# 30535 Applied Problem Set 1

*Peter Ganong*

*03/23/2020*

Due Friday April 24, 5:00PM Central.

Submit by pushing your code to your repo on Github Classroom.

This submission is our work alone and complies with the 30535 integrity policy.

Add your initials to indicate your agreement: **___**

Add names of anyone you discussed this problem set with: **___**

Late coins used this pset: 0. Late coins left after submission: 4.

Submit by pushing your code to your repo on Github Classroom: https://classroom.github.com/g/fnSknfLP.

## 1 Git merge conflicts (10 pts)

You and your partner will share a github remote repository and make changes and update the same files in your local repositories. This can be messy, but git has handy tools to deal with potential conflicts between your versions.

We are suggesting the following workflow to minimize conflicts. Divide the work into manageable chunks we refer to as "issues".

i. **Branch**: For each chunk, you will work on a branch. Pushing your changes to github often. We repeat. Add and commit code to your branch early and often!
ii. **Pull Request**: When you have a stable section of code, make a pull request.
iii. **Review**: Ask your partner to review your code (this should be a high priority step!)
iv. **Merge**: Merge your code into master, dealing with any merge conflicts immediately.

Repeat with the next issue of your assignment.

Now, to practice, you will create your first merge conflict. We created videos to support your through these steps and also will do a live demo in lab. Find the videos on canvas: https://canvas.uchicago.edu/courses/28237/files/folder/Videos/git%20primers/collaboration

*Prelude*

i. Play paper, scissors, rock to determine who goes first. Call that person *Partner 1*.
ii. Both partners, clone your groups `applied_ps_1`. On the master branch, *Partner 1* rename the file to `applied_ps_1.Rmd`, delete the html file, commit and push to github. *Partner 2* pull the changes. Now you both have the properly named file on your computers.

*Begin merge conflict practice*

i. 
    a. *Partner 1*, Start a branch called `merge_conflict_practice_1`. In `applied_ps_1.Rmd` replace "Yuxi Wu" with your name. Push your branch to github. (They call this "Publish" in github desktop).
    b. *Partner 2*, Start a branch called `merge_conflict_practice_2`. In `applied_ps_1.Rmd` replace "Yuxi Wu" with your name.

ii. *Partner 1* screen share and make a pull request.

iii. *Partner 2* screen share on github review the pull request. Accept your partners changes and merge the branch into `master`. Hooray! This is your first successful pull request!

iv. *Partner 2* make a pull request.

v. *Partner 1* screen share. On github review the pull request. There should be a merge conflict because you both changed the same line of the file. Adjust the file and then merge.

Switch roles and force a merge conflict with a different section of `applied_ps_1.Rmd` (e.g. the date or titles).

1. Succinctly explain, why did you have a merge conflict?

# 2 Flight Data: Part I (30 pts)

*An international trade organization is hosting a two-day convention in Chicago in 2020. The mayor's tourism office has asked for some planning help based on historical data from 2016.*

## 2.1 Download BTS data

- Download files for the 2016 calendar year data here: https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236
  - Warning: The whole file is huge, so this task will be faster and more manageable if you download just the columns and rows that you need. Limit the sample to Illinois and download only the columns that are relevant to the problem set.
  - Warning: The BTS site is representative of government data websites in that it is quite finicky. It seems to work best if you wait for one month of data to download completely before starting the next month.
- Read it into R using `il_flights <- read_csv("data.csv")` and use the `bind_rows` command to stitch together the monthly files into a yearlong file.
  - Note: Rmd files will look for data in the same folder as the file. You likely downloaded data into your Downloads directory. You can point R to look in downloads e.g. (for Mac and Linux) `il_flights <- read_csv("~/Downloads/data.csv")` or move the files to your current working directory.
  - Warning: `setwd()` does not work inside of Rmd chunks. See **this post** for an alternative solution.

## 2.2 Data Description (10 pts)

1. What is the unique identifier for each flight in the dataset?
2. R has six description methods: `print, head, str, glimpse, View, summary`. Apply them to `il_flights`
   1. Are any of the methods redundant, in the sense that you don't learn anything about the data from these commands that you didn't already know from the prior methods? Make a list of the non-redundant methods (giving preference to the command with prettier output).
   2. Of the non-redundant methods, write a note (max 2 lines per command) that will quickly help someone (perhaps future you!) recall how each command is useful.

## 2.3 Data Validation (20 pts)

1. You should have 675822 rows when you downloaded data for Illinois. Load the package `testthat` and then test that you have this many rows using the command `test_that("we have the right number of rows",expect_equal(nrow(data),675822))`

2. Because of the conditions you put into the webform, all flights should be to or from Illinois airports. Let's check this.
3. Drop flights to and from Midway and O'Hare. How many flights are left?
4. Among flights whose origin or destination is not Midway or O'Hare, what are the five most common origins? What are the five most common destinations? Where are these cities? Are these origins and destinations inside or outside Illinois? Can you explain why these are the most common origins and destinations?
5. Next, limit the sample to flights to or from Midway and O'Hare.
    1. How many rows do you think the dataset *should* have, approximately? Find at least two websites that estimate the number of flights into and out of each airport. Do these estimates agree with what is in the BTS dataset? Do these estimates agree with each other? If they disagree, why do you think they disagree?
6. Google to figure out the three highest-volume airlines, defined as number of flights leaving or arriving at an airport, at O'Hare and at Midway. Does this agree with what you find in the BTS data? If they disagree, why do you think they disagree?

# 3 Flight Data: Part II: When should they Mayor's tourism office host their convention? (60 points)

*Use the same data which you analyzed above. Limit the sample to flights to Midway and O'Hare.*

*For each question, please follow the four-part approach laid out in lecture. I have given you the question (step 1). You should write out your query (step 2), show the plot from this query (step 3), and write out the answer to the question in a sentence (step 4).*

## 3.1 Choose a month

1. When are average arrival delays into Chicago (measured using the arrival delay variable) the lowest? When are at least 80% of flights on-time? Make a single plot that answers both questions and write a sentence (or two) that answers these questions.

2. When are flights to Chicago most common? Make a plot to support your answer and write a sentence to answer the question.

3. What month do you recommend they have the convention and why? Consider both the number of flights to Chicago and that the tourism board would like attendees to arrive in Chicago on-time. Write a few sentences.

    a. In lecture, we covered the idea of "basic" plots and "sophisticated" plots. Make a "basic" plot which provides the minimal amount of information needed to support your written recommendation.
    b. Make a "sophisticated" plot that contains more information about flight delays. What are the sub-messages in the "sophisticated" plots that are not in the "basic" plot? If you could submit only one of the two plots to the mayor's office, which would you submit and why?
    c. You have (hopefully) reached the frontier of what you can do to answer this question with the data that you have. If you wanted to push the frontier further of figuring out when the convention should be, what are two other **public** datasets that would be useful in making a decision? Include links to the datasets and the names of the variables you would analyze. We do not expect you to actually analyze these other datasets.

## 3.2 Pick an airline

1. Now that you've decided on a date, recommend an airline to take in order to arrive on time. The attendees are not price-sensitive, so you don't need to worry about cost. Make a "basic" plot and a "sophisiticated" plot to support your recommendation. Which plot do you prefer and why?

## 3.3 Reconsider the date?

1. The trade organization sends an update. Some of its most important members are in Greer, South Carolina. Does that change your recommendation of when to host the convention? Make a plot that supports your new recommendation and shows why it is superior to your old recommendation.