William Lu
Shane Kelly
Kevin Crispie

/ mbtaPredict

/ Big Idea

We plan on developing a predictive tool for MBTA ridership based on station, date, and weather factors. Our minimum viable product is a program that takes in a station and uses historical data to predict the number of riders that will use that station on that particular day. Further iterations of the project would include a GUI that allows a user to query the date and the station name, and be able to access and visualize the data. Additional features of the project could include more detailed predictions based on time of day ridership.

/ Learning Goals

- Build statistical prediction skills (Team)
- Learn how to parse, analyze, and correlate large sets of data (Team)
- Manage large group (Team)
- Learn how to use pandas (Team)
- Learn some machine learning techniques/tools (Kevin)
- Make an awesome GUI (William and Shane)

/ Implementation Plan

We plan on using historical data published by the MBTA, along with weather data from Weather Underground. We plan on using this data to develop our prediction tool to see how many riders will use each station on a given day. If we have time, we then plan on developing a GUI, where the user can then query a station and date and figure out how busy the station will be that day. If we have time after that, we will spend our time improving the GUI and refining the prediction tool, perhaps adding more detailed predictions (based on time of day instead of just date).

/ Project Schedule (everything is pending access to data)

- Parse weather data (CSV / pending access to MBTA data to determine timeframe) 3/30/15
- Parse MBTA train location data (file format unclear) 3/30/15 (Pending access to data)
- Parse MBTA turnstile data (file format unclear) 4/4/15 (Pending access to data)
- Analyze / correlate data 4/10/15
- Implement machine learning 4/15/15
- Create an initial prediction algorithm 4/20/15
- Refined prediction algorithm 4/25/10
- Make initial GUI 4/30/15
- Implement data presentation libraries 4/30/15
- Draft website 4/31/15
- Finished GUI 5/5/15
- Additional Refinements/Website done 5/6/15

/ Collaboration Plan

For the most part, we plan on dividing our project into parts and then working individually on each part, checking in frequently (3-5 times per week) with each other to make sure that we are all on the same page. We also plan on doing some amount of pair programming, especially for the most difficult aspects of the project. When we do pair programming, we will make sure to allocate a large amount of time (at least 1-2 hours) to meet, in order to get the full benefits of pair programming. We will likely only pair program, as opposed to programming in a group of three.

/ Risks

We need to obtain information from the MBTA for their daily ridership data. This data is not directly available via the MBTA website, but is available upon request. We have contacted the MBTA, but have yet to receive the data. We should be able to get this data from them, but we do need it in order to complete the project.

/ Additional Course Content

Since our project primarily concerns data analytics and prediction, we think that learning some data analysis techniques for python would be very useful. Specifically, covering some machine learning and statistical analysis during class time (not just a toolbox) would be incredibly helpful.