**Exploratory Data Analysis**

Romita Dangol

Westcliff University

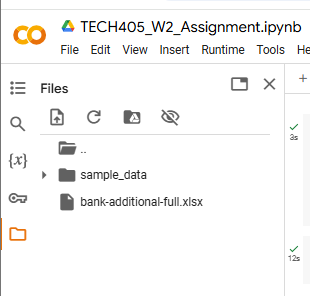
TECH 405: Artificial Neural Network and Deep Learning

Professor Acharya

November 2, 2024

**Exploratory Data Analysis**

For this assignment of report writing doing the exploratory data analysis of a dataset, I have chosen ‘Bank Additional Full’ dataset from Kaggle. In this report, the dataset that I have chosen is a bank client dataset which will be used for future assignments as well in the upcoming weeks that will involve implementing neural networks. Here, a proper EDA process is taken which helps us in understanding the patterns of the data, address any issues like missing values and overall prepare the data for machine learning. As mentioned by IBM (n.d.), EDA analyzes and investigates the data sets and sums up the major attributes by incorporating data visualization methods. Additionally, with visualization, it further helps in interpreting and understanding the data even better in a visualized way. I have done the EDA process in Google Colab using Python language. Below are all the steps for the EDA process.



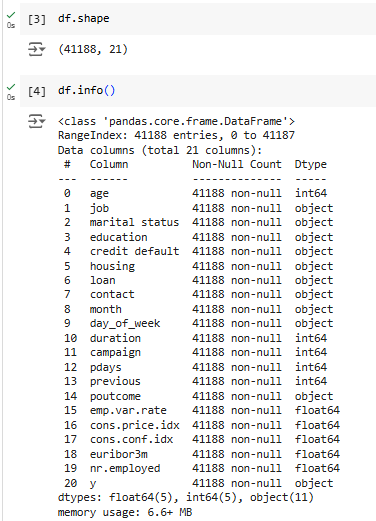
*Figure 1: Uploading the file in Google Colab*

In the files section of Google Colab, firstly, the excel file of the bank dataset was uploaded.



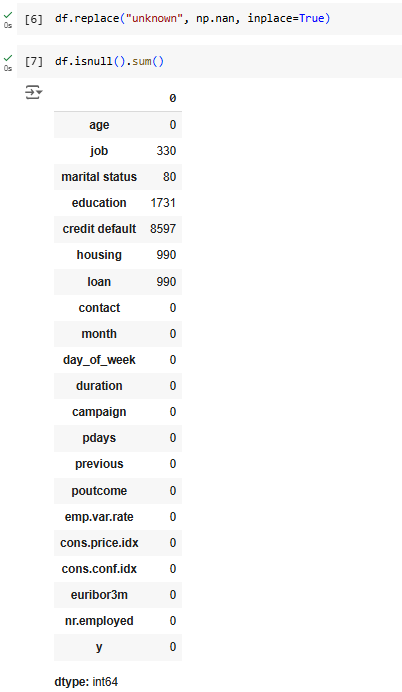
*Figure 2: Importing libraries and loading the dataset*

To execute the further process of EDA, all the necessary libraries like numpy, pandas, seaborn, matplotlib were imported. Then, the dataset was loaded which shows all the data in rows and columns.



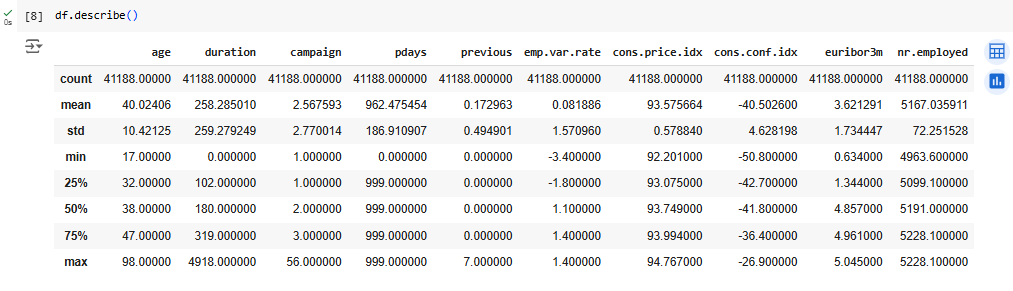
*Figure 3: Shape and information of the dataset*

The above given figure displays the shape of the dataset which is 41188, 22 where 41188 represents the rows and 21 represents the columns in the dataset. The information field shows the information regarding the dataset. It has all the 21 columns displayed with their respective information regarding the number of data they have and the type of data like object which represents categorical data, int64 represents integers and float64 represents decimal values. There are 11 objects in the dataset, 5 floats and 5 integers.



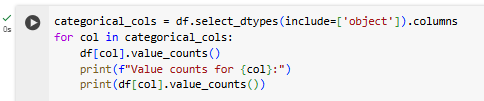
*Figure 4: Replacing and checking missing values*

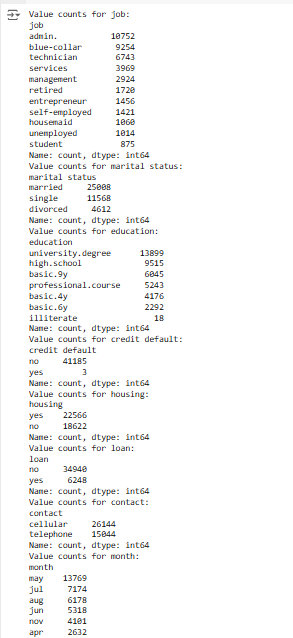
In this dataset, there were no blank spaces but in some columns, there were ‘unknown’ values which represent missing values. For such ‘unknown’ values in the dataset, they were replaced with ‘NaN’ to make handling the missing values easier. This step helps us to handle the missing values in further steps later.



*Figure 5: Statistical Summary of numerical columns*

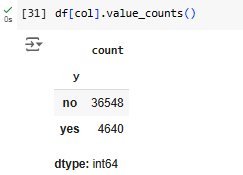
Here, for all the numerical columns in the dataset, a statistical summary is generated which has minimum, maximum values, mean, median, standard deviation, first, second and third quartiles. This step will help us while imputing the missing values with these values.





*Figure 6: Value count of categorical columns*

Here, for all the values that are in the categorical columns, it counts the frequency of each category for each categorical column. Like in the job column, there are 875 students, in the contact column, there are 26144 cellular devices registered and other such more.



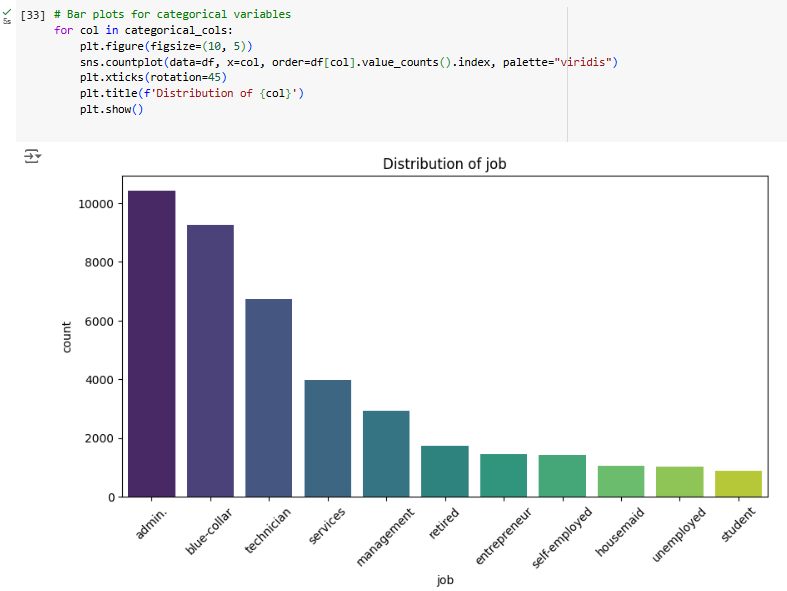
*Figure 7: Count of values in target column ‘y’*

As our target column is ‘y’, here it shows the number of counts of yes and no where there are 36548 no and 4640 yes. This output shows that there is an imbalance in the target variable as there are far more no and yes.



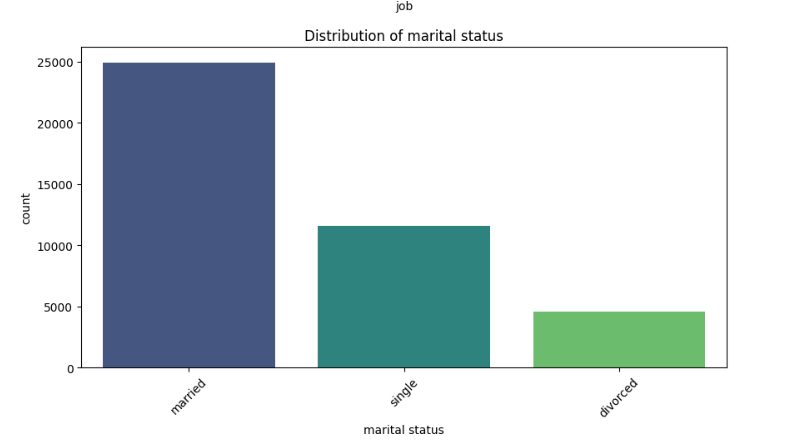
*Figure 8: Distribution of numeric features*

These diagrams show histograms of the distribution of the numeric features of the dataset. They have mixed distributionswith many being right skewed. The column age is right skewed where most of the clients are under the age of 60. Duration is also right skewed with many short call durations. Campaign is also right skewed where few clients have contacted multiple times. The pdays have most values all left skewed which indicates no prior contact. Previous is right skewed with only few repeat contacts and mostly all zero. For emp.var.rate, it indicates stable employment with some fluctuations. For cons.price.index, the peak in the histogram suggests stable periods which are possibly seasonal. For cons.conf.index, it is mostly between -50 and -30 which indicates low consumer confidence. The euribor.3m is left skewed and the peak suggests economic trends or shifts in policy. The nr.employed has distinct peaks which reflects quarterly employment data.



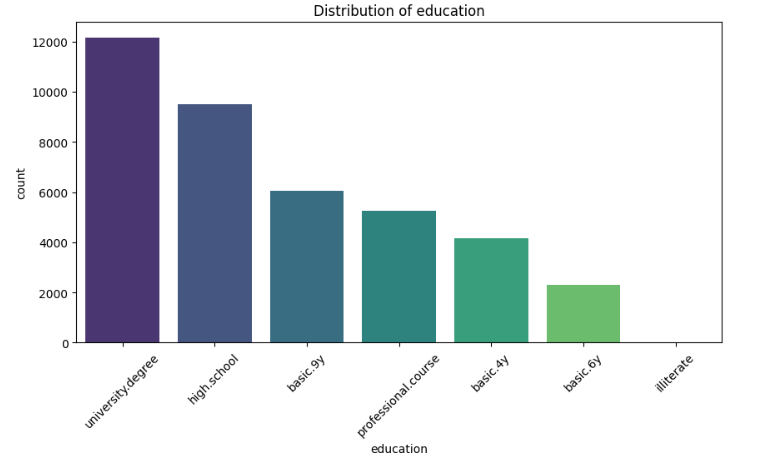
*Figure 9: Bar plots for categorical variables ‘job’*

In this bar plot, we can see that the highest number of job is of admin and the lowest is of students.

**

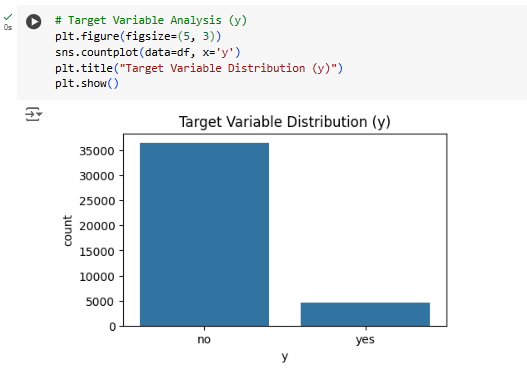
*Figure 10: Bar plot of marital status*

The number of married bank clients are more followed single and divorced which is very less.

**

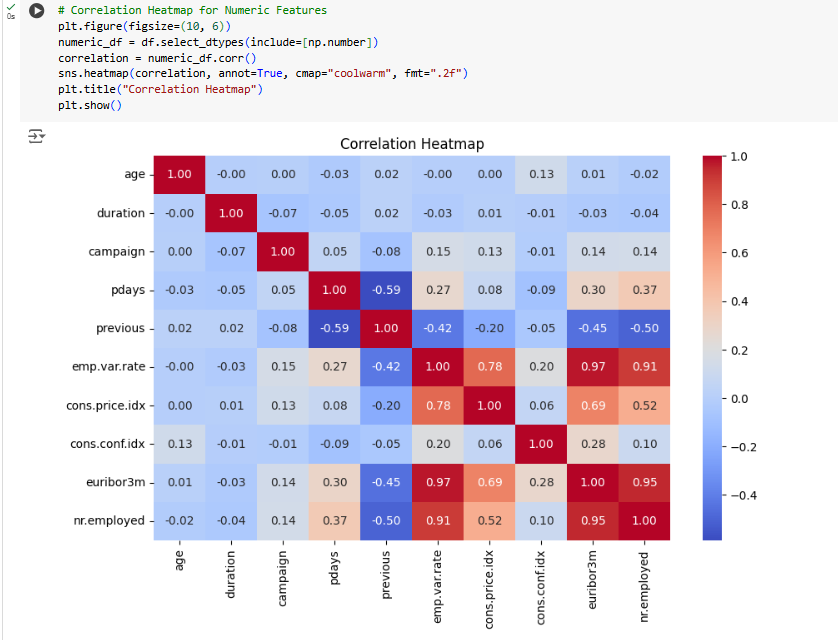
*Figure 11: Bar plot of education*

In terms of education level of the bank clients, most of them have university degree and almost very less are illiterate.



*Figure 12: Bar plot of y*

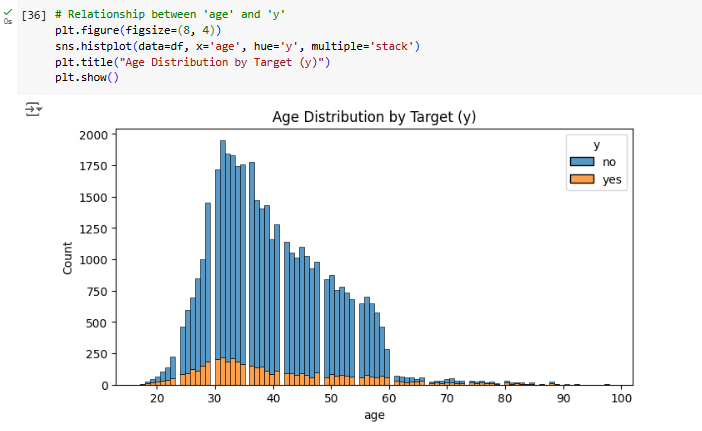
As y is our target variable, the number of customers saying no is more than of yes which shows that there are more customers that will not subscribe the bank term deposit than customers that will.



*Figure 13: Correlation Heatmap*

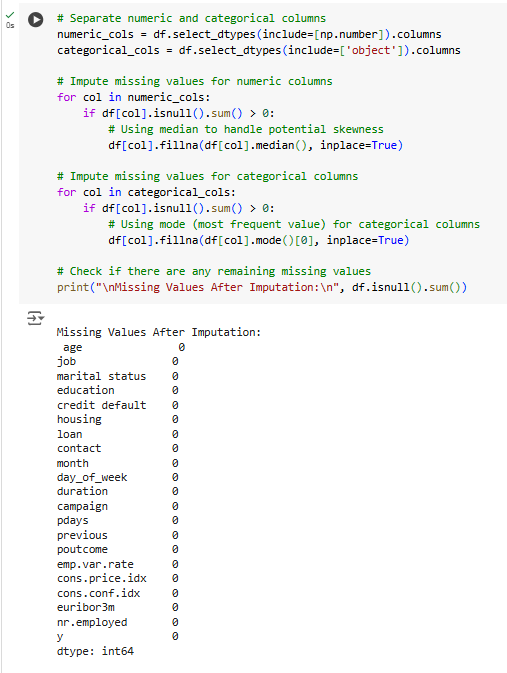
This is a correlation heatmap. A correlation heatmap shows the correlation between multiple variables as a color-coded matrix where each variable is represented by row and column and the cells between them with the value from 1 to -1 indicates how strong or weak the correlation is where 1 is the strongest and -1 being the weakest (Chip, 2022).

Here, the correlation is color coded where the dark red represents high correlation where as dark blue represents weak correlation. For instance, nr.employed and emp.var.rate have high correlation of 0.91 which indicates that as the employment variation increases, the number of employees also tends to increase.



*Figure 14: Age distribution by target ‘y’*

With the above histogram, the blue bars represent no and orange represents yes. We can see that the people with age below 60 have more data and will not subscribe to the bank term deposit compared to people who will.



*Figure 15: Imputing the missing values*

Earlier when we checked for the missing values, there were some missing values in few columns. To impute those values, first, the numerical and categorical columns were separated. For numerical columns, median was used to impute the missing values to handle potential skewness. As for categorical values, mode is used where the most frequent values in the column were placed in the places that had missing values. Lastly, once again, we checked in more any missing values and there were none.

The EDA process and steps that were done prepared our dataset for further processing. By handling the missing values, imputing, the visualizations we did made the understanding of the dataset even better for modeling and eliminating any potential issues of our data structure. The missing values were imputed that preserved the data integrity without reducing the dataset size and visualizations provided insights into the distribution of the data, relationships between variables and correlation. This will now be useful and ready for our next assignment while implementing neural networks on this preprocessed data.

**References**

Chip. (2022). How to read a correlation heatmap? *Data Fluency.* <https://www.quanthub.com/how-to-read-a-correlation-heatmap/#:~:text=What%20Is%20a%20Correlation%20Heatmap,closely%20related%20different%20variables%20are>.

IBM. (n.d.) What is exploratory data analysis (EDA)? *Topics*. <https://www.ibm.com/topics/exploratory-data-analysis>