#ETAPE 1

```
options(repos = c(CRAN = "https://cran.r-project.org"))



install.packages("readxl")
```

```
## le package 'readxl' a été décompressé et les sommes MD5 ont été vérifiées avec succés

##
## Les packages binaires téléchargés sont dans
##  C:\Users\hamza\AppData\Local\Temp\RtmpmW6r35\downloaded_packages
```

```
library(readxl)
setwd("C:/Users/hamza/Documents/iae/m1")
dataset <- read_excel("dataset_1.xlsx")
dataset
```

```
## # A tibble: 1,036 x 13
##       ID Marital_Status Gender Income Children Education   Occupation Home_Owner
##    <dbl> <chr>          <chr>   <dbl>    <dbl> <chr>       <chr>      <chr>
##  1 12496 M              F       40000        1 Bachelors   Skilled M~ Yes
##  2 24107 M              M       30000        3 Partial Co~ Clerical   Yes
##  3 14177 M              M       80000        5 Partial Co~ Professio~ No
##  4 24381 S              M       70000        0 Bachelors   Professio~ Yes
##  5 25597 S              M       30000        0 Bachelors   Clerical   No
##  6 13507 M              F       10000        2 Partial Co~ Manual     Yes
##  7 27974 S              M      160000        2 High School Management Yes
##  8 19364 M              M       40000        1 Bachelors   Skilled M~ Yes
##  9 22155 M              M       20000        2 Partial Hi~ Clerical   Yes
## 10 19280 M              M      120000        2 Partial Co~ Manual     Yes
## # i 1,026 more rows
## # i 5 more variables: Cars <dbl>, Commute_Distance <chr>, Region <chr>,
## #   Age <dbl>, Purchased_Bike <chr>
```

J'ai telecharge la bibliotheque 'readxl' qui nous permet d'importer notre data set, mais j'ai eu un probleme avec l'endroit ou se trouve mon fichier donc j'ai du le changer avec setwd() et du coup avec le read_excel() j'ai pu afficher le dataset.

```
duplicated_rows <- duplicated(dataset)
dataset[duplicated_rows, ]
```

```
## # A tibble: 34 x 13
##       ID Marital_Status Gender Income Children Education   Occupation Home_Owner
##    <dbl> <chr>          <chr>   <dbl>    <dbl> <chr>       <chr>      <chr>
##  1 13507 M              F       10000        2 Partial Co~ Manual     Yes
##  2 19280 M              M      120000        2 Partial Co~ Manual     Yes
##  3 22173 M              F       30000        3 High School Skilled M~ No
##  4 12697 S              F       90000        0 Bachelors   Professio~ No
##  5 11434 M              M      170000        5 Partial Co~ Professio~ Yes
##  6 25323 M              M       40000        2 Partial Co~ Clerical   Yes
```

```
##  7 23542 S              M       60000        1 Partial Co~ Skilled M~ No
##  8 20870 S              F       10000        2 High School Manual     Yes
##  9 23316 S              M       30000        3 Partial Co~ Clerical   No
## 10 12610 M              F       30000        1 Bachelors   Clerical   Yes
## # i 24 more rows
## # i 5 more variables: Cars <dbl>, Commute_Distance <chr>, Region <chr>,
## #   Age <dbl>, Purchased_Bike <chr>
```

Ici, jai utilise duplicated() pour voir les doublons et puis je les ai affiche. Ensuit, je vais les supprimer en gardant que les lignes uniques grace a '!' et on voit qu'il ya 1002 lignes maintenant au lieu de 1036:

```
dataset <- dataset[!duplicated(dataset), ]
nrow(dataset)
```

```
## [1] 1002
```

```
missing_va <- colSums(is.na(dataset))
missing_va
```

```
##               ID   Marital_Status           Gender           Income
##                0                0                0                1
##         Children        Education       Occupation       Home_Owner
##                0                1                1                1
##             Cars Commute_Distance           Region              Age
##                0                0                2                0
##   Purchased_Bike
##                0
```

La j'ai verifier combien de valeurs sont manquantes dans chaque colonne,pour le 'income' je prefere remplacer avec la moyenne sinon le reste c'est des valeurs non numerique donc j'utilise un remplacement par la valeur la plus frequente:

```
dataset$Income[is.na(dataset$Income)] <- mean(dataset$Income, na.rm = TRUE)
```

```
for (col in colnames(dataset)) {
  if (is.character(dataset[[col]]) || is.factor(dataset[[col]])) {
    # Calculer la fréquence des valeurs
    freq_table <- table(dataset[[col]])

    # Trouver la valeur la plus fréquente (le mode)
    mode_value <- names(freq_table)[which.max(freq_table)]

    # Remplacer les valeurs manquantes par le mode
    dataset[[col]][is.na(dataset[[col]])] <- mode_value
  }
}
```

Et du coup, la on verifie si les valeurs sont bonnes:

```
colSums(is.na(dataset))
```

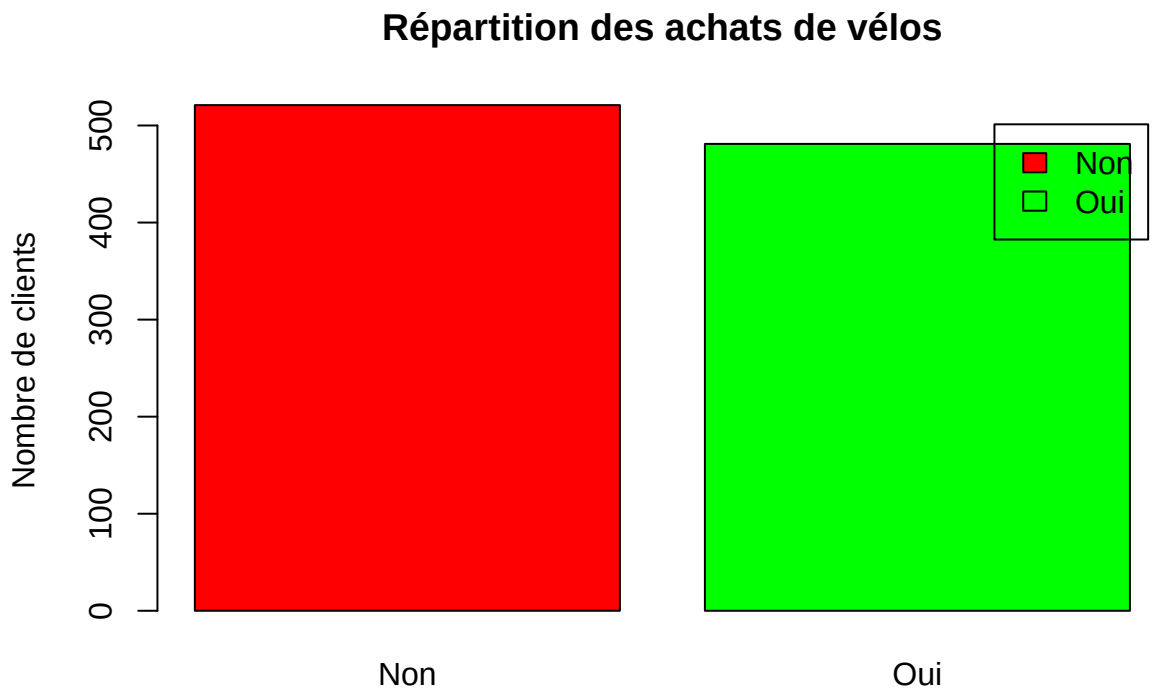
```
##                ID   Marital_Status           Gender           Income
##                 0                0                0                0
##          Children        Education       Occupation       Home_Owner
##                 0                0                0                0
##             Cars Commute_Distance           Region              Age
##                 0                0                0                0
##    Purchased_Bike
##                 0
```

#Etape 2

```
counts <- table(dataset$Purchased_Bike)
print(counts)
```

```
##
##  No Yes
## 521 481
```

```
barplot(counts,
        main = "Répartition des achats de vélos",
        col = c("red", "green"),
        legend = c("Non", "Oui"),
        ylab = "Nombre de clients",
        names.arg = c("Non", "Oui"))
```



```r
str(dataset)
```

```
## tibble [1,002 x 13] (S3: tbl_df/tbl/data.frame)
##  $ ID              : num [1:1002] 12496 24107 14177 24381 25597 ...
##  $ Marital_Status  : chr [1:1002] "M" "M" "M" "S" ...
##  $ Gender          : chr [1:1002] "F" "M" "M" "M" ...
##  $ Income          : num [1:1002] 40000 30000 80000 70000 30000 10000 160000 40000 20000 120000 ...
##  $ Children        : num [1:1002] 1 3 5 0 0 2 2 1 2 2 ...
##  $ Education       : chr [1:1002] "Bachelors" "Partial College" "Partial College" "Bachelors" ...
##  $ Occupation      : chr [1:1002] "Skilled Manual" "Clerical" "Professional" "Professional" ...
##  $ Home_Owner      : chr [1:1002] "Yes" "Yes" "No" "Yes" ...
##  $ Cars            : num [1:1002] 0 1 2 1 0 0 4 0 2 1 ...
##  $ Commute_Distance: chr [1:1002] "0-1 Miles" "0-1 Miles" "2-5 Miles" "5-10 Miles" ...
##  $ Region          : chr [1:1002] "Europe" "Europe" "Europe" "Pacific" ...
##  $ Age             : num [1:1002] 42 43 60 41 36 50 33 43 58 40 ...
##  $ Purchased_Bike  : chr [1:1002] "No" "No" "No" "Yes" ...
```

On voit que income et age sont de type numerique et Purcased_Bike est de type categorielle (ordinale ou pas) faire une matrice de corr pr tt les

```r
dataset$Purchased_Bike_num <- ifelse(dataset$Purchased_Bike == "Yes", 1, 0)

cor_income <- cor(dataset$Income, dataset$Purchased_Bike_num, use = "complete.obs")
print(paste("Corrélation entre Income et Purchased_Bike :", cor_income))
```

```
## [1] "Corrélation entre Income et Purchased_Bike : 0.0501084227178161"
```

```r
cor_age <- cor(dataset$Age, dataset$Purchased_Bike_num, use = "complete.obs")
print(paste("Corrélation entre Age et Purchased_Bike :", cor_age))
```

```
## [1] "Corrélation entre Age et Purchased_Bike : -0.106942078861221"
```

et pour Commute_distance, on va devoir changer les valeurs des distances pour donner des valeurs numeriques:

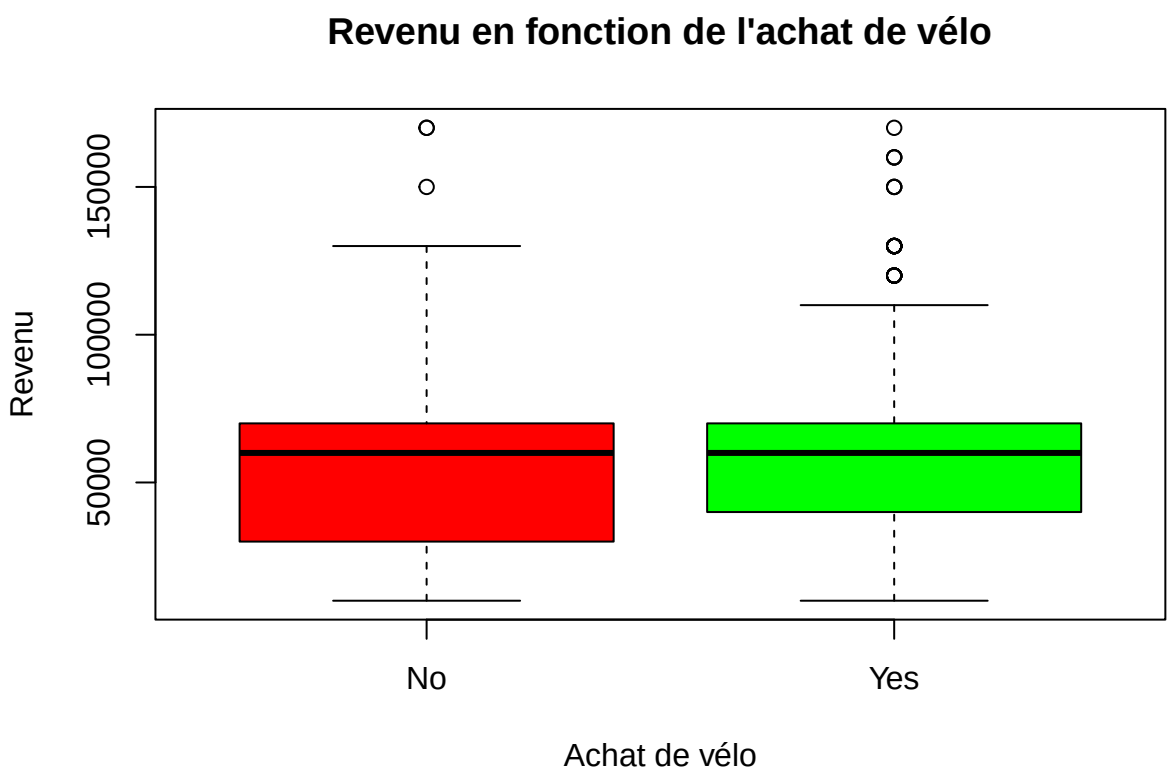
```r
convert_commute_to_numeric <- function(distance) {
  if (distance == "0-1 Miles") {
    return(0.5)
  } else if (distance == "1-2 Miles") {
    return(1.5)
  } else if (distance == "2-5 Miles") {
    return(3.5)
  } else if (distance == "5-10 Miles") {
    return(7.5)
  } else if (distance == "10-15 Miles") {
    return(12.5)
  } else if (distance == "15-20 Miles") {
    return(17.5)
  } else {
    return(NA)
  }
```

```
}

dataset$Commute_Distance_num <- sapply(dataset$Commute_Distance,convert_commute_to_numeric)


cor_commute <- cor(dataset$Commute_Distance_num, dataset$Purchased_Bike_num, use = "complete.obs")
print(paste("Corrélation entre Commute_Distance et Purchased_Bike :", cor_commute))
```

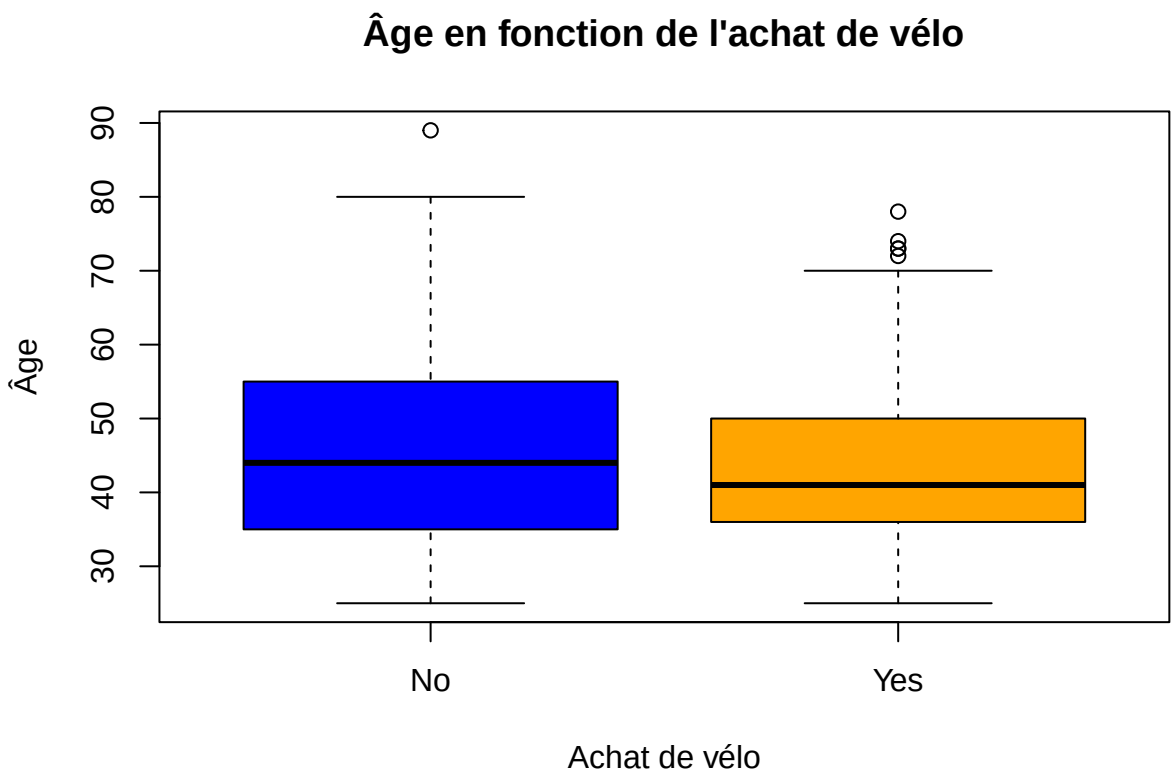
```
## [1] "Corrélation entre Commute_Distance et Purchased_Bike : -0.0919858578635445"
```

Voila des visualisations appropriées pour comprendre la repartition et les correlations des données:

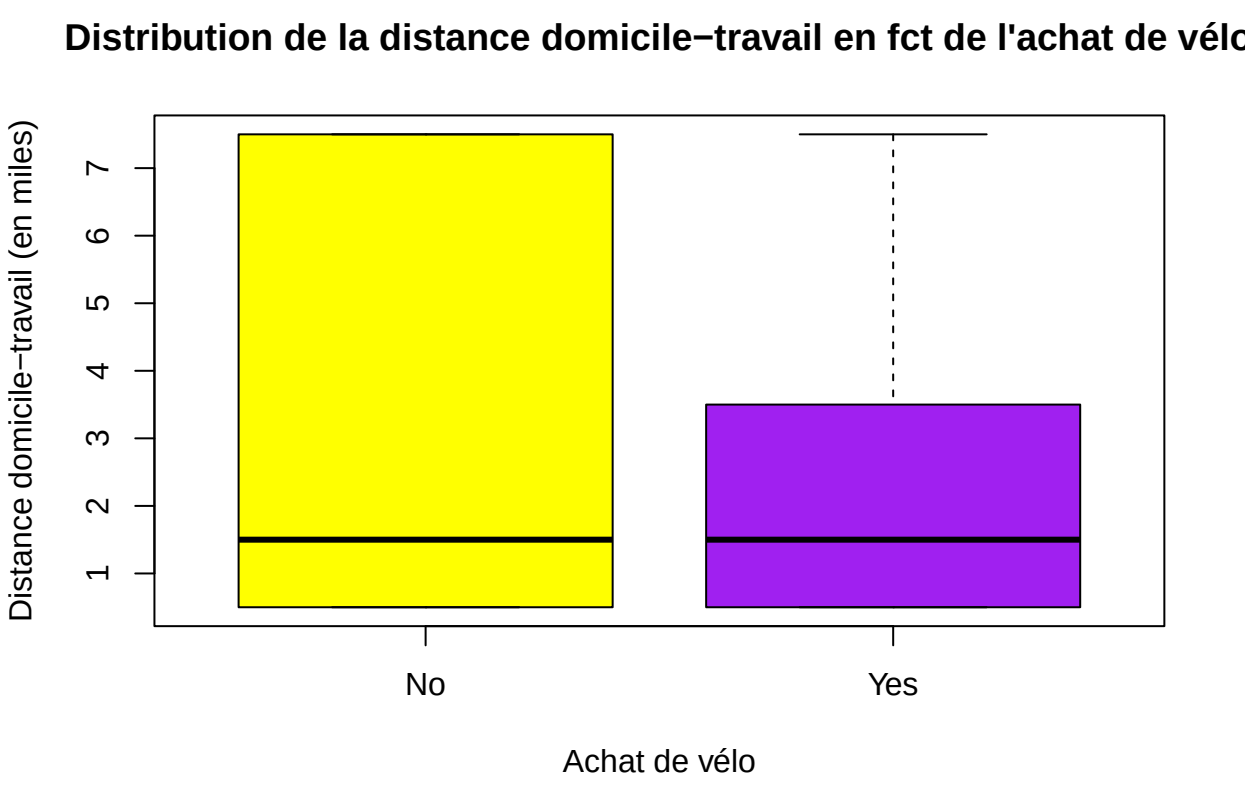
```
boxplot(Income ~ Purchased_Bike,
        data = dataset,
        main = "Revenu en fonction de l'achat de vélo",
        xlab = "Achat de vélo",
        ylab = "Revenu",
        col = c("red", "green"))
```




```
boxplot(Age ~ Purchased_Bike,
        data = dataset,
        main = "Âge en fonction de l'achat de vélo",
        xlab = "Achat de vélo",
        ylab = "Âge",
        col = c("blue", "orange"))
```

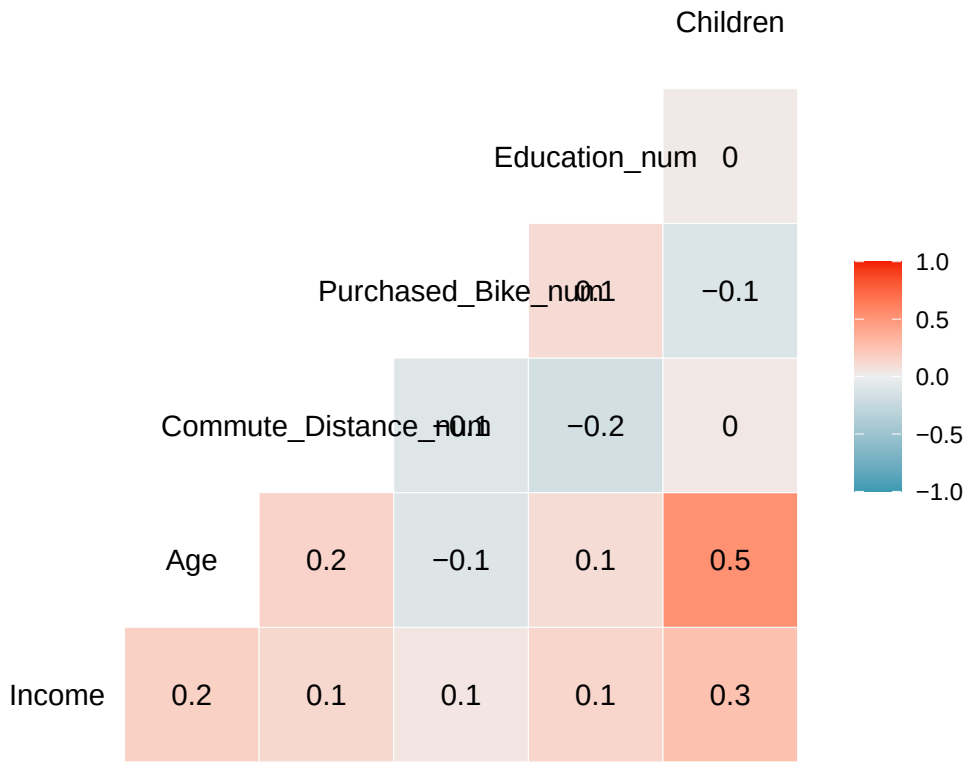



```r
boxplot(Commute_Distance_num ~ Purchased_Bike,
        data = dataset,
        main = "Distribution de la distance domicile-travail en fct de l'achat de vélo",
        xlab = "Achat de vélo",
        ylab = "Distance domicile-travail (en miles)",
        col = c("yellow", "purple"))
```



Du coup pour une correlation complete avec toutes les variables, jai pris les variables numeriques et les variables categorielles ordinales et je l'ai numeriser pour pouvoir faire une correlation:

```
library(dplyr)
library(GGally)

education_mapping <- c("Partial College" = 1, "Bachelors" = 2, "Masters" = 3, "PhD" = 4)
dataset$Education_num <- as.numeric(education_mapping[dataset$Education])

numeric_and_ordinal <- dataset %>%
  select(Income, Age, Commute_Distance_num, Purchased_Bike_num, Education_num, Children)

ggcorr(numeric_and_ordinal,
       method = c("pairwise", "pearson"),
       label = TRUE,
       label_size = 4,
       legend.size = ,
       title = "Corrélogramme avec variables numériques et ordinales")
```

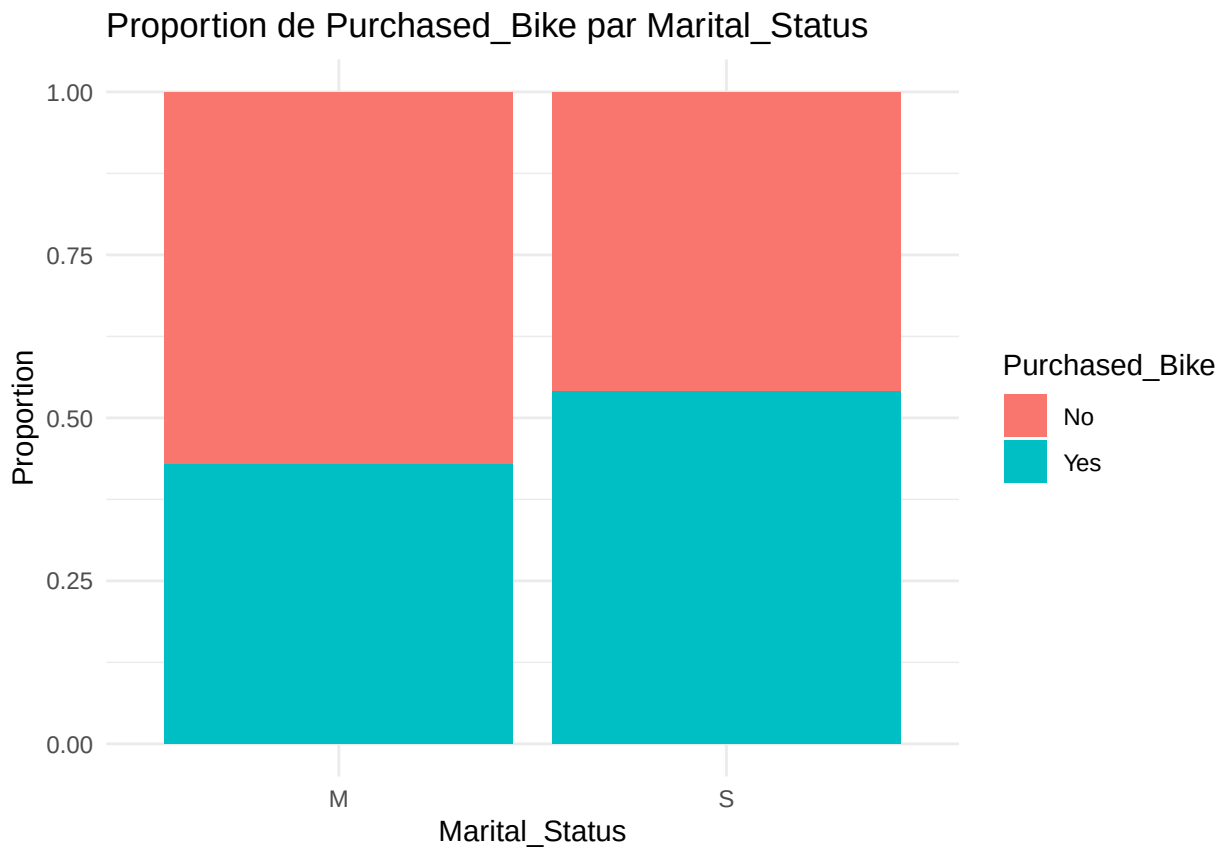


#ETAPE 3 J'ai fais une segmentation selon les 5 variables categorielles pertinentes en fonction des gens qui ont achetes un velo:

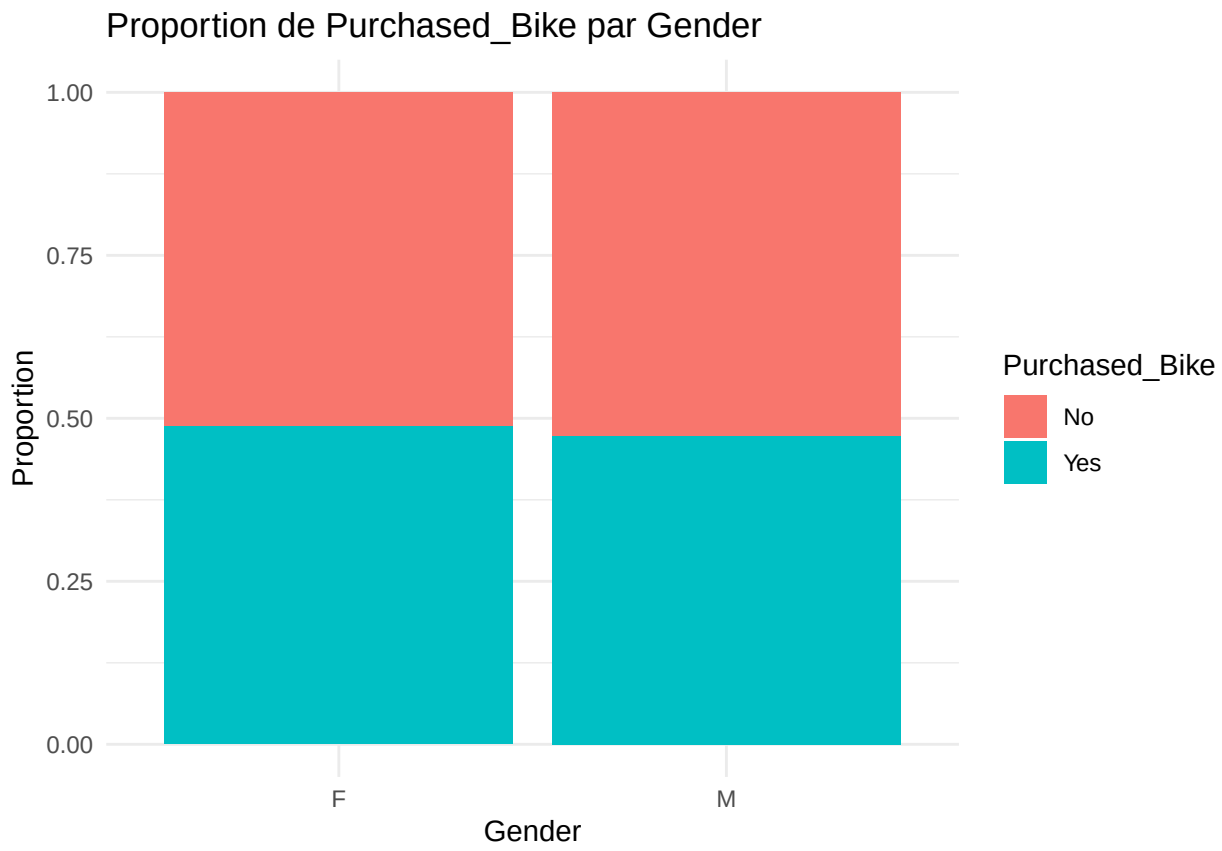
```
analyze_categorical <- function(data, target, cat_var) {
  cross_tab <- table(data[[target]], data[[cat_var]])
  print(cross_tab)

  cross_tab_df <- as.data.frame(cross_tab)

  ggplot(cross_tab_df, aes(x = Var2, y = Freq, fill = Var1)) +
    geom_bar(stat = "identity", position = "fill") +
    labs(
      x = cat_var,
      y = "Proportion",
      fill = target,
      title = paste("Proportion de", target, "par", cat_var)
    ) +
    theme_minimal()
}

cat_vars <- c("Marital_Status", "Gender", "Home_Owner", "Region", "Occupation")

for (cat_var in cat_vars) {
  print(paste("Analyse pour la variable :", cat_var))
  plot <- analyze_categorical(dataset, "Purchased_Bike", cat_var)
  print(plot)
```

```
}
```

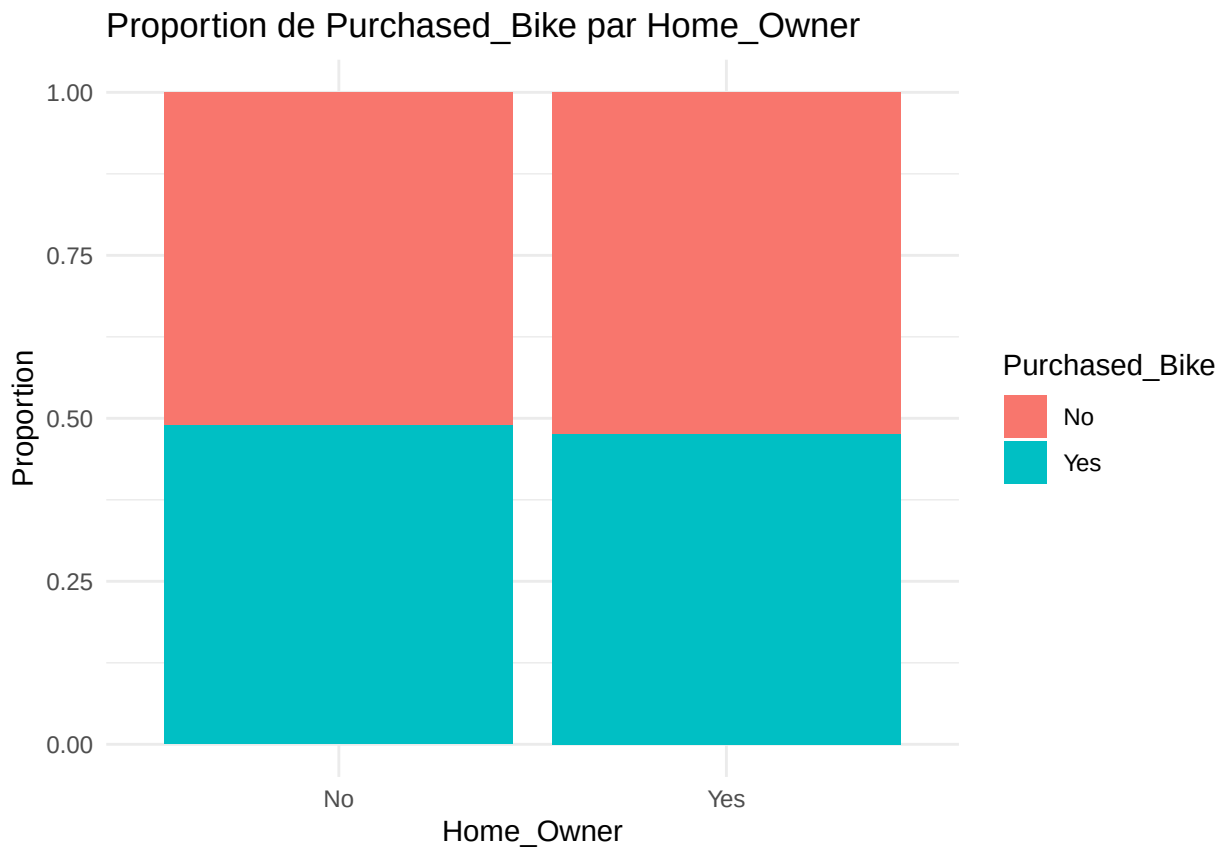
```
## [1] "Analyse pour la variable : Marital_Status"
##
##         M   S
##   No  308 213
##   Yes 231 250
```



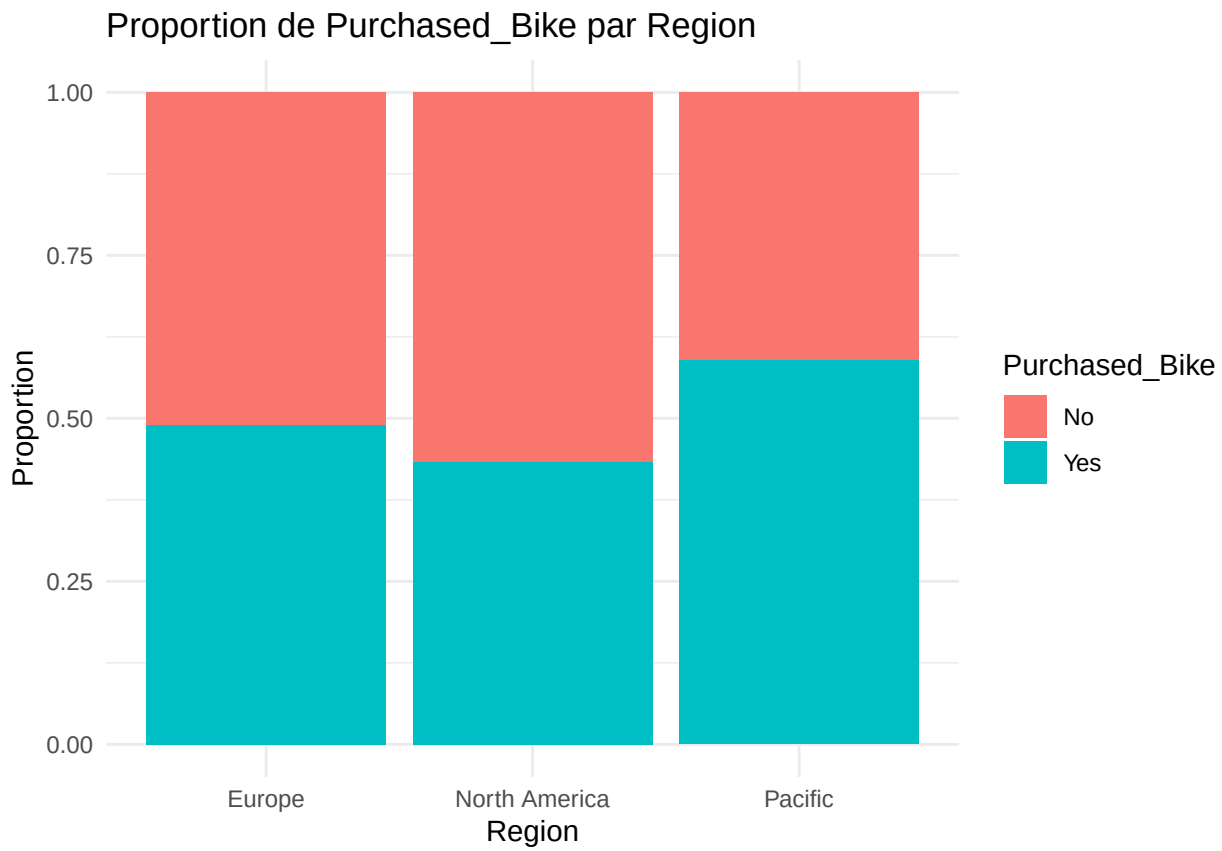

```
## [1] "Analyse pour la variable : Gender"
##
##         F   M
##   No  251 270
##   Yes 239 242
```



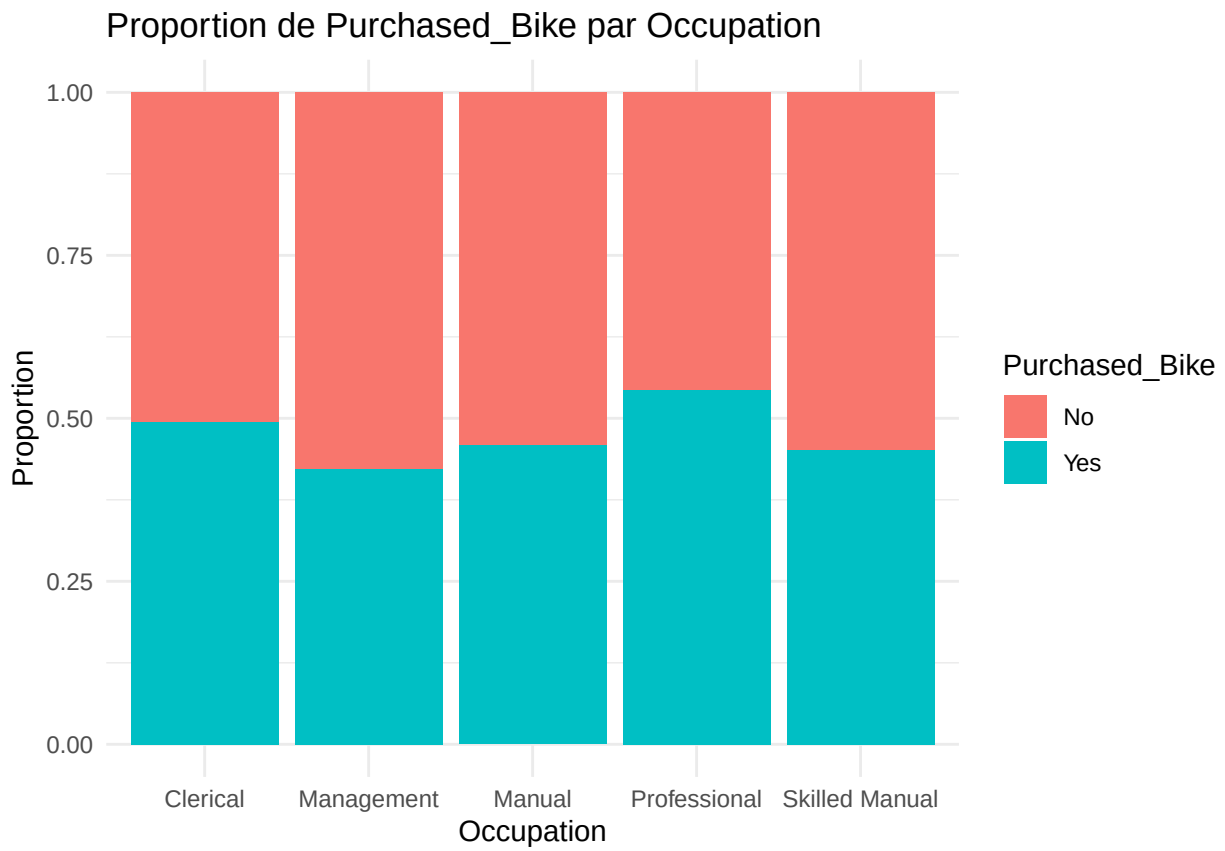

```
## [1] "Analyse pour la variable : Home_Owner"
##
##         No Yes
##   No   162 359
##   Yes 155 326
```




```
## [1] "Analyse pour la variable : Region"
##
##       Europe North America Pacific
##   No     153           289      79
##   Yes    147           221     113
```




```
## [1] "Analyse pour la variable : Occupation"
##
##       Clerical Management Manual Professional Skilled Manual
##   No        90        100     65          126             140
##   Yes       88         73     55          150             115
```



On peut passer a une segmentation a trois variables mais il ya deja des segmentations pertinentes avec deux variables ou on voit que les proportions sont differentes.

#ETAPE 4

Rapport d'analyse des données sur l'achat de vélos

1. Introduction L'objectif de cette analyse était de comprendre les facteurs influençant l'achat de vélos, d'identifier les segments de clients les plus susceptibles d'acheter un vélo et de formuler des recommandations marketing pour augmenter les ventes. Nous avons travaillé à partir d'un fichier de données contenant des informations sur les clients, telles que leur genre, statut marital, profession, région et statut de propriétaire, ainsi qu'une indication sur l'achat ou non d'un vélo.

2. Étapes de l'analyse

2.1 Chargement des données Les données ont été importées dans R à l'aide de la fonction `read.csv()`. Nous avons vérifié leur structure et effectué un prétraitement pour convertir certaines variables, comme `Purchased_Bike`, en facteur, afin de faciliter l'analyse.

```
data$Purchased_Bike <- factor(data$Purchased_Bike, levels = c("No", "Yes"))
```

2.2 Analyse de la distribution de la variable cible Nous avons commencé par examiner la distribution de la variable cible `Purchased_Bike` pour comprendre combien de clients avaient acheté un vélo et combien ne l'avaient pas fait. Une visualisation sous forme de graphique à barres a permis de constater que la majorité des clients avaient acheté un vélo.

2.3 Étude des relations avec d'autres variables Nous avons analysé les corrélations entre `Purchased_Bike` et les variables numériques telles que `Income`, `Age` et `Commute_Distance`. Des graphiques de dispersion ont été réalisés pour évaluer les tendances et comprendre l'impact de ces variables sur la décision d'achat.

2.4 Segmentation des clients Nous avons segmenté les clients en fonction de variables catégorielles comme `Marital_Status`, `Gender`, `Occupation`, `Home_Owner`, et `Region`. Pour chaque groupe, nous avons calculé le nombre total de clients, le nombre d'acheteurs de vélos et la proportion d'acheteurs.

Des tableaux croisés et des graphiques ont permis de visualiser les résultats, mettant en évidence les segments les plus intéressants, comme les célibataires, les professionnels, et les propriétaires de maison.

2.5 Identification des caractéristiques des acheteurs À partir des segments analysés, nous avons identifié les caractéristiques communes aux groupes ayant un taux d'achat élevé. Ces analyses ont permis de mieux comprendre les préférences et besoins des acheteurs.

3. Principales conclusions Les personnes mariées pourraient être plus ou moins enclines à acheter des vélos que les célibataires, en fonction des besoins familiaux ou des priorités budgétaires. Les achats de vélos peuvent varier selon les régions, en fonction de facteurs comme la topographie (ex. : zones urbaines vs rurales), la disponibilité des pistes cyclables, ou les modes de transport locaux.Ici on voit que dans la region du Pacifique il ya plus d'acheteurs de velos. Certaines professions peuvent être associées à des modes de vie plus actifs, ce qui pourrait influencer les achats de vélos mais ici on voit qu'il ya une legere difference d'achats de velos chez les professionnels, on dirait qu'ils en ont achetes plus que les autres.

4. Recommandations marketing

1. **Cibler les célibataires et les jeunes adultes** avec des vélos urbains et des promotions adaptées, comme des réductions ou des financements.
2. **Proposer des vélos haut de gamme ou électriques** pour les professionnels et travailleurs manuels qualifiés, mettant en avant leur praticité et robustesse.
3. **Adapter les campagnes publicitaires par région** en tenant compte des tendances locales, notamment en Amérique du Nord, Europe et Pacifique.
4. **Promouvoir des accessoires et services complémentaires** (porte-vélos, équipements de sécurité, garanties) pour attirer les propriétaires de maison.
5. **Intensifier la communication en ligne**, notamment via les réseaux sociaux, pour toucher des segments spécifiques en fonction de leur genre, profession ou région.

---

5. Conclusion Cette analyse a permis d'identifier les groupes les plus susceptibles d'acheter des vélos et de proposer des stratégies marketing adaptées. Une approche ciblée, basée sur les caractéristiques des acheteurs et les spécificités régionales, devrait permettre d'améliorer significativement les ventes tout en répondant aux attentes des clients.

#ETAPE 6 (VOITURE)

Combien de clients ont acheté une voiture ? Combien ne l'ont pas fait ? Cest quoi la moyenne de voitures achetes par personne?

```r
total_cars_purchased <- sum(dataset$Cars, na.rm = TRUE)

clients_with_cars <- nrow(dataset[dataset$Cars > 0, ])

clients_without_cars <- nrow(dataset[dataset$Cars == 0, ])

average_cars_per_client <- mean(dataset$Cars, na.rm = TRUE)

print(paste("Nombre total de voitures achetées :", total_cars_purchased))
```

```
## [1] "Nombre total de voitures achetées : 1443"
```

```r
print(paste("Nombre de clients avec au moins une voiture :", clients_with_cars))
```

```
## [1] "Nombre de clients avec au moins une voiture : 754"
```

```r
print(paste("Nombre de clients sans voiture :", clients_without_cars))
```

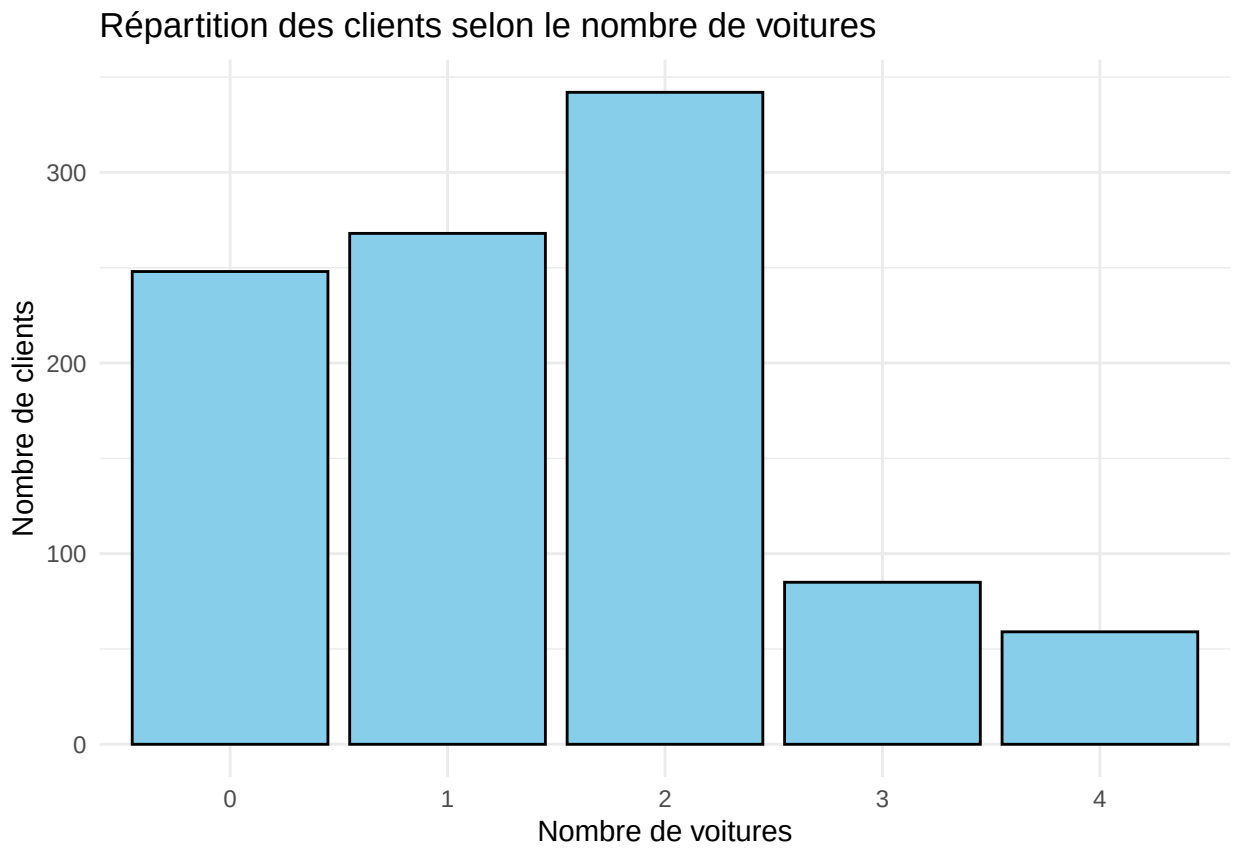
```
## [1] "Nombre de clients sans voiture : 248"
```

```r
print(paste("Moyenne de voitures achetées par client :", round(average_cars_per_client, 2)))
```

```
## [1] "Moyenne de voitures achetées par client : 1.44"
```

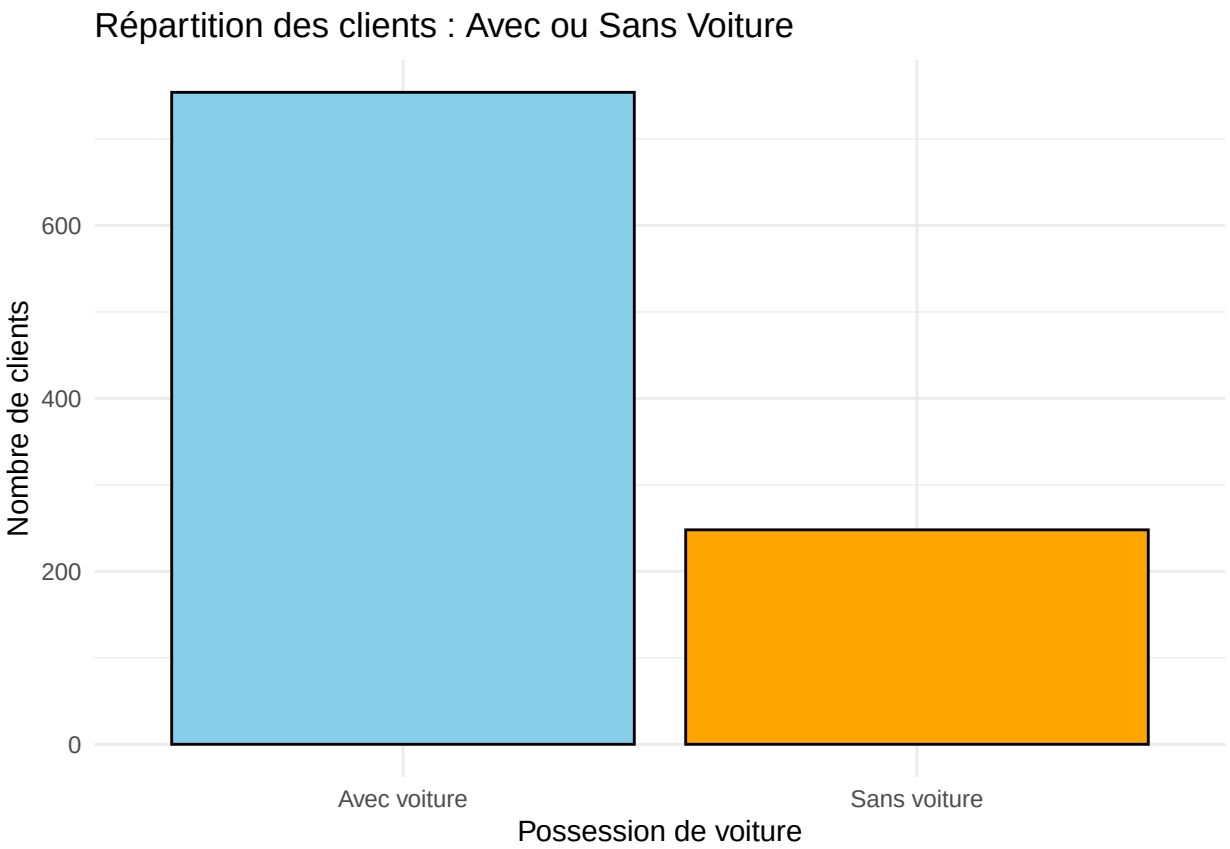
Ensuite on a fait un graphique avec la répartition des clients en fonction du nombre de voitures qu'ils possèdent.

```r
library(ggplot2)

car_distribution <- as.data.frame(table(dataset$Cars))

colnames(car_distribution) <- c("Number_of_Cars", "Frequency")

ggplot(car_distribution, aes(x = Number_of_Cars, y = Frequency)) +
  geom_bar(stat = "identity", fill = "skyblue", color = "black") +
  labs(title = "Répartition des clients selon le nombre de voitures",
       x = "Nombre de voitures",
       y = "Nombre de clients") +
  theme_minimal()
```



Ce graphique montre visuellement la proportion de clients ayant une voiture par rapport à ceux qui n'en possèdent pas:

```r
dataset$Has_Car <- ifelse(dataset$Cars > 0, "Avec voiture", "Sans voiture")

car_ownership_distribution <- as.data.frame(table(dataset$Has_Car))

ggplot(car_ownership_distribution, aes(x = Var1, y = Freq, fill = Var1)) +
  geom_bar(stat = "identity", color = "black") +
  labs(title = "Répartition des clients : Avec ou Sans Voiture",
       x = "Possession de voiture",
       y = "Nombre de clients") +
  scale_fill_manual(values = c("skyblue", "orange")) +
  theme_minimal() +
  theme(legend.position = "none")
```



Le graphique montre clairement une différence significative entre les deux groupes, avec une dominance des clients possédant une voiture.

Maintenant, on calcule la correlation entre les variables differentes:

```
cor_income <- cor(dataset$Income, dataset$Cars, use = "complete.obs")
print(paste("Corrélation entre Income et Cars :", cor_income))
```

```
## [1] "Corrélation entre Income et Cars : 0.428463382536988"
```

```
cor_age <- cor(dataset$Age, dataset$Cars, use = "complete.obs")
print(paste("Corrélation entre Age et Cars :", cor_age))
```

```
## [1] "Corrélation entre Age et Cars : 0.18576956046666"
```

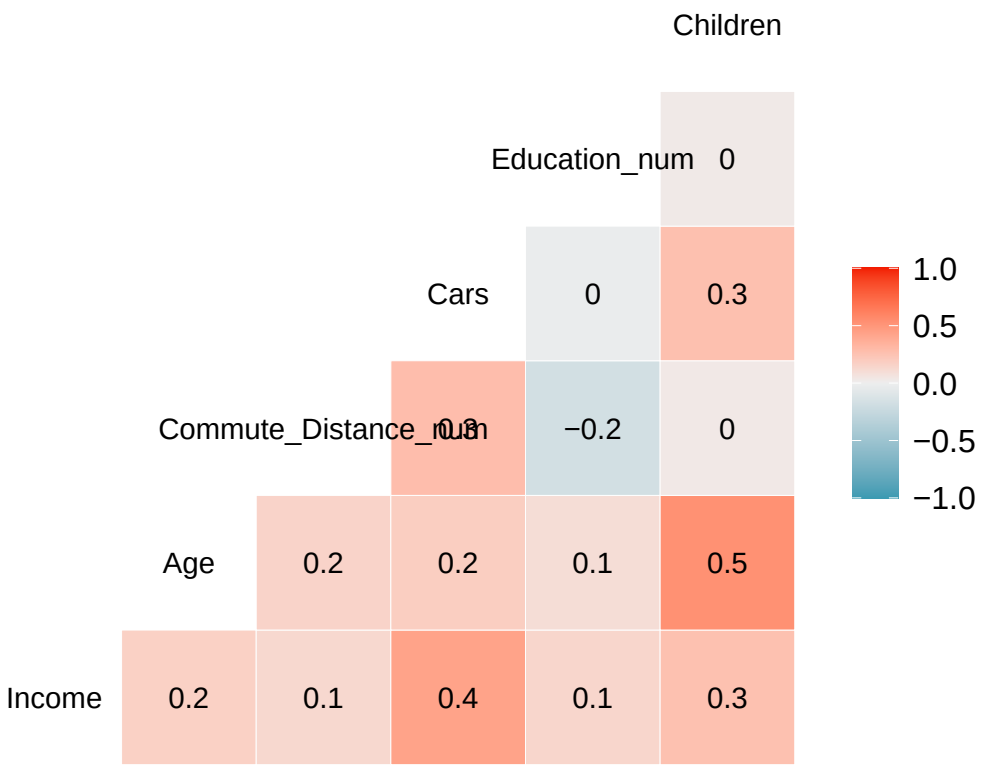
```
cor_commute <- cor(dataset$Commute_Distance_num, dataset$Cars, use = "complete.obs")
print(paste("Corrélation entre Commute_Distance et Cars :", cor_commute))
```

```
## [1] "Corrélation entre Commute_Distance et Cars : 0.279445301098936"
```

Et voila une matrice de correlation avec toutes les valeurs numeriques:

```
numeric_and_ordinal <- dataset %>%
  select(Income, Age, Commute_Distance_num ,Cars , Education_num, Children)
```

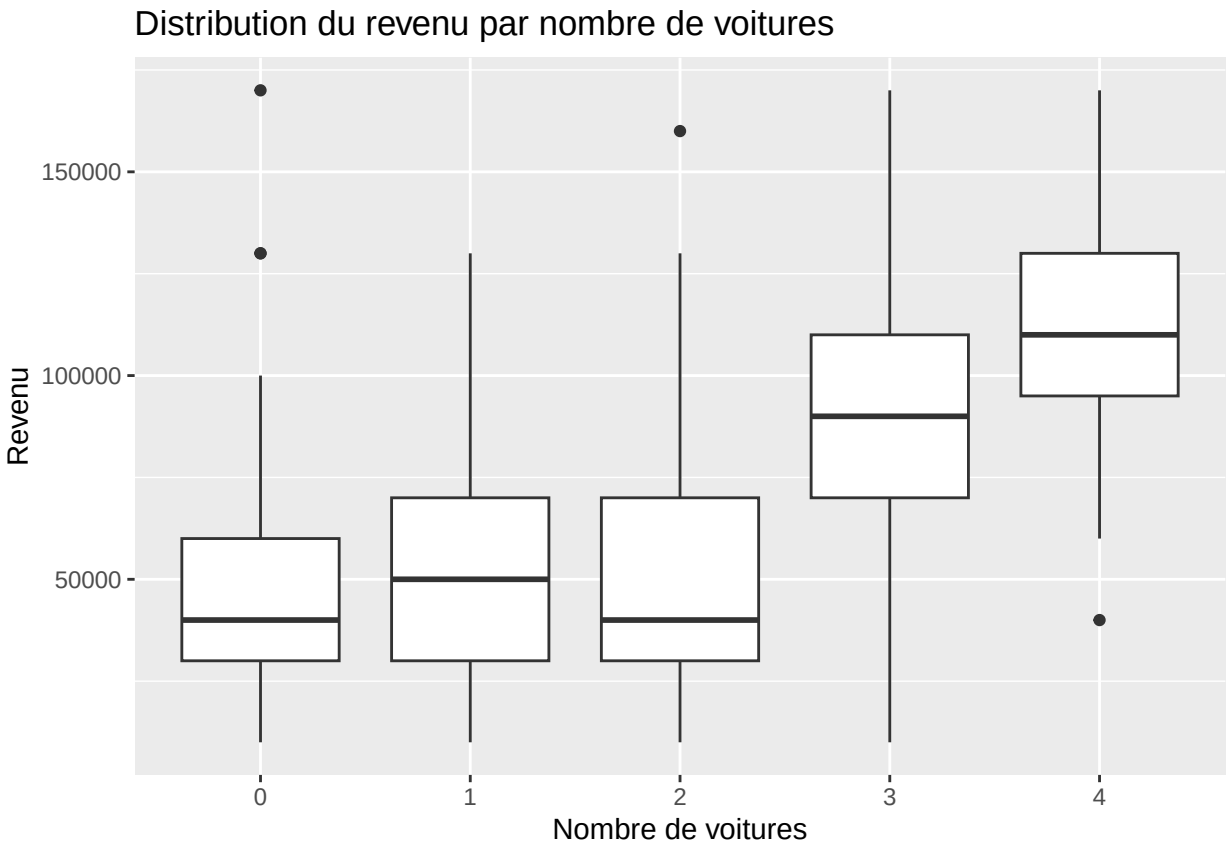
```
ggcorr(
  data = numeric_and_ordinal,
  method = c("pairwise", "pearson"),
  label = TRUE,
  label_size = 4,
  legend.size = 12,
)
```



Conclusion sur la matrice:

Le revenu et la distance domicile-travail influencent fortement la possession de voitures. L'âge a un impact faible. L'éducation est fortement corrélée au revenu, indiquant que les personnes plus diplômées gagnent mieux. Plus de voitures dans les familles nombreuses, mais la corrélation est faible. Donc on ne peut rien dire aussi. Du coup, voila quelques visualisations appropriées pour comprendre la repartition et les correlations des données:

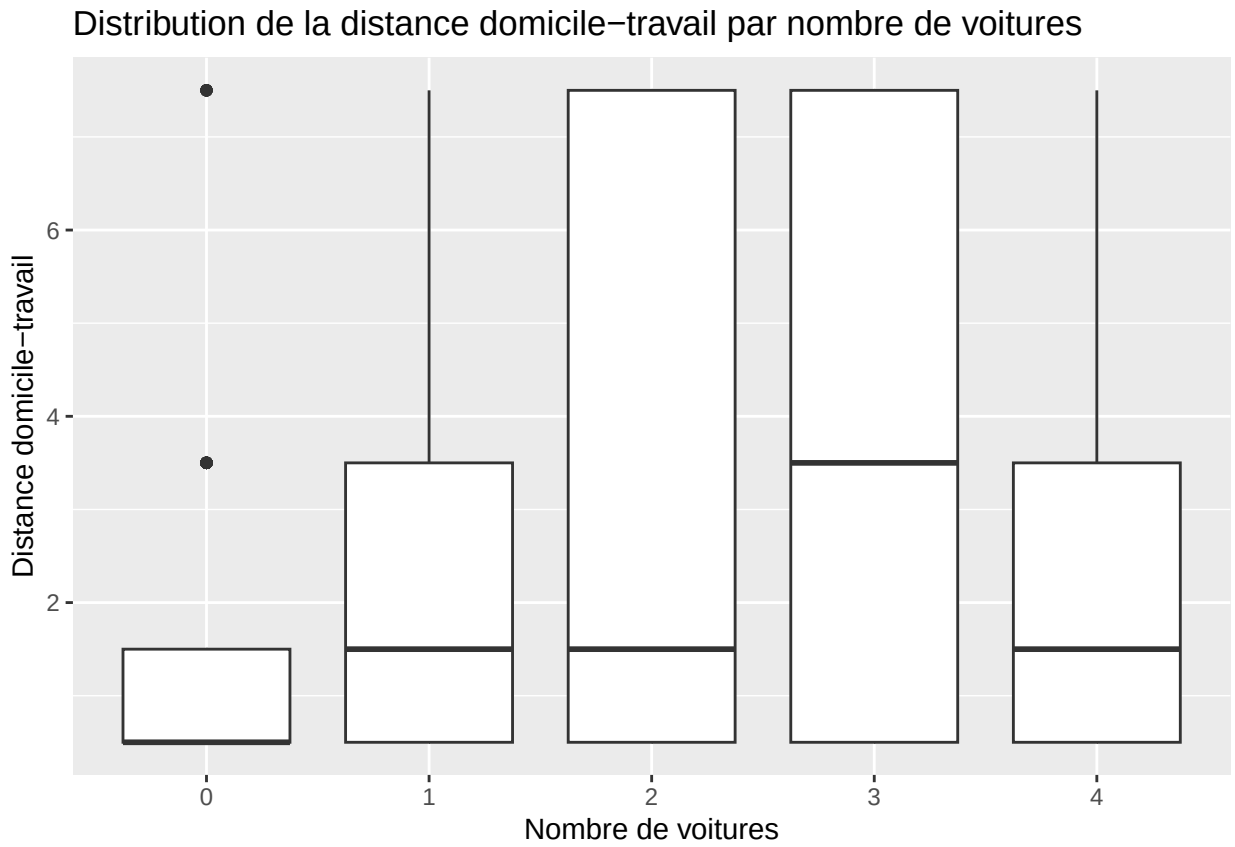
```
ggplot(dataset, aes(x = factor(Cars), y = Income)) +
  geom_boxplot() +
  labs(title = "Distribution du revenu par nombre de voitures", x = "Nombre de voitures", y = "Revenu")
```



-Les clients qui possèdent un plus grand nombre de voitures ont généralement des revenus plus élevés. -Une corrélation modérée (0.43) existe entre ces deux variables, indiquant que les clients à revenu élevé possèdent davantage de voitures. -Une faible corrélation avec l'âge (0.19) suggère que les jeunes adultes possèdent moins de voitures.

Nous allons maintenant examiner la relation entre le nombre de voitures possédées et la distance domicile-travail pour mieux comprendre les comportements de déplacement des clients.

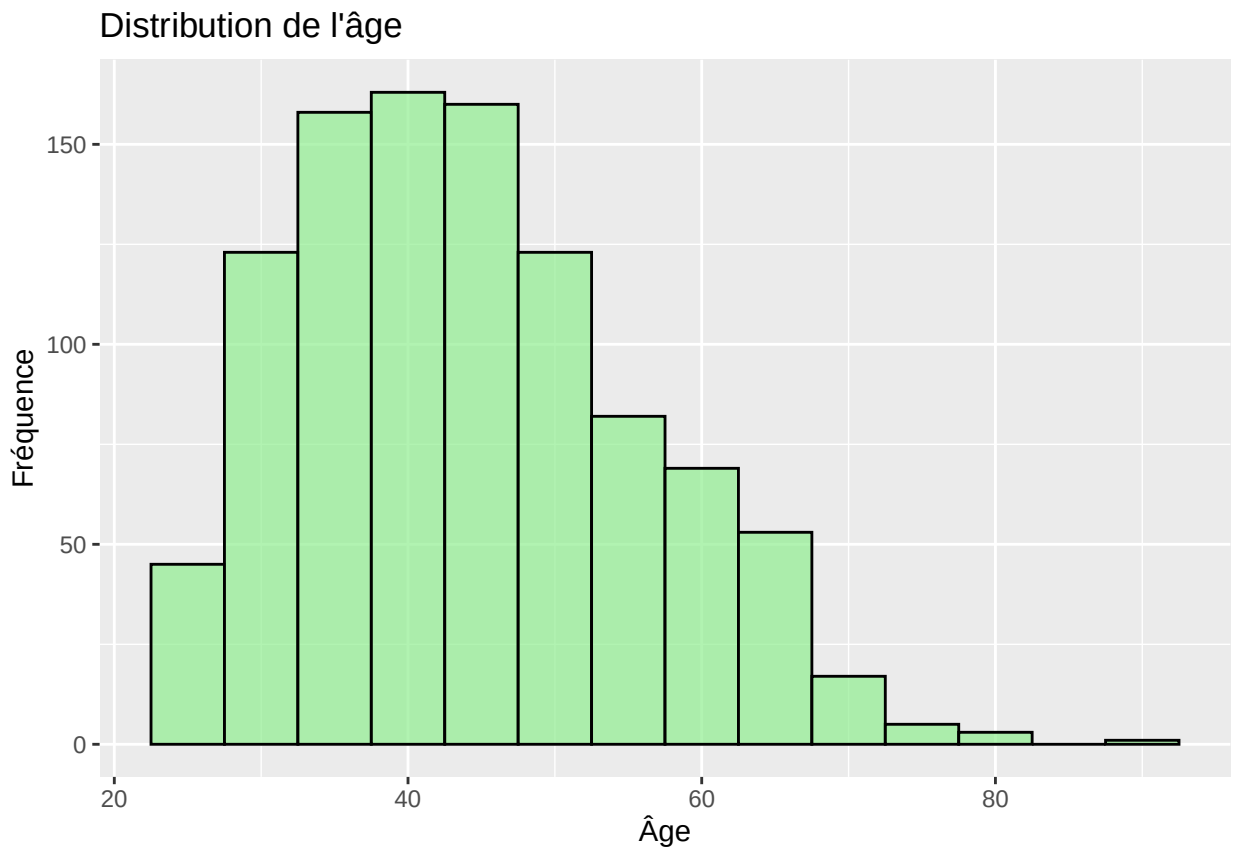
```
ggplot(dataset, aes(x = factor(Cars), y = Commute_Distance_num)) +
  geom_boxplot() +
  labs(title = "Distribution de la distance domicile-travail par nombre de voitures", x = "Nombre de voi
```



-Les clients possédant plus de voitures tendent à parcourir de plus grandes distances pour se rendre au travail. -Une augmentation notable de la médiane des distances est observée avec le nombre de voitures. -Les variations (écarts) dans les distances augmentent également pour les catégories ayant plusieurs voitures, indiquant une diversité de profils.

Nous allons maintenant identifier les groupes d'âge les plus représentés dans l'échantillon :

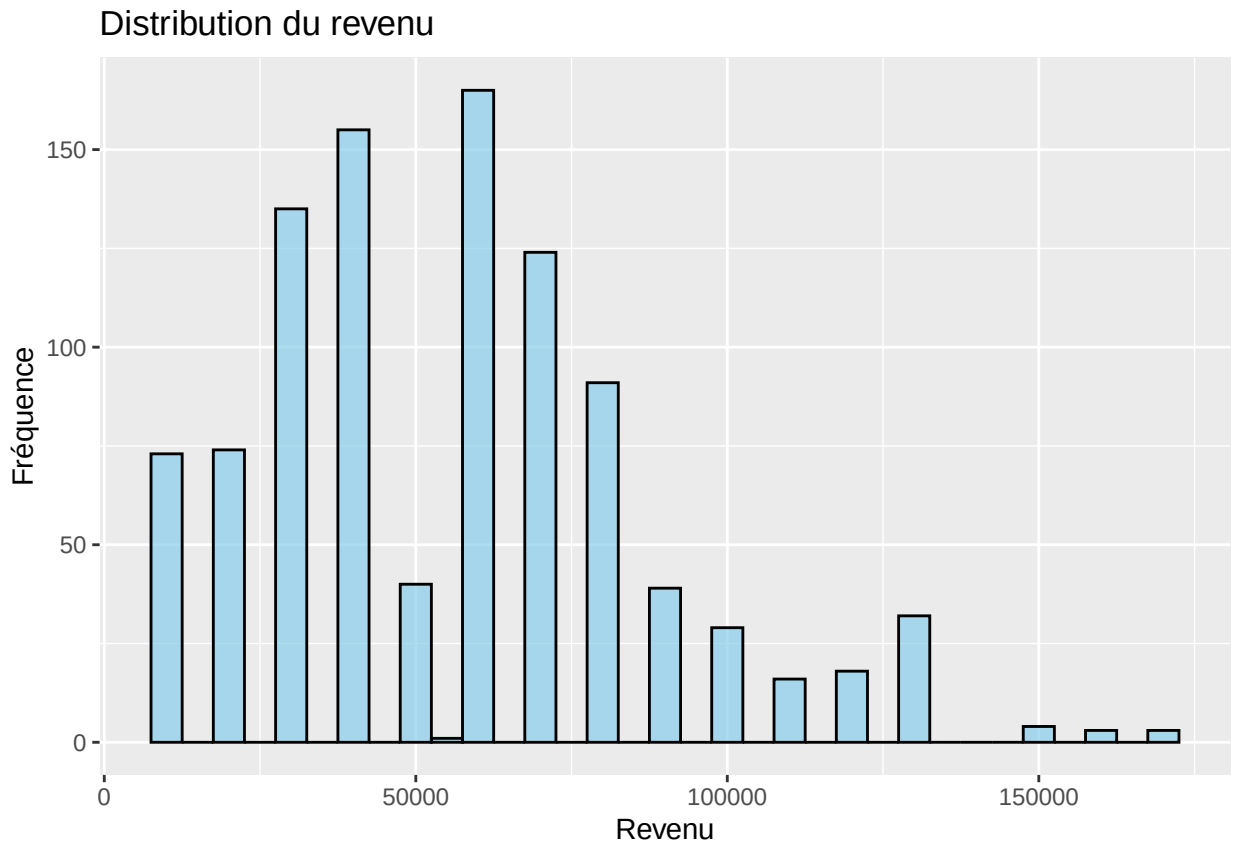
```
ggplot(dataset, aes(x = Age)) +
  geom_histogram(binwidth = 5, fill = "lightgreen", color = "black", alpha = 0.7) +
  labs(title = "Distribution de l'âge", x = "Âge", y = "Fréquence")
```



-La majorité des clients ont entre 35 et 52 ans (intervalle interquartile). -Peu de clients ont moins de 30 ans ou plus de 70 ans. Ce profil d'âge indique une clientèle principalement composée d'adultes d'âge moyen, avec une représentation moindre des jeunes et des seniors.

Maintenant,nous allons analyser la distribution des revenus des clients afin d'identifier les tendances générales et les éventuelles disparités :

```
ggplot(dataset, aes(x = Income)) +
  geom_histogram(binwidth = 5000, fill = "skyblue", color = "black", alpha = 0.7) +
  labs(title = "Distribution du revenu", x = "Revenu", y = "Fréquence")
```



On observe une forte densité dans les tranches de revenus moyens, ce qui suggère que la majorité des clients appartiennent à une catégorie socio-économique moyenne. Les clients avec des revenus très bas ou très élevés sont moins représentés, indiquant une diversité moindre aux extrêmes.

Maintenant, on va faire une segmentation pour avoir plus de precision et d'informations:

```
table(dataset$Marital_Status, dataset$Gender)
```

```
##
##       F   M
##   M 240 299
##   S 250 213
```

```
table(dataset$Occupation, dataset$Home_Owner)
```

```
##
##                   No Yes
##   Clerical        63 115
##   Management      36 137
##   Manual          49  71
##   Professional    97 179
##   Skilled Manual  72 183
```

```
table(dataset$Region, dataset$Gender)
```

```
##
##                    F   M
##   Europe         164 136
##   North America 240 270
##   Pacific         86 106
```

on a créé des tables croisées pour analyser les relations entre :

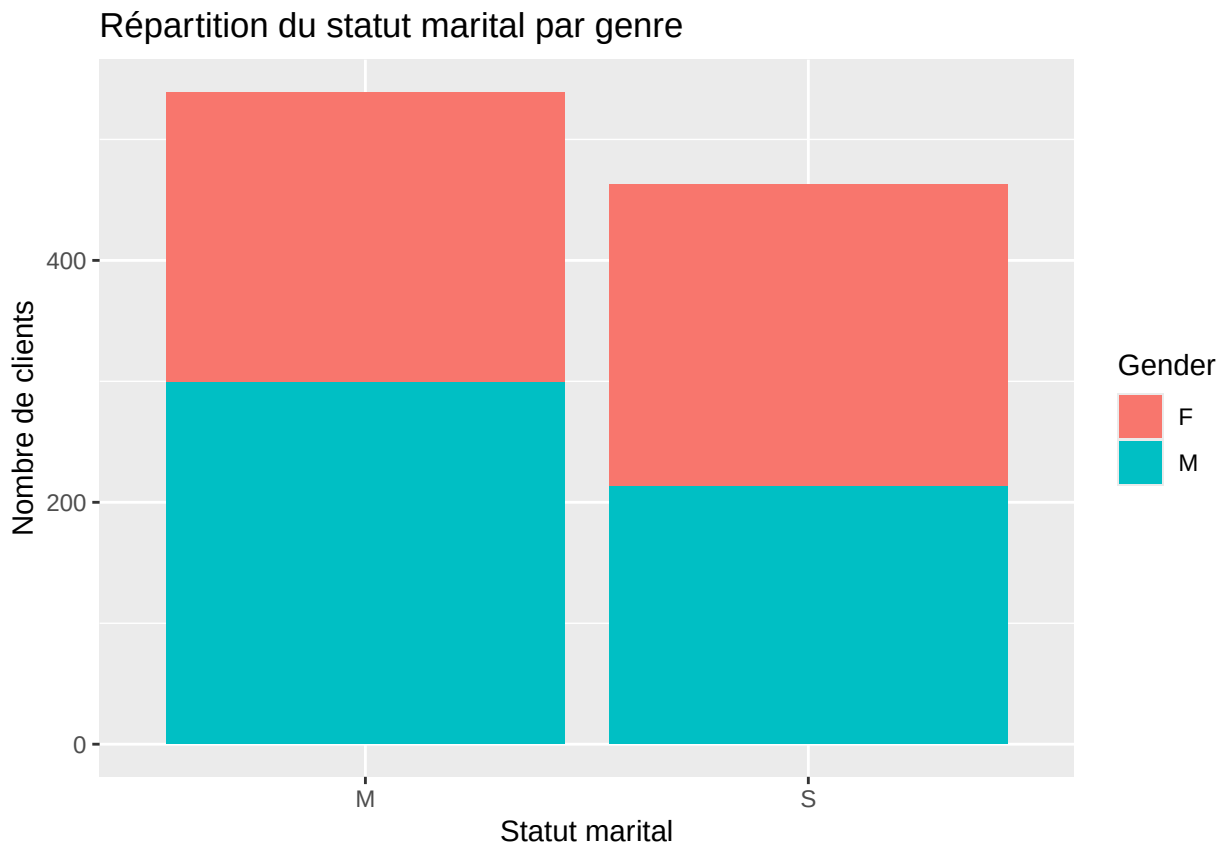
-Marital_Status et Gender. -Occupation et Home_Owner. -Region et Gender.

Analyse des résultats :

-Les genres sont équilibrés, avec quelques variations selon l'état civil et la région. -Les propriétaires sont plus nombreux parmi les professions qualifiées. -Les tendances varient légèrement entre régions et catégories.

Maintenant, nous allons représenter la répartition du statut marital en fonction du genre :

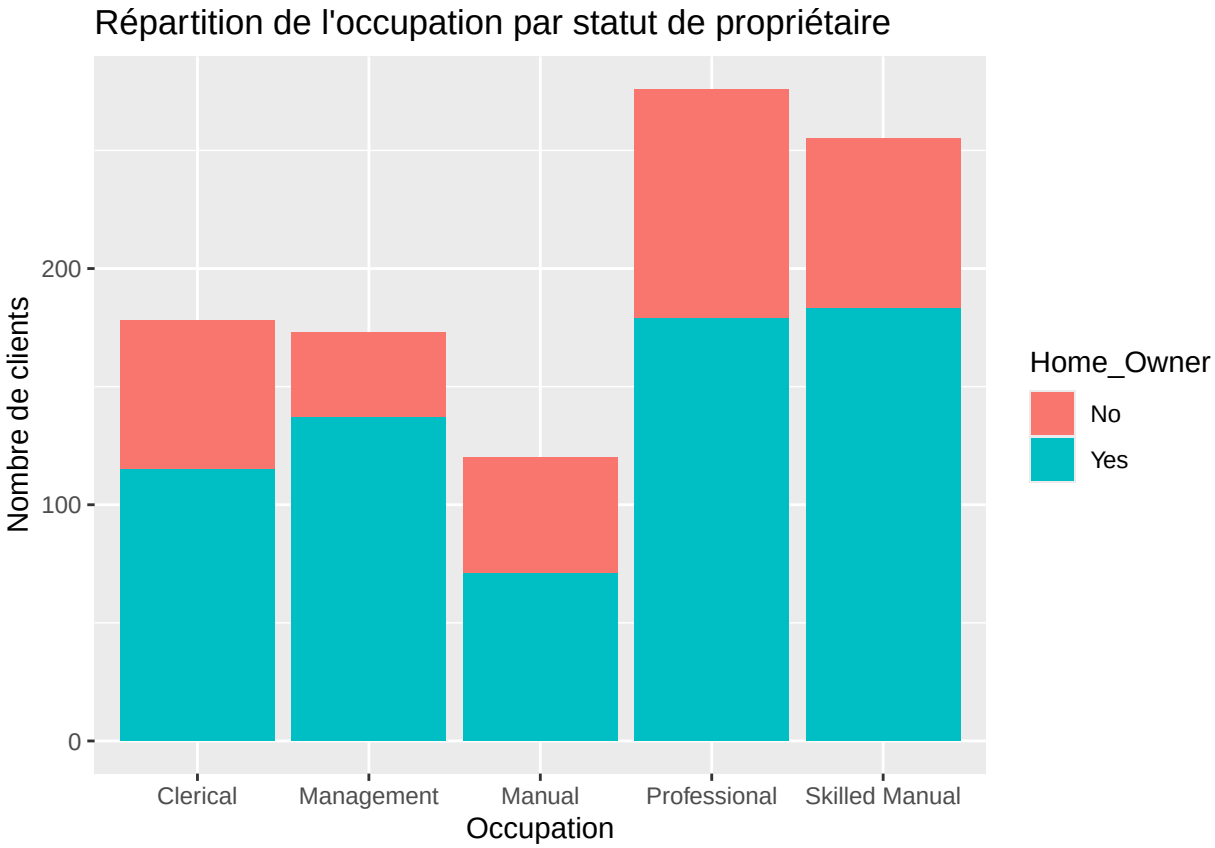
```
ggplot(dataset, aes(x = Marital_Status, fill = Gender)) +
  geom_bar(position = "stack") +
  labs(title = "Répartition du statut marital par genre", x = "Statut marital", y = "Nombre de clients")
```



Le graphique montre la proportion d'hommes et de femmes dans chaque catégorie de statut marital. Les genres sont équilibrés, mais certaines catégories présentent des variations légères (par exemple, plus d'hommes mariés ou plus de femmes célibataires).

Nous allons visualiser la répartition des occupations en fonction du statut de propriétaire ou non :

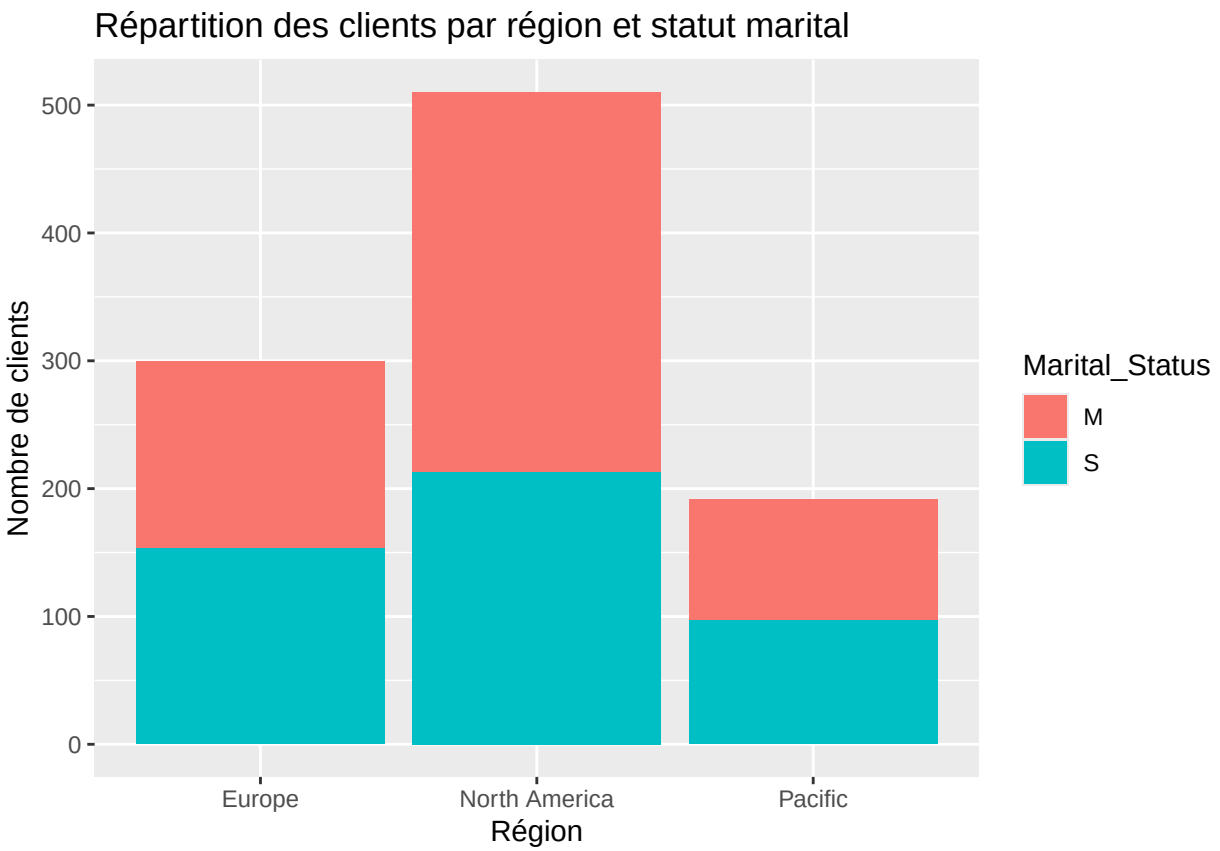
```r
ggplot(dataset, aes(x = Occupation, fill = Home_Owner)) +
  geom_bar(position = "stack") +
  labs(title = "Répartition de l'occupation par statut de propriétaire", x = "Occupation", y = "Nombre
```



Le graphique montre que les propriétaires sont plus nombreux dans toutes les catégories professionnelles, avec une proportion particulièrement élevée chez les professions qualifiées comme “Skilled Manual” et “Professional”. Les locataires sont moins représentés, surtout dans les métiers manuels.

Maintenant nous allons représenter la répartition des clients par région, en fonction de leur statut marital :

```r
ggplot(dataset, aes(x = Region, fill = Marital_Status)) +
  geom_bar(position = "stack") +
  labs(title = "Répartition des clients par région et statut marital", x = "Région", y = "Nombre de cli
```



Le graphique révèle les proportions de clients mariés et célibataires dans chaque région. Certaines régions, comme l'Europe ou le Pacifique, montrent un équilibre, tandis que d'autres peuvent avoir une légère prédominance d'un statut marital.

Et maintenant pour toutes les valeurs categorielles:

```r
library(ggplot2)
library(dplyr)

categorical_vars <- c("Marital_Status", "Gender", "Occupation", "Home_Owner", "Region")
analyze_categorical_vars <- function(dataset, categorical_vars) {
  for (i in 1:(length(categorical_vars)-1)) {
    for (j in (i+1):length(categorical_vars)) {
      var1 <- categorical_vars[i]
      var2 <- categorical_vars[j]

      cat("\nTableau croisé pour", var1, "et", var2, ":\n")
      print(table(dataset[[var1]], dataset[[var2]]))

      ggplot(dataset, aes_string(x = var1, fill = var2)) +
        geom_bar(position = "stack") +
        labs(title = paste("Répartition de", var1, "par", var2), x = var1, y = "Nombre de clients") +
        theme_minimal() +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))

      Sys.sleep(1)
    }
```

```
  }
}

analyze_categorical_vars(dataset, categorical_vars)
```

```
##
## Tableau croisé pour Marital_Status et Gender :
##
##       F   M
##   M 240 299
##   S 250 213

##
## Tableau croisé pour Marital_Status et Occupation :
##
##     Clerical Management Manual Professional Skilled Manual
##   M       86         98     43          167            145
##   S       92         75     77          109            110
##
## Tableau croisé pour Marital_Status et Home_Owner :
##
##      No Yes
##   M  97 442
##   S 220 243
##
## Tableau croisé pour Marital_Status et Region :
##
##     Europe North America Pacific
##   M    147           297      95
##   S    153           213      97
##
## Tableau croisé pour Gender et Occupation :
##
##     Clerical Management Manual Professional Skilled Manual
##   F       95         78     69          126            122
##   M       83         95     51          150            133
##
## Tableau croisé pour Gender et Home_Owner :
##
##      No Yes
##   F 156 334
##   M 161 351
##
## Tableau croisé pour Gender et Region :
##
##     Europe North America Pacific
##   F    164           240      86
##   M    136           270     106
##
## Tableau croisé pour Occupation et Home_Owner :
##
##                  No Yes
##   Clerical       63 115
```

```
##   Management      36 137
##   Manual          49  71
##   Professional    97 179
##   Skilled Manual  72 183
##
## Tableau croisé pour Occupation et Region :
##
##                  Europe North America Pacific
##   Clerical          114            39      25
##   Management         16           108      49
##   Manual            103             4      13
##   Professional       37           175      64
##   Skilled Manual     30           184      41
##
## Tableau croisé pour Home_Owner et Region :
##
##       Europe North America Pacific
##   No     105           141      71
##   Yes    195           369     121
```

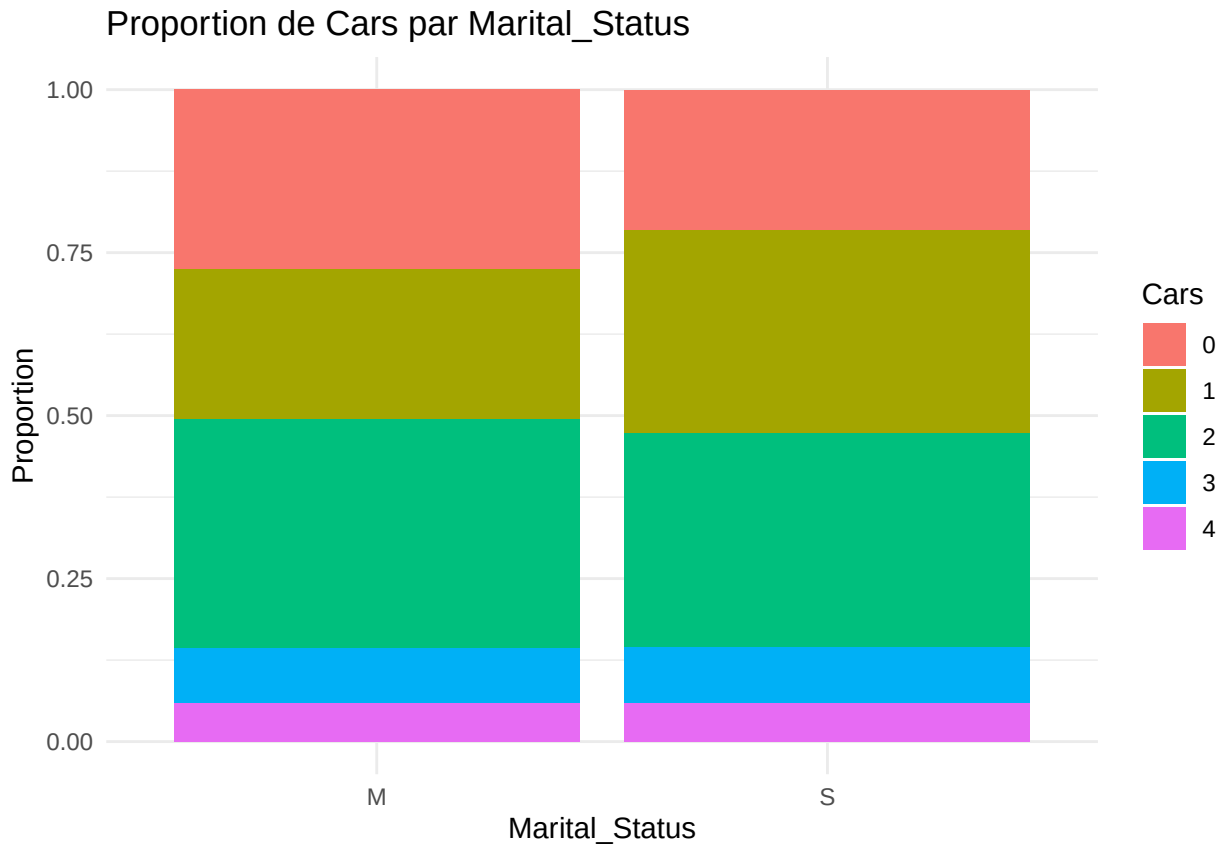
Les résultats des segments les plus pertinents proviendront de combinaisons où l'on observe une forte concentration ou des différences claires entre les catégories (par exemple, certaines professions étant majoritaires dans des régions spécifiques ou les propriétaires de maison étant plus nombreux dans certaines tranches de statut marital). Ces segments peuvent guider des stratégies de ciblage plus affinées.

Nous allons réaliser une segmentation sur des variables catégorielles pour explorer leur relation avec la possession de voitures ou non:

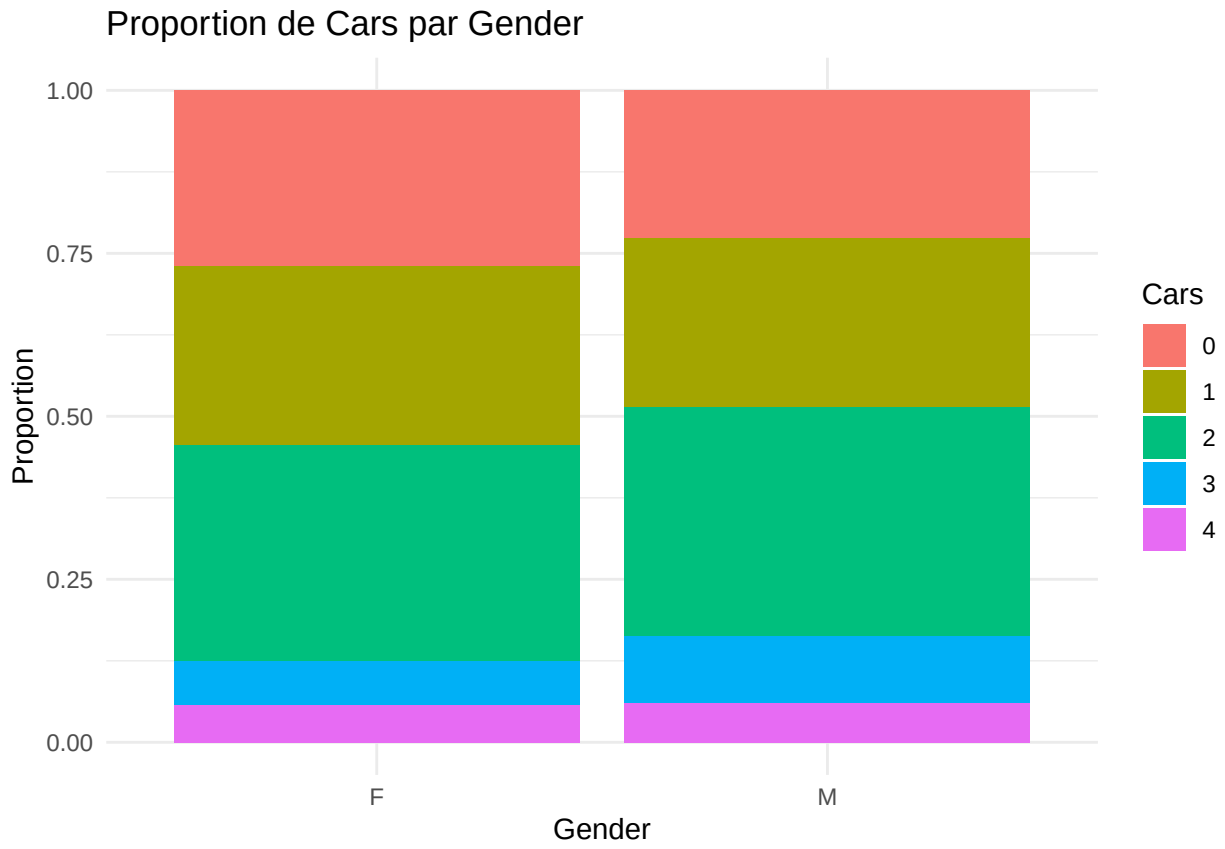
```r
analyze_categorical <- function(data, target, cat_var) {
  cross_tab <- table(data[[target]], data[[cat_var]])
  print(cross_tab)

  cross_tab_df <- as.data.frame(cross_tab)

  ggplot(cross_tab_df, aes(x = Var2, y = Freq, fill = Var1)) +
    geom_bar(stat = "identity", position = "fill") +
    labs(
      x = cat_var,
      y = "Proportion",
      fill = target,
      title = paste("Proportion de", target, "par", cat_var)
    ) +
    theme_minimal()
}

cat_vars <- c("Marital_Status", "Gender", "Home_Owner", "Region", "Occupation")

for (cat_var in cat_vars) {
  print(paste("Analyse pour la variable :", cat_var))
  plot <- analyze_categorical(dataset, "Cars", cat_var)
  print(plot)
}
```

```
## [1] "Analyse pour la variable : Marital_Status"
```

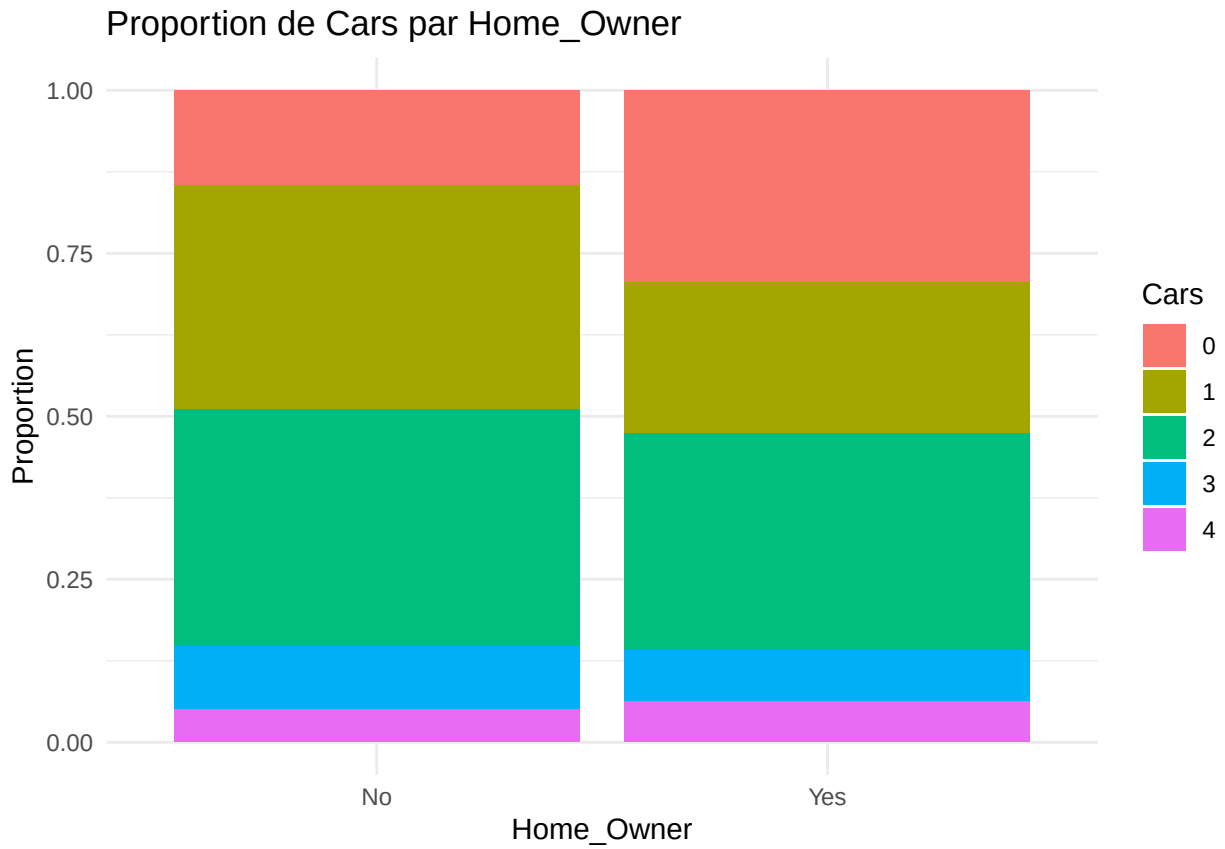
```
##
##       M   S
##   0 148 100
##   1 124 144
##   2 190 152
##   3  45  40
##   4  32  27
```



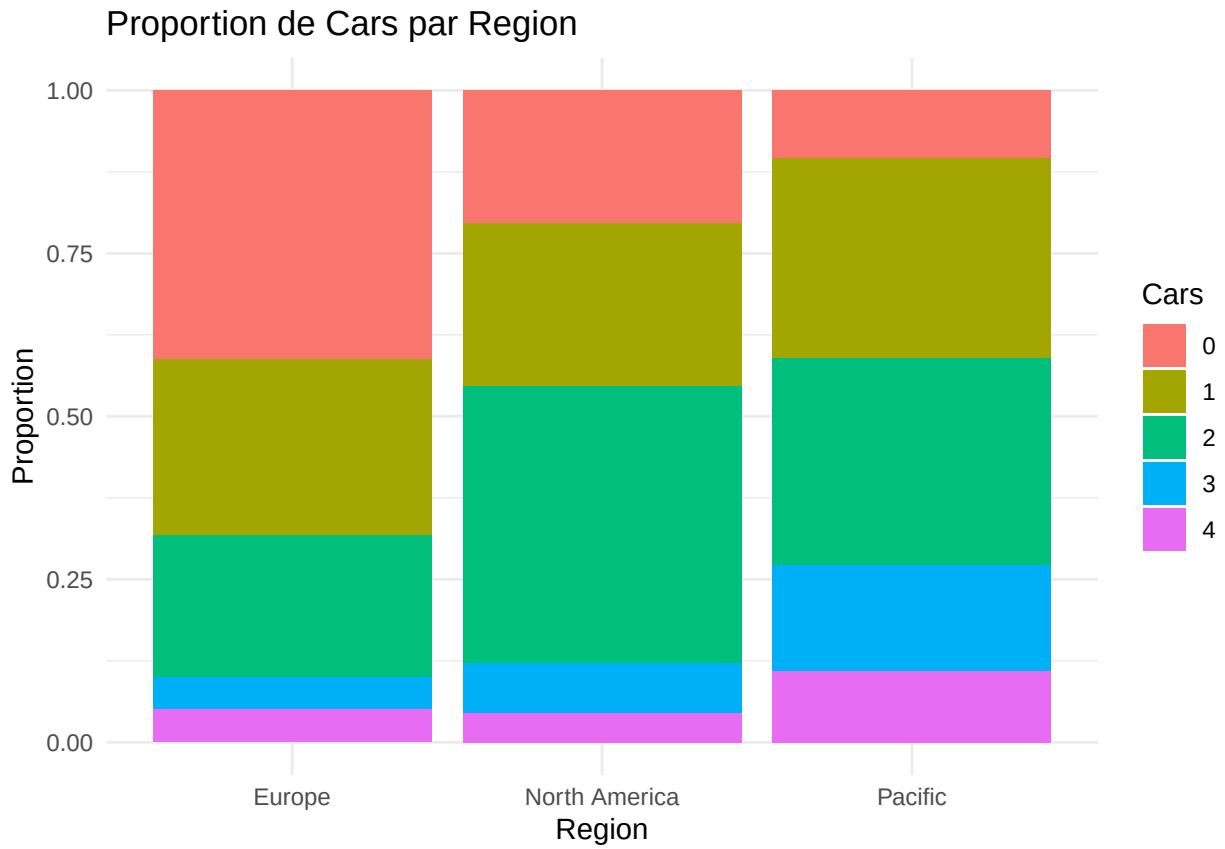

```
## [1] "Analyse pour la variable : Gender"
##
##       F   M
##   0 132 116
##   1 135 133
##   2 162 180
##   3  33  52
##   4  28  31
```



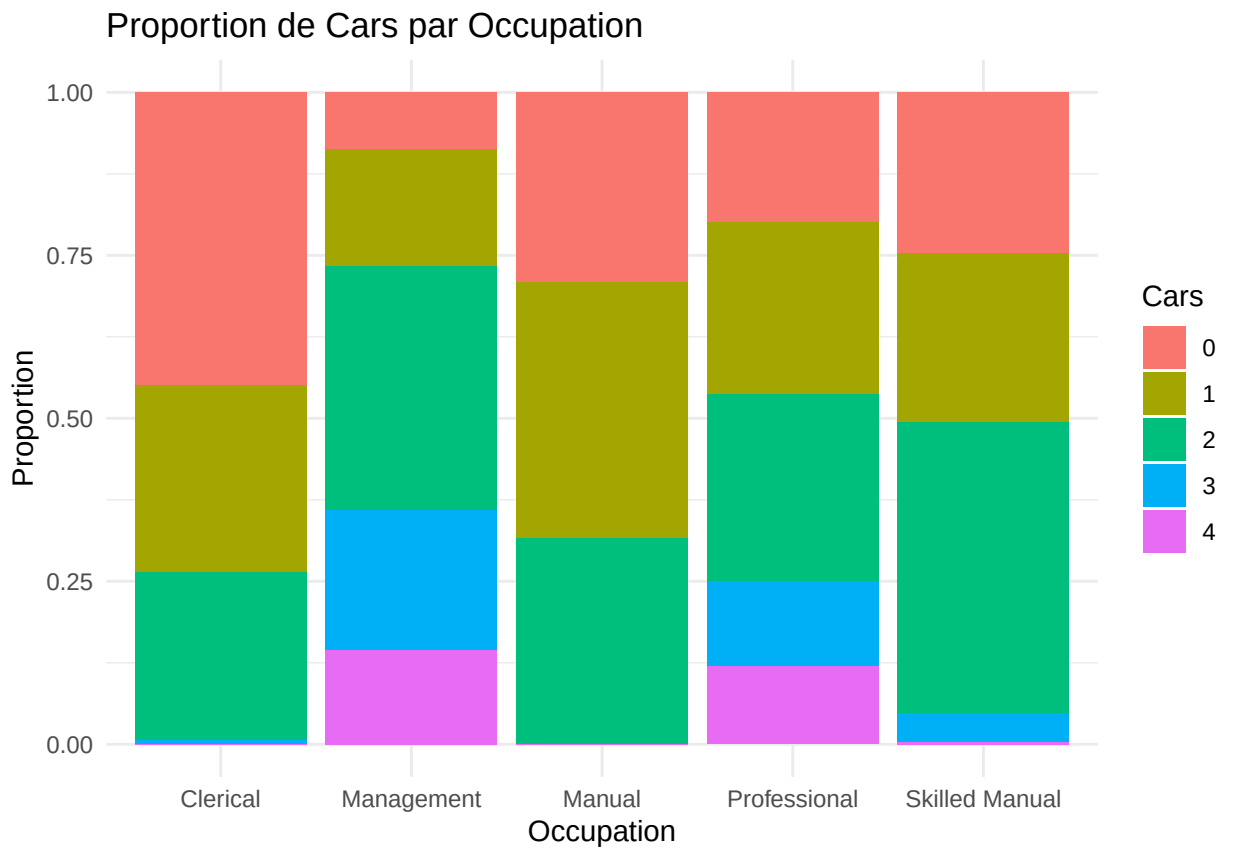

```
## [1] "Analyse pour la variable : Home_Owner"
##
##      No Yes
##   0  46 202
##   1 109 159
##   2 115 227
##   3  31  54
##   4  16  43
```




```
## [1] "Analyse pour la variable : Region"
##
##     Europe North America Pacific
##   0    124           104      20
##   1     81           128      59
##   2     65           216      61
##   3     15            39      31
##   4     15            23      21
```




```
## [1] "Analyse pour la variable : Occupation"
##
##     Clerical Management Manual Professional Skilled Manual
##   0       80         15     35           55             63
##   1       51         31     47           73             66
##   2       46         65     38           79            114
##   3        1         37      0           36             11
##   4        0         25      0           33              1
```



Pour les voitures, on voit clairement une difference partout a peu pres. Les propriétaires achètent bien plus de vélos que les locataires. L'Amérique du Nord domine dans les achats de vélos, suivie de la région Pacifique. Les professionnels et travailleurs qualifiés achètent le plus, tandis que les employés administratifs achètent peu. Les hommes achètent légèrement plus que les femmes. Pas de différence significative entre mariés et célibataires.

Rapport d'analyse des données sur la possession de voitures

1. Introduction L'objectif de cette analyse était de comprendre les facteurs influençant la possession de voitures, d'identifier les segments de clients les plus susceptibles de posséder une voiture, et de formuler des recommandations marketing pour augmenter les ventes ou améliorer les services associés. Nous avons travaillé à partir d'un fichier de données contenant des informations sur les clients, telles que leur genre, statut marital, profession, région, statut de propriétaire, et le nombre de voitures possédées.

2. Étapes de l'analyse

2.1 Chargement des données Les données ont été importées dans R à l'aide de la fonction `read_excel()`. Nous avons vérifié leur structure et effectué un prétraitement pour gérer les doublons et les valeurs manquantes. Les valeurs manquantes dans les variables numériques, comme `Income`, ont été remplacées par la moyenne. Les valeurs manquantes dans les variables catégorielles ont été remplacées par la modalité la plus fréquente.

2.2 Analyse de la distribution de la variable cible Nous avons commencé par examiner la répartition du nombre de voitures possédées (`Cars`) par les clients. Une visualisation sous forme d'histogramme a montré que la majorité des clients possèdent une ou deux voitures, avec un nombre plus réduit de clients possédant trois voitures ou plus.

2.3 Étude des relations avec d'autres variables Nous avons analysé les corrélations entre `Cars` et les variables numériques telles que `Income`, `Age`, et `Commute_Distance_num`.

- Une corrélation modérée positive entre le revenu (`Income`) et le nombre de voitures possédées.
- Une faible corrélation positive entre l'âge (`Age`) et le nombre de voitures.
- Une corrélation positive entre la distance domicile-travail (`Commute_Distance_num`) et le nombre de voitures.

Des visualisations telles que des boxplots ont été réalisées pour évaluer les tendances.

2.4 Segmentation des clients Nous avons segmenté les clients en fonction de variables catégorielles comme `Marital_Status`, `Gender`, `Occupation`, `Home_Owner`, et `Region`. Pour chaque groupe, nous avons calculé le nombre total de clients, le nombre de possesseurs de voitures, et la proportion correspondante.

Des tableaux croisés et des graphiques ont permis de visualiser les résultats, mettant en évidence les segments les plus intéressants, tels que les clients mariés, les propriétaires de maison, et les habitants de régions spécifiques.

2.5 Identification des caractéristiques des possesseurs de voitures

- Revenus élevés : Les clients avec des revenus plus élevés possèdent généralement plus de voitures.
- Statut marital marié : Les clients mariés ont tendance à posséder davantage de voitures, probablement en raison des besoins familiaux.
- Propriétaires de maison : Les propriétaires sont plus susceptibles de posséder plusieurs voitures.
- Profession : Les professionnels et les travailleurs manuels qualifiés possèdent plus de voitures.
- Région : Certaines régions montrent une proportion plus élevée de clients possédant plusieurs voitures.

3. Principales conclusions

- Le revenu est un facteur déterminant : Les clients à revenu élevé possèdent généralement plus de voitures.
- Les clients mariés et propriétaires de maison sont plus susceptibles de posséder plusieurs voitures, indiquant que les besoins familiaux et la stabilité résidentielle influencent la possession de véhicules.
- La distance domicile-travail influence la possession de voitures : Les clients parcourant de plus longues distances possèdent souvent plus de voitures.
- Les professionnels et travailleurs qualifiés possèdent davantage de voitures, probablement en raison de revenus plus élevés et de besoins de mobilité spécifiques.

4. Recommandations marketing

- Cibler les familles et les clients mariés avec des offres groupées, comme des réductions sur une deuxième voiture ou des accessoires familiaux.
- Proposer des modèles adaptés aux besoins des professionnels et des travailleurs qualifiés, mettant en avant le confort, la fiabilité, et les performances.
- Adapter les campagnes marketing par région en tenant compte des spécificités locales, notamment dans les régions où la possession de plusieurs voitures est plus courante.

- Offrir des solutions pour les propriétaires de maison, comme des options de financement avantageuses ou des services après-vente personnalisés.
- Mettre en avant les avantages pour les clients parcourant de longues distances, tels que des véhicules économes en carburant ou des programmes d'entretien pour les gros rouleurs.

5. Conclusion Cette analyse a permis d'identifier les groupes les plus susceptibles de posséder plusieurs voitures et de proposer des stratégies marketing adaptées. Une approche ciblée, basée sur les caractéristiques des possesseurs de voitures et les spécificités régionales, devrait permettre d'améliorer significativement les ventes et la satisfaction client.

Comparaison des résultats entre l'analyse des vélos et des voitures

Similarités

- Revenu :
  - Pour les vélos : Les clients avec des revenus moyens à élevés sont plus susceptibles d'acheter des vélos haut de gamme ou électriques.
  - Pour les voitures : Les clients à revenu élevé possèdent plus de voitures.
- Profession : Les professionnels et les travailleurs qualifiés sont des segments importants pour les deux produits.
- Segmentation régionale : Les campagnes marketing doivent être adaptées en fonction des spécificités locales dans les deux cas.

Différences

- Statut marital :
  - Pour les vélos : Les célibataires et les jeunes adultes sont plus enclins à acheter des vélos.
  - Pour les voitures : Les clients mariés sont plus susceptibles de posséder plusieurs voitures.
- Utilisation urbaine vs rurale :
  - Pour les vélos : Plus populaires en zones urbaines pour des trajets courts.
  - Pour les voitures : Essentielles en zones rurales ou pour de longues distances.
- Facteur d'âge :
  - Pour les vélos : L'âge a une influence notable, avec une clientèle plus jeune.
  - Pour les voitures : L'âge a un impact faible.

Conclusions sur les différences

- Stratégies marketing distinctes :
  - Les campagnes pour les vélos doivent cibler les jeunes adultes urbains, en mettant l'accent sur la mobilité douce et les modes de vie actifs.
  - Pour les voitures, cibler les familles, les professionnels, et les résidents des zones nécessitant une mobilité accrue.
- Messages publicitaires adaptés :
  - Pour les vélos : Insister sur la santé, l'écologie, et le côté pratique en ville.
  - Pour les voitures : Souligner le confort, la fiabilité, la sécurité, et les avantages pour les familles.

Points non pertinents pour les voitures mais pertinents pour les vélos :

- Cibler les jeunes adultes et les célibataires : Ce segment est crucial pour les vélos, moins pour les voitures.
- Accent sur les accessoires de loisirs et les modes de vie actifs : Les vélos sont associés à des activités de plein air et un mode de vie sain, ce qui est moins pertinent pour les voitures.