EMISSIONS BY VEHICLE PROPERTY

Exploring the Relationship of Vehicle Characteristics to Canada's Air Quality Ratings

Also Considering Global Factors

This data set has added attributes from the world; namely, which countries buy which parent vehicle brands more, which countries emit more CO2, which countries import more crude oil and which countries have a higher % of GDP from their transportation sector.

Arayana Janson
arayana.a.janson@wmich.edu

Transportation is often one of the most significant polluting sectors (next to Industrial production) of any economy, accounting for 28% of total greenhouse gas emissions in the U.S. in 2018.[i] Passenger travel accounts for 60% of CO2 emissions from transportation and freight comprises the rest.[ii] Flight emits a small percentage of global CO2 emissions, at 2.5%.[iii] Road traffic is thought to account for 10% of global CO2 emissions and cars are thought to emit more CO2 on average than planes because they consume more energy to transport the same number of passengers.[iv]

Perhaps related to their participation in the Paris Climate Agreement, many countries are interested in tracking and reducing CO2 emissions from transportation. Canada, for example, which is part of the Paris Agreement[v], tracks CO2 emissions by vehicle model.

Canada offers its findings on vehicle emissions at https://open.canada.ca/data/en/dataset/98f1a129-f628-4ce4-b24d-6f16bf24dd64#wb-auto-6, where the original data set for conventional and (general) hybrid cars, called "MY2020 Fuel Consumption Ratings", is found. This data set has been merged with the original data for 2020 plug-in hybrid electric vehicle models from the document, "MY2012-2020 Plug-in Hybrid Electric Vehicles", which was also located at the above website.[vi]

In addition, and, in the interest of obtaining a better picture of the larger context of CO2 emissions reduction, the following variables were pulled from the internet:

1. Parent Company[vii]

VWG	Volkswagen Group
FCA	Fiat Chrysler Automobiles
FMC	Ford Motor Co
HMG	Honda Motor Company
DAG	Hyundai Motor Group
BMWG	Daimler AG
RNMA	BMW Group
SC	Renault-Nissan-Mitsubishis Alliance
TMC	Subaru Corp
ZGHG	Toyota Motor Corp
HMC	Zhejiang Geely Holding Group
GM	General Motors
TM	Tata Motors
MMC	Mazda Motor Corp

2. The Country of Origin[viii] (or, where the parent brand of the model was first designed and/or produced, not necessarily where the mdoel is currently manufactured);
3. Country in which the Parent Company sells the most cars globally[ix] (or, "Country of Highest Sales by Parent Company");
4. Percentage of Global CO2 Emissions by Country[x] (in accordance with the country in which the Parent Company sells the most cars);
5. Percentage of Global Crude Oil Imports by Country[xi] (again, in accordance with the country in which the Parent Company sells the most cars);
6. Logistics Percentage of GDP by Country[xii] (also in accordance with the country in which the Parent Company sells the most cars); and,

7. Number of Vehicles in Use by Country in Thousands[xiii] (also in accordance with the country in which the Parent Company sells the most cars).

The plug-in hybrid electric vehicles have a combo fuel type of B/X (Battery and Regular Gasoline) or B/Z (Battery and Premium Gasoline).[xiv] The Fuel Type Codes List is:

X	Regular Gasoline
Z	Premium Gasoline
D	Diesel
E	Ethanol
B	Battery (Electric)

All variables in the cleansed data set are:

Excel CATEGORY NAME	R/.csv CATEGORY NAME(abbreviated)
Vehicle Type	VehicleType
Model	Model
Make	Make
Parent Company	ParentCo
Country of Origin	CountryofOrigin
Country of Highest Sales by Parent Brand	HighSalesCountry
Percentage Global CO2 Emissions by Country	PercentageGlobalCO2EmissionsbyCountry
Percentage Global Crude Oil Imports by Country	PercentageGlobalCrudeOilImportsbyCountry
Logistics Percentage of GDP by Country	LogisticsPercentageofGDPbyCountry
Number of Vehicles in Use in Thousands by Country	NoVehiclesCntry
Vehicle Class	Class
Engine Size	engineSize
Cylinders	Clndrs
Transmission	Transmission
Fuel Type	FuelType
Fuel Consumption Combo (L/100 km)	fuelConsumption
CO2 Emissions (g/km)	CO2Emissions
CO2 Rating	CO2Rating
Smog Rating	SmogRating

Before exploring methods of analysis and before exploring which questions the data can best answer, some plots will help visualize potential relationships and anomalies.

VISUALIZATIONS

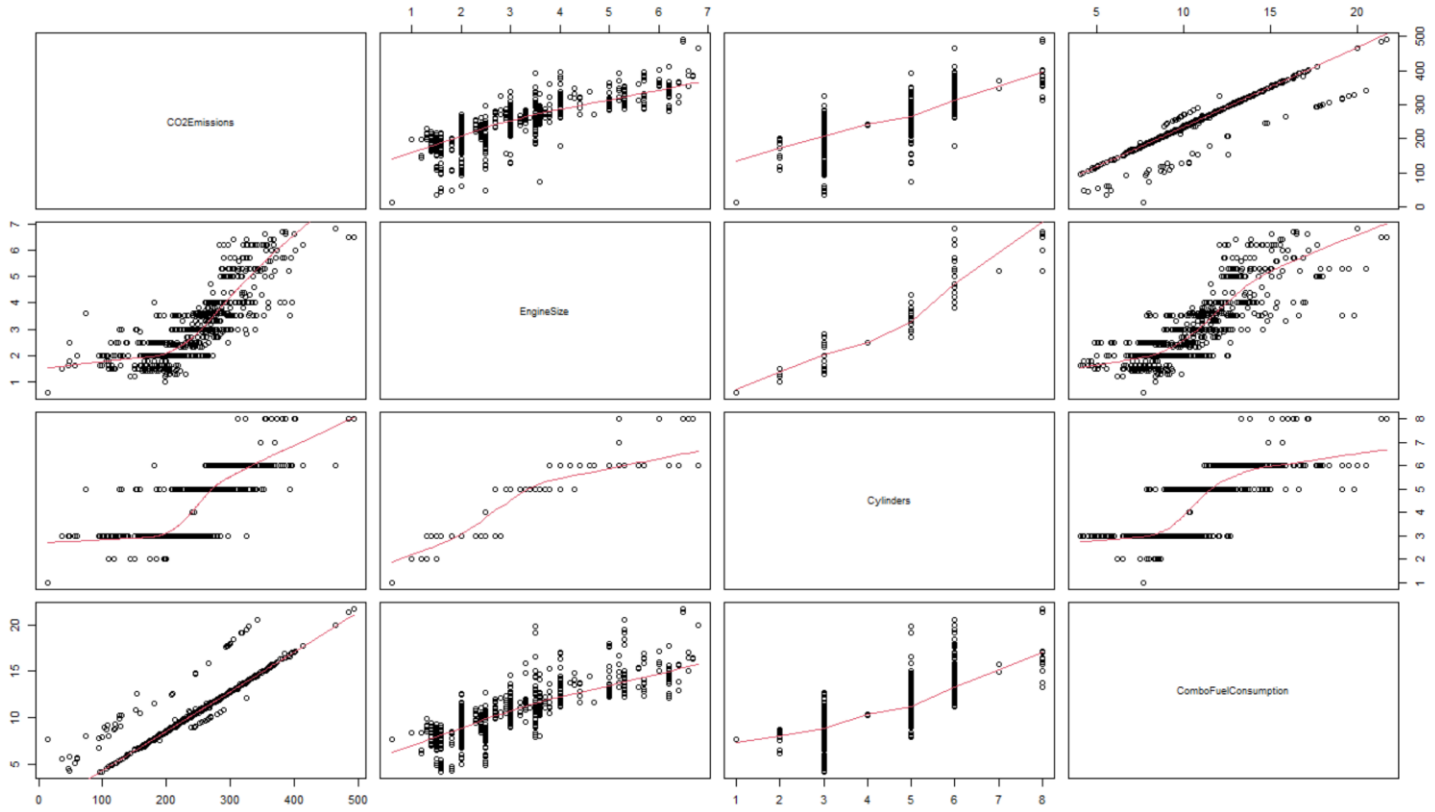


Figure 1: CO2Emissions, EngineSize, Cylinders and ComboFuelConsumption

Looking at Engine Size and the possible response variables, we find some potential collinearities.

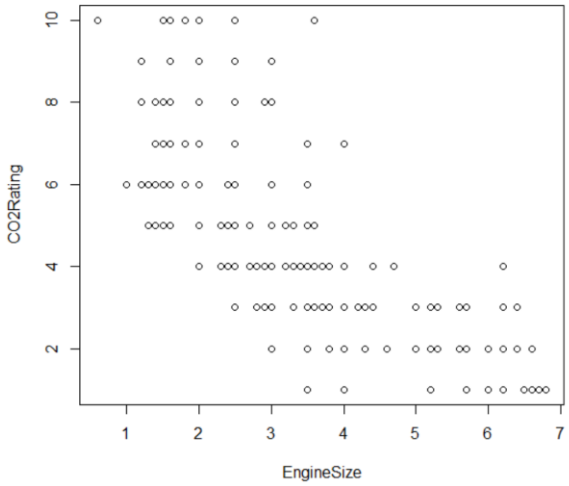


Figure 2: Better CO2 Ratings are awarded to vehicles with smaller Engine Size.

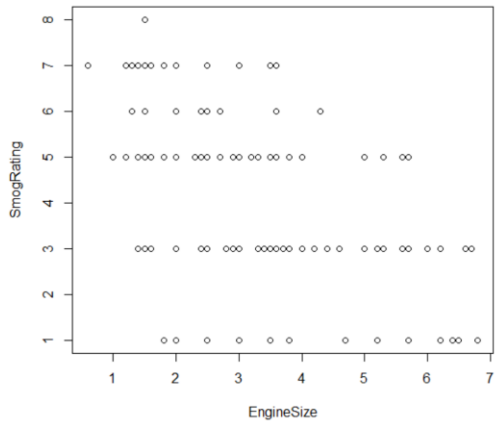


Figure 3: Better Smog Ratings tend to be awarded to vehicles with smaller Engine Size.

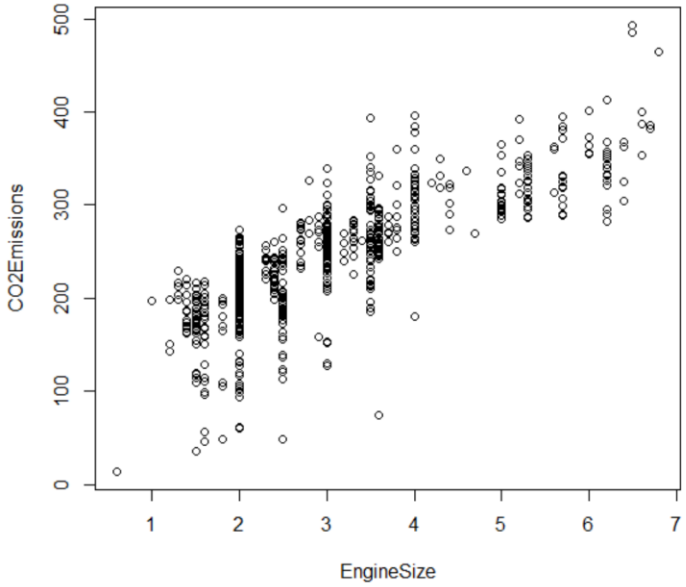


Figure 4: The larger the Engine Size, the higher the CO2 Emissions.

The Smog Rating is always either a 1, 3, 5, 6 or 7, with the exception of a single record with the Smog Rating of "8" in the original csv. This "8" outlier has now been removed, bringing the total number of records to 919 rather than 920.

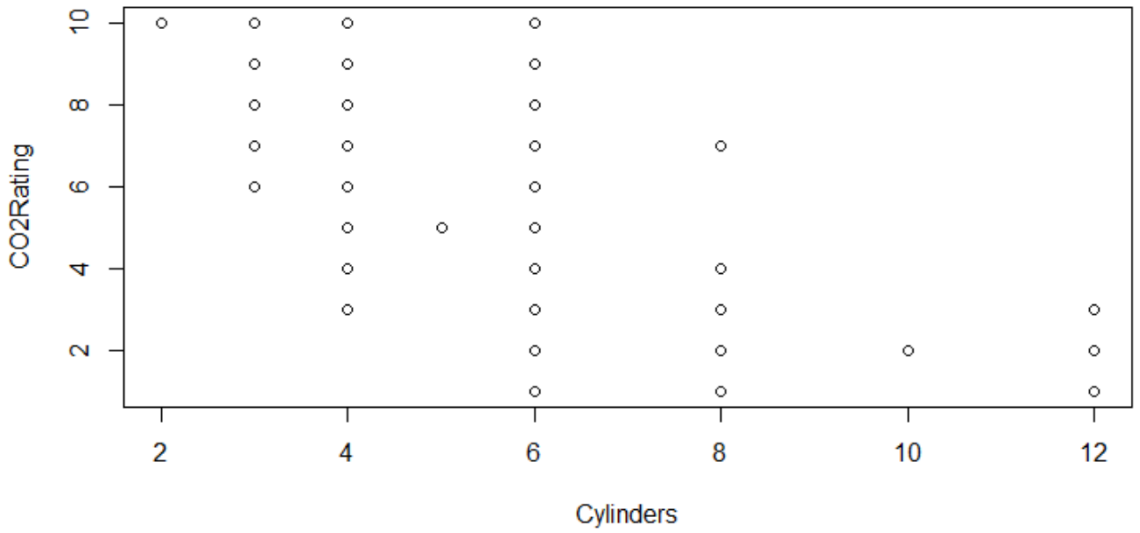


Figure 5: CO2 Rating by number of Cylinders.

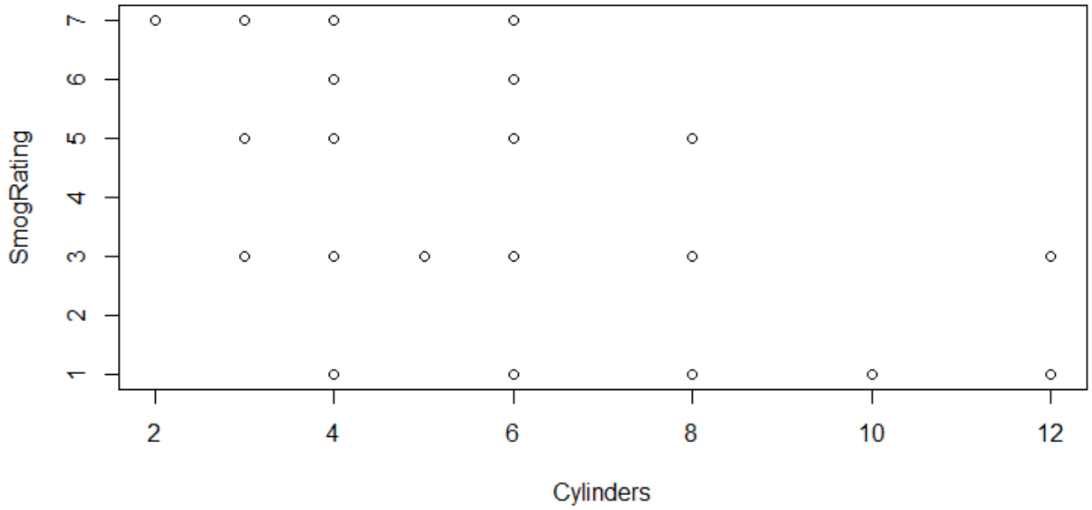
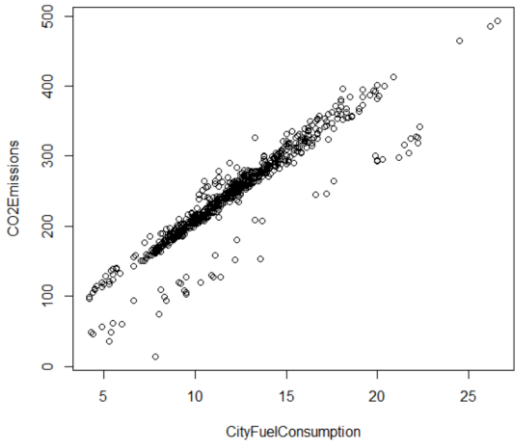


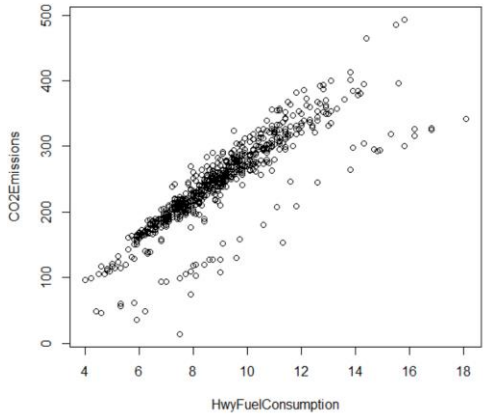
Figure 6: Smog Rating by number of Cylinders.

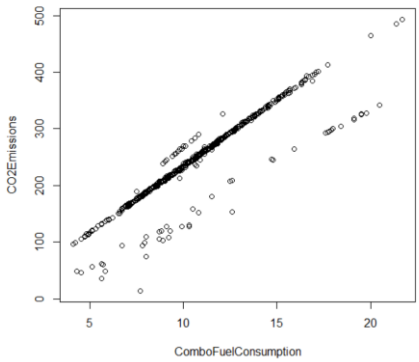
The dataset contains two records with 2 cylinders, two records with 5 cylinders and six records with 10 cylinders. The occurrence of vehicles with these respective amounts of cylinders are not high; however, 2, 5 and 10 cylinders happens more than once in each case. It looks like the rating process has rated each vehicle with two cylinders exactly the same in terms of CO2 Rating and Smog Rating. The same goes for each vehicle with 5 cylinders and each vehicle with 10 cylinders. The two 2-cylinder vehicles and the two 5-cylinder vehicles have to be removed for easier classification.

Looking at the different types of fuel consumption (city, highway or both as related to CO2 emissions, we find an interesting pattern of what resembles multiple regression lines. The offhand guess is that this stems from the difference between conventional/hybrid versus plug-in hybrid vehicle types. A similar diversion of records is seen later in CO2 Rating and Smog Rating, by both CO2 Emissions.

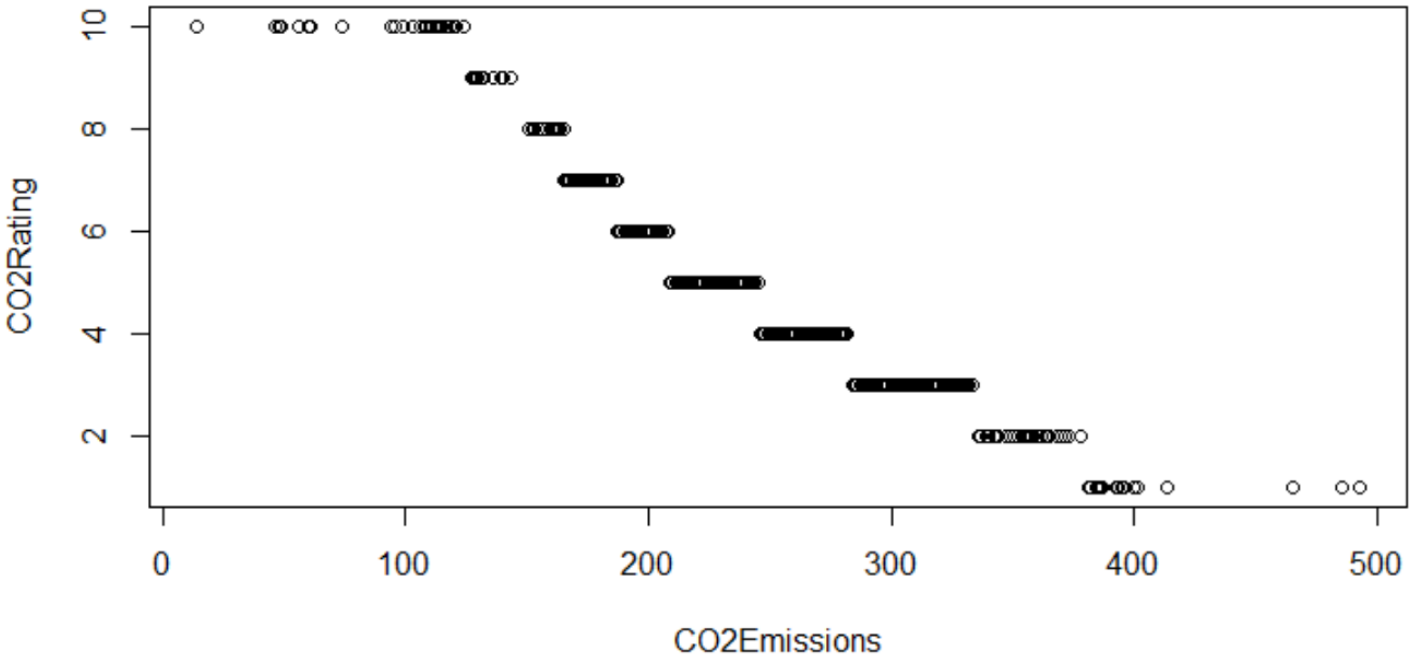
The City Fuel Consumption, Highway Fuel Consumption and Combo (combo of city and highway) Fuel Consumption are all very similar graphically, so the Combo Fuel Consumption column is the only one we'll use for analysis. The City and Highway Fuel Consumption columns have been removed from the csv dataset and Combo has been renamed to "FuelConsumption".

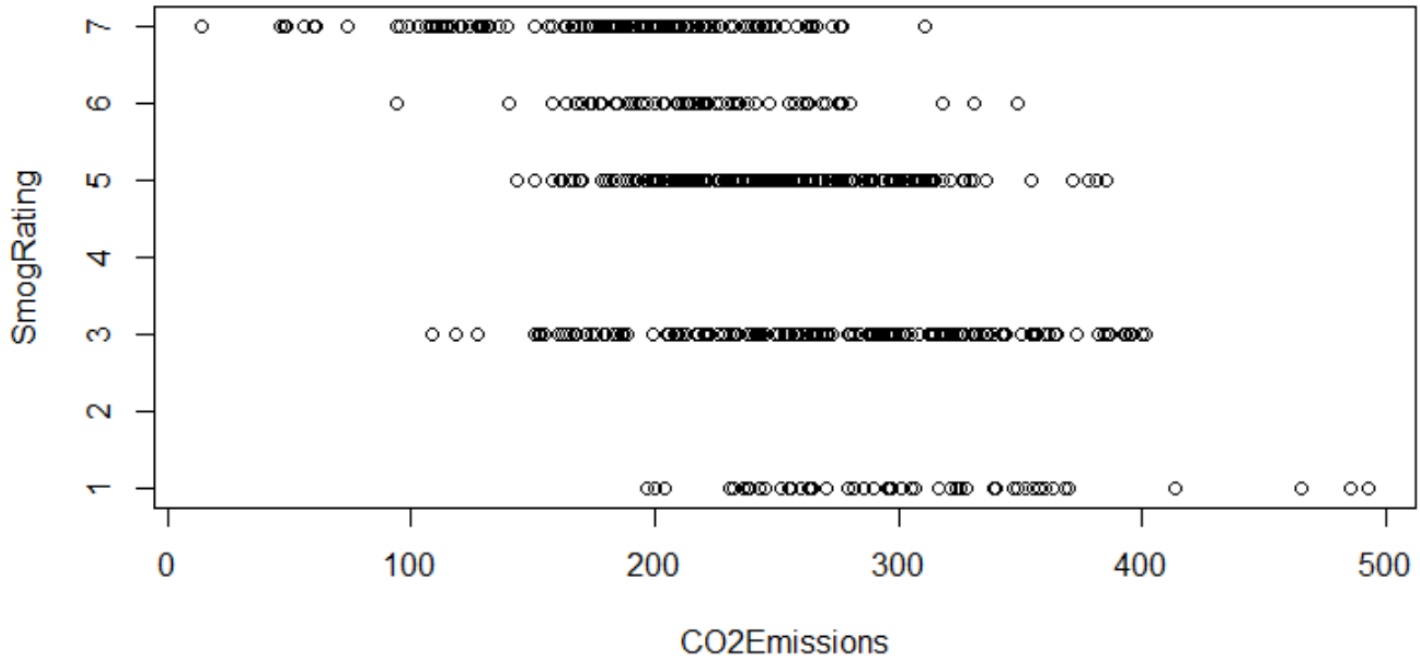


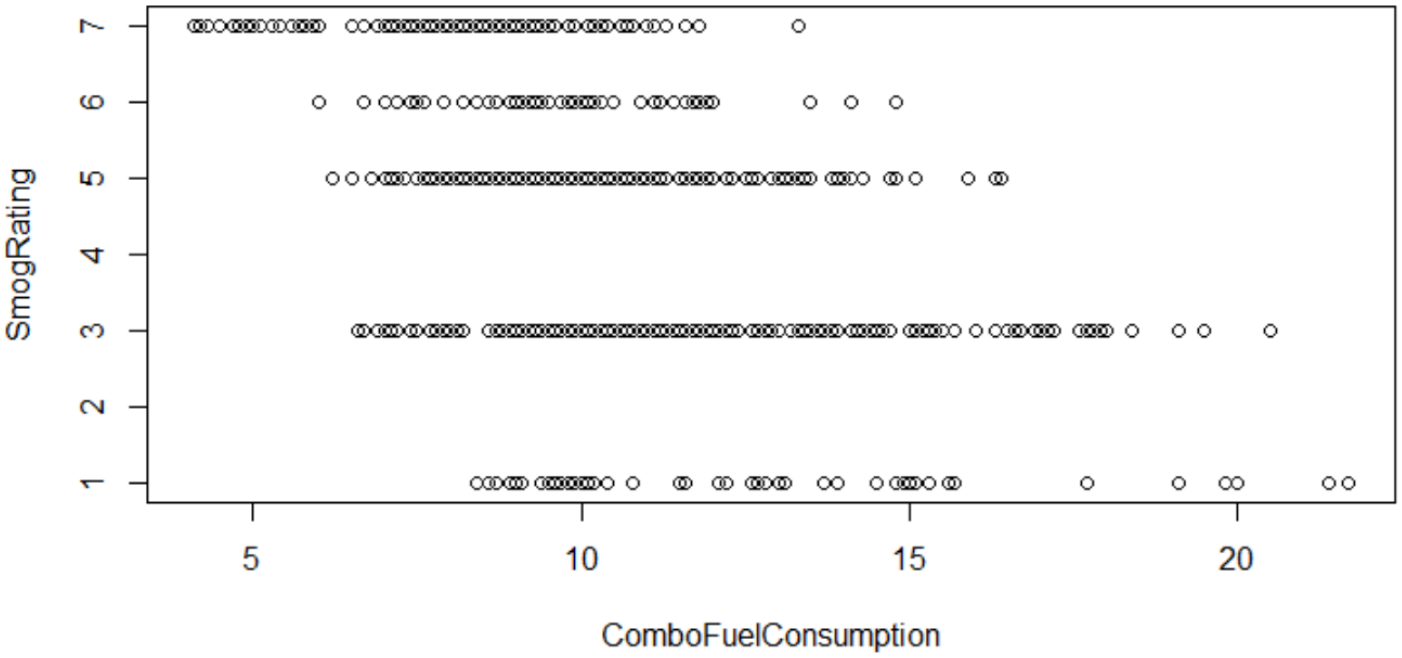


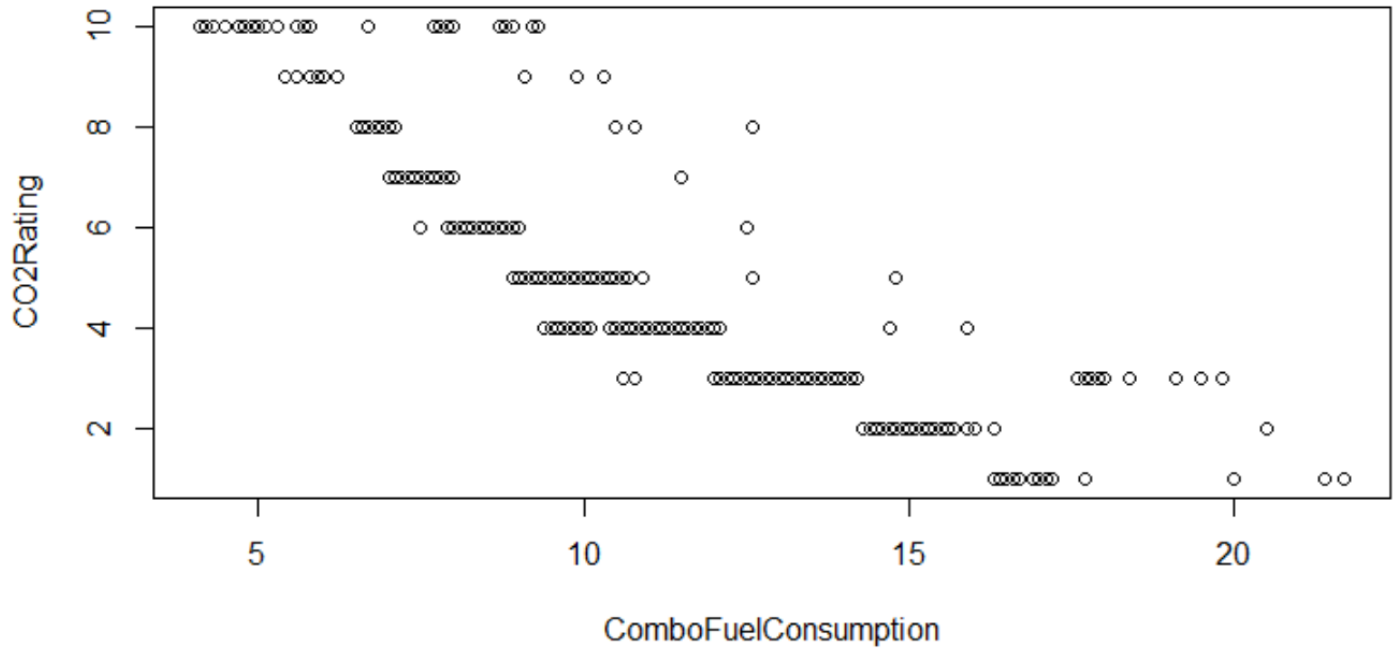


From our observations on some of the above plots, a question of bias in the ranking system might be worth exploring. We were not provided with a formula for CO2 Rating or Smog Rating; all we know is that 10 is best and 1 is worst. However, we also know that Smog Rating only falls on certain values within that ranking system. The median of CO2 emissions in the cleansed data set is 248, the median of CO2 Rating is 4 and the median of Smog Rating is 5.

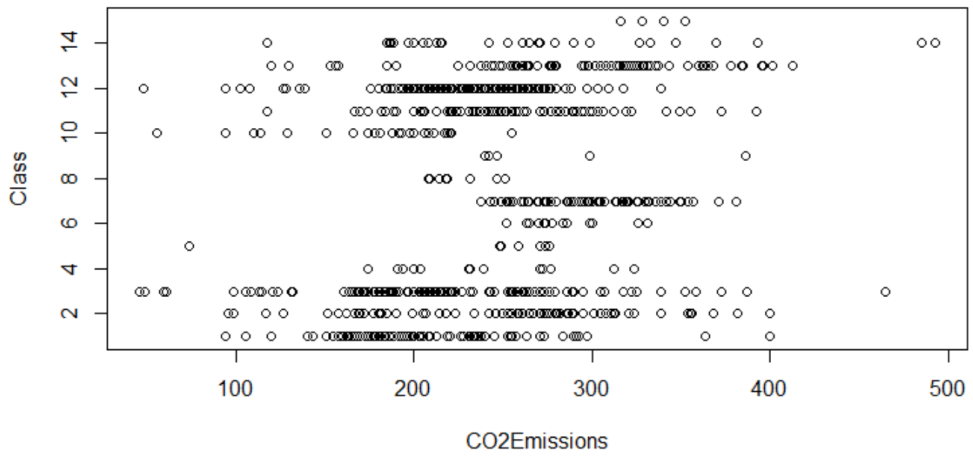


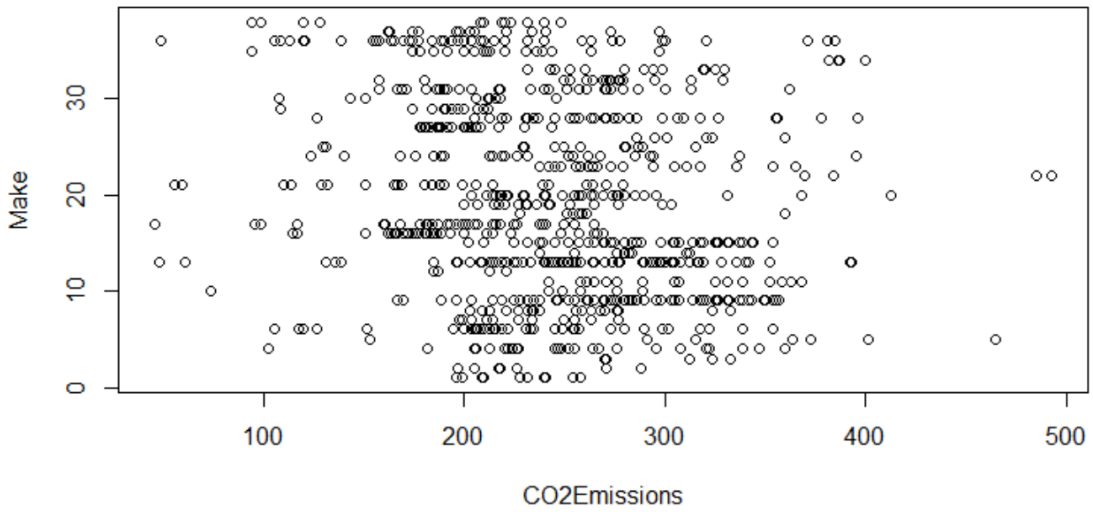




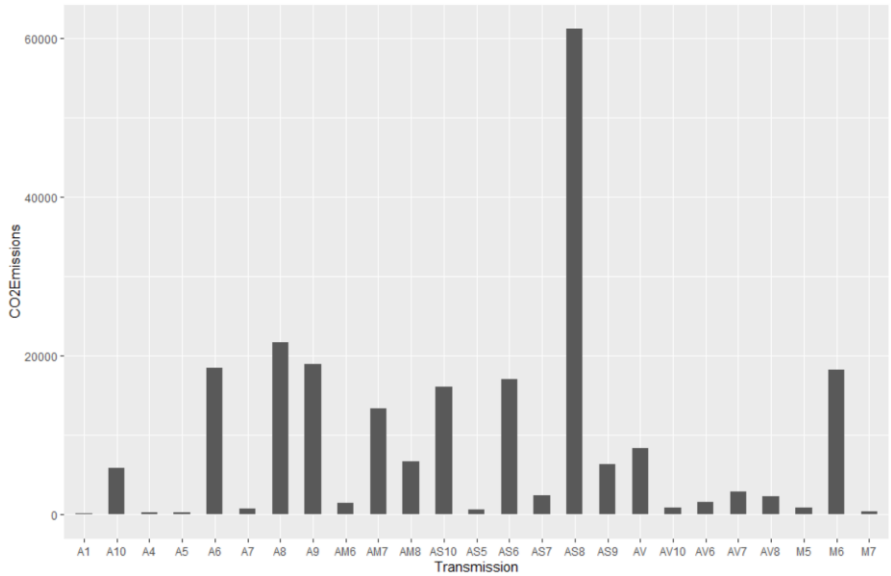


The above predictors explored visually—that is, Engine Size, Fuel Consumption and number of Cylinders—all seem to have a direct relationship with CO2 Emissions, CO2 Rating and/or Smog Rating. Vehicle Class and vehicle Make, however, do not seem to have much effect on the response variable (CO2 Emissions seems the most appropriate for a response variable).





Some transmission models are associated with higher CO2 emissions over others. After visualizing this with R and after going back to the transmission models to look at which transmissions are used on regular gasoline and premium gasoline fuel-type cars, we find that the only transmission model not available for regular and premium gasoline using cars is "A1" (which also is associated with the lowest emissions).



The information on Country of Origin—as well as which country a car brand is sold the most in—is pulled from the Internet; therefore, its reliability is questionable. However, we see in the below plot that the USA and Italy have the highest CO2 Emissions by country of origin, which may make sense, given both countries' history in sports and muscle car manufacturing, as well as their dedication to race car driving. We remember from the ISLR Auto data set that, in the 1980's, the USA was the country of origin that produced the least-best mpg. Although we do not have mpg in this data set, we will next explore Country of Origin and High Sales Country by Fuel Consumption. Fuel Consumption and Emissions are comparable by Country of brand Origin as well as by High Brand Sales Country.

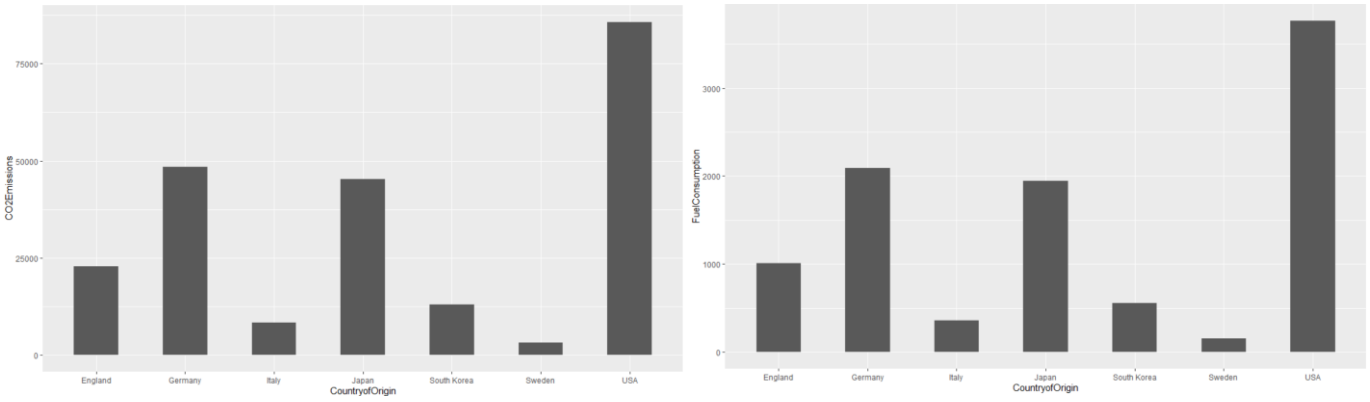


Figure 7: CO2 Emissions and Fuel Consumption by Country of Parent Brand Origin

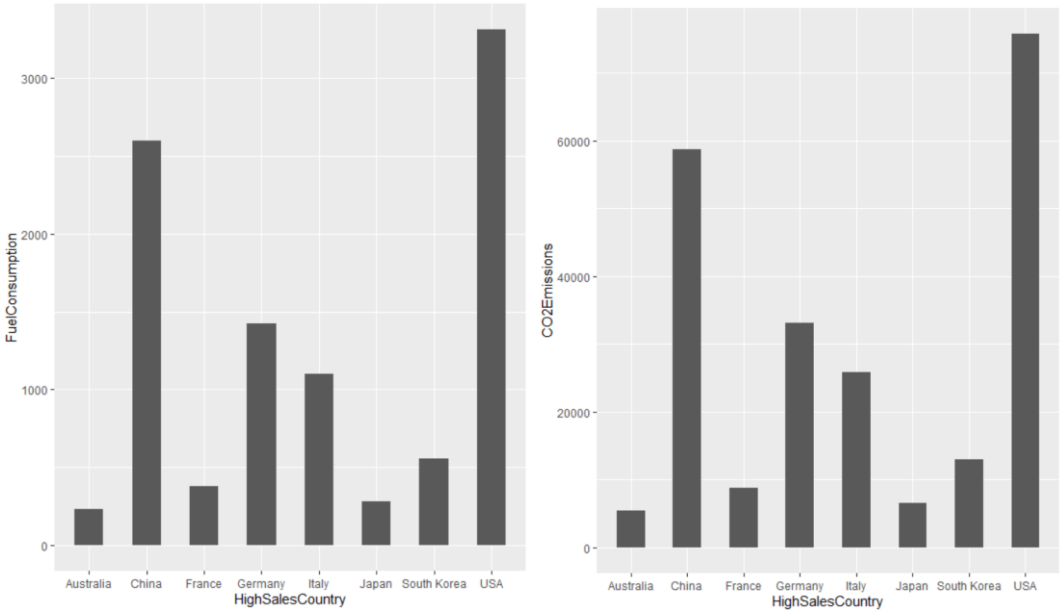
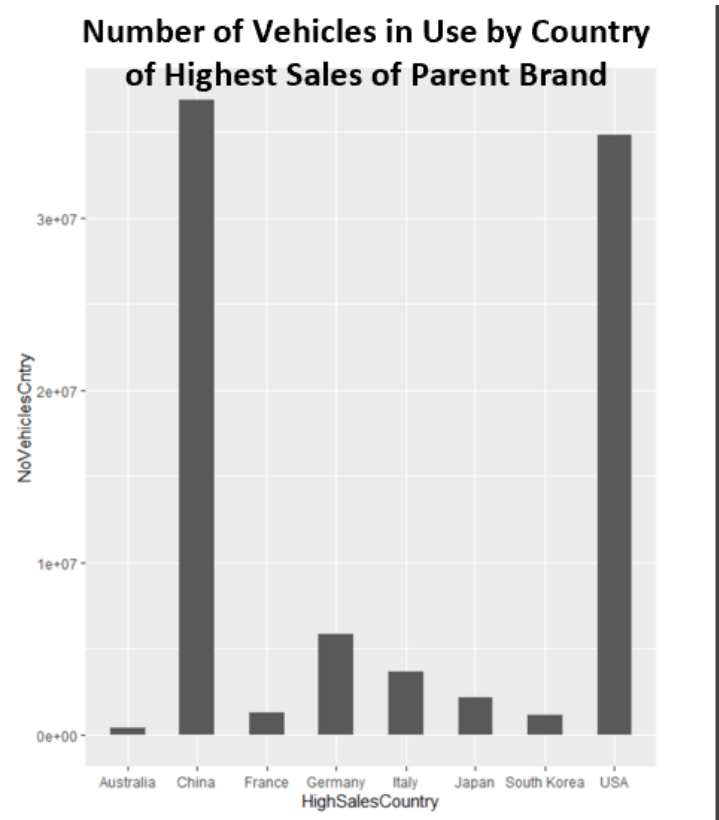
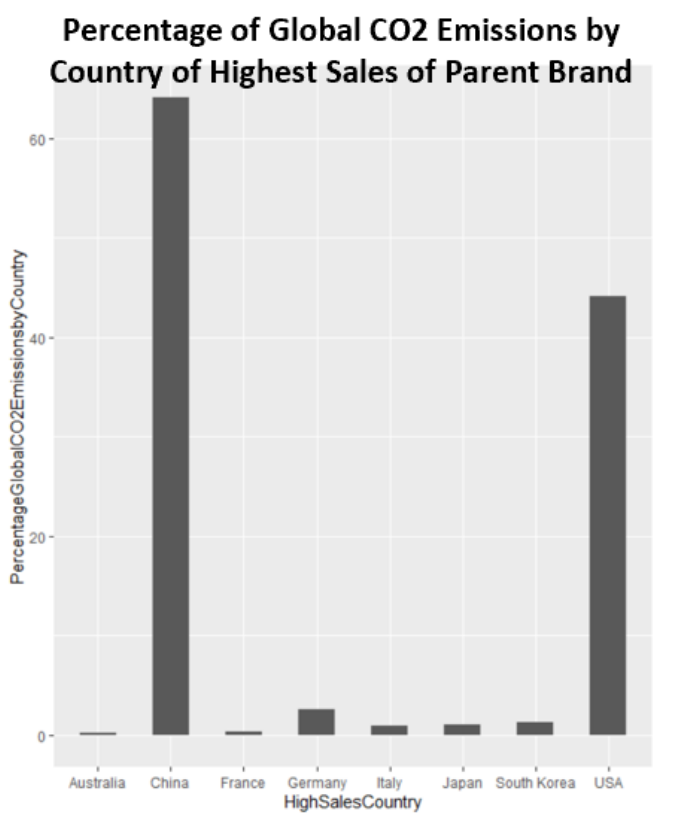
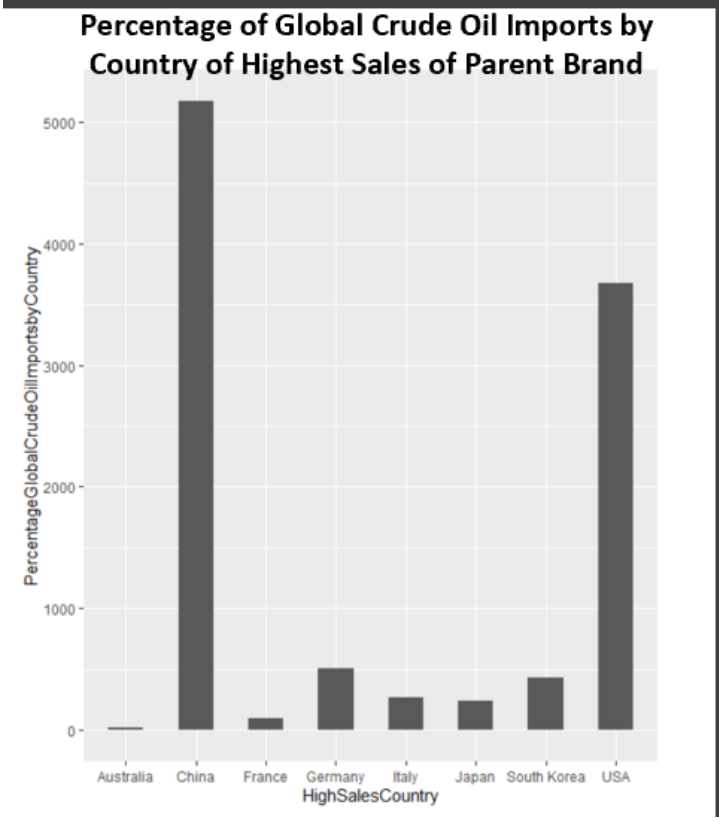


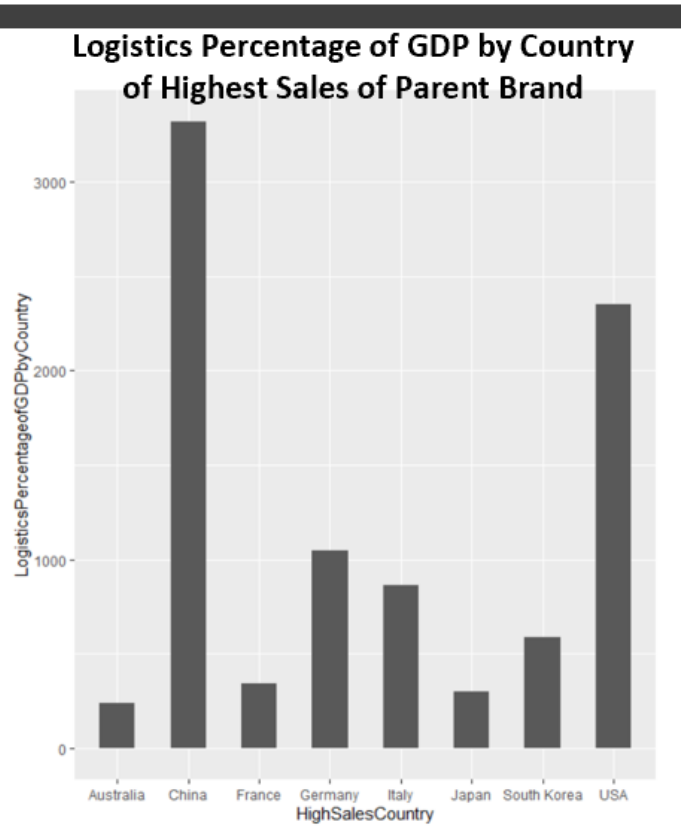
Figure 8:CO2 Emissions and Fuel Consumption by Country of Highest Parent Brand Sales

The columns pulled from Internet data can be viewed as a data set unto themselves and in fact should most likely be analyzed separately. The interplay between vehicle data from the Canadian government and international data from the internet would be more relevant if we had additional layers, like how much of passenger transportation contributes to national transportation-created CO2 emissions by country and how much of total car sales in each country is represented by each model or each parent brand in the Canadian data set. It may, however, be worthwhile to take a look at some of the internet-derived variables in order be better able to explain or contextualize what we find in the Canadian data set.









From these visuals on Countries of Highest Sales by Parent Brand, we see that—of course—China and the US lead in the number of actively used vehicles in each country, the percentage of global CO2 emissions they each produce and the percentage that logistics comprises of their national GDP. Interestingly, especially when contrasting to the percentage of global crude oil imports and the percentage of GDP from the logistics sector by country, we can see that the relationships for Germany and Italy are not as direct as they are for the US and China. Similarly, Germany has a

disproportion relationship of number of actively driven vehicles compared to the percentage of global emissions by country. Could these trends indicate that Germany might be leading in electric-only vehicles and transportation?

The overall view of the data on countries is based on the set of 2020 vehicle models that were tested (and therefore sold) in Canada. The statistics derived from online do not all come from the same year (by category), but they come from the latest year available. Additional layers of transparency on how much each vehicle model is driven by time, distance (both city and highway), idle time, emissions during idle time and as how many vehicles of each model from the Canadian data set (compared to all other vehicles sold) arere sold in all relevant countries would be needed to get a better overall view of the emissions contribution from a given year of sales. Such data is most likely proprietary or protected to some extent, by companies or by governments. In either case, the exploration of overall car model contributions to CO2 emissions extends beyond the scope of this exercise, which has the purpose of discovering which traits that we have data for are the most important predictors in CO2 emissions.

REGRESSION

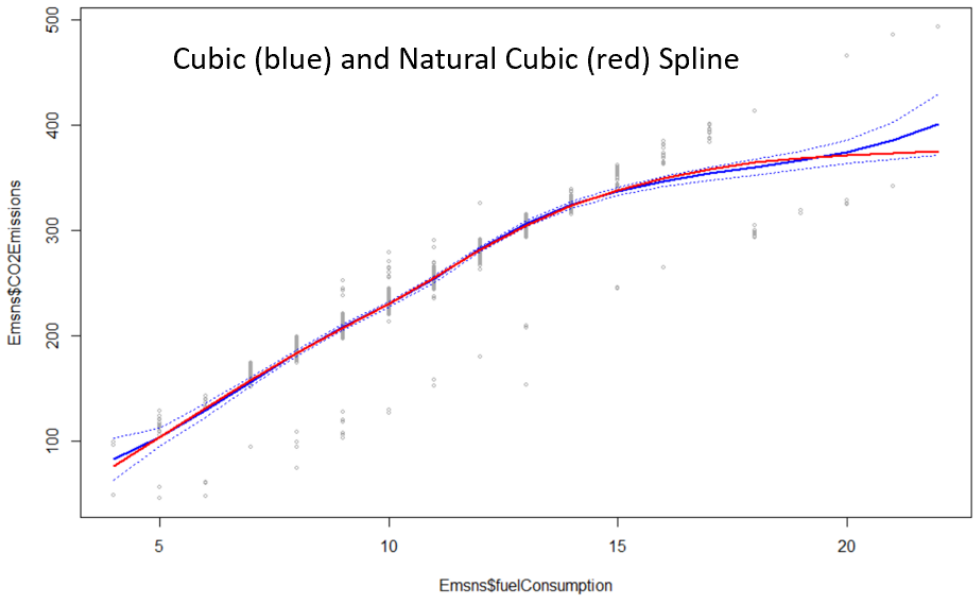
Multiple Linear Regression[xv]:

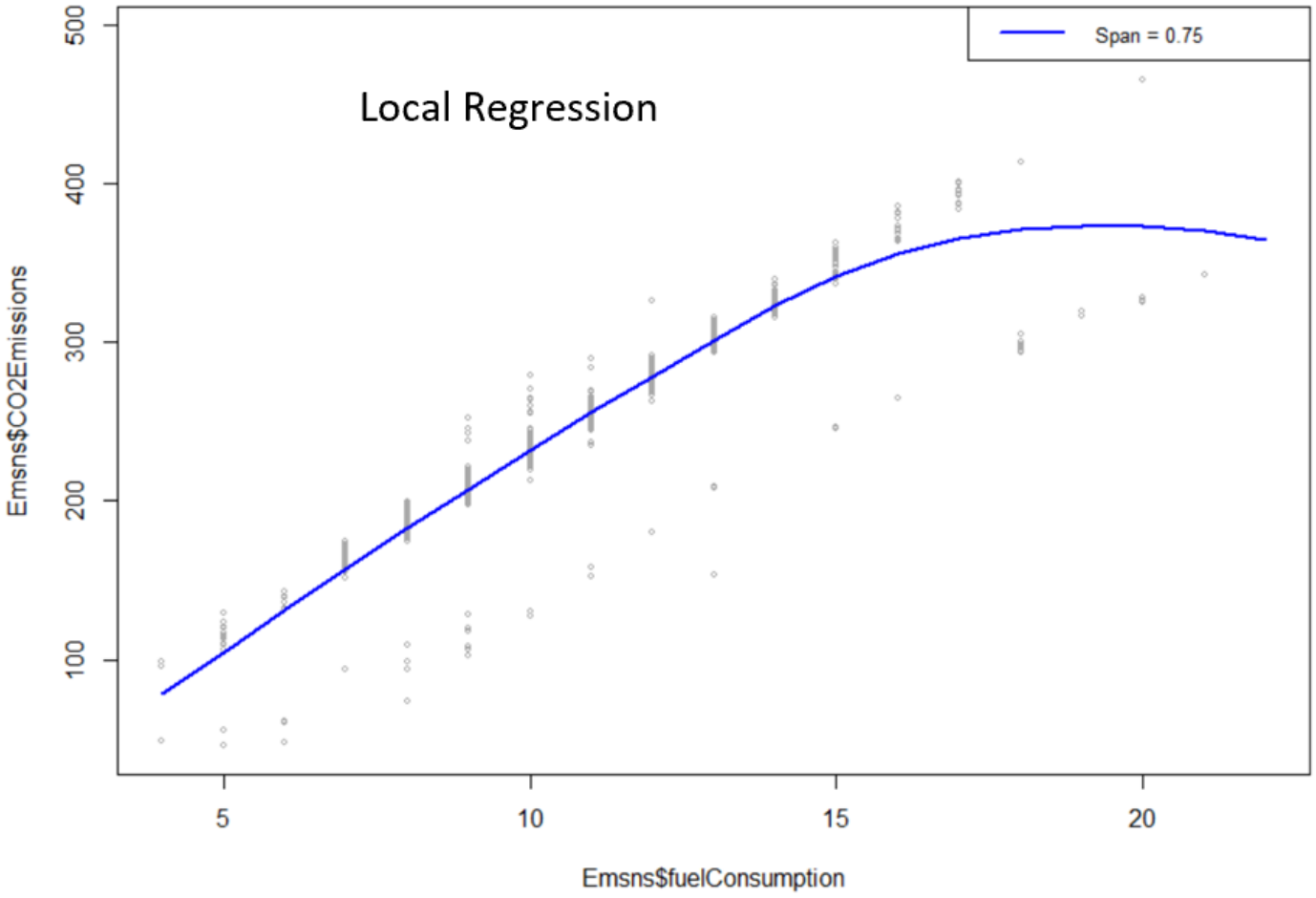
After fitting a linear regression model with the code lm(CO2Emissions ~ EngineSize + Cylinders + FuelConsumption, data = Emsns), we get a regression equation of:

CO2Emissions = 22.7510 + ((3.1423)EngineSize) + ((3.3626)Cylinders) + ((18.1984)FuelConsumption). All three predictors, Engine Size, number of Cylinders and amount of Fuel Consumed, seem to have a significant effect on CO2 Emissions. Noting that the adjusted R-squared value for the above model is .8649 and the MSE for this model is 587.5514.

Binary Logistic Regression:

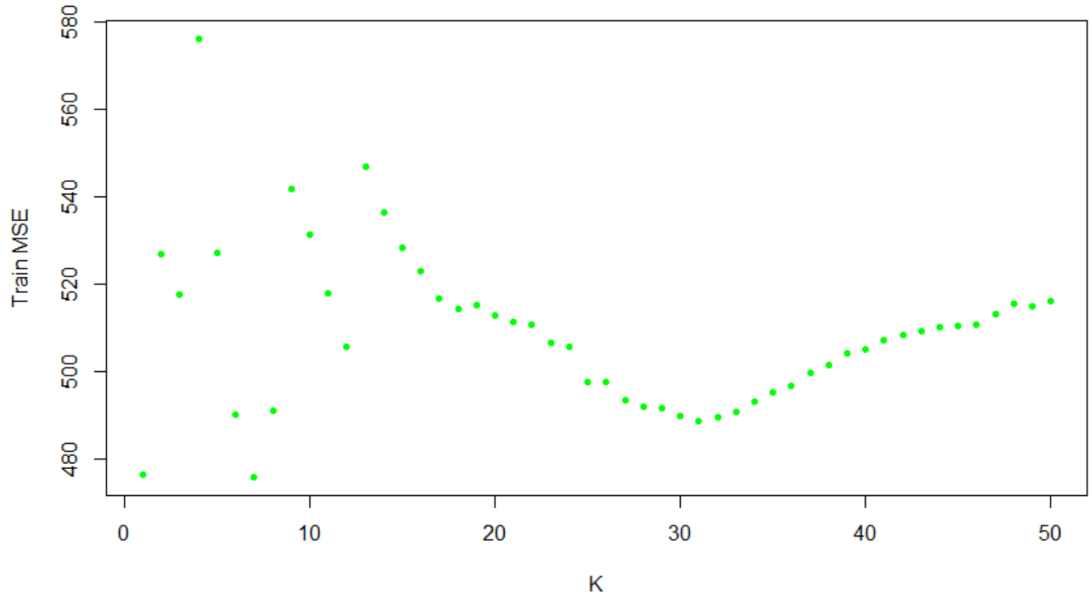
The logistic regression model for this data set—when classifying CO2 Emissions as a high level at over 248, the median CO2 Emissions number and as a low level if less than or equal to 248 and when using engineSize + Clndrs + fuelConsumption as the predictors—generated a prediction error rate of 7% and a success rate of 93%.[xvi]





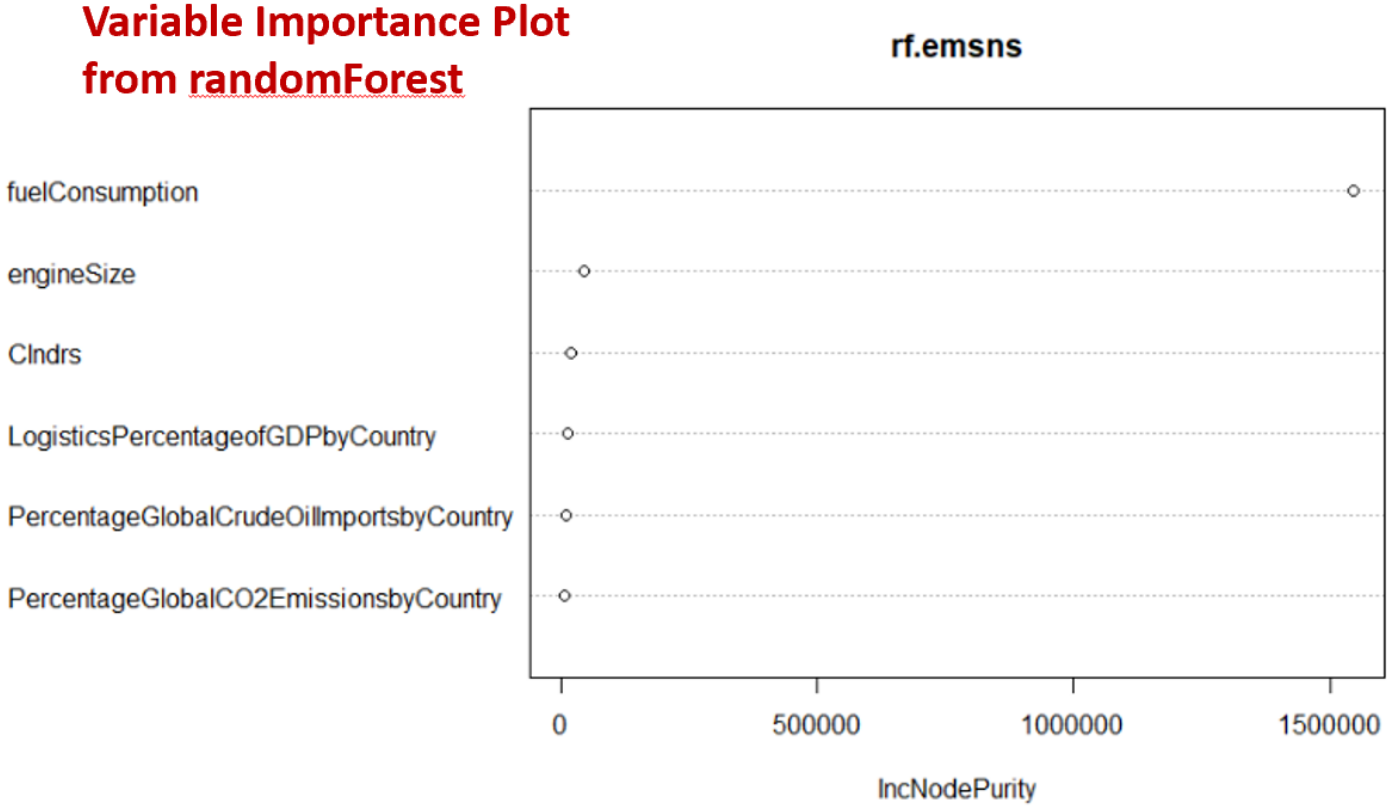
From the splines and the local regressions, we can begin to see that there isn't much to this data set as far as predictors. The outliers in almost every chart involving CO2 emissions and a dependent (numerical) variable—either engine size, number of cylinders or fuel consumption—are now most definitely from the plug-in hybrid models, but, since there is a small number of these records, the data is not too badly skewed.

KNN Regression:

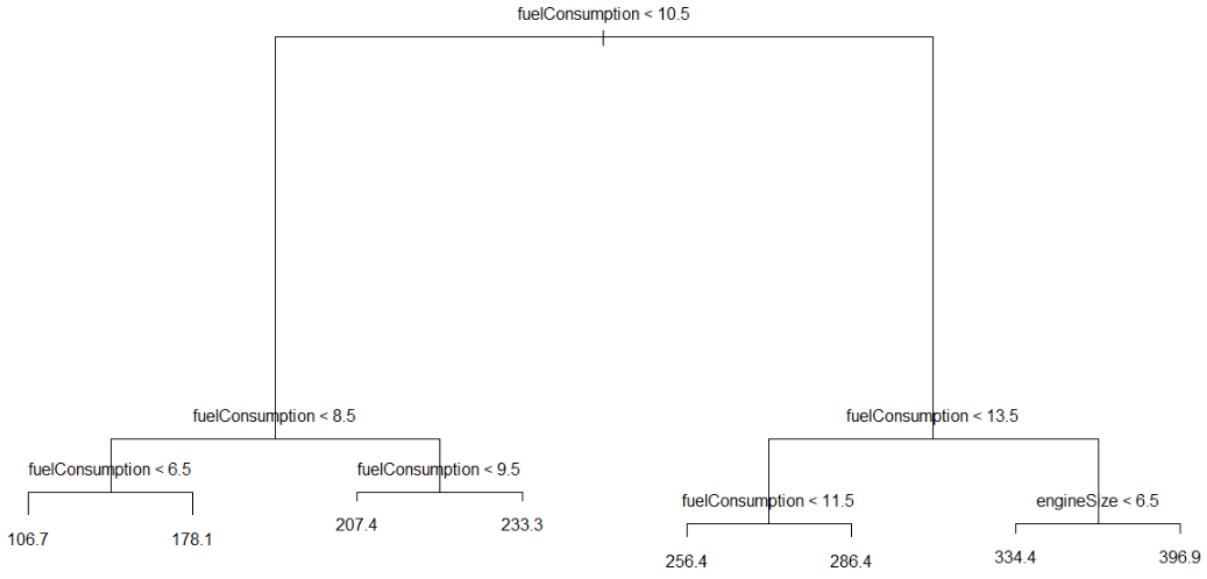


When we explore different k's for KNN regression, we see that optimal k is around 31 and optimal MSE is around 490. The high k and the high MSE here may be telling us that we do not have the variables necessary to accurately predict CO2 emissions by car model. The ISLR Auto data set, for example, had mpg and horsepower among other predictors that could be very useful in predicting emission by model. Skipping ahead, we see from the variable importance plot from randomForest that really only fuel consumption is of much significance in terms of its effect on CO2 emissions.

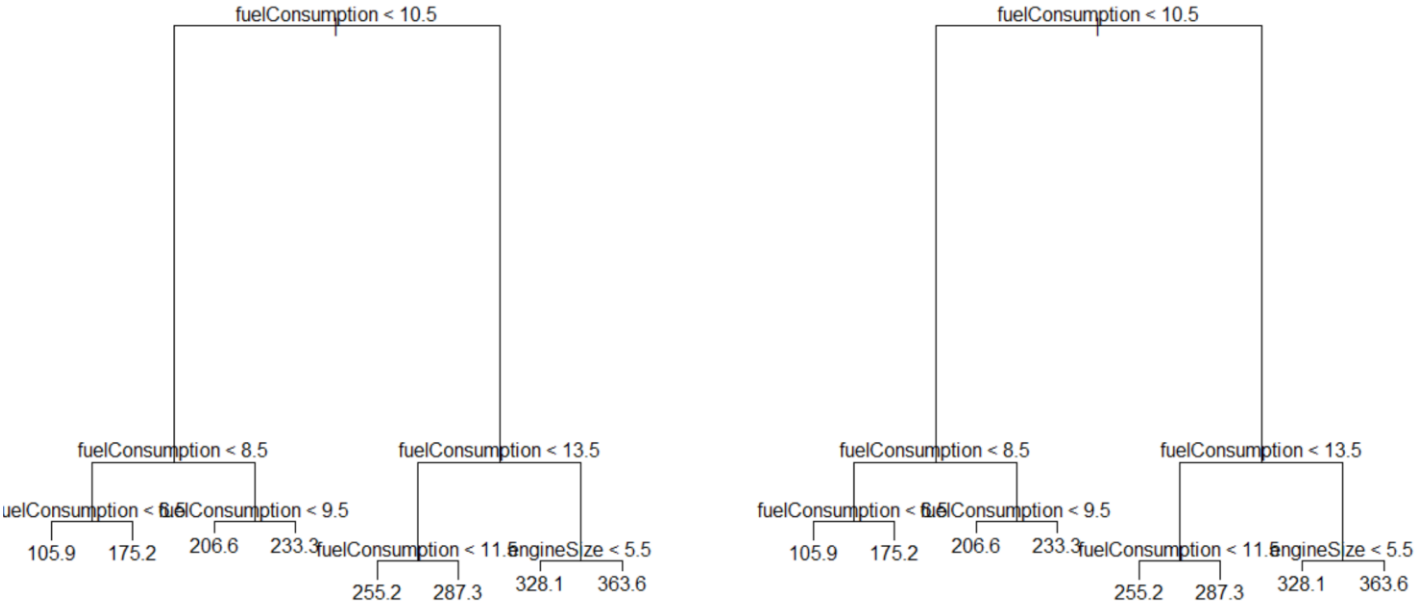
Variable Importance Plot from randomForest



DECISION TREE



The fact that fuel consumption is the only important predictor (usually) is reflected in the above decision tree, whose leaves branch out at various levels of fuel consumption, only once (out of four leaf breaks) making nodes at engine size. The MSE for this model is slightly improved over linear regression, at 578.2233. The MSE for the test set is 577.0371.



When pruning was run on this decision tree, no pruning could trim the tree or improve the results.

Bagging:

The MSE for bagging is slightly improved overall, at 469.4370.

Random Forest:

The MSE for random forest is slightly improved overall, at 470.6758.

RESULTS

Linear Regression	Decision Tree	Bagging	Random Forest
587.5514	577.0371	469.4370	470.6758

Bagging resulted from the best MSE, at slightly better than random forest.

In general, this dataset was interesting to build. However, it has also been a good learning opportunity for where the author's knowledge gap is on auto components and functions, and what vehicle-related variables contribute more (or don't contribute more) to carbon emissions. Fuel consumption in concept is not too different from the idea of mpg. If mpg, horsepower and maybe displacement or acceleration could also be added to this data set, as well as potentially some of the other variables mentioned above, we might be able to build a better picture of just how certain car models (used in Canada) contribute to CO2 emissions nationally (in Canada) and globally.

Further exploration should include the use of a GAM. A further-developed data set might be improved with the addition of other external variables, like emissions produced during car model manufacturing.

[i] https://www.epa.gov/greenvehicles/fast-facts-transportation-greenhouse-gas-emissions
[ii] https://www.planete-energies.com/en/medias/close/global-transportation-sector-co2-emissions-rise#:~:text=Passenger%20travel%20is%20responsible%20for,CO2%20emissions%20from%20fuel

[iii] https://ourworldindata.org/co2-emissions-from-aviation#:~:text=In%202018%2C%20it's%20estimated%20that,CO2%20emissions%20in%202018.&text=Aviation%20emissions%20have%20doubled%20since%20the%20mid%2D1980s.

[iv] https://youmatter.world/en/plane-or-cars-which-means-of-transport-pollutes-the-most/

[v] https://www.canada.ca/en/environment-climate-change/services/environmental-indicators/progress-towards-canada-greenhouse-gas-emissions-reduction-target.html

[vi] Although I also found 2020 model information, including variables called 'CO2 emissions', 'CO2 Rating' and 'Smog Rating' for electric battery vehicles in the document from the same site called "MY2012-2020 Battery Electric Vehicles", this information was largely useless in looking for the factors that contribute to greater pollution or that could be better improved in hybrid or petroleum-powered vehicles. This is in part because, for all electric battery vehicles, the Canadian government rated the CO2 Emissions level as "0", CO2 Rating as "10" and the Smog Rating as "10". Also, comparing models in the electric battery vehicles data set to hybrid or gas-powered vehicle models would have been difficult because the physical components of the different vehicle types are different; the electric vehicle set has a "Motor in kW" and a "Transmission" variable, whereas the hybrids and gas-powered models have "Engine Size" and "Cylinders", for example. Also different are the units of measure for the amount of energy (as well as the energy type) consumed by (100) kilometer(s).

[vii] https://www.consumerreports.org/cars-who-owns-which-car-brands/

[viii] https://www.canstarblue.com.au/vehicles/car-country-of-origin/

[ix] https://www.factorywarrantylist.com/car-sales-by-country.html

[x] https://www.ucsusa.org/resources/each-countrys-share-co2-emissions

[xi] http://www.worldstopexports.com/crude-oil-imports-by-country/

[xii] https://www.3plogistics.com/3pl-market-info-resources/3pl-market-information/global-3pl-market-size-estimates/

[xiii] https://www.nationmaster.com/nmx/ranking/global-passenger-cars-in-use

[xiv] A few things to note about the nature of the data:

1. For the plug-in hybrid electric vehicles there were two fuel types columns in the original data set; one was a combo type that combined the class of electric operating system/electric engine type with the type of conventional fuel the car also uses. I did not include the second fuel type column in my compiled data set, which was repeated information about the type of conventional fuel the car uses.
2. I used two ready-made reports from the site listed above. However, at https://fcr-ccc.nrcan-rncan.gc.ca/en, one can customize a report by "vehicle type" and other parameters. In 'vehicle type', the site generalizes conventional and hybrid vehicles together but also gives the option of plug-in hybrid electric vehicles as a (non-distinct) category. A Toyota Prius is listed under 'conventional/hybrid vehicles' in this tool, for example, when one also selects gasoline as the fuel type (and "all" is selected in the other categories under 'advanced search'). When the 'conventional/hybrid vehicle' type is selected and 'electricity' is selected as the fuel type (and "all" is selected in the other categories under 'advanced search'), there are no results. When 'plug-in hybrid electrical vehicle' type is selected and 'electricity' is selected as the 'fuel type' (and "all" is selected in the other categories under 'advanced search'), there are 161 results, all of which have 'combo fuel types' as described in bullet #1.
3. Vehicle manufacturers are not required to submit fuel consumption data all vehicles; for example, SUV's and passenger vans with large gross vehicle weight ratings (>10,000 lbs) are exempt. See https://www.nrcan.gc.ca/energy-efficiency/energy-efficiency-transportation/2020-fuel-consumption-guide/understanding-fuel-consumption-r/fuel-consumption-testing/21008 for more information on testing parameters, methods and metrics.

```
Call:
lm(formula = CO2Emissions ~ engineSize + Clndrs + fuelConsumption,
    data = Emsns)

Residuals:
     Min       1Q   Median       3Q      Max
-133.555   -2.344    3.724   10.656   65.282

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      23.5158     3.3746   6.969 6.14e-12 ***
engineSize        3.3399     1.4336   2.330     0.02 *
Clndrs            4.1431     1.0584   3.915 9.73e-05 ***
fuelConsumption  17.5508     0.4885  35.927  < 2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24.29 on 911 degrees of freedom
Multiple R-squared:  0.8558,    Adjusted R-squared:  0.8553
F-statistic:  1802 on 3 and 911 DF,  p-value: < 2.2e-16

> model1_summary <-summary(model1)
> mean(model1_summary$residuals^2)
[1] 587.5514
> |
```

xv

xvi

```
> contrasts(Emsns$EmsnsLevel)
     Low
High   0
Low    1
> glm.probs.Emsns <- predict(glm.fits.Emsns, type = "response")
> glm.probs.Emsns[1:10]
          1           2           3           4           5
0.990233374 0.003664567 0.990233374 0.990233374 0.246390109
          6           7           8           9          10
0.992251910 0.992251910 0.914921355 0.992082829 0.999974267
> glm.pred.Emsns <- rep("high.pred", dim(Emsns)[1])
> glm.pred.Emsns[glm.probs.Emsns > 0.8] = "low.pred"
> table(glm.pred.Emsns, Emsns$EmsnsLevel)

glm.pred.Emsns High Low
     high.pred  451  61
     low.pred     3 400
> (451 + 400) / dim(Emsns)[1] #rate of the diagonal term =
[1] 0.9300546
> #success rate = 93.00546
> (3 + 61) / dim(Emsns)[1] # training error rate or mis-classification rate =
[1] 0.06994536
> |
```