Final Project-Lin

Bin Lin

2016-12-15

1. Introduction: What is your research question? Why do you care? Why should others care?

The topic I want to study is to see if cars made by certain manufacturers tend to have higher carbon dioxide emission. The Volkwagen emissions scandal which erupted on September 2015 drives me to conduct this study. As most people know, the Volkwagen intentionally cheated on the laboratory emissions testing to have its vehicles' nitrogen oxide and carbon dioxide output meet US standards. Approximately 11 million cars worldwide are implicated. About 500K of them are sold in United States. I do care about the vehicles' emission issue, because I think the economy growth will not be sustainable unless we protect our enviroment in the meantime. The global warming is threatening a lot of costal cities including New York City as the result of rising sea level.Extreme weather will eventually hurt human beings, in the form of hurricane, thunderstorms and many other types of El Nino weather. Not only I should care about this issue, everyone living in this planet are responsible for taking care of it. No body can get away with it if the environment is devastated.

2. Data: Write about the data from your proposal in text form. Address the following points:

Data collection: Describe how the data were collected.

Data scientist Hadley Wickham uploads this data set, it can be downloaded from the following link: https://github.com/hadley/fueleconomy. After we have the dataset saved in the local machine, we can just use read.table method to load it into RStudio.

Cases: What are the cases? (Remember: case = units of observation or units of experiment)

The dataset I collected contains fuel economy data from the EPA, 1985-2015. There are total 34632 observations. Each observation represents a model of an vehicle with unique EPA identifier. There are total 74 variables. However, only some variables with complete data such as: id, make, year, class, trans, drive, cyl, displ, fuel, hwy, cty, co2TailpipeGpm, and so on.

Variables: What are the two variables you will be studying? State the type of each variable.

Two variables will get analyze heavily, which are the manufacturers and tail pipe carbon dioxide emission per 100 miles. the manufacturers are categorical variables. The tail carbon dioxide emission per 100 miles is a continuous numerical variable

Type of study: What is the type of study, observational or an experiment? Explain how you've arrived at your conclusion using information on the sampling and/or experimental design.

Observational Study, since we are not actually doing the experiment. There is no control or experimental group. All I am doing is analysing an existing data.

Scope of inference - generalizability: Identify the population of interest, and whether the findings from this analysis can be generalized to that population, or, if not, a subsection of that population. Explain why or why not. Also discuss any potential sources of bias that might prevent generalizability.

The population of interest is all vehicles in the United States and all its manufacturers. Due to the facts this dataset is from EPA (Environmental Protection Agency), it is generalizable to the population studied, because all the vehicles have to meet the EPA's emission standard or the stricter one before it can be sold in the United States.

One potential source of bias is the selection bias. According to Hadley Wickham,this datasets include models that have at least 10 years worth of data. Many new models may not be included in this datasets. However, they take up big market shares in the current auto market. So the selection bias might prevent generalizability.

Scope of inference - causality: Can these data be used to establish causal links between the variables of interest? Explain why or why not.

These data can not be used to establish causal links between two variables. We will have to design an experiment with both explanatory variable and response variable to establish causal connection. This study is only observational, we did not interfere with how the data arise. Only correlation can be established in this case.

3. Exploratory data analysis: Perform relevant descriptive statistics, including summary statistics and visualization of the data. Also address what the exploratory data analysis suggests about your research question.

The following code just to get the dataset clean and ready for the analysis

```
#Load any neccessary packages
library(IS606)
```

```
##
## Welcome to CUNY IS606 Statistics and Probability for Data Analytics
## This package is designed to support this course. The text book used
## is OpenIntro Statistics, 3rd Edition. You can read this by typing
## vignette('os3') or visit www.OpenIntro.org.
##
## The getLabs() function will return a list of the labs available.
##
## The demo(package='IS606') will list the demos that are available.
```

```
##
## Attaching package: 'IS606'
```

```
## The following object is masked from 'package:utils':
##
##     demo
```

```
library(psych)
library(vcd)
```

```
## Warning: package 'vcd' was built under R version 3.3.2
```

```
## Loading required package: grid
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.2
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
#Then I load the dataset into RStudio. Explore what kind of data are in this file. As we can see t
here are  total 34631 observations. In addition, I take subset of the data that are meaningful and
 related to the analysis.
raw_data <- read.table("vehicles.csv", sep = ",", stringsAsFactors = FALSE, header = TRUE)
nrow(raw_data)
```

```
## [1] 34631
```

```
str(raw_data)
```

```
## 'data.frame':    34631 obs. of  74 variables:
##  $ barrels08      : num  15.7 30 12.2 30 17.3 ...
##  $ barrelsA08     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ charge120      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ charge240      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ city08         : int  19 9 23 10 17 21 22 23 23 23 ...
##  $ city08U        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ cityA08        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ cityA08U       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ cityCD         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ cityE          : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ cityUF         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ co2            : int  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
##  $ co2A           : int  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
##  $ co2TailpipeAGpm: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ co2TailpipeGpm : num  423 808 329 808 468 ...
##  $ comb08         : int  21 11 27 11 19 22 25 24 26 25 ...
##  $ comb08U        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ combA08        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ combA08U       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ combE          : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ combinedCD     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ combinedUF     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ cylinders      : int  4 12 4 8 4 4 4 4 4 4 ...
##  $ displ          : num  2 4.9 2.2 5.2 2.2 1.8 1.8 1.6 1.6 1.8 ...
##  $ drive          : chr  "Rear-Wheel Drive" "Rear-Wheel Drive" "Front-Wheel Drive" "Rear-Wheel
 Drive" ...
##  $ engId          : int  9011 22020 2100 2850 66031 66020 66020 57005 57005 57006 ...
```

```
## $ eng_dscr        : chr  "(FFS)" "(GUZZLER)" "(FFS)" "" ...
## $ feScore         : int  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
## $ fuelCost08      : int  2600 5000 2050 5000 3150 2500 2200 2300 2100 2200 ...
## $ fuelCostA08     : int  0 0 0 0 0 0 0 0 0 0 ...
## $ fuelType        : chr  "Regular" "Regular" "Regular" "Regular" ...
## $ fuelType1       : chr  "Regular Gasoline" "Regular Gasoline" "Regular Gasoline" "Regular Gaso
line" ...
## $ ghgScore        : int  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
## $ ghgScoreA       : int  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
## $ highway08       : int  25 14 33 12 23 24 29 26 31 30 ...
## $ highway08U      : num  0 0 0 0 0 0 0 0 0 0 ...
## $ highwayA08      : int  0 0 0 0 0 0 0 0 0 0 ...
## $ highwayA08U     : num  0 0 0 0 0 0 0 0 0 0 ...
## $ highwayCD       : num  0 0 0 0 0 0 0 0 0 0 ...
## $ highwayE        : num  0 0 0 0 0 0 0 0 0 0 ...
## $ highwayUF       : num  0 0 0 0 0 0 0 0 0 0 ...
## $ hlv             : int  0 0 19 0 0 0 0 0 0 0 ...
## $ hpv             : int  0 0 77 0 0 0 0 0 0 0 ...
## $ id              : int  1 10 100 1000 10000 10001 10002 10003 10004 10005 ...
## $ lv2             : int  0 0 0 0 0 0 0 0 0 0 ...
## $ lv4             : int  0 0 0 0 14 15 15 13 13 13 ...
## $ make            : chr  "Alfa Romeo" "Ferrari" "Dodge" "Dodge" ...
## $ model           : chr  "Spider Veloce 2000" "Testarossa" "Charger" "B150/B250 Wagon 2WD" ...
## $ mpgData         : chr  "Y" "N" "Y" "N" ...
## $ phevBlended     : chr  "false" "false" "false" "false" ...
## $ pv2             : int  0 0 0 0 0 0 0 0 0 0 ...
## $ pv4             : int  0 0 0 0 90 88 88 89 89 89 ...
## $ range           : int  0 0 0 0 0 0 0 0 0 0 ...
## $ rangeCity       : num  0 0 0 0 0 0 0 0 0 0 ...
## $ rangeCityA      : num  0 0 0 0 0 0 0 0 0 0 ...
## $ rangeHwy        : num  0 0 0 0 0 0 0 0 0 0 ...
## $ rangeHwyA       : num  0 0 0 0 0 0 0 0 0 0 ...
## $ trany           : chr  "Manual 5-spd" "Manual 5-spd" "Manual 5-spd" "Automatic 3-spd" ...
## $ UCity           : num  23.3 11 29 12.2 21 ...
## $ UCityA          : num  0 0 0 0 0 0 0 0 0 0 ...
## $ UHighway        : num  35 19 47 16.7 32 ...
## $ UHighwayA       : num  0 0 0 0 0 0 0 0 0 0 ...
## $ VClass          : chr  "Two Seaters" "Two Seaters" "Subcompact Cars" "Vans" ...
## $ year            : int  1985 1985 1985 1985 1993 1993 1993 1993 1993 1993 ...
## $ youSaveSpend    : int  -1000 -13000 1750 -13000 -3750 -500 1000 500 1500 1000 ...
## $ guzzler         : chr  "" "T" "" "" ...
## $ trans_dscr      : chr  "" "" "SIL" "" ...
## $ tCharger        : logi  NA NA NA NA TRUE NA ...
## $ sCharger        : chr  "" "" "" "" ...
## $ atvType         : chr  "" "" "" "" ...
## $ fuelType2       : chr  "" "" "" "" ...
## $ rangeA          : chr  "" "" "" "" ...
## $ evMotor         : chr  "" "" "" "" ...
## $ mfrCode         : chr  "" "" "" "" ...
```

```
vehicles <- raw_data[,c("barrels08", "co2TailpipeGpm", "cylinders", "displ", "drive", "engId", "fu
elCost08", "fuelType", "fuelType1", "id", "make", "model", "trany", "UCity", "VClass", "year")]
head(vehicles)
```

```
##   barrels08 co2TailpipeGpm cylinders displ                      drive
## 1  15.68944       423.1905         4   2.0          Rear-Wheel Drive
## 2  29.95056       807.9091        12   4.9          Rear-Wheel Drive
## 3  12.19557       329.1481         4   2.2         Front-Wheel Drive
## 4  29.95056       807.9091         8   5.2          Rear-Wheel Drive
## 5  17.33749       467.7368         4   2.2 4-Wheel or All-Wheel Drive
## 6  14.96429       403.9545         4   1.8         Front-Wheel Drive
##   engId fuelCost08 fuelType        fuelType1    id       make
## 1  9011       2600  Regular Regular Gasoline     1 Alfa Romeo
## 2 22020       5000  Regular Regular Gasoline    10    Ferrari
## 3  2100       2050  Regular Regular Gasoline   100      Dodge
## 4  2850       5000  Regular Regular Gasoline  1000      Dodge
## 5 66031       3150  Premium Premium Gasoline 10000     Subaru
## 6 66020       2500  Regular Regular Gasoline 10001     Subaru
##                 model           trany   UCity          VClass year
## 1  Spider Veloce 2000    Manual 5-spd 23.3333     Two Seaters 1985
## 2          Testarossa    Manual 5-spd 11.0000     Two Seaters 1985
## 3             Charger    Manual 5-spd 29.0000 Subcompact Cars 1985
## 4 B150/B250 Wagon 2WD Automatic 3-spd 12.2222            Vans 1985
## 5    Legacy AWD Turbo    Manual 5-spd 21.0000    Compact Cars 1993
## 6              Loyale Automatic 3-spd 27.0000    Compact Cars 1993
```

```
#There are a couple of manufacturers. Some of them own very few models of vehicles. For the sake o
f explanation I only focuse on those manufacturers that own more than 1500 models. Five of giant c
ar producers end up on that list.
vehicles <- vehicles%>%
  group_by(make)%>%
  mutate(count = n())%>%
  filter(count >= 1500)
head(vehicles)
```

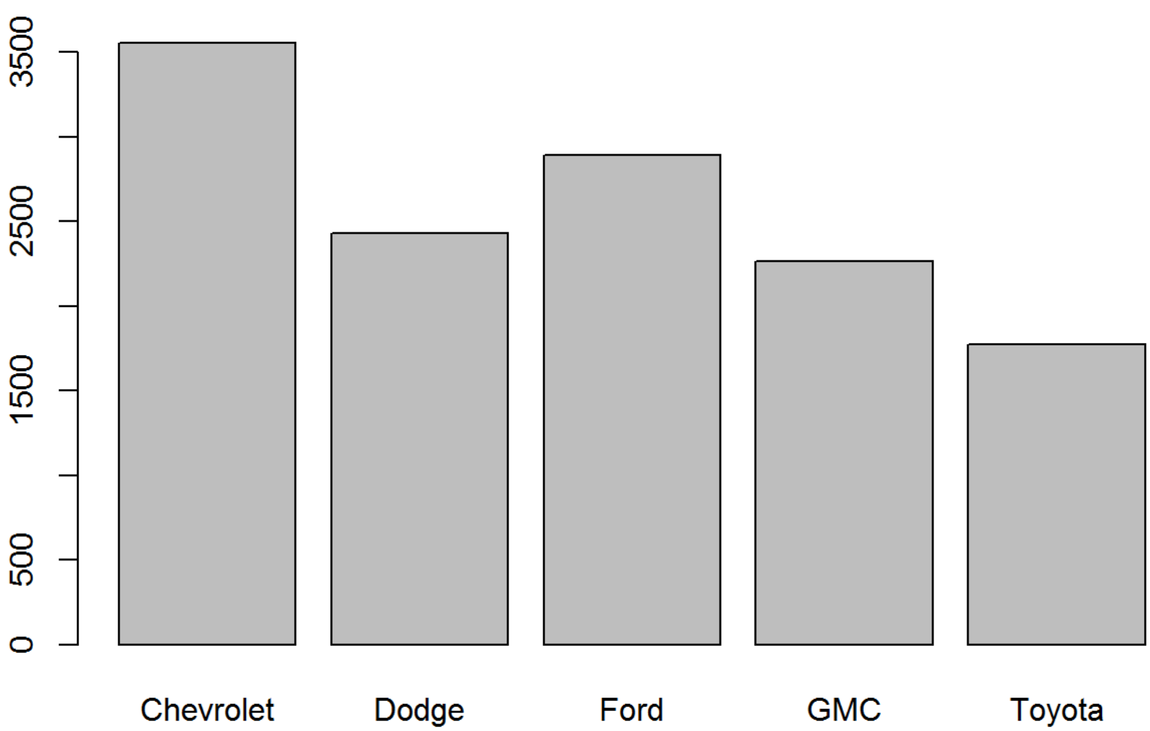
```
## Source: local data frame [6 x 17]
## Groups: make [2]
##
##   barrels08 co2TailpipeGpm cylinders displ             drive engId
##       <dbl>          <dbl>     <int> <dbl>             <chr> <int>
## 1  12.19557       329.1481         4   2.2 Front-Wheel Drive  2100
## 2  29.95056       807.9091         8   5.2  Rear-Wheel Drive  2850
## 3  13.73375       370.2917         4   1.6 Front-Wheel Drive 57005
## 4  12.65702       341.8077         4   1.6 Front-Wheel Drive 57005
## 5  13.18440       355.4800         4   1.8 Front-Wheel Drive 57006
## 6  12.65702       341.8077         4   1.8 Front-Wheel Drive 57006
## # ... with 11 more variables: fuelCost08 <int>, fuelType <chr>,
## #   fuelType1 <chr>, id <int>, make <chr>, model <chr>, trany <chr>,
## #   UCity <dbl>, VClass <chr>, year <int>, count <int>
```

Then I just create a table and barplot just to get an idea about what kind of data we are dealing with.

```
table(vehicles$make, useNA = "ifany")
```

```
##
## Chevrolet     Dodge      Ford       GMC    Toyota
##      3555      2431      2892      2267      1775
```

```
barplot(table(vehicles$make, useNA = "ifany"))
```

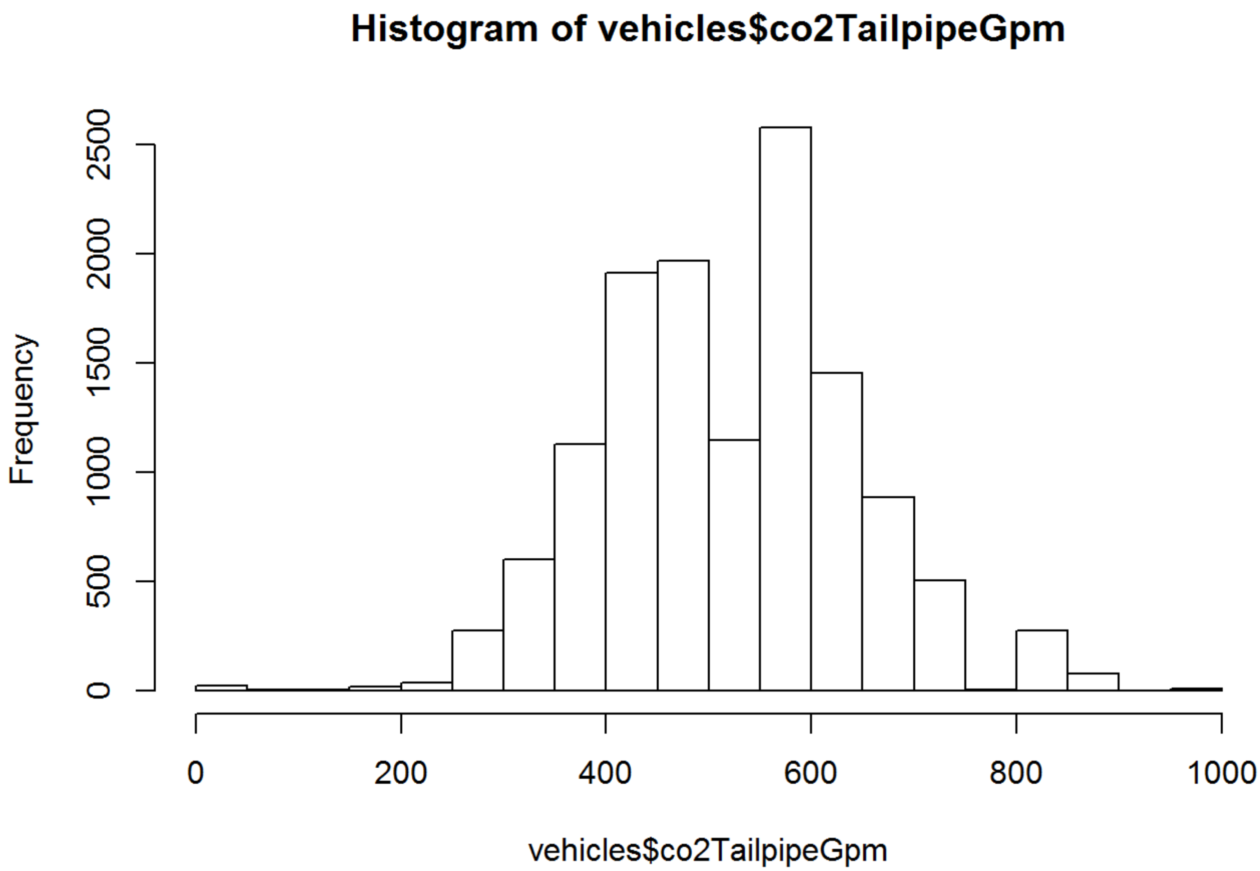


From the summary statistics, we found out that the tail pipe carbon dioxide emission per 100 miles ranges from 0 to 987.4 gallon. The histogram of this numerical data show a unimodel, symmetric, and bell-shaped curve. It correspond to normal distribution. To further prove this findings, I created a normal probability plot, so called qq plot, I found that all the data points kind line up and stay very closed to the line, with very few points that are deviated from the line. Therefore, we can conclude the carbon dioxide emission of vehicles are almost normally distributed.

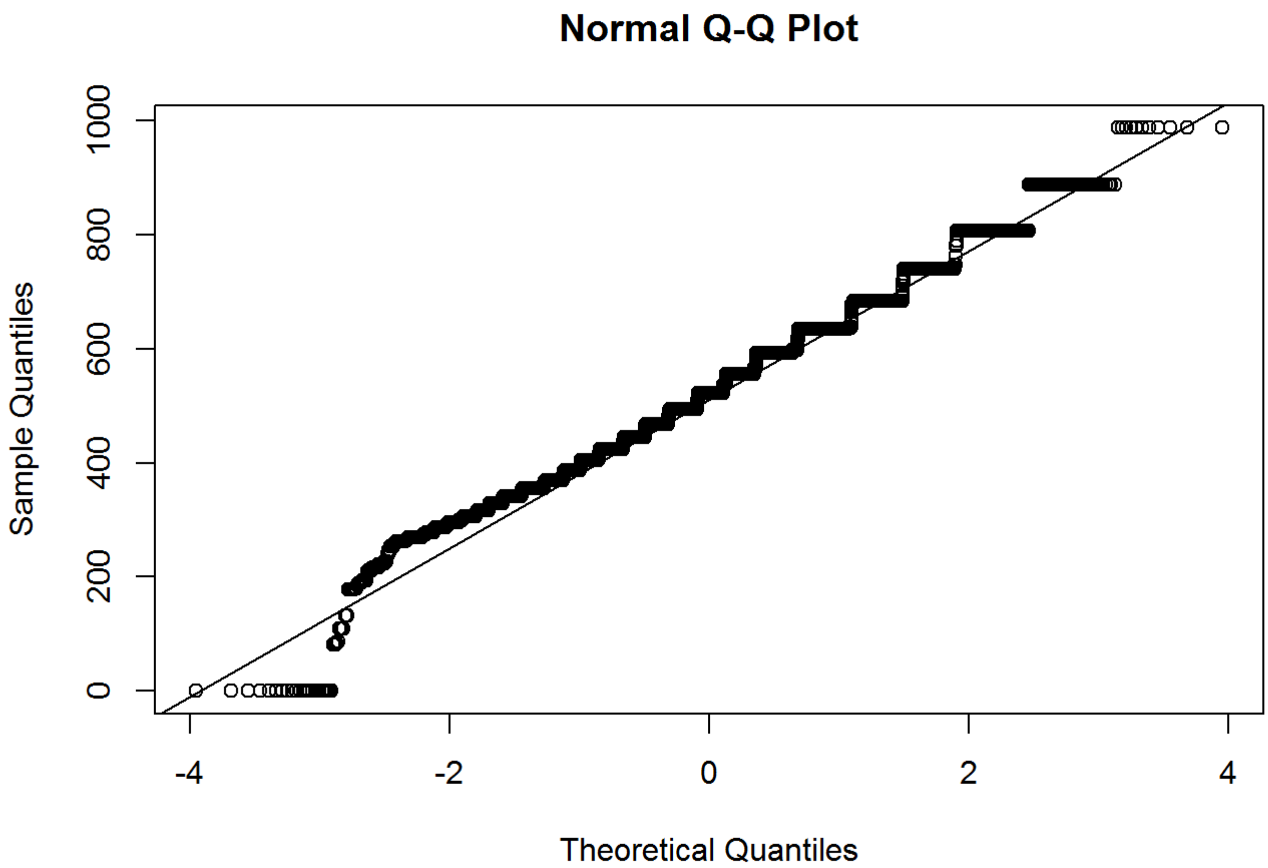
```
summary(vehicles$co2TailpipeGpm)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     0.0   423.2   522.8   523.8   598.8   987.4
```

```
hist(vehicles$co2TailpipeGpm)
```




```
qqnorm(vehicles$co2TailpipeGpm)
qqline(vehicles$co2TailpipeGpm)
```



4. Inference: If your data fails some conditions and you can't use a theoretical method, then you should use simulation. If you can use both methods, then you should use both methods. It is your responsibility to figure out the appropriate methodology. Check conditions

Theoretical inference (if possible) - hypothesis test and confidence interval Simulation based inference - hypothesis test and confidence interval Brief description of methodology that reflects your conceptual understanding

Theoretical Inference:

Since I am comparing means of multiple groups, I can not use either Z or T statistics. I need to use ANOVA test and F statistic.

Conditions: 1. The observations should be independent within and between groups

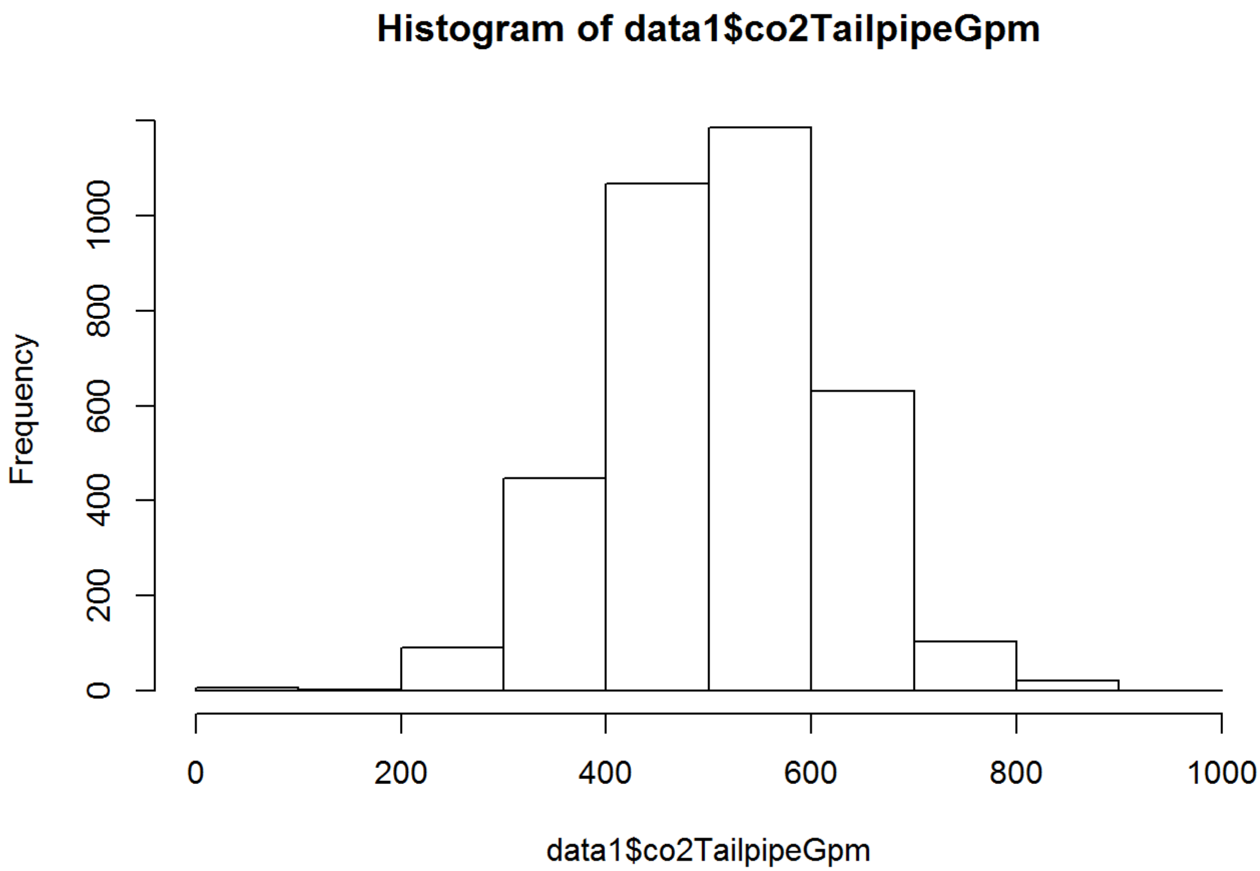
First of all, this data include majority of vehicles that are sold in the United States, therefore, it is definitely more than 10% of the total population. In addition, each group has very different sample size. That is why we have to think about within and between group independence individually.

Between groups: Since automobile industry is also a business. Therefore, auto companies probably not going to share their technologies to reduce the carbon dioxide emission. We will assume they are independent between groups. Within groups: Since each model is different from each other. It is impossible that one model will affect the other in terms of increasing or decreasing other models emission profile.

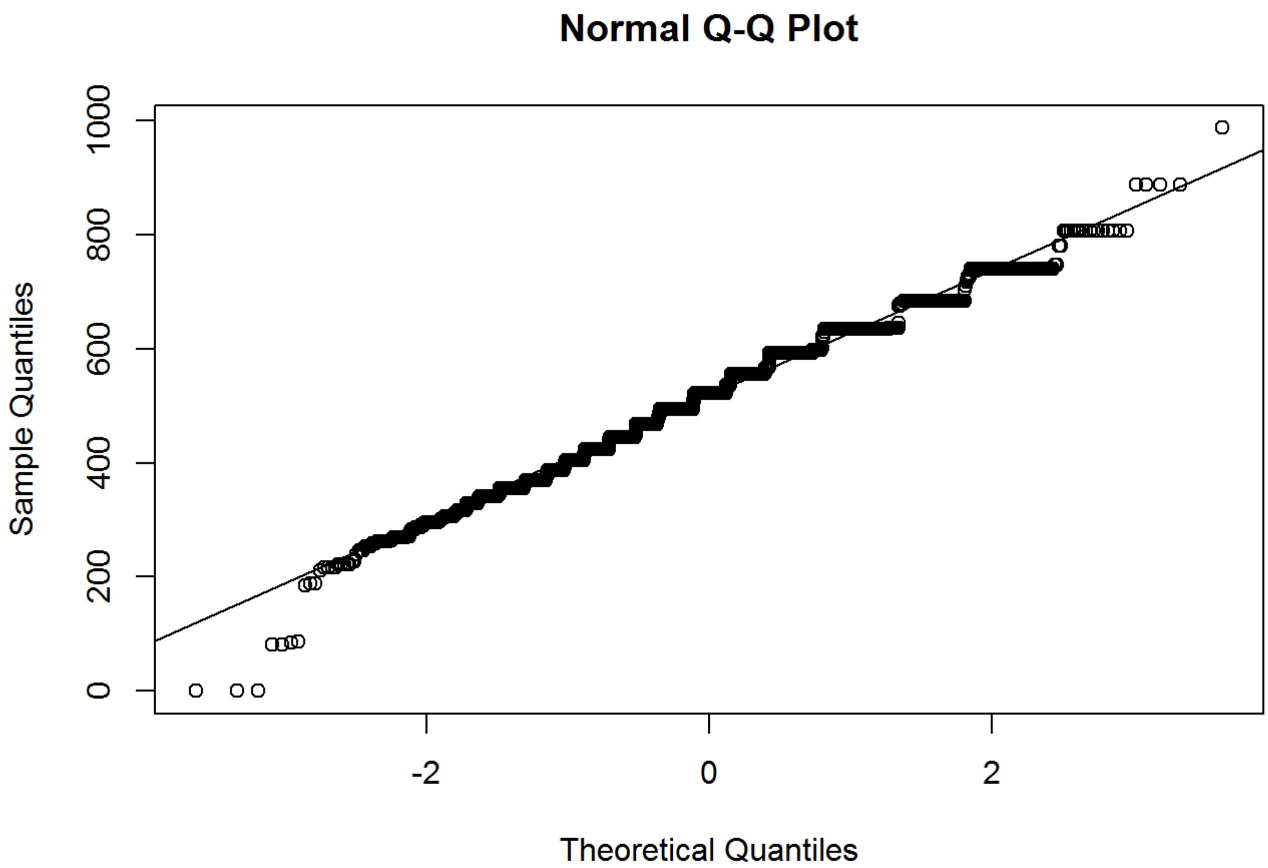
2. The observations within each group should be nearly normal. Especially important when the sample sizes are small. How do we check for normality?

The following code shows both histogram and qq plot of carbon dioxide emission for each manufacturers follow normal distribution.

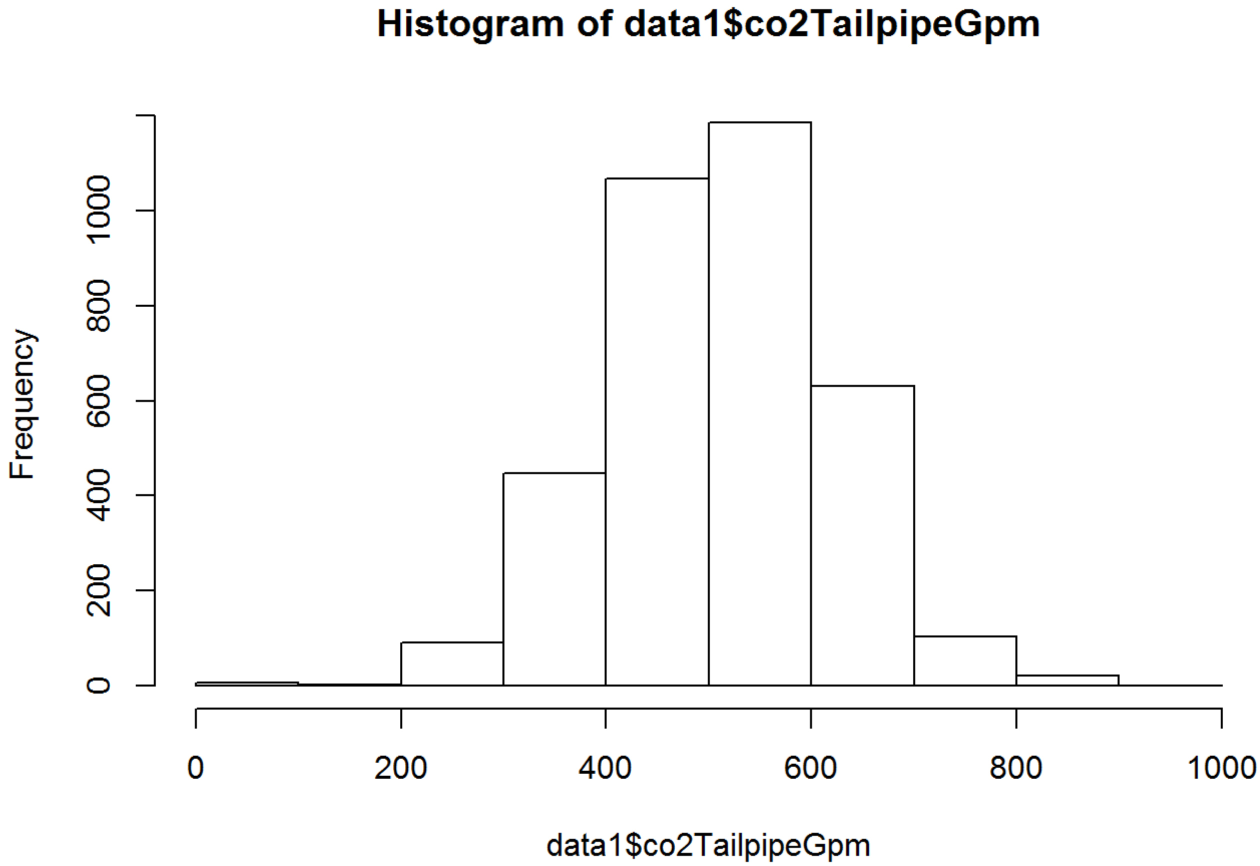
```
data1 <- subset(vehicles, vehicles$make == "Chevrolet")
hist(data1$co2TailpipeGpm)
```



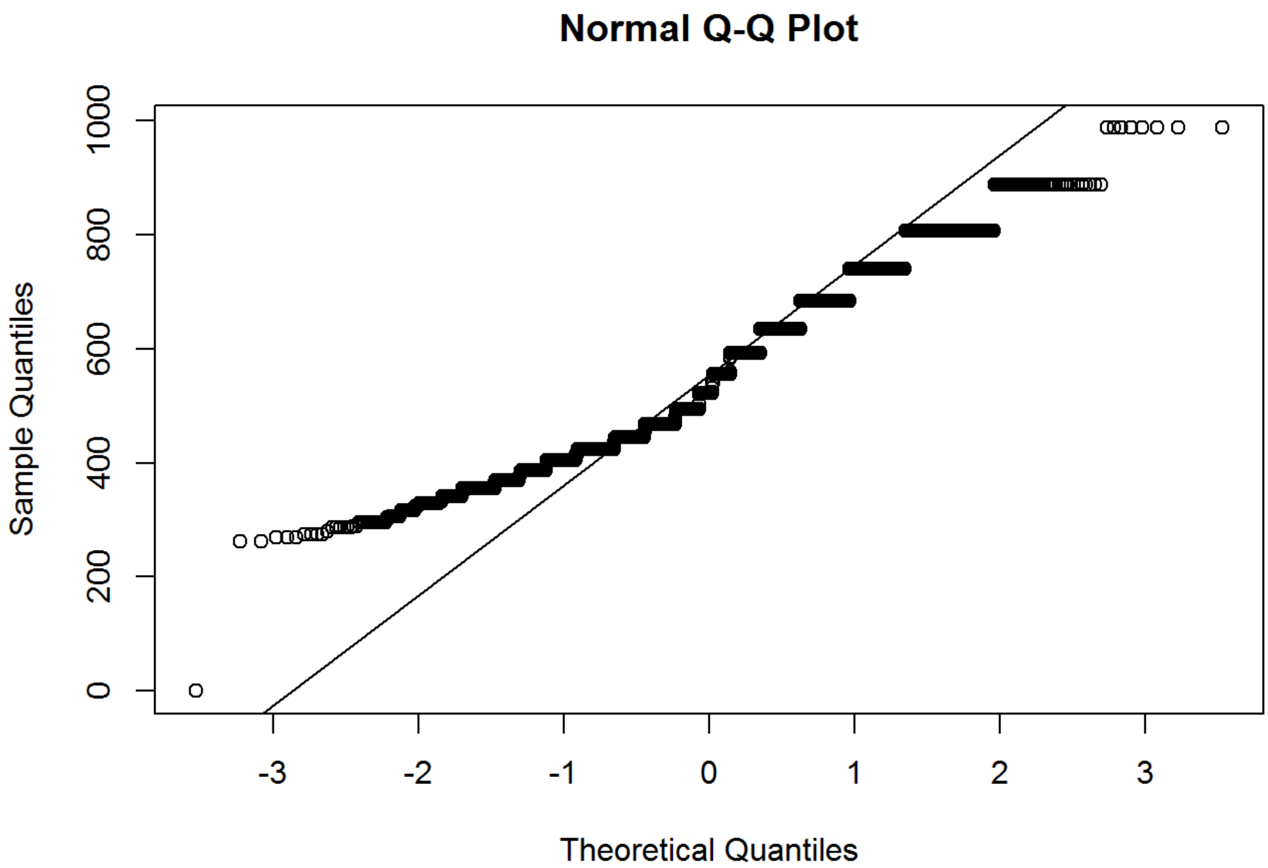

```
qqnorm(data1$co2TailpipeGpm)
qqline(data1$co2TailpipeGpm)
```



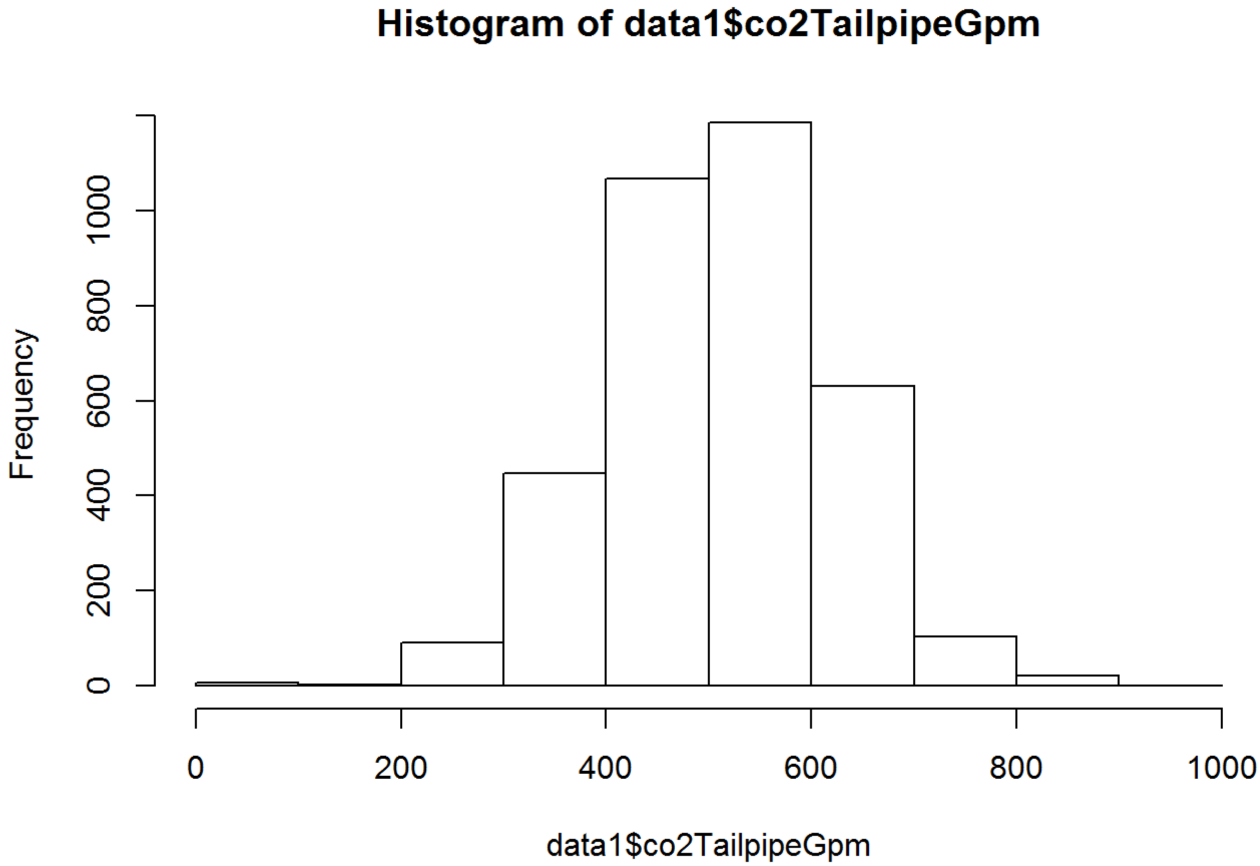

```
data2 <- subset(vehicles, vehicles$make == "Dodge")
hist(data1$co2TailpipeGpm)
```



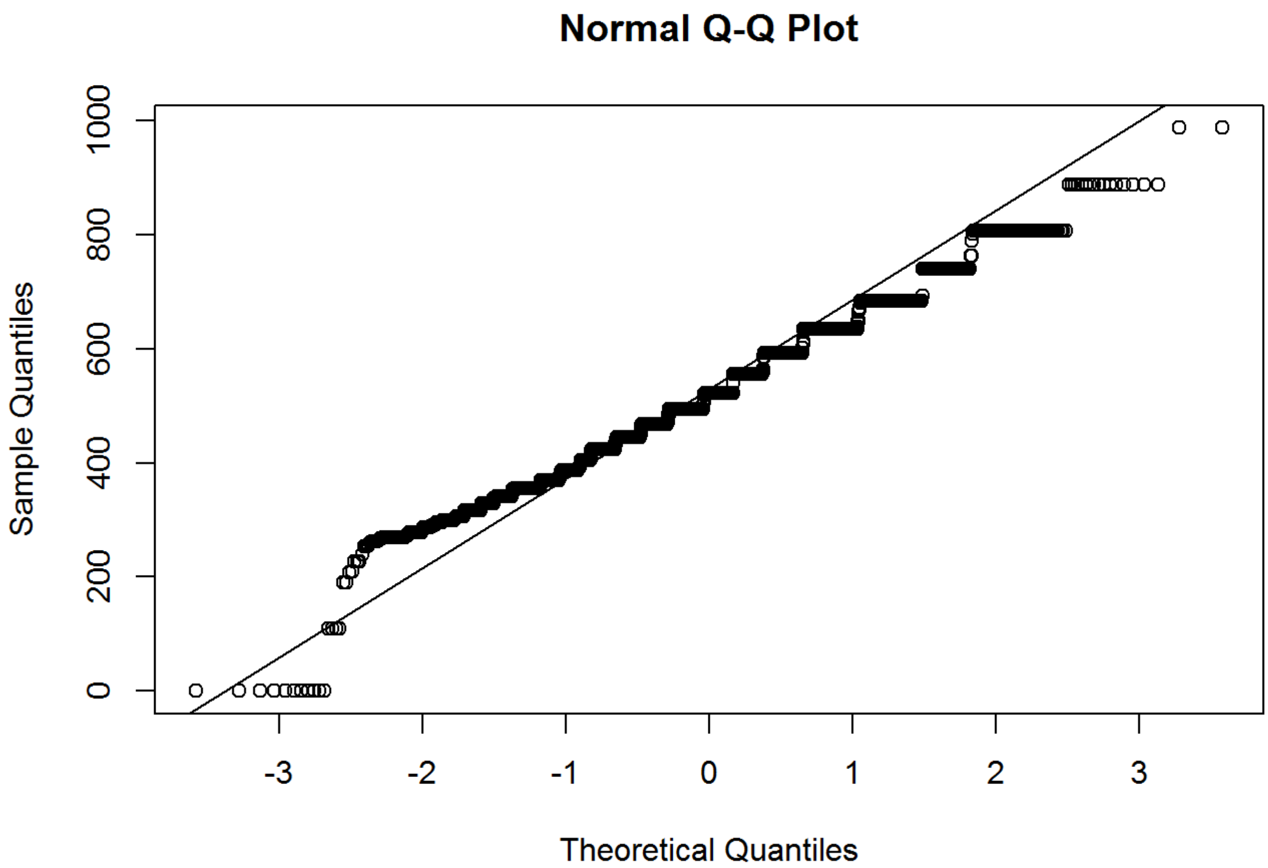

```
qqnorm(data2$co2TailpipeGpm)
qqline(data2$co2TailpipeGpm)
```



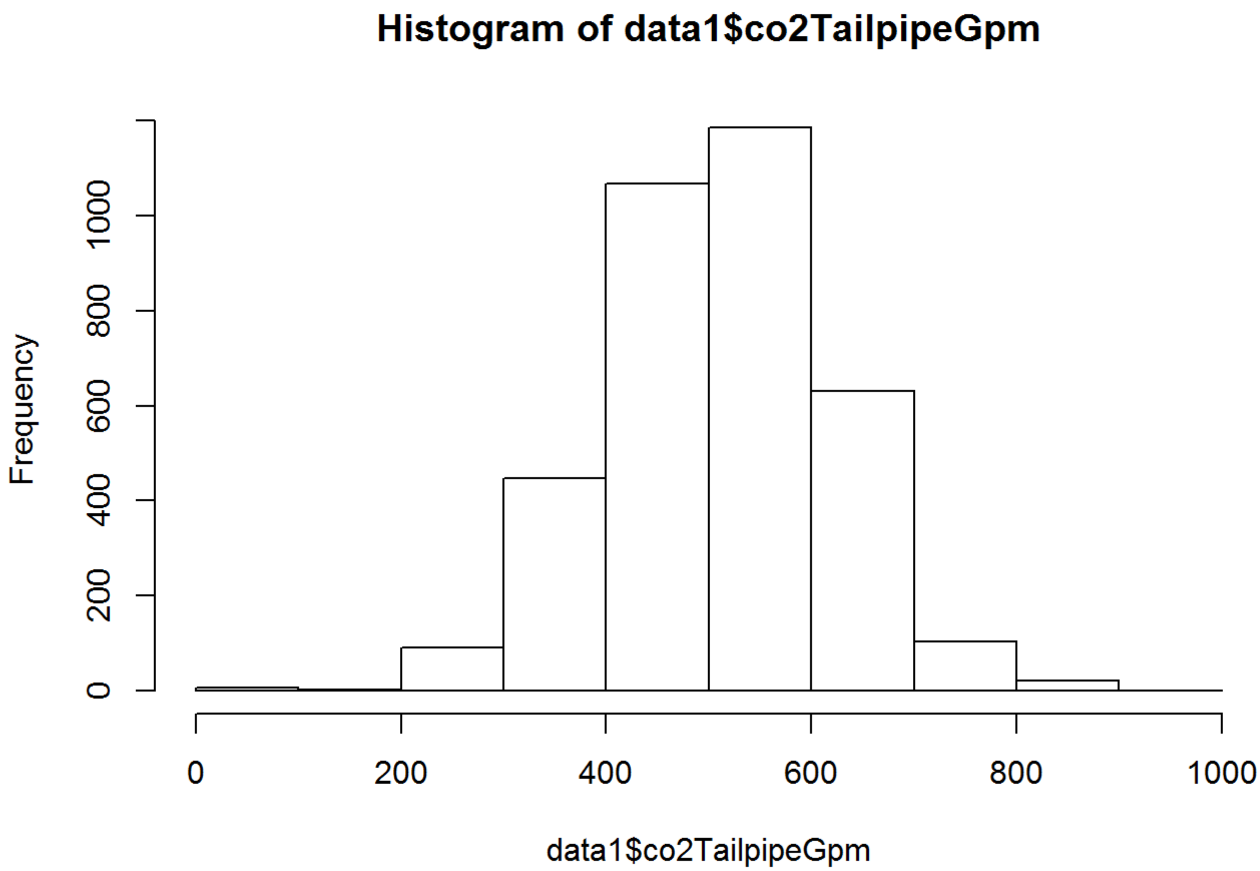

```
data3 <- subset(vehicles, vehicles$make == "Ford")
hist(data1$co2TailpipeGpm)
```



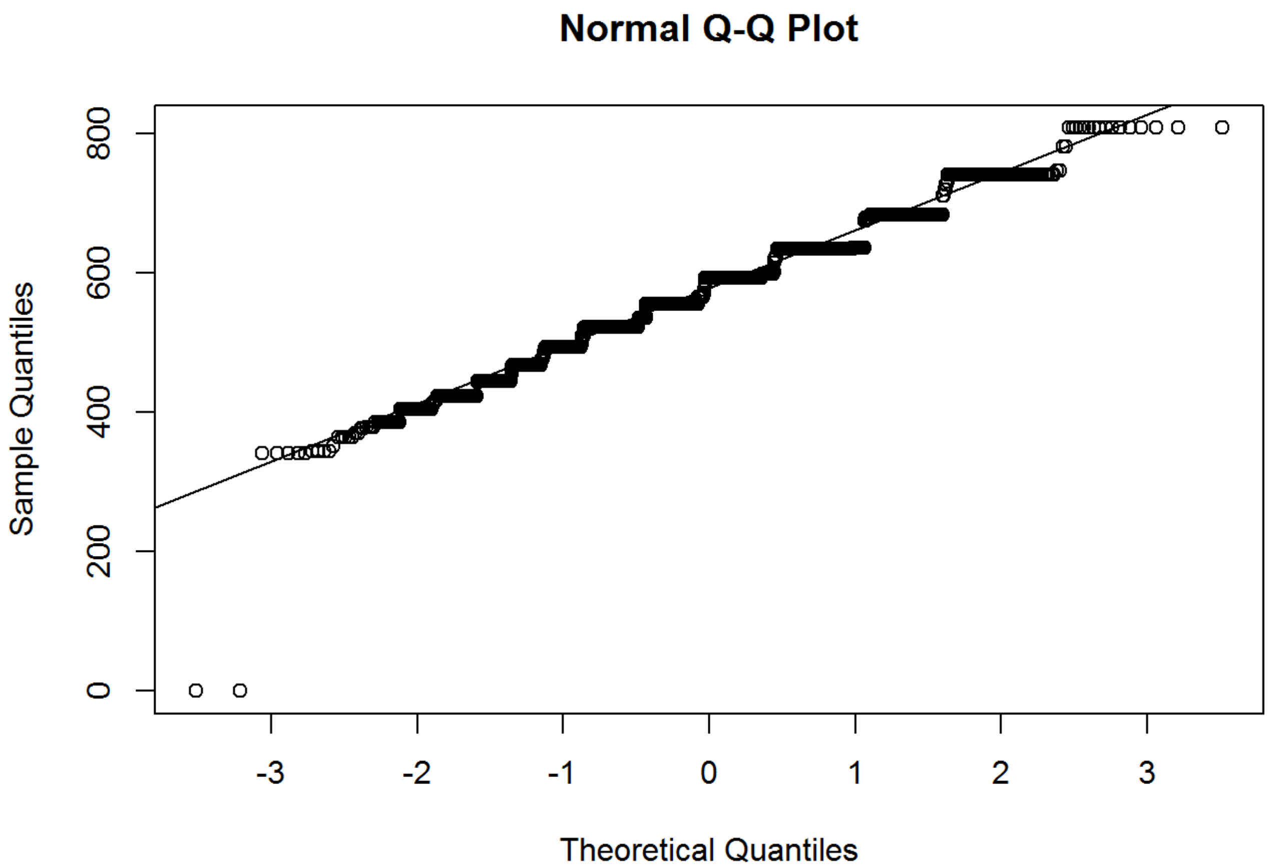

```
qqnorm(data3$co2TailpipeGpm)
qqline(data3$co2TailpipeGpm)
```



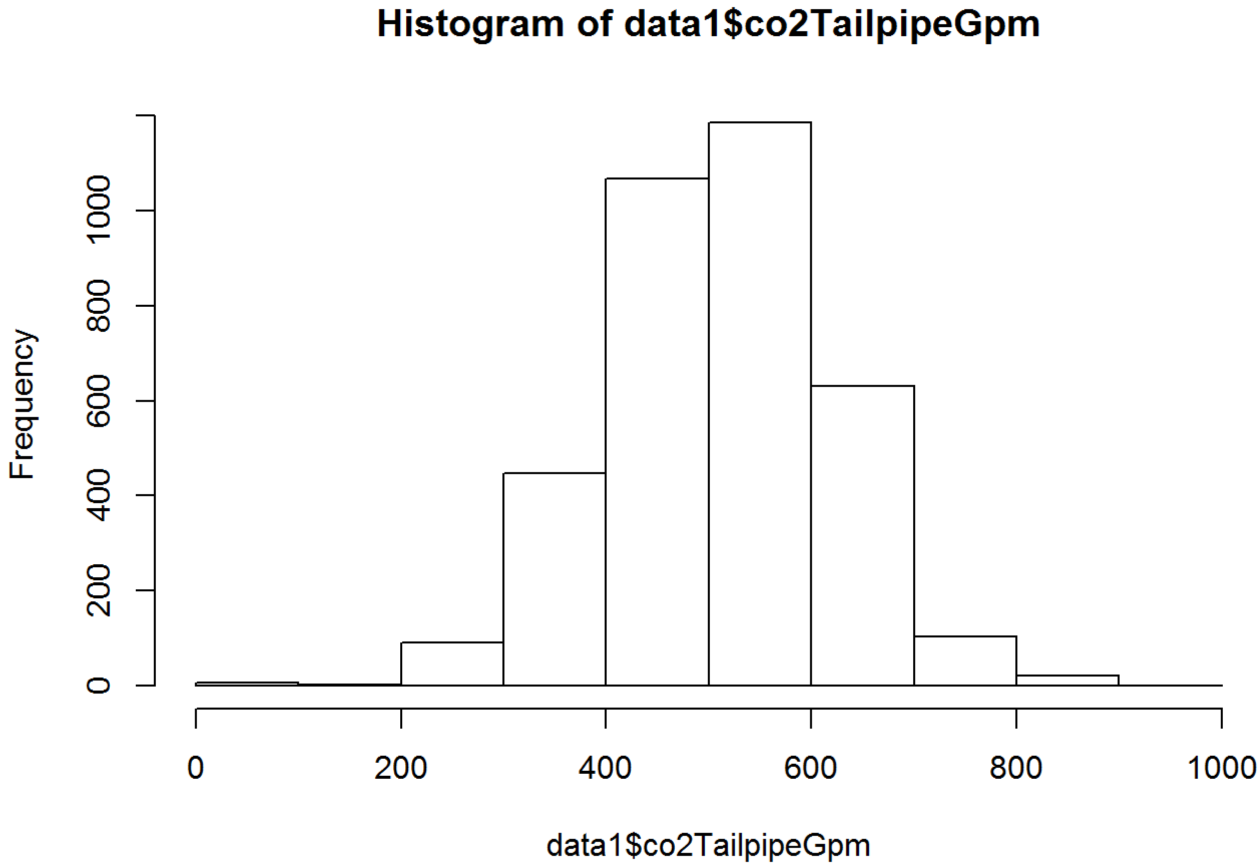

```
data4 <- subset(vehicles, vehicles$make == "GMC")
hist(data1$co2TailpipeGpm)
```



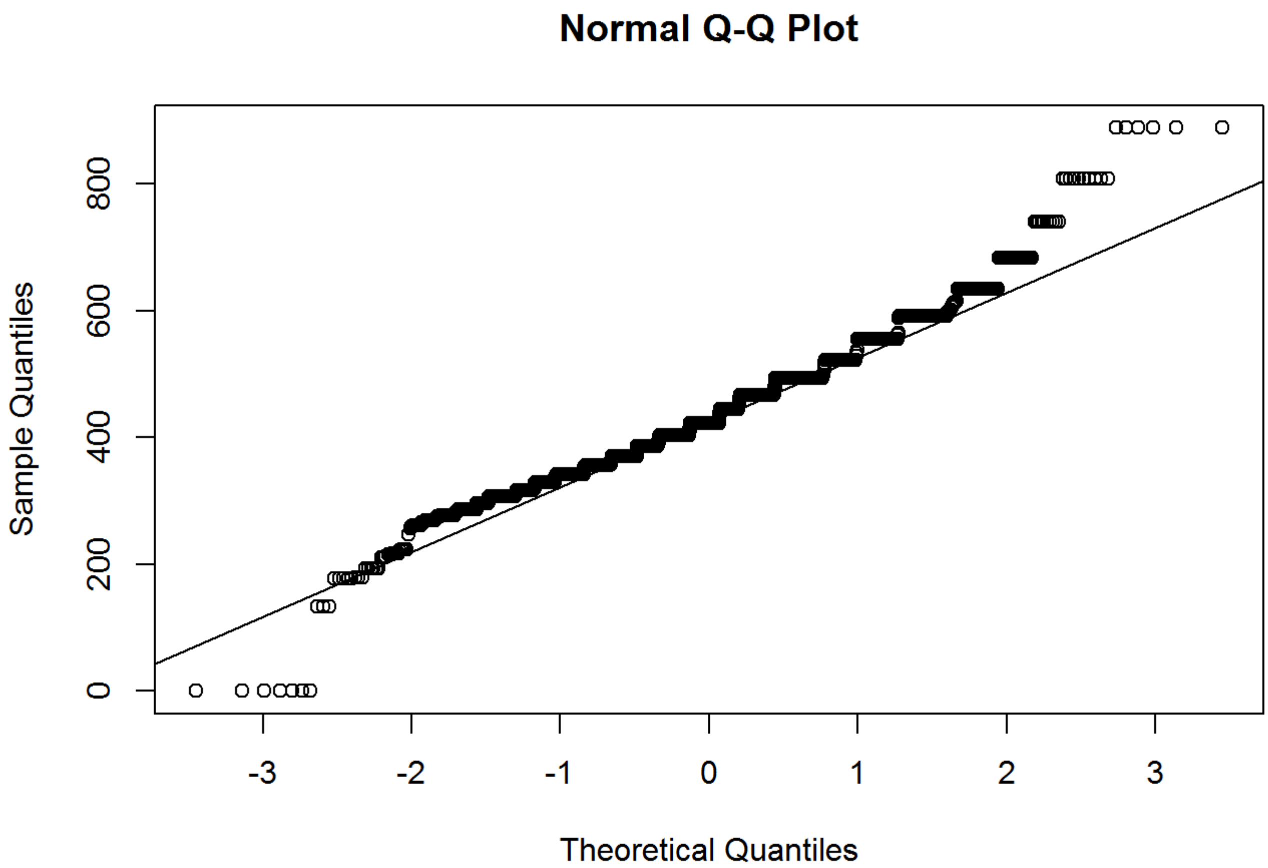

```
qqnorm(data4$co2TailpipeGpm)
qqline(data4$co2TailpipeGpm)
```




```
data5 <- subset(vehicles, vehicles$make == "Toyota")
hist(data1$co2TailpipeGpm)
```



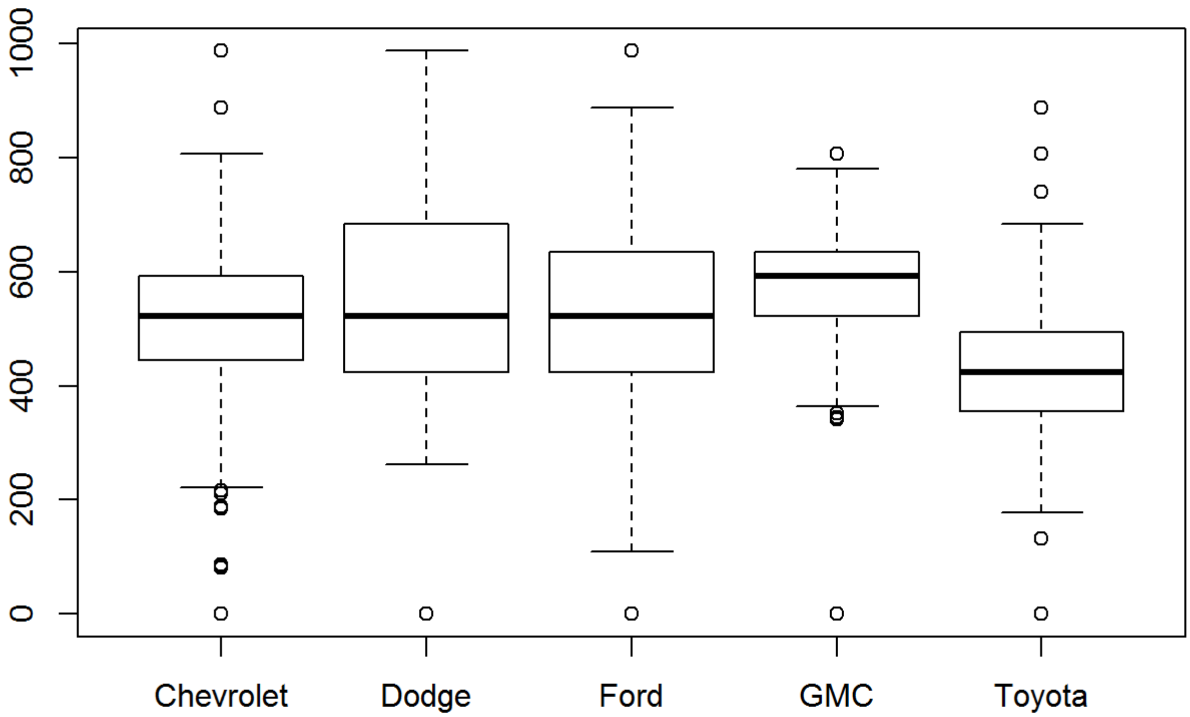

```
qqnorm(data5$co2TailpipeGpm)
qqline(data5$co2TailpipeGpm)
```



3. The variability across the groups should be about equal. Especially important when the sample sizes differ between groups.

From the boxplot of carbon dioxide emission for each manufacturers, it seems like each group has very different spread of data. For instance GMC, carbon dioxide emission of its models fall into a very narrow range. The opposite is true for dodge or ford. From summary statistics created by describeBy function, we can tell between groups there exist five standard deviations that are very different from each other. Some are in the 80's, some are more than 140's.

```
boxplot(vehicles$co2TailpipeGpm ~ vehicles$make)
```




```
describeBy(vehicles$co2TailpipeGpm, group = vehicles$make, skew = FALSE, mat = TRUE)
```

```
##     item   group1 vars    n     mean        sd min      max    range
## X11    1 Chevrolet    1 3555 517.1532 113.81648   0 987.4444 987.4444
## X12    2     Dodge    1 2431 553.4620 148.05770   0 987.4444 987.4444
## X13    3      Ford    1 2892 520.7487 132.60657   0 987.4444 987.4444
## X14    4       GMC    1 2267 574.7691  85.13641   0 807.9091 807.9091
## X15    5    Toyota    1 1775 436.5812 109.57771   0 888.7000 888.7000
##           se
## X11 1.908910
## X12 3.002884
## X13 2.465846
## X14 1.788091
## X15 2.600896
```

Therefore, since the data does not meet the third condition for conducting an ANOVA test. I will have to use the simulation based method.

Simulation Based Inference: Null Hypothesis: H0: (Chevrolet) = (Ford) = (Dodge) = (GMC) = (Toyota) Alternative Hypothesis: HA: (Chevrolet) != (Ford) != (Dodge) != (GMC) != (Toyota)

First step, I have to generate those random numbers that are normally distributed. From the summary statistics, I can obtain the mean and standard deviations for each group. And I also keep the sample size constant at 100.

```
set.seed(888)
chevrolet_carbon <- rnorm(100, mean = 517.153, sd = 113.816)
dodge_carbon <- rnorm(100, mean = 553.462, sd = 148.058)
```

```
ford_carbon <- rnorm(100, mean = 520.749, sd = 132.607)
gmc_carbon <- rnorm(100, mean = 574.769, sd = 85.136)
toyota_carbon <- rnorm(100, mean = 436.581, sd = 109.578)
```

Afterwards, I created an data frame that will be pass on as parameters for function aov, which will be responsible for creating ANOVA test.

```
carbon <- c(chevrolet_carbon, dodge_carbon, ford_carbon, gmc_carbon, toyota_carbon)
manufacturers <- c(replicate(100, "Chevrolet"), replicate(100, "Dodge"), replicate(100, "Ford"), r
eplicate(100, "GMC"), replicate(100, "Toyota"))
cars <- data.frame(carbon, manufacturers)
```

The result of the following code tells us the most information. From this summary statistics, we can know the degree of freedom of this test is 4. Sum of square between groups (SSG) is 981332 and sum of squares error (SSE) is 6632797. So the Sum of squares total(SST) is 7614129. The F value is equal to mean square errors between group (MSG) divided by mean square errors within group (MSE). 245333 / 13400 = 18.31 which is quite large. The p-value for F-statistic is closed to 0. Therefore, the differences between the mean of amount of carbon dioxide emission from vehicles of each manufacturers are statistically significant.

```
car_aov <- aov(cars$carbon ~ cars$manufacturers)
summary(car_aov)
```

```
##                      Df  Sum Sq Mean Sq F value Pr(>F)    
## cars$manufacturers   4 1264284  316071   22.84 <2e-16 ***
## Residuals          495 6851187   13841                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the two-way ANOVA test only tell us if the means are different across group, it does not tell me if the mean of one group is higher or lower than the other. The function called TukeyHSD shows us the result from comparing groups pair-wise. It also gives us the p-value to show if the differences are statistically significant. In addition, the confint function gives me the 95% confidence interval of that result. In this case, the R is using Chevrolet as a reference.

```
TukeyHSD(car_aov)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
## 
## Fit: aov(formula = cars$carbon ~ cars$manufacturers)
## 
## $`cars$manufacturers`
##                        diff          lwr        upr     p adj
## Dodge-Chevrolet    64.552525   19.0006265 110.104423 0.0011182
## Ford-Chevrolet     12.795943  -32.7559557  58.347841 0.9393668
## GMC-Chevrolet      57.578699   12.0268003 103.130597 0.0052783
## Toyota-Chevrolet  -75.516241 -121.0681394 -29.964343 0.0000695
## Ford-Dodge        -51.756582  -97.3084806  -6.204684 0.0168168
## GMC-Dodge          -6.973826  -52.5257246  38.578072 0.9935318
## Toyota-Dodge     -140.068766 -185.6206644 -94.516868 0.0000000
## GMC-Ford           44.782756   -0.7691424  90.334654 0.0566113
## Toyota-Ford       -88.312184 -133.8640822 -42.760285 0.0000017
## Toyota-GMC       -133.094940 -178.6468381 -87.543041 0.0000000
```

```
confint(car_aov)
```

```
##                                  2.5 %    97.5 %
## (Intercept)                  489.55067 535.78042
## cars$manufacturersDodge       31.86315  97.24190
## cars$manufacturersFord       -19.89343  45.48531
## cars$manufacturersGMC         24.88933  90.26807
## cars$manufacturersToyota    -108.20561 -42.82687
```

5. Conclusion: Write a brief summary of your findings without repeating your statements from earlier. Also include a discussion of what you have learned about your research question and the data you collected. You may also want to include ideas for possible future research.

According to the result we obtained from the above, Toyota is the winner among all in terms of the lowest carbon dioxide emission from the vehicles which it manufactures. If we directly compare the mean of carbon dioxide emission of Toyota's vehicle against thost of others, the differences are all statistically significant. (p-value less than 5%) and the confidence level way below 0.

From this research questions, I realize under the regulation of EPA on carbon dioxide emission, some car manufacturer not only meet the standard, but they are way better than the standard. If I got a chance to do further study on this dataset, I want to investigate which manufacturer produce cars that are most energy efficient.