STAT 550 Multivariate Statistical Analysis

Take-home Final

Andres Gonzalez

December 14, 2023

California State University - Long Beach
Departement of Mathematics and Statistics
1250 Bellflower Boulevard
Long Beach, California 90840

Part I

Introduction

This report examines the National Track Records for Women and Men, a topic covered in Chapter 8 of "Applied Multivariate Statistical Analysis" by Johnson and Wichern. These records, which show how fast athletes run different distances, are usually given in time – seconds for short races and minutes for longer ones like marathons. Our goal is to convert these times into a single, easy-to-understand measure: speed in meters per second. This makes it simpler to compare performances across all events and between women and men.

We've made the conversion process from time to speed straightforward and present our findings in a way that's easy to follow. By turning time records into speeds, we provide a new way to look at these athletic achievements, making them more relatable and interesting for everyone.

Methodology

In this analysis, we transformed the track event times from the National Track Records into a standardized speed metric, measured in meters per second. For shorter events (up to 400m), the original times, recorded in seconds, were directly used in the calculation. However, for longer distances (800m to the marathon), where times were noted in minutes, we first converted these into seconds by multiplying by 60. This conversion ensured uniformity in our calculations, allowing us to accurately compare performances across all events. Applying the formula Speed = Distance / Time, we then calculated the speeds for each event, providing a clear, consistent basis for comparison between the various distances and between women's and men's records.

Questions

(a) For each data, cluster the countries using both hierarchical (Single linkage and Complete linkage) and nonhierarchical (K-means) methods. Use the Euclidean distances as measures of (dis)similarity. Carefully characterize each resulted cluster. Compare the results from two data sets.

(b) Using your choice of data set (either Men or Women) perform Canonical Discrimination Analysis for the clusters (groups) found in (a). Carefully interpret the canonical variables and give a complete DA using the canonical variables. Plot the observations in the space of the canonical variable(s) and label the points according to the cluster to which they were assigned.

Cluster Analysis

Firstly, we focus on clustering the countries in our dataset, applying both hierarchical (Single linkage and Complete linkage) and non-hierarchical (K-means) methods. By utilizing Euclidean distances as a measure of similarity, we aim to group countries based on their track records, characterizing the unique attributes of each cluster. This approach allows us to draw comparisons between the clusters formed in the two datasets (Men and Women), offering insights into patterns and similarities across genders.

Women Track Records

Single Linkage Hierarchical:

This technique evaluates the distance between clusters as the minimum distance between any single point in one cluster and any single point in the other cluster, effectively chaining together nearby clusters.

Figure 1 presents the dendrogram resulting from this analysis. Each country's track record is represented by a leaf on the tree, and the 'height' at which two leaves merge represents the distance between clusters. It allows us to visualize the similarities between countries' track performances at a glance, revealing which countries have closely related track record profiles and which stand apart. For example, Western Samoa and Cook Islands, which are joined at a very low height, have more similar track performance profiles compared to the Dominican Republic, which joins at a higher point, indicating a less similar performance profile relative to the rest.

Complete Linkage Hierarchical:

This dendrogram (Figure 2) represents the hierarchical clustering of countries using the complete linkage method, which considers the maximum distance between any member of one cluster and any member of another cluster. This method typically results in more evenly sized and well-separated clusters.

The clusters formed here can provide valuable insights into which countries have similar athletic performance profiles and which are markedly different. It can also suggest potential groupings for more detailed analysis.

For instance, we observe Western Samoa and Cook Islands forming a cluster at a relatively low height, indicating a closer similarity in track performance. On the other hand, countries like the Dominican Republic, which join at higher heights, exhibit track records that are less similar to those of other countries, suggesting unique performance characteristics within this dataset.

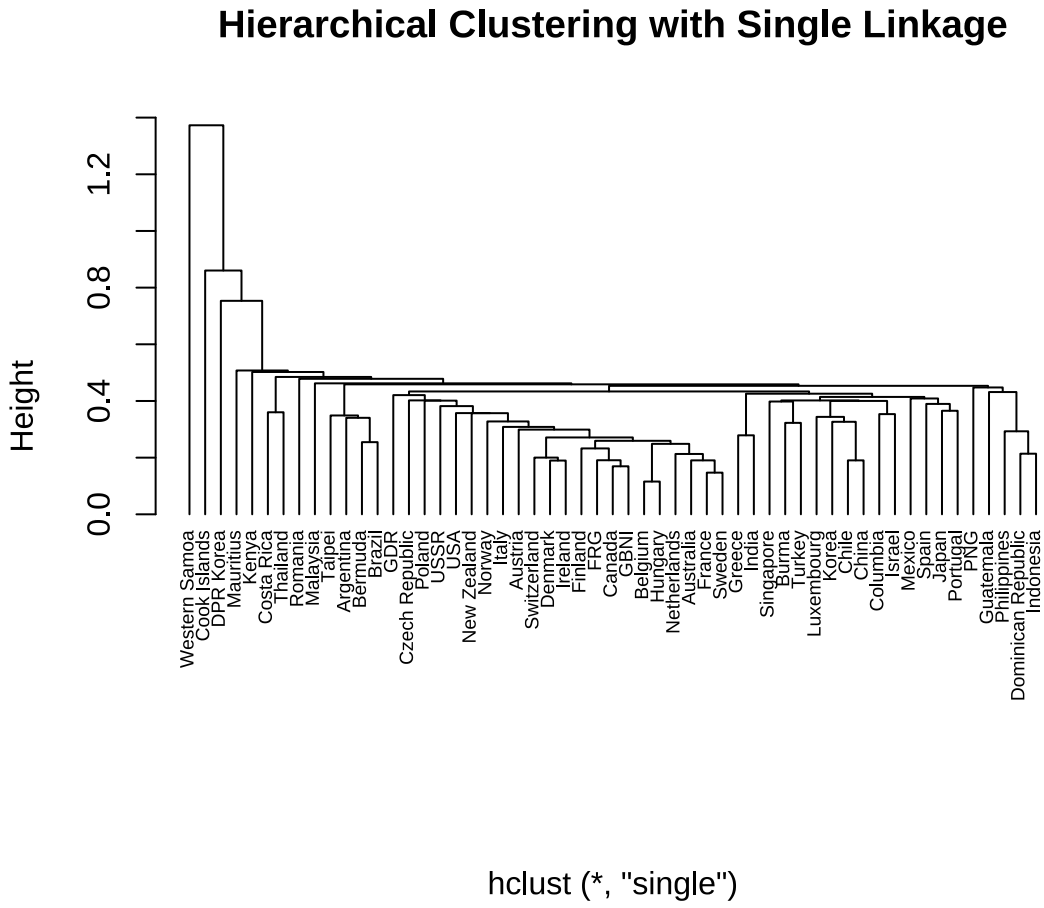

Figure 1: Dendrogram of Women's Track Records using Single Linkage Hierarchical Clustering

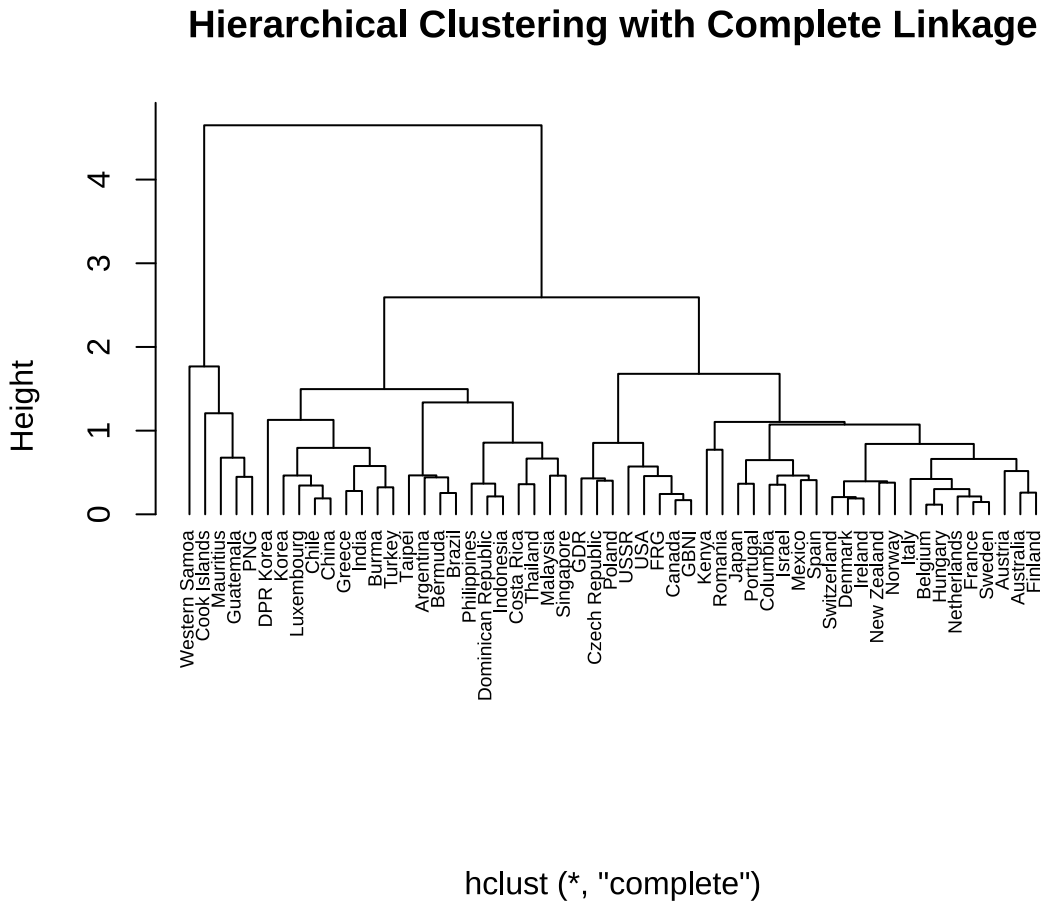

Figure 2: Dendrogram of Women's Track Records using Complete Linkage Hierarchical Clustering

K-means Nonhierarchical:

The K-means clustering algorithm was employed to segment the countries into groups based on their Women's Track Records. This nonhierarchical method organizes the data into clusters such that each point belongs to the cluster with the nearest mean, which is calculated as the average of the points within the cluster.

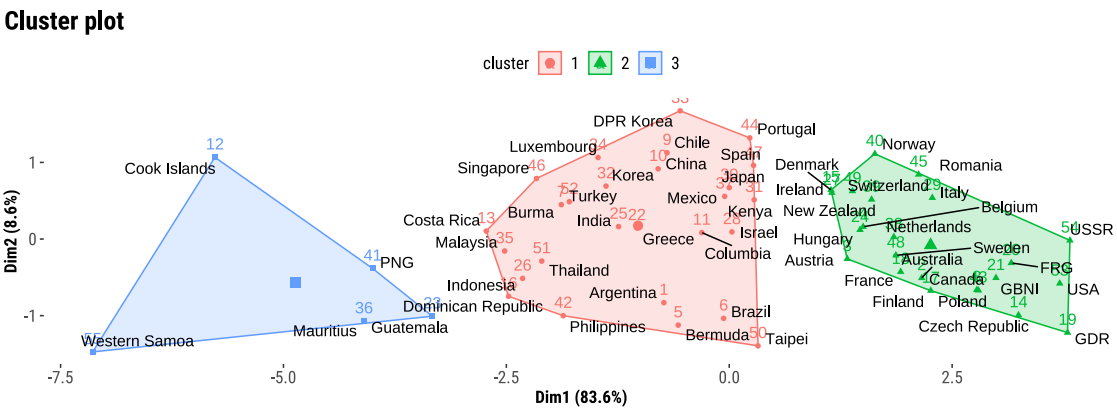

Figure 3: K-means Nonhierarchical Clustering of Women's Track Records

Figure 3 displays the clusters formed by the K-means algorithm. The two-dimensional plot, with axes likely representing the principal components of the dataset, reveals the distribution of countries into three distinct clusters. Each cluster is denoted by a unique color and contains countries whose track performance metrics are similar to each other and different from those in other clusters. The spread of the clusters across the principal components reflects the diversity in performance data.

For example, Western Samoa and Cook Islands are positioned closely within the same cluster, suggesting similarity in their track performance characteristics. In contrast, the Dominican Republic, located near the center of the plot, may represent average performance characteristics relative to the global dataset.

Men Track Records

Single Linkage Hierarchical:

Now, we applied single linkage hierarchical clustering to the Men's Track Records to identify how countries cluster based on their athletic performance data.

As shown in Figure 4, the dendrogram for the men's data replicates the 'chaining effect' seen in the women's data, where countries are joined together based on the closest distances between their records. The similarity in the structure of the men's and women's dendrograms suggests a consistency in track performance across genders that may be

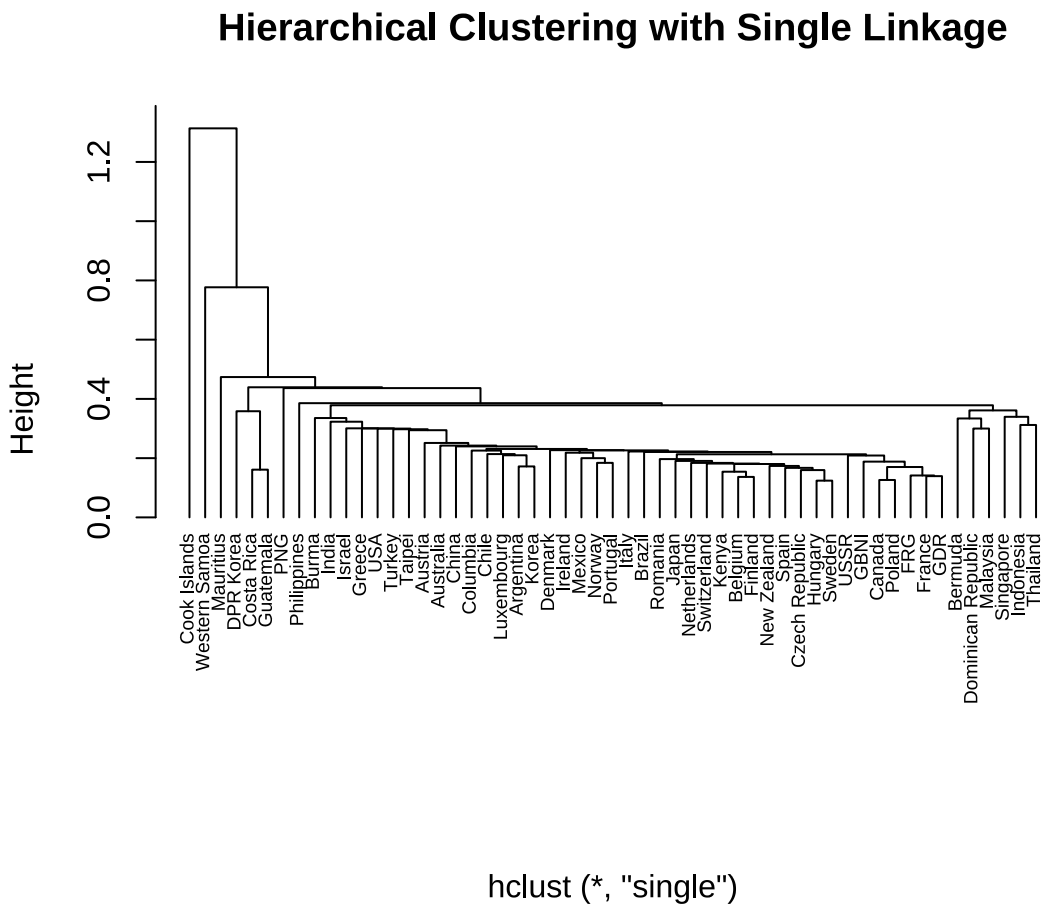

Figure 4: Dendrogram of Hierarchical Clustering with Single Linkage for men

influenced by factors such as national athletic policies, training standards, or geographic and cultural elements. The consistency between the men's and women's results may also provide an opportunity to delve into a deeper comparative analysis.

Complete Linkage Hierarchical:

In a similar process to the Women's Track Records, the Men's Track Records were subjected to complete linkage hierarchical clustering.

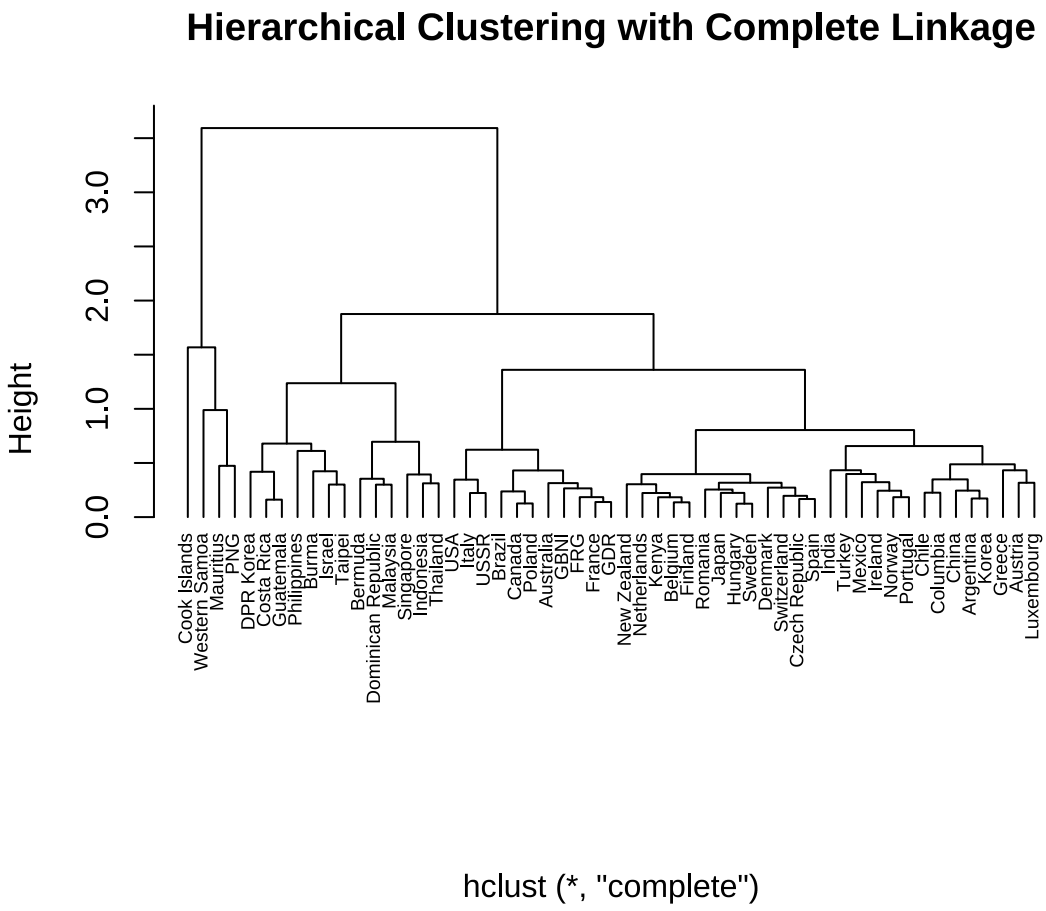

Figure 5: Dendrogram resulting from hierarchical clustering using complete linkage for men

The dendrogram presented in Figure 5 indicates how the countries are clustered based on the complete linkage criterion. The structure of the dendrogram reveals tight clusters where countries are closely related in terms of track performance, as well as more isolated countries that stand apart from the rest.

The similarity in the clustering patterns of the Men's Track Records to those of the Women's could suggest overarching trends in athletics that transcend gender, potentially pointing to broader national characteristics that impact track performance.

K-means Nonhierarchical:

As depicted in Figure 6, the cluster plot maps each country to a two-dimensional space defined by the first and second principal components, which explain a significant portion of the variance within the dataset. The countries are color-coded based on the cluster to which they are assigned, illustrating the grouping determined by the K-means algorithm.

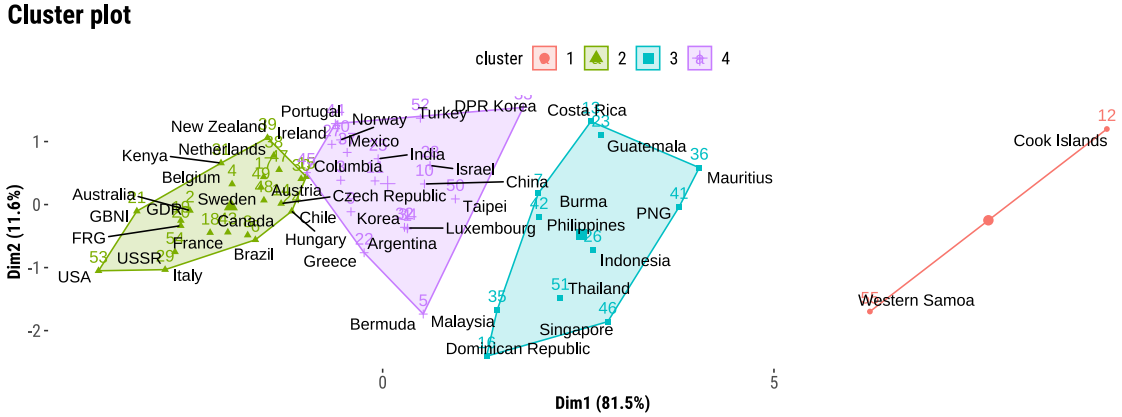

Figure 6: K-means Nonhierarchical Clustering of Men's Track Records

Based on Figure 6, countries such as Cook Islands and Western Samoa, shown at the periphery of the plot, may have unique performance profiles that starkly differ from the global trend.

"Dim1" (horizontal axis) accounts for 81.5% of the variance, while "Dim2" (vertical axis) explains 11.6%, signaling that 'Dim1' captures the bulk of the performance characteristics.

This clustering analysis provides insights into the similarities and differences in track performances among the countries and can be particularly revealing when compared to the clusters formed within the Women's Track Records.

Compare Results

Certain countries consistently appeared in similar clusters across both genders, which may indicate a universal strength or strategy in track and field within those nations.

Other countries showed distinct differences in their clustering between the men's and women's datasets, possibly reflecting gender-specific advantages or focus areas in their athletic training programs.

Canonical Discriminant Analysis

```
Call:
lda(Cluster ~ ., data = women_speed)

Prior probabilities of groups:
         1          2          3
0.49090909 0.41818182 0.09090909

Group means:
  Speed_100m Speed_200m Speed_400m Speed_800m Speed_1500m Speed_3000m
1   8.460229   8.306625   7.339998   6.302166    5.671996    5.183541
2   8.910258   8.816282   7.839236   6.728489    6.147980    5.683395
3   8.142769   7.843824   6.894379   5.838722    4.999349    4.469424
  Speed_Marathon
1       4.018826
2       4.595043
3       2.858931

Coefficients of linear discriminants:
                      LD1        LD2
Speed_100m     -2.1724127 -6.6742129
Speed_200m      3.7042128  2.7967028
Speed_400m     -0.4844898  0.6372165
Speed_800m      3.0081195 -4.0116979
Speed_1500m    -1.8572968  3.3058866
Speed_3000m     2.8398472 -1.8264413
Speed_Marathon  1.6441365  2.6067326

Proportion of trace:
   LD1    LD2
0.9591 0.0409
```

In the application of Canonical Discriminant Analysis to the cluster assignments derived from the women's speed data, we observe a pronounced distinction among the groups when analyzed through the lens of canonical variables, LD1 and LD2. The prior probabilities of the groups—approximately 49% for Cluster 1, 42% for Cluster 2, and 9% for Cluster 3—lay the groundwork for our expectations regarding group prevalence.

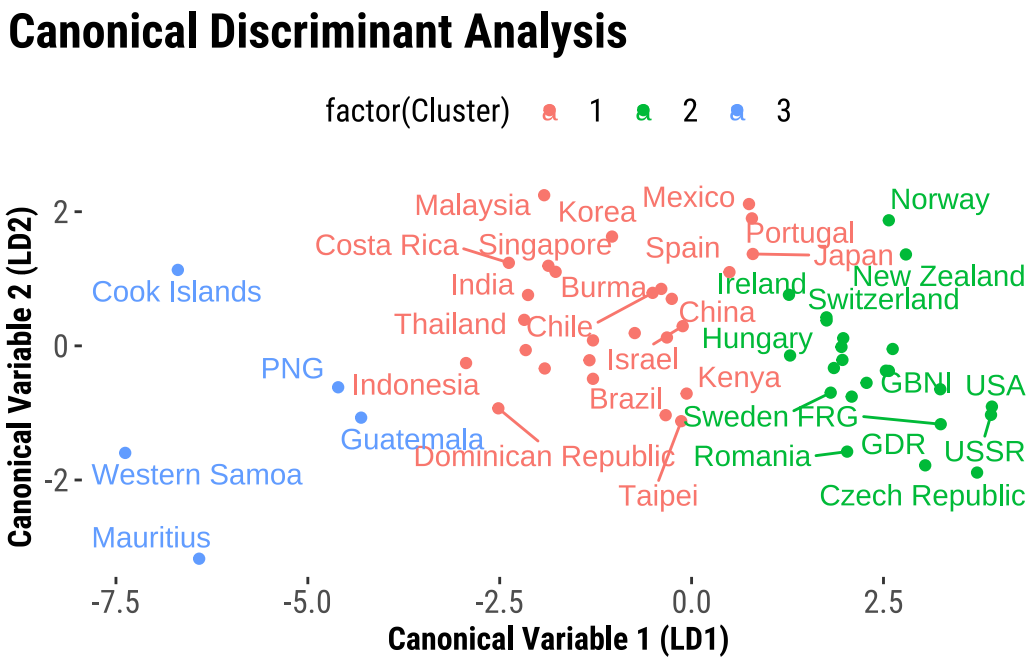

Figure 7: Canonical Discriminant Analysis of women's speed event data, with countries plotted according to their scores on the first two discriminant functions

The group means highlight distinct performance profiles across the speed events, with Cluster 1 exhibiting superior endurance as evidenced by faster marathon times, while Cluster 2 showcases consistently quicker times across all events, signaling a well-rounded athletic prowess.

The canonical coefficients reveal the relative influence of each event on the discriminant functions. LD1, which explains a substantial 95.91% of the variance between the groups, seems to encapsulate an endurance-sprint spectrum. Higher speeds in the marathon, 3000m, 800m, and 200m events draw countries towards the positive side of LD1, while quicker 100m, 400m, and 1500m times pull them towards the negative. This suggests that LD1 may be viewed as an axis contrasting sprint-oriented prowess with endurance-focused capability.

Conversely, LD2 accounts for a mere 4.09% of the group variance, yet it provides nuanced differentiation where shorter events like the 100m dash and 800m run inversely affect this discriminant, and longer events such as the 1500m and Marathon positively contribute to it.

Figure 7 illustrates the segregation of countries based on sprint and endurance capabilities along LD1, while LD2 further refines the distinction in intermediate athletic events. Notably, outliers such as the Cook Islands and Western Samoa reveal distinctive performance profiles, highlighting the global diversity of athletic talent.

Part II

Introduction

In this section of our analysis, we turn our attention to the US Crime Data, a comprehensive set of statistics reflecting the reported crime rates across the 50 states in 1985. This dataset encapsulates various dimensions of crime, from violent offenses such as murder and assault to property crimes like burglary and auto theft, contextualized by state-specific demographic information including land area and population. By examining these variables, we aim to uncover underlying patterns and regional trends in criminal activity, providing a quantitative foundation for understanding the distribution of crime across the United States. The insights garnered from this analysis hold the potential to inform policy decisions, focus crime prevention efforts, and enhance the allocation of resources for law enforcement.

Methodology

The dataset comprises 12 variables, ranging from land area and population to various crime categories, alongside regional classifications. Given the differences in state populations, crime figures were normalized per 100,000 inhabitants to calculate standardized crime rates. While the land area (X2) and population (X3) are included to provide context to the crime figures, identifiers such as the state name (X1), region (X11), and division (X12) are excluded to focus on the statistical relationships and patterns inherent in the crime data itself.

Questions

(a) Perform a preliminary analysis via covariance and correlation matrices.

(b) Perform a complete principal component analysis. Your analysis and discussion should include (but not limited) a choice of PCA on covariance or correlation matrix, number of PCs, adequacy of the number, interpretation of the selected PCs, scatter plots of eigenvectors and PCs and discussion on possible outliers and groups, and further discussions.

(c) Perform a complete factor analysis. Your analysis and discussion should include (but not limited) a choice of the number of factors and a chi-square test for the adequacy, interpretation of the rotated and unrotated factor loadings with plots, comparison of your findings with the preliminary analysis, interpretation of the communality and specific variance, scatter plots of factor scores and discussion on possible outliers and groups, and further discussions.

Preliminary Analysis: Covariance and Correlation

In our preliminary analysis of the US Crime Data, we employed covariance analysis to gauge the extent of variation between different types of crimes, complemented by a correlation analysis that revealed the strength and direction of their relationships. These statistical tools provided key insights into how various crime rates and demographic factors interrelate, laying the groundwork for a more detailed examination of the data.

Table 1: Covariance Matrix Table

	land	popu	murd	rape	robb	assa	burg	larc	auto
land	7816047096.35	32214277.93	83176.31	244806.30	-165593.37	976515.50	2159505.31	15889029.76	1453429.39
popu	32214277.93	25694559.62	5308.59	15571.44	288099.12	147326.05	784325.16	829507.56	380330.91
murd	83176.31	5308.59	14.81	14.70	119.68	213.15	384.46	176.95	84.36
rape	244806.30	15571.44	14.70	54.00	369.52	348.61	1804.50	3132.74	646.41
robb	-165593.37	288099.12	119.68	369.52	8316.23	3501.22	20485.88	28234.76	11232.25
assa	976515.50	147326.05	213.15	348.61	3501.22	4647.11	12816.31	15324.75	4495.59
burg	2159505.31	784325.16	384.46	1804.50	20485.88	12816.31	130356.94	205309.15	50455.50
larc	15889029.76	829507.56	176.95	3132.74	28234.76	15324.75	205309.15	503857.62	78605.91
auto	1453429.39	380330.91	84.36	646.41	11232.25	4495.59	50455.50	78605.91	39843.96

Table 2: Correlation Matrix Table

	land	popu	murd	rape	robb	assa	burg	larc	auto
land	1.000	0.072	0.244	0.377	-0.021	0.162	0.068	0.253	0.082
popu	0.072	1.000	0.272	0.418	0.623	0.426	0.429	0.231	0.376
murd	0.244	0.272	1.000	0.520	0.341	0.813	0.277	0.065	0.110
rape	0.377	0.418	0.520	1.000	0.551	0.696	0.680	0.601	0.441
robb	-0.021	0.623	0.341	0.551	1.000	0.563	0.622	0.436	0.617
assa	0.162	0.426	0.813	0.696	0.563	1.000	0.521	0.317	0.330
burg	0.068	0.429	0.277	0.680	0.622	0.521	1.000	0.801	0.700
larc	0.253	0.231	0.065	0.601	0.436	0.317	0.801	1.000	0.555
auto	0.082	0.376	0.110	0.441	0.617	0.330	0.700	0.555	1.000

Principal Component Analysis (PCA)

In our PCA, we chose the correlation matrix over the covariance matrix due to the differing scales of the variables in our dataset. This approach standardizes the variance across variables, ensuring a more balanced representation in the analysis.

```
Importance of components:
```

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	2.1125	1.2120	1.0745	0.85354	0.67288	0.52669	0.47401
Proportion of Variance	0.4958	0.1632	0.1283	0.08095	0.05031	0.03082	0.02497
Cumulative Proportion	0.4958	0.6590	0.7873	0.86827	0.91857	0.94940	0.97436

	PC8	PC9
Standard deviation	0.36513	0.31212
Proportion of Variance	0.01481	0.01082
Cumulative Proportion	0.98918	1.00000

The Principal Component Analysis (PCA) of the US Crime Data resulted in nine principal components (PCs), each explaining a portion of the total variance in the data. The selection of the number of components was guided by the cumulative proportion of variance explained. The first three components (PC1, PC2, and PC3) together account for approximately 78.73% of the total variance. This significant proportion suggests that these three components are adequate to capture the majority of the information in the dataset, representing a balance between simplifying the data and retaining its key features.

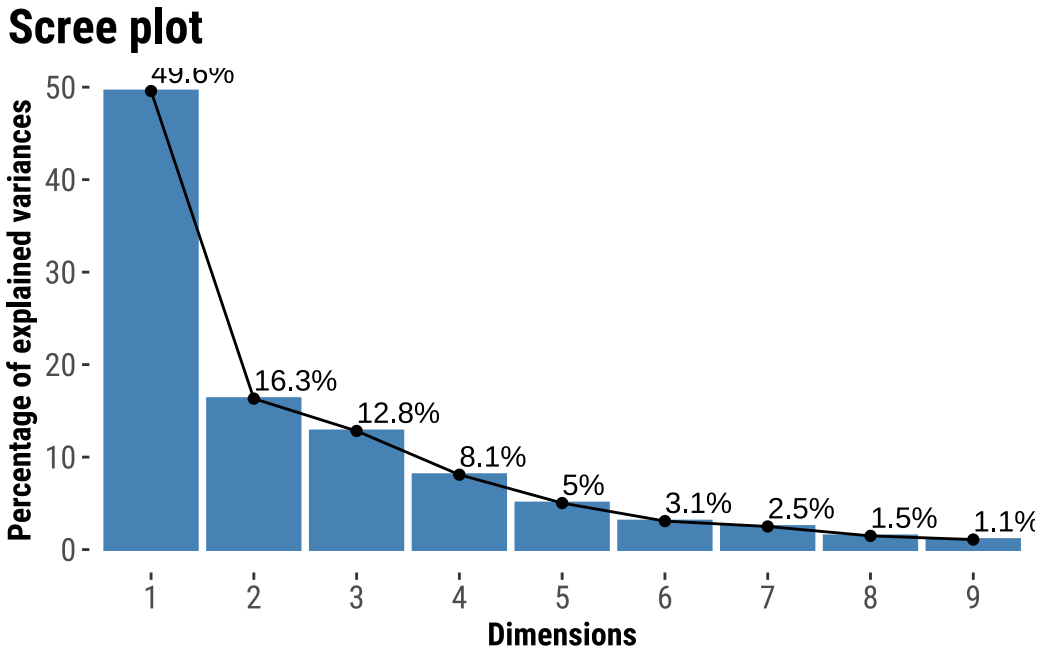

Figure 8: Elbow Plot of US Crime Data

	PC1	PC2	PC3	PC4	PC5	PC6
land	0.1239453	0.280415550	0.71559955	-0.51180272	0.21943989	-0.05266509
popu	0.2940490	-0.002926609	-0.37118074	-0.70021572	-0.34950103	0.38355074
murd	0.2655775	0.611943760	-0.10479048	0.22515812	0.18442892	0.24167319

```
rape 0.4030653  0.142736997  0.21154351  0.06870215 -0.27838202 -0.32301971
robb 0.3768923 -0.120446022 -0.34311330 -0.17088505  0.15570532 -0.72945727
assa 0.3699159  0.403917692 -0.15237811  0.24619137  0.01729497  0.04605531
burg 0.4066630 -0.279206589  0.06307445  0.24542860 -0.15096448  0.30222313
larc 0.3322011 -0.352250624  0.38630780  0.20251293 -0.33940695  0.03955487
auto 0.3325084 -0.383165722 -0.01589078 -0.03579352  0.74542202  0.24542849
             PC7         PC8         PC9
land  0.16469464  0.01029489 -0.23199602
popu -0.02756824 -0.04940562  0.11395761
murd  0.24849320  0.44363110  0.37727817
rape -0.73277594  0.12253100  0.18406840
robb  0.33890433  0.14488791 -0.06569495
assa  0.03287065 -0.70998444 -0.32955255
burg  0.05028771  0.43232065 -0.62361546
larc  0.43212933 -0.24043009  0.46261503
auto -0.26116083 -0.12590673  0.20283468
```

Interpretation of Selected PCs

- **PC1:** Being the most dominant component, PC1 explains about 49.58% of the variance. This component likely represents a general crime factor, possibly aggregating common features across various types of crime.
- **PC2:** The second component accounts for 16.32% of the variance and might capture aspects of crime that are orthogonal to those represented by PC1, perhaps distinguishing between violent and property crimes or other nuanced distinctions in crime patterns.
- **PC3:** With 12.83% variance explained, PC3 could be representing another distinct dimension of crime, potentially correlating with either socio-economic factors or specific types of crimes not encapsulated by the first two components.

Figure 9 presents a biplot of the PCA, where states are distributed according to their crime profiles against the first two principal components, revealing clusters and potential outliers. For instance, states like Mississippi (MS) are positioned distinctly, suggesting unique crime characteristics. The vectors for crime types, particularly “murd” and”assa”, indicate a strong positive correlation. Whereas, “burg” and “larc”, indicate a strong negative correlation amongst the variables, reflecting their significant contribution to the principal components.

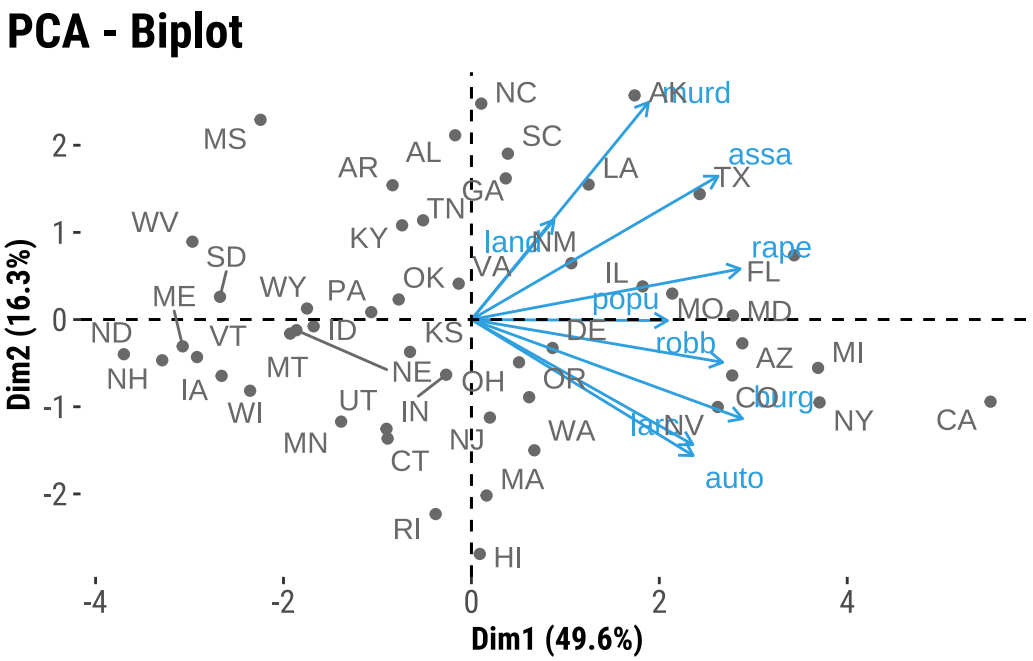

Figure 9: PCA Biplot of US Crime Data

Factor Analysis

This technique is designed to identify latent factors that can explain the patterns of correlation observed among various crime statistics. Factor analysis simplifies the data by reducing the number of variables into a smaller set of interpretable factors.

Unrotated Loadings:

```
Call:
factanal(x = crime_data_std, factors = 3, scores = "regression",     rotation = "none")

Uniquenesses:
 land  popu  murd  rape  robb  assa  burg  larc  auto
0.798 0.547 0.158 0.299 0.228 0.118 0.165 0.083 0.414

Loadings:
     Factor1 Factor2 Factor3
land  0.236          -0.382
popu  0.494   0.141   0.434
murd  0.530   0.727  -0.181
rape  0.823   0.113  -0.108
```

```
robb  0.714           0.507
assa  0.759   0.551
burg  0.874  -0.252
larc  0.783  -0.522  -0.179
auto  0.642  -0.244   0.338

                Factor1 Factor2 Factor3
SS loadings       4.134   1.266   0.792
Proportion Var    0.459   0.141   0.088
Cumulative Var    0.459   0.600   0.688

Test of the hypothesis that 3 factors are sufficient.
The chi square statistic is 19.31 on 12 degrees of freedom.
The p-value is 0.0814
```

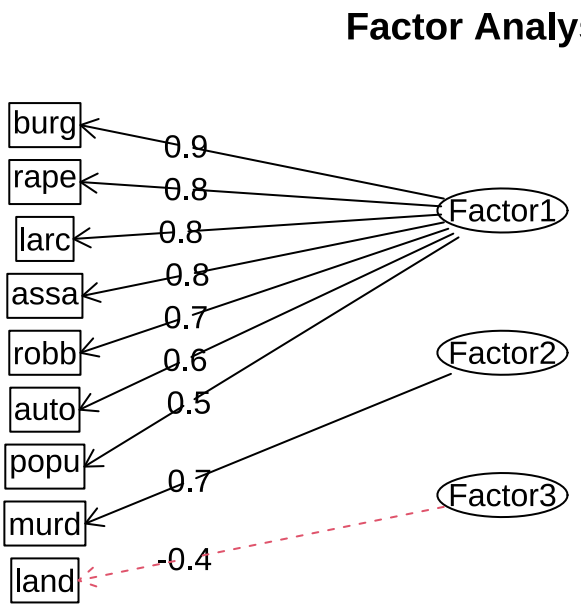

Figure 10: Unrotated Loadings for US Crime Data

The sum of squared loadings (SS loadings) for each factor demonstrates the amount of variance explained, with Factor1 explaining 45.9%, Factor2 contributing 14.1%, and Factor3 accounting for 8.8% of the total variance. Cumulatively, these factors explain 68.8% of the variance in the crime data, with the chi-square test for the hypothesis that three factors are sufficient yielding a p-value of 0.0814.

Rotated Loadings:

```
Call:
```

```
factanal(x = crime_data_std, factors = 3, scores = "regression",     rotation = "varimax"

Uniquenesses:
 land  popu  murd  rape  robb  assa  burg  larc  auto
0.798 0.547 0.158 0.299 0.228 0.118 0.165 0.083 0.414

Loadings:
     Factor1 Factor2 Factor3
land -0.169   0.253   0.332
popu  0.635   0.221
murd  0.135   0.907
rape  0.395   0.519   0.526
robb  0.827   0.249   0.162
assa  0.391   0.829   0.203
burg  0.601   0.183   0.664
larc  0.343           0.894
auto  0.663           0.383

                Factor1 Factor2 Factor3
SS loadings       2.362   1.988   1.841
Proportion Var    0.262   0.221   0.205
Cumulative Var    0.262   0.483   0.688

Test of the hypothesis that 3 factors are sufficient.
The chi square statistic is 19.31 on 12 degrees of freedom.
The p-value is 0.0814
```

The sum of squared loadings for Factor1 now accounts for 26.2% of the variance, Factor2 for 22.1%, and Factor3 for 20.5%. Collectively, these factors explain approximately 68.8% of the total variance in the dataset. The chi-square test for the hypothesis that three factors are sufficient remains at a p-value of 0.0814, supporting the adequacy of the three-factor solution, although with marginal certainty given the p-value is slightly above the conventional threshold for significance.

Interpretation of the Communality and Specific Variance:

The communalities from our factor analysis indicate that variables such as murder, assault, burglary, and larceny are largely explained by the underlying factors, reflecting that these aspects of crime data share common patterns. However, the specific variances for land and population size are comparatively high, suggesting unique influences on these variables not captured by the extracted factors.

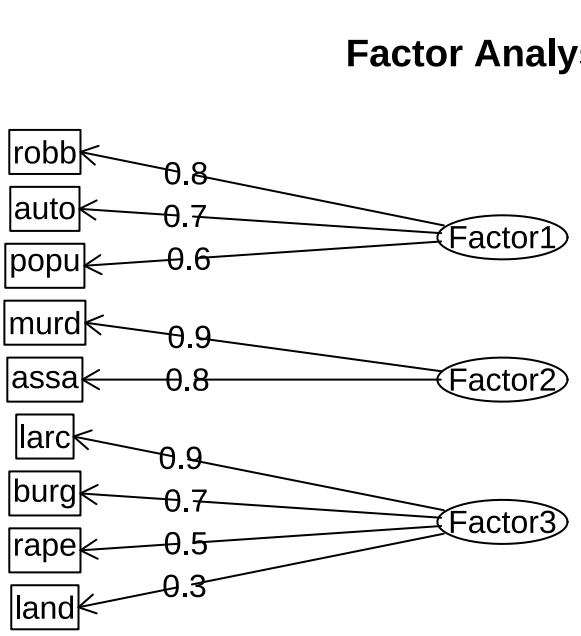

Figure 11: Rotated Loadings of US Crime Data

Table 3: Communalities and Specific Variances Table

	Communalities	Specific Variances
land	0.7975117	0.2024883
popu	0.5473288	0.4526712
murd	0.1580176	0.8419824
rape	0.2985404	0.7014596
robb	0.2278475	0.7721525
assa	0.1181258	0.8818742
burg	0.1648382	0.8351618
larc	0.0825234	0.9174766
auto	0.4137429	0.5862571

Scatter plots of factor scores for m = 3:

Based on Figure 12:

Factor 1 vs Factor 2: The position of states such as California (CA) and Arizona (AZ) with higher scores on Factor 3 suggests a higher prevalence of the types of crimes encapsulated by this factor.

Factor 1 vs Factor 3: Potential outliers such as California (CA), which appears far along Factor 2, may indicate a disproportionately high rate of the violent crimes that

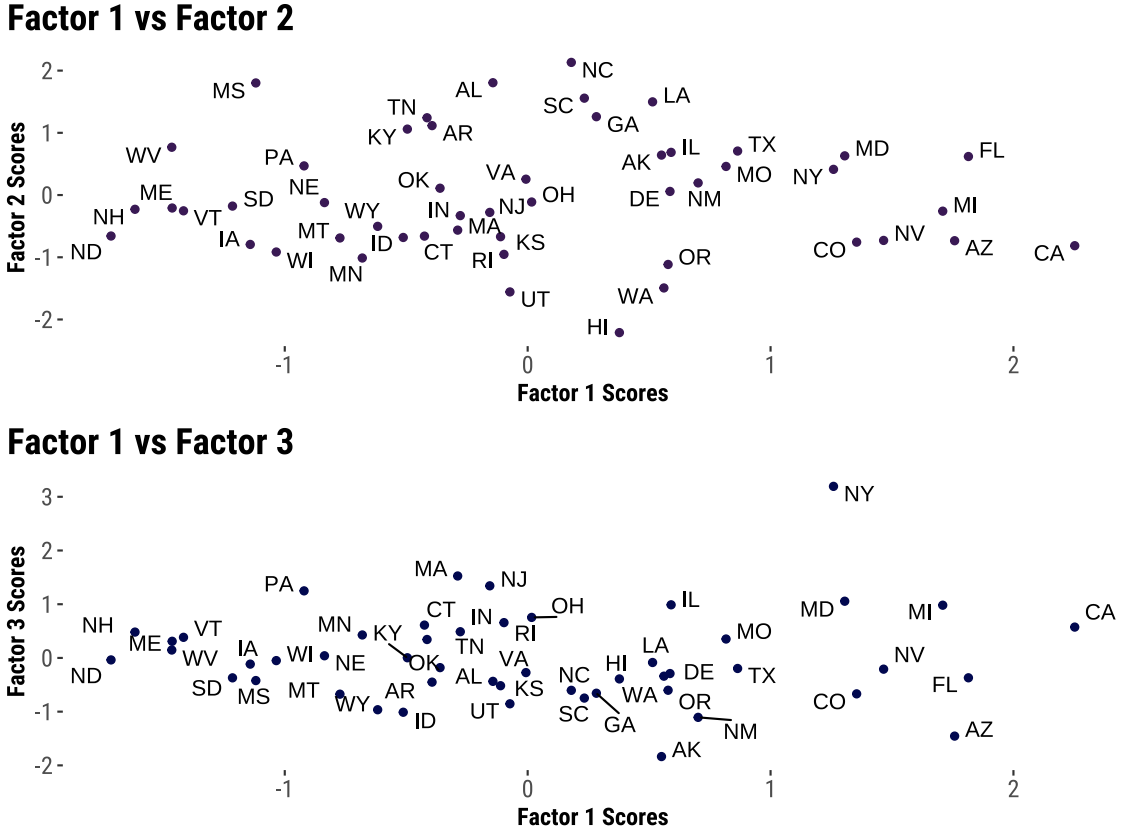

Figure 12: Scatter Plot of States by Factor 1 vs Factor 2 and Factor 1 vs Factor 3

Factor 2 represents.