Connor Eng

111612084

CSE 332 Project 3

Project 3 Report

Correlation Matrix:

Pros- Correlation matrices are good for determining which variables have a strong positive or negative correlation. This means that the correlation matrix will show the correlation coefficients between each variable/attribute where larger values such as .7 and -.7 show a strong positive and negative correlation, and values such as .1 and -.1 have a weak correlation. In turn, this will allow data scientists to determine whether certain variables have any effect on one another before diving deeper into studies on the dataset.

Cons- Correlation matrices are not good for deeper analysis. This means that it is hard to determine if the data has clusters, outliers, and why these variables have this type of correlation. The plot does not show any data points so it is impossible to create predictions or assumptions other than the relationship between two or more variables.

My Data- My correlation matrix shows that there are four attributes that have a high correlation to the rest of the 10 attributes. These attributes are points scored by the home team, points scored by the away team, field goal percentage by the home team, and field goal percentage by the away team. They seem to have around the same correlation coefficients because these attributes are the main premise of the game of basketball, to score the most points. The attributes with the weakest correlation coefficients with the 10 attributes are the free throw percentages by both home and away teams. This is because free throws are when the game is paused and cannot have any correlation with statistics such as assists or rebounds since the game is paused for a player to

shoot two to three shots. I found it interesting how rebounds by the away team had such a high correlation coefficient to field goal percentage by the home team. This is most likely because more rebounds by the away team must mean that the home team is missing their shots which allows rebounds by the away team. This correlation matrix cannot show which attribute relationship leads to a win or loss for the home team.

Scatter Plot Matrix:

Pros- Scatter plot matrices are similar to correlation matrices in that they are good for looking at the correlation between attributes or variables. The difference between the scatter plot matrix and the correlation matrix is that the scatter plot matrix is better at showing large quantities of data and clusters. This plot is good for looking at the larger picture of the dataset to see how all of the attributes correlate with each other in more detail than correlation matrices since it shows data points.

Cons- Scatter plot matrices are not good at showing accurate measurements of data. In addition to not showing accurate measurements, scatter plots are not able to show categorical data. Since there are so many data points and in a small area, it is hard to distinguish relationships or clusters that are too small to notice. In addition, the small size can alter some data points to make it less accurate.

My Data- My scatter plot matrix shows similar correlations as the correlation matrix. We can see that points by home/away teams and field goal percentages by home/away teams have the strongest relationship. The interesting part of this scatter plot matrix is that it shows which data points are when the home team wins or loses. The blue data points are statistics for when the home team wins and the red data points are statistics for when the home team loses. I found it

interesting how the rebounds by the away team had such consistent relationships. Most of the games played where the away team had large amounts of rebounds resulted in a home team loss. Before looking at this dataset, I thought the game of basketball was more about just out scoring the enemy team but now it seems as though defense has a very important role within the game as well. This type of plot is not good for looking at individual team performances because scatter plots cannot display categorical data.

Parallel Coordinates Plot:

Pros- Parallel coordinates plots are good for looking at trends throughout attributes in a dataset. The lines that it creates connecting the attributes with data points allows us to see the types of relationships between all of the attributes. This is especially important because we can visualize a dataset with a large number of dimensions. It is only restricted by the screen's horizontal length. In addition, stretching the plot vertically can also make it easier to spot trends and types of correlations.

Cons- Parallel coordinates plots are bad for looking in depth. These plots tend to have a lot of clutter due to the large amount of data points connecting all of the attributes. In addition to the clutter, it also may be difficult to visualize trends between attributes that are far apart in the plot. If the dataset contains many outliers, it will also make it difficult to visualize the trends because it will create colors that are blended together.

My Data- My parallel coordinates plot shows that there are strong positive correlations between points scored by the home/away teams and the field goal percentage of the home/away teams. This plot also consists of two colors determining whether the home team won(yellow) or loss(blue). Since the yellow color is more visible at first glance, we can see what the home team

statistics looks like when the home team wins. I found it interesting how the strongest correlation and trend for home team wins was not between the points scored and field goal percentage. The strongest correlation for home team wins was between field goal percentage of the home team and rebounds by the away team. The line between these two attributes has the strongest negative correlation in which all data points seem to have the same value. The second strongest correlation is between points scored by the away team and rebounds by the home team. The yellow line for home wins has a very distinguishable increase. This plot has a hard time showing trends where the data is cluttered. Trends tend to become a blur.

PCA Plot / Scree Plot:

Pros- PCA plots are good for removing correlated features of the dataset to help create principal components that are independent of each other. This helps to distinguish the differences between attributes to find clusters and relationships between clusters. In addition, this makes visualization easier because it helps to separate the data into smaller clusters to reduce overfitting. The scree plot helps to determine the top eigenvector components that can be used to create a biplot.

Cons- PCA plots are not good for independent variables and not standardized data. This is because when standardizing the data to create the components, the variance will be too high to achieve accurate results. In addition, independent variables will have lost data since the data will be turned into principal components that are linear components of the original dataset. Scree plots have no other use than to find variances for each component.

My Data- My PCA plot shows that principal component 1 has a larger range of values compared to principal component 2. In addition, I added color to show whether or not the home team won

or lost. This color scale helped to distinguish the 2 clusters of the PCA plot. I was surprised to find that the clusters were determined by principal component 2 and not principal component 1. Home team wins are positive values of principal component 2 and negative values of this component or the losses. We can also see that a large amount of the data points are centered around values of 0 for both principal components. This plot is not good for understanding trends and values of free throw percentages by both home and away teams because they were determined to be independent from the other attributes.

Biplot:

Pros- Biplots are good for showing the relationship between principal components, data points, and attributes. It serves the same purpose as the PCA plot but adds another dimension to the analysis of these attributes in relation to principal components. In addition, it helps to show the correlation between attributes.

Cons- Independent variables will be extremely difficult to visualize in this plot due to the small size of the vectors. In addition to the small vectors, independent variable data is also lost. This means that the plot will not show an accurate representation of all of the data and just data that is standardized and has some correlation.

My Data- My biplot consists of the top 2 PCA vectors, points scored by the home team and points scored by the away team. It was interesting to find that both points scored by the home and away team have a positive relationship to principal component 1. This means that data points on the right side of the graph have high values for principal component 1, points scored by the home team, and points scored by the away team. It makes sense that principal component 2 has a negative relationship with points by away team and positive relationship with points scored by

the home team. This is because more points scored by the home team should mean a win for the home team. The color distinguishes the two clusters for home team wins and losses just like the PCA plot.

MDS of Data:

Pros- MDS displays of data are good for visualizing large amounts of data in a 2D plot. It is too difficult to visualize data with more than 3 dimensions. MDS is used to help visualize data clusters and determine the different types of variation between the clusters. It is helpful to visualize the distance between data points. This means that the axises are not as important compared to the PCA plot. Therefore, the plot can be manipulated to help visualize the different clusters and data points as long as the distance between the points stays the same.

Cons- MDS is not as good when the dataset is more linear. This is because the PCA plot would be more effective to use due to the principal component axes. Lastly, MDS is hard to determine the causes of the clustering of some data points especially if there exists independent variables.

MDS Data- I found it interesting how the plot is basically an inverted version of the PCA plot. Instead of the home team wins having a positive relationship with dimension 2, it has a negative relationship. In addition, the dimension 1 scale is inverted compared to the PCA plot. The MDS plot has more outliers compared to the PCA plot, meaning that there are data points that overlap their respective cluster area. This is most likely because the plot is of all of the data points where free throw percentages of home and away teams are seen to be independent variables. This plot also cannot show the individual team performances for the data points since that attribute is categorical.

MDS Attributes:

Pros- The MDS of attributes is good to show the relationship between the given attributes. We are able to see the distance between the correlations of all of the attributes to determine the attributes that are most related to each other and the attributes that are independent.

Cons- The MDS of attributes cannot be used to accurately predict dataset trends because it is not the plot of the data and is only the plot of attributes. It also cannot be compared with other attributes that are categorical.

MDS Attributes- The MDS display of the attributes seems to have different outcomes compared to the previous plots. It shows that there is a cluster of four attributes in the top left and bottom right quadrants of the plot. This is interesting because I can determine by looking at the graph that the four attributes are the points scored by home/away teams, field goal percentage of home/away teams, assists of home/away teams, and rebounds of home/away teams. The home statistics would be one of the clusters and the away statistics would be the other. This is because their own statistics should have a similar correlation to one another. It is also interesting to see that my hypothesis on how free throw percentage was independent is seen in this plot. There are two data points that are outliers from the clusters.