PROJECT 2 PROPOSAL

Min Su Kim, Zach Kamran,
Varshant Dhar

Our project design centers around answering an array questions on bike-sharing usage on San Francisco, using an open sourced dataset in Kaggle[1]. We wish to visualize the bike-sharing patterns: more specifically, visualize whether or not different days concentrate the gradients of usage in one or more areas, and if so, which ones. Consequently, we would like to be able to answer: how do the bike sharing patterns change as a function of time? Are users more or less active on specific days (e.g., holidays or weekends)? Are certain areas more active based on the dates described above?

1.1 Dataset

The data was released on Kaggle, an online platform for predictive modeling and analytics that updates its databases on a near-daily basis. In this case, the dataset was aggregated to answer many different questions and therefore has a very wide range of information, uploaded by DataSF, an organization seeking to answer questions about SF based on data. We will selectively choose the data pertaining to the bike-sharing, named as follows:

- bikeshare_trips.csv - contains information on the start and end stations, date, and ride duration
- bikeshare_stations.csv - contains the respective locations of the bike-sharing stations, represented in latitude and longitude and identified by a unique ID and station name.

1.2 Data Transformations

The data preprocessing will require aggregating the number of trips taken from each station to all other stations per day. Due to the large amount of data this may require us to select a subset of the days in the dataset. Depending on whether users select a single day or a date range to view there may also be a simply processing requirement in javascript. Additionally if we allow the user to select not by day but by hour there may also be additional simple data aggregation requirements.

2 Visualization

Our ideal is to create a map of San Francisco with the stations and create a flow map with the width of each flow corresponding to the amount of traffic that each station gets. If skills and time permit, we would like to have dots moving and demonstrating the flow[2], but would otherwise be differentiated by color (e.g., blue for outgoing and red for incoming).
On top of this, we would like to add an interactive layer that controls the time variable. We would like to have the user control the start and end dates in a way that allows the user to answer the questions that were mentioned above.

[1] https://www.kaggle.com/datasf/san-francisco/data

[2] Inspired by: https://www.nytimes.com/interactive/2015/11/24/upshot/thanksgiving-flight-patterns.html