Comparing MLB Run Value and Awards

Andrew Eross, Owen Wassel, Nick McConnell

Introduction

Baseball has always been called "America's Pastime." However, beginning in the end of the 20th century and the early 2000's, the evolution of baseball analytics has greatly evolved the game. In the past, baseball people had just looked at offensive stats such as batting average, home runs, RBI, and strikeouts. But with the ability to track more and more data related to baseball, new stats have been developed to analyze and evaluate teams and players.

Many attribute the "Moneyball A's" as being a source renaissance for baseball analytics, where the early 2000's Oakland Athetics focused their roster building and talent acquisition by evaluating players on less popular statistics such as OBP (on base percentage), an effective way to acquire impact players who were overlooked by other teams. However, this age has now past, and new numbers drive baseball decision making.

MLB introduced Statcast in 2016, which, from the MLB website, is described as the "state-of-the-art tracking technology that allows for the collection and analysis of a massive amount of baseball data" through means such as cameras, radars, and other tracking devices. This data is accessible on Baseball Savant (an MLB licensed website). Baseball Savant takes these observations and calculates statistics related to how well players move and how balls are hit or thrown by players based on their spin, direction, and velocity. Some well-known Baseball Savant stats are hard-hit rates (how often a player hits a ball well) and xBA (the expected batting average of a player given how well they hit the ball).

The statistic we would like to analyze is Baseball Savant's run value, described on their website: "Every pitch is assigned a run value based on its outcome (ball, strike, home run, etc.). The sum of all of a player's contributions across a season, or multiple seasons, measures his overall batting or pitching run value. A positive value represents runs created for hitters, and runs prevented for pitchers." We are just looking at offensive run value, so how many runs hitters supposedly create at the plate (not including base running). This run value calculation is all "theoretical" meaning it is not derived by summing all the runs that score as a result of a hitter's plate appearance, rather how many *should* score based on how well they hit the ball, neutralizing factors such as opponent defense and teammate base running.

The calculation for run value is also not public. We have an idea of what data should be included in its derivation (such as on base percentage, home runs, etc), but Baseball Savant hides its exact calculation (or else others would copy its formula). Our report also includes research on estimating the weights of counting stats on run value, so seeing what stats might influence its calculation the most.

Run value is also calculated in each part of a hitter's zone. Depending on where the ball was pitched to the hitter, the associated run value may be attributed to the hitter's zone, shadow, chase, or waste run value. The image below shows where each of these zones are relative to the strike zone, as we will create visuals to compare hitters in different parts of the zone.

ENTER IMAGE

Research Questions

The reason why we want to use run value is to estimate the impact of a player's run value on their likelihood of winning an award. We would like to answer the following questions:

- 1.) Are the best offensive players (in terms of run value) more likely to finish atop the MVP (most valuable player) voting?
- 2.) Are other statistics (like WAR) a better predictor of who will place higher for MLB awards?
- 3.) How does this change when we analyze an award like Silver Slugger, an award that disregards defense and base running?

ENTER MORE

Data Provenance

As mentioned above, our *primary data set* is the run value data that comes from Baseball Savant, an official MLB website. However, we are also pulling the data from Baseball Reference, another online, reputable and accurate site to get the award voting data. This is our *secondary data set* and also includes many of the averages and counting stats that go with vote receivers. It is also worth noting that we are pulling this data from the 2019 season (just the regular season, not including playoffs). The reason for this was because this year was considered the "juiced ball year" meaning the physical baseball had extra bounce. This meant that the ball would bounce better off the bat and carry longer distances, helping hitters. Overall batting averages and runs scored were a relative high in 2019, meaning we have ample offensive data to draw from in a year known for its offense. Choosing any other year should reach similar conclusions (except for 2020 perhaps, where teams only played 60 games instead of 162).

Primary Data Set

- Source: Baseball Savant
- Description: display batter's run value data, calculated through Statcast measurables.
- Purpose: estimate a hitter's offensve imopact measured by calculating runs created.
- Cases: rows are hitters, colunmns have run value data overall and in each zone.

Secondary Data Set

- Source: Baseball Reference
- Description: displays award votes, who won each award, and basic statistics on these players.
- Purpose: give a basic description of a player's stats and where they finished for an award.
- Cases: rows are players who received votes for an award.
- Note: may include several tables from the same site depending on the award/conference of the player.

Data Wrangling

Because we are pulling the data from multiple data sources and combining into other data frames, several data wrangling steps must be taken. We downloaded all of the Baseball Savant run value and Baseball Reference voting data. Some important steps/ideas in the data wrangling process were as followed:

Headers

Extract the headers separately from the Baseball Reference cases. This ensures later we can better tidy the data by combining a separate header data frame to the rest of the data frame.

Tidying

In each table, make sure that column names/data are consistent. The way names appeared were different in both, so it was useful to separate names in a first and last name column. Also, rename columns to relevant, understandable names. It was also important to rename any values that contained accents, as loading the data frame would mess up those cell's formatting.

Filtering

It is important to filter out all of the pitchers who received award votes since they do not carry meaningful offensive data. We are just comparing hitters with their voting results.

Merging

An inner join was necessary on First_Name, Last_Name to combine the data frames to include run value data with voting data. Since we only wanted data on those receiving votes, only run vale data would be added to the names in the voting data frame.

Reproducibility

This process should be reproducible since it had to be repeated for data frame in both conferences. Avoiding "hard-coding" ensures that we are able to quickly change our code to produce other data frames.

FAIR Principles

Our data set must follow the FAIR principles so to ensure our data is trustworthy and valid to use.

Findable

Both data sets come from very well known baseball data pages. A quick google search of run value and 2019 award voting will take a user to these online data sets.

Accessible

Downloading the data is made very easy via the tables' settings/options on the website. This allows us to easily upload the data to our repo for our use.

Interoperable

For the most part, column names were very easy to understand what data came with it, and if not, we renamed those columns. The .csv files we downloaded from the sites were very easy to read into an R file, and understandable once doing so.

Reusable

Detailing our process should make the wrangling and EDA very reproducible for other researchers. There is adequate information on the structures of the data sets both in our documentation and from the website so other can use them for their own research.

CARE Principles

In most cases, analyzing professional sports data follows CARE principles because it is objective analysis on public data meant to study the highlights of the sport. Our research does so, uplifting the positives and accomplishments of baseball measurables. We understand we do not own any of these public data sets, and we provide proper reference information

Exploratory Data Analysis

To first tackle how well run value is related to MVP voting, we will plot a player's run value with their MVP vote points. Some context on how voting works- since MVP is conference specific, there is an award for both leagues. Therefore, we will be running this twice, one for each conference. Also, the y-axis is referenced as "Vote Points" because the way voting works is that each voter (30 in total) gets to vote for a 1st place finisher, 2nd place finisher, third place finisher, all the way down to a 10th place finisher. A player who is picked in first place gets 14 points, second place gets 9, 8 for third, and all the way down to 1 for 10th.

The number of vote points is summed for each player, and the player with the most points is the winner.

The **?@fig-nl-mvp-rv**, plots the player's run value vs vote points for the NL (National League). We see a pretty strong, positive correlation between vote points and run value. Our predictions were correct, in that players with higher run values are more likely to get more votes for the award. This of course makes sense, as the league's most valuable players should be great offensive players.

The points represent each player who received MVP votes. Originally, the player's name was listed above the point, but this was changed to improve readability.

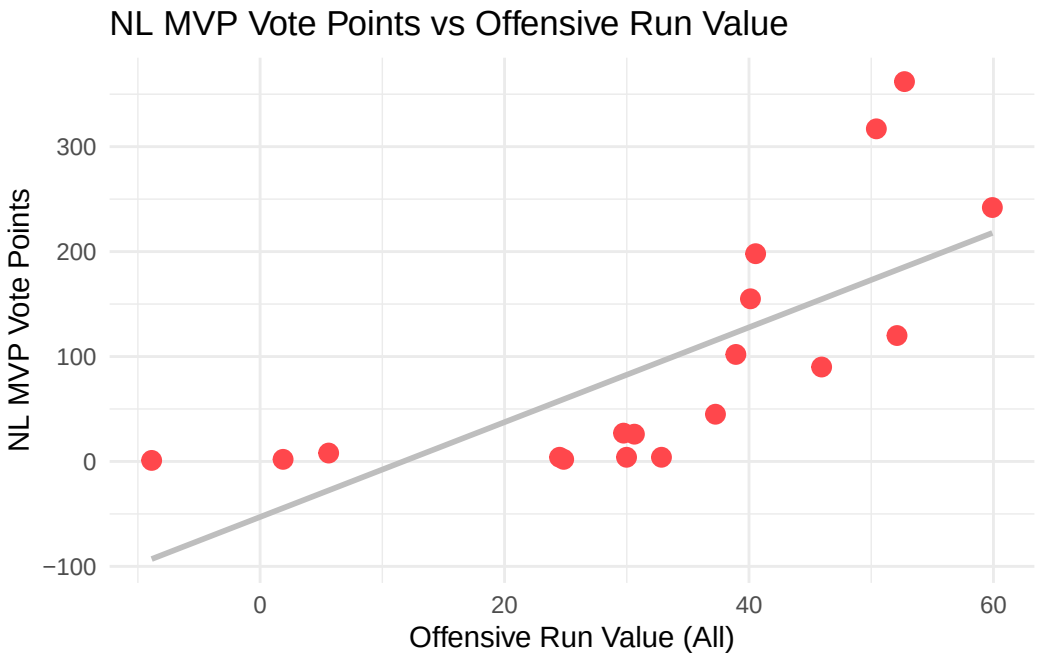

Figure 1: Run Value vs NL MVP Vote Points

The line of best fit is also graphed, with a calculated R squared value of **0.515**.

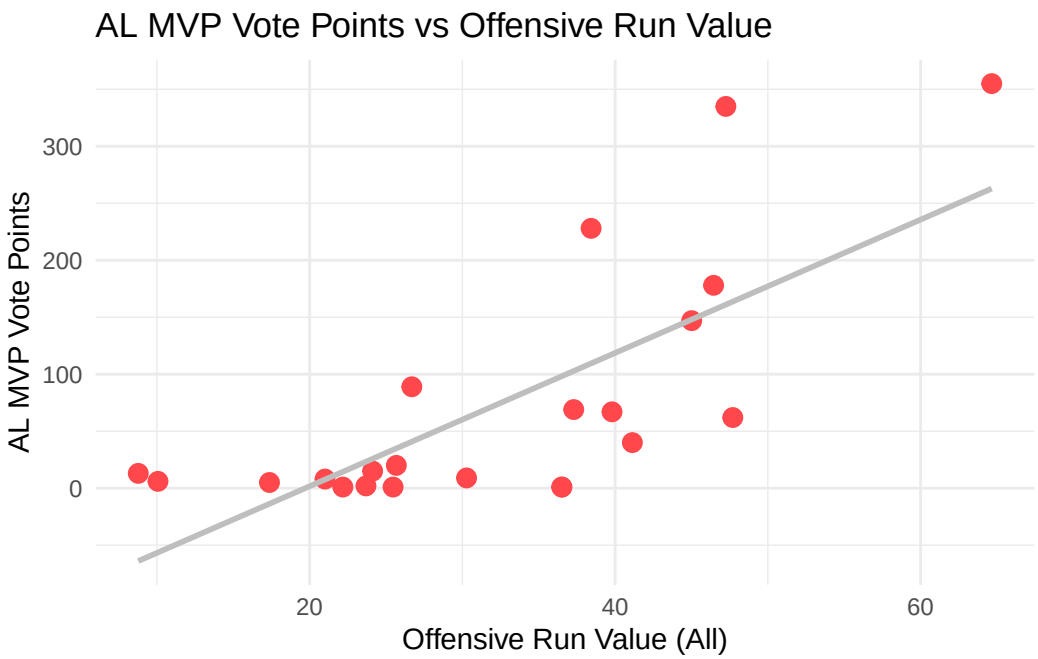

Figure 2: Run Value vs AL MVP Vote Points

We repeated this process for the AL MVP vote getters, with the similar figure shown. This produced an R squared value of **0.544**.

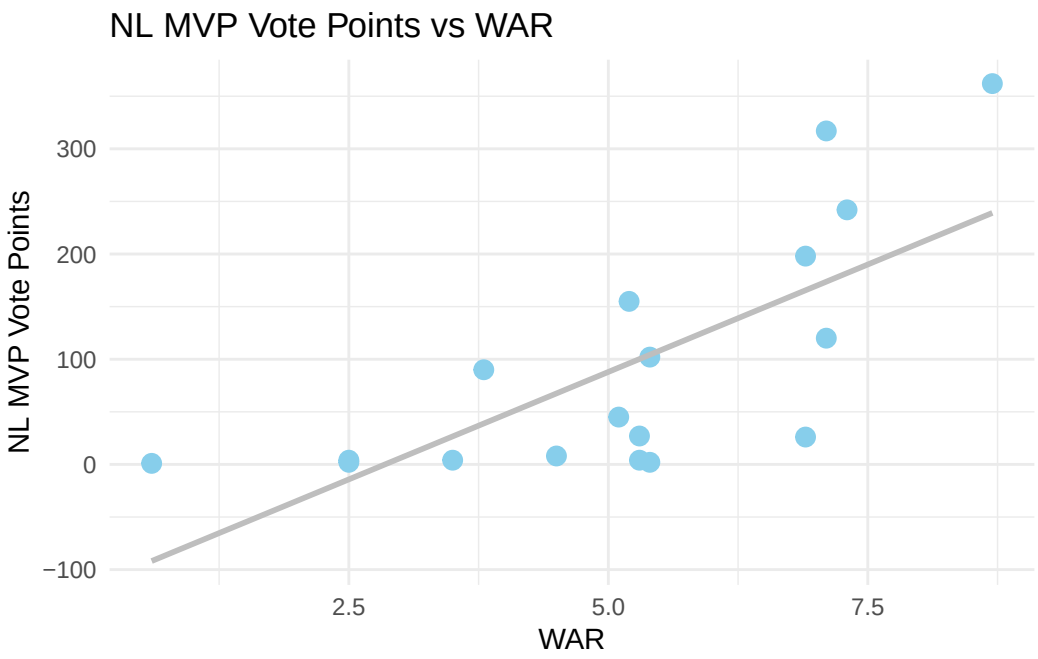

Figure 3: WAR vs NL MVP Vote Points

The R-squared value for the NL MVP model using WAR is 0.52.

The R-squared value for the AL MVP model using run value is 0.544.

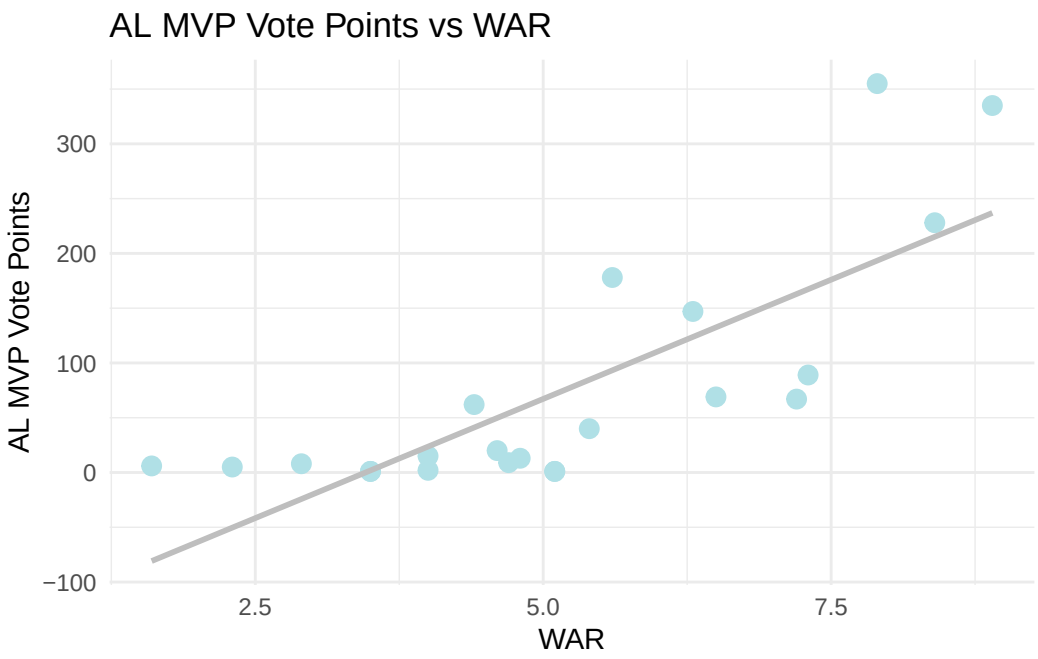

Figure 4: WAR vs AL MVP Vote Points

The R-squared value for the AL MVP model using WAR is 0.619.

Code Appendix

```
# ----DATA WRANGLING FOR NL MVP RUN VALUE DATA FRAME -------

# load packages
library(tidyverse)
library(ggplot2)

# Data Wrangling ---- Primary Data set
# read run value in from downloaded baseball savant
# table, use read csv
RVData <- read.table(
  file = "~/Documents/GitHub/MLB_Awards_Project/2019-batters.csv",
  sep =",",
  #include headers
  header = TRUE) %>%
```

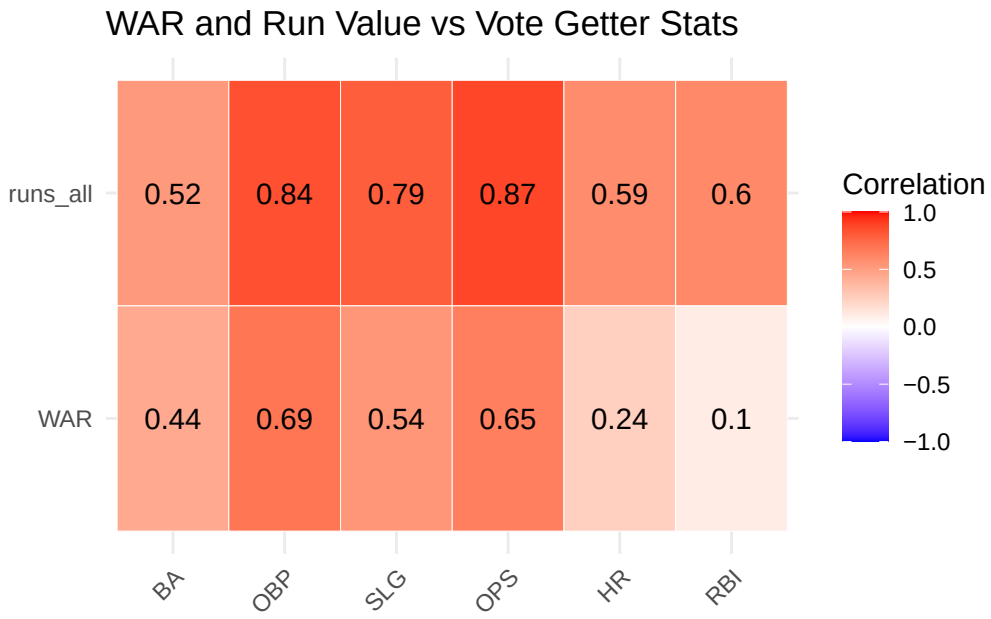

Figure 5: WAR and RV Correlations

```
  #separate first and last name into two columns
  separate(last_name..first_name, into = c("Last_Name", "First_Name"), sep = ", ")

# Data Wrangling ----- Secondary Data set

# Step 1: get headers from table, which is second row
NL_MVP_Header <- read.table(
  file = "~/Documents/GitHub/MLB_Awards_Project/2019NLMVP.csv",
  header = FALSE,
  sep = ",",
  skip = 1,
  nrows = 1
)

# Create large data frame through many piping steps
# firt get rest of cases, non-headers
NL_MVP <- read.table(
  file = "~/Documents/GitHub/MLB_Awards_Project/2019NLMVP.csv",
  header = FALSE,
  sep = ",",
```

```
  skip = 2,
  # use previous headers data frame to bind as headers
  col.names = as.character(unlist(NL_MVP_Header))
) %>%
  # select only offensive stats (remove pitching stats)
  select(1:19) %>%
  # filter out pitchers (any vote getters with less than 100 abs)
  filter(AB > 100) %>%
  # separate names into first and last
  separate(Name, into = c("First_Name", "Last_Name"), sep = " ") %>%
  # tidy ill-formatted names, fix accent errors
  mutate(Last_Name = ifelse(Last_Name == "SuÃ¡rez", "Suárez", Last_Name)) %>%
  mutate(Last_Name = ifelse(Last_Name == "AcuÃ±a", "Acuña Jr.", Last_Name)) %>%
  # rename column that makes more contextual sense
  rename('1st.Place.Votes' = X1st.Place) %>%
  # join on first name and last name with run value data frame
  inner_join(RVData, by = c("First_Name", "Last_Name")) %>%
  # drop unnecessary columns
  select(-player_id, -year, -pitches, -team_id)

#------PLOTTING RUN VALUE VS VOTE POINTS FOR NL MVP -------

#create plot for run value vs NL MVP vote points
# plotting from NL MVP data frame
# x axis is run value (all zones)
# y axis is vote points
ggplot(NL_MVP, aes(x = runs_all, y = Vote.Pts)) +
  # graph dots as points, light red for run value, size 3
  # note: not including names for readability
  geom_point(color = "#FF474C", size = 3) +
  # create line of best fit/regression line, light gray
  geom_smooth(method = "lm", se = FALSE, color = "gray") +
  # x, y, and graph labels
  labs(
    title = "NL MVP Vote Points vs Offensive Run Value",
    x = "Offensive Run Value (All)",
    y = "NL MVP Vote Points"
  ) +
  # minimize background to plain
  theme_minimal()
```

```r
# create model to calculate r squared value for vote points vs run value from nl mvp
model_nlmvp_rv <- lm(Vote.Pts ~ runs_all, data = NL_MVP)

# calculate r squared value from model
rsq_nlmvp_rv<- summary(model_nlmvp_rv)$r.squared

# I researched how to do this since it was not taught in course


# ----- PLOT WAR VS VOTE POINTS FOR NL MVP HITTERS --------

# similar approach to above, just now graphing for war vs vote points
# graph war vs vote points from NL MVP data set
ggplot(NL_MVP, aes(x = WAR, y = Vote.Pts)) +
  #color the points, will use a light blue for WAR
  geom_point(color = "#87CEEB", size = 3) +
  #create a regression line/line of best fit, color a light grat
  geom_smooth(method = "lm", se = FALSE, color = "gray") +
  # graph and axes labels
  labs(
    title = "NL MVP Vote Points vs WAR",
    x = "WAR",
    y = "NL MVP Vote Points"
  ) +
  #minimize background to white
  theme_minimal()
```

```r
#create model to get r squared value between vote and war
model_nlmvp_war <- lm(Vote.Pts ~ WAR, data = NL_MVP)
# calculate r squared value with model
rsq_nlmvp_war<- summary(model_nlmvp_war)$r.squared



# - CREATE PLOT FOR RUN VALUE VS VOTE POINTS FOR AL MVP ------

# create same plot of run value data vs vote points for the AL hitters
# data comes from AL MVP data frame, x axis is run value, y axis is vote points
ggplot(AL_MVP, aes(x = runs_all, y = Vote.Pts)) +
  # same idea with light red points for run value
  geom_point(color = "#FF474C", size = 3) +
  #create the line of best fit with a light gray line
```

```
  geom_smooth(method = "lm", se = FALSE, color = "gray") +
  #create graph, x, y labels
  labs(
    title = "AL MVP Vote Points vs Offensive Run Value",
    x = "Offensive Run Value (All)",
    y = "AL MVP Vote Points"
  ) +
  #minimize background
  theme_minimal()
```

```
#create model for r squared value vs AL run value vs vote points
model_almvp_rv <- lm(Vote.Pts ~ runs_all, data = AL_MVP)
#calculate r squared value using model
rsq_almvp_rv<- summary(model_almvp_rv)$r.squared


# ---- CREATE PLOT FOR WAR VS VOTE POINTS AL MVP --------

#same plot as above just change the x axis to war
# create plot from al mvp data frame, war is x axis and vote points are y-axis
ggplot(AL_MVP, aes(x = WAR, y = Vote.Pts)) +
  #create points with a light blue
  geom_point(color = "#B0E0E6", size = 3) +
  #line of best fit in a light grat
  geom_smooth(method = "lm", se = FALSE, color = "gray") +
  # graph and axes labels
  labs(
    title = "AL MVP Vote Points vs WAR",
    x = "WAR",
    y = "AL MVP Vote Points"
  ) +
  #minimize background
  theme_minimal()
```

```
# create model for vote points and war to calculate r squared value
model_almvp_war <- lm(Vote.Pts ~ WAR, data = AL_MVP)
# calculate r squared value with model
rsq_almvp_war<- summary(model_almvp_war)$r.squared



# ---- CREATE HEATMAP FOR CORRELATIONS ------
```

```r
#this is for a heat map to show correlations between RV and War for offensive stats
# this is beyond our in class curriculum so I researched the process of making a heat map

#load package
library(reshape2)

# combine hitters in both AL and NL into one data frame
MVP_Hitters <- rbind(AL_MVP, NL_MVP) %>%
  # can drop the rank column
  select(-Rank)

# rows are the derived stats (run value and WAR)
# columns are the counting stats that factor into these stats
rows <- c("WAR", "runs_all")
cols <- c("BA", "OBP", "SLG", "OPS", "HR", "RBI")

# create the correlation data frame with the hitters data set and given rows and columns
Correlation_DF <- MVP_Hitters[, c(rows, cols)]

# use the correlation data frame to compute the correlation matrix
Correlation_Matrix <- cor(Correlation_DF, use = "complete.obs")

# make a correlation subset using the matrix and given rows and columns
Correlation_Subset <- Correlation_Matrix[rows, cols]

# correlation long object by "melting" the subset
Correlation_Long <- melt(Correlation_Subset)

# plot the correlation long where x are the counting stats and y are the run value and war
# we are filling with the gradient of correlation
ggplot(Correlation_Long, aes(x = Var2, y = Var1, fill = value)) +
  # set tiles as white
  geom_tile(color = "white") +
  # create gradient, where low is -1 and high is 1, getting redder as correlation approaches
  #create colors from blue to white to red
  scale_fill_gradient2(low = "blue",
                       high = "red",
                       mid = "white",
                       midpoint = 0,
                       limit = c(-1, 1),
                       name = "Correlation") +
  # label with black text the actual correlation
```

```
  geom_text(aes(label = round(value, 2)), color = "black", size = 4) +
  # minimize background
  theme_minimal() +
  #axes and graph labels
  labs(x = "", y = "", title = "WAR and Run Value vs Vote Getter Stats") +
  # angle the x axis text for visual aesthetic
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```