Advanced Statistics HW1

Due date: October 9, 2018

Exercises 1

1. Run `ggplot(data = mpg)`. What do you see?
2. Make a scatterplot of `hwy` vs `cyl`.
3. What happens if you make a scatterplot of `class` vs `drv`? Why is the plot not useful?

Exercises 2

1. What's gone wrong with this code? Why are the points not blue?

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, color = "blue"))
```

2. Which variables in `mpg` are categorical? Which variables are continuous? (Hint: type `?mpg` to read the documentation for the dataset). How can you see this information when you run `mpg`?
3. Map a continuous variable to `color`, `size`, and `shape`. How do these aesthetics behave differently for categorical vs. continuous variables?
4. What happens if you map the same variable to multiple aesthetics?
5. What does the `stroke` aesthetic do? What shapes does it work with? (Hint: use `?geom_point`)
6. What happens if you map an aesthetic to something other than a variable name, like `aes(colour = displ < 5)`?

Exercises 3

1. What happens if you facet on a continuous variable?
2. What do the empty cells in plot with `facet_grid(drv ~ cyl)` mean? How do they relate to this plot?

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = drv, y = cyl))
```

3. What plots does the following code make? What does `.` do?

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy)) +
  facet_grid(drv ~ .)

ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy)) +
  facet_grid(. ~ cyl)
```

4. Take the first faceted plot in this section:

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy)) +
  facet_wrap(~ class, nrow = 2)
```

What are the advantages to using faceting instead of the colour aesthetic? What are the disadvantages? How might the balance change if you had a larger dataset?

5. Read `?facet_wrap`. What does `nrow` do? What does `ncol` do? What other options control the layout of the individual panels? Why doesn't `facet_grid()` have `nrow` and `ncol` arguments?
6. When using `facet_grid()` you should usually put the variable with more unique levels in the columns. Why?

Exercises 4

1. What geom would you use to draw a line chart? A boxplot? A histogram? An area chart?
2. Run this code in your head and predict what the output will look like. Then, run the code in R and check your predictions.

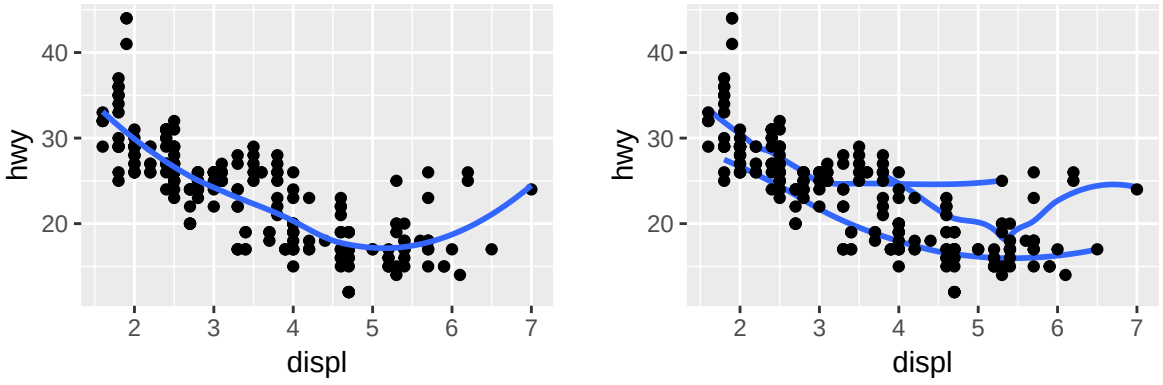
```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy, color = drv)) +
  geom_point() +
  geom_smooth(se = FALSE)
```

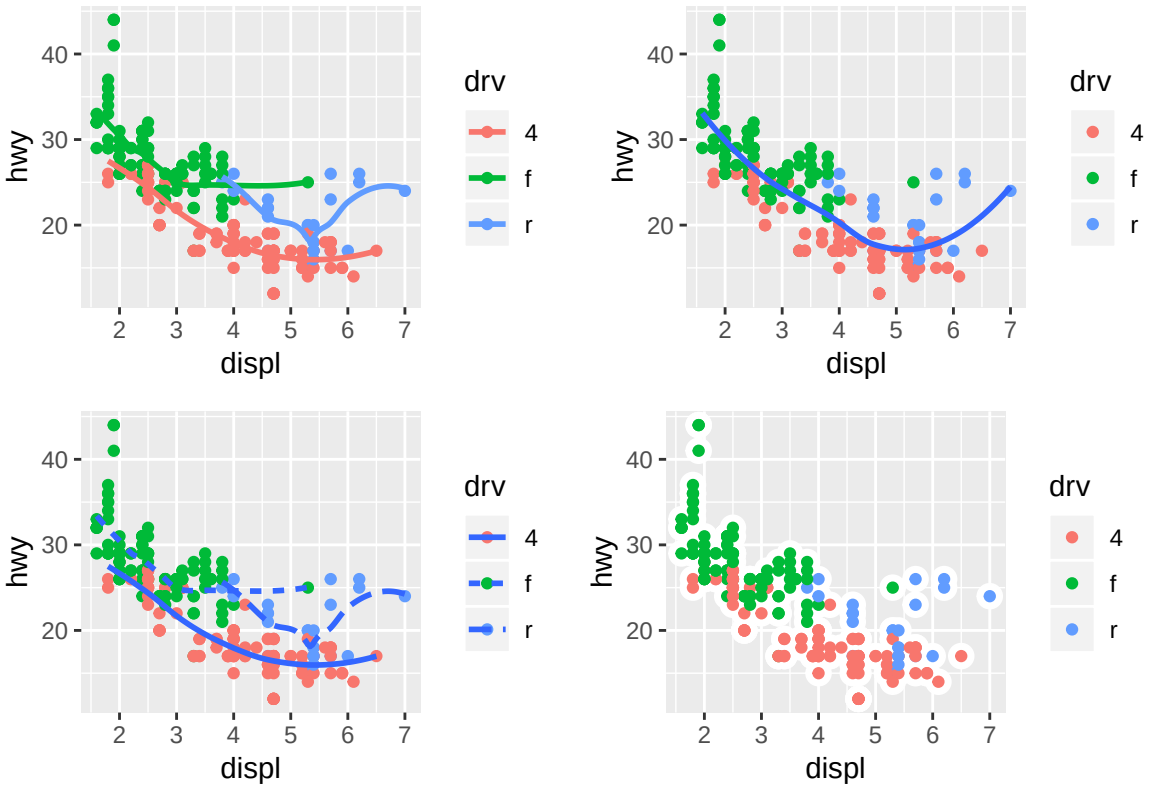
3. What does `show.legend = FALSE` do? What happens if you remove it?
   Why do you think I used it earlier in the chapter?
4. What does the `se` argument to `geom_smooth()` do?
5. Will these two graphs look different? Why/why not?

```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
  geom_point() +
  geom_smooth()

ggplot() +
  geom_point(data = mpg, mapping = aes(x = displ, y = hwy)) +
  geom_smooth(data = mpg, mapping = aes(x = displ, y = hwy))
```

6. Recreate the R code necessary to generate the following graphs.





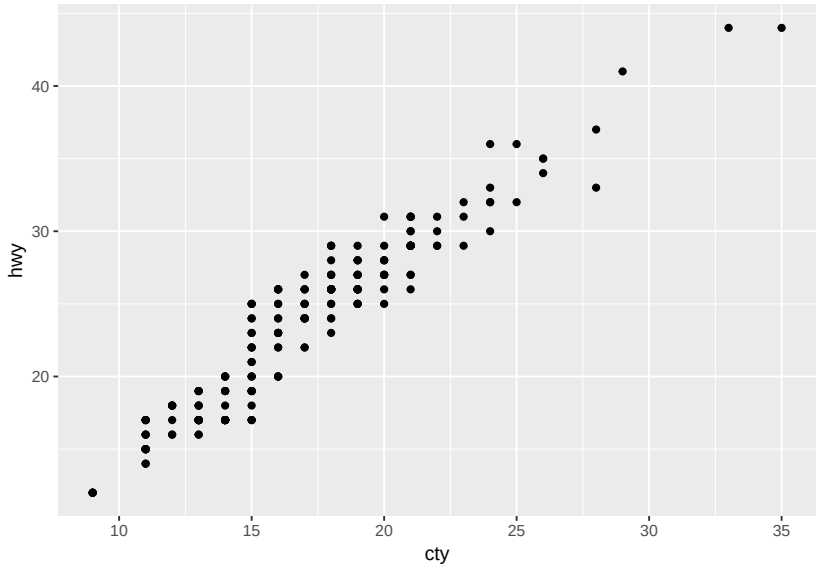
Exercises 5

1. What is the default geom associated with `stat_summary()`? How could you rewrite the previous plot to use that geom function instead of the stat function?
2. What does `geom_col()` do? How is it different to `geom_bar()`?
3. Most geoms and stats come in pairs that are almost always used in concert. Read through the documentation and make a list of all the pairs. What do they have in common?
4. What variables does `stat_smooth()` compute? What parameters control its behaviour?

Exercises 6

1. What is the problem with this plot? How could you improve it?

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



2. What parameters to `geom_jitter()` control the amount of jittering?
3. Compare and contrast `geom_jitter()` with `geom_count()`.
4. What's the default position adjustment for `geom_boxplot()`? Create a visualisation of the `mpg` dataset that demonstrates it.