

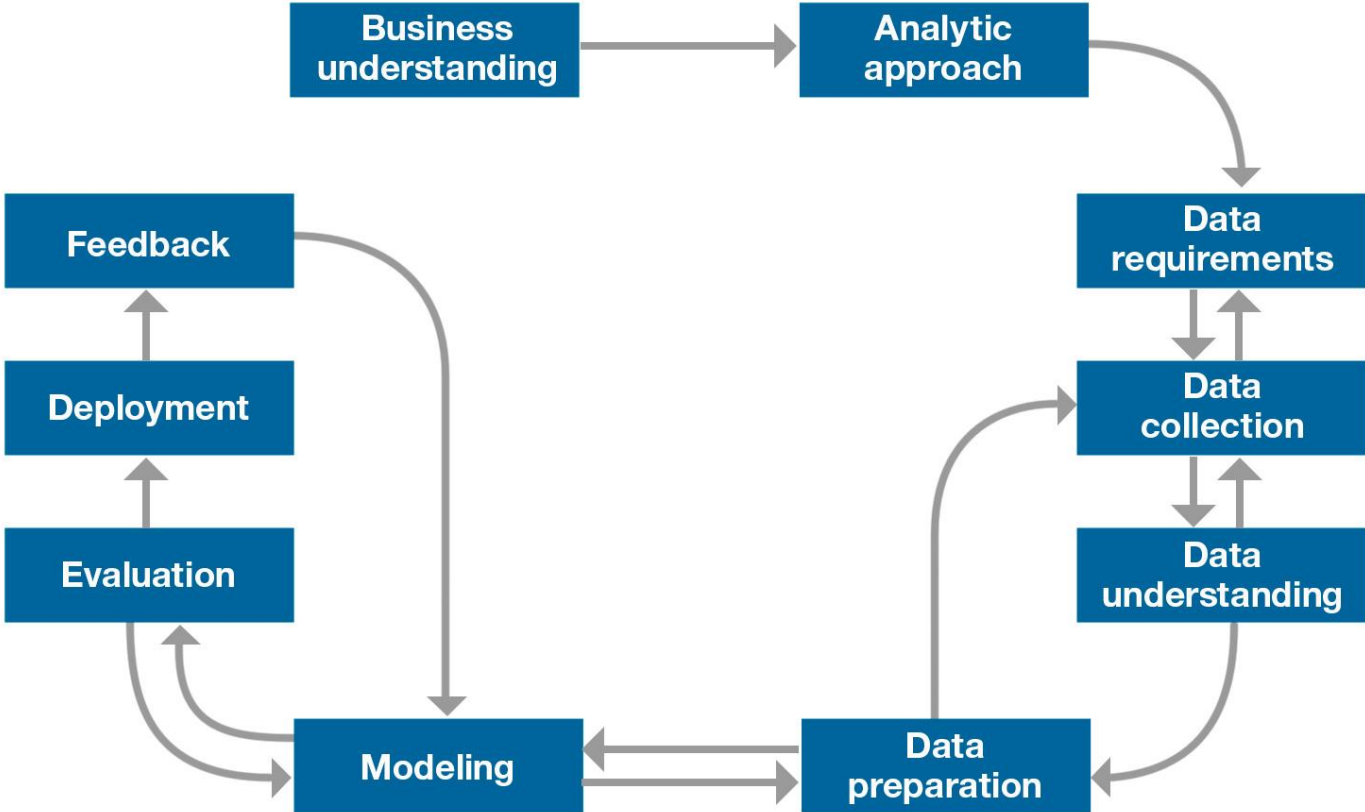


SILVER FOX DATA

CLEVER, CRAFTY, CONFIDENT

The Battle of the Neighborhoods
Author: Mark L Cichonski
IBM Data Science Capstone
1/26/2021

Roadmap



Abstract

This project is the capstone assignment for a 9 course IBM Professional Certificate in data science.

The scenario is that you live on the West side of the City of Toronto in Canada. You love your neighborhood mainly because of all the great amenities and other types of venues that exist in the neighborhood. Such as gourmet fast food joints, pharmacies, parks, grad schools, and so on. Now say you receive a job offer from a great company on the other side of the city with great career prospects. However, given the far distance from your current place, you unfortunately must move if you decide to accept the offer. Wouldn't it be great if you're able to learn about neighborhoods on the other side of the city? They are the same as your current neighborhood, and if not, perhaps similar neighborhoods that are at least closer to your new job.

In this capstone project, we had to be creative and produce our own idea or problem to solve using location data. I chose the scenario where my company is looking for office space and will use location data to do so. Hence the name of the capstone project is the battle of the neighborhoods.

Given my city of Charlotte, I will segment it into different neighborhoods using the geographical coordinates of the center of each neighborhood. And then, using a combination of location data and machine learning, I will group the neighborhoods into clusters and choose the optimal location for our office!

Business Understanding

•Seek clarification

You have the opportunity to be as creative as you want and come up with an idea to leverage Foursquare location data to explore or compare neighborhoods or cities of your choice or to come up with a problem that you can use the Foursquare location data to solve.

•What problem are you trying to solve

My company is opening a new office and we are looking for the prime location.

•Clearly defined question

What is the best location to open a consulting business in Charlotte North Carolina?

- What is the goal?

Find the prime location of an office for a consulting company.

- Objectives in support of the goal. The criteria for selection will be:

- A. Marketplace and demographics
- B. Competition
- C. Costs
- D. Growth Potential
- E. Safety
- F. Recruiting
- G. Brand Image

For this analysis, we will use FourSquare and additional data sources as needed to build the most comprehensive picture and select the best location for the office.

Analytic Approach

We will use location data and other data sources to build a picture of the Charlotte area and optimize our location based on the factors identified by the business as critical for the needs of a small business.

These factors will provide us with the location based on the availability of facilities, amenities, safety and the population makeup of any potential areas.

This will allow us to choose a location with the proper traffic and in accordance with the brand image we want the company to have.

To evaluate these factors we plan on using the information from FourSquare as primary data source, supported by other data as needed, to conduct segmenting and clustering of neighborhoods in Charlotte to determine the optimal location for our office.

Analytic Approach

	Expected Analytical Approach			
Factors	Descriptive	Diagnostic	Predictive	Prescriptive
Marketplace and demographics	X	X	X	
Competition	X	X	X	
Costs	X	X	X	X
Growth Potential	X	X	X	
Safety	X	X	X	
Recruiting	X	X	X	X
Brand Image	X	X	X	X

Descriptive-Current status, what is the current state?

Diagnostic-Why is the current state the way it is?

Predictive-What will happen in the area to determine future success?

Prescriptive-Where should we locate our office?

Data Requirements

	Data Sources (Recipe)			Data Mining Technique
Factors	4 Square	Additional Data	Additional Source	
Marketplace and demographics	X	X	https://www.neighborhoodscout.com/nc/charlotte/demographics	Table-screen scrape
Competition	X	X	https://www.expertise.com/nc/charlotte/business-consultants#:~:text=Here%20are%20the%20Picks:%201%20Balan%20lanyne%20Tax%20&,of%20consulting%20experience%20to%20business%20clients%20in%20Charlotte.	Table-screen scrape
Costs	X	X	https://www.crexix.com/properties?utm_term=%2Bcommercial%20%2Bestate%20%2Breal%20%2Bestate&utm_campaign=Bing_NonBrand_National_Sales&utm_source=bing&utm_medium=ppc&msclkid=b3dfd627eda714c838efb63e6cabe478	Data-screen scrape
Growth Potential	X			
Safety	X	X	https://www.neighborhoodscout.com/nc/charlotte/demographics	Table-screen scrape
Recruiting	X	X	http://blog.parkerlynch.com/jobs-report/charlotte-nc/	Data-screen scrape
Brand Image	X			

Data Collection

1. First step was to collect zipcodes:
 1. Data source: Zip Codes to Go
 2. Method: Pandas Read html
 3. Link: [ZIP Code List for North Carolina \(zipcodestogo.com\)](http://zipcodestogo.com)
2. Need longitude and latitude by zipcode
 1. Data source: opendatasoft
 2. Method: Download and read CSV
 3. Link: [US Zip Code Latitude and Longitude — Opendatasoft](https://opendatasoft.com/datasets/us-zip-code-latitude-and-longitude)
3. Foursquare for location analysis
 1. Data source: Foursquare
 2. Method: API/JSON
 3. Link: [Foursquare API](https://foursquare.com/docs/api)
4. Crime data for factor analysis
 1. Data source: Best Places
 2. Method: Manual lookup on finalists.
 3. Link: [Zip 28205 \(Charlotte, NC\) Crime \(bestplaces.net\)](http://bestplaces.net/crime/zip/28205)
5. Commute time for factor analysis
 1. Data Source: Best Places
 2. Method: Manual lookup on finalists
 3. Link: [Zip 28205 \(Charlotte, NC\) Commuting \(bestplaces.net\)](http://bestplaces.net/commute/zip/28205)
6. Cost/Pricing
 1. Data Source: Loop Net
 2. Method: Search
 3. Link: [28207 \(Charlotte\) Commercial Real Estate for Sale | LoopNet.com](http://loopnet.com/real-estate/charlotte/28207)

Data Understanding

	1.Zips	2. Location	3. Foursquare	4. Crime	5. Commute	6. Cost
Factors	X	X	X	X	X	X
Marketplace and demographics			X		X	
Competition						
Costs						X
Growth Potential			X	X		
Safety				X		
Recruiting						
Brand Image			X			

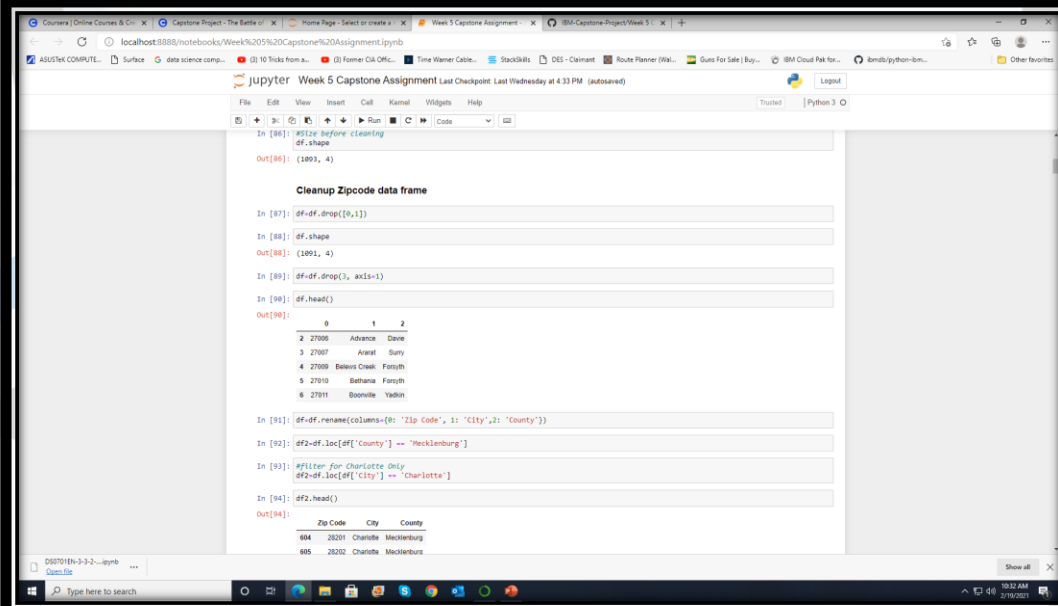
After looking at the available data, we decided that the amount of data available in the above 6 data sources was sufficient to make the location decision.

The zip code data for North Carolina had over 1,000 records that had to be narrowed to the Charlotte area. This was then narrowed to the 74 zip codes in the Charlotte area for analysis. The file with the lat/lon for the zip codes was over 43,000 records that had to be merged with the 74 Charlotte areas zip codes.

One challenge for this project was that the zip code names did not exist and had to be looked up manually by map.

Data Preparation

1. Download zip code data in CSV.
2. Drop un-needed columns.
3. Rename columns for understanding.
4. Filter to a new data frame for Charlotte only.
5. Download zip code lat/long information in CSV.
6. Convert code field to object type for merge with zip code data.
7. Merge files and drop un-needed columns.
8. Renamed columns of new data frame.
9. Removed incorrect zip codes.
10. Change cluster labels to integers.
11. After cluster analysis, had to drop NAs for decision tree.
12. Create dummy variables for Y, location in decision tree analysis.



The screenshot shows a Jupyter Notebook interface with the following code and output:

```
In [86]: #Size before cleaning
df.shape
Out[86]: (1893, 4)
```

Cleanup Zipcode data frame

```
In [87]: df=df.drop([0,1])
In [88]: df.shape
Out[88]: (1891, 4)
In [89]: df=df.drop(3, axis=1)
In [90]: df.head()
Out[90]:
```

	0	1	2
2	27006	Advance	Clare
3	27007	Arenal	Sury
4	27009	Bellevue Creek	Forsyth
5	27010	Bethania	Forsyth
6	27011	Boonville	Yadkin

```
In [91]: df=df.rename(columns={0: 'Zip Code', 1: 'City', 2: 'County'})
In [92]: df2=df.loc[df['County'] == 'Mecklenburg']
In [93]: #Filter for Charlotte Only
df2=df.loc[df['City'] == 'Charlotte']
In [94]: df2.head()
Out[94]:
```

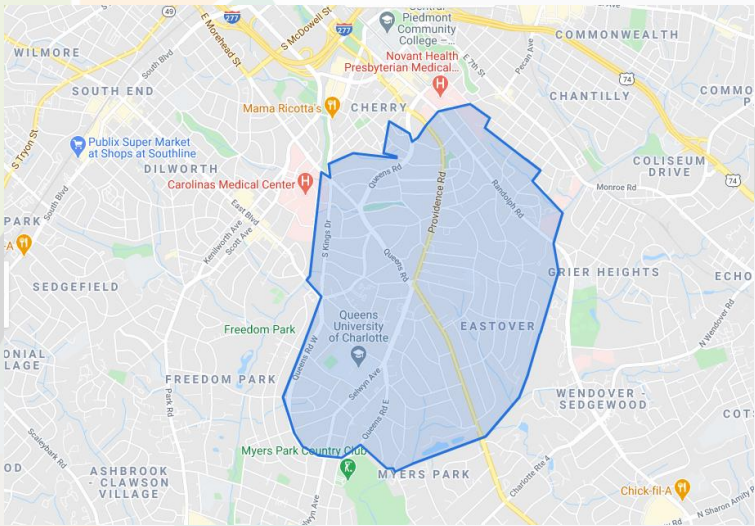
Zip Code	City	County
884	28201	Charlotte
885	28202	Charlotte

Modelling

Two models were developed for this analysis:

1. Kmeans Clustering. In the analysis, the average mean occurrence of each venue was calculated and stored in a data frame for the 41 neighborhoods. There were 137 venue types. The KMeans algorithm clusters data by trying to separate samples in n groups of equal variance, minimizing a criterion known as the inertia or within-cluster sum-of-squares. In this case, it is clustering the 41 neighborhoods based on the distance between venue occurrence means. I believe in this data set, there are so many 0 values, it is drastically effecting the results.
2. Decision Trees are a non-parametric supervised learning method used in this case for classification. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation. In this case the Y was the neighborhood and the X's were the factors (narrowed down); food, hotel, bar, business services, violent crime, property crime and commute time. The decision tree looked first at violent crime and split the 5 samples with $4 \leq 39.5$. Then it looked at food establishments which we did not think was critical. We went left to the cost and the tree split even with properties less than \$437500. We went back and looked at the 2 properties and chose the lowest cost, lowest crime and shortest commute, Eastover.

After clustering and decision tree classification, Eastover was selected as the optimal neighborhood.



Based on this analysis and our budget, we chose the following location:

3120 Latrobe Dr - Aberdeen Place
5,600 SF Office Condo Unit Offered at \$840,000 in Charlotte, NC



ABOUT 3120 LATROBE DR, CHARLOTTE, NC 28211 UNDER CONTRACT

Price \$840,000 Building Class B

Conclusion