Baseball Player Performance and its Factors

Vail Dorchester and Jackson Mediavilla

Project Motivation

Are there unexpected correlations between “outside” factors (temperature, stadium attendance, night/day, etc.) and baseball player performance metrics such as batting average?

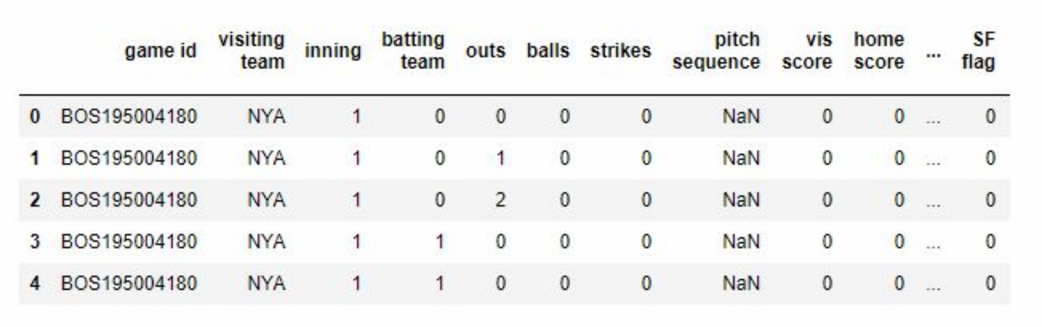
Which outside factors affect baseball player performance the most?

Tools Used

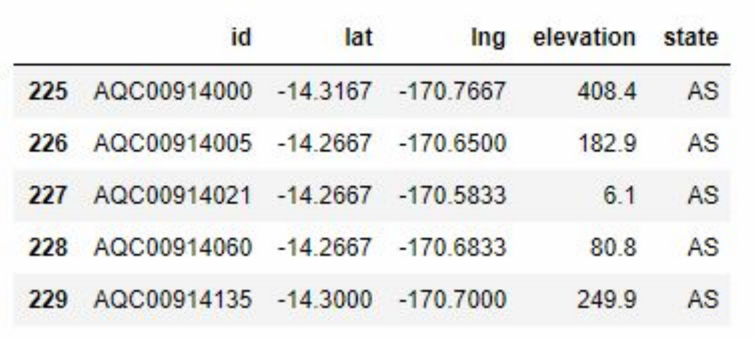
- Python
- Numpy
- Pandas - python data analysis library
- Git, GitHub
    - https://github.com/jackson-mediavilla/csci4502-project/
- Jupyter Notebook
- Sklearn
- Scipy
- MatPlotLib
- Excel
- Bash
- Retrosheet Scripts

Data Gathering

- Datasets
  - Event Data - 10.5M rows
  - Game Logs - 133K rows
  - City Data - 36K rows
  - Station Data - 106K rows
  - Weather Data - 1.4B rows
- Why all these datasets?
  - Connect game data to weather data



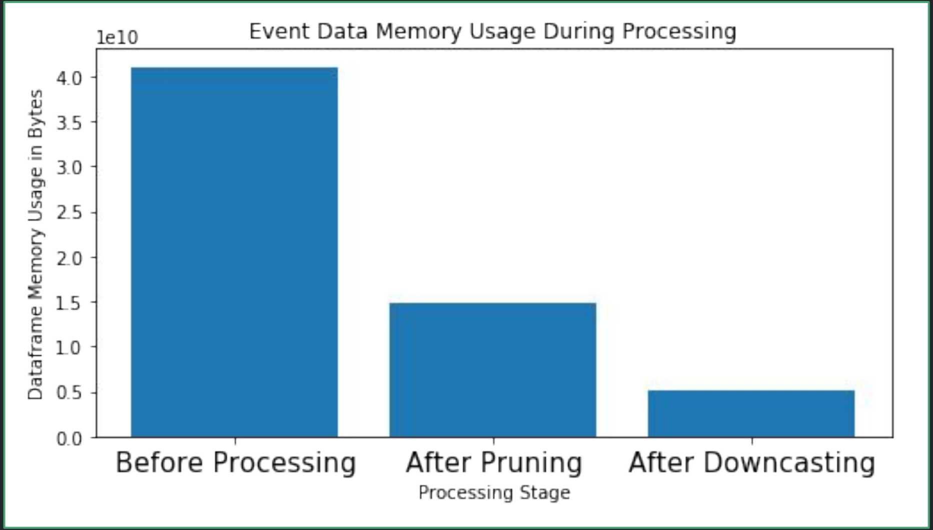
	game id	visiting team	inning	batting team	outs	balls	strikes	pitch sequence	vis score	home score	...	SF flag
0	BOS195004180	NYA	1	0	0	0	0	NaN	0	0	...	0
1	BOS195004180	NYA	1	0	1	0	0	NaN	0	0	...	0
2	BOS195004180	NYA	1	0	2	0	0	NaN	0	0	...	0
3	BOS195004180	NYA	1	1	0	0	0	NaN	0	0	...	0
4	BOS195004180	NYA	1	1	0	0	0	NaN	0	0	...	0



	id	lat	lng	elevation	state
225	AQC00914000	-14.3167	-170.7667	408.4	AS
226	AQC00914005	-14.2667	-170.6500	182.9	AS
227	AQC00914021	-14.2667	-170.5833	6.1	AS
228	AQC00914060	-14.2667	-170.6833	80.8	AS
229	AQC00914135	-14.3000	-170.7000	249.9	AS

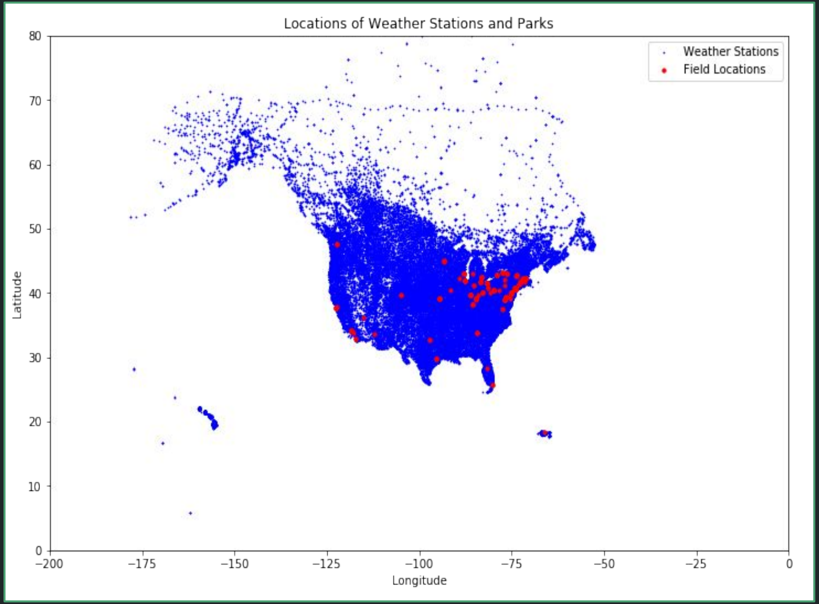
Data Pruning

- Data - 6GB
    - Main methods:
        - Pruning irrelevant columns
        - Converting objects to categories
        - Converting categories to ints
        - Downcasting numeric types
- All of the data we found had way more features than we needed
- Reduced memory usage by 89%



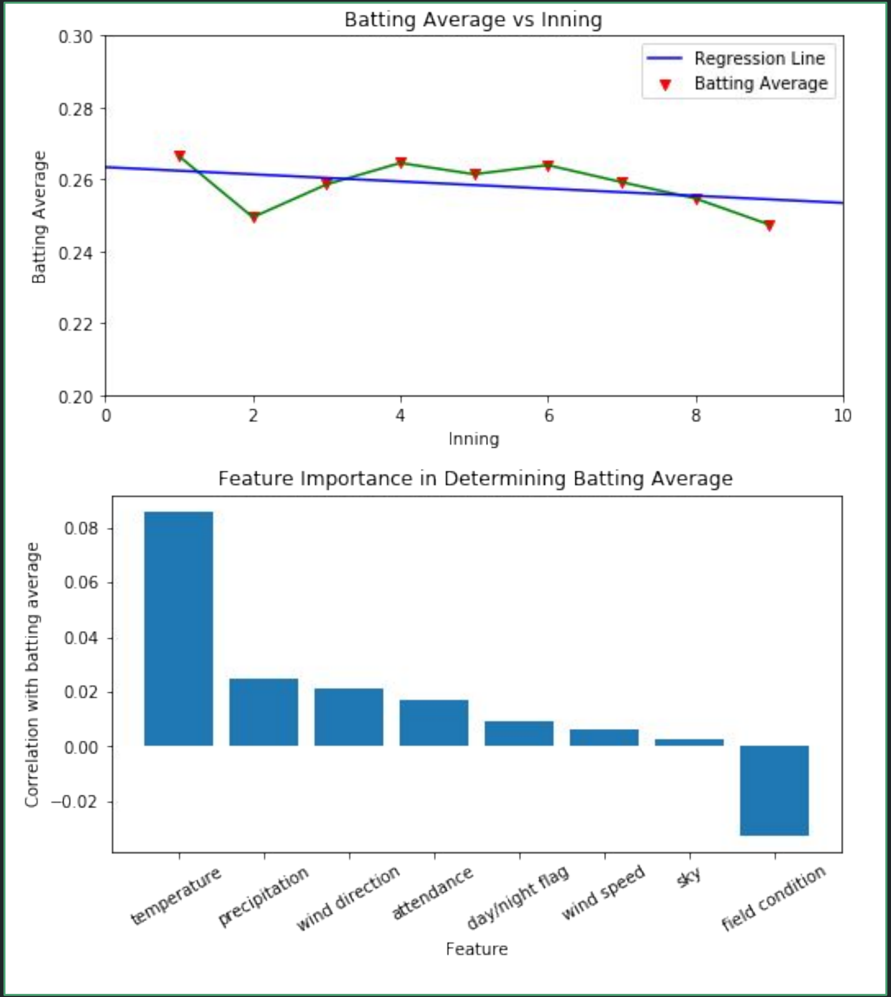
Connecting Relevant Data

- Connect weather data to game logs
  - Game logs list a city
  - Match city in game logs to city in city dataset to get coordinates
  - Match coordinates of game to nearest weather station
    - KNN
  - Use weather station and date of game to query for relevant weather data

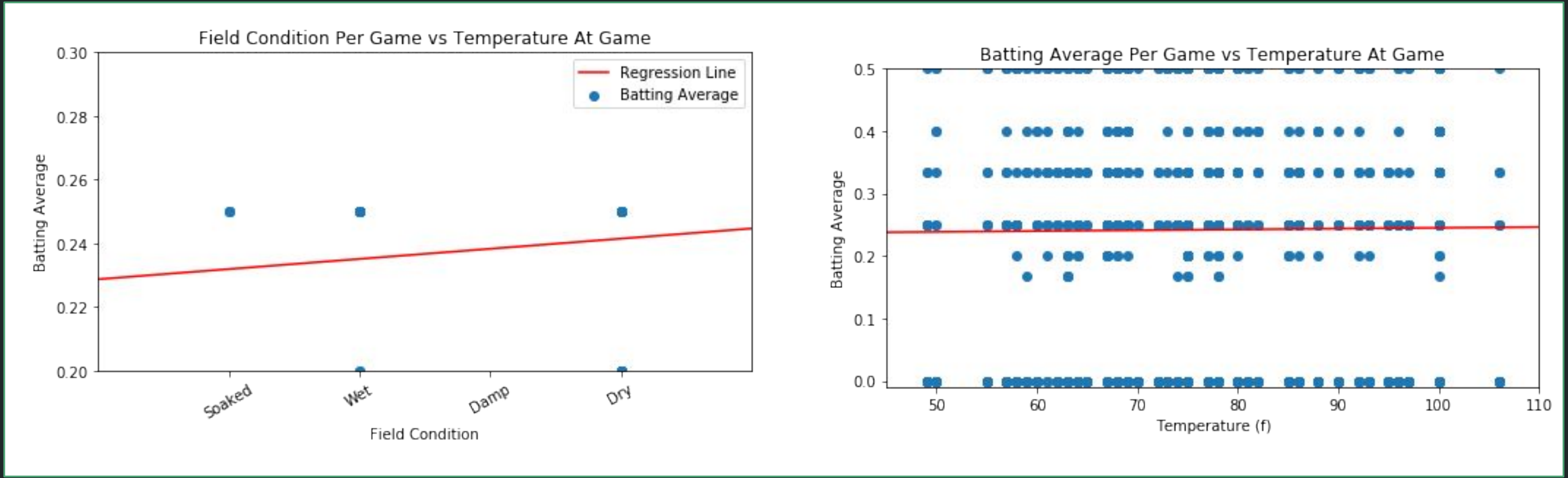


Data Analysis

- Looked at batting average vs inning
  - Linear Regression
- Determined feature importance
  - Random Forest Regressor
  - Extract features
  - Sort by importance
  - Calculate correlations
- Hypothesis Tests
  - Able to reject null hypothesis for every inning *except* the 9th



Feature Regression



Field Condition Per Game vs Temperature At Game

Batting Average Per Game vs Temperature At Game

Knowledge Gained

Outside factors and batting average are correlated

- Temperature, field condition, precipitation, wind direction, attendance, day vs. night, wind speed, sky conditions

Linear regression shows baseball players have lower batting average in later innings.

Application of Gained Knowledge

Baseball teams

- Coaching strategy
- Individual player strategy
- Scouting players

Baseball stadiums

- Adjust ticket prices based on weather factors

Sabermetrics

- Interesting data for baseball data enthusiasts
- Application in more complex sabermetric equations