COFFEE FOR THE NEWCOMERS: A CASE STUDY OF GENTRIFICATION IN BROOKLYN'S NEIGHBORHOODS

Rushil Mallarapu[1]

[1]Fairfield, Connecticut, USA
rushil.mallarapu@gmail.com

ABSTRACT

We study the relation between neighborhood median household income and neighborhood character in Brooklyn to quantify the changes gentrification brings to neighborhoods outside traditional standard-of-living metrics. We use clustering to find groupings of neighborhoods with similar character, in terms of the type of venues common in the region, and search for regression models between neighborhood venue frequency and median household income. This case study demonstrates the effect gentrification has on the urban community, and offers valuable insight for government planners and business owners looking to develop cities.

KEYWORDS

New York City, gentrification, data analysis, neighborhood clustering

1. INTRODUCTION

Urban life is a defining facet of modern societies. Over the past century, we have seen the effects of technological development, economic turmoil, and social upheaval on the institutions of city life. Across the world, the forces of progress coalesce around the city, for good reason. Cities bring large groups of people of diverse backgrounds in contact with jobs, resources, and most importantly, each other. For businesses, cities represent both trendsetters and target markets for investment and consumer goods. It is critical for citizens and stakeholders that cities develop in a way that spurs positive growth while being more inclusive and meeting the needs of their residents.

Inevitably, any discussion of the development of cities circles back to one key phrase: gentrification [1]. Defined by the CDC as the increase in value of a neighborhood, gentrification has been charged with driving up prices and forcing out a neighborhood's former residents – often lower-income people of color. City mayors shy from the word to show they do not support the displacement of people from their homes and communities, unwilling to acknowledge the unbreakable tie gentrification has with urban progress [2]. In truth, numerous studies, most recently by the Federal Reserve Bank of Philadelphia, have found that gentrification is not in itself a cause of displacement [3][4]. It is a symptom of an affordable housing problem, not a cause. Gentrification has been shown to reduce the exposure of residents to poverty, improving neighborhood conditions, and stimulating diversification of areas. Far less than tearing down low-income households to make coffee shops for newcomers, gentrification has allowed cities to modernize and provide people a better standard of living.

While the literature on the causes of gentrification is clear that it does not lead to the displacement of low-income communities, little research exists on the effects of gentrification on the community character of neighborhoods outside of traditional metrics (e.g. racial diversity, education access) [4]. Understanding how the modernization and influx of money to a

community sets it apart from low-income neighborhoods who have been overlooked by this wave of development is crucial to understanding the current development of urban society. City governments need to know where to allocate funds, and for what causes. Business owners need to know where ventures will be more successful. We perform a targeted analysis of neighborhoods in Brooklyn, New York City, finding the distribution of neighborhoods of similar characters and understanding how median household income and neighborhood character are related.

2. DATA

Data for this project was acquired from four sources. The overall project had three goals for its data. First, it had to provide information about the geography of Brooklyn's neighborhoods. Second, it had to show the median household income of Brooklyn on a neighborhood-wide resolution. Finally, it had to give information pertaining to the character of a neighborhood – the kinds of places and institutions that give a neighborhood its "feel." All data used in this project is on GitHub (https://github.com/sudo-rushil/Coursera_Capstone/).

2.1. Data Sources

Geographic data about neighborhoods in New York was acquired from New York University's Spatial Data Repository. This data was collected in 2014 by the New York City Department of City Planning [5].

GeoJSON files outlining the border of neighborhoods in Brooklyn was obtained from www.click-that-hood.com, an open-source project. The data was collected in 2014.

Neighborhood income data was acquired from the US Census Bureau 2018 American Community Survey 5-year estimates. Income data for all zip codes in Brooklyn was downloaded from table B19013 [6][7].

Neighborhood venue data, which was used as a surrogate for neighborhood character, was acquired through querying the Foursquare API. This data is recent as of the date of collection [8].

2.2. Data Cleaning

Geographic data on NYC neighborhoods was acquired as a GeoJSON file, and needed no further cleaning. Neighborhood names and locations in Brooklyn were extracted into a Pandas DataFrame.

GeoJSON data on Brooklyn neighborhood boundaries was processed to extract all the neighborhood names used, which were then used as a key for the processing of the income table. This data did not require reformatting.

Neighborhood income data from US Census Bureau table B19013 was gathered for all zip codes in New York City. From here, using the neighborhood names in the Brooklyn GeoJSON data, income data was paired with the corresponding neighborhood manually. This cleaned file was manually edited for consistency with other data sources and extraneous metadata was removed. It was then saved as a comma-separated values (CSV) file and imported into a Pandas DataFrame [6][7].

Neighborhood venue data, having been acquired from the Foursquare API, was processed live during exploratory data analysis, and was not cleaned prior to retrieval. A custom processing pipeline was written in python to extract venue information from each response body, which was performed in tandem with data retrieval. After initial retrieval, the DataFrame was written to memory as a CSV file for future consistency.

3. METHODOLOGY

This analysis of this research went through three distinct phases. First, data was acquired and prepared, as described above, and visualized. Second, neighborhoods were matched with venue data and clustered, and the relation between clusters and income was examined. Finally, correlations between venue frequencies and income were analyzed. This section details the process behind these steps. For the results of these two analyses, see the following section.

3.1. Data Processing

We first loaded the geographic data. This was processed from two GeoJSON files. The first defined the neighborhood locations, while the other defined the neighborhood boundaries in Brooklyn. This defines the scope of the research's area of study. We overlaid these two maps to give Figure 1, shown below.

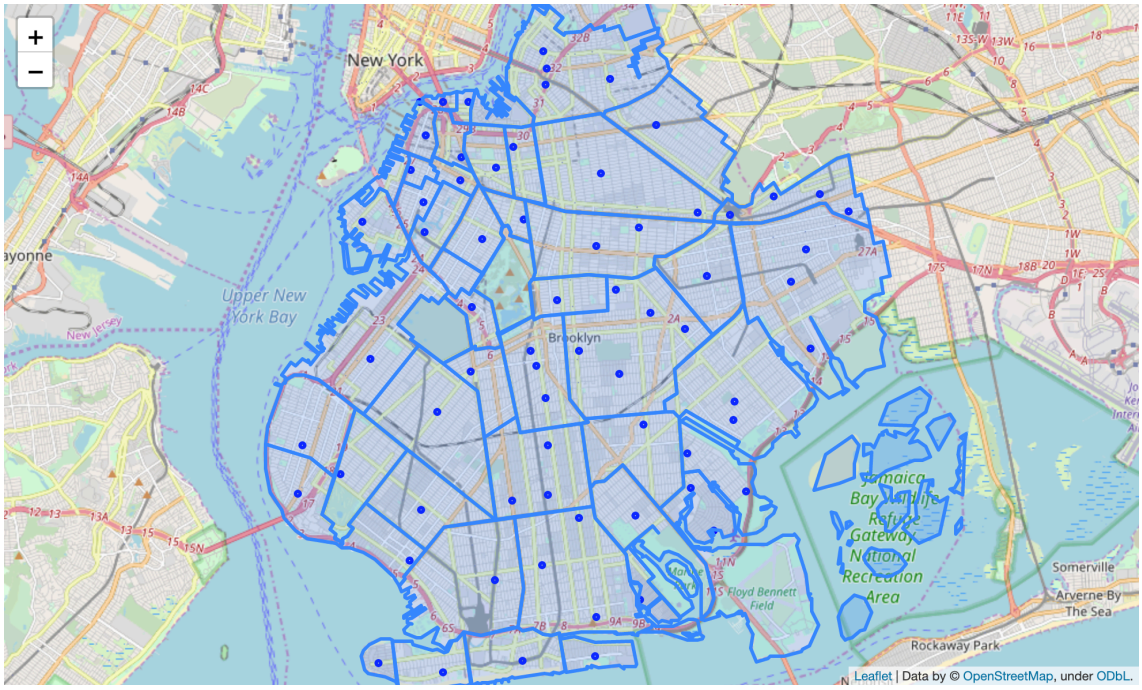

Figure 1. Neighborhoods and region boundaries. The blue dots represent neighborhoods for which venue data was collected, and the blue polygons represent regions of Brooklyn.

Next, we loaded income data after manually processing it by matching community zip codes to the zip codes in the median household income data from the US Census Bureau's 2018 American Community Survey 5-year estimate. The income dataset was created to ensure a lossless join to the region boundaries. We visualized these incomes using a choropleth map, to see the geographic distribution of wealth in Brooklyn. This is shown in Figure 2. On the image, darker green corresponds to higher median household incomes, and lighter green corresponds to lower median household incomes. Preliminarily, we see that the northwestern corner of Brooklyn, as well as the side facing out into the bay, are the wealthiest. Conversely, the poorest regions are concentrated to the east of the borough. Luckily, there is a variety of neighborhoods, often multiple in each district. As such, it is likely that some isolated neighborhoods will have a higher average income than the larger district they are found in.

We use the Foursquare API's explore endpoint to find the top venues in the vicinity of each of the neighborhoods in the neighborhood DataFrame. We use a custom request processing pipeline to turn each request body into a list of venues organized by neighborhood and category.

The information about how venue categories are distributed throughout neighborhoods is a good indicator of a neighborhood's "soft" character - the things that give its unique feel, its qualities outside of having good schools and clean streets.

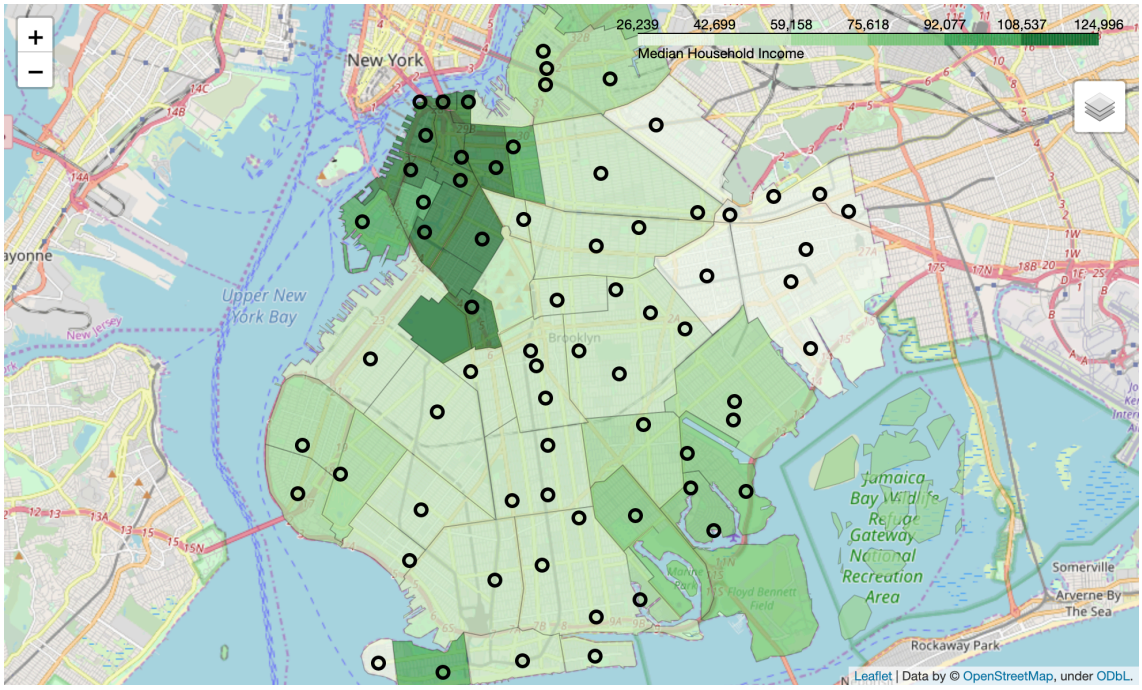

Figure 2. Median household income choropleth map. Dark green corresponds to high median household income, and light green corresponds to low median household income. Black circles represent neighborhoods of study.

3.2. Venue Clustering

In order to determine relations between neighborhood wealth and neighborhood character, we need to find clusters of neighborhoods with similar characteristics, as determined by their distribution of venues. After collecting the venue data, we generate a one-hot encoding of each venue category, and average these values for each neighborhood to find the frequency of each kind of venue in each neighborhood. We additionally find the top ten most frequent venue categories in each neighborhood, as a qualitative descriptor of neighborhood character.

We perform KMeans clustering over the venue category frequency dataset to group the neighborhoods in clusters. Figure 3 shows the cluster size distribution for each trial value of k, the number of clusters, between 3 and 6. From these distributions, we see that using $k > 4$ results in some outlier clusters of only one venue, which is not helpful. On the other hand, using $k = 4$ still provides more defined valid clusters, while preserving the major groupings observed for $k = 3$. As such, we choose to group the neighborhoods into four clusters. The results and analysis of this clustering is shown in Figure 4 and described in the following section.

3.3. Income Correlation Modeling

To see how well the frequency of different venues correlates with income, we first join the income and venue frequency datasets on the neighborhood key. This gives us, for each neighborhood, the median household income and the frequency of all categories of venues studied. Using the DataFrame correlation method in pandas, we found the top twenty venue

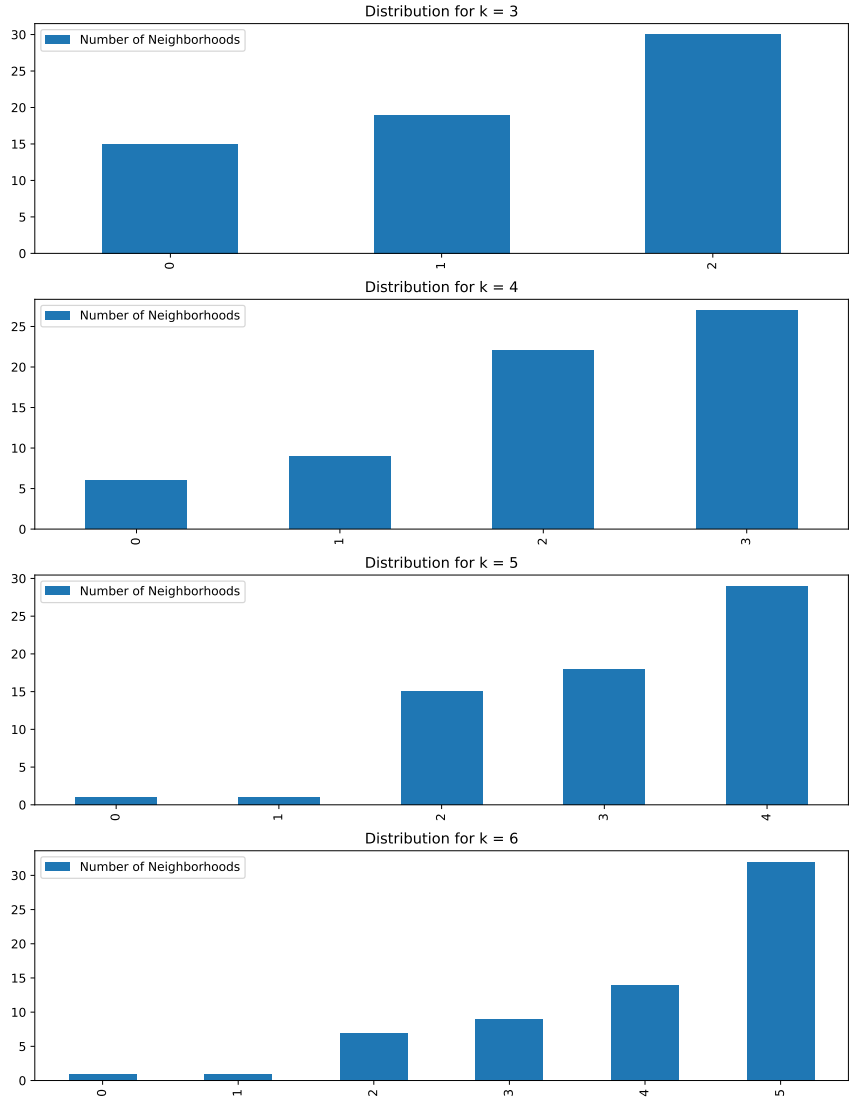

Figure 3. Cluster size distributions from KMeans clustering for varying values of *k*.

categories whose frequencies correlated with income. These venues are shown in Figure 5, while a correlation heat map of these venue categories and income is shown in Figure 6.

To determine whether this relationship could be modeled well, we ran experiments with ridge regression and linear regression models. We used a reduced frequency dataset consisting of only the top 19 most correlated venue categories as features, and the neighborhood incomes as the targets. The results of the model's predictions on training and test data are detailed in section 4.2. We determined that the data is inherently sparse, while showing reasonable regression R-squared values between venue category frequencies and income, which are depicted in Figure 7.

4. RESULTS

4.1. Neighborhood Clusters and Income

Figure 4 shows the distribution of neighborhood clusters overlaid onto the choropleth map of median household incomes in Brooklyn. We see visually that neighborhoods in more affluent parts of Brooklyn generally belong to different clusters, indicating they have some differences

in neighborhood character. We proceed to discuss the observed features of each of the four clusters in terms of their most common venues. See the project notebook for specific details on the most common venues (https://github.com/sudo-rushil/Coursera_Capstone/blob/master/project.ipynb).

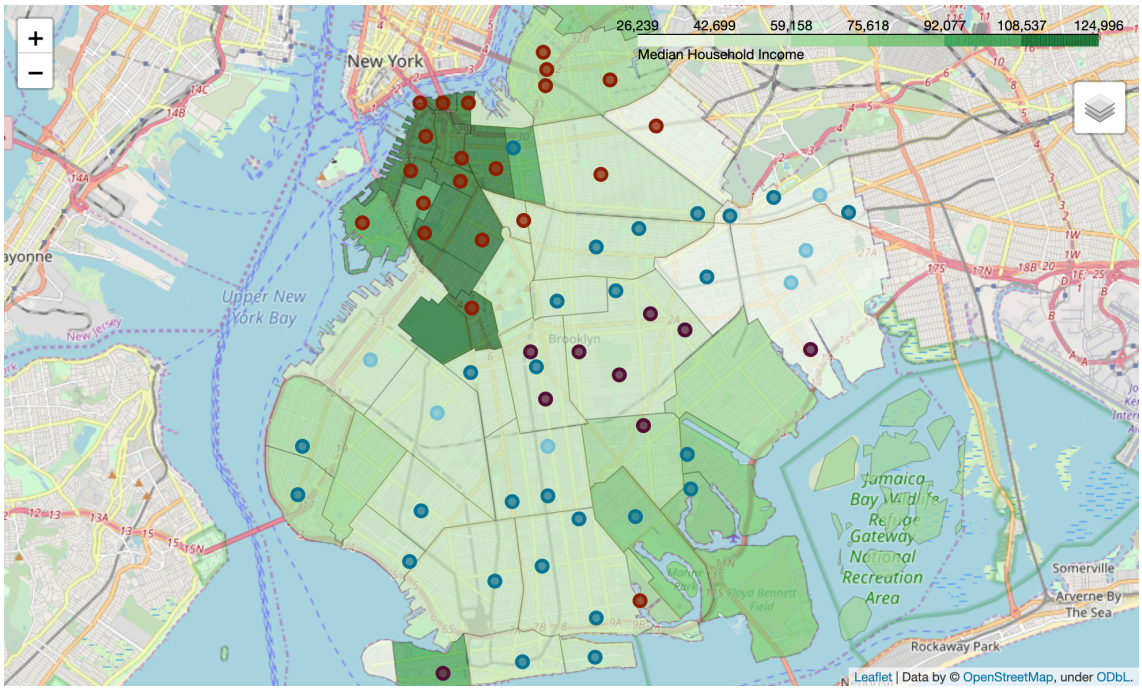

Figure 4. Neighborhood cluster distributions throughout Brooklyn overlaid onto median household income of neighborhoods.

The neighborhoods in the light red are mainly concentrated towards northwest Brooklyn, which has a much higher median household income than the rest of the borough. From the most common venue dataset, we observe there is a large frequency of bars, coffee shops, and cafés, which indicate an affluent, bohemian culture. We also see more traditional American style food, like pizza places and burger joint, as well as a variety of food and activities from other cultures, such as Chinese restaurants or yoga studios. These indicate that the residents of these neighborhoods have the capital to spend on homely food. This also shows that these areas are likely populated mainly by white residents, who like a mix of American meals and occasional foreign cuisine. This distribution of venues is in line with how gentrification brings a range of upscale venues to a region, helping diversify the area in terms of quality and regional background.

The neighborhoods in dark red are mainly in middle to low income regions, and are geographically close together. Looking at their venue distribution data, the most obvious factor is the prevalence of Caribbean restaurants, along with grocery stores and pizza places. This reflects one of the common trends in gentrification: When neighborhoods do not gentrify, i.e. when they remain at a low level of wealth, ethnic enclaves tend to solidify. This prevalence of one kind of restaurant indicates that this cluster likely represents a set of neighborhoods with a very similar ethnic background. This indicates that these neighborhoods have been passed up by the process of gentrification, as it tends to introduce a more diverse community character to a region.

The neighborhoods in dark blue are more spread out, and encompass a mixture of high, mid, and low income regions of Brooklyn. Their most prominent feature is their diversity of cuisine,

as well as their traditional urban locales of groceries and delis. Unlike the neighborhoods in dark red above, these places show a healthy range of ethnic diversity and price points. This indicates some level of gentrification, due to the ethnic mixing, and signals that these communities, with their diversity of incomes, are still in the process of urban development.

Finally, the points in light blue show are located in some of the lower income regions of Brooklyn. They mainly have low-budget venues, such as pizza places or fried chicken. There is also little diversity in the kinds of venues here. This again reflects how a lack of gentrification results in the deepening of regional homogeneity. These neighborhoods have not seen the same uptick in urban development as in some of the other clusters, thus resulting in communities that are characteristically barebones and underdeveloped.

4.2. Income Correlation Modeling Results

Figures 5 and 6 show the venue categories most correlated with high income, organized by absolute R-squared score, as well as a heat map of correlation strengths between these venue categories and income.

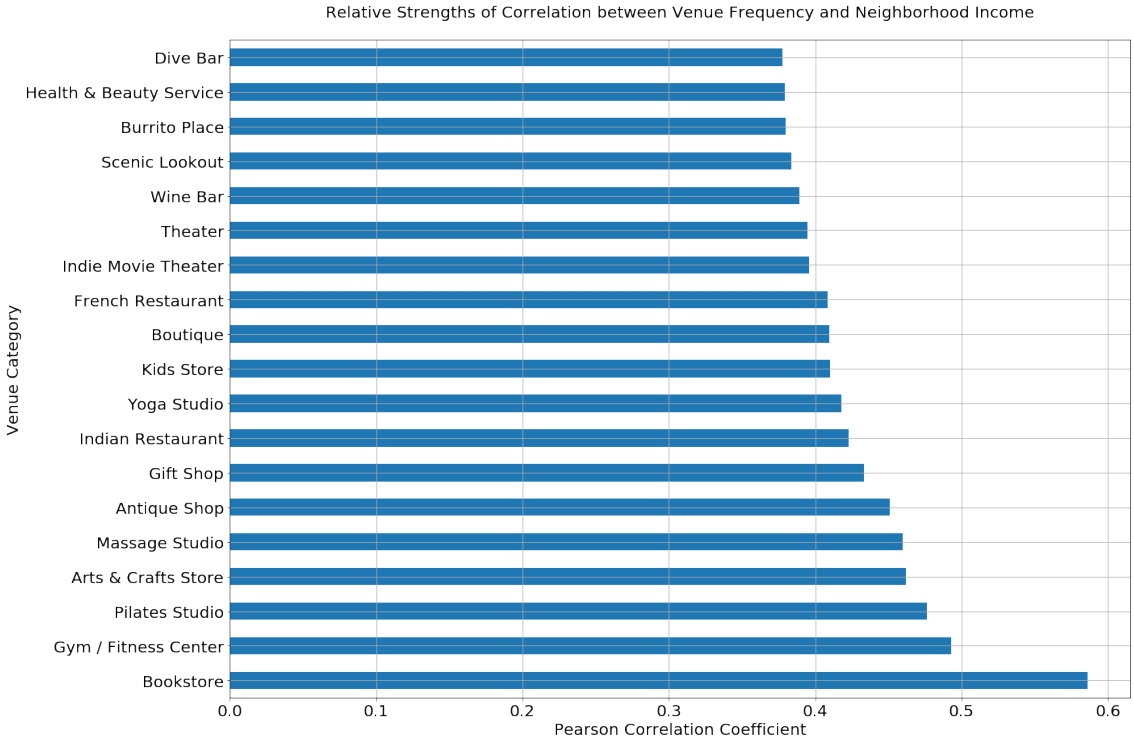

Figure 5. Top most correlated venue categories with income.

Observing the venues that are most correlated with income, we see that many of them include eclectic intellectual pursuits, such as antique stores, book stores, or indie movie theaters. This demonstrates how higher income communities often have more leisure time and can spend more on more abstract forms of entertainment. We also see a prevalence of fitness-related venues, like gyms, pilates studios, and yoga studios. This demonstrates how urban development enables the creation of more facilities for the betterment of public health. The final pattern to notice is the prevalence of foreign cuisine, like Indian and French restaurants. This reflects an earlier pattern on how gentrification leads to the ethnic diversification of a neighborhood as well as an increase in neighborhood income.

The heat map in Figure 6 shows that all these values have a decent correlation with income, as well as medium to strong correlations with each other. This indicates that there is likely a relation between these variables to be explored further, despite the data sparsity challenges.

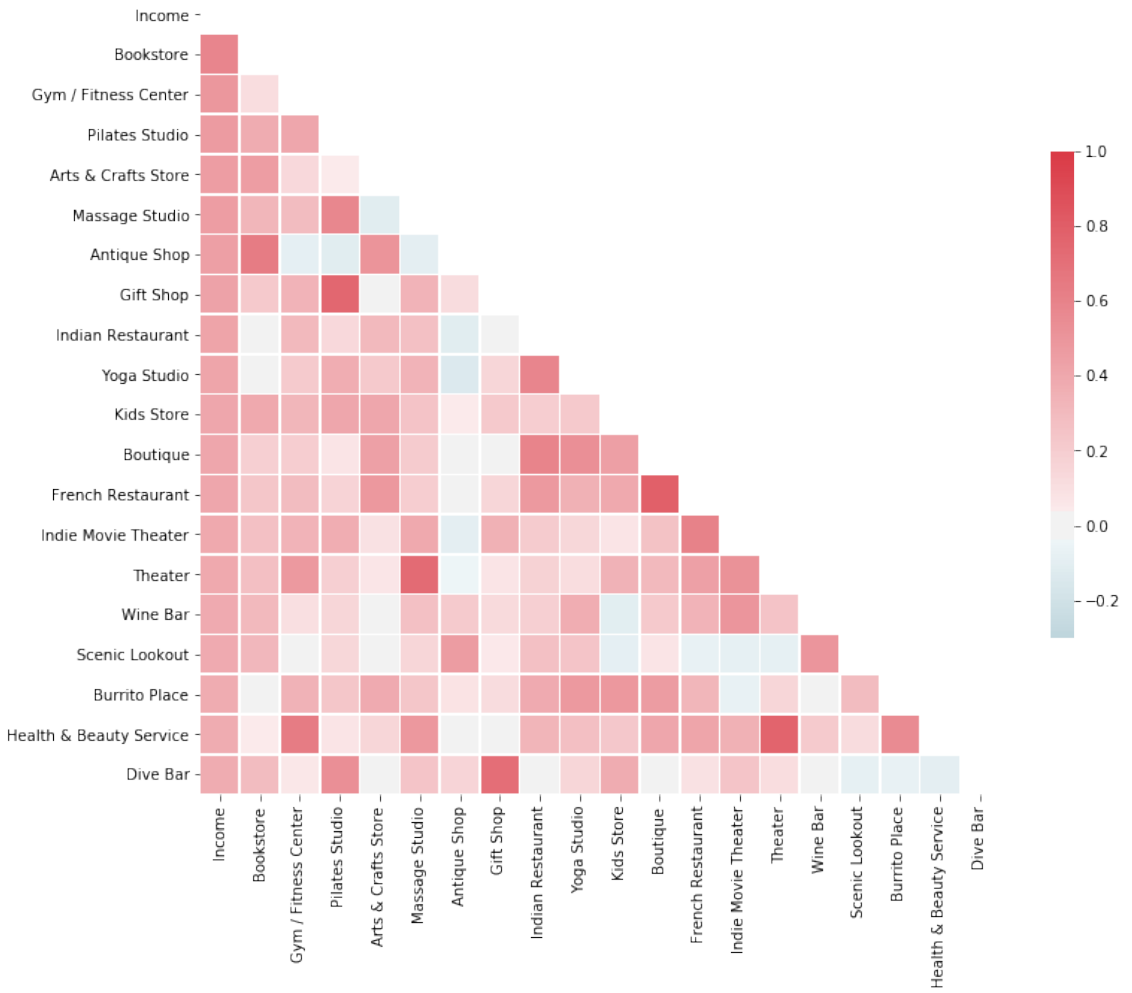

Figure 6. Heat map of correlations of top venue categories with income. Darker colors indicate higher correlations and paler colors indicate lower correlations.

We tested a ridge regression and linear regression model on the venue category/income data prepared as described in section 3.3. After fitting to the training data, both models had an R-squared of around 0.8 on the training data (ridge: 0.78, linear: 0.86). On the test data, the ridge regression model vastly outperformed the linear model in terms of median absolute error (ridge: 0.57, linear: 0.82). This is expected, due to ridge regression's better regularization than multiple linear regression. However, both of the models did only mildly well. This is primarily due to data sparsity, which we hope to resolve in future work on larger datasets with more venues under consideration.

Figure 7 is a plot of all the correlations between each of our regression variables and income, which helps visualize both the correlations and the data sparsity. These plots show that many of the correlations we observed might be weakened by the major data sparsity in our data. However, they are, to some degree present. In future work, it will be interesting to investigate whether these results are stronger after defining venues of similar characters themselves, thus preventing the data sparsity problem.

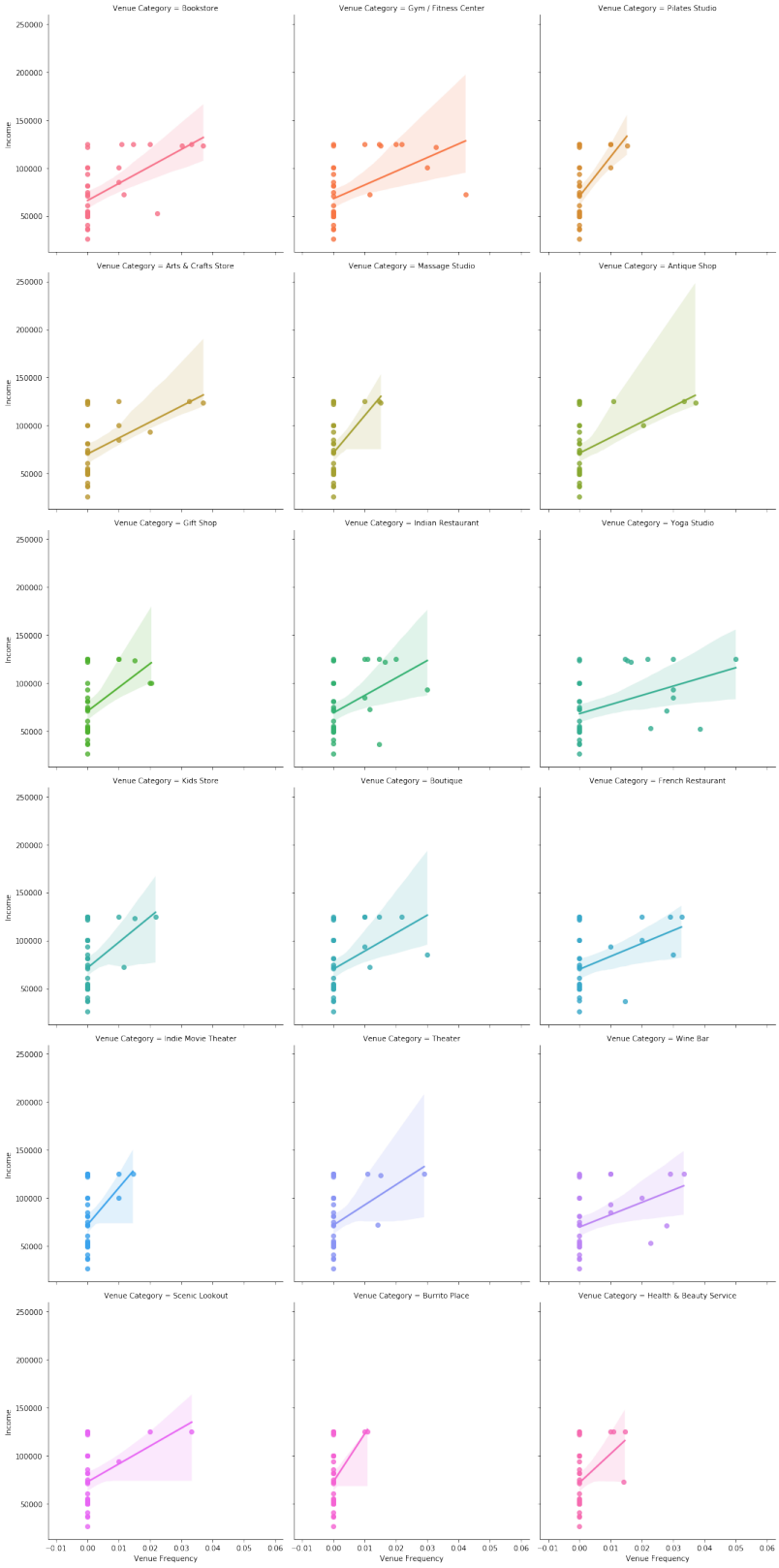

Figure 7. Regression plots of top venue categories with income.

5. Discussion

We sought to analyze how neighborhood characteristics and income overlap, as well as the relation between neighborhood venues and income. This information is indicative of the effects of gentrification, urban development, diversification, and resource influx into neighborhoods. It helps inform the planning of urban development projects as well as the analysis of their effects on communities.

Brooklyn is the largest borough of NYC by population, home to almost 2.6 million people, with an annual GDP of $91 billion [9][10]. As such, it is important for businesses to know where to invest resources into developing communities and raising the standards of living for all urbanites. This research shows how the presence and distribution of certain types of venues – a neighborhood's character – is related to income, highlighting how, as other research into the factual lack of housing displacement contends, gentrification leads to increased ethnic diversity and better facilities for the communities it affects. This information is the first step in getting a clearer picture over the kinds of communities our cities are home to, as well as what we should do to make them better.

6. Conclusion

We have examined the relation between neighborhood median household income and neighborhood character in Brooklyn to quantify the changes gentrification brings to neighborhoods outside traditional standard-of-living metrics. This research demonstrates the effect gentrification has on the urban community, and offers valuable insight for government planners and business owners looking to develop cities. Future work will use larger datasets over larger cities and hand-designed groupings of venues into super-categories to improve the data sparsity issue and find correlations between other indicators of a neighborhood's character and standards of living.

Acknowledgements

The author would like to thank his family for their constant support. He would also like to thank the instructors of the IBM Data Science specialization for their hard work in providing him the skills needed to complete this project.

References

[1] Governing. (2015, February). What, Exactly, Is Gentrification? Retrieved March 25, 2020, from https://www.governing.com/topics/urban/gov-gentrification-definition-series.html

[2] Richard Florida, C. L. (2015, September 16). This Is What Happens After a Neighborhood Gets Gentrified. Retrieved March 25, 2020, from https://www.theatlantic.com/politics/archive/2015/09/this-is-what-happens-after-a-neighborhood-gets-gentrified/432813/

[3] Buntin, J. (2015, January 15). Gentrification Is a Myth. Retrieved March 25, 2020, from https://slate.com/news-and-politics/2015/01/the-gentrification-myth-its-rare-and-not-as-bad-for-the-poor-as-people-think.html

[4] Capps, K., & Capps, K. (2019, July 22). The Hidden Winners in Neighborhood Gentrification. Retrieved March 25, 2020, from https://www.citylab.com/equity/2019/07/gentrification-effects-neighborhood-data-economic-statistics/594064/

[5] 2014 New York City Neighborhood Names. (2014). Retrieved March 25, 2020, from https://geo.nyu.edu/catalog/nyu_2451_34572

[6] Research Guides: New York City Data: Neighborhoods. (n.d.). Retrieved March 25, 2020, from https://guides.newman.baruch.cuny.edu/nyc_data/nbhoods

[7] U.S. Census Bureau. (2018). *Median Household Income in the Past 12 Months*. Retrieved from https://data.census.gov/cedsci/table?tid=ACSDT1Y2018.B19013&text=B19013&hidePreview=false&vintage=2018.

[8] Foursquare Developer. (n.d.). Retrieved from https://developer.foursquare.com/

[9] GDP by County, Metro, and Other Areas. (n.d.). Retrieved from https://www.bea.gov/data/gdp/gdp-county-metro-and-other-areas

[10] Population - Current and Projected Populations. (n.d.). Retrieved from https://www1.nyc.gov/site/planning/planning-level/nyc-population/current-future-populations.page