DSC520_Final_Project_Step2

Phil Han

8/8/2021

9.3 Final Project Step 2

- I imported a baseball dataset from Sean Lahman's website. It is pretty much the same dataset as the one from Keggle, but it includes the extra years of data from 2016 to 2020.
  I imported its csv file and removed any rows with NA. I also removed such columns that are not relavant to this research as "stints", "team ID", and "league ID." I also plan to condense the data size by taking a 50 year or 30 year worth of the data for an ideal sample size.

- Here is my first dataset:

```
##        playerID yearID G AB R H X2B X3B HR RBI SB CS BB SO IBB HBP SH SF GIDP
## 12630 aitchra01   1911 1  0 0 0   0   0  0   0  0  0  0  0   0   0  0  0    0
## 12701 camniha01   1911 2  0 0 0   0   0  0   0  0  0  0  0   0   0  0  0    0
## 12725 collaal01   1911 2  0 0 0   0   0  0   0  0  0  0 NA   0   0  0  0    0
## 12739 cottren01   1911 1  0 0 0   0   0  0   0  0  0  0  0   0   0  0  0    0
## 12831 griffha01   1911 1  0 0 0   0   0  0   0  0  0  0  0   0   0  0  0    0
## 13046 pucketr01   1911 1  0 0 0   0   0  0   0  0  0  0  0   0   0  0  0    0
##       X2B3B
## 12630     0
## 12701     0
## 12725     0
## 12739     0
## 12831     0
## 13046     0
```

- I still do not know how to impute values into empty cells, so I need to learn how to do that. Also, I am thinking about replacing the empty cells with 0.
- From the dataset, I realized that a category or variable for extra base hit+ (X2B3B) that combines doubles and tripes is missing. As such, I created one in (data1).
- One other step that I am planning is to join the attendance data from another dataset to the existing data frame.
- Also, I will investigate to see if the decreasing extra base hits (X2B or X3B) are correlated with decreasing attendance.I may plan on dropping the data from the pandemic shortened season of 2020 or find the projected 2020 data from a baseball data metric and replace the data for year 2020 with it.
- Another way to look at this data is whether or not strikeouts increased while the extra base hit+ decreased, which could explain why there are a lack of much action around the leagues.
  One other category that I will look into is stolen base.
- Stealing baese creates a lot of action around the bases. Thus, it is a valid category to research to see if it has a correlation with a decreasing trend of extra base hits. Lastly, I will also look at the number of home runs by year and the trend. There may be an inverse relationship between strikeouts and home runs, which also might explain why batters are going for the fenses instead of hitting for extra bases.
- Basically, I will summarize my data by displaying any correlations among extra base hits (2B & 3B), Strikeouts, Stolen Bases, Home Runs, and attendance.
- I plan on using histograms, scatter plots, residual plots (normal Q-Q plots and density plots). As for tables, DT within R Markdown as well as data.table package.
- I have yet to make a table out of the dataset that I've cleaned, but I plan on making a presentalbe table for the project where necessary.
- Once I learn about machine learning techniques in Week 10 in our DSC 520 class, I will consider incorporating some into my project.

Additional questions for the future steps

- I am seriously considering filtering the data by year so the results do not come out skewed since the data goes way back to Year 1871.
- Also, I have to figure out how to join the attendance data to the dataset in an approriate manner. And I will probably end up further removing some of the variables that may not be needed.
- Lastly, I will have to figure out how to calcute slugging % from the dataset and may consider adding it to the final dataset.