STAT 4355

Spring 2023

CO2 Emission Analysis

JUDY's Carbon Crusaders:

Jonathan Serrano, Umar Ali-Salaam, Daniel Li, Yaseen Mohammed

Table Of Contents

Introduction

Carbon dioxide (CO2) is a potent greenhouse gas that all motor vehicles emit. It is a byproduct of the combustion of fossil fuels in car engines, contributing significantly to global warming and climate change. The rising levels of CO2 in the atmosphere have caused severe environmental consequences, including rising temperatures in both the atmosphere and oceans, melting glaciers (which leads to rising sea levels), and more frequent severe weather/natural disasters worldwide.

This statistical report will focus on analyzing the state of CO2 emissions from various types of vehicles in Canada and their characteristics over a period of 7 years. The goal of our analysis is to determine which factors of a car, such as fuel consumption, the amount of cylinders, fuel type, and engine size, are best at modeling the amount of correlation to the CO2 emission.

By understanding which features of a car cause the CO2 we hope to provide insights that can inform the public on which vehicle and what characteristics in a car to look for that would produce lower carbon emissions and promote a sustainable future for the planet.

Data Description

We found our dataset on Kaggle, however the origin of the data is from the Canadian Government official open data website. Since the data originates from a government institution, we believe that the data is reliable. The data contains 7385 observations of vehicles from a period of 7 years, 6282 of them being unique. There are 12 columns that describe certain features of the car, statistics of the fuel, and CO_2 emissions. These features include both discrete and continuous types.

Variable Attributes

Our dataset contains 12 variables, where the CO2 emission will be our response variable. We believe that there are two variables, namely the Model and Fuel Consumption Combined in MPG, that are either redundant or may have minimal or no impact on predicting CO2 emission. By eliminating those 2 variables and using the 9 other variables as our predictor variables, we will create our linear regression models and check which predictors have a positive or negative correlation with CO2 emissions. These are the following variables we chose:

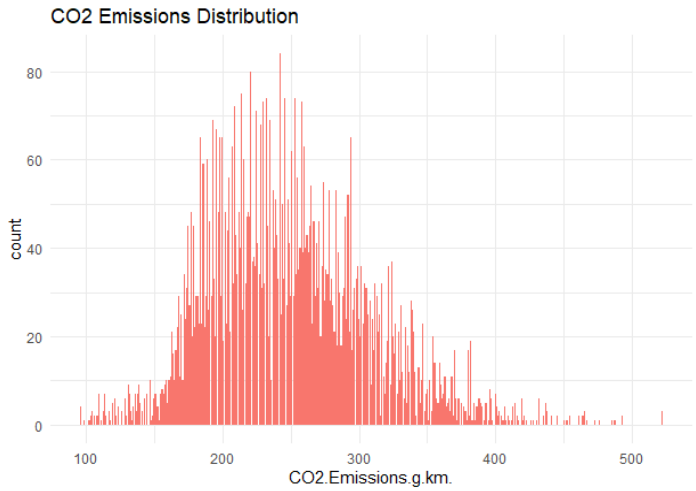
1. **CO2 Emissions(g/km) (Ranges from 96 to 522):** The tailpipe emissions of carbon dioxide (in grams per kilometer) for combined city and highway driving.
2. **Make (Discrete):** Make of car with details of manufacturer.
3. **Vehicle Class (Discrete):** Class of vehicle depending on their utility, capacity and weight.
4. **Engine Size (Ranges from 0.9 to 8.4):** Size of engine used in Litre.
5. **Cylinders (Discrete):** Number of cylinders.
6. **Transmission (Discrete):** Transmission type with number of gears.
7. **Fuel Type (Discrete):** Type of Fuel used.
8. **Fuel Consumption City (Ranges from 4.2 to 30.6):** Fuel consumption in city roads (L/100 km).
9. **Fuel Consumption Hwy (Ranges from 4 to 20.6):** Fuel consumption in highways (L/100 km).
10. **Fuel Consumption Comb (Ranges from 4.1 to 26.1):** The combined fuel consumption (55% city, 45% highway) is shown in L/100 km.

Data Analysis

Data Cleaning

Initially reviewing our data we searched for the presence of any NA values and duplicate values to check for missing data and the number of unique observations. We found that there were no NA values and about 1000 duplicate values. We decided not to remove the duplicates since having multiple observations with the same statistics allows multiple representations of data that are more likely to be aligned with their true values. We removed the “Fuel Consumption Comb (mpg)” column from our data, because it’s redundant because we have that same column “Fuel Consumption Comb (L/100km)” but in metric units not imperial. We wanted to keep the metric unit column since all the other columns utilized the metric system as well. We also removed the “Model” column, because as previously mentioned, the “Make” column will already consider the type of car it is. Finally, we added dummy variable columns (indicator variables) for the transmission, fuel type, and vehicle class where the value would be 1 if it is of that type, and 0 if it is not . We added these indicator columns since these columns are discrete values, and applying linear regression on discrete values requires analysis on indicator variables.

Exploratory Analysis

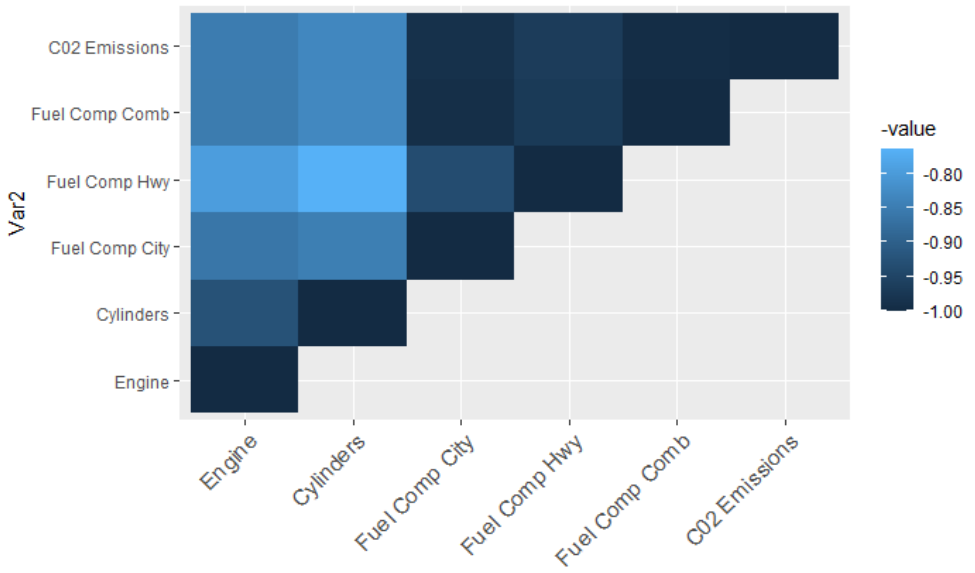


On this graph above, we can see the distribution of the emissions CO2 of every observation in our data. As seen on the graph, a majority of the emissions lie in between 150-300 g/km. The distribution also resembles a right skewed normal distribution with the average being 250 g/km and standard deviation being 58 g/km.

```{R}
length(unique(emissionsDF$Vehicle.Class) )# Could Work
length(unique(emissionsDF$Make) )# Too Many small values
length(unique(emissionsDF$Fuel.Type)) #  Could Work
length(unique(emissionsDF$Transmission) )# # Could Work
```

```
[1] 16
[1] 42
[1] 3
[1] 27
```

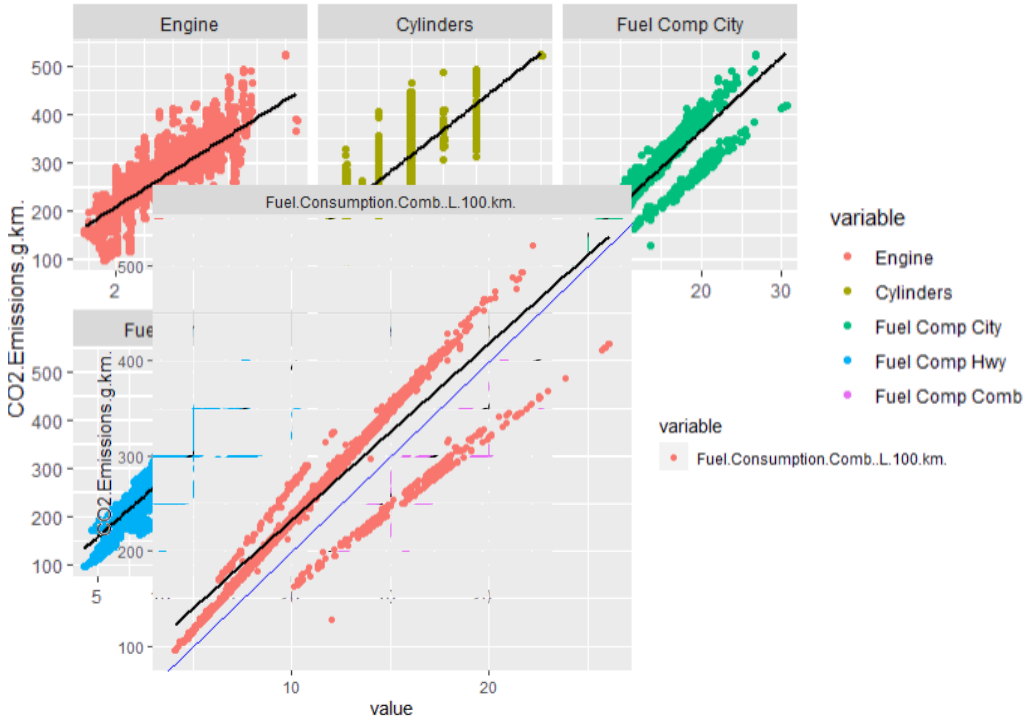
Upon examining the count of the discrete values, it was observed that there exists a substantial number of distinct categories for the "Make" variable, with some categories exhibiting considerably low frequency. Therefore, it was determined that the inclusion of such variables may result in overfitting. Consequently, these variables were excluded from the analysis.



	Engine	Cylinders	Fuel Comp City	Fuel Comp Hwy	Fuel Comp Comb	CO2 Emissions
Engine	1.0000000	0.9259199	0.8616495	0.7956075	0.8518779	0.8514051
Cylinders	0.9259199	1.0000000	0.8468025	0.7641587	0.8310884	0.8321540
Fuel Comp City	0.8616495	0.8468025	1.0000000	0.9353228	0.9926066	0.9882668
Fuel Comp Hwy	0.7956075	0.7641587	0.9353228	1.0000000	0.9709317	0.9662925
Fuel Comp Comb	0.8518779	0.8310884	0.9926066	0.9709317	1.0000000	0.9953828
C02 Emissions	0.8514051	0.8321540	0.9882668	0.9662925	0.9953828	1.0000000

This is a Correlation Matrix and their values for our continuous regressor variables and Cylinder count. We can see each variable a good amount of correlation with one another because all the values are 0.7 or higher. All the variables have a positive correlation with one another as well. We can also see that Fuel Consumption types

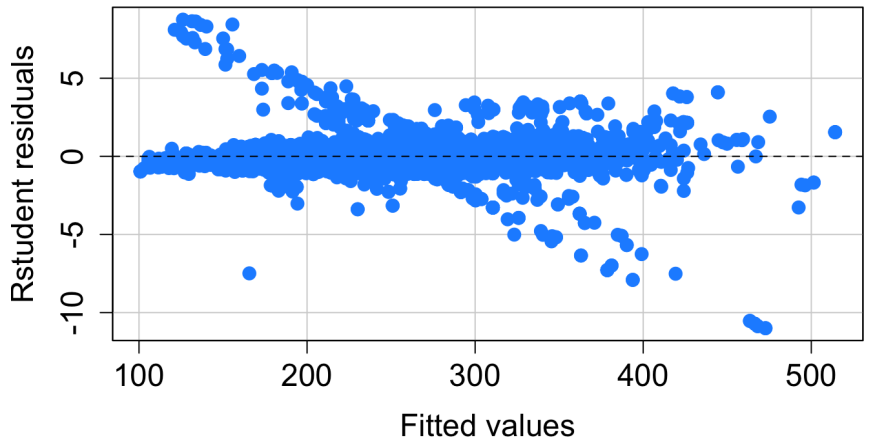
are the predictors that are most correlated with CO2 Emissions. This also shows that engine size, cylinders, and various types of fuel consumption have a negative effect by increasing emissions.

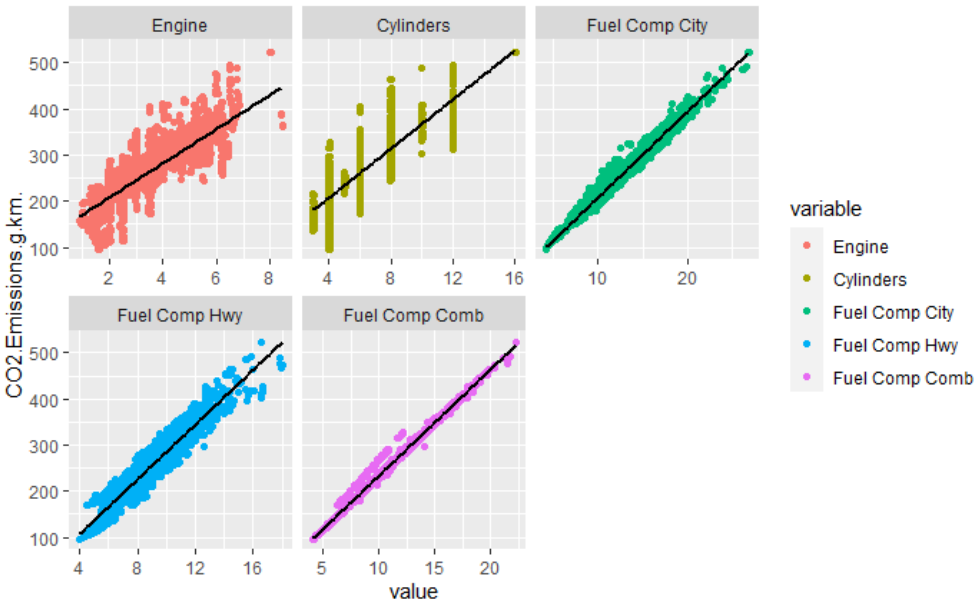


We also plotted the data by making scatter plots where each graph would correspond with one of the continuous variables or cylinder count on the x-axis and the CO2 emissions on the y-axis. This is a graphical representation of the correlation matrix we showed earlier. As expected, all the plots show a positively linear correlation with our outcome variable CO2 Emissions. This bodes well for linear regression. Even the cylinders count which is not continuous scales up with CO2 Emissions.

One thing we found and was skewing the results was a secondary distribution that was separated from the majority of the data. These observations seem to produce less carbon emissions for the same amount of their respected values. This secondary distribution caused our residual plot to have a non constant variance and non randomly distributed residuals because of a very clear line as shown below. This was somewhat alleviated if we added the model variable but that was an inadequate

solution due to the overfitting it may cause. Instead, since there were only a couple hundred observations that fell into this secondary distribution, it would be best to remove it, and only consider looking at the observations from where the majority of the data was.



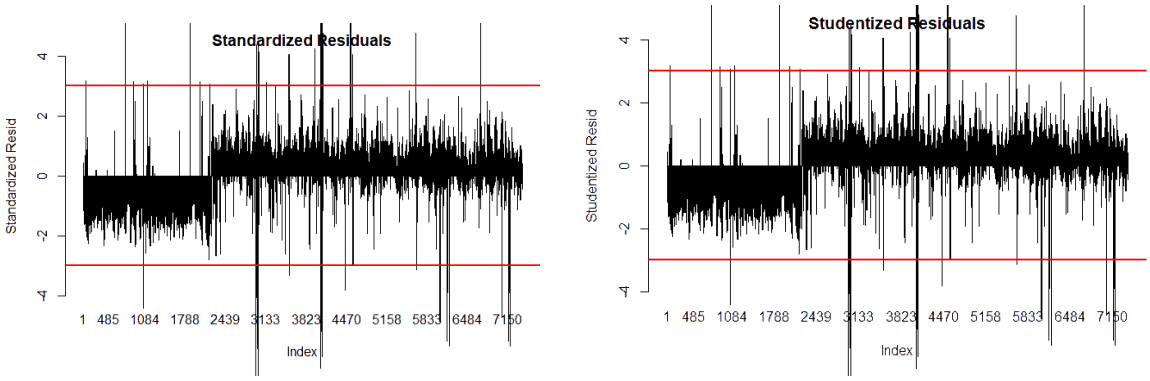


To solve this problem we decided to look at the Carbon Emissions vs Fuel Consumption scatter plot since the two distributions were clearly divided there. We used the lm() function in R to see what the linear model was and estimated a line where we could separate the 2 distributions. Once we found the line, we removed all

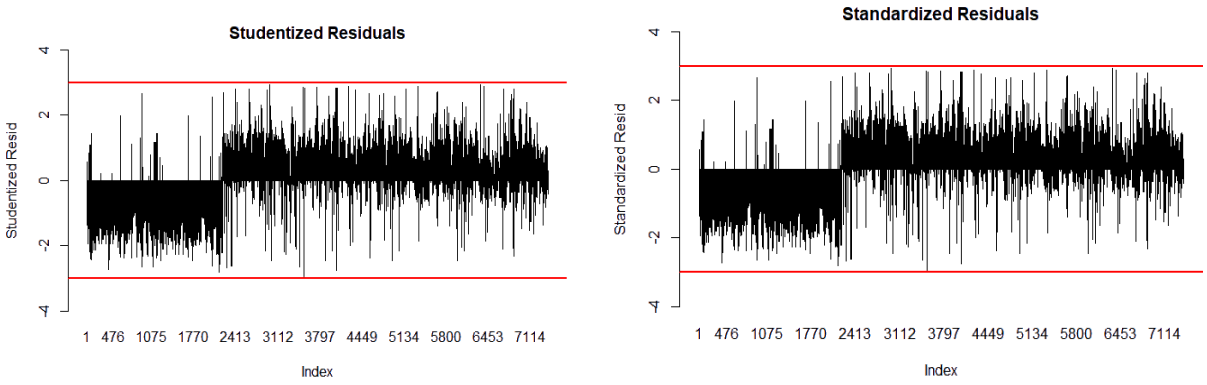
the observations from the secondary distribution. This process removed around 350 observations. We plotted the scatter plots made earlier after removing the observations and it is clearly seen that the data for each plot falls into a more centralized and compact distribution.

Residual Analysis

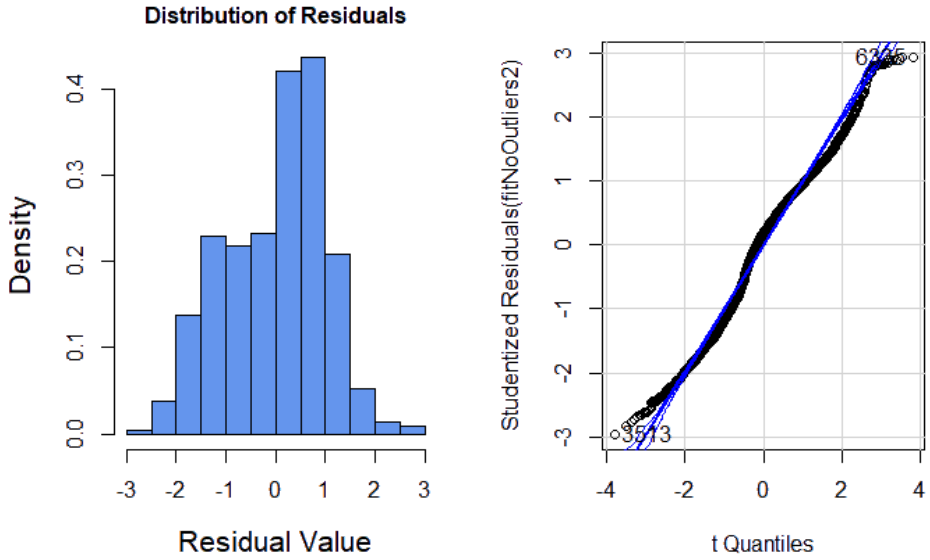
We created our residuals centered around the linear model which includes the outcome variable of CO2 Emissions. From that model, we plotted our standardized and studentized residuals. We set our cutoffs at 3 and -3 given the shape of the graph. As seen, there were some extreme outliers that were off the charts.



After removing, we're left with these simplified graphs shown below. Removing the outliers removed roughly 200 observations from our roughly 7000 observations that we had before. After removing outliers, some of the fuel types are removed from our data, which left us with regular gasoline and diesel.



Below is the plot of our residual distribution in a histogram, this graph is roughly normal and skewed slightly to the left. Our qqplot, which tests for normality, showed that our data is fairly normal, given that the points on the plot are close to the trend line middle.



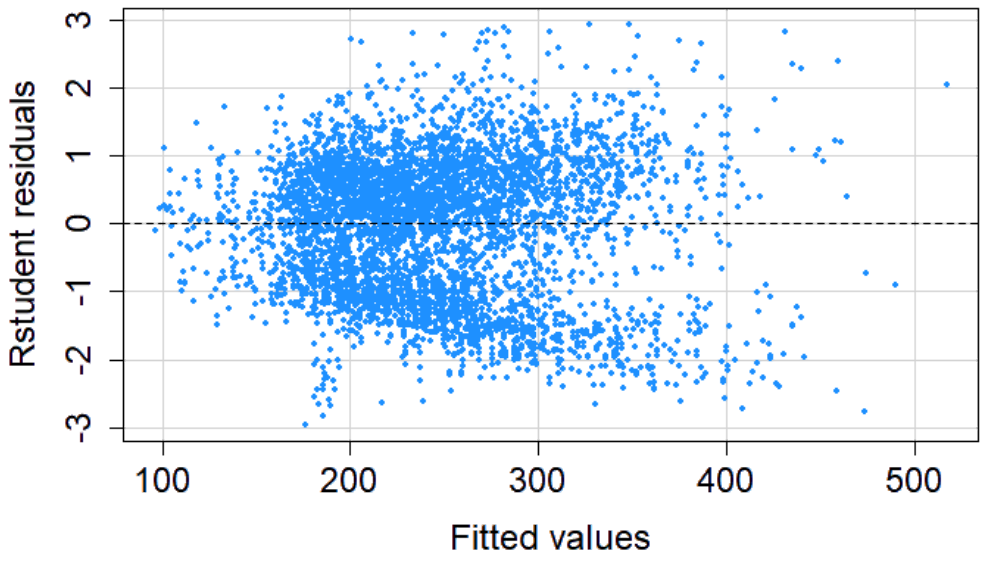
Lastly, our residual plot against fitted values is randomly distributed and forms a roughly contiguous band around the zero line. There is somewhat of a lobster claw

shape to the graph, but since this is real world data, it's not going to be exactly perfect. All in all, given its normality we think it's a satisfactory model.

Variable Selection

	GVIF
Engine.Size.L.	10.516585
Cylinders	9.638679
Fuel.Consumption.City..L.100.km.	1703.791684
Fuel.Consumption.Hwy..L.100.km.	438.338248
Fuel.Consumption.Comb..L.100.km.	3682.298415
Fuel.Type	2.620058
Transmission	13.712245
Vehicle.Class	17.165901

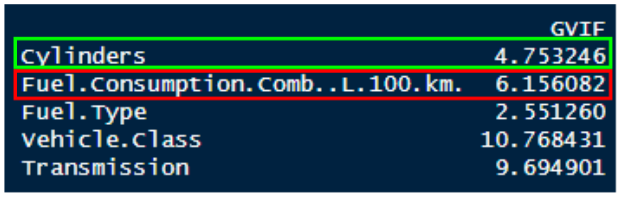
The presence of multicollinearity among the regressor variables in our multilinear regression model for predicting CO2 emissions was assessed using a VIF test. Initially,



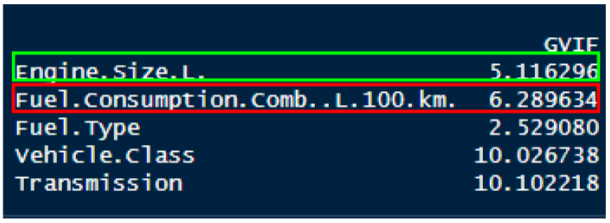
all predictor variables, including various fuel consumption types, cylinder count, vehicle class, transmission, and engine size, were included in the analysis.

The VIF values revealed that all fuel consumption types, cylinder count, vehicle class, transmission, and engine size had VIF values exceeding the recommended threshold of 10, indicating multicollinearity. To identify the specific variables causing multicollinearity, select regressor variables were removed one by one.

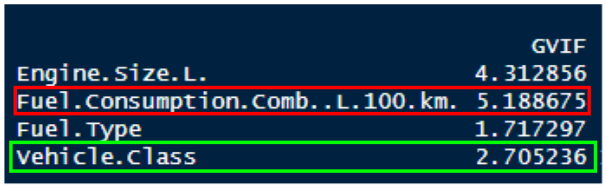
It was found that having multiple fuel consumption types, transmission options, vehicle classes, and considering both engine size and cylinder count simultaneously resulted in elevated VIF values. To address this issue, reduced models were considered with one representative variable from each category.



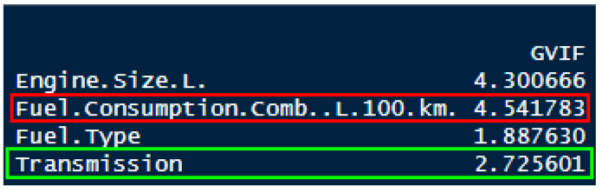
	GVIF
Cylinders	4.753246
Fuel.Consumption.Comb..L.100.km.	6.156082
Fuel.Type	2.551260
Vehicle.Class	10.768431
Transmission	9.694901



	GVIF
Engine.Size.L.	5.116296
Fuel.Consumption.Comb..L.100.km.	6.289634
Fuel.Type	2.529080
Vehicle.Class	10.026738
Transmission	10.102218



	GVIF
Engine.Size.L.	4.312856
Fuel.Consumption.Comb..L.100.km.	5.188675
Fuel.Type	1.717297
Vehicle.Class	2.705236



	GVIF
Engine.Size.L.	4.300666
Fuel.Consumption.Comb..L.100.km.	4.541783
Fuel.Type	1.887630
Transmission	2.725601

The reduced models showed lower VIF values, suggesting reduced multicollinearity. The specific VIF values for the reduced models were: around 5 and 6 for Fuel.Consumption.Comb..L.100.km., 3 for Transmission, 3 for vehicle.class, 4 for Engine.Size, and 5 for L. Cylinders. Based on all the permutations we tested, there were multicollinearity issues between all the fuel consumption types, vehicle class vs. transmission, and cylinders vs engine size. Therefore, to mitigate multicollinearity, we decided to include only one representative variable for each of the 3 multicollinearity issues we had in our models for ANOVA testing.

We also utilized forward selection, backward elimination, and stepwise regression to determine the optimal model for predicting CO2 emissions. Surprisingly, all initial regressor variables were consistently retained by these methods. However, this

contradicted the presence of multicollinearity identified through the VIF test. Therefore, it is essential to prioritize addressing multicollinearity by considering the valid model permutations identified earlier, rather than relying solely on the initial variable selection results.

Model Fitting

The final part in constructing our linear regression model involved ANOVA analysis on all our models to see if all the predictors were statistically significant. Based on the results of our variable selection, we made models for every possible permutation possible and looked at their F and p values. We used the summary() method to look at their corresponding F-statistic and p-values.

Based on all the permutations, we found that only 2 models where all the predictors had a p-value smaller than 0.05 and F-value greater than their F-statistic through ANOVA analysis. It was discovered that using fuel consumption in the city and vehicle class were the best predictors for our models. However, the 2 models differed where one included cylinders and the other included engine size. We opted for the one with the higher R squared value of 0.9898 (The model was the one with Cylinders).

We were also curious to see that since vehicle class was a discrete variable, it would not be necessary to include in our model. To check if it was, we made a reduced model excluding it and compared it with the full model including it. It turned out that it was significant since it produced a high F-value of around 280 using ANOVA analysis. Therefore we decided to keep the vehicle class in our final model.

Looking at our final model we discovered that not every single discrete dummy column that corresponded to the vehicle class discrete variable was significant. The standard pickup Truck had a p-value way higher than 0.05, so we decided to exclude that single dummy variable in our model.

Conclusion

```
Call:
lm(formula = CO2.Emissions.g.km. ~ Cylinders + Fuel.Consumption.City..L.100.km. +
    D + X + COMPACT + SUV...SMALL + MID.SIZE + TWO.SEATER + MINICOMPACT +
    SUBCOMPACT + FULL.SIZE + STATION.WAGON...SMALL + SUV...STANDARD +
    VAN...CARGO + VAN...PASSENGER + MINIVAN + SPECIAL.PURPOSE.VEHICLE +
    STATION.WAGON...MID.SIZE, data = noOutliersDF2)

Residuals:
    Min      1Q  Median      3Q     Max
-22.595  -3.944  -0.305   3.544  44.160

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     41.02966    0.58220  70.474  < 2e-16 ***
Cylinders                        0.38941    0.07856   4.957 7.33e-07 ***
Fuel.Consumption.City..L.100.km. 17.88499    0.05028 355.735  < 2e-16 ***
D                               21.07922    0.53604  39.324  < 2e-16 ***
X                               -1.96377    0.18058 -10.875  < 2e-16 ***
COMPACT                        -16.17567    0.35183 -45.976  < 2e-16 ***
SUV...SMALL                     -8.17506    0.32018 -25.532  < 2e-16 ***
MID.SIZE                       -16.53910    0.33827 -48.893  < 2e-16 ***
TWO.SEATER                     -13.67117    0.41073 -33.285  < 2e-16 ***
MINICOMPACT                    -13.23351    0.45683 -28.968  < 2e-16 ***
SUBCOMPACT                     -15.63740    0.38309 -40.819  < 2e-16 ***
FULL.SIZE                      -16.11333    0.36613 -44.010  < 2e-16 ***
STATION.WAGON...SMALL          -12.41030    0.47481 -26.137  < 2e-16 ***
SUV...STANDARD                  -3.81157    0.34821 -10.946  < 2e-16 ***
VAN...CARGO                      6.55083    1.72675   3.794  0.00015 ***
VAN...PASSENGER                  3.80004    0.92579   4.105 4.10e-05 ***
MINIVAN                        -10.06910    0.75428 -13.349  < 2e-16 ***
SPECIAL.PURPOSE.VEHICLE         -5.50251    0.77638  -7.087 1.50e-12 ***
STATION.WAGON...MID.SIZE       -11.56576    0.86137 -13.427  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.902 on 6921 degrees of freedom
Multiple R-squared:  0.9898,    Adjusted R-squared:  0.9897
F-statistic: 3.715e+04 on 18 and 6921 DF,  p-value: < 2.2e-16
```

<u>The Final Multi-linear Model:</u>

CO2 emissions = 41.03 + 0.39 * Cylinders + 17.88 * Fuel.Consumption.City..L.100.km. + 21.08 * D - 1.96 * X - 16.18 * COMPACT - 8.18 * SUV...SMALL - 16.54 * MID.SIZE - 13.67 * TWO.SEATER - 13.23 * MINICOMPACT - 15.64 * SUBCOMPACT - 16.11 * FULL.SIZE - 12.41 * STATION.WAGON...SMALL - 3.81 * SUV...STANDARD + 6.55 * VAN...CARGO + 3.80 * VAN...PASSENGER - 10.07 * MINIVAN - 5.50 * SPECIAL.PURPOSE.VEHICLE - 11.57 * STATION.WAGON...MID.SIZE

In this model, the intercept term is 41.03. The number of cylinders (Cylinders) , fuel consumption in the city (Fuel.Consumption.City..L.100.km.), and (D) diesel have positive coefficients of 0.39, 17.88, and 21.08 respectively, indicating that an increase in cylinders or city fuel consumption and usage of diesel leads to higher CO2 emissions.

The variables X (regular fuel), COMPACT, SUV...SMALL, MID.SIZE, TWO.SEATER, MINICOMPACT, SUBCOMPACT, FULL.SIZE, STATION.WAGON...SMALL, SUV...STANDARD, VAN...CARGO, VAN...PASSENGER, MINIVAN, SPECIAL.PURPOSE.VEHICLE, and STATION.WAGON...MID.SIZE have negative coefficients, indicating that their presence is associated with lower CO2 emissions.

This final model is derived from considering the effects of various vehicle characteristics on CO2 emissions. The 0.9898 R squared value indicates high accuracy with our specific data. However, it is important to note that this model assumes a linear relationship between the predictors and CO2 emissions and the result might vary from country to country.

We found that to lower emissions it is best to avoid having a high number of cylinders and high fuel consumption, and to avoid having vehicles that use diesel; regular gasoline cars, low fuel consumption and low number of cylinders were best at emitting fewer CO2 gasses. Also vehicle classes with typically smaller sized vehicles, such as (SUV Small, Subcompact, Minicompact) have lower emissions than larger sized vehicle classes (Van Cargo, Van Passenger).

Reflection

Looking back at our project, the CO2 Emission by Vehicles data set that we used was limited to only the select sample of cars that the Canadian government selected. If we were to improve upon our research, we would include a data set that is more inclusive to the distribution of the manufacturers and models of the world, instead of just Canada.

The dataset size is fairly small at about 7,000 observations, increasing the amount of observations should help reduce the amount of bias in the data. Also the amount of observations per model of car was very low, so in order to test for its significance, a large volume of data points per model is needed. To improve this, we would look for a data set that includes a higher volume of observations per model.

Referring back to our exploratory analysis, we had to cut out 371 rows, because it created its own linear distribution on the graph and skewed our data. Including this second distribution into its own regression analysis would be interesting, and may provide different insights into CO2 emissions.

Appendix

References

Chong, Jason. “Back to Basics - Linear Regression in R.” *Medium*, Towards Data Science, 28 May 2021, towardsdatascience.com/back-to-basics-linear-regression-in-r-3ffe4900482b.

Ibarrera, Written. “Why Duplicates Exist and How to Get Rid of Them?” *Data Ladder*, 18 Oct. 2022, dataladder.com/why-duplicates-exist-and-how-to-get-rid-of-them/. Accessed 10 May 2023.

Montgomery, Douglas C., et al. *Introduction to Linear Regression Analysis (5th Ed.)*. Wiley, 2012.

Podder, Debajyoti. “CO2 Emission by Vehicles.” *Kaggle*, 5 Aug. 2020, www.kaggle.com/datasets/debajyotipodder/co2-emission-by-vehicles.

Potters, Charles. “Variance Inflation Factor (VIF).” *Investopedia*, 12 Feb. 2023, www.investopedia.com/terms/v/variance-inflation-factor.asp.

Team Responsibilities

First, we determined how to clean the data as a group; then everyone was delegated a set of predictor variables to analyze the significance of correlation between them and the response variable. Each member then compared the strength of correlation of each predictor variable to the response variable and conducted a linear regression model. Finally, we came together to compare each model, decide which model is the best, and prepare a conclusion, presentation, and report.

Code

```
#1. Cleaning Data


# Read Data
emissionsDF <- read.csv("data/CO2 Emissions_Canada.csv")

# Look for NA's
which(is.na(emissionsDF))

# Add Dummy variables for Fuel Type exclude Z
lvl <- unique(emissionsDF$Fuel.Type)
res <- data.frame(fuel=emissionsDF$Fuel.Type,
            do.call(rbind,lapply(emissionsDF$Fuel.Type, function(x)
table(factor(x, levels=lvl)))),
                                        stringsAsFactors=FALSE)

# Add Dummy variables for Transmission exclude AV10
lvl2 <- unique(emissionsDF$Transmission)
res2 <- data.frame(Trasmission=emissionsDF$Transmission,
            do.call(rbind,lapply(emissionsDF$Transmission, function(x)
table(factor(x, levels=lvl2)))),
                                        stringsAsFactors=FALSE)

# Add Dummy variables for Vehicle Class exclude PICKUP.TRUCK...SMALL
lvl3 <- unique(emissionsDF$Vehicle.Class)
res3 <- data.frame(Class=emissionsDF$Vehicle.Class,
            do.call(rbind,lapply(emissionsDF$Vehicle.Class,
function(x) table(factor(x, levels=lvl3)))),
                                        stringsAsFactors=FALSE)

# Add Columns to DataFrame
emissionsDF <- cbind(emissionsDF, res[2:5])
emissionsDF <- cbind(emissionsDF, res2[2:27])
emissionsDF <- cbind(emissionsDF, res3[2:16])


# Remove MPG column
emissionsDF <- emissionsDF[,-11]
emissionsDF <- emissionsDF[,-2]
```

```
# Summary of data
summary(emissionsDF)



#2. Exploring Data

table(emissionsDF$Vehicle.Class) # Could Work
table(emissionsDF$Make) # Too Many small values
table(emissionsDF$Fuel.Type) #  Could Work
table(emissionsDF$Transmission) # # Could Work




names <- c("Engine", "Cylinders", "Fuel Comp City", "Fuel Comp Hwy",
"Fuel Comp Comb", "C02 Emissions")
# Look at correlations
corMatrix <- cor(emissionsDF[,c(3,4,7:10)])
colnames(corMatrix) <- names
rownames(corMatrix) <- names
corMatrix
corMatrix[6,]






library(ggplot2)
library(reshape2)

# Plot correlations

get_upper_tri <- function(cormat) {
    cormat[lower.tri(cormat)]<- NA
    return(cormat)
}
melted_cormat <- melt(get_upper_tri(corMatrix), na.rm = TRUE)

ggplot(data = melted_cormat, aes(x=Var1, y=Var2, fill=-value)) +
```

```
  geom_tile() +
  theme(axis.text.x = element_text(angle = -45, vjust = 1,
        size = 12, hjust = 1))

#count
ggplot(emissionsDF, aes(CO2.Emissions.g.km.)) +
  geom_bar(aes(fill = "blue"), show.legend = FALSE) +
  theme_minimal() +
  labs(title = "CO2 Emissions Distribution")

sd(emissionsDF$CO2.Emissions.g.km.)

# Plot correlations
emissionsDFPlot <- melt(emissionsDF[,c(3,4,7:10)], id.vars =
"CO2.Emissions.g.km.")

levels(emissionsDFPlot$variable) <- names



ggplot(emissionsDFPlot) +
  geom_jitter(aes(value,CO2.Emissions.g.km., color=variable)) +
  geom_smooth(aes(value,CO2.Emissions.g.km.),color="black",
method=lm, se=FALSE) +
  facet_wrap(~variable, scales="free_x")



firstfit <- lm(CO2.Emissions.g.km. ~ Engine.Size.L. + Cylinders +
            Fuel.Consumption.City..L.100.km. +
Fuel.Consumption.Hwy..L.100.km. +
            Fuel.Consumption.Comb..L.100.km. + Fuel.Type +
Transmission +
            Vehicle.Class,data=emissionsDF)


residualPlot(firstfit, type="rstudent", quadratic=F, col =
"dodgerblue",
pch=16, cex=1.5, cex.axis=1.5, cex.lab=1.5)



summary(lm(CO2.Emissions.g.km. ~ Fuel.Consumption.Comb..L.100.km.,
data = emissionsDF))
```

```
emissionsDF2 <- melt(emissionsDF[,c(9,10)], id.vars =
"CO2.Emissions.g.km.")
ggplot(emissionsDF2) +
  geom_jitter(aes(value,CO2.Emissions.g.km., colour=variable)) +
  geom_smooth(aes(value,CO2.Emissions.g.km.), colour="black",
method=lm, se=FALSE) +
  geom_abline(intercept = 0, slope = 20, color="blue") +
  facet_wrap(~variable, scales="free_x")



emissionsDF<- emissionsDF[!(emissionsDF$CO2.Emissions.g.km. < 20 *
emissionsDF$Fuel.Consumption.Comb..L.100.km.),]


emissionsDFPlot2 <- melt(emissionsDF[,c(3,4,7:10)], id.vars =
"CO2.Emissions.g.km.")

levels(emissionsDFPlot2$variable) <- names

ggplot(emissionsDFPlot2) +
  geom_jitter(aes(value,CO2.Emissions.g.km., colour=variable)) +
  geom_smooth(aes(value,CO2.Emissions.g.km.), colour="black",
method=lm, se=FALSE) +
  facet_wrap(~variable, scales="free_x")



#Look for Outliers


library(MASS)

firstfit <- lm(CO2.Emissions.g.km. ~ Engine.Size.L. + Cylinders +
            Fuel.Consumption.City..L.100.km. +
Fuel.Consumption.Hwy..L.100.km. +
            Fuel.Consumption.Comb..L.100.km. + Fuel.Type +
Transmission +
            Vehicle.Class,data=emissionsDF)


barplot(height = stdres(firstfit),
main = "Standardized Residuals", xlab = "Index",
ylab = "Standardized Resid", ylim=c(-4,4))
#Add cutoff values. Either 2 or 3 can be chosen.
```

```
abline(h=3, col = "Red", lwd=2)
abline(h=-3, col = "Red", lwd=2)


barplot(height = studres(firstfit),
main = "Studentized Residuals", xlab = "Index",
ylab = "Studentized Resid", ylim=c(-4,4))
#Add cutoff values. Either 2 or 3 can be chosen.
abline(h=3, col = "Red", lwd=2)
abline(h=-3, col = "Red", lwd=2)



#Remove Outliers

noOutliersBool <- abs(stdres(firstfit)) < 3

noOutliersDF <- emissionsDF[noOutliersBool,]

fitNoOutliers <- lm(CO2.Emissions.g.km. ~ Engine.Size.L. + Cylinders +
            Fuel.Consumption.City..L.100.km. +
Fuel.Consumption.Hwy..L.100.km. +
            Fuel.Consumption.Comb..L.100.km. + Fuel.Type +
Transmission +
            Vehicle.Class, data = noOutliersDF)



barplot(height = stdres(fitNoOutliers),
main = "Standardized Residuals", xlab = "Index",
ylab = "Standardized Resid", ylim=c(-4,4))
#Add cutoff values. Either 2 or 3 can be chosen.
abline(h=3, col = "Red", lwd=2)
abline(h=-3, col = "Red", lwd=2)


barplot(height = studres(fitNoOutliers),
main = "Studentized Residuals", xlab = "Index",
ylab = "Studentized Resid", ylim=c(-4,4))
#Add cutoff values. Either 2 or 3 can be chosen.
abline(h=3, col = "Red", lwd=2)
abline(h=-3, col = "Red", lwd=2)


#Remove Outliers 2
```

```
noOutliersBool2 <- abs(stdres(fitNoOutliers)) < 3

noOutliersDF2 <- noOutliersDF[noOutliersBool2,]

fitNoOutliers2 <- lm(CO2.Emissions.g.km. ~ Engine.Size.L. + Cylinders
+
            Fuel.Consumption.City..L.100.km. +
Fuel.Consumption.Hwy..L.100.km. +
            Fuel.Consumption.Comb..L.100.km. + Fuel.Type +
Transmission +
            Vehicle.Class, data = noOutliersDF2)


barplot(height = stdres(fitNoOutliers2),
main = "Standardized Residuals", xlab = "Index",
ylab = "Standardized Resid", ylim=c(-4,4))
#Add cutoff values. Either 2 or 3 can be chosen.
abline(h=3, col = "Red", lwd=2)
abline(h=-3, col = "Red", lwd=2)


barplot(height = studres(fitNoOutliers2),
main = "Studentized Residuals", xlab = "Index",
ylab = "Studentized Resid", ylim=c(-4,4))
#Add cutoff values. Either 2 or 3 can be chosen.
abline(h=3, col = "Red", lwd=2)
abline(h=-3, col = "Red", lwd=2)


#Look for Influential Points

library(car)
dfbetasPlots(fitNoOutliers2, intercept=T)
influenceIndexPlot(fitNoOutliers2)
myInf <- influence.measures(fitNoOutliers2)
summary(myInf)


#Distribution of Residuals

par(mfrow=c(1,2))
hist(studres(fitNoOutliers2), breaks=10, freq=F, col="cornflowerblue",
cex.axis=1.5, cex.lab=1.5, cex.main=2)
qqPlot(fitNoOutliers2)
```

```
residualPlot(firstfit, type="rstudent", quadratic=F, col =
"dodgerblue",
pch=16, cex=1.5, cex.axis=1.5, cex.lab=1.5)


residualPlot(fit2NoOutliers, type="rstudent", quadratic=F, col =
"dodgerblue",
pch=16, cex=1.5, cex.axis=1.5, cex.lab=1.5)

residualPlot(fitNoOutliers2, type="rstudent", quadratic=F, col =
"dodgerblue",
pch=16, cex=1.5, cex.axis=1.5, cex.lab=1.5)




table(noOutliersDF2$Vehicle.Class)
table(noOutliersDF2$Fuel.Type)
table(noOutliersDF2$Transmission)


#Multicollinearity Test


library(car)



fitOG <- lm(CO2.Emissions.g.km. ~ Engine.Size.L. + Cylinders +
            Fuel.Consumption.City..L.100.km. +
Fuel.Consumption.Hwy..L.100.km. +
            Fuel.Consumption.Comb..L.100.km. + Fuel.Type +
Transmission +
            Vehicle.Class, data=noOutliersDF2)

vif(fitOG)

#---------------------------- Check for Fuel Consumption------------#
# look at City
fit.city <- lm(Fuel.Consumption.City..L.100.km. ~ Engine.Size.L. +
Cylinders +
                 Fuel.Consumption.Hwy..L.100.km. +
                 Fuel.Consumption.Comb..L.100.km. + Fuel.Type +
Transmission +
                  Vehicle.Class, data=noOutliersDF2)
```

```
vif(fit.city)
summary(fit.city)

# look at hwy
fit.hwy <- lm(Fuel.Consumption.Hwy..L.100.km. ~ Engine.Size.L. +
Cylinders +
                Fuel.Consumption.City..L.100.km. +
                Fuel.Consumption.Comb..L.100.km. + Fuel.Type +
Transmission +
                 Vehicle.Class, data=noOutliersDF2)
vif(fit.hwy)
summary(fit.hwy)

# look at combo
fit.combo <- lm(Fuel.Consumption.Comb..L.100.km. ~ Engine.Size.L. +
Cylinders +
                Fuel.Consumption.City..L.100.km. +
                Fuel.Consumption.Hwy..L.100.km. + Fuel.Type +
Transmission +
                 Vehicle.Class, data=noOutliersDF2)
vif(fit.combo)
summary(fit.combo)
#-----------------------------------------------------------------#


fit <- lm(CO2.Emissions.g.km. ~ Engine.Size.L. + Cylinders +
            Fuel.Consumption.Comb..L.100.km. + Fuel.Type +
Transmission +
            Vehicle.Class, data=noOutliersDF2)

vif(fit)


#--Shows Engine Size and Cylinders Count are
correlated-----------------#
fit.engine <- lm(Engine.Size.L. ~  Cylinders +
            Fuel.Consumption.Comb..L.100.km. + Fuel.Type +
Transmission +
            Vehicle.Class, data=noOutliersDF2)

vif(fit.engine)
summary(fit.engine)
#-----------------------------------------------------------------
--#
```

```
fit <- lm(CO2.Emissions.g.km. ~ Engine.Size.L. +
            Fuel.Consumption.Comb..L.100.km. + Fuel.Type +
Transmission +
            Vehicle.Class, data=noOutliersDF2)

vif(fit) # Transmission and Vehicle Class are correlated


# With no Vehicle Class
fit <- lm(CO2.Emissions.g.km. ~ Engine.Size.L. +
            Fuel.Consumption.Comb..L.100.km. + Fuel.Type +
Transmission,
             data=noOutliersDF2)

vif(fit)

# With no Transmission
fit <- lm(CO2.Emissions.g.km. ~ Engine.Size.L. +
            Fuel.Consumption.Comb..L.100.km. + Fuel.Type +
Vehicle.Class,
             data=noOutliersDF2)

vif(fit)



#Multilinear Regression Models



# No significant results
fit0 = lm(CO2.Emissions.g.km.~1,data = noOutliersDF2)
stepAIC(fit0,direction="forward", scope=list(lower=fit0, upper=fitOG))
stepAIC(fitOG,direction="backward")
stepAIC(fit0,direction="both", scope=list(lower=fit0, upper=fitOG))







# -- Engine Size and Transmission ------------------#
f1.0  <- lm(CO2.Emissions.g.km. ~ Engine.Size.L. +
            Fuel.Consumption.City..L.100.km. + D + X +

A10+A4+A5+A6+A7+A8+A9+AM5+AM6+AM7+AM8+AM9+AS10+AS4+AS5+AS6+AS7+AS8+
```

```
            AS9+AV+AV6+AV7+AV8+M5+M6+M7,
             data=noOutliersDF2)

f2.0 <- lm(CO2.Emissions.g.km. ~ Engine.Size.L. +
            Fuel.Consumption.Comb..L.100.km. + D + X +

A10+A4+A5+A6+A7+A8+A9+AM5+AM6+AM7+AM8+AM9+AS10+AS4+AS5+AS6+AS7+AS8+
            AS9+AV+AV6+AV7+AV8+M5+M6+M7,
             data=noOutliersDF2)

f3.0 <- lm(CO2.Emissions.g.km. ~ Engine.Size.L. +
            Fuel.Consumption.Hwy..L.100.km. + D + X +

A10+A4+A5+A6+A7+A8+A9+AM5+AM6+AM7+AM8+AM9+AS10+AS4+AS5+AS6+AS7+AS8+
            AS9+AV+AV6+AV7+AV8+M5+M6+M7,
             data=noOutliersDF2)



#----Engine Size and Vehicle Class------------------#
f1.1 <- lm(CO2.Emissions.g.km. ~ Engine.Size.L. +
            Fuel.Consumption.City..L.100.km. + D + X +
            COMPACT + SUV...SMALL + MID.SIZE + TWO.SEATER +
MINICOMPACT +
            SUBCOMPACT + FULL.SIZE + STATION.WAGON...SMALL +
SUV...STANDARD +
            VAN...CARGO + VAN...PASSENGER + PICKUP.TRUCK...STANDARD +
MINIVAN +
            SPECIAL.PURPOSE.VEHICLE + STATION.WAGON...MID.SIZE,
             data=noOutliersDF2)

f2.1 <- lm(CO2.Emissions.g.km. ~ Engine.Size.L. +
            Fuel.Consumption.Comb..L.100.km. + D + X +
            COMPACT + SUV...SMALL + MID.SIZE + TWO.SEATER +
MINICOMPACT +
            SUBCOMPACT + FULL.SIZE + STATION.WAGON...SMALL +
SUV...STANDARD +
            VAN...CARGO + VAN...PASSENGER + PICKUP.TRUCK...STANDARD +
MINIVAN +
            SPECIAL.PURPOSE.VEHICLE + STATION.WAGON...MID.SIZE,
             data=noOutliersDF2)

f3.1 <- lm(CO2.Emissions.g.km. ~ Engine.Size.L. +
            Fuel.Consumption.Hwy..L.100.km. + D + X +
            COMPACT + SUV...SMALL + MID.SIZE + TWO.SEATER +
MINICOMPACT +
```

```
            SUBCOMPACT + FULL.SIZE + STATION.WAGON...SMALL +
SUV...STANDARD +
            VAN...CARGO + VAN...PASSENGER + PICKUP.TRUCK...STANDARD +
MINIVAN +
            SPECIAL.PURPOSE.VEHICLE + STATION.WAGON...MID.SIZE,
             data=noOutliersDF2)

# -- Cylinder Count and Transmission ------------------#
f1.2 <- lm(CO2.Emissions.g.km. ~ Cylinders +
            Fuel.Consumption.City..L.100.km. + D + X +

A10+A4+A5+A6+A7+A8+A9+AM5+AM6+AM7+AM8+AM9+AS10+AS4+AS5+AS6+AS7+AS8+
            AS9+AV+AV6+AV7+AV8+M5+M6+M7,
             data=noOutliersDF2)

f2.2 <- lm(CO2.Emissions.g.km. ~ Cylinders +
            Fuel.Consumption.Comb..L.100.km. + D + X +

A10+A4+A5+A6+A7+A8+A9+AM5+AM6+AM7+AM8+AM9+AS10+AS4+AS5+AS6+AS7+AS8+
            AS9+AV+AV6+AV7+AV8+M5+M6+M7,
             data=noOutliersDF2)

f3.2 <- lm(CO2.Emissions.g.km. ~ Cylinders +
            Fuel.Consumption.Hwy..L.100.km. + D + X +

A10+A4+A5+A6+A7+A8+A9+AM5+AM6+AM7+AM8+AM9+AS10+AS4+AS5+AS6+AS7+AS8+
            AS9+AV+AV6+AV7+AV8+M5+M6+M7,
             data=noOutliersDF2)

#---- Cylinder Count and Vehicle Class------------------#
f1.3 <- lm(CO2.Emissions.g.km. ~ Cylinders +
            Fuel.Consumption.City..L.100.km. + D + X +
            COMPACT + SUV...SMALL + MID.SIZE + TWO.SEATER +
MINICOMPACT +
            SUBCOMPACT + FULL.SIZE + STATION.WAGON...SMALL +
SUV...STANDARD +
            VAN...CARGO + VAN...PASSENGER + PICKUP.TRUCK...STANDARD +
MINIVAN +
            SPECIAL.PURPOSE.VEHICLE + STATION.WAGON...MID.SIZE,
             data=noOutliersDF2)

f2.3 <- lm(CO2.Emissions.g.km. ~ Cylinders +
            Fuel.Consumption.Comb..L.100.km. + D + X +
            COMPACT + SUV...SMALL + MID.SIZE + TWO.SEATER +
MINICOMPACT +
```

```
            SUBCOMPACT + FULL.SIZE + STATION.WAGON...SMALL +
SUV...STANDARD +
            VAN...CARGO + VAN...PASSENGER + PICKUP.TRUCK...STANDARD +
MINIVAN +
            SPECIAL.PURPOSE.VEHICLE + STATION.WAGON...MID.SIZE,
             data=noOutliersDF2)

f3.3 <- lm(CO2.Emissions.g.km. ~ Cylinders +
            Fuel.Consumption.Hwy..L.100.km. + D + X +
            COMPACT + SUV...SMALL + MID.SIZE + TWO.SEATER +
MINICOMPACT +
            SUBCOMPACT + FULL.SIZE + STATION.WAGON...SMALL +
SUV...STANDARD +
            VAN...CARGO + VAN...PASSENGER + PICKUP.TRUCK...STANDARD +
MINIVAN +
            SPECIAL.PURPOSE.VEHICLE + STATION.WAGON...MID.SIZE,
             data=noOutliersDF2)


#ANOVA


# Both 3.84
qf(p=.05, df1=1, df2=6909, lower.tail=FALSE)
qf(p=.05, df1=1, df2=6920, lower.tail=FALSE)


anova(f1.0)
anova(f2.0)
anova(f3.0)
anova(f1.1)
anova(f2.1)
anova(f3.1)
anova(f1.2)
anova(f2.2)
anova(f3.2)
anova(f1.3)
anova(f2.3)
anova(f3.3)

summary(f1.0)
summary(f2.0)
summary(f3.0)
summary(f1.1)
summary(f2.1)
summary(f3.1)
```

```
summary(f1.2)
summary(f2.2)
summary(f3.2)
summary(f1.3)
summary(f2.3)
summary(f3.3)

# In all expect two below, there were too low F-value (below 3.84)
# or high p-values (above 0.05)

# is 1.96
abs(qt(0.05/2, 6920, lower.tail=TRUE))


# Two best Ones
summary(f1.1) # City, Engine, Class
#  Engine has low t-value and PICKUP.TRUCK...STANDARD

summary(f1.3) # City, Cylinders, Class
#  PICKUP.TRUCK...STANDARD


f.reduced <- lm(CO2.Emissions.g.km. ~ Cylinders +
            Fuel.Consumption.City..L.100.km. + D + X,
             data=noOutliersDF2)

f.reduced2 <- lm(CO2.Emissions.g.km. ~ Engine.Size.L. +
            Fuel.Consumption.City..L.100.km. + D + X,
             data=noOutliersDF2)

# 1.67
qf(p=.05, df1=15, df2=6920, lower.tail=FALSE)

# Both are above
anova(f.reduced, f1.3)
anova(f.reduced2, f1.1)


anova(f.reduced)
summary(f.reduced)

f1.3.new <- lm(CO2.Emissions.g.km. ~ Cylinders +
            Fuel.Consumption.City..L.100.km. + D + X +
            COMPACT + SUV...SMALL + MID.SIZE + TWO.SEATER +
MINICOMPACT +
```

```
            SUBCOMPACT + FULL.SIZE + STATION.WAGON...SMALL +
SUV...STANDARD +
            VAN...CARGO + VAN...PASSENGER  + MINIVAN +
            SPECIAL.PURPOSE.VEHICLE + STATION.WAGON...MID.SIZE,
             data=noOutliersDF2)

anova(f1.3.new)
summary(f1.3.new)



#Conclusion

# f1.3 is best (exclude PICKUP.TRUCK...STANDARD tho)
avPlots(f1.3.new)
```