RESTAURANT FOR WASHINGTON COUNTY

IBM APPLIED DATA SCIENCE CAPSTONE PROJECT

Author
Nibedita Dash MBA, CSM

01/25/2019

Nivea4444@yahoo.com

Business problem:

This business problem is set in our locality of Washington County in rural Pennsylvania. Since last few years a progressive County executive has succeeded in making the county attractive to investments through business-friendly policies. As a result, we are currently seeing people originally from the county who had immigrated to metropolitan places in search of employment return to establish businesses here. Roger Goodrich is such an entrepreneur that left the county 15 years ago to finish his culinary training and has been working as a chef in nearby metropolis of Pittsburgh. Now, after saving up capital he is looking to open a restaurant in the county. His question for us is to decide type of restaurant and the location that would assure him the best return on investment. He wants to create food of a type that people are already acquainted with so that he is not trying to create a novel culinary demand in a rural community which has been fairly constrained in its choices. Secondly, he wants to be situated at a hub for business where lots of shopping took place. As an experienced chef he knows that people like to eat after or before they shop. Thirdly, he recognizes the fact that people "go out" to eat at places that offer maximum choices in types of restaurants; that way they have something of interest available if they change minds. We presumed that a choice location couldn't be gauged just on the absence of competitors per capita because people mostly flock to food courts where everyone else is visiting. They also generally avoid standalone places in the middle of nowhere. The strategy would be to stay within the competition and try to best it by providing something unique. Something with which people were already acquainted with but just could not find enough around because of lack of choices in the locality. Hence, the goal for solving the problem seems to lie in finding places where both food and shopping options are most clustered. Subsequently, to find out which food choices are lacking in such a cluster; a kind of choice which exists in the county but not enough in the chosen cluster.

Data:

To solve this problem, we would need to obtain data dividing the county into scalable units. The county is divided into various cities. Given the sparsely populated nature of the rural county a 'city' was determined to be a small enough unit to base our classifications on. We shall obtain latitude and longitudinal location data from this data source and try to concatenate it to data obtained from a local search-and-discovery service app to find restaurants and shopping area clusters within the county.

Due to a government shut down in the united states; the census and other such locality data could not be obtained from data.gov. Although not ideal, an alternative presented itself upon discovery of https://simplemaps.com/data/us-zips. The company Simple Maps states that data has "been built from the ground up using authoritative sources including the U.S. Postal Service™, U.S. Census Bureau, National Weather Service, American Community Survey, and the IRS". This was deemed reliable and accurate for purposes of the project and their free version of data was obtained in .csv format from the website. It provided latitude and longitude information for all states of the US divided into counties and cities.

We shall obtain location and features data for local businesses by using FourSquare, a local search-and-discovery service app. This app distinguishes itself by providing a reliable and free data collection for developers easily accessed by API. An account was created, and API was accessed to obtain data regarding categories of restaurants and shopping places in the county of Washington and the frequency with which they are visited.

Methodology:

The csv file obtained from https://simplemaps.com/ was stored on a local hard drive and was then uploaded as an asset to IBM Watson. From there the dataset was loaded on to the Jupyter Notebook using the inbuilt 'insert to code' function. The resulting dataset obtained was inspected. Along with information about location coordinates of each city it had population density, zip code and time zone data. The population density data had a lot of null values and was overall not representative, so we decided not to use it. A clean data frame with only cities in the Washington county of Pennsylvania State with their location coordinates was extracted from the parent data frame. There were no missing values so we did not need much more data wrangling to prepare the data frame for analysis. First, Geopy library was used to obtain latitude and longitude of Washington County and cities in it were superimposed on map of Washington (Figure 1). The relative location of cities in the county were appreciated. There are two areas of concentration that stand out on visual inspection, once alongside highway I-70 in the Eastern part and the other alongside highway I-79 towards the more central and western part of the county. Our goal is to find clusters with maximum shopping and restaurant areas for the client to base his business in.

Next, Foursquare API was used to explore the neighborhoods and segment them. In this process the foursquare credentials and version created earlier were passed on to the URL. The venues of interest were restricted only to restaurants by adding the parameter 'section' as 'food'. A radius of a thousand meters limited to hundred locations was deemed to produce an appropriate sample for each city. We were able to obtain an output of two hundred and fifty restaurants spread over fifty of the eighty-one cities in the county.

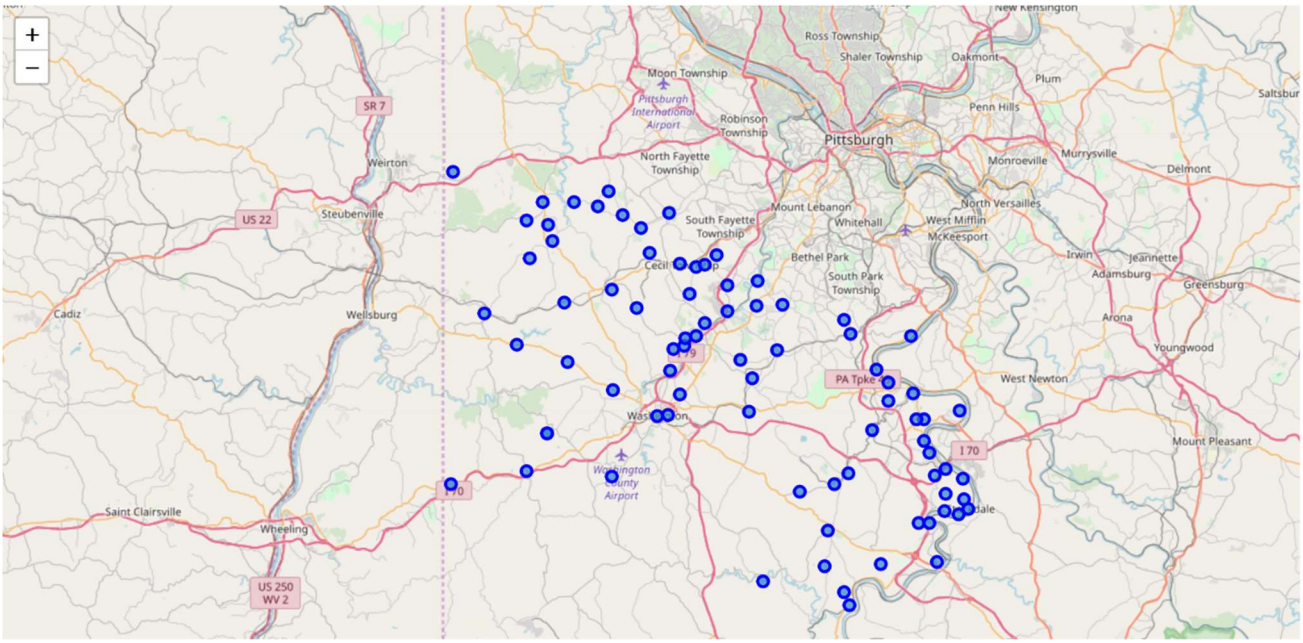

Figure 1: Map of Washington County with cities superimposed.

Thirty-one cities did not have any restaurants and were therefore dropped from the clustering process. We obtained a total of 34 unique varieties of restaurants in the whole county. Given our initial decision to restrict ourselves only to the types of restaurants that people in the county were used to, these 34 types shall form the set from which we shall eventually chose one. A dataset ‘food Group’ was created in the process containing all cities with at least one restaurant and the number of unique types of restaurants each had. A set of dummy variables for each type of restaurant was created and classified according to the city it was present in. Subsequently a list containing frequency of every type of restaurant in each city was created. This list was used to loop through all restaurants and sort them according to the least frequent type found in its city for all thirty-four types and create a data frame “citys_venues_sorted”.

It was decided to use unsupervised K means machine learning algorithm to cluster the cities according to similarity of restaurant concentration. The input was an unlabeled dataset with the varieties of restaurant types for each city with the city name removed. A total of 8 clusters were created after 350 iterations and clustering results were visualized on a map (Figure 2). Quick inspection of cluster data frames revealed the 5th cluster to have maximum number of restaurant dense cities. It was stored in a data frame "food_clustered". Prelimnary inspection of the dataset revelaed that generally sushi restaurants and steakhouses were lacking in most of the cities with high restaurant concentrations.

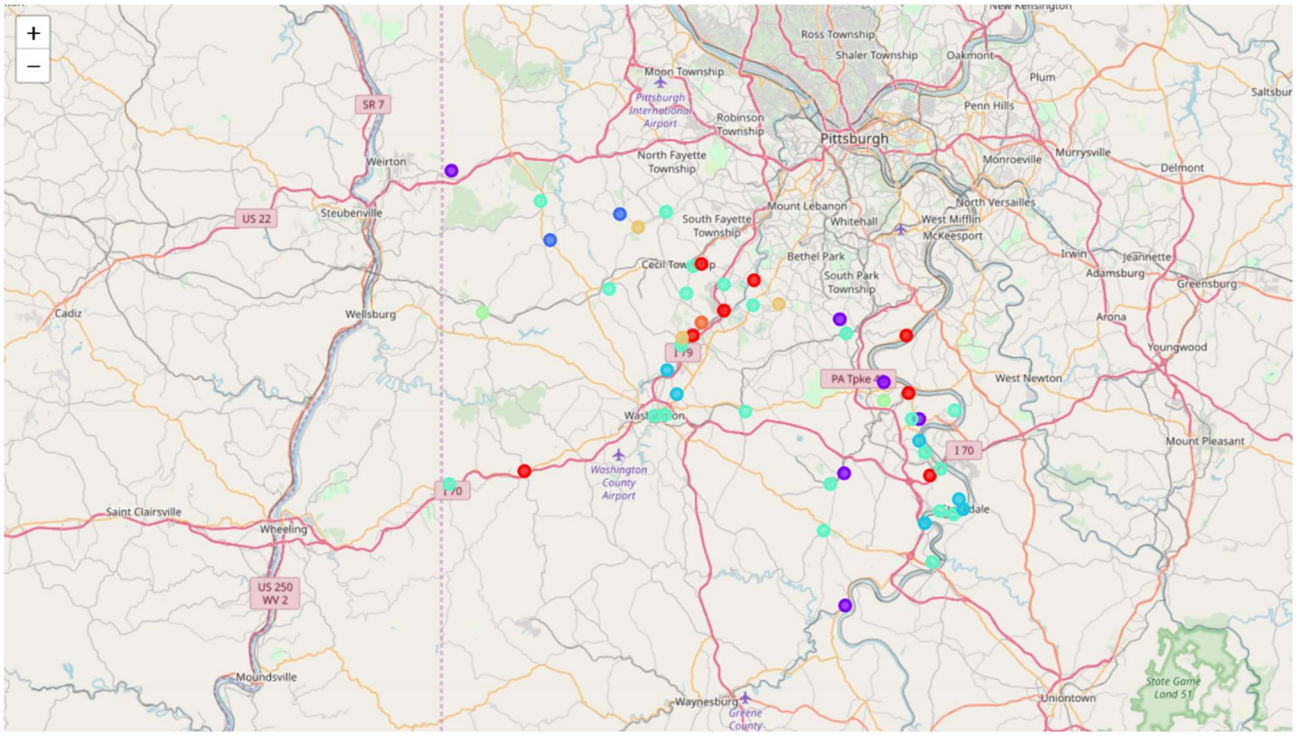

Figure 2: Map of Washington county with clusters of restaurant predominant cities superimposed.

Next, the Foursquare API was again used to explore shops in the Washington County by specifying section as shops in the URL. A radius of a thousand meters limited to hundred locations was deemed to produce an appropriate representative sample for each city. We were able to obtain an output of four hundred eighty-seven shops spread over seventy six of the eighty-one cities in the county. Five cities did not have any restaurants and were therefore dropped from the clustering process. We obtained a total of 76 unique varieties of restaurants in the whole county. A dataset 'shop_Group' was created in the process containing all cities with at least one shop and the number of unique types of restaurants each had. A set of dummy variables for each type of shops was created and classified according to the city it was present in. Subsequently a list containing ten most common of every type of shop in each city was created. This list was used to loop through all restaurants and sort them according to the most frequent type found in its city for all ten types and create a data frame "citys_shops_sorted". The decision to limit to 10 most common shops was taken because beyond the first few, isolated frequency of unique types were very often found to be zero.

It was decided to use unsupervised K means machine learning algorithm to cluster the cities according to similarity of shops concentration. The input was an unlabeled dataset with the varieties of shopping center types for each city with the city name removed. A total of 8 clusters were created after 350 iterations and clustering results were visualized on a map (Figure 3). Quick inspection of cluster data frames revealed the 1st cluster to have maximum number of shopping center dense cities. It was stored in a data frame "shops_clustered".

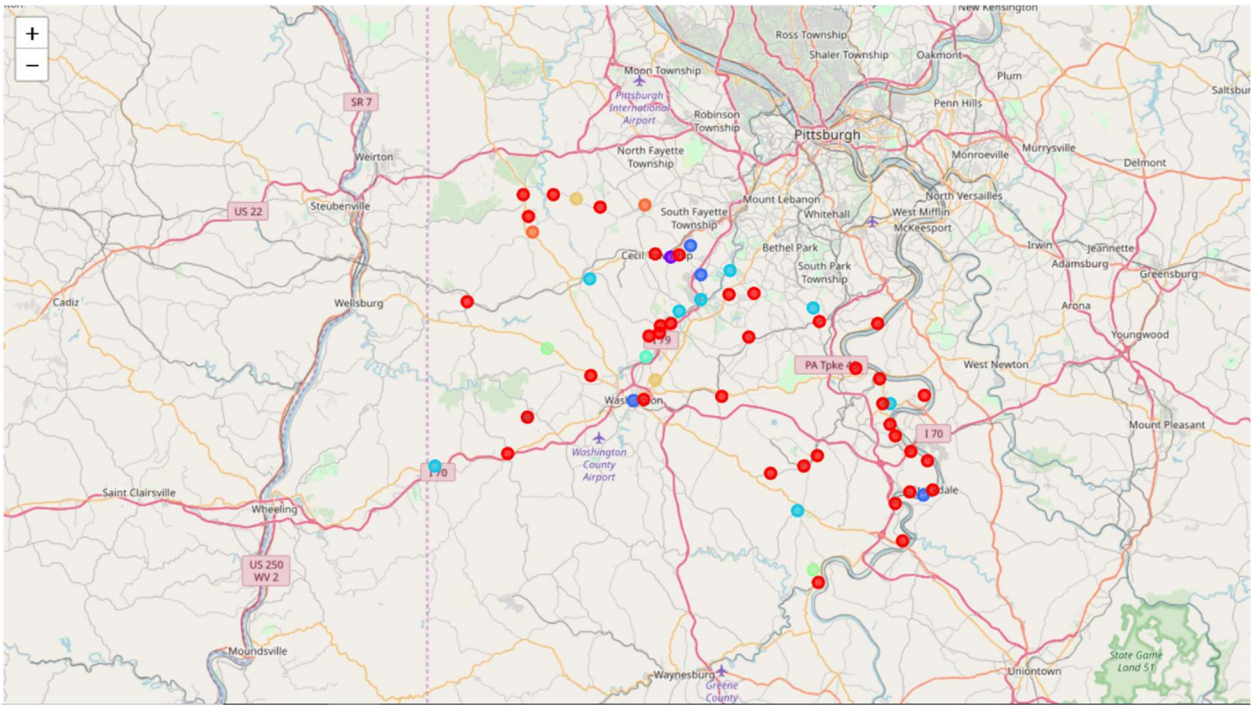

Figure 3: Map of Washington county with clusters of shopping center predominant cities superimposed.

Finally, a dataset of cities which were included in both the highest frequency clusters for restaurants and shopping centers was created by combining ‘food_clustered’ and ‘shops_clustered’ datasets. Cities that did not belong to both datasets were dropped. The number of unique restaurants and shopping centers was also added on by merging ‘shop_group’ and ‘food_group’ datasets. A new variable ‘Hip Places’ was created for each city by the sum of restaurants and shopping centers in that city. The resulting dataset ‘Washington_hip’ was cleaned to keep only the 3 least common restaurants in these cities and sorted according to the highest frequency of hip places in these cities.

Results:

During the process of preparing the data for analysis we obtained a total of 81 cities in the whole of Washington county of Pennsylvania. Out of these 31 cities did not have any restaurants and 5 cities did not have any shopping centers and were therefore excluded from analysis. The remaining cities had 250 restaurants of 34 types and 487 shopping centers of 76 types. For sake of simplicity we kept all the restaurant types but removed all but the 10 most common shopping centers from each city for our analysis.

During the whole process of analysis, we ran two k means clustering unsupervised algorithms with a maximal number of iterations of 350 to cluster both restaurants and shopping centers in the County. The cities that came in the intersection of highest frequency clusters of both shopping centers and restaurants are hypothesized to have the maximum return on investment. We got a resulting basket of 11 cities (Table 1). Out of these 11, Thompsonville had the largest number of shopping centers and restaurants at 55 besting the nearest rival city of East Washington by 40. Our analysis also resulted that the least common restaurant in Thompsonville was Fried Chicken Joint.

	1st Least Common Venue	2nd Least Common Venue	3rd Least Common Venue	Hip Places
Thompsonville	Fried Chicken Joint	Sushi Restaurant	Steakhouse	55.0
East Washington	Wings Joint	Korean Restaurant	Italian Restaurant	40.0
Charleroi	American Restaurant	Sushi Restaurant	Steakhouse	33.0
Burgettstown	Fried Chicken Joint	Sushi Restaurant	Steakhouse	19.0
Moninger	Fried Chicken Joint	Sushi Restaurant	Steakhouse	16.0
Finleyville	American Restaurant	Sushi Restaurant	Steakhouse	15.0
Donora	American Restaurant	Sushi Restaurant	Steakhouse	12.0
West Brownsville	American Restaurant	Sushi Restaurant	Steakhouse	8.0
Wickerham Manor	American Restaurant	Sushi Restaurant	Steakhouse	7.0
Speers	Fried Chicken Joint	Sushi Restaurant	Steakhouse	6.0
Eighty Four	American Restaurant	Sushi Restaurant	Steakhouse	5.0
Ellsworth	Fried Chicken Joint	Sushi Restaurant	Steakhouse	5.0
Elco	American Restaurant	Sushi Restaurant	Steakhouse	2.0

Table 1: Cities at the intersection of food and shopping clusters sorted in decreasing order of most number of unique venues.

Discussion:

The project sets out to find the most profitable kind of food business and the best place in the county to set it up at. We started out with a few constraints which were well defined by common sense principles and client request. It was assumed that we were to restrict our advice about the type of restaurant to the types that people of the county had exposure to. Being the first to enter the market as a unique product had its own disadvantages. Working in a rural area with mostly indigenous population increases the importance of launching business that people were familiar with. The rural area had vast stretches of very sparsely populated areas. As we found from the analysis, about 31 cities did not have a food outlet and 6 had no shopping areas operating out of them. This is a very significant finding which has a unique consideration for food type businesses.

We cannot simply locate a restaurant by choosing a city which doesn't have restaurants. This might be contradictory to the principles of demand and supply and may seem counterintuitive because of lack of competition in these places. However, a knowledge of food habits of people will quickly clarify this. People usually aggregate at places where they find other people and shopping centers. This is the rationale behind food courts in malls. Shopping and eating go together. Many places were left without any food outlets because they probably did not get enough business to self-sustain and therefore it would be very risky going into such unchartered territory. Also, the fact that people aggregate at places that gives them the maximum choice as far as food alternatives are concerned was considered. Therefore, it was decided to restrict our location preferences to the most crowded places with the highest number of shopping and restaurant alternatives. Next was the question that if we were to situate the business right in the middle of the place with the highest competition possible, how were we to distinguish ourselves from the rest of them. The answer lied in being able to provide a kind of food alternative that was lacking in such an area.

Our analysis yielded 34 unique types of restaurants that people from the county were acquainted with. As per assumptions, we tried to select one amongst these 34 types. Examination of intersection between the cluster with highest frequency food alternatives and the cluster with highest frequency shopping alternatives was done to narrow down the search to one most favorable city. This city of Thompsonville has a very low frequency of fried chicken joints along with sushi restaurants and steakhouses.

It was generally understood that we did not consider other factors that can influence the preference of location for a food business. Notably, it would have been ideal to have demographics, employment and median income data for the top few places considered. It is well known that an older and wealthier demography gravitates towards steakhouses whereas a younger and wealthier

demography gravitates towards sushi restaurants. Also, a less wealthy demography would mostly choose fast food like fried chicken. Using a supervised k nearest neighbor classification ML algorithm would have been ideal in the presence of such data. Our dataset did have some information about the median income and median age, but it was sparse with lots of missing values on examination for which we decided not to use it. We did attempt to obtain government census data from data.gov for analysis but were unsuccessful because of the government shutdown at the time of the project completion.

Conclusion:

As a result of our analysis we recommended Roger Goodrich to start a fried chicken joint in the city of Thompsonville.