FIND BEST LOCATION FOR NEW OFFICE

Course Name

Applied Data Science Capstone

Created By

Miraalfasa Sharma

Created Date

08 December 2019

INTRODUCTION

BACKGROUND

These days IT Services are much in demand, more and more new start-up companies arise. Most of the IT Service Companies in the Asia targets USA as Development Cost of Asian Companies are less than the US. Moreover, USA Business Persons are interested in getting IT Service done remotely at cheaper rates but as there are might Point of Contacts or Office in USA for support, sometimes they avoid dealing with which is ultimately loss to the Sales of IT Service. To increase their sales by having a new marketing office in USA we need to select the best place for their office where more people are expose to the office and marketing team can get the potential targeted audience.

PROBLEM

As we are looking for a perfect potential location in the USA, we need to know the high density areas where people spend more and where they spend more, so that, the people can marketing office and marketing team can contact them for the lead generation and can have a good sales conversion.
So, to increase the ultimate lead generation and sales it is required to analyze the best place to start a Marketing Office.

INTEREST

The CEO of IT Company is interested in starting a new marketing office in the best locality of all the cities in United states. According to him, following constraints defines a best locality.

- *Population density of a locality*
- *Per Capital income*
- *Population of each location*
- *Venues in each locality*

The category of the venues that he's interested in are,

- *Shops & Service*
- *Professional & Other places*

DATA

To find the best location for marketing office, we need to get the data from the following sources:

1. *List of all the cities in United States with population density and coordinates: https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population*
2. *List of all the cities in United States with Per Capita Income: https://en.wikipedia.org/wiki/List_of_United_States_counties_by_per_capita_income*
3. *Using Four Square API to get the following:*
    a. *List of all venues in each city.*
    b. *List of all venues in each locality in the selected city.*

By using the above data source, based on the values like Population density, per capita income of the state, number of venues our first step will be to select best city to proceed with. (We will be giving weights to each venue as per its category).

After selecting city our next step will be for the localities of that city. We will repeat the same approach i.e. based on the scores of venues in each locality.

DATA ACQUISITION AND CLEANING

We used 3 data sources to solve this problem and each of it need to be processed to have a useful and precise data.

FIRST DATA ACQUISITION CLEANING

List of all the cities in United States with population density and coordinates:
https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population

We get a list of all the cities in United States with population density and coordinates from Wikipedia by scraping the page using library BeautifulSoup. We got following data:

1. *"Rank",*
2. *"City",*
3. *"State",*
4. *"Estimated Population",*
5. *"Enumerated Population",*
6. *"Percentage Change",*
7. *"Sq. Area in Miles",*
8. *"Sq. Area in KM",*
9. *"Population density in Sq. Mi",*
10. *"Population density in Km2",*
11. *"Location"*

Out of above 11 properties we need to decide which is the actually useful and need to remove unnecessary properties.

We need only ***City, State, Population density in Km2, Location and Radius*** *so, we will drop following columns*

"Rank", "estimated population", "enumerated population", "percentage change", "Sq. Area in Miles", "Sq. Area in KM", "population density in Sq. Mi".

The data after removing unnecessary properties looks as below:

	City	State	Population density in Km2	Location	Radius
0	New York[d]	New York	10,933/km2	40.6635°N 73.9387°W	17.363755
1	Los Angeles	California	3,276/km2	34.0194°N 118.4108°W	21.649480
2	Chicago	Illinois	4,600/km2	41.8376°N 87.6818°W	15.076472
3	Houston[3]	Texas	1,395/km2	29.7866°N 95.3909°W	25.248762
4	Phoenix	Arizona	1,200/km2	33.5722°N 112.0901°W	22.750824

SECOND DATA ACQUISITION CLEANING

List of all the cities in United States with Per Capita Income:
https://en.wikipedia.org/wiki/ListofUnitedStatescountiesbypercapitaincome

We get a list of per capita income of Counties of United States from Wikipedia by scraping the page using library BeautifulSoup. We got following data:

1. *"Rank",*
2. *"County or county equivalent",*
3. *"State",*
4. *"Per Capita Income",*
5. *"Median Household Income",*
6. *"Median Family Income ",*
7. *"Population",*
8. *"Number of Households",*

Out of above 8 properties we need to decide which is the actually useful and need to remove unnecessary properties.

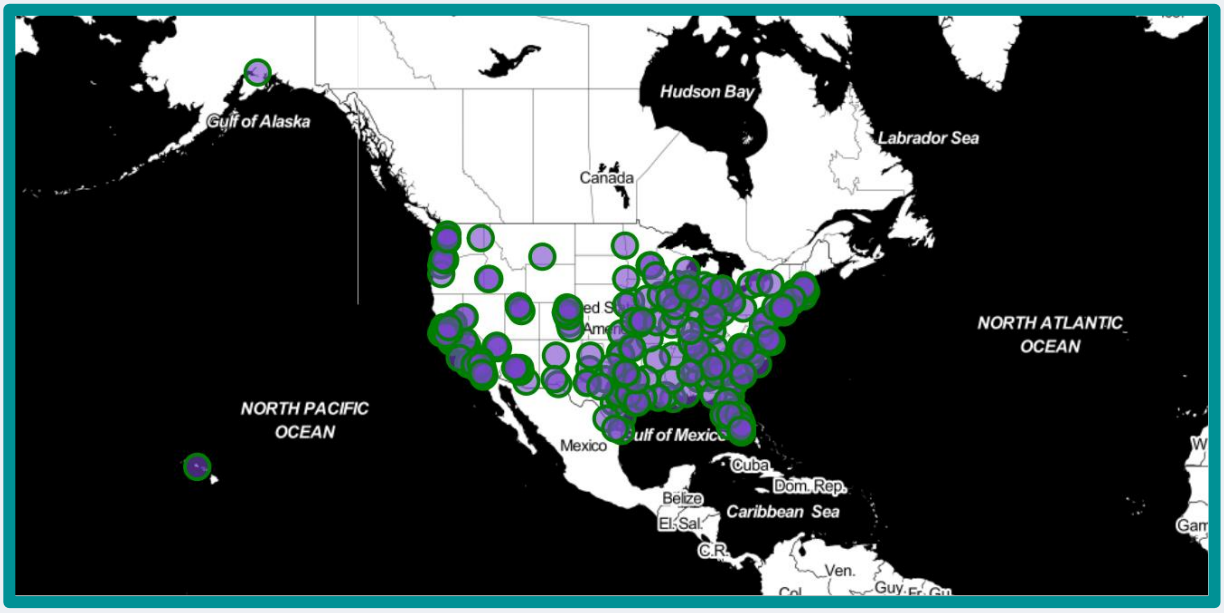
We need only ***County, State, Per Capita Income and Population*** *so, we will drop rest of the columns*

The data looks as below after removing unnecessary properties:

	Country-equivalent	State	Per capita income	Population
0	New York County	New York	$62,498	1,605,272
1	Arlington	Virginia	$62,018	214,861
2	Falls Church City	Virginia	$59,088	12,731
3	Marin	California	$56,791	254,643
4	Alexandria City	Virginia	$54,608	143,684

View Neighbors by plotting it on the map

As we have list of cities, lat, lng, and radius let's first plot the data on a map using Folium Library.



To get the best suitable place we will be using the Four Square API to get the details of the Venues in the cities and categories of Venues in the cities.

Add Data Using Four Square API To Find the Best Venues in The Cities.

Step – 1: Get the client ID, Secret Keys from the Four Square and add other details i.e. Versions, Radius and Limits.

Step – 2: Get the Category List from the Four Square API.

Step – 3: Get the Unique Category List from data.

Step – 4: Let's weight the prioritize categories.

Step – 5: Get the list of all venues as per the category priority.

Step – 6: Filter the venues and get the most relevant.

Step – 7: After getting the relevant venue list we need to get the mean values for each city type to find the most potential city.

Step – 8: After getting a potential cities we need to normalize the data using MinMaxScaler() from Sklearn library.

Step – 9: Get the best City Name & State Name.

Step – 10: Check city matching the per capital income criteria.

Step – 11: Get the lat, lng of the city and find the venues in the city using the FourSquare API.

Step – 12: Get the list of venues from in the city and set the weight to each category.

Step – 13: Plot the venues on Map

Step – 14: Using K Means algorithm to cluster the venues and calculating the weights for each cluster to decide which cluster would be the best area to have a new marketing office.

Step – 15: Get the clusters and group the list by cluster mean.

Step – 16: Get the lat, lng with the mean weight.

Step – 17: Plot the final map with the lat, lng of the best location.

RESULT

Based on the given constraints, a new marketing office should be in a place closer to the center of the circle to attract a greater number of diverse and potential customers to get huge revenue.

Conclusions

Based on the given constraints, a new marketing office should be in a place closer to the center of the circle to attract a greater number of diverse and potential customers to get huge revenue. We found California as a best city to have a New Marketing Office.

Thank You