# BARUCH COLLEGE, CUNY

# DEPARTMENT OF STATISTICS

Final Project: STA 9705
Semester: Spring, 2020

# A Multivariate Statistical Study on NYC Major Felony Offenses in 2019


## Authors:

Mukherjee, Tanay
Wei, Ziping
Wu, Allen
Zhu, Youran

# Table of Contents

Introduction 3
Multivariate Analysis of Variance 3
Discriminant Analysis 5
Principal Component Analysis 8
PCA Using Covariance Matrix 8
PCA Using Correlation Matrix 10
Factor Analysis 133
Appendix – Other Figures 18
Appendix – SAS Code 19
Works Cited 20

## Introduction

New York City is a city composed of five boroughs and each borough itself is a collection of their own unique neighborhoods. Manhattan is famous for its skyline and is home to many famous tourist attractions including Central Park, Times Square, and the Empire State Building. Brooklyn has developed into a popular hipster community housing artists, musicians, and creative individuals, while Queens is reported to have the most diverse group of residents in its borders (Lubin, 2017). Those familiar with NYC's past knows that drugs, crime, and corruption were a constant for those living here, but over the course of the late 80's and 90's, much has changed to transform NYC into its current iconic status (Compan, 2017). While crime rates have dropped over the past 30 years, there are still some stigmas about the Bronx being the most dangerous borough in NYC and a never-ending debate on the borough battle between Brooklyn and Queens.

In accordance with New York State laws, the New York City Police Department reports crime and offense data every year[1]. All crimes are grouped into 3 main law class, felony, misdemeanor, and violation, and each class is further subcategorized into broad crime and offense categories. In this analysis, we will look at the seven major felony offenses in 2019 to provide an evidence-based approach study the five boroughs. There are 77 observations with 8 variables defined as precinct number, murder and non-negligent manslaughter, rape, robbery, felony assault, burglary, grand larceny, and grand larceny of motor vehicle. An additional column was added to categorize boroughs with the following labels: M = Manhattan, X = Bronx, K = Brooklyn, Q = Queens, and S = Staten Island.

## Multivariate Analysis of Variance

As five boroughs of New York City have different demographic in terms of populations, GDP, density, etc., we predict that the reported crime statistics would be different across boroughs as well. Therefore, before applying other analysis, we conducted a MANOVA test to access whether the five boroughs are significantly different in terms of those seven felony offenses. In the MANOVA test, p = 7 (seven felony offenses), k = 5 (five boroughs of New York City) and the null hypothesis is stated below.

$$H_0: \mu_M = \mu_X = \mu_K = \mu_Q = \mu_S = 0$$
$$H_a: At\ least\ one\ inequality\ in\ H_0$$

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall Borough Effect
H = Type III SSCP Matrix for Borough
E = Error SSCP Matrix

S=4 M=1 N=32

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---|---|---|---|---|---|
| Wilks' Lambda | 0.26360401 | 3.83 | 28 | 239.39 | <.0001 |
| Pillai's Trace | 1.02353458 | 3.39 | 28 | 276 | <.0001 |
| Hotelling-Lawley Trace | 1.81510597 | 4.20 | 28 | 155.5 | <.0001 |
| Roy's Greatest Root | 1.11782385 | 11.02 | 7 | 69 | <.0001 |

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

*Figure 1 - MANOVA test on the 5 boroughs*

[1] https://www1.nyc.gov/site/nypd/stats/crime-statistics/historical.page

Since the p-values of all four MANOVA tests are smaller than 0.05, we can reject the null hypothesis and conclude that the five boroughs are significantly different in crime statistics, based on the seven major felony offenses. With our initial assumption proven correct, a more in-depth test on the differences between the boroughs can be performed using contrast tests. Two tests were designed where the first test's null hypothesis is:

$$H_0: \delta = \mu_K - \mu_Q = 0$$
$$H_a: \delta = \mu_K - \mu_Q \neq 0$$

To test if there is a difference in crime rates between Brooklyn and Queens and the second test's null hypothesis is:

$$H_0: 4\mu_X - \left( \mu_M + \mu_K + \mu_Q + \mu_S \right) = 0$$
$$H_a: 4\mu_X - \left( \mu_M + \mu_K + \mu_Q + \mu_S \right) \neq 0$$

This will signify us any underlying difference in crime rates between Bronx of the other 4 boroughs.

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall Brooklyn vs. Queens Effect
H = Contrast SSCP Matrix for Brooklyn vs. Queens
E = Error SSCP Matrix

S=1 M=2.5 N=32

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---|---|---|---|---|---|
| Wilks' Lambda | 0.90732231 | 0.96 | 7 | 66 | 0.4654 |
| Pillai's Trace | 0.09267769 | 0.96 | 7 | 66 | 0.4654 |
| Hotelling-Lawley Trace | 0.10214417 | 0.96 | 7 | 66 | 0.4654 |
| Roy's Greatest Root | 0.10214417 | 0.96 | 7 | 66 | 0.4654 |

*Figure 2 - Contrast test between Brooklyn and Queens*

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall Bronx vs. the other 4 Boroughs Effect
H = Contrast SSCP Matrix for Bronx vs. the other 4 Boroughs
E = Error SSCP Matrix

S=1 M=2.5 N=32

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---|---|---|---|---|---|
| Wilks' Lambda | 0.52271035 | 8.61 | 7 | 66 | <.0001 |
| Pillai's Trace | 0.47728965 | 8.61 | 7 | 66 | <.0001 |
| Hotelling-Lawley Trace | 0.91310542 | 8.61 | 7 | 66 | <.0001 |
| Roy's Greatest Root | 0.91310542 | 8.61 | 7 | 66 | <.0001 |

*Figure 3 - Contrast test between Bronx and the other 4 boroughs*

According to the test results in Figure 2 and 3, we fail to reject the null hypothesis of the test Brooklyn vs Queens, indicating that there is no significant difference between these two boroughs. However, we can reject the null hypothesis of Bronx being same as the average of the other four boroughs.

## Discriminant Analysis

Using our dataset, we can perform a discriminant analysis to identify which of the seven variables helps contribute the most to the definition of the five NYC boroughs. Using the formula $s = \min(k-1, p)$, we can see that there are going to be four non-zero eigenvalues for $\mathbf{E}^{-1}\mathbf{H}$, indicating there would be four uncorrelated discriminant functions describing the separation of the NYC boroughs based on felony crime rates. In Figure 4, we can see there are indeed four non-zero eigenvalues describing the matrix $\mathbf{E}^{-1}\mathbf{H}$, and that the first two eigenvalues already account for over 90% of the separation.

| Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq) | | | |
|---|---|---|---|
| Eigenvalue | Difference | Proportion | Cumulative |
| 1.1178 | 0.5811 | 0.6158 | 0.6158 |
| 0.5367 | 0.4194 | 0.2957 | 0.9115 |
| 0.1173 | 0.0740 | 0.0646 | 0.9762 |
| 0.0433 | | 0.0238 | 1.0000 |

*Figure 4 - Eigenvalues of $E^{-1}H$*

| Test of H0: The canonical correlations in the current row and all that follow are zero | | | | |
|---|---|---|---|---|
| Likelihood Ratio | Approximate F Value | Num DF | Den DF | Pr > F |
| 0.26360401 | 3.83 | 28 | 239.39 | <.0001 |
| 0.55826685 | 2.42 | 18 | 189.99 | 0.0016 |
| 0.85789918 | 1.08 | 10 | 136 | 0.3794 |
| 0.95853018 | 0.75 | 4 | 69 | 0.5638 |

*Figure 5 - Wilks test on separation significance*

We define the null hypothesis of this experiment as $H_0: a_i = 0, H_a: a_i \neq 0$, from i = 1 to 4. Utilizing Wilks' test, we can perform an iterative test to examine each discriminant function and their significance to the separation of the five boroughs. In Figure 5, we can see the $\Lambda_1 \to \Lambda_4$ values in the "Likelihood Ratio" column and their significance based on approximate F values in the "Pr > F" column. Although the p-value of $\Lambda_2$ is below our threshold of 5%, the p-value is based off an approximate F value. Using the table of the lower critical values for Wilks $\Lambda$, we can confirm that of $\Lambda_2$ is significant since $\Lambda_2 = 0.558 < \Lambda_{0.05,6,3,60} = 0.608$. We used $\nu_E = 60$ since $\nu_E = 72$ is not made available in the table and since $\nu_E = 60$ is a more conservative comparison, it is sufficient for our test. The results show that only the first 2 discriminant functions have significant coefficients in describing the group separation.

<u>Discriminant Functions for Group Separation:</u>

$$z_1 = -0.015x_1 + 0.02x_2 - 0.014x_3 + 0.003x_4 + 0.01x_5 - 0.002x_6 + 0.021x_7$$
$$z_2 = -0.042x_1 - 0.055x_2 + 0.009x_3 + 0.006x_4 - 0.01x_5 - 0.002x_6 + 0.006x_7$$
$$z_3 = -0.017x_1 + 0.026x_2 + 0.015x_3 - 0.011x_4 + 0.012x_5 - 0.001x_6 - 0.004x_7$$
$$z_4 = -0.081x_1 - 0.007x_2 - 0.295x_3 + 0.045x_4 - 0.695x_5 + 0.922x_6 + 0.868x_7$$

| Pooled Within-Class Standardized Canonical Coefficients | | | | |
|---|---|---|---|---|
| **Variable** | **Can1** | **Can2** | **Can3** | **Can4** |
| **murder** | -0.052446887 | -0.149254622 | -0.060075784 | -0.081400007 |
| **rape** | 0.264328800 | -0.714597224 | 0.333829971 | -0.007147148 |
| **robbery** | -1.363201239 | 0.913568293 | 1.444049827 | -0.295019792 |
| **assault** | 0.507599337 | 0.915767514 | -1.705278790 | 0.044955200 |
| **burglary** | 0.671969215 | -0.699067207 | 0.820702483 | -0.694781423 |
| **gl** | -0.792463541 | 0.126485124 | -0.279122446 | 0.922335803 |
| **glmv** | 0.855594329 | 0.243311399 | -0.165501431 | 0.867741153 |

*Figure 6 - Standardized coefficients of the discriminant analysis*

In Figure 6, the table provides the standardized coefficients, which we will use for comparison and interpretation. Since only the first two eigenvalues are significant, we will only focus on analyzing the first two discriminant functions, Can1 and Can2. Using the magnitudes of the coefficients, we can see that "robbery" (-1.36) makes the largest contribution for Can1, but there are also some non-insignificant contributions from the variables "grand larceny of motor vehicles" (0.86), "grand larceny" (-0.79), "burglary" (0.67), and "assault" (0.51). However, if we take the sign into consideration, we can see the weight of robberies and grand larcenies (-2.15) is similar to the weight of grand larcenies of motor vehicles, burglaries, and assaults (2.04). In the second discriminant function, we see the variables "assault" (0.92) and "robbery" (0.91) providing the largest contributions to the separation of the boroughs. "rape" (-0.71) and "burglary" (-0.70) also provide significant amount of contribution to the separation of the boroughs.

Since only two discriminant functions are significant to the separation of the five boroughs, we can visualize the flattening of our 7-variable and 5-group data onto a 2-dimensional plane of our group separation. In Figure 7[2], we plot the transformed observations of crime in each police precinct onto the axes of the first two discriminant functions. There is an easily distinguishable section that represents the Manhattan police precincts. However, the other 4 boroughs do not have such clear separation. Although Bronx seems to have a clear section described by the transformed observations, there are several Bronx precincts that appear much further away from that area. Between the Brooklyn and Queens precincts, there is a large overlap between these two boroughs. Finally, Staten Island is fully encased by either the Brooklyn or Queens area.

While this analysis does not provide us with any information on the overall safety of the five boroughs, we can see how the major felony crime statistics in each neighborhood help to describe the borough as a whole. Anecdotally, Bronx had reputation of being dangerous and our analysis does show that the crime statistics seem to group their neighborhoods together. However, we do see some Bronx precincts (48, 49, 50) have crime statistics more closely related to a Brooklyn neighborhood. There are also 2 non-Bronx precincts (75, 103) that have crime statistics like some of Bronx's neighborhood. These precincts are located in Brownsville and Jamaica, both having a poor reputation of being unsafe.

[2] The same group separation plot with precinct number labels can be found in the appendix as Figure 25

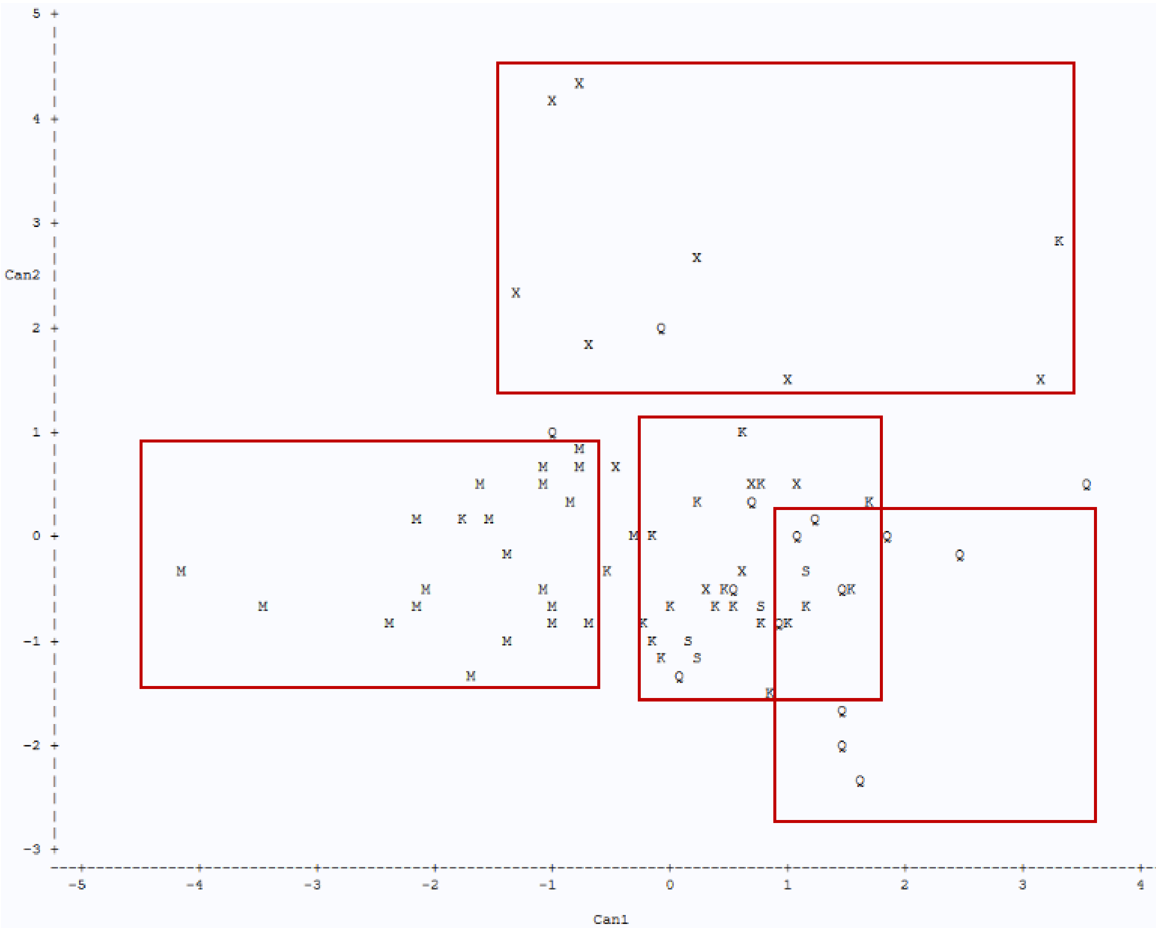


*Figure 7 - Plot of the first and second discriminant functions for group separation*

## Principal Component Analysis

Through principal component analysis (PCA), we are looking to maximize variance of a linear combination of all the variables. PCA analysis applied to our data with no groupings. Thus, we would consider each precinct on its own rather than grouping them into five boroughs. PCA could be used to reduce the number of dimensions and used as inputs for other analysis. It is particularly useful when the number of independent variables is large relative to the number of observations or the independent variables are highly correlated. It could also be an end in itself. For example, we could rank the safety of the neighborhood governed by the precinct with unequal weights for the felony crimes that spread the precincts out further on the scale and obtain a better ranking compared to simple average weighting.

| Covariance Matrix | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Murder | Rape | Robbery | Assault | Burglary | Larceny | LarcenyOfAutomotives |
| Murder | 14.1241 | 28.7162 | 251.3026 | 465.0395 | 96.4944 | -231.7162 | 90.0695 |
| Rape | 28.7162 | 202.4638 | 1268.6029 | 2132.3433 | 664.7061 | 753.1642 | 533.9978 |
| Robbery | 251.3026 | 1268.6029 | 12332.8660 | 18581.1328 | 5626.8433 | 8246.9067 | 3564.3756 |
| Assault | 465.0395 | 2132.3433 | 18581.1328 | 32876.5191 | 8064.7955 | 4310.3230 | 5980.7727 |
| Burglary | 96.4944 | 664.7061 | 5626.8433 | 8064.7955 | 4949.8274 | 11051.3681 | 2247.4927 |
| Larceny | -231.7162 | 753.1642 | 8246.9067 | 4310.3230 | 11051.3681 | 117629.8537 | 1420.8156 |
| LarcenyOfAutomotives | 90.0695 | 533.9978 | 3564.3756 | 5980.7727 | 2247.4927 | 1420.8156 | 2382.0424 |

*Figure 8 - Covariance matrix*

First, we exam the variance of each variables. We could see that the variances the of each variable are not in same scale. The variables “larceny” and “assault” have relatively large variances while the variable “murder” has relatively small variances. Thus, we will use both covariance matrix and correlation matrix to conduct PCA and compare the results.

## PCA Using Covariance Matrix

| Total Variance | 170387.69651 |
|---|---|

| Eigenvalues of the Covariance Matrix | | | | |
|---|---|---|---|---|
| | Eigenvalue | Difference | Proportion | Cumulative |
| 1 | 120003.357 | 73962.149 | 0.7043 | 0.7043 |
| 2 | 46041.208 | 43710.254 | 0.2702 | 0.9745 |
| 3 | 2330.954 | 1165.315 | 0.0137 | 0.9882 |
| 4 | 1165.639 | 371.108 | 0.0068 | 0.9950 |
| 5 | 794.530 | 749.256 | 0.0047 | 0.9997 |
| 6 | 45.274 | 38.539 | 0.0003 | 1.0000 |
| 7 | 6.735 | | 0.0000 | 1.0000 |

*Figure 9 - Eigenvalues based on covariance matrix*

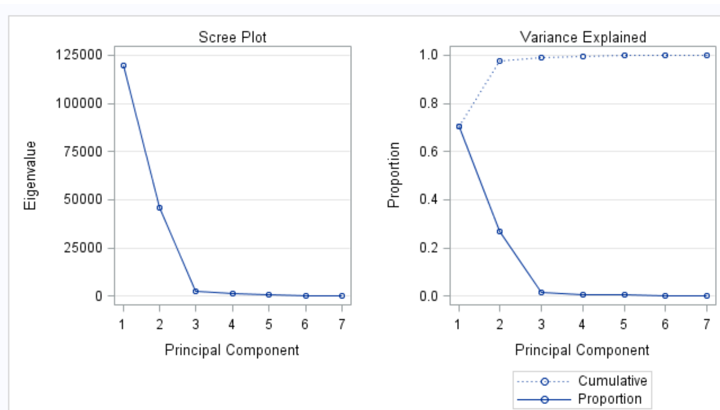


*Figure 10 - Scree plot based on covariance matrix*

To find the number of principal components (PCs) to retain, we will look at eigenvalues of the covariance matrix (S). There is total of seven eigenvalues and their percent of variance for each eigenvalue can be seen in Figure 9. We could see that the largest eigenvalue explains 70.43% of variance; the first two eigenvalues combined explain 97.45% of variance. Also, we could refer to the scree plot for better visualization in Figure 10. Since the first two eigenvalues capture a significant proportion of the total of the eigenvalues, we can retain the first two principal components.

$$z_1 = a_1^{'}x = -.001x_1 + .009x_2 + .096x_3 + .081x_4 + .106x_5 + .986x_6 + .021x_7$$
$$z_2 = a_2^{'}x = .012x_1 + .054x_2 + .475x_3 + .829x_4 + .199x_5 - .139x_6 + .159x_7$$

**Eigenvectors**

| | Prin1 | Prin2 | Prin3 | Prin4 | Prin5 | Prin6 | Prin7 |
|---|---|---|---|---|---|---|---|
| Murder | -.001289 | 0.012430 | -.000822 | 0.005965 | -.000693 | -.029552 | 0.999467 |
| Rape | 0.009329 | 0.054165 | 0.042651 | 0.059882 | 0.069879 | 0.992916 | 0.028423 |
| Robbery | 0.095773 | 0.475040 | 0.091126 | -.772976 | 0.399023 | -.012160 | -.001180 |
| Assault | 0.080645 | 0.828698 | -.349691 | 0.356698 | -.236205 | -.035436 | -.013830 |
| Burglary | 0.105531 | 0.199858 | 0.814690 | -.042720 | -.532282 | -.006795 | -.001995 |
| Larceny | 0.986234 | -.139153 | -.077295 | 0.038893 | 0.021878 | -.002316 | 0.002653 |
| LarcenyOfAutomotives | 0.020974 | 0.158752 | 0.144857 | 0.517992 | 0.704486 | -.108575 | -.007394 |

*Figure 91 - Coefficients of PC based on covariance matrix*

In Figure 11, we can see the largest coefficients in first PC ($z_1$) and second PC ($z_2$), 0.986 and 0.828, correspond to the variables with largest variances, "larceny" and "assault". But still, $z_1$ and $z_2$ are meaningful linear functions.

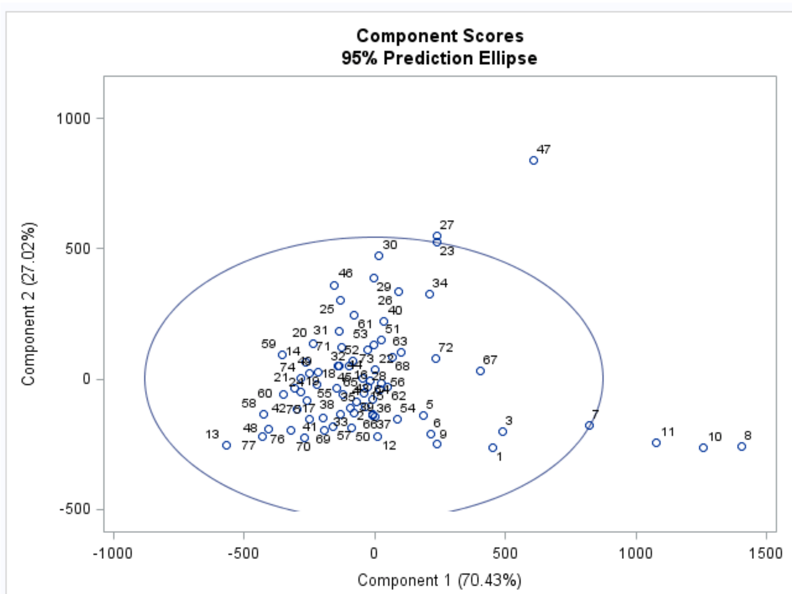


*Figure 102 - 95% prediction ellipse based on PCA using the covariance matrix*

We plot the first two principal components for each observation vector to detect outliers. There are several outliers – observation 47, 8, 10, 11 & 27, which are precinct 75, 14, 18, 19 & 44 respectively in Figure 12.

## PCA Using Correlation Matrix

| Simple Statistics | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Murder | Rape | Robbery | Assault | Burglary | Larceny | LarcenyOfAutomotives |
| Mean | 4.142857143 | 22.77922078 | 173.6363636 | 266.1818182 | 140.0389610 | 561.6753247 | 70.51948052 |
| StD | 3.758199057 | 14.22897653 | 111.0534377 | 181.3188328 | 70.3550098 | 342.9720888 | 48.80617152 |

| Correlation Matrix | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Murder | Rape | Robbery | Assault | Burglary | Larceny | LarcenyOfAutomotives |
| Murder | 1.0000 | 0.5370 | 0.6021 | 0.6824 | 0.3649 | -.1798 | 0.4910 |
| Rape | 0.5370 | 1.0000 | 0.8028 | 0.8265 | 0.6640 | 0.1543 | 0.7689 |
| Robbery | 0.6021 | 0.8028 | 1.0000 | 0.9228 | 0.7202 | 0.2165 | 0.6576 |
| Assault | 0.6824 | 0.8265 | 0.9228 | 1.0000 | 0.6322 | 0.0693 | 0.6758 |
| Burglary | 0.3649 | 0.6640 | 0.7202 | 0.6322 | 1.0000 | 0.4580 | 0.6545 |
| Larceny | -.1798 | 0.1543 | 0.2165 | 0.0693 | 0.4580 | 1.0000 | 0.0849 |
| LarcenyOfAutomotives | 0.4910 | 0.7689 | 0.6576 | 0.6758 | 0.6545 | 0.0849 | 1.0000 |

*Figure 113 - Correlation matrix*

To complete the principal component analysis, the next step is to replace the covariance matrix with the correlation matrix (R) in Figure 13. Using the correlation matrix will remove any issues with the scale of the variables and we can expect it to have a more balanced representation.

| Eigenvalues of the Correlation Matrix | | | | |
|---|---|---|---|---|
| | Eigenvalue | Difference | Proportion | Cumulative |
| 1 | 4.40664476 | 3.12148176 | 0.6295 | 0.6295 |
| 2 | 1.28516300 | 0.78989844 | 0.1836 | 0.8131 |
| 3 | 0.49526456 | 0.15127287 | 0.0708 | 0.8839 |
| 4 | 0.34399169 | 0.09035705 | 0.0491 | 0.9330 |
| 5 | 0.25363464 | 0.09753931 | 0.0362 | 0.9692 |
| 6 | 0.15609533 | 0.09688932 | 0.0223 | 0.9915 |
| 7 | 0.05920601 | | 0.0085 | 1.0000 |

*Figure 14 - Eigenvalues based on correlation matrix*

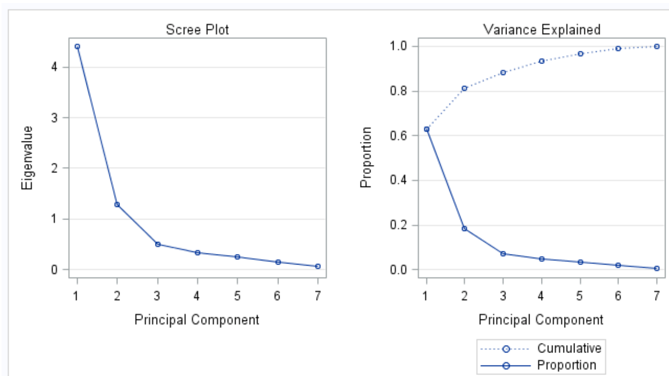


*Figure 15 - Scree plot based on correlation matrix*

In Figure 15, we can see the scree plots show a natural break point after the 2nd eigenvalue. But the first two eigenvalues combined only explain 81.31% of variance in this case. From the eigenvalues in Figure 14, we could see that we will need to retain first four principal components to explain more than 90% percent of variation. The first four principal components could be reached by eigenvectors table in Figure 16 as follows:

$$z_1 = a_1^{'}x = 0.328x_1 + .432x_2 + .443x_3 + .442x_4 + .385x_5 + .101x_6 + .394x_7$$
$$z_2 = a_2^{'}x = -.441x_1 - .014x_2 + .015x_3 - .139x_4 + .357x_5 + .811x_6 - .033x_7$$
$$z_3 = a_3^{'}x = .536x_1 - .245x_2 + .237x_3 + .199x_4 - .087x_5 + .328x_6 - .665x_7$$
$$z_4 = a_4^{'}x = .598x_1 - .301x_2 - .388x_3 - .389x_4 + .315x_5 + .137x_6 + .364x_7$$

| Eigenvectors | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Prin1 | Prin2 | Prin3 | Prin4 | Prin5 | Prin6 | Prin7 |
| Murder | 0.327802 | -.441149 | 0.535391 | 0.597898 | 0.188057 | -.119432 | -.064642 |
| Rape | 0.432034 | -.014132 | -.245098 | -.300796 | 0.378976 | -.712588 | -.105785 |
| Robbery | 0.443413 | 0.015305 | 0.236667 | -.388033 | -.180520 | 0.353298 | -.662694 |
| Assault | 0.442270 | -.139576 | 0.199643 | -.389318 | -.021397 | 0.246917 | 0.729428 |
| Burglary | 0.384511 | 0.356504 | -.087223 | 0.314539 | -.728493 | -.278983 | 0.099899 |
| Larceny | 0.100979 | 0.810705 | 0.327693 | 0.137186 | 0.439529 | 0.099889 | 0.056515 |
| LarcenyOfAutomotives | 0.394499 | -.033676 | -.665150 | 0.363519 | 0.253138 | 0.452059 | -.015167 |

*Figure 126 - Coefficients of PC based on correlation matrix*

We could see that all coefficients for PC 1 are positive. We could infer that the first component is an overall measure of felony situation. The second component is a shape component that contrasts the "robbery", "burglary" and "larceny" with "murder", "rape", "assault" and "grand larceny of motor vehicle". If we focus on large coefficients (> 0.1), this is quite interesting since the contrast is between "murder" & "assault" and "burglary" & "larceny". The former group are violent crimes while the latter group are property crimes.

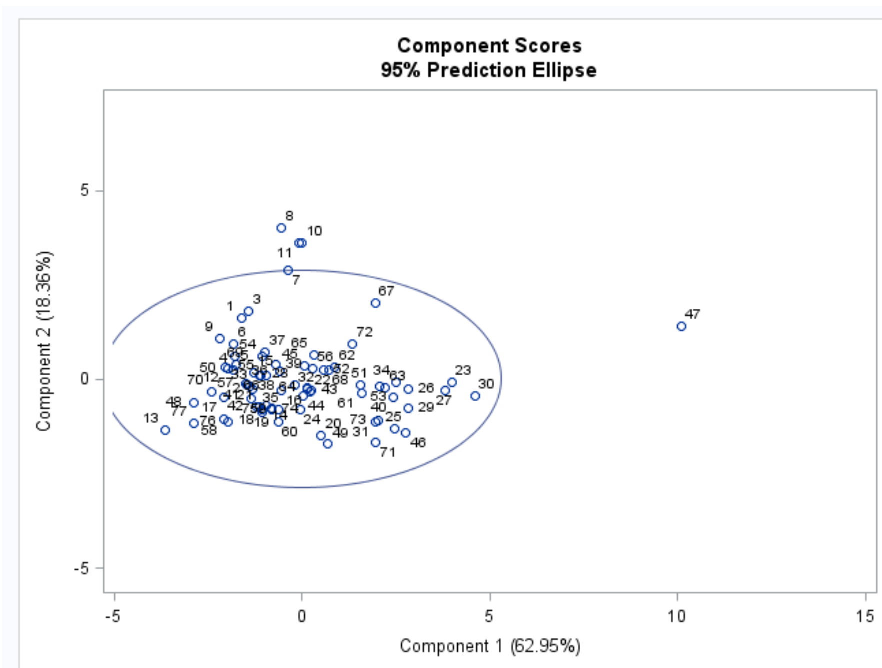


*Figure 137 - 95% prediction ellipse based on PCA using the correlation matrix*

We also plot the first two principal components for each observation vector. From the scatter plot in Figure 17, we could see the outliers are similar as before – observation 47, 8, 10, 11 which correspond to precincts 75, 14, 18, 19 respectively. We highlight these four police precincts to the NYC map for better visualization in Figure 18. We could see three of them are in mid-town Manhattan area and one of them in Brooklyn.

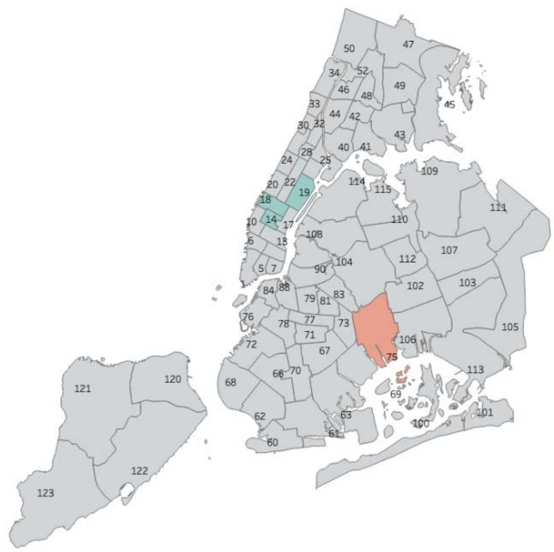


*Figure 148 - Outliers based on PC using correlation matrix*

Generally, using covariance matrix for PCA remains closer to the spirit since and intention of PCA is to maximize variance. It is more suitable if the components are to be used as input for further analysis and if components are in same scale. In our project, two principal components (PCs) are enough to cover over 90% of variance. However, they are dominant by two variables with large variances. Thus, we also conduct PCA from correlation matrix for better interpretation.

In general, PCs will be more interpretable if correlation matrix is used when the variances of variables differ a lot or if the measurement units are not commensurate. Since correlation matrix is scale-free, there will be a more balanced representation for each variable. However, the components from a given correlation matrix are not unique to that correlation matrix. In other words, the components depend only on the ratios (relative values) of the correlations, not on their actual values. Also, in our project, we need to retain four PCs to cover over 90% of total variance, which is less effective in terms of dimension reduction.

## Factor Analysis

In **factor analysis,** we represent the variables $y_1, y_2 \ldots\ y_p$ as linear combinations of a few random variables $f_1, f_2 \ldots\ f_m\ (m < p)$ called factors. The factors are underlying constructs or latent variables that "generate" the y's. Like the original variables, the factors vary from individual to individual; but unlike the variables, the factors cannot be measured or observed. The existence of these hypothetical variables is therefore open to question.

The two types: exploratory and confirmatory.

- Exploratory factor analysis is if you do not have any idea about what structure your data is or how many dimensions are in a set of variables.
- Confirmatory Factor Analysis is used for verification as long as you have a specific idea about what structure your data is or how many dimensions are in a set of variables.

In this dataset for NYC felony we are trying to observe how different factors are playing an important role in deciding the relationship between the features (in this case, different kinds of crimes) and which ones can be separated out as part of the analysis to reduce the dimensions.

Before we can decide on whether to use priors = SMC we must check the inverse correlation matrix and if it is invertible, we can decide. From the output below, we can confirm it is indeed invertible.

The FACTOR Procedure
Initial Factor Method: Principal Factors

| Inverse Correlation Matrix | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Murder | Rape | Robbery | Assault | Burglary | Larceny | LarcenyOfAutomotives |
| Murder | 2.09519 | 0.19092 | -0.07150 | -1.38123 | -0.07711 | 0.50967 | -0.18793 |
| Rape | 0.19092 | 4.43513 | -0.43302 | -2.17593 | -0.19151 | -0.18139 | -1.60801 |
| Robbery | -0.07150 | -0.43302 | 8.94126 | -7.01300 | -1.58467 | -0.69765 | 0.32415 |
| Assault | -1.38123 | -2.17593 | -7.01300 | 9.95969 | 0.45965 | 0.71611 | -0.12942 |
| Burglary | -0.07711 | -0.19151 | -1.58467 | 0.45965 | 3.19498 | -1.04416 | -1.08599 |
| Larceny | 0.50967 | -0.18139 | -0.69765 | 0.71611 | -1.04416 | 1.66479 | 0.40615 |
| LarcenyOfAutomotives | -0.18793 | -1.60801 | 0.32415 | -0.12942 | -1.08599 | 0.40615 | 2.87937 |

| Partial Correlations Controlling all other Variables | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Murder | Rape | Robbery | Assault | Burglary | Larceny | LarcenyOfAutomotives |
| Murder | 1.00000 | -0.06263 | 0.01652 | 0.30237 | 0.02980 | -0.27290 | 0.07651 |
| Rape | -0.06263 | 1.00000 | 0.06876 | 0.32739 | 0.05088 | 0.06675 | 0.44997 |
| Robbery | 0.01652 | 0.06876 | 1.00000 | 0.74316 | 0.29649 | 0.18083 | -0.06388 |
| Assault | 0.30237 | 0.32739 | 0.74316 | 1.00000 | -0.08148 | -0.17587 | 0.02417 |
| Burglary | 0.02980 | 0.05088 | 0.29649 | -0.08148 | 1.00000 | 0.45275 | 0.35805 |
| Larceny | -0.27290 | 0.06675 | 0.18083 | -0.17587 | 0.45275 | 1.00000 | -0.18550 |
| LarcenyOfAutomotives | 0.07651 | 0.44997 | -0.06388 | 0.02417 | 0.35805 | -0.18550 | 1.00000 |

*Figure 159 - Inverse correlation matrix*

**Communalities** (also known as $h^2$) are the estimates of the variance of the factors, as opposed to the variance of the variable which includes measurement error. Initially the communality estimates are set equal to the $R^2$ between each variable and all others.

The table below shows the initial estimates of communalities:

| Prior Community Estimates: SMC | | | | | | |
|---|---|---|---|---|---|---|
| Murder | Rape | Robbery | Assault | Burglary | Larceny | LarcenyOfAutomotives |
| 0.52271654 | 0.77452727 | 0.88815893 | 0.89959527 | 0.68700881 | 0.39932298 | 0.65270237 |

*Figure 20 - Communality estimates*

**Factor loadings:** Commonality is the square of the standardized outer loading of an item. Analogous to Pearson's r-squared, the squared factor loading is the percent of variance in that indicator variable explained by the factor. To get the percent of variance in all the variables accounted for by each factor, add the sum of the squared factor loadings for that factor (column) and divide by the number of variables. This is the same as dividing the factor's eigenvalue by the number of variables.

**Interpreting factor loadings:** By one rule of thumb in confirmatory factor analysis, loadings should be 0.7 or higher to confirm that independent variables identified a priori are represented by a particular factor, on the rationale that the 0.7 level corresponds to about half of the variance in the indicator being explained by the factor. However, the 0.7 standard is a high one and real-life data may well not meet this criterion, which is why some researchers, particularly for exploratory purposes, will use a lower level such as 0.6 for the central factor and 0.25 for other factors. In any event, factor loadings must be interpreted in the light of theory, not by arbitrary cutoff levels.

The table below shows the final estimates of loadings:

| Rotated Factor Pattern | | |
|---|---|---|
| | Factor1 | Factor2 |
| Murder | 0.71667 | -0.21302 |
| Rape | 0.85445 | 0.23022 |
| Robbery | 0.89387 | 0.27044 |
| Assault | 0.94941 | 0.07350 |
| Burglary | 0.63445 | 0.58835 |
| Larceny | 0.00585 | 0.67564 |
| LarcenyOfAutomotives | 0.75833 | 0.19833 |

*Figure 21 - Loadings after varimax rotation*

The factors that affect the question the most (and therefore have the highest factor loadings) are bolded. Factor loadings are similar to correlation coefficients in that they can vary from -1 to 1. The closer the factors are to -1 or 1, the more they affect the variable. A factor loading of zero would indicate no effect.

For Factor 1 all the crimes except “larceny” are strongly associated. Whereas for “burglary” it is a pretty close call between Factor 1 and Factor 2 but it doesn’t have a loading of 0.7 or higher for either of those factors and hence, we decide to go with Factor 2 as SAS by default uses the range (0.3 to 0.6) to differentiate factors in cases of conflict. As mentioned in the interpretation earlier, in real life situations a strong association with factors with a value of 0.7 or higher is not always expected in that case we lower the benchmark to 0.6, or better put, the one which is closest to the 0.7 gets the precedence.

The table below shows the proportion of variance explained by each factor:

| Eigenvalues of the Reduced Correlation Matrix: Total = 4.82403217 Average = 0.68914745 | | | | |
|---|---|---|---|---|
| | Eigenvalue | Difference | Proportion | Cumulative |
| 1 | 4.16931551 | 3.39801859 | 0.8643 | 0.8643 |
| 2 | 0.77129692 | 0.58305985 | 0.1599 | 1.0242 |
| 3 | 0.18823707 | 0.17609374 | 0.0390 | 1.0632 |
| 4 | 0.01214333 | 0.06586833 | 0.0025 | 1.0657 |
| 5 | -.05372500 | 0.03630030 | -0.0111 | 1.0546 |
| 6 | -.09002530 | 0.08318506 | -0.0187 | 1.0359 |
| 7 | -.17321036 | | -0.0359 | 1.0000 |

**2 factors will be retained by the NFACTOR criterion.**

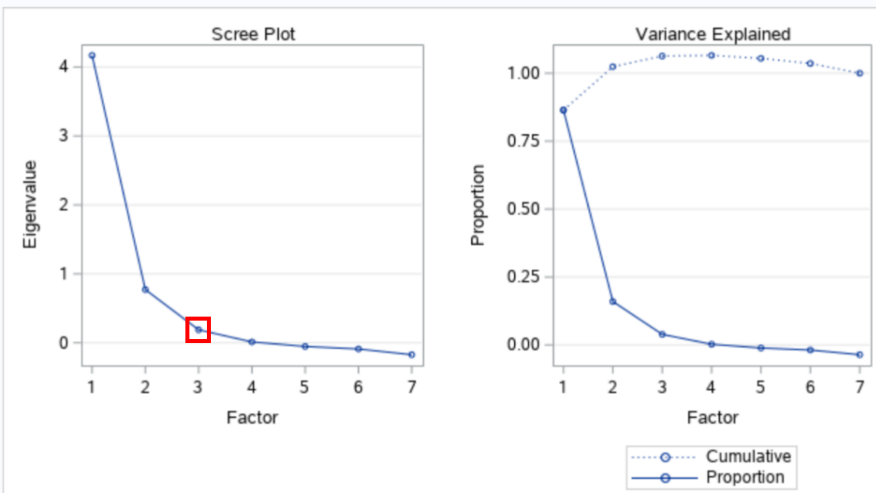


*Figure 22 – Eigenvalues with their proportions and scree plot*

From the scree plot above we see the natural break (also called as the elbow point) is seen at m=3. This also confirms that 2 factors are sufficient to proceed as after this, the share of variance will significantly vary for each feature.

**Rotation** serves to make the output more understandable, by seeking so-called "Simple Structure": A pattern of loadings where each item loads strongly on only one of the factors, and much weaker on the other factors. Rotations can be orthogonal or oblique (allowing the factors to correlate).

**Varimax rotation** is an orthogonal rotation of the factor axes to maximize the variance of the squared loadings of a factor (column) on all the variables (rows) in a factor matrix, which has the effect of differentiating the original variables by extracted factor. Each factor will tend to have either large or small loadings of any particular variable. A varimax solution yields results which make it as easy as possible to identify each variable with a single factor. This is the most common rotation option. However, the orthogonality (i.e., independence) of factors is often an unrealistic assumption. Oblique rotations are inclusive of orthogonal rotation, and for that reason, oblique rotations are a preferred method.

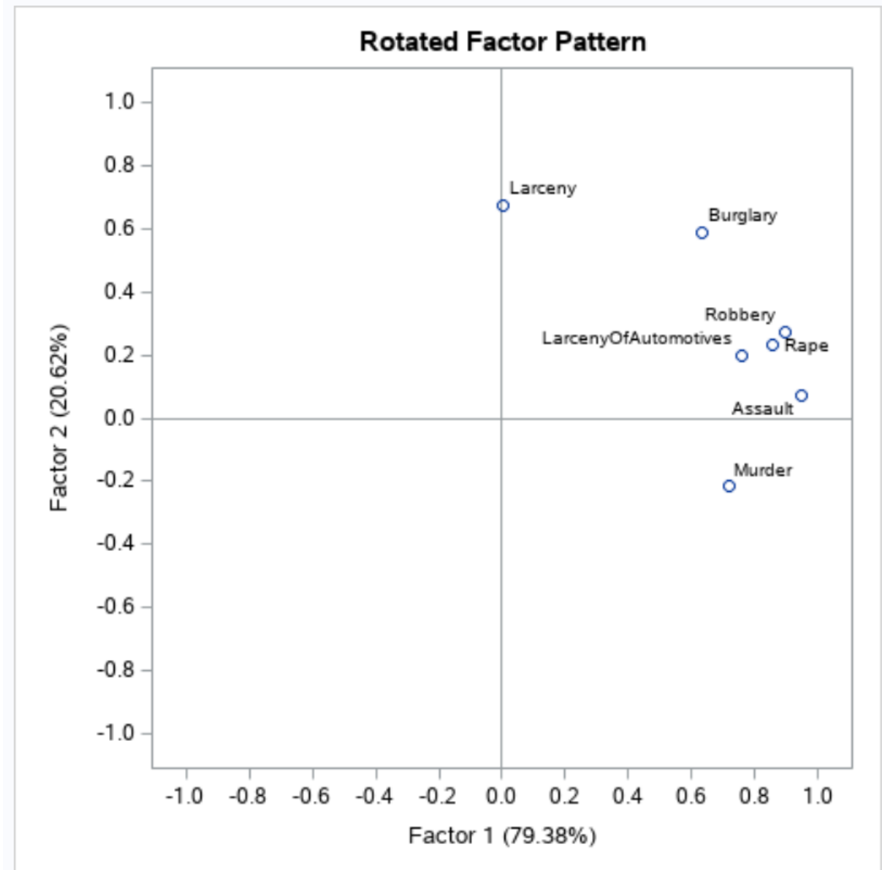


*Figure 163 - Plot of 7 variables on 2 factors after varimax rotation*

This rotated pattern simply tells us how the variance is shared across two factors. The Factor 1 on the x-axis represents the share of variance of ~80% amongst all features whereas for Factor 1 it is ~20%. As seen above, it is coming more clearly in this rotated factor pattern that ‘murder’ seems to be the strongest association in Factor 1 whereas for Factor 2 it is “larceny”. “Burglary” looks like an exception, but it is better and stronger fit to Factor 1 than Factor 2 based on out learning from the loading table above.

**Path analysis** is used to describe the directed dependencies among a set of variables. This includes models equivalent to any form of multiple regression analysis, factor analysis, canonical correlation analysis, discriminant analysis, as well as more general families of models in the multivariate analysis of variance and covariance analyses (MANOVA, ANOVA)

Path diagrams are like flowcharts. They show variables interconnected with lines that are used to indicate causal flow. Each path involves two variables (in either boxes or ovals) connected by either arrows (lines, usually straight, with an arrowhead on one end) or wires (lines, usually curved, with no arrowhead), or "slings" (with two arrowheads). Arrows are used to indicate "directed" relationships, or linear relationships between two variables. An arrow from X to Y indicates a linear relationship where Y is the dependent variable and X the independent variable.

In the following path diagram one can observe how it relates in the present case of NYC felony data. There are two factors and they detail on the respective dependencies. However, we see a conflict with variable 'Burglary' which has dependency on both the factors. The factor is 0.59 for Factor 1 and 0.63 for Factor 2. They are close enough but in cases of conflict we use the range (0.3 to 0.6) and by that logic Factor 1 is a better representation of the variable "burglary".

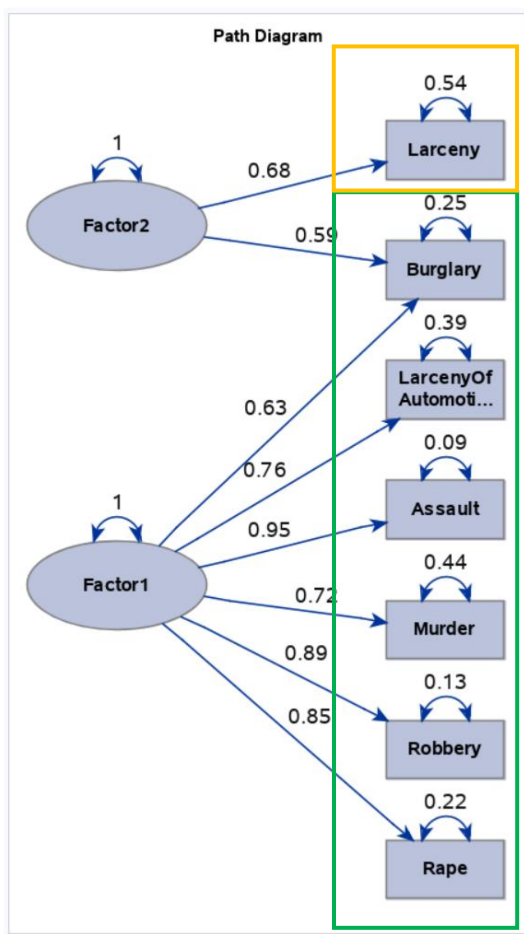


*Figure 174 - Factor analysis path diagram*

## Appendix – Other Figures

This figure is in line with Figure 7 above and includes labels for each precinct:

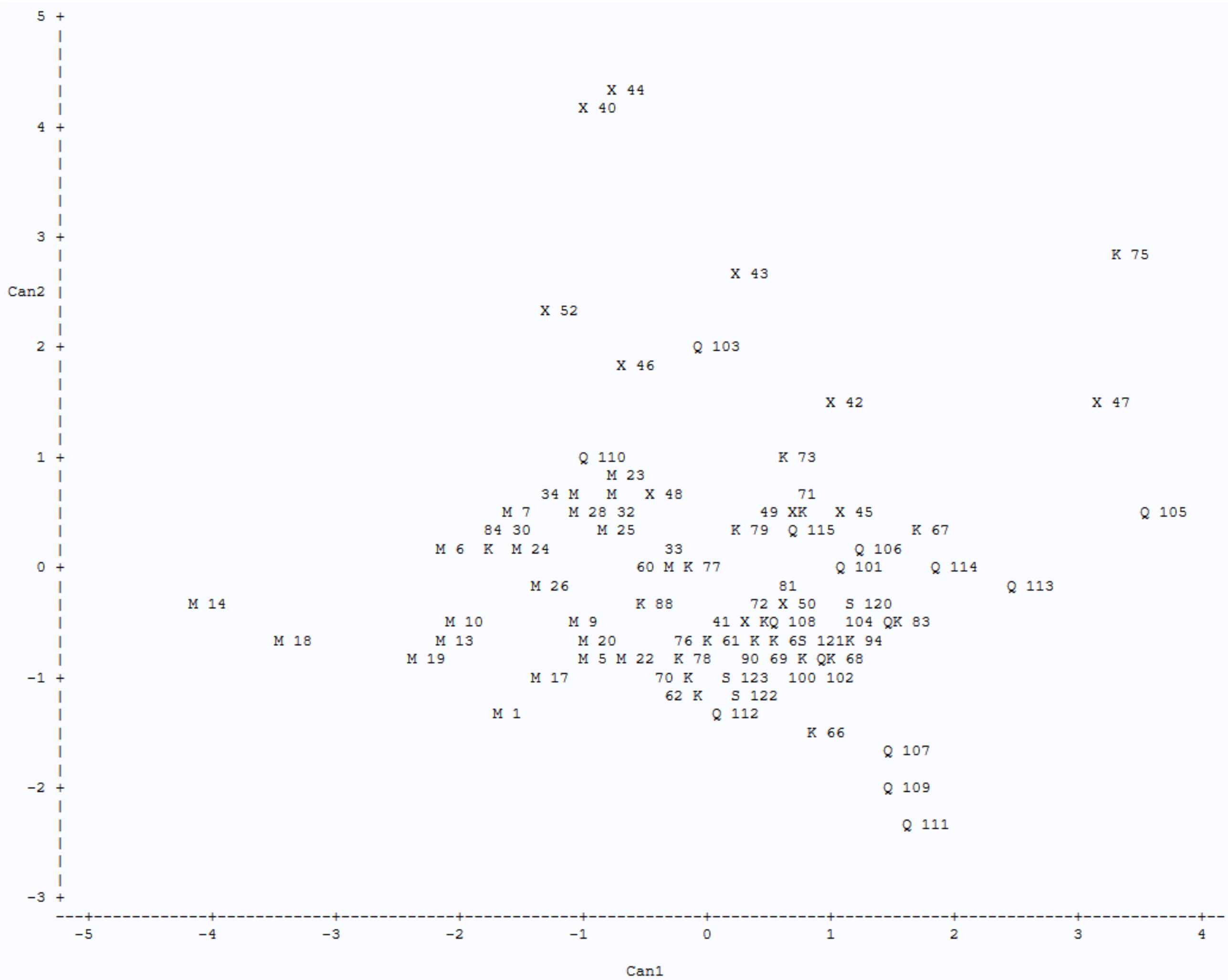


*Figure 185 - Plot of the first and second discriminant functions for group separation with precinct labels*

## Appendix – SAS Code

```
DATA NYC_Felony;
        INFILE 'MV.DAT';
        INPUT Borough $ Precinct Murder Rape Robbery Assault Burglary Larceny
                LarcenyOfAutomotives;

PROC GLM;
        /* MANOVA and Contrast Tests */
        CLASS Borough;
        MODEL Murder Rape Robbery Assault Burglary Larceny
                LarcenyOfAutomotives=Borough;
        MANOVA H=Borough/PRINTE PRINTH;
        CONTRAST 'Brooklyn vs. Queens' Borough 0 0 1 -1 0;
        CONTRAST 'Bronx vs. the other 4 Boroughs' Borough -1 4 -1 -1 -1;
        MANOVA H=Borough/PRINTE PRINTH;
        RUN;
----


PROC CANDISC OUT=NYC;
        /* Discriminant Analysis */
        CLASS Borough;
RUN;

PROC PRINT DATA=NYC;
RUN;

PROC PLOT DATA=NYC;
        /* Discriminant Analysis Plot */
        PLOT CAN2*CAN1=Borough;
        PLOT CAN2*CAN1=Borough $ Precinct;
        RUN;
----


PROC PRINCOMP COV OUT=RESULTS PLOT(NCOMP=2)=SCORE(ELLIPSE);
        /* Principal Component Analysis using Covariance Matrix */
        VAR Murder Rape Robbery Assault Burglary Larceny LarcenyOfAutomotives;
RUN;
```

```
PROC PRINT DATA=RESULTS;
	VAR PRIN1 PRIN2;
RUN;

PROC PRINCOMP OUT=RESULTS PLOT(NCOMP=2)=SCORE(ELLIPSE);
	/* Principal Component Analysis using Correlation Matrix */
	VAR Murder Rape Robbery Assault Burglary Larceny LarcenyOfAutomotives;
RUN;

PROC PRINT DATA=RESULTS;
	VAR PRIN1 PRIN2;
RUN;

----

PROC FACTOR METHOD=PRIN PRIORS=MAX ALL;
	/* Principal Factor Method */
	VAR Murder Rape Robbery Assault Burglary Larceny LarcenyOfAutomotives;
RUN;

PROC FACTOR METHOD=PRIN PRIORS=SMC NFACT=2 ROTATE=VARIMAX PLOTS=ALL;
	/* Principal Factor Method */
	VAR Murder Rape Robbery Assault Burglary Larceny LarcenyOfAutomotives;
RUN;

-----
```

## Works Cited

Compan, L. (2017, November 11). Retrieved from Great Future Stories: https://greatfuturestories.com/get-inspired-in-new-york/2017/11/13/new-york-city-back-in-the-1980s

Lubin, G. (2017, February 15). *Queens has more languages than anywhere in the world — here's where they're found*. Retrieved from Business Insider: https://www.businessinsider.com/queens-languages-map-2017-2