Coursera Capstone Project

Analysis of Sioux Falls restaurants

By: Allen Nash

April 2019

Introduction

The city of Sioux Falls, SD has been one of the highest rated cities in the United States for opening a business for many years. The state of South Dakota has no individual or corporate income tax, and business costs are more than 20 percent below the national average. Sioux Falls is also a rapidly growing city, in fact, its rate of population growth is nearly four times the national average. The healthcare sector is the cities biggest industry, but a close second is the financial sector, which consists of Citibank and Wells Fargo at the spearhead of this area. South Dakota has $3 trillion in bank assets, which is higher than any other state in the United States. Many secondary areas receive benefits from a growing and secure economy, one of which is the restaurant industry. This is especially important to an aspiring small business owner who may be cautious about creating a new restaurant without knowing what the future holds.

Business Problem

The intent of this project is to investigate the area of Sioux Falls SD while keeping several parameters in mind. These parameters that I will investigate in this report are:

1. What ratings are restaurants given as far as food safety is concerned?
2. Where are highest rated restaurants of Sioux Falls located?
3. Where would be the best location for the establishment of a new restaurant in the city?

Throughout this project, I will use many of the techniques and methods that I have learned in the data science courses. I will use webscraping, geographic data location via Foursquare, clustering models, and data analysis (kNN and logarithmic regression models) to create machine learning models to try to find the best location of a new restaurant.

My aim in this project is to help provide a new restaurant business with information to help them determine a possible location for a new business, whether the business owner intends to move a business to a new location, or start a business from scratch. The primary question that I wish to answer is: "Where is the absolute best location to open a new business based primarily on restaurant ratings?"

Target Audience

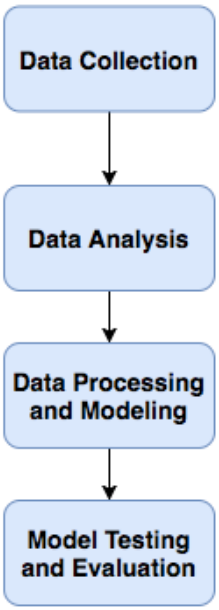
The primary audience in this report is any prospective business owners who would like to open a new restaurant in Sioux Falls. Also, any restaurant owners who wish to either expand or relocate their current business will find value in this report as well.

Data

For the food safety information of the report, the information will be primarily extracted from the state health department located at the web address :http://webapps.siouxfalls.org/inspections/restaurants.aspx. This site consists of all businesses in Sioux Falls that sell prepared food as well as bar establishments that sell alcoholic drinks. According to the state health department website, inspections are performed every 30 days, which would give a fairly accurate depiction of current conditions of the restaurants that are being evaluated. For the consumer ratings portion, I will be using Foursquare to find data about the various restaurants in the city. This will require web scraping, data cleaning and wrangling, K-means clustering, and map visualization using Folium.

Methodology

This portion of the report will describe the main components of the analysis and predictive modeling. The methodology follows the following flow chart which is separated into 4 parts:



1. **Data Collection**

The data was collected from the Sioux Falls Health Department website:

```
In [2]: url=requests.get('http://webapps.siouxfalls.org/inspections/restaurants.aspx').text
        soup = BeautifulSoup(url,'lxml')
```

2. **Data Analysis**

The data was organized into a data frame by looping and appending one row at a time. The code and output looks like the following:

```
In [5]: table_inspec = soup.find('table')
        fields = table_inspec.find_all('td')

        name = []
        address = []
        score = []
        score2 = []

        for i in range(1, len(fields), 4):
            name.append(fields[i].text.strip())
            address.append(fields[i+1].text.strip())
            score.append(fields[i+2].text.strip())
            score2.append(fields[i+2].text.strip())

        df_ins = pd.DataFrame(data=[name, address, score, score2]).transpose()
        df_ins.columns = ['Name', 'Address', 'Score','Score2']
        df_ins.head()
```

Out[5]:

	Name	Address	Score	Score2
0	18TH AMENDMENT (THE)	1301 W 41ST ST	78	78
1	1ST WOK	523 W 10TH ST	100	100
2	22 TEN KITCHEN & COCKTAILS	2210 W 69TH ST	91	91
3	AASHNA ASIAN MART	4713 E ARROWHEAD PKWY	97	97
4	ADDIS ABABA	818 E 8TH ST	90	90

The data was quite clean after this method, a second score column was added to help with the modeling and testing, which will be discussed later on in those sections of this report. By counting the indices, I noticed that there are 1047 rows and no null values.

3. Data Processing and Modeling—Data Processing

In order to create a machine learning technique, I chose to use Foursquare to use the addresses and create a coordinates column in a new data frame. I also created a loop to append the address column to add the city and state to restrict the addresses to the range within the city of Sioux Falls. The latitude and longitude was then split into two columns to allow analysis of these two features separately. The inspections data frame and the newly created geocode data frame were then merged in order to create a machine learning model that will use a single data frame. Refer to the coding to elaborate on all of the steps involved in this.

In [46]:
```
import os
from geopy import geocoders
from geopy.geocoders import Nominatim

#API_KEY = os.getenv( [redacted] )
#g = Nominatim(api_key=API_KEY)

geolocator = Nominatim(user_agent="foursquare_agent")
loc_coordinates = []
loc_address = []
for address in df_ins.Address:
    try:
        inputAddress = address
        location = geolocator.geocode(inputAddress, timeout=15)
        loc_coordinates.append((location.latitude, location.longitude))
        loc_address.append(inputAddress)
    except Exception as e:
        print('Error, skipping address...', e)


df_geocodes = pd.DataFrame({'coordinate':loc_coordinates,'address':loc_address})
```

```
Error, skipping address... 'NoneType' object has no attribute 'latitude'
Error, skipping address... 'NoneType' object has no attribute 'latitude'
```

The merged data frames looked like this:

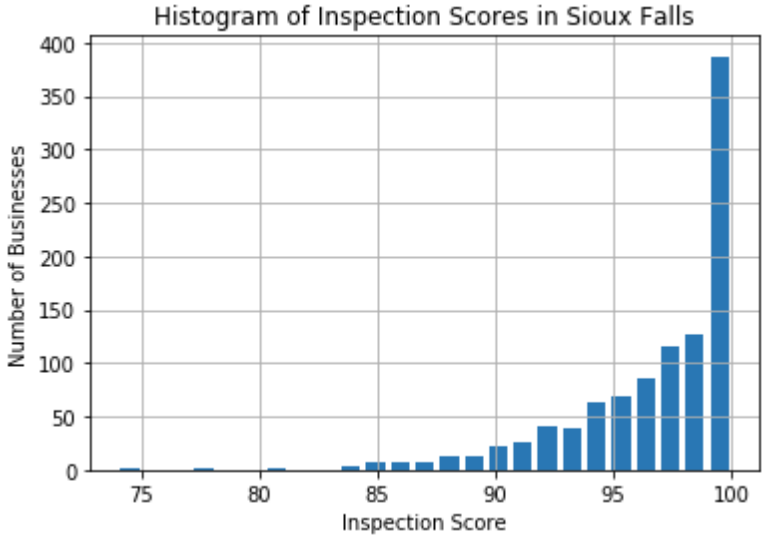
```
df_geo_ins2.head()
```

Out[37]:

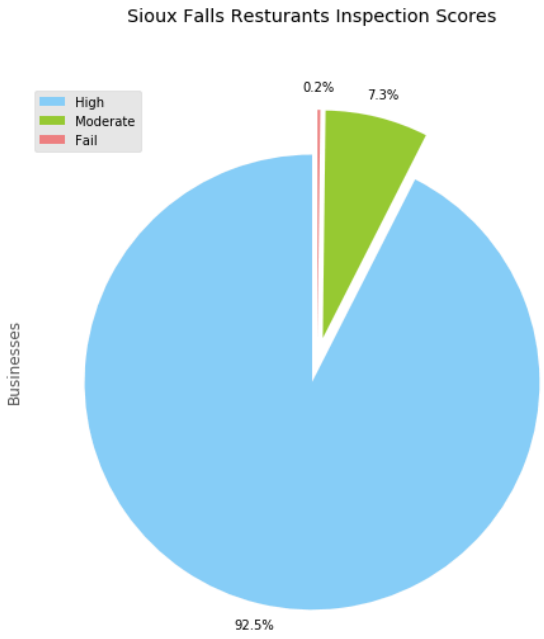
	address	coordinate	Name	Address	Score	Score2	Complete_Address	latitude	longitude
0	230 S PHILLIPS AVE, Sioux Falls, South Dakota	43.5449286, -96.7265849	1ST WOK	523 W 10TH ST	100.0	3.0	523 W 10TH ST, Sioux Falls, South Dakota	43.5449286	-96.7265849
1	5317 W 41ST ST, Sioux Falls, South Dakota	43.51472792, -96.791288	22 TEN KITCHEN & COCKTAILS	2210 W 69TH ST	91.0	3.0	2210 W 69TH ST, Sioux Falls, South Dakota	43.51472792	-96.791288
2	5020 S MARION RD, Sioux Falls, South Dakota	43.4999782708437, -96.7907087036966	AASHNA ASIAN MART	4713 E ARROWHEAD PKWY	97.0	3.0	4713 E ARROWHEAD PKWY, Sioux Falls, South Dakota	43.4999782708437	-96.7907087036966
3	5020 S MARION RD, Sioux Falls, South Dakota	43.4999782708437, -96.7907087036966	ADDIS ABABA	818 E 8TH ST	90.0	3.0	818 E 8TH ST, Sioux Falls, South Dakota	43.4999782708437	-96.7907087036966
4	5005 S WESTERN AVE, Sioux Falls, South Dakota	43.500235, -96.74808	AEROSTAY HOTEL	2821 N JAYCEE LN	100.0	3.0	2821 N JAYCEE LN, Sioux Falls, South Dakota	43.500235	-96.74808

3. Data Processing and Modeling—Modeling

The first model that I looked at was a histogram in order to create a better visual representation of the skewness of the scores.



As was very apparent without too much modeling since the data set was fairly small, most food establishments in the city during this period passed inspections. I then chose to break up the data further, into three groups, where the highest scores from 90 on up were in the top category with a rating of 3 in the "Scores2" column, followed by the mid-range group (between 80 and 89) that still passed inspection but by a narrower margin were in the second category (rating of 2) and finally the two businesses that did not pass inspection during this time (below 80) were in the last group (rating of 1). This categorical information was then used to create a pie graph to create another model for this data.



The primary model that I chose to use for testing was K-Nearest Neighbor using the latitude and longitude columns as the independent variables to see if there was any correlation with the scores that restaurants received. I believe with the geographical information that I was able to find, this method would yield the best results.

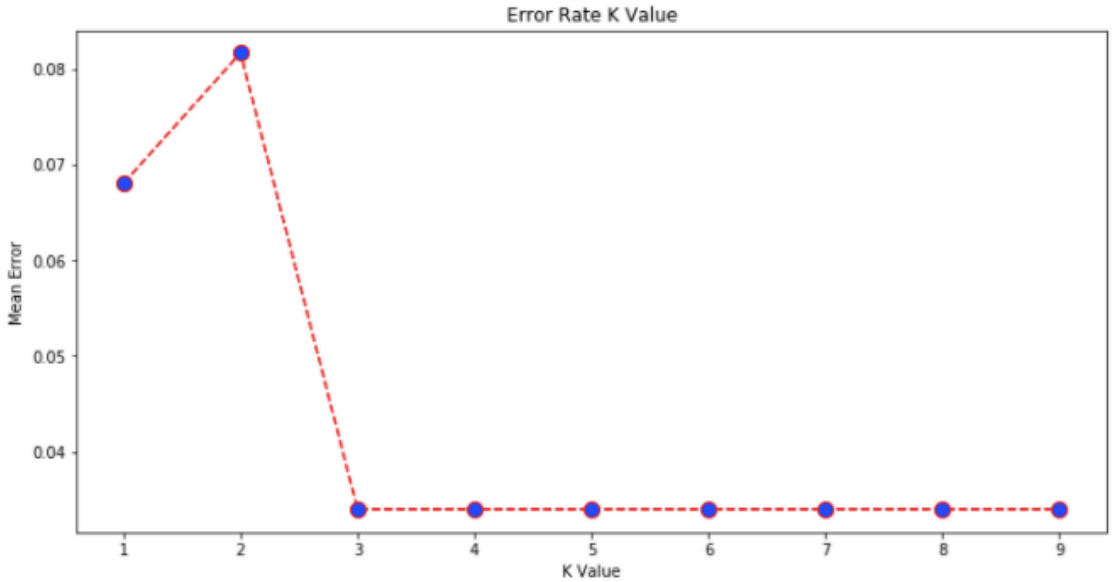
4. Model Testing and Evaluation—Model Testing

After organizing the data into a more usable data frame with latitude, longitude, and score, I was then able to use sklearn to create a training and test set over the data, with a test size of 0.15. Featured scaling was also used prior to using the NeighborsClassifier to try to find the best k value. I choose a K-Value of 5 initially, but then reduced this to 2 after evaluating my results.

4. Model Testing and Evaluation—Evaluation

After testing, I chose to use a confusion matrix to see the accuracy of the classifications of the KNN model. I noticed that the f1 score was at 95%, which is extremely high. This algorithm performed very well with this dataset.

As stated in the model testing phase, I chose to lower my K-Value when I calculated the mean error and plotted this value against the error of K-Values.



A K-Value of 2 was the highest spike in the graph with the greatest change in the graph occurring between the K-Values of 2 and 3.

Conclusion

This study was primarily focused on analyzing food inspection scores in the city of Sioux Falls and determining if the geographical location in the city of Sioux Falls has an influence on the ratings. I found that there seems to be a very high rating on average for the performance levels of Sioux Falls restaurants and food vendors, with a very low amount of restaurants failing inspection over the course of the last 30 days. Based on predictive modeling, I was able to create a machine learning model of this situation to predict with a very high degree of accuracy wether a restaurant will pass inspection or not.

Future Considerations

Even though my model testing of a confusion matrix performed very well, I would like to continue evaluating the food inspections over the course of a longer duration, perhaps a year or more. I would like to have a larger sample set to see how accurate my model stays. As there were very few restaurants that failed inspection (only 2) I believe a model that is truer would need to take place over a greater span of time.