IBM Coursera Capstone Project

Introduction:

Problem: If I were to develop an app that recommends local business to users, and I wanted to use the foursquare location data as my data source, what cities in the United States would be the best cities to launch this app in first.

How would be know what cities would be best to start in? The answer to this question can be found by analyzing the foursquare location data. We would want to start in cities where there are high numbers of venues that have been designated in the foursquare database. In order for an app to give useful recommendations, it needs to have a large dataset to draw from. There would be no point in having an app if it has only a few recommendations to give. We would also want to start in cities with large populations so we can get a large user base.

Data:

The data used for this project came from two sources. In order to choose the cities with large populations in the United States, the pandas package was used to read in a dataframe of the highest population cities with their corresponding locations from the website https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population. The imported dataframe was then cleaned and sorted to only include the top 50 cities by population using the pandas library in python. Using the longitudes and latitudes of these cities, requests were sent to the foursquare API to search for the number of venues in their database within a radius of 100 km of the city location. The retreived information was put into the pandas database and was ranked in descending order based on the number of venues the were retrieved. This allows us to see the top 50 cities where we should start the recommendation app availability if we are going to be using the foursquare location data for our app.

Methodology

I first used pandas library to read in a table from Wikipedia that contained US cities and their populations and location data. Once that table was read into python as a pandas dataframe, I cleaned the dataframe. This cleaning included dropping unnecessary columns, renaming columns, removing unneeded characters from the titles of columns. I also removed unnecessary characters from the cells of the dataframe (eg. units), so the cells only contained numbers. I then changed these numbers from strings to floats so that the dataframe could be sorted by these float numbers. I also had to split the location column into two separate columns (longitude and latitude). I then used that dataframe to create a new dataframe with only the 50 cities in the US with the highest

populations. I then defined my foursquare credentials and used the search API to find out how many foursquare venues were in each city. I used a radius of 100km from the location of the city as provided by the Wikipedia page. I defined a function that looped through the top 50 cities and counted how many venues were returned for each city and created a new dataframe showing these cities and their number of venues. I then sorted the cities based on the number of venues to show the cities with the highest number of venues.

Results:

	City	State	Number of Venues on Foursquare
0	Baltimore	Maryland	122
1	Seattle	Washington	122
2	Nashville	Tennessee	122
3	Long Beach	California	121
4	Jacksonville	Florida	120
5	Kansas City	Missouri	120
6	Albuquerque	New Mexico	120
7	San Diego	California	120
8	Portland	Oregon	119
9	Arlington	Texas	118
10	New York	New York	117
11	Memphis	Tennessee	116
12	Sacramento	California	114
13	San Francisco	California	114
14	Tampa	Florida	114
15	Tulsa	Oklahoma	112
16	Tucson	Arizona	112
17	Fort Worth	Texas	111
18	Dallas	Texas	110
19	Colorado Springs	Colorado	105

20	Louisville	Kentucky	104
21	Los Angeles	California	104
22	Atlanta	Georgia	102
23	Philadelphia	Pennsylvania	101
24	Columbus	Ohio	97
25	Oklahoma City	Oklahoma	97
26	Chicago	Illinois	96
27	Virginia Beach	Virginia	96
28	Phoenix	Arizona	93
29	Omaha	Nebraska	93
30	Minneapolis	Minnesota	79
31	San Antonio	Texas	79
32	Miami	Florida	77
33	Raleigh	North Carolina	77
34	Houston	Texas	74
35	San Jose	California	72
36	Fresno	California	68
37	New Orleans	Louisiana	66
38	Oakland	California	65
39	El Paso	Texas	65
40	Detroit	Michigan	62
41	Indianapolis	Indiana	54
42	Denver	Colorado	53
43	Charlotte	North Carolina	52
44	Austin	Texas	51
45	Milwaukee	Wisconsin	50
46	Las Vegas	Nevada	45
47	Washington	District of Columbia	40
48	Mesa	Arizona	28
49	Boston	Massachusetts	27

The table above shows the top 50 cities you should start in when making an app using foursquare data to locate venues as they have the most venue points within 100 km radius. These cities have the most venues returned when a search query is done using the location data from the website we chose for city and population data. These are also the cities within the top 50 populations in the United States.

Discussion:

When I ran the data analysis the first time, I was surprised by the results. For instance, San Francisco only returned 4 results. I would have guessed that a city like San Francisco, with such an active tech community, would use an app like foursquare a lot more. New York also had less results than I expected. When I looked back at the location data to try to determine what had happened, I looked up the longitude and latitude of San Francisco and New York City supplied by the website. When I punched that table location into google, it gave me locations that were slightly different than the city's actual locations. When I changed the table longitude and latitude for San Francisco and New York City to the more accurate google values, I got more predictable higher results. This leads me to believe that some of the other location data values may be inaccurate as well. If I were to do this project again (or if I had more time to refine it) I would use a different dataset for my location data.

Conclusion:

In conclusion, this project showed that it is not the cities you would expect that have the most robust foursquare location venue data available (eg. Nashville Tennessee was one of the highest), and this analysis helped to show that fact as long as the locations we used were correct. This analysis showed which highly populated cities in the United States would achieve the best results from an app recommending locations using the foursquare location database.