---
title: "PROJET DE MODELISATION STATISTIQUE DU PRIX DES VOITURES DE FABRICATION AMERICAINE SUR LE MARCHE DES ETATS-UNIS"
author: "Mohamed Falilou Fall"
date: "2024-07-17"
output: html_notebook

Les 11 variables: "Km" (kilométrage du véhicule), "Marque" (marque du véhicule), "Modèle" (modèle du véhicule), "Sous-modèle" (varient selon les options), "Type" (de carrosserie), "Cylindré" (moteur), "Litre" (mesure plus précise de la taille du moteur), "Portes" (Nombre de portes), "Régulateur", "HP" (haut-parleurs Dolby stéréo), "Cuir" (intérieur cuir).

Les Marques, "Buick" , "Cadillac" , "Chevrolet" , "Pontiac" , "Saturn", sont de fabrication americaine.
La marque "SAAB" est de fabrication suedoise.

1- Le DataFrame

1-1 Affichage des 7 premieres lignes du DataFrame
```{r}
head(Base_Projet_Introduction_Modelisation_2024,7)
```
1-2 Informations tirees du DataFrame

```{r}
# La dimension du DataFrame (804 voitures et 12 variables)
dim(Base_Projet_Introduction_Modelisation_2024)
```
```{r}
# Les Marques de voitures et leurs frequences
frequence_marque <- table(Base_Projet_Introduction_Modelisation_2024$Marque)
print(frequence_marque)
```

```{r}
# Compte des Modèles de voitures et leurs frequences
frequence_modele <- table(Base_Projet_Introduction_Modelisation_2024$Modèle)
print(frequence_modele)
```

```{r}

# `Valeure minimale d'une voiture = 5.182.840  F CFA`
# `Valeure maximale d'une voiture = 42.449.034 F CFA`

prix_minimal <- min(Base_Projet_Introduction_Modelisation_2024$Prix)
prix_maximal <- max(Base_Projet_Introduction_Modelisation_2024$Prix)



print(paste("Le prix minimal en $ d'une voiture de marque americaine est egal a:", prix_minimal, ", "  ,"soit:", 599.94 * prix_minimal, "F CFA"))



print(paste("Le prix maximal en $ d'une voiture de marque americaine est egal a:", prix_maximal, ", "  ,"soit:", 599.94 * prix_maximal, "F CFA"))

```


2- Identification des 3 variables les plus corrélées à la variable dépendante Prix.

2-1 - Seperation des variables numeriques `num_vars` des variables categorielles `cat_vars`

```{r}

num_vars <- Base_Projet_Introduction_Modelisation_2024[,
sapply(Base_Projet_Introduction_Modelisation_2024, is.numeric)]

cat_vars <- Base_Projet_Introduction_Modelisation_2024[,
sapply(Base_Projet_Introduction_Modelisation_2024, is.factor) | sapply(df,
is.character)]

```


2-2 Afichage de la base de donnees des variables numeriques `num_vars`
```{r}
head(num_vars,7)
```


2-3 Correlation entre les variables numeriques `num_vars`
```{r}
correlations <- cor(num_vars)
correlations

```
```{r}
# Heatmap Correlation entre les variables numeriques `num_vars`

library(corrplot)

corrplot(correlations, method = "color", type = "upper",
         tl.col = "black", tl.srt = 45,
         col = colorRampPalette(c("blue", "white", "red"))(200))

```


2-4 Extraction de la variable dependante `Prix`
```{r}
dependante_var <- "Prix"
dependante_var
```


2-5 Extraction des corrélations avec la variable dépendante `Prix`
```{r}
cor_avec_prix <- correlations[, dependante_var]
cor_avec_prix
```


2-7 Tri par ordre des corrélations avec la variable dépendante `Prix`
```{r}
tri_correlations <- sort(abs(cor_avec_prix), decreasing = TRUE)
tri_correlations
```


Interpretation :

**Les 3 variables les plus corrélées à la variable dépendante `Prix` sont `Cylindrée`,

`Litre` et `Régulateur`**

3 - Mise en place de 3 modèles de régression linéaire simple et Comparaisons des modèles

3-1 Ajustement du modèle de régression linéaire `Prix`, `Cylindrée`: le `model1`

```{r}
model1 <- lm(Prix ~ Cylindrée, data = num_vars)
summary(model1)
```

```{r}
library(ggplot2)
library(plotly)

p <- ggplot(num_vars, aes(x = Prix, y = Cylindrée)) +
  geom_point(color = 'blue') +
  geom_smooth(method = "lm", se = FALSE, color = 'red') +
  labs(title = "Régression linéaire entre Prix et Cylindrée", x = "Prix", y = "Cylindrée")

p_interactif1 <- ggplotly(p)

p_interactif1
```

3-2 Ajustement du modèle de régression linéaire `Prix`, `Litre`: le `model2`

```{r}
model2 <- lm(Prix ~ Litre, data = num_vars)
summary(model2)
```
```{r}

p <- ggplot(num_vars, aes(x = Prix, y = Litre)) +
  geom_point(color = 'blue') +
  geom_smooth(method = "lm", se = FALSE, color = 'red') +
  labs(title = "Régression linéaire entre Prix et Litre", x = "Prix", y = "Litre")

p_interactif2 <- ggplotly(p)

p_interactif2
```

3-3 Ajustement du modèle de régression linéaire `Prix`, `Régulateur`: le `model3`

```{r}
model3 <- lm(Prix ~ Régulateur, data = num_vars)
summary(model3)
```

```
```{r}
library(ggplot2)
library(plotly)


p <- ggplot(num_vars, aes(x = Prix, y = Régulateur)) +
  geom_point(color = 'blue') +
  geom_smooth(method = "lm", se = FALSE, color = 'red') +
  labs(title = "Régression linéaire entre Prix et Régulateur", x = "Prix", y =
"Régulateur")


p_interactif3 <- ggplotly(p)


p_interactif3
```


3-4 Comparaison des 3 models(`model1`, `model2`, `model3`)

3-4-1 Comparaison des `R-squared ($R^2$)` , `Adjusted R-squared` , `AIC` , `BIC` , `Coefficients` , `p-value`
```{r}
# Extraction des informations importantes pour la comparaison

models <- list(model1, model2, model3)
noms_models <- c("Model1 (Prix ~ Cylindrée)", "Model2 (Prix ~ Litre)", "Model3 (Prix ~
Régulateur)")

resultats_compares <- data.frame(
  Model = noms_models,
  R_squared = sapply(models, function(model) summary(model)$r.squared),
  Adjusted_R_squared = sapply(models, function(model) summary(model)$adj.r.squared),
  AIC = sapply(models, AIC),
  BIC = sapply(models, BIC),
  p_value = sapply(models, function(model) anova(model)$`Pr(>F)`[1])
)

print(resultats_compares)

```


Interpretation:

 - Le `R_Squared` : La proportion de la variance totale des données qui est expliquée par le `modèle1` est plus eleve que ceux des 2 autres models. Plus le `R-squared` est élevé, meilleur est l'ajustement du modèle aux données. Donc le `model1` est le meilleur model si nous nous basons uniquement sur le `R_Squared`.

 - `Adjusted_R_squared` : La valeure de l'`Adjusted R-squared` du `model1` indique un meilleur ajustement du modèle aux données car etant plus eleve. Donc le `model1` est le meilleur model si nous nous basons uniquement sur l'`Adjusted_R_squared`.

- Le `model1` a la plus petite valeur d'`AIC` (Akaike Information Criterion) donc il est le meilleur model.

- Le `BIC` (Bayesian Information Criteria) du `model1` est plus bas que celui des 2 autres donc nous choisisons le `model1`.

- Le `model1` a un variable prédictive `Cylindrée` ayant un p-value significativement
```

faible et est considéré comme le meilleur car il indique une meilleure adéquation du variable `Cylindrée` utilisée pour expliquer la variable `Prix `

3-4-2 Visualisation des residus

```{r}
# Le Trace les résidus vs les valeurs ajustées pour chaque modèle

par(mfrow = c(1, 3))  # Disposition des graphiques en une ligne et trois colonnes

plot(model1$fitted.values, model1$residuals, main = "Model1 (Prix ~ Cylindrée)",
     xlab = "Valeurs ajustées", ylab = "Résidus")
abline(h = 0, col = "red")

plot(model2$fitted.values, model2$residuals, main = "Model2 (Prix ~ Litre)",
     xlab = "Valeurs ajustées", ylab = "Résidus")
abline(h = 0, col = "red")

plot(model3$fitted.values, model3$residuals, main = "Model3 (Prix ~ Régulateur)",
     xlab = "Valeurs ajustées", ylab = "Résidus")
abline(h = 0, col = "red")

```

Interpretation :

- le `Model2` a une répartition uniforme et aléatoire des résidus autour de la ligne zéro (ligne rouge). Une distribution uniforme du `model2` suggère que les erreurs de prédiction sont réparties de manière équitable sur l'ensemble des donnees `num_vars`.

4 - Analyse multivariée pour décrire le processus d'attribution du Prix d'un véhicule en fonction des caractéristiques quantitatives.

4-1 Régression linéaire multiple (Interpretation du model)

4-1-1 Analyse du `summary` du model de regression multiple
```{r}
model_reg_multiple <- lm(Prix ~ Km + Cylindrée + Litre + Portes + Régulateur + HP + Cuir, data = num_vars)

summary(model_reg_multiple)
```

Analyse:

1 - `Estimate` (Estmations) :

- Pour chaque kilometre `Km` de plus dans le compteur d'une voiture, le Prix moyen diminue en moyenne de `1.698e-01 $ U.S` soit `101.88 FCFA` , si le taux de conversion est 1$ = 6oo F CFA.

- Pour chaque `Cylindre` de plus, le Prix moyen d'une voiture augmente en moyenne de `3.792e+03 $ U.S` soit `2.275.200 FCFA` , si le taux de conversion est 1$ = 6oo F CFA.

- S'il y'a presence de `Régulateur` (cruise control), le Prix moyen d'une voiture augmente en moyenne de `6.289e+03 $ U.S` soit `3.773.400 FCFA` , si le taux de conversion est 1$ = 6oo F CFA.

- Le fait que les sieges de la voiture sont en `Cuir` fait augmenter le Prix moyen de la voiture en moyenne de `3.349e+03 $ U.S` soit `2.009.400 FCFA`, si le taux de conversion est 1$ = 6oo F CFA.

- La dimunition d'une unite du nombre de `Portes` de la Voiture fait dimunier le Prix moyen en moyenne de `1.543e+03 $ U.S` soit `925.800 FCFA`, si le taux de conversion est 1$ = 6oo F CFA.

- Une non presence de haut-parleurs Dolby stéréo (`HP`) dans la voiture fait diminuer son Prix en moyenne de `1.994e+03 $ U.S` soit `1.196.400 FCFA`, si le taux de conversion est 1$ = 6oo F CFA.

2 - `Les codes de significativité` : d'apres les codes, les coefficients sont tous significatifs avec 3 etoiles (` ***`) sauf pour la variale `Litre`

4-1-2 Graphique des residus

```{r}
plot(model_reg_multiple)
```

4-1-2 Test de normalité des résidus

```{r}
shapiro.test(resid(model_reg_multiple))
```

Interpretation :

- Le `W = 0.92927` est tres proche de 1 donc les données semblent suivre une distribution normale. La proximite a 1 indique une adequation a la normalite.

- la `p-value = 0.00000000000000022` est inferieure au seuil `alpha=0,05` donc l'hypothese nulle `H0` est rejetee donc les donnees sont normalement distribuees.

5 - `38` prévisions avec des intervalles de `prévision` et `confiance` de `95%` et `99%` avec le `model1`

`fit=valeure ajustee` `lwr=valeure minimale` `upr=valeure maximale`

5-1 `38` Prédictions avec intervalle de `prédiction` 95%

```{r}
predictions1 <- predict(model1, interval = "prediction", level = 0.95)
head(predictions1,38)
```

5-2 `38` Prédictions avec intervalle de `prédiction` 99%
```{r}
predictions2 <- predict(model1, interval = "prediction", level = 0.99)
head(predictions2, 38)
```

5-3 `38` Prédictions avec intervalle de `confiance` de 95%
```{r}
predictions3 <- predict(model1, interval = "confidence", level = 0.95)
head(predictions3,38)
```

5-4 `38` Prédictions avec intervalle de `confiance` de 99%
```{r}
predictions4 <- predict(model1, interval = "confidence", level = 0.99)
head(predictions4,38)
```

5-5 Verification d'un quelconque ressemblance entre les differentes series de predictions

```{r}
# Verification d'un quelconque ressemblance
identical(predictions1, predictions3)
```

```{r}
# Verification d'un quelconque ressemblance
identical(predictions2, predictions4)
```
```{r}
# Verification d'un quelconque ressemblance
identical(predictions1, predictions2)

```

```{r}

identical(predictions3, predictions4)

```