Introduction to the GGPLOT Package

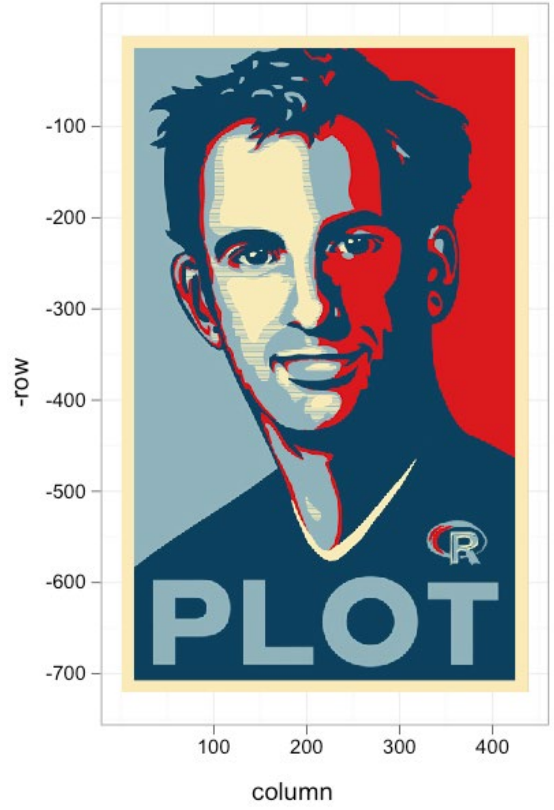
School of Information Studies
Syracuse University

Hadley Wickham

Chief scientist at R-Studio and self-described “data nerd”

GG: Grammar of Graphics

GGPLOT2 released in 2007, has more than 2 million downloads, most popular of the packages he wrote



ggplot2

ggplot2 divides plot into three different fundamental parts:

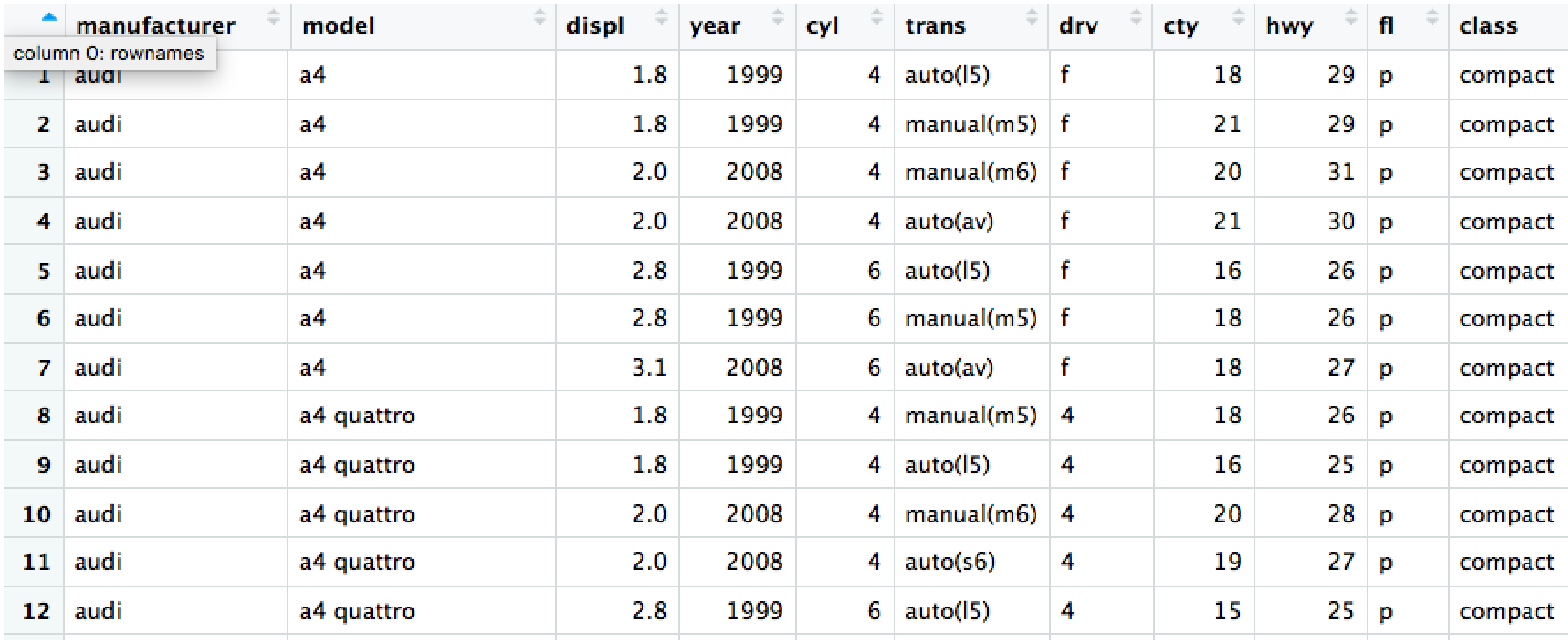
Plot = Data + Aesthetics + Geometry

Every plot can be defined as follows:

- **Data** is a dataframe.
- **Aesthetics** is used to indicate x and y variables. It can also be used to control the color, the size or the shape of points, the height of bars, etc.
- **Geometry** defines the type of display (histogram, box plot, line plot, density plot, dot plot, etc.)

Work With the Built-In MPG Data

View(mpg)



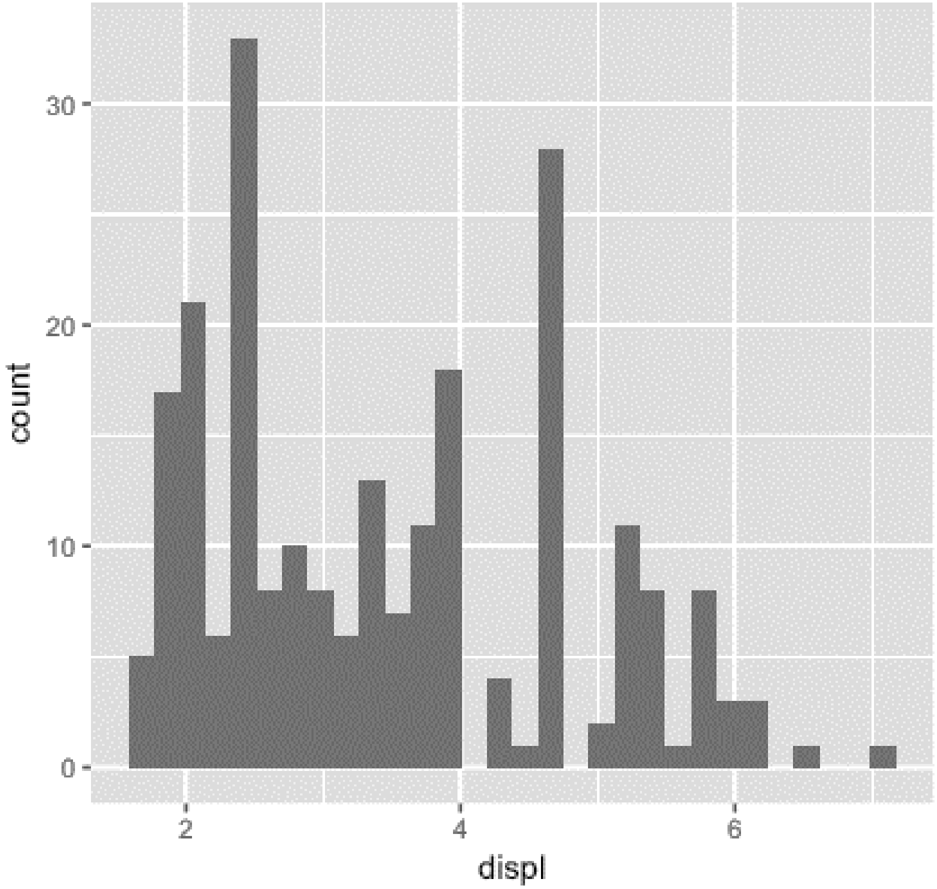
column 0: rownames

	manufacturer	model	displ	year	cyl	trans	drv	cty	hwy	fl	class
1	audi	a4	1.8	1999	4	auto(l5)	f	18	29	p	compact
2	audi	a4	1.8	1999	4	manual(m5)	f	21	29	p	compact
3	audi	a4	2.0	2008	4	manual(m6)	f	20	31	p	compact
4	audi	a4	2.0	2008	4	auto(av)	f	21	30	p	compact
5	audi	a4	2.8	1999	6	auto(l5)	f	16	26	p	compact
6	audi	a4	2.8	1999	6	manual(m5)	f	18	26	p	compact
7	audi	a4	3.1	2008	6	auto(av)	f	18	27	p	compact
8	audi	a4 quattro	1.8	1999	4	manual(m5)	4	18	26	p	compact
9	audi	a4 quattro	1.8	1999	4	auto(l5)	4	16	25	p	compact
10	audi	a4 quattro	2.0	2008	4	manual(m6)	4	20	28	p	compact
11	audi	a4 quattro	2.0	2008	4	auto(s6)	4	19	27	p	compact
12	audi	a4 quattro	2.8	1999	6	auto(l5)	4	15	25	p	compact

Layering ggplot Specifications: Univariate Display

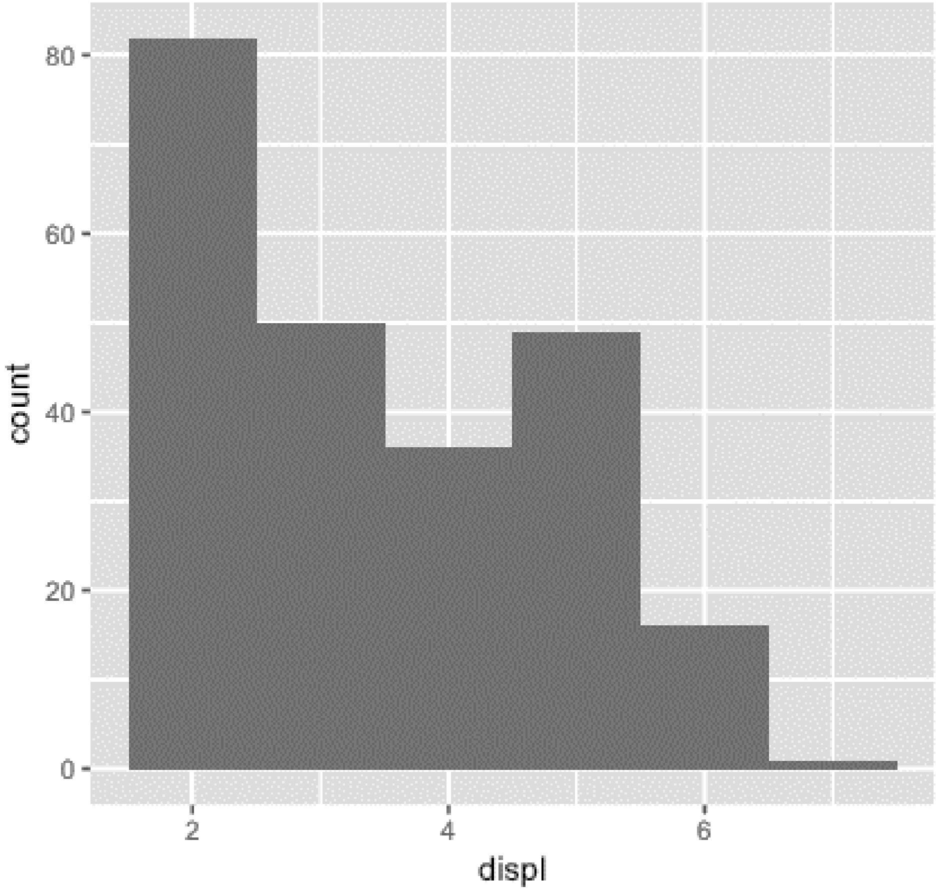
```
#The data
myPlot <- ggplot(mpg)

#The aesthetic
myPlot <- myPlot + aes(x=displ)

#The geometry
myPlot <- myPlot + geom_histogram()

#Invoke the plot to draw it
myPlot
```

Console Reports: `stat_bin()` Using `bins = 30.` Pick Better Value With `binwidth`



Univariate Display: Control binwidth

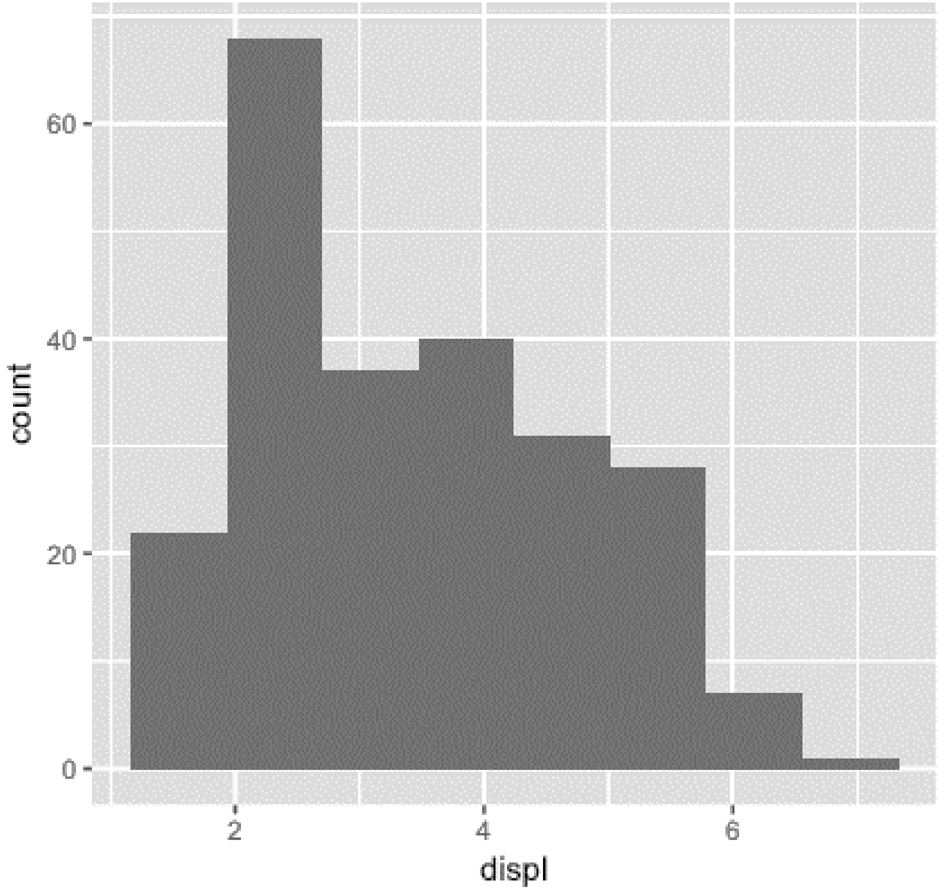
```
#The data
myPlot <- ggplot(mpg)

#The aesthetic
myPlot <- myPlot + aes(x=displ)

#The geometry
myPlot <- myPlot + geom_histogram(binwidth=1)

#Invoke the plot to draw it
myPlot
```



School of Information Studies
Syracuse University

Univariate Display: Set Bin Count

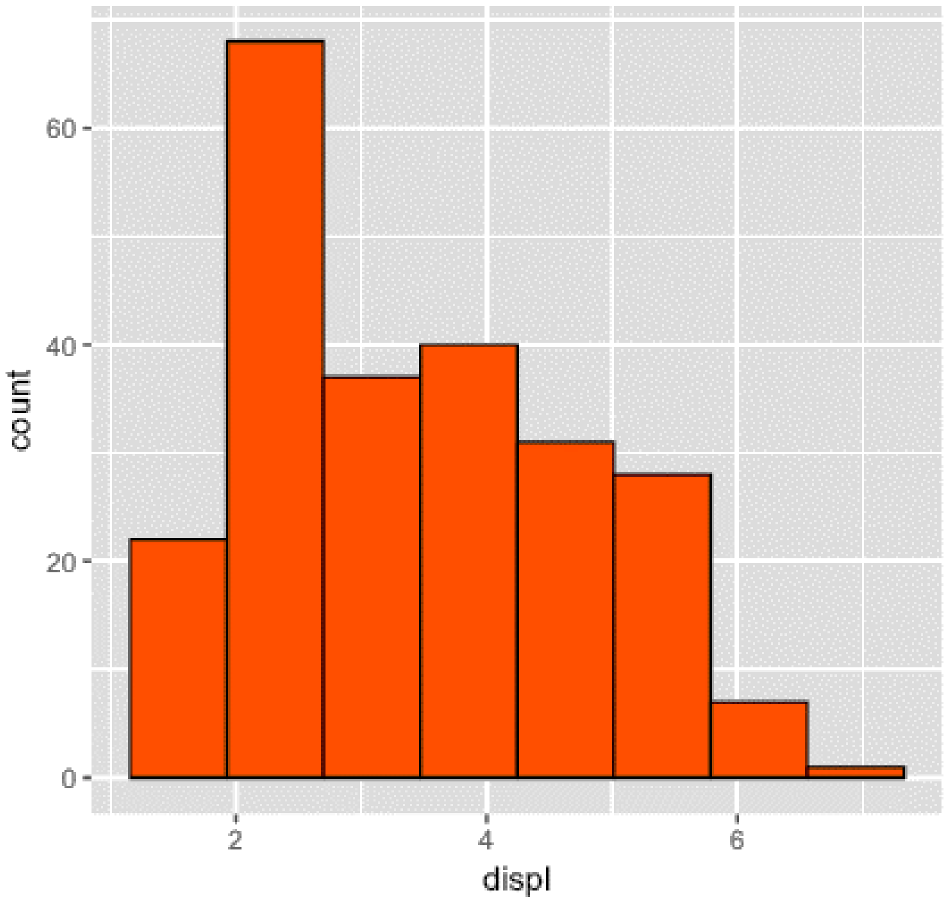
```
#The data
myPlot <- ggplot(mpg)

#The aesthetic
myPlot <- myPlot + aes(x=displ)

#The geometry
myPlot <- myPlot + geom_histogram(bins=8)

#Invoke the plot to draw it
myPlot
```



School of Information Studies
Syracuse University

Change Colors

```
#The data
myPlot <- ggplot(mpg)

#The aesthetic
myPlot <- myPlot + aes(x=displ)

#The geometry
myPlot <- myPlot +
          geom_histogram(bins=8,
             fill="red",col="black")

#Invoke the plot to draw it
myPlot
```



Note ggplot Code Can Fit on One Line

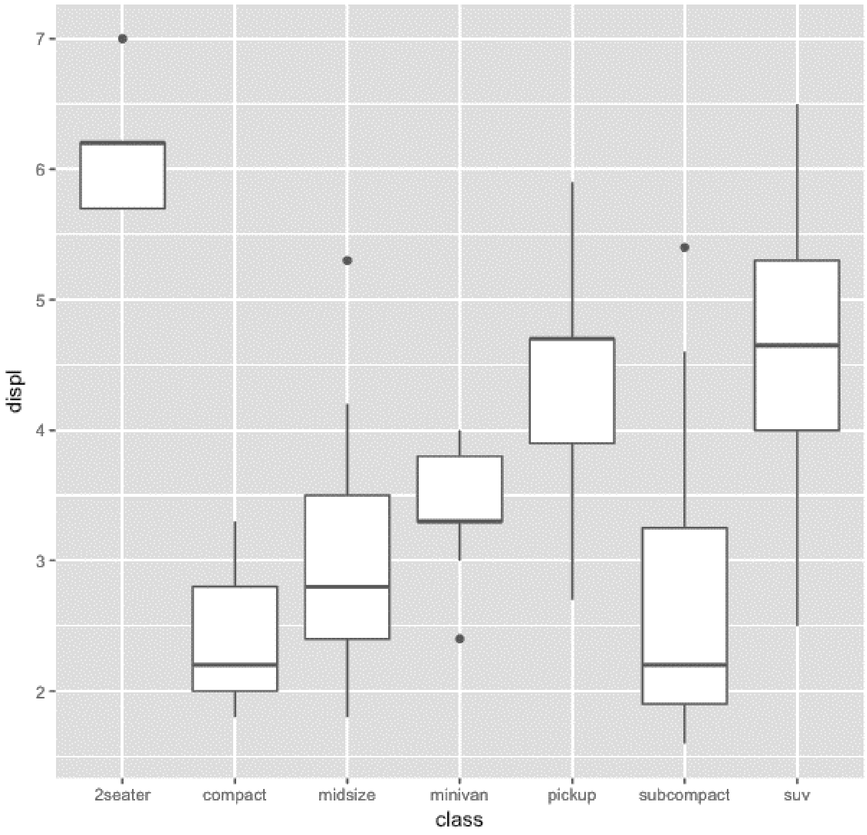
```
#Rather than storing the plot specs and building up piece by piece, you can combine everything into one command line.
ggplot(mpg)+aes(x=displ) + geom_histogram(bins=8, fill="red", col="black")
```

Box Plot

Make a boxplot of:
#displ (y-variable)

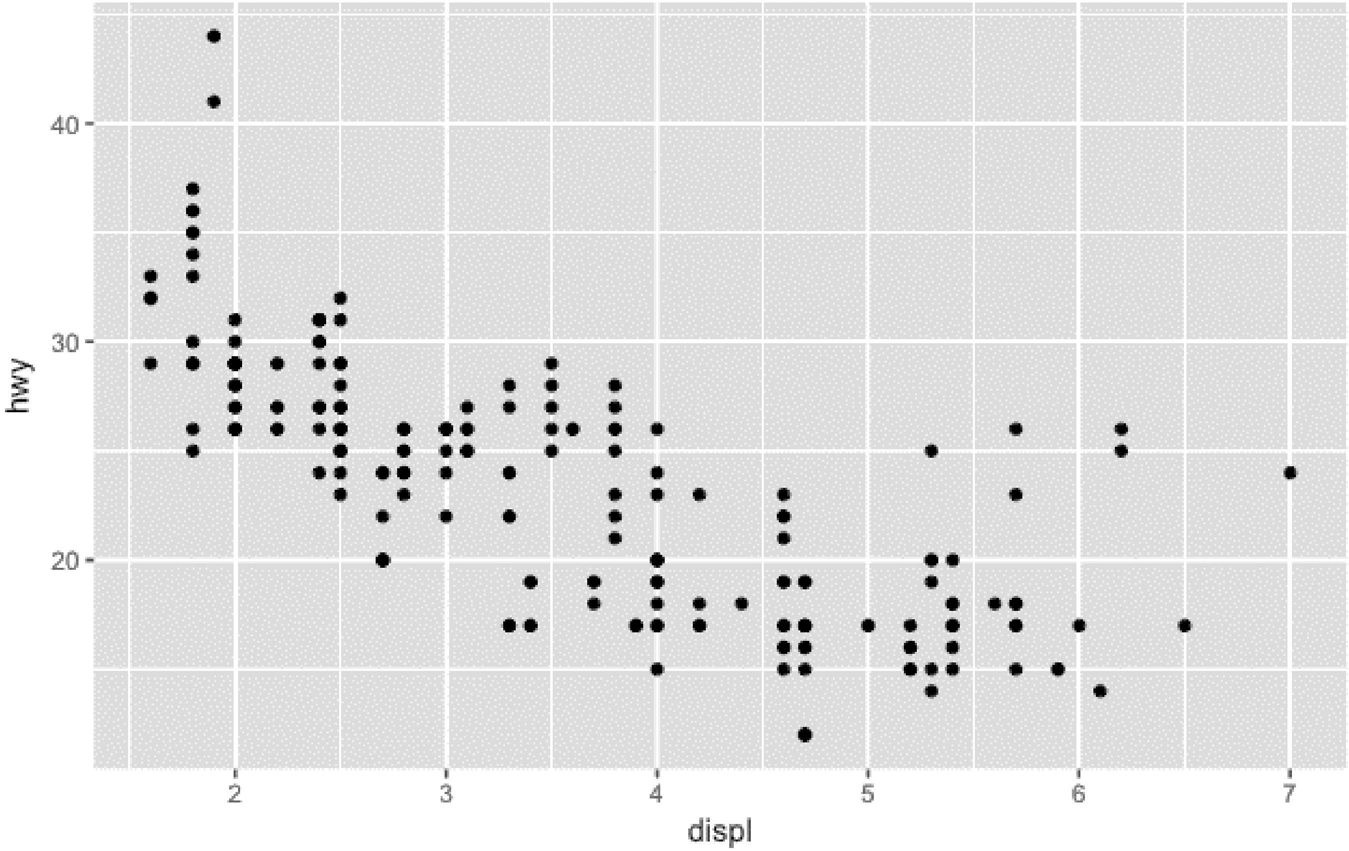
#cars class (x-variable)

```
ggplot(mpg) +
   aes(x=class,y=displ) +
   geom_boxplot()
```



Exploring More Than One Attribute With Scatter Plots

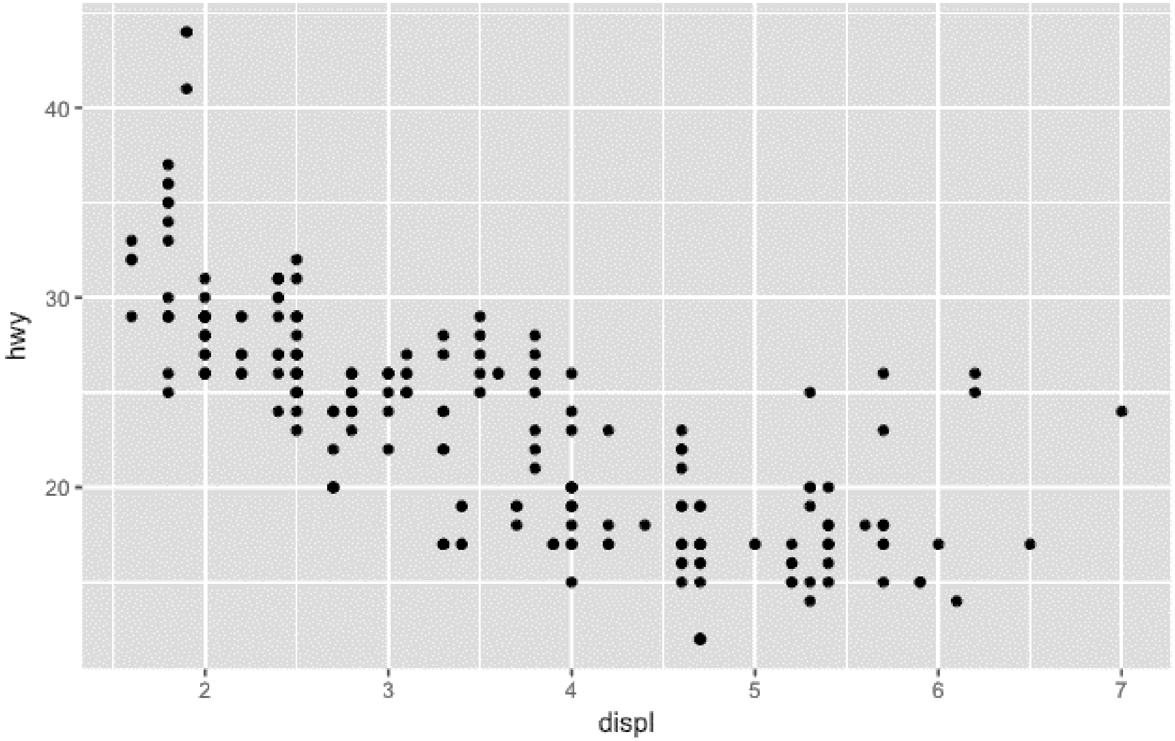
```
#The data
myPlot <- ggplot(mpg)

#The aesthetic
myPlot <- myPlot + aes(x=displ,y=hwy)

#The geometry
myPlot <- myPlot + geom_point()

#Invoke the plot to draw it
myPlot
```



Using Tidyverse Style

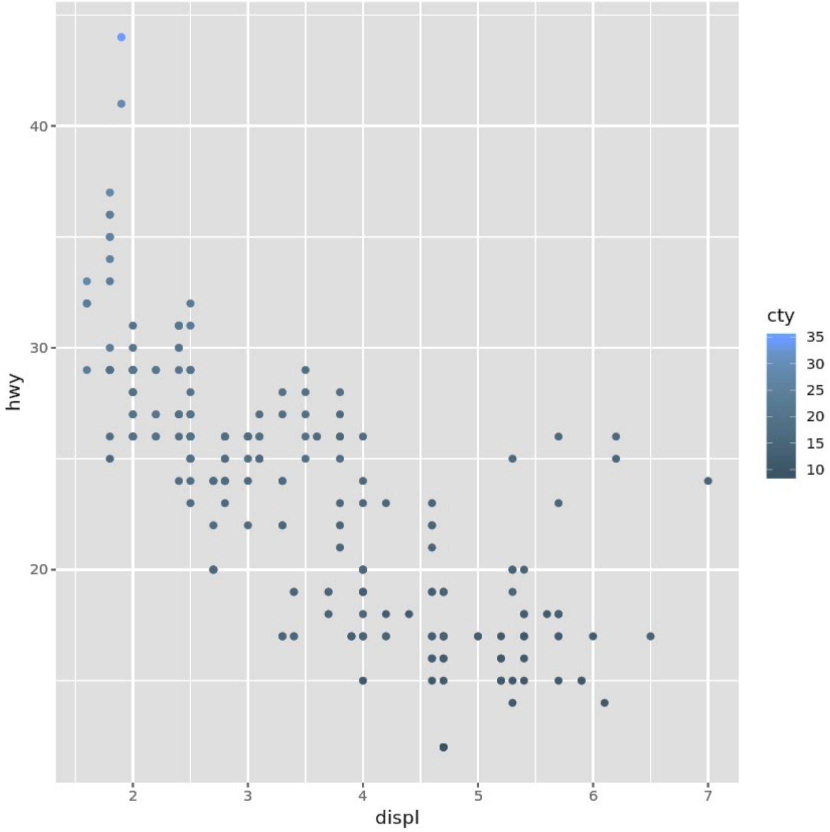
#Using Pipes and the '+' to generate a scatter plot

```
mpg %>%
  ggplot() +
    aes(x=displ,y=hwy) +
    geom_point()
```



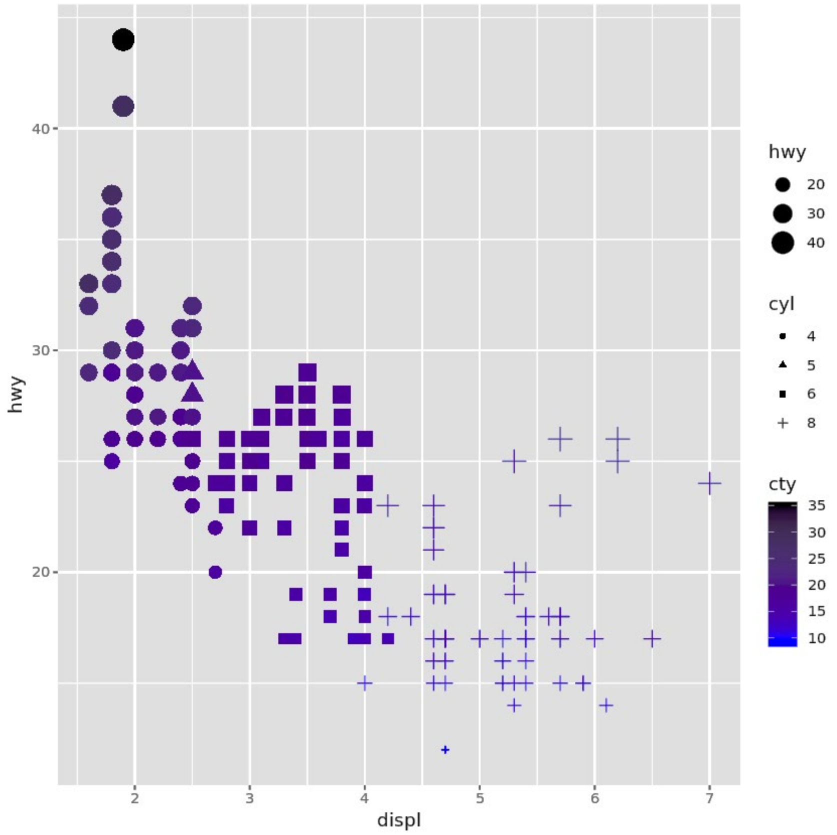
Adding Color

```
#Fill the points
mpg %>%
  ggplot() + aes(x=displ,y=hwy) +
  geom_point(aes(color = cty))
```



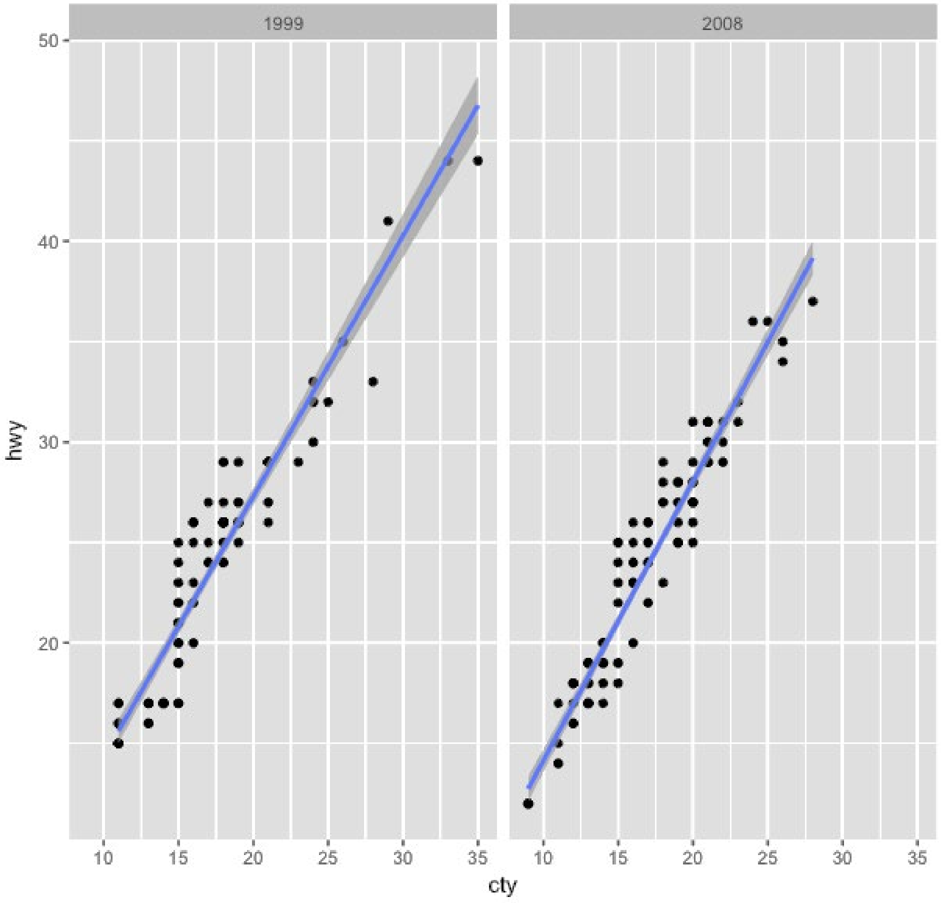
Adding More Attributes

```
#Why the need to mutate?
mpg %>%
  mutate(cyl=as.factor(cyl)) %>%
  ggplot() + aes(x=displ,y=hwy) +
  geom_point(aes(color = cty,
           shape=cyl, size=hwy)) +
  scale_color_gradient(low = "blue",
                    high = "black")
```



Dual Scatterplot With Fitted Line

```
myPlot <- ggplot(mpg)
#The aesthetic: city versus highway mpg
myPlot <- myPlot + aes(x=cty,y=hwy)
#The geometry: Points/Scatterplot
myPlot <- myPlot + geom_point()

#Compare two years with “facets” and best line fit
myPlot <- myPlot + facet_wrap(~year)
myPlot <- myPlot + geom_smooth(method="lm")

#Invoke the plot to draw it
myPlot
```



Your Turn!

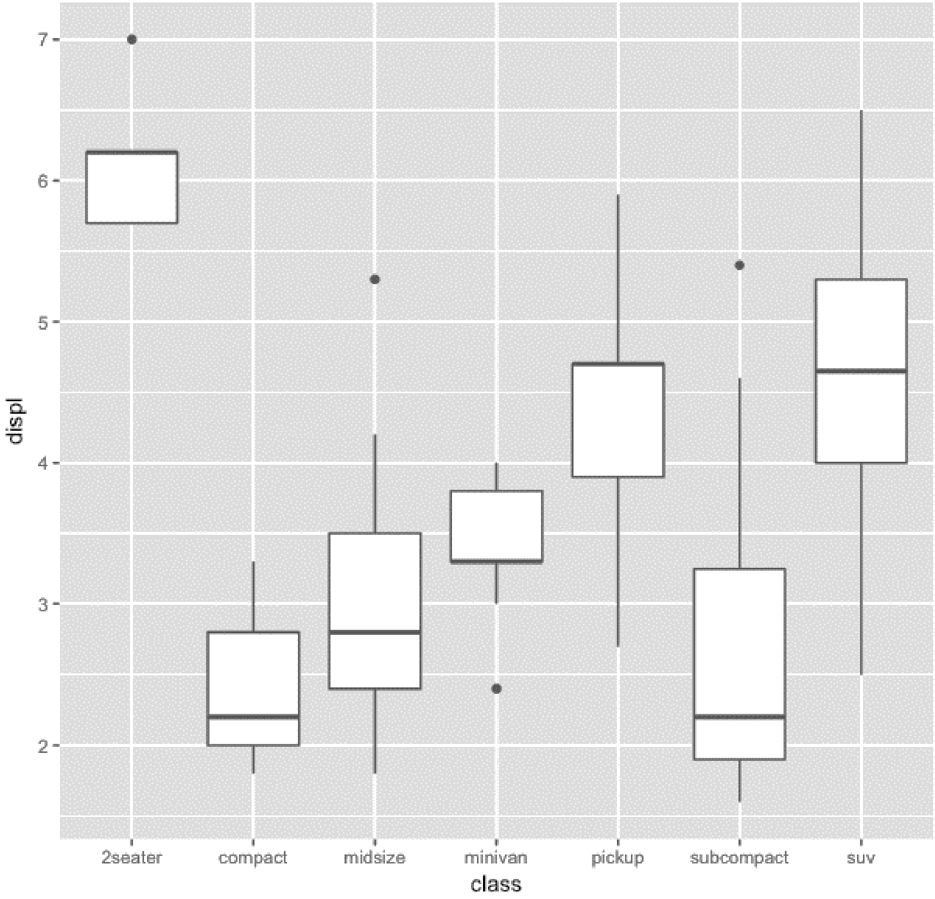
Open laptop, start R-studio, and install library ggplot2.

Make a boxplot of **displ** (y-variable) comparing cars by **class** (x-variable).

Important reminder: geom_boxplot() is generally a bivariate plot, with a “factor” (grouping) variable on the x-axis.

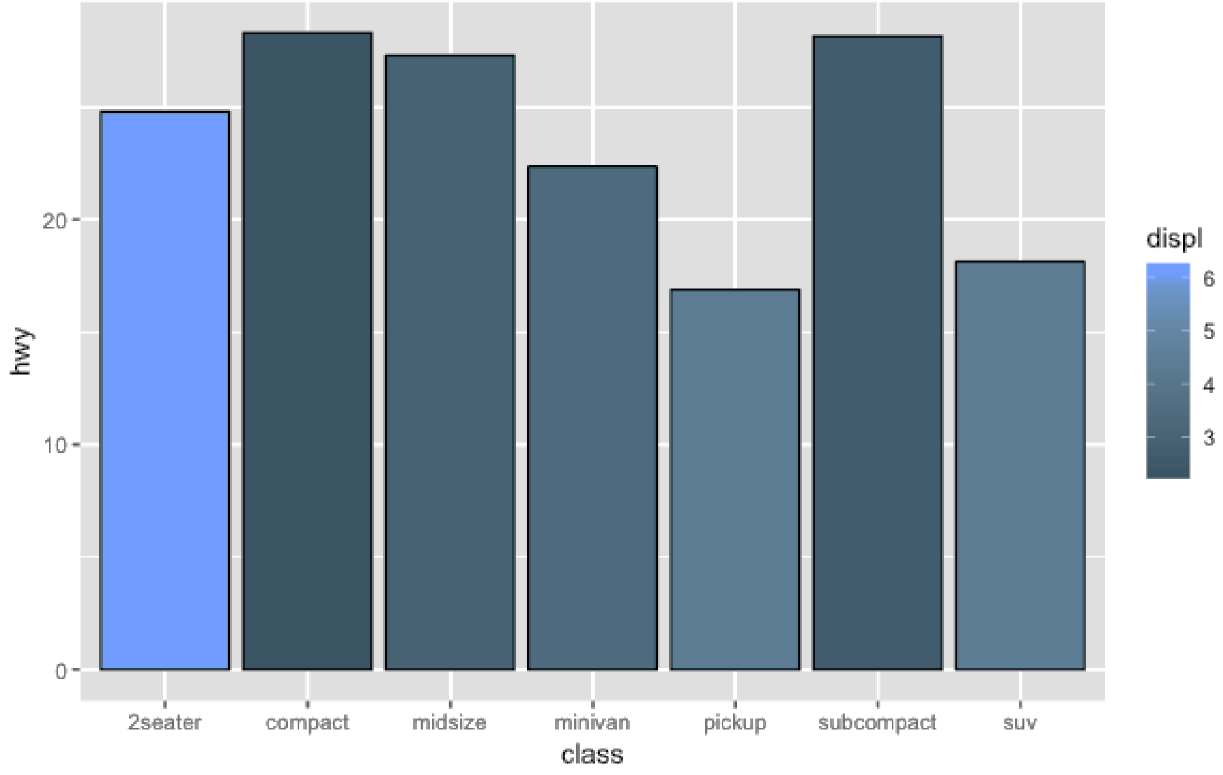
- Include both x and y in your aes() statement

Boxplot of displ by Class



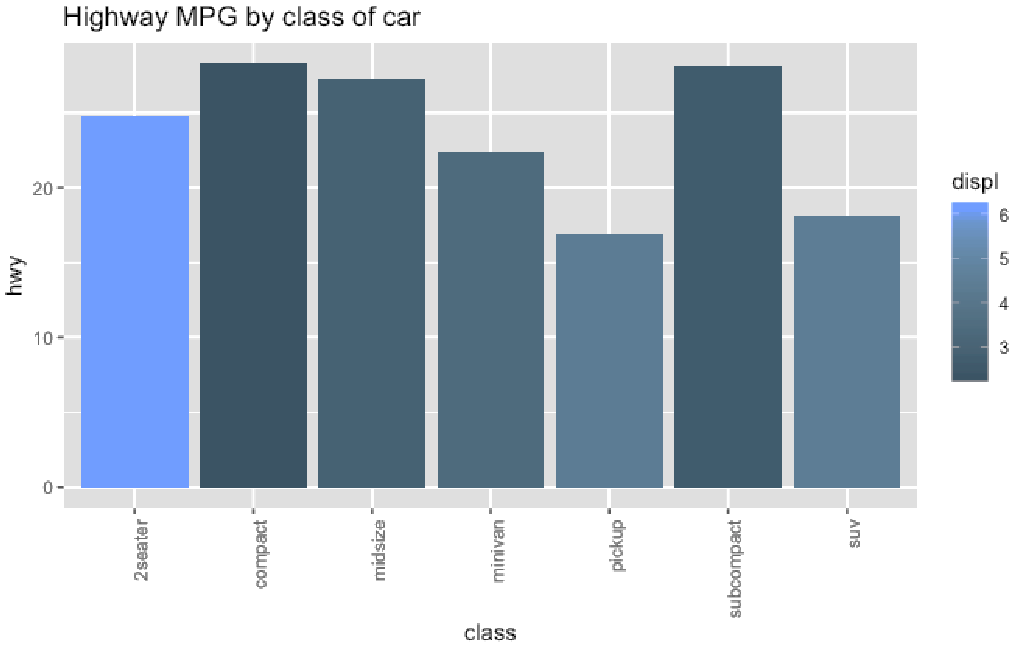
Bar Charts

```
#Use group_by to create data for bars

mpg %>%
      group_by(class) %>%
             summarize(hwy=mean(hwy),
                         displ=mean(displ))   %>%
       ggplot(aes(x = class, y=hwy))  +
             geom_col(color="black",
             aes(fill=displ))
```



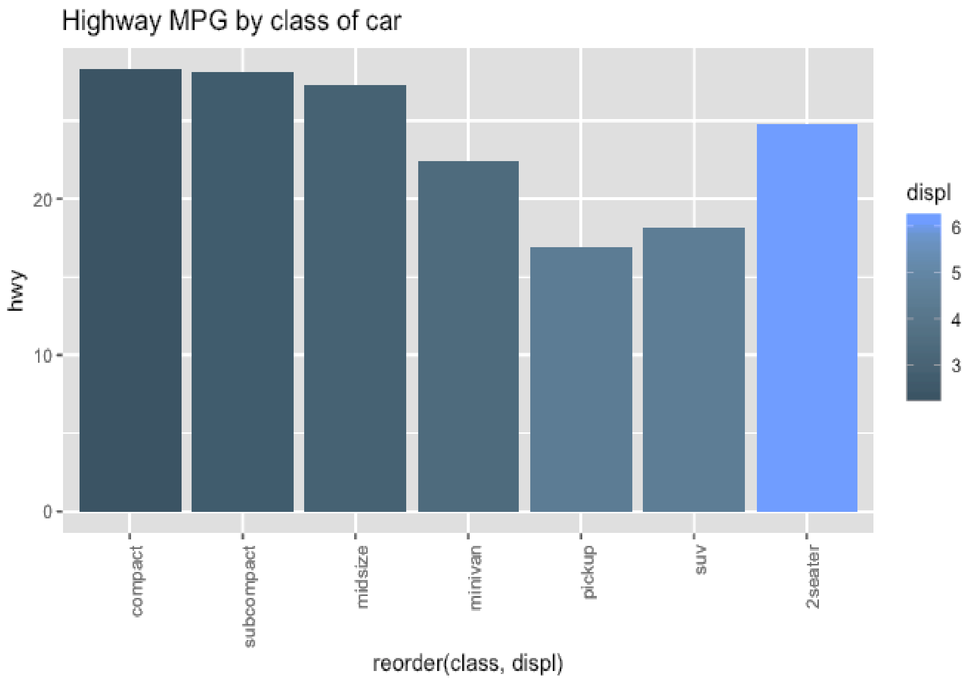
Bar Charts: Rotate Text

```
#Use theme to rotate text; also define title

mpg %>%
  group_by(class) %>%
  summarize(hwy=mean(hwy), displ=mean(displ))
%>%
     ggplot(aes(x = class, y=hwy))  +
          geom_col(aes(fill=displ)) +
         theme(axis.text.x =
               element_text(angle = 90, hjust = 1)) +
          ggtitle("Highway MPG by class of car")
```



Bar Charts: Reorder the Columns

```
#Use 'reorder'


mpg %>%
     group_by(class) %>%
     summarize(hwy=mean(hwy), displ=mean(displ))
%>%
     ggplot(aes(x = reorder(class, displ), y=hwy))  +
          geom_col(aes(fill=displ)) +
          theme(axis.text.x =
                   element_text(angle = 90, hjust = 1)) +
           ggtitle("Highway MPG by class of car")
```



School of Information Studies
Syracuse University

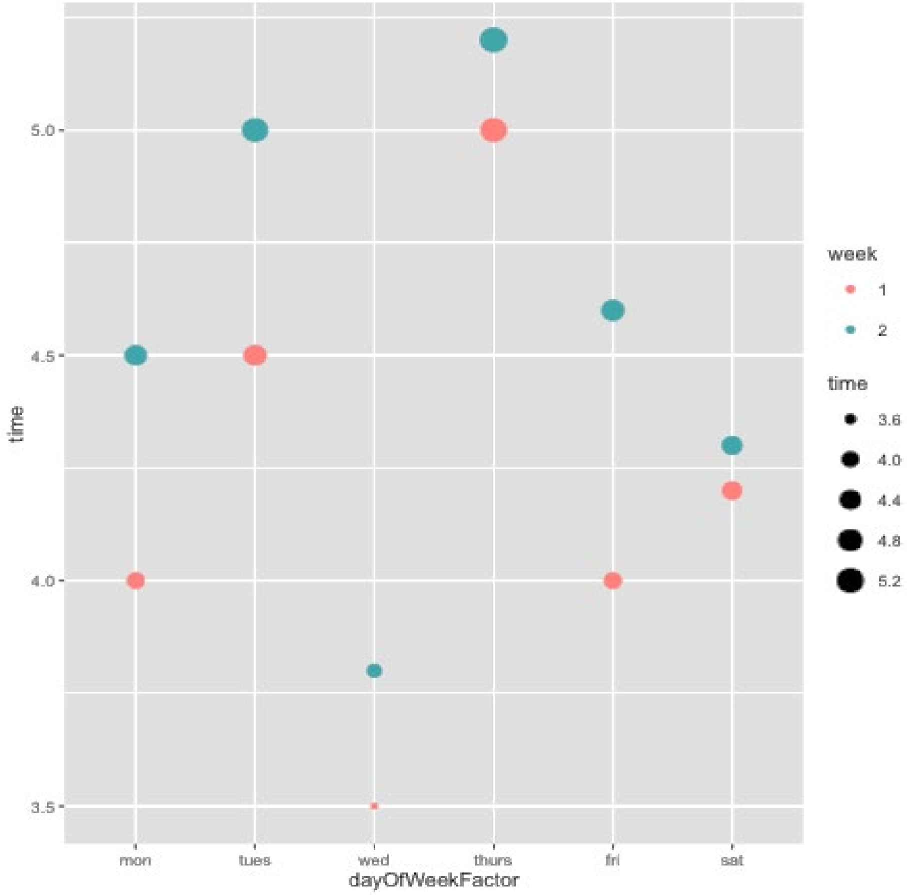
ggplot2: Timed Task

Define ‘travel.df’ to be the dataset →

	dayOfWeekFactor	time	week
1	mon	4.0	1
2	tues	4.5	1
3	wed	3.5	1
4	thurs	5.0	1
5	fri	4.0	1
6	sat	4.2	1
7	mon	4.5	2
8	tues	5.0	2
9	wed	3.8	2
10	thurs	5.2	2
11	fri	4.6	2
12	sat	4.3	2

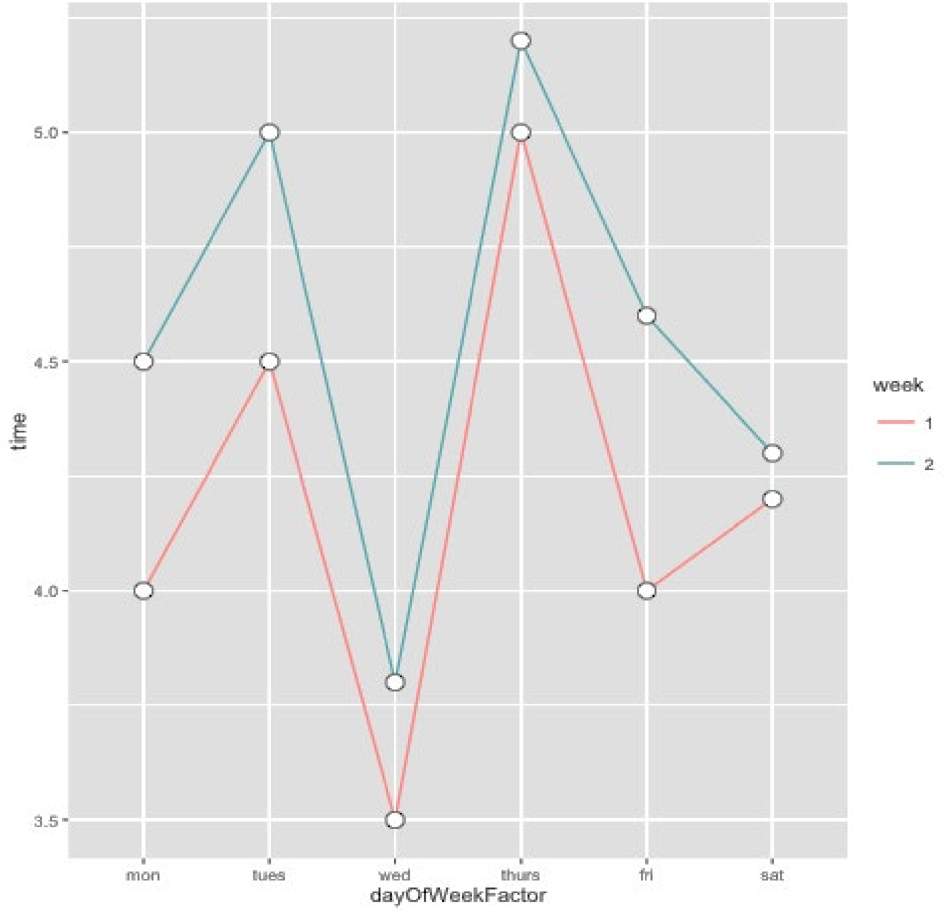
Show Points via a Scatter Plot

```
travel.df  %>%
   ggplot(
      aes(x=dayOfWeekFactor, y=time)) +
      geom_point(aes(size = time, color=week))
```



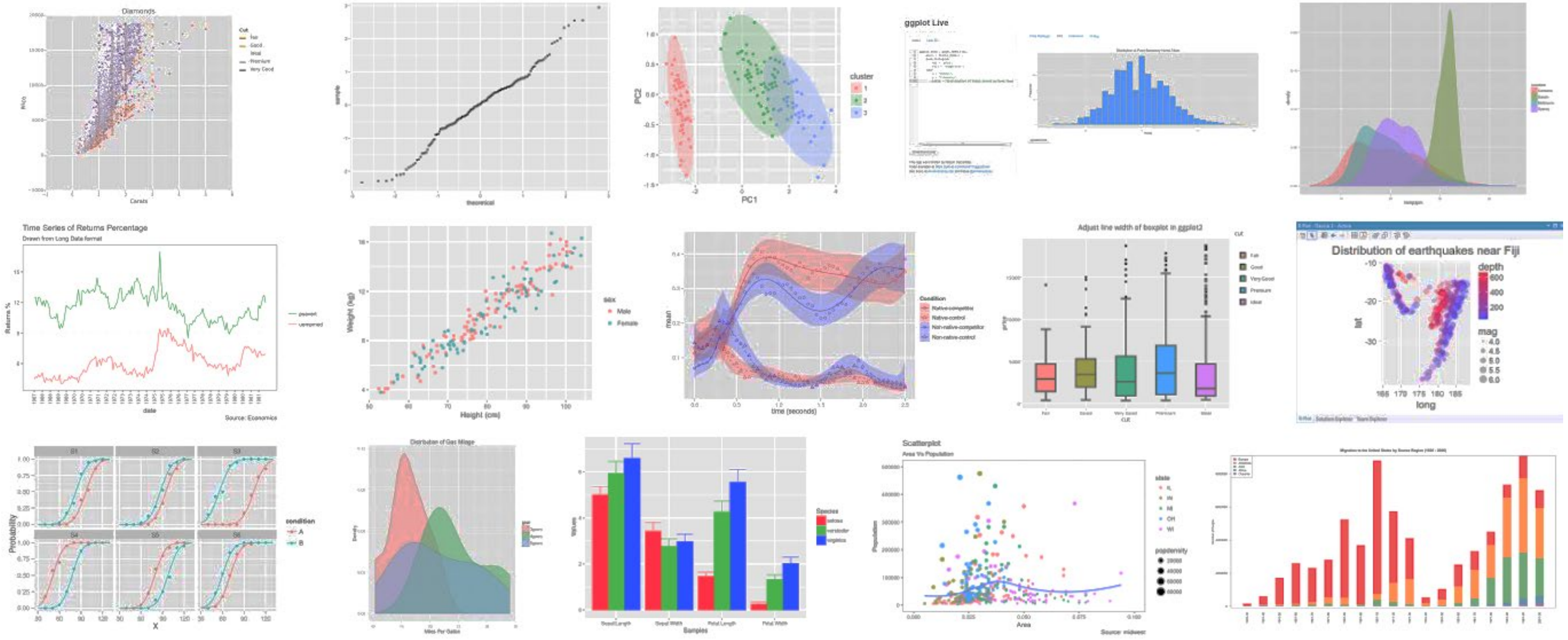
Show Line Plots

```
g <- ggplot(travel.df, aes(x = dayOfWeekFactor,
                          group=week, color=week)) +
          geom_line(aes(y = time))

g <- g + geom_point(y=time, colour="black",
           size=4, shape=21, fill="white”)

g <- g + ylab("time to NYC (in hours)") +
         ggtitle("compare weekly times")

g
```



Unexplored Potential

Using color and marker shapes; time series; labels; titling; grouping; mapping (next week)


Diamonds
Carats
PC1
PC2
cluster
ggplot Live
Time Series of Returns Percentage
Drawn from Long Data format
date
Source: Economics
Height (cm)
Weight (kg)
sex
Male
Female
time (seconds)
mean
Condition
Adjust line width of boxplot in ggplot2
cut
Distribution of earthquakes near Fiji
lat
long
depth
mag
Probability
condition
X
Distribution of Gas Milage
Values
Samples
Species
Scatterplot
Area Vs Population
Population
Area
Source: midwest