Final Project For Applied Statistics

Ben Cartwright

12/9/2021

#Question Prompt Fuel efficiency. Pick a data set from https://www.fueleconomy.gov/feg/download.shtml. Think about a question about fuel efficiency you would like to answer. One option is to look at whether there is a significant difference between the fuel efficiency of cars with automatic transmission and those with manual transmission, e.g. Visualize the data. Clean up your data set so that you retain just the information that you need. Conduct a hypothesis test choosing a significance level you find appropriate. What are your findings?

#Research Focus For my final project for M358 K, I will explore 2022 fuel economy data to determine if there is any statistically significant difference in the fuel efficiency of manual and automatic transmissions.

#Step 1: The first step in this project was to clean-up the data, removing any unnecessary information and creating a variable, "Trans", which indicates whether or not the transmission is manual or automatic

```
library(ggplot2)
messy_data <-
  read.csv("2022 FE Guide-release dates before 12-2-2021-no-sales -12-1-2021 for DOEpublic.csv")
data <- subset(messy_data,select =
                 c(Comb.FE..Guide....Conventional.Fuel, MFR.Calculated.Gas.Guzzler.MPG,Transmission))
Trans = c()
for( row in 1:nrow(data)){
  if ((data[row, "Transmission"] == "Manual(M7)")|(data[row, "Transmission"] == "Manual(M6)")
      |(data[row, "Transmission"] == "Manual(M5)")){
    Trans = append(Trans, "Manual")
  }
  else{
    Trans = append(Trans, "Automatic")
  }
}
names(data)[names(data) == "Comb.FE..Guide....Conventional.Fuel"] <- "mpg"
data = cbind(data, Trans)
head(data)
```

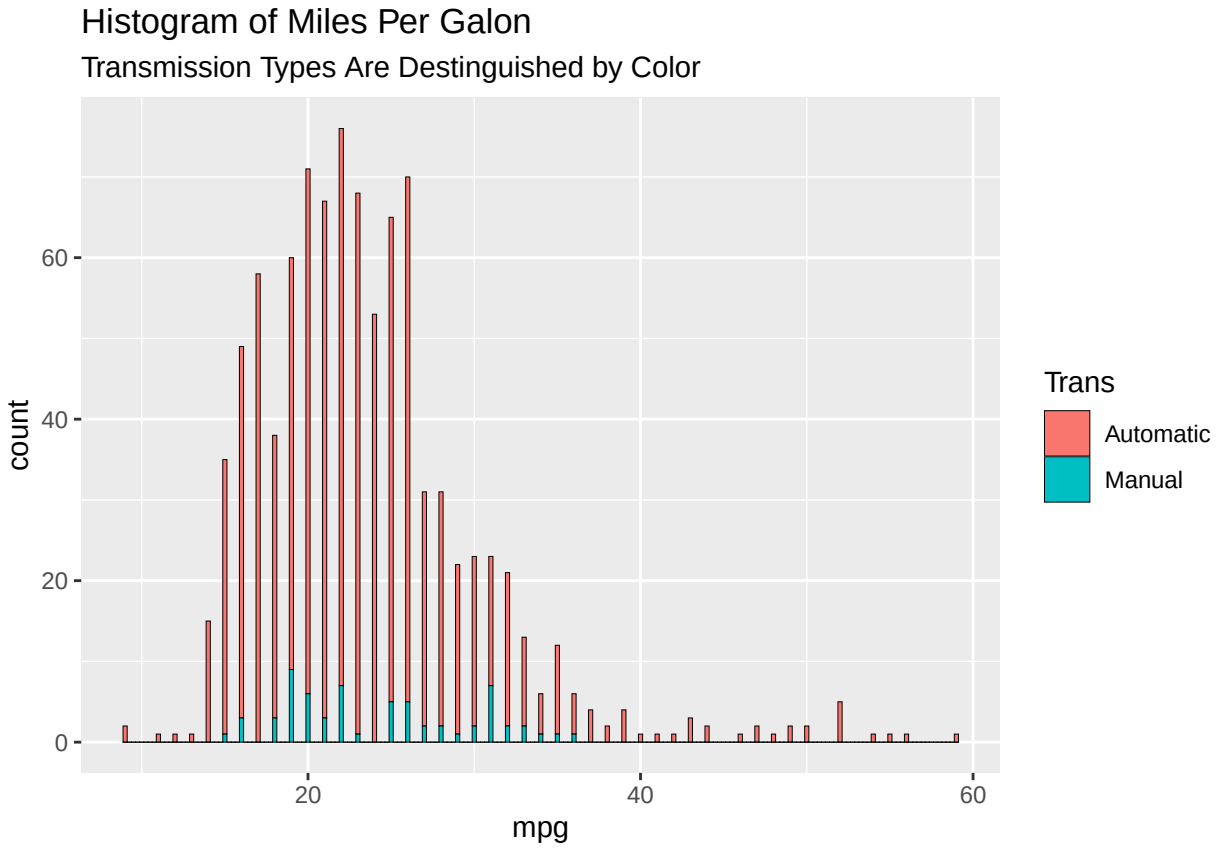
```
##   mpg MFR.Calculated.Gas.Guzzler.MPG Transmission     Trans
## 1  15                           19.1  Auto(AM-S7) Automatic
## 2  17                           20.7  Auto(AM-S7) Automatic
## 3  15                           19.1  Auto(AM-S7) Automatic
## 4  17                           20.7  Auto(AM-S7) Automatic
## 5  26                           33.5  Auto(AM-S7) Automatic
## 6  25                           32.9     Auto(S8) Automatic
```

#Step 2: Next, I created a histogram based on the mpg each car uses, and distinguished Automatic transmissions in red and manual transmissions in blue. Notice, at a first glance it we see that there are many

more automatic cars than manual cars being evaluated, and that most of the manual cars in question tend towards the center of the histogram.

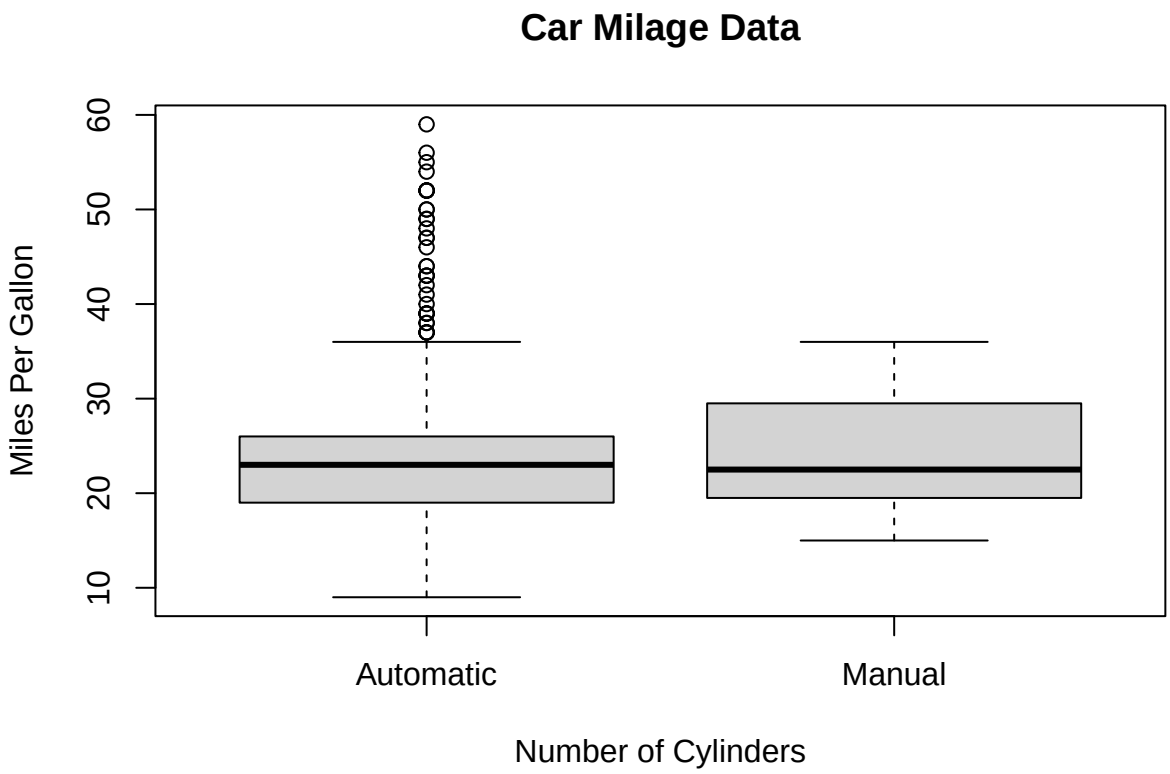
```
g <- ggplot(data, aes(mpg))

g + geom_histogram(aes(fill=Trans),
                   binwidth = .25,
                   col="black",
                   size=.1) +
  labs(title="Histogram of Miles Per Galon",
       subtitle="Transmission Types Are Destinguished by Color")
```

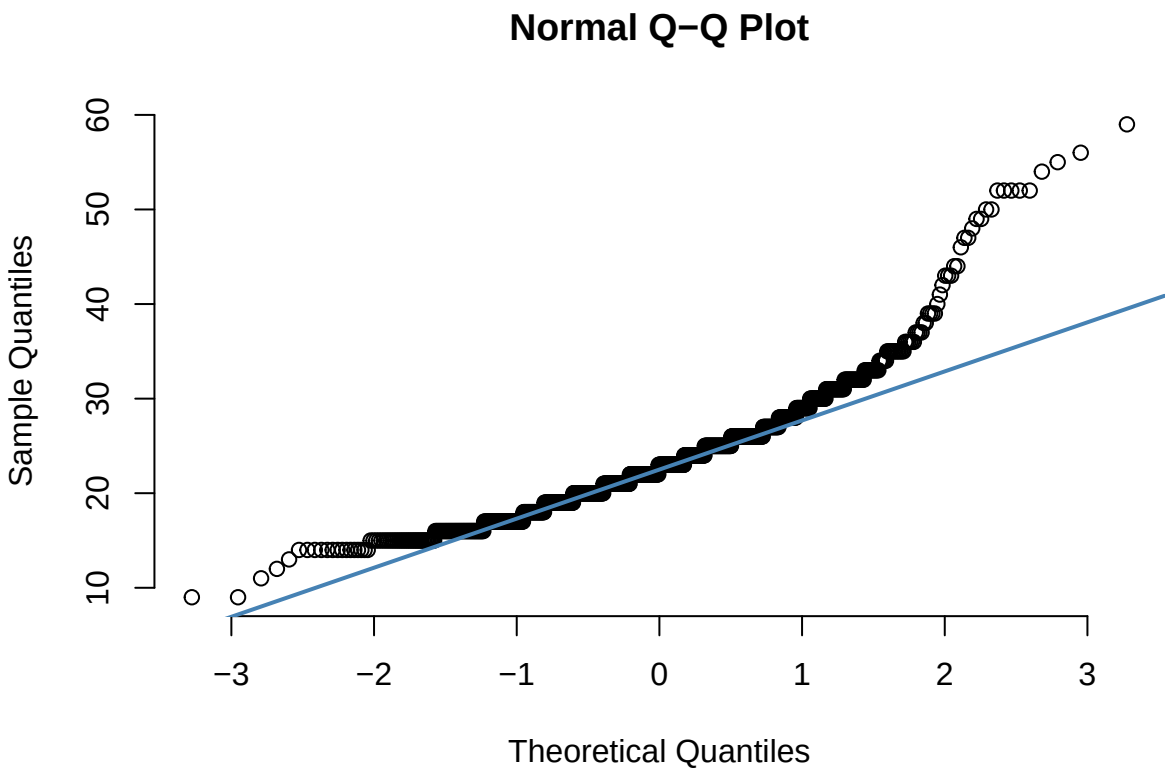


#Step 3: Now I check to see if the data is approximately normal in order to conduct hypothesis tests.

```
#With Outliers
#boxplot
boxplot(mpg~Trans,data=data, main="Car Milage Data",
   xlab="Number of Cylinders", ylab="Miles Per Gallon")
```

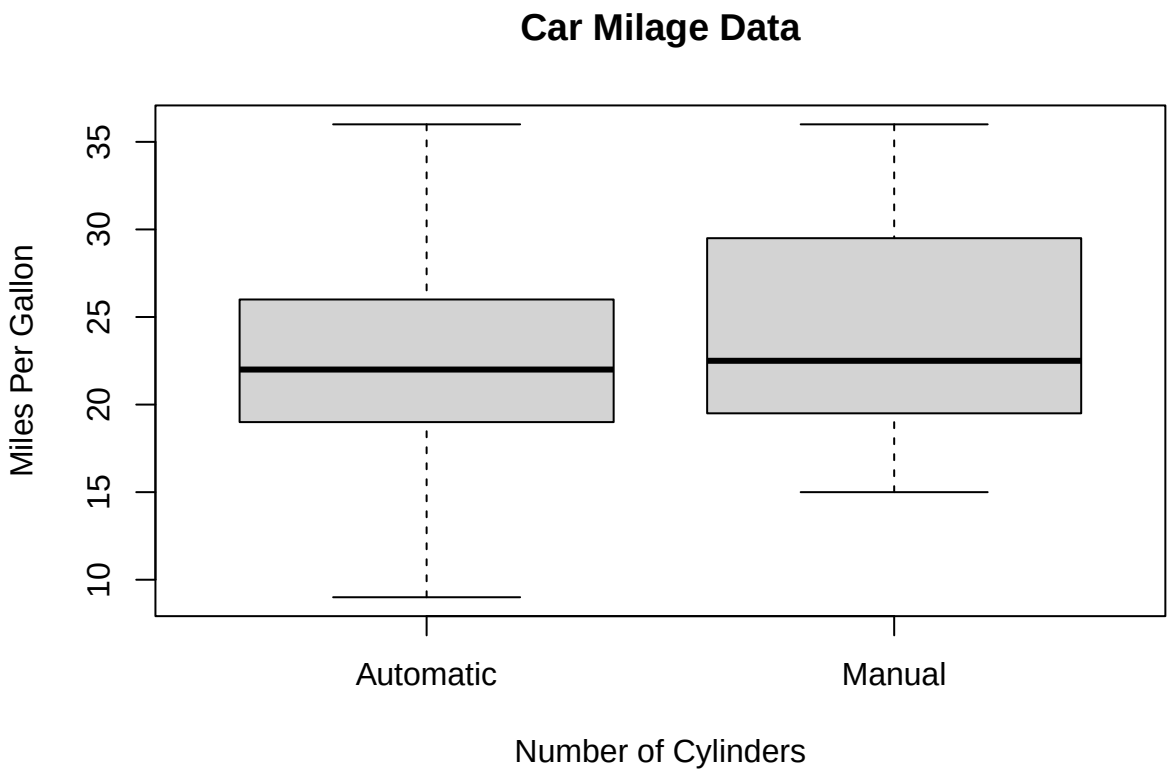



```
#qqplot
qqnorm(data$mpg, pch = 1, frame = FALSE)
qqline(data$mpg, col = "steelblue", lwd = 2)
```

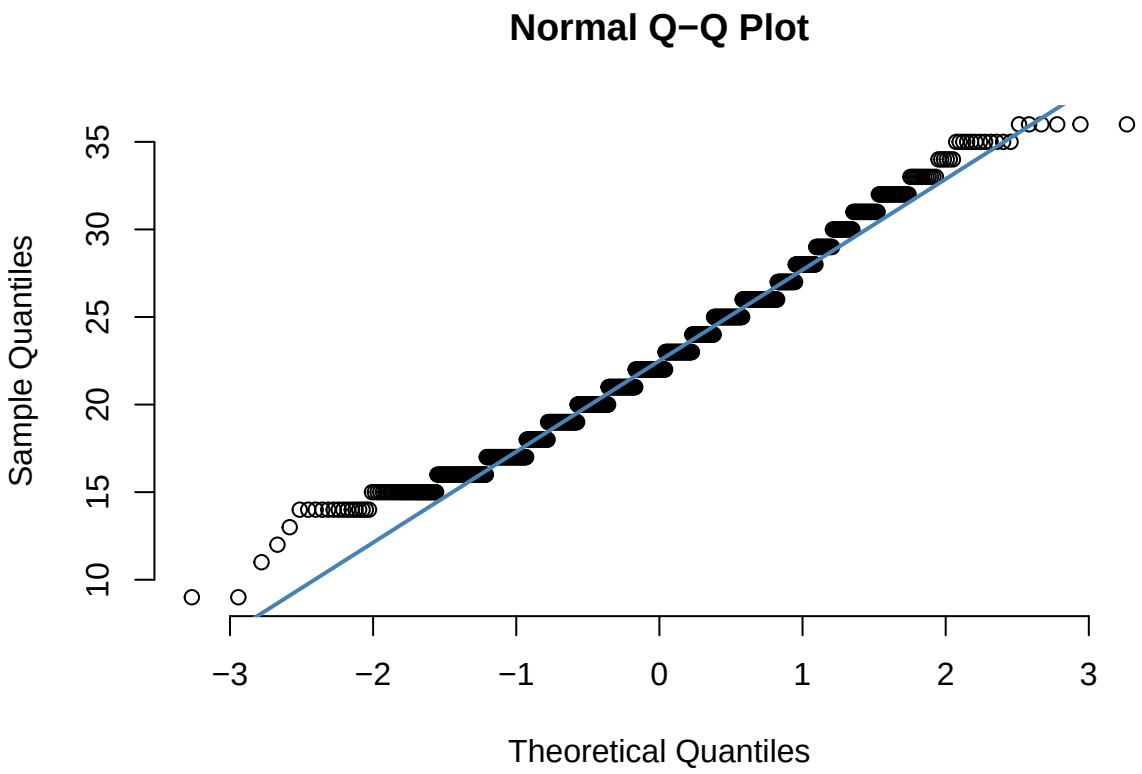



```
#remove outliers
out = boxplot.stats(data$mpg)$out
out_data_ind = which(data$mpg %in% c(out))
data_no_outs = data[-out_data_ind,]
data_no_outs = data.frame(data_no_outs)

#Without Outliers
#boxplot
boxplot(mpg~Trans,data=data_no_outs, main="Car Milage Data",
   xlab="Number of Cylinders", ylab="Miles Per Gallon")
```




```
#qqplot
qqnorm(data_no_outs$mpg, pch = 1, frame = FALSE)
qqline(data_no_outs$mpg, col = "steelblue", lwd = 2)
```

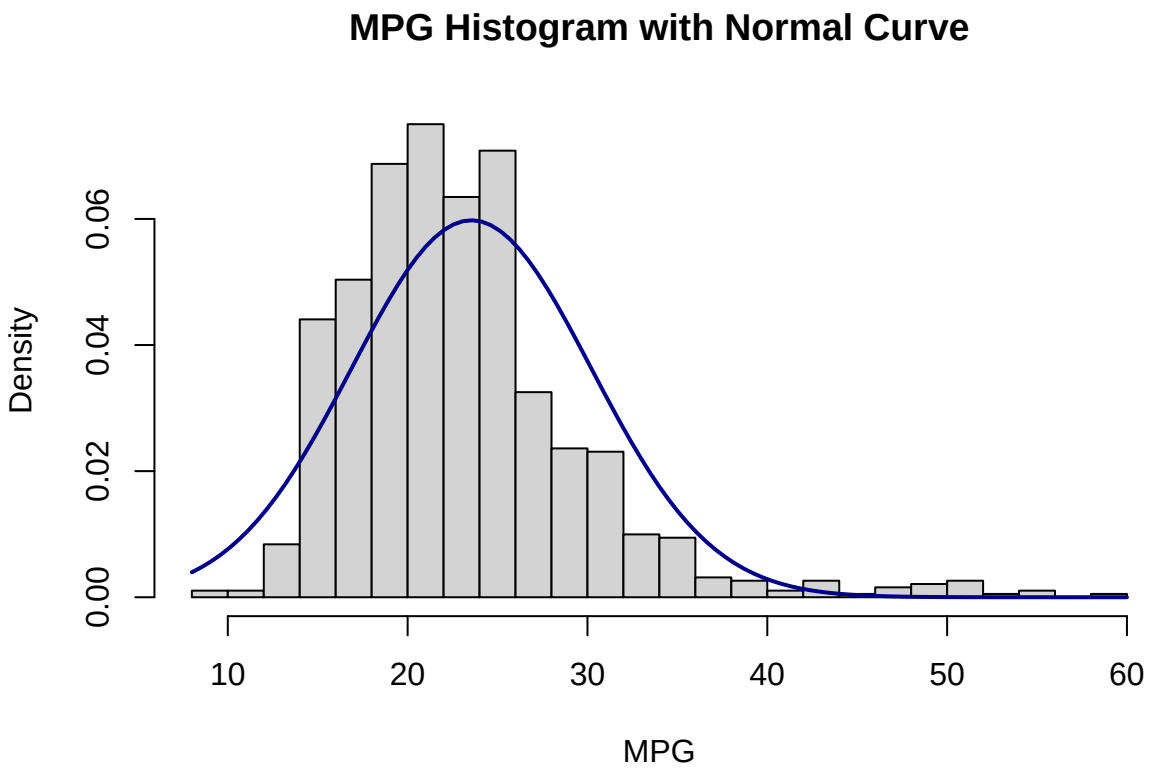


#Step 3 Conclusion: The boxplot and qqplot illustrate that there are outliers within the automatic transmission sub-population of the data which drastically impact the normality of the distribution. In terms of manual vs. automatic transmissions, the bax-plot of the data without outliers illustrates that these outliers bring the two medians closer together, and, without outliers, the two categories exhibit similar Q3 values while automatic transmissions have a much lower Q1 value. This suggests that the two categories have similar average mpg rates, but that cars with automatic transmissions have span a wider range of mpg values, including lower values than manual cars experience.

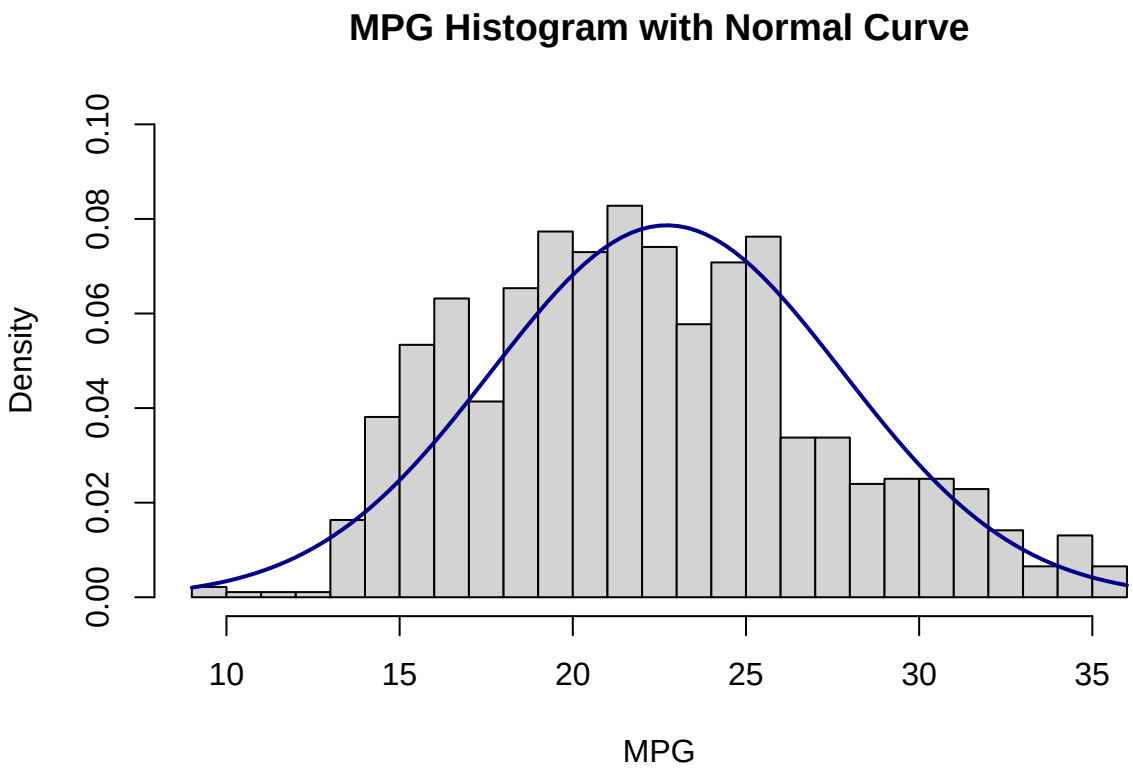
The following graphs superimpose the normal distribution onto the histogram of mpg and illustrate that the data is more normal without outliers.

```
m<-mean(data$mpg)
std<-sqrt(var(data$mpg))

hist(data$mpg, breaks=20, prob=TRUE,
     xlab="MPG", ylim=c(0, 0.075),
     main="MPG Histogram with Normal Curve")
curve(dnorm(x, mean=m, sd=std),
      col="darkblue", lwd=2, add=TRUE, yaxt="n")
```




```
m2<-mean(data_no_outs$mpg)
std2<-sqrt(var(data_no_outs$mpg))
hist(data_no_outs$mpg, breaks=20, prob=TRUE,
     xlab="MPG", ylim=c(0, 0.1),
     main="MPG Histogram with Normal Curve")
curve(dnorm(x, mean=m2, sd=std2),
      col="darkblue", lwd=2, add=TRUE, yaxt="n")
```



The following 2 tables illustrate how the means, grouped by transmission type, change with and without outliers.

```
aggregate(data$mpg, list(data$Trans), FUN=mean)
```

```
##     Group.1        x
## 1 Automatic 23.49381
## 2    Manual 24.28125
```

```
aggregate(data_no_outs$mpg, list(data_no_outs$Trans), FUN=mean)
```

```
##     Group.1        x
## 1 Automatic 22.59485
## 2    Manual 24.28125
```

#Note: At this point it is clear that outliers within the automatic transmission skew the data to the right, and it is important to note that there may be some common attribute to these outliers which explains the difference in there mpg.

```
#Hypothesis Test Conducted at alpha = 0.05 for data containing outliers, u1 is population mean for manu
#Null Hypothesis: u1 - u2 = 0
#Alternative Hypotheis: u1 - u2 > 0
manual_data = subset(data, Trans == "Manual")
auto_data = subset(data, Trans == "Automatic")
```

```r
x1 =  mean(manual_data$mpg)
x2 = mean(auto_data$mpg)
n1 = length(manual_data$mpg)
n2 = length(auto_data$mpg)
s1 = sd(manual_data$mpg)
s2 = sd(auto_data$mpg)
se = sqrt(s1^2/n1 + s2^2/n2)
t.value = (x1-x2)/se
p.value = 1 - pt(t.value, n1 - 1)
print(p.value)
```

```
## [1] 0.1413492
```

```r
#Hypothesis Test Conducted at alpha = 0.05 for data not containing outliers, u1 is population mean for
#Null Hypothesis: u1 - u2 = 0
#Alternative Hypotheis: u1 - u2 > 0
manual_data_out = subset(data_no_outs, Trans == "Manual")
auto_data_out = subset(data_no_outs, Trans == "Automatic")
x1 =  mean(manual_data_out$mpg)
x2 = mean(auto_data_out$mpg)
n1 = length(manual_data_out$mpg)
n2 = length(auto_data_out$mpg)
s1 = sd(manual_data_out$mpg)
s2 = sd(auto_data_out$mpg)
se = sqrt(s1^2/n1 + s2^2/n2)
t.value = (x1-x2)/se
p.value = 1 - pt(t.value, n1 - 1)
print(p.value)
```

```
## [1] 0.0104381
```

#Conclusion:

The hypothesis test conducted on the data containing outliers doesn't yield significant evidence against the null hypothesis, stating manual and automatic transmissions have similar fuel efficiency, because the observed p value is 0.1413492, above out 0,05 signifigance level.

However, if we remove outliers from the data, particularly vehicles with manual transmissions and fuel efficiency $>$ ~37mpg, our hypothesis test yields significant results. Specifically, we observe a p value of 0.0104381, below our significance level of 0.05, providing strong evidence against the null hypothesis in favor of manual transmissions being more fuel efficient than automatic transmissions.

These results seem to suggest that a portion of vehicles, with very high gas mpg rates, skew the data to the right. A further study could be done to examine if these vehicles have any similar attributes, say they are all hybrid engines or equipped with cruise control, which inherently make them have higher mpgs. If we can attribute some explanation to these outliers, than we can concllude that, amoung vehicles lacking this attribute, manual transmissions are more fuel effecient.