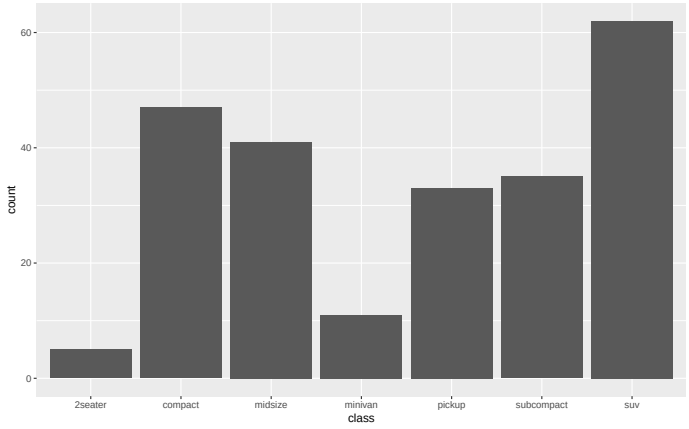
Introduction to Data Science
Homework 2: Due Wednesday September 5 at 2:00pm

Exercises:

1. Read the Structured Data handout and the first few sections through *Visualising distributions* of *R for Data Science*.
2. Find a data set or data sets that exemplify each of the terms: rectangular or tabular data, continuous, discrete, categorical, binary, ordinal.
3. Explain why it is important to have a taxonomy of data types.
4. True or False: The `diamonds` data set in the `ggplot2` package is a data frame. One way to answer this question is by typing the command `str(diamonds)` in R*.
5. True or False: The `diamonds` data set in the `ggplot2` package is "tidy."
6. How many variables are there in the `diamonds` data set?
7. Classify each of the variables in the `diamonds` data set. That is, state if the variable is continuous, discrete, categorical, etc.
8. Describe the difference between a bar plot and a histogram. Under what circumstances would you use each?
9. Explain the result of the command

```
ggplot(data = mpg) + geom_bar(mapping = aes(x=class))
```

   You should get a plot that looks like this:



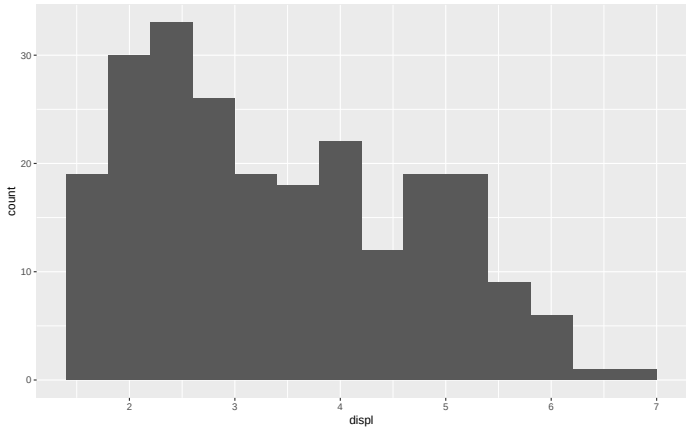
   Make sure to answer the question in the context of the data.

*If you want to look at the `diamonds` data set, make sure you have the `ggplot2` package installed and loaded.

10. Explain the result of the command

```
ggplot(data = mpg) + geom_histogram(mapping = aes(x=displ),binwidth=0.4)
```

You should get a plot that looks like this:



Make sure to answer the question in the context of the data.

11. Explain the meaning of the results obtained after running the R command[†]

```
summary(mpg)
```

What information does this tell you about the `mpg` data set?

[†]Make sure you have the `ggplot2` package installed and loaded.