PSTAT 231 HW1

2022-10-03

Q1: Supervised learning uses data that include an input and label(output). Using previous data labels, it learns to generalize for future predictions. Unsupervised learning doesn't include the label data and is used to identify clusters with similar patterns within a dataset. The main difference between the two is supervised learning learns patterns to predict unseen data, while unsupervised learning learns patterns to describe trends.

Q2: Regression models learn to predict continuous data, while classification models learn to predict qualitative data.

Q3: Regression: Mean Squared Error, Mean Absolute Error Classification: Precision, Recall

Q4: Descriptive - aims to provide general trends of data and provide the best visualizations to show patterns Inferential - aims to analyze which features are most significant(correlated) for prediction Predictive - aims to predict with least amount of error possible. Find all features that help build a better model.

Q5: Mechanistic uses concepts to tell stories about the real world. Empirical models uses real world events to derive a model.Both share the quality of using data to achieve its end goal. Empirically-driven models are easier to understand, since it is less mathematical in the begining process. With the collection of data, scientists may use their analysis to derive a model. The bias-variance trade off plays a role when trying to tries to not over generalize on one or the other. It tries to use mathematical concepts to generalize stories about the real world.

Q6: The firs question is predictive, since we a measuring some traits to retrieve information about another trait. The second question is inferential, since it considers the change in support by one trait if some other qualities were coinsidered.

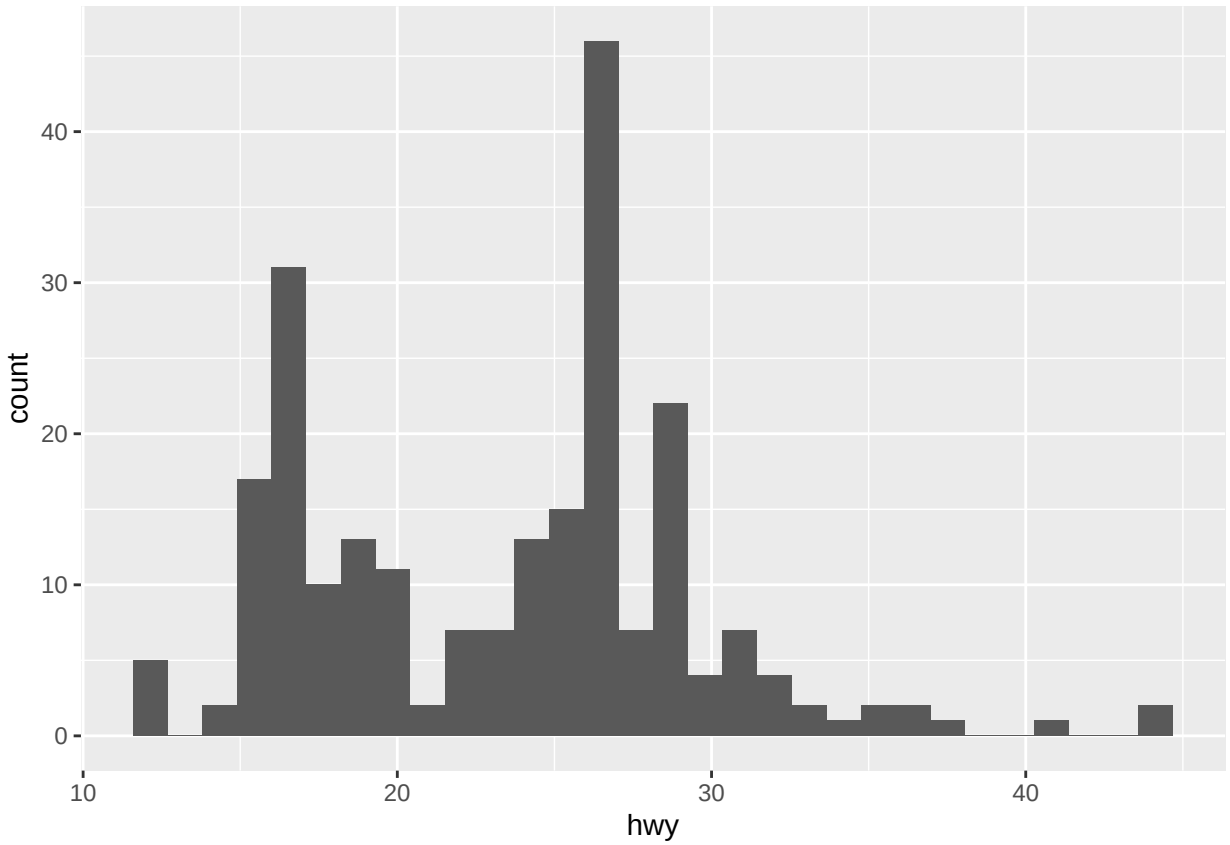
```
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
## v purrr   0.3.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

E1:

```
data(mpg)
ggplot(mpg, aes(x=hwy)) + geom_histogram()
```

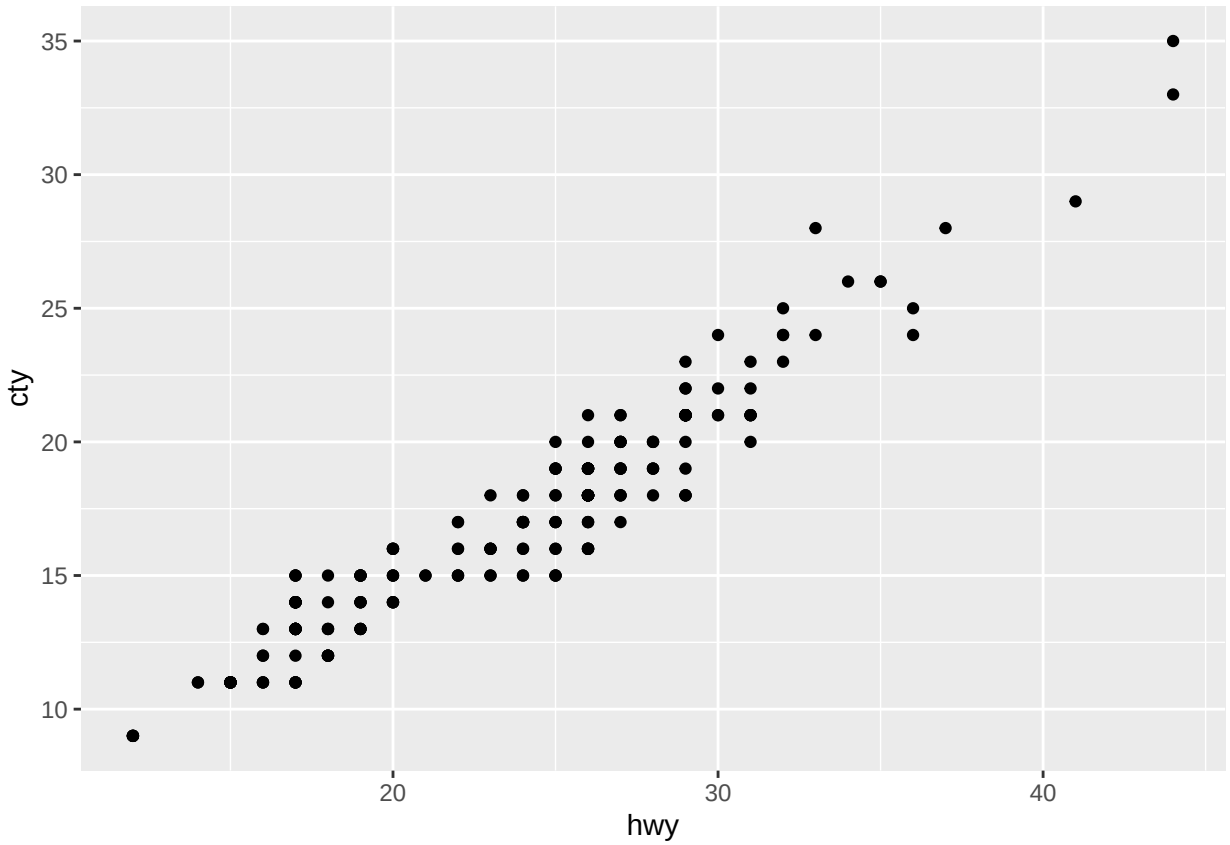
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Most vehicles have around 16 mpg or 26 mpg for highway performance.

E2:

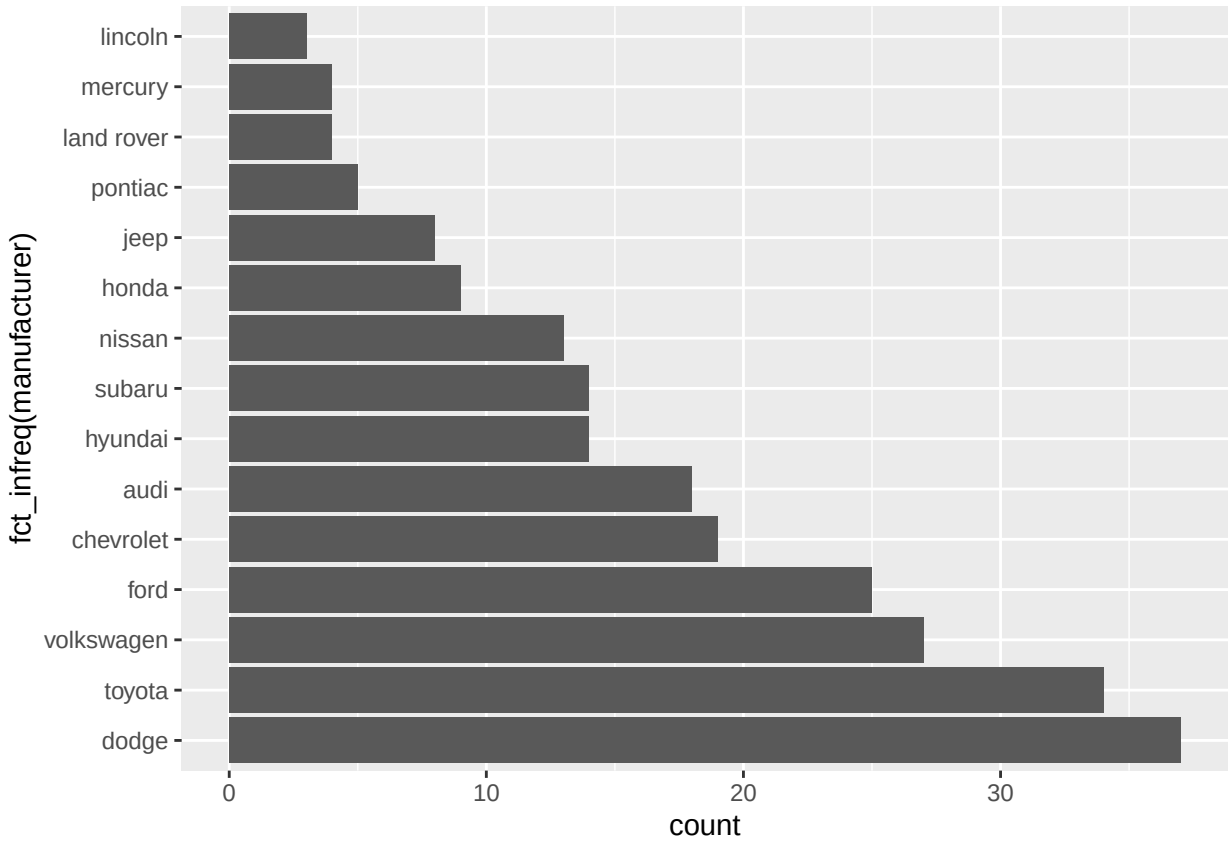
```
ggplot(mpg, aes(x=hwy, y=cty)) + geom_point()
```



There's an overall linear relation between city and highway performance. As one highway performance increases, the city performance goes up as well.

E3:

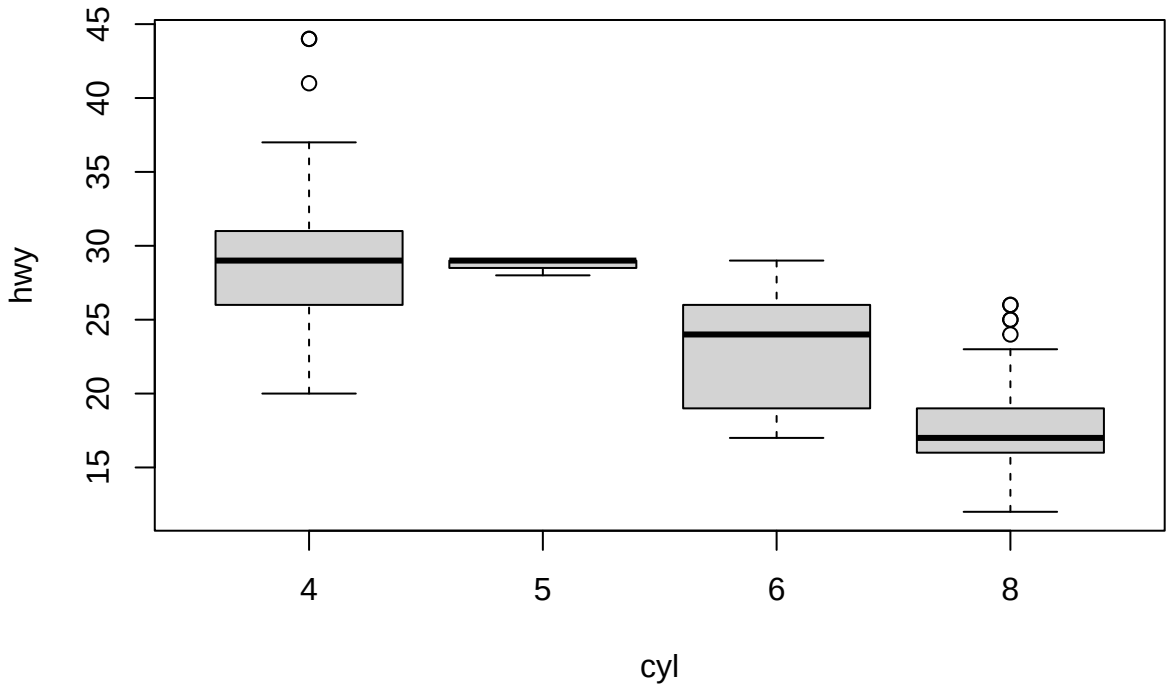
```
# Basic barplot
library(forcats)
p<-ggplot(data=mpg, aes(x=fct_infreq(manufacturer))) +
  geom_bar()
# Horizontal bar plot
p + coord_flip()
```



Dodge prduced the most cars and Lincoln produced the least.

E4:

```
boxplot(hwy~cyl, data=mpg)
```



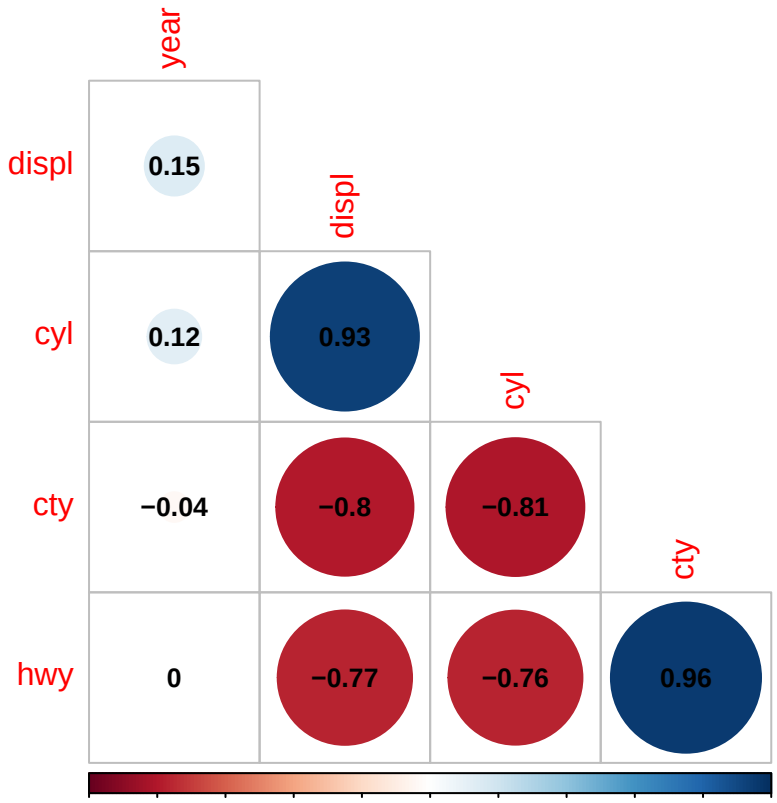
The higher the cylinder, the lower mpg on highway for a vehicle.

E5:

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

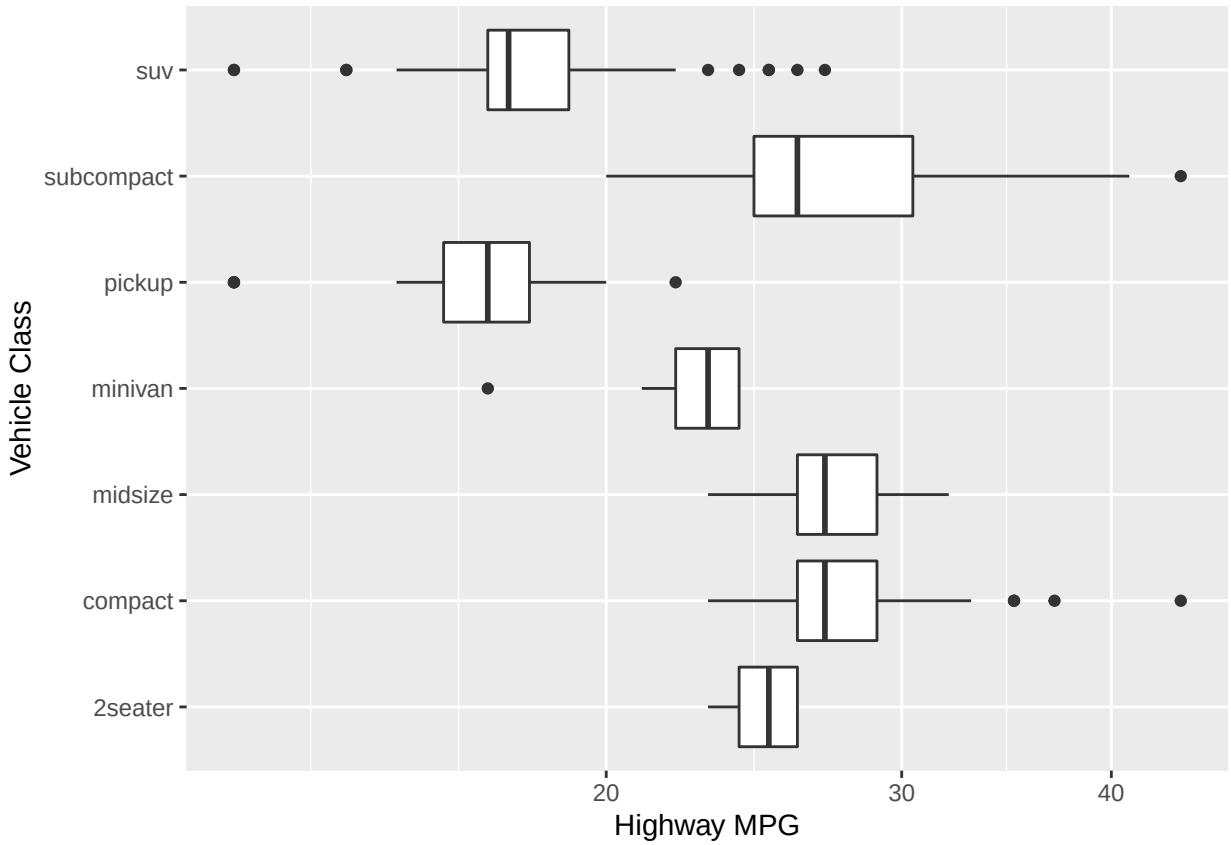
```
library(dplyr)

M <- select_if(mpg, is.numeric)

## leave blank on non-significant coefficient
## add significant correlation coefficients
corrplot(cor(M), method = 'circle', type = 'lower', insig='blank',
         addCoef.col ='black', number.cex = 0.8, order = 'AOE', diag=FALSE)
```



The pairs cyl/disply, cty/cyl,cty/displ, hwy/displ, hwy/cyl, and hwy/cty all hold significant correlations. Cyl/displ and hwy/cty hold psitive correlations, while the rest are negative. The positive correlations make sense. A vehicle's city/highway performance generally improve together, and their differences are only slightly better/worse. If a vehicle has more cylinders, It will carry more volume and hence, higher displacement. I would have assumed city and highway performance would improve with more cylinders/displacement. Those variables generally improve a vehicle performance but now looking back, it also chugs more gas.

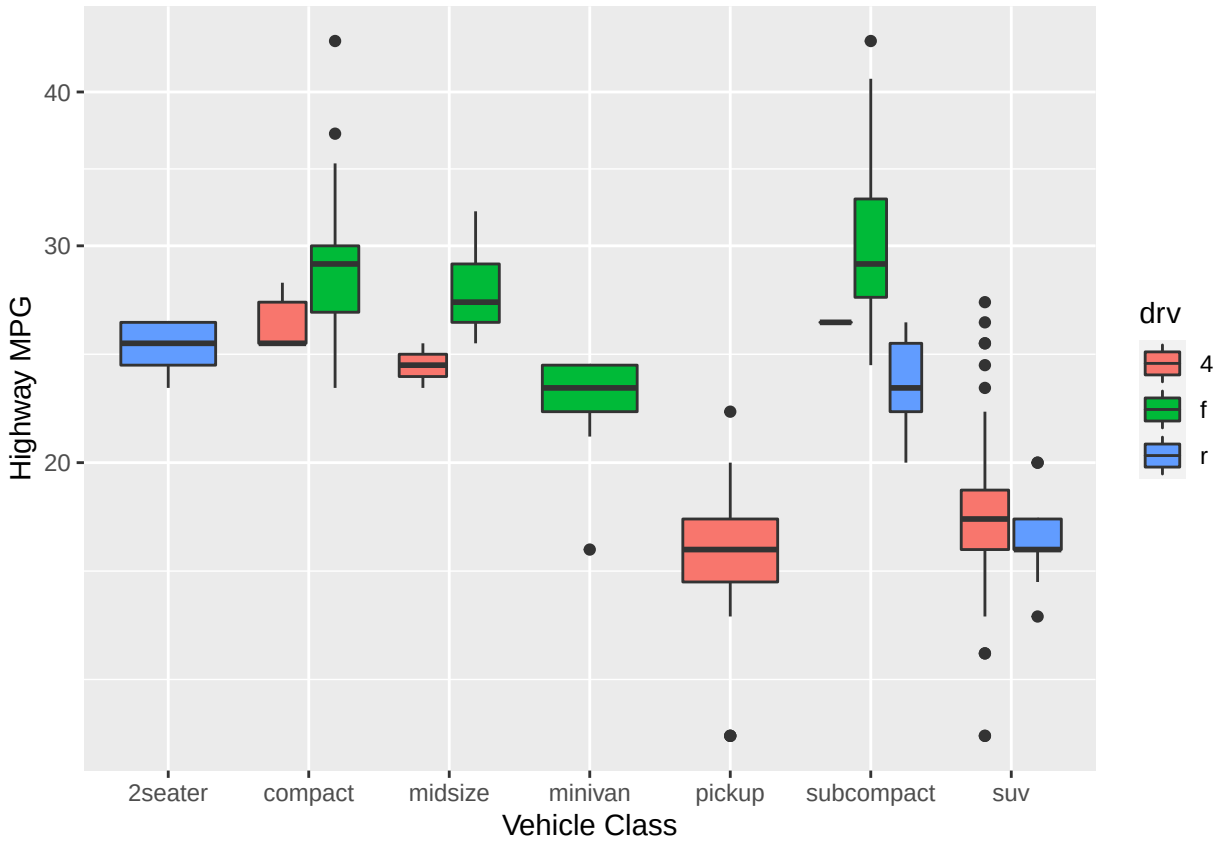
E6:

```
library(ggthemes)
  ggplot(data=mpg, aes(x=class, y=hwy)) +
  geom_boxplot()+
  coord_flip()+
  scale_y_log10()+
  labs(x="Vehicle Class", y="Highway MPG") +  scale_color_gdocs()#+ theme_stata(scheme = "s1mono")
```



E7:

```
ggplot(data=mpg, aes(x=factor(class), y=hwy, fill = drv)) +
geom_boxplot()+
scale_y_log10()+
labs(x="Vehicle Class", y="Highway MPG") + scale_colour_stata(scheme = "s2color")
```

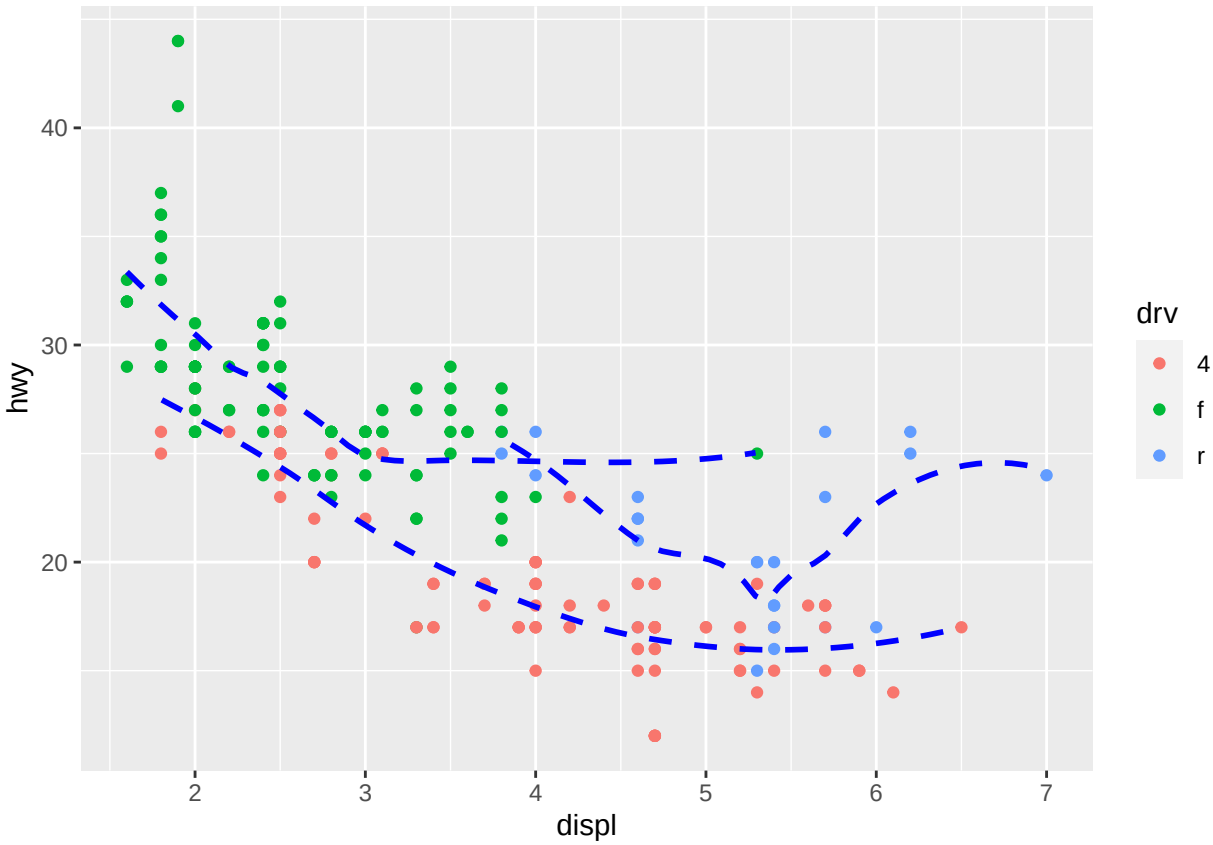


E8:

```
ggplot(mpg, aes(x=displ, y=hwy, color=drv)) +
  geom_point() +
  geom_smooth(aes(group=drv), method=loess, se=FALSE, fullrange=TRUE, linetype='dashed', color='blue')
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 69 rows containing missing values (geom_smooth).
```


hwy
40
30
20
2
3
4
5
6
7
displ
drv
4
f
r