Ross Brown

Capstone 1 Milestone Report

**Substance abuse treatment company struggles to manage staffing**

Twin Towns Treatment Centers is a growing network of centers offering outpatient treatment for drug and alcohol dependency. The work is labor intensive; staff salaries constitute the primary expense for the company. The company is also charting a growth plan, and seeks information on where it can successfully expand its business.

For these reasons, predicting patient demand is a critical consideration for Twin Towns. On the one hand, having qualified staff on hand for client intake and facilitating group therapy is a must, ensuring patients can get treatment, and staff don’t burn out from excessive overtime work. Conversely, overstaffing costs Twin Towns money and profits, potentially jeopardizing the company’s ongoing success and viability. For its growth planning, Twin Towns seeks to identify markets where the is current or existing demand, so it can fulfill its dual charter of serving the needs of patients and providing revenue and profit. For all these reasons, company executives seek a way to predict staffing needs in the present and near future.

As care providers familiar with the research literature in their field, Twin Towns’ executives know that studies have demonstrated a relationship between substance abuse rates and various external factors, such as unemployment and marital status. Similarly, an extensive body of research ties alcoholism to seasonal affective disorder (SAD). Sad is a mood disorder with characteristics of depression that is associated with features of seasonal winter weather, such as shorter days, cloudy skies, and weather conditions that may lead to homebound isolation, i.e., cold temperatures, and rain, sleet and snow.

Company executives hypothesize that the external factors exacerbating SAD may lead patients managing their alcoholism to experience an increase in symptoms that cause them to seek treatment during severe winter seasons. They also hypothesize that more severe winters may cause people who drink alcohol in a healthy way to develop alcoholism for the first time, a process that could manifest in the months following a severe winter. Either scenario would potentially drive more clients to Twin Towns during the periods noted, increasing staffing needs. A pilot study examining winter severity and alcoholism treatment admissions will demonstrate whether Twin Towns can make staffing decisions based on winter severity.

**Data: Winter Severity**

The [Accumulated Winter Season Severity Index] (<http://mrcc.isws.illinois.edu/research/awssi/indexAwssi.jsp#info>) quantifies winter severity. This index, from the National Oceanic and Atmospheric Administration and the National Weather Service, uses "daily meteorological parameters to quantify the severity of a winter season (<https://journals.ametsoc.org/doi/full/10.1175/JAMC-D-14-0217.1)>". The AWSSI is calculated for locations where weather reporting stations can provide daily temperature and snow measurements, which are used to calculate component temperature and snow indexes. In addition to these three indexes, the dates of winter start and end for each location are provided. These dates are based on observations of specific meteorlogical conditions, rather than calendar dates, so they are the two additional winter severity data points.

**Data: Alcoholism Admissions**

The Treatment Episode Data Set -- Admissions (TEDS-A) is a national census data system of annual admissions to substance abuse treatment facilities (https://catalog.data.gov/dataset/treatment-episode-data-set-admissions-teds-a-2013). Every patient admitted for substance abuse treatment by a program that receives public funding is recorded in this dataset, making it comprehensive, and perhaps the best proxy for new or relapse cases of alcoholism in the United States. Furthermore, alcoholism is the most common substance abuse disorder.

For each person admitted for substance abuse treatment, demographic characteristics listed, including gender, race, education, marital status, veteran status, and employment status. This target data allowed me to investigate whether winter severity had differential effects on alcoholism.

**Data Limitations**

*Winter severity*

1. The AWSSI data is limited to areas corresponding to weather station locations. The degree to which this is a problematic limitation will depend on the locations reported in other data sources for this study.
2. Although data starting in the 1950s is available, some daily measurements are missing. However, the number of missing snow and temperature readings is included with the index data, so the most complete data can be chosen.
3. Another potential limitation is the conceptual parameters it is based on. The index is a cumulative tally of each winter day's snow and temperature point scores. Point thresholds for snow and temperature "give greater weight to extreme or rare occurrences, which would have a higher impact, although the thresholds are admittedly somewhat arbitrary." Whether weighting from these thresholds creates an index that accurately represents winter severity is an unknown.

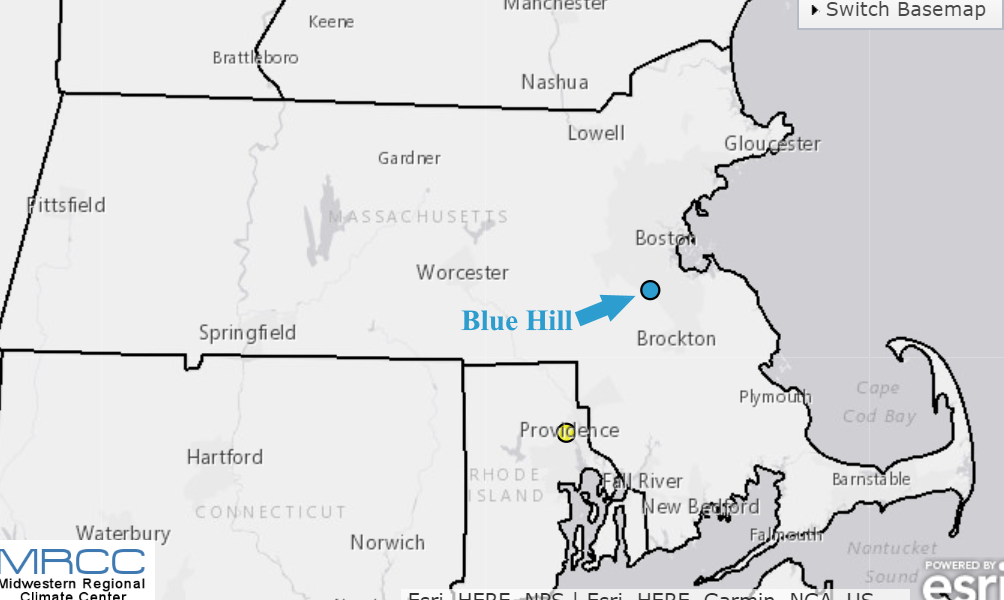
*Alcoholism Admissions*

1. The first of the annual datasets is for 1992; the latest is for 2014.
2. The TEDS-A case admissions are reported by several geographic designations, such as census division, census region, Primary Metropolitan Statistical Area, Core Based Statistical Area, metropolitan statistical area code, and census state codes. However, for some of these location categories, too much data is missing, up to 29%.

Limitations 1, 2, 4, and 5 taken together most likely preclude a lengthy longitudinal study on winter severity and alcoholism that encompasses all areas of the United States affected by winter. While it may be possible to match the geographies of the two datasets to include most of the United States, the timeframe for the admissions data is currently limited to 23 data points, i.e., target data for 23 years. It would be feasible to increase the area of the U.S. included in an analysis, thereby observing the 23 target data points in multiple locations to better discern patterns. However, due to the potential location mismatch for complete feature and target data, a study of multiple areas may still have limited generalizability.

Study Sample: Complete Data for a Large Sample from a Distinct Geography

After reviewing the map of AWSSI reporting stations and the TEDS-A documentation, I identified a weather reporting station and geographic area for a distinct location with a complete data set representing a large enough sample, making for an ideal pilot study dataset. 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 generally accounts for more than 4% of the overall admissions data. There are minimal missing data in the alcohol admissions tallies by state.



Graphic 1: Blue Hill weather reporting station (blue dot).

**Data Wrangling**

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 23 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 23 dataframes listing all alcoholism admissions in Massachusetts from 1992 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 1992-2014; component indexes (snow score and temperature score); the weather-based length, start date, and end date of each winter; and the number of missing readings for those years.

*Missing data*

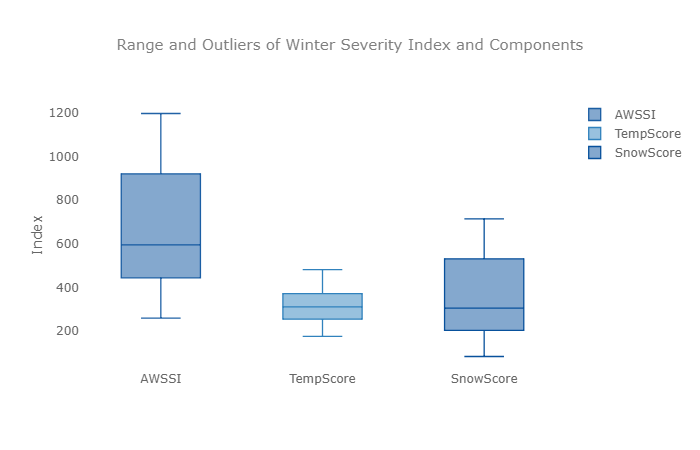
I next counted the number of cells for which there was missing data and listed the variables that had missing data. Of 38 instances of missing data, 32 were in columns included in the TEDS-A files that noted missing data; i.e., the indication that data was missing was missing. The remaining missing data 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.

*Feature engineering with datetime and timedelta*

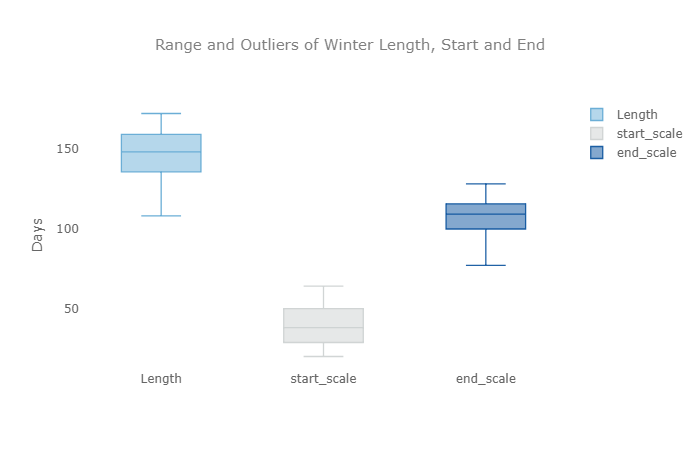
I used the end and start dates of winter to engineer two new features: Measurements in days of how early and how late the winter began and ended each year. Using datetime and timedelta, I first converted the string variables representing these dates into datetime objects. I created another datetime object representing the date of the winter solstice (December 21 most years). I subtracted the winter solstice datetime object from the winter start and end dates. I used the absolute value of the end of winter variable, so that for both scales, greater positive values indicated more: either a winter that started earlier or one that ended later. The two new features were on the same scale as the feature for the length of winter. Finally, I combined the alcoholism admissions dataframe and the winter severity index dataframe.

*Outliers*

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; the start of winter ranged from 20 to 64 days before the solstice, and the end of winter ranged 77 to 128 days after the solstice. None of the weather index feature variables had outliers.

\*

Graphic 2: Weather index box plots



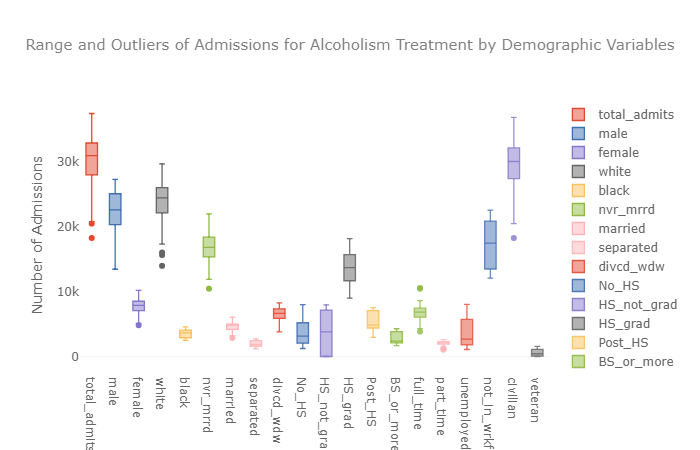
Graphic 3: Duration metrics box plots

There were outliers for seven of the target variables, the demographic breakdowns of patients admitted for alcoholism treatment. There were outliers in the count of all admissions, for females, whites, people who had never married, full-time workers, part-time workers, and civilians. Considering that these are the target variables, and that for each there are only 23 data points, no action was taken regarding these outliers.

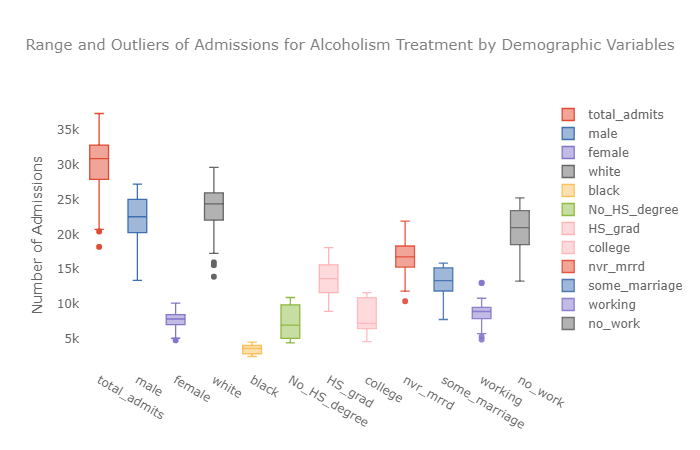
Subsequent analysis revealed that some of the target variables had counts that were possibly too small to be the basis of reliable inferences. For example, these subgroups with very low counts had correlations to the feature variables that were the opposite of the other target variables or close to zero. 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 (most cases were civilians, and some data was missing) 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 set of boxplots for these larger, combined demographic groups, showed outliers for the count of all admissions, for females, whites, people who had never married, and people who were working.



Graphic 4 a: initial target box plots



Graphic 4 b: Final target box plots

**Data Exploration**

*Final Features and Targets*

After engineering new features and consolidate some target subgroups, to review, below are the variable for this project.

Features:

The AWSSI data -- the feature in this analysis -- was measured at the Blue Hill, Massachusetts, weather reporting station from 1992 to 2014. The AWSSI data included the following elements:

\* AWSSI, the overall measure of winter severity

\* Temperature score, based on frequency and degree of cold temperatures

\* Snow score, based on frequency and amount of snowfall

\* Length of winter in days

\* Start scale, an engineered feature that indicates how early, in days, each winter started, relative to the winter solstice.

\* End scale, an engineered feature that indicates how late, in days, each winter ended, relative to the winter solstice.

Targets:

Massachusetts alcoholism admissions case data from 1992 to 2014 were collected from the federal government’s “Treatment Episode Data Set: Admissions” (TEDS-A). Case data included demographic characteristics for each admission. Demographic groups were included in the analysis. Some were derived by combining smaller subgroups, as described in the Data Wrangling report. Below are the groups that were analyzed:

\* Total number of admissions

\* Gender: male, female

\* Race: white, black

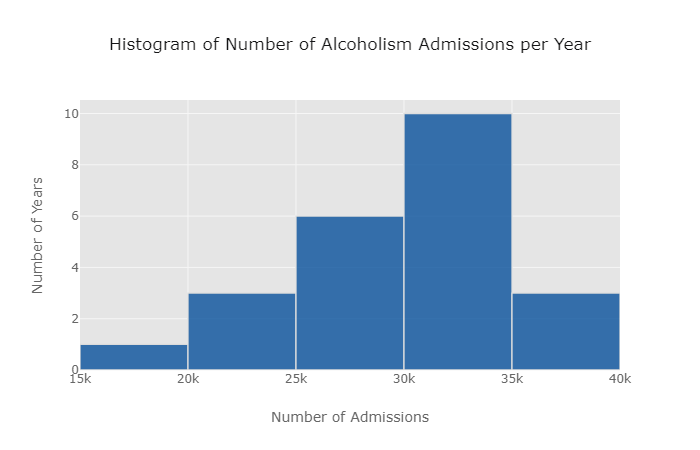
\* Education: Without a high school degree, high school graduate, attended college

\* Marital status: Never married, married now or in the past

\* Employment: working or not working

*Distribution of main target*

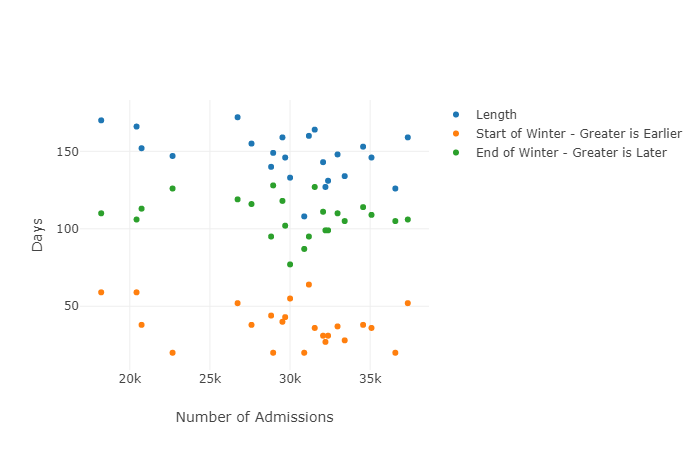
I made a histogram of the overall target variable, the total admissions by year for alcoholism treatment. With only 23 data points, it's hard to generalize from the histogram, but there were 10 years when Massachusetts saw between 30,000 and 34,999 patient admissions for alcoholism. The range on the low end of the histogram was greater than on the high end.



Graphic 5: Histogram of Number of Alcoholism Admissions per Year'

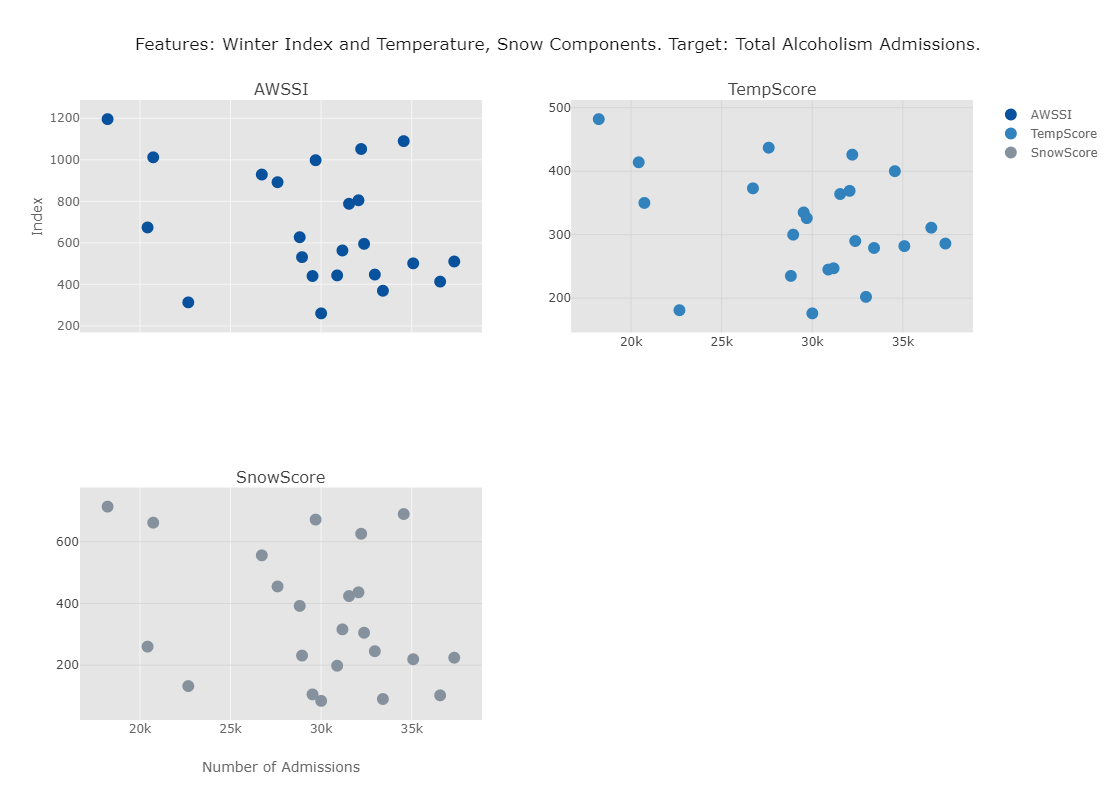
*Scatterplots*

All features were plotted against the main target: Total alcoholism admissions. When the three features relating to winter duration (length, start date, and end date) were plotted on a single chart, the only pattern evident was a slightly negative relationship between the features (in days) and the number of people admitted for alcoholism.

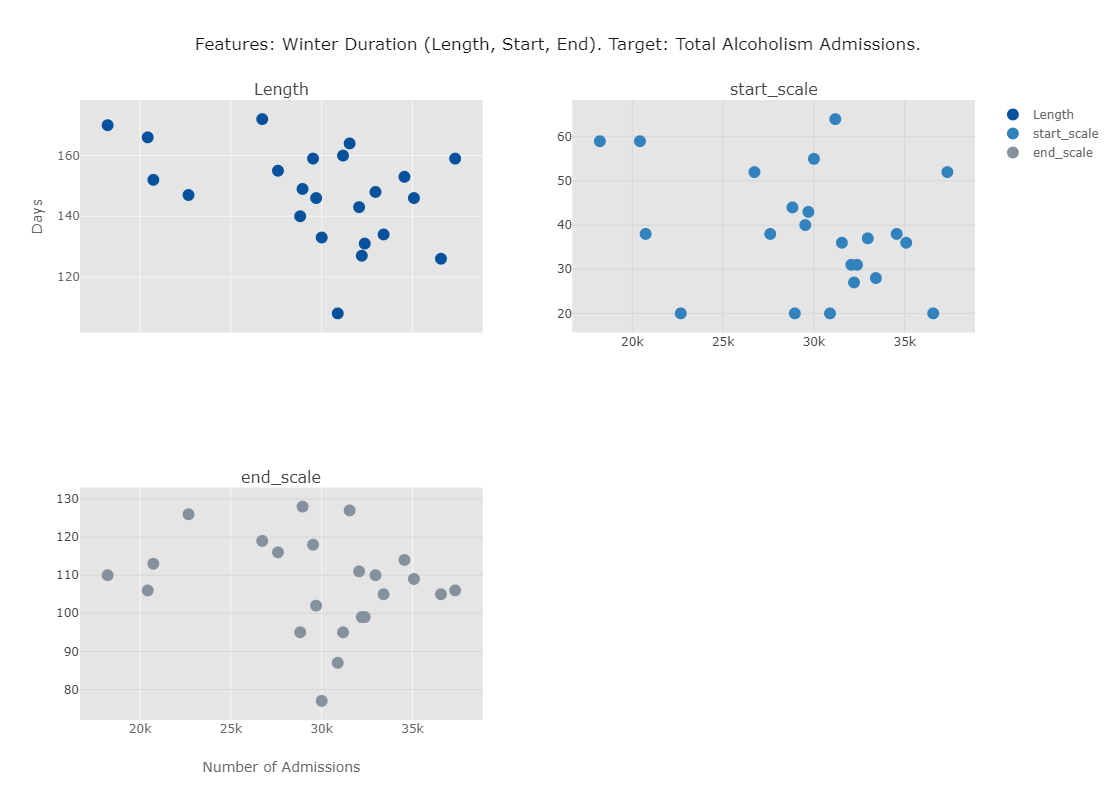


Graphic 6: All duration features plotted with total alcoholism admissions

I grouped the six features into two groups, one for the three winter duration measures, and one for the severity scale measures. Scatterplot matrices of these groups all demonstrated a slightly negative relationship between features and main target.



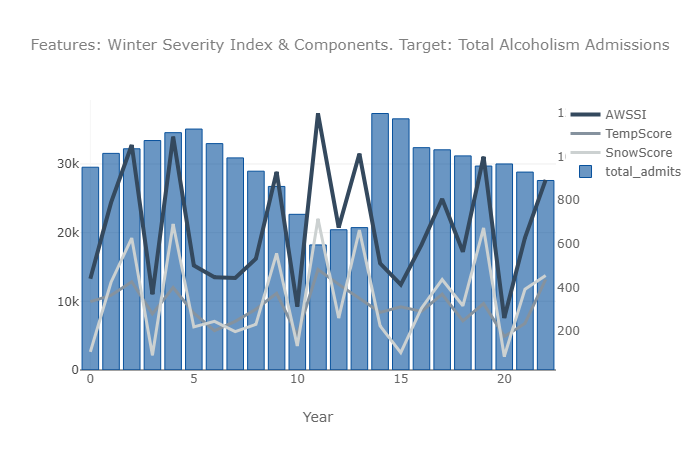
Graphic 7: Features: Winter Index and Temperature, Snow Components. Target: Total Alcoholism Admissions



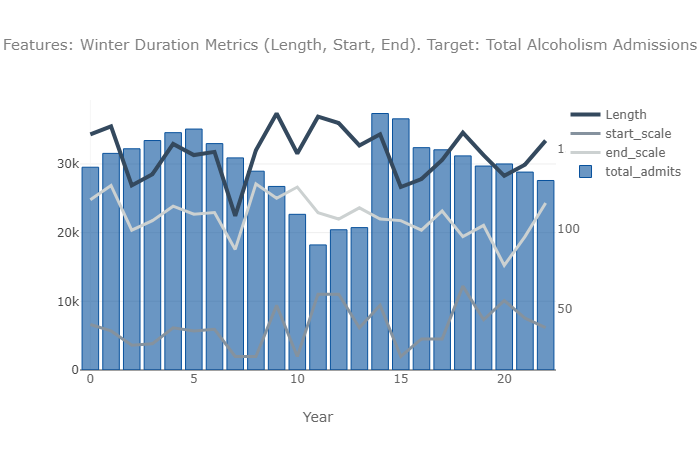
Graphic 8: Features: Winter Duration (Length, Start, End). Target: Total Alcoholism Admissions

*Target bar chart overlaid with feature line charts*

More evidence of the feature/target relationship was apparent when feature line charts were combined with target bar charts. One dip in admissions between 2001 and 2006 tracked moderately with a spike in winter severity for the years 2003 to 2005. A dip in winter severity for the years 2006 to 2008 coincided with an increase in admissions for those years.

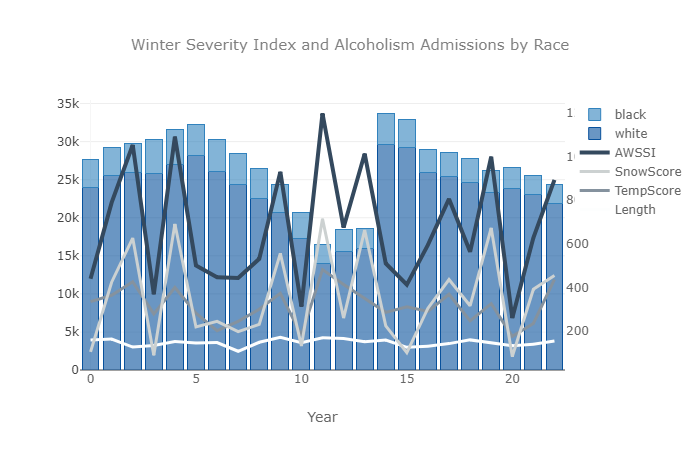


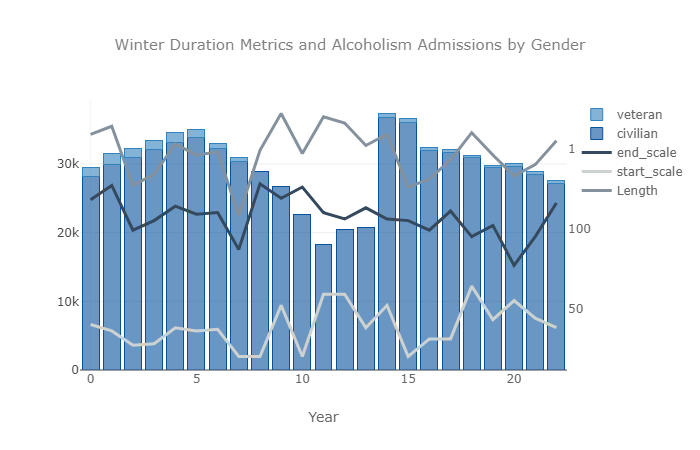
Graphic 9: Features: Winter Severity Index & Components. Target: Total Alcoholism Admissions



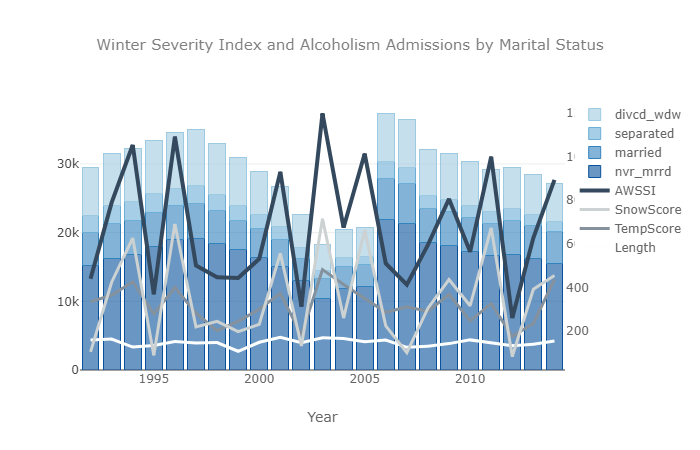
Graphic 10: Features: Winter Duration Metrics (Length, Start, End). Target: Total Alcoholism Admissions

Replacing the main target bars on these charts with stacked bars showing the components of the demographic classifications (i.e., female stacked on male, black stacked on white), there was no evidence of differential winter effects for different subgroups. Instead, changes in the demographic breakdowns over years were likely the result of other, independent trends affecting the overall population. For example, the percent of people admitted for alcoholism who were unemployed grew markedly after 2007, no doubt due to the increases in unemployment in the overall population during the Great Recession (*Graphic 14)*.

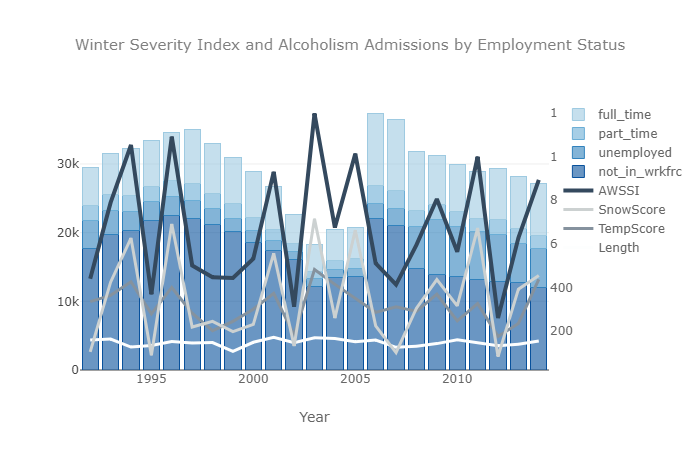
Graphic 11: Winter Severity Index and Alcoholism Admissions by Race



Graphic 12: Winter Duration Metrics and Alcoholism Admissions by Gender



Graphic 13: Winter Severity Index and Alcoholism Admissions by Marital Status



Graphic 14: Winter Severity Index and Alcoholism Admissions by Employment Status

*Revised hypothesis, machine learning methods*

Based on my exploratory data analysis and initial review of inferential statistics, I have changed the hypothesis for this pilot study: \*MILDER\* winter weather may be predictive of demand for treatment of alcoholism.The new hypothesis is also supported by summary statistics and correlation analysis that is not described here due space considerations.

At this point, regression models are clearly a machine learning starting point for this data.