Data Wrangling

Brian Hallberg

6/4/2018

Initial data loads

Load the data into a data frame and include all the libraries. The data files are all in CSV format. The files included are listed below:

- **player.csv** - data with all the players, including name key for all other files, full name, year start and end were the critical fields for me.
- **all_star.csv** - data with all-star information, includes players per year that played.
- **hall_of_fame.csv** - data per year with players names and whether they were inducted int the hall of fame.
- **batting.csv** - data with statistics per year for each player and team with information such as at bats, hits, doubles, triples, home runs, RBIs, games played, runs scored, stolen bases, caught stealing, were the most useful.
- **pitching.csv** - data with statistics per year for each player and his teams with information such as wins, losses, games, outs pitched (innings pitched * 3), saves, hits give up, home runs allowed, complete games, shutouts, earned runs, walks, strikeouts were the most useful.

Wrangle the data

Create some Summary Data sets

We'll create a summary of the all-star data that is based on the player and how many years they played in the all-star game. The existing files are by player and year. I want to know how many years the player made an All-Star game but I don't care which year.

```
sumallstar <- all_star %>%
  select(player_id, league_id, year) %>%
  group_by(player_id) %>%
  count(player_id)
```

Next a list of all the players that have made the Hall of Fame. The first elections were in 1936 and the last in this data set are were elected in 2016. The original data file includes players as well as managers, umpires, and executives. I only care about players. It also lists all the years when they voted on them for inclusion, but I only care about those that were elected to the hall.

```
sumhof <- hof %>%
  filter(category =="Player") %>%
  select(player_id, inducted) %>%
  filter(inducted == "Y")
```

The batting and pitching data file is unique for the combination of player_id/year/team since a player can play more than one year and within a year they may have played for multiple teams. I only care about the players career totals. So we group by player_id and add all the other columns which are numberic. I don't care about year or stint (which is just a index to the teams they played for in the year) so I drop them. We'll use these to do a join on this data and the player data to make one big dataset.

For hitting I add batting average (ba) which is hits (h) divided by at bats (ab).

```
sumhit <- hit %>%
  group_by(player_id) %>%
  summarise_if(is.numeric, sum, na.rm = TRUE) %>%
  mutate(ba = h/ab) %>%
  select(-year, -stint)
sumpitch <- pitch %>%
  group_by(player_id) %>%
  summarise_if(is.numeric, sum, na.rm = TRUE) %>%
  select(-year, -stint, -baopp, -era)
```

Now we build the final data set.

Include hall of fame data, all star data as well as pitching and hitting data for all players. This gets all the needed data in one place.

```
fullplay <- full_join(player, sumhof, by = "player_id")
fullplay <- full_join(fullplay, sumallstar, by = "player_id")
fullplay <- full_join(fullplay, sumhit, by = "player_id")
fullplay <- full_join(fullplay, sumpitch, by = "player_id")
```

Cleanup column names

Once pitching and hitting were added there are some similar fields, so we'll clean them up here.

```
fullplay <- rename(fullplay, allstar = n, rallow = r.y, hrallow = hr.y, bballow = bb.y, hallow = h.y)
fullplay <- rename(fullplay, gpitch = g.y, sopitch = so.y, batterhit = hbp.y)
fullplay <- rename(fullplay, game = g.x, run = r.x, hit = h.x, hr = hr.x, bb = bb.x, so = so.x, hbp = hb
```

Remove some missing data

- if inducted is missing, the value is N
- if allstar is missing, the valued is 0

```
fullplay <- fullplay %>% mutate_at(vars(inducted), funs(replace(., is.na(.), 'N')))
fullplay <- fullplay %>% mutate_at(vars(allstar), funs(replace(., is.na(.), 0)))
```

Reduce to Hall of Fame Eligible Players

In this section, I'm reducing the data set of players to those eligible for induction into the hall of fame. The rules say that the player must have played major league baseball for 10 years and they must be out of the game for 5 years. So we'll reduce the set of players to those that meet those requirements. There have been exceptions made in the past, but I think that is beyond the scope of this report.

```
eligible <- fullplay %>%
  mutate(timein = floor(difftime(final_game, debut, units="days")/365)) %>%
  filter(timein > 10) %>%
  filter(final_game <= "2011-01-01")
```

Write out the Data Sets to files

Save a few of the temporary files for later plotting and analysis. Saved as CSV files and rds

```r
write_csv(eligible, "eligible.csv")
saveRDS(eligible, "eligible.rds")
```

Some Notes on missing data

Of the 247 hall of fame players, 26 have data missing about either the debut or final game or both. We have dropped those from the final data set since there is not enough data about them. There are an additional 10 players dropped becuase they did not meet the calculated minimum play time and were elected for other reasons that are beyond the scope of this report.