**Bank Marketing Analytics Project Proposal**

**Introduction:**

The bank marketing data has information about the direct marketing campaigns used by a Portuguese banking institution. The marketing campaigns were based on phone calls. And to gauge whether the client would make a termed deposit, they often had to contact the customer multiple times. The data consists of numerical and categorical variables. The dataset has 20 predictor variables (features) and around 41K rows.

**Questions we are trying to answer:**

1. *The business questions you are trying to address and why this question is important.*

One of the most important questions we are trying to answer is whether the client will make a termed deposit(‘yes’) or not (‘no’).

We want to see whether the previous marketing campaigns(phone calls) has worked for the Banking Institution or not. To answer whether the customer will make a term deposit or not, we can look at the insights derived from the following questions:

1. Are there certain job types making more termed deposits? Were they targeted more?
2. Did the clients have to be contacted multiple times before they made the deposit?
3. How many contacts were made in the past for this client and what was the outcome of the previous campaigns? This would help answer how should those customers be contacted in the future
4. Is there a particular pattern observed for variables such as marital status and education? Does having a particular marital status and education affect how term deposits are made?

This question is important, because if we can correctly predict whether the customer will make a term deposit or not, then the Banking Institution can direct their marketing efforts and spends in a way to target those potential customers. Answering this question can help influence the company’s business and marketing decisions.

1. *Basic information about the data, including the source and the DGP of these data. Some basic summary statistics and some plots.*

This dataset is from: [http://archive.ics.uci.edu/ml/datasets/Bank+Marketing#](http://archive.ics.uci.edu/ml/datasets/Bank+Marketing)

**Basic information about the data and trying to understand the DGP as follows:**

***Numerical variables included in the data:***

Age: age of contacted person

Duration: last contact duration, in seconds

Campaign: number of contacts performed during this campaign and for this client

Pdays: number of days that passes by after the client was last contacted from a previous campaign

Previous: number of contacts performed before this campaign and for this client

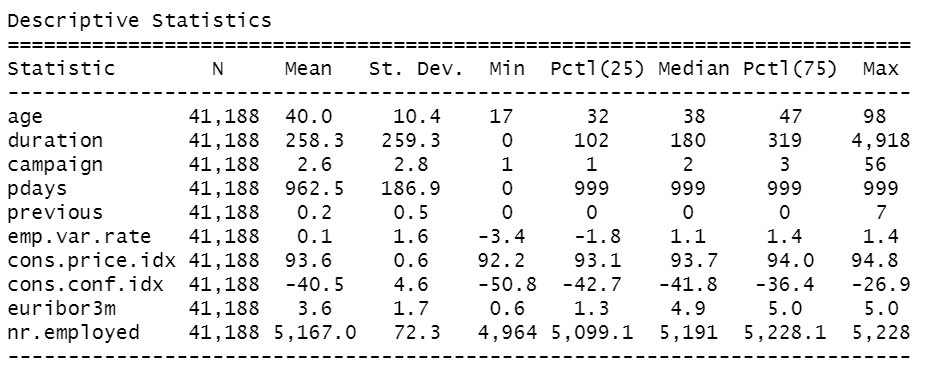
Emp.var.rate: employment variation rate (quarterly indicator)

Cons.price.idx: consumer price index (monthly indicator)

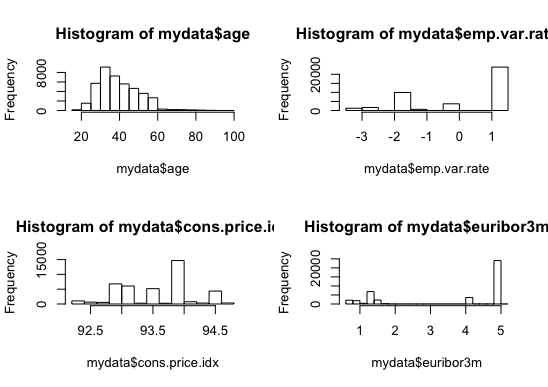
Con.conf.idx: consumer confidence index (monthly indicator)

Euribor3m: euribor 3 month rate (daily indicator)

Nr.employed: number of employees (quarterly indicator)

*A short summary for the continuous / numerical data can be seen as follows:*

We also try to understand the data, by plotting some of the variables. For eg, looking at the first graph of age, we see that the graph is mostly skewed towards the younger population, and the mean and median age is around 40 and 38 yrs old. Hence we can conclude that more of the younger aged people are contacted as compared to the older aged.



***Categorical variables included in the data:***

Job: type of job

Marital: marital status

Education: education level

Default: has credit in default?

Housing: has housing loan?

Loan: has personal loan?

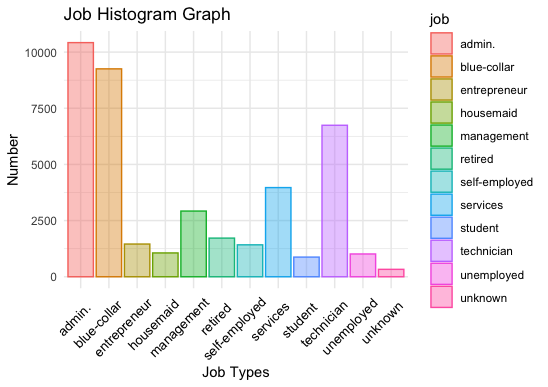
Contact: contact communication type

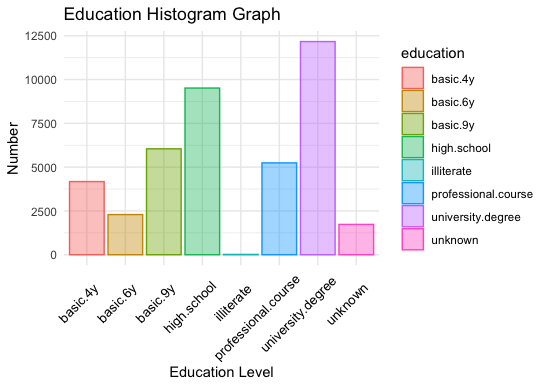
Month: last contact month of year

Day\_of\_week: last contact day of the week

Poutcome: outcome of the previous marketing campaign

By plotting the below two graphs, we observe that, majority of the people contacted have obtained a degree from some university. And also, the highest number of people contacted have admin and blue collared jobs.





***Binary Dependent Variable:***

y: has the client subscribed a term deposit? (will be converted to 0 and 1)

1. *Why these data are suitable to solve the business question?*

These data are suitable to help us solve the business question because the dataset contains numerous data points on customer demographics, such as job and marital status, which can help us answer which type of customers to target. In addition, the data contains variables on the marketing campaign itself, such as pdays, campaign, and more. We can see how each of the variables are correlated from the table below, and also understand how these variables and data can help us answer which marketing strategies work best for banking deposits. Therefore, this dataset allows us to approach the business question from both sides: the customer and the marketing strategy.

1. *Which tool is planned to solve the problem in data analysis*

We would be using the tool R to solve our business problem. Since the fundamental question that we are trying to answer is whether or not the customer would make a term deposit and hence it becomes our ‘Y’ variable. Our ‘Y’ variable takes only two values ‘yes’ and ‘no’ making it a categorical variable. When this is the case, we use Logistic Regression. And hence we will use the technique Logistic Regression using R to answer our business question. 