Attendance Forecasting

Michigan State University Basketball

Jake Leto

MSCA 31006 Final Project

Overview

- Problem Statement
- Source Data
- Data Properties & Processing
- Method & Results
- Learnings & Future Work

Problem Statement

Forecast Michigan State University (MSU) Basketball Attendance

Average monthly attendance

First quarter of 2018

Gain insight into long term trends

Quantify correlations between attendance and key variables

Source Data

Dataset:

- Obtained from NCAA public data (BigQuery Database)
- mbb_teams_games_sr: game summary data from the 2013-14 to 2017-18 seasons
- mbb_historical_teams_games: final scores from 1996-97 to 2017-18 seasons
- mbb_teams: team information including venue capacity, id, division name, etc.
- Summary tables contain attendance numbers

Challenges:

- Game level-of-detail (fouls, rebounds, shot percentage, etc.) only included in 2013-2018 seasons

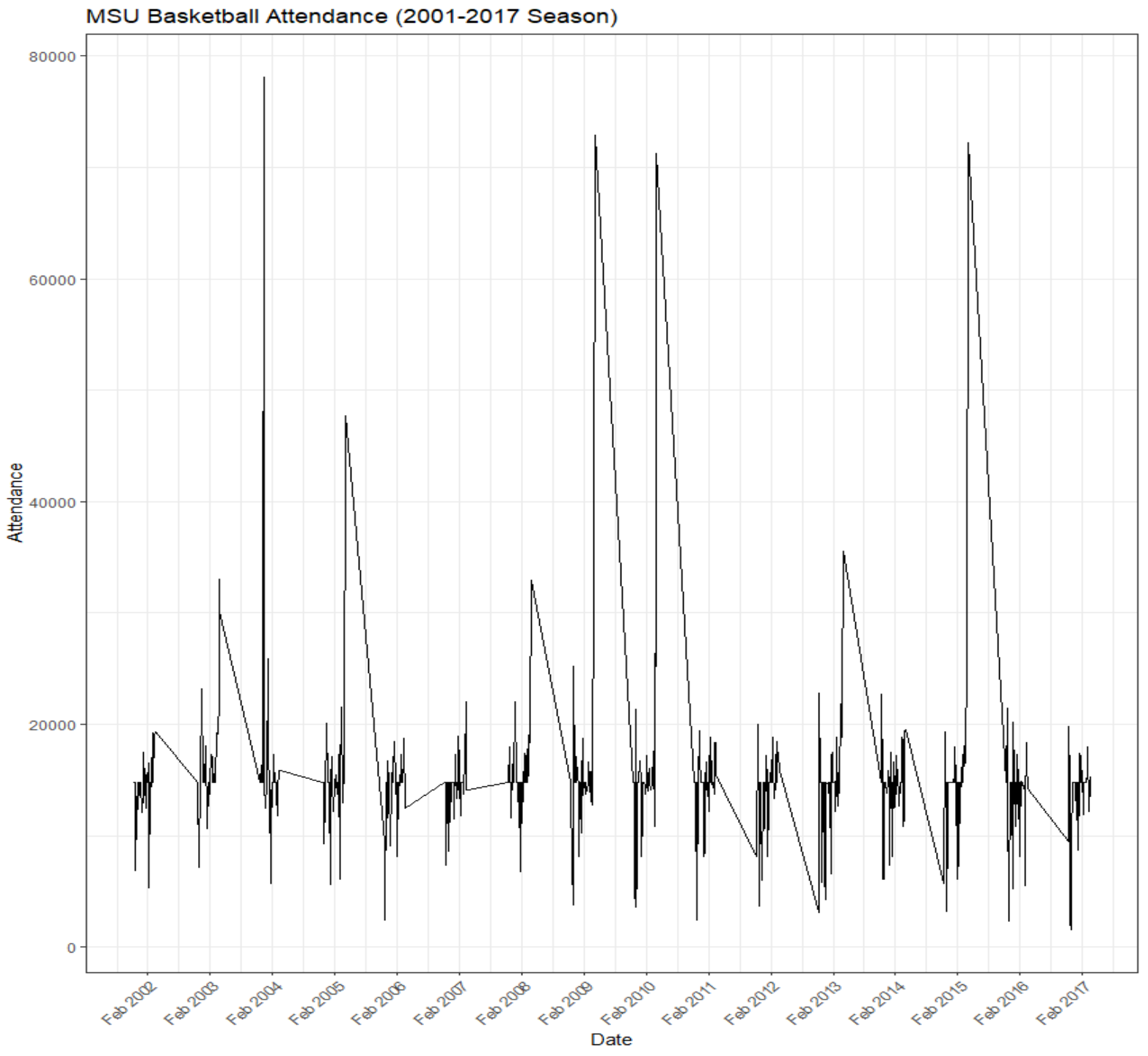
Data Properties & Processing

Filtered bigquery tables on team_id to obtain data for MSU only:

msu_game_summary_2001_2017.csv
msu_game_summary_2013_2018.csv

2001-2017 Dataset

- 563 games, 20 variables
- Game date, venue capacity, points scored (for both teams) are included
- Attendance timeseries has seven anomalies
  - Large spike in attendance
  - Likely tournament games



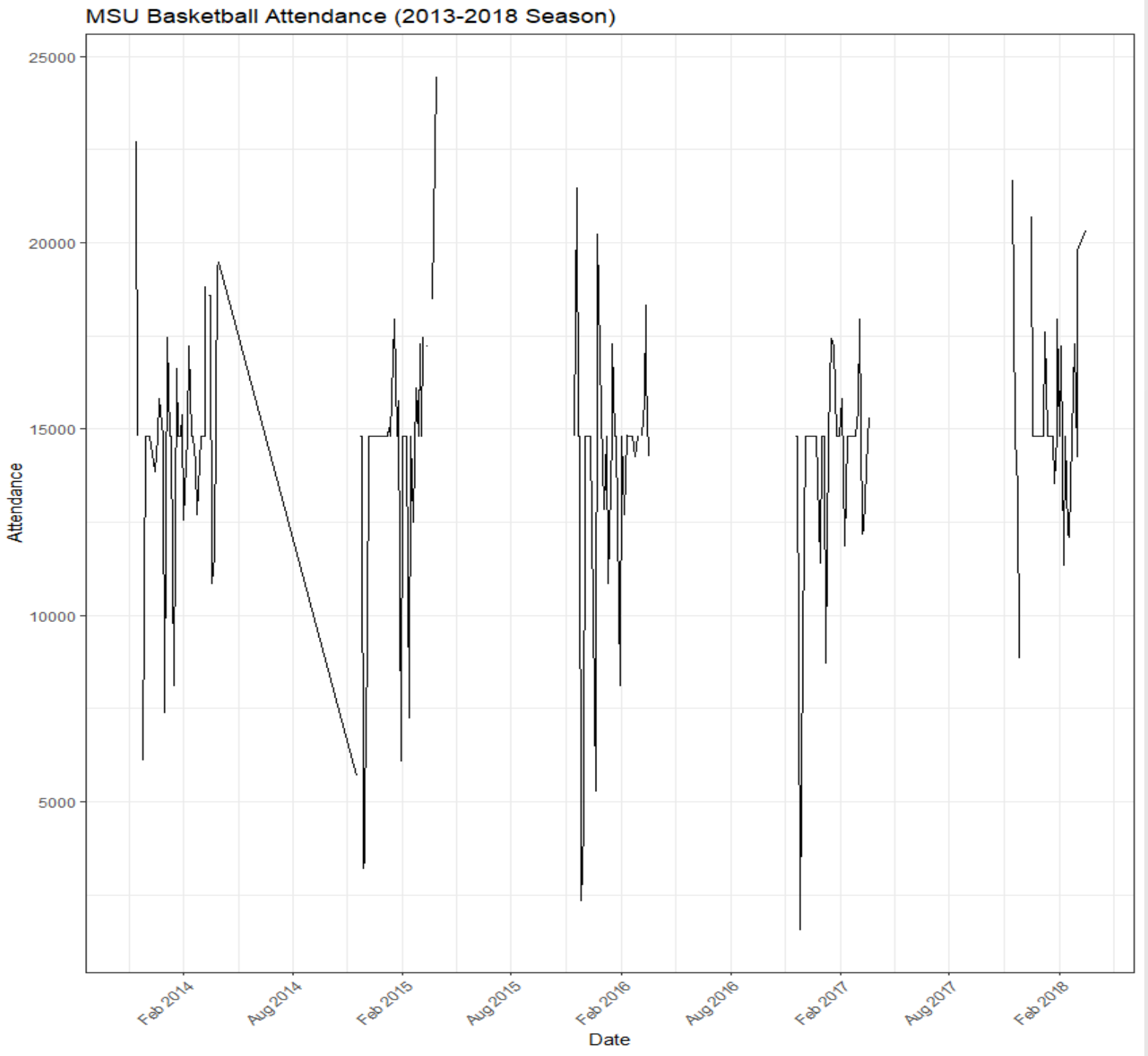
Data Properties & Processing

Filtered bigquery tables on team_id to obtain data for MSU only:

msu_game_summary_2001_2017.csv
msu_game_summary_2013_2018.csv

2013-2018 Dataset

- 182 games, 132 variables
- Game date, venue capacity, points scored (for both teams) are also included
- Game-level detail
- Attendance timeseries is non-uniform and discontinuous (missing values)



Data Properties & Processing

Addressing Anomalies & Non-Uniformity

- Normalize attendance by creating new metric (percent capacity)
- Percent capacity = attendance/venue capacity
- Linearly impute percent capacity for non-game days/missing values and aggregate to the monthly level
- Join 2017-18 season from 2013-2018 dataset to 2001-2017 dataset

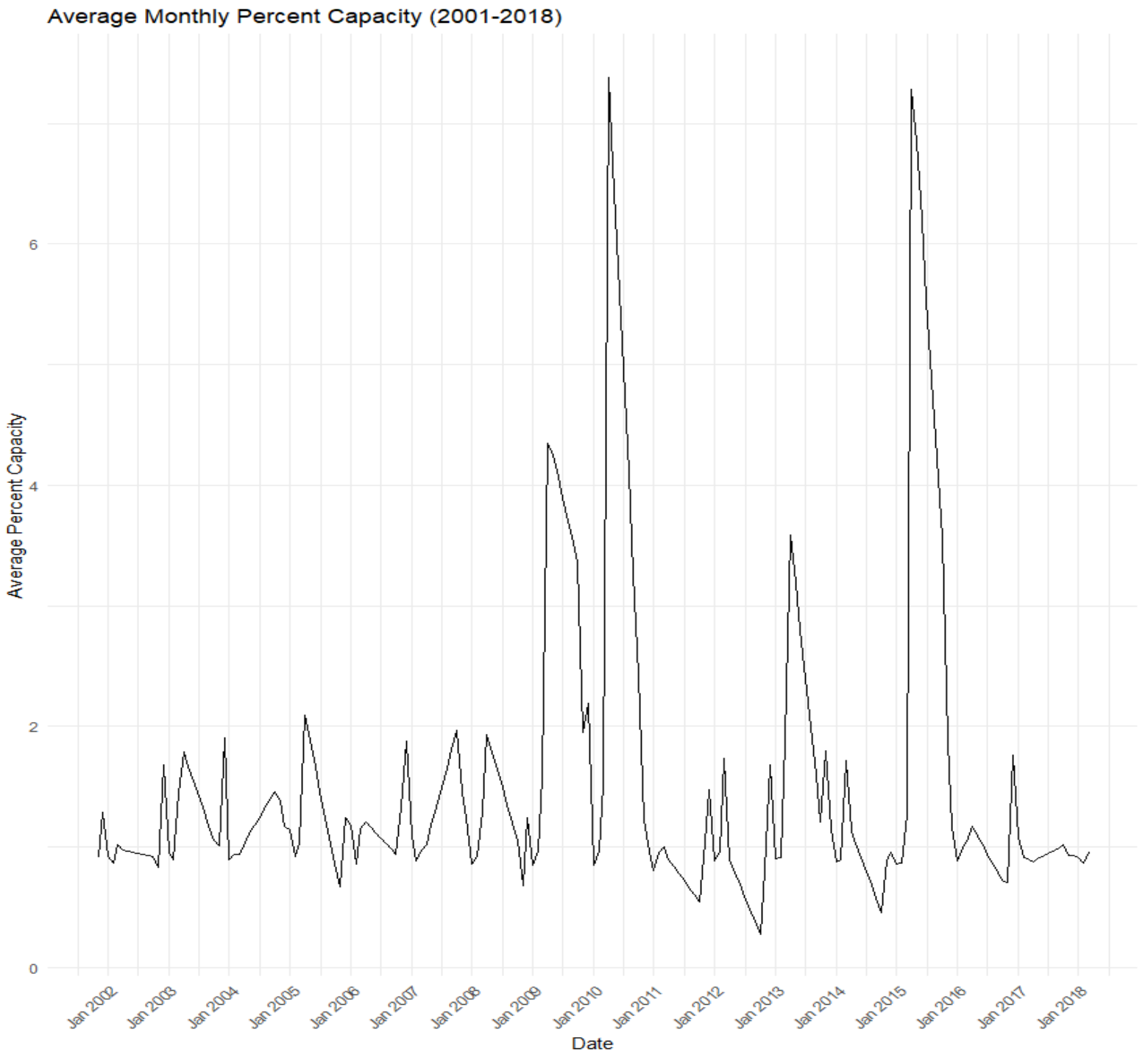
Challenges:

- Game venue is not included in 2001-2017 data
- Solution:
  - Randomly select 50% of games in each season and assign them as “away” games
  - Use opponent’s venue capacity in calculation for away games

Data Properties & Processing

2001-2018 Aggregated Data

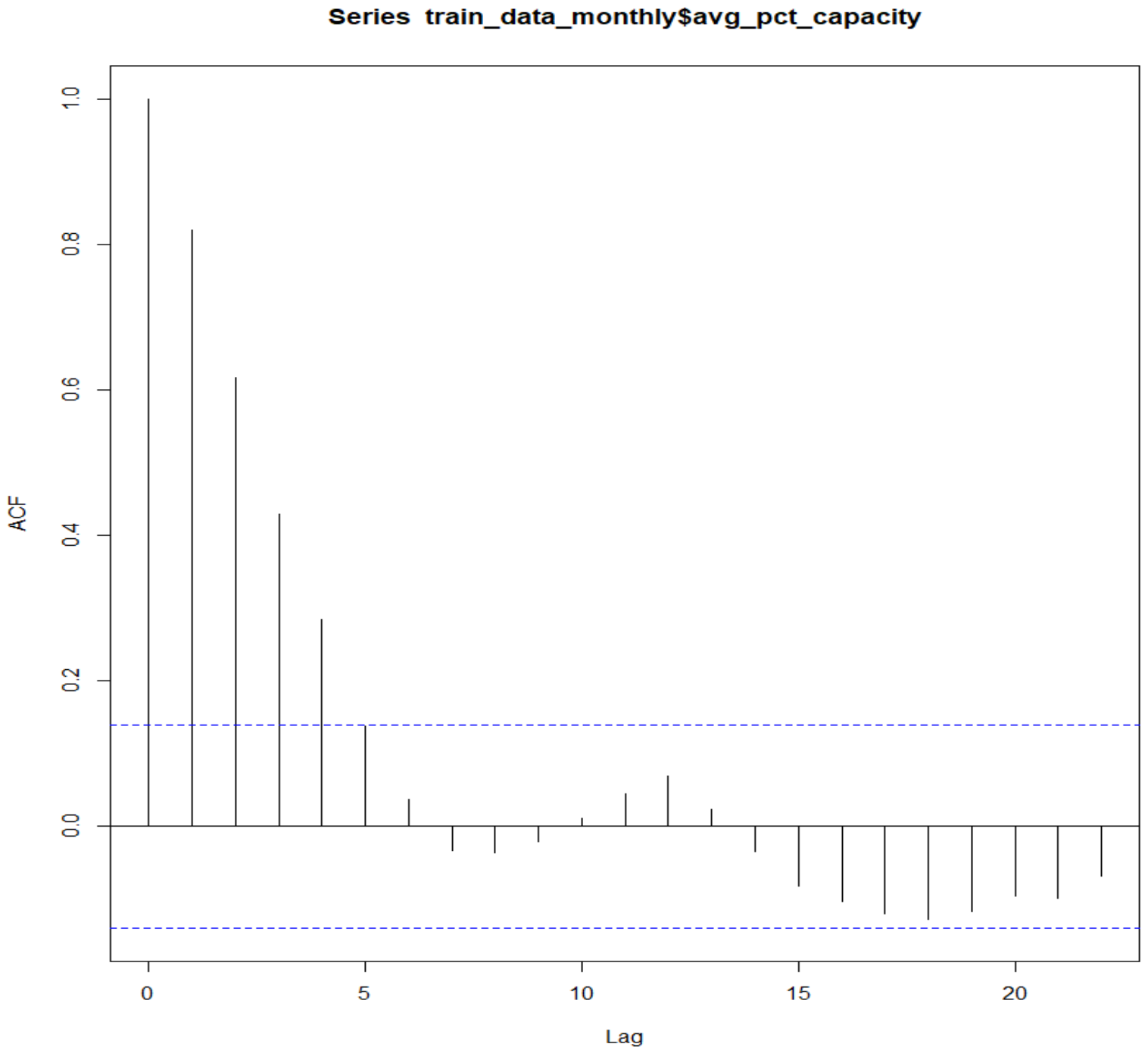
- Uniform timeseries – monthly
- Stationary
  - ADF Test p-value = 0.01
  - KPSS Test p-value = 0.1
- Notable Correlations (using game information from the 2013-2018 dataset):
  - Opponent Field Goal Percent = 0.18913
  - Rebounds = -0.13683
  - Blocks = 0.11768
- Anomalies: percent capacity > 1
  - Most likely due to random assignment of away games



Method & Results

Model Selection

- Stationarity and visual inspection of the autocorrelation function suggests that the timeseries in autoregressive
- Use auto.arima model as baseline
- Find optimal model parameters by varying p and q
- Measure accuracy of model
  - AIC / BIC
  - MAE
  - RMSE
  - MAPE



Method & Results

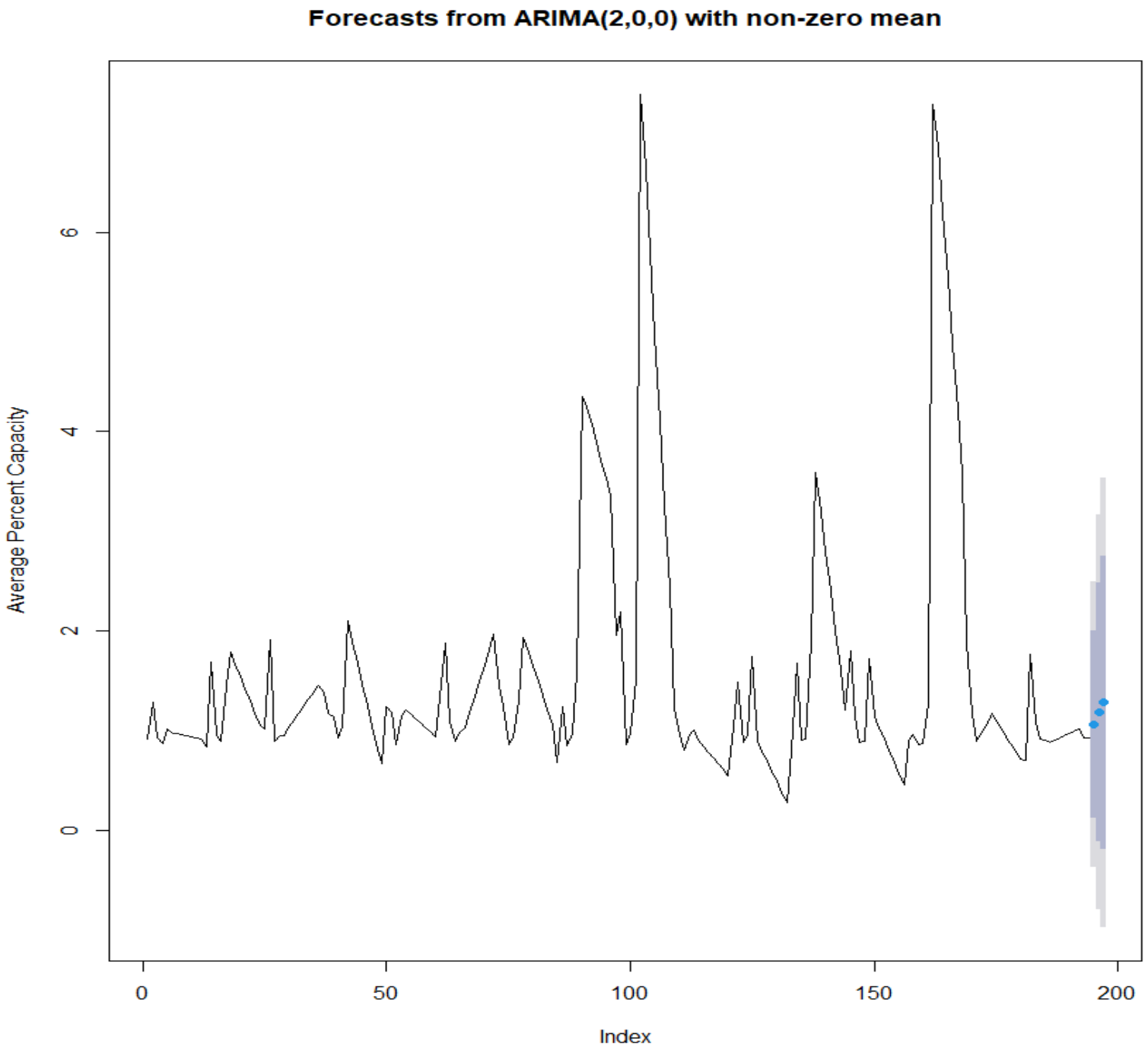
Baseline Model

Training:

- AIC = 436.31
- BIC = 436.31
- MAE = 0.344926
- RMSE = 0.727258
- MAPE = 24.45252

Test:

- MAE = 0.258987
- RMSE = 0.271528
- MAPE = 21.71277



Method & Results

Adjusted Model

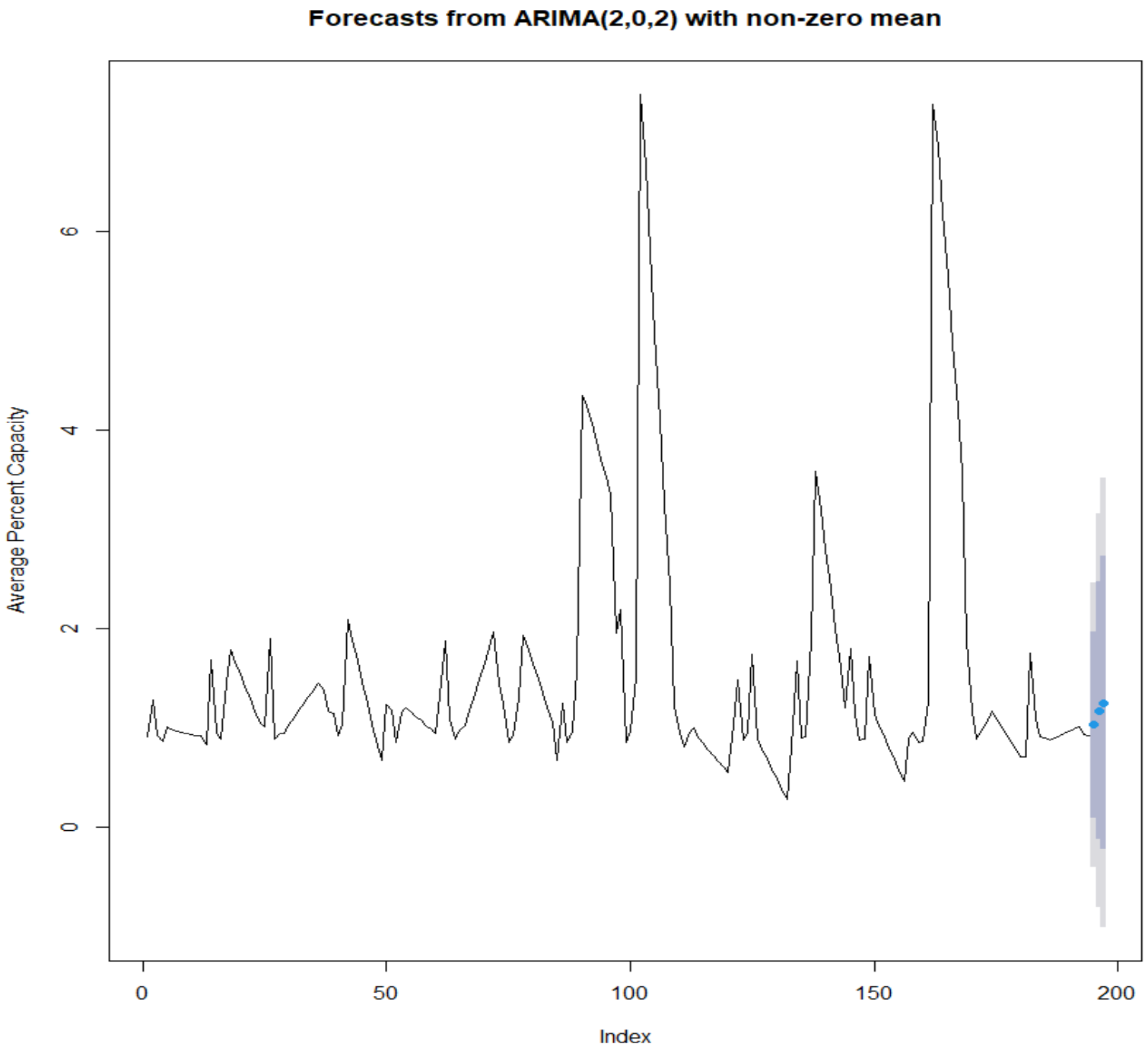
Training:

- AIC = 436.92
- BIC = 456.53
- MAE = 0.347656
- RMSE = 0.721031
- MAPE = 24.37465

Test:

- MAE = 0.238611
- RMSE = 0.253858
- MAPE = 20.26460

Marginal increase in accuracy when adjusting q to 2.



Learnings & Future Work

https://github.com/letojake/MSCA/tree/7fe9afc2acb7de94c20524dba74440d319b79f6a/MSCA31006/Final%20Project

Learnings:

- Methodology is largely driven by data structure
- Model selection process can be difficult
- Preprocessing and feature engineering are the most critical parts of the process

Future Work:

- Extend analysis to include dependence on game specific information (shot percentage, fouls, rebounds, etc.)
- Explore other transformation methods
  - Forecast Box-Cox transformed attendance
- Update dataset to include 2019-Current seasons