Université de Lille
1 SCIENCES ET TECHNOLOGIES

Licence 3 de Statistique

Option :

Mathématiques Statistiques et informatique décisionnelle

PROJET sous SAS

Régression linéaire multiple : Les miles par gallon d'une voiture

Travail réalisé par :
HADAD AHMED ALI
2022-2023

Table des matières

Chapitre 1

Modélisation du fichier prédictif

1.1 Introduction

Quant on achète une voiture parfois on compte l'efficacité et l'économie du carburant de la voiture. Les **miles par gallon (mpg)** d'une voiture mesurent jusqu'où une voiture peut aller avec un gallon de carburant. Dans ce projet, en tant que statisticien, je suis chargé modéliser une régression linéaire multiple permettant de **prédire les miles par gallon** d'une voiture. Durant ce travail, on essaye de construire un fichier prédictif. On crée successivement des fichier jusqu'à trouver le fichier qui nous donne un meilleur modèle. Nous n'allons pas visualiser les fichier créées, cependant pour visualiser les observation et les variable d'un fichier ou autre chose, veuillez exécuter un à un le code ci-joint (sous SAS). Notez que les photos que vous verez tout au long du rapport sont des captures d'ecran prise dans l'exécution du code en SAS.

1.2 Données à analyser

Nous disposons les données de spécifications techniques des voitures publiées à l'origine en 1983 pour **l'American Statistical Association Data Expo**, que vous pouvez le visualiser en exécutant le code. C'est un multiple des donnée de différents fichier , le but est de déterminer une régression linéaire qui selon un certains nombre de ces variables données, va me permettre d'estimer le **mile par gallon** d'une voiture. Il s'agit donc de chercher les variables qui me sera utiles, dites **significatives à la variable mpg** parmis toutes ces variables afin de construire une **équation linéaire** représentée comme suit :

$$mpg = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$$

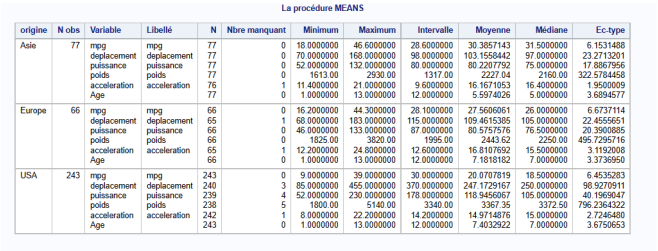
où α, β_1, β_2, β_3 sont des constantes réelles et $X_1, ..., X_n$ les variables **significative de mpg** qu'on les déterminera dans la suite.

1.3 Analyse descriptive des données

Pour mieux étudier le modèle, on est censé rassembler tout ces données dans un seul fichier. Plus simplement, avec des algorithmes ci-joint à la fin du rapport, on a crée un fichier qui est une concaténation sans modification de tout nos fichier. Ensuite, une création d'une variable Age (age de la voiture) qui remplace à son tour l'anné-du-model. On peut donc voir qu'on a **386 observations** sur **13 variables**.

1.3.1 Statistique descriptive

Voici les statistique descriptive de notre fichier :



La procédure MEANS

origine	N obs	Variable	Libellé	N	Nbre manquant	Minimum	Maximum	Intervalle	Moyenne	Médiane	Ec-type
Asie	77	mpg	mpg	77	0	18.0000000	46.6000000	28.6000000	30.3857143	31.5000000	6.1531488
		deplacement	deplacement	77	0	70.0000000	168.0000000	98.0000000	103.1558442	97.0000000	23.2713201
		puissance	puissance	77	0	52.0000000	132.0000000	80.0000000	80.2207792	75.0000000	17.8867956
		poids	poids	77	0	1613.00	2930.00	1317.00	2227.04	2160.00	322.5784458
		acceleration	acceleration	76	1	11.4000000	21.0000000	9.6000000	16.1671053	16.4000000	1.9500009
		Age		77	0	1.0000000	13.0000000	12.0000000	5.5974026	5.0000000	3.6894577
Europe	66	mpg	mpg	66	0	16.2000000	44.3000000	28.1000000	27.5606061	26.0000000	6.6737114
		deplacement	deplacement	65	1	68.0000000	183.0000000	115.0000000	109.4615385	105.0000000	22.4555651
		puissance	puissance	66	0	46.0000000	133.0000000	87.0000000	80.5757576	76.5000000	20.3900885
		poids	poids	66	0	1825.00	3820.00	1995.00	2443.62	2250.00	495.7295716
		acceleration	acceleration	65	1	12.2000000	24.8000000	12.6000000	16.8107692	15.5000000	3.1192008
		Age		66	0	1.0000000	13.0000000	12.0000000	7.1818182	7.0000000	3.3736950
USA	243	mpg	mpg	243	0	9.0000000	39.0000000	30.0000000	20.0707819	18.5000000	6.4535283
		deplacement	deplacement	240	3	85.0000000	455.0000000	370.0000000	247.1729167	250.0000000	98.9270911
		puissance	puissance	239	4	52.0000000	230.0000000	178.0000000	118.9456067	105.0000000	40.1969847
		poids	poids	238	5	1800.00	5140.00	3340.00	3367.35	3372.50	796.2364322
		acceleration	acceleration	242	1	8.0000000	22.2000000	14.2000000	14.9714876	15.0000000	2.7246480
		Age		243	0	1.0000000	13.0000000	12.0000000	7.4032922	7.0000000	3.6750653

On remarque qu'en **moyenne** les voitures Asiatique ont un **mpg** beaucoup plus que les voitures d'Europe et celles d'Europe ont un **mpg** beaucoup plus que celles d'USA. On peut meme voir que l'ordre est preservé pour le maximum d'**mpg**, c'est en Asie où on a le **mpg** le plus grand puis en Europe et en fin aux USA. Cependant, il ne s'agit donc pas le bon fichier prédictif, ce qui nous intéresse dans ce fichier c'est de remarquer qu'il y'a des valeurs manquantes. Nous avons ainsi remplacer ces valeurs manquantes par leurs médianes respective. Voici ce que présente le nouveau fichier :

La procédure MEANS

origine	N obs	Variable	Libellé	N	Nbre manquant	Minimum	Maximum	Intervalle	Moyenne	Médiane	Ec-type
Asie	77	mpg	mpg	77	0	18.0000000	46.6000000	28.6000000	30.3857143	31.5000000	6.1531488
		deplacement	deplacement	77	0	70.0000000	168.0000000	98.0000000	103.1558442	97.0000000	23.2713201
		puissance	puissance	77	0	52.0000000	132.0000000	80.0000000	80.2207792	75.0000000	17.8867956
		poids	poids	77	0	1613.00	2930.00	1317.00	2227.04	2160.00	322.5784458
		acceleration	acceleration	77	0	11.4000000	21.0000000	9.6000000	16.1701299	16.4000000	1.9373113
		Age		77	0	1.0000000	13.0000000	12.0000000	5.5974026	5.0000000	3.6894577
Europe	66	mpg	mpg	66	0	16.2000000	44.3000000	28.1000000	27.5606061	26.0000000	6.6737114
		deplacement	deplacement	66	0	68.0000000	183.0000000	115.0000000	109.3939394	105.0000000	22.2889271
		puissance	puissance	66	0	46.0000000	133.0000000	87.0000000	80.5757576	76.5000000	20.3900885
		poids	poids	66	0	1825.00	3820.00	1995.00	2443.62	2250.00	495.7295716
		acceleration	acceleration	66	0	12.2000000	24.8000000	12.6000000	16.7909091	15.5000000	3.0993164
		Age		66	0	1.0000000	13.0000000	12.0000000	7.1818182	7.0000000	3.3736950
USA	243	mpg	mpg	243	0	9.0000000	39.0000000	30.0000000	20.0707819	18.5000000	6.4535283
		deplacement	deplacement	243	0	85.0000000	455.0000000	370.0000000	247.2078189	250.0000000	98.3124921
		puissance	puissance	243	0	52.0000000	230.0000000	178.0000000	118.7160494	105.0000000	39.9029508
		poids	poids	243	0	1800.00	5140.00	3340.00	3367.45	3372.50	787.9682577
		acceleration	acceleration	243	0	8.0000000	22.2000000	14.2000000	14.9716049	15.0000000	2.7190133
		Age		243	0	1.0000000	13.0000000	12.0000000	7.4032922	7.0000000	3.6750653

Nous pouvons remarquer que ce nouveau fichier sans valeurs manquantes présente presque les même statistique descriptives que l'ancien fichier. Ceci est logique dans le sens où les valeurs manquantes n'influence pas sur la médiane.

1.4 Création du fichier prédictif

1.4.1 Rajout des variables continue

Une idée très pertinente est de créer les trois variables Asie USA Europe :

— USA=1 si le modèle est d'origine USA et 0 sinon.

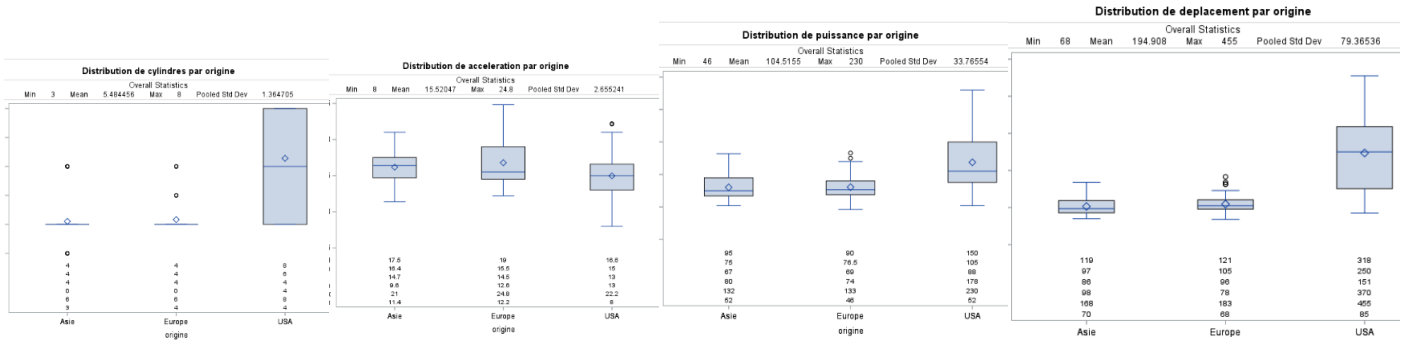
— Asie=1 si le modèle est d'origine Asie et 0 sinon.

— Europe=1 si le modèle est d'origine Europe et 0 sinon.

qui prennent des valeurs continues qui remplace la variable origine (non-continue) et me serviront à la modalisation de la régression. Vous pouvez visualiser en exécutant le code.

1.4.2 Elimination des valeurs atypiques

On remarque qu'il y'a des valeurs atypique qui rendent le modèle moins bon, il s'agit :



Après avoir éliminer ces valeurs notre fichier donne R carré 0.8431 et R ajusté de 0.8395 et une erreur inférieur à 0.0001 significatif, le modèle est mieux qu'avant.

1.4.3 Corrélation entre les variable

Coefficients de corrélation de Pearson, N = 352
Proba > |r| sous H0: Rho=0

	mpg	poids	puissance	acceleration	cylindres	deplacement	USA	Europe	Asie	age
mpg	1.00000	-0.86548 <.0001	-0.78573 <.0001	0.40838 <.0001	-0.82758 <.0001	-0.84754 <.0001	-0.59063 <.0001	0.20278 0.0001	0.53305 <.0001	-0.57589 <.0001
poids	-0.86548 <.0001	1.00000	0.86141 <.0001	-0.44950 <.0001	0.90079 <.0001	0.93101 <.0001	0.62087 <.0001	-0.31744 <.0001	-0.46596 <.0001	0.31349 <.0001
puissance	-0.78573 <.0001	0.86141 <.0001	1.00000	-0.68833 <.0001	0.84428 <.0001	0.90222 <.0001	0.49769 <.0001	-0.25976 <.0001	-0.36873 <.0001	0.41381 <.0001
acceleration	0.40838 <.0001	-0.44950 <.0001	-0.68833 <.0001	1.00000	-0.55189 <.0001	-0.58795 <.0001	-0.26797 <.0001	0.09664 0.0702	0.23766 <.0001	-0.28904 <.0001
cylindres	-0.82758 <.0001	0.90079 <.0001	0.84428 <.0001	-0.55189 <.0001	1.00000	0.95181 <.0001	0.62123 <.0001	-0.36523 <.0001	-0.42314 <.0001	0.36033 <.0001
deplacement	-0.84754 <.0001	0.93101 <.0001	0.90222 <.0001	-0.58795 <.0001	0.95181 <.0001	1.00000	0.64807 <.0001	-0.37074 <.0001	-0.45071 <.0001	0.36902 <.0001
USA	-0.59063 <.0001	0.62087 <.0001	0.49769 <.0001	-0.26797 <.0001	0.62123 <.0001	0.64807 <.0001	1.00000	-0.58792 <.0001	-0.68113 <.0001	0.09740 0.0680
Europe	0.20278 0.0001	-0.31744 <.0001	-0.25976 <.0001	0.09664 0.0702	-0.36523 <.0001	-0.37074 <.0001	-0.58792 <.0001	1.00000	-0.19181 0.0003	0.09456 0.0764
Asie	0.53305 <.0001	-0.46596 <.0001	-0.36873 <.0001	0.23766 <.0001	-0.42314 <.0001	-0.45071 <.0001	-0.68113 <.0001	-0.19181 0.0003	1.00000	-0.20376 0.0001
age	-0.57589 <.0001	0.31349 <.0001	0.41381 <.0001	-0.28904 <.0001	0.36033 <.0001	0.36902 <.0001	0.09740 0.0680	0.09456 0.0764	-0.20376 0.0001	1.00000

On remarque que les variables explicatives sont corréllées au **mpg** avec une $\mid corr \mid$>0.5, à l'exception des variables **acceleration** et **Europe** avec une $\mid corr \mid$<0.5. Entre les variables explicatives, beaucoups d'entre eux sont correllées $\mid corr \mid$>0.5 à voir par exemple **poids-puissance** dont $\mid corr \mid$=0.86141. Dans la suite, nous aurons besoin de la corrélation entre ces variables pour réduire, si c'est nécessaire selon l'évolution du **R carré** du modèle, ou non certains variables significative au **mpg**.

1.4.4 Variables significative

On est censé savoir les variables qui sont **significatives** à l'**mpg** pour donner des informations au mpg d'une voiture. Apres une étude algorithmique, de notre dernier fichier, on a :

Résultats estimés des paramètres

Variable	Libellé	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|	Inflation de variance
Intercept	Intercept	B	52.26724	1.84471	28.33	<.0001	0
poids	poids	1	-0.00437	0.00053256	-8.21	<.0001	10.52616
puissance	puissance	1	-0.00895	0.01067	-0.84	0.4019	8.61434
acceleration	acceleration	1	-0.28128	0.08876	-3.17	0.0017	2.48963
cylindres	cylindres	1	-0.49953	0.27349	-1.83	0.0686	11.26068
deplacement	deplacement	1	-0.00526	0.00628	-0.84	0.4025	22.19006
USA		B	-3.03076	0.47127	-6.43	<.0001	2.43409
Europe		B	-2.41394	0.52381	-4.61	<.0001	1.67355
Asie		0	0	.	.	.	.
age		1	-0.65360	0.04410	-14.82	<.0001	1.32819

Nous pouvons donc constater que seules les variables **poids acceleration Age USA Europe** qui sont **significative au mpg** avec une **p-value inférieur à 0.05**. Ceci dit que nous nous concentrons à ce cinq variables pour donner une information au **mpg**.

1.4.5 Élimination des eléments influant négativement le modèle

Le **Résidus studentisé et Distance de cook** nous montre que les valeurs dont les **obs** associés aux identifiants 321, 269, 328, 239, 327, 243, 324, 308, 382 et 348, influent négativement notre modèle. Après avoir supprimé ces valeurs on a constaté une forte augmentation du **R carrée** et du **R ajusté** qui passent de **0.84** à **0.8754** :

Analyse de variance					
Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	8	16944	2118.05552	301.16	<.0001
Erreur	343	2412.34447	7.03307		
Total sommes corrigées	351	19357			

Root MSE	2.65199	R carré	0.8754
Moyenne dépendante	22.89432	R car. ajust.	0.8725
Coeff Var	11.58364		

C'est donc jusqu'à là, mon meilleur fichier prédictif,vous pouvez visualiser le fichier en executant le fichier **auto-mpg-clean** dans le code. On constate que le fichier prédictif contient. **352** observation.

1.5 Explication du nouveau modèle prédictif

1.5.1 Analyse du modèle

Le modèle présente un erreur (**une p-value**) **global inférieur à 0.0001** ce qui montre que le modèle est bonne. Il présente aussi un **R carré** égale à **0.8752** et **R ajusté** egale à **0.8723** qui se rapprochent assez significativement et qui sont très proche de 1 ce qui veut dire que notre modèle est robuste. Seules les variables **poids acceleration Age USA Europe** qui sont **significatives** (**p-value inférieur à 0.05**) pour donner une information à la variable **mpg**. Ainsi, le modèle nous donne l'équation de régression suivante :

$$\textbf{mpg} = 52.26724 - 0.00437 \times \textbf{poids} - 0.28128 \times \textbf{acceleration} - 3.03076 \times \textbf{USA} - 0.65360 \times \textbf{age} - 2.41394 \times \textbf{Europe}$$

1.5.2 Analyse descriptive du fichier prédictif selon le mpg

Variable	Libellé	N	Nbre manquant	Minimum	Maximum	Intervalle	Moyenne	Médiane	Ec-type
mpg	mpg	352	0	9.0000000	40.8000000	31.8000000	22.8943182	22.0000000	7.4261394
poids	poids	352	0	1613.00	5140.00	3527.00	3009.32	2834.00	862.3492561
puissance	puissance	352	0	46.0000000	230.0000000	184.0000000	106.3323864	95.0000000	38.9541374
acceleration	acceleration	352	0	8.0000000	21.9000000	13.9000000	15.3332386	15.4500000	2.5162725
cylindres	cylindres	352	0	4.0000000	8.0000000	4.0000000	5.5568182	4.0000000	1.7368666
deplacement	deplacement	352	0	68.0000000	455.0000000	387.0000000	201.4502841	153.0000000	106.2262661
USA		352	0	0	1.0000000	1.0000000	0.6761364	1.0000000	0.4686148
Europe		352	0	0	1.0000000	1.0000000	0.1420455	0	0.3495937
Asie		352	0	0	1.0000000	1.0000000	0.1818182	0	0.3862436
age		352	0	1.0000000	13.0000000	12.0000000	7.2215909	7.0000000	3.6990474

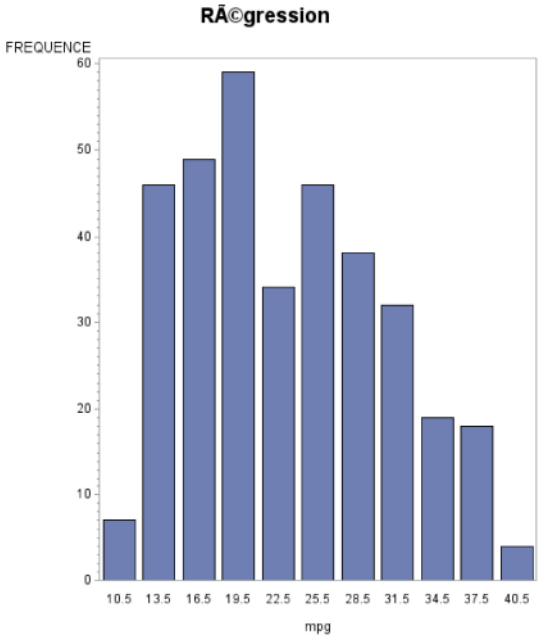
Nous constatons que globalement en **moyenne** les voiture ont un **mpg** égale à **22.8943182 Kilométrage/miles par gallon** dont le maximum est égale à **40.8 km/mg**, les **mpg** des voitures sont dispersé à **7.4261394**. Ce qui veut dire que par rapport au **miles par gallon**, une voiture ayant un **mpg** supérieur ou égale à **22.8943182 Kilométrage/miles par gallon** est meilleur.

origine	N obs	Variable	Libellé	N	Nbre manquant	Minimum	Maximum	Intervalle	Moyenne	Médiane	Ec-type
Asie	64	poids	poids	64	0	1613.00	2711.00	1098.00	2158.14	2132.50	266.5058960
		puissance	puissance	64	0	52.0000000	100.0000000	48.0000000	75.9062500	71.0000000	14.0867142
		deplacement	deplacement	64	0	71.0000000	144.0000000	73.0000000	100.0312500	97.0000000	16.7331709
		acceleration	acceleration	64	0	13.5000000	21.0000000	7.5000000	16.6000000	16.5000000	1.6871127
		age		64	0	1.0000000	13.0000000	12.0000000	5.6250000	5.0000000	3.8026724
		mpg	mpg	64	0	20.0000000	40.8000000	20.8000000	31.2796875	32.0000000	4.8667051
		cylindres	cylindres	64	0	4.0000000	4.0000000	0	4.0000000	4.0000000	0
Europe	50	poids	poids	50	0	1825.00	3270.00	1445.00	2337.52	2221.50	407.8784953
		puissance	puissance	50	0	46.0000000	115.0000000	69.0000000	81.5000000	78.0000000	17.4814771
		deplacement	deplacement	50	0	68.0000000	141.0000000	73.0000000	104.8000000	104.5000000	15.6218058
		acceleration	acceleration	50	0	12.2000000	21.9000000	9.7000000	15.9300000	15.5000000	2.2374594
		age		50	0	1.0000000	13.0000000	12.0000000	8.0800000	8.0000000	3.1676941
		mpg	mpg	50	0	18.0000000	37.3000000	19.3000000	26.5900000	26.0000000	4.6663034
		cylindres	cylindres	50	0	4.0000000	4.0000000	0	4.0000000	4.0000000	0
USA	238	poids	poids	238	0	1800.00	5140.00	3340.00	3379.35	3372.50	787.2183658
		puissance	puissance	238	0	63.0000000	230.0000000	167.0000000	119.7310924	105.0000000	39.6361505
		deplacement	deplacement	238	0	86.0000000	455.0000000	369.0000000	249.0273109	250.0000000	97.8029982
		acceleration	acceleration	238	0	8.0000000	21.0000000	13.0000000	14.9672269	15.0000000	2.6189704
		age		238	0	1.0000000	13.0000000	12.0000000	7.4705882	8.0000000	3.6660687
		mpg	mpg	238	0	9.0000000	39.0000000	30.0000000	19.8630252	18.1500000	6.3071036
		cylindres	cylindres	238	0	4.0000000	8.0000000	4.0000000	6.3025210	6.0000000	1.6563629

Ainsi en **moyenne** les voiture Asiatique ont un **mpg** significatif égale à **31.28 km/mg** par rapport aux autres continent, puis celles d'Europes ont en **moyenne** un **mpg** égale à **26.59 km/mg** mieux que les voitures des USA. Cet même ordre se présente au niveau des maximum des **mpg**.

1.5.3 Analyse histogramme de la variable mpg

Analysons l'histogramme de la variable **mpg** du fichier prédictif :



Nous constatons que on a plus des voitures dont leurs **mpg** se concentrent approximativement à **19.5 km/mg**, on a moin des voitures qui ont un **mpg** à **40.5 km/mp** approximativement.

Chapitre 2

Prédiction d'un mpg des nouvelles données des voiture

A partir de notre régression, nous allons prédire le **mpg** des voitures dont les information est données, voir le fichier auto-à-prédire ci-joint dans le code.

2.1 Construire le fichier à prédire

Ce dernier fichier n'est pas complet, il ne contient pas tout les variables nécessaire pour appliquer la régression de notre modèle, il s'agit donc de rajouter les variables **Age USA Europe Asie**. Avec les mèmes algorithmes utilisés pour la construction de notre modèle, on a construit à nouveau le fichier à prédire, voir le code ci-joint.

2.2 prédiction

Selon les donnés de notre fichier à prédire appliquées à l'équation de la régression, les résultats des **mpg** sont les suivants :

$$\textbf{mpg} = 52.26724 - 0.00437 \times \textbf{poids} - 0.28128 \times \textbf{acceleration} - 3.03076 \times \textbf{USA} - 0.65360 \times \textbf{age} - 2.41394 \times \textbf{Europe}$$

Obs.	Identifiant	cylindres	deplacement	puissance	poids	acceleration	origine	nom_de_la_voiture	age	USA	EUROPE	Asie	mpg
1	160	6	250	105	3897	18.5	USA	chevroelt chevelle malibu	8	1	0	0	21.7741
2	161	6	258	110	3730	19	USA	amc matador	8	1	0	0	22.3633
3	50	4	116	90	2123	14	Europe	opel 1900	12	0	1	0	28.7947
4	51	4	79	70	2074	19.5	Europe	peugeot 304	12	0	1	0	27.4618
5	293	4	86	65	1975	15.2	Asie	maxda glc deluxe	4	0	0	1	36.7466
6	302	4	85	65	2020	19.2	Asie	datsun 210	4	0	0	1	35.4249

On remarque que les deux voiture d'origine Asiatique ont un **mpg** significatif par rapport aux voitures des autres continent, ainsi plus la voiture a moins de poids mieux est l'**mpg**, plus la voiture à moins d'age mieux est l'**mpg**, moins de puissance mieux est l'**mpg** de la voiture, plus la cylindre est petite mieux est l'**mpg** de la voiture. Cet fichier à prédire conserve l'ordre de préférence qu'on a eu quant on a interprété le fichier prédictif.

2.3 Conclusion générale

Sur l'analyse descriptive du fichier prédictif, on a vu qu'en **générale** les voiture ont en **moyenne 22.8943182 Km/mg** ceci nous déconseille d'acheter une voiture d'un **mpg** moins de **22.8943182 km/mg** et on a vu encore que c'est en Asie qu'il y'a l'**mpg** le plus grand et une moyenne d'**mpg** significatif égale à **31.28km/mg**, ceci dit que vaut mieux acheter une voiture Asiatique de plus qui à

un **mpg** supérieure ou égale à **31.28km/mg** . On peut facilement constater ces résultat au niveau de la table à prédire où les deux voiture Asiatique ont un **mpg** significatif que les autres voitures. Ainsi, selon ce fichier qu'on a prédit, on voit que **mieux** est l'**mpg** pour une voiture de **moins** de **poids**, **mieux** est l'**mpg** pour une voiture d'une **petite cylindre**, **mieux** est l'**mpg** pour une voiture de **moins** de **puissance**.