Harvard University

CS171 Process Book

Weather and the MBTA

David Brown <david.brown@g.harvard.edu>
Filip Piasevoli <fpiasevoli@g.harvard.edu>
Aaron Zampaglione <azampaglione@g.harvard.edu>

CS171 - Spring 2015

Table of Contents

Overview and Motivation

Overview

Public Transportation and the Data Wave

Over the past several years, there has been a huge increase in the collection and use of data by all sorts of organizations – there has been a huge Data Wave. Public transit agencies have been collecting enormous amounts of data about their riders, but are faced with the challenge similar to so many others riding the Data Wave that transforming this data into useful, actionable information can be hard! Our goal is to develop visualizations that offer insight into how weather affects ridership which can be used by the Massachusetts Bay Transportation Authority (MBTA) to inform staffing.

Boston Transportation (MBTA)

Using data from the MBTA's fare collection system, we've been developing an initial predictive model of ridership patterns. We found that historical ridership patterns, the number of entries on the same day last week, two weeks prior, etc., to be an adequate baseline predictive model, but that we could improve upon this baseline by including weather information. This past winter, the city of Boston found out just how inconvenienced its commute could be by cataclysmic amounts of snow.

Motivation

Linear predictors like snowfall and temperature are easy enough to interpret in our forecast model because they are represented by coefficients. However this is not engaging and does not communicate the relationship effectively outside of a narrow technical audience. Our goal is to create an interface that allows for a more intuitive exploration of how varying amounts of snowfall and other weather conditions affect ridership for any MBTA rail station.

Related Work

- **MBTA Viz**
  - An interactive exploration of Boston's subway system.
  - http://mbtaviz.github.io
- **Making Data Matter**
  - The role of Information Design and Process in applying automated data to improve transit service.
  - http://dspace.mit.edu/bitstream/handle/1721.1/81640/859408312.pdf?sequence=1
- **Intermodal Passenger Flows on London's Public Transport Network**
  - Automated inference of full passenger journeys using fare-transaction and vehicle-location data.
  - http://dspace.mit.edu/bitstream/handle/1721.1/78242/830539087.pdf?sequence=1

Questions

Our intended audience is MBTA personnel. The tool is meant to help them explore the general question, "how does weather affect ridership on the MBTA?" By understanding how ridership changes with weather conditions, the MBTA can better assign personnel by having an idea of how many people will be entering each station at any given time.

After Exploratory Data Analysis (section below), we were able to list a more focused set of questions around which we built our visualization. Many weather variables do not have much correlation with ridership but large amounts of snowfall do. We created dozens of static plots to find that weather patterns other than snow make little difference on ridership, but our visualization will empower the user to explore the different weather conditions and reach the same conclusion (or prove us wrong). Some questions to motivate the user's exploration of the MBTA system are…

- At what level of snowfall can we start to see differences in number of riders entering the MBTA system?
- How do various amounts of snowfall affect the ridership of a particular station?
- Are stations near downtown Boston and in the Boston suburbs similarly affected by snow?
- Holding weather conditions fixed, do we see a different changes in ridership between the weekend and weekdays?
- Does rainfall have any impact on ridership?

Data and Processing

Data

MBTA Ridership Data

Through a partnership with the MBTA we have aggregated entries at 15-minute intervals for all MBTA railway stations going back to 2013. This data does not include light rail above ground Green Line trains on the B, C, D, and E branches. The data is relatively clean, containing a timestamp, station ID, number of entries, and number of exits. Our contact at the MBTA was not very strict about the release of the data since it does not contain any personally identifying information so we do not foresee issues with a publicly available project.

NOAA Weather Data

We have pulled Boston's weather data from the NOAA API going back to 2013. The readings are from Boston Logan airport so we believe the data to be fairly reliable for the local Boston area.

Processing

MBTA Ridership Data

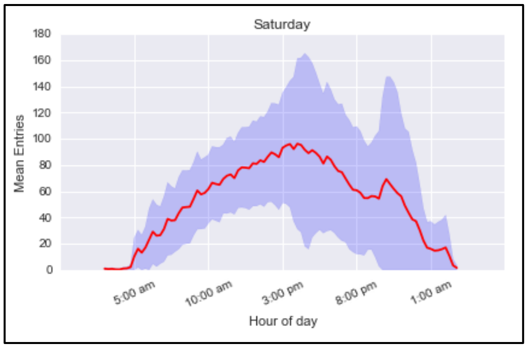
We plan to process this data in an iPython shell using the Pandas module. It provides functions to group our data set by station ID and average the entries over the 15-minute intervals to derive an accurate time series representation of each station. The output will then be made into a json file that also contains information for each station such as the station name, line, geographic coordinates, etc.

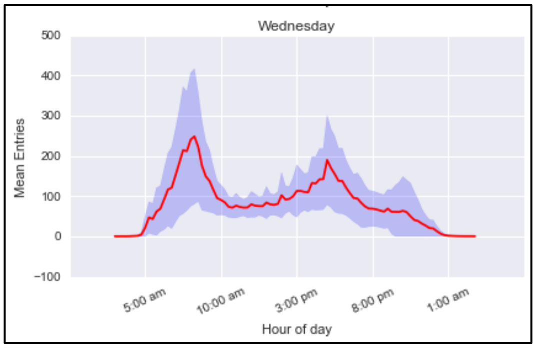
NOAA Weather Data

We will format the NOAA data in csv form so it can be loaded into Python via Pandas like above. There are a lot of climate features in the dataset that we will not be using so they can be removed. Using the ‘snowfalli’ feature that gives the amount of snowfall for a particular day in inches, we have added binary variables that correspond to bins of snowfall amount. For example, one variable is ‘2_to_4_snow’ that has a value of 1 if the snowfall for that day is between 2 inches and 4 inches and 0 otherwise. The bins have been hand-picked to have an even distribution of days within each bin since there are fewer days with 12-15 inches of snow than there are days with 2-4 inches of snow. This is expanded on in the exploratory data analysis section. This data can then be merged with the MBTA ridership data using the day portion of the timestamp as the key on which to merge.

Exploratory Data Analysis

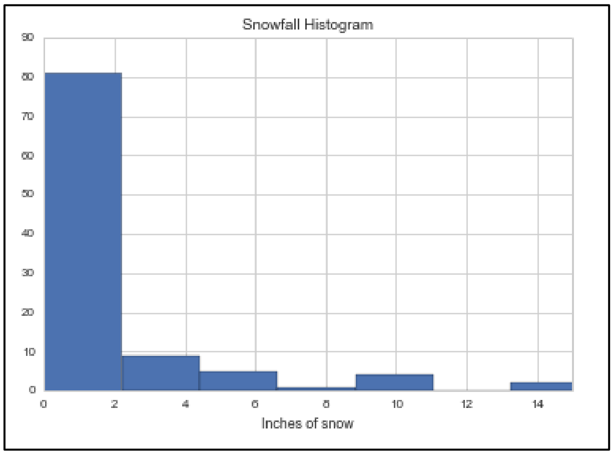
One of the first things we noticed was that for many stations, simply looking at historical trends gave a pretty accurate estimate of ridership for most weekdays. This made sense since the general population commutes to work and this is a habit with relatively low variation. Most stations had defined peaks in ridership corresponding to morning and afternoon rush hour with a mid-afternoon lull in ridership. Weekends, however, exhibited a completed different trend as ridership peaked around 3 pm with large amounts of variation across the day.





Some initial plots suggested that certain types of weather events, particularly snowfall, result in decreased ridership. Generally, more snow meant less people used the rail system. No matter how much snow fell, however, there were always some patrons that appear to be unperturbed by the snow and continued to use the MBTA. We thought that both patrons and MBTA personnel could benefit from visuals that allow them to explore how resilient particular stations are to snowfall.

Since our data only included two year's worth of ridership, we only had about 100 days where it snowed. About 75% of those days recorded snowfall between trace amounts and two inches with the other 25% ranging from three to twenty inches of snow. We had to make sure our design allowed users to explore various levels of snowfall that were observed in our data. We binned snowfall into amounts that made the most sense in terms of how people perceive amounts of snow and how many days would be included in each bin.



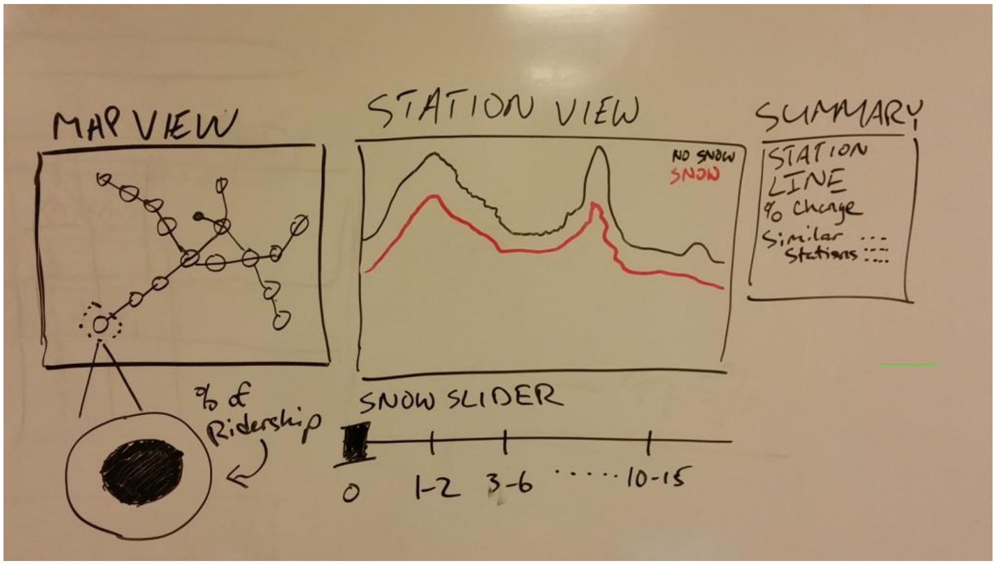
Patterns in other weather-related scenarios such as rainy days or days with variable temperatures weren't so apparent. Nonetheless, we decided to implement our design to allow users to explore various scenarios which they may face when planning their commutes.

Design Evolution

Initial Design

We will implement three views in our design. One view will be primarily dedicated to navigating the data slice. It will allow the user, using a map of the MBTA, to select a station of interest. This view will also enable exploration in the sense that the user will be able to select a particular level of snowfall and this view will display the change in ridership for each station. The high-level view will be one way the user can see which stations across the map are similarly affected. Each icon will be two concentric circles where the ratio of the inner circle area to the outer is proportional to the fraction of ridership for the user-selected snowfall. The second view will display two time series for the selected station; one series is the average entries over the day with no snowfall and the other is the average entries over the day for the user-selected snowfall. Lastly, there will be a summary view which will contain some general information about the station along with the percent change in ridership for the user-selected snowfall and a list of stations with similar changes in ridership.

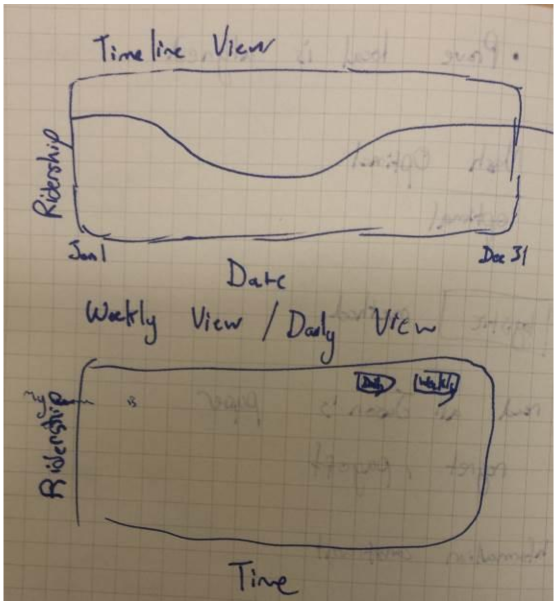
Initial Sketch



After developing a prototype and incorporating feedback from the design studio, we decided to add the following features to our implementation.

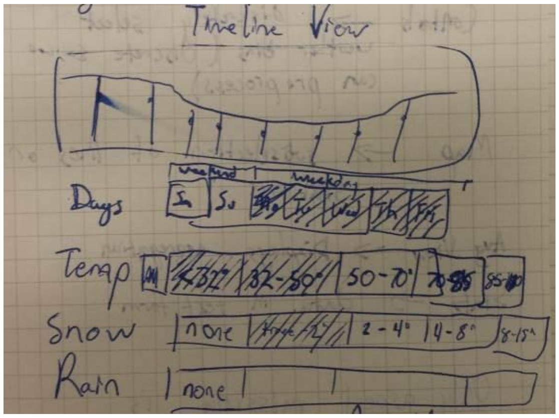
Add the ability to view the sub selection in time

After discussing the design with our TF, Benjy, it became clear that a it would be helpful to have a view showing which days were being used to generate the weather adjusted ridership. In the sketch below this is done in the top graph. We plan to use a timeline that shows the full two years of data and highlight the days that have the selected weather.



Add day of week controls

We think the user should be able to have more control than in the initial design. We considered allowing the full range of selections (show in the sketch below)



However, allowing so many user-options requires additional pre-processing and would detract from a focused user experience. We want to give the user enough freedom to make novel discoveries, but don't want to include so many options that the big conclusions lose their magnitude. Due to the size of the data, we pre-processed the data bins to provide timely interaction. We settled on allowing a weekday/weekend selection and one weather variable at a time.

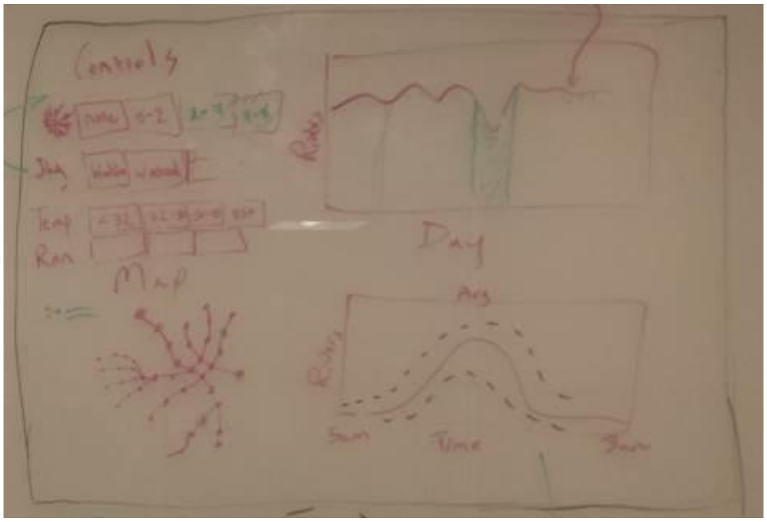
New Sketch

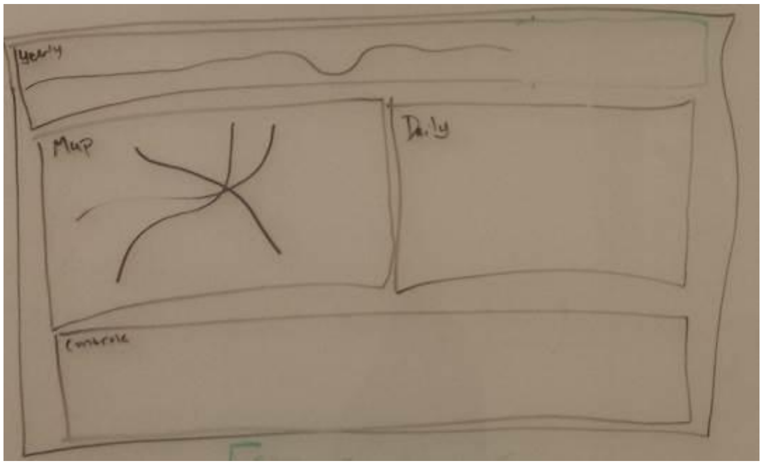
We have removed the text summary panel and now have four panels. Our layout will fit entirely onto the screen so as to be one cohesive and immersive experience. A timeline panel atop the layout will show the general trend in ridership across all MBTA stations and will be used to highlight the days aggregated in the station ridership panel. For example, if the user has selected the comparison of *0-2 inches of snow*, the timeline panel will render vertical bars at the days in our dataset when there was 0-2 inches of snow. This will help inform our user when our aggregation may contain one or two days as in the case of snowfall greater than 15 inches.

We initially planned on displaying the average ridership *of each line* in the timeline panel atop the visualization, but decided against it for a few reasons. While the questions motivating our implementation focused primarily on weather, we wanted the experience to guide the user from a high-level picture of MBTA ridership down into the idiosyncratic patterns of ridership for days with interesting weather patterns. A single trend showing average ridership across our dataset conveys the importance of seasonality across the whole dataset instead of focusing on the differences between lines. Boston, being a city largely composed of students, sees a large drop in ridership during the summer months before ridership sharply rises in September when students return to campus. Changes like this may not be so obvious on the Blue Line compared to its normal traffic since not many schools like in proximity to it, but these nuances are an important part of the MBTA system.

Below the timeline panel, we chose to divide the screen equally between the MBTA map selection panel and the change in average ridership panel. We debated how to best divide the real estate between the two windows since they both contained valuable insight. We wanted to encode information on the map when a click was registered, but the ridership panel was intended to be where the user saw the fruits of exploration so we didn't want it to be cannibalized by its neighbor. This decision forced us to decide how to intuitively encode information about similarly-affected stations into the map view's small icons representing each station. Initially, we considered having each icon consist of two concentric circles where the inner circle would fill up the parent circle in an amount proportional to the station’s change in ridership under the user-selected conditions. The icons were just too small given that we only wanted the map to take up half of the screen’s width. We decided a monochrome color scale depicting a station’s change in ridership would be a better way to relate similar stations given the size of the MBTA map panel.

Lastly, we decided to forego the station summary panel from the initial sketches in favor of a user control panel that would hold the buttons for the additional weather conditions we decided to include. Below are two rough sketches of the layout where the first has the controls above the map panel while the other has the controls at the bottom of the screen.

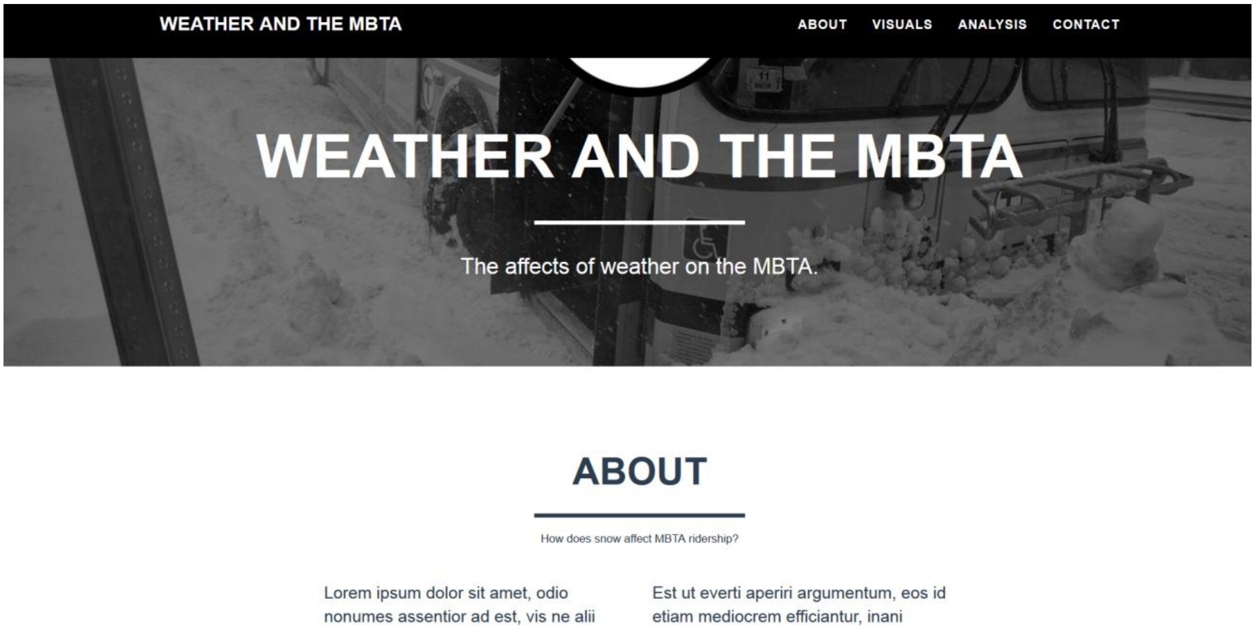




Implementation

Below we step through the various portions of our implementation and explain the technical choices and evolution of the particular modules.

Website Mockup **index.html**



WEATHER AND THE MBTA

The affects of weather on the MBTA.

ABOUT

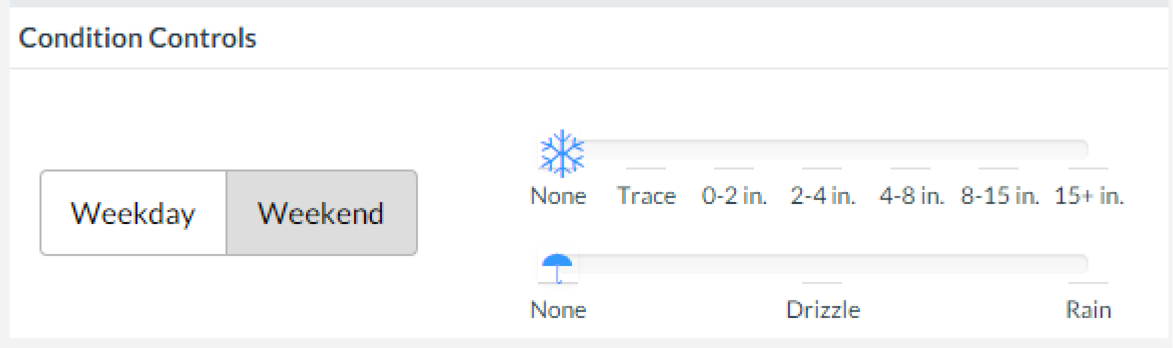
How does snow affect MBTA ridership?

Lorem ipsum dolor sit amet, odio nonumes assentior ad est, vis ne alii

Est ut everti aperiri argumentum, eos id etiam mediocrem efficiantur, inani

The landing page of our website contains all of our work on one page, allowing the user a free-flowing experience of traversing our visualization through scrolling, alone. We decided against using separate pages for the MBTA timeline panel and the mean ridership panel because we felt that it wouldn't present the issue motivating our visualization, ridership on days with poor weather conditions, in the context of normal MBTA travel patterns. In the final implementation, the visual panels and user-control panels all fit within the same window without the need to scroll down the page. Overall, we believed this decision would result in a cleaner user experience.

Control Panel



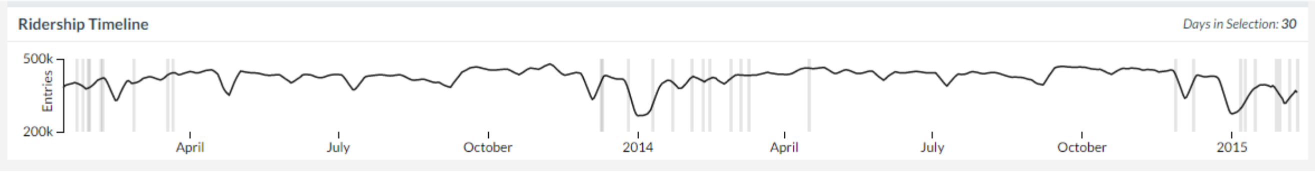
We initially designed our project to depict the way ridership changes as a function of snowfall. Earlier work suggested that weather patterns like rain and temperature and negligible effects on how willing people were to ride the T. As we continued to refine our implementation, we decided to include the capability to explore changes by weekend or weekday along with the added option of rainfall. Even though early work suggested that rain didn't affect ridership, we felt that it was important to give users the capability to come to the same conclusion on their own. Seeing negligible effects for varying amounts of rainfall makes the severe differences in ridership due to snowfall that much more impressive.

Changes in this control panel precipitate linked responses in the timeline, map selection, and ridership panels. See their respective panels below for explanations of how each panel responds to user events.

Notes:

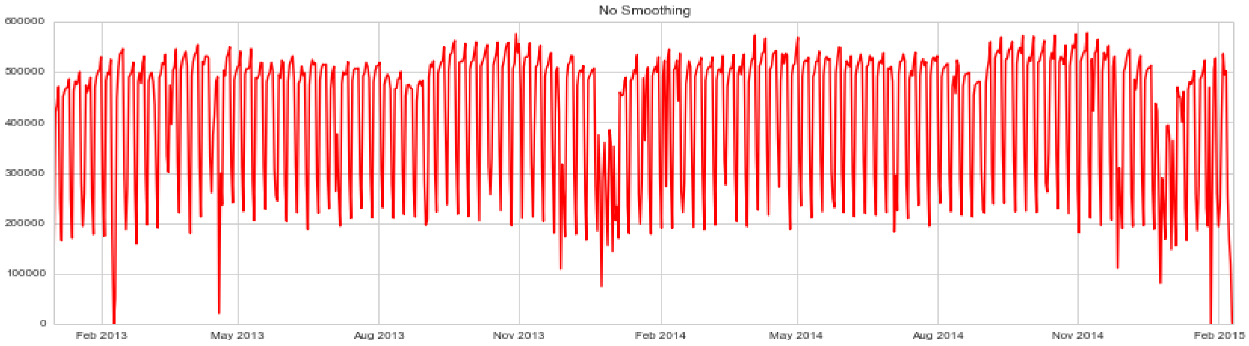
- The user cannot select to view ridership for both snowfall and rainfall at the same time; at least one slider must be set to 'None'.
- The comparison for ridership on days with no snow and days with user-selected amounts of snow are adjusted for seasonality. The time series for days without snow is computed from days in October-April without any snow since those are the months with snowfall in our data. Summer ridership is significantly different than Fall/Winter ridership so we wanted to make sure our implementation accurately represented the changes due to weather alone.

MBTA Timeline View



In our initial sketches, we didn't include the MBTA timeline view. After the first meeting with our TF Benjy, he suggested that we implement a way to show the users how many days we used to generate our plots. The entire dataset contains about 100 days on which it snowed, but 75% of those days only had 0-4 inches of snow. Upon allowing the user to select between weekdays and weekends, some scenarios like *weekday ridership for trace amounts of snow* had dozens of day's worth of data while *weekend ridership for 8-15 inches of snow* only included one day's worth of data. When the user selects a weather condition, the timeline renders vertical bands at the days which were used to construct the ridership plot. Even though we could do little to remedy underflow in the snowfall data, we felt it was important for the user to realize the inherent variability of data presented on only a handful of data points.

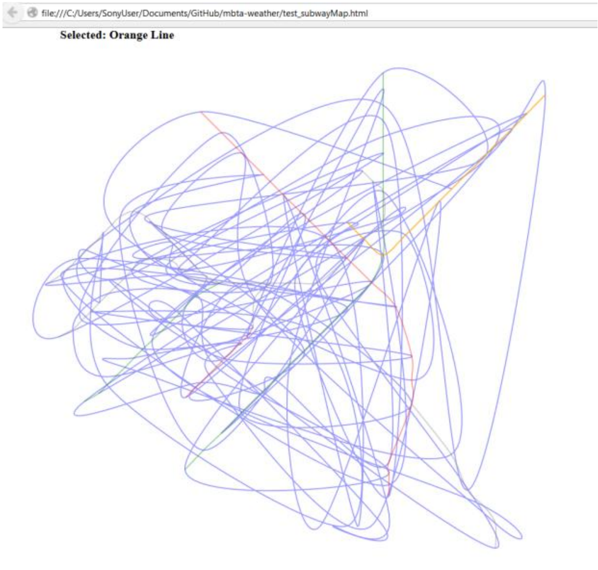
In addition to clarifying our other plots, this timeline is an important storytelling element to our implementation. The line represents the aggregated entries across all MBTA lines after smoothing via moving average. Below is a picture of the timeline without any smoothing. The large difference between weekday and weekend ridership causes the extreme periodicity. We first used a moving average window of 7 days to smooth the series then performed a second iteration of smoothing over 7 days to achieve a plot to our liking.



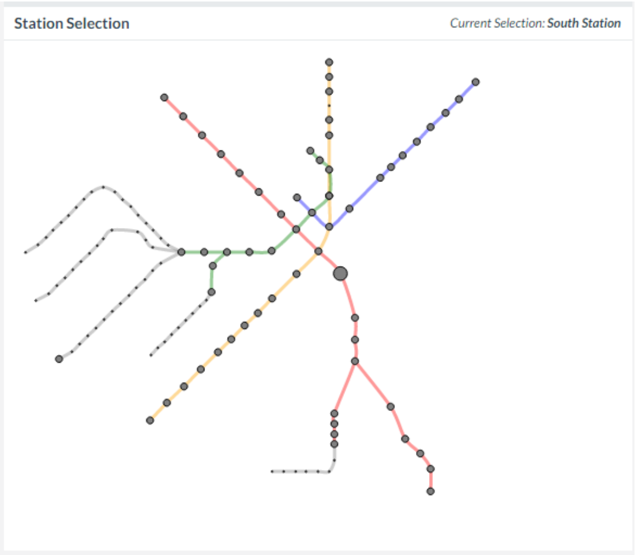
The smoothed timeline gently exposes the user to the idea of visualizing the MBTA. Several valleys correspond to familiar time periods when fewer people are using the MBTA; the large dip surrounding the New Year corresponds to the holiday season when many people are on vacation and there is a lull between June and September corresponding to colleges being out of session. The start of the school year in September sees a significant increase in ridership above what we see in the summer months. While this plot doesn't look to directly answer any of our questions regarding ridership under weather conditions, its place atop our visualization helps paint a cohesive picture of ridership before looking at the subset of ridership on days with certain weather conditions.

Map Selection Panel

Below is an initial map with rollover selection, rolled over downtown crossing. We chose to use the standard MBTA map layout due to its familiarity. Another less intuitive option would have been to allow selections via a Google Maps rendering of the stations, but since people are used to thinking of the MBTA in terms of the map, this was the obvious choice. Early implementations had a few bugs when scaling the station locations in images of the MBTA map to svg elements in our panel.

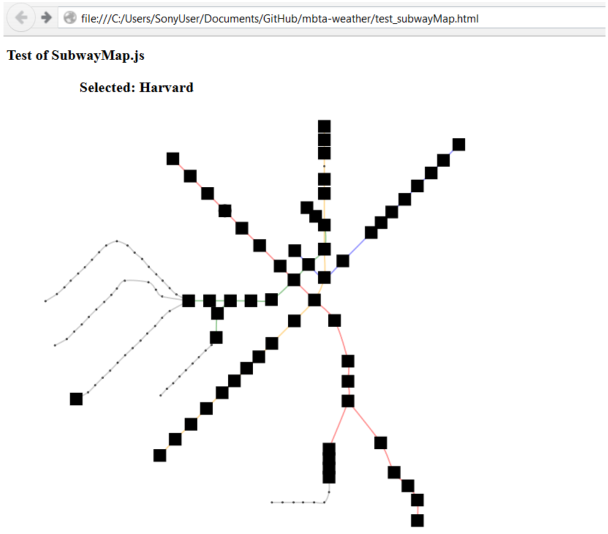

file:///C:/Users/SonyUser/Documents/GitHub/mbta-weather/test_subwayMap.html
Selected: Orange Line

Bugs with Scaling of Lines



Proper Scaling and Rendering

The initial implementation allowed users to change stations simply by rolling over an icon and displaying the station name atop the panel. Selecting a station changed the ridership panel to display the mean entries time series for the user selection. Later implementations opted for click selection instead of roll-over selection, and included the option to select an entire line instead of just one station. The darkened areas below show the areas that register mouse-over events for each station and each line.

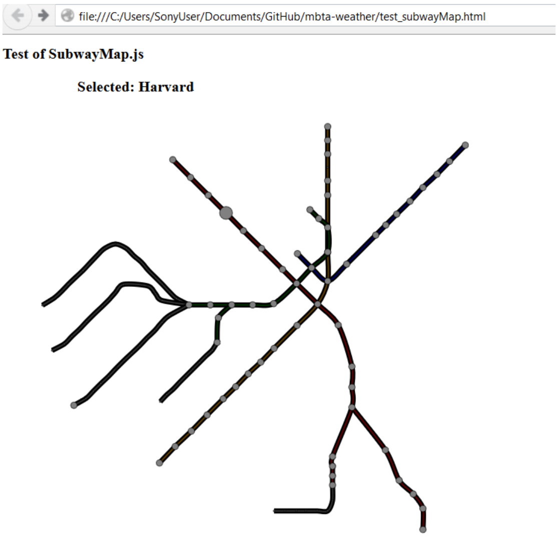


file:///C:/Users/SonyUser/Documents/GitHub/mbta-weather/test_subwayMap.html

Test of SubwayMap.js

Selected: Harvard

Mouse-over area for stations



file:///C:/Users/SonyUser/Documents/GitHub/mbta-weather/test_subwayMap.html

Test of SubwayMap.js

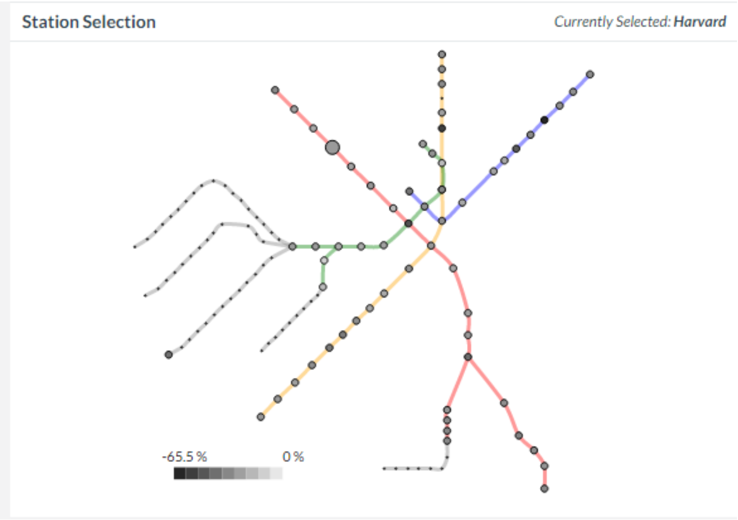
Selected: Harvard

Mouse-over area for lines

In order to make a cohesive final product, we decided to link this map selection panel to change according to user-selected weather conditions and convey information about stations that lose a similar proportion of ridership. The mean ridership panel does a great job of showing how one particular station is affected in for a given weather condition, but it's difficult to select two stations and interpret similarity. Humans aren't adept at interpreting the area under curves, computing the proportion between two areas, and then trying to relate that to another station's curves. Therefore, we chose to encode information about the percent-change in ridership at a station into the map selection panel.

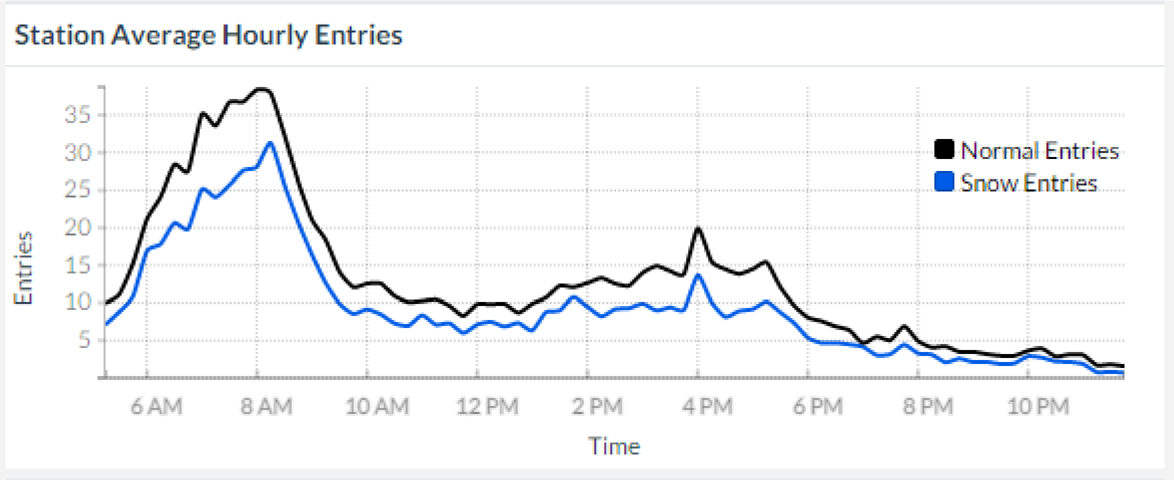
Initially we envisioned having each icon representing consisting of two concentric circles where the proportion of the inner area to the outer is equal to the proportion of ridership at the station given the user-selected weather scenario. However, the size of the icons made concentric circles and their areas a poor choice to convey meaningful information.

We opted in favor of using a monochromatic scale where darker gradients denote a greater loss in ridership and white denotes no change.



With information encoded this way, our map selection view is no longer just a navigational tool for the user. Now, changes in selected weather conditions trigger changes in the timeline panel (which days satisfy the weather conditions), the map selection panel (which stations lose a similar proportion of ridership), and the mean ridership panel (comparison curve showing ridership for days with user-selected weather). This was a major factor in making our implementation an immersive experience that gave the user multiple layers of information in order to build conclusions.

Mean Ridership Panel



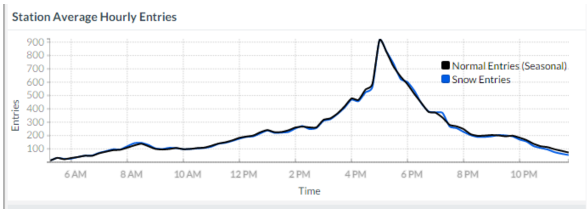
The mean ridership panel is the central focus of our implementation, showing the average entries aggregated over fifteen minute intervals. We set forth with the goal of seeing how an individual station’s ridership changes as a function of different weather conditions. This panel reflects changes to both the station of interest and the conditions of interest. Having weather options ‘Rain’ and ‘Snow’ set to ‘None’ displays the mean ridership pattern at the station on days without special weather patterns to influence rider demand. While changing between the various levels of weather severity, the user explores at what point people decide the weather is too bad to continue their normal routines.

This was the least-changed element of our implementation from inception to final product. Time series plots are relatively common and easy to interpret, so we dedicated our time to creating the clean look of the panel and to fleshing out interactive options built around this panel so that the user can explore a wider-variety weather conditions.

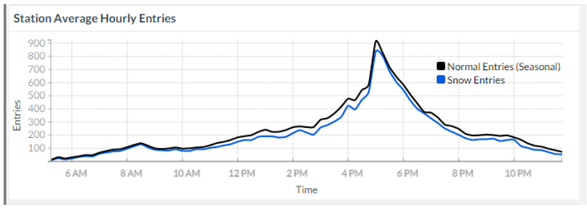
Evaluation

After how much snow does ridership start to decrease?

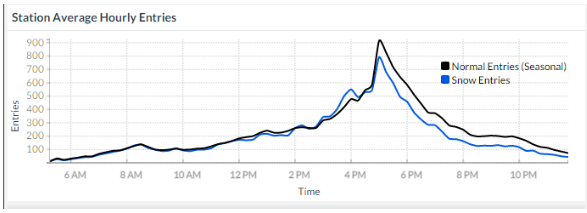
Using our visualization tool, we found that people did not ride the T less when snowfall ranged from trace amounts to 4 inches. Below is Harvard ridership for trace, 0-2 inches, 2-4 inches, and 4-8 inches of snow compared to normal ridership on weekdays. The control panel made it very easy to toggle between the different levels of severity while still having the ridership plot in the same window.



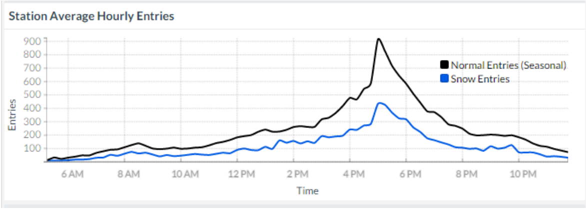
Trace Amounts



0-2 inches



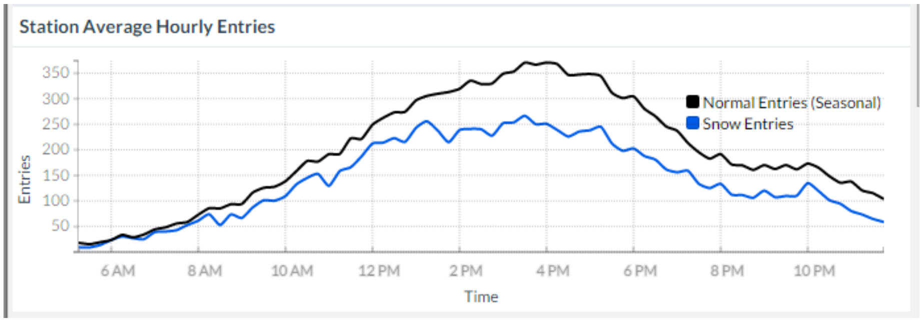
2-4 inches



4-8 inches

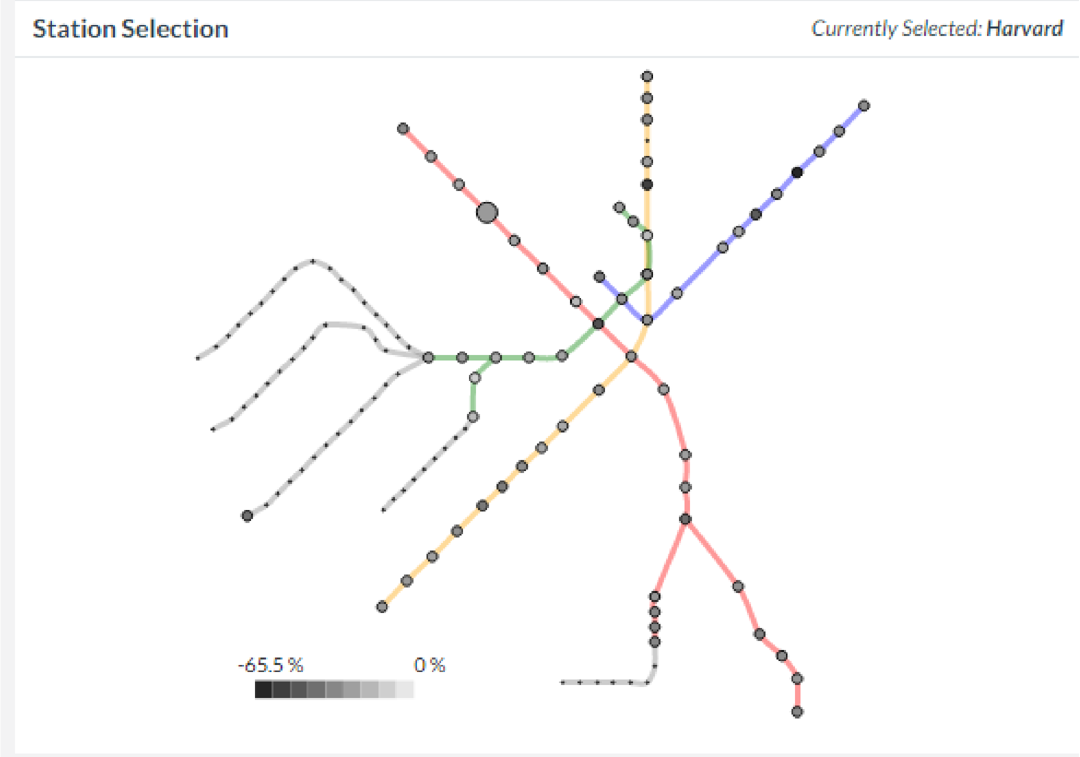
We saw this pattern at stations across the MBTA map, and the drop in ridership was even more significant for snowfall greater than 8 inches. This confirms intuitions about ridership from earlier research, and we were able to make more resounding conclusions because the tool made it easier to generate the plots for any station than it was when using python scripts.

Our tool also allowed us to explore the effects on weekend ridership, a subset of our data that was previously unexplored, and we were able to see that ridership noticeably decreased for much lower amounts of snow. Below is the weekend ridership at Harvard for trace amounts of snow. Recall that weekday ridership didn't experience losses like this until 4 inches of snowfall. It appears that the routine of the riders is a lot less compulsory than the weekday commute, and riders are much more willing to forego their travel if there are even little amounts of snowfall. However, not all stations were so sensitive to weekend snow.



Harvard Weekend Ridership, Trace Amounts of Snow

Do stations further away from the city lose a greater portion of their ridership when it snows compared to stations near downtown Boston?

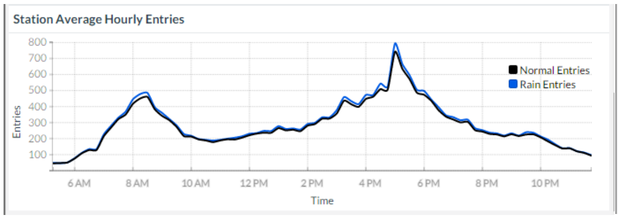


Ridership across the MBTA, 4-8 inches of snow

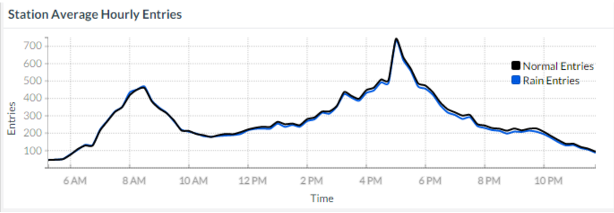
Looking at the map selection view for varying levels of snowfall, it's difficult to see any pattern of similarly affected stations. Those stations' icons furthest from the city on the red, orange, and blue lines are no more or less shaded than those closer to the city, suggesting that the loss in ridership at those stations are proportionally similar. If geographic location had any effect on how much ridership stations lost as a result of snow, one would expect to see drastically different shadings in the map selection view. One station with interesting ridership patterns for snowfall was Suffolk Downs. Even marginal amounts of snow led to significant decreases in entries at this station.

Do different amounts of rain have any effect on ridership?

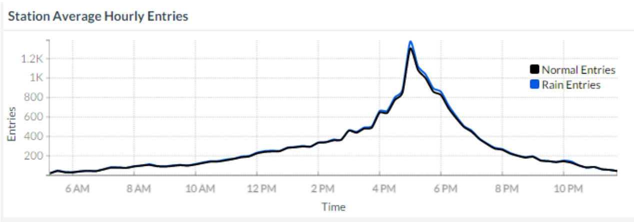
Previous work suggested that rain ridership didn't change significantly when it rained. However, it was difficult to definitively say that rain had no effect when we couldn't efficiently observe ridership at all stations in our dataset. Using our visualization, we can now say with much more certainty that patrons don't change their normal ridership patterns when it's raining. Below you can see that ridership at Harvard and Downtown Crossing doesn't significantly change for drizzle and more substantial rainfall.



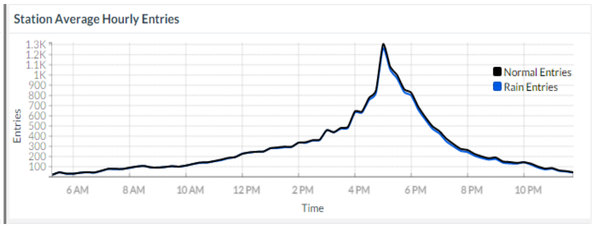
Harvard Drizzle



Harvard Rain



Downtown Crossing, Drizzle



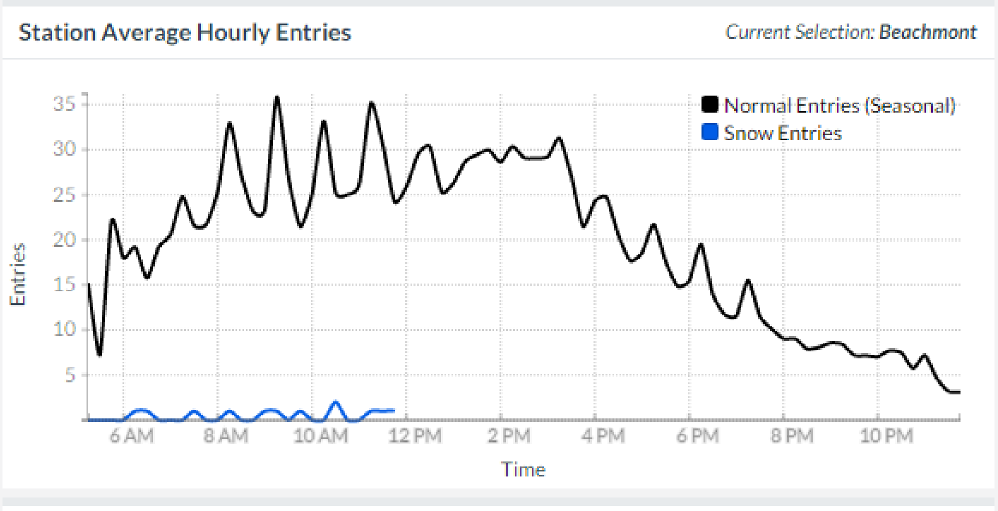
Downtown Crossing, Rain

Visualizing the 2013 February Nor'easter

Setting the user controls to 'weekend', '8-15 inches of snow', we see that only one day is in the selection. That lone day is Saturday February 9, 2013 which fell in the middle of a severe nor'easter snow storm. The MBTA stopped service the day before[1] so the blue line representing ridership on Saturday must be MBTA personnel entering at a rate of one or

[1] http://boston.cbslocal.com/2013/02/09/behemoth-storm-barrels-through-new-england/

two people per 15 minute interval. All stations either have no ridership or ridership that stops sometime before noon.



Beachmont Ridership During 2013 Nor'easter

Comments on the Visualization

Our completed visualization has a clean user interface, loads quickly, and seamlessly transitions as the user changes between weather conditions and weekday type. We're very happy that we built upon our initial sketches and included the timeline view and the option to toggle between rain/snow and weekend/weekday. The timeline view presents an intuitive high-level view of ridership while putting the user-selection and the data into a proper context to make confident conclusions.

Without even considering the effects of weather, the curious user can observe the normal ridership patterns of stations that reflect the common ideas of urban transportation. Statements like 'stations further away from the city center are most often entered by commuters in the morning' and 'weekend ridership peaks in the middle of the afternoon' are not earth-shattering conclusions. Having the data reflect our basic intuition about public transportation provides the user with a good introduction to visualizing the MBTA before diving into the weather data.

We believe that we were able to present a lot of information through a clean visual tool without the need to scroll or change the page. If we were to further develop the tool, we might consider allowing the selection of two stations while representing both plots stacked on top of one another. It wouldn't make sense to combine the ridership plots into one plot since each station comparison already generates two lines, and we wouldn't want to confuse the user with too many lines. One station may have entries that are two orders of magnitude higher than another, making comparisons between the stations difficult to display without scaling first.