BIOL 419/519

Homework 5, Winter 2016

Due on Saturday, Feb 13 by midnight

- **About this homework:** This homework is a departure from the previous homework assignments: you'll practice some research and hacking skills doing a data pull from a spreadsheet that's a little hairy. You'll also have to work with mixed data types (numbers and strings of characters) from that spreadsheet. This homework will help you practice for doing realistic data pulls from datasets like you may encounter for your course project—the skill of figuring how to do what you need by doing research on coding is very valuable!
- You are encouraged to inspect the spreadsheet of raw data and get to know its format. However, you are not allowed to modify the data spreadsheet in any way by hand.
- **Label your plots for full credit.**

Instructions: Please submit a write-up of your homework by uploading on Canvas. Your solutions must include the code your wrote to solve the problem as well as the output/answer. You should generate a write-up within Matlab using the "Publish" option to make a PDF file. Please also upload all .m files. **Failure to turn in all .m files will result in half credit for the entire assignment.**

In all exercises that ask you to plot or visualize something, label your plots for full credit.

Please seek help if you need it! You may ask questions at Friday's lab, come to office hours, and get together with your classmates to troubleshoot together.

Collaboration: As noted in the Syllabus, what you turn in should reflect your own understanding of the material. Collaboration with your classmates is encouraged, and I ask you to clearly indicate these collaborations in your homework.

1. (2 pts) **Pull in the data**

   Download the data spreadsheet from Canvas named `Neonatal_Mortality.xlsx`. This file contains the neonatal mortality data from all (or most) of the world's countries between the years of 1990 and 2015, as well as a bunch of other stuff about each country. You are encouraged to open the file in Excel or another spreadsheet software and look over the different columns of data.

   Use the `readtable` function to read in this spreadsheet. How many countries are in this spreadsheet? What are the units of the neonatal mortality data reported here?

   Hint: Read the documentation for the `table` variable types to learn how to manipulate tables and how to turn parts of tables into cell arrays or matrices of numbers. You may also find useful the `fieldnames` function to find the column labels of your table.

2. (3 pts) **Visualizing some distributions**

   Plot the histograms of the neonatal mortality rates of all countries in 2015 and in 1990. You should make one plot with the two histograms; each histogram should be a different color. The two histograms should have the same bin sizes.

   Hint: Read the documentation for the `histogram` function.

3. (2 pts) **Trends in time**

   Plot the neonatal mortality for Malawi for all years between 1990 and 2015.

   Hint: Some functions you may find useful to compare strings of characters (words) to other strings are `strcmp`, `strcmpi`, `strncmp` and `strncmpi`.

4. (3 pts) **Aggregate visualizations**

   The spreadsheet also contains information on whether each country is Low, Mid or High income in one of its columns. Actually, it specifies whether a country is "Upper Middle" or "Lower Middle" income, but we're going to combine these categories and consider all of these countries as "Mid" income.

   Plot the mean neonatal mortality for all Low, Mid, and High income countries for all years between 1990 and 2015. In other words, find all the Low (and Mid and High) income countries, compute their mean neonatal mortality for each year, and plot these as three different lines on one figure.

   Your plot should have "Mean neonatal mortality" on the vertical axis and "Year" on the horizontal axis.

   Hint: Some functions you may find useful to compare strings of characters (words) to other strings are `strcmp`, `strcmpi`, `strncmp` and `strncmpi`.

5. (Extra Credit) **Build a classifier**

   Build a classifier on all the data from 2000 to categorize a country as Low, Mid, or High income based on their neonatal mortality rates, using a Linear Discriminant Analysis as implemented by the `classify` function. Report the cross-validated accuracy.

6. (Informational) How many hours did you spend on this homework?