Analysis of Restaurants in Los Angeles and their Online Delivery System

1. Introduction

1.1. Background

Los Angeles officially the City of Los Angeles and often known by its initials L.A., is the most populous city in California; the second most populous city in the United States, after New York City; and the third-most populous city in North America, after Mexico City and New York City. With an estimated population of nearly four million people,[11] Los Angeles is the cultural, financial, and commercial center of Southern California. The city is known for its Mediterranean climate, ethnic diversity, Hollywood, the entertainment industry, and its sprawling metropolis.Los Angeles lies in a basin, adjacent to the Pacific Ocean, with mountains as high as 10,000 feet (3,000 m), and deserts. The city, which covers about 469 square miles is the seat of Los Angeles County, the most populous county in the United States. The Los Angeles metropolitan area (MSA) is the second-largest metropolitan area in the nation with a population of 13.1 million people. The Greater Los Angeles metropolitan area (CSA) is the second-most populous CSA metropolitan area with a 2015 estimate of 18.7 million people. Los Angeles has a diverse economy and hosts businesses in a broad range of professional and cultural fields. It is also famous for its movie, television, and recording industry. A global city, it has been ranked 6th in the Global Cities Index and 9th in the Global Economic Power Index. The Los Angeles metropolitan area also has a gross metropolitan product of $1.0 trillion (as of 2017), making it the third-largest city by GDP in the world, after the Tokyo and New York City metropolitan areas. Los Angeles hosted the 1932 and 1984 Summer Olympics and will host the 2028 Summer Olympics.

1.2. Problem

- To analyze all restaurant venues and extract their major delivery providers if any, to get information on major competitor and at the same time to focus on those restaurants which don't have any delivery providers and might be potential customers for any new delivery providers in future
- To analyze and categorize different cuisines which might provide valuable input for online delivery providers and to new startups planning to open new restaurants

1.3. Interest

- The project will provide in depth information of various restaurants and major competitors in the area.
- It will also be beneficial to startups who want to start any new restaurants in the area to get information regarding which major cuisines are there in the area. It could give valuable input on which cuisines are not present in a particular neighborhood and starting those cuisines in the areas might prove highly profitable as Los Angeles is a city of diversity and people tend to look for diverse cuisines.

2. Data Source, Acquisition and cleaning

2.1. Data Source

Geo spatial data of Los Angeles county was provided by datasets of LA county in this link Geo Spatial datasets of LA. From this link I took the latest Geojson file corresponding to neighborhoods of LA.

I have also referred to https://en.wikipedia.org/wiki/Los_Angeles#Demographics to analyze economy, cultural diversity, population distribution within LA. Since LA county has a large area distribution, I have used the above information to further filter and restrict the data to certain neighborhoods within central downtown region for my project.

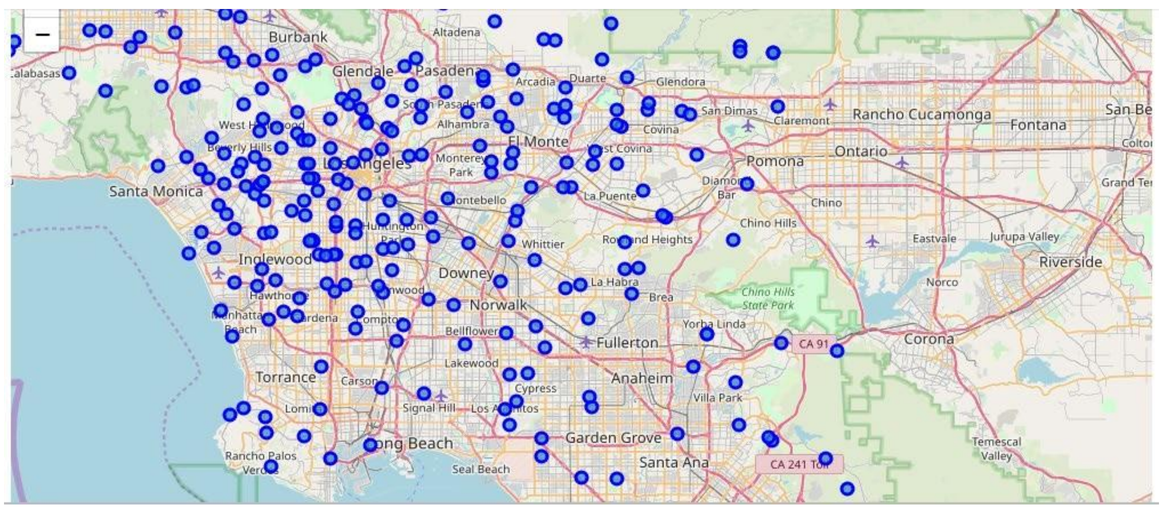
2.2. Data Acquisition and Cleaning

Using JSON library of Python, I have extracted the GeoJson file. Out of all the location features, regions of LA county, their corresponding Neighborhoods and their coordinates were extracted.

The coordinates given in the JSON file were for every Square Kilometer [km^2] of neighborhood. Hence, I have taken the mean Latitude and Longitude for every neighborhood while converting it to a data frame. Hence the initial data frame consists of regions, neighborhood and coordinate information. Sample records of dataframe extracted is as below:

	Region	Neighborhood	Latitude	Longitude
0	antelope-valley	Acton	34.539023	-118.207034
1	south-la	Adams-Normandie	34.037396	-118.308002
2	santa-monica-mountains	Agoura Hills	34.168157	-118.776212
3	northwest-county	Agua Dulce	34.488109	-118.378224
4	san-gabriel-valley	Alhambra	34.10504	-118.121747

I used the Folium map to create a map of LA county. For this I took coordinates using Geopy Geolocator library. Folium map was created to show different regions and neighborhoods of LA county. From this map I could visualize that downtown region falls within central LA region of dataset extracted. Hence, only central-LA region was selected for all future analysis and processing.



The subset dataframe corresponding to central-la region is shown below. It has 26 neighborhoods in total.

	Region	Neighborhood	Latitude	Longitude
0	central-la	Arlington Heights	34.052504	-118.31672
1	central-la	Beverly Grove	34.085755	-118.37249
2	central-la	Carthay	34.051684	-118.36664
3	central-la	Chinatown	34.067573	-118.22472
4	central-la	Downtown	34.061247	-118.24462
5	central-la	East Hollywood	34.098356	-118.28739
6	central-la	Echo Park	34.095686	-118.24486
7	central-la	Elysian Park	34.093903	-118.24245
8	central-la	Elysian Valley	34.108424	-118.24973
9	central-la	Fairfax	34.088949	-118.35112
10	central-la	Griffith Park	34.15884	-118.30434
11	central-la	Hancock Park	34.083488	-118.3266
12	central-la	Harvard Heights	34.05282	-118.31329
13	central-la	Hollywood Hills	34.152118	-118.31782
14	central-la	Hollywood Hills West	34.1292	-118.36951
15	central-la	Hollywood	34.105244	-118.32686
16	central-la	Koreatown	34.068982	-118.2868
17	central-la	Larchmont	34.076204	-118.32062
18	central-la	Los Feliz	34.118408	-118.27321
19	central-la	Mid-City	34.034962	-118.36859
20	central-la	Mid-Wilshire	34.068991	-118.34619
21	central-la	Pico-Union	34.052497	-118.28145
22	central-la	Silver Lake	34.112812	-118.26483
23	central-la	West Hollywood	34.097724	-118.36817

24	central-la	Westlake	34.053889	-118.25907
25	central-la	Windsor Square	34.076231	-118.31359

This dataframe was further filtered and foursqaure API was used to get restaurant venues, cuisine information and delivery providers. The final dataframe was then categorized, clustered and modelled. In depth analysis is explained in methodology section.

Final filtered dataframe with selected neighborhoods of downtown LA

	Region	Neighborhood	Latitude	Longitude
0	central-la	Chinatown	34.06757	-118.22472
1	central-la	Downtown	34.06125	-118.24462
2	central-la	Pico-Union	34.0525	-118.28145
3	central-la	Westlake	34.05389	-118.25907

Foursquare location API was used to process the above neighborhoods to get various restaurant venues. The obtained response was processed to get delivery provider information, cuisine category. Processing details have been explained in methodology section. Sample of processed dataframe is shown below:

	Neighborhood	Latitude	Longitude	Venueid	venuename	venueLatitude	venueLongitude	venuecategory	deliveryoption
0	Arlington Heights	34.0525	-118.32	52c77412498ee9acc1e90889	Liwan Restaurant & Hookah Lounge	34.05035	-118.437779	Mediterranean Restaurant	grubhub
1	Arlington Heights	34.0525	-118.32	4c00d9439cf52d7f636714e7	Hae Jang Chon Korean BBQ Restaurant	34.06389	-118.306075	Korean Restaurant	No Delivery option
2	Arlington Heights	34.0525	-118.32	4aff640ef964a520173822e3	Young Kyung Restaurant	34.05246	-118.30332	Chinese Restaurant	No Delivery option
3	Arlington Heights	34.0525	-118.32	4c7435bd2db5236a2eadb979	All Family Restaurant	34.05367	-118.323318	Korean Restaurant	No Delivery option

4	Arlington Heights	34.0525	-118.32	4ade63fcf964a5208f7521e3	Chung Kiwa Restaurant	34.05274	-118.31613	Korean Restaurant	No Delivery option
5	Arlington Heights	34.0525	-118.32	4b4501a3f964a5200c0126e3	Mi Rak Korean Restaurant 미락	34.04977	-118.308447	Korean Restaurant	No Delivery option
6	Arlington Heights	34.0525	-118.32	4c7177e1344437042a07275f	Nuevo Santaneco Restaurant	34.04693	-118.308123	Latin American Restaurant	No Delivery option
7	Arlington Heights	34.0525	-118.32	4c02cb6b0d0e0f47bdaa019a	V.I.P. CHINESE RESTAURANT국빈반점	34.05946	-118.309463	Chinese Restaurant	No Delivery option
8	Arlington Heights	34.0525	-118.32	4c748327db52b1f7ac7676dc	Olocuita Restaurant	34.06363	-118.309474	Latin American Restaurant	No Delivery option
9	Arlington Heights	34.0525	-118.32	4ada5935f964a520aa2121e3	El Cholo Restaurant	34.05018	-118.309141	Mexican Restaurant	grubhub
10	Arlington Heights	34.0525	-118.32	4b2940c7f964a520c29b24e3	Restaurant Namsan (남산)	34.06376	-118.300827	Korean Restaurant	grubhub

3. Methodology

The dataframe shown above consists of all restaurant venues corresponding to downtown LA which comprises of 4 neighborhoods. Once the raw dataframe is obtained its prepossessed and analyzed before modeling it. K-means approach is used for clustering.

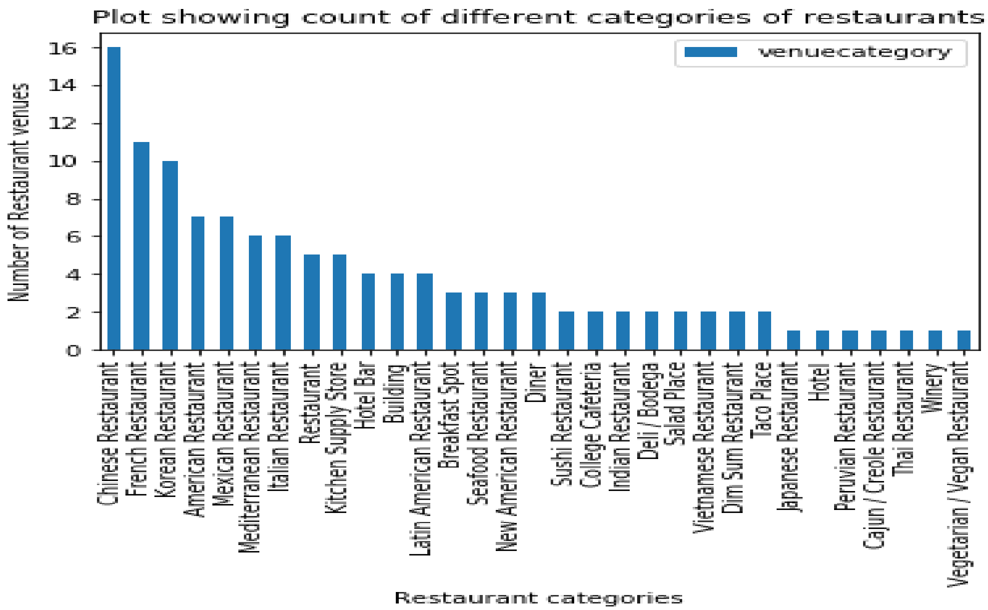
3.1. Preprocessing

a) The Neighborhood columns and venue ID, venue name, venue category are checked for Null values and corresponding rows are dropped if any

b) Since categorical variables can't be used for K means Clustering, they are converted to numeric using one-hot encoding approach. Venuecategory and Deliveryoption columns are encoded and merged with Venueid to get Restaurant_category dataframe

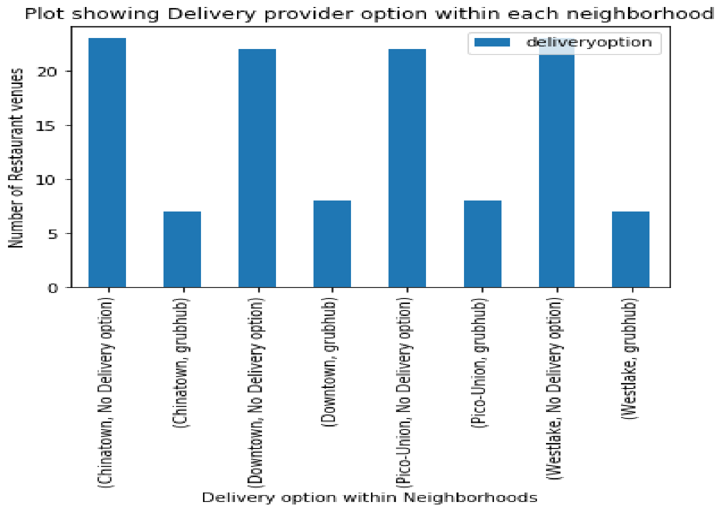
3.2. Exploratory Data Analysis

Restaurant Categories are grouped and analyzed with respect to total count of venues. A Bar chart showing the relationship is plotted.



We see that major categories are Chinese, Korean, French, Mexican and American. Various categories towards the lower end of graph are grouped into miscellaneous category.

Delivery Provider option within each neighborhood is also grouped and analyzed. A bar chart corresponding to the same is plotted.



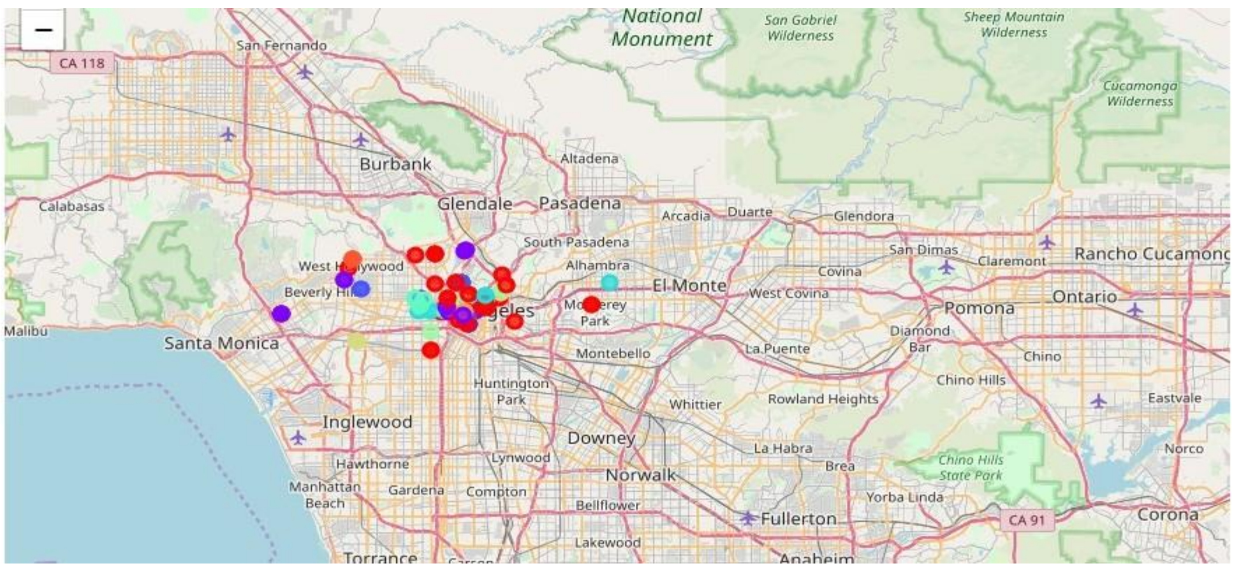
We see that within all neighborhoods, most of the restaurants don't have a major delivery provider.

3.3. Clustering of Restaurants

K-Means clustering is used due to its simplicity and accuracy. K value is chosen as 10 and clustered into 10 clusters. Cluster labels are generated for every Restaurant venue. A final merged restaurant data frame is created consisting of downtown LA restaurants merged with cluster labels for every restaurant. Sample set of dataframe is shown below:

	Neighborhood	Latitude	Longitude	Venueid	venuename	venueLatitude	venueLongitude	venuecategory	deliveryoption	labels
0	Chinatown	34.067573	-118.224722	49ebc74af964a5202b671fe3	Yang Chow Restaurant	34.062926	-118.238059	Chinese Restaurant	No Delivery option	4
1	Chinatown	34.067573	-118.224722	4415b1c5f964a520fd301fe3	Full House Seafood Restaurant	34.066155	-118.237916	Chinese Restaurant	No Delivery option	4
2	Chinatown	34.067573	-118.224722	52c77412498ee9acc1e90889	Liwan Restaurant & Hookah Lounge	34.050354	-118.437779	Mediterranean Restaurant	grubhub	1
3	Chinatown	34.067573	-118.224722	49c2818cf964a520f9551fe3	Boda Restaurant	34.075601	-118.217365	Vietnamese Restaurant	No Delivery option	0
4	Chinatown	34.067573	-118.224722	4aa8416bf964a5206c5020e3	Traxx Restaurant	34.056232	-118.236183	New American Restaurant	No Delivery option	0

All the clusters are vizualized using Folium map. Different clusters are shown with varying colors in map.



4. Results

After Clustering below are the properties of all clusters:

CLUSTER A – It consists of miscellaneous restaurants with no delivery option.

CLUSTER B – It consists of miscellaneous restaurants with grubhub as major delivery provider.

CLUSTER C and D – It consists of American and New American category of restaurants with no delivery option.

CLUSTER E – It consists of Mexican restaurants with no delivery provider.

CLUSTER F – It consists of Mediterranean Restaurants with grubhub as their major delivery provider.

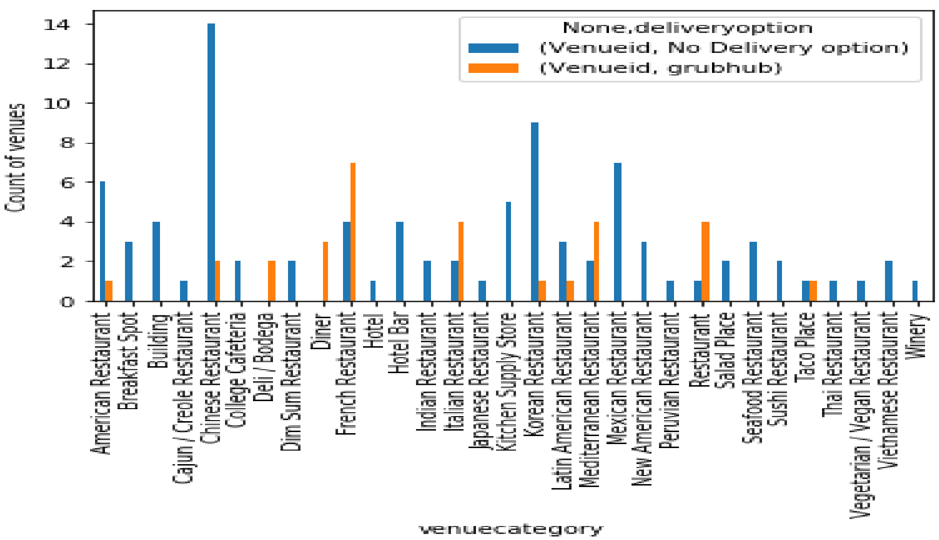
CLUSTER G – It consists of Korean restaurants with no delivery provider.

CLUSTER H – It consists of Chinese restaurants with no major delivery provider.

CLUSTER I – It consists of French restaurants with grubhub as their major delivery provider.

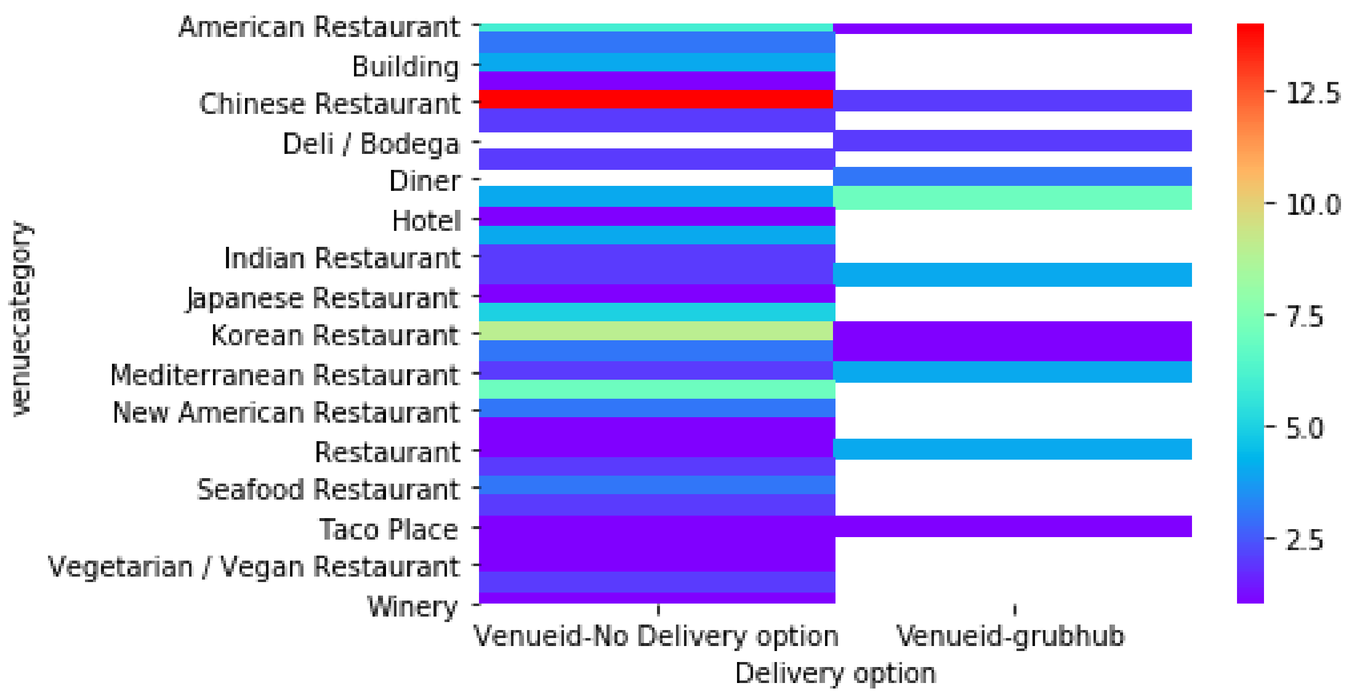
CLUSTER J – It consists of French restaurants with no major delivery provider.

Bar plot of various categories of restaurants vs. count of restaurants for different delivery options are plotted. This clearly shows categorical distribution of restaurants and their delivery provider options.



5. Discussions

Restaurant category and delivery provider option of dataframe was grouped and a pivot table was create to show number of venues corresponding to restaurant categories and delivery provider data. The Pivot table created was visualized using heat map.



It can be seen that the highest number of restaurants correspond to chinese and they don't have a major delivery provider. This corresponds to cluster 7 which can be the hottest attraction for new delivery providers. Its followed by Korean, American and New American categories.

The success probability of new delivery provider can be summarized as below:

cluster	Cuisines to be supported by Application	Delivery provider	success probability
0	Miscellaneous	None	Moderate
1	Miscellaneous	grubhub	Low
2	American	None	High
3	Mexican	None	High
4	New American	None	High
5	Mediterranean	grubhub	Low
6	Korean	None	High
7	Chinese	None	High
8	French	grubhub	Low
9	French	None	Moderate

High success scenarios are those clusters which don't have a current delivery provider and are specific to certain cuisine so it can be implemented and advertised easily.

French restaurants have moderate success as some already have a major delivery provider and the rest may start following the same provider.

Miscellaneous restaurants without any delivery provider have been assigned Moderate success as the delivery provider needs to support multiple cuisines.

Miscellaneous categories with existing delivery providers have least success rate. Also, Mediterranean categories have been assigned low success due to the same reason.

6. Conclusion

For starting a new delivery provider service to current restaurants, Chinese, Korean, American and Mexican restaurants will prove to be highly beneficial due to their large numbers and due to lack of any current major delivery provider competitors. French Restaurants will be moderately beneficial as some French restaurants are already using grubhub service provider and the rest may follow them too. Cluster 0 corresponding to miscellaneous cuisines may be profitable as they don’t have current delivery provider but might be challenging as multiple cuisines need to be supported. Cluster 1 restaurants will be least profitable as they have existing provider and also complex due to multiple miscellaneous categories.From perspective of providing new cuisines/specialties in the location Chinese/Korean/French/Mexican/American may not be beneficial as they are already present in large numbers. Indian, Japanese, Peruvian, Vietnamese are in very few numbers. Starting these cuisines may be beneficial.