ANLY 511 Final Project - T Tests/ Mann-Whitney U Tests

Mia Mayerhofer

2022-11-26

Data Preparation

```
# Load libraries
library(dplyr)
library(tidyverse)
library(RColorBrewer)
library(car)
```

```
# Create color palettes
Blues <- colorRampPalette(c("#0A146B", "#A9A3DA"))
Purples <- colorRampPalette(c("#3E1370", "#BDA3DA"))
GrBuPuPi <- c("#095826", "#0E7032", "#10913F", "#55A472", "#8CBF9E", "#8CBFB8",
              "#63B7AC", "#2D9A8B", "#137568", "#094E45", "#0B3C5C", "#17547C",
              "#2671A4", "#3C8CC1", "#72B1DB", "#96C3E1", "#B0CDE1", "#B0B3E1",
              "#858ACD", "#4F55AB", "#1923B3", "#0E1468", "#3C1075", "#5821A1",
              "#6B27C4", "#9455E5", "#A278D8", "#A990CA", "#ADA0BF", "#C1A5CB",
              "#B887CA", "#A35CBD", "#762594")
```

```
# Load in the cleaned csv data for nonelectric vehicles
gas <- read.csv("../data/cardata_nonelectric_clean.csv")
# Remove index columns
gas <- gas[,-1]
# View the data
head(gas)
```

```
##   Model.Year Vehicle.Manufacturer.Name Veh.Mfr.Code Represented.Test.Veh.Make
## 1       2018              aston martin          ASX              Aston Martin
## 2       2018              aston martin          ASX              Aston Martin
## 3       2018              aston martin          ASX              Aston Martin
## 4       2018              aston martin          ASX              Aston Martin
## 5       2018              aston martin          ASX              Aston Martin
## 6       2018              aston martin          ASX              Aston Martin
##   Represented.Test.Veh.Model Test.Veh.Displacement..L. Vehicle.Type
## 1                       DB11                       5.2          Car
## 2                       DB11                       5.2          Car
## 3                    DB11 V8                       4.0          Car
## 4                    DB11 V8                       4.0          Car
## 5                   Rapide S                       6.0          Car
## 6                   Rapide S                       6.0          Car
```

```
##   Rated.Horsepower X..of.Cylinders.and.Rotors Tested.Transmission.Type.Code
## 1              600                         12                           SA
## 2              600                         12                           SA
## 3              503                          8                           SA
## 4              503                          8                           SA
## 5              552                         12                           SA
## 6              552                         12                           SA
##   Tested.Transmission.Type X..of.Gears Transmission.Lockup. Drive.System.Code
## 1           Semi-Automatic           8                    Y                R
## 2           Semi-Automatic           8                    Y                R
## 3           Semi-Automatic           8                    Y                R
## 4           Semi-Automatic           8                    Y                R
## 5           Semi-Automatic           8                    Y                R
## 6           Semi-Automatic           8                    Y                R
##   Drive.System.Description Equivalent.Test.Weight..lbs.. Axle.Ratio N.V.Ratio
## 1     2-Wheel Drive, Rear                           4500       2.70      22.2
## 2     2-Wheel Drive, Rear                           4500       2.70      22.2
## 3     2-Wheel Drive, Rear                           4500       2.70      22.2
## 4     2-Wheel Drive, Rear                           4500       2.70      22.2
## 5     2-Wheel Drive, Rear                           4750       2.73      22.4
## 6     2-Wheel Drive, Rear                           4750       2.73      22.4
##   Test.Fuel.Type.Description THC..g.mi. CO..g.mi. CO2..g.mi. RND_ADJ_FE
## 1      Tier 2 Cert Gasoline   0.024700  0.418000     466.87       18.8
## 2      Tier 2 Cert Gasoline   0.001155  0.067334     285.00       30.9
## 3      Tier 2 Cert Gasoline   0.026500  0.070000     386.66       22.7
## 4      Tier 2 Cert Gasoline   0.000500  0.030000     259.74       33.8
## 5      Tier 2 Cert Gasoline   0.026900  0.500000     511.93       17.3
## 6      Tier 2 Cert Gasoline   0.000800  0.060000     296.63       29.9
##   DT.Inertia.Work.Ratio.Rating DT.Absolute.Speed.Change.Ratg
## 1                   -2.5300000                   -1.7300000
## 2                    1.3600000                    0.4400000
## 3                  -11.9900000                   -9.2600000
## 4                   -3.6400000                   -3.2100000
## 5                    0.5655838                    0.4420405
## 6                    0.5655838                    0.4420405
##   DT.Energy.Economy.Rating Target.Coef.A..lbf. Target.Coef.B..lbf.mph.
## 1               -1.7100000               40.94                  0.0169
## 2               -0.5900000               40.94                  0.0169
## 3               -7.7100000               40.94                  0.0169
## 4               -0.9600000               40.94                  0.0169
## 5               -0.2002973               32.66                  0.6085
## 6               -0.2002973               32.66                  0.6085
##   Target.Coef.C..lbf.mph..2. Set.Coef.A..lbf. Set.Coef.B..lbf.mph.
## 1                     0.0271           6.810               0.0807
## 2                     0.0271           6.810               0.0807
## 3                     0.0271          11.260               0.0919
## 4                     0.0271          11.260               0.0919
## 5                     0.0198           1.093               2.1980
## 6                     0.0198           1.093               2.1980
##   Set.Coef.C..lbf.mph..2. Aftertreatment.Device.Cd Aftertreatment.Device.Desc
## 1                  0.0245                      TWC         Three-way catalyst
## 2                  0.0245                      TWC         Three-way catalyst
## 3                  0.0251                      TWC         Three-way catalyst
## 4                  0.0251                      TWC         Three-way catalyst
```

```
## 5                  0.0280                       TWC         Three-way catalyst
## 6                  0.0280                       TWC         Three-way catalyst
##   Police...Emergency.Vehicle. Averaging.Method.Cd Averging.Method.Desc
## 1                           N                    N         No averaging
## 2                           N                    N         No averaging
## 3                           N                    N         No averaging
## 4                           N                    N         No averaging
## 5                           N                    N         No averaging
## 6                           N                    N         No averaging
```

Exploratory Data Analysis (EDA) for Test 1

Research Question: Is there a significant difference in the amount of carbon dioxide emissions between types of fuel, specifically the two most common fuel types?

How many observations are their for each type of fuel?

```
# Create a frequency table
frequencies <- data.frame(cbind(table(gas$`Fuel Type`)))
frequencies$`Fuel Type` <- row.names(frequencies)
frequencies$`Frequency` <- frequencies$cbind.table.gas..Fuel.Type...
frequencies <- frequencies %>% dplyr::select("Fuel Type", "Frequency")
rownames(frequencies) <- NULL
# Print table ordered by frequency
frequencies[order(frequencies$Frequency, decreasing = TRUE),]
```

```
## [1] Fuel Type Frequency
## <0 rows> (or 0-length row.names)
```

From the frequency table above, Tier 2 Cert Gasoline and Federal Cert Diesel 7-15 PPM Sulfur are the two most common gasoline types in the data set with 19,235 and 974 observations respectively.

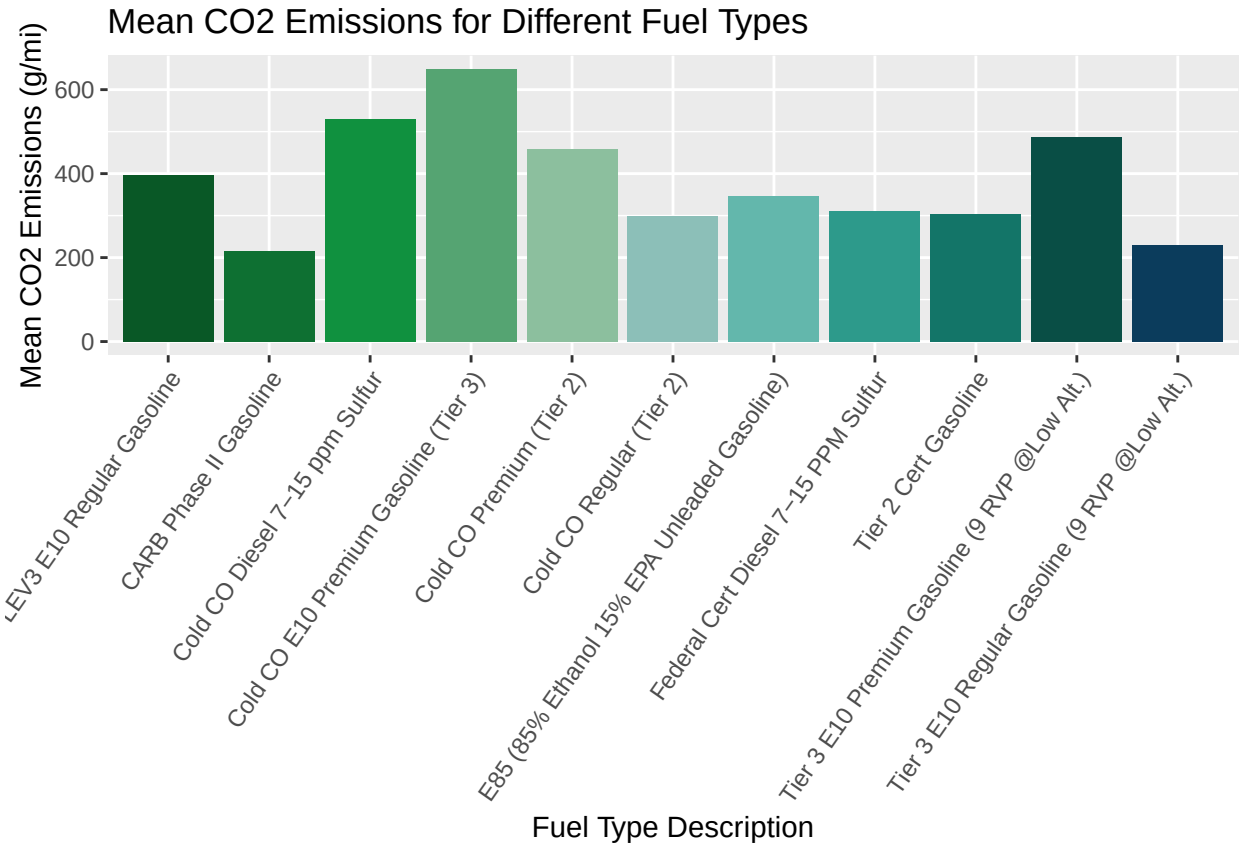
Which fuel type produces the most carbon dioxide emissions in this data set?

```
# Calculate the mean CO2 emissions for each fuel type
means_fuel <- gas %>% group_by(Test.Fuel.Type.Description) %>%
  summarise_at(vars(CO2..g.mi.), list(name = mean))
colnames(means_fuel) <- c("Fuel Type", "Mean CO2 Emissions")
# Print means ordered by mean
print(means_fuel[order(means_fuel$`Mean CO2 Emissions`, decreasing = TRUE),])
```

```
## # A tibble: 11 x 2
##    `Fuel Type`                                  `Mean CO2 Emissions`
##    <chr>                                                       <dbl>
##  1 Cold CO E10 Premium Gasoline (Tier 3)                        648.
##  2 Cold CO Diesel 7-15 ppm Sulfur                               530.
##  3 Tier 3 E10 Premium Gasoline (9 RVP @Low Alt.)                486.
##  4 Cold CO Premium (Tier 2)                                     458.
##  5 CARB LEV3 E10 Regular Gasoline                               397.
##  6 E85 (85% Ethanol 15% EPA Unleaded Gasoline)                  347.
##  7 Federal Cert Diesel 7-15 PPM Sulfur                          310.
##  8 Tier 2 Cert Gasoline                                         303.
##  9 Cold CO Regular (Tier 2)                                     298.
```

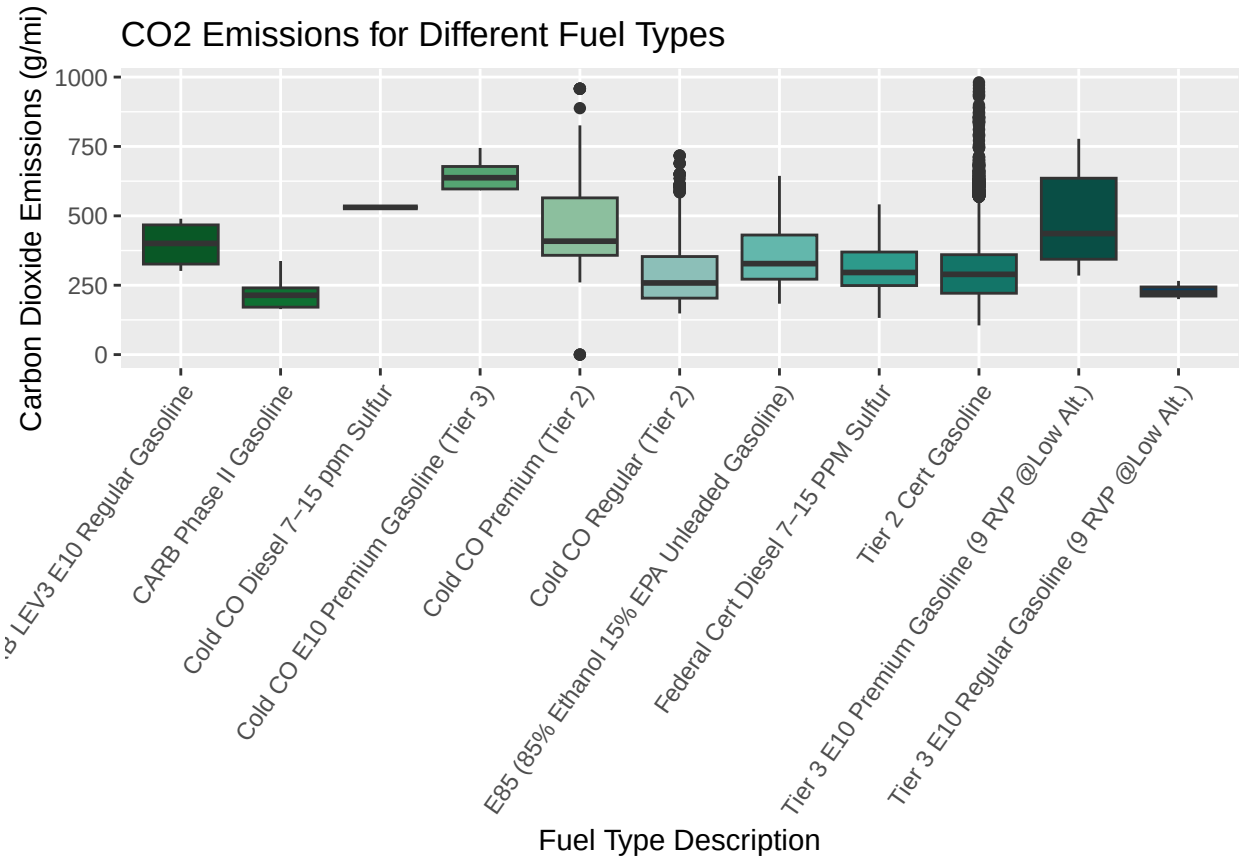
```
## 10 Tier 3 E10 Regular Gasoline (9 RVP @Low Alt.)                229.
## 11 CARB Phase II Gasoline                                       215.
```

```
# Plot a barplot of the means
means_fuel %>% ggplot(aes(x = `Fuel Type`, y = `Mean CO2 Emissions`,
                          fill = `Fuel Type`)) +
  geom_bar(stat = "identity", show.legend = FALSE) +
  ggtitle("Mean CO2 Emissions for Different Fuel Types") +
  xlab("Fuel Type Description") +
  ylab("Mean CO2 Emissions (g/mi)") +
  theme(axis.text.x = element_text(angle = 55, vjust = 1, hjust=1)) +
  scale_fill_manual(values = GrBuPuPi)
```



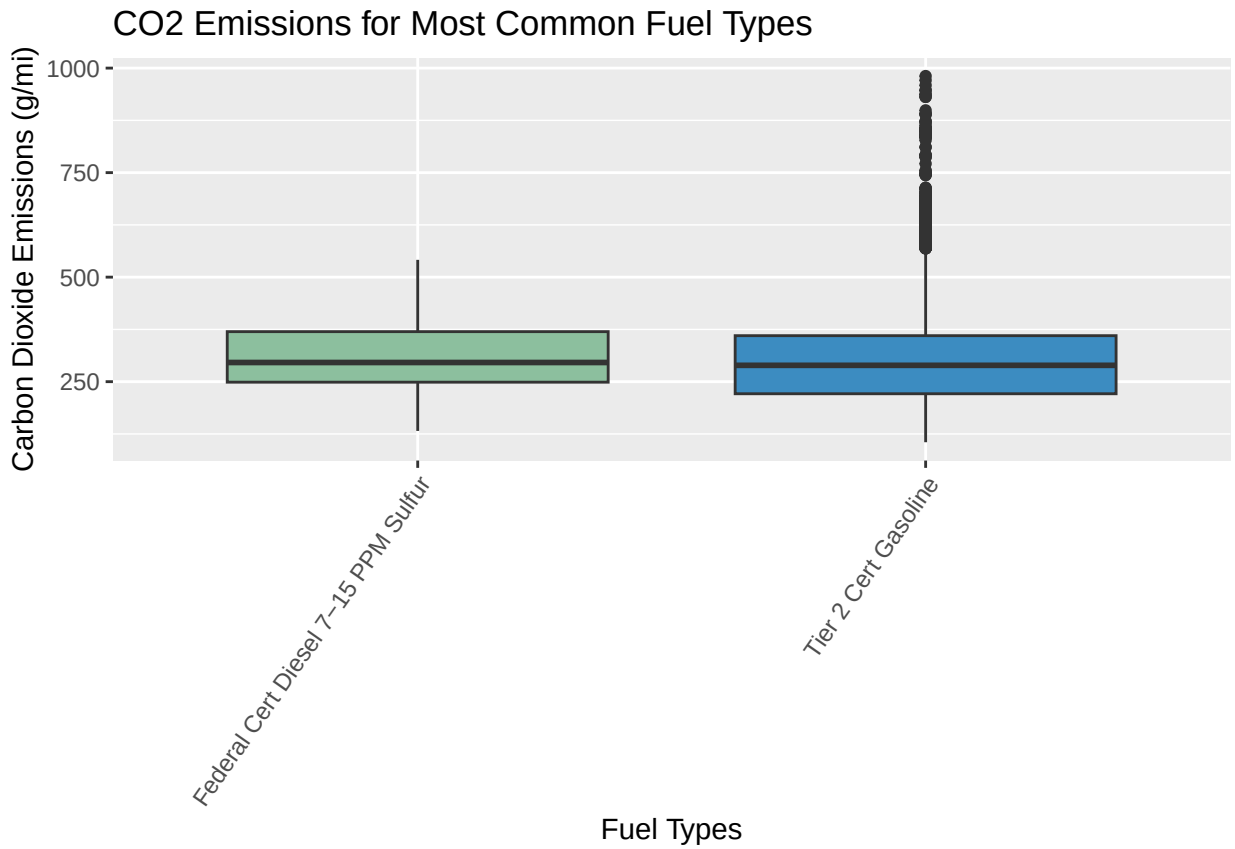
From the barplot and table above, it is clear that Cold CO E10 Premium Gasoline (Tier 3) produces the most carbon dioxide emissions out of all the different fuel types followed by Cold CO Diesel7-15 ppm Sulfur and Tier 3 E10 Premium Gasoline (9 RVP Low Alt.). The fuel types with the lowest mean carbon dioxide emissions are Cold CO Regular (Tier 2), Tier 3 E10 Regular Gasoline (9 RVP Low Alt.), and CARB Phase II Gasoline. Below, we will plot the boxplots of each fuel type to view the distributions and outliers.

```
names(gas)[names(gas) == 'Test.Fuel.Type.Description'] <- 'Fuel Type'
gas %>% ggplot(aes(x = `Fuel Type`, y = CO2..g.mi., fill =`Fuel Type`)) +
  geom_boxplot(show.legend = FALSE) +
  ggtitle("CO2 Emissions for Different Fuel Types") +
  xlab("Fuel Type Description") + ylab("Carbon Dioxide Emissions (g/mi)") +
  theme(axis.text.x = element_text(angle = 55, vjust = 1, hjust=1)) +
  scale_fill_manual(values = GrBuPuPi)
```



From the boxplots above, we can see that the mean carbon dioxide emissions varies depending on the fuel type. It is clear that some gasolines' mean carbon dioxide emissions differ more significantly than others. Tier 2 Cert Gasoline has the most outliers out of the fuel types. Let's look closer at the boxplots of just the top two fuel types: Tier 2 Cert Gasoline and Federal Cert Diesel 7-15 PPM Sulfur.

```
top2fueltypes <- gas[gas$`Fuel Type` %in% c("Federal Cert Diesel 7-15 PPM Sulfur",
                                            "Tier 2 Cert Gasoline"), ]
top2fueltypes %>% ggplot(aes(x = `Fuel Type`,
                             y = CO2..g.mi.,
                             fill = `Fuel Type`)) +
  geom_boxplot(show.legend = FALSE) +
  ggtitle("CO2 Emissions for Most Common Fuel Types") +
  xlab("Fuel Types") +
  ylab("Carbon Dioxide Emissions (g/mi)") +
  theme(axis.text.x = element_text(angle = 55,
                                   vjust = 1,
                                   hjust=1)) +
  scale_fill_manual(values = GrBuPuPi[c(5, 14, 20)])
```



We can see that the means are relatively similar; thus, the test below will determine whether or not there is a significant difference.

Test 1

Let us compare the mean emissions between the two most common fuel types in the data set. Below we will test to see if there are statistically significant difference in the mean emissions between Tier 2 Cert Gasoline and Federal Cert Diesel 7-15 PPM Sulfur. We will define the following null and alternative hypotheses:

Declaring Hypotheses and Significance Level

H_0 : The mean carbon dioxide emissions is the same for Tier 2 Cert Gasoline and Federal Cert Diesel 7-15 PPM Sulfur.

H_a : The mean carbon dioxide emissions is greater for Federal Cert Diesel 7-15 PPM Sulfur than Tier 2 Cert Gasoline.

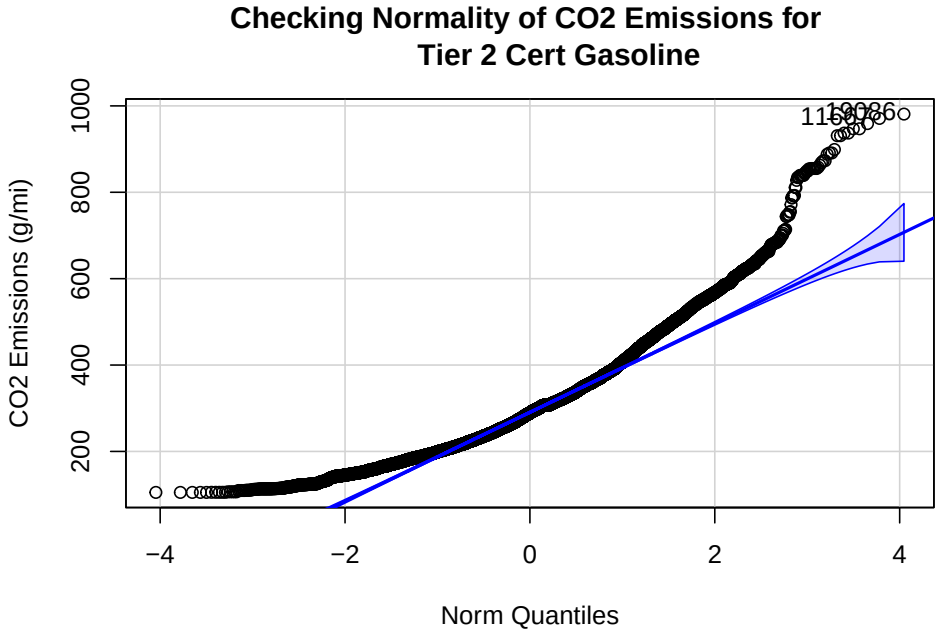
Significance Level: 1%

Checking Assumptions

```
# Separate into two data frames filtered by each type
tier2Cert <- gas %>%
  filter(`Fuel Type` == "Tier 2 Cert Gasoline")
```

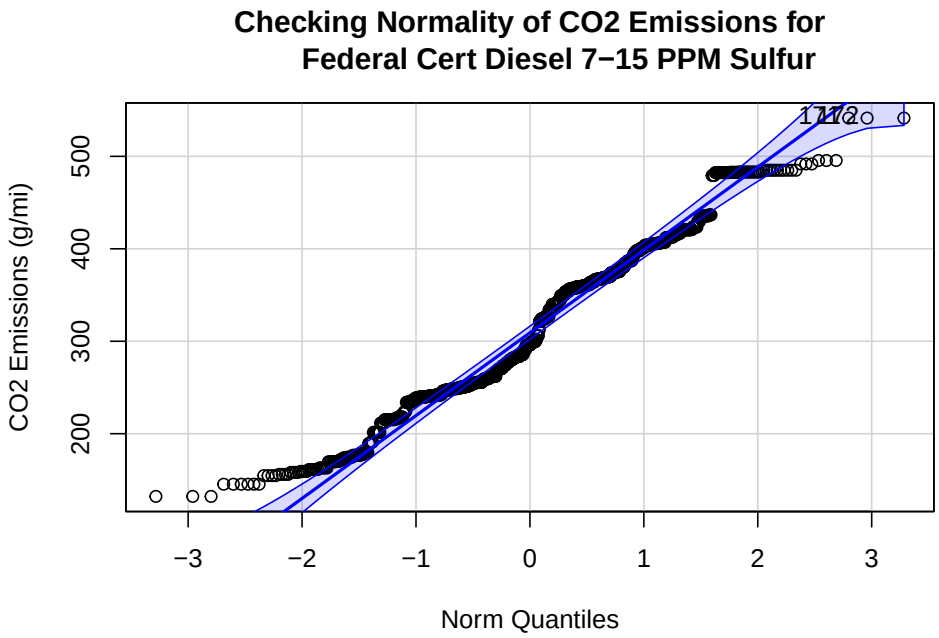
```r
fedCertDieselSulfur <- gas %>%
  filter(`Fuel Type` == "Federal Cert Diesel 7-15 PPM Sulfur")
```

```r
# Population 1: Tier 2 Cert Gasoline
qqPlot(tier2Cert$`CO2..g.mi.`,
       main = "Checking Normality of CO2 Emissions for
       Tier 2 Cert Gasoline",
       xlab = "Norm Quantiles",
       ylab = "CO2 Emissions (g/mi)")
```




```
## [1] 19086 11667
```

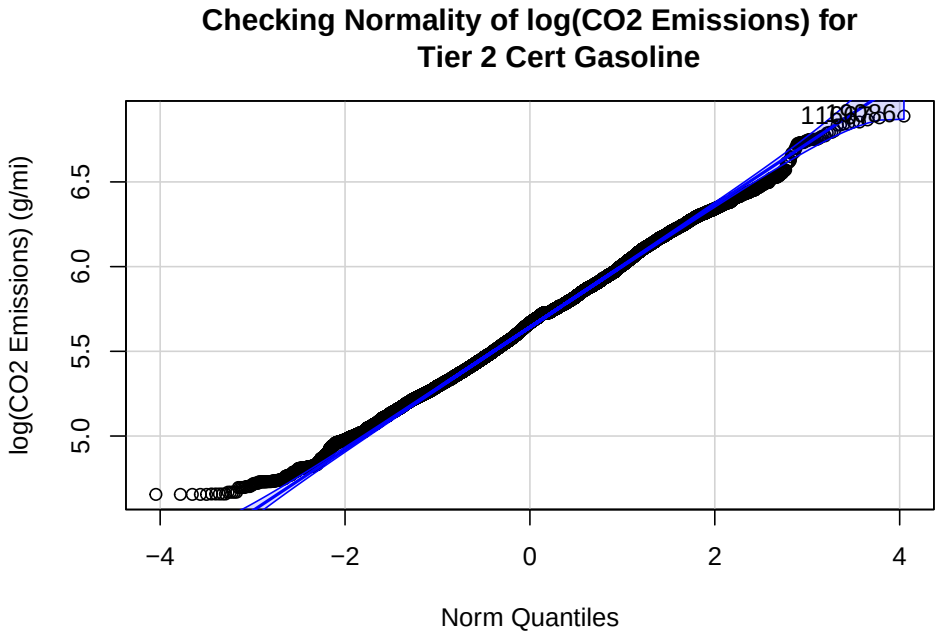
```r
# Population 2: Federal Cert Diesel 7-15 PPM Sulfur
qqPlot(fedCertDieselSulfur$`CO2..g.mi.`,
       main = "Checking Normality of CO2 Emissions for
       Federal Cert Diesel 7-15 PPM Sulfur",
       xlab = "Norm Quantiles",
       ylab = "CO2 Emissions (g/mi)")
```




```
## [1] 171 172
```

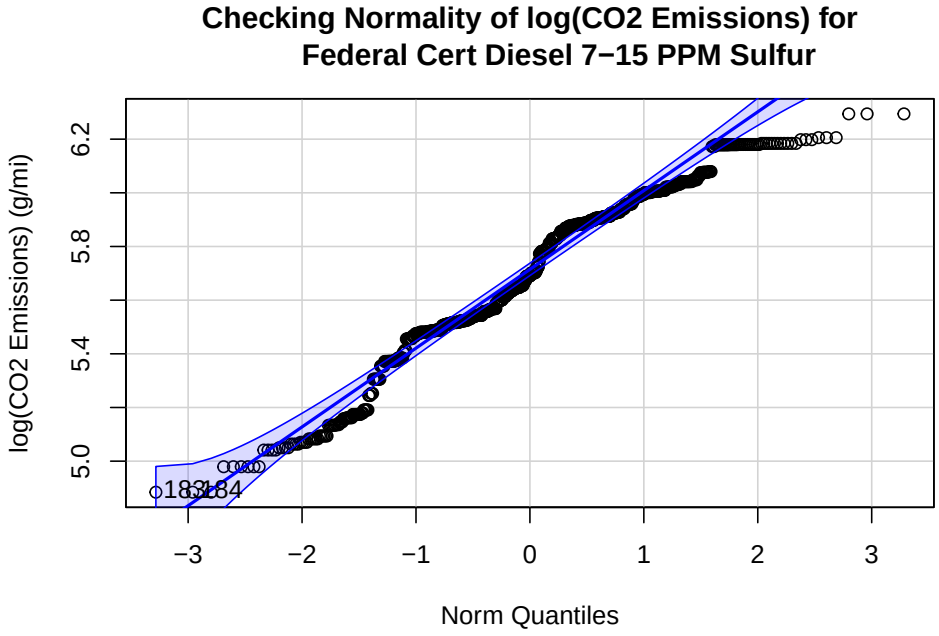
The QQ-plots show that the distribution of tier 2 cert gasoline carbon dioxide emissions is heavily skewed to the *right*, and the distribution of federal cert diesel carbon dioxide emissions is possibly *bimodal*. Thus, the normality assumption does not hold. Let us see if a log transformation normalizes the data:

```
# Population 1: Tier 2 Cert Gasoline
qqPlot(log(tier2Cert$`CO2..g.mi.`),
       main = "Checking Normality of log(CO2 Emissions) for
       Tier 2 Cert Gasoline",
       xlab = "Norm Quantiles",
       ylab = "log(CO2 Emissions) (g/mi)")
```




```
## [1] 19086 11667
```

```
# Population 2: Federal Cert Diesel 7-15 PPM Sulfur
qqPlot(log(fedCertDieselSulfur$`CO2..g.mi.`),
       main = "Checking Normality of log(CO2 Emissions) for
       Federal Cert Diesel 7-15 PPM Sulfur",
       xlab = "Norm Quantiles",
       ylab = "log(CO2 Emissions) (g/mi)")
```



```
## [1] 183 184
```

```
# Shapiro Test
shapiro.test(log(fedCertDieselSulfur$`CO2..g.mi.`))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  log(fedCertDieselSulfur$CO2..g.mi.)
## W = 0.96906, p-value = 1.548e-13
```

From the QQ-plots above, it is clear that the log transformation normalized the tier 2 cert gasoline data, but not the federal cert diesel data. The shapiro test result with a p-value of less than 0.01 confirms this result that the log of the federal cert diesel emissions is not normal. Thus, we will perform a Mann-Whitney U Test without the log transformation.

Mann-Whitney U Test

```
# Perform test
mw.test1 <- wilcox.test(fedCertDieselSulfur$`CO2..g.mi.`, tier2Cert$`CO2..g.mi.`,
                        na.rm = TRUE, paired = FALSE, exact = FALSE, conf.int = TRUE)
mw.test1
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  fedCertDieselSulfur$CO2..g.mi. and tier2Cert$CO2..g.mi.
## W = 10288865, p-value = 2.134e-07
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  10.61379 23.05800
## sample estimates:
## difference in location
##               16.96013
```

P-Value Analysis: Based on the test results above, the p-value $= 2.134e-07 < 0.01$ which is statistically significant. Thus, we reject the null hypothesis and conclude that the mean carbon dioxide emissions is greater for Federal Cert Diesel 7-15 PPM Sulfur than Tier 2 Cert Gasoline.

Confidence Interval Analysis: From the 95% confidence interval, we can conclude with 95% confidence that Federal Cert Diesel 7-15 PPM Sulfur, on average, produces between 10.61379 g/mi and 23.05800 g/mi more CO2 emissions than Tier 2 Cert Gasoline for the sample of vehicles in the data set.

Exploratory Data Analysis (EDA) for Test 2

Research Question: Is there a significant difference in the amount of carbon dioxide emissions between vehicle manufacturers, specifically the two most common manufacturers?

How many observations are their for each manufacturer?

```r
# Create a frequency table
frequencies <- data.frame(cbind(table(gas$Vehicle.Manufacturer.Name)))
frequencies$`Manufacturer` <- row.names(frequencies)
frequencies$`Frequency` <- frequencies$cbind.table.gas.Vehicle.Manufacturer.Name..
frequencies <- frequencies %>% dplyr::select("Manufacturer", "Frequency")
rownames(frequencies) <- NULL
# Print table ordered by frequency
frequencies[order(frequencies$Frequency, decreasing = TRUE),]
```

```
##            Manufacturer Frequency
## 9                    GM      2624
## 31               Toyota      2472
## 3                   BMW      2327
## 8                FOMOCO      2140
## 32  Volkswagen Group of      1703
## 10                Honda      1445
## 6           FCA US LLC      1399
## 23               Nissan      1299
## 11              Hyundai      1224
## 21       Mercedes-Benz       935
## 14                 Kia       899
## 25              Porsche       645
## 12  Jaguar Land Rover L       608
## 29               Subaru       446
## 19                MAZDA       430
## 7               Ferrari       272
## 33                Volvo       236
## 22 Mitsubishi Motors Co       195
## 18             Maserati       100
## 20   McLaren Automotive        85
## 26          Rolls-Royce        60
## 1          aston martin        50
## 27                Roush        37
## 17                Lotus        30
## 13  Karma Automotive, L        28
## 5             FCA Italy        15
## 28                  RUF         8
## 16          Lamborghini         7
## 2               Bentley         5
## 4                   EPA         4
## 24  Pagani Automobili S         4
## 30  SUBARU TECNICA INTE         4
## 15           Koenigsegg         2
```

From the frequency table above, the two most common manufacturers are GM and Toyota.

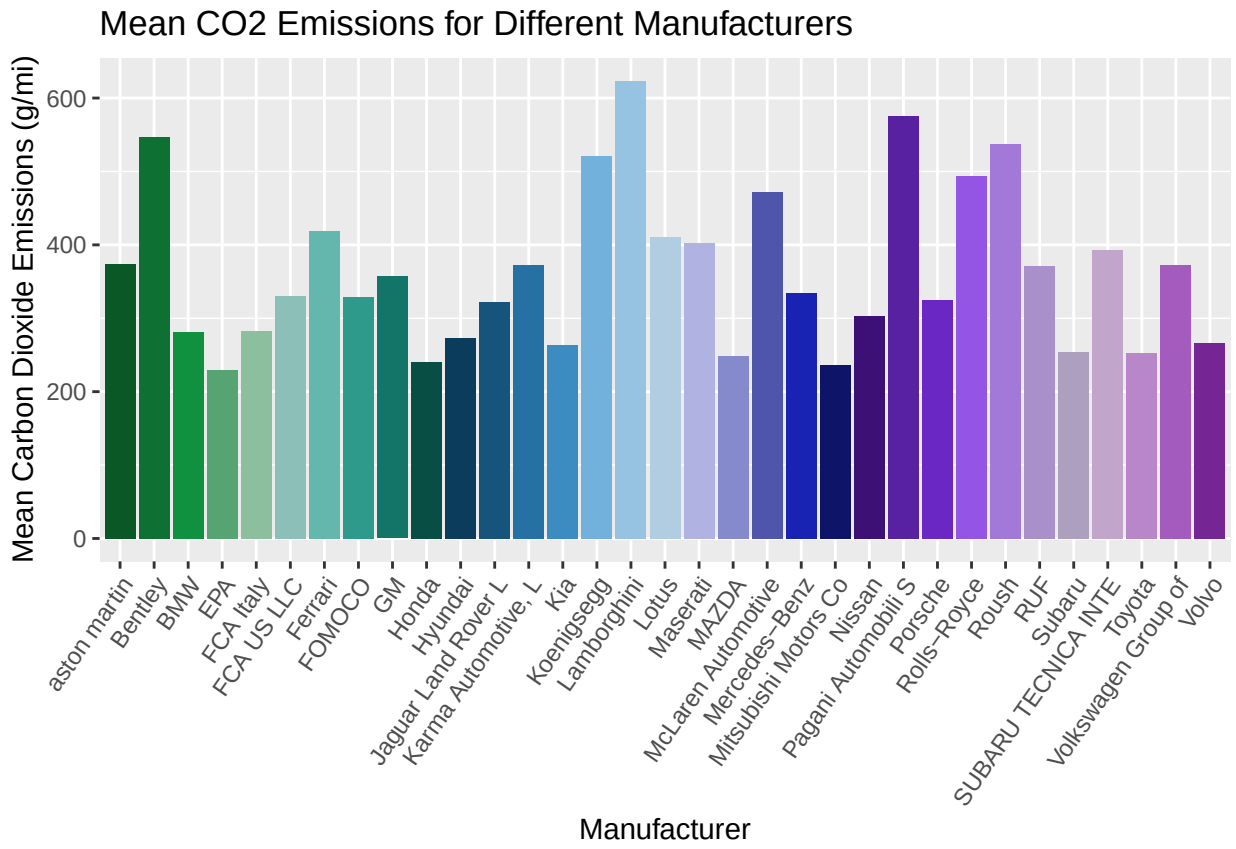
Which manufacturer produces the most carbon dioxide emissions in this data set?

```r
# Calculate the mean CO2 emissions for each fuel type
means_manufacturer <- gas %>%
  group_by(Vehicle.Manufacturer.Name) %>%
  summarise_at(vars(CO2..g.mi.), list(name = mean))
colnames(means_manufacturer) <- c("Manufacturer", "Mean CO2 Emissions")
```

```r
# Print means ordered by mean
print(means_manufacturer[order(means_manufacturer$`Mean CO2 Emissions`,
                               decreasing = TRUE),])
```

```
## # A tibble: 33 x 2
##    Manufacturer        `Mean CO2 Emissions`
##    <chr>                              <dbl>
##  1 Lamborghini                         624.
##  2 Pagani Automobili S                 575.
##  3 Bentley                             547.
##  4 Roush                               538.
##  5 Koenigsegg                          521.
##  6 Rolls-Royce                         494.
##  7 McLaren Automotive                  472.
##  8 Ferrari                             419.
##  9 Lotus                               411.
## 10 Maserati                            403.
## # ... with 23 more rows
```

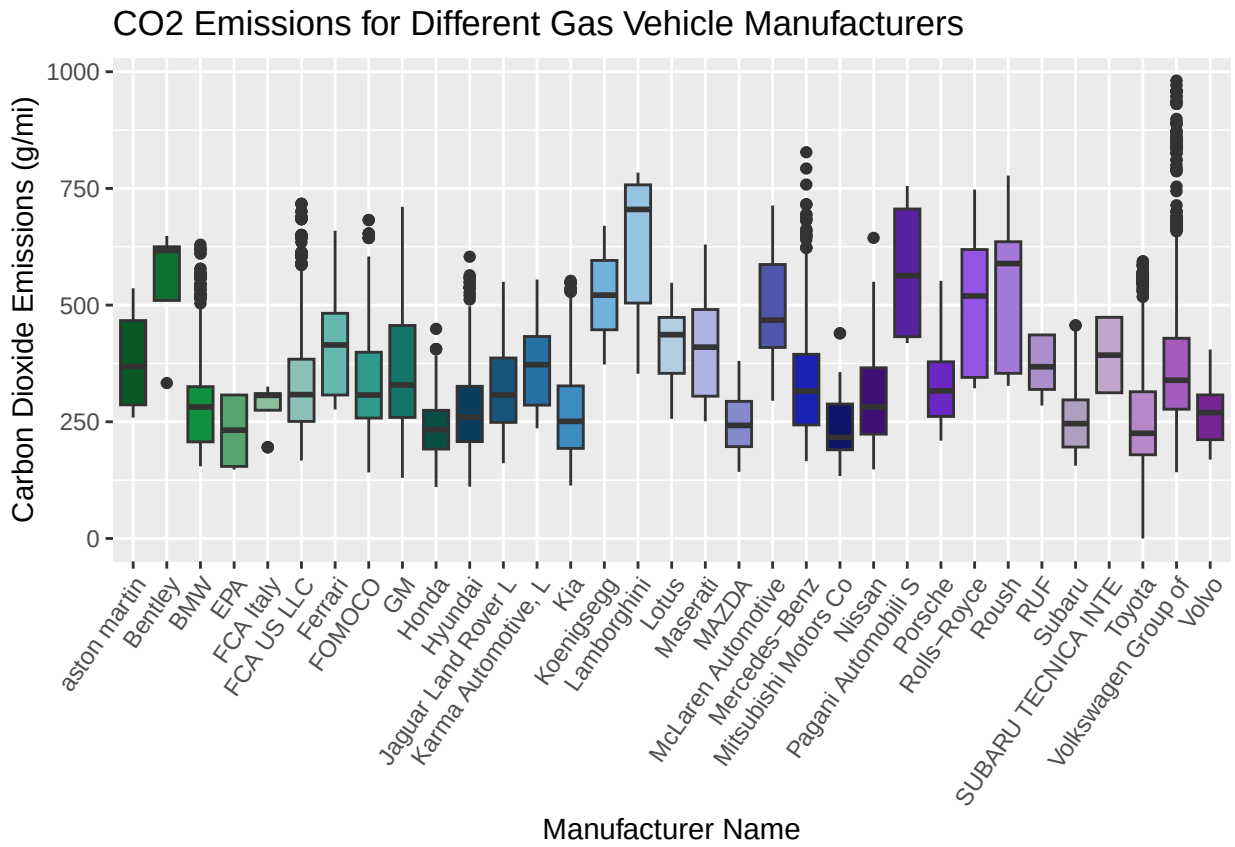
```r
# Plot a barplot of the means
means_manufacturer %>% ggplot(aes(x = Manufacturer,
                                  y = `Mean CO2 Emissions`,
                                  fill = Manufacturer)) +
  geom_bar(stat = "identity", show.legend = FALSE) +
  ggtitle("Mean CO2 Emissions for Different Manufacturers") +
  xlab("Manufacturer") +
  ylab("Mean Carbon Dioxide Emissions (g/mi)") +
  theme(axis.text.x = element_text(angle = 55,
                                   vjust = 1,
                                   hjust=1)) +
  scale_fill_manual(values = GrBuPuPi)
```



From the barplot and table above, the three manufacturers with the highest mean carbon dioxide emission in the data set are Lamborghini, Pagani Automobili S, and Bentley. The three manufacturers with the lowest mean carbon dioxide emission are Honda, Mitsubishi Motors Co, and EPA.

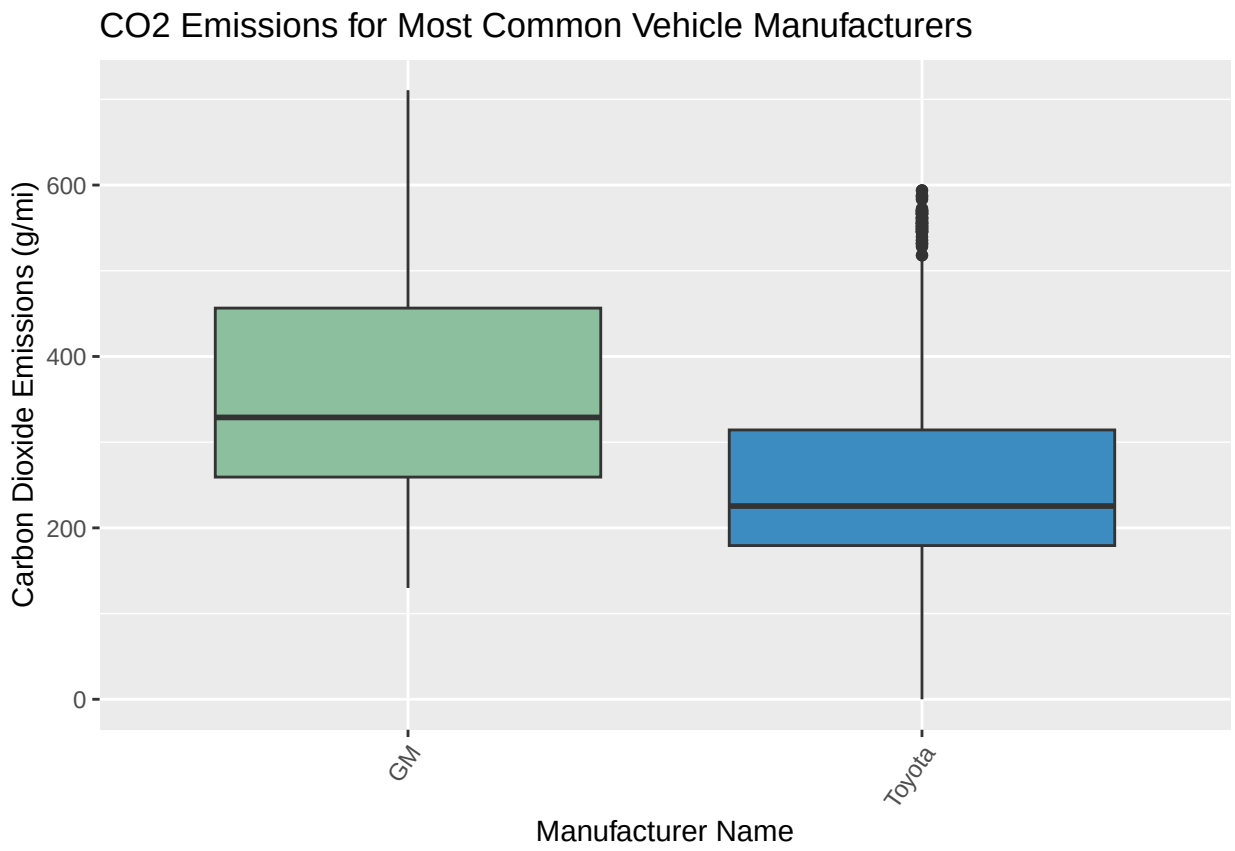
Below, we will plot the boxplots of carbon dioxide emissions for each manufacturer to view the distributions and outliers.

```
gas %>% ggplot(aes(x = Vehicle.Manufacturer.Name,
                   y = CO2..g.mi.,
                   fill = Vehicle.Manufacturer.Name)) +
  geom_boxplot(show.legend = FALSE) +
  ggtitle("CO2 Emissions for Different Gas Vehicle Manufacturers") +
  xlab("Manufacturer Name") +
  ylab("Carbon Dioxide Emissions (g/mi)") +
  theme(axis.text.x = element_text(angle = 55,
                                   vjust = 1,
                                   hjust=1)) +
  scale_fill_manual(values = GrBuPuPi)
```



From the boxplots above, we can see that the mean carbon dioxide emissions varies greatly between manufacturers. It is clear that some manufacturers' mean carbon dioxide emissions differ more significantly than others. FCA US LLC, Mercedes-Benz, and Volkswagen Group contain outliers with higher carbon dioxide emissions. Let's look closer at the boxplots of just the top two manufacturers: GM and Toyota.

```
top2manufacturers <- gas[gas$Vehicle.Manufacturer.Name %in% c("GM", "Toyota"), ]
top2manufacturers %>% ggplot(aes(x = Vehicle.Manufacturer.Name,
                                 y = CO2..g.mi.,
                                 fill = Vehicle.Manufacturer.Name)) +
  geom_boxplot(show.legend = FALSE) +
  ggtitle("CO2 Emissions for Most Common Vehicle Manufacturers") +
  xlab("Manufacturer Name") +
  ylab("Carbon Dioxide Emissions (g/mi)") +
  theme(axis.text.x = element_text(angle = 55,
                                   vjust = 1,
                                   hjust=1)) +
  scale_fill_manual(values = GrBuPuPi[c(5, 14, 20)])
```



Test 2

Let us compare the mean emissions between the two most common manufacturers in the data set: GM and Toyota. From the boxplot above, it appears that GM's mean carbon dioxide emission is higher than Toyota's, so we will test to see if this difference is significant below. We will define the following null and alternative hypotheses:

Declaring Hypotheses and Significance Level

H_0 : The mean carbon dioxide emissions is the same for GM and Toyota gasoline vehicles.

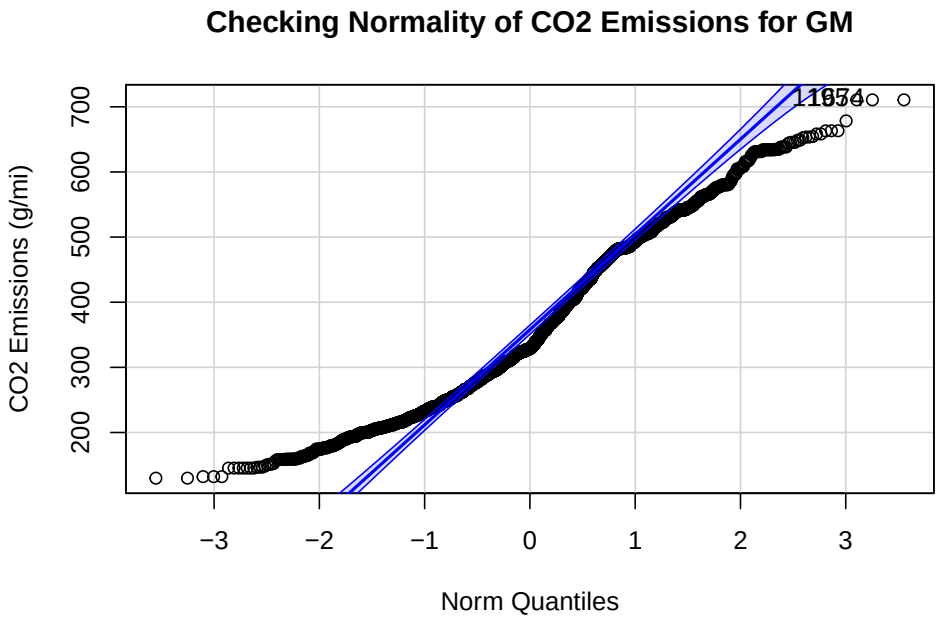
H_a : The mean carbon dioxide emissions is greater for GM gasoline vehicles than Toyota gasoline vehicles.

Significance Level: 1%

Checking Assumptions

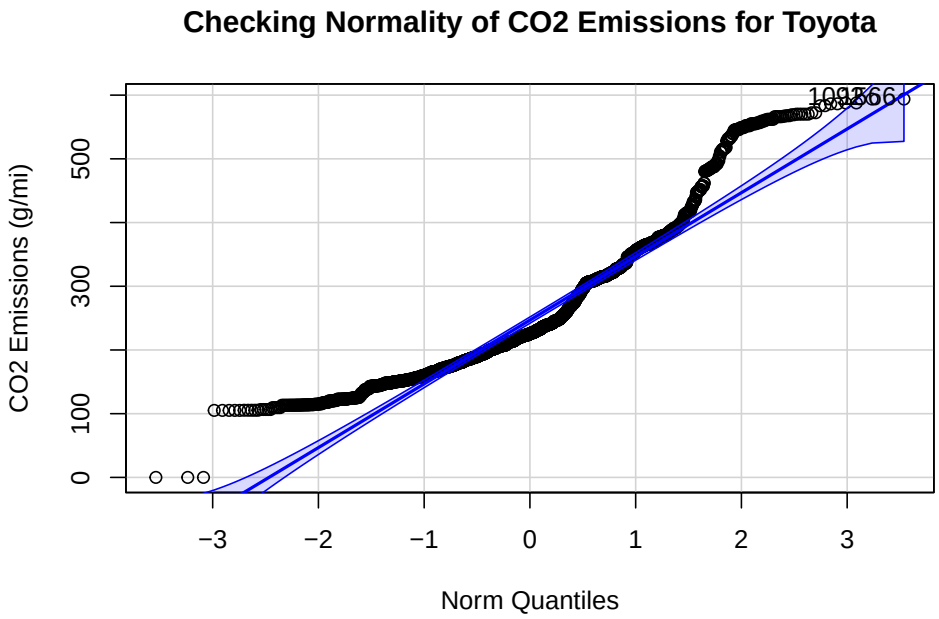
```
# Separate into two data frames filtered by each type
GM <- gas %>% filter(Vehicle.Manufacturer.Name == "GM")
Toyota <- gas %>% filter(Vehicle.Manufacturer.Name == "Toyota")
```

```r
# Population 1: GM
qqPlot(GM$`CO2..g.mi.`,
       main = "Checking Normality of CO2 Emissions for GM",
       xlab = "Norm Quantiles",
       ylab = "CO2 Emissions (g/mi)")
```




```
## [1] 1195 1674
```

```r
# Population 2: Toyota
qqPlot(Toyota$`CO2..g.mi.`,
       main = "Checking Normality of CO2 Emissions for Toyota",
       xlab = "Norm Quantiles",
       ylab = "CO2 Emissions (g/mi)")
```




```
## [1] 1092 1566
```

From the QQ-plots above, it is clear that both distributions are *not* normal. Thus, we will move forward with a Mann-Whitney U Test.

Mann-Whitney U Test

```
# Perform test
mw.test2 <- wilcox.test(GM$`CO2..g.mi.`, Toyota$`CO2..g.mi.`,
                        na.rm = TRUE, paired = FALSE,
                        exact = FALSE, conf.int = TRUE)
mw.test2
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  GM$CO2..g.mi. and Toyota$CO2..g.mi.
## W = 4902158, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##   94.60287 106.32853
## sample estimates:
## difference in location
##               100.4969
```

P-Value Analysis: Based on the test results above, the p-value $= 2.2e-16 < 0.01$ which is statistically significant. Thus, we reject the null hypothesis and conclude that the mean carbon dioxide emissions is greater for GM gasoline vehicles than Toyota gasoline vehicles.

Confidence Interval Analysis: From the 95% confidence interval, we can conclude with 95% confidence that GM gasoline vehicles, on average, produce between 99.16565 g/mi and 106.32853 g/mi more CO2 emissions than Toyota gasoline vehicles for the sample of vehicles in the data set.

Exploratory Data Analysis (EDA) for Test 3

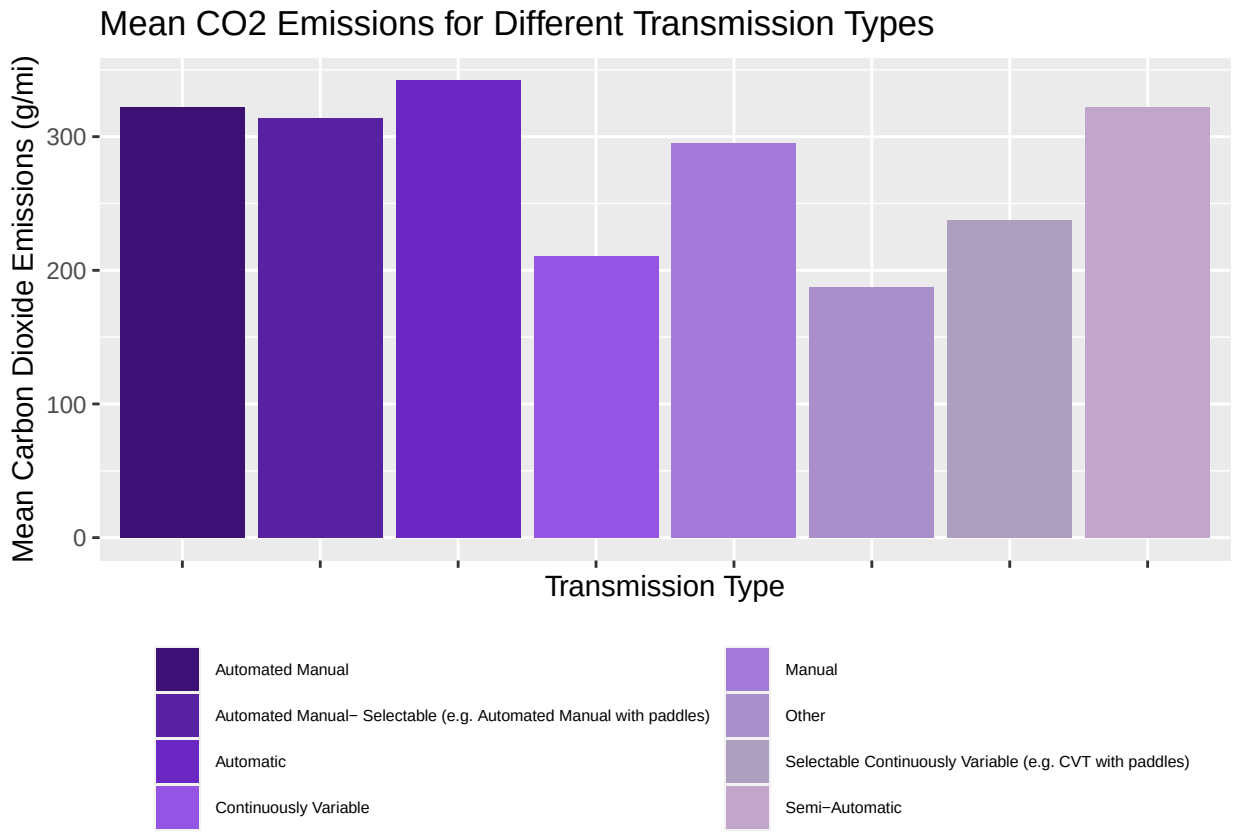
Research Question: Is there a significant difference in the amount of carbon dioxide emissions between vehicle transmission types, specifically manual and automatic vehicles?

Which transmission type produces the most carbon dioxide emissions in this data set?

```
# Calculate the mean CO2 emissions for each fuel type
means_transmissions <- gas %>%
  group_by(Tested.Transmission.Type) %>%
  summarise_at(vars(CO2..g.mi.), list(name = mean))
colnames(means_transmissions) <- c("Transmission Type", "Mean CO2 Emissions")
# Print means ordered by mean
print(means_transmissions[order(means_transmissions$`Mean CO2 Emissions`,
                                decreasing = TRUE),])
```

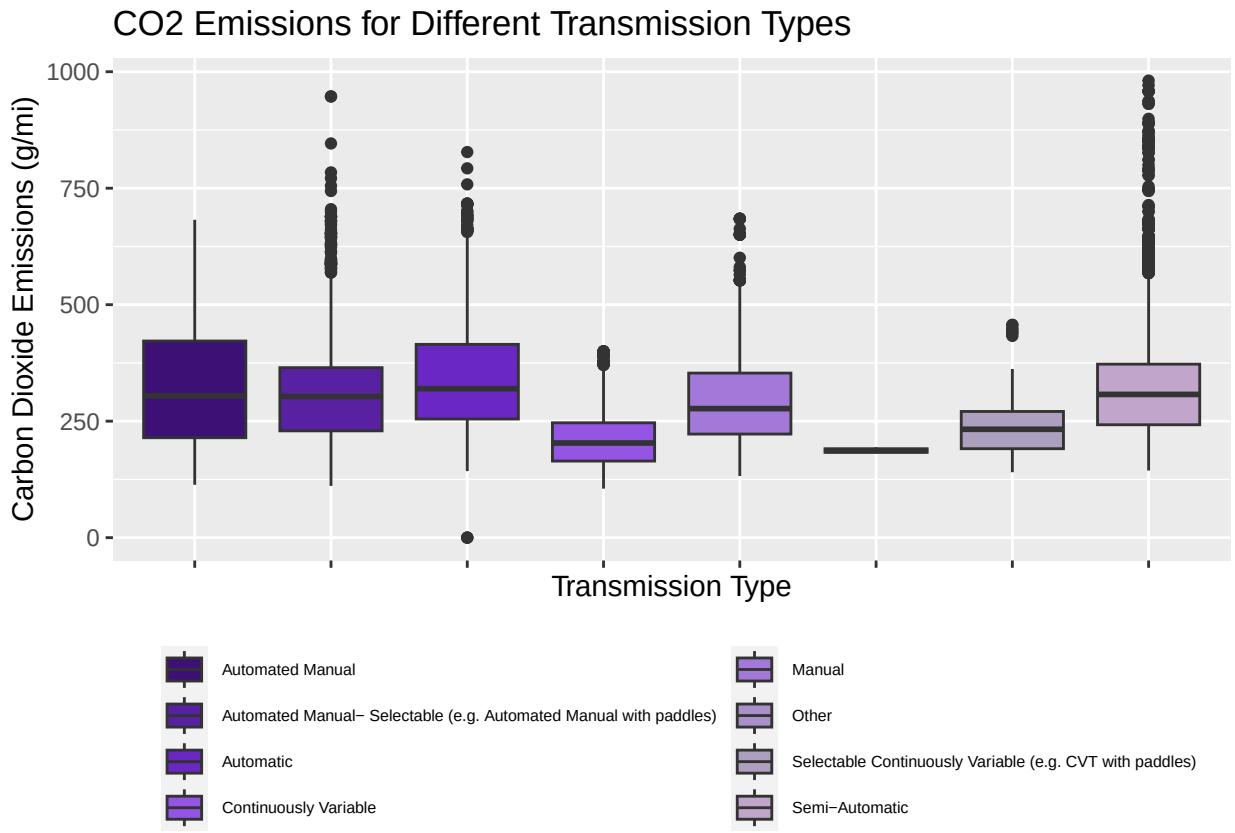
```
## # A tibble: 8 x 2
##   `Transmission Type`                                              Mean CO2 E~1
##   <chr>                                                                   <dbl>
## 1 Automatic                                                               342.
## 2 Automated Manual                                                        322.
## 3 Semi-Automatic                                                          322.
## 4 Automated Manual- Selectable (e.g. Automated Manual with paddles)       314.
## 5 Manual                                                                  295.
## 6 Selectable Continuously Variable (e.g. CVT with paddles)                238.
## 7 Continuously Variable                                                   211.
## 8 Other                                                                   187.
## # ... with abbreviated variable name 1: `Mean CO2 Emissions`
```

```
# Plot a barplot of the means
means_transmissions %>% ggplot(aes(x = `Transmission Type`,
                                   y = `Mean CO2 Emissions`,
                                   fill = `Transmission Type`)) +
  geom_bar(stat = "identity") +
  ggtitle("Mean CO2 Emissions for Different Transmission Types") +
  xlab("Transmission Type") +
  ylab("Mean Carbon Dioxide Emissions (g/mi)") +
  theme(axis.text.x = element_blank(),
        legend.position = "bottom",
        legend.text=element_text(size = 6)) +
  scale_fill_manual(values = GrBuPuPi[c(23,24,25,26,27,28,29,30)],
                    name = NULL) +
  guides(fill=guide_legend(ncol = 2))
```



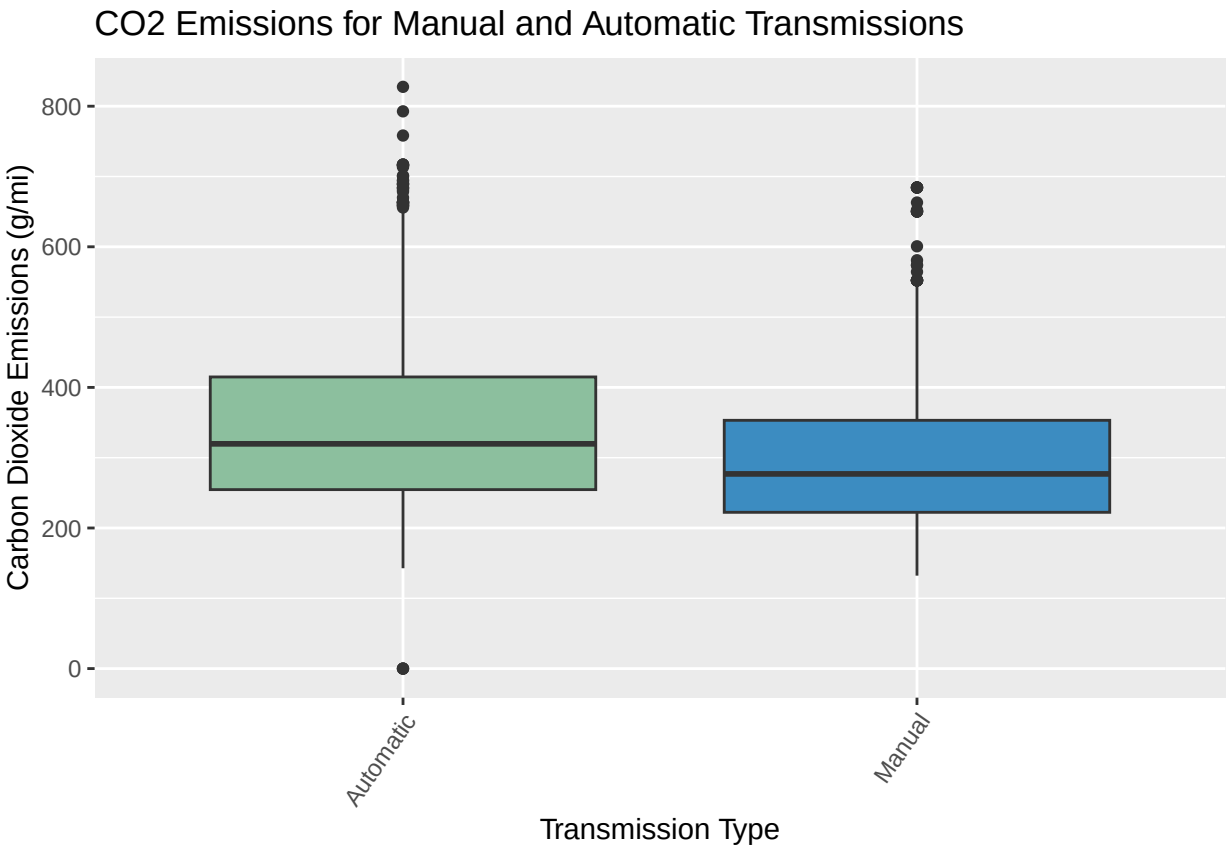
From the barplot and table above, we can see that the three transmission types with the highest mean carbon dioxide emissions are automatic, automated manual, and semi-automatic. The lowest three are selectable continuously variable, continuously variable, and other. Below, we will plot the boxplots of carbon dioxide emissions for each transmission type to view the distributions and outliers.

```
names(gas)[names(gas) == 'Tested.Transmission.Type'] <- 'Transmission Type'
gas %>% ggplot(aes(x = `Transmission Type`,
                   y = CO2..g.mi.,
                   fill = `Transmission Type`)) +
  geom_boxplot() +
  ggtitle("CO2 Emissions for Different Transmission Types") +
  xlab("Transmission Type") +
  ylab("Carbon Dioxide Emissions (g/mi)") +
  theme(axis.text.x = element_blank(),
        legend.position = "bottom",
        legend.text=element_text(size = 6)) +
  scale_fill_manual(values = GrBuPuPi[c(23,24,25,26,27,28,29,30)],
                    name = NULL) +
  guides(fill=guide_legend(ncol = 2))
```



From the boxplots above, we can see that the mean carbon dioxide emissions does not vary as much between transmission types as it did between manufacturer and fuel type. Many also contain several outliers that have higher carbon dioxide emissions. Let's look closer at the boxplots of just the manual and automatic vehicle CO2 emissions distributions.

```
df <- gas[gas$`Transmission Type` %in% c("Manual", "Automatic"), ]
df %>% ggplot(aes(x = `Transmission Type`,
                  y = CO2..g.mi.,
                  fill = `Transmission Type`)) +
  geom_boxplot(show.legend = FALSE) +
  ggtitle("CO2 Emissions for Manual and Automatic Transmissions") +
  xlab("Transmission Type") +
  ylab("Carbon Dioxide Emissions (g/mi)") +
  theme(axis.text.x = element_text(angle = 55,
                                   vjust = 1,
                                   hjust=1)) +
  scale_fill_manual(values = GrBuPuPi[c(5, 14, 20)])
```



From the boxplot above, it is clear that automatic transmissions have the slightly higher mean emissions; the test below will determine if this difference is significant.

T-Test 3

Let us compare the mean emissions between automatic cars and manual vehicles.

Declaring Hypotheses and Significance Level

H_0 : The mean carbon dioxide emissions is the same for automatic and manual vehicles.

H_a : The mean carbon dioxide emissions is greater for automatic vehicles is higher than for manual vehicles.

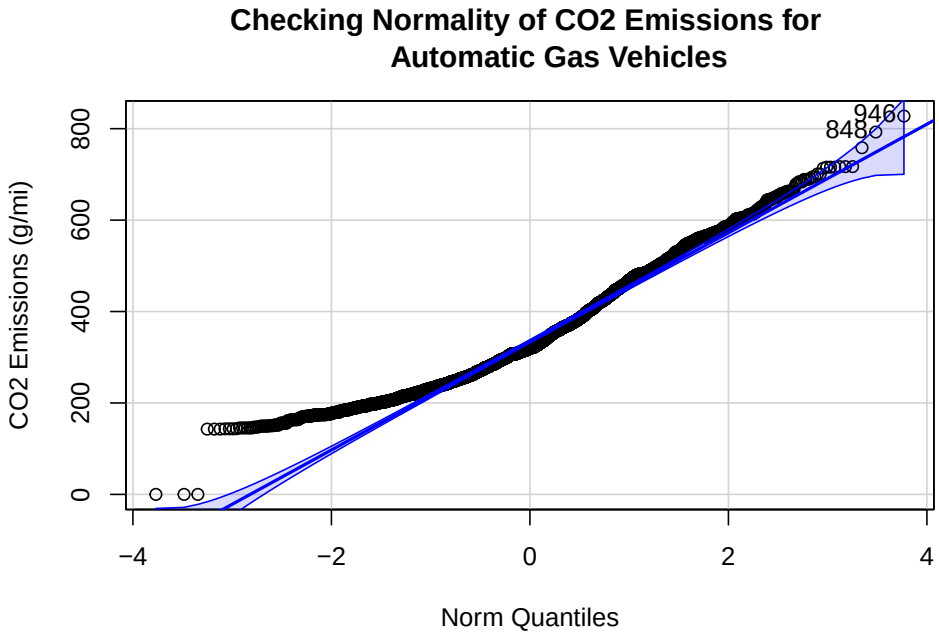
Significance Level: 1%

Checking Assumptions

```
# Separate into two data frames filtered by each type
automatic <- gas %>% filter(`Transmission Type` == "Automatic")
manual <- gas %>% filter(`Transmission Type` == "Manual")
```

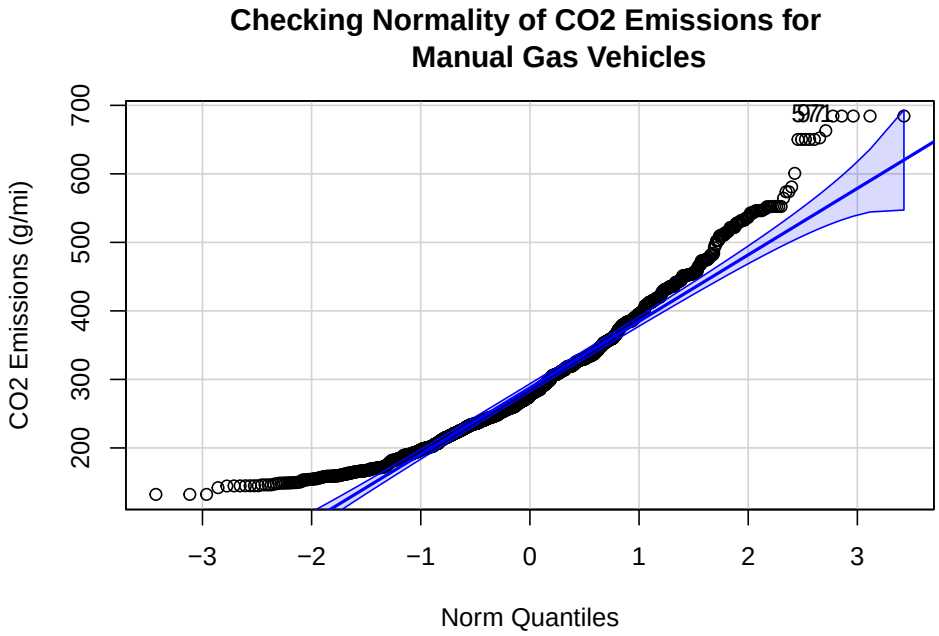
```
# Population 1: automatic
qqPlot(automatic$`CO2..g.mi.`,
       main = "Checking Normality of CO2 Emissions for
```

```
       Automatic Gas Vehicles",
       xlab = "Norm Quantiles",
       ylab = "CO2 Emissions (g/mi)")
```




```
## [1] 946 848
```

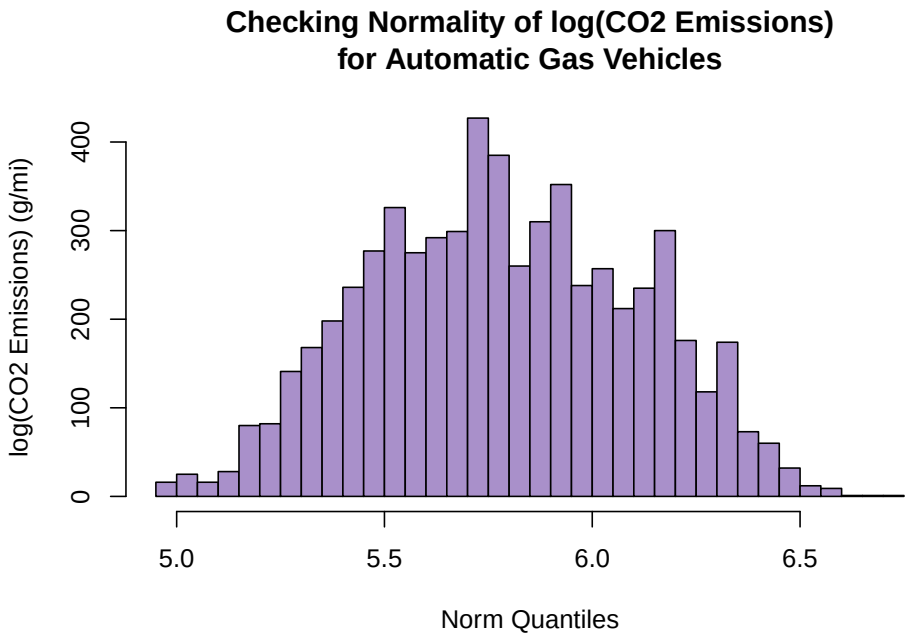
```
# Population 2: manual
qqPlot(manual$`CO2..g.mi.`,
       main = "Checking Normality of CO2 Emissions for
       Manual Gas Vehicles",
       xlab = "Norm Quantiles",
       ylab = "CO2 Emissions (g/mi)")
```




```
## [1]  97 571
```

Both distributions appear to be skewed to the right. Let us check if a log transformation is useful in normalizing the data.

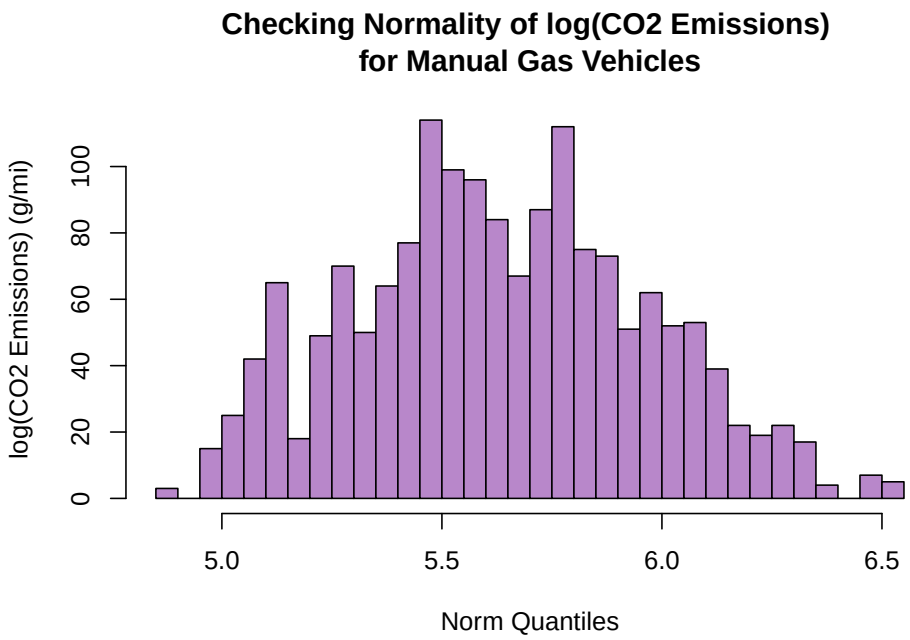
```
# Population 1: automatic
hist(log(automatic$`CO2..g.mi.`),
main = "Checking Normality of log(CO2 Emissions)
for Automatic Gas Vehicles",
xlab = "Norm Quantiles", ylab = "log(CO2 Emissions) (g/mi)",
col = GrBuPuPi[c(28)], breaks = 40)
```




```
data1 <- log(automatic$`CO2..g.mi.`)
data1[!is.finite(data1)] <- NA
shapiro.test(sample(data1, 5000))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sample(data1, 5000)
## W = 0.99112, p-value < 2.2e-16
```

```
# Population 2: manual
hist(log(manual$`CO2..g.mi.`),
main = "Checking Normality of log(CO2 Emissions)
for Manual Gas Vehicles",
xlab = "Norm Quantiles", ylab = "log(CO2 Emissions) (g/mi)",
col = GrBuPuPi[c(31)], breaks = 40)
```




```
data2 <- log(manual$`CO2..g.mi.`)
data2[!is.finite(data2)] <- NA
shapiro.test(data2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data2
## W = 0.99129, p-value = 2.697e-08
```

While the histograms show some improvement in normality from the log transformation, the Shapiro tests with very small p-values assert that the data still does not follow a normal distribution. Thus, we must move forward with a Mann-Whitney U Test.

Mann-Whitney U Test

```
# Perform test
mw.test3 <- wilcox.test(automatic$`CO2..g.mi.`, manual$`CO2..g.mi.`,
                        na.rm = TRUE, paired = FALSE,
                        exact = FALSE, conf.int = TRUE)
mw.test3
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  automatic$CO2..g.mi. and manual$CO2..g.mi.
## W = 6236277, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

```
## 95 percent confidence interval:
##  38.75607 49.62847
## sample estimates:
## difference in location
##               44.17741
```

P-Value Analysis: Based on the test results above, the p-value = $2.2e-16 < 0.01$ which is statistically significant. Thus, we reject the null hypothesis and conclude that the mean carbon dioxide emissions is greater for automatic gasoline vehicles than manual gasoline vehicles.

Confidence Interval Analysis: From the 95% confidence interval, we can conclude with 95% confidence that automatic gasoline vehicles, on average, produce between 38.75607 g/mi and 49.62847 g/mi more CO2 emissions than manual gasoline vehicles for the sample of vehicles in the data set.

This result shows that manual vehicles are more fuel efficient. This makes sense as manual vehicles are typically lighter and have a less complex engine set up.

References

1. R Color Codes:

https://www.rapidtables.com/web/color/RGB_Color.html

2. Barplots:

http://www.sthda.com/english/wiki/ggplot2-barplots-quick-start-guide-r-software-and-data-visualization

3. Boxplots:

http://www.sthda.com/english/wiki/ggplot2-box-plot-quick-start-guide-r-software-and-data-visualization

4. Legend Customization:

http://www.sthda.com/english/wiki/ggplot2-legend-easy-steps-to-change-the-position-and-the-appearance-of-a-graph-legend-in-r-software

5. ANLY 511 Lecture 10 Slides

6. QQ-Plot Documentation:

https://braverock.com/brian/R/PerformanceAnalytics/html/chart.QQPlot.html