Weather.com Practicum Report

Conor O'Sullivan[1], Anuj Saxena[2], Deeksha Chugh[3,*]
1 Conor O'Sullivan Analytics/School of Management/USF, San Francisco, CA
2 Anuj Saxena Analytics/School of Management/USF, San Francisco, CA
3 Deeksha Chugh Analytics/School of Management/USF, San Francisco, CA
∗ **E-mail: ccosullivan@dons.usfca.edu, asaxena2@dons.usfca.edu, dchugh@dons.usfca.edu**

Executive Summary

The target audience for this report is companies wishing to enhance their sales predictions with weather data. In this report, we show how we used machine learning and clustering to match different types of products by their seasonality and predict weekly sales of consumer products. The outcome of this research was the models we created that will eventually be used to demonstrate the worth of Weather.com's consulting arm. For information concerning where the data came from, see section 1. To read about the product clustering, see section 2. The report covers extra variables in section 3 and building heatmaps in section 4. Section 5 details the process we undertook for feature selection.

Introduction

How does weather affect the sales of consumer products? Our team worked to answer this question during the fall practicum project at Weather.com. This question was motivated by the fact that the data science group at Weather.com has branched out beyond weather forecasting. The company now provides consulting services to companies using its weather data. Our job was to investigate how helpful weather data can be in terms of providing an accurate forecast for products sold by these potential clients. This report details the process our team took to extract the correct information from Weather.coms API, engineer variables, perform clustering and finally model our data.

Section 1. Data Acquisition

We initially received two sets of data: one of weather variables and one of sales data. The sales data was purchased from Nielsen and contained the number of units sold and the prices for beer, bottled water, coffee, cough medicine, insect spray, lawn and garden supplies, lotion, soft drinks, sunscreen, and soup. This data was for the following cities: Atlanta, Boston, Chicago, Dallas, Detroit, Los Angeles, Minneapolis, New York City, Phoenix, Seattle. The sales amounts were recorded weekly from June 2009 to June 2012. The weather data was taken from Wunderground and Weather.coms own API, DataCloud. It was also in weekly form and contained the following variables: temperature, dewpoint temperature, feels like temperature, difference from normal temperature, humidity, precipitation, cloud cover percentage, and wind speed.

Our data was not complete, however. Thirteen weeks of weather data were missing from each of the cities. At the same time, Detroit weather data was missing for all 3 years. Our team had to first find the appropriate variables to collect from Wunderground and DataCloud, then finally collect the data through a python script. This turned out to be a valuable lesson in gaining enough domain knowledge to understand the data we needed to extract.

Section 2. Data Clustering

Our efforts began with clustering the products. We resorted to k-medoids clustering to find a stable result. We chose k to be three because there seemed to be three main seasonal trends in our products: summer, winter, and a less extreme beverage trend exhibited by beer and soft drinks. The clustering strategy that produced the best results was to cluster by each kind of product in each city. This method yielded the following as centroids: beer in Chicago, soup in Chicago, and insect spray in Atlanta. These results drove us to focus on these product categories for our models.

In English, we made 100 vectors and clustered them into three groups. Each vector was a possible item-city combination. These vectors had 156 dimensions, one dimension for each week's sale number. Some items, like insect spray and sunscreen, had similar purchasing trends so they ended up in the same cluster. There was a summer cluster, a winter cluster, and a third cluster specific to beverages like beer, bottled water, and soda.

To cluster cities, we again used k-medoids. We took each city's product clustering vector and clustered those to find which cities grouped their products similarly. This gave us three clusters: New York City, Los Angeles, and one containing the rest of the cities. We decided to choose Chicago as the representative of the large cities group because it was in two of the centroids of the product clustering. We chose Atlanta as a representative of southern, warm climate cities. Seattle was also chosen because it exhibited some unique seasonality as far as lotion is concerned, possibly indicating that its other products would exhibit some further unique seasonality when modeled.

On the next page, figure 1 displays the three main trends in our data. Its clear that soup sales has peaks in the winter, while beer and insect repellent sales peak in the summer. There is also a clear bump in beer sales again in the last two months of each year, probably provoked by the holiday season.

Section 3. Extra Variables

In order to add predictive power to our model, our team sought to engineer variables from our data and gather some from free online sources. We had a hunch that peoples sickness and soup sales were tied together, so we turned to Google Flu Trends for a weekly count of flu related searches in a metropolitan area. This exhibited a positive correlation of 0.65 with the number of soup cans sold and was later selected in a few of the models by MSE-minimizing feature selection. After this showed success, we looked at general economic indicators like the Consumer Price Index and Consumer Confidence Index, however these did not correlate with any of the sales numbers we had.

We also engineered some helpful time variables. Our personal experience reminded us that people often buy beer in holiday seasons. We added a variable that indicated the number of holidays per week, with 2 being the lowest amount per week. For holidays like Memorial Day, we increased it to 3 and then for Thanksgiving and the New Year, we increased it to 4. This variable also added predictive power and was picked up by some of the beer model feature selections. We brought in a factor for the month that the recording was made as well as the number of week in the year it was. Both of these were found to add to our random forest models, as well.

Section 4. Data Exploration

Images below show us linear relationships of the variables that are weather related, product related and extra variables. Since temperature is the most correlated with the unit sales, we present the powers of

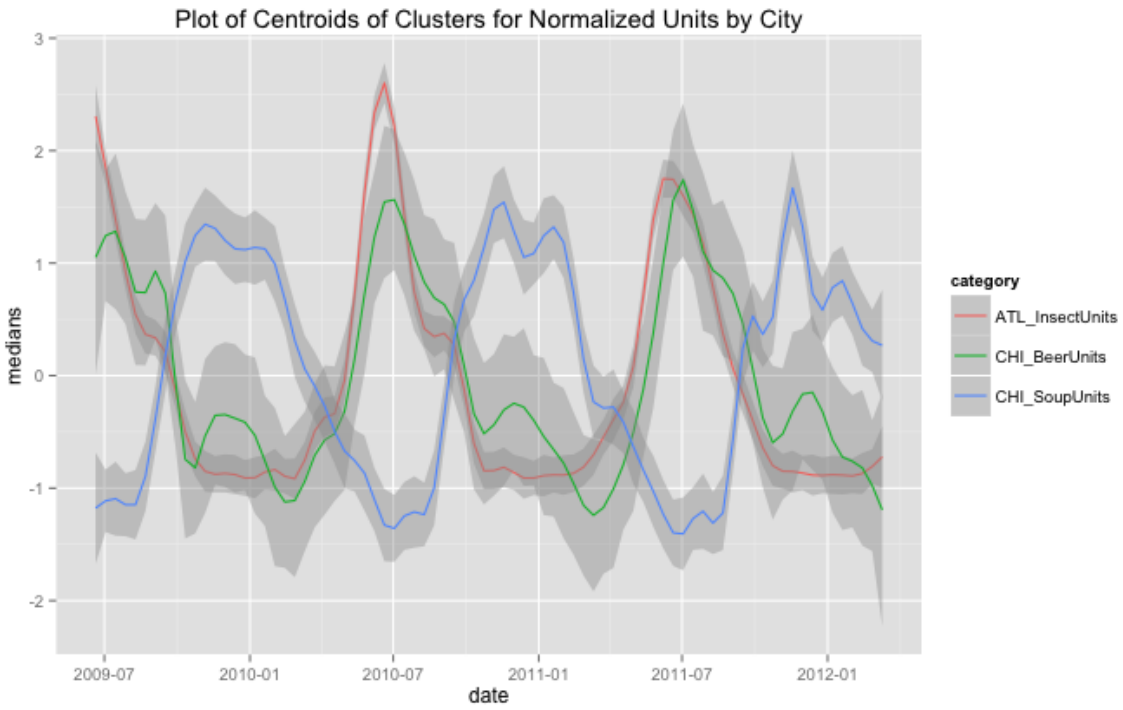
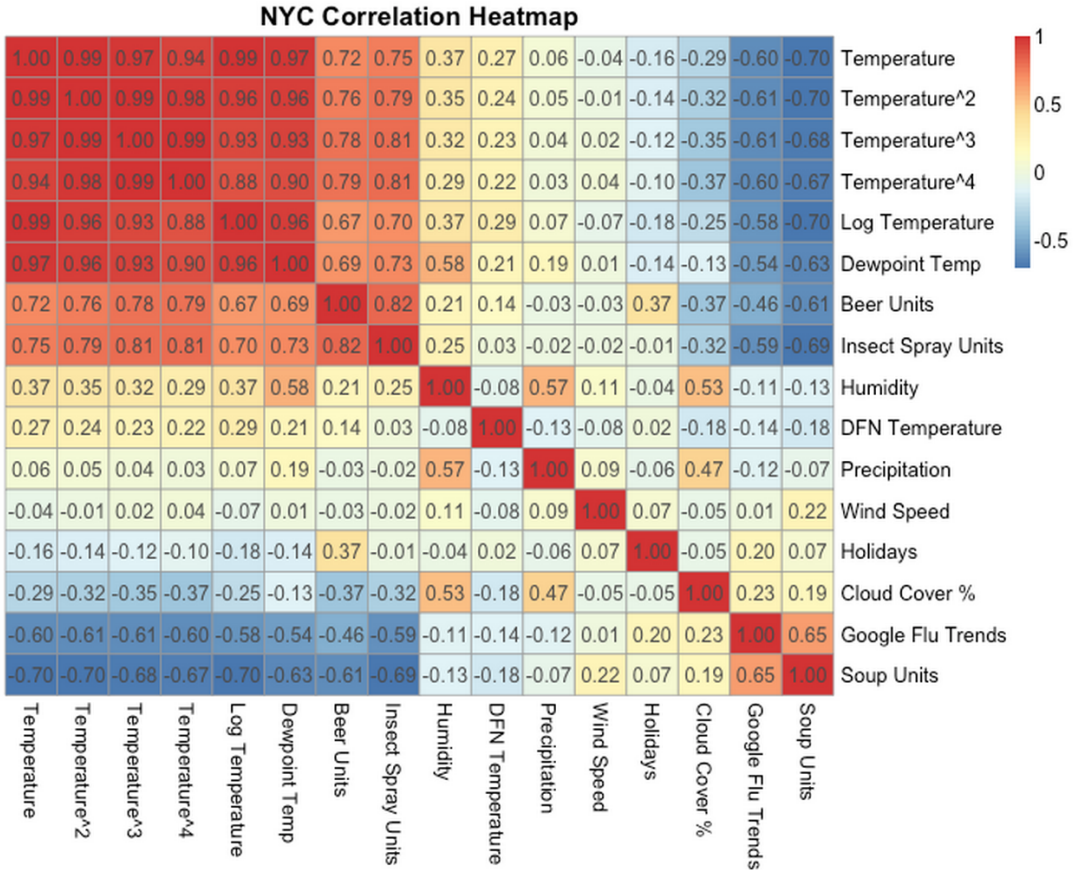
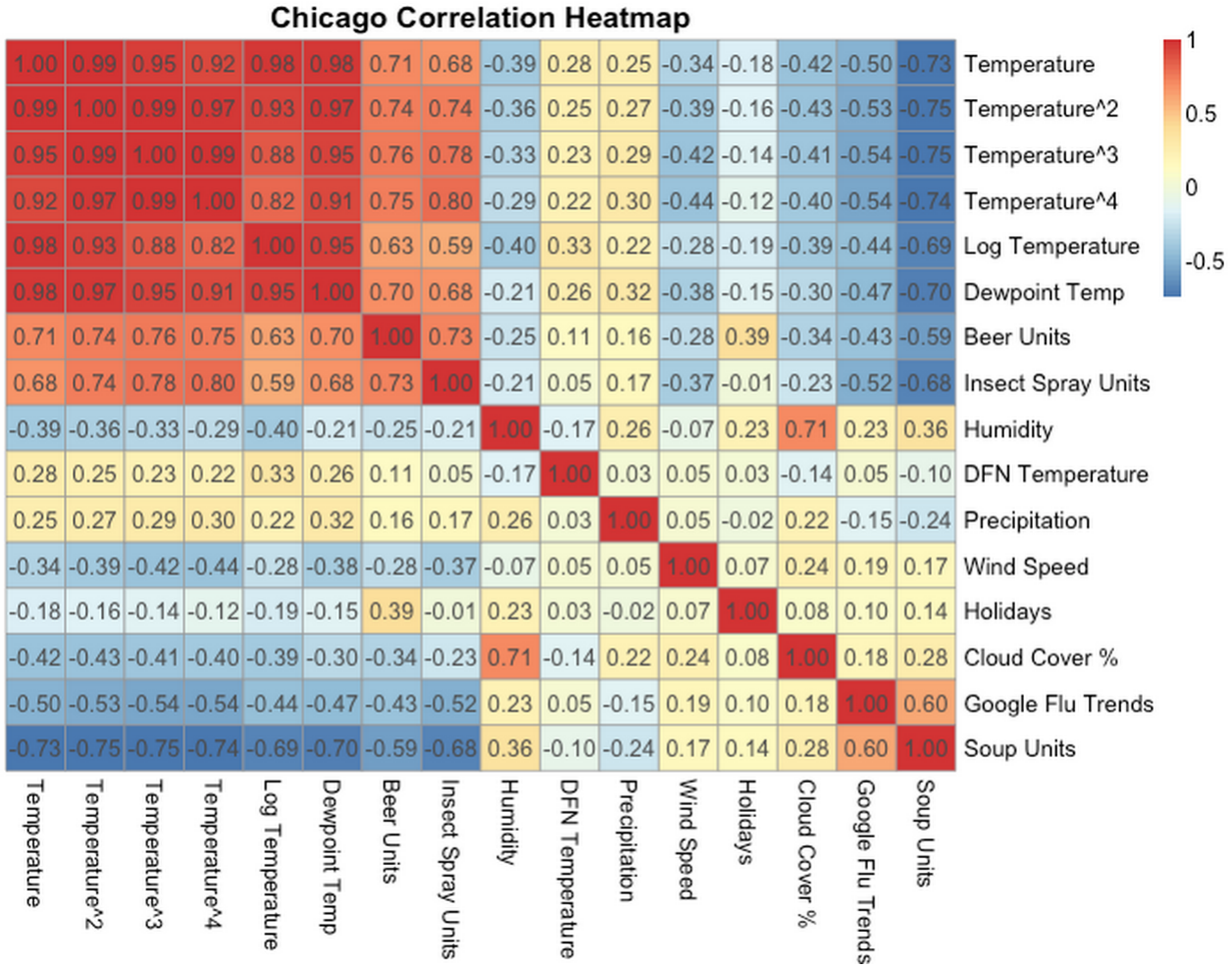


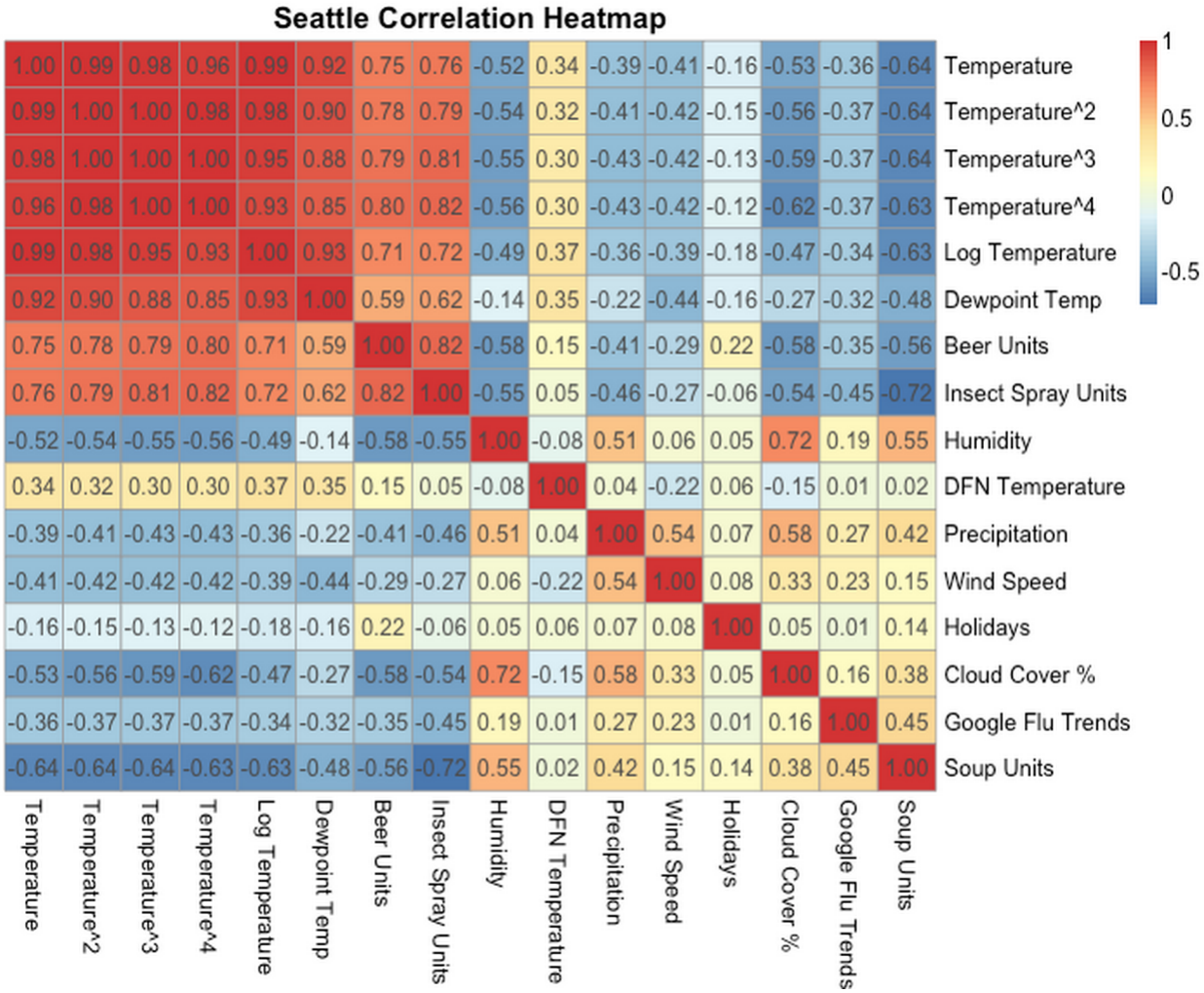
Figure 1. These three trends are the centroids of our three clusters, made from the 100 possible product-city combinations.

temperature as well and see its linear relationship with other variables. In general, it is easy to spot the summer and winter product categories here. Sales of cans of soup in New York city show strong negative correlation with temperature and different powers of temperature. Beer and insect repellent, on the other hand, clearly exhibit positive correlation with temperature. Soup units have a considerably strong correlation with the google flu trends. We can also see that the units of beer sold are mildly correlated with the holiday variable.

Looking at a few of the cities, we can see small differences that influenced our decision considering model building. In New York, the soup sales have a weak negative correlation with humidity (-0.13) whereas in Chicago, they are positively correlated (0.36). In Seattle, beer has a negative correlation with humidity (-0.58) and in New York, beer is weakly positively correlated with humidity (0.28). The weather in each of these cities affects sales in unique ways. Because of this, our team decided to create unique model for each product-city combination to capture these individual trends.




Chicago Correlation Heatmap
1.00 0.99 0.95 0.92 0.98 0.98 0.71 0.68 -0.39 0.28 0.25 -0.34 -0.18 -0.42 -0.50 -0.73 Temperature
0.99 1.00 0.99 0.97 0.93 0.97 0.74 0.74 -0.36 0.25 0.27 -0.39 -0.16 -0.43 -0.53 -0.75 Temperature^2
0.95 0.99 1.00 0.99 0.88 0.95 0.76 0.78 -0.33 0.23 0.29 -0.42 -0.14 -0.41 -0.54 -0.75 Temperature^3
0.92 0.97 0.99 1.00 0.82 0.91 0.75 0.80 -0.29 0.22 0.30 -0.44 -0.12 -0.40 -0.54 -0.74 Temperature^4
0.98 0.93 0.88 0.82 1.00 0.95 0.63 0.59 -0.40 0.33 0.22 -0.28 -0.19 -0.39 -0.44 -0.69 Log Temperature
0.98 0.97 0.95 0.91 0.95 1.00 0.70 0.68 -0.21 0.26 0.32 -0.38 -0.15 -0.30 -0.47 -0.70 Dewpoint Temp
0.71 0.74 0.76 0.75 0.63 0.70 1.00 0.73 -0.25 0.11 0.16 -0.28 0.39 -0.34 -0.43 -0.59 Beer Units
0.68 0.74 0.78 0.80 0.59 0.68 0.73 1.00 -0.21 0.05 0.17 -0.37 -0.01 -0.23 -0.52 -0.68 Insect Spray Units
-0.39 -0.36 -0.33 -0.29 -0.40 -0.21 -0.25 -0.21 1.00 -0.17 0.26 -0.07 0.23 0.71 0.23 0.36 Humidity
0.28 0.25 0.23 0.22 0.33 0.26 0.11 0.05 -0.17 1.00 0.03 0.05 0.03 -0.14 0.05 -0.10 DFN Temperature
0.25 0.27 0.29 0.30 0.22 0.32 0.16 0.17 0.26 0.03 1.00 0.05 -0.02 0.22 -0.15 -0.24 Precipitation
-0.34 -0.39 -0.42 -0.44 -0.28 -0.38 -0.28 -0.37 -0.07 0.05 0.05 1.00 0.07 0.24 0.19 0.17 Wind Speed
-0.18 -0.16 -0.14 -0.12 -0.19 -0.15 0.39 -0.01 0.23 0.03 -0.02 0.07 1.00 0.08 0.10 0.14 Holidays
-0.42 -0.43 -0.41 -0.40 -0.39 -0.30 -0.34 -0.23 0.71 -0.14 0.22 0.24 0.08 1.00 0.18 0.28 Cloud Cover %
-0.50 -0.53 -0.54 -0.54 -0.44 -0.47 -0.43 -0.52 0.23 0.05 -0.15 0.19 0.10 0.18 1.00 0.60 Google Flu Trends
-0.73 -0.75 -0.75 -0.74 -0.69 -0.70 -0.59 -0.68 0.36 -0.10 -0.24 0.17 0.14 0.28 0.60 1.00 Soup Units
Temperature
Temperature^2
Temperature^3
Temperature^4
Log Temperature
Dewpoint Temp
Beer Units
Insect Spray Units
Humidity
DFN Temperature
Precipitation
Wind Speed
Holidays
Cloud Cover %
Google Flu Trends
Soup Units
1
0.5
0
-0.5


Seattle Correlation Heatmap
1.00 0.99 0.98 0.96 0.99 0.92 0.75 0.76 -0.52 0.34 -0.39 -0.41 -0.16 -0.53 -0.36 -0.64 Temperature
0.99 1.00 1.00 0.98 0.98 0.90 0.78 0.79 -0.54 0.32 -0.41 -0.42 -0.15 -0.56 -0.37 -0.64 Temperature^2
0.98 1.00 1.00 1.00 0.95 0.88 0.79 0.81 -0.55 0.30 -0.43 -0.42 -0.13 -0.59 -0.37 -0.64 Temperature^3
0.96 0.98 1.00 1.00 0.93 0.85 0.80 0.82 -0.56 0.30 -0.43 -0.42 -0.12 -0.62 -0.37 -0.63 Temperature^4
0.99 0.98 0.95 0.93 1.00 0.93 0.71 0.72 -0.49 0.37 -0.36 -0.39 -0.18 -0.47 -0.34 -0.63 Log Temperature
0.92 0.90 0.88 0.85 0.93 1.00 0.59 0.62 -0.14 0.35 -0.22 -0.44 -0.16 -0.27 -0.32 -0.48 Dewpoint Temp
0.75 0.78 0.79 0.80 0.71 0.59 1.00 0.82 -0.58 0.15 -0.41 -0.29 0.22 -0.58 -0.35 -0.56 Beer Units
0.76 0.79 0.81 0.82 0.72 0.62 0.82 1.00 -0.55 0.05 -0.46 -0.27 -0.06 -0.54 -0.45 -0.72 Insect Spray Units
-0.52 -0.54 -0.55 -0.56 -0.49 -0.14 -0.58 -0.55 1.00 -0.08 0.51 0.06 0.05 0.72 0.19 0.55 Humidity
0.34 0.32 0.30 0.30 0.37 0.35 0.15 0.05 -0.08 1.00 0.04 -0.22 0.06 -0.15 0.01 0.02 DFN Temperature
-0.39 -0.41 -0.43 -0.43 -0.36 -0.22 -0.41 -0.46 0.51 0.04 1.00 0.54 0.07 0.58 0.27 0.42 Precipitation
-0.41 -0.42 -0.42 -0.42 -0.39 -0.44 -0.29 -0.27 0.06 -0.22 0.54 1.00 0.08 0.33 0.23 0.15 Wind Speed
-0.16 -0.15 -0.13 -0.12 -0.18 -0.16 0.22 -0.06 0.05 0.06 0.07 0.08 1.00 0.05 0.01 0.14 Holidays
-0.53 -0.56 -0.59 -0.62 -0.47 -0.27 -0.58 -0.54 0.72 -0.15 0.58 0.33 0.05 1.00 0.16 0.38 Cloud Cover %
-0.36 -0.37 -0.37 -0.37 -0.34 -0.32 -0.35 -0.45 0.19 0.01 0.27 0.23 0.01 0.16 1.00 0.45 Google Flu Trends
-0.64 -0.64 -0.64 -0.63 -0.63 -0.48 -0.56 -0.72 0.55 0.02 0.42 0.15 0.14 0.38 0.45 1.00 Soup Units
Temperature
Temperature^2
Temperature^3
Temperature^4
Log Temperature
Dewpoint Temp
Beer Units
Insect Spray Units
Humidity
DFN Temperature
Precipitation
Wind Speed
Holidays
Cloud Cover %
Google Flu Trends
Soup Units
1
0.5
0
-0.5

Section 5. Feature Selection

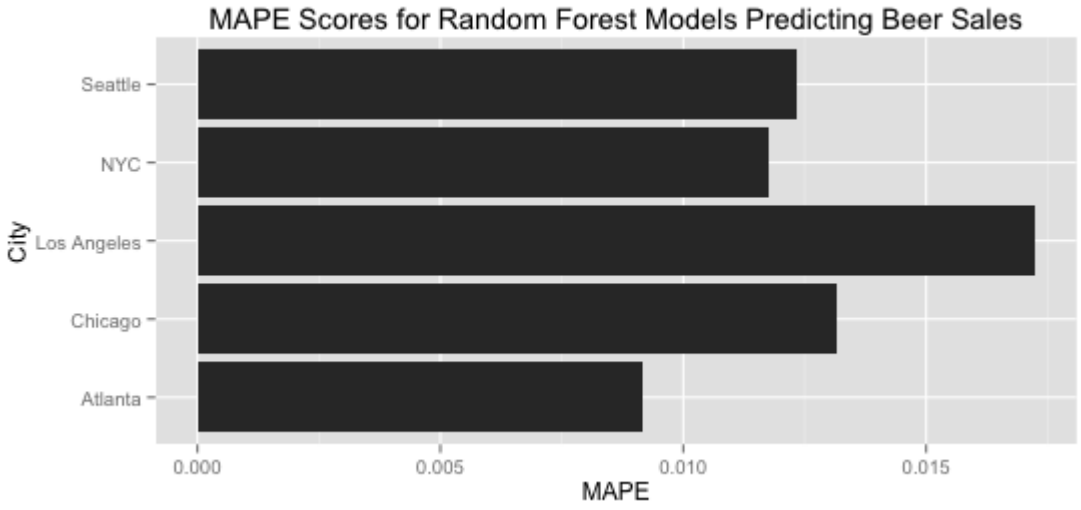
We then moved on to the model building phase and attempted to predict the number of units sold of our selected product categories. We were asked to find the optimal selection of features for each of the possible product-city combinations (soup in Chicago, soup in Los Angeles, etc.). The features we had to choose from were all of the weather data given to us by Weather.com as well as lagged units sold features (with lags of one through four weeks and fifty-two weeks) and any extra variables wed found online or made with our data (flu trends, holidays per week).

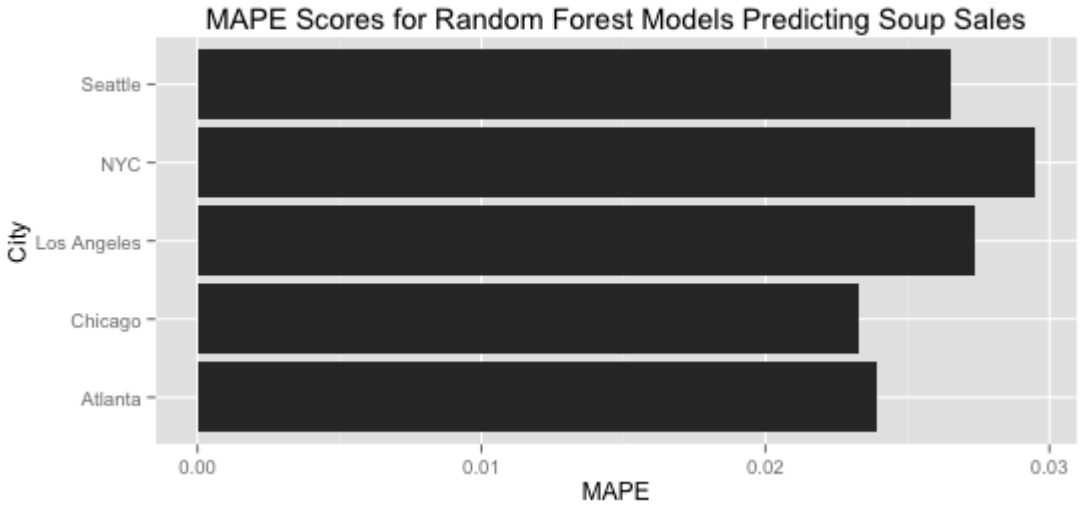
The data was split into training and testing sets of 69 and 36 recordings, respectively. We first took a product-city pair and created a random forest for it using all of the variables in the training data. The first two to four variables were saved after ordering the variables using mean increase in node impurity. We then recursively created random forests with the rest of the features, each round saving the variable that minimized the overall mean square error. The combination of mtry and starting variables that produced the lowest overall MAPE score on test data was selected as the best model. We found that selecting 2 to 3 starting variables and then having mtry set at around 13 to 16 produced the models with the lowest MAPE scores. These scores are displayed on the final page of this report. The scores for these optimal random forest models were then compared with a linear model containing all of the variables and a linear model containing only the selected variables. In all cases, the random forest model had lower MAPE scores.

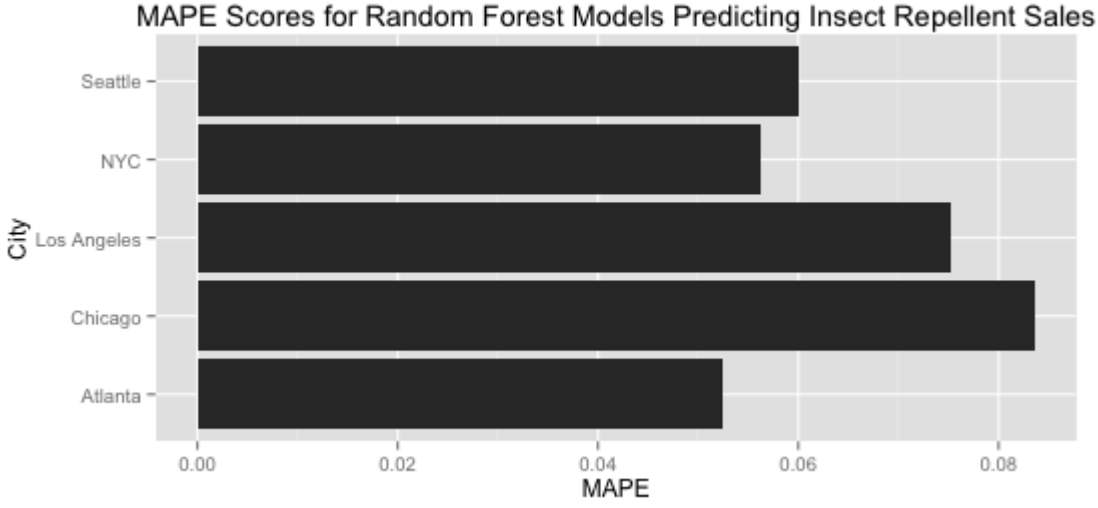
Conclusion and Recommendation

After clustering our data, we found the three main seasonal trends amongst our product categories. We then added variables to help predict these trends in our subset of cities. Judging by the MSE and MAPE scores of the models we produced, random forests with feature selection outperformed our baseline linear models. Were we to have more time, our team would be interested in experimenting with parameter tuning to determine the effects of individual changes in the weather variables, having all else held equal.

It's clear that weather can have a positive effect on predicting sales numbers. However, our team was limited by the fact that the data was weekly. Daily data would possbily illuminate more subtle changes in the weather having an effect on sales.


MAPE Scores for Random Forest Models Predicting Beer Sales
City
Seattle
NYC
Los Angeles
Chicago
Atlanta
0.000
0.005
0.010
0.015
MAPE


MAPE Scores for Random Forest Models Predicting Soup Sales
City
Seattle
NYC
Los Angeles
Chicago
Atlanta
0.00
0.01
0.02
0.03
MAPE


MAPE Scores for Random Forest Models Predicting Insect Repellent Sales
City
Seattle
NYC
Los Angeles
Chicago
Atlanta
0.00
0.02
0.04
0.06
0.08
MAPE