Université Chouaib Doukkali
Ecole Nationale des Sciences Appliquées d'El Jadida
Département Télécommunications, Réseaux et Informatique

Filière : **2ITE**
Niveau : **2ème Année**

MINI PROJET BUSINESS INTELLIGENCE

SUJET :

Mise en place d'une solution de business intelligence
Cas : entreprise de vente des voitures au Maroc

Réalisé Par :
BAKANZIZE SAAD
LAAZIZ AHMED
KARTBOUNI ANAS

Encadré par :
Prof. Mr.HANINE MOHAMMED

DEDICACES

A la mémoire de nos grands parents

Qui ont été toujours dans nos esprits et dans nos cœurs, on vous dédie aujourd'hui ce travail.

Que Dieu, le miséricordieux, vous accueille dans son éternel paradis.

A nos très chers pères

Autant de phrases et d'expressions aussi éloquentes soient-elles ne sauraient exprimer nos gratitudes et nos reconnaissances. On vous doit ce qu'on est aujourd'hui et ce qu'on sera demain et on fera toujours de notre mieux pour rester vos fiertés et ne jamais vous décevoir. Que Dieu le tout puissant vous préserve, vous accorde santé, bonheur, quiétude de l'esprit et vous protège de tout mal.

A nos très chères mères

Autant de phrases aussi expressives soient-elles ne sauraient montrer le degré d'amour et d'affection que on éprouve pour vous. Vous nous avez comblé avec votre tendresse et votre affection tout au long de notre parcours. Que Dieu le tout puissant vous préserve, vous accorde santé, bonheur, quiétude de l'esprit et vous protège de tout mal.

A nos chers amis

On vous dédie ce travail.

REMERCIEMENTS

Il nous est agréable de nous acquitter d'une dette de reconnaissance auprès de toutes les personnes, dont l'intervention au cours de ce projet a favorisé son aboutissement. Nos remerciements vont d'abord au Créateur de l'univers qui nous a maintenu en santé pour mener à bien cette année d'étude. Nous tenons aussi à adresser nos remerciements à nos familles qui nous ont toujours soutenus et poussés à continuer nos études. Ce présent travail a pu voir le jour grâce à leur soutien. Ainsi, nos remerciements les plus sincères vont à prof. HANINE MOHAMED, notre encadrant pédagogique, pour les conseils qu'il nous a prodigués, son soutien, son encadrement judicieux et son assistance tout au long de la réalisation du projet. Aussi nous tenons à remercier tout le cadre professoral de l'ENSAJ, pour la formation prodigieuse qu'il nous a prodigué. Que tous ceux et celles qui ont contribué, de près ou de loin, à l'accomplissement de ce travail, trouvent l'expression de nos remerciements les plus sincères.

RESUME

Après avoir extrait automatiquement une grande quantité de données des deux sites d'automobiles avito.ma et moteur.ma grâce au web scraping. C'est le moment de les exploiter à l'aide des solutions de Business intelligence.

Le présent rapport vient pour formaliser et présenter notre projet, il montre tous les résultats de notre travail et les méthodes suivies pour y arriver.

Ce rapport illustre toutes les étapes de la chaine décisionnelle. En se débutant par la préparation des données, et en se terminant par leur exploitation. Comme il expose aussi les définitions de plusieurs technologies (langages, bibliothèques, logiciels) utilisées en informatique décisionnelle ainsi que la différence entre eux.

ABSTRACT

After having automatically extracted a large amount of data from the two car sites avito.ma and moteur.ma thanks to web scraping. It is now time to exploit them with the help of business intelligence solutions.

This report comes to formalize and present our project, it shows all the results of our work and the methods followed to achieve it.

This report illustrates all the steps of the decision-making chain. Starting with the preparation of the data, and ending with their exploitation. As it also exposes the definitions of several technologies (languages, libraries, software) used in business intelligence and the difference between them.

TABLE DES MATIERES

TABLES DES FIGURES

LISTE DES ABREVIATIONS

ETL : Extract Transform load

SQL: Unified Modeling Language

CSV: Comma Separated Values

XML: Extensible Markup Language

PDI: Pentaho Data Integration

DW: Data Warehouse

BI : Business Intelligence

INTRODUCTION

Les données sont des éléments essentiels de toute recherche, qu'elles soient académiques, marketing ou scientifiques. Cependant ces données sont entrain de connaitre une croissance exponentielle, la quantité de données sur la planète est tout simplement énorme, cela crée une certaine difficulté pour gérer l'ensemble de ces données, et que la vraie question c'est de savoir comment obtenir la bonne information auprès de la bonne personne au bon moment. L'utilisation des bonnes données permet aux entreprises et aux scientifiques de prendre de meilleures décisions, comme faire des estimations, établir des prédictions mathématiques, ou encore effectuer des analyses de sentiments. Les entreprises sont prêtes à payer n'importe quel prix pour mettre la main sur des données relatives à leurs activités. C'est là qu'intervient le Business intelligence (informatique décisionnelle) qui est un ensemble de solutions informatiques permettant l'analyse des données de l'entreprise, afin d'en dégager les informations qualitatives nouvelles qui vont fonder des décisions, qu'elles soient tactiques ou stratégiques. Les systèmes de l'informatique décisionnelle sont utilisés par les décideurs pour obtenir une connaissance approfondie de l'entreprise et de définir et de soutenir leurs stratégies d'affaires, par exemple d'acquérir un avantage concurrentiel, d'améliorer la performance de l'entreprise, de répondre plus rapidement aux changements, d'augmenter la rentabilité, ou d'une façon générale la création de valeur ajoutée de l'entreprise.

Si Internet est une énorme bibliothèque de données, il reste à savoir où trouver des données utiles. Avec la quantité de données en jeu, il est tout simplement impossible de passer au crible et de trouver manuellement les "meilleures" informations. C'est un travail très fastidieux et prend beaucoup de temps. La solution est la technique de web scraping qui définit des programmes informatiques qui sont chargés d'extraire ces informations automatiquement dans un minimum de temps. Notre dataset est constituée de données issues des deux sites avito.ma

et moteur.ma grâce au web scraping. A l'aide des solutions du BI, ces données seront nettoyées, puis chargées dans un datawarehouse, et enfin analysées puis exploitées pour la génération des tableaux de bords.

Ce rapport contient trois chapitres. Dans le premier chapitre on va présenter notre dataset et son origine. Dans le second chapitre, on va définir la phase d'ETL ainsi que les technologies utilisées dans cette phase, et on va exposer comment on a fait notre ETL, et le datawarehouse dans laquelle on a chargé les données. Enfin dans le troisième chapitre, on va définir la phase de l'analyse et restitution ainsi que les technologies utilisées dans cette phase, et on va exhiber les résultats et les tableaux de bord.

I-DATASET

I-1 Définition d'une dataset

Un data set est une collection d'éléments de données connexes et discrets qui peuvent être consultés individuellement ou en combinaison ou gérés comme une entité entière.

Un ensemble de données est organisé en un certain type de structure de données. Dans une base de données, par exemple, un ensemble de données peut contenir des données commerciales (noms, salaires, coordonnées, chiffres d'affaires, etc.). La base de données elle-même peut être considérée comme un ensemble de données, tout comme les data set qu'elle contient et qui sont liés à un type particulier d'informations, comme les données de vente d'un service particulier de l'entreprise.[1]

I-2 Dataset

Notre dataset contient beaucoup de données sur le secteur automobile marocain. Ces données sont issues des deux sites AVITO.MA et MOTEUR.MA grâce au web scraping. Alors qu'est-ce que le web scraping ?

Le scraping définit de façon générale une technique permettant d'extraire du contenu d'un ou de plusieurs sites web de manière totalement automatique. Ce sont des scripts, des programmes informatiques, qui sont chargés d'extraire ces informations. Ce processus est comparable à un copier-coller automatique.

Le web scraping a plusieurs utilités. Il permet d'abord de réutiliser des contenus présents sur un site web pour l'afficher sur un autre site web, et ainsi multiplier sans effort le nombre de pages disposant d'un même contenu. Comme il permet aussi l'exploitation des données extraites après avoir les insérées dans des bases de données SQL, des fichier CSV, des fichiers EXCEL ou des fichiers XML. [2]

Voilà le lien vers notre dataset :

https://drive.google.com/file/d/1Q8YAUF2tCS_lzz0E_OIA1M7xyfJFNiX2/view

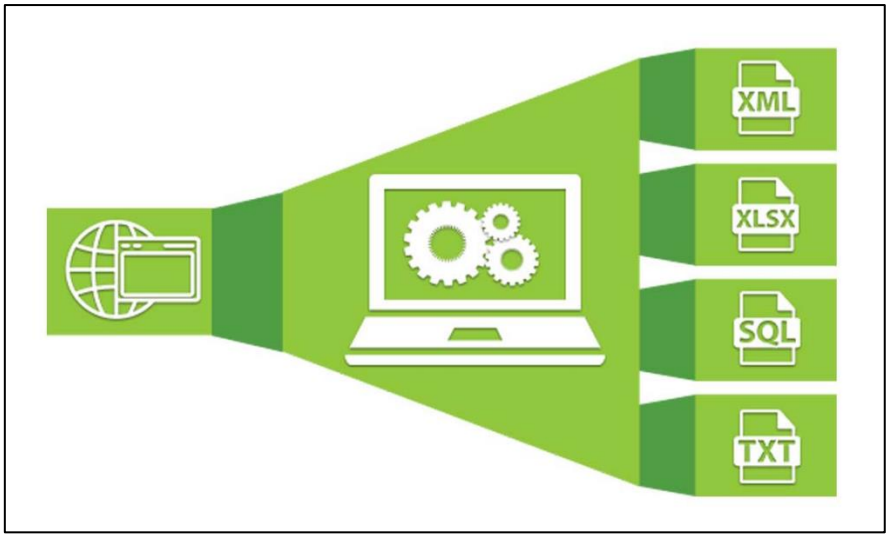


Figure 1: Processus du web scraping

SITE SCRAPEES

Dans ce projet, on a scrapé deux sites : MOTEUR.MA, AVITO.MA.

A. Moteur.ma Officiellement lancée en 2015, MOTEUR.MA Offre un large choix quantitatif de voitures d'occasion, et propose des centaines de modèles de marques neuves les plus récents au Maroc. [3]

Figure 2: Logo du site moteur.ma

B. Avito.ma Officiellement lancée en 2012, AVITO.MA est un site de vente qui contient plusieurs articles (voitures, vêtements, appartements…), mais pour nous on s'intéresse seulement à la section des voitures.

Figure 3: Logo du site Avito.ma

II-ETL

II-1 Définition

L'ETL est un processus d'intégration des données qui permet de transférer des données brutes d'un système source, de les préparer pour une utilisation en aval et de les envoyer vers une base de données, un entrepôt de données ou un serveur cible. Dans ce processus la transformation des données intervient sur un serveur intermédiaire avant le chargement sur la cible. Cette fonction s'avère particulièrement utile pour le traitement de vastes ensembles de données hétérogènes dans le cadre de l'analytique du Big Data et de l'informatique décisionnelle. Une variante est l'ELT qui permet le chargement des données brutes directement sur la cible, où elles seront alors transformées. L'un des principaux attraits de l'ELT tient à la réduction des délais de chargement par rapport au modèle ETL. En effet, tirer parti de la capacité de traitement intégrée à l'infrastructure d'un entrepôt de données permet souvent de diminuer le temps nécessaire au transfert des données et peut se révéler plus économique. ETL correspond à une surface de travail, il représente l'intermédiaire entre le système opérationnel et l'interface du système décisionnel. [6]

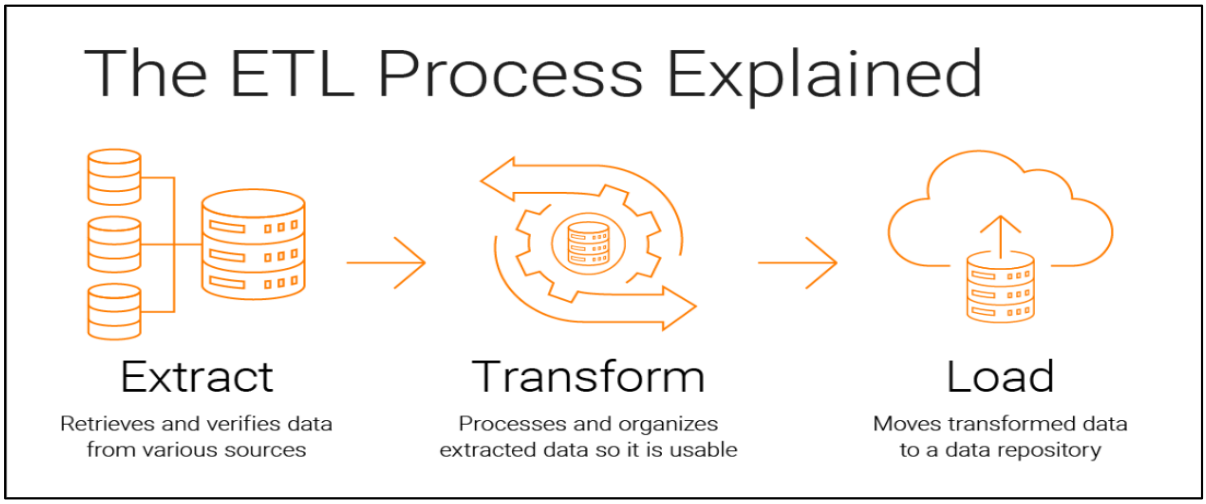


Figure 4: Processus d'un ETL

Dans notre projet, on va essayer de réaliser l'ETL avec deux méthodes, premièrement avec KETTLE et deuxièmement avec PYTHON, puis on va comparer entre les deux.

II-2 Kettle

PDI (Pentaho Data Integration), qui était auparavant connu sous le nom de Kettle, est un logiciel d'ETL (Extract, Transform, Load) Open Source qui permet la conception ainsi que l'exécution des opérations de manipulation et de transformation de données très complexes.

Son principal intérêt est de récupérer diverses sources dans divers formats, les traiter, les transformer, et former un résultat puis finalement exporter dans le format souhaité vers une destination souhaitée. Tout ceci se fait de visuellement en créant des étapes et en éditant le détail de chaque étape.[7]

Figure 5: Logo de Pentaho

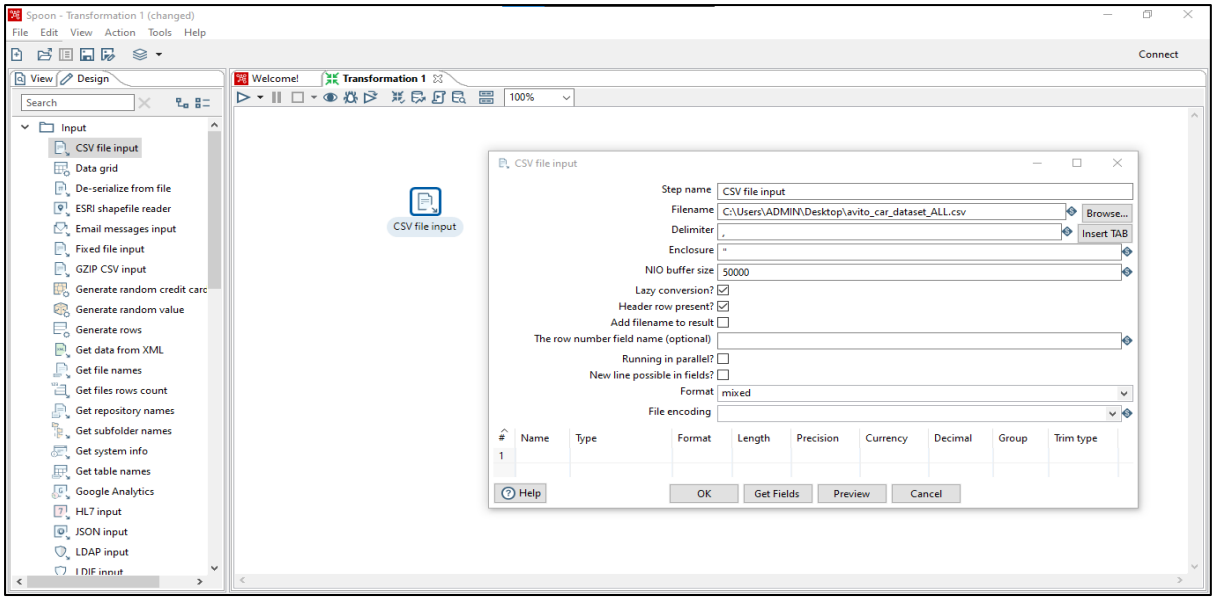


Figure 6: Stockage des données dans un fichier csv

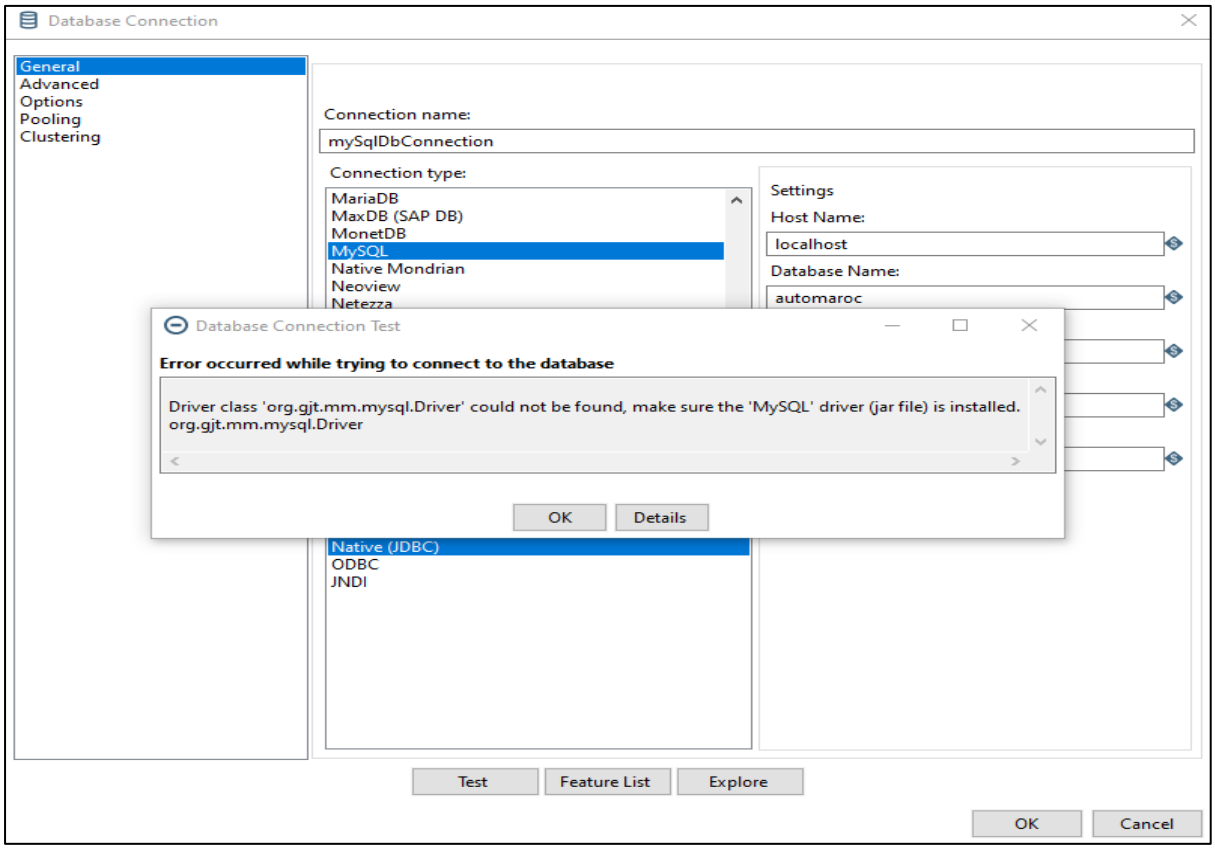


Figure 7: Illustrantion l'essaie de la connexion à la base de données MySQL

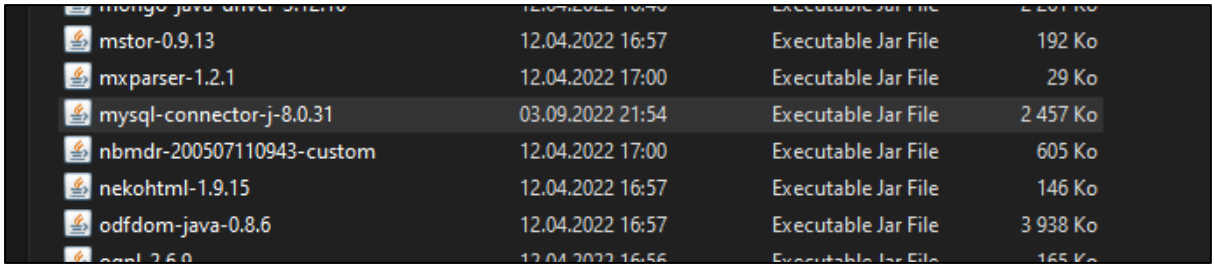


Figure 8: Illustration de la configuration de mysql-connector

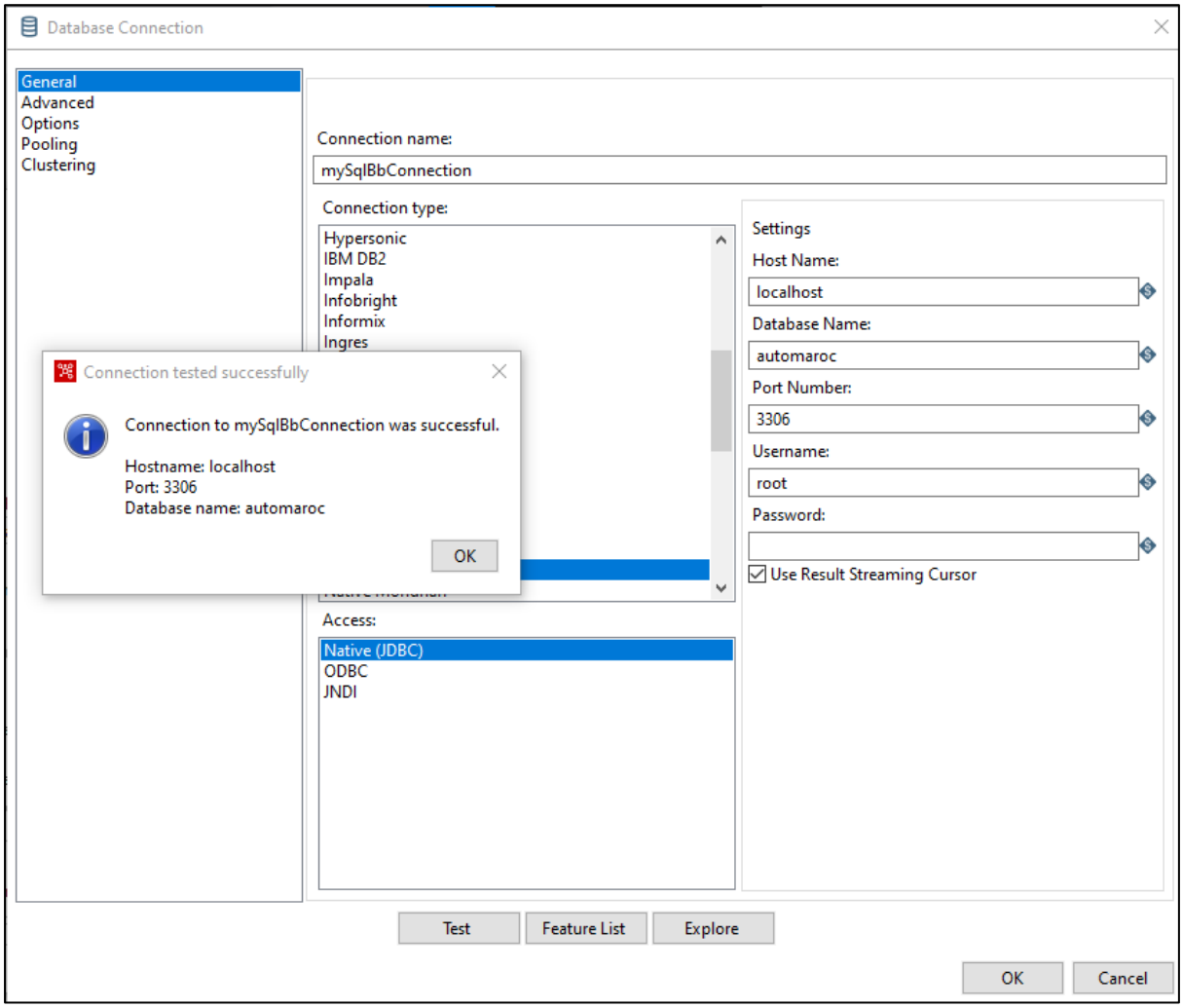


Figure 9: Illustration de la connexion avec la base de données

C'est le début de la phase ETL à l'aide de l'outil Kettle. En premier lieu on a extrait les données puis on les a stocké dans un fichier csv, après on a connecté Kettle à la base de données MySQL.

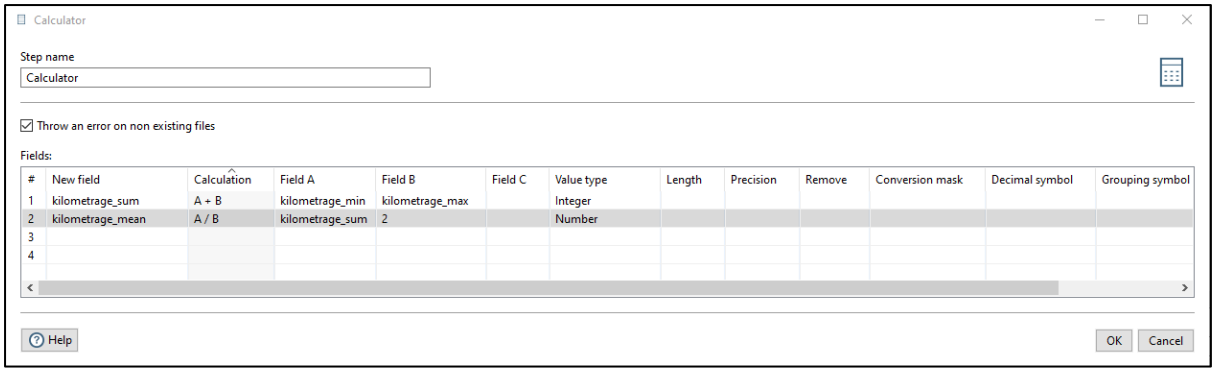


Figure 10: Illustration de la phase de préparation des données

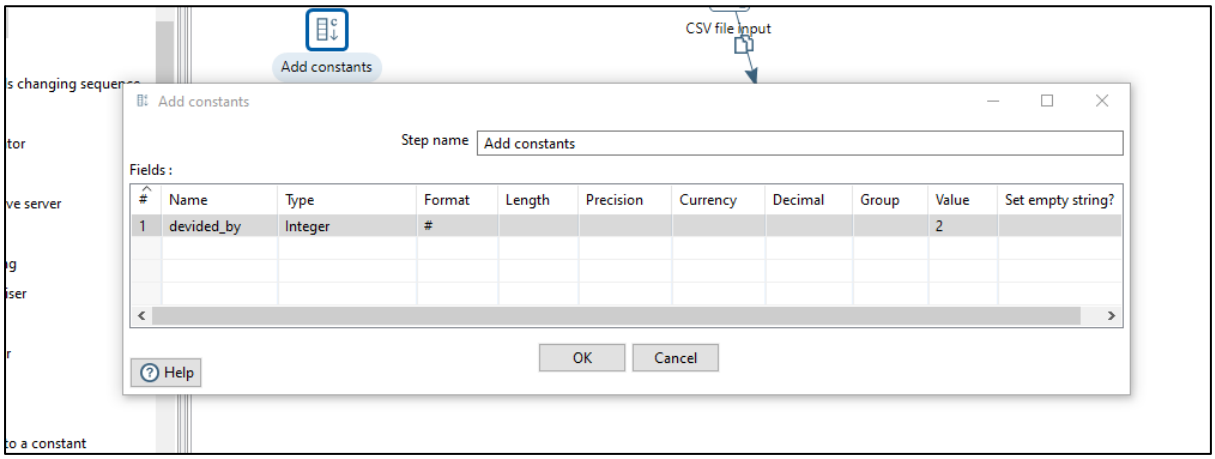


Figure 11: Illustration de l'ajout d'une colonne qui va nous aider dans le nettoyage des données

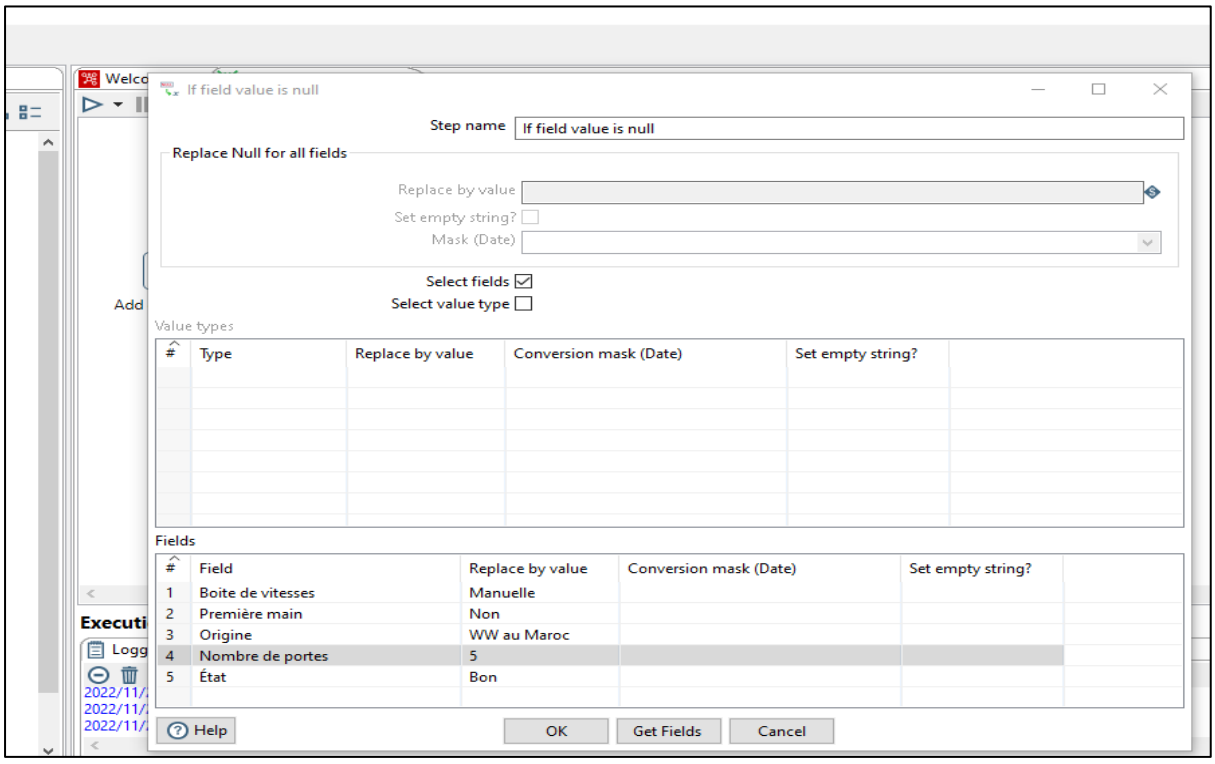


Figure 12: Remplacement des valeurs nulles par les valeurs les plus fréquentes dans notre dataset

En deuxième lieu et après la connexion avec la base de données, on a essayé de nettoyer le maximum les données, en remplaçant les valeurs nulles par des valeurs fréquentes dans notre dataset, et normalisant les données (remplacer un intervalle par sa valeur moyenne).

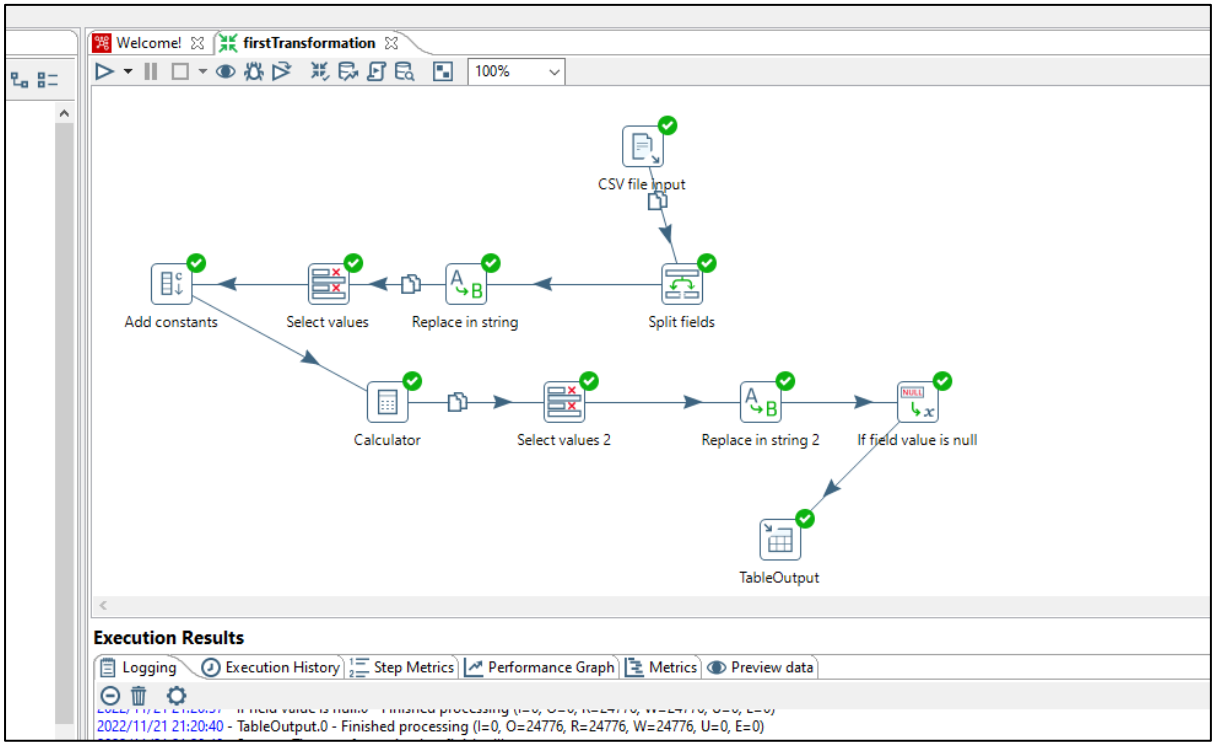


Figure 13: Schéma représentant l'ensemble des traitements ETL exécutés sur notre dataset

En troisième lieu, on a chargé les données dans notre datawarehouse qui est une base de données MySQL.

II-3 Python (Pandas)

Python est le langage de programmation open source le plus employé par les informaticiens. Ce langage s'est propulsé en tête de la gestion d'infrastructure, d'analyse de données ou dans le domaine du développement de logiciels. En effet, parmi ses qualités, Python permet notamment aux développeurs de se concentrer sur ce qu'ils font plutôt que sur la manière dont ils le font. Il a libéré les développeurs des contraintes de formes qui occupaient leur temps avec les langages plus anciens. Ainsi, développer du code avec Python est plus rapide qu'avec d'autres langages. [8]

Figure 14: Logo de Python

La bibliothèque logicielle open-source Pandas est spécifiquement conçue pour la manipulation et l'analyse de données en langage Python. Elle est à la fois performante, flexible et simple d'utilisation. Grâce à Pandas, le langage Python permet enfin de charger, d'aligner, de manipuler ou encore de fusionner des données. Cet outil permet aussi d'importer et d'exporter les données dans différents formats comme CSV ou JSON. Par ailleurs, Pandas offre aussi des fonctionnalités de Data Cleaning. [9]

Figure 15: Logo de la bibliothèque Pandas

```
import warnings # eliminer les erreurs de version
warnings.filterwarnings("ignore")
```

```
[ ] df=pd.read_csv("/content/drive/MyDrive/avito_car_dataset_ALL.csv", encoding='latin-1')
```

```
[ ] df.head()
```

	Unnamed: 0	Lien	Ville	Secteur	Marque	Modèle	Année-Modèle	Kilométrage	Type de carburant	Puissance fiscale	...	Caméra de recul	Vitres électriques	ABS	ESP	Régulateur de vitesse	Limiteur de vitesse	CD/MP3/Bluetooth	Ordinateur de bord
0	0	https://www.avito.ma/fr/massira_2/voitures/FIA...	Temara	Massira 2	Fiat	Punto	2007	200 000 - 249 999	Diesel	5	...	False	False	True	False	False	False	True	False
1	1	https://www.avito.ma/fr/temara/voitures/Dacia_...	Temara	NaN	Dacia	Dokker Van	2013	400 000 - 449 999	Diesel	6	...	False	False	False	False	False	False	False	False
2	2	https://www.avito.ma/fr/casablanca/voitures/Da...	Casablanca	NaN	Dacia	Dokker	2014	160 000 - 169 999	Diesel	6	...	False	False	False	False	False	False	False	False
3	3	https://www.avito.ma/fr/casablanca/voitures/to...	Casablanca	NaN	Volkswagen	Touareg	2005	0 - 4 999	Diesel	10	...	False	False	False	False	False	False	False	False
4	4	https://www.avito.ma/fr/dakhla/voitures/Toyota...	Dakhla	NaN	Toyota	Prado	2007	200 000 - 249 999	Diesel	12	...	False	False	True	False	False	False	True	False

5 rows × 32 columns

Figure 16: Récupération des données du fichier csv

```
# élimination des marque rare que nous n'avons pas assez d'informations sur
df.drop(df[ (df['Marque'] == "Suzuki") | (df['Marque'] == "mini")
          | (df['Marque'] == "Alfa Romeo") | (df['Marque'] == "Chevrolet")
          | (df['Marque'] == "Jeep") ].index, inplace=True)

# affichage des 5 premières ligne
df.head()
```

	Unnamed: 0	Lien	Ville	Secteur	Marque	Modèle	Année-Modèle	Kilométrage	Type de carburant	Puissance fiscale	...	Caméra de recul	Vitres électriques	ABS	ESP	Régulateur de vitesse	Limiteur de vitesse	CD/MP3/Bluetooth	Ordinateur de bord	Verro centr d
0	0	https://www.avito.ma/fr/massira_2/voitures/FIA...	Temara	Massira 2	Fiat	Punto	2007	200 000 - 249 999	Diesel	5	...	False	False	True	False	False	False	True	False	
1	1	https://www.avito.ma/fr/temara/voitures/Dacia_...	Temara	NaN	Dacia	Dokker Van	2013	400 000 - 449 999	Diesel	6	...	False	False	False	False	False	False	False	False	
2	2	https://www.avito.ma/fr/casablanca/voitures/Da...	Casablanca	NaN	Dacia	Dokker	2014	160 000 - 169 999	Diesel	6	...	False	False	False	False	False	False	False	False	
3	3	https://www.avito.ma/fr/casablanca/voitures/to...	Casablanca	NaN	Volkswagen	Touareg	2005	0 - 4 999	Diesel	10	...	False	False	False	False	False	False	False	False	
4	4	https://www.avito.ma/fr/dakhla/voitures/Toyota...	Dakhla	NaN	Toyota	Prado	2007	200 000 - 249 999	Diesel	12	...	False	False	True	False	False	False	True	False	

5 rows × 32 columns

Figure 17: Résultat après l'élimination des marques rares ayant un manque d'informations

Après la récupération des données de notre datawarehouse, on a commencé le nettoyage en éliminant les marques rares ayant un manque d’informations.

```
df.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 24178 entries, 0 to 24747
Data columns (total 32 columns):
 #   Column                             Non-Null Count  Dtype
---  ------                             --------------  -----
 0   Unnamed: 0                         24178 non-null  int64
 1   Lien                               24178 non-null  object
 2   Ville                              24178 non-null  object
 3   Secteur                            10474 non-null  object
 4   Marque                             24178 non-null  object
 5   Modèle                             24178 non-null  object
 6   Année-Modèle                       24178 non-null  int64
 7   Kilométrage                        24178 non-null  object
 8   Type de carburant                  24178 non-null  object
 9   Puissance fiscale                  24178 non-null  int64
 10  Boite de vitesses                  24178 non-null  object
 11  Nombre de portes                   19912 non-null  float64
 12  Origine                            19045 non-null  object
 13  Première main                      18143 non-null  object
 14  État                               17668 non-null  object
 15  Jantes aluminium                   24178 non-null  bool
 16  Airbags                            24178 non-null  bool
 17  Climatisation                      24178 non-null  bool
```

Figure 18: Affichage de la liste et format des différentes variables

Ici on a affiché la liste et le format des différentes variables, c'est-à-dire on a affiché les champs et leurs types.

```
df.isna().sum()

Unnamed: 0                             0
Lien                                   0
Ville                                  0
Secteur                            13704
Marque                                 0
Modèle                                 0
Année-Modèle                           0
Kilométrage                            0
Type de carburant                      0
Puissance fiscale                      0
Boite de vitesses                      0
Nombre de portes                    4266
Origine                             5133
Première main                       6035
État                                6510
Jantes aluminium                       0
Airbags                                0
Climatisation                          0
Système de navigation/GPS              0
```

Figure 19: Nombre de valeurs nulles par colonne

Ici on a vérifié le nombre de valeurs nulles par colonne pour qu'on puisse éliminé les colonnes ayant un grand nombre de valeurs nulls.

```
[ ]  df['Boite de vitesses'].value_counts()

     Manuelle       17129
     Automatique     4224
     --              2825
     Name: Boite de vitesses, dtype: int64

[ ]  df['Boite de vitesses'].value_counts()

     Manuelle       17129
     Automatique     4224
     --              2825
     Name: Boite de vitesses, dtype: int64

[ ]  df['Origine'].value_counts()

     WW au Maroc             14233
     Dédouanée                3834
     Importée neuve            915
     Pas encore dédouanée       63
     Name: Origine, dtype: int64

[ ]  df['État'].value_counts()

     Excellent      8559
     Très bon       7192
     Bon            1818
     Correct          90
     Pour Pièces       6
     Endommagé         3
     Name: État, dtype: int64

[ ]  df['Première main'].value_counts()

     Non    11514
     Oui     6629
     Name: Première main, dtype: int64
```

Figure 20: Nombre de valeurs contenues dans chaque colonne

```
[ ]  df['Boite de vitesses'].replace(to_replace='--',value='Manuelle',inplace=True)


     df['Origine'].replace(to_replace='',value='WW au Maroc',inplace=True)
     df['Nombre de portes'].replace(to_replace='',value='5',inplace=True)
     df['État'].replace(to_replace='',value='Bon',inplace=True)
     df['Première main'].replace(to_replace='',value='Non',inplace=True)

     # convertir les booléens en 0 et 1
     df.replace(to_replace=True,value=1,inplace=True)
     df.replace(to_replace=False,value=0,inplace=True)
```

Figure 21: Conversion des booléens en 0 et 1

```
[ ]  df['Boite de vitesses'].replace(to_replace='--',value='Manuelle',inplace=True)


     df['Origine'].replace(to_replace='',value='WW au Maroc',inplace=True)
     df['Nombre de portes'].replace(to_replace='',value='5',inplace=True)
     df['État'].replace(to_replace='',value='Bon',inplace=True)
     df['Première main'].replace(to_replace='',value='Non',inplace=True)

     # convertir les booléens en 0 et 1
     df.replace(to_replace=True,value=1,inplace=True)
     df.replace(to_replace=False,value=0,inplace=True)
```

```
[ ]  df.drop(columns=["Unnamed: 0","Lien","Secteur"],inplace=True) #suppression

     #drop any row with NaN values
     df = df.dropna()

     #checking null value
     df.isna().sum()
```

```
Ville                          0
Marque                         0
Modèle                         0
Année-Modèle                   0
Kilométrage                    0
Type de carburant              0
Puissance fiscale              0
Boite de vitesses              0
Nombre de portes               0
Origine                        0
Première main                  0
État                           0
Jantes aluminium               0
Airbags                        0
Climatisation                  0
```

Figure 22: Suppression des valeurs nulles

```
Ville                          0
Marque                         0
Modèle                         0
Année-Modèle                   0
Kilométrage                    0
Type de carburant              0
Puissance fiscale              0
Boite de vitesses              0
Nombre de portes               0
Origine                        0
Première main                  0
État                           0
Jantes aluminium               0
Airbags                        0
Climatisation                  0
Système de navigation/GPS      0
Toit ouvrant                   0
Sièges cuir                    0
Radar de recul                 0
Caméra de recul                0
Vitres électriques             0
```

Figure 23: Vérification de l'existence d'autres valeurs nulles

Après avoir visualisé le nombre de valeurs contenues dans chaque colonne, on a converti les booléens en 0 et 1 et on a supprimé les valeurs nulles. Après on a vérifié l'existence d'autres valeurs nulles (on n'a pas pu trouver aucune valeur nulle, ce qui montre qu'on a bien nettoyé nos données)

```
df.head()
```

	Ville	Marque	Modèle	Année-Modèle	Kilométrage	Type de carburant	Puissance fiscale	Boite de vitesses	Nombre de portes	Origine	...	Caméra de recul	Vitres électriques	ABS	ESP	Régulateur de vitesse	Limiteur de vitesse	CD/MP3/Bluetooth	Ordinateur de bord	Verrouillage centralisé à distance	Prix
0	Temara	Fiat	Punto	2007	200 000 - 249 999	Diesel	5	Manuelle	5.0	WW au Maroc	...	0	0	1	0	0	0	1	0	0	60000
1	Temara	Dacia	Dokker Van	2013	400 000 - 449 999	Diesel	6	Manuelle	3.0	WW au Maroc	...	0	0	0	0	0	0	0	0	0	70000
3	Casablanca	Volkswagen	Touareg	2005	0 - 4 999	Diesel	10	Automatique	5.0	WW au Maroc	...	0	0	0	0	0	0	0	0	0	90000
4	Dakhla	Toyota	Prado	2007	200 000 - 249 999	Diesel	12	Manuelle	5.0	WW au Maroc	...	0	0	1	0	0	0	1	0	0	97000
5	Khouribga	Volkswagen	Tiguan	2014	180 000 - 189 999	Diesel	8	Automatique	5.0	Dédouanée	...	0	0	1	1	0	0	1	0	0	255000

5 rows × 29 columns

```
[ ] df.shape
```

```
(13992, 29)
```

Figure 24: Visualisation de la dataset

```
    df['Année-Modèle'] = df['Année-Modèle'].astype(int)
    df['Nombre de portes'] = df['Nombre de portes'].astype(int)
```

```
[ ] df['Kilométrage']
```

```
0        200 000 - 249 999
1        400 000 - 449 999
3                0 - 4 999
4        200 000 - 249 999
5        180 000 - 189 999
               ...
24740    200 000 - 249 999
24741    150 000 - 159 999
24742    300 000 - 349 999
24745    190 000 - 199 999
24747      40 000 - 44 999
Name: Kilométrage, Length: 13992, dtype: object
```

```
[ ] splited = df['Kilométrage'].str.split("-", n = 1, expand = True)
    splited[0] = splited[0].str.replace(' ','').astype(int)
    splited[1] = splited[1].str.replace(' ','').astype(int)
    df['Kilométrage'] = (splited[1] + splited[0])/2
```

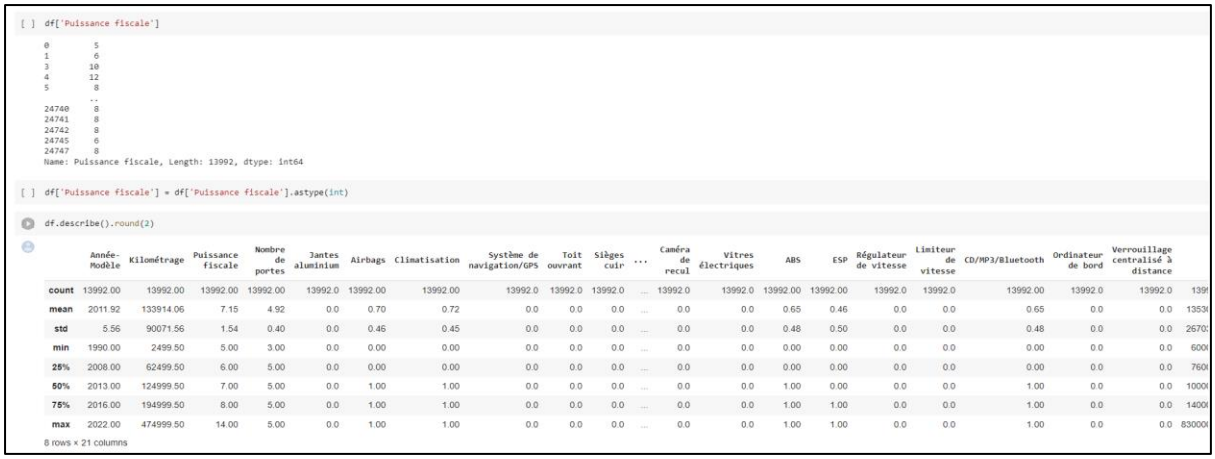
```
[ ] df.shape
```

```
(13992, 29)
```

```
[ ] # determining the name of the file
    file_name = 'Cars_transformed_DataSet.xlsx'

    # saving the excel
    df.to_excel(file_name)
    print('DataFrame is written to Excel File successfully.')
```

Figure 25: Suite de nettoyage des données



```
[ ] df['Puissance fiscale']

0        5
1        6
3       10
4       12
5        8
        ..
24740    8
24741    8
24742    8
24745    6
24747    8
Name: Puissance fiscale, Length: 13992, dtype: int64

[ ] df['Puissance fiscale'] = df['Puissance fiscale'].astype(int)

    df.describe().round(2)
```

	Année-Modèle	Kilométrage	Puissance fiscale	Nombre de portes	Jantes aluminium	Airbags	Climatisation	Système de navigation/GPS	Toit ouvrant	Sièges cuir	...	Caméra de recul	Vitres électriques	ABS	ESP	Régulateur de vitesse	Limiteur de vitesse	CD/MP3/Bluetooth	Ordinateur de bord	Verrouillage centralisé à distance	
count	13992.00	13992.00	13992.00	13992.00	13992.0	13992.00	13992.00	13992.0	13992.0	13992.0	...	13992.0	13992.0	13992.00	13992.00	13992.0	13992.0	13992.00	13992.0	13992.0	139
mean	2011.92	133914.06	7.15	4.92	0.0	0.70	0.72	0.0	0.0	0.0	...	0.0	0.0	0.65	0.46	0.0	0.0	0.65	0.0	0.0	1353
std	5.56	90071.56	1.54	0.40	0.0	0.46	0.45	0.0	0.0	0.0	...	0.0	0.0	0.48	0.50	0.0	0.0	0.48	0.0	0.0	2670
min	1990.00	2499.50	5.00	3.00	0.0	0.00	0.00	0.0	0.0	0.0	...	0.0	0.0	0.00	0.00	0.0	0.0	0.00	0.0	0.0	600
25%	2008.00	62499.50	6.00	5.00	0.0	0.00	0.00	0.0	0.0	0.0	...	0.0	0.0	0.00	0.00	0.0	0.0	0.00	0.0	0.0	760
50%	2013.00	124999.50	7.00	5.00	0.0	1.00	1.00	0.0	0.0	0.0	...	0.0	0.0	1.00	0.00	0.0	0.0	1.00	0.0	0.0	1000
75%	2016.00	194999.50	8.00	5.00	0.0	1.00	1.00	0.0	0.0	0.0	...	0.0	0.0	1.00	1.00	0.0	0.0	1.00	0.0	0.0	1400
max	2022.00	474999.50	14.00	5.00	0.0	1.00	1.00	0.0	0.0	0.0	...	0.0	0.0	1.00	1.00	0.0	0.0	1.00	0.0	0.0	83000

8 rows × 21 columns

Figure 26: Description des données présentes dans notre dataset

II-4 Synthèse

DATAWAREHOUSE

Un DW est une collection de données orientées sujet, intégrées, non volatiles, historisées, organisées pour la prise de décision. Ses objectifs sont :

• Regrouper, organiser des données provenant de sources diverses.

• Intégrer et stocker les données pour donner à l'utilisateur une vue orientée métier.

• Retrouver et analyser l'information selon plusieurs critères.

• Transformer un système d'information qui avait une vocation de production en un SI décisionnel.

• Doit contenir des données cohérentes

• Les données doivent pouvoir être séparées et combinées au moyen de toutes les mesures possibles de l'activité

• Le DW ne contient pas uniquement des données, mais aussi un ensemble d'outils de requêtes, d'analyse et de présentation de l'information. [4]

Pour notre projet on a utilisé MySQL comme DataWarehouse.

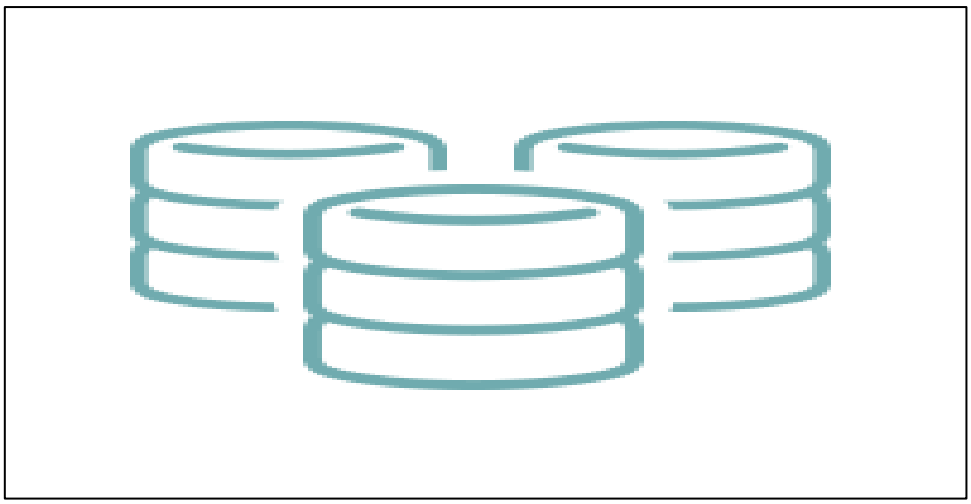


Figure 27: Illustration d'un datawarehouse

MySQL

MySQL est un système de gestion de bases de données relationnelles utilisant le langage de programmation SQL. Il propose une version open source qui permet à l'utilisateur d'accéder au code source et de le modifier, et une version entreprise permettant un accès aux dernières fonctionnalités du logiciel et au support fourni par Oracle, propriétaire et développeur actuel de MySQL. [5]

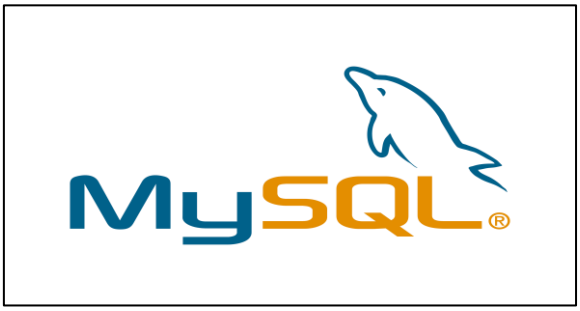
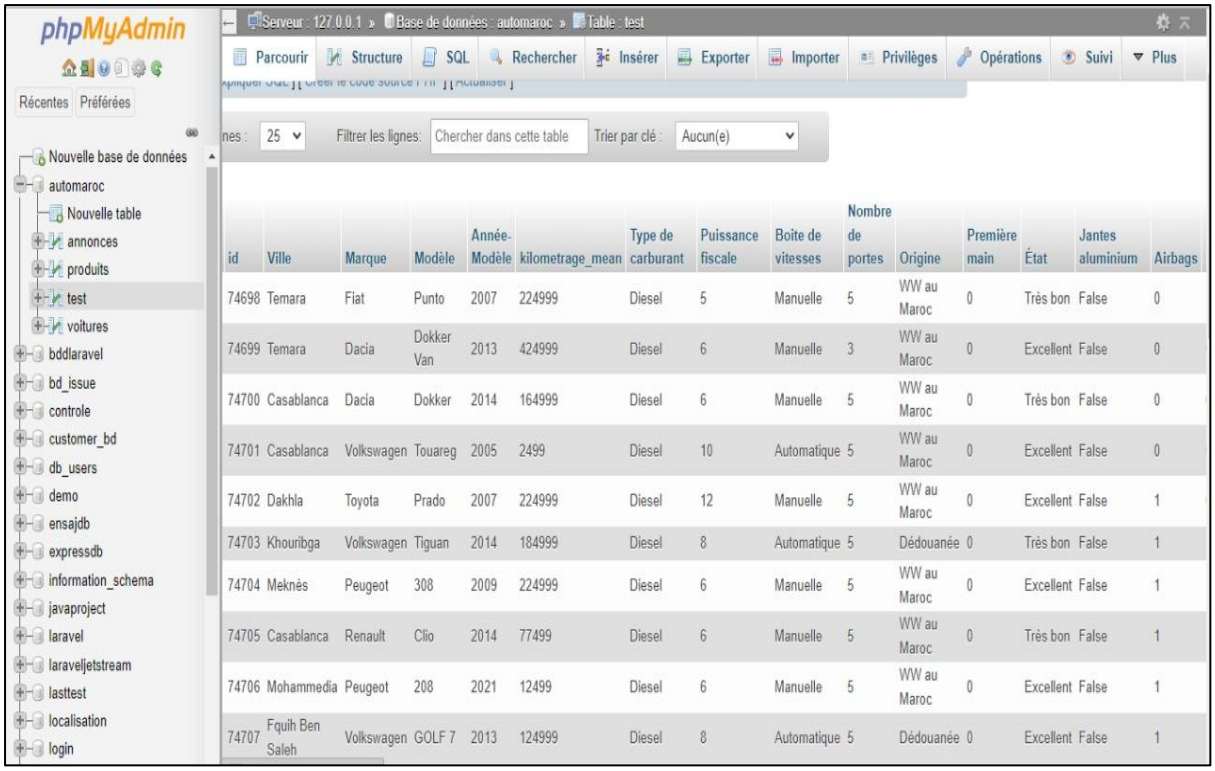


Figure 28: Logo de MySQL



id	Ville	Marque	Modèle	Année-Modèle	kilometrage_mean	Type de carburant	Puissance fiscale	Boite de vitesses	Nombre de portes	Origine	Première main	État	Jantes aluminium	Airbags
74698	Temara	Fiat	Punto	2007	224999	Diesel	5	Manuelle	5	WW au Maroc	0	Très bon	False	0
74699	Temara	Dacia	Dokker Van	2013	424999	Diesel	6	Manuelle	3	WW au Maroc	0	Excellent	False	0
74700	Casablanca	Dacia	Dokker	2014	164999	Diesel	6	Manuelle	5	WW au Maroc	0	Très bon	False	0
74701	Casablanca	Volkswagen	Touareg	2005	2499	Diesel	10	Automatique	5	WW au Maroc	0	Excellent	False	0
74702	Dakhla	Toyota	Prado	2007	224999	Diesel	12	Manuelle	5	WW au Maroc	0	Excellent	False	1
74703	Khouribga	Volkswagen	Tiguan	2014	184999	Diesel	8	Automatique	5	Dédouanée	0	Très bon	False	1
74704	Meknès	Peugeot	308	2009	224999	Diesel	6	Manuelle	5	WW au Maroc	0	Excellent	False	1
74705	Casablanca	Renault	Clio	2014	77499	Diesel	6	Manuelle	5	WW au Maroc	0	Très bon	False	1
74706	Mohammedia	Peugeot	208	2021	12499	Diesel	6	Manuelle	5	WW au Maroc	0	Excellent	False	1
74707	Fquih Ben Saleh	Volkswagen	GOLF 7	2013	124999	Diesel	8	Automatique	5	Dédouanée	0	Excellent	False	1

Figure 29: Données chargées dans notre datawarehouse

Python ETL vs outils ETL

La stratégie d'ETL doit être soigneusement choisie lors de la conception d'une stratégie d'entreposage de données. Une fois que vous avez choisi un processus ETL, vous êtes quelque peu enfermé, car il faudrait une énorme dépense d'heures de développement pour migrer vers une autre plate-forme. Cela est particulièrement vrai pour les entrepôts de données d'entreprise avec de nombreux schémas et architectures complexes.

Dans notre projet on a pu faire l'ETL avec l'outil KETTLE et avec Python, et on a arrivé à les mêmes résultats. Comparons donc l'utilité des outils Python ETL et ETL personnalisés pour éclairer ce choix.

- ✓ Coût

Le coût de la licence des outils ETL (en particulier pour les entrepôts de données des grandes entreprises) peut être élevé, mais cette dépense peut être compensée par le temps que cela permet à vos ingénieurs de gagner du temps pour travailler sur d'autres choses. Les petites entreprises ou les startups peuvent ne pas toujours être en mesure de payer le coût des licences

des plates-formes ETL. Dans un tel scénario, la création d'un ETL Python personnalisé peut être une bonne option. Mais il est également important de déterminer si ces économies de coûts valent le retard qu'elles entraîneraient dans la mise sur le marché de votre produit. Une autre considération pour les startups est que les plates-formes avec des prix plus flexibles comme Avik Cloud maintiennent le coût proportionnel à l'utilisation, ce qui le rendrait beaucoup plus abordable pour les startups en démarrage avec des besoins ETL limités.

- ✓ Taille et complexité de l'entrepôt de données

S'il s'agit d'un entrepôt de données volumineux avec un schéma complexe, l'écriture d'un processus ETL Python personnalisé à partir de zéro peut être difficile, en particulier lorsque le schéma change plus fréquemment. Dans ce cas, vous devez explorer les options de divers outils ETL qui correspondent à vos besoins et à votre budget.

- ✓ Simplicité et flexibilité

Si l'entrepôt de données est petit, vous n'aurez peut-être pas besoin de toutes les fonctionnalités des outils ETL d'entreprise. Il peut être judicieux d'écrire un processus ETL Python léger et personnalisé, car il sera à la fois simple et vous offrira une meilleure flexibilité pour le personnaliser en fonction de vos besoins.

- ✓ Évolutivité

La taille initiale de la base de données peut ne pas être importante. Mais si vous prévoyez une croissance dans un avenir proche, vous devez déterminer si votre pipeline ETL Python personnalisé pourra également évoluer avec une augmentation du débit de données. En cas de doute, vous voudrez peut-être examiner de plus près certains des outils ETL, car ils évolueront plus facilement.

- ✓ Convivialité

Pour utiliser Python pour votre processus ETL, comme vous pouvez le deviner, cela nécessite une expertise en Python. Mais les outils ETL ont généralement des interfaces graphiques conviviales qui facilitent leur utilisation, même pour une personne non technique. Encore une fois, c'est un choix à faire selon les exigences du projet.

- ✓ Ajout de valeur et assistance

Les outils ETL, en particulier ceux payants, offrent plus de valeur ajoutée en termes de fonctionnalités et de compatibilités multiples. Ils offrent également un support client, ce qui semble être une considération sans importance jusqu'à ce que vous en ayez besoin. Cependant, les outils open source disposent d'une bonne documentation et de nombreuses communautés en ligne qui peuvent également offrir une assistance.

.

III- ANALYSE ET RESTITUTION

III-1 Définition

La data visualisation (restitution) est une composante essentielle dans la chaine d'information décisionnelle. Elle s'inscrit pleinement dans la stratégie d'une entreprise autour de la mise à disposition et l'analyse de la donnée. Les entreprises ayant une ambition réfléchie et pérenne autour de la donnée plébiscitent les outils de data visualisation.

La phase d'analyse et restitution vise à mettre les données à la disposition des utilisateurs en prenant en compte leur profil et leur besoin métier. Dans cette phase les utilisateurs finaux vont analyser les informations qui leur sont fournies. Habituellement, les données sont modélisées par des représentations basées sur des requêtes pour construire des tableaux de bord ou des rapports via des outils d'analyse décisionnelle (Power BI, Tableau, Qlikview, etc.).

Les outils de data visualisation permettent de : Centraliser l'affichage de la donnée en un seul et même endroit (accessibilité), associer et croiser des données provenant de sources diverses, comprendre simplement et rapidement la donnée en lui conférant un 'sens', et créer des tableaux de bords personnalisés et les partager. Dans le but de :

- ✓ Mesurer les performances et identifier les tendances remarquables
- ✓ Aider les décisionnaires dans leur orientation stratégique et favoriser les innovations
- ✓ Optimiser les organisations et le chiffre d'affaires.

L'objectif de cette phase est d'assister au mieux l'utilisateur pour qu'il puisse analyser les informations mises à sa disposition et prendre des décisions. Cela passe notamment par le contrôle d'accès aux rapports, la prise en charge des requêtes et la visualisation des résultats [10].

III-2 Outils d'analyse et restitution

III-2-1 Google Data Studio

Google Data Studio est plutôt un outil de business intelligence à prendre en compte dans votre arsenal, à condition de bien identifier ce que vous pouvez en faire… et vos alternatives en la matière. Google Data Studio permet de transformer vos données en rapports esthétiques et surtout faciles à lire, à partager et à customiser. l se présente sous la forme d'une ou de

plusieurs pages vierges. Vous y ajoutez comme vous le souhaitez des tableaux, images, des textes, des graphiques, différents styles de mises en forme pour mettre en valeur vos données variées.[11]

Figure 30: Logo de Google Data Studio

RESULTATS:

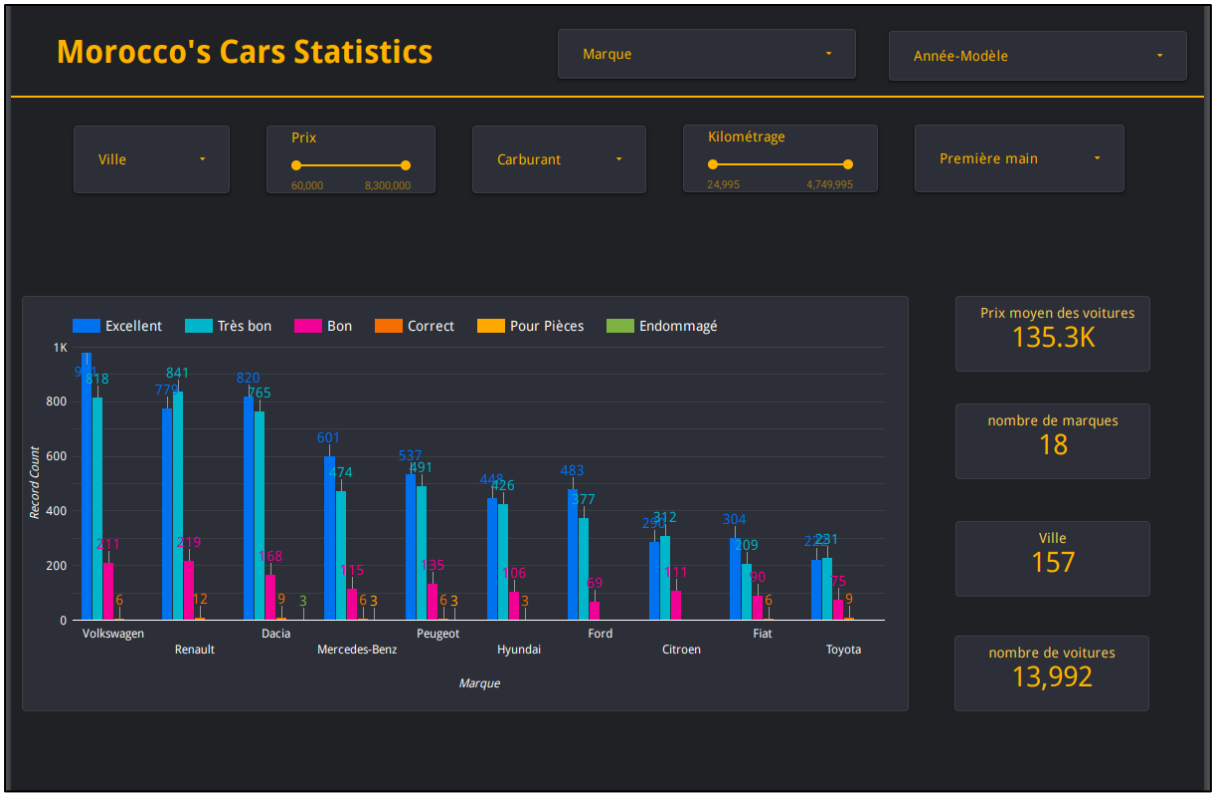


Figure 31: Premier tableau de bord

Figure 32: Deuxième tableau de bord

III-2-2 Python (Seaborn)

Seaborn est une bibliothèque permettant de créer des graphiques statistiques en Python. Elle est basée sur Matplotlib, et s'intègre avec les structures Pandas. Cette bibliothèque est aussi performante que Matplotlib, mais apporte une simplicité et des fonctionnalités inédites. Elle permet d'explorer et de comprendre rapidement les données. Des cadres de données complets peuvent être capturés, et les fonctions internes permettant la cartographie sémantique et l'agrégation statistique permettent de convertir les données en visualisations graphiques. Toute la complexité de Matplotlib est abstraite par Seaborn. Toutefois, il est possible de créer des graphiques répondant à tous vos besoins et vos exigences. [12]

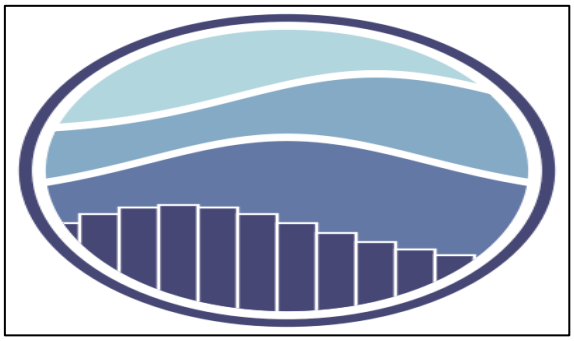


Figure 33: Logo de la bibliothèque Seaborn

RESULTATS:

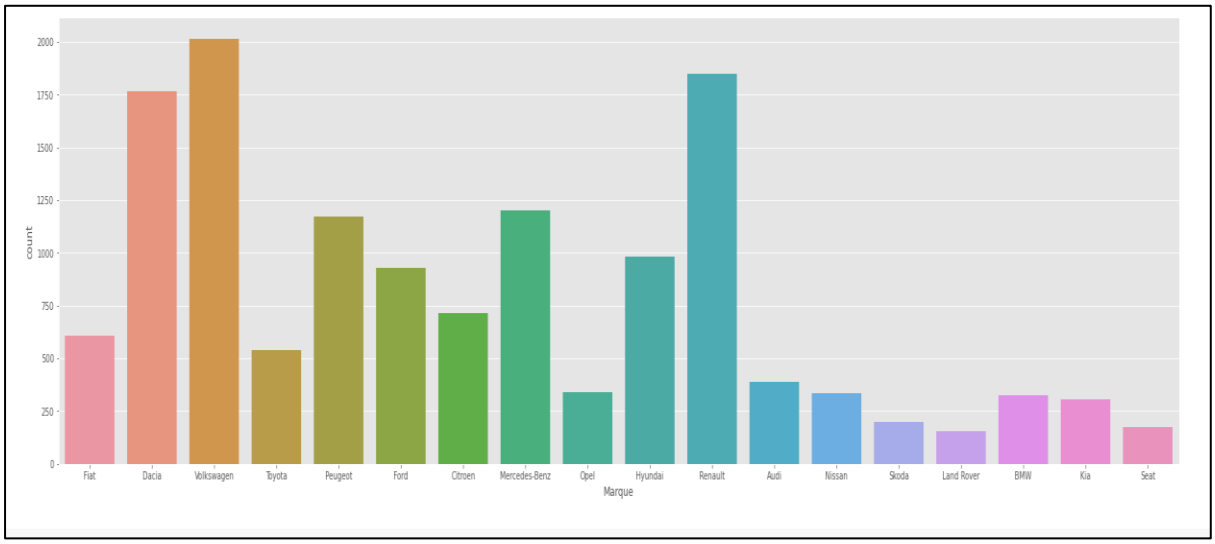


Figure 34: Diagramme en bâtons représentant le nombre des voitures par marque

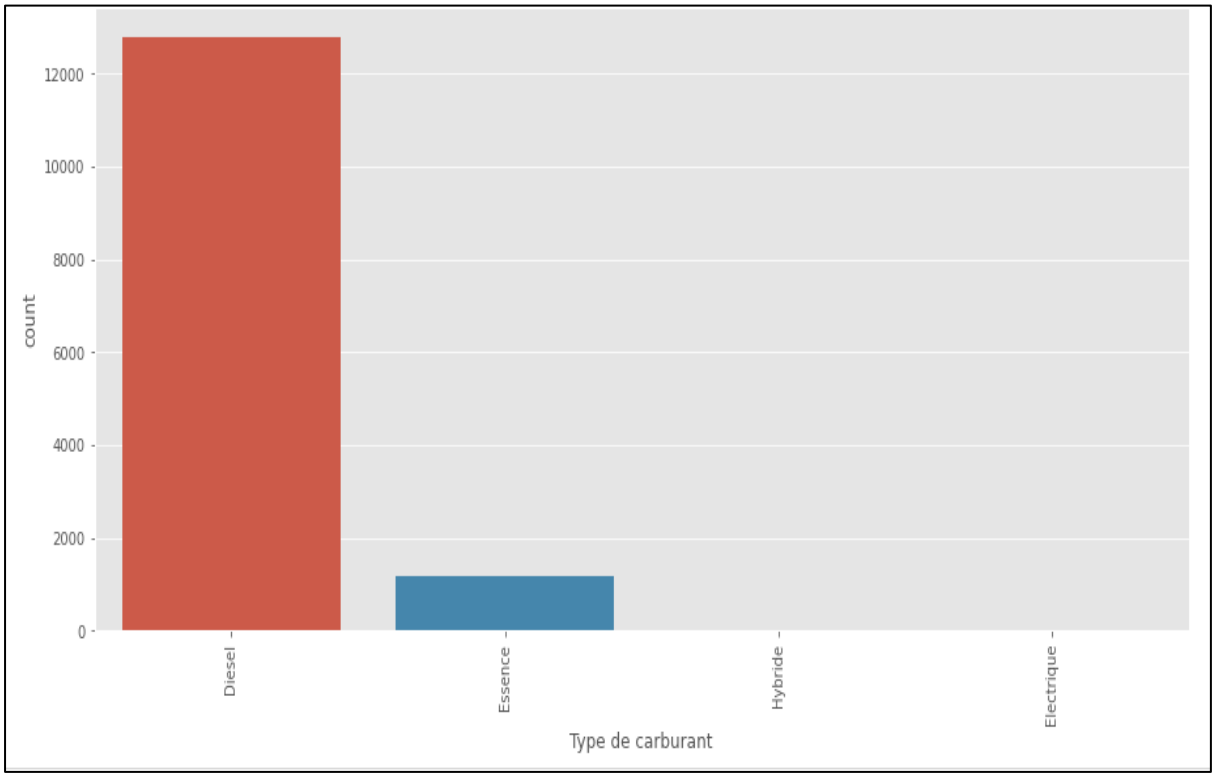


Figure 35: Diagramme en bâtons représentant le nombre de voitures par type de carburant

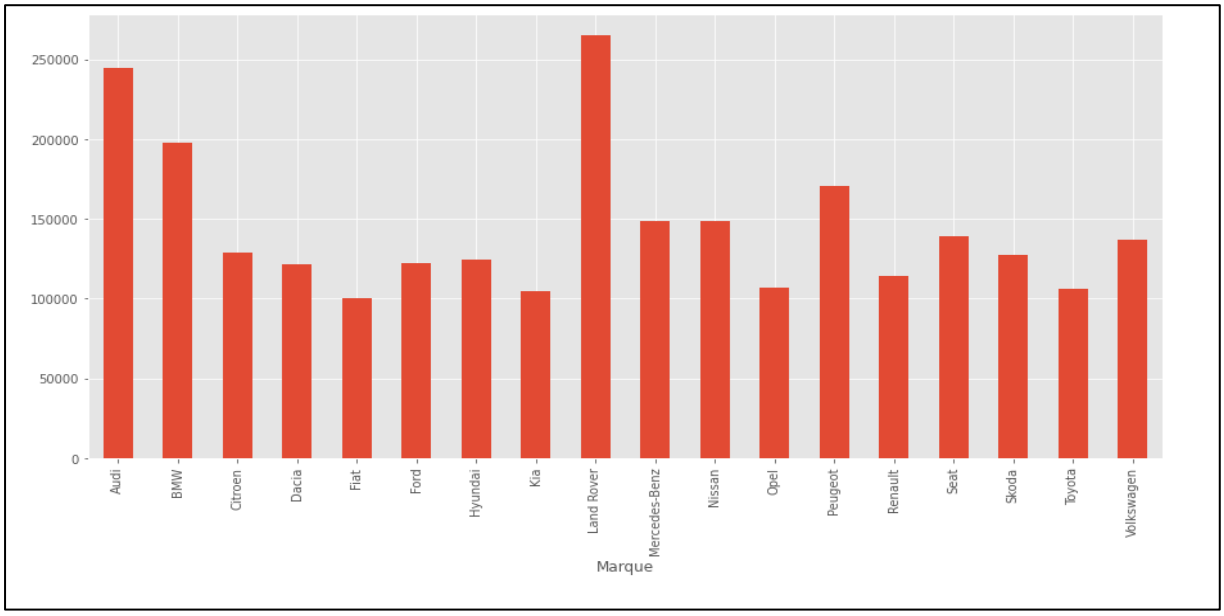


Figure 36: Diagramme en bâtons représentant le prix moyenne des voitures par marque

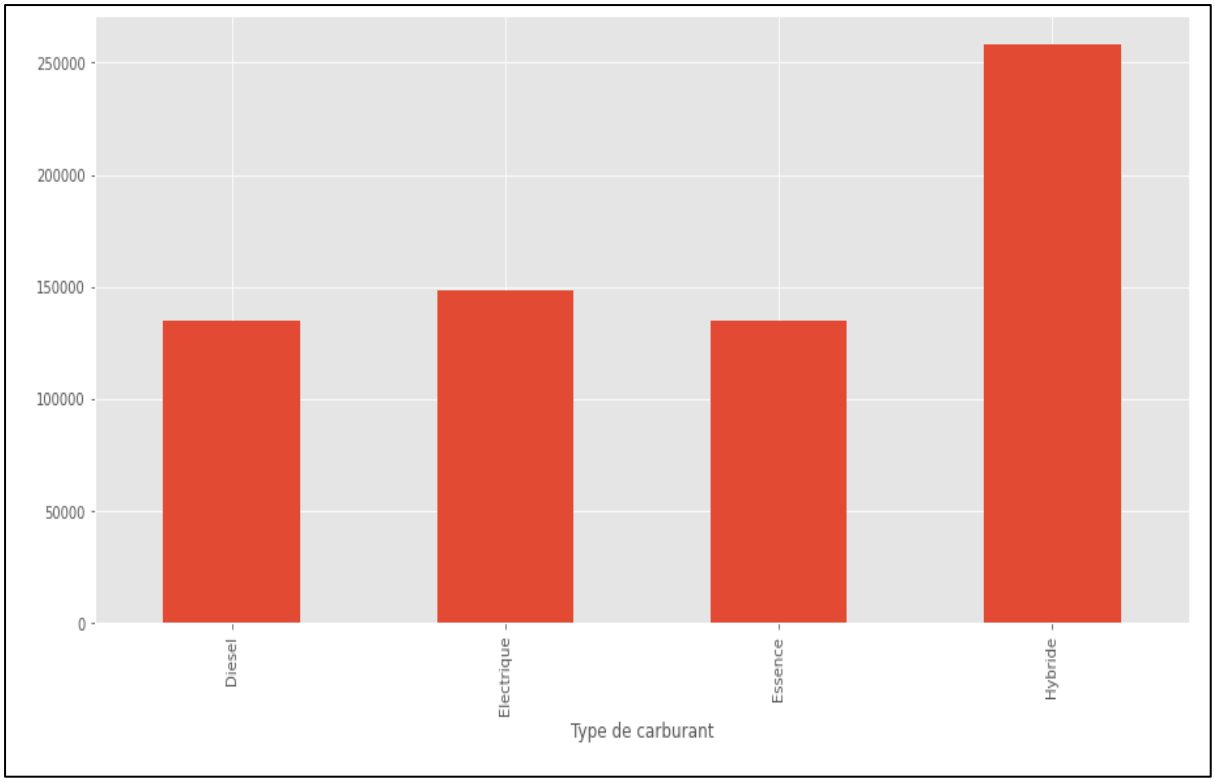


Figure 37: Diagramme en bâtons représentant le prix moyenne des voitures par type de carburant

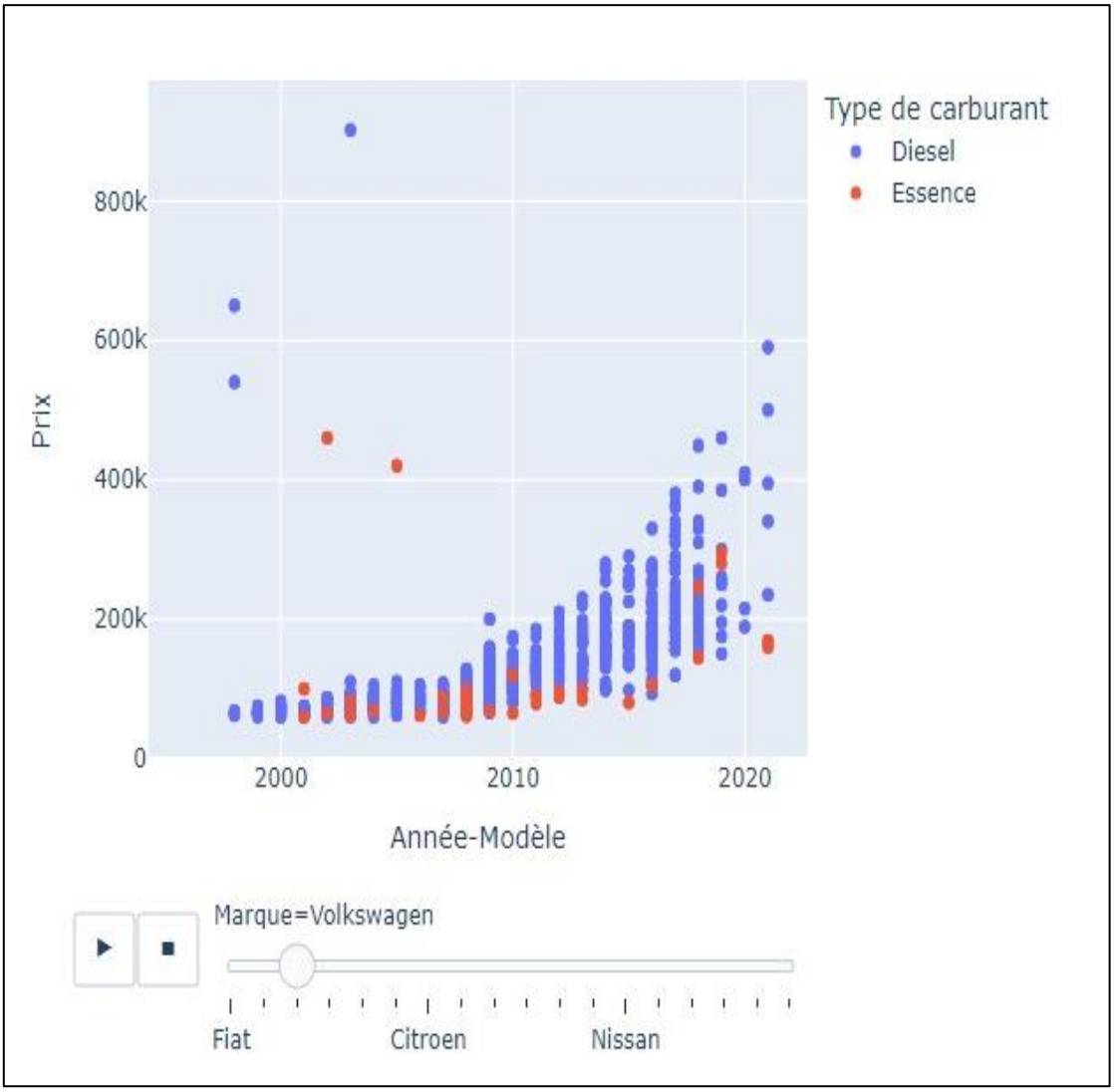


Figure 38: Scatter plot représentant le prix moyenne des voitures par année-modèle et type de carburant et marque

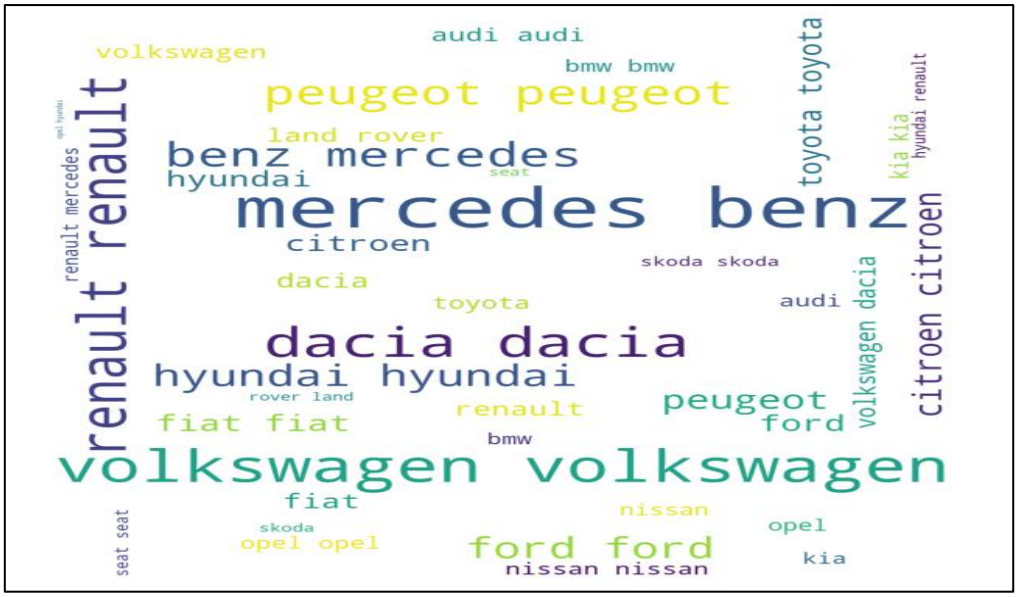


Figure 39: (Nuage de mots) Mots apparaissant souvent dans notre dataset

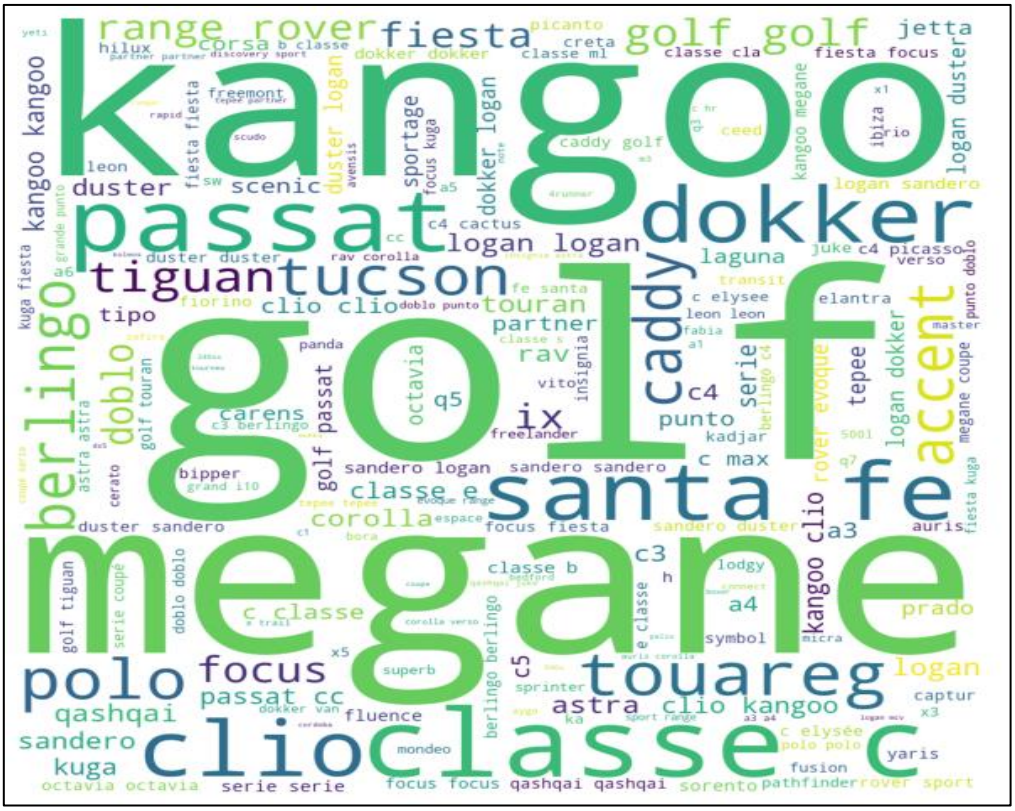


Figure 40: (Nuage de mots) Mots apparaissant souvent dans notre dataset

III-2-3 Power BI

Power BI est une solution de Business Intelligence (BI) développée par Microsoft pour permettre aux entreprises de consolider, d'analyser, de visualiser et diffuser leurs données. C'est un outil de data visualisation permettant de créer des tableaux de bord et de les diffuser au sein d'une entreprise. L'utilisation d'un tel outil est primordiale afin que les données internes et externes de l'entreprise puissent être comprises et analysées. Analyser des données dans une grille est difficile, mettre en forme pour mettre en valeur certaines informations permet en quelques secondes d'obtenir la donnée importante. [13]

Figure 41: Logo de Power BI

RESULTATS:

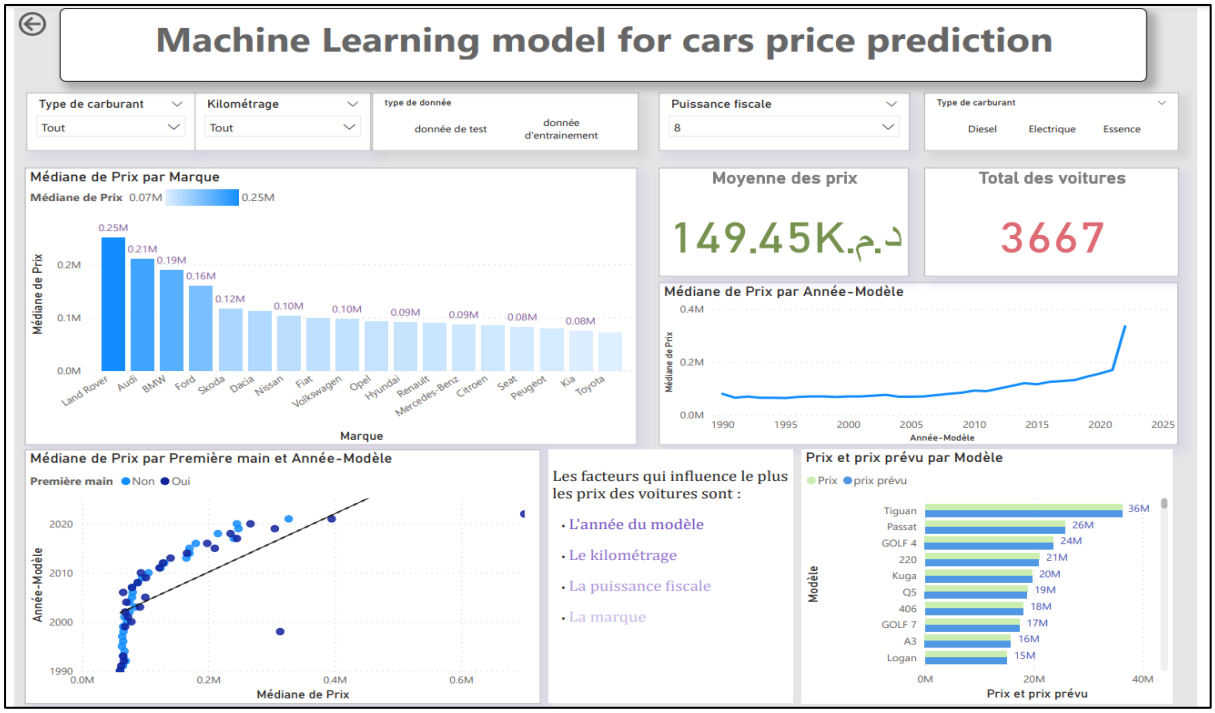


Figure 42: Tableau de bord

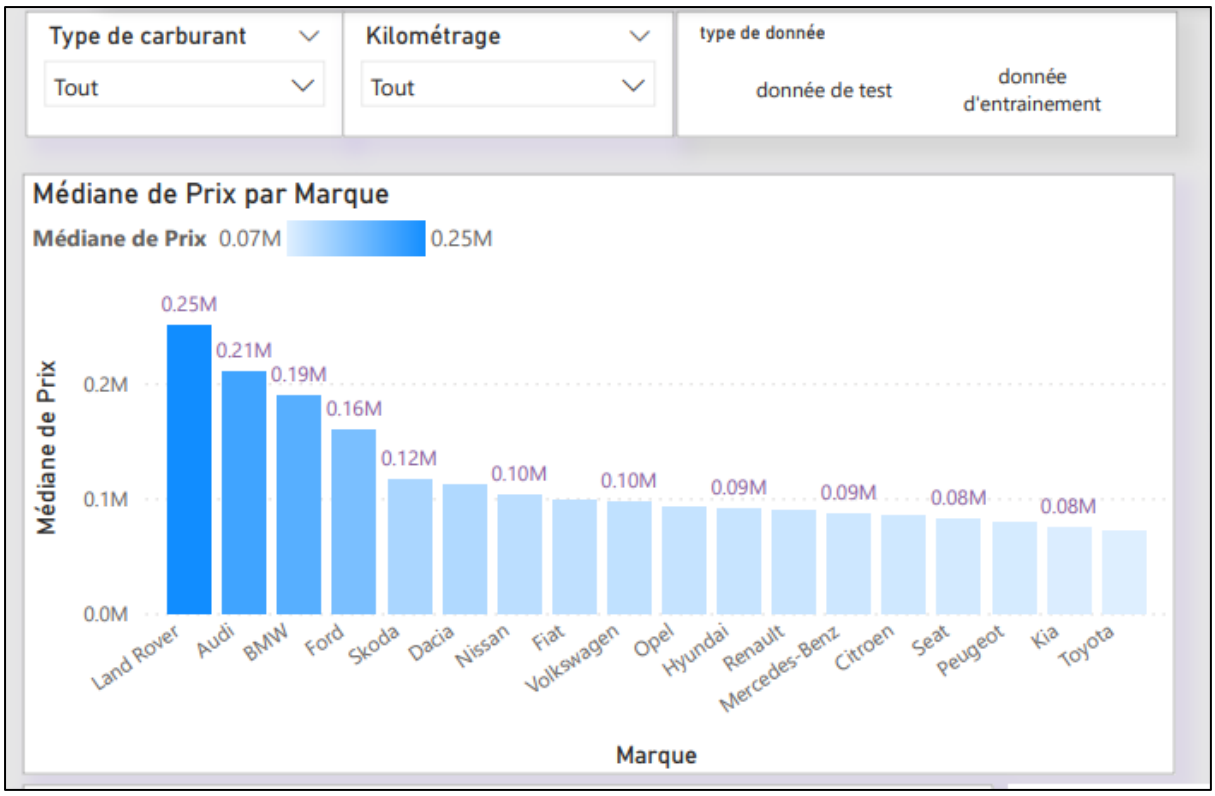


Figure 43: ZOOM 1 sur le tableau de bord

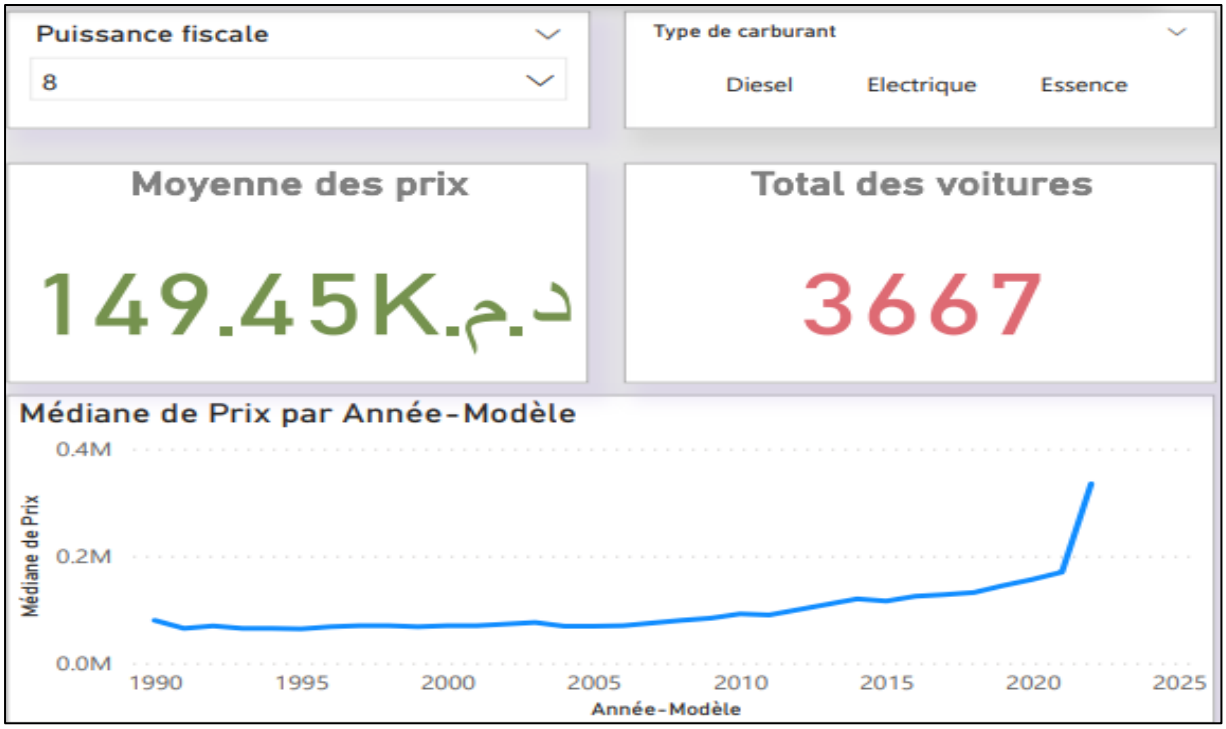


Figure 44: ZOOM 2 sur le tableau de bord

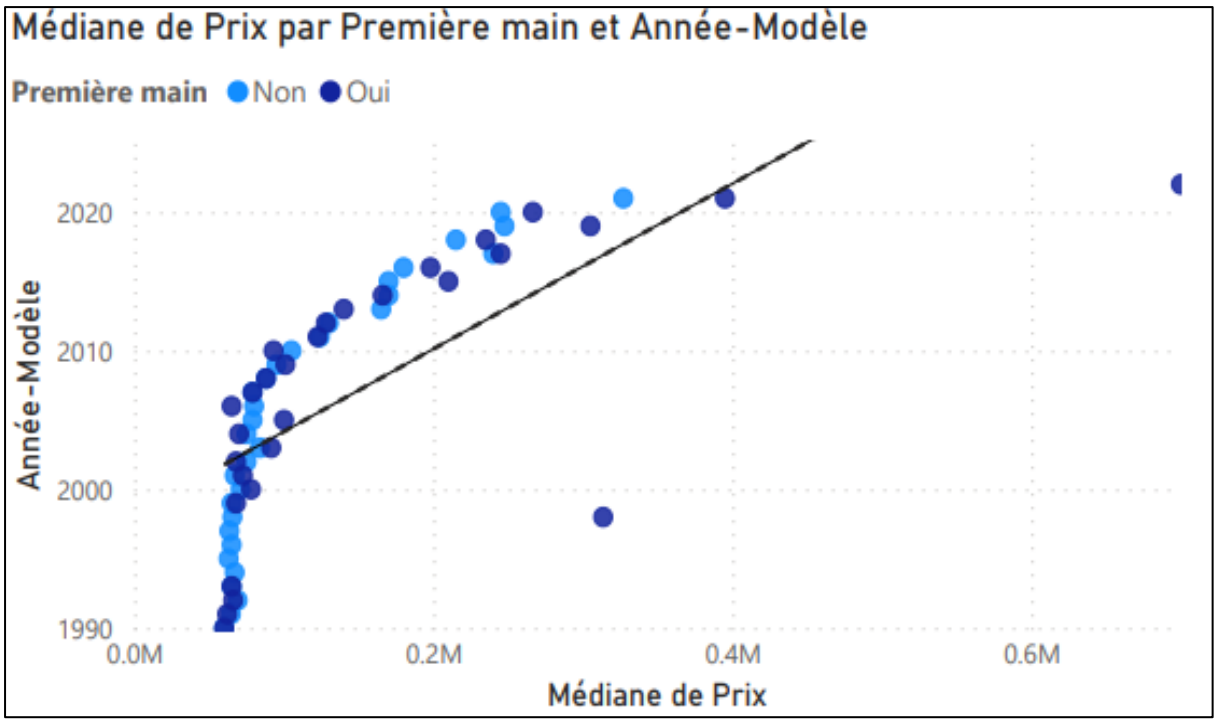


Figure 45: ZOOM 3 sur le tableau de bord

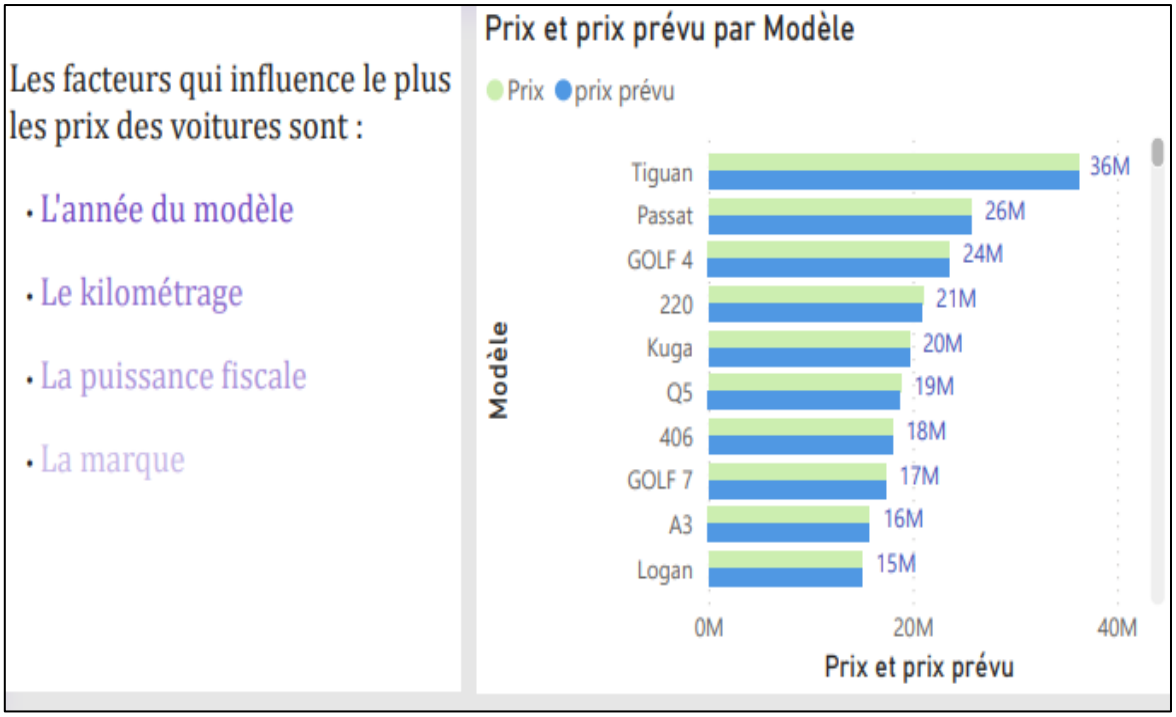
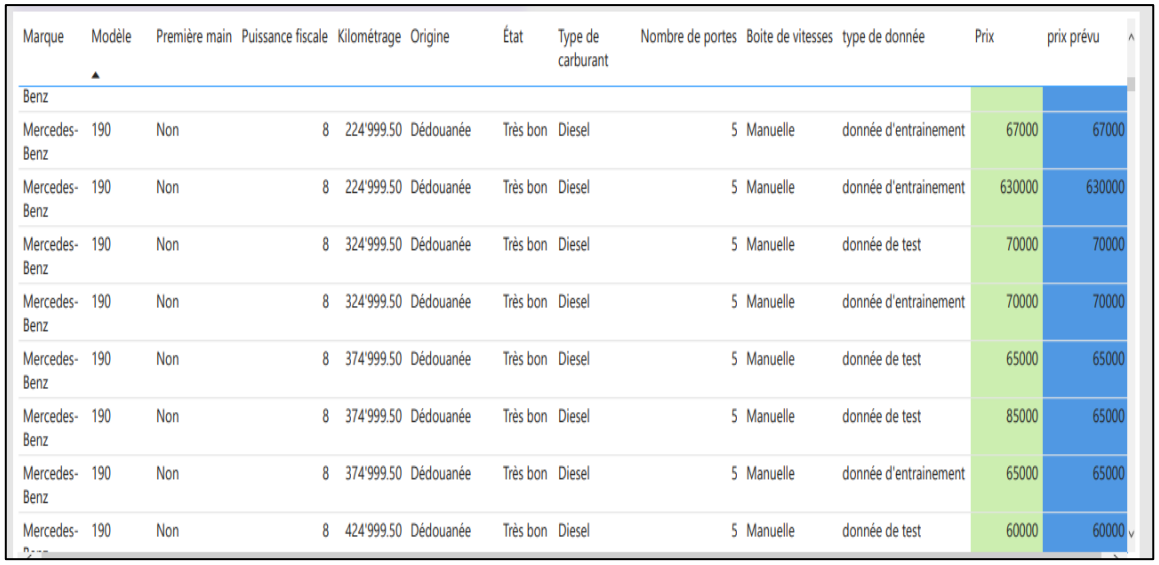


Figure 46: ZOOM 4 sur le tableau de bord



Marque	Modèle	Première main	Puissance fiscale	Kilométrage	Origine	État	Type de carburant	Nombre de portes	Boite de vitesses	type de donnée	Prix	prix prévu
Benz												
Mercedes-Benz	190	Non	8	224'999.50	Dédouanée	Très bon	Diesel	5	Manuelle	donnée d'entrainement	67000	67000
Mercedes-Benz	190	Non	8	224'999.50	Dédouanée	Très bon	Diesel	5	Manuelle	donnée d'entrainement	630000	630000
Mercedes-Benz	190	Non	8	324'999.50	Dédouanée	Très bon	Diesel	5	Manuelle	donnée de test	70000	70000
Mercedes-Benz	190	Non	8	324'999.50	Dédouanée	Très bon	Diesel	5	Manuelle	donnée d'entrainement	70000	70000
Mercedes-Benz	190	Non	8	374'999.50	Dédouanée	Très bon	Diesel	5	Manuelle	donnée de test	65000	65000
Mercedes-Benz	190	Non	8	374'999.50	Dédouanée	Très bon	Diesel	5	Manuelle	donnée de test	85000	65000
Mercedes-Benz	190	Non	8	374'999.50	Dédouanée	Très bon	Diesel	5	Manuelle	donnée d'entrainement	65000	65000
Mercedes-	190	Non	8	424'999.50	Dédouanée	Très bon	Diesel	5	Manuelle	donnée de test	60000	60000

Figure 47: ZOOM 5 sur le tableau de bord

III-3 Synthèse

Google Data Studio	Python	Power BI
- Une connexion multi-plateformes - Partage et travail collaboratif en temps réel -Personnalisation graphique poussée des rapports de performance (data visualization) -Filtrage des données intégré aux rapports de performance	-Agilité accrue -Flexibilité moyenne -Cout réduit -Des résultats plus rapides -Innovation améliorée	-Visualisation générale et/ou très détaillée en mode focus des données -Accès aux rapports et tableaux de bord via le web et tout appareil mobile Interaction sécurisée avec les tableaux de bord et les rapports -Partage rapide de rapports et tableaux de bord Ajout de commentaires directement dans un tableau de bord

CONCLUSION

Après avoir scrapé une grande quantité de données des deux sites d'automobiles avito.ma et moteur. C'est le moment de les bien exploiter à l'aide des solutions de Business intelligence. Ce rapport a illustré toutes les étapes qu'on a suivie et nos résultats. En se débutant par la préparation des données, et en se terminant par leur exploitation. Comme il a exhibé aussi les définitions de plusieurs technologies utilisées en informatique décisionnelle. Dans le premier chapitre on a présenté notre dataset et son origine. Dans le second chapitre, on a défini la phase d'ETL ainsi que les technologies utilisées dans cette phase, et on a exposé comment on a fait notre ETL, et le datawarehouse dans laquelle on a chargé les données. Enfin dans le troisième chapitre, on a défini la phase de l'analyse et restitution ainsi que les technologies utilisées dans cette phase, et on a exhibé les résultats à lesquelles on est arrivé et les tableaux de bord.

BIBLIOGRAPHIE

[1] https://actualiteinformatique.fr/data/definition-data-set

[2] https://www.ionos.fr/digitalguide/sites-internet/developpement-web/quest-ce-quele-web-scraping

[3] https://www.moteur.ma/fr/a-propos

[4] Support du cours

[5] https://datascientest.com/mysql-tout-comprendre

[6] https://www.lemagit.fr/definition/ETL-et-ELT

[7] https://www.next-decision.fr/editeurs-bi/etl/pentaho-pdi

[8] https://www.journaldunet.fr/web-tech/dictionnaire-du-webmastering/1445304-python-definition-et-utilisation-de-ce-langage-informatique/

[9] https://datascientest.com/pandas-python-data-science

[10] https://www.headmind.com/fr/informatique-decisionnelle/

[11] https://alphalyr.fr/google-data-studio-2/

[12] https://datascientest.com/seaborn

[13] https://datascientest.com/power-bi