly indicator (numeric)

18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)

19 - euribor3m: euribor 3 month rate - daily indicator (numeric)

20 - nr.employed: number of employees - quarterly indicator (numeric)

Output variable (desired target):

21 - y - has the client subscribed a term deposit? (binary: "yes", "no")

**Note:** The dataset chosen for further analysis and for ML model creation will be **bank-additional-full.csv** because it seems to be ideal for the purpose of this project since it is more recent and has more variables, which helps us to build an efficient model.

## 2.2 Describing the data

Below images display the meta-data of the data that we will be using along with the meta-data of it's attributes.

```
Meta-data for bank-additional-full.csv
-----
Number of rows: 41188
Number of columns: 21
Delimiter used: ;
Size of data file: 5.564617156982422 MB
```

```
Attribute Meta-data
-----
Number of Numerical features: 10
List of Numerical features:
['age', 'duration', 'campaign', 'pdays', 'previous', 'emp.var.rate', 'cons.price.idx', 'cons.conf.idx', 'euribor3m', 'nr.employed']
Number of categorical features: 11
List of categorical features:
['job', 'marital', 'education', 'default', 'housing', 'loan', 'contact', 'month', 'day_of_week', 'poutcome', 'y']
```

Below images show the first five records, the last five records and, random five records of the data. This helps us better understand the data at hand. We used the datassist library for the following output.

```
In [24]: ds.structdata.describe(df_bank_additional_full)
First five data points
  age      job  marital  ... euribor3m nr.employed  y
0  56  housemaid  married  ...    4.857      5191.0  no
1  57   services  married  ...    4.857      5191.0  no
2  37   services  married  ...    4.857      5191.0  no
3  40    admin.  married  ...    4.857      5191.0  no
4  56   services  married  ...    4.857      5191.0  no

[5 rows x 21 columns]

Random five data points
  age      job  marital  ... euribor3m nr.employed  y
39987  30  technician  single  ...    0.761      4991.6  no
18731  34  technician  divorced  ...    4.968      5228.1  no
24825  32   services  married  ...    4.153      5195.8  no
35646  40    admin.  single  ...    1.244      5099.1  no
1726   59    admin.  married  ...    4.855      5191.0  no

[5 rows x 21 columns]

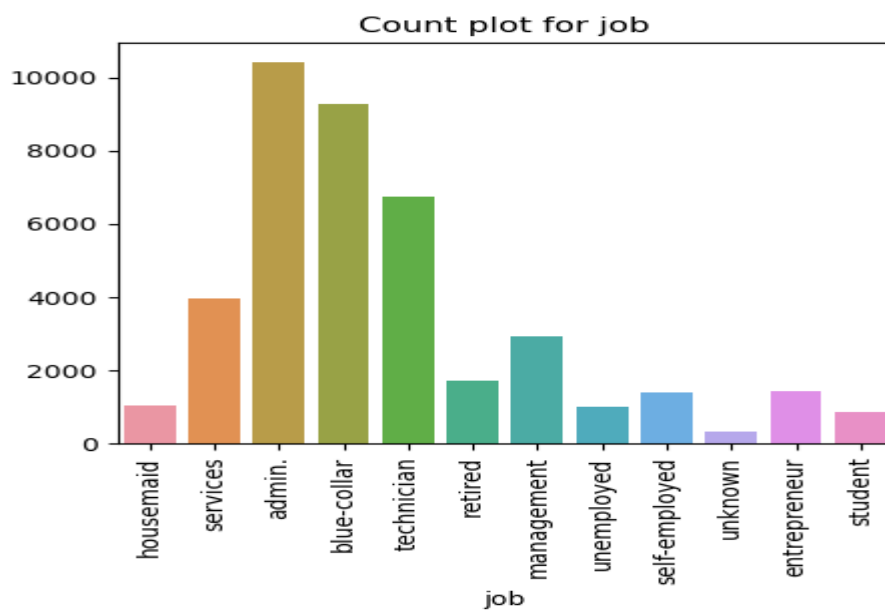
Last five data points
  age      job  marital  ... euribor3m nr.employed  y
41183  73    retired  married  ...    1.028      4963.6  yes
41184  46  blue-collar  married  ...    1.028      4963.6  no
41185  56    retired  married  ...    1.028      4963.6  no
41186  44  technician  married  ...    1.028      4963.6  yes
41187  74    retired  married  ...    1.028      4963.6  no
```

Lastly, we generated a report about the type of data in each attribute with the help of `pandas_dq` library. The following table shows the output of that report

Feature Name	Data Type	Missing Values%	Unique Values%	Minimum Value	Maximum Value
age	int64	0	0	17	98
job	object	0	0	admin.	unknown
marital	object	0	0	divorced	unknown
education	object	0	0	basic.4y	unknown
default	object	0	0	no	yes
housing	object	0	0	no	yes
loan	object	0	0	no	yes
contact	object	0	0	cellular	telephone
month	object	0	0	apr	sep
day_of_week	object	0	0	fri	wed
duration	int64	0	3	0	4918

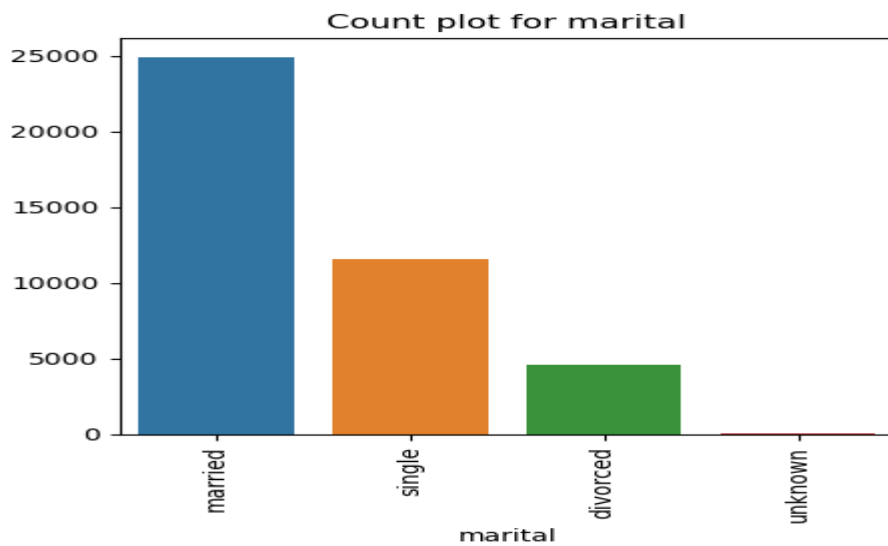
campaign	int64	0	0	1	56
pdays	int64	0	0	0	999
previous	int64	0	0	0	7
poutcome	object	0	0	failure	success
emp.var.rate	float64	0	NA	-3.4	1.4
cons.price.idx	float64	0	NA	92.201	94.767
cons.conf.idx	float64	0	NA	-50.8	-26.9
euribor3m	float64	0	NA	0.634	5.045
nr.employed	float64	0	NA	4963.6	5228.1
y	int64	0	0	0	1

## 2.3 Exploring the data

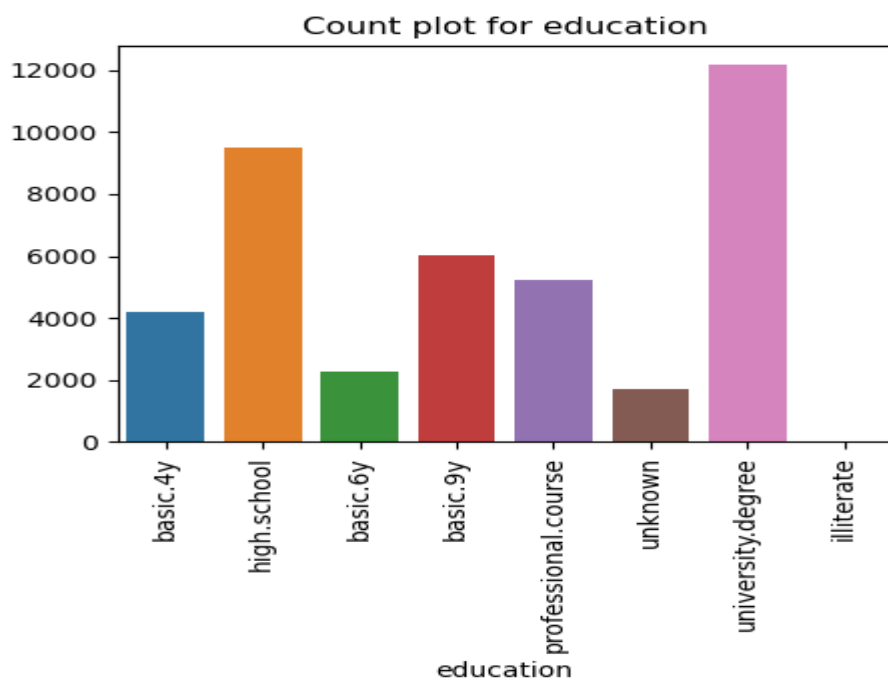


Observation: People with administrative job type were more involved in this project, followed by people with blue-collar jobs and technicians. The least number of people involved in this survey for the project were students, housemaid and the unemployed.

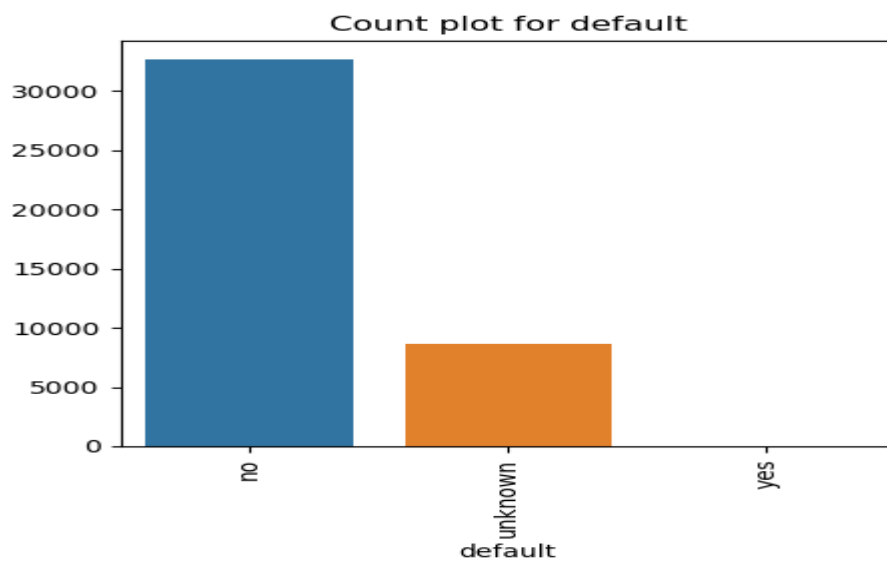




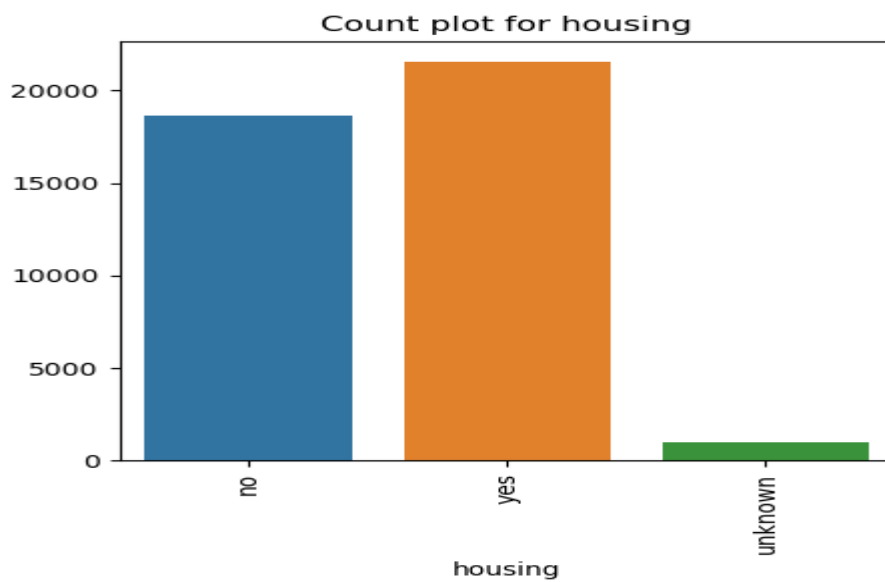
Observation: Married people were more involved in this project followed by single and the divorced which recorded the least number of people for this bank marketing campaign survey.



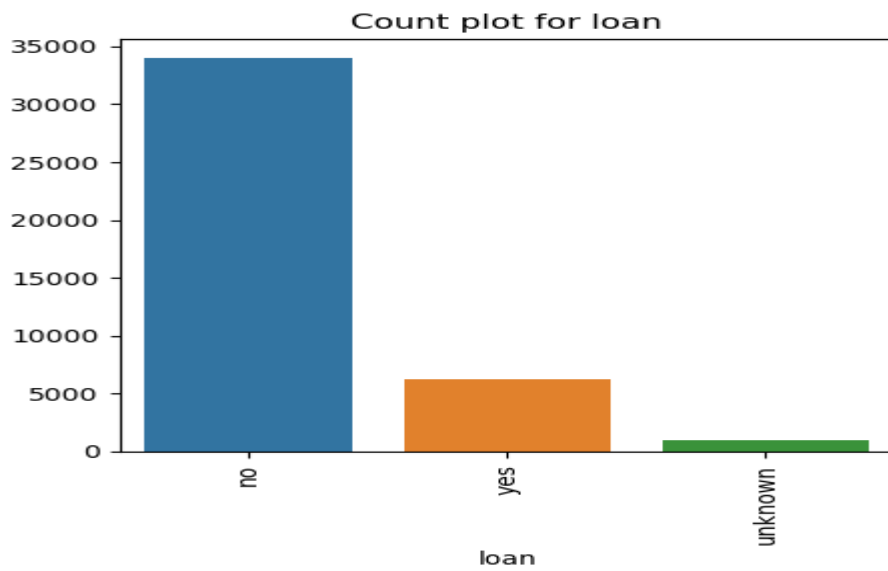
Observation: More people with a university degree took part in this bank marketing survey, followed by high school students. illiterate people recorded the least number considered in this bank marketing campaign survey.



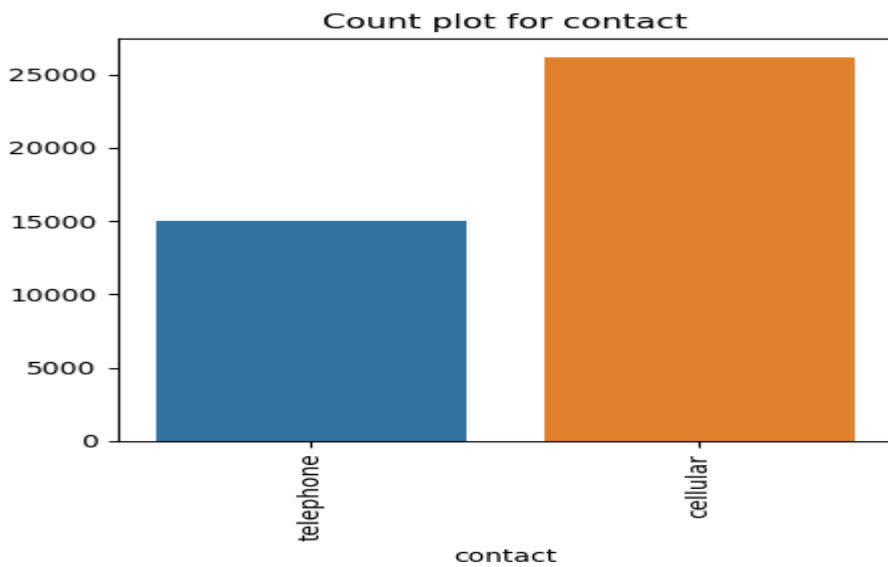
Observation: The number of people that defaulted on credit were much more than the people that did not creating a class imbalance which typically is always a fundamental issue when it comes to classification for machine learning process.



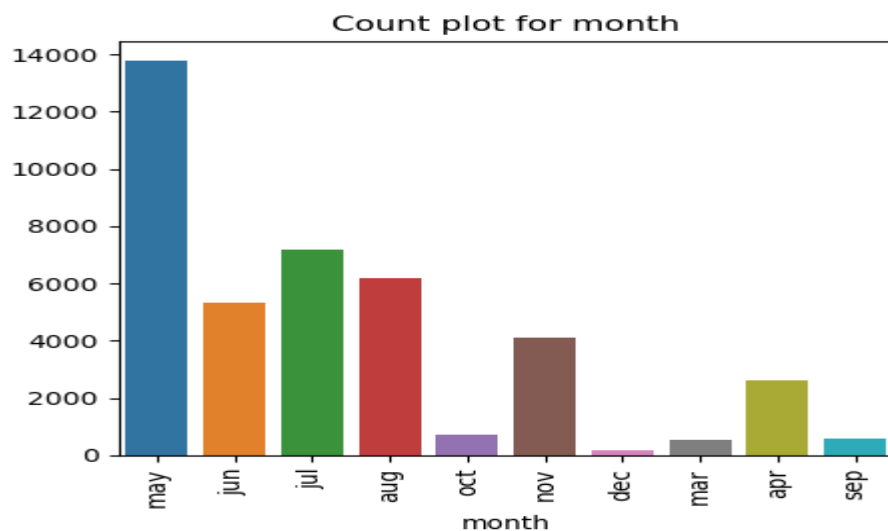
Observation: There were slightly more people with housing loan than people without.



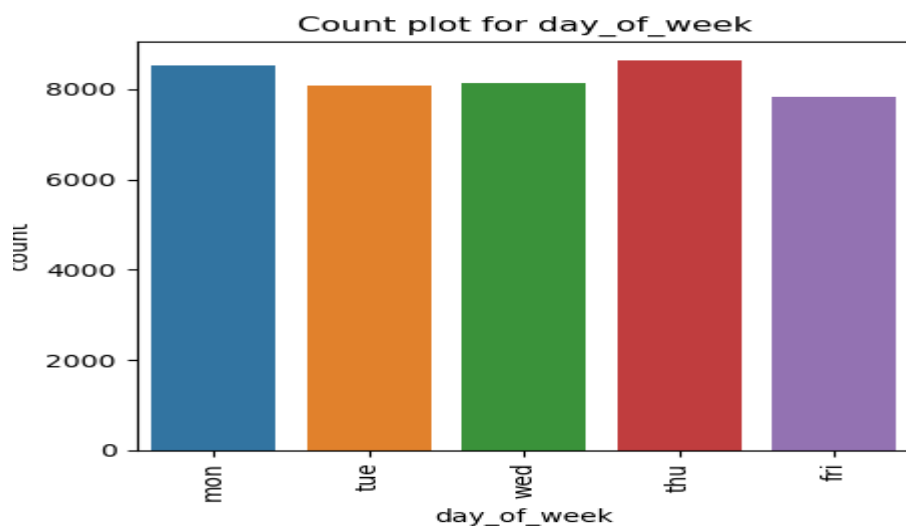
Observation: The number of people who had no loan were about ten times the people who had a loan, meaning more people without a loan took part in this survey than people with a loan.



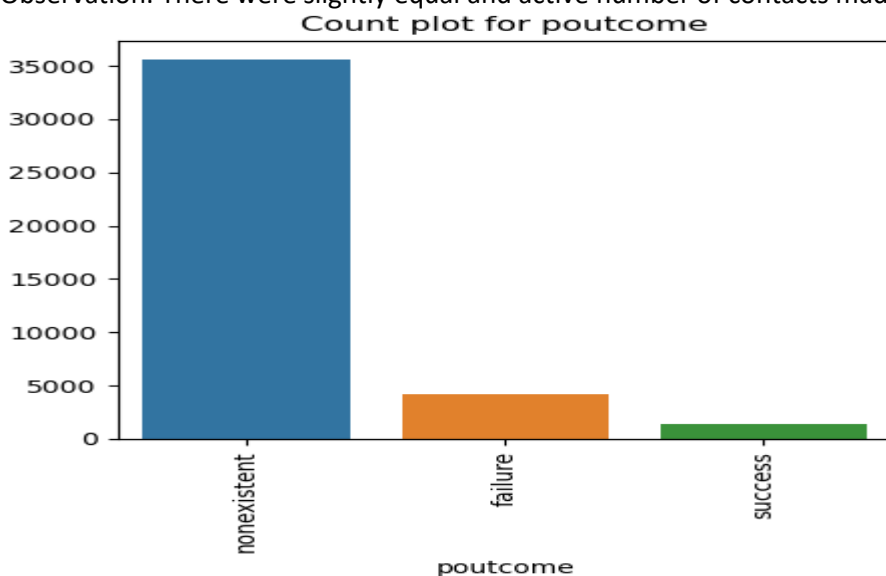
Observation: There was more contact made by cellular than telephone where for every three people contacted by telephone five more were contacted by cellular.



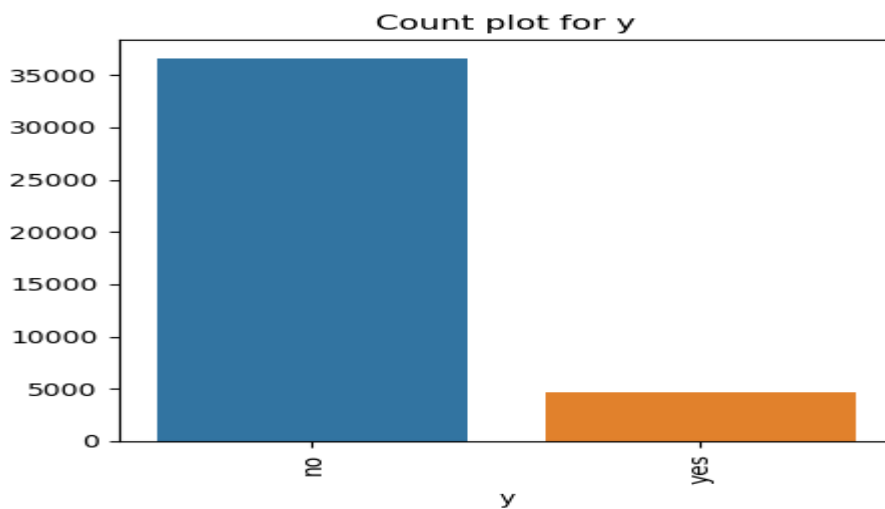
Observation: The month of May saw more last contacts been made than any month of the year according to the survey, followed by June and August with December recording the least number of contacts been made.



Observation: There were slightly equal and active number of contacts made for the days of the week.



Observation: There were more failures than success for the outcome of the previous campaign recorded with a high number of non-existent records with no activity.



Observation: The number of clients that subscribed for a term deposit were ten times more than clients that did not subscribe for a term deposit creating yet again an imbalance class problem.

## 2.4 Checking the quality of the data

### A) Null/ Missing Values:

Missing Values in Data			
	features	missing_counts	missing_percent
0	age	0	0.0
1	job	0	0.0
2	marital	0	0.0
3	education	0	0.0
4	default	0	0.0
5	housing	0	0.0
6	loan	0	0.0
7	contact	0	0.0
8	month	0	0.0
9	day_of_week	0	0.0
10	duration	0	0.0
11	campaign	0	0.0
12	pdays	0	0.0
13	previous	0	0.0
14	poutcome	0	0.0
15	emp.var.rate	0	0.0
16	cons.price.idx	0	0.0
17	cons.conf.idx	0	0.0
18	euribor3m	0	0.0
19	nr.employed	0	0.0
20	y	0	0.0

There are practically no missing values present in the data set, however there more unknown records present in the data set and this could be seen in the below for the different variables in the data set.

```
In [17]: df_bank_additional_full.job.value_counts()
Out[17]:
admin.          10422
blue-collar     9254
technician      6743
services        3969
management      2924
retired         1720
entrepreneur    1456
self-employed   1421
housemaid       1060
unemployed      1014
student         875
unknown         330
Name: job, dtype: int64
```

There are 330 unknown records for the job variable.

```
In [18]: df_bank_additional_full.marital.value_counts()
Out[18]:
married      24928
single       11568
divorced      4612
unknown        80
Name: marital, dtype: int64
```

There are 80 unknown records in the marital variable.

```
In [19]: df_bank_additional_full.education.value_counts()
Out[19]:
university.degree    12168
high.school          9515
basic.9y             6045
professional.course   5243
basic.4y             4176
basic.6y             2292
unknown              1731
illiterate            18
Name: education, dtype: int64
```

There are 1731 unknown records for the education variable.

```
In [20]: df_bank_additional_full.default.value_counts()
Out[20]:
no          32588
unknown     8597
yes           3
Name: default, dtype: int64
```

There are 8597 unknown records for the default variable.

```
In [28]: df_bank_additional_full.poutcome.value_counts()
Out[28]:
nonexistent    35563
failure         4252
success        1373
Name: poutcome, dtype: int64
```

There are 35563 non-existent unknown records for the poutcome variable.

## B) Duplicate Values:

```
In [39]: df_bank_additional_full.duplicated().sum()
Out[39]: 12
```

Twelve duplicate values were found in the dataset. But on further exploration, it was found that the duplicated records do have unique values in a few features, which means they are not entire duplicates. Hence, these records will not be removed.

## C) Outlier detection:

The following table shows the outliers for numerical features and rare categories for categorical features of the dataset. It also shows how we intend to deal with them.

Feature Name	Outliers/ Rare Categories	How to deal with them?
age	has 468 outliers greater than upper bound (69.5) or lower than lower bound(9.5).	Convert into categorical feature by binning the values.
job	7 rare categories: ['retired', 'entrepreneur', 'self-employed', 'housemaid', 'unemployed', 'student', 'unknown'].	These can be grouped into a single category.
marital	1 rare categories: ['unknown'].	This will not be changed/ transformed.
education	2 rare categories: ['unknown', 'illiterate'].	This will not be changed/ transformed.
default	1 rare categories: ['yes'].	This will not be changed/ transformed.
housing	1 rare categories: ['unknown'].	This will not be changed/ transformed.
loan	1 rare categories: ['unknown'].	This will not be changed/ transformed.
contact	No issue	
month	4 rare categories: ['oct', 'sep', 'mar', 'dec'].	This will not be changed/ transformed.
day_of_week	No issue	
duration	has 2963 outliers greater than upper bound (644.5) or lower than lower bound(-223.5).	Convert into categorical feature by binning the values.
campaign	has 2406 outliers greater than upper bound (6.0) or lower than lower bound(-2.0).	Records with values greater than 15 will be removed as outliers.
pdays	has 1515 outliers greater than upper bound (999.0) or lower than lower bound(999.0)	Since 999 is just a placeholder, it will be replaced by -1.
previous	has 5625 outliers greater than upper bound (0.0) or lower than lower bound(0.0).	This will not be changed/ transformed.
poutcome	1 rare categories: ['success'].	This will not be changed/ transformed.
emp.var.rate	No issue	
cons.price.idx	No issue	
cons.conf.idx	has 446 outliers greater than upper bound (-26.949999999999992) or lower than lower bound(-52.150000000000006).	This will not be changed/ transformed.
euribor3m	No issue	
nr.employed	No issue	

y	has 4639 outliers greater than upper bound (0.0) or lower than lower bound(0.0). Cap them or remove them.	Sampling methods will be used to deal with this imbalance
---	---	---

## References:

1. [Moro et al., 2011] S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimarães, Portugal, October, 2011. EUROSIS.
2. [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

## GitHub Repo Link:

<https://github.com/singhanuj695/Data-glacier-Group-Project>