INFOSYS 722 Assignment 4

What will affect the fuel consumption of the car

Iteration 4 - BDAS (Steps 1 - 8)

Mingwei(Tom) Wang | mwan676

Table of Contents

[1 Business Understanding 3](#_Toc52552148)

[1.1 Determining Business Objectives 3](#_Toc52552149)

[1.2 Assessing the Situation 3](#_Toc52552150)

[1.2.1 Resource Inventory 3](#_Toc52552151)

[1.2.2 Requirements, Assumptions, and Constraints 4](#_Toc52552152)

[1.2.3 Risks and Contingencies 4](#_Toc52552153)

[1.2.4 Cost/Benefit Analysis 4](#_Toc52552154)

[1.3 Determining Data Mining Goals 4](#_Toc52552155)

[1.3.1 Data Mining Goals 4](#_Toc52552156)

[1.3.2 Data Mining Success Criteria 5](#_Toc52552157)

[1.4 Project Plan 5](#_Toc52552158)

[2. Data Understanding 6](#_Toc52552159)

[2.1 Collecting Initial Data 6](#_Toc52552160)

[2.2 Describing Data 7](#_Toc52552161)

[2.3 Exploring Data 7](#_Toc52552162)

[2.4 Verifying Data Quality 16](#_Toc52552163)

[3 Data Preparation 18](#_Toc52552164)

[3.1 Selecting Data 18](#_Toc52552165)

[3.2 Cleaning Data 19](#_Toc52552166)

[3.3 Constructing New Data 22](#_Toc52552167)

[3.4 Integrating Data 22](#_Toc52552168)

[3.5 Formatting Data 23](#_Toc52552169)

[4 Data transformation 23](#_Toc52552170)

[4.1 Reduce the data 23](#_Toc52552171)

[4.2 Project the data 24](#_Toc52552172)

[5. Data-mining method(s) selection 26](#_Toc52552173)

[5.1 discussion of data mining methods 26](#_Toc52552174)

[5.2 Select the appropriate data-mining method 27](#_Toc52552175)

[6. Data-mining algorithm(s) selection 28](#_Toc52552176)

[6.1 Conduct exploratory analysis and discuss 28](#_Toc52552177)

[6.2 Select data-mining algorithms based on discussion 28](#_Toc52552178)

[6.3 Build/Select appropriate model(s) and choose relevant parameter(s) 29](#_Toc52552179)

[7. Data Mining 30](#_Toc52552180)

[7.1 Create and justify test designs 30](#_Toc52552181)

[7.2 Conduct data mining 31](#_Toc52552182)

[7.3 Search for patterns 33](#_Toc52552183)

[2.3 data exploring 33](#_Toc52552184)

[8. Interpretation 45](#_Toc52552185)

[8.1 Study and discuss the mined patterns 45](#_Toc52552186)

[8.2 Visualize the data, results, models, and patterns 46](#_Toc52552187)

[8.3 Interpret the results, models, and patterns 51](#_Toc52552188)

[8.4 Assess and evaluate results, models, and patterns 52](#_Toc52552189)

[8.5 Iteration 52](#_Toc52552190)

# 1 Business Understanding

## 1.1 Determining Business Objectives

Fuel is the power source for all kinds of engines, and one of the most wildly used fuel is a fossil fuel. Fossil fuel is the ancient plant, and animals remain after being pressured and high temperature heated for over hundreds of millions of years. Fossil fuel is primarily coal petroleum and natural gas; all of them are considered as non-renewable resources. As non-renewable resources, they can not regenerate themselves, or they regenerate so slowly that they can be ignored. In either way, it means they will be used up someday. As listed in the 17 global goals for sustainable development by the UN, 'affordable and clean energy' is one of the goals that focus on the environmental aspect. Furthermore, slowing down the consumption of fossil fuel is an excellent way to approach this purpose. Considering the air pollution produced by fuel combustion, 'climate action' is also can be approached.

The consumption of cars is the most related to everyone's daily life, and this research will use data mining to discuss what is related to cars' fuel consumption and how to reduce it.

Hence, the research will focus on the following objectives:

* Figure out what factors will affect the fuel consumption of a car and how closely their relationship between these factors and fuel consumption.
* Identify what the fuel consumption rate of the car is.

This research will be considered as successful when the following criteria are met:

* Identify the rough weight of each factor of influence on the fuel consumption

## 1.2 Assessing the Situation

### 1.2.1 Resource Inventory

When searching 'fuel consumption' on Kaggle, there are so many topics and datasets. And this research using the dataset from 'Vehicle Fuel Economy Estimates, 1984-2017' published by the US Environmental Protection Agency (<https://www.kaggle.com/epa/fuel-economy>).

Moreover, the GitHub is an excellent website to share and discuss any IT related work. And the AWS Educate account is provided by the Auckland University.

### 1.2.2 Requirements, Assumptions, and Constraints

According to the data publisher, "Fuel economy data are produced during vehicle testing at the Environmental Protection Agency's National Vehicle and Fuel Emissions Laboratory in Ann Arbor, Michigan, and by vehicle manufacturers with EPA oversight.". And since the data is published on Kaggle, which is a public data communication platform, there should be no legal issues for the data set. Therefore the data is reliable, authentic, legal to use, and mostly no risks.

### 1.2.3 Risks and Contingencies

|  |  |
| --- | --- |
| Risk | Contingency Plan |
| Data utility is low | During nine steps of the KDD process, data preparation and transformation will be processed |
| The result maybe can not meet the criteria | More take more iteration of the KDD process to meet the criteria |

(Table 1: risk and contingency plan)

### 1.2.4 Cost/Benefit Analysis

The data is gathered from the public website Kaggle; therefore, no cost for the data. Moreover, in this iteration, step 9 Action is not included, which means no cost for the resulting deployment.

By analyzing and understanding the data, will provide a better view of what affects the car's fuel consumption related to, and it will help with sustainable use of fuel. Also, it will help people reduce fuel costs in their daily life.

## 1.3 Determining Data Mining Goals

### 1.3.1 Data Mining Goals

The data mining goals are:

* Using importance to determine the relationship between the variables and the target
* Using data mining methods to classified the car into fuel economy score

### 1.3.2 Data Mining Success Criteria

In the study, will use the decision tree and clustering methods to determine the relationship between variables and the fuel consumption. The importance of each variable will be considered as the evidence of the relationship

## 1.4 Project Plan

The study is started on the 3rd of October and end on the 23rd of October. The detailed plan is shown in table 2. And also, the Gantt Chart of the time schedule is shown in figure 1.

|  |  |  |
| --- | --- | --- |
| Phase | Time | Risk |
| Business Understanding | 3/10 - 5/10 | Energy innovation |
| Data Understanding | 6/10 - 8/10 | Low data utility |
| Data Preparation | 9/10 - 12/10 | Data may not be cleaned enough |
| Data Transformation | 13/10 – 14/10 | Not remove all dirty data |
| Data Mining Method Select | 15/10 – 16/10 | The wrong method selected |
| Data Mining Algorithm Select | 17/10 – 18/10 | The wrong algorithm selected |
| Data Mining | 19/10 – 21/10 | Can not identify pattern clearly |
| Interpretation | 22/10 – 23/10 | Can not identify pattern clearly |

(Table 2: project plan)

(Figure 1: Gantt Chart of time schedule)

# 2. Data Understanding

## 2.1 Collecting Initial Data

On Kaggle, there are many data set related to the 'fuel consumption', and most of them contain variables like make, model, engine size, and fuel consumed. But the problems of these data set are they are not big enough for data mining purposes, too few variables, or no data description of column title. And after going through dozens of data set, I select the one that is published by the US Environmental Protection Agency (<https://www.kaggle.com/epa/fuel-economy>).

Text

Description automatically generated The dataset selected has 20 attributes and more than 38,000 rows. Therefore, there are enough data to identify their pattern and not too many attributes that may disturb the result. And the attributes are shown in figure 2:

(Figure 2: a result of Python's read )

In the dataset, there are some repeat attributes. Like 'Fuel Type', 'Fuel Type 1', and 'Fuel Type 2'. After browsing the dataset, it shows that the fuel type may be more than one type, and fuel type 1 and 2 will separate store the two different types of fuel. If there is only one type of fuel the car using, the 'Fuel Type 2' will have a null value. In this case, 'Fuel Type 1' and 'Fuel Type 2' are actually repeated. Therefore I will drop them. Attribute Vehicle ID is obviously an ID column. And Model and Make are actually the name of the car, can be dropped too.

## 2.2 Describing Data

The dataset has an attribute 'Fuel Economy Score', but all the scores before the year 2013 are -1. There are actually scores after 2013. Therefore, to use this attribute to the target, I need to only use the data since 2013.

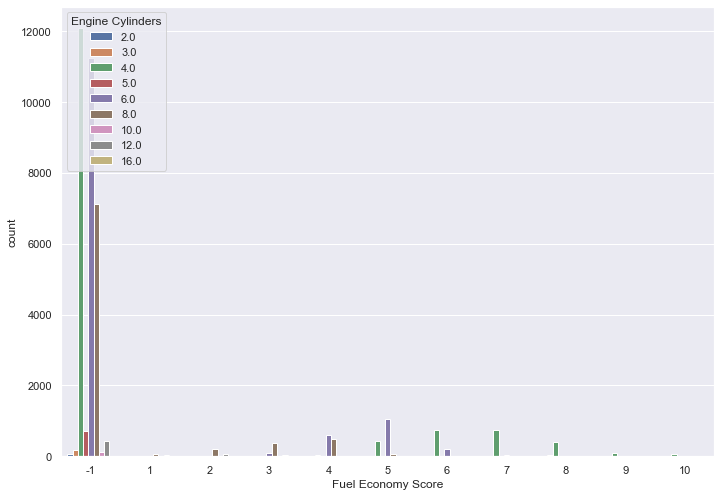
In this dataset, the fuel consumption is measured in miles per gallon(MPG). Therefore, the higher MPG, the less fuel the car consumes. In the dataset, there are three MPG attributes: city, highway, and combined. According to the publisher, 'combined' is an estimate that represents a combination of city driving (55%) and highway driving (45%).

In the dataset, there is continuous variable vehicle ID as record ID, and continuous variable year; nominal variables make, model, and class for the details of the vehicle. The nominal variable Drive classified the car is 2-wheel or 4-wheel drive; also, the car is front-driven or tail driven, and variable Transmission classified the manual or automatic and how many speed of the vehicle. Then engine cylinders and engine displacement are ordinal variables described the engine details. Turbocharger and Supercharger are flags of does the engine has a Turbocharger/ Supercharger or not(with 'T' or blank field). Three fuel type attributes are nominal variables that indicate what kind of fuel the vehicle using. The continuous variable 'annual fuel cost' is estimated as travel 15,000 miles per year (55% under city driving conditions and 45% under highway conditions) and that fuel costs $2.33/gallon for regular unleaded gasoline, $2.58/gallon for mid-grade unleaded gasoline, and $2.82/gallon for premium. At last, the attribute 'Gasoline/Electricity Blended' is a flag that shows the car is a hybrid car or not.

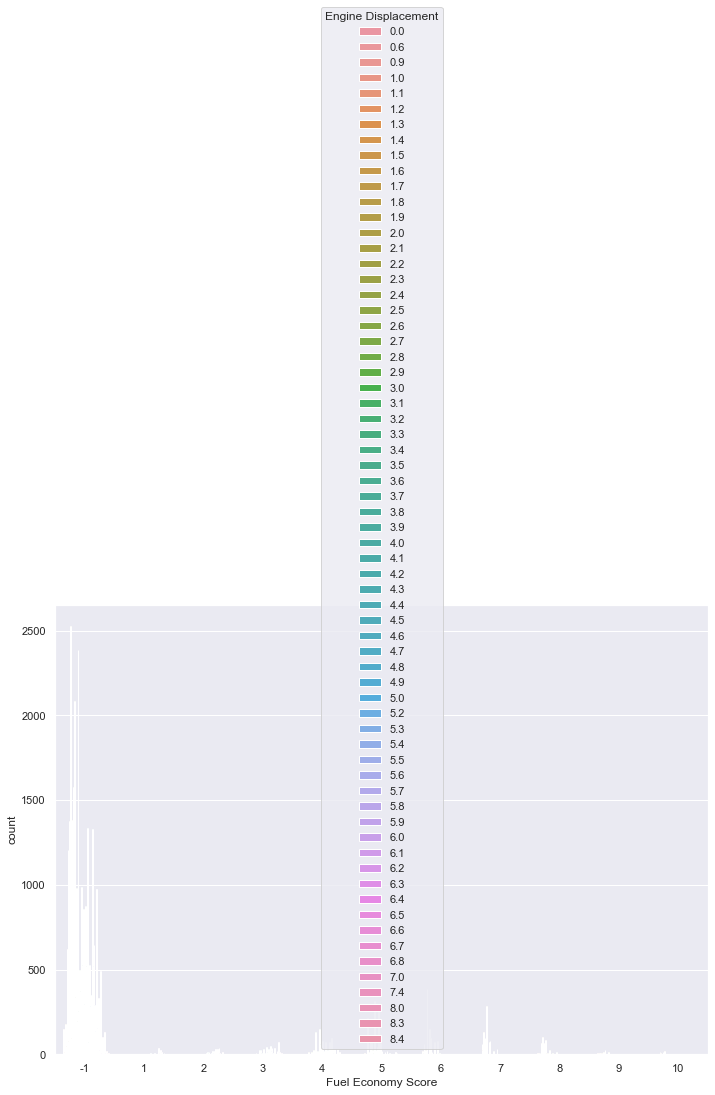
## 2.3 Exploring Data

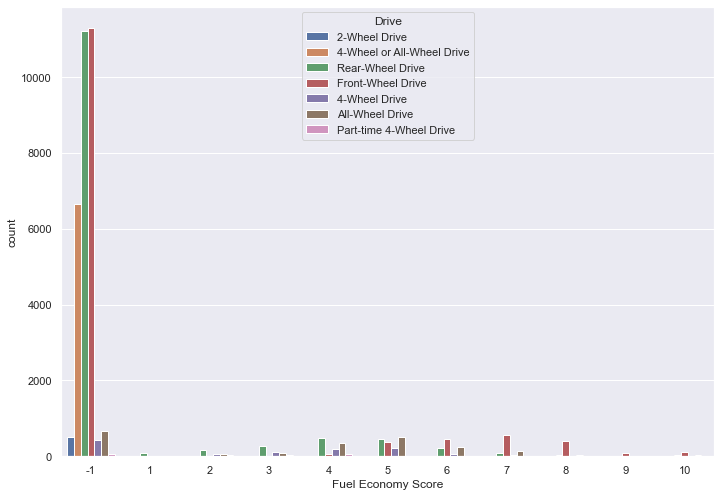
As fuel consumption of a car is a really common topic in everyone daily life, there are some initial hypotheses based on common sense:

* The 4-wheel Drive will consume more than 2-wheel Drive
* The more cylinders engine has, the more fuel will be consumed
* The more displacement engine has, the more fuel will be consumed
* Auto transmission can save fuel compared to the manual Transmission
* Turbocharger and supercharger will make the car cost more fuel

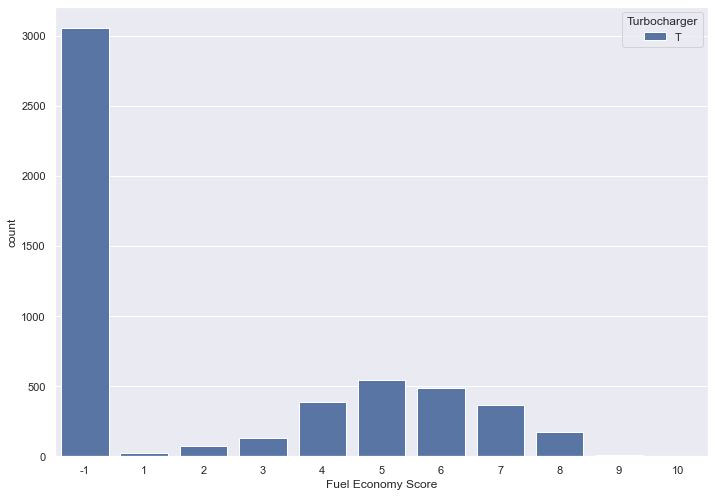
As the fuel economy score is used to measure fuel consumption, the problem transformed into finding the relationship of the fuel economy score. Therefore, creating countplot graphs of fuel economy score versus engine cylinders, engine displacement, turbocharge, drive, and Transmission is the most direct way to show their patterns. (all five graphs are displayed in figures 3-9).

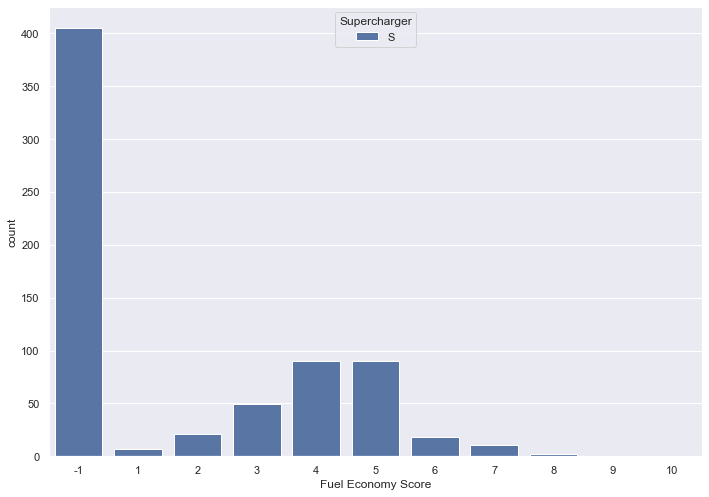
(figure 3: Fuel Economy Score-Engine Cylinders)

(figure 4: Fuel Economy Score-Engine Displacement)

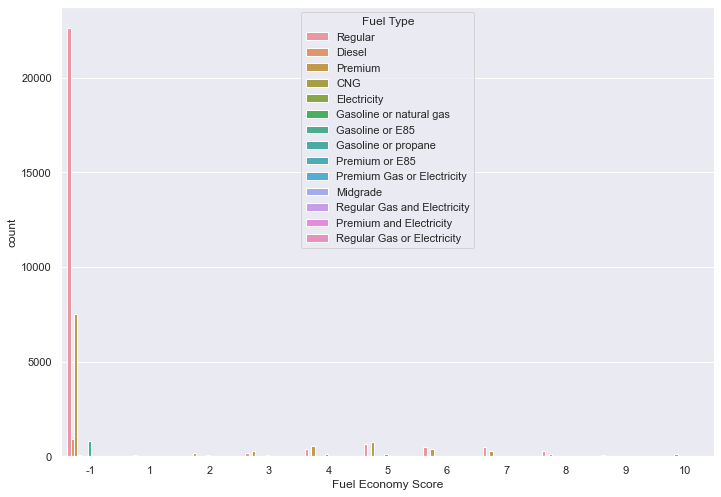
(figure 5: Fuel Economy Score-Drive)

(figure 6: Fuel Economy Score-Transmission)

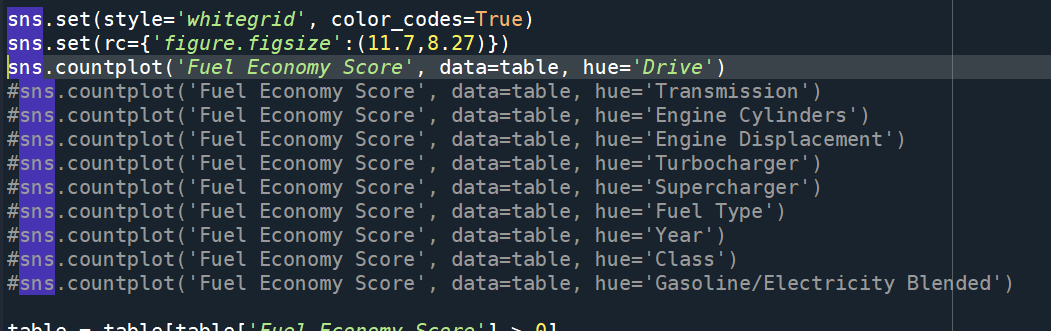


(Figure 7: Fuel Economy Score-Turbocharge)

(figure 8: Fuel Economy Score-Supercharge)

(figure 9: Fuel Economy Score-Fuel Type)

The above figures are created by the python code in figure 10 each line a time.



(figure 10: code of countplot graphs)

In the graph Fuel Economy Score – Displacement and Transmission, maybe because the diminutions are too many, the color of all the bar is white. So I use Tableau created another Fuel Economy Score – Displacement and Transmission plot graph.(figure 11-12)

Chart

Description automatically generated

(figure 11: Fuel Economy Score-Engine Displacement in Tableau)

Chart

Description automatically generated

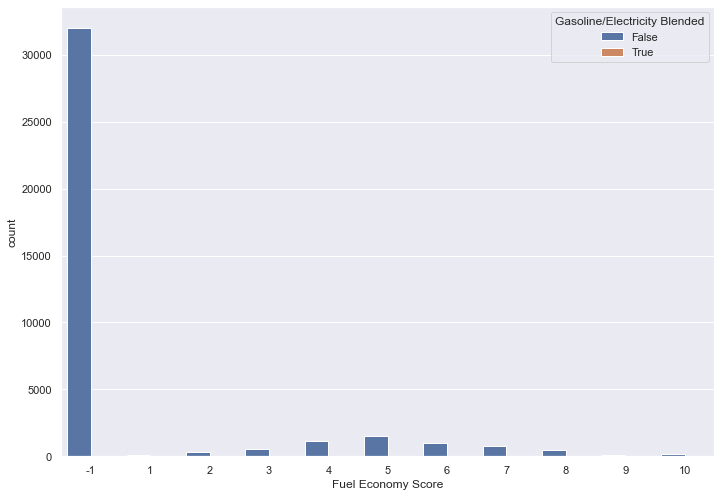
(figure 12: Fuel Economy Score-Engine Transmission in Tableau)

In the graph of engine displacement, it shows the pattern that when the displacement number is low, more displacement will provide more power and actually increased the fuel economy score, but then, more displacement will increase the unnecessary power usage and result in reducing the fuel economy score.

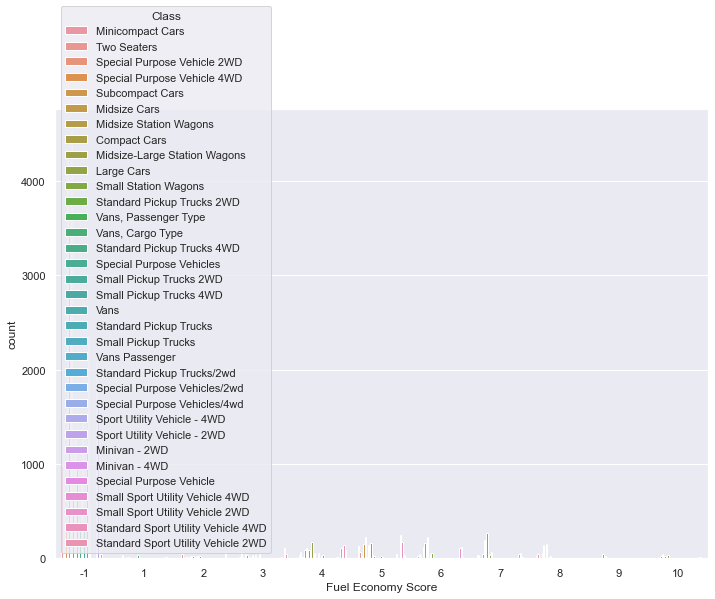
Figures of the turbocharger and supercharger are both showing the pattern that the cars have a turbocharger or a supercharger are mostly have fuel economy score in the medium range(4-6).

And the most obvious pattern is that most cars are having a Fuel Economy Score of '-1'.

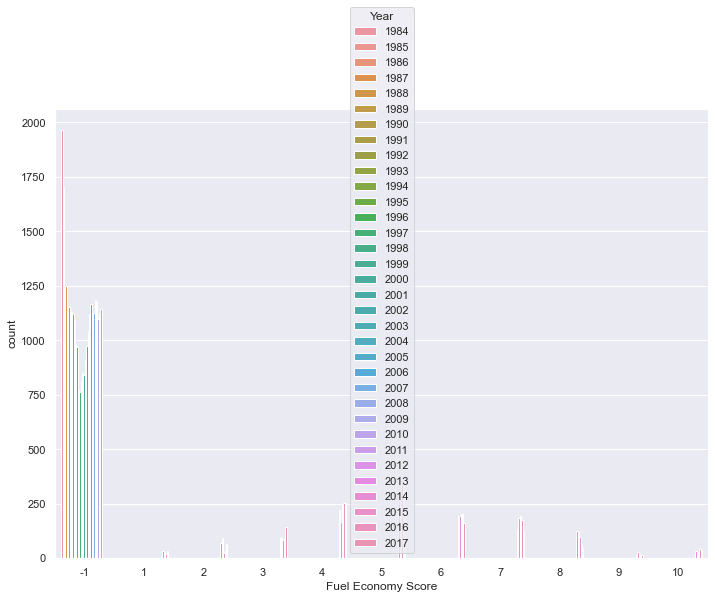
In further exploration, the rest attributes versus fuel economy score countplot graphs are generated. (figure 13-)



(figure 13: Fuel Economy Score- Gasoline/Electricity Blended)



(figure 14: Fuel Economy Score-Class)



(Figure 14: Fuel Economy Score-Year)

In these graphs, there is no obvious pattern shown, and mostly is because the number of -1 score data are too many and affect the graph badly.

## 2.4 Verifying Data Quality

After downloading the dataset, it was in .csv format, which is not supported by Tableau. Therefore, use Microsoft Excel to transform it into .xlsx format and then import it into Tableau.

First, in Python, I print out the data type of each column by using "print(table.dtypes)", 'table' is the variable I used in the code for the datasets. And the result is shown in figure 15.

Text

Description automatically generated

(Figure 15: data types)

The Engine Cylinders, Engine Displacement, and Fuel Economy Score attributes are the dimensions, so they don't have the concept of extreme value even they are integer or float type. Therefore, the column 'City MPG', 'Highway MPG', 'Combined MPG', and 'Annual Fuel Cost' columns should be done the process of removing extreme values. To check does each column has missing values or extreme values, I wrote the code as shown in figure 16 and result in figure 17.

Text

Description automatically generated

(figure 16: code of check missing and extreme values)

Text

Description automatically generated

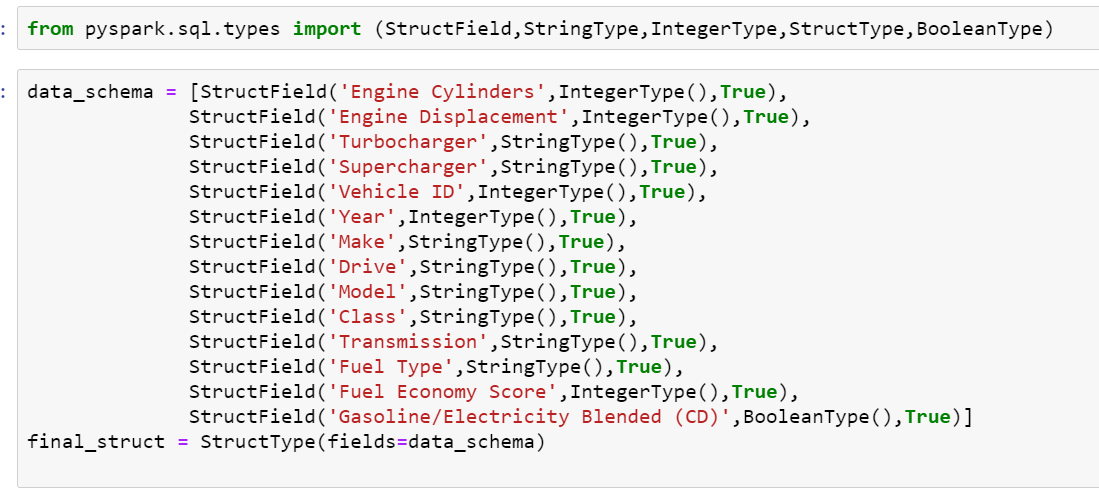
(figure 17: results of checking missing and extreme values)

# 3 Data Preparation

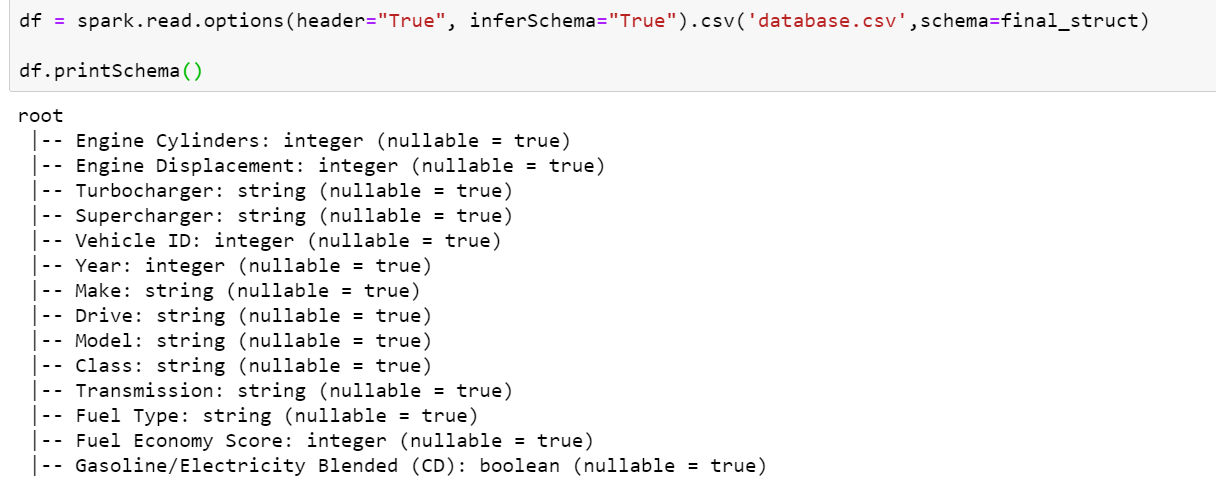
## 3.1 Selecting Data

As discussed above, the original dataset has some irrelevant attributes that should be dropped. In this study, the attributes 'Make', 'Model', 'Fuel Type 1', and 'Fuel Type 2' are dropped because they are repeated as record ID or another attribute. There are three different situations that MPG is measured, and the GMP is just another measurement of fuel economy score. Therefore, to show the importance of the other attributes, all MPG attributes should be ignored. And the same logic to the annual fuel cost, they have a strong connection and it will affect other attributes importance. So, block the annual fuel cost too.

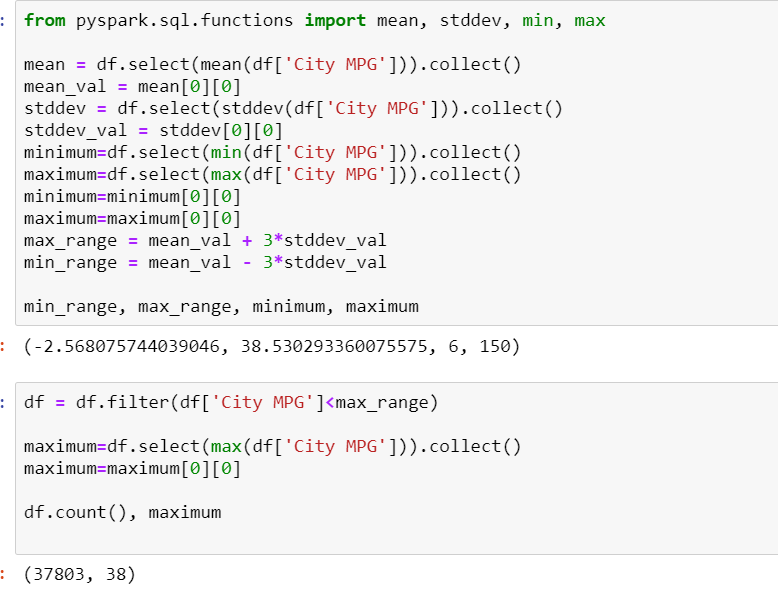
And because the Fuel Economy Score is actually recorded after 2013, and before that time, all the scores are recorded as -1, so all the invalid data should be removed.

 For step 3, I used Jupyter notebook PySpark to do the data preparation. The code to do these two steps is shown in the figure 18 and results in 19.

(figure 18: code of selecting data)

(figure 19: result of data selection)

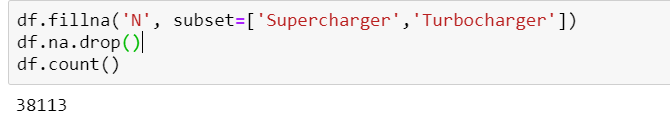
## 3.2 Cleaning Data

 The Engine Cylinders, Engine Displacement, and Fuel Economy Score attributes are the dimensions, so they don't have the concept of extreme value. Therefore, the column 'City MPG', 'Highway MPG', 'Combined MPG', and 'Annual Fuel Cost' columns should be done the process of removing extreme values. To do this, I used the code shown in figure 20.

(figure 20: code and result of removing extreme values)

I used the three-sigma method in this step, and this will cover 99.7% of the data. For the rest 0.3% extreme value in the data, the code will replace them with the maximum or the minimum number in the range. And the max and the min range are printed out too.

The next step is to remove or replace the missing values. In the columns' Turbocharger' and 'Supercharger', the value is left empty if the car does not has a turbocharger or a supercharger. Otherwise, the corresponding attributes will be marked 'T' or 'S'. In this case, to standardize the value, I fill all null values as 0 and 'T' or 'S' to 1. The code is shown in figure 21.



(figure 21: code and result of replacing values in Turbocharger and Supercharger)

After filled Supercharger and Turbocharger, when I do df.na.drop() which is a function that will drop a line with any null value in it. And when I count after this, it return the same number as the original data frame size (38113). This means that the data frame contains no missing value anymore.

## 3.3 Constructing New Data

As shown in figure 18 and 19 above, I defined all the fields I need one by one as string / integer or Boolean. Even they were not that data type before, I defined and changed them into what I want.

Also, I binned the fuel economy score into 3 groups. From 1-3 are group 1, 4-7 are group 2, and 8-10 are group 3, and named the new field as “bin\_score”. And I used the code in figure 22 to achieve this goal.

Text

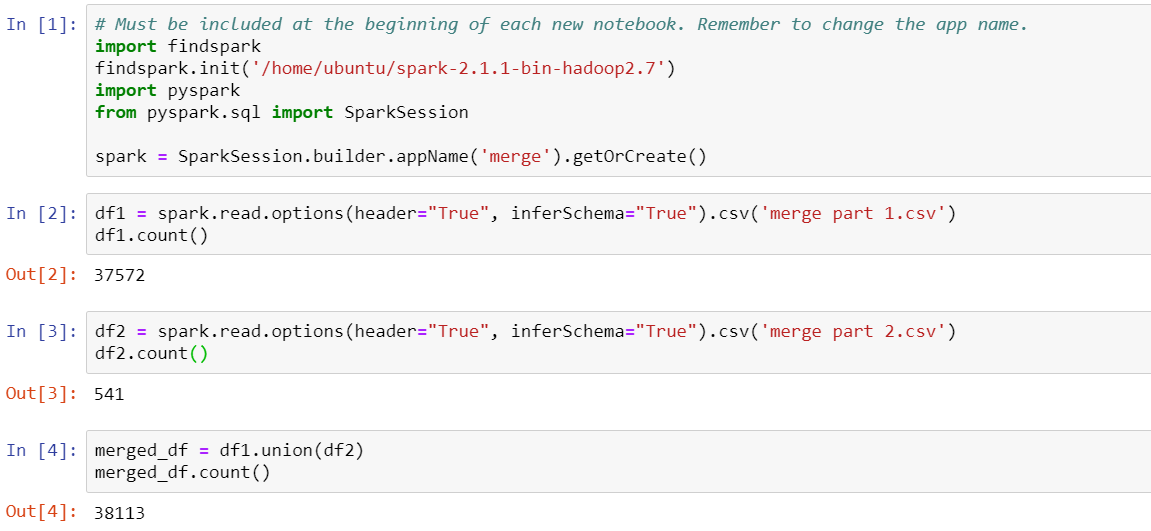
Description automatically generated

(figure 22: code and result of binning ‘Fuel Economy Score’ field)

## 3.4 Integrating Data

Since the dataset I'm using is enough for this study and I didn't collect any other dataset, I'll show steps on how to merge dataset that bifurcated from my original dataset.

Firstly, I divided my dataset into two parts. Create two new excel files named 'merge part 1' and 'merge part 2', copy the attributes title and pasted to the first row, then copy and paste the divided part to the 'merge part 1', and copy the other part of dataset into the 'merge part 2'. Then using the code in figure 22 to merge the two data frame into one named ”merged\_df”.

 With the steps above, the code showed that the merged data frame contains 38113 rows, and compared to the result in original file, they are same. And this indicated that the merge is succeed.

(Figure 23: code of mergeing excel files)

## 3.5 Formatting Data

It is not able to process the string variables like Class, Drive, Transmission, and Fuel Type in Python. Therefore, the method 'encode' is used in the code to transform the other types of variables into the integer variables that can be processed by the data mining algorithms. The code is shown in figure 27.



(Figure 24: code of encoding)

# 4 Data transformation

## 4.1 Reduce the data

A picture containing graphical user interface

Description automatically generatedAs mentioned before, I reduced the columns like make and model, fuel type 1 & 2, and all three MPG attributes, and annual fuel cost. For the reasons of either not relevant or too strong connection with the target that will affect the result of other attributes. And because I decided to use the fuel economy score as the target, I only selected the data after the year 2013. Besides, the attribute vehicle ID is just a sequence number of the car, and the class is actually describing how the car looks like. Therefore, these two columns are not related to the predicted target Fuel Economy Score logically, can be dropped too.

(figure 25: code and result of reducing data)

## 4.2 Project the data

The continuous attributes in the dataset are Year, Fuel Economy Score, and Vehicle ID. I used the code in figure 28 to do the normalization for the Fuel Economy Score, Engine Cylinders, and Engine Displacement and replaced the Engine Cylinders and Engine Displacement column by normalized data. Then, drow the hist diagram by running the code three times with one' plt.hist' function work each time as shown in figure 29-31.A picture containing graphical user interface

Description automatically generated

Chart, bar chart

Description automatically generated(Figure 26: code of normalization)

(Figure 27: hist diagram of normalized Engine Cylinder)Chart, histogram

Description automatically generated

(Figure 28: hist diagram of normalized Engine Displacement)Chart, histogram

Description automatically generated

(Figure 29: hist diagram of normalized Fuel Economy Score)

# 5. Data-mining method(s) selection

## 5.1 discussion of data mining methods

There are five common data mining methods used for almost all the data mining problems: classification, regression, clustering, outlier detection, and association rule learning.

Diagram

Description automatically generated

(Figure 30: Classification of Data Mining Techniques)

Follow to the steps in figure 32, both the clustering and association rule learning are unsupervised learning methods, which have no target data. And in this project, the target variable is the fuel economy score, which is numbers from 1 – 10 but is actually an ordinal variable. Then, the fuel economy score is just a score to roughly evaluate the fuel consumption of a car, so I identify the problem as a discriminant model. And there are definitely some rules behind the connection between the score and the parameters of the car.

And according to the data mining objective, what needs to de done is that base on the variables like Drive, Transmission, etc. to classify the car into the corresponding fuel economy score. And during the data mining process, building the model to figure out how the importance of each attribute is to affect the fuel economy score. Therefore, the classification method is suitable for this study.

## 5.2 Select the appropriate data-mining method

As discussed above, the objective of this study has a target data variable, which is an ordinal variable. So eliminated the clustering and association rule learning. And the target variable type rules out the regression method. Also, the objective is to find out the pattern, not basing on the pattern to find out the outlier. Therefore, only the classification is left.

The other way to show why using the classification method is that the data mining objective is to figure out the pattern and evaluate the importance or each variable. The target classes are well defined already, and using the data mining process to identify their location is the most suitable way to approach the objective.

But as the target value is a number, I will try the regression method as well.

# 6. Data-mining algorithm(s) selection

## 6.1 Conduct exploratory analysis and discuss

Since I decided to use the classification model, the decision tree is always the most classic algorithm to be used. Here are some discussions about several data mining algorithms.

C4.5: The C4.5 algorithm can manage numerical values, large data quantities, and datasets with missing values. The C4.5 algorithm uses a threshold value to divide the data into two ranges. The threshold value is selected to provide the most information from the raw data and is determined by sorting the attributes and selecting the average value of the attributes. C4.5 takes the training data and generates a single tree. It can work with continuous and categorical data and missing values. It also goes back over the tree to delete nodes or modify the internal structure.

Random forest: in this algorithm, the classification process uses more than one "tree". Each tree produces a classifier, and these classifiers vote to determine the algorithm that gets the most votes. This classification algorithm is then used to classify the dataset.

Advantage:

Random forests present estimates for variable importance, i.e., neural nets. They also offer a superior method for working with missing data. Missing values are substituted by the variable appearing the most in a particular node. Among all the available classification methods, random forests provide the highest accuracy.

The random forest technique can also handle big data with numerous variables running into thousands. It can automatically balance data sets when a class is more infrequent than other classes in the data. The method also handles variables fast, making it suitable for complicated tasks.

A decision tree is built on an entire dataset, using all the features/variables of interest, whereas a random forest randomly selects observations/rows and specific features/variables to build multiple decision trees from and then averages the results. After a large number of trees are built using this method, each tree "votes" or chooses the class, and the class receiving the most votes by a simple majority is the "winner" or predicted class.

Advantage:

A significant advantage of a decision tree is that it forces the consideration of all possible outcomes of a decision and traces each path to a conclusion. It creates a comprehensive analysis of the consequences along each branch and identifies decision nodes that need further analysis.

Compare these algorithms with the data, to achieve higher accuracy, the random forest may be the most suitable algorithm, but I will test a run for the rest algorithms too.

## 6.2 Select data-mining algorithms based on discussion

When to use to decision tree:

* When you want your model to be simple and explainable
* When you want a non-parametric model
* When you don't want to worry about feature selection or regularization or worry about multi-collinearity.
* You can overfit the tree and build a model if you are sure of validation or test data set is going to be a subset of the training data set or almost overlapping instead of unexpected.

When to use a random forest :

* When you don't bother much about interpreting the model but want better accuracy.
* The random forest will reduce the variance part of the error rather than the bias part, so on a given training data set decision tree may be more accurate than the random forest. But on an unexpected validation data set, Random forest always wins in terms of accuracy.

Based on the pros and cons of the methods, in this project, the higher accuracy is better, so I will use the random forest as the main algorithm, and I will also try the other algorithms.

In PySpark, the random forest algorithm has two related functions, RandoForestClassifier and RandomForestRegressor. As discussed above, I will do both, and also I will use a decision tree algorithm as DecisionTreeClassifier and DecisionTreeRegressor.

## 6.3 Build/Select appropriate model(s) and choose relevant parameter(s)

In the random forest algorithm, the main parameters are numTrees and numClasses, and maxDepth. As I group all features into one, therefore the class is binary classifier. And I will set bumTrees and maxDepth to default first and run it again with numTrees=100 and maxDepth=10 (default=4).

For the decision tree algorithms, the parameters of DecisionTreeRegressor and DecisionTreeClassifier are basically the same too, but what is different from the random forest is that for the decision tree algorithm, we don't need to set numTrees. Only change maxDepth to 10.

# 7. Data Mining

## 7.1 Create and justify test designs

In this iteration, the Jupyter notebook or the AWS serves has some problem, and to avoid these issues, I first output the cleaned datasets from the pervious notebook, then read it in a new one to reduce the workload in a single file.

In this study, the objective is to find out the relationship, and the statistic graphs and random forest algorithm result shows the linear correlation. That's the most strong evidence to prove that the success criteria of the objects have been met.

The reason is that when the dataset is split into train and test sets, there will not be enough data in the training dataset for the model to learn an effective mapping of inputs to outputs. There will also not be enough data in the test set to effectively evaluate the model performance.

To test the algorithm accuracy, I use partition node and divide 70% dataset as training data and 30 % as test data. And the reason for choosing a 70-30 ratio is to avoid using too many train data to cause the overfitting and also providing enough data to train the model.

The code to split the train and test dataset is shown in figure 33, result in figure 34.



(Figure 31: code of split train-test data)

Graphical user interface, text, application

Description automatically generated(Figure 32: code and result of train-test dataset count)

## 7.2 Conduct data mining

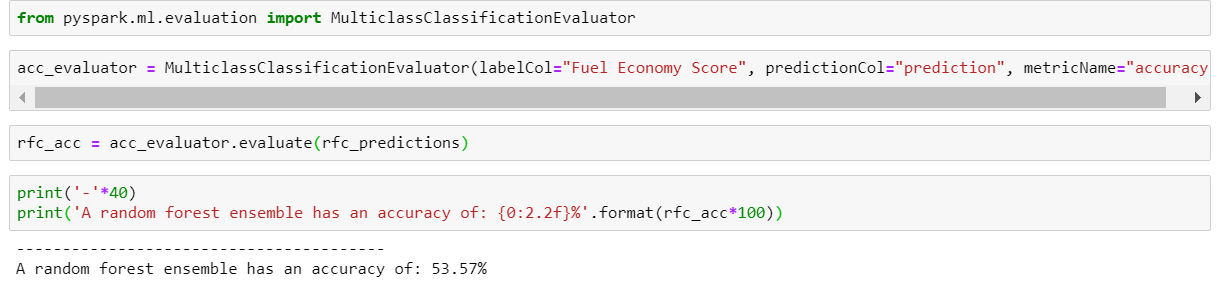
After splitting the train-test dataset, I started to work on mining the data. The code of four algorithms functions is shown in figure 35. The function I used to measure the predict correctness is the function binary classification evaluator. And the best situation is 1, the more close to 1, the higher correctness the model has.

I will use three different algorithms, so I used defaults to make the comparison "fair".



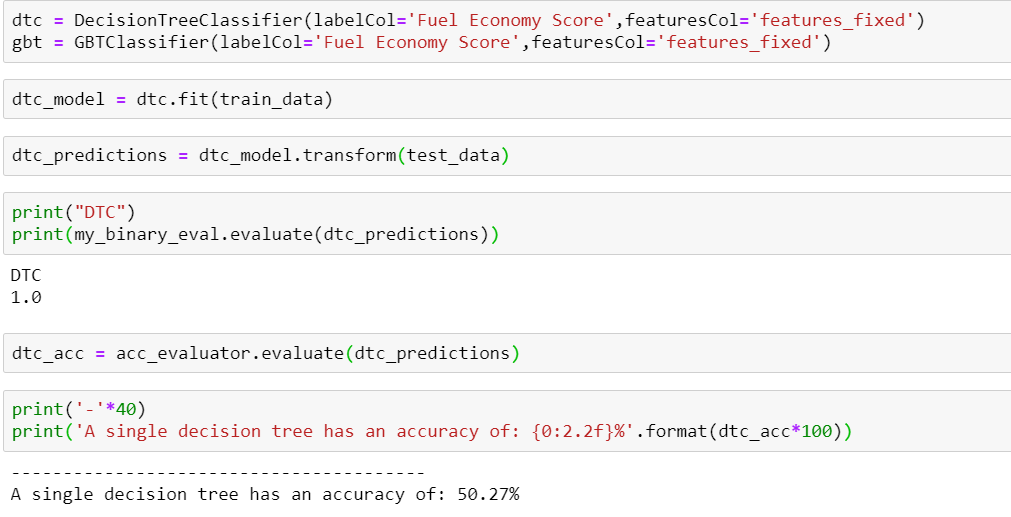
(Figure 33: Random forest data mining)

Then I set up multi-class classification and evaluator. (figure 37)



(Figure 34: Random forest data mining code and result)

Then I did the Decision tree classifier with defaults.



(Figure 35: Decision tree data mining)

Graphical user interface, text, application, email

Description automatically generated

(Figure 36: Random forest with 100 trees)

And then, add maxDepth=10.

Graphical user interface, text, application, email

Description automatically generated

(Figure 37: Random forest with 100 trees and max depth=10)

Graphical user interface, text, application, email

Description automatically generatedThen, I set the decision tree algorithm maxDepth to 10.

(Figure 38: Decision tree with maxDepth=10)

## 7.3 Search for patterns

To show the connection between the attributes, I go back to step

2.3 data exploring

here, skip the normalization part, and follow the code in figure 10, generate the countplot graphs for the cleaned data. (figure 41-50)

Chart

Description automatically generated

(Figure 41: Fuel Economy Score-Class)Chart, histogram

Description automatically generated

(Figure 42: Fuel Economy Score-Engine Cylinders)Chart

Description automatically generated

(Figure 43: Fuel Economy Score-Engine Displacement)Chart, bar chart

Description automatically generated

(Figure 44: Fuel Economy Score-Drive)Chart

Description automatically generated

(Figure 45: Fuel Economy Score-Fuel Type)Chart, bar chart

Description automatically generated

(Figure 46: Fuel Economy Score-Gasoline/Electricity Blended) Chart, bar chart

Description automatically generated

(Figure 47: Fuel Economy Score-Supercharger)A picture containing chart

Description automatically generated

(Figure 48: Fuel Economy Score-Transmission)Chart, bar chart

Description automatically generated

(Figure 49: Fuel Economy Score-Turbocharger)Chart, bar chart

Description automatically generated

(Figure 50: Fuel Economy Score-Year)

And using the code "table.to\_excel("C:/Users/tomwa/Desktop/INFOSYS 722/Research/Iteration 3/cleaned data.xlsx")”, I output the cleaned data to a new file. Import the clean data into the Tableau.

Pattern 1:

Chart

Description automatically generated

(Figure 51: Fuel Economy Score – Engine Displacement 3)

Based on the figure 51, the engine displacement has a quadric relationship with the fuel economy score. For the cars have 10 Fuel Economy Score, their Engine Displacement is higher than these have 9 and 8 scores. But if cars continueously have higher engine displacement, their scores are getting lower.

Pattern 2:

In figure 45, it shows some interesting pattern. Cars consume diesel are actually have a higher average fuel economy score than the regular or premium gas consume car. And electricity cars are all having 10.

The figure of Fuel Type shows that the car consuming electricity has a leading performance at the 10 score level, and most cars in the medium-low range score are using premium and gasoline or E85.

Pattern 3:

In figure 52, it shows the pattern that as the engine cylinders increasing, the fuel economy score decreased.

Chart, scatter chart

Description automatically generated

(Figure 52: Fuel Economy Score – Engine Cylinders)

Pattern 4:

In figure 50, as the year passed, the graphs of fuel economy score are shifting to the right slowly.

Pattern5：

The figure of Drive shows that the front-drive has more count in the high range of fuel economy score than the others.

# 8. Interpretation

## 8.1 Study and discuss the mined patterns

In the found pattern 1 and 3, there is an implication that 0 engine displacement or 0 engine cylinders mean the car is actually consuming electricity as power. So the fuel economy score at these two points is 10 for sure. Exclude the start point in both graphs of engine displacement and cylinders vs. fuel economy score; there is a small section that the engine displacement or cylinders are direct ratios to the fuel economy score. This is because, at the beginning phase, increasing displacement or cylinders is providing essential power needed. Extra displacement or cylinders beyond the critical point, are consuming fuel to do the extra unnecessary work.

In pattern 2, there are three attributes that only have one polt at the 10 score level. And all these three attributes are related to electricity. And other columns related to electricity are all having a higher average and bottom boundary of the fuel economy score. This indicates the hybrid engine or electricity engine are more environment friendly compared to the gas engine. And what's more interesting is that the diesel car having a higher fuel economy score. And my assumption for a reason why is the diesel engine has better combustion efficiency.

And for the pattern 4, the engine displacement is used to measure the size(power rate) of an engine. In describing how much power it can produce in unit time, regardless of how much fuel consumed. So higher engine displacement will make the car more powerful, and at the same time, it may consume more fuel to output the power.

## 8.2 Visualize the data, results, models, and patterns

Chart, scatter chart

Description automatically generated

(Figure 53: Fuel Economy Score - Class)Chart, scatter chart

Description automatically generated

(Figure 54: Fuel Economy Score – Engine Cylinders) Chart, scatter chart

Description automatically generated

(Figure 55: Fuel Economy Score – Engine Displacement)Chart, scatter chart

Description automatically generated

(Figure 56: Fuel Economy Score – Supercharger)A picture containing chart

Description automatically generated

(Figure 57: Fuel Economy Score – Transmission)Chart, scatter chart

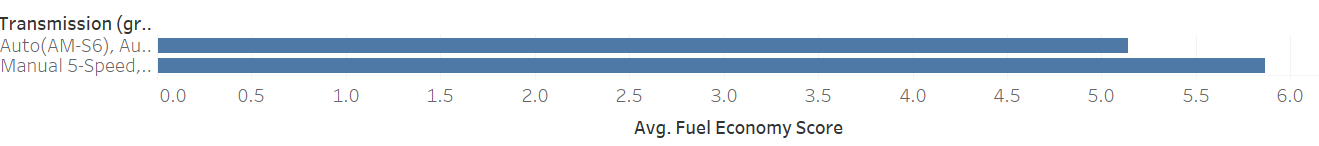
Description automatically generated

(Figure 58: Fuel Economy Score – Turbocharger)

## 8.3 Interpret the results, models, and patterns

The business objective of this study is to figure out what factors will affect the fuel consumption of a car and how closely their relationship between these factors and fuel consumption. And, identify what the fuel consumption rate of the car is.

Based on the patterns found, we can conclude that the fuel consumption rate will be strongly affected by engine displacement. Also, the engine cylinders, fuel type, Drive, Transmission, and even years can affect fuel consumption rate of a car. For example, 4-wheel Drive is less fuel economical, and automatical Transmission can consume more fuel. (as shown in figure below)



(figure 59: grouped transmission)

Graphical user interface, application

Description automatically generated(figure 60: grouped drive)

The other goal is to identify the fuel consumption rate, which is to classify the fuel economy score in this study. The success criteria are 80% of accuracy, but as a result, only 69% can be achieved based on the current dataset. To achieve the success criteria, more data should be collected, and more time is needed to do the deeper data mining.

## 8.4 Assess and evaluate results, models, and patterns

From the 3 iterations of data mining by using different parameters of Random Forest algorithm, the accuracy of multi-classes classification evaluator did increase a lot (53.57% - 53.95% - 71.01%). But there also a interested point that no matter what parameters set, or what algorithm used, the binary classification evaluator is always 1, which means no error. And with the parameters sets, the random forest tree algorithm is performed better than decision tree algorithm, as we expected in step 6.

## 8.5 Iteration

After I finished the data cleaning, I went back to the step 2.3 data exploring, and generate the graphs again with the cleaned data. Also for the data mining steps, I run the code several iterations to implement different parameters of data mining algorithms.

Also, I did the steps 2-3 and step 8 with the new target ‘Combined MPG’(figure 59-64). Assuming the Fuel Economy Score is linear relationship with the Combined MPG, then the figures all support the patterns and the results I got above.

Chart, line chart

Description automatically generated

(Figure 61: Combined MPG – Engine Cylinders)Chart

Description automatically generated

(Figure 62: Combined MPG – Engine Displacement)Chart, bar chart

Description automatically generated

(Figure 63: Combined MPG – Drive)Chart, bar chart

Description automatically generated

(Figure 64: Combined MPG – Fuel Type)Chart

Description automatically generated

(Figure 65: Combined MPG – Transmission)Chart, bar chart

Description automatically generated

(Figure 66: Combined MPG – Turbocharger, Supercharger, Gasoline/Electricity Blended)