

**MMA 860**

**Acquisition and Management of Data**

**Alex Scott**

**Project Report**

**18th April 2022 | 11:59PM**

**Team Adelaide**

|  |  |
| --- | --- |
| **Student Name** | **Student Number** |
| Ishaan Sinha | 20290167 |
| Jiali Bai | 20316342 |
| Juefei Chen | 20311141 |
| Kripa Shanker Nayak | 20315710 |
| Oluwaseun Adelana | 20310041 |
| Yichen Yuan | 20315046 |
|  |  |

## Order of files:

|  |  |  |
| --- | --- | --- |
| **Filename** | **Pages** | **Comments and/or Instructions** |
| MMA 860 Team Project Report | 2-8 |  |
| Appendix | 9-19 |  |
| References | 20 |  |
|  |  |  |
|  |  |  |

**Additional Comments:**

**CO2 Emission Rating**

**EXECUTIVE SUMMARY**

Canada has been facing increased pressure from its citizens to launch an enhanced CO2 emissions rating standard tool which helps consumers make smart decisions while purchasing a new or used vehicle. Canadian Emission Standard rating agency (CESRA), Canada’s automobile rating arm has hired us to develop a model which would provide a CO2 emission rating for all the vehicles available to consumers within Canada. CESRA has provided us access to past five years historical data for all vehicles purchased within Canada with the emissions data as provided by the manufacturer. Data provided by CESRA is unorganized and contains missing data hence we would perform various data cleaning and exploration methodologies to ensure usability of data for model development. Once the data was cleansed and deemed usable, we performed feature engineering and utilized the data to build a predictive model which helps in predicting the CO2 emissions based on significant variables such as make, model, class, fuel type, transmission etc. We also ran a linear hypothesis test before removing any variables which failed the t-test and didn’t show any explanatory power. As per our predictive model, vehicle types such as Sports cars, SUV, Vans and pickup trucks have high CO2 emission. Post our hypothesis validation we tested our model on a test data set and our model appears to be robust and accurate. Finally, we developed an executive dashboard to present our findings to CESRA board members.

**Introduction**

Climate change is a global environmental challenge. It is caused by the increased emission and concentration of greenhouse gases (GHGs) into the atmosphere largely because of human activities in Transportation and Energy. GHGs traps heat in the atmosphere, increasing the earth’s temperature and leading to extreme weather conditions thereby impacting the health and well-being of people and creating an economic and environmental crisis. Carbon dioxide (CO2) is the main greenhouse gas emitted from the transportation of people, goods, and services. According to Canada’s Environmental and Natural Resources, the oil and gas and transport sector were the largest GHG emitters in Canada. Together, they accounted for 52% of total emissions in the year 2019.

The United States Environmental Protection Agency also claims a typical passenger emits about 4.6 metric tons of carbon dioxide per year. This is an indication that CO2 emission is largely dependent on fuel consumption, type of fuel, and distance traveled which means humans can reduce the magnitude of these emissions by making better and informed choices in the type of vehicle they purchase and its fuel efficiency, opting for fuel with less carbon, purchasing an electric vehicle, carpooling, or using more of the public transit system.

Canada aims to achieve net-zero emissions by 2050 and automobile consumers are the key to achieving these emissions targets. Hence, Canadian Emission Standard rating agency (CESRA), Canada’s automobile emissions rating arm aims to provide consumers a tool which helps them make automobile purchase decisions resulting in minimal the impact on the environment. CESRA has hired us to develop a model which would provide a CO2 emission rating tool for all the vehicles available to consumers within Canada.

**Data Cleansing and Feature Engineering**

To form our dataset for analysis, we initially imported the data of different automobiles for the last five years and then combined them into a single data frame. Second step in our data analysis process was to cleanse the combined data frame. We dropped some of the variables that were not needed in the model, for instance, fuel consumption is an additional dependent variable in the data that is also collinear with CO2 emission. Next, we renamed the remaining variables and extracted relevant information from them, such as the model type of the vehicle (4-wheel drive, All-wheel-drive, Flexible-fuel vehicle, short wheelbase, long wheelbase, and Extended wheelbase) was identified from the vehicle model's name. Also, the data type of some of the variables was modified to model them better. Finally, we removed any duplicate rows and checked for NULL values.

As part of our data exploration, we used visualization techniques in R to observe the relationship between CO2 emission and other variables in the model: Scatterplots for ‘Engine size vs CO2 emission’ and ‘Cylinders vs CO2 emission’ and a Bar graph to observe the trend of CO2 emission over the years. From the plots, we noticed that both engine size and number of cylinders are positively related with CO2 emission, meaning that the large the engine size is, and with a greater number of cylinders, the high CO2 emission would be for that vehicle. It is commonly known that vehicles with larger engine size and more cylinders require more power for starting and/or speeding up, thus increasing its CO2 emission. Our plots results aligned with these facts.

After that, we populated a histogram about number of vehicles’ distribution for CO2 emission and found that most of the vehicles’ CO2 emissions are normally distributed and between 200-300 grams per kilometer Graph 1. This might be the most balanced choice vehicles can go with between being environmental-friendly and providing sufficient performance for customers. We also populated a histogram between year of production and CO2 emission however the distribution is relatively even and doesn’t provide strong evidence of any relationship between these two variables.

Further, we identified all the unique values for variables such as make (the brand of the automobile), class (SUV, Compact etc.), and model type. We also looked at the summary statistics of the data to gain a better understanding of each of the variables individually.

After merging and cleaning our data, our final model consists of variables as listed below:

* Model Year
* Make and Model
* Vehicle Class
* Engine size
* Cylinders
* Transmission
* Fuel Type and Fuel city,
* Fuel utilized on highways and
* CO2 emission

We leveraged feature engineering to modify some of the variables and create some new ones to suit our needs. We created variables for the age and the difference between fuel consumption on highways and in cities for different automobiles. Next, we created dummy variables as indicators to represent the categorical data in our model.

**Model Development and Testing**

Once the data had been prepared, we leveraged it to create a predictive model for CO2 emission. We used linear regression as a preferred method to build our predictive model. CO2 emission was kept as the dependent variable, and we initially ran a linear regression model keeping all the independent variables in the model. Through individual t-tests, we were able to identify the significance of all the variables. We then ran a linear hypothesis test to determine whether all the insignificant variables identified by us are collectively of significance or not. We found that some variables still belong in the model, and after trying different combinations, it was determined that vehicles of the brands Dodge and GMC should also be included. Now we ran a new linear regression model keeping all the significant variables identified by us. Using the Breusch-Pagan test in R, it was found that there was heteroskedasticity in our model. So, we used the HCCME method, available in the estimatR package to rectify this. We ran a new corrected regression model using the HC2 version of HCCME method. All the variables included passed the t-tests, and the error term was normally distributed. We did not find any outliers in our data as well.

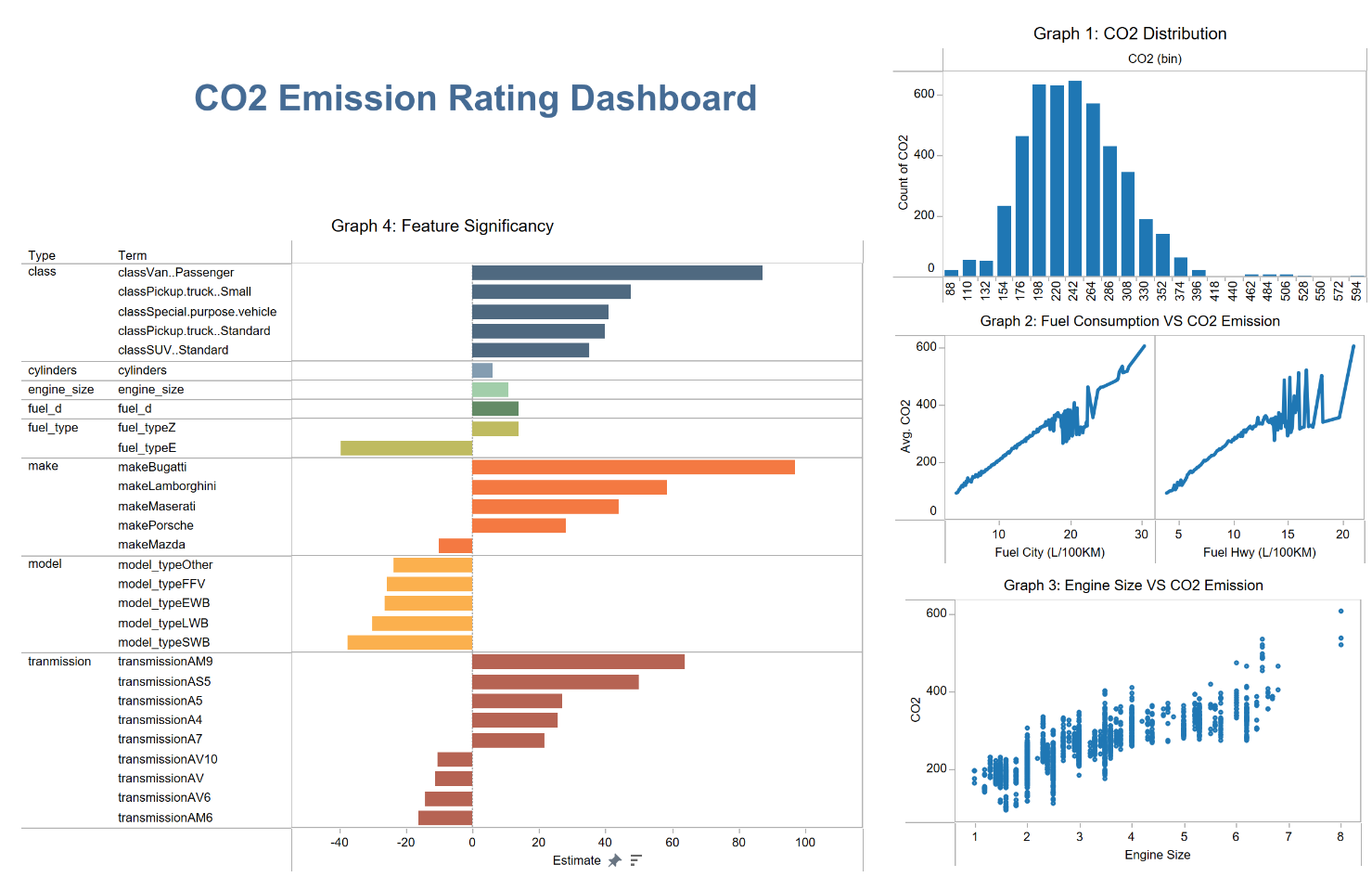
As per the corrected model’s output, the adjusted R-squared is 0.9005, which means that our model is correct and reliable. Among different vehicle makes such as Bugatti, Lamborghini and Maserati have the highest coefficients, meaning that they have high leverage in our model. Thus, we can conclude famous brands of supercars that show excellent performance in terms of speed result in high CO2 emission. On the other hand, Mazda has negative coefficient in our model, meaning that if we choose a Mazda car, it will have a low CO2 emission which it turn concludes that it is very environmentally friendly. If customers are looking for a fuel-saving and low emission vehicle, our model would recommend Mazda as the preferred choice.

Among vehicle’s class, we found vans and pickup trucks have more CO2 emission compared to other vehicle classes. Customers should not purchase these types of cars if they’re looking for some green solutions.

After fitting the model with the training dataset, we then used the testing dataset to understand the accuracy of our model. We predicted the values of the test dataset using the model that we developed. Then we compared our predicted values to the actual values of the test dataset by comparing the R2, RCME (average error), and MAE values with our model. These values are observed to be similar, which shows that our model is accurate. We also used Chow Test to see if the effect of engine size on the CO2 consumption of a vehicle is different for 2-seater or not. From the results, we observed that our null hypothesis is rejected and there in fact is a difference in CO2 emissions for 2-seater.

Finally, we also built an executive dashboard on Tableau with some plots highlighting the relationship between CO2 emission and several of the significant variables identified.

**Executive Dashboard**



**Conclusion and Recommendations**

Post our testing process, our predictive model was able to show accurate prediction based on significant variable selection. Variables such as make, model, class, cylinder, engine size, transmission and fuel type show a strong relationship with CO2 emissions. Our tool can be effectively shared with consumers so that they can predict the CO2 emission of all the vehicles available for purchase within Canada and make an informed decision.

We also recommend expanding the scope of the model so that data associated with variables such as horsepower, vehicle weight, acceleration and model enhancement trends are included. Also, we should capture data for electronic vehicles and global markets which would help enhance the model’s predictive power.

**Appendix**

**Part 1 – Merging Data and Cleaning**

1. Combined CSV file from 2018 to 2022 in R data frame
2. Rename columns
3. Remove columns that are not relevant: comb lkm, comb mpg, EO2 rating, smog rating.
4. Extract the below info from the model’s name, then remove the models’ name

|  |
| --- |
| 4WD/4X4 = Four-wheel drive |
| AWD = All-wheel drive |
| FFV = Flexible-fuel vehicle |
| SWB = Short wheelbase |
| LWB = Long wheelbase |
| EWB = Extended wheelbase |

1. Remove 470 duplicate rows
2. Correct data type: characters to numeric

**Output**

Table

Description automatically generated

**Part 2 – EDA and Data Visualization**

descriptive statistics**Table

Description automatically generated**

Chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart, histogram

Description automatically generated

Chart, bar chart

Description automatically generated

Counts of each category A picture containing table

Description automatically generated

**Part 3 - Feature Engineering**

1. Create a column for the difference between the model year and the current year
2. Create a column for the difference between the city consumption and the hwy consumption
3. Remove columns city consumption and the hwy consumption
4. Create dummy variables

**Part 4 – Fitting Model**

1. Fit a model with all variables
2. Run a linear hypothesis on the variables we want to remove
3. Remove variables does not belong in the model

**Output:**

Table

Description automatically generated

Table

Description automatically generated

**Part 5 – Testing and Tuning Model**

1. Plot to test model

Diagram

Description automatically generated

Chart, histogram

Description automatically generated

1. Test for heteroskedasticity

A picture containing timeline

Description automatically generated

There is heteroskedasticity in our current model.

Text

Description automatically generated with medium confidence

1. Correct the model using HCCME

**Part of the output**

Table

Description automatically generated

**Part 6 – Make Predictions**

1. Split data for train set (60%) and test set (40%)
2. Make prediction

**Output**

Text

Description automatically generated

**Part 7 – Chow Test: does a two-seater respond more to change in Engine Size?** Table

Description automatically generated with low confidence

**References**

Canada, E. and C. C. (2022, April 14). *Government of Canada*. Canada.ca. Retrieved April 16, 2022, from https://www.canada.ca/en/environment-climate-change/services/environmental-indicators/greenhouse-gas-emissions.html

Environmental Protection Agency. (n.d.). EPA. Retrieved April 16, 2022, from https://www.epa.gov/ghgemissions/overview-greenhouse-gases

Environmental Protection Agency. (n.d.). EPA. Retrieved April 16, 2022, from https://www.epa.gov/greenvehicles/fast-facts-transportation-greenhouse-gas-emissions

Environmental Protection Agency. (n.d.). EPA. Retrieved April 16, 2022, from https://www.epa.gov/greenvehicles/greenhouse-gas-emissions-typical-passenger-vehicle

*The official U.S. government source for fuel economy information.* www.fueleconomy.gov - the official government source for fuel economy information. (n.d.). Retrieved April 16, 2022, from https://www.fueleconomy.gov/