MLB ANALYSIS

Andrew Lujan
Regis University
M.S. Data Science
Practicum 1

Why?

REASON

Gain Experience applying models to a sports-related domain, an area that I wish to work in with data science.

PROBLEM

- Which statistics contribute the most to identifying whether or not a player is a power-hitter?
- How can we classify power-hitters?
- Which statistics are highly correlated with winning teams?

Data Source:

The data comes from a Kaggle dataset. It can be found at: https://www.kaggle.com/open-source-sports/baseball-databank

Additional information about the data:

- When working with the dataset I learned that there'saaan extension of the Open-Source sports dataset that is curated by Sean Lahman at the following link: http://www.seanlahman.com/baseball-archive/statistics/
- I intend to use this source of baseball data for future projects.

PROJECT LAYOUT

- Data acquisition
- Data cleaning
- Exploratory Data Analysis
- Data Visualization
- Model Building
- Model Testing
- 2 prediction problems
    - Power Classification
    - Predicting Wins
- Findings/Results

DATA ACQUISITION

- CSV files were downloaded from Kaggle pulled in using the pandas read csv function.

Import Data

```
8]: batting = pd.read_csv("/Users/drewsdesktop/Desktop/Data Science/Regis Classes/MSDS 692- Practicum/Datasets/Baseball Databank/Batting.csv")
    team = pd.read_csv("/Users/drewsdesktop/Desktop/Data Science/Regis Classes/MSDS 692- Practicum/Datasets/Baseball Databank/Teams.csv")
```

Data Cleaning

- Two datasets: batting, team.
  - both in CSV format.
- Utilized heatmaps to check for correlation in both datasets.
- Batting dataset
  - Filtered dataset to a time period starting from my lifetime, the year 1990-2015.
  - Missing hitting data for pitchers.
  - Feature engineering
  - Created classification for power hitters using binning.

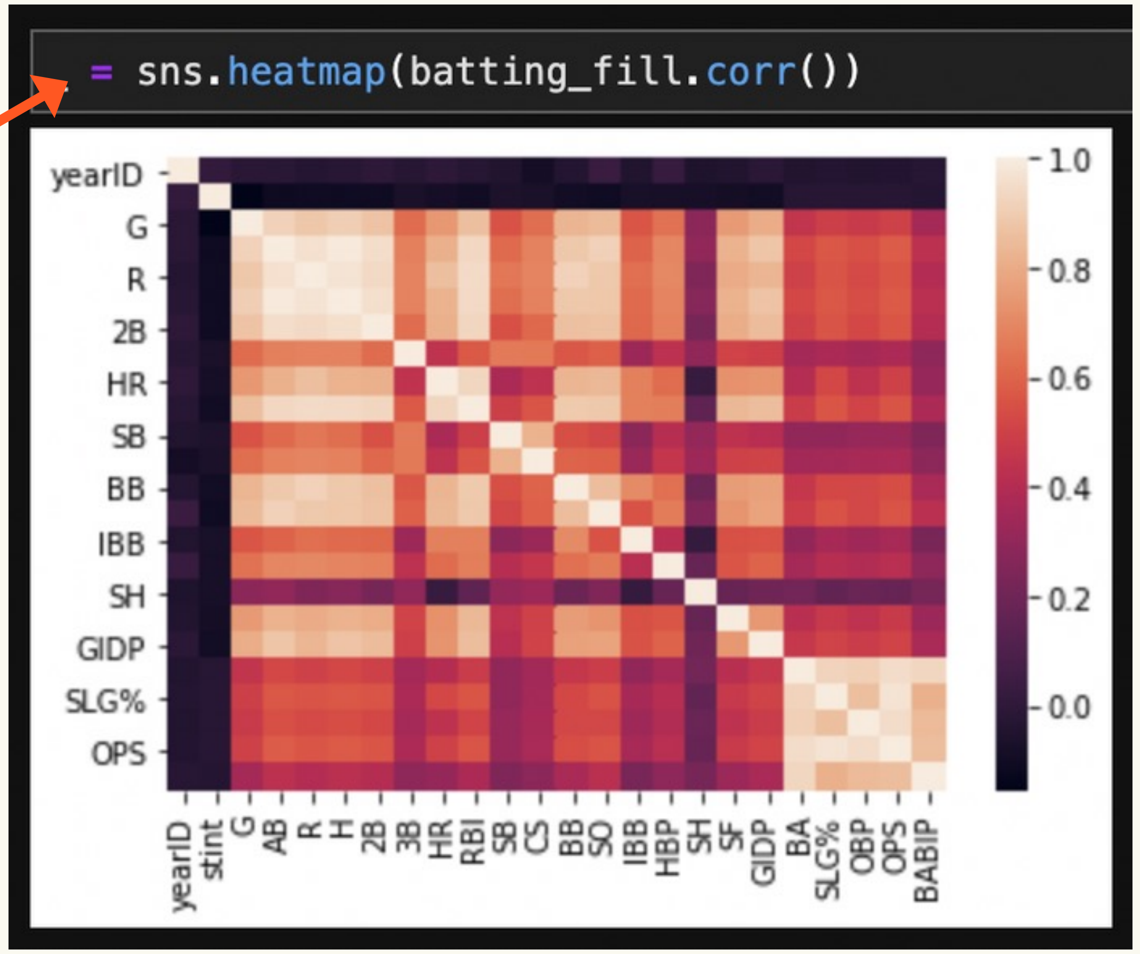
Data Cleaning

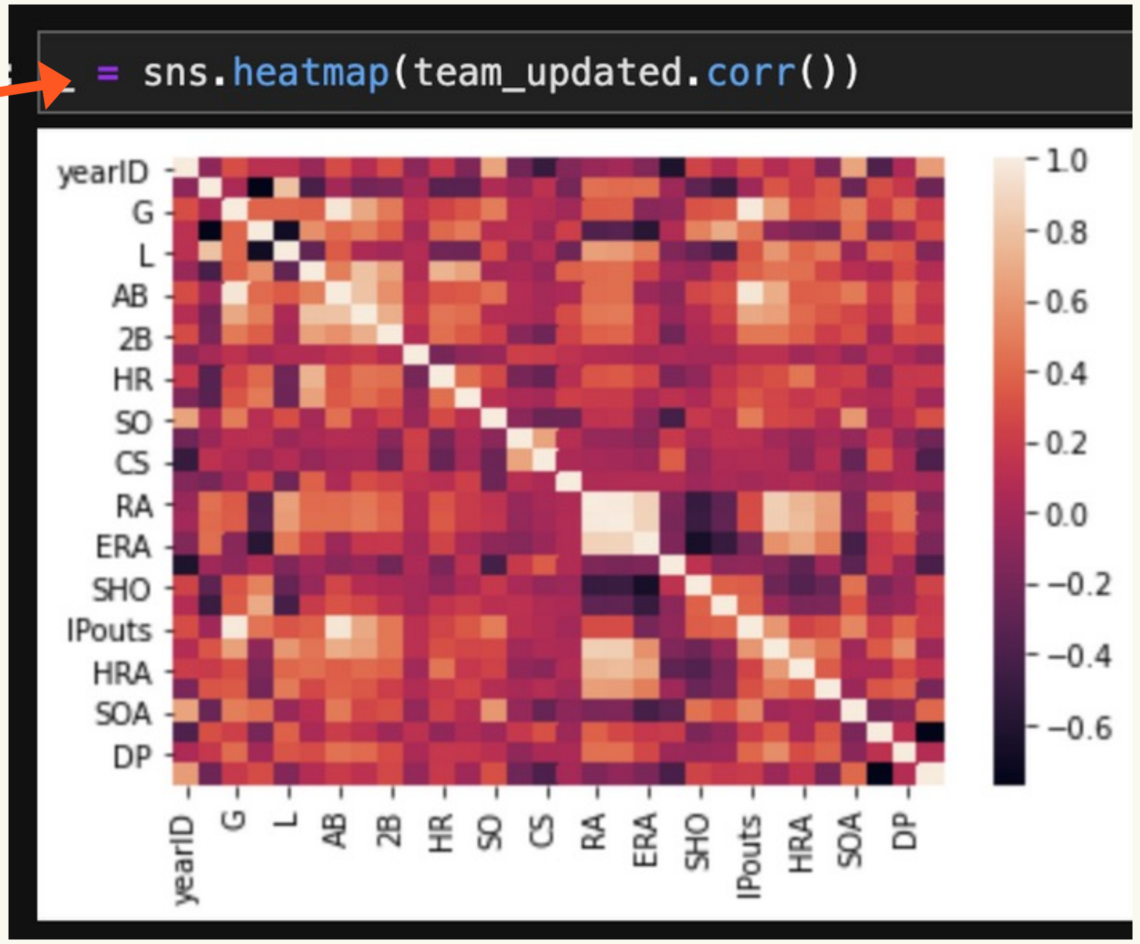
- Team dataset
  - Filtered dataset for time period 1990-2015.
  - Dropped unnecessary columns. Lots of identification columns.
  - Dropped some statistical columns where the null-values exceeded 30%.
  - Imputed some null values for more important statistical features.

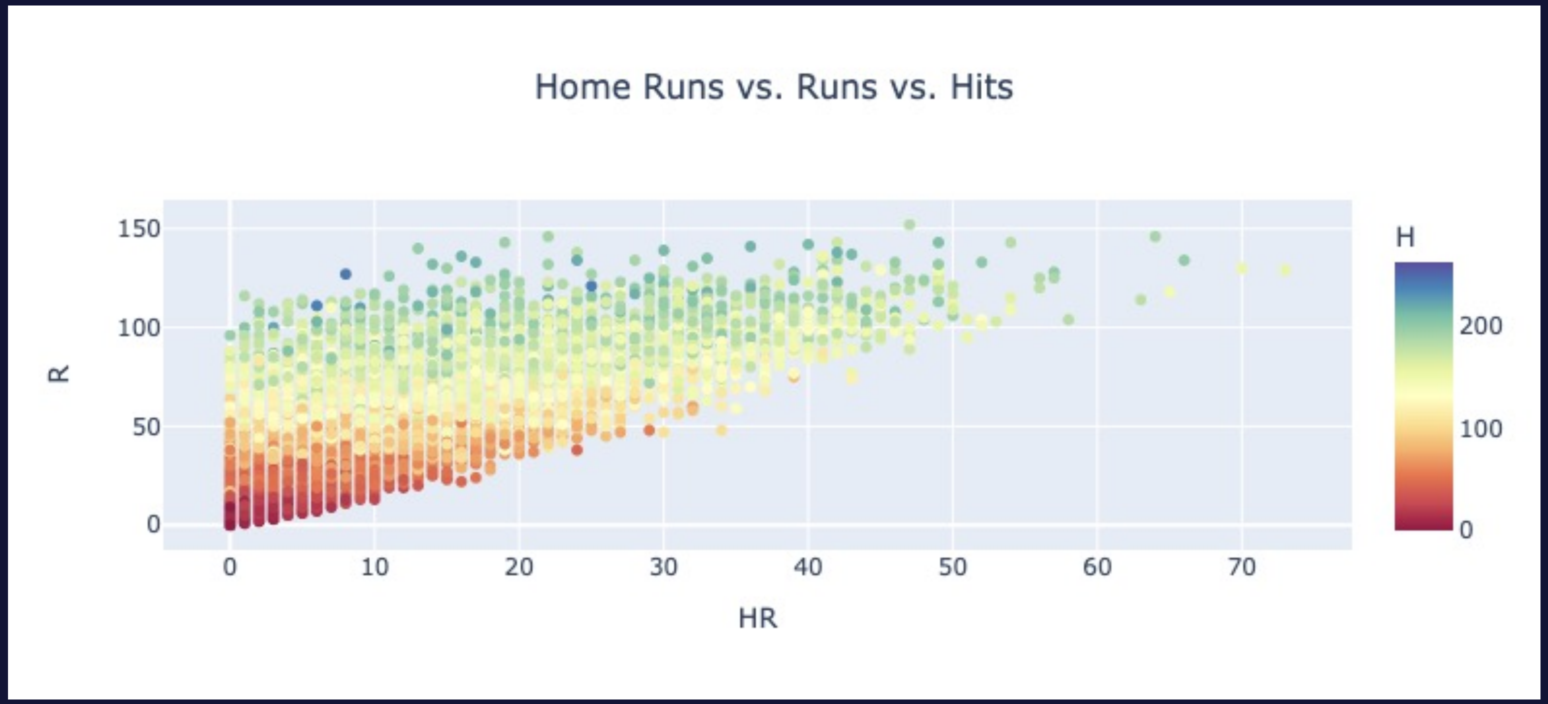
EXPLORATORY DATA ANALYSIS

Important statistical features:

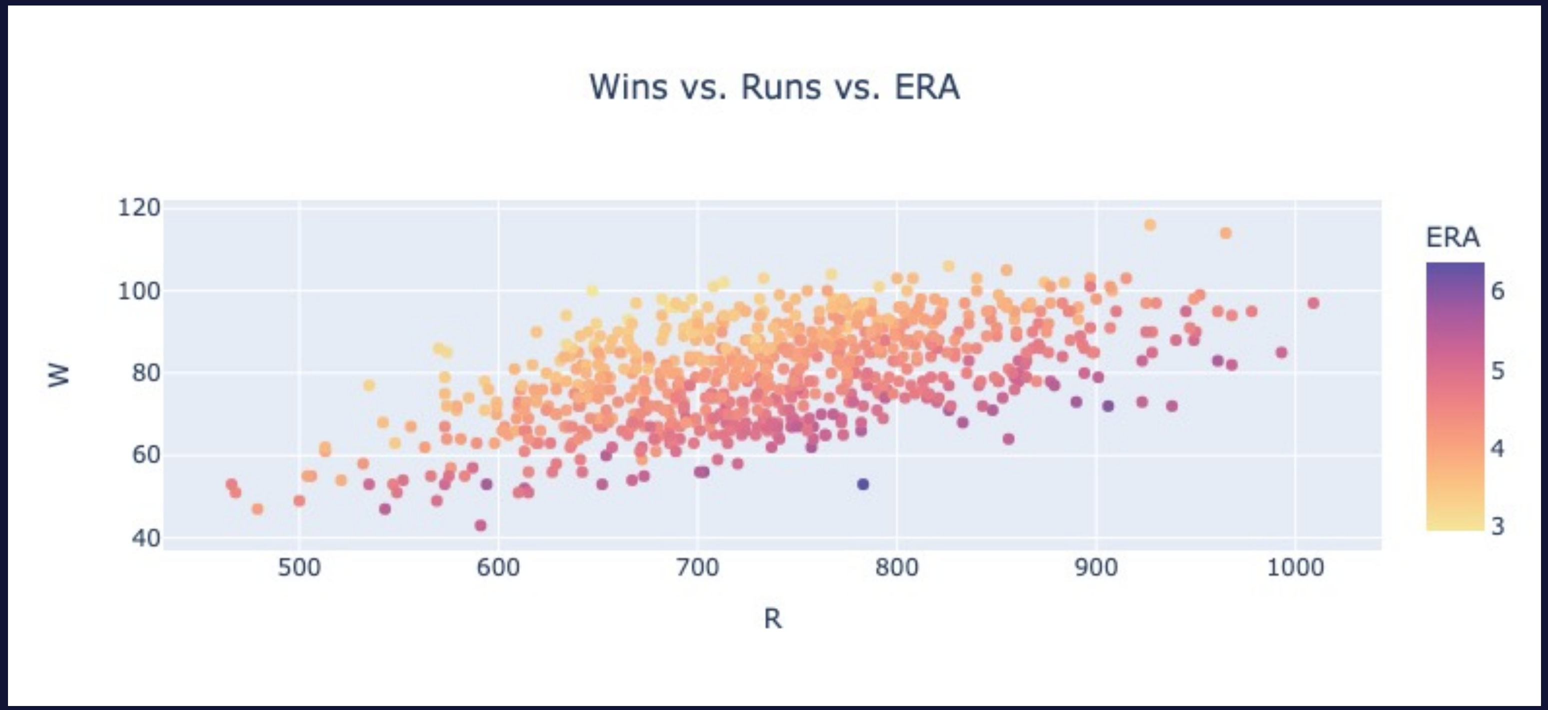
- Batting Dataset
  - Players with high totals for Runs, Hits, and Doubles tended to have higher home runs totals.
- Team Dataset
  - Runs, Home Runs, SOA (Strikeouts by pitchers) were the features that had the strongest correlation to win totals.







Batting Visualization



Team Visualization

Data Visualizations

***Visualizations were created to explore the relationships between the highly correlated variables and the predictor variables.**

MODEL BUILDING/ TESTING

Batting Dataset- (Accuracy Scores)

- Power classification
- Logistic Regression- (94.98)
- K-nearest neighbors- (95.27)
- Gaussian Bayes- (86.47)

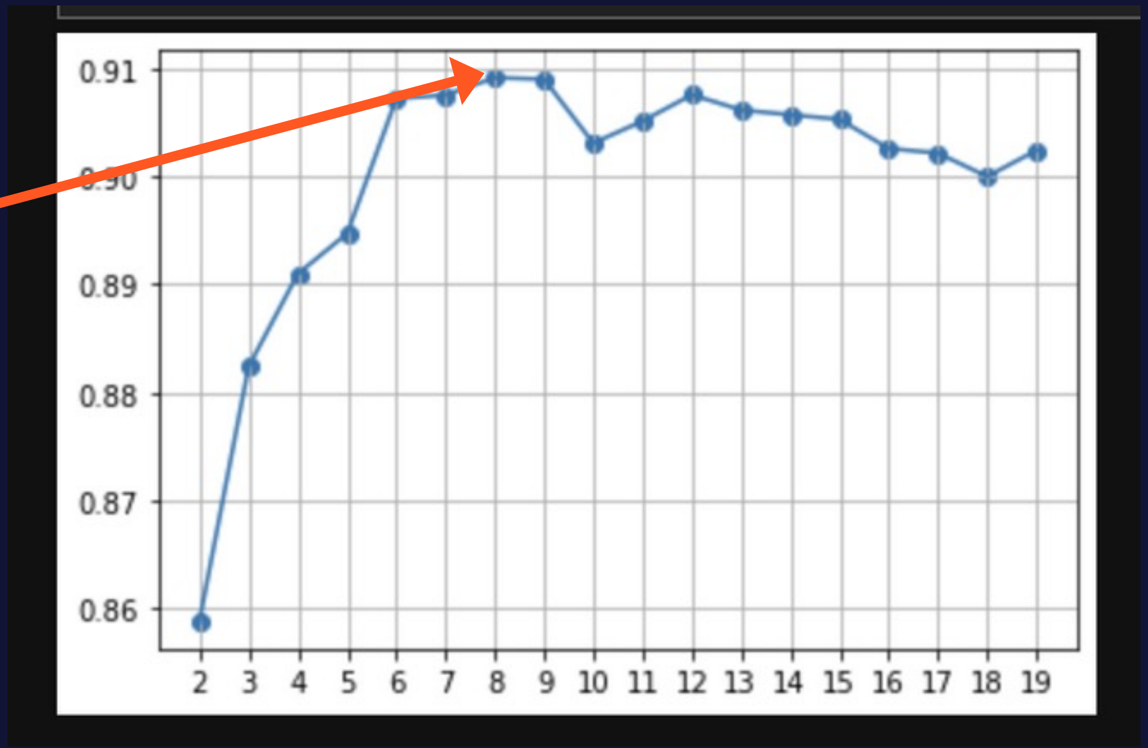
Team Dataset

- Predicting wins
- K-means clustering(.79)
- K Nearest Neighbors Regressor
  - Optimizing through normalization
- Random Forest Classifier- (.88)

Optimum clusters using silhouette score

```
Silhouette score for  2  clusters:  0.7898234908667836
Silhouette score for  3  clusters:  0.17264539737487875
Silhouette score for  4  clusters:  0.18132973300280172
Silhouette score for  5  clusters:  0.15279822494600523
Silhouette score for  6  clusters:  0.1344701214548308
Silhouette score for  7  clusters:  0.11550928429752681
Silhouette score for  8  clusters:  0.08494893473822426
Silhouette score for  9  clusters:  0.11201568329578987
Silhouette score for  10  clusters:  0.1035163727369197
Silhouette score for  11  clusters:  0.0984747436619074
Silhouette score for  12  clusters:  0.10374775762451482
Silhouette score for  13  clusters:  0.0820063537390603
Silhouette score for  14  clusters:  0.08228074461843092
Silhouette score for  15  clusters:  0.07518665776014691
Silhouette score for  16  clusters:  0.06105502345645081
Silhouette score for  17  clusters:  0.06968202896801265
Silhouette score for  18  clusters:  0.041440098507671594
Silhouette score for  19  clusters:  0.06966923898588959
Silhouette score for  20  clusters:  0.06500942777498112
Silhouette score for  21  clusters:  0.06956382872731404
Silhouette score for  22  clusters:  0.019891587582388244
Silhouette score for  23  clusters:  0.02434745694185978
Silhouette score for  24  clusters:  0.034825343581667084
Silhouette score for  25  clusters:  0.012148952702833482
Silhouette score for  26  clusters:  0.009286035281570841
Silhouette score for  27  clusters:  0.033312571584919425
Silhouette score for  28  clusters:  0.019221190239593452
Silhouette score for  29  clusters:  0.023809275166228085
```

Optimum neighbors by r2 score.



- **Original set = .797**
  - Normalized data model = .909

Findings/Results

Finding

- **Which statistics were most correlated with Home Runs?**
  - Three features that had high correlation with home runs were runs, doubles, and hits.
- **Can we classify power-hitters?**
  - We can classify power hitters by binning players based on their home run outputs
- **Which statistics are highly correlated with winning teams?**
  - **Runs, Home Runs, Strikeouts by Pitcher.**

Model Performance

- **Power Classification**
  - **KNN classifier**
- **Win Prediction**
  - **KNN classifier with normalized data**