

Kinduct Take Home Assessment

Kirk MacDonald

To complete all assigned questions, 3 notes were created.

DataLoader – Load and process the data. A Kaggle API called was used to originally download the csv file. While trying to load the data into a df, an issue was observed. A number of rows in the csv file contain a different number of columns. To resolve this, the function file_cleaner was created. This function reads the csv file line by line and checks how many columns each row has, if more than 23 (23 determined by exploring CSV file), the row gets written to a separate file for further investigation. All rows that pass the check, get written to another file, ready to be loaded into a dataframe. Further test cases can be added to the function to capture other data issue in the future.

FinalDF – This notebook creates a dateframe that contains tmID, year, Wins_agg, Losses_agg, GP_agg, Mins_over_GA_agg, GA_over_SA_agg, avg_percentage_wins. Note that the calculation for the fields were assumed to be exactly as described in the take assessment. Two groupbys are done in this notebook to calculate stats. One groupby per player per year and a second groupby per team per year. Joins were done assuming playerID and year is a unique key. No functions were created in this notebook as Pandas was able to do all transformations. If the calculated fields have to be reproducible for other tables, they could be turned into functions.

FinalDict – This notebook creates the two dictionaries the take home assessment requires. This was done by doing a groupby (playerID to find career totals) and then the required fields were calculated from the grouped data. To create the dictionary output, the dataframes were sorted and then sliced. No functions were created in this notebook as Pandas was able to do all transformations. If the calculated fields have to be reproducible for other tables, they could be turned into functions.