Predict Grocery Demand for Mobile Grocery Store based on Cuisine Type for Franklin County Neighborhoods in OH

Jagadeesh Sreeram

Dec 27th 2019

1. Introduction

1.1 Background

Raj my friend wants to start a Asian mobile grocery store that servers most of the Neighborhoods in Franklin County. His main target is restaurants in these neighborhoods. He procures and sells mostly 5 type of Asian grocery I.e Indian, Chinese, Thai, Japanese and Korean. He will have just one warehouse in Franklin county from where in the morning all his mobile store trucks for most of these neighborhoods will leave. So, he wants to know if a mobile grocery truck is leaving for a particular neighborhood what percent of each of above 5 type Asian grocery need to be carried in the mobile grocer store. His main target is Asian restaurants of above 5 types in these neighborhoods. He also doesn’t want the percent demand for each neighborhood at this point , but just want to keep to 5 groups.

1.2 Problem

Predicting Grocery demand based on restaurant concentration in ~25 mile radius of each neighborhood in Franklin county OH. Also minimize the category of grocery percentage groups to 5.

Assumptions: Due to non-availability of actual grocery demand data from each of these restaurants, we had to use only the number of similar category restaurants in the vicinity for the demand calculation.

1.3 Interest

Raj is very interested to know what type of Asian restaurants exists in his county and how much percentage quantity he need to carry in his mobile store trucks. Even though currently Raj is targeting restaurants in the neighborhood, later this also can be extended by taking different kind of population type in those areas, so that Raj can target residential areas too.

2. Data acquisition and cleaning

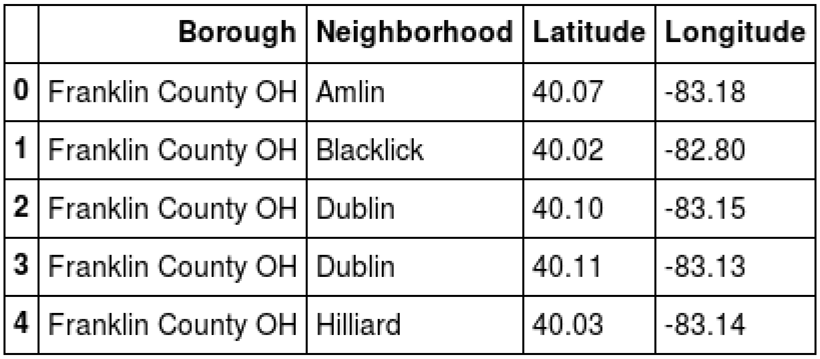
2.1 Data Source

The Franklin County neighborhood data is obtained from US Postal service website https://www.unitedstateszipcodes.org/zip-code-database/ . There are 67 neighborhoods in Franklin County. For restaurant data we used Foursquare venue search API with category filters to search restaurants in these neighborhood that are Indian, Chinese, Thai, Japanense and Korean. For each neighborhood we get above category restaurants within 25 mile radius
Foursquare API Ref :
https://developer.foursquare.com/docs/api/endpoints

2.2 Data Cleaning

The neighborhood information was readily available from US Postal website in csv format with all the required information for this project. We had to simple filter for Franklin county and upload the data in the project area. Below is snippet of Franklin county neighborhood data

Snapshot of just top 5:



	Borough	Neighborhood	Latitude	Longitude
0	Franklin County OH	Amlin	40.07	-83.18
1	Franklin County OH	Blacklick	40.02	-82.80
2	Franklin County OH	Dublin	40.10	-83.15
3	Franklin County OH	Dublin	40.11	-83.13
4	Franklin County OH	Hilliard	40.03	-83.14

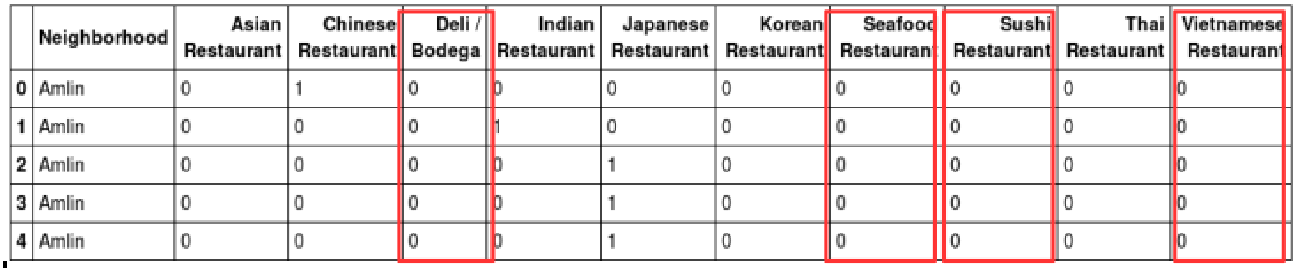
There are actually 67 neighborhoods in Franklin county and we use all for this project.

The category filter for the 5 type of restaurant types are obtained form Foursquare category doc and used in the project:
https://developer.foursquare.com/docs/resources/categories

2.3 Feature selection

Even though we provided the category for search API, foursquare gives restaurants for other types, since this data was very minimal and doesn’t impact overall model, we removed that data from the model as part of data cleanup. The data that was removed is for category types “Seafood Restaurant,Sushi Restaurant,Vietnamese Restaurant & Deli / Bodega” . Since the dataset was very low for these and these causes issue for our model we removed this data. Since Foursquare only gives ~100 venues for

each neighborhood, we had ~ 6700 restaurant information overall for the model.
Below are the highlighted categories that were removed.



	Neighborhood	Asian Restaurant	Chinese Restaurant	Deli / Bodega	Indian Restaurant	Japanese Restaurant	Korean Restaurant	Seafood Restaurant	Sushi Restaurant	Thai Restaurant	Vietnamese Restaurant
0	Amlin	0	1	0	0	0	0	0	0	0	0
1	Amlin	0	0	0	1	0	0	0	0	0	0
2	Amlin	0	0	0	0	1	0	0	0	0	0
3	Amlin	0	0	0	0	1	0	0	0	0	0
4	Amlin	0	0	0	0	1	0	0	0	0	0

The using neighborhood location information we retrieve the restaurant information from Foursquare for ~25 mile radius of each nighborhood and then compute what type of restaurants are more predominant in each neighborhood and rank them . And later using k-means approach , we categorize them into 5 categories and compute the average grocery demand for each of these 5 grocery type .

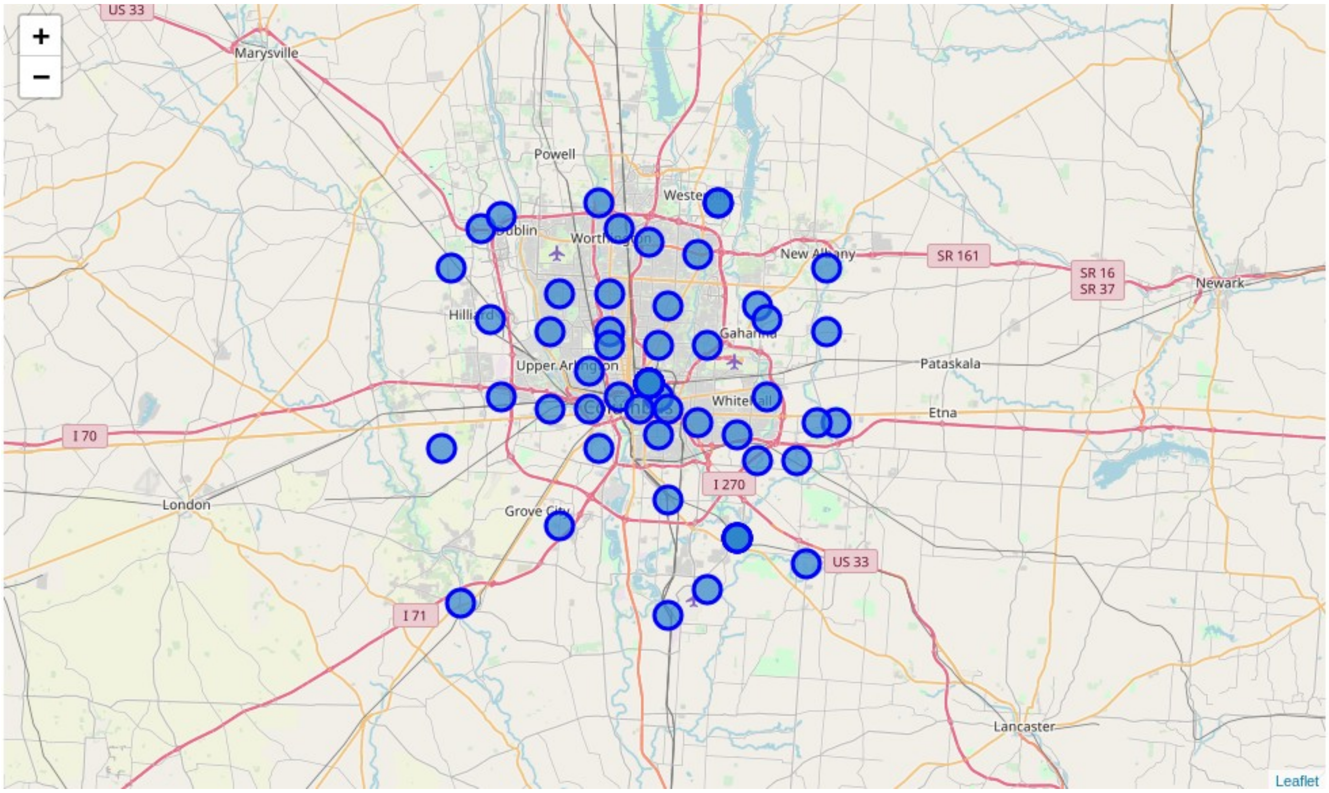
3. Methodology

The methodologies used for this project are getting restaurants for 5 categories with in 25 mile radius and ranking the neighborhood for each specific restaurant type. And the using K-mean methodology derive 5 clusters. These 5 clusters provide five variations of grocery demand distribution based on number of restaurants of a specific type in that neighborhood.

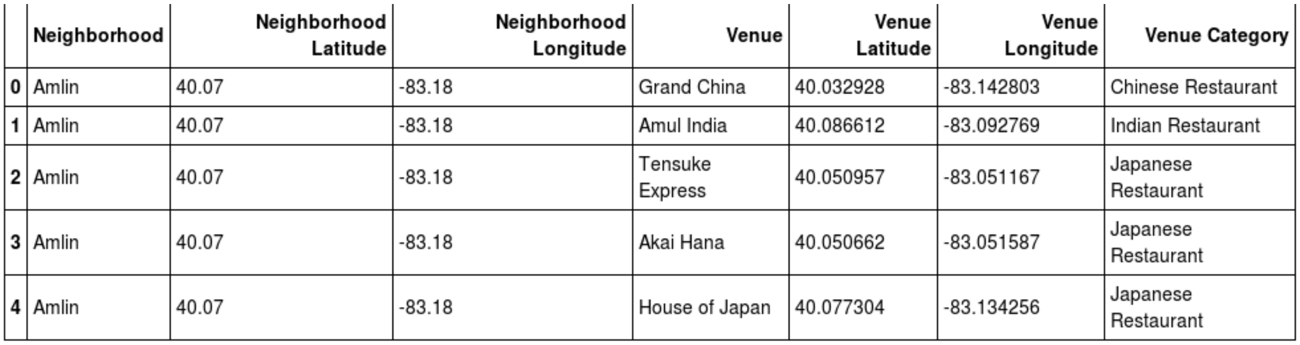
Due to non-availability of actual demand from each restaurant , we had to just rely on restaurant type for actual demand calculation. This is the main assumption we made.

3.1 Exploratory data analysis

Initially to analyze the data , I plotted a geo map of Franklin County and superimposed the postal data. This shows the data obtained from Us postal is valid and no outliers.



Also used one of the neighborhood and retrieved the restaurant information and validated that restaurants of preferred category are retrieved. There were few outliers we found and removed from the for more accurate model.



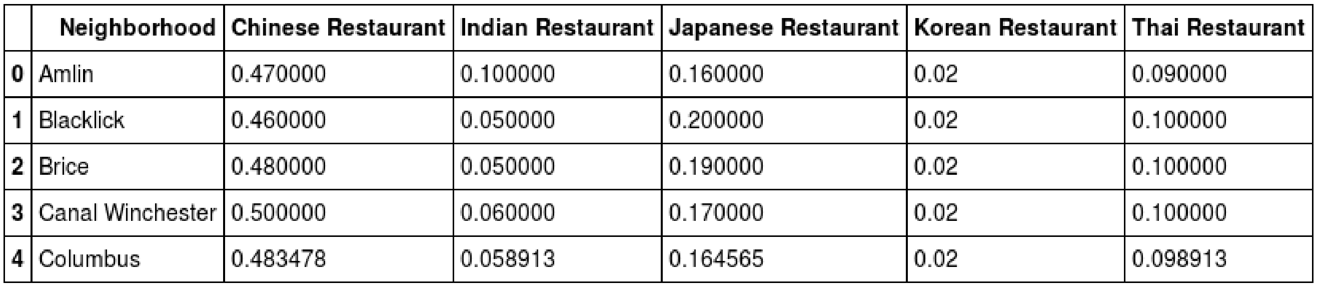
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Amlin	40.07	-83.18	Grand China	40.032928	-83.142803	Chinese Restaurant
1	Amlin	40.07	-83.18	Amul India	40.086612	-83.092769	Indian Restaurant
2	Amlin	40.07	-83.18	Tensuke Express	40.050957	-83.051167	Japanese Restaurant
3	Amlin	40.07	-83.18	Akai Hana	40.050662	-83.051587	Japanese Restaurant
4	Amlin	40.07	-83.18	House of Japan	40.077304	-83.134256	Japanese Restaurant

Relationship between Restaurant Type and Grocery Demand:

Given the data only relation we were able to draw between Restaurant type and grocer demand is if more the type of restaurant then more demand for that type of grocery.

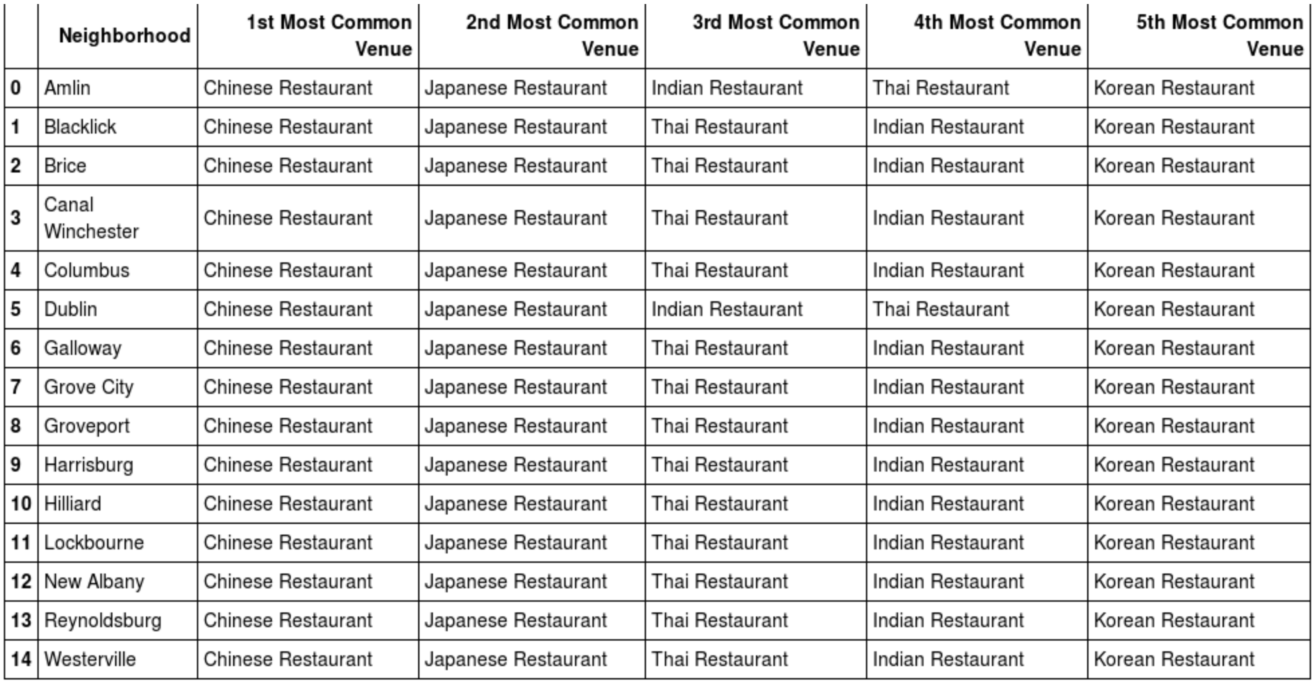
4. Predictive Modeling

Calculated simple mean for neighborhood for different type of restaurant in the list. This gave the demand for grocery type . Again the assumption made was that restaurant count is directly related to demand. And we have only calculated the weighted avg here.



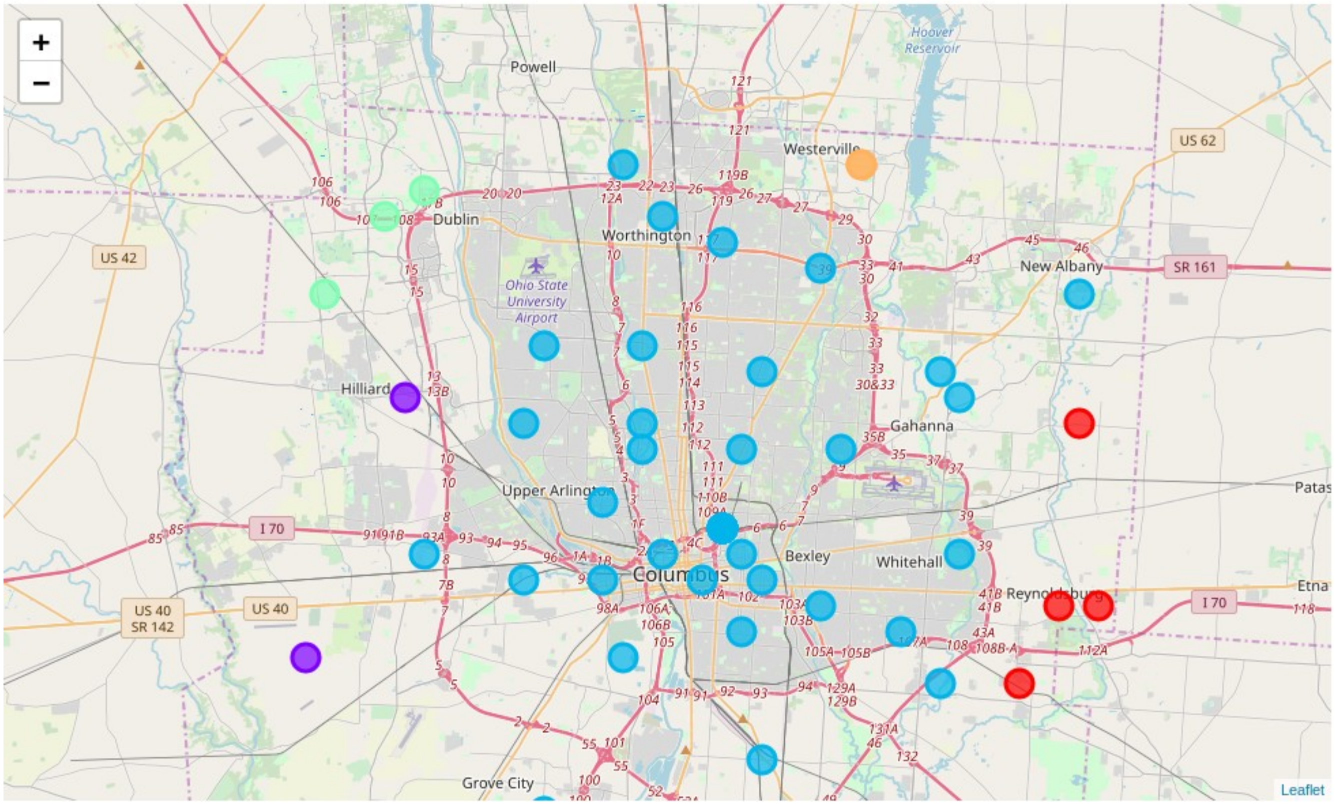
	Neighborhood	Chinese Restaurant	Indian Restaurant	Japanese Restaurant	Korean Restaurant	Thai Restaurant
0	Amlin	0.470000	0.100000	0.160000	0.02	0.090000
1	Blacklick	0.460000	0.050000	0.200000	0.02	0.100000
2	Brice	0.480000	0.050000	0.190000	0.02	0.100000
3	Canal Winchester	0.500000	0.060000	0.170000	0.02	0.100000
4	Columbus	0.483478	0.058913	0.164565	0.02	0.098913

After deriving this weighted average for each of this neighborhood we used k-means clustering approach to derive 5 clusters that client wanted. These clusters represent 5 groups of grocery distribution for 5 category types and represent which neighborhood belongs to which group. This helped Raj to load up his mobile store truck with right kind of grocery proportions which truck is leaving for the neighborhood.

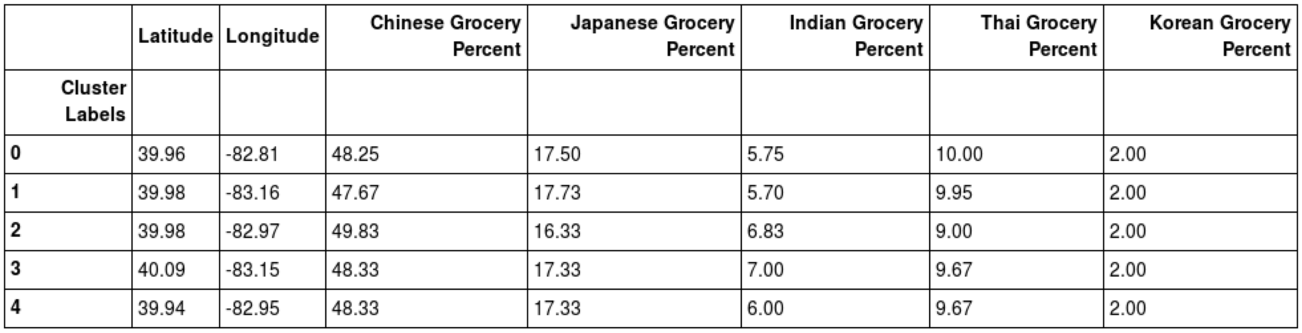


	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Amlin	Chinese Restaurant	Japanese Restaurant	Indian Restaurant	Thai Restaurant	Korean Restaurant
1	Blacklick	Chinese Restaurant	Japanese Restaurant	Thai Restaurant	Indian Restaurant	Korean Restaurant
2	Brice	Chinese Restaurant	Japanese Restaurant	Thai Restaurant	Indian Restaurant	Korean Restaurant
3	Canal Winchester	Chinese Restaurant	Japanese Restaurant	Thai Restaurant	Indian Restaurant	Korean Restaurant
4	Columbus	Chinese Restaurant	Japanese Restaurant	Thai Restaurant	Indian Restaurant	Korean Restaurant
5	Dublin	Chinese Restaurant	Japanese Restaurant	Indian Restaurant	Thai Restaurant	Korean Restaurant
6	Galloway	Chinese Restaurant	Japanese Restaurant	Thai Restaurant	Indian Restaurant	Korean Restaurant
7	Grove City	Chinese Restaurant	Japanese Restaurant	Thai Restaurant	Indian Restaurant	Korean Restaurant
8	Groveport	Chinese Restaurant	Japanese Restaurant	Thai Restaurant	Indian Restaurant	Korean Restaurant
9	Harrisburg	Chinese Restaurant	Japanese Restaurant	Thai Restaurant	Indian Restaurant	Korean Restaurant
10	Hilliard	Chinese Restaurant	Japanese Restaurant	Thai Restaurant	Indian Restaurant	Korean Restaurant
11	Lockbourne	Chinese Restaurant	Japanese Restaurant	Thai Restaurant	Indian Restaurant	Korean Restaurant
12	New Albany	Chinese Restaurant	Japanese Restaurant	Thai Restaurant	Indian Restaurant	Korean Restaurant
13	Reynoldsburg	Chinese Restaurant	Japanese Restaurant	Thai Restaurant	Indian Restaurant	Korean Restaurant
14	Westerville	Chinese Restaurant	Japanese Restaurant	Thai Restaurant	Indian Restaurant	Korean Restaurant

Below is the result from K-means clustering algorithm classifying the neighborhoods into 5 categories of same grocery demand .



Finally we calculated percent of each cluster for each grocery type.



	Latitude	Longitude	Chinese Grocery Percent	Japanese Grocery Percent	Indian Grocery Percent	Thai Grocery Percent	Korean Grocery Percent
Cluster Labels							
0	39.96	-82.81	48.25	17.50	5.75	10.00	2.00
1	39.98	-83.16	47.67	17.73	5.70	9.95	2.00
2	39.98	-82.97	49.83	16.33	6.83	9.00	2.00
3	40.09	-83.15	48.33	17.33	7.00	9.67	2.00
4	39.94	-82.95	48.33	17.33	6.00	9.67	2.00

Above table provided with information about what percent of each type of grocery need to be loaded to each mobile truck . So, that the demand is met correctly.

5. Conclusions

Using above approaches and based on data available we were able to predict grocery demand based of restaurant type in neighborhood I.e 1st place is Chinese grocery with 50% , then in 2nd place Japanese with ~17% , 3rd Thai with ~10 % and 4th Indian with ~ 6 – 7% and finally 5th Korean which is pretty constant at 2%.

We are able to conclude that demand varies at the ends of Franklin country , but at the central part of the county it remains pretty evenly distributed.

6. Future directions

Currently the model only considers the number of restaurants in the vicinity. To make the model more accurate we should include other grocery demand factors like season , occupancy information of each restaurant, menu and its ingredients etc.