STATS615finial_project_NBA

Group 4:Yuka Chen, Jack Lo, Adela Yang

2022-03-23

```
library(tidyverse)
library(broom)
```

1. Problem Statement:

Often time sports players would seem to have major contract with really high annual salary (some people would even think they should not get paid so much).

Since of one our group members is a super fan of basketball league, the National Basketball Association (NBA), in North America, he believes that those basketball players are paid by their season performance and they were not overpaid. However, other members in our group think otherwise.

Through this project, we want to find out whether these NBA players were overpaid, and do all of them have higher salary compared to previous contract they signed, or they salary were higher each year due to the inflation in the US dollar currency.

For this project, we would try to create a prediction model for NBA player's salary and help the players to predict their annual salary by their seasonal performance, such as scoring, rebound, field goal percentage etc.

##2. Identify a data set that is relevant to your question:

The two dataset were collected from Basketball Reference (https://www.basketball-reference.com/) by Chris Davis (https://data.world/datadavis).

The first dataset includes the information of NBA players backgroun information such as ID, Name, DOB, weights, heights, position, shoots etc, as well as their career performance.

```
players <- read_csv("players.csv") ##Rmarkdown and dataset are saved in same folder
## 24 variables and 4584 observations
## 20 characters and 4 doubles variables
#spec(players) #to check variable types
#str(players) # to check variable structure
names(players) #to check all variable names
```

```
##  [1] "_id"         "birthDate"   "birthPlace"  "career_AST"  "career_FG%"
##  [6] "career_FG3%" "career_FT%"  "career_G"    "career_PER"  "career_PTS"
## [11] "career_TRB"  "career_WS"   "career_eFG%" "college"     "draft_pick"
## [16] "draft_round" "draft_team"  "draft_year"  "height"      "highSchool"
## [21] "name"        "position"    "shoots"      "weight"
```

The second dataset includes the playerID, team name, season periods (year) and their salaries the 1984-1985 season to the 2017-2018 season.

```
salaries <- read_csv("salaries_1985to2018.csv") #same folder

## 7 variables and 14163 observations
## 4 characters and 3 double variables
```

```
#spec(salaries) #to check variable types
#str(salaries) # to check variable structure
names(salaries) #to check all variable names
```

```
## [1] "league"       "player_id"    "salary"       "season"       "season_end"
## [6] "season_start" "team"
```

3. Proposed Method of Analysis:

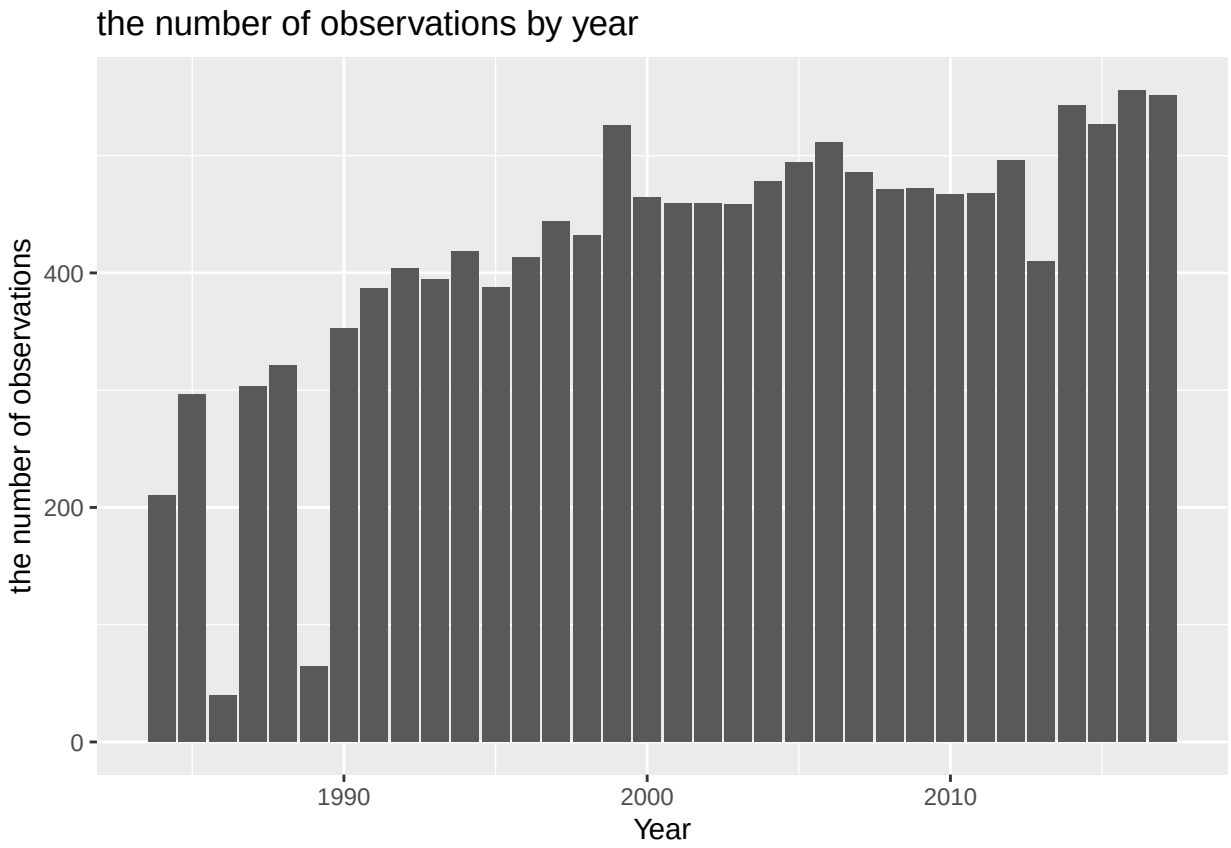
1. we would combine two dataset into one.
2. data exploratory analysis for variables visualization and see if there are any potential linear relationship or outlinears for the models that we are interested
3. to verify the linear relationship by t.test and other statistical information
4. using simple linear relationship to create a model for predicting salaries
5. using multiple linear relationship to see if there are other variables hae stronger relationship with salaries than the ones we expected
6. stepwise regression to find the best fited prediction variables for our models
7. visualization for our models

4. Exploratory Data Analysis (Still Working On It):

```
## combine two dataset
NBA <- salaries %>%
  inner_join(players, by = c("player_id" = "_id"))
```

For all of the plots and text below, "year" will refer to the year that the season started. For example, year 2017 refers to the 2017-2018 season.

```
salaries %>% group_by(season_start) %>%
  count() %>%
  ggplot(aes(season_start, n)) +
  geom_col() +
  labs(x = "Year", y = "the number of observations",
       title = "the number of observations by year")
```


the number of observations by year
the number of observations
0
200
400
1990
2000
2010
Year