ALY6010

Capstone-Report

Northeastern University, Vancouver, BC, Canada

Mohsen Soltanifar, PhD, AStat

Faculty Lecturer

December 2022

Bank Data Analysis

Shamim Sherafati

Graduate Student of Master of Professional Studies in Analytics
Northeastern University, Vancouver, BC, Canada

Abstract

The data is related to direct marketing campaigns of a banking institution. The classification goal is to predict if the client will subscribe to a term deposit, their account balance and the duration of their campaign's depository. The analysis on this dataset, will help to get most of the answers to these questions by predicting and identifying the right prospects at the right time, building customer loyalty, promoting efficiency across different departments, and getting detailed information on different sectors like loan amount. I used these methods for analyzing data: mean, standard deviation, regression, hypothesis testing, and correlation tests. The results which I gave from each question that I created, show the relations between variable, if the normality distributed or not and at the end, find the answer by hypothesis testing and different charts and figures which shows the results.

Key words: Regression, Q-Q plot, Statistical measure, Standard Error of Estimate plot, Residual plot, Pearson correlation, Duration of loan, and Campaign of depository.

1. Introduction

As technology and financial industry has advanced over the past years, there has been a huge surge in the availability of data. People use such data to improve services, generate business value, advanced science, and make more informed decisions. Based on this, this dataset included some information about its customers like: age, marital status, their job, education, the loan and housing, campaign and its duration and etc. As I previously study and searched on financial industry and the field of bank, so based on this I choose this subject as my dataset and analyzed on it because I think it's something that most of us nowadays familiar with it and became important in our life.

In this project, I created three main questions about the relation between variables in this dataset.
Question1: Only those who are above 40 years has the higher account balance more than 2000. Is it true? (Age is independent variable and Balance is dependent variable).
Question2: The duration of loans under the bank rate with 28 days is equal to 1000. Is it True? (Duration is independent variable and Day is dependent variable).
Question3: Is Campaign depositing for accounts with balance higher than 5000 higher than 10? (Campaign is independent variable and Balance is dependent variable).

2. Materials & Methods

2.1 Dataset

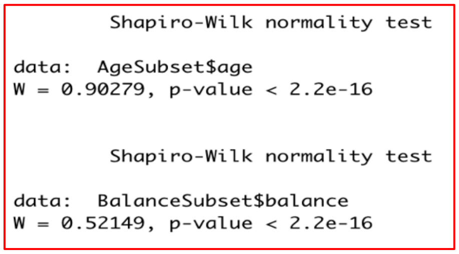
This bank's dataset which obtained from the previous milestone project, consists of both textual and numeric fields, and has a total of 11162 rows and 17 fields which after cleaning decreased to 1118 rows and 16 fields. I get this dataset from Kaggle website and used my previous; ALY 6010, Final project milestone 2 week 4 and referred to my previous r code for analyzing. (Source: file:///Users/shamimsherafati/Documents/Northeastern%20University/ALY%206010/Week%204/Milestone%202/Sherafati_M4--Milestone2.html)

2.2 Statistical Analysis

Statistical software used: For this project I used Excel and R notebook and create R script, RMD, MD and HTML output that helped me to achieve to these outputs.

2.2.1 Check normality with Shapiro test:

Question1: Only those who are above 40 years has the higher account balance more than 2000. Is it true?

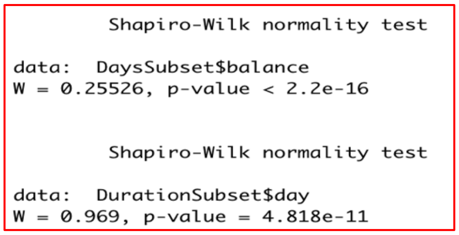


```
        Shapiro-Wilk normality test

data:  AgeSubset$age
W = 0.90279, p-value < 2.2e-16


        Shapiro-Wilk normality test

data:  BalanceSubset$balance
W = 0.52149, p-value < 2.2e-16
```

Getting the subset of vectors for 'age' and 'balance' that are more than 40 and more than 2000 respectively allows me to get the Shapiro test to check its normality which seems they are not normality distributed.

Figure 1: Shapiro test of question 1

Question2: The duration of loans under the bank rate with 28 days is equal to 1000. Is it True?

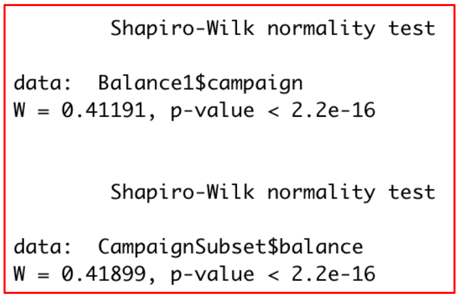


```
        Shapiro-Wilk normality test

data:  DaysSubset$balance
W = 0.25526, p-value < 2.2e-16


        Shapiro-Wilk normality test

data:  DurationSubset$day
W = 0.969, p-value = 4.818e-11
```

The subset of vectors for 'Duration and 'Day that is equal to 28 and more than 1000 respectively in the Shapiro test shows that they are not normality distributed.

Figure 2: Shapiro test of question 2

Question3: Is Campaign depositing for accounts with balance higher than 5000 higher than 10?



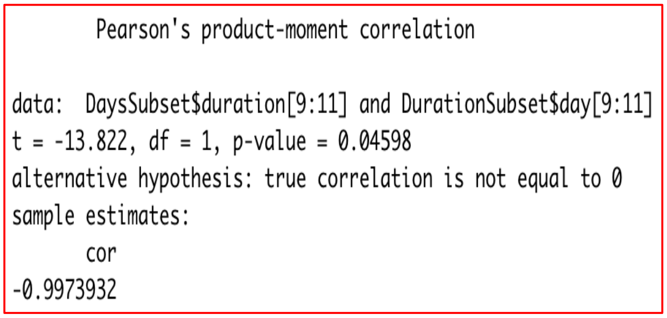
```
        Shapiro-Wilk normality test

data:  Balance1$campaign
W = 0.41191, p-value < 2.2e-16


        Shapiro-Wilk normality test

data:  CampaignSubset$balance
W = 0.41899, p-value < 2.2e-16
```

The subset of vectors for 'Campaign and 'Balance that is more than 10 and more than 5000 respectively and called CampaignSubset and Balance1 respectively in the Shapiro shows that they are not normality distributed.

Figure 3: Shapiro test of question 3

2.2.2 Pearson correlation test:

Question1: Only those who are above 40 years has the higher account balance more than 2000. Is it true?
I Check the relation between Age and Account Balance by using correlation table.

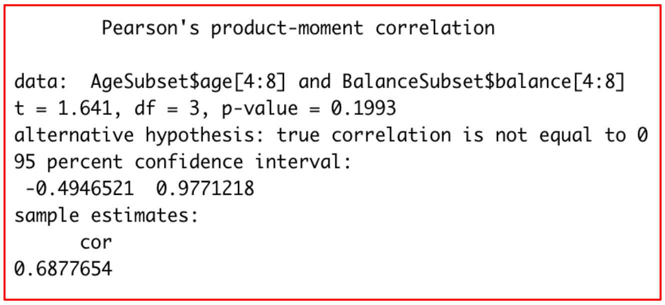


```
	Pearson's product-moment correlation

data:  DaysSubset$duration[9:11] and DurationSubset$day[9:11]
t = -13.822, df = 1, p-value = 0.04598
alternative hypothesis: true correlation is not equal to 0
sample estimates:
       cor
-0.9973932
```

Figure 4: Pearson correlation test of question 1

The t-test statistic value is 1.641, the degrees of freedom is 3, p-value is 0.1993, the confidence interval of the correlation coefficient at 95% is [-0.4946, 0.9771], and the correlation coefficient is 0.68 which means it has a moderate correlation.

Question2: The duration of loans under the bank rate with 28 days is equal to 1000. Is it True?

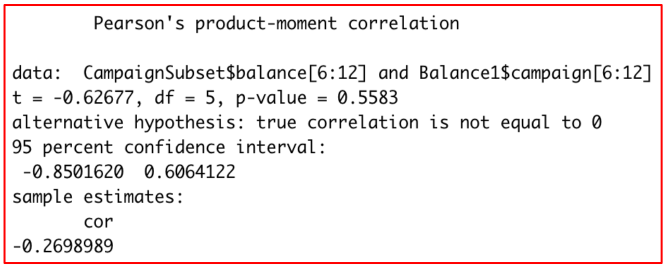


```
	Pearson's product-moment correlation

data:  AgeSubset$age[4:8] and BalanceSubset$balance[4:8]
t = 1.641, df = 3, p-value = 0.1993
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.4946521  0.9771218
sample estimates:
      cor
0.6877654
```

Figure 5: Pearson correlation test of question 2

After checking the relation between Duration and Day by using correlation table, got the following results:
the t-test statistic value is -13.82, the degrees of freedom is 1, p-value is 0.04598, and the correlation coefficient -0.99 which is very high.

Question3: Is Campaign depositing for accounts with balance higher than 5000 higher than 10?



```
	Pearson's product-moment correlation

data:  CampaignSubset$balance[6:12] and Balance1$campaign[6:12]
t = -0.62677, df = 5, p-value = 0.5583
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.8501620  0.6064122
sample estimates:
       cor
-0.2698989
```

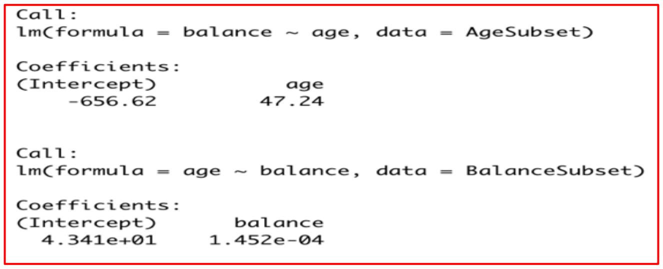
Figure 6: Pearson correlation test of question 3

The t-test statistic value is 0.62), the degrees of freedom is 5, p-value is 0.5583, the confidence interval of the correlation coefficient at 95% is [-0.850, 0.606] and the correlation coefficient -0.269 which is a weak correlation.

2.2.3 Regression Test:

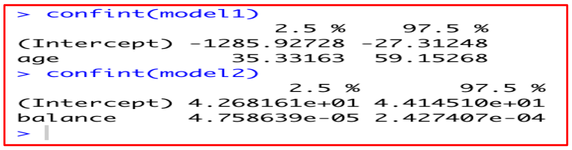
In the regression, dependent variable is correlated with the independent variable. This means, as the value of the independent variable changes, value of the dependent variable also changes.

Question1: Only those who are above 40 years has the higher account balance more than 2000. Is it true?



```
Call:
lm(formula = balance ~ age, data = AgeSubset)

Coefficients:
(Intercept)          age
    -656.62        47.24


Call:
lm(formula = age ~ balance, data = BalanceSubset)

Coefficients:
(Intercept)      balance
  4.341e+01    1.452e-04
```

Figure 6: Building a model as regression in question 1

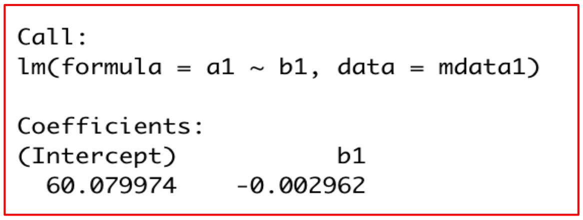


```
> confint(model1)
                   2.5 %    97.5 %
(Intercept) -1285.92728 -27.31248
age            35.33163  59.15268
> confint(model2)
                   2.5 %       97.5 %
(Intercept) 4.268161e+01 4.414510e+01
balance     4.758639e-05 2.427407e-04
> 
```

Figure 7: Confint a model as regression in question 1

In this stage, I got the Regression Test for this question to determine the relationship between the dependent and independent variables of the dataset.

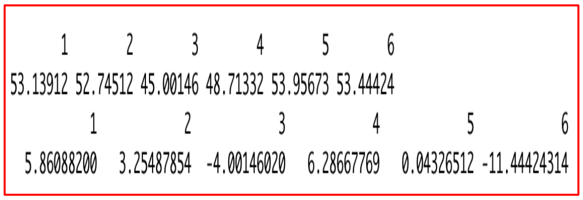
The mean value of the ‘age’ when all the predictor variables in the model are equal to zero is between -656.62 to 47.24 and the mean value of the ‘balance’ when all the predictor variables in the model are equal to zero is between 4.341e+01 to 1.452e-04.



```
Call:
lm(formula = a1 ~ b1, data = mdata1)

Coefficients:
(Intercept)           b1
  60.079974    -0.002962
```

Figure 8: Fit a model as regression in question 1

Here, I created a subset for age and balance in columns 1 to 6 and get its column bind function to fit linear models to data frames in the R that can be used to carry out regression and analysis of covariance to predict the value corresponding to data that is not in the data frame.

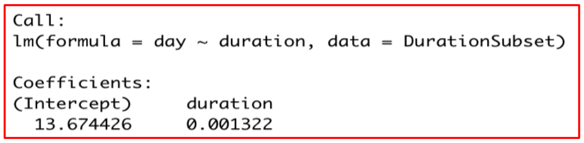


```
       1        2        3        4        5        6 
53.13912 52.74512 45.00146 48.71332 53.95673 53.44424 
          1           2           3           4           5           6 
 5.86088200  3.25487854 -4.00146020  6.28667769  0.04326512 -11.44424314 
```

Figure 9: residual a fit of a model as regression in question 1

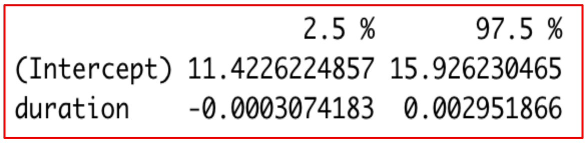
This residual (the difference between the observed outcomes and the predicted outcomes) which I created from the previous ‘Fit1’ table, extracts model from objects returned by modeling functions because all the objects which are returned by model fitting functions should provide a residual method.

Question2: The duration of loans under the bank rate with 28 days is equal to 1000. Is it True?



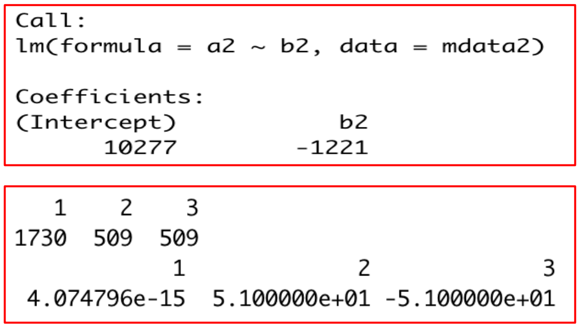
```
Call:
lm(formula = day ~ duration, data = DurationSubset)

Coefficients:
(Intercept)     duration
  13.674426     0.001322
```

Figure 10: Building a model as regression in question 2



```
                  2.5 %      97.5 %
(Intercept) 11.4226224857 15.926230465
duration    -0.0003074183  0.002951866
```

Figure 11: Confint a model as regression in question 2



```
Call:
lm(formula = a2 ~ b2, data = mdata2)

Coefficients:
(Intercept)           b2
      10277        -1221
```

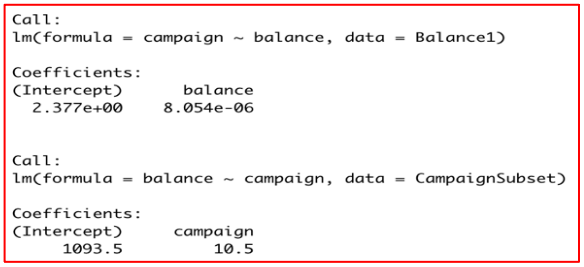
```
   1    2    3 
1730  509  509 
            1             2             3 
 4.074796e-15  5.100000e+01 -5.100000e+01 
```

Figure 12&13: Fit and residual a model as regression in question 2

The mean value of the ‘duration when all the predictor variables in the model are equal to zero is between 13.68 to 0.0013. With a subset for duration and day in columns 9 to 11 and get its column bind function to fit linear models to data frames in the R that can be used to carry out regression and analysis of covariance to predict the value corresponding to data that is not in the data frame.

This residual which I created from the previous ‘Fit2’ table, extracts model from objects returned by modeling functions because all the objects which are returned by model fitting functions should provide a residual method.

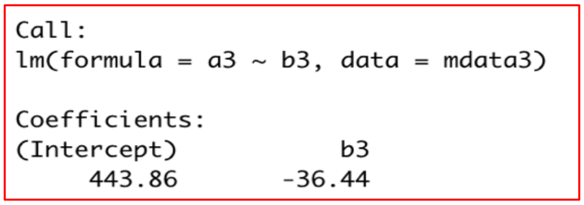
Question3: Is Campaign depositing for accounts with balance higher than 5000 higher than 10?



```
Call:
lm(formula = campaign ~ balance, data = Balance1)

Coefficients:
(Intercept)      balance
  2.377e+00    8.054e-06


Call:
lm(formula = balance ~ campaign, data = CampaignSubset)

Coefficients:
(Intercept)     campaign
     1093.5         10.5
```

Figure 14: Building a model as regression in question 3

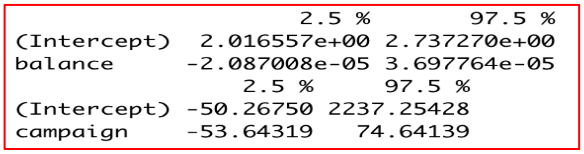
The mean value of the 'campaign' when all the predictor variables in the model are equal to zero is between 1093.5 to 10.5 and the mean value of the 'balance when all of the predictor variables in the model are equal to zero is between 2.377e+00 to 8.054e-06.



```
Call:
lm(formula = a3 ~ b3, data = mdata3)

Coefficients:
(Intercept)           b3
     443.86       -36.44
```

Figure 15: Fit a model as regression in question 3

A subset for campaign and balance in columns 6 to 12 and get its column bind function to fit linear models to data frames in the R that can be used to carry out regression and analysis of covariance to predict the value corresponding to data that is not in the data frame.

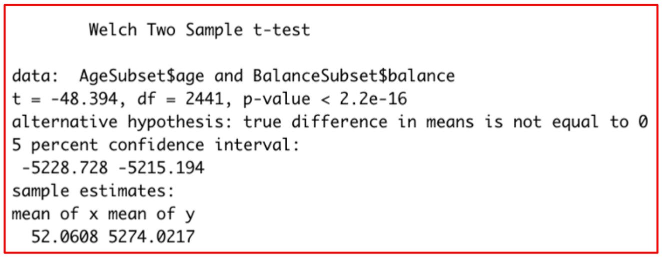


```
                    2.5 %       97.5 %
(Intercept)  2.016557e+00 2.737270e+00
balance     -2.087008e-05 3.697764e-05
                 2.5 %      97.5 %
(Intercept) -50.26750 2237.25428
campaign    -53.64319   74.64139
```

Figure 16: Confint a model as regression in question 3

2.2.4 Hypothesis Testing:

Question1: Only those who are above 40 years has the higher account balance more than 2000. Is it true?

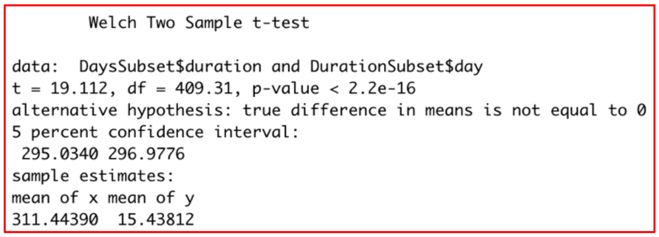


```
        Welch Two Sample t-test

data:  AgeSubset$age and BalanceSubset$balance
t = -48.394, df = 2441, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
5 percent confidence interval:
 -5228.728 -5215.194
sample estimates:
mean of x mean of y
  52.0608 5274.0217
```

H0= Above 40 years old's people, have the account balance higher than 2000.
H1= Having the higher account balance of more than 2000, is not limited to above 40 years old.

Figure 17: Hypothesis test for question 1

Question2: The duration of loans under the bank rate with 28 days is equal to 1000. Is it True?

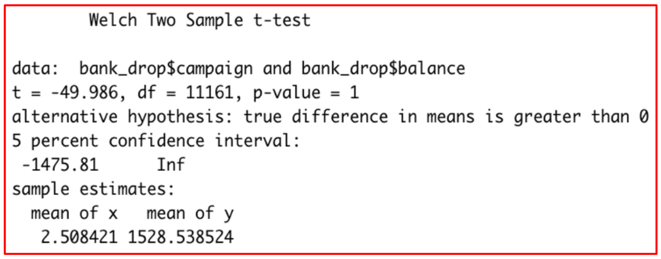


```
        Welch Two Sample t-test

data:  DaysSubset$duration and DurationSubset$day
t = 19.112, df = 409.31, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
5 percent confidence interval:
 295.0340 296.9776
sample estimates:
mean of x mean of y
311.44390  15.43812
```

H0= Duration of loans are equal to 1000 with 28 days.
H1= Duration of loans are not equal to 1000 with 28 days.

Figure 18: Hypothesis test for question 2

Question3: Is Campaign depositing for accounts with balance higher than 5000 higher than 10?



```
        Welch Two Sample t-test

data:  bank_drop$campaign and bank_drop$balance
t = -49.986, df = 11161, p-value = 1
alternative hypothesis: true difference in means is greater than 0
5 percent confidence interval:
 -1475.81      Inf
sample estimates:
  mean of x   mean of y
   2.508421 1528.538524
```

H0= Campaign depository with balance higher than 5000 is equal or higher than 10
H1= Campaign depository with balance higher than 5000 is not higher than 100.

Figure 19: Hypothesis test for question 3

3. Results

3.1 Exploratory Data Analysis (EDA)

Summary of Initial EDA: I've selected a few fields which I interested in exploring to see how they affect bank account situations. My fields of interests are Account balance, duration 0f loans, and campaign depository and after cleaning data start working on the questions. In this section, I provided some descriptive analysis for each question which can be seen below:

Question1: Only those who are above 40 years has the higher account balance more than 2000. Is it true?
I got the Shapiro test to check its normality and I found that both variables have the p-value lower than Sig.level which demonstrate that they are not normality distributed. Then, I checked their correlation with Pearson's test and saw that the p-value of the test is 0.1993 which is more than the significance level alpha = 0.05. We conclude that there is not a significant linear correlation between Age and Account Balance. After that, I created the model and get the confint of these variables and the results demonstrate that the confint of 'age' and 'balance' which can be seen, shows that when you repeat your experiment many times, the true parameter value will be below the CI in 2.5% of cases and above it in another 2.5% of cases - and the CI will cover it in 95% of cases, So the "standard" 95% CI consists of a lower 2.5% limit and an upper 97.5% limit.

Call	Residuals	Coefficients	Signif. codes	Residual standard error	Multiple R-squared	F-statistic
lm(formula = balance ~ age, data = AgeSubset)	Min -8505 1Q -1559 Median -1052 3Q 255 Max 77892	Estimate Std. -656.620 Error 321.003 t value -2.046 Pr(>|t|) -0.0409 * (Intercept) age 47.242 6.075 7.776 9.04e-15 ***	0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1	3862 on 4965 degrees of freedom	0.01203, Adjusted R-squared: 0.01183	60.47 on 1 and 4965 DF, p-value: 9.044e-15

Call	Residuals	Coefficients	Signif. codes	Residual standard error	Multiple R-squared	F-statistic
lm(formula = age ~ balance, data = BalanceSubset)	Min -24.428 1Q -10.714 Median -2.184 3Q 9.165 Max 51.255	Estimate Std. Error t value Pr(>|t|) (Intercept) 4.341e+01 3.732e-01 116.339 < 2e-16 *** balance 1.452e-04 4.976e-05 2.917 0.00356 **	0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1	13.11 on 2440 degrees of freedom	0.003476, Adjusted R-squared: 0.003067	8.51 on 1 and 2440 DF, p-value: 0.003564

Figure 20: Summary of model(1) for question 1
Figure 21: Summary of model(2) for question 1

The Multiple R-squared which tells us what percentage of the variation within our dependent variable that the independent variable is explaining is about to 0 in balance subset. Also, the average of residual standard error it makes which is about 13. At the end, I got the hypothesis test to check if those who are above 40 years old, have higher account balance of more than 2000 or not. As a result, the p value which is lower than significant level and reject the null hypothesis, so we can conduct that having the account balance more than 2000 is not limited to the 40 years age above.

Question2: The duration of loans under the bank rate with 28 days is equal to 1000. Is it True?
The Shapiro test for checking its normality shows the same results as the previous question that both variables have the p-value lower than Sig.level which demonstrate that they are not normality

distributed. Then, by checking their correlation with Pearson’s test, noticed that the p-value of the test is 0.04598 which is less than the significance level alpha = 0.05. We conclude that it exists a relationship between duration and day. After that, by creating a model for these variables in purpose of the getting the confint of these variables, in the results, the confint of ‘duration’ which can be seen, shows that when you repeat your experiment many times, the true parameter value will be below the CI in 2.5% of cases and above it in another 2.5% of cases - and the CI will cover it in 95% of cases, So the "standard" 95% CI consists of a lower 2.5% limit and
an upper 97.5% limit.

Call	Residuals	Coefficients	Signif. codes	Residual standard error	Multiple R-squared	F-statistic
lm(formula = balance ~ campaign, data = CampaignSubset)	Min -2169.0 1Q -1218.4 Median -974.5 3Q -192.7 Max 24070.5	Estimate Std. Error t value Pr(>|t|) (Intercept) 1093.49 580.17 1.885 0.0609 . campaign 10.50 32.54 0.323 0.7473	0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1	3253 on 208 degrees of freedom	0.0005004, Adjusted R-squared: - 0.004305	0.1041 on 1 and 208 DF, p-value: 0.7473

Figure 22: Summary of model(1) for question 2

Figure 23: Summary of model(2) for question 2

Call	Residuals	Coefficients	Signif. codes	Residual standard error	Multiple R-squared	F-statistic
lm(formula = campaign ~ balance, data = Balance1)	Min -2.031 1Q -1.433 Median -0.467 3Q 0.549 Max 60.550	Estimate Std. Error t value Pr(>|t|) (Intercept) 2.377e+00 1.836e-01 12.948 <2e-16 *** balance 8.054e-06 1.473e-05 0.547 0.585	0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1	3.101 on 769 degrees of freedom	0.0003884,Adjusted R-squared: - 0.0009115	0.2988 on 1 and 769 DF, p-value: 0.5848

The Multiple R-squared within our dependent variable is so little near to 0. Also, the average of residual standard error it makes which is about 3 in balance subset and 3252 in campaign subset. Finally, by getting the hypothesis test, I got the answer to the question which the lower amount of p-value in compare of the significant level, shows that we should reject the null hypothesis and it can be conclude that the duration of loans under the bank rate with 28 days is not equal to 100 and its below this amount.

Question3: Is Campaign depositing for accounts with balance higher than 5000 higher than 10? The Shapiro test for this question which demonstrate that they are not normality distributed as their p-value is lower than sig.level. Then, from the correlation test, we conclude that there is not a significant linear correlation between Campaign and Balance as the p-value of the test is 0.5583 which is more than the significance level alpha = 0.05. After that, the regression test which I used it by creating the model was showing that confint of ‘campaign’ and ‘balance’ shows when you repeat your experiment many times, the true parameter value will be below the CI in 2.5% of cases and above it in another 2.5% of cases - and the CI will cover it in 95% of cases.

Call	Residuals	Coefficients	Signif. codes	Residual standard error	Multiple R-squared	F-statistic
lm(formula = a2 ~ b2, data = mdata2)	1 2 3 4.075e-15 5.100e+01 -5.100e+01	Estimate Std. Error t value Pr(>|t|) (Intercept) 10277.00 678.51 15.15 0.042 * b2 -1221.00 88.33 -13.82 0.046 *	0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1	72.12 on 1 degrees of freedom	0.9948,Adjusted R-squared: 0.9896	191.1 on 1 and 1 DF, p-value: 0.04598

Figure 24: Summary of model for question 3

The Multiple R-squared is 99% means 99% of the variance of the dependent variable being studied is explained by the variance of the independent variable. Also, the average of residual standard error it makes which is about 72. Finally, the hypothesis testing shows that we cannot reject the null hypothesis as the p-value is equal to 1, so it can be concluded that the campaign depositing for accounts with balance higher than 5000 is higher than 10.

3.2 Question1: Age and Account Balance

Question1: Only those who are above 40 years has the higher account balance more than 2000. Is it true? In this section, I describe the analysis process with plots and charts to be more understandable. These two Q-Q plots which I created for check the normality. Their distribution is not followed by a normal trend as its not close to the straight line and even one of them form a curve instead of a straight line. The correlation plots show the relation between Age and Account Balance that shows there is not a significant linear correlation between Age and Account Balance**.** The ggcore function which is another way to find the relation between variables. For example, in age subset plot which we filter it for only above 40 years old, its relationship with balance is 0 that clearly shows they have no relation with each other.

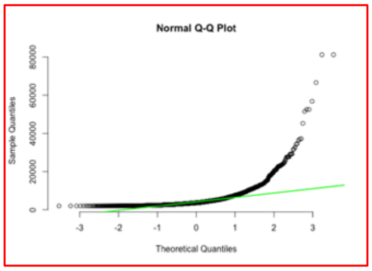


Figure 25: Q-Q plot.

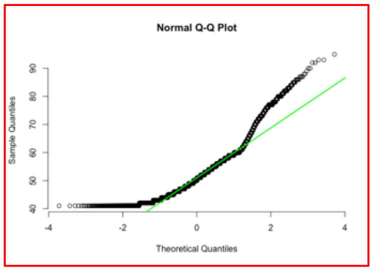


Figure 26: Q-Q plot

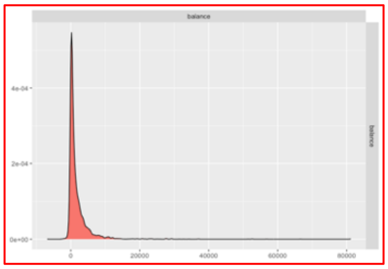


Figure 27: GGally function plot

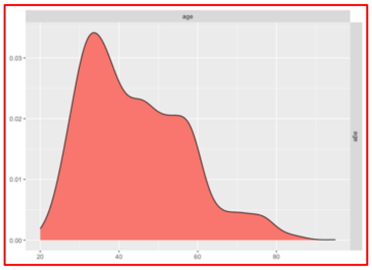


Figure 28: GGally function plot.

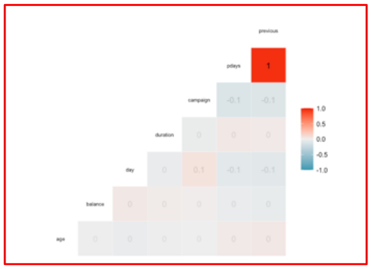


Figure 29: ggcore function plot

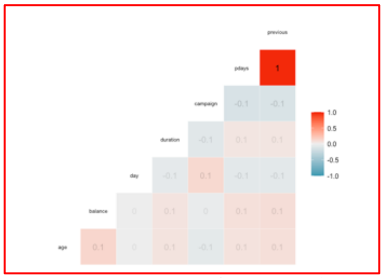


Figure 30: ggcore function plot

The two box plots which I created, is for understanding better the differences between groups in regression test. Now, with the density plot that I created, I measured these two variables statistically. We can see that in balance plot, it experienced downward trend while in age plot its fluctuating plot. The trend plot shows variables follow a positive trend. The higher the Output per age, the higher the balance, and the lower the Output per age, the lower the balance. This Standard Error of Estimate plot which measure the variation of an observation made around the computed regression line that the pink area is the area that measure the standard deviation of epsilons. Finally, I used the fitting function to Fit the Regression Line and its Residuals. As we can see, 4 circles which are under the line are small and have negative residual of about -1 and -6 for red circle but the two circles above the line are large and have positive residual of +6.

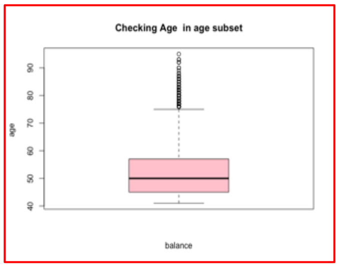


Figure 31: Box plot.

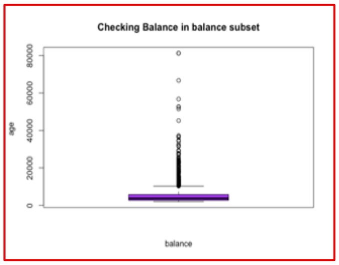


Figure 32: Box plot

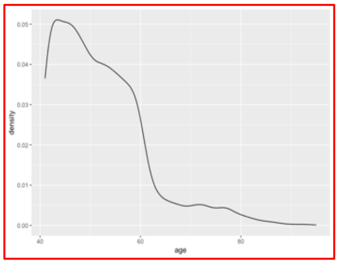


Figure 33: Density plot

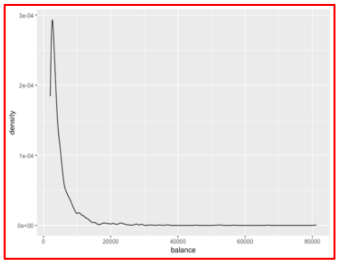


Figure 34: Density plot

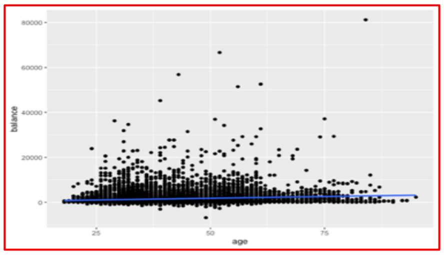


Figure 35: Trends plot.

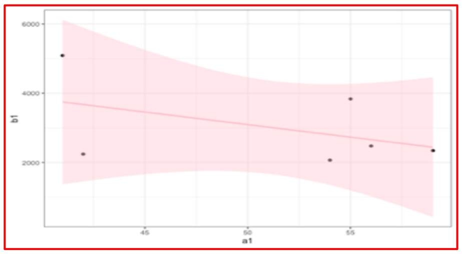


Figure 36: Standard of error estimation

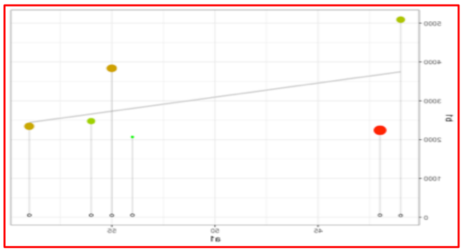


Figure 37:Residual plot

3.3 Question2: Duration and Day

Question2: The duration of loans under the bank rate with 28 days is equal to 1000. Is it True? In this section, I describe the analysis process with plots and charts to be more understandable. The two Q-Q plots which I created for check the normality. Their distribution is followed by a normal and positive trend as its close to the straight line. The correlation plots show the relation between duration and day that shows there is not a significant linear correlation between duration and day. This ggcore function which is another way to find the relation between variables. For example, in duration subset plot which we filter it for 28 days, its relationship with day is 0.1 that positively correlated.

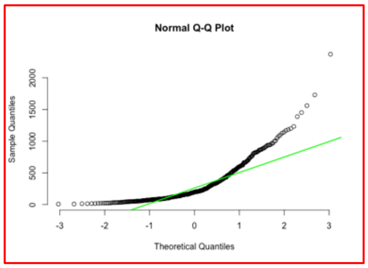


Figure 38: Q-Q plot.

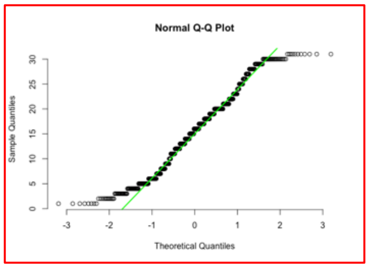


Figure 39: Q-Q plot

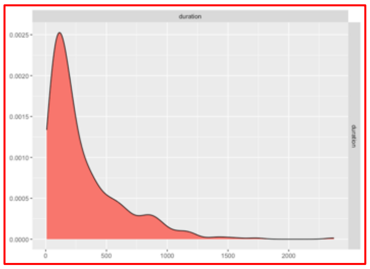


Figure 40: GGally function plot

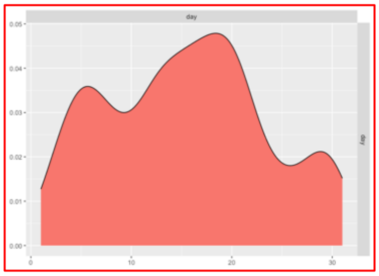


Figure 41: GGally function plot.

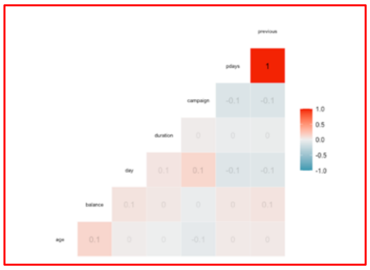


Figure 42: ggcore function plot

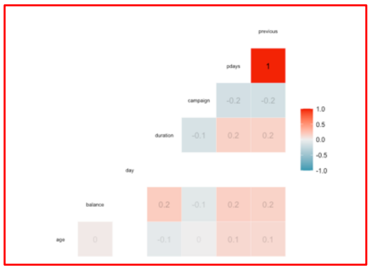


Figure 43: ggcore function plot

The two box plots which I created, is for understanding better the differences between groups in regression test. Now, with the density plot that I created, I measured these two variables statistically. We can see that in duration plot, it experienced downward trend while in day plot its fluctuating plot. The trend plot shows variables follow a positive trend. The higher the Output per day, the higher the duration, and the lower the Output per day, the lower the duration. The Standard Error of Estimate plot which measure the variation of an observation made around the computed regression line that the purple area is the area that measure the standard deviation of epsilons. Finally, I used the fitting function to Fit the Regression Line and its Residuals. As we can see, the two red circles are in the line which is considered has a residual of 0.

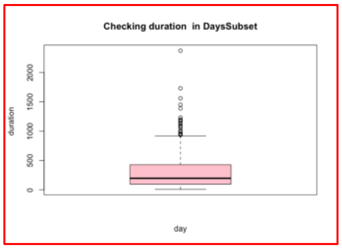


Figure 44: Box plot.

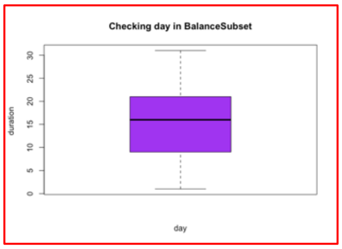


Figure 45: Box plot

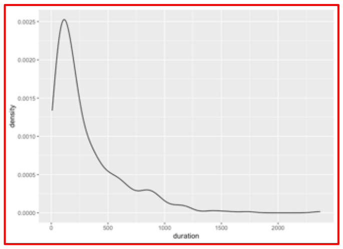


Figure 46: Density plot

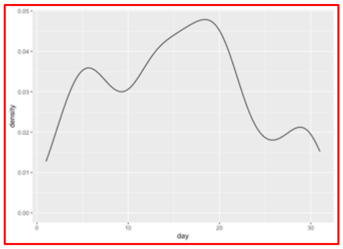


Figure 47: Density plot

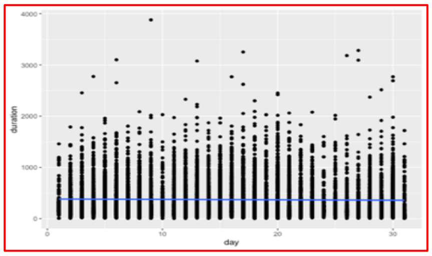


Figure 48: Trends plot.

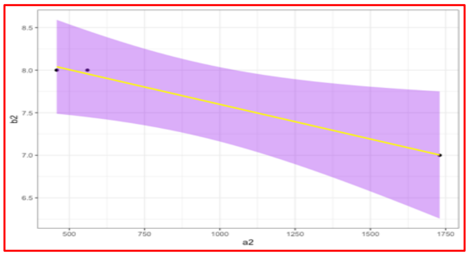


Figure 49: Standard of error estimation

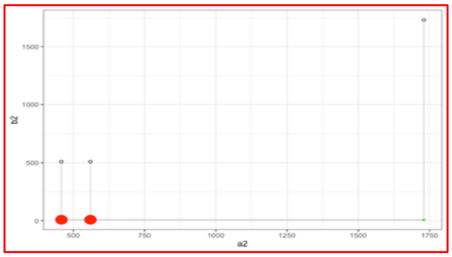


Figure 50:Residual plot

3.4 Question3: Campaign and Balance

Question3: Is Campaign depositing for accounts with balance higher than 5000 higher than 10? Here, I describe the analysis process with plots and charts to be more understandable. The two Q-Q plots which I created for check the normality. Their distribution is not followed by a normal trend as its not close to the straight. The correlation plots show the relation between campaign and balance that shows there is not a significant linear correlation between campaign and balance. The ggcore function which is another way to find the relation between variables. For example, in campaign subset plot which we filter it for less than 10, its relationship with balance is 0 which means there is no relation between them. Also, the campaign and duration relationship show that its -0.3 which means there are negatively correlated.

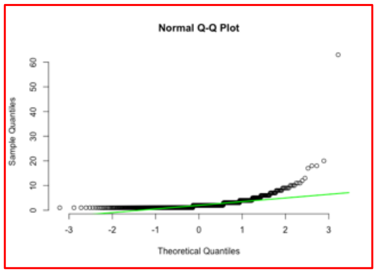


Figure 51: Q-Q plot.

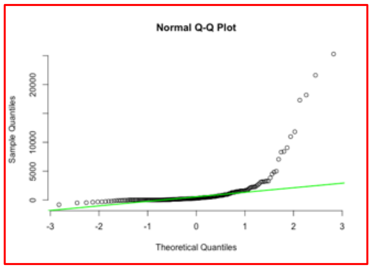


Figure 52: Q-Q plot

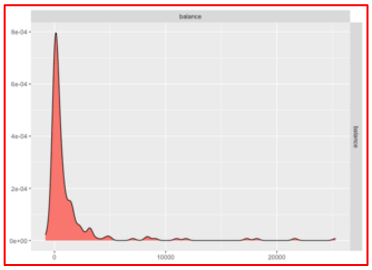


Figure 53: GGally function plot

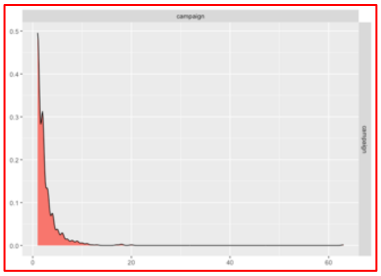


Figure 54: GGally function plot.

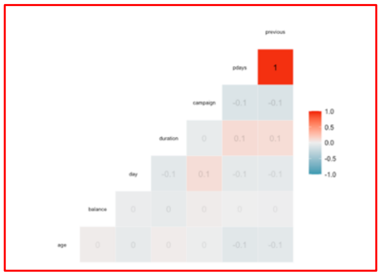


Figure 55: ggcore function plot

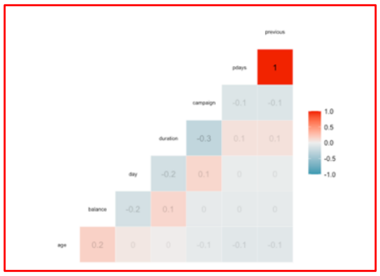
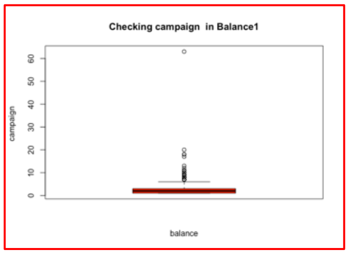
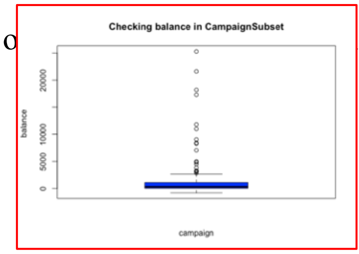
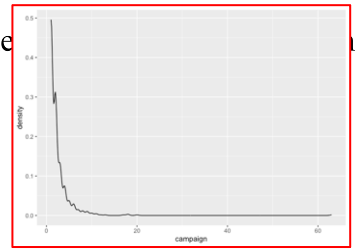


Figure 56: ggcore function plot

The two box plots which I created, is for understanding better the differences between groups in regression test. Now, with the density plot that I created, I measured these two variables statistically. We can see that in both campaign and balance plot there is a downward trend. The trend plot shows variables follow a positive trend as it has a straight line. The higher the Output per campaign, the higher the balance, and the lower the Output per campaign, the lower the balance. The Standard Error of Estimate plot which measure the variation of an observation made around the computed regression line that the yellow area is the area that measure the standard deviation of epsilons. At the end, I used the fitting function to Fit the Regression Line and its Residuals. As we can see, all of the red and green circles are in the line which is considered has a







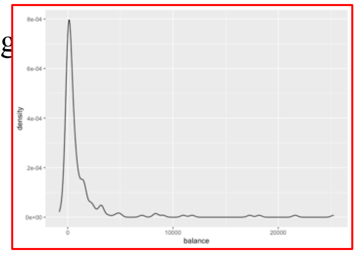
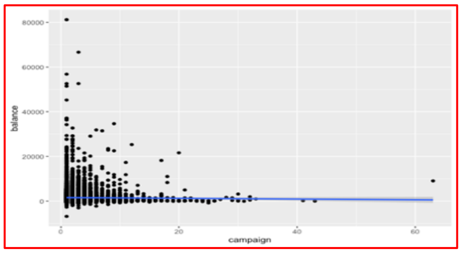


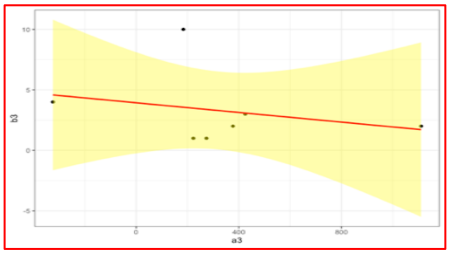
Figure 57: Box plot.

Figure 58: Box plot

Figure 59: Density plot

Figure 60: Density plot





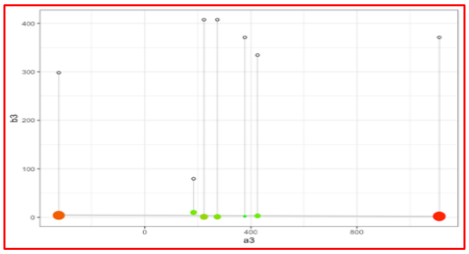


Figure 48: Trends plot.

Figure 49: Standard of error estimation

Figure 50:Residual plot

The reason of choosing these three questions for analysing on them?

As these are the questions which most of us are familiar with it in our everyday life, so will help to understand it better. I want to respond to each question step by step by checking their normality distribution, taking correlation tests on them, check their trend and measure of variation of an observation on each regression line and at the end, take hypothesis testing to find the answer for them.

4. Discussion

4.1 This work

From our investigation using the dataset obtained, we can make the following conclusions. Our first conclusion from the relationship between day and duration is that the higher output per day, the higher the duration. This trend was the same for age/balance and campaign/balance. The day and duration have a positive trend and a very high correlation and from the hypothesis generated we can conclude that they influence one another, i.e., change in one will result in to change in the other variable. In contrast, age/balance and campaign/balance don't have a normal trend and not significantly linear correlation. Also, from the second investigation, we can conclude that campaign and balance have a weak correlation and their correlation is not statistically significant and not followed by a normal trend. However, while conducting a t-test for the null hypothesis and as we accept the null hypothesis, we can conclude that campaign depository with balance higher than 5000 is higher than 10. Moreover, the correlation for age and balance show that they have a moderate correlation. As for regression taking balance as the response variable and campaign as predictor variable, we note that change in campaign will have a significant impact on balance.

4.2 Limitations

This study has two limitations. The first limitation was about the small domain of some variables which we selected to work on them. Because having the limited and small domain, limit the analysing data process and it makes a bit difficult. So, I applied more chart and table for make it more clear for visualization process. The second limitation is the limited variables of columns. As most of them are character and make it difficult to choose between them to find the best one for creating the questions and start work on them. Finally, we could not control for every possible factor, and the observational nature of this project leaves the possibility of residual confounding.

4.3 Future Work

Finally, based on the study and analyzing which I had on these variables of bank dataset, there are some recommend directions which I think if applied would be better in future work; t would be better if we increased the domain of variables with other useful numeric variables or even other type of data to reduce any limitation which may consist of to have a better data visualization on data. Also, the second direction which would be a good idea and can help us in the quality of analysing is to compare each of the variables with each other to see if they match with each other, their relationship and other factors that give us a better background of them and will be more prepare for creating an analyse on the data and desired variables.

5. References

1. Anderson, E. B. (1991). The Statistical Analysis of Categorical Data (2nd ed.). New York: Springer-Verlag.
2. Bates, D. M. and D. G. Watts (1988). Nonlinear Regression and Its Applications. New York: Wiley.
3. Chatterjee, S. and A. S. Hadi (1988). Sensitivity Analysis in Linear Regression. New York: Wiley.
4. Cook, R. D. and S. Weisberg (1982). Residuals and Influence in Regression. New York: Chapman & Hall.
5. Fox, J. (1997). Applied Regression Analysis, Linear Models, and Related Methods. Thousand Oaks, CA: SAGE Publications.
6. Freund, R. J. and P. D. Minton (1979). Regression Methods: A Tool for Data Analysis. New York: Marcel Dekker.
7. Graybill, F. A. (1969). Introduction to Matrices with Applications in Statistics. Belmont, CA: Wadsworth Publishing Company.
8. Hocking, R. R. (1976). The analysis and selection of variables in linear regression. Biometrics 32, 1–51.
9. Aitkin, Murray, Anderson, Dorothy, Francis, Brian and Hinde, John (1989) Statistical modelling in GLIM, Oxford University Press [182].
10. Atkinson, Anthony C. (1985) Plots, transformations and regression: An introduction to graphical methods of diagnostic regression analysis, Oxford University Press [166, 177].
11. Blyth, T. S. and Robertson, E. F. (2002a) Basic linear algebra, Springer, SUMS [vi, 36, 64, 80, 82, 114].
12. Collett, David (2003) Modelling binary data, 2nd ed., Chapman and Hall/CRC [191].

6. Appendix

```
1. bank = read.csv("~/Downloads/Archive (1)/bank.csv")
2. bank[order(-bank$age),]
3. bank_drop <- bank[, -9, -16]
4. na.omit(bank_drop)
5. names(bank_drop)[names(bank_drop) == 'contact'] <- 'Contact_Info'
6. bank_drop$marital <- gsub("single","SINGLE",as.character(bank$marital))
7. bank_drop$marital <- gsub("married","MARRIED",as.character(bank$marital))
8. bank_drop$marital <- gsub("divorced","DIVORCED",as.character(bank$marital))
9. bank <- dim(bank_drop)
10. str(bank_drop)
11. summary(bank_drop)
12. AgeSubset <- subset(bank_drop, age > 40)
13. BalanceSubset <- subset(bank_drop, balance > 2000)
14. shapiro.test(AgeSubset$age)
15. shapiro.test(BalanceSubset$balance)
16. qqnorm(AgeSubset$age, pch = 1, frame = FALSE)
17. qqline(AgeSubset$age, col = "green", lwd = 2)
18. qqnorm(BalanceSubset$balance, pch = 1, frame = FALSE)
19. qqline(BalanceSubset$balance, col = "green", lwd = 2)
20. GGally::ggpairs(AgeSubset, columns = 6, ggplot2::aes(color= "speciecs"))
21. GGally::ggpairs(BalanceSubset, columns = 1,
```

ggplot2::aes(color= "speciecs"))

22. GGally::ggcorr(AgeSubset, method = c("everything", "spearman"), label = TRUE, label_alpha = TRUE, size=2.5)
23. GGally::ggcorr(BalanceSubset, method = c("everything", "spearman"), label = TRUE, label_alpha = TRUE, size=2.5)
24. cor.test(AgeSubset$age[4:8], BalanceSubset$balance[4:8])
25. model1 <- lm(balance ~ age, data = AgeSubset)
26. model2 <- lm(age ~ balance, data = BalanceSubset)
27. summary(model1)
28. summary(model2)
29. confint(model1)
30. confint(model2)
31. sigma(model1)*100/mean(AgeSubset$age)
32. sigma(model2)*100/mean(BalanceSubset$balance)
33. boxplot(AgeSubset$age , main="Checking Age in age subset", xlab = "balance", ylab = "age", col = "pink")
34. boxplot(BalanceSubset$balance, main="Checking Balance in balance subset", xlab = "balance", ylab = "age", col = "purple")
35. mean(AgeSubset$age)
36. median(AgeSubset$age)
37. ggplot(AgeSubset, aes(age, fill = age)) geom_density(alpha = 1)
38. mean(BalanceSubset$balance)
39. median(BalanceSubset$balance)
40. ggplot(BalanceSubset, aes(balance, fill = balance)) , geom_density(alpha = 1)
41. ggplot(bank_drop, aes(age, balance)) + geom_point() + stat_smooth(method = lm)
42. a1 = AgeSubset$age[1:6]
43. b1 = BalanceSubset$balance[1:6]
44. mdata1 <- cbind.data.frame(a1,b1)
45. p1 <- ggplot(mdata1, aes(x=a1, y=b1))+ geom_point(color="black") theme_bw()
46. p2 <- p1 + geom_smooth(method = lm, color="red", se= FALSE)
47. p3 <- p1 + geom_smooth(method= lm, color= "pink", fill= "pink",se= TRUE)
48. fit <- lm(a1 ~ b1, data = mdata1)
49. mdata1$predicted <- predict(fit)
50. mdata1$residuals <- residuals(fit)

51. ggplot(mdata1, aes(x = a1, y = b1)) + geom_smooth(method = "lm", se = FALSE, color = "lightgrey") + geom_segment(aes(xend = a1, yend = predicted), alpha = .2) + geom_point(aes(color = abs(residuals), size = abs(residuals))) + scale_color_continuous(low = "green", high = "red") + guides(color = FALSE, size = FALSE) + geom_point(aes(y = predicted), shape = 1) + theme_bw()
52. fit1 <- lm(a1~ b1, data= mdata1)
53. summary(fit1)
54. res <- residuals(fit1)
55. fitted(fit1)
56. residuals(fit1)
57. test1 <- t.test(AgeSubset$age, BalanceSubset$balance , mu= 0 ,conf.level = 0.05)
58. attributes(test1)

59. DaysSubset <- subset(bank_drop, day == 28)
60. DurationSubset <- subset(bank_drop, duration > 1000)
61. shapiro.test(DaysSubset$balance)
62. shapiro.test(DurationSubset$day)
63. qqnorm(DaysSubset$duration, pch = 1, frame = FALSE)
64. qqline(DaysSubset$duration, col = "green", lwd = 2)
65. qqnorm(DurationSubset$day, pch = 1, frame = FALSE)
66. qqline(DurationSubset$day, col = "green", lwd = 2)
67. GGally::ggpairs(DaysSubset, columns = 11, ggplot2::aes(color= "speciecs"))
68. GGally::ggpairs(DurationSubset, columns = 9, ggplot2::aes(color= "speciecs"))
69. GGally::ggcorr(DaysSubset, method = c("everything", "spearman"), label = TRUE, label_alpha = TRUE, size=2.5)
70. GGally::ggcorr(DurationSubset, method = c("everything", "spearman"), label = TRUE, label_alpha = TRUE, size=2.5)

```
71. cor.test(DaysSubset$duration [9:11], DurationSubset$day [9:11])
72. model1 <- lm(duration ~ day, data = DaysSubset)
73. model2 <- lm(day ~ duration, data = DurationSubset)
74. summary(model1)
75. summary(model2)
76. confint(model1)
77. confint(model2)
78. sigma(model1)*100/mean(DaysSubset$duration)
79. sigma(model2)*100/mean(DurationSubset$day)
80. boxplot( DaysSubset$duration , main="Checking duration  in DaysSubset", xlab = "day", ylab = "duration ", col = "pink")
81. boxplot(BalanceSubset$day, main="Checking day in BalanceSubset", xlab = "day", ylab = "duration ", col = "purple")
82. mean(DaysSubset$duration)
83. median(DaysSubset$duration)
84. ggplot(DaysSubset, aes(duration, fill = duration)) + geom_density(alpha = 1)
85. mean(DurationSubset$day)
86. median(DurationSubset$day)
87. ggplot(DurationSubset, aes(day, fill = day)) + geom_density(alpha = 1)
88. ggplot(bank_drop, aes(day, duration)) + geom_point() + stat_smooth(method = lm)
89. a2 = DaysSubset$duration [9:11]
90. b2 = DurationSubset$day [9:11]
91. mdata2<-cbind.data.frame(a2,b2)
92. p1 <- ggplot(mdata2, aes(x=a2, y=b2))+ geom_point(color="black")+ theme_bw()
93. p2 <- p1 + geom_smooth(method = lm, color="yellow", se= FALSE)
94. p3 <- p1 + geom_smooth(method= lm, color= "yellow", fill= "purple",se= TRUE)
95. mdata2<-cbind.data.frame(a2,b2)
96. fit <- lm(a2 ~ b2, data = mdata2)
97. mdata2$predicted <- predict(fit)
98. mdata2$residuals <- residuals(fit)
99. ggplot(mdata2, aes(x = a2, y = b2)) + geom_smooth(method = "lm", se = FALSE, color = "lightgrey") + geom_segment(aes(xend = a2, yend = predicted), alpha = .2) +
100. geom_point(aes(color = abs(residuals), size = abs(residuals))) + scale_color_continuous (low = "green", high = "red") + guides(color = FALSE, size = FALSE) +  geom_point(aes(y = predicted), shape = 1) + theme_bw()

101. fit2 <- lm(a2~ b2, data= mdata2)
102. summary(fit2)
103. res <- residuals(fit2)
104. fitted(fit2)
105. residuals(fit2)
106. test2 <- t.test(DaysSubset$duration, DurationSubset$day, mu= 0 ,conf.level = 0.05)
107. attributes(test2)
108. Balance1 <- subset(bank_drop, balance > 5000 )
109. CampaignSubset <- subset(bank_drop, campaign > 10)
110. shapiro.test(Balance1$campaign)
111. shapiro.test(CampaignSubset$balance)
112. qqnorm(Balance1$campaign, pch = 1, frame = FALSE)
113. qqline(Balance1$campaign, col = "green", lwd = 2)
114. qqnorm(CampaignSubset$balance, pch = 1, frame = FALSE)
115. qqline(CampaignSubset$balance, col = "green", lwd = 2)
116. GGally::ggpairs(Balance1, columns = 12, ggplot2::aes(color= "speciecs"))
117. GGally::ggpairs(CampaignSubset, columns = 6, ggplot2::aes(color= "speciecs"))
118. GGally::ggcorr(Balance1, method = c("everything", "spearman"), label = TRUE, label_alpha = TRUE, size=2.5)
119. GGally::ggcorr(CampaignSubset, method = c("everything", "spearman"), label = TRUE, label_alpha = TRUE, size=2.5)
120. cor.test(CampaignSubset$balance [6:12], Balance1$campaign [6:12])
121. model1 <- lm(campaign ~ balance, data = Balance1)
```

```
122. model2 <- lm(balance ~ campaign, data = CampaignSubset)
123. summary(model1)
124. summary(model2)
125. confint(model1)
126. confint(model2)
127. sigma(model1)*100/mean(DaysSubset$duration)
128. sigma(model2)*100/mean(DurationSubset$day)
129. boxplot( Balance1$campaign , main="Checking campaign  in Balance1", xlab = "balance", ylab = "campaign ", col = "red")
130. boxplot(CampaignSubset$balance, main="Checking balance in CampaignSubset", xlab = "campaign", ylab = "balance ", col = "blue")
131. mean(Balance1$campaign)
132. median(Balance1$campaign)
133. ggplot(Balance1, aes(campaign, fill = campaign)) + geom_density(alpha = 1)
134. mean(CampaignSubset$balance)
135. median(CampaignSubset$balance)
136. ggplot(CampaignSubset,   aes(balance, fill = balance)) + geom_density(alpha = 1)
137. ggplot(bank_drop, aes(campaign, balance)) + geom_point() + stat_smooth(method = lm)
138. a3 = CampaignSubset$balance [6:12]
139. b3 = Balance1$campaign [6:12]
140. mdata3<- cbind.data.frame(a3,b3)
141. p1 <- ggplot(mdata3, aes(x=a3, y=b3)) + geom_point(color="black")+ theme_bw()
142. p2 <- p1 + geom_smooth(method = lm, color="red", se= FALSE)
143. p3 <- p1 + geom_smooth(method= lm, color= "red", fill= "yellow",se= TRUE)
144. fit <- lm(a3 ~ b3, data = mdata3)
145. mdata3$predicted <- predict(fit)
146. mdata3$residuals <- residuals(fit)
147. ggplot(mdata3, aes(x = a3, y = b3)) + geom_smooth(method = "lm", se = FALSE, color = "lightgrey") +geom_segment(aes(xend = a3, yend = predicted), alpha = .2) + geom_point(aes(color = abs(residuals), size = abs(residuals))) + scale_color_continuous(low = "green", high = "red") + guides(color = FALSE, size = FALSE) + geom_point(aes(y = predicted), shape = 1) +theme_bw()

148. fit3 <- lm(a3~ b3, data= mdata3)
149. summary(fit3)
150. res <- residuals(fit3)
151. fitted(fit3)
152. residuals(fit3)
153. test3 <- t.test(bank_drop$campaign, bank_drop$balance, mu= 0 ,conf.level = 0.05, alternative = "greater")
154. attributes(test3)
```