JB, Ted Stetzel
Multivariate Final Project
Brooklyn Housing Analysis

1. Objective of Analysis

The diversity of New York City's residential buildings and housing programs provides many opportunities for low- and moderate-income households to live in the city even with the overall high cost for housing. Despite that variety and a multitude of intergovernmental interventions that include direct and indirect housing subsidies, tax incentives for new construction and redevelopment of existing housing stock, and rent increase regulations for multi-family buildings, housing affordability remains elusive for many New Yorkers. Supportive housing programs are controversial not only because of the underlying ideologies supporting or opposing safety nets, but also for the very way affordability and housing burden are measured. Affordability should take into context people and places, looking at scales of density, real estate costs, household income, and housing typology.

Multivariate statistics can provide meaningful measures of those differences among the separate and diverse communities within the city so that appropriate policies may be developed to preserve important attributes of neighborhoods and help New Yorkers maintain a foothold in their homes. This analysis explores the characteristics of Brooklyn households using a range of statistical methods. The primary objective is applying methods of multivariate statistics to better quantify and classify the similarities and differences of household characteristics across Brooklyn's 18 Community Districts. More effective policies addressing affordability may be developed from the accurate analysis of its many determinants and the way in which affordability varies across neighborhoods and households.

Affordability is most commonly measured by the housing cost burden borne by households, evaluating household income and total housing costs. A share of no more than 30 percent of income is the most widely used measure of affordability and provides a test typically applied in mortgage lending for assessing the credit worthiness of borrowers. Households that spend more than 30 percent of income on housing are considered rent burdened. This standard is derided for being infeasible in many markets where the cost of a home has outpaced wages and where they are high to start. According to *Housing New York: A Five-Borough, Ten-Year Plan*, approximately 55 percent of all New York renter households were considered rent-burdened in 2012, representing an increase of more than 11 percent since 2000[1].

The following analysis begins with a MANOVA test of the variance within and across Brooklyn's community districts. The MANOVA test is helpful to perform as a preliminary test to determine if the housing characteristics chosen for analysis exhibit differences among Brooklyn's neighborhoods beyond chance. A similar approach to MANOVA called discriminant analysis is used in the second method. Discriminant analysis uses linear functions of variables to determine the difference between groups. In other words, it looks for the relative contribution of each variable to explain group separation and to help illustrate configuration of those groups. The study then explores classification analysis, a close cousin of discriminant analysis used to assign new observations to pre-determined groups. The study concludes with cluster analysis in attempting to create taxonomy of Brooklyn's neighborhoods based on household data.

Patterns that emerge in this analysis are not exhaustive but provide a formidable foundation for additional research into the topic. These small PUMA geographies allow us to compare and contrast

[1] http://www.nyc.gov/html/housing/assets/downloads/pdf/housing_plan.pdf, page 17.

housing characteristics throughout Brooklyn's rapidly changing neighborhoods and would enhance qualitative methods which consider community desires for new construction and housing redevelopment.

2. Data description

Given the contrast among Brooklyn's neighborhoods, the variety of its commerce, and the mosaic of its built environments, uniform programs may be less potent in addressing the affordable housing shortage. Reliable data are essential to developing accurate analysis and good housing policy. A paucity of non-proprietary data capturing information on both owner-occupied and rental units linked to demographic data complicates research.

The most widely used data available for demographic and housing analysis are captured in the American Community Survey (ACS). The ACS replaced the long-form survey administered with the Decennial Census in 2000, and approximately 1 in 38 households now participate in the ACS every year. A subset of survey results are stripped of personal identifiers and made public, packaged as the Public Use Microdataset Sample (PUMS) in statistical geographies known as Public Use Microdataset Areas (PUMAs). These Census geographies approximate New York City's Community Districts, though, the boundaries are not coterminous. Data for this study is focused on PUMAs comprising the Borough of Brooklyn (Kings County) using annual data from 2005 to 2013. Incomplete responses to the survey were removed, providing 21,620 records.

The variables used throughout the study are defined as:

- CD = Community District (derived from PUMA code)
- HPER represents the dependent variable percentage of household income spent on housing for renters and owners. This is our affordability metric – a lower percentage indicates greater affordability and a higher percentage indicates a greater housing cost burden.
- NP = Number of persons per household
- Pop = Population of area
- WGTP = Number of Housing Units
- BDS = Bedrooms
- RMS = Rooms in housing unit
- Tenure = Designates the occupant as a renter or an owner of the unit
- Time = Time lived in unit

3. Variance Analysis (MANOVA)

3.1 Introduction

The first step in the analysis is to determine if the means of each of our independent variables are different among Brooklyn's community districts. This model is described as a one-way Multivariate Analysis of Variance (MANOVA), which compares the "between" sample covariance matrix to the "within" sample covariance matrix across a single set of groups[2]. Like ANOVA, MANOVA examines the difference of means, but rather than examining differences of two or more groups, MANOVA calculates

[2] Yu, Yue, Chapter 6b

the difference among multiple vectors of means. MANOVA tests have limitations that should be considered, primarily sensitivity to outliers and multicolinearity in the dependent variables.

The multiple vectors for this study are those associated with the independent variables of each group and their error matrix (**E)** and hypothesized effects matrix (**H**), which are defined as:

$\mathbf{H} = \mathrm{n}\sum_{\mathrm{i}=1}^{\mathrm{k}}(\bar{\mathrm{y}}_{\mathrm{i}} - \bar{\mathrm{y}}\ldots)(\bar{\mathrm{y}}_{\mathrm{ij}} - \bar{\mathrm{y}}\ldots)'$; and $\mathbf{E} = \sum_{\mathrm{i}=1}^{\mathrm{k}}\sum_{\mathrm{j}=1}^{\mathrm{n}}(\bar{\mathrm{y}}_{\mathrm{i}} - \bar{\mathrm{y}}\ldots)\left(\bar{\mathrm{y}}_{\mathrm{ij}} - \bar{\mathrm{y}}\ldots\right)'$

Where y_{ij} is associated with the *j*th observation of the *i*th population within the *k*th group.

This section will examine the independent variables of

- PL = Number of people living at residence
- HINCP = Household income
- yearsOwned = Years that the resident has been living at this residence
- HPER = Estimated housing burden
- Pop = Population of the Community District

3.2 Analysis of Variance

The MANOVA test compares the equality of the means among the matrices. The null hypotheses can be stated as $H_0 = \mu_1 = \mu_2 = \ldots \mu_k$ versus the alternative statement (H_a) that at least one inequality exists. Four tests should be run to determine if the difference is real, meaning beyond chance.

The **E**, **H,** and **E Inverse * H** produced by SAS are provided in Appendix A of the report, as well as the eigenvalues used in each MANOVA test described below.

Wilks' Λ test statistic:

$\Lambda = |E| / |E+H|$, rejecting H_0 if $\Lambda \leq \Lambda\alpha(p,vH,vE) = .882 \leq .924$ (5,12,1000)(VH=17 approximated at 12; vE= 389,142 approximated at 1,000), so we reject H_0 in Wilks' test.

n = number of observations; p = number of variables; vH = k−1: degrees of freedom for hypothesis; vE = k(n−1): degrees of freedom for error.

Roy's largest root:

$\theta = \lambda 1 / 1 + \lambda 1$ (.105/1.105 = .095), rejecting if H_0 if $\theta > \theta\ \theta\alpha(s,m,N) = .095 > .082\ \alpha(5,5,240)$, so we reject H_0 for Roy's test.

s = min(vH,p); m = ½*(|vH −p|−1); N = ½*(vE −p−1)

Pillai statistic:

V (s) = λ1/(1+λ1), rejecting H_0 if V (s) ≥ V (s) α (s,m,N) = .121 < 2.271 (5,5,10). So we fail to reject H_0 in the case of Pillai.

Lawley-Hotelling (Hotelling's generalized T2) statistic:

U(s) = λ1 = θ/1- θ, rejecting H_0 for large values of $(vE/vH)*U^{(s)}$ = (389,142/17)*(.131) = 2998.68. Because p<VH, we test (p, VH, VE), use (5,7,∞) instead of (5,17,389142) = 6.6902 < 2998.68, so we reject H_0 in Lawley-Hotelling's test.

vH = 1(k = 2), s = 1 and $T2 = (n1 + n2-2)U^{(1)}$

3.3 Variance Analysis Conclusion

Using the equation: $\frac{\lambda 1}{\sum \lambda} = \frac{.1048}{0.1048+ 0.0199+ 0.0042+ 0.0017+ 0.0003}$ = .80, we can conclude that the essential dimensionality of the space of the mean vectors is equal to 1 or is collinear. This means that the relative powers of the tests are: $\theta \geq U \geq \Lambda \geq V^s$. So although we failed to reject the null hypotheses in the Pillai test, Pillai is the weakest test when there is collinearity. Based on this determination, we can conclude that the household characteristics are different across Brooklyn's neighborhoods.

4. Discriminant Analysis

4.1 Introduction

In order to better understand how the burden of housing costs affect the residents of Brooklyn in different community districts, a discriminant analysis of the Brooklyn Housing Data was conducted across the 18 community districts across in the borough. The goal of this analysis is to better understand how the selected variables differ across different neighborhoods, especially those that are in consideration for re-zoning as part of Mayor Bill de Blasio's affordable housing plan.

Knowing what separates groups is useful for determining how different they are (in terms of relative distance from in each other on a multi-dimensional plane) and which variables contribute most to this separation. The method is well suited for continuous variables for data sets where the number of observations is greater than the number of variables. In order to differentiate between the 18 communality districts in Brooklyn we will evaluate 20,000, where each group has just over 1000 records the using the variables

- PL = Number of people living at residence
- HINCP = Household income
- yearsOwned = Years that the resident has been living at this residence
- HPER = Estimated housing burden
- Pop = Neighborhood population

First, the discriminant functions for each variable are determined. To do so, each of our group's observations are transformed into the vector $\mathbf{y}_{ij}$ (where i = the group (CD) and j = is the number of the observation). Each of the $\mathbf{y}_{ij}$ vectors is then transformed to obtain the discriminant function $\bar{z}_i$ which is expressed by the equation:

$$\bar{z}_i = a'\bar{y}_i$$

Where

$$\bar{y}_i = \sum_{j=1}^{ni} y_{ij}/n_i$$

$\bar{z}_i$ is important because it represents a linear combination that shows the differences between the variables across observations. These values will be used to conduct discriminant analysis, which identifies the vectors that maximize the distance between these z values.

Once these are determined, the **E** and **H** (just as in MANOVA analysis in section 3) are evaluated to the relationship between error within each group compared to error as a whole. These resulting proportions are called eigenvalues and are found using the following formula:

$$\lambda = \frac{SSH(z)}{SSE(z)}$$

These values are used to evaluate the relative importance of each z_i

4.2. Eigenvectors of $\boldsymbol{E^{-1}H}$

Using the CANDISC function, SAS provides the following Eigenvalues $\lambda_1 = 0.1048$, $\lambda_2 = 0.0199$, $\lambda_3 = 0.0042$, $\lambda_4 = 0.0017$, $\lambda_5 = 0.0003$

While none the eigenvalues are greater than 1, the first value (λ_1) seems to be the most significant as it accounts for a significant proportion of the total. The proportion for λ_1 is determined by dividing λ1 by the sum of the eigenvalues as displayed below:

$$\frac{\lambda 1}{\sum \lambda} = \frac{.1048}{0.1048 + 0.0199 + 0.0042 + 0.0017 + 0.0003} = .80$$

Since the remaining 4 eigenvalues only account for about 19% combined of the total proportion, we can assume that the mean vectors line in a single dimension. This means that only one discriminant function will be needed to describe the separation between the community groups and the rest of the eigenvalues can be ignored.

Each of these Eigenvalues has a corresponding Eigenvector, which explains the composition of the eigenvalue on a per variable basis expressed as **a**. Knowing that only a single function will be needed to

describe 80% of the difference between groups, we can then turn our attention to the first eigenvector, $\mathbf{a_1}$ as shown below:

$$\boldsymbol{a_1} = \boldsymbol{S}_{pl}^{-1}(\bar{y}_1 - \bar{y}_2) = (-0.3316\ ,0.00001, .0961, -0.0011, -0.0004\ )'$$

An inspection of the values in the eigenvector shows that 3 are effectively 0 and contribute nothing to the model. This is because the variables used for this analysis are not measured on the same scale, with comparable variances, and the data needs to be standardized. To standardize data the $\mathbf{a_1}$ vector is multiplied by the square root of each diagonal of the pooled covariance matrix as shown below:

$$\boldsymbol{a_1^*} = \sqrt{diag(\boldsymbol{S}_{pl})\boldsymbol{a_1}} = (\text{-}0.539, 0.918, 0.078, \text{-}0.033, \text{-}0.114)'$$

The resulting vector can then be evaluated for to see which vectors values have the highest absolute value. The variables that these correspond to have the highest influence on the separation of the groups. From this output we can see that the second variable (HINCP) has the most influence with .918 and has NP notable influence with -.539.

4.3 Test of Significance

Finally the discriminant functions can be tested to determine if they describe a significant difference between groups or if these difference can be explained by random variance. Our test will focus on the Wilks' Lambda value of 0.8819 provided in the Multivariate Statistics and F Approximations table in Appendix A.

First $\mathbf{a_1}$ will be tested using the Hypothesis: H_0: $\mathbf{a_1} = 0$, meaning that the difference is not significant. If this hypothesis is rejected then each additional a vector will be tested until will be tested until the null hypothesis cannot be rejected.

After conducting the hypothesis test we find that
We reject H_0: $\mathbf{a_1} = 0$ since $\boldsymbol{\Lambda_1} \leq \boldsymbol{\Lambda}_{\boldsymbol{\alpha}(5,17,1000)} = .882 < .964$
We fail reject H_0: $\mathbf{a_2} = 0$ since $\boldsymbol{\Lambda_2} > \boldsymbol{\Lambda}_{\boldsymbol{\alpha}(5,17,1000)} = .9744 > .937$
(*full work is included in Appendix 3*):

So we conclude that $\mathbf{a_1}$ is significant and describes separation between groups and all sequential **a** vectors do not significantly contribute to separation.

4.4 Visual Examination of Discriminant Functions: 2D Scatter plot

A visual inspection of a two dimensional scatter plot can be performed by plotting the results of the first and second linear function on a 2D plane. Although more than two dimensions are required to explain all of the variance, an overwhelming majority of difference between groups is explained in the first two variables, (especially the first) so this method is effective for visually determining separation.

Displaying a scatter plot of all +20k observations across 18 groups would not likely be of any use so a subset of 150 samples across groups 2, 5, and 9 was randomly selected and a scatter of the first (x) and second (y) discriminant functions was created and can be viewed in Appendix 1 Part 4. Although separation is not particularly strong across all three groups, a high concentration of group 2 is clustered in the upper left and group 9 is concentrated lower on the y-axis than group 5.

4.5 Discriminant Analysis Conclusion

The differences between the 18 groups can be mostly determined using a single discriminant function. The differences are statistically significant and are driven predominantly by household income and inversely by the number of people per household.

5. Classification Analysis

5.1 Introduction

If the Community District were unknown or if a separate study where conducted were a single observation needed to be assigned to a group such as meeting the criteria of being rent burdened, classification analysis is a process similar to discriminant analysis that could prove very useful. Instead of analyzing variables to describe their contributions to the differences in pre-determined groups, classification analysis uses a pre-established model to predict which groups new observations will likely belong to. In this section the model that was built using the discriminant will be used to predict group membership. Depending on the covariance matrices for the data set, classification can be linear (when covariance is the same) or quadratic (unequal proportions between variables).

5.2. Linear Classification Functions

In order to classify each observation using liner classification each variable is calculated using a linear for each corresponding group. The function with the highest value is the predicted group. Below is a subset of the linear functions for linear classification of this data set:

$$\boldsymbol{L_1}(x_0) = 0.82666(NP_1) + 0.0000306(HINCP_1) + 13.55983(Time_1) + 0.0826(HPER_1) - 0.000484(Pop_1) - 64.11$$

$$\boldsymbol{L_2}(x_0) = 0.81419(NP_1) + 0.0000359(HINCP_1) + 13.65582(Time_1) + 0.08284(HPER_1) - 0.000116(Pop_1) - 765.1$$

…

$$\boldsymbol{L_{17}}(x_0) = 0.77044(NP_1) + 0.000031(HINCP_1) + 13.6235(Time_1) + 0.08361(HPER_1) - 0.000435(Pop_1) - 64.62$$

$$\boldsymbol{L_{18}}(x_0) = 0.81258(NP_1) + 0.0000296(HINCP_1) + 13.61568(Time_1) + 0.08295(HPER_1) - 0.000871(Pop_1) - 64.67$$

To determine if classification analysis is useful, we'll use the variables from section 4 to create a linear classification chart.

A classification matix was created using SAS and can be viewed in Appendix 1 Part 5.The error rate is determined by adding the non-diagonal values and dividing by the total number of observations:

$$\frac{8222}{9381} = .876$$

Given that there is a 10% chance that someone would randomly pick the correct group, a 12.4% correct prediction rate is not of much use. Additionally this same matrix was created using a quadratic classification function and resulted in slightly less accurate prediction rate.

5.4 Classification Analysis Conclusion

For the purposes of this data set, neither linear nor quadratic classifications are helpful given the large number of groups and inadequate separation to accurately predict groups based on the variables provided.

5. Cluster Analysis

5.1 Introduction and data prep

While discriminant analysis is a useful tool to determine the differences between groups, additional insight can also be gained from evaluating the similarities between groups. In order to better understand the relationship between the different community groups, a cluster analysis of the data was also conducted. Cluster analysis is useful to assess geographic context, helping us to determine if neighborhoods that are contiguous also exhibit similarities in household characteristics.

In order to evaluate community districts using cluster analysis a second dataset was created using pivot tables in Microsoft Excel. The new data set contains a row for each community district and the columns hold the average value for each of the variables in the original data set on a per group bases. Another column was added named 'percentOwned', which is the proportion of all those surveyed that own their apartment or house instead of renting.

5.2 Hierarchical Clustering

The simplest and most efficient form of clustering is hierarchical clustering which uses the algorithm below to group (i.e. cluster) each observation based on its closest neighbor.

$$N(n,g) = \frac{1}{g!}\sum_{k=1}^{g}\left(\binom{g}{k}(-1)^{g-k}k^n\right)$$

This recursive algorithm begins by grouping the two closes observations to form a cluster. The distance of each group is then measured from its center in comparison to other observations and other clusters. All observations and resulting clusters of observations are merged until there is only one cluster left which contains all observations.

Since a single cluster provides no information about differences between clusters, a cutoff number of cluster (*g*) needs to be determined. A popular way of determining a cut off is to use the Mojena method which utilizes the formula:

$\propto_j > \overline{\propto} + ks_{\propto}$

$\overline{\propto}$ is the mean of all distances
$\propto_j$ is the value at which the cut off will be conducted. All clusters prior to this distance will remain clustered
$s_{\propto}$ is the standard deviation of all distances = $\sqrt{\frac{\sum(\propto_j - \overline{\propto})^2}{n}}$

Where k is a best practice constant. 1.25 will be used here based on Milligan and Cooper (1985)

After standardizing the data setting mean =0 and std =1 a cluster analysis using SAS (The Cluster History chart included in Appendix A was produced in SAS) the following number of clusters was established using the Mojena method:

$\overline{\propto} + ks_{\propto}$ = 1.61 +(1.25*.4) =2.11

The Mojena method would make the cutoff point here at 3 clusters. However we'll choose to round up one additional cluster due to most observations being in a single cluster. Interestingly even after rounding up another cluster, there is still one observation that has not been classified (District 4). This means that this district (Bushwick) is very different than the rest of the group. Additionally districts 2 and 6 have been classified as a single cluster indicating that Downtown Brooklyn and Park Slope are also very different than the rest of the districts.

5.4 Cluster Analysis Conclusion

An examination of the cluster analysis chart (Appendix 1 part 6) supports the hypotheses that districts closer to Manhattan would be very different than the other districts. Additionally the chart shows some slight differences between the northern and southern regions of Brooklyn. CL12 (highlighted in blue) in the dendrograph in the appendix) is composed of the neighborhoods north east of Prospect Park.

Districts 2, 4 and 6 seem to be outliers in terms of the groups. This is likely due to their close proximity to Manhattan which has likely increased the average income household income for residents of these neighborhoods. Surprisingly, CD 1 (Williamsburg/Greenpoint) was not specified as an outlier given its close proximity to Manhattan and the other outlier CDs were not grouped in the primary cluster.

6. Conclusion

Overall there seems to be a large difference in average household income across the 18 community districts. The neighborhoods closer to Manhattan are outliers in this regard with the rest of Brooklyn much closer in terms of this variable. While the burden associated with paying rent or a mortgage slightly increases as income decreases, it does not vary as drastically between neighborhoods as income does. This could be because the number of people per household tends to increase as income decreases. These characteristics should be taken into account when re-zoning areas of Brooklyn to foster economic growth. Instead of trying to copy the economic growth of neighborhoods such as Downtown Brooklyn and Park Slope, which are comprised of high income households with fewer individuals, growth initiatives should be designed to benefit larger, lower income households which already bear a slightly higher housing cost burden.

Appendix 1: SAS Code & Output

Import from excel

```
PROC
IMPORT OUT = Nyc.CDFull DATAFILE="C:\Users\ts149267\Desktop\BK05_13_full.xlsx"
DBMS=xlsx REPLACE;
SHEET = "BK05_13";
GETNAMES=YES;
RUN;

PROC
IMPORT OUT = Nyc.CDAVG DATAFILE="C:\Users\ts149267\Desktop\BDRMS_CDs.xlsx"
DBMS=xlsx REPLACE;
SHEET = " BDRMS_CD";
GETNAMES=YES;
RUN;


PROC
IMPORT OUT = Nyc.CDsamp4
DATAFILE="C:\Users\Midwest\Dropbox\Ted_school_stuff\STA_9705_MV\project\data\BK05_13_full.xls
x"
DBMS=xlsx REPLACE;
SHEET = "samp";
GETNAMES=YES;
RUN;
```

***Part 3: MANOVA;**

```
PROC
IMPORT OUT = CDFull
DATAFILE="/folders/myfolders/Inputs/Project/BK05_13_full.xlsx"
DBMS=xlsx REPLACE;
SHEET = "BK05_13";
GETNAMES=YES;
RUN;

**Set up MANOVA Test on the full dataset to produce the E and H
tables;

Title 'MANOVA test on BK CD Housing Data';
PROC GLM;
CLASS CD;
MODEL NP HINCP yearsOwned HPER POP = CD;
MANOVA H=CD/PRINTE PRINTH;
RUN;
```

The GLM Procedure

Multivariate Analysis of Variance

E = Error SSCP Matrix					
	NP	HINCP	yearsOwned	HPER	Pop
NP	57096.649881	548761043.4	26.710465725	-84552.48807	6814089.6648
HINCP	548761043.4	9.9466677E13	-57428294.66	-19001559552	45241309650
yearsOwned	26.710465725	-57428294.66	14385.150073	1653.92234	74838.011168
HPER	-84552.48807	-19001559552	1653.92234	18000868.923	-5779258.884
Pop	6814089.6648	45241309650	74838.011168	-5779258.884	1957666476.2

H = Type III SSCP Matrix for CD					
	NP	HINCP	yearsOwned	HPER	Pop
NP	1715.6685279	-58746650.2	-33.39954991	17701.466056	210111.4794
HINCP	-58746650.2	7.0958762E12	3314002.7201	-1364700646	-11981325494
yearsOwned	-33.39954991	3314002.7201	6.6506203459	-687.7178072	-5391.376154
HPER	17701.466056	-1364700646	-687.7178072	328433.69288	2962424.3259
Pop	210111.4794	-11981325494	-5391.376154	2962424.3259	35040716.325

Characteristic Roots and Vectors of: E Inverse * H, where H = Type III SSCP Matrix for CD E = Error SSCP Matrix						
Characteristic Root	Percent	Characteristic Vector V'EV=1				
		NP	HINCP	yearsOwned	HPER	Pop
0.10483077	**80.04**	-0.00225631	0.00000009	0.00065386	-0.00000766	-0.00000257
0.01990347	**15.20**	0.00429575	0.00000005	0.00028858	0.00008815	-0.00001045
0.00417976	**3.19**	-0.00276055	0.00000004	0.00039159	0.00011215	0.00002599
0.00172618	**1.32**	0.00050003	-0.00000002	0.00047344	-0.00022167	0.00000913
0.00033949	**0.26**	0.00012794	-0.00000000	0.00829725	0.00001043	-0.00000162

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall CD Effect H = Type III SSCP Matrix for CD E = Error SSCP Matrix S=5 M=5.5 N=10798					
Statistic	**Value**	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Wilks' Lambda	0.88193641	32.34	85	104412	<.0001
Pillai's Trace	0.12062400	31.41	85	108010	<.0001
Hotelling-Lawley Trace	0.13097967	33.28	85	81675	<.0001
Roy's Greatest Root	0.10483077	133.21	17	21602	<.0001

***Part4: Discriminant Analysis;**

```
PROC CANDISC DATA=nyc.cdfull OUT=CAND;
CLASS CD;
VAR NP HINCP yearsOwned HPER Pop;
run;
```

	Eigenvalue	Difference	Proportion	Cumulative
1	0.1048	0.0849	0.8004	0.8004
2	0.0199	0.0157	0.152	0.9523
3	0.0042	0.0025	0.0319	0.9842
4	0.0017	0.0014	0.0132	0.9974
5	0.0003		0.0026	1

Raw Canonical Coefficients	
Variable	**Can1**
NP	-0.331623795
HINCP	0.000013531
yearsOwned	0.096102425
HPER	-0.001125870
Pop	-0.000378265

Pooled Within-Class Standardized Canonical Coefficients					
Variable	**Can1**	**Can2**	**Can3**	**Can4**	**Can5**
NP	-0.539143058	1.026464767	-0.659631332	0.119482291	0.030571864
HINCP	0.918174356	0.487976122	0.433815117	-0.226031091	-0.035845801
yearsOwned	0.078423201	0.034611163	0.046966630	0.056784048	0.995156120
HPER	-0.032500337	0.374004818	0.475838939	-0.940475114	0.044261270
Pop	-0.113872301	-0.462424298	1.149973250	0.403872167	-0.071672617

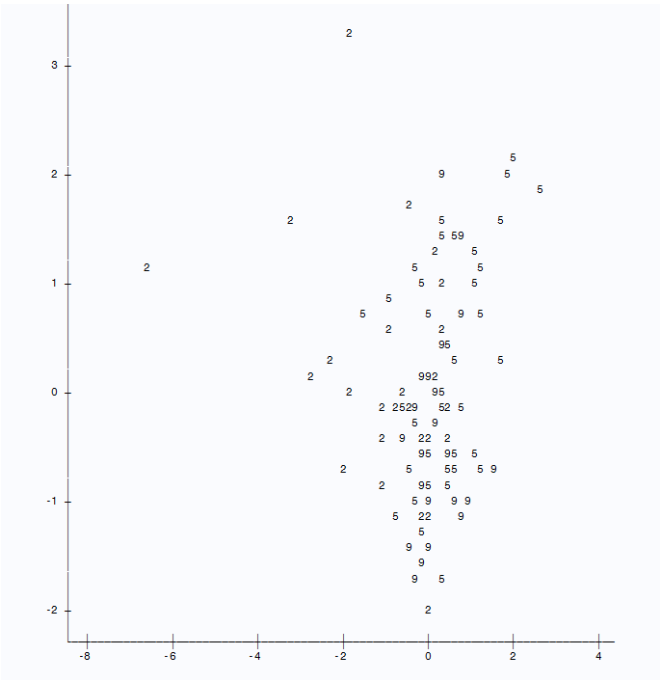
Multivariate Statistics and F Approximations					
Statistic	**Value**	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Wilks' Lambda	0.88193641	32.34	85	104412	<.0001
Pillai's Trace	0.12062400	31.41	85	108010	<.0001
Hotelling-Lawley Trace	0.13097967	33.28	85	81675	<.0001
Roy's Greatest Root	0.10483077	133.21	17	21602	<.0001

Part4: Discriminant Scatter chart;

```
PROC CANDISC data=nyc.cdsamp5 OUT=CAND;
  CLASS CD ;
  VAR NP HINCP yearsOwned HPER Pop;
run;
PROC PRINT DATA=CAND;
RUN;
PROC PLOT DATA=CAND;
  PLOT CAN2*CAN1=CD;
RUN;
```

***Part 5: Classification;**

```
PROC DISCRIM data=nyc.Cdfull LIST
CROSSVALIDATE;
   CLASS CD;
VAR NP HINCP yearsOwned HPER Pop;
RUN;
```



```
PROC CANDISC DATA=nyc.cdsamp4 OUT=CAND;
CLASS CD;
VAR NP HINCP yearsOwned HPER Pop;
run;
PROC CANDISC OUT=CAND;
  CLASS CD;
run;
PROC PRINT DATA=CAND;
RUN;
```

Subset of 18 variables:

Variable	1	10	11	12	13	14	15	16
Constant	-64.113	-65.101	-64.753	-65.444	-63.718	-64.876	-64.894	-64.789
NP	0.82666	0.81419	0.94371	1.20445	0.72543	0.8296	0.85925	0.91605
HINCP	3.1E-05	3.6E-05	3.2E-05	3.1E-05	2.9E-05	3.3E-05	3.3E-05	2.7E-05
years	13.5598	13.6558	13.596	13.5908	13.5791	13.618	13.6391	13.6258
HPER	0.0826	0.08284	0.08531	0.08903	0.07645	0.08505	0.08141	0.08306
Pop	-0.0005	-0.0012	-0.001	-0.0014	-0.0004	-0.0004	-0.0009	-0.0005

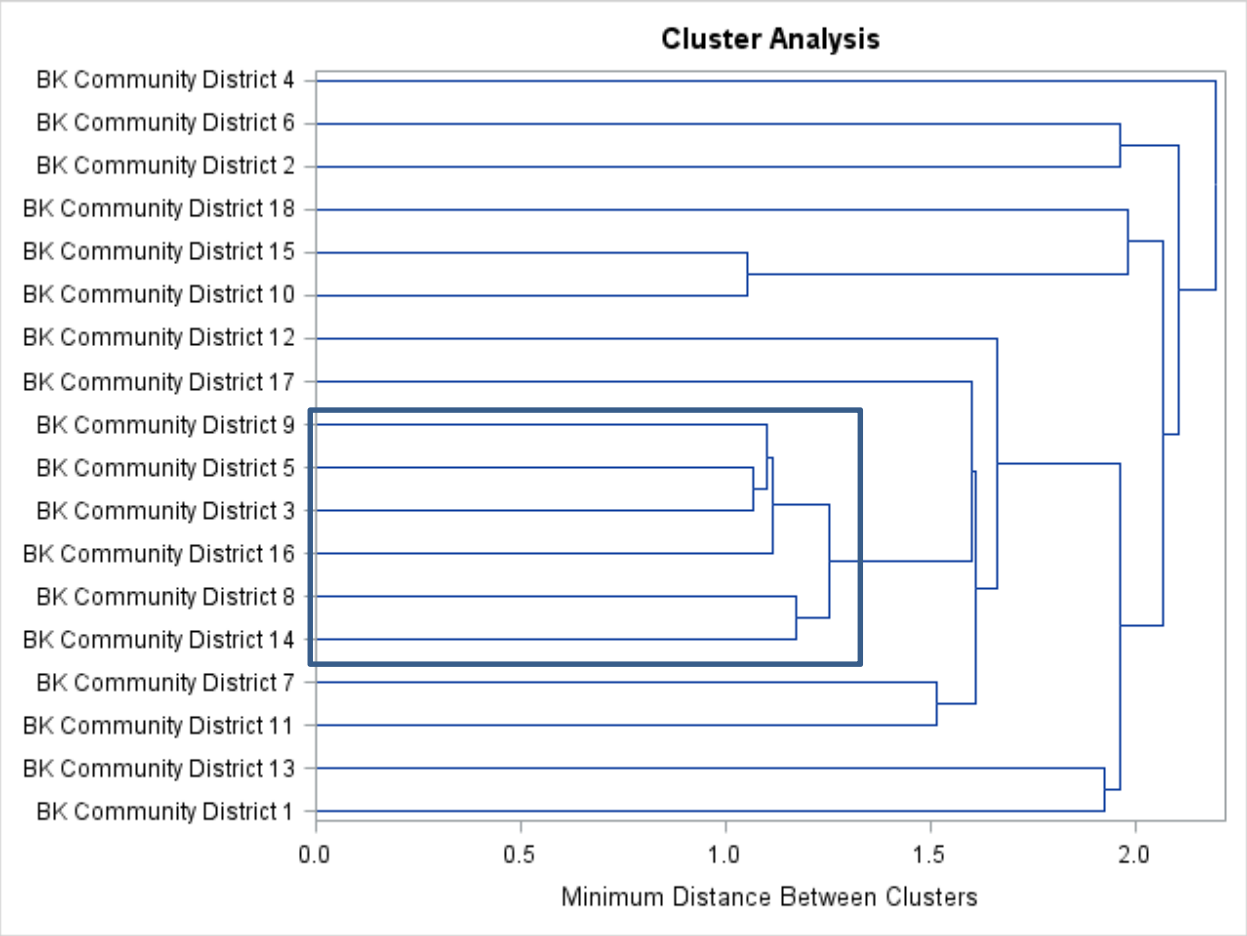
Below is a prediction matrix using a subset of 10 the total 18 groups. The diagonal has been bolded to show correct predictions.

Predicted Group vs Actual		Predicted group									
Actual group	# observations	1	10	11	12	13	14	15	16	17	18
1	646	**0**	8	1	87	548	2	0	164	14	39
10	671	2	**22**	3	60	572	12	3	160	16	88
11	786	2	32	**0**	97	649	6	2	283	38	134
12	634	1	19	2	**171**	438	3	3	195	17	62
13	612	4	12	0	35	**552**	9	2	150	21	42
14	650	4	20	1	76	543	**6**	2	217	15	58
15	739	6	30	2	81	608	12	**3**	155	27	79
16	465	2	5	0	46	411	1	0	**173**	14	28
17	511	7	13	0	49	438	4	3	177	**37**	85
18	733	5	42	0	100	574	12	3	165	65	**195**

***Part 6: Cluster Analysis;**

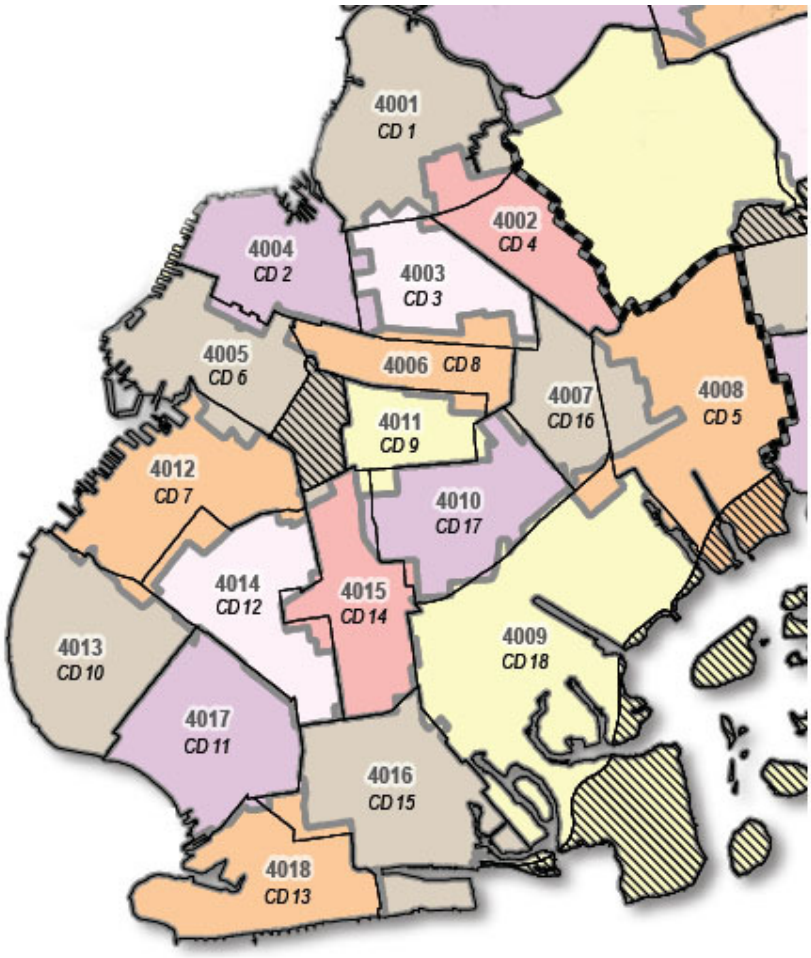
```
proc standard data= nyc.cdavg out= nyc.cdavg
mean=0 std=1;
var ANP AHINCP AyearsOwned AHPER APop
PercentOwn;
run;

proc cluster data=nyc.cdavg
outtree=treehousing method=single nonorm;
      VAR ANP AHINCP AyearsOwned AHPER APop
PercentOwn;
   id CD;run;
proc tree data=treehousing;
id CD;run;
```

Cluster History				
Clusters	**Clusters Joined**		**Freq**	**MinDist**
17	**10**	**15**	2	1.0517
16	**3**	**5**	2	1.0695
15	**CL16**	**9**	3	1.1006
14	**16**	**CL15**	4	1.1159
13	**14**	**8**	2	1.1719
12	**CL13**	**CL14**	6	1.2518
11	**11**	**7**	2	1.5139
10	**CL12**	**17**	7	1.6023
9	**CL11**	**CL10**	9	1.6124
8	**CL9**	**12**	10	1.6648
7	**1**	**13**	2	1.9257
6	**CL7**	**CL8**	12	1.9628
5	**2**	**6**	2	1.9654
4	**CL17**	**18**	3	1.9844
3	**CL6**	**CL4**	15	2.0677
2	**CL3**	**CL5**	17	2.1056
1	**CL2**	**4**	18	2.1969



Northern Brooklyn Districts Highlighted in Blue box

Appendix 2: Map of Brooklyn Community Districts



Brooklyn

1	4001	Greenpoint & Williamsburg	10	4013	Bay Ridge & Dyker Heights
2	4004	Brooklyn Heights & Fort Greene	11	4017	Bensonhurst & Bath Beach
3	4003	Bedford-Stuyvesant	12	4014	Borough Park, Kensington & Ocean Parkway
4	4002	Bushwick	13	4018	Brighton Beach & Coney Island
5	4008	East New York & Starrett City	14	4015	Flatbush & Midwood
6	4005	Park Slope, Carroll Gardens & Red Hook	15	4016	Sheepshead Bay, Gerritsen Beach & Homecrest
7	4012	Sunset Park & Windsor Terrace	16	4007	Brownsville & Ocean Hill
8	4006	Crown Heights North & Prospect Heights	17	4010	East Flatbush, Farragut & Rugby
9	4011	Crown Heights So., Prospect Lefferts & Wingate	18	4009	Canarsie & Flatlands

Appendix 3: Full Discriminant Hypothesis Tests

To test the first vector, the following test is conducted using Wilks Lambda:

H_0: $a_1 = 0$
H_a: $a_1 \neq 0$

Reject if $\boldsymbol{\Lambda_1} \leq \boldsymbol{\Lambda_\alpha}(\mathrm{p}, \mathrm{p}-\mathrm{m}+1,\ \mathrm{N}-\mathrm{k}-\mathrm{m}+1)$
where p = 5 k = 18, m = 5.5, N=10798

The value of Wilks' Lambda provided by SAS can be calculated by using the following formula that uses our 5 Eigenvalues:

$$\boldsymbol{\Lambda_1} = \prod_{i=1}^{5} \frac{1}{1+\lambda_i}$$

$$= \frac{1}{1+.1048} * \frac{1}{1+.0199} * \frac{1}{1+.0042} * \frac{1}{1+.0017} * \frac{1}{1+0.0003}$$

$$= .882$$

Find the critical value

$$= \boldsymbol{\Lambda_{.05}}(5, 17,\ 1000)$$

$$\approx .964$$

We reject H_0: $\mathbf{a_1} = 0$ since

$\boldsymbol{\Lambda_1} \leq \boldsymbol{\Lambda_{\alpha(5,17,1000)}} = .882 < .964$

Since there was proven significance with $\mathbf{a_1}$, the remaining a vectors are testing with $\mathbf{a_1}$ removed.

H_0: $a_2 = 0$
H_a: $a_2 \neq 0$

Reject if $\boldsymbol{\Lambda_2} \leq \boldsymbol{\Lambda_\alpha}(\mathrm{p}, \mathrm{p}-\mathrm{m}+1,\ \mathrm{N}-\mathrm{k}-\mathrm{m}+1)$
where p = 4, k = 18, n >1000

$$\boldsymbol{\Lambda_1} = \prod_{i=1}^{4} \frac{1}{1+\lambda_i}$$

$$= \frac{1}{1+.0199} * \frac{1}{1+.0042} * \frac{1}{1+.0017} * \frac{1}{1+0.0003}$$

$$= .9744$$

Find the critical value

$$= \boldsymbol{\Lambda_{.05}}(4, 16, \ 1000)$$

$$\approx .937$$

We fail reject H_0: $\mathbf{a_2} = 0$ since

$\boldsymbol{\Lambda_2} > \boldsymbol{\Lambda_{\alpha(5,17,1000)}} = .9744 > .937$