Text

Description automatically generated

**Data Management and Big Data - ALY6110**

**Basic Analysis and Visualizations**

**Final Project**

**Group 5:**

Himanshu Kumar

Zihan Ma

**Date:** 1st Dec’2023

**Professor:** Dr. Hema Seshadri

**About the Dataset:**

***(Source:*** *https://www.kaggle.com/datasets/krantiswalke/bankfullcsv?select=bank-full.csv****)***

The dataset is taken from Kaggle and has a total of *45,211* entries, each entry representing an individual record. It comprises 17 variables providing a range of demographic and behavioral data about the individuals. The dataset essentially captures information related to banking clients and their interaction with marketing campaigns, with a particular emphasis on term deposit subscriptions. It details the personal demographics, financial situation, and banking-related behaviors of clients, with variables including age, job, marital status, education level, whether they have defaulted on credit, their balance, whether they have a housing or personal loan, and more.

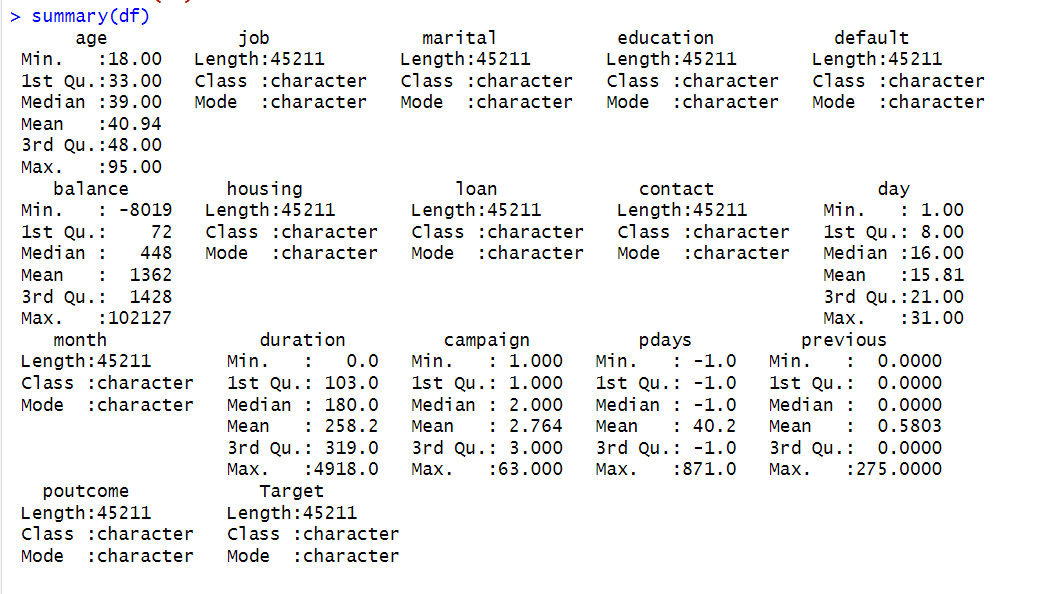
Additionally, it includes data on the method and frequency of contact with clients, as well as information on previous marketing campaigns and their outcomes. The 'Target' variable indicates whether a client has subscribed to a term deposit or not, serving as a key outcome measure. In essence, the dataset provides a comprehensive picture of banking clients' profiles, their financial behaviors, and their responses to banking marketing campaigns. This data could potentially be utilized to optimize future marketing strategies, improve customer service, and enhance financial product offerings.

**Variable Characteristics:**

The dataset has 17 variables in total out of which 10 are categorical variables and 7 are numerical variables, the split is as follows:

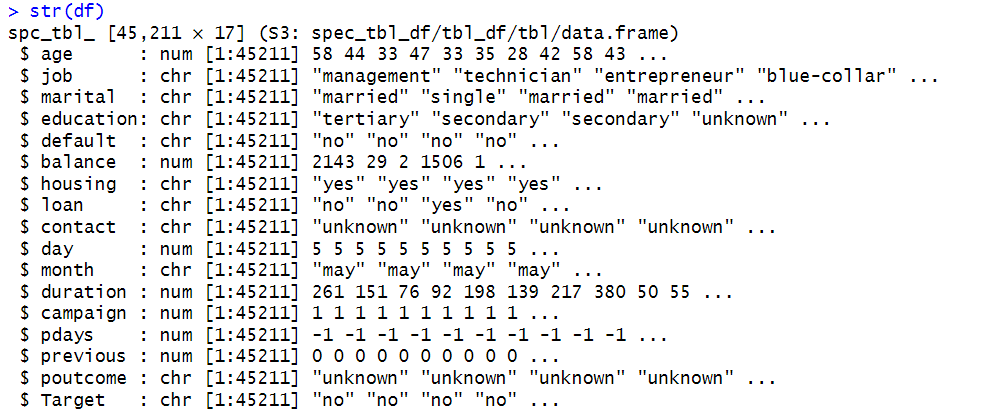
* ***Categorical Variables:*** Job, Marital, Education, Default, Housing, Loan, Contact, Month, Poutcome, and Target.
* ***Numerical Variables:*** *Age, Balance, Day, Duration, Campaign, Pdays, and Previous.*

**Data Exploration:**

****

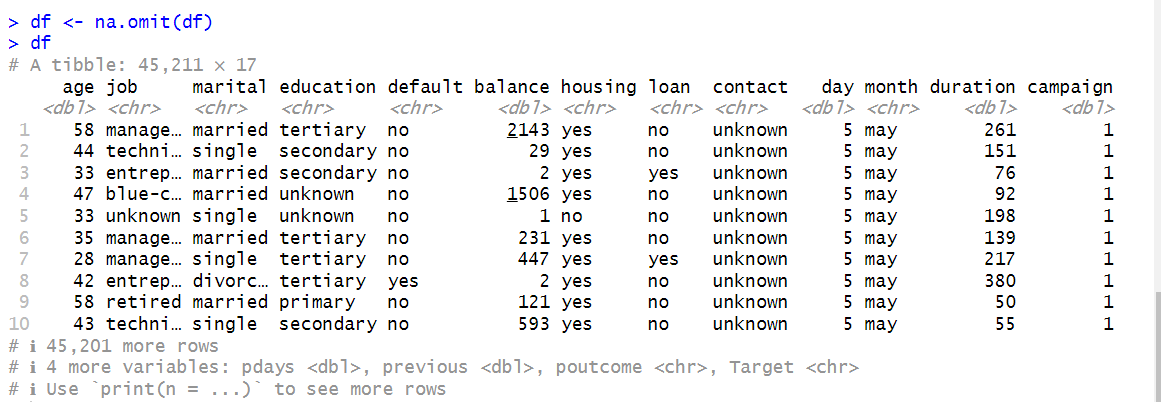
For the exploration of the dataset, the data was analyzed using R. The initial step involved loading the dataset head and gathering relevant information. Subsequently, descriptive statistics were performed to gain insights into the dataset. This involved examining measures such as mean, median, mode, and range to summarize the central tendency and spread of the data. Additionally, the standard deviation was calculated to understand the dispersion or variability within the dataset. These exploratory analyses were crucial in understanding the characteristics and distribution of the data, allowing for a thorough examination of the dataset. Overall, these steps laid the foundation for further analysis.

**Data Entries:**

****

The dataset, sourced from Kaggle, includes a substantial number of 45,211 entries, where each entry signifies a unique record. It encompasses 17 different variables, each offering various demographic and behavioral information about the individuals in question.

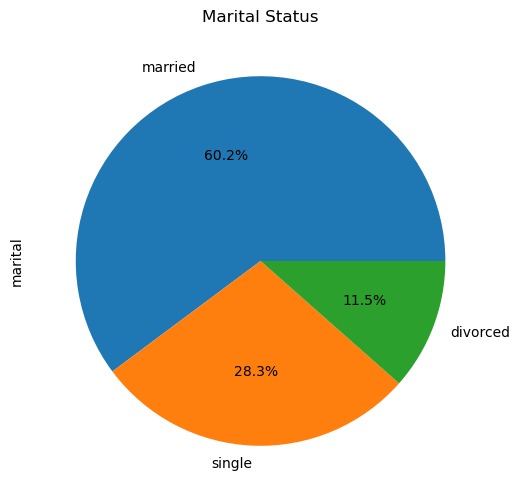
**Data Cleaning and Duplicate Values:**

****

The dataset we analyzed was very clean and well-organized. There were no missing values, which means we had all the necessary information for our analysis. This is important because missing data can cause problems and lead to inaccurate results. We also didn't find any duplicated entries, so each piece of information was unique and didn't repeat. This is good because it means we didn't have any unnecessary repetition in our dataset. Having clean and complete data will make our analysis more reliable and help us draw accurate conclusions from the dataset.

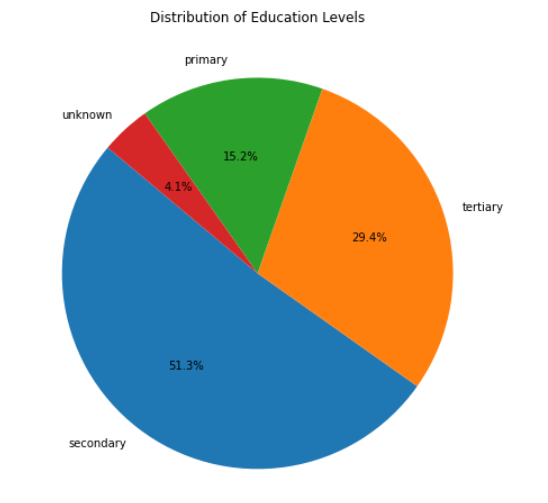
**Data Visualizations:**

**Pie Chart of Marital Status**



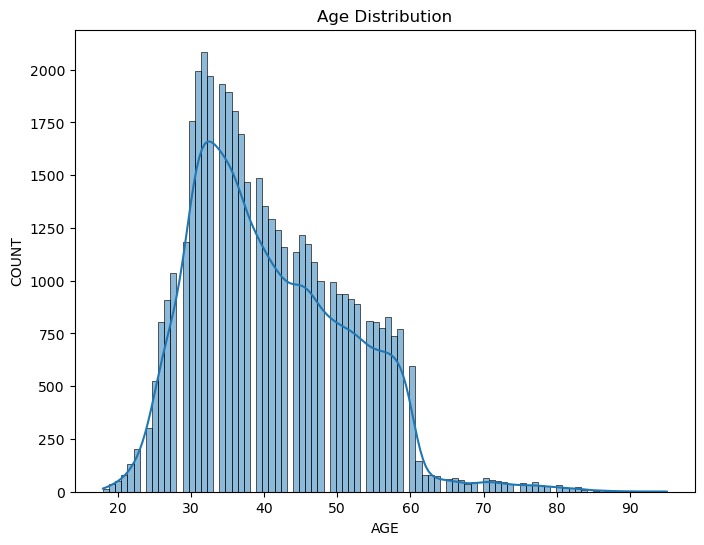
The above pie chart shows the marital status of all the customers who were contacted by the bank during their campaign. 60% of the customers are married while 28% are single. Also, 11% of the customers have their marital status as divorced. This breakup of marital status will help the bank to customize the services according to the needs of the customers such as type of loan and financial budget etc.

**Pie Chart of Education Level**



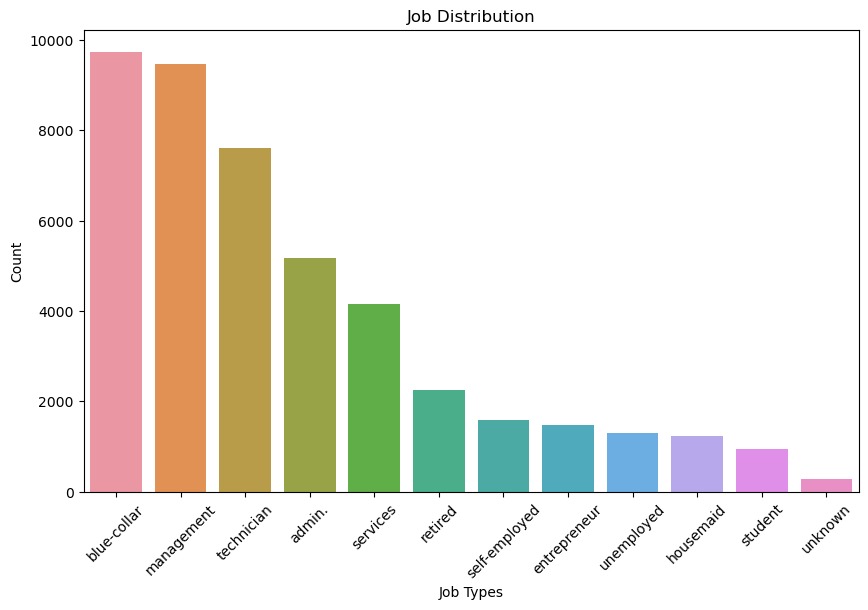
The above pie chart shows the education level of all the customers who were contacted by the bank during their campaign. 51% of the customers hold a secondary level of education while 29% of them have a tertiary level of education. Also, 15% of the customers have basic primary education while 4% of customers have their education level as unknown.

**Histogram of Count concerning age distribution**



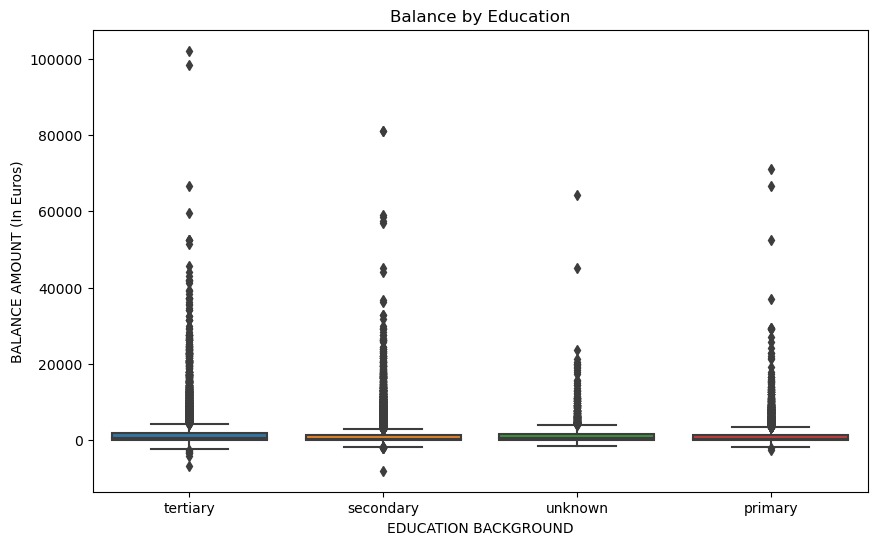
The above histogram shows how many clients are in different age groups. Most of the clients in the dataset are between 25 and 55 years old. This will help the bank to segment the target audience where there are the most clients and how to customize their services to enhance customer experience.

**Bar Plot of Count concerning Job Type**



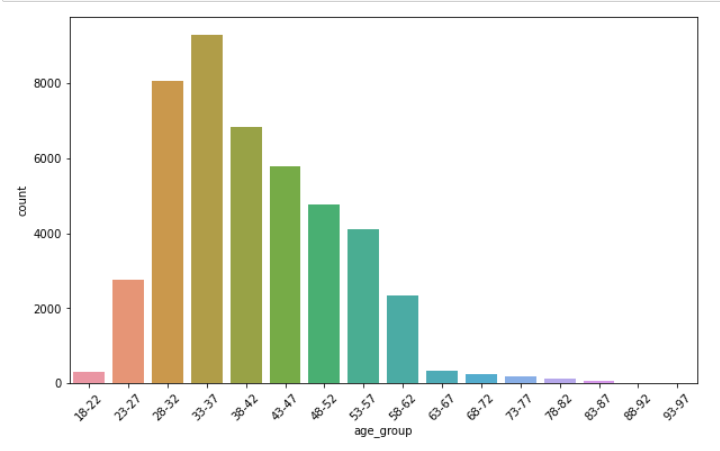
The above bar plot shows the breakup of occupation within 12 different categories. Most of the people do blue-collar jobs which count to around 9900 in number followed by customers practicing management in their jobs with almost the same number. The data has a good diversity of customers falling into different categories who have different financial backgrounds. The bank can use this information in shaping their campaigns based on the occupation followed by an individual.

**Box Plot of Balance Amount v/s Education Background**



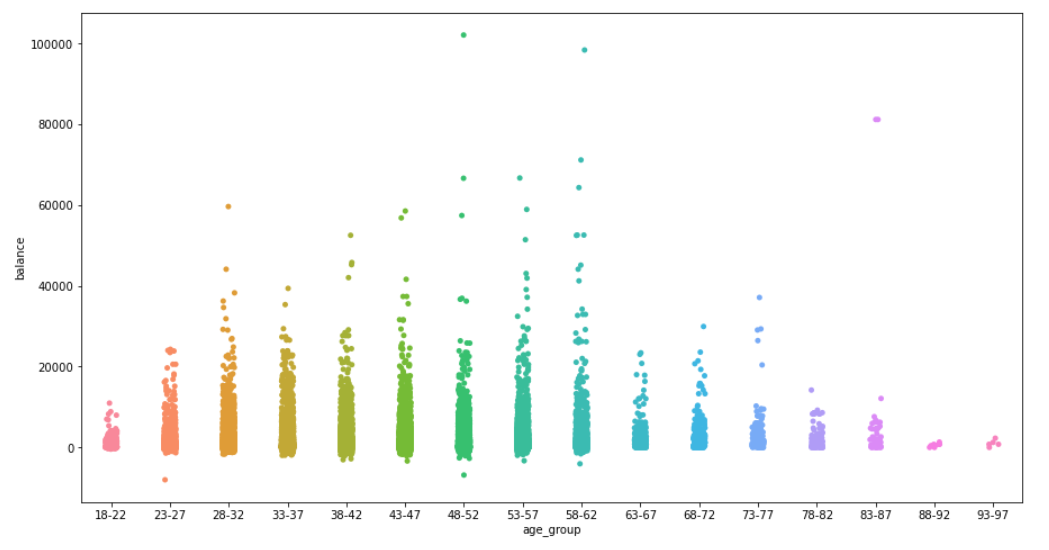
The boxplots depicting education categories (primary, secondary, tertiary, and unknown) against client balances reveal the presence of outliers in each group. These outliers represent individuals with exceptionally high or low balances based on their respective education levels. The boxplots effectively highlight the variations in balances within each education category, showcasing instances where certain clients have significantly different account balances compared to the majority within their educational group.

**Bar Chart of Customers (Age wise)**



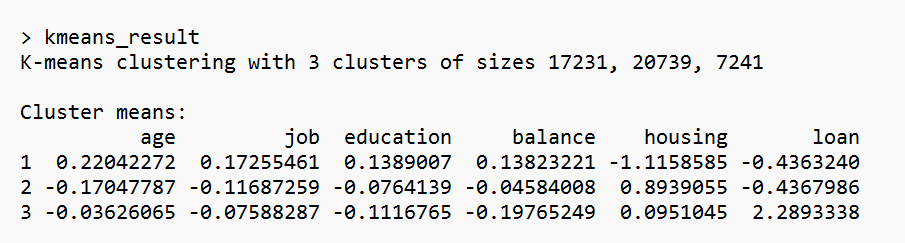
The above bar chart shows the age groups of the customers in the bank. The majority of the customers who were contacted during the marketing campaign belong to the age group of 33 to 37 years. This is a very important visualization from the bank’s perspective as they may segregate their customers based on the targeted campaign which may have a huge impact on that particular section.

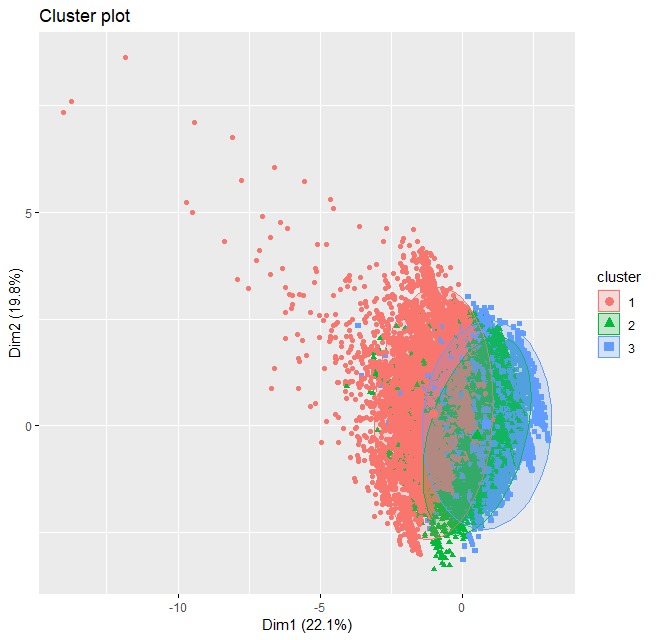
**Scatter Plot of Age Group and Balance**



Continuing the age group bar graph from the previous figure this scatter plot bar chart represents the age group of the customers on the x-axis and their bank balance on the y-axis. It can be seen that the bank balance is on the higher side by the age group of people belonging to 48 to 62 years. This means that bank can target their age group for their campaigns or schemes.

**Cluster Plot:**

****



So using the clustering method through R, we have developed three categories of customers based on variables like age, job, education, balance, housing, and loan. Each category represents the likelihood of a customer enrolling for the term deposit scheme that the bank is marketing to its clients. The customers who have the red dots are more prone to be enrolled in the bank scheme as they maintain a high level of balance and belong to a median age range. Customers with green dots depicting the borderline which can rest at the mid-level in the targeting funnel of the bank. The least likely customers are represented through blue dots, which the bank can omit from their list and can save their cost of marketing.