The original idea of my project was to answer the question, How far do 1st round draft picks to make it in their MLB careers? I was having issues with some of the original questions I planned on answering doing this project. The main issue I was running into was finding enough player data which showed which high school they were picked from. A lot of the information I was finding was inaccurate or simple didn’t account for the player's high school name. I found myself taxing a large amount of time to just find some clean data and joining the datasets I did find together properly and in a way that made sense. Due to this, I chose to focus my project on MLB but with the main hypothesis focusing on the winning percentage of a baseball team to its run differential.

I pulled my dataset from Retrosheet. This website contained game logs from various years going back to 1865. I chose to focus my data from the years 2010-2019. To begin, I chose to create a linear regression model to relate the win percentage of a baseball team to its run differential. To train the model, I chose the years 2010-2016 which would help with evaluating the performance of the model on test data from 2017. I then used a logistic regression model to predict the outcome of a game based on the number of total hits by the home team.

When I started the project, I was skeptical about how accurate I could get this model. After conducting a linear regression analysis, I was able to get an R-squared for linear fit = 0.8891311544013384, which I found to be a pretty good linear fit. Another interesting aspect of the analysis was that the error between the predicted results and the actual dataset wasn’t that far off. Percentage error on test set = 1.7380849888242396% vs Percentage error on training set 1.925846685228569%

I faced some minor challenges with this project such as not having the correct .zip files in my directory which caused an error every time I tried to import a new .csv file. I think another issue I ran into (which I plan to work on in the future) was to predict the 2020 season. For my project, I found myself confirm my accuracy through the 2019 game logs which I obviously couldn’t do since no games have been played. Overall, I think the project was a success and I was happy with the results I got in the end. I look forward to continuing to build on this model and hopefully be able to predict future seasons.