Exploring Gentrification and Displacement Through User-Generated Geographic Information

Final Capstone Project Report, NYU CUSP 2019
Sponsor and mentor: Karen Chapple, UC Berkeley
Team Members: Kent Pan, Tiffany Patafio, Manrique Vargas, Jiawen Wan, Tiancheng Yin

Abstract

Gentrification and displacement are pressing issues for many cities today, as urban populations continue to grow and neighborhoods change rapidly in response. While gentrification can bring new businesses, resources, and other positive changes to a neighborhood, the rapid change can be destabilizing to long-term, low-income residents already in the area. Thus, it is important to better understand the activities and behaviors in these changing areas to better provide insight for the local community and public officials. This study utilizes two such novel data sources, Twitter and Foursquare, to explore gentrification and displacement risk for neighborhoods within the 31-county NY metro region. Using methodology established by the UC Berkeley Urban Displacement Project, and expanding on last year's CUSP capstone project, we utilize both administrative census data and user-generated social media data, to model the gentrification phenomena and in changing neighborhoods. We show that Foursquare and Twitter have the potential of improving prediction of gentrification, but their power for modelling gentrification alone is still weak compared to benefits derived from Census data. We propose alternative definitions of gentrification related to Supergentrification, people gentrification (using college education and income) and place gentrification (using housing and rent price). With these alternative, more specific definitions of change, we were able to see more impact from Foursquare and Twitter datasets on certain change types. Our results show that it is easier to model 'people' or 'place' gentrification than combining the two.

NYU | Center for Urban Science + Progress

Table of Contents

1. Introduction

1.1. Problem Statement

Gentrification can be defined as "simultaneously a spatial and social practice that results in the transformation of a working-class or vacant area of the central city into middle- class residential or commercial use" (Lees, 2013). While gentrification can bring new businesses, resources, and other positive changes to a neighborhood, the rapid neighborhood change associated with gentrification can be destabilizing to long-term, low-income residents already in the area.

The UC Berkeley Urban Displacement Project (UDP) developed a gentrification typology index to quantify and identify the state of change of any given neighborhood. This index, which relies entirely on administrative census data, was applied to the NY metro region by the 2018 CUSP capstone group (Chapple, 2018). This capstone project aims to extend upon that work by using novel data sources from online platforms beyond the typical census data sources, to explore the relationships between modern online data sources and gentrification. We aim to evaluate whether this new data can improve on the existing typology and determine if these sources can support more timely, near real-time predictions.

The large availability of data in online platforms has the potential to add enormous value to urban studies. By measuring these social needs stored in online media servers, a new layer of the city is defined and thus, it is available for analysis and eventually intervention. Ultimately, this will help develop a deeper, more nuanced understanding of different factors that influence gentrification, and will allow city officials and local community members to adequately prepare for future transformations in their neighborhood.

1.2. Literature Review

To predict neighborhood change in London, Reades et al (2018) created a score of ‘socioeconomic status’ by combining four variables: household income, property sale value, occupational share, and qualifications; principal components analysis (PCA) to choose a singular measure of socioeconomic status. Their model is built on the characteristics from the 2001 Census to ‘predict’ the 2011 scores, and then use the same model with the 2011 Census data to predict outcomes in 2021. It is possible to envision useful revisions to their approach to incorporate more ‘timely’ data – such as from Foursquare (business data) or Twitter (useful as a marker of cultural change) – to develop the kind of real-time ‘early warning system’ anticipated by Chapple and Zuk (2016).

In a 2018 study, Chapple, Poorthois and others used Twitter data to understand mobility and visitor patterns in gentrifying neighborhoods, specifically what predicts these visits and whether they can help identify areas at risk of change. The study, originally conducted in San Francisco, used the same gentrification typology and further identified a subset of 108 at-risk tracks with heightened visitor activity-- many of which were adjacent to gentrifying areas.

Real Estate prices are widely used in research to understand neighborhood change. Recent studies from Steif et al (2016) used longitudinal data from the census to predict neighborhood change with regards to home pricing. Using random forests and other machine learning models, the study able to predict home prices with a median error of 8% for 29 US cities. In a separate study, Steif uses changes in home prices, incomes, and bachelor’s attainment as proxies for neighborhood change. In this study, he found that the

house price change regressions had the highest explanatory power, with 12 selected census features explaining 60% of house price changes.

Cerron et al. (2018) used Foursquare data to serve as a tool to describe and operationalize Jan Gehl's traditional observations on activities people engage in public urban space. They extracted Foursquare and Instagram data to describe contemporary urban processes - specifically by implementing indices to better describe experiential qualities and opportunities of urban neighborhoods.

2. Data

2.1. Data Description

Through coordination with our project sponsor, three sources of online user-generated data were initially identified to act as proxies for human activity across different domains: Twitter for mobility, Zillow for housing, and Foursquare for businesses. The Zillow data was ultimately not obtained for the project due to server issues, so analysis was only continued on Twitter and Foursquare data. The table below summarizes the characteristics of the datasets in our project scope. Further detailed information on the individual datasets are described in our previous progress report.

Table 1: Datasets provided by sponsor for use in research and modeling

	Twitter	**Foursquare**	**Census**
Time Range	2012-2015	2009-2018	1990-2016
Geographic Range	Entire study region (31-county NYC metro)	NYC and NYC-adjacent areas only	Entire study region (31-county NYC metro)
Data Type	Point (individual tweets)	Point (individual businesses)	Polygon (census tract)
Attributes	Raw tweet metadata (user, location, and time of tweet).	Business categories, types, class; # of users, check-ins, visits.	Various demographic and housing attributes (income, education, median home value, etc.)
Data processing requirements	Data has been pre-processed to aggregate user data to the census tract level, and extract time-spatial (e.g., # of local and visitor tweets) and demographic features for each tract.	Required additional processing to align with the rest of our data (at polygon/census tract level).	Data from different time periods required processing to reconcile changing census tract boundaries over time.
Limitations	Limited time range of dataset, gentrification is	Only contains 30% sample of businesses in	The longer time range of the census data captures

	a long term process that may not be captured from a limited time range. Geographic range of aggregated data is much more limited to the NYC-adjacent region.	NYC region, assumption that sampling is consistent throughout the region. All data is aggregated to a single value from the entire time range.	the long-term nature of gentrification, but the older features do not directly correspond to the later dates of the other datasets.

All of these datasets are joined to the UDP 2018 typology index identified for each census tract in the study region, for further analysis. Throughout the model development, additional public data sources (including NYC PLUTO data and additional census features) were also obtained to augment certain models, described in Section 3. In cases where the data availability was limited, the particular model was restricted to the NYC region only. More detailed information about the data, including feature descriptions, are included in **Appendix A**.

2.2. Data Exploration

An initial round of exploratory analysis was done on the datasets to gain an understanding of the profile of the data, distribution of the data across our target geography, and identify attributes of the data that can be feature engineered to include into our models. Detailed clustering, correlation, and visual analysis were conducted to achieve these goals.

Foursquare data was initially clustered by income level, with the results shown in **Appendix B**. High income neighborhoods were found to have more businesses in general, as well as a higher number of businesses per person. In addition, certain business types were found in different proportions between high vs low income areas. For example, donut shops and fast food restaurants made up a significantly larger proportion of business in low income neighborhoods, while coffee shops and Italian restaurants were more prevalent in high income areas.

In gentrifying areas specifically, some business types (such as sandwich shops and Mexican restaurants) were proportionally more significant. Along the spectrum of gentrifying typologies, certain businesses, like coffee shops and bars, became noticeably more prevalent with further gentrification. Looking at Foursquare data classes (necessary, optional) and categories (nutrition, social, consumption) more broadly, similar trends occur with increased gentrification -- decreasing percentage of necessary businesses, with increasing optional businesses and decreasing nutritional percentage with increasing social businesses.

Similarly, Twitter data was clustered using the percentage of tweets by each user type (neighbor, local, visitor) and 5 clusters, identified as the optimal using the elbow method. These clusters identified different profiles of twitter behavior, and were then compared with the UDP typologies to identify potential patterns in user behavior in typologies of interest.Most tracts associated with gentrifying neighborhoods were found to have predominantly local tweets, with increasing visitor tweets as they move towards exclusion. Additional details of these results are available in **Appendix C**.

3. Methodology

3.1. Feature Engineering & Selection

Although the provided Twitter and Foursquare data contained various features that would be important inputs to the model, the exploratory analysis revealed underlying trends that would be valuable to make explicit in the models. Techniques to extract additional features included calculation of distances to particular Foursquare business types, calculation of visitor demographics in the Twitter data, and standardizing by census tract population, as well as statistical techniques including principal component analysis and nearest neighbor analysis. Further details of the various techniques are documented in **Appendix D**.

3.2. Initial Modeling

The initial round of models focused on attempting to model the spectrum of gentrification, using multiclass classification to predict all eight gentrification typologies in the original UDP index. The chosen techniques for each model iteration included support vector machines, logistic regression, decision trees, and random forest models. The first baseline model used census features identified by the previous year's capstone group and UDP. Two separate models, described in Table 2 below, were built to evaluate the impact of adding in Foursquare and Twitter data, respectively, to evaluate if the new data improved the accuracy of the model. The F1 score was chosen as the primary metric for evaluating our model performance to capture the true false positive/negative rates of the model.

Table 2. Initial multiclass classification model approach.

Model	**Data**	**Target Variable**
Model 1A: Census data (baseline)	Census features	8-part gentrification & typology index, as provided by the UDP
Model 1B: Census data + Foursquare	Census features with Foursquare features	
Model 1C: Census data + Twitter	Census features with Twitter features	

The gentrification typologies were difficult to capture through this model design, with low and irregular F1 scores and only slightly better performances when compared to the baseline model. Based on these results, we re-evaluated the modeling approach of the project through additional literature review and feedback from our project sponsor/mentor. Two major changes were made to address the poor performance of the models. First, we shifted from a multiclass classification approach to a binary classification approach - this simplification would target only the ongoing and advanced gentrification tracts, as successfully identifying gentrifying and changing areas is of much greater interest than identifying areas that are already stable and generally known to not be changing.

Second, we started to delve into different metrics of gentrification, namely the distinction between people gentrification, place gentrification, and hybrid gentrification approaches. The "people gentrification" idea as argued by Ellen & O'Regan (2011) states that gentrification is driven mainly by new high-income, higher-education populations, and results in new housing demand for neighborhood real estate. The "place gentrification" theory, as argued by Smith (1979) states that gentrification is driven not by people, but primarily by land & housing markets - real estate investment and development occurs initially and then sets off the process of neighborhood change. A hybrid approach (as adopted by the UDP), adopts and combines elements from the two separate ideas. Based on these distinct ideas, we adopt several binary target variables to capture these distinct variations on the gentrification concept.

3.3. Revised Models - Binary Approach

Seven binary variables were chosen to represent the different aspects of gentrification, shown in Table 3 below.

Table 3. The selected seven binary target variables for the revised modeling approach.

Target Variable #	Variable Description	Gentrification Type	Description
1	% change in college education population	People gentrification	2000-2016 time range
2	% change in household income (2000-2016)		2000-2016 time range
3	% change in median home value	Place gentrification	2000-2016 time range
4	% change in median rental cost		2000-2016 time range
5	Ongoing and advanced gentrification (1990-2016)	Hybrid gentrification	1990-2016 time range, from UDP gentrification/typology index
6	Ongoing and advanced gentrification (2000-2016)		2000-2016 time range, from UDP; includes tracts that gentrified after 2000 only, to align more closely with modern data sources
7	"Super gentrification"		1990-2016 time range, from UDP; includes tracts that gentrified after 1990 and currently have household incomes > $200,000

The distribution of the census tracts with these new target variables are shown in **Appendix E**. Each of these target variables were modeled using census features, including both those used to define last year's typologies and additional features identified through literature review, as a baseline. As in the initial round of modeling, Foursquare and Twitter data were added individually to models to understand the impact of these datasets on the identified target variables. A combined model to evaluate the impact of both datasets simultaneously was also added, and all model variations are described in Table 4 below.

Table 4. Revised models for binary classification

Model Group	**Data**	**Target Variables**
Model 2A: Census data	Expanded Census features	7 binary target variables as described in Table X
Model 2B: Census data with Foursquare	Expanded Census features with Foursquare features	
Model 2C: Census data with Twitter	Expanded Census features with Twitter features	
Model 2D: Census data with Foursquare and Twitter	Expanded Census features with selected Foursquare and Twitter features	

4. Results

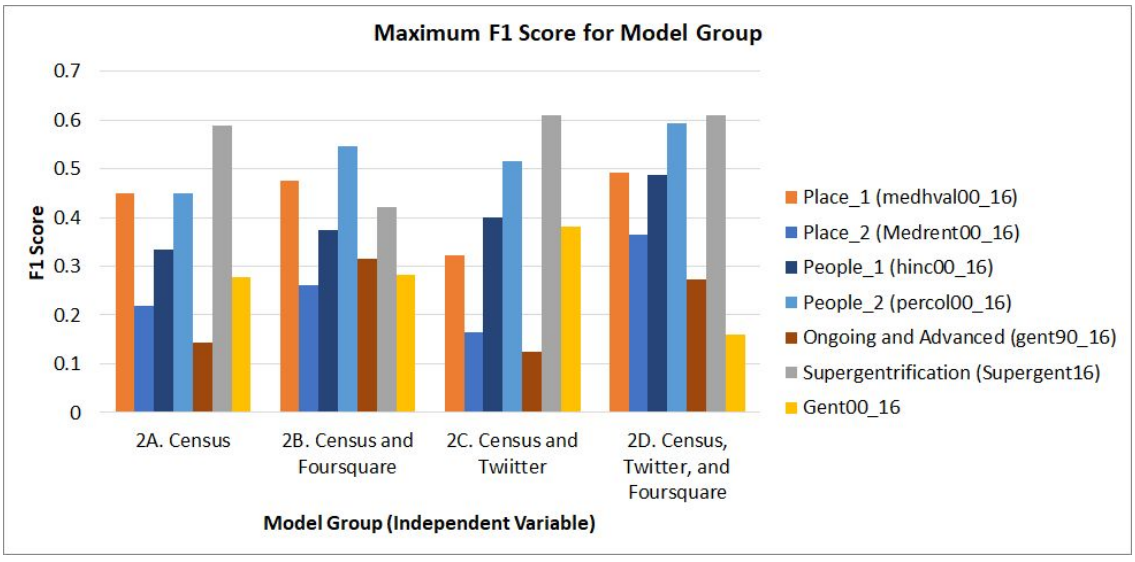


Figure 1. Maximum F1 Score (All models were reduced to Greater New York Metro Region)- The F1 scores above represent the scores of our best performing models for each featureset and target variable combination. Interestingly, Twitter's greatest improvements beyond census are on people change variables whereas Foursquare also outperforms census alone for place change variables.

Figure 1 displays the F1 scores for each model group described in Table 4. Further results are discussed for each model group below.

4.1. Modeling Typologies using Census only

Census features proved most robust for some of the Decision Tree models, but were particularly poor across all model types in predicting the Ongoing and Advanced Gentrification (1990-2016) variable. While the census featureset decision tree models for Ongoing and Advanced Gentrification (2000-2016) and percentage change in Median Home Value, it was outperformed by models which combined other data for every variable. Consistently, household income from 2000 was consistently amongst the top weights, although the percent of renters from the 2000 census seemed to be more important in the modelling Ongoing and Advanced Gentrification between 2000 and 2016.

4.2. Modeling Typologies using Census and Foursquare Data

As shown in the model feature importance in Appendix E, the top 10 most relevant features were median housing value, housing income, percentage college education, median rent, car commuters, median rent, distance to American Restaurants, percentage of renters, Boat Rental business types neighbours (refer to Appendix D for more details regarding Foursquare feature engineering). It can be seen that the improved features became important and that our feature engineering was effective. Again, decision trees and logistic regression performed the best.

4.3. Modeling Typologies using Census and Twitter Data

Models with Twitter features consistently outperformed Census data alone, with the exception of On-going and Advanced Gentrification between 2000 and 2016. The improvements from Twitter data were particularly marked for Supergentrification and people gentrification indicators, percent change in income and college education. Decision Trees performed the best overall for Twitter modelling, with Random Forest and Logit performing well for only certain featuresets.

	Featureset			
Binaryvariable	Census Only	Census + Twitter Raw	All + Distance Features	All + Derived Twitter Features
Gentrification (90-16)	0.0000	0.1439	0.1878	0.0000
% Change College Educated	0.2857	0.3008	0.3129	0.5213
% Change Housenhold Income	0.0124	0.2816	0.1961	0.4592
% Change Med Home Value	0.2883	0.4154	0.2178	0.2105
% Change Med Rent	0.0000	0.0233	0.0400	0.2901
Gentrification (00-16)	0.3883	0.0000	0.3802	0.0000
Supergentrification (00-16)	0.6341	0.6111	0.7143	0.7027

Figure 2: The figure above shows Fscores for Decision Tree models for all variable, featureset combinations. While generally F1 scores were low, Decision tree models of Supergentrification consistently outperformed other models across featuresets, reaching above a .7 test F-score once the complete set of Twitter raw and engineered features had been added. Household Income change showed the biggest improvement with the inclusion of all twitter features with a F-score improvement of over 3500% from census alone.

More interesting than the pure improvements from Twitter data, which we expected when beginning our modelling journey, are the drivers of these improvements. Looking at feature weights, we can begin to uncover the features, and therefore behaviors, apparent in changing neighborhoods. For example, in neighborhoods identified as having people gentrification, we see indications that there is also strong visitor patterns in the percentage of highly educated visitors to that area. (See Figure below)

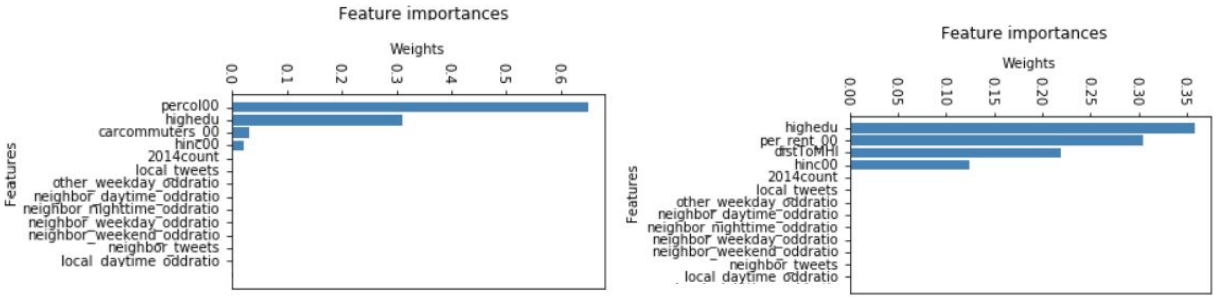


Figure 3: The plots above show feature weights for two of our strongest performing Decision Tree models on people gentrification indicators, with percent change in college educated on the left and percent change in income on the right. In neighborhoods where the percentage of college educated individuals has increased most between 2000 and 2016, we see indications that the percentage of highly educated visitors to that area has also increased. Unsurprisingly, these highly educated visitors also appear as the most important feature for identifying areas where income has increased the most over the 2000 to 2016 time period, followed by percent of renters and distance to Medium-High Income tracts.

The total number of local users and number of neighbor tweets were important weights in improving the modelling of On-going and Advanced Gentrification between 1990-2016-- this would align to research on gentrifying neighborhoods since active Twitter usage is often associated with the age range typical of gentrifying neighborhoods and visitors from surrounding neighborhoods would become more frequent as the neighborhood improves. For Supergentrification, presence of local tweeters consistently proved important to the model. Visitor tweets ("other tweets") proved to be amongst the most influential weights from Twitter across census and Twitter models-- which was expected given our earlier findings that visitor tweets increase in exclusion neighborhoods.

4.4. Modeling Typologies using Foursquare, Census and Twitter Data

The combined model feature importance is dominated by Census data. However, Foursquare enhanced features do emerge among the top ten most important features, outperforming Twitter data. This was found consistently throughout other test cases using PCA and considering commercial areas in Manhattan only. As the data became more non-linear, decision trees performance decreased, SVM performance increased and logistic regression performed the best.

5. Conclusions

Though the difference in time frames of our datasets was a challenge, the different time spans actually allow us to understand more about the phases of change in a neighborhood. The changes in census data

over a longer timespan tell us about overall changes in trends, while Twitter and Foursquare data helps us see the impact of those changes on actual people behavior and businesses in the neighborhood. This combination has allowed us to pull together more granular insights on human behavior in changing neighborhoods to add to the collection of gentrification research, like the types of visitors frequenting high change areas and the most important business types.

The modelling has also shown us the power of being specific in defining neighborhood change, with typologies revolving around single factors and the very specific "Supergentrification" typology far outperforming the models which tried to combine all factors in a single variable. This methodology also allowed us to look at people and place separately which highlighted the value Foursquare and business data can bring to place gentrification analysis, whereas Twitter's biggest improvements were unsurprisingly related to people gentrification. Our results show that it is easier to model people or place than combining the two. This suggests that the phenomenon is complex, and people-place gentrification are not necessarily linked in a simple way as the typology maps originally seem to imply.

We show that Foursquare and Twitter have the potential of improving prediction of gentrification, but their contribution is still weak compared to benefits derived from Census data. As online sources and social media platforms such as Foursquare and Twitter continue to amass data, the long term changes and trends in behavior can be compared to the parallel changes we see in census data for the same period more holistically. These sources can not only help us understand the impact of change, but might be able to identify specific types of change in its early stages so that early tactics proven to support the development of a thriving neighborhood for all can be put in place. We believe that as Foursquare and Twitter continue to accumulate data and record rates of change throughout time, models including Foursquare and Twitter may perform better in the future.

References

1. Cerrone, Damiano & Lehtovuori, Panu & López Baeza, Jesús. (2018). Integrative Urbanism: Using Social Media to Map Activity Patterns for Decision-Making Assessment. 10.13140/RG.2.2.24650.36802.
2. Chapple, K., Ate, P,. Matthew,. Z, Eva,. P (2019). Monitoring Streets through Tweets: Using User-generated Geographic Information to Predict Gentrification and Displacement.
3. Chapple, K.; Bianco, Federica, Kleinman, N.; Sabolevsky, S.; Chermesh Reshef, D.; Xi, H.; Rodrigo Vazquez, G.; Hambardzumyan, R. (2018). Map of Gentrification and Displacement for The Greater New York. NYU, CUSP.
4. Chapple, K and Zuk, M (2016). Forewarned: The use of neighborhood early warning systems for gentrification and displacement. Cityscape: A Journal of Policy Development and Research 18(3): 109–130.
5. Ellen, Ingrid Gould & O'Regan, Kathy. (2011). How Low Income Neighborhoods Change. NYU Furman Center for Real Estate and Urban Policy.
6. Hamnett, Chris. *Unequal City: London in the Global Arena*. Routledge, 2005.
7. James, G., D. Witten, T. Hastie and R. Tibshirani (2013). An introduction to statistical learning. London: Springer, 102, pp.303–368.
8. Lees, L., Slater, T., and Wyly, E. (2013). Gentrification. Routledge.
9. López Baeza, Jesús & Cerrone, Damiano & Männigo, Kristjan. (2017). Comparing two methods for Urban Complexity calculation using Shannon-Wiener index. WIT Transactions on Ecology and Environment. 226. 369-378. 10.2495/SDP170321.
10. Luc Anselin and Sergio J. Rey and Wenwen Li. (2014). Metadata and provenance for spatial analysis: the case of spatial weights. International Journal of Geographical Information Science. 28, 11, 2261-2280, 2014, Taylor & Francis. https://doi.org/10.1080/13658816.2014.917313
11. Reades, J., De Souza, J., & Hubbard, P. J. (2018). Understanding urban gentrification through Machine Learning: Predicting neighbourhood change in London.
12. Smith, Neil. (1979) Toward a Theory of Gentrification A Back to the City Movement by Capital, not People, Journal of the American Planning Association, 45:4, 538-548, DOI: 10.1080/01944367908977002
13. Steif, Ken. A. M. (2016). *Predicting gentrification using longitudinal census data.* Philadelphia: http://urbanspatialanalysis.com/
14. Zukin, Sharon. *Loft Living: Culture and Capital in Urban Change*. Rutgers University Press, 2014.

Appendix A: Description of Datasets

UDP Typology Criteria

Table 1

Typology	Typology Criteria
Not Losing Low-Income Households (Low Income)	• Pop in 2000>500 • Low Income Tract in 2016 • Not classified as At Risk or Ongoing Gentrification or Displacement
At Risk of Gentrification (Low Income)	• Pop in 2016 > 500 • Low Income Tract in 2016 • Vulnerable in 2016 (Defined in Appendix) • "Hot market" from 2000 to 2016 • Not currently undergoing displacement or ongoing gentrification
Ongoing Displacement of Low-Income Households (Low Income)	• Pop in 2000 > 500 • Low Income Tract in 2016 • Loss of LI households 2000-2016 (absolute loss) • Few signs of gentrification occurring
Ongoing Gentrification (Low Income)	• Pop in 2000 or 2016 > 500 • Low Income Tract in 2016 • Gentrified in 1990-2000 or 2000-2016 (Defined in Appendix)
Advanced Gentrification (Moderate to High Income)	• Pop in 2000 or 2016 > 500 • Moderate to High Income Tract in 2016 • Gentrified in 1990-2000 or 2000-2016 (Defined in Appendix)
Stable Exclusion (Moderate to High Income)	• Pop in 2000>500 • Moderate to High Income Tract in 2016 • Not classified as Ongoing Exclusion
Ongoing Exclusion (Moderate to High Income)	• Pop in 2000 > 500 • Moderate to High Income Tract in 2016 • Loss of LI households 2000-2016 (absolute loss) • LI migration rate (percent of all migration to tract that was LI) in 2016 < in 2009
Super Gentrification or Exclusion (Very High Income)	• Pop in 2000 > 500 • Median household income > 200% of regional median in 2016 • Indicators of Gentrification or Exclusion

Twitter Data Details

Table 2.1: Twitter Data Census Tract Level Features

Field Name	Description

total_tweets	The number of tweets in a tract
count_missing_ht	The number of tweets with N/A home_tract
miss_home_tract	The ratio of tweets with N/A home_tract (missing home_tract) to the total_tweets
total_users	The number of users in a tract
userwith5_count	The number of users with more than 5 tweets in a tract
tweets_hometract	The number of tweets with a home_tract This is total_tweets – count_missing_ht
tweets_sent_from_home	The number of tweets where home_tract = tract This is the same as local_tweets
percent_of_tweets_ht	The ratio of tweets with a home_tract to the total distribution of the dataset (all 75368830 tweets in the dataset)
percent_of_local_tweets_byalltwitter	The ratio of tweets where home_tract = tract to the total distribution of the dataset (all 75368830 tweets in the dataset)
percent_of_local_tweets _bytract	The ratio of tweets where home_tract = tract to the total tweets in a tract
percent_of_users	The ratio of users in a tract to the total users in the whole distribution (163724 unique users)
percent_of_users_5_tweet	The ratio of users with more than 5 tweets in the tract to the total users in the tract (userwith5_count/total_users)

Table 2.2: Twitter Data Census Time Level Features

Field Name	Descriptio

oddratio_12AM to oddratio_11PM	The ratio of tweets that occurred in the hour mentioned in the variable name in the tract to the total tweets in the tract. There's an odd ratio for every hour from 12AM – 11PM.
Monday_oddratio to Sunday_oddratio	The ratio of tweets that occurred on the day mentioned in the variable name in the tract to the total tweets in the tract. There's an odd ratio for each day from Monday to Sunday.
weekend_oddratio	The ratio of tweets that occurred on a weekend (from Friday 8PM until Monday 3:59AM) in the tract to the total tweets in the tract
weekday_oddratio	The ratio of tweets that occurred on a weekday (from Monday 4AM to Friday 7:59PM) in the tract to the total tweets in the tract
nighttime_oddratio	The ratio of tweets that occurred from 7PM to 3:59AM in the tract to the total tweets in the tract
daytime_oddratio	The ratio of tweets that occurred from 4AM to 6:59PM in the tract to the total tweets in the tract
local_tweets	The count of tweets where home_tract = tract for that tract This is equal to tweets_sent_from_home
local_oddratio_12AM to local_oddratio_11PM	The ratio of local tweets that occurred in the hour mentioned in the variable name in the tract to the total tweets in the tract. There's an odd ratio for every hour.
local_Sunday_oddratio to local_Saturday_oddratio	The ratio of local tweets that occurred on the day mentioned in the variable name in the tract to the total tweets in the tract. There's an odd ratio for each day.
local_weekend_oddratio	The ratio of local tweets that occurred on a weekend (from Friday 8PM until Monday 3:59AM) in the tract to the total tweets in the tract
local_weekday_oddratio	The ratio of local tweets that occurred on a weekday (from Monday 4AM to Friday 7:59PM) in the tract to the total tweets in the tract
local_nighttime_oddratio	The ratio of local tweets that occurred from 7PM to 3:59AM in the tract to the total tweets in the tract

local_daytime_oddratio	The ratio of local tweets that occurred from 4AM to 6:59PM in the tract to the total tweets in the tract

Table 2.3: Twitter Data Census Spatial Level Features

Field	Description
neighbor_tweets	The count of tweets where the variable tract is in the neighboring tract of the home_tract
other_tweets	The count of tweets where home_tract != tract and tract is not a neighboring tract of home_tract for that tract (total_tweets – local_tweets – neighbor-tweets = other_tweets)
percent_of_other_tweets_pertract	The ratio of other tweets (defined above) in the tract to total tweets in a tract
percent_of_neighbor_tweets_pertract	The ratio of neighbor tweets in the tract to total tweets in a tract
neighbor_weekend_oddratio	The ratio of neighbor tweets sent on the weekend in the tract to the total tweets in a tract
neighbor_weekday_oddratio	The ratio of neighbor tweets sent on a weekday in the tract to the total tweets in a tract
neighbor_nighttime_oddratio	The ratio of neighbor tweets sent at night in the tract to the total tweets in a tract
neighbor_daytime_oddratio	The ratio of neighbor tweets sent during the day in the tract to the total tweets in a tract

other_weekend_oddratio	The ratio of other tweets sent on the weekend in the tract to the total tweets in a tract weekend_oddratio –local_weekend_oddratio-neighbor_weekend_oddratio = other_weekend_oddratio Note: local_weekend_oddratio + local_weekday_oddratio + neighbor_weekend_oddratio + neighbor_weekday_oddratio + other_weekend_oddratio + other_weekday_oddratio = 1
other_weekday_oddratio	The ratio of other tweets sent on a weekday in the tract to the total tweets in a tract weekday_oddratio –local_weekday_oddratio-neighbor_weekday_oddratio = other_weekday_oddratio
other_nighttime_oddratio	The ratio of other tweets sent at night in the tract to the total tweets in a tract nighttime_oddratio –local_nighttime_oddratio-neighbor_nighttimeoddratio = other_nighttime_oddratio Note: local_nighttime_oddratio + local_daytime_oddratio + neighbor_nighttime_oddratio + neighbor_daytime_oddratio + other_nighttime_oddratio + other_daytime_oddratio = 1
other_daytime_oddratio	The ratio of other tweets sent during the day in the tract to the total tweets in a tract daytime_oddratio –local_daytime_oddratio-neighbor_daytime_oddratio = other_daytime_oddratio

Foursquare Data Details

Table 3.

Field Name	Description
id	A unique string identifier for this venue
Place Name	The best known name for this venue.
Rating	Numerical rating of the venue (0 through 10). Not all venues will have a rating.
Checkins	Amount of times that a venue has been visited using the ‘check-in’ button in the apps Foursquare or Swarm.
Users	Amount of (different) Foursquare users that have visited the venue.
Visits	Amount of times that a venue has been visited, according to GPS tracking and third-party data sources.
Pricing	An object containing the price tier from 1 (least pricey) - 4 (most pricey)
Type	Types of the venue provided by venue owner
Category	11 Categories based on SPIN Unit Urban Activity Wheel
Class	Activities classified by Optional and Necessary. Being Optional are those performed when one has free time (gym, museum, spa, etc.), Necessary, those integrated in the daily routine (school, office, market, etc.).
geometry	Geographic coordinates of this venue

Appendix B: Exploratory Analysis of Foursquare Data

The Foursquare data provided is a sample dataset, featuring a mix of high and low income census tracts with more high income tracts represented. In this dataset, we find many more businesses in these high income areas:

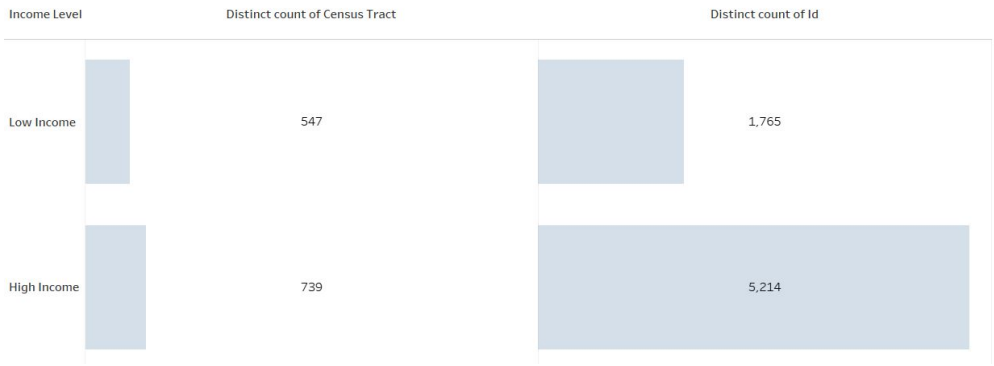


Figure B.1: The chart above shows the number of distinct census tracts and business IDs by high and low income census tracts. You can see that while the split of census tracts are close, there are many more businesses from high income areas in the dataset.

The following visualizations show the average percentage of businesses within each class (Figure A.2) and category (Figure A.3) for all census tracts within the typology, relating to the discussion in the results section above.

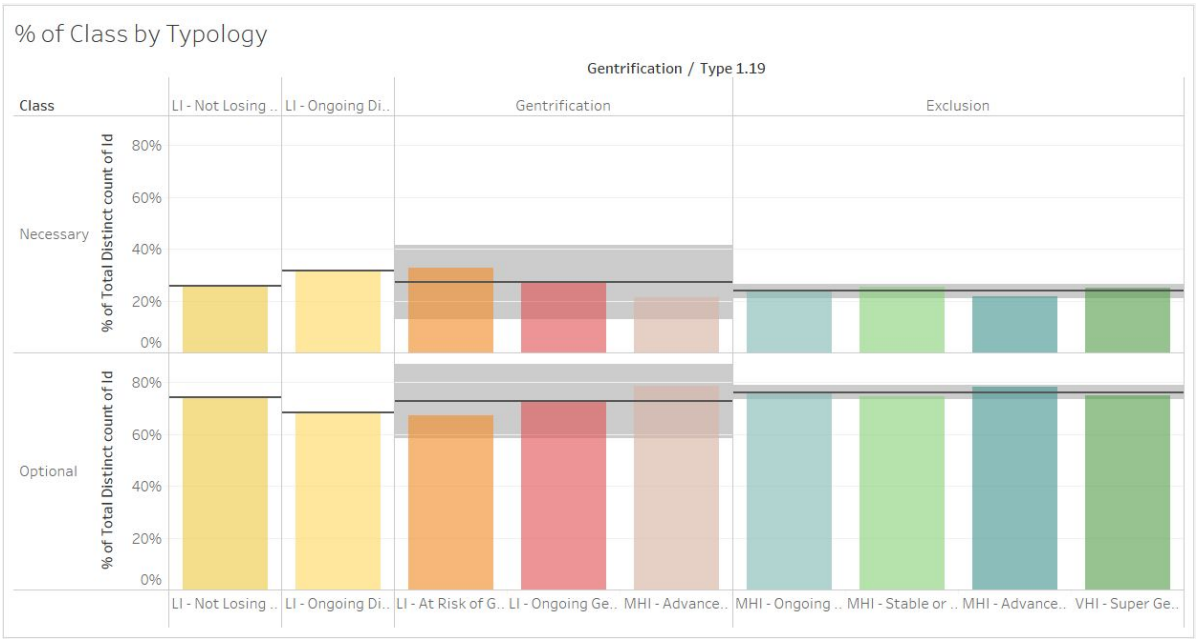


Figure B.2: Necessary and Optional business types vary greatly across the gentrifying typologies, with necessary businesses decreasing to below exclusion areas and optional increasing to levels above.

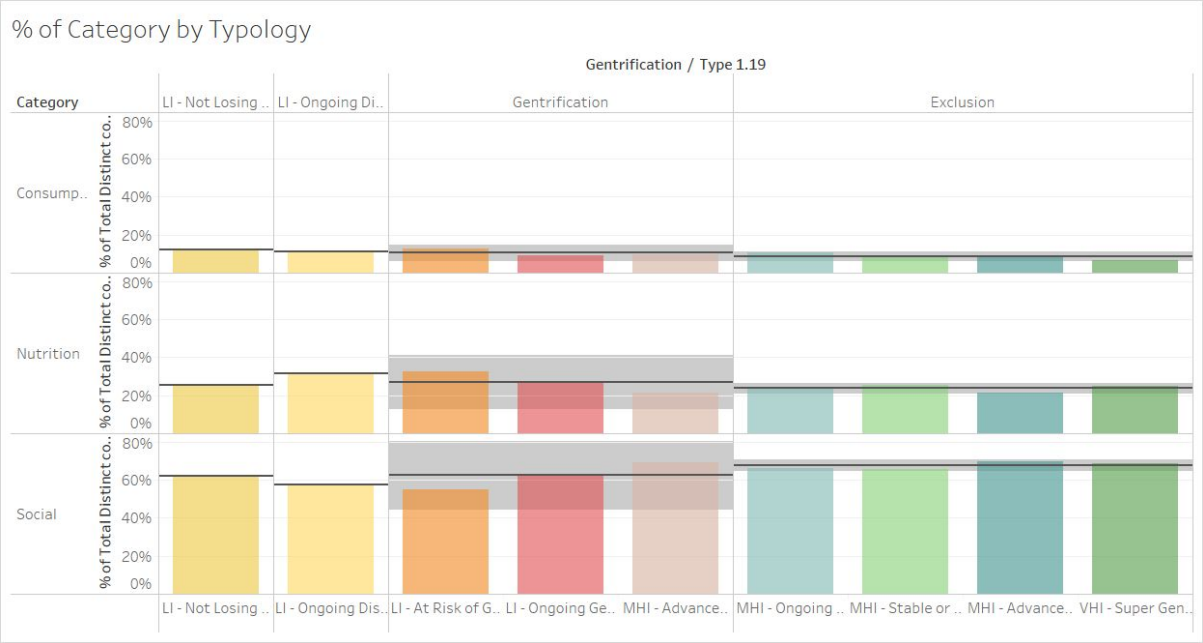


Figure B.3: Consumption businesses represent similar percentages across typologies, while nutrition and social show distinct patterns across the spectrum of gentrifying neighborhoods with social increasing and nutrition decreasing.

The following figures show category and class by income based on clustering of census tracts based on income into 5 groupings. These results have also been represented as business per person by dividing business by the population of the census tract.

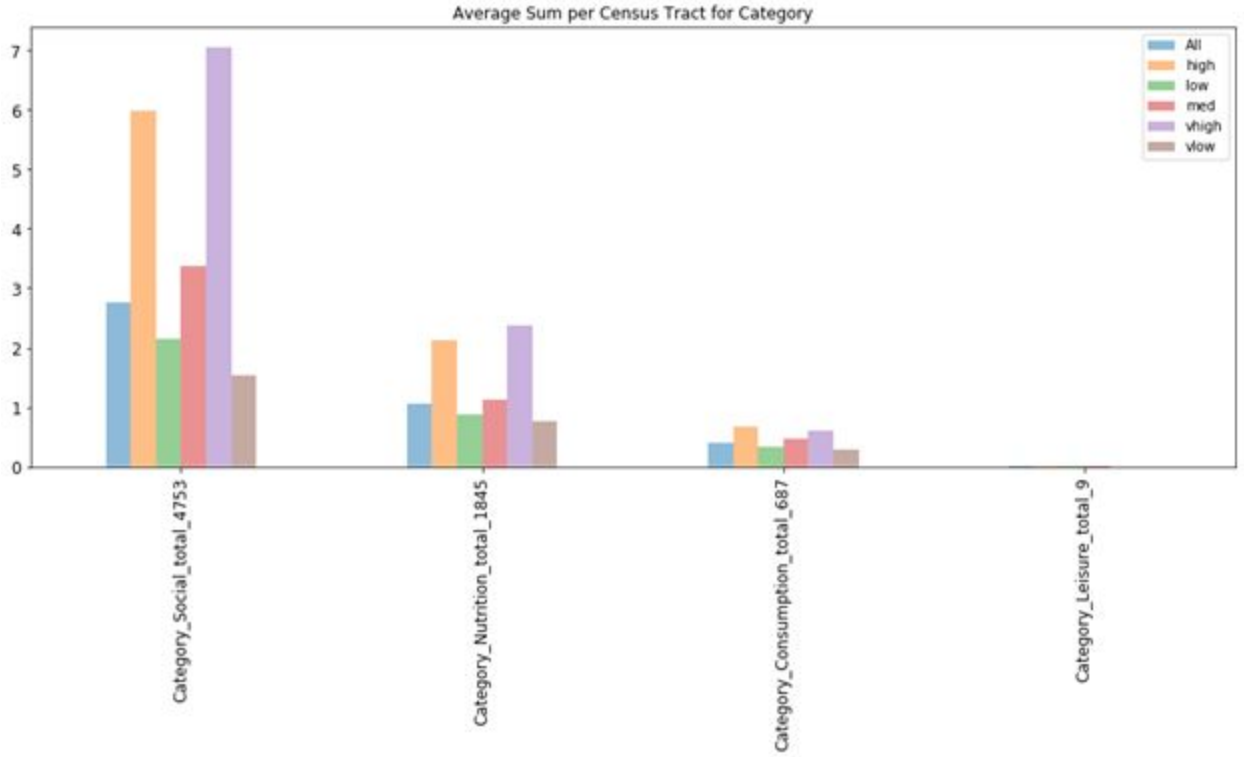


Figure B.4: Distribution of total average sum of 'Category' types per Census Tract per income level

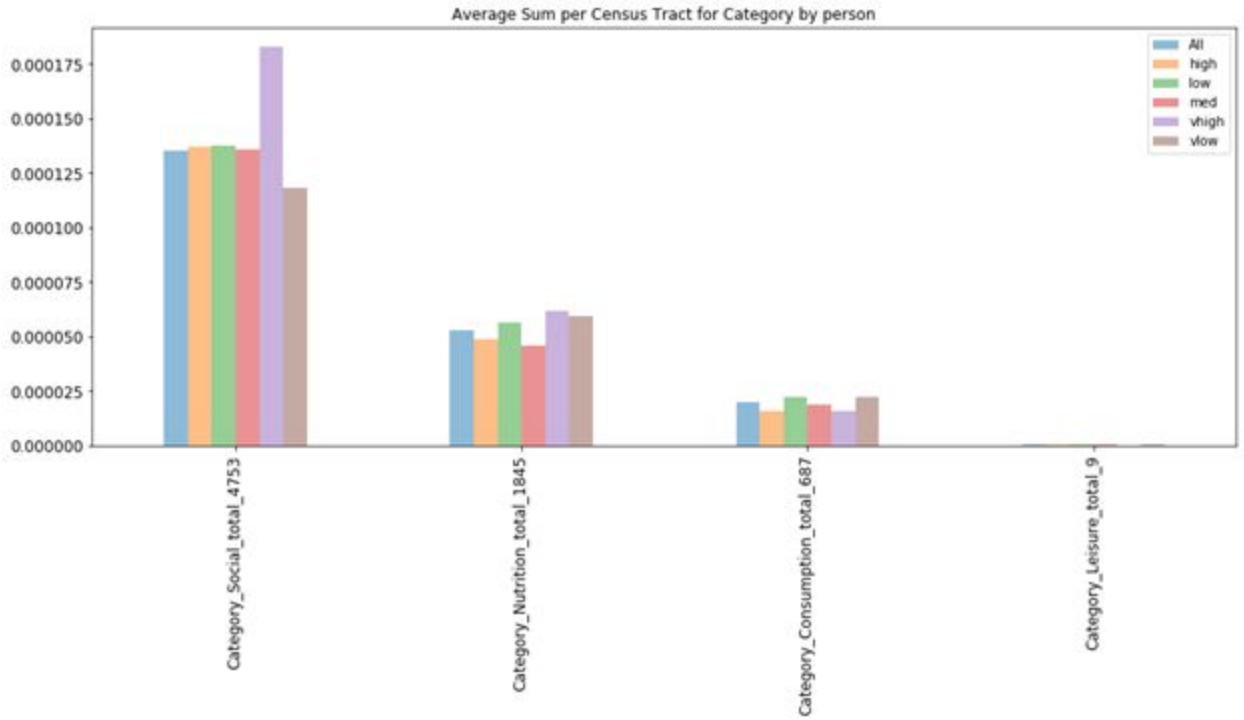


Figure B.5: Distribution of total average sum of 'Category' types by person per income level

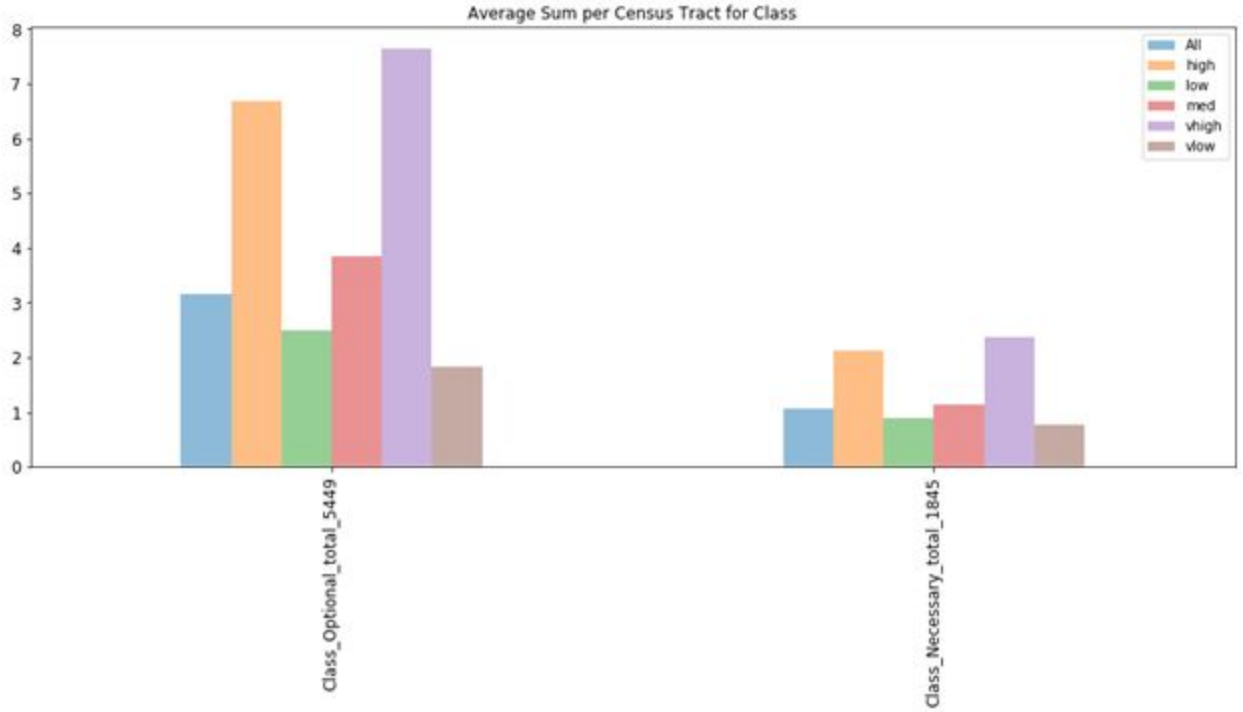


Figure B.6: Distribution of total average sum of 'Class' types per Census Tract per income level

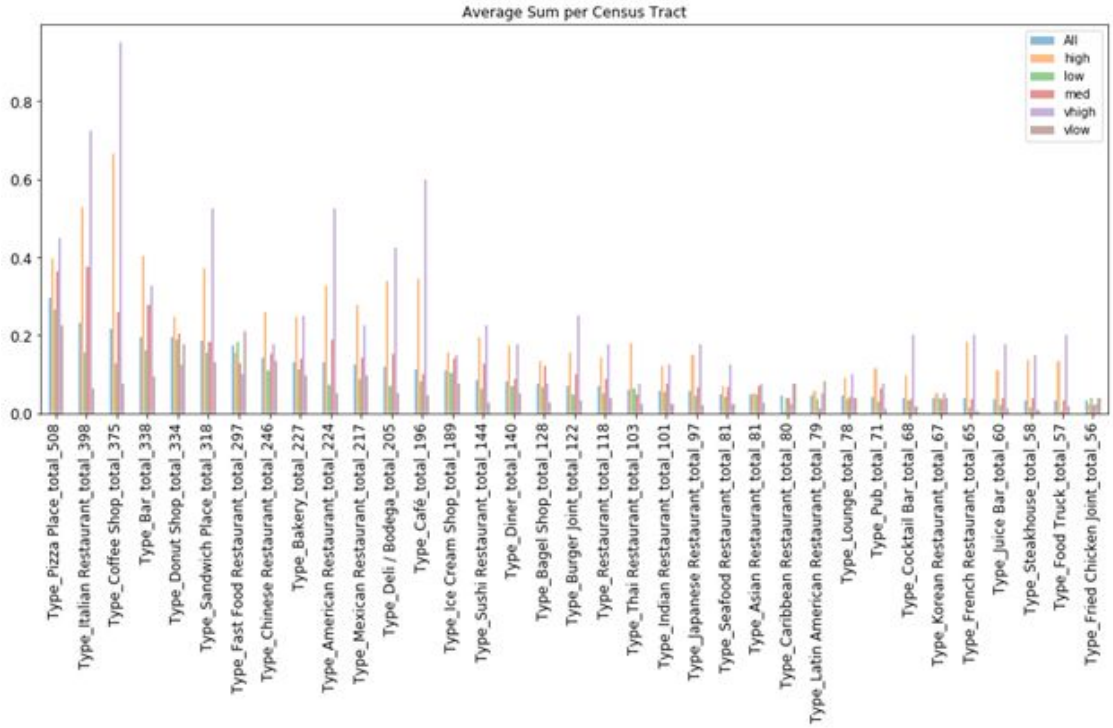


Figure B.7: Distribution of total average sum of 'Type' per Census Tract per income level

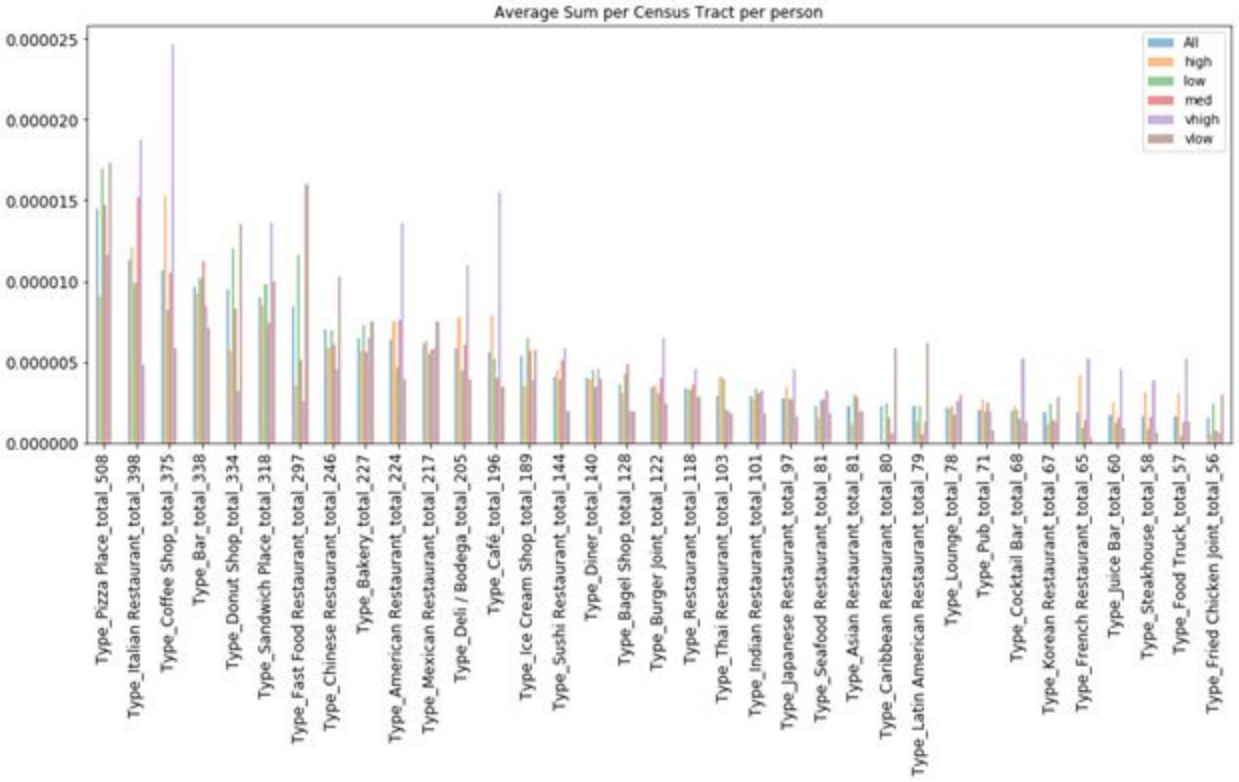


Figure B.8: Distribution of total average sum of 'Type' per Census Tract per income level

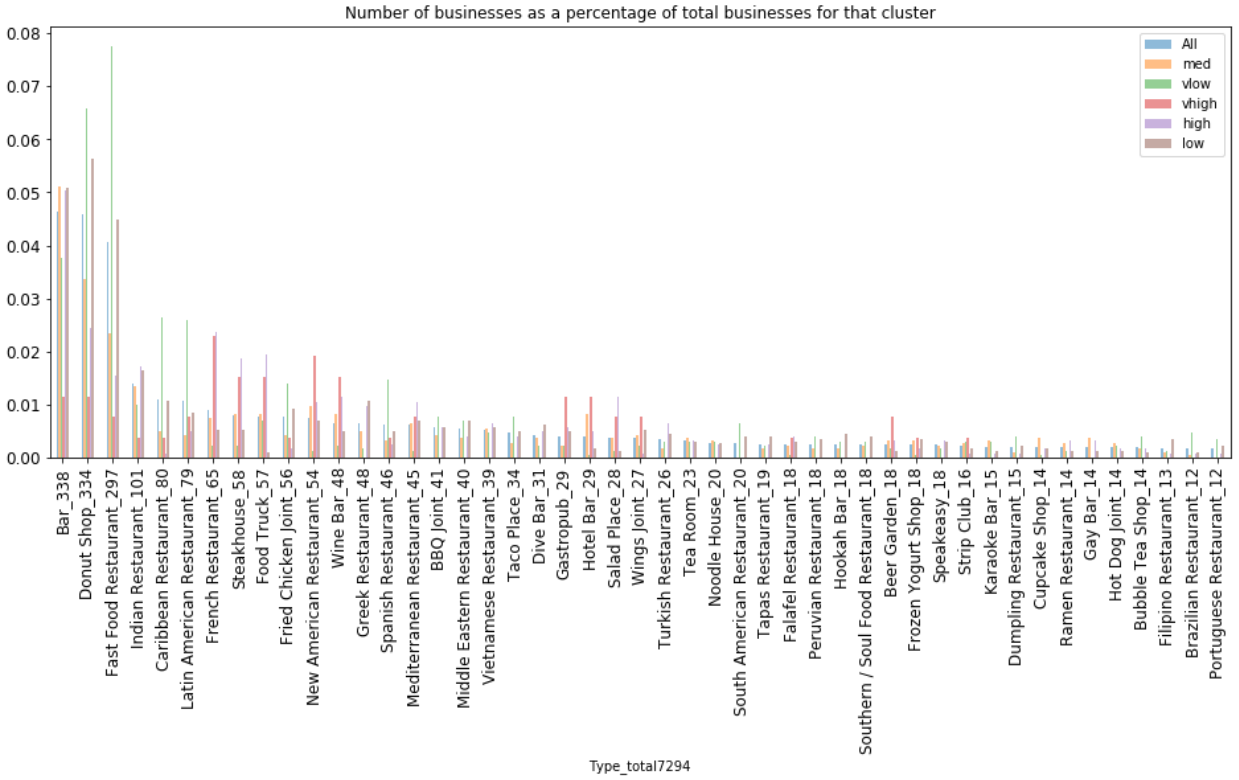


Figure B.9: Distribution of total of business 'Type' as a percentage of the total number of businesses

Mean Values

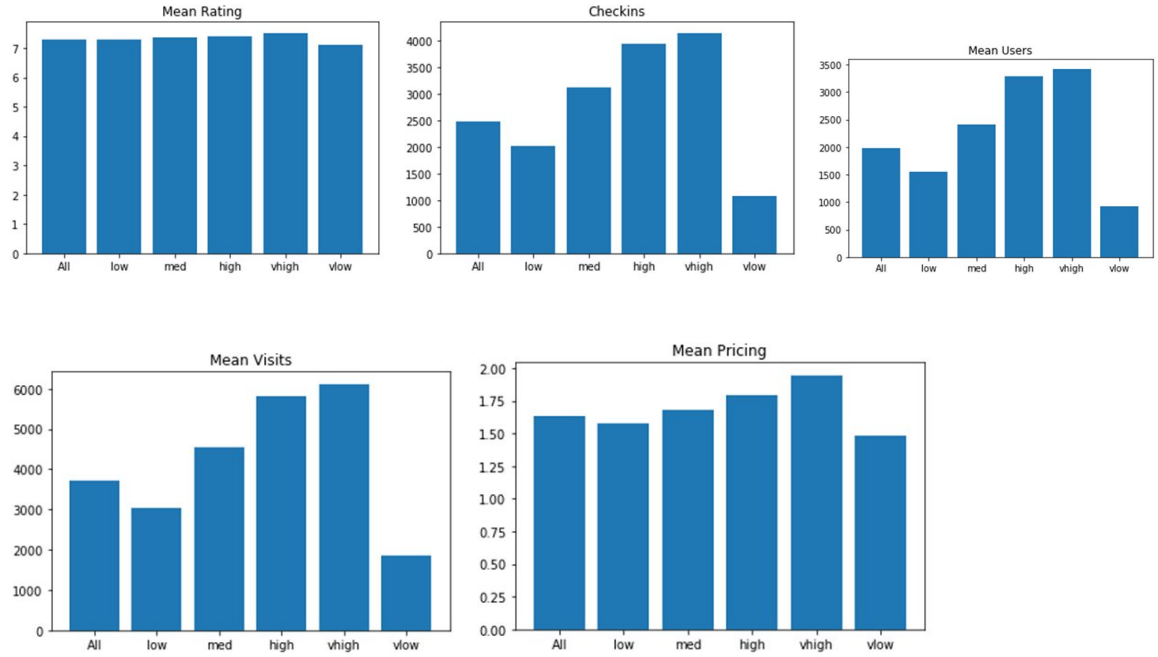


Figure B.10: Distribution of mean number of business 'Rating', 'Check-ins', 'Users', 'Visits', and 'Pricing'

Mean Values per Person

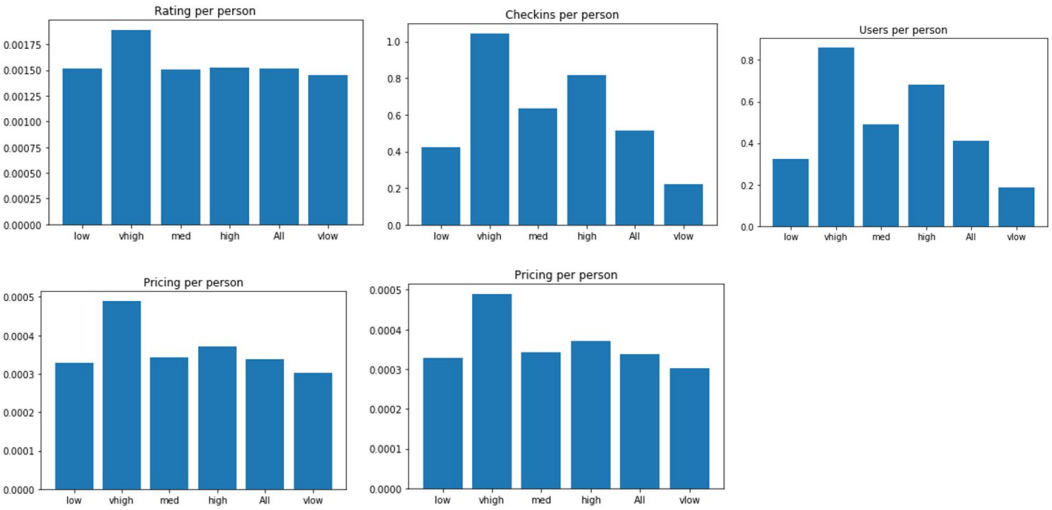


Figure B.11: Distribution of mean number of business 'Rating', 'Check-ins', 'Users', 'Visits', and 'Pricing' per person

Appendix C: Exploratory Analysis of Twitter Data

Following on from correlation exploration detailed in prior reports, the team moved on to clustering techniques to try to profile Twitter behavior in neighborhoods. Percentage of Local Tweets, Neighbor (from a surrounding census tract) Tweets, and Visitor Tweets were the metrics used to cluster census tracts. Using the elbow method (Figure C.1 below), we were able to identify the ideal number of clusters to provide sufficient detail to our team as 5. Silhouette score was also reviewed to ensure no significant issues for a single cluster arose.

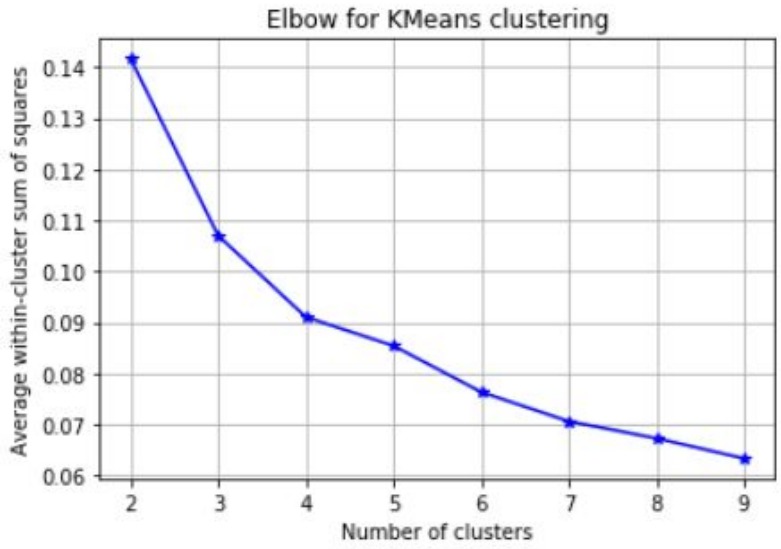


Figure C.1: Elbow plot for K-means clustering of Twitter data. There is a clear drop in performance after cluster 5.

Results are detailed in Figure B.2 below:

Figure C.2: Interactive dashboard which enables exploration of the cluster results. Clusters 0 and 2 represent the highest population of census tracts, dominated by Local Tweets. Clusters 1 and 4 represent an interesting subsection dominated by visitor and neighbor tweets, respectively.

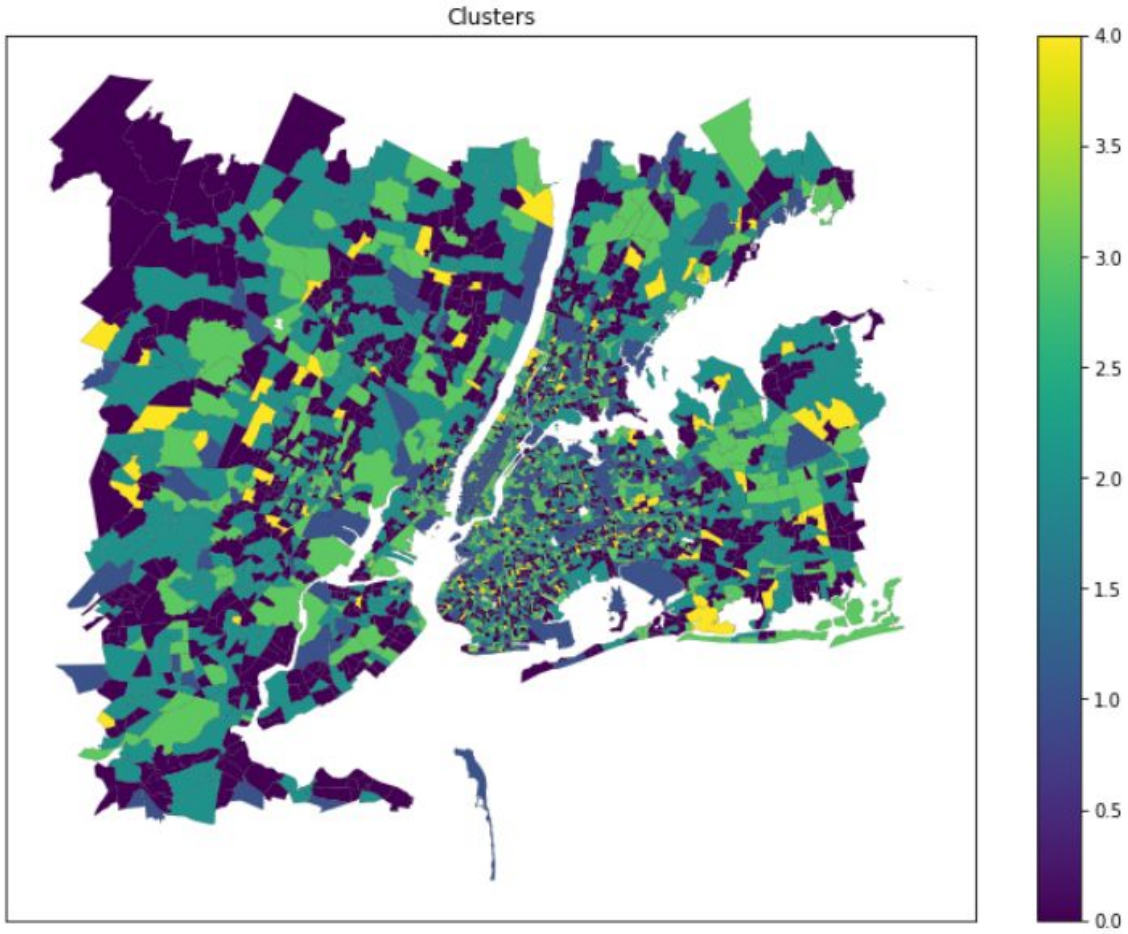


Figure C.3: A map of the cluster results in NYC and the surrounding counties. You can see that a large portion of Manhattan is Cluster 1, which is predominantly visitor tweets, as well as chunks of Brooklyn and Queens. Cluster 4, neighbor-dominated tweets, show up mostly scattered throughout Brooklyn, NJ, the Bronx, and Long Island.

Figure C.4 shows how the typologies related to each cluster. You can see that they are pretty widely distributed. However, it is worth noting that the presence of cluster one seems significantly higher in high income areas.

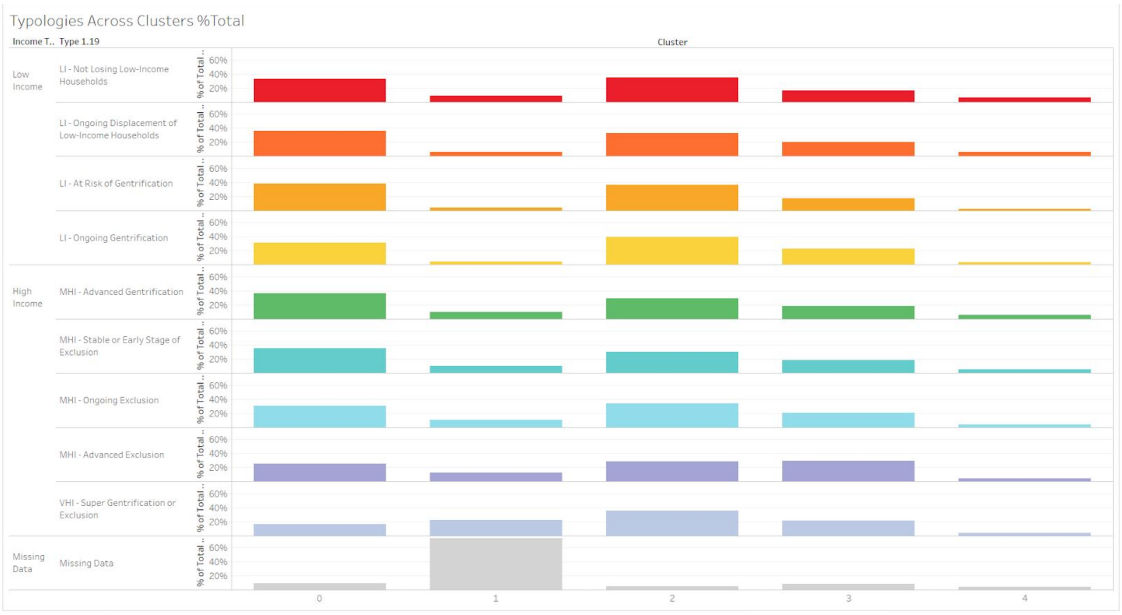


Figure C.4: This visualization shows the percentage of each typology that falls into the various clusters defined above.

Looking deeper at the neighborhoods in Typology 'LI- At Risk of Gentrification,' we can see that Clusters 1 and 4 correspond with a unique set of income and median rent ranges compared to the other regions:

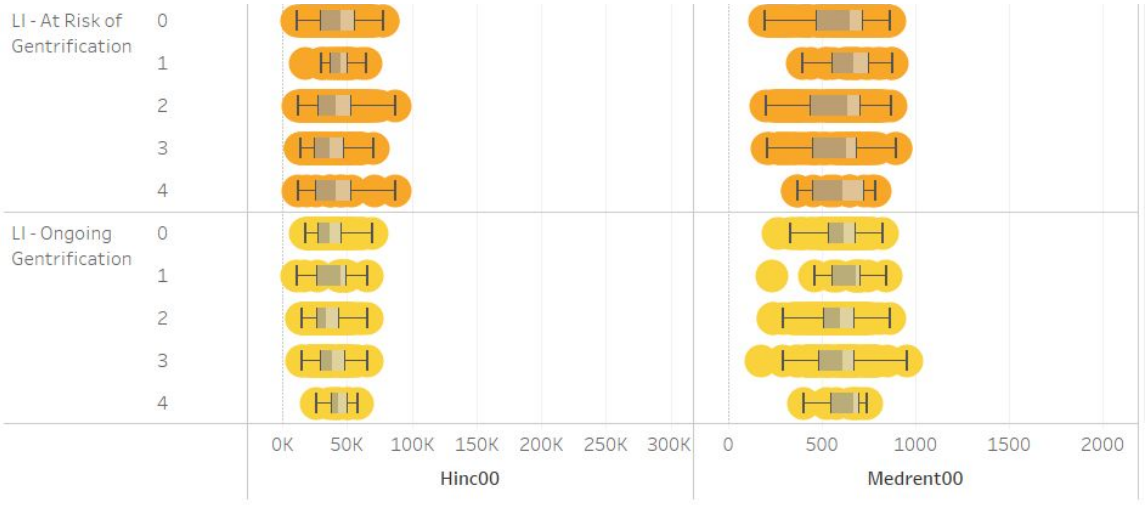


Figure C.5: This visualization shows a bar and whisker plot of Income and Median Rent for each census tract within the two target gentrifying typologies for each cluster. You can see that Cluster 0, 2, and 3 exhibit similar trends on these metrics, while clusters 1 and 4 seem to diverge. We have begun to dive into these anomalies to understand what underlying behavior is driving these findings.

Appendix D: Feature Engineering and Selection

Foursquare

D.1. Initial feature enhancements

For the Foursquare model, the following features were derived from the aggregated data:

- business per population
- business by income level
- Business type as a percentage of total businesses
- business type per population
- business type by income level

D.2. Nearest neighbors analysis

Spatial weights matrices were used to assess the effect of neighbouring census tract geometries - queens weights were used, as illustrated in the figure below (Anselin Rey & Li 2014). 10 iterations were used to achieve a larger catchment area for each particular census tract.

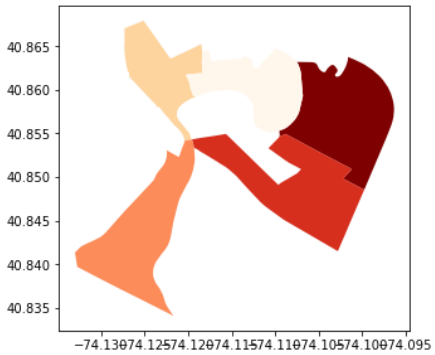


Figure D1. Example of a spatial weights matrix to assess the effect of neighbouring census tract geometries.

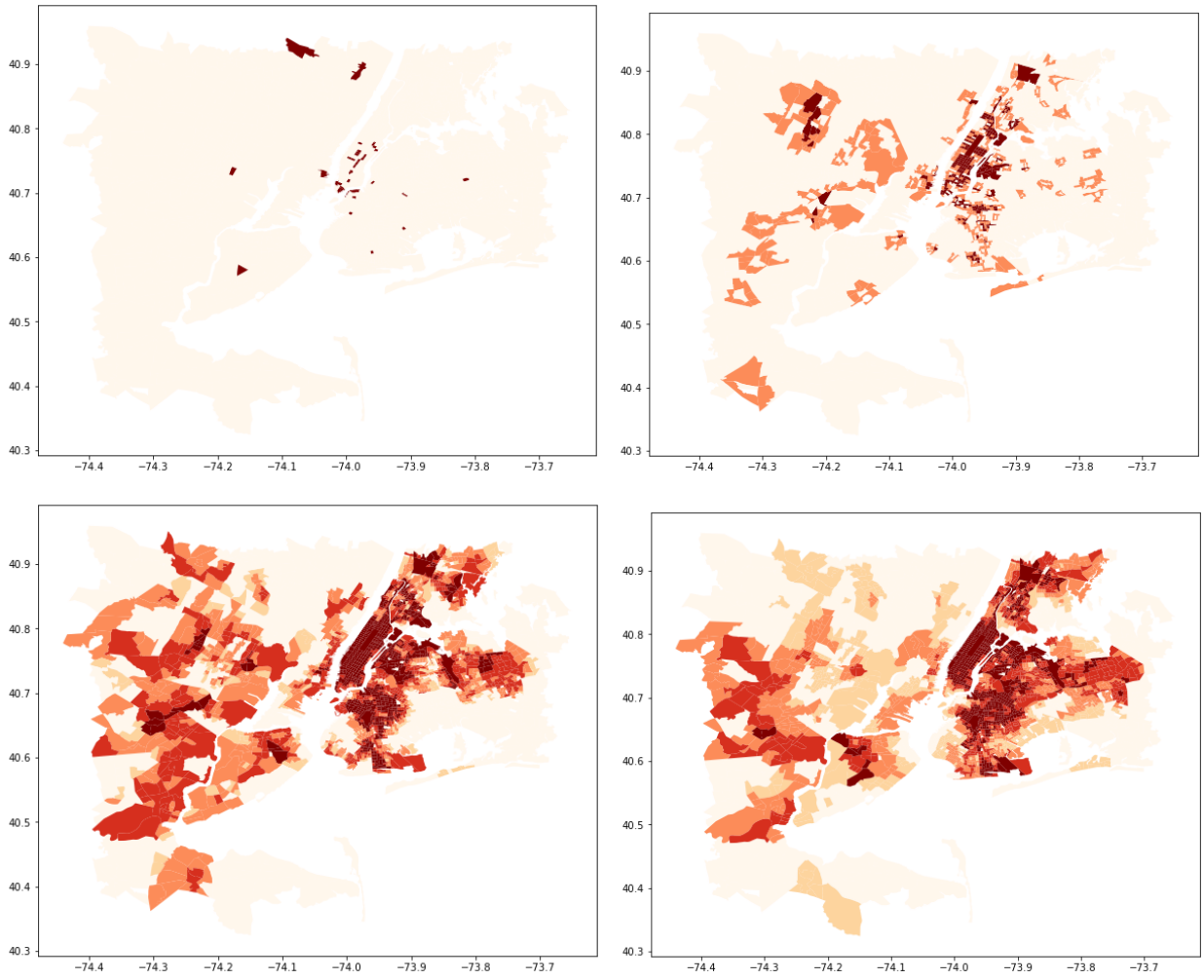


Figure D2. Example of Feature Enhancement using Learning from Neighbours. We added the sum of the business from the surroundings. The figure on the top left shows the raw data. After 1 iteration learning from the surrounding Census tracts we obtain the figure in top right. However we needed more iterations to increase the catchment area and produce useful features. The highest accuracy was obtained when conducting 10 iterations. This way, what used to look like figure on the top left becomes what is observed on the figure on the bottom right

D.3. Distance to Foursquare business types/class/category

After the exploratory analysis, we found the business types that change the most among income levels. The following business types were considered more present in high income neighborhoods:

- 'Type_American Restaurant',
- 'Type_Juice Bar',
- 'Type_Coffee Shop',
- 'Type_Deli / Bodega',
- 'Type_Food Truck',
- 'Type_Salad Place',
- 'Type_New American Restaurant',
- 'Type_Gastropub',
- 'Type_Steakhouse',
- 'Type_French Restaurant',

- 'Type_Café',
- 'Type_Wine Bar',
- 'Type_Cocktail Bar',
- 'Type_Beer Garden',

The following businesses were considered more present in low income neighborhoods:

- 'Type_Latin American Restaurant',
- 'Type_Bar',
- 'Type_Donut Shop',
- 'Type_Taco Place',
- 'Type_Fast Food Restaurant',
- 'Type_Spanish Restaurant',
- 'Type_Fried Chicken Joint',
- 'Type_Caribbean Restaurant'

In addition to the above types, we calculated the distance to all classes and categories for a total of 45 features.

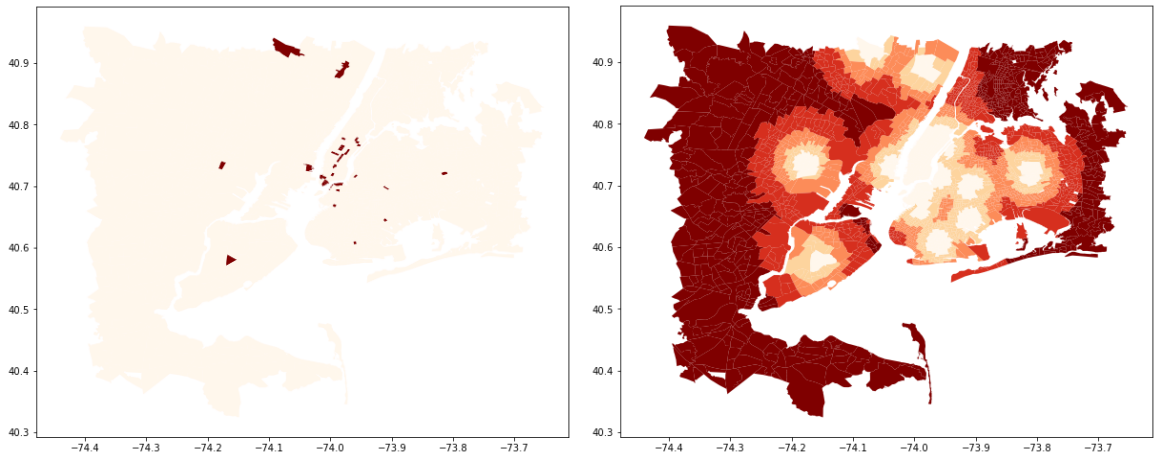


Figure D3. Example of Feature Enhancement using Learning from Neighbours. The left plot shows the census tracts in which a particular business type (salad places) is present, while the right shows the calculated distances for each census tract to the nearest tract with that business.

D.4. Principal component analysis

With the one-hot encoding of the Foursquare data business types and categories, 712 features were added to the dataset due to the large variety of business types. Combined with the additional distance features from the nearest neighbor analysis, the Foursquare data ultimately had 1463 features. Dimensionality reduction was a challenge for this dataset, and principal component analysis was used to reduce the dimensionality of the data. Even with 80 main principal components we can only explain less than 50% of the variance.

Twitter

D.5 Twitter raw data feature enhancements

For the Twitter model, the following features were derived:

- Local, visitor, and neighbor tweet counts, using a derived home location following Karen Chapple, et al, methodology as detailed above
- Tweet timing ratio features detailed above (e.g. weekday vs weekend, daytime tweets, day of week, etc.) following Chapple, et al, methodology
- Percentage of visitor, local, and neighbor tweets
- Tweet counts from each year 2012, 2013, 2014

D.6 Visitor Demographic Features

In addition to these raw data manipulations, the following features were created to get an understanding of the type of visitors to each neighborhood. These features were derived by looking at the home tract of a visitor tweet and assigning the attributes of that tract to the visitor:

- percentage of visitors from high, medium, and low income census tract
- percentage of visitors from highly educated census tract, as well as those from medium and low education tracts

D.7 Spatial Features

The final set of features were created to understand proximity to local hot spots, using the total number of tweets, medium high income tracts, and the percentage of visitor tweets (think: tourist locations, but also areas with airports were found to have high visitor percentages) to define these hotspots. From there, the distance from the tract to the nearest hotspot was measured:

- distance to high tweet tracts
- distance to high visitor percentage tracts
- distance to medium high income tract

D.8 Binary Typologies

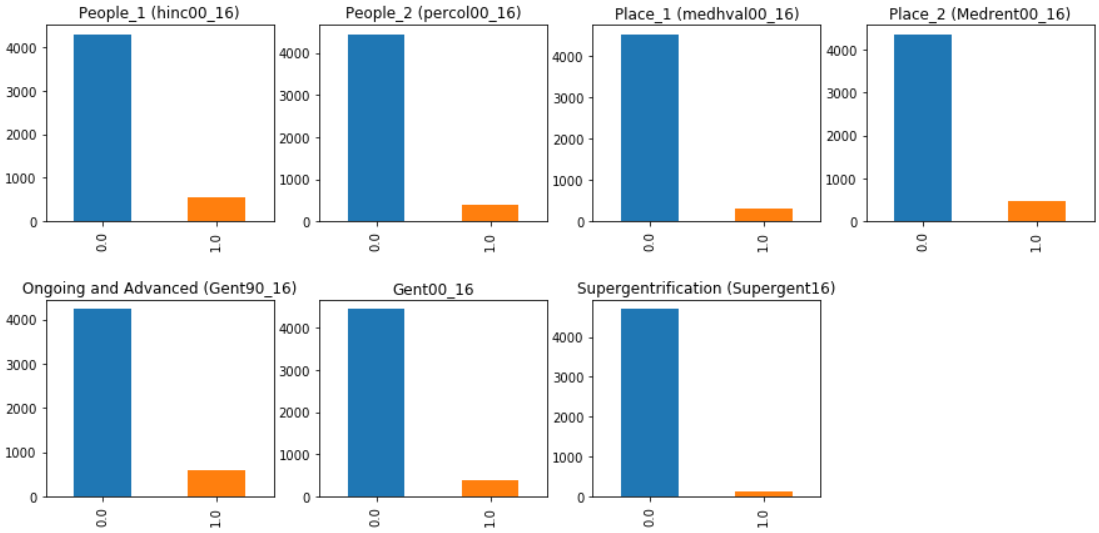


Figure D4. Distribution of the 7 Target Binary Typologies. It can be seen that there is a large class imbalance in our models. Supergentrification particularly has very few samples for gentrification.

Appendix E: Modeling Results

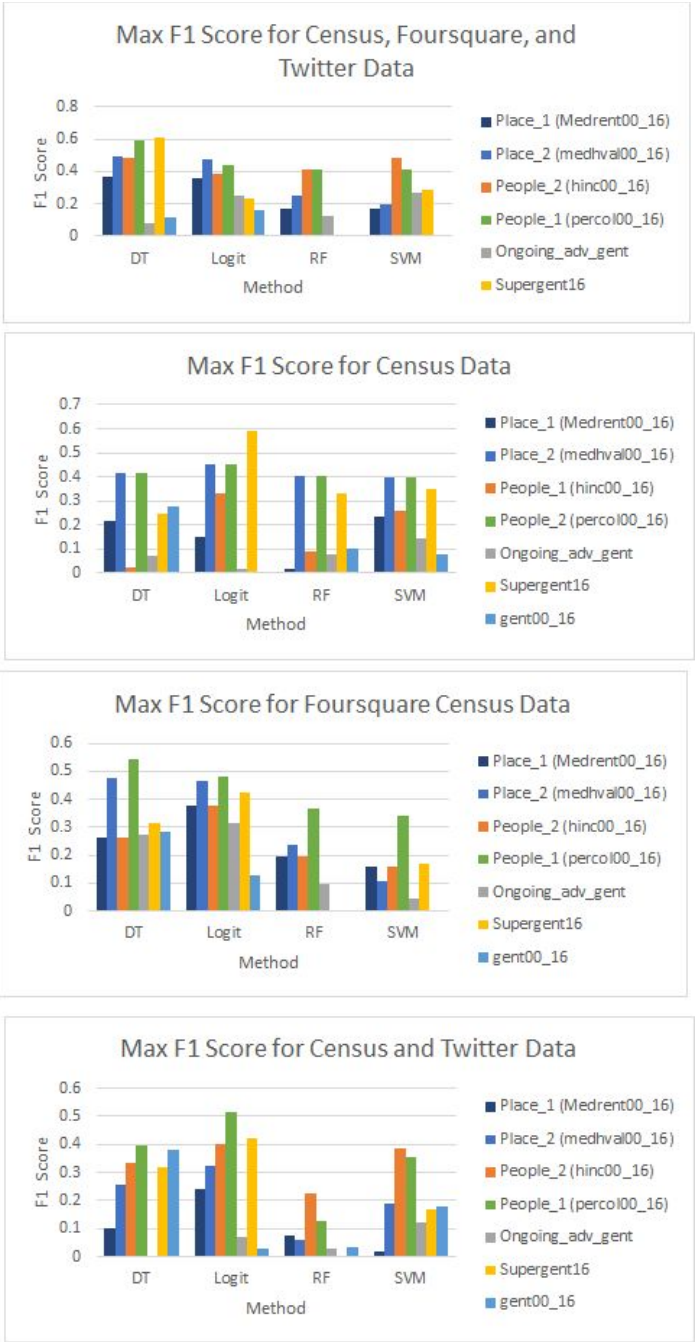
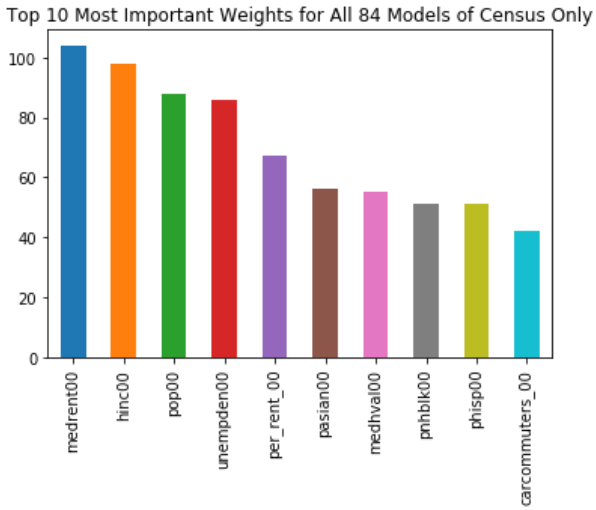
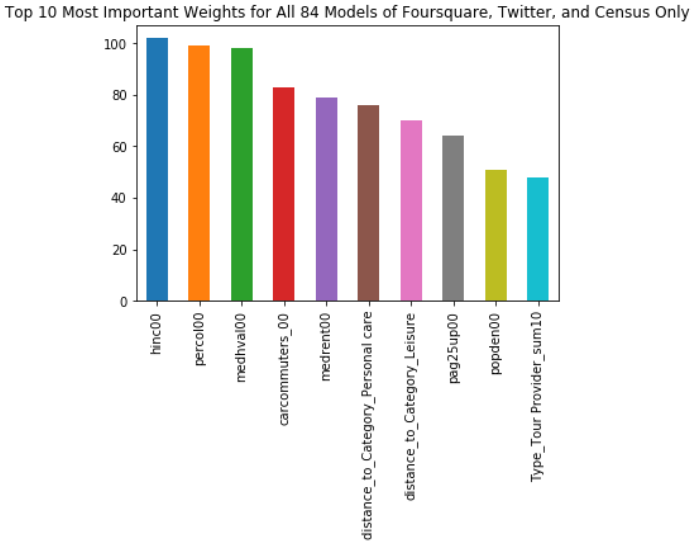
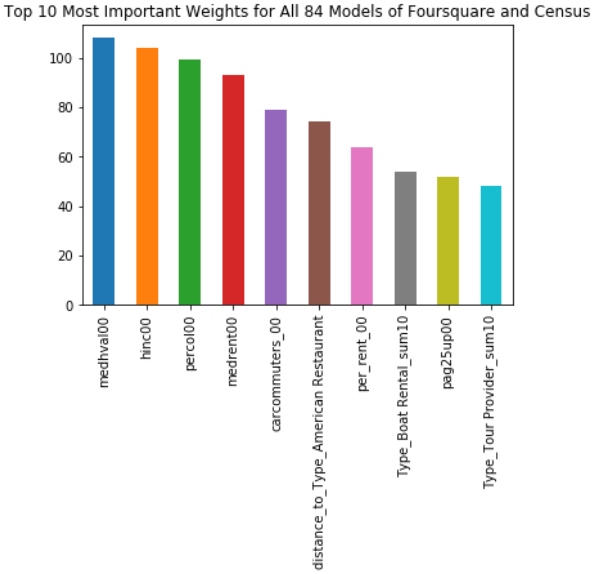


Figure E1. F1 Scores for Reduced Greater New York Area. In general decision trees and logisitc regression perform the best. It can be seen that Decision Tree perform the best with Foursquare data and the Combined Model.







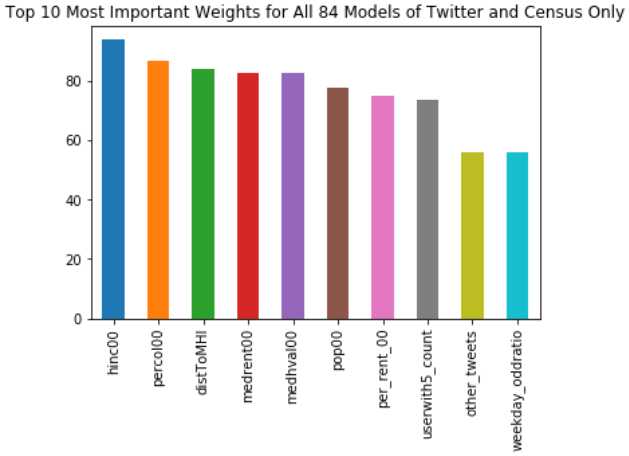


Figure E2.Weight Importance for Reduced New York Area Analysis.

Table E1. F Score on Test data for Twitter Models Using Decision Tree Model

Model Number	1.1	1.2	1.3	1.4	1.5	1.6	1.7
Binary Variable	binary	people_target	people2_target	place1_target	place2_target	gen16_target	supergen_target
Census Data Only	0.0	0.076	0.566	0.452	0.811	0.010	0.735
Twitter Data Only	0.019	0.070	0.027	0.034	0.581	0.062	0.055
Census and Twitter Data	0.212	0.042	0.597	0.546	0.822	0.439	0.802
Census and Twitter Data and Distances	0.014	0.441	0.523	0.331	0.990	0.450	0.780
All Twitter Data	0.256	0.069	0.563	0.320	0.990	0.434	0.780
All Twitter Data and New Census Data	0.069	0.441	0.555	0.271	0.990	0.010	0.796
Average	**0.094**	**0.189**	**0.471**	**0.325**	**0.864**	**0.234**	**0.656**

Table E2. F Score on Test data for Twitter Models Using Logistic Regression Model

Model Number	1.1	1.2	1.3	1.4	1.5	1.6	1.7
Binary Variable	binary	people_target	people2_target	place1_target	place2_target	gen16_target	supergen_target
Census and Twitter Data	0.004	0.0	0.0	0.0	0.0	0.232	0.802
Census and Twitter Data and Distances	0.0	0.116	0.142	0.086	0.729	0.010	0.224
All Twitter Data	0.0	0.181	0.157	0.20	0.768	0.0	0.340
All Twitter Data and New Census Data	0.032	0.04	0.270	0.195	0.740	0.0	0.178

Average	**0.009**	**0.084**	**0.142**	**0.120**	**0.559**	**0.60**	**0.386**

Table E3. F Score on Test data for Twitter Models Using Random Forest Modell

Model Number	1.1	1.2	1.3	1.4	1.5	1.6	1.7
Binary Variable	binary	people_target	people2_target	place1_target	place2_target	gen16_target	supergen_target
Census and Twitter Data	0.004	0.0	0.0	0.0	0.0	0.232	0.802
Census and Twitter Data and Distances	0.0	0.0	0.477	0.0	0.941	0.0	0.307
All Twitter Data	0.0	0.0	0.345	0.015	0.976	0.021	0.471
All Twitter Data and New Census Data	0.0	0.015	0.253	0.005	0.921	0.011	0.385
Average	**0.001**	**0.004**	**0.268**	**0.005**	**0.710**	**0.066**	**0.491**

Table E4. Out of Sample Accuracy on Test data [%] for Twitter Models

Model Number	**1.1**	**1.2**	**1.3**	**2.1**
Model Name	**Baseline**	**Population**	**Income**	**Binary**
Logistic Regression	0.451	0.351	0.398	0.434
Decision Tree	0.325	0.298	0.567	0.702
Random Forest	0.280	0.245	0.569	0.673
SVM	0.368	0.409	0.323	0.417
Average	**0.356**	**0.326**	**0.464**	**0.567**

Appendix F: Other Resources and Tools used during the project

The following Github Repositories were used in the development of this project:
https://github.com/mv1742/UDPNY
https://github.com/patafiot/Gentrification-Capstone

The team created a website with content about our project as well as an interactive map visualization which can be found here:
https://ace-gabriel.github.io/twitter_gentrification/

There are also interactive visualizations used during the Data Exploration phase, which can be found here:
https://public.tableau.com/views/TwitterDeepDive/Aboutthisworkbook?:embed=y&:display_count=yes&:origin=viz_share_link
https://public.tableau.com/profile/tiffany.patafio#!/vizhome/FoursquareDeepDive/WhatsintheData

Appendix G: Team Collaboration Statement

Below are the major tasks that were completed by each team member.

- Kent Pan
    - UDP 2018 Typology SME
    - Interactive map for website
    - Spatial feature engineering and data processing
    - Coordination with project sponsor/mentor
    - Final report writing and website project management
- Tiffany Patafio
    - Twitter data SME
    - Twitter data exploration
    - Twitter feature engineering and modelling
    - Interpretation of Twitter modeling and results
    - Zillow data coordination
- Manrique Vargas
    - Foursquare data SME
    - Foursquare data exploration
    - Foursquare feature engineering and combined modelling
    - Interpretation of Foursquare modeling and results
    - Interpretation of combined modeling datasets and results
- Jiawen Wan
    - Foursquare data exploration
    - Census data retrieval
- Tiancheng Yin
    - Twitter feature engineering
    - Website design & layout