

Homework 1

Due: Wednesday January 30 at 11:59pm

Use RStudio for all exercises, and use either SQL commands or operations on `big.matrix` objects. Efficiency is important. Use efficient programming techniques discussed in class, and use the data objects we have already created (creating a new database or `big.matrix` backing file when the data is already available in another format is inefficient because it costs both time and memory).

You should provide one script (a `.R` or `.rmd` file) that contains all the code and includes code comments noting which code is for which exercises. You will also need to show and comment on the results, so place the results in a Word (or Open Office or HTML or PDF) document and write sentences to answer the questions, or use `knitr` to programmatically create your document. **Script files must be the actual script files**, not unevaluated code pasted into some other document.

Include your name in the name for each file submitted ('<Your-First-Name> <Your-Last-Name> HW#.R', e.g. 'JaneDoeHW1.R').

Any code based on code from elsewhere (e.g. code provided with the text) **must reference in code comments** the source of the original code.

Exercises for All Students

- 1) This exercise is for aggregate departure delay information for flights from 1987 to 1989 in the data.
 - a) Using SQL, obtain the total number of flights in the data in the 1980s.
 - b) Using SQL, obtain the number of flights with departure delayed by more than 15 minutes in the 1980s in the data.
 - c) Comment on the percentage of flights with departure delayed by more than 15 minutes during that time period.
- 2) Now we look at the similar delay information by month during that period. (Note: This is just by month, not by month and year. For instance, flights for January 1987, January 1988, and January 1989 will be aggregated together.)
 - a) Obtain a table for the total number of flights in our data by month in the 1980s from the data.
 - b) In a separate table, obtain the number of flights by month with departure delayed by more than 15 minutes in the 1980s in the data.
 - c) From the results in parts **a** and **b**, programmatically calculate the percentage of flights delayed by more than 15 minutes by month of year during that time period, and comment on how the monthly rates compare to the overall rate found in exercise 1.

- 3) Now we look at aggregate flight data for 2007 and 2008.
 - a) Obtain the total number of flights in 2007 and 2008, the number of flights delayed by more than 15 minutes during that time period, and the percentage of flights delayed by more than 15 minutes during that time period.
 - b) Comment on how this delay rate compares with the rate found for the 1987-1989 flights.
- 4) Now we look at the delay rate per year for 2007 and 2008.
 - a) For each year from 2007 to 2008, calculate the number of flights and the number of flights delayed by more than 15 minutes. (You should have counts for 2007 and counts for 2008.) Be sure to use efficient programming techniques.
 - b) Compute the percentage of flights with departure delayed by more than 15 in each of those two years and compare the annual rates with the aggregate rate found in exercise 3.

Additional Exercises for Graduate Students

- 5) This exercise is to compare delay rates by day of week from 1987 to 1989 with delay rates by day of week from 2007 to 2008 within the data provided.
 - a) Calculate the percentage of flights delayed by more than 15 minutes for each day of the week for the period from 1987 to 1989 in the data provided.
 - b) Repeat part **a** for 2007 and 2008 data.
 - c) Comment on similarities and differences in the delay rate on particular days of week between the two time periods.