Segregation and Consumption Options in Chicago City

---- Explore the Underlying Patterns of Chicago Public Consumption Spaces with Machine-Learning Methods

Victoria Wang

1 Introduction

1.1 Background

Segregation has always been a critical issue for Chicago City. While the city's overall population is fairly evenly divided -- approximately 32 percent white, 30 percent black, and 29 percent Hispanic (ACS 2017) these people are segmented into 77 different communities, living starkly different lives -- and there are striking patterns involving who lives in which neighborhood, and why.

"The place that you live shapes so profoundly all aspects of your life," said Maria Krysan, a professor of sociology at the University of Illinois at Chicago. Indeed, Chicago citizens with different racial or economic backgrounds not only dwell in divergent residential environments but are also exposed to extremely different public spaces and resources, which in turn shapes their ways of living. This project looks into one specific aspect of the life of Chicago residents -- their consumption options in their respective living spaces.

1.2 Question

Consumption works as a way through which social belonging and distinction are built and expressed, and these destinations, targeting specific users, include some while excluding others (Bolzoni 2016). It is reasonable to assume that segregation in Chicago City can be reflected by not only people's living places but also their daily consumption -- in other words, their options on where to spend money on food, retail goods, activities, etc. This project aims to verify this assumption. **Apart from residential segregation, are citizens with different social backgrounds exposed to different public consumption places?**

The project does not assume or attempts to deduct a causal relationship between segregation and consumption options, but simply tries to show that some correlation exists between the two.

1.3 Interests

This can be a preliminary step in further studies of socioeconomic segregation in the U.S. urban areas. Previous studies are mostly focused on the residential distribution of the urban population, while this project intends to find patterns of segregation that are related to the use and accessibility of public consumption spaces. The result of the study will provide a snapshot of the segregation underlying consumption resources in U.S. urban areas. It can be of interest to scholars looking for a new lens to inspect segregation. It can also be helpful to policymakers, social workers and/or activists looking to reduce segregation in the U.S. urban areas.

2 Methodology

2.1 Data Acquisition, Wrangling and Clustering

In this project, a community's accessibility to a type of venue is measured by the prevalence -- or the frequency of occurrences -- of the type of venues. If certain types of venues are commonly present in a community area, it can be inferred that residents in this community are more exposed to these types of venues. Resident characteristics of individual communities are given by assorted indicators obtained from government-sourced census data. These indicators give sufficient information on residents' personal development, financial well-being, and racial identities in each community.

Venue information on the 77 Chicago community areas is obtained from the FourSquare API (a sample of venue information shown in table 2.1). I searched for venues within a 2.5km radius about each community center, with an upper limit of 100 venues per community. It is worth noting that the 2.5km search radius is not completely satisfactory for obtaining venues for small communities (community area radius less than 2.5km), which may lead to a false inclusion of venues belonging to their adjacent communities. A more detailed discussion about this

limitation will be presented in the discussion section.

	Community	Community Latitude	Community Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Rogers Park, Chicago, IL	42.010531	-87.670748	Morse Fresh Market	42.008087	-87.667041	Grocery Store
1	Rogers Park, Chicago, IL	42.010531	-87.670748	El Famous Burrito	42.010421	-87.674204	Mexican Restaurant
2	Rogers Park, Chicago, IL	42.010531	-87.670748	Rogers Park Social	42.007360	-87.666265	Bar
3	Rogers Park, Chicago, IL	42.010531	-87.670748	Glenwood Sunday Market	42.008525	-87.666251	Farmers Market
4	Rogers Park, Chicago, IL	42.010531	-87.670748	Lifeline Theatre	42.007372	-87.666284	Theater

The frequency of occurrence for each venue category in each community area is then calculated and the top 10 most frequently occurred venue types are picked out for each community area. Part of the resulting dataset is shown in table 2.2.

	Community	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Albany Park, Chicago, IL	Park	Coffee Shop	Pizza Place	Mexican Restaurant	Middle Eastern Restaurant	Korean Restaurant	Gym	Grocery Store	Bar	Bakery
1	Archer Heights, Chicago, IL	Mexican Restaurant	Discount Store	Mobile Phone Shop	Sandwich Place	Bank	Grocery Store	Pizza Place	Bar	Video Store	Seafood Restaurant
2	Armour Square, Chicago, IL	Chinese Restaurant	Pizza Place	Bar	Mexican Restaurant	Coffee Shop	Grocery Store	Bakery	Asian Restaurant	Art Gallery	Salon / Barbershop
3	Ashburn, Chicago, IL	Sandwich Place	Pizza Place	Park	Ice Cream Shop	Donut Shop	Shoe Store	Discount Store	Mexican Restaurant	Pharmacy	Chinese Restaurant
4	Auburn Gresham, Chicago, IL	Discount Store	Fast Food Restaurant	Sandwich Place	Pharmacy	Lounge	Seafood Restaurant	Fried Chicken Joint	Bar	Train Station	Southern / Soul Food Restaurant

2.2 K-Means Clustering by Venues

To put communities with similar types of venues into the same groups, I applied the K-Means Clustering method to the venue dataset.

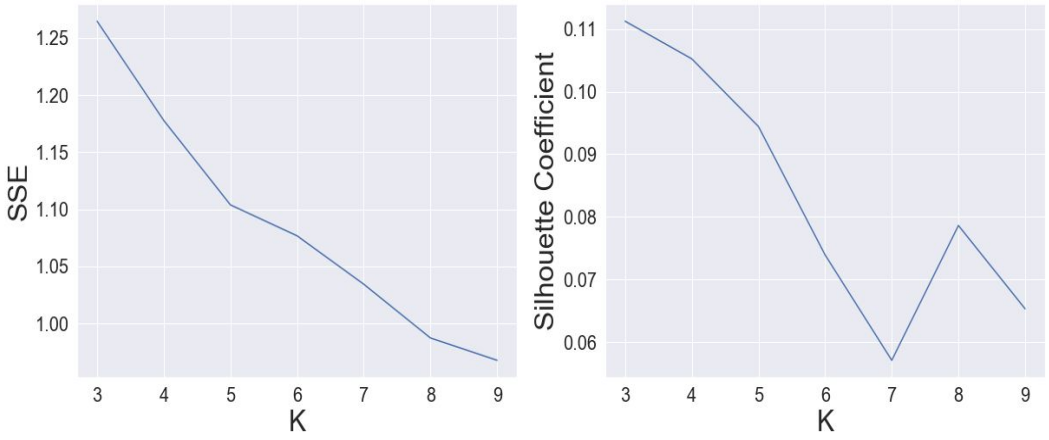
The accuracy of a K-Means model depends heavily on the number of clusters chosen for the model, namely, the value of K. To determine the optimal K-value for this model, I applied two methods to cross-evaluate K at values 3 to 10[1]:

1. The **Elbow Criterion** method. This method aims to find the smallest k-value at which the SSE (Sum of Standard Error -- distances of samples to their closest cluster centroid) turns

[1] I chose range 3 to 10 as k-values to be tested because the analysis in the next step would require more than 2 clusters for meaningful results.

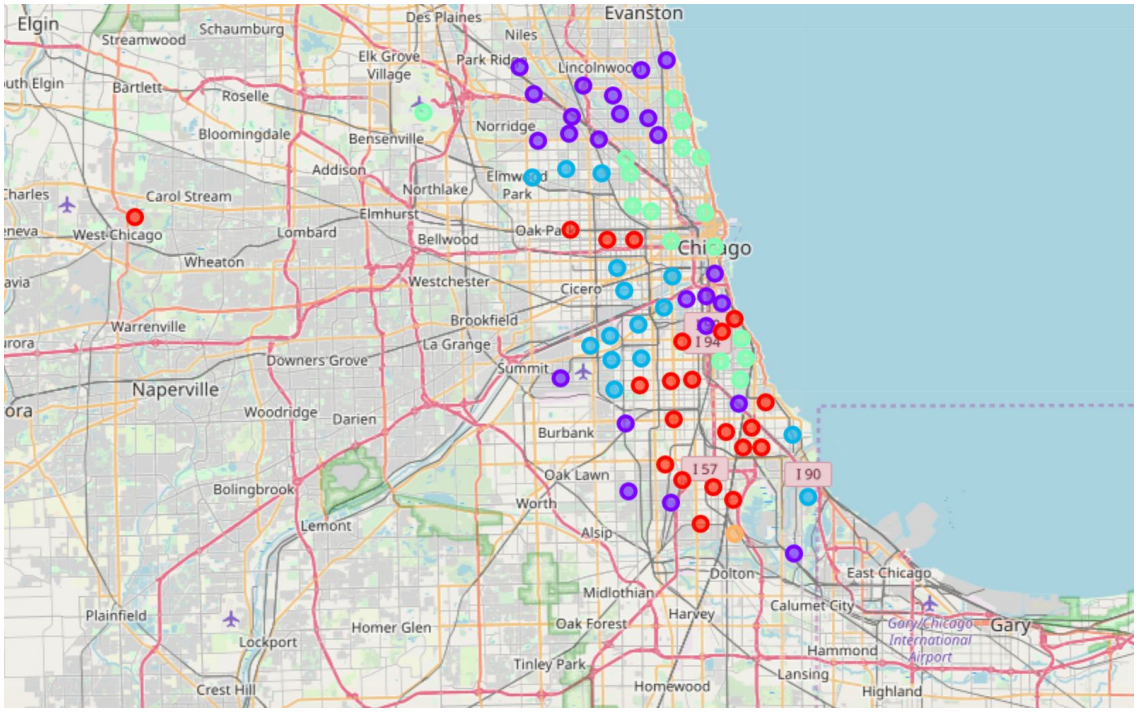
sharply close to 0. As shown in the first plot in figure 2.3, while the elbow shape is not fairly obvious, a comparatively optimal SSE shows at k = 5 or 6.

2. The **Silhouette Coefficient** measurement. The coefficient measures the closeness of data from its cluster centroid in relation to its distance from other cluster centroids. In other words, a higher Silhouette Coefficient indicates denser clusters and more distinct boundaries in between each cluster. As shown in the second plot in figure 2.3, while all Ks yield fairly low Silhouette Coefficients, the number is higher when k = 3, 4 or 5.



Given the results of the above two measures, it is most optimal if Chicago communities are put into 5 clusters. Map 2.4 shows the 77 Chicago communities colored by cluster labels. Intuitively, communities marked in the same color on this map should have similar venue occurrence patterns.

Visually, the five clusters are somewhat geographically segmented, despite some outliers in the mix. Cluster 0 (red) is mostly in downtown Chicago; clusters 1 and 3 (purple and green) in middle-uptown, while cluster 2 (blue) is in the western part of the city. Cluster 4 (orange) is one community on its own, located in the very south of the city.



The second part of the data collection is to obtain resident characteristics of each community area. For this part, I derived data from two sources: the *Chicago 2017 Census data* and the *Chicago Community Snapshots*. From the two datasets, I extracted the following resident attributes of the communities: median per-capita income; the percentage of residents who are under 18 or over 64 (i.e. not in the potential labor force); the percentage of residents who are over 16 and unemployed; the percentage of residents over 25 without a high school diploma (i.e. low education); the percentage of residents who are white, black, Hispanic, or Asian. These features are chosen empirically for they are the typical economic and social segregating determinants in Chicago communities.

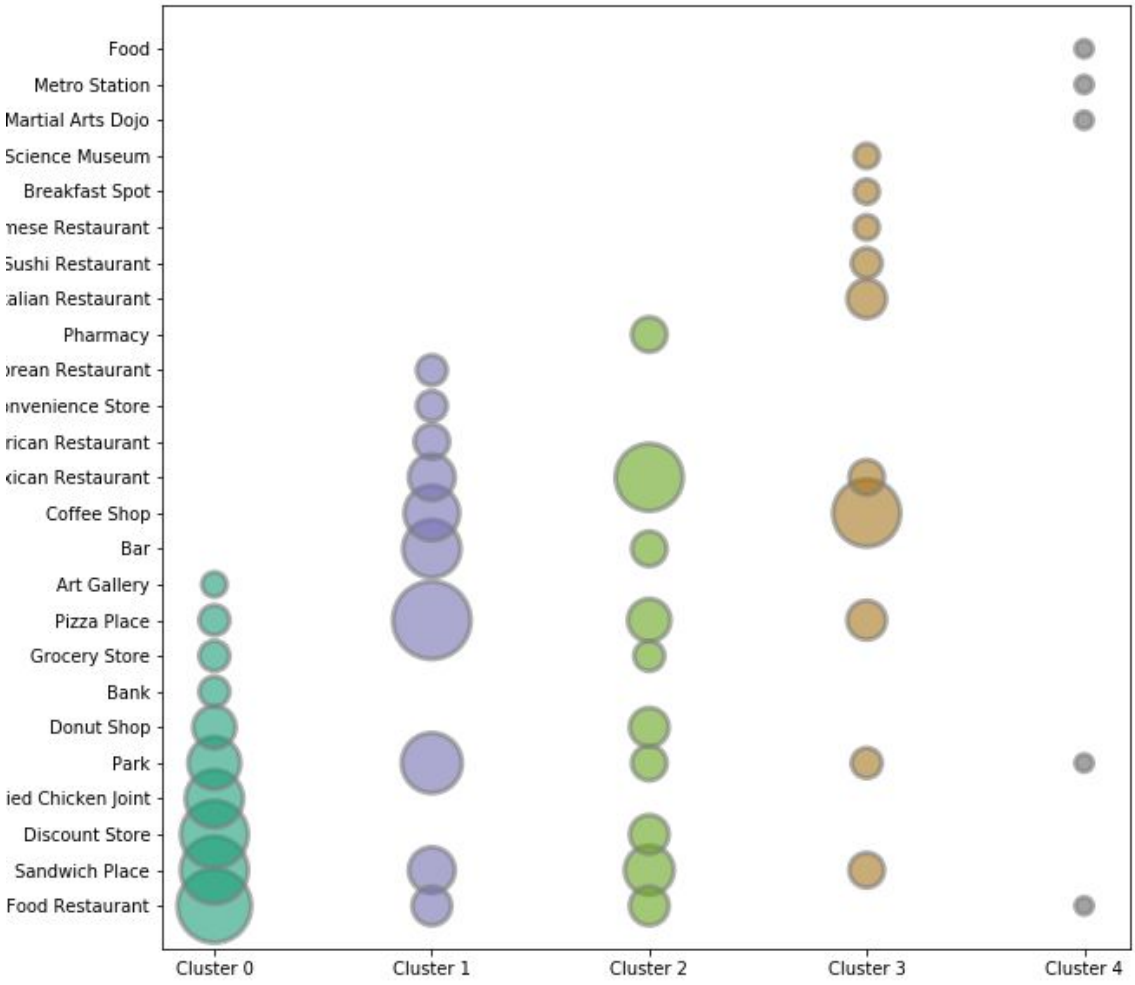
Cluster labels generated in the previous step are then inserted into the resident features dataset. Part of the final dataset is shown in table 2.5.

	ID	Community	Cluster Labels	%White	%Hisp	%Black	%Asian	per_capita_income	%below_poverty	hardship_index	%under18_over64
42	1	South Shore	0	0.03	0.01	0.94	0.01	19398	31.1	55	35.7
25	2	West Garfield Park	0	0.02	0.02	0.94	0.00	10934	41.7	92	43.6
24	3	Austin	0	0.05	0.13	0.81	0.01	15957	28.6	73	37.9
67	4	Englewood	0	0.01	0.03	0.95	0.00	11888	46.6	94	42.5
65	5	Chicago Lawn	0	0.04	0.50	0.45	0.00	13231	27.9	80	40.6

2.3 Explore Venue Patterns Between Clusters

First, I look at the clustering data on its own. I find the top 5 most frequently occurred venue types in each of the five clusters to examine in-cluster patterns and between-cluster divergence. Figure 2.6 shows a visualization of the result. It is quite apparent that different clusters of community areas are presented with distinct types of consumption choices -- and the

most diverging among which are food options.



As figure 2.6 shows, communities in cluster 0 and cluster 2 have almost solely fast-food restaurants, which, in common sense, is very unhealthy but filling, more convenient, and most importantly, cheaper. It is worth noting that cluster 2, in particular, not only has a frugal consumption environment overall but also has a large number of Mexican restaurants as well.

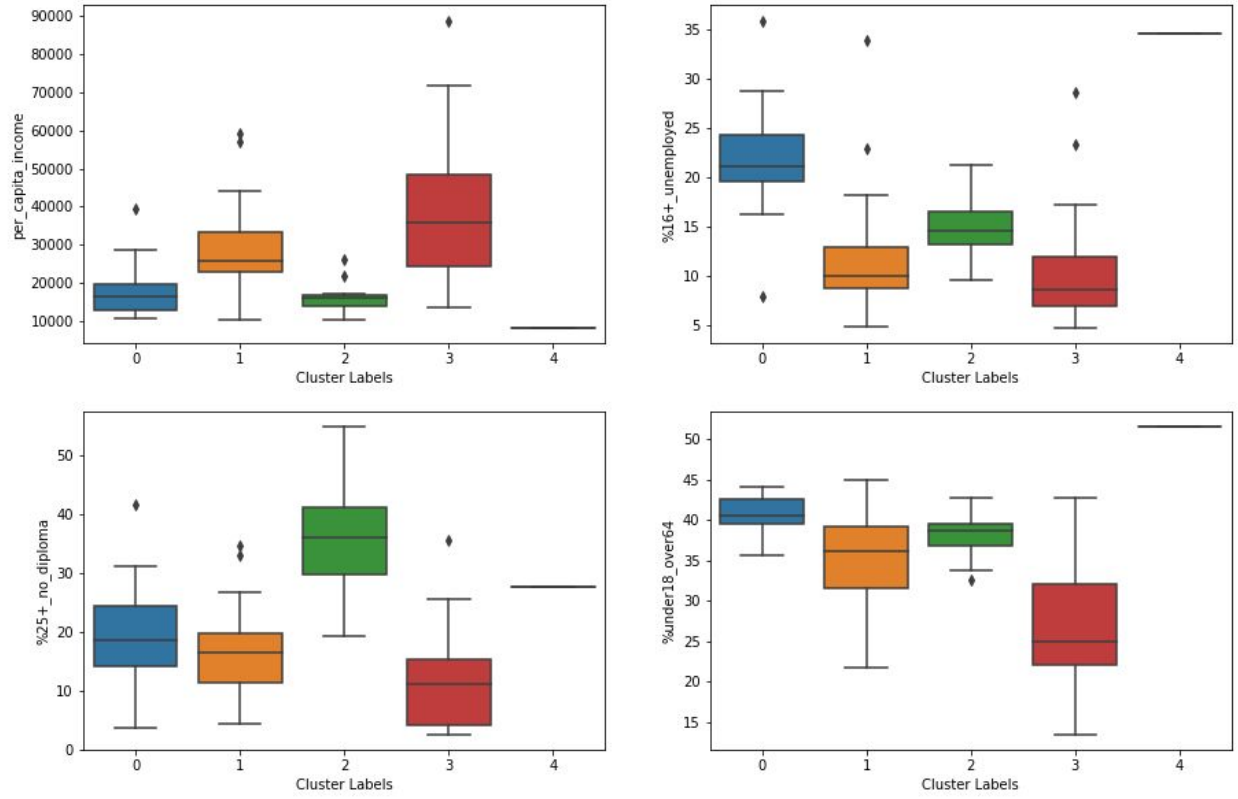
By comparison, communities in cluster 3 are well-indulged in coffee shops, middle-to-high-end Italian and Japanese restaurants, and hotels. Meanwhile, communities in Cluster 1 and 4 have more evenly distributed venues, no distinct patterns are to be noted yet.

2.4 Explore Community Resident Attributes By Cluster

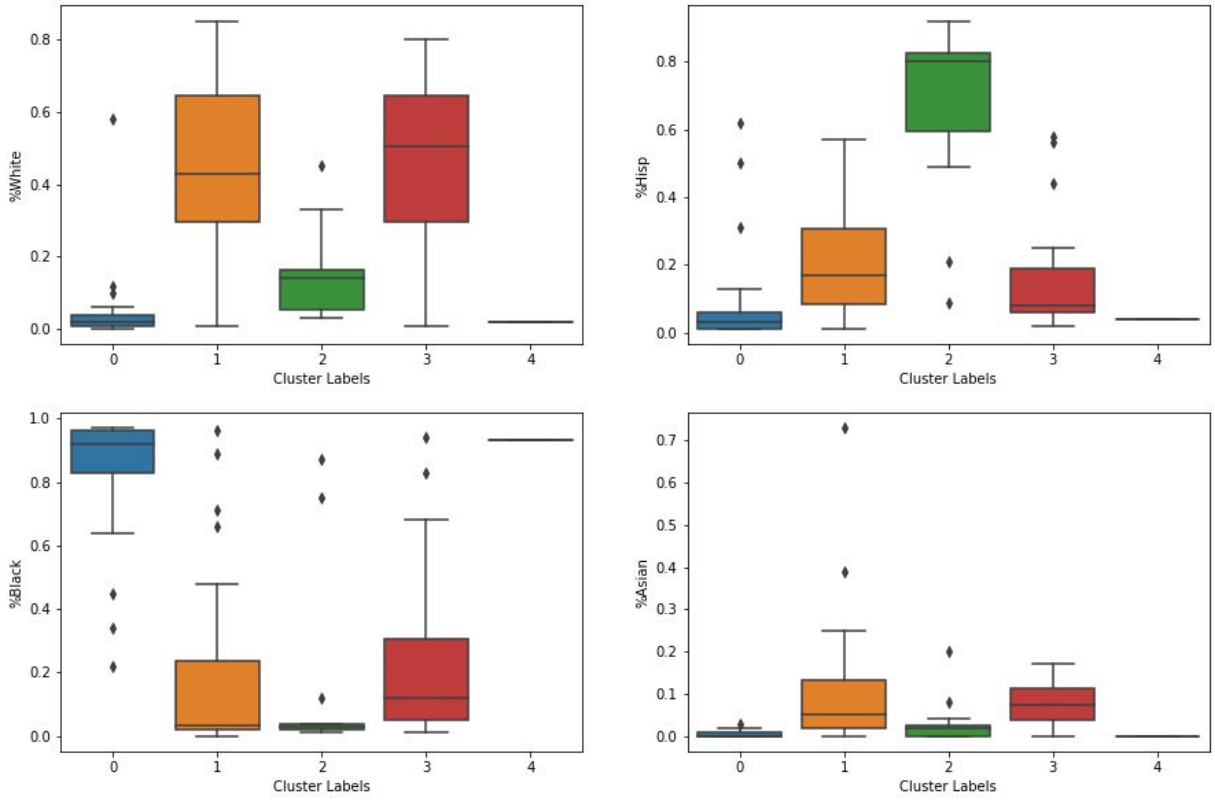
This outcome of clustering by common venue types is then examined jointly with data that denote resident attributes of each community. First, indicators for communities' overall financial situation, education status, employment rates, and age composition. The results are shown as boxplots in figure 2.7.

The first boxplot shows the communities' median per capita incomes in each cluster. Clearly, communities in cluster 0, 2, and 4 have lower-income dwellers. For cluster 1 and 3, while low-income communities exist, the average financial status of communities in these two clusters are high. Cluster 0 and 4 have distinctively high rates of unemployment; Cluster 2 is particularly high in low-education rates. Finally, as the last boxplot shows, while the first four plots have similar percentages of the youngs and olds, cluster 4 is particularly high in this part.

These results correspond to the findings from the previous section, which suggests that communities in clusters 1 and 3 have relatively high-priced, high-quality consumption spaces while communities in the remaining clusters have the opposite.



Next, I examine the residents' racial make-up of communities in each cluster. The result is shown as boxplots in figure 2.8. Albeit a few outliers, these graphs exhibit a clear racial segmentation between clusters. Communities with a high percentage of white residents, as presented in the first plot, are mostly incorporated in clusters 1 and 3, where high-end consumption spaces are prevalent. Clusters 0 and 4 are filled with black communities, some of which the percentage of black residents are over 95%; Cluster 2 shows the same trait for Hispanic residents. These three clusters of communities are the fast-food and discount store congregations. Cluster 2, as noted in the previous section, moreover has a large number of Mexican restaurants. While there's not an obvious trend in the distribution of Asian-centric communities, cluster 1 does include two outlying communities with high percentages of Asian residents. As cases in other U.S. cities suggest, Asian residents are smaller in total population and tend to reside collectively in one or two small, centralized spaces. These two communities are presumably such spaces for Chicago.



In short, the exploratory analysis tells that some discrepancies of resident characteristics exist between these venue-defined clusters of Chicago communities. Specifically, for communities that tend to have higher-end while higher-priced consumption options, residents are on average highly-educated, well-paid adults who are also mostly white. On the other hand, in communities overfilled with fast-food joints and thrift shops, a high number of residents -- who are often non-white -- receive lower education and income.

2.5 Logistic Regression Analysis

Is there a true relationship between public consumption spaces (i.e. venues) and resident economic/racial characteristics? Based on conclusions from the exploratory analysis above, I use a **multinomial logistic regression model** to examine if resident features examined in the previous section are truly related to the venue-generated clustering of communities. The test hypotheses are as follows:

H0: There is no true relation between venue-generated clusters and the resident features of chicago communities[2].

Ha: There is a true relation between the two.

Multinomial logistic regression is applicable to this dataset for two reasons. First, it allows the dependent variable categories to have more than two levels. In this case, Cluster labels are treated as the dependent variable, which has five levels since there are five clusters; resident features are treated as independent variables. Second, it does not assume normality, linearity, or homoscedasticity of the independent variables. In this case, the independent variables have relatively high correlations to one another. Income/social-status-based segregation in society is generally commensurate with racial segregation, therefore, the effects of social/financial status on the community clustering cannot be fully exempted from being effected by racial status, and vice versa.

[2] Here, resident features are the ones discussed in the exploratory analysis: median per-capita income; the percentage of residents who are under 18 or over 64 (i.e. not in the potential labor force); the percentage of residents who are over 16 and unemployed; the percentage of residents over 25 without a high school diploma (i.e. low education); the percentage of residents who are white, black, Hispanic, or Asian.

Because my goal is to examine all resident features as a package, I only analyze the outcome for the full model, in which all independent variables are taken into account.

3 Results

The summary statistics of the logistic regression model delivers a somewhat mixed message, but on the whole, the outcome favors reasonably the rejection of H0.

Table 3.1 shows the overall summary of the full model. The Pseudo-R-Square is computed based on the ratio of the maximized log-likelihood function for the null model "M0" and the full model "M1". Pseudo-R-Square values close to 0 means the model does not significantly improve the sureness of estimating dependent variables' categorization, while being close to 1 means the model fits perfectly and the log-odds ratio is maximized to 0. In this model, this value is at ~0.6, which indicates that the full model works slightly better than mediocre.

The LLR indicators assess the performance of the full model versus the null model (in which no indicators are taken into account). The LLR p-value, in particular, indicates the probability of observing the test statistic assuming the null hypothesis (H0) where the population coefficient is zero. In this model, the p-value is significantly low at 4.166e^-14, which suggests that we can reject the null hypothesis that the null model works better than the full model.

MNLogit Regression Results

Dep. Variable:	y	**No. Observations:**	77
Model:	MNLogit	**Df Residuals:**	41
Method:	MLE	**Df Model:**	32
Date:	Sat, 17 Aug 2019	**Pseudo R-squ.:**	0.6040
Time:	10:02:48	**Log-Likelihood:**	-43.274
converged:	False	**LL-Null:**	-109.28
		LLR p-value:	4.166e-14

While looking into the performance of each specific regressors, however, the statistics do not look so satisfactory. The z-values, the ratio of the coefficient estimate divided by the standard error of the estimator, are low; their corresponding p-values are not low enough.

Z-values and p-values indicate the amount of uncertainty around a point estimate. Larger z and smaller p usually means less uncertainty for the specific point estimation in the model. The p-values for single regressors (look at table A1 in the Appendix) are barely less than 0.1, let alone 0.005 where a rejection of the null hypotheses can be reasoned.

However, these estimates cannot fully reflect the "partial marginal effect" of each indicator on the dependent variables, but merely the effects of the probabilities around the dependent variable. Moreover, due to the non-linear structure in the logistic regression model, the effect of the one indicator is not immediately separable from the effect of the others. Therefore, the low score is not fully representative of the effects of individual regressors.

In short, the model results show that, collectively, all resident features analyzed in the full model have a significant relationship with venue dispositions in the Chicago communities. However, there is no sufficient proof that any of the resident features, on their own, are significant influencers to the venue types in the communities.

4 Discussion

The first half of the analysis is to apply a K-means Clustering method to the venue data of community areas in Chicago City. The performance of this model, which translates to the accuracy in categorizing communities and examining the venue types that represent them, depends greatly on the quality of data acquired. However, a drawback at this stage potentially undermines data quality. While obtaining venues from the FourSquare API, a uniform search radius (2.5km about the community center) was used for all Chicago communities. None of the Chicago communities are perfectly round-shaped and the sizes of them vary from over 7 km to less than 2 km. This may affect purity of venue data for smaller communities, for venues outside the communities but within a 2.5km search radius may be falsely put under their names.

Exploratory analysis gives an impression that strong links exist between the residents' social, economic, and cultural attributes and their respective consumption space types in Chicago community areas. The logistic regression results, while showing that consumption space types

are truly related to all residents' attributes as a whole, fails to detect its correlation with most of the individual independent variables (i.e. resident attributes) separately.

Such an outcome is likely due to **multicollinearity**, a statistical phenomenon in which predictor variables in a logistic regression model are highly correlated. The existence of collinearity inflates the variances of the parameter estimates which affects confidence intervals and consequently incorrect inferences about relationships between explanatory and response variables (Midi et. al 2010). Unavoidably, social, racial, and economic segregation are deeply intertwined phenomena, imposing great influence on one another. Therefore, using these variables indicating segregation in these three aspects puts the model under certain risk of multicollinearity, blocking the model to observe any significance of each indicator on the dependent variable.

Future studies may employ more rigorous approach, such as using ridge regression or Principal Component Analysis (CPA) to minimize the confounding influence of multicollinearity; or to break up the full model and compare the performance of different partial models: each time, only some of the resident features should be selected as indicators, then the LLR test can be used to compare the performances of different combinations.

5 Conclusion

Daily experience tells us all that different consumption spaces serve different groups in society, and this project provides a rough proof to this common sense.

The project performs a fairly comprehensive, though rudimentary, analysis on the relationship between public consumption options and resident social, economic, and cultural segmentation in Chicago communities. A sign of correlation is drawn between the two, though it is not perfectly attested by the following logistic regression analysis.

The analysis finds that certain food and activity consumption options are more prevalent in low-income and racial minority resident communities, such as fast-food restaurants, bars and clubs. Higher-income and mostly white community groups, on the other hand, are presented with options to spend on fine restaurants and finer entertainment like hotels and museums. Future

studies may conduct more rigorous investigation into the underlying reasons of why different types of venues open in different communities. Research can also inspect more deeply into the effects of such consumption option divergence onto aspects of Chicago citizens' lives, such as their physical, mental, or social well-being.

6 Appendix

Table A1: Partial marginal effect summary, by levels of dependent variable

MNLogit Regression Results						
y=1	coef	std err	z	P>|z|	[0.025	0.975]
const	75.5586	65.971	1.145	0.252	-53.743	204.860
%White	-66.9461	65.890	-1.016	0.310	-196.088	62.195
%Hisp	-74.6018	65.187	-1.144	0.252	-202.366	53.162
%Black	-74.8109	65.824	-1.137	0.256	-203.823	54.201
%Asian	-35.6904	77.250	-0.462	0.644	-187.097	115.716
per_capita_income	-0.0001	0.000	-1.105	0.269	0.000	9.81E-05
%25+_no_diploma	-0.0091	0.149	-0.061	0.952	-0.301	0.283
%16+_unemployed	-0.0865	0.159	-0.545	0.586	-0.398	0.225
%under18_over64	0.0038	0.229	0.017	0.987	-0.445	0.453
----------						
y=2	coef	std err	z	P>|z|	[0.025	0.975]
const	0.2961	89.483	0.003	0.997	-175.088	175.680
%White	-9.1284	92.975	-0.098	0.922	-191.357	173.100
%Hisp	-10.9261	91.190	-0.120	0.905	-189.655	167.803
%Black	-11.8628	92.281	-0.129	0.898	-192.730	169.005
%Asian	19.4063	100.063	0.194	0.846	-176.715	215.527
per_capita_income	1.939E-05	0.000	0.122	0.903	0.000	0.000
%25+_no_diploma	0.2314	0.195	1.184	0.236	-0.152	0.614
%16+_unemployed	-0.2360	0.253	-0.932	0.351	-0.732	0.260
%under18_over64	0.1985	0.300	0.661	0.508	-0.390	0.787
----------						
y=3	coef	std err	z	P>|z|	[0.025	0.975]
const	169.7944	76.228	2.227	0.026	20.391	319.198
%White	-155.8278	76.112	-2.047	0.041	-305.004	-6.652
%Hisp	-157.6310	74.173	-2.125	0.034	-303.008	-12.254
%Black	-159.4006	75.314	-2.116	0.034	-307.014	-11.787

%Asian	-142.2882	86.615	-1.643	0.100	-312.051	27.474
per_capita_income	-9.548E-05	0.000	-0.852	0.394	0.000	0.000
%25+_no_diploma	-0.0052	0.182	-0.029	0.977	-0.362	0.351
%16+_unemployed	0.0905	0.183	0.495	0.620	-0.268	0.449
%under18_over64	-0.3976	0.257	-1.547	0.122	-0.901	0.106
----------						
y=4	coef	std err	z	P>|z|	[0.025	0.975]
const	-96.0030	1.5E+08	-6.38E-07	1.000	-2.95E+08	2.95E+08
%White	-128.6503	3.8E+08	-3.39E-07	1.000	-7.44E+08	7.44E+08
%Hisp	-168.7042	4.13E+08	-4.09E-07	1.000	-8.09E+08	8.09E+08
%Black	-161.2817	3.64E+08	-4.43E-07	1.000	-7.13E+08	7.13E+08
%Asian	-142.0406	1.3E+08	-1.09E-06	1.000	-2.55E+08	2.55E+08
per_capita_income	-0.0033	7324.623	-4.47E-07	1.000	-1.44E+04	1.44E+04
%25+_no_diploma	0.7946	1.2E+06	6.6E-07	1.000	-2.36E+06	2.36E+06
%16+_unemployed	-1.9802	6.28E+05	-3.16E-06	1.000	-1.23E+06	1.23E+06
%under18_over64	6.8429	1.99E+06	3.45E-06	1.000	-3.89E+06	3.89E+06

7 Bibliography

Bolzoni, Magda. 2016. “Spaces of distinction, spaces of segregation -- Nightlife and consumption in a central neighbourhood of Turin”.

Midi, Habshah and Sarkar, S.K. and Rana Sohel. 2010. “Collinearity diagnostics of binary logistic regression model”. *Journal of Interdisciplinary Mathematics*, 13:3, 253-267, DOI: 10.1080/09720502.2010.10700699

United States Census Bureau. 2017. *2013-2017 American Community Survey 5-Year Estimates*. https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=CF