Food Truck Location Planner

Business Problem

This project aims to analyze open data for a city to determine the best places for a food truck business. In this case we are assuming that a friend has a food truck that serves tacos. They want to find out where in the city they should plan to spend most of their time. So they have the following requirement,

- The area should be high population density.
- The area crime rate should be low.
- The number of competing restaurants should be low.

After spending countless hours online researching this and not getting anywhere they turned to me for help.

Data Sources

The first step was to see what data is available freely for the area. In this case the city was Denver. The following was the data source used and where this dataset is available.

Available on the Denver city website, (https://www.denvergov.org/opendata/search?mode=table)

- Denver Census Data.
- Denver Crime Data.

FourSquare data

- Venues - restaurants, bars, parks.

The data will be used to determine the following,

1. Census Data
    - Find the population density of various neighborhoods.
    - Find the age groups living in those areas.
    - The primary areas of focus would be areas where age groups between 21 - 40.
    - The assumption is that this group is more outgoing and has the ability to spend on eating out instead of making food at home.
2. Crime Data
    - Determine where most of the crimes are taking place
    - Determine the crime density
    - Map this data to give a visual of the ideal locations
3. Location Specific Data
    - Based on the above data get venues for the top 10 places.
    - Determine what is the most common venue in that location.
    - Avoid areas with high concentration of Mexican restaurants.

Methodology

Data Analysis

The data was collected from various sources. This data needed a lot of cleaning and processing to allow for analysis that was intended. The section below describes the processing done on the data sources.

Population Data

The original file had a lot of information about that was split in various groups. However, it was determined that we need only the following columns from that data

- Neighborhood names
- Population data
- Male and Female population
- Age groups
    - Under 20
    - 21-40
    - 41-60
    - 60+

In order to get this data the extra columns were dropped and an additional column containing neighborhood names in lowercase was added. This would later help to merge the data frame with other data frames.

	population	num_houses	MALE	FEMALE	nbr_lower
neighborhood					
Montbello	30348.0	8516.0	15111.0	15237.0	montbello
Gateway / Green Valley Ranch	29201.0	9900.0	14160.0	15041.0	gateway / green valley ranch
Hampden	17547.0	9917.0	8438.0	9109.0	hampden
Westwood	15486.0	4717.0	7928.0	7558.0	westwood
Capitol Hill	14708.0	11757.0	8331.0	6377.0	capitol hill

For the age groups separate data frames were created as the original dataset contained these groups in separate columns with 5 year increments (5-10, 15-20 and so on). These data frames were then summed across rows to give the total population in that group. So for example 5-10, 15-20 groups were added to one group called under 20.

All these data frames were then merged into a data frame called df_pop.

	population	num_houses	MALE	FEMALE	nbr_lower	under20	21-40	41-60	above 60
neighborhood									
Montbello	30348.0	8516.0	15111.0	15237.0	montbello	12138.0	9031.0	6324.0	2855.0
Gateway / Green Valley Ranch	29201.0	9900.0	14160.0	15041.0	gateway / green valley ranch	10755.0	9827.0	6849.0	1770.0
Hampden	17547.0	9917.0	8438.0	9109.0	hampden	3113.0	5391.0	4643.0	4400.0
Westwood	15486.0	4717.0	7928.0	7558.0	westwood	6176.0	4752.0	3065.0	1493.0
Capitol Hill	14708.0	11757.0	8331.0	6377.0	capitol hill	562.0	9436.0	3216.0	1494.0

Crime Data

The crime data was very detailed and spanned 5 years. However, the traffic related incidents were also reported that were removed.

	OFFENSE_ID	OFFENSE_TYPE_ID	OFFENSE_CATEGORY_ID	REPORTED_DATE	INCIDENT_ADDRESS	GEO_LON	GEO_LAT	PRI
0	2018869789239900	theft-other	larceny	12/27/2018 4:51:00 PM	2681 N HANOVER CT	-104.866156	39.755561	
1	2015664356544100	traffic-accident	traffic-accident	11/13/2015 8:38:00 AM	4100 BLOCK W COLFAX AVE	-105.040760	39.739991	
2	20176005213239901	theft-bicycle	larceny	6/12/2017 8:44:00 AM	1705 17TH ST	-104.999264	39.753669	
3	20196012240230800	theft-from-bldg	larceny	12/9/2019 1:35:00 PM	1350 N IRVING ST	-105.029208	39.738134	
4	2018861883501600	violation-of-restraining-order	all-other-crimes	12/22/2018 10:00:00 PM	13625 E RANDOLPH PL	-104.828868	39.797750	

This was done using the column ‘CRIME_IS’. Once this was removed the data was filtered to get crimes committed after 2018. To do this the date column in the data frame had to be changed since the original data frame was stored as an object.

The column was set to date using the pd.to_date function. A series was created on which the to_date function was applied. This was added to the main crime data frame as year and then it was possible to filter out the dates.

The next step was to group these statistics by neighborhood using the count function. This gave us a simple data frame containing neighborhood, latitude and longitude and crime count.

	neighborhood	GEO_LAT	GEO_LON	Crime_Count	nbr_lower
0	west colfax	39.738134	-105.029208	1892	west colfax
1	montclair	39.729818	-104.919802	564	montclair
2	westwood	39.700625	-105.047350	1839	westwood
3	five points	39.754698	-104.988366	5759	five points
4	lincoln park	39.739501	-105.000542	2055	lincoln park

Income Data

The city maintains a community survey that contains data about household income and poverty by different neighborhoods. This dataset was relatively easy to clean as only 4 columns were sliced out of the original data. These were

- Neighborhood
- Median Household income
- Median Family Income
- Percentage poverty

This data was then merged with the population data frame, df_pop.

neighborhood	population	num_houses	MALE	FEMALE	nbr_lower	under20	21-40	41-60	above 60	hh_income	family_income	PCT_POVERTY
Montbello	30348.0	8516.0	15111.0	15237.0	montbello	12138.0	9031.0	6324.0	2855.0	53788	52312	20.666667
Hampden	17547.0	9917.0	8438.0	9109.0	hampden	3113.0	5391.0	4643.0	4400.0	54240	70033	10.950000
Westwood	15486.0	4717.0	7928.0	7558.0	westwood	6176.0	4752.0	3065.0	1493.0	35135	35640	31.550000
Capitol Hill	14708.0	11757.0	8331.0	6377.0	capitol hill	562.0	9436.0	3216.0	1494.0	45296	85005	17.333333
Hampden South	14370.0	7787.0	6974.0	7396.0	hampden south	2828.0	4575.0	3708.0	3259.0	69930	99149	4.833333

Combining Data

The final step in data processing was to merge df_crime with df_pop. Remember the nbr_lower column we created at the start. This column is used here to combine the population data with crime data. The new data frame was called df_combined.

neighborhood	population	num_houses	MALE	FEMALE	under20	21-40	41-60	above 60	hh_income	family_income	PCT_POVERTY	lat	lon	Crime_Count
montbello	30348.0	8516.0	15111.0	15237.0	12138.0	9031.0	6324.0	2855.0	53788	52312	20.666667	39.773785	-104.812508	3213
hampden	17547.0	9917.0	8438.0	9109.0	3113.0	5391.0	4643.0	4400.0	54240	70033	10.950000	39.670597	-104.875513	1473
westwood	15486.0	4717.0	7928.0	7558.0	6176.0	4752.0	3065.0	1493.0	35135	35640	31.550000	39.700625	-105.047350	1839
capitol hill	14708.0	11757.0	8331.0	6377.0	562.0	9436.0	3216.0	1494.0	45296	85005	17.333333	39.735581	-104.977276	3315
hampden south	14370.0	7787.0	6974.0	7396.0	2828.0	4575.0	3708.0	3259.0	69930	99149	4.833333	39.626882	-104.894224	1156

Machine Learning - Applying K-Means

The data set had quite a bit of features and it was not very practical to evaluate these using graphical methods. A faster way to achieve this objective was to use clustering techniques. The combine dataset was simplified to contain 8 features that were deemed the most important.

- Population
- Age groups
    - Under 20
    - 21-40
    - 41-60
    - 60+
- Household income
- Family income (these two differ because not all households are family units)
- Percentage poverty
- Crime count
- Percentage of 21-40 year old in population

neighborhood	population	under20	21-40	41-60	above 60	hh_income	family_income	PCT_POVERTY	Crime_Count	pct_21-40
montbello	30348.0	12138.0	9031.0	6324.0	2855.0	53788	52312	20.666667	3213	29.758139
hampden	17547.0	3113.0	5391.0	4643.0	4400.0	54240	70033	10.950000	1473	30.723201
westwood	15486.0	6176.0	4752.0	3065.0	1493.0	35135	35640	31.550000	1839	30.685781
capitol hill	14708.0	562.0	9436.0	3216.0	1494.0	45296	85005	17.333333	3315	64.155562
hampden south	14370.0	2828.0	4575.0	3708.0	3259.0	69930	99149	4.833333	1156	31.837161

The next step was to determine the ideal number of clusters. The first test was done with only 4 clusters however this resulted in clusters that were too generalized and hard to distinguish on features. The determining factors used were

- Distribution of income.
- 21-40 population percentage.
- Crime count and crime per capita.

Increasing the number of clusters to 5 resulted in slightly better results. So 6, 7, 8 and 10 clusters were tried. The best results were achieved at 7 clusters. At this point the neighborhoods were very closely arranged according to income, 21-40 population and crime profile. The clusters were plotted on a map but

the map did not reveal much of a pattern or information to decide which clusters to pick. The map is shown in the figure below.

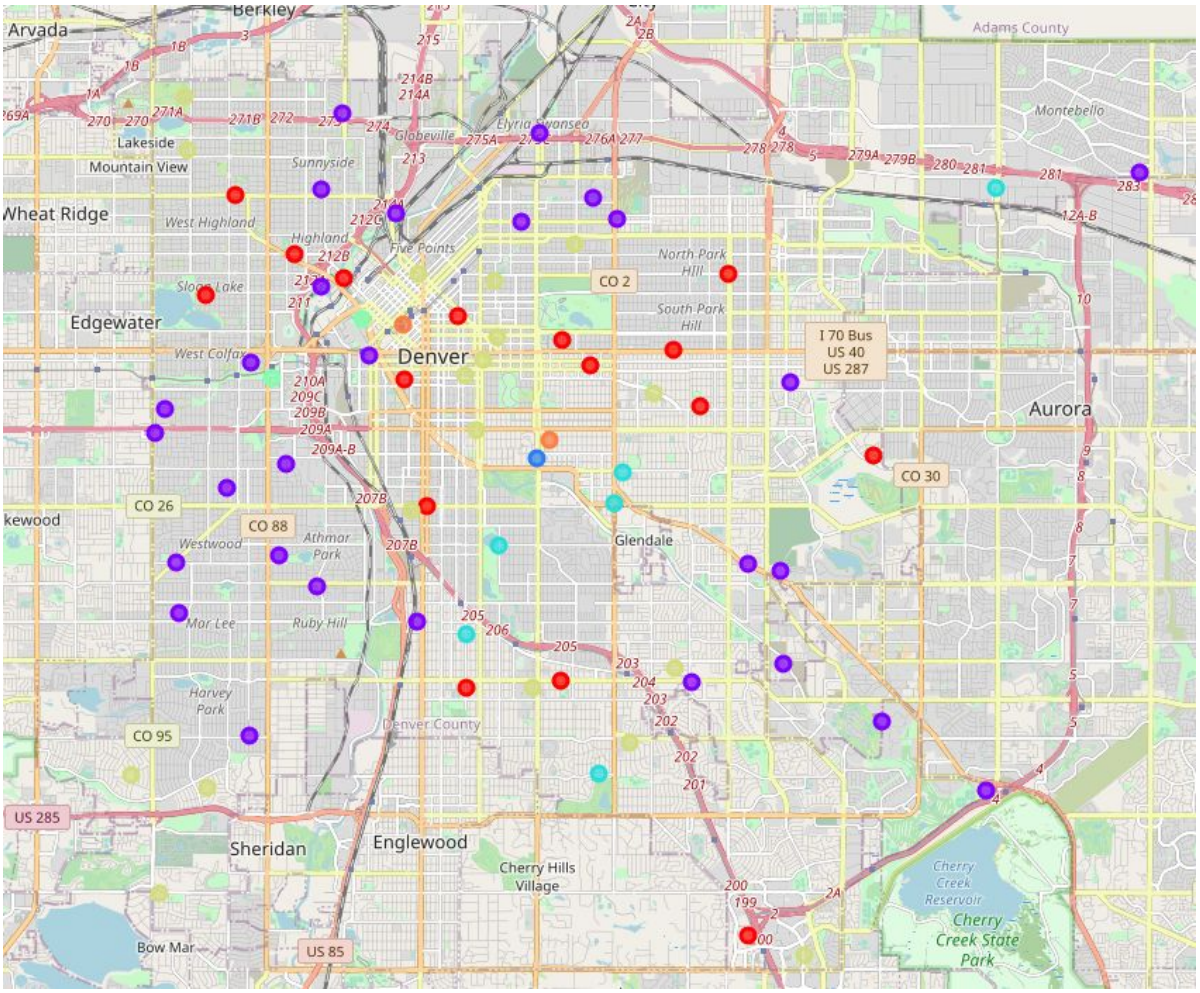
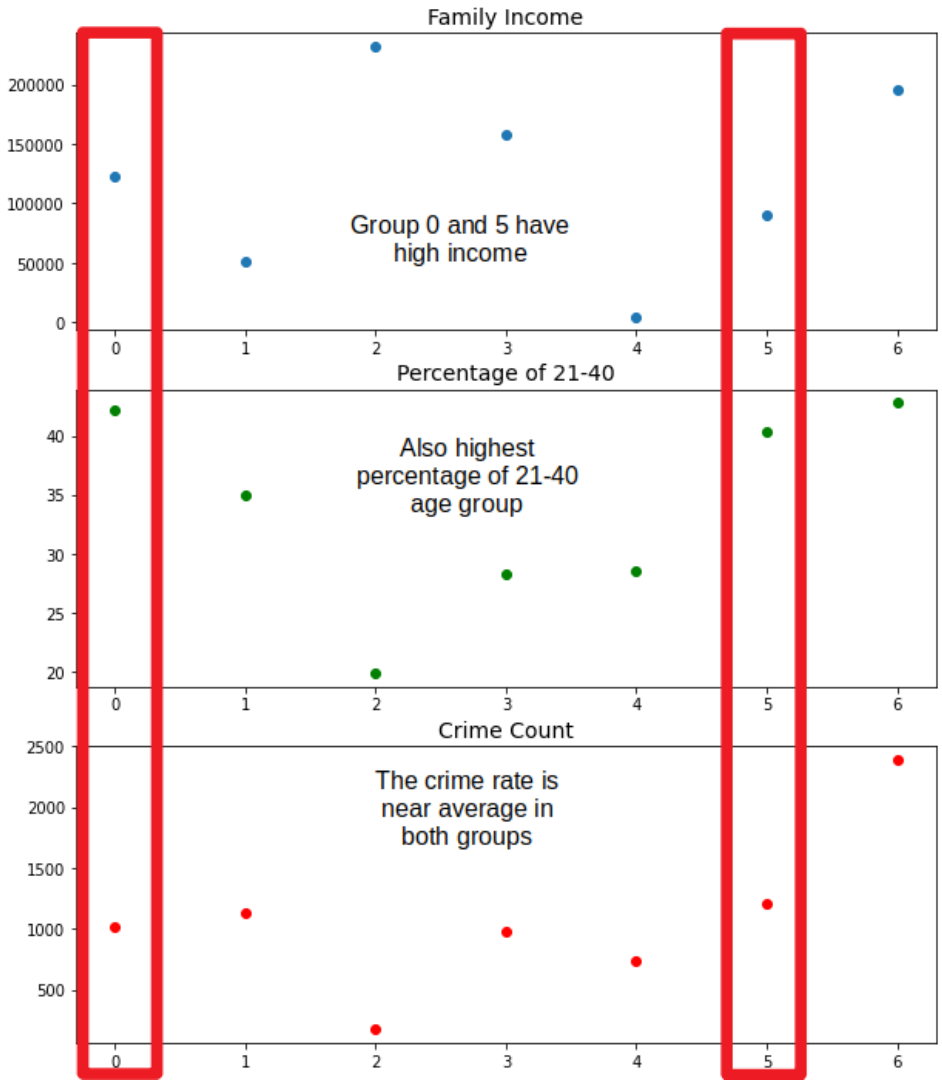


Fig 1: Geographical distribution of the neighborhood clusters

Selecting the right cluster

Once these clusters were established the next step was to select the right clusters. The clusters were grouped by mean values and then combined to form a dataframe. The following parameters were then plotted on scatter plots

- Family income
- Percentage of 21-40 age group
- Crime count



It was determined that cluster 0 and 5 had the best features. This was because of the following

- Both areas have a high percentage of 21-40 age groups (above 40%).
- The income levels in both areas are high enough to be able to spend on eating out.
- The crime count is close to mean.

This criteria helped eliminate some areas that had high income but lower number of people in the right age group or generally higher levels of crime.

Getting Venues

The next step was to get venues in the selected neighborhoods. The two clusters were combined and the coordinates for these were passed to FourSquare API. The search radius was set to 2000m and 50 venues were searched.

```
[ ]  # organizing results in a dataframe
     df_venues = pd.DataFrame([item for venue_list in venues_list for item in venue_list])
     df_venues.columns = ['Neighborhood', 'Neighborhood Latitude', 'Neighborhood Longitude'
     df_venues.shape
```

```
(1720, 7)
```

These venues were then one hot encoded (again using the function from the previous exercises). Using these encodings the number of unique venues for each neighborhood was determined.

	Neighborhood	ATM	Alternative Healer	American Restaurant	Arcade	Art Gallery	Art Museum	Arts & Crafts Store
0	baker	0.000000	0.00	0.060000	0.00	0.04	0.00	0.02
1	bear valley	0.042553	0.00	0.000000	0.00	0.00	0.00	0.00
2	berkeley	0.000000	0.00	0.000000	0.00	0.00	0.00	0.00
3	capitol hill	0.000000	0.00	0.060000	0.00	0.02	0.02	0.00

This data was collected in a neighborhoods_venues_sorted data frame.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	baker	Bar	American Restaurant	Marijuana Dispensary	Breakfast Spot	Coffee Shop
1	bear valley	Mexican Restaurant	Hotel	Convenience Store	ATM	Coffee Shop
2	berkeley	Brewery	Coffee Shop	Pizza Place	Park	Mexican Restaurant
3	capitol hill	Sandwich Place	American Restaurant	Yoga Studio	Breakfast Spot	Brewery
4	cheesman park	Sandwich Place	American Restaurant	Pizza Place	Mexican Restaurant	Breakfast Spot

Final Selection

The final selection step was simple. The locations that had the first most common venues as a bar or brewery were selected. This was done making the assumption that these places usually have frequent foot traffic and people that are willing to spend money on street vendors. The next step was to see if any of these had Mexican restaurants as top 5 venues.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	baker	Bar	American Restaurant	Marijuana Dispensary	Breakfast Spot	Coffee Shop
2	berkeley	Brewery	Coffee Shop	Pizza Place	Park	Mexican Restaurant
9	five points	Brewery	Pizza Place	Bar	Restaurant	Coffee Shop
20	regis	Brewery	Pizza Place	Bar	Coffee Shop	Park
34	whittier	Brewery	American Restaurant	Pizza Place	Coffee Shop	Bar

The final selection was made on the population and age groups. High population and high percentage of 21-40 age groups were the ideal location.

neighborhood	population	num_houses	MALE	FEMALE	under20	21-40	41-60	above 60	hh_income	family_income	PCT_POVERTY	lat	lon	Crime_Count
baker	4879.0	2697.0	2565.0	2314.0	823.0	2208.0	1299.0	549.0	67996	83877	16.800000	39.710262	-104.989994	1530
berkeley	8112.0	4322.0	3884.0	4228.0	1410.0	3027.0	2232.0	1443.0	77070	104158	8.400000	39.777990	-105.044876	738
five points	12712.0	7147.0	7440.0	5272.0	1615.0	6878.0	3159.0	1060.0	63316	90983	22.233333	39.754698	-104.988366	5759
regis	3934.0	1686.0	1884.0	2050.0	950.0	1333.0	1006.0	645.0	68342	88272	16.900000	39.787971	-105.045320	454
whittier	4831.0	2150.0	2380.0	2451.0	1226.0	1962.0	1104.0	539.0	74611	79196	15.700000	39.753426	-104.969991	553

Results

Based on the above analysis the best location for the food trucks was determined to be the ‘Five Points’ area. This area had the following features that made it a good choice.

- It has a large population out of which at least 54% is between the age of 21-40.
- It also has a large concentration of bars and breweries that are good places for a food truck to find a location to park and serve food.
- The top 5 locations in that area do not include any Mexican restaurant which means less competition for the taco truck.
- In terms of the number of crimes that neighborhood appears to be in the top 10 areas however per capita crime rate is much lower than the nearby neighborhoods.
- The other main advantage of this location is close proximity to several high population density areas which means the total foot traffic could potentially be much higher.

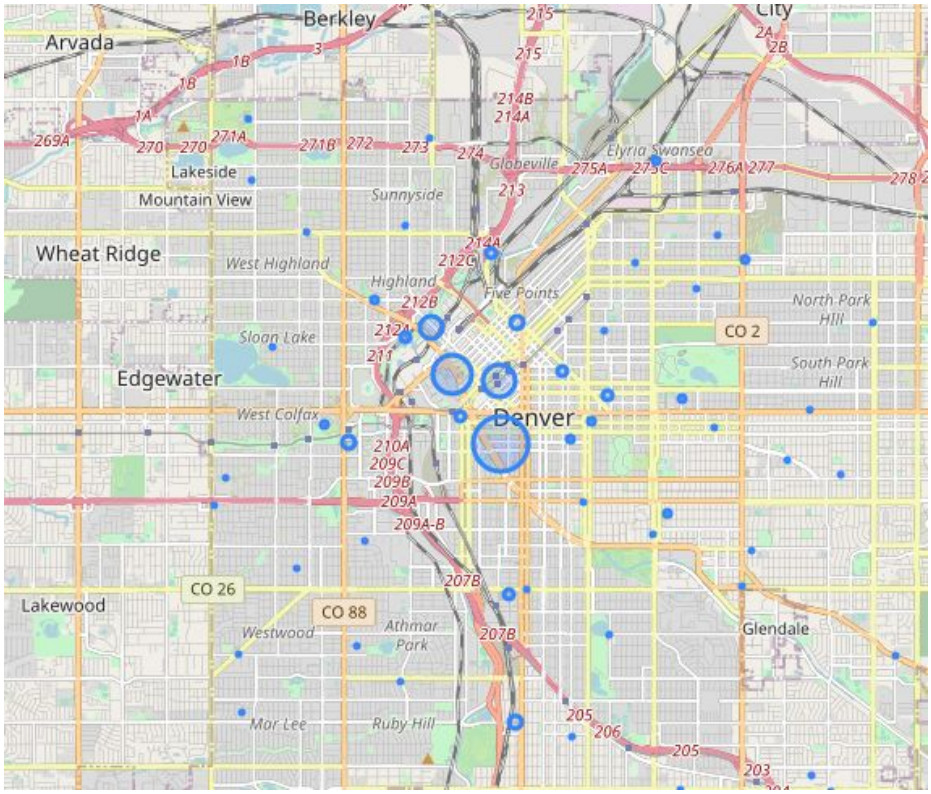


Fig 2: crime/population map for different neighborhoods

Discussion

The data for this project was solely public domain data which had many features distributed in a wide variety of sources. This made the data collection and analysis the most time consuming part of the project.

The problem solved here can actually be expanded upon especially if data is available on different businesses in different areas about revenues generated, foot traffic density, special events etc. However, even without that data it is possible to get useful insights that can help to position a business in a better location and possibly capture a bigger market than it otherwise would have.

During the analysis several observations were made that have not been all included in the above sections. For example the five points area did have a high crime count but its crime rate per capita was much lower. Also it was noted that some areas like CBD actually had a much larger population density of 21-40 age groups but its overall population was small and the crime rate in the area was much higher.

Because the business problem asked to select an area for the business a clustering approach was deemed suitable as we were not asked to make a prediction about a feature but only find the best combination of features.

It is also possible to expand this approach to solve other problems such as the ideal place to rent an apartment or buy a house. It is also possible to determine locations for community services or charities of different kinds.

Conclusion

To sum up the project, we were asked to determine the best location for a business based on some simple assumptions and public domain data. It was decided based on the available data to select eight features to run the analysis on and select the best location.

These features were fed into a clustering algorithm and two clusters were chosen that had most attractive features. The final selection was done purely on competitive advantage and higher probability of customer turn around (high population, no competitive vendors and close proximity to locations to conduct business).

It is suggested to review the recommendations based on the observations after a certain time has elapsed (3 months would be a good time frame). The observations that we should look to collect would include traffic patterns, age groups, average amount spent, favorite items, temporary set up in various locations etc. This data could help fine tune the recommendations further and possibly allow for much better business recommendations.