Wrangle Report

pass-2016-wrangle

The gather phase of the project was done in the following steps: I loaded the csv file 'pass-2016.csv' into a dataframe called df_original. I then made a copy called df for wrangling purposes.

The cleaning portion of the project was done in the following steps: I found that there was extra data in the 'Player' column that appeared to be some sort of id. I also found asterisks in some of the rows as well, with no discernable meaning. In both cases, I removed the data from the column. I found that there were multilpe data points in the column 'QBrec'. I out each of these data points into their own column, the three being called 'QBwin', 'QBlose', and 'QBtie'. I found that there was data in the 'Pos' column that contained the same data in both lowercase and uppercase form. I converted the entire column to uppercase for consistency.

Finally, I put the cleaned dataframe into a csv file called 'pass-2016-master'.

run-2016-wrangle

The gather phase of the project was done in the following steps: I loaded the csv file 'Career_Stats_Rushing.csv'into a dataframe called df_original. I then made a copy called df for wrangling purposes.

The cleaning portion of the project was done in the following steps: I found that the symbol '--' was used to indicate missing data. I changed this to NaN. I found that The columns 'Rushing Attempts', 'Rushing Yards', 'Yards Per Carry', 'Rushing Yards Per Game', 'Rushing TDs', 'Longest Rushing Run', 'Rushing First Downs', 'Percentage of Rushing First Downs', 'Rushing More Than 20 Yards', 'Rushing More Than 40 Yards', and 'Fumbles' were of the data type 'object'. I changed this to float through the to_numeric function. I found that all of the colum names that had multiple words also had spaces. I changed all of the spaces to underscores. I found that one row had a value in the column 'Percentage of Rushing First Downs' was greater than 100. Since this is a percentage column, this was in error. I went to ESPN's website (linked on page), and found that the error was commited there too. Since I could not validate any of the data for this player, I deleted the entire row.

Finally, I put the cleaned dataframe into a csv file called 'Career_Stats_Rushing_master.csv'.