# MATH 4323 Project: Final Report

Group members: Noah Rhodes (2012376), Karthik Kallakuri(2073194), Alex Flores(1678014)

## Introduction

We selected the women's 2023 WTA Women's Tour data because we are interested in performing cluster analysis. We want to know what the natural groupings are for match winners. **How many types of match winners are there?**

We believe a natural number of *K* clusters for this dataset is 2. The 2 clusters correspond to the following groups. Perhaps group 1 corresponds to the higher ranked player winning the match and group 2 corresponds to the lower ranked player winning the match.

### Description of all variables:

WTA = Tournament number (women)
Location = Venue of tournament
Tournament = Name of tournament (including sponsor if relevant)
Date = Date of match (note: prior to 2003 the date shown for all matches played in a single tournament is the start date)
Series = Name of ATP tennis series (Grand Slam, Masters, International or International Gold)
Tier = Tier (tournament ranking) of WTA tennis series.
Court = Type of court (outdoors or indoors)
Surface = Type of surface (clay, hard, carpet or grass)
Round = Round of match
Best of = Maximum number of sets playable in match
Winner = Match winner
Loser = Match loser
WRank = ATP Entry ranking of the match winner as of the start of the tournament
LRank = ATP Entry ranking of the match loser as of the start of the tournament
WPts = ATP Entry points of the match winner as of the start of the tournament
LPts = ATP Entry points of the match loser as of the start of the tournament
W1 = Number of games won in 1st set by match winner
L1 = Number of games won in 1st set by match loser
W2 = Number of games won in 2nd set by match winner
L2 = Number of games won in 2nd set by match loser
W3 = Number of games won in 3rd set by match winner
L3 = Number of games won in 3rd set by match loser
W4 = Number of games won in 4th set by match winner
Wsets = Number of sets won by match winner
Lsets = Number of sets won by match loser

Comment = Comment on the match (Completed, won through retirement of loser, or via Walkover)
B365W = Bet365 odds of match winner
B365L = Bet365 odds of match loser
B&WW = Bet&Win odds of match winner
B&WL = Bet&Win odds of match loser
CBW = Centrebet odds of match winner
CBL = Centrebet odds of match loser
EXW = Expekt odds of match winner
EXL = Expekt odds of match loser
LBW = Ladbrokes odds of match winner
LBL = Ladbrokes odds of match loser
GBW = Gamebookers odds of match winner
GBL = Gamebookers odds of match loser
IWW = Interwetten odds of match winner
IWL = Interwetten odds of match loser
PSW = Pinnacles Sports odds of match winner
PSL = Pinnacles Sports odds of match loser
SBW = Sportingbet odds of match winner
SBL = Sportingbet odds of match loser
SJW = Stan James odds of match winner
SJL = Stan James odds of match loser
UBW = Unibet odds of match winner
UBL = Unibet odds of match loser
MaxW= Maximum odds of match winner (as shown by Oddsportal.com)
MaxL= Maximum odds of match loser (as shown by Oddsportal.com)
AvgW= Average odds of match winner (as shown by Oddsportal.com)
AvgL= Average odds of match loser (as shown by Oddsportal.com)

# Methodology

We will use K-Means and Hierarchical clustering to find natural groupings.

Advantages of K-Means clustering:

- We can use cluster validation to select the optimal number of *K* clusters. This is helpful to avoid finding noise patterns and to compare this solution with the hierarchical clustering solution.
- If we choose to perform PCA, cluster validation will also help us to compare cluster solutions in the original predictor space.

Advantages of Hierarchical clustering:

- We do not need to specify a number of *K* clusters beforehand.
- This solution results in a tree-like representation of the observations, called a dendrogram.

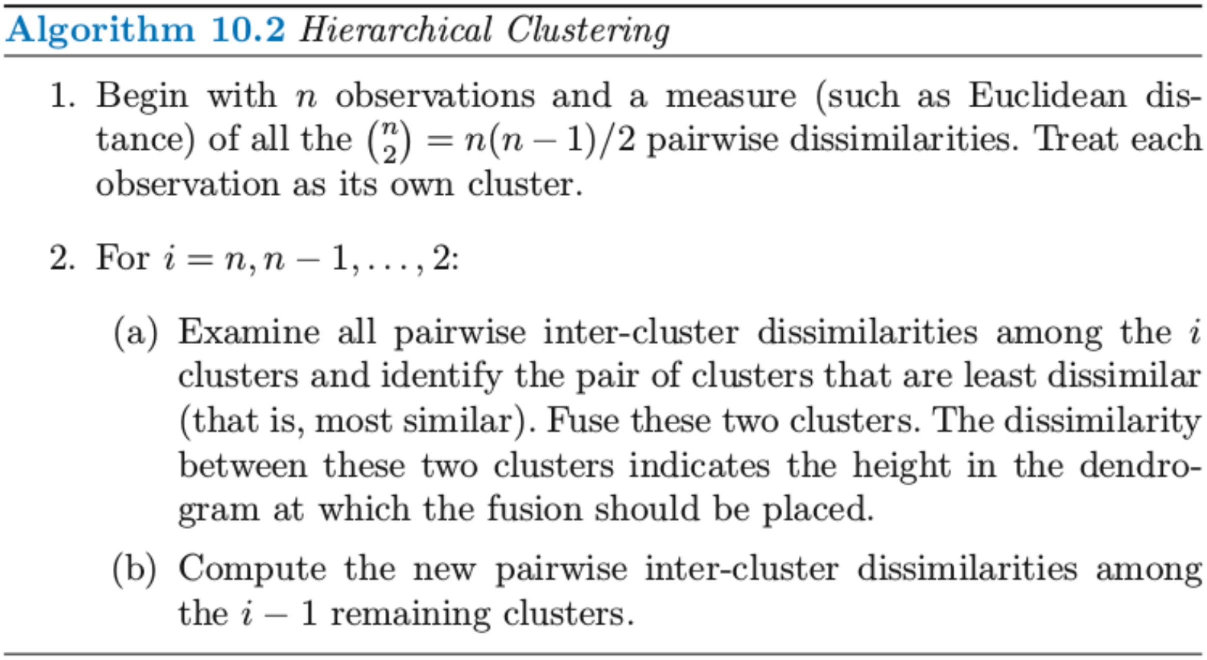
We now compare the performance of the two models. It is important to note that hierarchical clustering always results in a single cluster. Thus, we will take advantage of K-Means cluster validation approaches to help us determine the optimal number of *K* clusters. With this optimal *K*, we can then cut the dendrogram produced from our hierarchical clustering such that we obtain an equal number of *K* clusters in this solution, as well. Then we can compare the performance of each solution for a given value of *K*.

Explicit formulas for each model

- K-Means Clustering

$$\underset{C_1,\ldots,C_k}{minimize}\left\{\sum_{k=1}^{K} W(C_k)\right\}$$

- Hierarchical Clustering with "complete" linkage



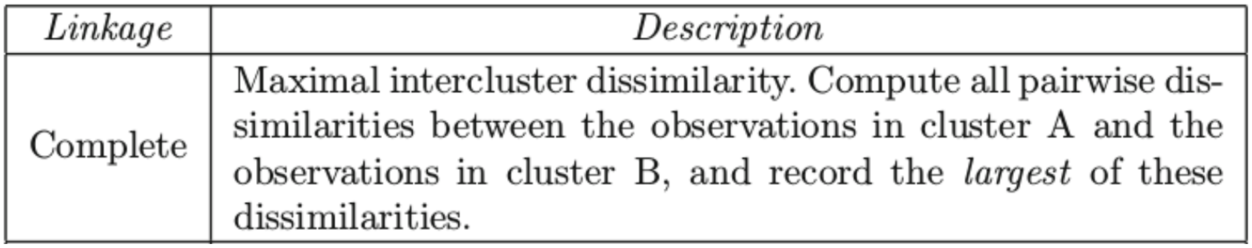
**Algorithm 10.2** *Hierarchical Clustering*

```
1. Begin with $n$ observations and a measure (such as Euclidean distance) of all the $\binom{n}{2} = n(n-1)/2$ pairwise dissimilarities. Treat each observation as its own cluster.

2. For $i = n, n-1, \ldots, 2$:

   (a) Examine all pairwise inter-cluster dissimilarities among the $i$ clusters and identify the pair of clusters that are least dissimilar (that is, most similar). Fuse these two clusters. The dissimilarity between these two clusters indicates the height in the dendrogram at which the fusion should be placed.

   (b) Compute the new pairwise inter-cluster dissimilarities among the $i - 1$ remaining clusters.
```

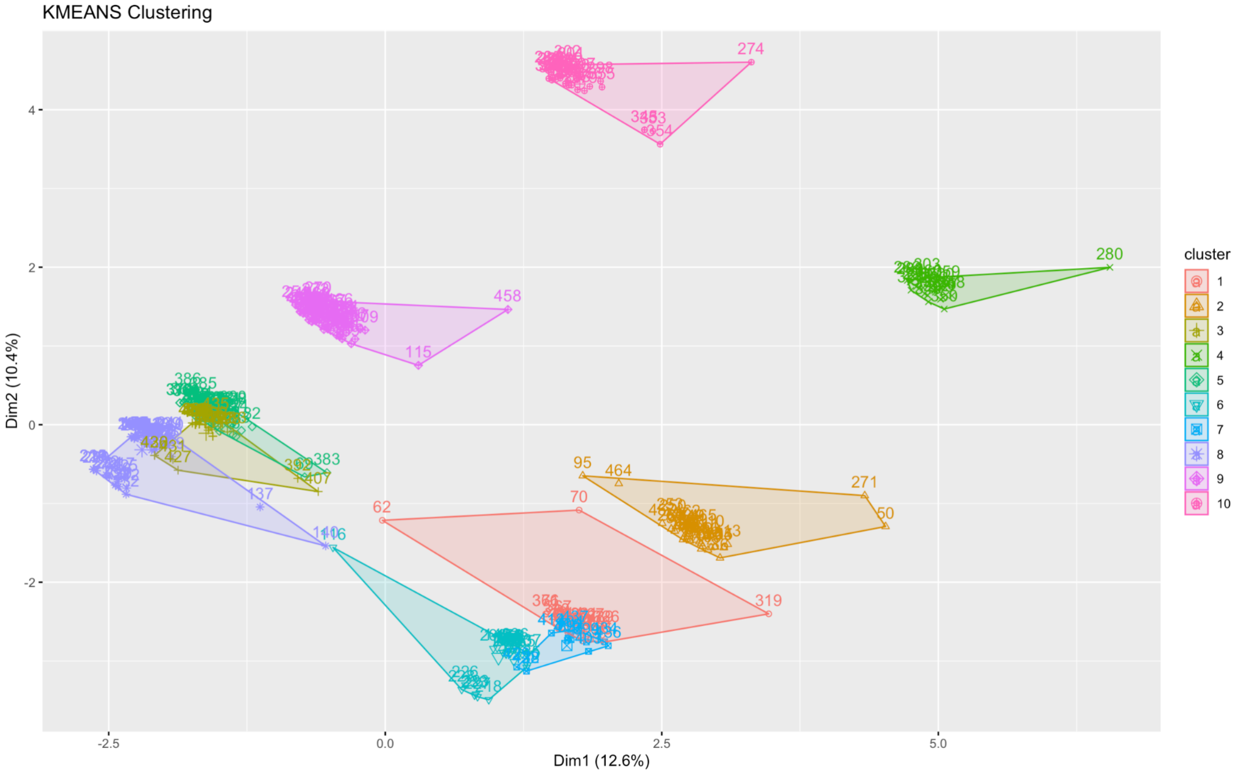


| *Linkage* | *Description* |
|---|---|
| Complete | Maximal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *largest* of these dissimilarities. |

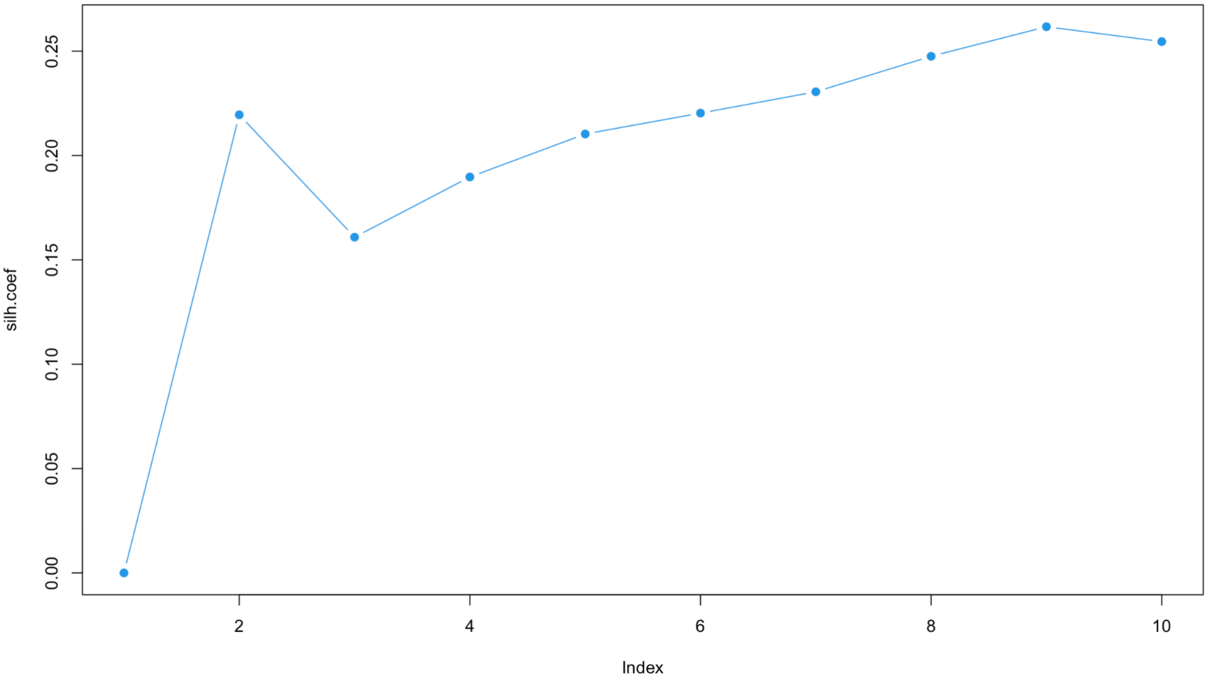
# Data Analysis

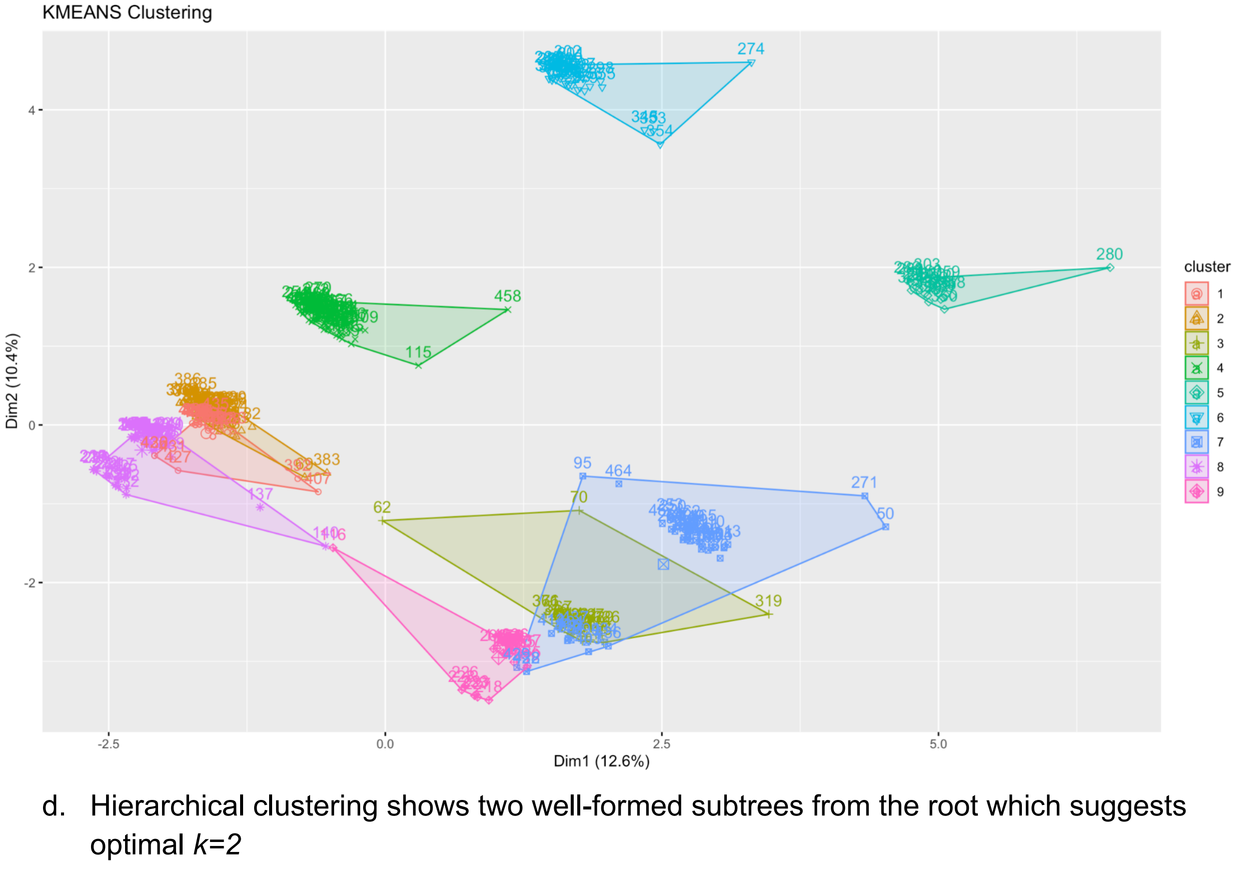
a. We chose to drop the following columns: *WTA*, *Tournament*, *Wset, Lset, Date*, *Surface*, *Best.of*, *Winner*, *Loser* since they are unnecessary. There are several winner, loser column pairs with numeric values that are easily used to create a single column with binary categorical values. **The data is modified below such that a 1 indicates the winner having the better column value.** For example, if the *WRank* (rank of the match winner)is < *LRank* (rank of the match loser) then create a new column *Rank* with a 1 for this row. Another example, if the *WPts* (ATP Entry points of the match winner as of the start of the tournament) is > *LPts* (ATP Entry points of the match loser as of the start of the tournament) then create a new column *Pts* with a 1 for this row. Yet another example, if *B356W* (Bet356 betting odds of match winner) is < *B365L* (Bet365 odds of match loser) then create a new column *B365* with a 1 for this row.

b. Cluster validation using gap statistic suggests optimal *k=10*.
   *nstart = 100 & nboot = 100*

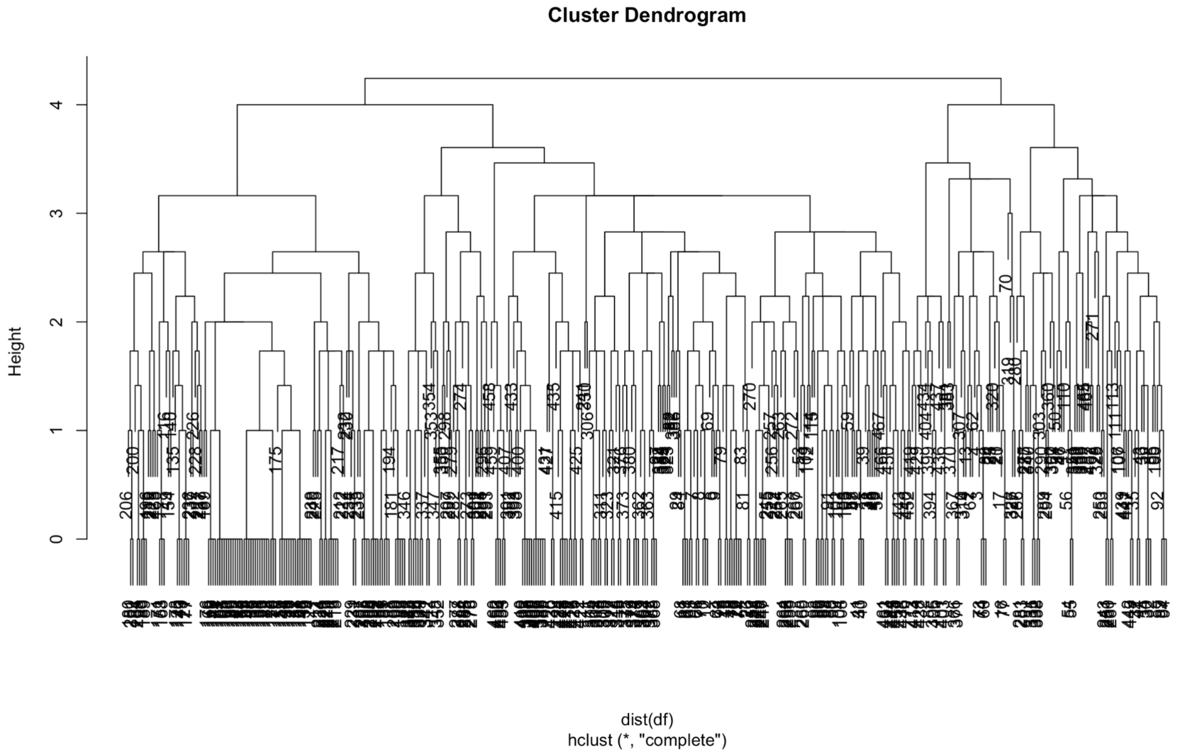


c. Cluster validation with silhouette coefficient suggests optimal *k=9*
   *Nstart = 100*





d. Hierarchical clustering shows two well-formed subtrees from the root which suggests optimal *k=2*



e. Interpretation

Consider the clustering solution obtained when cutting the dendrogram to obtain two clusters. The first cluster corresponds to the case where the winner has the better odds for most of the observations. In fact, in 89.9858% of the observations, the winner has better odds, on average. The second cluster corresponds to the case where the winner has worse odds. In fact, in 0.6392045% of the observations the winner has worse odds, on average.

We tried to make similar interpretations for the K-Means clustering solutions where K=9 and K=10, but they did not make much sense. The observations were not well separated according to rank, points or odds.

# Conclusion

Gap statistic suggests there are 10 types of match winners while silhouette coefficient suggests that there are 9 types of winners. However, these clustering solutions are not easily interpreted. Hierarchical clustering with complete linkage suggests there are 2 types of match winners. This clustering solution is easily interpreted as follows. The first cluster corresponds to the case where the winner has the better odds for most of the observations. The second cluster corresponds to the case where the winner has worse odds.

# References

- http://www.tennis-data.co.uk/alldata.php