**Determining Number of Groups: Hierarchical Analysis**

We used SPSS v22 to conduct the Cluster Analysis.

|  |  |
| --- | --- |
| **Step 1: Import Data** | * Using SPSS, import the csv file “Cluster\_Feature\_Vectors\_top.csv” * File Composition:   + Rows: 173 Rows of unique Issuer IDs   + Columns:     - Issuer\_ID: Unique ID Label to distinguish rows     - 60 Feature Columns: All scalar numerals   + Value: Each number in the feature vector represents the aggregated total number of tickets issued during the given calendar time period for the given feature |
| **Step 2: Define Hierarchical Analysis** | * **Agglomerative:** We determined that an agglomerative method would best allow us to determine an appropriate number of cluster groups for the top performers via a dendrogram. * **Non-Normalize:** We chose not to normalize the data because all of the feature variables have the same similar scale on the dimensions. We wanted to preserve the relative size of more popular features versus smaller ones to emphasize dis-similarity. * **Linkage Rules:** We chose a Ward’s Method as a criterion method to determine clustering because all of the feature variables are scalar numbers and do not have binary numbers involved. * **Distance:** We chose a squared Euclidian distance formula to give more progressively greater weight on clusters that are further apart |
| **Step 3: Output** | * SPSS Generates the following output: Cluster\_Analysis\_Kmeans\_Agglomerative.xlsx   + Proximity Matrix: Calculates the distance between each feature/row entry   + Agglomeration Schedule: Determines the ordering of which rows are added to each other to create the cluster. It lists the nodes combined and the distance coefficients   + Dendrogram: Visual representation of the clustering layered on a scaled distance |
| **Step 4:**  By plotting the agglomeration schedule one can see that there is a noticeable elbow curve within the last clusters. | Elbow |
| **Interpreting Dendrogram:**  Using the dendrogram we spotted 5 potential clusters.    3  5  4  2  1 | |

**Generating Clustering Groups: K-Means**

We used SPSS v22 to conduct the Cluster Analysis.

|  |  |
| --- | --- |
| **Step 1: Import Data** | * Using SPSS, import the csv file “Cluster\_Feature\_Vectors\_top.csv” * File Composition:   + Rows: 173 Rows of unique Issuer IDs   + Columns:     - Issuer\_ID: Unique ID Label to distinguish rows     - 60 Feature Columns: All scalar numerals   + Value: Each number in the feature vector represents the aggregated total number of tickets issued during the given calendar time period for the given feature |
| **Step 2: Define K Means** | * **Data Selection:** We choose all 60 features and used “issuer\_id” as the label * **Number of Cluster:** Using the insight from the Hierarchical clustering we chose a cluster count of 5 * **Defined Iteration Count:** 40 |
| **Step 3: Output** | SPSS Generates the following output:   * Cluster Membership: The assignment of each of the 173 issuer\_ids to one of the 5 clusters. It also has the node distance from the centroid. * Iteration History: Shows each of the K-means convergence steps. Running the k-means a couple of times, the five clusters consistently converged within less than 10 iterations. * Anova Table: The analysis for each of the 60 feature contribution to the overall variance. * Final Centroid Totals (“kmeans5\_3Feature\_Final”): For each of the 5 clusters, outputs the feature total |
| **Step 4: Interpretation** | Running the k-means a couple of times, the five clusters consistently converged within less than 10 iterations.   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | Iteration | Change in Cluster Centers | | | | | | 1 | 2 | 3 | 4 | 5 | | 1 | 1121.972 | 1128.145 | 1304.358 | 1220.947 | 1122.462 | | 2 | 144.968 | 155.421 | 95.306 | 43.900 | 33.288 | | 3 | 37.132 | 133.567 | 48.715 | 100.735 | 0.000 | | 4 | 41.536 | 50.871 | 31.960 | 18.242 | 0.000 | | 5 | 56.208 | 50.386 | 19.555 | 43.208 | 62.191 | | 6 | 0.000 | 24.783 | 18.367 | 24.477 | 0.000 | | 7 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | |
| **Step 4: Interpretation** | The Centroid output reveals the aggregate total tickets for each cluster. To improve comparability, we converted each centroid total into the percentage of the Centroid aggregate total. We color coded the entries to reveal which features contributed the most to the overall number of tickets. This generated the corresponding color table which clear distinguishes differences between the five clusters.   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | |  | **Cluster Centroid % of Total** | |  |  |  | |  | **1** | **2** | **3** | **4** | **5** | | Week\_AM\_14 | 8.89% | 1.90% | 10.64% | 2.71% | 37.59% | | Week\_AM\_38 | 2.56% | 2.28% | 5.76% | 0.78% | 0.31% | | Week\_AM\_69 | 0.17% | 0.50% | 5.94% | 3.16% | 7.38% | | Week\_AM\_21 | 39.41% | 1.57% | 15.74% | 1.02% | 1.13% | | Week\_AM\_37 | 0.26% | 1.63% | 2.95% | 0.37% | 0.11% | | Week\_AM\_20 | 6.88% | 0.68% | 5.09% | 0.13% | 1.22% | | Week\_AM\_31 | 0.07% | 0.18% | 1.78% | 0.81% | 8.96% | | Week\_AM\_16 | 2.39% | 1.54% | 7.51% | 0.10% | 2.52% | | Week\_AM\_46 | 2.65% | 0.50% | 2.77% | 0.21% | 3.17% | | Week\_AM\_40 | 6.48% | 0.39% | 2.95% | 0.21% | 2.04% | | Week\_AM\_47 | 0.00% | 0.03% | 4.19% | 1.20% | 6.16% | | Week\_AM\_19 | 2.72% | 0.27% | 2.11% | 0.29% | 2.63% | | Week\_AM\_42 | 0.02% | 0.15% | 0.54% | 0.60% | 0.62% | | Week\_AM\_71 | 2.25% | 0.39% | 1.75% | 0.29% | 0.57% | | Week\_AM\_17 | 0.65% | 0.30% | 1.39% | 0.23% | 0.99% | | Week\_PM\_14 | 0.76% | 10.23% | 1.54% | 19.26% | 4.01% | | Week\_PM\_38 | 0.69% | 11.65% | 2.11% | 3.50% | 0.06% | | Week\_PM\_69 | 0.02% | 2.25% | 2.02% | 11.59% | 0.90% | | Week\_PM\_21 | 0.79% | 0.24% | 0.21% | 0.08% | 0.03% | | Week\_PM\_37 | 0.29% | 9.75% | 2.20% | 3.24% | 0.06% | | Week\_PM\_20 | 0.43% | 8.42% | 1.42% | 1.36% | 0.48% | | Week\_PM\_31 | 0.02% | 2.31% | 0.93% | 16.52% | 1.38% | | Week\_PM\_16 | 0.41% | 7.20% | 1.06% | 1.33% | 0.96% | | Week\_PM\_46 | 0.91% | 4.80% | 0.42% | 2.11% | 0.99% | | Week\_PM\_40 | 0.48% | 4.36% | 0.54% | 0.99% | 0.54% | | Week\_PM\_47 | 0.00% | 0.15% | 0.69% | 4.41% | 0.45% | | Week\_PM\_19 | 0.53% | 2.16% | 0.42% | 1.36% | 0.40% | | Week\_PM\_42 | 0.00% | 1.04% | 0.87% | 6.60% | 0.59% |   The greater the green color, the larger contribution the specific row feature had to the total aggregate tickets in the cluster. |
| **Interpretation** | The highlighted green features were also reinforced within the ANOVA table as contributing the highest to the overall variations.   |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | **Feature** | **Mean Square** | **df** | **Mean Square2** | **df3** | **F** | **Sig.** | | Week\_AM\_21 | 11703701.23 | 4 | 62317.68 | 168 | 187.807 | 0 | | Week\_AM\_14 | 6149960.361 | 4 | 48930.274 | 168 | 125.688 | 0 | | Week\_PM\_14 | 3201941.114 | 4 | 53693.625 | 168 | 59.634 | 0 | | Week\_PM\_31 | 2765263.061 | 4 | 35097.749 | 168 | 78.787 | 0 | | Week\_PM\_69 | 1226829.305 | 4 | 27064.219 | 168 | 45.33 | 0 | | Week\_PM\_38 | 950459.981 | 4 | 36947.14 | 168 | 25.725 | 0 | | Week\_PM\_37 | 658201.412 | 4 | 38422.366 | 168 | 17.131 | 0 | | Week\_PM\_20 | 527691.37 | 4 | 17377.61 | 168 | 30.366 | 0 | |

**Conclusion of Clustering: Real World Interpretation**

Using the significant features from the K-Means clustering, we were able to create personas for each of the 5 clusters. You can reference the “ViolationCodes\_match” tab in the excel file.

|  |  |  |
| --- | --- | --- |
| **Persona Name** | **Description of Behavior** | **Feature Importance** |
| **Butch ‘The Cleaner’**  **the-cleaner-01.jpg** | Street-cleaning: Meter people who issue “Street cleaning” tickets during weekday mornings. (12.14% of issuers). | Cluster 1:  “Week\_am\_21” = 40% of centroid totals.  Violation 21 = Street Cleaning: No parking where parking is not allowed by sign, street marking or traffic control device. |
| **Curbside Terror’ Terry**  **ticket.jpg** | Curbside: Meter people who issue expired-meter, general no parking, and general curbside parking tickets in the afternoon. Top weekend performers. (27.75% of issuers) | Cluster 2:  Both “Week\_PM\_38” , “Week\_PM\_37” features were the most prominent of all the clusters.  Violation Code 37: Parking in excess of the allowed time  Violation Code 38: Failing to show a receipt or tag in the windshield |
| **‘Early-Birdie’ Rose**  **187640_1.jpg** | Morning generalists: Street-cleaning, general no standing or parking, time violation in zone, fail to show receipt, parked in truck unloading zone, etc. during weekdays (24.86% of issuers) | Cluster 3:  This was the most unremarkable cluster with broad distributions of tickets across code in the morning. |
| **Shawn ‘The Shield’**  **1327339.jpg** | Afternoon Commercial enforcers: No parking or standing, truck unloading  zone, and commercial metered zone tickets during weekday afternoons. (24.28% of issuers) | Cluster 4:  Highest contribution in “Week\_PM\_31” and “Week\_PM\_69”.  Code 31: Standing of a non-commercial vehicle in a commercial metered zone.  Code 69: Failing to show a muni-meter receipt, commercial meter zone. |
| **‘Mafioso’ Benetto**  **Jennifer+Garner+Getting+Parking+Ticket+Brentwood+p_U5ntVrcgEl.jpg** | No standing in weekday mornings: Meter people who issue “General No Standing: Standing or parking where standing is not allowed by sign, street marking or; traffic control device.” tickets during weekday mornings. (10.98% of issuers). | Cluster 5:  “Week\_AM\_14”: Had the most significant feature contribution.  Code 14: General No Standing: Standing or parking where standing is not allowed by sign, street marking or; traffic control device |