SIO-6003 - Techniques de forage de données

Projet de mi-session

Loïc Artino (536 756 361)

8 mars 2022

Question 1

Dans le cadre de ce projet, nous avons tout d'abord procédé à l'analyse descriptive du jeu de données.

La première étape est le téléchargement des données, ce qui permet de voir **503** observations réparties dans **12** colonnes, parmi lesquelles $ValeurAchat$ est la variable d'intérêt. Nous pouvons voir une partie de ces données dans le tableau ci-dessous :

Tableau 1 : Aperçu des premières lignes

Genre	Age	Revenu	InvestBourse	InvestBitcoin	NbRetours	Carburant	Hypotheque	AchatPrecedent	ValeurAchat	AmznPrime	Fidelite
M	48	1016	106	108	4	83	1709	118	135	Yes	No
M	53	1021	383	434	4	79	2496	405	161	No	No
M	74	1053	145	444	5	84	2420	437	156	No	No
M	59	1086	201	495	1	56	1931	280	118	No	No
M	78	1092	196	353	5	72	1210	374	156	No	No

De plus, pour s'assurer de l'intégrité de toutes les valeurs des **503** observations, une vérification des valeurs manquantes a été faites. Cette opération nous a permis de constater qu'il y avait **0** valeur(s) manquante(s) dans notre table de données. Par la suite, nous allons observer un résumé du jeu de données :

Tableau 2 : Résumé du jeu de données par la moyenne de chaque variable

Genre	Age	Revenu	InvestBourse	InvestBitcoin	NbRetours	Carburant	Hypotheque	AchatPrecedent	ValeurAchat
M	55.58964	5409.54	297.9343	302.7649	3.410359	74.7749	1487.805	294.7709	297.002
Masculin	56.00000	7549.00	146.0000	323.0000	5.000000	90.0000	710.000	342.0000	476.000

Tableau 3 : Résumé des statistiques pour les variables numériques

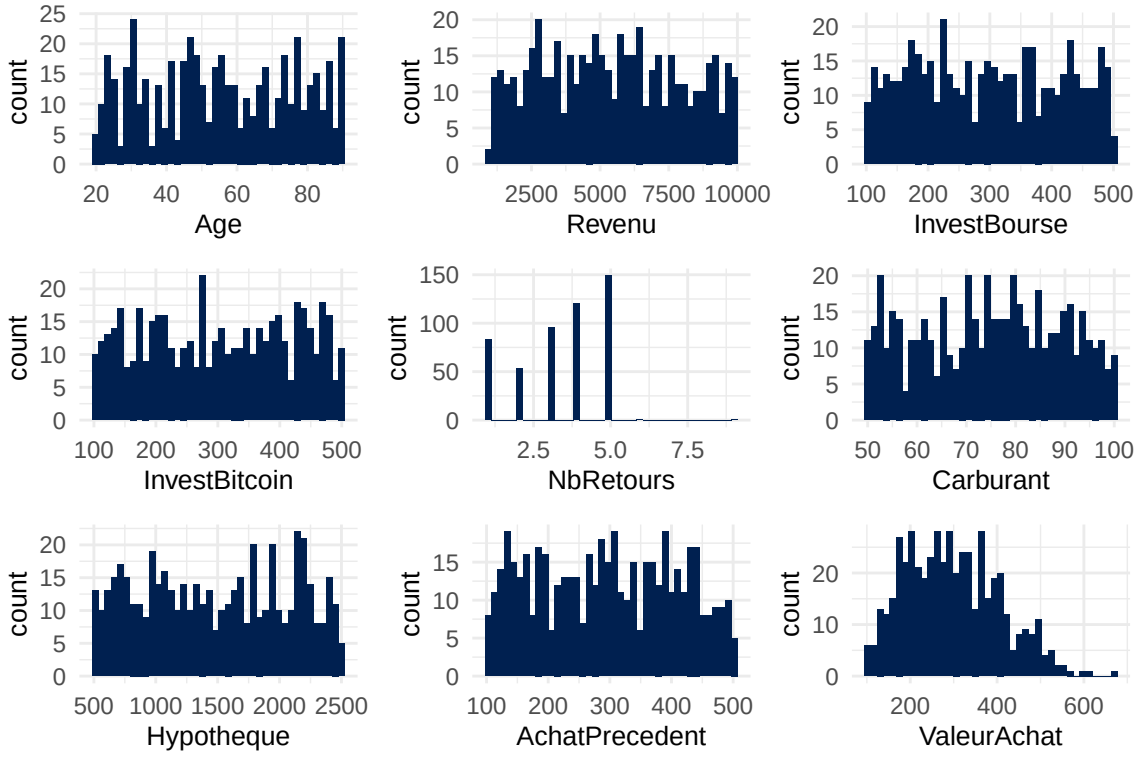
	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Age	1	503	55.590457	20.61522	55	55.600496	26.6868	20	90	70	0.0062127	-1.2210253	0.9191873
Revenu	2	503	5413.793241	2528.94993	5348	5390.488834	3215.7594	1016	9997	8981	0.0729643	-1.1135128	112.7603040
InvestBourse	3	503	297.632207	116.47392	294	296.863524	151.2252	101	500	399	0.0629273	-1.2356838	5.1933153
InvestBitcoin	4	503	302.805169	117.75076	303	303.404467	151.2252	101	499	398	-0.0333773	-1.2509754	5.2502468
NbRetours	5	503	3.413519	1.45311	4	3.503722	1.4826	1	9	8	-0.3320980	-0.7671367	0.0647910
Carburant	6	503	74.805169	14.31802	76	74.898263	17.7912	50	100	50	-0.0814283	-1.1019508	0.6384091
Hypotheque	7	503	1486.258449	585.49045	1472	1485.280397	747.2304	504	2496	1992	0.0031317	-1.2732056	26.1057291
AchatPrecedent	8	503	294.864811	113.56758	296	294.133995	146.7774	100	500	400	0.0143563	-1.1820727	5.0637281
ValeurAchat	9	503	297.357853	107.24418	288	292.044665	120.0906	104	672	568	0.4138706	-0.3717551	4.7817816

Nous remarquons que le premier résumé contient deux lignes, **M** et **Masculin**, qui correspondent à la même information, soit le genre masculin. Ce "doublon" a pour conséquence d'influencer les statistiques appliquées sur les autres variables. Nous souhaitons les harmoniser par la suite. Voici le nouveau résumé, nous observons que la moyenne est désormais identique dans les deux résumés.

Tableau 4 : Résumé après modification

Genre	Age	Revenu	InvestBourse	InvestBitcoin	NbRetours	Carburant	Hypotheque	AchatPrecedent	ValeurAchat
M	55.59046	5413.793	297.6322	302.8052	3.413519	74.80517	1486.258	294.8648	297.3579

Le second tableau présente différentes statistiques appliquées sur les variables numériques (nous excluons spécifiquement les variables catégorielles $Genre$, $AmznPrime$ et $Fidelite$). À première vue, il est possible de voir que la majorité des variables sont relativement dispersées, comme en témoignent les écarts-types de chacune. De même, un bon nombre présentent une légère asymétrie vers la gauche ou vers la droite lorsqu'on observe la colonne $skew$ du tableau, ainsi qu'une distribution aplatie ($kurtosis < 1$). Ceci est davantage visible lorsqu'on observe les histogrammes ci-dessous :



Ci-dessous est présenté la matrice de corrélation pour chacune des variables. Le gradient de couleur indique l'intensité de la corrélation. Nous remarquons quelques corrélations modérées entre $Revenu$ et $ValeurAchat$, mais également entre $Fidelite$ et $NbRetours$.

Tableau 5 : Matrice de corrélation

	Age	Revenu	InvestBourse	InvestBitcoin	NbRetours	Carburant	Hypotheque	AchatPrecedent	ValeurAchat	AmznPrime	Fidelite
Age	1.00	0.04	-0.05	0.05	-0.04	-0.09	0.05	0.05	-0.02	0.01	-0.07
Revenu	0.04	1.00	0.10	0.07	0.09	-0.01	-0.10	0.00	0.60	0.07	0.12
InvestBourse	-0.05	0.10	1.00	0.06	0.11	0.02	-0.03	0.07	0.06	0.02	0.03
InvestBitcoin	0.05	0.07	0.06	1.00	0.04	-0.01	-0.04	0.04	-0.29	0.01	0.01
NbRetours	-0.04	0.09	0.11	0.04	1.00	-0.08	-0.02	0.07	0.00	-0.12	0.50
Carburant	-0.09	-0.01	0.02	-0.01	-0.08	1.00	-0.05	-0.03	0.22	0.00	-0.05
Hypotheque	0.05	-0.10	-0.03	-0.04	-0.02	-0.05	1.00	0.06	-0.19	-0.07	-0.09
AchatPrecedent	0.05	0.00	0.07	0.04	0.07	-0.03	0.06	1.00	0.06	-0.08	0.05
ValeurAchat	-0.02	0.60	0.06	-0.29	0.00	0.22	-0.19	0.06	1.00	0.42	0.13
AmznPrime	0.01	0.07	0.02	0.01	-0.12	0.00	-0.07	-0.08	0.42	1.00	-0.05
Fidelite	-0.07	0.12	0.03	0.01	0.50	-0.05	-0.09	0.05	0.13	-0.05	1.00

Question 2

En résumé, les données présentent des distributions légèrement asymétriques, sauf dans le cas de $ValeurAchat$, qui est asymétrique à droite (skewness > 0). De plus, les variables sont fortement aplaties, ce qui est notamment une conséquence des valeurs d'écarts-types observées élevées. D'autre part, certaines variables sont modérément corrélées entre elles, ce qui montre une influence réciproque.

Ainsi, il semblerait par exemple que si le revenu du client augmente, la valeur de son achat aussi et vice-versa (corrélation positive). De même, l'abonnement au programme de fidélité semble influencer la fréquence à laquelle les clients retournent leurs achats.

L'abonnement à Amazon Prime semble également influencer la valeur de l'achat, tout comme le montant payé pour du carburant. Toutefois, l'interaction entre ces deux dernières est plus faible. En revanche, nous observons que la plupart des variables ont une influence très faible sur l'investissement en Bitcoin, puisque les coefficients de corrélation sont proches de zéro.

Question 3

Voici les pourcentages des clients possédant ou non un abonnement à Amazon Prime, ainsi qu'une carte de fidélité, ainsi que les diagrammes en boites des différentes variables numériques :

Tableau 6 : Tableau croisé des pourcentages de clients adhérant aux abonnements ou non

AmznPrime	Fidelite		
	No	**Yes**	**Total**
No	22.86	32.01	54.87
Yes	21.27	23.86	45.13
Total	44.13	55.87	100.00

Les résultats montrent qu'une plus grande proportion de clients adhèrent au programme de fidélité, soit **55.87%**, indépendamment de leur adhésion à Amazon Prime. D'autre part, peu de clients ne sont uniquement abonnés à ce dernier (soit **21.27%**).

Tableau 7 : Influence des variables AmznPrime et Fidelite

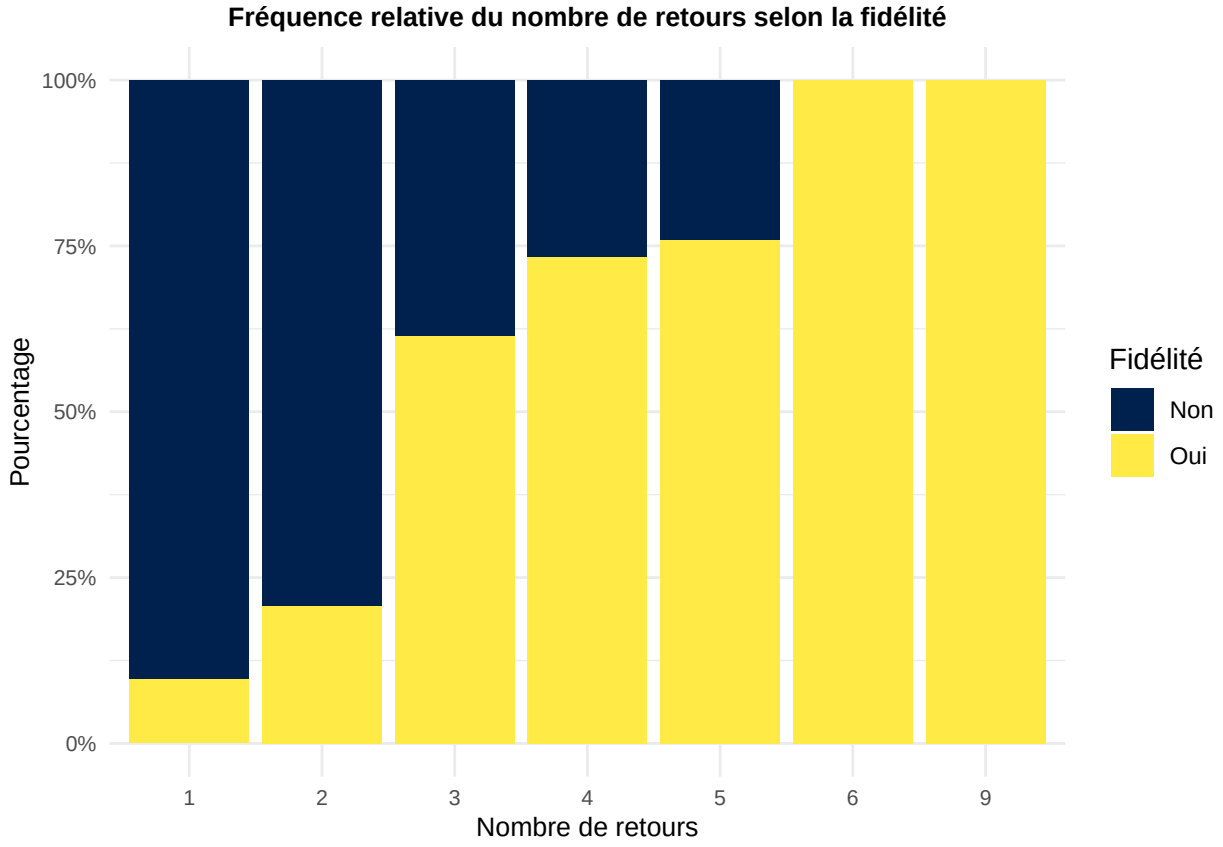
	AmznPrime	Fidelite
AmznPrime	1.00	-0.05
Fidelite	-0.05	1.00

Si nous reprenons la matrice de corrélation pour les variables $AmznPrime$ et $Fidelite$ uniquement, celle-ci nous indique que la corrélation entre les variables est de **-0.055**, ce qui correspond à une interaction négative très faible. Les deux variables sont pratiquement indépendantes l'une de l'autre, ce qui ce confirme avec le test de corrélation de Pearson, pour lequel H_0 est l'hypothèse sous laquelle les coefficients de corrélation sont égaux à zéro et H_1 l'hypothèse alternative sous laquelle les coefficients sont différents de zéro :

```
    Pearson's product-moment correlation

data:  df_cor$AmznPrime and df_cor$Fidelite
t = -1.2288, df = 501, p-value = 0.2197
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.14156721  0.03276829
sample estimates:
        cor
-0.05481722
```

Les résultats du test indiquent une $p - value$ de $0.219715 > 0.05$, nous rejettons donc H_0. Les coefficients sont différents de zéro, toutefois, la corrélation demeure faible.

Question 4



Il semblerait que les retours soient plus fréquents lorsque le client n'est pas abonné au programme de fidélité de l'entreprise. Par exemple, près de 75% des clients ayant effectué 5 retours étaient abonnés au programme de fidélité de l'entreprise. Au-delà de 5, tous les clients y sont abonnés.

Question 5

Ici, l'intervalle de confiance à 95% pour la proportion π des clients ayant l'adhésion au programme de fidélité est de $[0.514; 0.602]$. Nous l'obtenons suite à l'inversion de la table de proportions.

Question 6

Nous souhaitons comparer la moyenne des valeurs d'achat des clients qui ont l'adhésion au programme « fidélité » avec celle des clients qui n'ont pas cette adhésion. Soit :

- H_0 l'hypothèse sous laquelle les moyennes des deux groupes ne sont pas significativement différentes.
- H_1 l'hypothèse sous laquelle les moyennes des deux groupes sont significativement différentes.

```
    Welch Two Sample t-test

data:  df$ValeurAchat[df$Fidelite == "Yes"] and df$ValeurAchat[df$Fidelite == "No"]
t = 2.8544, df = 455.08, p-value = 0.004508
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  8.584346 46.526872
sample estimates:
mean of x mean of y
 309.5196  281.9640
```

Le test renvoie une $p - value$ de $0.0045084 < 0.05$ à l'intervalle de confiance de 95%. Nous acceptons donc l'hypothèse H_0.

Question 7

a. Modèle de régression linéaire

Il s'agit ici de construire un modèle de régression linéaire multiple, avec $ValeurAchat$ étant la variable dépendante. Nous décidons d'omettre la variable $Genre$ car celle-ci n'est composée du genre masculin uniquement et n'aura pas d'effet sur la performance du modèle.

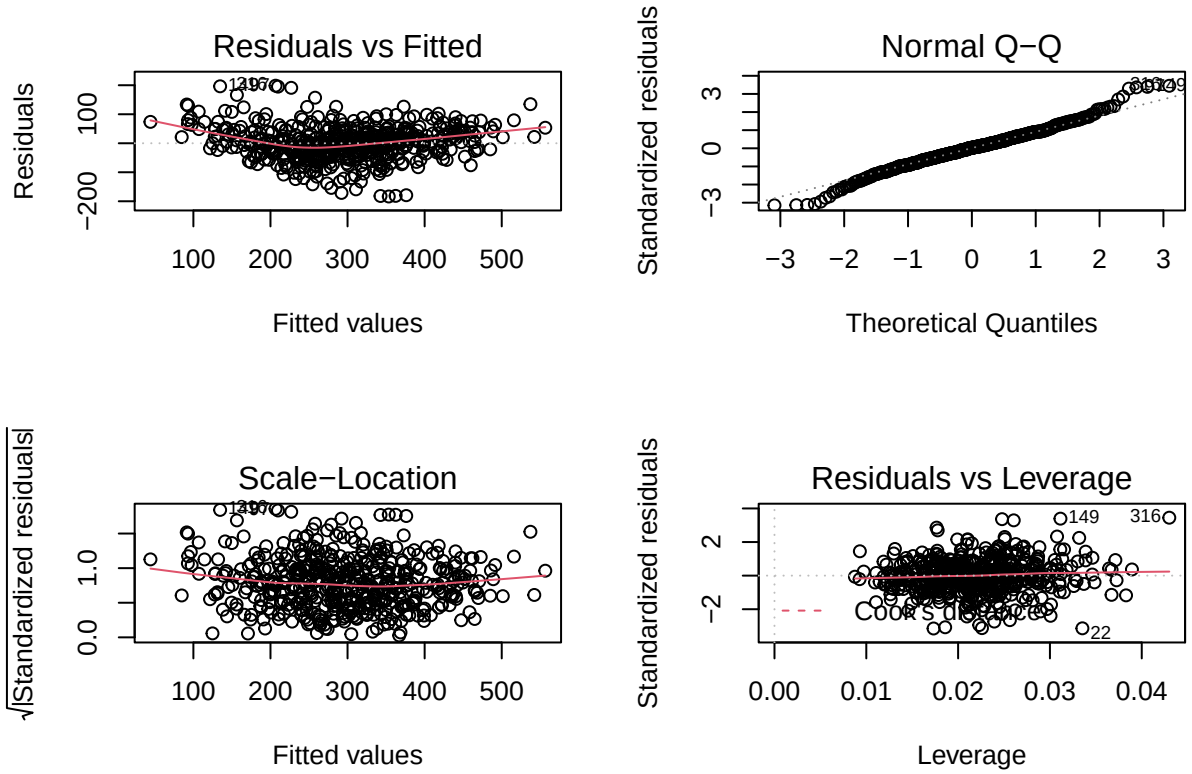
b. Discussion des résultats

```
Call:
lm(formula = ValeurAchat ~ ., data = subset(df_reg, select = -Genre))

Residuals:
     Min       1Q   Median       3Q      Max
-183.472  -34.324    1.939   35.964  198.393

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    83.117064  22.715547   3.659 0.000281 ***
Age            -0.024126   0.128959  -0.187 0.851675
Revenu          0.024799   0.001060  23.390  < 2e-16 ***
InvestBourse   -0.002972   0.022901  -0.130 0.896791
InvestBitcoin  -0.308104   0.022454 -13.722  < 2e-16 ***
NbRetours      -2.034401   2.113325  -0.963 0.336194
Carburant       1.716845   0.185026   9.279  < 2e-16 ***
Hypotheque     -0.018949   0.004558  -4.157 3.80e-05 ***
AchatPrecedent  0.111744   0.023389   4.778 2.35e-06 ***
AmznPrime      84.832157   5.351113  15.853  < 2e-16 ***
```

```
Fidelite       19.605370   6.147484   3.189 0.001518 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 58.81 on 492 degrees of freedom
Multiple R-squared:  0.7053,    Adjusted R-squared:  0.6993
F-statistic: 117.7 on 10 and 492 DF,  p-value: < 2.2e-16
```

Nous remarquons que seules les variables $Revenu$, $InvestBitcoin$, $Carburant$, $Hypotheque$, $AchatPrecedent$, ainsi que les abonnements au programme de fidélité et au service Amazon Prime semblent être significatifs, puisque leur $p-value$ est inférieure au seuil de confiance de 0.05. D'autre part, la $p-value$ de $1.2059925 \times 10^{-123} < 2.2e^{-16} < 0.05$ et le R^2 ajusté est de **0.6992639**. Le modèle semble pertinent mais les prédicteurs n'expliquent qu'à 70% la valeur du prochain achat.

c. Diagnostic du modèle



Les graphiques ci-dessus nous montrent que l'homoscédasticité est violée : en effet, le nuage de points sur le graphique des résidus en foncttions des valeurs ajustées n'est pas uniforme et prend la forme d'un entonnoir. Cela montre que les variances des résidus ne sont pas constantes. En revanche, la normalité des résidus est relativement respectée. Enfin, les valeurs 149 et 316 semblent être aberrantes.

Question 8

a. Modèle de régression linéaire

Il s'agit ici de construire un modèle de régression linéaire multiple, avec $InvestBitcoin$ étant la variable dépendante. Nous décidons d'omettre la variable $Genre$ car celle-ci n'est composée du genre masculin uniquement et n'aura pas d'effet sur la performance du modèle.

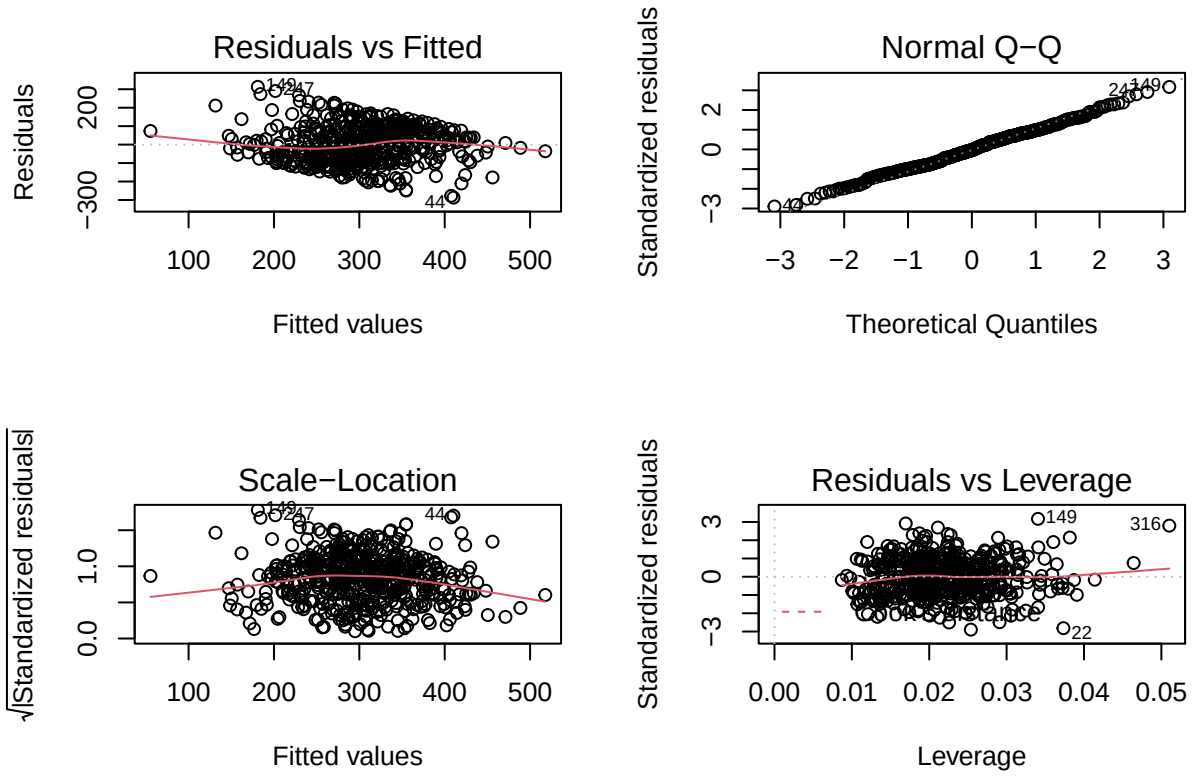
b. Discussion des résultats

```
Call:
```

```
lm(formula = InvestBitcoin ~ ., data = subset(df_reg, select = -Genre))

Residuals:
     Min       1Q   Median       3Q      Max
-287.269  -76.705   -2.988   70.430  312.912

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    253.884524  37.608299   6.751 4.15e-11 ***
Age              0.197382   0.220029   0.897  0.37012
Revenu           0.024222   0.002394  10.119  < 2e-16 ***
InvestBourse     0.032567   0.039076   0.833  0.40502
NbRetours        0.533851   3.611866   0.148  0.88256
Carburant        1.521378   0.335527   4.534 7.27e-06 ***
Hypotheque      -0.022068   0.007856  -2.809  0.00517 **
AchatPrecedent   0.130563   0.040428   3.230  0.00132 **
ValeurAchat     -0.898318   0.065467 -13.722  < 2e-16 ***
AmznPrime       78.403466  10.660222   7.355 8.05e-13 ***
Fidelite        13.486200  10.587467   1.274  0.20334
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 100.4 on 492 degrees of freedom
Multiple R-squared:  0.2871,    Adjusted R-squared:  0.2727
F-statistic: 19.82 on 10 and 492 DF,  p-value: < 2.2e-16
```

Nous remarquons que seules les variables $Revenu$, $Carburant$, $Hypotheque$, $AchatPrecedent$, $ValeurAchat$, ainsi que l'abonnement au service Amazon Prime semblent être significatifs, puisque leur $p-value$ est inférieure ou proche du seuil de confiance de 0.05. D'autre part, la $p-value$ de $7.7457953 \times 10^{-31} < 2.2e^{-16} < 0.05$ et le R^2 ajusté est de **0.2726585**. Le modèle semble pertinent, mais performe mal ; les prédicteurs n'expliquent qu'à 27% l'investissement des clients en Bitcoin.

c. Diagnostic du modèle



Les graphiques ci-dessus nous montrent que la normalité des résidus est respectée, cependant, l'homoscédasticité est violée : le nuage de points sur le graphique des résidus en fonction des valeurs ajustées n'est pas uniforme et courbe. Cela montre que les variances des résidus ne sont pas constantes. Enfin, les valeurs 22, 44, 149 et 316 semblent aberrantes, ce qui affecte la qualité du modèle.

Annexe

```
# INSTALLATION DE PACKAGES COMPLEMENTAIRES

packages<-function(x){
  x<-as.character(match.call()[[2]])
  if (!require(x,character.only=TRUE)){
    install.packages(pkgs=x,repos="http://cran.r-project.org")
    require(x,character.only=TRUE)
  }
}
packages(kableExtra)
packages(leaps)
packages(ggplot2)
packages(tidyverse)
packages(rmarkdown)
packages(viridis)
packages(corrplot)
packages(egg)
packages(psych)

# kableExtra est un outil permettant de gérer l'affichage des tableaux dans le rapport
# corrplot est un outil permettant de facilement visualiser les matrices de corrélation
# viridis est un outil de gestion des couleurs dans les visules de type ggplot
# la fonction ggarrange() d'egg est un outil permettant de combiner plusieurs objets de type ggplot
```

```r
# psych est un outil offrant davantage d'outils statistiques

# QUESTION 1

df = read.csv("Projet-Mi-session-DataSet.csv", sep = ";", header = TRUE)

summary_1 = df %>%
  select_all() %>%
  group_by(Genre) %>%
  summarize_if(is.numeric, .funs = mean)

summary_2 = df %>%
  select(-Genre, -AmznPrime, -Fidelite) %>%
  describe()

# On cherche le numéro de la ligne pour laquelle la variable Genre
# est égale à "Masculin" et on la remplace par "M"

which(df$Genre == "Masculin")
df$Genre[381] = "M"

# Mise à jour du résumé

table_cont_new = df %>%
  select(-AmznPrime, -Fidelite) %>%
  group_by(Genre) %>%
  summarize_all(., .funs = mean)

# Pourcentages des variables booléennes

perc_df = df %>%
  select(AmznPrime, Fidelite) %>%
  table(dnn = c("AmznPrime", "Fidelite")) %>%
  prop.table() %>%
  round(., 4)*100

perc_df = addmargins(perc_df, FUN = sum)
rownames(perc_df)[3] = "Total"
colnames(perc_df)[3] = "Total"

# On affiche les distributions de toutes les variables sauf celles booléennes

# Boxplots

plots = list()

for(i in names(df[,2:10])){
plots[[i]] = ggplot(df) +
  geom_boxplot(aes_string(y = i), position = position_dodge(width = .60), show.legend = "none") +
          theme_minimal() +
  theme(axis.title = element_text(size = 10))
}

# Distributions

plots_1 = list()
```

```r
for(i in names(df[,2:10])){
plots_1[[i]] = ggplot(df) +
  geom_density(aes_string(x = i)) +
          theme_minimal() +
  theme(axis.title = element_text(size = 10))
}

# ggarrange(plots[[1]], plots[[2]], plots[[3]], plots[[4]], plots[[5]],
#           plots[[6]], plots[[7]], plots[[8]], plots[[9]])

ggarrange(plots_1[[1]], plots_1[[2]], plots_1[[3]], plots_1[[4]], plots_1[[5]],
          plots_1[[6]], plots_1[[7]], plots_1[[8]], plots_1[[9]])

# Corrélation

df_cor = df

df_cor$AmznPrime = ifelse(df$AmznPrime == "Yes", 1, 0)
df_cor$Fidelite = ifelse(df$Fidelite == "Yes", 1, 0)

# Création d'une matrice de corrélation

mat_1 <- round(cor(df_cor[,2:12]),2)

# Affichage de la matrice sous forme d'un corrélogramme

# corrplot(mat_1, method = "circle")

df_mat = data.frame(mat_1)

# QUESTION 3

# Test de Pearson

test1 = cor.test(df_cor$AmznPrime,
                 df_cor$Fidelite)
test1

# QUESTION 4

# Graphique à barres empilées égales

ggplot(df) +
  geom_bar(aes(as.factor(NbRetours), fill = as.factor(Fidelite)), position = "fill") +
  scale_fill_viridis(discrete = TRUE, labels = c("Oui", "Non"), option = "E") +
  scale_y_continuous("Pourcentage", labels = scales::percent) +
  theme_minimal() +
  ggtitle("Fréquence relative du nombre de retours selon la fidélité") +
  theme(plot.title = element_text(family = 'Helvetica', face = 'bold',
                                  hjust = 0.5, size = 10),
        axis.title.x = element_text(size = 10),
        axis.title.y = element_text(size = 10),
        axis.text.y = element_text(size = 8),
        axis.text.x = element_text(size = 8),
        legend.position = "right") +
```

```r
  labs(x = "Nombre de retours", fill = "Fidélité")

# QUESTION 5

# Intervalle de confiance

prop.test(table(df$Fidelite), conf.level = 0.95)$"conf.int"

# Inversion de la table de proportions

table_inv = df$Fidelite %>%
  factor(., levels = c("Yes", "No")) %>%
  table()

prop_test_int = prop.test(table_inv, alternative = "two.sided", conf.level = 0.95)$"conf.int"

# QUESTION 6

# T.test

welch_test = t.test(df$ValeurAchat[df$Fidelite == "Yes"],
       df$ValeurAchat[df$Fidelite == "No"],
       alternative='two.sided',
       conf.level = 0.95)
welch_test

# QUESTION 7

# Transformation des variables catégorielles

df_reg = df

df_reg$AmznPrime = ifelse(df_reg$AmznPrime == "Yes", 1, 0)
df_reg$Fidelite = ifelse(df_reg$Fidelite == "Yes", 1, 0)

# Modèle entier

reg_model = lm(ValeurAchat ~ ., data = subset(df_reg, select = -Genre))
summary(reg_model)

# Plots

par(mfrow = c(2,2))
plot(reg_model)

# QUESTION 8

# Modèle entier

reg_model_btc = lm(InvestBitcoin ~ ., data = subset(df_reg, select = -Genre))
summary(reg_model_btc)

# Plots

par(mfrow = c(2,2))
plot(reg_model_btc)
```