

CS 171

MBTA Sporting Events: Process Book

Lyla Fadden & Micah Lanier

GitHub: github.com/micahlanier/cs171-gameday

Web: micahlanier.github.io/cs171-gameday

Overview & Motivation

As one of the largest public transportation agencies in the United States, the Massachusetts Bay Transit Authority (MBTA) serves the needs of the greater Boston area and facilitates an average of 1.3 million trips (by bus and rail) every weekday. Delays and overcrowding are widely felt, and are often exacerbated by spikes in demand brought on by major public events.

Our project aims to visualize public transportation demand before and after major sporting events in Boston. The relationship between ridership and events—and the variability thereof—is difficult to quantify effectively. Observations depend on a variety of related factors: the teams and sport involved, venues' public transit connections, game length, the stakes of the contest (e.g., playoff games), and exogenous factors (e.g., weather, holidays) all hold apparent influence over ridership. For these reasons, not only is visualization an effective narrative medium for explaining their connections, but it is also extremely useful for *discovering* connections that are not apparent in the numbers. We believe that ours is an effective solution for both facets of this challenge.

Related Work

Our project stems from work that we have taken on for [AC 297r](#), a capstone class for students in Computational Science & Engineering. We are working with two other students (both of whom are using MBTA data for a separate project in CS 171) to develop a predictive demand model for MBTA managers to apply to their day-to-day work (see our [project website](#) for more). We found in our work that MBTA ridership is typically predictable, but can vary widely due to events like weather and sporting events. This particular visualization project grew out of that observation and the difficulties we've encountered when attempting to engineer useful summaries of sports schedules to include in our machine learning models.

With the increasing availability of public APIs and open datasets that pertain to public services, many others have applied their analytical faculties and visualization skills to topics like our own.

[Visualizing MBTA Data](#)—developed by two WPI students and published in 2014—is a superbly-executed D3-based examination of MBTA train traffic and ridership statistics. They include observations about several basketball games in their analysis and note an apparent connection between station congestion and train delays. Our analysis primarily focuses on the “demand side” of this phenomenon (i.e., we will not make use of any train arrival data); but their evidence demonstrates the importance of understanding demand when attempting to provide optimal service.

Outside of Boston, there is no shortage of public transit visualizations that rely on D3. These include [multiple ways](#) of displaying New York subway data. Though few if any of these visualizations specifically address questions similar to our own. As such, we hope that ours will provide a novel perspective both for Boston and for public transportation services generally.

Questions

The core question that underlies our analysis is simple and intuitive: what ridership patterns distinguish gameday (and gametime) demand for public transit compared to days without games? As suggested above, we have undertaken this project both to highlight relationships that we *know* to exist between ridership and sporting events, and to provide a means for discovering dynamics that are not yet apparent to us.

Several variables describe ridership pattern changes that occur in response to sporting events. Different concentrations of ridership lift (or perhaps decline in isolated cases) might occur at different times, perhaps in response to a boring game or an exciting one that runs far longer than anticipated. Some fans might be apt to leave early and catch the next train, while others might stay in the neighborhood for several hours after each game. Predicting all of these variables is an important logistical task for MBTA planners concerned with saving money while still providing optimal service at peak times.

At a higher level, we are interested in what features of games (day of week, time of day, specific opponents, or others) might predict certain ridership distributions. On a weeknight, fans might rush home immediately after games; on Saturday evenings we might not expect them to be in such a rush. Important games—say, playoffs and rivalries—might also influence fans to arrive early and stay to the very end of games, while low-stakes contests may see late arrivals and early exits.

As suggested, our visualization aims to answer the above questions both by emphasizing the trends that we have observed and by providing an exploratory method that users may employ to discover novel relationships.

Data

Our project ultimately aligns two primary groups of datasets: home game schedules for Boston's major sports teams, and MBTA ridership statistics.

Sports Data

Our sporting data features home games and associated information for three major sports teams that play their home games in Boston: the Bruins (NHL), Celtics (NBA), and Red Sox (MLB). Records go back to the beginning of 2013 (the earliest date for which we have MBTA ridership data).

To obtain these schedules, we found it easiest to manually download each team's schedules from espn.com. Those schedules are readily available in a fairly clean HTML format (see the [2013 Red Sox schedule](#) for an example) that we were able to process using a text editor and several regular expressions. For all sports, we were able to easily extract dates, times, home games, and opponents; we manually added identifiers for playoff games. While not a particularly technically-sophisticated process, this proved fast and effective.

The following is an example of several Boston Bruins games from 2013:

<u>Datetime</u>	<u>Playoff</u>	<u>Opponent</u>
2013-01-19 19:00	0	NY Rangers
2013-01-21 13:00	0	Winnipeg
2013-01-25 19:00	0	NY Islanders
2013-01-29 19:00	0	New Jersey
2013-01-31 19:00	0	Buffalo
2013-02-12 19:30	0	NY Rangers
2013-02-28 19:00	0	Ottawa
2013-03-02 13:00	0	Tampa Bay
2013-03-03 19:30	0	Montreal

We performed some additional “wrangling” of the data that is ultimately served to visitors, but the final data format largely follows the above.

MBTA Ridership Data

As part of our aforementioned capstone project, we received ridership data since the start of 2013 for MBTA trains (subways and Green Line trains, not commuter rail) and busses. Each dataset shows aggregate “entries” (bus data breaks down those who use Charlie Cards and those who do not) at fifteen-minute intervals, with varying amounts of geographic detail (discussed below).

Our underground train dataset covers 320 million trips (“station entries” to be precise) over the last two years. This is an example of several rows from April of 2014:

<u>locationid</u>	<u>servicedate</u>	<u>servicetime</u>	<u>entries</u>	<u>exits</u>
1002	2014-04-01	500	5	1
1002	2014-04-01	515	81	8
1002	2014-04-01	530	60	15
1002	2014-04-01	545	68	13
1002	2014-04-01	600	90	15
1002	2014-04-01	615	135	23
1002	2014-04-01	630	121	17
1002	2014-04-01	645	90	26
1002	2014-04-01	700	117	11

The “locationid” column corresponds to particular locations, which we can look up (along with associated line colors) in an associated metadata CSV file. In terms of data volume, we have almost 4 million records in 110 MB of data.

Bus and above-ground Green line data constitute 11.1 million records, clocking in at 286 MB of data. The file structure resembles the above, omitting exits.

Processing

Obviously, our raw dataset is far too voluminous and “messy” to serve to users’ browsers. To that end, we have preprocessed data using Python, and saved nine summarized files of game data, each available in the “data” directory of our Github repository.

Team schedules and MBTA data have been modified to remove certain holidays and service outages, as those days show deep dropoffs in ridership that we can safely intuit to *not* be associated with sports games themselves. Such days with no games also skew expectations of train traffic on gameless days, so removing them will improve those estimates as well.

When the data is processed, we join team schedules ridership data aggregated at several levels—at various times by station, line, date, day of week, and time of day. From there, we generate a comparison dataset that aggregates ridership by time of day and day of week for days

that did not feature any sporting events. We restrict the time range of this comparative aggregation to dates during the season so that we do not inadvertently pick up seasonal effects in our comparison (say, by comparing Red Sox gameday ridership to ridership in January). The difference between ridership on each day and comparison ridership for the same day of week and time of day are used in the visualization.

Exploratory Data Analysis

Naturally, we undertook exploratory analysis to determine whether or not there actually exist observable and significant trends in ridership demand before and after sporting events. Detecting those trends requires consideration of other variables, such as weekly ridership cyclicalities and the extent to which holidays influence ridership (and potentially affect comparisons between gamedays and gameless days).

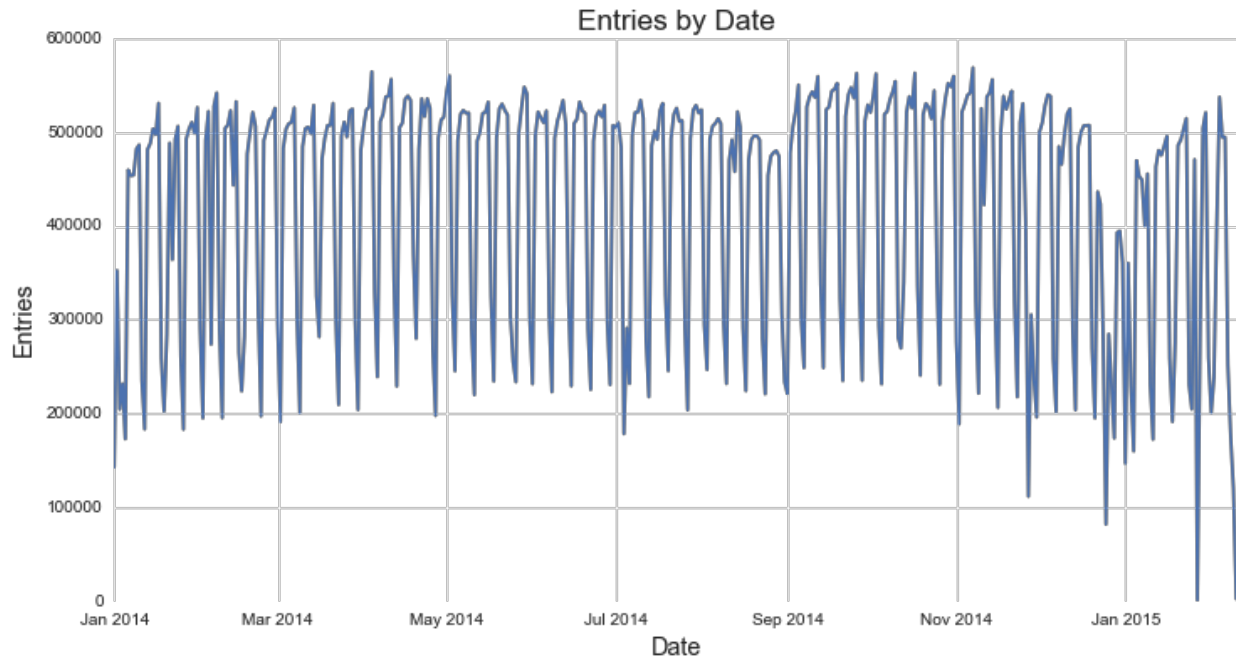
The particular distribution of sporting events will also guide our analysis. Naturally, league executives and team management schedule games for profitability, both in terms of attracting local crowds and drawing television viewers. This manifests in particular game scheduling trends, like the preponderance of prime time games on weeknights (as opposed to poorly-attended afternoon games) and differing league traditions of playing on or around major holidays (e.g., the days after Thanksgiving or Christmas).

Finally, this section will also illustrate the high-level ridership trends that a user might analyze with our visualization. This section will focus on exploratory analysis; later sections will use observations from our visualization itself.

General Ridership Trends

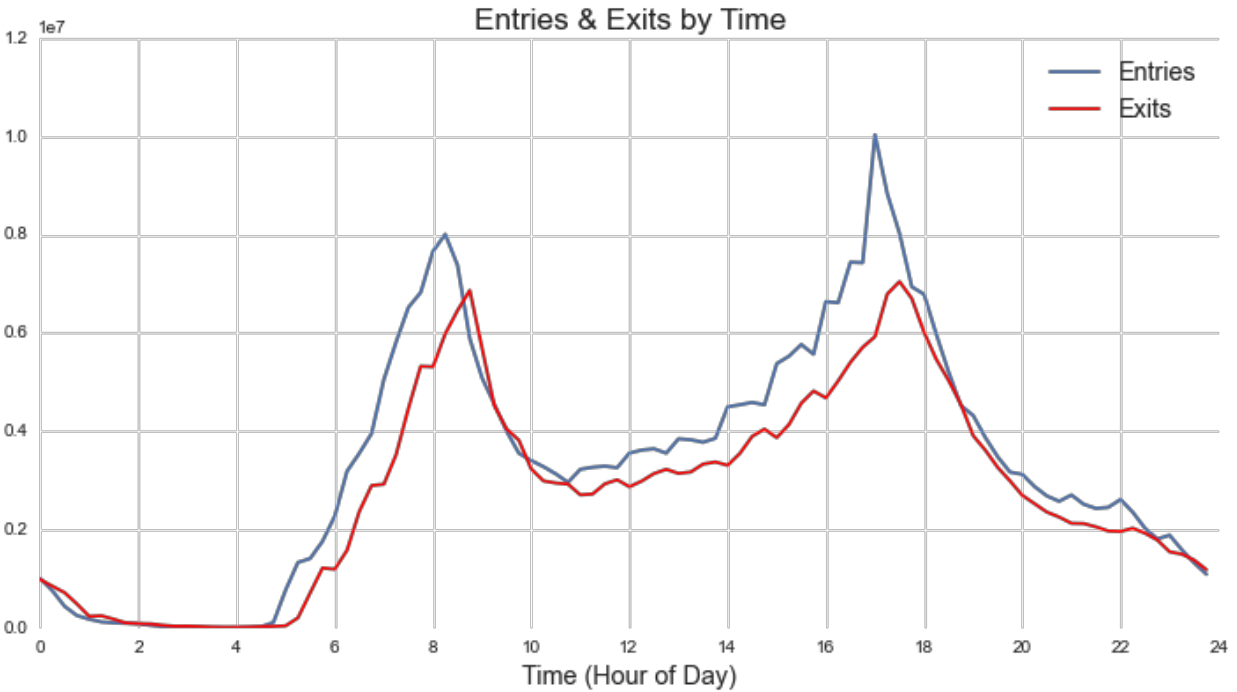
At a high level, it is worthwhile to examine aggregate ridership statistics. Doing so may reveal broad patterns applicable to more targeted analyses of traffic around sports games.

For starters, 2014-15 train data shows an obvious weekly cycle:



The regular dips occur around weekends, suggesting that we must account for day of week when comparing gametime traffic to general days. There are also apparent seasonal effects, such as students' arrivals in early September and dips in ridership around Thanksgiving and late December. The fact that TD Garden (where the Bruins and Celtics both play) is located near a major downtown station (North Station) suggests that we ought to take care when including games around holidays in our analysis.

We can also observe regular commuting patterns in our dataset:

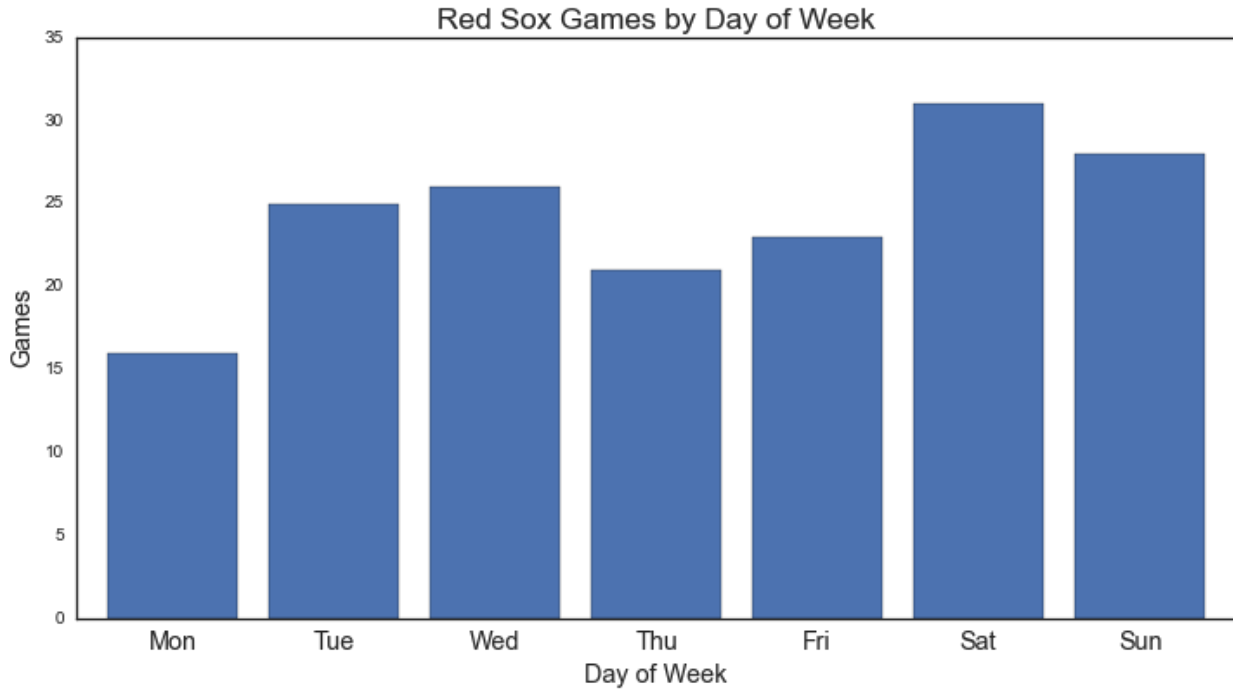


Many games occur during the “downward slope” of afternoon rush hour. Intuition suggests that many people—not just game-goers—might alter their behavior in response to a game, perhaps commuting at different times to avoid crowds or using alternative routes.

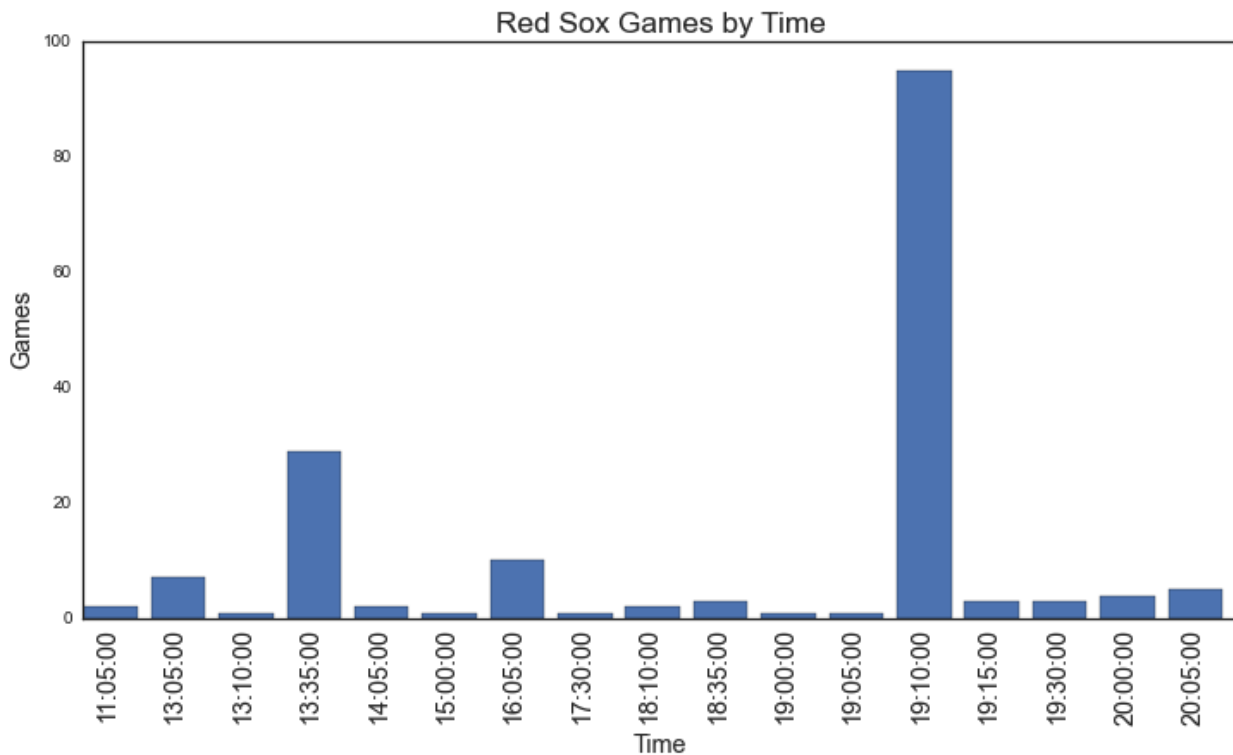
Sporting Event Characteristics

Sporting events are best analyzed in two groups: Red Sox games by themselves and Bruins/Celtics games together.

Red Sox games take place consistently throughout the week, suggesting that we have enough data to make useful conclusions about particular weekdays:

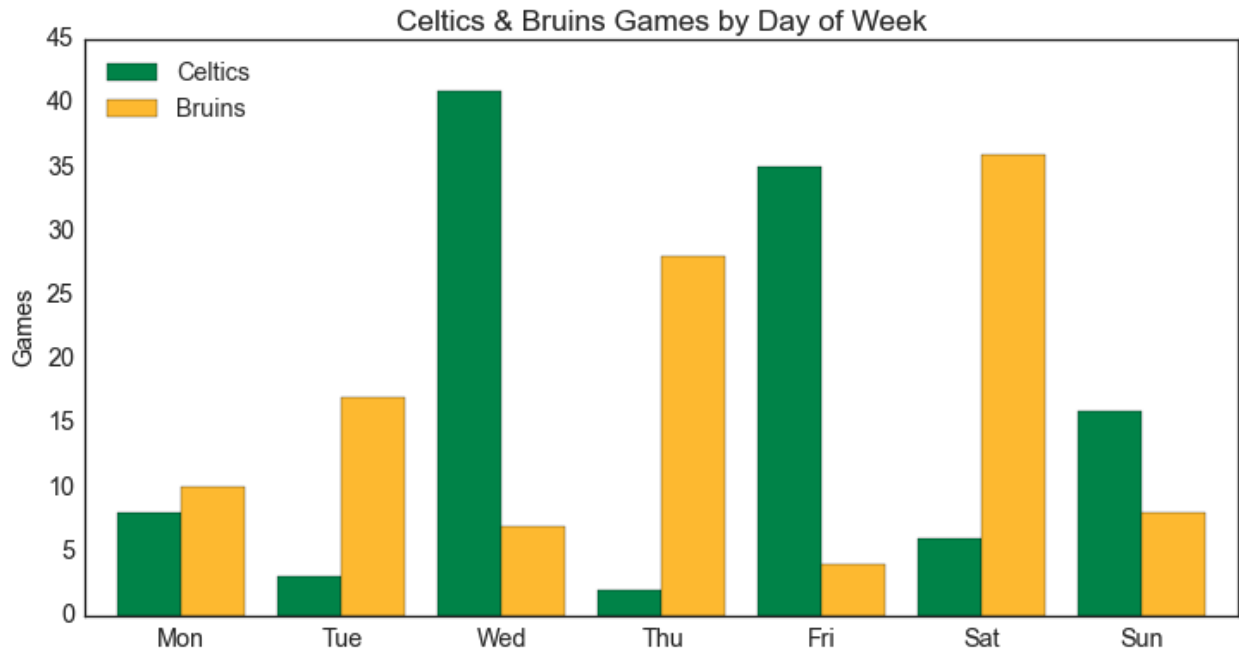


Red Sox games show an interesting distribution of times as well:



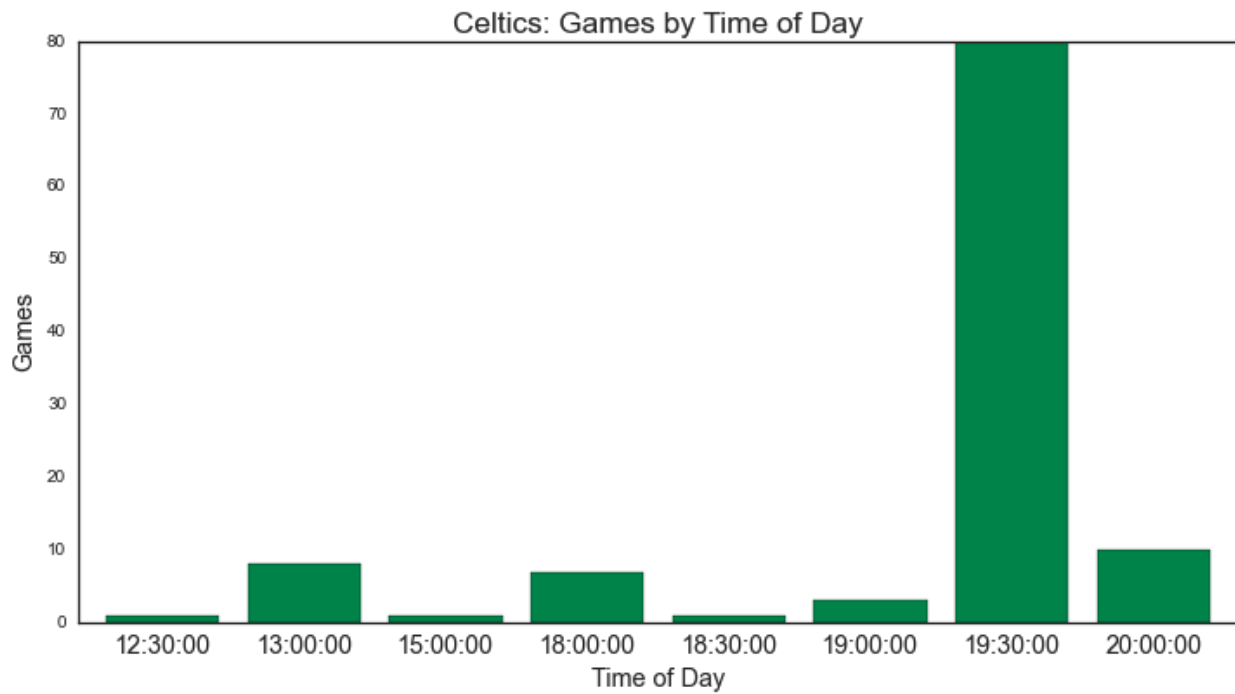
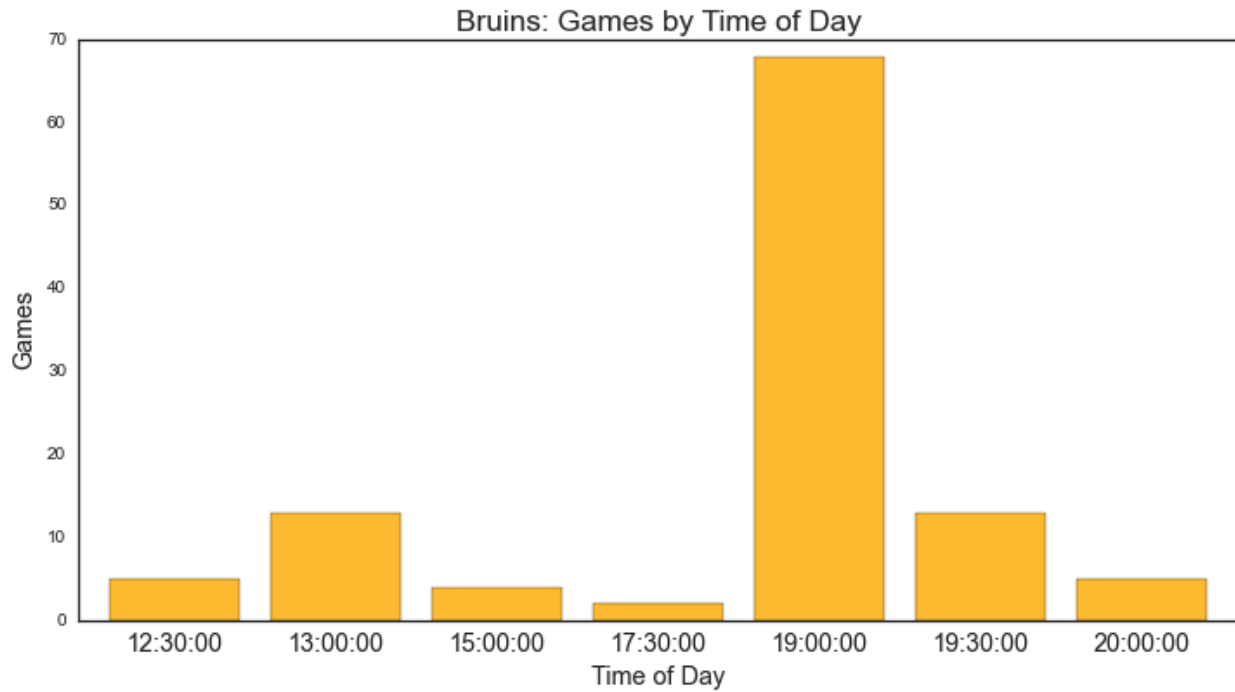
Though there is a distinct preponderance of prime time games, there is also a wide distribution of daytime games as well. Observations suggest that those tend to be held on weekends. This distinction offers a useful opportunity to distinguish among both gametimes in our analysis.

Because the Bruins and Celtics share TD Garden and their seasons take place at essentially the same time, their home game schedules show distinct patterns:



Tuesdays/Thursdays/Saturdays are distinctly Bruins days, while Wednesday and Friday show the bulk of Celtics games. This might affect the specific groups of games to which we draw the user's attention, and suggests that we will have to take care to ensure we only compare game traffic to those days where *neither* team held a game at TD Garden.

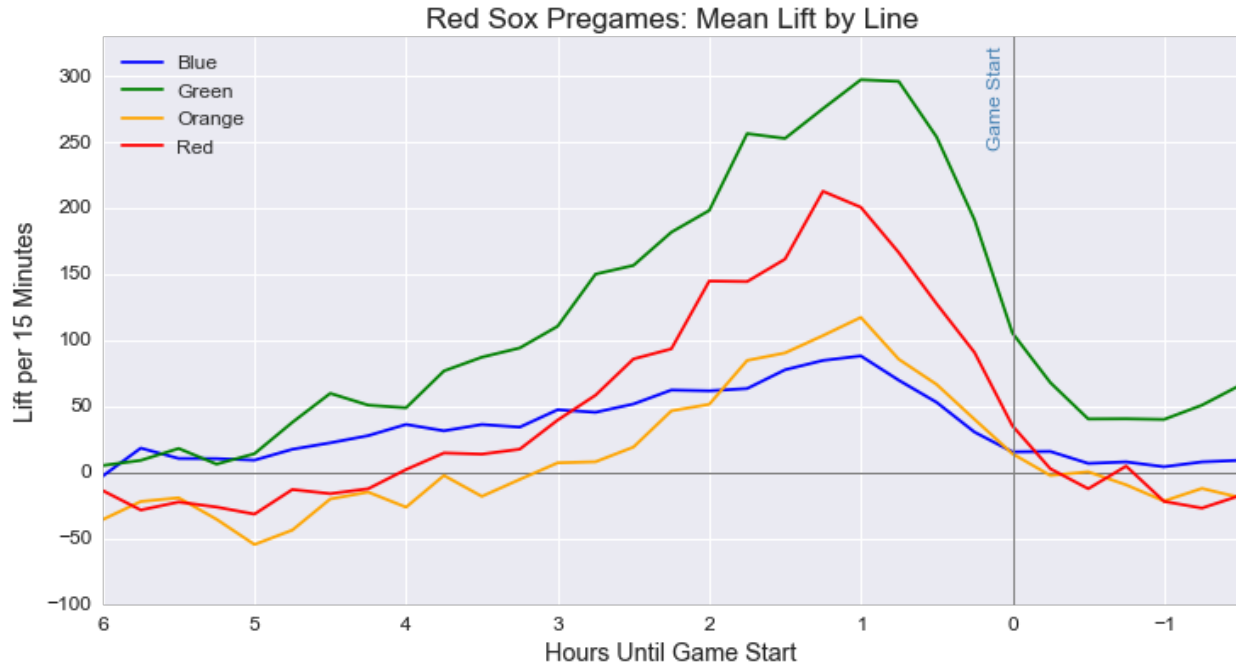
Time-of-day trends suggest a distinct preference for prime time games among both teams:



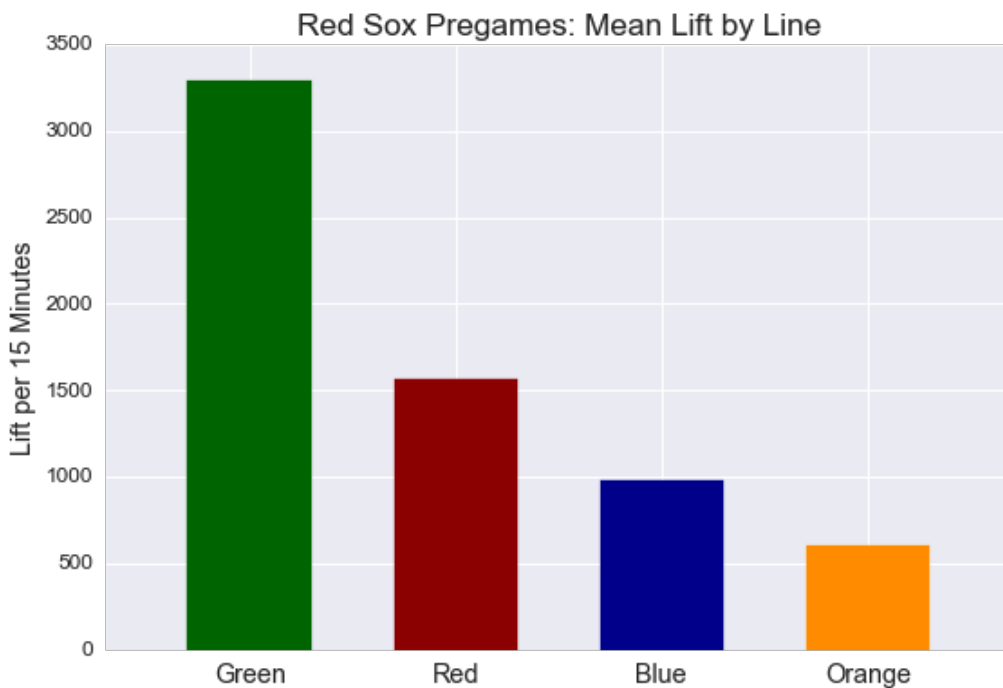
There may not be much in the way of an interesting pattern of ridership traffic by time of day, but early-day games may still be common enough that their traffic is worth understanding.

Ridership Before Sporting Events

The Boston Red Sox show ridership increases quite a while before games actually start:



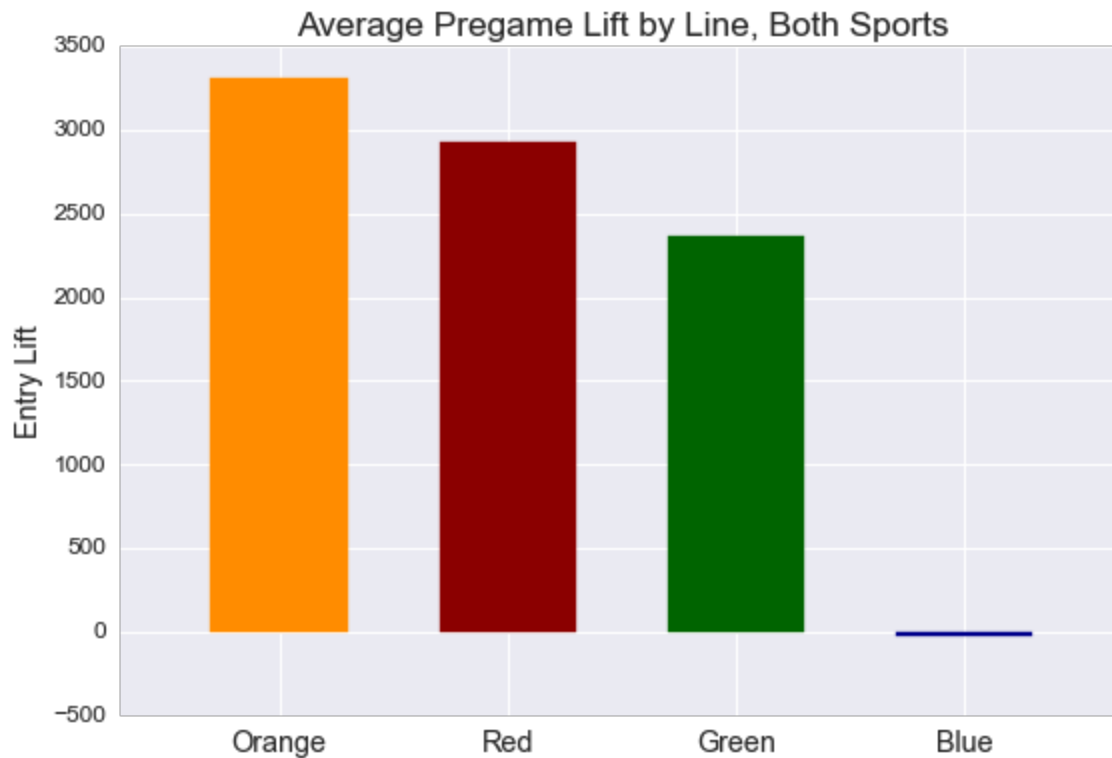
The following chart confirms the evident popularity of the Green Line:



The above chart may even understate Green Line ridership, as those who enter stations at the other lines likely transfer to the Green somewhere along the line (assuming all of these people are indeed going to the game).

Pregame lift patterns for Bruins and Celtics games (combined here for ease of analysis) also begin quite a while before games themselves:

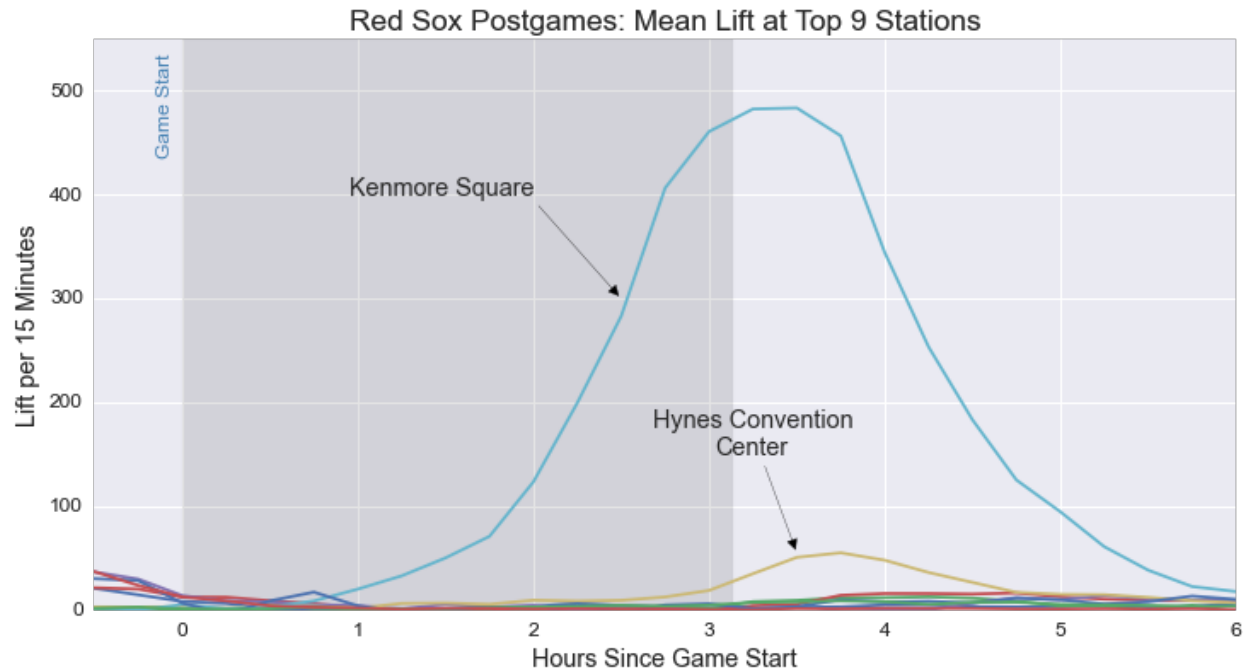
The Orange and Green lines are predictably popular (both visit North Station), but the Blue attracts few (if any) additional passengers for a typical game:



Interestingly, despite playing in a larger venue, the Red Sox appear associated with less pre-game lift than the Bruins/Celtics. This may be a function of fanbase that simply *changes* its ridership behavior rather than exclusively using MBTA services on game days; it may also result from non-game-goers being more apt to avoid crowded public transit during games.

Ridership After Sporting Events

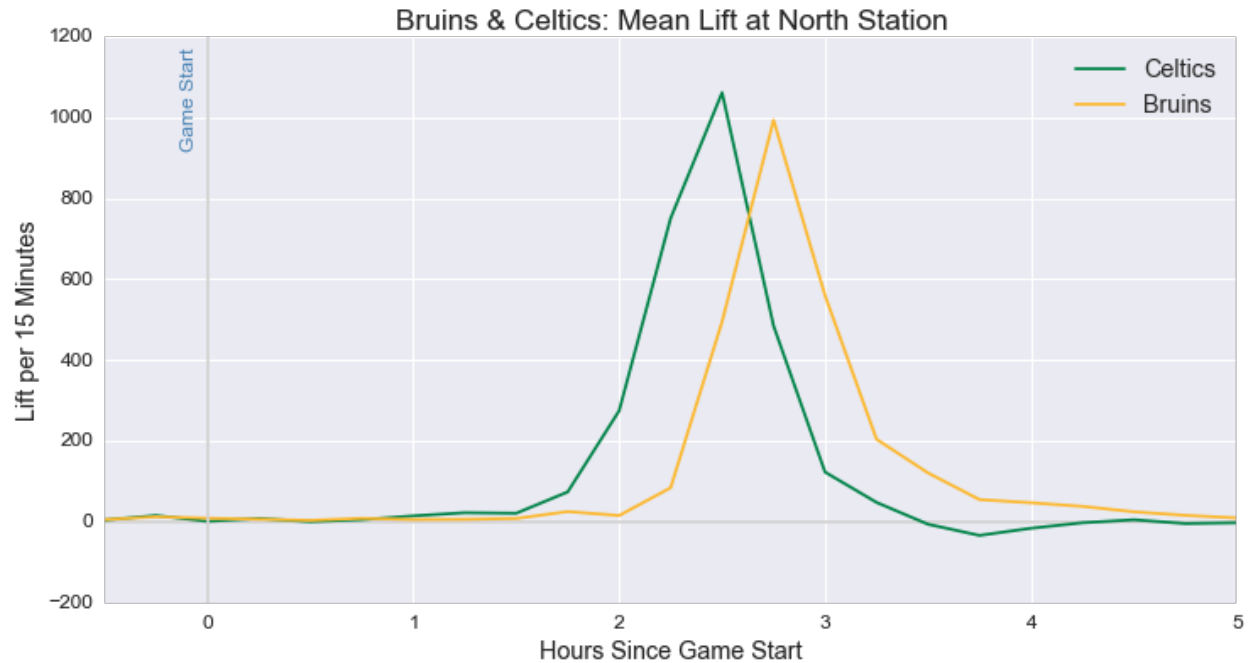
Postgame ridership increases are more moderate for both sports. Red Sox ridership peaks at two particular nearby stations:



This defied our intuition that other stations in the vicinity of Fenway Park would show a ridership boost.¹ In all, the “typical” ridership distribution is fairly spread-out. This is likely both a function of people leaving games early and the inherent time variability of baseball games.

In contrast to the Red Sox’ postgame ridership pattern, the Celtics and Bruins show distinct spikes immediately following their games:

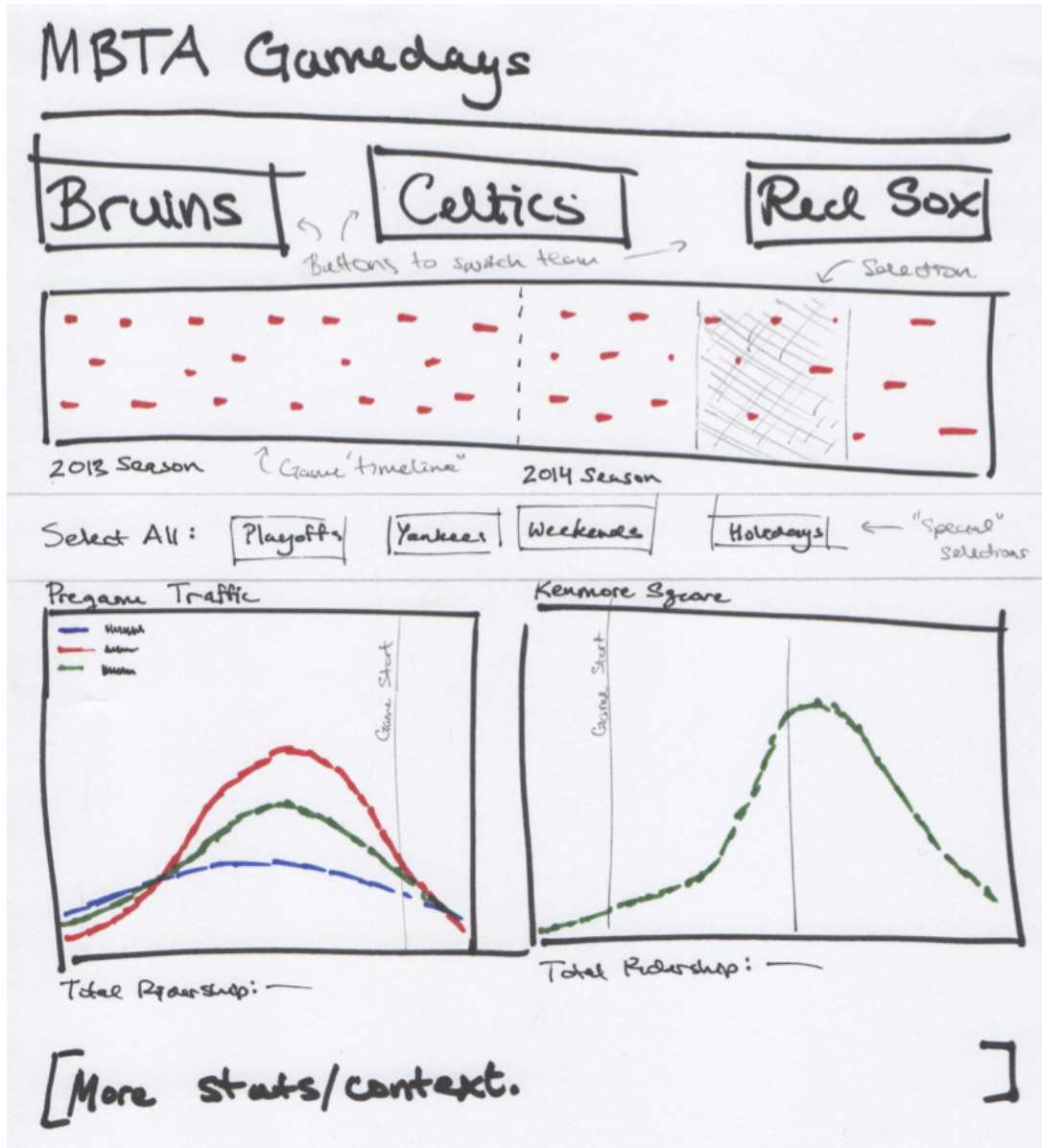
¹ Note that we have no way to detect the bulk of Green Line entrances at the Fenway Park station itself. The MBTA uses a different boarding process for which we have no data. All we can say is that Fenway is a popular postgame destination for Red Sox fans, but that it is not more popular than Kenmore.



Note that hockey games supposedly tend to be slightly longer. It is difficult to attribute this to a particular cause, but we suspect that it occurs due to more predictable game lengths. Bruins and Celtics games also have problems measuring ridership. Particularly busy times often see a relaxation on ticket requirements—we cannot observe ridership reliably on these days. For that reason, analysis of the “shape” of the ridership distribution might be more useful than raw numbers themselves.

Design Evolution

Our visual design largely followed the original paper mockup that we designed as part of our project proposal. That is reproduced below for reference:



This design intended to allow users to select teams, brush over their games (divided by season) or select specific non-contiguous game groups (e.g., weekend games or games against a particular rival), and see distributions of pre-game traffic across the entire system and post-game traffic at stations around each sporting venue. We ultimately chose to add additional charts where indicated in the “more stats/context” area.

Given the geographic components of our dataset, we were interested in ways that we might display traffic across the entire MBTA system, rather than just at particular, pre-defined points in the network. We discussed this idea with another group in our design studio session and confirmed our skepticism about the value that this would add to the visualization. We revisited this option over the course of our development process but never decided to pursue it.

Milestone 1

For the first milestone, we implemented the brushable timeline, an extensible system for defining the “dynamic” filter buttons that sit below it, and placeholders for the two major charts that received input but did not render charts.

These components functioned as we originally intended them to, with the exception of the team-specific filter buttons. We considered a set of static HTML buttons that could be easily shown/hidden when the user changes sports teams. Rather than implement this design, we opted for a D3-based solution wherein we define filter buttons dynamically, add them to the DOM, and bind special filtering functions to them. This implementation is described in detail below.

Rounding Out the Original Design

We implemented pregame and postgame traffic designs much as envisioned in our proposal. The time axes in each mockup proved useful in the final visualization, and we were able to settle on our final design for each chart with minimal iteration.

The only feature that we opted to change is the indicator for game start and end in the postgame visualization. We originally wanted to include traffic from the start of each game after observing that—for the Red Sox in particular—traffic often ramps up before the typical game is over. Our initial design called for simple lines marking the start and end of a typical game for the selected sport. During the development process, we determined that adding a shaded region in between both lines would be a more effective means of indicating gamespan.

Once again, a complete examination of each implemented component appears below.

Further Changes

Having developed the components described, we began to think about how best to round out our project and present the most useful information about T ridership. This prompted two observations about the visualization components that we had already implemented:

1. Pregame line-by-line traffic varies significantly across games, and there is no easy way to evaluate overall differences between lines. This is particularly true when pregame traffic lift along a line dips below 0—i.e., those games show *less* traffic than expected for equivalent gameless days. To calculate aggregate differences by line, the user would have to mentally evaluate the area between the curve and the “0” line, which is impossible to do with precision (in this problem, at least).
2. The “shape” of postgame traffic was remarkably stable across games for each team. The interesting difference between games was thus not the shape of the distribution, but rather the *scale*. Because our visualization components change lift scales to match the current

selection—a feature that we considered relaxing but ultimately decided was useful—users could not easily understand the distribution of postgame traffic volume.

In response to the first concern, we found a simple solution in our exploratory analysis above: a bar chart showing aggregate ridership lift by line for the selected games. We determined that an effective visualization would have several features:

- The scale should change dynamically depending on the selection.
- It should clearly show lines that show *decreases* in aggregate ridership with “negative” bars.
- The bars ought to change order based on the lift for each line.
- We ought to emphasize precision, ensuring that users can easily understand absolute lift for each line.

As the final design shows, we indeed implemented this design. Details of its implementation are discussed below. While implementing it, we determined that there was value in ensuring that each page element also featured animated transitions, as the effect of bars suddenly switching places was jarring. This led to our decision to implement other transitions where appropriate.

In response to the second concern above, we noted that games show a wide aggregate distribution of postgame ridership. To demonstrate how this distribution changes with selections, we also implemented a histogram of aggregate postgame ridership for each selected game. This and other visualization components are discussed below.

Implementation

Technical Background

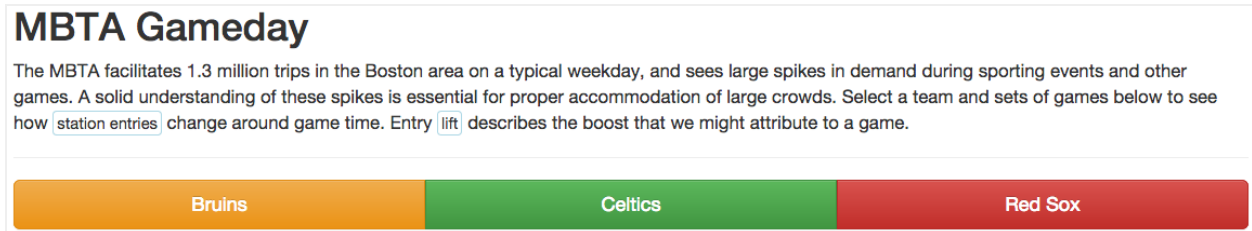
Our data processing and cleaning was performed primarily with Python. Specifically, we used [Pandas](#) to whittle down our dataset and arrange it into a format useful for visualization, and augmented that with [Matplotlib/Seaborn](#) to perform exploratory analysis and verify that our data will tell a worthwhile story.

Much like the rest of our work product in this course, our visualization is built using [D3](#). Our visualizations are primarily “pure” D3, though we utilized Mike Bostock’s [queue](#) library to load our data. We also use [jQuery](#) in several contexts: event handling, tool tips, non-charting DOM updates, and others. Finally, we developed much of our page design using [Bootstrap](#), and drew heavily from its set of stock elements when designing non-visualization page elements.

As stated above, we rely on jQuery bindings and triggers for event orchestration. The various components of our visualization are designed to be modular and are stored as separate Javascript files to ease collaborative development.

Page Load & Team Selection

When loading our project, the user is presented with a simple explanation of our work and given the option of selecting a particular team for visualization. We also include several tooltips to provide a slightly expanded introduction for unfamiliar users.



This design aims to introduce the topic and make it clear that toggling among different teams will affect the data displayed below. If we had selected a “starting” team by default, we risked presenting the user with a lot of information and no intuitive indication that they could change teams. This design establishes a clear hierarchy of interaction that makes team switching intuitive.

Selecting a team for the first time reveals the rest of our visualization. If a team had been selected in the past, all filters (i.e., brushes or other selections) are removed and the visual elements are updated with data for the team selected.

Technical Details

This component uses Bootstrap’s button designs to create buttons that match the colors of each team. When a user clicks one, it calls our primary D3 event handler with the selected team. The event handler then informs the various interested components of the change. In the case of the first team selection, a click also “unhides” all other page components.

Game Brush

The game selection “bar” changes appearance with each team selected. It is shown here for the Boston Celtics:



Starting at the top of the component: the title changes with each team selection, giving a straightforward indication of the selected team. Further instructions clarify the potential

interactions and significance of the vertical axis. Selecting games adds a clear brush region and indicates the games that are selected and those that are not:



Technical Details

The game selector uses a standard `d3.svg.brush` object for user interaction. Before rendering, it employs a system of D3 scales to map game dates to the horizontal axis, with appropriate divisions of and padding between seasons. This system dynamically adapts to each particular team and the data available for it, allow us to (for example) vary the number of seasons between the Bruins/Celtics (each with three seasons in the data) and the Red Sox (with two).

Additional Game Filters

In addition to brushing, the user can manipulate game selections using our special filter buttons (shown below for the Bruins):



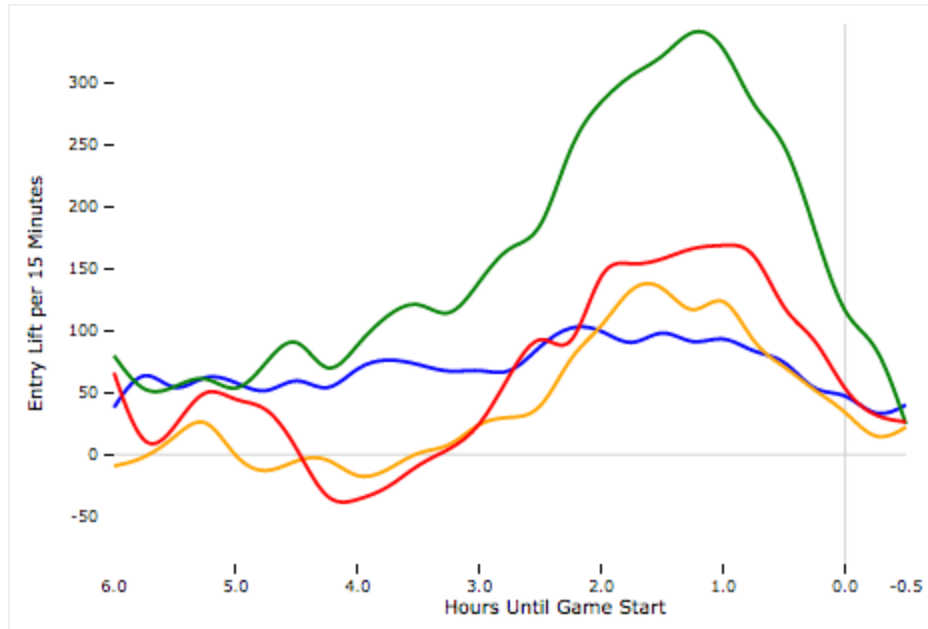
The selection of options updates with each team selection. The three teams share several button sets (the first five buttons shown above) and each has at least one “special” button for interesting games; we intended this to allow users to ask intuitive questions (e.g., “how does traffic vary when the Red Sox play the Yankees?”) that cannot be easily specified with brushing behavior.

Technical Details

Rather than defining these buttons statically, we defined them as groups of Javascript objects that are converted to DOM `<button>`s on page load. Each button has an associated function that can be supplied to a Javascript array `filter()` function, making it very easy to specify filtering rules programmatically. Ultimately, we did not find many groupings of games that showed notable ridership trends, but this design would enable us to add more groupings if some proved interesting.

Pregame Ridership by Line, Over Time

The most interesting page components are below the filter buttons. The first chart in the first row shows average pregame ridership (versus same-season gameless days) across all underground trains and above-ground Green Line trains for several hours before each game. We did not start with the intuition that so many hours of data would be useful, but then discovered that there is indeed a large difference in ridership that begins several hours before each game. The chart below shows traffic for the Red Sox when playing the Yankees:



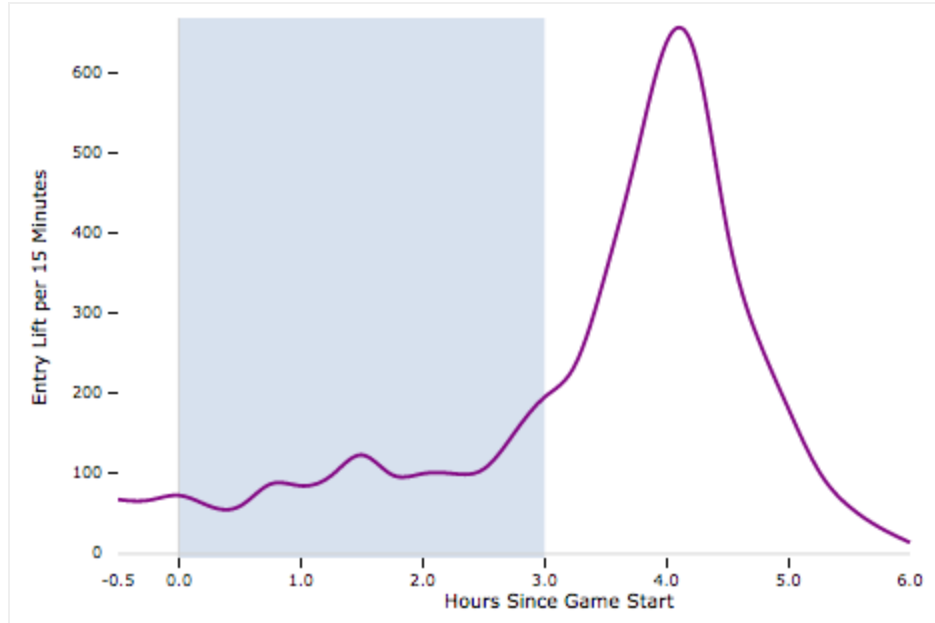
Ultimately, we chose to forego coloring keys as the colors of the plotted lines convey the T line that each represents. And as traffic can dip below the 0 lift line, we also indicate that line visually in order to ensure that users do not get confused and interpret lines relative only to the bottom axis.

Technical Details

We use `d3.svg.line` to generate both line charts in our design. Data is available at 15-minute increments, so we also employed basis interpolation for the lines themselves, giving them a smooth appearance.

Postgame Ridership Over Time

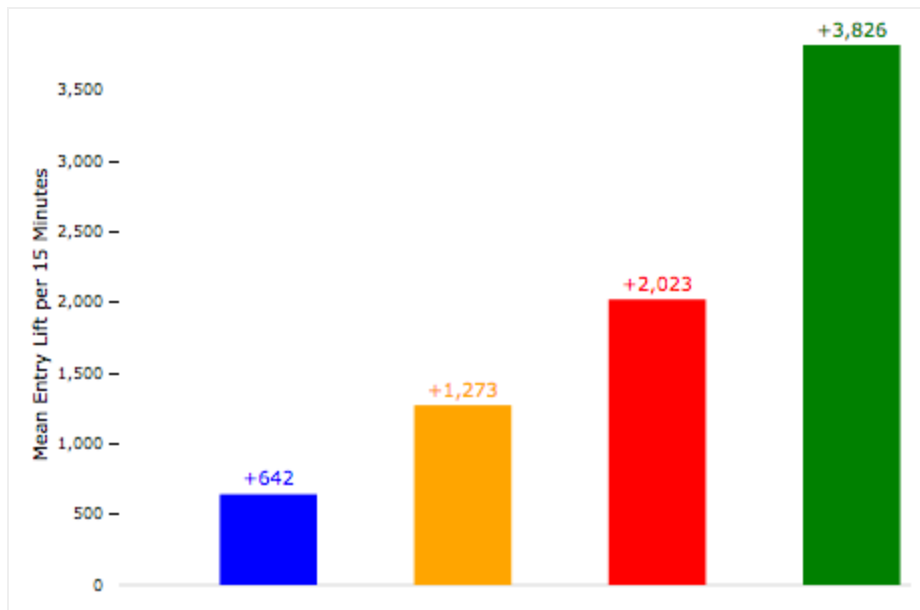
The first chart in the right-hand column shows ridership at stations near the selected team's venue during and after the game. As before, we settled on the time range presented after examining data and finding that—particularly for the Red Sox—nearby entries begin to peak before the typical game ends. The following shows postgame ridership at Hynes and Kenmore when the Red Sox play in the playoffs:



In order to make game chronology clear, we added a shaded region for the typical length of a game in each sport.

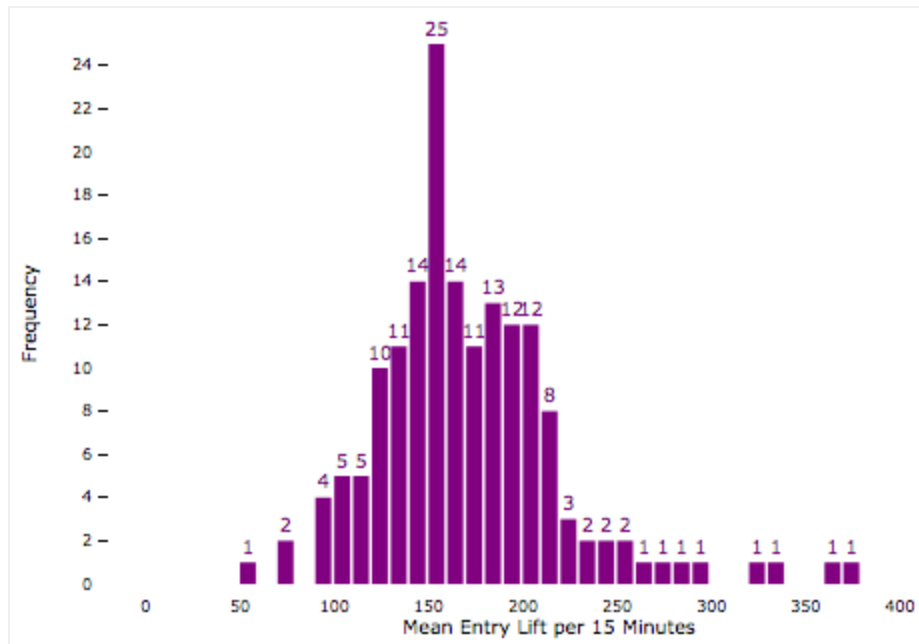
Pregame Ridership by Line, Aggregate

As suggested above, it is difficult to analyze aggregate ridership increases before games. This is true both due to the difficulty of analyzing the area between each curve and the 0 axis, and the resolution of the data, which is only available at 15 minute increments. This chart (shown for all Red Sox games below) aggregates the lift displayed in the first chart in a manner that is easier for aggregate interpretation.



Postgame Ridership Histogram

The final chart—in the lower-right corner—displays a histogram of postgame entry lift *by game* (shown below for all Red Sox games). In a sense, this chart attempts to solve the same problem as the last: the difficulty of calculating the area under the curve of the “postgame lift” chart. But because that chart features no breakdown by line, we can show the distribution across selected games.



Evaluation

With our visualization complete, we began examining the sorts of insights that we can gather from it. A collection of several appear below. Each represents an insight that we would have had difficulty reaching absent an interactive visualization system.

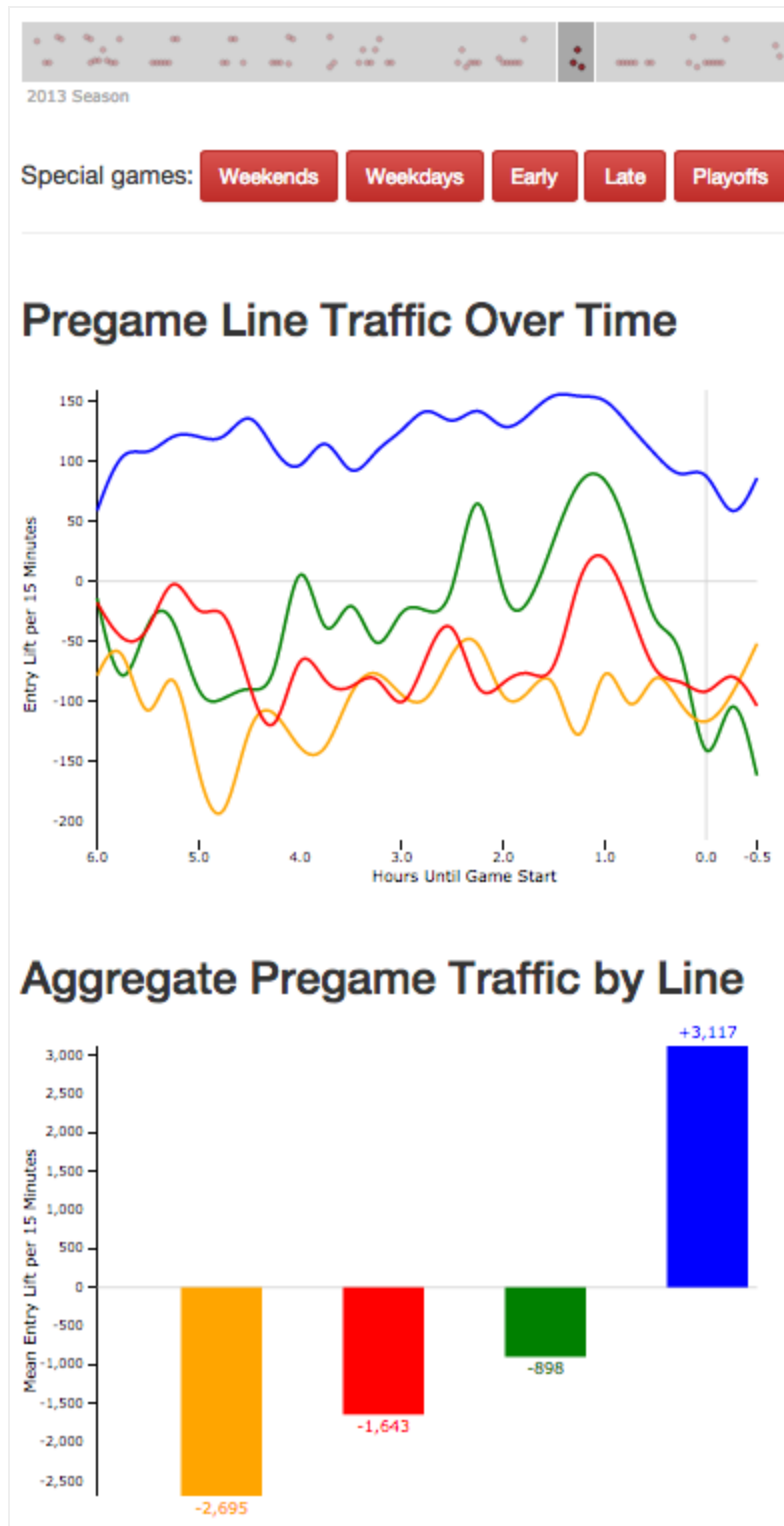
Wide Swings in Pregame Traffic

One guiding question of our analysis is the extent to which the “shape” of pregame and postgame traffic varies among games. Exploring this for different time periods and sets of games is difficult absent an interactive system. Pregame traffic by line is *somewhat* consistent among games for each sport, but sometimes shows wild swings that defy our understanding of line-specific trends.

For example, all sports show very little increase in traffic on the Blue Line before games.² But the Red Sox show several games approximately two thirds of the way through the 2013 season

² This is particularly true among Bruins and Celtics games. In the aggregate, the average Celtics game actually shows *negative* lift during games (i.e., *fewer* passengers compared to gameless days).

where pregame traffic along other lines plummets dramatically while Blue Line traffic increases. A screenshot of the left half of the visualization (with all relevant games selected) exploring the trend appears below:

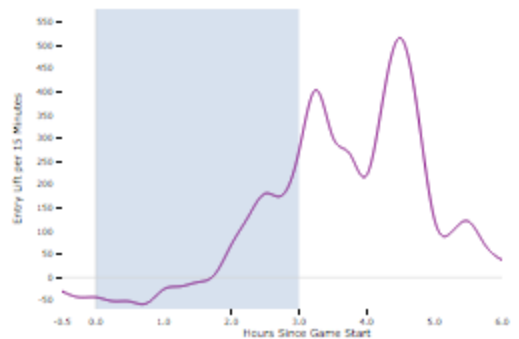


We have no explanation for this trend, though there is likely an underlying cause for this observation; i.e., it defies common sense that this anomaly appeared randomly.

In this sense, our visualization might help us locate anomalous games that should be excluded from the visualization due to their non-representativeness. We already cleaned some data from the dataset for this reason. Days that featured major service interruptions or holidays tended to skew overall measurements of mean ridership to the extent that traffic we can reasonably attribute to games became indiscernible. For example, the Bruins played games on days like December 26th and Martin Luther King Day. Because the TD Garden arena is located downtown, the underlying decrease in ridership at North Station made it impossible to reason about the lift in traffic that could be attributed to the hockey game.

Consistent Postgame Traffic

Postgame traffic showed the opposite of the phenomenon observed above for pregame traffic. As the histogram of game traffic illustrates, there is variation in *overall volume* of postgame traffic at nearby T stations. But the shape of that traffic varies very little. There are occasional edge cases where games do show odd postgame traffic distributions (see an example of such a Red Sox game to the left). But overall, postgame traffic shows a very predictable runup, peak, and decline with each game.

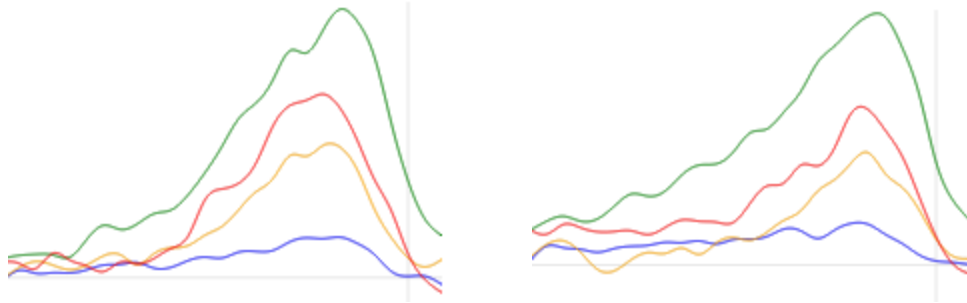


This was somewhat more surprising for Red Sox games. The aggregate distribution of postgame traffic is more gradual than for the other teams. We speculate that this occurs because people are more willing to leave baseball games early, and because games may vary more in length than basketball/baseball games. Despite this intuition, consistency in the shape of postgame traffic tended to remain consistent.

Early & Late Games

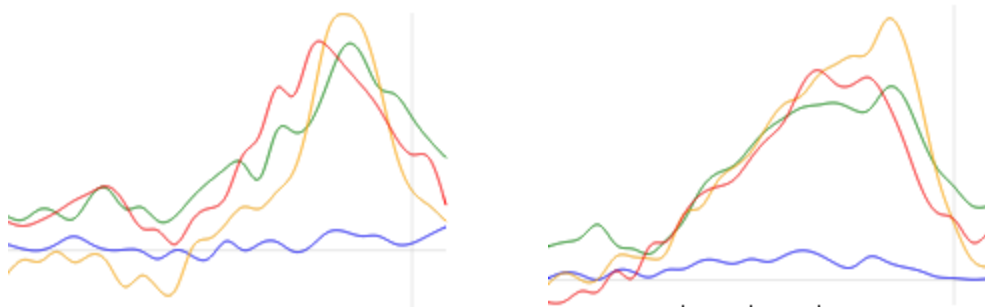
Finally, we were particularly curious about differences in T ridership for early and late games. Each sport has particular customs around game scheduling (particularly when making up cancelled games), and the Red Sox show significantly more games played early in the day than the other two teams. Our intuition suggested that early games for any sport would show diminished traffic.

Interestingly, the Red Sox show little difference between the two game times, and even show a slightly higher bump in pregame traffic before *early* games:

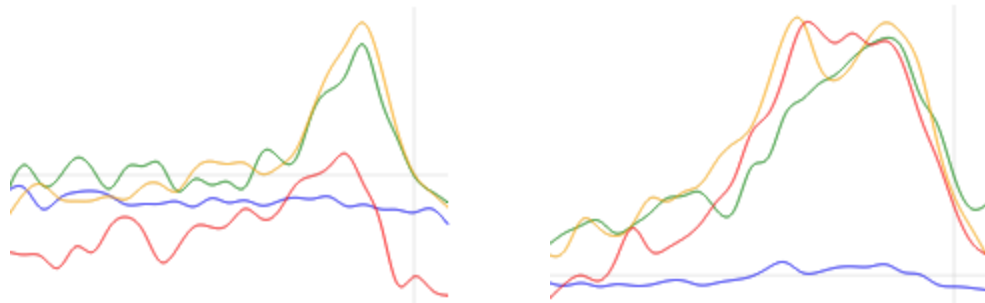


Pregame Red Sox traffic. Left: early games; right: late games.

In contrast, the Bruins and Celtics show much larger traffic ridership increases before late games than before early games:



Pregame Celtics traffic. Left: early games; right: late games. The most pronounced difference is along the Orange Line: early games are much slower to show an increase in ridership, while later games show a much earlier uptick in ridership along that line.



Pregame Celtics traffic. Left: early games; right: late games. All lines show diminished traffic, with the Red Line showing particularly low ridership.

This sort of observation is exactly the kind of insight that we hoped to facilitate with our visualization. It indicates to decision-makers that they ought think about meeting gameday demand differently depending on sports and game times.

Future Steps

As suggested above, there might be more work that we can do to ensure the integrity of our dataset. This may require more specific knowledge of traffic patterns for particular days than we currently have.

Also, our selection of filter buttons remains fairly naïve. It follows our intuition about certain games—times, days of week, specific opponents, etc.—but we were unable to derive much insight about new filters to add over the course of our design process. Our clients at the MBTA might have more insight into what kinds of games would be worthy of analysis, based on a more learned understanding of crowd control on game days. We might also want to take a “kitchen sink” approach, defining special filters for as many special games as we can think of, and paring that list back just to those that display interesting trends.

Finally, our selection mechanism does not allow arbitrary selections. The user must view all team games, a specific time range of games, or a particular set that we have defined as a filter button. Enabling more granular filters—say, through a rule-building interface or allowing users to click on games to select/unselect them—might make for a more powerful visualization. This is not a trivial task and would require careful consideration of the information to represent to users making those selections, but might create a system that could be used for regular decision-making if executed well.

Appendix 1: Project Proposal

For reference, our original project proposal is reproduced below.

Background & Motivation

We are members of a four-person team of Harvard students working alongside MBTA staff to analyze and better predict demand for public transportation. The primary goal of that project is a machine learning model to predict subway demand, accounting for recent past ridership, seasonal/cyclical effects, and sporadic events like severe weather and major events. When building these models, we realized that sporting events in particular precipitate acute changes in demand that we can detect at key stations before and after games.

We wish to use our insight about sporting events (Bruins, Celtics, and Red Sox games in particular) to visualize MBTA traffic flow by line and station (for stations proximate to venues) before and after games. We will allow users to intuitively select sets of games—likely by brushing and providing filters for non-contiguous games (e.g., weekend games) or particularly interesting sets of games (playoff games in particular). Beyond visualizing the specific games selected, we will provide useful visual comparisons to other events (so that users can detect notable games) and to days without games.

Note that our work with the MBTA is part of our capstone course (AC 297r), but this project is *not* a required component of that work.

Project Objectives

As suggested above, our primary objective is to help users understand how MBTA traffic changes before, during, and after major sporting events. Fans travel to games from different parts of the city and surrounding area (in the latter case, they often travel from far-flung T stations like Riverside and Alewife). They might come early or stay late in order to spend time in the surrounding neighborhood, and they might change in number or behavior for special games—say, for Patriot’s Day or a late-season series against the Yankees.

Beyond simple entertainment, planners might find this tool useful for anticipating large crowds or timing shifts and train runs to best accommodate swells of riders along specific lines. For this reason, we hope to add a measure of precision to our interface, reporting relevant statistical observations with our pretty pictures. We have found these traffic patterns quite difficult to analyze, and we will consider our project a success if it is not only aesthetically pleasing, but useful as well.

Data

There are two important data sources that we will use: MBTA traffic data and sports schedules.

We have MBTA traffic data going back to 2013. It shows entries (and exits) for every subway station and above-ground Green Line route in 15-minute increments. Some example raw data for April 1 of last year:

<u>locationid</u>	<u>servicedate</u>	<u>servicetime</u>	<u>entries</u>	<u>exits</u>
1002	2014-04-01	500	5	1
1002	2014-04-01	515	81	8
1002	2014-04-01	530	60	15
1002	2014-04-01	545	68	13
1002	2014-04-01	600	90	15
1002	2014-04-01	615	135	23
1002	2014-04-01	630	121	17
1002	2014-04-01	645	90	26
1002	2014-04-01	700	117	11

We also have useful metadata with information about each station, including the lines that each station serves and latitude/longitude measurements for geographic plotting (should we decide to use it).

Additionally, we have sports schedules for Bruins, Celtics, and Red Sox home games since 2013. All were manually scraped from [ESPN](#) schedule listings. A few example rows for Boston Bruins games in 2014:

<u>Datetime</u>	<u>Playoff</u>	<u>Opponent</u>
2013-01-19 19:00	0	NY Rangers
2013-01-21 13:00	0	Winnipeg
2013-01-25 19:00	0	NY Islanders
2013-01-29 19:00	0	New Jersey

2013-01-31 19:00	0	Buffalo
2013-02-12 19:30	0	NY Rangers
2013-02-28 19:00	0	Ottawa
2013-03-02 13:00	0	Tampa Bay
2013-03-03 19:30	0	Montreal

We hope that opponents may prove to be an interesting dimension for analysis. Fan reactions to the Yankees' or Canadiens' arrivals may cause particularly interesting traffic patterns to emerge.

At this point, the only thing missing from our dataset is Red Sox playoff games. We expect that those will be easy to retrieve manually.

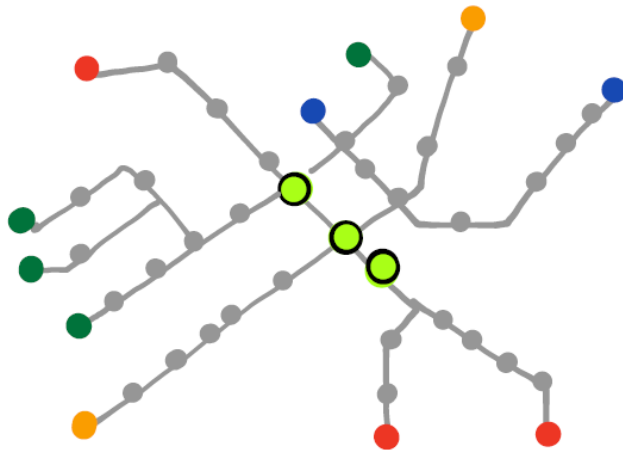
Data Processing

All of our data is very clean and low-dimensional, so we do not expect to have to contend with serious cleanliness or missing data issues. However, our MBTA is granular and voluminous—it spans several hundred megabytes and billions of individual records—it is wholly unsuitable for web visualization in its raw form. For visualization, we will have to summarize it into a compact form that still retains information pertinent to our display.

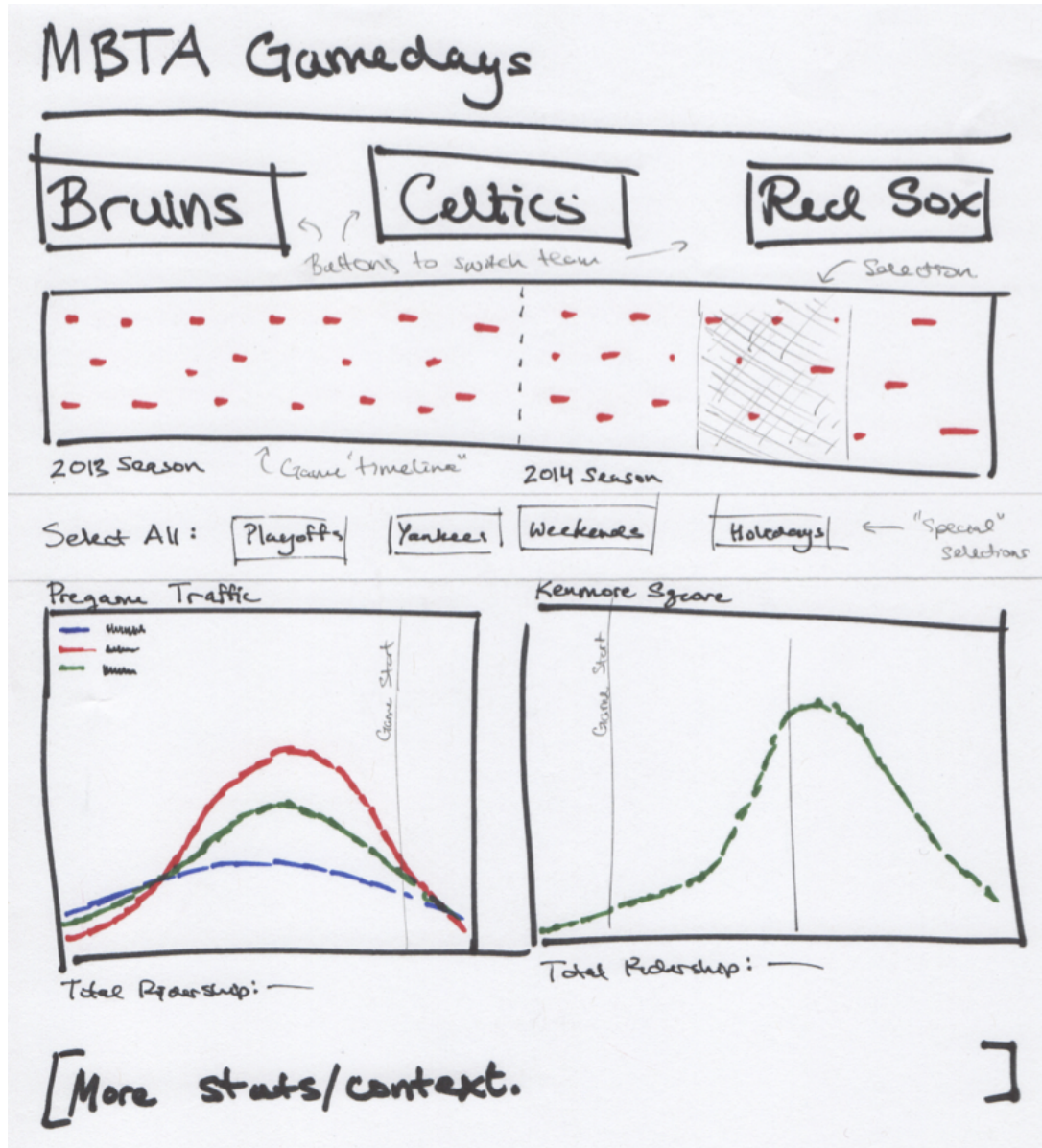
Having spent some time working with the data, we expect to be able to summarize game day traffic by line (for pre-game visualization) and by stations close to Fenway Park and TD Garden Arena (for post-game visualization). We can incorporate comparisons to average traffic in those calculations and avoid including that data in our final deliverable. This way, we can essentially “throw away” data irrelevant to gamedays themselves. Back-of-the-envelope calculations suggest that this will yield observations in the thousands—not a small amount, but an amount suitable for visualization.

Visualization

The selection of rail lines and stations could be done by providing a map representation of the MBTA rail system with the ability of selecting single or multiple stations or full lines:



The basic workflow will include a method for selecting the type of game (Bruins, Celtics, Red Sox) and the game(s) of interest. The game schedule can be displayed in a time frame indicating date/season on the horizontal axis and game start time on the vertical axis. The user can select a single game of interest or multiple ones by brushing. A different way of selecting groups of games for analysis is by selecting predefined filters including interesting sub-selections of game types such as playoffs, weekend games, and holidays. An additional grouping based on opponent will also be available. Visualizations of pre-game and post-game MBTA entries based on all selections and relevant statistics will provide a way of examining and understanding traffic trends.



Must-Have Features

In order to realize our project objective, several important features will need to be included in our final product as outlined below:

- Method for viewing and selecting specific rail station(s) of interest
- Method for viewing and selecting specific rail line(s) of interest
- View of game dates (schedule) for each type of sporting event (Bruins, Celtics, Red Sox games) including opponent information
- Method for selecting game or set of games in time range of interest
- Visualization of pre-game traffic by station or line for selected games

· Visualization of post-game traffic at stations near the sporting event for the selected games
Additional components of the traffic visualizations include:

- Indicator of starting/ending time of the sporting event taking place where appropriate
- Comparison of average traffic to game day traffic

Optional Features

After observing some final projects from last year, we realized that it might be interesting to add a “narrative” element to the visualization that highlights notable events that precipitated interesting demand patterns. But we do not yet know what those events might be (but using our visualization might help us find them!), so we do not know if such a feature would be valuable.

We also hope to make use of live updating as much as possible (e.g., charts will transform depending on brush extents). However, due to data volume this might not be possible to do elegantly. We will have to determine feasibility once we have a better handle on our data.

Project Schedule

Week of April 6:

- Gather final missing data (Red Sox playoff dates).
- Solidify main visualization page components.
- Begin page design and component layout.
- Summarize game day traffic for all games.

Week of April 13 (includes Milestone 1):

- Implement basic visualization framework, primary page elements (charts of traffic by line, traffic at nearby stations).
- Improve cosmetic appearance of visualization.

Week of April 20:

- Add ancillary interactions (hovering effects, additional contextual information).
- Add additional metrics and visual elements depending on fit with elements developed. These may include histograms, confidence intervals, summary statistics, etc.

Week of April 27:

- Polish visualization appearance (typography, artwork, colors, etc.).
- Potentially: add narrative page elements.

Week of May 4: By this point, we only anticipate minor polishing work before final submission.