Preparation

Initial file inspection of the csv file to be used shows that the delimiter used is a semicolon, and uses categorical values.

bank-full.csv

CS174-M1 Project > bank > bank-full.csv

Age";"job";"marital";"education";"default";"balance";"housing";"loan";"contact";"day";"month";"duration";"campaign";"pdays";"previo
58;"management";"married";"tertiary";"no";2143;"yes";"no";"unknown";5;"may";261;1;-1;0;"unknown";"no"
44;"technician";"single";"secondary";"no";29;"yes";"no";"unknown";5;"may";151;1;-1;0;"unknown";"no"
33;"entrepreneur";"married";"secondary";"no";2;"yes";"yes";"unknown";5;"may";76;1;-1;0;"unknown";"no"
47;"blue-collar";"married";"unknown";"no";1506;"yes";"no";"unknown";5;"may";92;1;-1;0;"unknown";"no"
33;"unknown";"single";"unknown";"no";1;"no";"no";"unknown";5;"may";198;1;-1;0;"unknown";"no"
35;"management";"married";"tertiary";"no";231;"yes";"no";"unknown";5;"may";139;1;-1;0;"unknown";"no"
28;"management";"single";"tertiary";"no";447;"yes";"yes";"unknown";5;"may";217;1;-1;0;"unknown";"no"
42;"entrepreneur";"divorced";"tertiary";"yes";2;"yes";"no";"unknown";5;"may";380;1;-1;0;"unknown";"no"
58;"retired";"married";"primary";"no";121;"yes";"no";"unknown";5;"may";50;1;-1;0;"unknown";"no"
43;"technician";"single";"secondary";"no";593;"yes";"no";"unknown";5;"may";55;1;-1;0;"unknown";"no"
41;"admin.";"divorced";"secondary";"no";270;"yes";"no";"unknown";5;"may";222;1;-1;0;"unknown";"no"
29;"admin.";"single";"secondary";"no";390;"yes";"no";"unknown";5;"may";137;1;-1;0;"unknown";"no"
53;"technician";"married";"secondary";"no";6;"yes";"no";"unknown";5;"may";517;1;-1;0;"unknown";"no"

This can be fixed by replacing the semicolons to a comma using notepad find and replace all

bank_full.csv - Notepad

File Edit View

;

,

Replace Replace all

```
"age";"job";"marita                                                    uration";"campaign"
;"pdays";"previous"
58;"management";"ma                                                    "no"
44;"technician";"si                                                    o"
33;"entrepreneur";"married";"secondary";"no";2;"yes";"yes";"unknown";5;"may";76;1;-1;0;"unknown";"no"
47;"blue-collar";"married";"unknown";"no";1506;"yes";"no";"unknown";5;"may";92;1;-1;0;"unknown";"no"
33;"unknown";"single";"unknown";"no";1;"no";"no";"unknown";5;"may";198;1;-1;0;"unknown";"no"
35;"management";"married";"tertiary";"no";231;"yes";"no";"unknown";5;"may";139;1;-1;0;"unknown";"no"
28;"management";"single";"tertiary";"no";447;"yes";"yes";"unknown";5;"may";217;1;-1;0;"unknown";"no"
42;"entrepreneur";"divorced";"tertiary";"yes";2;"yes";"no";"unknown";5;"may";380;1;-1;0;"unknown";"no"
58;"retired";"married";"primary";"no";121;"yes";"no";"unknown";5;"may";50;1;-1;0;"unknown";"no"
43;"technician";"single";"secondary";"no";593;"yes";"no";"unknown";5;"may";55;1;-1;0;"unknown";"no"
41;"admin.";"divorced";"secondary";"no";270;"yes";"no";"unknown";5;"may";222;1;-1;0;"unknown";"no"
29;"admin.";"single";"secondary";"no";390;"yes";"no";"unknown";5;"may";137;1;-1;0;"unknown";"no"
53;"technician";"married";"secondary";"no";6;"yes";"no";"unknown";5;"may";517;1;-1;0;"unknown";"no"
58;"technician";"married";"unknown";"no";71;"yes";"no";"unknown";5;"may";71;1;-1;0;"unknown";"no"
57;"services";"married";"secondary";"no";162;"yes";"no";"unknown";5;"may";174;1;-1;0;"unknown";"no"
51;"retired";"married";"primary";"no";229;"yes";"no";"unknown";5;"may";353;1;-1;0;"unknown";"no"
45;"admin.";"single";"unknown";"no";13;"yes";"no";"unknown";5;"may";98;1;-1;0;"unknown";"no"
57;"blue-collar";"married";"primary";"no";52;"yes";"no";"unknown";5;"may";38;1;-1;0;"unknown";"no"
60;"retired";"married";"primary";"no";60;"yes";"no";"unknown";5;"may";219;1;-1;0;"unknown";"no"
33;"services";"married";"secondary";"no";0;"yes";"no";"unknown";5;"may";54;1;-1;0;"unknown";"no"
28;"blue-collar";"married";"secondary";"no";723;"yes";"yes";"unknown";5;"may";262;1;-1;0;"unknown";"no"
56;"management";"married";"tertiary";"no";779;"yes";"no";"unknown";5;"may";164;1;-1;0;"unknown";"no"
```

Ln 5, Col 77 100% Unix (LF) UTF-8

The columns are now recognized properly

cleaner.py bank_full.csv

bank_full.csv

Age	Job	Marital	Education	Default	Balance	Housing	Loan	Contact	Day	Month
Z	management	married	tertiary	no	2143	yes	no	unknown	5	may
44	technician	single	secondary	no	29	yes	no	unknown	5	may
33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may
47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may
33	unknown	single	unknown	no	1	no	no	unknown	5	may
35	management	married	tertiary	no	231	yes	no	unknown	5	may
28	management	single	tertiary	no	447	yes	yes	unknown	5	may
42	entrepreneur	divorced	tertiary	yes	2	yes	no	unknown	5	may
58	retired	married	primary	no	121	yes	no	unknown	5	may
43	technician	single	secondary	no	593	yes	no	unknown	5	may
41	admin.	divorced	secondary	no	270	yes	no	unknown	5	may

Encoding the categorical columns

```python
from sklearn.preprocessing import LabelEncoder

dataTypes = df_bankFull.dtypes

df_bankFull.head()

for x, column in enumerate(df_bankFull):
    if not((str(dataTypes[x])) == 'int64'):
        print(column , df_bankFull[column].unique())
        encoder = LabelEncoder()
        values = encoder.fit_transform(df_bankFull[column])
        df_bankFull[column] = values

df_bankFull.head()
```

✓ 0.1s Python

```
job ['management' 'technician' 'entrepreneur' 'blue-collar' 'unknown'
 'retired' 'admin.' 'services' 'self-employed' 'unemployed' 'housemaid'
 'student']
marital ['married' 'single' 'divorced']
education ['tertiary' 'secondary' 'unknown' 'primary']
default ['no' 'yes']
housing ['yes' 'no']
loan ['no' 'yes']
contact ['unknown' 'cellular' 'telephone']
month ['may' 'jun' 'jul' 'aug' 'oct' 'nov' 'dec' 'jan' 'feb' 'mar' 'apr' 'sep']
poutcome ['unknown' 'failure' 'other' 'success']
y ['no' 'yes']
```

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign
0	58	4	1	2	0	2143	1	0	2	5	8	261	1
1	44	9	2	1	0	29	1	0	2	5	8	151	1
2	33	2	1	1	0	2	1	1	2	5	8	76	1
3	47	1	1	3	0	1506	1	0	2	5	8	92	1
4	33	11	2	3	0	1	0	0	2	5	8	198	1

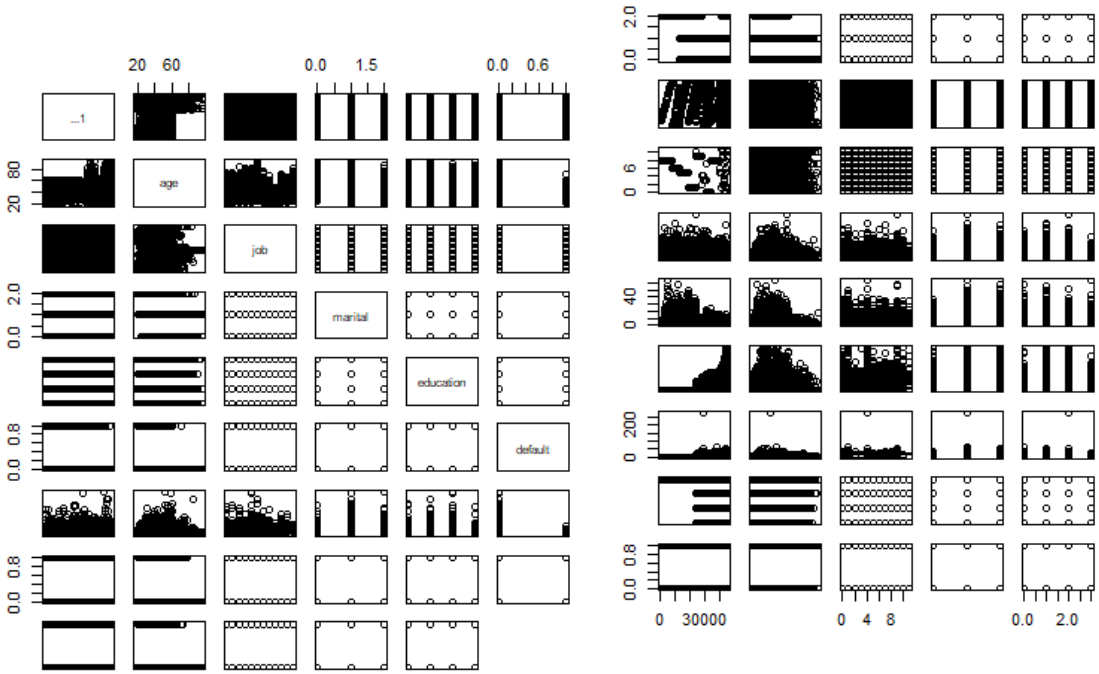
Exploratory Data Analysis

Performing exploratory data analysis to give us a better overview or context about our data.

```
> print(head(data))
# A tibble: 6 × 17
    age   job marital educa…¹ default balance housing  loan contact   day month durat…² campa…³
  <dbl> <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl> <dbl>   <dbl> <dbl> <dbl>   <dbl>   <dbl>
1    58     4       1       2       0    2143       1     0       2     5     8     261       1
2    44     9       2       1       0      29       1     0       2     5     8     151       1
3    33     2       1       1       0       2       1     1       2     5     8      76       1
4    47     1       1       3       0    1506       1     0       2     5     8      92       1
5    33    11       2       3       0       1       0     0       2     5     8     198       1
6    35     4       1       2       0     231       1     0       2     5     8     139       1
# … with 4 more variables: pdays <dbl>, previous <dbl>, poutcome <dbl>, y <dbl>, and
#   abbreviated variable names ¹education, ²duration, ³campaign
# ℹ Use `colnames()` to see all variable names
```

```
> summary(data)
      age             job           marital        education        default
 Min.   :18.00   Min.   : 0.00   Min.   :0.000   Min.   :0.000   Min.   :0.00000
 1st Qu.:33.00   1st Qu.: 1.00   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.00000
 Median :39.00   Median : 4.00   Median :1.000   Median :1.000   Median :0.00000
 Mean   :40.94   Mean   : 4.34   Mean   :1.168   Mean   :1.225   Mean   :0.01803
 3rd Qu.:48.00   3rd Qu.: 7.00   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:0.00000
 Max.   :95.00   Max.   :11.00   Max.   :2.000   Max.   :3.000   Max.   :1.00000
    balance          housing            loan            contact            day
 Min.   : -8019   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   : 1.00
 1st Qu.:    72   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 8.00
 Median :   448   Median :1.0000   Median :0.0000   Median :0.0000   Median :16.00
 Mean   :  1362   Mean   :0.5558   Mean   :0.1602   Mean   :0.6402   Mean   :15.81
 3rd Qu.:  1428   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:2.0000   3rd Qu.:21.00
 Max.   :102127   Max.   :1.0000   Max.   :1.0000   Max.   :2.0000   Max.   :31.00

     month           duration         campaign          pdays          previous
 Min.   : 0.000   Min.   :   0.0   Min.   : 1.000   Min.   : -1.0   Min.   :  0.0000
 1st Qu.: 3.000   1st Qu.: 103.0   1st Qu.: 1.000   1st Qu.: -1.0   1st Qu.:  0.0000
 Median : 6.000   Median : 180.0   Median : 2.000   Median : -1.0   Median :  0.0000
 Mean   : 5.523   Mean   : 258.2   Mean   : 2.764   Mean   : 40.2   Mean   :  0.5803
 3rd Qu.: 8.000   3rd Qu.: 319.0   3rd Qu.: 3.000   3rd Qu.: -1.0   3rd Qu.:  0.0000
 Max.   :11.000   Max.   :4918.0   Max.   :63.000   Max.   :871.0   Max.   :275.0000
    poutcome          y
 Min.   :0.00   Min.   :0.000
 1st Qu.:3.00   1st Qu.:0.000
 Median :3.00   Median :0.000
 Mean   :2.56   Mean   :0.117
 3rd Qu.:3.00   3rd Qu.:0.000
 Max.   :3.00   Max.   :1.000
```

```
> cor(data)
                    age           job       marital     education       default       balance
age        1.000000000 -0.0218679434 -0.403240136 -1.068066e-01 -0.017879304  0.097782739
job       -0.021867943  1.0000000000  0.062045485  1.667067e-01 -0.006853085  0.018231515
marital   -0.403240136  0.0620454852  1.000000000  1.085761e-01 -0.007023365  0.002121918
education -0.106806594  0.1667067239  0.108576125  1.000000e+00 -0.010717690  0.064514043
default   -0.017879304 -0.0068530852 -0.007023365 -1.071769e-02  1.000000000 -0.066745057
balance    0.097782739  0.0182315155  0.002121918  6.451404e-02 -0.066745057  1.000000000
housing   -0.185513082 -0.1253628132 -0.016095882 -9.079024e-02 -0.006025218 -0.068768316
loan      -0.015655273 -0.0330039210 -0.046892524 -4.857353e-02  0.077234241 -0.084350246
contact    0.026221067 -0.0820633039 -0.039201423 -1.109276e-01  0.015404140 -0.027272944
day       -0.009120046  0.0228555732 -0.005261364  2.267105e-02  0.009423899  0.004502585
month     -0.042357405 -0.0928695791 -0.006990661 -5.730383e-02  0.011485783  0.019777231
duration  -0.004648428  0.0047436409  0.011852173  1.935105e-03 -0.010021461  0.021560380
campaign   0.004760312  0.0068386259 -0.008994100  6.255137e-03  0.016821531 -0.014578279
pdays     -0.023758014 -0.0244550401  0.019172254  5.235498e-05 -0.029979361  0.003435322
previous   0.001288319 -0.0009106174  0.014973243  1.756963e-02 -0.018329405  0.016673637
poutcome   0.007366903  0.0110103583 -0.016850456 -1.936137e-02  0.034898194 -0.020967337
y          0.025155017  0.0404380188  0.045587526  6.624056e-02 -0.022418966  0.052838410
-
               housing         loan     contact          day        month     duration
age       -0.185513082 -0.015655273  0.02622107 -0.009120046 -0.042357405 -0.004648428
job       -0.125362813 -0.033003921 -0.08206330  0.022855573 -0.092869579  0.004743641
marital   -0.016095882 -0.046892524 -0.03920142 -0.005261364 -0.006990661  0.011852173
education -0.090790237 -0.048573533 -0.11092757  0.022671046 -0.057303833  0.001935105
default   -0.006025218  0.077234241  0.01540414  0.009423899  0.011485783 -0.010021461
balance   -0.068768316 -0.084350246 -0.02727294  0.004502585  0.019777231  0.021560380
housing    1.000000000  0.041322866  0.18812289 -0.027981649  0.271480739  0.005075449
loan       0.041322866  1.000000000 -0.01087301  0.011370158  0.022144853 -0.012411972
contact    0.188122888 -0.010873011  1.00000000 -0.027936231  0.361144884 -0.020839303
day       -0.027981649  0.011370158 -0.02793623  1.000000000 -0.006027676 -0.030206341
month      0.271480739  0.022144853  0.36114488 -0.006027676  1.000000000  0.006313636
duration   0.005075449 -0.012411972 -0.02083930 -0.030206341  0.006313636  1.000000000
campaign  -0.023598707  0.009979846  0.01961438  0.162490216 -0.110030865 -0.084569503
pdays      0.124178400 -0.022753639 -0.24481646 -0.093044074  0.033064690 -0.001564770
previous   0.037076150 -0.011043488 -0.14781140 -0.051710497  0.022727145  0.001203057
poutcome  -0.099970667  0.015457767  0.27221380  0.083459682 -0.033038191  0.010925350
y         -0.139172702 -0.068185035 -0.14839488 -0.028347777 -0.024471438  0.394521016
              campaign         pdays      previous     poutcome           y
age        0.004760312 -2.375801e-02  0.0012883192  0.007366903  0.02515502
job        0.006838626 -2.445504e-02 -0.0009106174  0.011010358  0.04043802
marital   -0.008994100  1.917225e-02  0.0149732426 -0.016850456  0.04558753
education  0.006255137  5.235498e-05  0.0175696313 -0.019361368  0.06624056
default    0.016821531 -2.997936e-02 -0.0183294048  0.034898194 -0.02241897
balance   -0.014578279  3.435322e-03  0.0166736367 -0.020967337  0.05283841
housing   -0.023598707  1.241784e-01  0.0370761497 -0.099970667 -0.13917270
loan       0.009979846 -2.275364e-02 -0.0110434883  0.015457767 -0.06818503
contact    0.019614376 -2.448165e-01 -0.1478113997  0.272213798 -0.14839488
day        0.162490216 -9.304407e-02 -0.0517104967  0.083459682 -0.02834778
month     -0.110030865  3.306469e-02  0.0227271448 -0.033038191 -0.02447144
duration  -0.084569503 -1.564770e-03  0.0012030569  0.010925350  0.39452102
campaign   1.000000000 -8.862767e-02 -0.0328552897  0.101587641 -0.07317201
pdays     -0.088627668  1.000000e+00  0.4548196355 -0.858361643  0.10362149
previous  -0.032855290  4.548196e-01  1.0000000000 -0.489751865  0.09323577
poutcome   0.101587641 -8.583616e-01 -0.4897518649  1.000000000 -0.07784038
y         -0.073172006  1.036215e-01  0.0932357728 -0.077840384  1.00000000
```



We now check if our chosen variables meet the assumptions of multiple regression.

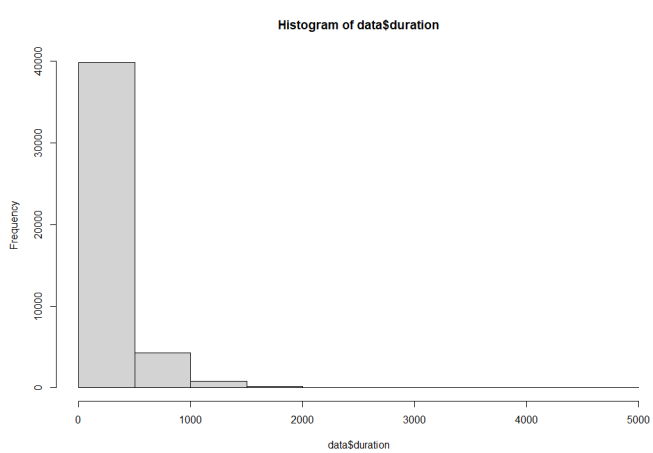
The dependent variable is correlated with each of the independent variables.

```
> cor(data$duration, data$age)
[1] -0.004648428
> cor(data$duration, data$balance)
[1] 0.02156038
```

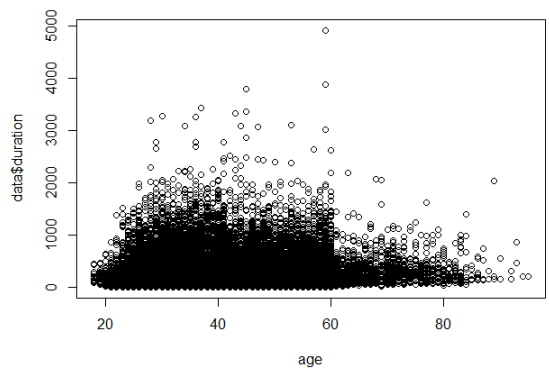
The independent variables are not correlated with each other.

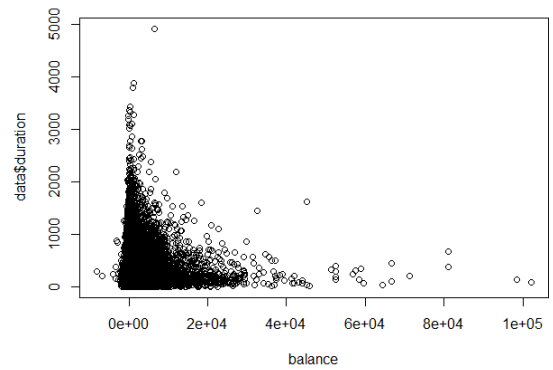
```
> cor(data$balance, data$age)
[1] 0.09778274
```

```
> hist(data$duration)
```



Linearity of dependent with each independent variable





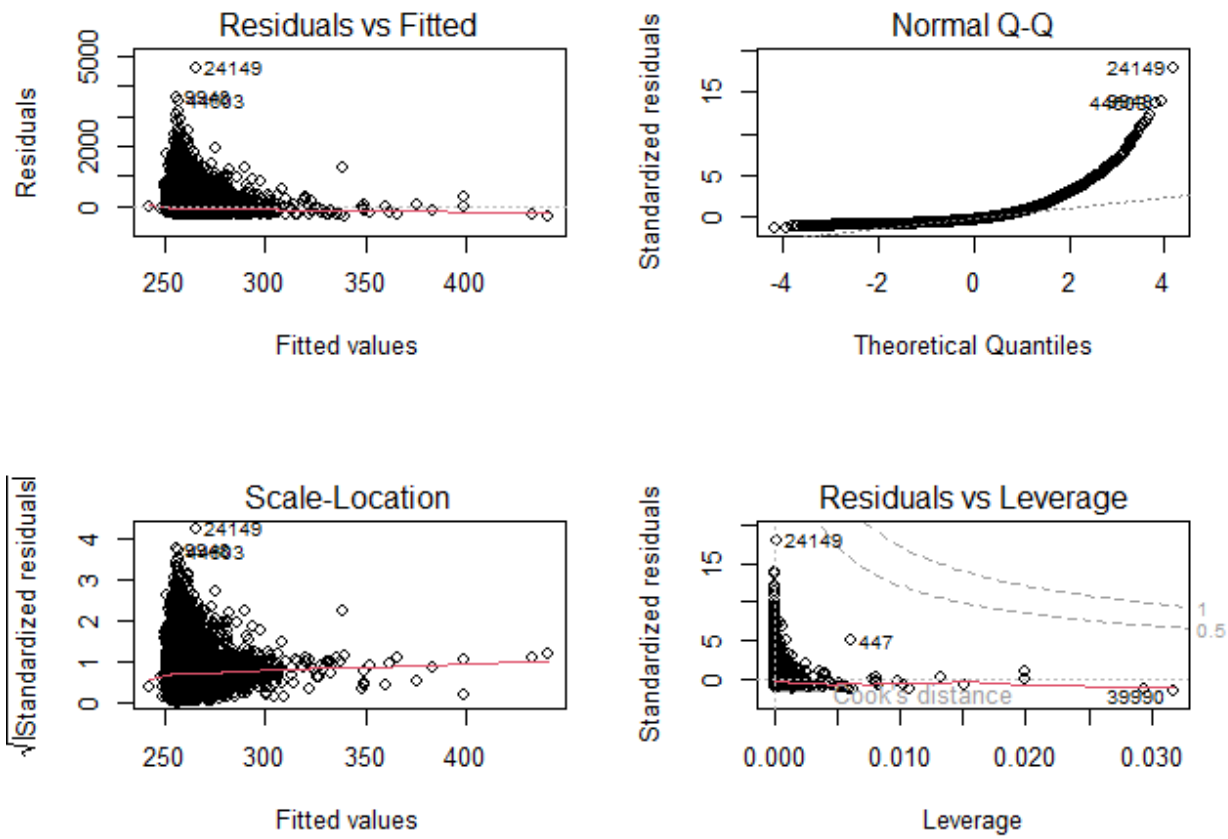

```
> dim(train)
[1] 34573    17
> print(head(train))
  age job marital education default balance housing loan contact day month duration campaign pdays previous poutcome y
1  58   4       1         2       0    2143       1    0       2   5     8      261        1    -1        0        3 0
2  44   9       2         1       0      29       1    0       2   5     8      151        1    -1        0        3 0
3  33   2       1         1       0       2       1    1       2   5     8       76        1    -1        0        3 0
5  33  11       2         3       0       1       0    0       2   5     8      198        1    -1        0        3 0
6  35   4       1         2       0     231       1    0       2   5     8      139        1    -1        0        3 0
7  28   4       2         2       0     447       1    1       2   5     8      217        1    -1        0        3 0
> dim(test)
[1] 10638    17
> print(head(test))
   age job marital education default balance housing loan contact day month duration campaign pdays previous poutcome y
4   47   1       1         3       0    1506       1    0       2   5     8       92        1    -1        0        3 0
8   42   2       0         2       1       2       1    0       2   5     8      380        1    -1        0        3 0
9   58   5       1         0       0     121       1    0       2   5     8       50        1    -1        0        3 0
14  58   9       1         3       0      71       1    0       2   5     8       71        1    -1        0        3 0
21  28   1       1         1       0     723       1    1       2   5     8      262        1    -1        0        3 0
25  40   5       1         0       0       0       1    1       2   5     8      181        1    -1        0        3 0
```

```
> model_all=lm(duration~balance+age, data=train)
> summary(model_all)

Call:
lm(formula = duration ~ balance + age, data = train)

Residuals:
   Min     1Q Median     3Q    Max 
-351.3 -155.6  -79.1   60.0 4652.6 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.607e+02  5.570e+00  46.793  < 2e-16 ***
balance      1.830e-03  4.592e-04   3.986 6.74e-05 ***
age         -1.229e-01  1.326e-01  -0.927    0.354    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 259.4 on 34570 degrees of freedom
Multiple R-squared:  0.0004674, Adjusted R-squared:  0.0004096 
F-statistic: 8.083 on 2 and 34570 DF,  p-value: 0.0003092

> print(coef(model_all))
  (Intercept)       balance           age 
260.658178325   0.001830427  -0.122879524 
```

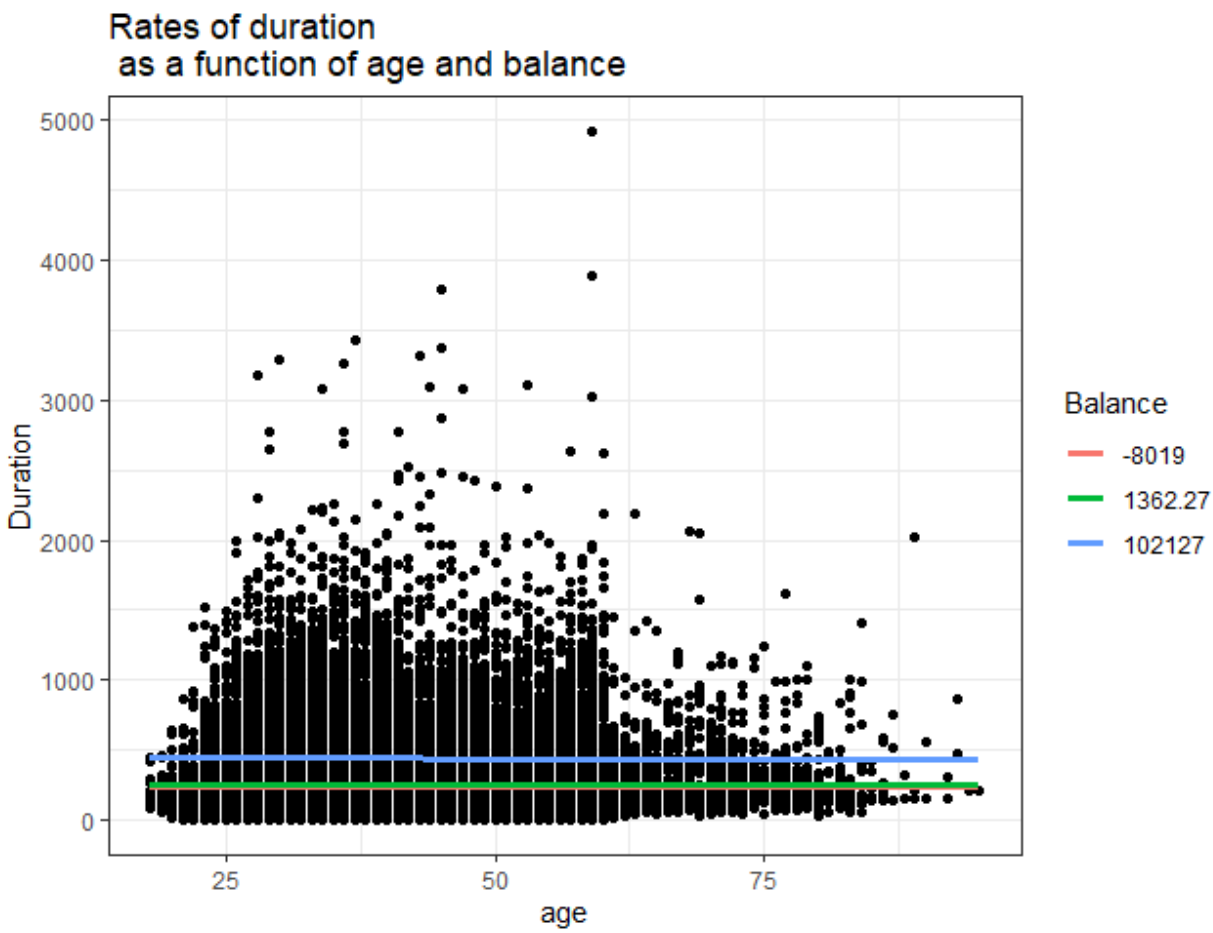
Our equation for out regression model given the balance(x1) and age(x2) to predict the duration value(Y) would be

$$Y = 260.658178325 + (0.001830427) * x_1 + (-0.122879524) * x_2$$

Make sure that the model fits the homoscedasticity assumption. We check if our model is actually a good fit for the data, and that we don't have large variation in the model error.



We can see that the red lines which represent the means are all centered around zero, meaning there are no outliers or biases in our data to make our regression invalid. Although the normal Q-Q forms a curve towards the end which means that the sample data are skewed towards that value of the data.



Seeing the line regression lines in an almost flat line, we can say that there are barely any associations with the variables.

We perform prediction on the test dataset using the model.

```
> pred <- predict(model_all,test)
> pred
       2        7       11       16       19       24       28       33       36       41
255.3046 258.0358 256.1143 254.8105 253.3952 257.6777 254.4753 253.3568 253.7694 259.9850
      45       50       53       58       62       67       70       75       79       84
253.7069 257.0947 256.7260 255.4430 255.9559 253.4684 278.7307 258.9199 254.6371 257.6970
      87       92       96      101      104      109      113      118      121      126
253.8593 257.2500 256.4194 254.0178 253.7359 253.4632 253.7147 256.4805 253.8794 256.7953
     130      135      138      143      147      152      155      160      164      169
```

```
> rmse_val <- sqrt(mean(pred-test$age)^2)
> rmse_val
[1] 216.9893
>
> SSE = sum((pred-test$age)^2)
> SST = sum((pred-mean(test$age))^2)
> r2_test = 1 - SSE/SST
> print(r2_test)
[1] -0.002828329
```

In the bank dataset, we found that there is little relationship between the age and duration, and the balance and duration ($p < 0$ and $p < 0.022$, respectively).

Specifically we found a 0.18% increase(± 0.003) in the duration for every 1% increase in balance, and a 12.287% decrease (± 0.009524) in the duration for every 1% increase in age. This means that the model is unreliable to use when predicting the duration given the time and age. This can be due to the skewed nature of the dataset, and lack of correlation of each of the independent variables with the dependent variable. Observing the correlation values from the initial data exploration, we observe that there is not much correlation between the variables, therefore it was hard for the data to meet the multiple regression assumptions.

We can see that our model has an RMSE (Root Mean Squared Error) value of 216.9893 means that on average, the model's predictions are off by 216.9893 units. The model also has an R-squared value of -0.002828329 indicates that the model does not fit the data well. A negative R-squared value indicates that the model fits the data worse than a horizontal line through the mean of the dependent variable would. This means that the model is not providing any useful information to predict the outcome variable. In other words, the model may not be appropriate for the data, the relationship between the dependent and independent variables may not be linear, or there may be other factors affecting the outcome variable that are not captured by the model. Therefore, a negative R-squared value suggests that the model should be revised or that additional variables should be considered to improve the model's predictive power.

References:

[1] Abhinav Agarwal. 2022. How to perform multiple linear regression in R -. *ProjectPro*. Retrieved March 1, 2023 from https://www.projectpro.io/recipes/perform-multiple-linear-regression-r

[2] Rebecca Bevans. 2020. Linear Regression in R. Scribbr. Retrieved March 2, 2023 from https://www.scribbr.com/statistics/linear-regression-in-r/

[3] Rohit Sharma. 2022. Multiple linear regression in R [with graphs & examples]. *UpGrad*. Retrieved March 1, 2023 from https://www.upgrad.com/blog/multiple-linear-regression-in-r/