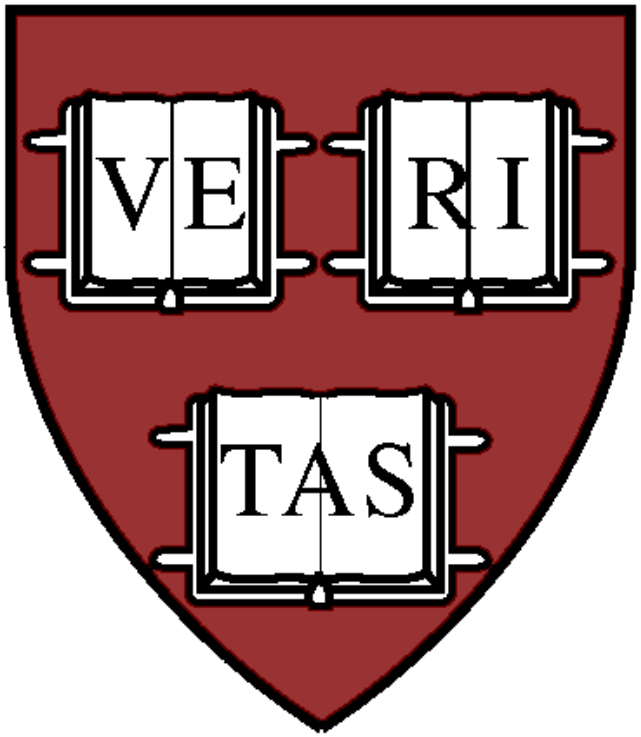
Harvard University



CS171 Project Proposal

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

Background and Motivation

Background

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 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 there affect on ridership. Using data scraped from NOAA (National Oceanic and Atmospheric Administration), we found ridership to generally decrease with larger amounts of snow.

Motivation

Linear predictors like snowfall and temperature are easy enough to interpret in our predictive models, but the actual effect of snow on ridership, holding all other predictors constant, is best understood visually. The number of entries for a particular station can be viewed as a time series with measurements of the aggregated entries at 15-minute intervals. Averaging over all weekdays, we can get a fairly accurate picture of average ridership patterns for a particular station. Filtering our data to a subset of days when it snowed, we

can compare average ridership patterns for days without snow to days with a level of snowfall selected by the user. Looking at a coefficient in a linear model corresponding to the amount of snowfall is a purely quantitative way of understanding snow's effect on ridership, but it is not the most engaging or descriptive. The goal 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.

Project Objectives

- Visualize how various amounts of snowfall affected the ridership of a particular station
- Determine what level of snowfall can we start to see differences in number of riders entering the MBTA system
- Across all MBTA stations, identify those stations which experience similar changes in ridership for equal amounts of snowfall. We have an initial hypothesis that those stations further away from the city lose more riders when it snows than stations closer to downtown, but will have to test it with data.

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 will pull 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.

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.

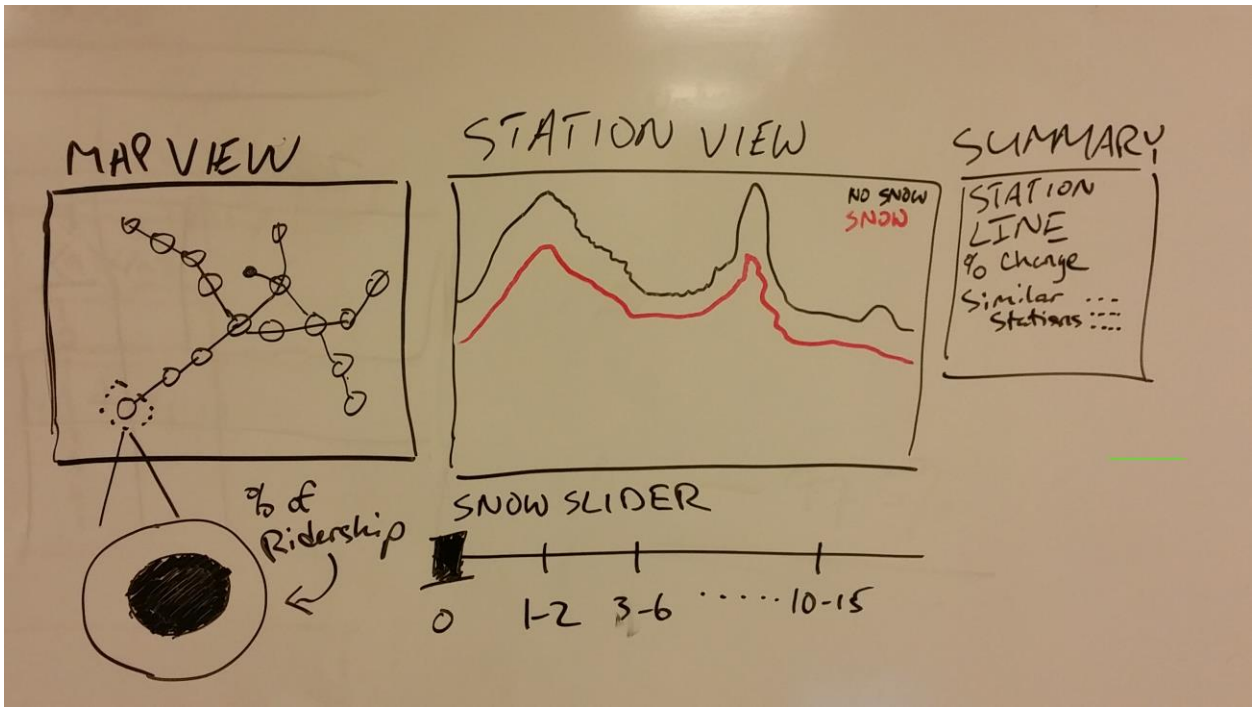
NOAA Weather Data

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, we will create binary features that bin our days according to snowfall. For example, one feature may be '2_to_4_snow' that has a value of 1 if the snowfall for that day is

between 2 inches and 4 inches. The bins will be hand-picked so 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 data can then be merged with the MBTA ridership data using the day portion of the timestamp as the key on which to merge.

Visualization

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.



Must Have Features

- Ability to select snowfall amount interactively
- Coordinate views with all interaction
  - Map station selection → time series graph changes
  - Map station selection →summary view changes
  - Snowfall selection change → map view global data view changes
  - Snowfall selection change →time series graph changes
  - Snowfall selection change → row update in summary view

Optional Features

- List stations with similar ridership for given amount of snowfall in summary view (need to define "similar")
- Ability to select from other weather conditions besides snowfall (rain, temperature, etc.) – (need to be careful of seasonal dependence)

Project Schedule

Make sure that you plan your work so that you can avoid a big rush right before the final project deadline, and delegate different modules and responsibilities among your team members. Write this in terms of weekly deadlines.

April 10

- Cleaning MBTA data in json format (Fil)
- Scraping weather data (Aaron)
- Cleaning and converting weather data in json format (Fil / Aaron)
- MBTA map as an svg (Dave)
- Initial website mockup (Aaron)

April 17 (milestone)

- Choosing snowfall bin sizes (Fil)
- Independent Map View (Dave)
- Independent Summary View (Fil)
- Independent Time Series View (Aaron)

April 24

- Initial coordinated view (all)

May 5 (project)

Project due

- Website
- Screencast
- Process Book