

Dataset links:

all_seasons.csv: <https://www.kaggle.com/datasets/justinas/nba-players-data>

games.csv: <https://www.kaggle.com/datasets/nathanlauga/nba-games?resource=download>

- Both are csv
- Entries for all_seasons.csv: 12844
 - Columns for all_seasons.csv: 22
- Entries for games.csv: 26651
 - Columns for games.csv: 21

gamesdf.dtypes		seasonsdf.dtypes	
[18]	✓ 0.0s	[9]	✓ 0.0s
...	GAME_DATE_EST object	...	Unnamed: 0 int64
	GAME_ID int64		player_name object
	GAME_STATUS_TEXT object		team_abbreviation object
	HOME_TEAM_ID int64		age float64
	VISITOR_TEAM_ID int64		player_height float64
	SEASON int64		player_weight float64
	TEAM_ID_home int64		college object
	PTS_home float64		country object
	FG_PCT_home float64		draft_year object
	FT_PCT_home float64		draft_round object
	FG3_PCT_home float64		draft_number object
	AST_home float64		gp int64
	REB_home float64		pts float64
	TEAM_ID_away int64		reb float64
	PTS_away float64		ast float64
	FG_PCT_away float64		net_rating float64
	--		oreb_pct float64

- game.csv has a different format for the season column than the all_seasons.csv
- The game.csv needs to have the game date reformatted to be consistent with the years of all_seasons.csv
- The all_seasons.csv can be cleaned on the I, J, K columns where “Undrafted” is a value among all the other int values in the column.
- There do not seem to be missing values, but there are inconsistent data formats between the two datasets (i.e. season column)
- Merging with seasons column after cleaning
- We are uniquely positioned to work with these files because Harsh is a big basketball fan. Aaron and Minji are not very familiar with basketball, so they could help gather an unexpected insight.