Pré-analyse des données de cartes de crédit

Boujamaa Atrmouh - Christophe WANG

20 novembre 2023

Contents

Introduction

Source des données

Les données utilisées dans cette analyse proviennent du jeu de données “Credit Card Customers” disponible sur Kaggle à l’adresse suivante : Credit Card Customers Dataset.

Outils Utilisés

- Langage de programmation : R
- Bibliothèques : ggplot2, skimr, VIM, outliers

Objectifs de l’analyse

L’objectif principal de cette analyse est de comprendre les caractéristiques des clients de cartes de crédit et d’identifier des tendances ou des facteurs qui pourraient être liés à la résiliation de services de carte de crédit. Nous allons commencer par explorer les données, détecter les éventuelles valeurs manquantes ou aberrantes, et définir des problématiques spécifiques pour guider notre analyse.

Exploration des Données

Afficher les premières lignes du dataframe

Ce segment de la présentation offre un aperçu initial de nos données à travers trois sections du dataframe `tab`. Chaque section représente un échantillon, permettant une visualisation succincte des premières observations. Cette approche préliminaire facilite la compréhension initiale de la structure et de la distribution des données dans le cadre de notre analyse.

Les deux dernières variables ne sont pas importantes pour le moment, elles représentent les caractéristiques utilisées par le modèle Naive Bayes pour la classification.

```
head(tab[1:21])
```

```
##   CLIENTNUM    Attrition_Flag Customer_Age Gender Dependent_count
## 1 768805383 Existing Customer           45      M               3
## 2 818770008 Existing Customer           49      F               5
## 3 713982108 Existing Customer           51      M               3
## 4 769911858 Existing Customer           40      F               4
## 5 709106358 Existing Customer           40      M               3
## 6 713061558 Existing Customer           44      M               2
##   Education_Level Marital_Status Income_Category Card_Category Months_on_book
## 1     High School        Married     $60K - $80K          Blue             39
## 2        Graduate         Single  Less than $40K          Blue             44
## 3        Graduate        Married    $80K - $120K          Blue             36
## 4     High School        Unknown  Less than $40K          Blue             34
## 5      Uneducated        Married     $60K - $80K          Blue             21
## 6        Graduate        Married     $40K - $60K          Blue             36
##   Total_Relationship_Count Months_Inactive_12_mon Contacts_Count_12_mon
## 1                        5                      1                     3
## 2                        6                      1                     2
## 3                        4                      1                     0
## 4                        3                      4                     1
## 5                        5                      1                     0
## 6                        3                      1                     2
##   Credit_Limit Total_Revolving_Bal Avg_Open_To_Buy Total_Amt_Chng_Q4_Q1
## 1        12691                777           11914                1.335
## 2         8256                864            7392                1.541
## 3         3418                  0            3418                2.594
## 4         3313               2517             796                1.405
## 5         4716                  0            4716                2.175
## 6         4010               1247            2763                1.376
##   Total_Trans_Amt Total_Trans_Ct Total_Ct_Chng_Q4_Q1 Avg_Utilization_Ratio
## 1            1144             42                1.625                 0.061
## 2            1291             33                3.714                 0.105
## 3            1887             20                2.333                 0.000
## 4            1171             20                2.333                 0.760
## 5             816             28                2.500                 0.000
## 6            1088             24                0.846                 0.311
```

Information sur les types de données et les valeurs manquantes

Un dépassement se produit lorsque j'affiche les deux dernières variables, ce qui nous empêche de l'afficher, mais ce sont des types `num`.

```r
str(tab[1:21])
```

```
## 'data.frame':    10127 obs. of  21 variables:
##  $ CLIENTNUM               : int  768805383 818770008 713982108 769911858 709106358 713061558 810347
##  $ Attrition_Flag          : chr  "Existing Customer" "Existing Customer" "Existing Customer" "Existi
##  $ Customer_Age            : int  45 49 51 40 40 44 51 32 37 48 ...
##  $ Gender                  : chr  "M" "F" "M" "F" ...
##  $ Dependent_count         : int  3 5 3 4 3 2 4 0 3 2 ...
##  $ Education_Level         : chr  "High School" "Graduate" "Graduate" "High School" ...
##  $ Marital_Status          : chr  "Married" "Single" "Married" "Unknown" ...
##  $ Income_Category         : chr  "$60K - $80K" "Less than $40K" "$80K - $120K" "Less than $40K" ...
##  $ Card_Category           : chr  "Blue" "Blue" "Blue" "Blue" ...
##  $ Months_on_book          : int  39 44 36 34 21 36 46 27 36 36 ...
##  $ Total_Relationship_Count: int  5 6 4 3 5 3 6 2 5 6 ...
##  $ Months_Inactive_12_mon  : int  1 1 1 4 1 1 1 2 2 3 ...
##  $ Contacts_Count_12_mon   : int  3 2 0 1 0 2 3 2 0 3 ...
##  $ Credit_Limit            : num  12691 8256 3418 3313 4716 ...
##  $ Total_Revolving_Bal     : int  777 864 0 2517 0 1247 2264 1396 2517 1677 ...
##  $ Avg_Open_To_Buy         : num  11914 7392 3418 796 4716 ...
##  $ Total_Amt_Chng_Q4_Q1    : num  1.33 1.54 2.59 1.41 2.17 ...
##  $ Total_Trans_Amt         : int  1144 1291 1887 1171 816 1088 1330 1538 1350 1441 ...
##  $ Total_Trans_Ct          : int  42 33 20 20 28 24 31 36 24 32 ...
##  $ Total_Ct_Chng_Q4_Q1     : num  1.62 3.71 2.33 2.33 2.5 ...
##  $ Avg_Utilization_Ratio   : num  0.061 0.105 0 0.76 0 0.311 0.066 0.048 0.113 0.144 ...
```

Vérification de la taille et des noms de colonnes

Le tableau contient n = 23 et p = 1027, avec les individus en ligne et les variables bancaires en colonne.

```r
colnames(tab)
```

```
##  [1] "CLIENTNUM"
##  [2] "Attrition_Flag"
##  [3] "Customer_Age"
##  [4] "Gender"
##  [5] "Dependent_count"
##  [6] "Education_Level"
##  [7] "Marital_Status"
##  [8] "Income_Category"
##  [9] "Card_Category"
## [10] "Months_on_book"
## [11] "Total_Relationship_Count"
## [12] "Months_Inactive_12_mon"
## [13] "Contacts_Count_12_mon"
## [14] "Credit_Limit"
## [15] "Total_Revolving_Bal"
## [16] "Avg_Open_To_Buy"
## [17] "Total_Amt_Chng_Q4_Q1"
## [18] "Total_Trans_Amt"
## [19] "Total_Trans_Ct"
## [20] "Total_Ct_Chng_Q4_Q1"
## [21] "Avg_Utilization_Ratio"
```

```
## [22] "Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Educa
## [23] "Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Educa
```

```
dim(tab)
```

```
## [1] 10127    23
```

Classe du tableau

```
class(tab)
```

```
## [1] "data.frame"
```

DÉTECTION DES VALEURS MANQUANTES

Détecter les valeurs manquantes

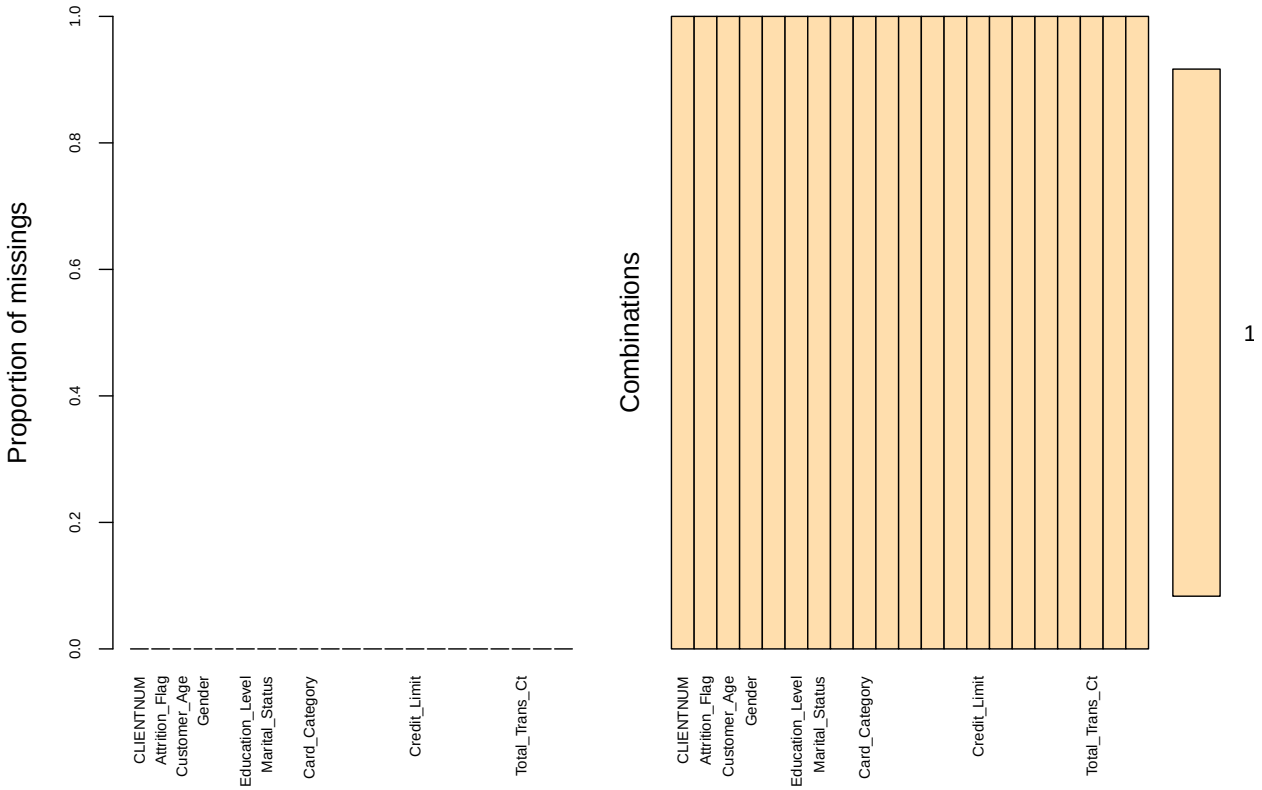
Ce script vise à détecter la présence de valeurs manquantes (NaN). Si au moins une valeur manquante est identifiée, le code affiche ces valeurs et génère un message d'arrêt indiquant à l'utilisateur de nettoyer les données avant de poursuivre son analyse.

```
if (sum(is.na(tab)) > 0) {
  print(tab[is.na(tab)])
  stop("Des valeurs NaN ont été détectées. Veuillez nettoyer les données avant de continuer.")
}
```

Visualisation des valeurs manquantes

Il est observé que notre ensemble de données ne contient aucune valeur manquante. Cette constatation est appuyée par le graphique de gauche, où aucune absence de données n'est visuellement identifiable, confirmant ainsi cette affirmation.

```
par(mfrow = c(1, 1))
aggr(tab[1:21], col = c("navajowhite1", "navajowhite3"),
     numbers = TRUE,
     sortVars = TRUE,
     labels = names(tab),
     cex.axis = 0.7,
     gap = 3,
     pch = 19)
```




```
##
##  Variables sorted by number of missings:
##                  Variable Count
##                 CLIENTNUM     0
##            Attrition_Flag     0
##              Customer_Age     0
##                    Gender     0
##           Dependent_count     0
##           Education_Level     0
##            Marital_Status     0
##           Income_Category     0
##             Card_Category     0
##            Months_on_book     0
##  Total_Relationship_Count     0
##    Months_Inactive_12_mon     0
##     Contacts_Count_12_mon     0
##              Credit_Limit     0
##       Total_Revolving_Bal     0
##           Avg_Open_To_Buy     0
##       Total_Amt_Chng_Q4_Q1     0
##           Total_Trans_Amt     0
##            Total_Trans_Ct     0
##       Total_Ct_Chng_Q4_Q1     0
##     Avg_Utilization_Ratio     0
```

DÉTECTION DES VALEURS ABERANTES

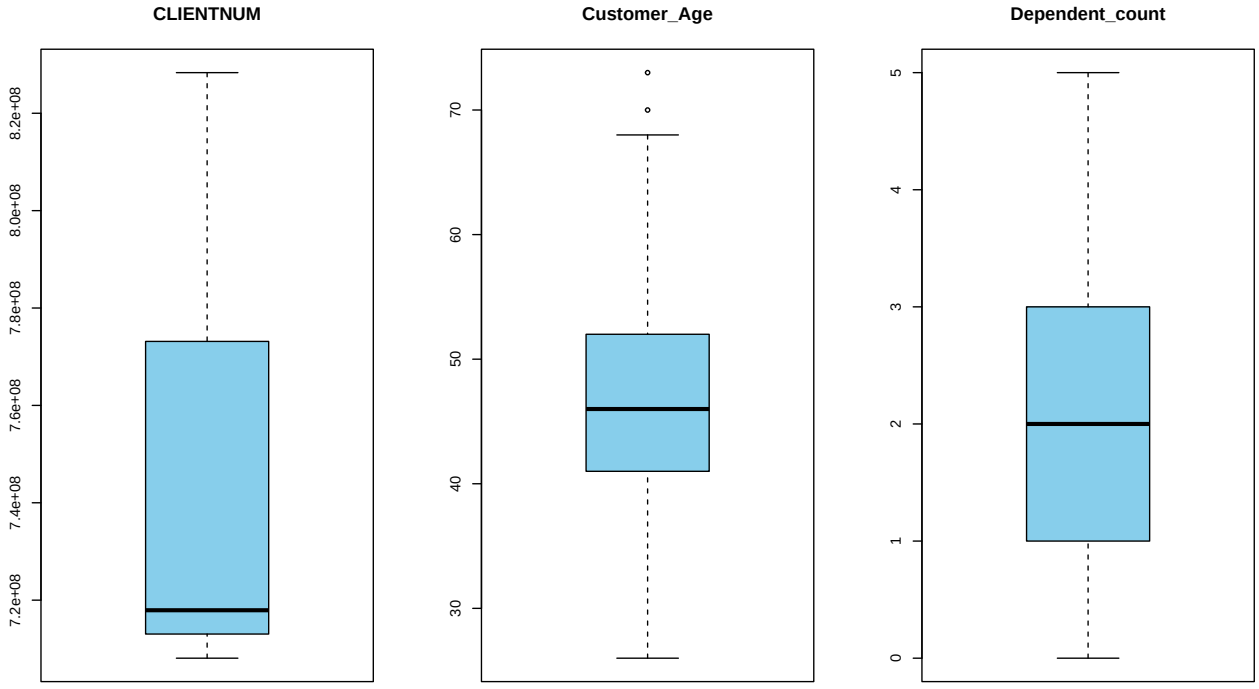
Dans cette analyse, la librairie R `outliers` est employée pour détecter d'éventuelles valeurs aberrantes. Après une évaluation approfondie, aucune valeur aberrante n'est observée initialement dans les données.

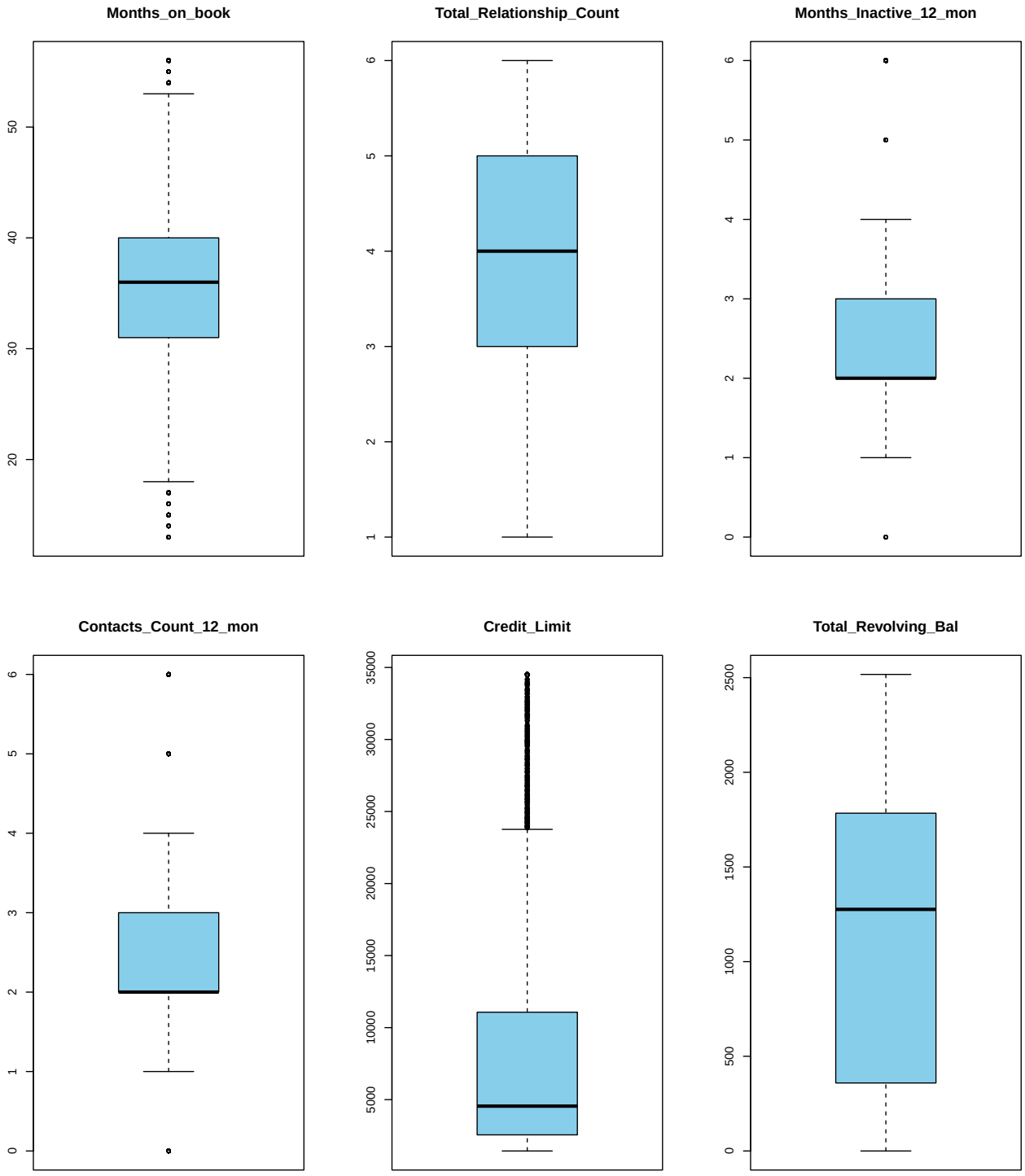
```
# Liste pour stocker les outliers détectés pour chaque colonne
all_outliers <- list()

# Détection des valeurs aberrantes pour chaque colonne
numeric_columns <- sapply(tab, is.numeric)

for (col in colnames(numeric_columns)) {
  outliers_test <- grubbs.test(tab[[col]])
  outliers_detected <- outliers_test$outliers
  all_outliers[[col]] <- outliers_detected
}

# Afficher les valeurs aberrantes détectées pour chaque colonne
print(all_outliers)
```

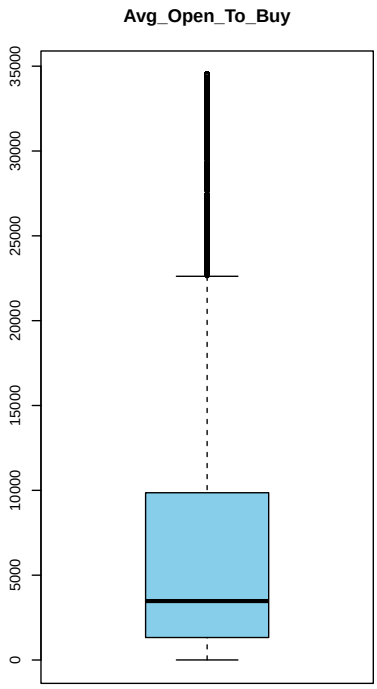
ANALYSES DES VARIABLES QUANTITATIVES ET CATÉGORIELLES

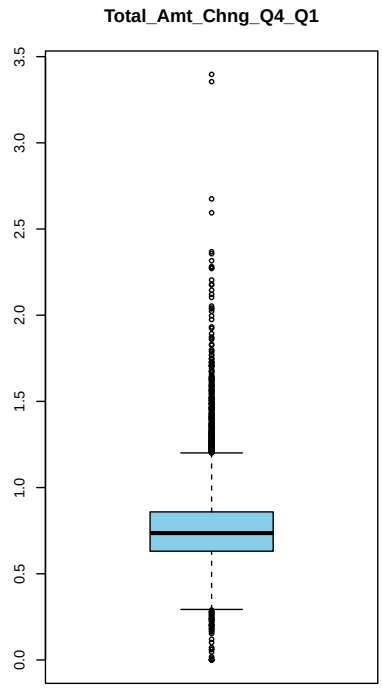
Les informations extraites à l'aide de la fonction `str(tab)` nous permettent de classifier deux types de données dans un tableau (ou un dataframe) : les variables quantitatives, qui représentent des quantités numériques telles que des nombres entiers ou à virgule flottante, et les variables catégorielles, qui représentent des catégories ou des labels tels que des chaînes de caractères. Cette classification est basée sur les types de données observés dans chaque colonne du tableau.

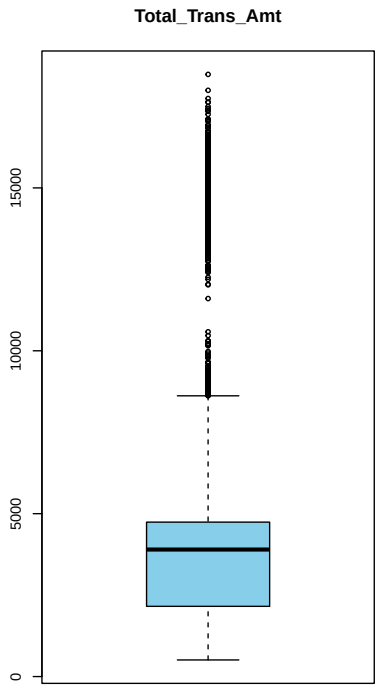
Analyse des variables quantitatives :

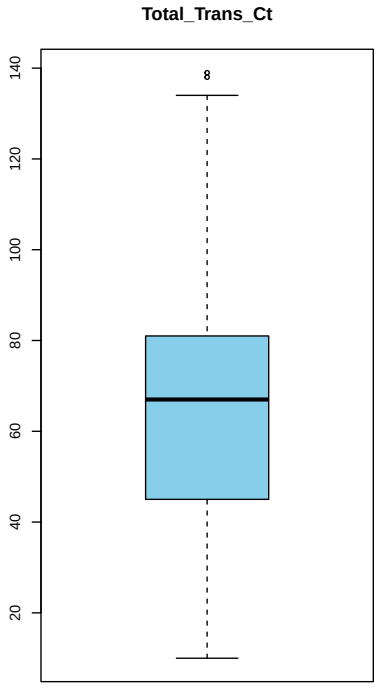


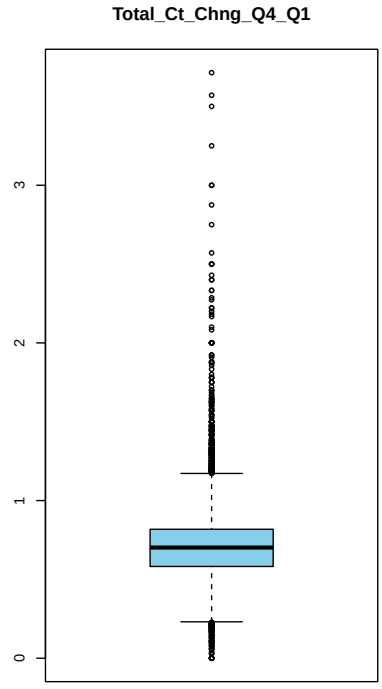

Months_on_book
Total_Relationship_Count
Months_Inactive_12_mon
Contacts_Count_12_mon
Credit_Limit
Total_Revolving_Bal

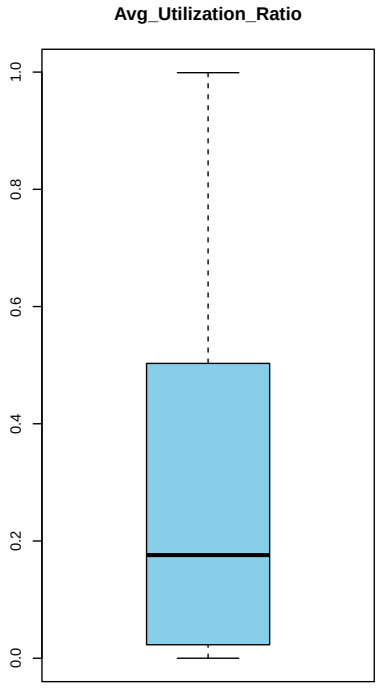

Avg_Open_To_Buy
0
5000
10000
15000
20000
25000
30000
35000


Total_Amt_Chng_Q4_Q1
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5

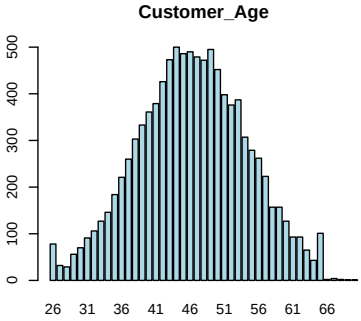

Total_Trans_Amt
0
5000
10000
15000

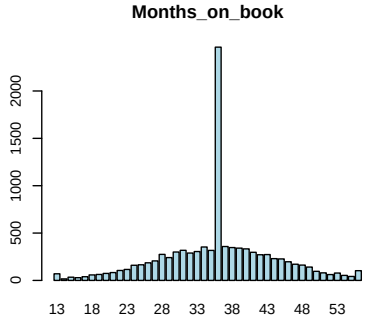

Total_Trans_Ct
20
40
60
80
100
120
140

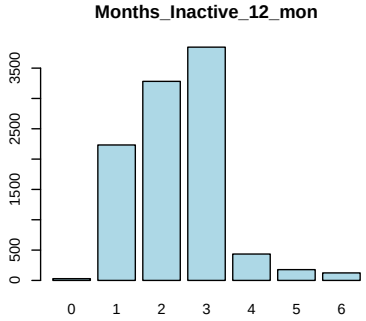

Total_Ct_Chng_Q4_Q1
0
1
2
3

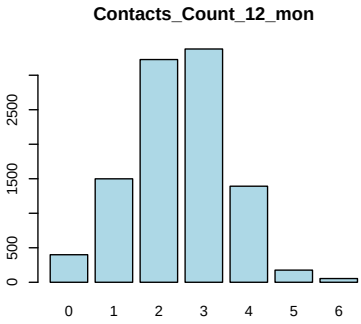

Avg_Utilization_Ratio
0.0
0.2
0.4
0.6
0.8
1.0

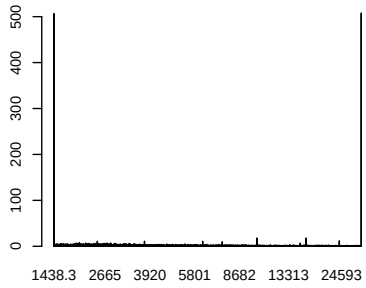
Analyse des valeurs possiblement abérantes sur le boxplot :

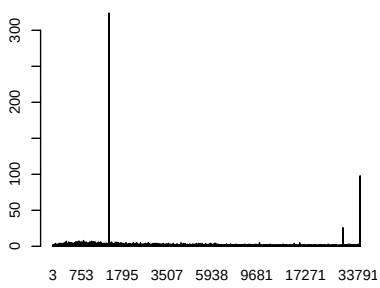


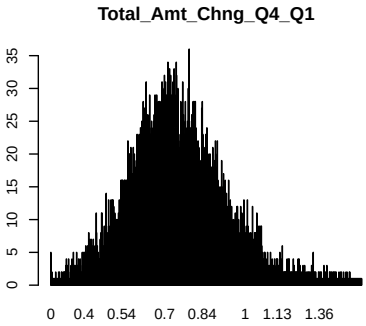


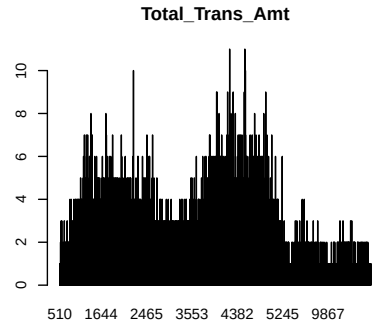


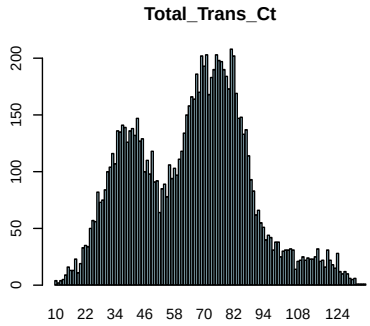


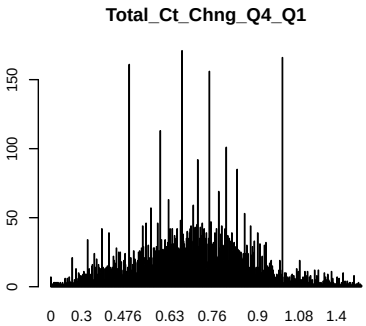






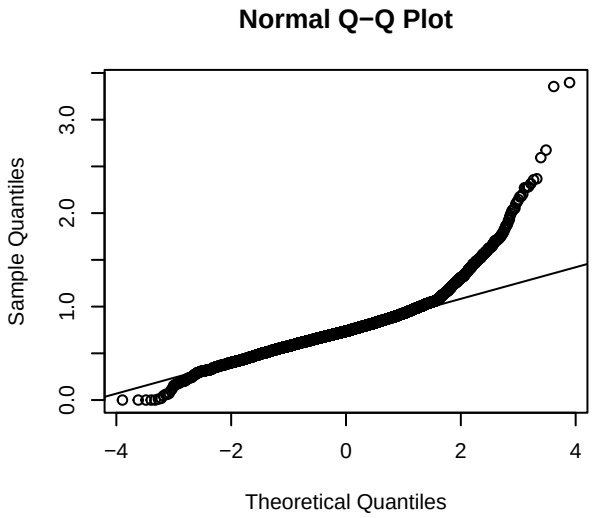






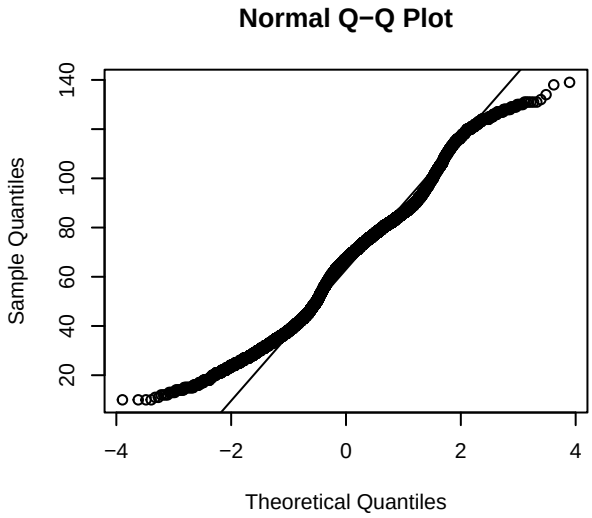
Interprétations des boxplot avec outliers :

1. **Customer_Age :** Aucun problème évident n'est détecté, bien que l'on observe un léger dépassement sur le boxplot, suggérant la nécessité d'une inspection plus approfondie.
2. **Months_on_book :** Aucun problème majeur n'est apparent, probablement lié à un ajustement de calibrage. Il est recommandé de consulter le résumé pour une confirmation supplémentaire.
3. **Months_Inactive_12_mon :** Aucun problème significatif n'est observé, probablement dû à un calibrage particulier. Une vérification du résumé peut fournir des détails supplémentaires.
4. **Contacts_Count_12_mon :** Des observations similaires sont notées, indiquant probablement un ajustement de calibrage.
5. **Credit_Limit :** Après l'analyse du barplot, les données semblent propres et homogènes.
6. **Avg_Open_To_Buy :** L'analyse du barplot révèle deux pics importants, mais la nature de la donnée ne semble pas aberrante.
7. **Total_Amt_Chng_Q4_Q1 (Valeurs Aberrantes : 3.397, 3.355) :** Malgré une impression de distribution normale d'après le barplot, l'étude des quartiles identifie deux valeurs aberrantes à l'extrême droite.



8. **Total_Trans_Amt :** Aucune valeur aberrante flagrante n'est détectée après l'analyse du barplot.

9. **Total_Trans_Ct :** Bien que le barplot suggère une distribution normale, l'étude des quartiles ne confirme pas de manière concluante la présence de valeurs aberrantes.

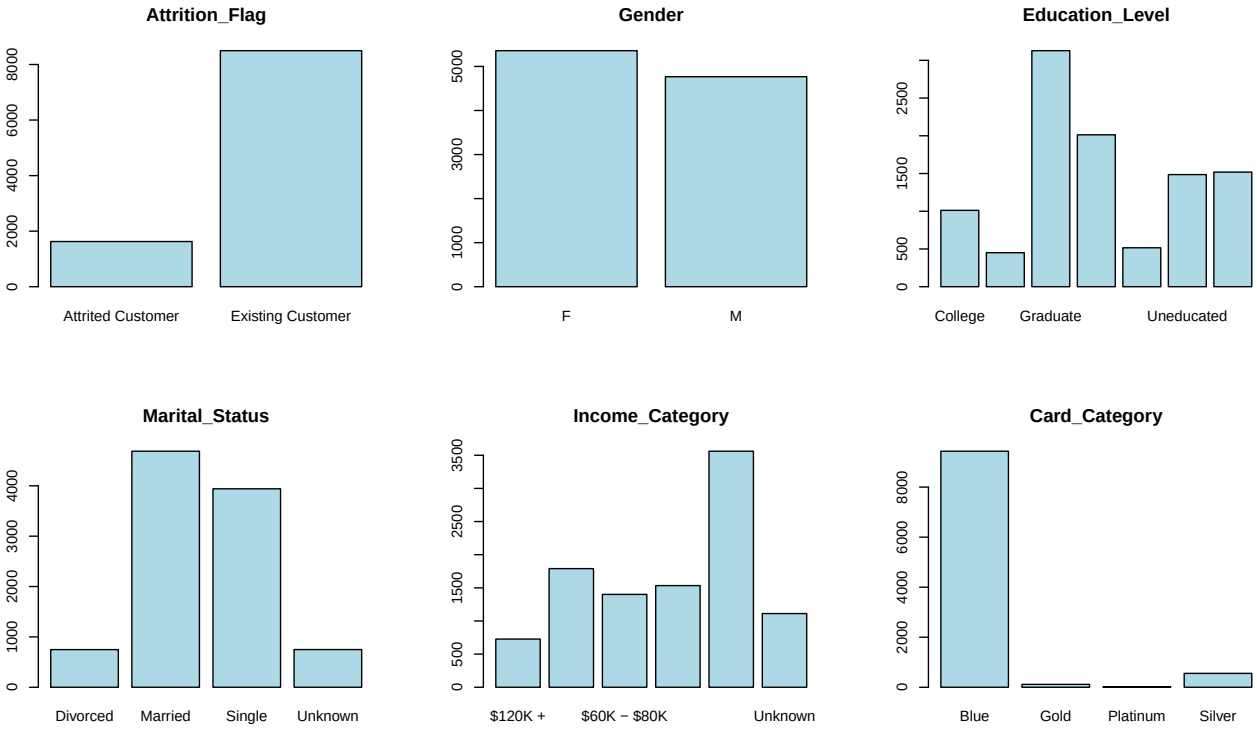


10. **Total_Ct_Chng_Q4_Q1 :** L'analyse du barplot, combinée à la nature de la variable, indique qu'il n'est pas possible d'exclure les changements dans le nombre de transactions comme des valeurs aberrantes.

Actions sur la donnée - Traitement des valeurs aberrantes :

```
# Supprimer les lignes avec les deux valeurs (3.397, 3.355)
tab <- tab[!(tab$Total_Amt_Chng_Q4_Q1 %in% c(3.397, 3.355)), ]
```

Analyse des variables catégorielles :



Interprétations :

- Les lignes contenant des valeurs 'Unknown' doivent été supprimées de l'ensemble de données, car cette catégorie ne correspond à aucune information pertinente et peut introduire des incohérences dans les analyses, affectant ainsi la qualité des résultats. Ces valeurs se retrouvent dans `Education_Level`, `Marital_Status` et `Income_Category`.
- La variable `Avg_Open_To_Buy` présente principalement des nombres entiers, mais parmi eux, un nombre décimal unique, 1438.3, a été identifié. Pour maintenir la consistance des données, il est nécessaire de convertir ce nombre décimal en un nombre entier.

Actions sur la donnée :

Traitement des valeurs inconnues :

```r
# Supprimer les lignes avec des valeurs "Unknown"
tab <- tab[rowSums(tab == "Unknown", na.rm = TRUE) == 0, ]

# Vérifier les dimensions du dataset après la suppression
tables_data_after_cleaning <- lapply(
  c("Attrition_Flag", "Gender", "Education_Level",
    "Marital_Status", "Income_Category", "Card_Category"),
                                    function(x) table(tab[[x]]))
list(tables_data_after_cleaning)
```

```
## [[1]]
## [[1]][[1]]
##
## Attrited Customer Existing Customer
##              1113              5966
##
## [[1]][[2]]
##
##    F    M
## 3375 3704
##
## [[1]][[3]]
##
##       College     Doctorate      Graduate   High School Post-Graduate
##           843           358          2591          1653           431
##    Uneducated
##          1203
##
## [[1]][[4]]
##
## Divorced  Married   Single
##      569     3564     2946
##
## [[1]][[5]]
##
##        $120K +    $40K - $60K    $60K - $80K   $80K - $120K Less than $40K
##            572           1412           1102           1201           2792
##
## [[1]][[6]]
##
##     Blue     Gold Platinum   Silver
##     6596       81       11      391
```

```r
# Remplacer la seule valeur float en valeur int
tab$Avg_Open_To_Buy[tab$Avg_Open_To_Buy == 1438.3] <- 1438

# Afficher les lignes avec des valeurs de Avg_Open_To_Buy égale à 1438.3
rows_greater_than_1438 <- tab[tab$Avg_Open_To_Buy == 1438.3, ]
print(rows_greater_than_1438$Avg_Open_To_Buy)
```

Correction d'une valeur décimale à entier :

```
## numeric(0)
```

```
summary(tab$Avg_Open_To_Buy)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       3    1248    3244    7323    9490   34516
```

VARIABLES DISCRETES ET CONTINUES

Significations :

Variables Discrètes :

1. **CLIENTNUM:** Identifiant unique du client - Discrète. Il s'agit d'un numéro d'identification unique attribué à chaque client.
2. **Attrition_Flag:** Variable binaire indiquant si le compte est fermé (1) ou non (0) - Discrète. C'est une variable binaire avec des valeurs distinctes.
3. **Gender:** Variable catégorielle indiquant le genre du client (M=Male, F=Female) - Discrète. Il s'agit d'une variable catégorielle avec des catégories distinctes.
4. **Dependent_count:** Nombre de personnes à charge - Discrète. C'est un nombre entier représentant le nombre de personnes à charge.
5. **Education_Level:** Niveau d'éducation du titulaire du compte - Discrète. Il s'agit d'une variable catégorielle avec des catégories distinctes.
6. **Marital_Status:** État civil du titulaire du compte - Discrète. Il s'agit d'une variable catégorielle avec des catégories distinctes.
7. **Income_Category:** Catégorie de revenu annuel du titulaire du compte - Discrète. Il s'agit d'une variable catégorielle avec des catégories distinctes.
8. **Card_Category:** Type de carte de crédit (Blue, Silver, Gold, Platinum) - Discrète. Il s'agit d'une variable catégorielle avec des catégories distinctes.
9. **Months_on_book:** Période de relation avec la banque - Discrète. Il s'agit d'un nombre entier représentant le nombre de mois.
10. **Months_Inactive_12_mon:** Nombre de mois d'inactivité au cours des 12 derniers mois - Discrète. Il s'agit d'un nombre entier.
11. **Contacts_Count_12_mon:** Nombre de contacts au cours des 12 derniers mois - Discrète. Il s'agit d'un nombre entier.

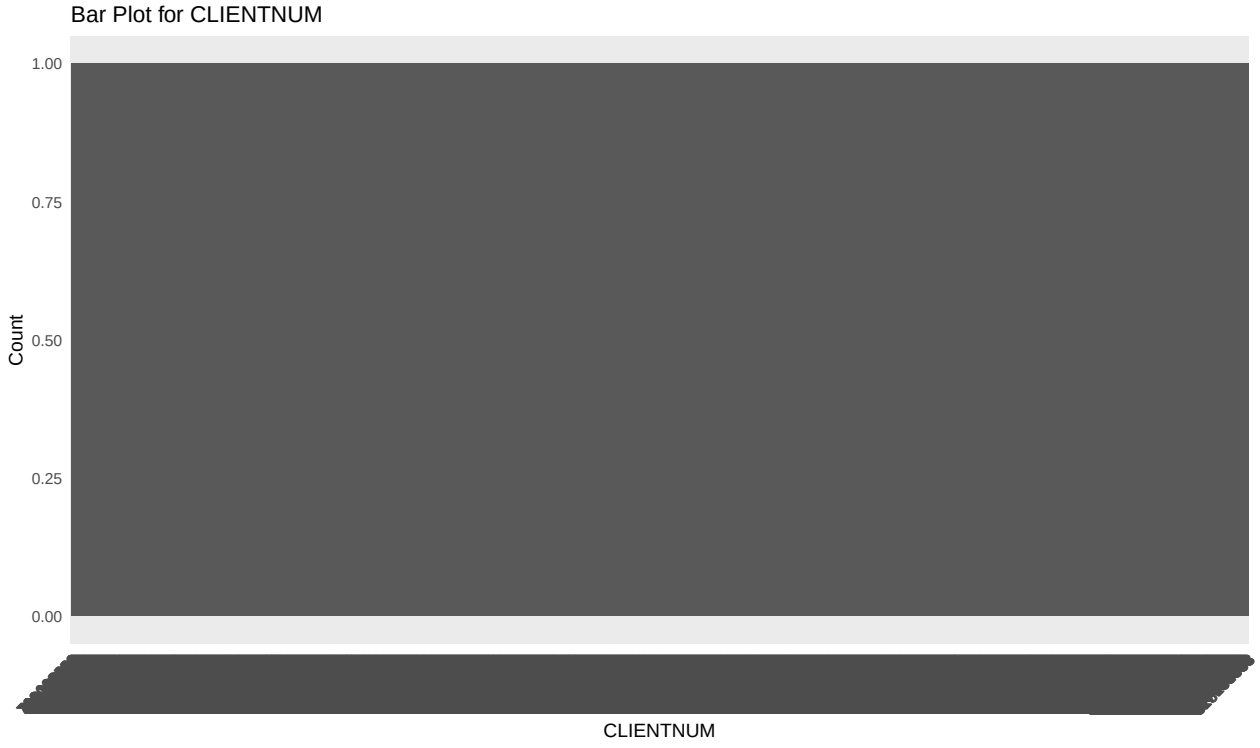
Variables Continues :

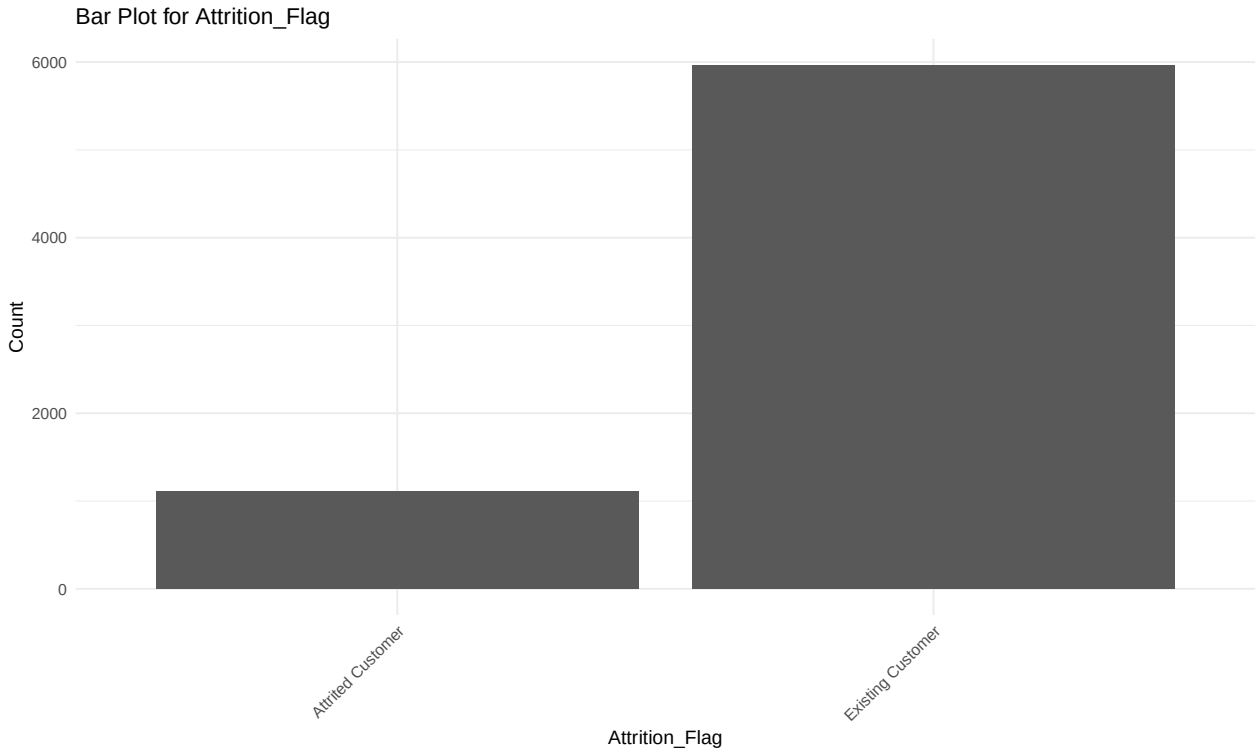
1. **Customer_Age:** Âge du client en années - Continue. L'âge est une variable continue car elle peut prendre n'importe quelle valeur dans un intervalle.
2. **Credit_Limit:** Limite de crédit sur la carte de crédit - Continue. La limite de crédit peut prendre n'importe quelle valeur dans un intervalle.
3. **Total_Relationship_Count:** Nombre total de produits détenus par le client - Continue. Le nombre total de produits peut prendre n'importe quelle valeur dans un intervalle.
4. **Total_Revolving_Bal:** Solde total en rotation sur la carte de crédit - Continue. Le solde en rotation peut prendre n'importe quelle valeur dans un intervalle.
5. **Avg_Open_To_Buy:** Montant disponible pour les achats sur la carte (moyenne des 12 derniers mois) - Continue. Cette moyenne peut prendre n'importe quelle valeur dans un intervalle.
6. **Total_Amt_Chng_Q4_Q1:** Changement du montant des transactions (Q4 sur Q1) - Continue. Le changement du montant peut prendre n'importe quelle valeur dans un intervalle.

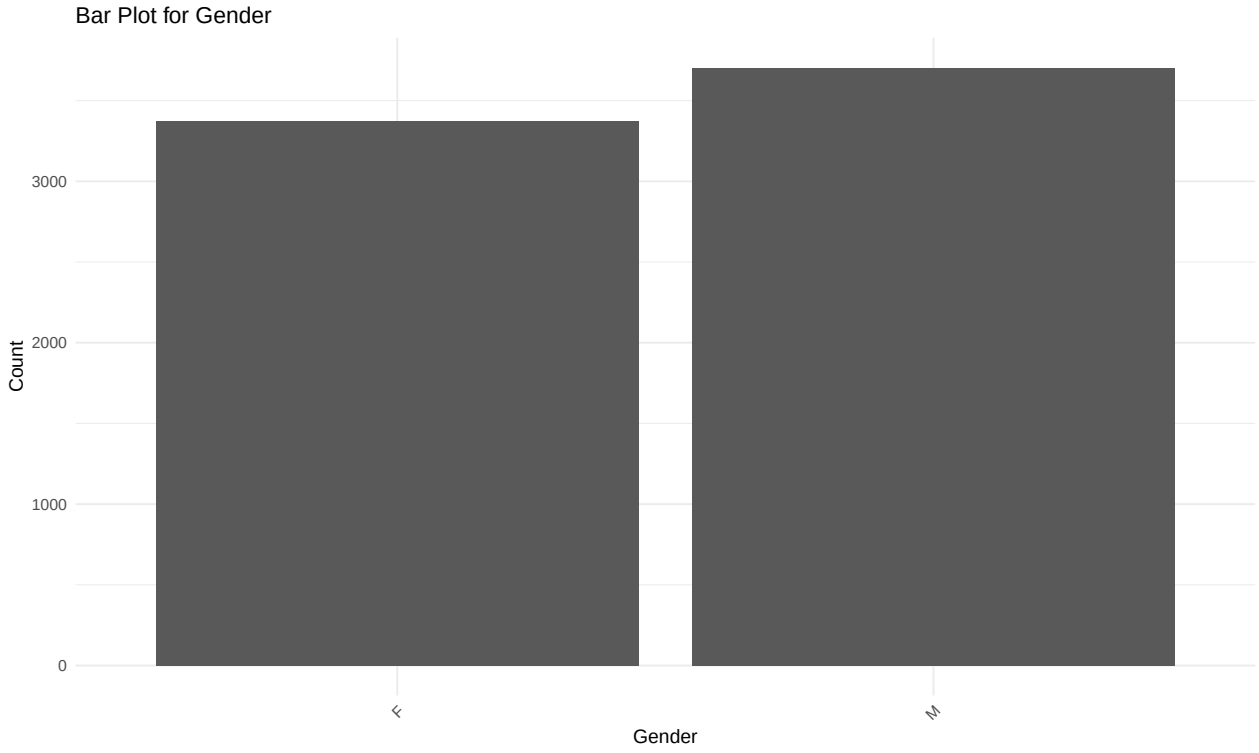
7. **Total_Trans_Amt:** Montant total des transactions (12 derniers mois) - Continue. Le montant total des transactions peut prendre n'importe quelle valeur dans un intervalle.
8. **Total_Trans_Ct:** Nombre total de transactions (12 derniers mois) - Continue. Le nombre total de transactions peut prendre n'importe quelle valeur dans un intervalle.
9. **Total_Ct_Chng_Q4_Q1:** Changement du nombre de transactions (Q4 sur Q1) - Continue. Le changement du nombre de transactions peut prendre n'importe quelle valeur dans un intervalle.
10. **Avg_Utilization_Ratio:** Ratio moyen d'utilisation de la carte de crédit - Continue. Le ratio moyen peut prendre n'importe quelle valeur dans un intervalle.

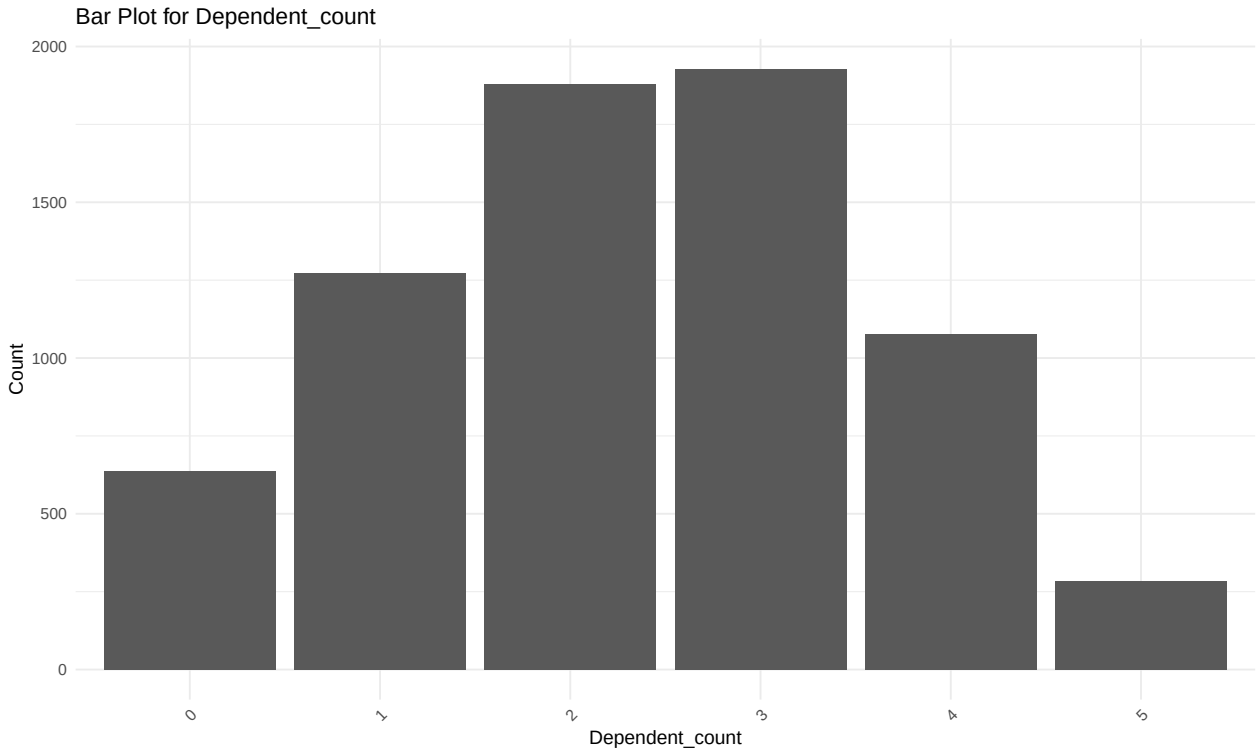
Visualisation :

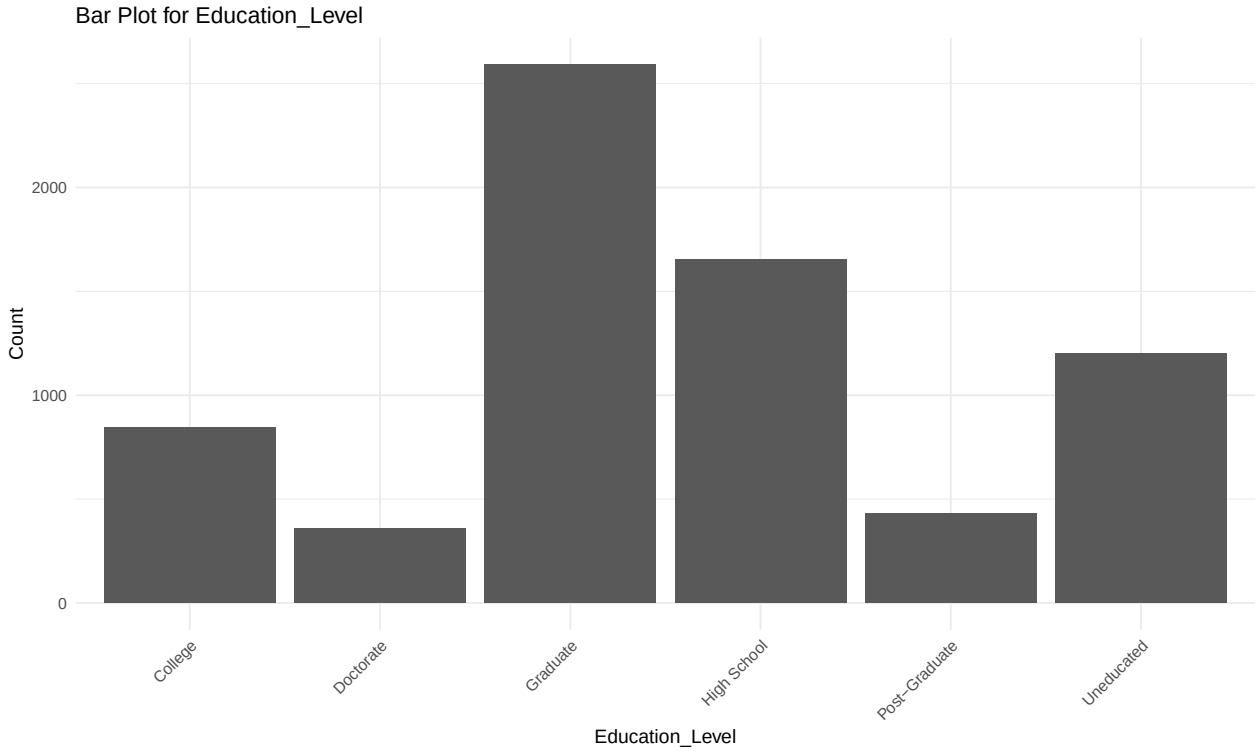
Visualisations pour les variables discrètes

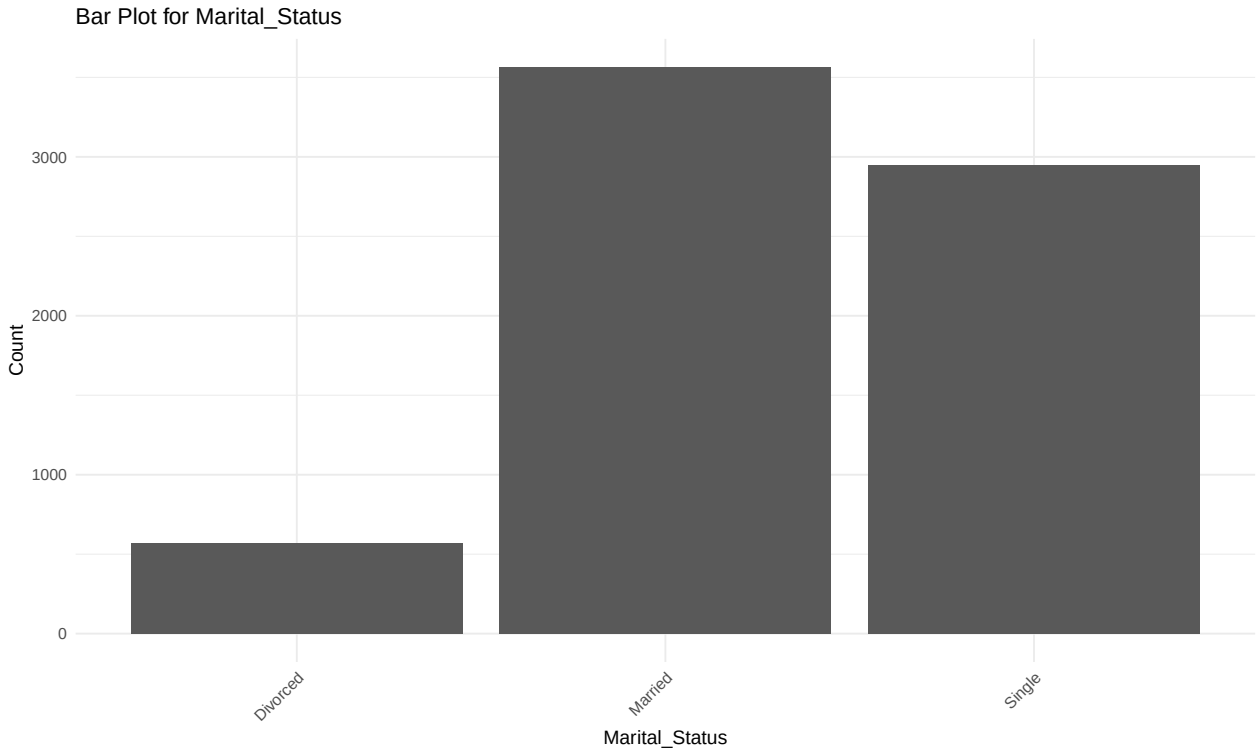


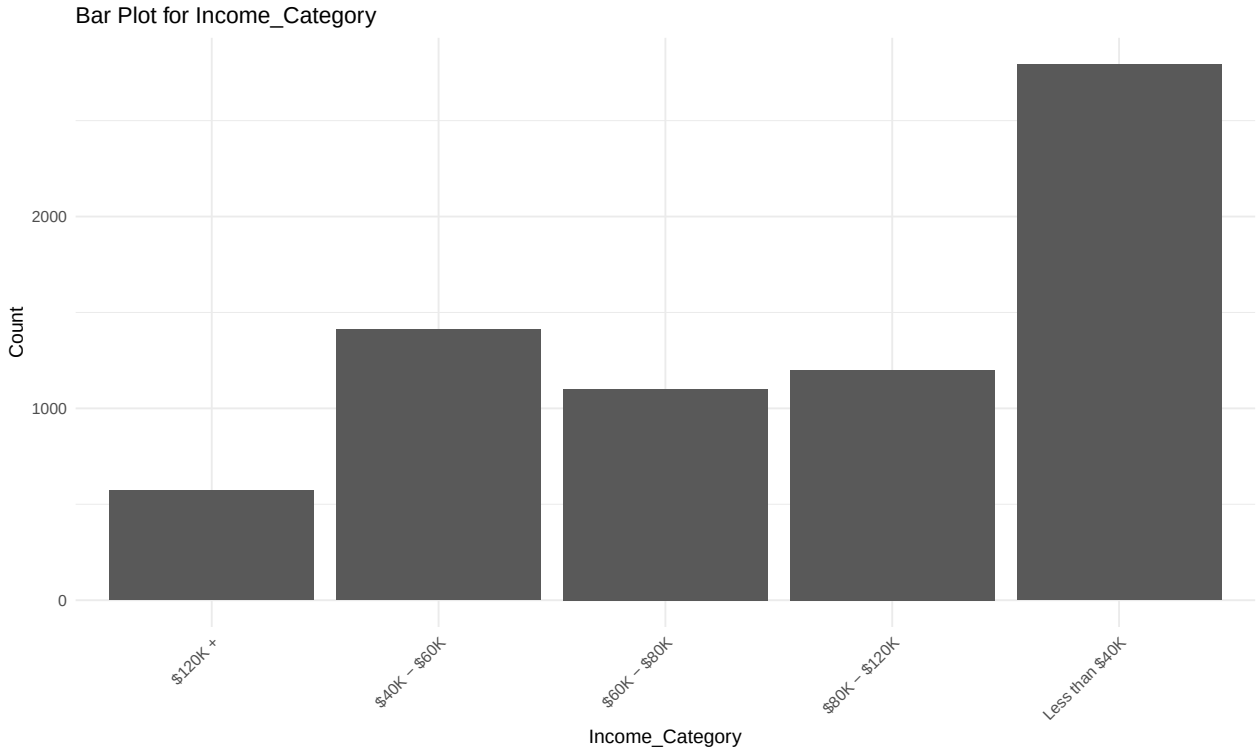

Bar Plot for Attrition_Flag
6000
4000
2000
0
Count
Attrited Customer
Existing Customer
Attrition_Flag

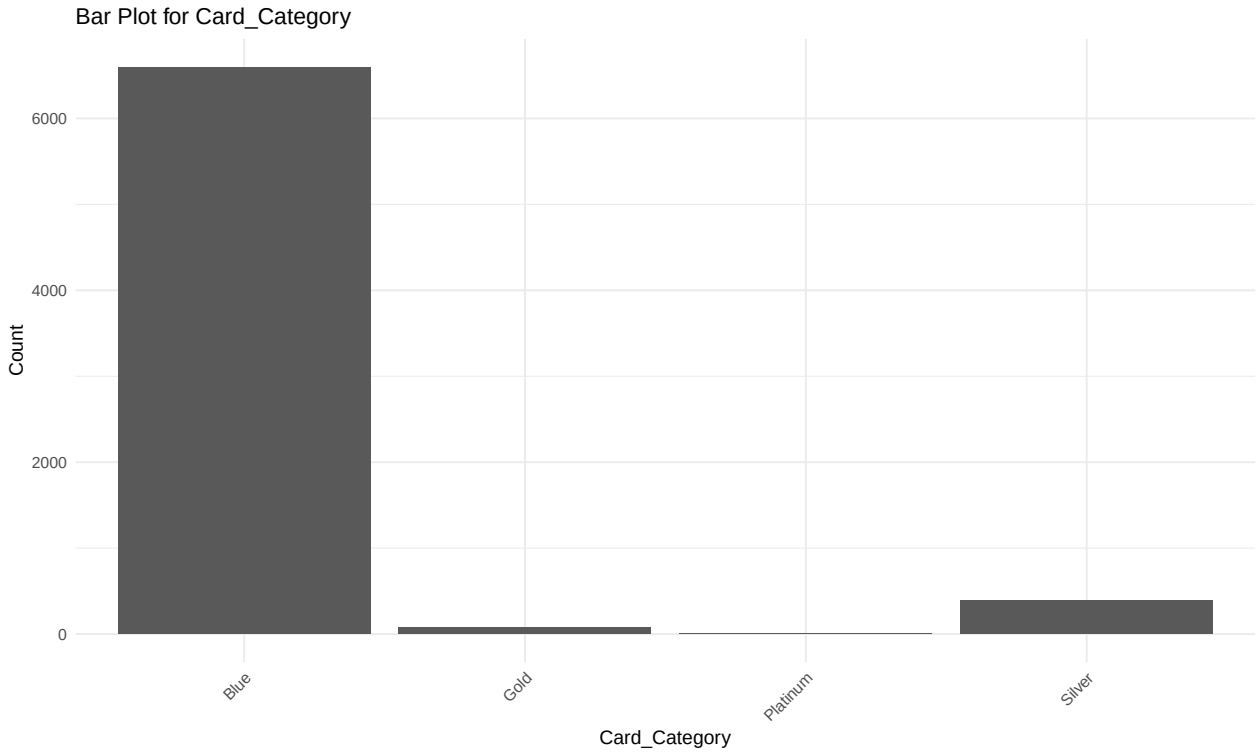

Bar Plot for Gender
3000
2000
1000
0
Count
F
M
Gender

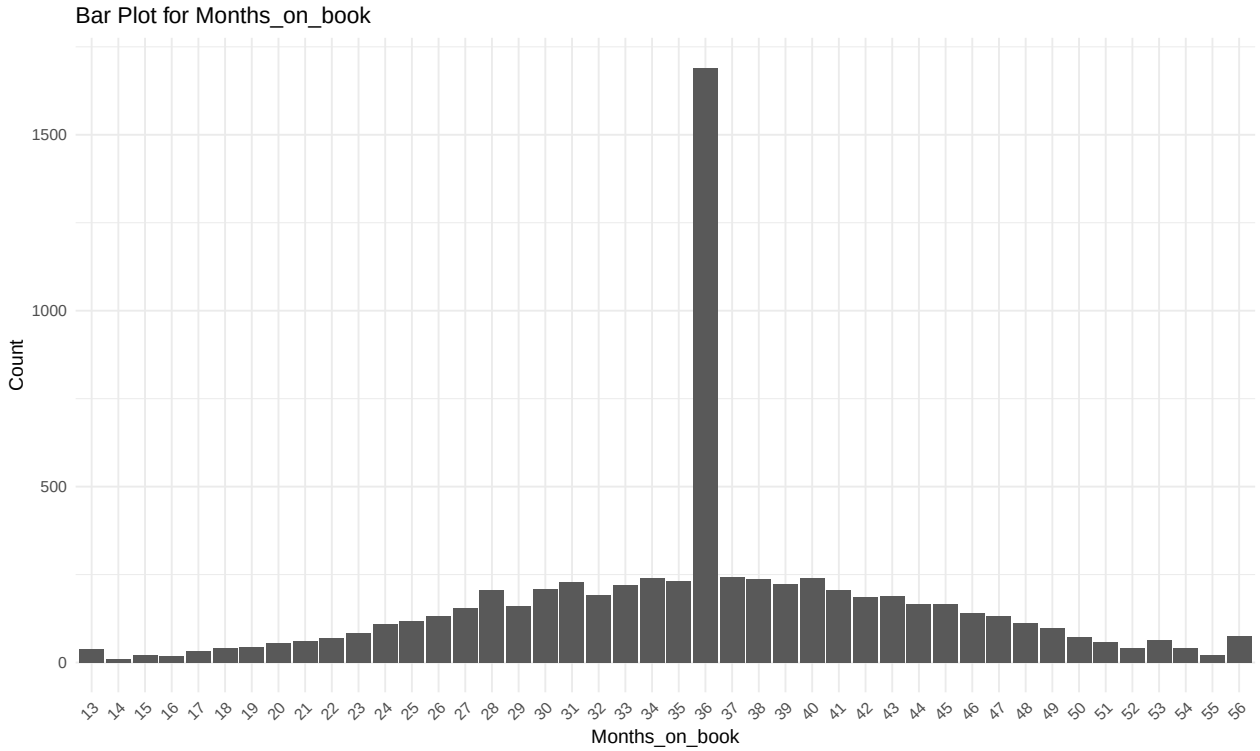

Bar Plot for Dependent_count
Count
2000
1500
1000
500
0
0
1
2
3
4
5
Dependent_count

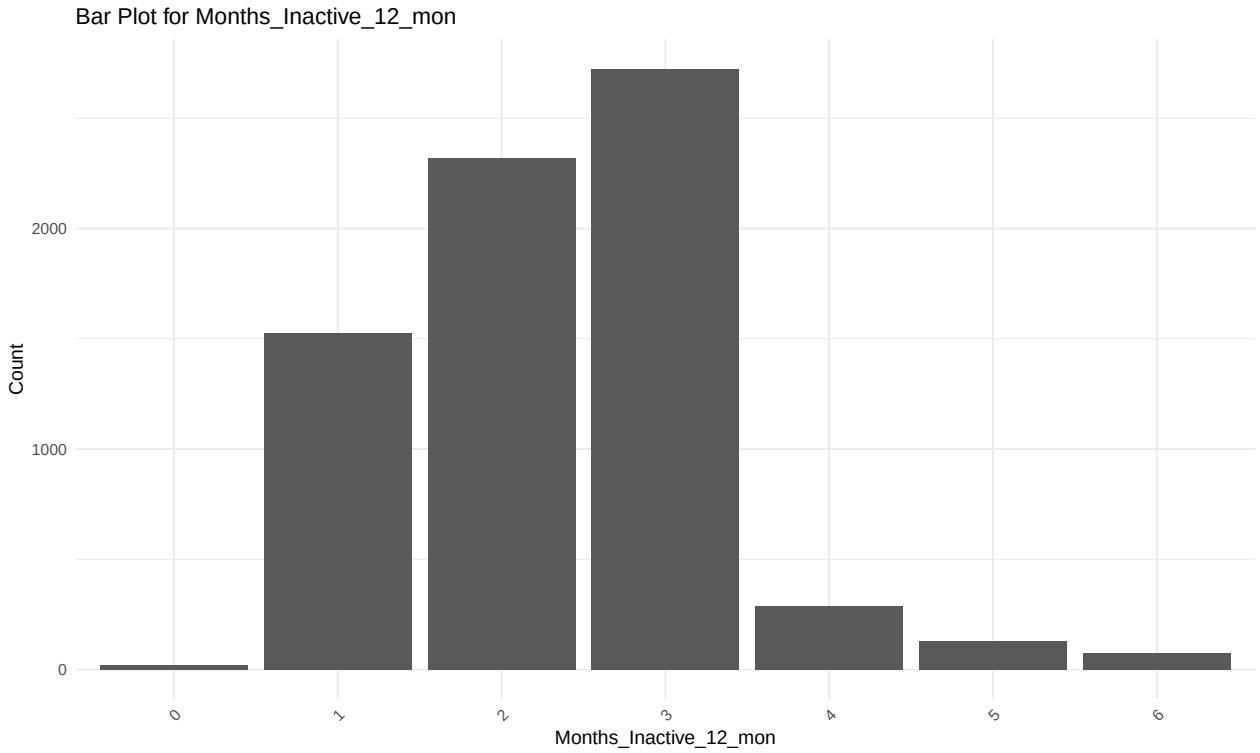

Bar Plot for Education_Level
Count
2000
1000
0
College
Doctorate
Graduate
High School
Post-Graduate
Uneducated
Education_Level

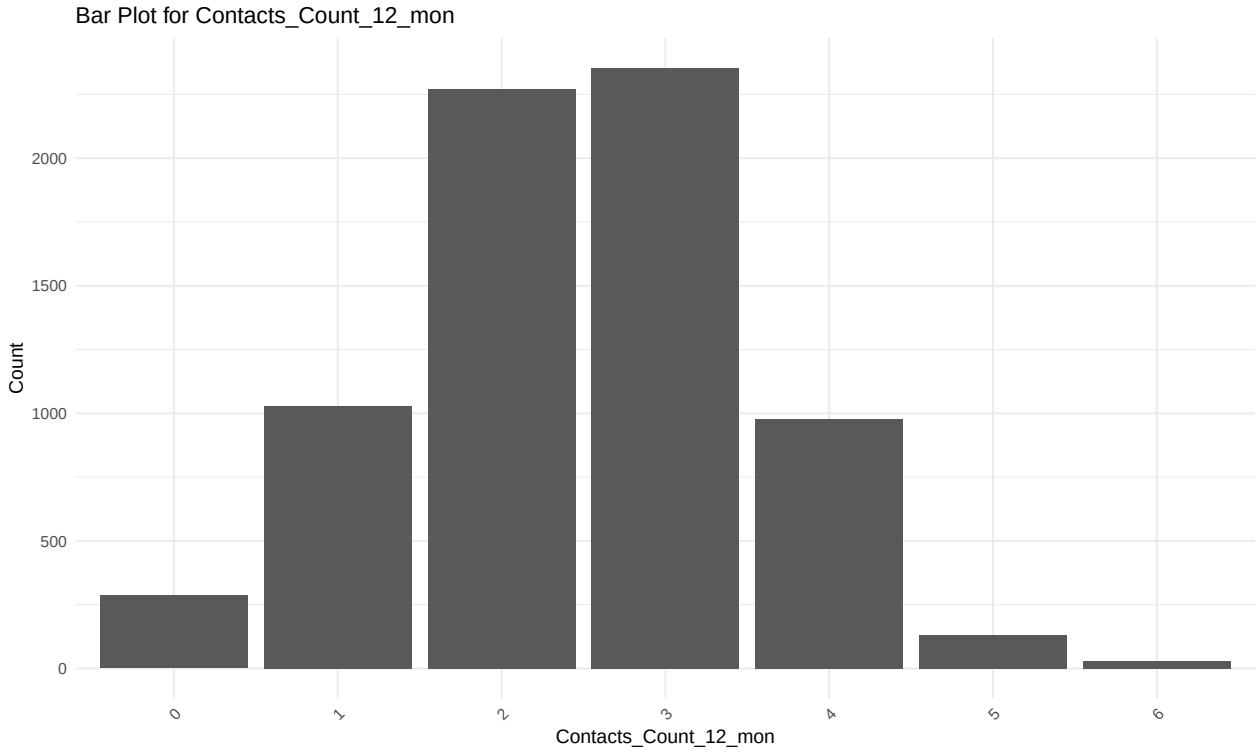

Bar Plot for Marital_Status
Count
3000
2000
1000
0
Divorced
Married
Single
Marital_Status


Bar Plot for Income_Category
Count
2000
1000
0
$120K +
$40K – $60K
$60K – $80K
$80K – $120K
Less than $40K
Income_Category

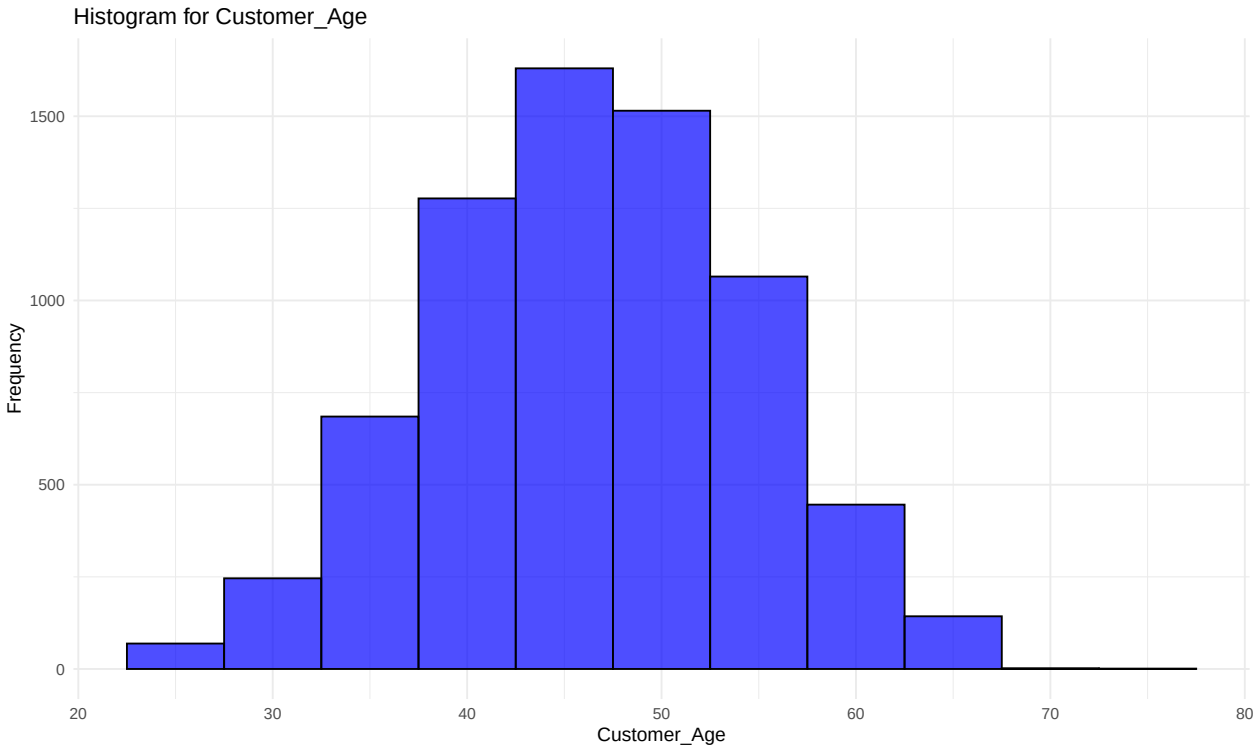

Bar Plot for Card_Category
6000
4000
2000
0
Count
Blue
Gold
Platinum
Silver
Card_Category

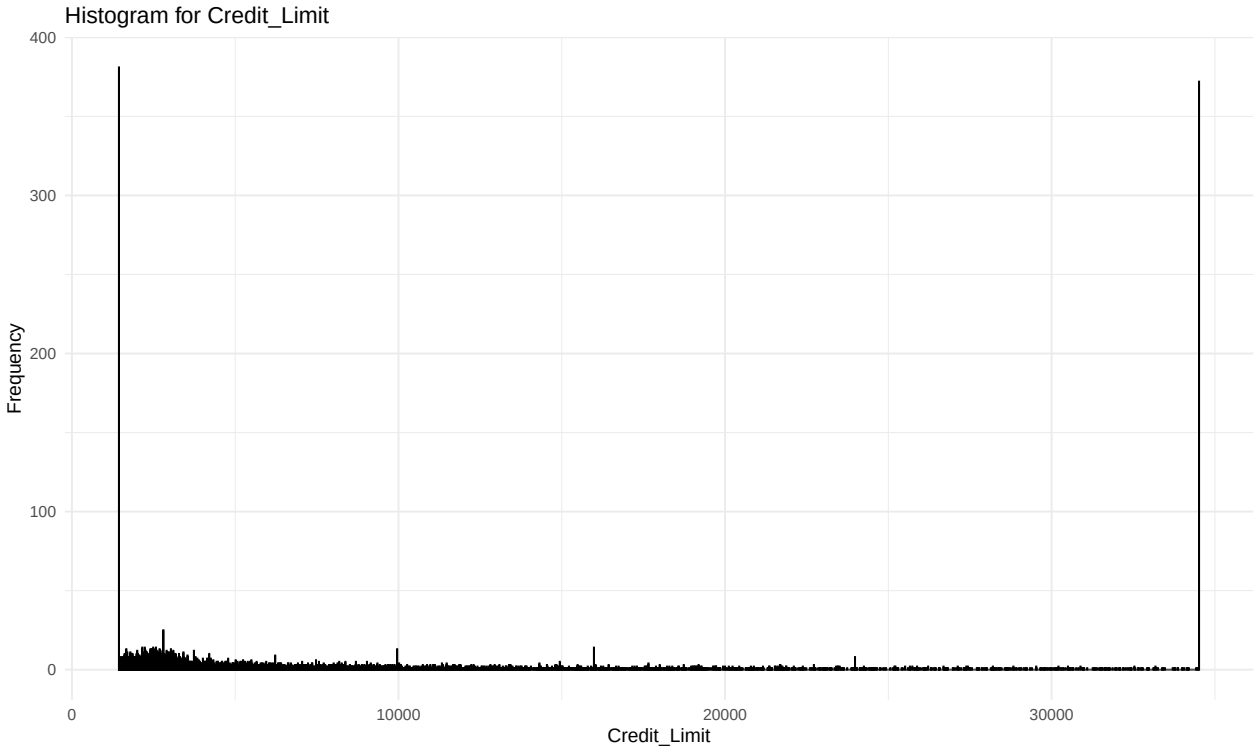

Bar Plot for Months_on_book
1500
1000
500
0
Count
13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56
Months_on_book

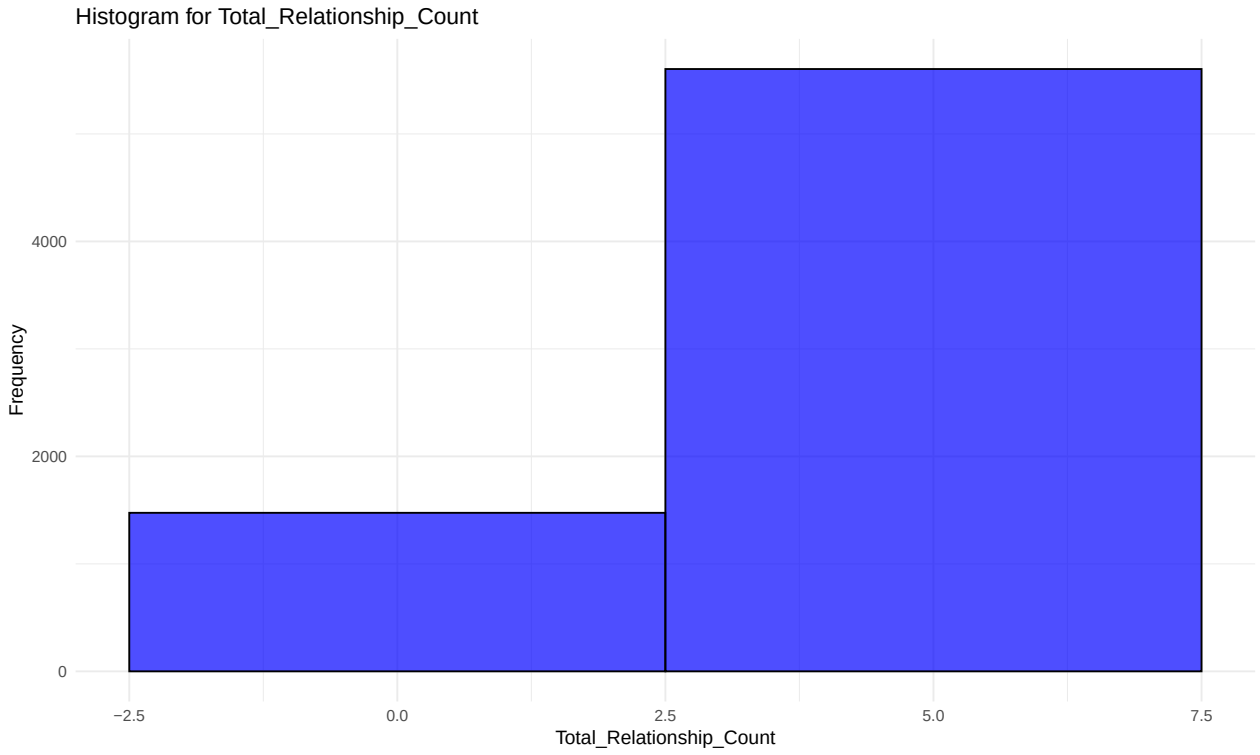

Bar Plot for Months_Inactive_12_mon
Count
2000
1000
0
0
1
2
3
4
5
6
Months_Inactive_12_mon

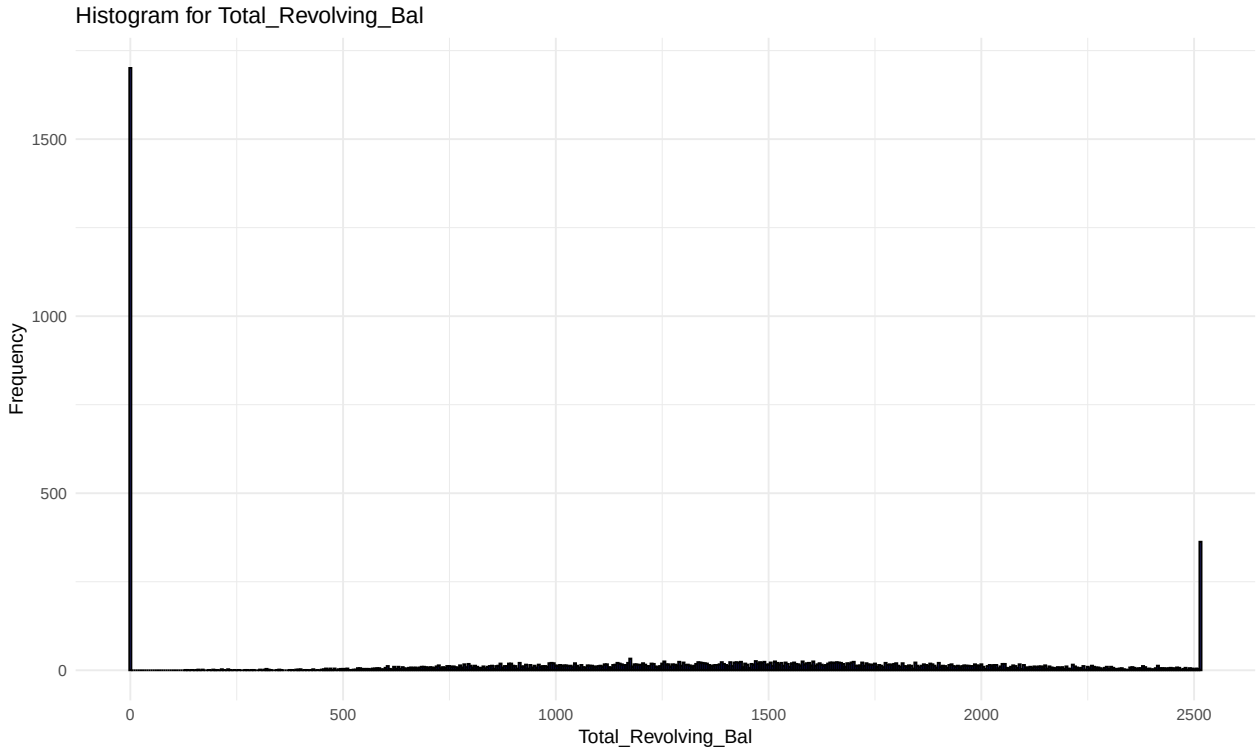

Bar Plot for Contacts_Count_12_mon
Count
2000
1500
1000
500
0
0
1
2
3
4
5
6
Contacts_Count_12_mon

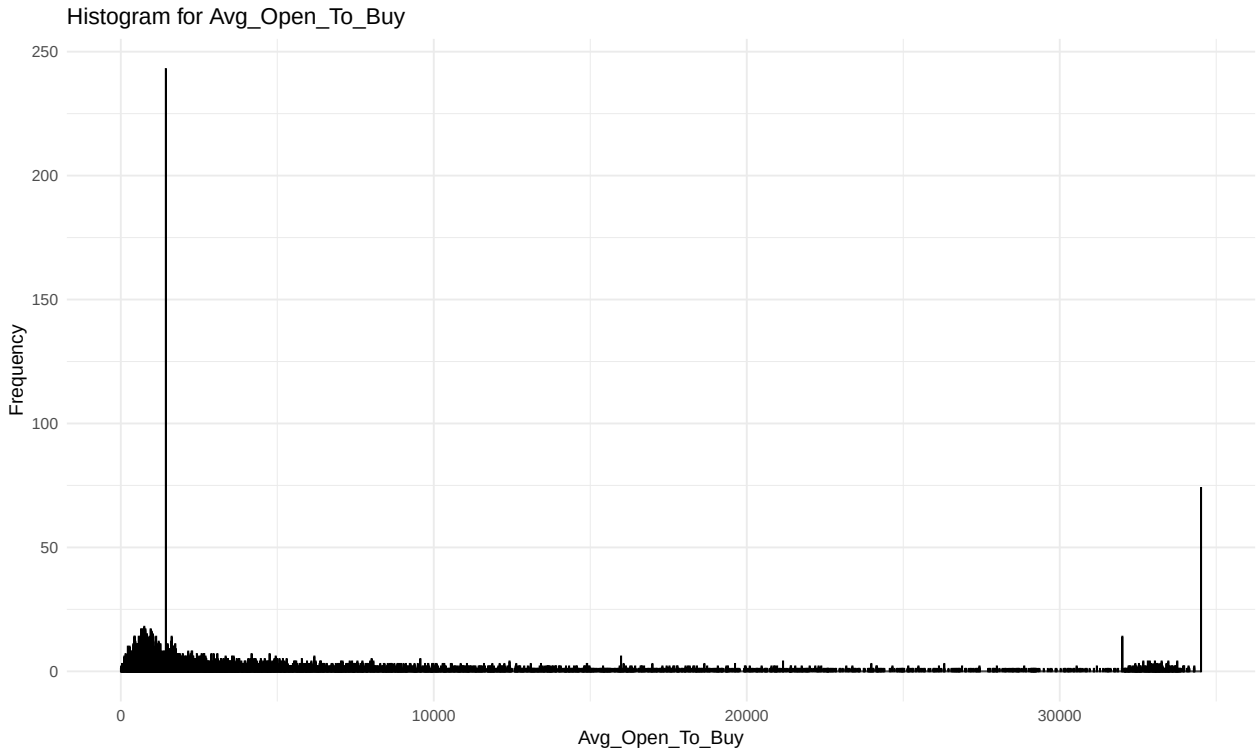
Visualisations pour les variables continues

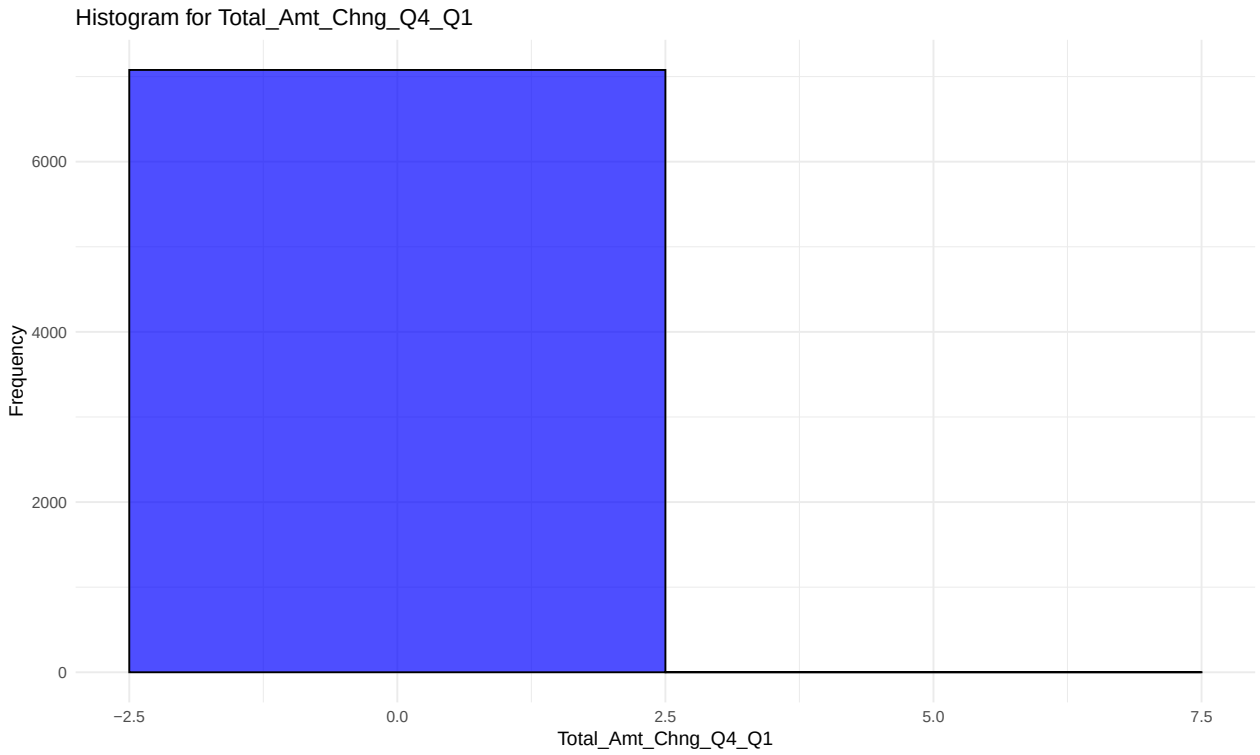


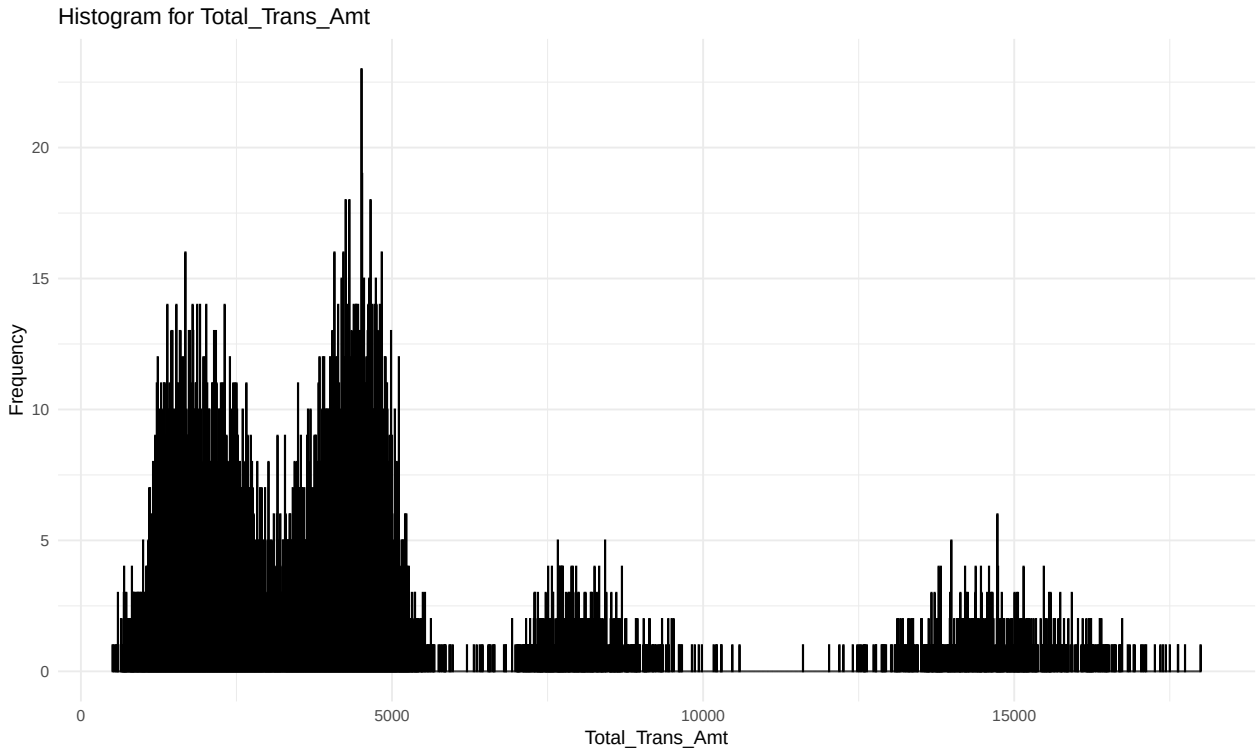


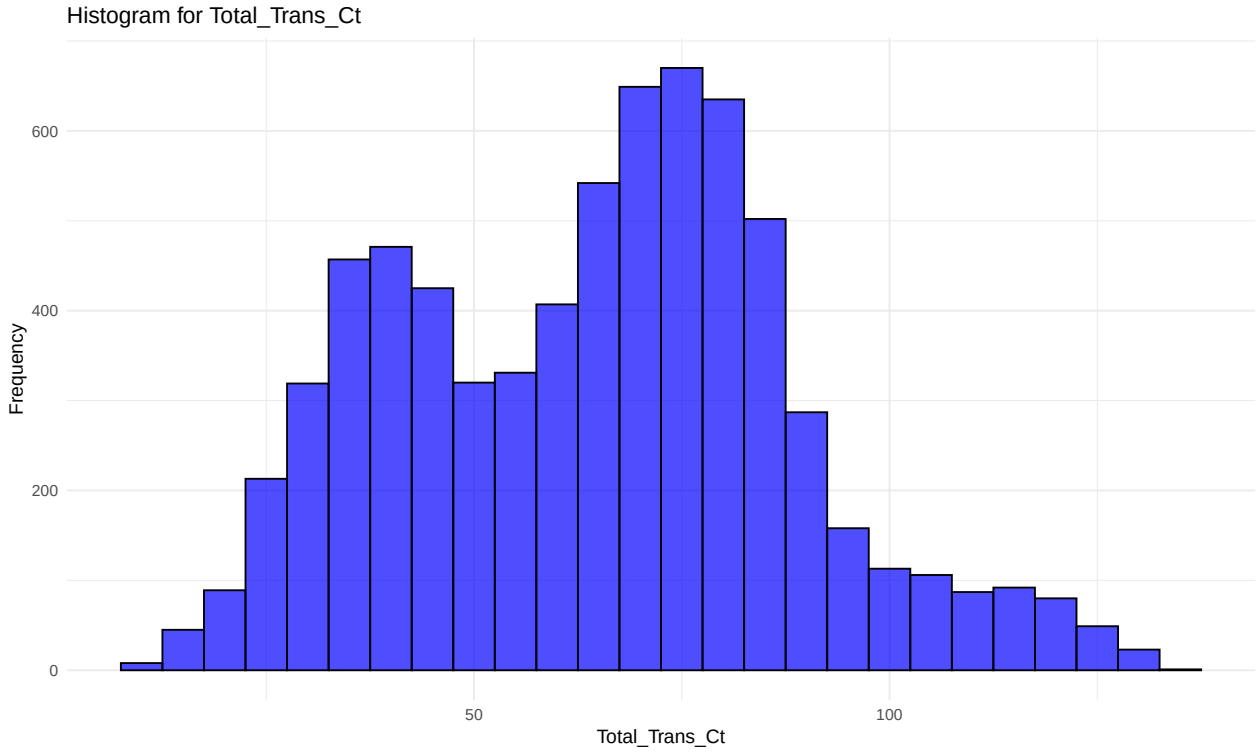

Histogram for Total_Relationship_Count
Frequency
4000
2000
0
−2.5
0.0
2.5
5.0
7.5
Total_Relationship_Count

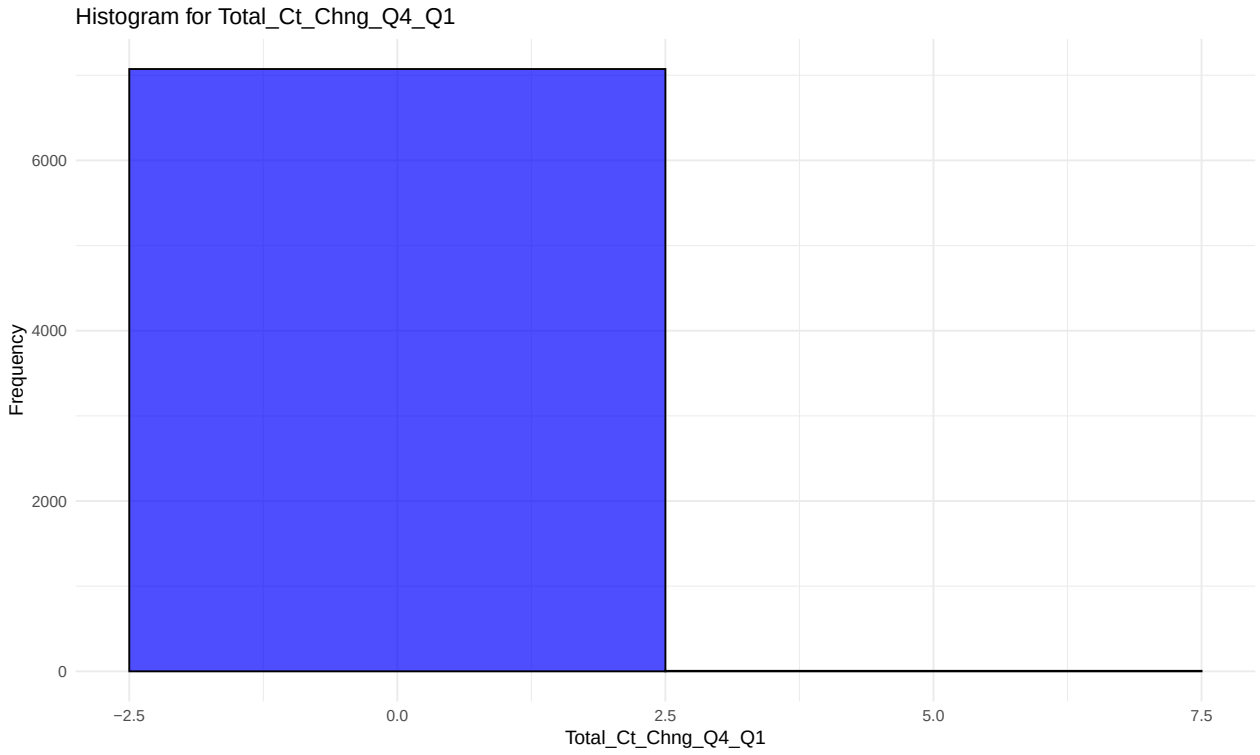

Histogram for Total_Revolving_Bal
Frequency
1500
1000
500
0
0
500
1000
1500
2000
2500
Total_Revolving_Bal

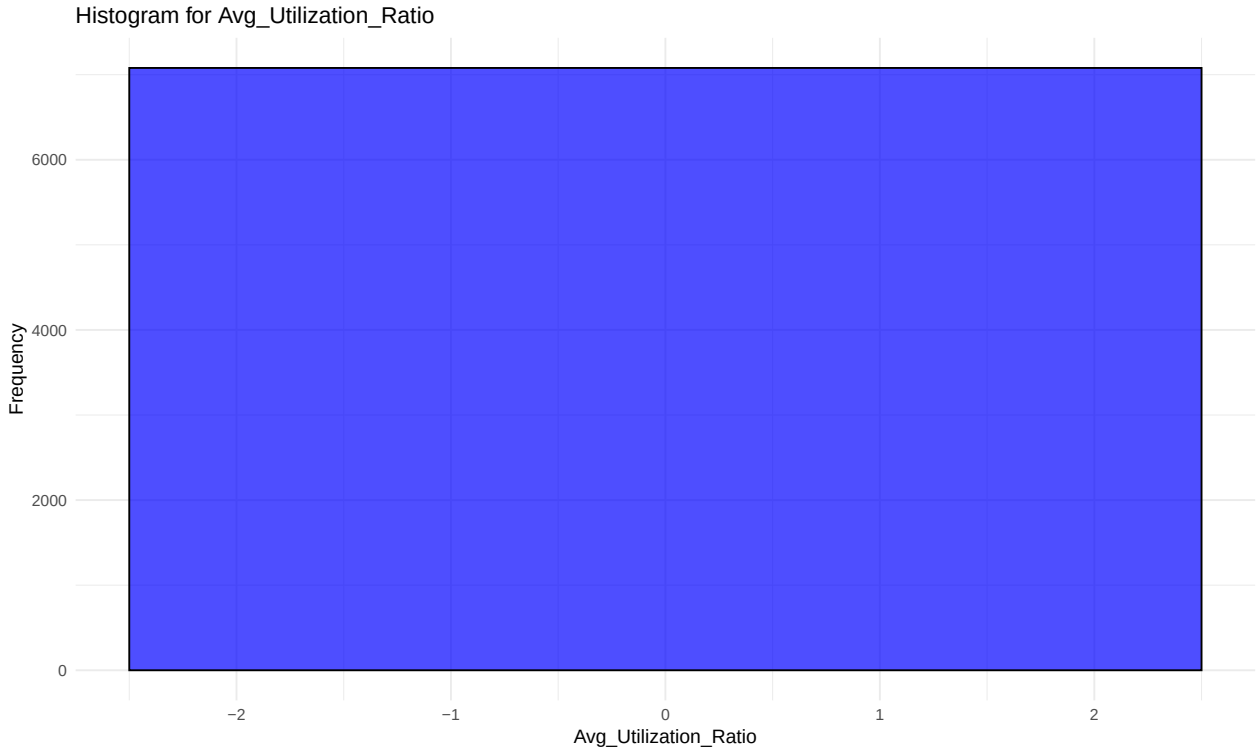

Histogram for Avg_Open_To_Buy
Frequency
250
200
150
100
50
0
0
10000
20000
30000
Avg_Open_To_Buy


Histogram for Total_Amt_Chng_Q4_Q1
Frequency
6000
4000
2000
0
−2.5
0.0
2.5
5.0
7.5
Total_Amt_Chng_Q4_Q1


Histogram for Total_Trans_Amt
Frequency
0
5
10
15
20
0
5000
10000
15000
Total_Trans_Amt


Histogram for Total_Trans_Ct
Frequency
0
200
400
600
50
100
Total_Trans_Ct


Histogram for Total_Ct_Chng_Q4_Q1
Frequency
6000
4000
2000
0
−2.5
0.0
2.5
5.0
7.5
Total_Ct_Chng_Q4_Q1


Histogram for Avg_Utilization_Ratio
Frequency
6000
4000
2000
0
−2
−1
0
1
2
Avg_Utilization_Ratio

ANALYSE DES STATISTIQUES DESCRIPTIVES

Résumé sur l'ensemble des variables :

```
## -- Data Summary ------------------------
##                            Values
## Name                       tab[, -c((ncol(tab) - 1):...
## Number of rows             7079
## Number of columns          21
## _________________________
## Column type frequency:
##   character                6
##   numeric                  15
## _________________________
## Group variables            None
##
## -- Variable type: character -----------------------------------------------------
##   skim_variable   n_missing complete_rate min max empty n_unique whitespace
## 1 Attrition_Flag          0             1  17  17     0        2          0
## 2 Gender                  0             1   1   1     0        2          0
## 3 Education_Level         0             1   7  13     0        6          0
## 4 Marital_Status          0             1   6   8     0        3          0
## 5 Income_Category         0             1   7  14     0        5          0
## 6 Card_Category           0             1   4   8     0        4          0
##
## -- Variable type: numeric -------------------------------------------------------
##    skim_variable            n_missing complete_rate          mean           sd
##  1 CLIENTNUM                        0             1 739099926.    36854571.
##  2 Customer_Age                     0             1        46.3          8.04
##  3 Dependent_count                  0             1         2.34         1.29
##  4 Months_on_book                   0             1        36.0          8.00
##  5 Total_Relationship_Count         0             1         3.82         1.54
##  6 Months_Inactive_12_mon           0             1         2.34         0.994
##  7 Contacts_Count_12_mon            0             1         2.46         1.10
##  8 Credit_Limit                     0             1      8490.        9126.
##  9 Total_Revolving_Bal              0             1      1167.         812.
## 10 Avg_Open_To_Buy                  0             1      7323.        9131.
## 11 Total_Amt_Chng_Q4_Q1             0             1         0.760        0.219
## 12 Total_Trans_Amt                  0             1      4395.        3469.
## 13 Total_Trans_Ct                   0             1        64.5         23.8
## 14 Total_Ct_Chng_Q4_Q1              0             1         0.711        0.237
## 15 Avg_Utilization_Ratio            0             1         0.282        0.279
##            p0           p25           p50           p75          p100 hist
##  1 708082083  713015058     717846108     773253520.    828298908
##  2        26         41            46            52           73
##  3         0          1             2             3            5
##  4        13         31            36            40.5         56
##  5         1          3             4             5            6
##  6         0          2             2             3            6
##  7         0          2             2             3            6
##  8      1438.      2496.         4287         10708         34516
##  9         0        464.         1282          1781         2517
## 10         3       1248.         3244          9490.        34516
## 11         0          0.629         0.735         0.858        2.59
```

```
## 12       510       2090          3831          4740         17995
## 13        10         44            67          80.5          134
## 14         0      0.583           0.7         0.818         3.71
## 15         0      0.026         0.186         0.516        0.999
```

Problématique

Problématique : Quels facteurs et profils de clients sont associés à la résiliation des services de cartes de crédit, et comment peut-on les prédire?

Objectifs :

1. **Compréhension des caractéristiques démographiques :**
   - **Question :** Quels sont les profils démographiques des clients résiliant leurs services de cartes de crédit?
   - **Méthodologie :** Analyses descriptives univariées pour chaque variable démographique.
2. **Étude des relations entre les variables :**
   - **Question :** Existe-t-il des relations significatives entre les différentes variables démographiques et transactionnelles?
   - **Méthodologie :** Analyses bivariées avec des tests de corrélation pour évaluer les relations.
3. **Facteurs sous-jacents à la résiliation :**
   - **Question :** Quels sont les facteurs sous-jacents qui contribuent le plus à la décision de résilier?
   - **Méthodologie :** Analyses factorielles pour réduire la dimensionnalité des données.
4. **Modélisation prédictive :**
   - **Question :** Peut-on développer un modèle de régression prédictive pour estimer la probabilité de résiliation en fonction des variables disponibles?
   - **Méthodologie :** Utilisation de méthodes de régression.
5. **Classification des clients :**
   - **Question :** Peut-on classer les clients en groupes distincts en fonction de leurs comportements et caractéristiques, en particulier en ce qui concerne la résiliation?
   - **Méthodologie :** Utilisation de méthodes de classification.

Conclusion

Notre pré-analyse des données de cartes de crédit a révélé un ensemble de données complet, sans valeurs manquantes et avec peu de valeurs aberrantes. Nous avons exploré les variables quantitatives et catégorielles, visualisé leurs distributions, et formulé une problématique axée sur la résiliation des services de cartes de crédit.

Les prochaines étapes de notre analyse incluront l'exploration des profils démographiques, l'étude des relations entre les variables, l'identification des facteurs de résiliation, la création de modèles prédictifs, et la classification des clients en groupes distincts. Cette pré-analyse fournit une base solide pour approfondir notre compréhension des comportements des clients de cartes de crédit.