# Week 2 Assignment 1 – EDA

By: Zach Adair

Regis University

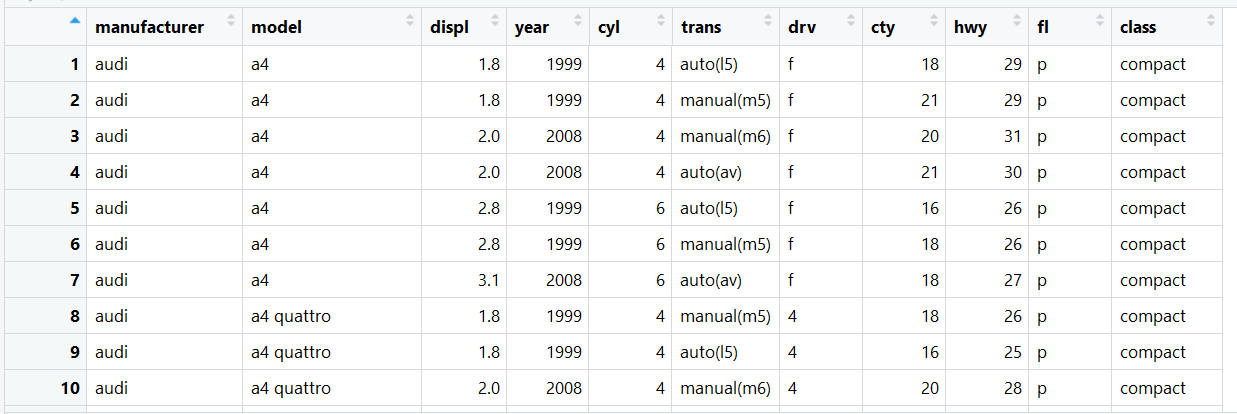
MSDS – 650 Data Analytics

## Assignment Objective

For Assignment 1, Week 2 we need to find a multivariate dataset of interest and create a data science working document. Start with preliminary analysis, than move onto trying to discovering answers to questions of interest through analysis. Finally summarize your findings and include the code used to throughout the analysis.

## Dataset of Choice: MPG

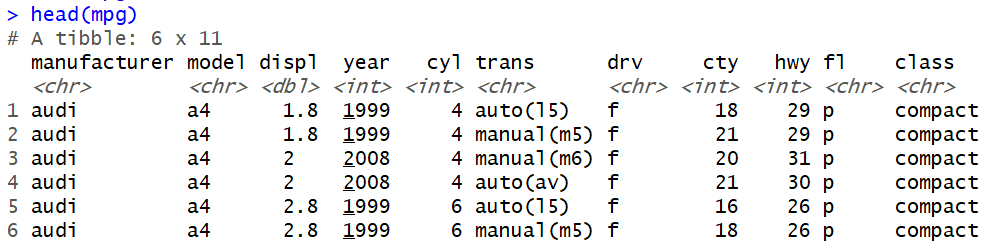
MPG is a dataset under the ggplot2 library, and is a multivariate dataset which takes a look at the fuel economy data from 1999 and 2008 for 38 car models. After using the command View(mpg) I got to look at the data in tabular form, here is a snapshot of that:



## Summary Analysis of the MPG Dataset

The MPG dataset is something I’ve never really got to look at before within my limited knowledge of RStudio, so doing some summary style commands will be a good idea before doing any further analysis on the data itself.

I’ll start with using the head function, just to again take a look at my data, kind of like how I could with the view command from earlier:

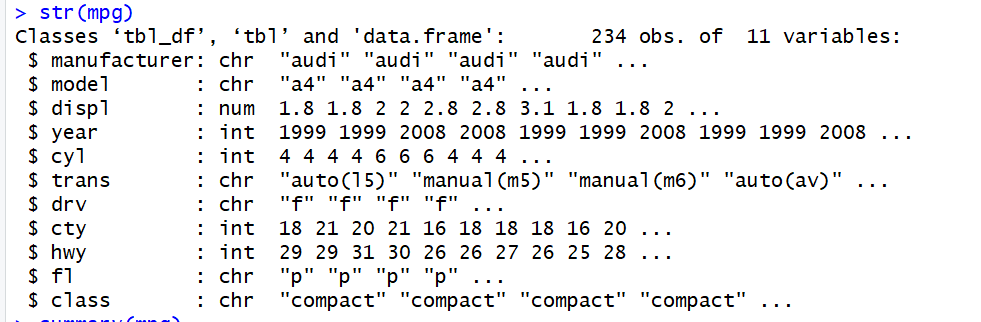


Next, check out the dimensions of the table:



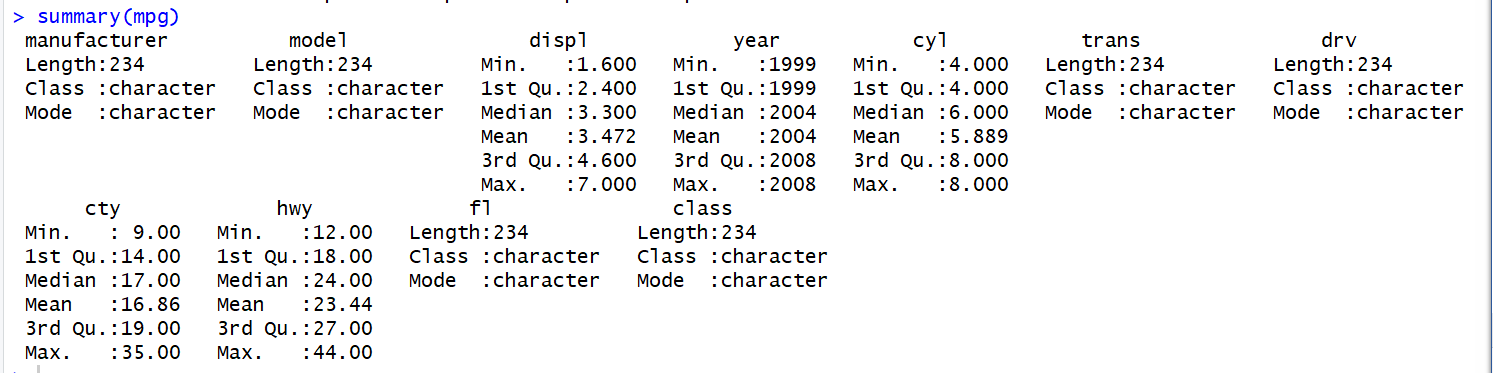
After seeing this we know there are 234 rows and 11 columns within the mpg dataset.

Now I want to see the internal structure of the mpg dataset:



Like the head command before it just flipped and shown in another way, both commands are good at understanding how the dataset is constructed.

The summary of the data is what I looked at next, to get a good understanding of each column individually and see a summary of each:



For the integer columns were able to get a better break down than the columns of characters but nonetheless good information from here to help find questions for further analysis.

From the summary commands I was able to find some good information about the construct of the dataset and know what kind of information is in it. It has also helped me form some questions that I would like to solve with some deeper analysis.

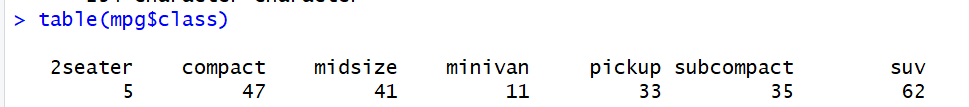
There are a couple questions I’d like to solve about the mpg dataset.

1. Which class of vehicle has the worst mpg on the highway and which has the worst in the city?
2. Which model has the best mpg?
3. Does cyl or displ seem to positively or negatively affect city or highway mileage?

## Answering those Questions

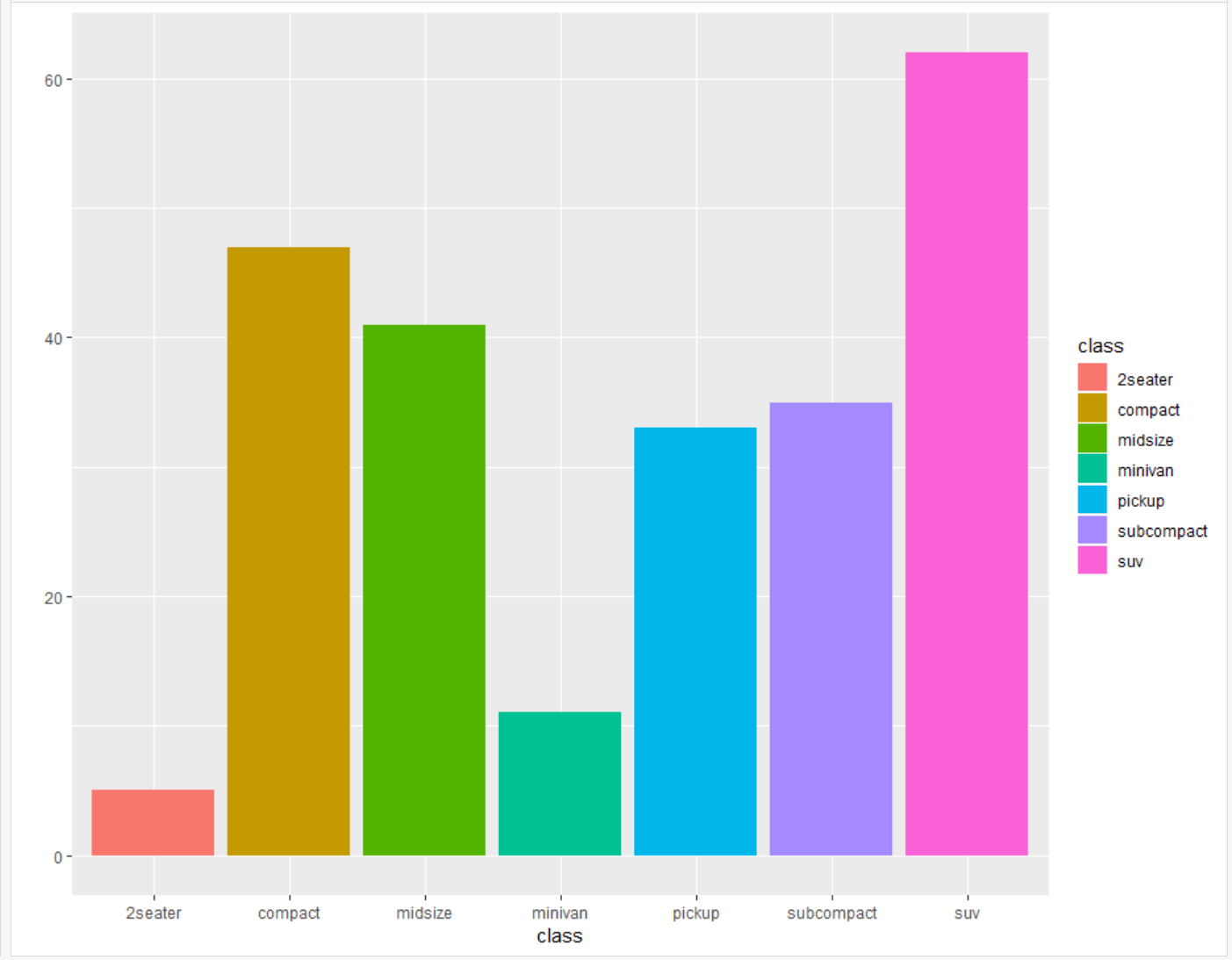
### Which class of vehicle has the worst mpg on the highway and which had the worst in the city?

Start with looking at how many vehicles I have of the Vehicle Classes:

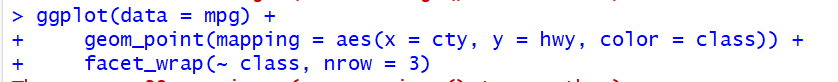


I have a wide range of different types of vehicles, the most I have is SUV’s with 62 and the least I have are the two seater cars with 5. Just from personal experience I see the 2 seaters less than I see any other vehicle so I’m not surprised the data holds this notion also. This is easier to show though through a better visual representation. Here I will create a bar chart of the table above.





Next, I want to make a grid of the data, which will show me each point plotted for class and if the class is higher or lower, from there it should be easier to answer which class has the worst highway mpg and which vehicle class has the worst city mpg.



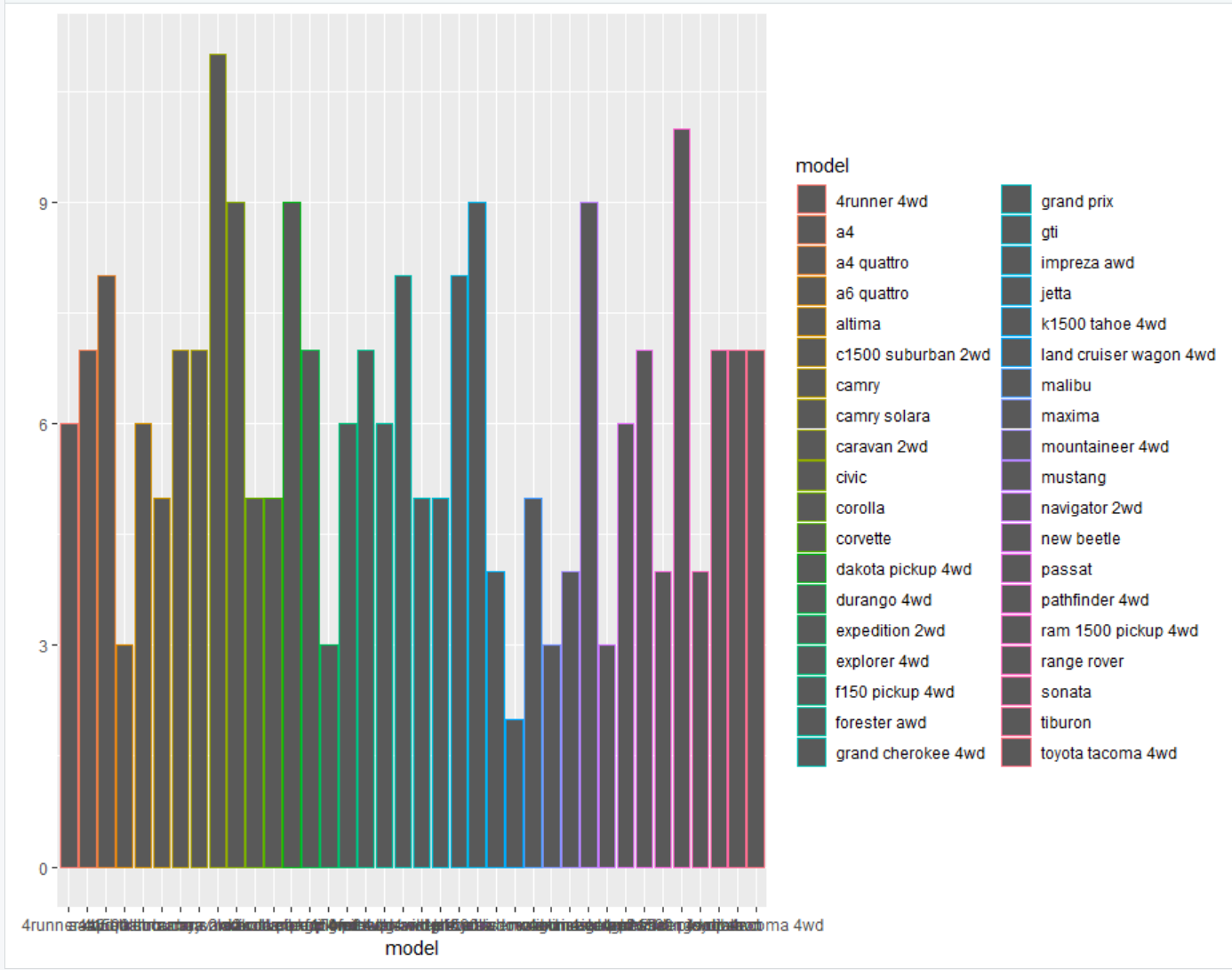


Since the data point for city and highway mpg was already given to us so the best way to display which is worst per class is to separate them out by class and then create a scatter plot for each class by the city and highway efficiency to see what sticks out visually and what doesn’t. For my question of which class is the worst, it is obvious that for highway and city mileage pickups are the worst. They are the closest to the bottom, even though SUV has a wider variance of vehicles which are better performing on the highway and in the city they still are generally bad performers too. From this group of plots what surprised me was the poor performance of 2 seater cars in the city. I know there isn’t a lot of 2 seaters to map here but it is the only one to not besides pickups to have an efficiency of over 17.5 with at least one of it’s vehicles.

### Which model has the best mpg?

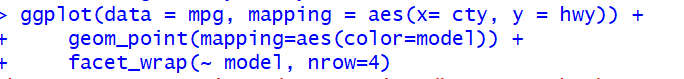
The search for this answer is similar to question 1, because they are pretty similar questions, so I started by just creating a bar chart of all the different car models we had and how many we have of them in the data.

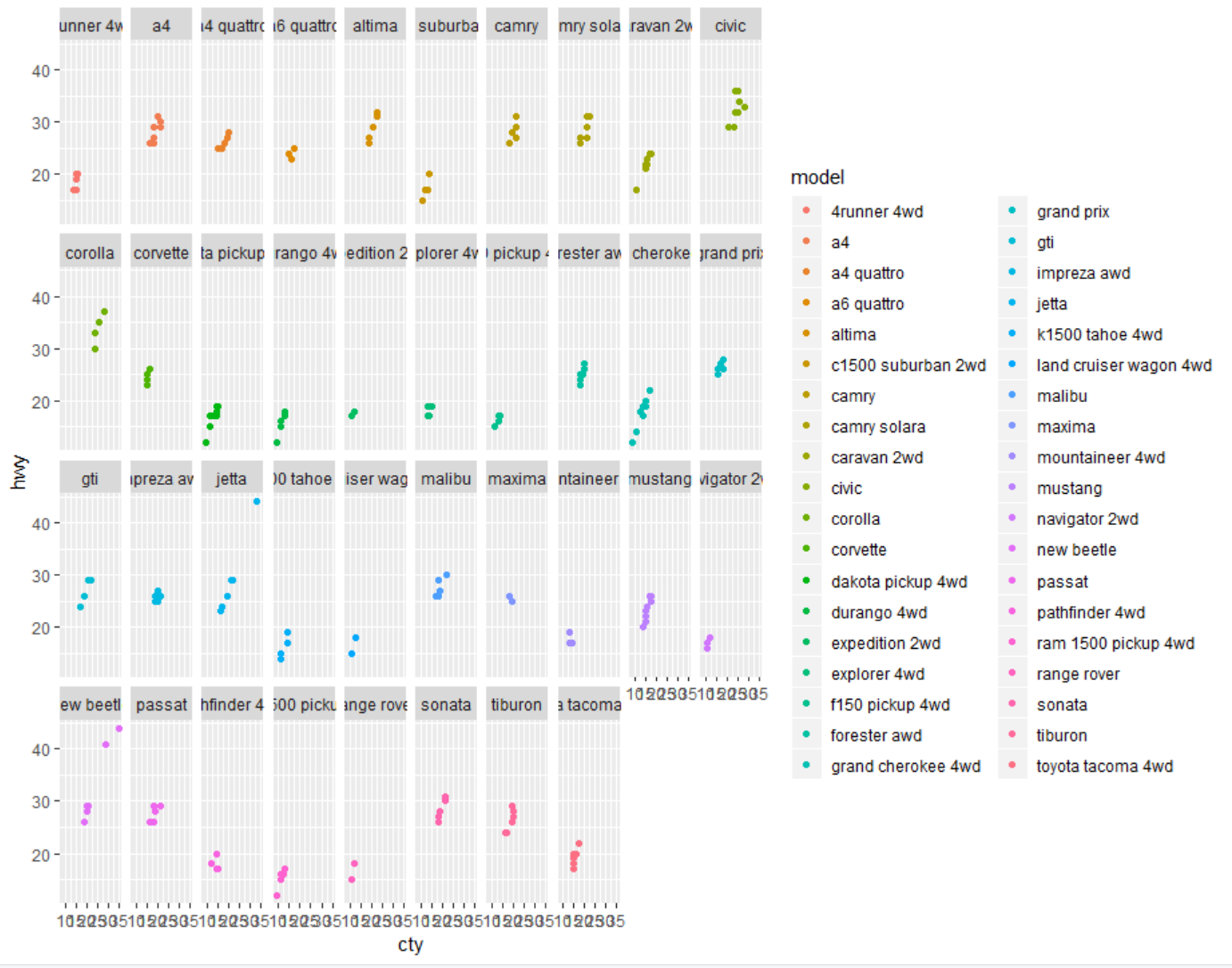




There are a lot of different models so it makes the data harder to read in Rstudio which is why the spacing on the x-axis is such a mess.

From here, did a similar step since each model already had the mpg for the city and highway in the data table, I separated out the models and then mapped them by their highway and city efficiency, from there it is easier to denote which model gets the best mileage overall.

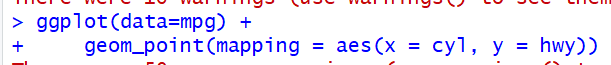


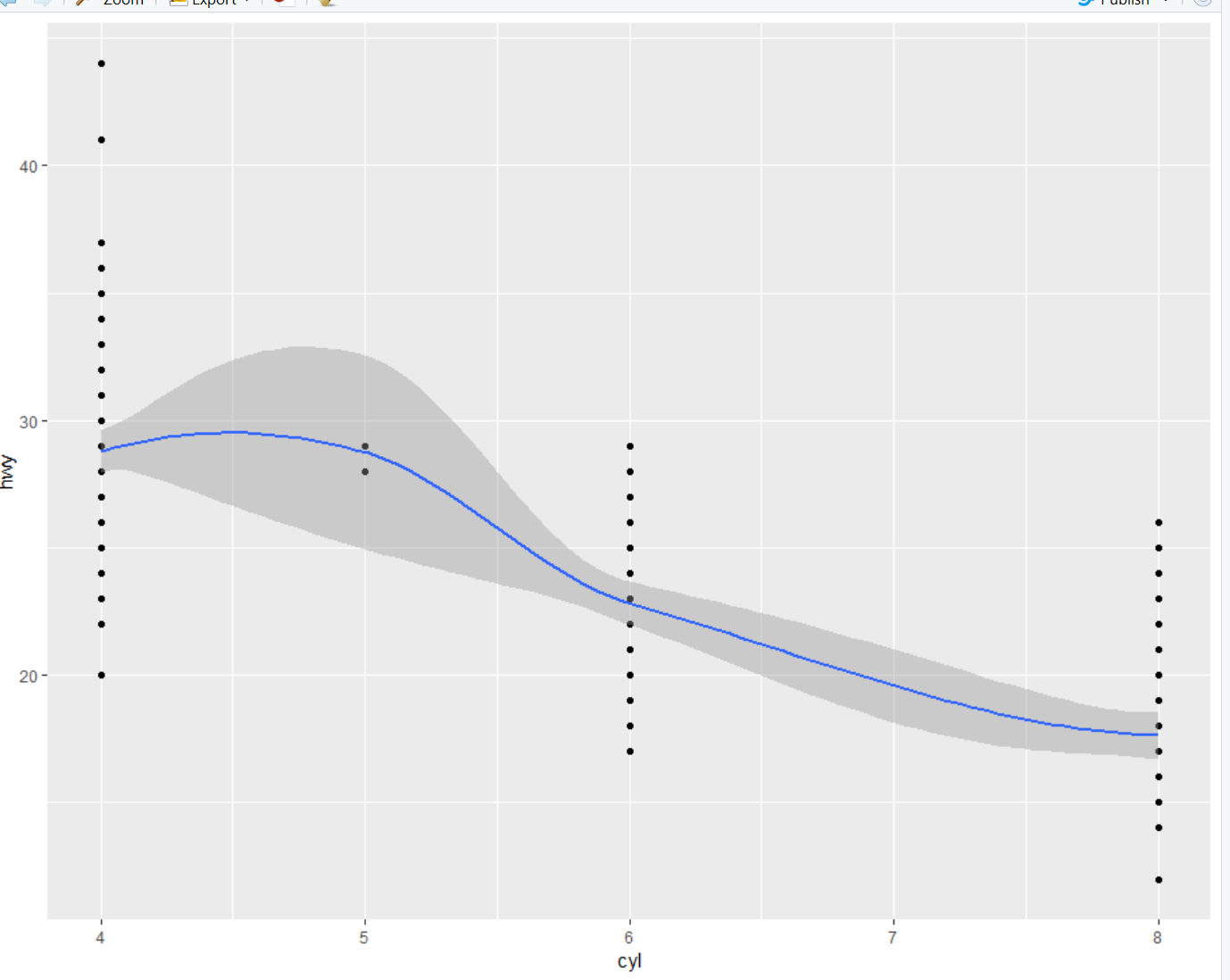


From the data above it is pretty easy to spot that the Civic is the best performing car based on mpg. The way we can tell is by it’s individual box it has each of its models clustered closely to the top right of the graph, insinuating each model has a high city and highway efficiency. Other good performing models are the corolla and also the new beetle had a couple of models that performed fantastic, but overall the mpg of the model wasn’t as high.

### Do cylinders and/or displacement positively or negatively affect highway efficiency?

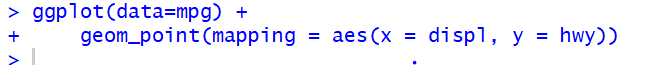
I’ll start with the cylinders effectiveness first

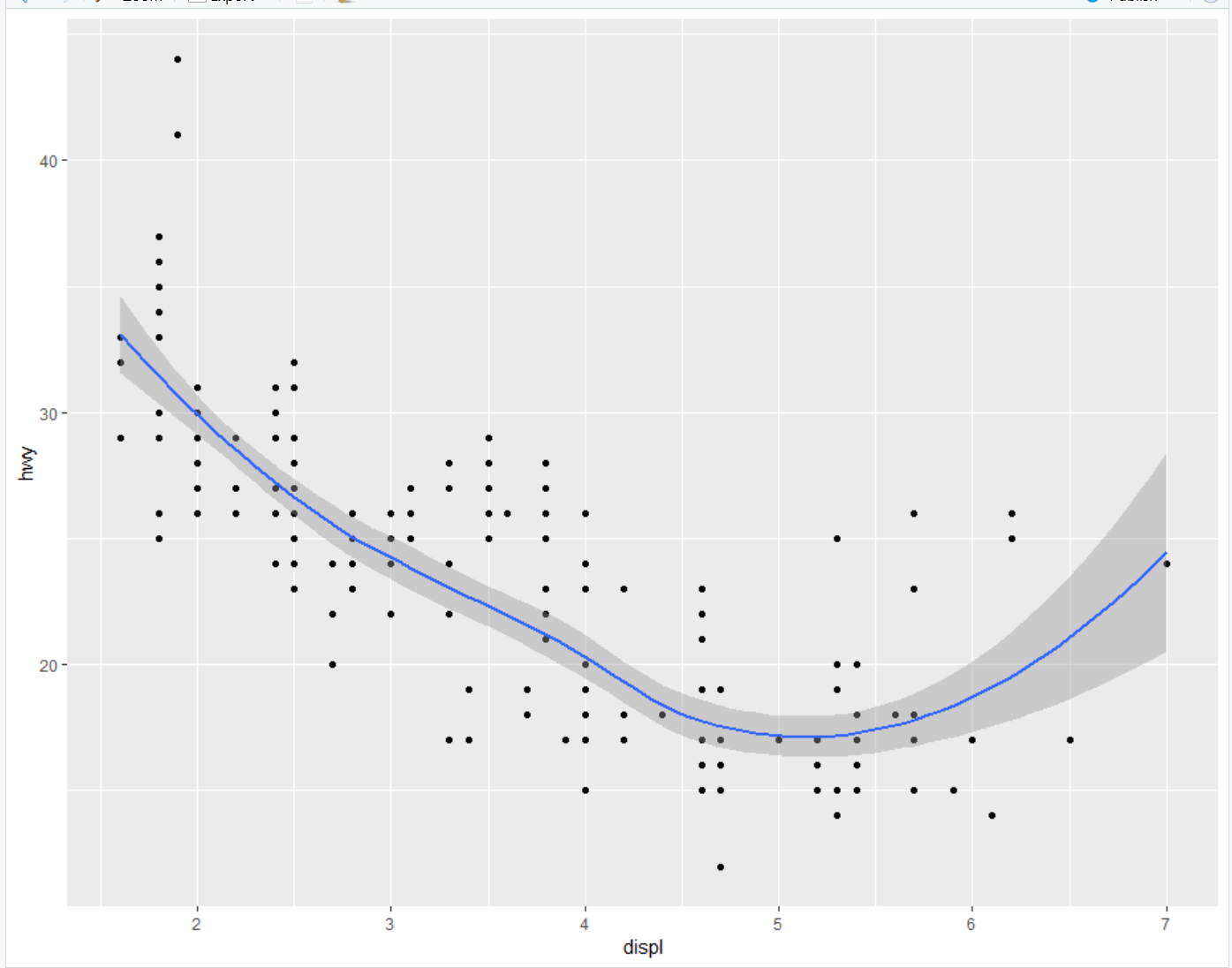




For each cylinder in the engine, it seems like there is a sacrifice in highway mpg’s, but there are still some 4 cylinder engines with some pretty low mpg’s so it is not always the case.

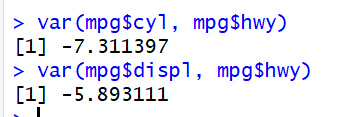
Now I’ll look at displacement and it’s effect on highway mpg’s.





Again for displacement we are seeing a downward trending scatterplot which would suggest the highway efficiency is negatively effected by displacement of the vehicle.

If we also look at the variance of highway efficiency vs both displacement and again versus cylinders we may understand why both graphs are negative trending.



The variance for both is trending which makes a lot more sense why the data is trending negative above. Which leads me to conclude that each the cylinders and displacement on has a negative effect on highway performance as they grow the highway performance generally tends to take a dip in performance.

## Summary

From my EDA I was able to find out all the different things I wanted to find. The information I learned from the questions I asked didn’t surprise me really, but I was hoping that it might. Question one I found it was pretty obvious, pickups and SUV’s have long been the worst performers when it comes to fuel efficiency but I did think it was interesting the lack of city mpg the 2 seater vehicles had, I just figured they were always good cars for moving around anywhere on any kind of road. Question two wasn’t much of a surprise either because civics and corolla’s are well known for their fuel efficiency so it would have been surprising if they didn’t perform well. For question three that also wasn’t much of a surprise, the number of cylinders and displacement usually mean there is less fuel efficiency in the vehicle so seeing data trend negatively and reaffirm that notion just validates that adage about vehicles. Overall, I think it is a really neat dataset because it does have a lot of different parts that can be looked at an analyzed.