Week11: EDA presentation for business users

Name: Rayhanul Islam Rumel
Batch Code: LISUM17
Submission Date: 16/03/2023
Submission To: Data Glacier

Name: Amogh Vig
Batch Code: LISUM17
Submission Date: 16/03/2023
Submission To: Data Glacier

Agenda

EDA for Bank Full Data

EDA for Bank Additional Full Data

Model Recommendation

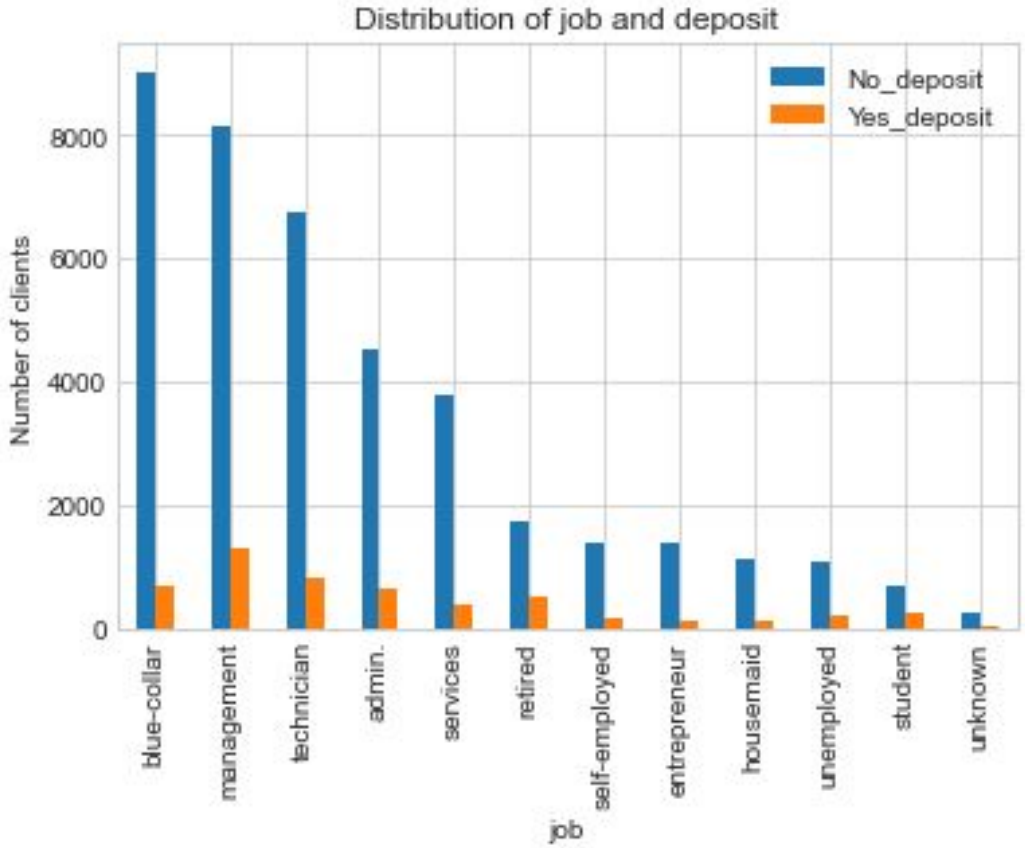
EDA for Bank Full Data

EDA for Bank Full Data

- **Basic Overview of the dataset:**
  - In this dataset, we have in total 45211 observations, and 17 features: 16 input features, 1 output feature
  - There are no duplicate, and/ or null values in the dataset
  - The input features can be classified into three groups: Client Data, Campaign-related data, and other attributes
  - The output feature tells us whether a client has subscribed to a term deposit or not

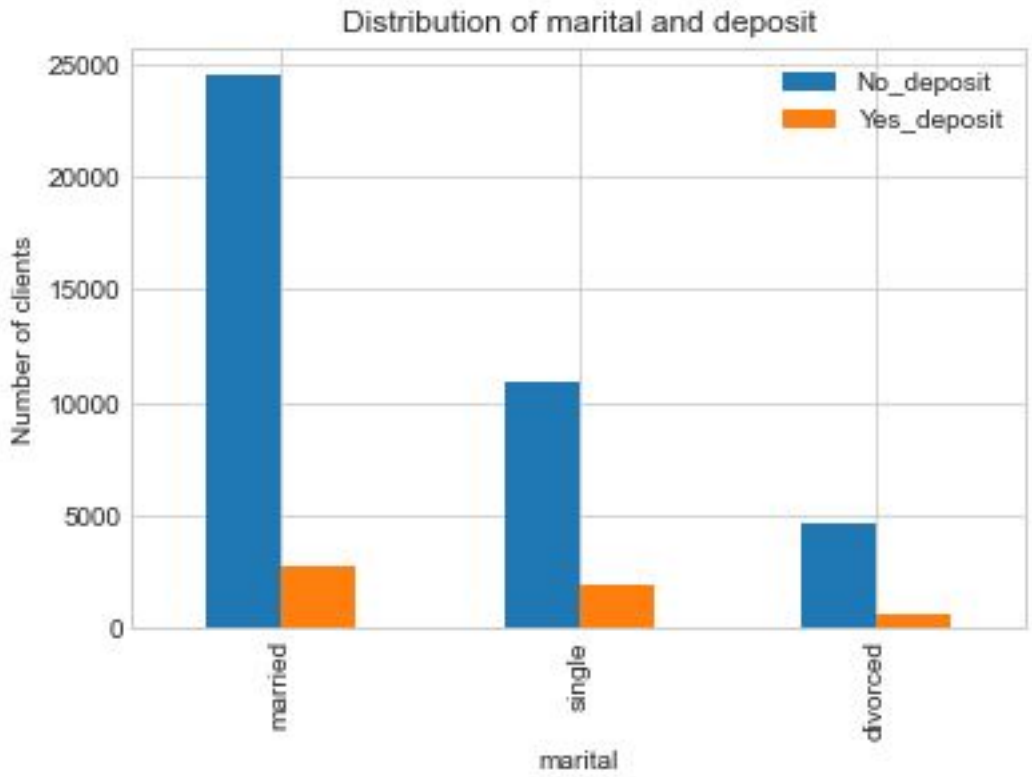
EDA for Bank Full Data

- **Distribution of Jobs and Deposit status of the clients**
- From the visualization, we clearly observe that clients who are in management position purchased the highest number of term deposits
- Second highest purchase came from the clients who are technician by profession
- The highest number of refuse came from the clients who are blue-collar by profession



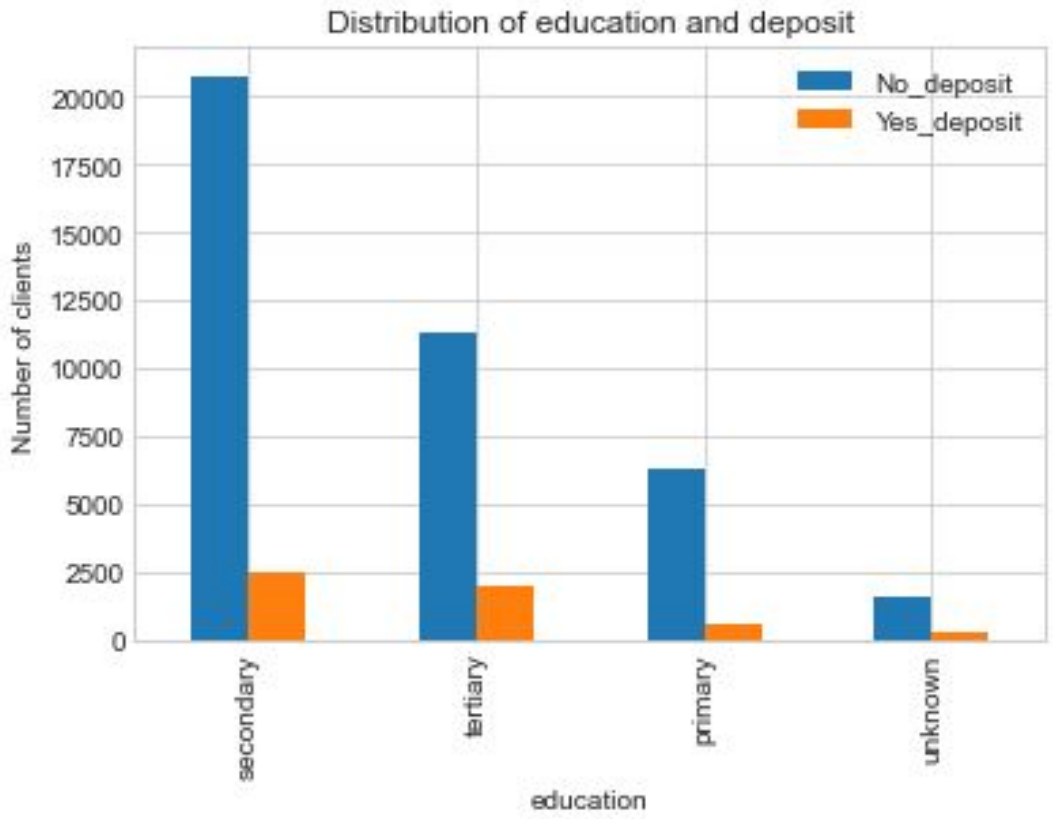
EDA for Bank Full Data

- **Distribution of Marital and Deposit status of the clients**
- From the visualization, we see that clients who are married purchased the highest number of term deposits
- Second highest purchase came from the clients who are single
- The lowest number of purchase came from the clients who are divorced



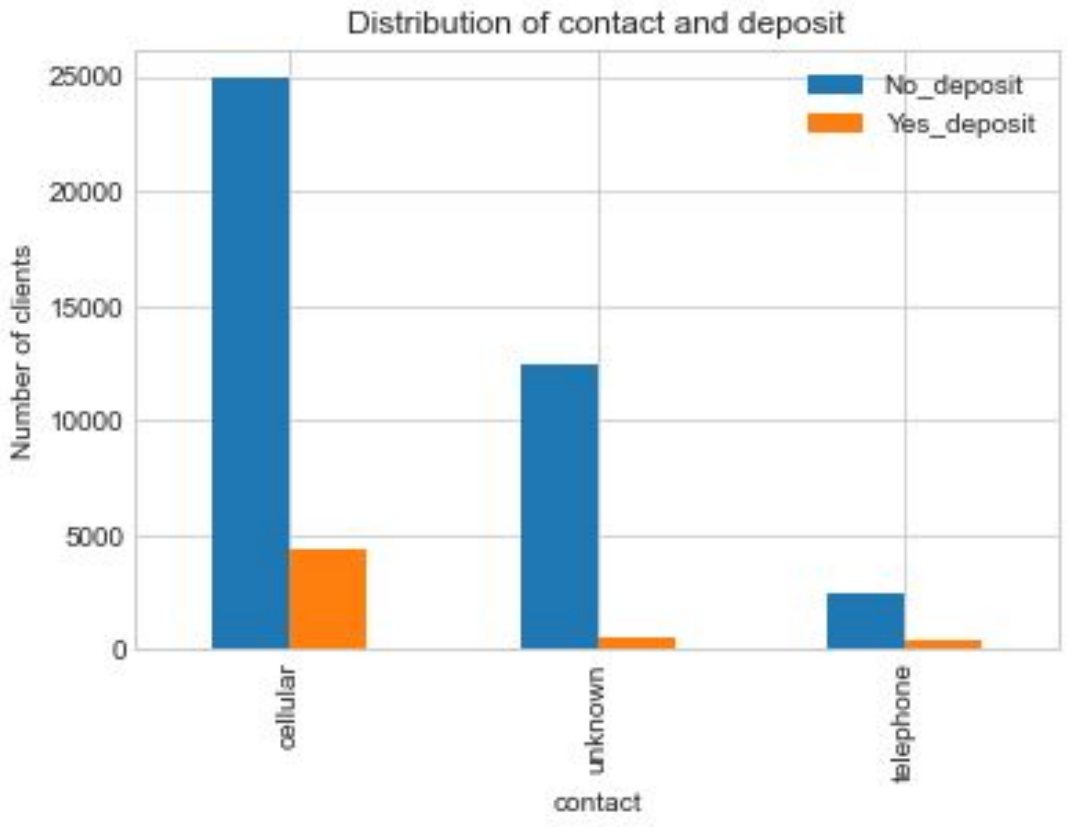
EDA for Bank Full Data

- **Distribution of Education and Deposit status of the clients**
- we see that clients who qualified secondary education purchased the highest number of term deposits
- Second highest purchase came from the clients who hold tertiary education
- The lowest number of purchase came from the clients who hold primary education



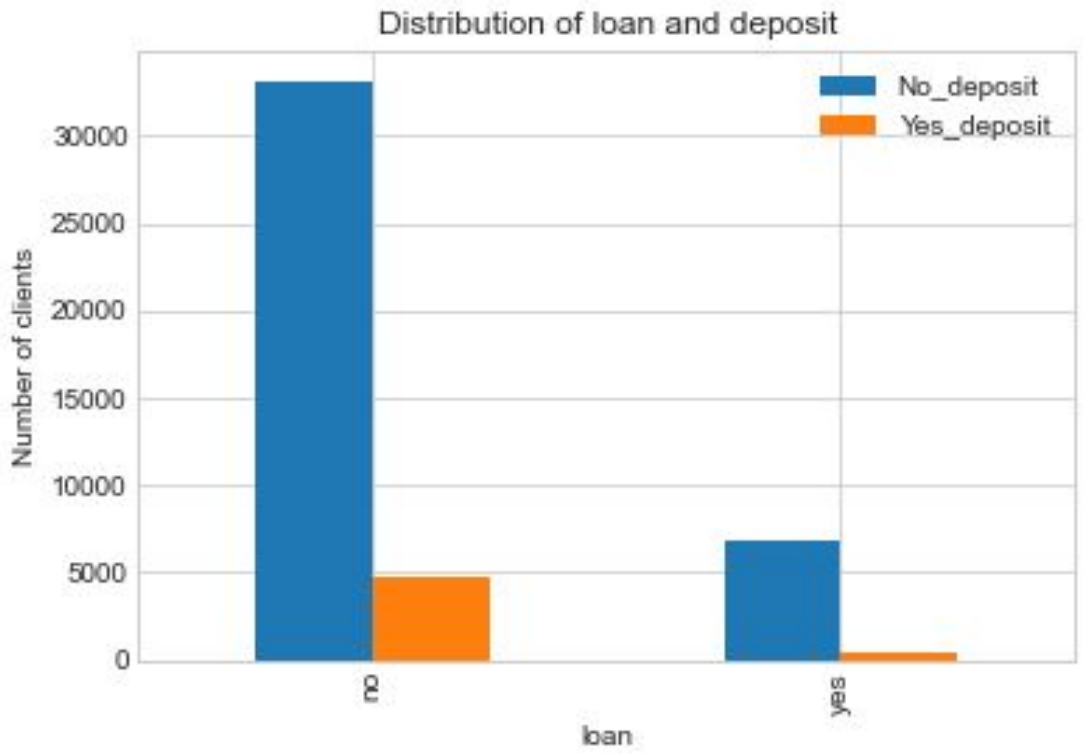
EDA for Bank Full Data

- **Distribution of Contact and Deposit status of the clients**
- we observed that clients who were being contacted via cellular purchased the highest number of term deposits
- Second highest purchase came from the clients whose contact source are unknown
- The lowest number of purchase came from the clients who are being contact via telephone



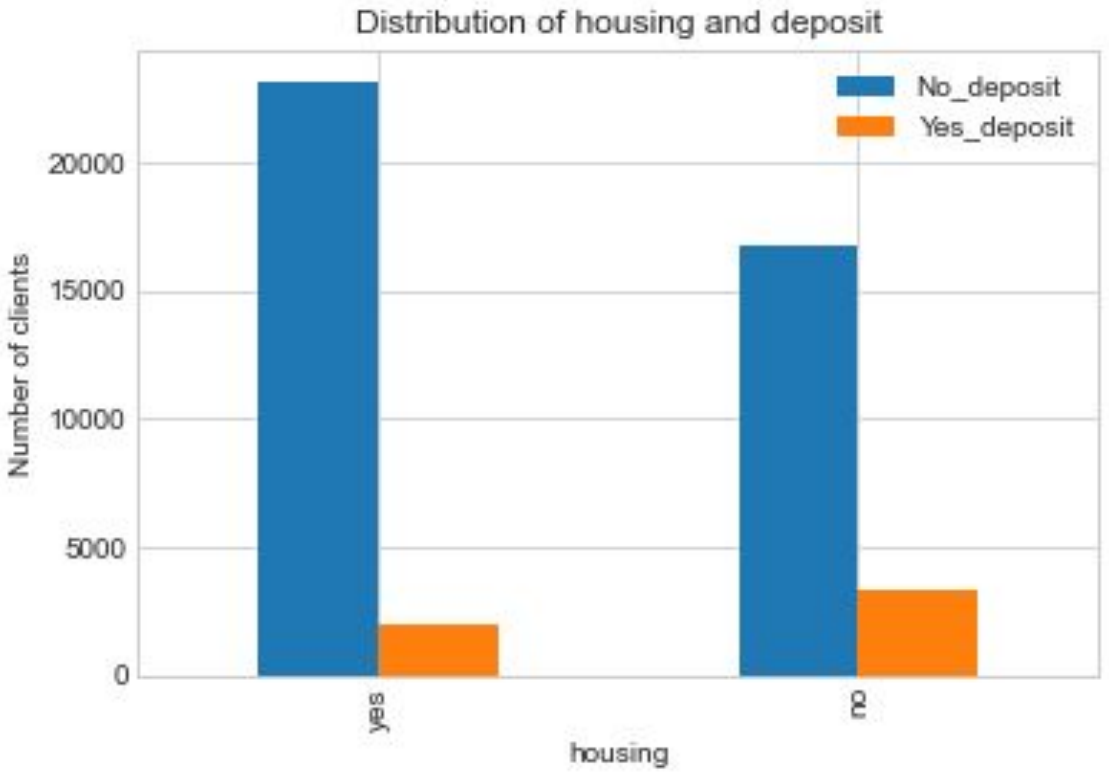
EDA for Bank Full Data

- **Distribution of Loan and Deposit status of the clients**
- we observed that clients with no loan purchased higher number of term deposits than clients with loan



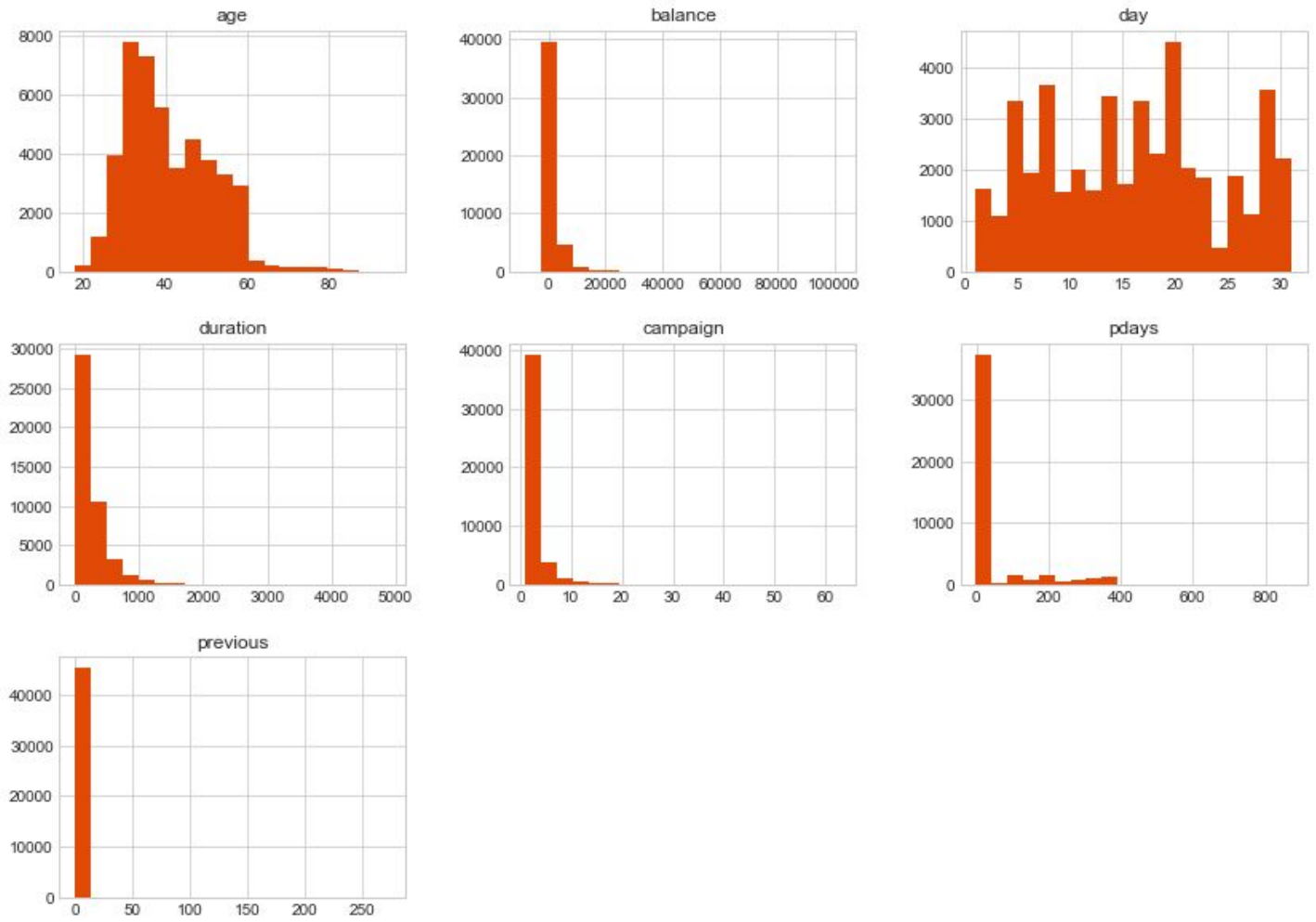
EDA for Bank Full Data

- **Distribution of Housing and Deposit status of the clients**
- We observed that clients with no housing loans purchased higher number of term deposits than clients with housing loans.



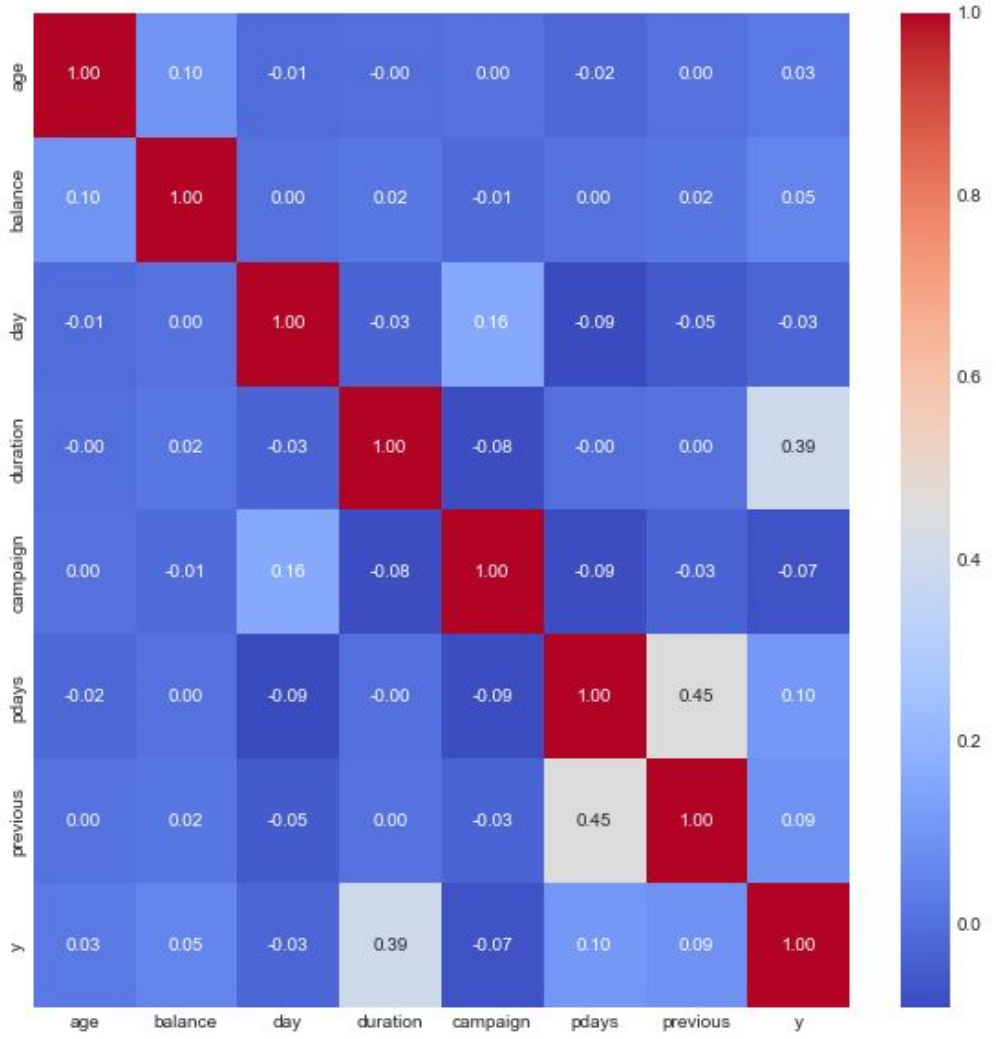
EDA for Bank Full Data

- **Distribution of Numeric data**
- Except Age and Day, the distribution of all other data is positive skewed or right skewed distribution
- Since we would be expecting all of our data to be normally distributed, we can use IQR method to deal with the outliers of the skewed distributions



EDA for Bank Full Data

- **Distribution of Numeric data**
- In the correlation chart, we really don't really observe any two features that are strongly/ significantly correlated



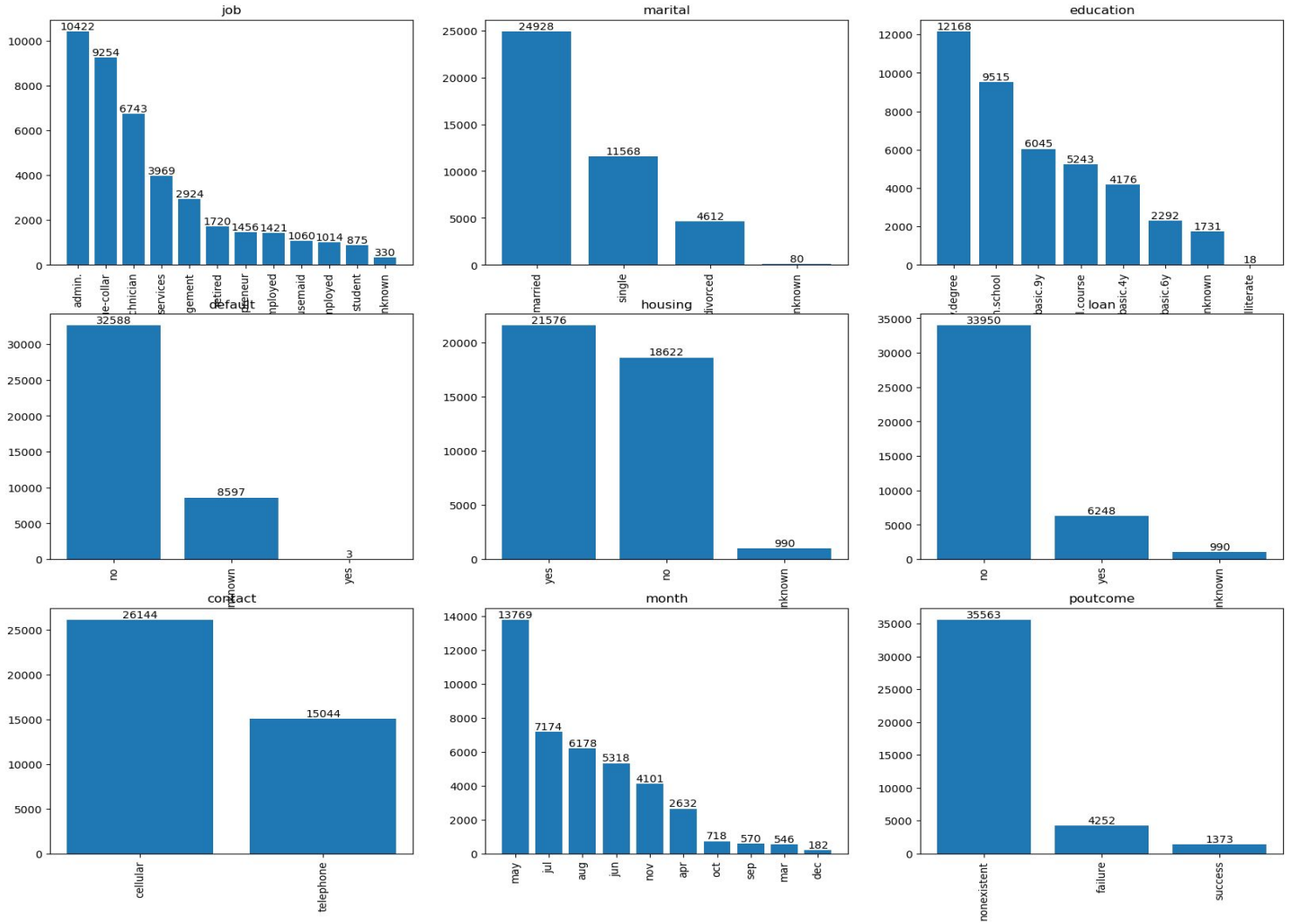
EDA for Bank Additional Full Data

EDA for Bank Additional Full Data

- **Basic Overview of the dataset:**
  - In this dataset, we have in total 41188 observations, and 21 features: 20 input/ independent features, 1 output/ dependent feature
  - There are no duplicate, and/ or null values in the dataset
  - The independent features can be classified into three groups: Client Data, Campaign-related data, and other attributes
  - The output feature tells us whether a client has subscribed to a term deposit or not

EDA for Bank Additional Full Data

- **Categorical Feature Exploration**



EDA for Bank Additional Full Data

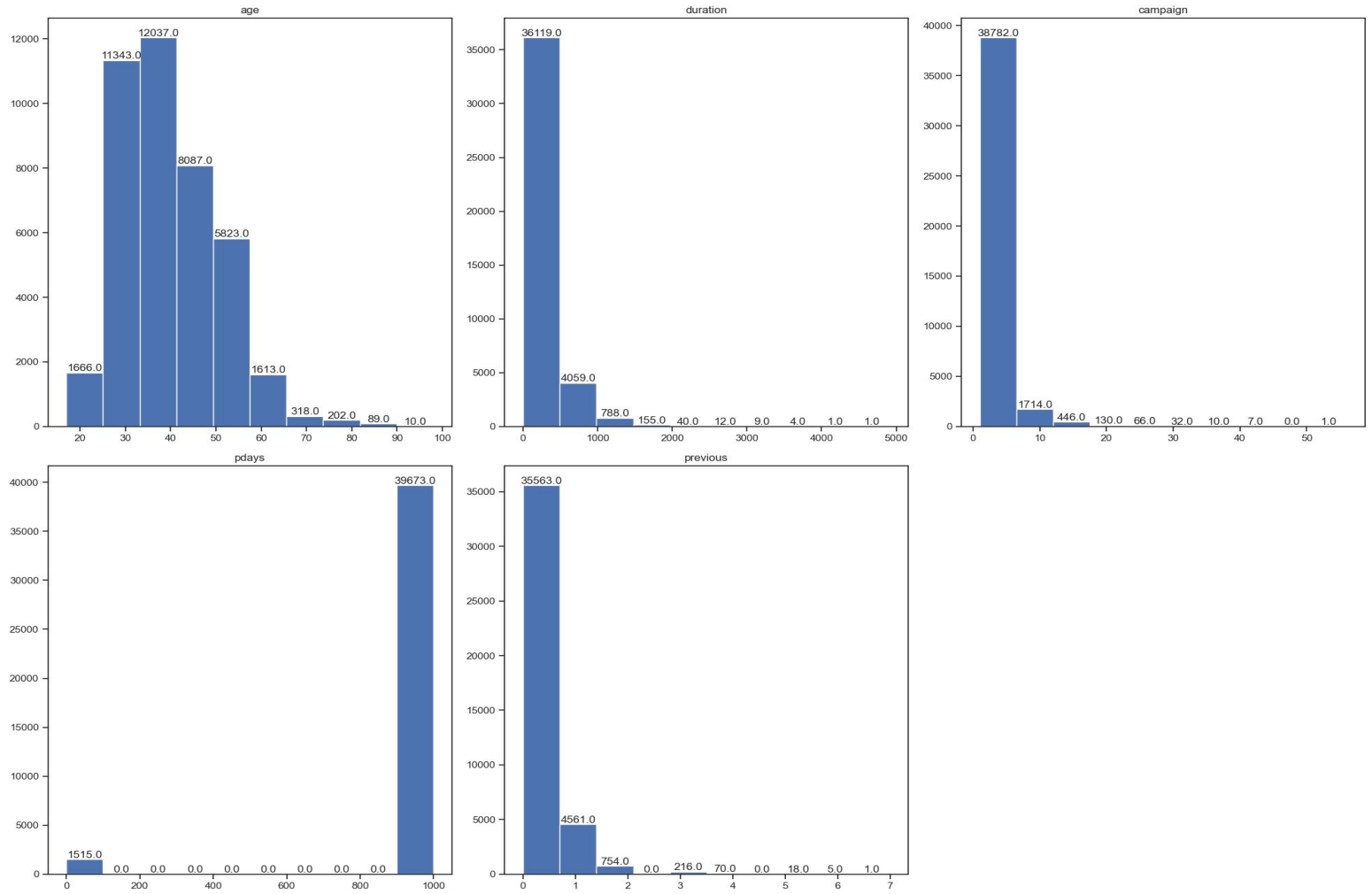
- **Categorical Feature Exploration**
- **Job:** Majority of bank customers belong to admin, blue-collar, and technician categories respectively. Admin category has the largest number of customers, followed by blue-collar and technician.
- **Marital:** Majority of customers in the dataset are classified as married, while the number of customers who are divorced or widowed is relatively small.
- **Education:** A large proportion of customers have obtained university degrees. Default: Majority of customers have a clean credit history, with no instances of defaulting on credit payments.
- **Housing:** Difference in proportions between customers who have a housing loan and those who do not is relatively small. Number of customers who have a housing loan is greater than those who do not have one. About 990 people have unknown information.

EDA for Bank Additional Full Data

- **Categorical Feature Exploration**
- **Loan:** Number of customers who have loans other than housing loans is significantly lower than those who have housing loans along with other types of loans.
- **Contact:** Most frequently used method of telemarketing campaign was to contact customers through their cellular telephones.
- **Month:** May had the highest number of customer contacts, presumably as part of the most recent marketing campaign.
- **Poutcome:** Proportion of customers who subscribed to term deposits from the previous marketing campaign was the lowest, whereas the highest proportion of customers had a nonexistent subscription status.

EDA for Bank Additional Full Data

- **Numerical Feature Exploration**

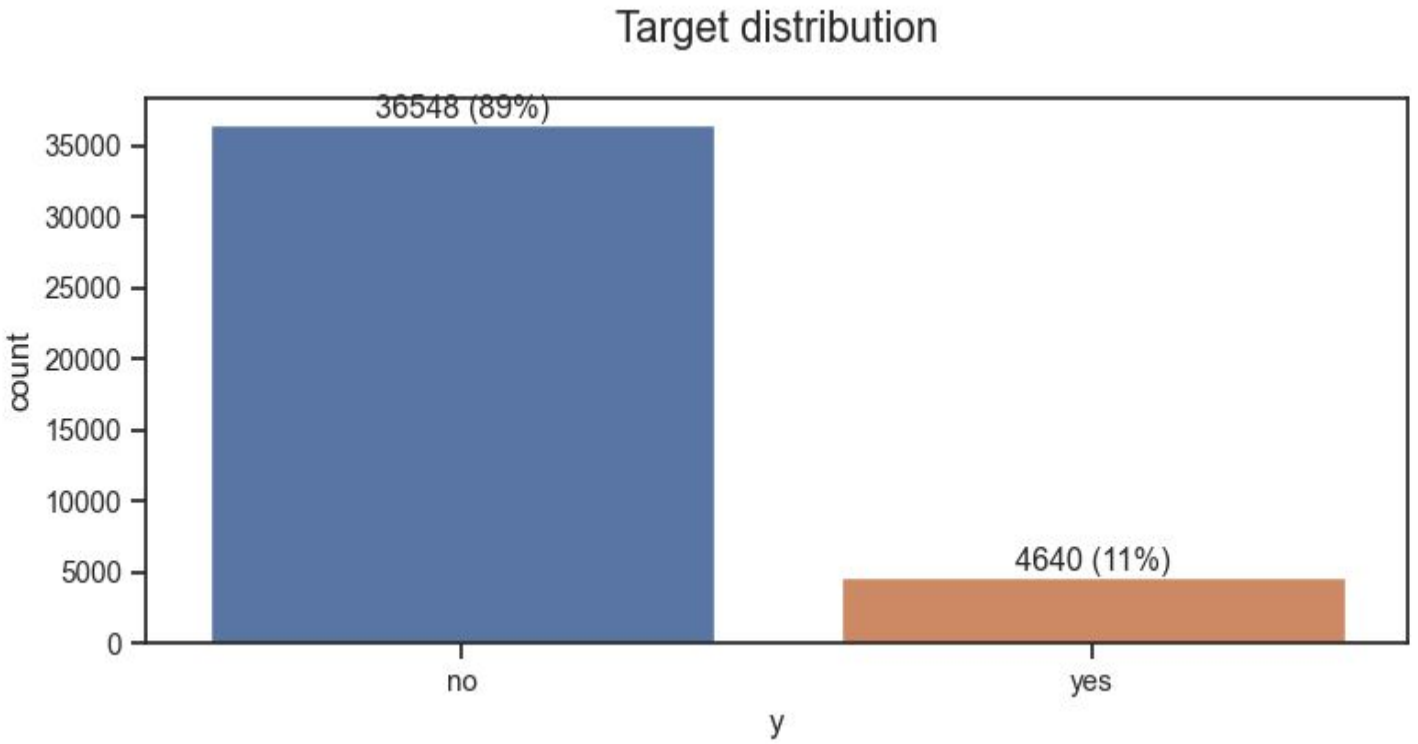


EDA for Bank Additional Full Data

- **Numerical Feature Exploration**
- **Age:** Largest customer segment falls within the age range of 30-50 years old. Occupation with youngest distribution of ages is "student", whereas occupation with oldest distribution of ages is "retired".
- **Duration:** Maximum duration of last contact was 36119 seconds.
- **Campaign:** Total of 38782 contacts were made during the course of this marketing campaign.
- **Pdays:** Highest number of days elapsed since a client was last contacted from a previous campaign was 39673, while the lowest number was 1515 days.
- **Previous:** Highest number of recorded contacts, totaling 35563, was made prior to the current marketing campaign.

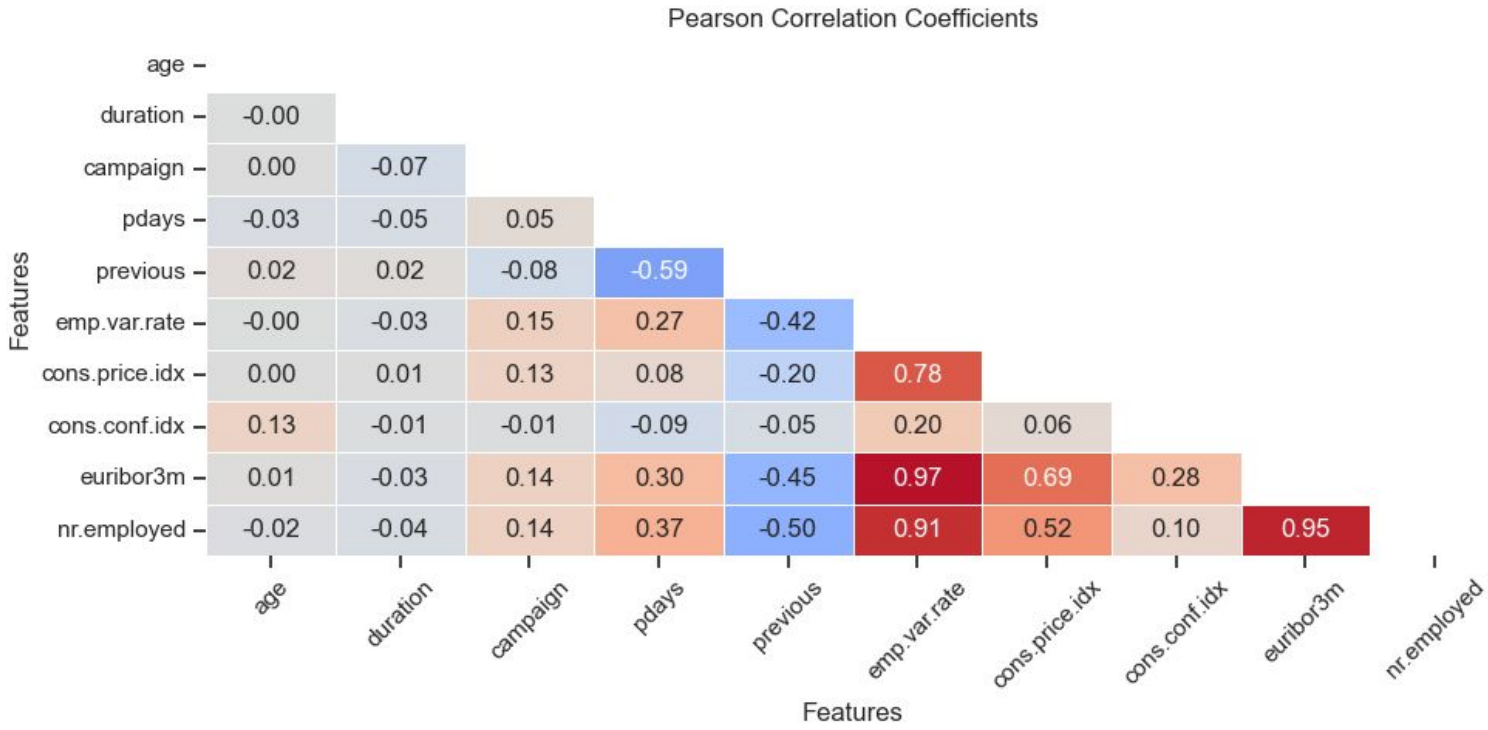
EDA for Bank Additional Full Data

- **Distribution of Housing and Deposit status of the clients**
- Number of customers who declined the term deposit offer in the current campaign is significantly greater (89%) than the number of customers who accepted it (11%). Only a few thousand customers accepted the offer.



EDA for Bank Additional Full Data

- **Distribution of Housing and Deposit status of the clients**
- From the chart, we observe that there is a strong correlation between euribor3m and emp.var.rate (.97), nr.employed and euribor3m (.95), and nr.employed and emp.var.rate (.91). All of these correlations are positive.



EDA for Bank Additional Full Data

- **Outliers in the dataset**
- Outliers are present in all of the data
- The interquartile range for the numeric variables, except for age, is relatively narrow
- Since we would be expecting all of our data to be normally distributed, we can use IQR method to deal with the outliers of the skewed distributions

Model Recommendation (For Technical users only)

Model Recommendation (For Technical Users only)

- **Model Recommendation (For Technical Users only)**
- After doing EDA on both bank-full-data and bank-additional-full-data, we would suggest to use classification algorithms to consider as the model for the bank.
- A few classification models which we can consider in our project are mentioned below:
    - Logistic Regression
    - K Nearest Neighbour Classifier
    - Random Forest Classifier
    - Gradient-boosting Classifier

Thank You