Determining the Optimal Location for a New Chipotle Restaurant in Queens, New York

Peter Jandovitz

March 29, 2021

1 Introduction

The "fast-casual" restaurant style has grown rapidly in popularity over the past decade. These restaurants straddle the border between fast food and traditional full table service restaurants, promising higher quality ingredients than at fast food restaurants at a lower price point than full-service restaurants.

In this report, we use demographic data as well as collected user preference and existing restaurant location data to try to determine attractive areas for opening a new fast-casual restaurant location. Specifically, we try to find areas in Queens, New York where we expect the chain Chipotle Mexican Grill to be popular and face relatively less competition. Accordingly, this report is targeted at the regional management of Chipotle. However, the data and analysis are relevant to a variety of stakeholders in the restaurant industry.

2 Data

The two main sources of data we use are demographic data from the US Census Bureau, and venue data from Foursquare.

We were able to obtain census-tract-level demographics using the US Census Bureau online API. Specifically, we used data from the 2019 American Community Survey 5-year estimates. We pulled tract-level data from the Detail Tables and Subject Tables to get information on the population's education, income, age, and ethnicity.

We also obtained geometric data for the census tracts using the Census TIGERweb REST API. This data is used for map visualizations, Foursquare location searches, and calculating population density.

We used the Foursquare Places dataset for two purposes. First, we used the venues/listed endpoint to determine the categories of venues that occurred in lists along with Chipotle. We used the most frequently occurring categories to serve as an estimate of the competitors of Chipotle. We then used the venues/search endpoint to determine the number of competitor venues in the vicinity of each census tract.

3 Methodology

3.1 Target and Feature Creation

The first part of our analysis consisted of computing our target variable of "Chipotle-like" venue density. As mentioned previously, we chose to use venues that appeared alongside Chipotle in user-created lists on Foursquare. We did not determine the actual subjects of the lists or the implied axis of similarity. Instead, we simply treat it as general measure of similarity among customer-base. The results of this analysis are shown in Table 1.

Categories	Count
Mexican Restaurant	707
Asian Restaurant	316
Bar	122
Pizza Place	100
American Restaurant	96
Italian Restaurant	94
Food & Drink Shop	92
Vegetarian / Vegan Restaurant	81
Seafood Restaurant	79
Dessert Shop	73

Table 1: Counts of venue categories in lists with Chipotle

The most frequently occurring venue category is "Mexican Restaurant," which is not surprising because that is the primary category of Chipotle in

Foursquare. The rest of the frequently occurring categories are fairly varied and might include venues not considered to be direct competitors to Chipotle. However, the categories are similar enough in purpose and customer-base that it makes sense to use as our measure of venue supply.

Once we obtained the list of similar categories, we used the Foursquare API to count, for each census tract, the number of venues within a 500m radius belonging to the top 5 categories. The radius and number of categories were chosen to return a large enough number of results while retaining enough locality. The final target variable of “venue density” was calculated simply as the venue count divided by the total tract population.

Some features were similarly calculated. Education or “bachelors degree” was calculated as the fraction of population having a bachelor’s or more advanced degree. Population density was calculated as the total population divided by total land area, not including non-residential parks and bodies of water. The rest of the features are self-explanatory.

3.2 Descriptive Statistics

Figure 1 shows the correlation matrix of our chosen features and target. The largest correlations are education with income, and education with ethnicity, both at 0.5. Income also shows a large negative correlation (-0.39) with population density. Venue density shows small to moderate correlations with all features except median age.

Since we observed some moderate correlations between our features, we calculated the principal components of the standardized feature set to see if there were redundancies we should remove to reduce the dimensionality. We determined that the smallest component explained 7% of the total variance. We decided that the variance was well-enough spread across our features to proceed without projection onto the principal components.

Figure 2 shows choropleth maps showing the values of some of our feature variables for each census tract. We can see some of the correlations discussed above, as well as geographic patterns in the data. Ethnicity is noticably clustered, with a large cluster of non-white population in the southeast, and smaller clusters in the center, and clusters of high-percentage white population in the north and southwest. Interestingly, income seems to vary on smaller scales than education or ethnicity. Population density is skewed by a high-density outlier caused by a small tract containing only high-density residential buildings. We removed this outlier, as well as outliers in median age

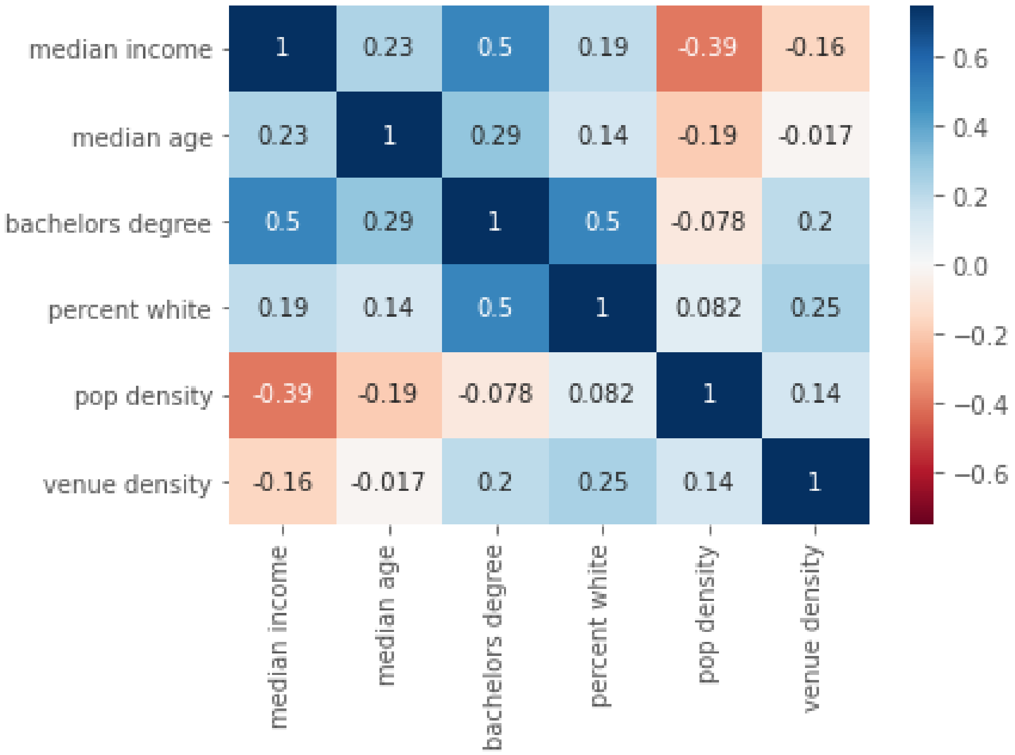


Figure 1: Correlation matrix

and median income, before proceeding with further analysis and modeling.

Figure 3 shows the calculated venue density for each tract. We see that most tracts have low densities, with small clusters of high density in the center and northwest regions.

3.3 Cluster Analysis

We performed a cluster analysis to group census tracts by demographic similarity. We ran both a hierarchical agglomerative clustering and a k-means clustering. We chose the number of clusters by eye, using a dendrogram for the agglomerative model and the "elbow" plot for k-means. We chose n=5 for the agglomerative model, and n=4 for k-means, although the number of clusters was not strongly indicated by either model, as shown in Figure 4.

The results of clustering are shown in Figures 5 and 6. Both models

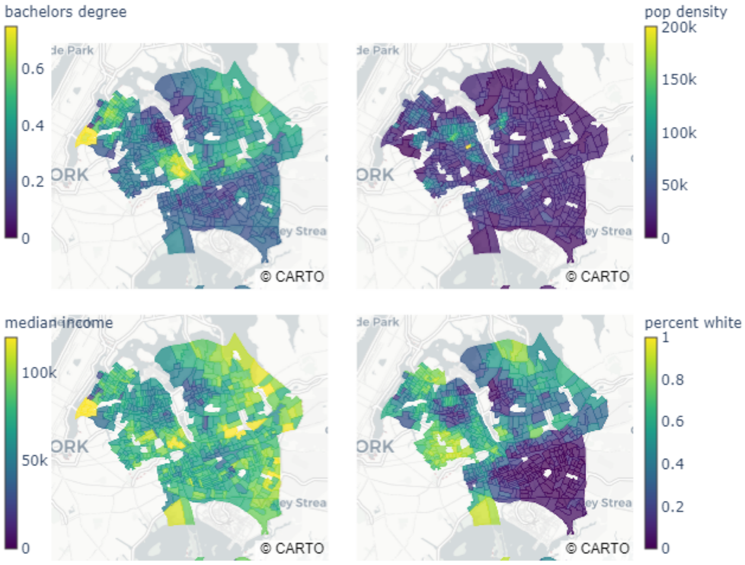


Figure 2: Choropleth maps of features

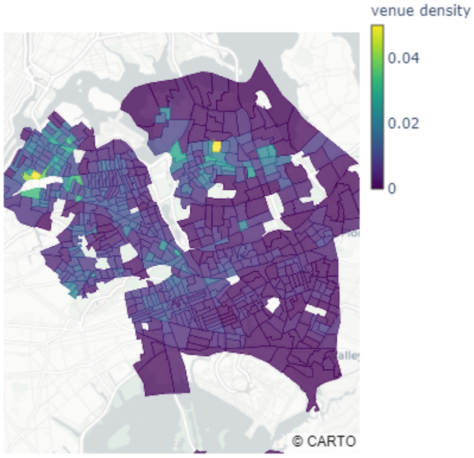
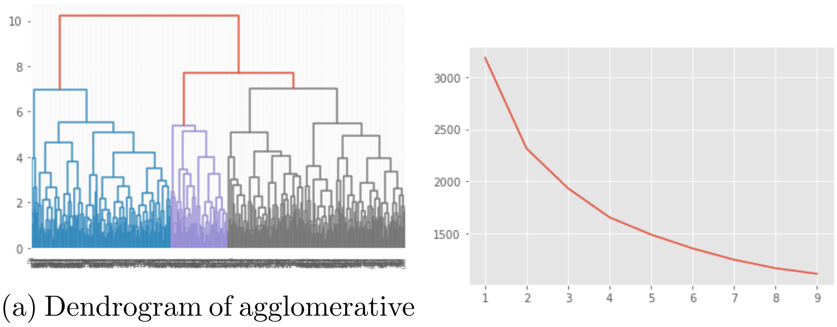


Figure 3: Choropleth map of venue density

showed similar results, but we decided to proceed with the n=4 k-means



(a) Dendrogram of agglomerative model

(b) Elbow plot for k-means

Figure 4: Choosing number of clusters

model because it produced better geographical separation and had a more even distribution of cluster size.

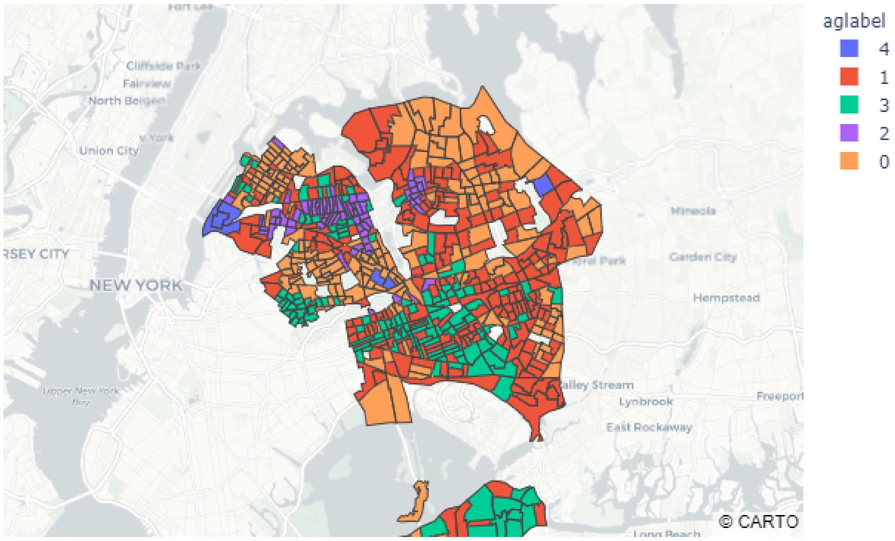


Figure 5: N=5 Agglomerative Clustering

Table 2 shows the mean values of features and target for each cluster

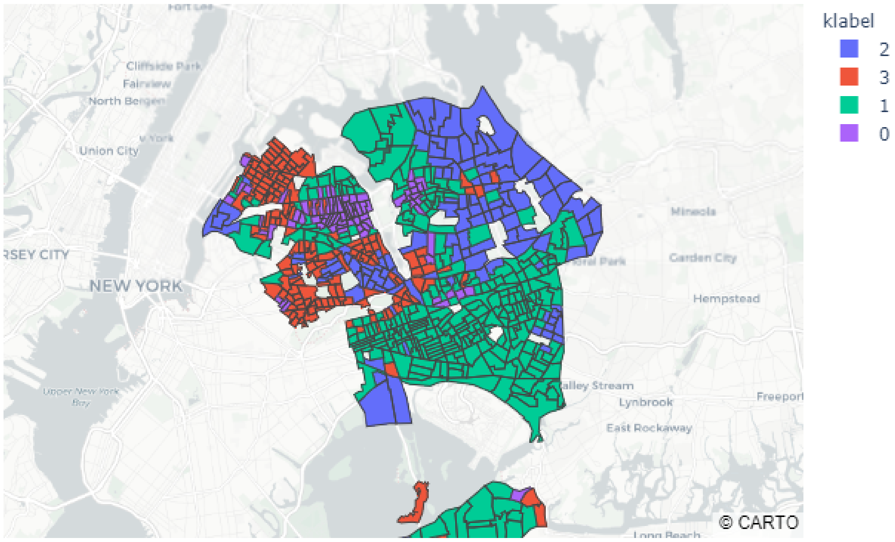


Figure 6: N=4 K-means Clustering

from the k-means models. We used these cluster mean values to calculate deviations of venue density for each tract. We searched for tracts with large, negative deviations on the assumption that demand for Chipotle-like venues is similar within a cluster, so a negative deviation represents demand unmet by current supply.

klabel	median income	median age	bachelors degree	percent white	pop density	venue density
0	50,200	36.3	0.188	0.317	8.09e+04	0.00752
1	70,700	36.7	0.226	0.171	2.76e+04	0.00582
2	95,800	45.4	0.441	0.529	1.87e+04	0.00563
3	73,900	36.8	0.397	0.665	4.44e+04	0.0107

Table 2: Mean values for each cluster

Figure 7 shows the negative deviation from cluster mean of venue density for each tract. The largest deviation occurs in the center of the map, next to the southeast end of Flushing Meadows Park. A band of deviation extends east from this point. There's a similar, but less intense band running along the northern border of the borough. The rest of the large deviation tracts are more scattered, with the most notable occurring in the western and northwestern sections.

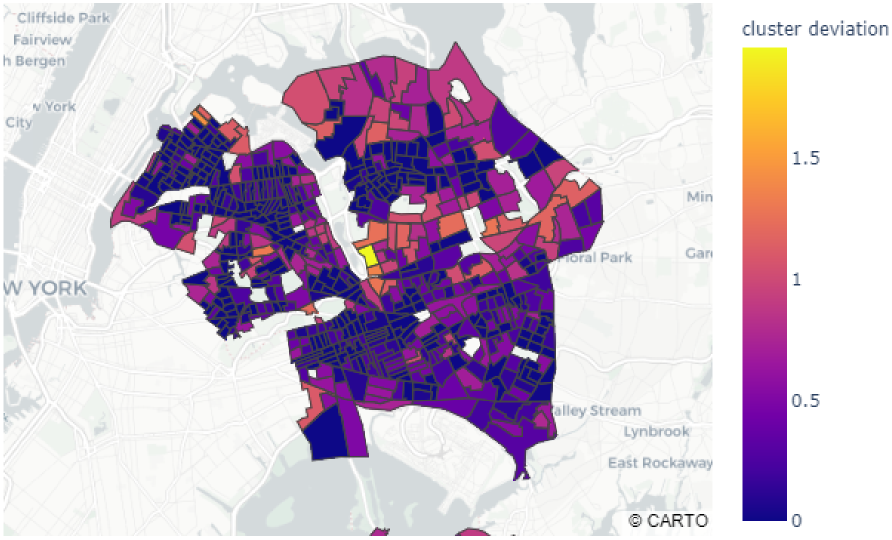


Figure 7: Cluster Deviation of Venue Density

3.4 Regression Analysis

We next created a linear regression model to predict venue density from the demographic data. We ran a grid search with 10-fold cross-validation over regularization parameter and polynomial degree. As shown in Figure 8, none of the models fit the data particularly well, with a max mean R^2 of 0.17. The best combination of high mean R^2 and relatively low standard deviation was the model with polynomial degree 2 and regularization parameter 100. We proceeded to train that model on the entire dataset and then picked out the census tracts with large negative residuals.

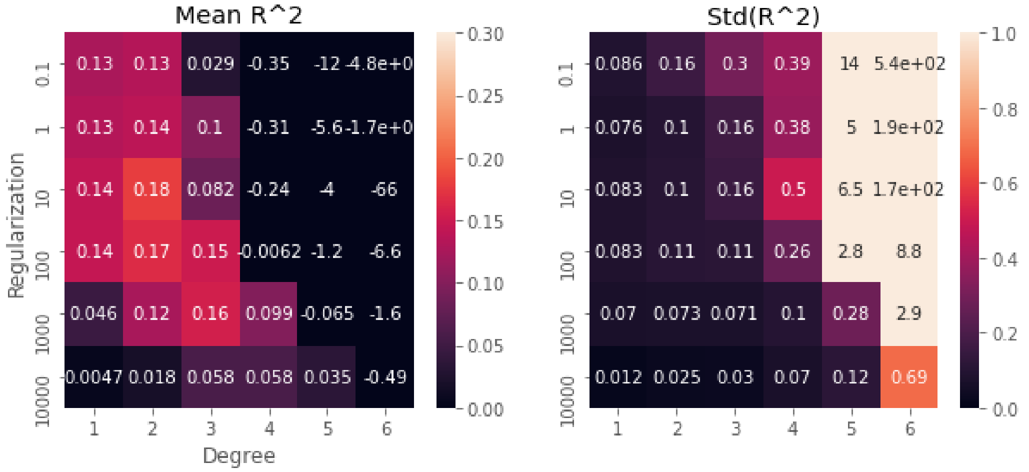


Figure 8: Grid search results

Figure 9 shows the negative residuals for each tract. The residuals show a similar pattern as the cluster deviations, with notable bands running east from the center. For the residuals the cutoff is sharper and fewer scattered tracts appear.

4 Results

Our final result is shown in Figure 10, which shows existing Chipotle locations along with our two measures of unmet demand. According to the Foursquare dataset, there are no Chipotle locations in the northeast quadrant of the map. We conclude that the band running east from Flushing Meadows Park, and the band to the north of that, are excellent candidates for a new location because they show up in both models of unmet demand, and there are no existing locations nearby.

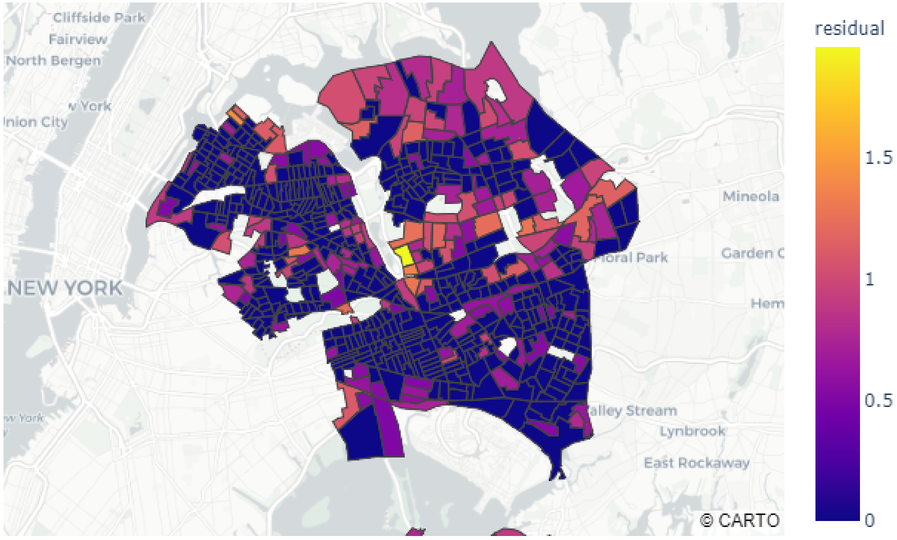


Figure 9: Residuals from linear model

5 Discussion

Making a decision on opening a new location would require more information than was considered here, including company financials, real estate price and availability, and proximity to transportation. Nevertheless, the results presented here should serve as a useful starting point for identifying areas for further research and analysis. Therefore, we recommend the central-eastern band shown in the maps for further research and analysis for opening a new location.

6 Conclusion

We used demographic data from the US Census Bureau and venue data from Foursquare to predict demand for Chipotle. We used the supply of venues with similar customer bases as Chipotle, based on Foursquare data, as a proxy

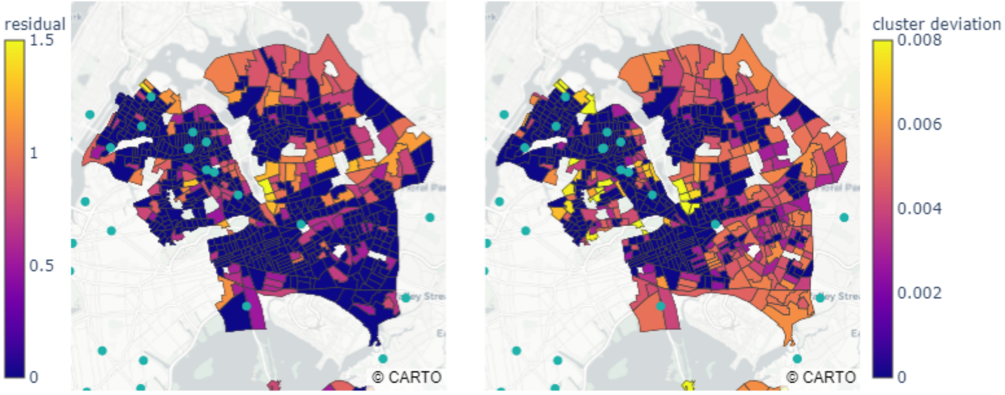


Figure 10: Maps showing existing locations overlayed on models of unmet demand

for demand for Chipotle. We created a k-means clustering model and a linear regression model based on the demographic data. We were successfully able to identify promising areas for a new location based on tracts that deviated from these models and had no existing locations nearby.