36103 Statistical Thinking for Data Science

Assessment 1 | Exploratory Data Analysis

Nathan Collins

12062131

Assessment 1:	Exploratory Data Analysis
Type:	Individual Assessment
Deliverables:	Jupyter Notebook (x1) Experiment Report (x1) Final Report **496 words**
Weight:	100 pts
Due:	Sunday, 27 August, 23:59

Assessment Criteria:

- Exploring, interpreting and visualising data
- Clarity and brevity in explaining data issues, and the appropriateness of exploratory data analysis.
- Designing and managing data investigations
- Depth of insight and applying a minimum of three distinct exploratory data analysis techniques to gain preliminary insights from the data.
- Clarity and fluency in writing
- Clarity and fluency in communicating your findings to a technical target audience.

Section 1: Business Understanding

[1.1] Business Objective & Data Mining Goals

A telecommunications company has launched a novel marketing campaign, promoting a subscription plan to their customers. As the consultant data scientist, the company seeks to identify customer segments that indicate high responsiveness to their recent campaign. This may require querying recurring cohort traits, their frequencies and influences. An Exploratory Data Analysis (EDA) is ultimately to be conducted on a provided dataset to identify cohorts and corresponding traits.

Ethical Considerations

Ethical implications can take place from misuse of data. While limited personal information is provided, mindfulness in exercising impactful consequences fashioned by third parties must always be applied. These may come in the form of the ***privacy*** *of those incorporated in the dataset, the* ***outcome*** *of use and ensuing decision* ***biases****.*

[1.2] Hypothesis

Null | *The marketing campaign has no significant influence on a new subscription uptake among customers.*

Alternate | *The marketing campaign has a significant influence on the uptake of the new subscription plan among customers. There are customer segments that display high sensitivity to the marketing campaigns.*

Section 2: Data Understanding and Preparation

[2.1] Understanding the Data

The dataset consists of **41180** observations across **21** variables.

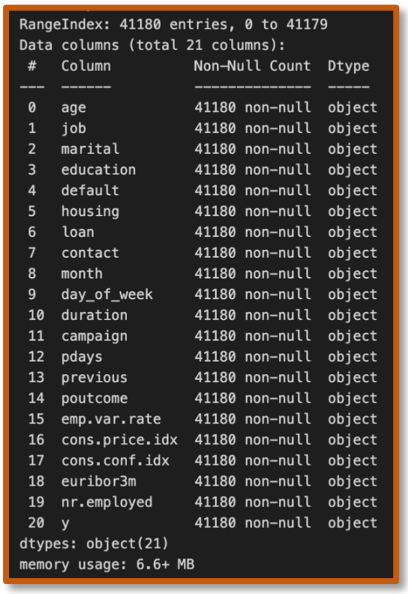
> Observations represent individual customers and features associated with that customer. Some features include the target subscription ("**y**", where a "yes" indicates they have subscribed)**,** alongside their **age**, **education** background, contacts performed during a **campaign**, and **employment variation rate**.

< See the appendix for a complete list of the features. >

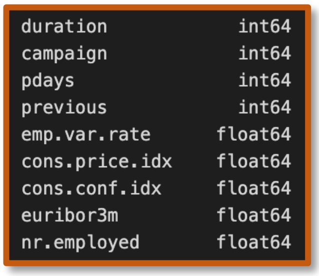
[2.2] Data Preparation

Prior to analysis, data was organised by applying conventional cleaning and manipulation techniques; first by transformation into a pandas data frame, followed by specific feature conversion into numerical formatting, (Figure 1, list of transformed variables).

< For descriptions of pre-processing actions, see the attached notebook. >



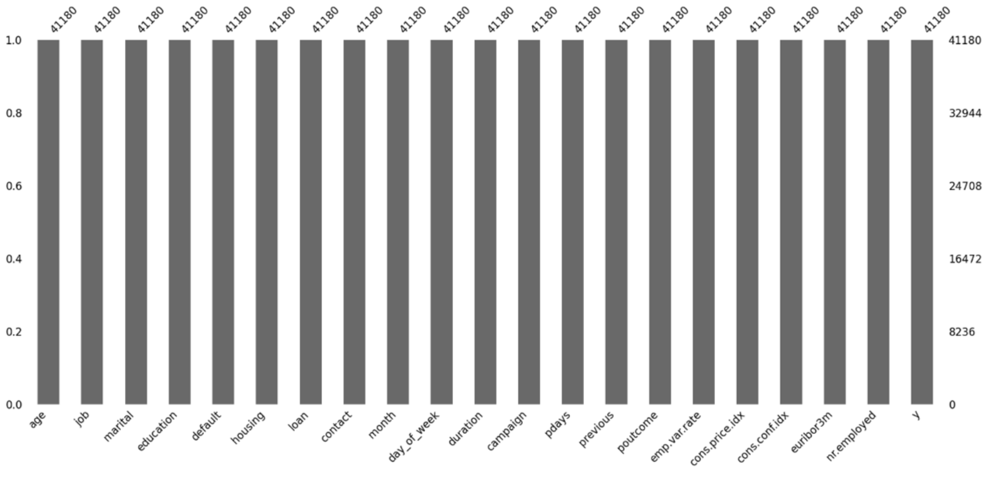
```
RangeIndex: 41180 entries, 0 to 41179
Data columns (total 21 columns):
 #   Column          Non-Null Count  Dtype 
---  ------          --------------  ----- 
 0   age             41180 non-null  object
 1   job             41180 non-null  object
 2   marital         41180 non-null  object
 3   education       41180 non-null  object
 4   default         41180 non-null  object
 5   housing         41180 non-null  object
 6   loan            41180 non-null  object
 7   contact         41180 non-null  object
 8   month           41180 non-null  object
 9   day_of_week     41180 non-null  object
 10  duration        41180 non-null  object
 11  campaign        41180 non-null  object
 12  pdays           41180 non-null  object
 13  previous        41180 non-null  object
 14  poutcome        41180 non-null  object
 15  emp.var.rate    41180 non-null  object
 16  cons.price.idx  41180 non-null  object
 17  cons.conf.idx   41180 non-null  object
 18  euribor3m       41180 non-null  object
 19  nr.employed     41180 non-null  object
 20  y               41180 non-null  object
dtypes: object(21)
memory usage: 6.6+ MB
```

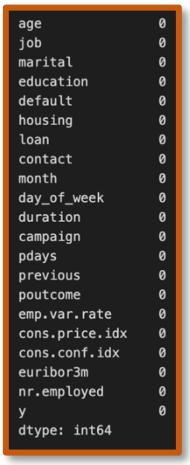


```
duration            int64
campaign            int64
pdays               int64
previous            int64
emp.var.rate      float64
cons.price.idx    float64
cons.conf.idx     float64
euribor3m         float64
nr.employed       float64
```

Figure 1 Transformations of variable data type from object, to int64 and float64.

While 12 duplicate rows were identified, naught were omitted as machine learning would not be an objective. No "NaN" or missing values were identified (Figure 2), nor were any outliers excluded (Figure 6, boxplots i, ii, iv, vii). All implausible "**999**" values in **pdays** were, however, replaced with NaN values, (Figure 6, boxplot iii; Figure 9, appendix).





```
age               0
job               0
marital           0
education         0
default           0
housing           0
loan              0
contact           0
month             0
day_of_week       0
duration          0
campaign          0
pdays             0
previous          0
poutcome          0
emp.var.rate      0
cons.price.idx    0
cons.conf.idx     0
euribor3m         0
nr.employed       0
y                 0
dtype: int64
```

Figure 2 NaN values visualised with Missingno.

Section 3: Exploratory Data Analysis

[3.1] Categorical Variables

All categorical features were visualised and examined. Inspecting the target variable (**y**) revealed only **11.3%** of the cohort were subscribed, this denoted a class imbalance (Figure 3). The data also represents a predominantly early 30s cohort (Figure 4). While it may appear the majority of subscribers reside within this age bracket, the distribution originally surveyed **also** includes a predominately 30s age bracket.

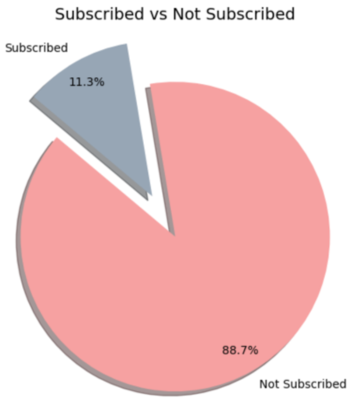


Figure 3 Pie graph visualisation of the target variable "y".

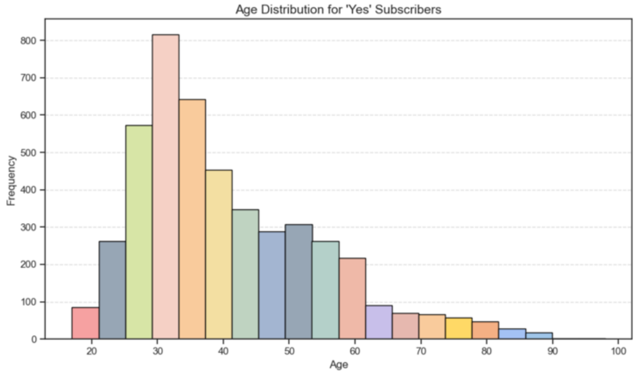


Figure 4 Bar chart illustrating the age distribution of subscribers.

Bar charts were constructed to compare categorical features against individuals who were subscribed and not subscribed. Among several insights, principal takeaways indicated an increased interest from:

- **admin**, **blue-collar** and **services** professions (Figure 5, bar chart ii),
- the **university-educated** (bar chart iv),
- **cellular** users (bar chart viii),
- and higher reception as **Autumn** concludes (bar chart ix).

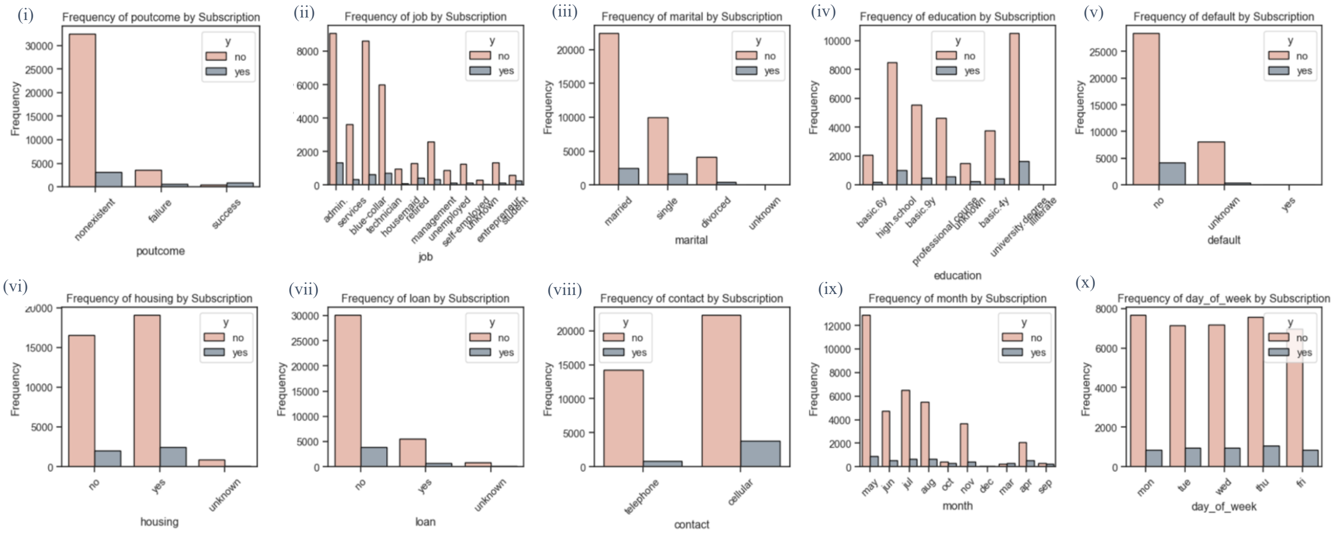


Figure 5 A series of bar charts illustrating the categorical features of subscribers and non-subscribers.

[3.2] Numerical Variables

All numerical features were likewise visualised with boxplots for discrepancies in "**y**". Among interpretations, subscribers were seen to have held:

- *longer mean "contact"* ***durations*** *(Figure 6, boxplot i)*,
- *fewer mean* ***campaign*** *"contact" counts (Figure 6, boxplot ii)*,
- *increased mean* ***previous*** *contact counts (Figure 6, boxplot iv)*.
- *and variation in employment and rate indexes (Figure 6, boxplot viii & ix)*.

It's important to recognise that the class imbalance may influence these differences.

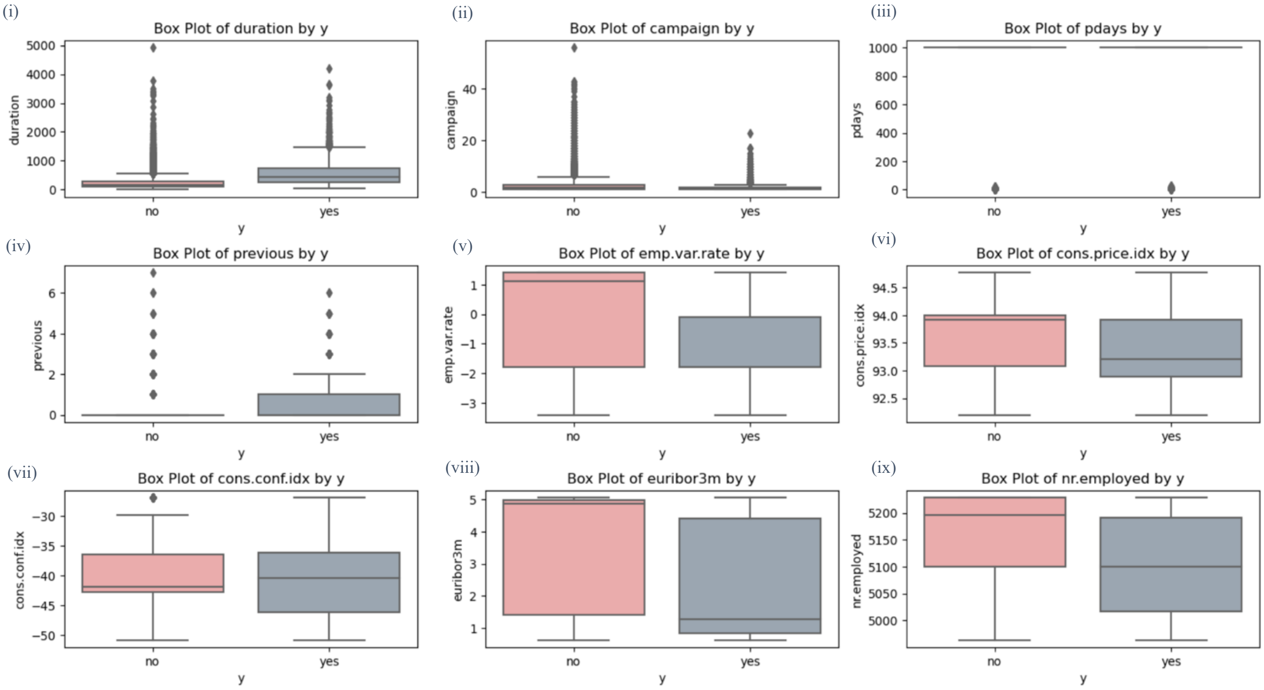


Figure 6 A series of box and whisker plots illustrating the numerical features of subscribers and non-subscribers.

[3.3] Correlations

Heatmaps were employed to assess correlation (**r**) between numerical features. While not always significant, subscribers do lean towards an increased correlation between features. These are apparent through the **duration**, **campaign** and **previous** features, and less so, for the **price**, **rate** and **employed** indexes. (Figure 7).

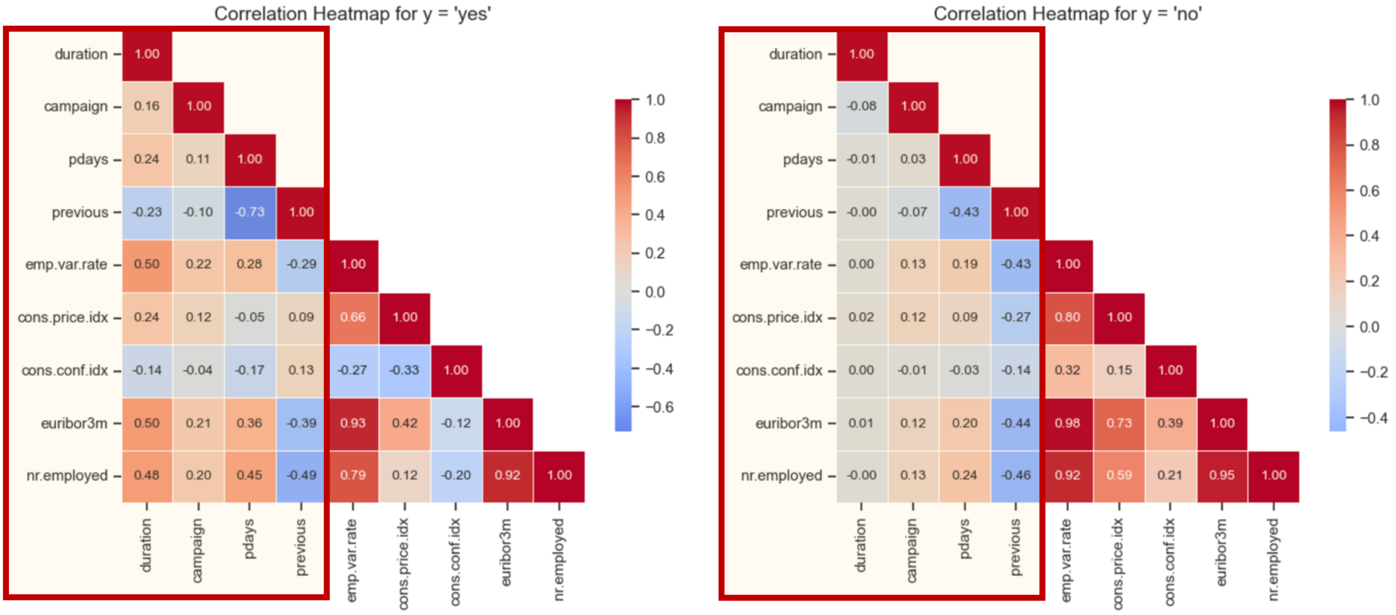


Figure 7 Heatmaps illustrating the correlation between numerical features of subscribers and non-subscribers.

[3.4] Campaign Assessment | Previous Vs. Current

By comparing campaigns, an assessment of the success of each in attaining subscribers may be understood.

> *The 'previous' campaign saw a **24.4%** success in subscriber acquisition. When excluding this count from the 'current' or, otherwise the recently acquired subscribers count, the new campaign is seen to only have attracted **9.4%** of the possible pool.*

As a result of both campaigns, **11.13%** represent the presently subscribed, though this doesn't consider the users who have unsubscribed since the former campaign, (Figures 11 & 12, appendix).

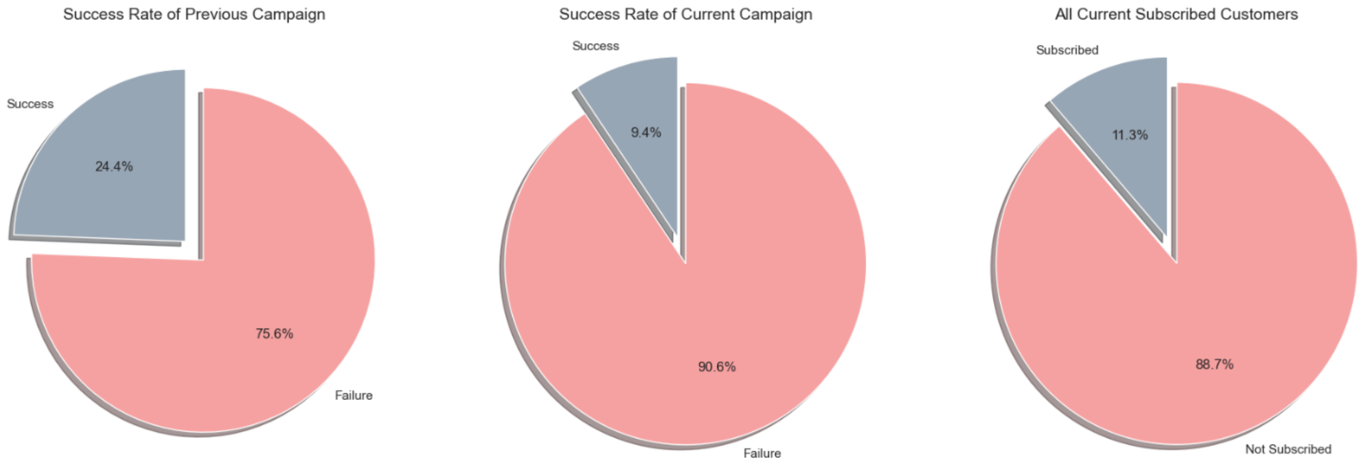
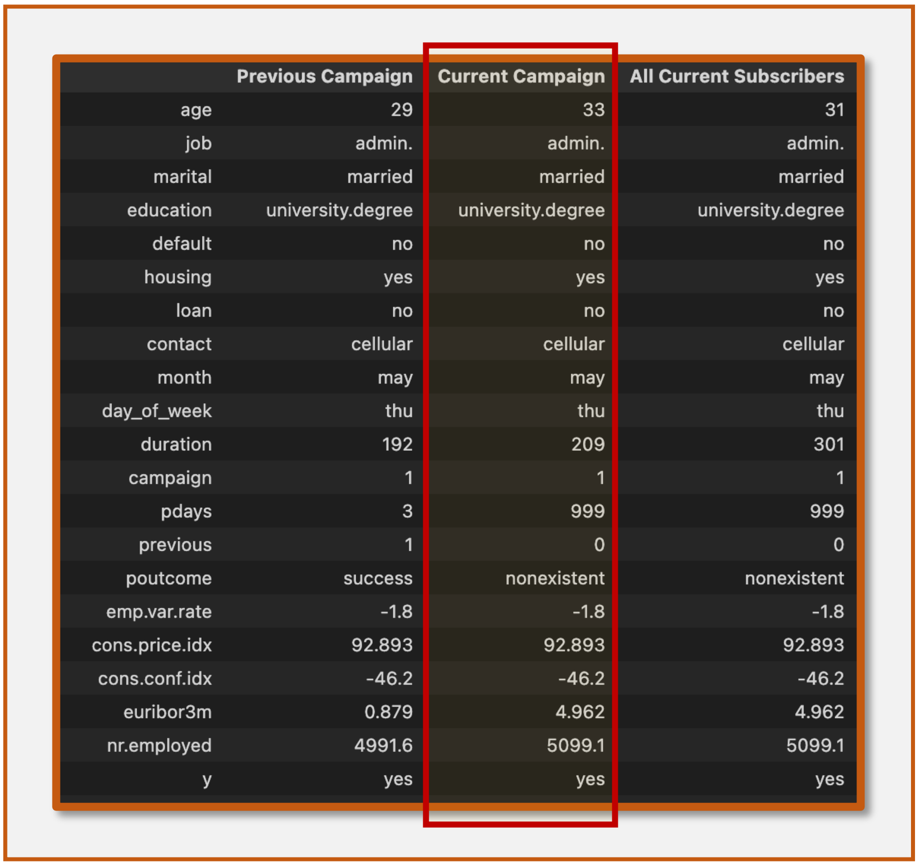


*Figure 8 A series of pie graphs comparing the success rate of the previous campaign, to the current campaign, to all present subscribers. **Further visuals available in the appendix – see Figures 11 & 12.**.*

Section 4: Outcome

4.1 Results

To achieve the business requirements, insights into the most responsive customer segments to the latest campaign were requested. Through an EDA, several qualities of the principal subscriber have been compiled for assessment (Figure 8). A more comprehensive list, including value counts, is available within the document appendix (Figure 13).



	Previous Campaign	Current Campaign	All Current Subscribers
age	29	33	31
job	admin.	admin.	admin.
marital	married	married	married
education	university.degree	university.degree	university.degree
default	no	no	no
housing	yes	yes	yes
loan	no	no	no
contact	cellular	cellular	cellular
month	may	may	may
day_of_week	thu	thu	thu
duration	192	209	301
campaign	1	1	1
pdays	3	999	999
previous	1	0	0
poutcome	success	nonexistent	nonexistent
emp.var.rate	-1.8	-1.8	-1.8
cons.price.idx	92.893	92.893	92.893
cons.conf.idx	-46.2	-46.2	-46.2
euribor3m	0.879	4.962	4.962
nr.employed	4991.6	5099.1	5099.1
y	yes	yes	yes

Figure 8 A series of data frames comparing the traits of subscribers from the previous campaign, subscribers from the current campaign, and all present subscribers. This figure illustrates the most sensitive to the new campaign is a ***married, early 30s and university educated individual, who has access to housing*, primarily uses cellular and has had at least one contact during the campaign****. This individual is most receptive on a Thursday, at the conclusion of Autumn, and was not exposed to the previous campaign. At this stage, index ranges (bottom 5 variables) don't express immediate significant influences on subscription rates.*

[4.2] Conclusion

Following analysis, it is understood that business requirements have been delivered and downstream evaluation may be actioned. The data is considered to be suitable for modelling following the encoding of categorical variables (Figure 1).

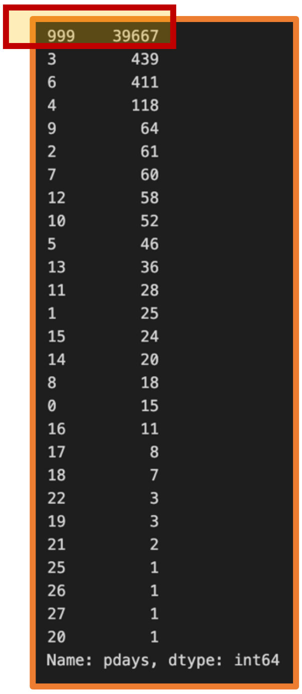
Modelling is advised prior to conducting upcoming campaigns.

Appendix

Table 1 The provided data dictionary by the Telcom client.

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? [Feature of interest]*

Accompanying Visualisations



```
999    39667
3        439
6        411
4        118
9         64
2         61
7         60
12        58
10        52
5         46
13        36
11        28
1         25
15        24
14        20
8         18
0         15
16        11
17         8
18         7
22         3
19         3
21         2
25         1
26         1
27         1
20         1
Name: pdays, dtype: int64
```

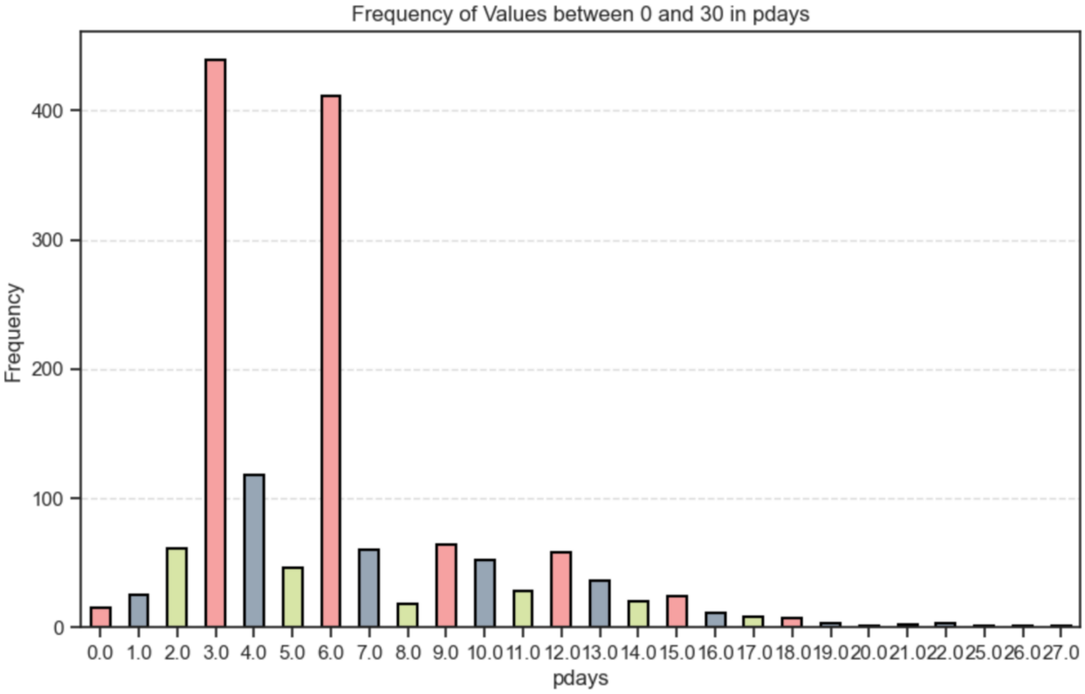


Figure 9 Following transformation of (39667x) "999" values in pdays, the average frequency of days passing after contact with the client from the previous campaign, is visualised.

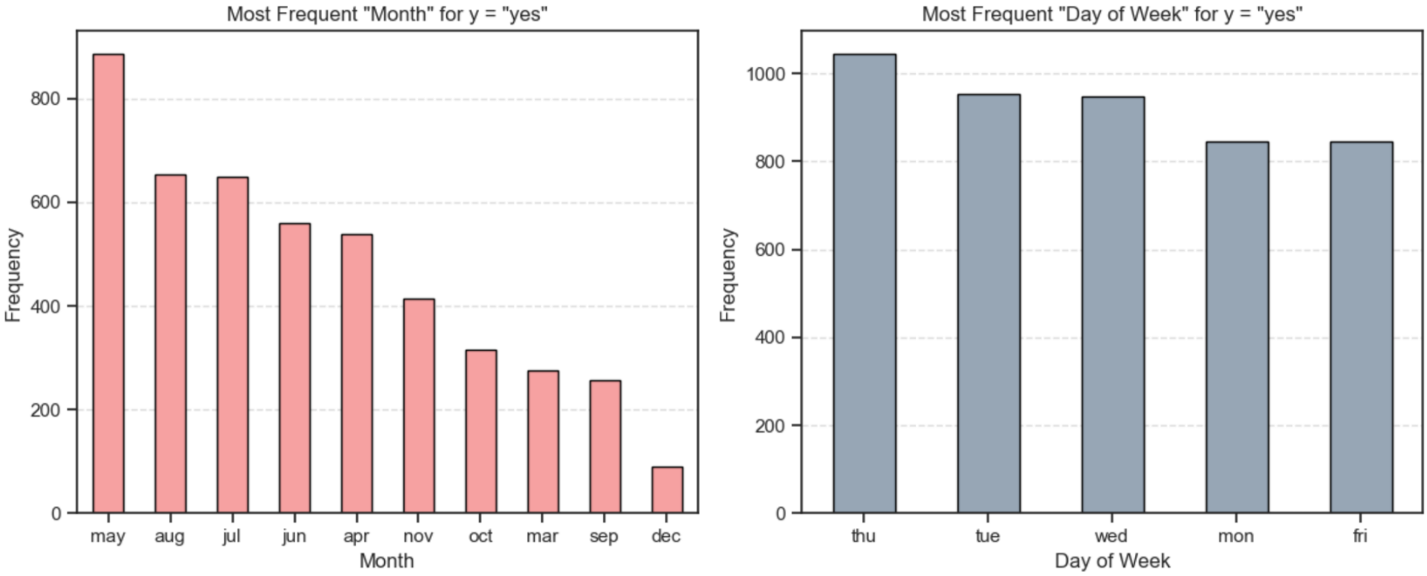
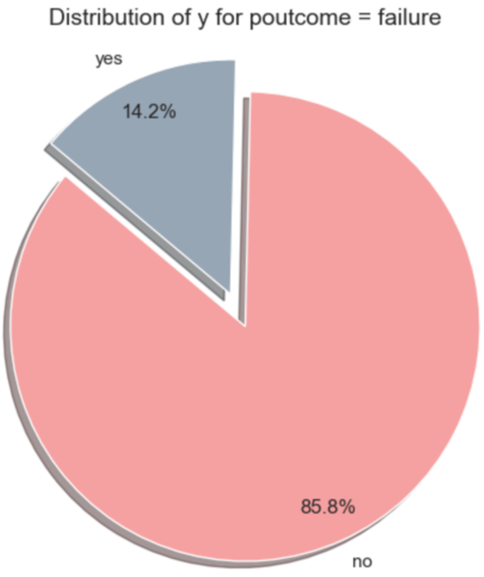


Figure 10 Visualisations for Month and Day frequencies. Charts illustrate when it's best for contacts to achieve higher reception for subscriptions.



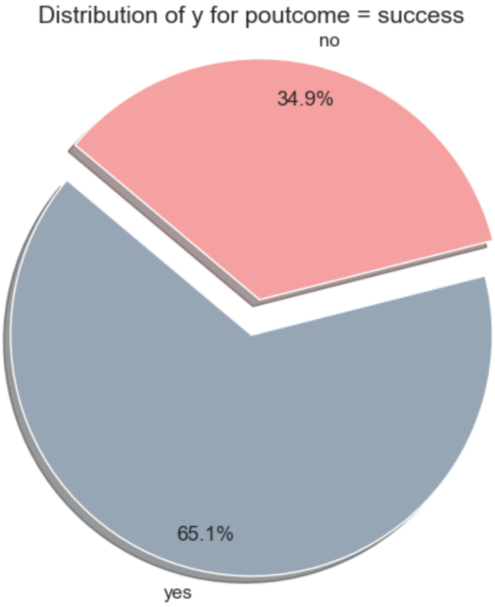
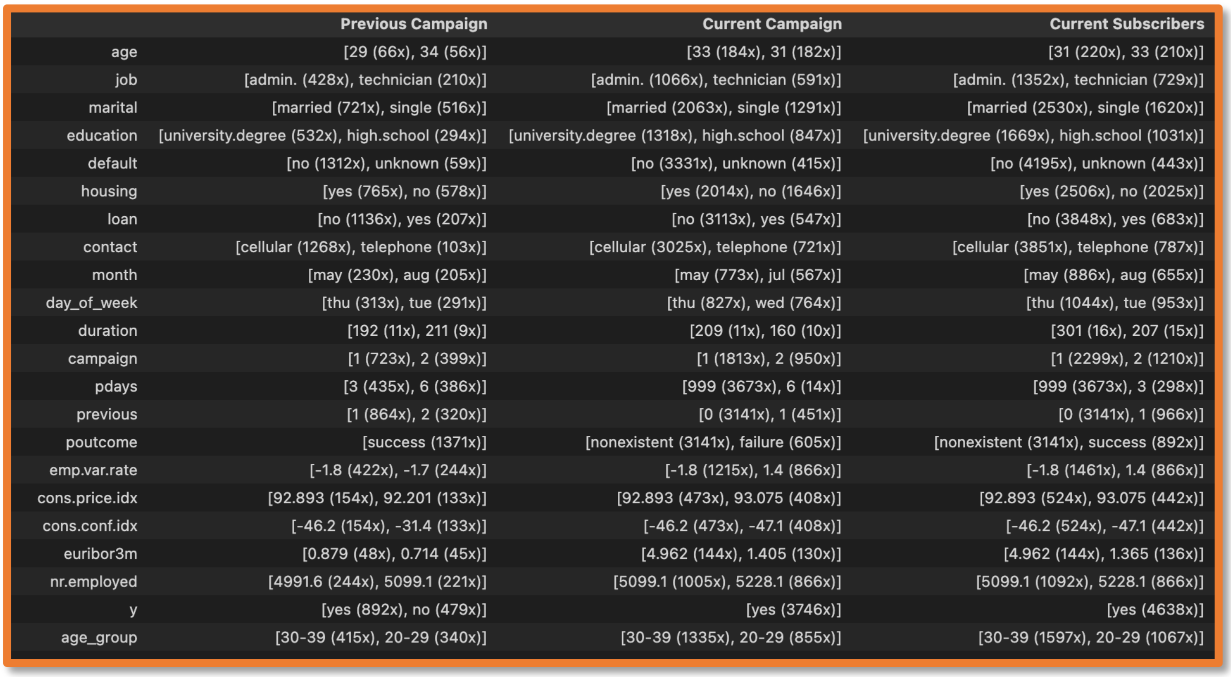


Figure 11 A pie chart illustrating the customer percentages where the previous campaign failed, and the current was successful.

Figure 12 A pie chart illustrating the customer percentages who were subscribed as a result of the previous campaign and have since unsubscribed.



	Previous Campaign	Current Campaign	Current Subscribers
age	[29 (66x), 34 (56x)]	[33 (184x), 31 (182x)]	[31 (220x), 33 (210x)]
job	[admin. (428x), technician (210x)]	[admin. (1066x), technician (591x)]	[admin. (1352x), technician (729x)]
marital	[married (721x), single (516x)]	[married (2063x), single (1291x)]	[married (2530x), single (1620x)]
education	[university.degree (532x), high.school (294x)]	[university.degree (1318x), high.school (847x)]	[university.degree (1669x), high.school (1031x)]
default	[no (1312x), unknown (59x)]	[no (3331x), unknown (415x)]	[no (4195x), unknown (443x)]
housing	[yes (765x), no (578x)]	[yes (2014x), no (1646x)]	[yes (2506x), no (2025x)]
loan	[no (1136x), yes (207x)]	[no (3113x), yes (547x)]	[no (3848x), yes (683x)]
contact	[cellular (1268x), telephone (103x)]	[cellular (3025x), telephone (721x)]	[cellular (3851x), telephone (787x)]
month	[may (230x), aug (205x)]	[may (773x), jul (567x)]	[may (886x), aug (655x)]
day_of_week	[thu (313x), tue (291x)]	[thu (827x), wed (764x)]	[thu (1044x), tue (953x)]
duration	[192 (11x), 211 (9x)]	[209 (11x), 160 (10x)]	[301 (16x), 207 (15x)]
campaign	[1 (723x), 2 (399x)]	[1 (1813x), 2 (950x)]	[1 (2299x), 2 (1210x)]
pdays	[3 (435x), 6 (386x)]	[999 (3673x), 6 (14x)]	[999 (3673x), 3 (298x)]
previous	[1 (864x), 2 (320x)]	[0 (3141x), 1 (451x)]	[0 (3141x), 1 (966x)]
poutcome	[success (1371x)]	[nonexistent (3141x), failure (605x)]	[nonexistent (3141x), success (892x)]
emp.var.rate	[-1.8 (422x), -1.7 (244x)]	[-1.8 (1215x), 1.4 (866x)]	[-1.8 (1461x), 1.4 (866x)]
cons.price.idx	[92.893 (154x), 92.201 (133x)]	[92.893 (473x), 93.075 (408x)]	[92.893 (524x), 93.075 (442x)]
cons.conf.idx	[-46.2 (154x), -31.4 (133x)]	[-46.2 (473x), -47.1 (408x)]	[-46.2 (524x), -47.1 (442x)]
euribor3m	[0.879 (48x), 0.714 (45x)]	[4.962 (144x), 1.405 (130x)]	[4.962 (144x), 1.365 (136x)]
nr.employed	[4991.6 (244x), 5099.1 (221x)]	[5099.1 (1005x), 5228.1 (866x)]	[5099.1 (1092x), 5228.1 (866x)]
y	[yes (892x), no (479x)]	[yes (3746x)]	[yes (4638x)]
age_group	[30-39 (415x), 20-29 (340x)]	[30-39 (1335x), 20-29 (855x)]	[30-39 (1597x), 20-29 (1067x)]

Figure 13 A Comprehensive set of data frames, listing the most common two traits of subscribed customers, from each campaign.