HW 3, STAT 450

Due: Friday, September 27

Directions: This assignment should be completed using Quarto and submitted to Canvas as a self-contained HTML or PDF file.

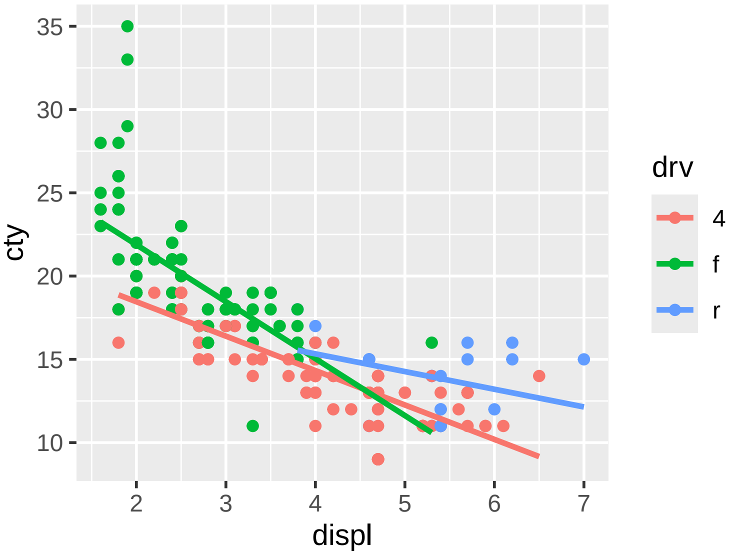
Reading: Chapters 1 and 9 from R for Data Science (2e)

```r
library(tidyverse)
```

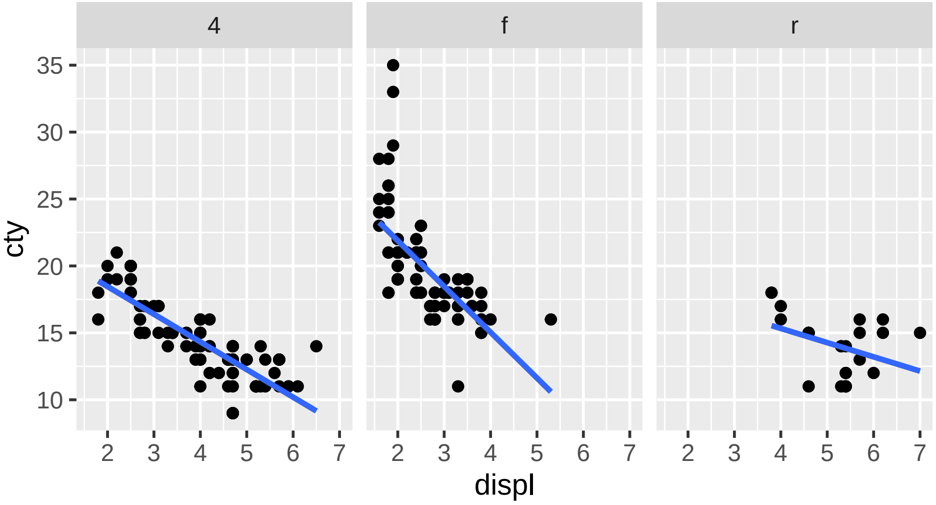
Exercise 1

Using the `mpg` data frame, recreate the R code necessary to make the following plots. In your submission include both the code and the plot.

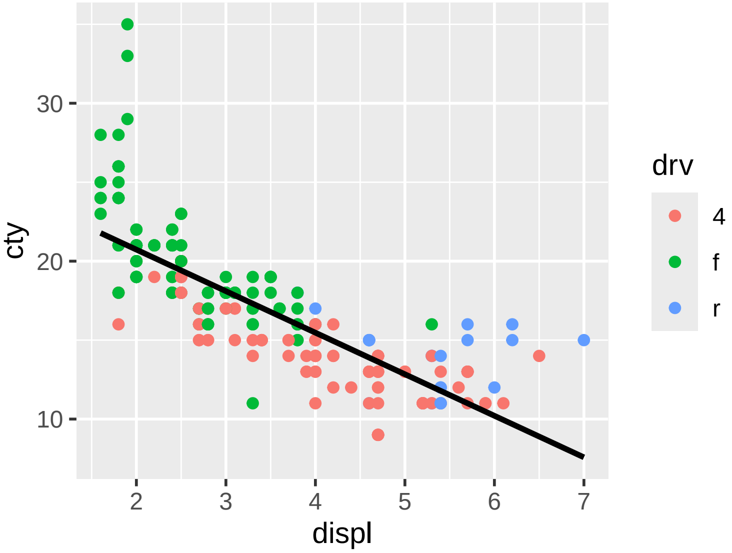
a



b

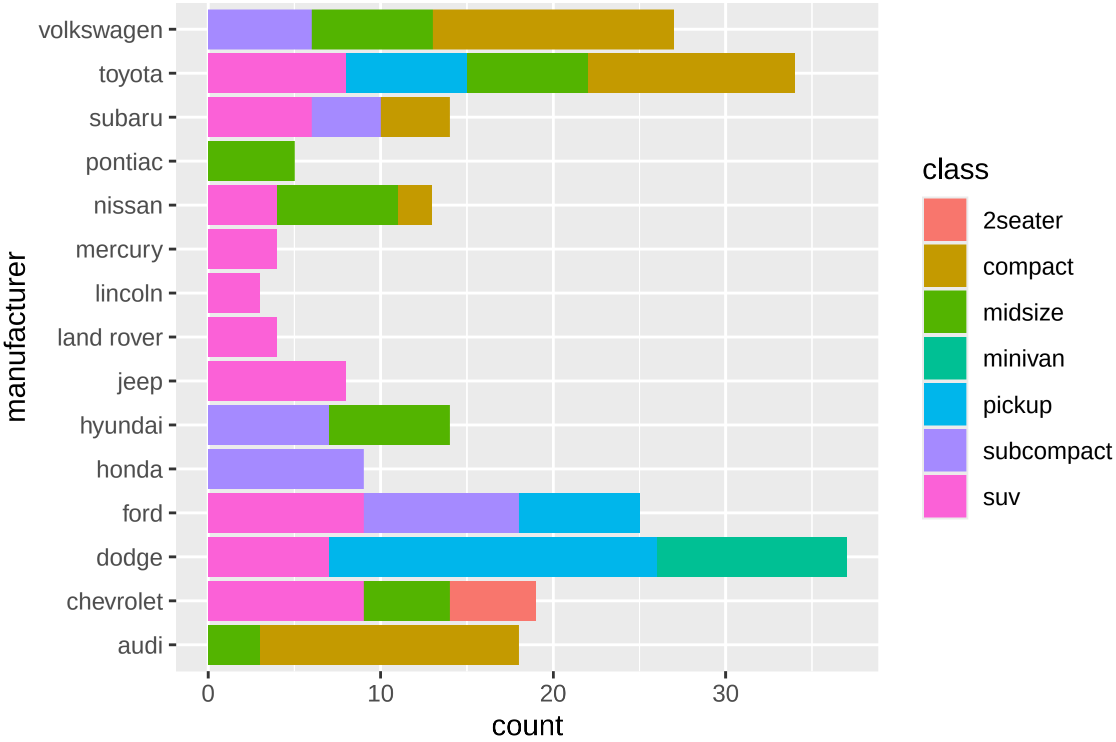


c



cty
30
20
10
2
3
4
5
6
7
displ
drv
4
f
r

d



manufacturer
volkswagen
toyota
subaru
pontiac
nissan
mercury
lincoln
land rover
jeep
hyundai
honda
ford
dodge
chevrolet
audi
0
10
20
30
count
class
2seater
compact
midsize
minivan
pickup
subcompact
suv

Exercise 2

For this exercise use the CPS85 data frame from the mosaicData package. Use ggplot2 to create all graphics.

```r
library(mosaicData)
```

```r
head(CPS85)
```

```
##   wage educ race sex hispanic south married exper union age   sector
## 1  9.0   10    W   M       NH    NS Married    27   Not  43    const
## 2  5.5   12    W   M       NH    NS Married    20   Not  38    sales
## 3  3.8   12    W   F       NH    NS  Single     4   Not  22    sales
## 4 10.5   12    W   F       NH    NS Married    29   Not  47 clerical
## 5 15.0   12    W   M       NH    NS Married    40 Union  58    const
## 6  9.0   16    W   F       NH    NS Married    27   Not  49 clerical
```

```r
dim(CPS85)
```

```
## [1] 534  11
```

A description of this data set is provided in the help menu.

```r
help(CPS85)
```

a

Make a histogram and density plot of `wage`. For the histogram, set the argument `binwidth = 3`. Describe the shape of the distribution.

b

Make side-by-side box plots to look at the distribution of `wage` for each category of `sector`. Which sectors have the highest median wages? Which sector has the greatest variability in wages?

c

Make a bar plot of `sector`. Which sector has the highest number of employees?

d

Make a stacked bar plot that shows the relationship between `sector` and `sex`. Map the `sex` variable to the fill color of the bars.

e

Repeat part **d**, but display proportions instead of counts. Which sectors have roughly the same proportion of male and female employees?

Exercise 3

In this exercise you'll make a map of Alameda County. First, make sure to load the relevant map packages:

```
library(maps)
library(mapproj)
```

a

Run the following code to make a map of California with county boundaries.

```
ca <- map_data("county", "ca")
ggplot(data = ca, aes(x = long, y = lat, group = group)) +
  geom_polygon(fill = "white", color = "black") +
  coord_map()
```

b

The object `ca` is a data frame that contains the coordinates for the polygons of each county in California. Here is a preview of the first several rows:

```
head(ca)
```

```
##        long      lat group order     region subregion
## 1 -121.4785 37.48290     1     1 california   alameda
## 2 -121.5129 37.48290     1     2 california   alameda
## 3 -121.8853 37.48290     1     3 california   alameda
## 4 -121.8968 37.46571     1     4 california   alameda
## 5 -121.9254 37.45998     1     5 california   alameda
## 6 -121.9483 37.47717     1     6 california   alameda
```

Run the following two commands, and explain what you think each command is doing.

```
unique(ca$subregion)
length(unique(ca$subregion))
```

c

Use the `dplyr` function `filter()` to extract the rows of the `ca` data frame that correspond to Alameda County. Store the subset in a new data frame called `alameda_ca`.

d

Use the subsetted data frame from part **c** to make a map of Alameda County with `ggplot2`.

Bonus

Make a map of the nine counties in the Bay Area (Alameda, Contra Costa, Marin, Napa, San Francisco, San Mateo, Santa Clara, Solano, and Sonoma).