Using R and the Tidyverse to Play Fantasy Baseball

Angeline Protacio
Protacio Analytics, LLC
July 12, 2019
useR! 2019 – Toulouse, France
Real and fantasy baseball

In real baseball, only runs matter!

In fantasy baseball, every statistic (including runs) matters!

### Team Comparisons

<table>
<thead>
<tr>
<th>TEAM</th>
<th>R</th>
<th>HR</th>
<th>RBI</th>
<th>SB</th>
<th>AVG</th>
<th>OPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rizzo my Winker</td>
<td>34</td>
<td>11</td>
<td>36</td>
<td>1</td>
<td>.254</td>
<td>.791</td>
</tr>
<tr>
<td>Dropped Third Strike</td>
<td>44</td>
<td>13</td>
<td>34</td>
<td>2</td>
<td>.303</td>
<td>.956</td>
</tr>
</tbody>
</table>
Key questions

1. Which positions contribute most to scoring categories?

2. How do I rank players based on their projected contributions?

Using R to draft your fantasy team

```r
library(dplyr)
library(purrr)

pos_files <- c("data/depth_1B.csv", "data/depth_2b.csv", "data/depth_3b.csv",
               "data/depth_SS.csv", "data/depth_OF.csv",
               "data/depth_C.csv")
pos_names <- c("first_base", "second_base", "third_base", "short", "outfield",
               "catcher")
batters <- map2_df(pos_files, pos_names, load_data) %>%
            select(Name, Team, playerid, position, PA, R, HR, RBI, SB, OPS, SO, WAR)
```
Home runs by position

```r
library(ggplot2)
library(wesanderson)

ggplot(batters, aes(x = position, y = HR, fill = position)) +
  geom_violin() +
  scale_fill_manual(values = wes_palette("FantasticFox1", 6, type = "continuous")) +
  theme_bw() +
  theme(legend.position="none") +
  scale_y_continuous(limits = c(0, 60), expand = c(0, 0))
```
library(ggplot2)
library(wesanderson)

ggplot(batters, aes(position, HR, fill = position)) +
  geom_violin() +
  scale_fill_manual(values = wes_palette("FantasticFox1", 6, type = "continuous")) +
  theme_bw() +
  theme(legend.position="none") +
  scale_y_continuous(limits = c(0, 60), expand = c(0, 0))
Runs by position

```r
ggplot(batters, aes(position, R, fill = position)) +
  geom_violin() +
  scale_fill_manual(values = wes_palette("FantasticFox1", 6, type = "continuous")) +
  theme_bw() +
  theme(legend.position="none")
+ scale_y_continuous(limits = c(0, 125), expand = c(0, 0))
```
Runs batted in (RBI) by position

ggplot(batters, aes(position, RBI, fill = position)) + geom_violin() + scale_fill_manual(values = wes_palette("FantasticFox1", 6, type = "continuous")) + theme_bw() + theme(legend.position="none") + scale_y_continuous(limits = c(0, 150), expand = c(0, 0))
Key questions

1. Which positions contribute more to winning scoring categories?
   - Outfielders
   - First and third basemen
   - Definitely not catchers

2. How do I rank players based on their projected contributions?
Z-scores, explained

<table>
<thead>
<tr>
<th>position</th>
<th>Name</th>
<th>HR</th>
<th>HR_z</th>
</tr>
</thead>
<tbody>
<tr>
<td>outfield</td>
<td>Giancarlo Stanton</td>
<td>58</td>
<td>4.391960</td>
</tr>
<tr>
<td>first_base</td>
<td>Joey Gallo</td>
<td>42</td>
<td>2.612658</td>
</tr>
<tr>
<td>third_base</td>
<td>Joey Gallo</td>
<td>42</td>
<td>2.612658</td>
</tr>
<tr>
<td>outfield</td>
<td>Aaron Judge</td>
<td>41</td>
<td>2.501452</td>
</tr>
<tr>
<td>first_base</td>
<td>Cody Bellinger</td>
<td>39</td>
<td>2.279039</td>
</tr>
<tr>
<td>third_base</td>
<td>Nolan Arenado</td>
<td>39</td>
<td>2.279039</td>
</tr>
<tr>
<td>outfield</td>
<td>Mike Trout</td>
<td>39</td>
<td>2.279039</td>
</tr>
<tr>
<td>outfield</td>
<td>J.D. Martinez</td>
<td>39</td>
<td>2.279039</td>
</tr>
<tr>
<td>outfield</td>
<td>Cody Bellinger</td>
<td>39</td>
<td>2.279039</td>
</tr>
</tbody>
</table>
Generating z-scores

bat_z <- batters %>%
  filter(PA >= 300) %>%
  select(playerid, position, Name, Team, R, HR, RBI, SO, SB, OPS, WAR) %>%
  mutate(  
    R_z = z_score(R),
    HR_z = z_score(HR),
    RBI_z = z_score(RBI),
    SO_z = -z_score(SO),
    SB_z = z_score(SB),
    OPS_z = z_score(OPS),
    tot_z = round((R_z + HR_z + RBI_z + SO_z + SB_z + OPS_z), 3)
  )

z_score <- function(stat) {
  (stat - mean(stat, na.rm = TRUE))/sd(stat, na.rm = TRUE)
}
Top five players by total z-score

bat_z %>%
  top_n(., 5, tot_z) %>%
  arrange(desc(tot_z)) %>%
  select(position, Name, R, HR, RBI, SO, SB, OPS, WAR, tot_z)

<table>
<thead>
<tr>
<th>position</th>
<th>Name</th>
<th>R</th>
<th>HR</th>
<th>RBI</th>
<th>SO</th>
<th>SB</th>
<th>OPS</th>
<th>WAR</th>
<th>tot_z</th>
</tr>
</thead>
<tbody>
<tr>
<td>outfield</td>
<td>Mike Trout</td>
<td>114</td>
<td>39</td>
<td>105</td>
<td>131</td>
<td>22</td>
<td>1.027</td>
<td>8.2</td>
<td>11.889</td>
</tr>
<tr>
<td>outfield</td>
<td>Giancarlo Stanton</td>
<td>109</td>
<td>58</td>
<td>140</td>
<td>171</td>
<td>2</td>
<td>1.029</td>
<td>6.4</td>
<td>11.869</td>
</tr>
<tr>
<td>third_base</td>
<td>Nolan Arenado</td>
<td>97</td>
<td>39</td>
<td>118</td>
<td>101</td>
<td>3</td>
<td>0.937</td>
<td>5.0</td>
<td>8.766</td>
</tr>
<tr>
<td>outfield</td>
<td>Bryce Harper</td>
<td>100</td>
<td>35</td>
<td>102</td>
<td>122</td>
<td>10</td>
<td>0.984</td>
<td>5.6</td>
<td>8.646</td>
</tr>
<tr>
<td>first_base</td>
<td>Anthony Rizzo</td>
<td>97</td>
<td>34</td>
<td>107</td>
<td>98</td>
<td>9</td>
<td>0.927</td>
<td>4.7</td>
<td>8.343</td>
</tr>
</tbody>
</table>
Key questions – answered!

1. Which positions contribute more to winning scoring categories?
   • Outfielders
   • First and third basemen
   • Definitely not catchers !!

2. How do I rank players based on their projected contributions?
   • If a good index metric doesn’t exist, create your own
     (I used z-scores)

Also important: Draft based on data, not fandom!
Thanks, tidyverse!

Ideal for:
- Data cleaning
- Exploratory analysis
- Plotting your data
- Others adapting your analysis
Contact me!

Website: www.angelineprotacio.com
Twitter: @dataangeline
Email: email@angelineprotacio.com
Code: https://github.com/angelinepro/useR_july2019