Solent University

Faculty of Business, Law, and Digital Technologies

Machine learning prediction of carbon emission radiation from vehicles and engines and how to increase performance of the algorithm.

Author : **Q15798411**

Course Title : **Dissertation**

Supervisor : **Shadi Eltanani (PhD)**

Date : **9th September 2022**

Table of Content

Abstract

Climate change is a global menace. It is one of the greatest problems the world is currently battling with. High level of carbon emission, because of increasing human industrial activities, is one of the major culprits of climate change. The effect of climate change cannot be underestimated. Extreme heat, rising sea level, massive flooding and cyclone, and drought are some of the challenges that arose with climate change. There are many ways through which carbon dioxide is been emoted into the atmosphere, but this study focuses on emission from automobile vehicles.

Machine learning, artificial intelligence and data science are widely used across various sectors like finance, healthcare, agriculture and so on, and this has turnaround our modern lifestyle. In recent time, machine learning knowledge has now been applied to climate change. Machine learning algorithm can used to predict different aspects of carbon accounting based on various available historical data.

This work used linear algorithm model to predict carbon emission of various brands of vehicles and it has a performance level of eighty-four percent.

Chapter one

1.1 Background of the Study

In the summer of 2022, according to the Met Office (the national metrological service for the UK), UK recorded the highest temperature ever in history. A provisional temperature of 40.3°C was recorded at Coningsby on 19th July 2022, and this beat the previous record of 38.7°C set in 2019 by 1.5°C. (https://www.metoffice.gov.uk/about-us/press-office/news/weather-and-climate/2022/red-extreme-heat-warning-ud). The greatest problem the world is battling with right now is climate change. An overwhelming body of scientific evidence indicates that the earth's climate is rapidly changing, predominantly because of increases in greenhouse gases caused by human activities. (Nicholas Stern 2007). Human activities are changing the composition of the atmosphere and its properties

According to the United Nations, climate change refers to the long-term change in temperature and weather patterns. (https://www.un.org/en/climatechange/what-is-climate-change).
Human activities have been the major driver of climate change. This is primarily due to a phenomenon otherwise known as greenhouse effect caused by greenhouse gases. Greenhouse gases include carbon monoxide, carbon dioxide, methane, nitrous oxide, and fluorinated gases. For this project, we shall be emphasising on carbon dioxide because carbon dioxide alone accounts for seventy-nine percent of greenhouse gases and the aim of the project is to develop a machine learning algorithm that can predict carbon emission radiation from vehicles and engines.

Machine learning has been used in the recent time to turnaround our modern lifestyle. This is evident in activities like computer vision, image processing, autonomous vehicles to mention a few. In the same vein, machine learning professionals are now beginning to use the knowledge of machine learning to solve one of the biggest problems that mankind is battling with now, which is climate change. Machine learning is a subdivision of artificial intelligence which extracts patterns from available data and makes prediction of possible future outcome from the data by filling in missing information. With the aid of machine learning, patterns, observations, and predictions can be deduced from the vast historical climate data and carbon emission data, thereby equipping us with the right knowledge of how to mitigate the climate impact in the years ahead.

1.2 Statement of Problem

There is no planet B. Earth has been scientifically proven to be the only habitable planet that can supports human life. However, man's only home is on the brink. The planet is grossly threatened because of climate change which is orchestrated by global warming because of the effect of greenhouse gases, of which carbon dioxide takes a major percentage. This project is targeted towards developing a machine learning algorithm to predict carbon emission radiation of vehicles and engines. If emission can be accurately predicted, measures can be easily provided to mitigate the volume and effect of carbon emission in the future, thereby saving the planet.

1.3 Aims and Objectives

The aims and objectives of this project include:

1. To develop machine learning algorithms to predict carbon emission radiation from vehicles and engines.
2. To compare the prediction accuracy of the developed algorithms.
3. To combine a few algorithms using ensemble method and therefore carry out performance evaluation of the ensembled algorithm.

Chapter two

2.1 What is Pollution and why is it important?

Humans are massively changing the earth. (Vitousek et al, 1997). Between one-third and one-half of the land surface has been transformed by human actions. Since the beginning of industrial revolution, carbon dioxide concentration in the atmosphere has increased by nearly thirty percent. Earth is changing faster than we understand it due to our activities. Pollution is the introduction of any harmful substance into the environment. The harmful substances that are released into the environment are called pollutants. Pollutant can be solid, for example solid wastes and trashes; it may be liquid, like in the case of runoffs generated by factories, sewage produced from domestic activities, runoff of chemicals used on farmlands to kill weeks and pests and so on; pollutant may also be gaseous like in the case of exhaust from car engines and industrial activities. Pollutants contaminate air, water, and soil. All living things depend on earth's supply of air and water. When these resources are continuously polluted, all forms of life become endangered over time.

It should be noted that people and industries do not deliberately contaminate the environment. Pollution occurs because no process is 100% efficient. (Marquita K. Hill 2004). Almost any chemical, substance, or material, irrespective of how it is generated, be it naturally or by human beings can cause pollution. A good example of natural pollution is a volcano. When a volcano erupts, massive amount of ash, chlorine, sulphur dioxide and other chemicals are released into the environment, thereby endangering the lives of organisms in its environs.

Fig 1:

Source:

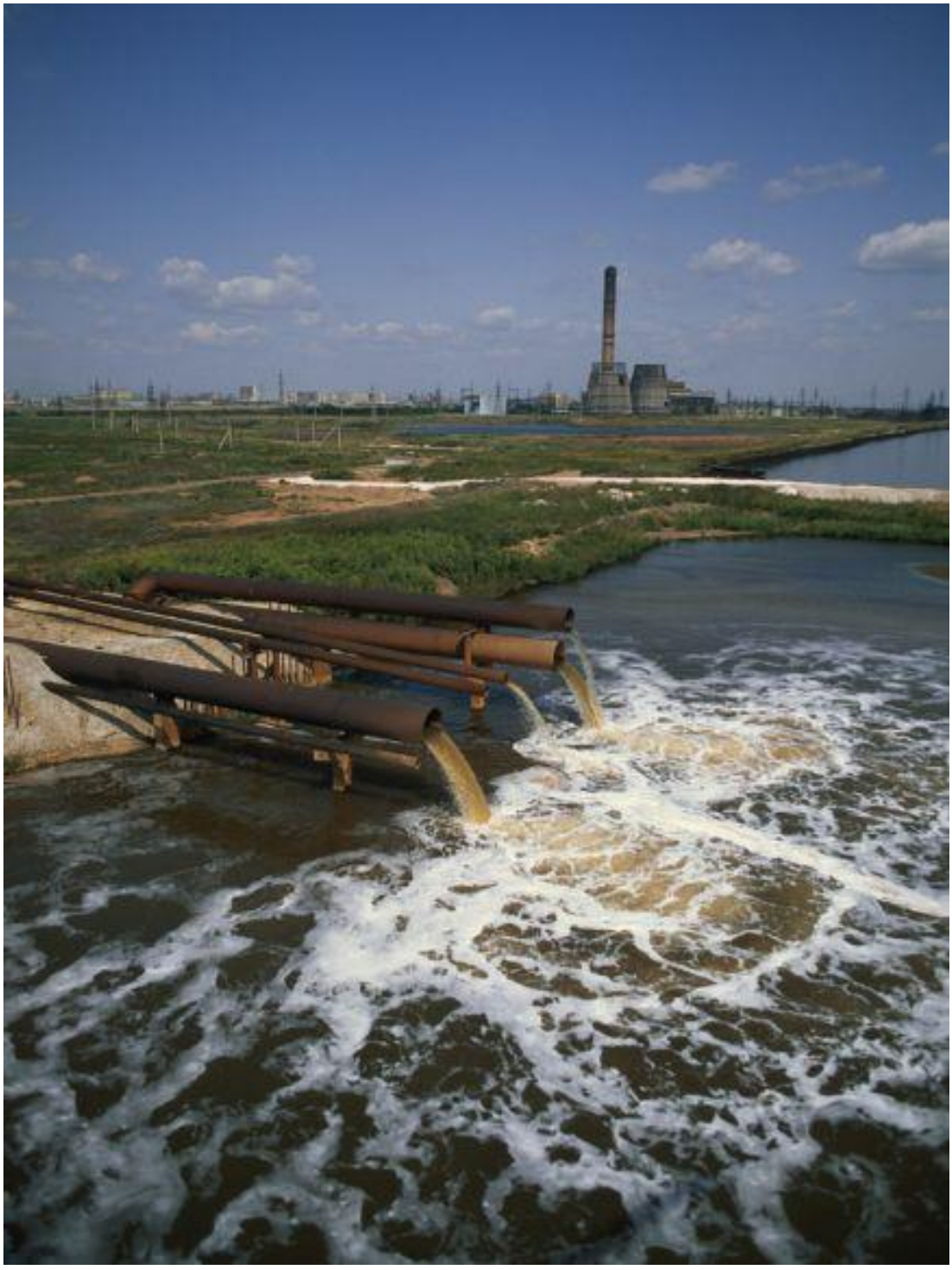


Fig 2:

Source:


6311
ALLIED
The Careful Movers
MARTIN STORAGE CO., INC.

2.2 Understanding Climate Change

Different teams of scientists have independently gathered and combed through more than one hundred years' worth of temperature records. Their analyses all point to a rise of 0.8°C in the average surface air temperature of earth over the last century (Henson R. 2006). The planet is indeed warning up. While a degree increase in surface air temperature may sound like it is negligible, the effect of the rise has indeed been higher in certain locations like the Arctic region. The dynamics of world's climate has drastically changed because of the release of enormous volume of carbon dioxide and other greenhouse gases to the atmosphere over the last 150 years. These gases absorb the heat that is radiated by the earth, but only release part of that heat to the space and retain the rest, thereby resulting in a warmer atmosphere. About 26 billion metric tonnes of CO_2 is released to the atmosphere annually; this is more than four metric tonnes per person per year. Climate change has led to major sea-level rise, increased flooding and droughts, rising ambient temperature, more major hurricanes and extinction of some animal species

2.3 The greenhouse effect

Joseph Fourier, a French mathematician, and physicist was one of the foremost proponents of the earth energy balance. In 1820, Fourier was able to show through his calculations the stark temperature difference between an airless earth and the one we enjoy. Fourier knew that the energy reaching earth as sunlight must be balanced by energy returning to space, some of it in a different form. Although he could not pin down the exact process, but he suspected that some of this outgoing energy

is intercepted by the atmosphere, thus keeping us warmer than we would otherwise be. Hypothesising after the results by other scientists on how glass box traps heat was known, the concept of greenhouse effect was born. The gases that are responsible for greenhouse effect are carbon dioxide, nitrous oxide, methane, ozone and hydrofluorocarbon. Carbon dioxide takes the larger percentage of the greenhouse gases. Greenhouse gases is made up of 53% (380ppm) Carbon dioxide, 17% (1.8ppm) methane, 13% (0.03ppm) near-surface ozone, 12% (0.3ppm) nitrous oxide and 5% (1% ppm) HCFs.

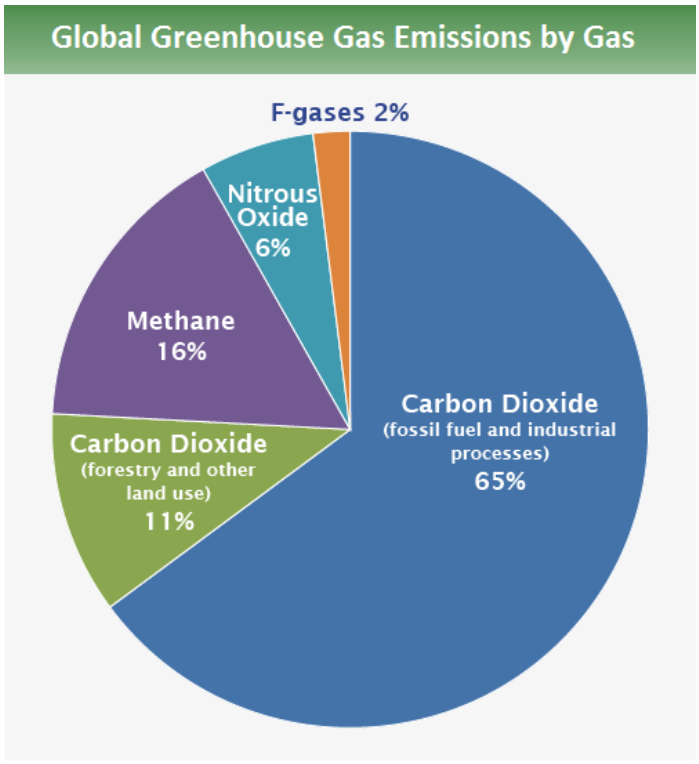


Fig 2. distribution pie chart of greenhouse gases Source: IPCC (2014)

Carbon dioxide (CO_2): This is the chief culprit among the greenhouse gases. It accounts for about 380 of every million molecules in the air

2.4 What is machine learning?

Machine learning is the science and art of programming computers so they can learn from Data (Geron A 2019). Machine learning enables computers with the capacity to learn with being explicitly programmed (Samuel A, 1959). In machine learning, patterns are basically extracted from data. A machine learning algorithm utilises a dataset to develop a model which is further trained to make predictions based on new data that were not initially in the original dataset.

Tom Mitchell (1997) gave a more engineering-oriented definition of machine learning as follows:

"A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P improves with experience E."

The spam filter in an email is a typical example of a machine learning program. The system learns to classify mails flagged by the users as spam, but the regular mails are classified as non-spam mails. In this example, the flagged emails that the system uses to learn is referred to as the training set while each training example is called a training instance (or sample).

Some other examples of machine learning tasks are as enumerated as follows:

1. Analysing images of products on a production line to automatically classify them using convolutional neural networks (CNN).
2. Detection of credit card fraud through an anomaly detection.
3. Recommending a product, a customer may be interested in buying, based on his/her past purchases and some other known

information about the client. This becomes possible through an artificial neural network (ANN).

4. Creating a chatbot or a personal assistant using natural language processing (NLP), natural language understanding (NLU) and some question-answering modules.
5. Detecting tumours in brain scan using convoluted neural networks (CNN). In this case, there is a semantic segmentation of the scanned image and each pixel in the image is classified.

Machine learning can be classified based on the amount of supervision they get during training as supervised, unsupervised, semi supervised and reinforcement learning depending on the level of involvement of human supervision.

2.4.1 Supervised Learning

In supervised machine learning, the training set that is fed into the algorithm includes the desired solutions which are otherwise called labels. The labelled datasets are classified by the algorithm. The weights of the input data are adjusted as they are being fed into the model until the model has been fitted appropriately; this occurs as a part of cross-validation process. Many organizations have been able to use supervised learning to solve a variety of real-world problems at scale, an example is classification of spam mails in a separate folder from a regular mail inbox. The spam filter is trained with many example emails along with their class (spam), and it must learn how to classify new emails.

Some of the most common supervised learning algorithms include:

1. K-Nearest Neighbours
2. Linear Regression

3. Logistic Regression
4. Support Vector Machines (SVMs)
5. Decision Tree and Random Forests
6. Neural Networks

2.4.1.1 K-Nearest Neighbours

K-nearest neighbour is otherwise called KNN algorithm. This algorithm classifies data points based on their proximity and association to other available data. It works on the assumption that similar data points can be found near each other. Distance between data points are calculated through Euclidean disctance, and the data points are assigned a category based on the most frequent average.

KNN has low calculation time and relatively easy to use, hence it is one of the preferred algorithms frequently used by data scientists. However, as the datasets grow, the processing times increases, and it subsequently become less appealing for classification task. It is most often used for image recognition task.

2.4.1.2 Linear Regression

Linear regression identifies the relationship between a dependent variable and one or more independent variables. It could be a simple or a multiple linear regression, depending on the number of independent variables. If there is only one independent variable, it is referred to as simple linear regression, but when there are more than one independent variable, it is referred to as multiple linear regression. The aim of a linear regression algorithm is to locate a line of best fit, and this is calculated through least squares method. As the name linear implies,

unlike other regression models, linear regression is always a straight line when plotted on a graph.

2.4.1.3 Logistic Regression

Logistic regression is used when the dependent variables are categorical unlike in linear regression where the dependent variables are continuous. Categorical variables are the variables that have binary output only, for example “yes” or “no” and “true” or “false” are a typical categorical variable. In summary, logistic regression seeks to solve mainly binary classification problem.

2.4.1.4 Support Vector Machines (SVMs)

A support vector machine (SVM) can be used for both linear and nonlinear classification, regression and even outlier detection. It is a highly powerful and versatile machine learning algorithm, and one of the most popular models in machine learning.

2.4.1.5 Decision Trees and Random Forest

Like support vector machines (SVMs), decision trees are also versatile machine learning algorithms that can perform both classification and regression tasks, and even multi-output tasks. They are capable of fitting complex datasets. decision trees and random forests are somehow inter-related. Decision trees are a fundamental component of random forests. Random forest can also solve both classification and regression problems. The “forest” references a collection of uncorrelated decision trees, which are merged to reduce variance and create more accurate data predictions.

2.4.1.6 Neural networks

Neural networks imitate the interconnectivity of the human brain through layers of nodes. Nature has inspired many inventions, and it seems logical to look at the brain's architecture for an inspiration on how to build an intelligent machine. It is this logic that sparked artificial neural networks (ANNs). An ANN is a machine learning model that is at the very core of deep learning. They are highly versatile, powerful, and scalable. They have the capacity to tackle large and highly complex machine learning task such as powering speech recognition services (Apples' Siri and Amazon's Alexa), classification of billions of images (Google images), recommendation of the best video to watch for millions of users everyday (YouTube) and so on. Each node is made up of inputs, weights, a bias (threshold), and an output. If an output value exceeds a given threshold, it activates the node and passes the data to the next layer in the network.

2.4.2 Unsupervised learning

Unsupervised machine learning uses machine learning algorithm to analyse and cluster unlabelled datasets. The algorithm infers hidden patterns or data grouping without any reference to knowns or labelled outcome and does not have a need for any human intervention. In unsupervised learning, there is no correct answer and there is no teacher. The algorithms are left to their own devices to discover and present the hidden interesting structures and patterns in the data. Unlike in supervised learning, unsupervised machine learning cannot be directly applied to a regression or classification problem because it has

no idea what the values for the output data might be. Unsupervised learning can instead be used to discover the underlying structure of the data. It can further be grouped into clustering, association problem and dimensionality reduction. Its capability to discover similarities and differences in information makes it the ideal solution for exploratory data analysis, cross-selling strategies, customer segmentation and image recognition. It is applicable in computer vision, medical imaging, anomaly detection, recommendation engines and customers personas.

2.4.2.1 Clustering

This data mining technique groups unlabelled dataset according to their similarities or differences. Raw and unclassified data are processed into groups which are represented by structures or patterns in their information. It may be exclusive, overlapping, hierarchical or probabilistic.

K-means clustering is an exclusive clustering method. In this method, data points are assigned into groups and K represents the number of clusters based on the distance from each group's centroid. Data points closest to a given centroid are clustered in the same category. A large K value means the groupings are smaller and with more granularity, whereas smaller K value implies that the grouping are larger and with less granularity. K-means is most frequently used in image segmentation, image compression, document clustering and market segmentation.

2.4.2.2 Association Rule

This is a rule-based method that is used to find relationship between variables in each dataset. It is commonly used market basket analysis and enables companies to better understand the relationship between different products. There are few algorithms that are used to generate association rules but the most used is apriori algorithms.

Apriori algorithms are used within transactional datasets to identify frequent collection of items and the likelihood of consuming a product given the consumption of another product.

2.4.2.3 Dimensionality reduction

The performance of a machine learning algorithm can be negatively impacted if the volume of the data is too large. Ordinarily, more data should generally yield a more accurate result, but oftentimes more data leads to overfitting which subsequently makes it difficult to visualise dataset. It is for this reason that dimensionality reduction was developed. Dimensionality reduction is a technique used when the number of features in each dataset is too large to process. It reduces the data inputs as much as possible while still preserving the integrity of the dataset as much as possible.

The commonly used dimensionality reduction methods are principal component analysis (PCA), singular value decomposition (SVD) and autoencoders.

As beneficial as unsupervised machine learning is, it also has some challenges too since it executes without any human intervention. Some of the challenges include:

1. Longer training times
2. Higher risk of inaccurate results
3. Computational complexity due to high volume of training data
4. Lack of transparency into the basis on which data was clustered
5. Human intervention may be needed to validate output variables

2.4.3 Semi-supervised learning

This sits in between supervised and unsupervised machine learning. It happens in a situation where one has many input data (X) and only some of the data is labelled as (Y). many real-world machine learning problems fall into this category. Summarily in semi-supervised machine learning, some data are labelled but most of it is unlabelled. A mixture of supervised and

Chapter three

3.1 Understanding the dataset.

Having imported all the necessary libraries, the dataset was imported as follows:

```
dataset = pd.read_csv("CO2_Emissions_Canada.csv")
```

Afterward, the attributes of the dataset were checked. The dataset has twelve features which are as follows:

Make of car, that is company that manufactured the car

Model of the car

Vehicle class, depending on their utility, capacity and weight.

Engine size

Number of cylinders

Transmission type with number of gears

Type of fuel used

Fuel consumption on the city roads

Fuel consumption on the highways

Combined fuel consumption for both city roads and highways (ltrs/100Km)

Combined fuel consumption for both city roads and highways (mpg)

CO_2 emission

It should be noted that CO_2 is our dependent variables while the rest of the features are the independent variables.

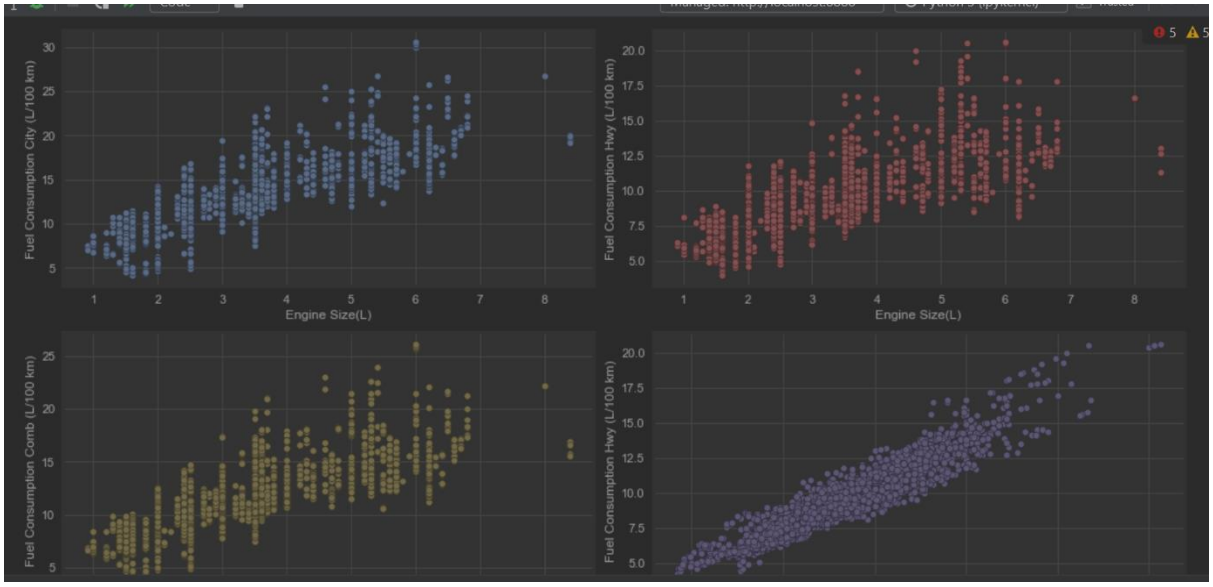
3.2 Data Visualisation

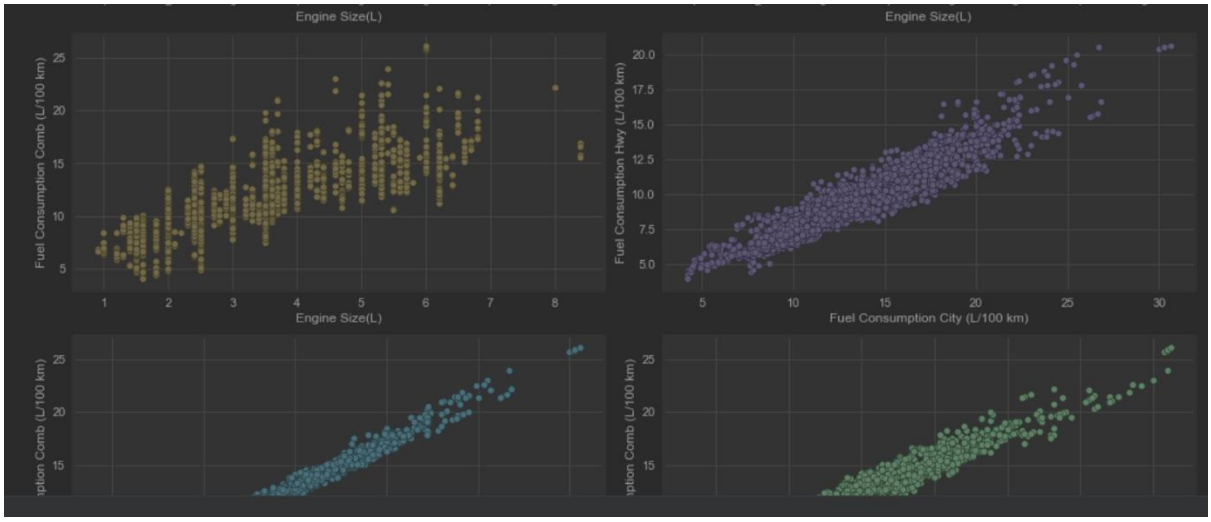
Having imported Autoviz library, a single line code was executed to carry out visualisation of the dataset used.

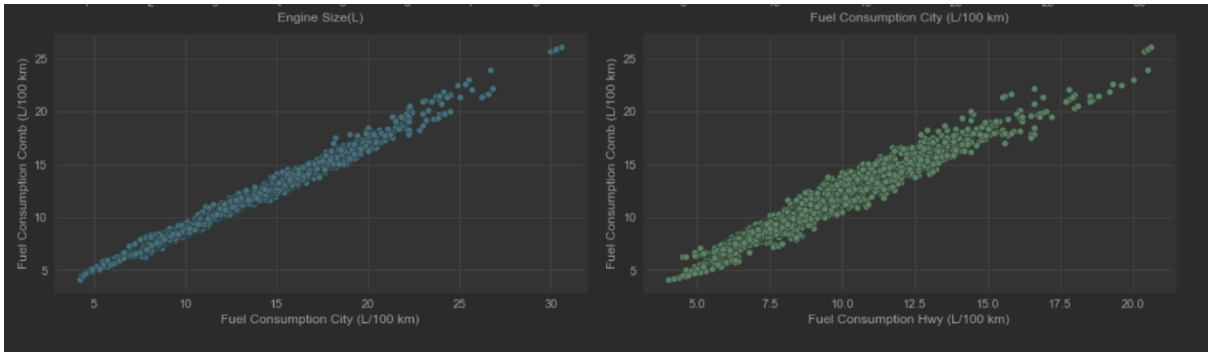
```
from autoviz.AutoViz_Class import AutoViz_Class
%matplotlib inline
AV = AutoViz_Class()
viz = AV.AutoViz('CO2_Emissions_Canada.csv', sep=',')
```

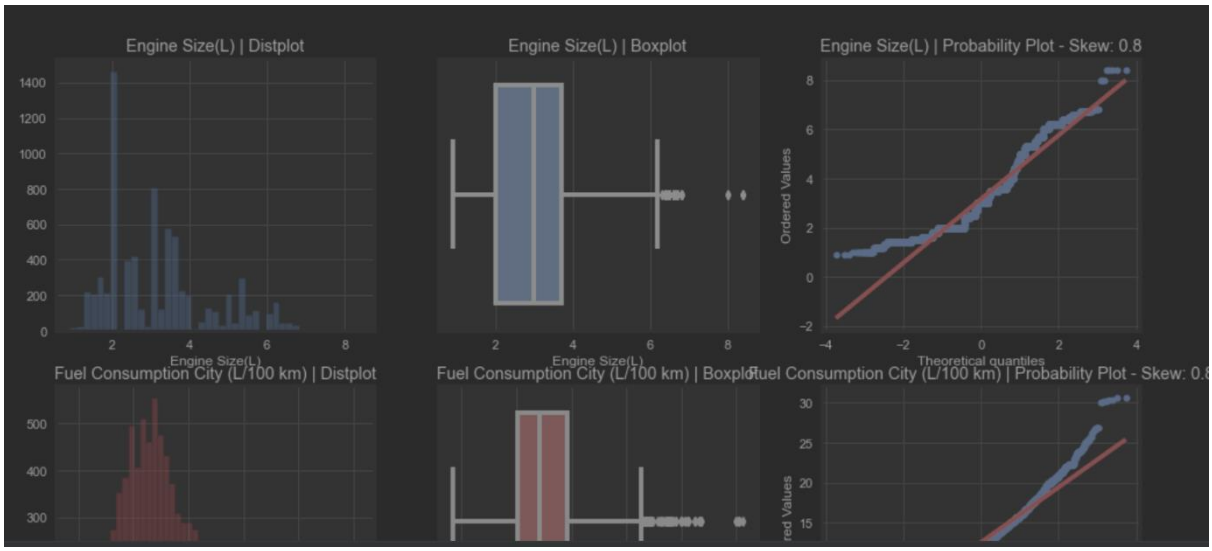
Snippet for visualisation code

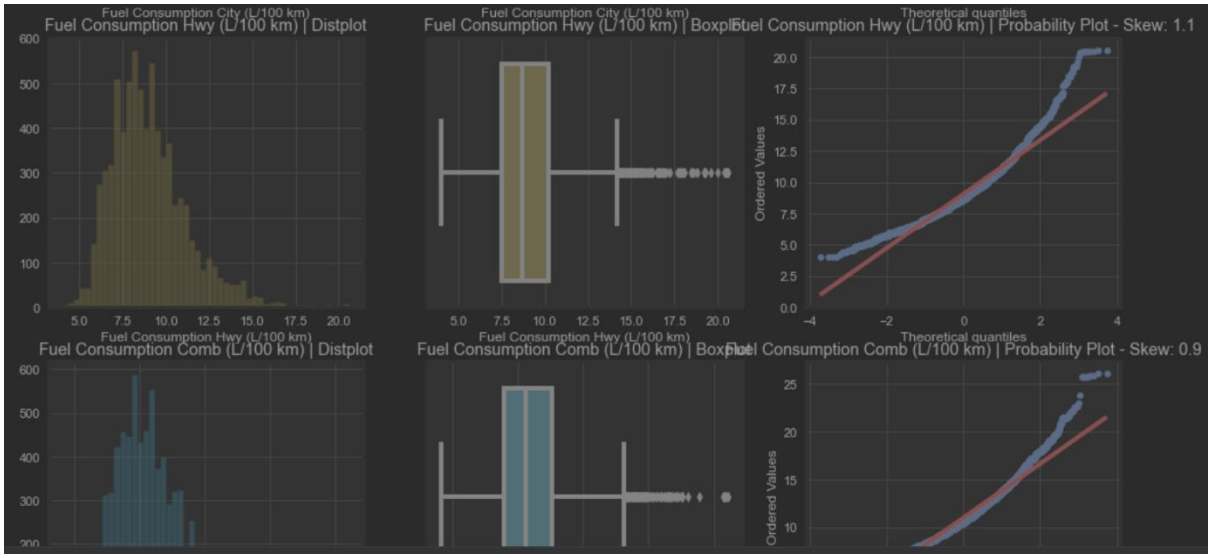
Below are some of the visualisations that were performed by the function.

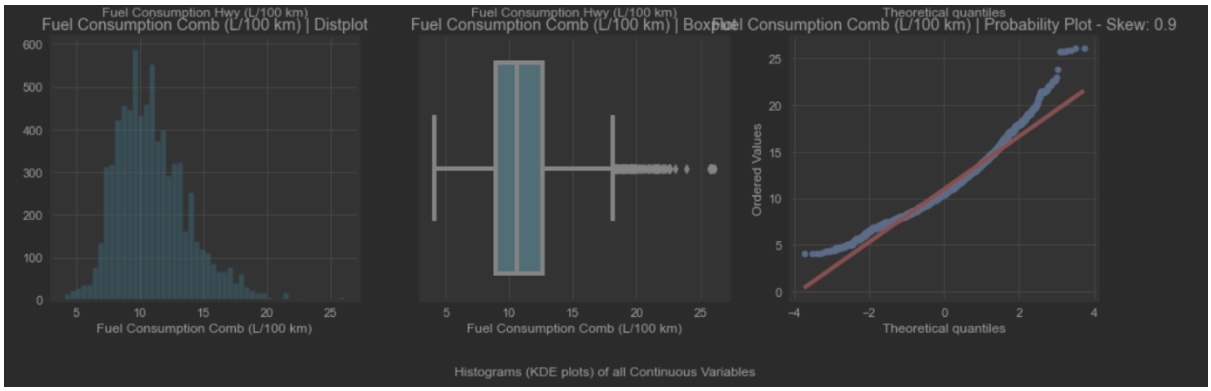


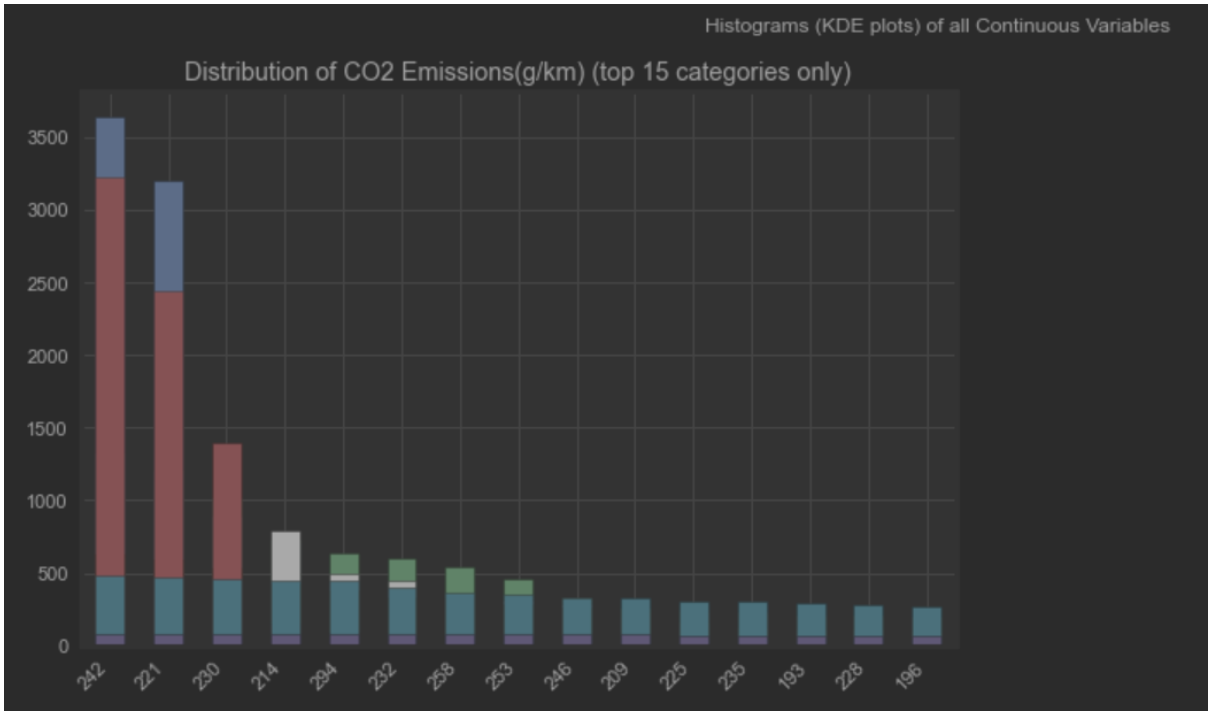

Engine Size(L)
Fuel Consumption Comb (L/100 km)
Engine Size(L)
Fuel Consumption Hwy (L/100 km)
Engine Size(L)
Fuel Consumption City (L/100 km)

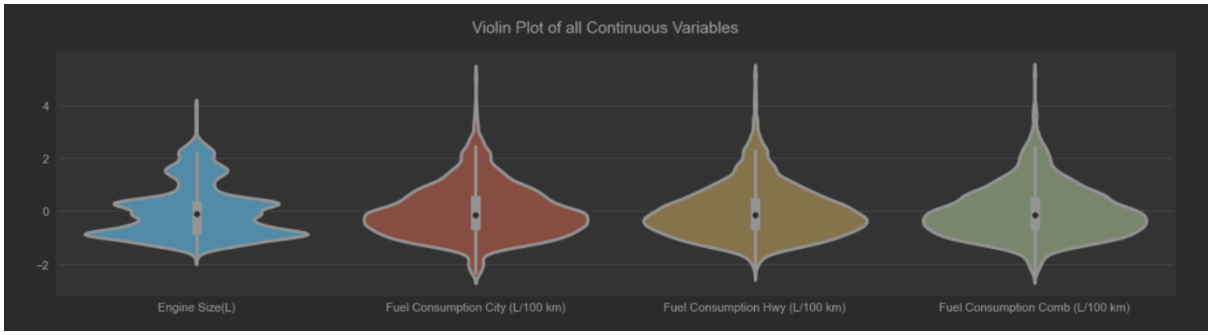

Engine Size(L)
Fuel Consumption Comb (L/100 km)
Fuel Consumption City (L/100 km)
Fuel Consumption City (L/100 km)
Fuel Consumption Comb (L/100 km)
Fuel Consumption Hwy (L/100 km)

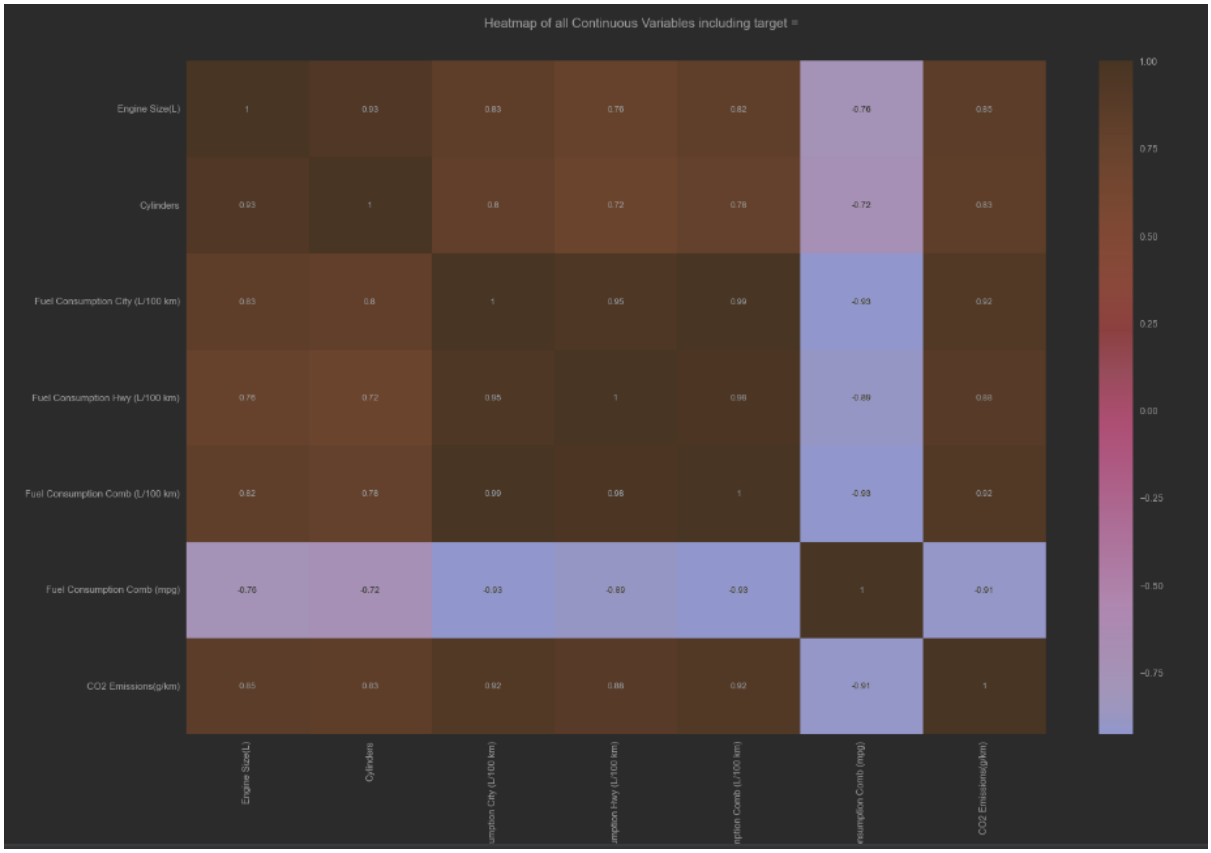

Engine Size(L) | Distplot
Engine Size(L) | Boxplot
Engine Size(L) | Probability Plot - Skew: 0.8
Ordered Values
Theoretical quantiles
Fuel Consumption City (L/100 km) | Distplot
Fuel Consumption City (L/100 km) | Boxplot
Fuel Consumption City (L/100 km) | Probability Plot - Skew: 0.8

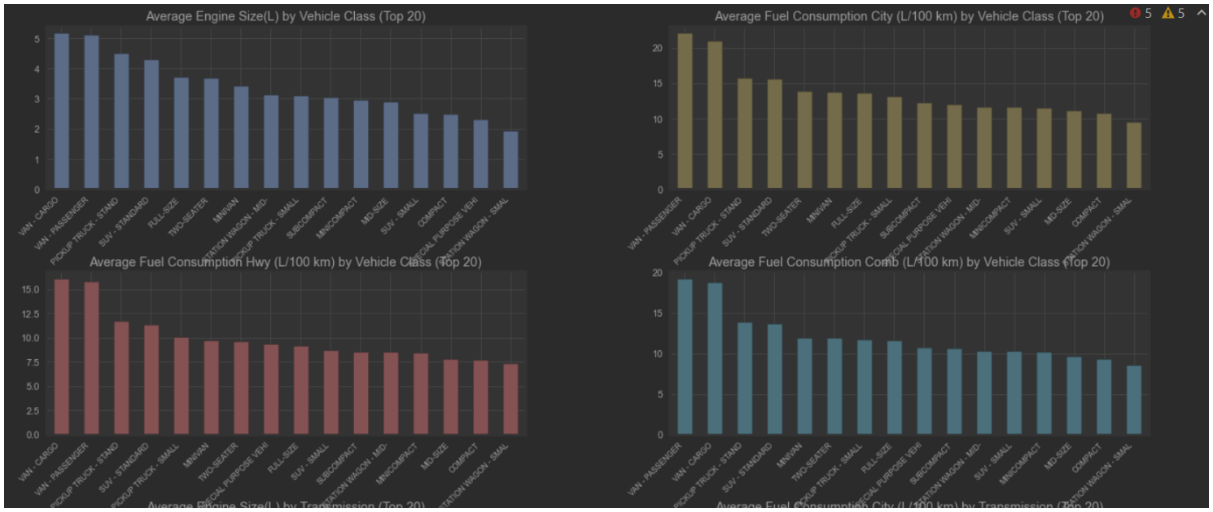

Fuel Consumption City (L/100 km)
Fuel Consumption Hwy (L/100 km) | Distplot
Fuel Consumption Hwy (L/100 km)
Fuel Consumption City (L/100 km)
Fuel Consumption Hwy (L/100 km) | Boxplot
Fuel Consumption Hwy (L/100 km)
Theoretical quantiles
Fuel Consumption Hwy (L/100 km) | Probability Plot - Skew: 1.1
Ordered Values
Theoretical quantiles
Fuel Consumption Comb (L/100 km) | Distplot
Fuel Consumption Comb (L/100 km) | Boxplot
Fuel Consumption Comb (L/100 km) | Probability Plot - Skew: 0.9
Ordered Values

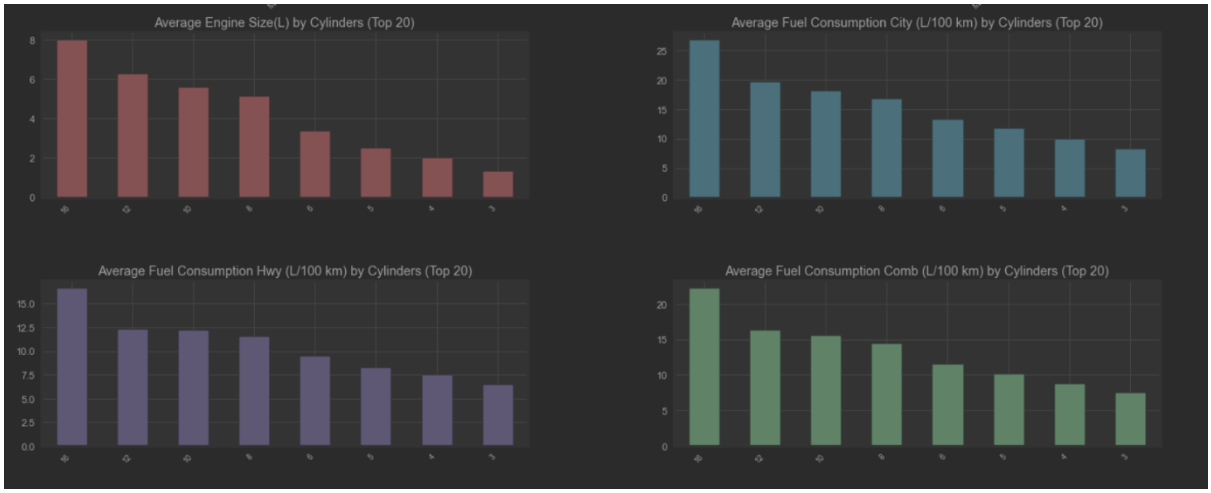

Fuel Consumption Hwy (L/100 km)
Fuel Consumption Comb (L/100 km) | Distplot
Fuel Consumption Comb (L/100 km)
Fuel Consumption Hwy (L/100 km)
Fuel Consumption Comb (L/100 km) | Boxplot
Fuel Consumption Comb (L/100 km)
Theoretical quantiles
Fuel Consumption Comb (L/100 km) | Probability Plot - Skew: 0.9
Ordered Values
Theoretical quantiles
Histograms (KDE plots) of all Continuous Variables

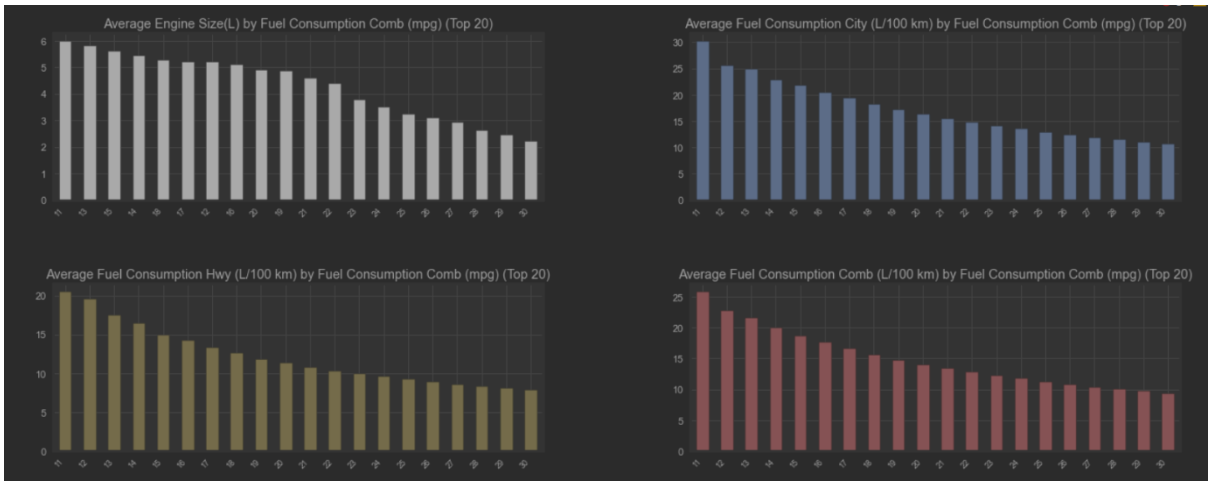

Histograms (KDE plots) of all Continuous Variables
Distribution of CO2 Emissions(g/km) (top 15 categories only)
3500
3000
2500
2000
1500
1000
500
0
242
221
230
214
294
232
258
253
246
209
225
235
193
228
196


Violin Plot of all Continuous Variables
4
2
0
−2
Engine Size(L)
Fuel Consumption City (L/100 km)
Fuel Consumption Hwy (L/100 km)
Fuel Consumption Comb (L/100 km)


Heatmap of all Continuous Variables including target =
Engine Size(L)
Cylinders
Fuel Consumption City (L/100 km)
Fuel Consumption Hwy (L/100 km)
Fuel Consumption Comb (L/100 km)
Fuel Consumption Comb (mpg)
CO2 Emissions(g/km)
1.00
0.75
0.50
0.25
0.00
-0.25
-0.50
-0.75


Average Engine Size(L) by Vehicle Class (Top 20)
Average Fuel Consumption City (L/100 km) by Vehicle Class (Top 20)
Average Fuel Consumption Hwy (L/100 km) by Vehicle Class (Top 20)
Average Fuel Consumption Comb (L/100 km) by Vehicle Class (Top 20)





3.1 Linear regression

Training model means setting of parameters so that the model best fits the training set. In linear regression, the motive is to find the minimum parameter vector that minimizes the root mean square error.

The dependent and the independents variables were first reduced into a 1X1 dimension and was then splited into training data and test data.

```
X = dataset['Fuel Consumption Comb (L/100 km)'].values.reshape(-1,1)
X
```

Snippet of independent variable “fuel consumption”

```
Y = dataset['CO2 Emissions(g/km)'].values
Y
```

Snippet of dependent variable “CO_2 emission”

```
X_train, X_test, y_train, y_test = train_test_split(X, Y, train_size=0.8,
test_size=0.2, random_state=2020)
X_train
```

Snippet of the code splitting dataset into training data and test data

Chapter four

4.1 Results and Discussion

It was observed that the higher the fuel consumption, the higher the carbon dioxide emission.

Carbon emission can be controlled to acceptable level that will not affect the environment

Chapter five

5.1 Conclusion and Recommendation

We do not have a second planet; hence we have got to do what it takes to save the only place we know as home. Data science and artificial intelligence have a critical role to play in tackling climate crisis and ultimately save the planet.

References

Berners-Lee, M. (2019): “There is no Planet B; A Handbook for the make or Break Years”, Cambridge University Press, Cambridge

Cline, W.R (1992): “The Economics of Global Warming” Washington DC: Institute for International Economics

EMC Education Services (2015): “Data Science and Big Data Analytics; Discovering, Analyzing, Visualizing and Presenting Data”, John Wiley & Sons, Indianna

Fang, D., X. Zhang, Q. Yu, T.C. Jin, L. Tian (2018): “A novel method for carbon dioxide emission forecasting based on improved Gaussian processes regression”,

Geron, A. (2019): “hands-on Machine with Scikit-learn, Keras and TensorFlow”, 2nd edition, O’Reilly, Farnham

Hansen, J., m. Sato, R. Ruedy et al. (2006): Global temperature change, Proceedings of the National Academy”

Henson, R. (2006): “The Rough Guide to Climate Change”, Rough Guides, London

Hill, M.K. (2004): “Understanding Environmental Pollution”, 2nd edition, Cambridge University Press, Cambridge

Maslin, M. (2006): “Global warming: a very short introduction” Oxford University Press, New York

Schar, C., P.L. Vidale, D. Luthi, et al. (2004): “The role of increasing temperature variability in European summer heatwaves” Nature

Stern, N. (2019): The Economics of Climate Change (The Stern Review). University Press, Cambridge

(https://www.metoffice.gov.uk/about-us/press-office/news/weather-and-climate/2022/red-extreme-heat-warning-ud).