Jacob Martin
DSBA 6520 Network Science
6/15/2021

San Francisco Bay Bike Share Data Analysis

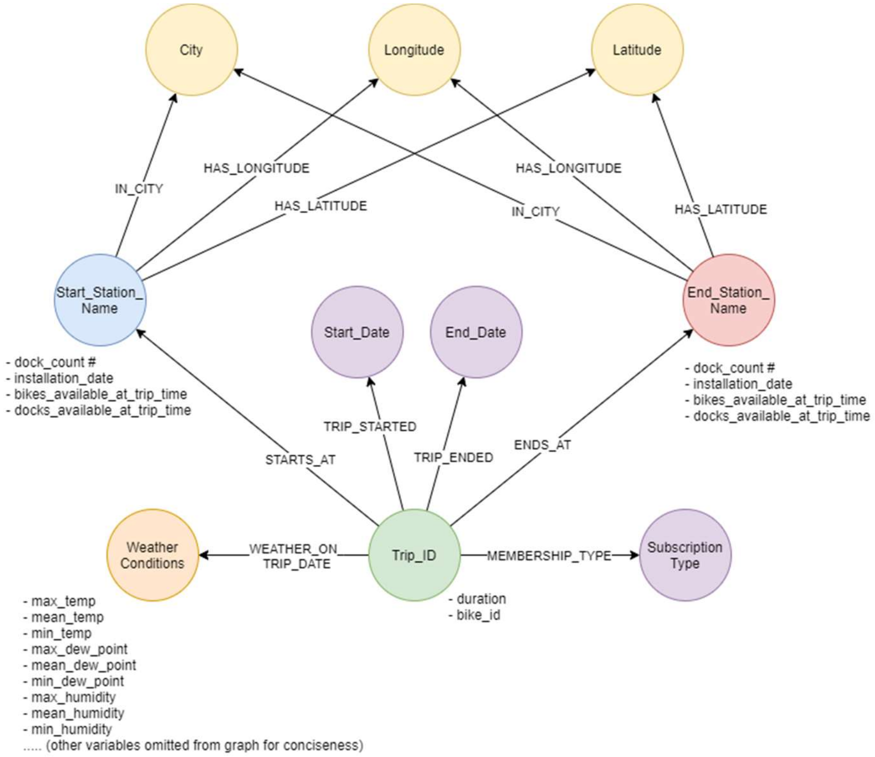
Dataset & Business Use Case

The data set chosen comes from Kaggle.com and is representative of a bike share program in the San Francisco Bay Area. It spans from bike trips from August 2013 until August 2015. The data set includes information pertaining to bike stations, individual bike trips, and weather data in 4 different csv files. Over the period there are a total of 670,000 trips recorded. Each trip has a start and ending station, duration, if the individual was a subscriber, and the bike's id. The weather data consists of temperature, dew point, humidity measures, and other values pertaining to the day's weather in each of the cities.

Our goal is to create an analysis of the data to answer some basic questions about the current ridership. These include understanding what bike stations are the most and least important in the network. It will also be useful to know if all the stations create a single network or if the network is segmented in multiple components. There may be conditions such as rivers or highways which naturally divide the ability to get a bike between stations. Understanding issues like these could aid in developing a more robust network as well as knowing what makes stations better than others. After profiling the data, it shows that 15% of rides are by customers and 85% are by subscribers. We will investigate if certain stations are responsible for a disparity between those types.

Then we will want to create a model that predicts the estimated ride duration based on a series of variables. Creating a model based on duration has utility for understanding the SF Bay Area ridership and ensuring stations will have enough bikes stationed at them to meet demand. The weather data included will be used in the predictive model to help standardize poor weather condition days as we would expect they will adversely affect ridership.

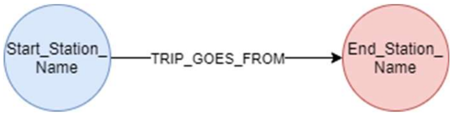
Graph Data Model



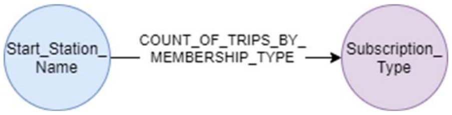
The Graph model is centric around individual trips. Stations share common locations so the end and start stations will be linked to common nodes for latitude, longitude, and city. All the weather conditions will share a common node, which represents the data for that day. Our weather data is not granular to the time of the trip but rather to the day the trip was taken. The weather data is quite extensive so for the sake of conciseness all the properties were not listed on the graph.

Graph Projections

Our first projection from our graph model is to help us learn more about where individuals are going. We have a bipartite projection from the start station to the ending station that is linked through the trip ID node.



As mentioned previously, 85% of trips are taken by subscribers. Understanding what stations they typically start at can aid in understanding our users' habits. This projection goes from the start station to the subscription type based on a count of the trip id field.



Links

Dataset: https://www.kaggle.com/benhamner/sf-bay-area-bike-share

Github: https://github.com/Troxoboxo/SF_Bike_Share