Analyse du prix des maisons avec R

1. Introduction

Ce rapport présent une analyse exploratoire du jeu de données **challenge sale price** réalisée avec le langage R.
L'objectif est d'identifier les facteurs qui influencent le plus le prix de vente des maisons et de préparer les données pour une éventuelle modélisation prédictive.

---

2. Packages utilisés

Les packages suivants ont été utilisés pour la manipulation, la visualisation et l'exploration des données :

- **tidyverse** – manipulation et visualisation des données
- **VIM** – visualisation des valeurs manquantes
- **corrplot** – analyse de corrélation
- **skimr** – statistiques descriptives

```
install.packages("tidyverse")
install.packages("VIM")
install.packages("corrplot")
install.packages("skimr")

library(tidyverse)
library(VIM)
library(corrplot)
library(skimr)
library(readr)
```

3. Chargement et Analyse exploratoire des données

Le fichier `train.csv` a été importé dans R à l'aide de la fonction `read.csv()`.

```
train <- read.csv("C:\\Users\\Admin\\Documents\\train.csv", stringsAsFactors =
FALSE)


#Explorer le dataset
head(train)
str(train)
glimpse(train)
```

Le dataset contient **1460 observations** et **81 variables** décrivant les caractéristiques des maisons ainsi que leur prix de vente (SalePrice)

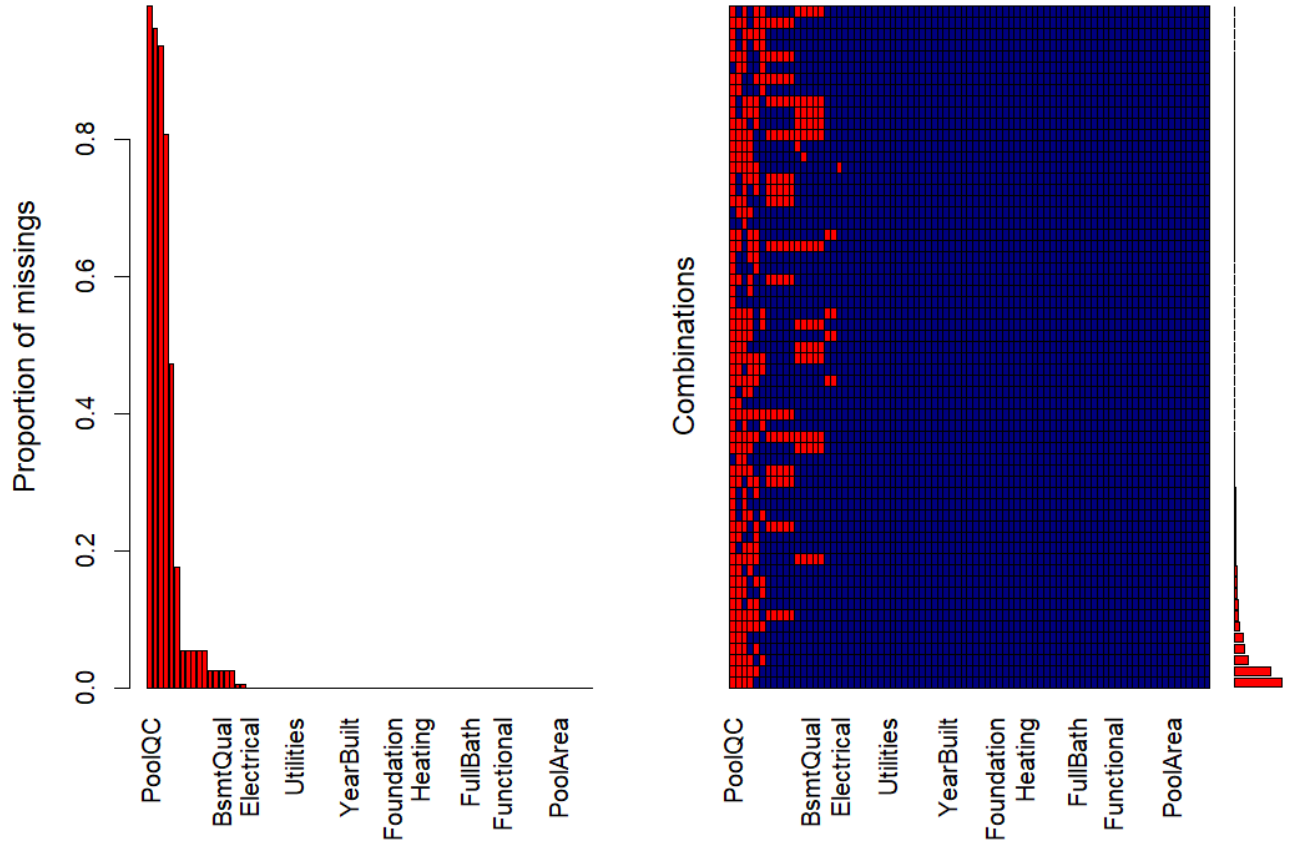
```
#Statistiques descriptives
skim(train)
```

4. Nettoyage et préparation des données

4.1. Visualisation des valeurs manquantes

```
VIM::aggr(train, col = c('navyblue', 'red'), numbers = TRUE, sortVars = TRUE)
```

```
#Resultat

Variables sorted by number of missings:
      Variable        Count
        PoolQC 0.9952054795
   MiscFeature 0.9630136986
         Alley 0.9376712329
         Fence 0.8075342466
   FireplaceQu 0.4726027397
   LotFrontage 0.1773972603
    GarageType 0.0554794521
   GarageYrBlt 0.0554794521
  GarageFinish 0.0554794521
    GarageQual 0.0554794521
    GarageCond 0.0554794521
  BsmtExposure 0.0260273973
  BsmtFinType2 0.0260273973
      BsmtQual 0.0253424658
      BsmtCond 0.0253424658
  BsmtFinType1 0.0253424658
    MasVnrType 0.0054794521
    MasVnrArea 0.0054794521
    Electrical 0.0006849315
```



4.2. Nettoyage des données

Certaines variables contiennent de nombreuses valeurs manquantes (PoolQC, Alley, Fence, MiscFeature…).
Les colonnes **catégorielles** avec des valeurs manquantes ont été remplacées par "None", et les colonnes **numériques** par 0.

```
# Colonnes où NA = "None"

train <- train %>%
  mutate(
    across(
      c("Alley", "BsmtQual", "BsmtCond", "BsmtExposure", "BsmtFinType1",
        "BsmtFinType2", "FireplaceQu", "GarageType", "GarageFinish",
        "GarageQual", "GarageCond", "PoolQC", "Fence",
"MiscFeature","MasVnrType","Electrical"),
      ~ ifelse(is.na(.), "None", .)))


# Colonnes numériques où NA = 0

train <- train %>%
  mutate(
    across(
      c("BsmtFinSF1", "BsmtFinSF2", "BsmtUnfSF", "TotalBsmtSF","MasVnrArea",
        "BsmtFullBath", "BsmtHalfBath", "GarageCars",
"GarageArea","LotFrontage","GarageYrBlt"),
      ~ ifelse(is.na(.), 0, .)))
```

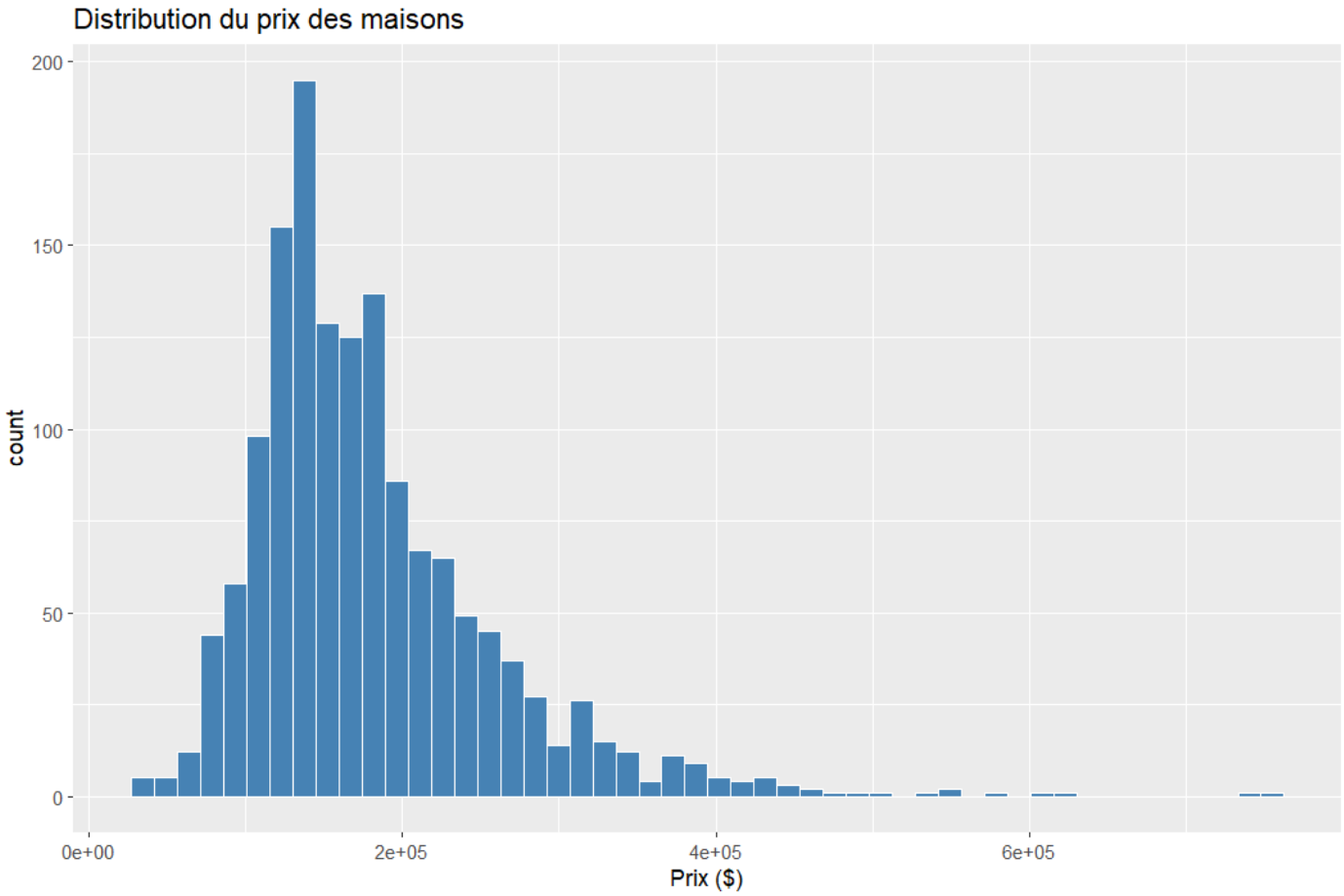
5. Statistiques descriptives et indicateurs clés (KPIs)

5.1. Distribution du prix de vente (SalePrice)

```
summary(train$SalePrice)

#Resultat
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  34900  129975  163000  180921  214000  755000
```

Le prix moyen est d'environ 180 000 $, avec une médiane autour de 163 000 $.

```
#Visualisation
ggplot(train, aes(x = SalePrice)) +
  geom_histogram(bins = 50, fill = "steelblue", color = "white") +
  labs(title = "Distribution du prix des maisons", x = "Prix ($)")
```



La distribution est **asymétrique à droite** — quelques maisons très chères tirent la moyenne vers le haut.

5.2. Prix moyen par type de bâtiment

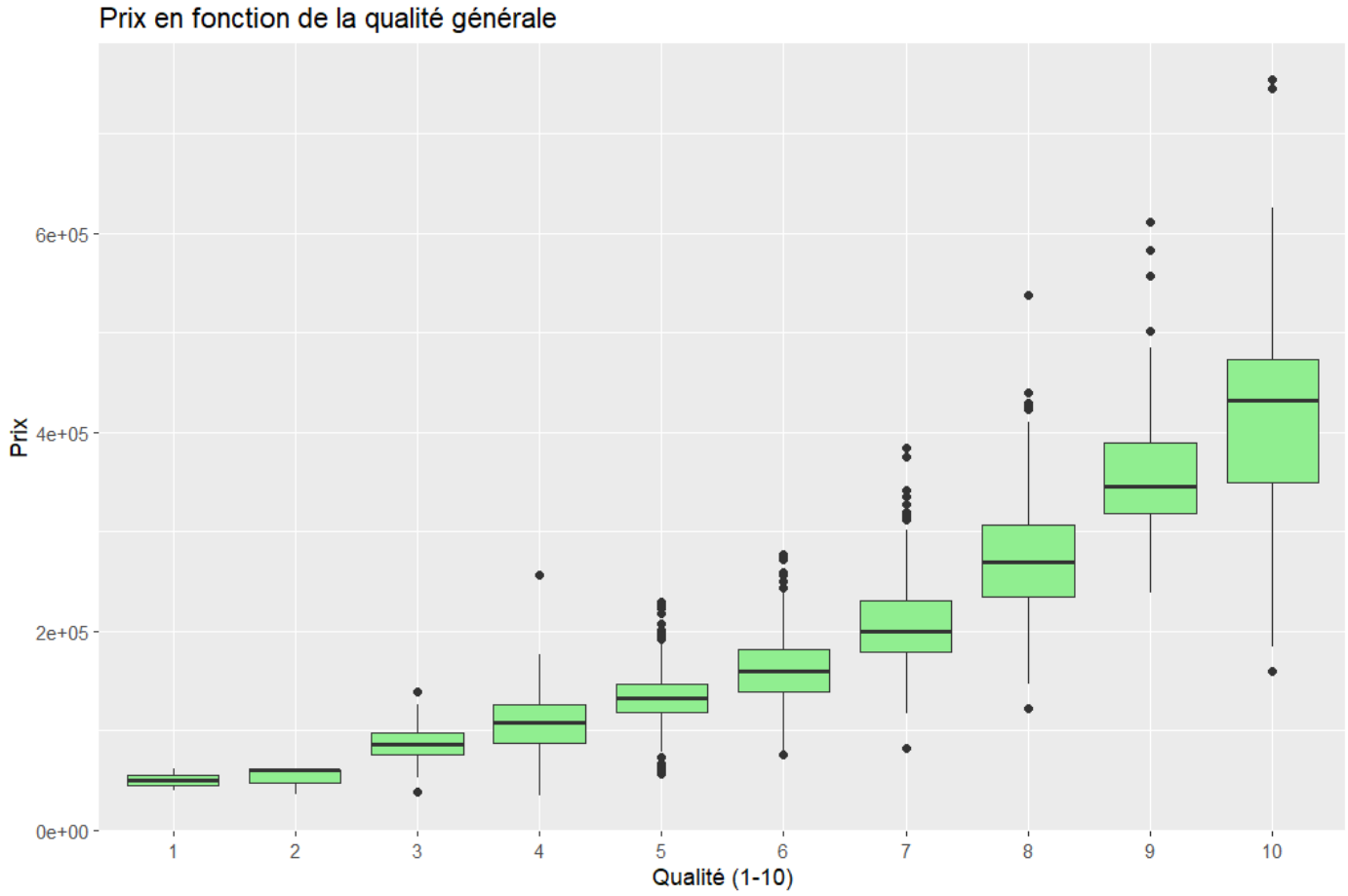
```
train %>%
  group_by(BldgType) %>%
  summarise(avg_price = mean(SalePrice), count = n()) %>%
  arrange(desc(avg_price))
```

```
#Résultat
 BldgType avg_price count
   <chr>       <dbl> <int>
1 1Fam       185764.  1220
2 TwnhsE     181959.   114
3 Twnhs      135912.    43
4 Duplex     133541.    52
5 2fmCon     128432.    31
```

Les maisons individuelles (`1Fam`) sont à la fois les plus chères et les plus représentées dans le dataset.

5.3. Prix en fonction de la qualité générale de la maison

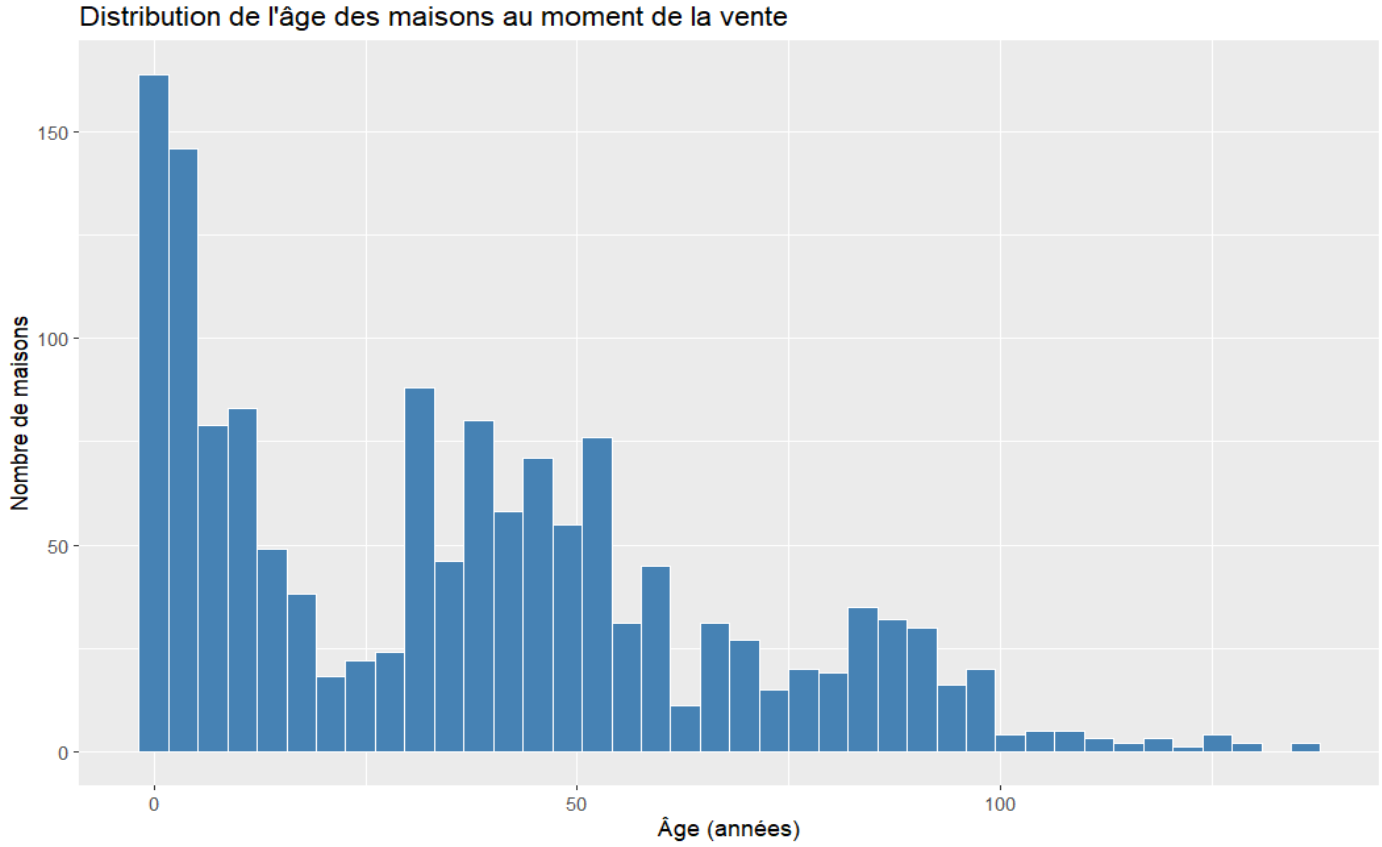
```
# Visualisation
ggplot(train, aes(x = factor(OverallQual), y = SalePrice)) +
  geom_boxplot(fill = "lightgreen") +
  labs(title = "Prix en fonction de la qualité générale", x = "Qualité (1-10)", y
= "Prix")
```



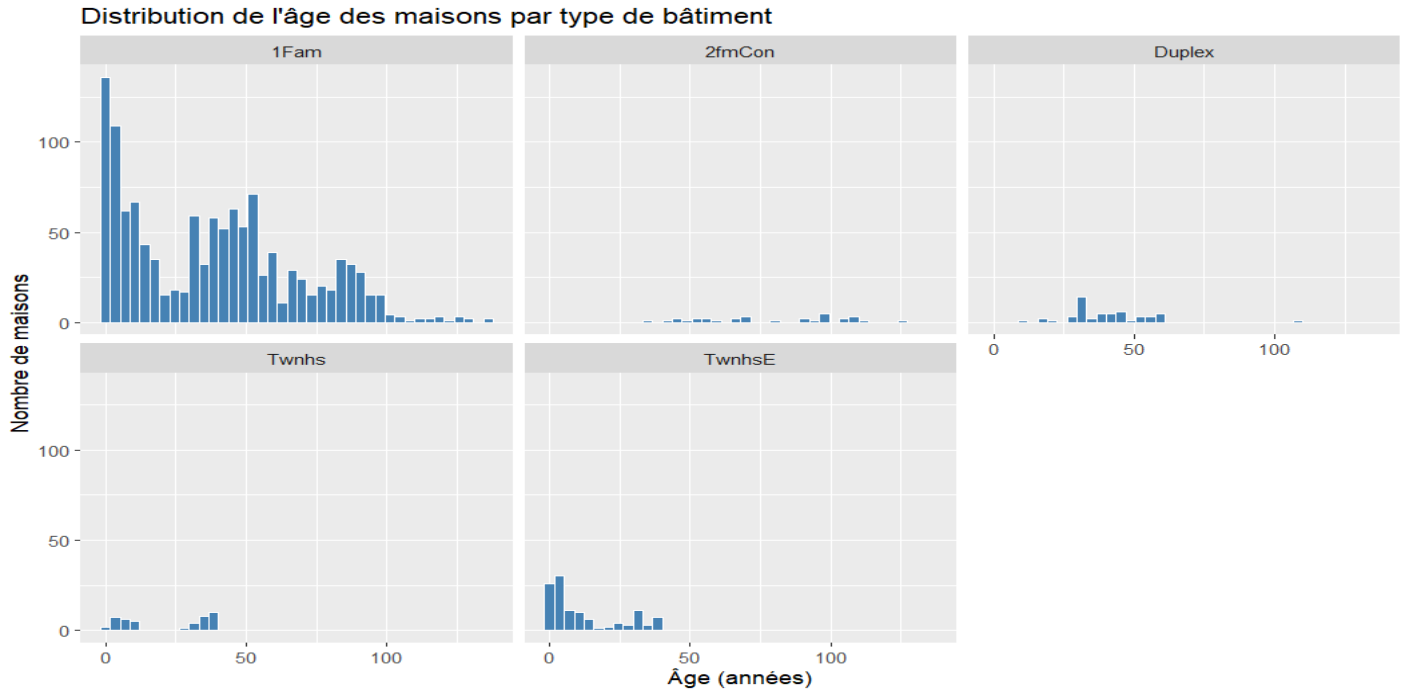
5.4. Répartition de l'âge des maisons

```
# Créer la colonne "HouseAge"
train <- train %>%
  mutate(HouseAge = YrSold - YearBuilt)


# Statistiques clés
summary(train$HouseAge)


#Résultat
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    8.00   35.00   36.55   54.00  136.00
```

```
# Visualisation
ggplot(train, aes(x = HouseAge)) +
  geom_histogram(bins = 40, fill = "steelblue", color = "white") +
  labs(title = "Distribution de l'âge des maisons au moment de la vente",
       x = "Âge (années)", y = "Nombre de maisons")
```




```
# Visualisation
ggplot(train, aes(x = HouseAge)) +
  geom_histogram(bins = 40, fill = "steelblue", color = "white") +
  facet_wrap(~BldgType) +
  labs(title = "Distribution de l'âge des maisons par type de bâtiment",
       x = "Âge (années)", y = "Nombre de maisons")
```



La majorité des maisons ont entre **20 et 60 ans**, avec une concentration de maisons plus récentes dans certains quartiers.

6. Corrélations avec SalePrice

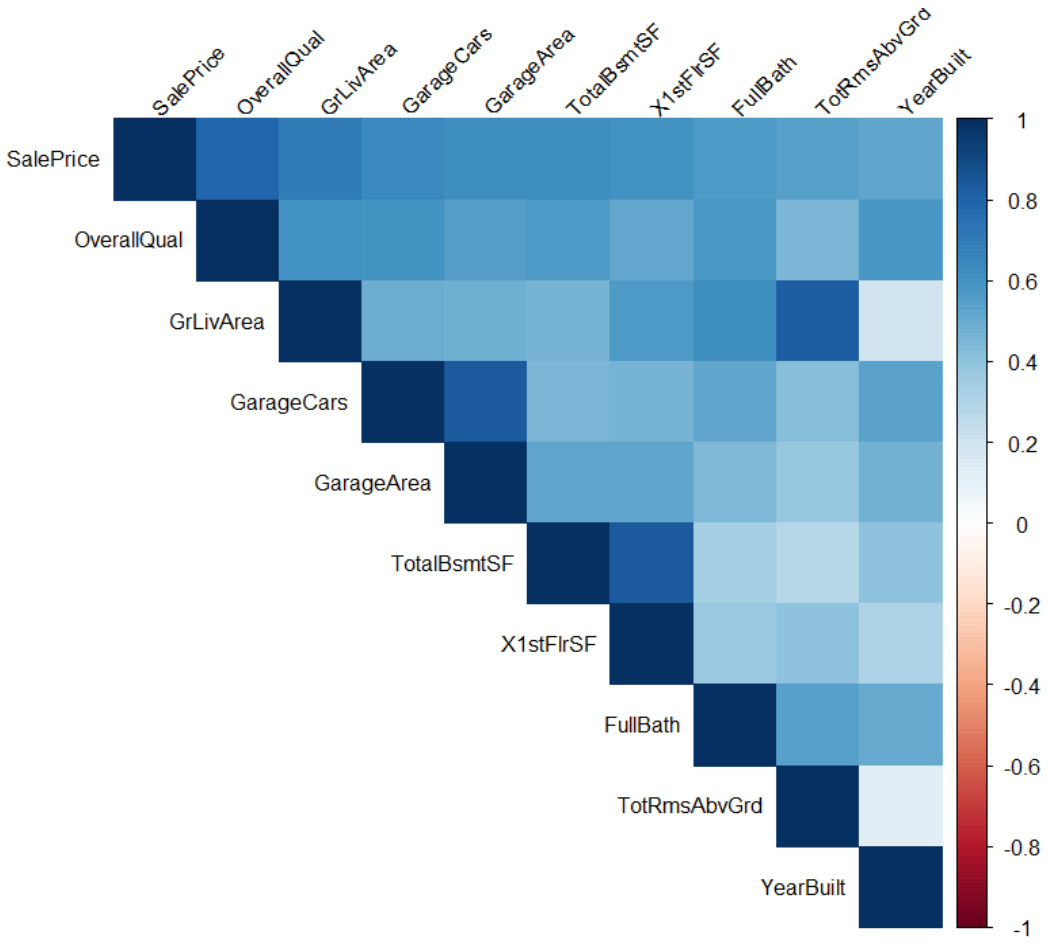
6.1. Corrélations numériques

```
# Sélectionner les colonnes numériques
numeric_data <- train %>% select_if(is.numeric)


# Calculer la matrice de corrélation
cor_matrix <- cor(numeric_data, use = "complete.obs")


# Extraire les corrélations avec SalePrice
saleprice_cor <- cor_matrix[, "SalePrice"] %>%
  sort(decreasing = TRUE)


# Afficher les 10 plus fortes corrélations
head(saleprice_cor, 10)
```

```
#Résultat
SalePrice  OverallQual    GrLivArea   GarageCars   GarageArea
   1.0000000    0.7978807    0.7051536    0.6470336    0.6193296
 TotalBsmtSF    X1stFlrSF     FullBath TotRmsAbvGrd    YearBuilt
   0.6156122    0.6079691    0.5666274    0.5470674    0.5253936
```

6.2. Visualisation des corrélations

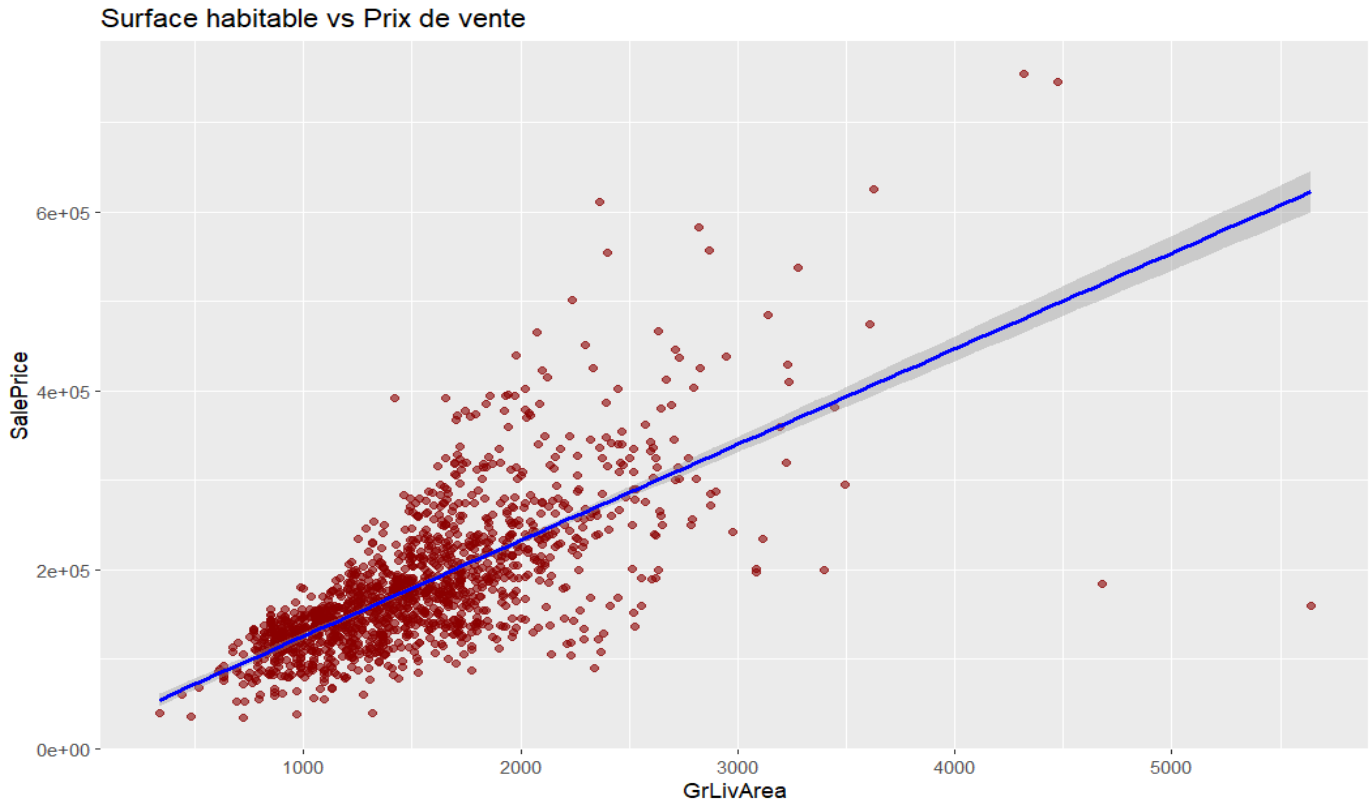
```
# Top 10 variables corrélées avec SalePrice
top_vars <- names(head(saleprice_cor, 10))
sub_matrix <- cor_matrix[top_vars, top_vars]


#Visualisation
corrplot(sub_matrix, method = "color", type = "upper",
         tl.cex = 0.8, tl.col = "black", tl.srt = 45)
```



La qualité générale, la surface habitable et la capacité du garage sont les principaux facteurs influençant le prix.

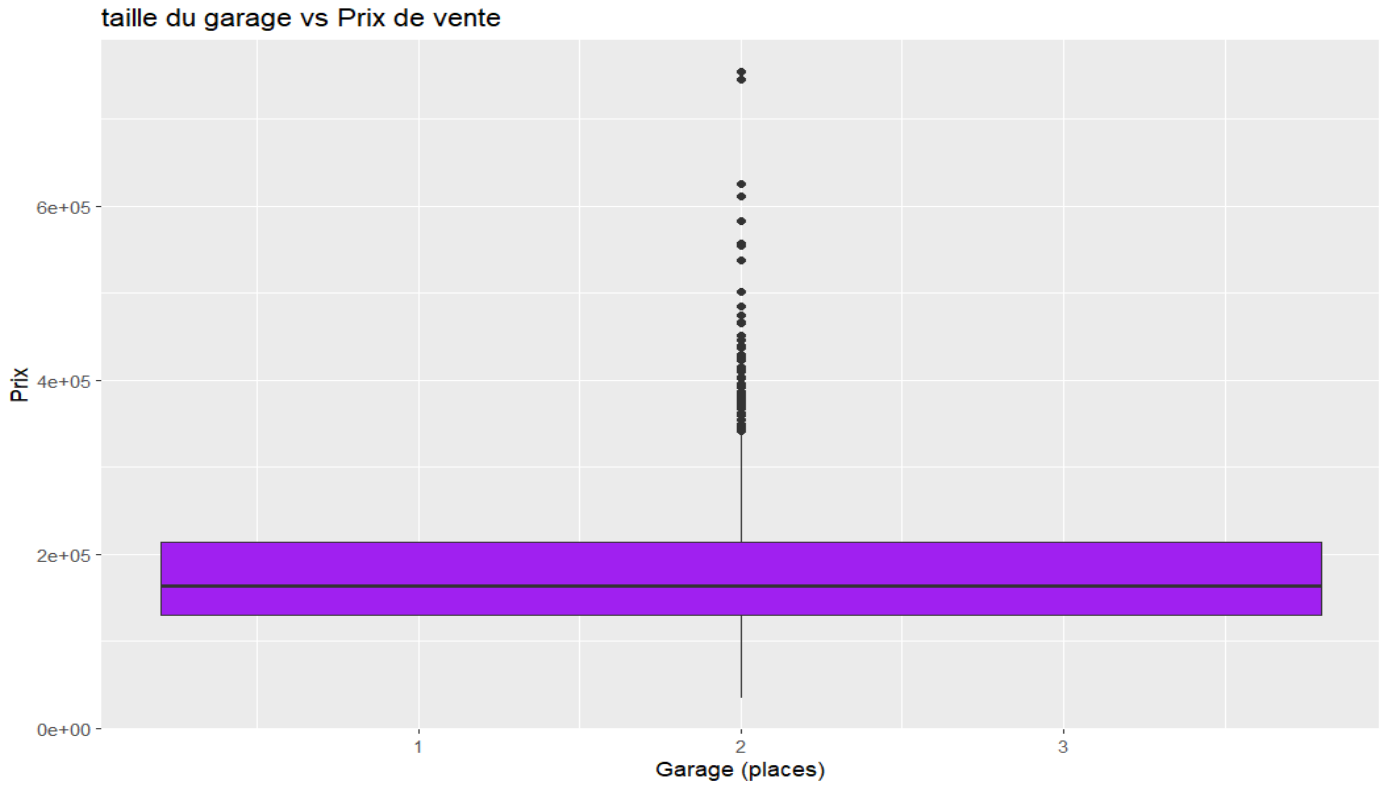
6.3. Visualisation des variables les plus corrélées

```
# prix vs surface habitable
ggplot(train, aes(x = GrLivArea, y = SalePrice)) +
  geom_point(alpha = 0.6, color = "darkred") +
  geom_smooth(method = "lm", color = "blue") +
  labs(title = "Surface habitable vs Prix de vente")
```



GrLivArea vs SalePrice ⟶ tendance linéaire nette

```
# Visualisation Prix vs taille du garage
ggplot(train, aes(x = GarageCars, y = SalePrice)) +
  geom_boxplot(fill = "purple") +
  labs(title = "taille du garage vs Prix de vente", x = "Size of garage in car
capacity", y = "Prix")
```



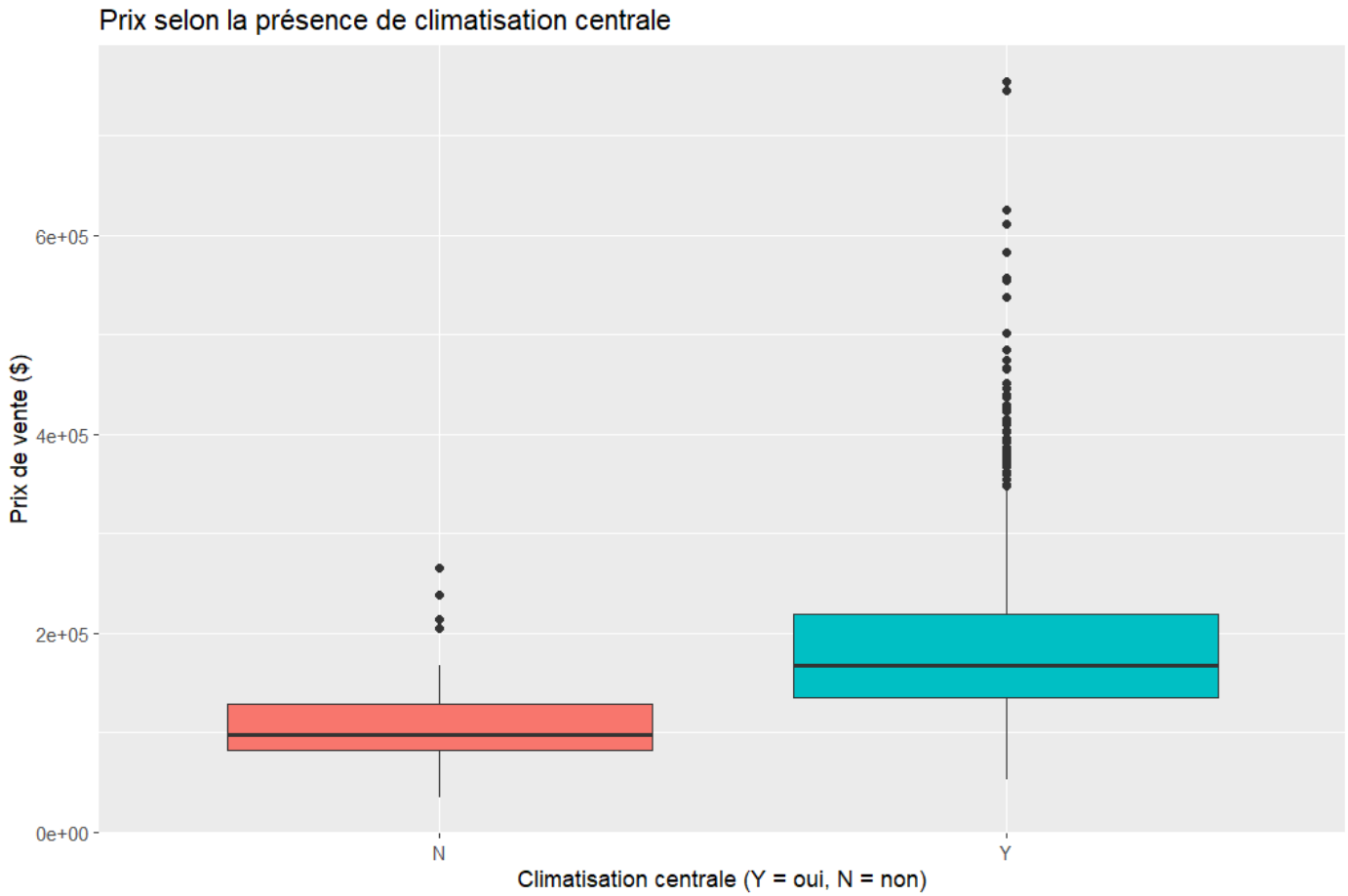
GarageCars vs SalePrice ⟶ effet positif jusqu'à 3 places

7. Categorical Variable Analysis

```
# Convertir les facteurs
train <- train %>%
  mutate_if(is.character, as.factor)


# Vérifier
str(train)


#Nombre des variables categorielles
train %>%
  select(where(is.factor)) %>%
  ncol()


#Resultat
[1] 43


# par Neighborhood
train %>%
  group_by(Neighborhood) %>%
  summarise(
    avg_price = mean(SalePrice, na.rm = TRUE),
    count = n()
  ) %>%
  arrange(desc(avg_price))


#Résultat
Neighborhood avg_price count
   <fct>            <dbl> <int>
 1 NoRidge        335295.    41
 2 NridgHt        316271.    77
 3 StoneBr        310499     25
 4 Timber         242247.    38
 5 Veenker        238773.    11
 6 Somerst        225380.    86
 7 ClearCr        212565.    28
 8 Crawfor        210625.    51
 9 CollgCr        197966.   150
10 Blmngtn        194871.    17
```

On observe de fortes différences de prix selon le quartier.
Les zones telles que **NoRidge**, **NridgHt** et **StoneBr** regroupent les maisons les plus onéreuses**.**

```
# Visualisation par présence de Climatisation
ggplot(train, aes(x = CentralAir, y = SalePrice, fill = CentralAir)) +
  geom_boxplot() +
  labs(title = "Prix selon la présence de climatisation centrale",
       x = "Climatisation centrale (Y = oui, N = non)",
       y = "Prix de vente ($)") +
  theme(legend.position = "none")
```



Les maisons équipées de climatisation centrale (Y) présentent des prix de vente plus élevés.
Cela reflète probablement un meilleur confort et un standing plus moderne.

8. Conclusion

En conclusion, les variables les plus déterminantes du prix sont la qualité générale (OverallQual), la surface habitable (GrLivArea), et le quartier (Neighborhood).
Ces résultats offrent une base solide pour le développement d'un modèle de régression prédictif.