Project Description

Pipe it Up! will use the Kaggle dataset "Social Power NBA" to explore the following research question: Is there a relationship between measures of basketball success (such as win percentage and offensive/defensive ratings) and Internet popularity, measured in the number of Twitter followers? If we find that there is a relationship between measures of basketball success and social media popularity, we plan to conduct further analysis to determine which of these measures is the best predictor of Internet popularity. Our hypothesis regarding this question is that there will be at least a weak positive correlation between basketball success (which will be discussed in the following paragraph) and the number of Twitter followers; in other words, players with relatively high offensive and defensive ratings will tend to also have more Twitter followers. We also predict that a player's salary, points scored, and win percentage will be particularly strong predictors of Twitter popularity.

Recently, researchers from many different fields have begun using data science techniques to extract information from social media data. A 2019 journal article found that scholars in the United States alone have published 7548 papers that incorporate social media data analysis (Esfahani et al., 2019). Our desire to explore this research question stems from the increasing relevance of social media in today's hyperconnected world. Social media provides a platform for people to share stories and opinions, and influential people, especially athletes, have considerable sway over large segments of the population. Additionally, since the NBA is a star-driven league due to its small team size and worldwide recognition, social media presence is significant for basketball players. The original creator of our Kaggle dataset, UC Davis and Northwestern professor Noah Gift, conducted analyses to determine which factors relating to basketball success social media and internet popularity could accurately predict. He searched for correlation between those factors and a number of response variables, including arena attendance, endorsements, salary, and NBA performance. Our team was fascinated by the statistician's conclusions, but we wanted to take the reverse approach and determine which factors (relating to basketball success) can predict the social media popularity that he claims is so important.

The dataset we wish to explore includes on-court performance data for NBA players in the 2016-2017 season, along with their salary, Twitter engagement, and Wikipedia traffic data. Because we are examining the relationship between player statistics and the number of Twitter followers, we decided to only consider players who had a Twitter account, by filtering for values where `TWITTER_HANDLE` is not 0 (NA). After filtering, we have 95 observations.

The data was originally collected from ESPN, Basketball-Reference, Twitter, Five-ThirtyEight, and Wikipedia. We assume that basketball players in the dataset were not chosen randomly because, while the creators did not specify how they chose which players to include, many (if not all) big name basketball players are included in the dataset. This led to us to assume the creators first chose the best basketball players and then selected "random" lesser-known (less talented) basketball players. Moreover, the players in the dataset are not all from the same team and the dataset includes at least one player from all 30 NBA teams.

While the sample was not collected randomly, there is a sufficiently large sample size (95) and all groups are represented with no team dominating the sample; thus, the dataset has independent observations. So, the number of Twitter followers of one player does not affect the number of Twitter followers of another player. The dataset also has independent groups, so the number of Twitter followers of a player from one team does not affect the number of Twitter followers of a player from another team.

As a note, due to the NBA's salary cap guidelines, the salaries of players on the same team influence each other; thus, we have some reservations about including `SALARY_MILLIONS` as a predictor.

More information about the dataset can be found at this link: https://www.kaggle.com/noahgift/social-power-nba

While the original dataset included 63 variables, we chose to only use 15 by selecting statistics we thought were indicative and all-encompassing of a player's basketball performance and Twitter followers since our research question attempts to predict the number of Twitter followers. We ignored variables, like `RANK`, because they are discrete numerical variables and instead selected continuous numerical variables, like percentage ratios, because they are more indicative of a player's skill (and we can use more advanced statistical methods). Below we have listed the variables we are planning to use (descriptions available in Data Dictionary):

<<<<<<< HEAD PLAYER_NAME: Player's name (categorical) TEAM_ABBREVIATION: Abbreviation for the team the player is on (categorical) AGE: Player age (quantitative) W_PCT: Percentage of games played won (quantitative) OFF_RATING: Player offensive rating calculated using the formula - Offensive Production Rating = (Points Produced / Individual Possessions) x OAPOW × PPG + FTM/FT * 3pt% + FG% (quantitative) DEF_RATING: Player defensive rating - Defensive Player Rating = (Players Steals*Blocks) + Opponents Differential= 1/5 of possessions - Times blown by + Deflections* OAPDW(Official Adjusted Players Defensive Withstand) (quantitative) NET_RATING: Average of the offensive/defensive rating (quantitative) AST_RATIO: Assists-to-turnovers ratio (quantitative) REB_PCT: Total rebounds (quantitative) USG_PCT: Usage percentage, an estimate of how often a player makes team plays (quantitative) PIE Player impact factor, a statistic roughly measuring a player's impact on the games that they play that's used by nba.com (quantitative) SALARY_MILLIONS: Salary in millions (quantitative) TWITTER_FOLLOWER_COUNT_MILLIONS: Number of Twitter followers (quantitative) ACTIVE_TWITTER_LAST_YEAR: Whether or not the player was active (posted) on Twitter last year. (categorical) PTS: Points scored (quantitative) ======= PLAYER_NAME TEAM_ABBREVIATION AGE W_PCT OFF_RATING DEF_RATING NET_RATING AST_RATIO REB_PCT USG_PCT PIE SALARY_MILLIONS TWITTER_FOLLOWER_COUNT_MILLIONS TWITTER_HANDLE PTS >>>>>>> 4fc25bdbce251b560f05a402e31eea7768bf7ca3

The Data

```r
library(tidyverse)
library(broom)
library(stringr)
library(knitr)
```

```r
nba_social_power <- read_csv("data/nba.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_integer(),
##   PLAYER_NAME = col_character(),
##   TEAM_ABBREVIATION = col_character(),
##   W_PCT = col_double(),
##   MIN = col_double(),
##   OFF_RATING = col_double(),
##   DEF_RATING = col_double(),
##   NET_RATING = col_double(),
##   AST_PCT = col_double(),
##   AST_TO = col_double(),
##   AST_RATIO = col_double(),
##   OREB_PCT = col_double(),
##   DREB_PCT = col_double(),
##   REB_PCT = col_double(),
##   TM_TOV_PCT = col_double(),
##   EFG_PCT = col_double(),
##   TS_PCT = col_double(),
##   USG_PCT = col_double(),
##   PACE = col_double(),
##   PIE = col_double(),
##   FGM_PG = col_double()
##   # ... with 8 more columns
```

```
## )
## See spec(...) for full column specifications.
```

Finally, we will examine the relevant variables in the dataset after removing players without Twitter handles (since social power cannot be easily measured for these players):

```r
nba_social_power_mod <- nba_social_power %>%
  filter(TWITTER_HANDLE != "0") %>%
  select(PLAYER_NAME,
         TEAM_ABBREVIATION,
         AGE,
         W_PCT,
         OFF_RATING,
         DEF_RATING,
         NET_RATING,
         AST_RATIO,
         REB_PCT,
         USG_PCT,
         PIE,
         SALARY_MILLIONS,
         ACTIVE_TWITTER_LAST_YEAR,
         TWITTER_FOLLOWER_COUNT_MILLIONS,
         PTS)

glimpse(nba_social_power_mod) # examine modified dataset for analysis
```

```
## Observations: 95
## Variables: 15
## $ PLAYER_NAME                     <chr> "Russell Westbrook", "Demetriu...
## $ TEAM_ABBREVIATION               <chr> "OKC", "BOS", "NOP", "HOU", "G...
## $ AGE                             <int> 28, 22, 24, 27, 28, 32, 32, 26...
## $ W_PCT                           <dbl> 0.568, 0.200, 0.413, 0.667, 0....
## $ OFF_RATING                      <dbl> 107.9, 124.2, 104.2, 113.6, 11...
## $ DEF_RATING                      <dbl> 104.6, 117.8, 102.5, 107.3, 10...
## $ NET_RATING                      <dbl> 3.3, 6.3, 1.7, 6.3, 16.0, 7.7,...
## $ AST_RATIO                       <dbl> 23.4, 31.1, 7.3, 27.6, 18.4, 2...
## $ REB_PCT                         <dbl> 0.167, 0.103, 0.170, 0.123, 0....
## $ USG_PCT                         <dbl> 0.408, 0.172, 0.326, 0.341, 0....
## $ PIE                             <dbl> 0.230, 0.194, 0.192, 0.190, 0....
## $ SALARY_MILLIONS                 <dbl> 26.54, 1.45, 22.12, 26.50, 26....
## $ ACTIVE_TWITTER_LAST_YEAR        <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, ...
## $ TWITTER_FOLLOWER_COUNT_MILLIONS <dbl> 4.500, 0.049, 1.220, 4.470, 16...
## $ PTS                             <dbl> 31.6, 2.0, 28.0, 29.1, 25.1, 2...
```

Because we are concerned with satisfying the independence assumption, we will fit a multiple linear regression model predicting the number of Twitter followers based on the 15 predictor variables we chose:

```r
model <- lm(TWITTER_FOLLOWER_COUNT_MILLIONS ~ AGE + OFF_RATING + DEF_RATING + PIE + REB_PCT + USG_PCT +
            data = nba_social_power_mod)
tidy(model)
```

```
## # A tibble: 10 x 5
##    term                     estimate std.error statistic p.value
##    <chr>                       <dbl>     <dbl>     <dbl>   <dbl>
##  1 (Intercept)              -36.4      16.0      -2.27    0.0257
##  2 AGE                        0.205     0.119     1.72    0.0889
```

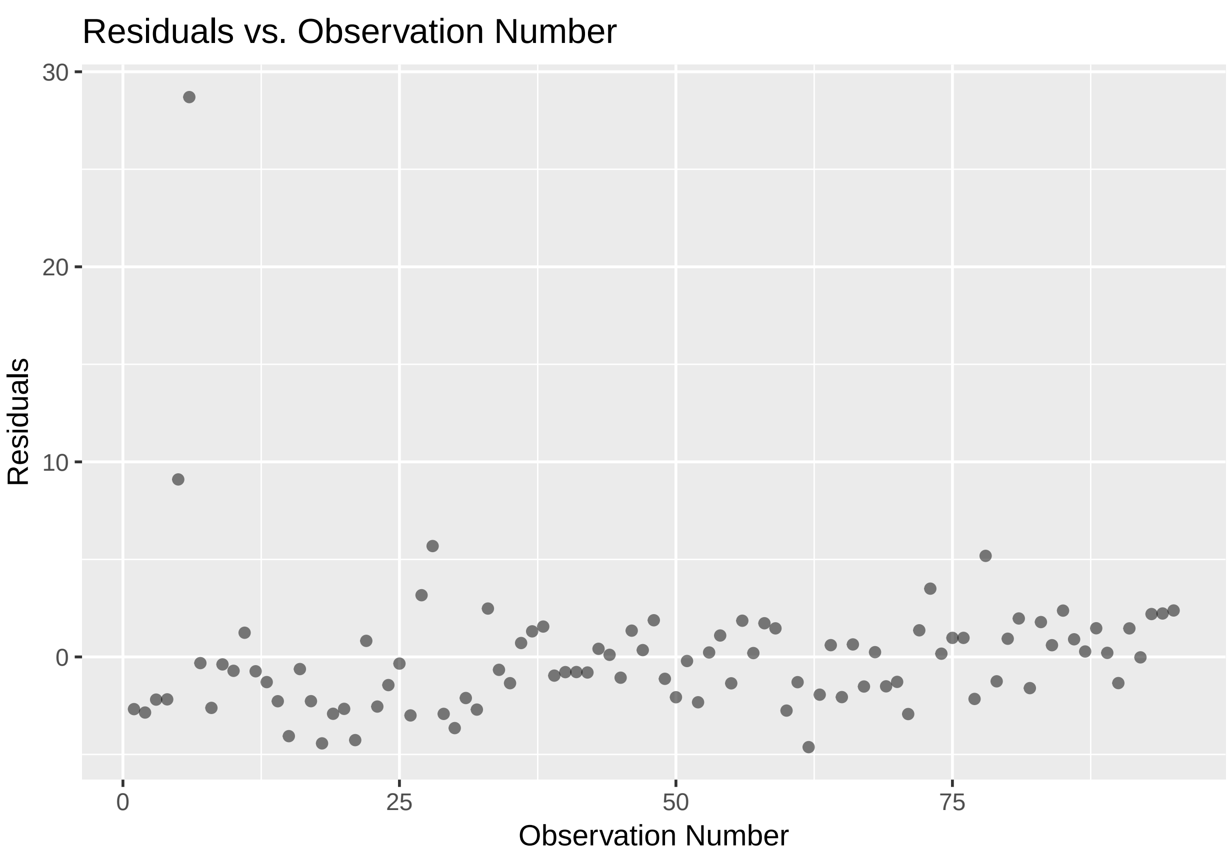
```
##   3 OFF_RATING                 0.103      0.0980     1.05     0.299
##   4 DEF_RATING                 0.0918     0.122      0.751    0.455
##   5 PIE                       35.6       25.2        1.41     0.161
##   6 REB_PCT                    2.12       8.62       0.247    0.806
##   7 USG_PCT                   -0.984     10.3       -0.0960   0.924
##   8 SALARY_MILLIONS            0.145      0.0577     2.50     0.0142
##   9 W_PCT                      3.04       3.50       0.868    0.388
## 10 ACTIVE_TWITTER_LAST_YEAR   3.47       2.84       1.22     0.225
```

The linear model is: -36.407 + 0.205 * `AGE` + 0.103 * `OFF_RATING` + 0.092 * `DEF_RATING` + 35.618 * `PIE` + 2.125 * `REB_PCT` - 0.984 * `USG_PCT` + 0.145 * `SALARY_MILLIONS` + 3.037 * `W_PCT` + 3.472 * `ACTIVE_TWITTER_LAST_YEAR`.

Now, we will look at a residual plot of observation number:

```
model_aug <- augment(model)
model_aug <- model_aug %>%
  mutate(obs_num = 1:nrow(model_aug))

ggplot(data = model_aug, aes(x = obs_num, y = .resid)) +
  geom_point(alpha = 0.5) +
  labs(x = "Observation Number",
       y = "Residuals",
       title = "Residuals vs. Observation Number")
```



It is clear, based on the plot displayed above, that there are some deviations from independence because the residuals are not "randomly" distributed.

Hence, our group has discussed the possibility of excluding certain high-profile players (e.g. Lebron James) to resolve this violation of independence. We would appreciate feedback about this aspect.

References

Esfahani, H. J., Tavasoli, K., & Jabbarzadeh, A. (2019). Big data and social media: A scientometrics analysis. International Journal of Data and Network Science, 145–164. doi: 10.5267/j.ijdns.2019.2.007