**Design and Implementation of Analytics System:**

***Analyzing Baseball Data & Predict the Winners of 2020 Series***

**Project Background:**

Baseball has been referred to as America’s national pastime. Since its start in the mid-19th century, the game has grown in popularity across the globe. Many professional leagues have come and gone since the game began, with Major League Baseball (MLB) becoming the premier league. MLB held its first game in 1869 and continues to play games to this day. Financially, the MLB generates a great deal of money, and according to Forbes annual valuation of the MLB, the league generated $10.5 billion in revenue during the 2019 season. However, this year is different. Due to coronavirus, the 2020 season has been vastly changed.

The season’s opening day was pushed back by three months and the number of games to be played is now at 60 versus the usual 162 game season. Additionally, fans are not allowed to be in the stadiums when the games are played. Through September 2nd, 41 games have been postponed due to Covid-19 concerns. With all of these changes affecting the game and its players, we want to see if we can accurately predict several data points for the 2020 MLB season.

Forbes annual valuation URL:

[*https://www.forbes.com/sites/mikeozanian/2020/04/09/despite-lockdown-mlb-teams-gain-value-in-2020/#711e0e162010*](https://www.forbes.com/sites/mikeozanian/2020/04/09/despite-lockdown-mlb-teams-gain-value-in-2020/#711e0e162010)

**Scope:**

Baseball is a game of trends. With normal seasons being 162 games, where each team plays most other teams across multiple series, it’s normal to get a large sample size of regular season data which can be used to predict post season success. While the post season matchups are much shorter in length and they can be subject to short term variance and predicted winners could end up losing, the teams that are believed to be the best regular season teams are likely to be the best post season teams. However, 2020 has changed that. With MLB having only 37% of their regular games, it might be much more difficult to predict the post seasons successes this year. Additionally, MLB has been considering a reduction in the number of games played for a while now. Perhaps they could see this unique season as a working hypothesis to determine the effect that fewer games would have. Many different groups of people would be interested in answering these questions: Major League Baseball, individual baseball teams, casinos, even personal sports betters would be very interested in the outcome of the 2020 MLB season, and what prediction power the shortened season has.

**Research Questions:**

* Who will win the 2020 Cy Young for each league?
* Who will win the 2020 MVP Winners for each league?
* Who will win the National League? American League?
* Who will win the World Series in 2020?

**Tools & Technologies:**

* Scripting – R, Python / Jupiter-Notebooks
* ETL – KNIME / API for Current Year Data
* MongoDB for Datastore (NoSQL scalable database)
* Data Visualization – Tableau

Possible source of 2020 stats <https://www.baseball-reference.com/leagues/MLB/2020.shtml>