Data Questions & Models

Noah Rhodes, Alex Flores, Binh Pham, Lien Pham, Karthik Kallakuri

1. Main goals
    a. We would like to obtain natural groupings of observations in our data. We believe a natural number of *K* clusters for this dataset is 3. The 3 clusters correspond to the following groups. Group 1 is the case where the odds of the match winner are less than the odds of the match loser. Group 2 is the case where the odds of the match winner is equal to the odds of the match loser. Group 3 is the case where the odds of the match winner are greater than the odds of the match loser.

2. Models and methods
    a. We will use K-Means and Hierarchical clustering to find natural groupings.
        i. Advantages of K-Means clustering:
            1. 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.
            2. If we choose to perform PCA, cluster validation will also help us to compare cluster solutions in the original predictor space.
        ii. Advantages of Hierarchical clustering:
            1. We do not need to specify a number of *K* clusters beforehand.
            2. This solution results in a tree-like representation of the observations, called a dendrogram.
    b. 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*.

3. We will distribute the workload as follows (tentative).
    a. We will perform data preprocessing so that we can perform each clustering algorithm with mixed data type.
    b. Noah & Lien will implement and interpret K-Means clustering with cluster validation.
    c. Alex, Binh & Karthik will implement and interpret Hierarchical clustering.