Ice Cream Shops of the Commonwealth

Matthew Owen

March 23, 2020

Introduction

Deciding whether to open a food service business can be a difficult decision, as success can depend on many factors that can be hard to assess ahead of time. One of those factors is the potential demand within the local community for the business. In-depth analysis can be used to assess this demand, but this can be time consuming and expensive. A prospective business owner who has many options of where to locate may need to narrow down these options before performing market analysis to make their final decision. A simple way to find good candidate communities for a new business is to determine which ones appear to be underserved for that type of business. Put simply, which communities currently have a fewer number of that business type than they "should" based on easy to determine demographic and economic factors?

Problem

A good candidate for this type of analysis is ice cream shops, since they are a popular yet niche food service business. Thus, a simple model can be trained to predict how many ice cream shops a given town should have and the output of this model can be compared with how many shops that town currently has. Towns with many more predicted shops than actual ones should be attractive candidates for entrepreneurs interested in opening a new shop. I will therefore construct a model that will predict which towns in the state of Massachusetts are the best candidates for a new ice cream shop.

Data

Data Sources

My model of ice cream shop demand will be built using general demographic data for individual towns. I will use data from the state of Connecticut to train the model, because it is a state of roughly equivalent size that is in the same region as Massachusetts. Town by town demographic data for Connecticut is publicly available on a state website in the form of tables that can be downloaded as csv files. From these I will pull out the following information about each town: the total population, the median resident age, the per capita income, the average household size, and the population density. For the model to make a prediction, I will need the equivalent demographic data for the state of Massachusetts. These can be found on the website for Boston.com in tabular form that can be gathered using HTML parsing. The Massachusetts data is not as up to date as the Connecticut data but is less than 10 years old and so should be recent enough for this application. I expect that population size will be the strongest predictor of the number of ice cream shops and a quick look at the data indicates that the town sizes in Connecticut range from 873 to around 146,000, while in Massachusetts they range from 76 to 625,000. The other component needed to build my model will be information on the current number of ice cream shops in each town in Massachusetts and Connecticut. This information will be acquired using the Foursquare API. I can use the search endpoint to look for venues "near" to each town in each state that have a category designation of 'Ice Cream Shop'.

Data Cleaning

The demographic data from Connecticut was organized to include categories such as age groups and ethnic background that are not relevant to this study, so I sliced the dataframes to only give one value per town for each demographic factor. These values were then combined into a single dataframe. Likewise, the Massachusetts data had some extraneous categories that were excluded when combining the demographic data into a single dataframe. One of the Massachusetts tables contained an extra row that repeated the column headers, so this row was dropped. I determined that towns of very small size would be highly unlikely to be able to support a niche business such as an ice cream shop, so all towns with a population less than 1500 were dropped from the analysis.

The data on ice cream shop venues from Foursquare contained many duplicates since a search for each town could return venues from neighboring towns. Thus, these duplicate entries were dropped from the dataframe. The primary category designation for each venue had to be retrieved from the more complex formatting returned by Foursquare. These categories, the name of each venue and its town were used to populate a new dataframe. Despite searching specifically for the 'Ice Cream Shop' category, Foursquare returned venues with a variety of category designations. For this study, I chose to keep only those with a designation of 'Ice Cream Shop' and 'Frozen Yogurt Shop'. Finally, some of the town designations returned by Foursquare did not match the official town names, either because they were misspelled or because they referred to a more specific location within a town. I constructed a dictionary that paired each of these designations with the correct town designation for them and then used this to correct the values in the dataframe.

Methodology

I examined the distributions of the different demographic parameters for each state, which were generally very similar. As can be seen from the Massachusetts data below (Fig 1, top row), the distributions for median age and average household size appear to approximate normal distributions, however the other 3 variables had distributions skewed heavily to low values. Thus, I performed a natural log transformation on each of the parameters and observed the new distribution (Fig 1, bottom row). The distributions for the total population, per capita income, and population density all appear to be much more normally distributed following this transformation.

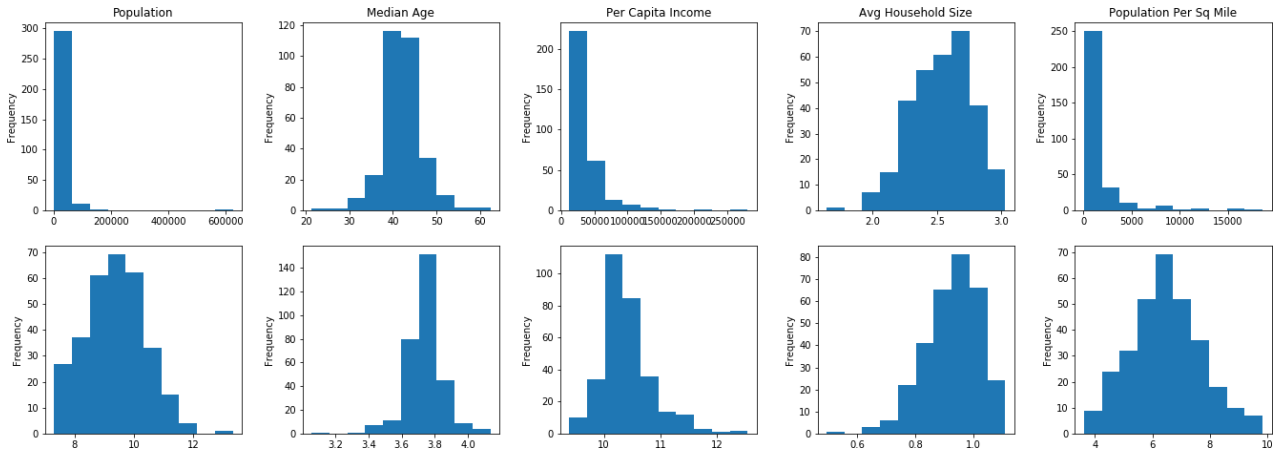

Figure 1: Top row - histograms display the distribution of Massachusetts towns for each demographic variable. Bottom row – histograms display the log transformed distributions.

I next examined the relationship between each independent variable and the number of ice cream shops. I used the log transformations of the data for total population, per capita income, and population density. For the Connecticut data (Fig 2), there appear to be reasonable positive correlations between population and population density and the number of shops, whereas there is a negative correlation between median age and the number of shops. Per capita

income and average household size appear to be weakly or not at all correlated with the number of ice cream shops.

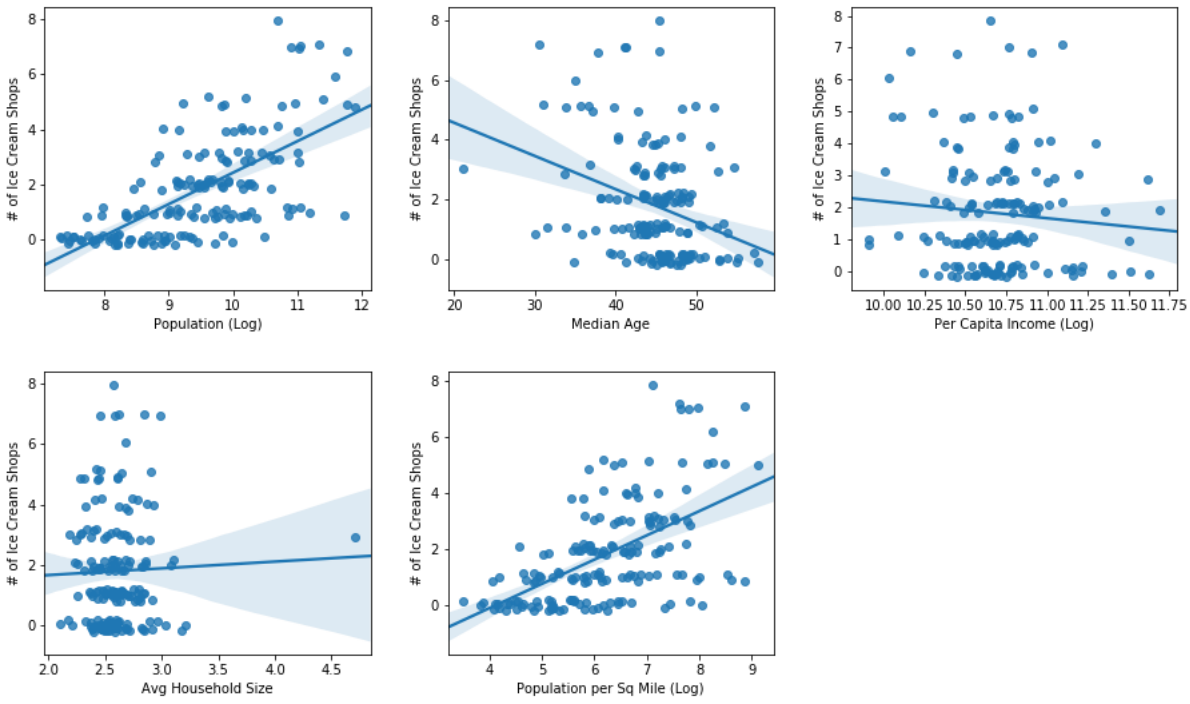

Figure 2: Scatter plots for the Connecticut town data of each demographic variable (or log transformation) plotted against the number of ice cream shops.

Looking at the distributions of the Massachusetts data (Fig 3), the positive correlations of ice cream shop numbers with total population and population density are similar to the Connecticut data, as is the lack of correlation with per capita income. However, whereas the Connecticut data showed a negative correlation between median age and ice cream shops, the Massachusetts data shows virtually no correlation between these variables. Likewise, where Connecticut showed no correlation between average household size and the number of shops, the Massachusetts data shows a negative correlation. The disagreement between the datasets on

these points suggests that the Connecticut dataset is not an ideal training set to predict outcomes in Massachusetts.

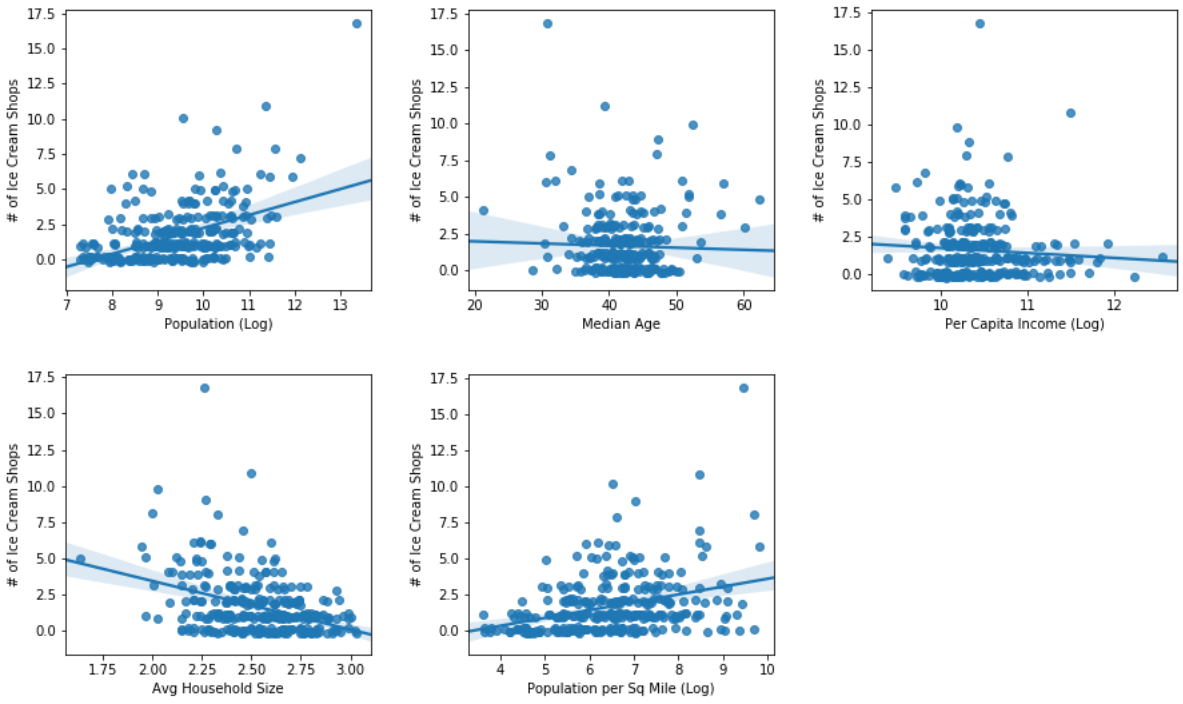

Figure 3: Scatter plots for the Massachusetts town data of each demographic variable (or log transformation) plotted against the number of ice cream shops.

Given that most of the independent variables in these datasets appear to have a linear relationship with the dependent variable, I decided to build my model using multiple linear regression. This approach not only allows for a model that can predict the outcomes required for the main goal of the analysis, but also allows me to determine the individual contribution of each of the independent variables to the outcome. For this regression analysis, I am using the log transformed versions of the total population, per capita income, and population density. Further, I am also standardizing all of the input variables by converting them to z-scores before fitting the model.

Results

The multiple linear regression model returned a best fit equation of:

$y = 1.8 + 1.74x_1 + 0.08x_2 + 0.04x_3 – 0.19x_4 – 0.48x_5$

where x_1 is the total population, x_2 is the median age, x_3 is the per capita income, x_4 is the average household size, and x_5 is the population density. The R-squared value was 0.46, suggesting a relatively weak fit with the data. The equation suggests that total population has the largest contribution to the number of ice cream shops in a town, specifically that a doubling of a town's population would lead to an average of an additional 1.67 ice cream shops. Surprisingly, even though population density on its own positively correlates with the number of shops, once the other parameters are factored in the model predicts that an increase in density would lead to a decrease in the number of ice cream shops.

After using the model to predict the expected number of ice cream shops in each Massachusetts town, I found that there was at least a positive correlation between the number of predicted shops and the actual number (Fig 4). Taking the predicted number of shops (negative predicted values were changed to zero) and subtracting the actual number gives a value that should correspond to how underserved a given town is in terms of its number of ice cream shops. I sorted the Massachusetts towns in descending order for this value (Fig 5) to find the most underserved towns. At the top of the list, with 3.9 fewer ice cream shops than expected by the model was Lynn, Massachusetts, a town with a population of 91,000 people and zero ice cream shops.

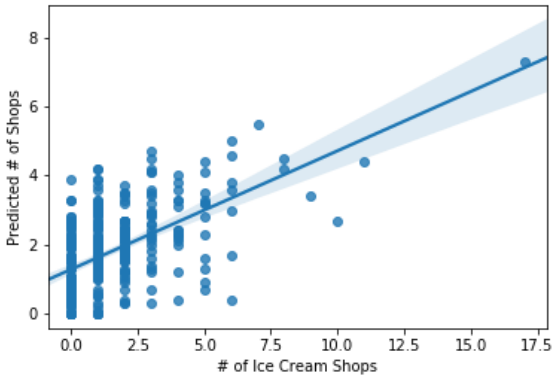

Figure 4: Scatter plot and fit line for the number of ice cream shops in each Massachusetts town versus the number predicted by the model.

Town	# of Ice Cream Shops	Predicted # of Shops	Predicted - Actual
LYNN	0	3.9	3.9
MALDEN	0	3.3	3.3
ANDOVER	0	3.3	3.3
FRAMINGHAM	1	4.2	3.2
BROCKTON	1	4.2	3.2
WALTHAM	1	3.9	2.9
WEYMOUTH	1	3.9	2.9
STOUGHTON	0	2.8	2.8
WATERTOWN	0	2.8	2.8
EASTON	0	2.7	2.7

Figure 5: The top 10 Massachusetts towns that are underserved for the number of ice cream shops.

Discussion

The model constructed for this analysis was a simple one and one that was hampered by the fact that several of the chosen demographic variables (especially per capita income) did not have much predictive power with regard to the outcome of interest. However, the model output did identify many towns that seem to have greatly fewer ice cream shops than would be expected. This model was not designed to be the final determinant of the best location to open a

new ice cream shop. Instead, the intent was to pare down the number of potential options so that more in-depth market research could be done. In this regard, I believe the model was successful, as the top candidates noted in Figure 5 seem to have strong potential.

In examining the rankings of Massachusetts communities underserved for ice cream shops it was notable that many of the most overserved towns were beach towns on Cape Cod or Martha's Vineyard that have heavy seasonal tourism is the summer. It is not surprising that these towns would have more ice cream shops than others, all else being equal. However, I decided to examine further to see whether this phenomenon might have impacted the analysis.

I designated as beach towns all towns on Cape Cod, Martha's Vineyard, and Nantucket that were not otherwise excluded from the analysis. Looking at the relative distributions of the beach towns (Fig 6, orange dots), it is notable that in addition to having a disproportionately large number of ice cream shops, they also have a disproportionately high median age and a disproportionately low average household size. Thus, it is plausible that these towns may be skewing the analysis somewhat.

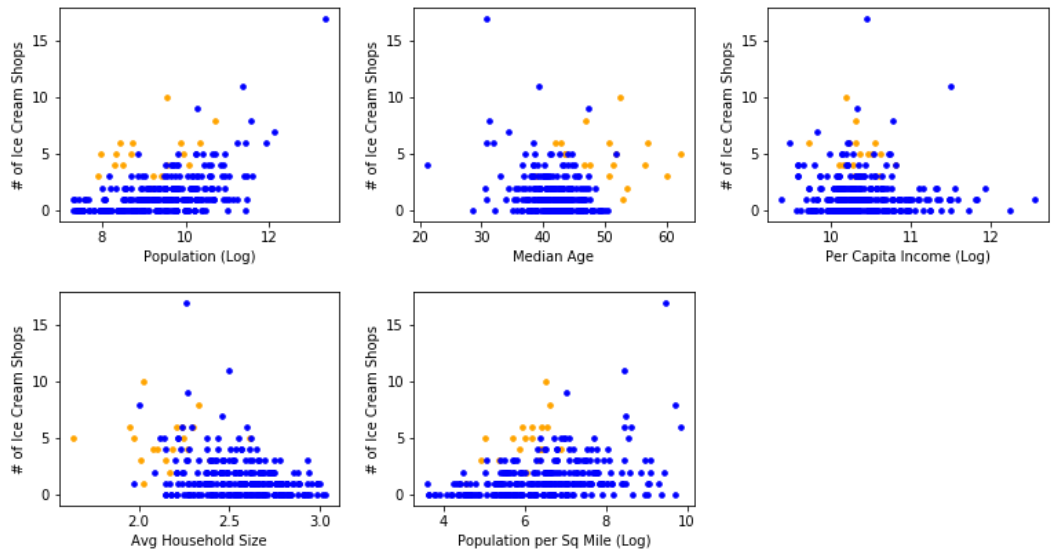

Figure 6: Scatter of each demographic variable vs the number of ice cream shops for each Massachusetts town with beach towns highlighted in orange.

To test what the effect of the Massachusetts beach towns may be, I re-ran the analysis with these towns excluded. As expected, this resulted in a significant change in the relationship between the number of ice cream shops and the average household size, and an even greater change between the number of shops and the median age. Excluding the beach towns shifts the fit line of this relationship from a virtually flat one to one that is almost the same as the one for the Connecticut data (Fig 7), meaning that the model based on the Connecticut data should be better at predicting the Massachusetts data when the beach towns are excluded. This is verified by comparing the actual number of ice cream shops with the number predicted by the model. When the beach towns are excluded, the value of R-squared for this correlation increases from 0.30 to 0.33 (Fig 8).

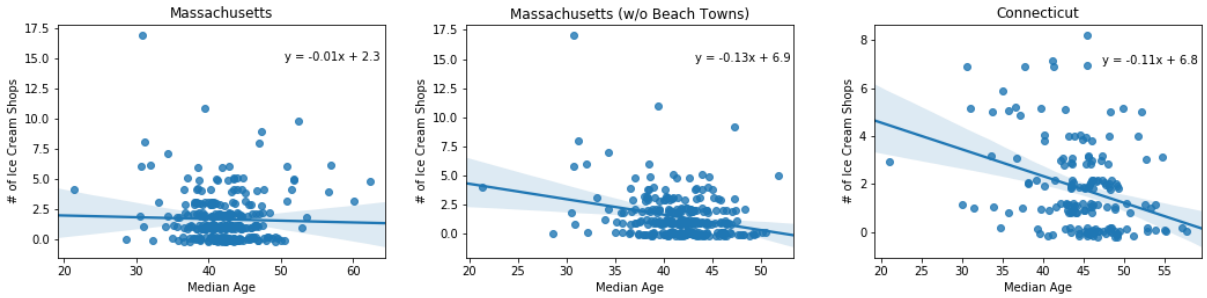

Figure 7: Scatter plots and fit lines of median age vs the number of ice cream shops for the indicated groups of towns.

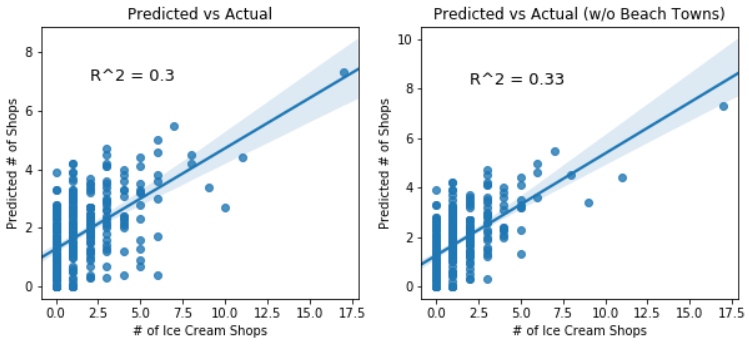

Figure 8: Scatter plots and fit lines relating the predicted number of ice cream shops and the actual number for all Massachusetts towns (left) and Massachusetts towns excluding beach towns.

Conclusion

In this analysis, I used multiple linear regression and Connecticut demographic data to build a model that predicted the number of ice cream shops in each town in Massachusetts. The model showed that the total population of a town was the biggest determinant of the number of shops it has. The output of the model did a decent job of predicting the number of Massachusetts shops, with an R-squared of 0.30. By comparing the predicted number of shops with the actual number, I was able to identify which communities were the most potentially underserved in terms of ice cream shops, with Lynn, Massachusetts being the top candidate. While the model did a competent job at identifying underserved towns, a final determination of the best place to open a new shop would require more detailed analysis of more complex data.