Math 748: Professor Tao He

Course Project: Progress Report II

Who Did We Stop This Time?

Gabrielle Salamanca

November 22, 2024

Contents

1. Introduction 2
2. Dataset Cleaning and Editing 2
3. Feature Selection: Unsupervised Learning 3
    - 3.1. Principal Component Analysis 3
    - 3.2. K-Means Clustering 3
4. Results: Unsupervised Learning 4
    - 4.1. Principal Unsupervised Learning 4
    - 4.2. K-Means Clustering 6
5. Feature Selection: Supervised Learning 7
    - 5.1. Decision Tree 7
6. Results: Supervised Learning 8
    - 6.1. Decision Tree 8
7. Plots and Visuals 9

I. Introduction

There can be multiple reasons why a police officer would stop someone on the road. Perhaps the license plate has expired registration tags, someone was speeding, traffic violations, or someone was driving under the influence to name a few. And any of these can be issued a citation, meaning the culprit is being legally charged with violating the traffic law.

For this project, I want to find out if there's a pattern to receiving a citation, if certain characteristics or circumstances are more likely in receiving one.

II. Dataset Cleaning and Editing

There's a column that I sincerely wanted to split apart, and it was the date column. It was written as year, month, day; and I personally wanted only the month. So, I had to use the transform function to split the date into three columns: day, month, year. This subset would be named policeDate, which would later have its problem column turned into binary values. If the person was stopped for a traffic reason, they received a 1; and if they were stopped for suspicious, they received a 0. I also made another subset called polDate where I removed the year and date, because we already know these reports are from 2017, and the date will keep giving us trouble for one of the methods. The final subset was numPol, where I removed categorical columns.

I did this, because I was planning to do a mix of supervised and unsupervised learning. numPol was especially made, because Principal Component Analysis does not take categorical data. PCA can only take numerical data, because it calculates either the covariance or correlation matrix, and it relies on linear algebra operations like eigenvalue decomposition. If there was a possibility to change the column values into a binary value or integer, I would have done so.

Gender unfortunately could not be, because it had three values: Male, Female, Unknown. I don't think I could assign proper weights to each one, and this included preRace and Race.

III. Feature Selection: Unsupervised Learning

Unsupervised learning was my main objective for this project. As stated previously, I wanted to specifically see if there was some sort of pattern with the Minneapolis police stopping cars, and so I have decided to use PCA and K-Means Clustering.

A. Principal Component Analysis

PCA is a dimensionality reduction technique in statistics and machine learning. It transforms the original data into a lower-dimensional space, which means transforming it with possibly correlated features into a smaller set of uncorrelated features called principal components. As this happens, it retains most of the original variance in the data.

B. K-Means Clustering

K-Means Clustering is a machine learning algorithm used to partition data into a specified number clusters (k) based on similarity. It groups the unlabeled dataset into different clusters, where the datapoints in the same cluster or more similar to each than to those in other clusters.

IV. Results: Unsupervised Learning

A. Principal Component Analysis

Importance of Components:

	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6	Comp 7
Standard deviation	1.2540	1.1457	1.0385	1.0173	0.9627	0.8156	0.6396
Proportion of Variance	0.2247	0.1875	0.1541	0.1479	0.1324	0.0950	0.0584
Cumulative Proportion	0.2246	0.4122	0.5662	0.7141	0.8465	0.9416	1

The goal of PCA is to retain enough components to explain most of the variance while also reducing dimensionality, so we want a cumulative proportion threshold sto be between 80-95%. We see in this summary of PCA, we find that by Component 5, 84.65% of the variance can be explained. If we wanted closer to 95%, then up Component 6 would suffice.

It must be also noted that by Component 3, 56.62% of the variance can be explained, meaning they capture the majority of the patterns in the data. Since I'm satisfied with having 84.65% of the variance explained, we can drop Component 6 and 7.

Loadings:

	C1	C2	C3	C4	C5	C6	C7
problem		0.722		0.134		0.640	0.218
citationIssued	0.136	0.641	-0.222	-0.255	0.184	-0.647	
personSearch	0.694	-0.179					0.692
vehicleSearch	0.704			0.106		0.155	-0.684
policePrecinct		0.113	0.336	0.867		-0.339	
day			-0.655	0.277	-0.700		
month		0.149	0.636	-0.276	-0.687	-0.151	

	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6	Comp 7
SS loadings	1	1	1	1	1	1	1
Proportion Var	0.143	0.143	0.143	0.143	0.143	0.143	0.143
Cumulative Var	0.143	0.286	0.429	0.571	0.714	0.857	1

Loadings represent the contribution of each variable to a principal component, indicating how strongly each variable correlates with the corresponding component. In a loading, if the value is either a large positive or negative, it means that the variable heavily influence the component. If it's close to zero, the variable doesn't contribute as much to the component.

The variables with large positive contributions to Component 1 are personSearch and vehicleSearch For Component 2, it's problem and citationIssued both positively affect it; Component 3, day affects it negatively and month affects it positively; Component 4, only policePrecinct affects it positively; Component 5, day and month both negatively affect it;

Component 6, problem positively affects it and citationIssued negatively affects it; Component 7, personSearch positively affects it and vehicleSearch negatively affects it.

Both personSearch and vehicleSearch may imply a a trend with Component 1, perhaps a higher frequency contribute to the variation or there are areas with greater perceived risk. Perhaps the type of problem being issued a citation affects Component 2. For Component 3, perhaps certain days are less common for the type of problem, while month implies there's certain seasonal trends for spikes in the stops. For Component 4, perhaps only certain police precincts, perhaps they have a larger district or have more police activity, contribute to the higher score. For Component 5, certain days and months may have lower incidents. For Component 6, there's a possibility of capturing problematic areas that are highly impacted by incidents but have fewer citations, or areas with higher issued citations yet fewer incidents. Finally, Component 7 may imply a pattern of law enforcement focus. personSearch could have more targeted interventions, while vehicleSearch is more routine.

B. K-Means Clustering

To find out the optimal number of clusters for K-Means Clustering, I ran the fviz_nbclust function to visualize it by using the silhouette method.The silhouette method assesses how similar each point is to its own cluster and how different it is from other clusters. The silhouette score will range from -1 and +1: +1 indicates the point is well-matched to its own cluster, 0 indicates that it's on or very near the boundary between 2 clusters, and -1 indicates that it might be misclassified and is closer to another cluster. All the values didn't fall under 0, but there was one value that was far above the rest: k = 2. So, I ran that value through the kmeans function, printed its centers, and visualized its clusters.

Centers:

	problem	Citation Issued	Person Search	Vehicle Search	Police Precinct	Day	Month
1	-0.2079	0.1062	2.3403	1.724	-0.1092	-0.0171	0.0024
2	0.0333	-0.0170	-0.3756	-0.2767	0.0175	0.0027	-0.0004

Every center has a negative and positive value in each variable. A negative value suggests that on average, the values of that variable in the data points within that cluster are below the overall mean of that variable if it's standardized. If it's a positive value, then they're above the overall mean on average. And since all centroids have a negative and positive value, this suggests that there's a contrast in how the clusters differ with respect to that variable.

V. Feature Selection: Supervised Learning

I decided to include supervised learning for this, because I can predict if a citation was issued or not. So, I wanted to try a method that I could have done on my Math 448 project, and use one of the familiar ones.

A. Decision Tree

A Decision Tree is used for both classification and regression tasks, which we will be using for the former. It splits the data insto subsets based on the value of input features. It then builds a model in the form of a tree. Each node in the tree represents a decision on an attribute, each branch represents an outcome of the test, and each leaf represents the final prediction.

VI. Results: Supervised Learning

A. Decision Tree

The tree started with the entire dataset with 0.17 probability of the majority class at this node. It then split into two subsets, but it's unclear what feature or condition caused the split. Finally, it has two leaves: 0 and 1. 0 has 83% proportion of samples, and 1 has 17% proportion of samples. So, we have an imbalance class, thus the tree struggled to find meaningful splits for the minority class. So, I will have to downsample my dataset, and when I did, it was basically the same.

I did try a confusion matrix, and here are the results:

Confusion Matrix:

	No	Yes
No	406	132
Yes	286	556

It has an 69.7% accuracy, meaning it can correctly predict approximately 70% of all cases. It has an 80.8% precision, meaning when the model predicts "Yes", it's right 80.8% of the time. 66% recall means it correctly identified 66% of all actual "yes" cases. 75.5% specificity means it correctly identified the actual "No" cases 75.5% of the time. An F1-score of 72.5% means the model strikes a reasonable balance between identifying "Yes" cases and avoiding false positives.

VII. Plots and Visuals

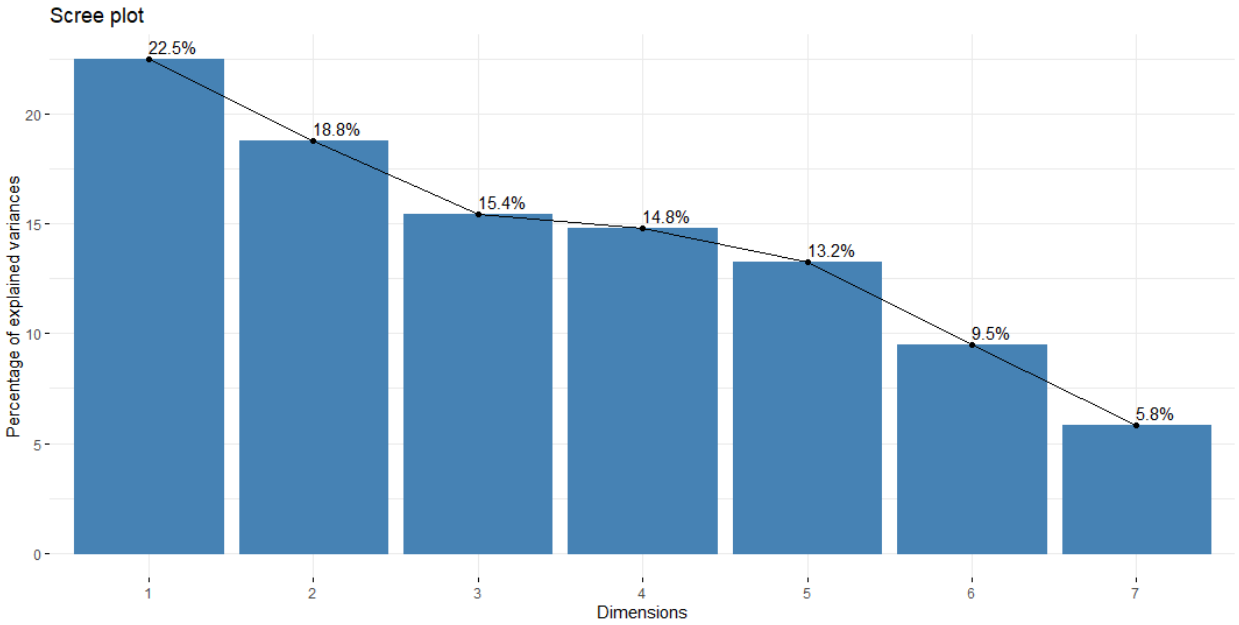

Figure 1: PCA Scree Plot of polCA

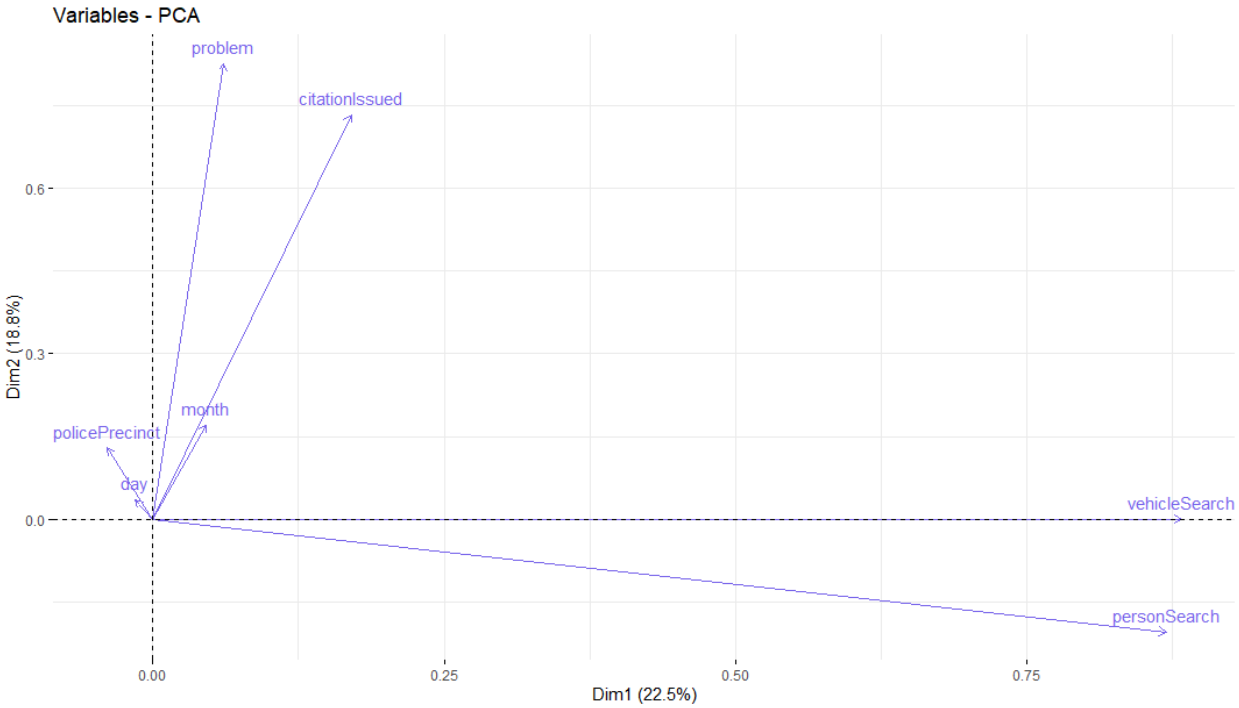

Figure 2: PCA Variable Plot of polCA

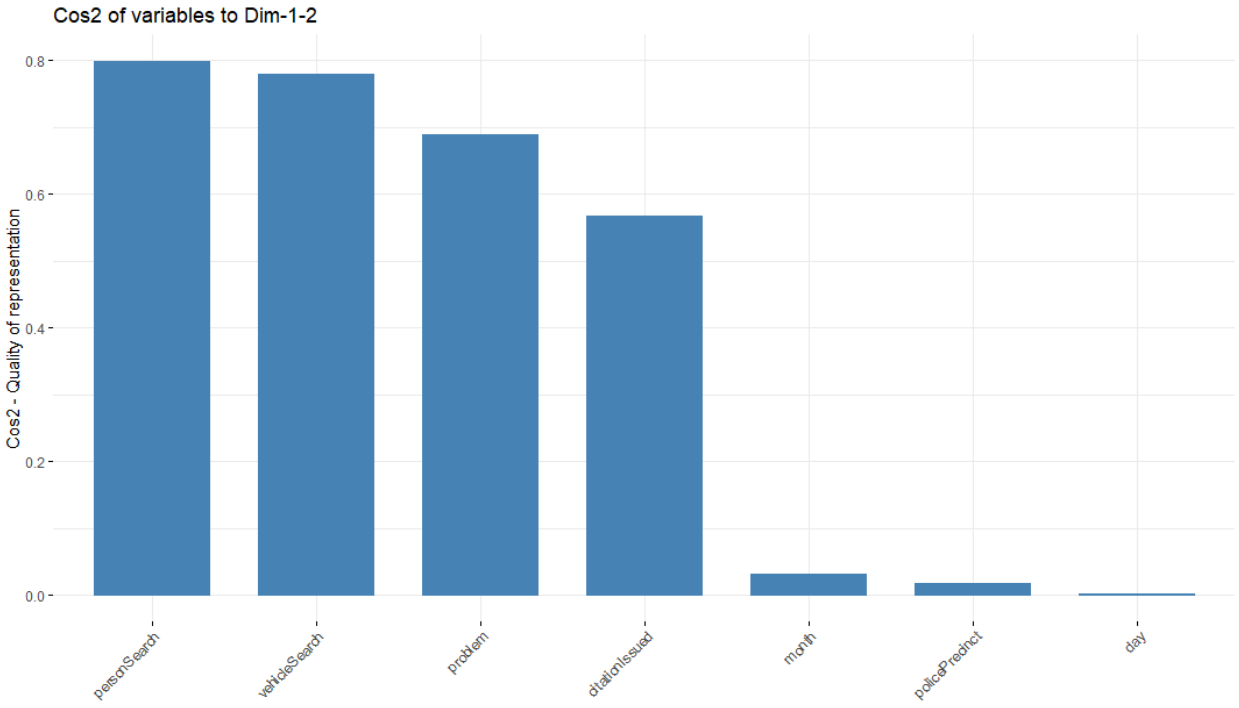

Figure 3: PCA Cos2 Plot of polCA

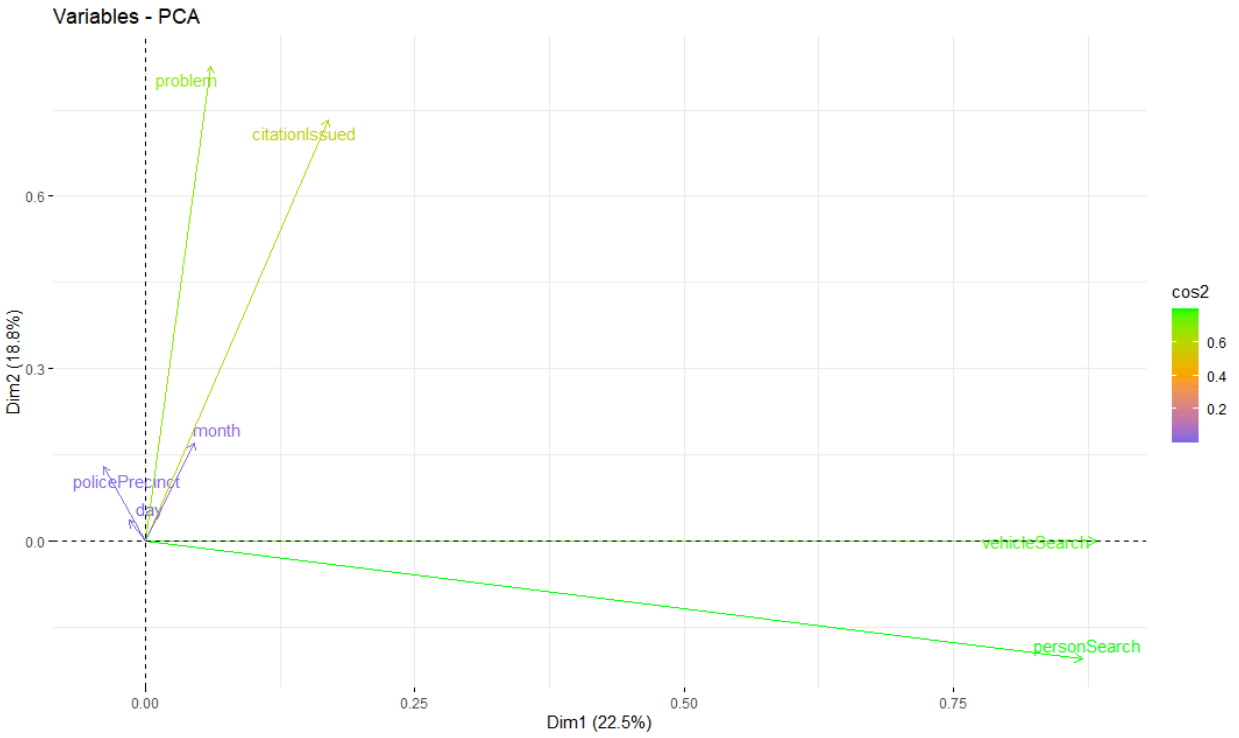

Figure 4: PCA Cos2 Variables Plot of polCA

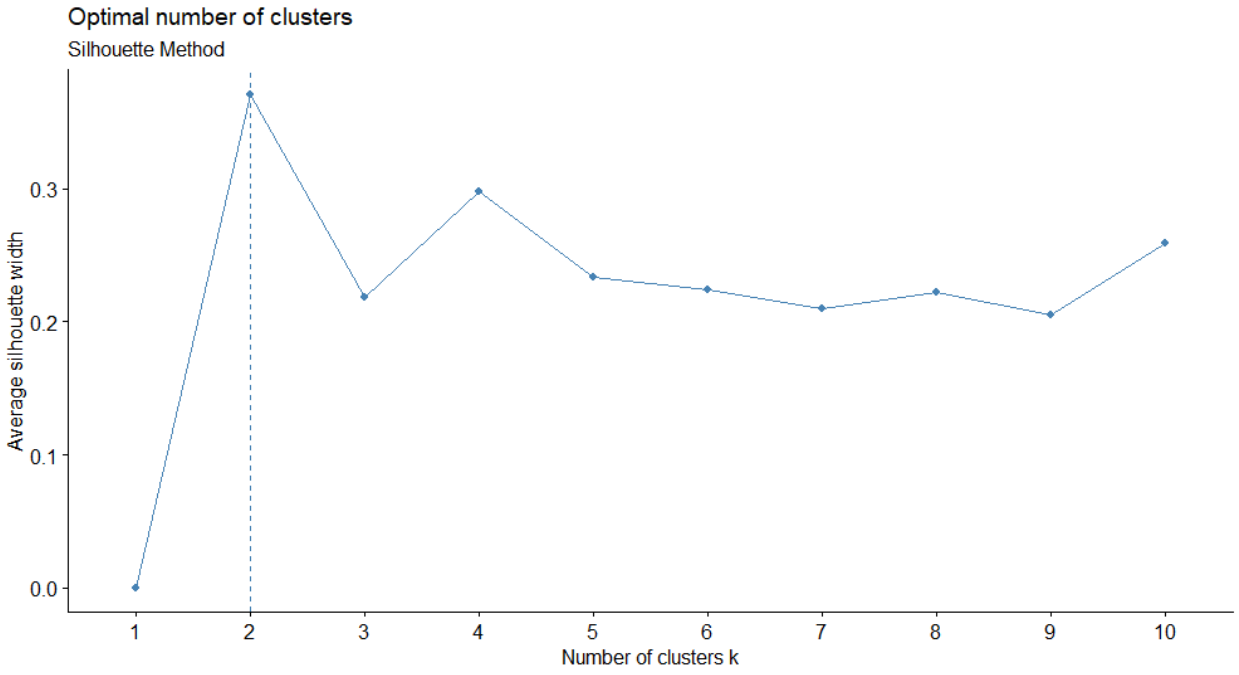

Figure 5: K-Means Cluster Plot of optimal number of clusters

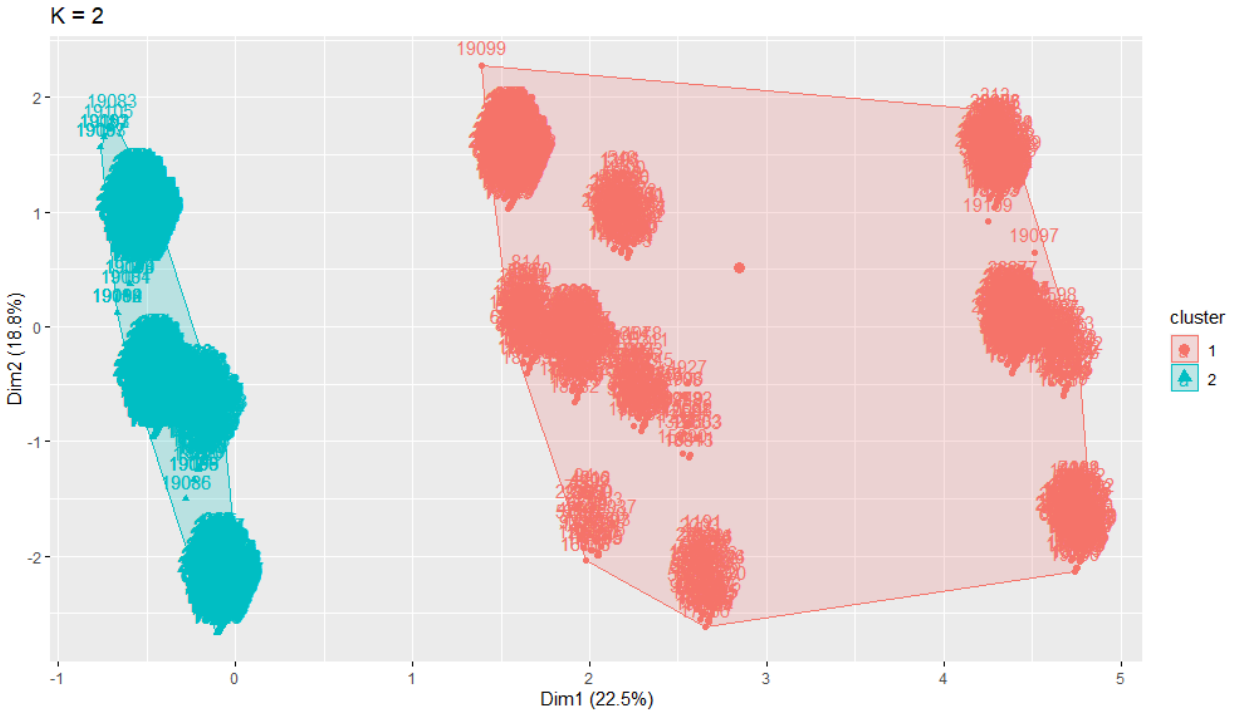

Figure 6: K = 2 Cluster Plot

Figure 7: Decision Tree Plot (Model 1)