Homework 2

Due: Wednesday February 6 at 11:59pm via compass2g

Use RStudio for all exercises. Efficiency is important. Use efficient programming techniques discussed in class, and use the data objects we have already created when possible.

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. 'JaneDoeHW2.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. All exercises are based on the 2007-2008 airline data we have used in class.

Exercises for All Students

1) A traveler is planning a trip for July 2009 and wonders about the amount of departure delay they might encounter. They have the data from 2007 and 2008 and want to look at delays that are at least of median length. Obtain the 50th through 99th percentiles for July data in those years and interpret what the results tell us about magnitudes and frequency of delayed departures in July during those two years.
2) The traveler is also curious about differences in departure delay percentiles for July during those two years. Compute and compare the 50th through 99th percentiles for July 2007 and July 2008. Provide an informative visualization along with interpretation of similarities and differences in the delay quantiles.
3) Consider month and day of week as continuous linear predictors for departure delay. Obtain a linear regression model for departure delay as a function of month and day of week using the 2007-2008 data. Interpret what the model suggests about the relationship between delay time and the day of week and month. Comment on the usefulness of the model and any issues with using this model.

Additional Exercises for Graduate Students

4) Rather than a straight linear trend, it is suggested that delays might be much worse in winter and not as bad in summer. Likewise, it is suggested that delays might get increasingly worse as the week goes on.

   To model this suggested behavior, consider a linear model where departure delay is equal to $\beta_0 + \beta_1(Month - 6)^2 + \beta_2 DayOfWeek^2$ (Note: See the `I` function in R to

include the squared terms in the model). Again use the 2007-2008 data, interpret the model, comment on how useful it is, and compare it to the model from exercise 3 which is linear in `Month` and `DayOfWeek`.