UTS UNIVERSITY OF TECHNOLOGY SYDNEY

ASSESSMENT TASK 1: EXPLORATORY DATA ANALYSIS.

Agam Singh Saini
ID:25531702

Submitted to:

Dr. Alice Dong

on

2nd September 2025

Table Of Contents

Section 1: Problem Formulation

1.1 Analysis Context:

The dataset pertains to a telecommunication company's marketing campaign aimed at promoting a new subscription plan. The analysis focuses on understanding customer behavior and identifying segments most responsive to the campaign.

1.2 Specific Problem:

The goal is to analyze the dataset to uncover patterns, relationships, and insights that can guide marketing strategies. This includes identifying key features influencing customer responses and addressing data quality issues.

1.3 Pertinent Questions:

1. What are the characteristics of customers who respond positively to the campaign?
2. Which features (e.g., demographic, economic, or campaign-related) are most predictive of customer responses?
3. How do missing or ambiguous values (e.g., "unknown") impact the analysis?
4. Are there any outliers or anomalies in the data that need special handling?

1.4 Hypotheses:

1. Customers with longer call durations are more likely to subscribe.
2. Economic indicators (e.g., employment variation rate, consumer confidence index) significantly influence customer responses.
3. Certain demographic groups (e.g., job, education) are more responsive to the campaign.
4. Customers with a successful outcome in a prior marketing campaign are more likely to subscribe to the current campaign compared to those with failed or nonexistent prior outcomes.

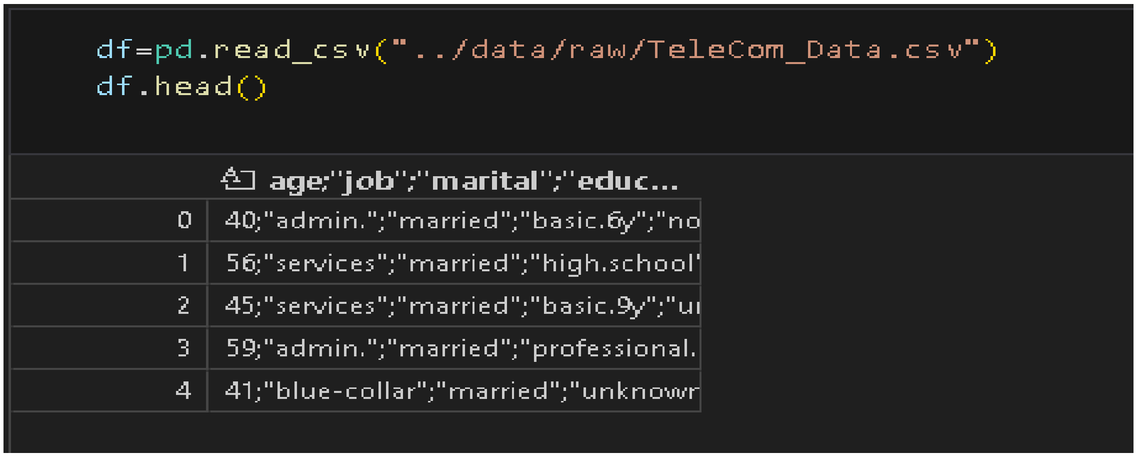
Section 2: Data Preprocessing

2.0 Data Introduction:

- Data Dictionary
  - 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?
- The dataset contains 41,180 entries and 21 columns.
- Data types include 5 float columns, 5 integer columns, and 11 object (categorical) columns.
- Key features cover demographics (age, job, marital, education), financial status (default, housing, loan), campaign details (contact, month, day_of_week, duration, campaign, pdays, previous, poutcome), economic indicators (emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed), and the target variable (y).

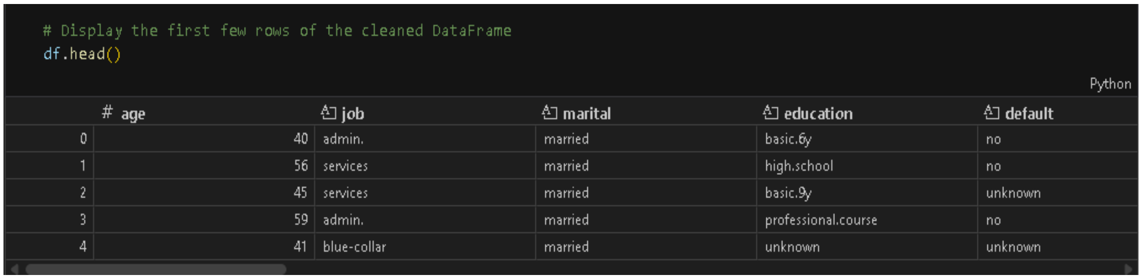
2.1 Data Import and Initial Cleaning



```
df=pd.read_csv("../data/raw/TeleCom_Data.csv")
df.head()
```

```
      age;"job";"marital";"educ...
   0  40;"admin.";"married";"basic.6y";"no
   1  56;"services";"married";"high.school"
   2  45;"services";"married";"basic.9y";"u
   3  59;"admin.";"married";"professional.
   4  41;"blue-collar";"married";"unknow
```

- The raw dataset was found to be incorrectly formatted, with issues related to quotes and delimiters.
- String manipulation techniques were applied to correct the formatting, ensuring proper structure for data import.



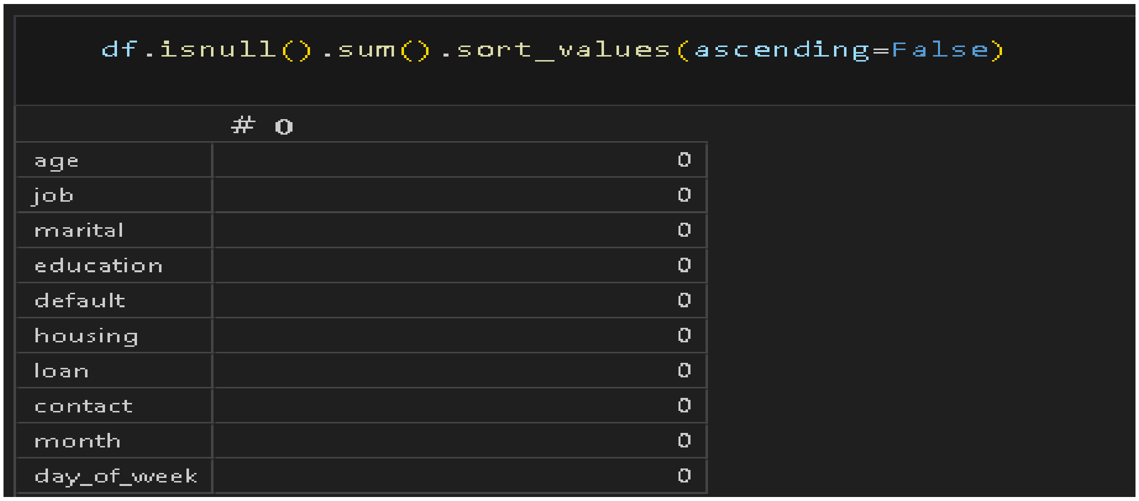
```
# Display the first few rows of the cleaned DataFrame
df.head()
```

Python

	# age	job	marital	education	default
0	40	admin.	married	basic.6y	no
1	56	services	married	high.school	no
2	45	services	married	basic.9y	unknown
3	59	admin.	married	professional.course	no
4	41	blue-collar	married	unknown	unknown

- The cleaned data was loaded into a pandas DataFrame for further analysis.
- It was ensured that each feature had the correct data type.

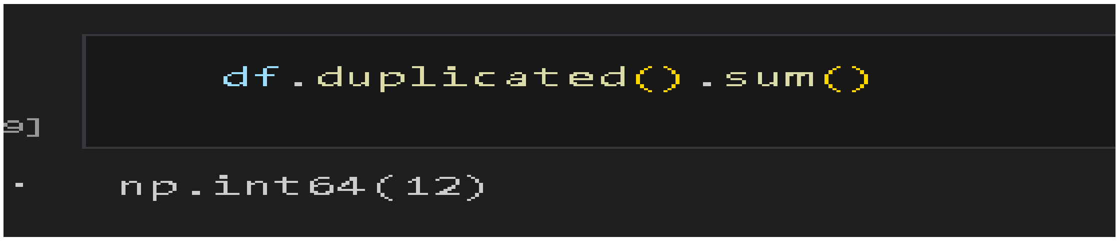
2.2 Handling Missing and Duplicate Values



```
df.isnull().sum().sort_values(ascending=False)
```

	# 0
age	0
job	0
marital	0
education	0
default	0
housing	0
loan	0
contact	0
month	0
day_of_week	0

- The dataset was checked for missing values, and it was confirmed that there are no null values present.

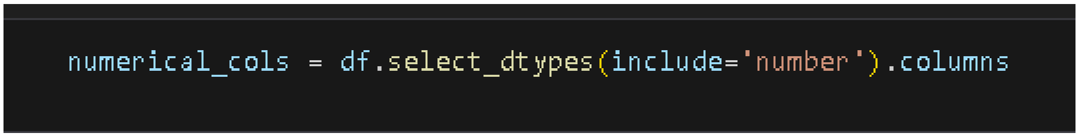


```
df.duplicated().sum()
```

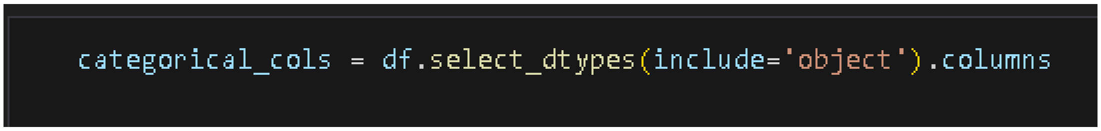
```
np.int64(12)
```

- Duplicate rows were identified (12 in total) and removed to ensure data integrity.

2.3. Feature Identification and Overview



```
numerical_cols = df.select_dtypes(include='number').columns
```



```
categorical_cols = df.select_dtypes(include='object').columns
```

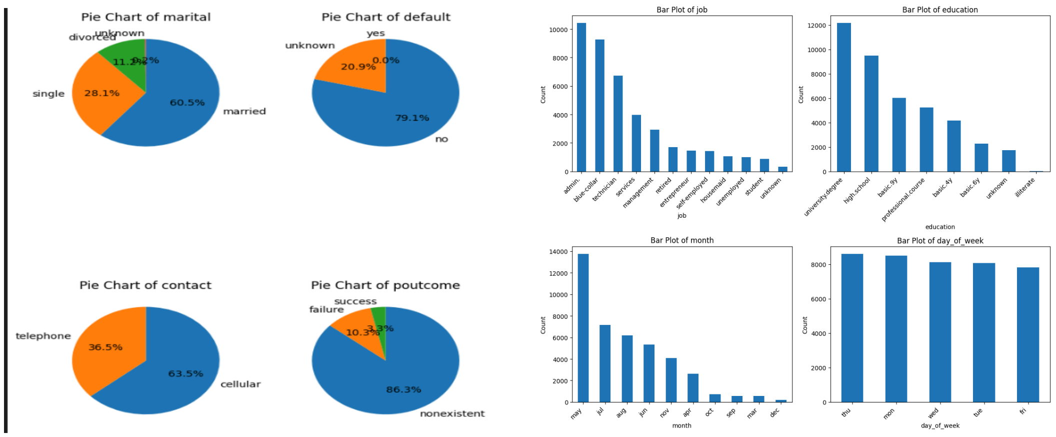
- Categorical and numerical features were identified and described, providing a clear and easy overview of the dataset's structure.

2.4. Final Dataset Validation

- These preprocessing steps ensured the dataset was error-free and accurately processed, ready for subsequent analysis.

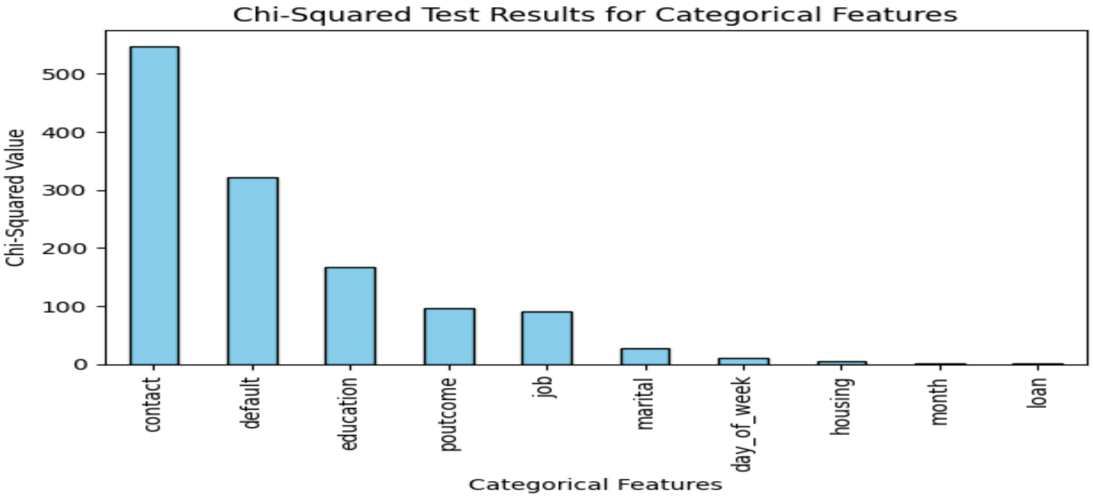
Section 3: Exploratory Data Analysis

3.1 Categorical Feature Analysis



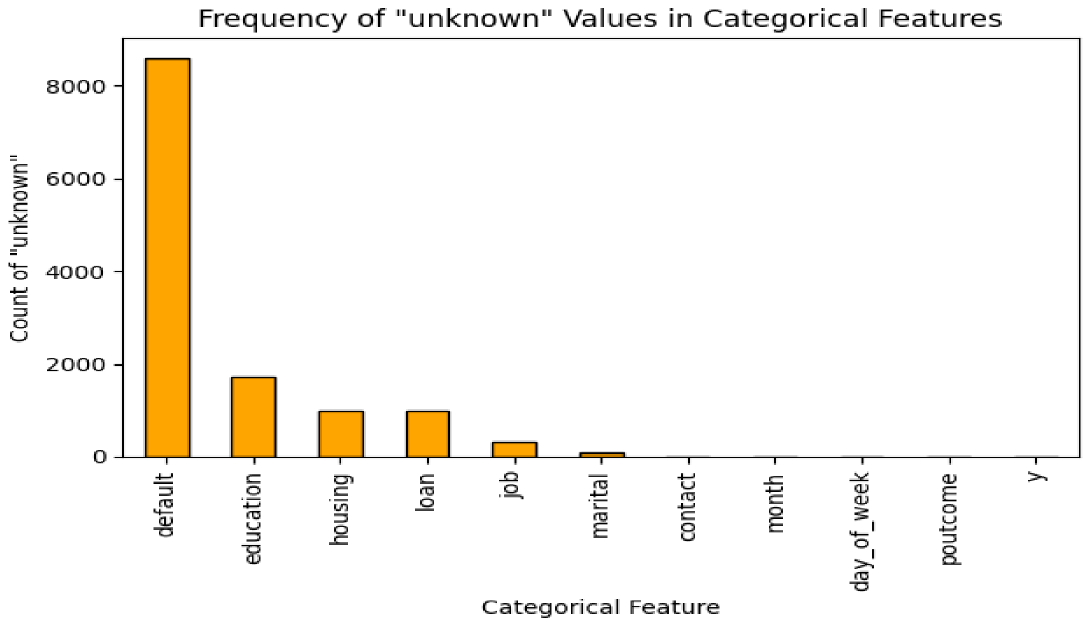
- Categorical features were described and visualized using pie and bar charts.
- Key findings include dominance of certain categories (e.g., 'admin.' in job, 'married' in marital, 'university.degree' in education) and class imbalance in the target variable.
- Seasonality and distribution patterns were observed in features like month and day_of_week.

3.1.1 Chi-Squared Test



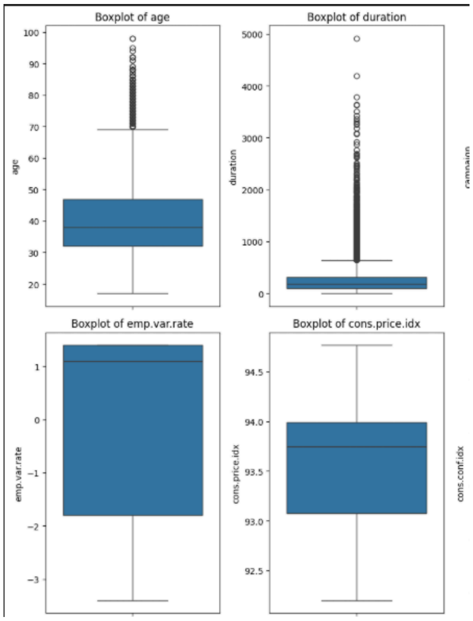
- A chi-squared test was done to assess the association between categorical features and target.
- Contact method, default status, and education showed the strongest associations with the target.

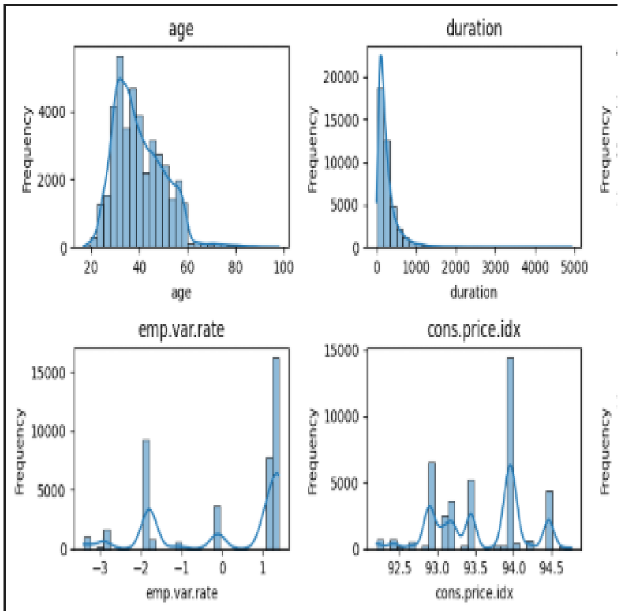
3.1.2 Analysis of 'Unknown' Values



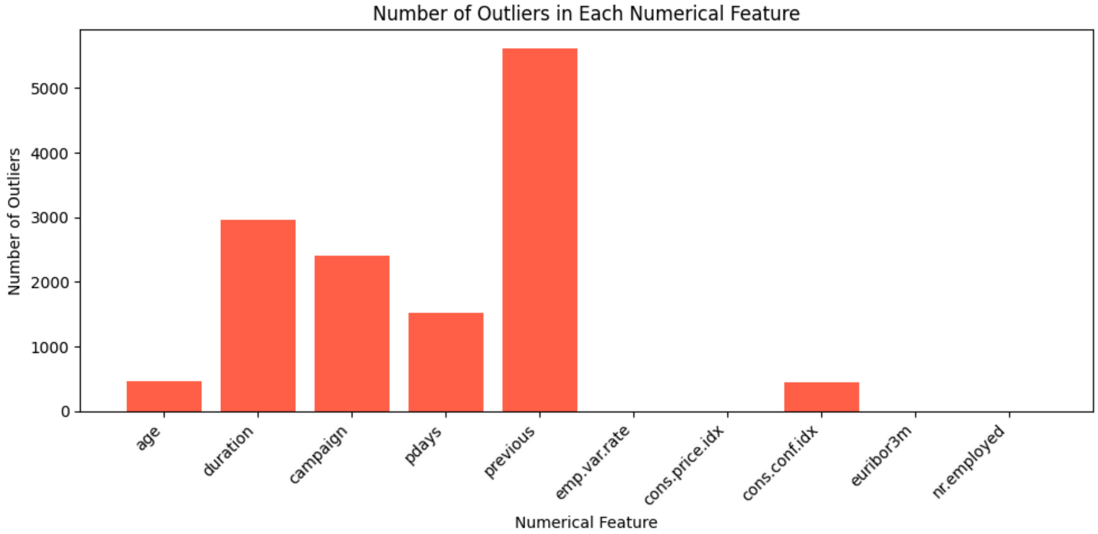
- The frequency and percentage of 'unknown' values in each categorical feature were calculated and visualized.
- The 'default' feature had the highest proportion of 'unknown' values, followed by education, housing, and loan.
- Most other features had very few or no 'unknown' values.

3.2 Numerical Feature Analysis



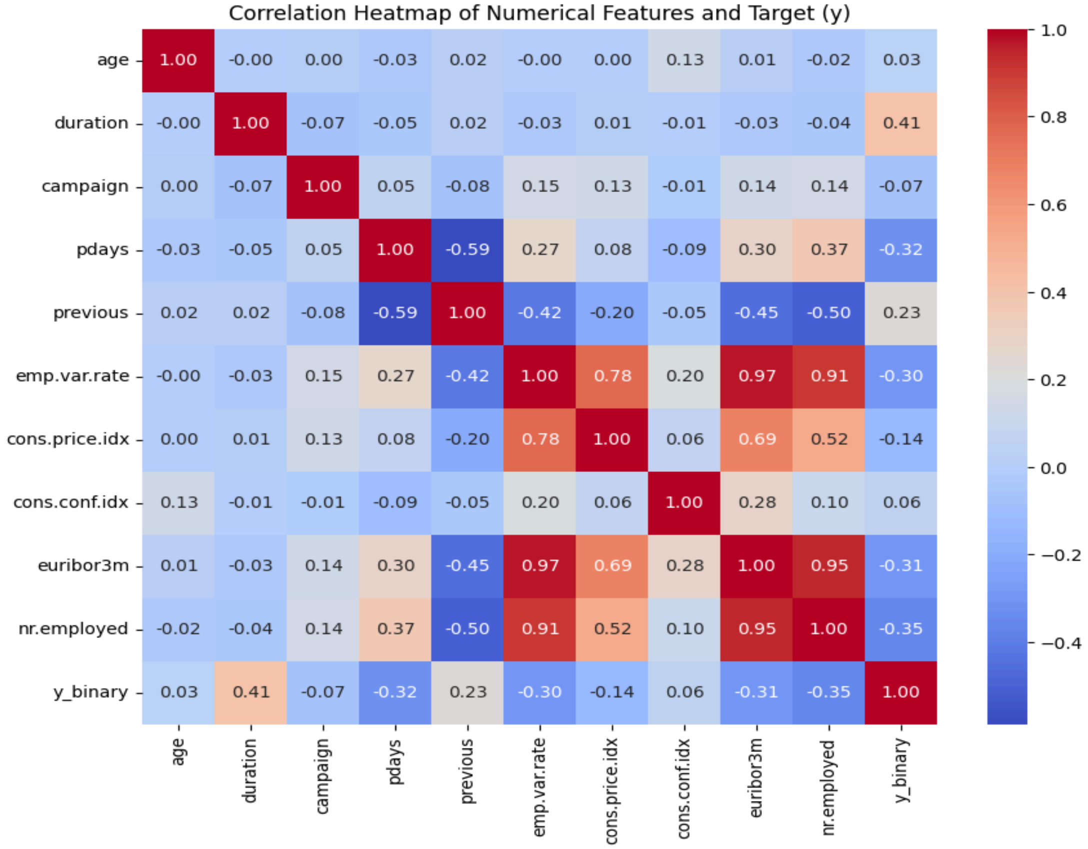


- Descriptive statistics, histograms, and boxplots were used to explore numerical features.
- Many features (e.g., duration, campaign, pdays, previous) were right-skewed and contained outliers.
- Economic indicators showed multimodal or discrete distributions.



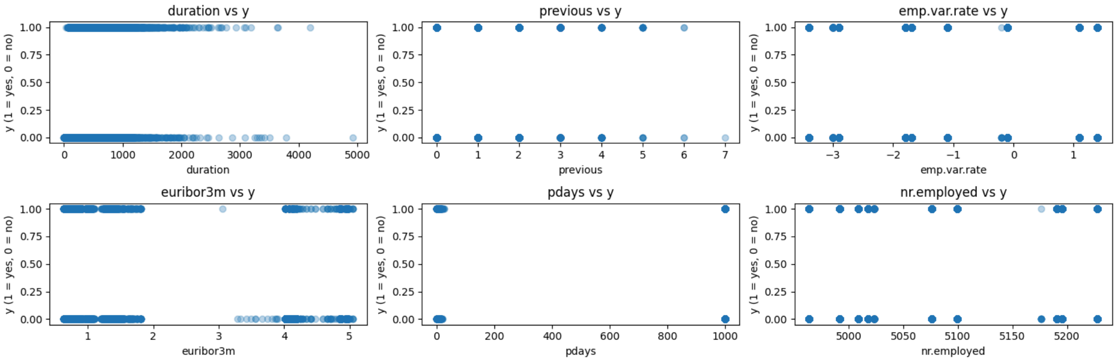
- Outlier counts were calculated and visualized for each numerical feature.

3.2.1 Correlation and Feature Relationships



	age	duration	campaign	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed	y_binary
age	1.00	-0.00	0.00	-0.03	0.02	-0.00	0.00	0.13	0.01	-0.02	0.03
duration	-0.00	1.00	-0.07	-0.05	0.02	-0.03	0.01	-0.01	-0.03	-0.04	0.41
campaign	0.00	-0.07	1.00	0.05	-0.08	0.15	0.13	-0.01	0.14	0.14	-0.07
pdays	-0.03	-0.05	0.05	1.00	-0.59	0.27	0.08	-0.09	0.30	0.37	-0.32
previous	0.02	0.02	-0.08	-0.59	1.00	-0.42	-0.20	-0.05	-0.45	-0.50	0.23
emp.var.rate	-0.00	-0.03	0.15	0.27	-0.42	1.00	0.78	0.20	0.97	0.91	-0.30
cons.price.idx	0.00	0.01	0.13	0.08	-0.20	0.78	1.00	0.06	0.69	0.52	-0.14
cons.conf.idx	0.13	-0.01	-0.01	-0.09	-0.05	0.20	0.06	1.00	0.28	0.10	0.06
euribor3m	0.01	-0.03	0.14	0.30	-0.45	0.97	0.69	0.28	1.00	0.95	-0.31
nr.employed	-0.02	-0.04	0.14	0.37	-0.50	0.91	0.52	0.10	0.95	1.00	-0.35
y_binary	0.03	0.41	-0.07	-0.32	0.23	-0.30	-0.14	0.06	-0.31	-0.35	1.00

- Correlation matrices and heatmaps were generated for numerical features and the binary target.



- Features with strong correlations to the target were identified and visualized using scatter plots.
- Key relationships included longer call durations and recent contacts being associated with higher subscription rates.

3.3 Key Insights

The EDA in the attached notebook systematically explores the Telecom Campaign dataset to support the goal of predicting term deposit subscription. Key findings include:

- Class Imbalance: The target variable y is highly imbalanced, with most clients not subscribing. This highlights the need for careful model evaluation and possibly resampling techniques.
- Outliers: Numerical features such as duration, campaign, pdays, and previous show strong right-skewness and contain significant outliers, as seen in histograms and boxplots. These outliers may affect model performance and require preprocessing.
- Missingness in Categorical Features: Several categorical features (e.g., default, education, housing, loan) have substantial 'unknown' values, visualized in bar plots. Handling these is important for reliable modeling.
- Feature Predictiveness: Chi-squared tests and correlation analysis reveal that contact method, duration, and economic indicators (e.g., emp.var.rate, euribor3m, nr.employed) are most predictive of the target. These features should be prioritized in modeling.
- Data Quality: The notebook documents cleaning steps, including string manipulation and duplicate removal, ensuring a robust foundation for analysis.
- Visualization: Pie charts, bar plots, and heatmaps provide clear insights into feature distributions and relationships.

References

[1] GitHub Copilot: GitHub Copilot. (n.d.). Retrieved September 2, 2025, from https://github.com/features/copilot

[2] UTS Canvas: University of Technology Sydney. (n.d.). Canvas. Retrieved September 2, 2025, from https://canvas.uts.edu.au/