Bay Area Bike Distribution - Project Proposal

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| James Woo V00816927 [jameswoo@uvic.ca](mailto:jameswoo@uvic.ca) | | Ramunas Wierzbicki V00806259 mwierz@uvic.ca | | Jakob Roberts V00484900 jakob.m.roberts@gmail.com | |
|  | Cole McGinn V00780893 cmcginn@uvic.ca | | Jordan Vlieg V00779758 jvlieg@uvic.ca | |  |

***Abstract*—This document is a proposal for data analysis in regards to bike sharing systems.**

# Problem Description & Problem Overview

Biking has become a popular choice for individuals to commute to work or tour around a city. This gives them the exercise they desire and peace of mind in terms of their contribution to cleaner alternatives. Some commuters may choose to purchase a bike that fits their needs. However, they need to maintain the bike and ensure that their bike is parked in a safe location. Additionally, tourists generally do not want to purchase a bike for their trip and instead rely on companies to supply bikes for them to enjoy city amenities. To solve this problem, companies have started Bike Sharing Systems. Commuters no longer have to worry about maintenance and theft, while tourists, through a small fee, can also enjoy the benefits to tour around. The bike sharing systems have become an international business and can be found in cities across Europe, Asia, and America.

Bay Area Bike Share is a public bike share program where riders can rent from 700 bikes in 70 locations. This project focuses on determining how many bikes a location will require. This will ensure that the 700 bikes are distributed appropriately based on usage statistics and expected availability. In turn, the company profits from the optimization and riders are more likely to rent bikes.

# Related Work

There are many bike share systems set up in major cities. Although we are focusing on the San Francisco bike share, there is another major one in Boston that we were also considering. The system in Boston, called Hubway[1], hosted a data visualization challenge that produced a lot of data visualizations for their bike sharing data set. Since they made their data so readily available, some data predictions emerged.

A related data analysis project to our own proposal was created[2]. The original analysis focused around the availability of bikes at certain bike docks depending on their location.This mainly mentioned if the bikes are downtown or uptown. The downtown docks were more full during the day, as workers biked to work downtown. Uptown bike stops were empty during the day as workers took the bikes from these locations. The issue was not having enough bikes uptown during the day, so if a user needed a bike, the docks could be empty. Also, during the day downtown, the docks could be full and the user would have to find another place to store the bike.

The Bay Area Bike Share has so much data available to the public that there are many data predictions based on it. The information is available on Kaggle [3] which spurred a few analyses. These included a time-based exploration that determined if the usage was increasing or decreasing and if there are patterns based on time of day or year. Other analyses include a map of how the bike docks are connected based on how many trips between each dock there are.This also displays popularity of these locations. There is also weather data available which was used to determine which type of weather causes users to bike.

# Data Description

The data for the Bay Area Bike Share System can be found online. There are 3 data sources: 2014 data, 2015 data, and a live feed. Each trip is anonymized and includes the following information [4]:

* Bike number
* Trip start day and time
* Trip end day and time
* Trip start station
* Trip end station
* Rider type – Annual or Casual (24-hour or 3-day member)
* If an annual member trip, it will also include the member’s home zip code
* Daily temperatures
* Daily precipitation
* Wind Speed
* Station location
* Station construction date
* Station bay availability

# Proposed Project

The data for this project was obtained online. The public data was split between Year 1, Year 2, and Live. To format and store the data the team is planning to use Python. The plan for the project is to train and cross-validate based on Year 1 and test on Year 2. When doing an initial sweep of the data, we found that there were instances where there were either very few bikes or empty slots available at a station. This could cause issues when people approach a station to rent a bike to commute but are unable to rent one. This could lead to lost revenue and slower adoption of bike share due to reduced reputation. As for the algorithm the group would like to train, test, and predict using several algorithms such as decision trees, naive Bayes. Through this, the highest accuracy algorithm will be chosen the algorithm which results in the highest accuracy. The focus of this project is to determine the number of bikes a certain location requires based on the day of week and weather results.

We are interested in multivariable regression analysis which would allow us to predict bike needs based on historical data. Written in python, pandas is another possible prediction tool which was designed to forecast based on time series data. This tool will help with the multivariable regression analysis.

# Estimated Timeline

|  |  |
| --- | --- |
| **4-Oct-16** | Proposal Due |
| **21-Oct-16** | Data Compiled & Algorithms Tested |
| **28-Oct-16** | Algorithm Chosen & New data predicted |
| **4-Nov-16** | Presentation Draft Done |
| **8-Nov-16** | Mid-Term Report Due |
| **18-Nov-16** | Predicted Data Compared to Real-Time Feed |
| **23-Nov-16** | Presentation |

# Distribution of Tasks Among Members

|  |  |
| --- | --- |
| **Democratic Project Manager -** | Ramunas Wierzbicki |
| **Executive Data Acquisition -** | Jordan Vlieg |
| **Executive Code Review -** | James Woo |
| **Executive Algorithm Design -** | Cole McGinn |
| **Executive Domain Expert -** | Jakob Roberts |

We assigned each other a main role for us to focus on to ensure that each task is perfectly completed. We plan on tackling and completing each task as a group, but each member will make the final decisions based on their main roles. With leaders ensuring each task is completed to the best of their ability, the team will strive as a whole for a high-quality project.

##### References

[1] Hubway, "Hubway Data Visualization Challenge," in Hubway, Hubway Data Visualization Challenge, 2012. [Online]. Available: http://hubwaydatachallenge.org/. Accessed: Oct. 2, 2016.

[2] Lauren Alexander, Gabriel Goulet-Langlois, Joshua Wolff and C. 109 Hubway, "CS109 Hubway project," in github.io. [Online]. Available: http://cs109hubway.github.io/classp/#analysis. Accessed: Oct. 2, 2016.

[3] Kaggle, "SF Bay Area Bike Share," in Kaggle, 2016. [Online]. Available: https://www.kaggle.com/benhamner/sf-bay-area-bike-share/kernels. Accessed: Oct. 2, 2016.

[4] Bay Area Bike Share, "Introducing bay area bike share, your new regional transit system," in Bay Area Bike Share, 2013. [Online]. Available: http://www.bayareabikeshare.com/open-data. Accessed: Oct. 2, 2016.