DATA CLEANING AND SUMMARISING

By Hari Hara Priya Kannan

# TABLE OF CONTENTS

1. **Data Preparation** 
   1. Load Data
   2. Format data
   3. Correct typos
   4. Remove whitespaces
   5. Cast text
   6. Sanity Checks
   7. Deal with missing Values
2. **Data Exploration** 
   1. Visualizations
   2. Hypothesis
   3. Scatter Matrix

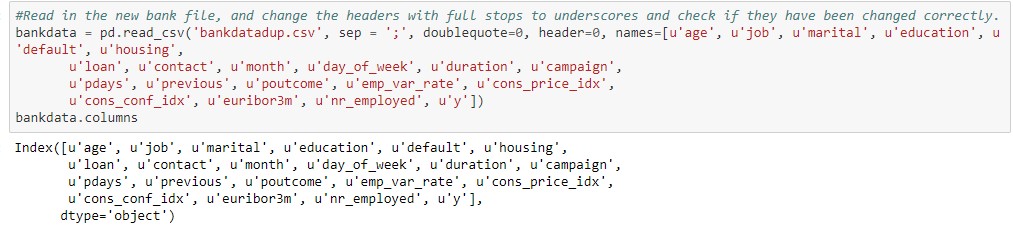
# DATA PREPARATION

## 1.1 Load Data

To load the data, initially the following modules were imported.

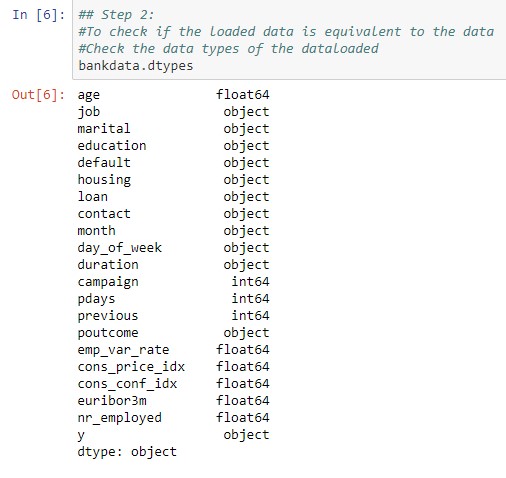
* csv
* numpy
* pandas

At first a duplicate copy of the dataset ‘bankdatadup’ was created. Later using the read\_csv function, the data was loaded into a data frame called ‘bankdata.csv’.

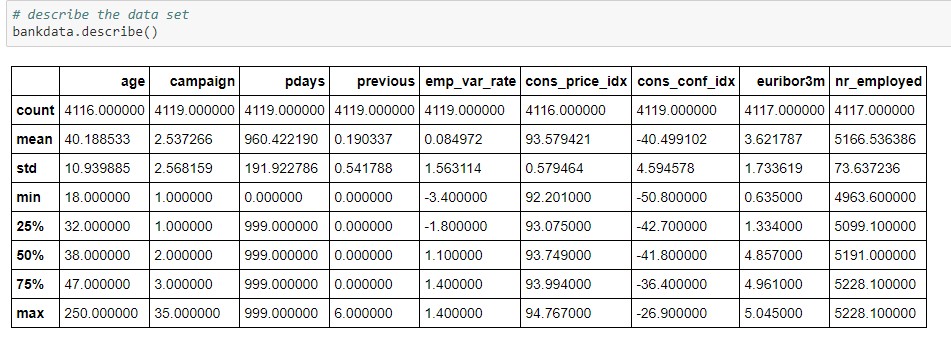


## 1.2 Format Data

After loading the dataset, the data frame was checked for the format of the data types and the summary was obtained.



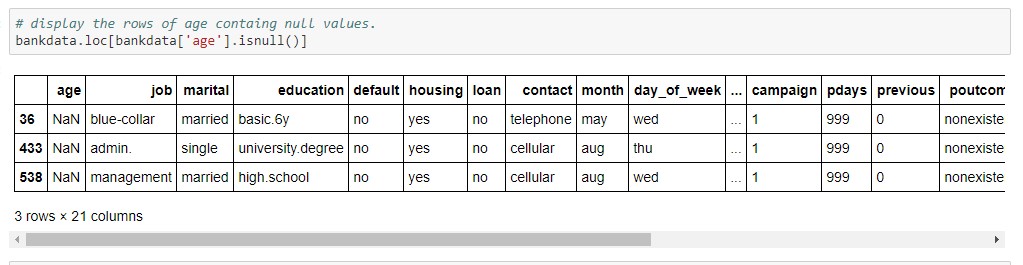
Then, the summary of the numerical values were obtained.



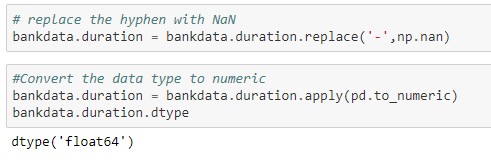
From the above table, we can observe the following:

1. *The count values of are not uniform.*
2. *The maximum value of age is observed to be 250 which is not possible.*
3. *Most of the values in pdays are 999*
4. *Most of the values in previous are 0*

The **age** column is first explored and we can find three values that are null and also the maximum value to be 250. The null values are the filled with the mean value of the age column. An age mask is created to correct any erroneous values more than 120.



The type of **duration** is found to be object and there is a hyphen in the dataset. The hyphen is changed to null and the column type is changed to numeric.



## 1.3 Correct Typos

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Before** | **After** |
| **Job**- ‘admin.’  Was changed to admin |  |  |
| **Marital**- ‘divorcede d’ was changed to  ‘divorced’ |  |  |
| **Education**-‘basic.6yes ’ was changed into ‘basic.6y’ |  |  |
| **Education**Replaced fullstops to underscor es |  |  |
| **Housing**- Replaced  ‘yess’ to  ‘yes’ |  |  |

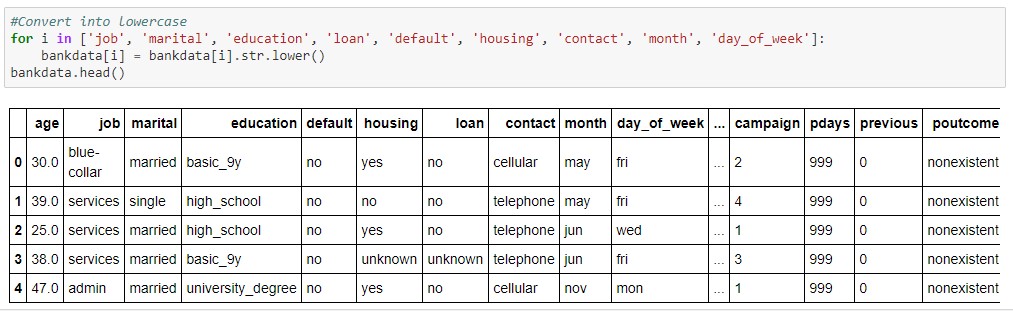
## 1.4 Remove whitespaces

Extra whitespaces were removed using the str.trim function for all categorical variables.



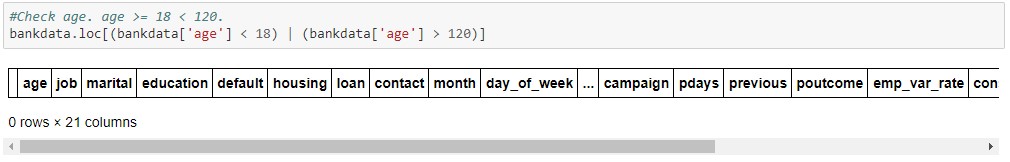
## 1.5 Cast text

All the categorical variables were converted to lower case using the str.lower function

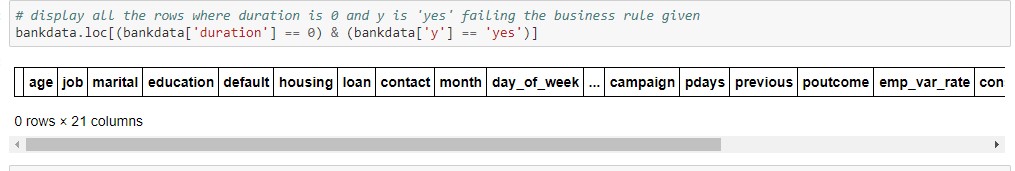


## 1.6 Sanity Checks

The age value was checked for any invalid values.



The duration value was checked if it wasn’t failing the business rule by searching for entries that violated the business rule.



## 1.7 Deal with missing Values

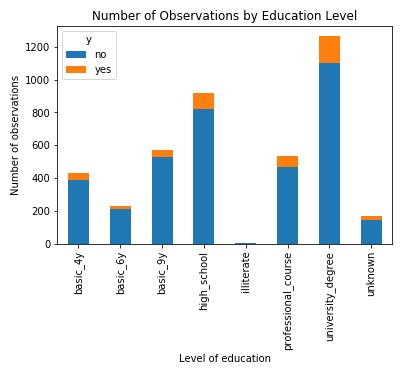
The variables containing null values were found and were replaced with their respective column mean values.



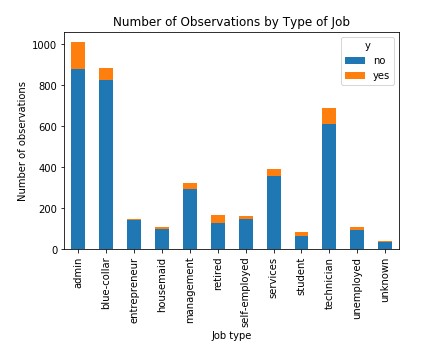
# DATA EXPLORATION

2.1 Visualizations

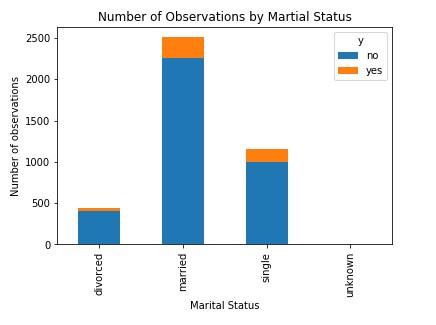
The count of all the categorical values were represented as a stacked bar chart grouped by the ‘y’ value.



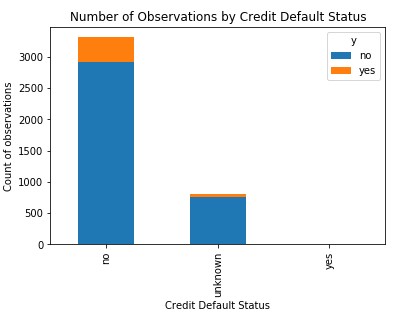
We can see that people with education levels ‘university\_degree’, ‘professional\_course’ and ‘high\_school’ were more successful in choosing yes.



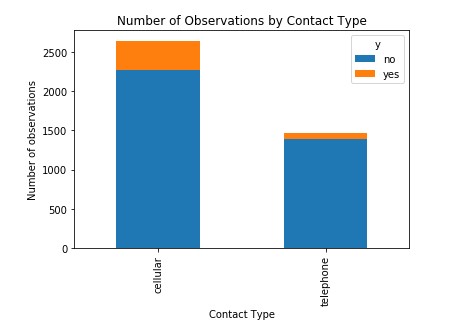
People with job type ‘admin’, ‘blue-collar’ and ‘technician’ were more likely to respond ‘yes’.



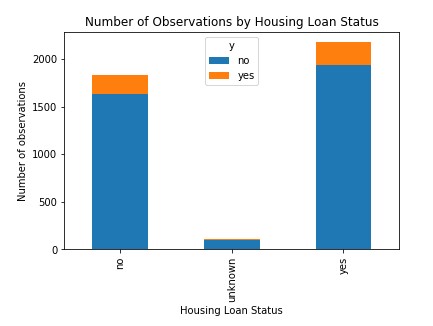
Married and single people were the more successful in subscribing to the term deposit.



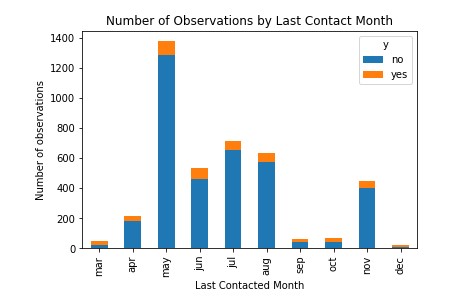
People who did not have a credit default were active in subscribing to the term deposits.



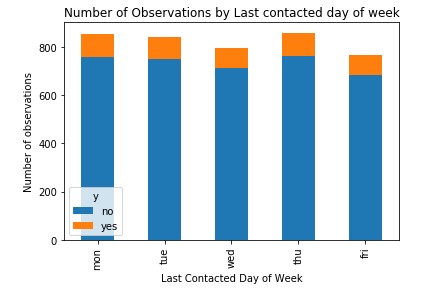
More number of campaigns that were contacted through cellular were successful.



People who did not have a housing loan were more active in responding yes to the campaign.

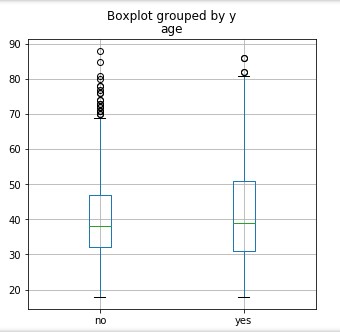


We can observe that more number of contacts were made between the months of may to august but no particular trends are prominent.

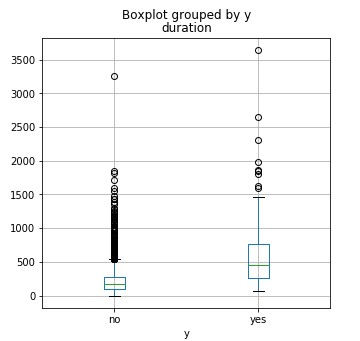


No observable patterns were found in the above graph for the number of observations by last contact day of the week.

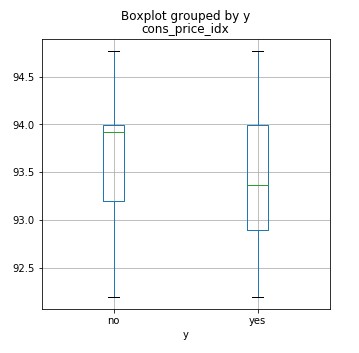
Boxplots were used for numerical values.



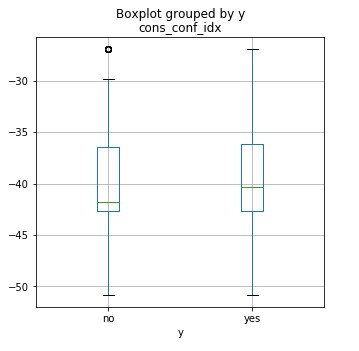
We can the see mean value of age is between 35-40 with very few outliers.



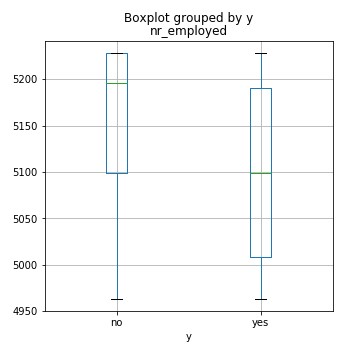
The boxplot for the durarion shows that more number of successful campaigns had more duration of the call than those that responded no.



This boxplot shows the consumer price index grouped by the target variable ‘y’.



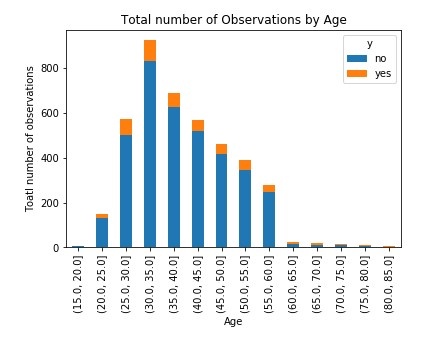
This boxplot shows the customer confidence index grouped by ‘y’. No observable trends are present.



This boxplot shows that people who were employed were more likely to respond ‘yes’.

## 2.2 Hypothesis

Hypothesis 1: People from age groups 25-45 are targeted more and have a higher success rate during the campaign.

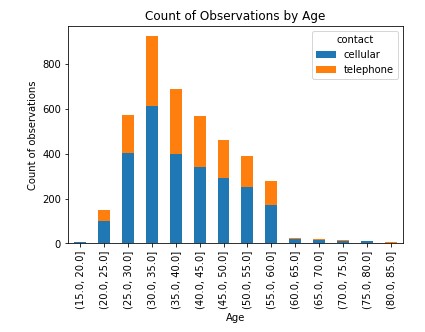


We can observe from this barchart that the people with the age groups 25.0-30.0, 30.0-

35.0, 35.0-40.0 and 40.0-45.0 were more likely to respond yes.

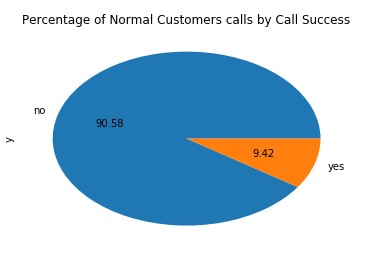
Hypothesis 2: Communication through cellular has more success rate by age group

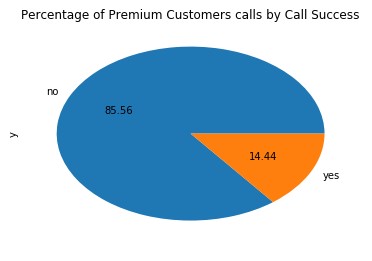
The below plot shows the communication grouped by age.



Hypothesis 3: Customers can be grouped into premium and normal customers.Customers with better backgrounds are more likely to be successful and can be called premium customers. Premium customers can be classified as below.

* job in admin, blue-collar, technician or retired
* marital status as married or single
* education level with high\_school, professional\_course or university\_degree • has no credit defaults





From the above two graphs we can notice that a higher percentage of people who were a part of the premium customers responded yes.

## 2.3 Scatter Plot

The scatter plot shows all the numerical values grouped by y.

