Capstone 3 Documentation

Data Collection and Cleaning

The dataset is made up of four seperate csv files. The data was downloaded as csv files from multiple sources. In-game stats and attendance data comes from Baseball Reference League Index. Information on ticket prices comes from the Economic History Association's The Economic History of Major League Baseball and Average ticket price in Major League Baseball 2006-2020. Data from the last two sources was compiled by hand using Excel before loading as a pandas DataFrame.

As the 2021 season was still in progress at the time of data collection, the Baseball Reference data (in the `bat`,`pitch`, and `misc` DataFrames) is incomplete. Due to the COVID-19 pandemic, the 2020 season was shortened to approximately 1/3 of the usual length. To ensure complete and easily comparable data, drop 2020 and 2021 from each dataset. Additonally, append attendance per game to each DataFrame in order to facilitate finding correlations between attendance and various other features of the data.

Problem statement

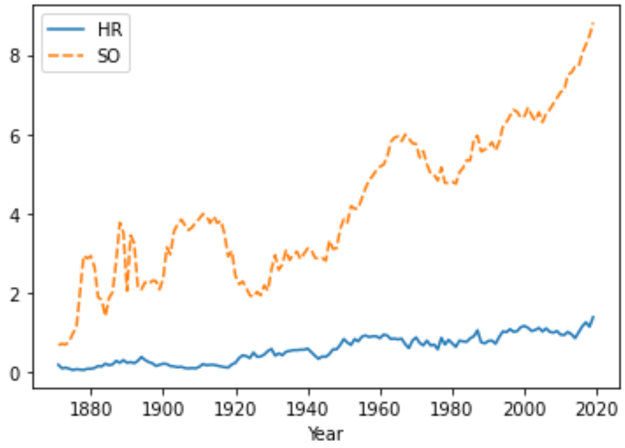
In the past five to ten years, there has been a noticeable drop in attendance at professional baseball games. In order to investigate the change in attendance, this analysis attempts to investigate the following questions:

- Is attendance correlated to any aspect of the game played on the field?
- Can the drop in attendance be predicted, or is it an anomaly?

EDA

Attendance per game is most correlated with strikeouts and homeruns. Let's plot both over time.

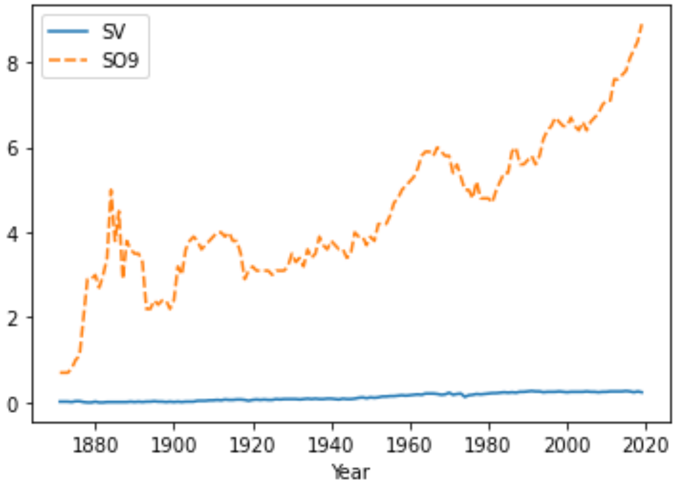
Strikeouts and Homeruns



Both strikeouts and home runs rise over time. Strikeouts look to increase steadily over time. Home runs have ups and downs over time, but rise more dramatically in recent years.

On the pitching side, the general rise in attendance is most correlated with saves and strikouts per 9 innings. There is also a high correlation with games finished, but this is a less interesting stat as it does not necessariy have anything to do with the games themsevles, just how many there are.

Saves and strikeouts over nine innings



Saves were not commonly recorded until some time in the 1950s. Furthermore, there can be no more than one save in any game. Nevertheless, the rise in saves is corrlated with the rise in attendance (**r = .94**).

The correlation coefficient between attendance per game and average ticket prices is approximately **r = .63**. By convention, this is a moderate correlation, although not as high as others I have focused on.

A common hypothesis among baseball fans and media pundits is that baseball is becoming less popular because the games take too long. Next, I will take a look at how long an average game has taken over time.

Here is a a sample of average game lengths (many times are missing):

Attendance per game is highly corrleated (**r = .95**) with game time. This observation does not square with the idea that people do not prefer longer games.

Modeling

Correlations

The correlations above provide an adequate picture of changes over time, but we can also investigate further.

Persistence Model

The next model is a simple persistence model. Each point of observation is predicted simply on the basis of the previous observation. Using this model provides a relative baseline for evaluating other models.

ARIMA Model

The final model is a simple ARIMA model. First, I checked the shape of the distribution then performed a power transformation using the Box-Cox method to give the data a more normal shape. In order to determine the parameters for the ARIMA model, I used the Augmented Dickey-Fuller test to determine stationarity and observed autocorrelation and partial autocorrelation plots to determine corrleation within the series. The model needed to be differenced 1 time. The size of the moving average window was determined by partial autocorrelation to be 2. The ideal number of lag observations was something like 10 -12, but 0 was used to simplify the model to ensure that the algorithm converges.

Results

The results of the final ARIMA model show that attendance is continuting to drop below the predicted average. However, during modeling, the predicted average was observed to move around depending on the number of lag observations. In some cases, attendance was observed to drop below the predicted average, while in others it was seen to fluctuate around the predicted average (as is to be expected). It was not seen to rise above the predicted average. Based on these observations, I provide the following conlcusions:

Correlations

As quantified above, attendance does seem to be corrleated with high profile statistics such as homeruns, strikeouts, and pitching saves. As game time has increased so has attendance. Despite league and media attention to shortening game time, game time itself does not seem to adversely affect attendance. Ticket price was correlated, moderately, with attendance.

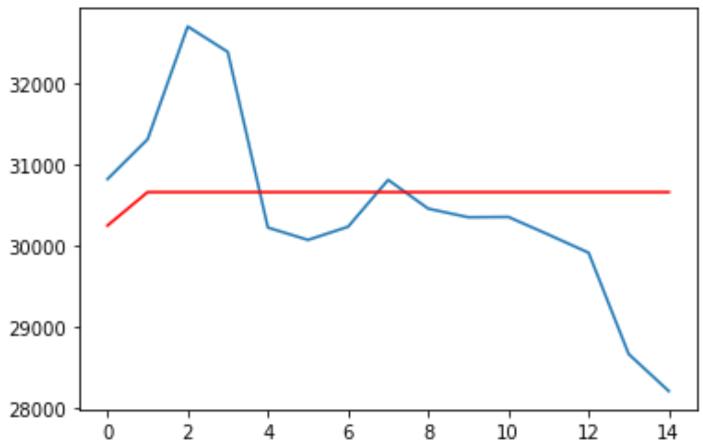
Persistence Model

The persistence model teaches us little apart from providing a baseline for evaluating other models. Upon validation, this model performed with a root mean squared error of 1591.918.

ARIMA Model

The final model showed that attendance is dropping below the predicted average with a root mean squared error of 1146.070 — an improvement over the persistence model.

Predicted Average and Actual Data



Conclusions and Next Steps

A rise in certain in-game measures does highly correlate with a rise in attendance. My first recommendation is that nothing needs to fundamentally be altered in how the game of baseball is played. Although the game has changed in recent years to increase these stats, people seem to appreciate the change.

Next, in recent years, attendance is dropping below the predicted average or, at best, remaining steady. It could still be the case that decreasing ticket prices could increase attendance. If the goal is to soley increase attendance and maintain interest in baseball, lower ticket prices. If the goal is to make money for team owners, no change is necessary, unless attendance continues to drop.

Finally, the recent drop or steadying off of attendance should be investigated further. In-game statistics did not prove greatly fruitful in this investigation. Perhaps, further investigating ticket prices or other exogenous variables will prove to be more promising.