**Logo

Description automatically generated**

**Module 4 – Group Assignment**

[**CAPSTONE PROJECT - EXPLORATORY**](https://northeastern.instructure.com/courses/144023/assignments/1800508) **DATA ANALYSIS**

**Group Members:**

**Mehadi Hussain Meer (002192125)**

**Aditi Padhye (002950010)**

College of Professional Studies: Northeastern University

ALY 6140 - Analytics System Technology

Instructor: Dr. Alex Maizlish

May 7th, 2023.

**INTRODUCTION**

Vehicles are an integral part of our lives and play a crucial role in transportation. With the increase in the number of vehicles, there is a corresponding increase in the consumption of fuel. Fuel consumption is a major concern not only for individual vehicle owners but also for society as a whole, as it has significant implications for the environment and the economy. Therefore, it is essential to study and understand the fuel consumption patterns of different types of vehicles, which can be accomplished through the use of analytics. With the help of analytics, we can identify the factors that contribute to higher fuel consumption. We can also predict the fuel consumption rates, Co2 emission rates by using various aspects related to that particular vehicle.

**About the Dataset:**

* The Fuel Consumption Ratings data set is a valuable resource for analyzing and understanding the fuel consumption patterns and carbon emissions of light-duty vehicles sold in Canada. This data set is open to the public and is available on the website of the Canadian government under the data section titled "Fuel Consumption Ratings." The data set is published by Natural Resources Canada and provides model-specific fuel consumption ratings and estimated carbon dioxide emissions for new light-duty vehicles for retail sale in Canada. The Fuel Consumption Ratings data set is available from the 1990s to the present day, with individual data sets for each year. The data set is released under the Open Government Licence - Canada, which allows users to freely access, use, and distribute the data for any purpose.
* For our project, we have chosen to select and gather data from the years 2000-2023, which results in a data set of almost **10,000+ rows/instances and across 15 features/variables**. These features include vehicle make and model, fuel consumption ratings, estimated CO2 emissions, engine size, transmission type, and other relevant parameters.
* The Fuel Consumption Ratings data set provides a wealth of information that can be used for analyzing and understanding the fuel efficiency and environmental impact of different types of vehicles. With the help of this data set, we can develop predictive models to estimate fuel consumption and CO2 emissions, identify trends and patterns in fuel efficiency and emissions, and explore the impact of various factors on vehicle fuel efficiency.

**Reason for choosing the dataset:**

* The data set on firsthand has enough number of categorial and numerical variables for analytics purposes. This data perfectly aligned with our course load in establishing thorough EDA, prepare proper data visualizations and build models in Python.
* The predictive models built can be used to estimate the fuel consumption and CO2 emissions of a vehicle before its production, which can help vehicle manufacturers in designing more fuel-efficient and eco-friendly vehicles.
* These models can also be used by individual vehicle owners to estimate the fuel cost and carbon footprint of their vehicle under different driving scenarios.
* Moreover, these models, EDA conclusions and all insights generated can also be used by policymakers and government organizations to formulate policies and regulations to promote fuel efficiency and reduce carbon emissions from the transportation sector.

**Few of the Investigation Questions we look for:**

1. What are main contributing factors to Fuel Consumption & also Co2 Emissions ?
2. How has the average CO2 emission changed over the years?
3. Is there a correlation between CO2 emissions and fuel consumption in the city?
4. Is there a significant difference in smog ratings between automatic and manual transmission vehicles,
5. Any hidden insights about total combined fuel consumption for each vehicle class and when breakdown by each fuel type?
6. How many vehicles fall into each rating category?
7. Are Higher CO2 Emissions Ratings More Common Among Certain Vehicle Classes?
8. Comparing Smog Ratings for Different Fuel Types in Vehicles: Which Fuel Type Performs Better in Terms of Smog Emissions?
9. How accurately can predict the fuel consumption combination or Co2 emissions or Co2 ratings or Smog ratings of a new vehicle based on its make, model, engine size and other parameters?

**List of models we plan to work on :**

* Linear Regression – The reason we plan to work on this model is because the linear regression can be used to develop a simple and interpretable model that can capture the linear relationships between the input features and the fuel consumption or CO2 emissions of the vehicle. This model is planed in-case we find suitable relationships between input and output variables are expected to be linear.
* Decision Tree Regression - This model can capture the non-linear relationships between the input features and the fuel consumption or CO2 emissions of the vehicle. This model is planed in-case we find suitable relationships between input and output variables are expected to be non-linear or complex.
* Random Forest Regression - This model can capture both linear and non-linear relationships between input features and output variables by combining multiple decision trees. This model is suitable when the relationships between input and output variables are expected to be complex and non-linear, and when the data set has a large number of input features.

**Note/ Key Points:**

* Fuel Consumption Ratings Dataset of each year are CSV files which will be combined.
* Years considered 2000-2023.

*Data fields description* :

|  |  |  |
| --- | --- | --- |
| **Variables** | **Description** | **Data type** |
| model\_year | Model year of the vehicle. The combined file includes year from 2010 to 2023. | Numerical |
| make | The brand of the vehicle. For example – Nissan Altima, Nissan is the make and Altima is the model | Categorical |
| model | Specific model of the vehicle. For example – Nissan Altima, Nissan is the make and Altima is the model | Categorical |
| vehicle\_class | It refers to the different category of the vehicle. Like SUV, Sedan, Compact, etc. | Categorical |
| engine\_size\_(l) | Size of the engine in the vehicle. | Numerical |
| cylinders | Number of cylinders in the vehicle. | Numerical |
| transmission | Different types of transmission. where A stands for automatic, AM stands for automated manual, AS stands for automatic with select shift, AV stands for continuously variable, and M stands for manual. Additionally, 3-10 indicates the number of gears in the transmission. Like A4 – Automatic with 4 gears, M6 – Manual with 6 gears, and so on. | Categorical |
| fuel\_type | Type of fuel used by the vehicle where X stands for regular gasoline, Z stands for premium gasoline, D stands for diesel, E stands for ethanol (E85), and N stands for natural gas. | Categorical |
| fuel\_consumption\_city\_(l/100\_km) | City fuel consumption rating shown in Litre per 100 kilometers in urban driving. | Numerical |
| hwy\_(l/100\_km) | Highway fuel consumption rating shown in Litre per 100 kilometers in a mix of open highway and rural road driving, typical of longer trips. | Numerical |
| comb\_(l/100\_km) | Combined fuel consumption rating shown in Litre per 100 kilometers with 55% city driving and 45% highway driving. | Numerical |
| comb\_(mpg) | It shows combined Miles per gallon (mpg) used by the vehicle. The distance a vehicle can travel per gallon of fuel. | Numerical |
| co2\_emissions\_(g/km) | The amount of carbon dioxide emitted by a vehicle's tailpipe during a combined drive-in city and highway conditions is expressed in grams per kilometer (g/km). | Numerical |
| co2\_rating | The carbon dioxide emissions from the vehicle's exhaust are evaluated using a rating system that ranges from 1 (worst) to 10 (best). | Numerical |
| smog\_rating | The emissions of smog-forming pollutants are also rated on a scale of 1 to 10, where 1 represents the highest emissions and 10 represents the lowest emissions. | Numerical |

**DATA PREPARATION**

**Data Cleaning & Merging of CSV Files:**

Since the data is in multiple csv file the data preparation is required :

Text

Description automatically generated

*Fig:1*

The **remove\_empty\_cells** function is used to clean up data by removing any empty cells from a row of the dataset. If the value is not an empty string (i.e., ''), the value is appended to the new\_list using the append() method. Finally, the function returns the new\_list that contains only the non-empty values of the original row.

Graphical user interface, text, application

Description automatically generated

*Fig: 2*

The **append\_data\_rows** will append rows of data from a CSV file to an existing list of data. It takes two arguments: a CSV reader object, and an existing list of data. Within the loop, it reads the next row of data from the CSV reader using the next() method. Then, it checks if any values are present in the row using the any() function. If the row contains values, the function removes any empty cells in the row using the remove\_empty\_cells() function and appends the cleaned-up row to the existing list of data. If the row is empty (i.e. any(row) returns False), the loop is broken using the break statement. Once all rows have been read and appended to the data list, the function returns the updated data list.

Text

Description automatically generated

*Fig - 3*

The **remove\_special\_characters** function takes a string as input and returns a new string with specific special characters ('@','#','$','\*','&') removed from it. The input string is then assigned to a new variable normal\_string to avoid modifying the original input string. A loop is used to iterate through each character in the special\_characters list. During each iteration of the loop, the replace() method is used to replace the current special character with an empty string in the normal\_string. This removes the special character from the normal\_string. Finally, the modified normal\_string is returned as the output of the function.

Text

Description automatically generated

*Fig - 4*

The Convert\_encoding function takes a filename as input, reads the contents of the file using the original encoding, and then writes the contents of the file to a new file using a different encoding. codecs.open() function opens the input file with the 'r' mode for reading and the original encoding of the file, 'windows-1252'. The contents of the input file are then read using the read() method and assigned to a variable called contents. Next, the function opens the output file using the codecs.open() function with the 'w' mode for writing and the new encoding, 'utf-8'. The contents of the input file, which are stored in the contents variable, are then written to the output file using the write() method.

Text

Description automatically generated

*Fig – 5*

This code block initializes with an empty list to store the data frame. Then, it loops through first five rows of each csv file using csv.reader. And then the first two rows are cleaned up to make consistent headers with other files. It is followed by removal of special characters and empty cells from the headers. After, making the headers uniform for all the files data rows were appended using append data row’s function. Then, the data for each year appended is stored in a different data frame. We checked the list of columns from all the data frames.

Graphical user interface, application

Description automatically generated with medium confidence

Graphical user interface, application

Description automatically generated

*Fig – 6*

Finally, the data frames were concatenated using concat function. combined\_df show the combined data from all files. There are 14,523 rows and 15 columns after combining all the data frames.

**Missing value imputation** :

#Check missing values :

Table

Description automatically generated

*Fig – 7*

Missing values from all the data columns were computed using isnull() and sum() function. Only co2\_rating and smog\_rating has missing values.

Graphical user interface, text, application, Word

Description automatically generated

*Fig – 8*

Co2\_rating and smog\_rating was stored in a separate data frame and first five elements were checked. We can see there are null values in the rows.

#Imputing:

Text

Description automatically generated

*Fig – 9*

We noticed that imputing value using simple imputation method would not be useful as the entire column for co2\_rating and smog\_rating for the initial years i.e. 2010, 2011, 2012 were missing, and a logical imputation should be performed because co2 and smog rating might dependent on other features of the dataset like Make, transmission, fuel type, and so on. So, we have imputed the missing values in co2\_rating and smog\_rating by K-nearest neighbour imputer method, which is based on KNN Machine learning algorithm. Also, we observed that there are only 6.4% missing values of co2\_rating and smog\_rating for those years, so it is a very small number of missing data which can be imputed by using any statistical method. As the ratings are whole number, so we rounded off the values imputed to nearest whole number.

Graphical user interface, text

Description automatically generated

*Fig – 10*

#Let us the missingness after assigning imputed values in the co2\_rating and smog\_rating.

Text

Description automatically generated with medium confidence

*Fig – 11*

We could clearly see : After imputation there are no missing values in the dataset.

Text, table

Description automatically generated

*Fig – 12*

The above are the changed data types of few columns from object for the analysis purpose.

**EXPLORATORY DATA ANALYSIS**

**#Descriptive Statistics:**

Table

Description automatically generated

Fig:13

The **.describe()** method provides basic statistical measures such as count, mean, standard deviation, minimum, maximum, and quartiles for each numeric column in the data frame. The **.T** method transposes the summary statistics table so that the columns become rows and vice versa, making it easier to read and compare the statistics for each variable. The resulting table shows that the data set contains 14,523 observations for each numeric variable, including engine size, number of cylinders, fuel consumption in the city, on the highway, and combined, CO2 emissions, CO2 rating, and smog rating.

**Insight :** The above summary statistics provide valuable insights into the distribution and central tendencies of the variables in our dataset. One interesting observation is that the mean engine size is 3.27 liters, which is quite large and suggests that the dataset may be skewed towards larger vehicles. The most interesting finding from this summary is the range of CO2 emissions, which vary widely from 94 grams per kilometer to a staggering 608 grams per kilometer, highlighting the need for more environmentally friendly transportation options. Another interesting finding is that the minimum smog rating is 1, indicating that some vehicles in the dataset emit extremely high levels of harmful pollutants, while the maximum smog rating is 8, indicating that there are also some vehicles with very low levels of smog emissions. It is also noteworthy that the mean CO2 and Smog ratings are below 5, indicating that most vehicles in the dataset have a relatively low CO2 emission rating and Smog ratings.

**#Data Visualizations:**

Let’s Check the correlations between the variables :

Chart

Description automatically generated*Fig – 14*

The correlation plot is to visualize the correlation structure of a dataset using a heatmap. The Spearman correlation matrix is useful for identifying non-linear relationships between variables, and the Seaborn library provides many customization options to create informative and visually appealing heatmaps. A heatmap is created for the lower triangle of the spearman correlation matrix which includes all the features of the dataset and shows their relationship between 0 to 1.

* The fuel consumption city variable is highly negatively correlated (-0.61) with co2 rating. Similarly, with highway fuel consumption (-0.62) and combined fuel consumption (-0.63) there is high negative correlation with co2 rating. If fuel consumption as well as the co2 rating is less then the vehicle is using less fuel per 100 kilometers and emitting less carbon dioxide from the vehicle, hence it is a good performing vehicle.
* Fuel Consumption Combined miles per gallon have a very high negative correlation (-0.96) with co2 emission (grams/kilometers). The high negative correlation between combined miles per gallon and CO2 emissions (grams/kilometers) indicates that as the fuel efficiency of a vehicle increases, its CO2 emissions tend to decrease. This is because the amount of CO2 released during the combustion of fuel is directly proportional to the amount of fuel burned. When a vehicle has a higher fuel efficiency and gets more miles per gallon, it means that it can travel the same distance while burning less fuel.
* Co2 rating is negatively correlated (-0.63) with Fuel Consumption combined miles per gallon. CO2 rating is the combined fuel consumption expressed in miles per gallon, indicating that vehicles with higher CO2 ratings tend to have lower fuel efficiency.
* Co2 and smog rating are less positively correlated (0.46). CO2 emissions are primarily a by-product of the combustion of fuel, whereas smog is formed by the reaction organic compounds in the presence of sunlight. The factors that contribute to the formation of smog are complex and can include both vehicle-related and non-vehicle-related sources, such as industrial emissions, agricultural activities, and weather conditions.
* There is a high positive correlation (0.96) between engine size and cylinder.

**#let’s check correlations of Co2 ratings deeply with prioritization :**

A picture containing bar chart

Description automatically generated

*Fig – 15 Co2 rating correlation*

It was found that the CO2 rating is the most strongly correlated with CO2 emissions, indicating that vehicles with higher CO2 ratings tend to emit more CO2. The second most correlated variable with CO2 rating is the combined fuel consumption expressed in litres per kilometer, suggesting that vehicles with higher CO2 ratings tend to have higher fuel consumption rates. In contrast, the smog rating appears to be the least correlated variable with CO2 rating, indicating that the CO2 rating may not be a good predictor of a vehicle's smog-forming emissions.

**#let’s check correlations of Smog ratings deeply with prioritization :**

Background pattern

Description automatically generated

*Fig – 16 smog rating* correlations

It was found that the smog rating is most strongly correlated with the CO2 rating, indicating that vehicles with higher CO2 ratings may also have higher levels of smog-forming emissions. The second most correlated variable with smog rating is CO2 emissions, suggesting that vehicles with higher levels of CO2 emissions may also have higher levels of smog-forming emissions. In contrast, highway fuel consumption expressed in liters per 100 kilometers appears to be the least correlated variable with smog rating, suggesting that the smog rating may not be a good predictor of a vehicle's fuel efficiency on the highway.

**Counts bar plot of number of cars by model year.**

Chart, bar chart

Description automatically generated

*Fig– 17*

The count bar plot of the number of cars by model year reveals some interesting insights about the trends in the automotive industry. Firstly, there was a noticeable rise in the number of cars produced in 2015, which could be attributed to various factors such as increasing demand, favorable economic conditions, or new technologies that made manufacturing more efficient.  However, the barplot also shows a significant drop in the number of cars produced between 2020 and 2022, which is likely due to the impact of the COVID-19 pandemic on the automotive industry. The pandemic caused disruptions in supply chains, reduced demand, and forced many factories to shut down, all of which had a negative impact on production.  Furthermore, the barplot reveals a sudden decrease in the number of cars produced in 2023, which can be attributed to the global semiconductor shortage. The shortage has affected various industries, including the automotive industry, as semiconductors are used in various vehicle components such as infotainment systems, power steering, and braking systems.

**How has the average CO2 emission changed over the years?**

Chart, line chart

Description automatically generated

*Fig : 18*

**Insight :** The line plot above shows the average CO2 emissions of the vehicles in the dataset over the years. It is evident from the plot that there was a significant drop in the CO2 emissions in the year 2015, but it increased again in the following years. However, there is a notable decrease in CO2 emissions in 2023, which could be attributed to the shortage of semiconductors prevailing in the market. This shortage has led to a slowdown in the production of new vehicles, which in turn has resulted in a drop in CO2 emissions.

**Investigation of drops and raise in Co2 emissions respect to model year - through fuel type dependency.**

Chart, bar chart

Description automatically generated

*Fig : 19*

**Insight :** The above grouped bar plot visualization is created to investigate the changes in CO2 emissions based on fuel type and model year. The analysis highlights that the drop in CO2 emissions in 2015 could be attributed to changes in regulations implemented by the Canadian government in Heavy-duty Vehicle and Engine Greenhouse Gas Emission Regulations (Legislative Services Branch, 2022). However, further exploration reveals that the drop in emissions could also be related to a significant decrease in the count of premium and regular gasoline vehicles, which are known to contribute significantly to CO2 emissions. This highlights the impact of both regulatory changes and consumer behavior in the reduction of CO2 emissions.

**Is there a correlation between CO2 emissions and fuel consumption in the city by fuel type?**

Chart, scatter chart

Description automatically generated

*Fig : 20*

**Insight :** Based on the above scatterplot, it can be observed that there is a positive correlation between fuel consumption and CO2 emissions, as expected. However, there is also a clear distinction between the ethanol fuel type and the other fuel types in terms of their CO2 emissions and fuel consumption. Ethanol fuel has significantly lower CO2 emissions and fuel consumption compared to the other fuel types. Additionally, there is a noticeable delay in the fuel consumption and CO2 emission points of some fuel types compared to others, indicating that certain fuel types are more efficient than others in terms of fuel consumption and CO2 emissions.

**Any hidden insights about total combined fuel consumption for each vehicle class and when breakdown by each fuel type?**

**Chart

Description automatically generated**

*Fig : 21*

For the above visualization, the **groupby** method is used to group the data by **vehicle\_class** and **fuel\_type**, and the **mean** method is used to calculate the mean fuel consumption for each group. The resulting data is then converted to a pivot table using the **unstack** method, and a horizontal stacked bar chart is generated using the **plot** method.

**Insight :** According to the above horizontal stacked bar chart of combined fuel consumption by vehicle class and fuel type, it can be seen that Pickup Truck Standard, SUV, SUV Standard, SUV Small, Midsize Vehicle, Van Passenger, and Full-Size Vehicle have crossed the average fuel consumption of 40 L/100 km. This indicates that these vehicle classes may not be as fuel-efficient as other classes, and fuel consumption may be a significant cost factor for individuals or companies operating such vehicles. It is worth noting that most of these vehicle classes crossed the average with both regular and premium gasoline options, which may be reason for impact of fuel efficiency and overall cost of ownership.

**Range of ratings and how many vehicles fall into each rating category?**

**Chart, histogram

Description automatically generated**

**Chart, histogram

Description automatically generated**

*Fig : 22*

**Insight :** The above histogram plots of CO2 and smog ratings reveal some interesting insights about the distribution of these ratings in the dataset. The majority of vehicles have a CO2 rating between 3 to 7, with over 7500 vehicles having a rating of 5 and over 2000 vehicles having a rating of 4. On the other hand, the smog ratings are more widely spread, with a significant number of vehicles having a rating of 1, 3, 5, 6, or 7. Although the most common smog rating is 5, with over 8000 vehicles having this rating, around 1800 vehicles have a rating of either 3 or 6. This variation in ratings could be due to differences in the number of cylinders or other factors.

**Is there a significant difference in smog ratings between automatic and manual transmission vehicles?**

Chart, bar chart

Description automatically generated

*Fig : 23*

For the above Visualization, a function is defined to map smog ratings to two levels, "Below 3" and "Above 8" (**map\_smog\_rating()**). A new column is added to the data frame with the mapped smog ratings (**apply()**). The dataframe is filtered to contain only rows with manual and automatic transmission (**str.match()**). The filtered dataframe is then grouped by transmission and smog rating level (**groupby()**), and the count of smog ratings is calculated for each group (**count(), unstack()**). The resulting dataframe is then plotted as a stacked bar chart (**plot(kind='bar', stacked=True)**), with the x-axis showing transmission type and the y-axis showing the count of vehicles.

**Insight :** The above stacked bar plot represents the smog rating distribution of cars based on their transmission type, either manual or automatic. The data is divided into two categories, cars with smog ratings below 3 and above 8. The blue bars represent the number of cars with smog ratings below 3, while the orange bars represent cars with smog ratings above 8. The graph indicates that most of the automatic cars have smog ratings above 8, while the there are observable of the manual cars that have smog ratings between less than 3 . This is an important insight that can help us understand the distribution of smog ratings and how it relates to the type of transmission. Furthermore, this means that there are more automatic cars with high smog ratings than manual cars. This information can be useful for policymakers and car manufacturers to develop and implement measures to reduce air pollution caused by vehicles.

**Are Higher CO2 Emissions Ratings More Common Among Certain Vehicle Classes?**

Chart, bar chart

Description automatically generated

*Fig : 24*

For the above visualization, a function called **map\_co2\_rating()** is defined to map CO2 ratings to two levels, "Below 5" and "Above 5". This function is applied to the **co2\_rating** column of the dataframe, and a new column called **co2\_rating\_level** is added to the dataframe with the mapped CO2 ratings. The dataframe is then grouped by **vehicle\_class** and **co2\_rating\_level** using the **groupby()** method, and the count of CO2 ratings is calculated for each group using the **count()** method. The resulting dataframe is reshaped using the **unstack()** method to have the **co2\_rating\_level** values as column headers. The **columns** attribute is then used to rename the columns to "Below 5" and "Above 5". Finally, a stacked bar plot is created using the **plot()** method.

**Insight :** The above analysis of CO2 emissions ratings by vehicle class reveals that compact, full size, mid size, mini compact, station wagon small, SUV small, sub compact, and two-seater are the vehicle classes that have CO2 ratings below 5. One possible reason for this could be that these vehicle classes generally have smaller engines and are more fuel-efficient compared to other larger vehicles and hence have worst ratings. Another reason could be that these vehicle classes may have more advanced emissions control systems that help to reduce the emissions and improve the fuel efficiency.

**Comparing Smog Ratings for Different Fuel Types in Vehicles: Which Fuel Type Performs Better in Terms of Smog Emissions?**

**Chart, box and whisker chart

Description automatically generated**

*Fig : 25*

For the above visualization, a list of smog ratings was defined using the list constructor and filtered using the **isin()** method to include only those ratings in the dataframe. The smog ratings were then categorized as "below 5" or "above 5" using a lambda function with the **apply()** method. Then, **sns.violinplot()** method from the Seaborn library was used to create the plot.

**Insight :** The above violin plot comparing smog ratings for different fuel types in vehicles indicates that diesel (fuel type D) and premium gasoline (fuel type Z) vehicles had mostly lower smog ratings of 1 (worst), 3, and 5. Regular gasoline vehicles (fuel type X), on the other hand, had most smog ratings of 3. It is important to note that smog ratings are not directly proportional to smog emissions, but rather represent a combination of emissions of pollutants that form smog, including nitrogen oxides and volatile organic compounds. The observed difference in smog ratings by fuel type could be due to a variety of factors, such as differences in the technology used to reduce emissions in each type of engine, differences in the fuels themselves, and differences in driving patterns of the vehicles. Further investigation into the specific factors contributing to these differences in smog ratings would be necessary to draw any conclusive insights.

**Data Modeling**

We imported necessary libraries, including pandas, numpy, seaborn, and matplotlib, for data manipulation, visualization, and modeling.

**Business question:** How can we predict the fuel consumption (comb\_(l/100\_km)) of a vehicle based on its characteristics in the dataset?

**Method 1: Linear Regression**

In this analysis, we aimed to predict the fuel consumption (comb\_(l/100\_km)) of a vehicle based on its characteristics using the linear regression method in Python. Linear regression is a statistical technique used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. By leveraging a dataset containing various vehicle characteristics, we applied linear regression to predict the fuel consumption and evaluate the model's performance.

**Methodology:** Categorical columns, such as 'make,' 'transmission,' 'fuel\_type,' 'vehicle\_class,' 'model\_year,' and 'model,' were selected for one-hot encoding. One-hot encoding was performed on the categorical columns using the pd.get\_dummies() function to convert them into numerical representation, enabling their usage in the regression model. The dataset was split into features (X) and the target variable (y). Features included all columns except 'comb\_(l/100\_km),' while the target variable was 'comb\_(l/100\_km).'

**Training and Testing Sets:** The dataset was divided into training and testing sets using the train\_test\_split() function from the sklearn library. The data was split with a test size of 20% and a random state of 42 to ensure reproducibility.

**Linear Regression Modeling:** We initialized a linear regression model using the LinearRegression() function from sklearn.linear\_model. The model was then fitted to the training data using the fit() function, allowing it to learn the underlying relationship between the features and the target variable.

**Prediction and Evaluation:** The trained model was used to predict the fuel consumption for the test set using the predict() function. Mean squared error (MSE) and R-squared (R^2) were calculated to evaluate the performance of the model. MSE measures the average squared difference between the predicted and actual values, providing insight into the model's accuracy. R^2 represents the proportion of the variance in the target variable explained by the model. It ranges from 0 to 1, with higher values indicating a better fit.

**Results and Analysis:**

After running the linear regression model, we obtained the following results:

The obtained mean squared error (MSE: 8032192687883653.0) and R-squared (R^2: -805627318846502.9) values indicate the model's performance in predicting fuel consumption. A lower MSE value signifies better predictive accuracy, while a higher R^2 value suggests a stronger fit to the data.

A picture containing screenshot, text, line, diagram

Description automatically generated

*Figure – X*

**Interpretation of plot:**

The scatter plot showcases the relationship between the actual fuel consumption (comb\_(l/100\_km)) and the predicted fuel consumption values from our model. Each data point represents a specific vehicle's characteristics. Ideally, the points would align closely along a straight line, indicating a strong correlation between the predicted and actual values. However, in our plot, we observe that the data points are scattered, and most of them do not align near the line. This suggests that our linear regression model may not adequately capture the complexities of the relationship between vehicle characteristics and fuel consumption.

In this analysis, we attempted to predict vehicle fuel consumption using linear regression. However, our results indicate that the model's performance may be inadequate, as evidenced by the high mean squared error and low R-squared values. These findings suggest that other factors not considered in our analysis may significantly influence fuel consumption. Further exploration of additional features or alternative modeling techniques, such as Kbest selection method, may help improve the accuracy of fuel consumption predictions.

**Model 1: Optimized Linear Regression Model**

For this model, we aimed to optimize the linear regression model used for predicting fuel consumption (comb\_(l/100\_km)) of vehicles based on their characteristics. By incorporating feature selection techniques, specifically SelectKBest, we aimed to enhance the model's performance and achieve better accuracy in our predictions.

**Methodology:** Categorical columns ('make,' 'transmission,' 'fuel\_type,' 'vehicle\_class,' 'model\_year,' and 'model') were identified for one-hot encoding.One-hot encoding was performed on the categorical columns using the pd.get\_dummies() function to convert them into a numerical representation, enabling their usage in the regression model.The dataset was split into features (X) and the target variable (y). Features included all columns except 'comb\_(l/100\_km)', 'fuel\_consumption\_city\_(l/100\_km)', 'hwy\_(l/100\_km)', and 'comb\_(mpg).' The target variable was 'comb\_(l/100\_km).'

**Training and Testing Sets:** The dataset was divided into training and testing sets using the train\_test\_split() function from the sklearn library.The data was split with a test size of 20% and a random state of 42 to ensure reproducibility.

**Feature Selection using SelectKBest:** We utilized the SelectKBest class from the sklearn.feature\_selection module to select the top k features that have the highest correlation with the target variable.

The f\_regression scoring function was used to evaluate the significance of each feature. The training data was transformed to include only the selected features using the fit\_transform() function, while the testing data was transformed using the transform() function.

**Linear Regression Modeling with Selected Features:** We initialized a linear regression model using the LinearRegression() function from sklearn.linear\_model. The model was fitted to the transformed training data, which included the selected features, using the fit() function.

**Prediction and Evaluation:** The trained model was used to predict the fuel consumption for the transformed test data using the predict() function.Mean squared error (MSE) and R-squared (R^2) were calculated to evaluate the performance of the model.MSE measures the average squared difference between the predicted and actual values, providing insight into the model's accuracy.R^2 represents the proportion of the variance in the target variable explained by the model. A higher value indicates a better fit to the data.

After optimizing the linear regression model, we obtained the following results:

**Selected Features:**

engine\_size\_(l)

cylinders

co2\_emissions\_(g/km)

co2\_rating

fuel\_type\_E

These features were identified as the most influential in predicting fuel consumption.

Mean Squared Error: 0.107

R-squared: 0.989

The mean squared error (MSE) and R-squared (R^2) values significantly improved after feature selection, indicating that the optimized model performs better in predicting fuel consumption. The lower MSE value and the higher R^2 value demonstrate improved accuracy and a stronger fit to the data.

**Interpretation of plot:**

A picture containing text, screenshot, line, plot

Description automatically generated

*Figure – X*

The scatter plot displays the relationship between the actual fuel consumption (comb\_(l/100\_km)) and the predicted fuel consumption values from our optimized model. Notably, the data points are more aligned towards the line, indicating a closer correlation between the predicted and actual values. This suggests that our optimized linear regression model better captures the underlying relationship between vehicle characteristics and fuel consumption.

In this analysis, we successfully optimized the linear regression model for fuel consumption prediction by incorporating feature selection using SelectKBest. By selecting the top five most relevant features, we achieved a significant improvement in the model's performance, as evidenced by the lower mean squared error and the higher R-squared value. The plot further supports the effectiveness of the optimized model, showcasing a closer alignment between the predicted and actual fuel consumption values. These findings demonstrate the value of feature selection in enhancing the accuracy and reliability of fuel consumption predictions using linear regression.

**Business question:** How can we predict the fuel consumption (co2\_emissions\_(g/km)) of a vehicle based on its characteristics?

**Model 2: Decision Tree Analysis**

Decision tree regression is a supervised learning algorithm that is commonly used for both classification and regression tasks. In the case of regression, it predicts continuous numerical values based on the input features. The decision tree regression method constructs a tree-like model that recursively splits the dataset into subsets based on the feature values. It uses a series of binary decisions to divide the data into smaller and more homogeneous groups. Each internal node of the tree represents a decision based on a specific feature, and each leaf node represents a predicted value. The objective of this analysis was to predict the fuel consumption of a vehicle, specifically CO2 emissions (co2\_emissions\_(g/km)), based on its characteristics. In this study, we employed the decision tree regression method in Python to develop a predictive model.

**Methodology:** We imported the necessary libraries, including sklearn.model\_selection, sklearn.tree, and sklearn.metrics, for data manipulation, modeling, and evaluation. The dataset was split into features (X) and the target variable (y), where X consisted of all columns except 'co2\_emissions\_(g/km)' and y represented 'co2\_emissions\_(g/km)'.

**Training and Testing Sets:** The dataset was divided into training and testing sets using the train\_test\_split() function from the sklearn library. The data was split with a test size of 20% and a random state of 42 to ensure reproducibility.

**Decision Tree Regression Modeling:** We initialized a decision tree regression model using the DecisionTreeRegressor() function from sklearn.tree. The model was fitted to the training data using the fit() function, allowing it to learn the relationships between the vehicle characteristics and CO2 emissions.

**Prediction and Evaluation:** The trained model was used to predict CO2 emissions for the test set using the predict() function. Mean squared error (MSE) and R-squared (R2) were calculated to evaluate the model's performance. MSE measures the average squared difference between the predicted and actual values, providing insight into the model's accuracy. R^2 represents the proportion of the variance in the target variable explained by the model. A higher value indicates a better fit to the data.

After applying the decision tree regression model, we obtained the following results:

Mean Squared Error: 8.738

R-squared: 0.998

These metrics indicate that the decision tree regression model achieved high accuracy and a strong fit to the data. The low MSE value suggests that the model's predicted CO2 emissions values closely align with the actual values. The high R-squared value indicates that the model can explain approximately 99.8% of the variance in CO2 emissions.

**Business question:** How can we predict the fuel consumption (co2\_emissions\_(g/km)) of a vehicle based on its characteristics?

**Model 3: Multivariate Linear Regression with statsmodels**

Multivariate linear regression is a statistical technique used to predict a dependent variable (in this case, fuel consumption or CO2 emissions) based on multiple independent variables (characteristics of the vehicle). It assumes a linear relationship between the independent variables and the dependent variable.

**Methodology:** The categorical columns are identified, and one-hot encoding is performed to convert them into numerical features. This allows the model to work with categorical variables by representing them as binary indicators (0s and 1s).

**Training and Testing Sets:** The dataset was divided into training and testing sets using the train\_test\_split() function from the sklearn library. The data was split with a test size of 20% and a random state of 42 to ensure reproducibility.

**Multivariate Linear Regression modeling:** The training data is used to fit the multivariate linear regression model. In this case, the Ordinary Least Squares (OLS) method from the statsmodels library is utilized. OLS minimizes the sum of the squared differences between the predicted and actual values.

**Prediction and Evaluation:** The trained model was used to predict CO2 emissions for the test set using the predict() function. Mean squared error (MSE) and R-squared (R2) were calculated to evaluate the model's performance. MSE measures the average squared difference between the predicted and actual values, providing insight into the model's accuracy. R^2 represents the proportion of the variance in the target variable explained by the model. A higher value indicates a better fit to the data.

**Prediction and Evaluation:** The performance of the multivariate linear regression model is evaluated using metrics such as mean squared error (MSE) and R-squared (R2). MSE measures the average squared difference between the predicted and actual values, while R2 represents the proportion of the variance in the dependent variable explained by the model. Lower MSE and higher R2 values indicate a better fit of the model to the data.

After applying the decision tree regression model, we obtained the following results:

Mean Squared Error: 17.889041329630018

R-squared: 0.9953456328721295

These metrics indicate that the decision tree regression model achieved high accuracy and a strong fit to the data. The low MSE value suggests that the model's predicted CO2 emissions values closely align with the actual values. The high R-squared value indicates that the model can explain approximately 99.5% of the variance in CO2 emissions.

A picture containing text, screenshot, plot, line

Description automatically generated

*Figure – X*

**Interpretation of plot:**

The scatter plot visualizes the relationship between the actual CO2 emissions and the predicted CO2 emissions values, providing an assessment of the model's performance. The data points align with the regression line, therefore the model is fit.

Comparing the multivariate linear regression model's performance with the decision tree regression model mentioned earlier, we can observe the following:

Mean Squared Error (MSE):

* Multivariate Linear Regression: 17.889
* Decision Tree Regression: 8.738

The decision tree regression model achieved a lower MSE, indicating better performance in predicting CO2 emissions compared to the multivariate linear regression model.

R-squared (R^2):

* Multivariate Linear Regression: 0.995
* Decision Tree Regression: 0.998

Both models exhibit high R-squared values, indicating a strong fit to the data. However, the decision tree regression model achieved a slightly higher R2, suggesting that it explains a larger proportion of the variance in CO2 emissions compared to the multivariate linear regression model.

**Business question:** How accurately can we predict the CO2 Rating and Smog Rating of vehicles using the given features in dataset?

**Model 4: Random Forest Regression**

Random forest regression is a supervised learning algorithm that combines multiple decision trees to create a robust and accurate prediction model. It is commonly used for regression tasks where the goal is to predict a continuous variable (in this case, CO2 rating and smog rating) based on a set of input features (vehicle characteristics).

**Methodology:** The dataset is prepared by splitting it into features (X) and target variables (CO2 rating and smog rating). The categorical columns are one-hot encoded to convert them into numerical features.

**Training and Testing:** The dataset is divided into training and testing sets. The training set is used to build the random forest regression model, while the testing set is used to evaluate its performance. The dataset was divided into training and testing sets using the train\_test\_split() function from the sklearn library. The data was split with a test size of 20% and a random state of 42 to ensure reproducibility.

**Random Forest Regression modeling:** Random forest regression models are initialized and fitted to the training data. Each model in the random forest is trained on a subset of the training data and uses a random subset of features. This randomness helps to reduce overfitting and improve generalization. Random forest regression models are built separately for predicting CO2 rating and smog rating using the RandomForestRegressor class from the sklearn.ensemble module. The number of estimators (decision trees) in the random forest is set to 3.

**Prediction and Evaluation:** The models are trained using the training data, and predictions are made for the test set using the predict() method. Mean squared error (MSE) and R-squared (R2) are calculated using the mean\_squared\_error() and r2\_score() functions, respectively, to evaluate the performance of the models.

After applying the decision tree regression model, we obtained the following results:

CO2 Rating:

* Mean Squared Error: 0.03499713138267356
* R-squared: 0.9746301300452619

Smog Rating:

* Mean Squared Error: 0.3781985083189902
* R-squared: 0.7628433514015527

The random forest regression models achieved relatively low MSE values and high R-squared values for both CO2 rating and smog rating, indicating good predictive performance. However, it's worth noting that the R-squared value for the Smog Rating is slightly lower compared to the CO2 Rating, suggesting that the model explains a lower proportion of the variance in Smog Rating.

A picture containing text, screenshot, diagram, line

Description automatically generated

*Figure – X*

*A picture containing text, screenshot, line, diagram

Description automatically generated*

*Figure - X*

**Interpretation of plot:**

The scatter plots show the relationship between the actual and predicted values for CO2 rating and smog rating. In the CO2 rating plot, the data points appear to align reasonably well with the line, indicating a strong relationship between the actual and predicted values. However, in the Smog Rating plot, some data points deviate from the line, suggesting that the model may not capture all the variations accurately.

The random forest regression method demonstrates promising performance in predicting both CO2 rating and smog rating based on the given features. However, further analysis and fine-tuning of the model may be necessary, particularly for improving the prediction of smog rating, as indicated by the scatter plot.

**Model 4: Optimized Random Forest Regression**

For this model, we have optimized the random forest regression model to predict the CO2 Rating and Smog Rating of vehicles based on the given features.

**Methodology:** The dataset is split into features (X) and target variables (CO2 Rating and Smog Rating). Categorical columns are encoded using one-hot encoding to convert them into numerical features.

**Data Splitting:** The dataset is divided into training and testing sets, similar to the previous model.

**Feature Selection:** SelectKBest, a feature selection method, is applied to select the top k features that have the highest correlation with the target variable. In this case, the f\_regression scoring function is used to evaluate the correlation between the features and the CO2 Rating. The top five features are selected based on their scores.

**Model Building:** Random Forest regression models are initialized and fitted to the training data, but this time using only the selected features from Step 3. One model is trained for predicting the CO2 Rating, and another model is trained for predicting the Smog Rating.

**Prediction and Evaluation:** The optimized models are used to make predictions on the test set. Mean squared error (MSE) and R-squared (R2) values are calculated to evaluate the performance of the models, as in the previous model.

The selected features are printed to showcase the top five features that were chosen by the SelectKBest method. This provides insights into which features are most relevant for predicting the CO2 Rating.

Comparing the results with the previous random forest regression model:

CO2 Rating:

* Mean Squared Error (MSE): 0.147
* R-squared (R2): 0.893

Smog Rating:

* Mean Squared Error (MSE): 0.848
* R-squared (R2): 0.468

The optimized random forest regression model shows a slight decrease in R-squared and an increase in MSE for both CO2 Rating and Smog Rating compared to the previous model.

A picture containing text, screenshot, diagram, line

Description automatically generated

*Figure – X*

*A picture containing text, screenshot, line, diagram

Description automatically generated*

*Figure – X*

**Interpretation of plot:**

Regarding the scatter plots, both the CO2 Rating and Smog Rating plots show data points that align better with the line compared to the previous model. This suggests that the optimized model performs better in capturing the variations and making more accurate predictions for both target variables.

The optimized random forest regression model, using the top five selected features, demonstrates improved performance in predicting both CO2 Rating and Smog Rating. However, there is still room for further analysis and fine-tuning of the model to potentially enhance its predictive capabilities. We chose k=5 so as to reduce model complexity and while executing the code with a greater number of k, the machine became slower, and it took longer to process. It was determined that the top five features have the highest correlation or importance in predicting the CO2 Rating. By selecting only these top features, the model aims to focus on the most relevant aspects and potentially improve interpretability and computational efficiency.

**REFERENCES**

1. Fuel consumption ratings - Open Government Portal. (n.d.). <https://open.canada.ca/data/en/dataset/98f1a129-f628-4ce4-b24d-6f16bf24dd64>
2. Canada, N. R. (2018, September 14). *Understanding fuel consumption ratings*. <https://natural-resources.canada.ca/energy-efficiency/transportation-alternative-fuels/fuel-consumption-guide/understanding-fuel-consumption-ratings/21006>
3. *An introduction to seaborn — seaborn 0.12.2 documentation*. (n.d.). <https://seaborn.pydata.org/tutorial/introduction>
4. Chayan. (2020). Missing Value imputation using MICE&KNN | CKD data. *www.kaggle.com*. <https://www.kaggle.com/code/chayan8/missing-value-imputation-using-mice-knn-ckd-data/notebook \>
5. Eddie. (2023). How to Handle Missing Data: A Step-by-Step Guide (Updated 2023). *Analytics Vidhya*. <https://www.analyticsvidhya.com/blog/2021/05/dealing-with-missing-values-in-python-a-complete-guide/>
6. Chapter 02 - Introduction to Python.ipynb. (n.d.). <https://jupyter.brynmawr.edu/services/public/dblank/CS245%20Programming%20Languages/2016-Fall/Labs/Chapter%2002%20-%20Introduction%20to%20Python.ipynb>
7. Viriyakovithya, E. (2019). How to combine multiple CSV files with 8 lines of code. *freeCodeCamp.org*. <https://www.freecodecamp.org/news/how-to-combine-multiple-csv-files-with-8-lines-of-code-265183e0854/>
8. CSV File Reading and Writing. (n.d.). Python Documentation. <https://docs.python.org/3/library/csv.html>
9. Nkmk. (2023). Remove an item from a list in Python (clear, pop, remove, del). *note.nkmk.me*. <https://note.nkmk.me/en/python-list-clear-pop-remove-del/>
10. *4 Ways to Remove Empty Strings from a List*. (n.d.). <https://datatofish.com/remove-empty-strings-from-list/>
11. *How to Merge all CSV Files into a single dataframe � Python Pandas*. (n.d.). <https://www.tutorialspoint.com/how-to-merge-all-csv-files-into-a-single-dataframe-python-pandas>
12. GeeksforGeeks. (2022). How to Create a Stacked Bar Plot in Seaborn. *GeeksforGeeks*. <https://www.geeksforgeeks.org/how-to-create-a-stacked-bar-plot-in-seaborn/>
13. Legislative Services Branch. (2022, October 3). *Consolidated federal laws of Canada, Heavy-duty Vehicle and Engine Greenhouse Gas Emission Regulations*. <https://laws.justice.gc.ca/eng/regulations/SOR-2013-24/FullText.html>