# PRELIMINARY INSIGHTS – TELECOM DATASET

*AT1: Exploratory Data Analysis*

36103 – Statistical Thinking for Data Science

University of Technology Sydney

Thanh Nhan Le – 11919925

27/08/2023

## Problem formulation

This report aims to give preliminary insights into a marketing campaign dataset via exploratory data analysis (EDA). The goal of this analysis is to gain a current understanding of

(1) Any data quality issues, such as missing values and errors,
(2) The general structure of the dataset, such as variables' distributions, relationship between independent variables and the response variable, or amongst independent variables.

The answers to these questions will aid with next steps in the modelling process, such as preprocessing and applying predictive models, which will ultimately provide the business with insights into the customer segments that show the most excitement towards the new telecom subscription plan.

## Data cleaning and processing

The dataset has 41,180 rows and 21 columns, and there are no missing values upon initial checking. The raw file is poorly delimited, in that multiple delimiters are found in both column names and values:

| Single semicolon | Semicolon in double-quotes | Semicolon in double double-quotes | Semicolon with half double-quotes |
|---|---|---|---|
| ; | “;” | ““;”” | ;” |

The raw file was parsed by including all these delimiters. Afterwards, some minor clean-up was performed on individual columns' names and values (`age`, `job`, `day_of_week`, `poutcome`, and `y`) to remove the extra double quotes that remained.

Upon a second check, it was revealed that the `pdays` column contained '999', which was usually to indicate a missing value. This was undetected by the first check using `data.isna().any()`, but was revealed using descriptive statistics. These were imputed via the random hot deck technique (Andridge & Little, 2010), using values in the `previous` column as reference:

- If `previous = 0`: `pdays = -1`, because there was no previous contact
- If `previous != 0`: sample `pdays` randomly from the empirical conditional probability mass function

$$p(pdays \mid previous = i), for\ i \in \{1, 2, \dots, 7\}$$

# Exploratory data analysis (EDA)

## Descriptive statistics

Using descriptive statistics, I noticed that the data has no errors, and that the '999' value in the `pdays` column denotes a missing value. The data has no other missing values elsewhere.

| | 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 1*

| | job | marital | education | default | housing | loan | contact | month | day_of_week | poutcome | y |
|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 41180 | 41180 | 41180 | 41180 | 41180 | 41180 | 41180 | 41180 | 41180 | 41180 | 41180 |
| unique | 12 | 4 | 8 | 3 | 3 | 3 | 2 | 10 | 5 | 3 | 2 |
| top | admin. | married | university.degree | no | yes | no | cellular | may | thu | nonexistent | no |
| freq | 10422 | 24921 | 12166 | 32581 | 21571 | 33943 | 26140 | 13765 | 8622 | 35559 | 36542 |

*Figure 2*

The imputation technique described in part II preserved the distribution of the data, with the mean, variance and shape of the distribution remaining largely the same.

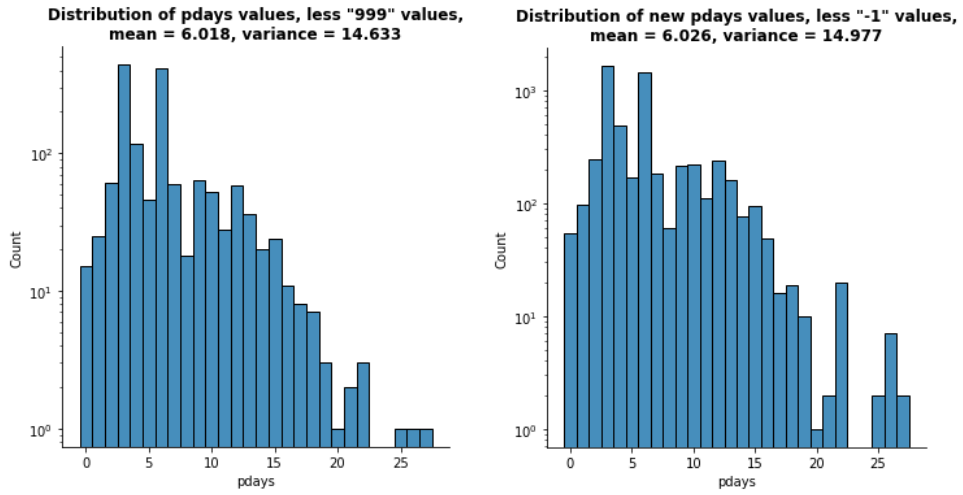


*Figure 3*

## Distribution plots

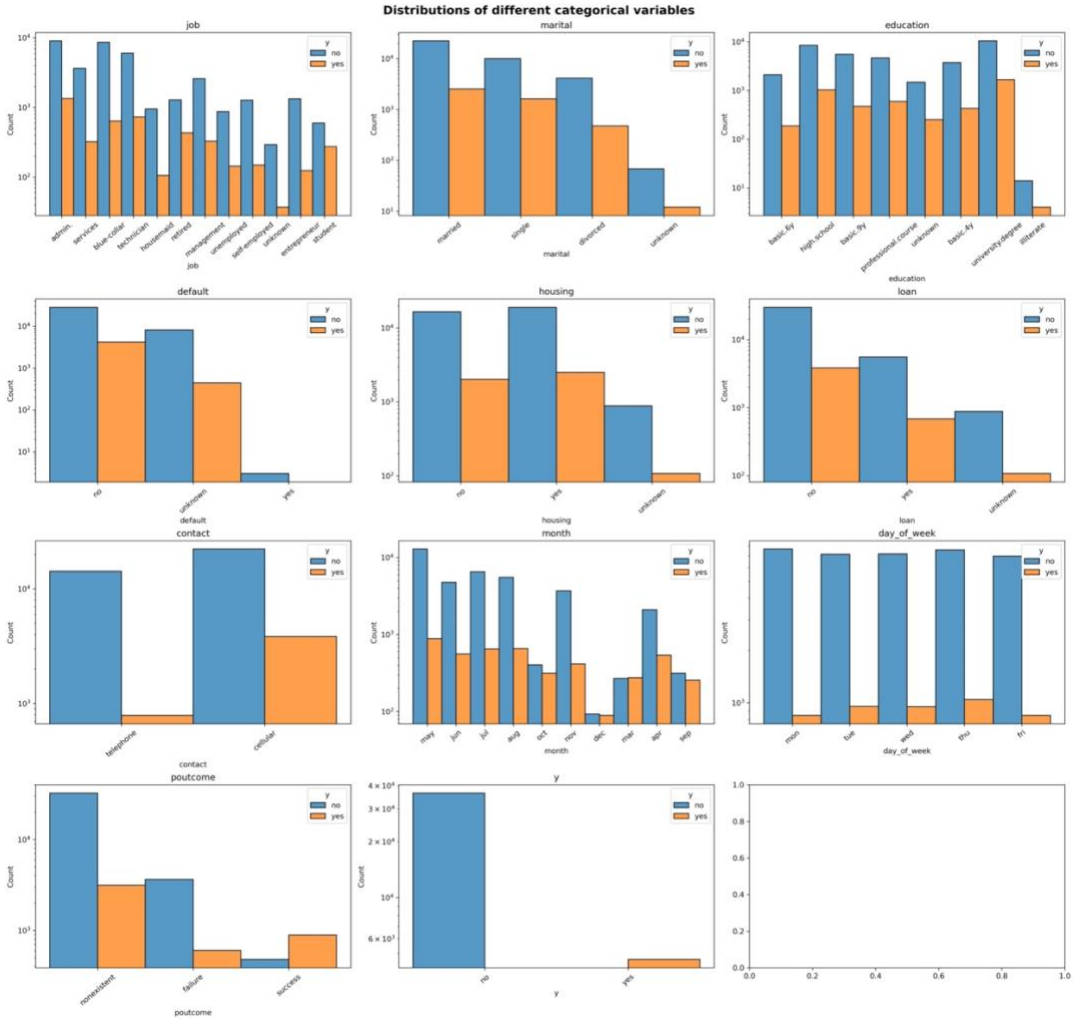


*Figure 4*

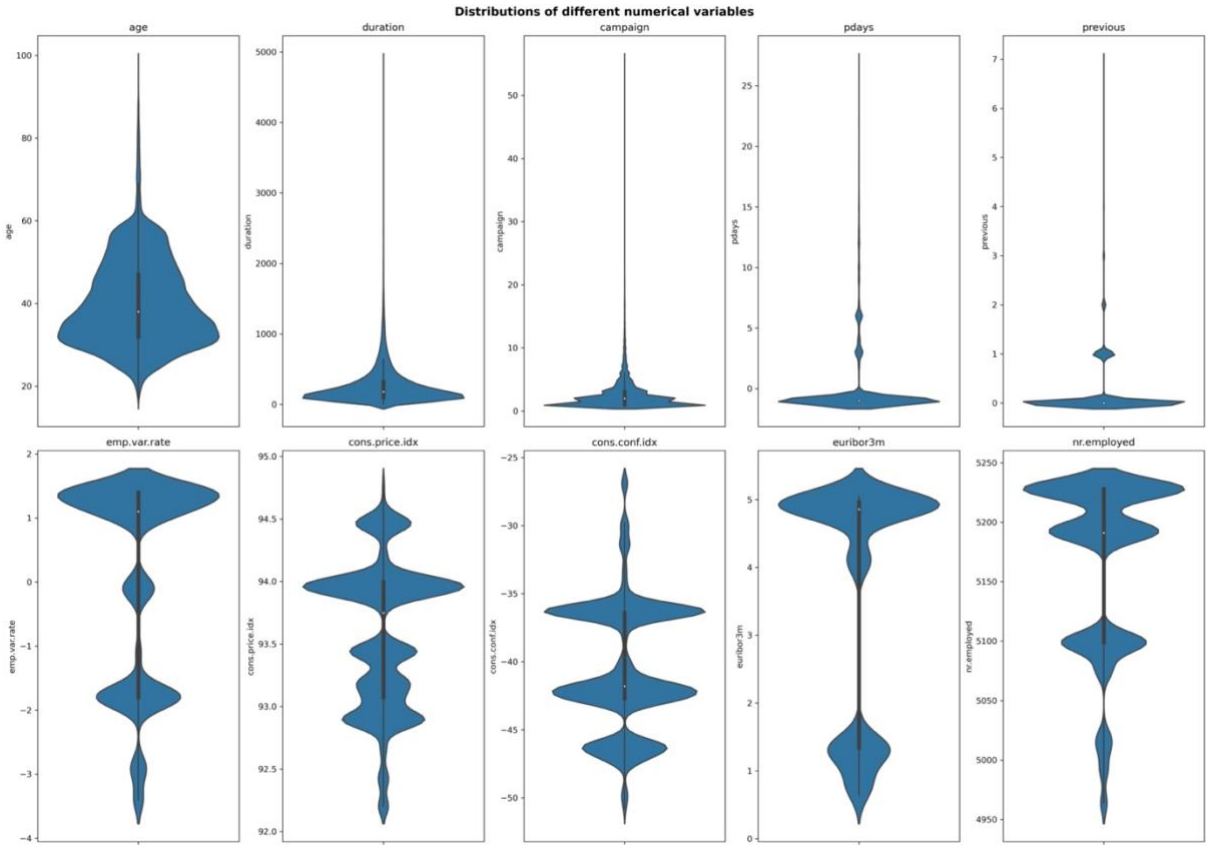


*Figure 5*

Per figure 4, the segments showing highest potentials to the new product are revealed. It's also an extension of figure 2, where complete distributions of all categorical variables are shown. Top subscriber groups hold admin jobs, are married, have a university degree, a housing loan and no personal loans, have no credit in default, and did not participate in any previous campaign.

Figure 5 is the numerical counterpart. Most numerical variables do not follow any standard distribution, and the 5 distributions on the top row are highly skewed.

## Bivariate relationship plots

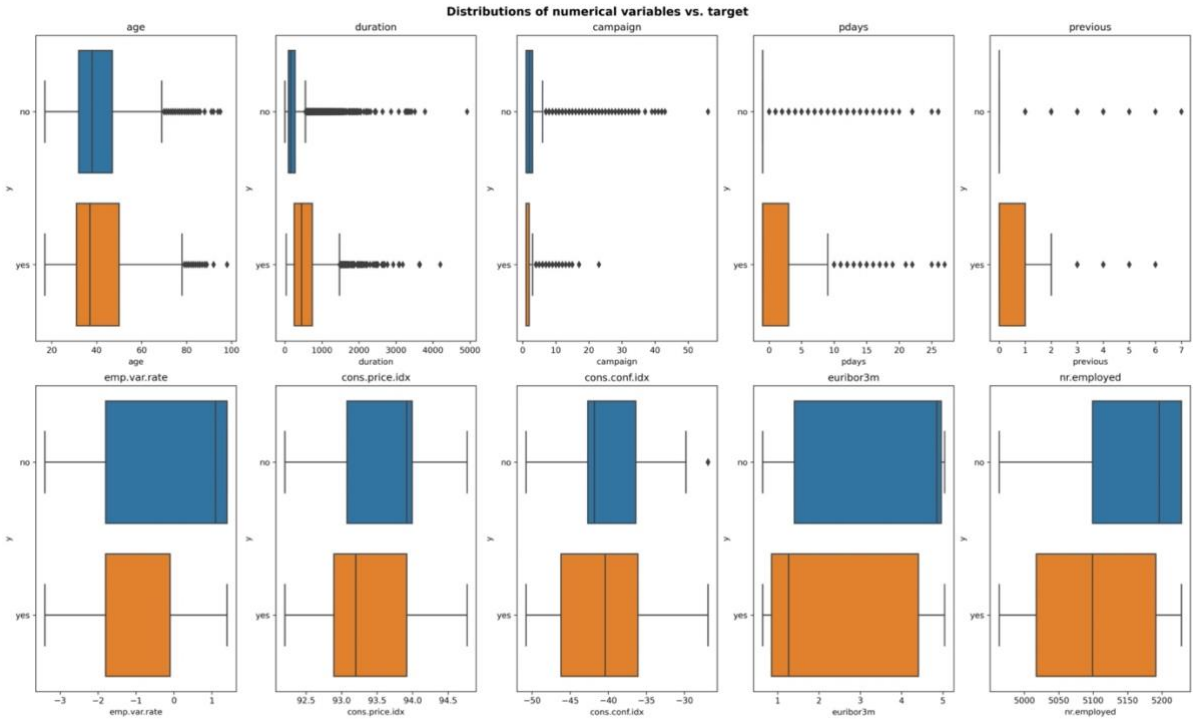


*Figure 6*

Figure 6 shows the variables that may be able to predict the outcome 'y' of the campaign: `duration`, `pdays`, `previous`, `emp.var.rate`, and `nr.employed`, for their wide discrepancies between outcomes. These can fit a sigmoid curve fairly well, so logistic regression may be used for modelling.

## Dependence measures

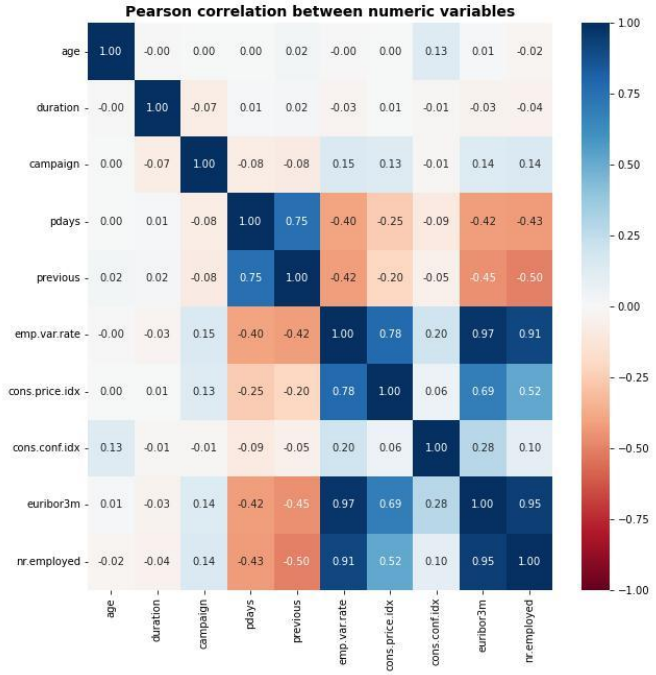


*Figure 7*

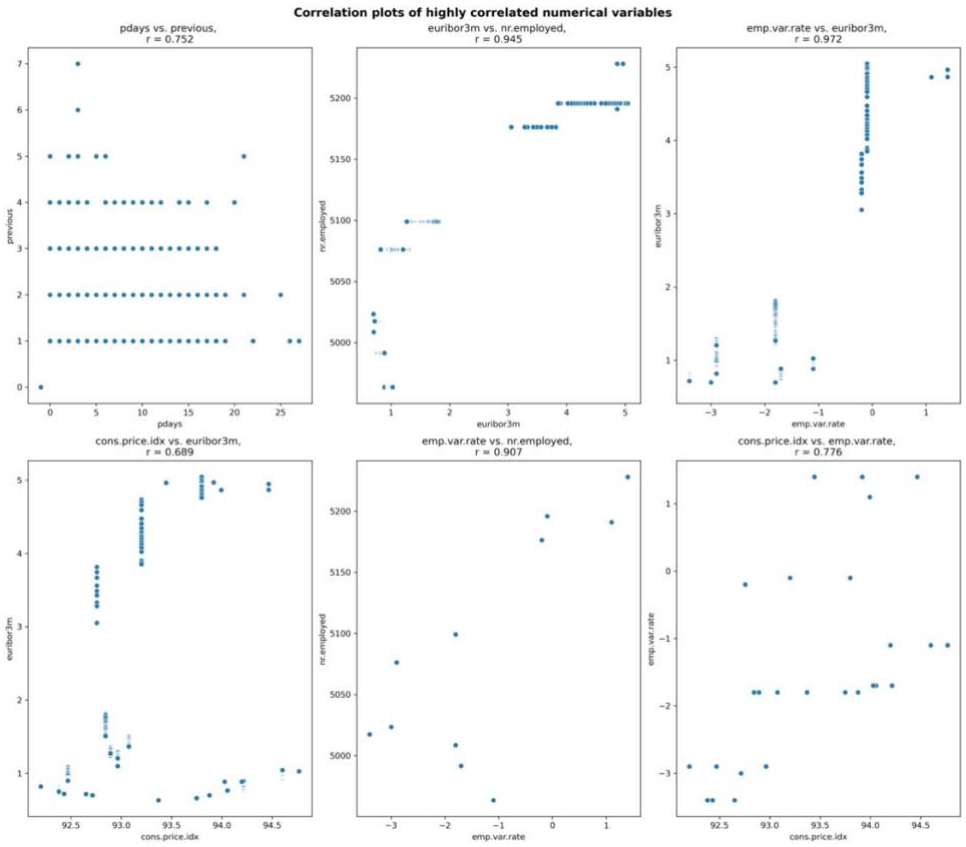


*Figure 8*

Some off-diagonal pairs have very high correlation coefficients (figure 7), but not all show clear linearity (figure 8), e.g. pdays vs. previous. Other pairs in figure 8 generally exhibit a linear relationship with some noise.

## Conclusion

The bivariate plots suggest some linear relationship between explanatory variables and response variable, so a generalised linear model, e.g. logistic regression, may be a good starting point for modelling. Correlation analysis suggests eliminating highly linear explanatory variables to reduce multicollinearity.

We have also seen from categorical distributions the customer segments that are the most responsive to this marketing campaign.

# References

Andridge, R. R., & Little, R. J. (2010). A Review of Hot Deck Imputation for Survey Non-response. *International statistical review = Revue internationale de statistique*, 78(1), 40–64. https://doi.org/10.1111/j.1751-5823.2010.00103.x

# Appendix 1: Figure names

*Figure 1: Descriptive statistics for numerical data.*

*Figure 2: Descriptive statistics for categorical data.*

*Figure 3: Distribution of pdays values, less '999' (left) and new pdays (original + imputed) values, less '-1' (right), with means and variances noted.*

*Figure 4: Distribution of different categorical variables, with counts split into 'yes' and 'no' in terms of the response variable.*

*Figure 5: Distribution of different numerical variables.*

*Figure 6: Bivariate relationships between numerical variables and the response variable.*

*Figure 7: Correlation matrix between explanatory variables.*

*Figure 8: Bivariate correlation plots for the most correlated variable pairs (above 0.6).*