High Anticipation

Exploring Trends Between Public Perception and Player Value

Alana Willis, Fiona Dunn, and Sahana Rayan
October 25, 2020

Guiding the Research

~ What is the best metric for defining player performance?

~ How do we quantify public perception?

~ Do any players or subset of players stand out?

~ Can we predict public perception using player performance and vise versa?

Area of Focus

- Narrowed NBA players to 2018 first round draft class
  - All data had to be collected separately
  - Includes a randomized mix of players
    - International
    - Highly anticipated
    - Underrated

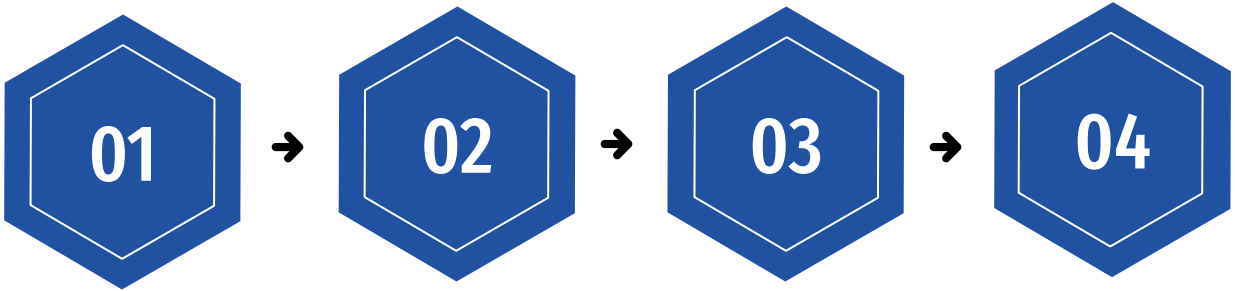
Constructing the Data Set

Reddit

Scraped Reddit user comments using RedditExtractoR

Game Data

Scraped Box score data from NBA stats page during the 2018 - 2019 season



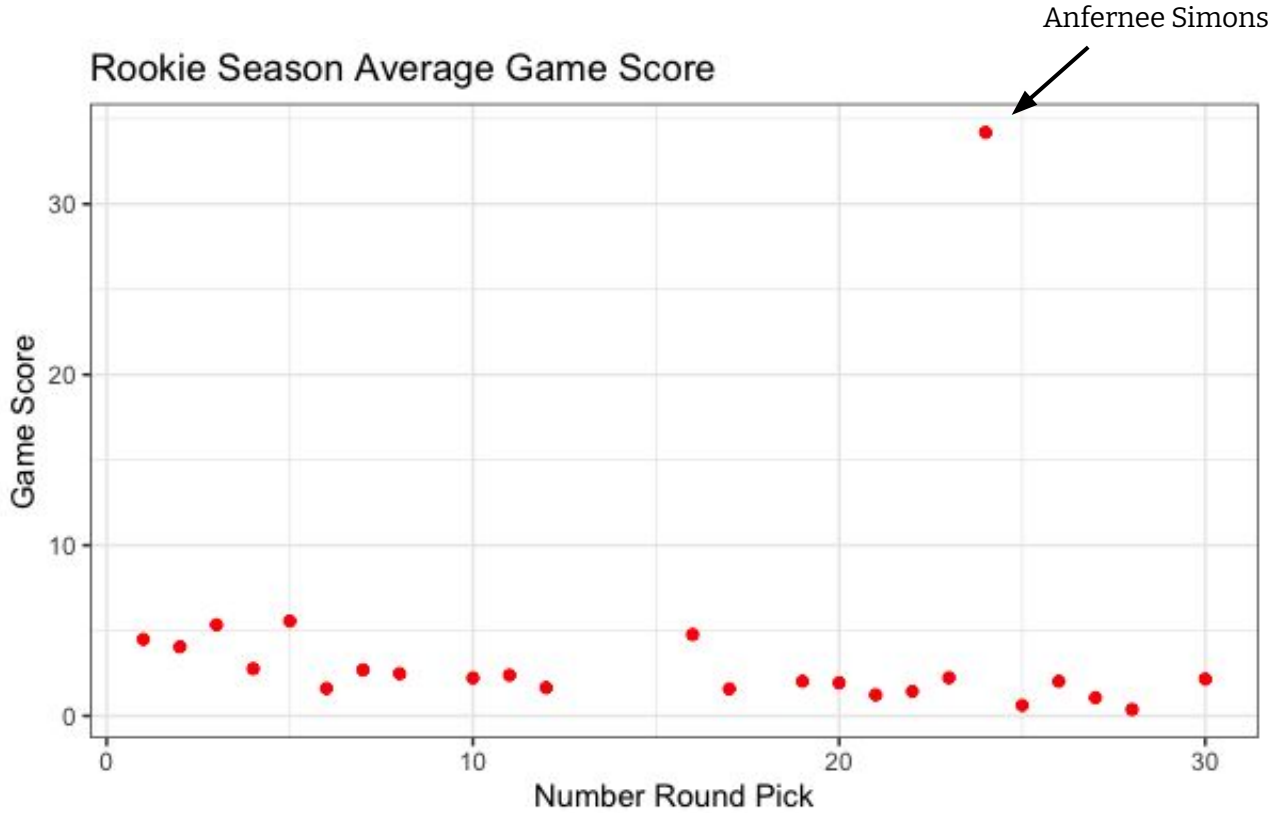
Web Hits

Scraped views from YouTube, Wikipedia, and Google (Images and News)

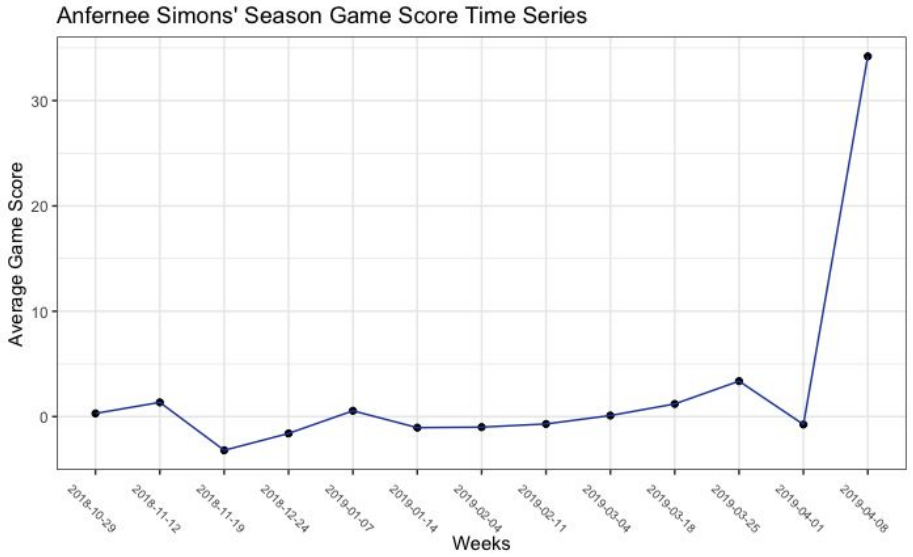
Combined

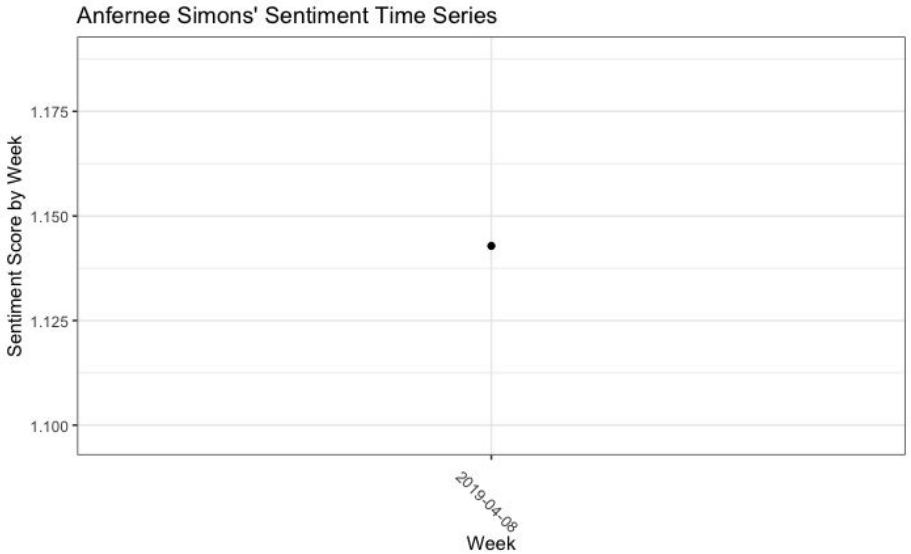
Rows aligned by weeks a player participated in at least one game over rookie season.

General EDA

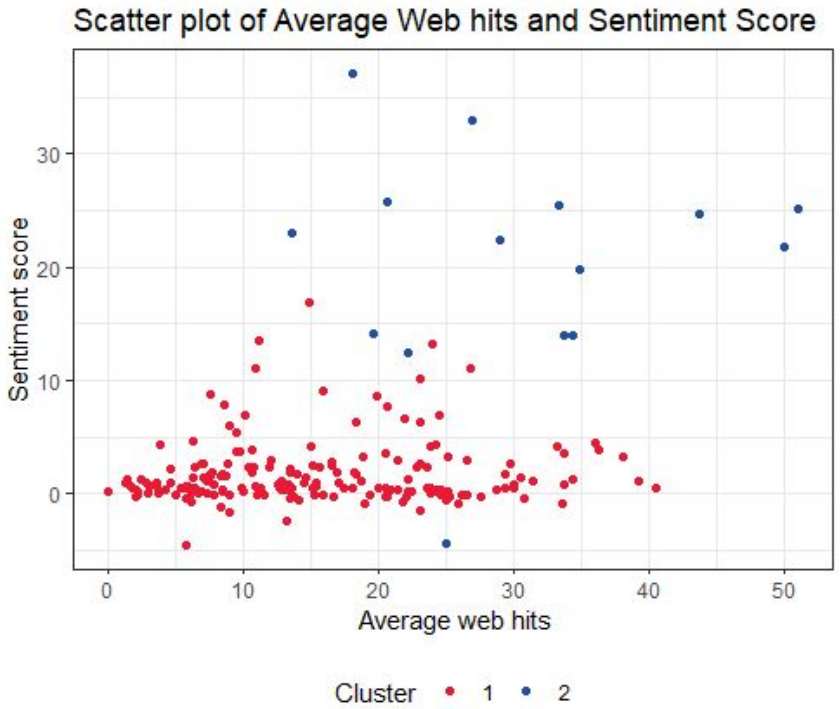

Anfernee Simons
Rookie Season Average Game Score
Game Score
0
10
20
30
0
10
20
30
Number Round Pick

Interesting Player Trends



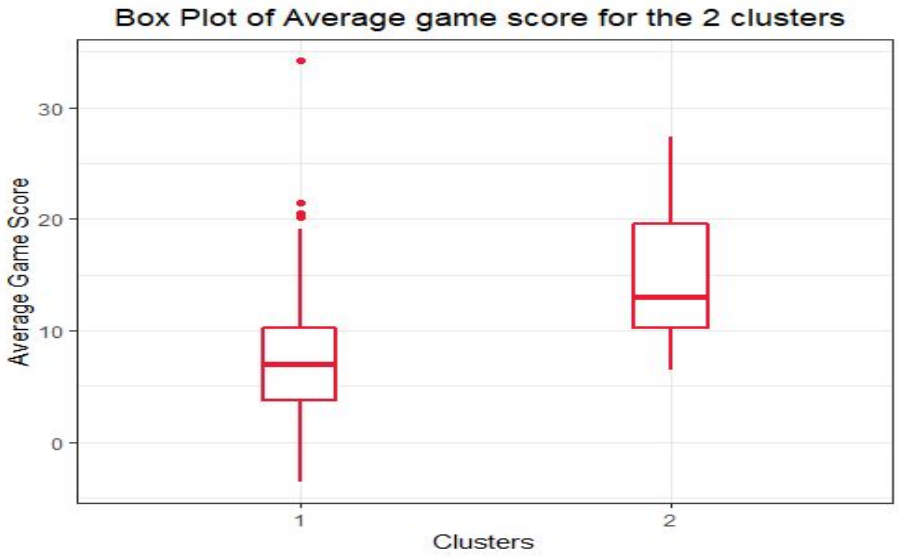


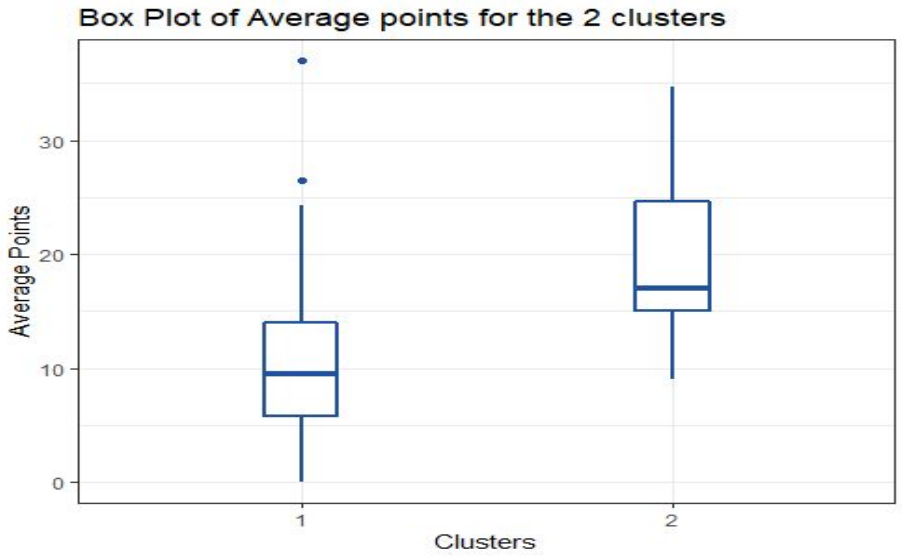
Clustering



- Hierarchical Clustering was done using positive score, negative score, Wiki views, Google web hits, Google news hits, and Youtube hits
- Cluster 1: Not very popular, low sentiment
- Cluster 2: A bit more popular, high sentiment

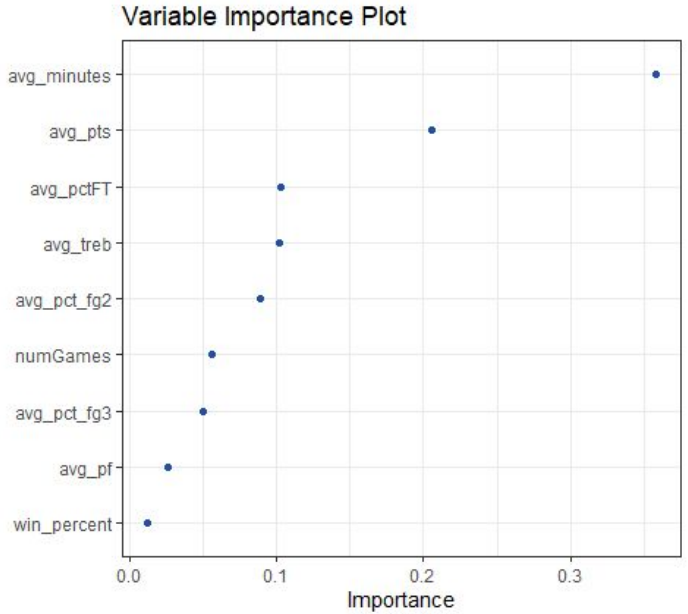
More EDA with the clusters



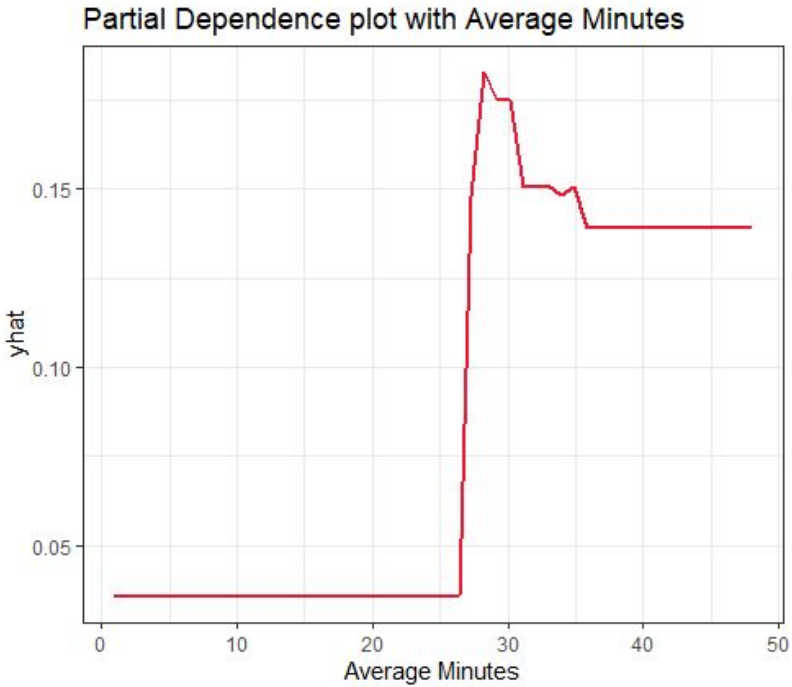


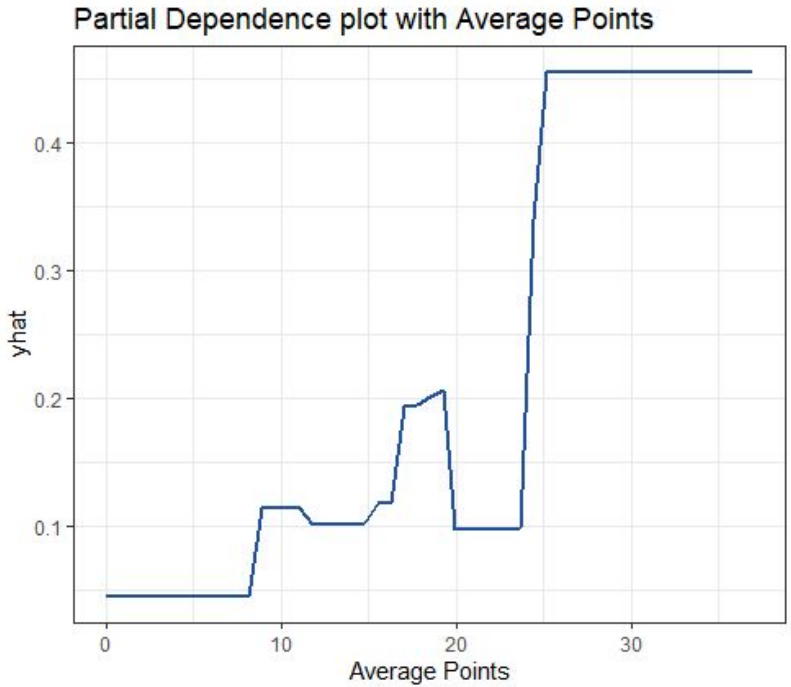
XGBoost: Predicting Clusters

- These clusters were used as a response variable for a binary classification problem.
- XGboost model was trained using AUC as an evaluation metric and the test predictions has an AUC of 0.887



Partial Dependence Plots





Final Thoughts

- Data wrangling and cleaning is HARD!
- There is a relationship between a player's performance, public sentiment, and overall popularity.
- One metric needs to be used to define public perception: Social Score and Clusters
- Average minutes played and average points scored are the most important in predicting popularity and sentiment clusters

Continuing the Research

- Introduce more variables to the models.
  - Reddit: comment score and controversiality
- Expand to more draft classes and to more social media/chat rooms
  - Other subreddits, blogs, Facebook, ect
- Explore specific comments in the Reddit data
- Predicting public perception onto future draft classes based on college or international play
- Principal Component Analysis with sentiment and popularity data

Contacts

Alana Willis

- Winston-Salem State University
- awillis117@rams.wssu.edu
- linkedin.com/in/alana-willis-7a8014188/

Fiona Dunn

- Kenyon College
- dunn2@kenyon.edu
- linkedin.com/in/fiona-dunn-706035138/

Sahana Rayan

- Purdue University
- srayan@purdue.edu
- linkedin.com/in/sahana-rayan/

References

Bresler, Alex. “NbastatR v0.1.1503.” *NbastatR Package | R Documentation*, www.rdocumentation.org/packages/nbastatR/versions/0.1.1503.

Game Score, www.nba.com/resources/static/team/v2/thunder/statlab-gamescore-191201.pdf .

Massicotte, Philippe. “GtrendsR.” *Function | R Documentation*, www.rdocumentation.org/packages/gtrendsR/versions/1.3.5/topics/gtrends.

Meissner, Peter. “Wikipediatrend v2.1.6.” *Wikipediatrend Package | R Documentation*, www.rdocumentation.org/packages/wikipediatrend/versions/2.1.6.

Rivera, Ivan. *Package ‘RedditExtractoR,’* 1 May 2019, cran.r-project.org/web/packages/RedditExtractoR/RedditExtractoR.pdf.

Silge, Julia, and David Robinson. *Text Mining with R*, 7 Mar. 2020, www.tidytextmining.com/index.html.

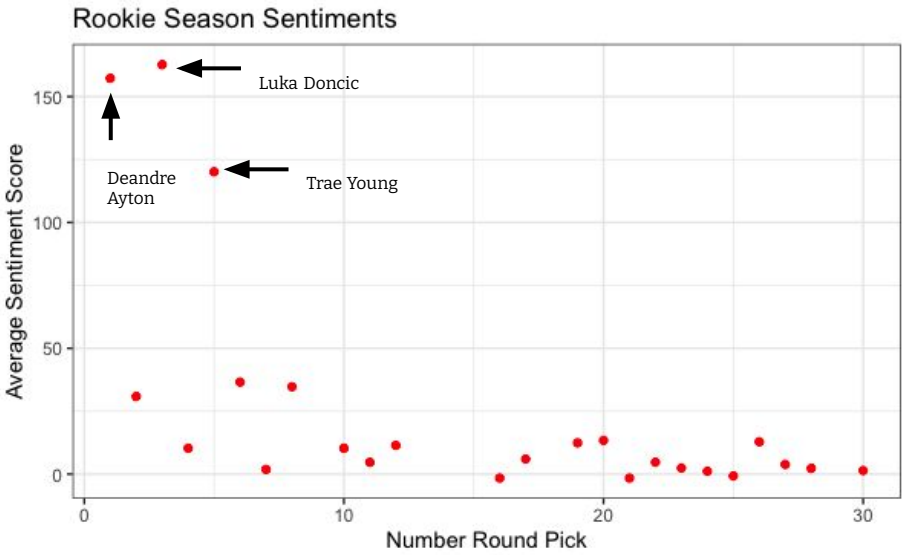
Thank You!

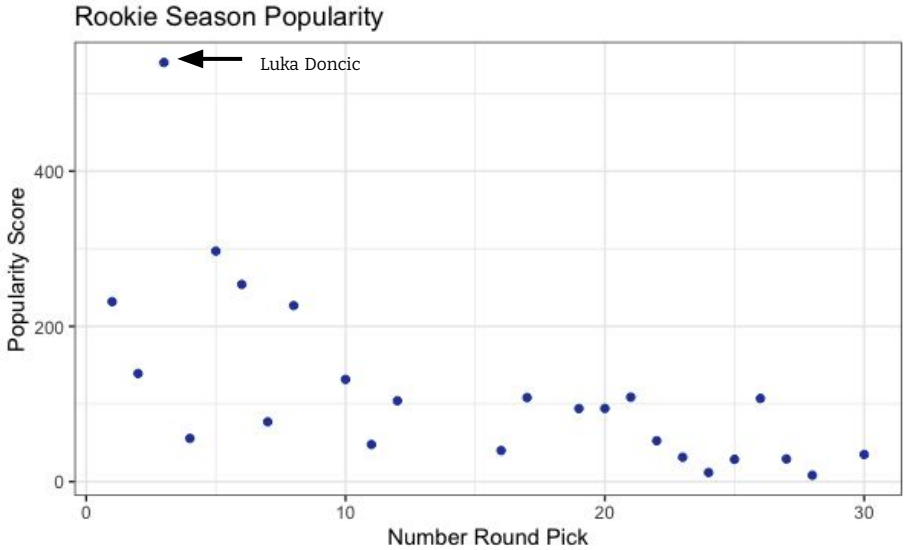
- Carnegie Mellon University
  - Rebecca Nugent
  - Ron Yurko
  - Beomjo Park
  - Pratik Patil
  - 2020 Cohort
- Atlanta Hawks
  - Maksim Horowitz

Variables of Interest

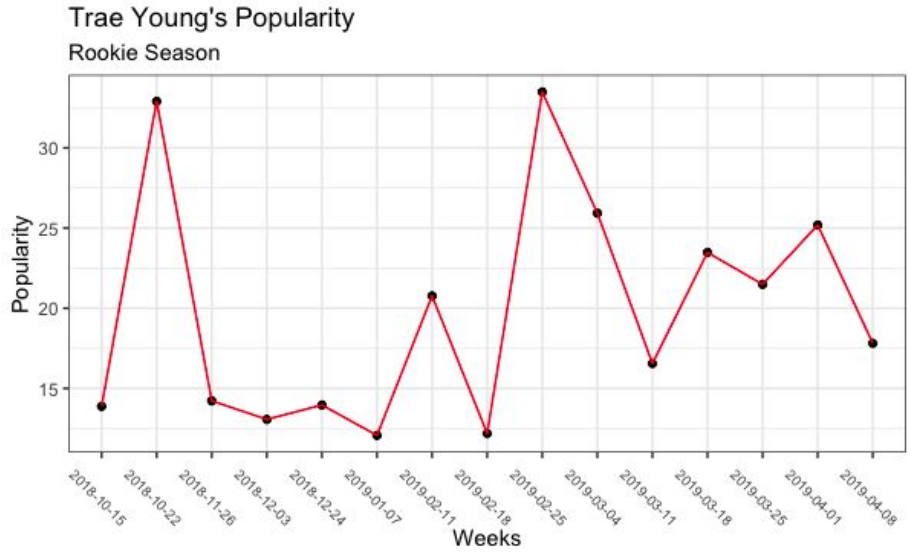
1. **Week:** Mon-Sun of a week in the regular season of the NBA
2. **Game Score:** a rough measure of a player's productivity for a single game
3. **Popularity** = (wiki_views + avg_web_hits + avg_image_hits + avg_news_hits + avg_yt_hits) / 5
4. **Social Score** = (sentiment_week + 5 * Popularity) / 6

General EDA

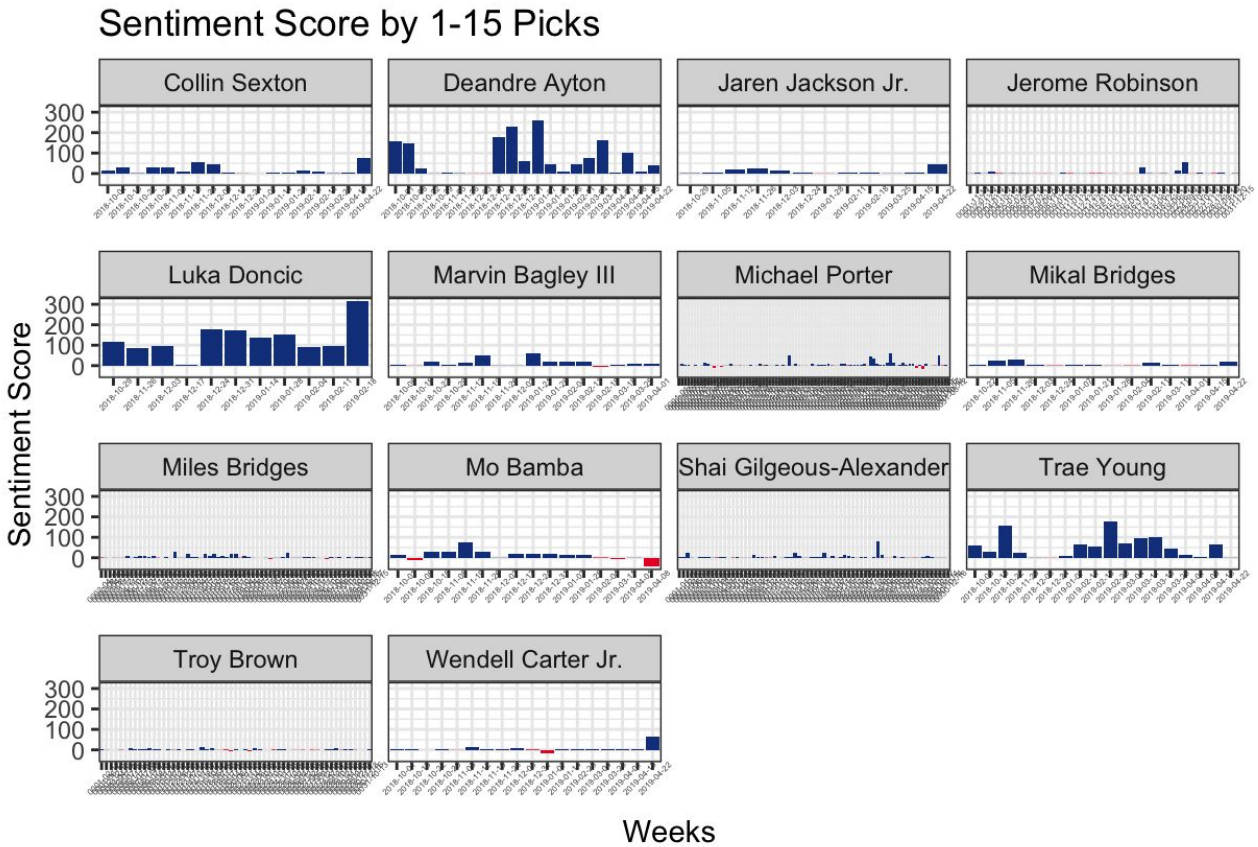

Rookie Season Sentiments
Luka Doncic
Deandre Ayton
Trae Young
Average Sentiment Score
150
100
50
0
0
10
20
30
Number Round Pick


Rookie Season Popularity
Luka Doncic
Popularity Score
400
200
0
0
10
20
30
Number Round Pick

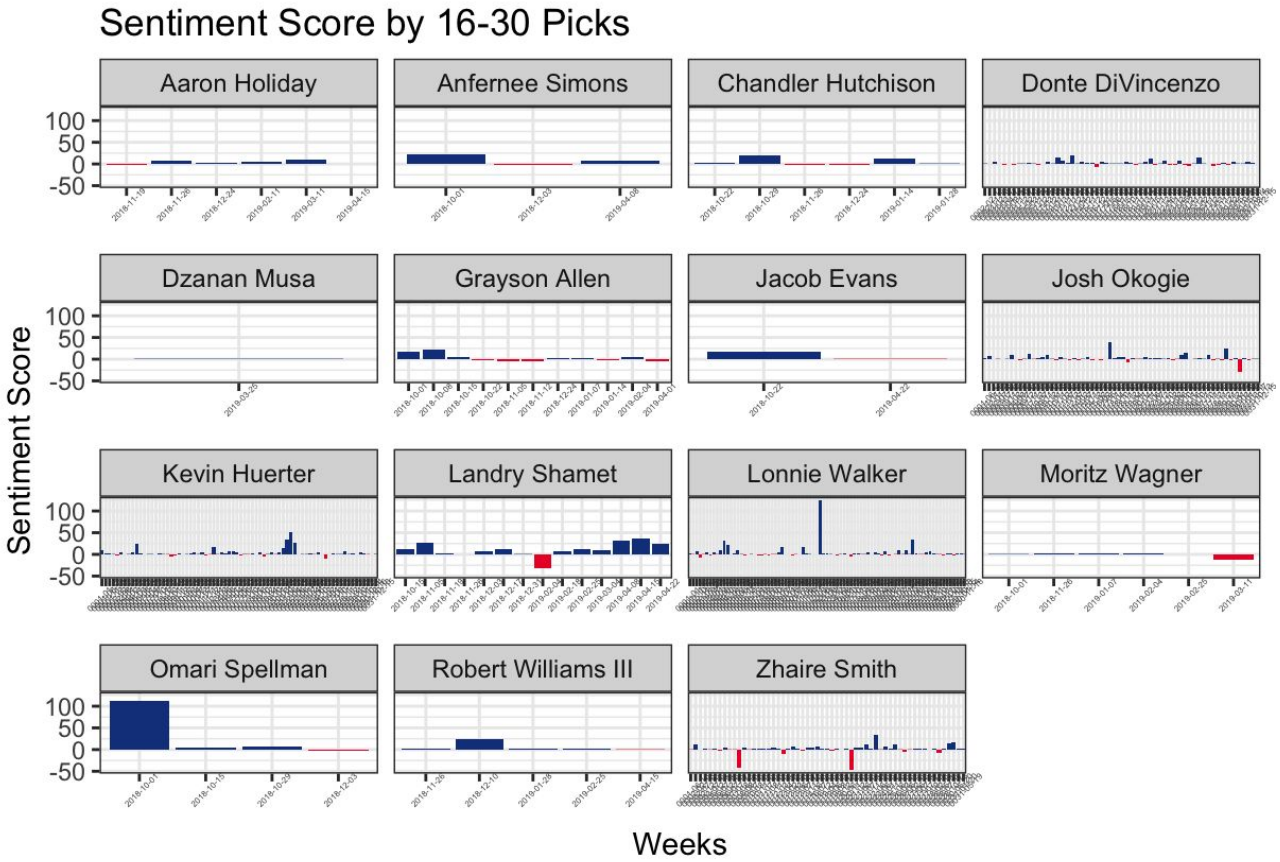
Interesting Player Trends



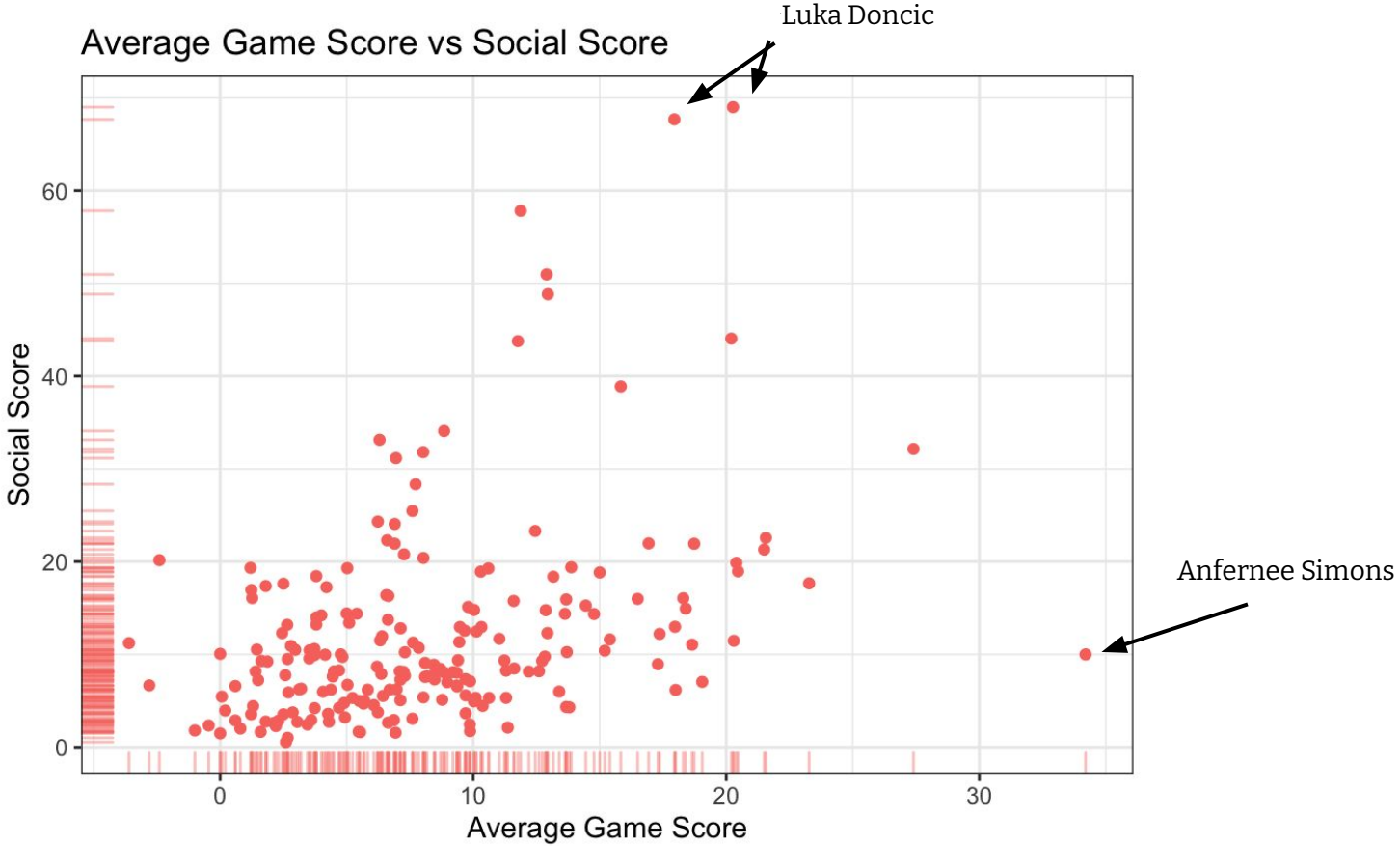
General EDA


Sentiment Score by 1-15 Picks
Collin Sexton
Deandre Ayton
Jaren Jackson Jr.
Jerome Robinson
Luka Doncic
Marvin Bagley III
Michael Porter
Mikal Bridges
Miles Bridges
Mo Bamba
Shai Gilgeous-Alexander
Trae Young
Troy Brown
Wendell Carter Jr.
Sentiment Score
300
200
100
0
Weeks

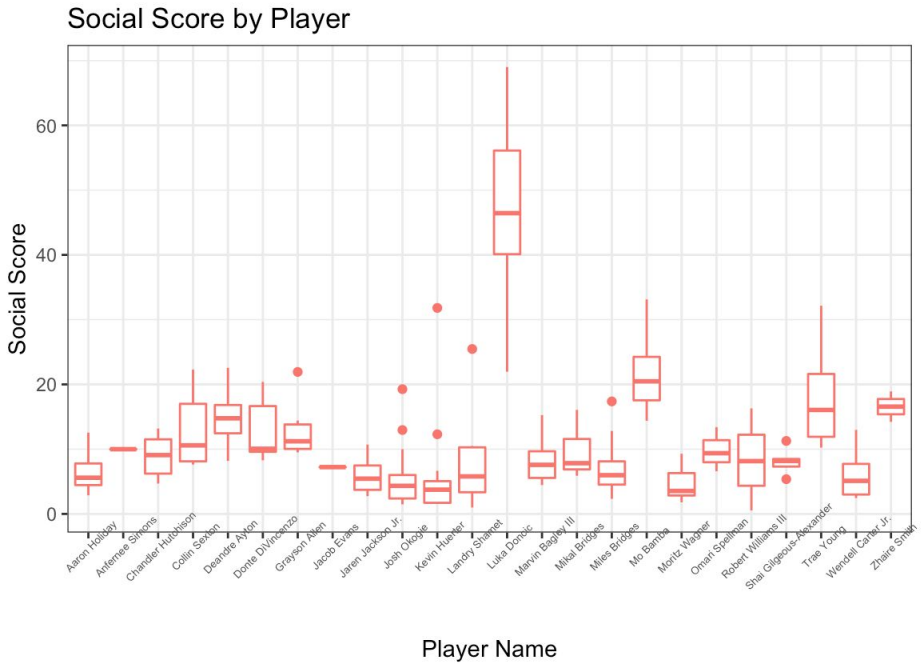
General EDA

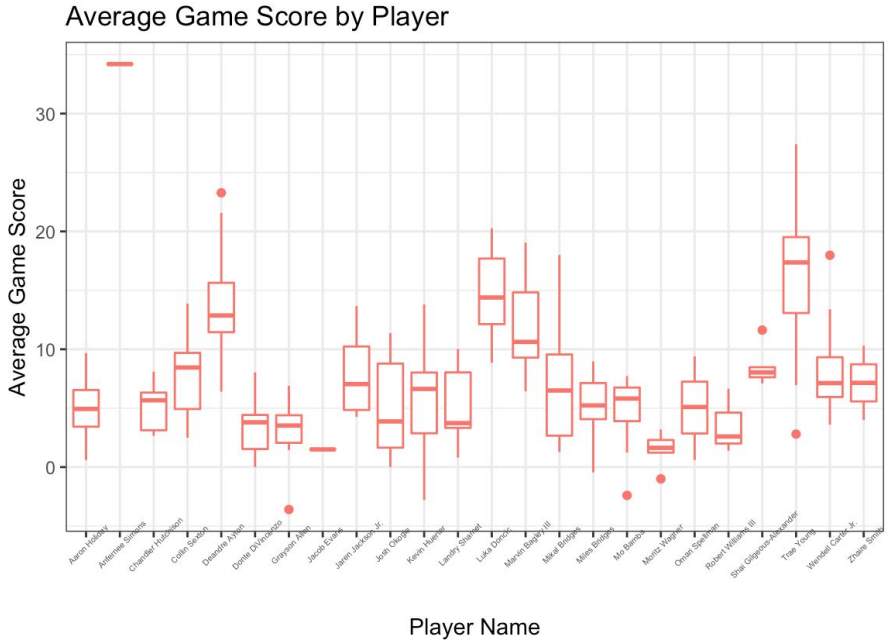

Sentiment Score by 16-30 Picks
Aaron Holiday
Anfernee Simons
Chandler Hutchison
Donte DiVincenzo
Dzanan Musa
Grayson Allen
Jacob Evans
Josh Okogie
Kevin Huerter
Landry Shamet
Lonnie Walker
Moritz Wagner
Omari Spellman
Robert Williams III
Zhaire Smith
Sentiment Score
100
50
0
-50
Weeks

General EDA

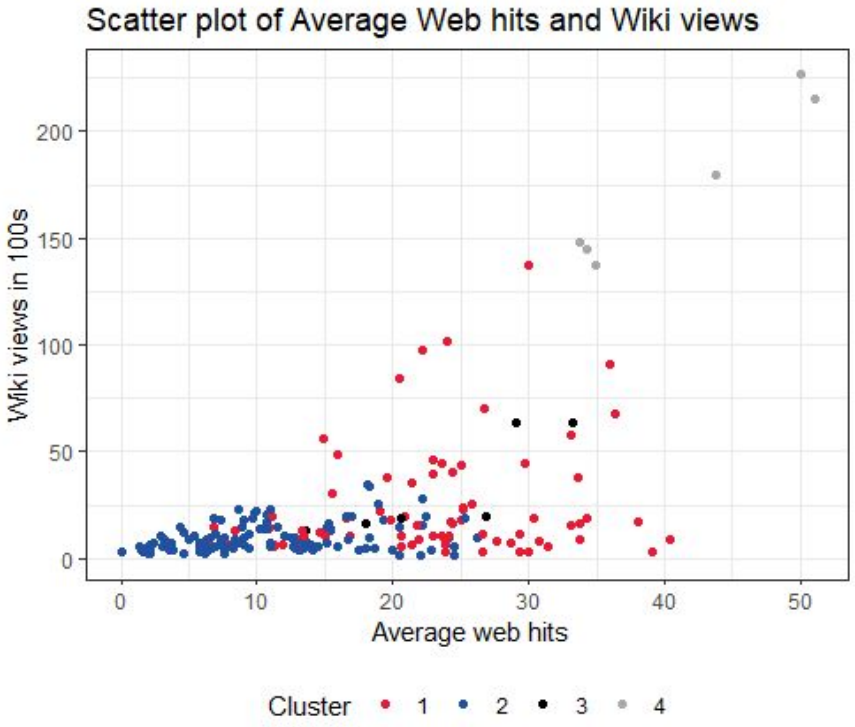

Average Game Score vs Social Score
Luka Doncic
Anfernee Simons
Social Score
0
20
40
60
0
10
20
30
Average Game Score

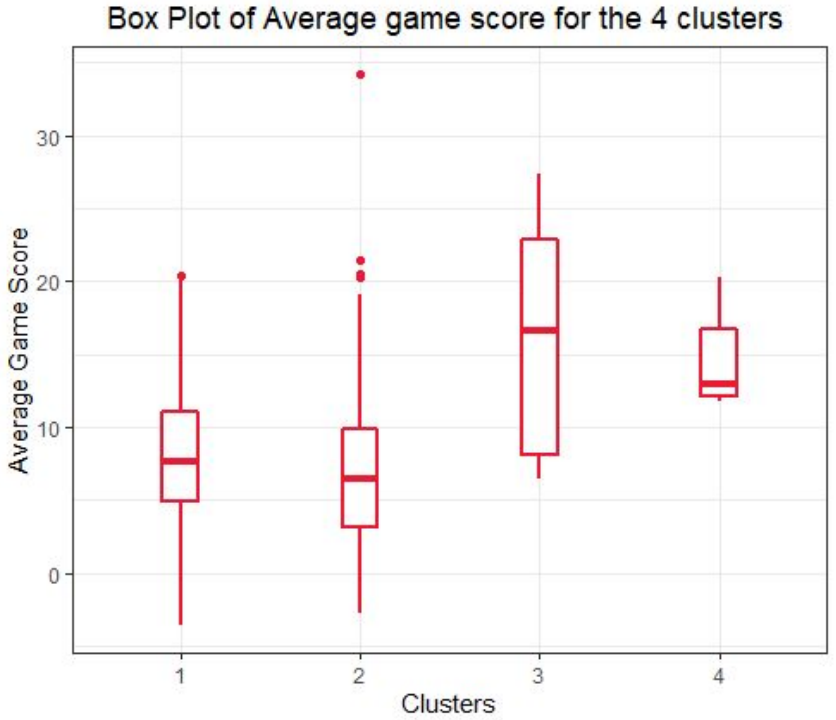
Interesting Player Trends



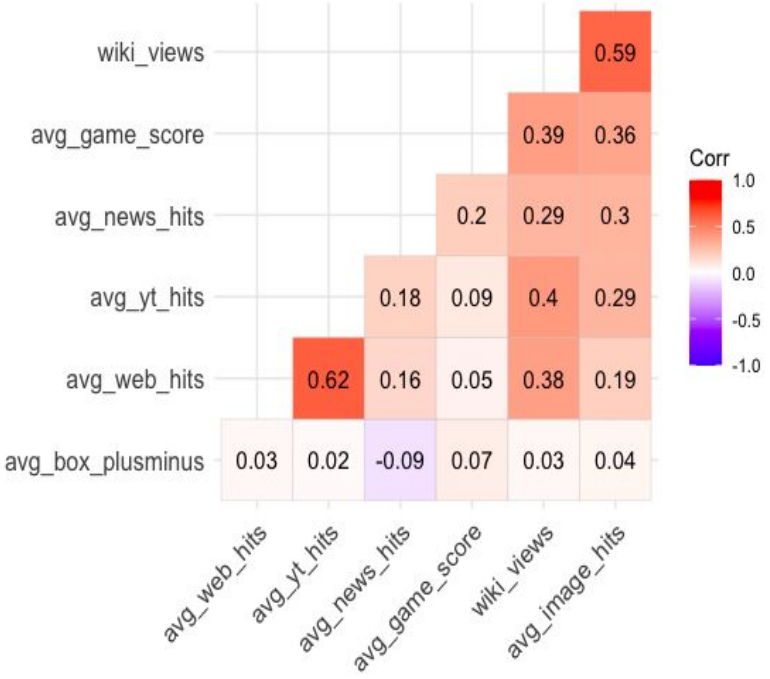


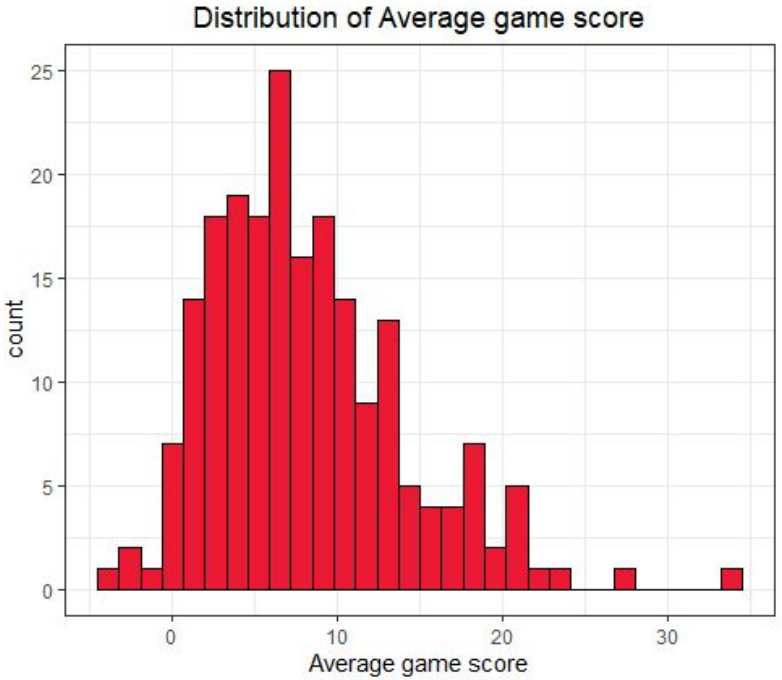
Clustering for Popularity Metrics



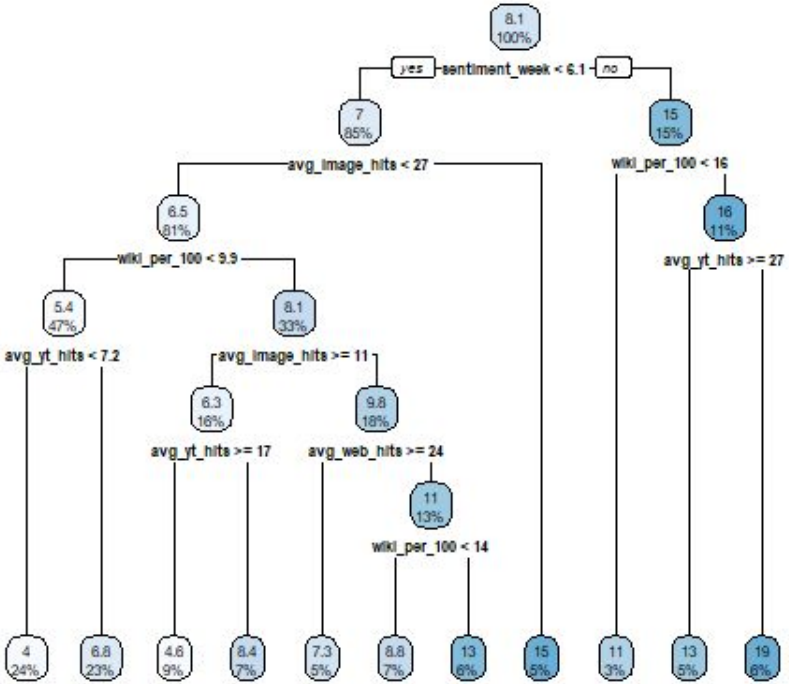


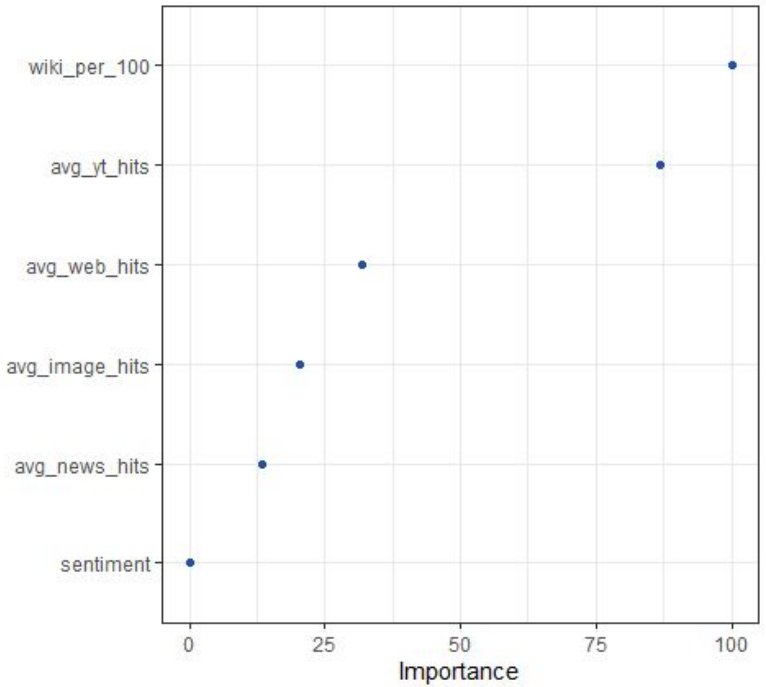
Before Modeling



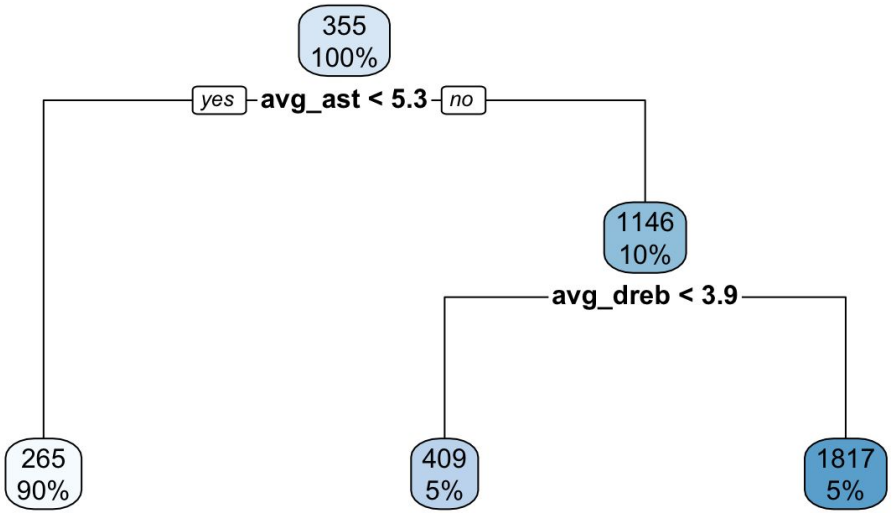


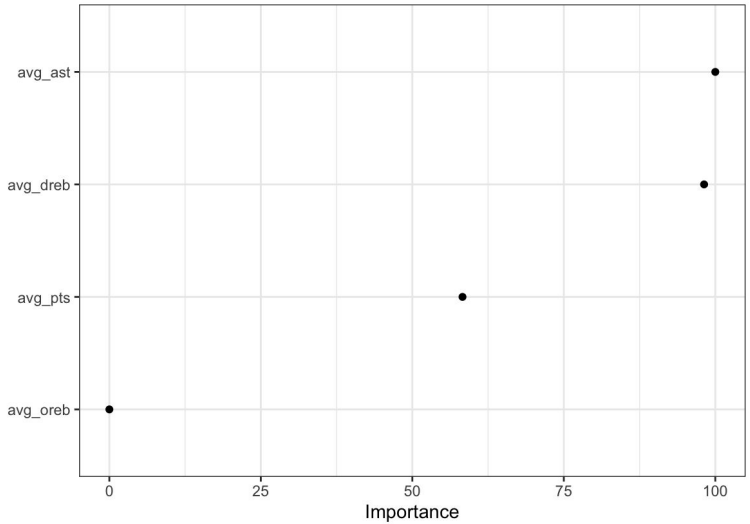
Decision Tree: Predicting Avg Game Score





Decision Tree: Predicting Social Score





In The Beginning

Initial Research Questions

How do players interact with their fan bases conditioned on the demographic of their home market?

Can we quantify "home town bias" among local media outlets for each NBA team?

Comparing public perception vs. actual player value (using player stats and social media data)

Packages

TidyCensus

NBAStatR

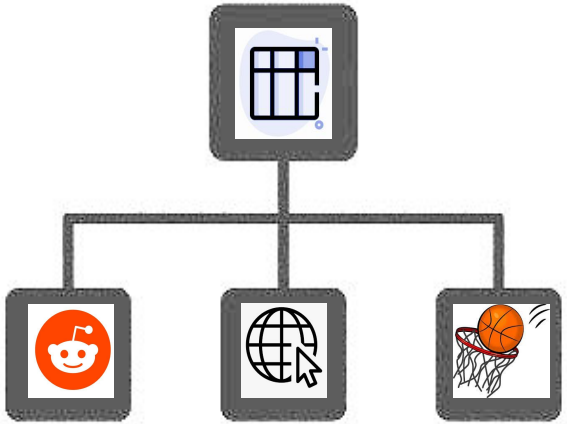
TwitteR

Stage 1 : Reddit Data

- Scraped all Reddit posts within subreddit r/nba using *RedditExtractoR*
  - Filtered the posts and comments to match the dates of the rookie season (~October 2018 to April 2019)
- Used bing sentiment analysis on each of the comments to create a sentiment score
  - Resulted in over 2 million rows
  - Grouped comments about each player by week to reduce rows to 296

Stage 4: Final Dataset

- The game data, Reddit data, and web data were combined to form a data set where each row's unique identifiers were the player name and the week
  - 206 observations by 25 variables
- This data set comprises of 3 types of data; Sentiment data (Reddit), popularity data (Wikipedia and Google trends), and basketball performance data (NBA stat)



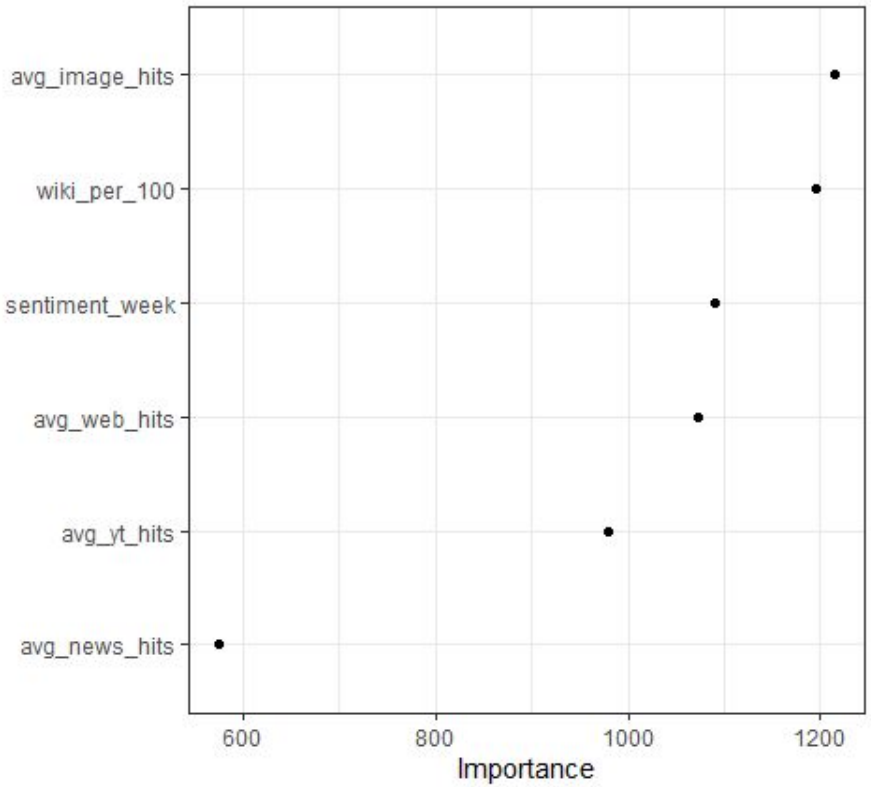
Stage 2 : Web Data

- Used *wikipediatrend* package to scrape daily Wikipedia page views for each player
  - Filtered by the dates of the rookie season (~October 2018 to April 2019)
- Used *gtrends* package to scrape daily Google image hits, web hits, news hits and YouTube hits for each player
  - Filtered by the dates of the rookie season (~October 2018 to April 2019)
  - Google trends assigns scores from 0 to 100

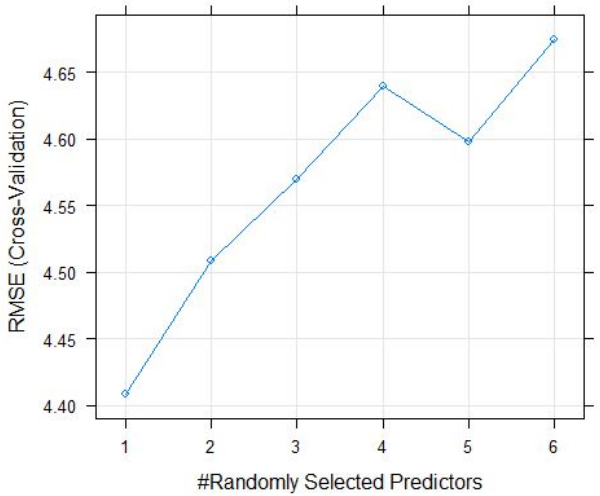
Stage 3: Game Data

- Scraped box score data from the NBA stats page
  - Only looking at 2018-2019 season for players drafted in top 30
- Each row in this data showed the player's performance for a game played in the season
- Game score: a measure of player performance in a given game

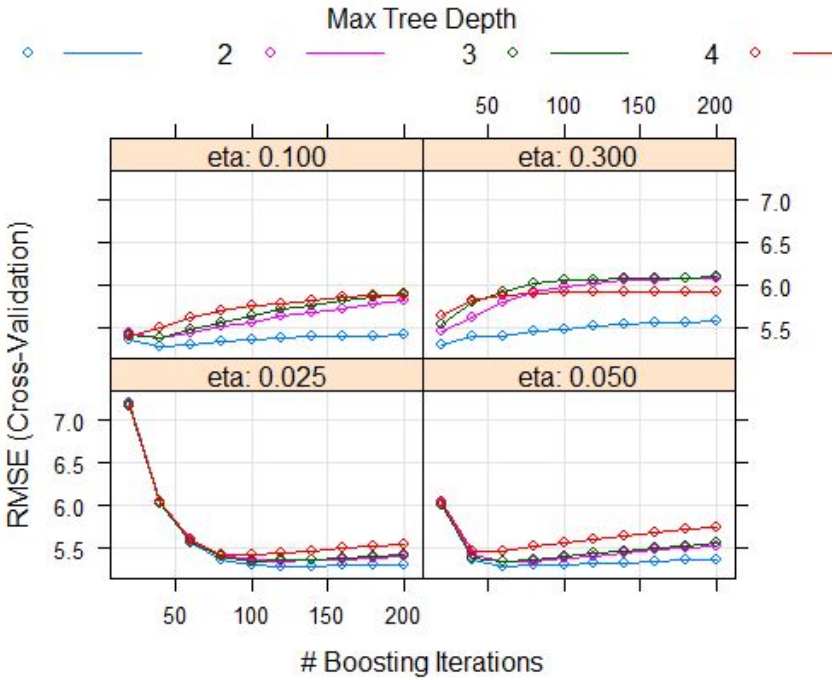
Random Forest: Predicting Game Score



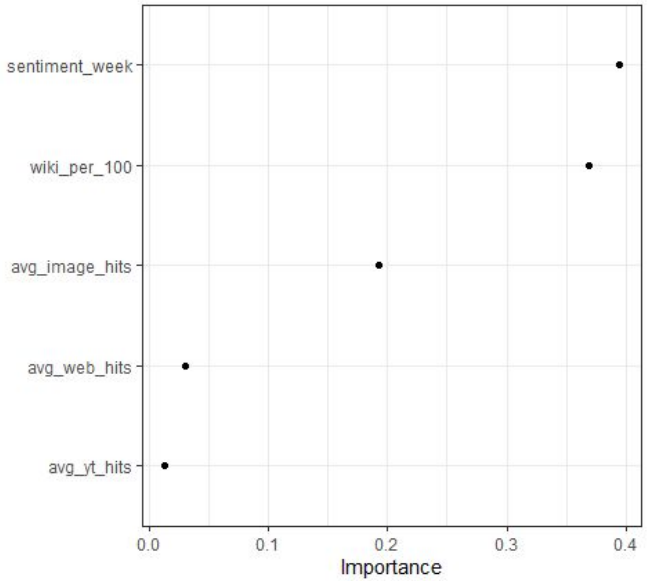
- N_trees = 40 - 80 with increments of 5
- M_try = 1 to 6
- Min_node_size = 3, 5



XGBoost: Predicting Avg Game Score



Best Tuning Parameters:
nrounds = 120, eta = 0.025, max tree depth = 1



Clustering for Sentiment Metrics