CSCI 381 - Project 3 Proposal Avenash Ramnarine

For my Project 3 I would like to explore the statistics of Major League Baseball and inspect certain attributes of my data set. More specifically I would like to explore the correlation between the amount of homeruns a baseball team makes during the year and how that affects their probability of winning a championship. I was lucky enough to find a data set that includes statistics for every single MLB team that has won their division leading all the way back to 1876. However this data set is massive and I will be cutting out a large portion of entries that are either too old or contain information that I don’t plan on using. You can take a look at the data set information here <https://www.openintro.org/data/index.php?data=mlb_teams> . Or if you would like to see it organized on my github account, you can take a look here <https://github.com/AvenashRamnarine/CSCI381/blob/main/mlb_teams.csv> .

The range I am planning to use is anything from about 1950-2020 which includes about 2500 rows. I am also planning on pulling out about 9 columns of data that will be relevant to my study of Homeruns vs Wins (including championship probability) which will most likely be: year, team\_name, games\_played, wins, losses, runs\_scored, hits, homeruns and world\_series\_winner. I am choosing to use this as my project idea as I am a huge fan of baseball and I have always wanted to do a project relating to this area of study. Also as you may know analytics has become a sort of revolution within baseball over the last few years. So being able to analyze and work with this kind of data to solve my very own hypothesis is something that I really would love to do.