Initial Location Analysis in San Francisco for Restaurante La Tierra De Los Tacos

By: Daniel Garcia
Date: January, 2020

Overview

1. Background
2. Preprocessing
3. Analysis
4. Clustering
5. Conclusion

1. Background

Methods Overview

- Business Understanding
- Data Collection
- Data Preprocessing
- Data Analysis
- Clustering
- Conclusions and Report

Problem: Choosing a Location For Restaurante La Tierra De Los Tacos

- Late stages in business plan with aims of establishing a restaurant in San Francisco
- Needs assistance in filtering out the neighborhoods that don't meet their specific requirements.
- “Neighborhood” = Census tracts designated to the SF county

Location Requirements

Contain a population of 5,000 + (increase opportunity of traffic and visibility)

Near at least one college and surrounded by at least three other types of schools (partnerships with schools and other organizations, and to provide catering/delivery options)

Rent at or below the city's average -$3,000

Median income of $50,000 or more (safeguard for tough financial times)

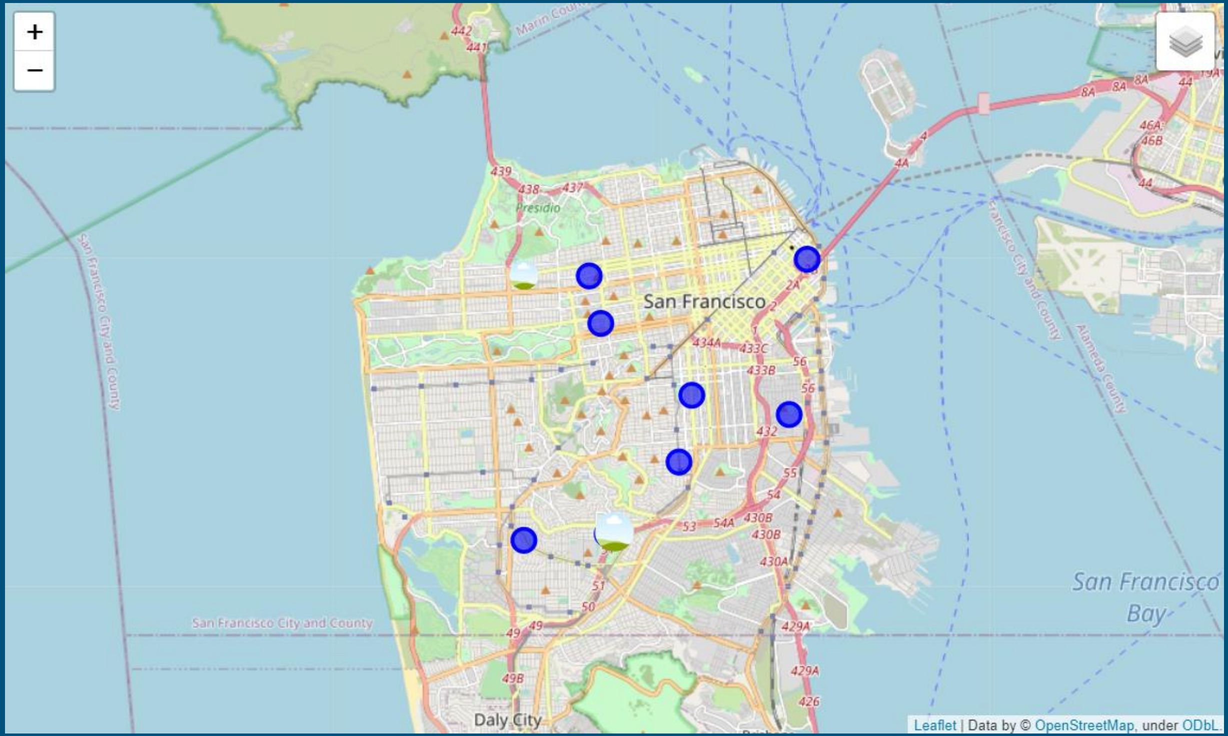
With minimal competition (Mexican restaurants) and similar venues (ie. restaurants) nearby

Data Sources

1. ACS 5 Year Estimates (2013-2017) -- Selected Characteristics of the Total and Native Populations in the United States (Census Table ID: 601)
1. ACS 5 Year Estimates (2014-2018) -- Selected Housing Characteristics (Census Table ID: DP04)
1. Colleges (2011)
1. Analysis Neighborhoods - 2010 Census Tracts Assigned to Neighborhoods
1. Venues retrieved from FourSquare API call

Solution: 7 Potential Census Tracts

1. Tract 154
2. Tract 165
3. Tract 514
4. Tract 615
5. Tract 217
6. Tract 215
7. Tract 309



2. Data Preprocessing

Main Preprocessing steps

- Duplicate, missing, and irrelevant data as well as columns were dropped
- Merging of three datasets into one dataframe
- Changed data types of certain columns from the merged dataframe

Filtering Based on Business Requirements

- Also filtered out data based on the first three business requirements
    - More than 5,000 population
    - Median income >= $50,000
    - Gross rent >= $3,000
- Was left with 18 tracts out of the original 197 tracts
- Retrieved venues using Four Square ‘recommended venues’ API call

Dataframes After Preprocessing

Out[23]:

	Tract	Total Pop	Estimated Median Gross Rent	Estimated Median Values (Owner-Occupied Units)	Median Income	Median Age	Total Male %	Total Female %	lat	long
0	Census Tract 154	5877	2267	1574000	70083	36.0	46.0	54.0	37.7843608	-122.4510358
1	Census Tract 165	5760	1865	1215900	60518	33.2	43.8	56.2	37.7741958	-122.4477884

In [24]:

```
#change the types of the following columns
SF_Summary=SF.astype({'Total Pop':'int64', 'Median Income':'int64','Total Male %':'float64','Total Female %':'float64',
                     'Median Age':'float64', 'Estimated Median Gross Rent': 'int64',
                     'Estimated Median Values (Owner-Occupied Units)': 'int64'})
```

In [25]:

```
SF_Summary.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16 entries, 0 to 15
Data columns (total 10 columns):
Tract                                             16 non-null object
Total Pop                                         16 non-null int64
Estimated Median Gross Rent                       16 non-null int64
Estimated Median Values (Owner-Occupied Units)    16 non-null int64
Median Income                                     16 non-null int64
Median Age                                        16 non-null float64
Total Male %                                      16 non-null float64
Total Female %                                    16 non-null float64
lat                                               16 non-null object
long                                              16 non-null object
dtypes: float64(3), int64(4), object(3)
memory usage: 1.4+ KB
```

The main SF dataframe
That resulted after preprocessing.

```
VENUES.head(2)
```

[33]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Census Tract 165	37.774196	-122.447788	Home Service Market aka "George's"	37.774063	-122.445976	Convenience Store
1	Census Tract 165	37.774196	-122.447788	Soothe	37.773662	-122.447404	Massage Studio

Lastly we print out how many items were deleted from the venues dataset.

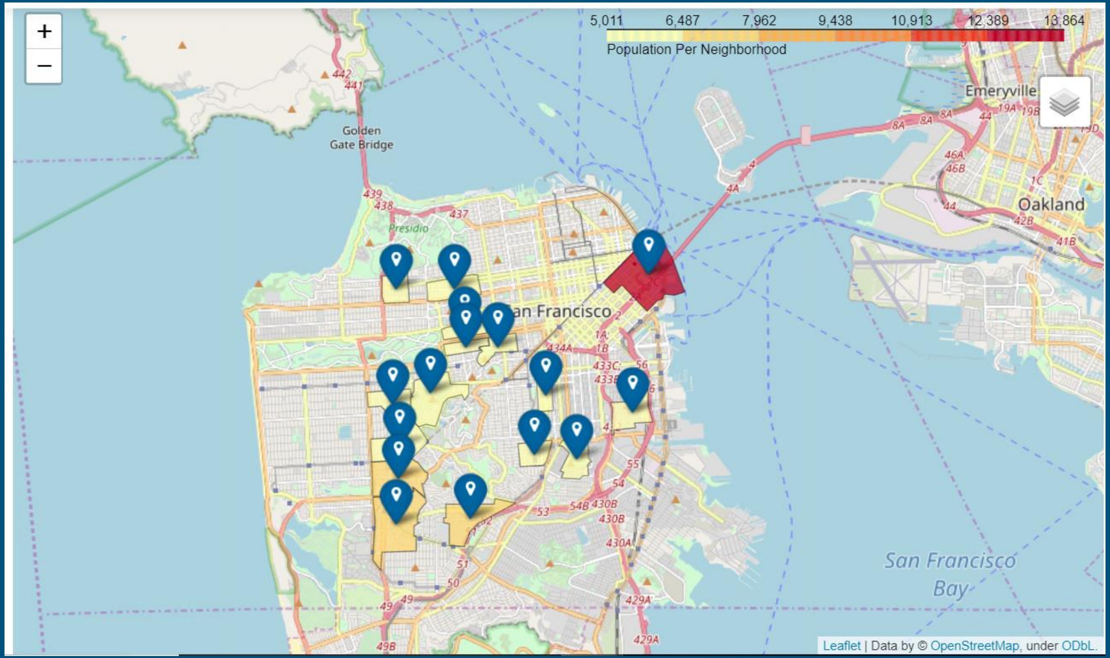
```
print('There are {} uniques categories.'.format(len(VENUES['Venue Category'].unique())))

print('Percentage of rows kept: {:.2%}'.format(len(VENUES)/len(dr)))
print('Dataframe rows before preprocessing: {} \n Dataframe rows after preprocessing: {}'.format(len(dr),VENUES.shape[0]))
```

```
There are 155 uniques categories.
Percentage of rows kept: 59.42%
Dataframe rows before preprocessing: 764
 Dataframe rows after preprocessing: 454
```

The venues data that was retrieved from FourSquare and then cleaned.

3. Analysis

Initial Folium Map Visualizations

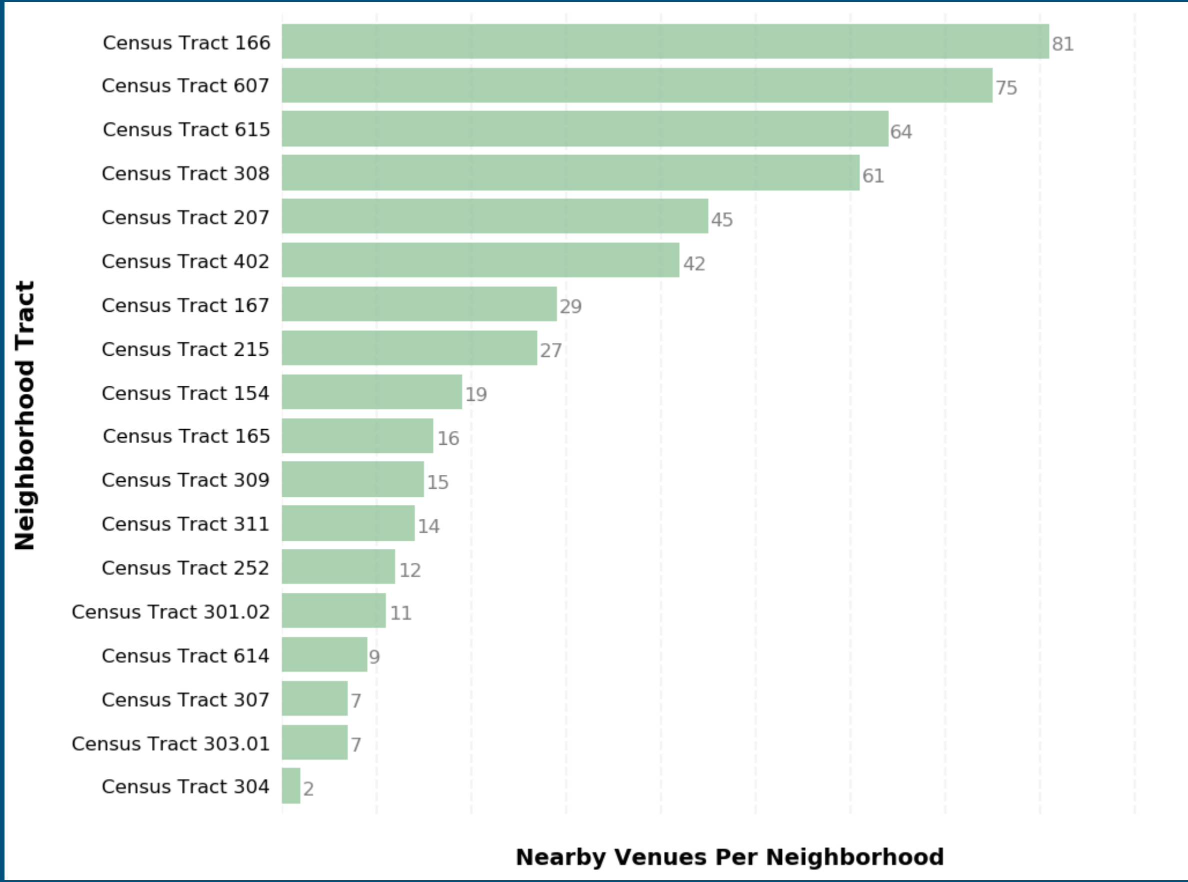


- Population choropleth map
- Select Census tract markers
  - Population and median income as popup info.

Most of the districts that meet the first three criterias are mostly roughly located around the center and upper east side of SF.

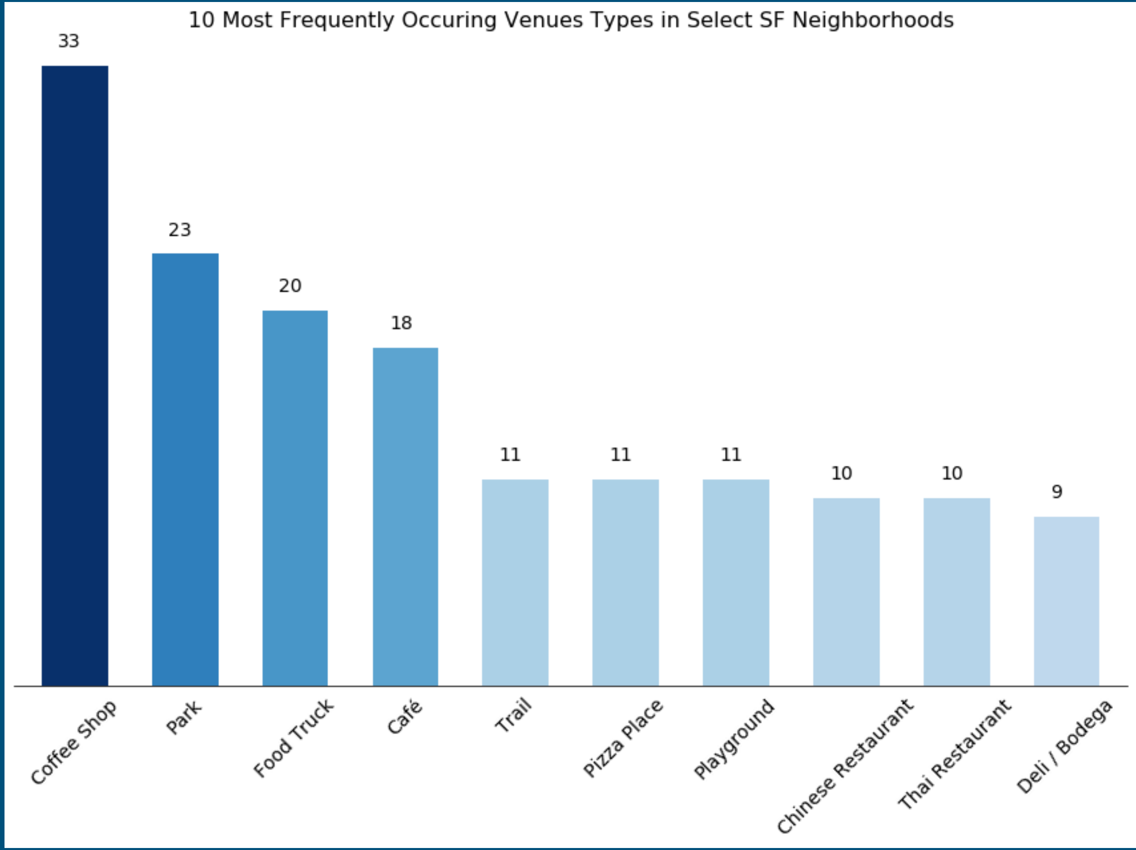
Venue Analysis

- Expected a good chunk to be close to 100 given the many restaurants.
- 81 nearby venues
- Roughly half of the tracts having a nearby venue count that is less than 19.



Venues Analysis continued ...

- Coffee shops (in combination to cafes) by far the most frequently occurring venues.
- Parks were also quite common.



4. Clustering

Clustering

- Segment tracts according to their venue categories
- Applied one-hot encoding technique for the venue categories
- Grouped by their venue categories and then took each one of their mean

```python
# Lastly create a dataframe with grouped by the mean of each venue category per SF tract
sf_grouped = venue_onehot.groupby('Neighborhood').mean().reset_index()
sf_grouped.head(3)
```

	Neighborhood	Accessories Store	Adult Boutique	Alternative Healer	American Restaurant	Arcade	Art Gallery	Arts & Crafts Store	Asian Restaurant	BBQ Joint	...	Toy / Game Store	Trail	Tunnel	Vegetarian / Vegan Restaurant	Vic St
0	Census Tract 154	0.000000	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	0.0	...	0.0	0.0	0.052632	0.000000	0.0C
1	Census Tract 165	0.000000	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	0.0	...	0.0	0.0	0.000000	0.000000	0.06
2	Census Tract 166	0.024691	0.0	0.0	0.012346	0.0	0.0	0.012346	0.012346	0.0	...	0.0	0.0	0.000000	0.012346	0.0C

3 rows × 156 columns

Clustering continued ...

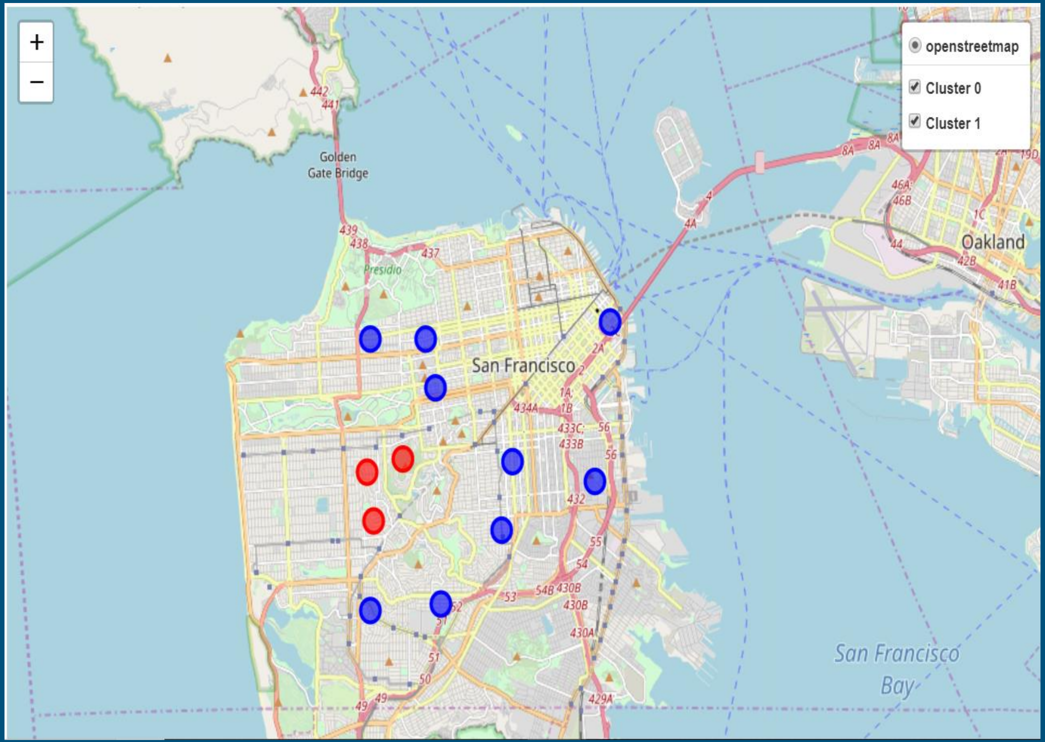
- Used the K-Means Algorithm
- Given the small size of tracts (16), chose 2 as the K clusters

```
kclusters = 2
# to be able to cluster only the venue categories
temp = sf_grouped.drop('Neighborhood', 1)
# run Kmeans with different centorid seeds and select the best n-starting point
kmeans = KMeans(n_init=50, n_clusters=kclusters, random_state=0).fit(temp)

print(len(kmeans.labels_))
print(kmeans.labels_)

# insert the resulting cluster labels to the datframe containing the top 10 most commmon venues.
neighborhoods_venues_sorted.insert(0, 'ClusterLabels', kmeans.labels_)
```

```
16
[0 0 0 0 0 0 0 1 1 1 0 0 0 0 0 0]
```

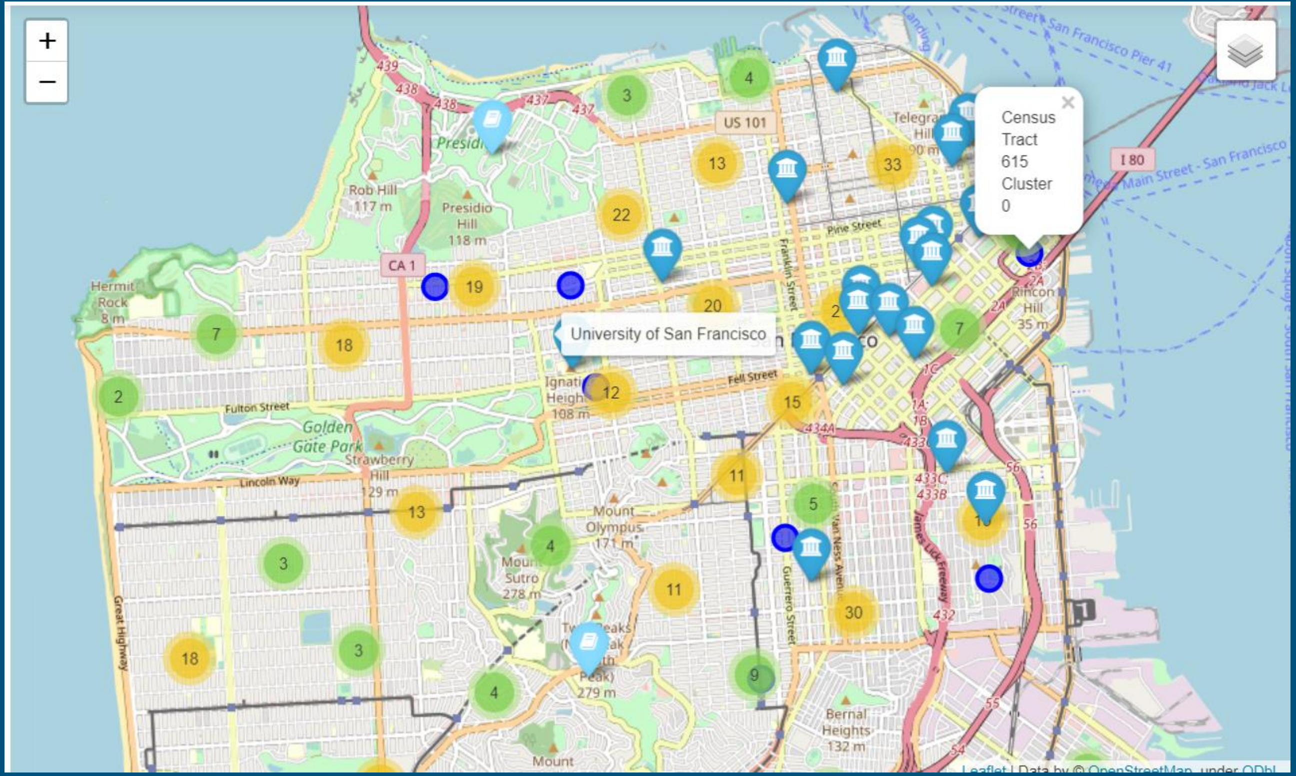
Cluster Map

- Chose cluster 0 (blue) as the target cluster
- Made some further filters by excluding tracts in the target cluster that contained a Mexican food-related venue



Schools in San Francisco Map

Filtered out tracts 402 and 311 as they didn’t meet the last business requirement.



5. Conclusion

Key Points

- Found seven tracts/neighborhood candidates for the business to choose from
    - Characterized by coffee shops, outdoor venues such as parks, and asian restaurants.
    - Scattered around the mid-center of the city.
    - Affluent with a median income mean of 77,701 dollars, and an average gross rent of 2,229 dollars.
- Tract 615 stood --high population size, high median income (103,451 dollars), and proximity to several colleges.

Conclusion

- Candidate tracts (154,165,54,615,217,215, and 309) are potential good locations for the company to look further into as the requirements outlined in the introduction of this report.

Questions?