MLB Pitch Prediction

(Sprint 2)
Ty Painter, Connor Mignone, Dong Chen,
Daron Greenblatt

Objectives

- Create a model and eventually dashboard that will help the offensive team (the hitting team) in pitch anticipation during an at-bat.
- Help predict pitch type based on batter, pitcher, runners on base, count during the at-bat, pitch sequence, etc.
- This dashboard would help teams predict what will be thrown therefore hopefully increasing the chances the batter will be successful.

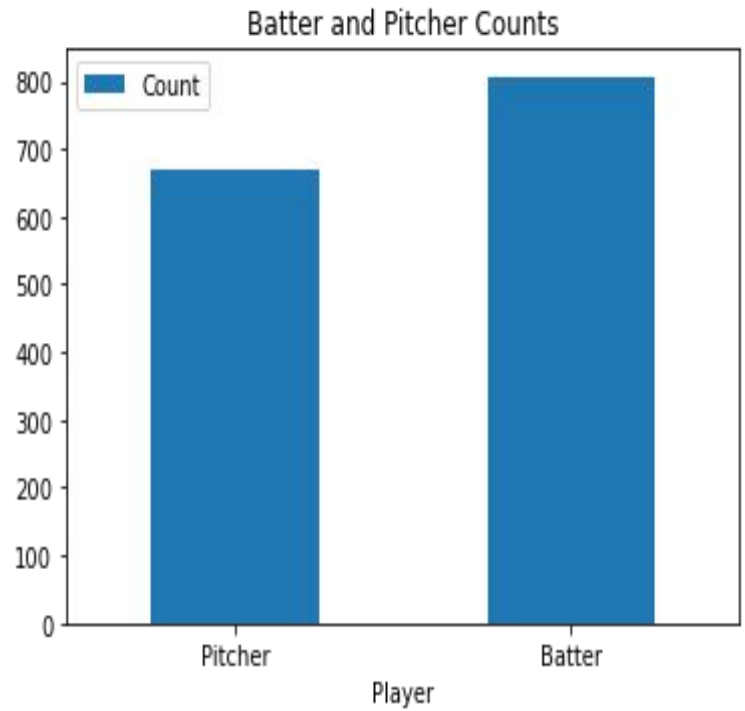
Data Summary

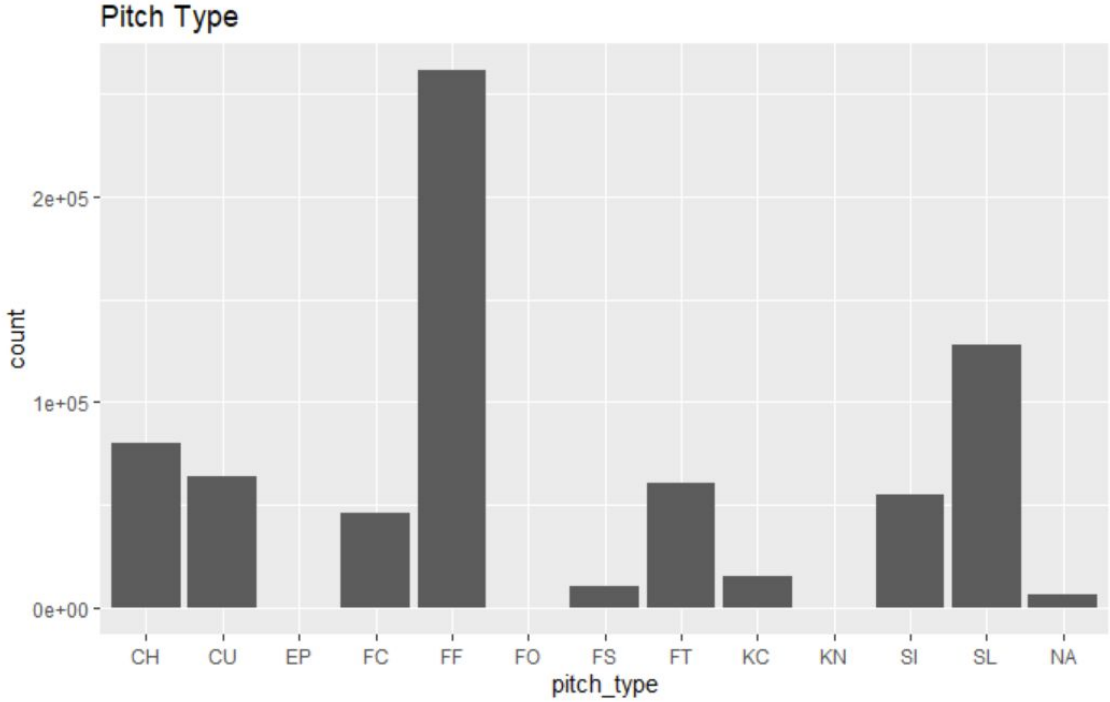
- Every pitch thrown during the Major League Baseball (MLB) 2019 season
- 8 total datasets, 1.06GB
- Merged 3 datasets together
  - Pitch data
  - At-Bat data
  - Player data
- Merged dataset
  - ~600K rows
  - 54 columns
  - Too big for GitHub - revised down to 25 columns

Data Cleaning

Seeing inaccurate or incomplete data:

- Pitch count often not changing after a pitch is thrown - will be difficult to include this important feature in our model if it is unreliable
- Many other columns have questionable data but we will likely use filler values such as mean or mode


Batter and Pitcher Counts
Count
800
700
600
500
400
300
200
100
0
Pitcher
Batter
Player


Pitch Type
2e+05
1e+05
0e+00
count
CH
CU
EP
FC
FF
FO
FS
FT
KC
KN
SI
SL
NA
pitch_type

count
40000
30000
20000
10000
0
Batter Interference
Bunt Groundout
Bunt Lineout
Bunt Pop Out
Catcher Interference
Caught Stealing 2B
Caught Stealing 3B
Caught Stealing Home
Double
Double Play
Fan Interference
Field Error
Fielders Choice
Fielders Choice Out
Flyout
Forceout
Game Advisory
Grounded Into DP
Groundout
Hit By Pitch
Home Run
Intent Walk
Lineout
Passed Ball
Pickoff 1B
Pickoff 2B
Pickoff 3B
Pickoff Caught Stealing 2B
Pickoff Caught Stealing 3B
Pickoff Caught Stealing Home
Pickoff Error 1B
Pop Out
Runner Double Play
Runner Out
Sac Bunt
Sac Bunt Double Play
Sac Fly
Sac Fly Double Play
Single
Stolen Base 2B
Strikeout
Strikeout Double Play
Triple
Triple Play
Walk
Wild Pitch
Play Event

Next Steps

- Feature engineering
  - Batter
  - Pitcher
  - Count
  - Base runners
  - Score
- More EDA
- Split train/test data
  - Eliminate pitchers/hitter with below a threshold of pitches thrown