Report of the Artificial Intelligence Project

Student Name : Rami Soussi

Student Number : 0583600

Option of the project : Option 1

Project goal and structure :

The second hand's car market is a very huge business around the world.

When someone buys a car and wants to sell it later, he could sell it at a wrong price, and had a lot of losses.

This problem may be encountred by the seller who may be unable to estimate the real cost of the car.

The goal of my project is to find the most important characteristics (=features in the dataset) that have the biggest effect in evaluating the price of the car.

The dataset I used is a dataframe of 6019 cars with 11 characteristics and their prices. The cars in the table are samples of the cars sold in different Indian towns between 1998 and 2019.

I divided my feature processing project into three parts :

The first part is the data cleaning step. I find some Nan values in my inintial dataset, and there were numerical features written with units as strings. Besides, I allocated this part to normalize the data, in order to have a better estimation afterwards.

In the second part, I used three statistical methods (Pearson correlation, Mutual information, One-way Anova) to evaluate the importance of each feature in the dataset, and then I verified the results using the simple linear regression model and a built-in sklearn function named SelectKBest.

In the third part, I added new features to the dataset based on the principal components analysis and I checked the improvement of the predictability by the multivariate linear regression model.

Description of the machine Learning techniques:

1) Peason correlation :

The correlation between two variables gives an insight on their dependence. It refers to the degree to which they are linearly related. They are useful because they can indicate a predictive relationship. It measures the strength of the linear relationship between two variables. It is unit-less and ranges between –1 and 1. The closer to +1 (-1), the stronger the positive (negative) linear relationship.

2) Mutual information :

Mutual information is one of many quantities that measures how much one variable tells us about another. It can be thought of as the reduction in uncertainty about one variable given knowledge of another. High mutual information indicates a large reduction in uncertainty; low mutual information indicates a small reduction.

3) one-way ANOVA :

One-Way ANOVA ("analysis of variance") compares the means of two or more independent groups in order to determine whether there is statistical evidence that the associated population means are significantly different.

4) Simple Linear Regression :

Linear Regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. It has an equation of the form Y=a+bX, where X is the explanatory variable and Y is the dependent variable. The slope of the line is b, and a is the intercept.

5) Multivariate Linear Regression :

This is quite similar to the simple linear regression model, but with multiple independent variables contributing to the dependent variable and hence multiple coefficients to determine and complex computation due to the added variables.

6) PRINCIPAL COMPONENT ANALYSIS :

Principal Component Analysis, or PCA, is a dimensionality-reduction method that is used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

Project Content and Results :

First part :

In the first part of the project, I changed the type of string inputs into floats by droping their units.

Then I used a normalisation function to set them up between 0 and 1 in order to have coherent cofficients in next steps.

Besides, I found that there were categorical features which are : Location of the car, name of the car, Fuel type, type of transmission and the owner type that means the number of previous owners.

Those features have the following counts :

```
count_Location :  11
count_Name :  1876
Count_Fuel_type :  5
Count_Owner_type :  4
Count_Transmission :  2
```

I decided to drop the column ‘Name’ because it has many categories and won’t be significant for the regression model.

Also, I decided to drop the column ‘Location’ because it has a lot of categories compared to the remaining features.

Afterwards, I changed the remaining categorical features into numerical ones.

For the changement of the Fuel type feature, I used the mean of prices belonging to each category in order to have a better fitting in the regression model.

```
Fuel_Type
LPG           2.487000
CNG           3.516786
Petrol        5.701100
Diesel       12.840605
Electric     12.875000
Name: Price, dtype: float64
```

Second part :

In the second part, the three statistical methods gave me the following ranking of feature importances :

	Pearson Correlation	Mutual Information	Anova one-way	importance ranking
Power	1	1	3	1
Engine	2	2	4	2
Transmission	3	5	5	3
Mileage	5	3	7	4
Kilometers_Driven	9	8	1	5
Fuel_Type	4	6	8	6
Owner_Type	7	9	2	7
Year	6	4	9	8
Seats	8	7	6	9

NB : the importance ranking column gives the importance of each feature based on the sum of its ranking by the three statistical methods.

I found that the power and the engine of the cars determines the most its price. Then comes the other characteristics like transmission and mileage. We remark that the number of seats in car doesn’t affect largely its price.

To verify the correctness of the results found statistically, I implemented a simple linear regression model that has to be trained using the dataset samples and then predict the price of cars.

I used the mean square error to evaluate the regression model, which is the difference between true price and predicted price, to the power 2.

I found the following results :

```
Power                54.428636
Engine               72.722804
Transmission         82.169533
Fuel_Type           112.597413
Mileage             113.415600
Year                113.481728
Owner_Type          123.957597
Seats               124.799647
Kilometers_Driven   125.132163
Name: MSE_of_each_feature, dtype: float64
```

We remark that the three first features (Power, Engine and Transmission) give a low mean square error compared to other features which is coherent with the results found previously with statistical methods.

The last step in the second part is the implementation of the sklearn built-in function SelectKBest that selects the best k features in the dataset.

This function gives the following results :

```
best 1 features are:  ['Power']
best 2 features are:  ['Power', 'Engine']
best 3 features are:  ['Power', 'Engine', 'Transmission']
best 4 features are:  ['Power', 'Engine', 'Transmission', 'Year']
best 5 features are:  ['Power', 'Engine', 'Transmission', 'Year', 'Fuel_Type']
best 6 features are:  ['Power', 'Engine', 'Transmission', 'Year', 'Fuel_Type', 'Mileage']
best 7 features are:  ['Power', 'Engine', 'Transmission', 'Year', 'Fuel_Type', 'Mileage', 'Seats']
best 8 features are:  ['Power', 'Engine', 'Transmission', 'Year', 'Fuel_Type', 'Mileage', 'Seats', 'Kilometers_Driven']
best 9 features are:  ['Power', 'Engine', 'Transmission', 'Year', 'Fuel_Type', 'Mileage', 'Seats', 'Kilometers_Driven', 'Owner_
Type']
```

We found again that the three best features are the same : Power, Engine and Transmission.

The less significative features are : Seats, Kilometers driven and Owner type.

This classification is logic because the price depends the most on the measurement of engine power which is the brake horsepower. It varies in our dataset between 40bhp and 500 bhp.

The next important feature is the engine displacement which is the cylinder volume expressed using the cubic centimetres (cc = millilitres). In our dataset, the cylinder volume is between 100 cc and 6000 cc .

For this reason, the engine displacement and its power are often used in advertising.

The third important characteristic is the type of transmission which is either manual or automatic. The cars with automatic transmission are more expensive than those with manual transmission.

This can be shown by the mean of prices of each category.

```
data.groupby('Transmission').mean()['Price'].sort_values()
```

```
Transmission
Manual        5.332703
Automatic    19.843971
Name: Price, dtype: float64
```

Among the less important features, we found the number of seats, which is logic since the majority of cars have 5 seats.

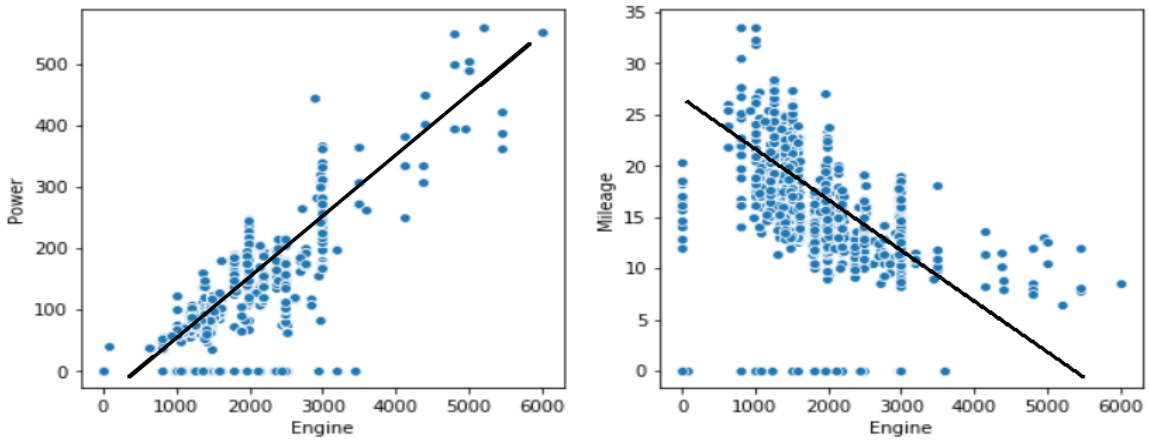
```
data.groupby('Seats').count()
```

Seats	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power
0.0	1	1	1	1	1	1	1	1
2.0	16	16	16	16	16	16	16	16
4.0	99	99	99	99	99	99	99	99
5.0	5014	5014	5014	5014	5014	5014	5014	5014
6.0	31	31	31	31	31	31	31	31
7.0	674	674	674	674	674	674	674	674
8.0	134	134	134	134	134	134	134	134
9.0	3	3	3	3	3	3	3	3
10.0	5	5	5	5	5	5	5	5

The simple linear regression and SelectKbest show that the mutual information and pearson correlation methods are more precise than the anova method.

Pearson correlation and mutual information give almost the same results as simple linear regression and SelectKbest function, which is not the case for the analyse of variance method (one-way anova)

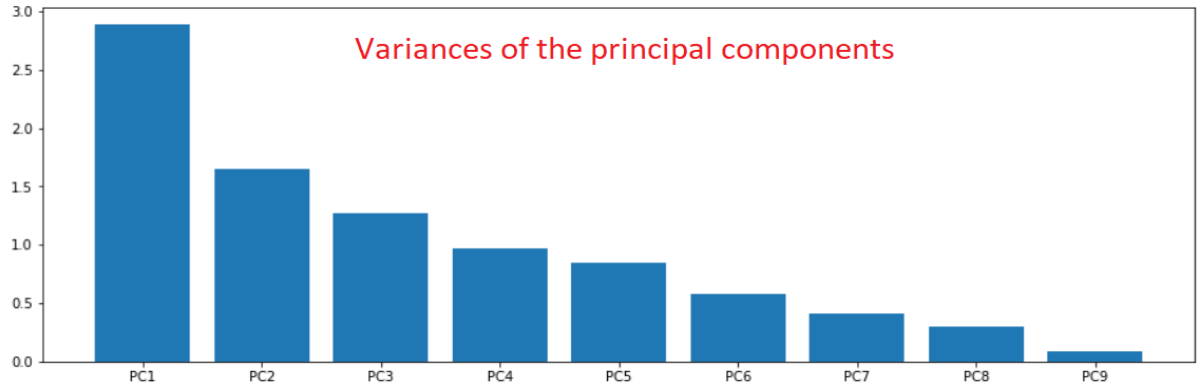
Third part:



As we can see in this scatterplots, the engine is positively correlated with power and negatively correlated with mileage. Those axes are called axes of variation.

The whole idea of Principal Components Analysis (PCA) is describing the data with its axes of variation instead of describing it with its original features. The axes of variation become the new features. For example, we can describe cars by 'Engine*Power' or 'Engine / Mileage'.

PCA also tells us the amount of variation in each component. We can see from the figure that the components are classified according to their variances, high variance could lead to a better prediction.



Using the PCA techniques, we aim to improve the predictability of our regression model.

The Machine Learning algorithm I used in prediction is Multivariate Linear Regression, and I used Mean Square Error to evaluate its prediction.

I won't use the principal components directly as features because it doesn't change the MSE of the prediction, as shown here :

```
Entrée [85]: train_test_split_and_predict(X_pca,y)

             Mean Square Error is  37.63868639934753
```

Mean Square Error of Principal Components

```
Entrée [62]: train_test_split_and_predict(data,y)

             Mean Square Error is  37.63868639934721
```

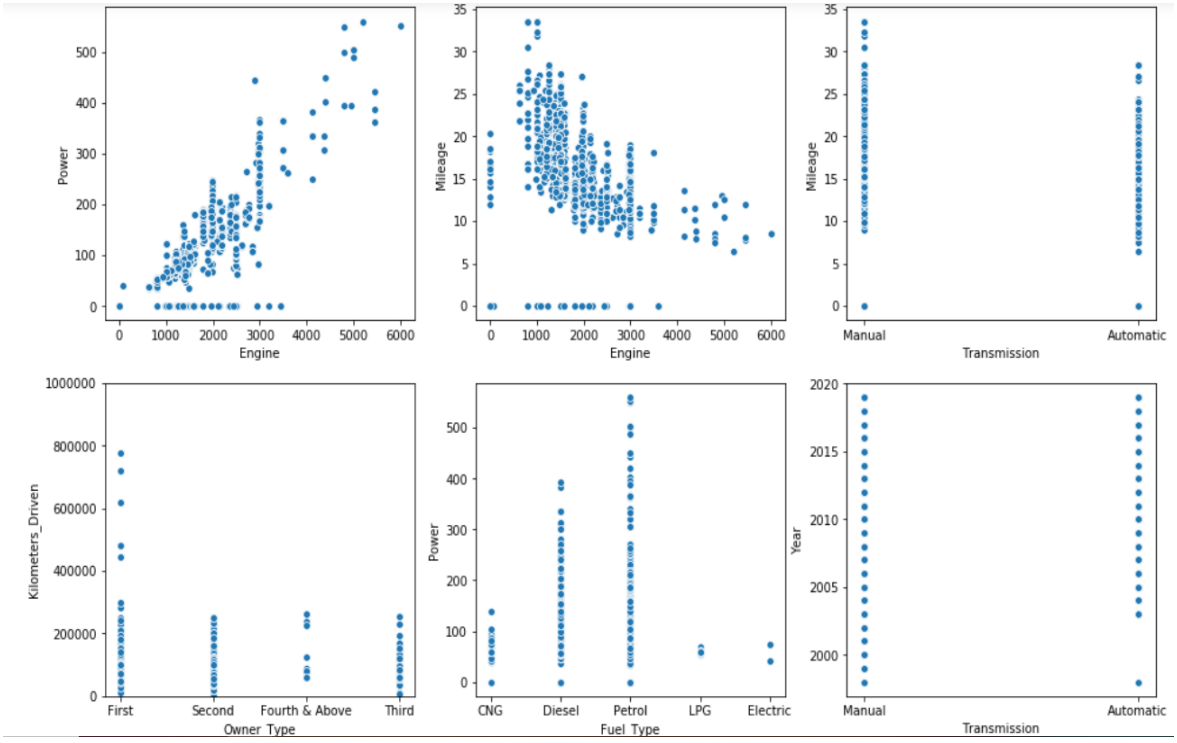
Mean Square Error of Original Features

In fact, the main use of PCA technique in my project is creating new features based on the coefficients of the original features in the principal components, which are called loadings.

```
loadings(data)
```

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Year	-0.027803	0.651518	0.025822	-0.038058	0.058484	-0.658152	-0.096830	0.352870	-0.046464
Kilometers_Driven	0.067990	-0.246540	0.347906	0.693442	0.534456	-0.215849	-0.000099	0.019781	0.008292
Fuel_Type	0.258440	0.246948	0.494129	0.274261	-0.465450	0.364262	0.201011	0.385422	0.120652
Transmission	0.390967	0.146936	-0.435677	0.217598	-0.018701	-0.117372	0.721925	-0.214074	-0.085219
Owner_Type	0.044353	-0.537973	-0.035013	0.124964	-0.617613	-0.549055	-0.056859	0.074954	-0.004231
Mileage	-0.382380	0.348411	0.141855	0.353686	-0.322141	-0.017840	-0.111265	-0.673371	-0.131442
Engine	0.553725	0.024489	0.062647	-0.034877	-0.016518	0.057480	-0.363258	-0.096677	-0.736729
Power	0.515381	0.134404	-0.199852	0.114992	-0.027607	-0.009697	-0.460460	-0.237762	0.627490
Seats	0.233555	-0.051959	0.615713	-0.489869	0.087677	-0.261926	0.263533	-0.394703	0.148813

The following scatterplots confirm the different relations existing between features that have been calculated by PCA.



We remark in PC1 that Power and Engine have the biggest coefficients. Thus, I added a new feature which is their product.

```
X1 = data.copy()
X1["Feature1"] = X1.Power * X1.Engine
train_test_split_and_predict(X1,y)
```

```
Mean Square Error is  33.209276620109016
```

We remark that the MSE decreased (33,2 < 37,63), which proves the efficiency of the added feature.

I added then new columns to the dataset to improve the predictability based on the loadings of original features.

```
X2 = data.copy()
X2["Feature1"] = X2.Power * X2.Engine
X2["Feature2"] = X2.Power / (0.1+X2.Kilometers_Driven)
X2["Feature3"] = X2.Engine / (0.1+X2.Mileage)
train_test_split_and_predict(X2,y)
```

```
Mean Square Error is  30.732571039509246
```

```
X3 = data.copy()
X3["Feature1"] = X3.Power * X3.Engine
X3["Feature2"] = X3.Power / (0.1+X3.Kilometers_Driven)
X3["Feature3"] = X3.Engine / (0.1+X3.Mileage)
X3["Feature4"] = X3.Kilometers_Driven / (0.1+X3.Year)
train_test_split_and_predict(X3,y)
```

```
Mean Square Error is  29.942499215748928
```

We remark that MSE decreased when I added new columns. That's the point of PCA method which improves the predictability of Machine Learning models.

Conclusion:

The project gives an insight into the second hand's car market in india, and it can be the case of this business in the whole world. I hope it helps sellers,and buyers as well, to evaluate the real cost of their cars.

References :

https://www.kaggle.com/learn/feature-engineering

https://www.kaggle.com/iabhishekmaurya/used-car-price-prediction

https://en.wikipedia.org/wiki/Mutual_information

https://en.wikipedia.org/wiki/One-way_analysis_of_variance