The World AB Problem

In this work, I will expose:

I The Data structure.

II The interpretation of relations between the variables.

Also, I will use R to compute some indicators.

I- The Data structure:

We have two datasets to manipulate during this work. In fact, they are two sets of historical data for the same campaign that come from two different data providers.

Our famous files are:

- dataset_A.csv
- dataset_B.csv

The first thing to do was to discover the variables contained in each one of our CSV files. I started with dataset_A (100 000 observations and 12 variables):

- Call_ID
- Sale
- Agent_ID
- Age
- List_ID
- Phone_code
- First_Name
- Last_Name
- Area_Code
- Gender
- Call_Count

Then, I moved to dataset_B (100 000 observations and 10 variables):

- Call_ID
- Sale
- Agent_ID
- Age
- Phone_code
- First_Name
- Last_Name
- Area_Code
- Gender
- Call_Count

I noticed that we have some missing variables in dataset_B which are:

- List_ID
- Timezone

➡ Our datasets have both 100 000 observations and they have 10 common variables.
Also, in our data, there are some Not Assigned values that we should take care of when we compute any indicator in the next section.

Now, let's select the significant variables. In fact, some variables may be constant so, they will not have a remarkable effect on the prediction model.

```
> summary(A)
    Call_ID              Sale              Agent_ID            Age              List_ID     
 Min.   : 8432514   Mode :logical         : 2315   Min.   : 43.00   Min.   :142.0  
 1st Qu.: 9230592   FALSE:91248     4955   : 1969   1st Qu.: 73.00   1st Qu.:146.0  
 Median : 9527831   TRUE :8720      5077   : 1946   Median : 91.00   Median :147.0  
 Mean   : 9562981   NA's :32        5271   : 1911   Mean   : 95.57   Mean   :147.2  
 3rd Qu.: 9919197                   5001   : 1854   3rd Qu.:111.00   3rd Qu.:149.0  
 Max.   :10476640                   5146   : 1852   Max.   :221.00   Max.   :151.0  
                                    (Other):88153   NA's   :374                     
    Timezone        Phone_code     First_Name       Last_Name        Area_Code        Gender   
 Min.   :0.000   Min.   :37    Ma     : 1625   Ma     : 2938   4068   : 1967   A:52002  
 1st Qu.:2.000   1st Qu.:37    Me     : 1374   MA     : 2136   4092   : 1653   B:46996  
 Median :2.000   Median :37    NL     : 1198   Me     : 2135   3201   : 1520   U: 1002  
 Mean   :1.999   Mean   :37    Sa     : 1089   ME     : 1502   4001   : 1499            
 3rd Qu.:2.000   3rd Qu.:37    Se     : 1044   Mi     : 1482   1632   : 1226            
 Max.   :2.000   Max.   :37    (Other):93134   (Other):89280   7441   : 1226            
                 NA's   :28    NA's   :  536   NA's   :  527   (Other):90909            
   Call_Count    
 Min.   : 1.000  
 1st Qu.: 1.000  
 Median : 2.000  
 Mean   : 3.269  
 3rd Qu.: 4.000  
 Max.   :55.000  
```

The first conclusion is that there are some variables that are almost constant.
These variables are:

- Timezone
- Phone_code

Also, there some variables that are specific for each observation and they are:

- Call_ID
- Agent_ID
- First_Name
- Last_Name

So, I implemented a new data frame after eliminating the previous six variables.

As we can see also, there are some missing values in the following variables:

- Sale
- Age
- Phone_code
- First_Name
- Last_Name

The problem of missing values can be solved by just deleting the corresponding rows or by replacing these NAs by the mean value of the variable or by zeros. These different solutions depend on the studied case.

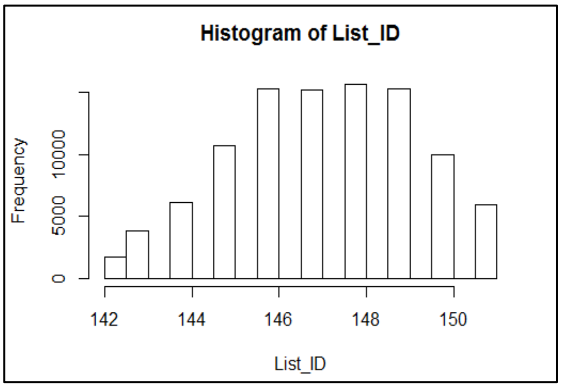
In our case, when I omitted the rows containing NAs, I obtained 99 626 observation which represents 99.62 % of the first data.

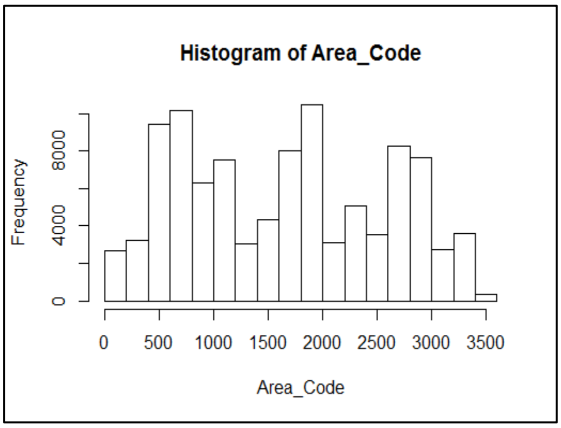
In the following work, I will use the new created data frame named ATN.
The correlation matrix of our data frame is presented by the following screenshot

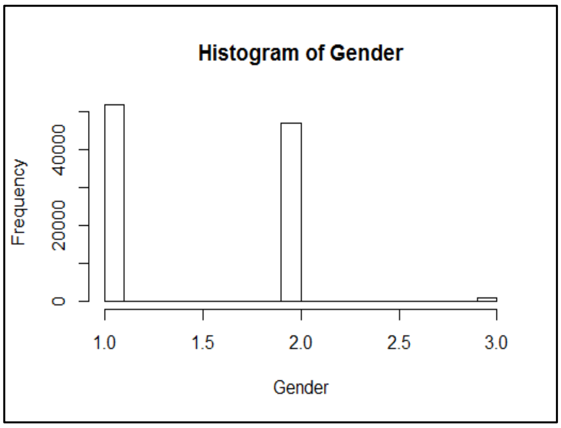
```
> cor(ATN)
                 Sale          Age      List_ID    Area_Code       Gender   Call_Count
Sale        1.00000000  0.050857265  0.013546654 -0.021610615 -0.010576000 -0.045305472
Age         0.05085726  1.000000000 -0.103939843  0.058231135  0.048182120 -0.002358446
List_ID     0.01354665 -0.103939843  1.000000000 -0.002605388 -0.022217057 -0.404145546
Area_Code  -0.02161062  0.058231135 -0.002605388  1.000000000  0.014463867 -0.002978880
Gender     -0.01057600  0.048182120 -0.022217057  0.014463867  1.000000000 -0.003931136
Call_Count -0.04530547 -0.002358446 -0.404145546 -0.002978880 -0.003931136  1.000000000
```

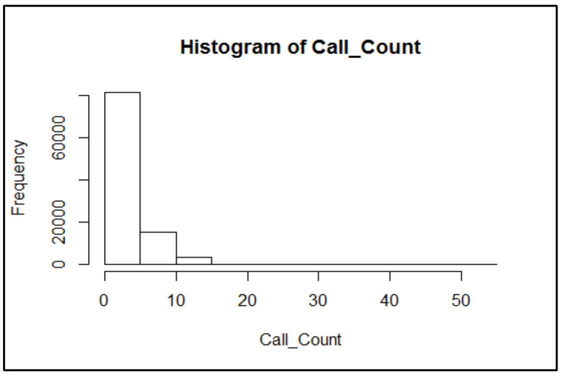
We can conclude that the correlation of the variables with our result Sale is low. So, we can't conclude a trivial relation between one of the variables and Sale.

The histograms of variables are shown in the following screenshots:

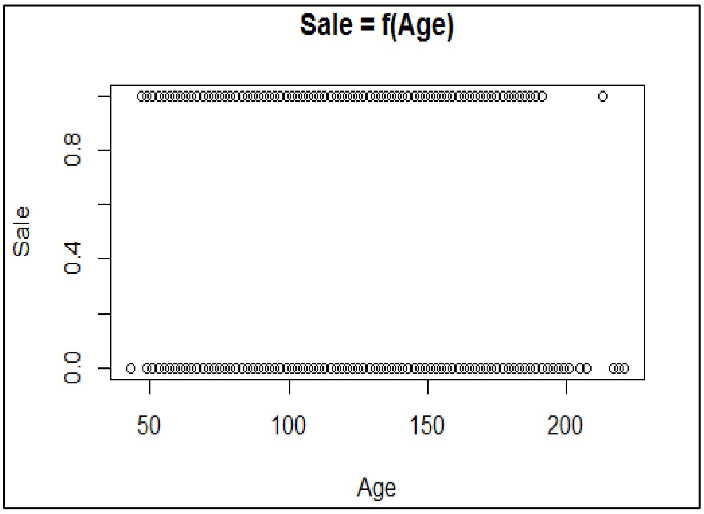


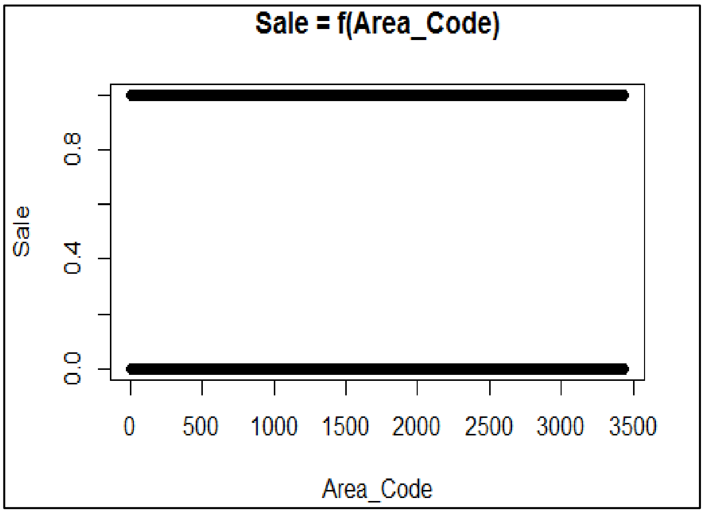


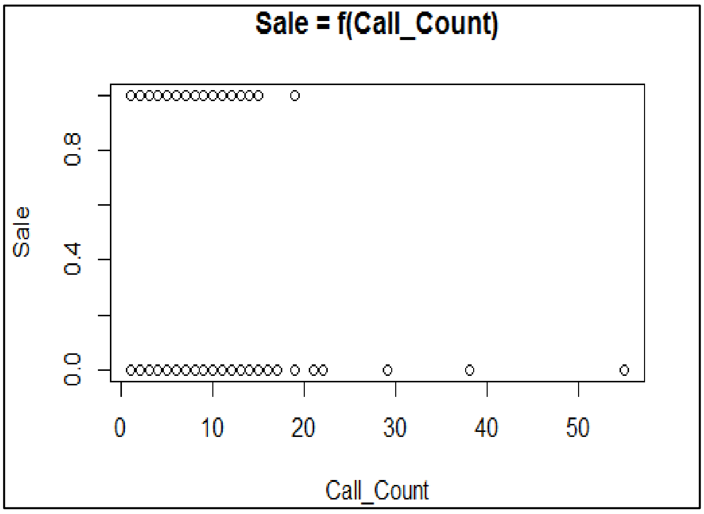




And the plots of Sale in function of each variable are:







II- The proposed model:

I will use the logistic regression for predicting the result of this problem.

I will recommend Fitting Generalized Linear Models for this task. In fact, the model that should be developed for our case, should predicts a probability of making a yes/no choice (Bernoulli variable). Such model is less suitable as a linear-response model, since probabilities are bounded on both ends. Also, the generalized linear models (GLMs) are a broad class of mo dels that include linear regression, ANOVA, Poisson regression, log-linear models etc.

The final result will have the following representation: $\mathbf{Y = \alpha + \beta x1 + \gamma x2 + \ldots}$

In which Y is the expectation of target variable and the second part is the linear combination of predictors (α,β,γ to be predicted).

There are three components to any GLM:

- **Random Component**: refers to the probability distribution of the response variable Y
- **Systematic Component**: specifies the explanatory variables (X_1, X_2, ... X_k) in the model, more specifically their linear combination in creating the so called *linear predictor*
- **Link Function, η or g (μ)**: specifies the link between random and systematic components. It says how the expected value of the response relates to the linear predictor of explanatory variables.

The developed code for the previous parts is in the attached file.