STAT 405: Presentation

Ahmad Razi

April 26, 2016

Contents

1 Introduction

1.1 Background

The data being used is the Fuel Economy data of various vehicles retrieved from the United States' Department of Energy's website, fueleconomy.gov (specifically, https://www.fueleconomy.gov/feg/download.shtml), called `vehicles.csv`. According to the source, the data was collected from testing done by the Environmental Protection Agency's National Vehicle and Fuel Emissions Laboratory (Ann Arbor, Michigan) and by vehicle manufacturers (with EPA oversight). The data set includes model years as old as 1984 and as recent as 2017 (as designated by vehicle manufacturers).

1.2 The variables of interest

There are a number of unique variables in the `vehicles.csv`. The variables that are the most valuable insightful are listed below:

- `co2TailpipeGpm` - tailpipe CO2 in grams/mile
- `comb08` - combined MPG (measured or estimated to 2008 testing standards)
- `cylinders` - number of cylinders
- `displ` - engine displacement in liters
- `drive` - drive axle type
- `fuelType` - engine descriptor, values include:
  - Various types of conventional gasoline: `Regular`, `Premium`, `Gasoline or natural gas`, `Gasoline or E85`, `Gasoline or propane`, `Premium or E85`, `Midgrade`, `Premium Gas or Electricity`, `Regular Gas and Electricity`, `Premium and Electricity`, and `Regular Gas or Electricity`
  - `CNG` - compressed natural gas
  - `Diesel` - Diesel fuel
  - `Electricity` - powered by stored electricity
- `make` - vehicle manufacturer
- `model` - model name
- `trany` - transmission
- `vclass` - EPA vehicle size class
- `year` - model year
- `youSaveSpend` - Dollars (USD) consumer saves/spends over 5 years compared to an average car because of fuel economy

Transformations include:

- discarding duplicate entries
- removing incomplete entries
- converting codes to more legible labels
- deciding on which metrics are the most appropriate
- converting some obscure units into more understandable units.

1.3 The Question

Given the data on the fuel economy of vehicles, what trends can be seen regarding the environmental impact of vehicles (in terms of fuel consumption and emissions) with changes in model years, manufacturers, transmissions, vehicle size class, drive axle, and the other variables mentioned above? The public is often told from advertising that certain factors (such as less cylinders) lead to greater fuel efficiency, but it is important to break down the data and give a quantitative answer to this question. Also, the push for more fuel efficient and environmentally friendly cars necessitates understanding the trend of fuel economy across various factors and which factors work best to create an ideal car, in terms of fuel economy and environmental impact. Has the United States really moved towards more environmentally friendly vehicles, or has it fallen short of expectations?

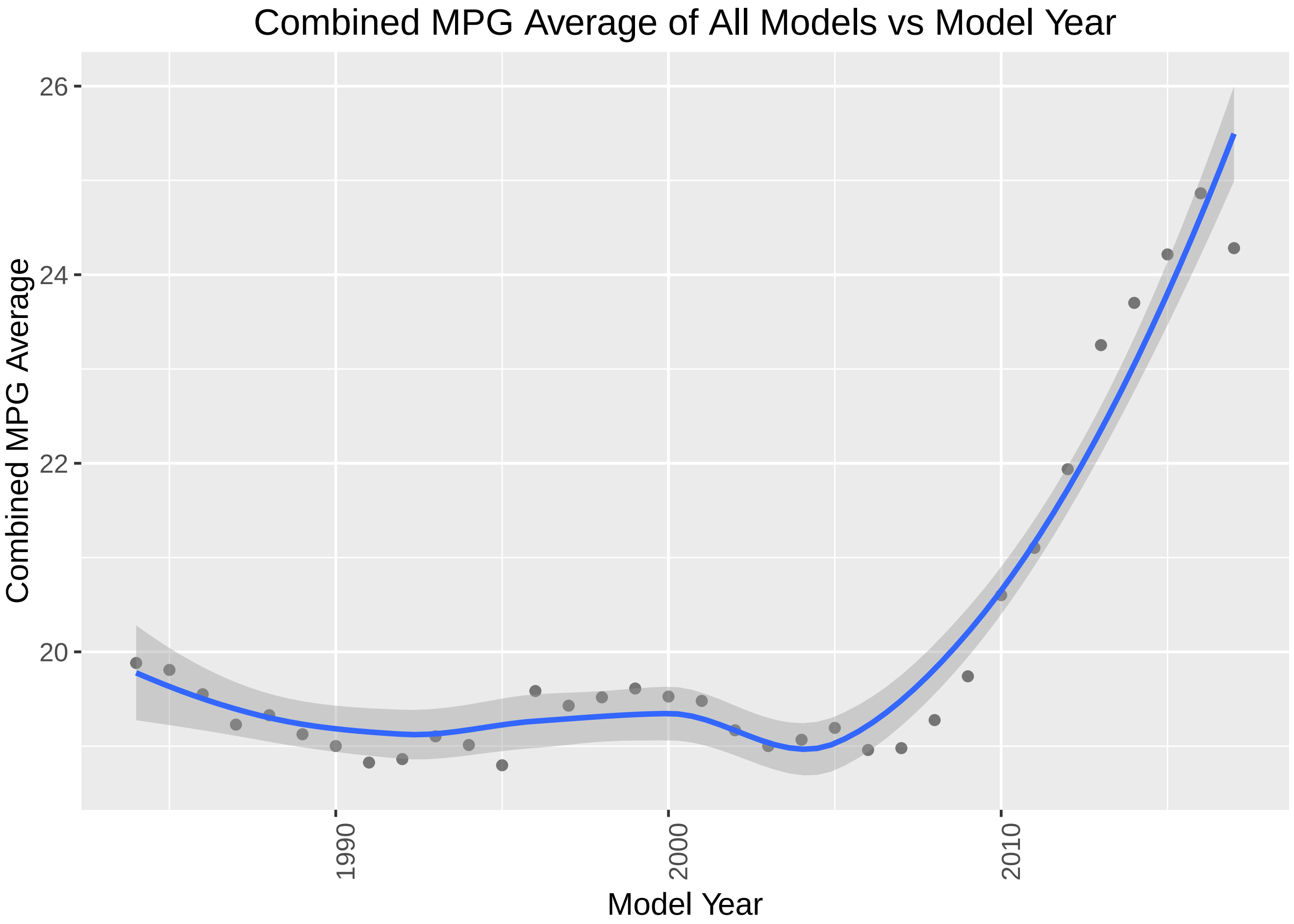
The main interest regarding this dataset is the insight it gives in the manufacturing trends of U.S. cars and how they have changed with time and other specifications. From the insights, possible explanations can be formulated and possibly tested further.

2 Exploring the Data

2.1 Trends with Time

One of the most important variables in a large dataset such as `vehicles.csv` is time. Thanks to the variable `year`, we are able to see various trends in the variables of interest with respect to time.

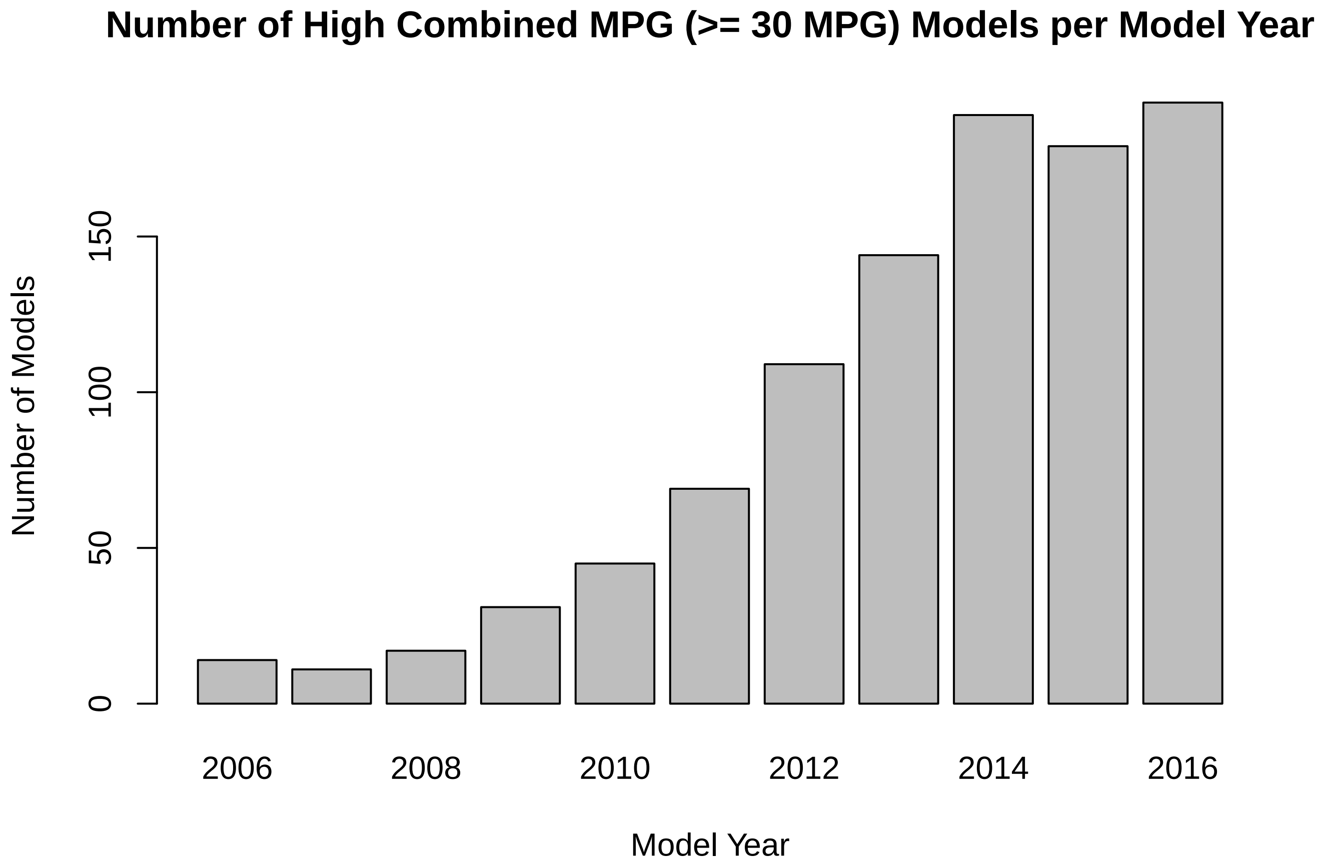
A simple plot of the average combined MPG versus model year for all the vehicles in the dataset is an excellent way to start off:



As this plot is analyzed, it is important to recognize the trend in the increasing MPG and hypothesize what is causing this. Given the recent societal emphasis on energy efficient vehicles, a good possible explanation is that there has been a recent increase in high MPG vehicles. This explanation is explored and tested in the next section by analyzing the high MPG vehicle subset of the dataset.

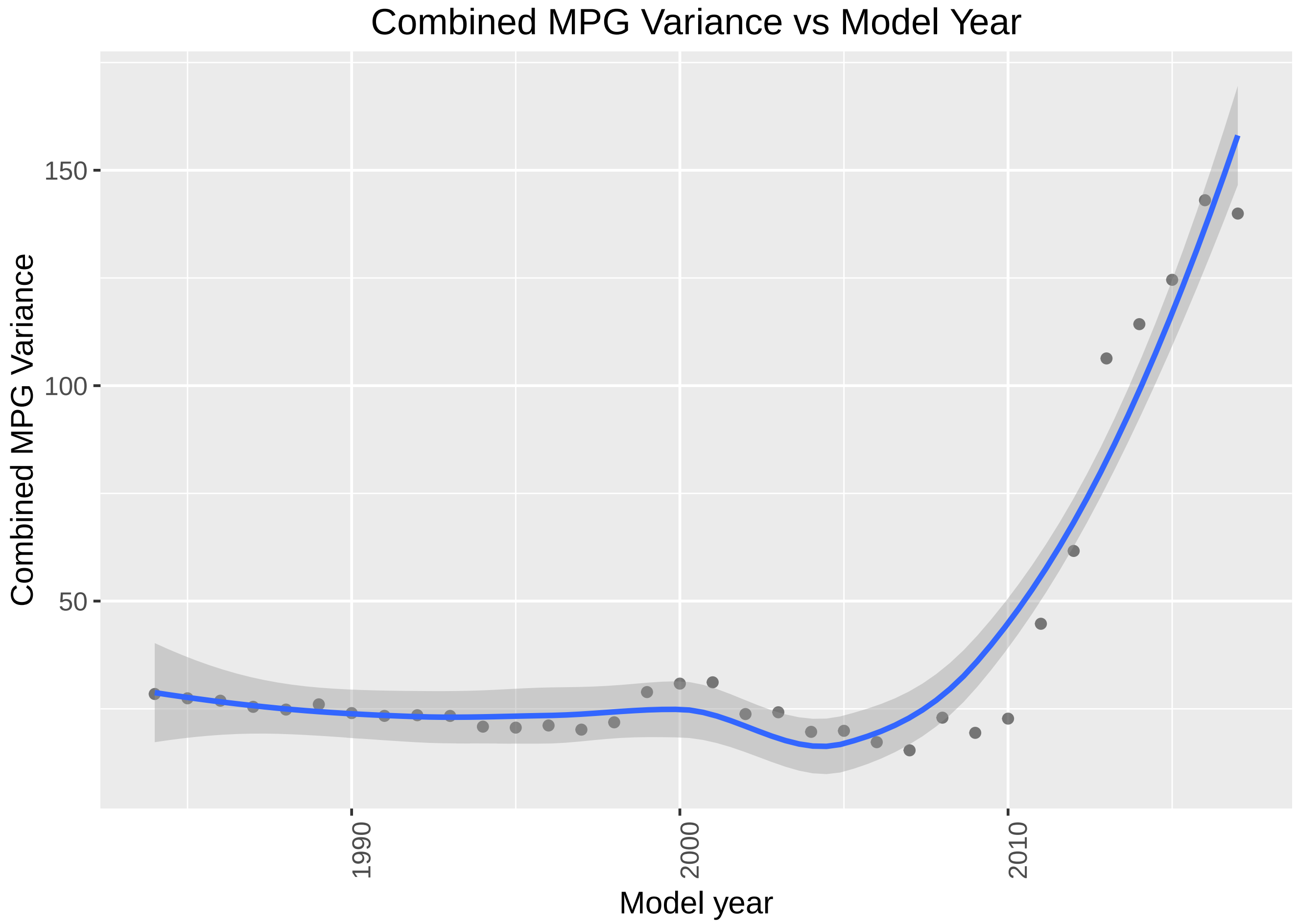
2.2 High MPG Vehicles

The last decade has seen an large rate of increase in the average combined MPG of all EPA tested models compared to the years before then, thanks to public awareness of climate change, scientific studies, and limited natural (e.g. petroleum) resources. A bar graph is appropriate to see the amount of high MPG (defined by me as greater than or equal to 30 miles per gallon) models tested by the EPA in the past decade.



This bar plot also confirms that the number of high MPG models is increasing with time. In fact, the number of high MPG vehicles being offered by manufacturers to Americans is more than triple than the number offered just six years ago in 2010. These high MPG models are likely contributing to the overall increase of the combined MPG average of all vehciles offered. However, are all vehicles following a trend of higher efficiency? Or is there some variance in the MPG ratings?

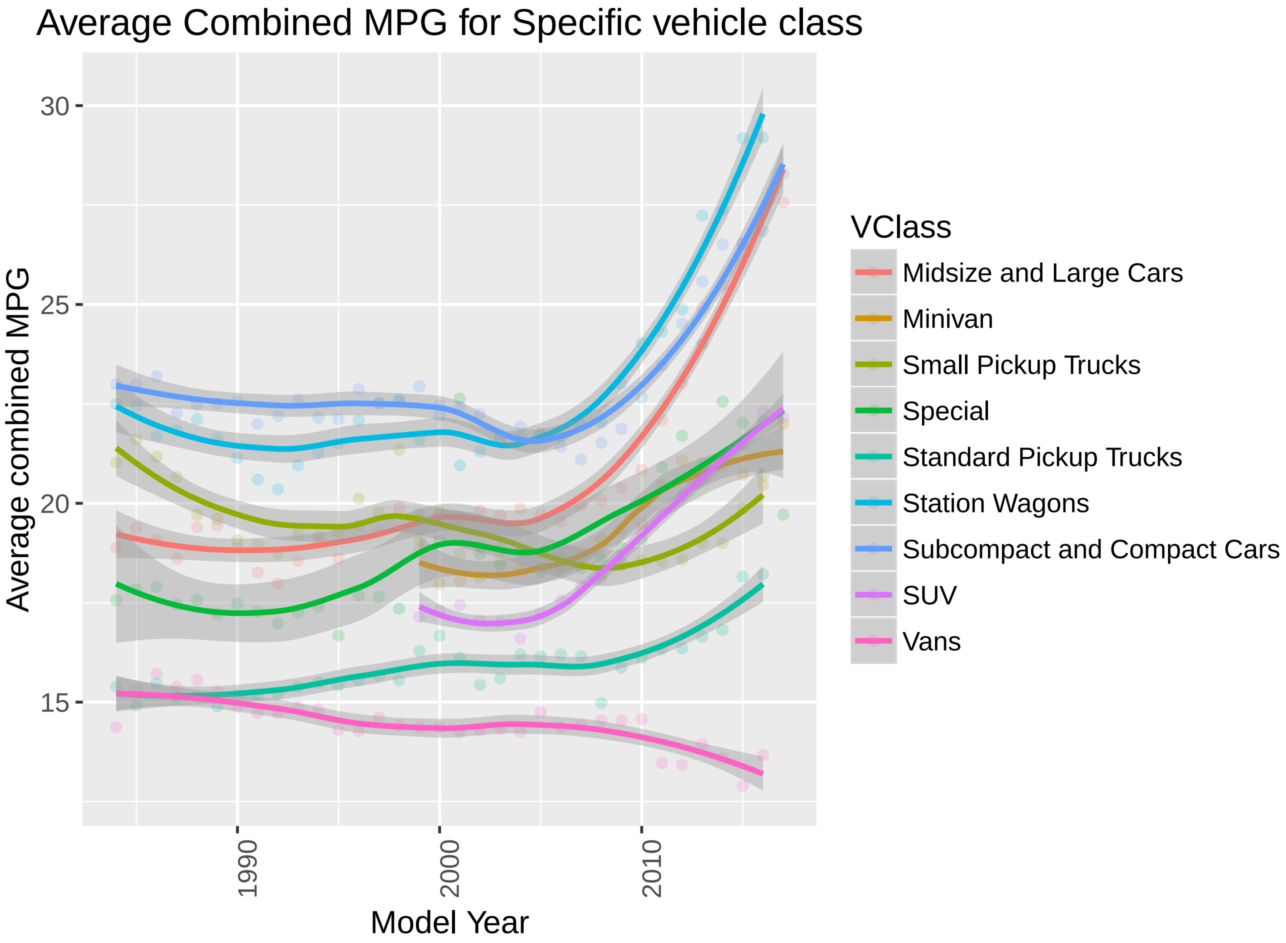
With these new high MPG models, lower MPG models continue to exist, and this leads to an increase of variance of the years, as evidenced by the combined MPG variance versus time plot:



The plot of combined MPG variance with model years helps show that in the past decade, the spread of combined MPG values for EPA tested vehicles has increased dramatically. The number of vehicles with high MPG ratings is increasing, but there are also still a significant number of normal and low MPG vehciles that lead to an increase in the variance in combined MPG ratings.

2.3 Trends in MPG ratings with respect to vehicle class

Another important trend to investigate is the trend of MPG ratings among vehcile classes. SUVs and heavy duty pickup trucks carry the social stigma of being ”gas guzzlers” and overall fuel inefficient, whereas compact cars and other small vehicles are touted as more efficient due to their smaller size, lighter weight, and less required power. In order to investigate how much the EPA-designated vehicle class plays a role in the MPG rating.



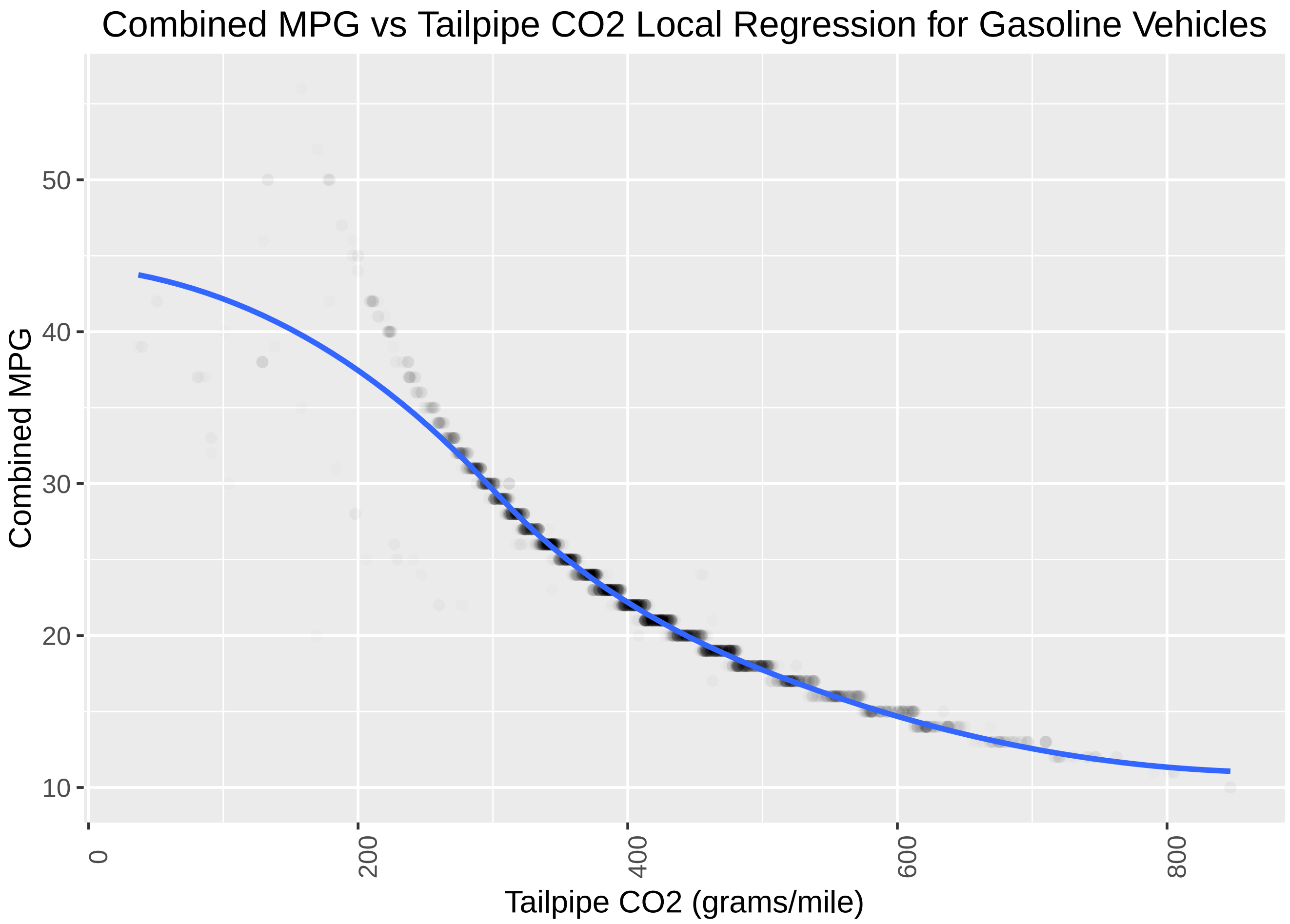
As was expected , lighter cars such as station wagons and (sub)compact cars have overall higher average MPG ratings than SUVs, trucks, and vans. One interested item to note, however, is how midsize and large cars have almost "caught up" to (sub)compact cars in terms of average combined MPG ratings in recent years. Despite the class and size difference, the avaerages are very close. This could be attributed to the fact that changes in manufacturing techniques have helped made larger cars comparable in weight and power to (sub)compact cars, casuing the closing of the gap between the two vehicle classes MPG-wise.

Another interesting relationship that also surprises a little is the relationship between fuel economy and emissions.

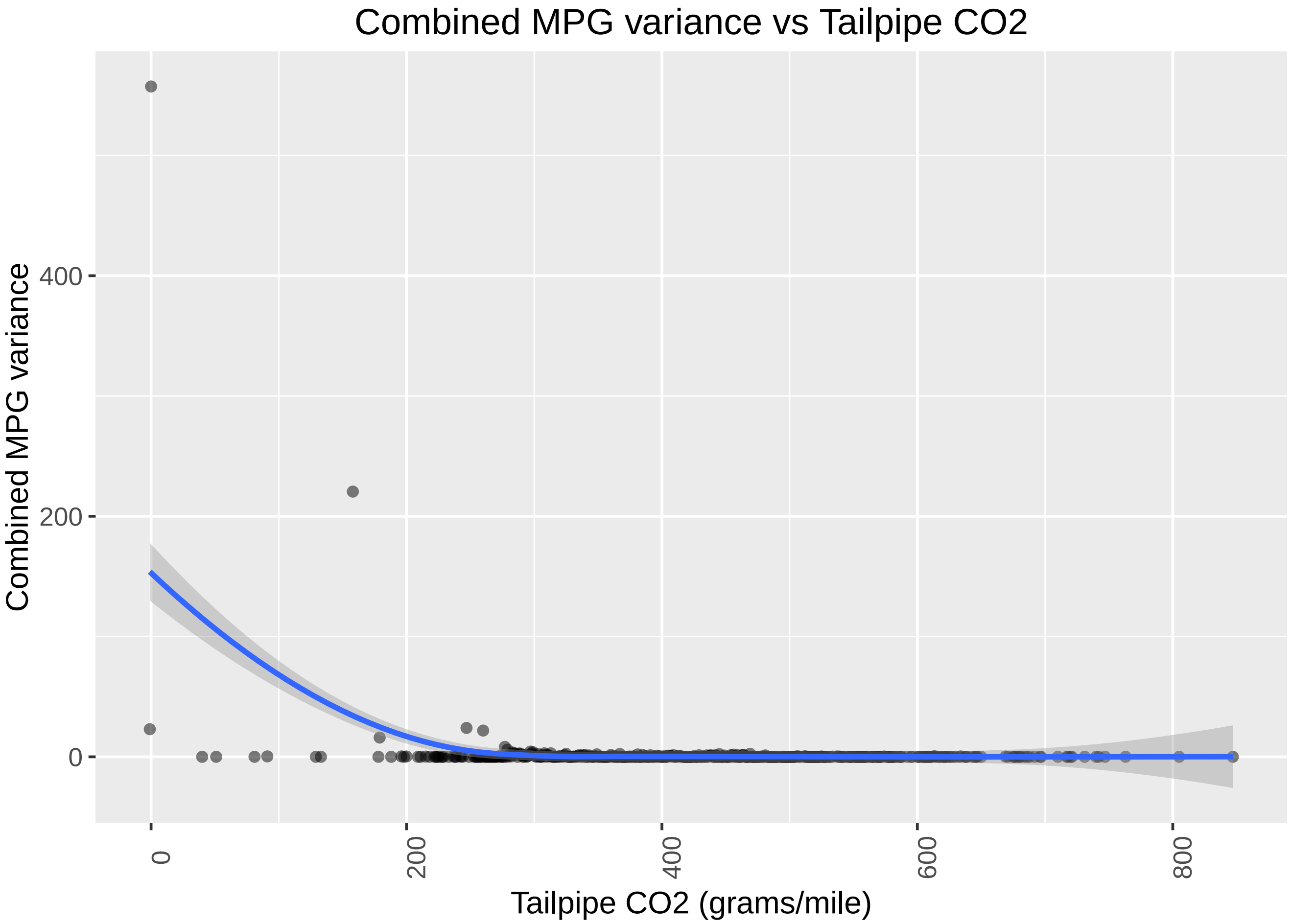
2.4 Emissions Analysis

MPG ratings and emissions amounts have often been thought to go hand in hand. The more fuel burned seems to automoatically equate to more emissions, especially as the sensitive mechanism of disposing these emissions has become regulated and standardized in the automotive industry.

Therefore, the plot of the relationship between combined MPG and carbon dioxide emissions from the exhaust pipe for gasoline vehicles can be analyzed to see if this trends holds any water:



The relationship is not that novel in a sense: vehicles with higher combined MPG ratings have lower tailpipe CO2 emissions; however, the relationship is not exactly as straightforward as one might expect. Rather than appearing almost linear, there is an obvious increase in variation in combined MPG towards the lower end of the tailpipe CO2 spectrum. This variance can be observed more properly here in an explicit graph of the variance:



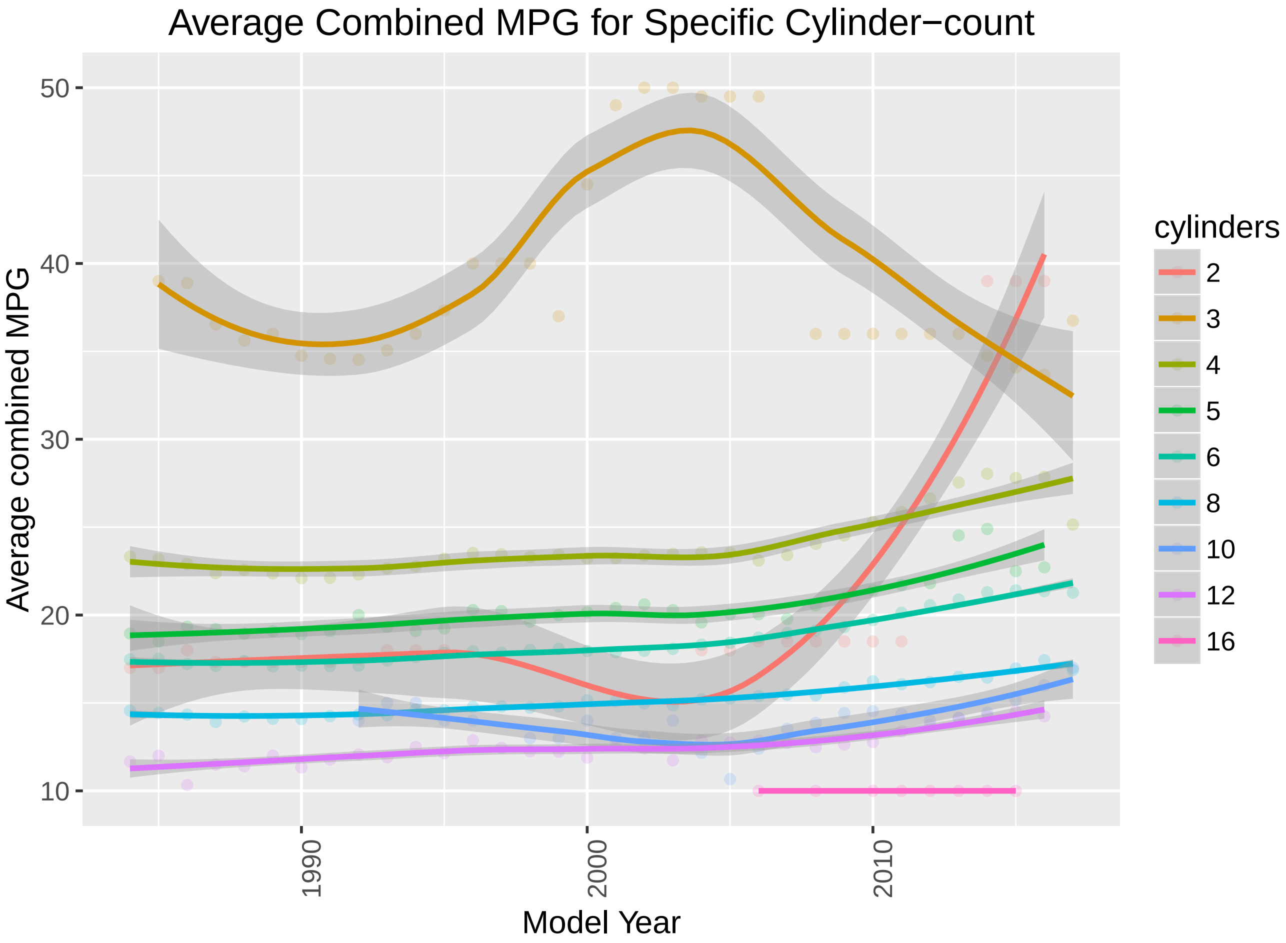
Here, it can be seen that, relative to the rest of the data set, lower tailpipe CO2 tested vehicles have a much larger amount of variance in combined MPG ratings. This might draw into question exactly how direct the relationship between fuel economy and exhaust emissions is and if the public opinion that high MPG means lower emissions is truly valid. This could also mean that since there are too few (relatively) high MPG vehciles compared to the rest of the vehicle offerings, variation amongst the high MPG vehicles themselves can lead to this observed phenomenom.

In a related topic, we can investigate the effect of the number of cylinders (more cylinders likely means more fuel used) to investigate MPG trends.

2.5 Trends with respect to Cylinders

Another popular perception is that fewer cylinders means higher MPG ratings. In fact, some vehicle manufacturers offer two version of a model, one with less cylinders than the other, and the model with less cylinders sometimes qualifies for tax breaks while the other model does not.

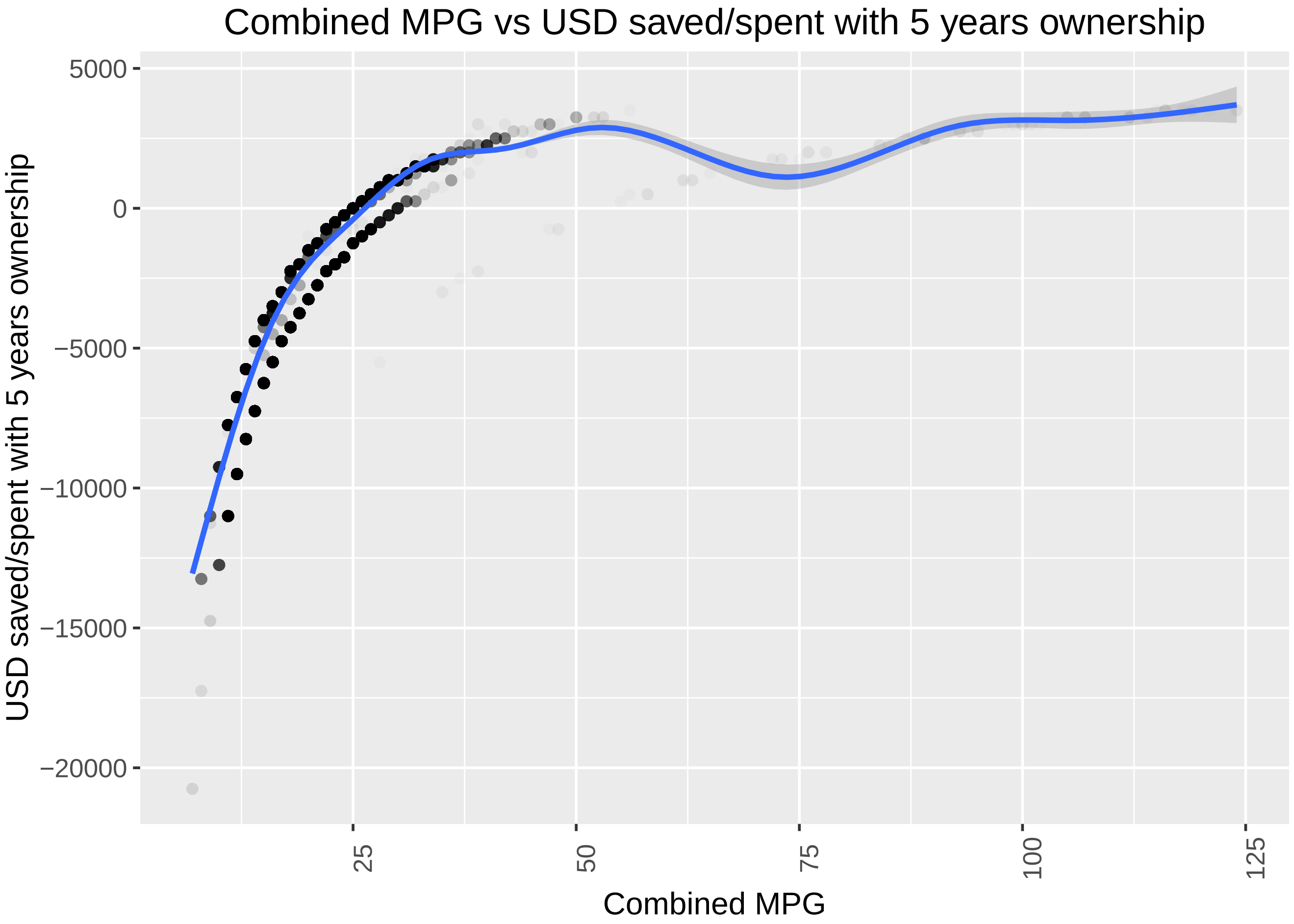
Here, a simple plot of the average MPG over the years by clinder quantity was constructed:



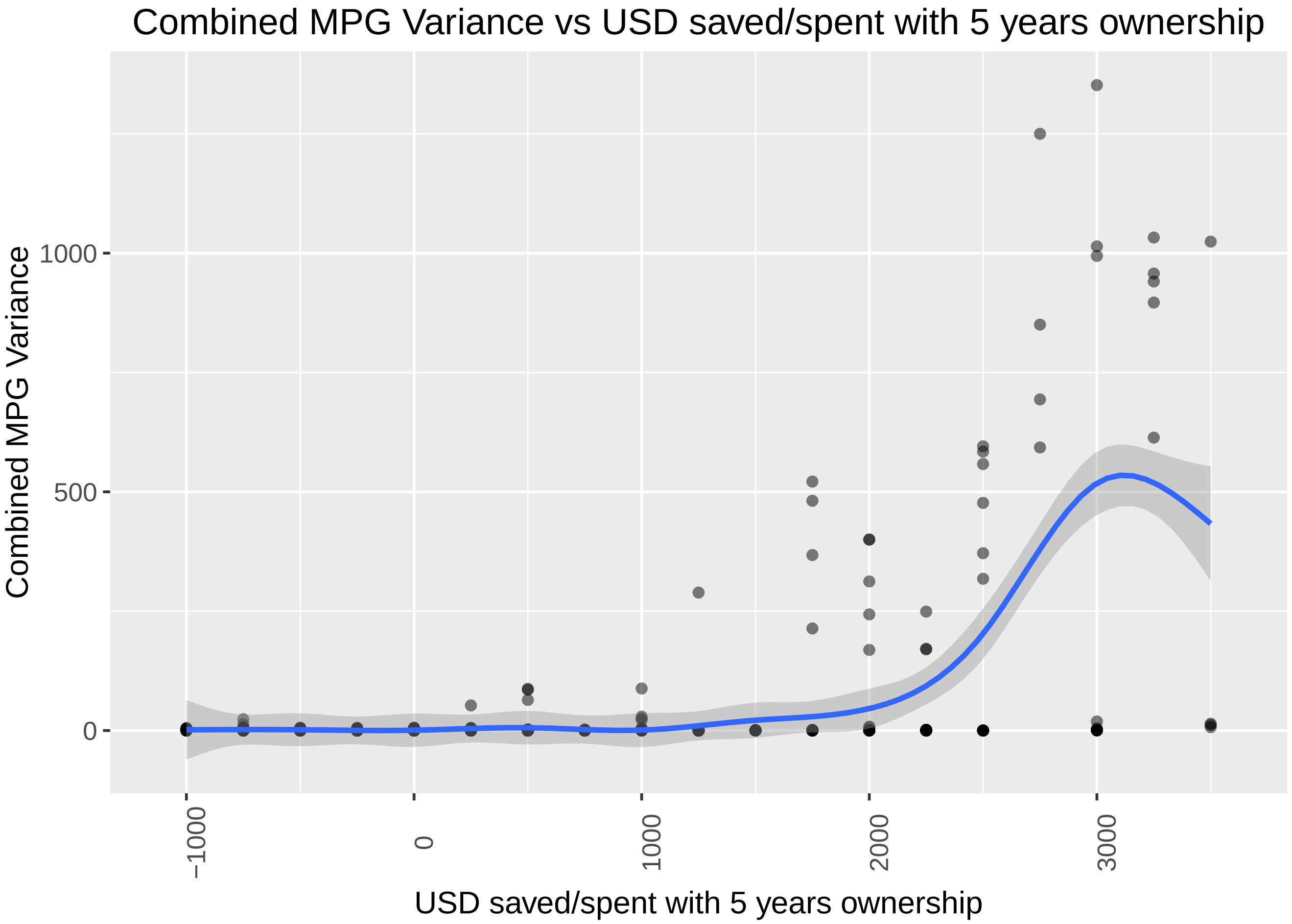
Here, the results are unsuprising for the most part: the fewer the cylinders, the higher the average MPG ratings. This holds true for cylinder quantities 4 or more, but there has been interesting fluctuations in the MPg ratings for 2 and 3 cylinder vehicles, which can be attributed to there being fewer of such vehicles (and those vehicles being special-type vehicles) and therefore prone to flutuations.

2.6 Trends with cost of ownership.

The `youSaveSpend` variable shows the amount of money in U.S. dollars an owner of a particular car model would save (or spend, if the value is negative) over five years over an average car. The costs is supposedly based off of the MPG ratings, but a closer analysis raises eyebrows



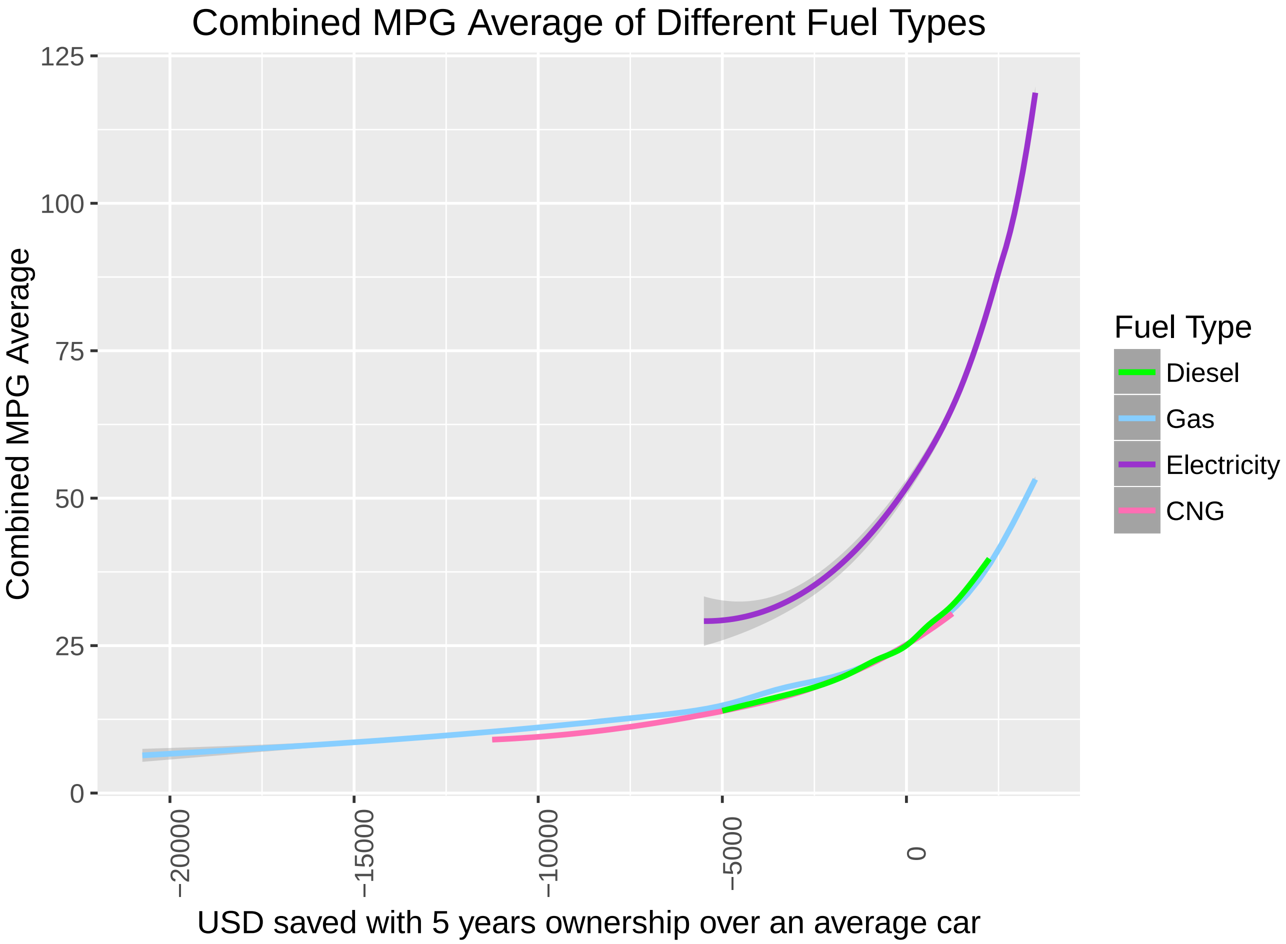
The relationship for lower MPG-rated vehicles seems to be normal, but higher MPG vehicles display interesting behaviors. The fewer amount of points inhigh MPG regions lend insight into what is causing this behavior: lower amount of high MPG vehicles compared to normal and low MPG vehicles. This effect can be seen in the variance of the cost of ownership:



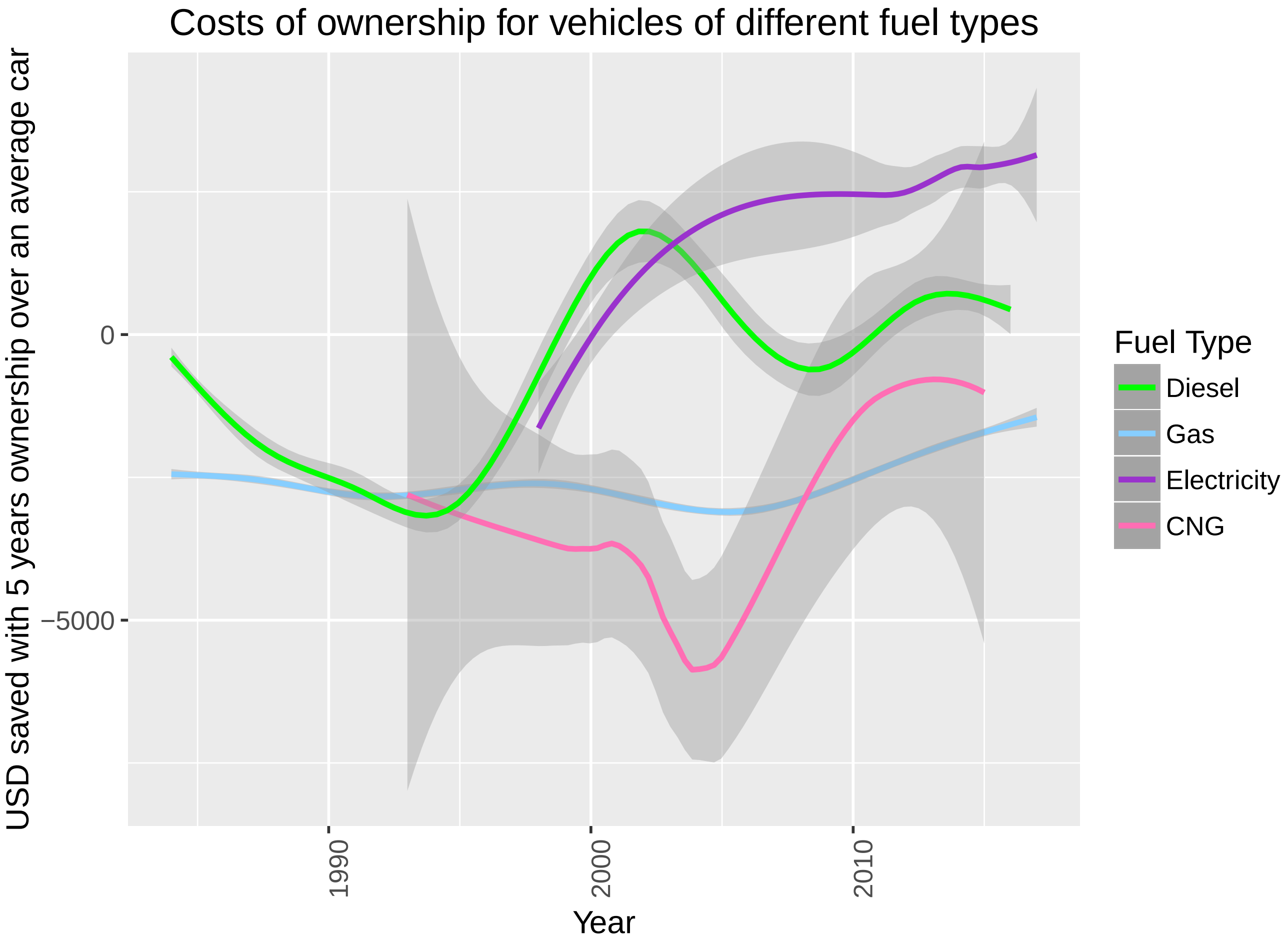
The cost of ownership values can also be examined with regards to different fuel types.

2.7 Fuel Types and Ownership Cost

Diesel, natural gas, and electricity powered cars have often been touted as being more environmentally friendly than gasoline cars, but does this environmental friendliness translate to lower ownership costs. The plot for this is straightforward in telling us the answer:



Here, it is surpising to see that only electrical cars have a notable fuel economy advantage over cars of other fuel types. But that finding seems to be overshadowed by the spread of the ownvership costs. Despite having higher MPG (determined from equivalent electrical usage), the costs of ownership seem to higher for electrical vehicles. The following plot confirms it.



Despite concerns that fuel costs and other consideration will make gasoline cars more expensive to own that electric cars, electric cars are actually the most expensive of the lot. The higher cost of purchasing an electric vehicle does not seem to be recouped. This is likely due to the amount of work that still has to go into refining the electric vehicle tecnology. Also, the rarity of CNG vehicles nowadays also likely contributes to the higher cost of ownership for CNG vehicles over gasoline vehicles.

3 Conclusions

3.1 Based off the data

Based off the data, there has been definitely been some misconceptions (both societal and personal) as to what contributes to more fuel efficient vehicles.

It has been shown that the effect of high MPG veicles has actaully not made quite the impact as one might imagine. In addition, factors such as vehicle class do not have as a drastic impact on MPG ratings in some cases, and that there are more similarities than differences. And that despite how far we've come in electric car technology, we have yet to make it more cost effective than gasoline vehicles.

But despite these findings, some common beliefs have been confirmed. Emissions and MPG ratings do share similar behavior at normal and low MPGS ratings. Cylinder quantities greater than 3 show a decreasing MPG rating systematically. And higher MPG gasoline vehicles have lower costs of ownership.

From this project, we are able to walk away with a greater understanding of the fuel economy trends of vehicles being manufactured for the U.S. public today.

3.2 Skepticism

Of course, when dealing with such large amounts of data, it is as important to know what the limits are as it is to know what cane be done. The data here is from the EPA's testing of vehicles. The EPA typically tests each vehicle model once per year and therefore this dataset can't give any information that directly reflects the fact that some cars are more popular than others. Were I to graph the average tailpipe CO2 for all cars in the United States in grams/mile, popular cars would cause larger influences on the data than less popular cars; in the dataset of this study however, a 2016 Toyota Camry holds the same significance as a 2016 Ferrari Enzo. Another aspect of this limitation is how the number of cars owned has increased significantly in recent years. The environmental impact of one 1990 Buick Regal is less than 100 Toyota Yarises.

Nonetheless, if we were to look at this data with the understanding that car manufacturers are faced against the need to make their vehicles more fuel efficient and more acclimated to today's world, this project has been able to elucidate the trends in the cars that come off the manufacturing belt today and have come off since 1984.

4 Appendix: Code

A

Loading GGplot2

```
#load ggplot2 and plyr
suppressWarnings(library(ggplot2))
library(ggplot2)
library(plyr)
#Loading data
vehicles = read.csv("vehicles.csv")
opts_chunk$set(fig.width=7, fig.height=5)
```

B

Combined MPG Average of All Models vs Model Year

```
avgyear <- ddply(vehicles, .(year), summarise, avg=mean(comb08))
ggplot(data=avgyear)+
aes(x=year, y=avg)+
geom_point(alpha = 0.5) +
theme(text = element_text(size=12), axis.text.x = element_text(angle=90, vjust=1))+
geom_smooth()+
ggtitle("Combined MPG Average of All Models vs Model Year")+
xlab("Model Year")+
ylab("Combined MPG Average")
```

C

Number of High Combined MPG (¿= 30 MPG) Models per Model Year

```
decade <- vehicles[vehicles$comb08 >= 30 & vehicles$year >= 2006 & vehicles$year != 2017,]
vehdec <- ddply(decade, .(year), summarise, count=length(model))
barplot(vehdec$count, vehdec$year, names.arg=c(2006:2016),
        main = "Number of High Combined MPG (>= 30 MPG) Models per Model Year",
        xlab = "Model Year", ylab = "Number of Models")
```

D

Combined MPG Variance vs Model Year

```
varyear <- ddply(vehicles, .(year), summarise, variance=var(comb08))

ggplot(data=varyear)+
aes(x=year, y=variance)+
geom_point(alpha = 0.5) +
# geom_smooth(method="lm", se=FALSE)+
theme(text = element_text(size=12), axis.text.x = element_text(angle=90, vjust=1))+
geom_smooth()+
ggtitle("Combined MPG Variance vs Model Year")+
xlab("Model year")+
ylab("Combined MPG Variance")
```

E

Average Combined MPG for Specific vehicle class

```
# simplifying vehicle classes vehclassmod <- data.frame(vehicles,
# stringsAsFactors=FALSE)
vehclassmod <- vehicles[, c(16, 63, 64)]
vehclassmod[, c(2)] <- sapply(vehclassmod[, c(2)], as.character)

# midsize and large cars
vehclassmod$VClass[vehclassmod$VClass == "Midsize Cars" | vehclassmod$VClass ==
    "Large Cars"] <- "Midsize and Large Cars"

# minivans
vehclassmod$VClass[vehclassmod$VClass == "Minivan - 2WD" | vehclassmod$VClass ==
    "Minivan - 4WD"] <- "Minivan"

# small pickup trucks
vehclassmod$VClass[vehclassmod$VClass == "Small Pickup Trucks" | vehclassmod$VClass ==
    "Small Pickup Trucks 2WD" | vehclassmod$VClass == "Small Pickup Trucks 4WD"] <- "Small Pickup Trucks"

# special
vehclassmod$VClass[vehclassmod$VClass == "Special Purpose Vehicle" | vehclassmod$VClass ==
    "Special Purpose Vehicle 2WD" | vehclassmod$VClass == "Special Purpose Vehicle 4WD" |
    vehclassmod$VClass == "Special Purpose Vehicles" | vehclassmod$VClass ==
    "Special Purpose Vehicles/2wd" | vehclassmod$VClass == "Special Purpose Vehicles/4wd" |
    vehclassmod$VClass == "Two Seaters"] <- "Special"

# Standard Pickup Trucks
vehclassmod$VClass[vehclassmod$VClass == "Standard Pickup Trucks" | vehclassmod$VClass ==
    "Standard Pickup Trucks 2WD" | vehclassmod$VClass == "Standard Pickup Trucks 4WD" |
    vehclassmod$VClass == "Standard Pickup Trucks/2wd"] <- "Standard Pickup Trucks"

# Station Wagons
vehclassmod$VClass[vehclassmod$VClass == "Midsize Station Wagons" | vehclassmod$VClass ==
    "Midsize-Large Station Wagons" | vehclassmod$VClass == "Small Station Wagons"] <- "Station Wagons"

# Subcompact and Compact Cars
vehclassmod$VClass[vehclassmod$VClass == "Minicompact Cars" | vehclassmod$VClass ==
    "Subcompact Cars" | vehclassmod$VClass == "Compact Cars"] <- "Subcompact and Compact Cars"

# SUV
vehclassmod$VClass[vehclassmod$VClass == "Small Sport Utility Vehicle 2WD" |
    vehclassmod$VClass == "Small Sport Utility Vehicle 4WD" | vehclassmod$VClass ==
    "Sport Utility Vehicle - 2WD" | vehclassmod$VClass == "Sport Utility Vehicle - 4WD" |
    vehclassmod$VClass == "Standard Sport Utility Vehicle 2WD" | vehclassmod$VClass ==
    "Standard Sport Utility Vehicle 4WD"] <- "SUV"

# Vans
vehclassmod$VClass[vehclassmod$VClass == "Vans" | vehclassmod$VClass == "Vans Passenger" |
    vehclassmod$VClass == "Vans, Cargo Type" | vehclassmod$VClass == "Vans, Passenger Type"] <- "Vans"

avgclass <- ddply(vehclassmod, .(VClass, year), summarise, avg = mean(comb08))
avgclass <- na.omit(avgclass)
```

```r
avgclass$VClass <- factor(avgclass$VClass, levels = unique(avgclass$VClass))

ggplot(avgclass, aes(x = year, y = avg, colour = VClass)) + geom_point(alpha = 0.2) +
    geom_smooth() + theme(text = element_text(size = 12), axis.text.x = element_text(angle = 90,
    vjust = 1)) + ggtitle("Average Combined MPG for Specific vehicle class") +
    xlab("Model Year") + ylab("Average combined MPG")
```

F

Combined MPG vs Tailpipe CO2 Local Regression for Gasoline Vehicles

```r
ggplot(data = vehicles[vehicles$co2 > 10 & vehicles$fuelType !=
    "CNG" & vehicles$fuelType != "Diesel" & vehicles$fuelType !=
    "Electricity", ]) + aes(x = co2, y = comb08) +
    geom_point(alpha = 0.01) + geom_smooth(method = "loess",
    se = FALSE) + theme(text = element_text(size = 12),
    axis.text.x = element_text(angle = 90, vjust = 1)) +
    ggtitle("Combined MPG vs Tailpipe CO2 Local Regression for Gasoline Vehicles") +
    xlab("Tailpipe CO2 (grams/mile)") + # geom_jitter()+
ylab("Combined MPG")
```

G

Combined MPG variance vs Tailpipe CO2

```r
varemissions <- ddply(vehicles, .(co2), summarise, variance=var(comb08))
ggplot(data=varemissions)+
aes(x=co2, y=variance)+
geom_point(alpha = 0.5) +
geom_smooth()+
theme(text = element_text(size=12), axis.text.x = element_text(angle=90, vjust=1))+
ggtitle("Combined MPG variance vs Tailpipe CO2")+
xlab("Tailpipe CO2 (grams/mile)")+
ylab("Combined MPG variance")
```

H

Average Combined MPG for Specific Cylinder-count

```r
avgcyl <- ddply(vehicles, .(cylinders, year), summarise, avg=mean(comb08))
avgcyl <- na.omit(avgcyl)

avgcyl$cylinders <- factor(avgcyl$cylinders, levels = unique(avgcyl$cylinders))

ggplot(avgcyl, aes(x = year, y = avg, colour=cylinders))+
geom_point(alpha=0.2)+
geom_smooth()+
theme(text = element_text(size=12), axis.text.x = element_text(angle=90, vjust=1))+
ggtitle("Average Combined MPG for Specific Cylinder-count")+
xlab("Model Year")+
ylab("Average combined MPG")
```

I

Combined MPG vs USD saved/spent with 5 years ownership

```
# youSaveSpend and comb08
ggplot(data=vehicles)+
aes(x=comb08, y=youSaveSpend)+
geom_point(alpha = 0.02) +
#geom_smooth(method="loess", se=FALSE)+
geom_smooth()+
#geom_jitter()+
theme(text = element_text(size=12), axis.text.x = element_text(angle=90, vjust=1))+
ggtitle("Combined MPG vs USD saved/spent with 5 years ownership")+
xlab("Combined MPG")+
ylab("USD saved/spent with 5 years ownership")
```

J

Combined MPG Variance vs USD saved/spent with 5 years ownership

```
yssvaryear <- ddply(vehicles, .(year, youSaveSpend), summarise, yssvariance=var(comb08))

ggplot(data=yssvaryear)+
aes(x=youSaveSpend, y=yssvariance)+
geom_point(alpha = 0.5) +
# geom_smooth(method="lm", se=FALSE)+
geom_smooth()+
theme(text = element_text(size=12), axis.text.x = element_text(angle=90, vjust=1))+
xlim(-1100,3600)+
ggtitle("Combined MPG Variance vs USD saved/spent with 5 years ownership")+
xlab("USD saved/spent with 5 years ownership")+
ylab("Combined MPG Variance")
```

K

Combined MPG Average of Different Fuel Types

```
ggplot() + geom_smooth(data = vehicles[vehicles$fuelType != "CNG" &
    vehicles$fuelType != "Diesel" & vehicles$fuelType != "Electricity",
    ], aes(youSaveSpend, comb08, color = "green"), show.legend = TRUE) +
    geom_smooth(data = vehicles[vehicles$fuelType == "CNG", ],
        aes(youSaveSpend, comb08, color = "skyblue1")) + geom_smooth(data = vehicles[vehicles$fuelType ==
    "Diesel", ], aes(youSaveSpend, comb08, color = "darkorchid3")) +
    geom_smooth(data = vehicles[vehicles$fuelType == "Electricity",
        ], aes(youSaveSpend, comb08, color = "hotpink1")) +
theme(text = element_text(size = 12), axis.text.x = element_text(angle = 90,
    vjust = 1)) + geom_smooth() + ggtitle("Combined MPG Average of Different Fuel Types") +
    xlab("USD saved with 5 years ownership over an average car") +
    ylab("Combined MPG Average") + scale_colour_manual(values = c("green",
    "skyblue1", "darkorchid3", "hotpink1"), name = "Fuel Type",
    labels = c("Diesel", "Gas", "Electricity", "CNG"))
```

L

Costs of ownership for vehicles of different fuel types

```
ggplot() + geom_smooth(data = vehicles[vehicles$fuelType != "CNG" &
    vehicles$fuelType != "Diesel" & vehicles$fuelType != "Electricity",
    ], aes(year, youSaveSpend, color = "green"), show.legend = TRUE) +
    geom_smooth(data = vehicles[vehicles$fuelType == "CNG", ],
        aes(year, youSaveSpend, color = "skyblue1")) + geom_smooth(data = vehicles[vehicles$fuelType ==
    "Diesel", ], aes(year, youSaveSpend, color = "darkorchid3")) +
    geom_smooth(data = vehicles[vehicles$fuelType == "Electricity",
        ], aes(year, youSaveSpend, color = "hotpink1")) + theme(text = element_text(size = 12),
    axis.text.x = element_text(angle = 90, vjust = 1)) + geom_smooth() +
    ggtitle("Costs of ownership for vehicles of different fuel types") +
    xlab("Year") + ylab("USD saved with 5 years ownership over an average car") +
    scale_colour_manual(values = c("green", "skyblue1", "darkorchid3",
        "hotpink1"), name = "Fuel Type", labels = c("Diesel",
        "Gas", "Electricity", "CNG"))
```