Final Project Instructions
BIO5312, Fall 2017

For the final project, you will ask **four scientific questions** about a dataset. Each question should be addressed using a statistical approach learned in class, and each analysis should contain a corresponding figure and brief figure interpretation. **An example project has been prepared for you. I strongly recommend you go over this example project to see what is expected. Failure to review this document may result in deductions.**

You must submit the assignment via Canvas, as an RMarkdown document *with your data uploaded as well*. If you are unable to upload your data to Canvas, please contact me by email.

You have a choice of two deadlines for this project:

1. **Early:** Tuesday December 12th by 11:59 pm. You will receive a **10 point bonus** (5% points) if you submit by this deadline.
2. **Regular:** Tuesday December 19th by 7:45 pm. This corresponds to the end time of the official Final Exam slot for this class. *Projects not submitted by this deadline will not be accepted and will earn a grade of 0. There will be **no exceptions**.*

The final project is worth **200 points** and will be graded according to the following guidelines:

1. **Introduction (30 points)**
    a. This section should contain 1-2 paragraph *written in your own words* describing the dataset with any citations (if there are citations, please include a properly-formatted “References” section at the end of the RMarkdown document). This text can be the same as that used for the project proposal. Note that if references are used but not included in a references section, points will be deducted.
    b. You must include (similar to what you have seen in previous homeworks) a *bullet point list* list of all variables in your dataset.
    c. Indicate the alpha value you will use to assess significance throughout the project.

2. **Data Preparation/Wrangling (10 points)**
    a. This section should be used to show all code used to prepare data for use, for example reading from a database, tidying data, or merging different datasets with join() or similar.
    b. If no preparation in R was necessary, simply use this section to read in your dataset. *You should not read in the dataset in any other section of the assignment besides this one*.
    c. Note that this section *should not* be used to wrangle or manipulate data as it pertains to individual questions. Use this section only to tidy data for subsequent use.

3. **Questions (30 points each = 120 points)**
    a. Each question will have these main components (see the template and the example project!).
        i. The header (four hashtags) should state the question
        ii. Next, there should be a single bold sentence indicating the method used
        iii. A code chunk showing statistical analysis
        iv. A paragraph interpreting your analysis and *answering the question*
        v. A code chunk showing your data visualization, *preceded* by a brief interpretation of the figure. Figures should have meaningful and clear axis names and use aesthetics correctly.
    b. Any reported information without corresponding code will be regarded as incorrect; All code must be shown for any reported results. In addition, if your answers are determined by printed R output, this R output must actually be printed from the R code itself (although you must avoid printing large dataframes, see section 4 below).
    c. Note that methods using random sampling (bootstrap, k-fold cross validation, permutation tests, etc.) will give a slightly different results each time the code is run. This means your “previewed” R chunks may not match the final output. Points will **not** be deducted for minor inconsistencies between reported values and those in the R code, provided the differences are due to random sampling.

4. **Organization and Presentation (40 points)**
    a. Your document must be *proofread for spelling and grammatical errors.* Typos and egregiously incorrect grammar (specifically, non-sentences like sentence fragments) will result in point deductions. Note that RStudio has

a (limited) spell-check feature: The “ABC” checkmark button next to the “Knit” button. *This is not a substitute for proofreading but may be useful!*

b. Write your project in active voice, not passive voice. In other words, use phrasing like “I ran a t-test,” **not** “A t-test was run.” Points will be deducted if passive voice overly-predominates the document. *Writing in first-person is therefore required*. Please do not hesitate to contact me with any questions on this point.

c. Your document must adhere to the provided project structure. Points will be deducted for deviating from this template or imposing radically different formatting.

d. All libraries must be loaded in the “global_options” R chunk at the top of the document. The template loads a minimal set of libraries, so you may need to add more.

e. Points will be deducted for extraneous code, code that prints full data frames or large amounts of output (always print dataframes with head()), and for printing of excessive warnings (>1 line) in the Rmarkdown output. Please consult the R markdown documentation (or the example project!) to see how to turn off warning output if you run into it.

f. Points will be deducted for documents that fail to knit.

g. *To reiterate, It is essential that you knit and carefully proofread examine your final document in HTML format to ensure that you do not lose points here.*

5. **Finally,** be aware that the following issues will result in a final project grade of 0:
    1. Failure to include data, or failure for your submitted data to fully reproduce your analysis. If you are unsure about what data to submit (i.e. in cases where data was obtained from a database or cases where substantial data wrangling is needed) please contact me ahead of submission to clarify.
    2. ***Any** use of language that is not your own*. If you are unsure if certain language will be acceptable, please contact me ahead of submission to clarify.