



A QUANTITATIVE ANALYSIS OF MLB BASEBALL

RYAN SHEA

DECEMBER 18, 2021



~~OVERVIEW~~

- Background in baseball, play at the collegiate level
- Always had a passion for the game as well as the analytics
- Focus on 3 major questions



DATA

- All data from Baseball Savant
- Utilize NumPy, Pandas in Python in order to preprocess/clean



TOOLS FOR VISUALIZATIONS

- Tableau for map visualizations
- Matplotlib, Seaborn, shap in Python



QUESTION I:WHAT ARE THE MOST IMPORTANT ATTRIBUTES ON AN EFFECTIVE SLIDER?

- Quantitative statistics
- Manually label, feed into XGBoost classification ML model
- Derive feature importances using SHapley Additive exPlanations



QUESTION 2: WHICH ARE THE MOST IMPORTANT COUNTS IN DETERMINING WIN EXPECTANCY

- Database has a change in win expectancy for each pitch
- Use similar approach for the first question, this time utilizing a regression approach



QUESTION 3: HOW DOES SEASONALITY AFFECT CERTAIN STATISTICS?

- How does the month change batting average, fastball velocity, etc.
- Something I personally deal with as a pitcher in the cold
- See if it holds up at the MLB as well

