STAT 360
Dr. Straight
Quiz 1, Due Monday, February 13

Note: *You are on your honor to do your own work and not help others!*
For each question, submit a PDF file of your R Notebook.

General Instructions: Use the `flights` data frame from the `nycflights13` package. Create a data frame `lag` of flights from LaGuardia Airport (LGA), and a data frame `lag_to_atlanta` of flights from LaGuardia Airport to Atlanta (ATL). Use `select/filter` to produce data frames that contain only the columns/rows of interest. Also, be sure to filter out any rows with NAs in any of the numeric columns.

1. For this question, use `lag_to_atlanta` and a random sample of 29 values of the `arr_delay` variable. Seed the random generator using 21323.
(5 points)
   (a) Use the naïve method to compute the 70th percentile.
   (b) Use the `quantile()` function to compute the same thing.
   (c) Use the naïve method to compute the median and the first and third quartiles.
   (d) Use `favstats()` to compute the same thing.
   (e) Produce a box plot. Are there any outliers?

2. For this question, add a column `status` to the `lag_to_atlanta` data frame. It has three possible values: `early`, `on_time`, and `late`. Define a flight to be "on time" if it is not more than 15 minutes late. Use the `tally()` command to produce a contingency table with a row for each airline and a column for each value of `status`. Be sure to show the row and column totals.
(5 points)

3. Continuing with `lag_to_atlanta`:
(5 points)
   (a) Produce a histogram for `air_time`.
   (b) Reseed the random number as above. Pick two of the three airlines DL, FL, and MQ; denote them by A and B. Take a random sample of 100 flights for A, and let $\hat{p}_A$ be the proportion of late flights (in the sample). Compute $\hat{p}_B$ in a similar way. Label the airlines so that $\hat{p}_A > \hat{p}_B$. Produce a confidence interval for $p_A - p_B$.

4. Use `lag`.
(5 points)
Include those airlines with more than 1000 departures. Using a chi-square test, test the hypothesis that whether a flight is late or on time is independent of the airline. Hint: Set up a contingency table where the rows are "late" and "on time" (not late) and the columns are the airlines. Then the expected value for the cell in row i and column j is the product of the total for row i and the total for column j, divided by the overall total. The number of degrees of freedom

for the chi-square statistic is $(r - 1)(c - 1)$, where r is the number of rows and c is the number of columns.