Student Number: D20124995

Name: Mina Jamshidian

Programme Code: Probability And Statistical Inference Math 9102(TU059)

For this project the dataset of data mining assignment was used.(Bank Marketing)

Table of contents: 1

Section 1 - Research Question(s)

Research Question

This analysis is about predicting whether a client with different individual and social characteristics will subscribe to term deposit or not based on last contact duration?

This analysis would help us to find which factors could be considered to be perfect for prediction of determining a subscription situation.
Furthermore this is a project that concerns whether clients' education, their age, last contact duration with two contact types can be influenced on subscription of term deposit by clients.

Hypothesis

- Hypothesis 1:
  - H0: There will be no significant predictor for last call duration that leads clients to subscribed term deposits by different age groups?
  - H1: There will be a significant predictor for last call duration that leads clients to subscribed term deposits by different age groups?

- Hypothesis 2:
  H0: There will be no significant predictor for last call duration that leads clients to subscribed term deposits by different age and different education?
  H1: There will be a significant predictor for last call duration that leads clients to subscribed term deposits by different age and different education?

Section 2 - Dataset

For this project the dataset of data mining assignment was used which is bank marketing, this dataset is about the Portuguese banking direct marketing, the marketing was based on calling with phone. The result of this marketing shows whether the customers will subscribe to term deposit or not.

21 variables exist in a dataset that has 11 categorical variables and 9 numerical variables with 41188 rows. The y variable is the target variable which should be predicted. The datasets variables were divided to following parts:

1. Bank client data: 1- age, 2- job, 3- marital, 4- education, 5- default, 6- housing, 7- loan.
2. Depends on the last contact of the current campaign: 8- contact, 9- month, 10- day_of_week,11- duration.
3. Other attributes: 12- campaign,13- pdays, 14- previous, 15- poutcome.

4. Social and Economic Context Attributes: 16- emp.var.rate, 17- cons.price.idx, 18- cons.conf.idx, 19- euribor3m, 20- nr.employed.
5. Target Variable: 21- y

The dataset is csv format which is *bank-additional-full.csv*, it is imported to the variable with the name of the **bank**.

Loading Libraries:

```
library(lmSupport)
library(stargazer)
library(dplyr)
library(rcompanion)
library(pastecs)
library(lavaan)
library(pander)
library(psych)
library(ggplot2)
library(varhandle)
library(effsize)
library(Hmisc)
library(stringr)
library(lmtest)
library(sampler)
library(gmodels)
library(readr)
library(ROSE)
library(tidyr)
library(gplots)
library(lsr)
library(stargazer)
library(car)
```

Importing Data:

```
# importing the  bank marketing data csv file
bank = read.csv("bank-additional-full.csv",sep = ";")
```

After that it would be better to treat the dataset and and preparing some variables for further analysing as explained in following part:

1. Creating two new variables related to the project question:
   The age_label variable which is dividing the age variable in five categories (1. age > 60 for senior-citizen, 2. age >45 for mid-old 3. age>30 for Mid-age 4. age>15 for Young 5. age<15 for Children)

```
bank= bank%>% mutate(age_label= if_else(age > 60,"senior-citizen",if_else(age>45,"mid-old",if_else(age>30,"Mid-age",if_else(age>15,"Young","Children")))))
```

2. Finding the unknown values then doing the best solution for each variable based on their correlation by the target variable by using the cross table function("CrossTable function | R Documentation", 2021).

```
#searching for unknowns throughout the dataset
colSums(bank == "unknown")
```

```
      age          job      marital    education      default      housing         loan
        0          330           80         1731         8597          990          990
  contact        month  day_of_week     duration     campaign        pdays     previous
        0            0            0            0            0            0            0
 poutcome emp.var.rate cons.price.idx cons.conf.idx  euribor3m  nr.employed            y
        0            0            0            0            0            0            0
```

Figure 1

It was found the *unknown* values for the default (8579), education (1731), housing (990), loan (990) and job (330) and marital (80) variables as it shows in Figure 1.
It was decided to remove the *unknown values for* job, marital and housing then change the *unknown* value for education variable to *university.degree* value based on the correlation with target variable(*y*) as it is obvioused in Figure 2 and because of the number unknown value for default is a lot, the column was removed.
At the end it was checked again for unknown value ,the result was shown in Figure 3.

```
CrossTable(bank$job,bank$y_new)
CrossTable(bank$marital,bank$y_new)
CrossTable(bank$education,bank$y_new)
CrossTable(bank$housing,bank$y_new)

bank = bank %>% filter( job!= "unknown")
bank = bank %>% filter( marital!= "unknown")
bank$education[bank$education=="unknown"]="university.degree"
bank = bank %>% filter( housing!= "unknown")
#removing column default & duration
bank = bank[-c(5)]

#Checking the unknown values again
colSums(bank == "unknown")
```

bank$job	bank$y no	yes	Row Total
admin.	9070	1352	10422
	3.423	26.961	
	0.870	0.130	0.253
	0.248	0.291	
	0.220	0.033	
blue-collar	8616	638	9254
	19.926	156.951	
	0.931	0.069	0.225
	0.236	0.138	
	0.209	0.015	
entrepreneur	1332	124	1456
	1.240	9.767	
	0.915	0.085	0.035
	0.036	0.027	
	0.032	0.003	
housemaid	954	106	1060
	0.191	1.507	
	0.900	0.100	0.026
	0.026	0.023	
	0.023	0.003	
management	2596	328	2924
	0.001	0.006	
	0.888	0.112	0.071
	0.071	0.071	
	0.063	0.008	
retired	1286	434	1720
	37.814	297.849	
	0.748	0.252	0.042
	0.035	0.094	
	0.031	0.011	
self-employed	1272	149	1421
	0.097	0.767	
	0.895	0.105	0.035
	0.035	0.032	
	0.031	0.004	
services	3646	323	3969
	4.375	34.458	
	0.919	0.081	0.096
	0.100	0.070	
	0.089	0.008	
student	600	275	875
	40.090	315.775	
	0.686	0.314	0.021
	0.016	0.059	
	0.015	0.007	
technician	6013	730	6743
	0.147	1.156	
	0.892	0.108	0.164
	0.165	0.157	
	0.146	0.018	
unemployed	870	144	1014
	0.985	7.758	
	0.858	0.142	0.025
	0.024	0.031	
	0.021	0.003	
unknown	293	37	330
	0.000	0.001	
	0.888	0.112	0.008
	0.008	0.008	
	0.007	0.001	
Column Total	36548	4640	41188
	0.887	0.113	

bank$marital	bank$y no	yes	Row Total
divorced	4136	476	4612
	0.464	3.652	
	0.897	0.103	0.112
	0.113	0.103	
	0.100	0.012	
married	22396	2532	24928
	3.450	27.174	
	0.898	0.102	0.605
	0.613	0.546	
	0.544	0.061	
single	9948	1620	11568
	9.778	77.021	
	0.860	0.140	0.281
	0.272	0.349	
	0.242	0.039	
unknown	68	12	80
	0.126	0.990	
	0.850	0.150	0.002
	0.002	0.003	
	0.002	0.000	
Column Total	36548	4640	41188
	0.887	0.113	

bank$education	bank$y no	yes	Row Total
basic.4y	3748	428	4176
	0.486	3.829	
	0.898	0.102	0.101
	0.103	0.092	
	0.091	0.010	
basic.6y	2104	188	2292
	2.423	19.088	
	0.918	0.082	0.056
	0.058	0.041	
	0.051	0.005	
basic.9y	5572	473	6045
	8.065	63.527	
	0.922	0.078	0.147
	0.152	0.102	
	0.135	0.011	
high.school	8484	1031	9515
	0.198	1.561	
	0.892	0.108	0.231
	0.232	0.222	
	0.206	0.025	
illiterate	14	4	18
	0.244	1.918	
	0.778	0.222	0.000
	0.000	0.001	
	0.000	0.000	
professional.course	4648	595	5243
	0.004	0.032	
	0.887	0.113	0.127
	0.127	0.128	
	0.113	0.014	
university.degree	10498	1670	12168
	8.292	65.317	
	0.863	0.137	0.295
	0.287	0.360	
	0.255	0.041	
unknown	1480	251	1731
	2.041	16.079	
	0.855	0.145	0.042
	0.040	0.054	
	0.036	0.006	
Column Total	36548	4640	41188
	0.887	0.113	

bank$housing	bank$y no	yes	Row Total
no	16596	2026	18622
	0.312	2.461	
	0.891	0.109	0.452
	0.454	0.437	
	0.403	0.049	
unknown	883	107	990
	0.023	0.184	
	0.892	0.108	0.024
	0.024	0.023	
	0.021	0.003	
yes	19069	2507	21576
	0.305	2.400	
	0.884	0.116	0.524
	0.522	0.540	
	0.463	0.061	
Column Total	36548	4640	41188
	0.887	0.113	

Figure 2

age	job	marital	education	housing	loan	contact	month	day_of_week	duration	campaign	pdays
0	0	0	0	0	0	0	0	0	0	0	0

previous	poutcome	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed	y	age_label
0	0	0	0	0	0	0	0	0

Figure 3

3. After analysing it was found that the campaign column which is about the number of contacted time, it was found the customer after more than 10 times calling didn't subscribed the term deposit so it would be better to filter the dataset with less than 10 times.

```
#Filter the dataset by campaign of less than 10
bank= bank%>%filter(campaign<10)
```

The variables of interest used in this research are shown below in Figure 4:

Concept	Variable Name	Statistical Type	Possible Values
last contact duration, in seconds	duration	Numerical	Range 0 to 4918
client age	age_label	Categorical	1.age>60(senior-citizen) 2.age>45(mid-old) 3.age>30(Mid-age) 4. age>15(Young) 5.age<15(Children)
client subscribed a term deposit?	y	Categorical	yes and no
contact communication type	contact	Categorical	cellular, telephone
client education	education	Categorical	basic.4y to university.degree

Figure 4

In this step the sample should be chosen, the sample was chosen by Sampling Design & Analysis advised by (Ziegel & Lohr, 2000).

Choosing the sample came in the following part with the name of sbank2 which is the final data that we want to use for further analysis. And the final data balancing came in Figure 4.

```
size <- rsampcalc(nrow(sbank), e=5, ci=95,p=0.5, over=0.1)
sbank2<-ssamp(sbank, size, y_new,over=0.1)
sbank2 %>% count(y_new)
```

y <chr>	n <int>
no	409
yes	53

Figure 4

Statistical Measurement

Each of the variables of interest would be inspected. The numeric variables of interest, that represent the last contact duration with the customer, in seconds. it would be inspected for normality by checking standardised scores for skewness and kurtosis and considering the percentage of standardised scores for the variables fell outside of expected boundaries and creating histograms and QQ plots. Decisionsing about the skewness and the kurtosis came from the advice of (George & Mallory, 2011) that categorizing the distribution as normal when the relevant standardised scores of the skewness and the kurtosis fall in the range +/- 2 with the advice of (Field, Miles & Field, 2012) which categorizing the distribution as normal when 95% of the the variable standardised scores fall within the +/-3.29 bounds for a dataset larger than 80 cases. The categorical variables summary statistics would be identified for analysis.

Duration variable:

Inspecting the Duration variable and its Normality by code:

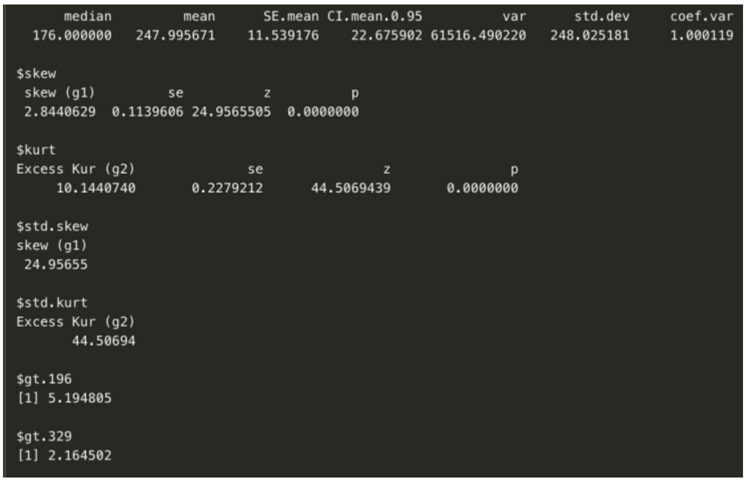
```
# Descriptive statistics
# getting summary statistics for duration variable
duration<- sbank2$duration
duration_d<-describe(duration,omit = TRUE, IQR = TRUE)
duration_s<-list(pastecs::stat.desc(duration, basic = FALSE))

skew        <- semTools::skew(duration)
kurt        <- semTools::kurtosis(duration)
stdskew     <- skew[1] / skew[2]
stdkurt     <- kurt[1] / kurt[2]
zscore      <- abs(scale(duration))
gt196       <- FSA::perc(as.numeric(zscore), 1.96, "gt")
gt329       <- FSA::perc(as.numeric(zscore), 3.29, "gt")

duration_s$skew <- skew
duration_s$kurt <- kurt
duration_s$std.skew <- stdskew
duration_s$std.kurt <- stdkurt
duration_s$gt.196 <- gt196
duration_s$gt.329 <- gt329
duration_s

# Distribution of Age variable with visualization
ggplot(sbank2,aes(x=duration))+labs(x='duration', y='Density')+
  geom_histogram(binwidth = 30,colour='black',aes(y=..density..,fill=..count..))+
  scale_fill_gradient("Count",low="#132B43", high="#56B1F7")+
  stat_function(fun = dnorm,color="red",args = list(mean=mean(duration,na.rm =
TRUE),sd=sd(duration,na.rm = TRUE)))+
  ggtitle('Figure 1: Distribution of Duration')+
  theme(plot.title = element_text(size=10))
```

```
# Create QQ Plot
qqnorm(duration, main = "Figure 2: QQplot of Duration")
qqline(duration,col=2) # show line on the plot
```



```
      median         mean      SE.mean CI.mean.0.95          var      std.dev     coef.var
  176.000000   247.995671    11.539176    22.675902 61516.490220   248.025181     1.000119

$skew
 skew (g1)         se          z          p
 2.8440629  0.1139606 24.9565505  0.0000000

$kurt
Excess Kur (g2)             se              z              p
     10.1440740      0.2279212     44.5069439      0.0000000

$std.skew
skew (g1)
 24.95655

$std.kurt
Excess Kur (g2)
       44.50694

$gt.196
[1] 5.194805

$gt.329
[1] 2.164502
```

Figure 5

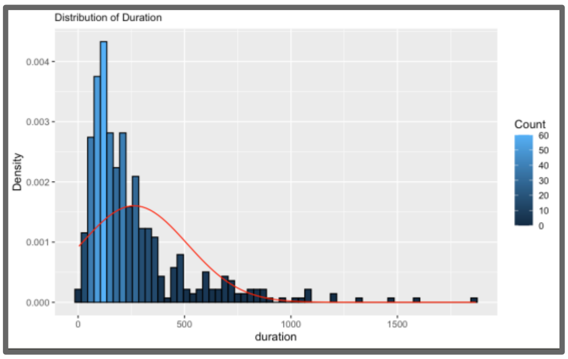


Figure 6

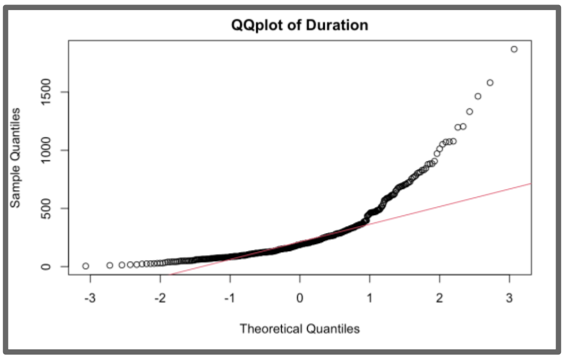


Figure 7

Report of normality analysis for duration variable:

The Duration is represented by a numeric variable in the dataset. Inspection of the standardised scores for skewness and kurtosis reveal that the kurtosis score (*kurtosis* = 44.5, *SE* = .227) and the skewness score (*skewness* = 24.95, *SE* = .113) is out of that range which is the range of range of -2 and 2 . This implies that kurtosis and skewness is not normal. For further inspection using plots such as histogram and normality plot (figure 8 and figure 8), we found that the distribution is positively skewed and not normalized. there are various values deviating from the normality line. On inspection of the count of outliers, there was found 2.1% standardised scores were outside the acceptable range of [-3.29, +3.29] that shows none of the values is outside the 95% Confidence Interval. In total, based on all the tests, it can be said that the data for duration variable will not be treated as a normal within this analysis (*Median*=176, *IQR*=185, *M* = 247.99, *SD* = 248.02, *N* = 462).

Transforming the duration variable:

After inspecting the duration variable,it was found that it was treated as non normal. In this analysis, the Log transformation was performed to this non normal data to convert it into the normal data before doing the linear parametric tests. This transformation approach was chosen because the Parametric test transformed normal data is considered more powerful compared to non parametric test on untransformed non normal data. So, the transformed duration variable would be used for result analysis and manipulation.

Transforming the duration variable and Checking Normality by code:

```
#Using the Log function for transforming duration
tduration<- log(sbank2$duration+10)

#adding the duration transformation to dataset table as new column by the name tduration
sbank2 <- sbank2 %>% mutate(tduration)

tduration_d<-describe(tduration,omit = TRUE, IQR = TRUE)
tduration_s<-list(pastecs::stat.desc(tduration, basic = FALSE))

skew         <- semTools::skew(tduration)
kurt         <- semTools::kurtosis(tduration)
stdskew      <- skew[1] / skew[2]
stdkurt      <- kurt[1] / kurt[2]
zscore       <- abs(scale(tduration))
gt196        <- FSA::perc(as.numeric(zscore), 1.96, "gt")
gt329        <- FSA::perc(as.numeric(zscore), 3.29, "gt")

tduration_s$skew <- skew
tduration_s$kurt <- kurt
tduration_s$std.skew <- stdskew
tduration_s$std.kurt <- stdkurt
tduration_s$gt.196 <- gt196
tduration_s$gt.329 <- gt329
tduration_s

# Distribution of transformed duration(tduration) variable with visualization After
Transformation
ggplot(sbank2,aes(x=tduration))+labs(x='Tranformed Duration', y='Density')+
  geom_histogram(binwidth = 0.5,colour='black',aes(y=..density..,fill=..count..))+
  scale_fill_gradient("Count",low="#132B43", high="#56B1F7")+
  stat_function(fun = dnorm,color="red",args = list(mean=mean(tduration,na.rm =
TRUE),sd=sd(tduration,na.rm = TRUE)))+
  ggtitle('Distribution of Tranformed Duration')+
  theme(plot.title = element_text(size=10))

# Create QQ Plot
qqnorm(tduration, main = "QQplot of Tranformed Duration")
qqline(tduration,col=2) # show line on the plot
```

```
     median         mean      SE.mean CI.mean.0.95          var      std.dev     coef.var
 5.22574667   5.23719034   0.03624910   0.07123396   0.60706689   0.77914497   0.14877156

$skew
skew (g1)        se         z         p
0.1408975 0.1139606 1.2363704 0.2163209

$kurt
Excess Kur (g2)              se                z                p
      0.2651529       0.2279212        1.1633535        0.2446861

$std.skew
skew (g1)
  1.23637

$std.kurt
Excess Kur (g2)
       1.163354

$gt.196
[1] 5.194805

$gt.329
[1] 0
```

Figure 8

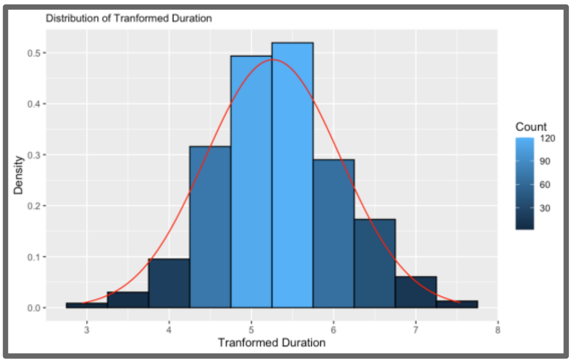


Figure 9

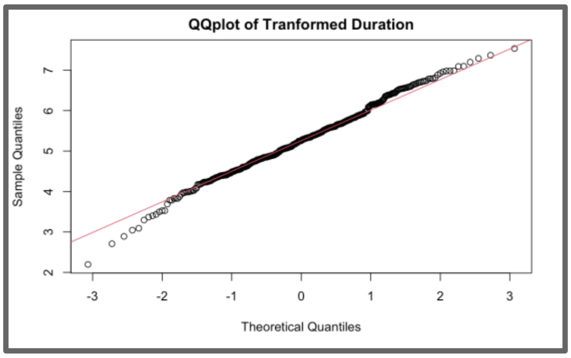


Figure 10

Report of normality analysis for transformed duration variable:

Transformed duration variable is represented by a numeric variable which was calculated by doing a Log function on the duration variable in the dataset. Inspecting the standardized scores for skewness (*skewness* = -.62, *SE* = .113) and kurtosis (*kurtosis* = 1.48, *SE* = .227) shows that both of the skewness and the kurtosis value fall within the standardized score range of -2 and 2 which implies that both skewness and kurtosis are normal. For further inspection by using plots such as histogram and normality plot (Figure 9 and Figure 10), it was found that the distribution is normal. On inspection of the count of outliers, we found the 0.2% standardised scores were outside the acceptable range of [-3.29, +3.29]. which shows that none of the values is outside of the 95% Confidence Interval. In total, it was found base all the test that was done the data for transformed duration variable has a normal distribution by this analysis (*M* =5.23, *SD* = .77, *N* = 462).

Has the client subscribed to a term deposit?

Has the client subscribed a term deposit(y) is a nominal variable in the bank marketing dataset. The sample dataset contains data from 53 clients who subscribed 'Yes' and 409 clients who did not subscribe 'No'. The variable is representative of a sample which the clients will subscribe to the term deposit.

Inspecting the variable by code:

```
y<-table(sbank2$y)
y

# report basic summary statistics by a grouping variable
describeBy(tduration,sbank2$y)

#remove NA from if there is exist
dy<-data.frame(sbank2$y,tduration)
dy<-na.omit(dy)
names(dy)<-c("y","duration")

# Create box plot for the variable
medhelp_graph<-ggplot(dy,aes(y,duration))
medhelp_graph+stat_summary(fun.y = mean,geom = "bar",fill="blue",colour="black",na.rm =
TRUE)+stat_summary(fun.data = mean_cl_normal,geom = "pointrange",na.rm =
TRUE)+labs(x="Client subscribtion of term deposit",y="Total Anxiety",title="Mean call
duration by Client subscribtion of term deposit")
```

```
 Descriptive statistics by group
group: no
   vars   n mean   sd median trimmed  mad  min max range  skew kurtosis   se
X1    1 409 5.14 0.72   5.18    5.14 0.73 2.83 7.2  4.37 -0.05      0.2 0.04
------------------------------------------------------------------------------------
group: yes
   vars  n mean   sd median trimmed  mad  min  max range skew kurtosis   se
X1    1 53 5.99 0.81      6    5.98 1.08 4.29 7.46  3.17 0.05    -1.02 0.11
```

Figure 14

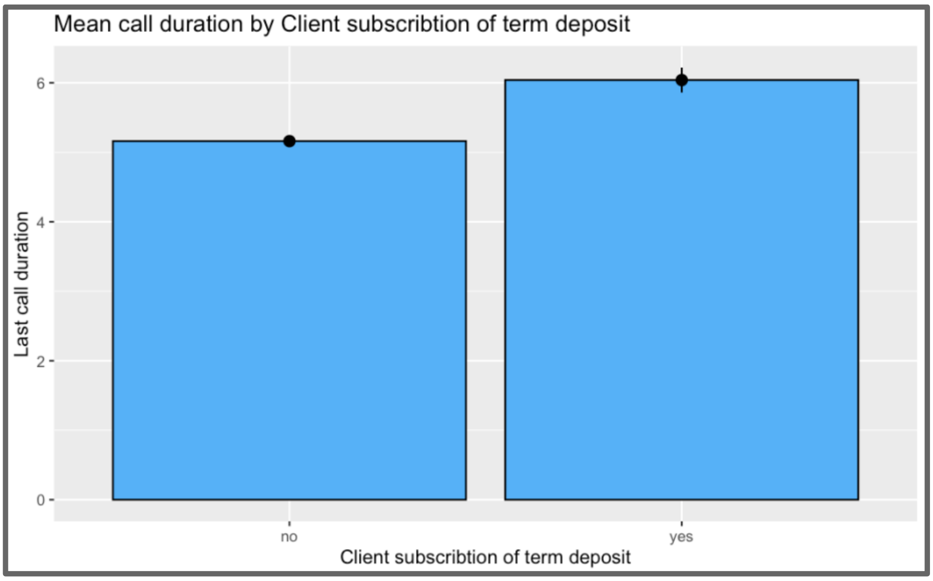


Figure 15

Age of client:

Age of client is a categorical variable in the bank marketing dataset. The sample dataset has data from 9 clients in the senior-citizen group(age>60), 265 clients in the Mid-age group(age >45), 117 clients in the mid-old group(age>30), 71 clients in the young group(age>15). The variable is representative of a sample which is a client from a Portuguese banking institution.

Inspecting the variable by code:

```
age_label<-table(sbank2$age_label)
age_label

# report basic summary statistics by a grouping variable
describeBy(tduration,sbank2$age_label)

#remove NA from if there is exist
dage<-data.frame(sbank2$age_label,tduration)
dage<-na.omit(dage)
names(dage)<-c("age_label","duration")

# Create box plot for the variable
medhelp_graph<-ggplot(dage,aes(age_label,duration))
medhelp_graph+stat_summary(fun.y = mean,geom = "bar",fill="#56B1F7",colour="black",na.rm =
TRUE)+stat_summary(fun.data = mean_cl_normal,geom = "pointrange",na.rm =
TRUE)+labs(x="Clients Age",y="Last call duration",title="Mean Last call duration by Clients
age")
```

```
 Descriptive statistics by group
group: Mid-age
   vars   n mean  sd median trimmed  mad  min  max range skew kurtosis   se
X1    1 247 5.22 0.8   5.18    5.19 0.75 2.83 7.46  4.62 0.32      0.3 0.05
----------------------------------------------------------------------------
group: mid-old
   vars   n mean   sd median trimmed  mad  min  max range  skew kurtosis   se
X1    1 112 5.34 0.73   5.34    5.34 0.59 3.53 7.25  3.72 -0.03     0.21 0.07
----------------------------------------------------------------------------
group: senior-citizen
   vars  n mean   sd median trimmed  mad  min  max range skew kurtosis   se
X1    1 11 5.58 0.74   5.26    5.58 0.64 4.43 6.74  2.31 0.16    -1.52 0.22
----------------------------------------------------------------------------
group: Young
   vars  n mean   sd median trimmed  mad  min  max range  skew kurtosis   se
X1    1 92 5.12 0.77   5.17    5.12 0.78 2.94 6.79  3.84 -0.14    -0.01 0.08
```

Figure 16

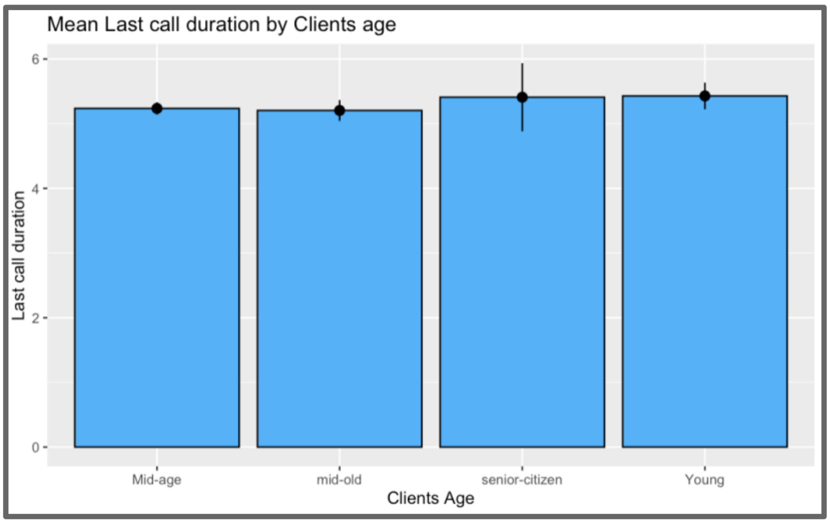


Figure 17

Clients Education:

Client Education is a categorical variable in the bank marketing dataset. The sample dataset has data from 42 clients in basic.4y group education, 30 clients in the basic.6y education group, 81 clients in the basic.9y education group, 110 clients in the high.school education group, 58 clients in the professional.course education group, 141 clients in the university.degree education group. The variable is representative of a sample which is a client from a Portuguese banking institution.

Inspecting the variable by code:

```
education<-table(sbank2$education)
education

# report basic summary statistics by a grouping variable
describeBy(tduration,sbank2$education)

#remove NA from if there is exist
dedu<-data.frame(sbank2$education,tduration)
dedu<-na.omit(dedu)
names(dedu)<-c("education","duration")

# Create box plot for the variable
medhelp_graph<-ggplot(dedu,aes(education,duration))
medhelp_graph+stat_summary(fun.y = mean,geom = "bar",fill="#56B1F7",colour="black",na.rm =
TRUE)+stat_summary(fun.data = mean_cl_normal,geom = "pointrange",na.rm =
TRUE)+labs(x="Clients Education",y="Last call duration",title="Mean Last call duration by
Clients Education")
```

```
 Descriptive statistics by group
group: basic.4y
   vars  n mean   sd median trimmed mad  min  max range skew kurtosis   se
X1    1 44  5.4 0.78   5.29    5.36 0.6 3.93 7.25  3.32 0.37    -0.15 0.12
------------------------------------------------------------------------------------
group: basic.6y
   vars  n mean   sd median trimmed  mad  min  max range  skew kurtosis   se
X1    1 24 5.09 0.78    5.2    5.12 0.94 3.37 6.25  2.88 -0.27    -1.04 0.16
------------------------------------------------------------------------------------
group: basic.9y
   vars  n mean   sd median trimmed  mad  min  max range  skew kurtosis   se
X1    1 57 5.04 0.63   5.12    5.09 0.68 2.83 6.19  3.36 -0.86     1.14 0.08
------------------------------------------------------------------------------------
group: high.school
   vars  n mean   sd median trimmed  mad  min  max range skew kurtosis   se
X1    1 91 5.35 0.88   5.33    5.34 0.87 3.18 7.46  4.28 0.12    -0.33 0.09
------------------------------------------------------------------------------------
group: illiterate
   vars n mean sd median trimmed mad  min  max range skew kurtosis se
X1    1 1 5.23 NA   5.23    5.23   0 5.23 5.23     0   NA       NA NA
------------------------------------------------------------------------------------
group: professional.course
   vars  n mean   sd median trimmed  mad  min  max range skew kurtosis   se
X1    1 51 5.33 0.81   5.18     5.3 0.82 3.85 6.94  3.09 0.33    -0.78 0.11
------------------------------------------------------------------------------------
group: university.degree
   vars   n mean   sd median trimmed  mad  min  max range skew kurtosis   se
X1    1 194  5.2 0.75    5.2    5.19 0.71 2.94 7.43  4.48  0.1     0.46 0.05
```

Figure 18

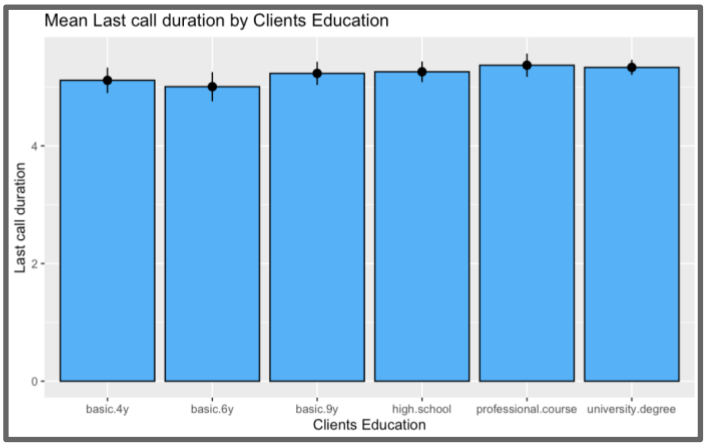


Figure 19

Section 3 - Result

An alpha level of .05 was adopted for the Pearson Correlation Test and Cohen's rules on effect size (coefficient which is r) were adopted. For correlation, according to Cohen, the effect size is low if the r value varies around .1, medium if r varies around .3, and high if r varies more than .5 (Field, Miles & Field, 2012).A p-value lower than the level of significance of .05 indicates that the null hypothesis is clear evidence to reject it. The effect size is low if the value of eta-squared varies around .01, medium if eta-squared varies around .06, and high if eta-squared varies more than .14, as per Cohen, for difference (Field, Miles & Field, 2012).

Hypothesis 1:

- H0: There will be no significant predictor for last call duration that leads clients to subscribed term deposits in different age groups?
- H1: There will be a significant predictor for last call duration that leads clients to subscribed term deposits in different age groups?

Statistical Evidence

Check the difference in the last call duration for clients by different age groups.

```
# Descriptive statistics by clients age group
durage<-na.omit(data.frame(sbank2$tduration,sbank2$age_label))
names(durage)<-c('tduration','age_label')

# Descriptive statistics by client age groups
describeBy(durage$tduration,durage$age_label)
mean_durage<-round(tapply(durage$tduration,durage$age_label,FUN =mean),digits = 2)
mean_durage
```

```
 Descriptive statistics by group
group: Mid-age
   vars   n mean  sd median trimmed  mad  min  max range skew kurtosis   se
X1    1 247 5.22 0.8   5.18    5.19 0.75 2.83 7.46  4.62 0.32      0.3 0.05
------------------------------------------------------------------------------
group: mid-old
   vars   n mean   sd median trimmed  mad  min  max range  skew kurtosis   se
X1    1 112 5.34 0.73   5.34    5.34 0.59 3.53 7.25  3.72 -0.03     0.21 0.07
------------------------------------------------------------------------------
group: senior-citizen
   vars  n mean   sd median trimmed  mad  min  max range skew kurtosis   se
X1    1 11 5.58 0.74   5.26    5.58 0.64 4.43 6.74  2.31 0.16    -1.52 0.22
------------------------------------------------------------------------------
group: Young
   vars  n mean   sd median trimmed  mad  min  max range  skew kurtosis   se
X1    1 92 5.12 0.77   5.17    5.12 0.78 2.94 6.79  3.84 -0.14    -0.01 0.08
```

Figure 20

```
     Mid-age        mid-old senior-citizen          Young
        5.22           5.34           5.58           5.12
```

Figure 21

The plot (Figure 22) shows how mean values of last call duration changes with different client age groups and the number of people belonging to each group as well.

```
plotmeans(durage$tduration~durage$age_label,digits = 2,
          ccol = 'red',mean.labels = TRUE,xlab = 'Clients age group',ylab = 'Last call
duration',
          main='Plot of Last call duration Mean by Clients age group')
```

From the graph(Figure 22), we can understand that the mean value of last call duration differs for different groups.
Last call duration with 'young' group having the lowest mean and Group 'senior-citizen' having the highest mean.

the boxplot analysis for further hypothesis testing was performed.

```
boxplot(durage$tduration~durage$age_label,
        main='Plot of Last call duration Score by Clients age group (dot is mean)',
        xlab = 'Clients age group',ylab = 'Last call duration',col=rainbow(6))
points(mean_durage,col='black',pch=18)
```

As it was obvious in boxplots(Figure 23), it was inferred that each client in different age groups has a different amount of variation in last call duration and there is a lot of overlap among values for different groups. But this information is not enough to provide evidence to simply affirm or reject null hypothesis as it does not give information whether the differences are statistically significant. To determine statistical significance, we need to assess the confidence intervals for the differences of means. We further investigate the difference in mean values, considering there is a lot of overlap of last call duration for different clients groups, just because of variation within groups or variation among the groups. The ANOVA Test would be done.

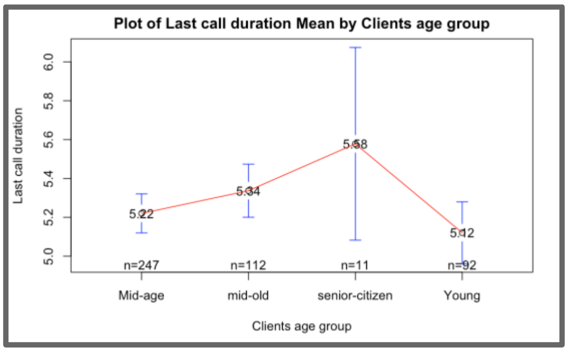


Figure 22

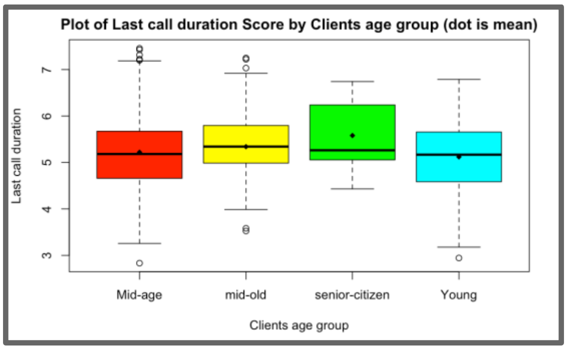


Figure 23

Doing the Bartlett test of homogeneity of variances for last call duration and clients in different age groups.

```
bartlett.test(durage$tduration,durage$age_label)
```

As it was aboviused in Figure 23 the p-value = .71 > .05 so the null hypothesis of the test is accepted and should be said that the variance of different groups can be assumed to be equal.

```
        Bartlett test of homogeneity of variances

data:  durage$tduration and durage$age_label
Bartlett's K-squared = 1.3557, df = 3, p-value = 0.7159
```

Figure 24

Doing the assumption of homoscedasticity for Anova test for last call duration and clients in different age groups.

```
aov_durage<-aov(durage$tduration~durage$age_label)
summary(aov_durage)
```

```
                  Df Sum Sq Mean Sq F value Pr(>F)
durage$age_label   3   3.72  1.2393   2.055  0.105
Residuals        458 276.14  0.6029
```

Figure 25

Since the value of F-statistic=2.05 > 1 (significant) and p-value=0.105 > 0.05 in Figure(24), this shows that the variation among the groups and the variation within groups is high, so the mean values for different groups are not significantly different. Therefore, we could not reject the null hypothesis of the test which means the values for different groups are equal. In total, it was concluded that for the confidence interval the null hypothesis can not be rejected that there is no significant difference in last call duration for clients in different age groups.

Doing the calculation of eta square, the result came in Figure 25.

```
etaSquared(aov_durage)
```

```
                      eta.sq eta.sq.part
durage$age_label 0.01328459  0.01328459
```

Figure 25

Report of Difference Analysis:

A Bartlett's test was done and the equality of variance for Last call duration for all clients in different age group was indicated *K-squared* = 1.35, *P* = 0.7.
A one-way between-groups analysis of variance was conducted to last call duration for clients in different age groups. clients were divided into four groups according to their age (Group 1 : senior-citizen, Group 2 : Mid-age, Group 3 : mid-old, Group 4 : young).
There was no statistically significant difference level in last call duration mean for different clients age (*F(3, 458)* = 2.05, *p* = .105).
The effect size, calculated using eta squared was .01.
The test results indicate there is evidence to support accepting the null hypothesis that there is no difference in last call duration for clients in different age groups.

Checking the difference in last duration call leads clients to subscribed a term deposit.

```
# Descriptive statistics
contact<-as.factor(sbank2$y)

#Conduct Levene's test for homogeneity of variance in library car
ltest<-car::leveneTest(tduration ~ y, data=sbank2)
#Pr(F) is the probability
ltest

#Conduct the t-test from package stats
#You can use the var.equal = TRUE option to specify equal variances and a pooled variance
estimate
stats::t.test(tduration~y,var.equal=TRUE,data=sbank2)
#Effect Size

effsize::cohen.d(tduration,y, alpha = 0.05, na.rm=TRUE)
# effet size=-19.23 large difference
```

Report of Difference Analysis

A Levene's test was conducted and indicated equality of variance for Last call duration for clients who subscribed the term deposit (*F-value* = 3.02, *P* = .08). A t-test analysis of variance was conducted to explore last call duration for clients who subscribed term deposit. Participants were divided into groups according to which clients will subscribe to the term deposit(Group 1 : Yes, Group 2 : No). There was a statistically significant difference in last call duration mean scores for clients who subscribed to a term deposit.The p-value for two sample tests is very small which means it is significant and can reject the null hypothesis.

The effect size, calculated using Cohen's d was -19.23 which implies there is a strong standardised mean difference for both groups. The test results indicate there is evidence to reject null alternative hypothesis which is no difference in Last call duration for clients who subscribed term deposit. The result came in Figure 27.

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value  Pr(>F)
group   1   3.027 0.08256 .
      460
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

        Two Sample t-test

data:  tduration by y
t = -7.9077, df = 460, p-value = 1.963e-14
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.0548170 -0.6349074
sample estimates:
 mean in group no mean in group yes
         5.140269          5.985131


Cohen's d

d estimate: -19.23453 (large)
95 percent confidence interval:
    lower     upper
-21.09966 -17.36939
```

Figure 27

Model 1

Building the linear regression models
Baseline Model last call duration predicted that clients in different age groups will subscribe to term deposit.

```
#dummy code
df<-data.frame(sbank2)
which(names(df)=='age_label')
which(names(df)=='y')
which(names(df)=='education')
df<-df[,c(1,21,5)]
df$tduration<-tduration
df<-na.omit(df)
df
df$age_label=recode(df$age_label,"senior-citizen"="1","Mid-age"='2',"mid-
old"='3',"Young"='4')
df$education=recode(df$education,'basic.4y'='1','basic.6y'='2','basic.9y'='3','high.school'
='4','professional.course'='5','university.degree'='6')
df$y=recode(df$y,'yes'='1','no'='2')
df
```

Creating model1

```
model1=lm(df$tduration~df$y+df$age_label)
anova(model1)
summary(model1)
```

```
Analysis of Variance Table

Response: df$tduration
              Df  Sum Sq Mean Sq F value  Pr(>F)
df$y           1  33.491  33.491 63.1625 1.5e-14 ***
df$age_label   3   4.049   1.350  2.5451 0.05552 .
Residuals    457 242.318   0.530
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 28

```
Call:
lm(formula = df$tduration ~ df$y + df$age_label)

Residuals:
     Min       1Q   Median       3Q      Max
-2.30485 -0.51309  0.00773  0.48589  2.06161

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.89867    0.22319  26.429  < 2e-16 ***
df$y2         -0.88126    0.11034  -7.987 1.14e-14 ***
df$age_label2  0.12066    0.23226   0.520    0.604
df$age_label3  0.24884    0.23812   1.045    0.297
df$age_label4 -0.03087    0.23837  -0.129    0.897
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7282 on 457 degrees of freedom
Multiple R-squared:  0.1341,    Adjusted R-squared:  0.1266
F-statistic:  17.7 on 4 and 457 DF,  p-value: 1.612e-13
```

Figure 29

```
stargazer(model1, type="text") #Tidy output of all the required stats
# Figure 29
plot(model1)
```

```
===============================================
                        Dependent variable:
                    ---------------------------
                             tduration
-----------------------------------------------
y2                           -0.881***
                              (0.110)

age_label2                     0.121
                              (0.232)

age_label3                     0.249
                              (0.238)

age_label4                    -0.031
                              (0.238)

Constant                      5.899***
                              (0.223)

-----------------------------------------------
Observations                    462
R2                             0.134
Adjusted R2                    0.127
Residual Std. Error      0.728 (df = 457)
F Statistic           17.699*** (df = 4; 457)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

Figure 30

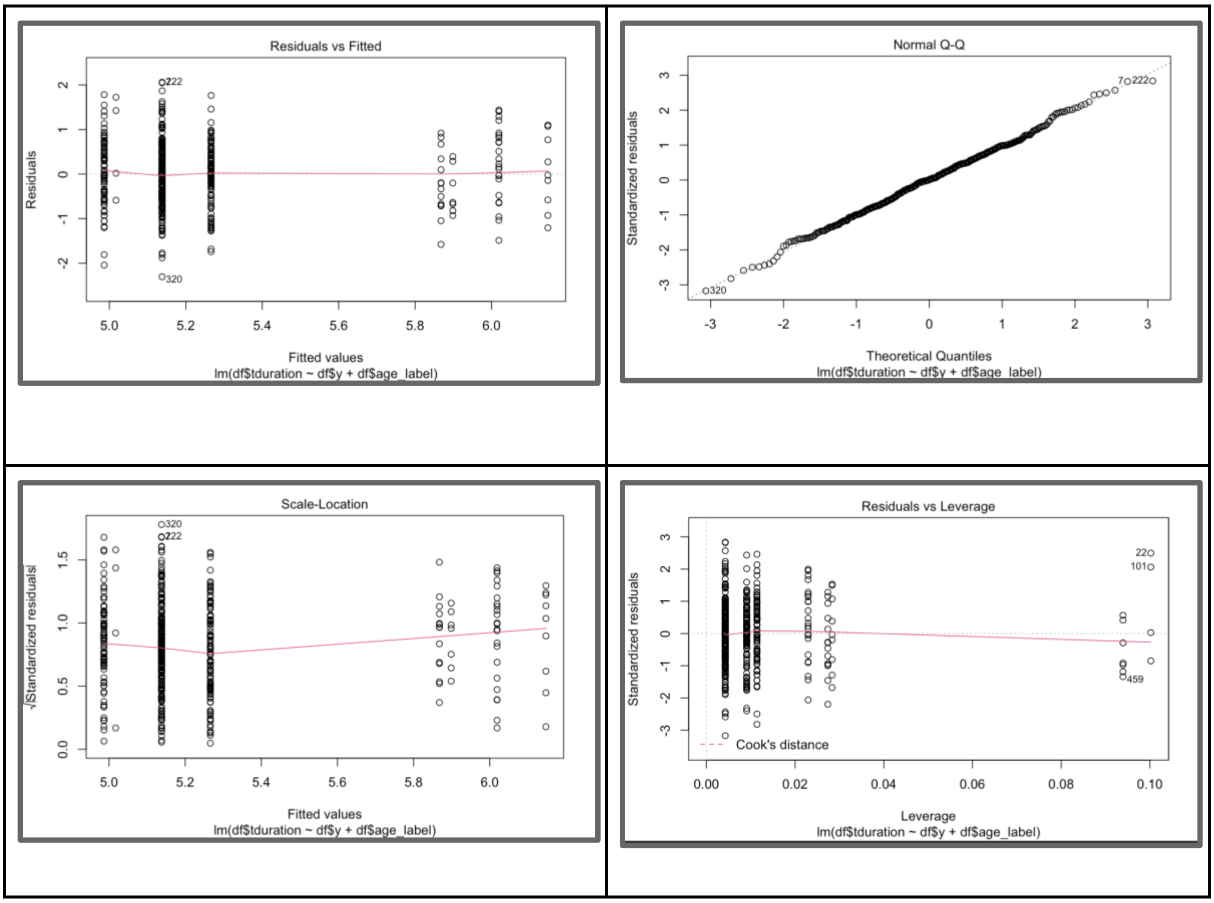


Figure31

```
#Check assumptions
# List of   residuals
resid(model1)
#A density plot of the residuals
plot(density(resid(model1)))
```

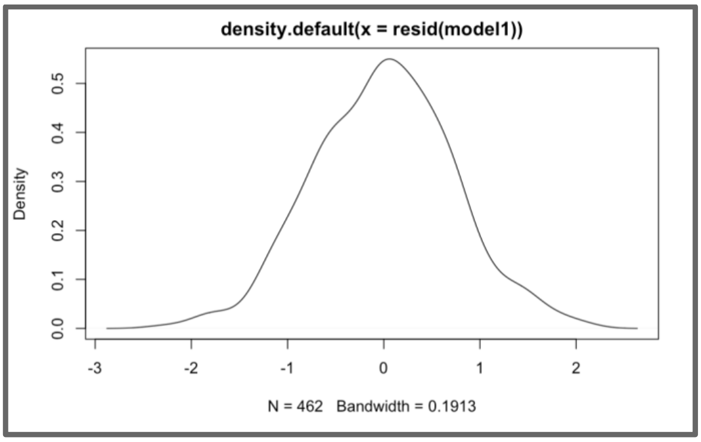


Figure 32

```
# leverage plots
leveragePlots(model1)
```

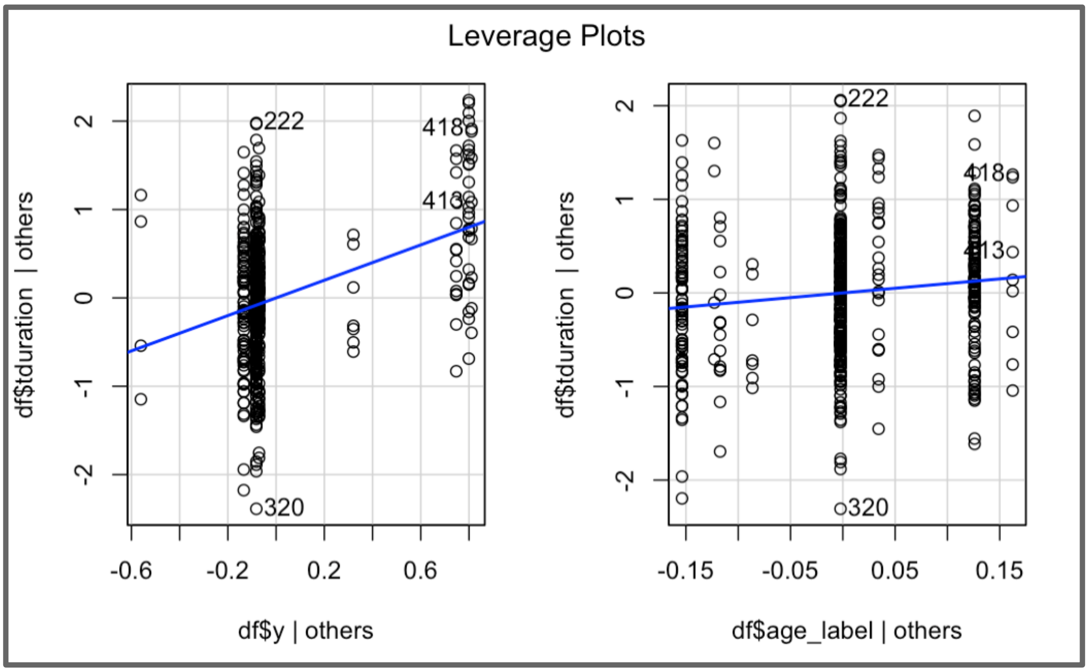


Figure 33

```
#Cook's distance
cooks.distance(model1)
#Plot Cook's distance
plot(cooks.distance(model1), ylab="Cook's statistic")
# none of the values is greater than 1 so no influential values
```

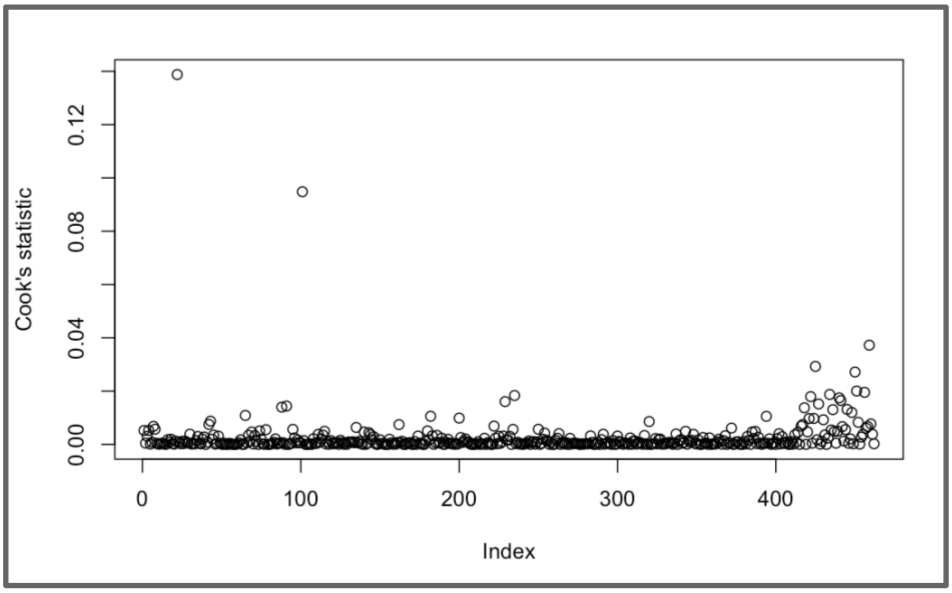


Figure 34

```
# Collinearity
vifmodel<-vif(model1)
vifmodel
# value < 2.5 not problem
1/(vifmodel)
# values > .4 not problem
```

```
                      GVIF Df GVIF^(1/(2*Df))
sbank2$y          1.104606  1        1.051002
sbank2$age_label  1.146879  3        1.023104
sbank2$education  1.091177  5        1.008764
```

Figure 35

Report of Linear Modelling Analysis

A multiple linear regression analysis was conducted to determine whether last call duration and clients age lead the customer to subscribed the term deposit. A significant regression equation was found (*F(4,457)*= 17.7, *p*=1.612e-13), with an Multiple R-squared=.1341.

Examination of the histogram, normal P-P plot of standardised residuals and the scatterplot of the dependent variable, last call duration, and standardised residuals showed that no outliers existed and the residuals followed normal distribution. Also, examination of the standardised residuals showed that none of the values was outside the standard range (95% within limits of -3.29 to +3.29) as the minimum and maximum values are -2.3 and 2.06 respectively further affirming that there were no outliers. Also, none of the Cook's distances were found to be more than 1, hence there are no influential values.

Examination for multicollinearity showed that the tolerance and variance influence factor measures were within acceptable levels (tolerance >0.4, VIF <2.5) as outlined in Tarling (2008). The scatter plot of standardised residuals showed that the data met the assumptions of homogeneity of variance and linearity. The data also meets the assumption of non-zero variances of the predictors.

Because all the assumptions for the model 1 have been proven true and 13.41% of the variance in last call duration is explained by the considered predictors. On checking the significance levels for each of the main terms (in this case the coefficients associated with the constant, y2), we found that there is evidence that each of these terms are adding something to the model (they are statistically significant as p<.05). Hence, these statistical values provide enough evidence to reject that null hypothesis which is no significant prediction of last call duration that lead clients to subscribed term deposits in different age groups.

Hypothesis 2:

- H0: There will be no significant predictor for last call duration that leads clients to subscribed term deposits by different age and different education?
- H1: There will be a significant predictor for last call duration that leads clients to subscribed term deposits by different age and different education?

Statistical Evidence

Check difference in the last call duration for clients with different education

```
# Descriptive statistics
duredu<-na.omit(data.frame(tduration,sbank2$education))
names(duredu)<-c('tduration','education')
# Descriptive statistics by education
describeBy(duredu$tduration,duredu$education)
# check mean for each education group
mean_duredu<-round(tapply(duredu$tduration,duredu$education,FUN =mean),digits = 2)
mean_duredu
```

```
 Descriptive statistics by group
group: basic.4y
   vars  n mean   sd median trimmed  mad  min  max range skew kurtosis   se
X1    1 45  5.4 0.77   5.28    5.36 0.57 3.93 7.25  3.32 0.38    -0.09 0.12
----------------------------------------------------------------------------
group: basic.6y
   vars  n mean   sd median trimmed  mad  min  max range  skew kurtosis   se
X1    1 24 5.09 0.78    5.2    5.12 0.94 3.37 6.25  2.88 -0.27    -1.04 0.16
----------------------------------------------------------------------------
group: basic.9y
   vars  n mean   sd median trimmed  mad  min  max range  skew kurtosis   se
X1    1 57 5.04 0.63   5.12    5.09 0.68 2.83 6.19  3.36 -0.86     1.14 0.08
----------------------------------------------------------------------------
group: high.school
   vars  n mean   sd median trimmed  mad  min  max range skew kurtosis   se
X1    1 91 5.35 0.88   5.33    5.34 0.87 3.18 7.46  4.28 0.12    -0.33 0.09
----------------------------------------------------------------------------
group: professional.course
   vars  n mean   sd median trimmed  mad  min  max range skew kurtosis   se
X1    1 51 5.33 0.81   5.18     5.3 0.82 3.85 6.94  3.09 0.33    -0.78 0.11
----------------------------------------------------------------------------
group: university.degree
   vars   n mean   sd median trimmed  mad  min  max range skew kurtosis   se
X1    1 194  5.2 0.75    5.2    5.19 0.71 2.94 7.43  4.48  0.1     0.46 0.05
```

Figure 36

```
basic.4y           basic.6y           basic.9y        high.school professional.course   university.degree
    5.40               5.09               5.04               5.35                5.33                5.20
```

Figure 37

The plot (Figure 38) shows how mean values of last call duration changes with different client age groups and the number of people belonging to each group as well.

```
plotmeans(duredu$tduration~duredu$education,digits = 2,
          ccol = 'red',mean.labels = TRUE,xlab = 'Clients education',ylab = 'Last call
duration',main='Plot of Last call duration Mean by Clients educationh')
```

From the graph (Figure 38), it was founded that the mean value of last call duration differs for different education groups.
Last call duration with 'basic.9y' group having the lowest mean and group 'basic.4y' having the highest mean.

the boxplot analysis for further hypothesis testing was performed.

```
boxplot(duredu$tduration~duredu$education,
        main='Plot of last call duration by Clients educationh (dot is mean)',xlab = '
Clients educationh',ylab = 'Last call duration',col=rainbow(11))
points(mean_duredu,col='black',pch=18)
```

As it was obvious in boxplots(Figure 39), it was inferred that each client in different age groups has a different amount of variation in last call duration and there is a lot of overlap among values for different groups. But this information is not enough to provide evidence to simply affirm or reject null hypothesis as it does not give information whether the differences are statistically significant. To determine statistical significance, we need to assess the confidence intervals for the differences of means. We further investigate the difference in mean values, considering there is a lot of overlap of last call duration for different clients groups, just because of variation within groups or variation among the groups. The ANOVA Test would be done.

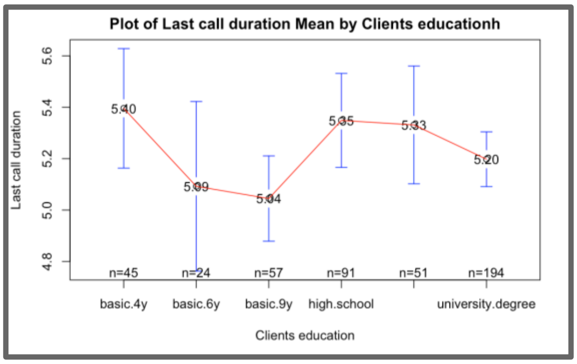


Figure 38

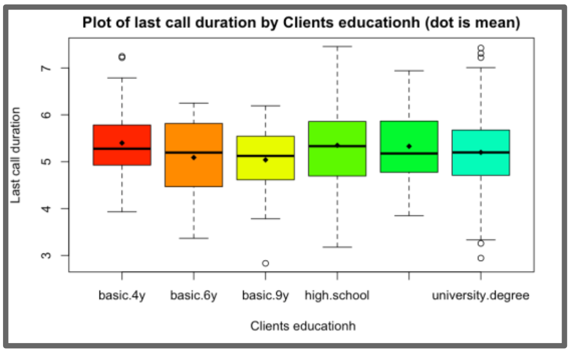


Figure 39

Doing the Bartlett test of homogeneity of variances for last call duration and clients with different education.

```
bartlett.test(duredu$tduration,duredu_new$education)
```

As it was aboviused in (Figure 40) the p-value = .14 > .05 so the null hypothesis of the test is accepted and should be said that the variance of different groups can be assumed to be equal.

```
        Bartlett test of homogeneity of variances

data:  duredu$tduration and duredu_new$education
Bartlett's K-squared = 8.1246, df = 5, p-value = 0.1495
```

Figure 40

Doing the assumption of homoscedasticity for Anova test for last call duration and clients in different age groups.

```
aov_durage<-aov(durage$tduration~durage$age_label)
summary(aov_durage)
```

```
                   Df Sum Sq Mean Sq F value Pr(>F)
duredu$education    5   5.63  1.1254   1.871 0.0979 .
Residuals         456 274.23  0.6014
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 41

Since the value of F-statistic=1.87 > 1 (significant) and p-value=0.097 > 0.05 in (Figure41), this shows that the variation among the groups and the variation within groups is high, so the mean values for different groups are not significantly different. Therefore, we could not reject the null hypothesis of the test which means the values for different groups are equal. In total, it was concluded that for the confidence interval the null hypothesis can not be rejected that there is no significant difference in last call duration for clients with different education.

Doing the calculation of eta square, the result came in Figure 41.

```
etaSquared(aov_durage)
```

```
                      eta.sq eta.sq.part
duredu$education 0.02010697  0.02010697
```

Figure 42

Report of Difference Analysis:

A Bartlett's test was done and the equality of variance for Last call duration for all clients in different age group was indicated *K-squared* = 8.12, *P* = .14.
A one-way between-groups analysis of variance was conducted to last call duration for clients with different education. clients were divided into six groups according to their education (Group 1 : basic.4y, Group 2 : basic.6y, Group 3 : basic.9y, Group 4 : high school,Group 5 : professional.course,Group 6 :university.degree).
There was no statistically significant difference level in last call duration mean for different clients education (*F(5, 456)* = 1.87, *p* = .09).
The effect size, calculated using eta squared was .02.
The test results indicate there is evidence to support accepting the null hypothesis that there is no difference in last call duration for clients with different education.

Model 2

Building the linear regression models
Model last call duration predicted that clients by different age groups and different education will subscribe to term deposit.

Creating model 2

```
model2<-lm(df1$tduration~df1$y+df1$age_label+df1$education)
anova(model2)
summary(model2)
```

```
Analysis of Variance Table

Response: df1$tduration
               Df  Sum Sq Mean Sq F value    Pr(>F)
df1$y           1  33.491  33.491 63.1038 1.572e-14 ***
df1$age_label   3   4.049   1.350  2.5428   0.05571 .
df1$education   5   2.428   0.486  0.9151   0.47088
Residuals     452 239.890   0.531
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 43

```
Call:
lm(formula = df1$tduration ~ df1$y + df1$age_label + df1$education)

Residuals:
     Min       1Q   Median       3Q      Max
-2.20230 -0.50565  0.00438  0.50858  1.95609

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.94960    0.23864  24.931  < 2e-16 ***
df1$y2          -0.85290    0.11178  -7.630  1.4e-13 ***
df1$age_label2   0.10867    0.23414   0.464    0.643
df1$age_label3   0.23069    0.23911   0.965    0.335
df1$age_label4  -0.04488    0.24044  -0.187    0.852
df1$education2  -0.13573    0.18641  -0.728    0.467
df1$education3  -0.16985    0.14811  -1.147    0.252
df1$education4   0.03823    0.13571   0.282    0.778
df1$education5   0.01021    0.15140   0.067    0.946
df1$education6  -0.10260    0.12342  -0.831    0.406
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7285 on 452 degrees of freedom
Multiple R-squared:  0.1428,    Adjusted R-squared:  0.1257
F-statistic: 8.368 on 9 and 452 DF,  p-value: 1.4e-11
```

Figure 44

```
stargazer(model2, type="text") #Tidy output of all the required stats
stargazer(model1, model2, type="text") #Quick model comparison
plot(model2)
```

```
===============================================
                        Dependent variable:
                    ---------------------------
                             tduration
-----------------------------------------------
y2                           -0.853***
                              (0.112)

age_label2                    0.109
                              (0.234)

age_label3                    0.231
                              (0.239)

age_label4                   -0.045
                              (0.240)

education2                   -0.136
                              (0.186)

education3                   -0.170
                              (0.148)

education4                    0.038
                              (0.136)

education5                    0.010
                              (0.151)

education6                   -0.103
                              (0.123)

Constant                     5.950***
                              (0.239)

-----------------------------------------------
Observations                    462
R2                             0.143
Adjusted R2                    0.126
Residual Std. Error      0.729 (df = 452)
F Statistic           8.368*** (df = 9; 452)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

```
===================================================================
                              Dependent variable:
                    -----------------------------------------------
                                   tduration
                             (1)                     (2)
-------------------------------------------------------------------
y2                        -0.881***               -0.853***
                           (0.110)                 (0.112)

age_label2                  0.121                   0.109
                           (0.232)                 (0.234)

age_label3                  0.249                   0.231
                           (0.238)                 (0.239)

age_label4                 -0.031                  -0.045
                           (0.238)                 (0.240)

education2                                         -0.136
                                                   (0.186)

education3                                         -0.170
                                                   (0.148)

education4                                          0.038
                                                   (0.136)

education5                                          0.010
                                                   (0.151)

education6                                         -0.103
                                                   (0.123)

Constant                   5.899***                5.950***
                           (0.223)                 (0.239)

-------------------------------------------------------------------
Observations                 462                     462
R2                          0.134                   0.143
Adjusted R2                 0.127                   0.126
Residual Std. Error   0.728 (df = 457)        0.729 (df = 452)
F Statistic         17.699*** (df = 4; 457) 8.368*** (df = 9; 452)
===================================================================
Note:                                   *p<0.1; **p<0.05; ***p<0.01
```

Figure 45

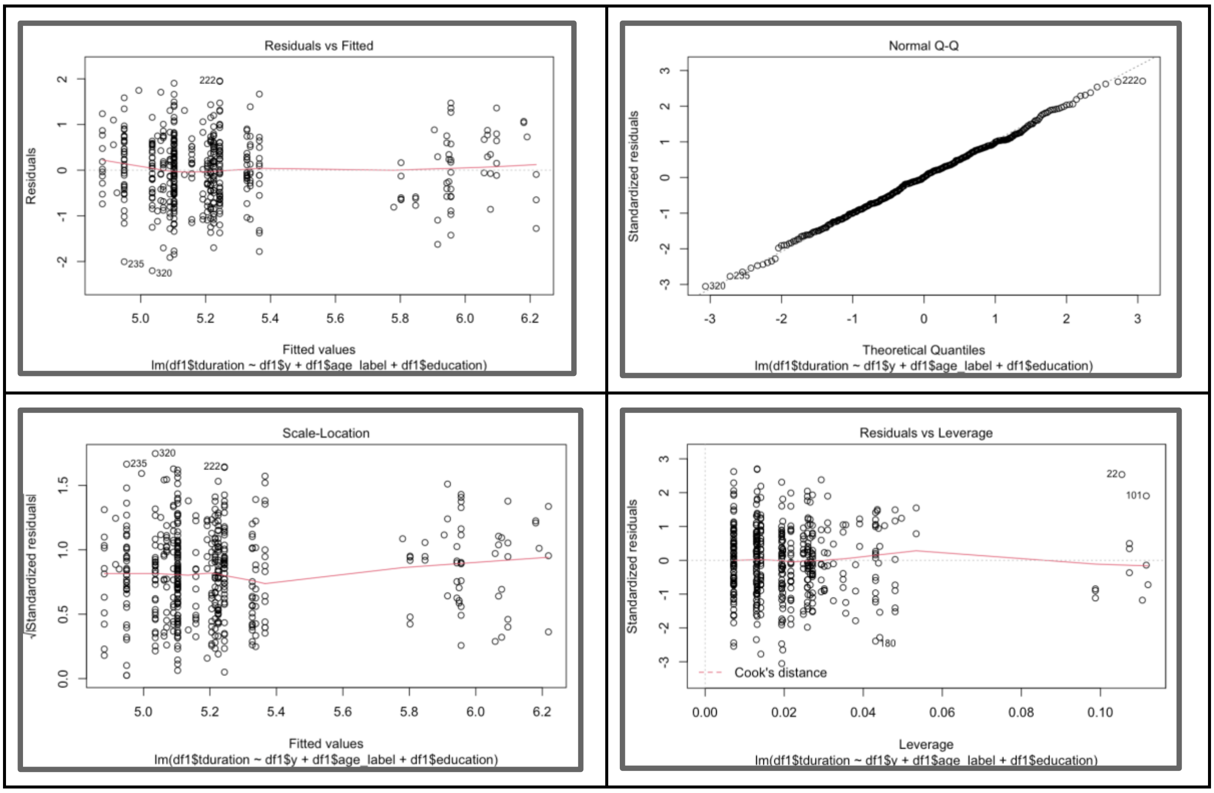


Figure 46

```
#Check assumptions
# List of residuals
resid(model1)
#A density plot of the residuals
plot(density(resid(model1)))
```

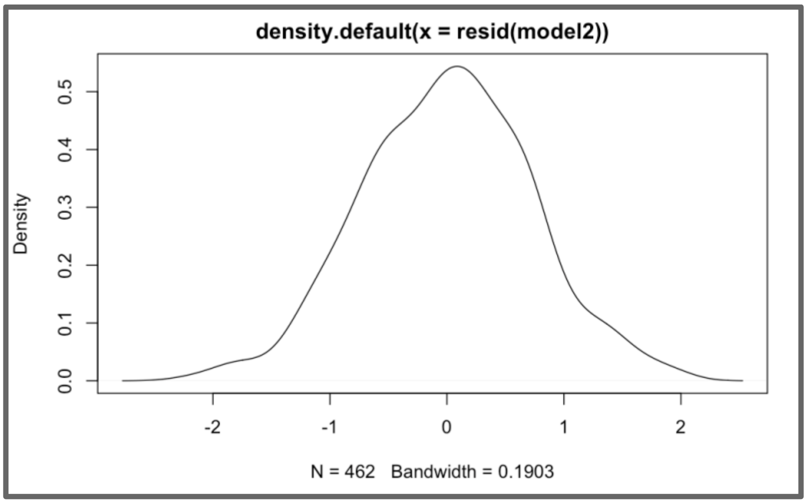


Figure 47

```
# leverage plots
leveragePlots(model1)
```

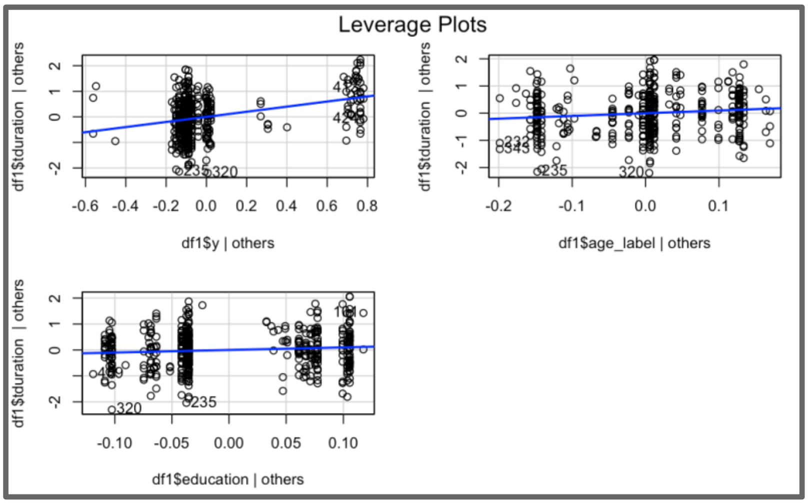


Figure 47

```
#Cook's distance
cooks.distance(model1)
#Plot Cook's distance
plot(cooks.distance(model1), ylab="Cook's statistic")
# none of the values is greater than 1 so no influential values
```

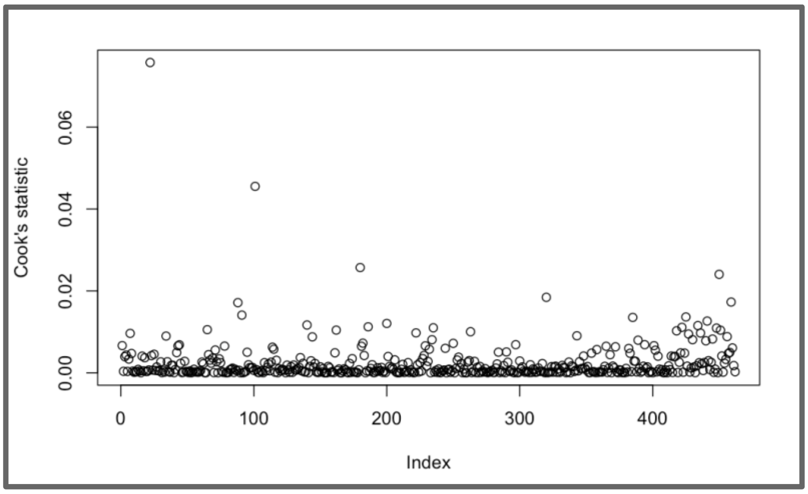


Figure 48

```
# Collinearity
vifmodel<-vif(model2)
vifmodel
# value < 2.5 not problem
1/(vifmodel)
# values > .4 not problem
```

```
                   GVIF Df GVIF^(1/(2*Df))
df1$y          1.104606  1        1.051002
df1$age_label  1.146879  3        1.023104
df1$education  1.091177  5        1.008764
                    GVIF        Df GVIF^(1/(2*Df))
df1$y          0.9053000 1.0000000       0.9514726
df1$age_label  0.8719319 0.3333333       0.9774182
df1$education  0.9164419 0.2000000       0.9913123
```

Figure 49

Report of Linear Modelling Analysis

A multiple linear regression analysis was conducted to determine whether last call duration and clients age lead the customer to subscribed the term deposit. A significant regression equation was found (*F(9,452)*= 8.368, *p*= 1.4e-11), with an Multiple R-squared=.1428.

Examination of the histogram, normal P-P plot of standardised residuals and the scatterplot of the dependent variable, last call duration, and standardised residuals showed that no outliers existed and the residuals followed normal distribution. Also, examination of the standardised residuals showed that none of the values was outside the standard range (95% within limits of -3.29 to +3.29) as the minimum and maximum values are -2.2 and 1.95 respectively further affirming that there were no outliers. Also, none of the Cook's distances were found to be more than 1, hence there are no influential values.

Examination for multicollinearity showed that the tolerance and variance influence factor measures were within acceptable levels (tolerance >0.4, VIF <2.5) as outlined in Tarling (2008). The scatter plot of standardised residuals showed that the data met the assumptions of homogeneity of variance and linearity. The data also meets the assumption of non-zero variances of the predictors.

Because all the assumptions for the model 1 have been proven true and 14.28% of the variance in last call duration is explained by the considered predictors. On checking the significance levels for each of the main terms (in this case the coefficients associated with the constant, y2), we found that there is evidence that just y2 and constant are adding something to the model (this statistically significant as p<.05). Hence, these statistical values provide enough evidence to reject that null hypothesis which is no significant prediction of last call duration that lead clients to subscribed term deposits by different age and different education.

Summary Comparison

Compare Model 1 and Model 2

```
anova(model1,model2)
```

```
Analysis of Variance Table

Model 1: df1$tduration ~ df1$y + df1$age_label
Model 2: df1$tduration ~ df1$y + df1$age_label + df1$education
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    457 242.32
2    452 239.89  5    2.4284 0.9151 0.4709
```

Figure 50

Model Comparison Results

It was obvious that the p-value obtained for second model (*p*=.47) is not statistically significant (more than .05), so addition of a new variable significantly not improved the fit over model 1. therefore, we should not reject model 1.

Section 4 – Discussion/Conclusion

In this study, we wanted to determine which factors could be considered as the best predictors for determining the last call duration that can lead customers to subscribed term deposits . This was conducted by using Multiple Linear Regression where we firstly tried to establish evidence that the various predictors chosen can be used for modelling. We investigated whether a client's age, their education,and customer subscribed term deposit. We found that there were statistically significant differences in last call duration with customers with different client age, education. Furthermore, different models were built to determine which variables are the best for prediction.

The results of the baseline first model are analysed that determine whether last duration call can lead customers with different age to subscribed to the term deposit can be used as predictor for output variable. Since the p-value for the model obtained is statistically significant (1.612e-13 <.05), it was shown the model is good to fit as it performs better than the average score method for prediction. As per the analysis of R-squared value which is found to be .1341, we can say that 13.41% of the variance in the Last Call duration is explained by the considered predictors. On checking the significance levels for each of the main terms (in this case the coefficients associated with the constant,y2), we found that there is evidence that each of these terms are adding something to the model (they are statistically significant as p<.05). There were found no outliers, residuals, leverage points and influential values for the model. This model explained the% of variance in the Last call duration.

The results of the second model are analysed which determine whether last call duration can lead clients with different age and education to subscribed term deposit can be used as predictor for output variable. Since the p-value for the model obtained is statistically significant (1.4e-11<.05), it gives us evidence to suggest that the model is good to fit as it performs better than the average score method for prediction. As per the analysis of R-squared value which is found to be .1428, we can say that 14.28% of the variance in Last call duration is explained by the considered predictors. On checking the significance levels for each of the main terms (in this case the coefficients associated with the constant, y2), we found that there is evidence that each of these terms are adding something to the model (they are statistically significant as p<.05). There were no outliers, residuals, leverage points and influential values found for the model. Overall, the second model can not be considered better than the baseline model as it is statistically significant based on comparison results and addresses a few higher amounts of variance in the output variable, Last call duration which implies it is not better at making predictions.

From this analysis, there is evidence to conclude that Model 2 is the better compared to the model because of 14.28%% of the variance in Last call duration but this is not very much different between model 1 to model 2, so if we want to decide we should say the model 1 is better because this differences is not a lot.

Reference

1. *George, D. (2011).*

2. *Ziegel, E., & Lohr, S. (2000). Sampling: Design and Analysis. Technometrics, 42(2), 223. doi: 10.2307/1271491*
3. *Field, A., Miles, J., & Field, Z. (2012).*
4. *Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 62, 22-31. doi: 10.1016/j.dss.2014.03.001*
5. *UCI Machine Learning Repository: Bank Marketing Data Set. (2021). Retrieved 6 January 2021, from https://archive.ics.uci.edu/ml/datasets/Bank+Marketing*
6. *CrossTable function | R Documentation. (2021). Retrieved 4 January 2021, from https://www.rdocumentation.org/packages/gmodels/versions/2.18.1/topics/CrossTab e*