Ice Cream

Which location offers the best opportunity for a new ice cream shop

Final project for IBM Data Science course

Aug-2020

Introduction – Based on Rollins' Methodology

Problem

Business Understanding

How Will Data Help

The question presented is: which location presents the best opportunity for a new ice cream shop? This question would be of interest to a range of parties, including those considering opening an ice cream shop (e.g. Unilever, an avid home ice cream maker), those considering investing in or lending to a new ice cream business, and those

The process of 'business understanding' adds the constraint that only the cities of Houston, Miami, New York, and Chicago are in consideration. Further, finding the best **zip code** is the goal.

Data provides insight into the current state of the ice cream market – specifically competitors and customers.

Introduction (2/3)

What Data is Needed

The problem requires data concerning: the number of ice cream stores now, their locations and facts about those locations, and the demographics of the communities in which they are located.

Data Sources

That data can be sourced publicly:

Information Needed	Data Source	Method
Zip codes	zip-codes.com	Search by city, extract zip codes
Ice cream shops	Foursquare	Venue query on ice cream, extract zip code
Weather	USClimateData.com	Request by city (note: DarkSky cut free api after Apple's purchase, NOAA was difficult to use)
Income, Population	US Census/ACS	Zip code search, extract median household income, population
Geo data	WolframAlpha	Zip code search, extract area
Geo data	USZipCode from PyPi	Need to associate zip code with longitude and latitude

Introduction (3/3)

Is Data Representative

The data are from credible sources, collected systematically, and available for all the points under consideration.

Manipulation

That data will need to be formatted – the disparate sources suggest this could be a challenge.

Deriving the Answer

Further, visual/graphical tools will be used to examine both physical location and ice cream shop density in the metric space defined by the various factors that may provide insight into the best business opportunity.

Further, a regression model will suggest a relationship among the independent variables (generally speaking, the market) and the dependent variable (the number of ice cream shops).
By identifying areas that are under-served (as indicated by the model) by current ice cream shops, we can identify the location of greatest opportunity for a new ice cream shop.

In the hypothetical context of this problem, once the best location is identified, the business problem is resolved (i.e., the proposed solution can be implemented)

Current Competitors – Where are the ice cream shops now
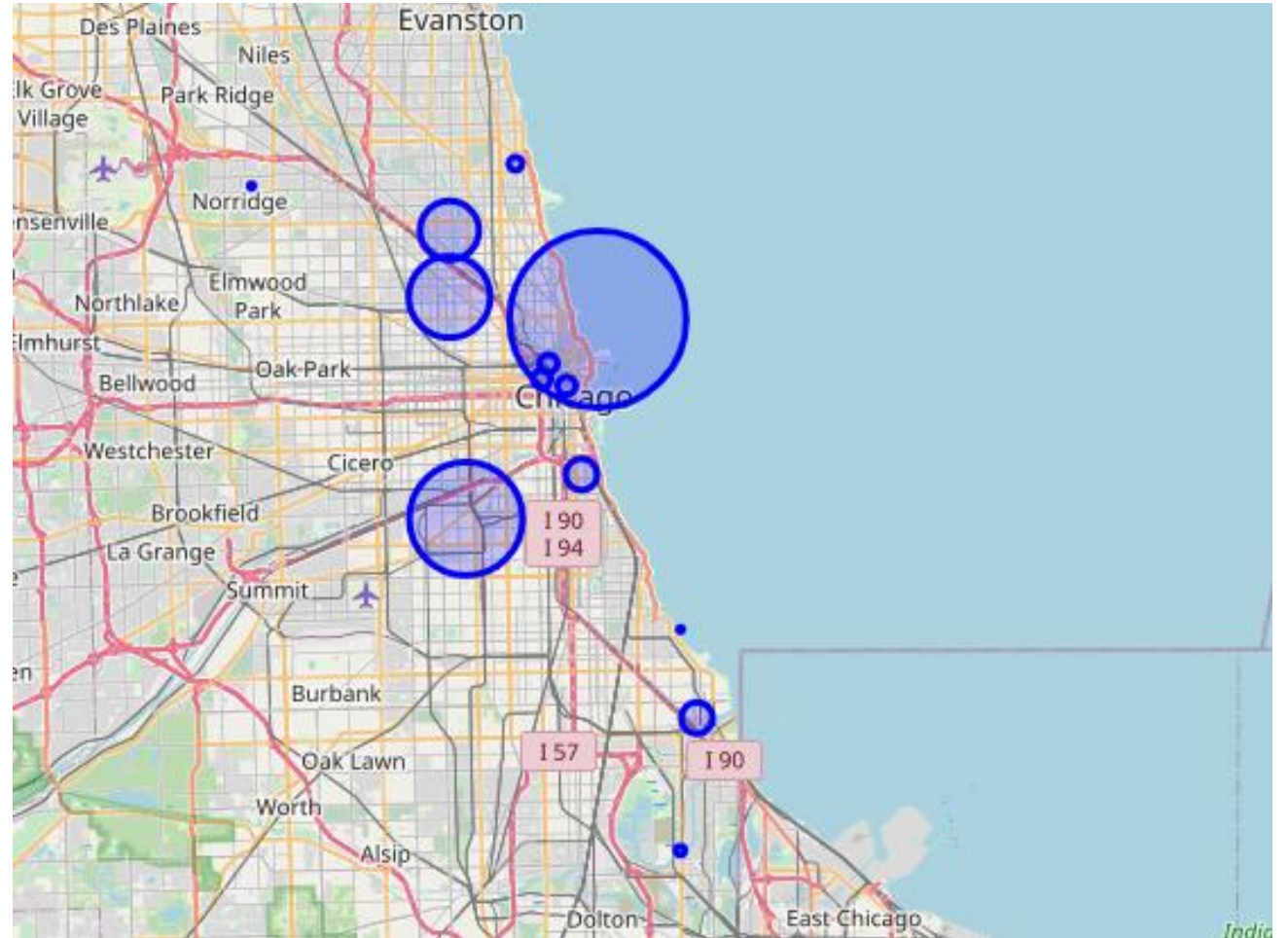
Chicago

Where are current ice cream stores?

The Folium map provides popups with details, and allows exploration of the area.

Zip codes comprised by the city are identified by a `requests.get` call to **zipcode.org**.

There are 85 zip codes in the city of Chicago.

Those zip codes are used to for a venue query of Foursquare.com.

The latitude and longitude of each zip code is found using the Python library `uszipcode`.

Houston

Where are current ice cream stores?

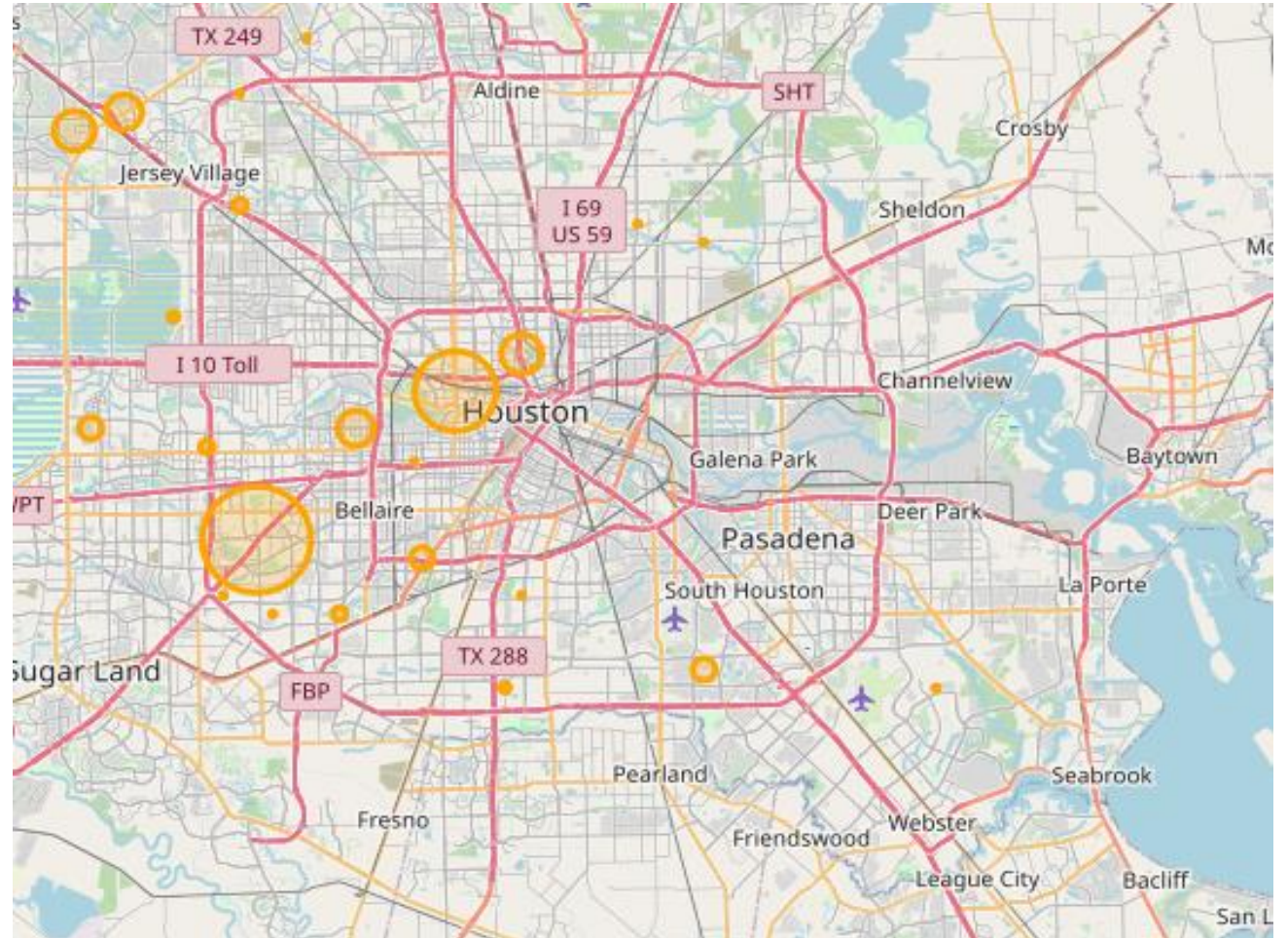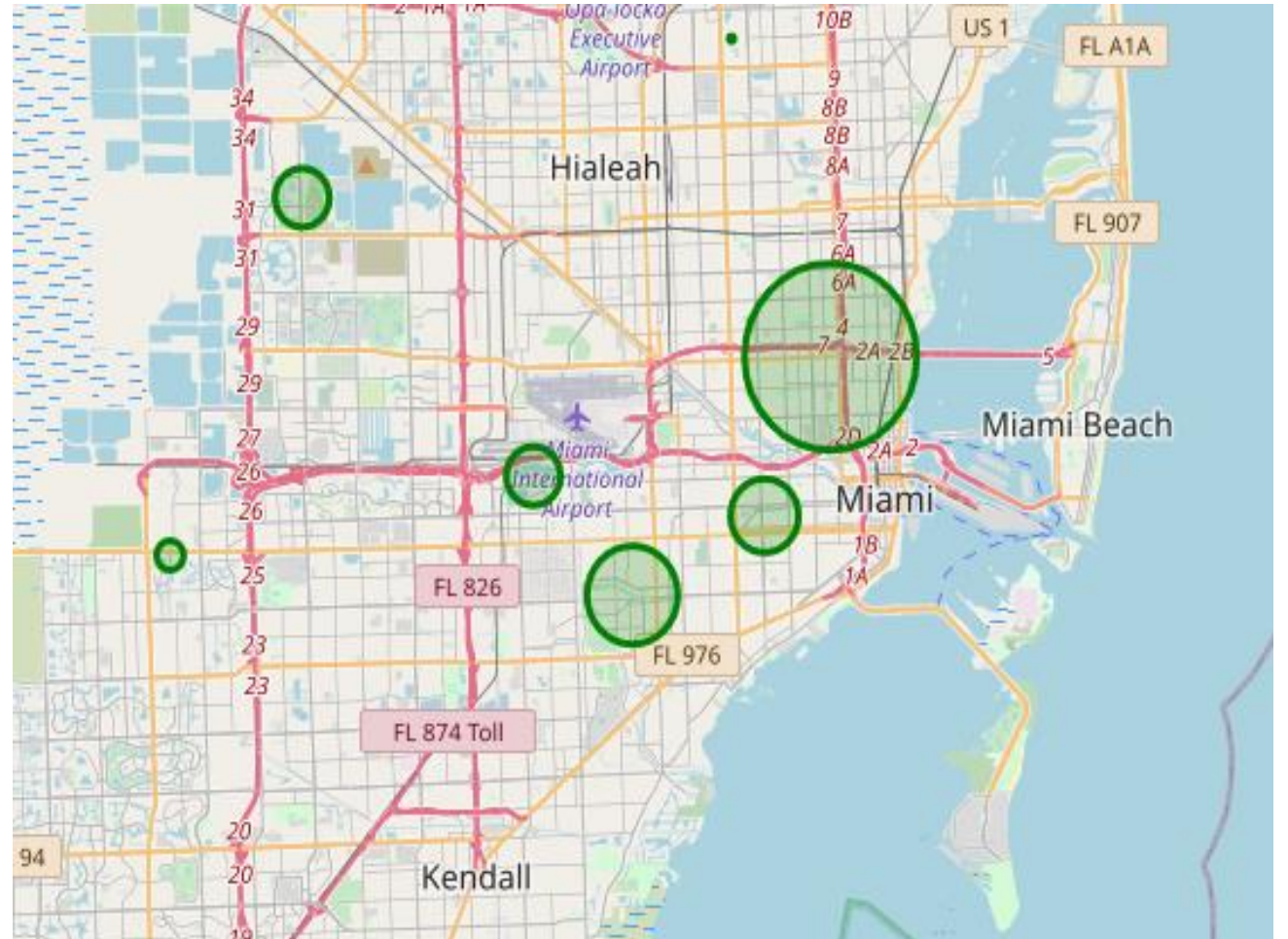The Folium map provides popups with details, and allows exploration of the area.

Zip codes comprised by the city are identified by a `requests.get` call to **zipcode.org**.

There are 177 zip codes in the city of Houston.

Those zip codes are used to for a venue query of Foursquare.com.

The latitude and longitude of each zip code is found using the Python library `uszipcode`.

.

Miami

Where are current ice cream stores?

The Folium map provides popups with details, and allows exploration of the area.

Zip codes comprised by the city are identified by a `requests.get` call to **zipcode.org**.

There are 92 zip codes in the city of Miami.

Those zip codes are used to for a venue query of Foursquare.com.

The latitude and longitude of each zip code is found using the Python library `uszipcode`.

New York

Where are current ice cream stores?

The Folium map provides popups with details, and allows exploration of the area.

Zip codes comprised by the city are identified by a `requests.get` call to **zipcode.org**.

There are 150 zip codes in the city of New York.

Those zip codes are used to for a venue query of Foursquare.com.

The latitude and longitude of each zip code is found using the Python library `uszipcode`.
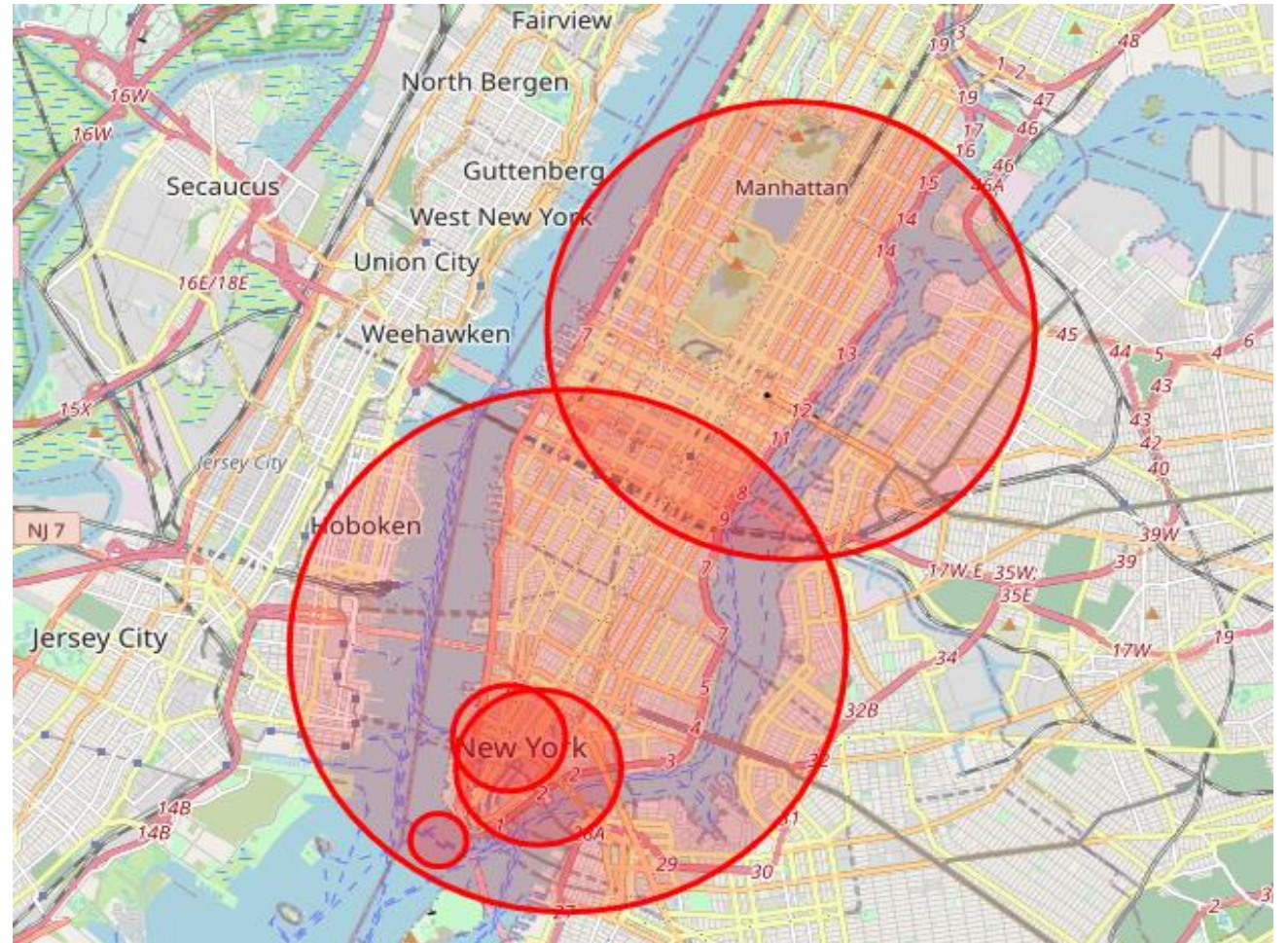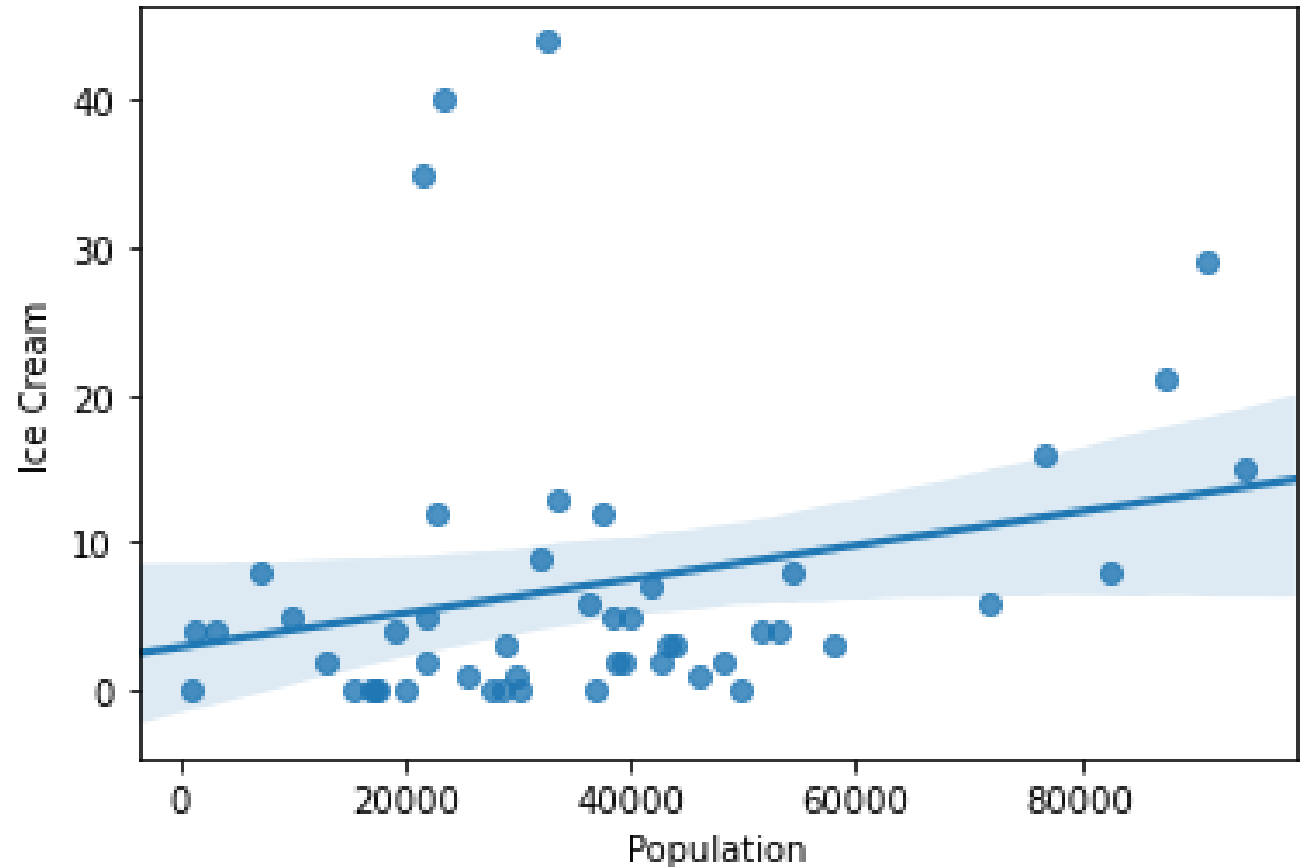
What Drives Ice Cream Location

Single factor analysis

Population

Are the number of people resident in a zip code a strong predictor of the number of ice cream shops in that zip code?

The Seaborn `regplot` demonstrates that it is not.

Correlation given by `.corr()` is only 0.26. This is a counterintuitive result.
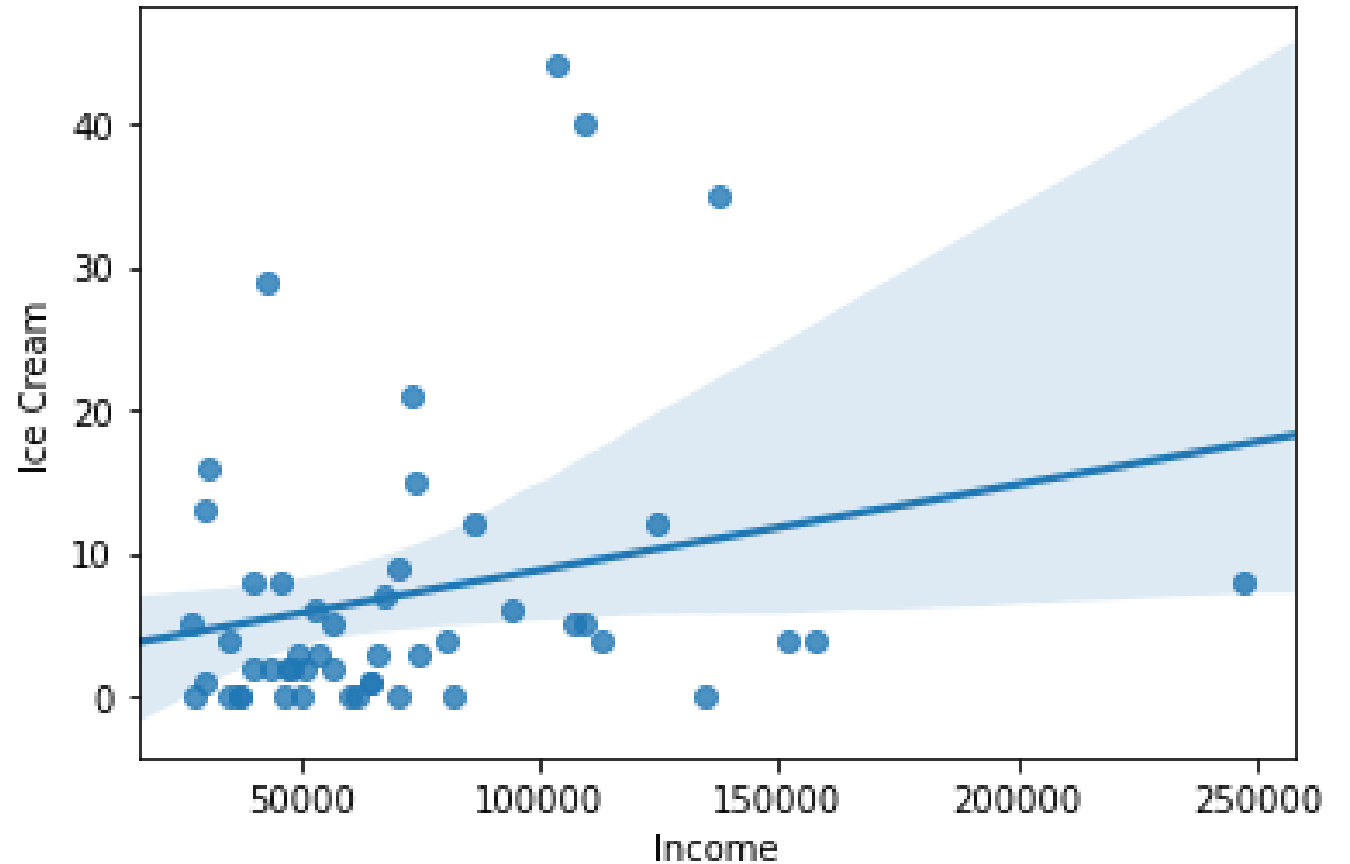
Income

Is local (to the zip code) household income a strong predictor of the number of ice cream shops in the zip code?

The Seaborn **regplot** demonstrates that it is not.

Correlation given by **.corr()** is only 0.24. This is not particularly surprising, as people in many income brackets consume ice cream.

Income and Ice Cream per Capita

Is local (to the zip code) household income a strong predictor of the number of ice cream shops **per person** in the zip code?

The Seaborn `regplot` demonstrates that it is to some degree.

Correlation given by `.corr()` is 0.62. While income is a weak predictor of the number of ice cream shops in a location, it is a stronger predictor of ice cream shops per person in a location.
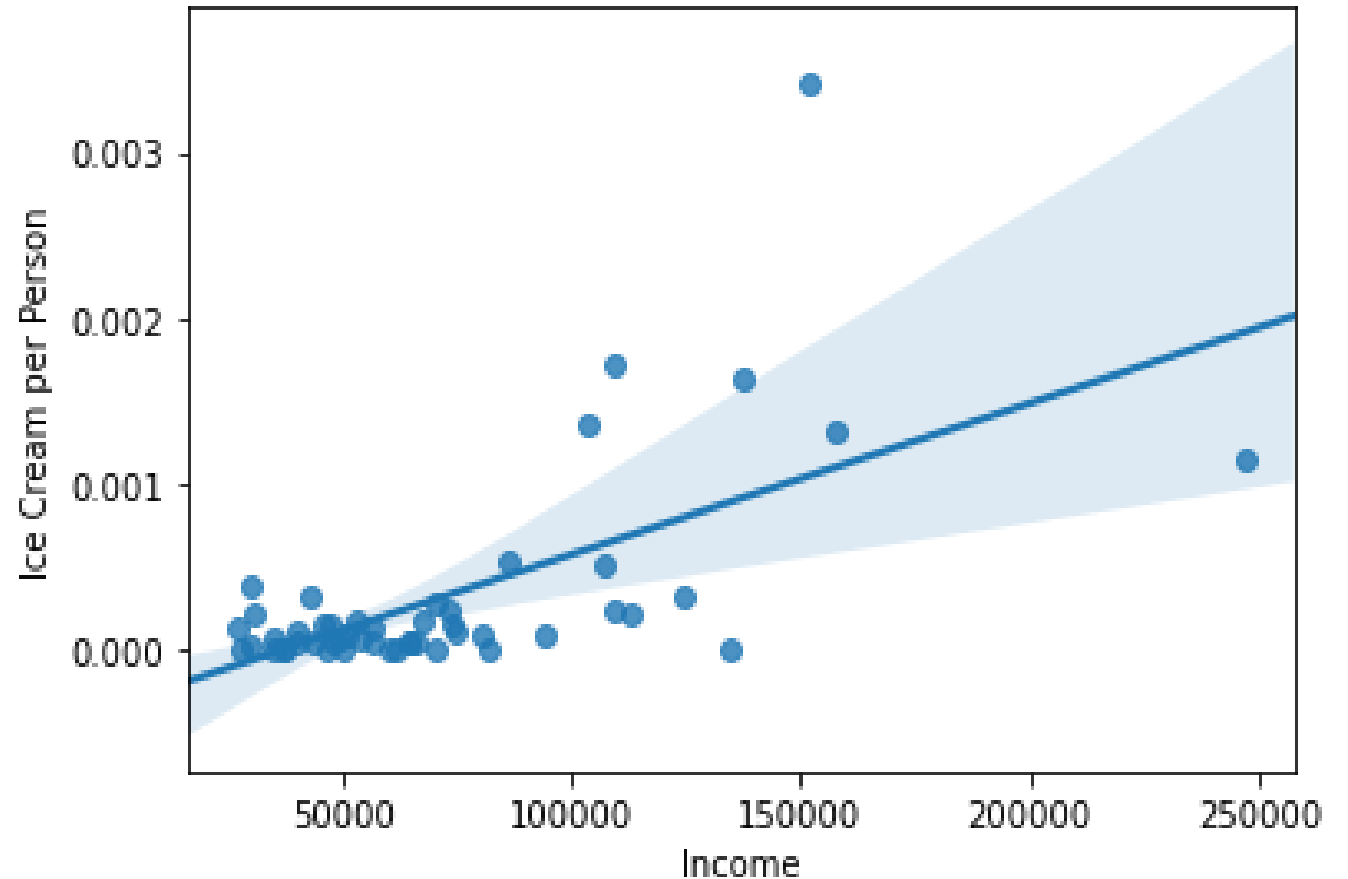
Temperature

Is local (to the zip code) average high temperature a strong predictor of ice cream shops?

These Seaborn `regplot` demonstrate that it is not, regardless of whether one considers the simple number of ice cream shops in the zip code, or the ice cream shops per capita, or the ice cream shops per square mile.

This is a counterintuitive result – one would expect there to be more demand for ice cream in warm locations.
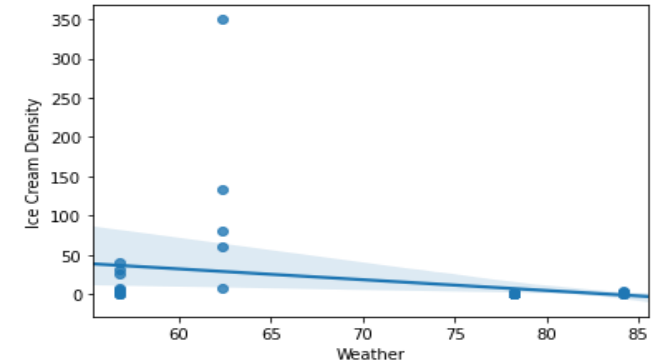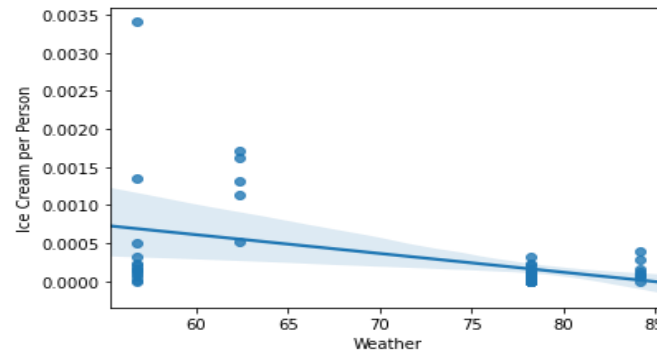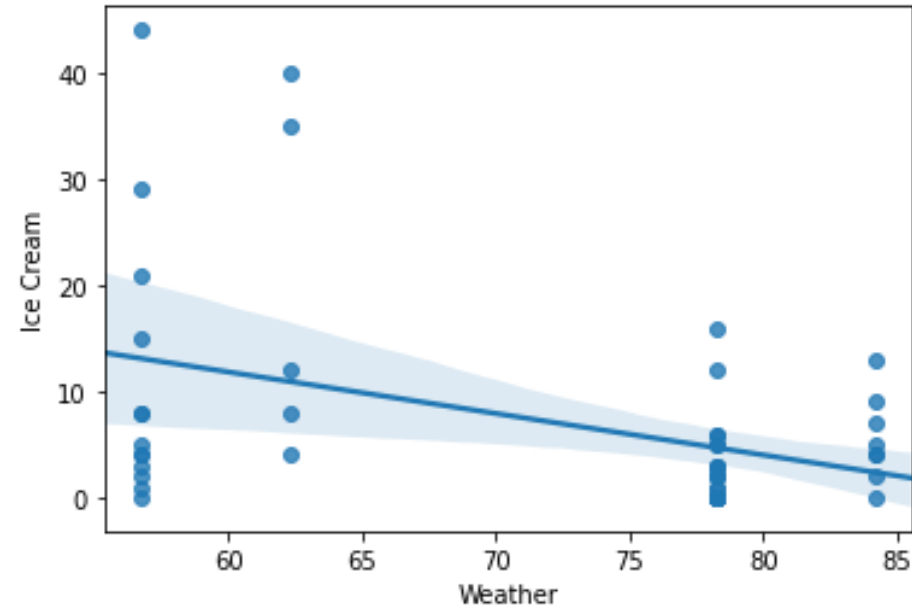
What Drives Ice Cream Location

Multi-factor analysis

Multiple Linear Regression

There are a number of factors that can plausibly drive ice cream shop location. The prior analysis shows that there is no one dominant driver. (Indeed there are at best weak correlations for any factor considered by itself.)

Let us then use all of our independent variables and fit a model that will predict the number of ice cream shops in a zip code.

If the actual number is significantly different from the number predicted by the model, we will conclude that the zip code is ripe for another shop.

```python
#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)
```

```python
#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
```

Solution

The best place for the next ice cream shop

New York City!

Zip 10007 offers the best opportunity, as measured by the greatest difference between the predicted and actual ice cream shops.

Best Opportunity	
City	New
Zip	10007
Population	7023
Income	246813
Avg Temperature	62.3333
Area	0.1
Current Ice Cream Shops	8
Modelled Ice Cream Shops	19.1204
Opportunity	11

Best Opportunity is: 10007