BANKING DATA ANALYSIS CASE PROJECT

Ishan Singh

CONTENTS

- Data Assessing And Cleaning
- Exploratory Data Analysis
- Key Insights

DATA ASSESSING AND CLEANING

BRIEF SUMMARY OF THE DATASET

The dataset consists of records from direct marketing campaigns conducted by a Portuguese banking institution between May 2008 and November 2010. With 45,211 rows and 18 columns, the primary objective is to predict whether clients subscribed to term deposits following phone-based marketing efforts. Each row represents a client contacted during these campaigns, with the target variable (y) indicating subscription outcome (yes or no). The dataset likely includes a range of features such as demographic data, client financial details, and specifics of the marketing interactions (e.g., number of calls, duration). The campaigns utilized various outreach strategies including email, advertisements, and telephonic marketing, with emphasis on the latter due to its perceived effectiveness despite requiring significant resources. The dataset aims to enable analysis aimed at optimizing marketing strategies, targeting potential subscribers more accurately, and potentially reducing costs associated with large-scale telephonic campaigns.

COLUMN DESCRIPTIONS

- **age**: Numeric variable indicating the age of the bank client in years.
- **job**: Categorical variable indicating the type of job the client has (e.g., "admin.", "unknown", "unemployed", "management", etc.).
- **marital**: Categorical variable indicating the marital status of the client (e.g., "married", "divorced", "single").
- **education**: Categorical variable indicating the level of education of the client (e.g., "unknown", "secondary", "primary", "tertiary").
- **default**: Binary variable indicating whether the client has credit in default (options: "yes" or "no").
- **balance:** Numeric variable representing the average yearly balance in euros for the client.

- **housing**: Binary variable indicating whether the client has a housing loan (options: "yes" or "no").
- **loan:** Binary variable indicating whether the client has a personal loan (options: "yes" or "no").
- **contact**: Categorical variable indicating the type of communication used to contact the client (e.g., "unknown", "telephone", "cellular").
- **day**: Numeric variable representing the last contact day of the month.
- **month**: Categorical variable representing the last contact month of the year (e.g., "jan", "feb", "mar", etc.).
- **duration**: Numeric variable representing the duration of the last contact in seconds.

- **campaign**: Numeric variable representing the number of contacts performed during this campaign for this client.
- **pdays**: Numeric variable representing the number of days that passed by after the client was last contacted from a previous campaign (-1 means the client was not previously contacted).
- **previous**: Numeric variable representing the number of contacts performed before this campaign for this client.
- **poutcome**: Categorical variable representing the outcome of the previous marketing campaign (e.g., "unknown", "other", "failure", "success").
- **y**: Binary variable indicating whether the client has subscribed to a term deposit (target variable: "yes" or "no").

ISSUES

- Dirty Data (Data With Quality Issues):
  - Completeness Issues
    - Year column is missing.
    - Education and marital columns, each have 3 missing values.
  - Validity Issues
    - Few columns have mismatched datatype.
    - 5 duplicate rows exist.

REMOVAL

- Year column added using the info that the dataset provided contains columns ordered by date (from May 2008 to November 2010).
- Missing values in education column shifted to unknown category and those in marital column filled with "Data Not Available".
- Datatypes of few columns were changed to categorical.
- Duplicate rows were dropped.

ISSUES

- Messy Data (Data With Structural Issues):
  - Apart from day, month additional column day-month exists.
  - 2 duplicated columns exist i.e., marital and marital_status.

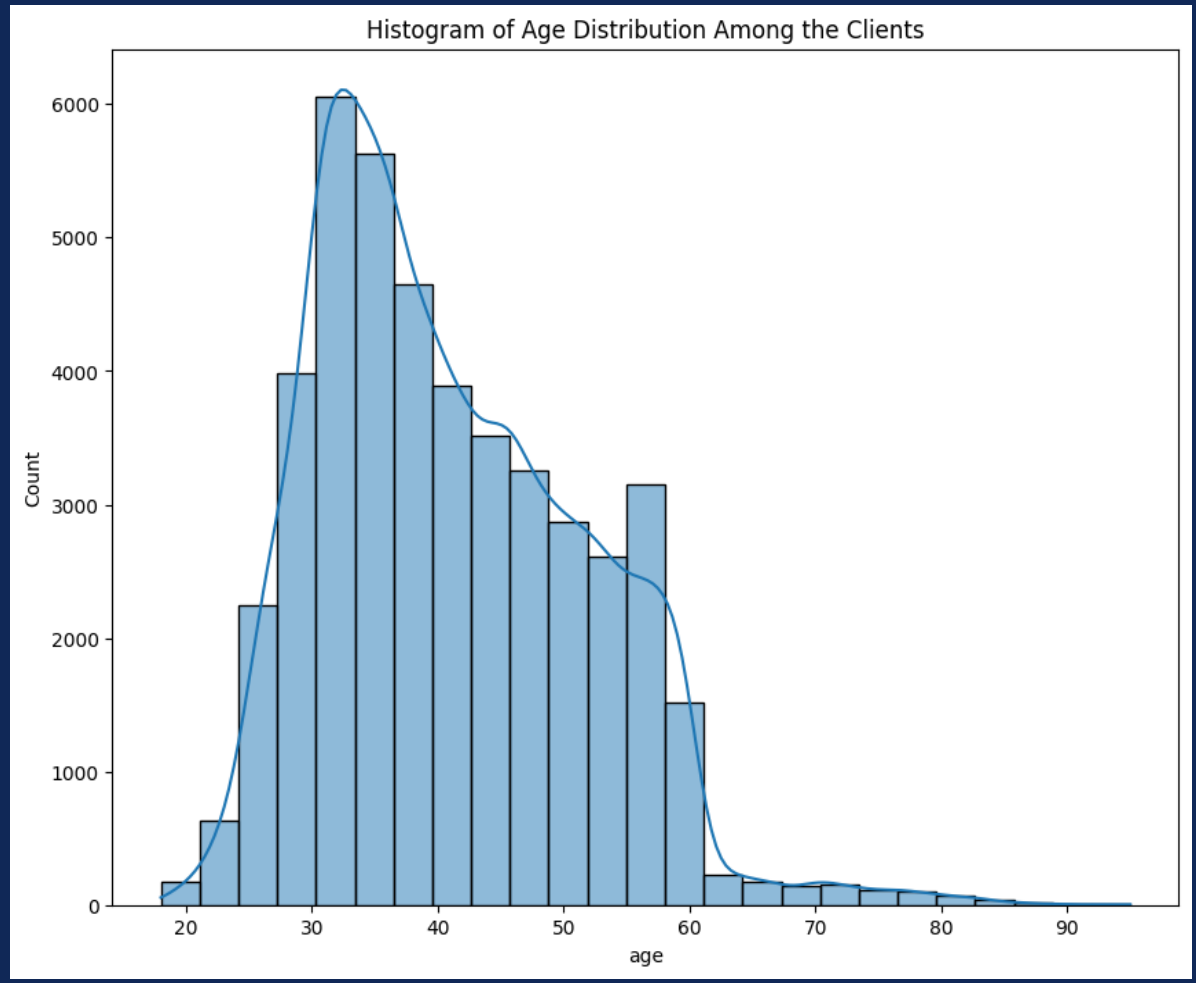
REMOVAL

- After adding year column; day, month and year column were merged into date column and day, month, year and day_month columns were removed.
- marital_status column was dropped.

EXPLORATORY DATA ANALYSIS

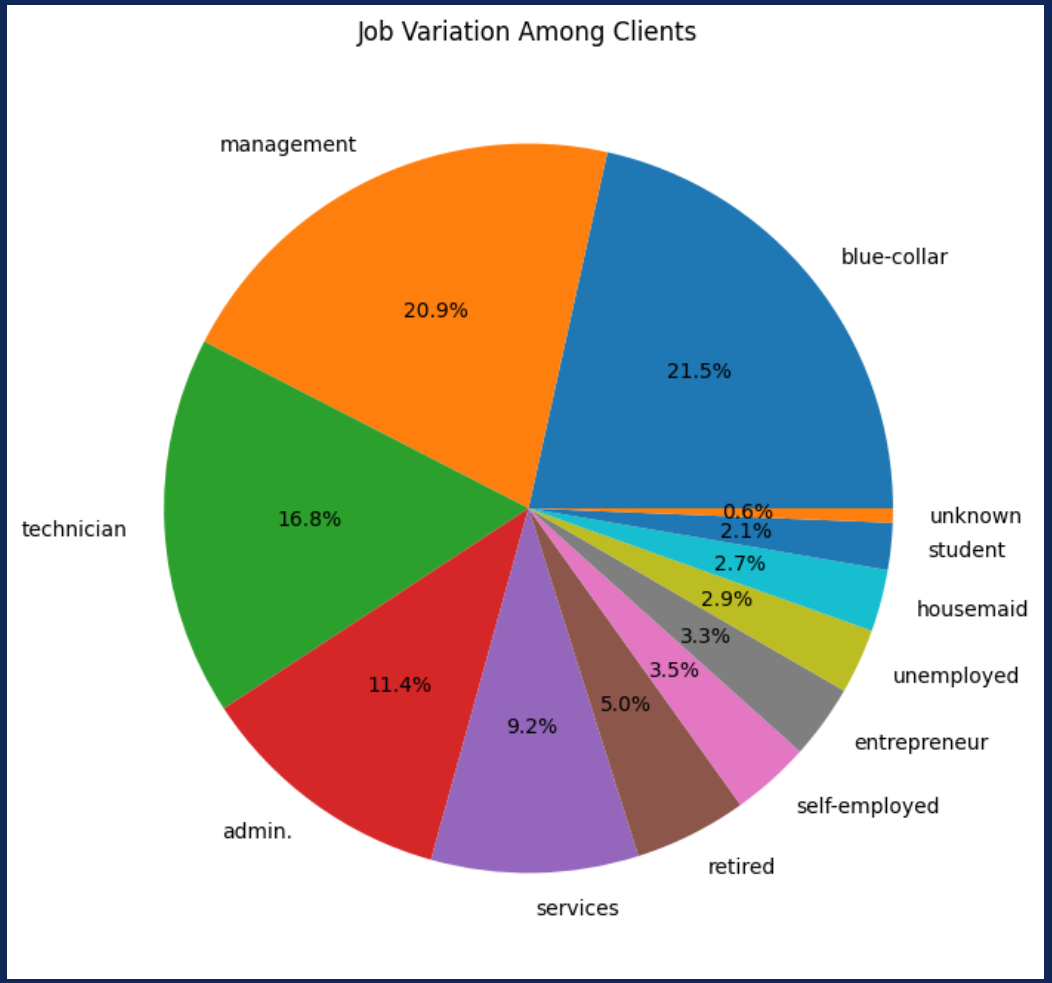
AGE DISTRIBUTION

- Most of the clients (nearly 97%) have their ages between 20 to 60.
- Median of the ages of the clients is 39 years.
- The distribution is right skewed.



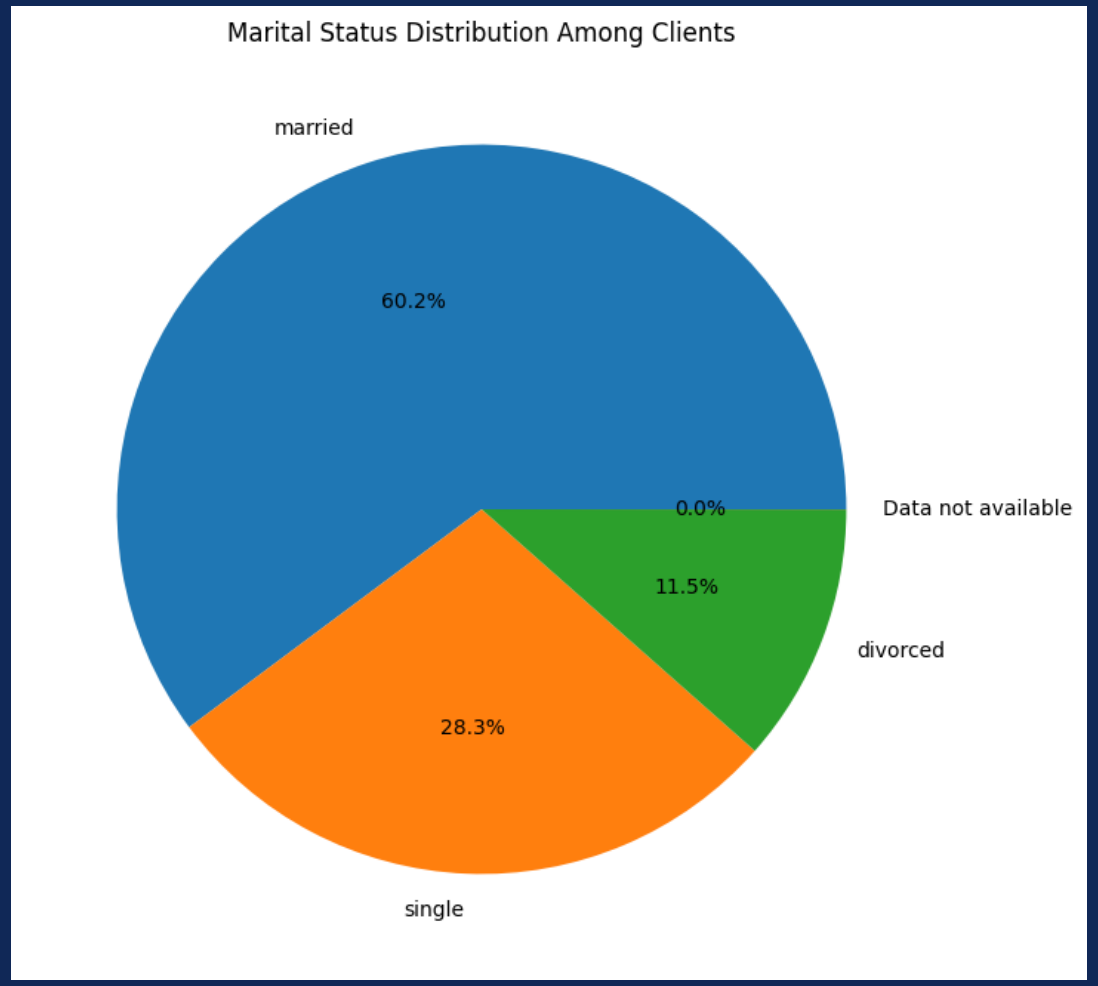
JOB DISTRIBUTION

- **unknown** category has smallest no. of clients(0.6%) followed by **student** category which has 2.1%
- Most of the clients have **blue collar** and **management** jobs (nearly 42.4%)
- The clients who are **self-employed, entrepreneurs, unemployed**, **housemaids**, and **students** are relatively less.



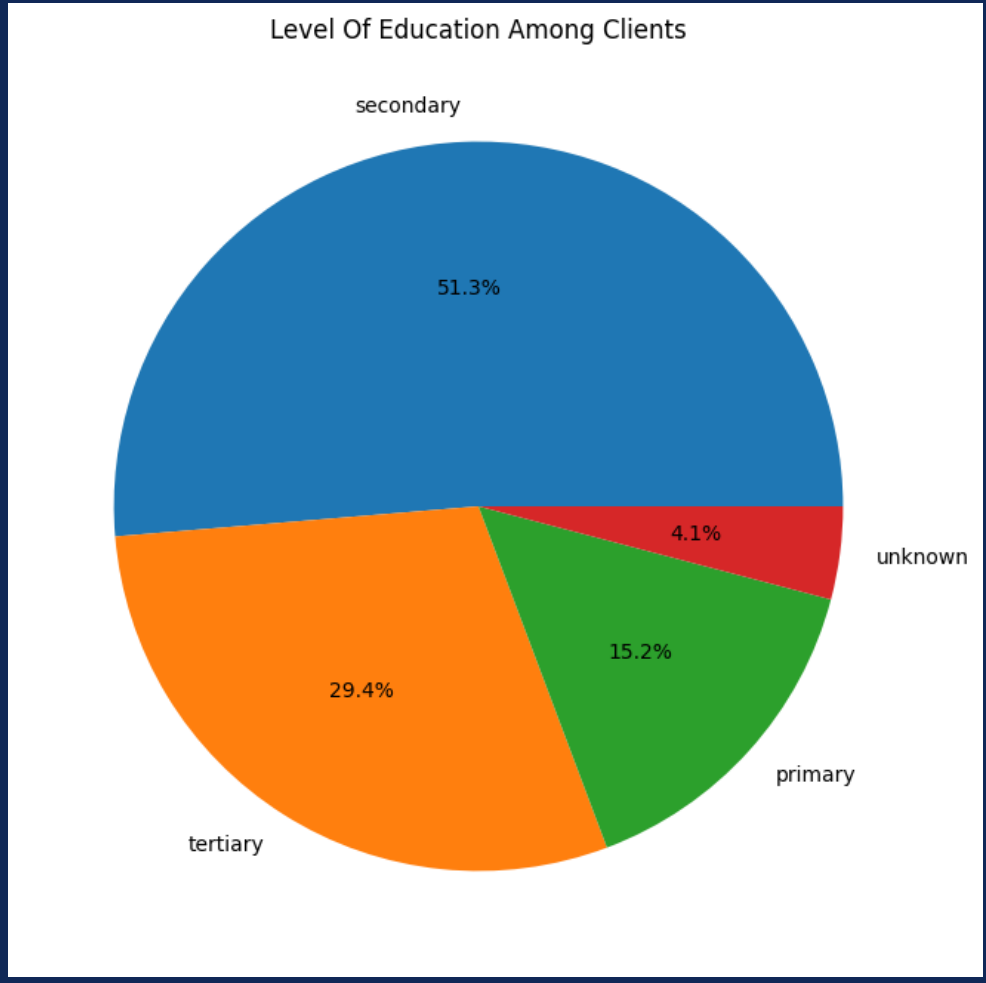
MARITAL STATUS DISTRIBUTION

- Most of the clients are **married** (nearly 60.2%) followed by **single** clients(28.3%) and **divorced** clients(11.5%).
- There are very few clients with **"Data not available"** indicating their marital status is not known.



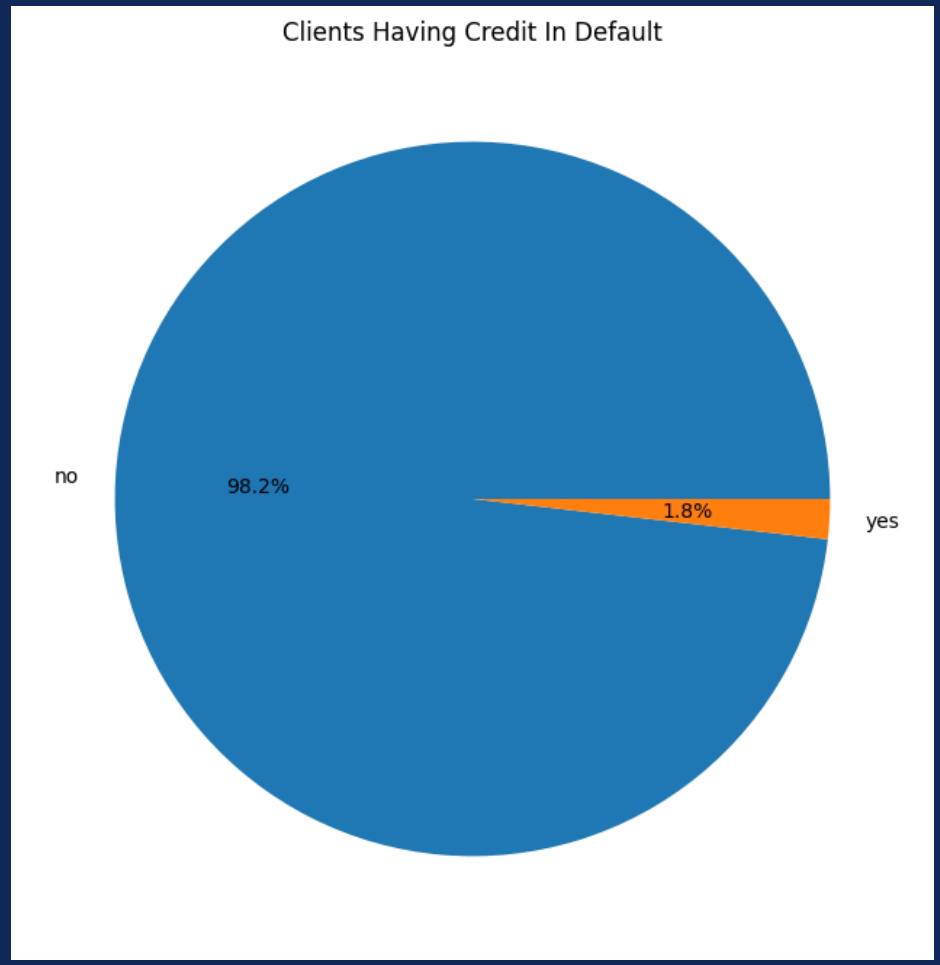
LEVEL OF EDUCATION

- Most of the clients have **secondary** level of education (nearly 51.3%) followed by those having **tertiary** (29.4%) and **primary** levels of education (15.2%).
- There are a few clients with **"unknown"** (nearly 4.1%) indicating their level of education is not known.



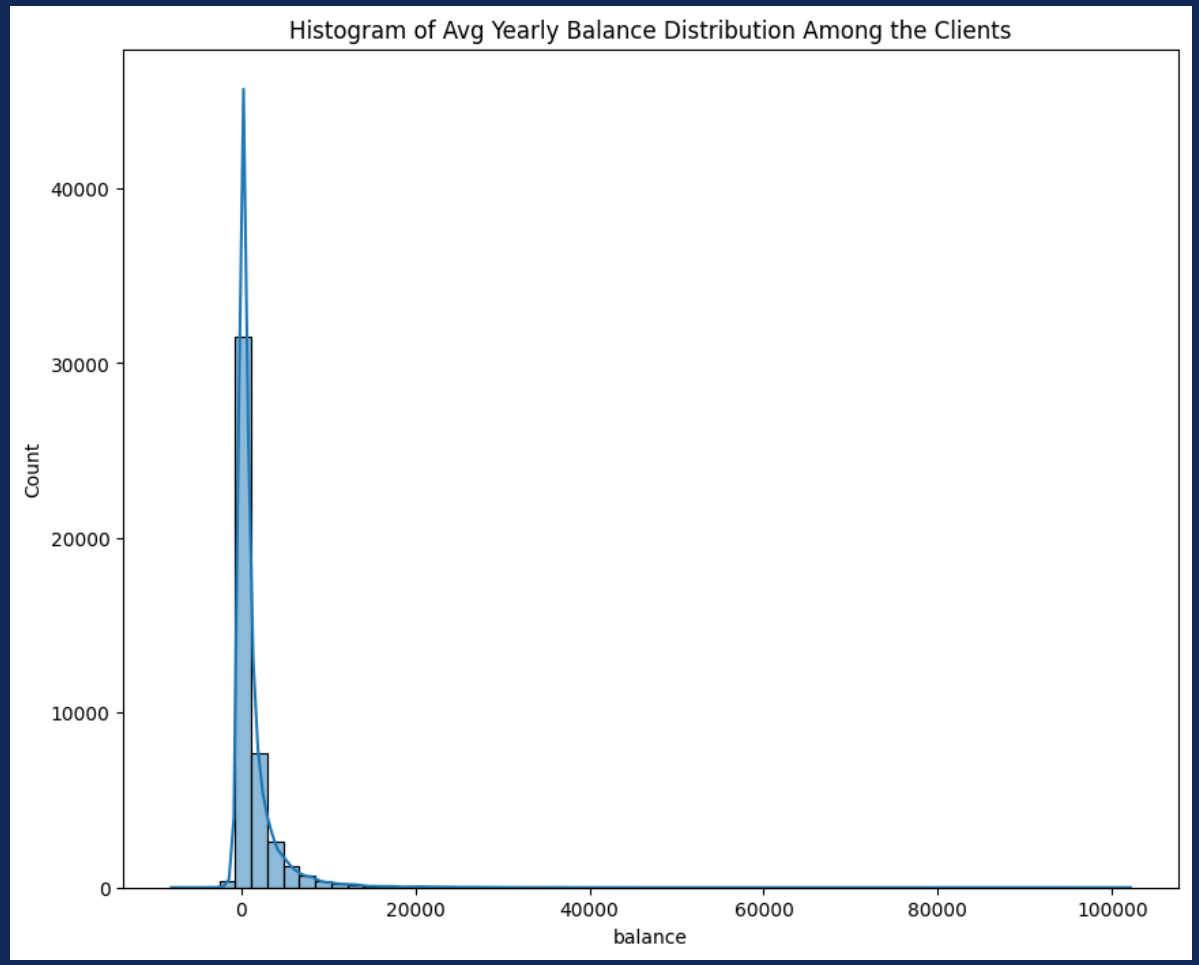
CLIENTS HAVING CREDIT IN DEFAULT

- As we can see from the pie chart that only 1.8% (815 clients) of clients have their credits in default.



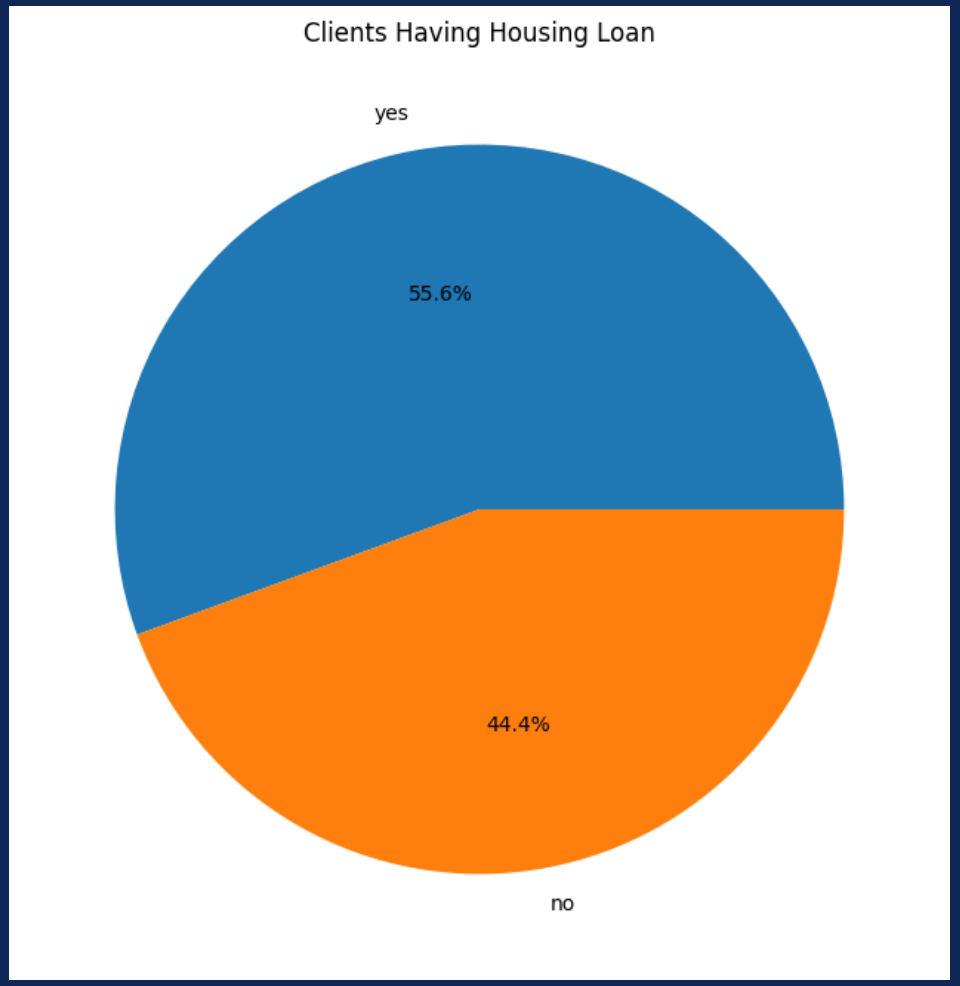
BALANCE DISTRIBUTION

- Most of the clients have relatively low average yearly **balance**.
- Few (3766) clients have negative average yearly **balance**.
- The median average yearly **balance** is 448 which is relatively low.
- There are very few clients having high average yearly **balance**. For example: no. of clients having average yearly **balance**>20000 is 193 which is very less.



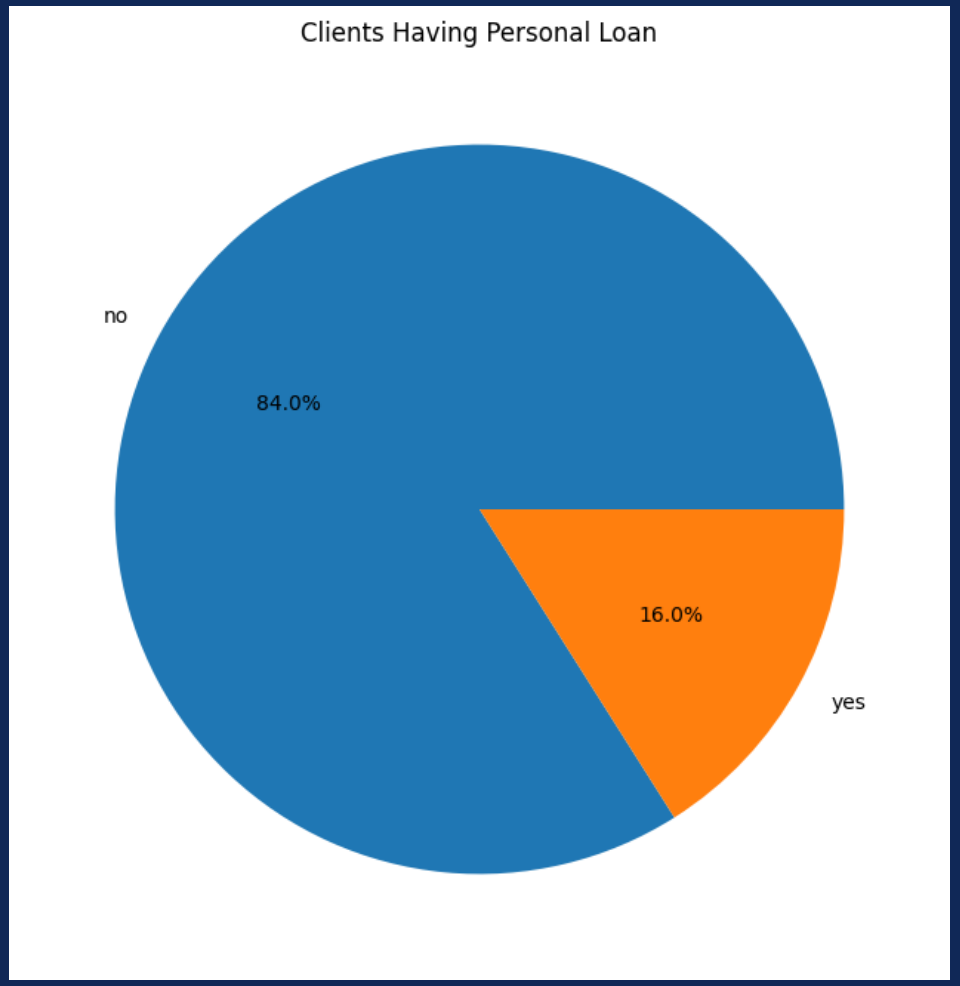
HOUSING LOAN DISTRIBUTION

- As we can see from the pie chart that majority of clients i.e. 55.6% (25130) clients have housing loan.



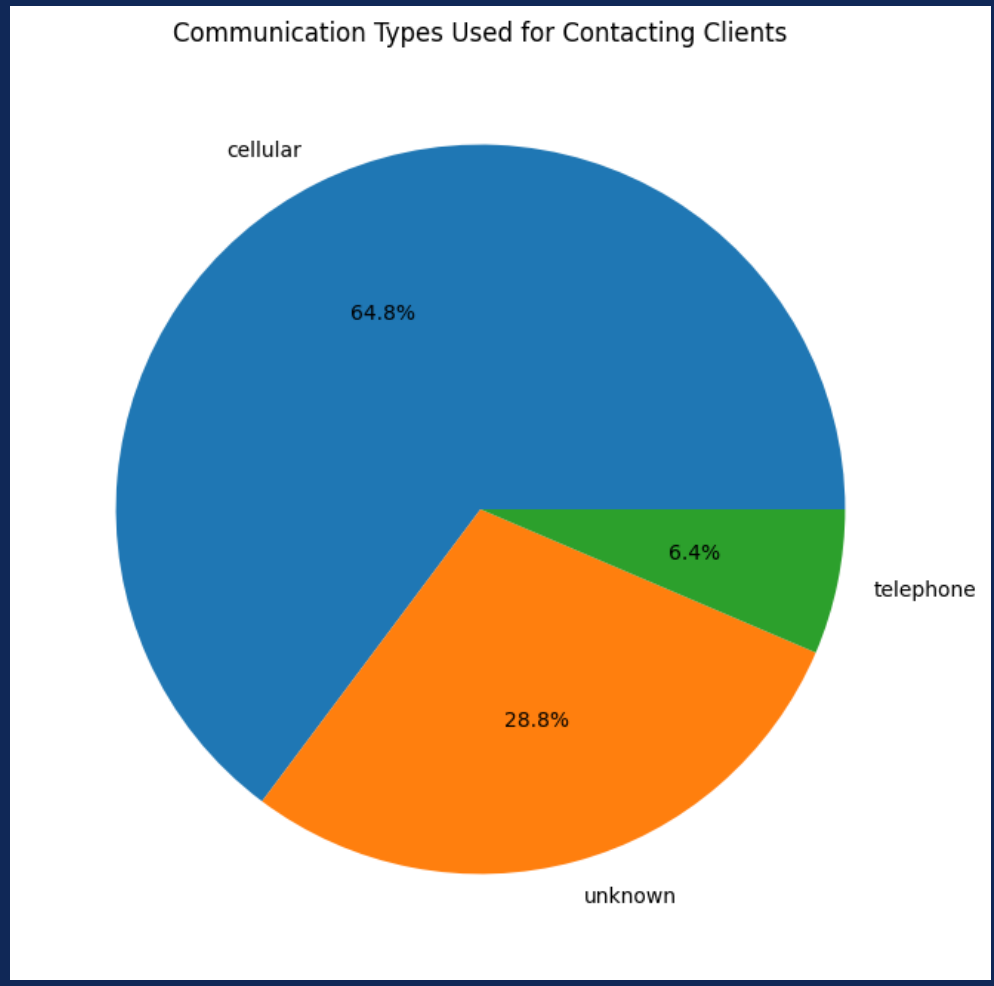
PERSONAL LOAN DISTRIBUTION

- As we can see from the pie chart that few clients i.e. 16% (7244) clients have personal loan.



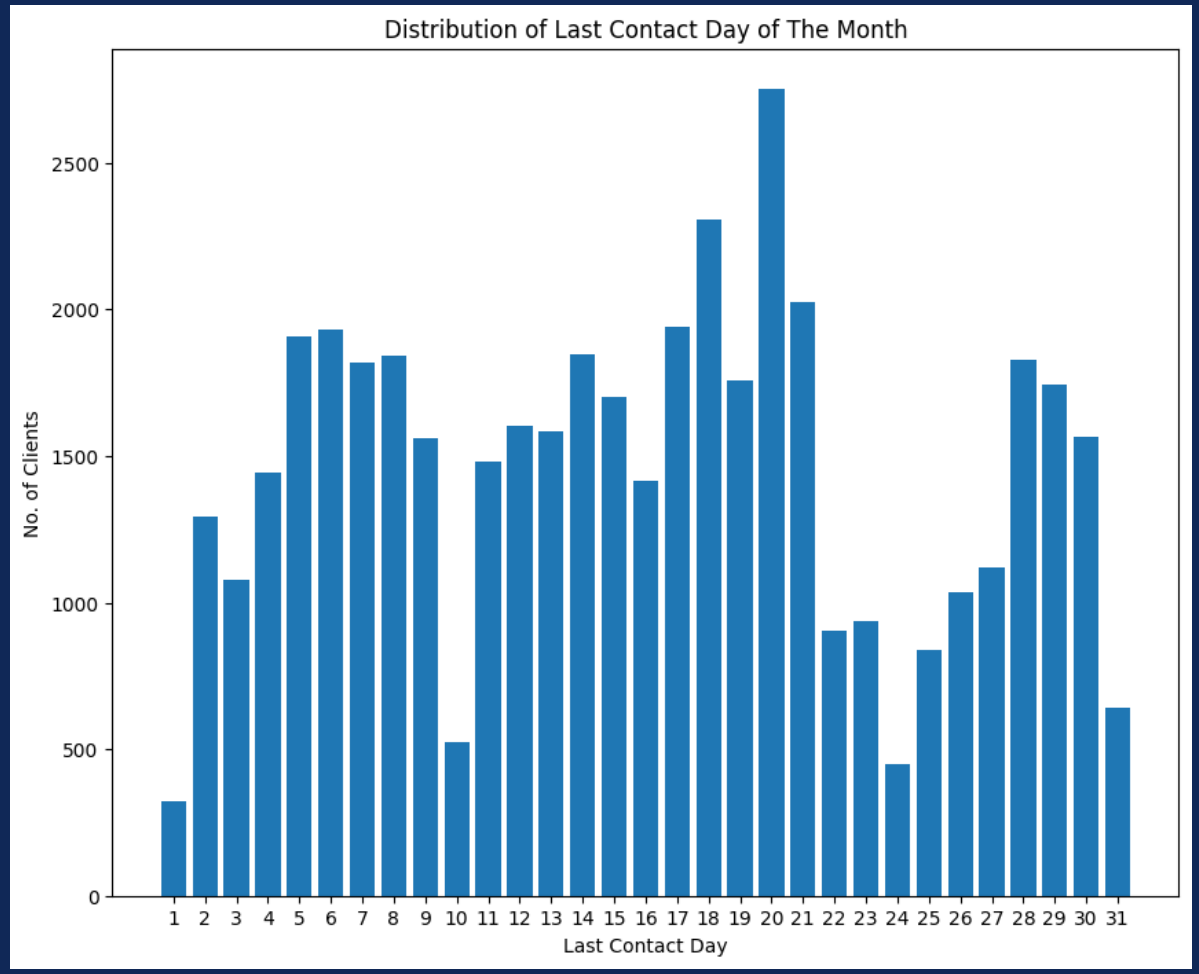
TYPES OF COMMUNICATION

- Mostly **cellular** communication was used to contact clients(nearly 64.8%).
- There are clients with **"unknown"** communication type (nearly 28.8%) indicating that the communication types used for contacting those clients is not known.
- Few clients were contacted through **telephone** (nearly 6.4%)



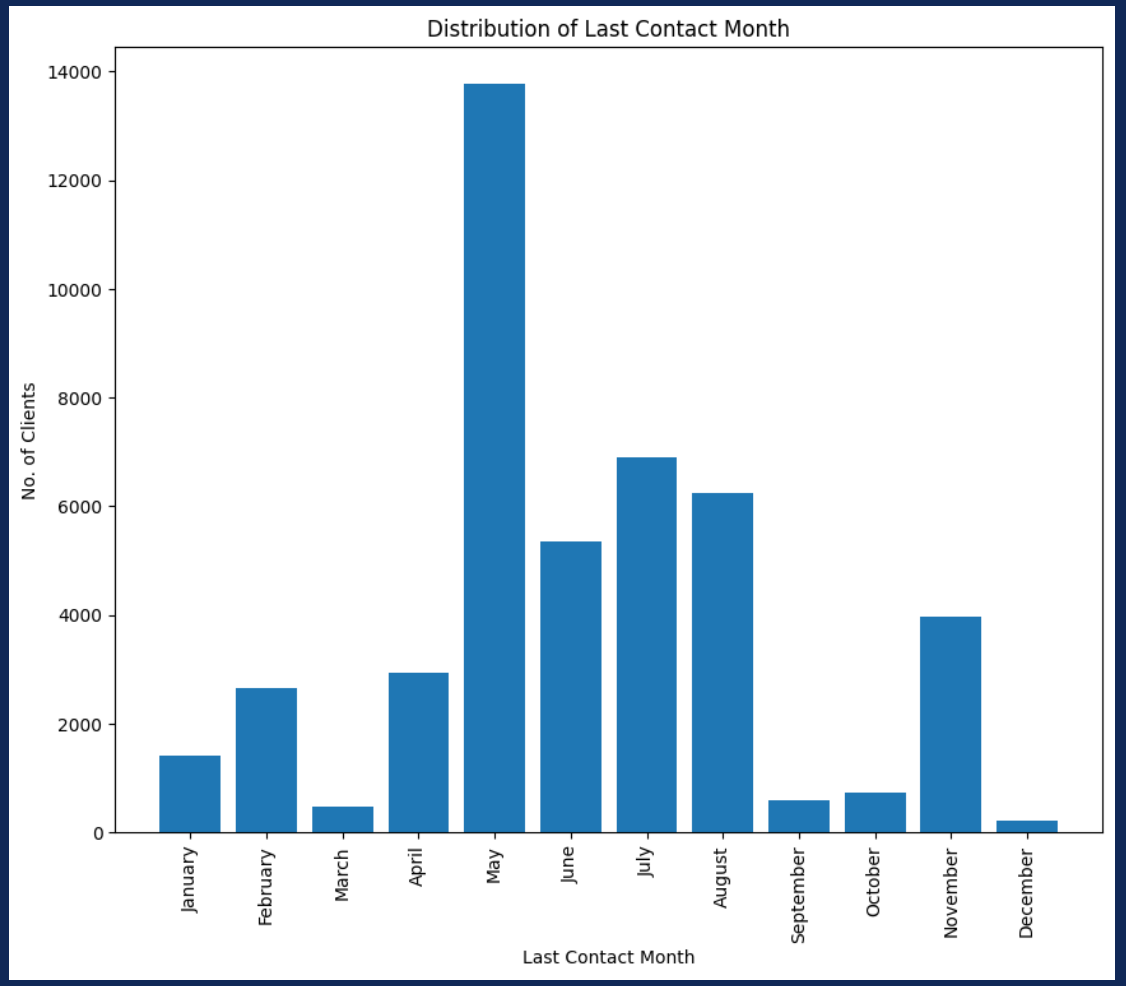
DISTRIBUTION OF LAST CONTACT DAY

- Days 1,10,24 and 31 have lower activity.
- This distribution is non uniform across the month.
- In the midle of the month there is significant increase in activity (especially on day 20).



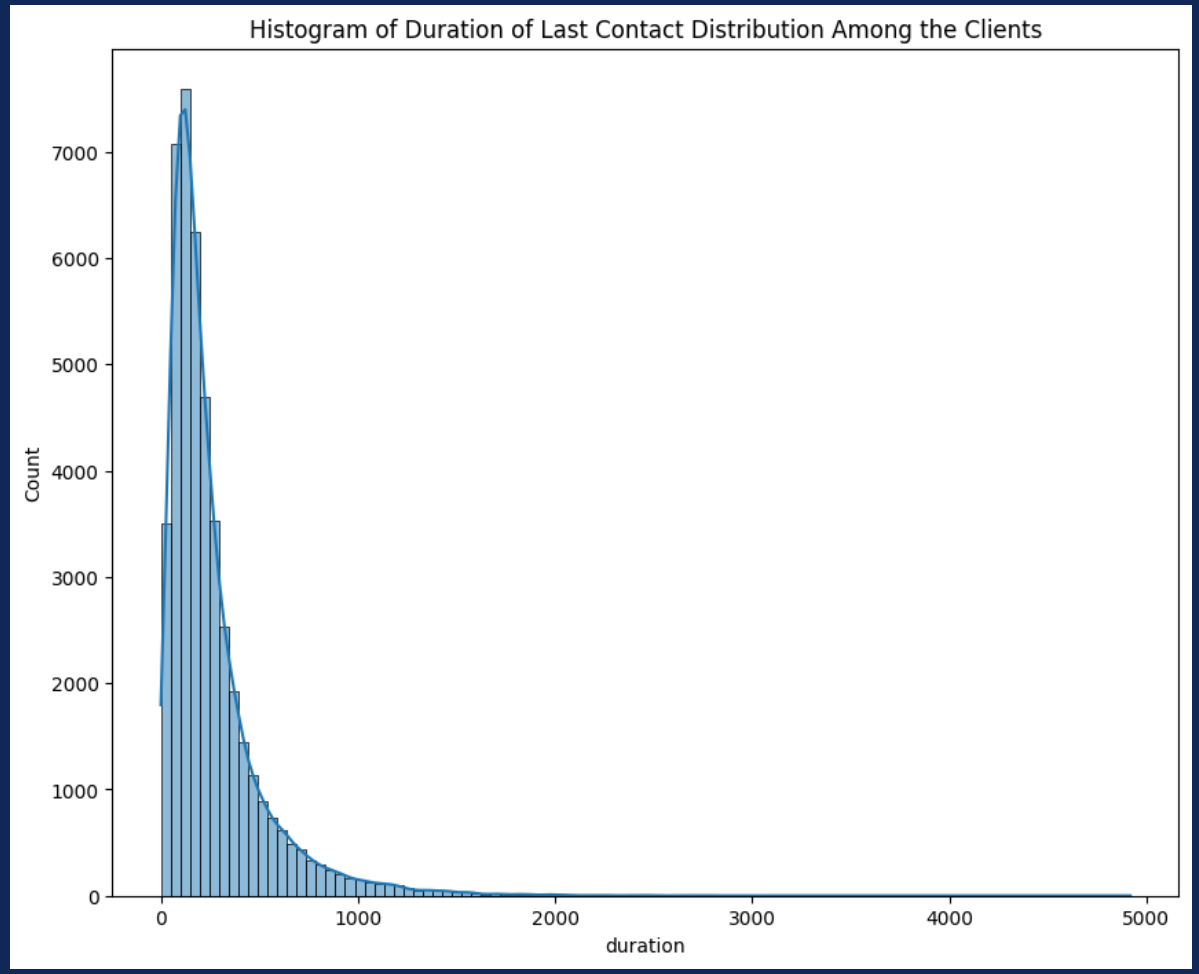
DISTRIBUTION OF LAST CONTACT MONTH

- **May** has the highest number of contacts with client suggesting a boom in marketing activities.
- **March**, **September**, **October** and **December** has relatively lower number of contacts with client.
- Rest months have moderate number of contacts.



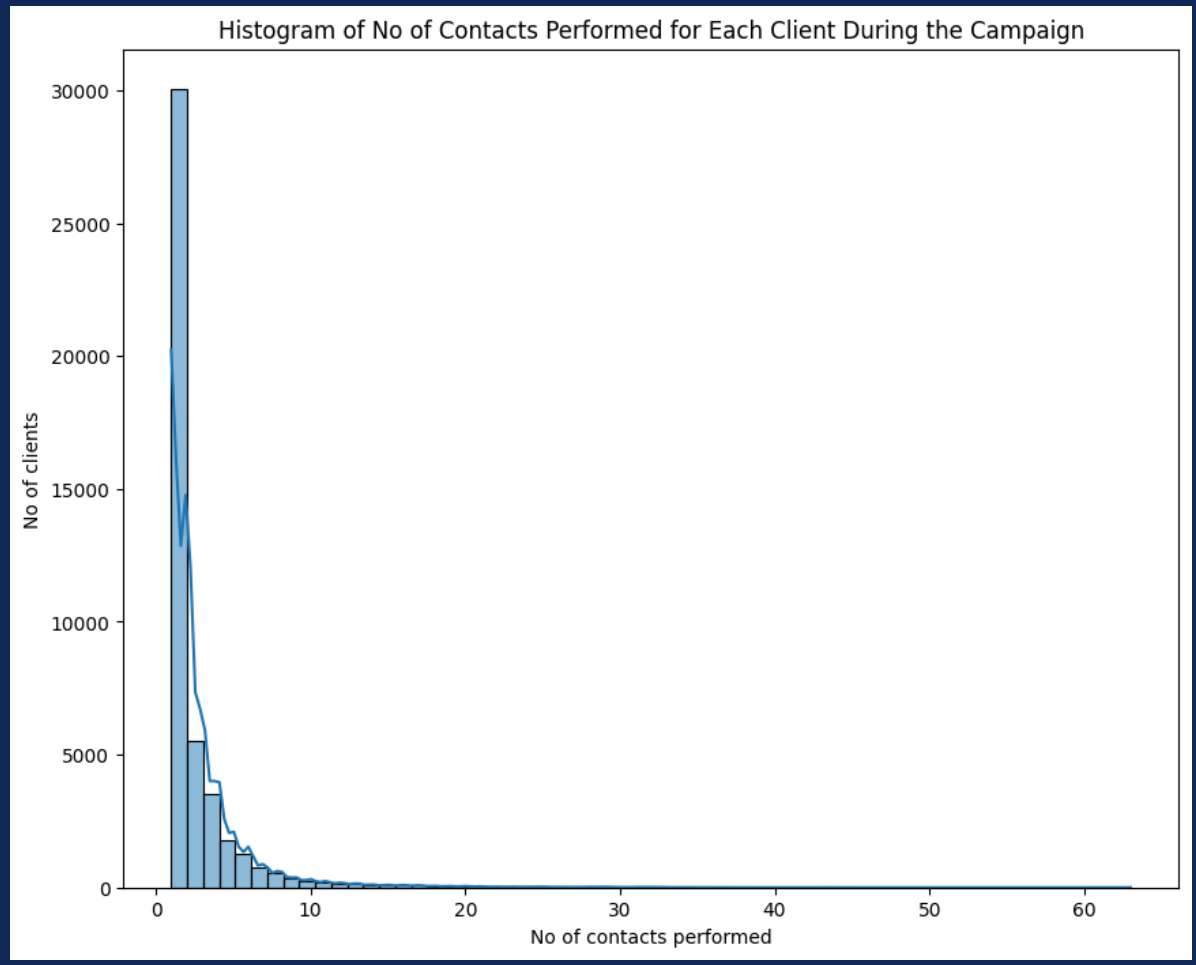
DISTRIBUTION OF DURATION OF LAST CONTACT

- The median of contact **duration** is 180 seconds.
- Most of the calls were of short **durations**, showing focus on quick interactions.
- Only few calls have very long **duration**.
- This distribution is highly right skewed as number of calls decreases rapidly as call **duration** increases.



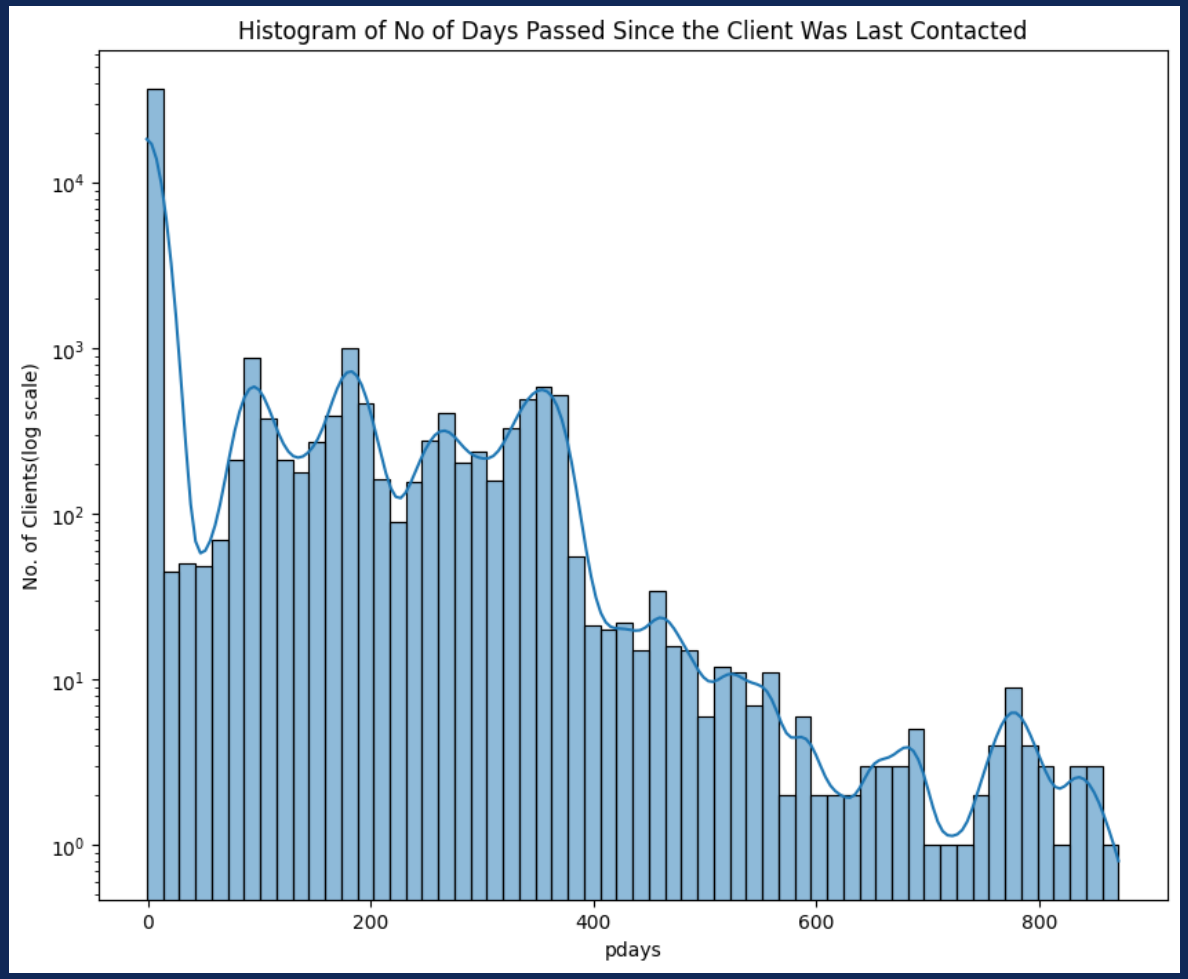
NUMBER OF CONTACTS PERFORMED DURING CAMPAIGN

- This distribution is also highly right skewed.
- The median of no of contacts performed during campaign is 2.
- Large number of clients have been contacted during campaign very few times.
- Only a few clients have high number of contacts during campaign.



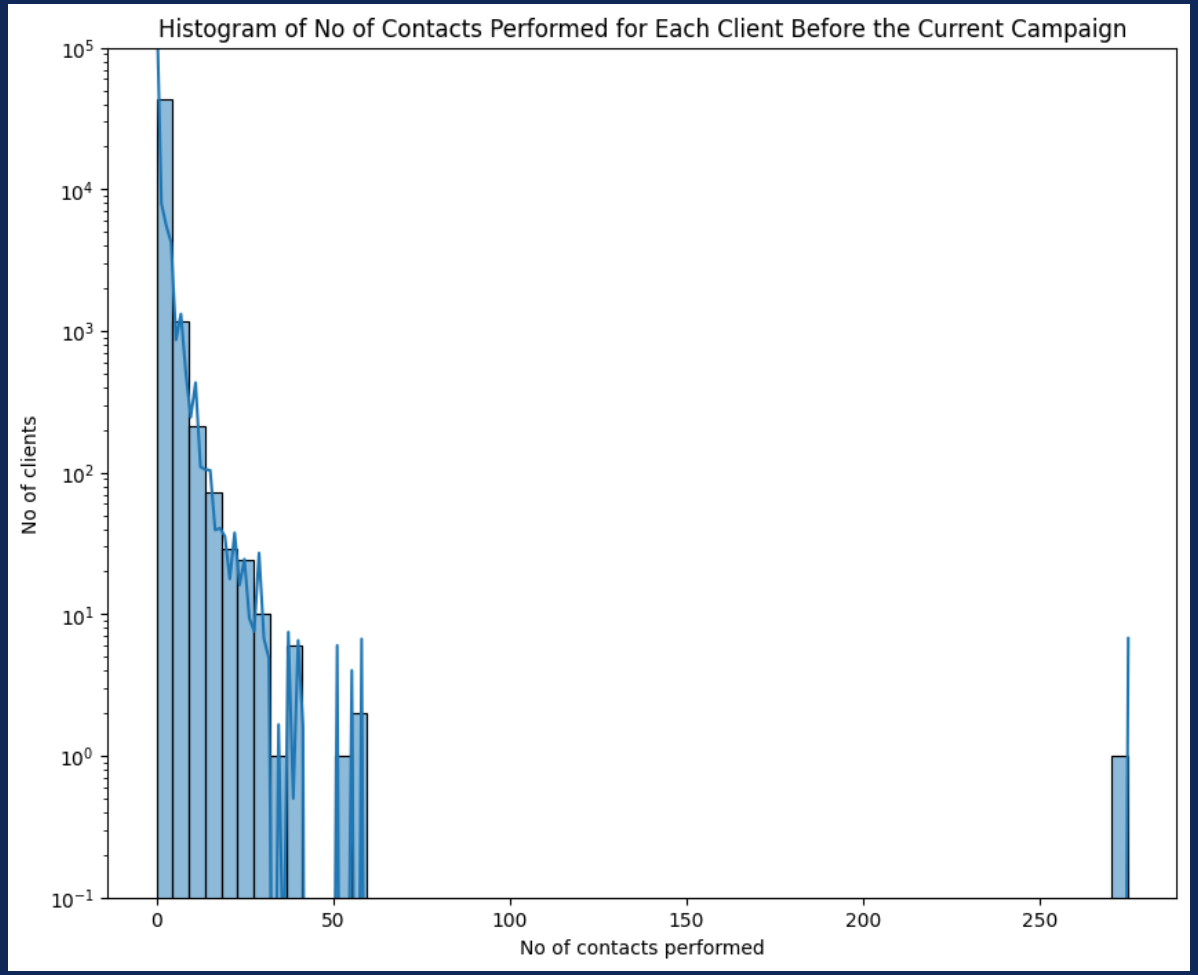
NO OF DAYS PASSED SINCE CLIENT WAS LAST CONTACTED

- This distribution is also highly right skewed. It indicates that most of the clients have been contacted recently or never have been contacted before.
- Median of pdays is -1 i.e. never contacted before, it is also shown through the data that most of the clients (nearly 81.73%) have never been contacted before.



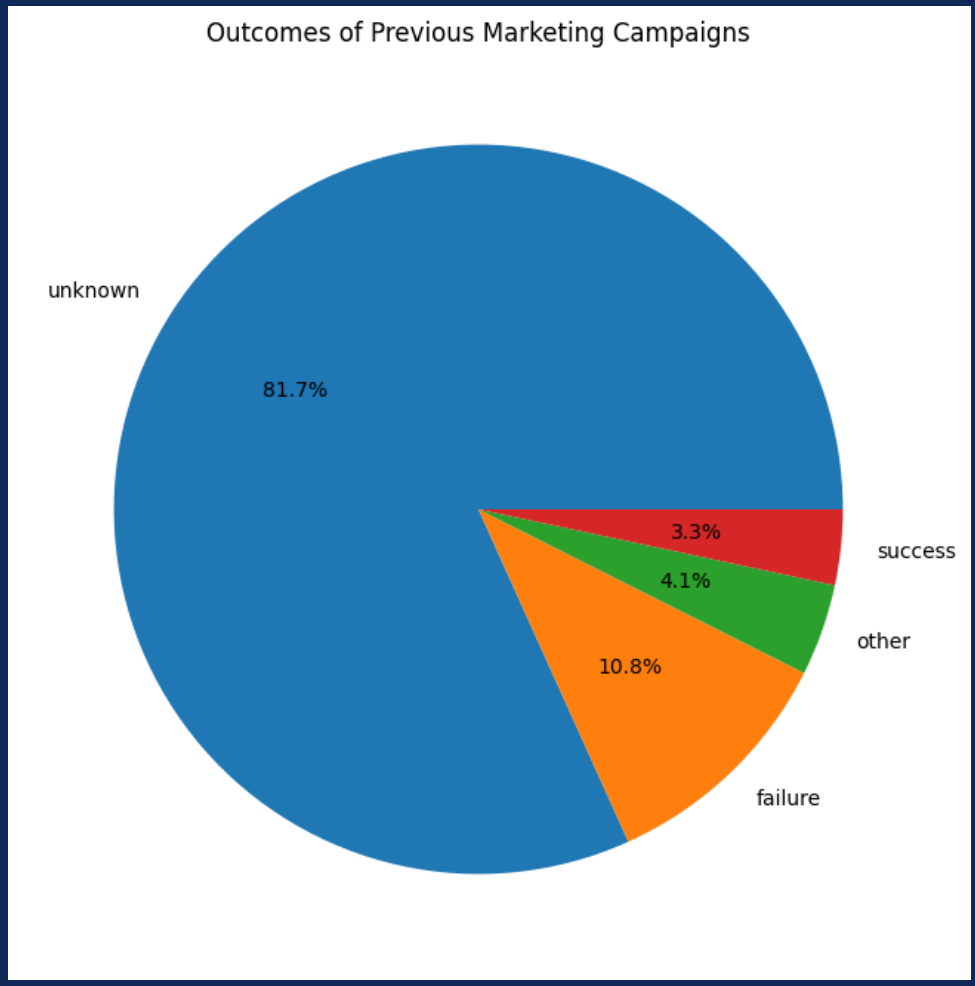
NUMBER OF CONTACTS PERFORMED BEFORE CURRENT CAMPAIGN

- As seen in **pdays** distribution we saw that most of the people (nearly 81.73%) have never been contacted before and so the median of no of **previous** contacts is as expected, i.e., 0.
- This distribution is also highly right skewed indicating only few clients being contacted repeatedly.



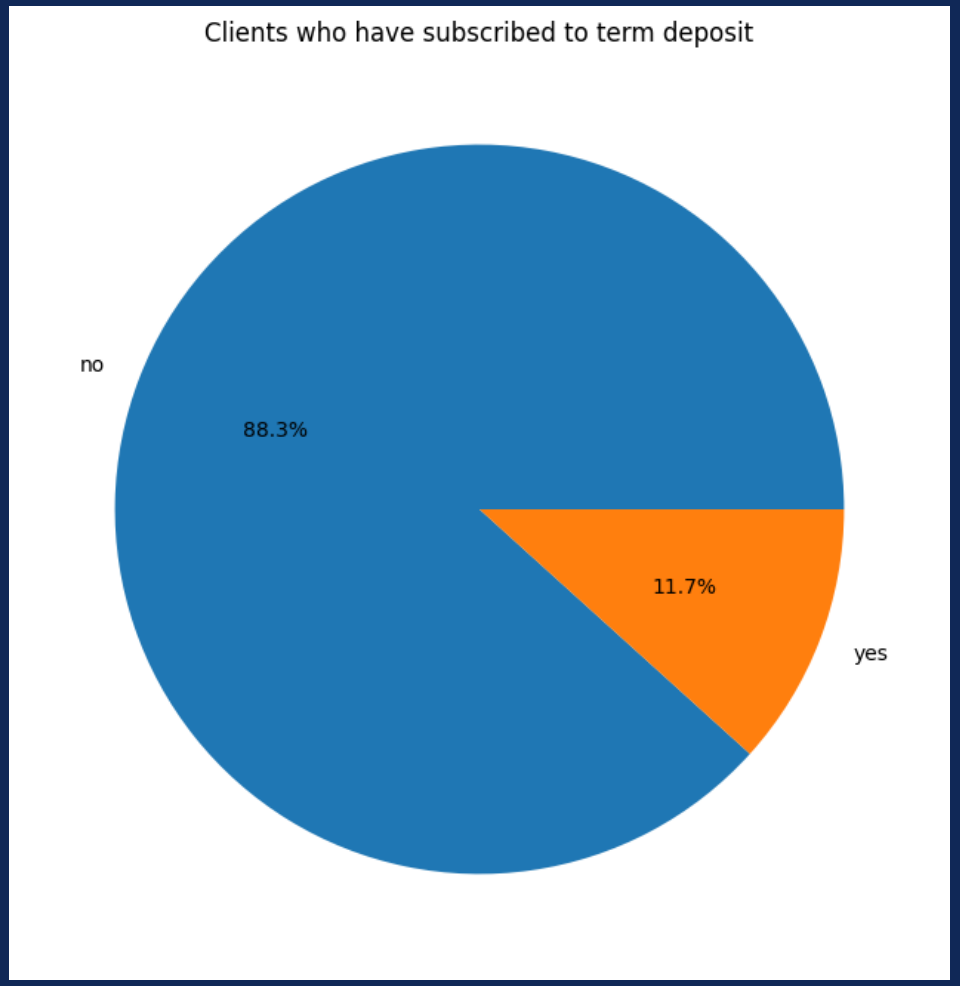
OUTCOMES OF PREVIOUS MARKETING CAMPAIGNS

- Since we saw previously that most of the clients (81.73%) have never been contacted before, so most of their contact outcome is unknown, thus majority of outcomes (81.7%) are unknown.
- Only a small part of outcomes have actual results i.e. failure(10.8%) and success(3.3%)
- 4.1% have other as the outcome indicating neither success nor failure.



DISTRIBUTION OF SUBSCRIPTION TO TERM DEPOSIT

- Only a few clients (11.7%) have subscribed to term deposit.
- While majority of them (88.3%) did not subscribe to a term deposit.



KEY INSIGHTS

- Students show high chances of subscription to term deposit while blue collar and entrepreneur show low chances of subscription.
- Single clients have higher chances of subscribing to term deposit while married clients have lower chances of subscription to term deposit, maybe due to more responsibilities. Divorced clients exhibit relatively high subscription rates, suggesting a greater demand for financial security after divorce.
- Clients with tertiary level of education have higher chance of subscribing to a term deposit than those with secondary level of education. Clients with primary level of education have lowest chance of subscribing.
- Clients which have no credit in default have higher chance in subscribing to term deposit. This shows that default is an important aspect while measuring subscription rates for term deposit.
- Clients having lower balance (<20000) have relatively low chance of subscribing to term deposit (nearly 11.7%).

- Clients which don't have housing loans and personal loans have higher chance in subscribing to term deposit as compared to those which have it. This shows that loan is also an important aspect while measuring subscription rates for term deposit as clients with no loans have more financial freedom.
- Communication through cellular phone can lead to higher chances of subscription of term deposit whereas communication through telephone has lesser chance of subscription.
- March, December, October and September have higher chance of getting subscriptions while rest months have significantly lower chances.
- Clients with lower age have a lower chance of subscribing to a term deposit while clients with age above 60 are more likely to subscribe to a term deposit. It shows that people with higher age are less likely to take risk and thus settle with a lower risk financial option i.e term deposit.

THANK YOU