Ice Cream: weather and wealth

Final project for IBM Data Science course

Aug-2020

Contents

Introduction

The question examined in this report is: What is the most promising location[1] (as defined by zip code) to open a new ice cream shop?

This question is of potential interest to various parties. Entrepreneurs in casual dining may be searching for fruitful locations. Investors and lenders to businesses may want to evaluate the potential of different geographies to support an ice cream business. Established chains of ice cream shops (e.g. Baskin-Robbins, Ben & Jerry's) may be considering expansion.

For purposes of the following analysis, certain geographic constraints have been imposed. These are plausible in that a genuine business-expansion decision is likely to be constrained, however the motivation in this case was simply to manage the amount of data extracted for the sources (see below), particularly in order to stay within the limits imposed by data providers on their non-paying users.

Based on these data and the subsequent analysis, the study finds that the most promising zip code among those considered for a new ice cream shop is 10007.

Data and Methodology

Conceptualizing the Approach

The analysis proceeds by identifying potential factors influencing the demand for ice cream stores ("environmental factors"), evaluating whether in fact those environmental factors are useful predictors (by reference to existing ice cream stores), and then finding locations in which the actual number of ice cream stores is fewer than that predicted by the best model of how locations with particular environmental factors support ice cream shops.

Specific Data Needed

The data needed can be conceptualized as falling into either the category of primary data or secondary data. Data called *primary data* are data directly relevant to modelling environmental factors and to identifying current ice cream shops. Data called *secondary data* are sourced because they are needed in order to extract primary data from a data source.

For example, zip codes (more precisely, US Census ZCTAs) are the basic geographical unit used in the analysis, however certain data is coded by city, and so it is necessary to find the city associated with a zip code in order to retrieve weather data, an environmental factor used in the analysis. In this case, the zip code is considered primary and the city is considered secondary data.

Data required includes: Zip codes for the relevant geographies, weather, population, household income, area (that is, the physical area of the relevant geography, useful for considerations of density and proximity), and existing ice cream shops.

[1] For purposes of this study, only the cities of Miami, Houston, Chicago, and New York are considered. The rationale for this is discussed herein.

Finally, keep in mind that a constraint on the analysis is that only locations in Chicago, Houston, New York City, and Miami are considered.

Data Sources and Associated Challenges

This data was sourced from freely available internet resources.

I. Weather data
   a. Weather data for each of the four relevant cities are found at USClimateData.com
   b. This was a complicated data sourcing problem. It appears that the NOAA website offers APIs, but the process is complicated and time consuming to learn. DarkSky shut off their free data very recently, coincident with their acquisition by Apple Inc.

II. Zip code data
   a. Zip codes comprised by each of the four relevant cities are found using an API call to zip-codes.com.

III. Longitude and latitude data
   a. The PyPi package USZipCode was used to extract lon/lat coordinates for each zip code.

IV. Population data
   a. Population for each zip code was retrieved from the US Census department's ACS survey data by API call.

V. Physical area data
   a. The area of each zip code was retrieved from WolframAlpha by API call.

VI. Household income data
   a. Household income for each zip code was retrieved from the US Census department's ACS survey data by API call.

VII. Ice cream shop data
   a. Foursquare venue calls were used to identify existing ice cream shops

The principle organization for the data was by zip code. For each zip code found to be associated with one of the four cities studied, that zip code was associated with weather, population, etc. data. The data was cleaned: any zip code that did not have an associated area was discarded[2]. Any zip code that did not have a population was discarded[3].

Analysis

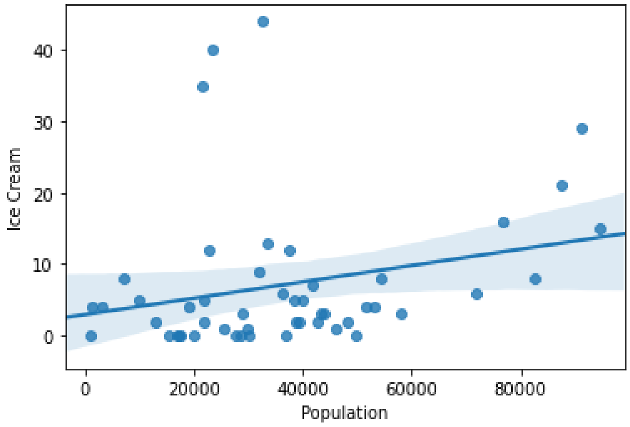
Data Exploration

Viewed independently, the environmental variable identified are (at best) weakly predicative of the presence of ice cream shops in a particular location. While intuition might suggest that locations with higher-income households would have more ice cream shops and might suggest that warmer locations would have more ice cream shops, these conclusions are not supported by the sample data.

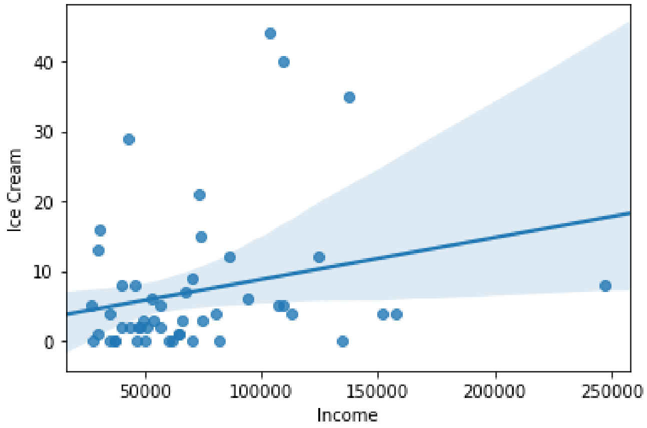
[2] Spot checks of these zip codes suggest that they are post office box locations.

[3] Spot checks of these zip codes suggest that they are post office box locations.

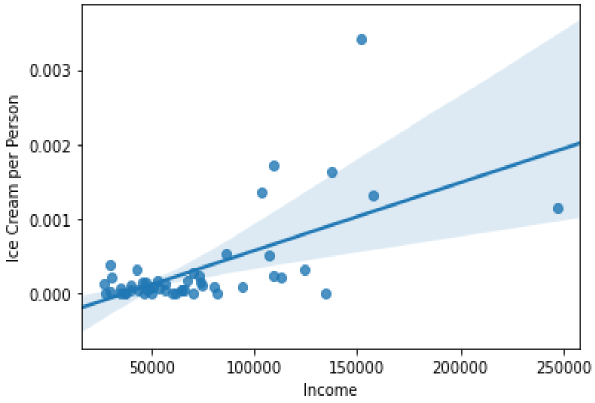
Population is also not a particularly good indicator of the presence of ice cream shops in a location, as shown below.



Income is a weak explanatory factor of the presence of ice cream shops, as shown below.

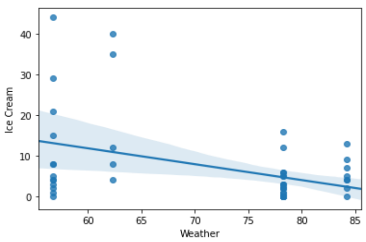


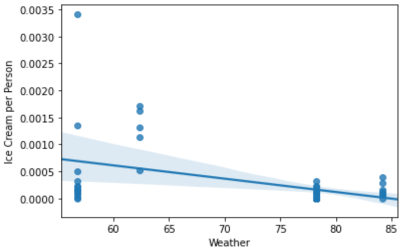
Ice cream shops per capita has a stronger relationship with income, as shown below.

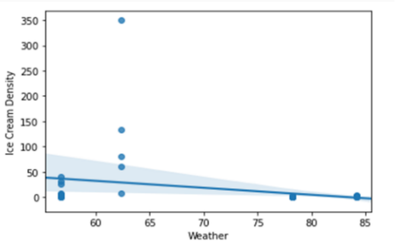


We may understand the difference between these last two graphs as follows. High income doesn't result in demand for ice cream by itself because it isn't dollars of income that buy ice cream. Rather, people buy ice cream, and people with more money are served by more ice cream shops.

Counterintuitively, weather (average high temperature) is (weakly) inversely correlated with ice cream shop presence. This is true for total ice cream ships in a location, for ice cream shops per person and for ice cream shops per square mile. This can be seen in the graphs below.







Modelling

Pairwise relationships between particular environmental factors and ice cream shop presence are weak, as we saw in the prior section. However, by combining available environmental factors, we may develop a better model for the relationship.

In this case, we choose to model the relationship as a multiple regression on the environmental factors, namely weather (average high temperature), location (lat/lon), city, population, household income, and area. From these independent variables, we fit a multiple regression to the observed ice cream shops.

Note, were we to ask a slightly different research question, which neighborhoods are likely to have many ice cream shops, we could use a tool like K-Nearest Neighbors.

The model selected and fit is coded as shown in Appendix One

Limitations and Further Study

There are several limitations to this study; the most significant are listed below.

The regression model we have chosen explains only 44% of the variation in ice cream shop presence. No examination of residuals was conducted to assess bias. No other models were evaluated. The data concerning existing ice cream shops was limited by the number of Foursquare returns available to non-paying users – thus certain areas may have more ice cream shops that the data indicate.

A significant conceptual limitation of this approach is that we assume that the presence of ice cream shops tells us something about ice cream shop business success. While this isn't an outlandish assumption, it could certainly be the case that existing ice cream shops are failing (in a non-uniform way across the sample).

Numerous other potentially relevant factors have been ignored, including the presence of businesses that compete with ice cream shops (e.g candy stores) and businesses that are not coded as ice cream shops in Foursquare but which sell ice cream.

Conclusion

The zip code for which the current number of ice cream shops is most below the model's predicted number of ice cream shops is 10007, in New York City. This zip code has a only 8 ice cream shops, but based on environmental factors[4], our model suggests it should have 19 shops.

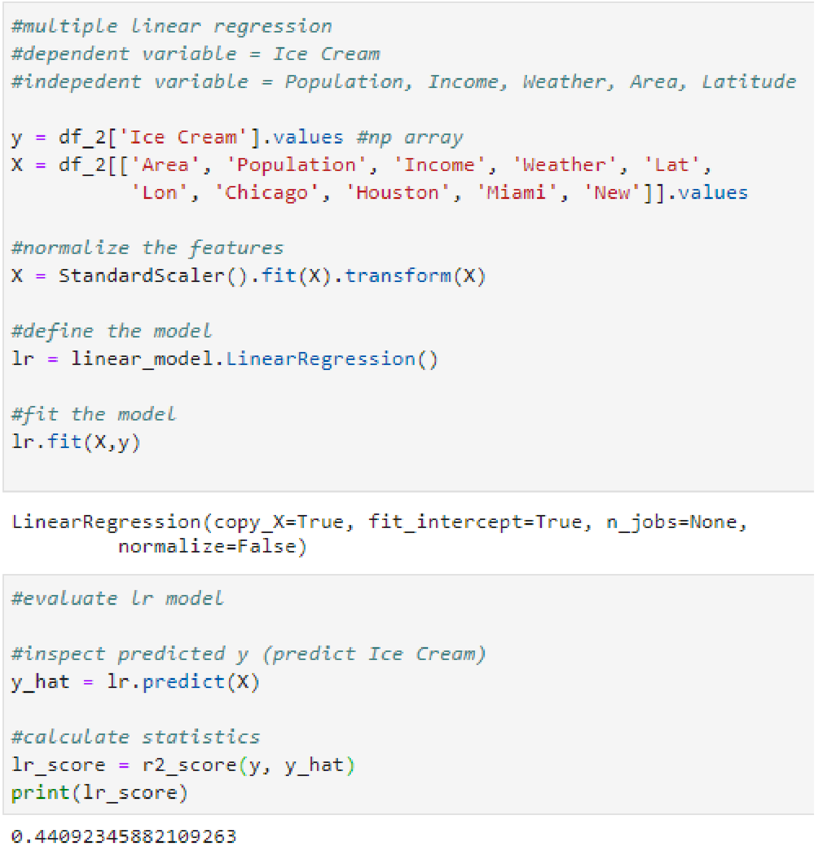
Best Opportunity is: 10007
Hoboken
New York
Manhattan Bridge
East River
Brooklyn Navy Yard
Brooklyn
Holland Tunnel
West Street
5th Avenue
2nd Avenue
1st Avenue
Avenue A
Avenue B
FDR Drive
Williamsburg Bridge
Brooklyn Bridge
Wythe Avenue
Park Avenue
Furman Street

The best location in Chicago, Houston, Miami, or New York City for a new ice cream shop is 10007.

[4] Population=7023, income=$246,813, temperature=62.3 degrees, area = 0.1 sq. mile, lat = 40.71 deg, lon=-74.01 deg.

Appendix One

Python coding of model fit:



```
#multiple linear regression
#dependent variable = Ice Cream
#indepedent variable = Population, Income, Weather, Area, Latitude

y = df_2['Ice Cream'].values #np array
X = df_2[['Area', 'Population', 'Income', 'Weather', 'Lat',
          'Lon', 'Chicago', 'Houston', 'Miami', 'New']].values

#normalize the features
X = StandardScaler().fit(X).transform(X)

#define the model
lr = linear_model.LinearRegression()

#fit the model
lr.fit(X,y)
```

```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None,
         normalize=False)
```

```
#evaluate lr model

#inspect predicted y (predict Ice Cream)
y_hat = lr.predict(X)

#calculate statistics
lr_score = r2_score(y, y_hat)
print(lr_score)
```

```
0.44092345882109263
```