

# Data Analysis on Bank Marketing Data

**Project Analysis** - Praveen Kumar Peddabudi

**Qualification** - B. Tech in Computer Science

**Skills** - Python, EDA, SQL, Power BI

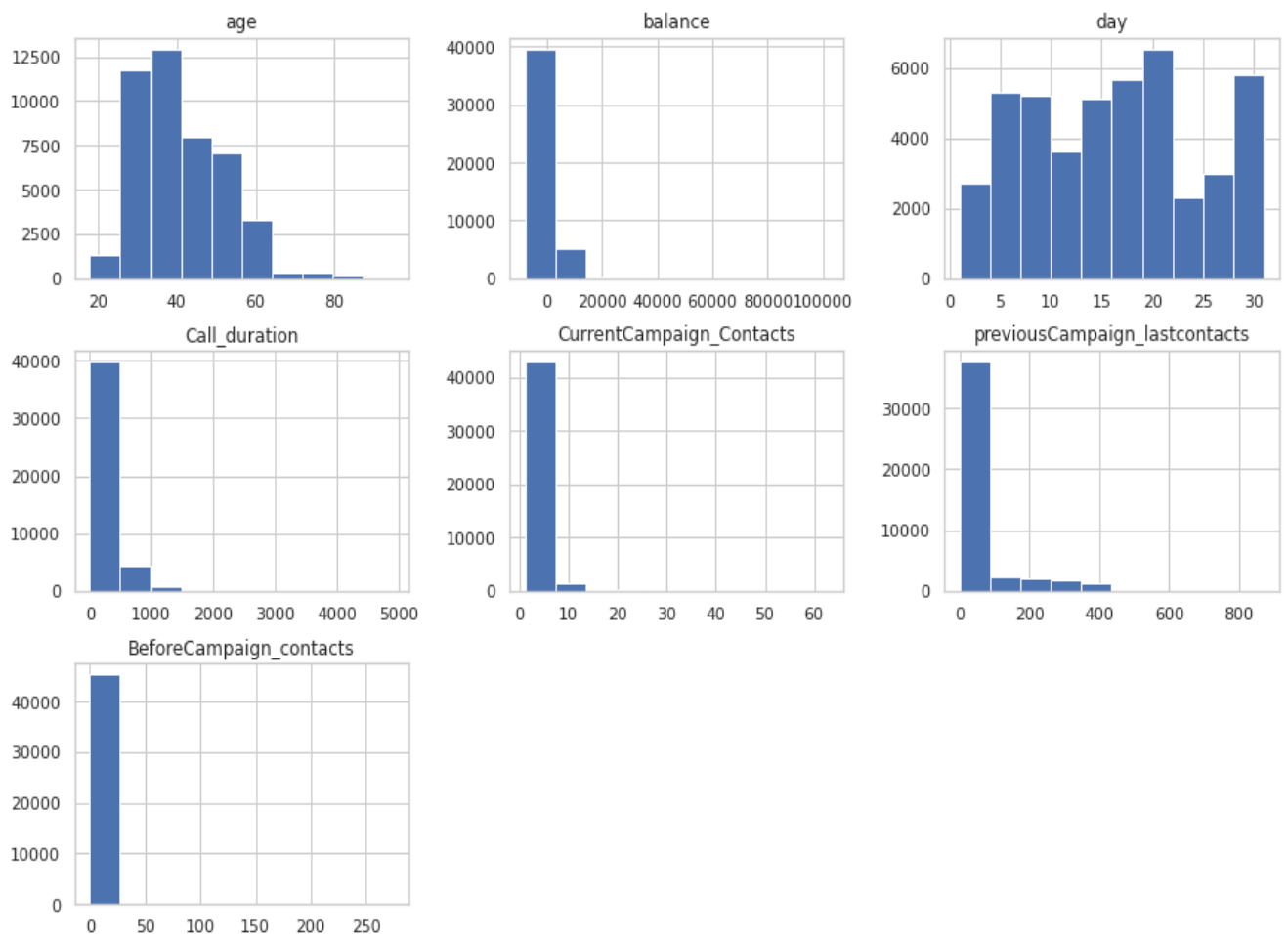
## PROJECT PROBLEM STATEMENT

We need to build a profile of customers who are more likely to get a term deposit from the bank by identifying factors which will make bank to develop more targeted marketing campaigns

## PROJECT PAIN POINTS

Treating Skewness because we have 5 numerical variables highly skewed which will introduce bias and plotted using histogram as shown below

Histogram plot for Numerical Variables



**Explanation:** By looking at the histogram subplots we can find Balance, Call duration, Current Campaign Contacts, Previous Campaign last contact, Before Campaign contact independent variables are highly skewed towards right.

## Data Exploration

head () and tail () - View top or bottom rows of panda's data frame

shape - View dimensionality of the dataset

describe () - Provides descriptive information about the data.

info () - Shorter version of dataset summary.

dtypes() - Give data types of each feature.

```
bm.describe(include='category') # shows the categorical columns with statistical data
```

	job	marital	education	Credit_Status	Housing_loan	Personal_loan	Communication_type	Contact_month	PreviousOutcome	Term_deposit
count	45211	45211	45211	45211	45211	45211	45211	45211	45211	45211
unique	12	3	4	2	2	2	3	12	4	2
top	blue-collar	married	secondary	no	yes	no	cellular	may	unknown	no
freq	9732	27214	23202	44396	25130	37967	29285	13766	36959	39922

If we see top frequency is high in factors like Credit\_Status, PreviousOutcome and Term\_deposit which shows imbalance in the data

```
bank_marketing.describe() # shows the numerical columns with statistical data
```

	age	balance	day	duration	campaign	pdays	previous
count	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000
mean	40.936210	1362.272058	15.806419	258.163080	2.763841	40.197828	0.580323
std	10.618762	3044.765829	8.322476	257.527812	3.098021	100.128746	2.303441
min	18.000000	-8019.000000	1.000000	0.000000	1.000000	-1.000000	0.000000
25%	33.000000	72.000000	8.000000	103.000000	1.000000	-1.000000	0.000000
50%	39.000000	448.000000	16.000000	180.000000	2.000000	-1.000000	0.000000
75%	48.000000	1428.000000	21.000000	319.000000	3.000000	-1.000000	0.000000
max	95.000000	102127.000000	31.000000	4918.000000	63.000000	871.000000	275.000000

If we observe the difference between mean and standard deviation is very high for balance, pdays columns and also between 75th percentile and max value is very high which indicates skewness for duration, campaign, pdays and previous columns

## Preparing data

After getting familiar with the data, we can prepare our data. We can ask the questions like

- Are there any unnecessary columns?
- Are there any duplicate columns?
- Are the columns naming, correct?

By finding answers of the above questions, we can clean our dataset and keep only those variables which can be used later.

**df.duplicated.sum ()** Returns number of duplicate rows in the dataset. If the number is higher, we can assume some data collection error and remove the duplicates.

**df.rename ()** If the naming does not make sense for any variables, we can rename it to develop more understanding

**df.drop()** Removes columns from the dataset.

## Variable Understanding

In this step we focus more on individual variables. This is also known as univariate analysis

This analysis can be done with different ways for

Categorical variables - Value of the variable is from one of the predefined categories

- (e.g., Occupation - Doctor, Engineer, Scientist)

Numerical variables - Value of the variable is come from a numerical range

- (e.g., Age of a person, Height in cm)

Categorical variables are further divided into nominal and ordinal.

- Nominal - Country, Gender, Race
- Ordinal - Salary bucket, Rank in the class

Independent Variables	Dependent Variables
Age	Term deposit
job	Credit Status
marital	Housing Loan
education	Personal Loan
Balance	Previous Outcome
Communication Type	
Contact Month	
Call_duration	
Day of Contact	
PresentCampaign_contacts_made	
Previous Campaign last contacted	
Before Campaign Contact	

## Variable Understanding - useful functions

### Categorical Variables

**value counts ()** - Show the values and the number of occurrences in descending order.

**unique ()** - Count number of unique values

**plot ()** - We can plot the value counts to view the result in graphical format.

### Numerical Variables

**hist ()** - to plot histogram

**skew ()** - skewness of the variable

**Kurt ()** - Kurtosis of the variable

**df.isnull(). sum ()** - shows the total number of null values in each feature. Applicable to both categorical and numerical features.

### Variable Relationship

Relation of variables and influence on target variable.

**Correlation** - Relationship between two variables

**Positive correlation** - If value of one variable increases, value of second variable also increases.

**Negative correlation** - If value of one variable increases, value of second variable decreases.

**Zero or near zero correlation** - Variables are independent of each other.

**Scatterplot** - Understanding relationship of two numerical variables.

**Boxplot** - Understand relationship of numerical and categorical variables.

**Bar chart** - Understand relationship of two categorical variables.

### Variable Relationship - Useful functions

**seaborn. pair plot ()** - To view all variables relationship in one chart. Not useful when the number of features is high.

**seaborn. cat plot ()** - To generate boxplots.

**seaborn. Scatterplot ()** - To generate scatter plots between the data of two variables.

**df. corr ()** - To get correlation matrix of variables.

**seaborn. Heatmap ()** - To visualize values in a colored matrix.

## Univariate Analysis

To plot all categorical features in a single plot



## **Key Observations from Univariate Analysis on Categorical Columns**

- We can identify Client who are married are contacted more by the bank and divorced Client have been contacted least followed by single category.
- In Job Factor, we can observe that Client with blue-collar, management and technician jobs have been contacted more by the bank which are more than 6k
- If we observe the plot, secondary education category is contacted most.
- The plot shows the client with default credit value status as 'no' are the most who have been contacted by the bank for the deposits. Client with default status 'yes' have not been contacted by the bank at all.
- Housing loan category clients are more in number which may help in finding relationships with target variable.
- Personal loan category shows nearly 7K customers are contacted.
- The Clients who are last contacted are majorly in the month of May followed by July and august. Bank has least contacted the clients in the december followed by march, september and october.
- The Plot on contact communication type explains bank has contacted clients mostly through cellular communication and least contacted through telephone communication. Also, there are nearly 14k clients where bank is unable to reach customers.
- The Plot on outcome of the previous marketing campaign contains 37K unknown values which is more imbalanced.

## **APPROACH ADOPTED**

### **Skewness Treatment or Outlier Analysis**

#### **Skewness**

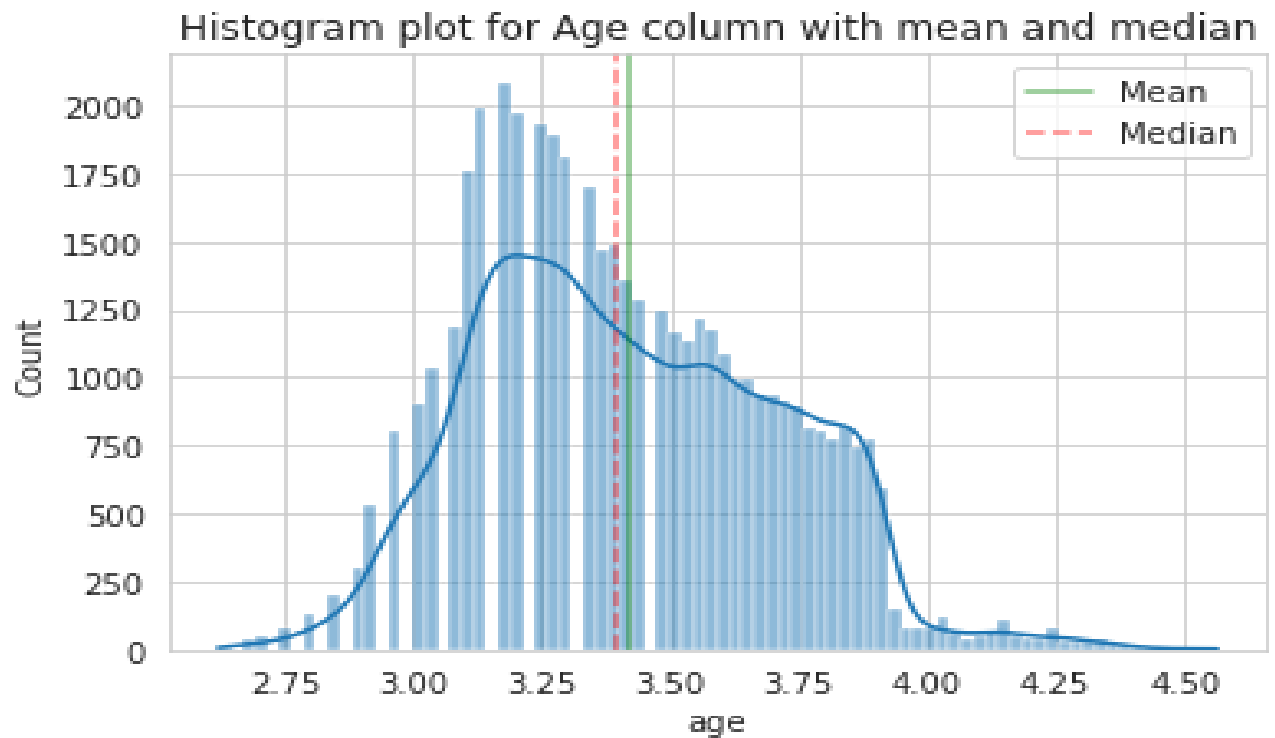
- If skewness is between -0.5 and 0.5, the distribution is approximately symmetric
- If skewness is between -1 and -0.5 or between 0.5 and 1, the distribution is moderately skewed
- If skewness is less than -1 or greater than 1, the distribution is highly skewed

#### **Kurtosis**

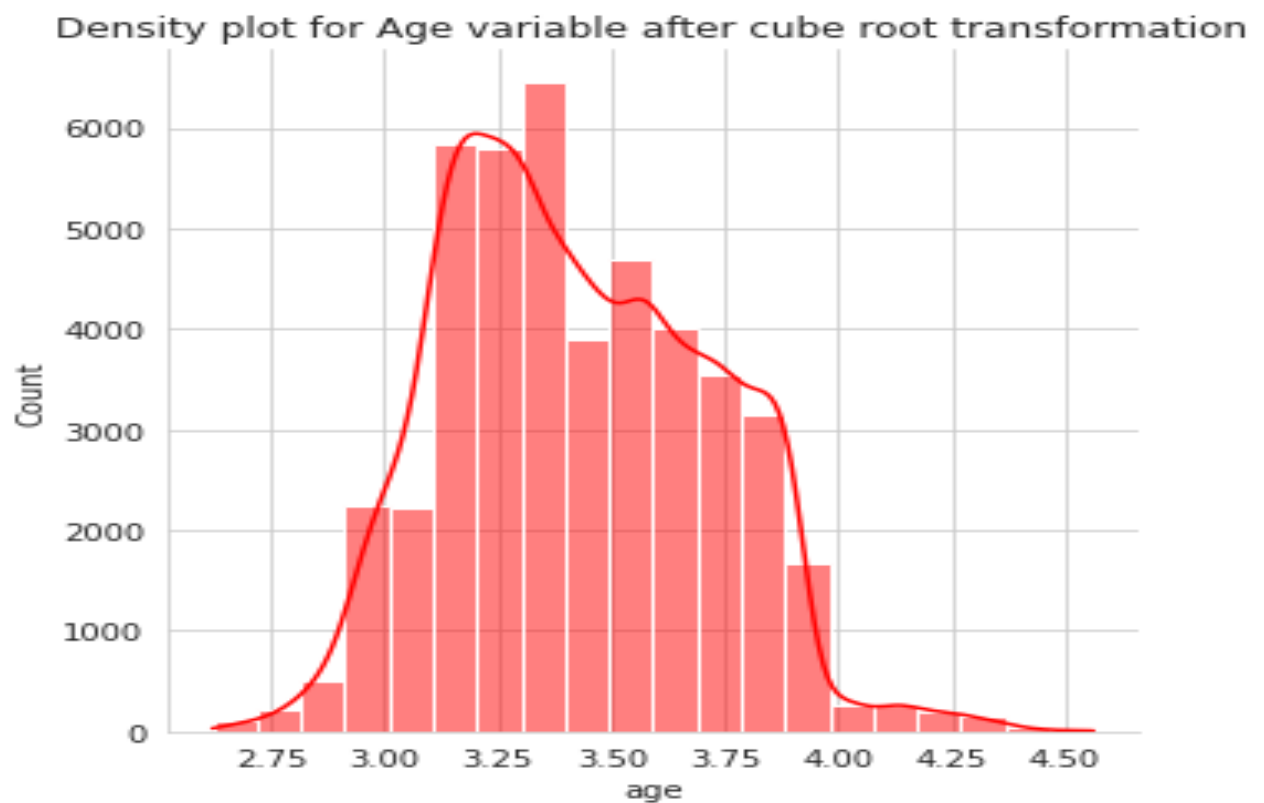
relates to the degree of presence of outliers in the distribution

In finance, kurtosis is used as a measure of financial risk. A large kurtosis indicates high level of risk for an investment because it indicates that there are high probabilities of extremely large and extremely small returns. Also, a small kurtosis signals a moderate level of risk because the chances of extreme returns are relatively low.

As there is skewness observed in the plot applying transformations to make the variables change to symmetrical distribution

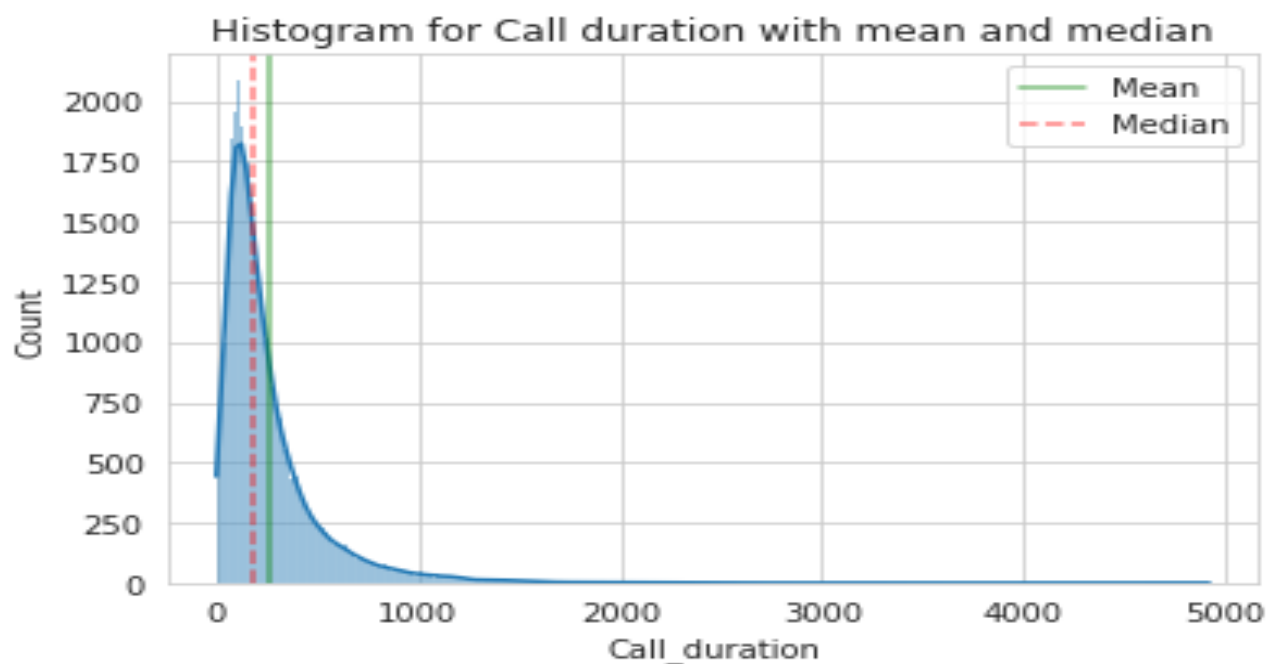


**Explanation:** If we observe the plot, first of all mean > median so the curve is right skewed and for that applying cube root transformations as log and square root is not suitable

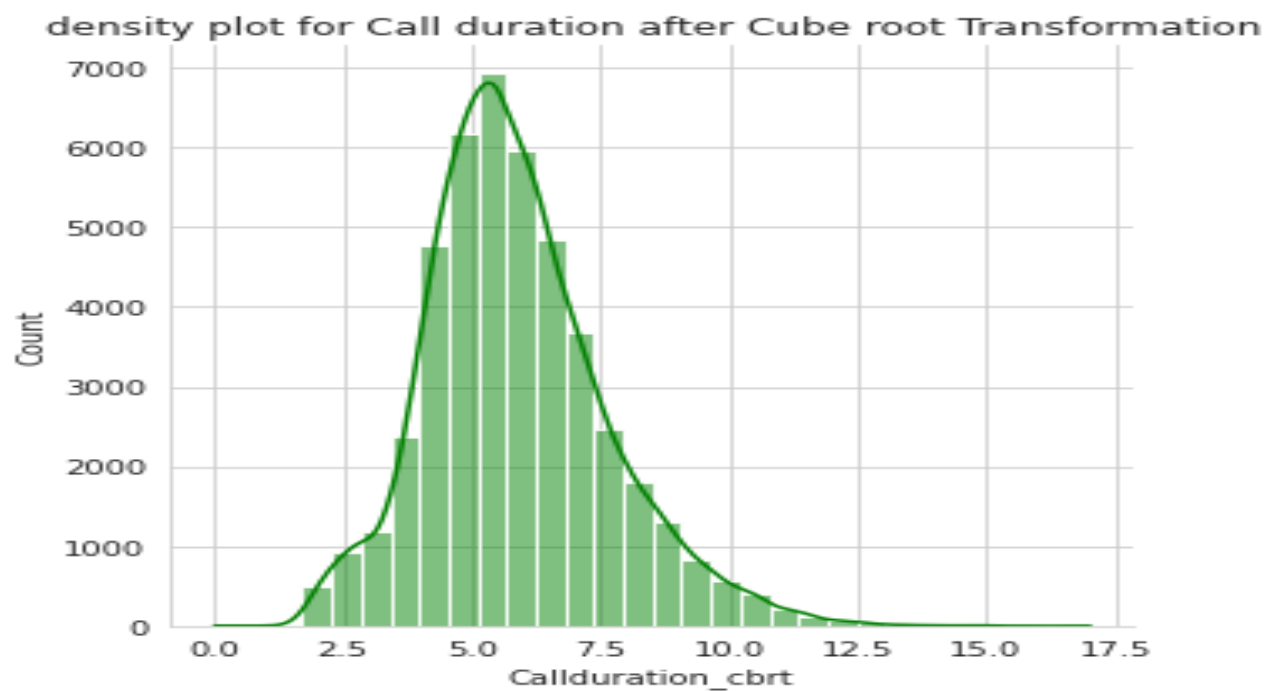


**Explanation:** Here, if we observe the plot is slight normally distributed but not very accurately

In the same way, applied histogram plot for call duration as figure below which is right skewed and mean > median



**Explanation:** As the curve is skewed so performed cube root transformation to make distribution normal as below where plotted a density for Call duration



**Explanation:** If we see the difference of the skewness is changed when compared before and after cube root transformation



## TECHNIQUES AND TOOLS

As project requires Used Python and Google Colab notebook

Libraries imported to perform Exploratory Data Analysis

- *Pandas* - For data loading and manipulation
- *Numpy* - Mathematical calculations, extra faster functions
- *Matplotlib* - Visualization tool (Can create complex custom charts)
- *Seaborn* - Visualization tool (in-build high quality charts)

### Functions to read data

**read\_csv()** - Read comma delimited file into pandas dataframe.

**read\_excel()** - Read excel files into pandas dataframe.

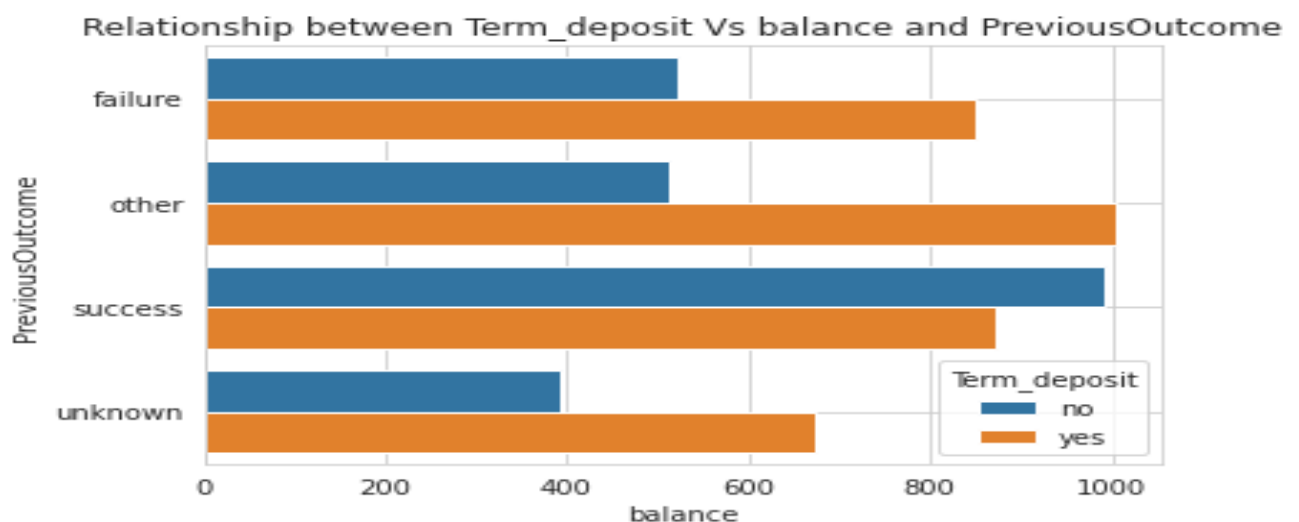
Example: `pandas.read_csv("./data/myfile.csv")`

### Additional parameters

We can also give parameters like column separator, read headers, column names, date converters etc

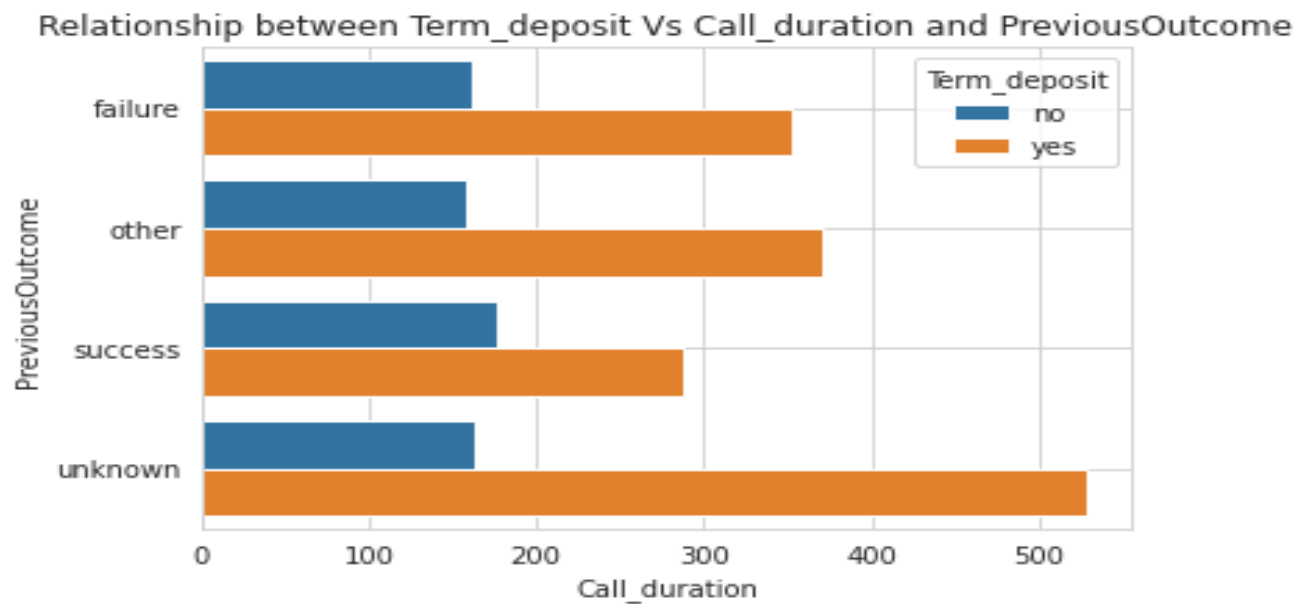
## PROJECT ANALYSIS

Multivariate analysis:



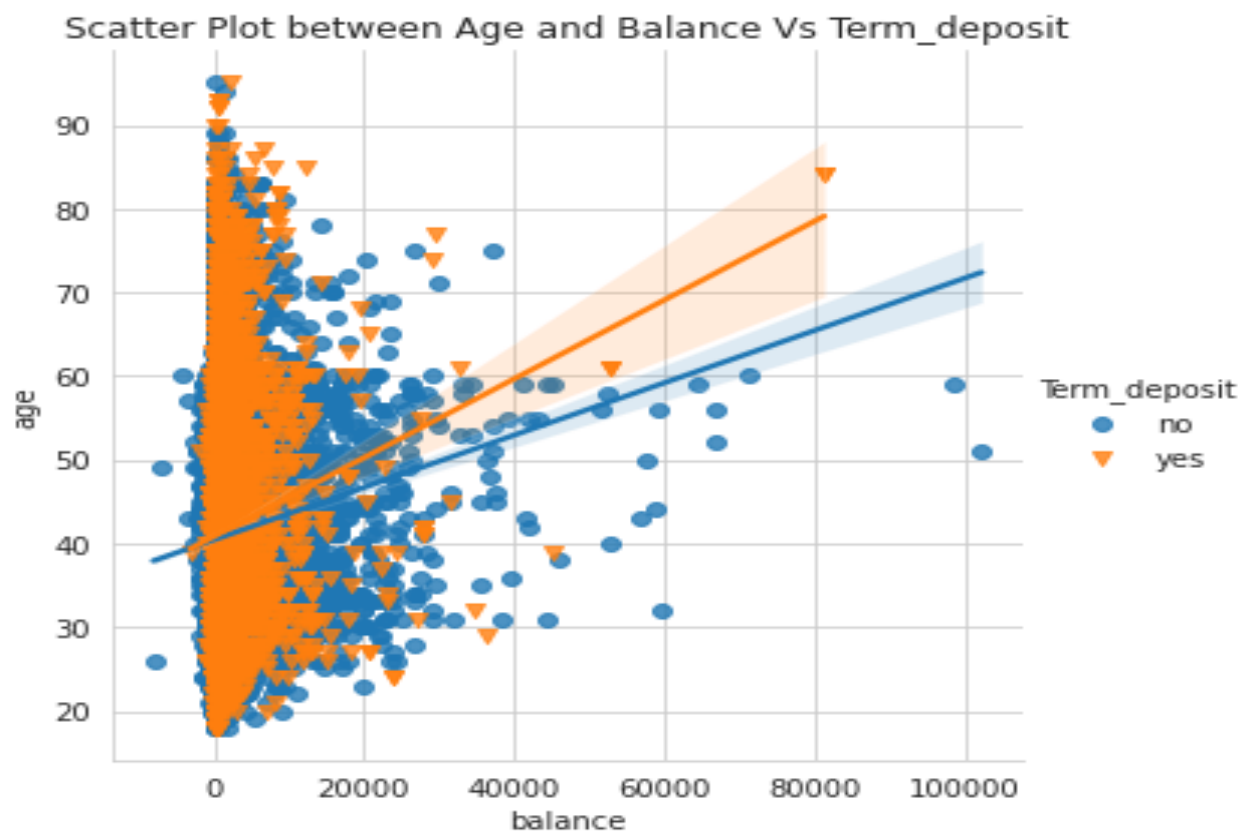
**Explanation:**

The Success Outcome possibility is directly proportional to the balance client is managing and at the same time failure rate is equal to the half of the other samples taken



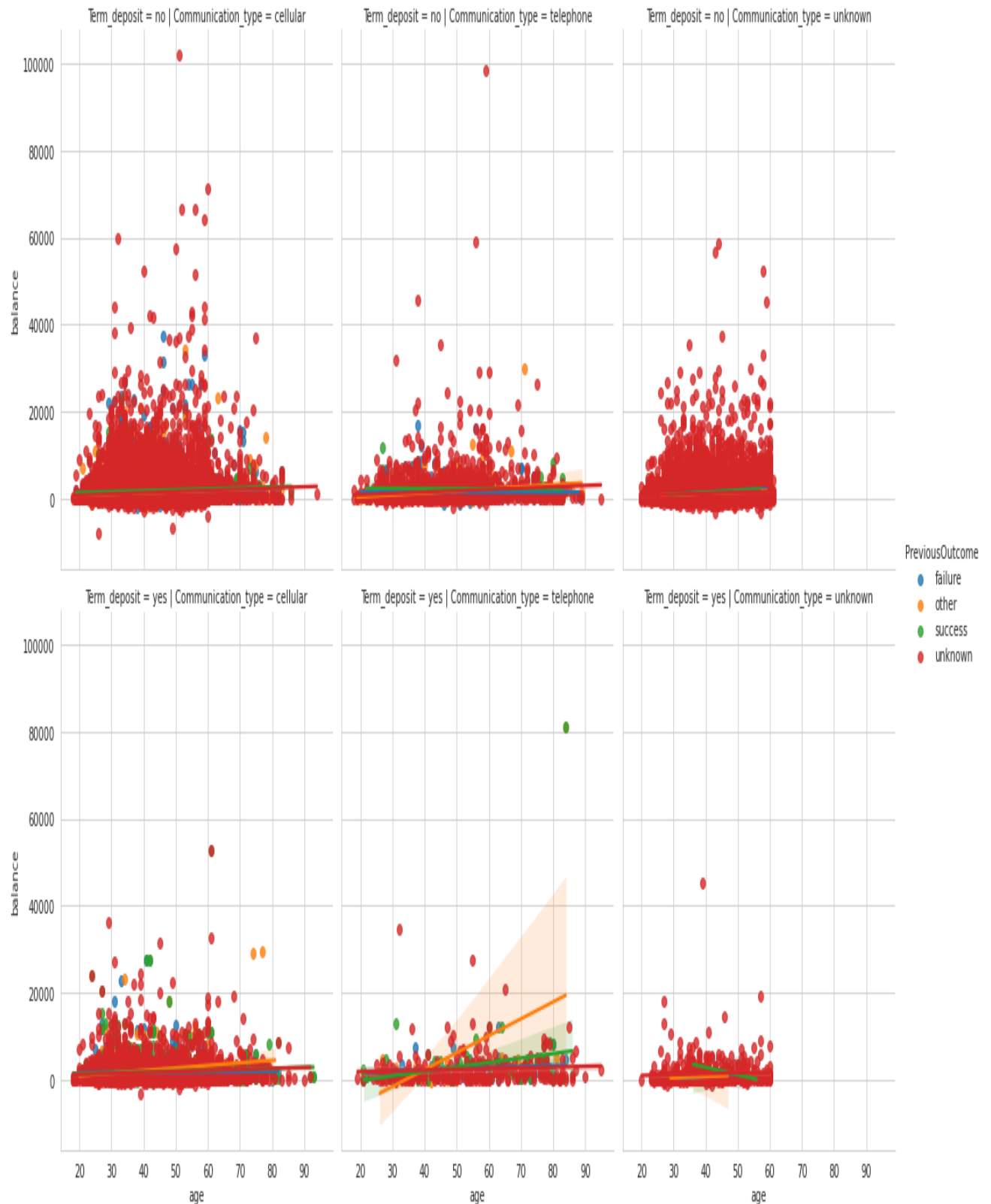
#### Explanation:

Bar plot depicts ratio of success possibility of previous Outcome campaign somewhere near to 4.5 minutes on an average



#### Explanation:

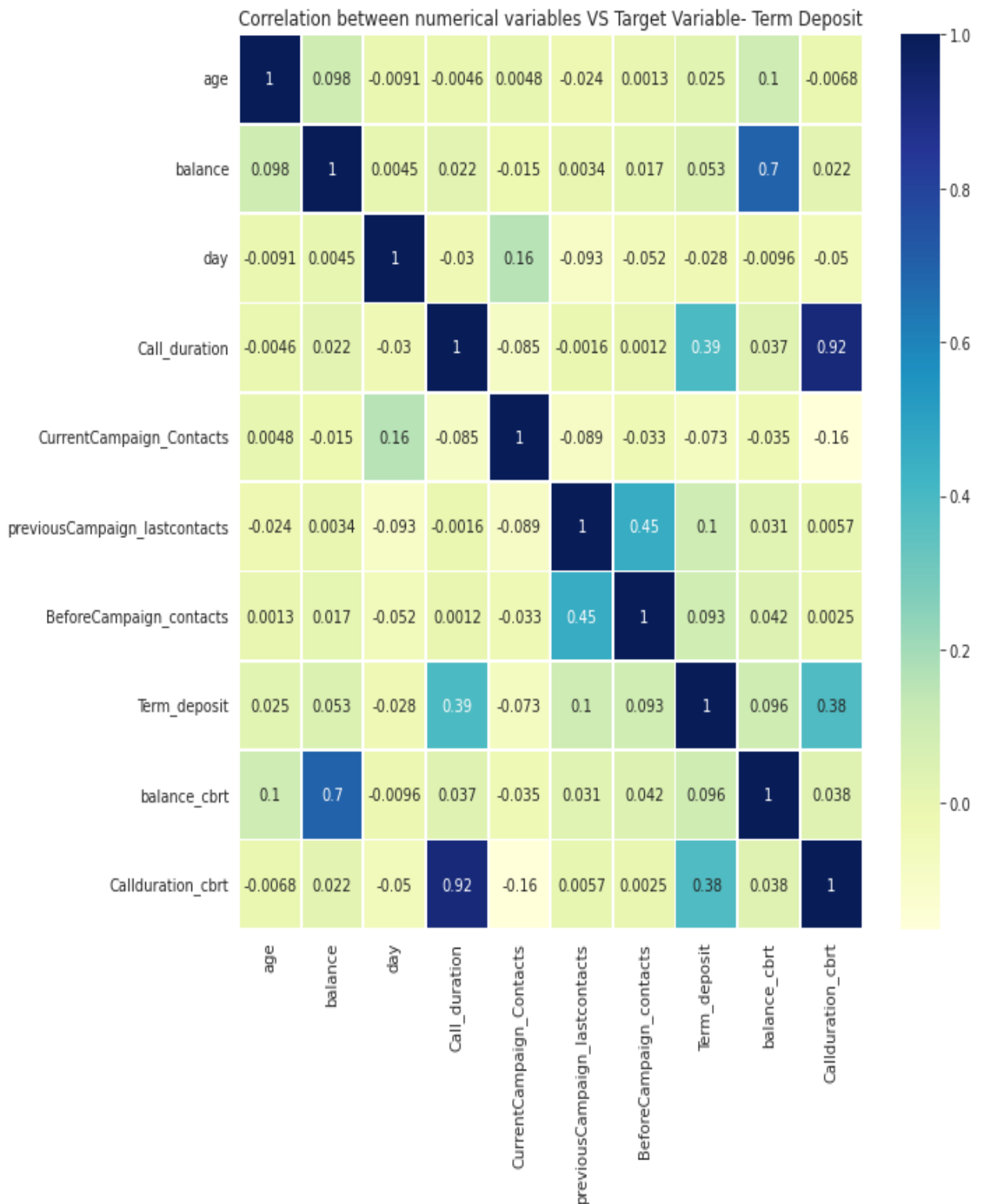
If we see the plot age has positive linear correlation with balance in term deposit activation and are important factors in analysis going further.



### Explanation:

The Plot shows the term deposit in communication telephone category plot with age and balance factors are positively linear correlation and success in the outcome of previous campaign

### Heatmap Correlation:



### Explanation:

From the Correlation among the numerical columns we can observe that call duration is correlated with term deposit whereas balance, previous Campaign last contacted, Before Campaign contact are also nearly correlated. But after performing analysis it is considerable to drop columns to make a better model.

## **Multivariate Analysis key Observations**

- Entrepreneurs who are contacted in March and April are mostly liked to get term-deposits which is important observation
- Cellular and Telephonic both categories client includes married, divorced are showing interested towards to get a term deposits
- Clients with job categories including entrepreneur, services, retired and housemaid in tertiary educated are more interested towards taking term deposit
- Admin, blue-collar, retired clients in secondary education are more interested towards taking term deposit











## **Key Insights drawn from the Analysis**

- Clients with job categories including entrepreneur, services, retired and housemaid in tertiary educated are more interested towards taking term deposit
- Professions of Admin, blue-collar, retired clients in secondary education are more interested towards taking term deposit
- Cellular and Telephonic both categories client includes married, divorced are showing interested towards to get a term deposits
- If we can see that callduration is symmetrical where balance and age is right skewed
- For balance column data is skewed also no correlation with target so this not an important factor anymore.
- Contact\_month explains clients who are contacted are not interested in any months except in the month of May(who are interested to buy term deposits) but more capacity is towards "no" so this states Contact\_month is not an important factor
- Clients who are holding personal or house loans are interested again to subscribe term deposit also bank should approach people who are not having active loans in the bank

This concludes data which is collected from previous campaign is more useful, housing and personal loan factors are very important

## CONCLUSION

Below are the major factors and client's profile having to be taken under below categories which will helpful for the bank to build and design effective campaigns to get more clients who will take term deposits.

-  Age
-  Communication type
-  Credit Status
-  Education
-  Housing loan
-  Job
-  Marital
-  Personal loan
-  Call duration
-  Previous Outcome