final project data demo

David Hong

2023-04-12

```
data1 <- read.csv('131dataset.csv')
any(is.na(data1))
```

```
## [1] FALSE
```

An overview of your dataset

My dataset consists with nba 2k ratings for different players. It includes every basic stats of a player, such as points per game, 2k rating, and etc. The dataset is obtained from kaggle(https://www.kaggle.com/datasets/willyiamyu/nba-2k-ratings-with-real-nba-stats?resource=download). There are 2412 observations and roughly 31 predictors(but i will omit some of them because some are redundant). The data types are mostly continuous. Luckily, I do not have any missing data due to integrated environment of Kaggle.

An overview of your research question(s)

I am interested in predicting the ratings of a player. I wish that I am able to make prediction based on my model on new datasets. My response variable is continuous, and it is a numerical variable that is designed to describe the goodness of a basketball player(The higher the better). The Questions will be answered in regression approach. I think points per game and field goal percentage are especially useful. The goal of my model is predictive. The reason behind it is i am trying to predict the 2k ratings of players based on their statistics with my model.

Your proposed project timeline

I am planning to start doing eda in week3 and by the end of midterm i will try to select an appropriate model to fit my dataset. By week 8 i will finish my project and make adjustments to my project in week 9. After all, i will check everything and submit it in week 10.

Questions or concerns

Dear professor Coburn, do you think this dataset will be too easy to work with? Intuitively, the first idea that comes to my mind to build the model is using multilinear regression. However, that is more like the knowledge from pstat126. Are we going to learn more stuff on new regression algorithms? I really want to apply new knowledge to this project.