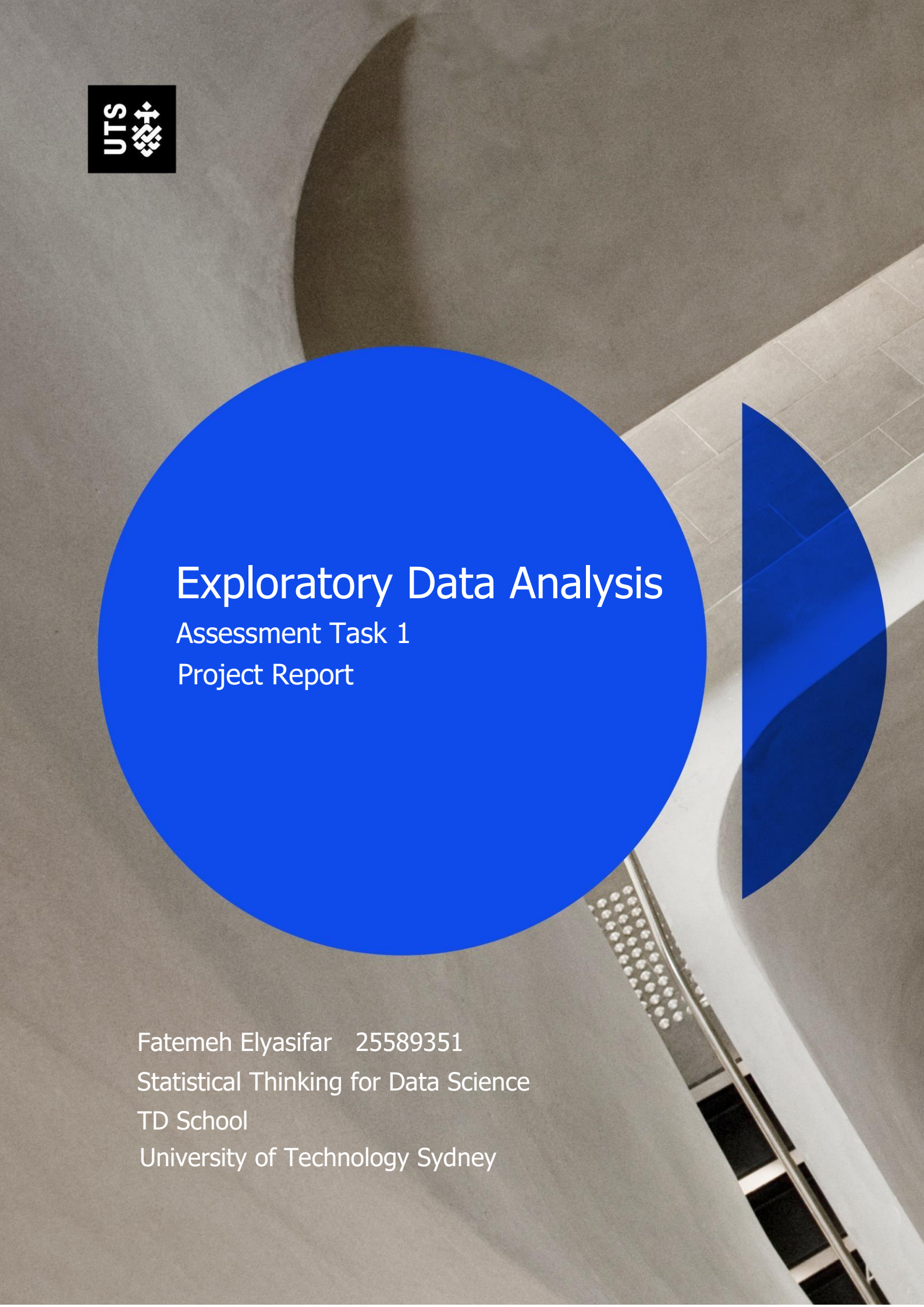

Exploratory Data Analysis

Assessment Task 1

Project Report

Fatemeh Elyasifar 25589351
Statistical Thinking for Data Science
TD School
University of Technology Sydney

Table of Contents

Summary and Introduction

Summary

This project analyses a marketing campaign implemented by a telecommunication company to promote a new subscription plan. The company aims to identify customer segments that are most responsive to the campaign.

To address this, the project uses exploratory data analysis (EDA) to understand the relationship between various customer behaviors and the possibility of subscribing to the new plan. Through EDA, the project provides useful insights by analysing patterns within data.

Key findings reveal that most customers tend not to subscribe to the new plan; however, customers around the ages of 60 and 20 displayed a higher subscription rate compared to other age groups. Additionally, customers who were engaged in previous campaigns are more likely to subscribe to the new campaign. This shows the importance of customer history.

By understanding and targeting the right customer segments, the company can significantly enhance the effectiveness of its marketing efforts. These insights will enable the telecommunication company to improve customer acquisition.

Introduction

Problem Statement

The telecommunication company launched a marketing campaign to promote a new subscription plan to attract customers. However, understanding customer behaviour and identifying which segments are most responsive to the campaign is a complex challenge. Without clear insights, the company risks investing resources into ineffective marketing strategies.

Rationale

Effective marketing requires a deep understanding of customer preferences and behaviours. By analysing past campaign data, companies can refine their targeting strategies and improve customer engagement. This exploratory data analysis (EDA) is critical for uncovering patterns and insights that are not immediately apparent so the company can increase its subscription rates.

Project Aims and Objectives

The main aim of this project is to conduct an exploratory data analysis (EDA) of the marketing campaign dataset to identify key customer responsiveness to the campaign. The specific objectives of research are as follows:

1. Which age groups are most likely to subscribe to the new campaign?
2. What is the relationship between customers' occupations and the likelihood of subscription?
3. Does higher education lead to an increase in subscription rate?
4. How does the duration of last contact affect the subscription rate?

This analysis indicates that most customers are unlikely to subscribe to the new campaign which suggests that the company needs to revise its strategies.

In summary, this analysis will equip the telecommunication company with the knowledge needed to enhance the effectiveness of their marketing efforts and improve customer acquisition.

Methodology

Methods Overview

To conduct the analysis, the dataset was first uploaded to Google Collab, an online platform that facilitates data processing and analysis.

The initial step involved addressing missing values and detecting data errors. These were systematically identified and replaced with appropriate values to ensure the integrity and completeness of the dataset.

Following data cleaning, various visualisations, such as charts and graphs, were generated to explore and analyse the data. These visual tools provided insights into customer behaviour and helped identify trends related to the likelihood of subscription to the new campaign.

Methods Details

Data Acquisition

The dataset contains different types of data, including numerical and categorical data, as follows:

- Age, campaign (discrete data)
- Cons.price.idx, emp.var.rate (continuous data)
- Education (ordinal data)
- Marital (nominal data)

Relevant libraries, including Pandas and NumPy, were imported to facilitate data processing. The dataset, stored in a CSV (Comma-Separated Values) format, was then loaded, with each entry separated by a comma within the rows. Subsequently, the data was converted into a data frame, a structured format that facilitates further analysis.

Data Preprocessing

The initial step involved an overview of the dataset's descriptive statistics, including the mean, minimum, maximum, and count for each column. Following this, the dataset was checked for duplicate rows, and 12 duplicates were identified. Eleven of these were removed as they were redundant.
The next phase involved checking for invalid data entries. While no null values were found, some columns contained "unknown" entries, specifically in the marital status, job, education, default, housing, and loan columns. Additionally, 'nonexistent' values were identified in the 'poutcome' column. These invalid values were then replaced.
For the marital status column, all "unknown" entries were replaced with the mode of the column due to their limited number. A similar approach was applied to the job, housing, and loan columns. For the education column, values were assigned by age: 'high-school' for those under 18, 'university.degree' for those aged 18 to 25, and 'professional.course' for everyone else.
Due to the large number of unknown values in the default column (over 8,000) and the significant number of 'nonexistent' values in the 'poutcome' column (over 35,000), these entries were considered as a new category. This approach was taken to avoid potential bias or inaccuracies in subsequent analyses because altering these values without additional data could lead to incorrect conclusions.

Exploratory Analysis

Most contacts during this campaign were performed on Monday. Also, most contacts were performed from May to July. This reveals customers' preferences to subscribe the campaign during these months.

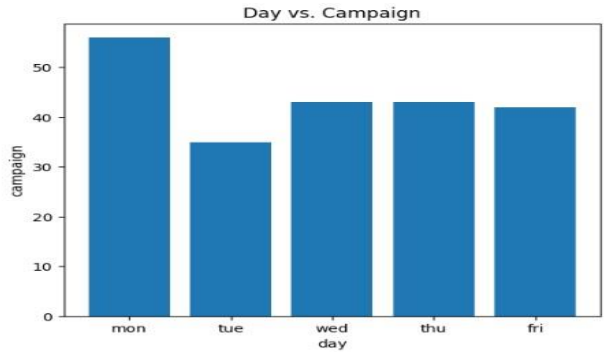

Figure 1. Day vs. Campaign

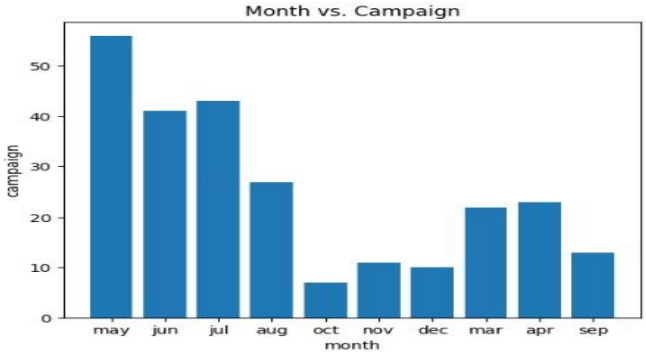

Figure 2. Month vs. Campaign

Based on the descriptive statistics, the majority of customers were between the ages of 30 and 40 and the most frequent level of education observed was a university degree.

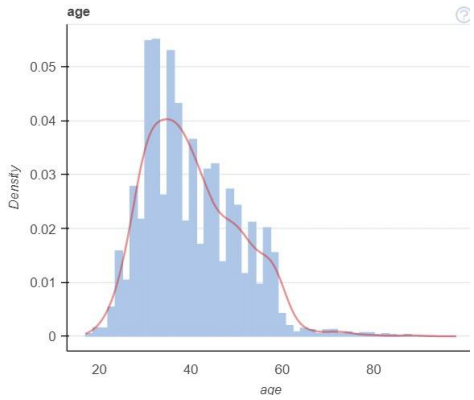

Figure 3. Age Distribution

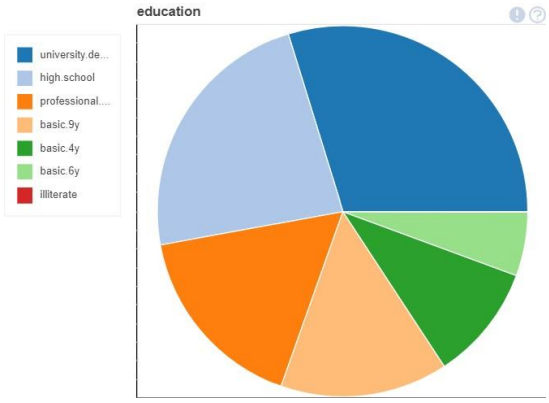

Figure 4. Education Distribution

Both Figure 5 and Table 1 indicate that students and retirees were more likely to subscribe to the campaign, as further supported by Figure 8 and Table 2 (Appendix A). After excluding outliers, specifically customers over the age of 80, the data revealed that individuals around the ages of 60 and 20 were more likely to respond positively to the campaign. Removing the outliers is the best approach due to the limited number of customers in this age group; conclusions based on a single 98-year-old participant would not be reliable.

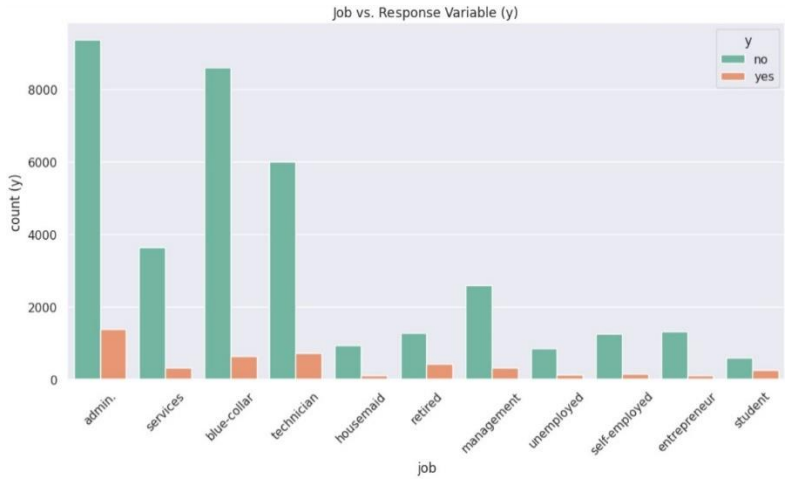

Figure 5. Job vs. Response Variable

y	no	yes
job		
student	68.571429	31.428571
retired	74.766900	25.233100
unemployed	85.798817	14.201183
admin.	87.087171	12.912829
management	88.778652	11.221348
technician	89.180766	10.819234
self-employed	89.514426	10.485574
housemaid	89.990557	10.009443
entrepreneur	91.483516	8.516484
services	91.853720	8.146280
blue-collar	93.104194	6.895806

Table 1. Distribution of Positive and Negative Responses Across Different Job Categories

Additionally, the analysis shows a positive correlation between the duration of the last contact and the likelihood of subscribing to the campaign, with most of these contacts lasting over 500 seconds.

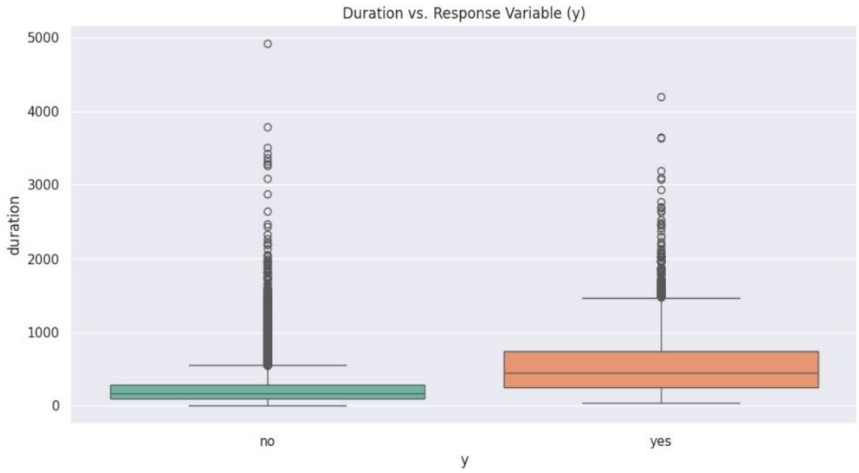

Figure 6. Duration of Calls vs. Response Variable

According to Figure 7, retired customers between the ages of 50 and 80 are the best candidates to respond positively to the campaign. There is an equal possibility of responding positively or negatively among other age groups with different jobs.

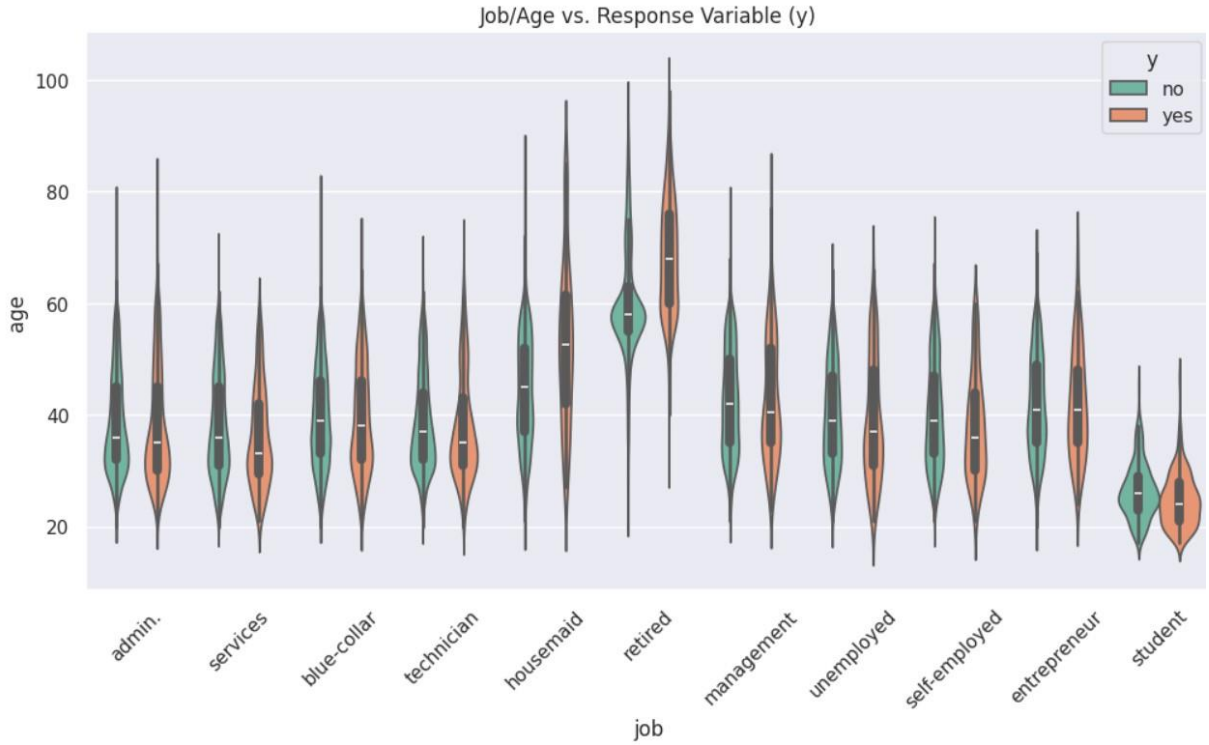

Figure 7. Distribution of Positive and Negative Responses Across Different Job Categories and Age groups

Finally, categorical columns were encoded as numerical values to facilitate the display of the correlation matrix and to more effectively identify the relationships between variables.

Results

Key Findings

- The age group 60 and 20 are more likely to subscribe to the campaign.
- The higher the last contact duration, the higher the possibility of subscription.

In-depth Results

1. Which age groups are most likely to subscribe to the new campaign?
   Clients around the ages of 60 and 20 are more likely to respond positively to the campaign (Appendix C, Figure C1).

2. What is the relationship between customers' occupations and the likelihood of subscription?
   There is almost no correlation between clients' occupations and the likelihood of subscription (Appendix C, Figure C1); however, retired clients are more likely to respond positively to the campaign.

3. Does higher education lead to an increase in subscription rate?
   Although there is almost no correlation between the education and the subscription rate, most clients with university degree tend to subscribe to the campaign (Appendix B, Figure B1)

4. How does the duration of last contact affect the subscription rate?
   There is a slight positive correlation between the duration of last contact and the subscription rate (Appendix C, Figure C1). The higher the last contact duration, the higher the possibility of subscription (Figure 6).

5. What is the relationship between the number of contacts before this campaign and this number during this campaign?
   There is a slight negative relationship suggesting that the number of contacts before this campaign negatively impacts the number of contacts during this campaign (Appendix C, Figure C1). Specifically, as the number of pre-campaign contacts increases, the number of contacts during the campaign decreases.

6. How does the outcome of the previous campaign affect the subscription rate?
   There is a slight negative relationship between the outcome of the previous campaign and the subscription rate.

7. What is the relationship between the number of days that passed by after the client was last contacted and the subscription rate?
   There is a negative relationship between the number of days that passed by after the client was last contacted from a previous campaign and the likelihood of subscription (Appendix C, Figure C1) suggesting the more the days past after the last call, the more the clients tend to subscribe to the new campaign.

Conclusion and References

Conclusion

Based on the analysis of the marketing campaign data, several key insights emerge regarding client behavior and how effective the campaign is.
According to age group analysis, clients around the ages of 60 and 20 appear to be the most responsive targets. Even though occupation doesn't show a strong link to whether people subscribe, retired clients may play a significant role in customising the campaign. Also, interacting with clients for a longer duration during previous contacts may lead to a more positive response to the campaign. On the other hand, too much contact beforehand might lead to a more negative response to the campaign. Moreover, how previous campaigns went can affect current subscription rates. Clients who were contacted recently are less likely to subscribe, suggesting that the company might need to improve its contacts strategies.
In summary, focusing on specific age groups and retired clients and optimising contact strategies could enhance the effectiveness of future campaigns. Modifying the approach based on these insights may lead to higher subscription rates and greater market success.

References

Wickham, H. (2014). Tidy data. Journal of Statistical Software, Articles, 59(10), 1–23. https://doi.org/10.18637/jss.v059.i10

Tukey, J. W. (1977). Exploratory data analysis. Addison-Wesley.

Alice Dong. (2024). TeleCom_Data_1.csv [Dataset]. The real-world dataset provided by the professor.

Appendices

Appendix A: Distribution of Positive and Negative Responses Across Different Age Groups

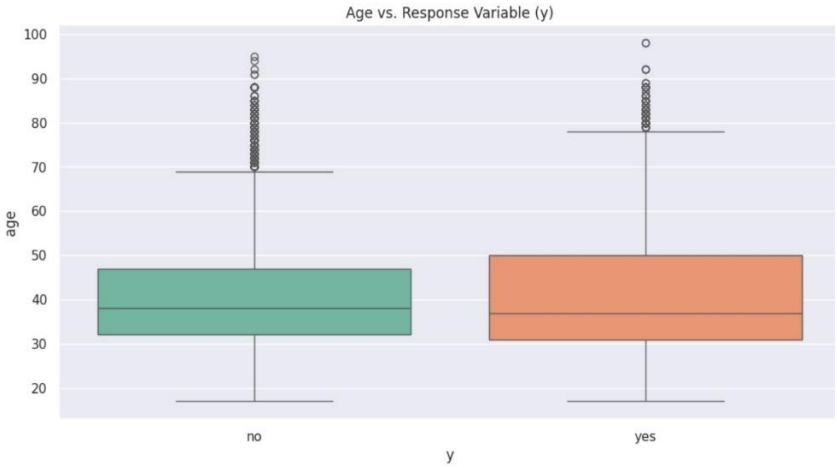

Figure A1. Age vs. Response Variable

y	no	yes
age		
65	47.727273	52.272727
66	49.122807	50.877193
19	52.380952	47.619048
64	52.631579	47.368421
68	54.545455	45.454545
61	56.164384	43.835616
18	57.142857	42.857143
67	57.692308	42.307692
69	58.823529	41.176471
17	60.000000	40.000000

Table A1. Distribution of Positive and Negative Responses Across Different Age Groups

Appendix B: Relationship Between Education and Subscription Likelihood

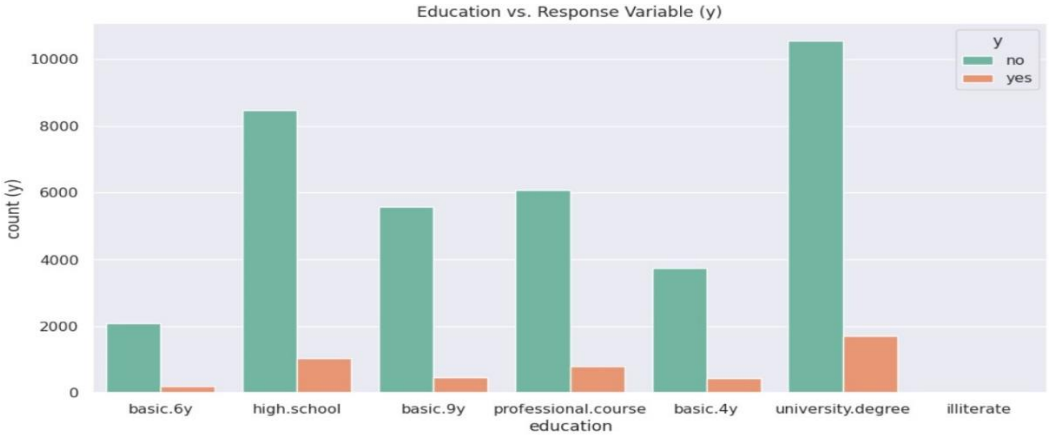

Figure B1. Distribution of Positive and Negative Responses Across Different Education Levels

Appendix C: Relationship Between Input Variables and Subscription Likelihood

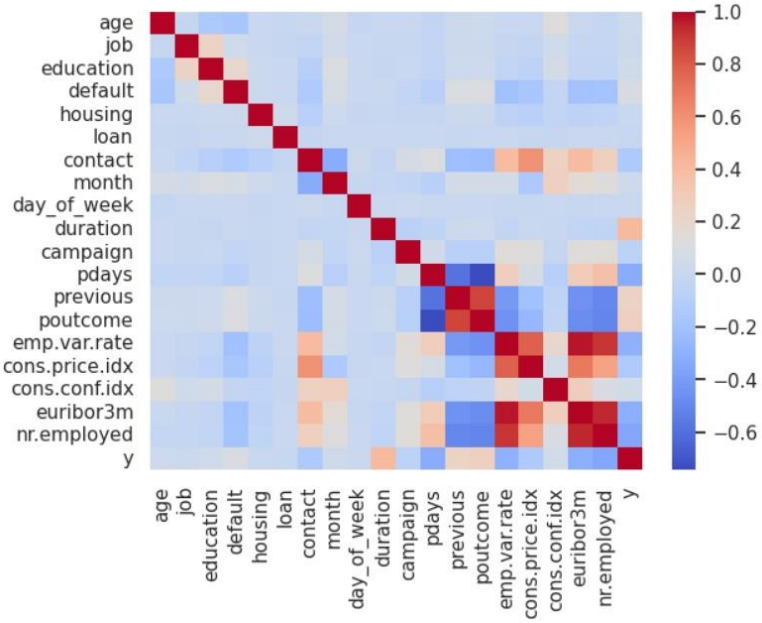

Figure C1. Correlation Matrix