Final Project Part 2

Andrew Jordan

2022-05-20

Import and Clean my data

```
library(tidyverse)
draft_df <- read.csv("data/draft.csv")
draft_df <- filter(draft_df, position == "K")
draft_df <- draft_df[c("draft","round","pick","draftTeam","nameFull")]
head(draft_df)
```

```
##   draft round pick draftTeam         nameFull
## 1  1977     4   89       ATL   Leavitt, Allan
## 2  1977     5  127       BUF O'Donoghue, Neil
## 3  1977     9  239        SF     Posey, David
## 4  1977    10  258        NO  Septien, Rafael
## 5  1977    10  278       MIN      Beaver, Dan
## 6  1977    11  302        LA     Long, Carson
```

```
fieldGoal_df <- read.csv("data/field-goal-stats.csv")
fieldGoal_df <- filter(fieldGoal_df, Year >= 1977)
fieldGoal_df <- fieldGoal_df[c("Name","Year","Team","Games.Played","FGs.Made","FGs.Attempted","FG.Percer
fieldGoal_df <- fieldGoal_df %>% mutate_at(c("FGs.Made","FGs.Attempted","FG.Percentage"),as.integer)
fieldGoal_df <- aggregate(cbind(Games.Played,FGs.Made,FGs.Attempted) ~ Name, data = fieldGoal_df, sum)
fieldGoal_df$Career_FG_Percentage <- with(fieldGoal_df, (FGs.Made / FGs.Attempted) * 100)
head(fieldGoal_df)
```

```
##              Name Games.Played FGs.Made FGs.Attempted Career_FG_Percentage
## 1   Abbott, Vince           23       21            34             61.76471
## 2 Aguayo, Roberto           16       22            31             70.96774
## 3   Aguiar, Louie           16        1             2             50.00000
## 4    Akers, David          237      386           477             80.92243
## 5   Allegre, Raul           92      137           186             73.65591
## 6 Alvarez, Wilson            4        3             7             42.85714
```

Importing and cleaning my data starts with bringing in each csv file as a dataframe, identifying the most relevant variables, removing all non-relevant variables, and finally combining duplicate rows to better represent career statistics (as opposed to individual season statistics).

What does the final data set look like?

The final data set will be a combination of the imported and cleaned csv files. It will include only players who appear in both of the above datasets to remove any kicker statistics for which there is not corresponding

draft information, as well as the player's number of games played, points scored, kick percentage, and the year and round in which they were drafted.

Questions for future steps.

How should I go about importing and merging data regarding team statistical performance, which is measured by season, with my current dataset measuring a players' performance over their entire career? Should I include the year in which a player was drafted when determining return on investment? Have the values of draft picks changed over time? How can I best combine field goals made, field goals attempted, and field goal percentage into a single value that reflects a kicker's performance?

What information is not self-evident?

The format in which the modern NFL draft occurs does not appear to be the format in which it has always occurred. In the modern draft there are 7 rounds, but my dataset shows older drafts with over 10 rounds of picks. I need to determine whether I will have a large enough sample size by including data only from the modern draft era. If the sample size is not large enough, it may be necessary to first condense the rounds beyond 7 of older drafts by determining their equivalent valuation in the modern 7 round draft era.

What are different ways to look at this data?

Instead of combining player statistics to examine career performance, I could separate them by year and just include columns keeping a running total of their career statistics. Doing so would allow the measure of team success to better be compared against a kicker's performance, as well as allowing for kicker and team performance to be compared across multiple teams that a kicker may play for.

How can I summarize my data to answer key questions?

Mean, maximum, and minimum values would provide a great deal of insight into the distribution of kicker career/seasonal performance. Multiple regression analysis will be the best way in which to measure the importance of draft pick values and the impact field goals made/missed/percentage have on team success. Once the impact of field goals is known I can then measure draft pick value against the average expected kicker performance (field goals made/missed/percentage), which should provide an answer to the question of when is it best to draft a kicker?

Types of plots and tables I plan to use

As much of my information will be a result of multiple regression analysis, I expect the best way in which to present said results will be with scatter plots. Using the R2 value I can ensure that I am using the best fitting line for my scatter plots which, in conjunction with the other results of my regression analysis, will allow for the cleanest representation of the data. With the knowledge of average expected kicker performance and draft pick value, I can also cleanly illustrate any outliers by examining a residual vs fitted plot.

Machine Learning Techniques

Given the relatively small sample size, and static nature of the data, I do not currently plan to use any machine learning techniques with which I am familiar. If I am able to gain a better understanding of machine learning techniques that would be useful for this project, I would be happy to modify my work to include it.