The Best Location for a New Starbucks in Philadelphia

Cathryn Pierce

April 2, 2019

1. Introduction

1.1 Background

Starbucks is undoubtedly one of the largest coffee shops with over 30,000 locations worldwide. One aspect imperative to Starbucks' success is where to place a new location. Starbucks announced last year that they would be closing 150 poorly performing company-operated stores in 2019. The affected stores were located in mostly urban areas that were already densely populated with Starbucks locations. With Philadelphia being the 6th most populous U.S. city, it would be a great city for a new Starbucks location, if coffee shop density was taken into account. It is advantageous for Starbucks to look at different postal codes within Philadelphia to determine which postal code would be best to open a new, successful Starbucks location.

1.2 Problem

The data that would most useful in determining the best location to open a new Starbucks would be postal code data describing the most common venues within each postal code and population data of each postal code. This project takes an in-depth look of Foursquare data to determine which postal codes are the least populated with coffee shops and have the greatest population in order to determine the best postal code to open a new Starbucks.

1.3 Interest

With the announcement of 150 stores closing in the year 2019, Starbucks would be highly interested in the data presented in this project. It would assist them in the best place to open a new location with the highest probability of being successful.

2. Data Acquisition and Cleaning

2.1 Data Sources

The postal codes of Philadelphia were scraped from http://ciclt.net/sn/clt/capitolimpact. The latitude and longitude data was a csv file retrieved from https://public.opendatasoft.com. The population data was scraped from https://www.zipdatamaps.com/zipcodes-philadelphia-pa. Lastly, Foursquare was utilized to retrieve a list of venues based on the coordinates of Philadelphia and the coordinates of each postal code.

2.2 Data Cleaning

The postal code data had repeated rows for certain postal codes that were located in multiple parts of the city in Philadelphia. These were grouped together so that only one postal code existed for each. All of the data was merged together on the postal codes and all unnecessary columns were dropped (Table 1). The csv file used to provide the latitude and longitude of the postal codes had to be filtered and all postal codes not found in Philadelphia dropped along with columns not relevant to this study (Table 1). Postal codes with a population of zero (indicating a P.O. Box) were dropped from the dataframe in addition to city labels (Table 1).

Foursquare venue data was limited to 100 venues per postal code. It was then narrowed down to the top ten venues per postal code. Then only postal codes that didn't contain a coffee shop in their top 9 most common venues were chosen for further population analysis.

Table 1. Feature Selection during Data Cleaning

Kept features	Dropped features	Reason for dropping
PostalCode	City	All of the postal codes are going to be located in Philadelphia.
PostalCode, Latitude, Longitude	City, State, Timezone, Daylight savings time flag, geopoint	Repetitive data, unnecessary columns for this study
PostalCode, Population	Zip Code Name	Repetitive data

2.3 Feature Selection

After grouping the postal codes together I ended up with 55 postal codes in total. Further merging on the postal code with the other data resulted in 47 postal codes to analyze. These postal codes were used in the in the Foursquare data to find the list of venues and ultimately the most common venues for each postal code.

3. Methadology

3.1.1 Postal Code Retrieval

The postal code table scraped from http://ciclt.net/sn/clt/capitolimpact was placed in the df_phil dataframe. The postal codes were grouped together and joined the 'City' column joined. After grouping the data, there were 61 unique postal codes.

	PostalCode	City
0	19019	Philadelphia
1	19101	Philadelphia
2	19102	Mid City East
3	19102	Middle City East
4	19102	Philadelphia

3.1.2 Latitude and Longitude Retrieval

The latitude and longitude csv file read from https://public.opendatasoft.com were placed in the df_lonlat dataframe. All unnecessary columns were dropped and then the dataframes closely resembled each other.

	Zip	Latitude	Longitude
0	19121	39.981062	-75.17450
1	19154	40.091460	-74.97719
2	19161	40.001811	-75.11787
3	19119	40.053511	-75.18858
4	19120	40.033944	-75.12118

3.1.3 Population Retrieval

The population data was scraped from https://www.zipdatamaps.com/zipcodes-philadelphia-pa and placed into the dataframe df_pop.

	PostalCode	Population
0	19102	4705
1	19103	21908
2	19104	51808
3	19106	11740
4	19107	14875

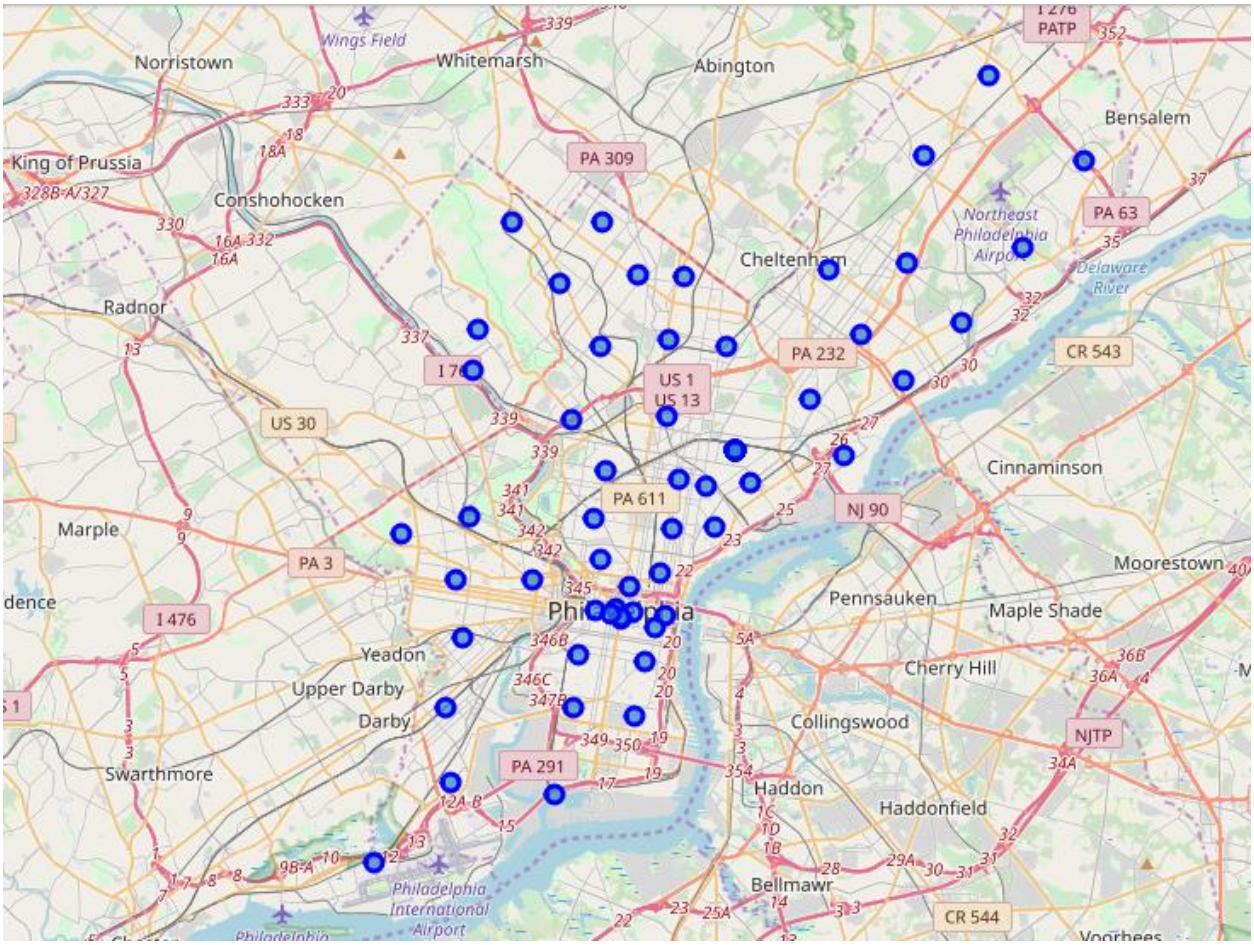
3.1.4 Merging df_phil and df_pop

The two dataframes, df_phil and df_pop were merged together on the PostalCode column and placed in new dataframe df_merged. After the postal codes with a population of zero were dropped, 47 postal codes remained for further analysis.

	PostalCode	City	Latitude_x	Longitude_x	Latitude_y	Longitude_y	Latitude	Longitude	Population
2	19102	Mid City East,Middle City East,Philadelphia	39.952962	-75.16558	39.952962	-75.16558	39.952962	-75.16558	4705
3	19103	Mid City West,Middle City West,Philadelphia	39.952162	-75.17406	39.952162	-75.17406	39.952162	-75.17406	21908
4	19104	Philadelphia	39.961612	-75.19957	39.961612	-75.19957	39.961612	-75.19957	51808
6	19106	Philadelphia	39.951062	-75.14589	39.951062	-75.14589	39.951062	-75.14589	11740
7	19107	Philadelphia	39.952112	-75.15853	39.952112	-75.15853	39.952112	-75.15853	14875

3.2 Folium Map Representing Postal Codes

I used the folium library to visualize the geographical data of Philadelphia. It represents the postal codes as the blue icons using the latitude and longitude data.



3.3 Foursquare Data for Postal Codes

The foursquare data was used to gather a limit of 100 venues within a 500 meter radius of each of the provided postal codes.

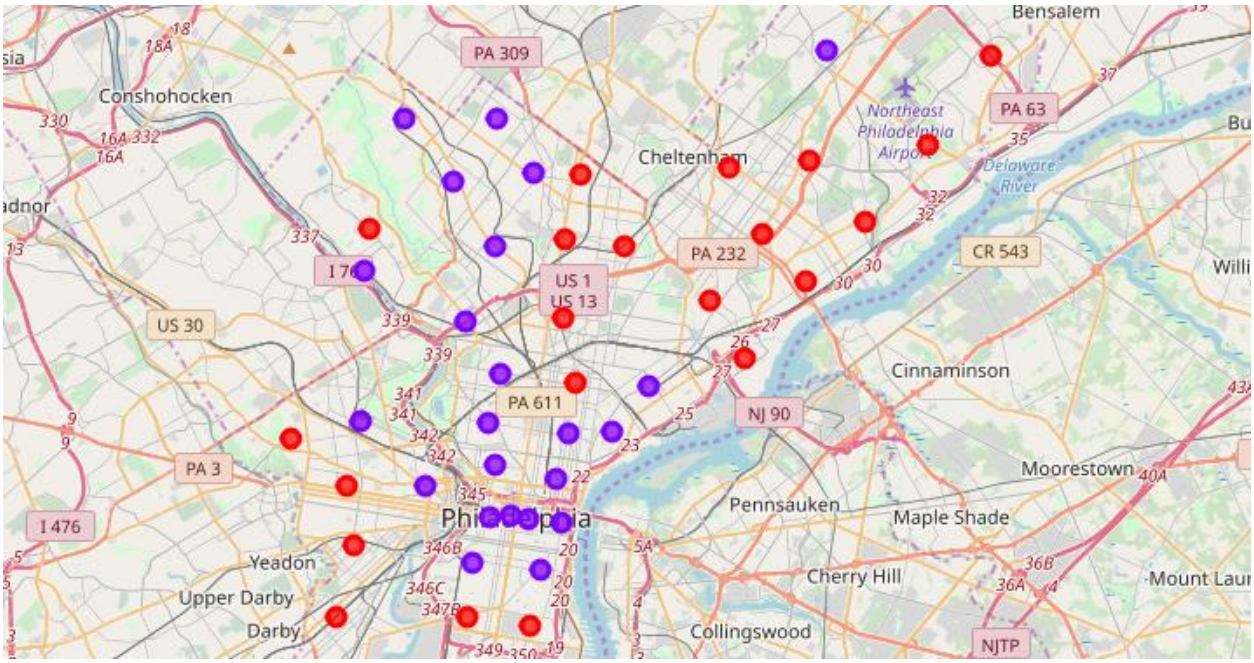
	PostalCode	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	19102	39.952962	-75.16558	Dilworth Park	39.952772	-75.164723	Park
1	19102	39.952962	-75.16558	La Colombe Coffee Roasters	39.951659	-75.165238	Coffee Shop
2	19102	39.952962	-75.16558	City Hall Courtyard	39.952484	-75.163592	Plaza
3	19102	39.952962	-75.16558	JFK Plaza / Love Park	39.954123	-75.165303	Plaza
4	19102	39.952962	-75.16558	One Liberty Observation Deck	39.952740	-75.168068	Scenic Lookout

There were 230 unique categories of venues returned by Foursquare so the data was organized to only show the top 10 venues for each postal code.

	PostalCode	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue
0	19102	Coffee Shop	Hotel	Italian Restaurant	Cosmetics Shop	American Restaurant	Seafood Restaurant	Yoga Studio	Salad Place
1	19103	American Restaurant	Coffee Shop	Deli / Bodega	Vegetarian / Vegan Restaurant	New American Restaurant	Bar	Bakery	Sushi Restaurant
2	19104	Pizza Place	Bakery	Department Store	Deli / Bodega	Piano Bar	Cosmetics Shop	Coffee Shop	Sandwich Place
3	19106	History Museum	Coffee Shop	Historic Site	Boutique	Italian Restaurant	American Restaurant	Art Gallery	Bar
4	19107	Bakery	Sandwich Place	Hotel	Chinese Restaurant	Mediterranean Restaurant	Gastropub	Hot Dog Joint	Shanghai Restaurant

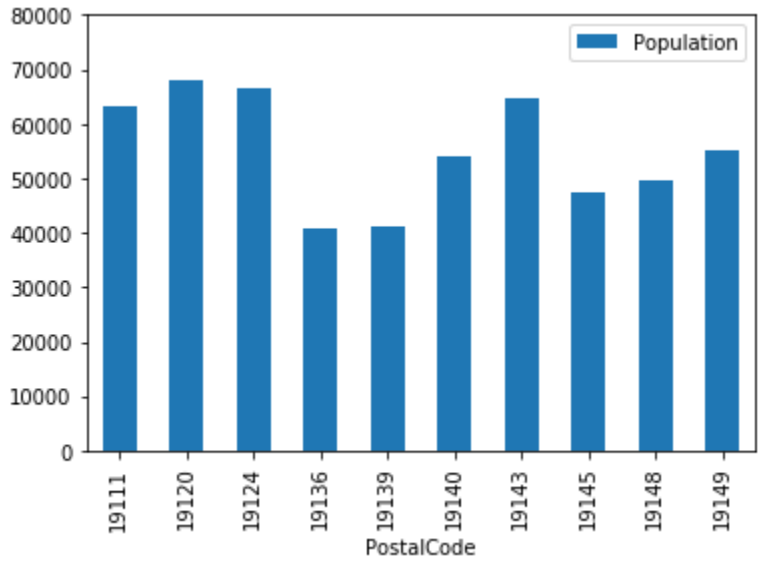
3.4 K-means Clustering

A cluster map was then made segmenting the postal codes into clusters. Cluster 1, indicated in red, appears to be clustered by a high volume of restaurants. Cluster 2, in purple, has more bars, retail stores, and fast food shops. Cluster 3, not pictured here, only contained one postal code and was segmented based on its venues that contained food in their classification of venue and contained a zoo exhibit.



3.5 Bar Chart Representing Population

Lastly, the postal codes with a coffee shop as their top 9 most common venues were dropped from the data. The remaining data was then placed into a bar plot by population size. Only postal codes with a population of more than 40,000 were included in the bar plot for easier interpretation.



The bar plot makes it easy to see which postal code would be the best to select to build a new Starbucks location.

4. Conclusion

Over the course of this study, I analyzed various postal codes throughout the greater Philadelphia area to extrapolate the most optimal location for Starbucks to open a new store. I concluded that the variables most useful in making my determination would be the population, as well as, what most common venues were for each respective postal code. Postal codes with a population of zero were excluded from the analysis because they were P.O. boxes or similar. The ten most common venues were found for each postal code and from that data, the postal codes that did not have a coffee shop in their top nine most common venues were selected for population analysis. The postal code 19120 would be the best place to open a new Starbucks.