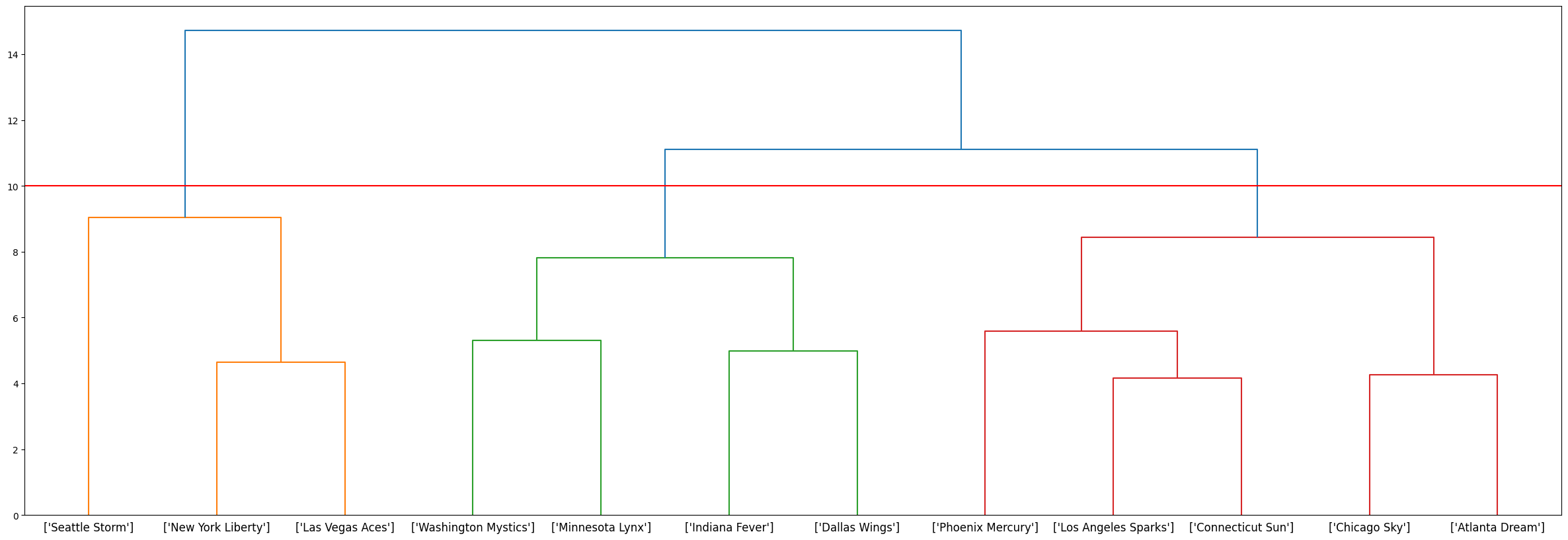
**Understanding the Problem and Data**

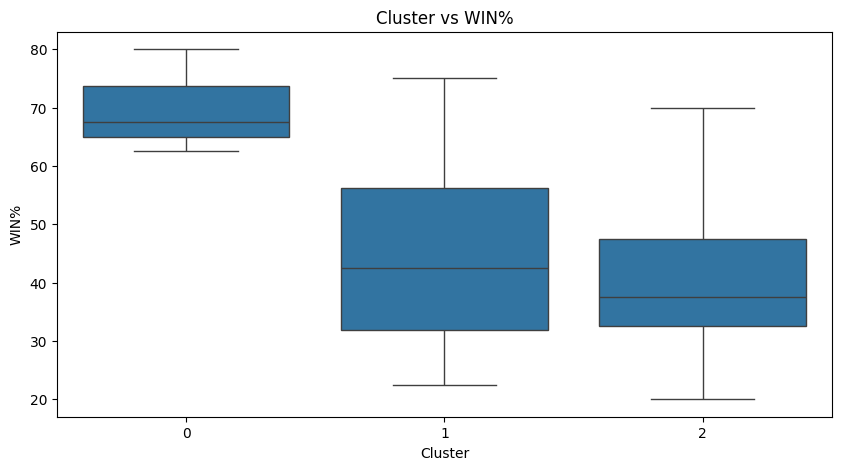
The dataset contains information regarding WNBA teams with statistics like average points per game, field goal percentage, win percentage in the regular season, etc. The goal of this analysis is to create multiple dendrograms, determine the best out of these, and be able to group the data according to the analysis that we wish to pursue.

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Maximum | Minimum | Mean |
| PTS | 86.4 | 77.0 | 81.675 |
| FG% | 45.4 | 40.8 | 43.625 |
| 3P% | 38 | 28.8 | 33.533 |
| FT% | 84 | 74.2 | 78.592 |
| OR | 10.9 | 5.6 | 8.225 |
| DR | 28.5 | 24.3 | 26.167 |
| AST | 23.0 | 18.4 | 20.525 |
| STL | 9.3 | 5.9 | 7.458 |
| BLK | 5.2 | 3.2 | 4.142 |
| TO | 15.1 | 10.8 | 13.283 |

**Model Estimation I**  
Below is the dendrogram using hierarchical clustering with complete linkage, with an integer-valued cut at 10, and three colored clusters.



* Cluster 0:
  + New York Liberty
  + Las Vegas Aces
  + Seattle Storm
* Cluster 1:
  + Minnesota Lynx
  + Indiana Fever
  + Washington Mystics
  + Dallas Wings
* Cluster 2:
  + Connecticut Sun
  + Phoeniz Mercury
  + Atlanta Dream
  + Chicago Sky
  + Los Angeles Sparks



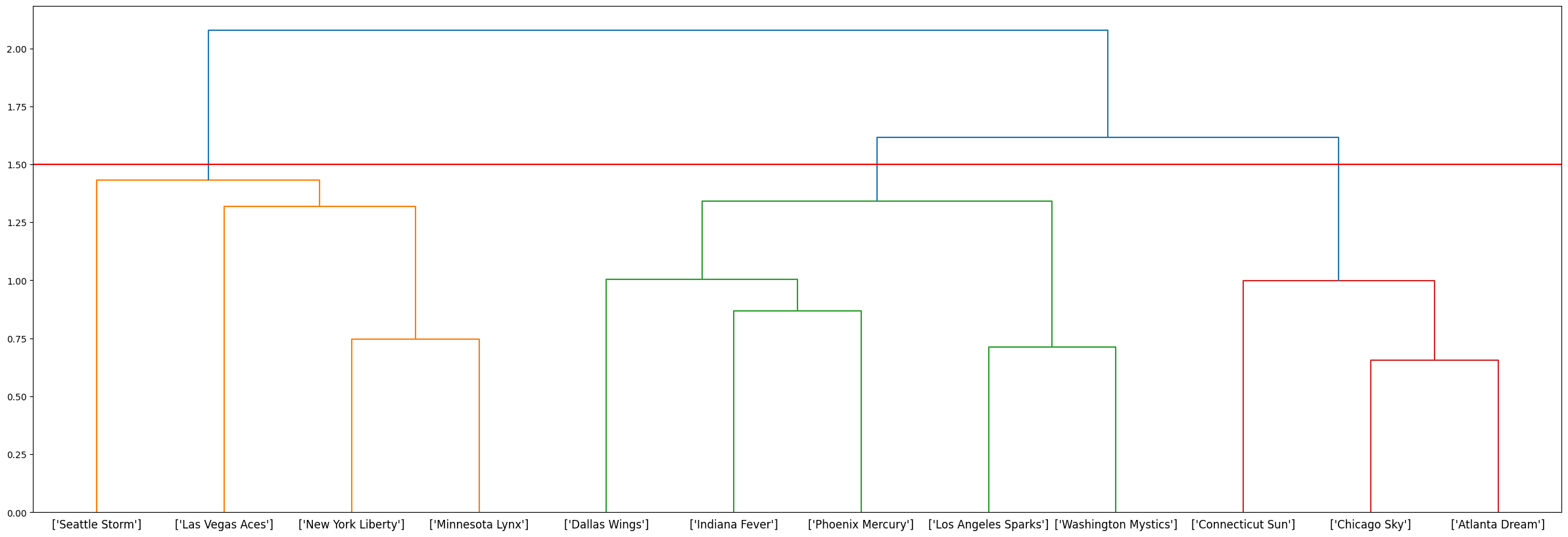
The variables that help push the teams in cluster 0 above the others are 3P%, FT%, and PTS. The teams in cluster 0 are above the average for the entire data set in these variables, 34.3, 81.8, and 84.3 for 3P%, FT%, and PTS respectievly. To win basketball games you need to score more points than the opposing team, and being above average in the categories that score points is a good indicator of how well a team will do.

**Model Specification II**

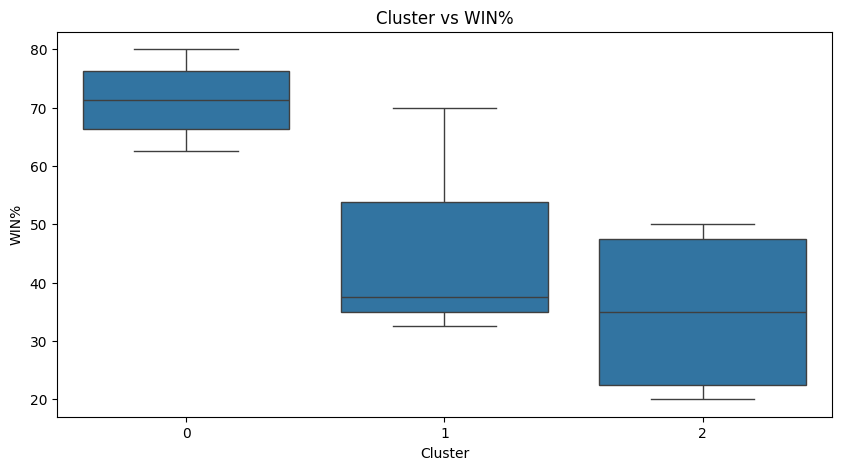
We might need to normalize the data because an increase of one in a certain variable may have more weight than an increase of one in another variable. For example, a one percent increase in field-goal percentage is not as impactful as an increase of one in a variable like steals or blocks because they have smaller ranges. This could lead to different clustering because if a team has a lot of steals, but not a high field-goal percentage, then they would appear close to teams with a few less steals, but far from teams with a ten or fifteen percent lead in field-goal percentage. Normalizing scales the data so that the distances between each variable are equivalent.

**Model Estimation II**

The most similar teams with the normalized data are the Chicago Sky and the Atlanta Dream as they join first in the normalized dendrogram pictured below. This differs from the original data where the closest teams were the Los Angeles Sparks and the Connecticut Sparks.



* **Cluster 0:**
  + New York Liberty
  + Minnesota Lynx
  + Las Vegas Aces
  + Seattle Storm
* **Cluster 1:**
  + Connecticut Sun
  + Atlanta Dream
  + Chicago Sky
* **Cluster 2:**
  + Indiana Fever
  + Phoenix Mercury
  + Washington Mystics
  + Dallas Wings
  + Los Angeles Sparks

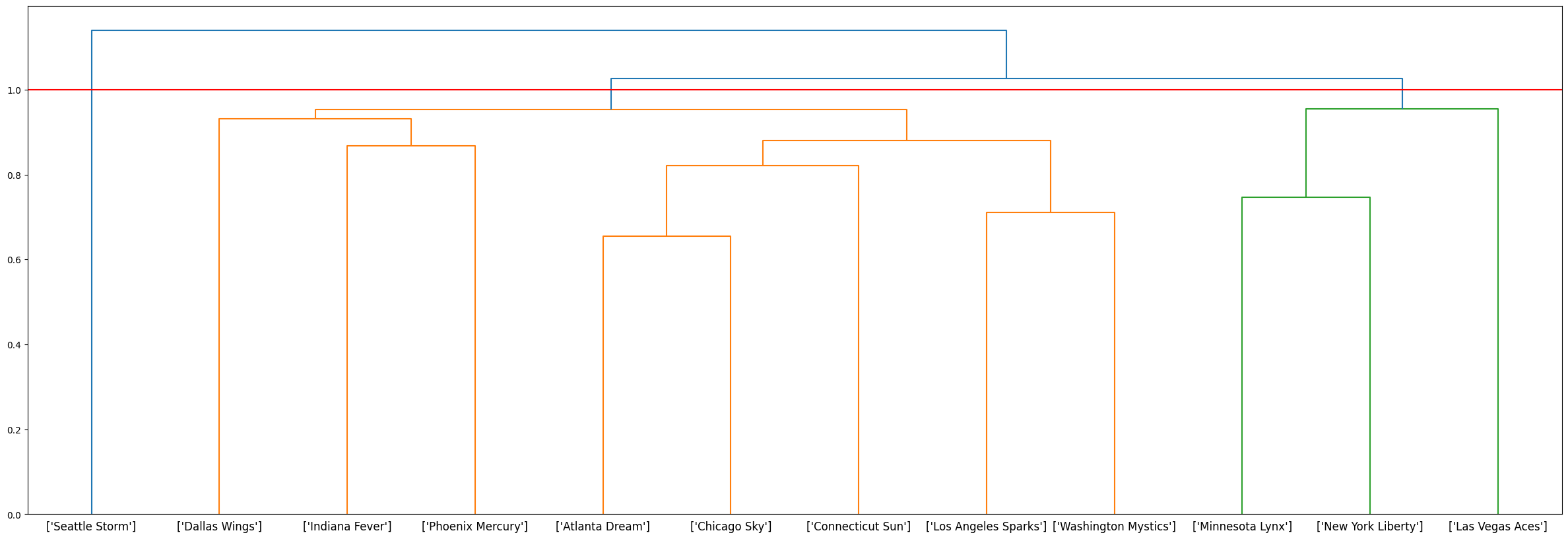
Cluster 0 has the highest win percentage on average.

**Model Specification III**

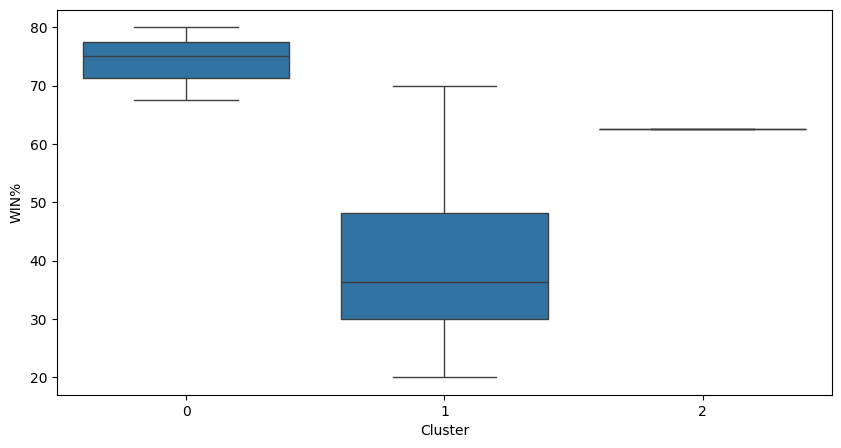
In complete linkage used earlier, it calculated the Euclidean distance between a point and the nearest cluster based on the farthest point in that cluster. This will not change the first connection, however, once clusters are formed it does change when they join. In single linkage, the distance between a point and the nearest cluster is calculated based on the closest point in that cluster.

**Model Estimation III**

A cut that leads to three clusters using single linkage is 1 as shown below with the red line.



* **Cluster 0:**
  + New York Liberty
  + Minnesota Lynx
  + Las Vegas Aces
* **Cluster 1:**
  + Connecticut Sun
  + Indiana Fever
  + Pheonix Mercury
  + Atlanta Dream
  + Washington Mystics
  + Chicago Sky
  + Dallas Wings
  + Los Angeles Sparks
* **Cluster 2:**
  + Seattle Storm



In all three cases, Cluster 0 appears to be the best performing cluster while the other two clusters remain somewhat similar, but Cluster 1 being better on average. However, in the final model with single linkage, Cluster 2 only has one team which is better than the average of eight teams that are in Cluster 1.

**Model Assessment**

Moving forward I chose the dendrogram from Model Estimation II. First, this Model does not contain a cluster with only one team like the dendrogram from Model Estimation III does. Additionally,

**Model Deployment**

There does not exist a clustering where each cluster either has all teams in the playoffs, or all teams not in the playoffs. This is because of how closely related the Atlanta and Chicago teams are, in the model I chose they joined first, and Atlanta made the playoffs while Chicago did not. This makes it impossible for a clustering to exist where all teams either did or did not make the playoffs unless you put each team in their own cluster.

Based on the dendrogram, our analysis does not support the teams in the semi-finals. New York, Minnesota, and Las Vegas, the 1st 2nd and 4th seeds respectively, join each other relatively early, however the 3rd seed the Connecticut Sun joins Chicago (who didn’t make the playoffs), and Atlanta (the lowest seed in the playoffs). Additionally, the 5th seed Seattle joins the cluster with the top teams and the 3rd seed Connecticut only joins the best cluster when all the teams are in one cluster. Essentially, our model rates the 3rd seed, the Connecticut Sun, much lower than the WNBA website does.

The largest number of clusters where the New York Liberty and Minnesota Lynx are in the same cluster is 9. This tells us that these teams have similar stats and are likely to be better than the other competition.