Ross Brown Capstone 1 Report

Substance abuse treatment company struggles to manage staffing

Ross Brown

Capstone 1 Report

Data Science Challenge: Can an alcoholism treatment center's patient intake levels be predicted based on winter weather?

Patient intake numbers vary month to month. Executives ask a data scientist to see if their hypothesis – that winter weather exacerbates alcoholism – can be the basis of a predictive algorithm that they can use to improve staffing decisions.

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.

Predicting patient demand is a critical consideration for Twin Towns:.

- Having enough trained, competent 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.
- 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 with existing or expected 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 near future and beyond.

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 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. And finally, the Mayo Clinic indicates a component of winter severity -- living far from the equator and therefore having less sunlight during the season – is a risk factor for SAD.

Without relying on causal relationships, chaining research findings suggest this relationship, which the Twin Town executives hope to leverage:

More Severe Winter → Increased SAD → Increased Alcoholism

Company executives hypothesize that for patients with nascent or previously treated alcoholism, external factors that exacerbate SAD (such as winter severity) may lead to increase treatment needs:

- Patients managing their alcoholism may experience an increase in symptoms, causing them to seek treatment during severe winter seasons.
- More severe winters may have an effect on treatment demand during the months following that winter, as casual drinkers may become habituated to excessive drinking, and develop alcoholism.
- An unknown combination of these relationships may also affect the epidemiology of alcoholism.

Twin Towns hired this data scientist to conduct a pilot study to determine if the effects could be predictive. There were some challenges when it came to identifying subsets of the data that would produce reliable and generalizable findings from a pilot study.

Data Wrangling

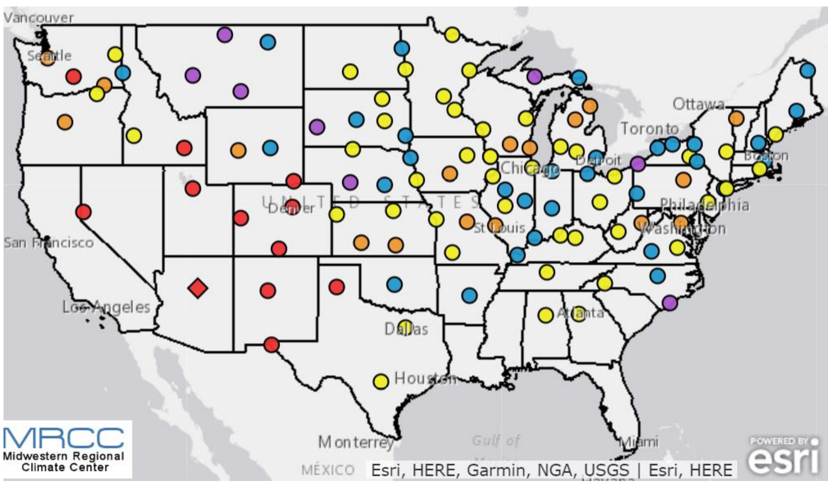
Feature data: winter severity

The Accumulated Winter Season Severity Index 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." 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 meteorological conditions (i.e., hitting a temperature or snow threshold marks the start of any particular winter season by this metric, as does retrospective measurements mark the end), rather than calendar dates, so they are the two additional winter severity data points.

Target 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. 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.

TEDS-A includes demographic characteristics of each person admitted for substance abuse treatment, 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.



Graphic 1

Feature data limitations

1. The AWSSI data is limited to areas corresponding to weather station locations (*Graphic 1*). The degree to which this is a problematic limitation depends on the locations reported in TEDS-A.
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.

Target data limitations

4. The first of the annual datasets is for 1992; the latest is for 2014.

5. 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 (i.e., using alcoholism admissions data for an area that encompasses multiple winter weather patterns) may have limited generalizability.

Selecting project dataset

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 2: Blue Hill weather reporting station.

Read and match data files, construct dataframes

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 23 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

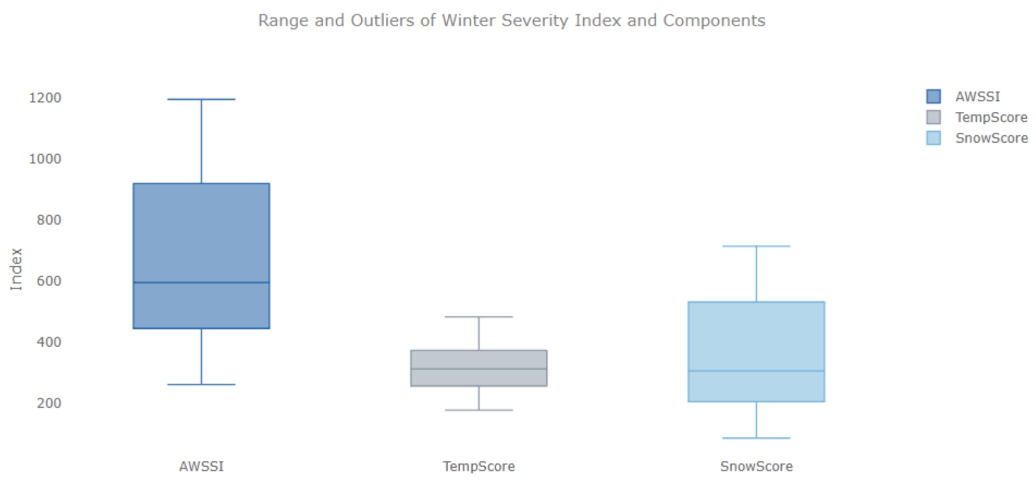
I next counted the number of cells for which there was missing data and listed the variables that had missing data. Of 38 instances of missing data, 32 were in columns included in the TEDS-A files that noted missing data; i.e., the indication that data was missing was missing. The remaining missing data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

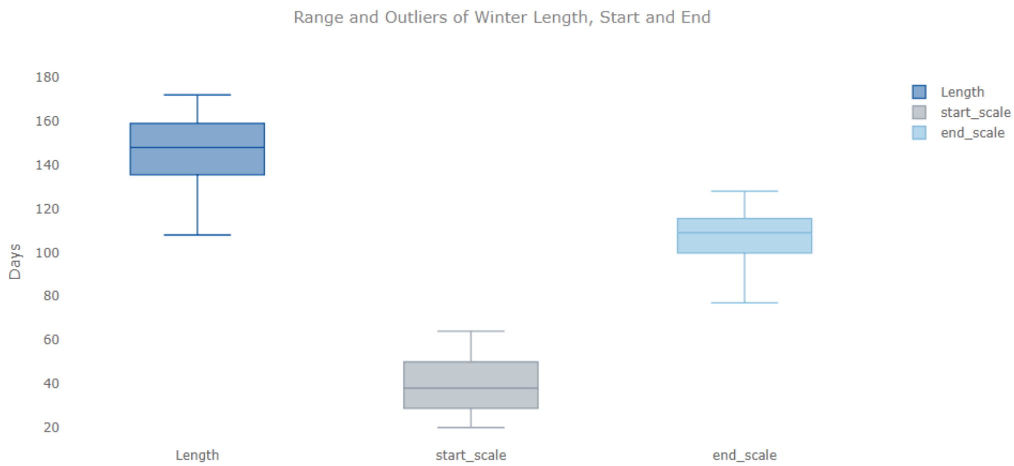
I used the end and start dates of winter to engineer two new features: Measurements in days of how early and how late the winter began and ended each year. Using datetime and timedelta, I first converted the string variables representing these dates into datetime objects. I created another datetime object 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, boxplots

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 (Graphic 3, below. The Python for the following graphics are in the Jupyter notebook labeled “01_Data Wrangling.). 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 3: Weather index box plots



Graphic 4: Duration metrics box plots

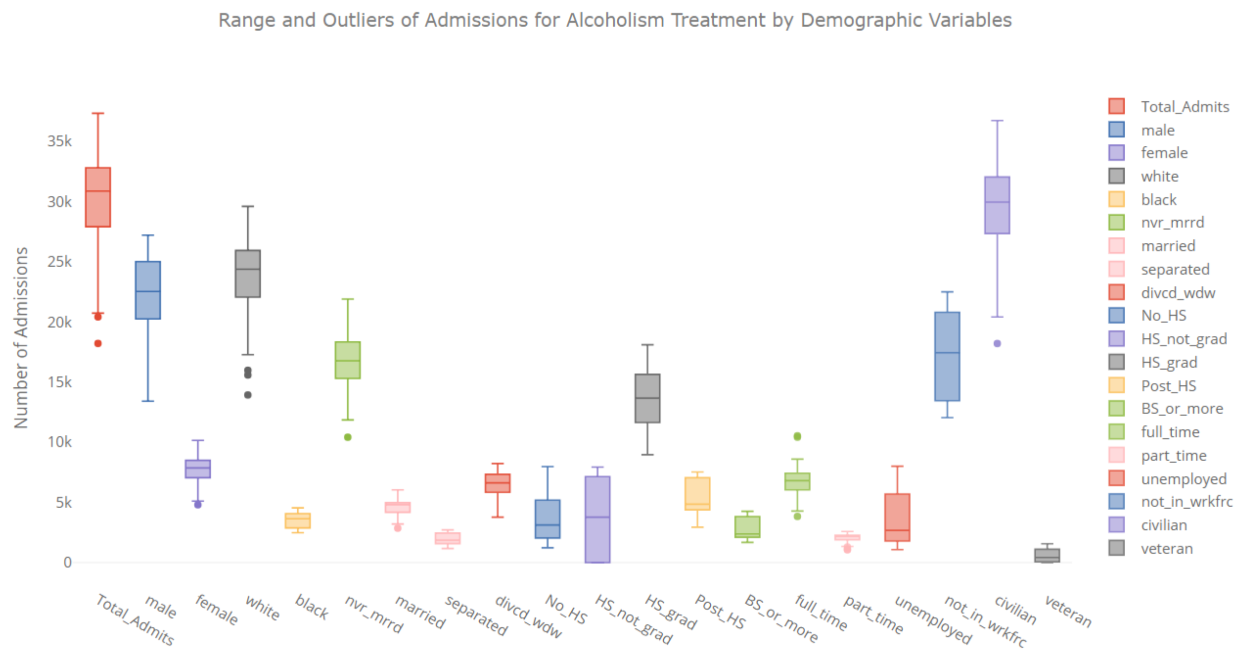
There were outliers for seven of the target variables, the demographic breakdowns of patients admitted for alcoholism treatment. There were outliers in the count of all admissions, for females, whites, people who had never married, full-time workers, part-time workers, and civilians. Considering that 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.

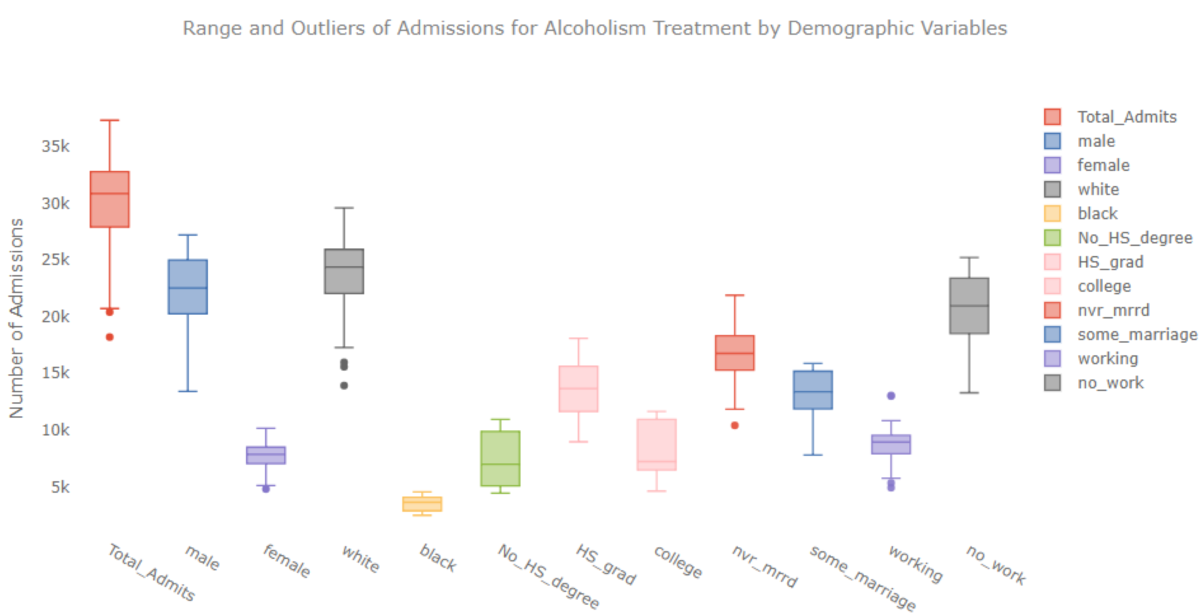
Deleting and consolidating target subgroups

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 5a: Initial target box plots.



Graphic 5b: Final target box plots.

Data Exploration

Final Features and Targets

To review: After engineering new features and consolidating some target subgroups, below are the variable for this project.

<u>Features</u>:

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:

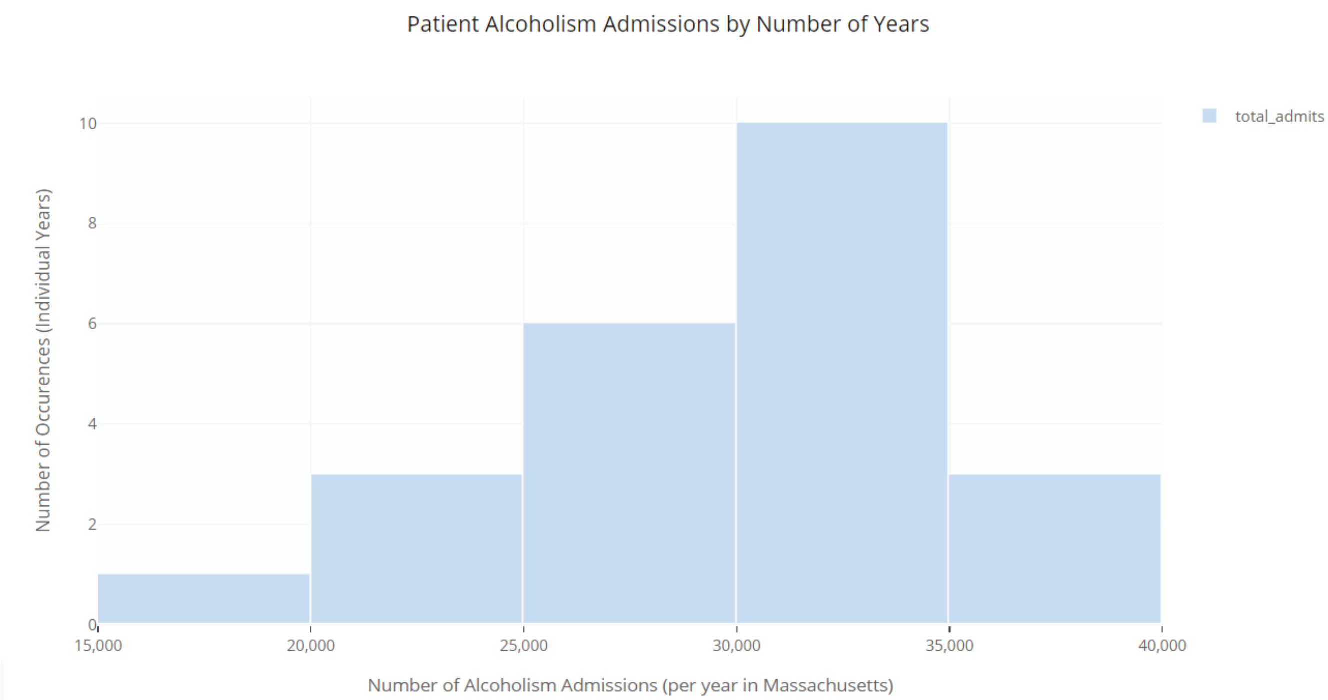
* AWSSI, the overall measure of winter severity
* Temperature score, based on frequency and degree of cold temperatures
* Snow score, based on frequency and amount of snowfall
* Length of winter in days
* Start scale, an engineered feature that indicates how early, in days, each winter started, relative 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

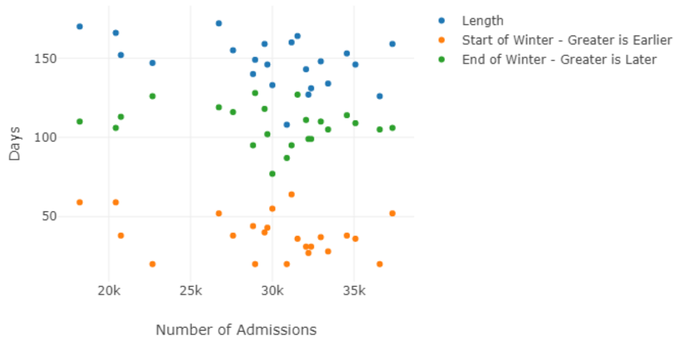
Distribution of main target

I made a histogram of the overall target variable, yearly admissions 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 6: 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.

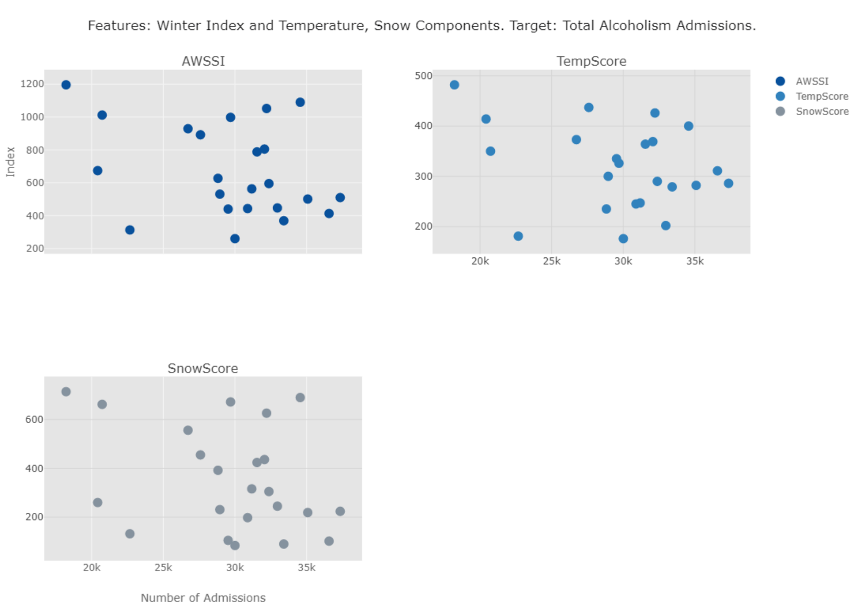


Graphic 7: All duration features plotted with total alcoholism admissions

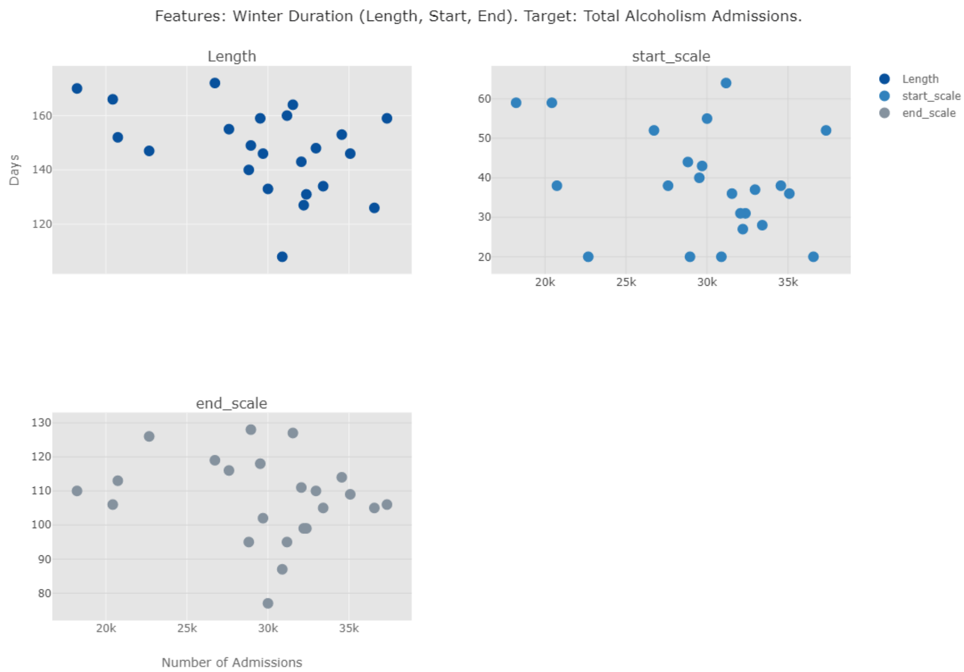
I grouped the six features into two groups:

- the three winter duration measures
- the three winter severity scale measures

Scatterplot matrices of these groups all demonstrated a slightly negative relationship between features and main target.



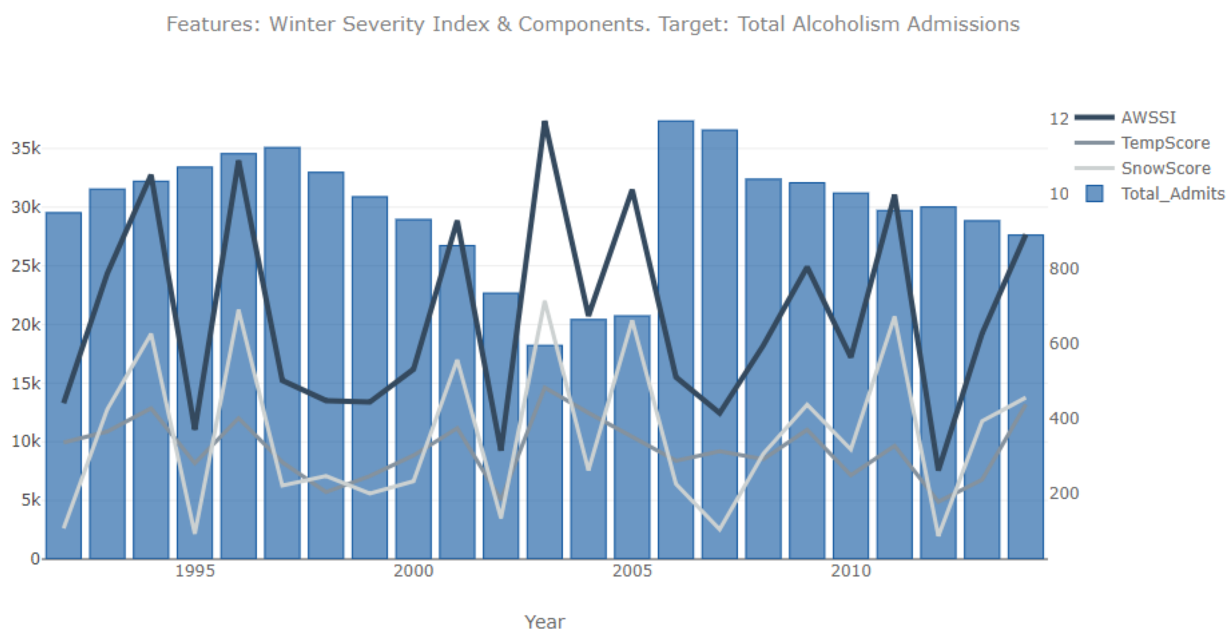
Graphic 8: Features: Winter Index and Temperature, Snow Components.
Target: Total Alcoholism Admissions



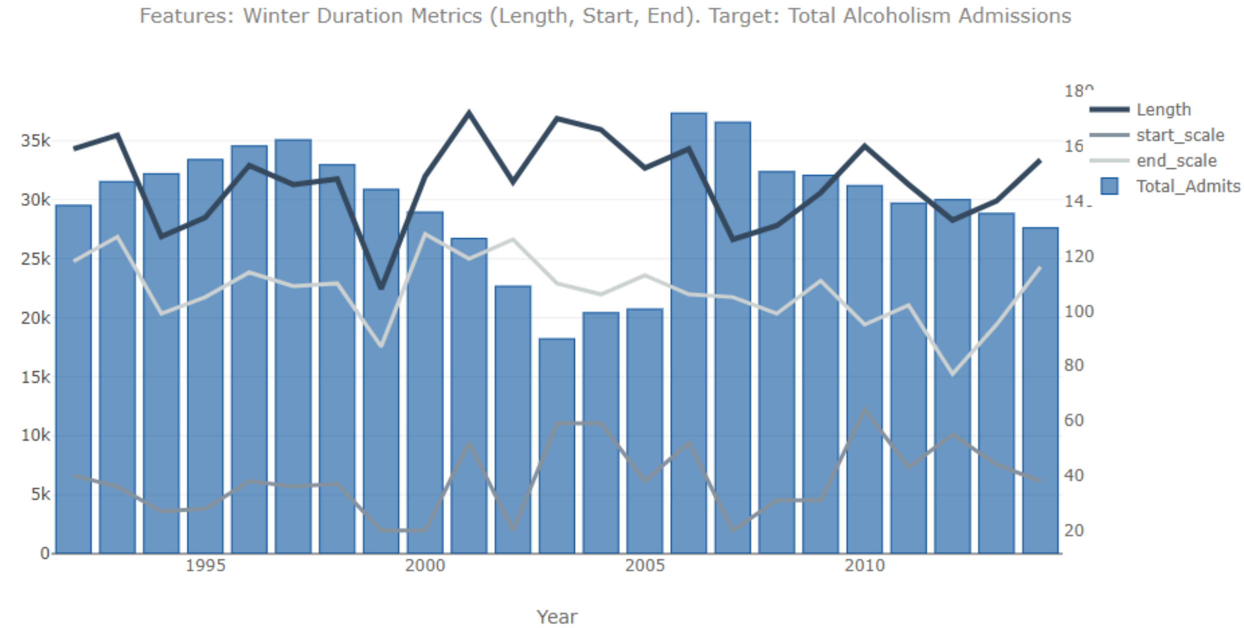
Graphic 9: Features: Winter Duration (Length, Start, End). Target: Total Alcoholism Admissions

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.



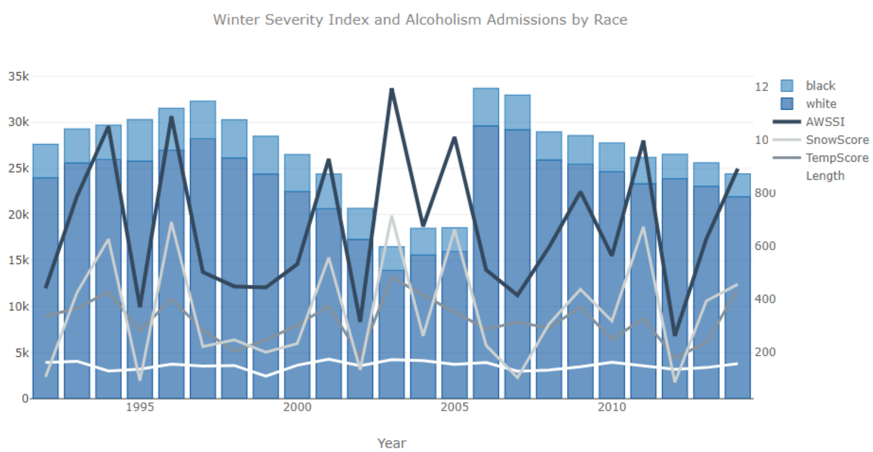
Graphic 10 Features: Winter Severity Index & Components. Target: Total Alcoholism Admissions



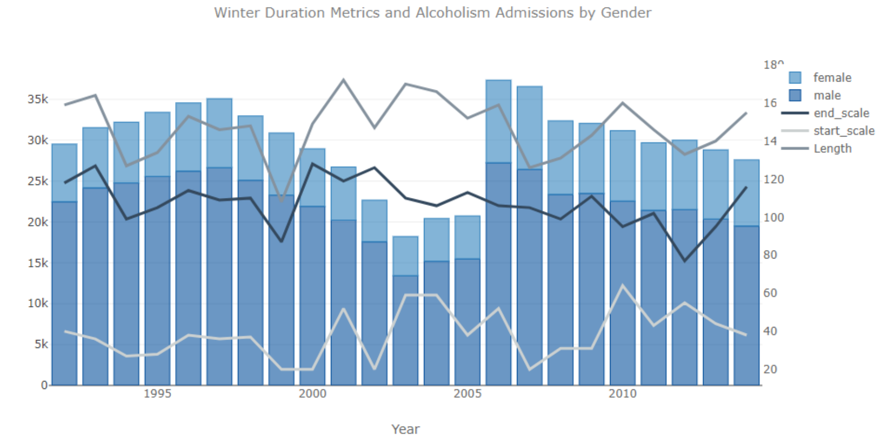
Graphic 11 Features: Winter Duration Metrics (Length, Start, End). Target: Total Alcoholism Admissions

Stacked target subgroup bar charts overlaid with feature line charts

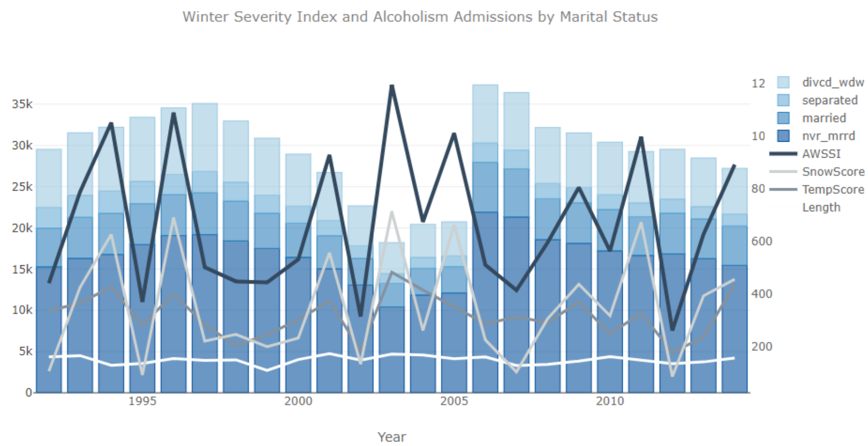
Replacing the main target bars on these charts with stacked bars showing the components of the demographic classifications (i.e., female stacked on male, black stacked on white), there was no evidence of differential winter effects for different subgroups. Instead, changes in the demographic breakdowns 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)*.



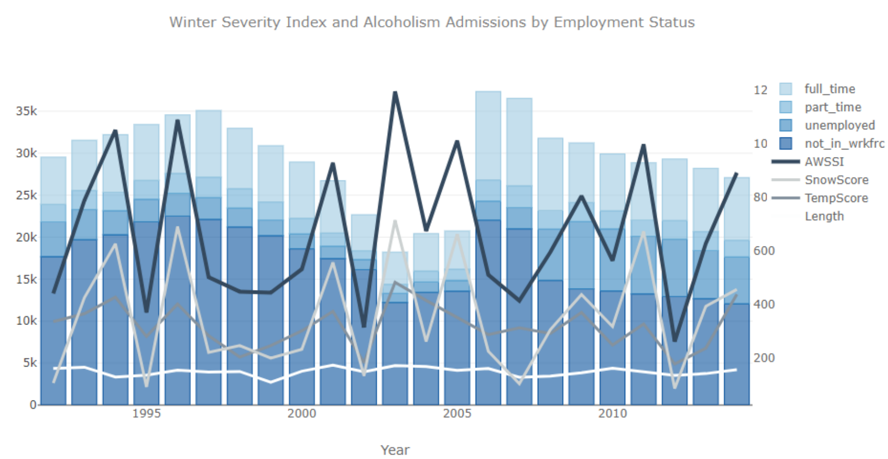
Graphic 12: Winter Severity Index and Alcoholism Admissions by Race



Graphic 13: Winter Duration Metrics and Alcoholism Admissions by Gender



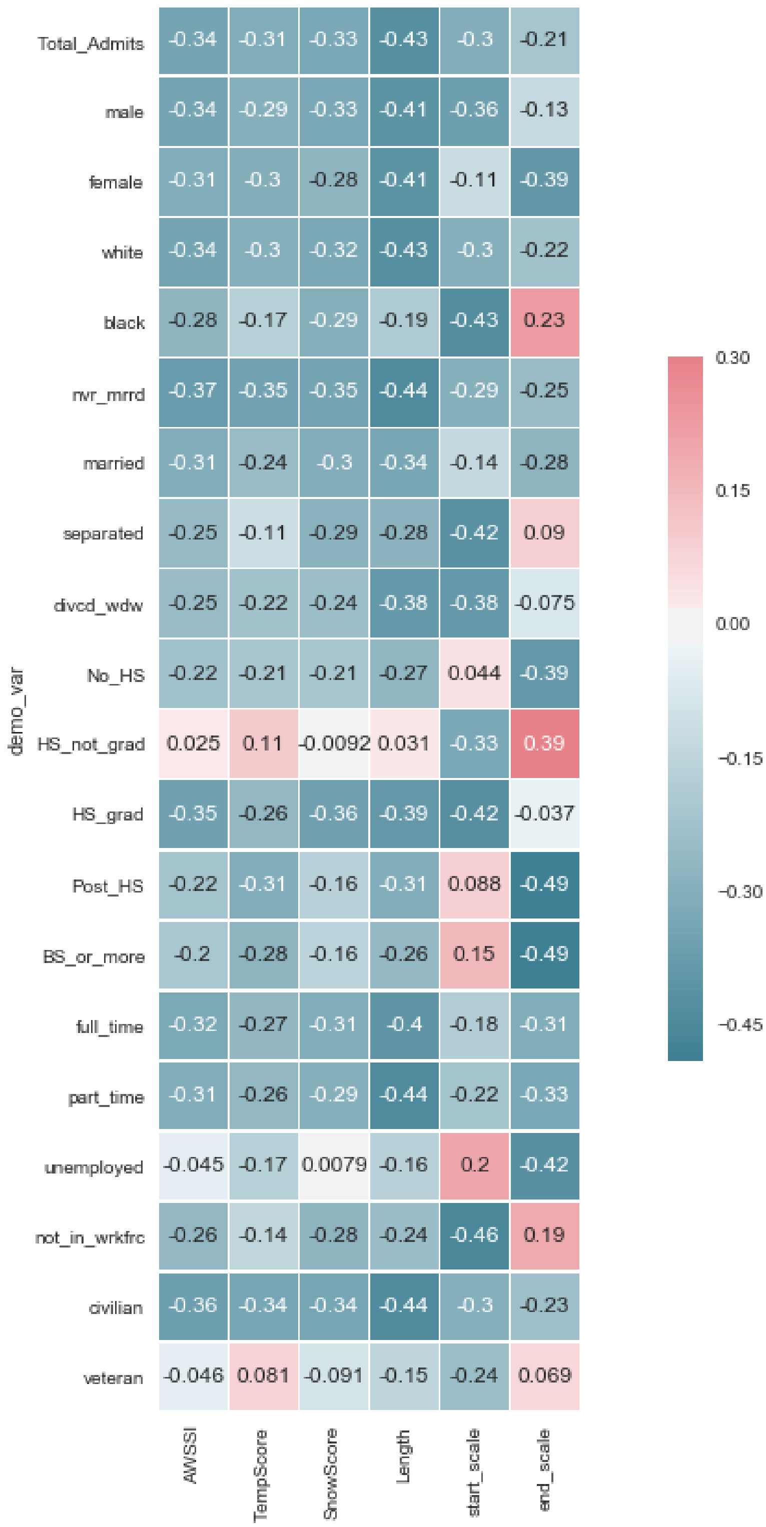
Graphic 14: Winter Severity Index and Alcoholism Admissions by Marital Status



Graphic 15: Winter Severity Index and Alcoholism Admissions by Employment Status

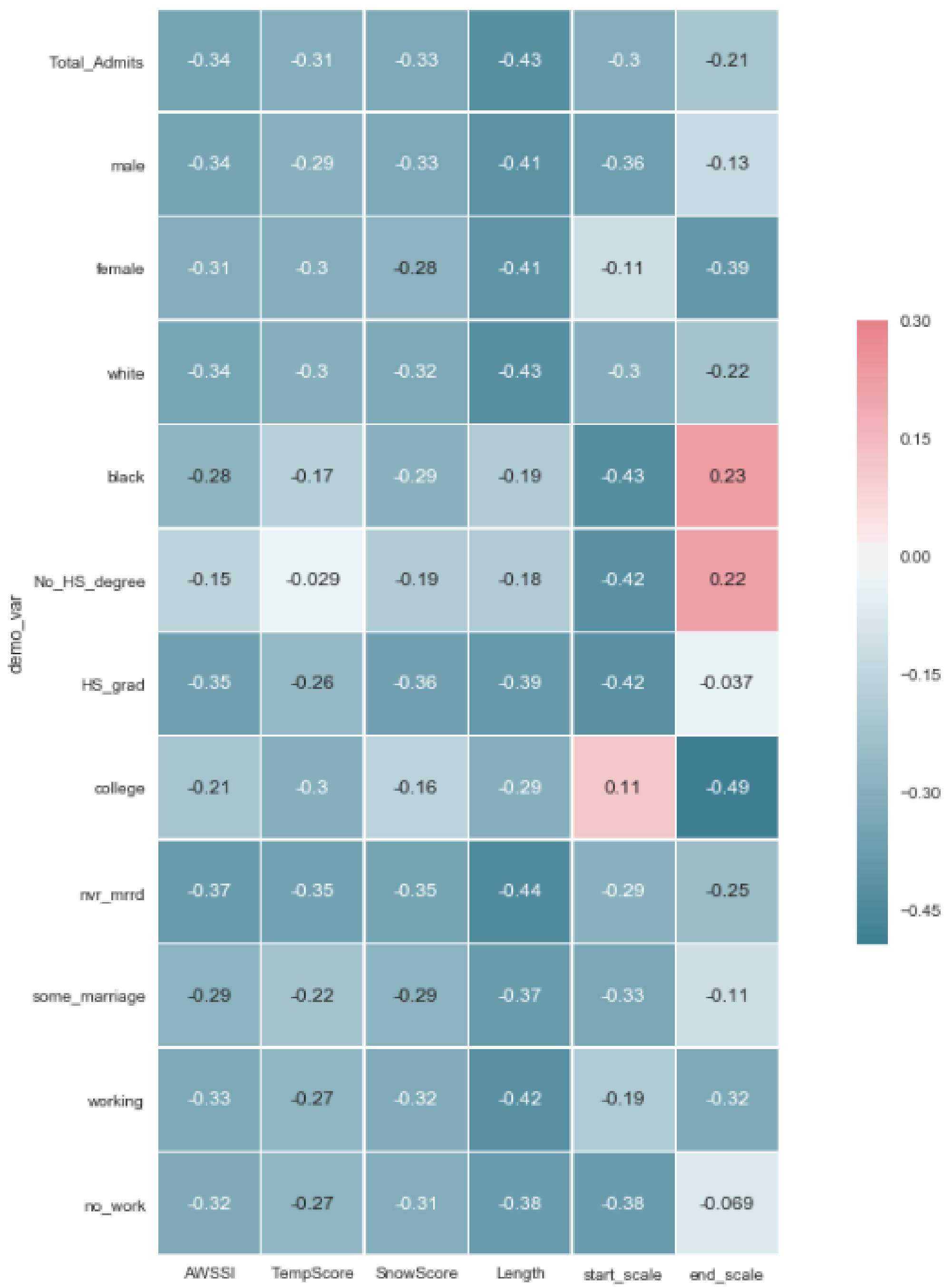
Correlation heatmaps

The preceding visualizations don't support the initial hypothesis; rather these visualizations suggest the relationship between alcoholism and winter severity, if it exists, is in the other direction: *MILDER* winters may be predictive of increased alcoholism admissions. To explore this further, I next created heatmaps showing correlations between features and targets. An initial heatmap, using a diverging color scheme to highlight the directions of correlations, showed correlations between the six main features and the more granular subdivisions of the target (Graphic 16). The majority of correlations were weakly negative (green), and those that were positive, were mostly even weaker than the negative correlations. It's also noteworthy that the demographic subgroups with the smallest *n*s were more likely to have positive correlations, such as those who attended high school but did not graduate, and veterans. This suggests positive correlations for these subgroups may be due to random variance having a greater effect on these small groups, rather than the actual effect of winter severity metrics.



Graphic 16: Heatmap of correlations between six winter severity features and targets, overall admits and the smaller demographic subgroups.

Next I constructed a heatmap of correlations between the same six winter severity features and the larger, consolidated target demographic groups. When the tinier target subgroups in the heatmap above were combined for fewer subgroups, the results supported the hypothesis that positive correlations above were more likely due to random variance. The larger groups showed fewer and smaller positive correlations between the features and targets.



Graphic 17: Correlation heatmap between winter weather features and larger grouped targets.

The 69 out of 72 negative correlations in Graphic 17 range from weakly negatve to moderately negative. The size of the correlations doesn't bode well for developing a predictive model. The largest negative correlation, .49, suggests that only 24% of the variance in the target (in this case, alcoholism admissions for people with some college education) is due to the feature (in this case, the winter end scale). Most of the correlations are smaller, and therefore account for even less of the target variance.

Graphic 18: Correlations between winter weather measures.

A final correlation heat map (Graphic 18), constructed from the various winter severity metrics, verified exptected relationships. Long winters were correlated with winters that started earlier and those that ended later. The winter severity index was correlated with its components. The weakest correlations were across the two groups of winter severity measures, i.e., AWSSI and the length, start, and end scales.

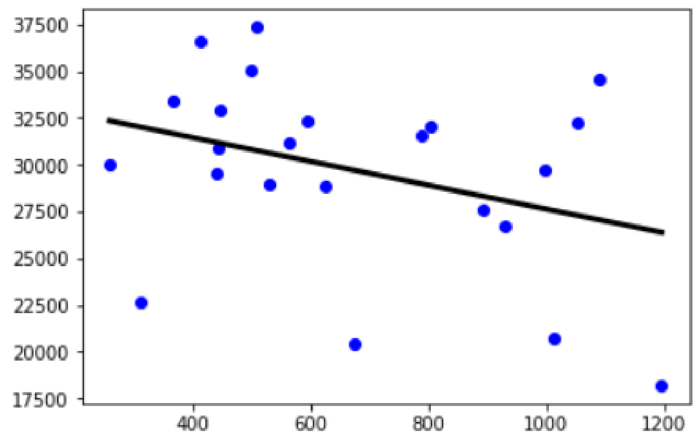
Based on my exploratory data analysis and review of inferential statistics, I have changed the hypothesis for this pilot study: *MILDER* winter weather may be predictive of demand for treatment of alcoholism.

Machine Learning

The Python code for this section is in the Jupyter notebook titled, "03_Machine Learning." With a continous variable for a target, machine learning for this project is limited to regression.

1. Linear regression with a single feature, the overall winter severity index

The revised hypothesis – that the weather index is negatively predictive of admissions for alcoholism – was tested by conducting regression using only the AWSSI and overall alcoholism admissions. The plot of that regression line, Graphic 19, shows that the AWSSI data are sparse and not especially predictive of the target.



Graphic 19: Regression line for AWSSI and alcoholsim admissions.

The R^2, at 0.12, indicates that with linear regression using only the primary feature, the overall winter severity index, very little variance in the target is predicted by the feature. A train/test split analysis computed a test set score of only -0.01, a harbinger of all subsequent analyses. The conclusion for this first analysis is that the winter severity index alone has no value for developing a predictive algorithm for alcoholism admissions.

For this and subsequent regression analyses, linear regression coefficients (slopes) and intercepts were computed. Since there is a negative relationship between all features (winter weather severity index, temperature score, snow score, length, start scale, and end scale) and the target and its larger subgroups (overall alcoholism admissions and demographic breakdowns of the overall count), most of the regression coefficients were negative; exceptions are noted.

2. *Ordinary least squares regression using all available features indiscriminately*

   a) *Large RMSE*

The six features are: AWSSI, TempScore, SnowScore, Length, start_scale and end_scale. This regression yielded worse results: An R2 of -0.03. Additionally, the root mean squared error, the standard deviation of the differences between predicted values and observed values in the same units as the response variable (in this case, in units of admissions for alcoholism), was quite high. It was 4,712, for a target with values ranging from 18,000 to 35,000.

   b) *Improved training score due to overfit*

This analysis generated a training set score of 0.34 and a test set score of -0.14. Compared to to linear regression using only one feature, using all available features increases the training set score, as would be expected with a more complex model that more accurately describes the specific characteristics of the training sample, but does poorly on another sample, i.e., the test data. This is because the model overfits to the training set, and is not generalizable to the test set, or other data.

   c) *Small coefficients reflect minor effects of most features.*

The coefficients are shown below. Most were very small, reflecting small effects on the target, and the positive slopes were the smallest in terms of absolute value. These small coefficients reflect the minimal effect on the target these features have when combined with all available features in this regression analysis. Notably, in this analysis, the features for length of winter and the start of winter were larger by a factor of 10 relative to all but one of the other coefficients, and by a factor of 6 for that last one.

AWSSI	-6
TempScore	-11.61
SnowScore	5.61
Length	-70.94
Start_scale	-76.93

End_scale	6

The intercept is: 48716.86

3. Ridge regression

Considering the small coefficients calculated in the linear regression analysis above, ridge regression would be expected to produce similar train/test scores. Indeed, they're identical: training set score of 0.34, test set score of -0.14.

That's because ridge regression coefficients are chosen to be as small as possible, as well as predict on the training data. The features used for this project have such small predictive strength as it is – which means the coefficients are small – the addition of that as a coefficient selection criteria in the ridge regression analysis made little difference.

4. Lasso regression

Using an alpha of 1.0 and increasing the maximum iterations to 100,000 produced a model that converged and showed the features for length of winter and start of winter to be the two most predictive under the regression model that constrains some coefficients to zero, i.e., ignores them entirely. Obviously, the outcome from the lasso model, indicating the relative predictive strength of winter length and the start date, is just what ordinary least squares regression indicated. These similarities reflect the minimal predictive power of the features in this analysis; the results all come down to the same meaningless regression outcomes.

Conducting ordinary least squares regression using the two most predictive features identified by the lasso regression resulted in the coefficients and intercept below, shown with the coefficients and intercept resulting from OLS that included all features (2, above):

	Two Features Only	With Other Four Features
Length	-104.85	-70.94
Start_scale	-59.81	-76.93
Intercept:	47749.63	48716.86

Similar intercepts and slightly more variance associated with length and start of winter for the regression using these features alone is to be expected. The variance associated from the other for features that the lasso regression model reduced to zero is now associated with length and start scale.

All results:

lr.coef_	-104.85, -59.81
lr.intercept_	47749.63
Training set score	0.27
Test set score	-0.05
R^2	-0.05
Root Mean Squared Error	6315.72

The RMSE for this model, with only two of the six features, is almost 50% greater than the RMSE for the regression model using all available features. This is explained by the overfit of the all-feature model to the train data set. This model was more generalizable, with a R^2 that was greater than the overfitting all-feature model. However, this difference was only .02, and the two models are both beyond useless for predictive purposes, so the difference is essentially meaningless, and likely well with the error associated with the parameters.

Takeaways, implementing findings, next steps

Data disaster

The characteristics of the data chosen for this project (a continuous target variable) restricted the number of machine learning techniques that could be implemented to regression models, which failed to produce predictive algorithms. The potential effect size of winter severity on alcoholism was not likely to be apparent in a sample consisting of 23 data points. The data was not available for a longer period, nor was it available in increments smaller than annual tallies; if such data were available, an effect large enough to be predictive may have been evident.

The data for the project was chosen for characteristics relating to study design, as well as novelty, i.e., the winter severity index and the manner by which it is calculated was interesting, and did not seem to have been used in other projects or research. In the future, I'll be sure to weigh these considerations against those more important to data science, such as the amount of data and if a classification is possible.

Considering the weakness of the effect found by this project, it is possible that a multiyear economic factor, the Great Recession, may have introduced too much variance in the target relative to that from the features studied.

Improved prospects for future research

Studies seeking to build on this one could seek out a proxy for alcoholism admission as a target; perhaps hospital data could support inferences about overall alcoholism. More granular data may be available from smaller government entities, such as cities and states, or private organizations, such as health insurance companies or providers. While I first thought the Blue Hill, MA, weather data and the comprehensive and complete Massachusetts alcoholism data were the best possible alignment of feature and target that precluded confounding effects from multiple weather stations or weather variance within an alcoholism geographic reporting designation (i.e., regions, groups of states), now I think there may be a better approach. Taking all of the weather reporting data for areas affected by winter and selecting alcoholism reporting categories to match that may be aligned and complete.

Future research may benefit from more qualitative data: Surveys attempting to capture the processes related to weather and the development of alcoholism, or ratings by patients to capture degrees of effects from things like temperature, snow, duration, and onset and end of winter.

Client benefits

Twin Towns benefits from the findings that

- the effect may be the opposite of what was expected
- may be a weak effect
- the effect was largely consistent across target subgroups and feature components.

With this knowledge, Twin Towns can better

- target its efforts to identify the implications on its own operations (it may have been missed if Twin Towns were looking for an effect in the opposite direction)
- expect winter weather to be a less important operational factor
- understand that if the effect were identified, it would likely cross all patient demographics.

Concrete steps for Twin Towns.

TT should first determine whether the weak-to-moderate winter severity/alcoholism relationship will have any bearing on its costs, staff morale, and patient treatment outcomes. Overlaying the publicly available winter severity data with its own patient admissions and staffing needs would tell Twin Towns whether the benefits in these areas are limited to the margins, or are worth making changes to achieve operational benefits.

TT should collect and track the daily, cumulative winter severity data that is publicly available, and enter it in an Excel spreadsheet formula that will automatically generate line charts showing ongoing changes in winter severity on a daily basis. Tweak staffing levels based on these trends and monitor outcomes in terms of the match between staff levels and patient demands, noting staff surpluses and shortages as they occur. Review and revise as needed.

Winter start date had was one of the most predictive winter severity metrics, and it is available prospectively; the date is announced when it occurs. Twin Towns need only consider the date relative to historic mean, median, and outliers, as well as rolling averages, and use this information to make staffing decisions. Review and revise as needed.

Use brainstorming and anecdotal data from staff observations to develop simple questions or survey response items – directly relating to winter weather effects – that can be easily incorporated into intake and treatment protocols. Simple, affordable analysis of the data, with close attention to periods of mild winter weather and heavy patient intake, can point to actionable patterns. Survey patients about the dynamics of alcoholism onset, tying survey questions to winter weather patterns.