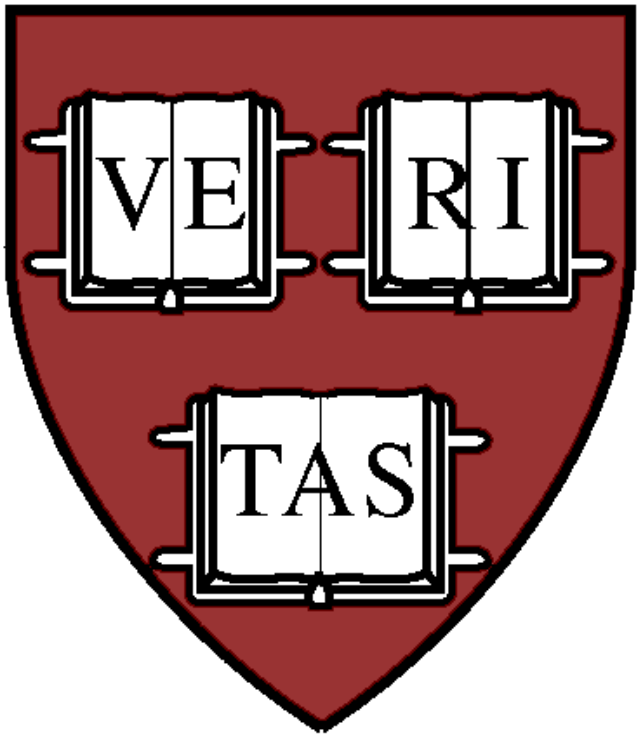
Harvard University



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. Since the city of Boston is not usually hit by such historical amounts of snow, we looked at more typical amounts of snowfall in previous years and their effect on ridership.

Motivation

Linear predictors like snowfall and temperature are easy enough to interpret in our forecast model, they are represented by a coefficient. However this is not engaging and does not communicate the relationship effectively outside of a narrow technical audience. Our goal with this project is to create an interface that allows for a more intuitive exploration of how varying amounts of snowfall 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

- Initial Question: How does weather affect ridership on the MBTA?

After Exploratory Data Analysis (section below) the questions changed:

Many weather variables do not have much correlation with ridership but a large amounts of snowfall does. We will want to show at least one other weather variable along with snow so it can be compared. We also want to investigate

- 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?

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.

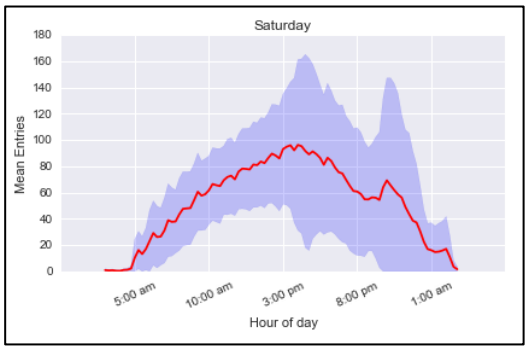
NOAA Weather Data

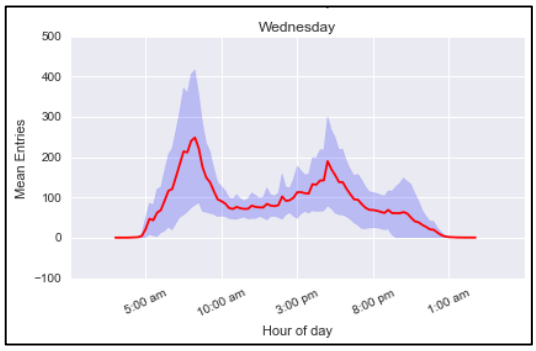
We will format the NOAA data in csv form so it can be loaded into Python via Pandas like above. There are a lot of climate features in the dataset that we will not be using so they can

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

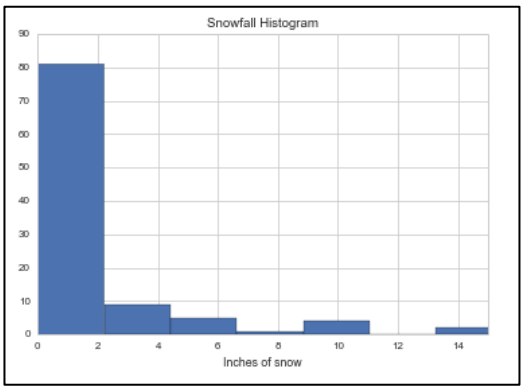
One of the first things we noticed was that for many stations, simply looking at historical trends gave a pretty accurate estimate of ridership for most weekdays. This made sense since the general population commutes to work and this is a habit with relatively low variation. Most stations had defined 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.



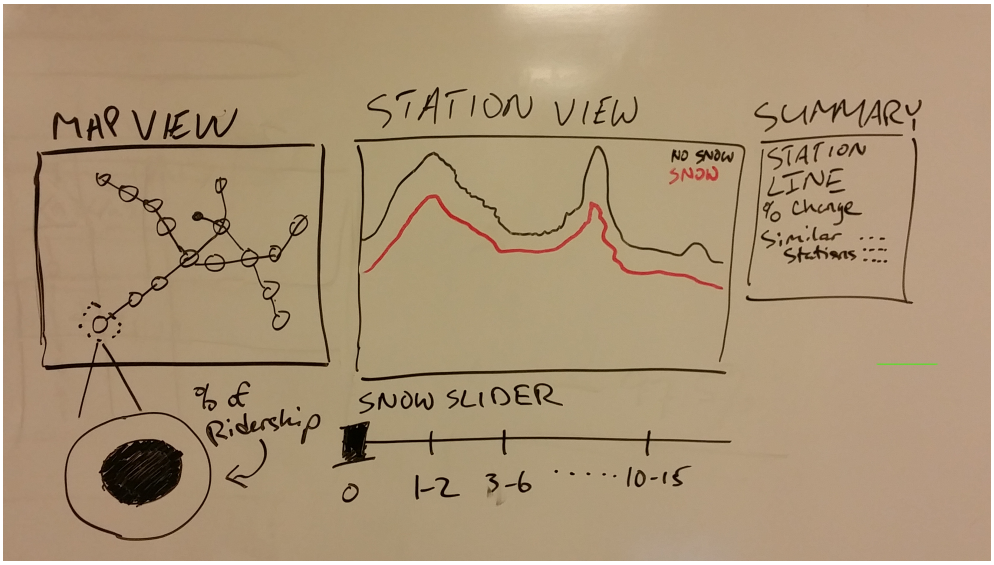
Patterns in other weather-related scenarios such as rainy days or days with variable temperatures weren't so apparent. Nonetheless, we decided to implement our design to allow users to explore various scenarios which they may face when planning their commutes.

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.

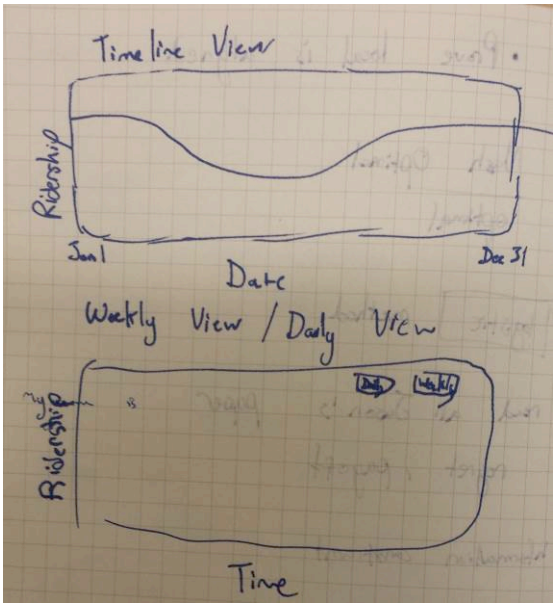
Initial Sketch



After prototyping our design and examining it critically, including in the peer review design studio, the idea has evolved.

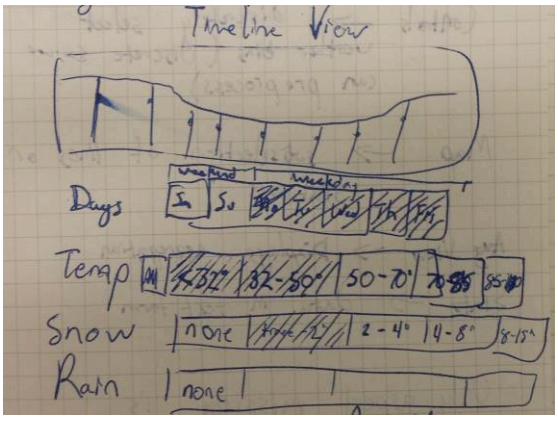
Add the ability to view the sub selection in time

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

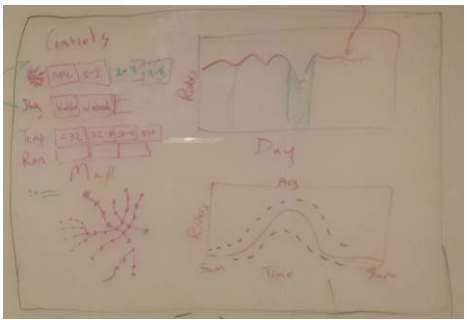
We think the user should be able to have more control then in the initial design. We considered allowing the full range of selections (show in the sketch below)

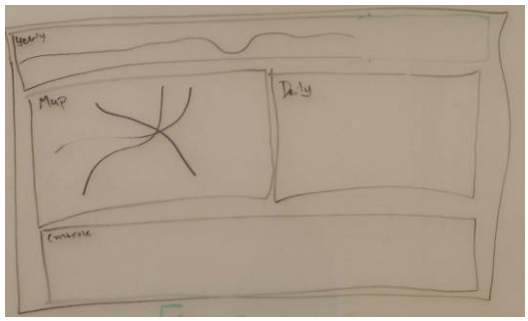


However this prevents effective pre-processing of data and given our audience we should favor clear targeted information over unlimited exploration. Due to the size of the data we want to pre-process the data bins to provide timely interaction. We settled on allowing a weekday / weekend selection and one weather variable at a time

New Sketch:

We have removed the text summary pane and now have four divisions. A control panel, a map selection panel, a timeline panel, and a station average ridership panel. We are considering two layouts of these four panels shown in the two sketches below. We haven't decided between the two yet.

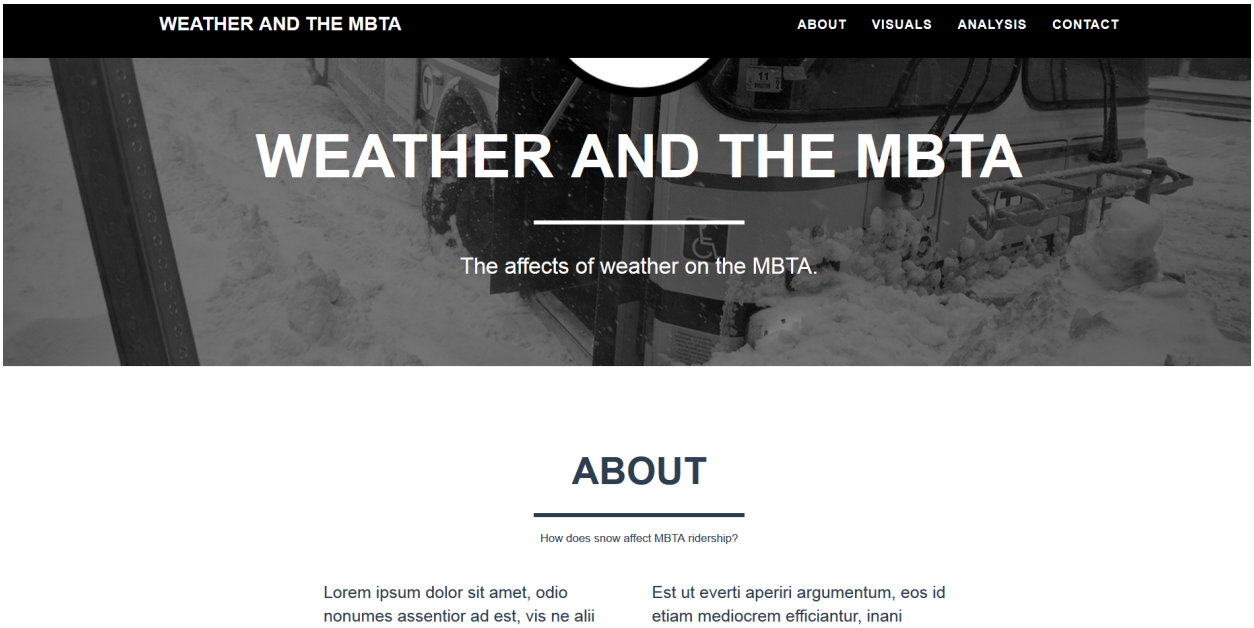




Implementation

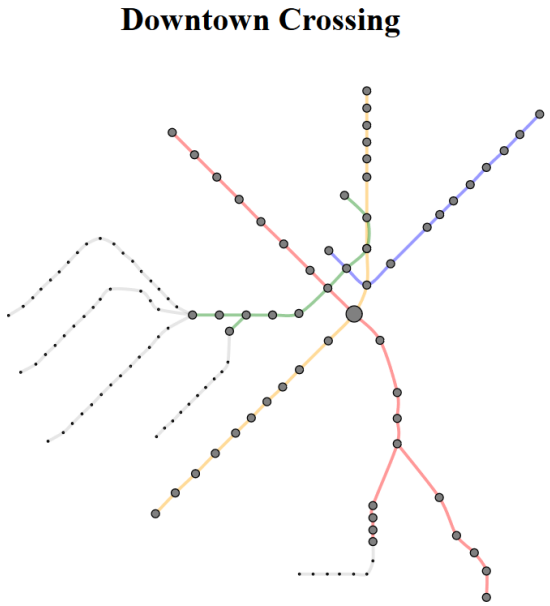
Section will be updated when we reach this stage of the process

Website Mockup **index.html**



test_subwayMap_simple.html

Initial map with rollover selection, rolled over downtown crossing. After feedback we will add rollover context with click selection



Evaluation

Section will be updated when we reach this stage of the process