# Homework 2: Stats with R

*Mike Lopez*

*September 2017*

***General instructions for homeworks***:

- Make a new R Markdown file (.Rmd) referring to the assignment on the course Github page
- Change the heading to include your author name
- Save the R Markdown file (named as: [MikeID]-[Homework01].Rmd – e.g. "mlopez-Lab01.Rmd") to somewhere where you'll be able to access it later (zip drive, My Documents, Dropbox, etc)
- Your file should contain the code/commands to answer each question in its own code block, which will also produce plots that will be automatically embedded in the output file
- **Each answer must be supported by written statements (unless otherwise specified) as well as any code used**: In other words, if the answer is `24`, you should write "The answer is 24" (as opposed to just showing the code and output).
- Include the names of anyone you collaborated with at the top of the assignment
- I recommend copying the raw .Rmd code from the Github page as a start
- Homeworks are due at the start of class – please print the HTML and hand in.

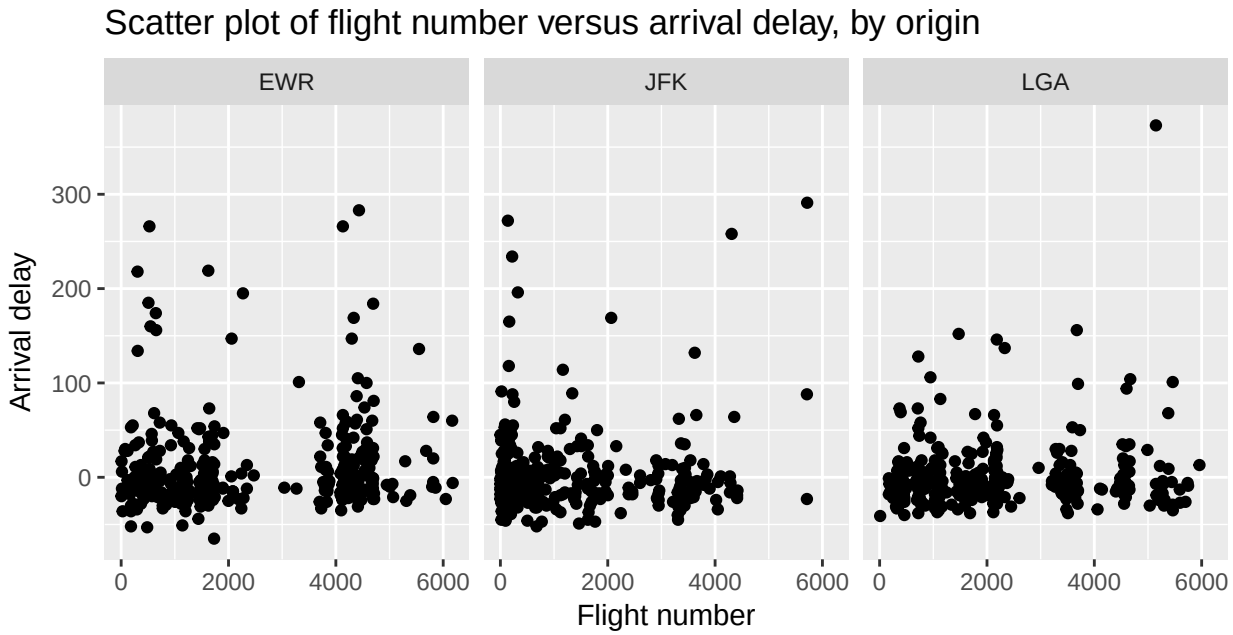
The two data sets that we'll be using for this lab are (i) `mpg` and (ii) a random sample of `flights`, the latter of which can be found in the `nycflights13` package. See class notes for exact details on the flights data set, or enter `?flights` to use the Help tab.

Start be executing the following code

```
library(tidyverse)
library(nycflights13)
set.seed(0)
flights <- flights %>% sample_n(1000)
```

1. How many rows and columns are in `flights`?
2. What does the `flight` variable describe in flights?
3. Make a scatter plot of `flight` versus arrival delay, and summarize the association. Does it surprise you that there (is/is not) an association between flight and arrival delay?
4. There's something peculiar about the distribution of `flight`. What do you think it is?
5. What happens if you make a scatter plot of `flight` versus `tailnum`. Why is this plot not useful?
6. Make the following plot:

```
## Warning: Removed 28 rows containing missing values (geom_point).
```




7. Here's two sets of code. Describe the difference between mapping color to your graph when using continuous versus categorical variables.

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

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

8. Here's two sets of code. Describe the difference between mapping shape to your graph when using continuous versus categorical variables (note: one set of code may not work – explain why it doesn't!)

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

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

9. Explain what's wrong with the following code:

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

10. One option is to color each car type by class. What are the advantages to using the faceting below? What are the disadvantages? How might the balance change if you had a larger dataset?

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

11. Run this code in your head and predict what the output will look like. Check your predictions. What is the new code that is used here?

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