Michael Cissokho

4/5/20

Capstone Data Wrangling/Cleaning Description

For my project of evaluating past NBA match data and using a regression analysis model to determine categorically the winner and loser of each match going forward based on key statistics several steps were involved in cleaning/wrangling the data and making it suitable for evaluation.

The data was copy and pasted directly from the NBA's website into an excel spreadsheet. All the headers for each column needed to be rewritten in order to be accepted by Microsoft Excel. In addition I needed to include a column to represent the home and away team for each match as there were two lines for each match representing each of the team's statistics. Next I needed to include another column with game id's to identify each match going forward in the dataset. One problem that was encountered in organizing the data this way was that because there were two lines for each game, the home team's statistics and then the away team's statistics, duplicate lines needed to be created for game id's for each game. The way this was done was to manually input the first two rows game id's (0001 and 0001 respectively) and then use a formula that could then be used to fill in the remaining cells in the column (a4 = a2+1). This was effective as there would be a gap of two rows consistently in the spreadsheet as the formula made its way down the column. The last step involved looking at the date column which represented the date the game took place in. The problem there was that the NBA website attached a hyperlink to the date in each row in order to refer back to the match itself on its website. Upon removing the hyperlinks however the datetime format would be changed to an unrecognizable set of numbers. A spot check in a jupyter notebook however using python, allowed me to see that the date would show up fine in the dataframe and no hyperlink would be attached.

There were no missing values in the spreadsheet so that wasn't a problem. There were outliers in a few games with regards to minutes played as some games

occasionally go to overtime, and therefore the statistics for the match are impacted, and it can be inferred that sometimes teams have unusually good or bad games that those statistics would be outliers, however as the purpose of the model was simply to determine the winner and loser of the game, both teams playing would play the same minutes, and there is no way of knowing minutes played prior to the match. With stats of individual teams being outliers there is no way of knowing that without running an individual regression of all 30 teams prior to creating this model. The data frame was read into the jupyter notebook using pandas, and read in correctly and from a spot check looked fine.