STAT 360
Dr. Straight
Midterm Exam, Due Tuesday, April 4

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

1. Produce a "thematic" map of the Middle East. Include important bodies of water. Each country should be labelled with its name, and these names should be well-placed. Color the countries according to some attribute (e.g., per capita GDP, major religion(s), type of government). For full credit, your use of color most approach "Madigan quality."

2. For this question, use the `Batting` data frame from the `Layman` package. We are interested in classifying batters as "low RPPPA" or "high RPPPA", where RPPPA is runs produced per plate appearance. INCUDE LOTS OF COMMENTS! Explain what you are doing and what the results mean.
First, "clean up" the data frame. We are interested in the years from 2010 to 2019. We're going to look only at National League batters, and only those with a minimum of 100 plate appearances. (We'll treat each "stint" by a batter in a season as a separate instance). Select the columns of interest and fix their types, if needed. You'll need to add the following three columns (and perhaps others): (1) PA (number of plate appearances), defined as AB + BB + HBP + SF; (2) RP (runs produced), defined as R + RBI – HR; (3) RPPPA (runs produced per plate appearance), defines as RP/PA. (Note: A home run counts as both a run scored and a run batted in, which is why we're subtracting HR from the sum of R and RBI.)

(a) Calculate the median RPPPA, and use it to classify each batter as "low RPPPA" or "high RPPPA". Add a column to your data frame with this classification.

(b) Form the training and testing data frames. Use an 80% - 20% split.

(c) Determine the proportion of batters in each class and determine the accuracy of the null model. (Do not actually construct the null model – simply use common sense!)

(d) Perhaps batting average (H/AB) is a good predictor. Build a classifier with this variable as the only predictor; call it `model1`. How does it do (on the training set)? Compute its accuracy and its confusion matrix.

(e) Determine another variable that might be a good predictor. Using `tally()` and/or appropriate graphs, show that this variable may be a good predictor. Add this predictor to `model1` to obtain `model2`. How does it do? In addition to computing its accuracy and confusion matrix, produce a summary using `broom::tidy(model2)`. Interpret it.

(f) Apply `model2` to the testing set. How does it do?

3. Use the flights data frame from the nycflights13 package. Create a data frame ny_to_dfw of flights from LaGuardia Airport (LGA) or Newark Airport (EWR) to Dallas/Fort Worth (DFW). We are interested in flights that depart before noon. Use select/filter to produce a data frame that contains only the columns of interest and rows for the three major airlines that fly this route. Also, be sure to filter out any rows with NAs in any of the numeric columns.

(a) Produce a contingency table with five columns, one for each weekday (Monday – Friday), and three rows, one for each value of the variable status (early, on-time, late). Use the tally() command, and be sure to show the row and column totals.

(b) Is there a difference in the distribution of status values from one day to another? For example, are there more late flights on Monday than on Friday? Perform an appropriate chi-square test of independence.

(c) Produce a confidence interval for the mean air_time. Produce a confidence interval for the difference in mean air_time for two of the days.

(d) You consult for companies that use the two airports EWR and LGA whose representatives frequently travel to Dallas/Fort Worth on business. Typically, they need to be at the designated meeting location by 2:00 pm. (Note: DFW is in the Central time zone.) Looking at the confidence interval for air_time found in part (c), and allowing 1 hour to get from the DFW airport to the meeting location, come up with a recommendation/rule for the latest possible scheduled departure time. Then use sampling from the appropriate data frame to test your rule. That is, using your rule, how frequently will representatives be late for their meetings, and by how much? Note: Your "rule" should not be overly restrictive. That is, it's OK for some folks to be late to their meetings, but the percentage should be reasonable.

4. Do something interesting with supervised learning. Include at least two of the following model types: decision tree, nearest neighbor, neural network.

Here is an idea. Use the Pitchers data frame. We want to classify a pitcher as "low ERA," "medium ERA," or "high ERA," depending on whether the pitcher's ERA is less than *x*, between *x* and *y*, or greater than *y*. (You should choose *x* and *y* so that each category contains roughly the same number of pitchers.) As predictors, use variables such as those in Quiz 3. Suggestion: Consider the 2010 – 2019 seasons, and group by playerID and yearID. Optional: Create a new variable, pitcherID, by combining playerID and yearID. Include only those pitchers with a relatively high number of total appearances (that is, total value of G.)