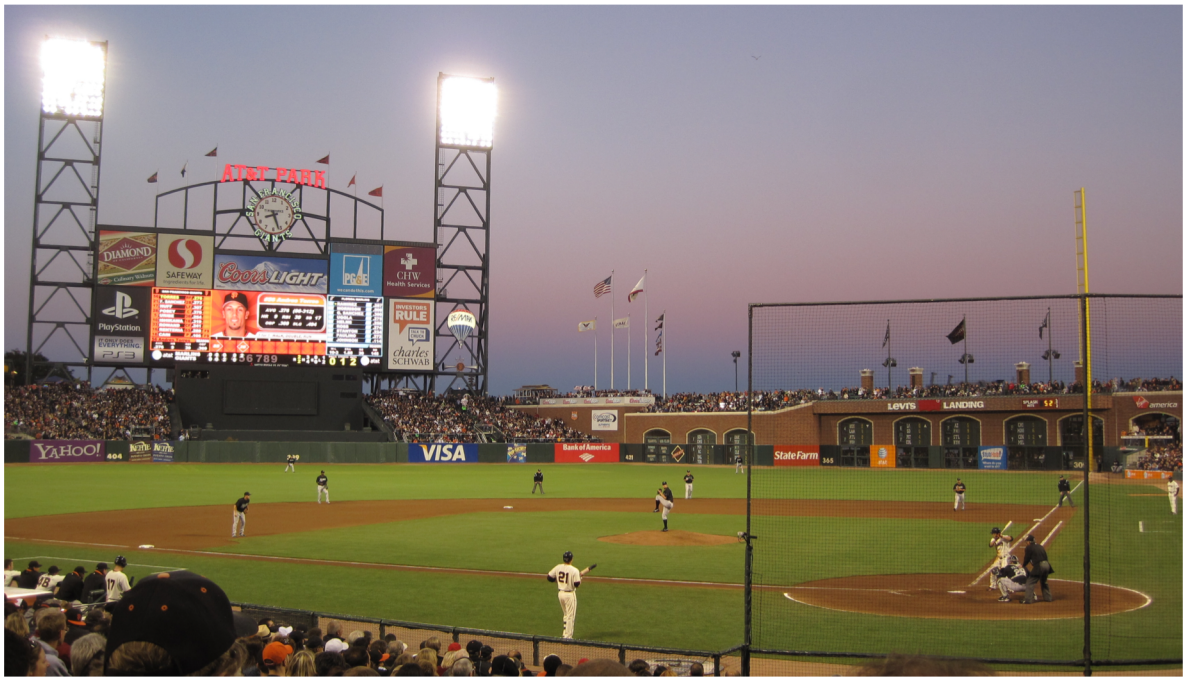
Using R and the Tidyverse to Play Fantasy Baseball



Angeline Protacio
Protacio Analytics, LLC
July 12, 2019
useR! 2019 – Toulouse, France

Real and fantasy baseball

In real baseball, only runs matter!

San Francisco Giants
(31 - 39)

3 - 2

Los Angeles Dodgers
(48 - 25)

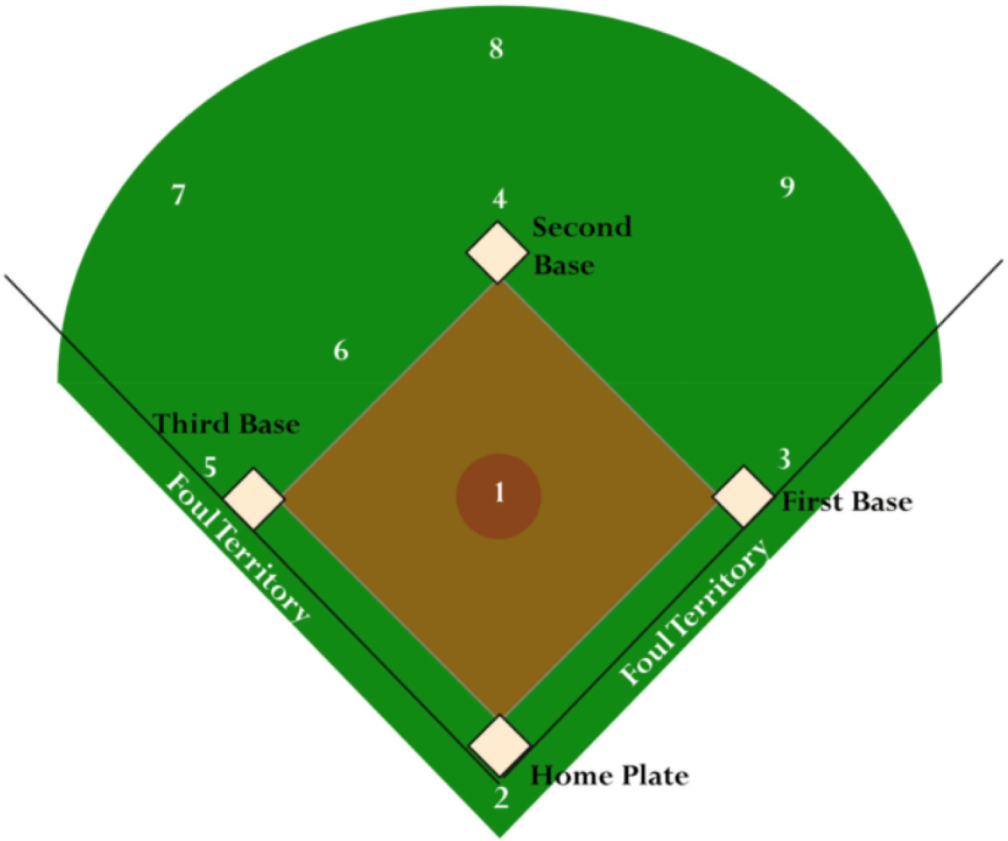
In fantasy baseball, every statistic (including runs) matters!

TEAM	R	HR	RBI	SB	AVG	OPS
Rizzo my Winker	34	11	36	1	.254	.791
Dropped Third Strike	44	13	34	2	.303	.956

Key questions

1. Which positions contribute most to scoring categories?

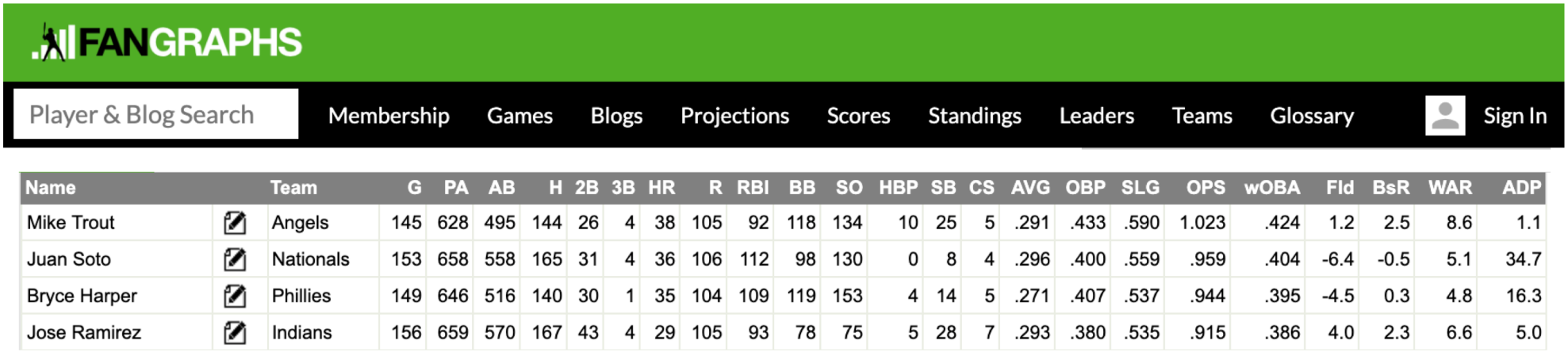
2. How do I rank players based on their projected contributions?



https://www.researchgate.net/figure/FIGURE-A1-Diagram-of-baseball-diamond-courtesycliparthutcom_fig11_296688489

Using R to draft your fantasy team

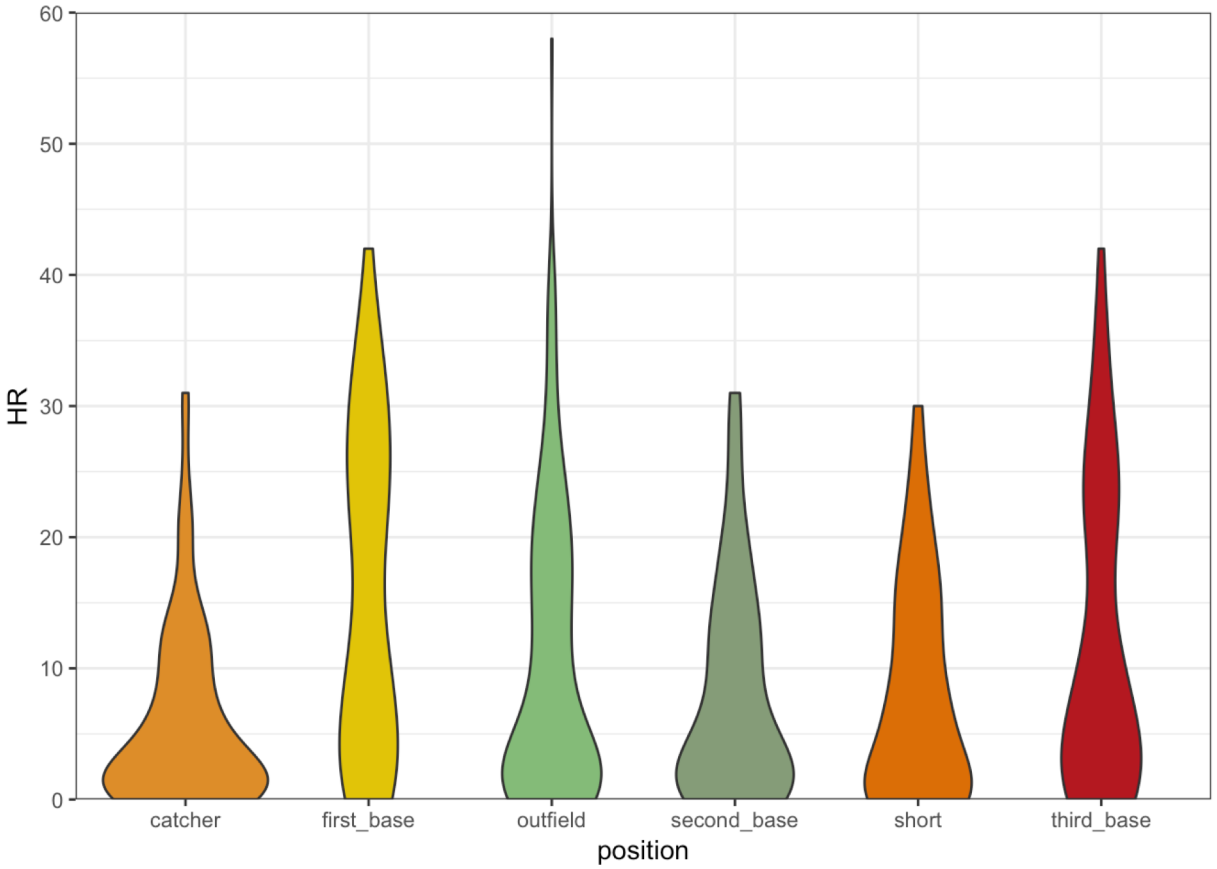
```
library(dplyr)
library(purrr)

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)
```



Name		Team	G	PA	AB	H	2B	3B	HR	R	RBI	BB	SO	HBP	SB	CS	AVG	OBP	SLG	OPS	wOBA	Fld	BsR	WAR	ADP
Mike Trout		Angels	145	628	495	144	26	4	38	105	92	118	134	10	25	5	.291	.433	.590	1.023	.424	1.2	2.5	8.6	1.1
Juan Soto		Nationals	153	658	558	165	31	4	36	106	112	98	130	0	8	4	.296	.400	.559	.959	.404	-6.4	-0.5	5.1	34.7
Bryce Harper		Phillies	149	646	516	140	30	1	35	104	109	119	153	4	14	5	.271	.407	.537	.944	.395	-4.5	0.3	4.8	16.3
Jose Ramirez		Indians	156	659	570	167	43	4	29	105	93	78	75	5	28	7	.293	.380	.535	.915	.386	4.0	2.3	6.6	5.0

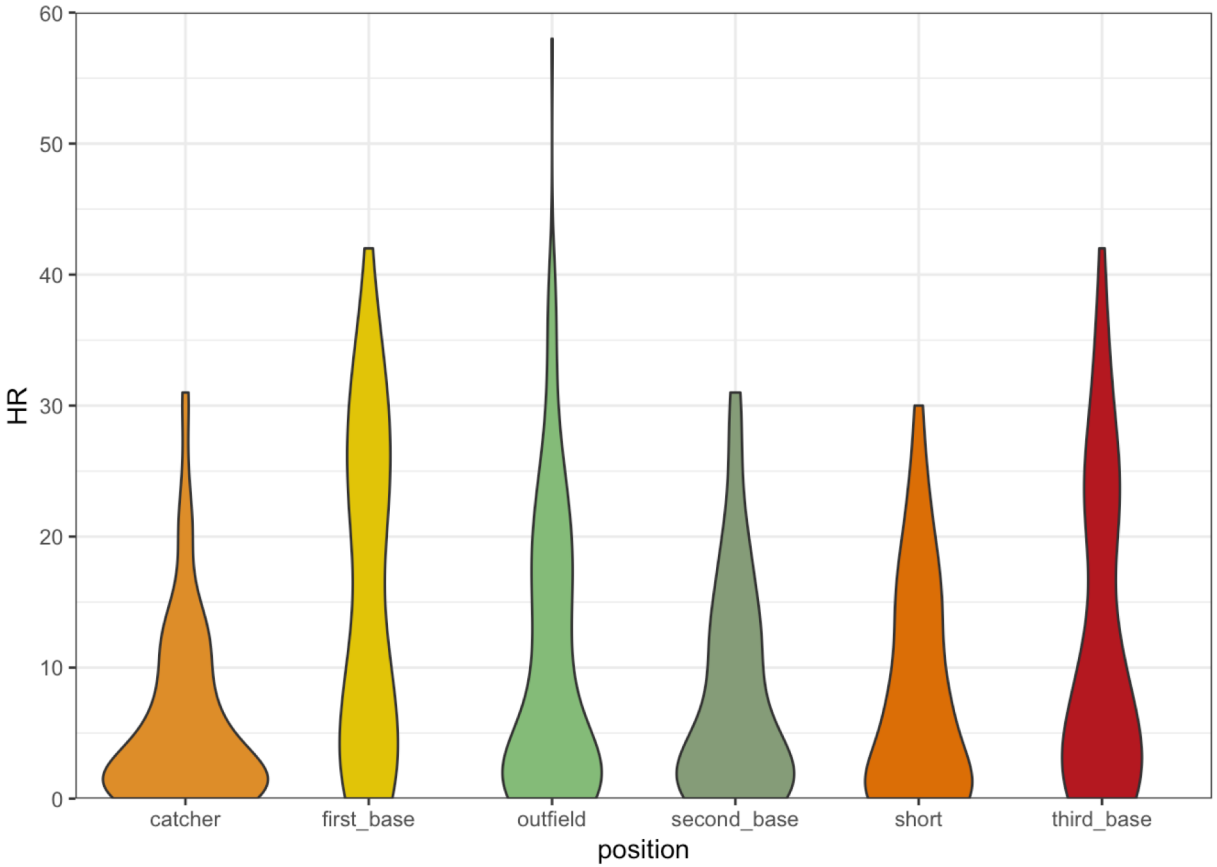
Home runs by position

```
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)
```



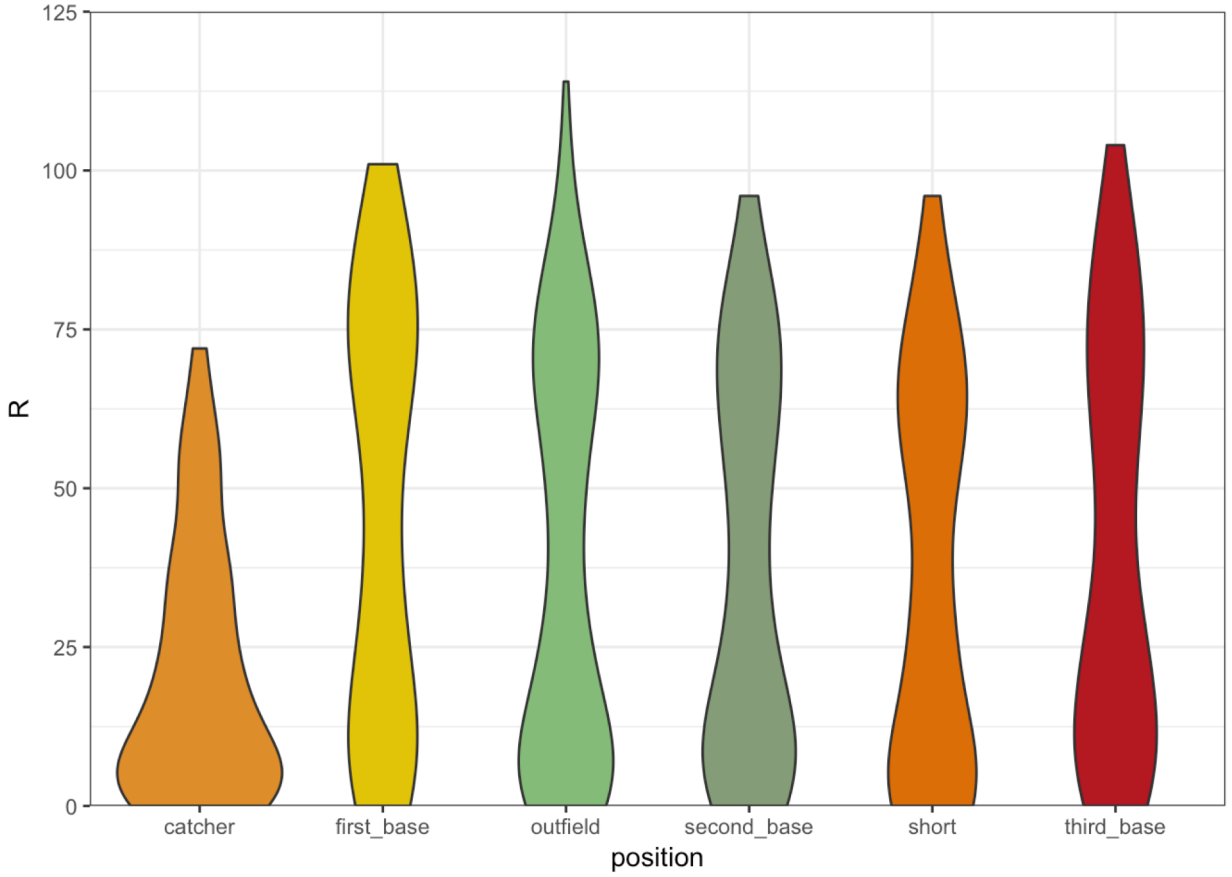
Home runs by position

```
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)
```



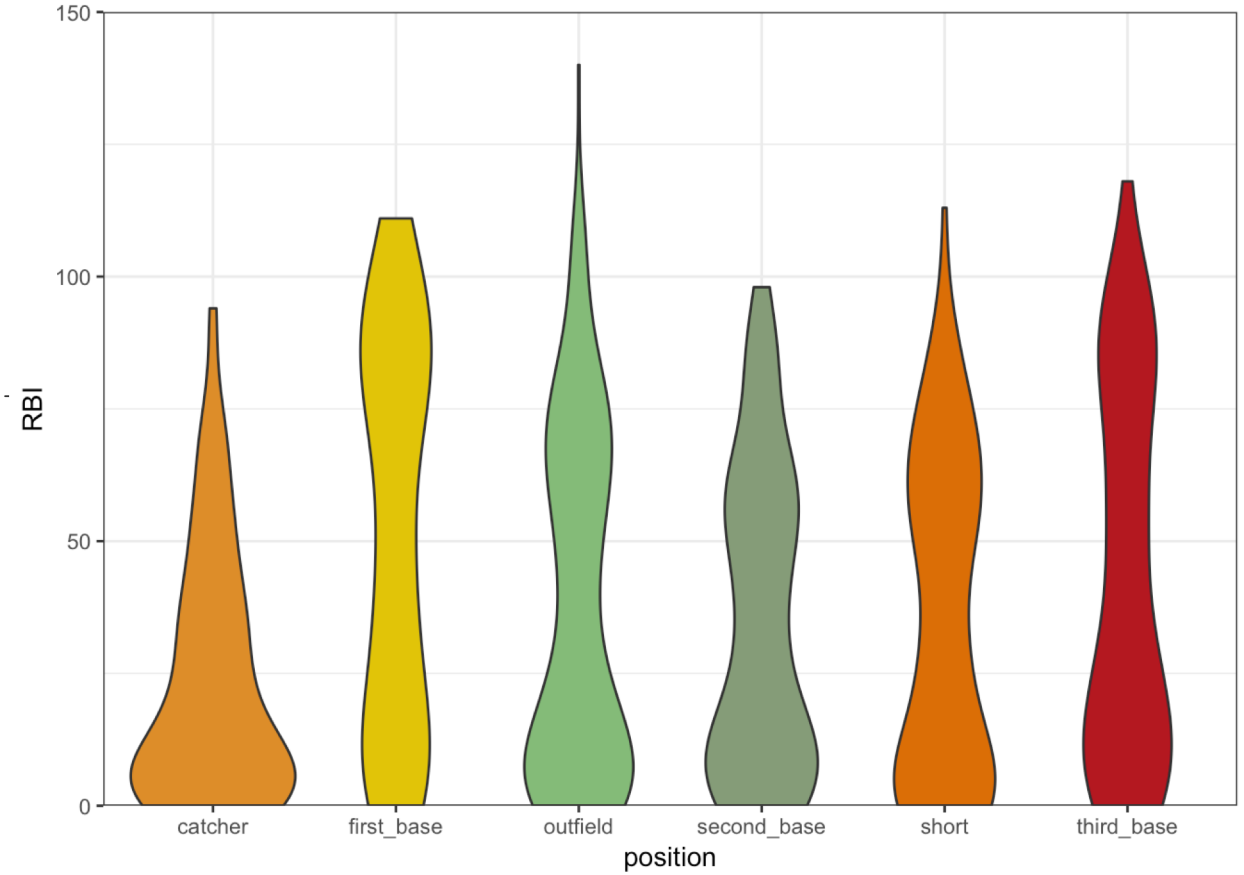
Runs by position

```
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))
```

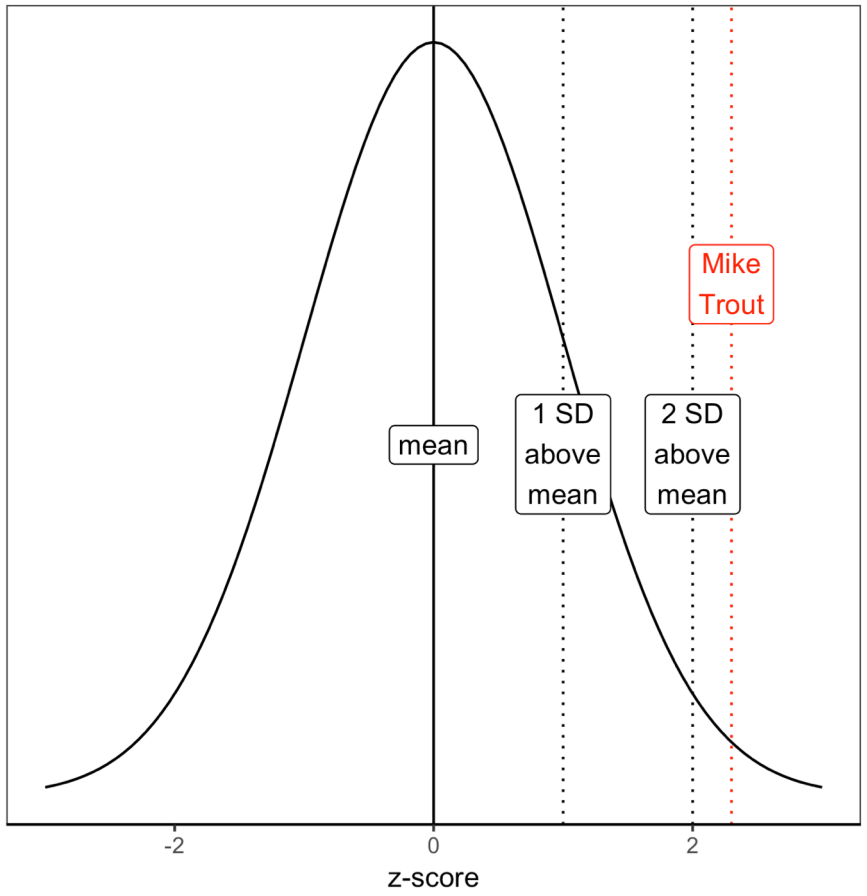


Key questions

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



position	Name	HR	HR_z
outfield	Giancarlo Stanton	58	4.391960
first_base	Joey Gallo	42	2.612658
third_base	Joey Gallo	42	2.612658
outfield	Aaron Judge	41	2.501452
first_base	Cody Bellinger	39	2.279039
third_base	Nolan Arenado	39	2.279039
outfield	Mike Trout	39	2.279039
outfield	J.D. Martinez	39	2.279039
outfield	Cody Bellinger	39	2.279039

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)
```

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

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