E392: Problem Set 1

Introduction to R & Data Visualization

Spring 2018

Answer Key

Answers are included below in italic font. I graded Part 5 from Question 1 and Question 2.2. I commented on your project ideas if something came to my mind, but did not formally grade this question.

Please work on the following questions and hand in your solutions in groups of at most 2 students. You are asked to answer all questions, but we will only select 2 (sub)questions randomly to grade.

Part 1: R questions

Question 1: First Steps with R

Finish working through the R Tutorial uploaded on Canvas answering all the questions. *See separate document.*

Question 2: Data Visualization

The following questions use the `mpg` data set than comes with the `tidyverse` library.

Question 2.1: Visualization Basics

1. Run `ggplot(data = mpg)`. What do you see and why? *You will only see an empty plot since this command just indicates which data set to use, but not which plot to create. Add a mapping and a geom too see an actual plot.*
2. What does the `drv` variable describe? Read the help for `?mpg` to find out. *The variable describes whether a car has front-wheel, rear-wheel or all-wheel drive.*
3. What happens if you make a scatterplot of `class` vs `drv`? Why is the plot not useful? *The problem with this plot is that you only see points where combinations of `class` and `drv` are present in the data. However since both variables are categorical, points will perfectly overlap and it is impossible to see how many cars there are for each combination. Therefore, the plot only conveys very limited information.*

Question 2.2: Aesthetic Mappings

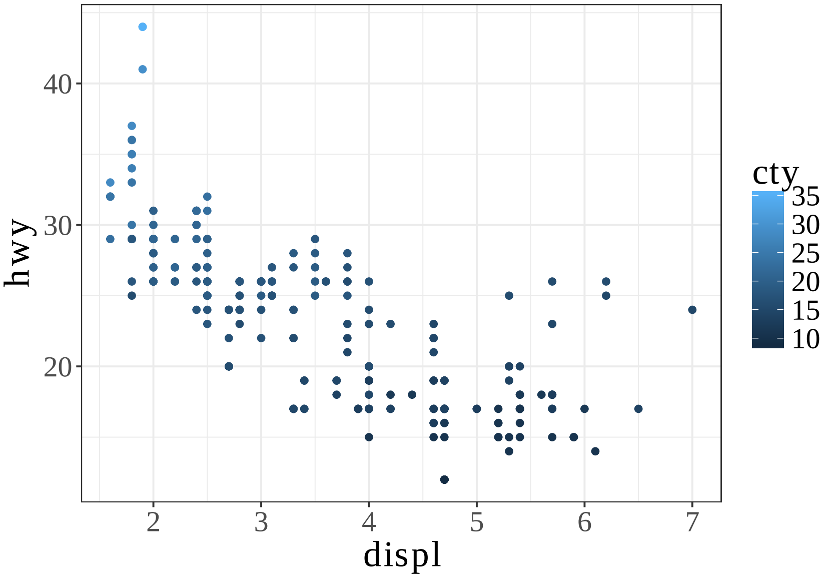
1. What's gone wrong with this code? Why are the points not blue? *Here, the* `color` *attribute is not part of the mapping, but the layout definition of the plot. In order to make the points blue, simply move it out of the mapping function.*

   ```
   ggplot(data = mpg) +
     geom_point(mapping = aes(x = displ, y = hwy, color = "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? *When you print the* `mpg` *data frame, each column displays which variable type it contains. For example,* `doubles` *are very likely to be continuous variables, while* `character` *vectors or factor variables typically indicate categrocial variables.* `displ, cty, hwy` *are clearly continuous variables, the other variables are interpreted as categorical. Sometimes, the distinction is not clear-cut, for example, you might interpret the year-variable also as continuous.*

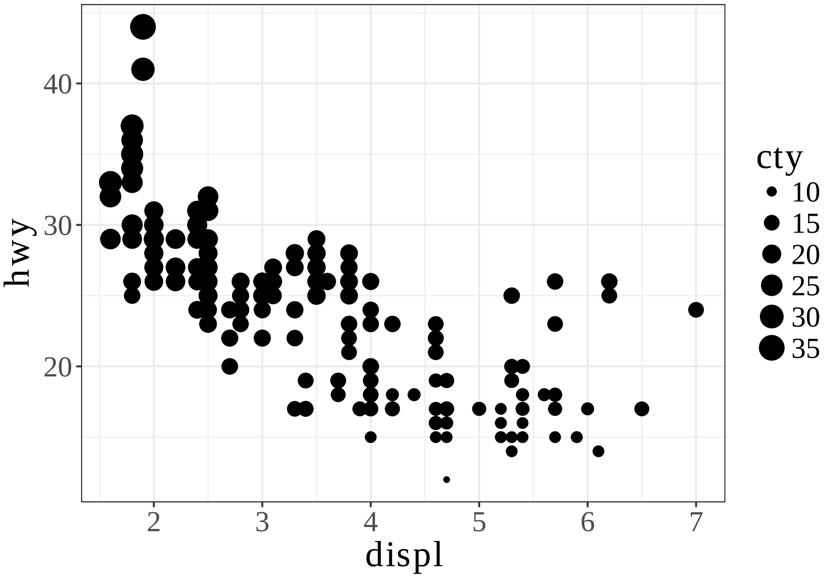
3. Map a continuous variable to `color`, `size`, and `shape`. How do these aesthetics behave differently for categorical vs. continuous variables?

   ```
   # Map cty to color.
   # We get points colored in a color continuum.
   # I think this looks quite good.
   ggplot(data = mpg) +
         geom_point(mapping = aes(x = displ, y = hwy, color=cty))
   ```



   ```
   # Map cty to size.
   # Similarly, we get points with size varying continuously.
   # Here, it does not look very nice and clutters our graph.
   ```

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



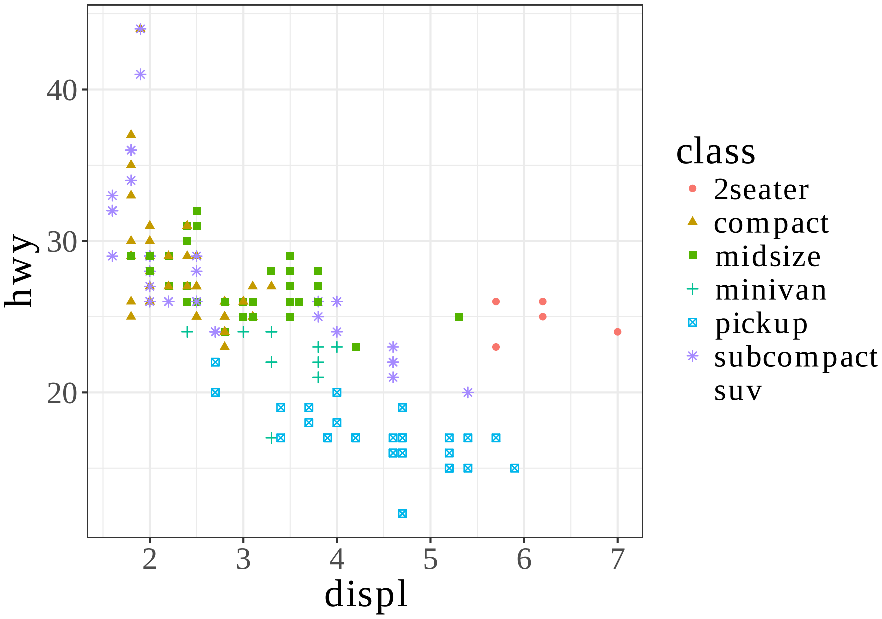
```
# Map cty to shape.
# This will fail, since R only has a finite (actually very limited) number of shap
# ggplot(data = mpg) +
#       geom_point(mapping = aes(x = displ, y = hwy, shape=cty))
```

4. What happens if you map the same variable to multiple aesthetics? *That's not a problem at all, the variable just determines both aesthetics. However, this double-mapping is usually not very informative.*

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

```
## Warning: The shape palette can deal with a maximum of 6 discrete values
## because more than 6 becomes difficult to discriminate; you have 7.
## Consider specifying shapes manually if you must have them.
```

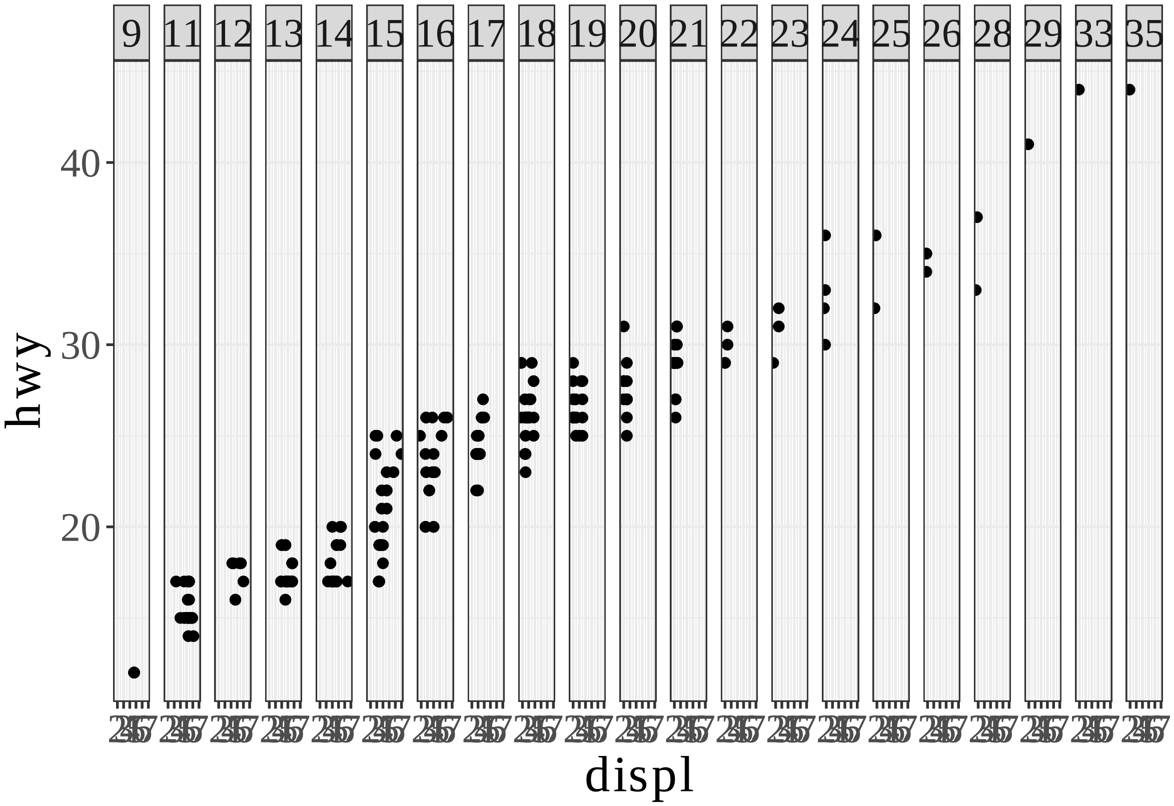
```
## Warning: Removed 62 rows containing missing values (geom_point).
```



Question 2.3: Facets

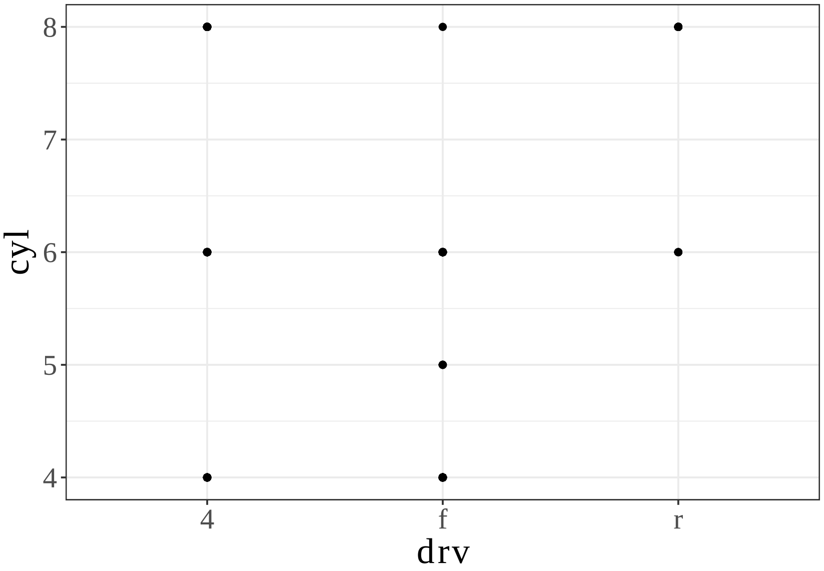
1. What happens if you facet on a continuous variable? *You will get one facet for each realization of the variable that's present in the data set. So this is rarely what you want...*

```
# For example faceting on cty-fuel consumption.
# Don't do this!
ggplot(data = mpg) +
      geom_point(mapping = aes(x = displ, y = hwy)) +
      facet_grid(~cty)
```

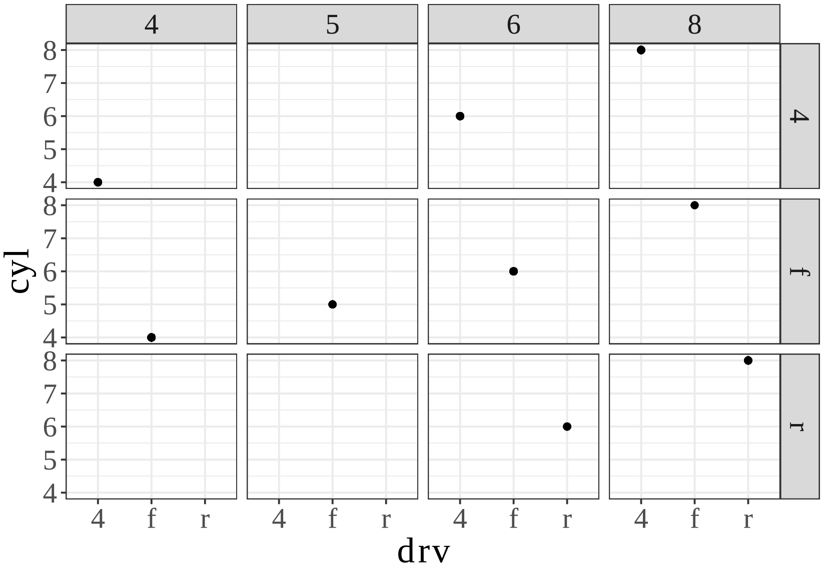


2. What do the empty cells in plot with `facet_grid(drv ~ cyl)` mean? How do they relate to this plot? *They indicate combiantions of `drv`-`cyl` that are not present in the data. In the first plot below, these combinations are displayed by empty grid points. In the second plot, variable values that are not present at all are automatically dropped, for example there are no 7 cylinder cars, so there is now subplot row for 7 cylinders. However, we still get empty subplots when a combination is not present, but at least one of the variable values is, for example 4-cylinder rear-wheel drive cars.*

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

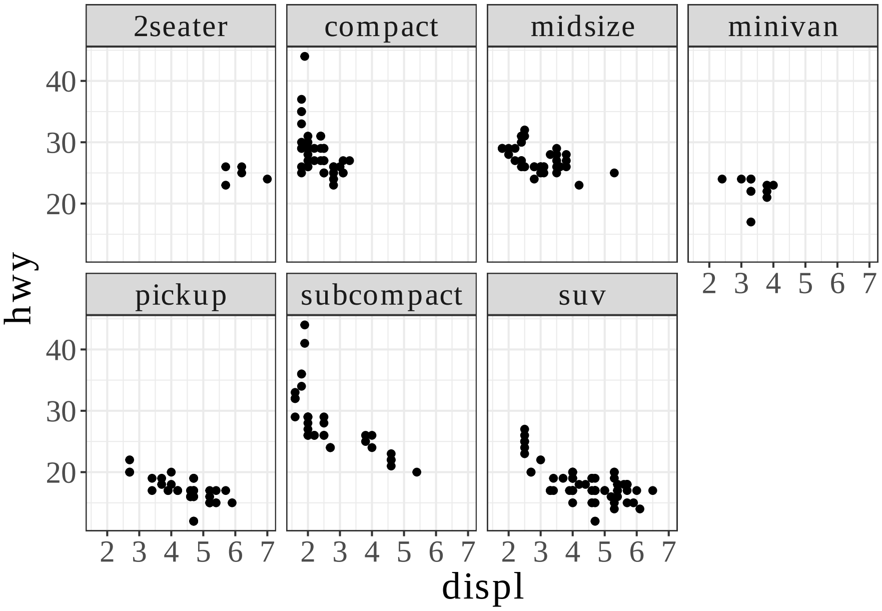


```r
ggplot(data = mpg) +
    geom_point(mapping = aes(x = drv, y = cyl)) +
    facet_grid(drv ~ cyl)
```



3. Take the following faceted plot:

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



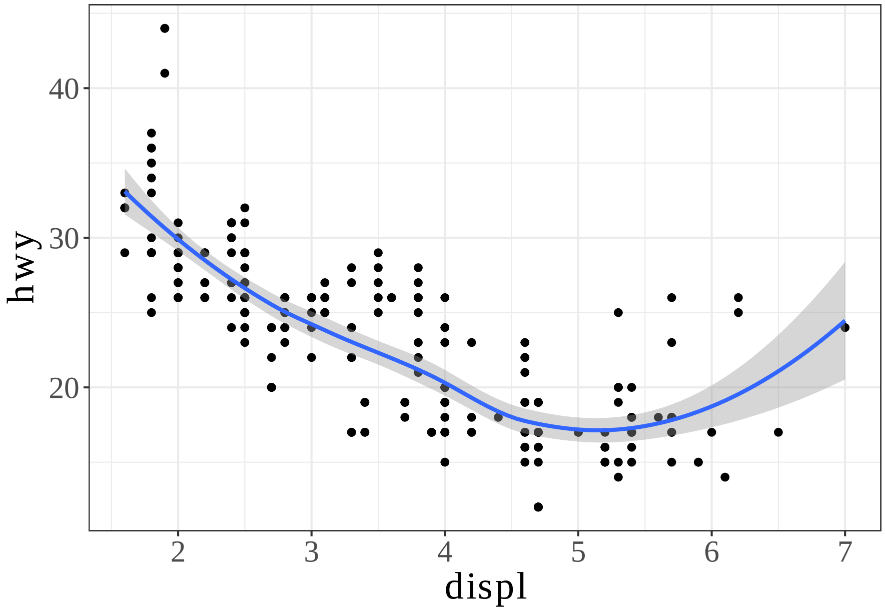
What are the advantages to using faceting instead of the colour aesthetic? What are the disadvantages? How might the balance change if you had a larger dataset? *Generally, the color aesthetic might be more useful if you are mostly interested in the global relationship between two variables since all points are displayed in one graph. If you are more interested in how the relationship looks for specific groups and potentially comparing the within-group relationships across groups, facets might be the better choice. In addition, a single graph can easily be look cluttered when you have a very large data sets, splitting the plot into several subplots might then allow you to see more of the actual information.*

Question 2.4: Geometric Objects

1. What `geom` would you use to draw a line chart? A boxplot? A histogram? An area chart? *Simply check the ggplot-cheatsheet: `geom_boxplot()`, `geom_histogram()`, `geom_area()`.*
2. Will these two graphs look different? Why/why not? *As we can see, there's no difference at all in the output. In the second plot we provide exactly the same mapping to both geoms. In the first one we define a global mapping and leave the local mapping for each graph empty so that each geom-function uses the global mapping. The first version of the code looks cleaner and has less room for typos, so use global mappings wherever you can!*

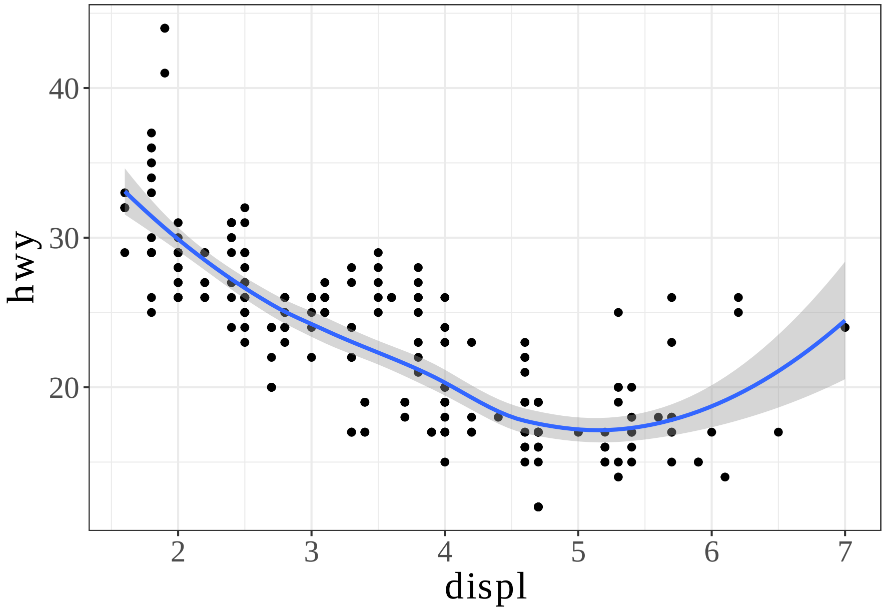
```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
  geom_point() +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'loess'
```



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

```
## `geom_smooth()` using method = 'loess'
```



Part 2: Your project

Question 3

Get started on your project! Think about an empirical question that you find exciting and start looking for available data to answer this question. It is very likely that many questions/project ideas will fail eventually, so it's a good idea to get started by just brainstorming to get several ideas. Then start to invstigate the 2 or 3 ideas that you like most in more detail.

If you cannot come up with a specific question, you can also start the other way round: start browsing the Internet for interesting data sets. Check out what is available, what kind of information the data contains and how these data could be used for an interesting analysis.

Write up to one page about this process and what you think the most promising idea for your project is and why.