Homework 2 - STAT 405

Frank Portman

January 24, 2013

1 Introduction

For this assignment, I explored some of the mpg2 dataset. It contains over 30,000 observations of various car data including manufacturer, year, model, highway MPG, and much more.

I decided to focus my analysis on a small portion of the data in order to thoroughly investigate it. Specifically, I considered the effect of transmission type and engine charge on fuel economy. The two types of engine transmissions I worked with are 'automatic' and 'manual' - I grouped all transmission types to fall in those two categories. The engine types I considered were: 'Naturally Aspirated', 'Supercharged', and 'Turbocharged'. Supercharged and turbocharged engines use different mechanisms to increase the density of air supplied to the engine. Since this enables the car to go faster it was intuitive to assume these two traits could have a negative effect on fuel economy.

After some initial observations regarding fuel economy, transmission type, and engine type, I deepened my exploration of transmission types to explain some of the contrasts I encountered.

2 Plots and Analysis

To explore the relationship between engine charge type, transmission, and highway MPG, I had to create a few new variables in the data and clean out NA terms. I first searched for all instances of the words 'Auto' and 'Man' in the 'tran' variable of the data. Then, I created a new variable named 'tranClean' which simply recorded whether the transmission was manual or automatic. This took care of cleaning different spellings of the same transmission types and grouped together same transmissions with different amounts of gears.

```
autoTrans <- grep("Auto", mpg2$tran)
manTrans <- grep("Man", mpg2$tran)

mpg2.transmissions <- mpg2
mpg2.transmissions$tranClean <- "filler"

mpg2.transmissions$tranClean[autoTrans] <- "Automatic"
mpg2.transmissions$tranClean[manTrans] <- "Manual"
tranFillers <- which(mpg2.transmissions$tranClean == "filler")
mpg2.transmissions <- mpg2.transmissions[-tranFillers, ]
```

With this cleaned data, I was able to produce my first plot:

```
qplot(hwy, ..density.., geom = "freqpoly",
      binwidth = 3, color = charger, data = mpg2.transmissions) +
      xlim(10, 50) + facet_wrap(~ tranClean) + xlab("Highway MPG") +
      ggtitle("Density of Highway MPG by Engine Type")
```

Figure 1 indicates that cars with manual transmission have a higher density of their cars on the upper end of the Highway MPG range, as compared to automatic cars. Within each transmission type, supercharged engines are more densely populated on the lower end of the Highway MPG spectrum. Turbocharged engines are very densely populated around the mid 20 MPG range for both transmission types, while naturally aspirated engines have a smoother distribution - most of the very fuel efficient cars (Highway MPG ≥ 30) are naturally aspirated.

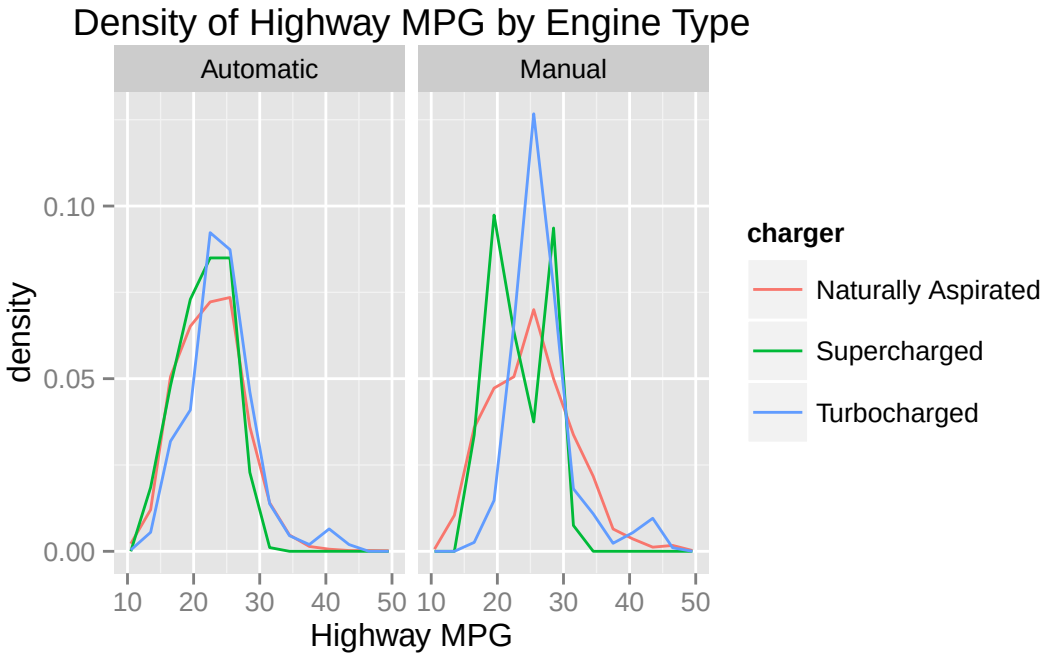


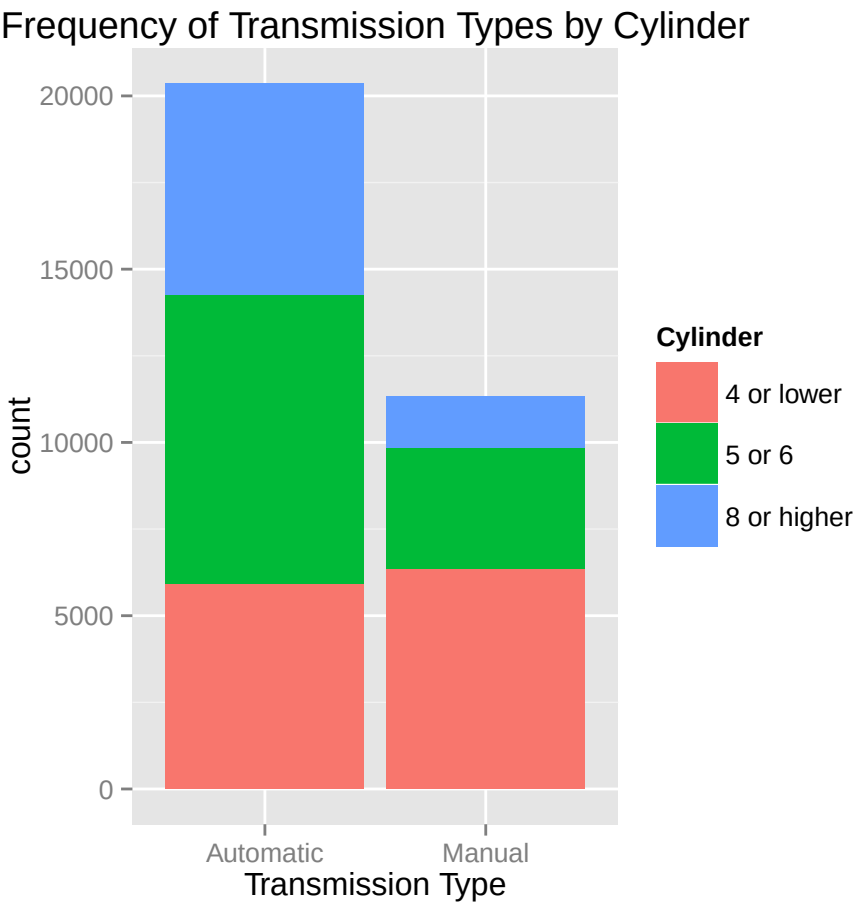
Figure 1: Above: A density plot outlines the relationship between transmission type, fuel economy, and engine charge.

It was interesting to note that automatic cars all had a very similar distribution of fuel economy across all engine types, while manual cars, albeit more fuel efficient, did not share the homogeneous results of their non-stick shift counterparts. One reason for this phenomenon that came to mind right away was simply the difference in sample sizes of automatic cars versus manual cars. There are many more automatic cars in this dataset, therefore some of the jaggedness in the portion of the graph referring to manual transmissions can be attributed to the fewer observations.

In order to explain some of the reasons why manual cars tended to be more fuel efficient, I delved deeper into the characteristics of manual cars. I considered the possibility that cylinder size would affect fuel economy and that manual cars had more fuel efficient cylinder sizes. To see if this was the case I had to clean the 'cyl' column a bit and then plot a simple histogram detailing the frequency of transmission types by cylinder.

```
naCyl <- which(is.na(mpg2.transmissions$cyl))

mpg2.cylinders <- mpg2.transmissions
mpg2.cylinders <- mpg2.cylinders[-naCyl, ]

dashCyl <- which(mpg2.cylinders$cyl == "-")
```

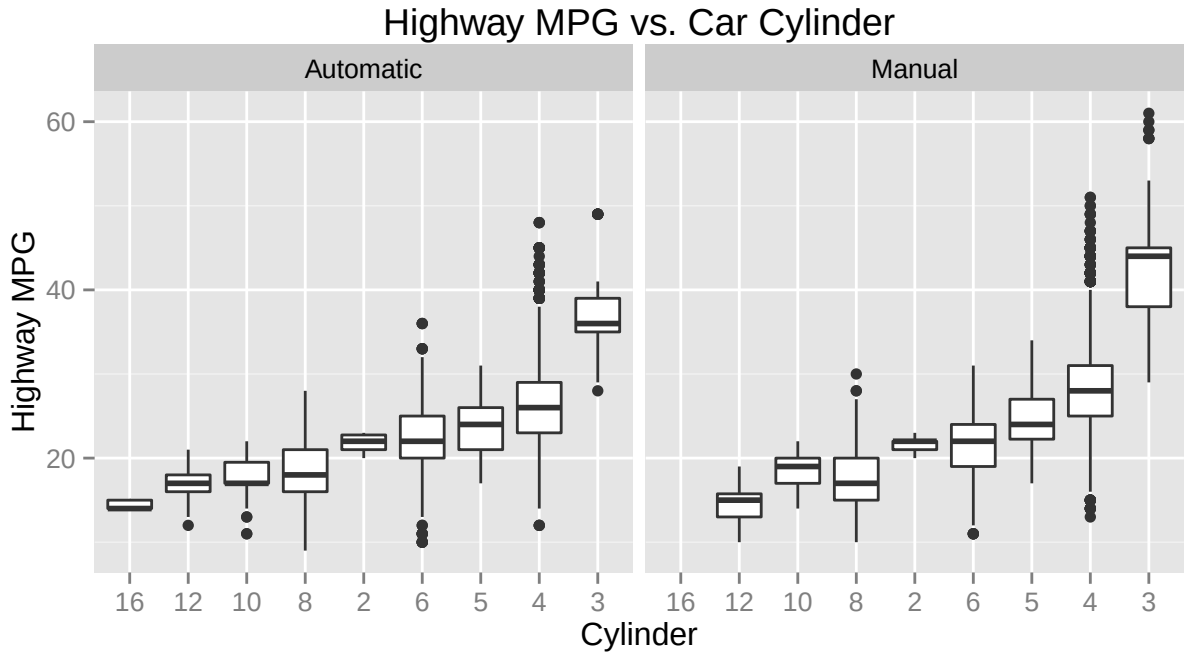
```
mpg2.cylinders <- mpg2.cylinders[-dashCyl,]

cyl.labels <- c("2" = "4 or lower", "3" = "4 or lower", "4" = "4 or lower",
                "5" = "5 or 6", "6" = "5 or 6", "8" = "8 or higher",
                "10" = "8 or higher", "12" = "8 or higher", "16" = "8 or higher")

mpg2.cylinders$Cylinder <- cyl.labels[mpg2.cylinders$cyl]

qplot(tranClean, fill = Cylinder, data = mpg2.cylinders) +
      xlab("Transmission Type") +
      ggtitle("Frequency of Transmission Types by Cylinder")
```



The above histogram confirms my previous suspicions. Manual cars are, on average, much lower cylinder than automatic ones. It then seemed natural to determine whether lower cylinder cars are more fuel efficient. I turned to a boxplot to answer my questions:

```
qplot(reorder(as.character(cyl), hwy), hwy, data = mpg2.cylinders,
      geom = "boxplot") + xlab("Cylinder") + ylab("Highway MPG") +
```

```
ggtitle("Highway MPG vs. Car Cylinder") + facet_wrap(~ tranClean)
```



In the boxplot above, it becomes clear that cars with lower cylinders are more fuel efficient. Our previous histogram, on page 3, indicated that manual cars had mostly lesser cylinders. Furthermore, due to the facet wrap in the boxplot directly above, we know that this relationship holds regardless of transmission type. Thus, it seems reasonable that my hypothesis that manual cars are more fuel efficient due to their, on average, lower cylinder content holds credibility.

3 Conclusion

All in all, after some data manipulation and plotting we were able to see that manual cars were, on average, more fuel efficient than automatic cars. Inside those two categories, naturally aspirated engines covered the largest range of highway MPGs while turbocharged and supercharged engines were both mostly in the high teens - mid twenties range. In terms of charged engines, turbocharged ones were seen to be slightly more fuel efficient on the highway. While our data for automatic cars took a nice shape, I am skeptical of whether we have conclusive results, given our 'manual' plot distribution. The manual portion of Figure 1 is jagged and inconsistent - leading me to believe that we either have too small of a sample size and/or that the relationship I am drawing isn't as strong as I would hope. If we were to consider more manual cars, some progress towards a better answer could be made.

A deeper exploration of highway MPG vs. cylinders lessened some of my worries. I was able to find that the number of cylinders was negatively related to highway MPG. Also, with a quick implementation of 'histogram' and 'fill', I saw that almost all manual cars were 4 cylinder or lower, while automatic cars had large amounts of higher cylinder cars. Therefore, it does not seem unreasonable to suggest that the higher fuel economy of manual types as illustrated in Figure 1 can partially be attributed to the number of cylinders in most manual cars. However, some more search as to the significance of the 'cyl' variable as a predictor of highway MPG might yield some results that go against my findings. Also, the smaller sample size of manual cars leaves much to be desired in terms of making definitive statements.