Data Wrangling section 3

Jusitn Huang

To make sure there were no mistakes I did each year from 2000 to 2019 in their own separate R script file. I then later rbind them and renamed columns names as well as did a little bit of cleaned up of the data.

Loaded the different libraries needed to clean and sort the data. And set the directory to nba folder.

Dplyr, tidyr, stringr

Loaded the data.

2019\_age.csv is from stats.nba.com

hoops\_hype\_salary\_2019.csv from hoopshype.com

2019\_stats.csv from basketball-reference.com

Cleaned up the stats, salary and age data by removing unnecessary columns that weren’t going to be used. Also made sure when doing a full join changed the names column to player\_name for all 3.

There were some problems when doing an initial full join. Had to clean up names to remove “-“, “\*”, “,” from the name column for each one. Each website also had different spellings for some of the players. They also would include Jr or III after the name.

Used grep and gsub to find and replace these names so they would all be matching.

There were also issues of having duplicate names due to player trades and movement. Used which and duplicate to remove the duplicate rows and kept the total. When doing the full join(age will have the team column) it will keep the last team they played on for the season.

After the full join. Changed the height column so that it would not be a date like June 6th which should be 6’6”. Used gsub and which to change this.

After there was still some NA’s for salary for several rows. Had to create another csv file and go through hoopshype.com and <https://www.eskimo.com/~pbender/misc/salaries09.txt>. To fill in the missing salary and missing NA columns. Loaded in the data removed the old NA’s and rbinded the new missing\_salary\_2019.csv.

Used the stringr package to make sure the format of the salary column was correct and changed to a numeric.

Then looked through the rest of the NAs in the stats criteria and age criteria columns. Did a google search on most of the players to see if they played that season. If they were waived, injured, did not play that season they were removed from the data.

For each year set player\_id for that year. Loaded in tax and set in the location tax for each team, loaded in team wins and inflation for each year. All included by full\_join().

For the award section I used previous year data. The reason why is I wanted to see if the awards would affect the players performance or salary the following year after receiving the award.

Created a new column for each award such as MVP, NBA first team, sixth\_man, etc…

Filled all the data with NO first. Then used <https://www.landofbasketball.com/awards/>

To fill in the “NO” as “YES” if the player for the 2019 data received an award for that criteria. Used grep to find the player and filled it into a value which located and put “YES” if they won the award.

Then wrote the csv.

This process is repeated for 2000-2019. Then was loaded into another R script to rbind all the data.