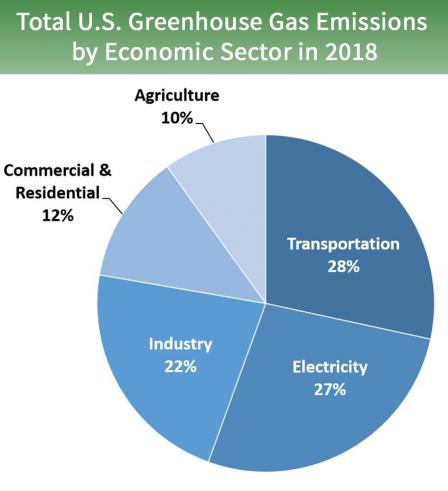
**Introduction**

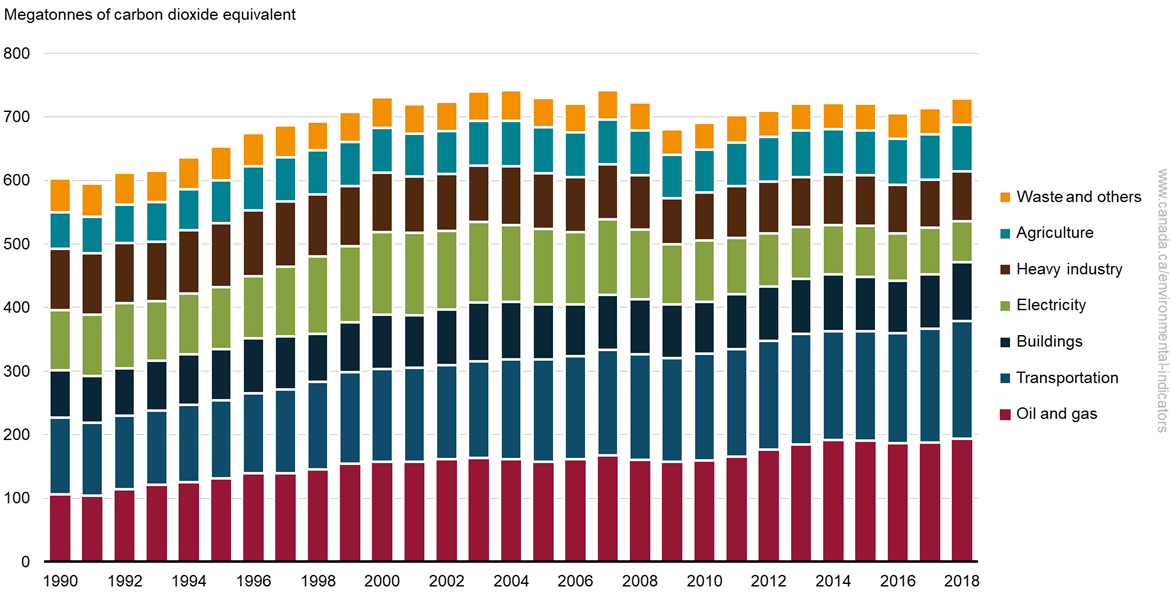
Carbon dioxide, or CO2, is known as a greenhouse gas; that is, it warms Earth’s climate by trapping solar radiation akin to a greenhouse. This accelerates the rising of sea levels, the increase in hurricanes, and the destruction of natural habitats, to name a few detriments. In response, many governments have set emission limits for this gas, from the power generation sector to the transportation sector.

Transportation is by far one of the biggest culprits of CO2 emission is the transportation sector. The US EPA estimates that currently active vehicles account for approximately 28% of all greenhouse gas emissions (similar rate for Canada), a third of which are caused by personal motor vehicles. Despite the emergence of electric vehicles, mitigating CO2 emissions remains the practical solution for climate change in the near term due to the lack of battery range and charging infrastructure. Vehicle manufacturers are targeting CO2 emissions only, as all vehicles built since 1975 (1988 in Canada) are equipped with catalytic converters to render harmless all other pollutant gases such as nitrous oxide, methane, and carbon monoxide.

US CO2 Emissions by Sector:



Canadian CO2 Emissions by Sector:



My goal for this project is to predict the carbon dioxide emissions of vehicles per distance travelled by using various vehicle engine data, and to find out which engine parameters result in a minimum CO2 emission rate. With an emission prediction model, companies can more easily comply with laws and international agreements, safely reduce vehicle costs, and more importantly, determine if a new vehicle is safe for the environment to mitigate the aforementioned consequences of a warming climate. In addition, consumers can use such a model to make informed decisions on what eco-friendly cars to purchase.

**Research Questions**

The following are the questions I plan to ask with my dataset so far:

1. Can CO2 emissions be accurately predicted using our variables? If so, which engine variables weigh the most in the model?
2. Is there really a non-negligible difference in fuel consumption between car using regular gasoline (octane rating 87) and premium gasoline (octane rating 91)?
3. Is diesel fuel worse for the environment as many people claim?
4. What car manufacturers are the biggest culprits of CO2 emissions?
5. Is it also possible to predict fuel consumption rates using the same variables?
6. What type of car engine is the most common (e.g., 3.0L V6)?

**How I plan to address the problem statement**

First, I will have to replace each categorical variable with one separate Boolean vector per category corresponding to whether an observation has that category. This will make the dataset much easier to use regressions models on.

For Question 1, I will generate a multiple regression model and determine if it is accurate. I will start by plotting scatter plots of CO2 emissions versus the continuous variables (boxplots for CO2 emissions versus the categorical variables) to visually and numerically determine any correlations. If the correlations are found to be linear, I will generate a multiple linear regression model from the training data (70% of my dataset). If some correlations are not linear, I will assess whether it is appropriate to use a quadratic or cubic term for that variable in the regression model. I will use the RMSE, R2 value, and a few of the casewise diagnostics from our previous exercises to determine how accurate the model really is.

To answer Question 2, I will use a two-sample t-test to determine if the difference between fuel consumption of regular and premium gas cars is, in fact, statistically significant. For Question 3, I will run another two-sample t-test, this time on the CO2 emissions data of diesel fuel and regular gas.

Question 4 can be answered by constructing a box plot of CO2 emissions versus car brand. I will recommend that consumers avoid the top 10 polluting brands and only purchase from the top 10 cleanest CO2 emitters.

Question 5 will use a similar setup as Question 1, but with fuel consumption rather than CO2 emissions

Question 6 can be answered by using histograms. I will plot histograms of all the engine data (bar charts for categorical variables) to get a general idea as to which engine one could find if a car was picked at random from the street.

**Dataset**

The dataset I will use is a CSV file containing observations on 7,385 cars from the 2013-2020 model years. Each row corresponds to one car make and model and each column corresponds to one of 12 engine variables. This dataset was compiled by the Department of Natural Resources Canada and is free to download from the Canadian government’s website. There are no missing values from this dataset, as it is mandatory to report all vehicle specifications before sale. The variables in the dataset are as follows:

* Make: A string displaying the manufacturer of each car.
* Model: A string displaying the brand sold by the manufacturers (I won’t use this as it is just a marketing name).
* Vehicle Class: A string displaying the overall shape of the vehicle (SUV, station wagon, etc). This variable takes weight into account.
* Engine Size: A float showing the total volume swept by all the pistons in liters.
* Cylinders: An integer showing the number of pistons within the engine.
* Transmission: A string displaying the mechanism used to transmit engine power to the wheels. Five categories exist: Automatic, Manual, Automated Manual, Automatic Hybrid, and CVT.
* Fuel Type: A string displaying the fuel used to power the engine. Five categories exist: gasoline (Z), premium gasoline (X), diesel (D), natural gas (N), and ethanol (E).
* City Fuel Consumption: A float displaying the estimated fuel consumed using city driving habits, measured in liters per 100 kilometers traveled.
* Highway Fuel Consumption: A float displaying the estimated fuel consumed using highway driving habits, measured in liters per 100 kilometers traveled.
* Combined Fuel Consumption: A float displaying an estimate of the real-world fuel economy from the car, combining the previous values. Note: this is NOT an arithmetic mean of the city and highway fuel consumptions.
* CO2 Emissions: An integer displaying the amount of carbon dioxide emitted by each vehicle, measured as grams CO2 per kilometer traveled.

Here is a link to the dataset.

<https://open>.canada.ca/data/en/dataset/98f1a129-f628-4ce4-b24d-6f16bf24dd64#wb-auto-6

**How my approach will partially address the prediction problem**

My approach should generate a prediction model that should be appropriate to all countries’ cars, but I anticipate it will only be partially accurate due to lacking many other engine variables. For example, the dataset does not contain vehicle engine power which I feel will heavily impact emissions. There is only engine size and cylinders as a proxy for it.

I feel that the anticipated model would become truly universal if data such as power, torque, and drivetrain type are found.

**Packages**

I will use the following R packages in my analysis so far:

*ggplot2:* data visualizations

*dplyr, tidyr:* data tidying and segregation

*GGally,* possibly. It will make generating several correlation coefficients much quicker.

**Plots**

I will be using the following plots and tables in my analysis: Histograms, scatter plots, box plots, and possibly a correlation matrix.

**Further Information Required**

Currently, I will need to learn how to quantify the correlation between continuous and categorical variables. Currently, I only know how to visually assess the correlation via box plots.

References:

Babad, Michael (1985). *Canada to require catalytic converters on US auto imports.* United Press International. Retrieved on November 5, 2020 from https://www.upi.com/Archives/1985/03/09/Canada-to-require-catalytic-converters-on-US-auto-imports/5059479192400/

Brinson, Linda C. *How much air pollution comes from cars?* HowStuffWorks. Retrieved on November 5, 2020 from https://auto.howstuffworks.com/air-pollution-from-cars.htm#:~:text=According%20to%20the%20EPA%2C%20motor,percent%20of%20greenhouse%20gas%20emissions.

*Greenhouse Gas Emissions.* Canadian Government. Retrieved on November 5, 2020 from https://www.canada.ca/en/environment-climate-change/services/environmental-indicators/greenhouse-gas-emissions.html.

Loveday, S. (2020). *11 Reasons People Don’t Buy Electric Cars.* US News and World Report. Retrieved on November 5, 2020 from https://cars.usnews.com/cars-trucks/why-people-dont-buy-electric-cars.

Sources of Greenhouse Gas Emissions. United States Environmental Protection Agency. Retrieved on November 5, 2020 from https://www.epa.gov/ghgemissions/sources-greenhouse-gas-emissions#:~:text=In%202018%2C%20greenhouse%20gas%20emissions,of%20U.S.%20greenhouse%20gas%20emissions.

Zeng, W., Miwa, T., and Morikawa, T. (2016). *Prediction of vehicle CO2 emission and its application to eco-routing navigation.* ScienceDirect Journal. Retrieved on November 5, 2020 from https://www.sciencedirect.com/science/article/pii/S0968090X16300110.