# Capstone Project 1 Data Wrangling

Much of the data wrangling for this project took place before any files were read into Python. The goal of the project is to conduct a pilot study for a network of alcoholism treatment centers to see if the severity of winter weather, as measured by the Accumulated Winter Season Severity Index (AWSSI), could be used to project patient demand. The study is based on the hypothesis that severe winter weather may exacerbate symptoms of alcoholism, leading more people to seek treatment, and thereby increase demand at treatment centers.

The AWSSI, from the National Oceanic and Atmospheric Administration and the National Weather Service, is calculated for 110 U.S. weather reporting stations where snow and temperature data are collected. Alcoholism treatment data were from the federal government’s “Treatment Episode Data Set: Admissions” (TEDS-A), which “provides annual data on the number and characteristics of persons admitted to public and private substance abuse treatment programs that receive public funding.”

In reviewing the two data sets to identify the data for this pilot study, there were several factors to consider to ensure the findings were reliable and generalizable. First, only locations that experienced winter in a meaningful way were considered. Second, locations that were too small to produce results upon which to draw conclusions were eliminated.

The TEDS-A is reported by several geographic designations: census division, census region, Primary Metropolitan Statistical Area, Core Based Statistical Area, metropolitan statistical area code, and census state codes. The amount of missing data for each of these is included in the documentation, and for some, much of the data are missing. Core Based Statistical Area, a collective term for both metro and micro areas, had 29% missing data for some years. Therefore, another consideration for choosing the data set was which data sets were most complete.

After reviewing the map of AWSSI reporting stations and the TEDS-A documentation, I found a weather reporting station and geographic area that most met the criteria outlined above. Massachusetts is a compact but populous state affected by winter. There is a single AWSSI weather reporting station in Massachusetts (in Blue Hill), and the state accounts for more than 4% of the overall data. There are no missing data in the alcohol admissions tallies by state.

The TEDS-A data consists of large files with a row for each patient admission for substance abuse treatment in a given year. Some files are comma-separated, others are tab-separated. I used pandas to read in the file for each of the 20 years included in the study. Using the codes listed in the TEDS-A documentation designating the substance abuse being treated and the state, I created a dataframe for each year listing all cases of alcohol treatment admissions in Massachusetts. For each case, I included demographic variables that might produce different findings across subgroups of patients. These were gender, race, education, marital status, employment status, and veterans.

Using pandas, I took the 20 dataframes listing all alcoholism admissions in Massachusetts from 1995 to 2014 and created dataframes that counted the number of cases overall, and the number of cases of each of the categories of the demographic variables, i.e., number of admissions for men, women, whites, blacks, married people, divorced people, etc. I combined these 20 dataframes into a single dataframe.

The AWSSI data for Blue Hill was in a single csv file that I read in using pandas. I created a dataframe of the Accumulated Winter Season Severity Index values for 1995-2014, component indexes (snow score and temperature score), the length of each winter, and the number of missing readings for those years. I combined the alcoholism admissions dataframe and the winter severity index dataframe.

I next counted the number of cells for which there was missing data and listed the variables that had missing data. Of 35 instances of missing data, 20 were in columns included in the TEDS-A files that noted missing data; i.e., the indication that data was missing was missing. The remaining missing data indicated that for six years, the number of veterans admitted for treatment of alcoholism was not noted, and that for nine years, there were no admissions reported in an education subgroup, people who attended high school but did not graduate. The missing data on admissions for veterans will affect the interpretability of the findings relating to that group. However, I believe the missing data for the high school attendees who did not graduate is a finer designation that likely was included in other categories, and that it would likely be combined into another category in analysis I conduct, i.e., included with admissions of people without a high school degree, which would also include those who did not attend high school. All NaN cells were recoded as zero in the data frame.

I made boxplots to visualize the range of all feature and target variables. The winter severity index ranged from 260 to 1196 and was skewed toward the higher values. The temperature score component of the winter severity index had the smallest range of the weather indexes, from 176 to 482, with a seemingly normal distribution. Like the overall Winter Severity index, the component snow score was skewed toward the higher values; it ranged from 84 to 714. The length of winter in days ranged from 108 to 172. None of the weather index feature variables had outliers.

There were outliers for four of the target variables, the demographic breakdowns of patients admitted for alcoholism treatment. There were outliers for whites, full-time workers, civilians and veterans. Considering that these are the target variables, and that for each there are only 20 data points, no action was taken regarding these outliers.

Subsequent analysis revealed that some of the target variables had counts that were possibly too small to be the basis of reliable inferences. For example, these subgroups with very low counts had correlations to the feature variables that were the opposite of the other target variables or close to zero. Based on means, the categories veterans, part-time workers, and people separated from their spouses accounted for 1.5% to 7% of the total.

I decided to delete the veteran and civilian variables and combine some of the other subgroups into larger groups with similar characteristics. For education, five subgroups were reduced to three: those without high school degrees, those with high school degrees, and those with any amount of college education. For employment status, four groups were combined into two: those working (part-time or full-time), and those not working (unemployed or not in the workforce). Four marital status groups were combined into to: those never married, and those who were currently or formerly married.

After combining the subgroups, new boxplots were created. The first set of boxplots, comprised of the smaller subgroups, had four outliers for full-time workers, three that were less than the lower fence, and one greater than the upper fence. This pattern was replicated when full-time workers were combined with part-time workers into a single group of working people. Removing outliers may be appropriate for subsequent analyses.