**Lab 5**

**Preparing Data**

In this lab, you will learn how to:

\*Obtain demographic information from the US Census

\* Use Microsoft Excel to clean and format data for use with Tableau

\*Create visualizations in Tableau using information from different data sources

**Data**

When creating visualizations, half the work is usually the process of obtaining the data to input to Tableau or a similar program. We will work through the process of cleaning a data file in order to use it to create a variety of plots.

***USstates\_Income.xlsx*** – The finished data file we will be creating for practice. The tab ‘Income’ in the Excel file has cleaned data from the US Census showing median household income by county.

**Videos**

<http://www.tableausoftware.com/learn/tutorials/on-demand/data-blending-0>

<http://www.tableausoftware.com/learn/tutorials/on-demand/join-types>

<http://www.tableausoftware.com/learn/tutorials/on-demand/data-blending-0?reg-delay=true>

**Practice**

**Cleaning Data**

The Census Bureau website provides a wide range of county level data.

<http://www.census.gov/support/USACdataDownloads.html>

Navigate to this site and look at the different types of data available. You will have to choose one of the available datasets for the homework.

For now, download these files:

1. Mastdata.xls (Reference Information Files)
2. INC01.xls (Income, Money (U.S. Census Bureau).
3. USstates.xlsx (from Canvas)

The file Mastdata.xls is a list of all the available variables from the Census Bureau. In addition, it is recommended to read the help page to understand the content of the Excel files that are available (<http://www.census.gov/support/USACdata.html>).

We now go through the process of cleaning the Excel file INC01.xls.

* First, open both INC01.xls and Mastdata.xls.
* We can see that INC01.xls has a large number of columns, most of which are superfluous for our purposes. Delete the column labeled ‘STCOU’ as well as the columns which are full of zeros. You should be left with a column called ‘Area\_name’ and about 10 columns of variables.
* We don’t need every column, so just keep the first three data columns (‘INC110179D’, ‘INC1101189D’, and ‘INC110199D’). Note that right now we don’t really know which variables we have, as they have unintelligible names.
* Take the name of each column, and search for it in the Mastdata.xls file. (Hint: use Ctrl + F (Cmd + F on a mac) to search within Mastdata.xls).
* Name each data column in INC01.xls based on what the variable represents. (For example, you might re-label ‘INC110179D’ to ‘Median Income 1979’)
* Make a new column called ‘Abv’. We can see that the last two letters of the first column correspond to the state abbreviation. In the second cell of the column you named ‘State’, type ‘=Right(A2,2)’. This copies the state abbreviation to this column.
* Fill this formula all the way down to the bottom of the Excel sheet by dragging the black box at the bottom right side of the cell.
* Go through and delete all the rows that do NOT correspond to counties. (UNITED STATES, ALABAMA, ALASKA, …). Be patient, this can take a little time.
* Tableau likes all the data to be in the same file. Make a new tab in USstates.xlsx called ‘Income’. Now copy the data from INC01.xls to this tab, and save the resulting file as USstates\_Income.xlsx.

**Tableau**

* Open Tableau and import USstates\_Income.xslx
* Join the two tables using inner joins:
  + Join State Data and Income with an inner join. This associates each state with its abbreviation. This yields additional data about each individual state, with income data in addition to the demographic data which we have used previously.
* Create a barplot of median income in 1999 by state. Sort in decreasing order.
* Create a scatterplot of median income in 1999 vs. human development index.

**Assignment**

**Due Date: Feb 16, 2018, 6:00 pm. Value: 10 points**

1. Download a data file of interest to you from the Census Bureau from <http://www.census.gov/support/USACdataDownloads.html> along with the file Mastdata.xls.
2. Prepare the data for use in Tableau as described in the ‘Cleaning Data’ section. Add the data to USstates.xlsx and save the resulting file as ‘USstates\_blabla.xlsx’, where the file name depends on your chosen topic from the Census data. **You will submit this file on eCommons** in addition to the .pdf of the Tableau dashboard. (5 pts.)
3. Load the data into Tableau, and join the data as described above. Generate a scatterplot using data from each table. Place this scatterplot in a new dashboard names ‘Exercise 1’. What is the relationship between the two variables you chose based on your scatterplot? Does this relationship make sense? Discuss the meaning of your plot by making comments on the dashboard. (5 pts.)

**Remember to upload both the .pdf of the ‘Exercise 1’ dashboard and the Excel file to Canvas!**