CARBON DIOXIDE EMISSIONS PREDICTION REPORT

Prepared by: Sidra Kamal
August 2022

Data Source: *https://www.kaggle.com/datasets/debajyotipodder/co2-emission-by-vehicles*

INTRODUCTION:

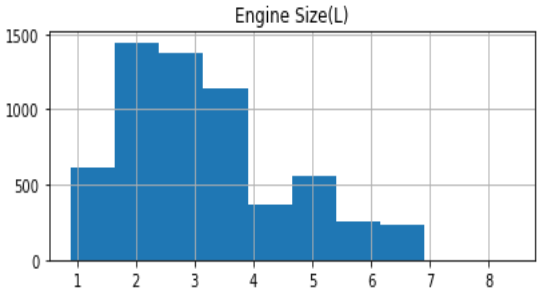
The transportation sector is responsible for 27 percent of greenhouse gas (GHG) emissions in Canada. The usage of cars has increased over the years, which in turn has led to increase in CO_2 emissions. This is a serious issue as CO_2 is a principle GHG that is linked with global warming and climate change. Canada is committed to reducing its GHG emissions and is therefore finding methods of reducing its global carbon footprint.
This dataset is obtained from Government of Canada and captures the details of how CO_2 emissions by a vehicle can vary with different features. It contains information on 7385 vehicles of different car brands and their respective features such as type of fuel, engine size(L), cylinders, fuel consumption(L/km), transmission type and vehicle class.

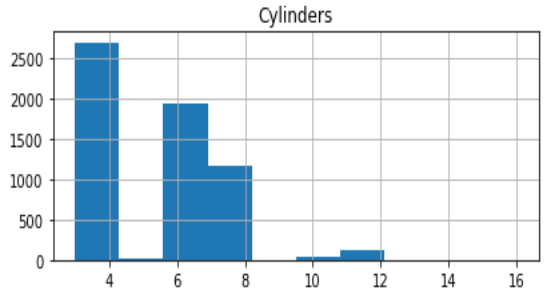
PROBLEM:

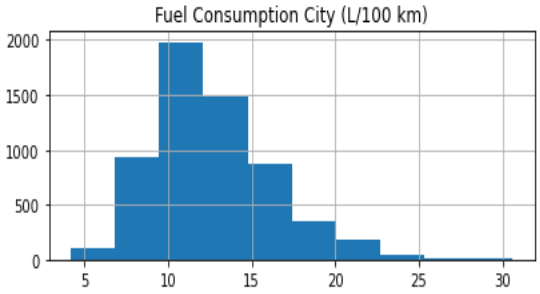
There are two ways to reduce CO_2 emissions from cars: by making vehicles more efficient or by changing the fuel used. Cars and SUVs are a significant contributor to Canada's overall greenhouse gas (GHG) emissions, which are pushing the world into a climate crisis. Light-duty vehicles produce more than four times the GHG emissions of all domestic aviation, according to Canada's 2019 national greenhouse gas inventory. Light duty vehicles and light duty trucks such as cars, pickups, SUVs and smaller vans all account for nearly half of all GHG emissions from the transportation category. Heavy-duty vehicles, such as 18-wheelers and larger pickup trucks, make up the other big chunk, at 35 per cent.

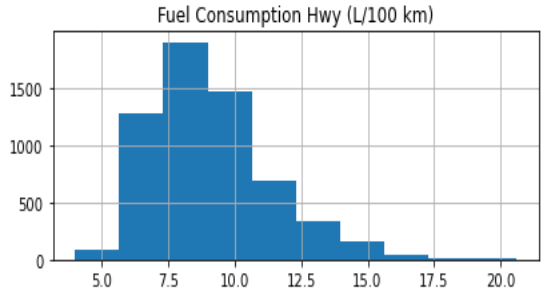
DATA WRANGLING:

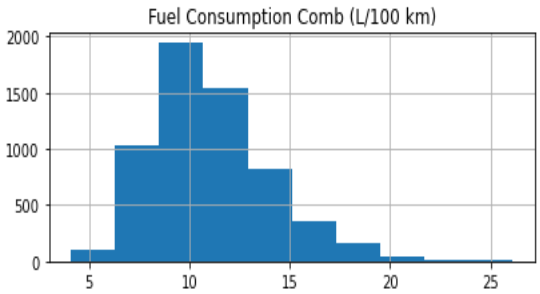
The data originally came as two CSV files. Fuel Type feature values were changed because it had abbreviated version of the fuel. Mapping dictionary was created that mapped the abbreviation with the Fuel and changed the values. Missing values for each column were checked as it would affect the regression model if not handled beforehand. For this dataset, there weren't any missing values. Transmission feature values were changed for any value that started with 'A' as 'Automatic' and 'M' as Manual. Unnecessary columns such as 'Fuel Consumption in mpg' were removed because another column with similar feature in S.I unit already exists in this dataset. 1394 duplicate data values were found and removed, and then descriptive statistics was performed.

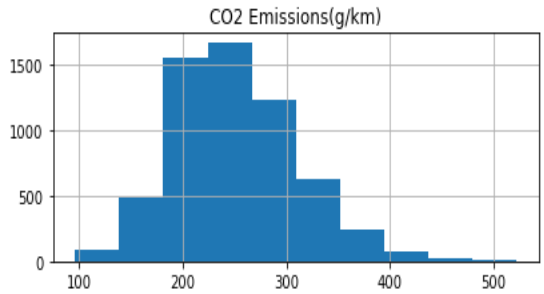




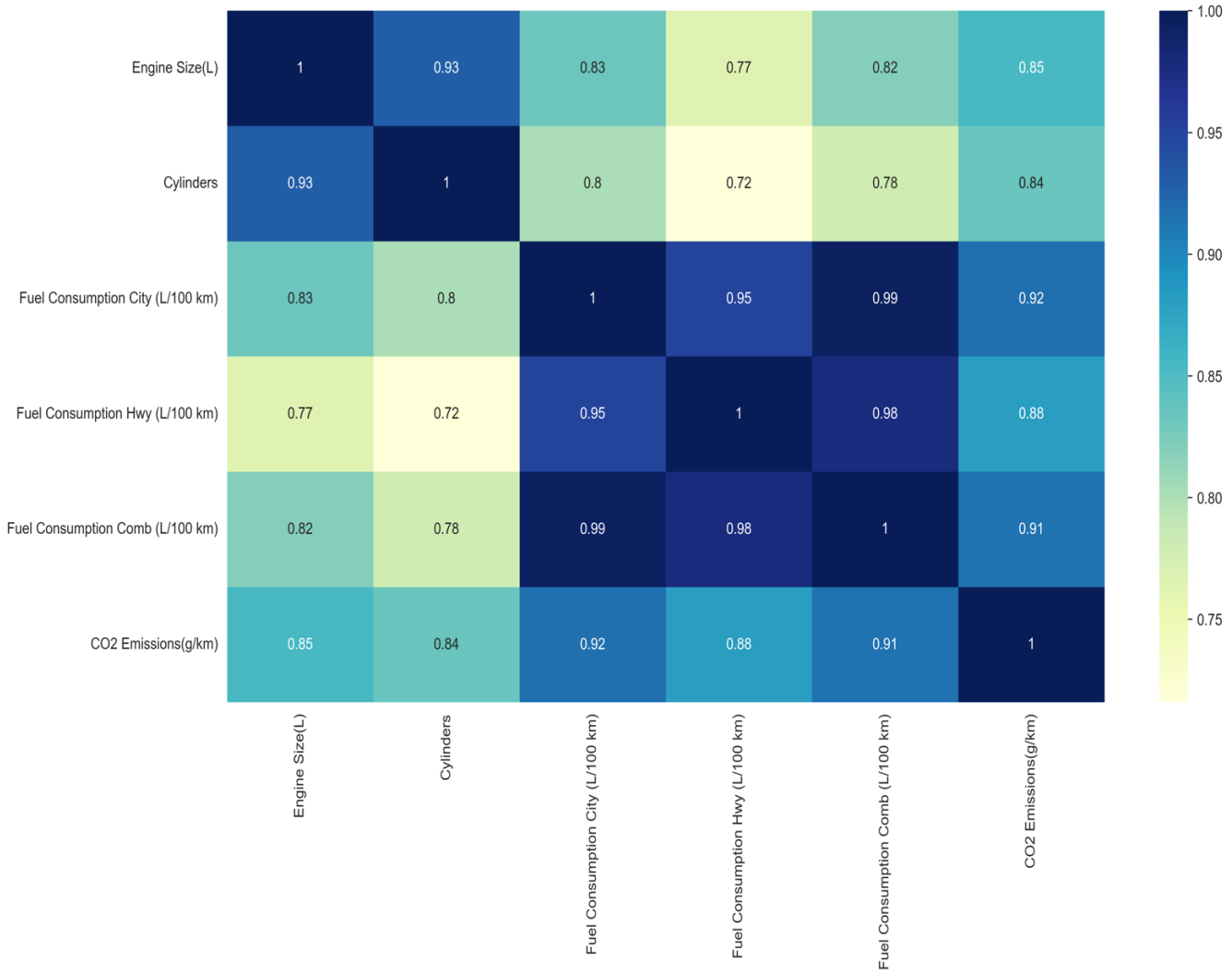








Histogram of distributions of features is plotted to investigate any obvious outliers. There aren't any features that stand out in terms of being outliers, however cylinders seem to be clustered towards the lower end. This could be due to most of the cars having 4 or less cylinders on average.



The heatmap shows a positive correlation between Fuel consumption/CO_2 Emissions and Engine Size(L) and Fuel Consumption/CO_2 Emissions and cylinders.

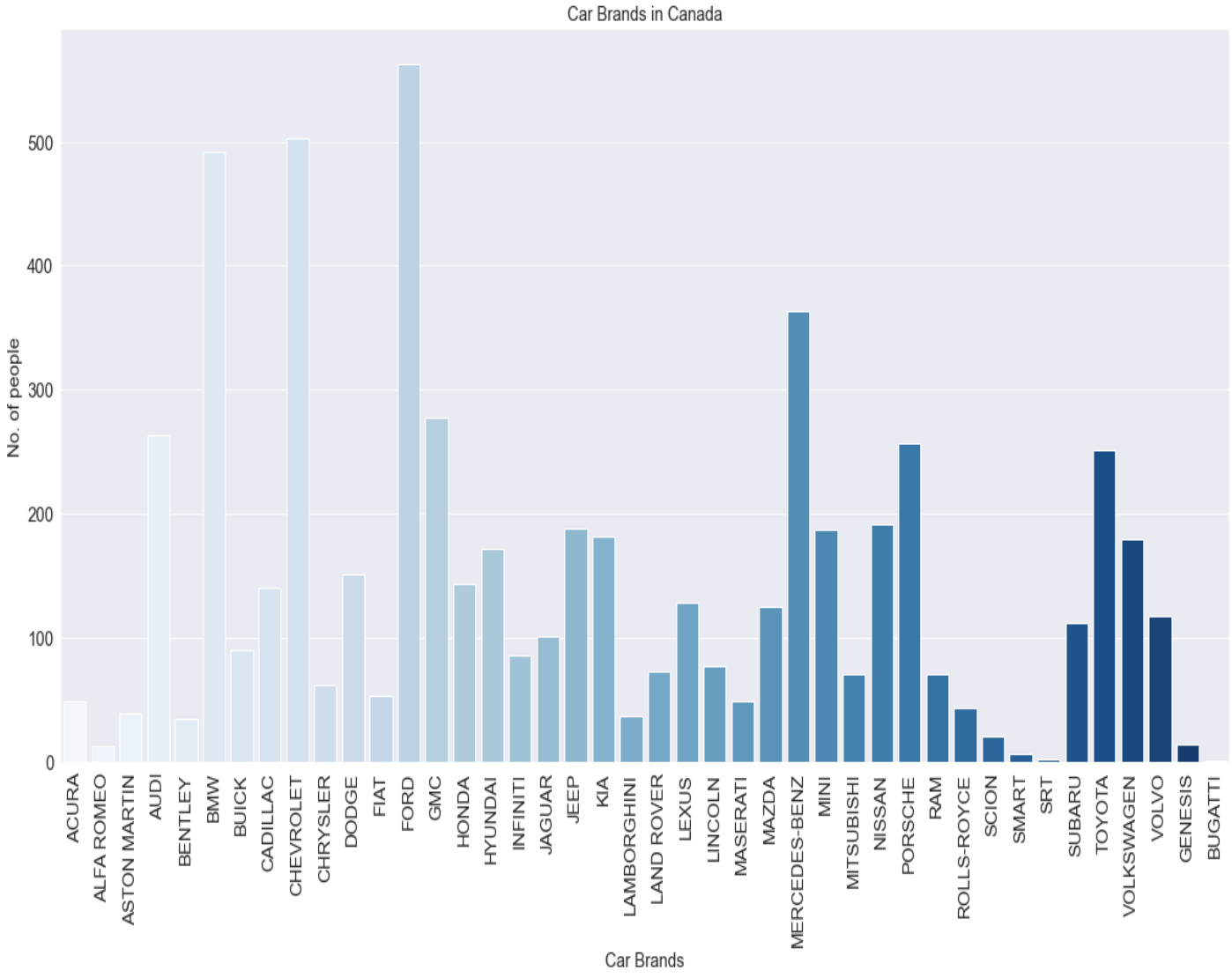
EXPLORATORY DATA ANALYSIS:

Before starting this process, some questions and hypothesis that need to be addressed are:

1. Which auto brand is used the most in Canada?
2. Which model and class have the highest CO_2 emissions?
3. Does increase in Engine Size(L) have an impact on CO_2 emissions?
4. Which fuel type results in highest CO_2 emissions?
5. Automatic transmission is better than manual transmission for fuel consumption because of the design of their system
6. Does increase in number of cylinders affect fuel consumption and CO_2 Emissions?
7. City has higher CO_2 emissions than highway.

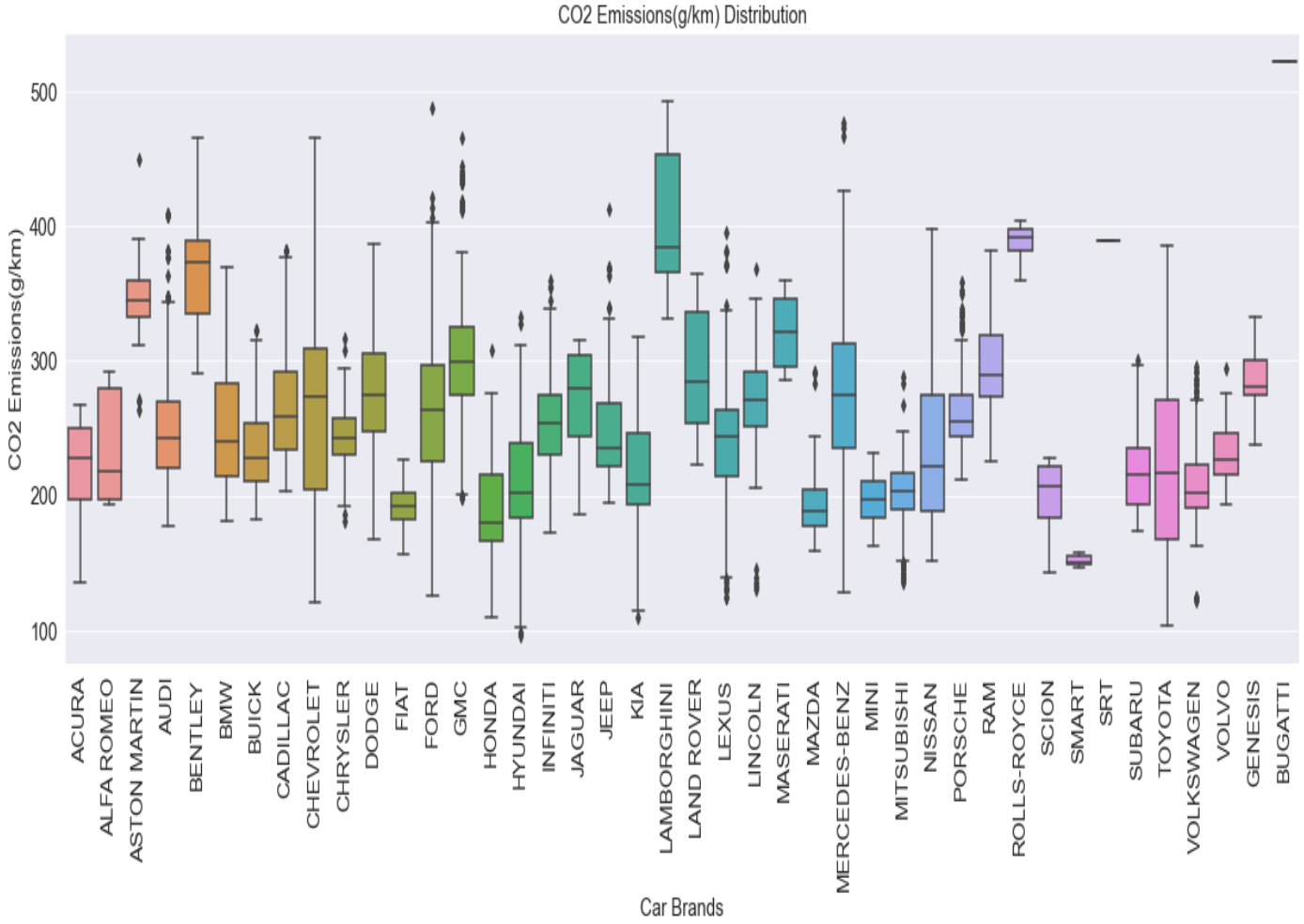
Let's address these questions, claims and hypothesis:

1. Which auto brands are most popular in Canada?



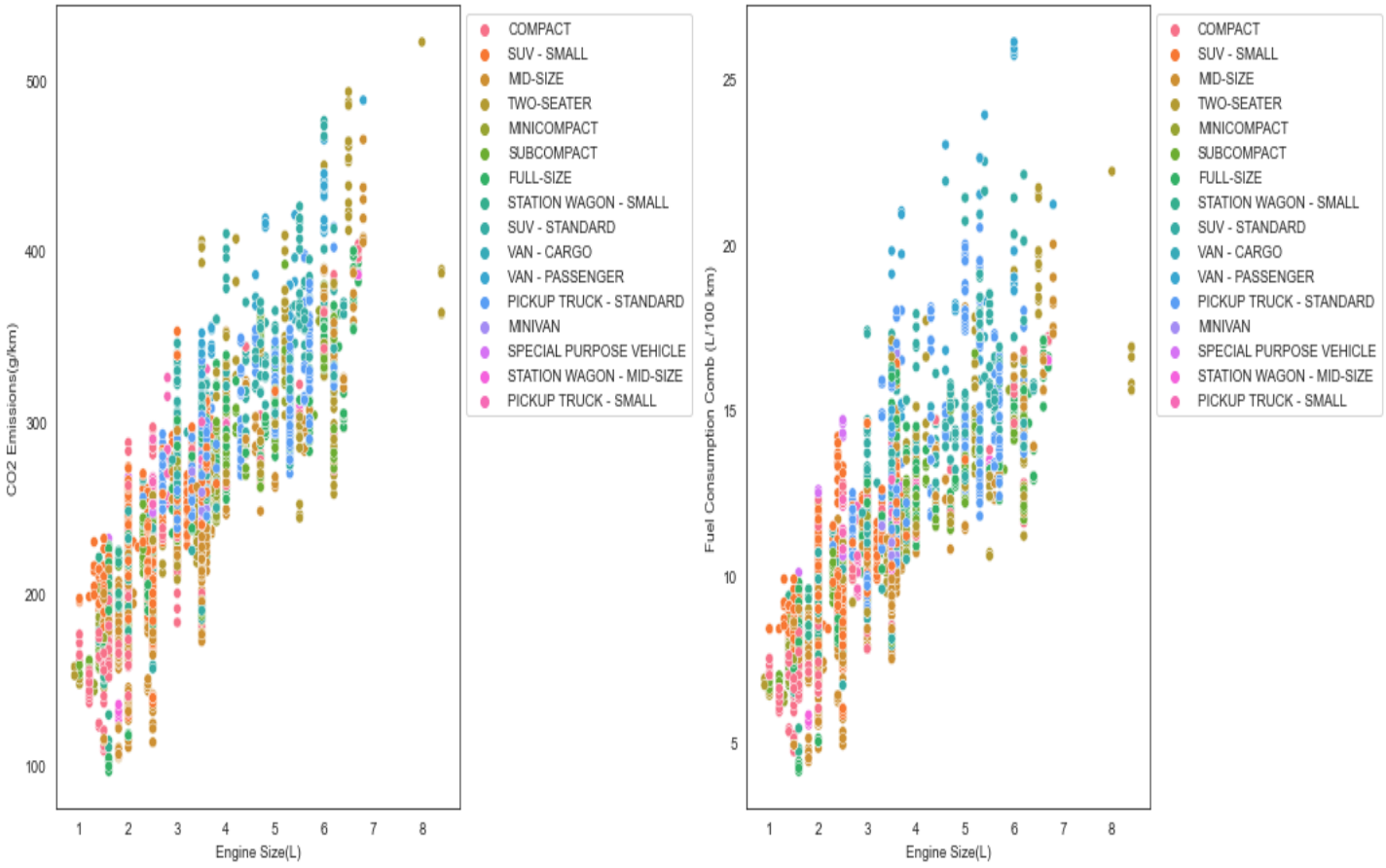
This bar chart shows that 'Ford' auto brand seems to be the most popular choice in Canada primarily because of the Ford F-series, followed by 'Chevrolet' and 'BMW'. Due to harsh winter conditions in Canada, most of the people prefer all-wheel drive to drive through snow and bad road conditions.

2. Which model and class have the highest CO2 Emissions?



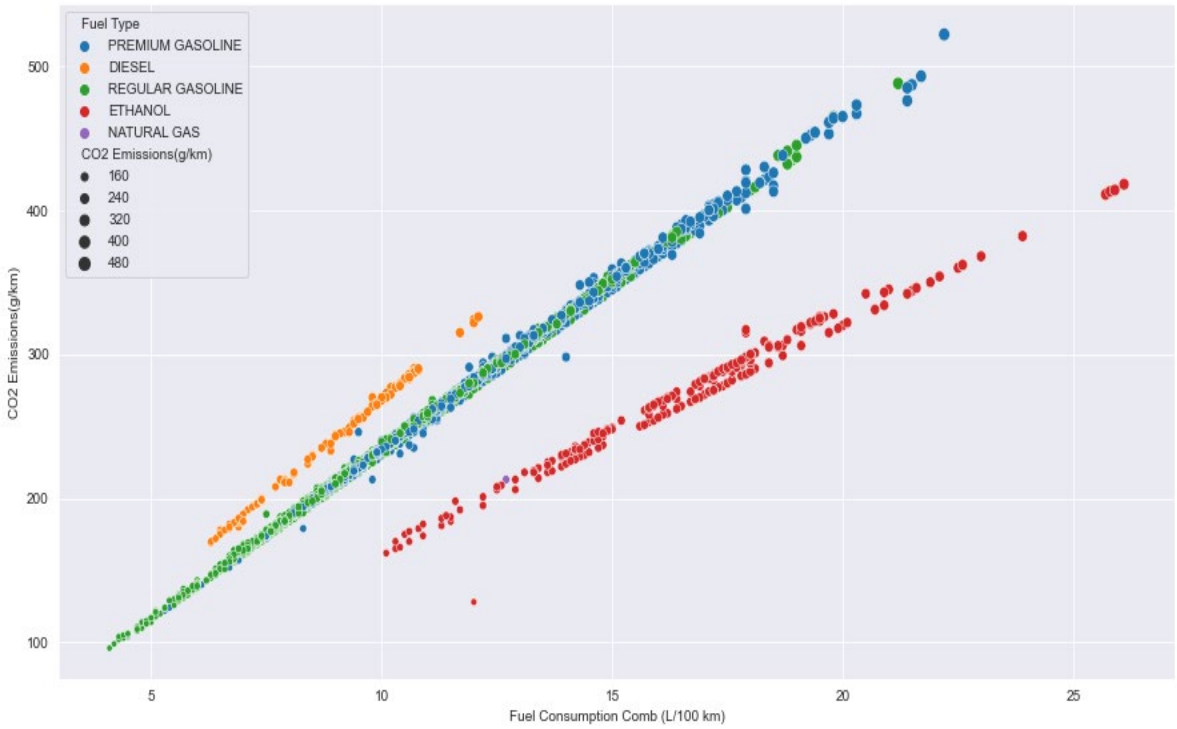
From this boxplot, it seems that Lamborghini and Bugatti have the highest CO_2 Emissions. This is because Lamborghini and Bugatti are sports car, and they tend to have large engine size(L), high horsepower and worst fuel economy. GMC, JEEP, AUDI, LEXUS, MERCEDES-BENZ and FORD have outliers. There is variation of CO_2 Emissions for every auto brand because there are different models for each brand that have different specifications which in turn leads to different amount of emissions.

3. Does increase in Engine Size(L) have an impact on CO_2 emissions?



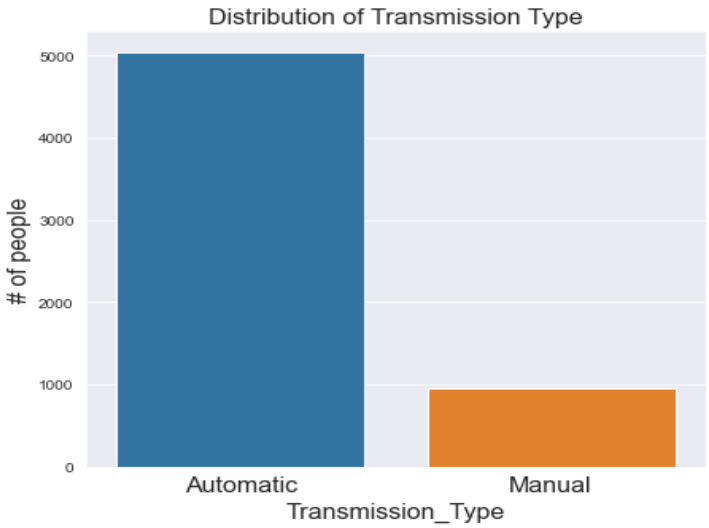
There appears to be a somewhat linear relationship between engine size, fuel consumption (L/100km) and CO_2 Emissions(g/km). As a larger engine is generally able to burn more fuel and produce more power, a car with a larger, more powerful engine is likely to be able to accelerate faster and have higher emissions.
By plotting the two on a scatterplot and adding a color coding of Vehicle Class, we can see certain clustering patterns emerge. Small Vehicles have smaller engines and therefore consume less fuel. Two-Seaters have the highest fuel consumption, followed by Vans and SUV. Larger engine size is generally able to burn more fuel and produce more power and accelerate faster.

4. Which fuel type results in highest CO_2 emissions?

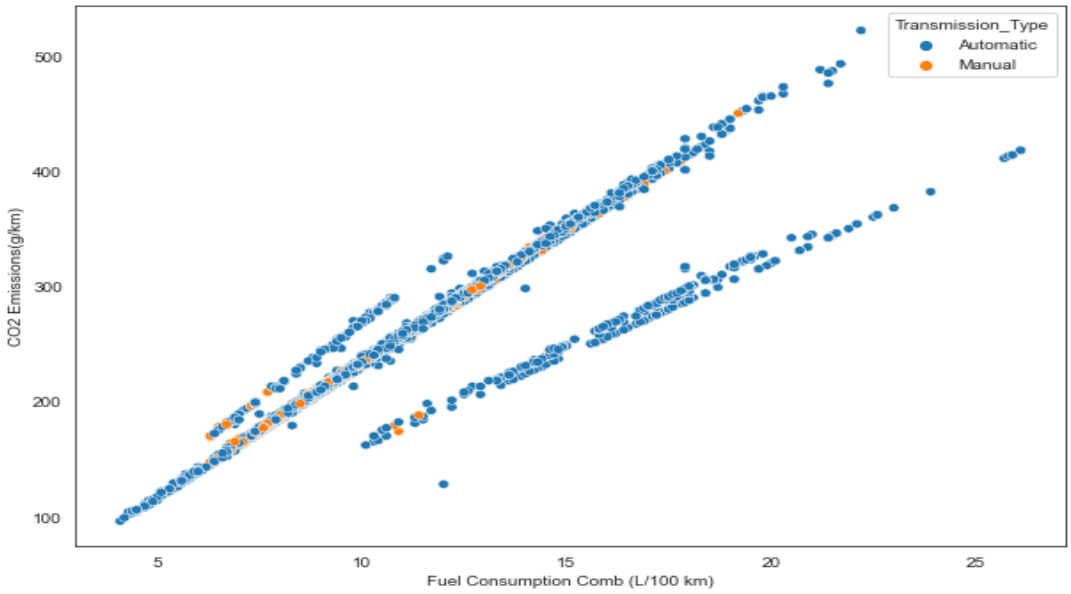


Fuel Consumption Combined in Highway and City is plotted against the emissions and hue adjusted as the fuel type. Distinct slope can be seen for each fuel type used. The steeper slope shows higher emissions for lower fuel consumption. Diesel, Natural Gas, Premium Gasoline and Regular Gasoline have steeper curves.
Looking at the scatterplot of the different fuel shows that ethanol-based fuels on average result in the lowest CO_2 emissions and higher fuel consumption. E85 contains 83% ethanol content and has about 27% less energy per gallon than gasoline so therefore it would consume more fuel in comparison to gasoline to go the same distance(L/km). Ethanol based fuels also have the lowest emissions because plants that are made into renewable fuels absorb carbon dioxide from the atmosphere as they grow, and that same amount of carbon dioxide is re-released when the fuel is produced and combusted in an engine. In this way, ethanol and other renewables simply recycle atmospheric carbon.

5. Hypothesis: “Manual transmission is better than automatic transmission for fuel consumption because of the design of their system”

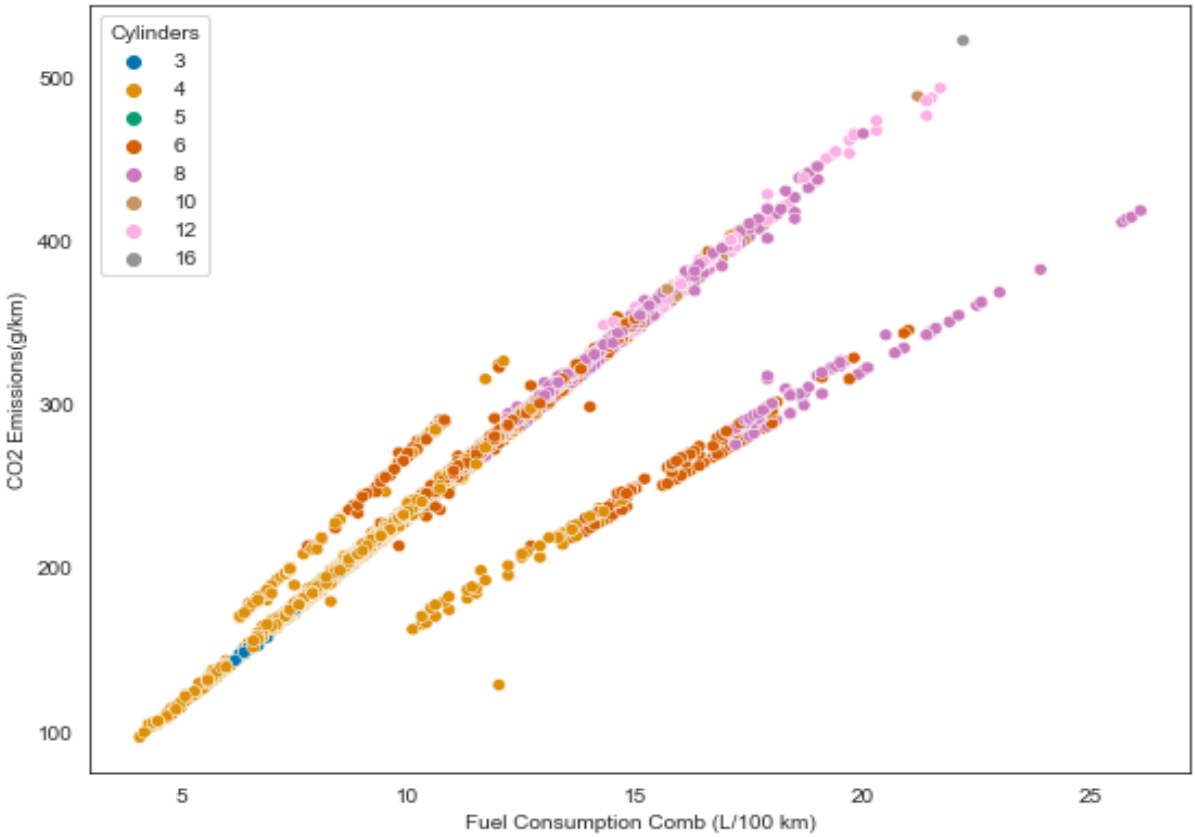


From this bar graph, Automatic transmission is a popular choice for people in Canada.



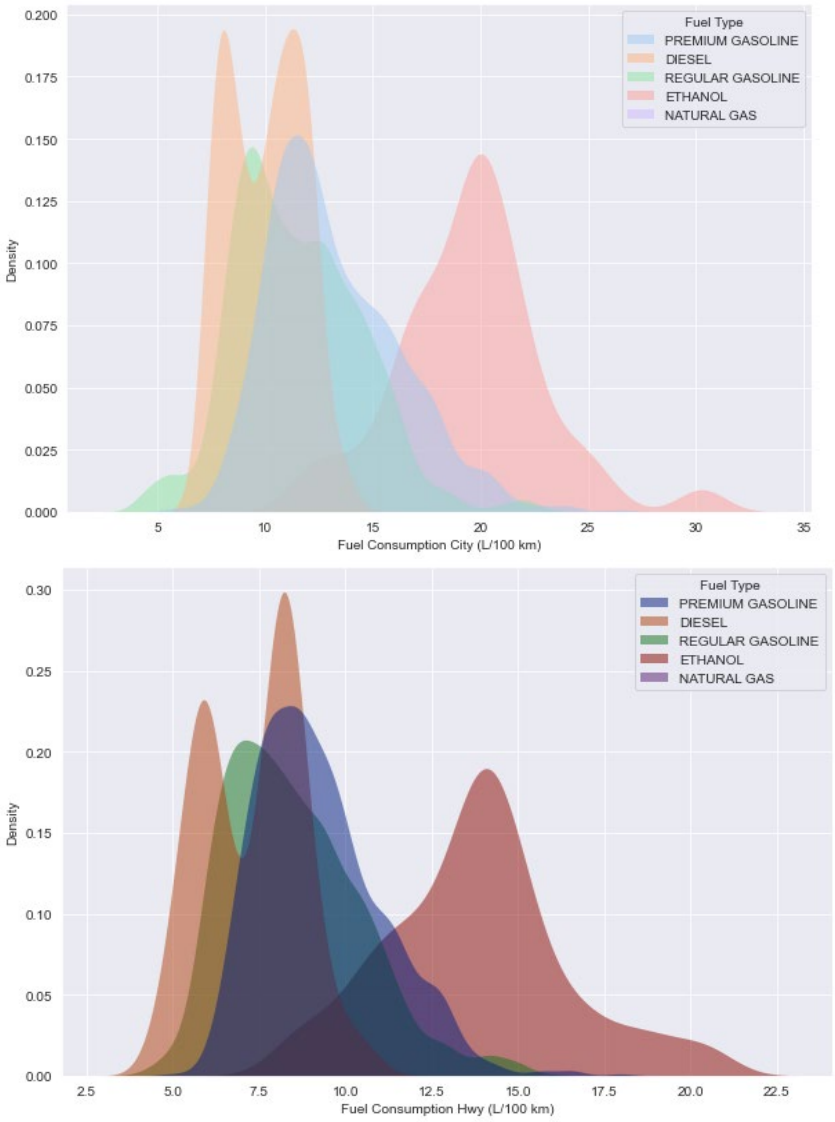
The visualizations shown above does not go into detail regarding any biases that may exist in the data. The number of automatic vehicles in the dataset are significantly larger than the manual vehicles. More data needs to be gathered to be more specific regarding the transmission type being a factor for CO2 emissions impact.

6. Does increase in number of cylinders affect fuel consumption and CO_2 Emissions?



Cylinders are the powerhouse of a car engine. Lesser number of cylinders consumes a lesser amount of fuel in running the engine. Essentially, the larger the volume of the cylinder, the more room there is for air and fuel inside it, which dictates how much power it can produce and how much fuel it consumes. Combined Fuel consumption is plotted against CO_2 emissions and number of cylinders is color coded. It is clearly seen that with increase in the number of cylinders, the CO_2 emissions and fuel consumption also increases.

7. Hypothesis: City has higher CO_2 emissions in comparison to highway



The Kernel Density Plot shows that Fuel consumption is higher in City in comparison to Highway for each fuel type. It is usually higher on highways compared to city since city driving necessitates slower speeds, idling and braking.

MACHINE LEARNING:

All the categorical features were transformed to one-hot numeric arrays using get_dummies from sklearn library in python. Feature Scaling was performed to improve the convergence of steepest descent algorithm like gradient descent. It was also used to make sure that the data ranges are uniform and to compare the measurements that have different units (e.g: fuel consumption(L/km) and engine size(L)) and allows for increases in performance.

METRICS:

The coefficient of determination denoted R^2 and MSE (Mean Squared Error) will be used to evaluate each model. R^2 is a measure of the goodness of fit of a model. In regression, the R^2 coefficient of determination is a statistical measure of how well the regression predictions approximate the real data points. MSE will be representing the average of the squared difference between the original and predicted values for the CO_2 emissions in the data set. MAE will measure the average of the absolute difference between the actual and predicted values.

MODELS:

Data Modeling: Nine different models were tried and tested. Fine hyperparameter tuning on each of them was performed through cross validation and grid search to obtain high coefficient of determination and low Mean Absolute Error and prevent overfitting of the training dataset. The data was split into 30% testing and 70% training. The nine models that were tested are as follows:

- **Lasso Regression (L1 Regularization):** This adds a penalty term equal to the absolute value of the magnitude of the coefficients. The lambda coefficient is tuned and decided upon how strong the penalty should be. This model introduced hyperparameter, alpha, the coefficient to penalize weights. The alpha chosen was 10 and cross validation was performed.

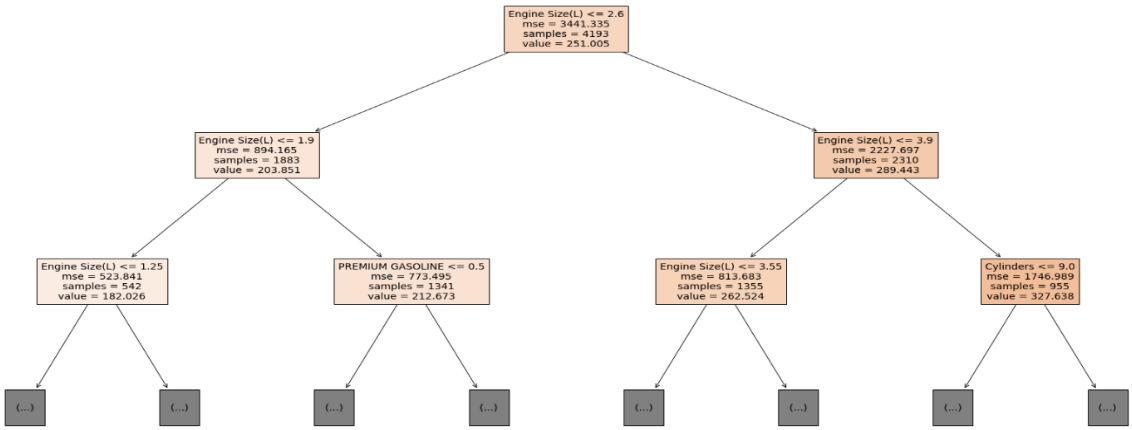
- **Ridge Regression (L2 Regularization):** This adds a penalty equal to the square of the magnitude of the coefficients. All coefficients are shrunk by the same factor, and it doesn't necessarily eliminate the coefficients. Thus, the weights not only tend to have smaller absolute values, but also really tend to penalize the extremes of the weights, resulting in a group of weights that are more evenly distributed. After performing k-fold cross-validation fit to the training set, the optimal alpha obtained was 10.

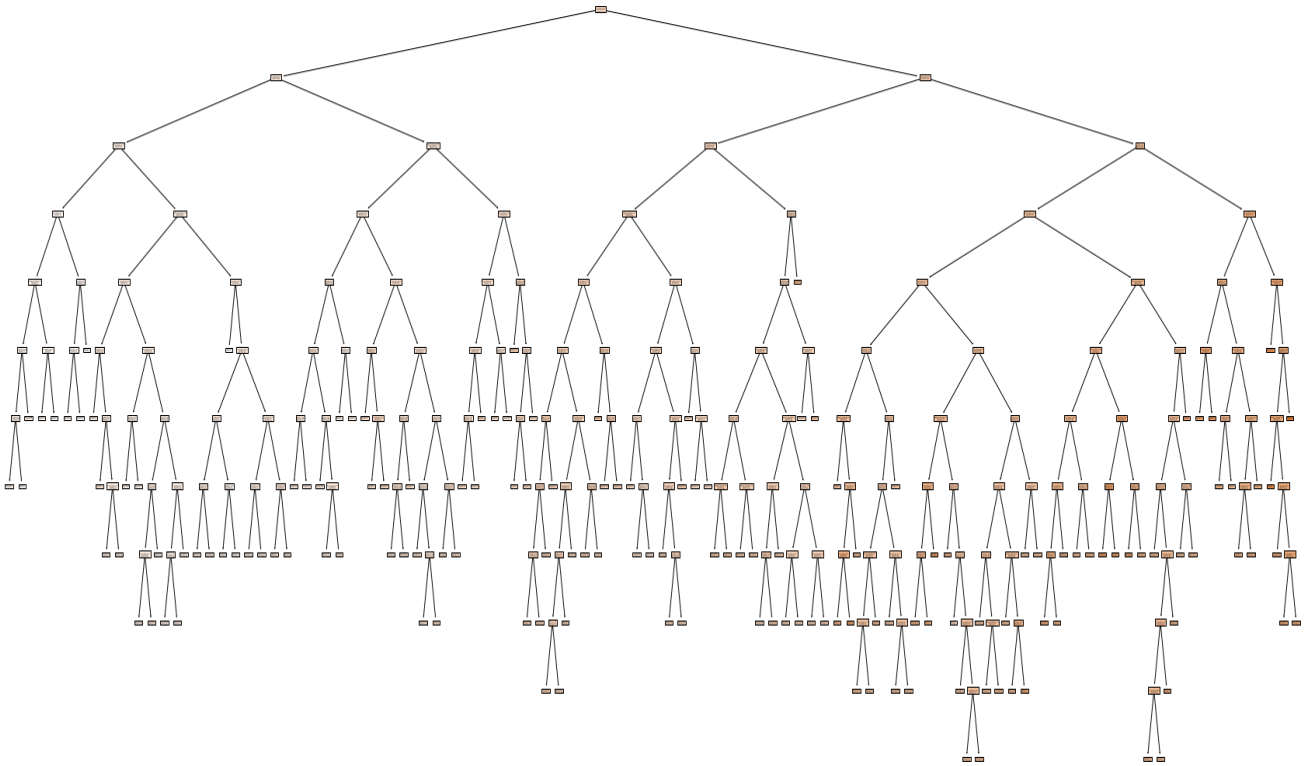
- **Elastic Net Regression:** The absolute value penalization and squared penalization are combined in this model with the addition of an alpha parameter deciding the ration between them. The L1_ratio parameter (elastic net mixing parameter chosen was 0.95 after performing grid search and cross validation.

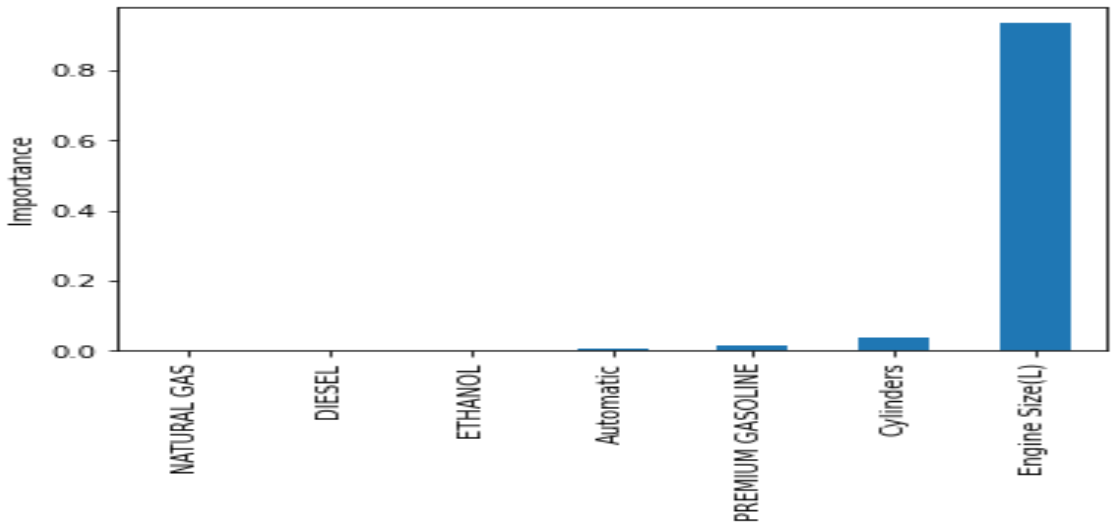
- **Linear Regression:** This model is not penalized for its choice of weights therefore during the training stage, if the model feels like one feature is particularly important, the model will place a larger weight to that feature which may lead to overfitting. To our surprise, the model gave the same score as the ridge regression model. This maybe because some features are important than others in predicting $C0_2$ Emissions in the dataset.

```
In [35]: print('Accuracy of linear regression on test set: {:.2f}'.format(reg_all.score(X_test, y_test)))
         Accuracy of linear regression on test set: 0.75

In [36]: regression_score=(reg_all.score(X_test, y_test))

In [37]: #finding the residuals
         test_res = y_test - linear_y_pred
         type(test_res)

Out[37]: pandas.core.series.Series

In [38]: from sklearn.metrics import mean_squared_error
         from sklearn.metrics import mean_absolute_error

In [39]: #Mean Squared Error, Root Mean Squared Error
         MAE = mean_absolute_error(y_test,linear_y_pred)
         MSE = mean_squared_error(y_test,linear_y_pred)
         RMSE = np.sqrt(MSE)

In [41]: MAE,RMSE

Out[41]: (22.99259787072235, 30.178428293529244)
```

- **Decision Tree Regression:** Grid Search was performed and the max_depth selected from this grid search was 12 with cross validation=5







The feature of importance for this model was found to be Engine Size(L) with importance score of above 0.8

- **Random Forest Regression:** Random Search with cross validation was performed and the best parameters obtained for this model after performing hyper-parameters tuning are:

jupyter C02 Emissions-Modelling Data Last Checkpoint: 07/27/2022 (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

Run Code

```
                                random_state=42, verbose=2)
```

In [719]:
```
rf_random.best_params_
```

Out[719]:
```
{'n_estimators': 200,
 'min_samples_split': 2,
 'min_samples_leaf': 1,
 'max_features': 'sqrt',
 'max_depth': 50,
 'bootstrap': True}
```

In [720]:
```
rf_random_pred = rf_random.predict(X_test)
rf_random_pred
```

Out[720]:
```
array([202.97135102, 269.67426458, 218.75509034, ..., 316.02660437,
       333.08108677, 217.80035696])
```

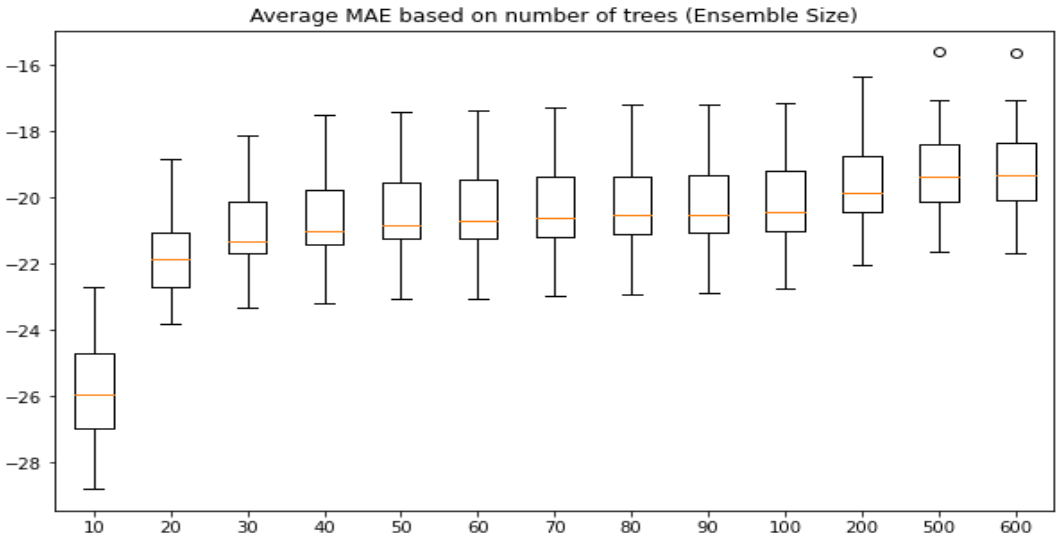
In [721]:
```
print('Accuracy of RF cross validation regression on test set: {:.2f}'.format(rf_random.score(X_test, y_test)))
```

```
Accuracy of RF cross validation regression on test set: 0.83
```

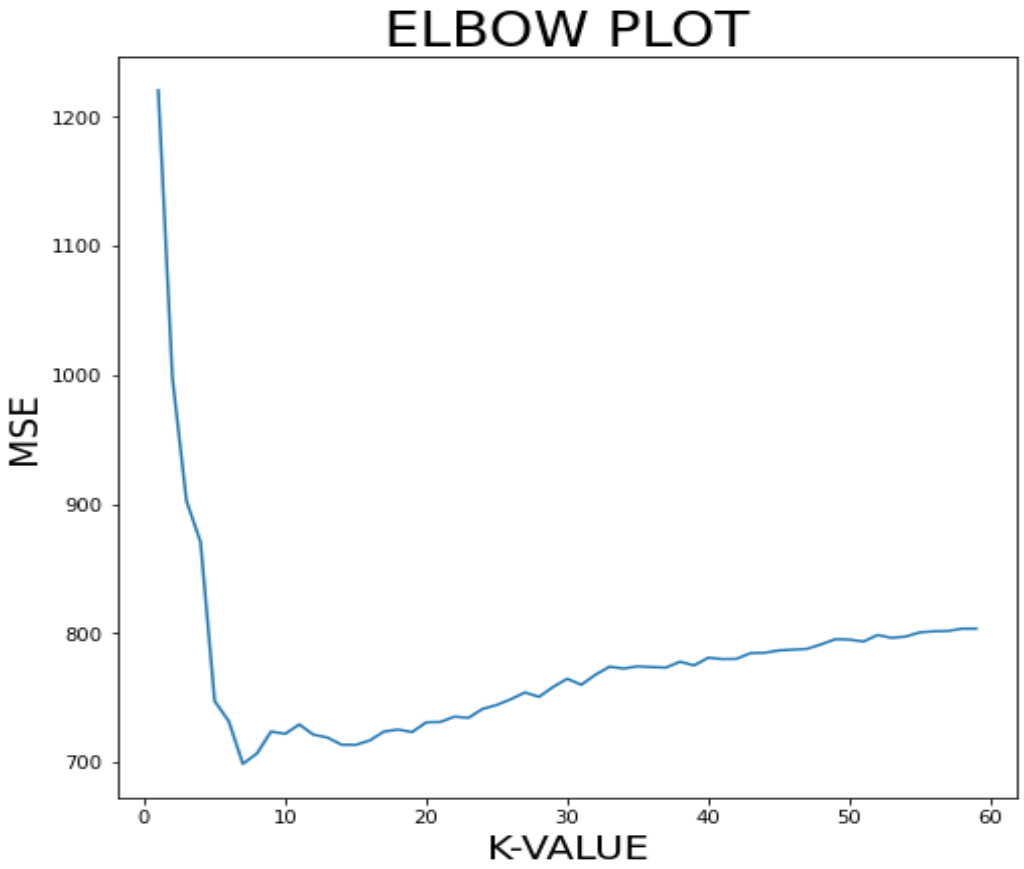
In [722]:
```
rf_random_score=rf_random.score(X_test, y_test)
```

In [723]:
```
data ={'Model':['linear_regression','ridge_regression', 'lasso_regression','elastic_regression',
                'knn_regression', 'svm_regression', 'decision_tree_regression','gradient_boosting_regression',
                'random_forest_regression'],'Model_Score':[regression_score,ridge_score, lasso_score,elastic_score,
               knn_score, svm_model, decision_tree_score,gradient_boosting,
                random_forest]}

df_model= pd.DataFrame(data=data);
df_model
```

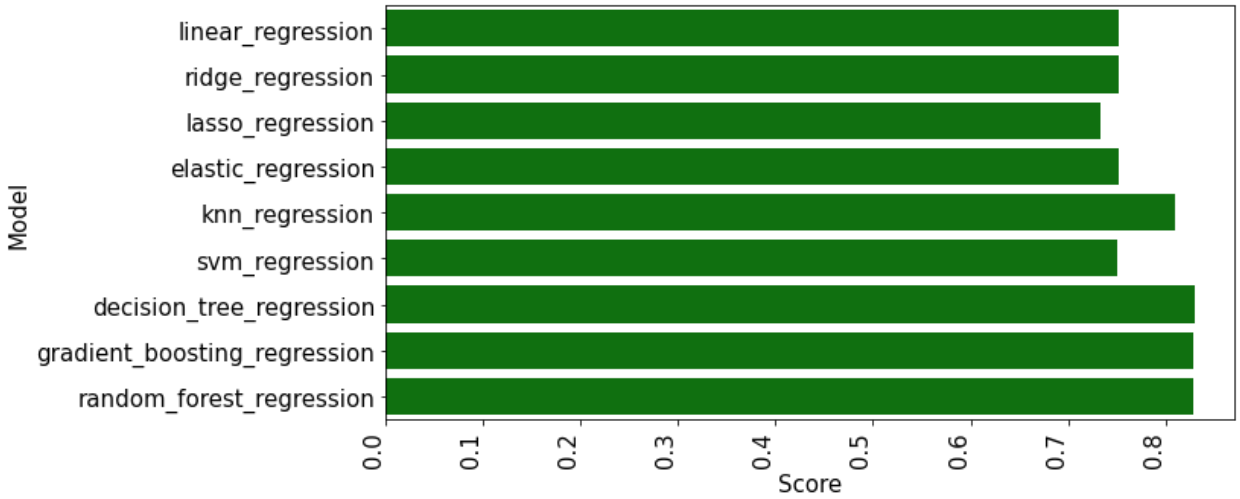
- **Gradient Boosting:** Individual models were trained in a sequential way. Each individual model learned from mistakes made by the previous model. With this model, the best parameters we obtained were maximum depth of 4 and number of estimators (random subsets=500 from the training set)



- **KNN REGRESSOR:** Constructed an Elbow plot to determine the optimal value of K that yields lowest MSE. From the elbow plot, we deduced that the k-value of 7 yielded the lowest MSE.



	Model	Model_Score
0	linear_regression	0.751553
1	ridge_regression	0.751638
2	lasso_regression	0.732634
3	elastic_regression	0.751755
4	knn_regression	0.809349
5	svm_regression	0.750893
6	decision_tree_regression	0.829381
7	gradient_boosting_regression	0.828529
8	random_forest_regression	0.827929



From all the models tested, it was found that decision tree, random forest regression and gradient boosting had the best coefficient of determination score of above 0.8. They also had the lowest MAE and MSE. This shows that these three models are the most reliable for predicting the CO_2 Emissions from motor vehicles given the features of fuel type, transmission, engine size(L) and number of cylinders.

CONCLUSION

Data Visualization was done to see the correlation between different features and find out any pattern that could show relationship between any car features.

By building machine learning models, we were able to figure out that Engine Size(L) had the most impact on the performance of the model and that random forest, decision tree and gradient boosting were the best models for predicting the emissions.

These models could be improved by incorporating more features like horsepower, weight of vehicle, type of car wheels used, additional car features like A/C or heater and how much of it is used by the driver etc.

Government can use these models to predict CO_2 Emissions and set criteria for regulations on the type of vehicle used in the city and rural area. Likewise, auto brands can use these models to predict CO2 Emissions based on the design of the car and its features.

Decreasing Engine Size(L) and using biofuels like ethanol seems to be the best choice for lowering the emissions in the city and highways.