Project Title: Data Science:: Bank Marketing (Campaign) -- Group Project
Date: January 25, 2024

Group Name: Data Department_1
- **Name**: Minseok Kim
- **Email**: mxk230041@utdallas.edu
- **Country**: United States
- **College/Company**: The University of Texas at Dallas
- **Specialization**: Data Science

Problem Description:
The project aims to develop a predictive model to identify potential customers for a bank's term deposit product. Challenges include addressing data imbalances and extracting actionable insights from customer demographics and past interactions with the bank.

Data Understanding:
The dataset comprises demographic details, financial behaviors, and contact information from past marketing campaigns. Key variables include job type, education level, housing loans, and contact outcomes.

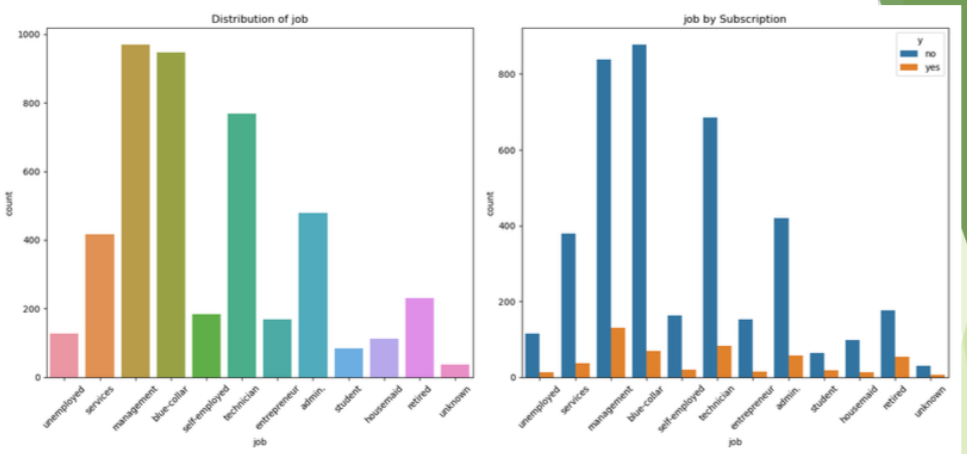
Data Challenges and Approaches:

NA Values: Imputation strategies, like using median values for continuous variables or mode for categorical ones, are considered to maintain data integrity.
Outliers: Robust scaling techniques or transformations are planned to mitigate their impact.
Skewness: Data normalization or log transformations to stabilize variance across the dataset.

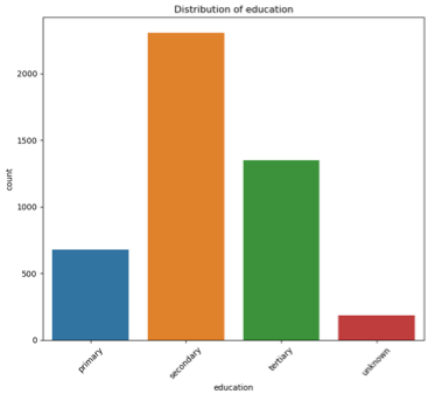
The data has revealed meaningful insights, such as management and technicians being more likely to subscribe, and those with housing loans showing less propensity to subscribe. Specific age groups and education levels also correlate with subscription likelihood. The aim is to apply ML techniques that can handle these nuances in the dataset effectively.

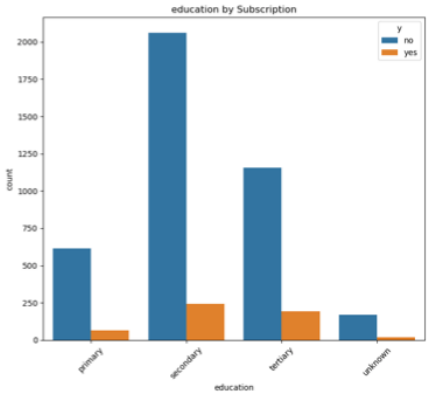
GitHub Repository: https://github.com/N0VA-code/Data-Glacier-Data-Based-Consulting-Project.git

BankMarketting_ML project_EDA

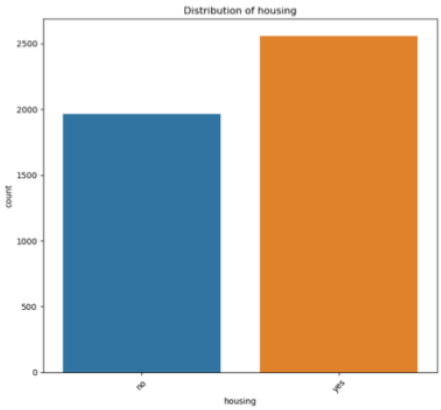


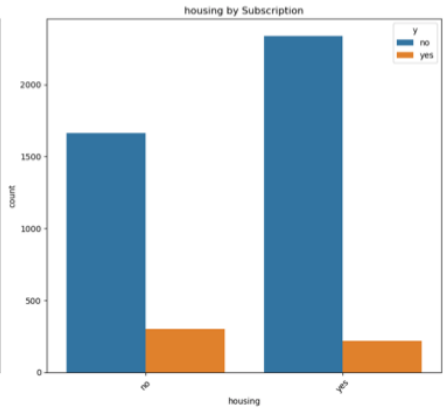
Observation: Management, Blue-collar, Technician (3 positions) were the majority (58%) by following admin and service position. However, even tough there almost same number of people in management and blue-collar section, more people in management occupation has reported Y(yes) to the subscription by almost double. Which means there are more chances of selling our product to management, than clue-collar despite the total number is almost the same. Additionally In terms of proportion,



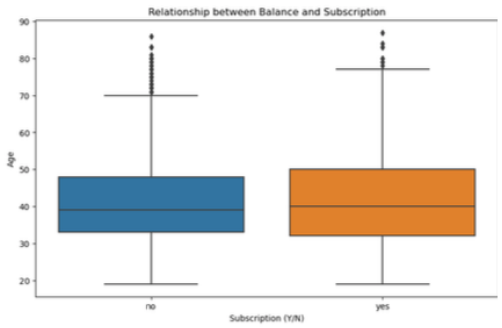


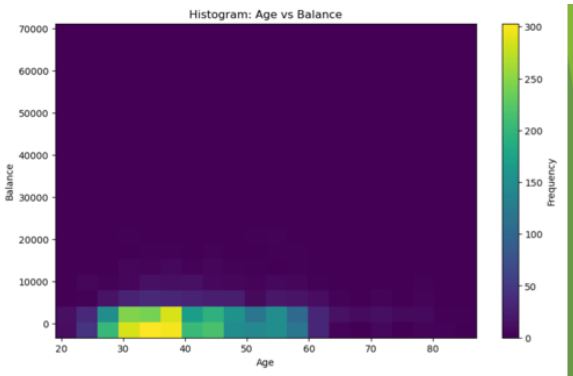
Observation: People who completed tertiary education has reported Y the most(proportionally), and secondary(in numbers). Even tough there's a difference in numbers (Between Secondary and Tertiary)by double, tertiary education has reported almost same numbers of Y s. In conclusion, our main consumers are most likely to be Tertiary(and Secondary)



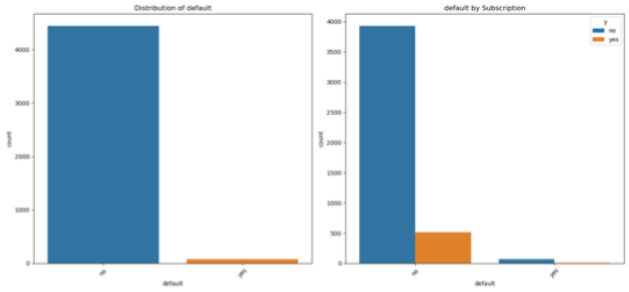


Observation: 56% of total replied they have housing loan, and the others replied they don't. However more people who do not have housing loan subscribed(proportionally). In conclusion, People who do not have housing loan are more likely to subscribe than those people who do. But due to the little numeric relationship, housing factor should not be considered rigorously as others. (Related but not as much as others are)

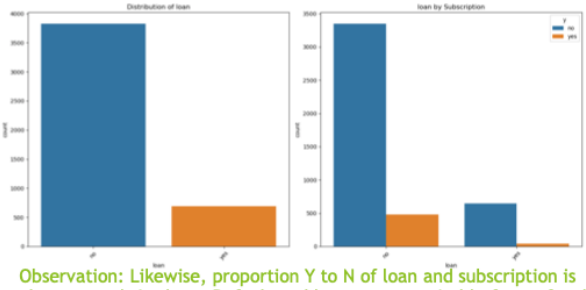




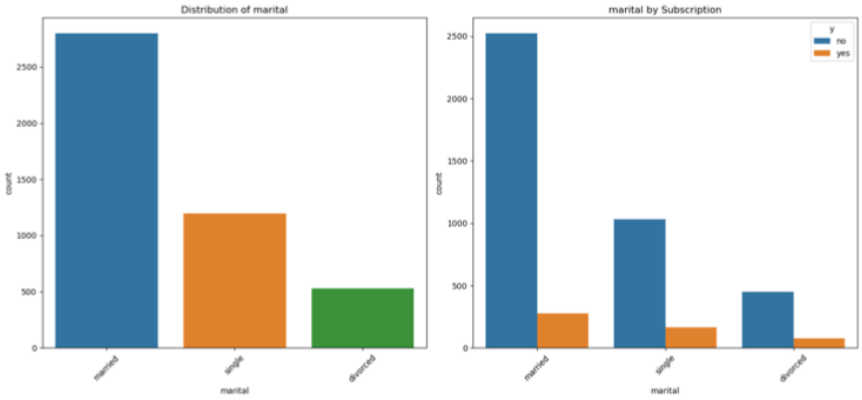
Observation: As it sees, people with higher balance would subscribe more, and has slightly higher median age of early 40s. And the chart on the right-hand side says mid 30s to late 30s has the most in both balances and numbers. So mid 30s to early 40s would fit the most for our customers.



Observation: Number of people who has credit in default is too low to take account and find the relationship with subscription(Y/N)

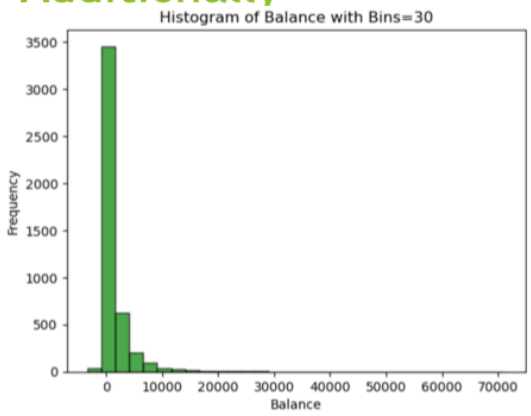


Observation: Likewise, proportion Y to N of loan and subscription is almost equal. In short, Default and loan are not suitable factor for the ML model



Observation: I see very little correlation between marital status and subscription. So marital status is not impactful to subscription, but I can see there's a VERRY LITTLE chance of selling out product to single than married.

Additionally



Additionally, since the most of population has a balance of less than 1500, the minimum for the bank term deposit should be lower than that