Multivariate Analysis

US OPEN - 2013

Papineni Anusha

**Introduction:**

The datasets selected for this project are based on Tennis major tournaments match statistics. The datasets contained separate excel files for both Men and Women Tennis players match statistics in US-Open 2013. Our motivation for selecting this dataset to perform the multivariate analysis is to study how men tennis player performances differ from women tennis players in a major tennis tournament and to study which attributes are correlated and significant to determine match win in men or women tennis match.

These two data files have the same 38 attributes comprising of numerical, and categorical data. The total number of rows is equal to the number of matches that occurred in the US open 2013 tournament. Each row contains information on a single match and the two players in the match.

The attributes name in the dataset are abbreviated. Following are the attribute description for the columns used in the analysis:

1. Player 1: Name of Player
2. FSP.1: First Serve Percentage for player (Real Number)
3. FSW.1: First Serve Won by player (Real Number)
4. SSP.1: Second Serve Percentage for player (Real Number)
5. SSW.1: Second Serve Won by player (Real Number)
6. ACE.1: Aces won by player (Numeric-Integer)
7. DBF.1: Double Faults committed by player (Numeric-Integer)
8. BPC.1: Break Points Created by player (Numeric)
9. BPW.1: Break Points Won by player (Numeric)
10. ST1.1: Set 1 result for player (Numeric-Integer)
11. ST2.1: Set 2 result for player (Numeric-Integer)

**Objective:**

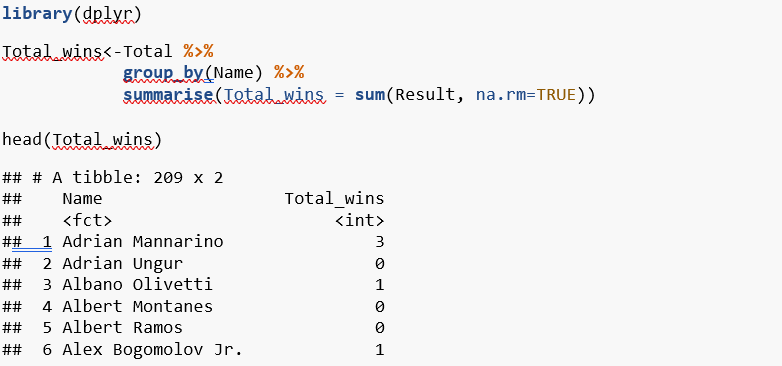
The aim of analysis in this project is to:

1. Find difference in men and women style of play in Tennis
2. Capture latent variable that affect the results in a match
3. Compare players’ performances
4. Find key variables that explains the results in the data

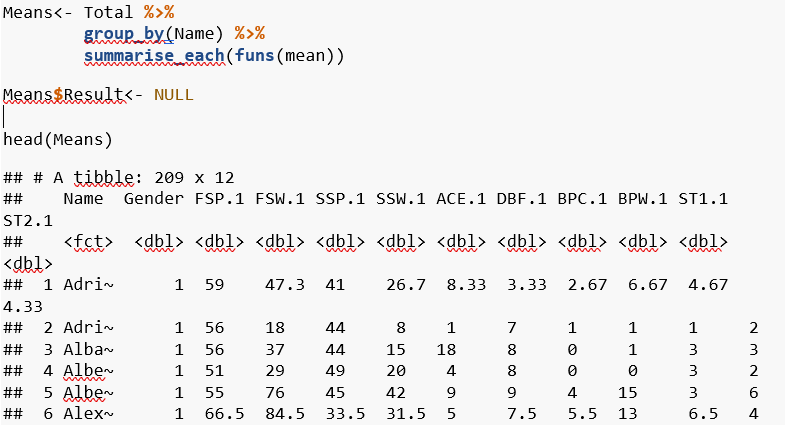
**Cleaning:**

The original datasets contained observations pertained to each match and contained scores of both the players of a match in a single row. To perform analysis on these datasets few data cleaning activities are necessary. Each row in the data set must be split into two rows each representing the information of one player. For analysis purpose only few variables which are common to both the datasets are considered and the columns had to be renamed to be consistent in both the sets. The rows in both the data sets are then combined using rbind() function. A new column of “Total\_wins” is calculated using “dplyr” package which represents the total number of matches for each player.

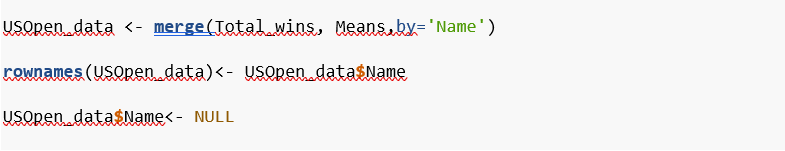
The package “dplyr” must be installed before using it.



The scores of each player are then aggregated by calculating the mean of all the matches played by that player.



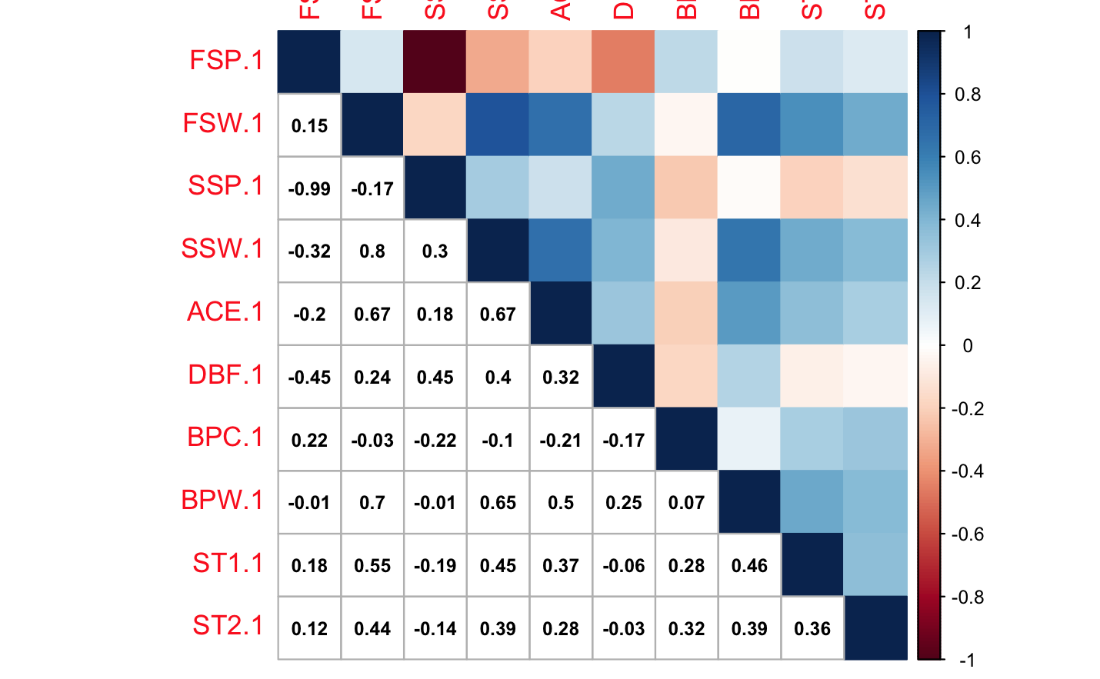
Both above calculated data frames are now merged by the column ’Name’ then the same column ’Name’ is made as the index for the merged data frame.



**Correlation matrix:**

A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. A correlation matrix is used to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses.





The above correlation matrix is the correlation matrix of this project. It can be observed that the following variables are negatively correlated:

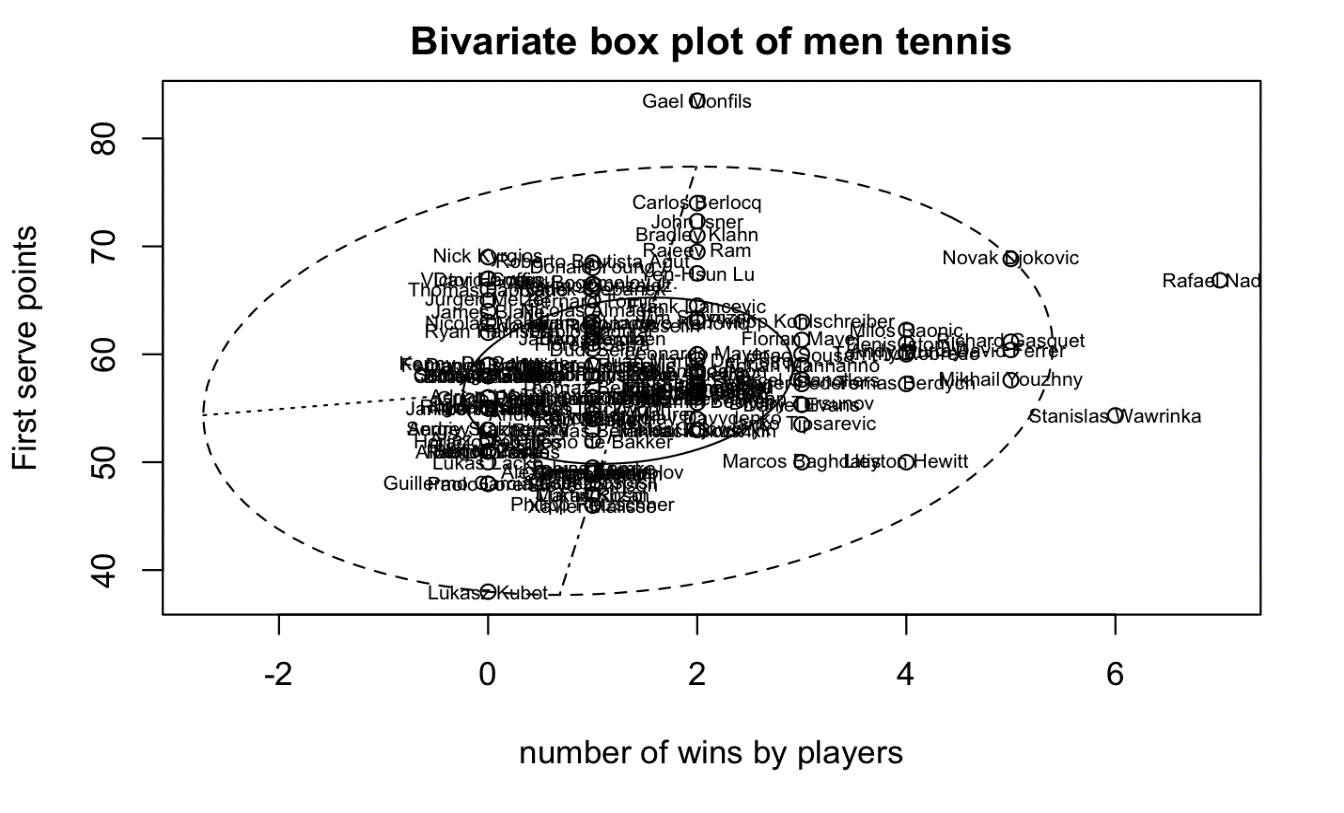
* FSP (first serve percentage) and SSP (second serve percentage)
* DBF (double Faults) and FSP (first serve percentage)

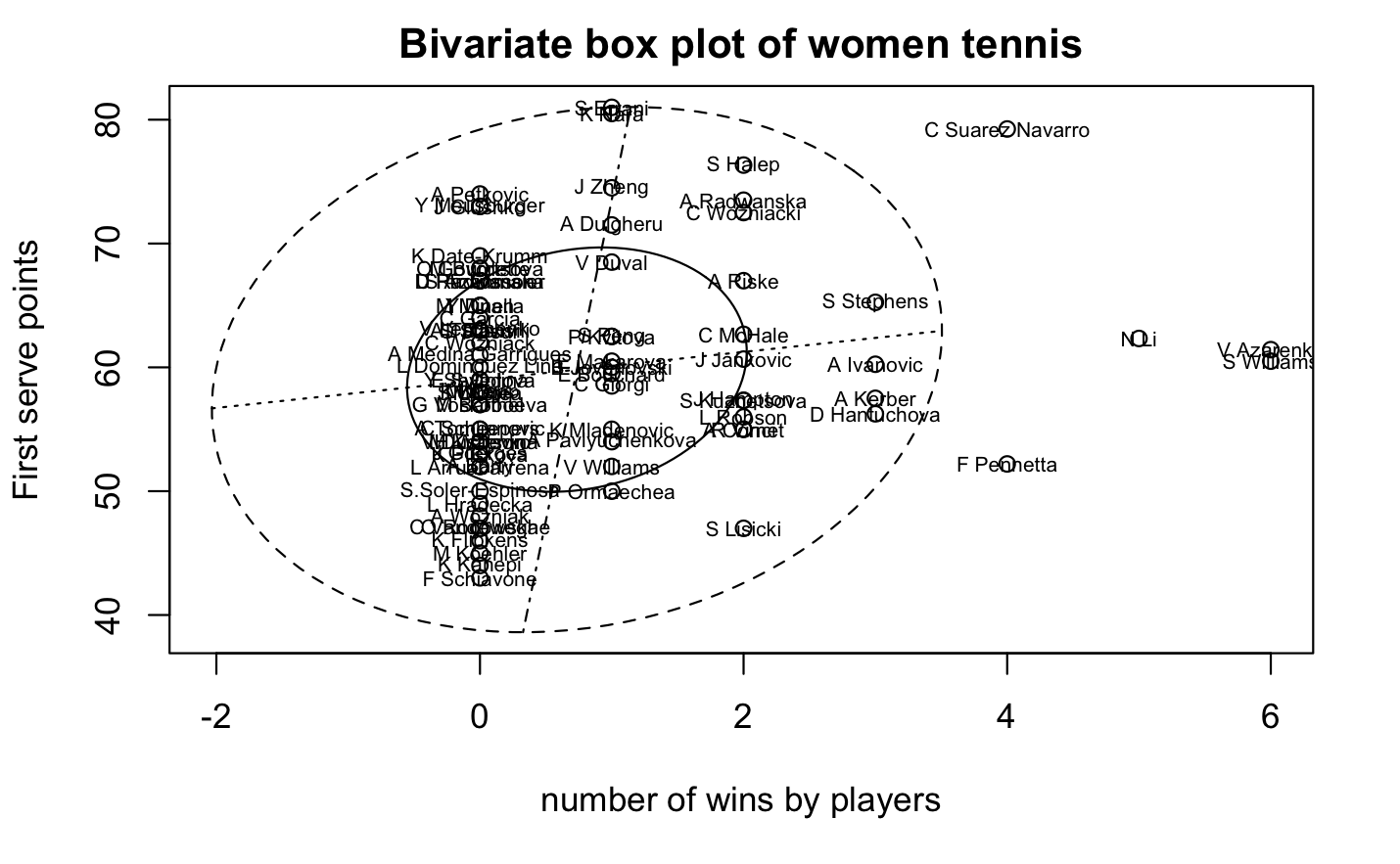
The following variable are positively correlated:

* ACE and SSW (second serve win)
* ACE and FSW (first serve win)
* BPW (break point wins) and SSW (second serve win)

**Visualizations**

**Bivariate Box Plot**

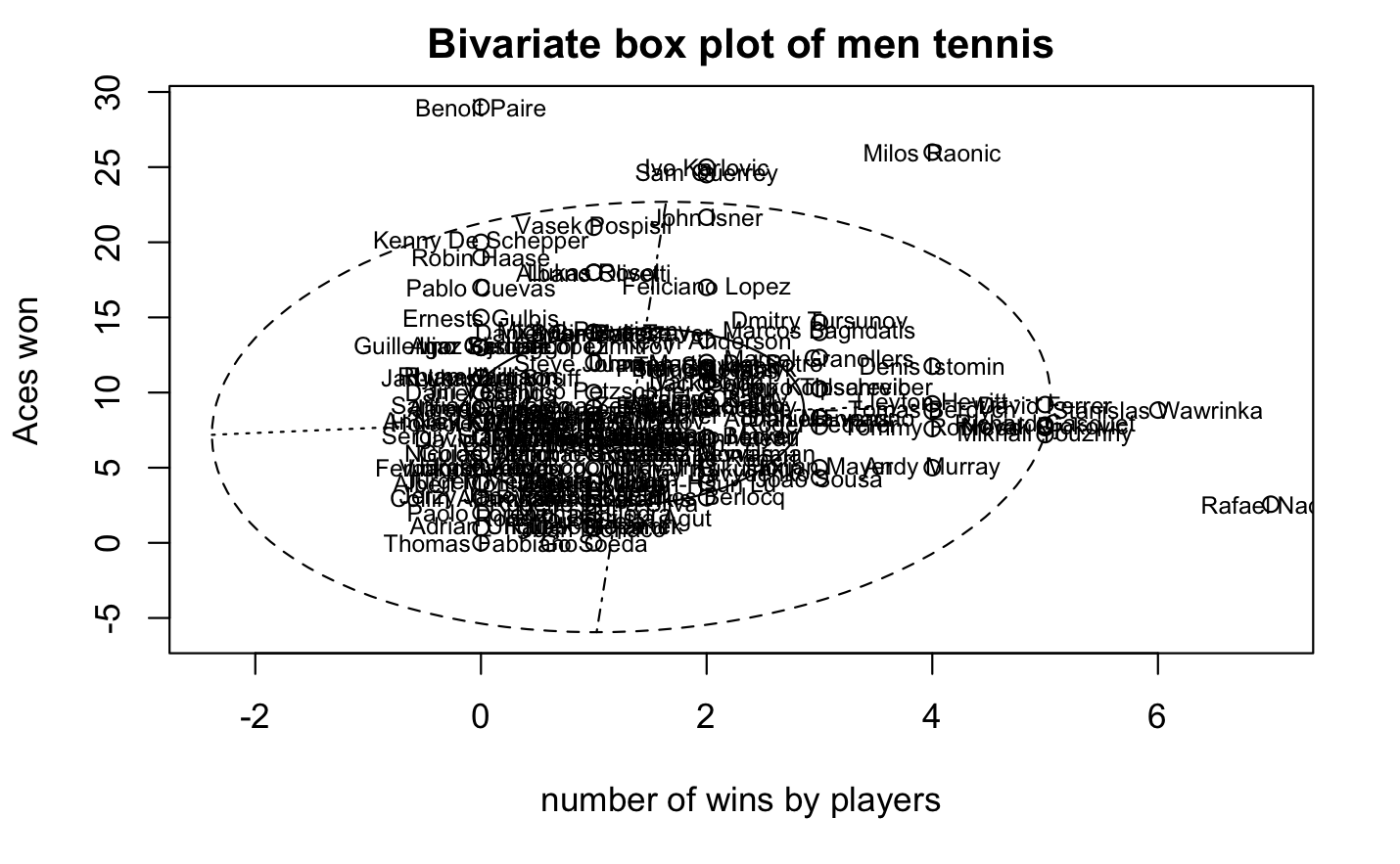


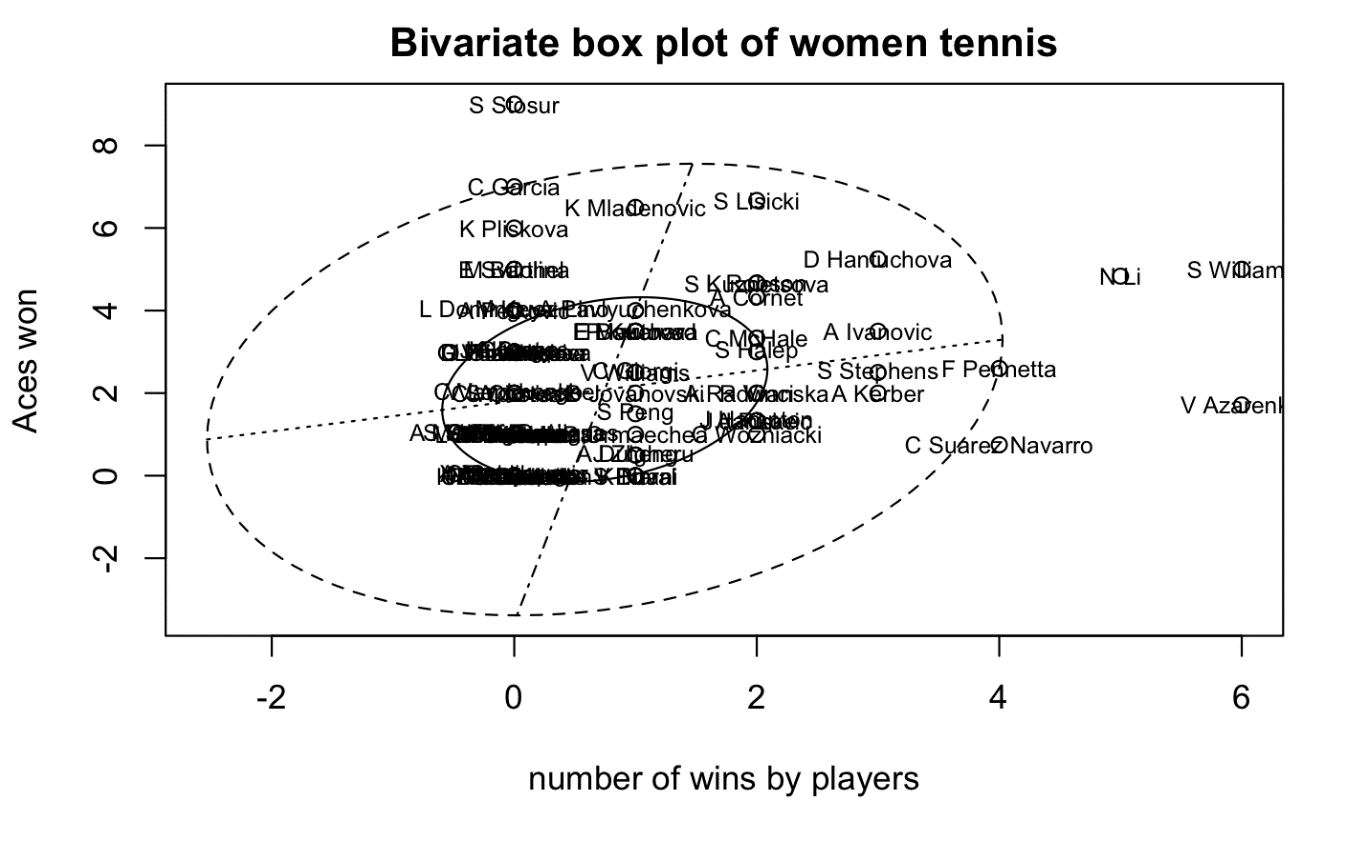


Bivariate boxplots Creates diagnostic bivariate quelplot ellipses (bivariate boxplots) using the method of Goldberg and Iglewicz (1992). The output can be used to check assumptions of bivariate normality and to identify multivariate outliers.

In the above bivariate box plot for men tennis players, Rafael Nadal and Stanislas Wawrinka are the two players which are outliers in this data, as they have won maximum matches where they also had maximum first serve wins

In the above bivariate box plot for women tennis players, Victoria Azarenka and Serena Williams, and N Li are the three players which are outliers in this data, as they have won maximum matches where they also had maximum first serve wins.

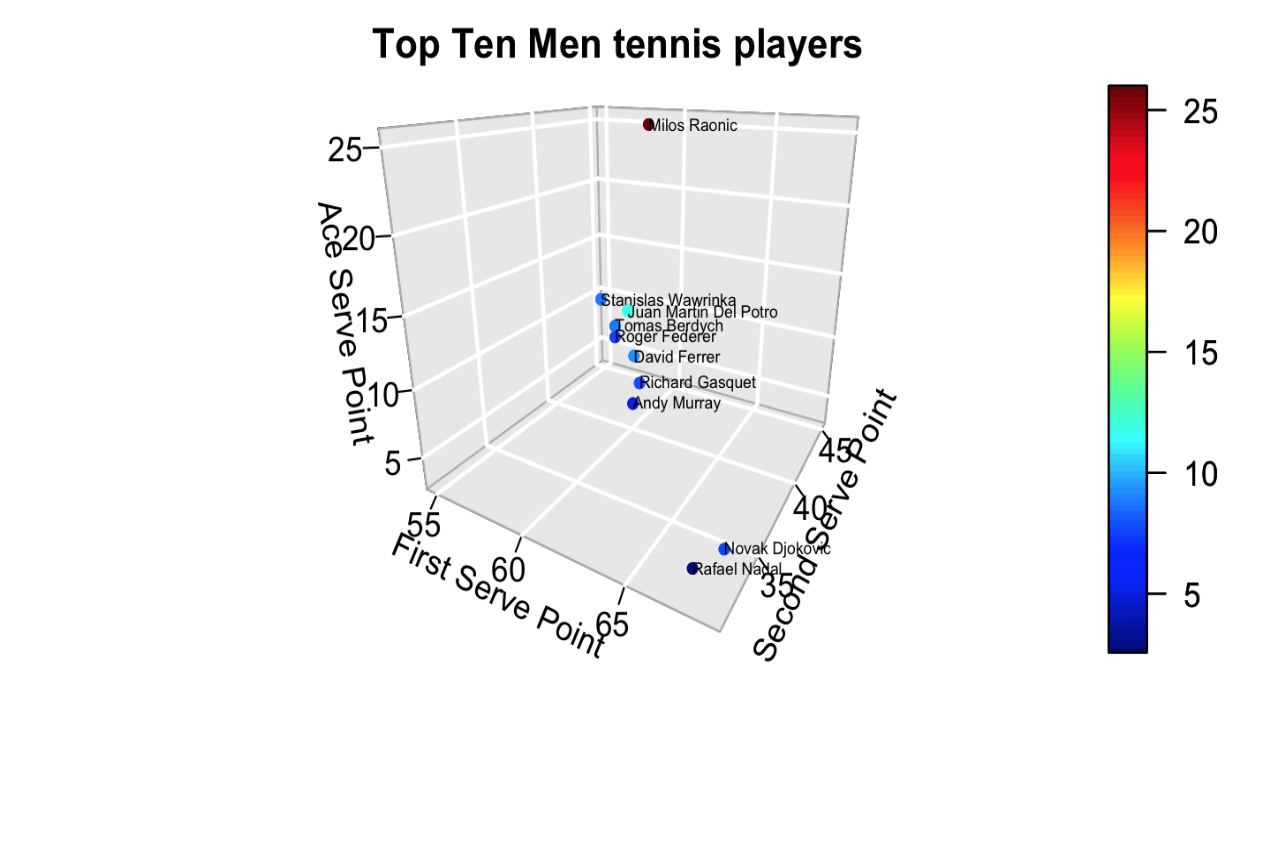


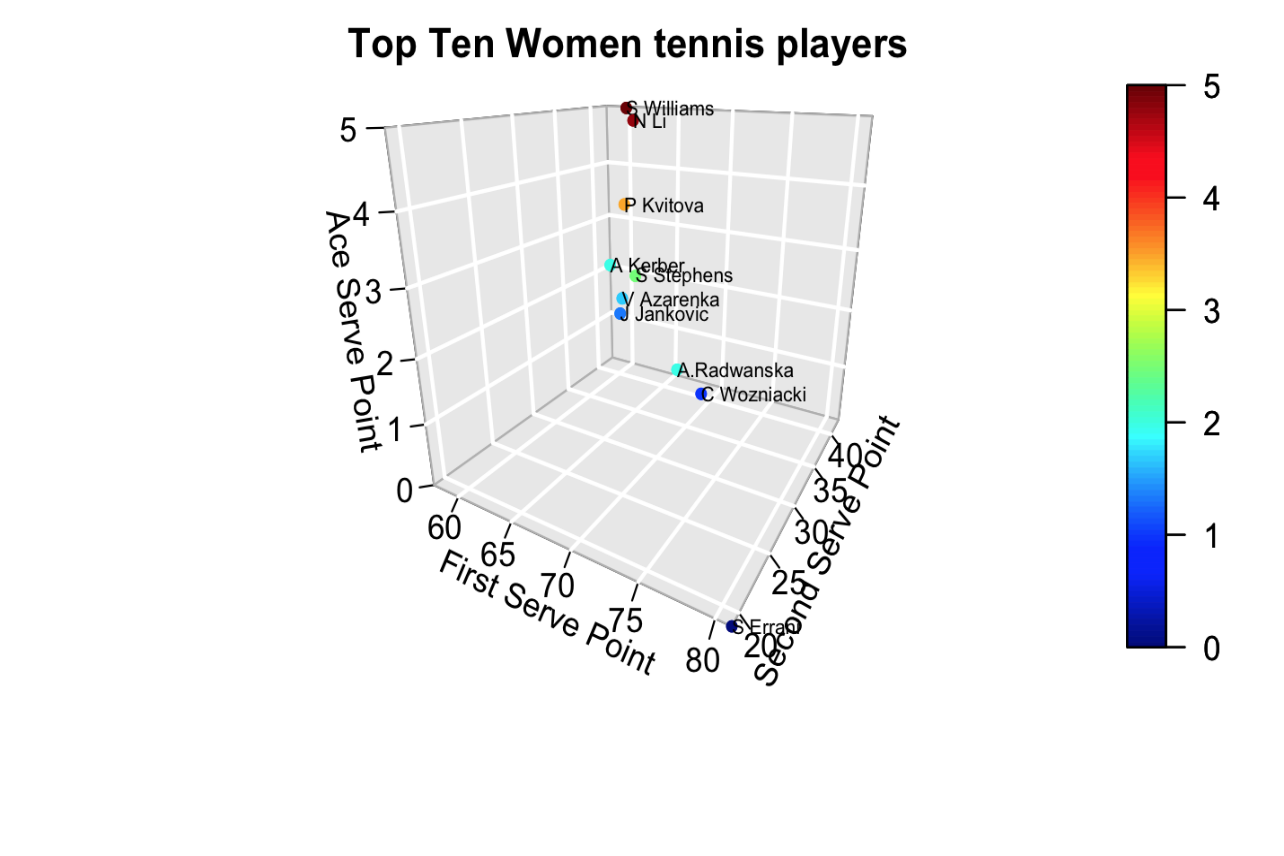


In the above bivariate box plot for men tennis players, Rafael Nadal has the highest number of wins and maolis baonic is the player who has the highest aces won in the tournament

In the above bivariate box plot for women tennis players, Victoria Azarenka and Serena Williams, and N Li are the three players which are outliers in this data, however Samantha Stosur, N li and serena Williams have the highest aces won in the tournament. And Victoria Azarenka has the highest match wins in the tournament.

**Scatter 3D Plot**





3D scatter plots are used to plot data points on three axes in the attempt to show the relationship between three variables. A fourth variable can be set to correspond to the color or size of the markers, thus adding yet another dimension to the plot.

In the above 3D scatter plot, the objective is to observe how top 10 world men and women players compared in winning first serve point, second serve point, and aces.

In top ten men tennis players graph, it is observable that Novak Djokovic, and Rafael Nadal achieved the highest first serve points, but milos Raonic have the highest aces and second serve points.

In top ten women tennis players graph, it is observable that Sarah Errani (Serrani) achieved the highest first serve points, and Serena Williams and Ni have the highest aces and second serve points.

**Dimension Reduction Analysis**

As data generation and collection keeps increasing, visualizing it and drawing inferences becomes more and more challenging. One of the most common ways of doing visualization is through charts. Suppose we have 2 variables, X and Y. We can use a scatter or line plot between X and Y and visualize their relationship easily. Now consider a case in which we have, say 100 variables (p=100). In this case, we can have 100(100-1)/2 = 5000 different plots. It does not make much sense to visualize each of them separately. In such cases where we have many

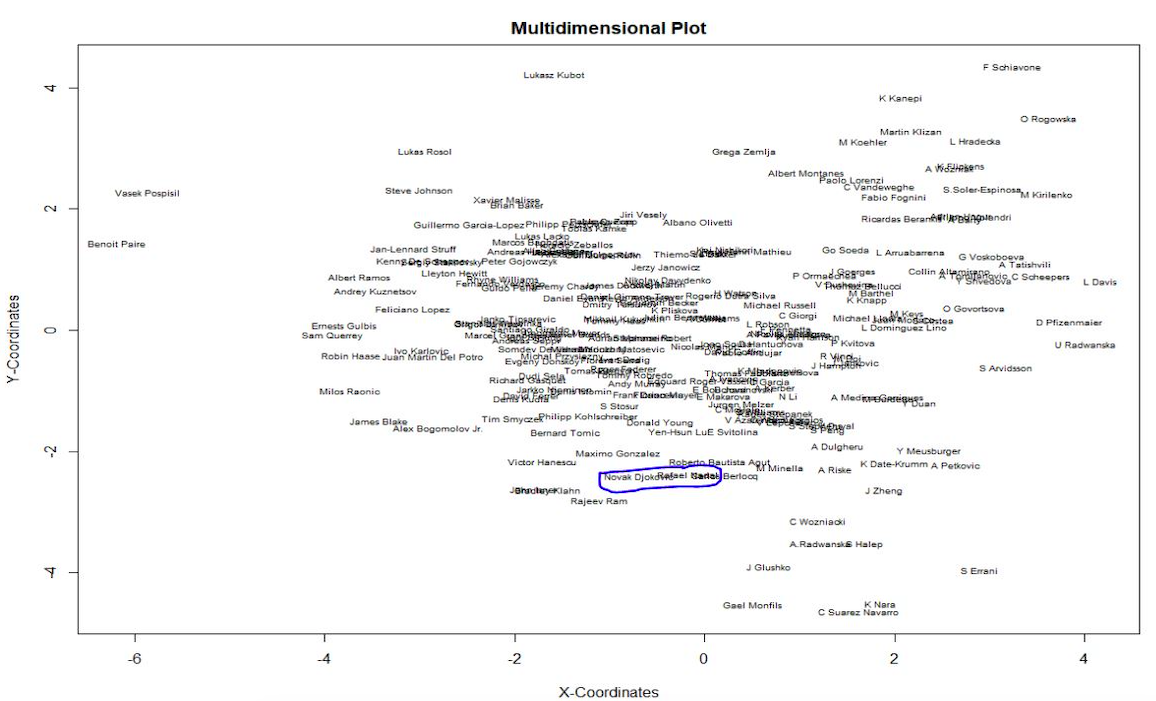
variables, it is better to select a subset of these variables (p<<100) which captures as much information as the original set of variables. This process of reducing the number of columns which can explain largest amount of variance as in the original data set is called as Dimension Reduction.

In our analysis we performed two techniques of dimension reduction, Firstly Multidimensional Scaling and Secondly, Principal Component Analysis.

**Multidimensional Scaling**

Multidimensional Scaling (MDS) is a technique of dimension reduction. It is used to supply a visual representation of the pattern similarities or dissimilarities among a set of objects. MDS plots the objects that are similar, near to each other and the objects that are not similar far from each other. MDS uses distance matrix as an input and are usually used to plot the 2-D representation of this distance matrix. This creates a cluster, the objects that are similar would be placed close to each other.

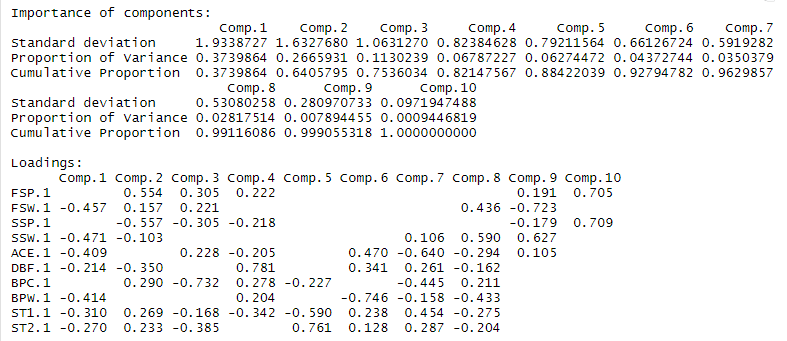
In the below multidimensional plot, we can see that players that have similarities in their play style for this tournament are clubbed together. As we can see Novak Djokovic and Rafael Nadal are near to each other and they were the finalist in the US Open 2013 men’s category.



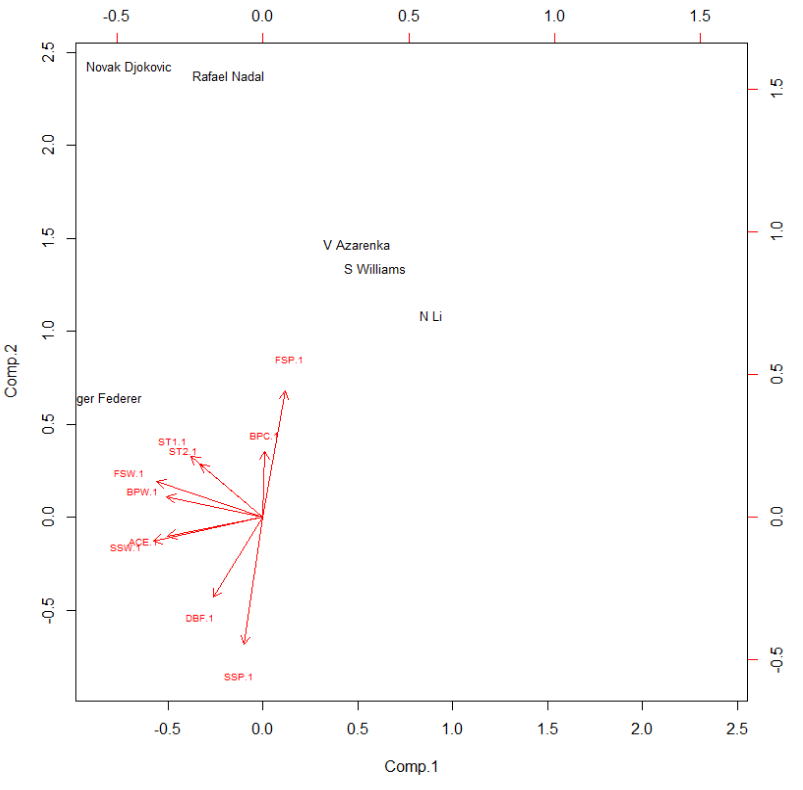
**Principal Component Analysis**

The central idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of many interrelated variables, while retaining as much as possible of the variation present in the data set. This is achieved by transforming to a new set of variables, the principal components (PCs), which are uncorrelated, and which are ordered so that the first few retain most of the variation present in all of the original variables. *[Jolliffe, Principal Component Analysis, 2nd edition]*

In the PCA analysis below it is understood that the first three components have a cumulative proportion of 75.3%, this tells us that the first three components can show the 75% of variance of the original data which had 10 variables.



In the biplot below the variables FSP and SSP are highly negatively correlated which is also visible from the correlation plot. The variables BPC and FSP are highly correlated which implies that most of the break points that were created in the tournament were through first serve. Similarly, FSW and BPW are very alike and correlated with high order, this implies that in the tournament most of the break points were in first serve which is practically very true. Novak Djokovic and Rafael Nadal show high value for both component 1 and 2 and have similar characteristics, however Azarenka and Williams have both moderate values for both the components.



**Cluster Analysis:**

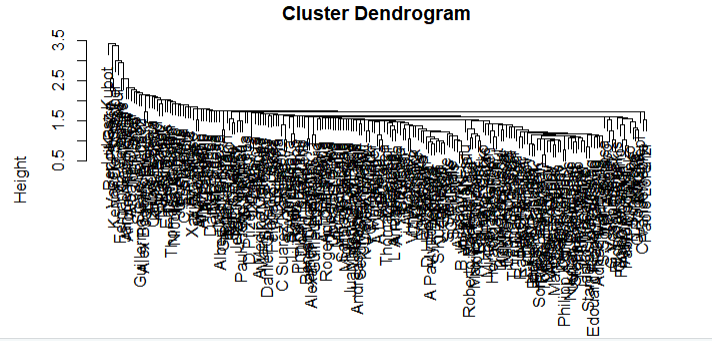
Cluster is a collection of objects, similar objects are placed or grouped in same cluster. This is an unsupervised technique. A good clustering method will produce high quality clusters that have low inter-class similarity and high intra-class similarity. The quality of the clustering method is measured by its ability to discover the hidden patterns in the data. *[*[*http://www.stat.columbia.edu/~madigan/W2025/notes/clustering.pdf*](http://www.stat.columbia.edu/~madigan/W2025/notes/clustering.pdf)*]*

**Clustering Techniques used in our Analysis**

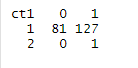
1. **Hierarchical Clustering:** Hierarchical clustering is an algorithm used for cluster analysis. Hierarchical clustering can use both raw data or distance matrix as an input. This process works by treating each observation as a separate cluster and then performs two steps: (1) identify the two clusters that are close to each other, and (2) merge the two closest clusters. The output of the hierarchical cluster is a dendrogram which represents the hierarchical relationship between the clusters.

* **Single linkage:** In single linkage hierarchical clustering, the distance between the two clusters is defined by the shortest distance between the two points in the cluster.

Below is the dendrogram for single linkage.

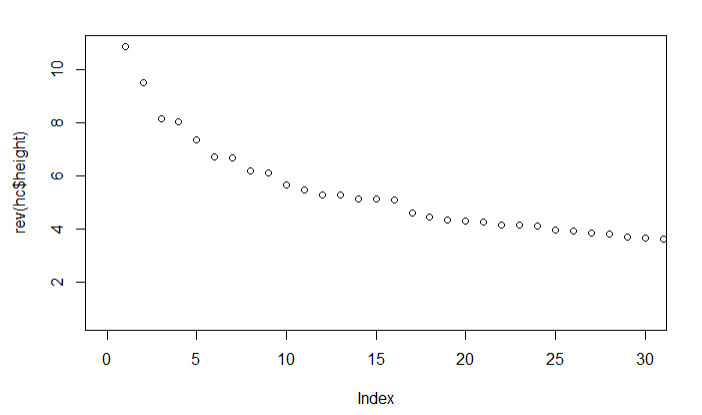


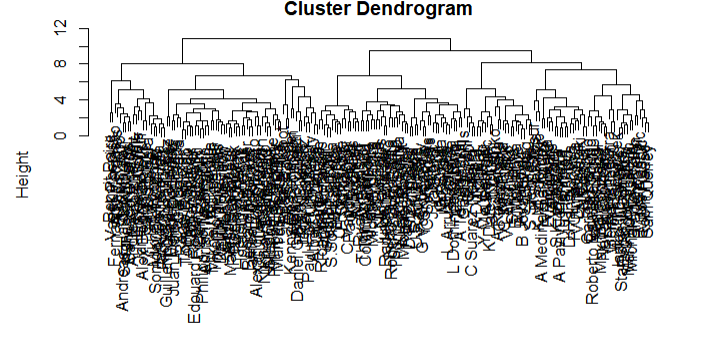
When the single linkage clustering is cut into 2 clusters and contingency table is created between Clusters and the Gender cluster 1 has most of the players and cluster 2 has just one player. This clustering doesn’t represent our data appropriately.

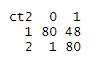


* **Complete linkage:** In complete linkage hierarchical clustering, the distance between two clusters is defined by the longest distance between the two points in the cluster.

Below is the scree plot for H-clust complete linkage.

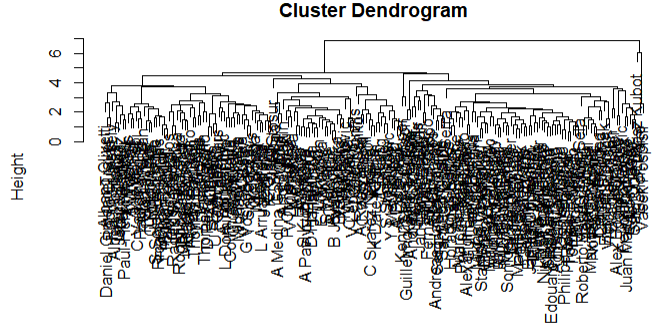
Below is the dendrogram which shows there are two major clusters.

The clustering when cut into 2 clusters and contingency table is created among cluster and Gender, cluster1 contained 80 women and 48 men and cluster2 contained 1 woman and 80 men. This method nearly divides the data into required clusters appropriately.



* **Average linkage:** In Average linkage hierarchical clustering, the distance between two cluster is defined by the average distance between each point in one cluster to every point in other cluster.

Below is the dendrogram for average linkage clustering.

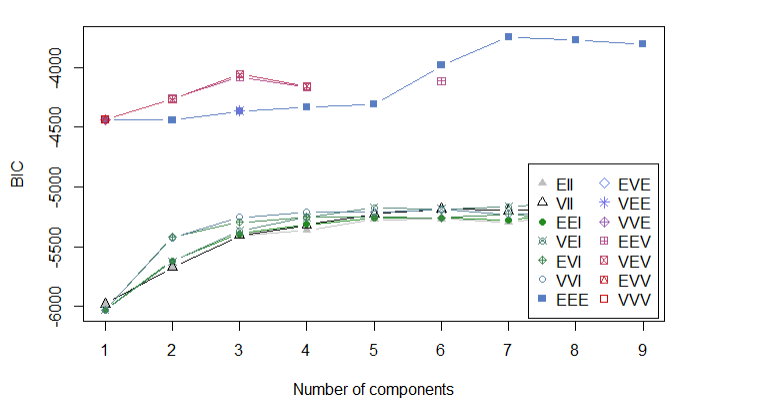


When the average linkage clustering is cut into 2 clusters and contingency table is created between Clusters and the Gender cluster 1 has most of the players and cluster 2 has just 3 players. This clustering doesn’t represent our data appropriately.

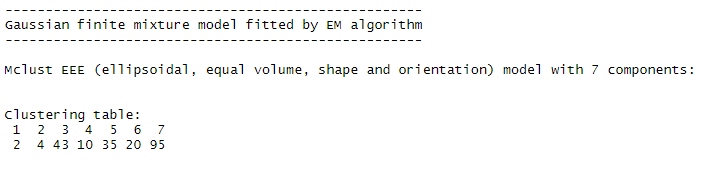


1. **Model Based Clustering:** In model-based clustering data is coming from a mixture of density (multivariate distribution). The model parameters can be estimated using the Expectation-Maximization (EM) algorithm initialized by hierarchical model-based clustering. The best model is selected using the Bayesian Information Criterion or BIC. A large BIC score indicates strong evidence for the corresponding model.

Below is the BIC plot for model-based cluster which shows there are 7 clusters.

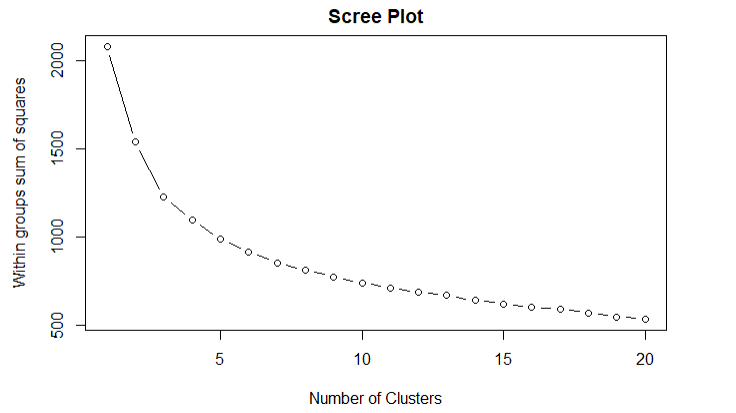


Result of model-based clustering:



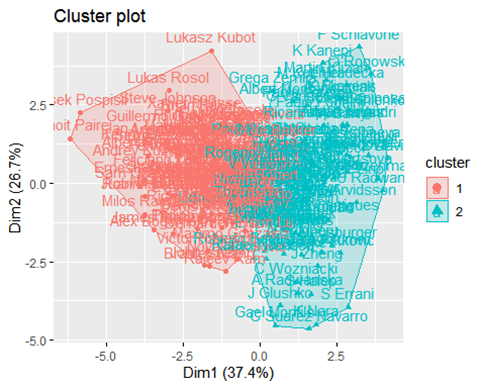
1. **K-Means Clustering:** The most used implementation of the k-means clustering is one that tries to find the partition of the *n* observations (rows of data) into *k* groups that minimizes the within-group sum of squares (WGSS) over all variables.

The Scree Plot for K-Means clustering suggests 2-3 clusters.

We proceeded with 2 clusters and applied K-Means clustering of two centers with nstart of twenty iterations to get a more precise result. In our cluster table we can see that cluster1 consists of 96 men and 2 women and cluster2 consists of 79 women and 32 men.



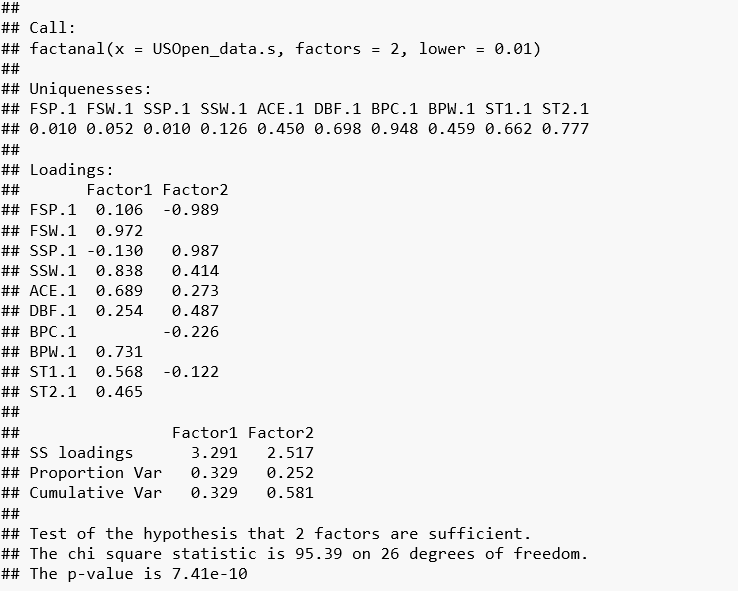
The package “factoextra” helps in plotting the clusters. It calculates the principal components and plots clusters based on first two principal components. In the below cluster plot the orange cluster represents the men’s category and blue represents the women’s category. The plot also shows in the center there were some crossovers which we had also observed in our cluster table.

 From the above cluster methods, we concluded that K-Means method is best suitable for our dataset which divided the data into appropriate clusters of men and women with a little deviation from expectation.

**Confirmatory Factor Analysis**

Confirmatory Factor Analysis (CFA) model may arise from theoretical considerations or be based on the results of an exploratory factor analysis where the investigator might wish to postulate a specific model for a new set of similar data. We first proceeded with the Exploratory Factor Analysis (EFA). In EFA we first find the unobservable latent variables. This is used to investigate the relationship between the manifest variables and the factors; however, CFA is used to test whether the specific factor model provides an adequate fit for the correlation between the variables.

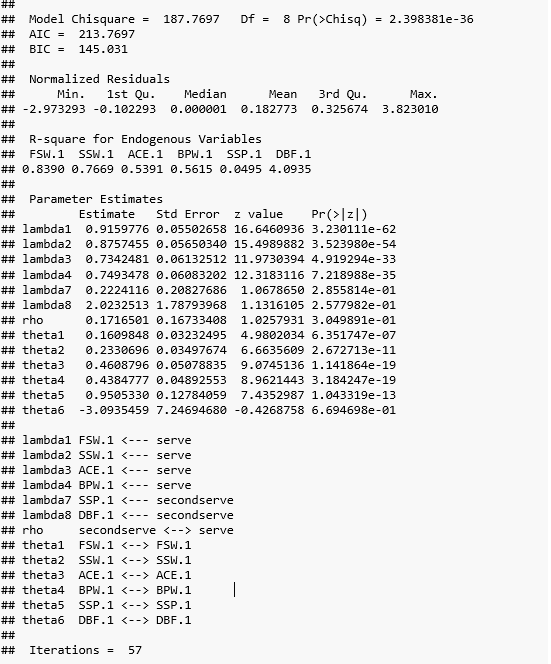
We started by determining the latent variables of our data. The Exploratory Factor Analysis (EFA) gave us the below output. As it can be observed below the variance of the error for FSP.1, FSW.1 and SSP.1 is smaller than other variables. We tried multiple factors ranging from 2-5, among them two factor results were more conclusive. The first latent factor which we termed as “Serve Characteristics” which was dominated by FSW, SSW, ACE and BPW all these are related to serve in tennis and the second latent factor corresponds to the “Second Serve Characteristics” dominated by FSP, SSP and DBF. The p-value is less than 0.05. This made us check the Confirmatory Factor Analysis on the two factors found below.



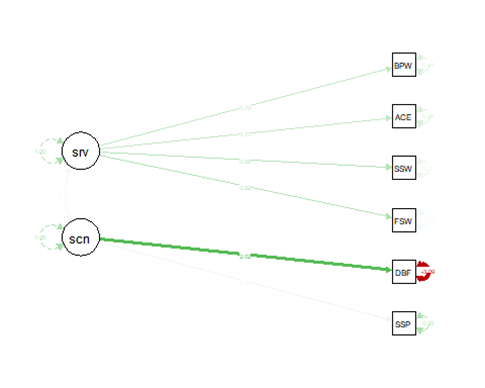
In this data we have found two latent variables using EFA, using this we have designed our factor model which we test using the CFA. We created the model based on the two latent variables.



The correlation between the factors is found to be very less 0.17, however the p-value of all the components is observed to be very less than 0.05. This result was produced after 57 iterations. Variance of error is least for FSW and estimate of “Serve Characteristic” being a latent factor for FSW is 0.91 with standard error of 0.05. The chi-square value of the model is large, and p-value is very less. The GFI and AGFI is very less hence we concluded that this model is not a good fit.



Variance of error is least for FSW and estimate of “Serve Characteristic” being a latent factor for FSW is 0.91 with standard error of 0.05. The chi-square value of the model is large, and p-value is very less. The GFI and AGFI is very less hence we concluded that this model is not a good fit.



**Conclusion**

After performing all the analysis, we were able to draw similarities in our data using the dimension reduction techniques. Our Principal component analysis showed that three components were enough to show almost the same variability as our data. In cluster analysis we tried multiple methods, but only k-means clustering was able to distinguish better results compared to other methods. In our factor analysis we were able to find two latent variables, which we tried to test by implementing the factor model using these to latent factors; however, the model was not a good fit as the GFI and AGFI was below the acceptable threshold. We can use this model in contrasting current player performances in different tennis tournaments.

**References**

1. <https://www.analyticsvidhya.com/blog/2018/08/dimensionality-reduction-techniques-python/>
2. <http://www.analytictech.com/borgatti/mds.Htm#N_3_>
3. <http://www.stat.columbia.edu/~fwood/Teaching/w4315/Fall2009/pca.pdf>
4. <https://www.displayr.com/what-is-hierarchical-clustering/>
5. <https://www.saedsayad.com/clustering_hierarchical.htm>