Analysis and Prediction of Carbon Dioxide Emissions for New Light-Duty Vehicles

Jiajun Bao

CSE 163 Final Project

06 June 2021

Research questions

(1) What is the relationship between engine size and carbon dioxide emissions?

Answer: There is a strong correlation between engine size and carbon dioxide emissions. As the engine size(L) increases, the CO2 emissions(g/km) also increases, and the relationship appears to be linear. Please refer to the Result Part for the detailed explanation.

(2) What is the relationship between number of cylinders and carbon dioxide emissions?

Answer: There is a strong correlation between the number of cylinders and CO2 emissions(g/km). The carbon dioxide emissions increase as the number of cylinders of the cars increases. Besides, there are more variabilities of CO2 emissions within the groups of cars that have 4, 6, 8 cylinders than groups of cars with 3 cylinders and 10 cylinders. Overall, there is a positive relationship between the number of cylinders and carbon dioxide emissions. Please refer to the Result Part for the detailed explanation.

(3) What is the relationship between fuel consumption(city, highway, combined) and carbon dioxide emissions?

Answer: There is a strong correlation between carbon dioxide emissions and fuel consumption for all three types of fuel consumption(city, highway, combined). The relationship appears to be positive and linear. As fuel consumption increases, the carbon dioxide emissions also increase. Please refer to the Result Part for the detailed explanation.

(4) What is the relationship between vehicle classification and carbon dioxide emissions?

Answer: There is not a precise correlation between vehicle classification and carbon dioxide emissions. However, if we summarize characteristics of groups, then compact, mid-size, small SUVs, and subcompact tend to have lower CO2 emissions. On the other hand, full-size, standard pickup trucks, standard SUVs, two-seaters, and vans tend to have higher CO2 emissions. Please refer to the Result Part for the detailed explanation.

(5) What type of vehicle classification, engine size, number of cylinders, and fuel consumption should people look for in order to reduce carbon dioxide emissions?

Answer: Suppose people are looking to buy cars that help with carbon dioxide emissions reduction. In that case, they should look for vehicles with small engine sizes, few cylinders, and low fuel consumption(city, highway, combined). Depending on the consumer needs, compact, mid-size, small SUV, and subcompact are all good choices in terms of choosing vehicle class.

(6) If the car manufactures provides information about features of a new releasing car and its carbon dioxide emissions, can we predict carbon dioxide emissions and crosscheck the credibility of the provided information from the car manufacturers?

Answer: Based on the result from the machine learning implementation, if we are provided with all the 11 features of a car[1], we can predict the CO2 emission with an R squared score over 0.96 on average. However, sometimes it is hard to collect all the features, so we also explore which features are the most informative for how a decision is made based on the prediction accuracy of each feature. Then, We found that using using just three Fuel Consumption features (City, Hwy, Combo) can also provide insight on verification of numbers from manufacturers with an R squared score over 0.93 on average. Please refer to the Result Part for the detailed explanation.

Motivation and Background

Global warming endangers people's health, threatens the ecosystem on the earth, and negatively impacts the environment, such as rising seas and increasing temperature. Our personal vehicles are one of the primary causes of global warming. Collectively, cars and trucks account for nearly one-fifth of all US emissions, emitting around 24 pounds of carbon dioxide and other global-warming gases for every gallon of gas. About five pounds comes from the extraction, production, and delivery of the fuel, while the great bulk of heat-trapping emissions—more than 19 pounds per gallon—comes right out of a car's tailpipe.[2]

Although electric vehicles are getting more and more popular and ease the situation, they currently make up only 3% of car sales worldwide.[3] So it is important to let both car manufacturers and consumers get a better picture of what type of car will help reducing carbon dioxide emissions to help with global warming. Only with combined effort can we solve the problem of global warming, so car manufacturers should take the responsibility to design cars that fight global warming, and consumers should also consider carbon dioxide emissions as one of the factors when buying cars.

In this project, I will explore four factors: engine size, numbers of cylinders, fuel consumption(city, highway, combined) and vehicle class that potentially affect carbon dioxide emissions of automobiles. Based on the finding, I will generate a list of specifications of cars people should look at if they want to buy new light-duty vehicles that protect our environment and provide some factors for car manufacturers should consider when they design cars to fight global warming. Additionally, I will utilize machining learning algorithms to predict the carbon dioxide emission with given information for future car release to confirm the manufacturer's statistics and explore on what features are most informative in terms of making decision in machine learning models.

[1] Brand, Vehicle Class, Engine Size(L), Cylinders, Transmission, Fuel Type, Fuel Consumption City(L/100 km), Fuel Consumption Hwy(L/100 km), Fuel Consumption Comb(L/100 km), CO2 Rating, Smog Rating.

[2] https://www.ucsusa.org/resources/car-emissions-global-warming

[3] https://policyadvice.net/insurance/insights/electric-car-statistics/

Dataset

The dataset I use for this project is from the Government of Canada official website, and the publisher is Natural Resources Canada. The datasets provide model-specific fuel consumption ratings and estimated carbon dioxide emissions for new light-duty vehicles for retail sale in Canada from 1995 to 2021 model year vehicles. I choose to use this dataset because it is from the official government website. And according to the website, the data was collected using standard, controlled laboratory testing and analytical procedures, which ensures the consistency and accuracy of data among all car models. The dataset I used is the most recent dataset - 2021 Fuel Consumption Ratings (2021-04-30).

- Link to the website that contains the dataset:
https://open.canada.ca/data/en/dataset/98f1a129-f628-4ce4-b24d-6f16bf24dd64#wb-auto-6
- Link to dataset:
https://www.nrcan.gc.ca/sites/nrcan/files/oee/files/csv/MY2021%20Fuel%20Consumption%20Ratings.csv

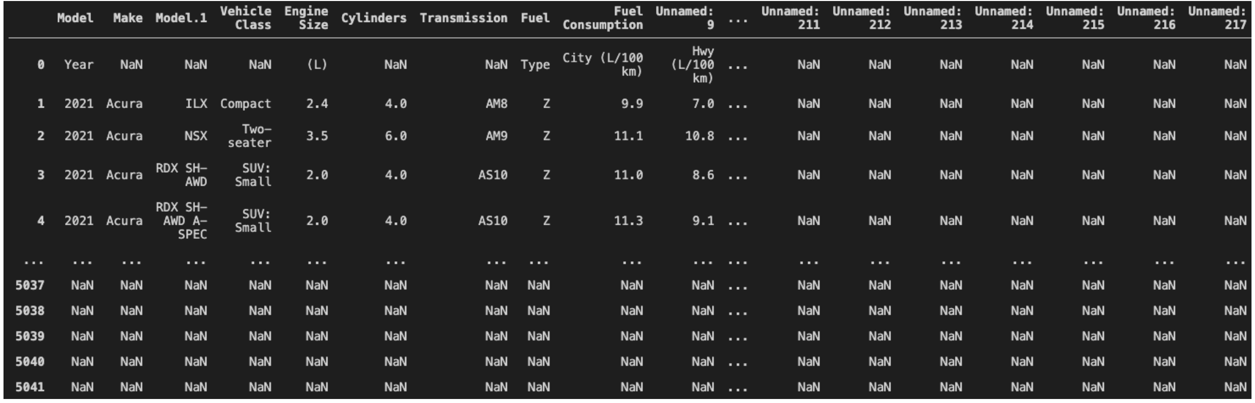
Method

1) Preparing the Data

Goal: This will prepare the dataset into the format we need for later plotting and machine learning algorithm analysis to answer the research questions

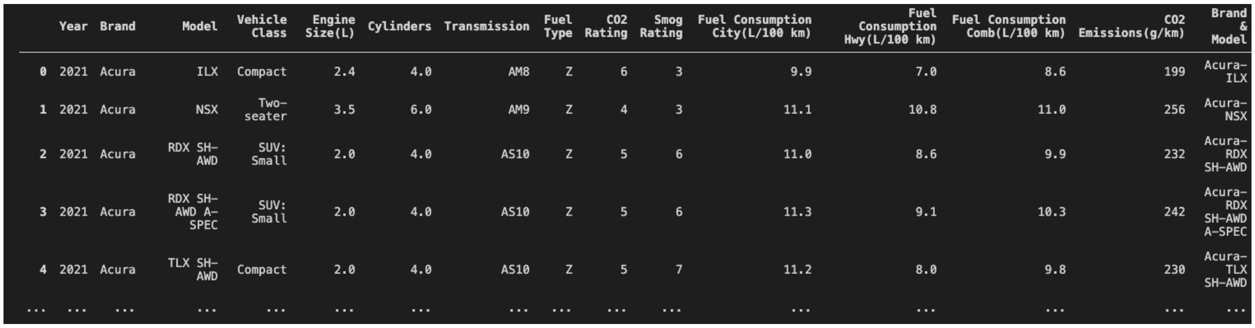
- load the data in the format of cvs and save it to a pandas data frame
- combine the first two rows representing labels to one row for later plotting purpose
- adjust column names for better description purpose after combing first two rows
- create and run test file to check whether the filtered data meet the expected number of rows and columns

Raw dataset before cleaning:



	Model	Make	Model.1	Vehicle Class	Engine Size	Cylinders	Transmission	Fuel	Fuel Consumption	Unnamed: 9	...	Unnamed: 211	Unnamed: 212	Unnamed: 213	Unnamed: 214	Unnamed: 215	Unnamed: 216	Unnamed: 217
0	Year	NaN	NaN	NaN	(L)	NaN	NaN	Type	City (L/100 km)	Hwy (L/100 km)	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	2021	Acura	ILX	Compact	2.4	4.0	AM8	Z	9.9	7.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	2021	Acura	NSX	Two-seater	3.5	6.0	AM9	Z	11.1	10.8	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	2021	Acura	RDX SH-AWD	SUV: Small	2.0	4.0	AS10	Z	11.0	8.6	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	2021	Acura	RDX SH-AWD A-SPEC	SUV: Small	2.0	4.0	AS10	Z	11.3	9.1	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5037	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5038	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5039	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5040	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5041	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Dataset after clearing:



	Year	Brand	Model	Vehicle Class	Engine Size(L)	Cylinders	Transmission	Fuel Type	CO2 Rating	Smog Rating	Fuel Consumption City(L/100 km)	Fuel Consumption Hwy(L/100 km)	Fuel Consumption Comb(L/100 km)	CO2 Emissions(g/km)	Brand & Model
0	2021	Acura	ILX	Compact	2.4	4.0	AM8	Z	6	3	9.9	7.0	8.6	199	Acura-ILX
1	2021	Acura	NSX	Two-seater	3.5	6.0	AM9	Z	4	3	11.1	10.8	11.0	256	Acura-NSX
2	2021	Acura	RDX SH-AWD	SUV: Small	2.0	4.0	AS10	Z	5	6	11.0	8.6	9.9	232	Acura-RDX SH-AWD
3	2021	Acura	RDX SH-AWD A-SPEC	SUV: Small	2.0	4.0	AS10	Z	5	6	11.3	9.1	10.3	242	Acura-RDX SH-AWD A-SPEC
4	2021	Acura	TLX SH-AWD	Compact	2.0	4.0	AS10	Z	5	7	11.2	8.0	9.8	230	Acura-TLX SH-AWD
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

2) Plotting for Different Independent Variables & Carbon Dioxide Emissions

- utilize the library Altair to create interactive scatter plots for:
    - engine size vs. carbon dioxide emissions
    - numbers of cylinders vs. carbon dioxide emissions
    - fuel consumption(city/highway/combined) vs. carbon dioxide emissions
    - vehicle classification vs. carbon dioxide emissions

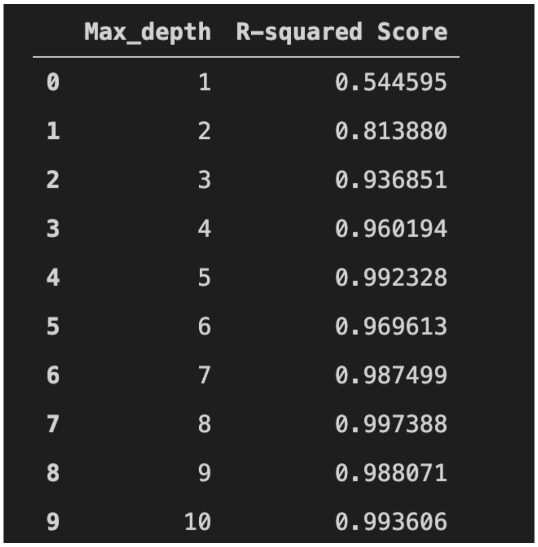
Please refer to the later Result part for the plots and later Challenge goals part on the interactive feature of the scatterplot

3) Develop Machine Learning Model

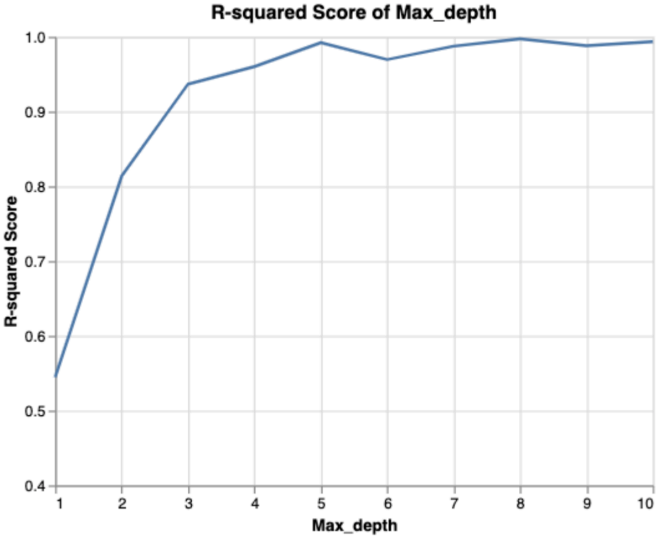
Part(a)

Part(a) Goal: Create the machine learning model to predict the carbon dioxide emission with given information for future car release to confirm the manufacturer's statistics.

- construct the general model with all the given features: Brand, Vehicle Class, Engine Size(L), Cylinders, Transmission, Fuel Type, Fuel Consumption City(L/100 km), Fuel Consumption Hwy(L/100 km), Fuel Consumption Comb(L/100 km), CO2 Emissions(g/km), CO2 Rating, Smog Rating and separate CO2 Emissions(g/km) for the label. And we do not take the year column into consideration as the data is all from 2021.
- perform a one-hot encoding on variables such as brand, fuel type, vehicle class, and transmission to transform the categorical features in order to use DecisionTreeRegressor
- The decision tree Max_depth is decided (Max_depth=5) after plotting the relationship between r-squared score of the mode and the decision tree Max_depth.



	Max_depth	R-squared Score
0	1	0.544595
1	2	0.813880
2	3	0.936851
3	4	0.960194
4	5	0.992328
5	6	0.969613
6	7	0.987499
7	8	0.997388
8	9	0.988071
9	10	0.993606



- Assess its accuracy using a random split to break the dataset up randomly into a training set and test set (20% of the rows to be the test set) with two examining standard: R-squared Score and mean-squared error

Please refer to the later Result part for the evaluation of the general model

Part(b)

Part(b) Goal: Discover which features are the most informative for how a decision is made and explore on models with different combinations of features

- plot and interpret which features are the most informative for how a decision is made based on the R-squared Score and mean-squared error
- explore on combination of features: combination one that include just three most informative features based on previous step and combination two that include every other features except the three most informative ones. Then compare and contract two different models performance

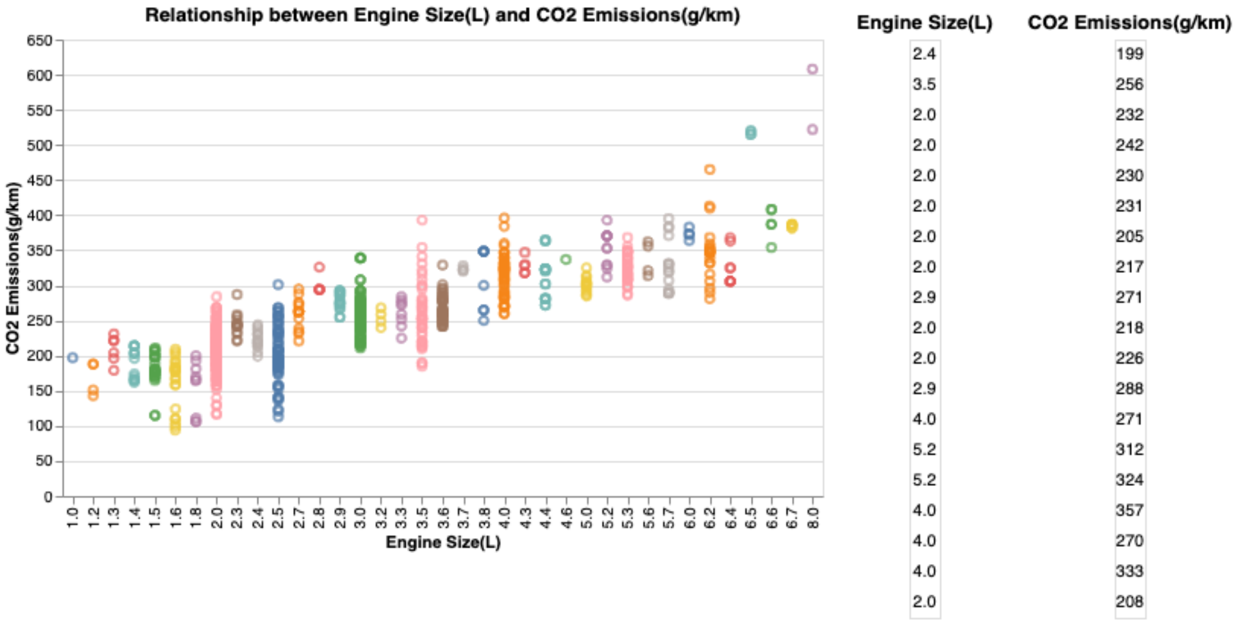
Please refer to the later Result part for the evaluation of each feature and model along with plots

Results

The research breaks down into 6 different questions:

(1) What is the relationship between engine size and carbon dioxide emissions?

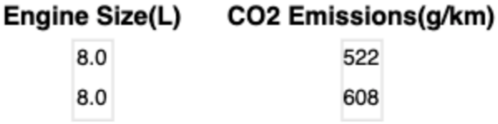
Since we have over 500 car data from the dataset, so it will be hard to explore the relationship by looking at the pandas data frame table. So I choose to plot the relationship between engine size and carbon dioxide emissions:



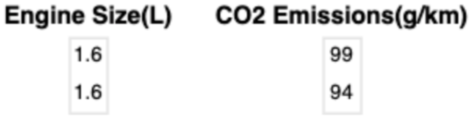
Engine Size(L)	CO2 Emissions(g/km)
2.4	199
3.5	256
2.0	232
2.0	242
2.0	230
2.0	231
2.0	205
2.0	217
2.9	271
2.0	218
2.0	226
2.9	288
4.0	271
5.2	312
5.2	324
4.0	357
4.0	270
4.0	333
2.0	208

Plot-1 Relationship between Engine Size and Carbon Dioxide Emissions

After visualizing the data, we can see that there is a strong linear correlation between engine size and carbon dioxide emissions. As the engine size(L) increases, the CO2 emissions(g/km) also increases. Since we incorporate the interactive feature to the plot, we can click on the data point to see the CO2 emission and their car model name. We extract the information of two cars from the dataset that have the most CO2 emission per kilometer and two vehicles from the dataset that have the least CO2 emission per kilometer:



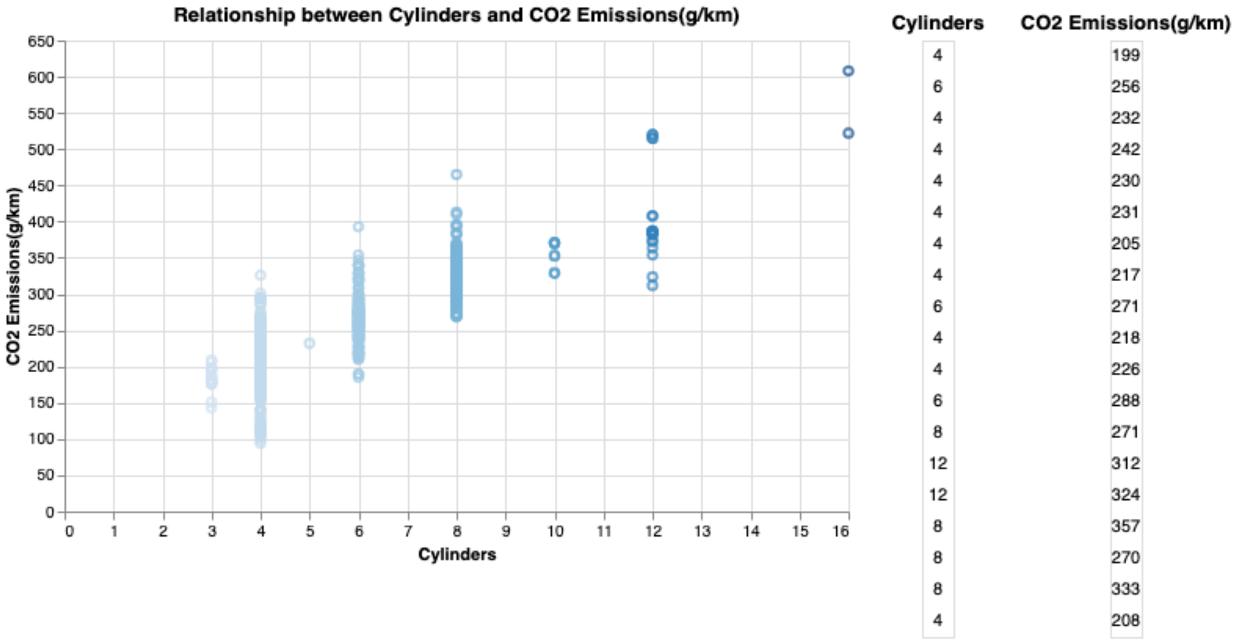
Engine Size(L)	CO2 Emissions(g/km)
8.0	522
8.0	608



Engine Size(L)	CO2 Emissions(g/km)
1.6	99
1.6	94

The two cars *Bugatti Chiron* and *Bugatti Chiron Pur Sport*, with engine size of 8.0L, have about 5 times CO2 emission per kilogram of the two vehicles with the least CO2 emission in the dataset *Hyundai IONIQ Blue* and *Hyundai IONIQ.* We can see that car manufacturers of supercars such as Bugatti and Lamborghini design cars that have extremely high performance with the cost of excess CO2 emissions. Such car manufacturers should take more responsibility to help with global warming and consider environmental factors rather than pursuing high performance that are powered by combustion engines. And with more and more people choosing electric vehicles for commutation, electric supercars are also rising nowadays. They can achieve similar performance as those exotic cars powered by gasoline, but more beneficial to our environment. The government and consumers should support the electric car industries. For example, the motorsport organizations across world could set up more electric racing events that use only electric cars or hybrid cars. And governments should come up with more beneficial policies such as tax deductions for promoting the electric car market.

(2) What is the relationship between number of cylinders and carbon dioxide emissions? Following the similar idea as question(1), we plot the relationship between number of cylinders and carbon dioxide emissions:



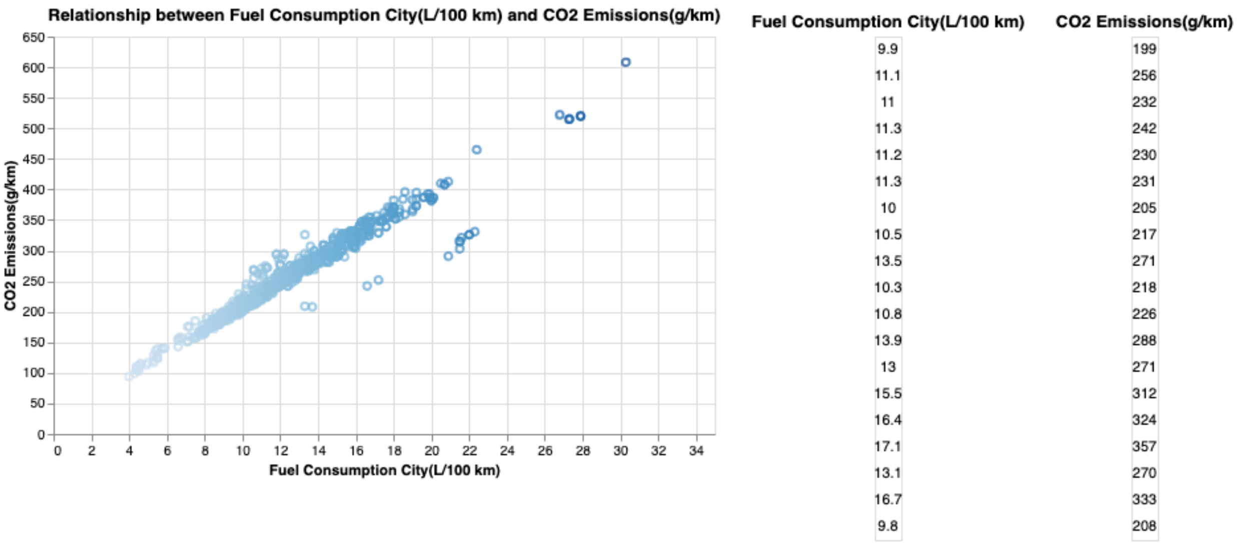
Cylinders	CO2 Emissions(g/km)
4	199
6	256
4	232
4	242
4	230
4	231
4	205
4	217
6	271
4	218
4	226
6	288
8	271
12	312
12	324
8	357
8	270
8	333
4	208

Plot-2 Relationship between number of Cylinders and carbon Dioxide Emissions

Based on the plot-2 above, we can wee that there is a strong linear correlation between the number of cylinders and CO2 emissions(g/km). The carbon dioxide emissions increase as the number of cylinders of the cars increases. Besides, there are more variabilities of CO2 emissions within the groups of cars that have 4, 6, 8 cylinders than groups of cars with 3 cylinders and 10 cylinders. Overall, there is a positive relationship between the number of cylinders and carbon dioxide emissions.

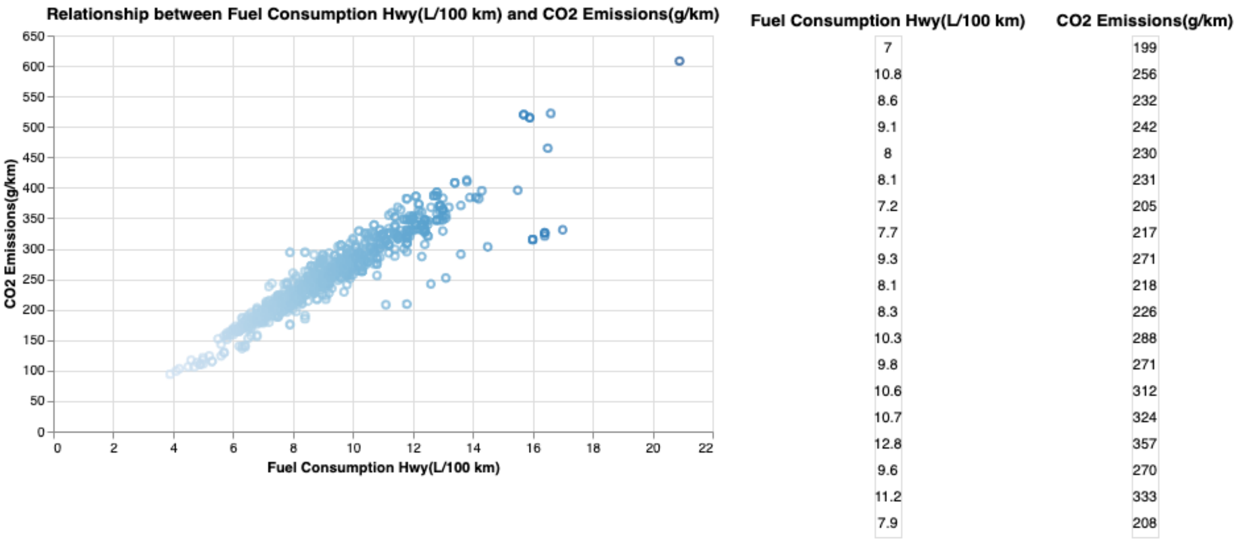
(3) What is the relationship between fuel consumption(city, highway, combined) and carbon dioxide emissions?

Following the similar idea as question(1), we plot the relationship between fuel consumption(city, highway, combined) and carbon dioxide emissions:



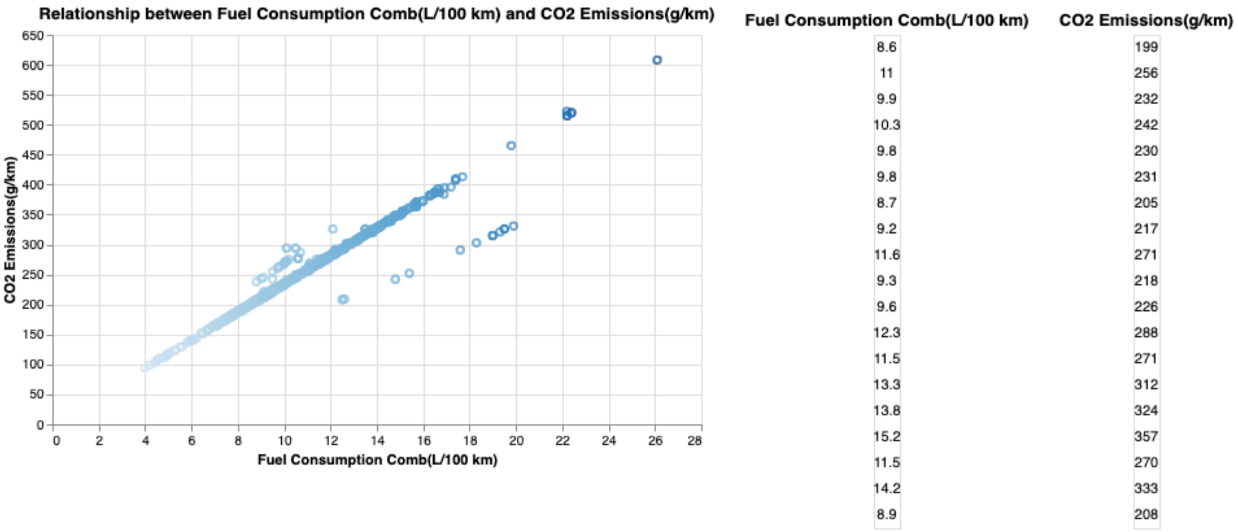
Fuel Consumption City(L/100 km)	CO2 Emissions(g/km)
9.9	199
11.1	256
11	232
11.3	242
11.2	230
11.3	231
10	205
10.5	217
13.5	271
10.3	218
10.8	226
13.9	288
13	271
15.5	312
16.4	324
17.1	357
13.1	270
16.7	333
9.8	208

Plot-3(a) Relationship between City Fuel Consumption and Carbon Dioxide Emissions



Fuel Consumption Hwy(L/100 km)	CO2 Emissions(g/km)
7	199
10.8	256
8.6	232
9.1	242
8	230
8.1	231
7.2	205
7.7	217
9.3	271
8.1	218
8.3	226
10.3	288
9.8	271
10.6	312
10.7	324
12.8	357
9.6	270
11.2	333
7.9	208

Plot-3(b) Relationship between Highway Fuel Consumption and Carbon Dioxide Emissions

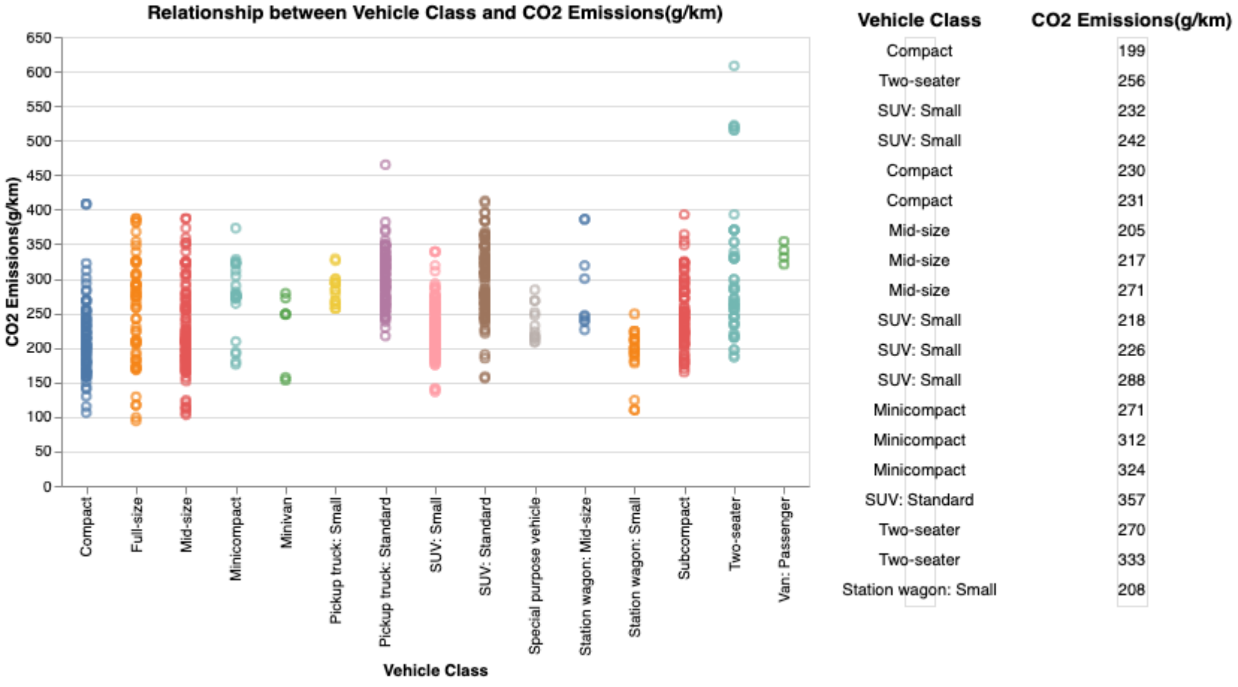


Fuel Consumption Comb(L/100 km)	CO2 Emissions(g/km)
8.6	199
11	256
9.9	232
10.3	242
9.8	230
9.8	231
8.7	205
9.2	217
11.6	271
9.3	218
9.6	226
12.3	288
11.5	271
13.3	312
13.8	324
15.2	357
11.5	270
14.2	333
8.9	208

Plot-3(c) Relationship between Fuel Consumption Combination and Carbon Dioxide Emissions

Based on plot3 (a)(b)(c), we can see that there is a strong correlation between carbon dioxide emissions and fuel consumption for all three types of fuel consumption (city, highway, combined). The relationship appears to be positive and linear. As fuel consumption increases, the carbon dioxide emissions also increase, which is reasonable since the more fuel a car spends per kilogram, the more carbon dioxide emissions produced from the exhaust.

(4) What is the relationship between vehicle classification and carbon dioxide emissions? Following the similar idea as question(1), we plot the relationship between vehicle classification and carbon dioxide emissions:



Vehicle Class	CO2 Emissions(g/km)
Compact	199
Two-seater	256
SUV: Small	232
SUV: Small	242
Compact	230
Compact	231
Mid-size	205
Mid-size	217
Mid-size	271
SUV: Small	218
SUV: Small	226
SUV: Small	288
Minicompact	271
Minicompact	312
Minicompact	324
SUV: Standard	357
Two-seater	270
Two-seater	333
Station wagon: Small	208

Plot-4 Relationship between Vehicle Classification and Carbon Dioxide Emissions

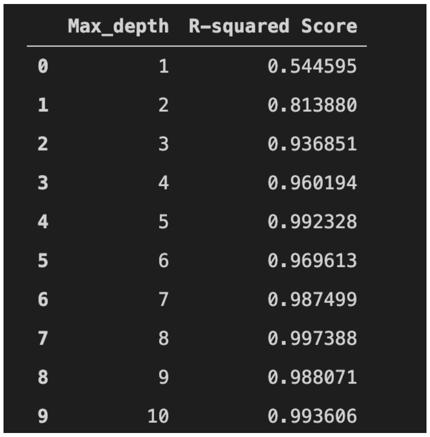
Based on plot-4, we can see that there is not a precise correlation between vehicle classification and carbon dioxide emissions. However, if we summarize the characteristics of groups by the mean of each group, then compact, mid-size, small SUVs, and subcompact tend to have lower CO2 emissions. On the other hand, full-size, standard pickup trucks, standard SUVs, two-seaters, and vans tend to have higher CO2 emissions on average. Since CO2 emissions are spread out among and within the groups, we cannot explain the relationship between vehicle classification and CO2 emission linearly like in previous questions. This is reasonable because we cannot say much detail about a car if we only know its vehicle classification. For example, a full-size SUV could be a hybrid Toyota Highlander(158 g/km) with less CO2 emission than a MiniCooper Convertible(176 g/km); or a full-size SIV could be Jeep Grand Cherokee Track-hawk that have a 410 g/km CO2 emission. Therefore, vehicle classification is not an ideal indicator of CO2 emissions compare with the previous three featured we analyzed.

(5) What type of vehicle classification, engine size, number of cylinders, and fuel consumption should people look for in order to reduce carbon dioxide emissions?

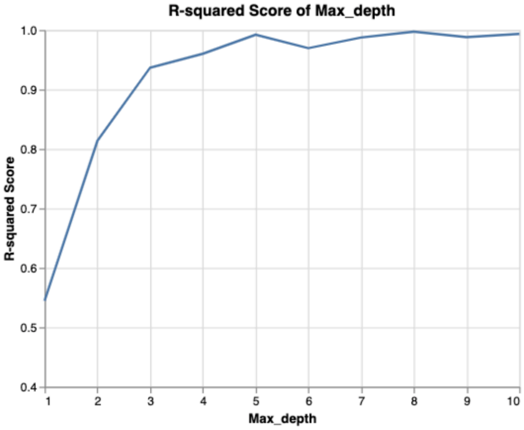
Based on the finding from question (1)-(4), suppose people are looking to buy cars that help with carbon dioxide emissions reduction. In that case, they should look for vehicles with small engine sizes, few cylinders, and low fuel consumption(city, highway, combined). Depending on the consumer needs, compact, mid-size, small SUV, and subcompact are all good choices in terms of choosing light-duty vehicle class.

(6) If the car manufactures provides information about features of a new releasing car and its carbon dioxide emissions, can we predict carbon dioxide emissions and crosscheck the credibility of the provided information from the car manufacturers?

Following the previous steps from Part Method-Develop Machine Learning Model, we first create the general machine learning model that utilize all the features given.[4] to predict the carbon dioxide emission for future car release to confirm the manufacturer's statistics. After about 20 simulations, we choose max_depth = 5 here as on average, it gives us a good R-squared and the performance doesn't improve significantly with high max_depth.

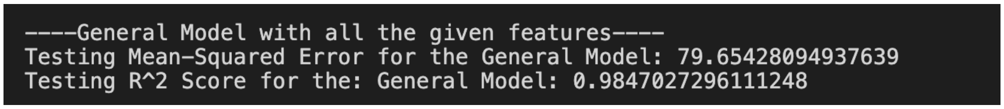


```
   Max_depth  R-squared Score
0          1         0.544595
1          2         0.813880
2          3         0.936851
3          4         0.960194
4          5         0.992328
5          6         0.969613
6          7         0.987499
7          8         0.997388
8          9         0.988071
9         10         0.993606
```



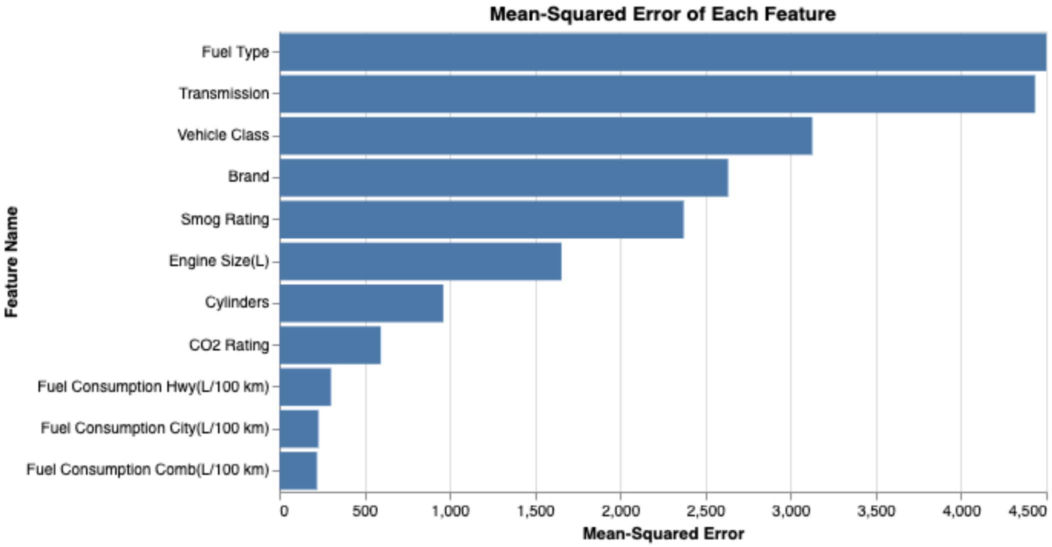
[4] With all the given features: Brand, Vehicle Class, Engine Size(L), Cylinders, Transmission, Fuel Type, Fuel Consumption City(L/100 km), Fuel Consumption Hwy(L/100 km), Fuel Consumption Comb(L/100 km), CO2 Rating, Smog Rating.

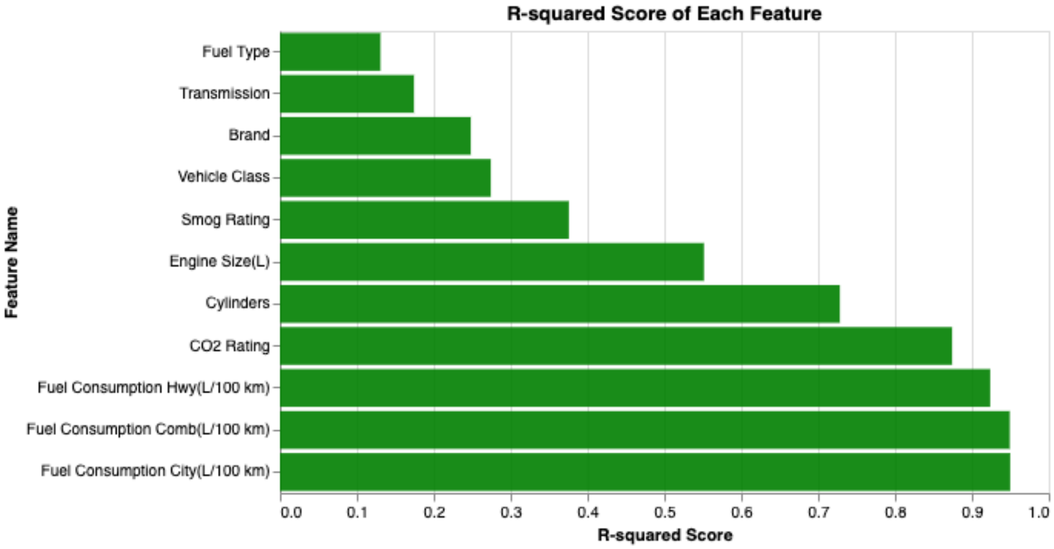
Then we assess its accuracy using a random split to break the dataset up randomly into a training set and test set (20% of the rows to be the test set) with two examining standards: R-squared Score and mean-squared error. We choose R-squared Score because it represents the proportion of the variance for a dependent variable explained by an independent variable or variables in a regression model, so the higher the R squared(closer to 1), our model is more accurate. We also choose another way to evaluate the performance of decision tree regressor models by mean-squared error, which measures the average of the squares of the errors. The smaller the mean-squared error, the more accurate our model is. And we get:



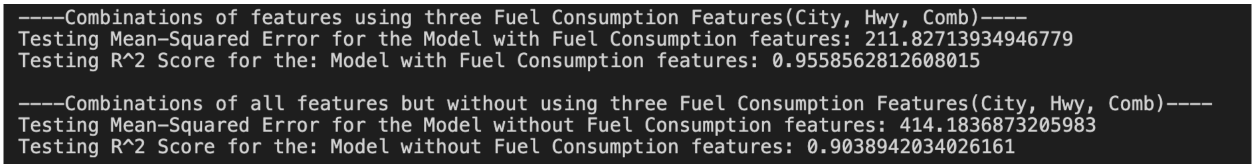
```
----General Model with all the given features----
Testing Mean-Squared Error for the General Model: 79.65428094937639
Testing R^2 Score for the: General Model: 0.9847027296111248
```

Therefore, based on the result from the machine learning implementation, if we are provided with all the 11 features of a car, we can predict the CO2 emission with an R squared score over 0.96 on average. However, sometimes it is hard to collect all the features, so we also explore which features are the most informative for how a decision is made based on the prediction accuracy of models of each indicia feature:





Based on the plots above, we can see that, on average, Fuel Consumption features (City, Hwy, Comb) are the three most informative features. So then, we build a model using just three Fuel Consumption features (City, Hwy, Comb) , and its R squared score is over 0.93 on average. Additionally, if we choose to use all the given features that include every other feature except the three most informative ones(Brand, Vehicle Class, Engine Size(L), Cylinders, Transmission, Fuel Type, CO2 Rating, Smog Rating) to build another model, we find the first model that only use three most informative features tends to outperform the second model on average. So we can see that the three fuel consumption features play important roles in predicting the CO2 emissions.



```
----Combinations of features using three Fuel Consumption Features(City, Hwy, Comb)----
Testing Mean-Squared Error for the Model with Fuel Consumption features: 211.82713934946779
Testing R^2 Score for the: Model with Fuel Consumption features: 0.9558562812608015

----Combinations of all features but without using three Fuel Consumption Features(City, Hwy, Comb)----
Testing Mean-Squared Error for the Model without Fuel Consumption features: 414.1836873205983
Testing R^2 Score for the: Model without Fuel Consumption features: 0.9038942034026161
```

Challenge goals

(1) Machine learning

In this project, I applied machine learning methods to gain insights into the data we have and predict carbon dioxide emissions for car releases. Interpreting which features are the most informative for making a decision is a challenging goal for me.

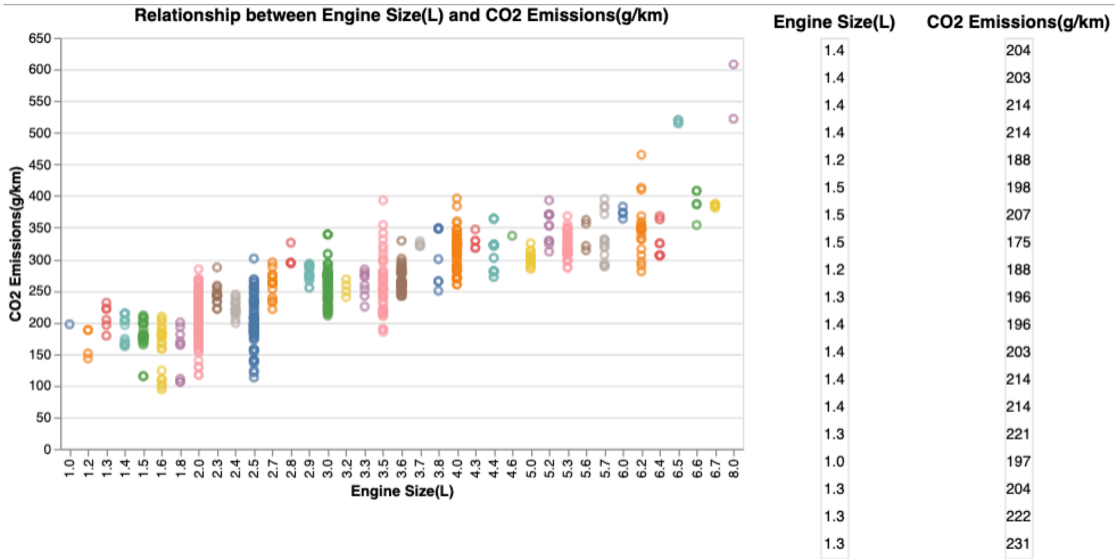
Before interpreting the most effective feature, I created the general model based on all the available features, including make/car brand, vehicle class, engine size, numbers of cylinders, fuel consumption, transmission, fuel type, and fuel consumption. This can be useful for people to check the credibility of the manufacturer's information about carbon dioxide emission of the new releasing car. Then one challenging goal is to determine what max-depth for the decision tree regressor I should use, and I decide to use max_depth = 5 after plotting the relationship between the testing R^2 Score.

Then to assess if I meet the expedition prediction, I use a random split to break the dataset up randomly into a training set and test set (20% of the rows to be the test set). Secondly, the challenging part is to explore which feature is the most informative. To do this, I separated different features of cars into different models to predict the CO2 emissions, i.e., Model #1 for car's brand, Model #2 for vehicle class, Model #3 for engine size, Model #4 for fuel type, etc. Based on the prediction accuracy, I concluded that important features for CO2 emission prediction are. Besides, I also explore if we could utilize fewer combinations of features to achieve similar testing accuracy than the general model that uses all the available features from the dataset. Then I was able to use only the most three informative features to predict the CO2 emission with only about 5% less compared with the general model that use all the features. Please refer to the Result Part for the plots comparing different features and conclusions on which features are the most informative. Therefore, I created new models and analyzed how different feature sets affect the model's performance, and this challenging goal has been achieved.

(2) External library

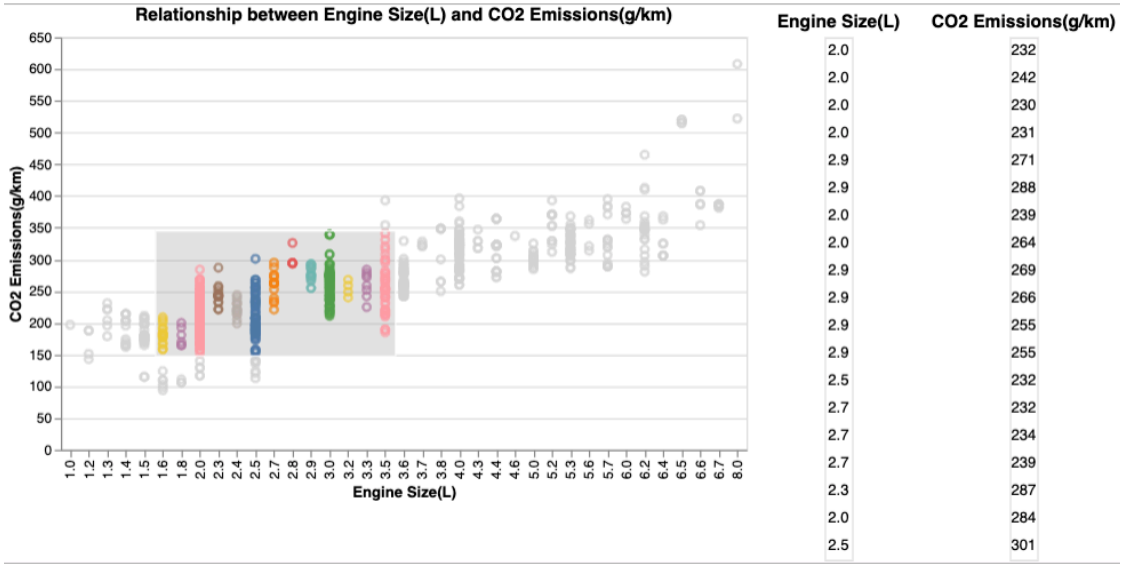
In this project, one of the challenging goals for me is to learn a new library(Altair) and utilize the library to produce concise and informative graphical representations of over 800 car models. Based on the previous result parts, I have successfully created four interactive plots according to the four factors I am interested in (engine size, numbers of cylinders, fuel consumption, and vehicle classification) and their relationship to CO2 emissions. The most challenging part was incorporating the interactive features of the plots that allow people who view this code from Jupyter Notebook to select a specific region of datapoint to see the accurate data and move the mouse pointer to any given point to check the brand and mode info of the given car. The following demonstrations are screenshots of interactive plots. Therefore, the plots with desired functions have been created, and this challenging goal has been achieved. Please refer to the Jupyter Notebook for the interactive process.

Demo: General trend with no selection or mouse pointer selection



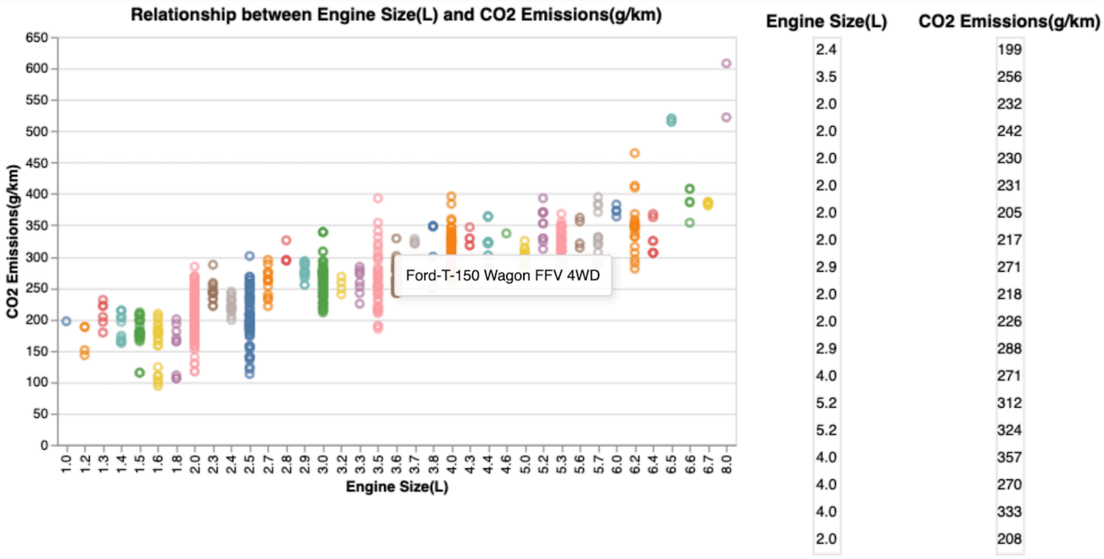
Engine Size(L)	CO2 Emissions(g/km)
1.4	204
1.4	203
1.4	214
1.4	214
1.2	188
1.5	198
1.5	207
1.5	175
1.2	188
1.3	196
1.4	196
1.4	203
1.4	214
1.4	214
1.3	221
1.0	197
1.3	204
1.3	222
1.3	231

Demo: General trend with selection:



Engine Size(L)	CO2 Emissions(g/km)
2.0	232
2.0	242
2.0	230
2.0	231
2.9	271
2.9	288
2.0	239
2.0	264
2.9	269
2.9	266
2.9	255
2.9	255
2.5	232
2.7	232
2.7	234
2.7	239
2.3	287
2.0	284
2.5	301

Demo: General trend with with mouse pointer selection:



Engine Size(L)	CO2 Emissions(g/km)
2.4	199
3.5	256
2.0	232
2.0	242
2.0	230
2.0	231
2.0	205
2.0	217
2.9	271
2.0	218
2.0	226
2.9	288
4.0	271
5.2	312
5.2	324
4.0	357
4.0	270
4.0	333
2.0	208

Work Plan Evaluation

Task 1: Preparing the Data and Cleaning data
Expected : 3 hours | Actual : 3 hours

- reformat the first two rows representing labels to one row for later plotting purpose
- update the data frame according to columns we will use for later analysis
- test the data frames sizes, content, etc meet expectations

Task 2: Plotting for Different Independent Variables & Carbon Dioxide Emissions
Expected : 8 hours | Actual : 12 hours

- learn the new library Altair and create the plots as described in Method part
- check plots meet expectations and interpret the plotting result

Task 3: Develop Machine Learning Model
Expected : 6 hours | Actual : 8 hours

- separate the data into features and labels, train different models on that data, and assess its accuracy
- plot the DecisionTreeRegressor for demonstration
- interpret which features are the most informative for how a decision is made based on the prediction accuracy of each feature feature

Task 4: Summarize all the findings
Expected : 4 hours | Actual : 7 hours

The prediction for how long I would spend for task 1 - preparing the data and cleaning data was very accurate since the cleaning process is straightforward and I have build some knowledge of the dataset when doing research in finding dataset. Thus, the preparing data part is smooth. For task 2, learning how to use a new library Altair and incorporating the interactive feature takes longer than I anticipated. Especially getting the details of plotting matching my expectations, such as the size of the plot and the color of the plot, are more time-consuming than I thought because each plotting library vary slightly in terms of modifying axis, colors. I have to adapt the commands Altair used by going though the documentations from the Altair library. This teaches me to that the next time I start a project, I need to consider both the big picture ideas and the time to finish potential details. For part 3, the time for building the machine learning was similar to what I expected, but then I realized it takes more than just changing the parameter from the function to decide what tree Max_depth is suitable for this particular project since this would not be a good way to present to the readers. So I took the extra 2 hours to create plots of the relationships between tree Max_depth and model performance. For future projects, I will think more thoroughly about making a work plan and setting aside time for exploration. I was surprised that part 4 take much longer than I expect. Then I realized that it takes time to come up with ways to synthesize all the codes' findings and summarize findings with supporting plots generated from the codes.

Testing

(1) Testing the project codes
For testing the project codes to see the dataset has been prepared and cleaned for analysis, I created another Jupyter Notebook and utilized the assert_equals functions from cse163_utils.py. Firstly, I check if all 883 car models are included after cleaning by examining the row numbers of the data frame. Secondly, I check if all the features and the label for building the machine learning model are included by check whether the sorted list of features and label matched the sorted data frame columns. After confirming the dimension matched my expectations, I pick two car models from the dataset and check whether each row's entires match the original dataset before filtering to ensure the consistency of dataset.

(2) Examines the machine learning mode performance
For testing whether the machine learning model is effective, I use a random split to break the dataset up randomly into a training set and test set (20% of the rows to be the test set) with two examining standards: R-squared Score and mean-squared error. I choose R-squared Score because it represents the proportion of the variance for a dependent variable explained by an independent variable or variables in a regression model. The higher the R squared(closer to 1), our model is more accurate. I also choose another way to evaluate the performance of decision tree regressor models by mean-squared error, which measures the average of the squares of the errors. The smaller the mean-squared error, the more accurate our model is.

Collaboration

I did not collaborate with anybody.