Université de Carthage

école Supérieure de la statistique et de l'analyse de l'information

Rapport de projet de fin d'année

WebScraping et Analyse comparative des prix

Réalisé Par
Azza DAKHLI

Encadrante :
Mme. Haifa BEN MASSOUD

Entreprise d'étude :

Février-Mai 2021

Remerciements

Tout d'abord, je tiens à remercier mon encadrante de projet Madame Haifa BEN MASSOUD, pour son accompagnement durant toute la période de travail, ainsi que pour avoir pris le temps de veiller à l'avancement de mon projet de fin d'année.

Enfin, je tiens également à remercier toutes les personnes qui ont participé de près ou de loin à la réalisation de ce travail.

Table des matières

Introduction **4**

1 Présentation du produit d'étude : **5**
1.1 Présentation de la société Mercedes 6
1.2 Description du produit 6
1.3 Les concurrents de la marque Mercedes 6

2 Webscraping **8**
2.1 Introduction 9
2.2 Les librairies utilisées 9
2.3 Les méthodes 10
2.3.1 Extraction de la base 10
2.3.2 Nettoyage de la base 11
2.4 La Base de données finale 12

3 Le tableau de bord **13**
3.1 Power bi 14
3.2 Le tableau de bord 14
3.3 Interprétations des résultats 14

Conclusion **16**

Bibliographie **18**

Annexe **20**

Table des figures

3.1 Installation et chargement des packages 21
3.2 Script d'accés au site web 21
3.3 Script d'extraction des données 22
3.4 Script du nettoyage des données 22
3.5 Un aperçu de la base de données extraite 23
3.6 Représentation du prix en $ CAD en fonction de l'accélération 0-100km/h 23
3.7 Représentation du prix en $ CAD en fonction de la puissance(ch) 24
3.8 Représentation de l'accélération 0-100km/h de la voiture en fonction de la puissance(ch) 24
3.9 Deuxième interface du tableau de bord 25

Introduction

Dans le cadre de mes études à l'école supérieure de la statiqtiques et de l'analyse de l'information, j'ai été amené à effectuer un projet de fin d'année.Une analyse comparative des prix des voitures SUV de Mercedes est le sujet du projet.

Comparer les prix de la concurrence est une technique de marketing utilisée par les grandes entreprises , ayant pour but de comprendre la position concurrentielle de sa propre offre par rapport à celle des concurrents. Une technique qui permet à l'entreprise de conquérir un marché aussi puissant et dynamique que celui des voitures.

Dans le cadre de ce projet , j'ai découvert le **Webscrpaing** et ses outils de base et comment l'exploiter pour effectuer une analyse comparative des prix. Un vrai plus pour mon projet professionnel , puisque ce projet permet de conquérir à la fois le domaine du marketing et le domaine de l'informatique . En outre , ce stage est une opportunité pour renforcer mes compétences pratiques en tant qu'ingénieur en statistiques et analyse de l'information, d'ailleurs , j'ai eu l'avantage de pratiquer les langages informatiques , notamment le langage R , ainsi que pratiquer les modules de statistiques descriptive appris lors de la première année des études d'ingénieurs ,

Ce document contient une présentation de l'analyse descriptive de la concurrence des voitures SUV avec une brève présentation sur la marque d'étude **Mercedes** dans le premier chapitre.

Chapitre 1

Présentation du produit d'étude :

1.1 Présentation de la société Mercedes

Mercedes-Benz est un constructeur allemand d'automobiles, de camions, d'autocars et d'autobus indépendant fondé en 1926 par trois autres constructeurs : Daimler-Motoren-Gesellschaft, Mercedes et Benz & Cie . Son siège social est situé à Stuttgart en Allemagne. Au cours de l'année 2020 l'entreprise réalise un chiffre d'affaires de $ 5 188 757 274,98 CAD [1].

1.2 Description du produit

Mercedes possède différentes type de carosserie :cabriolet , citadine, Berline.etc. Pour cette étude j'ai choisi de travailler sur les **SUV**, un type de carosserie existant chez tous les constructeurs d'automobiles.
Les **SUV :**

- *sport*
- *utility*
- *car*

sont principalement caractérisées par leur forme bicorps, inspirée des breaks, leur hauteur et leur volume importants. Par rapport aux autres types de véhicules automobiles de même taille, l'espace intérieur est souvent plus important et permet aussi bien de transporter des objets assez grands que de voyager en famille, à la manière d'un monospace ou d'un ludospace. La puissance d'une voiture **SUV** de Mercedes varie entre 221Ch-416Ch quant à son accélération 0-100km/h varie entre 6.7s - 4.9s

1.3 Les concurrents de la marque Mercedes

Une entreprise aussi puissante que Mercedes ne peut prendre qu'une place de leader sur le marché canadien ou même mondial , elle fait donc face à plusieurs concurrents. Les constructeurs suivants sont ordonnés selon leurs intensités concurrentielles :

- **BMW :** Un constructeur de véhicules allemand fondé en 1916 par Gustav Otto et Karl Friedrich Rapp. Dans cette étude je me contente d'explorer la gamme X de BMW. Son chiffre d'affaire de l'année 2020 est de $142,41 CAD.
- **Audi :** Un constructeur d'automobiles allemand fondé par August Horch en 1909.La gamme étudiée est la gamme Q. En 2020, son chiffre d'affaire atteint $73 milliards en CAD.
- **Jeep :** Jeep est un constructeur automobile américain Jeep a commencé à produire des véhicules pour le marché civil en 1945.Le chiffre d'affaire se sa maison mére Fiat Chrysler Automobiles est de $126,50 CAD en 2020.

1. Dollar canadien

— **Toyota :** Officiellement Toyota Motor Corporation est un constructeur automobile originaire du Japon. Toyota possède la valorisation la plus élevée au monde dans le secteur automobile et la huitième mondiale toutes activités confondues avec $49,71 milliards de dollars (2017).
— **Alfa romeo :** Un constructeur issu de l'organisation mère Stellantis italienne fondé en 1910 . Son chiffre d'affaire en 2020 est de $196,37 CAD.

Chapitre 2

Webscraping

2.1 Introduction

Le Web scraping consiste à extraire des données de sites Internet et à les enregistrer afin de les analyser ou de les utiliser de toute autre façon. Le scraping permet de collecter des informations de nature bien différente. Il peut par exemple s'agir de coordonnées comme des adresses e-mail ou des numéros de téléphone, mais aussi de mots-clés individuels ou d'URL. Ces informations sont alors rassemblées dans des bases de données locales ou des Data frame comme le cas de la base de donnée de cette étude .Ce chapitre est une préparation pour l'étude réalisée.Je présente ci-dessous les différentes librairies utilisées ainsi que les méthodes et la démarche d'extraction et de nettoyage des données.

2.2 Les librairies utilisées

Dans cette étude , le langage utilisé est le langage **R** , un langage destiné aux statisticiens et aux data scientist [1]. L'installation d'une librairie se fait sous R par le biais de la commande *install.package(...)* (voir annexe Figure 3.1) Pour effectuer le webscraping , j'ai eu recours à cinq librairies y compris les librairies d'extraction et de nettoyage de la base

- **Rselenium :** Une bibliothèque qui sert comme outil pour automatisation du navigateur, elle fournit des extensions afin de reproduire des interactions utilisateur avec les navigateurs. RSelenium est exclusivement basée sur le HTML et le JavaScript et permet aux développeurs de tester et enregistrer les interactions avec une application Web afin de les répéter ensuite aussi souvent que souhaité, de façon entièrement automatisée.
- **Rvest :**il s'agit d'un package qui vous permet de parser (autrement dit de parcourir et d'aller chercher) le contenu d'une page web, pour le rendre exploitable par R.Par exemple de créer une liste depuis une page Wikipédia, récupérer un texte sur une page , le transformer le tableau html extrait en data.frame.Cette bibliothéque est indispensable à l'étape de l'extraction des données qui seront ensuite stockée dans le data frame.
- **Stringr :**Aprés extraction des données sous forme de chaîne de caractére, nettoyer les informations extraites est necéssaire pour une manipulation plus facile de la base et pour mieux représenter les graphiques necéssaire à l'analyse comparative.Cette librairie fournit des commandes de manipulation des chaînes de caractères permettant de combiner,remplacer et effectuer des recherches dans celles-ci.
- **Xml2** La librairie xml2 est une liaison à libxml2 ,une librairie de R.Il s'agit d'un package qui facilite le travail avec les langages HTML et XML à l'aide d'une interface simple et cohérente.
- **Writexl :** Il s'agit d'une librairie d'extraction de donnée : il permet au développeur de prélever les données obtenues du webscraping et les sauvegarder sous un fichier excel pour mieux pouvoir ,par la suite,les importer

1. mot anglais désignant un spécialiste de la science des données

aux logiciels de visualisation des données comme Powerbi,Dash.etc.

2.3 Les méthodes

2.3.1 Extraction de la base

L'extraction de la base se fait principalement grâce aux deux librairies mentionnées ci-dessus **Rselenium** et **rvest**.Je commence par ouvrir le serveur de navigation par la commande try(rsDriver(port=4444L,browser='..')) [1] .J'ai entamé par la suite la navigation. Pour ce faire , j'ai créé une variable remDr en lui affectant la commande remoteDriver() [1].Cette commande communique avec le serveur RSelenium. Si une erreur se produit lors de l'exécution de la commande, le serveur renvoie un code d'erreur HTTP avec une réponse codée JSON qui indique le code d'erreur de réponse précis.

Puis , j'ai eu recours à remDr$open() [1] et remDr$navigate("..") [1] en affectant l'URL du site web à scraper ,ceci a pour but de naviguer dans le site web désiré.En outre ,Pour faire défiler la page web jusqu'à la fin j'ai utilisé remDr$findElement() [1] et sendKeysToElement() [1].Puis ,il faut bloquer l'exécution de ces commandes afin de charger la page web toute complète et ceci est par le biais de la commande Sys.sleep(5) [1].

J'ai affecté remDr$getPageSource() [1] à un objet html pour extraire le code html de la page et j'ai utilisé read_html() [2] qui sert à lire ce fichier html (Voir annexe Figure 3.2).Finalement,j'ai choisi le xpath de l'élément à extraire grâce à la commande html_nodes() [2] et puis le converir en un fichier text par la commande html_text() [2](Voir annexe Figure 3.3).Pour conclure ,Voir tableau ci-dessous.

1. Une commande de la librairie Rselenium
2. Une commande de la librairie Rvest

Commande	Librairie	Utilité
try(rsDriver(port=4444L,browser='..')	Rselenium	Ouvrir le serveur
remoteDriver()	Rselenium	Communiquer avec le serveur
remDr$open() et remDr$navigate()	Rselenium	Ouvrir le lien de la page web
remDr$findElement() et sendKeysToElement()	Rselenium	faire défiler la page web
Sys.sleep(5)	Base	Bloquer l'exécution du script jusqu'à la fin du téléchargement de la page
remDr$getPageSource()	Rselenium	Obtenir le code html
read_html()	rvest	lire le code html
html_nodes()	Rvest	extraire l'élément dont le xpath est donné entre parenthèse
html_text()	Rvest	Convertir les informations extraites en texte

TABLE 2.1 – Commandes d'extraction des données

2.3.2 Nettoyage de la base

2.3.2.0.1 Extraire les données est compliqué ,néanmoins, ce n'est pas l'étape la plus difficile car le format des informations extraites n'est pas toujours optimal , les informations peuvent contenir des caratères futiles qui peuvent générer des erreurs au niveau des études statistiques.Voici un exemple de la data frame extraite du site **Mercedes**.

```
[1] "$42,400\n                    \n   \n   \n   \n   \n   \n   \n\n   \n      \n   \n      \n
  \n          \n                  Disclaimer*"
[2] "$46,500\n                    \n   \n   \n   \n   \n   \n   \n\n   \n      \n   \n      \n
  \n          \n                  Disclaimer*"
[3] "$49,900\n                    \n   \n   \n   \n   \n   \n   \n\n   \n      \n   \n      \n
  \n          \n                  Disclaimer*"
[4] "$53,900\n                    \n   \n   \n   \n   \n   \n   \n\n   \n      \n   \n      \n
  \n          \n                  Disclaimer*"
[5] "$69,900\n                    \n   \n   \n   \n   \n   \n   \n\n   \n      \n   \n      \n
  \n          \n                  Disclaimer*"
[6] "$94,500\n                    \n   \n   \n   \n   \n   \n   \n\n   \n      \n   \n      \n
  \n          \n                  Disclaimer*"
[7] "$101,900\n                    \n   \n   \n   \n   \n   \n   \n\n   \n      \n   \n      \n
   \n          \n                  Disclaimer*"
[8] "$199,400\n                    \n   \n   \n   \n   \n   \n   \n\n   \n      \n   \n      \n
   \n          \n                  Disclaimer*"
[9] "$154,900\n                    \n   \n   \n   \n   \n   \n   \n\n   \n      \n   \n      \n
   \n          \n                  Disclaimer*"
>
```

Pour nettoyer cette base j'ai eu recours à la librairie **stringr** et **stringi**.Pour commencer j'ai utilisé la commande str_trim() [1] pour supprimer les espaces et stri_replace_all_regex() [2] pour remplacer tout caratère inutile par un caratère

1. Commande de la librairie stringr
2. Commande de la librairie stringi

vide et puis l'appliquer sur toutes les chaînes de caratères de notre data frame par la commande **lapply()**.(Voir annexe Figure 3.4)

2.4 La Base de données finale

Pour extraire la base de donnée finale , il faut répéter la démarche mentionnée ci-dessus, pour chacun des sites des concurrents afin d'obtenir la base de donnée finale(voir Annexe Figure 3.5).

Chapitre 3

Le tableau de bord

Dans ce chapitre, je vais commencer par donner une brève introduction sur le logiciel de travail **Power bi** ,puis, je vais présenter le tableau de bord de l'analyse comparative.

3.1 Power bi

Power bi Desktop est une application d'analyse de données de **Microsoft**.Ce logiciel permet à son utilisateur de se connecter à des données, de les transformer et de les visualiser.Il permet de créer des tableaux de bord personnalisés ainsi que des rapports intéractifs. Ces visualisations peuvent être par la suite publiées sur l'espace de travail et partagées entre plusieurs utilisateurs.A part les fonctionnalités déjà existants sur Power bi comme les graphes et les cartes , le logiciel fournit des éléments visuels pour les scripts R ou python qui peuvent être utiles pour les développeurs.

Pour commencer il faut importer la base de données ou les bases de données de travail , pour ce faire , Power bi offre une variété de type de fichier de base de données : Excel , Serveur SQL , jeux de données,etc.Il permet aussi d'effectuer des modifications sur les colonnes de la base.

3.2 Le tableau de bord

Le tableau de bord effectué se compose de trois pages . Une première page introductive contenant le titre et deux boutons : Un premier bouton conduisant à une page contenant des visualisations des données sous forme des graphes, ces graphes sont des représentations linéaires des différentes variables de la base de données.Dans un premier lieu, j'ai étudié la variation du prix en fonction de l'accélération(0-100km/h) (voir annexe Figure 3.6) ainsi qu'en fonction de la puissance(Voir annexe Figure 3.7) et finalement pour montrer la corrélation entre ces deux caractéristiques d'une voiture ,j'ai représenté la variation de l'accélération en fonction de puissance (voir annexe Figure 3.8), ceci a pour but de faciliter la lecture des graphes de la comparaison effectués par la suite.
La deuxième page contient l'anlyse comparative.Elle est composée de 3 diagrammes de Kiviat(ou graphique en toile d'araignée),un type de graphique utilisée le plus souvent dans les analyses comparatives entre les produits et les concurrents, accompagnées d'un segment intéractif : l'utilisateur peut choisir de la liste l'ensemble de voiture à comparer en terme d'accélération , de puissance et de prix.(voir annexe Figure 3.9)

3.3 Interprétations des résultats

Difficile, voire impossible, de classer tous les SUV compacts disponibles sur le marché , du meilleur au moins recommandable. Pourquoi ? Parce qu'ils sont très nombreux, parce qu'ils ne se négocient pas tous au même tarif, ou parce que tout simplement, parce que les acheteurs n'ont pas tous la même sensibilité

à certains détails.Si je me restreint alors à la base de données extraite et aprés avoir effectuer l'étape du webscraping , le plus important est d'analyser ces données afin de se renseigner sur le concurrent le plus puissant face à notre marque d'étude **Mercedes**.Chacun de ses concurrents se caractérise par des points faibles et des points forts de sa marque.D'aprés la base de données de notre étude , j'ai choisi la meilleure voiture de chaque marque qui se caractérise avec un rapport qualité-prix optimal.L'ordre donné ci-dessous n'est influencé que par les trois caractéristiques accélération(0-100km/h) , puissance(Ch) et prix(en CAD) :

- **Audi Q5** :Elle Vient en premier lieu, avec une puissance 261 cheveux , une accélération de 5 secondes et un prix de 55 400$ CAD.
- **Toyota Prime SE AWD** :Elle est en 2ème lieu , avec une accélération de 6.4 accélération , une puissance de 302 Cheveux et un prix de 44 990$ CAD.
- **Mercedes GLC Coupé** :Le produit d'étude vient en 3ème lieu se caractérisant par une accélération 6.2 seconde et une puissance de 255 Ch .Son prix est de 53 900$ CAD.
- **BMW X4** : Cette voiture de la marque de la plus grande concurrence avec Mercedes vient en 4ème lieu avec son accélération de 6.3 seconde et une puissance de 248Ch. Son prix est dans la même marge de prix que la Mercedes GLC coupé avec un prix de 56 300$ CAD.
- **Alfa Romeo Sprint** :Contrairement au reste des voitures de la marque Alfa Romeo, cette voiture est caractérisée par une performance (accélération,puissance) remarquable,elle est classée 5éme parmi les produits d'étude, avec une accélération de 5.6 secondes et une puissance de 280 Ch.Son prix est d'environ 53 095$ CAD.
- **Jeep Wrangler** : Cette voiture italienne vient en dernier lieu avec une accélération de 8.1 secondes , une puissance de 281 et un prix aux alentours de 35 122 $ CAD.

Conclusion

Ce projet de fin d'année est enrichissant d'un point de vue pratique ainsi que de point de vue personnel.

D'un point de vue pratique, le projet m'a permis de mettre en avant mes compétences en programmation en langage R et de les enrichir . J'ai, d'autre part , découvert le monde des voitures et la puissance de la concurrence entre les marques, ainsi que les différents moyens ainsi quer les graphiques adéquats utiles pour effectuer une analyse comparative.

De surcroît, les difficultés trouvées à scraper les sites sont utiles à faire évoluer les connaissances en ce domaine sur les différents moyens utilisés par les grands développeurs du web pour surmonter ces difficultés. D'un point de vue plus personnel, ce projet aide, tout d'abord , à pratiquer la communication à l'oral grâce aux présentaions hebdomadaires , on a aussi pratiquer les bonne manière de gestion du temps : le travail a été divisé sur plusieurs semaines du semestre.

Il est par ailleurs évident que le travail effectué n'est qu'une étape primaire aussi bien pour d'autres projets plus approfondis.

Bibliographie

[1] https ://www.r-project.org/

Annexe

```
install.packages("RSelenium")
install.packages("rvest")
install.packages("stringr")
install.packages("xml2")
install.packages("writexl")

library(RSelenium)
library(rvest)
library(stringr)
library(xml2)
```

FIGURE 3.1 – Installation et chargement des packages

```
try(rsDriver(port=4444L,browser='firefox'))

remDr<-remoteDriver()
remDr$open()
remDr$navigate("https://www........ca/en/home.html")


Sys.sleep(5)

webElem <- remDr$findElement("css", "body")
webElem$sendKeysToElement(list(key = "end"))

Sys.sleep(5)

morereviews <- remDr$findElement(using = 'xpath', "/html/body/...")
morereviews$clickElement()

Sys.sleep(5)

webElem <- remDr$findElement("css", "body")
webElem$sendKeysToElement(list(key = "end"))

Sys.sleep(5)
html <- remDr$getPageSource()[[1]]
html1<- read_html(html)
```

FIGURE 3.2 – Script d'accés au site web

```r
get_prices<-function(html){
  html %>%
    html_nodes(xpath='....') %>%
    html_text()%>%
    str_trim()%>%
    unlist()
}
lp<-get_prices(html1)
names(lp)<-c(1:12)
get_name<-function(html){
  html %>%
    html_nodes(xpath='...') %>%
    html_text()%>%
    str_trim()%>%
    unlist()
}

ln<-get_name(html1)
ln
```

FIGURE 3.3 – Script d'extraction des données

```r
remove<-function(l1){
  stri_replace_all_regex(l1, "[\n Disclaimer]", "")
}
l<-remove(l)
l<-sapply(l, pas)
```

FIGURE 3.4 – Script du nettoyage des données

noms	prix	puissance	accélération
Mercedes GLA SUV	42 400	221	6,7
Mercedes GLB SUV	46 500	221	6,9
Mercedes GLC SUV	49 900	255	6,2
[illegible]	[illegible]	[illegible]	[illegible]
Toyota LE FWD	28 590	203	6,8
Toyota LE AWD	30 690	203	6,9
Toyota XLE FWD	32 190	203	6,8
Toyota Hybrid LE AWD	32 950	219	6,3
Jeep Wrangler	35 122	284	8,1
Jeep Cherokee	30 557	185	8,8
Jeep Compass	26 394	180	8,9
BMW X1	39 990	228	6,5
BMW X2	43 275	228	6,5
BMW X7	102 900	335	6,1
Audi e-tron	85 600	402	5,7
Audi Q8	82 550	335	6,0
Alfa Romeo Sprint	53 095	280	5,6
Alfa Romeo Stelvio Ti	57 595	280	5,6
Alfa Romeo Stelvio Ti Sport	61 145	280	5,6
Alfa Romeo Stelvio Quadrifoglio	97 845	505	3,6

FIGURE 3.5 – Un aperçu de la base de données extraite

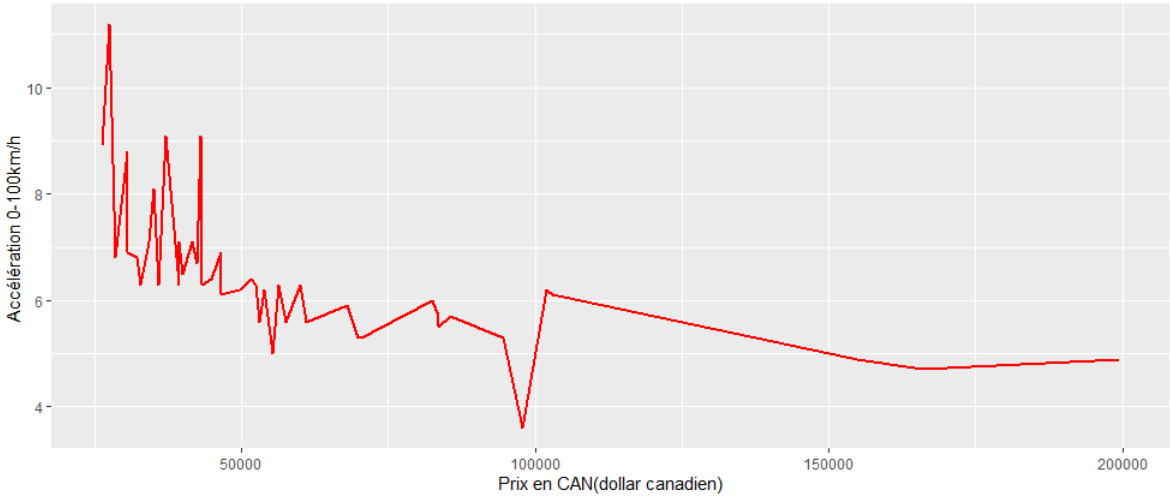

FIGURE 3.6 – Représentation du prix en $ CAD en fonction de l'accélération 0-100km/h

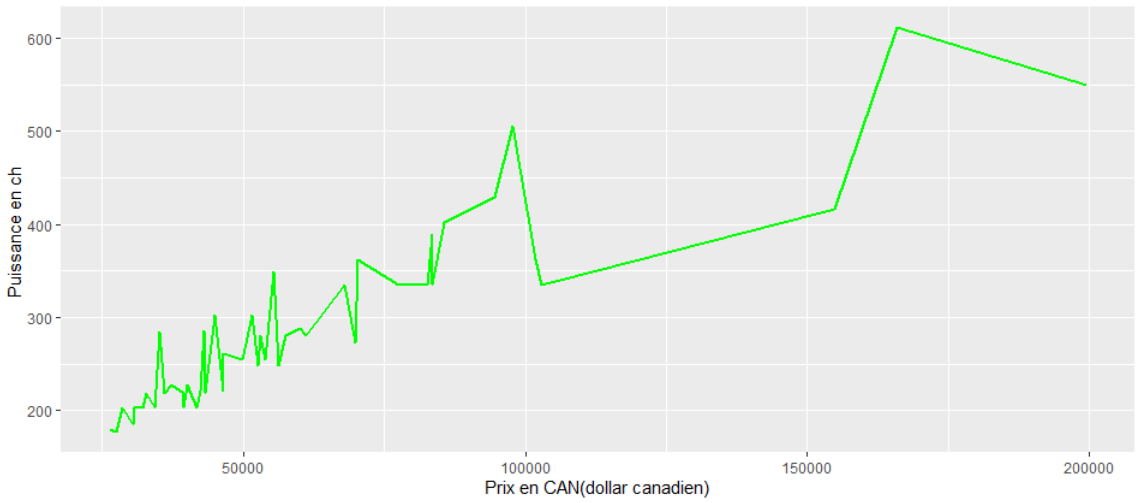

FIGURE 3.7 – Représentation du prix en $ CAD en fonction de la puissance(ch)

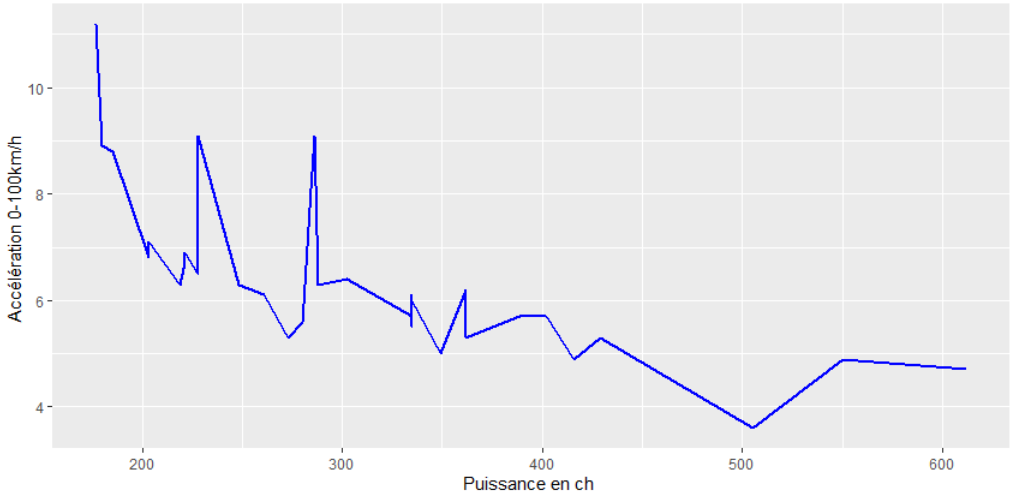

FIGURE 3.8 – Représentation de l'accélération 0-100km/h de la voiture en fonction de la puissance(ch)

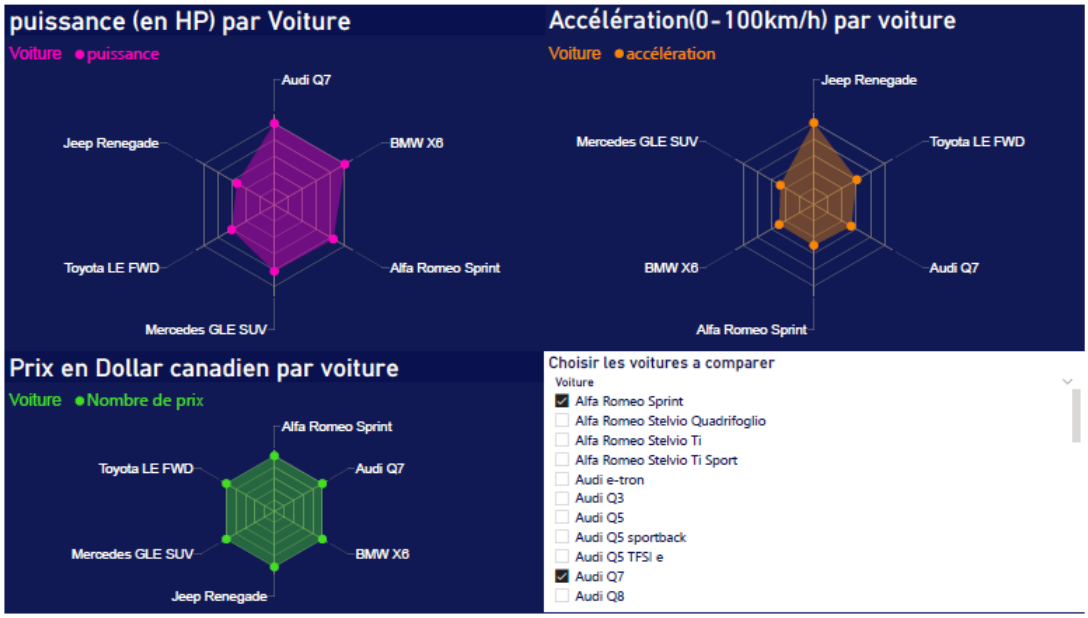

FIGURE 3.9 – Deuxième interface du tableau de bord