Table des matières

Introduction

Au début du XXème siècle, Henry Ford déclarait : « Le client peut choisir la couleur de sa voiture, pourvu que ce soit noir ». C'est sans doute sa citation la plus connue. En effet, jusqu'en 1914, la Model T était disponible en plusieurs teintes extérieures. Mais lorsque H. Ford s'est rendu compte que le séchage de la peinture était l'opération la plus longue du processus de production, il décida d'adopter uniquement pour la couleur qui sèche le plus vite : le noir.

De nos jours, l'industrie automobile est un secteur très dynamique soumis à plusieurs facteurs qui contribuent à la formation des prix des véhicules. Il est essentiel pour les consommateurs, pour les fabricants ou encore pour les décideurs économiques, de comprendre les mécanismes souterrains qui déterminent le prix des voitures. Dans cette étude économétrique, nous nous pencherons sur les divers éléments susceptibles d'influencer ces prix.

L'analyse nous permettra d'évaluer empiriquement l'impact de la couleur noire sur le prix de vente des voitures à l'aide de notre variable explicative et de nos variables de contrôles. L'objectif est donc d'identifier les déterminants les plus significatifs et de quantifier leurs effets respectifs. A travers notre étude, nous chercherons à répondre à la problématique suivante : "Qu'est-ce qui influence le plus le prix d'une voiture ? Est-ce plus économique d'acheter une voiture noire ?".

Pour notre étude, nous avons choisi d'utiliser une base de données générée selon nos instructions par une IA qui contient 12344 observations de véhicules et 14 variables portant sur les différentes caractéristiques de ces véhicules.

Explications des variables

- Prix : Prix de vente du véhicule en dollars
- Ford : Variable binaire qui vaut 1 si l'entreprise qui a fabriqué le véhicule est Ford, 0 sinon.
- Année : Année de production
- Catégorie : Type de véhicule
- Intérieur_cuir : Variable binaire valant 1 si l'intérieur du véhicule est en cuir et 0 sinon
- Ess : Variable binaire qui vaut 1 si le carburant utilisé par le véhicule est de l'essence , 0 sinon.
- Cylindrée : Volume du moteur
- D_Km : Nombre de kilomètres du véhicule divisé par 10 000
- Roues_motrices : Quelles sont les roues motrices du véhicule
- Turbo : Variable binaire indiquant si le moteur du véhicule dispose d'un turbo ou pas
- Conduite_droite : Variable binaire valant 1 si le volant est à droite et 0 sinon
- Boîte_auto : Variable binaire valant 1 si la voiture est équipée d'une boîte automatique et 0 sinon
- Black : Variable binaire valant 1 si la voiture est de couleur noire et 0 sinon
- D_Km2 : Carré de la variable "D_Km"

Nettoyage et statistiques descriptives

La première étape pré-nettoyage fût de nous approprier la base, en binarisant certaines variables catégorielles, renommant les variables, en bref nous l'avons modelé pour qu'elle réponde précisément à notre problématique.

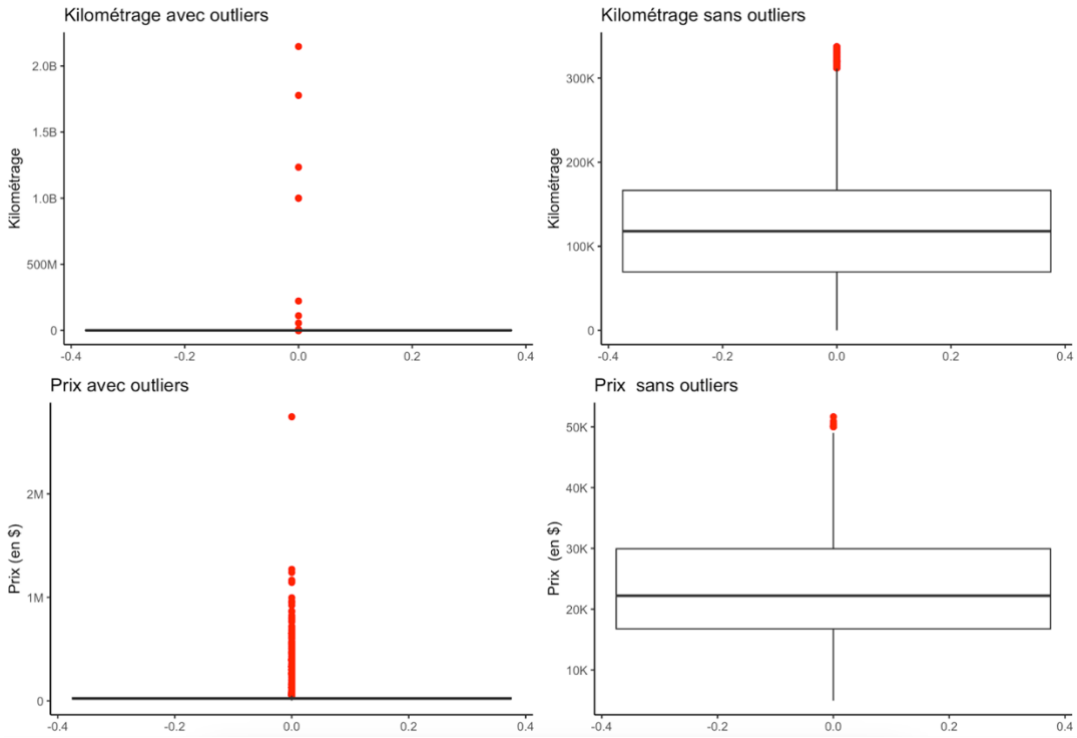
Nous nous sommes premièrement concentrés seulement sur les voitures non commerciales que nous possédions (en excluant par exemple les limousines et les camionnettes) pour nous focaliser sur les voitures accessibles au grand public. Ce qui nous fît tomber à 10687 observations.

Nous avons continué en filtrant les observations aberrantes au sein de notre base sur les deux principales variables que sont le Kilométrage et le Prix avec la formule suivante :

$$Outlier = \{x | x < Q1 - 1.5 * IQR \; ou \; x > Q3 + 1.5 * IQR\}$$
$$avec \; Q1 = premier \; quartile$$
$$Q3 = 3e \; quartile$$
$$IQR = Q3 - Q1$$

Grâce à cette détection des valeurs aberrantes, notre base passe à 7699 observations qui sont cette fois-ci non-aberrantes et pertinentes.

Ci-dessous la distribution des deux variables évoquées précédemment avant et après filtration des outliers.



Une fois notre base filtrée et adaptée à notre problématique nous avons pu nous atteler à la description des variables de cette dernière afin de mieux comprendre la base
Regardons dans un premier temps le tableau ci-dessous qui regroupe les statistiques générales de notre base de données

Statistiques générales						
	mean	sd	median	min	max	range
Prix	23689,35	8720,25	22222	4948	51680	46732
Année	2011,73	4,26	2012	1999	2020	21
Intérieur_cuir	0,67	Variables binaires				
Ford	0,06					
Turbo	0,08					
black	0,24					
Conduite_droite	0,1					
Ess	0,8					
Boîte_auto	0,76					
Cylindrée	2,02	0,57	2	0,4	3,8	3,4
D_km	12,3	7,26	11,8	0	33,76	33,76

Premièrement, nous constatons que les voitures présentes au sein de notre base de données sont vendues entre 4 948$ et 51 680$ avec une moyenne de 23 689$. Nous travaillons sur des modèles produits entre 1999 et 2020 avec autant de voitures fabriquées avant et après 2012. Ensuite, nous remarquons qu'il y a 24% de véhicules de couleur noire dans notre base. Cette donnée nous intéresse particulièrement puisqu'il s'agit de notre variable explicative. Nous chercherons à savoir si ces 24% de voitures noires sont significativement moins chères que les autres couleurs. De plus, il y a 6% de voitures fabriquées par Ford parmi les fabricants dans notre base. Enfin, on note que la base contient des voitures neuves (0km) et à l'inverse des voitures très usées (337 600km). La moyenne est d'environ 123 000km, ce qui est assez élevé.

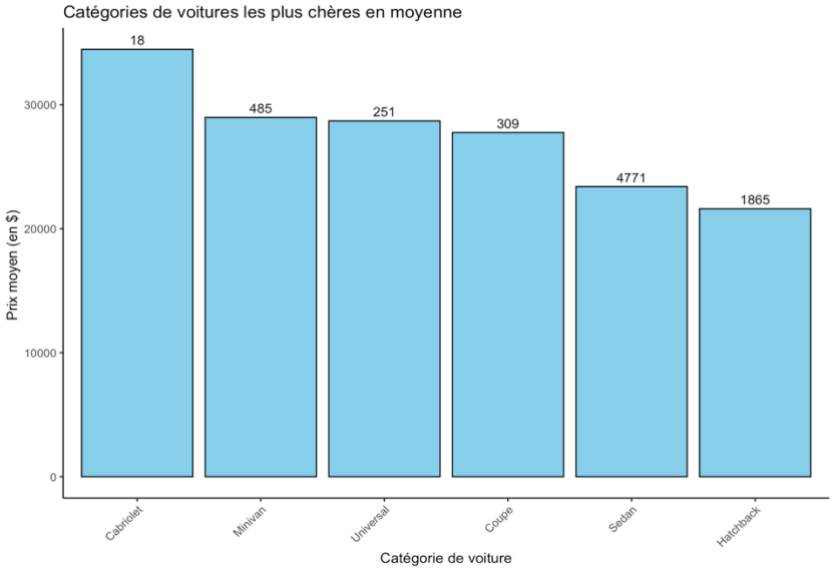
En ce qui concerne les voitures non noires, nous remarquons que leur prix moyen est de 23 531,52$ avec des valeurs allant de 4 948 à 50 565$. De plus, nous pouvons voir que 64% de ce type de voitures possède un intérieur cuir ou que 78% de celles-ci sont équipées d'une boîte automatique. Enfin, leur kilométrage moyen est d'environ 124 000km avec des valeurs allant de 0 à 337 600km.

Statistiques des voitures non noires (black == 0)						
	mean	sd	median	min	max	range
Prix	23531,52	8641,19	21879	4948	50565	45617
Année	2011,67	4,25	2012	1999	2020	21
Intérieur_cuir	0,64	Variables binaires				
Ford	0,06					
Turbo	0,07					
Conduite_droite	0,11					
Ess	0,79					
Boîte_auto	0,78					
Cylindrée	1,97	0,54	1,8	0,6	3,8	3,2
D_km	12,39	7,28	11,84	0,00	33,76	33,76

Statistiques des voitures noires (black == 1)						
	mean	sd	median	min	max	range
Prix	24184,12	8947,91	22279	7407	51680	44273
Année	2011,92	4,29	2013	1999	2020	21
Intérieur_cuir	0,75	Variables binaires				
Ford	0,06					
Turbo	0,12					
Conduite_droite	0,07					
Ess	0,82					
Boîte_auto	0,70					
Cylindrée	2,18	0,64	2	0,4	3,8	3,4
D_km	12,03	7,19	11,7	0,0	33,6	33,6

Pour les voitures noires, la moyenne de prix est supérieure à celle des voitures non noires. En effet, elle est de 24 184,12$ avec des valeurs allant de 7 407 à 51 680$. Il serait donc intéressant de voir si cette tendance se confirme lors des régressions et de l'ajout des variables de contrôle. Nous pouvons aussi constater qu'il y a plus d'intérieur cuir que pour les voitures non noires mais moins de boîtes automatiques. En effet, 75% des voitures noires possèdent un intérieur en cuir et seulement 70% de celles-ci sont équipées d'une boîte automatique. Enfin, leur kilométrage moyen est inférieur à celui des voitures non noires (environ 120 000km avec des valeurs allant de 0 à 336 000km).

Après avoir étudié les tableaux des statistiques descriptives générales et comparé les sous populations des voitures noires et non noires, nous allons commenter le graphique ci-dessous montrant le prix moyen d'une voiture en fonction de sa catégorie.

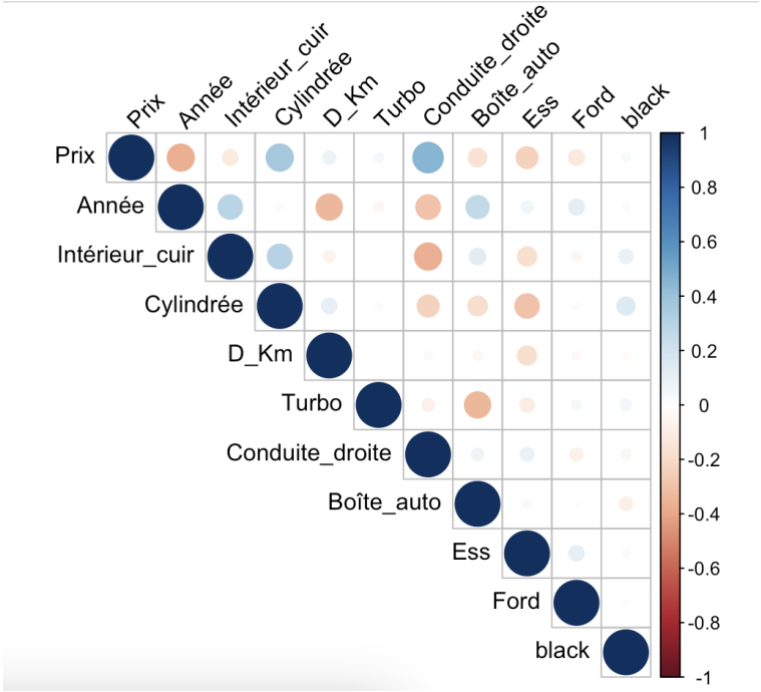


Ce graphique nous indique donc la répartition du prix moyen par catégorie de voiture.
On peut voir également le nombre de voitures par catégorie (485 minivans par exemple). Les cabriolets sont les voitures les moins représentées mais ce sont également les plus chères avec un prix moyen d'environ 35 000$. Les voitures les plus vendues sont les Sedan, avec plus de 60% d'observations. Cependant, elles coûtent en moyenne un peu plus de 20 000$, ce qui fait d'elles la deuxième moins chère derrière la catégorie Hatchback (Citadines).

Ce graphique nous permet de nous rendre compte de la corrélation entre nos différentes variables, ce qui est nécessaire avant d'entamer la partie suivante qui portera sur les régressions.

Il en ressort majoritairement que le Prix, la Cylindrée et la conduite à droite sont corrélés positivement et la variable année quant à elle est corrélée négativement au prix.

Nous ne voyons pas de corrélation excessive ce qui est un bon présage pour nos régressions linéaires.



Régressions linéaires

Nous sommes premièrement partis de la régression naïve suivante :

$$ln(Prix) = \beta_0 + \beta_1\, black + \varepsilon$$

Le β_1 ici égal à 0.024 nous donne ainsi l'écart moyen de prix (en%) entre les voitures noires et les autres, ici les voitures noires coûtent en moyenne 2.4% plus chères que les autres.

Nous avons donc « amélioré » notre régression linéaire petit à petit en incorporant les variables de contrôle et ainsi arriver au modèle final suivant :

$$\begin{aligned} ln(Prix) = {} & \beta_0 + \beta_1\, black + \beta_2 Caté + \beta_3 Ford + \beta_4 D_{Km} + \beta_5 Km2 + \beta_6 An \\ & + \beta_7 An * Km + \beta_8 Ess + \beta_9 ln(Cyl) + \beta_{10} Turbo + \beta_{11} Cuir \\ & + \beta_{12} Boîte + \beta_{13} Roues + \beta_{14} Cond + \beta_{15} D_{Km2} + \beta_{16} Ann * D_{Km}\, c \\ & + \beta_{17} Ford * black + \varepsilon \end{aligned}$$

Au fur et à mesure de l'incorporation de nos variables de contrôles par « blocs » ainsi que de nos transformations, observons la variation de notre $\beta_1$black :

Bloc 1	Année, Catégorie, Ford , D_Km
Bloc 2	Ess, Cylindrée, Turbo
Bloc 3	Intérieur_cuir
Bloc 4	Boîte_auto, Roues_motrices, Conduite_droite
Transformations	D_Km2, Année*D_Km, black*Ford

Nous avons pensé nos « blocs » comme ci-suit :
Le Bloc 1 correspond aux « spécificités du véhicule », le 2 à celles de son moteur, le 3 aux options que le véhicule peut avoir et pour finir, le 4, aux spécifications techniques plus « générales » du véhicule.

De la régression simple dite "naïve" au modèle final						
	Black	**Black + bloc 1**	**Black + bloc 1:2**	**Black + bloc 1:3**	**Black + bloc 1:4**	**+ Transformation**
β de black sur ln(Prix)	0,024	0,021	-0,006	0,002	-0,014	-0,017
Significativité de Black	**	**			**	**
R2 ajusté du modèle	0,1%	18,0%	26,2%	28,7%	48,9%	50,4%
Significativité globale du modèle (statistique F)	**	***	***	***	***	***
*p<0.1; **p<0.05; ***p<0.01						

Pourquoi avoir opté pour certaines de ces transformations ? Pour le fun ? Oui mais pas toutes …

- Premièrement, nous avons passé notre variable Prix en log afin de pouvoir minimiser les éventuels problèmes d'hétéroscédasticité et d'avoir un modèle en Log-niveau et pouvoir ainsi interpréter des pourcentages qui dans notre cas du Prix sont plus parlants.

- Nous avions l'intuition que l'effet marginal du kilométrage sur le prix pourrait ne pas être le même à tous les niveaux de kilométrage (en effet, une voiture qui passe de 90 à 100k Km perd logiquement plus de valeur qu'une voiture qui passe de 20 à 30k Km) c'est pour ce faire que nous avons intégré le carré de notre variable D_Km à la régression, afin de capturer l'effet non linéaire de cette variable dans notre modèle.

- Le modèle de double différence black*Ford, nous l'avons inclus par pur clin d'œil à notre problématique, afin d'avoir encore plus de précision et de savoir si les voitures Ford se vendent moins chères quand elles sont noires.

- L'interaction entre Année*D_Km a été incluse à notre modèle afin de savoir si l'Année de notre véhicule influe ou non sur l'effet qu'a le Kilométrage sur le Prix (une voiture récente décote-t-elle de la même manière qu'une vieille voiture à cause de son kilométrage ?).

- Le modèle Log-Log entre le log (Prix) et le log(Cylindrée) nous a permis de nous rendre compte de la variation du Prix en pourcentage pour chaque augmentation de 1% de la Cylindrée. Nous avons fait ce choix de transformation en log de la cylindrée car bien qu'il s'agisse de valeurs relativement « petites » il est beaucoup plus intuitif de rendre compte d'une augmentation du % de la cylindrée du moteur que d'augmentation dans l'unité du moteur (il y a une grosse différence entre un moteur de 1L de cylindrée et un moteur de 2L de cylindrée)

Attardons-nous désormais plus spécifiquement à notre régression finale afin d'analyser et discuter certains de ses principaux coefficients de notre régression :

	Dependent variable:
	log(Prix)
black	-0.017**
	(0.007)
CatégorieCoupe	-0.158**
	(0.064)
CatégorieHatchback	-0.335***
	(0.063)
CatégorieMinivan	-0.290***
	(0.064)
CatégorieSedan	-0.239***
	(0.063)
CatégorieUniversal	-0.143**
	(0.065)
Ford	-0.106***
	(0.015)
D_Km	1.980***
	(0.165)
Année	0.002
	(0.001)
Ess	-0.158***
	(0.009)
log(Cylindrée)	0.461***
	(0.015)
Turbo	-0.024**
	(0.012)
Intérieur_cuir	-0.055***
	(0.008)
Boîte_auto	-0.034***
	(0.008)
Roues_motricesFront	-0.118***
	(0.017)
Roues_motricesRear	0.020
	(0.018)
Conduite_droite	0.627***
	(0.012)
D_Km2	0.0004***
	(0.00005)
D_Km:Année	-0.001***
	(0.0001)
black:Ford	0.021
	(0.029)
Constant	7.026***
	(2.660)
Observations	7,699
R2	0.505
Adjusted R2	0.504
Residual Std. Error	0.264 (df = 7678)
F Statistic	392.343*** (df = 20; 7678)
Note:	*p<0.1; **p<0.05; ***p<0.01

- Les voitures noires sont 1,7% moins cher en moyenne, toutes choses égales par ailleurs.

- La catégorie de voitures les moins chères toutes choses égales par ailleurs sont les « hatchbacks » (citadines), elles sont moins chères d'environ 33,5% par rapport à la catégorie de référence qui est les Cabriolets.

- Les voitures "Ford" se vendent 10,6% moins cher en moyenne, toutes choses égales par ailleurs, que les voitures qui ne sont pas de la marque Ford.

- En moyenne chaque 10k de Km parcouru, toutes choses égales par ailleurs, augmente la valeur de la voiture de 198 % (Valeur très surprenante voire même louche, possibles explications dans les limites !!).

- Log cylindré : Une augmentation de 1% de la cylindrée, toutes choses égales par ailleurs, est associée à une augmentation d'environ 46,1% du prix (encore une fois valeur quelque peu surprenante que nous développerons dans les limites …).

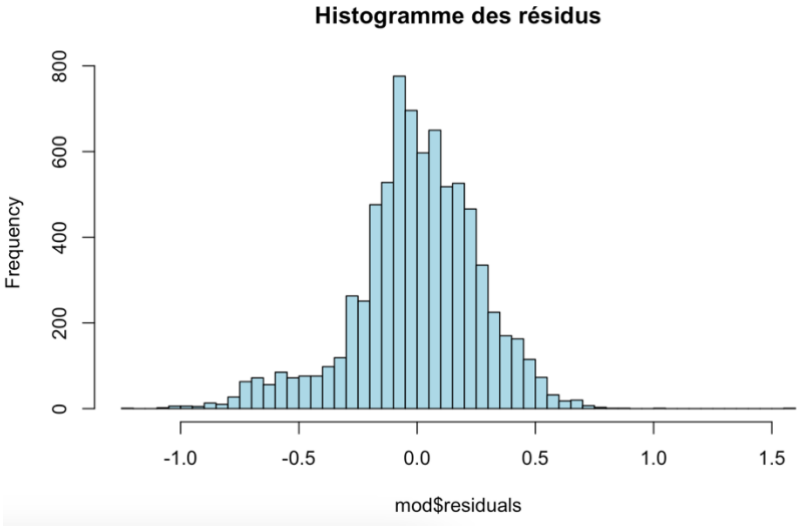
- Le coefficient de D_Km2 nous indique que l'effet non linéaire de D_Km sur le prix est très significatif.

- Le coefficient négatif de l'interaction entre D_Km:Année nous permet d'identifier le phénomène suivant : à mesure que l'année de la voiture augmente, l'effet du kilométrage sur le prix tend à diminuer (une voiture jeune et kilométrée se vendra plus chère qu'une vieille voiture kilométrée).

Que dire de notre R2 ? avec notre R2 ajusté de 50,4% on peut donc voir que notre modèle explique 50,4% de la variance de notre variable dépendante.

Comment payer sa voiture le moins cher possible me diriez-vous ? Grâce à notre modèle et sur notre base, nous pourrions envisager une voiture de catégorie "Hatchback", de marque "Ford", de couleur noire, avec un kilométrage relativement faible (!!!), utilisant de l'essence, sans turbo, avec intérieur en cuir et boîte automatique. Cependant, il est important de noter que d'autres facteurs non-inclus dans le modèle peuvent également influencer les prix, et ces résultats sont basés sur les données spécifiques à notre échantillon.

Intéressons-nous à notre modèle, commençons par analyser ses résidus que voici :



Comme nous pouvons le voir, bien qu'il semblerait que les résidus suivent une loi normale (à l'allure relativement « en forme de cloche » de cette distribution), les résidus ne suivent malheureusement pas une loi normale (le test de Shapiro nous l'a confirmé).

En analysant plus en profondeur notre modèle, nous pouvons voir grâce au test de Breusch-Pagan qu'il y a de l'hétéroscédasticité malgré la transformation en log de notre variable dépendante …

Il pourrait donc être judicieux de nous diriger vers un modèle robuste aux violations de ces prérequis afin d'avoir des estimations plus robustes de nos différents coefficients.

Conclusion

Pour conclure, nous avons réalisé une analyse des statistiques descriptives et une régression linéaire afin de trouver un lien entre la couleur noire et le prix des voitures. Alors que l'analyse descriptive semblait montrer une moyenne de prix plus élevée pour les voitures noires, la régression a permis de prouver que ces dernières sont significativement moins chères que les autres toutes choses égales par ailleurs, lorsqu'on ajoute les variables de contrôle. Ainsi, nous pouvons affirmer que la couleur impacte, bien que très légèrement, le prix des voitures.

Cependant, la base de données contient des limites. En effet, cette base a été créée par une IA et semble avoir généré une base qui ne ressemble pas à la réalité qu'on pourrait connaître.
L'augmentation du kilométrage augmenterait énormément le prix d'une voiture.
Or, la réalité paraît opposée. Plus une voiture a de kilométrage et moins elle se vend chère. Le résultat peut donc être biaisé, soit par la présence de voitures de collection ou bien par la présence en quantité déséquilibrée de voitures neuves et d'occasions qui ne sont pas comparables entre elles. Il aurait été préférable d'avoir une base uniquement avec des voitures neuves ou une base avec des voitures d'occasions.
Il pourrait également y avoir un biais de variables omises qui semble très plausible.
On note aussi un coefficient étrangement élevé au niveau de la cylindrée mais celui-ci s'explique par le fait que c'est une variable "petite" qui est mise en logarithme. De ce fait, son interprétation bien que plus intuitive dans notre cas, peut être quelque peu faussée.

Annexe (Code R)

```
rm(list = ls())
#installer et charger les packages
packages_a_installer <- c("tidyverse","readr","ggplot2","ggpubr","dplyr",
"gmodels","stargazer","DataExplorer","cowplot","htmltools","psych","crosstable",
"flextable","knitr","kableExtra","pROC","htmlTable","shiny","outliers","car","lmtest", "VIM","MASS","margins")
for (package in packages_a_installer)
  {if (!requireNamespace(package, quietly = TRUE)) {
    install.packages(package)}}
for (package in packages_a_installer) {library(package, character.only = TRUE)}
rm(package, packages_a_installer)

#charger la base de données
url <- "https://dl.dropboxusercontent.com/scl/fi/5he67pzzxa9ce4h8o90m7/car_base.csv?rlkey=udczfyeyq8z1tfwhcsd7r0ucb&"
df <- read_csv(url, col_names = TRUE)

#on regarde le nombre de modalités dans différentes catégories
sapply(df, function(x) nlevels(factor(x)))

#944 modalités dans model, on ne la gardera pas, on supprime les autres variables inutiles
#on voit également en regardant l'ID qu'il y a des doublons (moins d'ID que d'observations)
df <- df %>%
  distinct() %>%
  dplyr::select(-ID,-Model )
colnames(df)

#renommer les variables
nom_colonnes <- c("Prix", "Fabricant","Année","Catégorie","Intérieur_cuir","Carburant",
"Cylindrée","Kilométrage","Type_boite","Roues_motrices","Côté_conduite","Couleur")
colnames(df)[1:12] <- nom_colonnes
rm(nom_colonnes, url)
str(df)

#on nettoie la colonne kilométrage en enlevant les caractères non numériques et on fait une colonne de dizaine de Km
df$Kilométrage <- as.numeric(gsub("[^0-9.]", "", df$Kilométrage))
df$D_Km <- df$Kilométrage/10000

#on renomme les modalités des carburants pour mieux les comprendre
df <- df %>%
  mutate(Carburant = ifelse(Carburant == "Diesel", "Diesel","Essence"))

#création variable binaire Turbo, boite auto, intérieur cuir et conduite à droite, …  pour gagner en précision
df$Turbo <- ifelse(grepl("Turbo",df$Cylindrée),1,0)
df$Cylindrée <- as.numeric(gsub("[^0-9.]", "", df$Cylindrée))
df$Conduite_droite <-  ifelse(df$Côté_conduite== "Right-hand drive", 1, 0)
df$Côté_conduite <- NULL
df$Boîte_auto <- ifelse(df$Type_boite =="automatic" , 1, 0)
df$Type_boite <- NULL
df$Intérieur_cuir <- ifelse(df$Intérieur_cuir =="Yes", 1, 0)
df$Ess <- ifelse(df$Carburant =="Essence",1, 0)
df$Ford <- ifelse(df$Fabricant=="FORD", 1,0)
df <- df %>%
  mutate(black = ifelse(Couleur == "Black", 1, 0))
df$Couleur <- NULL

# Filtrer le dataframe pour ne conserver que les catégories de voitures pour les particuliers
categories_a_garder <- c("Cabriolet", "Coupe", "Hatchback", "Sedan", "Minivan", "Universal")
df <- df %>%
  filter(Catégorie %in% categories_a_garder)

#détection et suppression des outliers
variables <- c("Prix", "Kilométrage","Cylindrée")
for (variable in variables) {
  Q1 <- quantile(df[[variable]], 0.25)
  Q3 <- quantile(df[[variable]], 0.75)
  IQR <- Q3 - Q1
  limite_superieure <- Q3 + 1.5 * IQR
  limite_inferieure <- Q1 - 1.5 * IQR}

#on fait un df2 qui est sans outlier
df2 <- df %>%
  filter(across(all_of(variables), ~ between(., quantile(., 0.25) - 1.5 * IQR(.), quantile(., 0.75) + 1.5 * IQR(.))))

#on identifie les valeurs aberrantes pour le prix et le kilométrage (avant suppression et après)
#graphiquement
 a <-ggplot(df, aes(y = Kilométrage)) +
   geom_boxplot(outlier.colour = "red", outlier.size = 2) +
   labs(y = "Kilométrage", title = "Kilométrage avec outliers") +
   theme_classic2() +
   scale_y_continuous(labels = scales::label_number_si())
b <-  ggplot(df, aes(y = Prix)) +
  geom_boxplot(outlier.colour = "red", outlier.size = 2) +
  labs(y = "Prix (en $)", title = "Prix avec outliers") +
```

```
  theme_classic2() +
  scale_y_continuous(labels = scales::label_number_si())
c <- ggplot(df2, aes(y = Kilométrage)) +
  geom_boxplot(outlier.colour = "red", outlier.size = 2) +
  labs(y = "Kilométrage", title = "Kilométrage sans outliers") +
  theme_classic2() +
  scale_y_continuous(labels = scales::label_number_si())
d <- ggplot(df2, aes(y = Prix)) +
   geom_boxplot(outlier.colour = "red", outlier.size = 2) +
   labs(y = "Prix  (en $)", title = "Prix  sans outliers") +
   theme_classic2() +
   scale_y_continuous(labels = scales::label_number_si())
plot_grid(a,c,b,d, ncol = 2, nrow = 2)
rm(list=setdiff(ls(), c("df", "df2")))
df2$Kilométrage <- NULL

#autres firgures pour rendre compte de la distribution : répartition des ventes par fabricant
summary_data <- df2 %>%
  group_by(Fabricant) %>%
  summarise(count = n()) %>%
  arrange(desc(count))
a <- ggplot(summary_data, aes(x = reorder(Fabricant, -count), y = count)) +
  geom_bar(stat = "identity", fill = "skyblue", color = "black") +
  geom_text(aes(label = count), vjust = -0.5, color = "black") +
  labs(title = "Nombre de voitures par fabricant",
       x = "Fabricant",
       y = "Nombre de voitures") +
  theme_classic2() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
# et catégories de voiture les plus chères en moyenne
#histogramme des catégories les plus chères et nombre d'observation dans chaque catégorie
summary_data <- df2 %>%
  group_by(Catégorie) %>%
  summarise(mean_price = mean(Prix), count = n())
summary_data <- summary_data %>%
  arrange(desc(mean_price))
b <- ggplot(summary_data, aes(x = reorder(Catégorie, -mean_price), y = mean_price)) +
  geom_bar(stat = "identity", fill = "skyblue", color = "black") +
  geom_text(aes(label = count), vjust = -0.5, color = "black") +
  labs(title = "Catégories de voitures les plus chères en moyenne",
       x = "Catégorie de voiture",
       y = "Prix moyen (en $) ") +
  theme_classic2() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
{variables_numeriques <- df2[, sapply(df2, is.numeric)]
matrice_correlation <- cor(variables_numeriques)
c <- corrplot::corrplot(matrice_correlation, method = "circle", type = "upper", tl.col = "black", tl.srt = 45)
c}
plot_grid(a,b, ncol = 2, nrow = 1)

#statistiques conditionnelles avec noir =1 ou =0
group_black_1 <- df2 %>% filter(black == 1)
group_black_0 <- df2 %>% filter(black == 0)
summary(group_black_0)
summary(group_black_1)
summary(df2)

#modèle de régression simple dite naîve
mod12 <- lm(log(Prix)~ black , data= df2)
stargazer(mod12, type= "text")

#pour capturer la non linéarité potentielle du kilométrage
df2$D_Km2 <- df2$D_Km^2

# on fera un modèle de double différence pour savoir si les voitures noires de chez ford sont les moins chères (CF introduction)
#régression finale avec toutes les variables de contrôle + interractions, etc. incorporées
mod <- lm(log(Prix) ~ black  + Catégorie + Ford + D_Km +   Année +
          Ess + log(Cylindrée) + Turbo +  Intérieur_cuir + Boîte_auto +
          Roues_motrices + Conduite_droite + D_Km2 +Année*D_Km +Ford*black , data = df2)
stargazer(mod, type = "text")
summary(margins(mod))

#on vérifie si les résidus suivent une loi Normale
hist(mod$residuals, main = "Histogramme des résidus", col = "lightblue", border = "black", breaks = "FD")
#ça n'en as pas trop l'air et Shapiro test confirme cela

#on vérifie l'homoscédasticité
bptest <- lmtest::bptest(mod)
print(bptest)
#il y a une hétéroscédasticité !! attention !!

#certains prérequis aux MCO ont été violés un modèle robuste serait peut-être plus approprié ?
mod2 <- rlm(log(Prix) ~ black  + Catégorie + Ford + D_Km +   Année +
           Ess + log(Cylindrée) + Turbo +  Intérieur_cuir + Boîte_auto +
           Roues_motrices + Conduite_droite + D_Km2 +Année*D_Km +Ford*black , data = df2)
stargazer(mod2, type = "text")
```

```
rm(list = ls())
#installer et charger les packages
packages_a_installer <- c("tidyverse","readr","ggplot2","ggpubr","dplyr",
"gmodels","stargazer","DataExplorer","cowplot","htmltools","psych","cros
stable",
"flextable","knitr","kableExtra","pROC","htmlTable","shiny","outliers","ca
r","lmtest", "VIM","MASS","margins")
for (package in packages_a_installer)
  {if (!requireNamespace(package, quietly = TRUE)) {
    install.packages(package)}}
for (package in packages_a_installer) {library(package, character.only =
TRUE)}
rm(package, packages_a_installer)

#charger la base de données
url <-
"https://dl.dropboxusercontent.com/scl/fi/5he67pzzxa9ce4h8o90m7/car_bas
e.csv?rlkey=udczfyeyq8z1tfwhcsd7r0ucb&"
df <- read_csv(url, col_names = TRUE)

#on regarde le nombre de modalités dans différentes catégories
sapply(df, function(x) nlevels(factor(x)))

#944 modalités dans model, on ne la gardera pas, on supprime les autres
variables inutiles
#on voit également en regardant l'ID qu'il y a des doublons (moins d'ID que
d'observations)
df <- df %>%
  distinct() %>%
  dplyr::select(-ID,-Model )
colnames(df)

#renommer les variables
nom_colonnes <- c("Prix",
"Fabricant","Année","Catégorie","Intérieur_cuir","Carburant",
"Cylindrée","Kilométrage","Type_boite","Roues_motrices","Côté_conduite
","Couleur")
colnames(df)[1:12] <- nom_colonnes
rm(nom_colonnes, url)
str(df)

#on nettoie la colonne kilométrage en enlevant les caractères non
numériques et on fait une colonne de dizaine de Km
df$Kilométrage <- as.numeric(gsub("[^0-9.]", "", df$Kilométrage))
df$D_Km <- df$Kilométrage/10000

#on renomme les modalités des carburants pour mieux les comprendre
df <- df %>%
  mutate(Carburant = ifelse(Carburant == "Diesel", "Diesel","Essence"))

#création variable binaire Turbo, boite auto, intérieur cuir et conduite à
droite pour gagner en précision
df$Turbo <- ifelse(grepl("Turbo",df$Cylindrée),1,0)
df$Cylindrée <- as.numeric(gsub("[^0-9.]", "", df$Cylindrée))
df$Conduite_droite <-  ifelse(df$Côté_conduite== "Right-hand drive", 1,
0)
df$Côté_conduite <- NULL
df$Boîte_auto <- ifelse(df$Type_boite =="automatic" , 1, 0)
df$Type_boite <- NULL
df$Intérieur_cuir <- ifelse(df$Intérieur_cuir =="Yes", 1, 0)

#idem pour le carburant et la marque de la voiture et pour la couleur noire
df$Ess <- ifelse(df$Carburant =="Essence",1, 0)
df$Ford <- ifelse(df$Fabricant=="FORD", 1,0)
df <- df %>%
  mutate(black = ifelse(Couleur == "Black", 1, 0))
df$Couleur <- NULL
```

```
# Filtrer le dataframe pour ne conserver que les catégories de voitures pour
les particuliers
categories_a_garder <- c("Cabriolet", "Coupe", "Hatchback", "Sedan",
"Minivan", "Universal")
df <- df %>%
  filter(Catégorie %in% categories_a_garder)

#détection et suppression des outliers
variables <- c("Prix", "Kilométrage","Cylindrée")
for (variable in variables) {
  Q1 <- quantile(df[[variable]], 0.25)
  Q3 <- quantile(df[[variable]], 0.75)
  IQR <- Q3 - Q1
  limite_superieure <- Q3 + 1.5 * IQR
  limite_inferieure <- Q1 - 1.5 * IQR}

#on fait un df2 qui est sans outlier
df2 <- df %>%
  filter(across(all_of(variables), ~ between(., quantile(., 0.25) - 1.5 * IQR(.),
quantile(., 0.75) + 1.5 * IQR(.))))
str(df2)

#on identifie les valeurs aberrantes pour le prix et le kilométrage (avant
suppression et après)
#graphiquement
 a <-ggplot(df, aes(y = Kilométrage)) +
   geom_boxplot(outlier.colour = "red", outlier.size = 2) +
   labs(y = "Kilométrage", title = "Kilométrage avec outliers") +
   theme_classic2() +
   scale_y_continuous(labels = scales::label_number_si())
b <-  ggplot(df, aes(y = Prix)) +
  geom_boxplot(outlier.colour = "red", outlier.size = 2) +
  labs(y = "Prix (en $)", title = "Prix avec outliers") +
  theme_classic2() +
  scale_y_continuous(labels = scales::label_number_si())
c <- ggplot(df2, aes(y = Kilométrage)) +
  geom_boxplot(outlier.colour = "red", outlier.size = 2) +
  labs(y = "Kilométrage", title = "Kilométrage sans outliers") +
  theme_classic2() +
  scale_y_continuous(labels = scales::label_number_si())
d <- ggplot(df2, aes(y = Prix)) +
   geom_boxplot(outlier.colour = "red", outlier.size = 2) +
   labs(y = "Prix  (en $)", title = "Prix  sans outliers") +
   theme_classic2() +
   scale_y_continuous(labels = scales::label_number_si())
plot_grid(a,c,b,d, ncol = 2, nrow = 2)
rm(list=setdiff(ls(), c("df", "df2")))
df2$Kilométrage <- NULL

#autres firgures pour rendre compte de la distribution : répartition des
ventes par fabricant
summary_data <- df2 %>%
  group_by(Fabricant) %>%
  summarise(count = n()) %>%
  arrange(desc(count))
a <- ggplot(summary_data, aes(x = reorder(Fabricant, -count), y = count)) +
  geom_bar(stat = "identity", fill = "skyblue", color = "black") +
  geom_text(aes(label = count), vjust = -0.5, color = "black") +
  labs(title = "Nombre de voitures par fabricant",
       x = "Fabricant",
       y = "Nombre de voitures") +
  theme_classic2() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```