Introduction to Data Science
Homework 2: Due Wednesday September 18 at 2:00pm

Exercises:

1. Visit `http://data.gov`, and identify five data sets that sound interesting to you. For each write a brief description and propose three interesting things you might do with them.

2. Read the first few sections through *Visualising distributions* of *R for Data Science* `https://r4ds.had.co.nz/`.

3. 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*.

4. How many variables are there in the `diamonds` data set?

5. Classify each of the variables in the `diamonds` data set. That is, state if the variable is continuous, discrete, categorical, etc.

6. Describe the difference between a bar plot and a histogram. Under what circumstances would you use each?

7. Explain the result of the command

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

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

8. Explain the result of the command

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

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

9. 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?

10. Describe what each of the listed functions do, be detailed

    - `list.files`

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

- `dir.create`
- `file.create`
- `file.info`
- `file.rename`
- `file.copy`

11. Read through section 4 of the article *Excuse me, do you have a moment to talk about version control?*, then answer the following questions.

    (a) What is Git?

    (b) What is GitHub?

    (c) Why should someone in data science care about these?