FUNDAMENTALS OF DATA SCIENCE – DSC 441

# HOMEWORK 1

*Turn your assignment in as a PDF file to the D2L submission folder by the due date associated with that folder (unless otherwise specified explicitly by the Instructor). Late assignments will incur a penalty. Please ask if you need an extension and come to office hours if you need assistance.*

*Academic integrity is key to maintaining the impact of your degree. You may discuss the problems and methods with others, but everything you submit must be your own work. For violating, you may incur significant sanctions including failing the course or penalties from the University. If in doubt, ask first.*

*To get credit, you must clearly label your responses with the questions they answer. Include results and visualizations that you generate, and clearly written answers to any questions included in the problem. Your code file must be submitted alongside your document in D2L.*

This homework assignment covers material from Modules 1 and 2. You will get practice manipulating and exploring data, including pivoting and visualization.

## Problem 1 (15 points):

For this question, we will use the US census dataset from 1994, which is in *adult.csv*.

a. First, we look at the summary statistics for all the variables. Based on those metrics, including the quartiles, compare two variables. What can you tell about their shape from these summaries?
b. Use a visualization to get a fine-grain comparison (you don't have to use QQ plots, though) of the *distributions* of those two variables. Why did you choose the type of visualization that you chose? How do your part (a) assumptions compare to what you can see visually?
c. Now create a scatterplot matrix of the numerical variables. What does this view show you that would be difficult to see looking at distributions?
d. These data are a selection of US adults. It might not be a very balanced sample, though. Take a look at some categorical variables and see if any have a lot more of one category than others. There are many ways to do this, including histograms and following tidyererse *group_by* with *count*. I recommend you try a few for practice.

e. Now we'll consider a relationship between two categorical variables. Create a cross tabulation and then a corresponding visualization and explain a relationship between some of the values of the categoricals.

## Problem 2 (15 points)

In this question, you will integrate data on different years into one table and use some reshaping to get a visualization. There are two data files: *population_even.csv* and *population_odd.csv*. These are population data for even and odd years respectively.

a. Join the two tables together so that you have one table with each state's population for years 2010-2019. If you are unsure about what variable to use as the key for the join, consider what variable the two original tables have in common. (Show a *head* of the resulting table.)
b. Clean this data up a bit (show a *head* of the data after):
    a. Remove the duplicate state ID column if your process created one.
    b. Rename columns to be just the year number.
    c. Reorder the columns to be in year order.
c. Deal with missing values in the data by replacing them with the average of the surrounding years. For example, if you had a missing value for Georgia in 2016, you would replace it with the average of Georgia's 2015 and 2017 numbers. This may require some manual effort.
d. We can use some tidyverse aggregation to learn about the population.
    a. Get the maximum population for a single year for each state. Note that because you are using an aggregation function (*max*) across a row, you will need the *rowwise()* command in your tidyverse pipe. If you do not, the max value will not be individual to the row. Of course there are alternative ways.
    b. Now get the total population across all years for each state. This should be possible with a very minor change to the code from (d). Why is that?
e. Finally, get the total US population for one single year. Keep in mind that this can be done with a single line of code even without the tidyverse, so keep it simple.

## Problem 3 (15 points)

Continuing with the data from Problem 2, let's create a graph of population over time for a few states (choose at least three yourself). This will require another data transformation, a reshaping. In order to create a line graph, we will need a variable that represents the year, so that it can be mapped to the x axis. Use a transformation to turn all those year columns into one column that holds the year, reducing the 10 year columns down to 2 columns (year and population). Once the data are in the right shape, it will be no harder than any line graph: put the population on the y axis and color by the state.

One important point: make sure you have named the columns to have only the year number (i.e., without popestimate). That can be done manually or by reading up on string (text) parsing (see the *stringr* library for a super useful tool). Even after doing that, you have a string version of the year. R is seeing the 'word' spelled

two-zero-one-five instead of the number two thousand fifteen. It needs to be a number to work on a time axis. There are many ways to fix this. You can look into *type_convert* or do more string parsing (e.g., *stringr)*. The simplest way is to apply the transformation right as you do the graphing. You can replace the *year* variable in the ggplot command with `as.integer(year)`.

## Problem 4 (15 points)

This problem is short answer questions only. No code is needed.

a. Describe two ways in which data can be dirty, and for each one, provide a potential solution.
b. Explain which data mining functionality you would use to help with each of these data questions.
    a. Suppose we have data where each row is a customer and we have columns that describe their purchases. What are five groups of customers who buy similar things?
    b. For the same data: can I predict if a customer will buy milk based on what else they bought?
    c. Suppose we have data listing items in individual purchases. What are different sets of products that are often purchased together?
c. Explain if each of the following is a data mining task
    a. Organizing the customers of a company according to education level.
    b. Computing the total sales of a company.
    c. Sorting a student database according to identification numbers.
    d. Predicting the outcomes of tossing a (fair) pair of dice.
    e. Predicting the future stock price of a company using historical records.