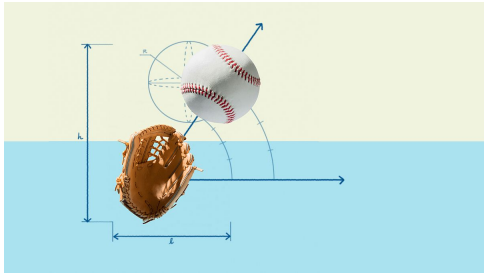
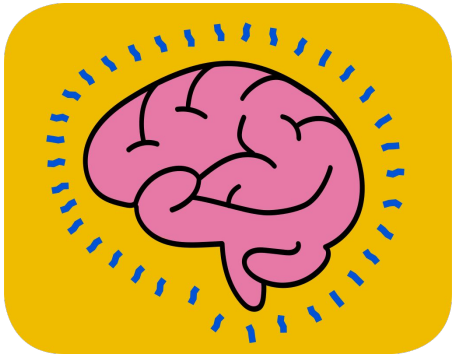
Baseball Attendance Data for Time Series Analysis and Forecasting

Springboard Data Science Capstone Project

Isaac Paulson



Presentation Goals



I can…

- Communicate data findings
- Understand the data science process
- Use Python for time series analysis and forecasting
- Show that I have learned something

Problem

Baseball attendance has been declining

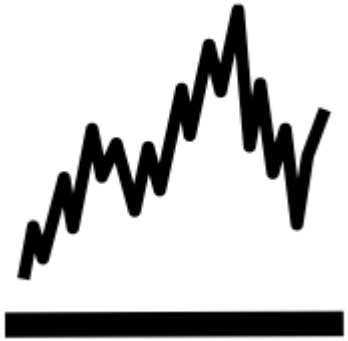
Hypotheses

1. Attendance is affected by on-field measures (e.g. strikeouts, homeruns)
2. Attendance is more affected by off-field measures (e.g. ticket prices, time of game)

The Data

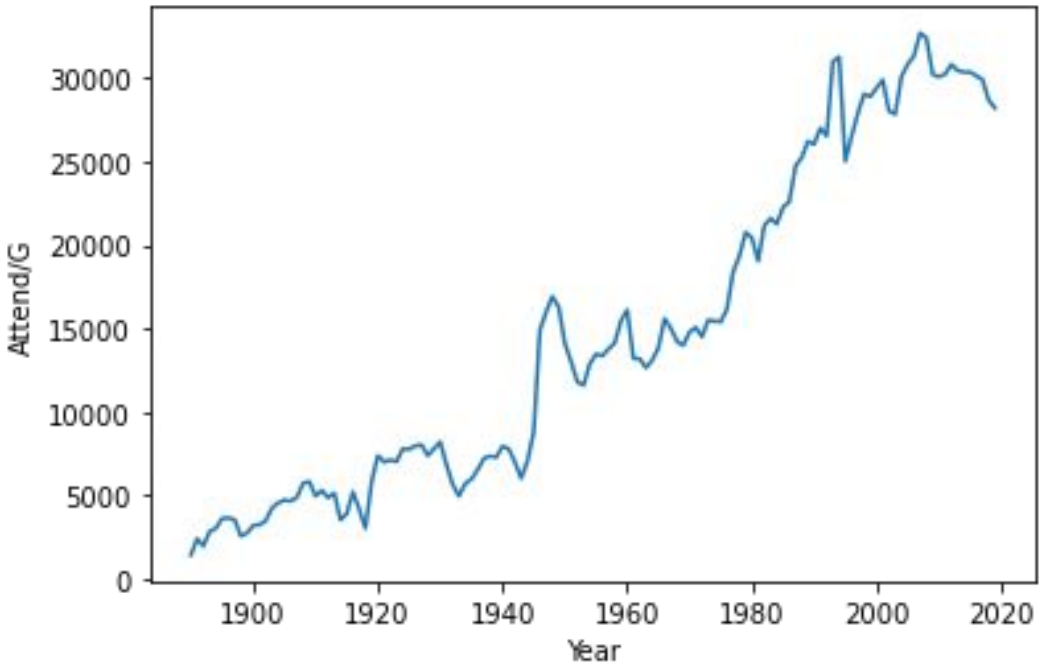
Multivariate time series (by year)

- Attendance per game
- In-game measures (e.g. home runs, strikeouts, etc.)
- Time of game
- Cost per ticket



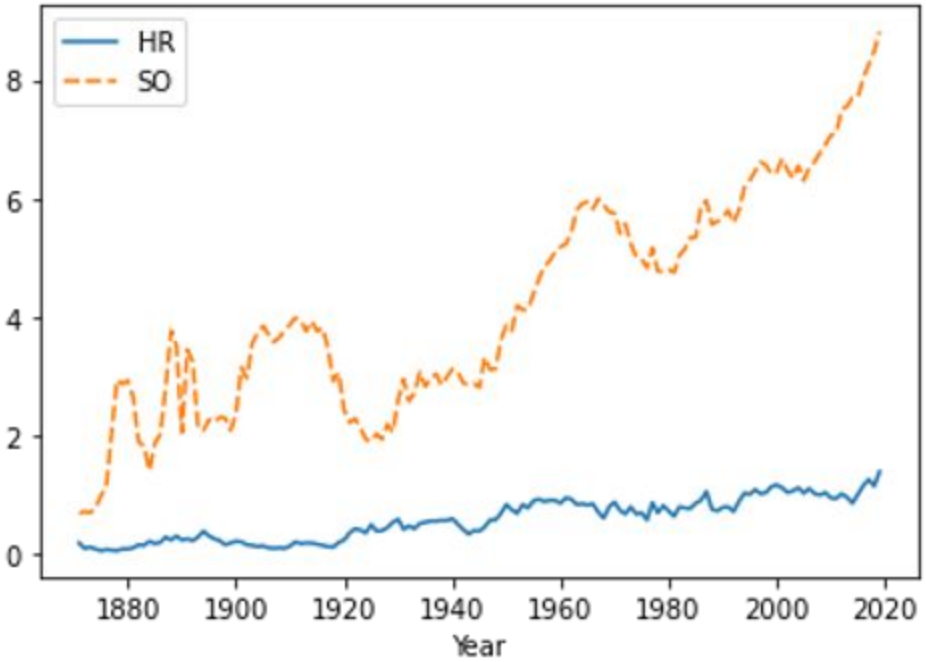
EDA

Attendance per game



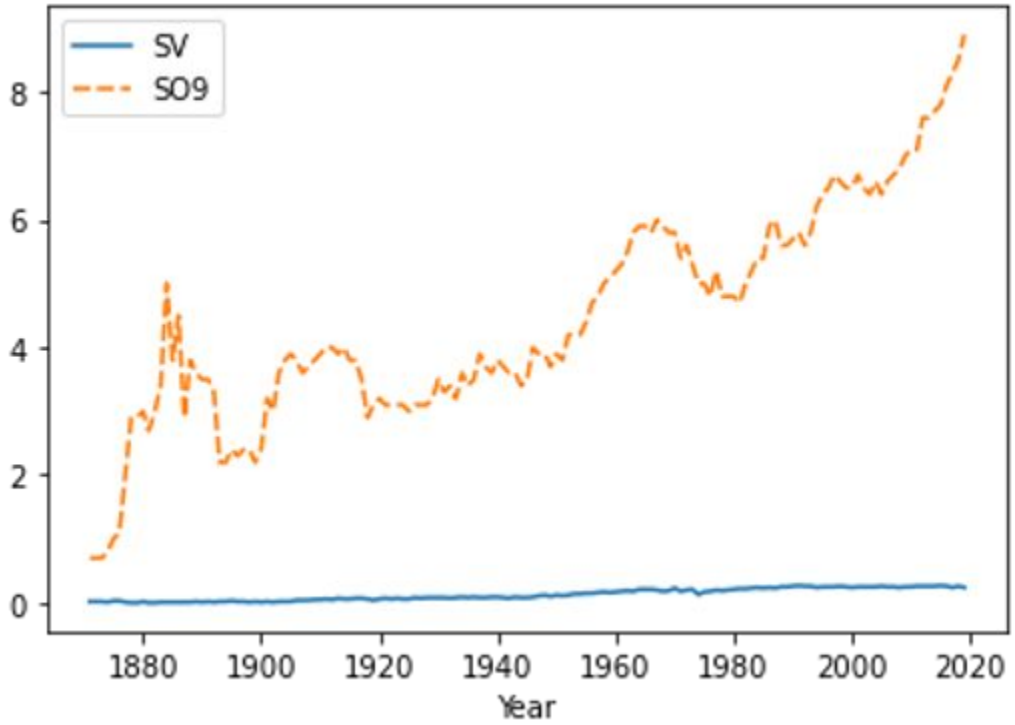
EDA

Strikeouts and Home Runs



EDA

Saves and Strikeouts over 9 Innings



Correlations

- Saves and attendance: **r = .94**
- Ticket prices and attendance: **r = .63**
- Game time and attendance: **r = .95**

Persistence Model

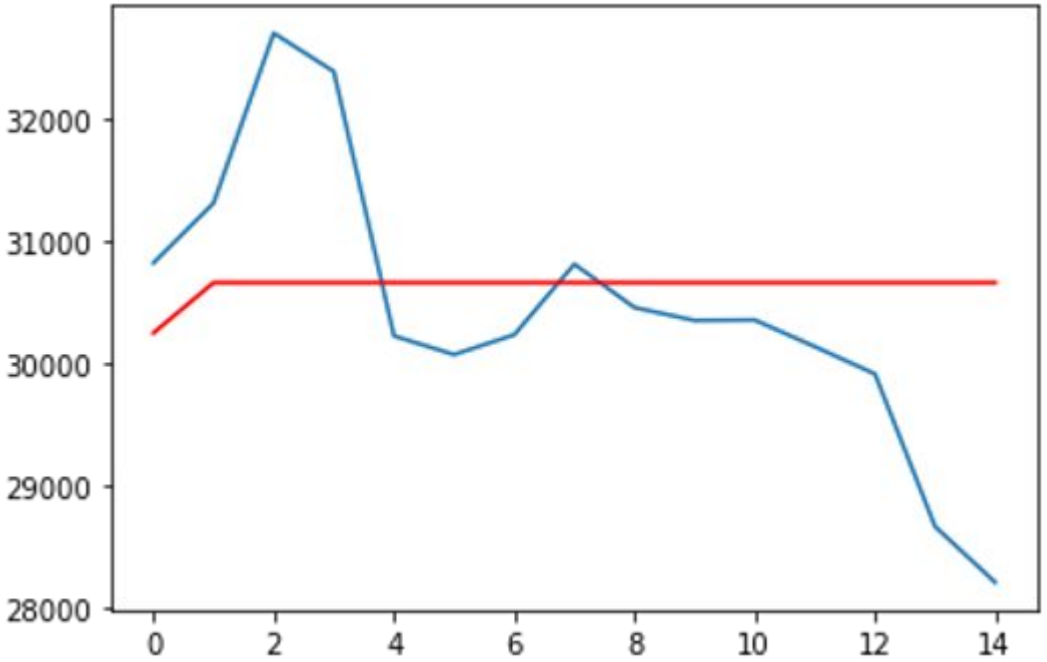
- t - 1
- RMSE = 1591.918

ARIMA Model

- Lag order: p = 0
- Degree of differencing: d = 1
- Order of moving average: q = 2
- Root mean squared error: RMSE = 1146.070

ARIMA Model

Predicted Average and Actual Data



Conclusions and Next Steps

- Changes in in-game statistics correlate to rising attendance
- Attendance is dropping (or maybe holding steady)
- Why is attendance dropping?

What I Learned

- No such thing as a crystal ball
- Working with time series is harder than it looks!