

# **Analysis of Data Related to Direct Marketing Campaigns(Phone Calls) of a Banking Institution**

## **By Ebenezer Omari**

### **Descriptive Introduction**

The data is related to direct marketing campaigns (phone calls) of a banking institution. The data is related with direct marketing campaigns of a banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

The main objectives are to:

1. Identify patterns, correlations, and any necessary data preprocessing steps, such as handling missing values, outliers, and data normalization.
2. Evaluate which features might be most relevant to predicting client subscription and consider creating new features if applicable.
3. Use a machine learning algorithm of choice to build a model predicting the subscription outcome.
4. Use appropriate metrics such as accuracy, precision, recall, and F1 score to assess model effectiveness. Consider any imbalanced classes and adjust accordingly, possibly using techniques like oversampling, undersampling, or adjusting the class weights
5. Summarize key findings from the EDA and insights from the model, such as which features were most impactful, common characteristics of clients likely to subscribe, and actionable recommendations for the marketing team.

### **Data Preprocessing**

To prepare the data for analysis, the following steps were taken:

1. Checked for duplicates
2. Checked for any rows with missing values
3. Checked the structure of the data set
4. Converted features to a factor
5. Split the data into training(70%) and testing(30%) sets for our data model

### **Descriptive Statistics Exploratory Data Analysis of Age**

<b>Min.</b>	<b>1st Qu.</b>	<b>Median</b>	<b>Mean</b>	<b>3rd Qu.</b>	<b>Max.</b>
<b>18.00</b>	<b>33.00</b>	<b>39.00</b>	<b>40.94</b>	<b>48.00</b>	<b>95.00</b>

#### **Analysis**

The summary of the age variable gives insights to the distribution of ages in the data set. The age ranges from 18 years to 95 years. The average year is 41 years. Half of the clients are younger than 39 years.

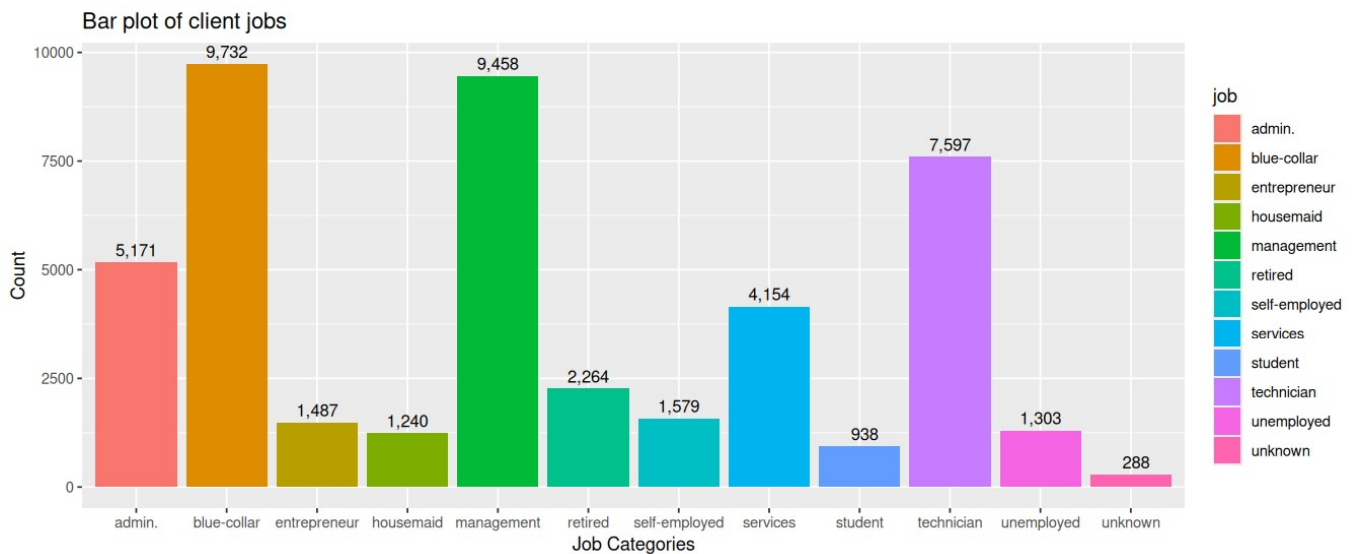
1st quartile: 25% of the clients are below 33 years.

2nd quartile: 75% of the clients are below 48 years.

The distribution of the age data is positively skewed due to the mean greater than the median.

## Exploratory Data Analysis (EDA)

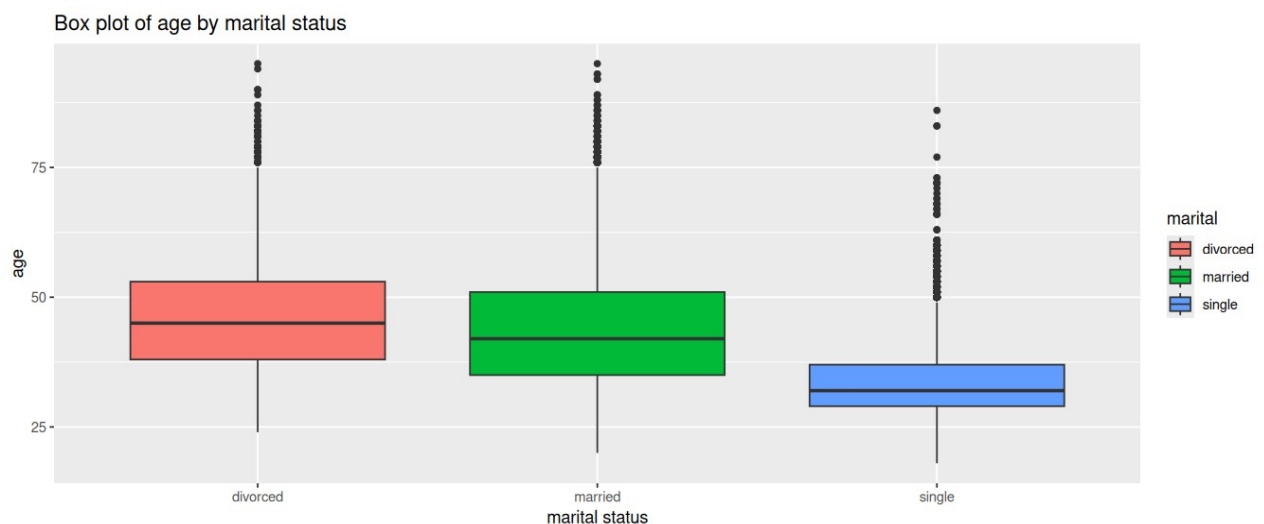
### Data Visualization on client jobs using bar plot



### Analysis

The bar graph shows a bar plot of client jobs. The x axis represent the different job categories, while the y-axis shows the count ranging from 0 to 10,000. The blue collar job category has the highest number of clients recording 9,732, per the evidence from the plot. The management job category has the second highest number of clients recording 9,458. The technician and admin job category has the third and fourth highest number of clients respectively. The lowest number of job category happens to be the unknown with record of 288.

### Data Visualization on age by marital status using box plot

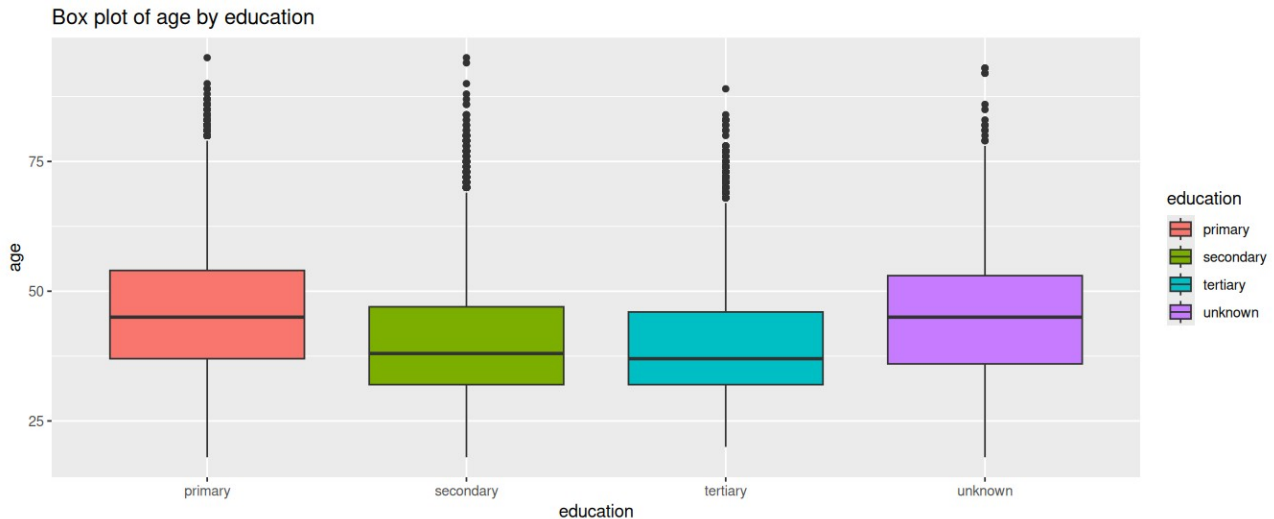


### Analysis

This box plot compares age of clients and marital status. The y-axis represents the age and the x-axis represents the marital status. Each box represents the inter-quartile range for the marital status. From the plot, clients who are divorced has the higher median age followed by clients who are married with a slightly larger box than the divorced, indicating more age variability among

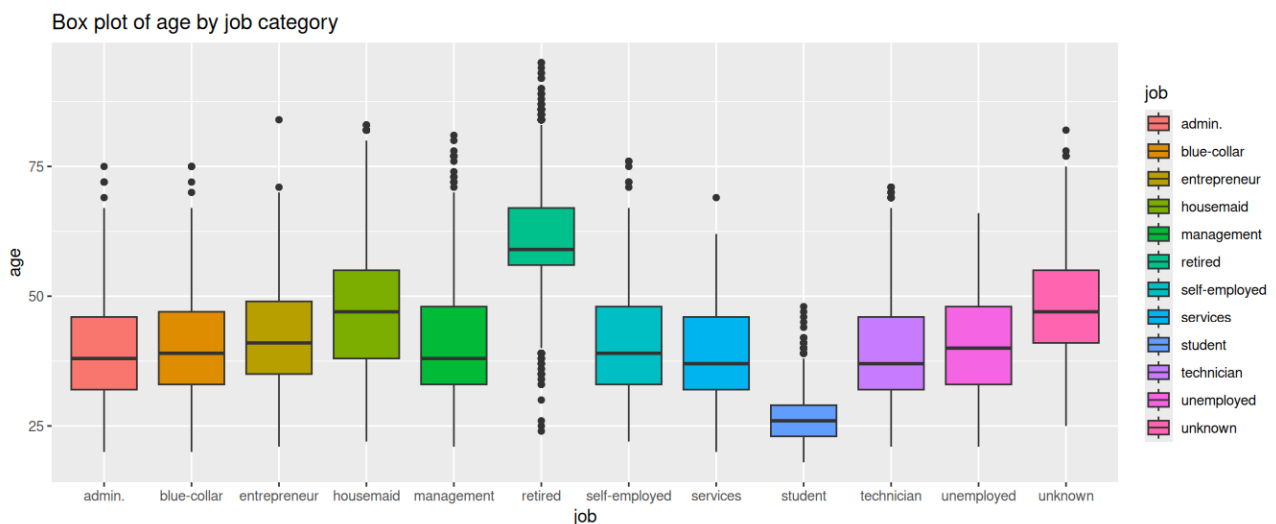
those who are married. Clients who are single has the lowest median with small box indicating less variability in age. All marital status have outliers beyond the whiskers. The single status appears to be negatively skewed whilst the divorced and married appears to be evenly distributed. The presence of outliers suggest age alone may not be the predictor of the marital status.

## Data Visualization on age by education using box plot



## Analysis

This box plot compares age of clients and education. The y-axis represents the age and the x-axis represents the education. Each box represents the inter-quartile range of the education categories. From the plot, clients who belong to the unknown category have the highest median and the largest range indicating more variability in age. The unknown category tends to have a slightly positive skewness in distribution whilst the primary category has an even distribution. The tertiary category has the lowest median and a negative skewness in distribution. All categories show some outliers above their distribution.



## **Analysis**

This box plot compares age of client and their jobs. The y-axis represents the age and the x-axis represents the jobs. All categories show some outliers above their distribution. The student category has the smallest median age with an even distribution.