Techniques de Forage des données
SIO-6003
Session automne 2022

Projet de mi-session

Destinataire
Mohammed Lassaad Ammari

Date de remise : 30 octobre 2022

Salah Eddine Khalil

1. Analyse descriptive :

Comme le montre la figure ci-dessous, Notre jeu de données contient 503 enregistrements avec 12 variables dont 3 sont catégorielles et 9 sont quantitatives. On peut remarquer aussi que notre jeu ne contient aucune valeur manquante :

Basic Statistics

Raw Counts

Name	Value
Rows	503
Columns	12
Discrete columns	3
Continuous columns	9
All missing columns	0
Missing observations	0
Complete Rows	503
Total observations	6,036
Memory allocation	35.1 Kb

Statistiques ‘’ValeurAchat’’ :

La plage des données pour cette variable est entre 104 et 610 avec 50% des données qui sont concentrées dans l’intervalle interquartille (209.5 , 371).

```
> summary(data$ValeurAchat)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  104.0   209.5   288.0   296.6   371.0   610.0
```

Statistiques ‘’Revenu’’ :

```
> summary(data$Revenu)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1016    3190    5348    5414    7508    9997
```

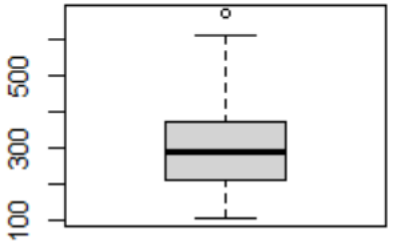
Statistiques ‘’InvestBitcoin’’ :

```
> summary(data$InvestBitcoin)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  101.0   201.5   303.0   302.8   405.5   499.0
```

Pré-traitement des données :

- L’enregistrement 381 ne respecte pas la norme de définition du genre donc on va l’adapter aux autres données : `data$Genre[381]="M"`

- ValeurAchat présente une valeur abberante (672) tel que montré dans le box plot :



Qu’on va remplacer par la moyenne :

```
> summary(data$ValeurAchat)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  104.0   209.5   288.0   297.4   371.0   672.0
> data$ValeurAchat[476]=297.4
> data$ValeurAchat[476]
[1] 297.4
>
```

- On supprime la variable genre puisque tous les individus sont masculins : `data <- data[,-1]`
- On transforme Les deux variables catégorielles ‘’AmznPrim’’ et ‘’Fidélité’’ en des variables indicactrices qui vont nous permettre de faire la régression par la suite.

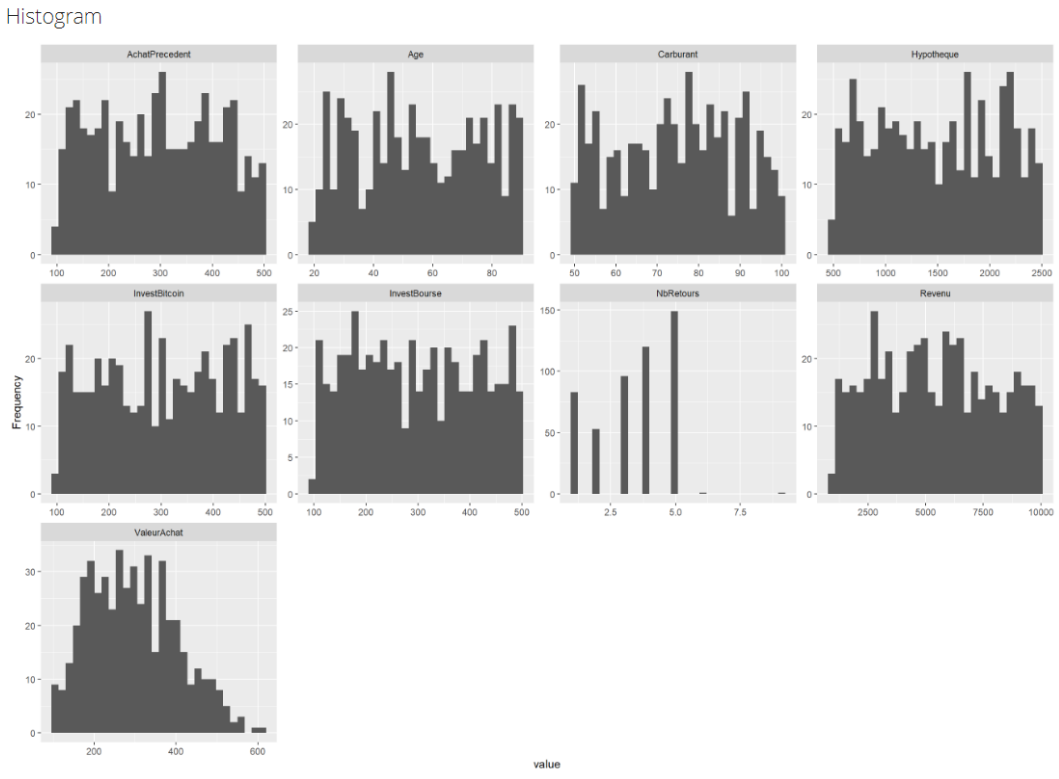
Le jeu de données après pré-traitement :

Filter

	Age	Revenu	InvestBourse	InvestBitcoin	NbRetours	Carburant	Hypotheque	AchatPrecedent	ValeurAchat	AmznPrime	Fidelite
1	48	1016	106	108	4	83	1709	118	135	1	0
2	53	1021	383	434	4	79	2496	405	161	0	0
3	74	1053	145	444	5	84	2420	437	156	0	0
4	59	1086	201	495	1	56	1931	280	118	0	0
5	78	1092	196	353	5	72	1210	374	156	0	0
6	57	1101	265	214	1	80	1013	343	236	0	0
7	74	1101	308	272	1	90	1851	264	171	0	0
8	84	1110	113	353	3	63	682	295	104	0	0
9	24	1112	396	209	2	92	1211	285	121	1	0
10	45	1125	438	410	1	55	1311	385	159	0	0

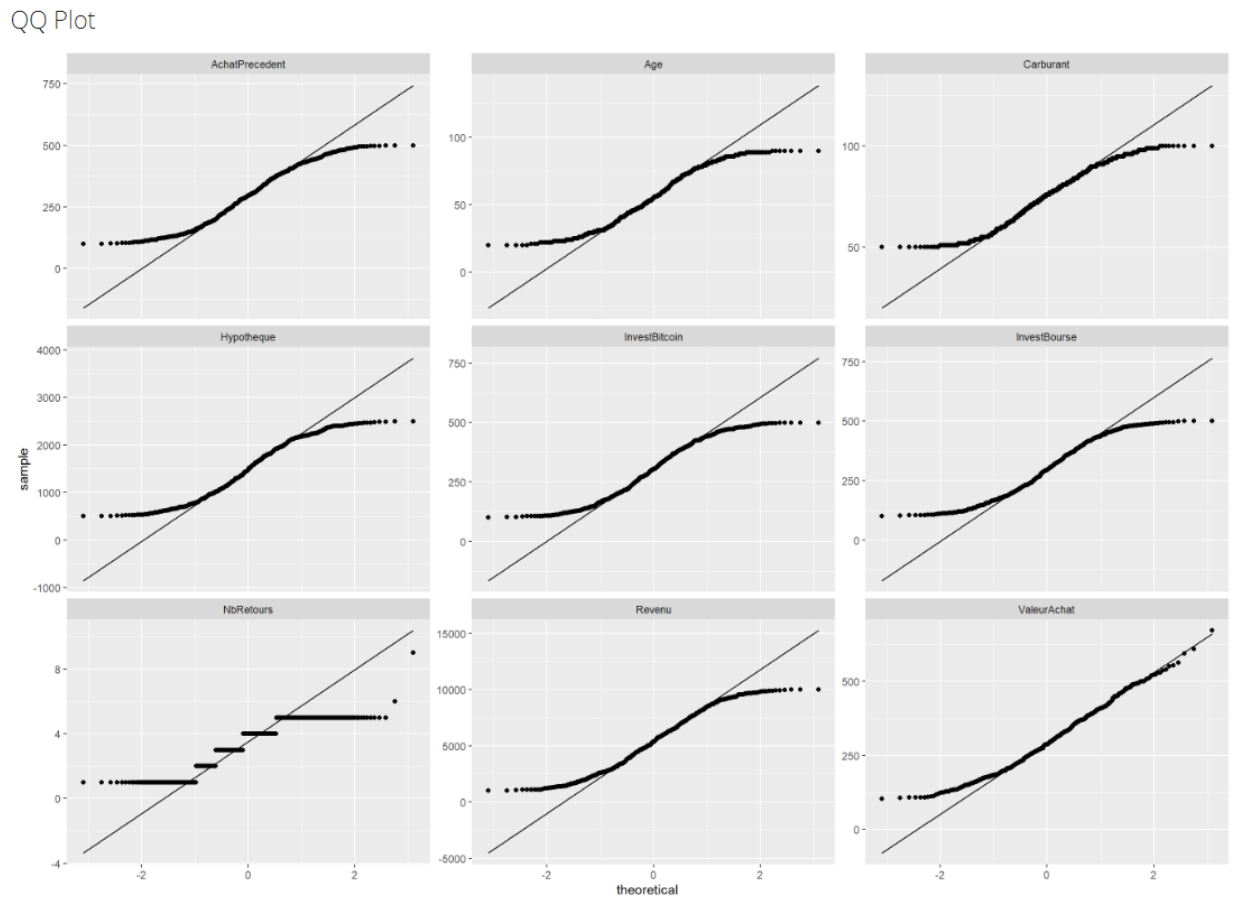
2. Visualisations et Conclusions:

La Distribution des variables sous forme d’Histogrammes nous donne :

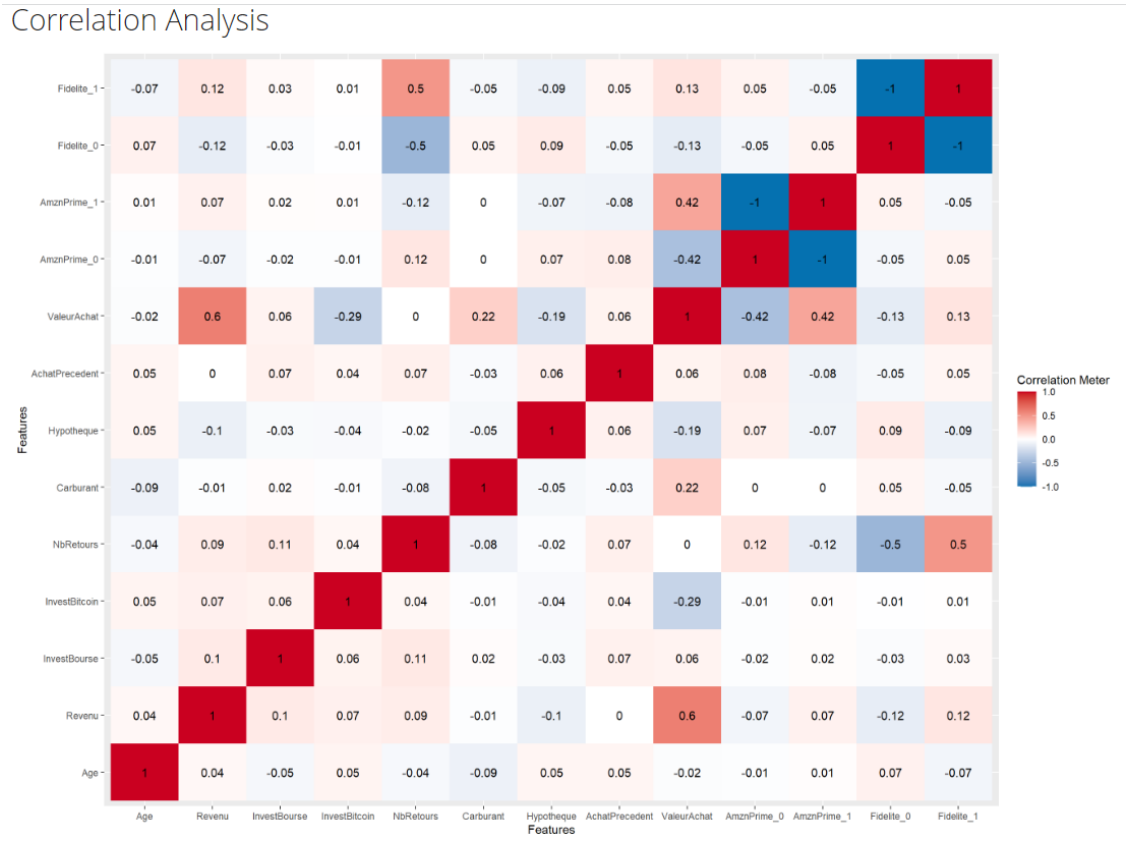


D’où on pourrait conclure qu’aucune variable n’a une distribution normal. pourtant la variable ValeurAchat semble avoir une distribution presque normale et présente une valeur abberante qui est bien isolée au niveau de l’histogramme.

Le Q-Q plot met en évidence lui aussi la non normalité de toutes les variables :



La matrice de corrélartion :



La plupart des variables de notre jeu ne sont pas corrélées, et il n'y a pas non plus une corrélation parfaite, mais on peut remarquer qu'ils existent certaines corrélations cor(ValeurAchat, Revenu)=0.6 et cor(NbRetours, Fidelité)= 0.5 .

3. Influence de ''AmznPrim'' sur ''Fidélité'' :

Le tableau de contingence :

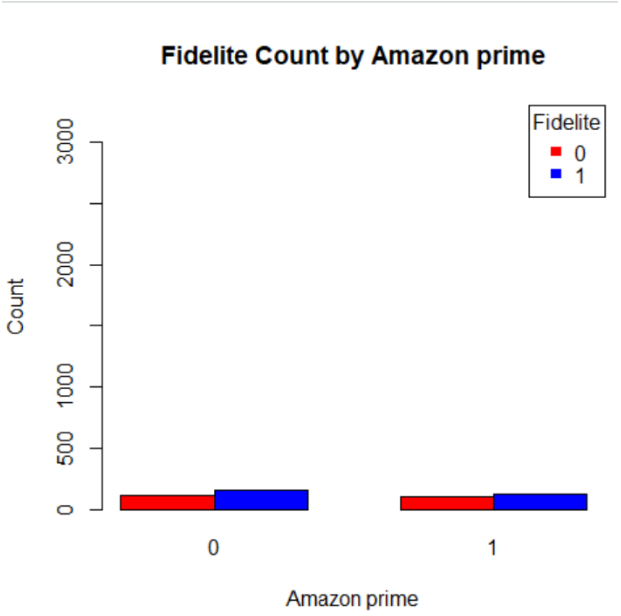
Fidelite	Amazon Prime 0	1
0	115	107
1	161	120

Le tableau de contingence en terme de pourcentages :

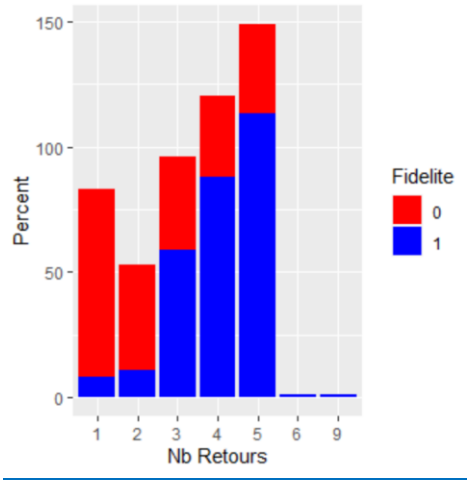
Fidelite	Amazon Prime 0	1
0	22.86	21.27
1	32.01	23.86

On remarque que les clients ayant un abonnement ''AmznPrim'' ne sont pas trop adhérés au programme de fidélité tels que ceux qui n'ont pas l'abonnement ''AmznPrim'', donc on peut constater que le degré de fidélité est très grand chez les non adhérés à ''AmznPrim'' par rapport aux adhérés.

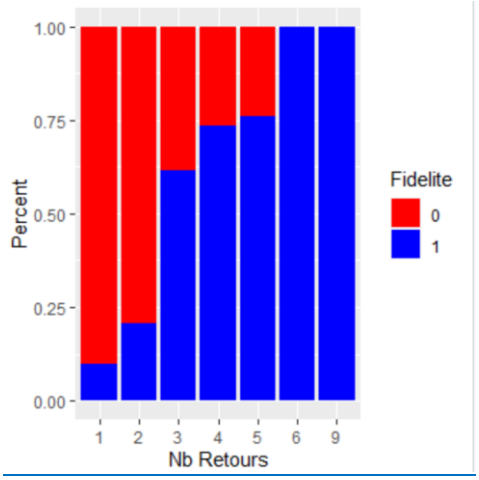
Graphique :



4. Superposition de ''NbRetours'' et ''Fidélité'' :



On remarque que plus le nombre de retours par les clients augmente, l'adhéssion au programme de fidélité augmente aussi, qui est encore bien expliqué en terme de diagramme en pourcentages suivant :



5. IC à 95% pour la proportion les clients adhérés au programme de fidélité :

```
> #Intervalle de confiance
> prop.test(table(data$Fidelite==0),conf.level = 0.95)$"conf.int"
[1] 0.5139759 0.6024110
attr(,"conf.level")
[1] 0.95
```

On est certain à 95% que la proportion des clients adhérés au programme de fidélité est entre 51.39% et 60.24% . ce qui est bien logique puisque le nombre de clients adhérés au programme de fidélité(281) est plus grand de celui qui ne le sont pas (222).

6. Comparaison des moyennes de valeurAchat des clients ''fidèles'' vs ''non fidèles'' :

On subdivisant notre table en deux tables séparés, l'une qui contienne seulement les fidèles et l'autre seulement les non fidèles, on trouve après faire le test sur la différence entre les moyennes de la valeur d'achat entre les deux tables :

```
> t.test(subfidel1$ValeurAchat, subfidel2$ValeurAchat,alternative='two.sided',conf.level = 0.95);

        Welch Two Sample t-test

data:  subfidel1$ValeurAchat and subfidel2$ValeurAchat
t = 3.0856, df = 462.2, p-value = 0.002153
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 10.63444 47.93623
sample estimates:
mean of x mean of y
 309.5196  280.2342
```

p-value =0,002 < alpha=0,05

donc on rejette l'hypothèse nulle c'est-à-dire qu'on est certain à 95% que les deux moyennes sont différentes, ce qui est bien justifié par les deux moyennes :

mean(ValeurAchat pour fidèles) = 309.51

mean(ValeurAchat pour non fidèles) = 280.23

7. Régression ''Valeur d'achat'' :

Variable dépendante : ValeurAchat

Variable indépendante : tous les autres variables

```
Call:
lm(formula = ValeurAchat ~ Revenu + Hypotheque + AchatPrecedent +
    Age + InvestBourse + InvestBitcoin + NbRetours + Carburant +
    AmznPrime + Fidelite, data = data)
```

Les résultats obtenues :

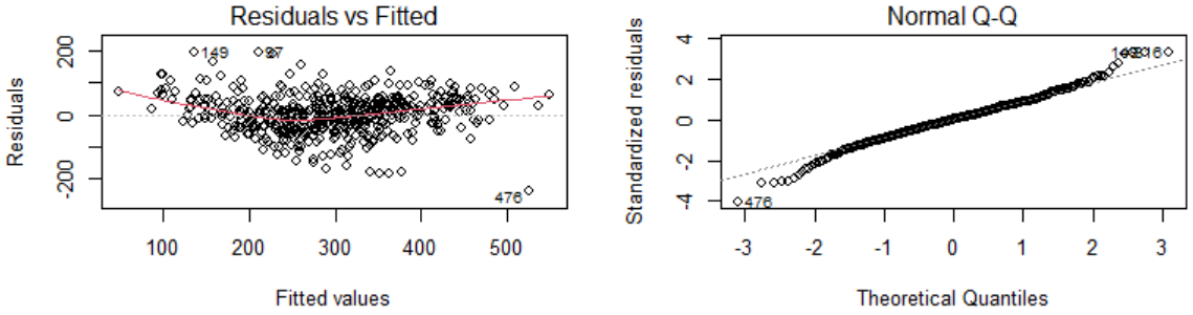
```
Residuals:
     Min       1Q   Median       3Q      Max
-237.431  -34.046    2.165   37.862  194.106

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    90.389703  22.977197   3.934 9.56e-05 ***
Revenu          0.024286   0.001072  22.645  < 2e-16 ***
Hypotheque     -0.018260   0.004611  -3.960 8.59e-05 ***
AchatPrecedent  0.104782   0.023659   4.429 1.17e-05 ***
Age            -0.075640   0.130445  -0.580  0.56227
InvestBourse   -0.009572   0.023164  -0.413  0.67962
InvestBitcoin  -0.297501   0.022712 -13.099  < 2e-16 ***
NbRetours      -1.294564   2.137668  -0.606  0.54506
Carburant       1.649476   0.187157   8.813  < 2e-16 ***
AmznPrime1     83.560847   5.412750  15.438  < 2e-16 ***
Fidelite1      20.451105   6.218295   3.289  0.00108 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 59.49 on 492 degrees of freedom
Multiple R-squared:  0.6909,    Adjusted R-squared:  0.6846
F-statistic:   110 on 10 and 492 DF,  p-value: < 2.2e-16
```

D’après les p-values, on peut déduire que toutes les variables explicatives sont significatives au niveau de cette régression pour un alpha= 5% sauf **Age** et **InvestBourse** qui ne sont pas significatives puisqu’ils ont des p-values supérieur à 5%, Et c’est bien logique parce que l’age et investissement en bourse n’ont pas une vraie influence sur les valeurs d’achats des clients.

Avec un erreur résidual standard de 59,49

Et un modèle qui s’ajuste à 69,09% aux données

Diagnostic :



En observant le diagramme de probabilité normale (Q-Q plot) on constate que :

- La plupart des points du graphique de probabilité normale s'alignent sur une ligne droite. pourtant, ils y a plusieurs valeurs extrêmes qui s’écartent de la droite.
- nous concluons l’hypothèse de normalité n’est pas respectée.

D’après Le deuxième diagramme qui montre la distribution des résidus par rapport aux valeurs prédites on constate que :

- La propagation verticale des points semble uniforme : l’hypothèse de la variance constante (homoscédasticité) est respectée.
- Il existe une courbure évidente (en rouge): l’hypothèse de linéarité n’est pas respectée.
- L’existence de points abberantes qui sont numérotés selon leur ordres dans le jeu de données.

8. Régression ‘’investissement bitcoin’’ :

Variable dépendante : InvestBitcoin

Variable indépendante : tous les autres variables

```
Call:
lm(formula = InvestBitcoin ~ Revenu + Hypotheque + AchatPrecedent +
    Age + InvestBourse + ValeurAchat + NbRetours + Carburant +
    AmznPrime + Fidelite, data = data)
```

Les résultats obtenues :

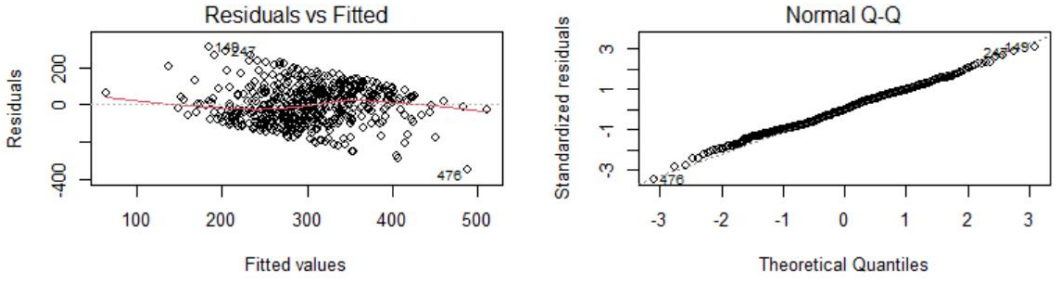
```
Residuals:
    Min      1Q  Median      3Q     Max
-346.80  -76.98   -2.62   72.00  308.51

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    262.291108  38.092324   6.886 1.76e-11 ***
Revenu           0.023101   0.002404   9.611  < 2e-16 ***
Hypotheque      -0.021042   0.007949  -2.647  0.00838 **
AchatPrecedent   0.122008   0.040867   2.985  0.00297 **
Age              0.158832   0.222917   0.713  0.47648
InvestBourse     0.027805   0.039580   0.703  0.48270
ValeurAchat     -0.869112   0.066351 -13.099  < 2e-16 ***
NbRetours        1.295738   3.654600   0.355  0.72308
Carburant        1.412159   0.338277   4.175 3.53e-05 ***
AmznPrime1      74.876269  10.754279   6.962 1.07e-11 ***
Fidelite1       13.544741  10.727151   1.263  0.20731
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 101.7 on 492 degrees of freedom
Multiple R-squared:  0.2692,    Adjusted R-squared:  0.2543
F-statistic: 18.12 on 10 and 492 DF,  p-value: < 2.2e-16
```

D'après les p-values, on peut déduire que toutes les variables explicatives sont significatives au niveau de cette régression pour un alpha= 5% sauf **Nbretours** qui n'est pas significative puisqu'il a un p-value supérieur à 5%, Et c'est bien logique parce que le nombre de retours n'a pas une vraie influence sur l'investissement en bourse des clients.

Avec un erreur résidual standard de 101,7

Et un modèle qui s'ajuste à 26,92% aux données (un très mauvais modèle).

Diagnostic :



En observant le diagramme de probabilité normale (Q-Q plot) on constate que :

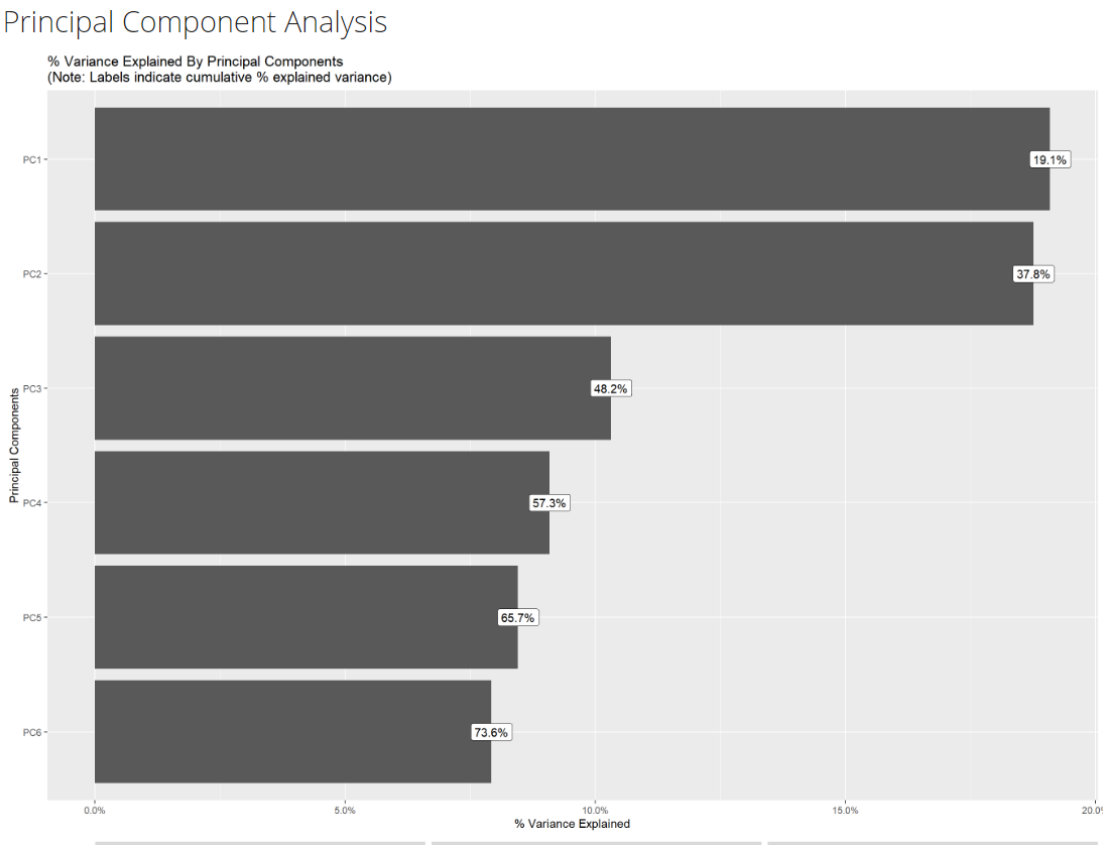
- La plupart des points du graphique de probabilité normale s'alignent sur une ligne droite.

  nous concluons l'hypothèse de normalité est bien respectée.

D'après Le deuxième diagramme qui montre la distribution des résidus par rapport aux valeurs prédites on constate que :

- La propagation verticale des points semble uniforme : homoscédasticité est respectée.
- l'hypothèse de linéarité n'est pas respectée.
- L'existence de points abberantes.

Analyse en Composantes principales :



Analyse en composantes principales permet de bien faciliter l'analyse exploratoire des données en créant un nouvel espace de dimensions réduites tout en gardant le plus d'informations possibles sur nos données, dans cet exemple on choisit 6 composantes principales puisque c'est à partir de la 6ème composante qu'on va avoir plus que 70% de variance totale des données expliquée.

Allons plus en détail et cherchons quelles sont les variables de notre jeu de données qui contribuent le plus à la formation de chaque axe principale :