Exploring Gentrification and Displacement Through User-Generated Geographic Information

Capstone Second Progress Report, CUSP NYU 2019
Sponsor and mentor: Karen Chapple, UC Berkeley
Students: Kent Pan, Tiffany Patafio, Manrique Vargas, Jiawen Wan, Tiancheng Yin

Abstract

Our capstone project will study gentrification and displacement risk for neighborhoods within the NY metro region. Using methodology established by the UC Berkeley Urban Displacement Project, and expanding on last year's CUSP capstone project, we will refine the methodology of the project and expand the scope of the model to incorporate not only administrative census data, but also user-generated social media data (geotagged Twitter data) and other sources of real estate, business, and transportation data. We will use classification models such as random forests with the collected data to better understand granular patterns and variances in activity across changing neighborhoods from those at risk to those in advanced gentrification and exclusion states. With an ever-increasing interest in previously disenfranchised areas as costs within cities climb and investors look for new domain, this area of study is extremely important in helping to shape our cities with residents in mind. The overall goal of our work is to better understand the activities and behaviors in these changing areas so that we can provide insight to help the community, public officials, and other interested parties better manage and plan for the cycle of change.

Abstract **1**

Introduction **3**

Data **3**

Methodology **4**

Data Exploration 4

Modelling Approach & Scope Changes 5

Challenges 6

Results **7**

Foursquare Results 7

Twitter Results 9

Results Discussion **10**

Conclusion **11**

References **12**

1. Introduction

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.

The majority of the previous literature studying gentrification and displacement have used "traditional" sources of census data, such as household income and demographic information. Recently, studies have begun to leverage other novel data sources, such as Chapple et al. (2018) utilizing Twitter data in San Francisco to study mobility patterns, or Cerron et al. (2018) using Foursquare & Instagram data to describe how people engage in public urban spaces.

The Urban Displacement Project (UDP) at UC Berkeley, along the 2018 CUSP capstone group, developed and refined a methodology to categorize the state of gentrification/displacement at the individual census tract level along an eight-stage gentrification typology index. Using this index, this project aims to extend upon that work and apply the methodology to the NY metro region using three different sources of near real-time data sources from online platforms: Twitter, Foursquare, and Zillow. Drawing upon methodologies from similar previous studies, we will apply models to these data sources to understand different factors that influence gentrification beyond traditional census data, such as human mobility and business interaction. Understanding these factors will allow city officials and local community members to adequately prepare for future transformations in their neighborhood.

2. Data

Through coordination with our project sponsor, we've identified three different available online data sources to act as proxies for human activity across three different domains: Twitter for mobility, Zillow for housing, and Foursquare for businesses. These distinct sources will be used to evaluate whether or such generated data can help us improve our understanding of more granular gentrification patterns and activity in changing neighborhoods.

The table below summarizes the characteristics of each of the three datasets in our project scope. Further, more detailed information on the individual datasets are described in our previous progress report.

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

	Zillow	**Twitter**	**Foursquare**
Time Range	1990-2016	2012-2015	2009-2018
Geographic Range	Entire study region	Entire study region	NYC region only

Data Type	Polygon (census tract)	Polygon (census tract)	Point (individual businesses)
Attributes	Transaction data & property attributes (e.g. sales price, building age, property use type) aggregated to census tract level).	Raw tweet metadata (user, location, and time) aggregated to the census tract level.	Business categories, types, class; # of users, check-ins, visits.
Other Considerations	Only a small sample of dataset has currently been provided (3 counties). Full dataset is being processed by project sponsor.	Data has been processed to extract time-spatial (e.g., # of local and visitor tweets) and demographic features for each tract.	Only contains 30% sample of businesses in NYC region; All data is aggregated from entire time range.

Each of these three datasets are joined to the census data and typology index that has already been identified for each census tract in the study region for further analysis.

Although we are still working with our project sponsor to obtain the Zillow dataset, we have begun exploratory analysis and modeling on the Twitter and Foursquare datasets, described in the sections below.

3. Methodology

3.1. Data Exploration

In order to set a solid foundation for our analysis and understand the limitations of the various datasets, we used simple visualizations as well as basic mathematical techniques to explore our currently available datasets, FourSquare and Twitter. This work included gaining an understanding of the profile of the data, the distribution of the data across our target geography, as well as an in depth look at some simple relationships within the datasets. Detailed clustering, correlation, and visual analysis were conducted to achieve these goals. Results and visuals presented in section 4 and the appendices.

We clustered Foursquare data by income level. Appendix A shows the results of our data exploration. It can be observed that high income neighborhoods have more business in general. When analyzing number of businesses per population, higher income neighborhoods also have a higher number of businesses per person. We believe that adding the feature businesses per population will improve our model. Briefly, based on this data exploration, the following features will be added:

- Business per population
- Business type as a percentage of total businesses

Similarly, Twitter data was clustered using several different metric combinations, with the best results coming from clustering 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, detailed in the results section below, and were then compared with the known typologies to identify potential patterns in user behavior in typologies of interest. Results are discussed in section 4.2 and Appendix B.

3.2. Modelling Approach & Scope Changes

Our goal is to model the typologies defined by last year's capstone detailed in prior papers using new data sources, including Twitter, Foursquare, and Zillow. The typologies are inherently divided between high income and low income neighborhoods, which will impact our model design to be similarly split. Our goal, therefore, will be to classify the types of neighborhood change (typologies) undergone in high income and low income areas using our other data sources, Zillow, FourSquare, and Twitter, with the hope of better understanding patterns and activity in neighborhoods across the spectrum of typologies.

We model our problem to output a multiclass classifier of the typologies. The output classifier is represented as y_i, where i corresponds to every census tract in our study. Data pre-processing has been critical to setting a strong foundation for our models. In particular, a large number of possible features in the datasets and a high likelihood for correlation in features required exploration for some feature engineering and dimensionality reduction techniques, such as normalization and regularization, respectively as well as multicollinearity testing. Additionally, handling of categorical variables through factorization will be critical given the frequency of categorical features across datasets.

Three different models were explored using different features for each dataset, as well as two predictive models (Models 1.4 and 1.5 for census/Foursquare only), with the target variable being the gentrification typology. Each model was evaluated with four methods: logistic regression, decision trees, support vector machines, and random forests. The details for these models are presented in Tables 2 and 3 below.

Table 2: Foursquare Models

Foursquare Model	**Features**
Model 1.1: Baseline Foursquare model	Raw Foursquare data features
Model 1.2: Foursquare data by population	Raw Foursquare data features and relevant census population features
Model 1.3: Foursquare data by income level	Raw Foursquare data features as percentage of income level
Model 1.4: Census data (predictive model)	Census features from years 1990 and 2000
Model 1.5: Foursquare data with census data (predictive model)	Raw Foursquare data and census features from years 1990 and 2000

Table 3: Twitter Models

Twitter Model	**Features**

Model 1.1: Baseline Twitter model	Raw Twitter data features
Model 1.2: Twitter data by population	Raw Twitter data features and relevant census population features
Model 1.3: Twitter data by income level	Raw Foursquare data features as percentage of income level

The results of this modeling are presented in Section 4 of the report below. After reviewing these results that attempt to model all 8 typologies, we are going to refine our models in a few ways. First, we have begun to focus our analysis to more specifically look at the gentrified areas in particular, as this most closely aligns to our goals and would be most useful in helping to understand changing communities. To do so, we consolidate the 8-part index into 2 simplified typologies: gentrified and not gentrified. This allows us to more effectively target the specific features that impact neighborhood change, rather than those that are prevalent in the other more stable typologies. We have used this design with all four methods detailed above (random forest, logistic regression, decision trees, support vector machines), and identified it as Model 2.1 for both Foursquare and Twitter datasets (see Table 4 below).

Table 4: Binary model for Foursquare and Twitter

Model	**Features**	**Target**
Model 2.1: Binary model	Raw Foursquare or Twitter data features	2 simplified typologies: gentrified and not gentrified

Secondly, we will repeat the models executed on all 8 typologies on high income typologies and low income typologies separately as income is largely the basis of the typologies and prior research, including a study from Reades et al. (2018) indicate that income alone is the biggest component in understanding neighborhood change. Lastly, we will continue exploring potential new features identified during data exploration exercises, including the percentage of identified business types relative to total businesses in an area and features around visitor tweets. As a final step, a combined model including the most important features from each dataset will be tested and evaluated.

3.3. Challenges

There have been several key challenges for our team to overcome in executing this analysis. Data availability is the single largest challenge in painting a full picture of neighborhood change. Datasets from different sources are at different levels of detail and represent different timeframes-- for example, Foursquare data is only available as a snapshot from March 2018 while Twitter data covers 2012-2015. The Foursquare data provided is also only a sample (30% of the full population), with more high income areas represented as shown in Figure A.1 in Appendix A. Additionally, the Zillow data has taken longer than expected to be delivered and the team only has sample data for a few counties. For this reason, Zillow analysis has not been executed for this report and is at risk for final delivery, pending the reception of the data.

Modelling and identification of appropriate features has also presented some challenges, which we have used extensive data exploration to mitigate. Understanding the relationships between features within and across datasets is crucial to further refining our models. For example, with the FourSquare data, we have been exploring questions such as is the number of visitors to a certain type of business truly varied across neighborhood type or is there a bias in check-in user activity at different businesses across neighborhoods? Similarly, relationships between features like income and user activity and correlation within a single dataset (e.g. number of twitter users and number of tweets) must be accounted for in our modelling and interpretation.

4. Results

4.1. Foursquare Results

During the data exploration phase, several findings around business type stood out as significant. For example, donut shops and fast food restaurants made up a significantly larger proportion of business in low income neighborhoods than in high income areas. Additionally, the presence of coffee shops and Italian restaurants was noticeably larger in high income areas.

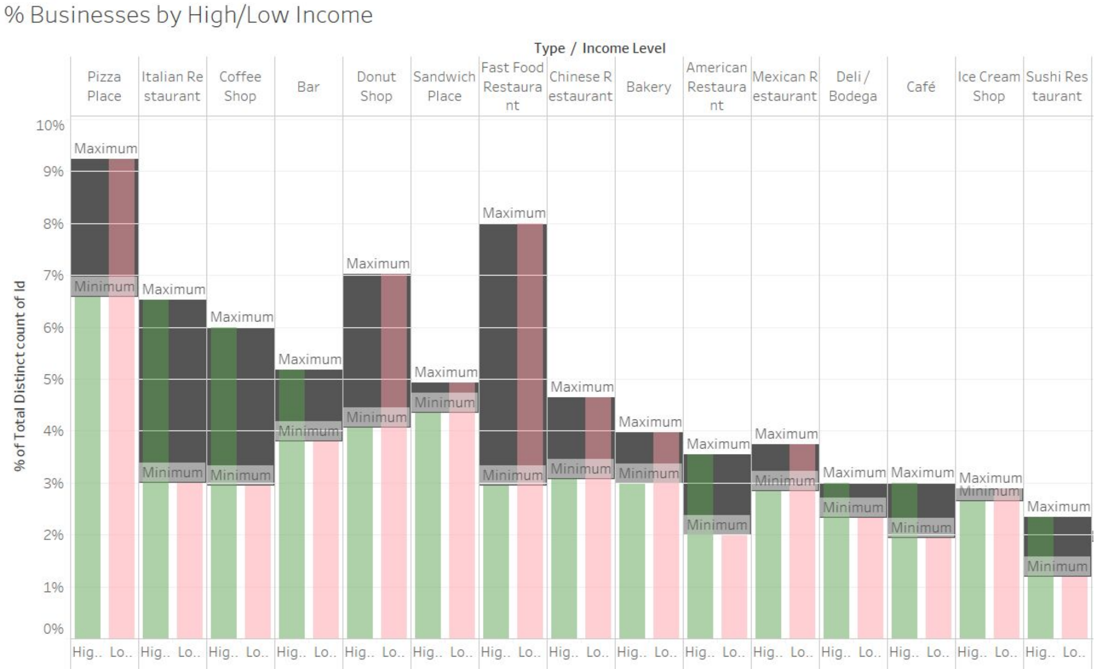


Figure 1: The chart above shows the percentage of businesses of each type by high (green) and low (pink) income typologies. The dark fill represents the difference between the two-- so businesses with larger dark bars represent those with greater disparities.

Upon further exploration, we were also able to identify some business types which made up larger shares of gentrifying areas, such as Sandwich Places and Mexican Restaurants. The Foursquare data also highlighted interesting trends as you moved along the spectrum of gentrifying typologies-- for example, an increase in coffee shops and bars with further gentrification. As a last portion of our exploratory analysis, we examined the Foursquare data classes (necessary, optional) and categories (nutrition, social, consumption). Looking across the phases of gentrification, we again see a trend-- decreasing percentage of necessary businesses, with increasing optional businesses and decreasing nutritional percentage with increasing social businesses. These results are detailed in Appendix A.

As discussed in previous methodology section, we conducted six different models for Foursquare data. The results are shown in Table 5 below.

Table 5. Out of Sample Accuracy on Test data [%] for Foursquare Models

Model Number	**1.1**	**1.2**	**1.3**	**1.4**	**1.5**	**2.1**
Model Name	**Baseline model**	**By Population**	**By Income Level [Pop16]**	**Using Census Only [1990 and 2000 data only]**	**Using Census with Foursquare**	**Binary**
Logistic Regression	0.270	0.305	0.302	0.337	0.337	0.700
Decision Tree	0.283	0.290	0.441	0.530	0.529	0.704
Random Forest	0.285	0.291	0.403	0.516	0.530	0.690
SVM	0.282	0.278	0.280	0.286	0.286	0.667
Average	**0.280**	**0.291**	**0.357**	**0.417**	**0.420**	**0.690**

It can be observed that the accuracy increases as we improve the features of our model (left to right in Table 5). As expected, logistic regression is a reliable classifier but not the best with non-linear data. Surprisingly, decision trees method performs the best in some cases. SVM should perform better or similar to other classification techniques. The poor performance of SVM is most likely due to the lack of normalization. Future models will normalize the data before conducting SVM classification.

Appendix D shows the feature importance found in the random forest models. It can be observed that in models 1.4 and 1.5 that uses Census data, the Census data features are more important than any Foursquare feature, which explains why the model improves. On the other hand, while check-ins, visits and users have the greatest importance overall among Foursquare data attributes, their importance diminishes as we improve our features.

4.2. Twitter Results

The Twitter clustering based on percentages revealed 5 distinct profiles of neighborhood Twitter activity, with the majority of census tracts falling in locally dominated clusters 0 and 2. Clusters 1 and 4 show unique activity patterns with a predominance of visitor and neighbor tweets respectively.

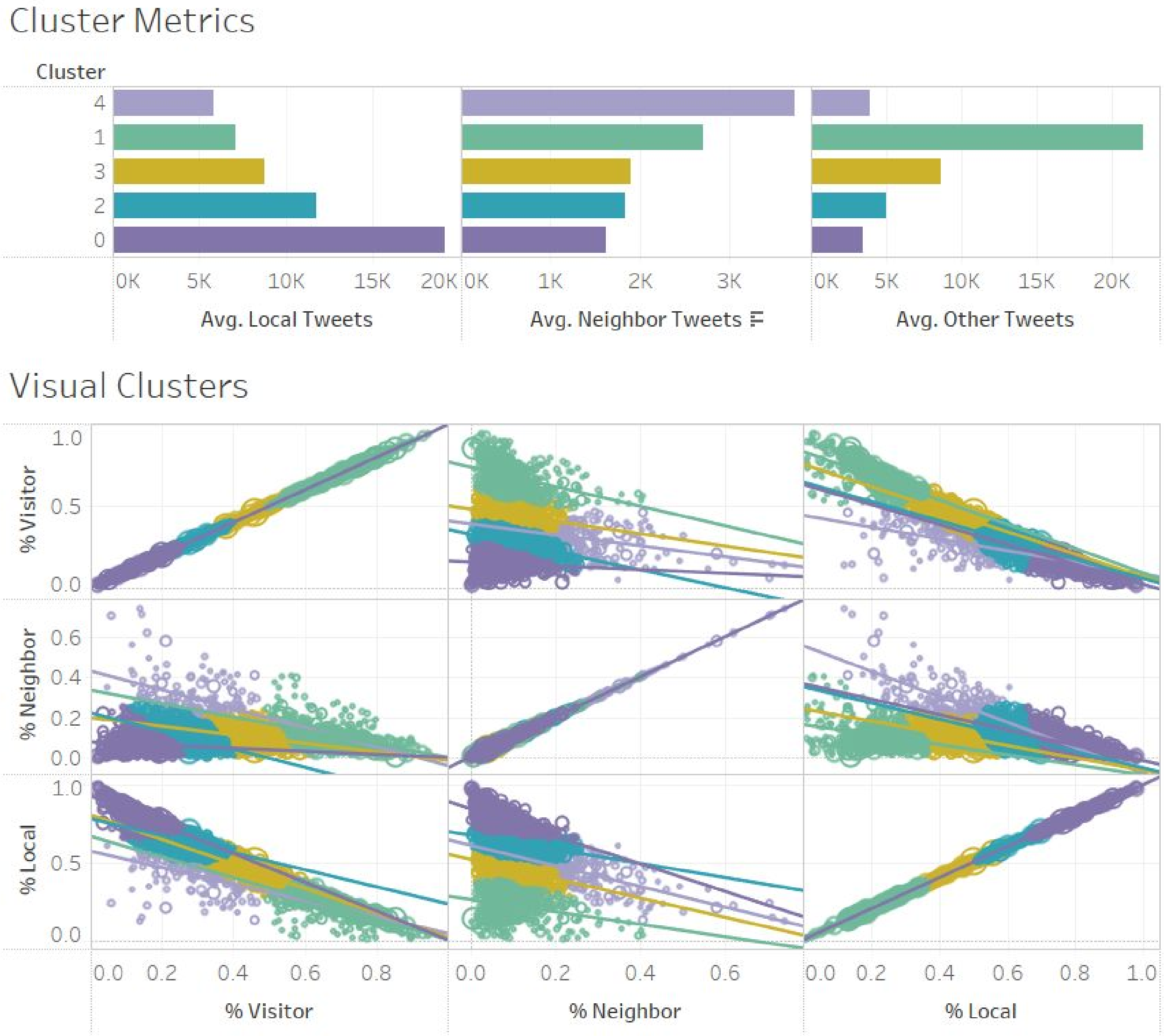


Figure 2. The figure above represents the results of the clustering results. The top visual provides average metrics for the census tracts within each cluster, while the bottom shows the relative positioning across percentage metrics with each census tract represented by a point whose size is the total number of tweets and color is the associated cluster.

Comparing these results to the typologies unveiled that most tracts associated with gentrifying neighborhoods have predominantly local tweets, with increasing visitor tweets as they move towards exclusion. Further digging revealed at risk tracts within clusters 1 and 4 show different median income and percentage rent levels than at risk neighborhoods in other clusters. Additional result details available in Appendix B. We will continue to explore these anomalies as well as increased exploration of the demographics of visitors to incorporate into our modelling and final results.

Modeling the relationship between the Twitter dataset and gentrification typologies (using the four different models as detailed in Section 3.2) led to results shown in Table 6 below.

Table 6. 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**

In general, model performance is poor across all four. Twitter models did not show improvement with the inclusion of population data, but improved considerably with the addition of income data as expected. Based on a study by Chapple et. al (2019), we expect visitor tweet behavior and demographics to provide even better insights into behavior and modelling. The addition of this data and exploration of anomalies discussed above should reveal more interesting features for use in our models.

4.3. Results Discussion

The additional features using population and income improved the performance of the Foursquare data. From our data exploration, it was evident that businesses per population and income provide insights about the neighborhoods studied. The accuracy improved from 0.280 for model 1.1 to 0.357 for model 1.3 (see table 5). On the other hand, as expected, Twitter models didn't improve with population given early exploration showed no correlation between these features and volume of tweets.

Appendix C shows the distribution of typology types. It can be seen that even a model with a strong prior that assumes all neighborhoods to be classified as MHI - Stable or Early Stage of Exclusion, which represents 0.37 of total typologies, can have a stronger performance than our base model that obtained 0.28 accuracy only. Similarly, the best performing model using Foursquare data on binary classes (gentrified and not gentrified) has an out of sample accuracy of 0.70, while the largest class (0 - 'No Gentrification') represents 0.72 of the total two classes.

In order to improve our results performance, we must continue to improve our features. We will extract Shannon-Wiener Index of complexity of activities as well as gravity and centrality indexes to better describe experiential qualities of public spaces and development opportunities of urban spaces and neighborhoods. In addition, in order to learn from the effects of the neighboring Census Tracts, we will map the average census tract and user generated data features spatial lag deciles (the average of the values for the neighbouring regions) using queen's weights (Anselin, Rey & Li 2014).

5. Conclusion

While the addition of certain features and tweaks to our initial models showed some improvement, it is clear from the initial set of results that the models are not yet performing at sufficient levels to establish strong findings. Moving forward, we intend to focus more targeted efforts towards understanding the relationships and activity specifically within gentrifying neighborhoods through additional iterations of the binary model. This focused effort will allow us to identify a subset of features with strong relationships to this type of neighborhood change which is at the core of our research objectives. Development of new features based on our data exploration findings will also greatly improve model performance.

Through the next month, we will continue to iterate through independent Foursquare and Twitter models identifying the most important features and honing in on the methods that provide the clearest results. With the delivery of Zillow data, we expect to follow a similar targeted approach given the limited timeframe for exploration-- focussing specifically on gentrifying neighborhoods. Finally, we will test whether the combination of these three datasets into a single model could provide stronger insight. The ultimate goal will be to provide interactive maps and visualizations for the community and interested parties to explore, highlighting interesting case studies, findings, and modelling results. The details of the workplan are available in Appendix E.

6. References

1. Chapple, K., Ate, P,. Matthew,. Z, Eva,. P (2019). Monitoring Streets through Tweets: Using User-generated Geographic Information to Predict Gentrification and Displacement.
2. 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.
3. 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.
4. 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
5. Steif, Ken. A. M. (2016). *Predicting gentrification using longitudinal census data.* Philadelphia: http://urbanspatialanalysis.com/
6. Reades, J., De Souza, J., & Hubbard, P. J. (2018). Understanding urban gentrification through Machine Learning: Predicting neighbourhood change in London.

Appendix A

Foursquare Data Exploration Plots

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:

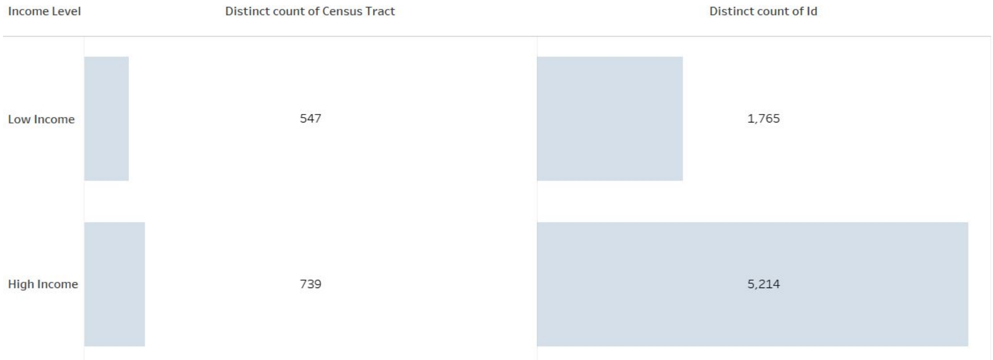


Figure A.1: The chart above shows the number of distinct census tracts and business IDs by high and low income census tracts. You can see that while the split of census tracts are close, there are many more businesses from high income areas in the dataset.

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.

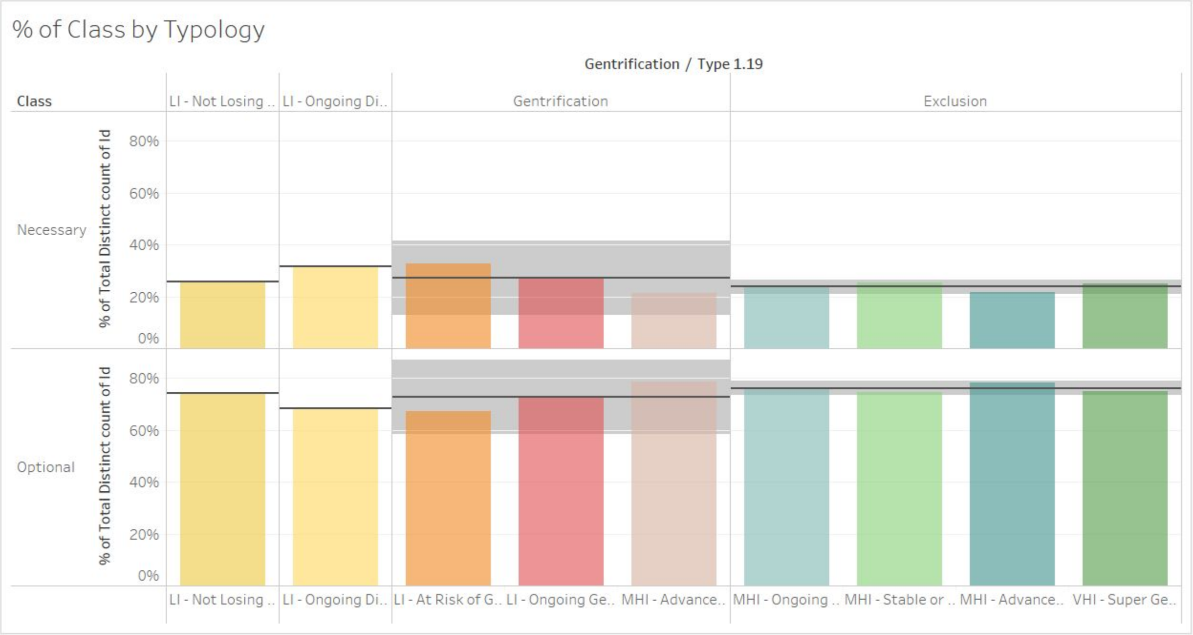


Figure A.2: Necessary and Optional business types are vary greatly across the gentrifying typologies, with necessary businesses decreasing to below exclusion areas and optional increasing to levels above.

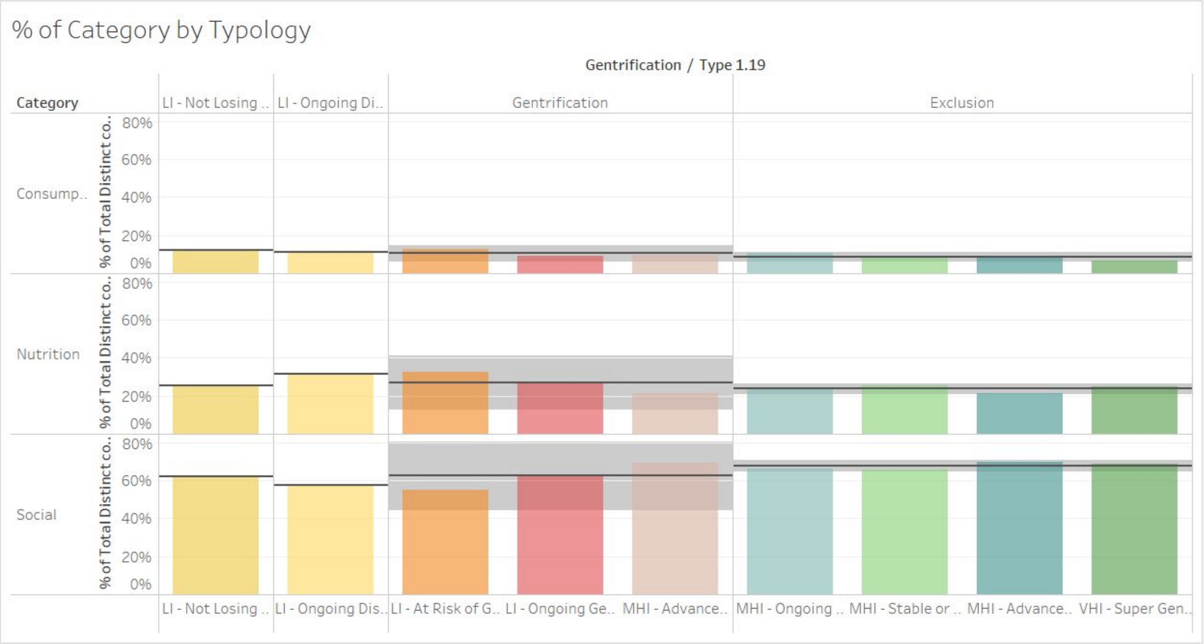


Figure A.3: Consumption businesses represent similar percentages across typologies, while nutrition and social show distinct patterns across the spectrum of gentrifying neighborhoods with social increasing and nutrition decreasing.

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.

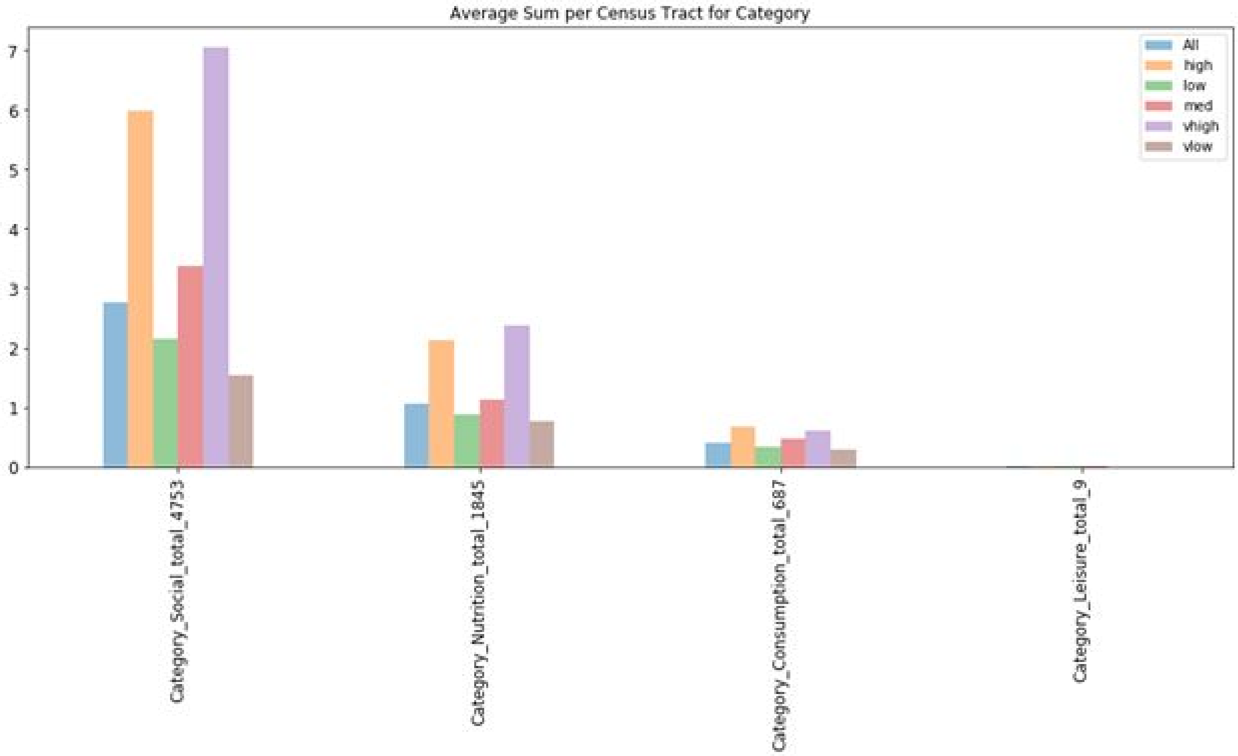


Figure A.4: Distribution of total average sum of ‘Category’ types per Census Tract per income level

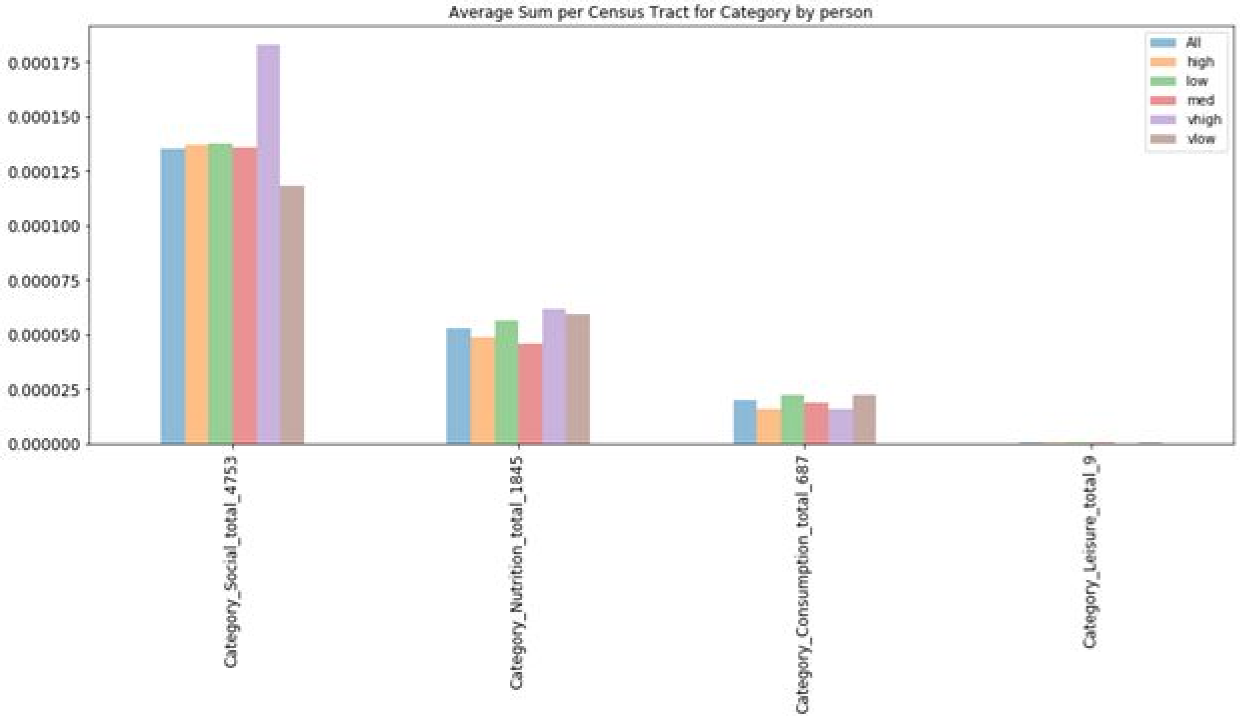


Figure A.5: Distribution of total average sum of ‘Category’ types by person per income level

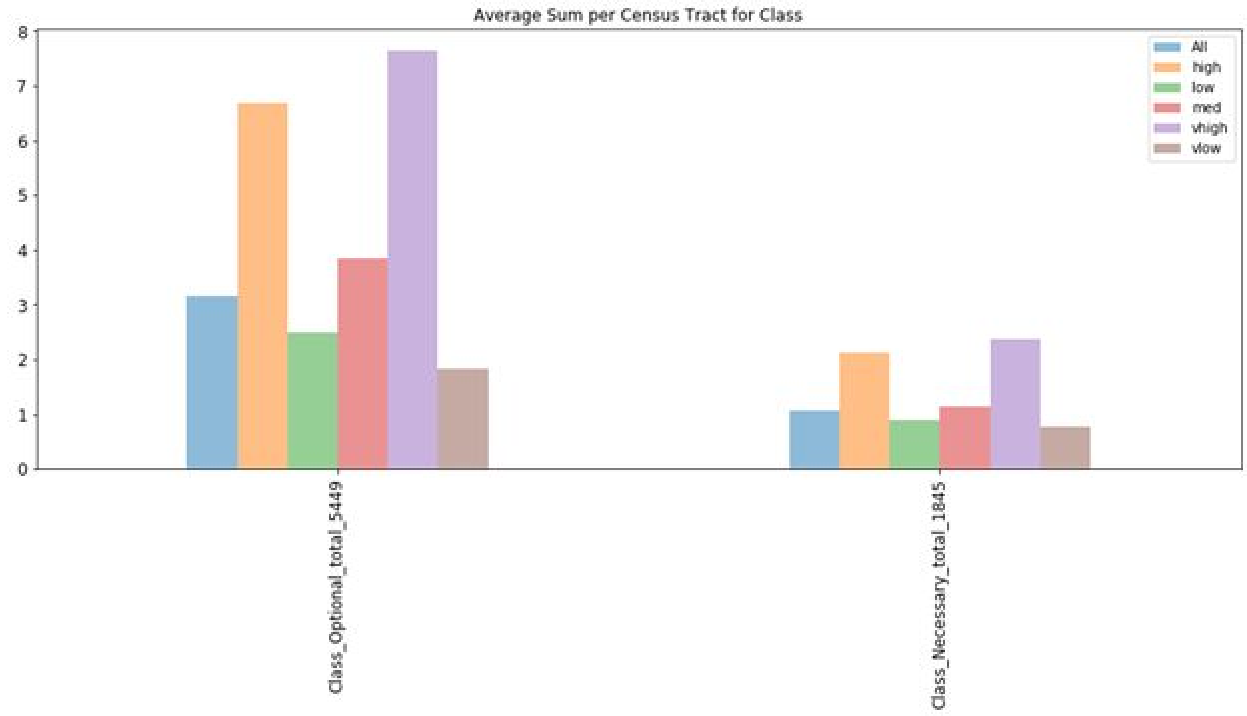


Figure A.6: Distribution of total average sum of ‘Class’ types per Census Tract per income level

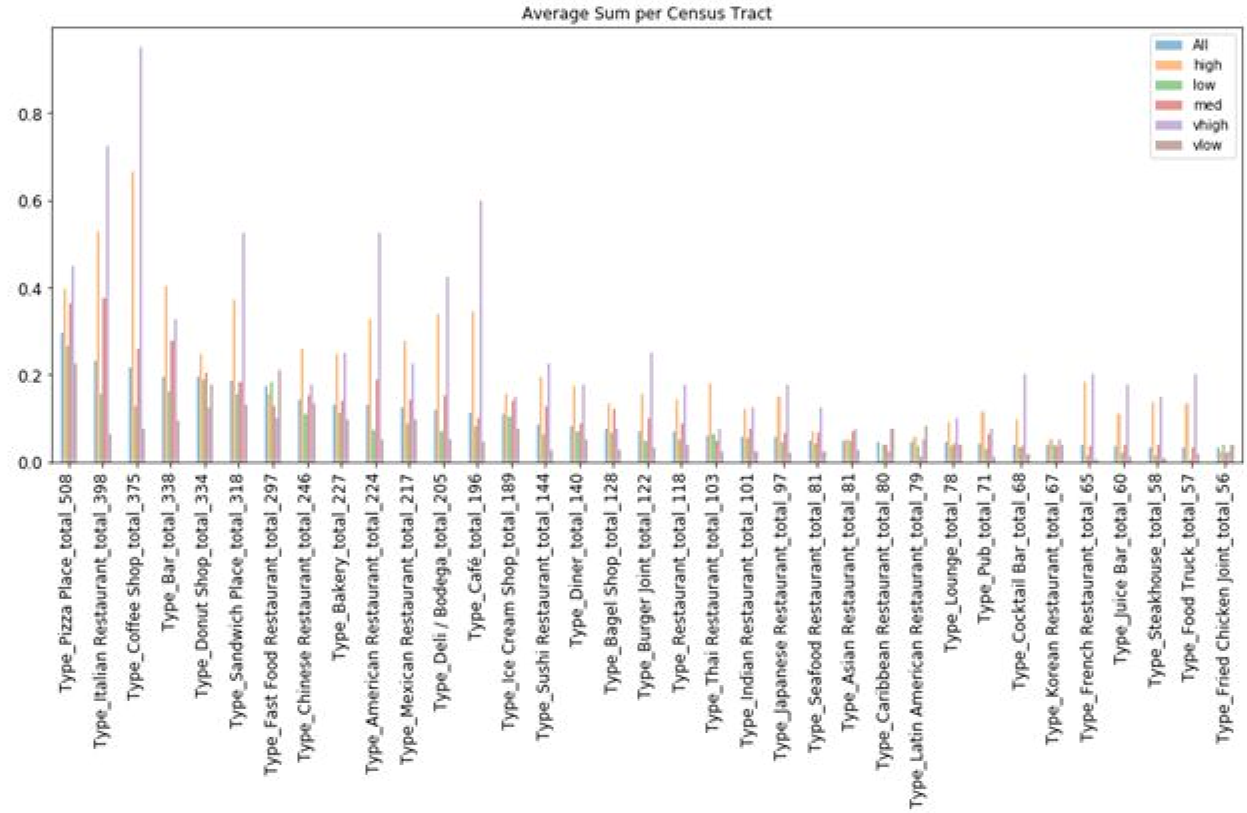


Figure A.7: Distribution of total average sum of ‘Type’ per Census Tract per income level

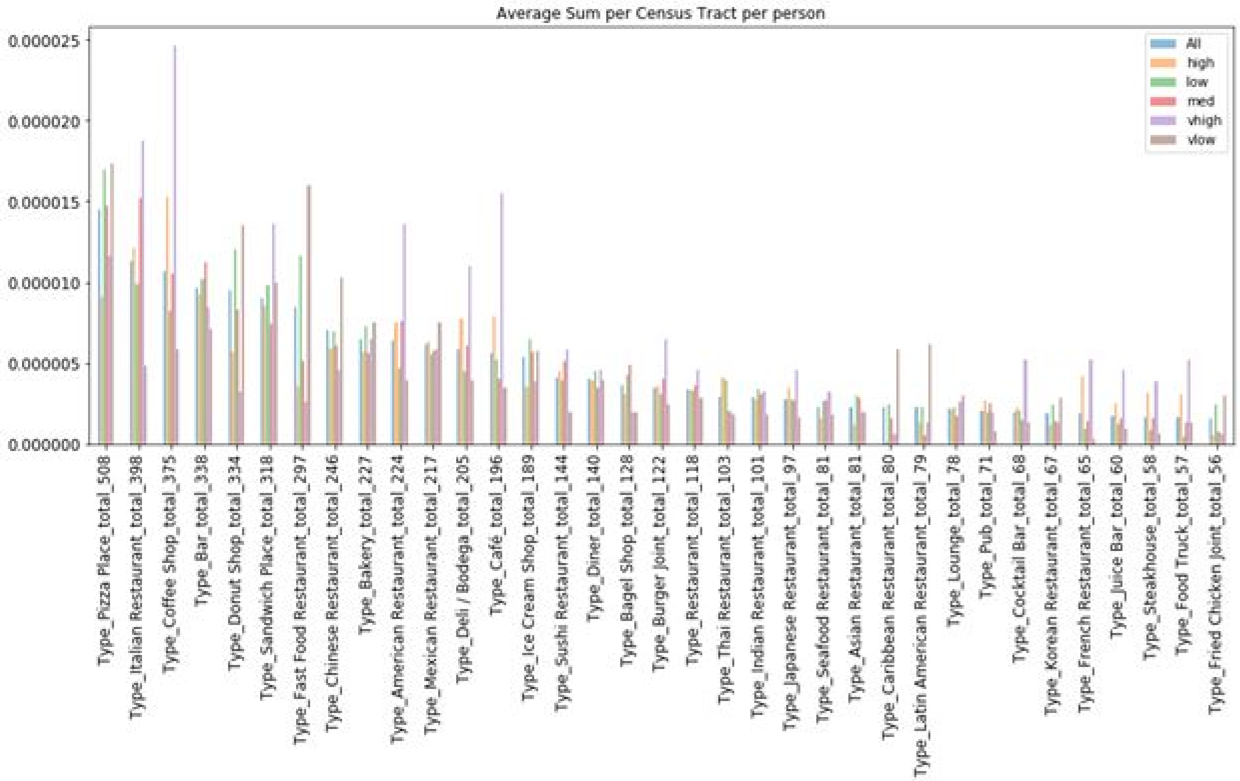


Figure A.8: Distribution of total average sum of ‘Type’ per Census Tract per income level

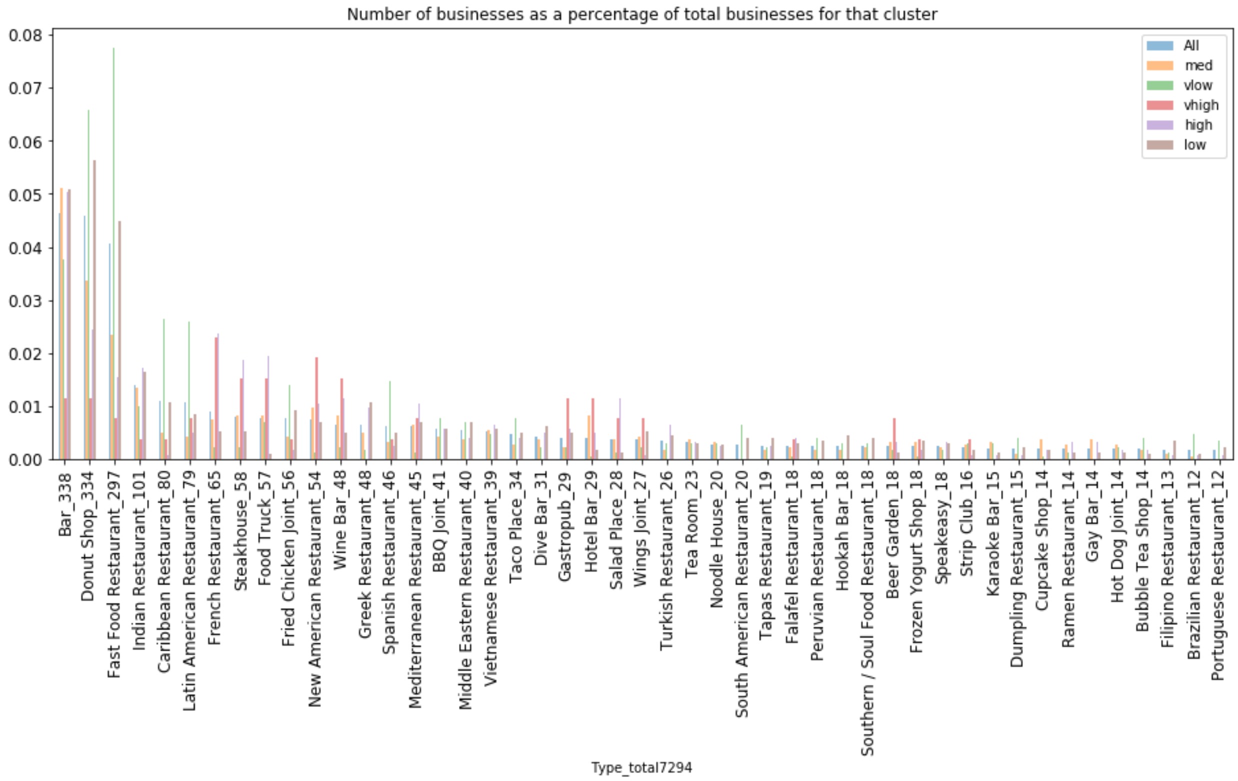
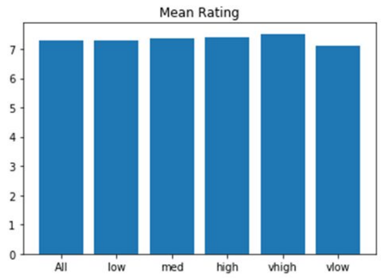
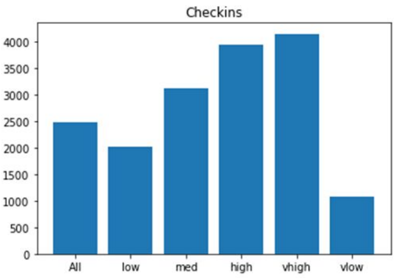
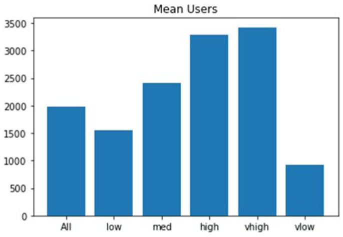


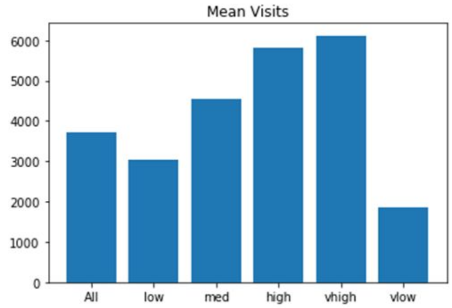
Figure A.9: Distribution of total of business 'Type' as a percentage of the total number of businesses

Mean Values









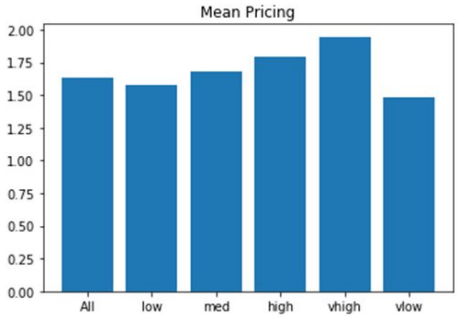


Figure A.10: Distribution of mean number of business 'Rating', 'Check-ins', 'Users', 'Visits', and 'Pricing'

Mean Values per Person

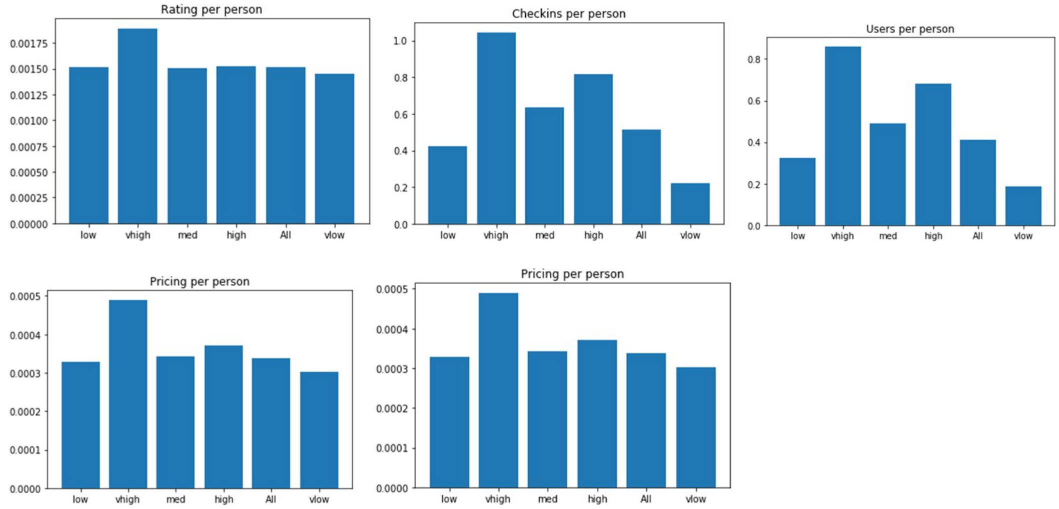


Figure A.11: Distribution of mean number of business 'Rating', 'Check-ins', 'Users', 'Visits', and 'Pricing' per person

Appendix B

Twitter Results

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 B.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.

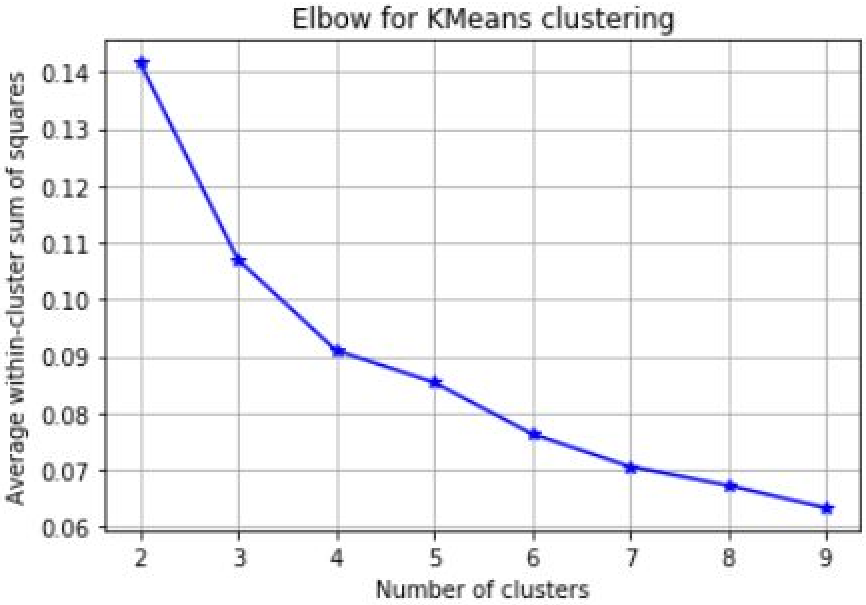


Figure B.1: Elbow plot for K-means clustering of Twitter data. There is a clear drop in performance after cluster 5.

Results are detailed in Figure B.2 below:

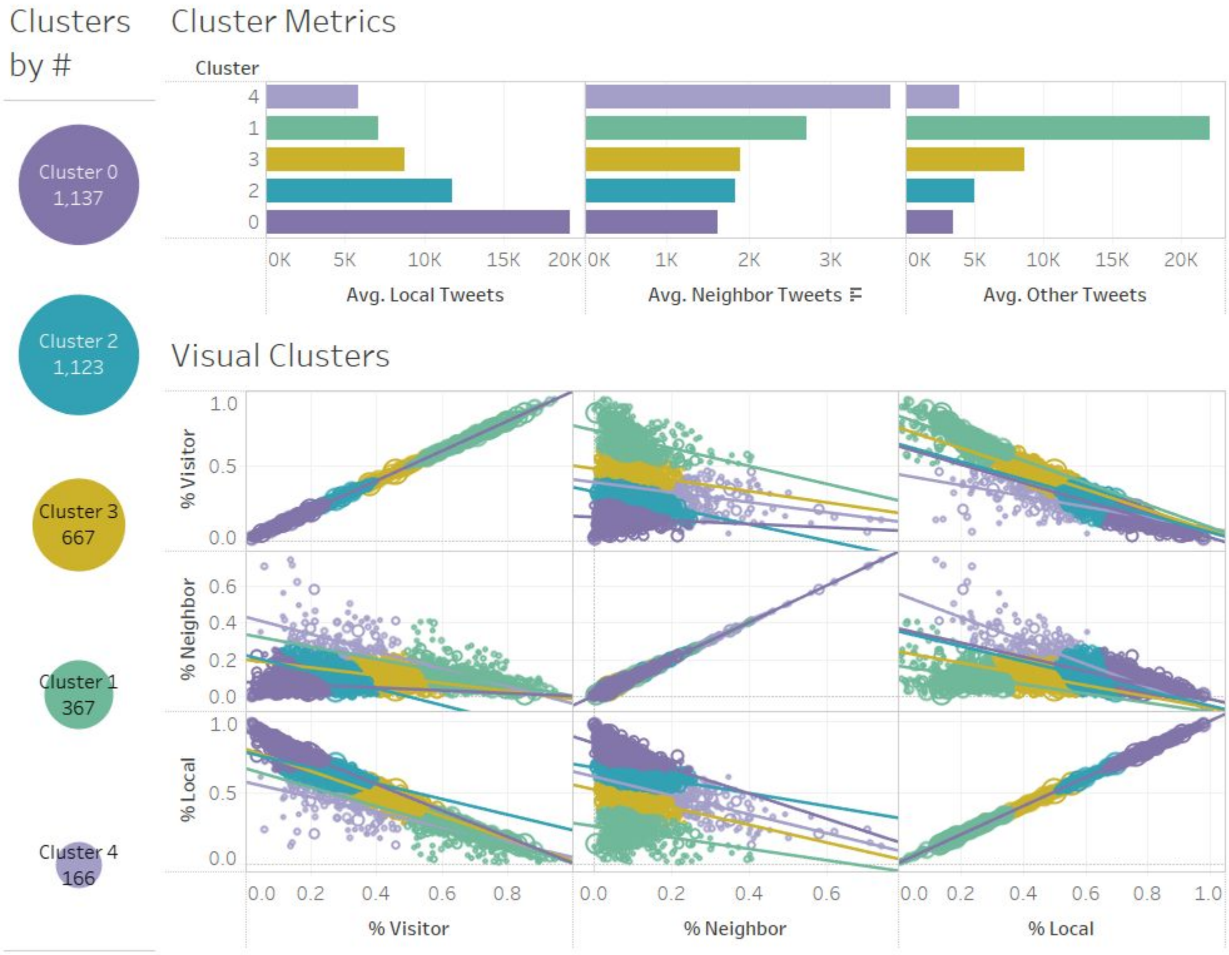


Figure B.2: Interactive dashboard which enables exploration of the cluster results. Clusters 0 and 2 represent the highest population of census tracts, dominated by Local Tweets. Clusters 1 and 4 represent an interesting subsection dominated by visitor and neighbor tweets, respectively.

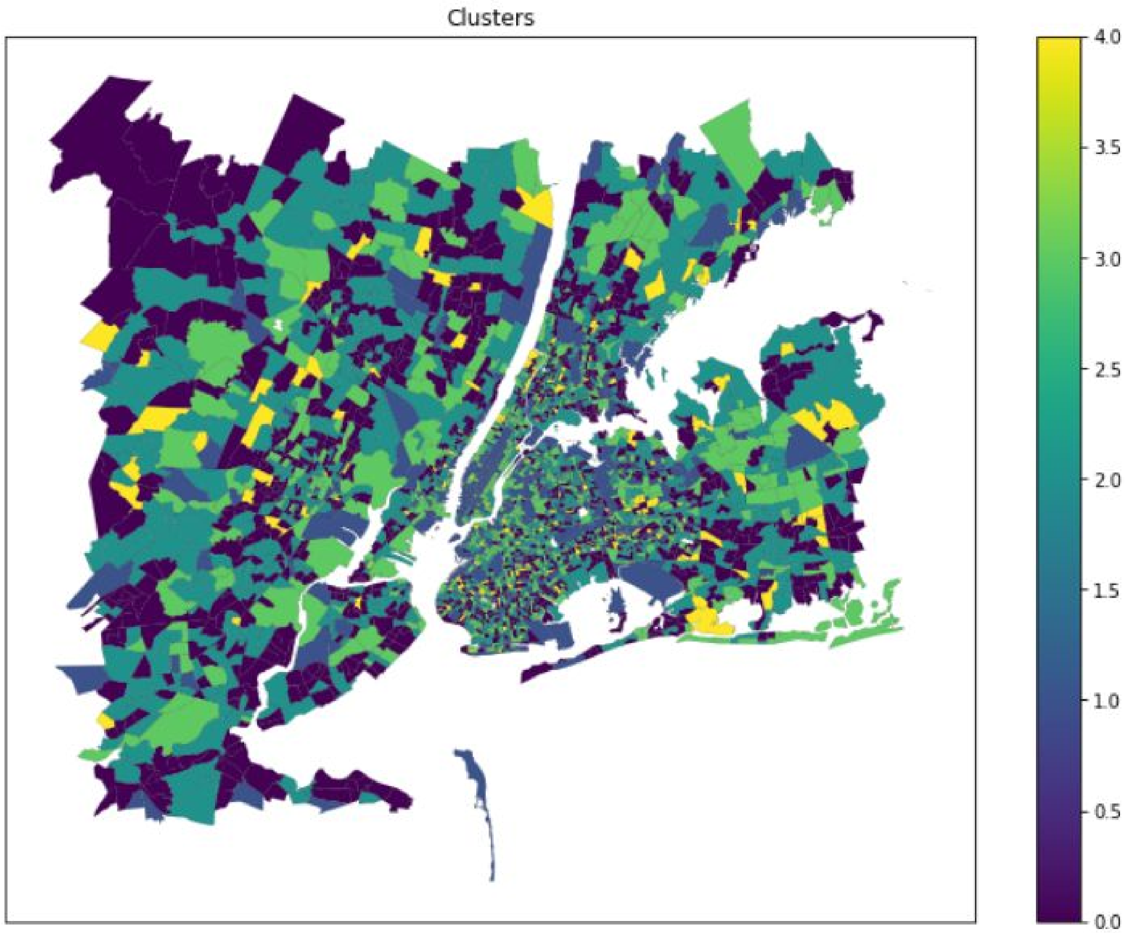


Figure B.3: A map of the cluster results in NYC and the surrounding counties. You can see that a large portion of Manhattan is Cluster 1, which is predominantly visitor tweets, as well as chunks of Brooklyn and Queens. Cluster 4, neighbor-dominated tweets, show up mostly scattered through Brooklyn, NJ,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

Figure B.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.

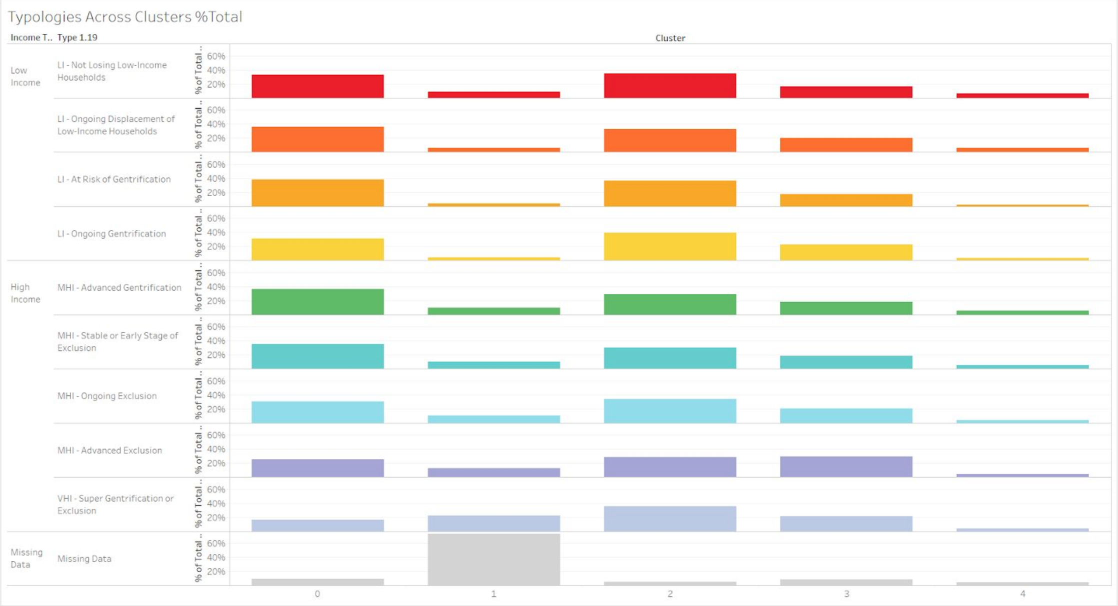


Figure B.4: This visualization shows the percentage of each typology that falls into the various clusters defined above.

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:

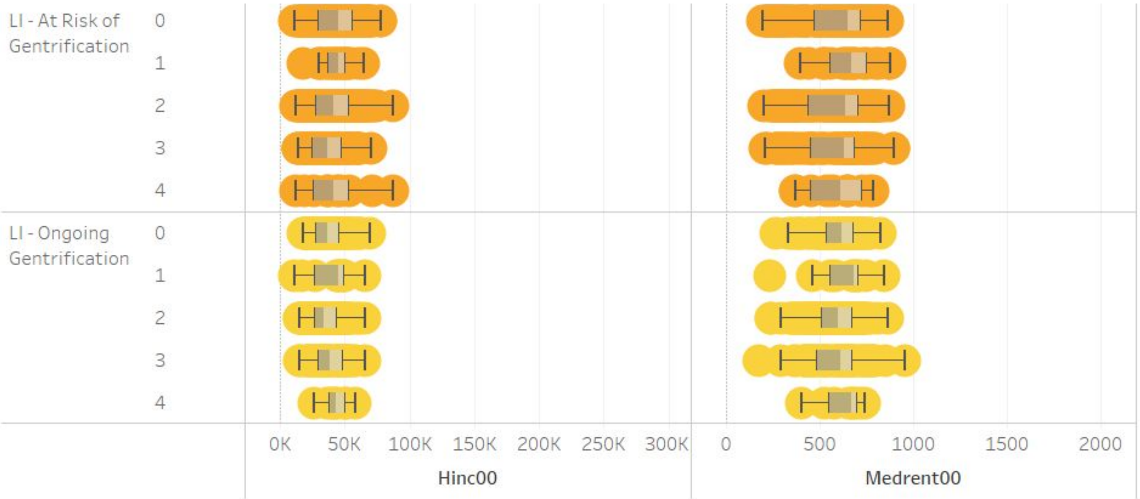


Figure B.5: This visualization shows a bar and whisker plot of Income and Median Rent for each census tract within the two target gentrifying typologies for each cluster. You can see that Cluster 0, 2, and 3 exhibit similar trends on these metrics, while clusters 1 and 4 seem to diverge. We have begun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

A draft version of interactive results are available at:
https://public.tableau.com/views/TwitterDeepDive/Aboutthisworkbook?:embed=y&:display_count=yes&:origin=viz_share_link

Appendix C

Distribution of Typologies for Foursquare Data

Table C.1 and Figure C.1: Distribution of Gentrification Typologies for Foursquare Data (All 8 Classes)

Typology	Percentage
MHI - Stable or Early Stage of Exclusion	0.379
MHI - Ongoing Exclusion	0.144
LI - At Risk of Gentrification	0.099
LI - Not Losing Low-Income Households	0.091
LI - Ongoing Displacement of Low-Income Households	0.087
MHI - Advanced Gentrification	0.070
VHI - Super Gentrification or Exclusion	0.059
LI - Ongoing Gentrification	0.052
Missing Data	0.019

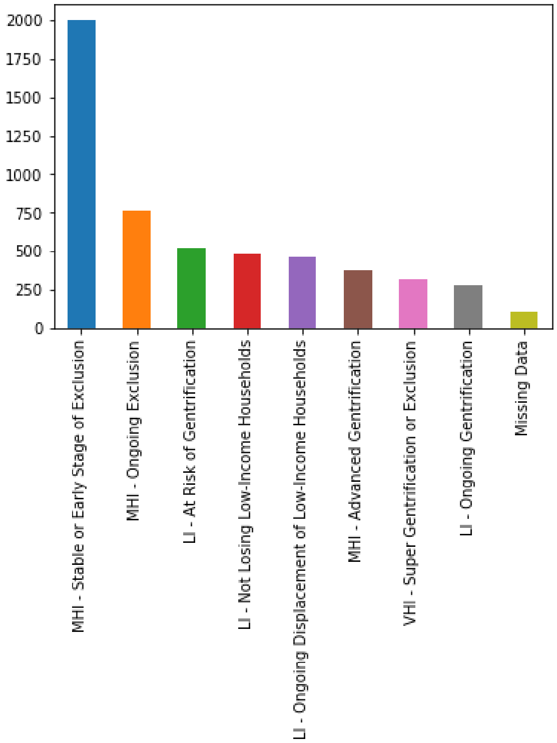
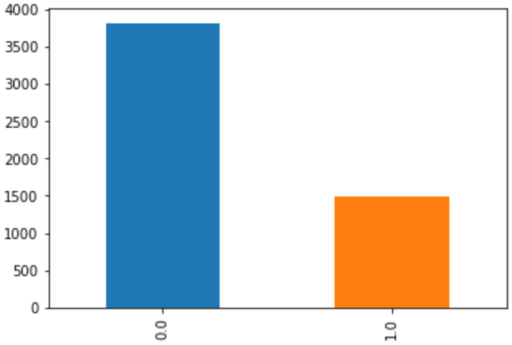


Table C.2 and Figure C.2: Distribution of Gentrification Typologies for Foursquare Data (Binary Classes)

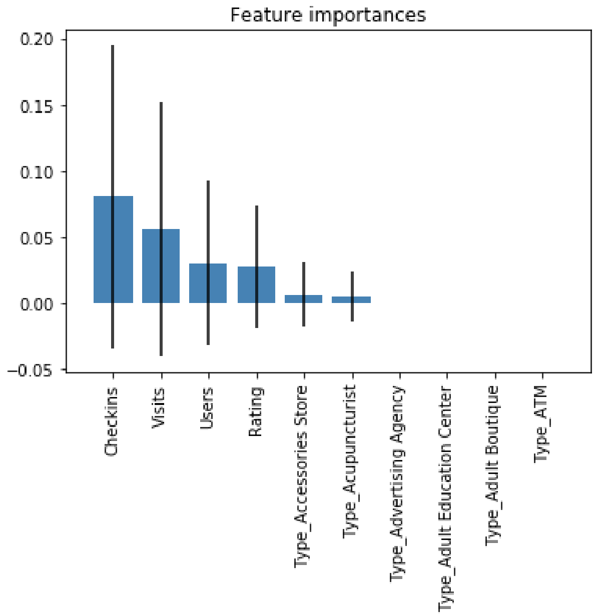
Typology	Percentage
No Gentrification (0)	0.720
Gentrification (1)	0.279



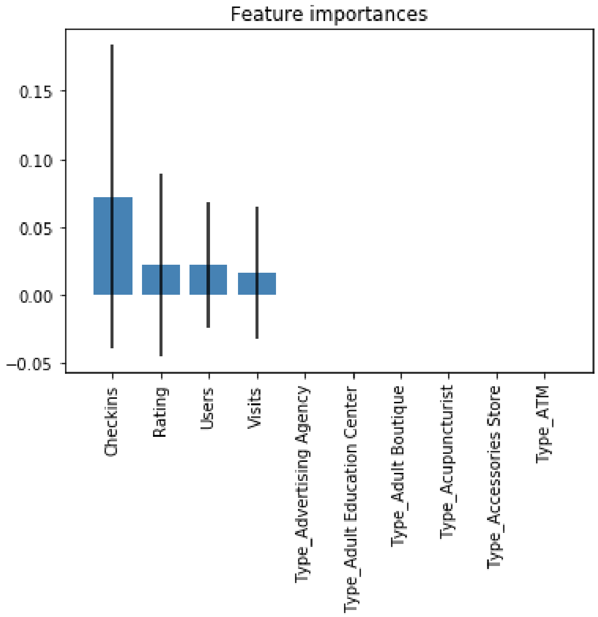
Appendix D

Modeling Results

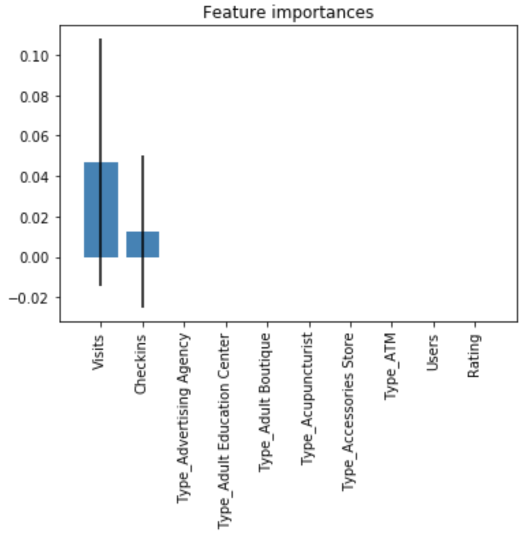
Feature Importance for Random Forest Models



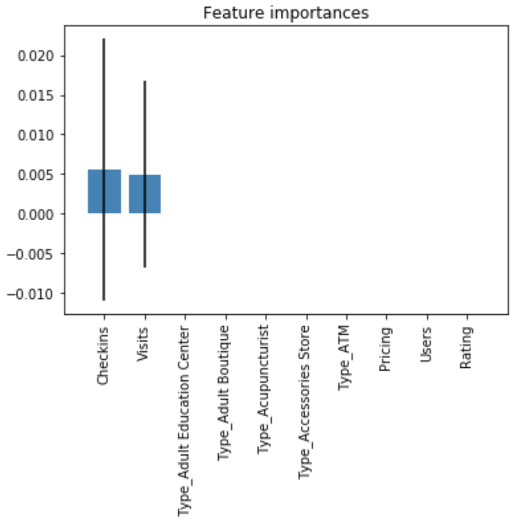
Model 1.1: Baseline Foursquare model



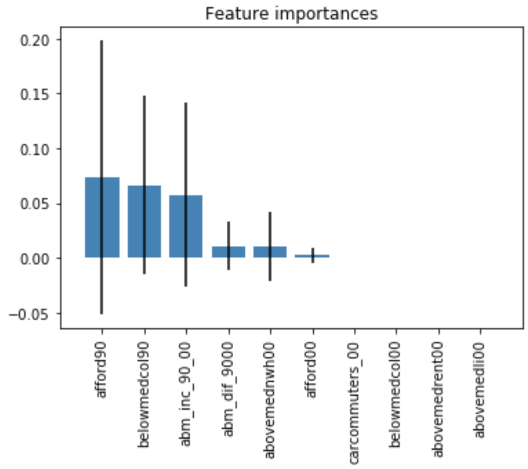
Model 2.1: Binary Model



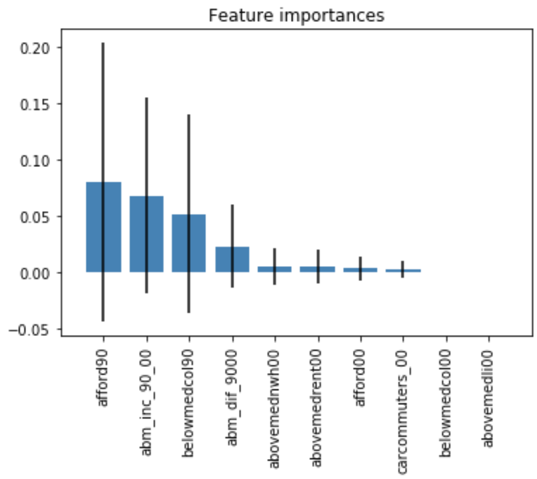
Model 1.2: Foursquare data by population



Model 1.3: Foursquare data by income level



Model 1.4: Census data (predictive model)

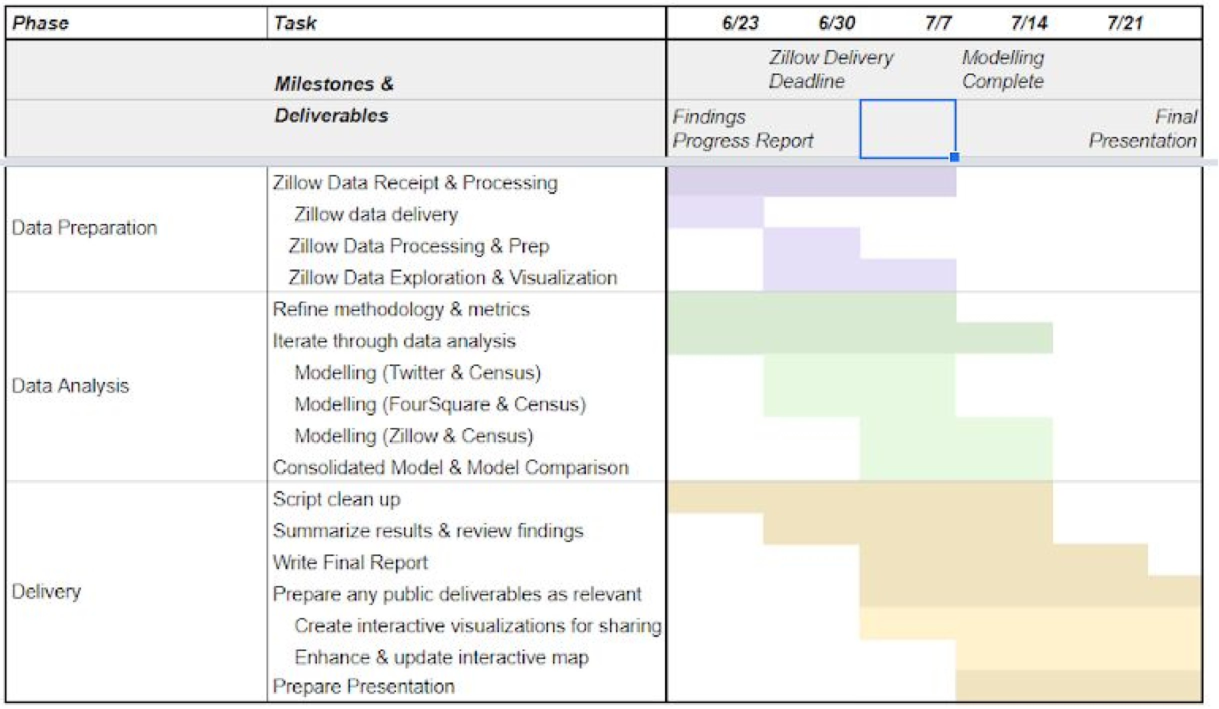


Model 1.5: Foursquare data with census data (predictive model)

Appendix E

Detailed Workplan

Gantt View Project Plan:



Phase	Task	6/23	6/30	7/7	7/14	7/21	
	Milestones &		Zillow Delivery Deadline		Modelling Complete		
	Deliverables	Findings Progress Report					Final Presentation
Data Preparation	Zillow Data Receipt & Processing	■	■	■			
	Zillow data delivery	■					
	Zillow Data Processing & Prep		■				
	Zillow Data Exploration & Visualization		■	■			
Data Analysis	Refine methodology & metrics	■	■	■			
	Iterate through data analysis	■	■	■	■		
	Modelling (Twitter & Census)		■	■			
	Modelling (FourSquare & Census)		■	■			
	Modelling (Zillow & Census)			■	■		
	Consolidated Model & Model Comparison			■	■		
Delivery	Script clean up	■	■	■	■		
	Summarize results & review findings		■	■	■		
	Write Final Report			■	■	■	
	Prepare any public deliverables as relevant			■	■	■	■
	Create interactive visualizations for sharing			■	■	■	■
	Enhance & update interactive map				■	■	■
	Prepare Presentation				■	■	■

The above project plan represents key activities and milestones to enable us to complete this research project. The next several weeks will be focussed on completing the Twitter and Foursquare analysis and enhancing existing models. In the final phase, we will create a consolidated model of our best features and compare performance across models. These results will be combined with our results to date and input into the final report and public facing web-based deliverables.

Key Risks:

- Zillow data has not been delivered to date. If Zillow data is not received by month end, the team will not have sufficient time to fully integrate it into the models and may have to reduce the scope of analysis.
- Models today have not provided convincing results and, while there is a plan to further enhance these models, there is a risk that results will still fall short of expectations by project end. However, great learnings and case studies have already been identified in this initiative that will still provide benefit to our sponsor and the public.