Worcester Red Sox X HDAG Final Presentation

Monday, December 18th, 2023

Dries, Sarah, Diego, Kevin, Max

Presentation Agenda

01

The Fall in overview.

Overview of project goals, redirection towards classification models and tools built.

02

Singles game visualizations

Key actionable insights for singles tickets, start of our analysis.

03

Modelling

Single-level models, game-level models, time-series models

04

Progress on geographic visualization and future direction

15-20 minutes for Q&A, reflections from analysts, feedback exchange, and discussion of next steps (including code handoff).

Team Introductions

Dries Rooryck (‘26)
Case Team Lead

Sarah Cao (‘25)
Associate

Diego Gonzalez Gauss (‘27)
Analyst

Kevin Liu (‘27)
Analyst

Max Wagner (‘27)
Analyst

Project Timeline

Onboarding & data acquisition

Acquiring, validating, and cleaning data needed to achieve project objectives; clarifying project goals & building understanding of previous work

EDA & Visualization

Gaining familiarity with datasets & variables of interest; converting previous data visualizations from Excel to Python and augmenting using county demographic data.

Midpoint Presentation

Update on current progress, space to redirect and refocus attention and receive feedback.

Further Analysis & Modelling

Using current insights to guide further data analysis; conducting outside research & implementing models as needed; developing recommendations

Final Presentation

Presenting key insights and final recommendations

Key Objectives from Midpoint

End Goal: maximize turnout at weekday games

1. **Understanding the demographics of fan base:** Explore market behavior at different times during season and driving factors for demand
2. **Weather regression and forecasting:** the effect of rainouts, temperature, and other predictable
3. **Solidifying code in notebooks for production:** Three notebooks are available and ready with interactive features.

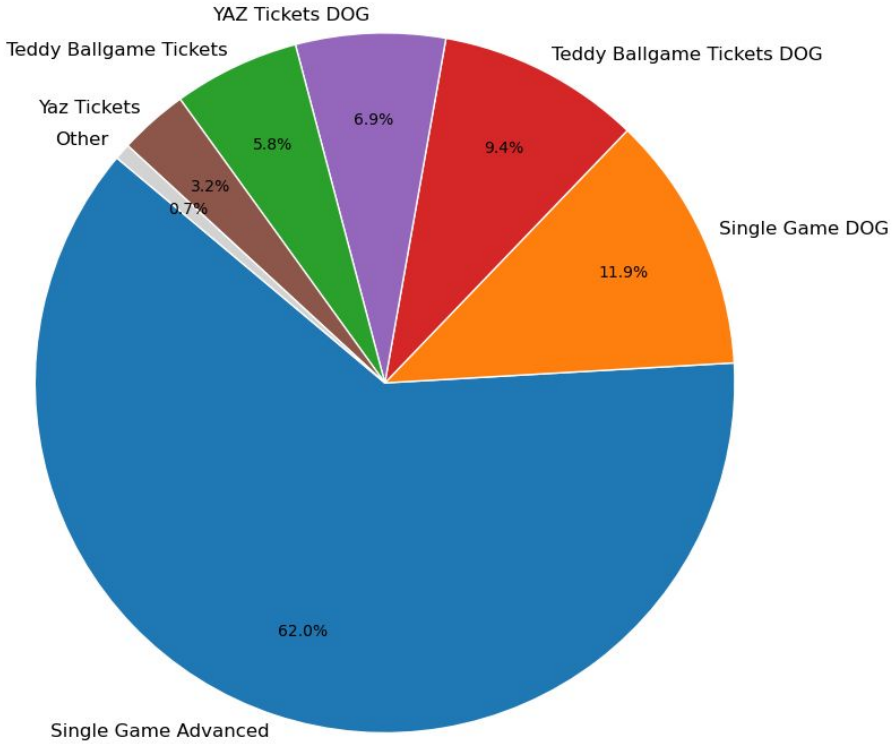
Three projects

Technical deliverables

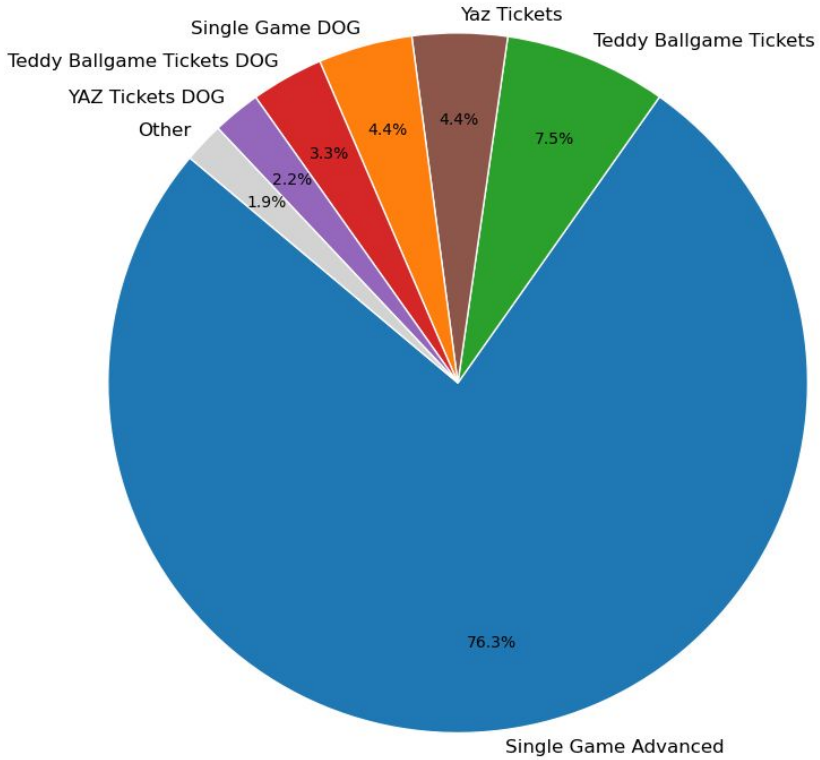
1. **Interactive geographic data visualization:** RShiny web-application
2. **Person-level modelling:** Build a model to predict ticket-to-turnstile ratios at games.
3. **Game-level modelling:** Model ticket-to-turnstile ratios at games using by-game data.

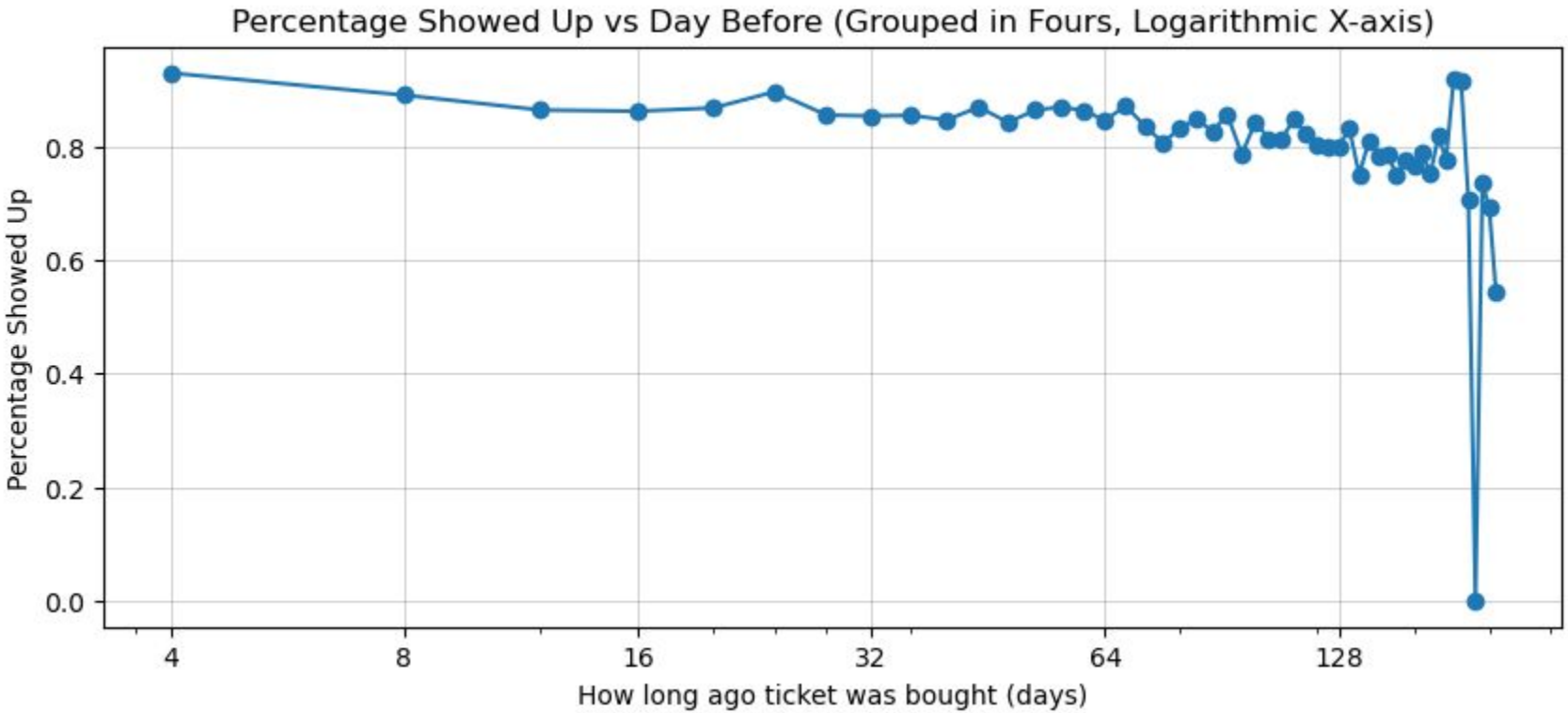
(I) Single Tickets Visualizations

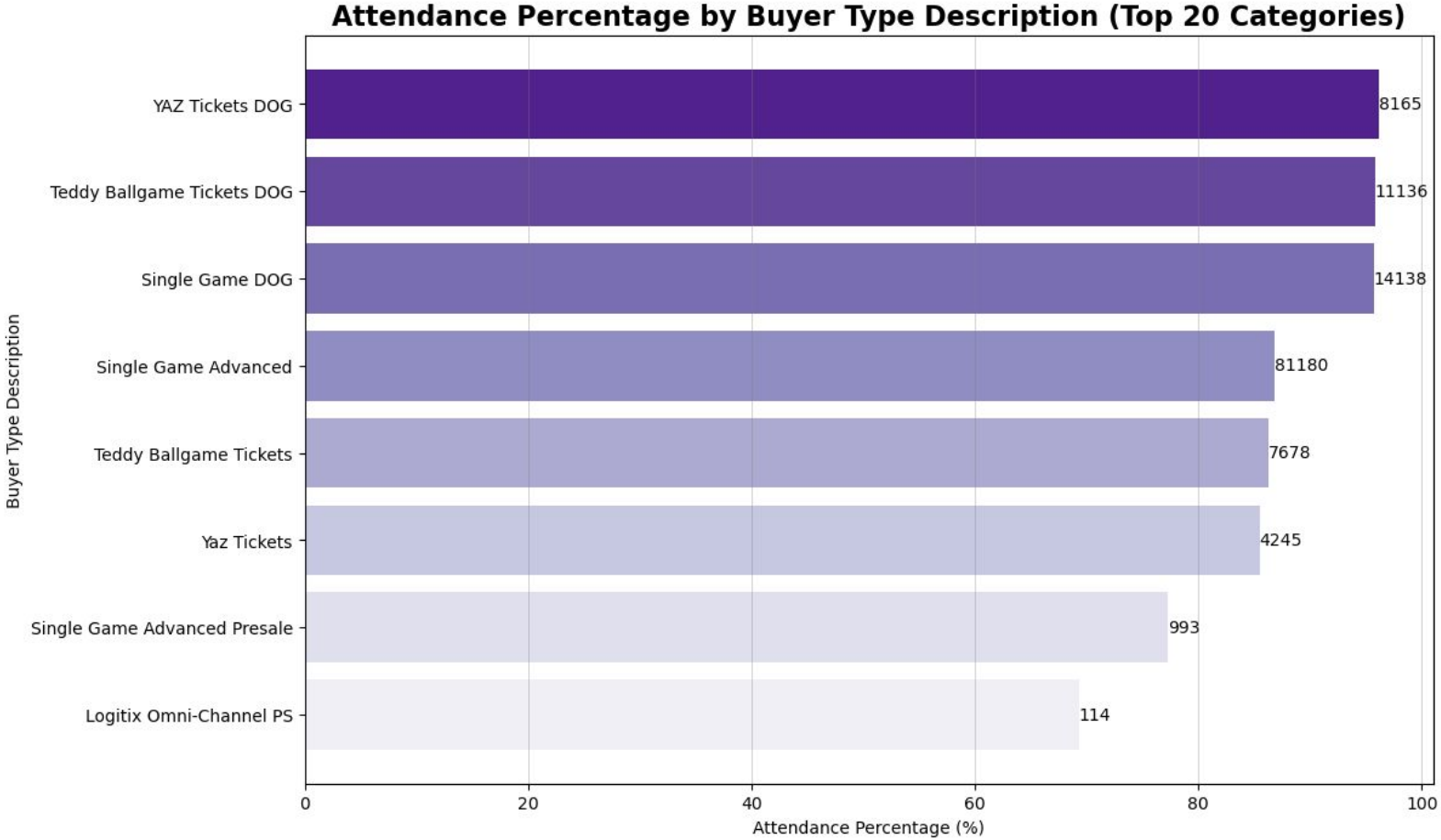
Distribution of Buyers Who Showed Up



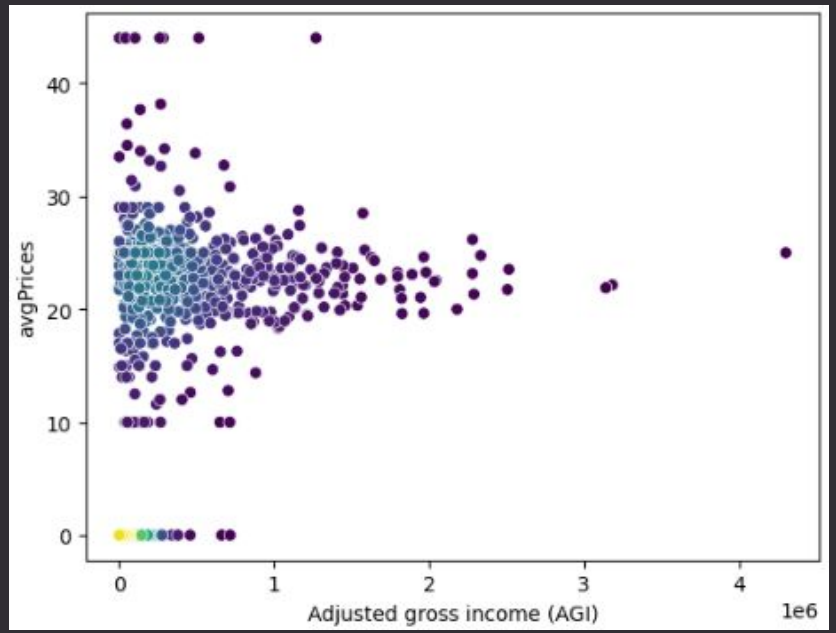
Distribution of Buyers Who Did Not Show Up



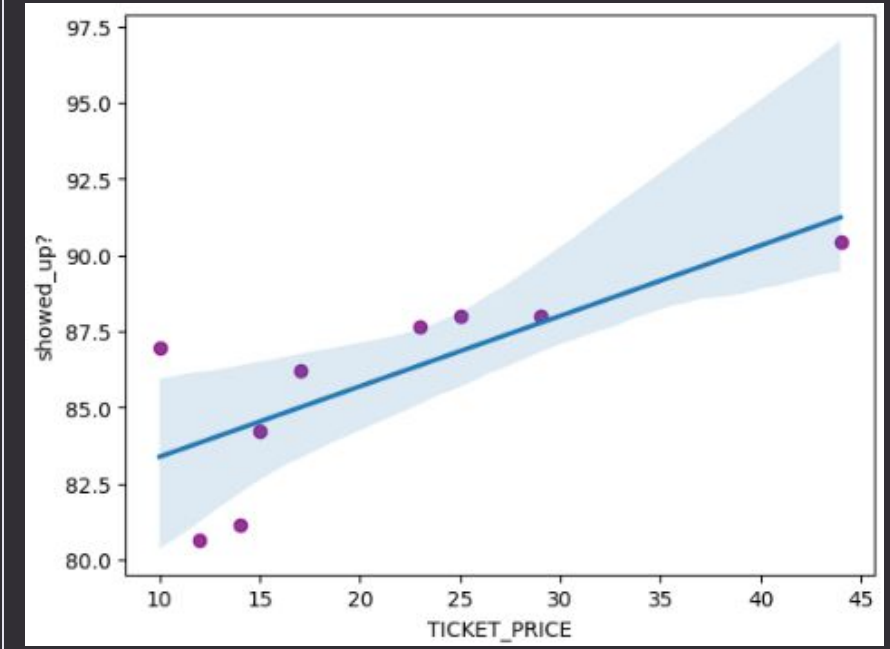

Percentage Showed Up vs Day Before (Grouped in Fours, Logarithmic X-axis)
Percentage Showed Up
0.0
0.2
0.4
0.6
0.8
4
8
16
32
64
128
How long ago ticket was bought (days)


Attendance Percentage by Buyer Type Description (Top 20 Categories)
YAZ Tickets DOG
Teddy Ballgame Tickets DOG
Single Game DOG
Single Game Advanced
Teddy Ballgame Tickets
Yaz Tickets
Single Game Advanced Presale
Logitix Omni-Channel PS
8165
11136
14138
81180
7678
4245
993
114
Buyer Type Description
0
20
40
60
80
100
Attendance Percentage (%)

Ticket Price vs. Income



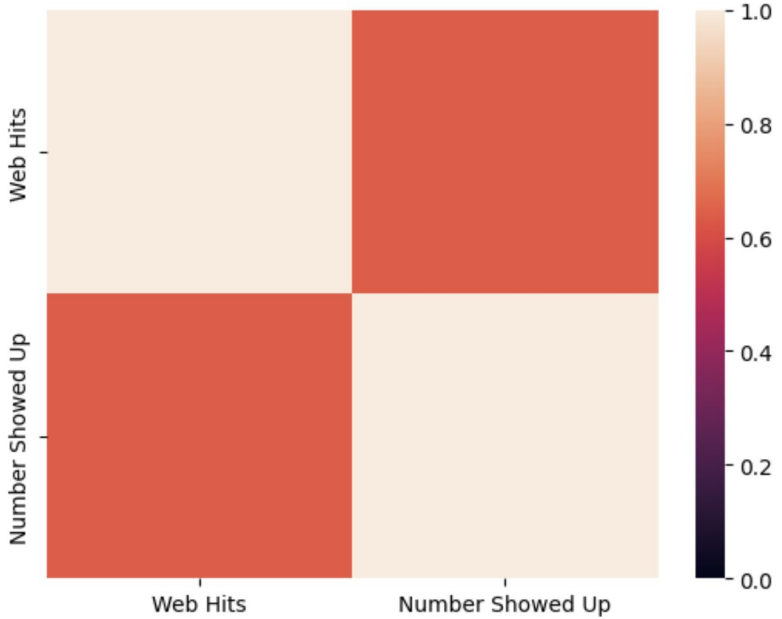
Ticket Price vs. Attendance



Relating to Web Hits

- Do number of web hits (visiting ticket page) correlate with people showing up to games?

- Zip Code Based
  - # of fans:
    - Showed up to WooSox game per zip code
    - Visited ticket page per zip code
- Correlation: r = 0.637



(2) ML modeling work:

(a) on the level of the individual:

What causes fans to not attend games they buy tickets for?

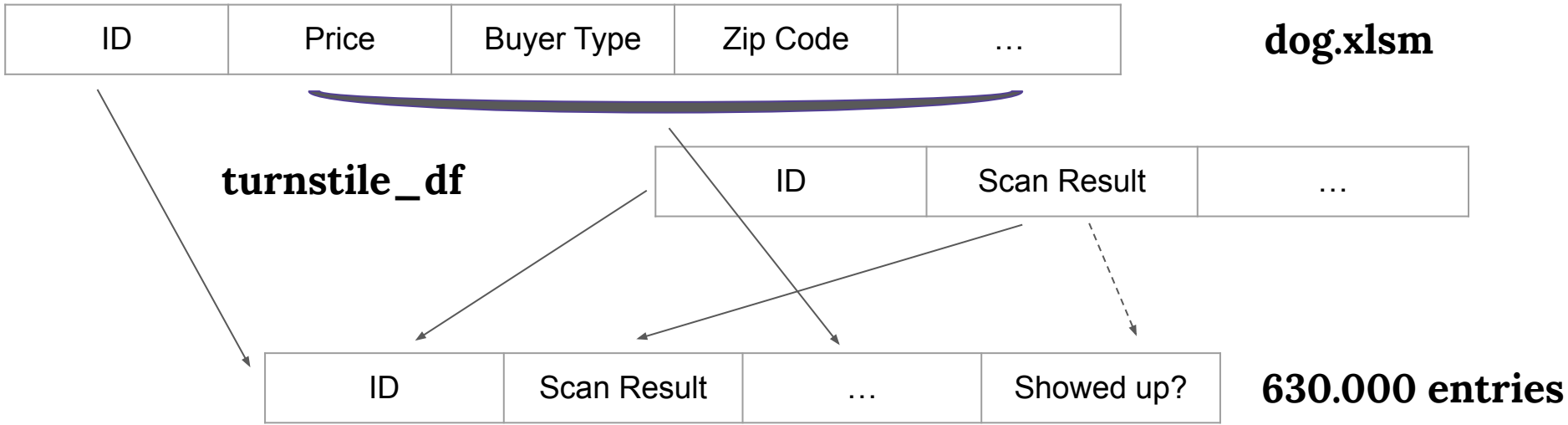
Finding Factors that Influence Attending Game Decision

- **Ticket Price**
  - Will a more expensive ticket compel a fan to attend a game?
- **Day Before**
  - How many days before did the fan buy the ticket before the game?
    - Plans could change
- **Series Week**
  - Time of the year, opponent, etc. could be influencing factors
- **Weekly Promo**
  - Specific days of the week have different promos, which could influence a fan's excitement towards the game

Finding Factors that Influence Attending Game Decision

- Zip Code Dependent Factors
  - **Driving Time**
    - How far is Polar Park from the zip code where the fan bought the ticket, on average?
    - Longer driving time could decrease a fan's willingness to attend a game
  - **Zipcode AGI** (average adjusted gross income)
    - Fans from different zip codes come from different backgrounds and wealth levels

Merging ticket sales and turnstile



Data Preparation

- First preparing the data:
  - SMOTE sampling (Synthetic Minority Over-sampling Technique)
    - Notice the high imbalance between **people NOT showing up to games** and **people showing up to games**
    - Creating ‘synthetic’ data points in the minority (people not showing up to games)

	Ticket Price	Day Before	Series Week	Weekly Promo	FINANCIAL_POSTAL_CODE	Driving Time	Zipcode AGI	showed_up?
91316	8.0	56.0	1.0	5	1608.0	2.6833	47.525466	1
91317	8.0	56.0	1.0	5	1608.0	2.6833	47.525466	1
91318	8.0	16.0	1.0	5	1608.0	2.6833	47.525466	1
91319	8.0	16.0	1.0	5	1608.0	2.6833	47.525466	1
91320	8.0	0.0	1.0	5	1608.0	2.6833	47.525466	1

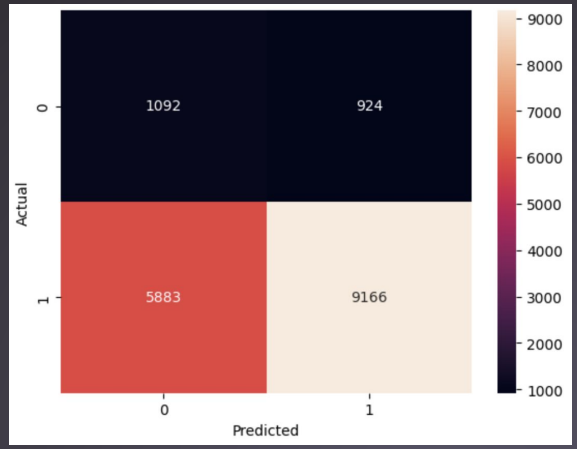
Initial Models

Features:

```
['Ticket Price', 'Day Before',
'Series Week', 'Weekly Promo',
'Driving Time', 'Zipcode AGI']
```

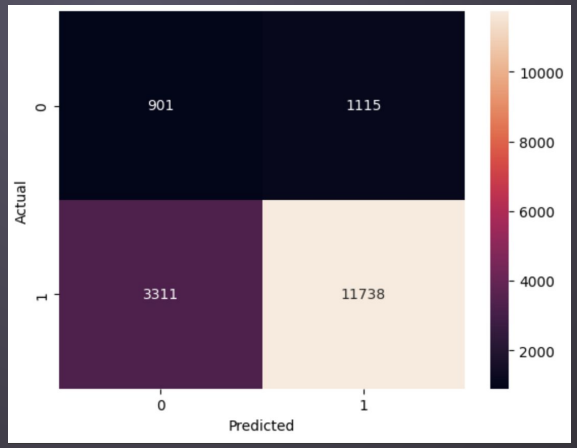
Multiple Logistic Regression: **60.11% accuracy**

AdaBoostClassifier: **74.06% accuracy**



Logistic Regression

AdaBoost



Random Forest Model

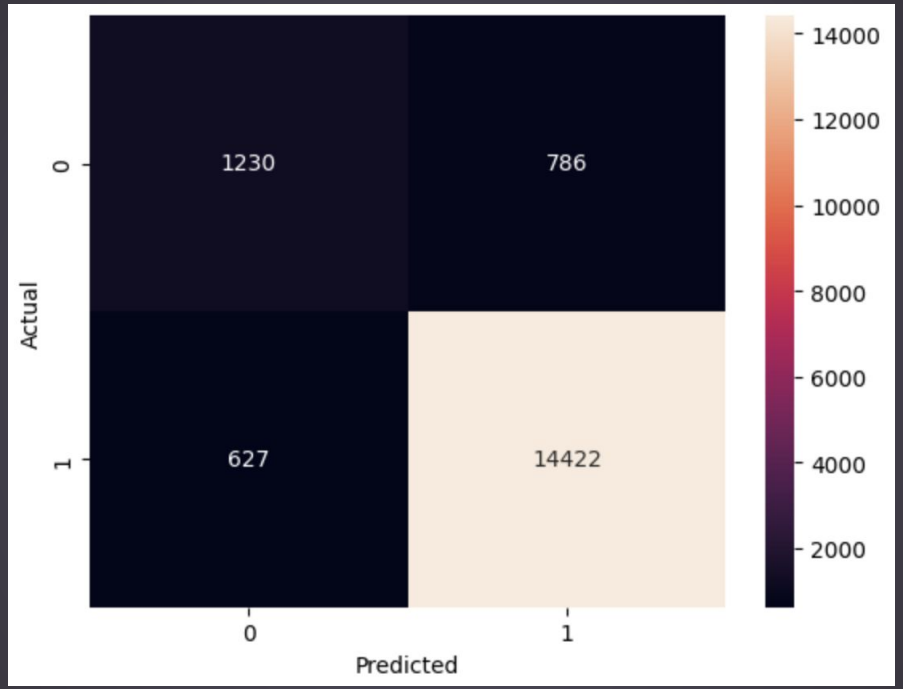
Excellent model for predicting

Chooses a random subset of features and changes them in different ways

Features:

```
['Ticket Price', 'Day Before',
'Series Week', 'Weekly Promo',
'Driving Time', 'Zipcode AGI']
```

91.66% accuracy

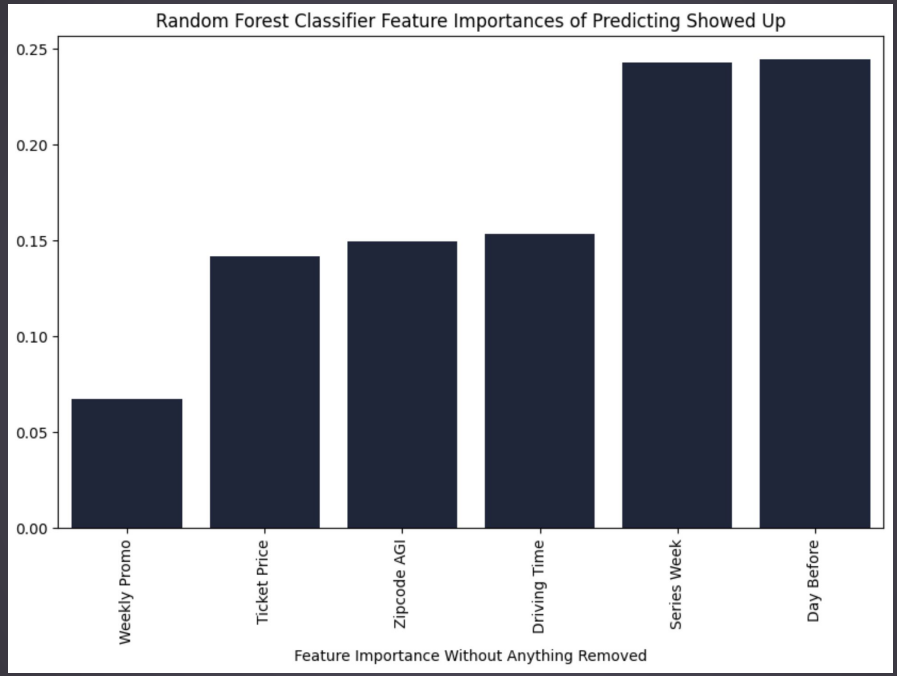


Random Forest Model Feature Importance

- ***"Day Before"*** feature and ***"Series Week"*** are most influential, according feature importance graph

- ***"Driving Time", "Zipcode AGI", "Ticket Price"*** are all relatively the same importance, however severe drop off from earlier

- ***"Weekly Promo"*** does not influence willingness to attend games



Random Forest Model Robustness

Sensitivity Analysis

Used to test how robust (resistant to change) the model is.

Dropping one input feature at a time

As a result, how do the results change?

Analysis of Random Forest Model

Change in Rank of Each Feature Influencing Fans Showing up to Games

Day Before	-	1	0	0	1	0
Series Week	-1	-	0	1	-1	0
Driving Time	-1	-1	-	-1	0	0
Zipcode AGI	-1	-1	-1	-	0	0
Ticket Price	-1	-1	-1	-1	-	0
Weekly Promo	-1	-1	-1	-1	0	-

(b) on the level of the ballgame:

What impacts attendance on a gamewide level?

Some features that could influence overall game attendance

- **Weather variables**
  - Could encourage/discourage possible attendees
  - Seasonal!
- **Date, series week**
  - Shapes possible attendees' idea of a game's "stakes"
  - Also seasonal!
- **Competitive events** e.g. MLB, the Revolution
  - Could attract possible WooSox attendees
  - (Less) seasonal

More data preparation

- Data cleaning
- Building intuition for data
  - What can be linearized?
  - Finding attendance ratios
- Handling rainout games and other extremities
  - Consider both

days since rev	days until rev	days to MLB g	days to Revol	Series Week	NT_USAGE_D	WeeklyPromo	Ballpark Then	Special WooS	Package	Group	Single	Comp	Ticket Sold	Turnstile	Ratio	tempmax
13	1	1	1	1	**45016**	**FIREWORK**	Opening Day		3220	1168	3519	1109	**7907**	**5497**	0.6952067788	48.9
0	0	0	0	1	**45017**	**CATCH**	Opening Weekend		3020	353	5027	381	**8400**	**2951**	0.3513095238	63
1	6	0	1	1	**45018**	**FUNDAY**	Opening Weekend; Wepas de		2990	761	5200	580	**8951**	**4528**	0.5058652665	51.3
3	11	3	3	2	**45027**	**T&T**			2817	348	3618	389	**6783**	**2470**	0.3641456583	72.1
4	10	2	4	2	**45028**	**WOOF**			2918	213	3708	495	**6839**	**2561**	0.3744699517	72.1
5	9	1	5	2	**45029**	**THROW**	WooU Night	413 Night	2996	495	4275	475	**7766**	**3851**	0.4958794746	86.1
6	8	0	6	2	**45030**	**FIREWORK**	Wepas de Worcester		2816	988	5633	766	**9437**	**6743**	0.714527922	89.1
7	7	0	7	2	**45031**	**CATCH**	Bark in the Par	2:49 p.m. Mara	2789	741	5200	679	**8730**	**5258**	0.6022909507	74
8	6	0	6	2	**45032**	**FUNDAY**	Marvel Super Hero Day		2761	582	4680	567	**8023**	**4217**	0.5256138602	59.1
3	4	3	3	3	**45041**	**T&T**	Portuguese Heritage Night		2874	128	3472	450	**6474**	**1823**	0.28158789	56
4	3	2	3	3	**45042**	**WOOF**	Shrewsbury Town Takeover		2997	179	3487	446	**6663**	**1859**	0.2790034519	55.2
5	2	1	2	3	**45043**	**THROW**	Framingham Tr	Derek Lowe's r	2886	433	3393	404	**6712**	**2180**	0.3247914184	55.2
6	1	0	1	3	**45044**	**FIREWORK**	**Deaf & Hard of Hearing Awar**		3223	1134	1671	740	**6028**	**4676**	0.7757133378	60.3
0	0	0	0	3	**45045**	**CATCH**	**Autism Acceptance**		2798	1440	1373	508	**5611**	**3009**	0.5362680449	52.2
1	27	0	1	3	**45046**	**FUNDAY**	Grafton Town Takeover		2949	698	946	389	**4593**	**1798**	0.3914652732	51.3
3	25	0	3	4	**45048**	**T&T**	Irish Heritage Night		2908	569	3249	398	**6726**	**1684**	0.2503716919	51.3
4	24	0	4	4	**45049**	**WOOF**			3057	523	3414	391	**6994**	**1360**	0.1944523878	51.3
5	23	0	5	4	**45050**	**THROW**	May the Fourth Be With You		3122	618	3268	432	**7008**	**2095**	0.2989440639	48
6	22	1	6	4	**45051**	**FIREWORK**	Wepas de Worcester II		2911	881	931	526	**4723**	**3053**	0.6464111793	59
7	21	2	7	4	**45052**	**CATCH**	**Breast Cancer Awareness**		2630	1303	2550	548	**6483**	**5174**	0.7980873053	70.2

Some extreme % attendances

14.0%

September 13, 2023

- Lowest attendance day
- Final week, rain, gusty, low visibility

35.1%

April 1, 2023

- Lowest attendance of opening weekend
- MLB and Revolution games, rain/snow

87.8%

July 7, 2023

- Highest attendance day
- ~July 4th, good weather, fireworks promo

Regression modeling

- Model: OLS regression
  - Conventional linear regression

- Features: `['index', 'Series Week', 'Ticket Sold', 'singlesattendance', 'days since holiday', 'days to holiday', 'days since MLB game', 'days until MLB game', 'days since revolution game', 'days until revolution game', 'temp', 'feelslike', 'precip', 'windspeed', 'cloudcover', 'visibility']`

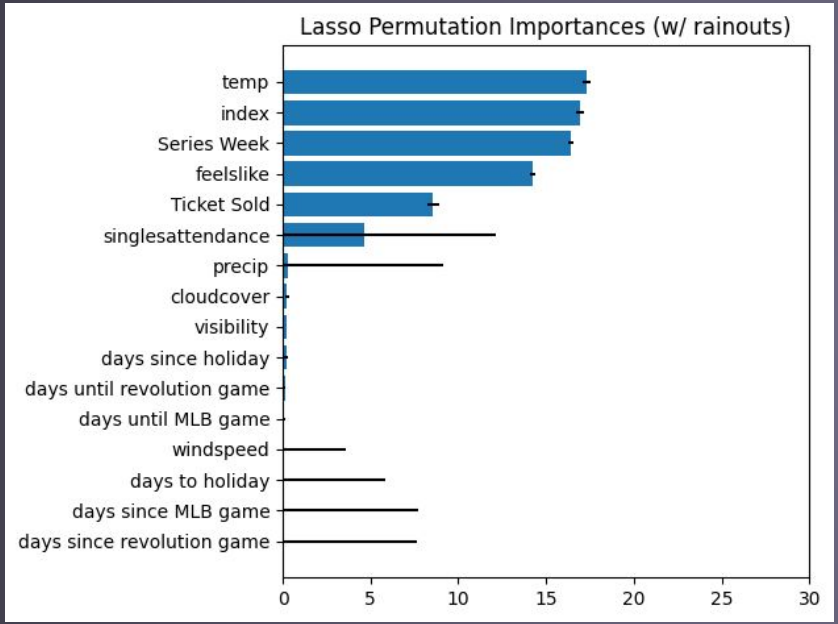
- Baseline RMSE: **899 people**
  - Problem: not parsimonious

Model: Lasso linear regression

- Regularization method for linear regression
  - Assumes only a few features needed for forecasting ⇒ parsimonious
    - Determined by regularization constant

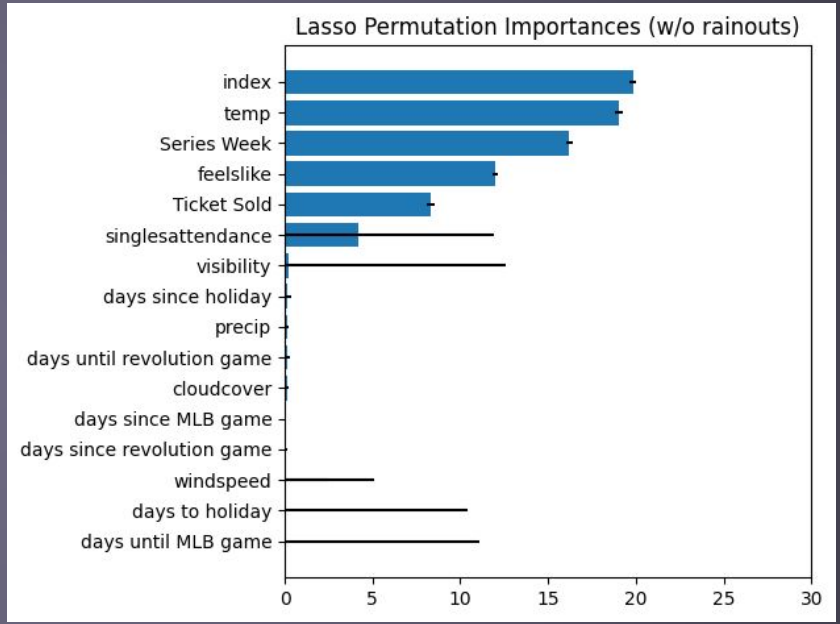
- RMSE with rainouts (constant at 10): ~**643 people**
- RMSE sans rainouts (constant at 10): ~**881 people**

Permutation importances
w/ rainouts



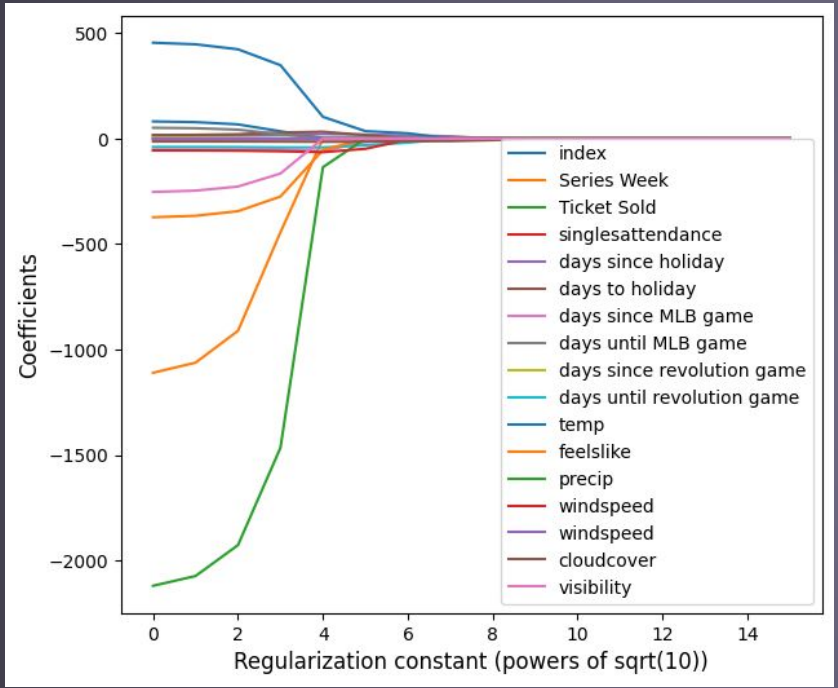
The important features:
['index', 'Series Week', 'Ticket Sold', 'temp', 'feelslike']

w/o rainouts

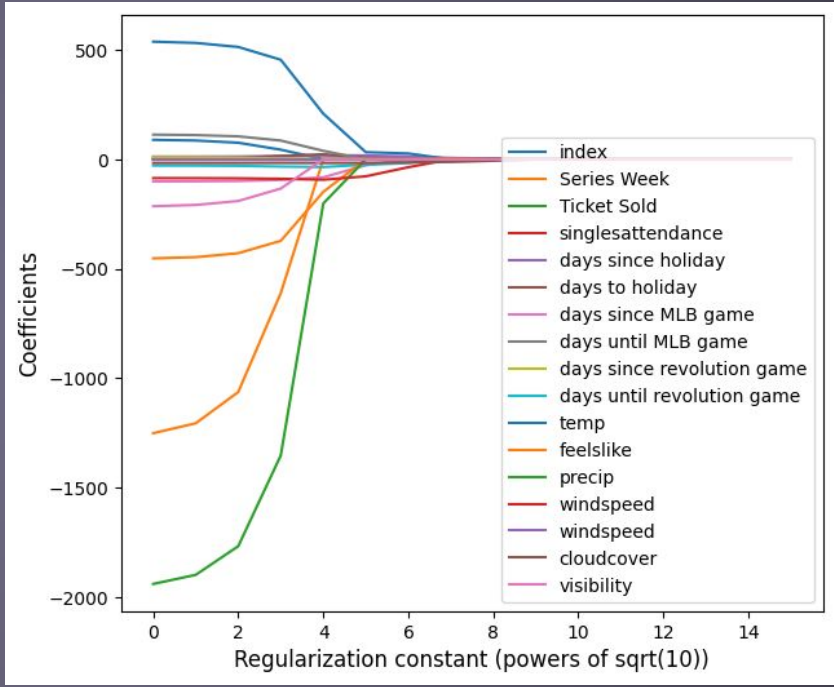


Lasso coefficients v. regularization

w/ rainouts

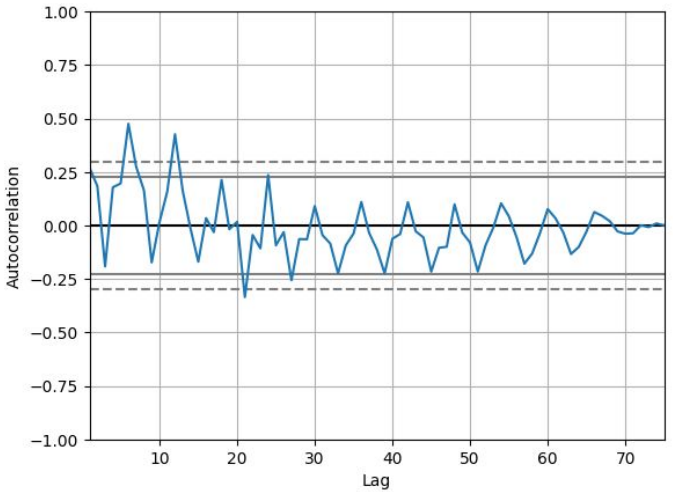


w/o rainouts



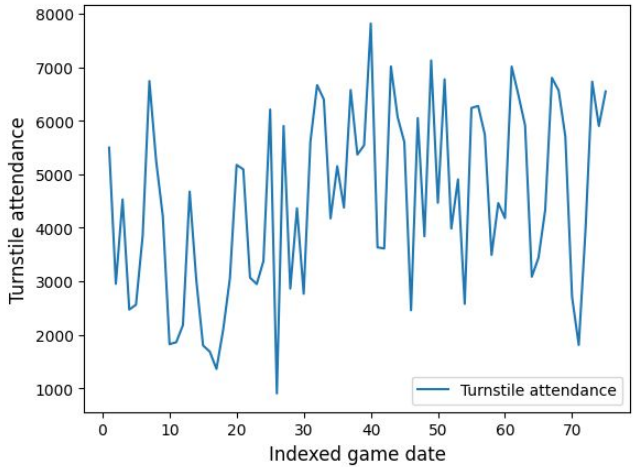
Time series

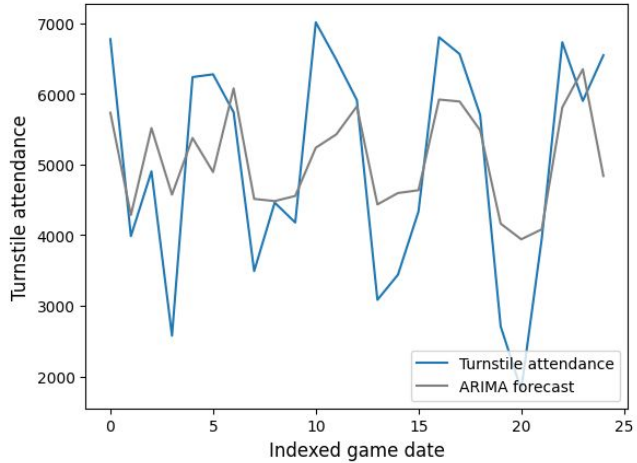
- Motivation
  - Embedding information about date in a different way
  - Seasonal features were very important
  - Useful for forecasting
- Some time series to use
  - Time series with mean attendance imputed in non-gamedays
  - Time series of just gamedays



Example model: rolling-forecast ARIMA

- Model: ARIMA
  - Time series forecasting method
  - Combines seasonal prediction and moving average (MA), with differencing
- Rolling-forecast
  - ARIMA retrained as new data becomes available - simple way of forecasting series





Some sample code:

Building our dataset of Lasso coefficients w.r.t. our regularization constant

```
k = 16

cdf = pd.DataFrame(index=X.columns, columns=range(k))
for i in range(k):
  lasso = Lasso(alpha=10**(i/2), max_iter = 10000)
  lasso.fit(X_train, y_train)
  cdf.iloc[:,[i]] = lasso.coef_.reshape((17, 1))
print(cdf)
```

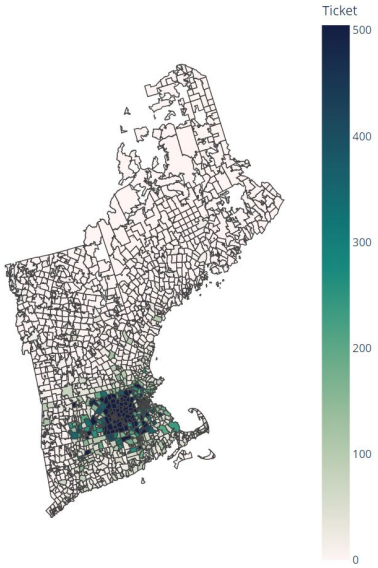
0	0	1	2	\
index	89.114677	85.936567	75.893247	
Series Week	-1251.15527	-1205.914545	-1062.947271	
Ticket Sold	1.219826	1.220599	1.223038	
singlesattendance	-0.990832	-0.989585	-0.985641	
days since holiday	12.118659	12.316716	12.942745	
days to holiday	1.84228	2.789786	5.783265	
days since MLB game	-100.984047	-100.500422	-98.972307	
days until MLB game	112.87445	110.968665	104.943257	
days since revolution game	10.052419	9.574524	8.064494	
days until revolution game	-28.47913	-28.789717	-29.770953	
temp	537.718032	531.890973	513.451862	
feelslike	-452.454457	-446.805432	-428.931947	
precip	-1940.005764	-1898.537056	-1767.401722	
windspeed	-85.819879	-86.033827	-86.710011	
windspeed	-0.0	-0.0	-0.0	
cloudcover	-16.568261	-16.524289	-16.38527	
visibility	-214.097642	-208.392146	-190.348645	

Building our example walk-forward ARIMA

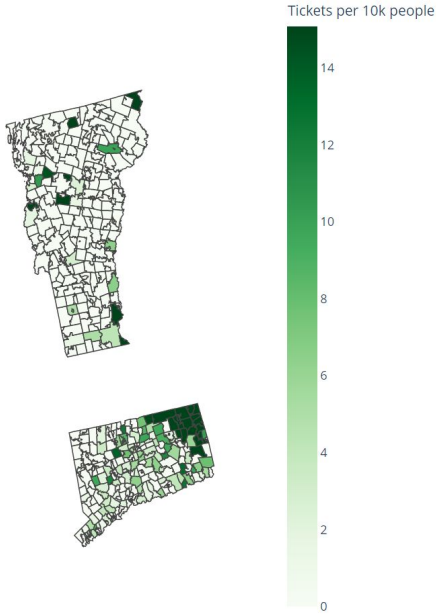
```
for t in range(len(val)):
  arima = ARIMA(series, order=(6,1,0))
  arima_fit = arima.fit()
  output = arima_fit.forecast()

  yhat = output[0]
  pred.append(yhat)
  series.append(val[t])
```

(3) Geographic Visualization & R Shiny

Progress on Cloropleths from Midpoint

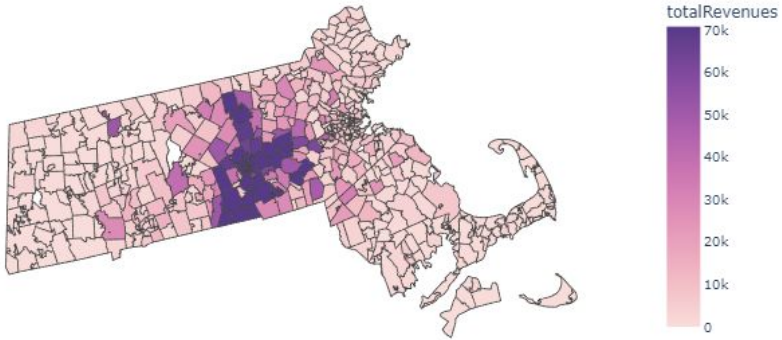


Regionalization

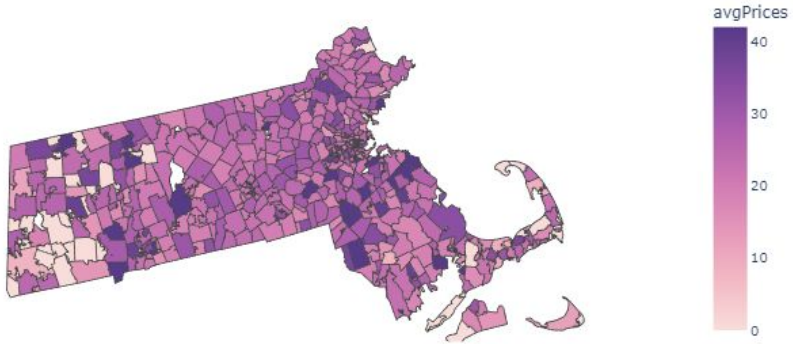


Seeing socioeconomic indicators

Price sales analysis by geographical location

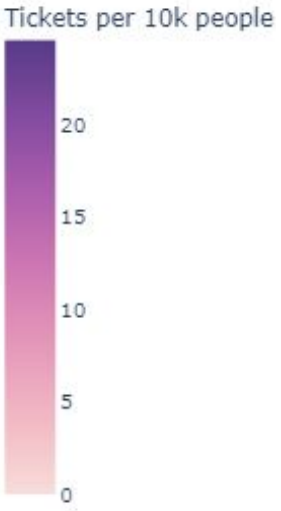


Total Revenue by Zipcode

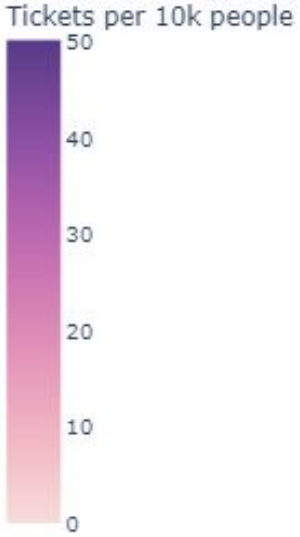


Average Ticket Price by Zipcode

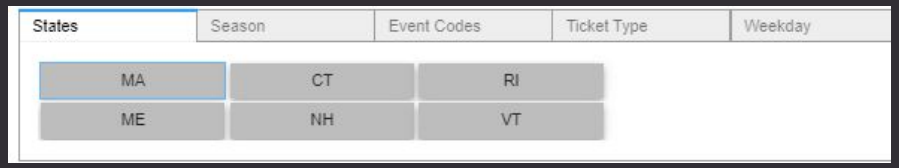
Vermont



New Hampshire

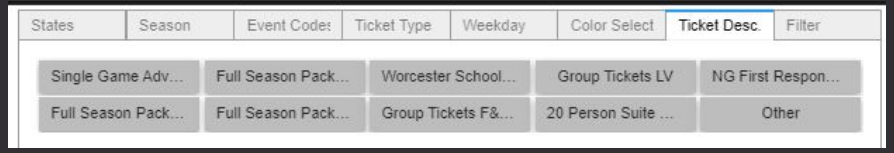


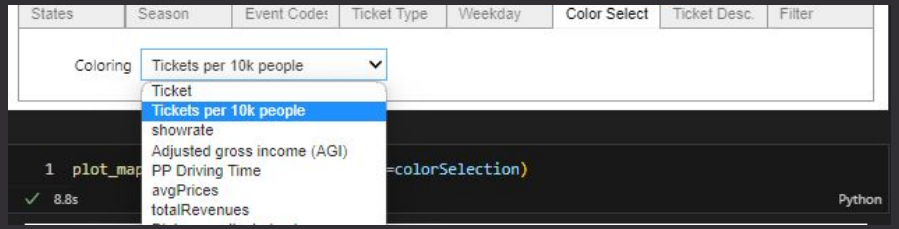
Easy Data Filtration



State Selection

Ticket Type Selection





Map Coloration Selection

Interactive Ticket Sales Map of New England

Forthcoming work.

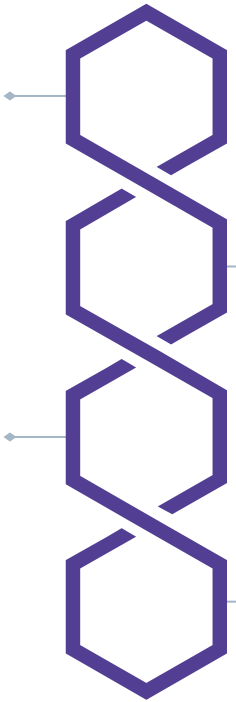
Future work and direction

Time-series improvements

Multivariate analysis
Ex. structural equation modeling

More data fine-tuning

Drink and beverage purchases, ticket-price adjustment to region and demographics.

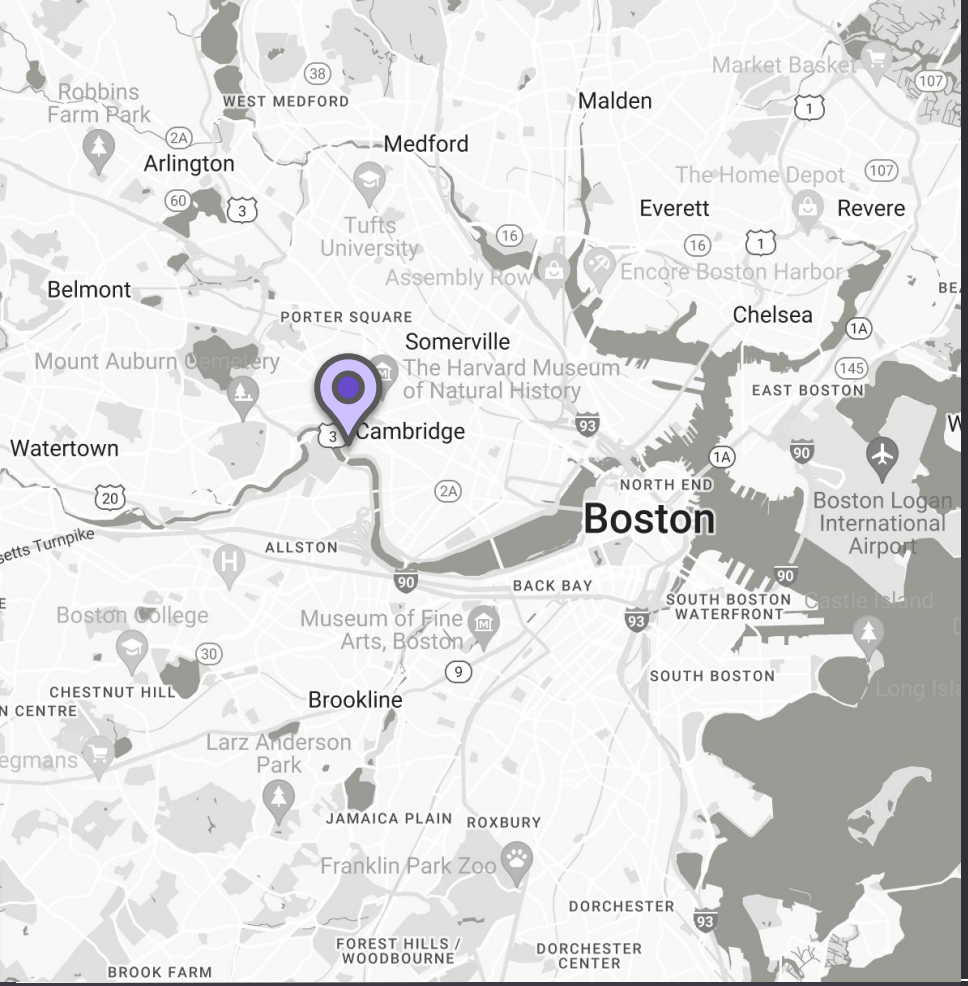


Geographic Visualization App

Progress towards an interactive web-app for real-time ticket-sale visualization

Interseasonal analysis

Looking at trends from Pawtucket to Worcester, improve year-by-year ticket outcomes with more data leverage



Contact Us

332 Eliot Mail Ctr
Cambridge MA 02138 USA

WEBSITE
harvardanalytics.org

EMAIL
partnership@harvardanalytics.org

dries_rooryck@college.harvard.edu

INSTAGRAM
@harvardanalytics

Thank you

We welcome any and all questions or feedback at this time.