Bay Area Bike Distribution - Midterm Report

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 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 the midterm report for data, prediction model, and evaluation analysis in regards to bike sharing systems.**

# Problem Description & Problem Overview

Biking has become a popular choice for individuals to commute to work, tour around a city, or simply get from point A to B. 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 expensive 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. The Bay Area Bike Share system currently faces a challenge: determining which locations will require a certain number of bikes in order to service as many riders as possible. On any given day, the rideability of the day can change. Rideability is defined in terms of weather forecast, events happening in the city, and whether the current time is within tourist season. Poor weather would indicate lower rideability, whereas good, clear weather would indicate higher rideability. If there is an event within the city, the team is assuming that more people are likely to rent bikes, thus increasing rideability. Lastly, it is assumed that if a given time is within the tourist season, more bikes will be rented, thus increasing rideability. Using the rideability attribute, this project focuses on determining how many bikes a location will require in advance and during the business day, or how many empty stations will be needed. First, a prediction can be made to determine how many bikes are required at each location at the start of the day and Bay Area Bike Share can perform the load balancing before bikes are rented out. Additionally, further predictions can be made during peak rental hours and Bay Area Bike Share can perform further load balancing. This will ensure that the available 700 bikes are distributed appropriately based on usage statistics and expected availability. In turn, Bay Area Bike Share profits from the optimization, where a high percentage of the bikes are rented out, and riders are more likely to rent bikes due to increased availability.

# Execution Plan

Currently, Bay Area Bike Share has plans to expand the system. However, existing customers are unhappy with current service. Searching on the Bay Area Bike Share Facebook page, an existing customer has complained, seen in Figure 1 below.



Figure 1: Bay Area Bike Share Complaint [2]

Additionally, the team went through Yelp reviews to determine how many customers are experiencing similar problems. After reading through several reviews, the team discovered that roughly 10% of the 133 reviews contained complaints regarding the issue of availability. An example of the review is shown in Figure 2.

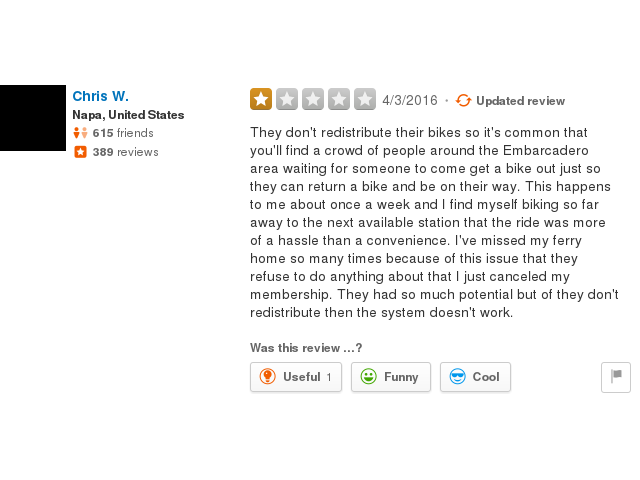


Figure 2: Bay Area Bike Share Yelp Review [3]

Based on real customer feedback, it was determined that the problem defined in this report exists. Data analysis must be done to determine when, where, and why the problem occurs.

Firstly, the team performed an initial analysis from the dataset for the Bay Area Bike Share, which is available online[1]. The project proposal covered this analysis. In summary, the team found that the bikes were distributed in a way that customers could neither dock nor rent a bike at particular locations and times. The initial analysis allowed the team to determine what problems can be solved. One such issue is determining how many bikes a particular station will need before day start and if there is additional need for further load balancing during the day.

Secondly, the team wants to investigate and determine whether a bike station will be full or empty at a specific hour given the rideability at that hour. The goal is to prevent unavailability cases as they prevent customers from docking or renting bikes, respectively.

# Data Analysis

During the initial data analysis for 2016 (year 3 data, 35,517,186 unique entries), it was found that, on average, any station is empty 0.9% (12.96 minutes) of the time and full 0.4% (5.76 minutes) of the time in a day. This information indicated that the program, when viewed as an average, has good service when it comes to usage and ability to dock bikes.

The year 3 dataset contains the following information:

* Station ID
* Bikes Available
* Docks Available
* Time

The team discovered that this dataset is inconsistent with Bay Area Bike Share station information. In the dataset, the station IDs sometimes do not align with the station information. Additionally, there are some stations that do not exist in the station information data, but appear as a station in the dataset. To alleviate this issue, data with an invalid station ID is ignored.

Firstly, the year 3 dataset will be analyzed to determine at which hours of the day stations are empty or full. Using the dataset, it can be predicted which stations will require a certain number of bikes or if docking locations are needed. For example, the team predicts that many commuters will bike downtown to work in the morning. Therefore, it needs to be ensured that the bike stations downtown have available empty docks.

Secondly, the dataset will be analyzed to determine when a given station was either empty or full for more than thirty minutes. This will help determine if there are problematic stations and what problem they might have. Problem stations are defined as having a high number of occurrences of empty or full status over the year, in relation to other stations. In addition, problem stations are characterized as being empty more often than full or vice-versa. For example, it is anticipated that commuters will often find a full station at their workplace or find an empty station when trying to find a bike to get to work.

This midterm report will focus on the percentage of how full or empty a station is. By figuring out at what times a station is completely full or empty, it can be determined when these stations are a problem and what needs to be addressed. Predictions, given the rideability, can be made which can help prevent these issues.

To summarize, the input dataset contains how many bikes are at each bike station at a given minute. Predictions will be made on how full or empty each bike station will be at a specific time in the future. This forecast can help add or remove bikes from problematic stations to ensure there will be bikes available for customers or empty spot for a user to park a bike.

# Results

First, the team was able to generate an infographic of Bay Area Bike Share year 3 dataset. Provided in Appendix A, Figure A1, this infographic describes how many minutes at the given hour any station was empty or full over the whole year. From the graph, it is clear that there are empty and full bike stations at all times of the day. This indicates that prediction work is necessary all the time. There are also very clear spikes which correlate to standard commuting times and these times are good targets to focus the team’s prediction on. Furthermore, it should be noted that stations are empty more often than they are full, which shows that more customers are in transit than bikes are sitting in docks. Stations which are empty very often, but are not empty for very long are of special interest. The team believes that these stations are so in demand that it is beneficial for customers to form makeshift queues to access a bike, since customers are continually returning bikes to the station. To increase customer satisfaction and decrease waiting time, Bay Area Bike Share should monitor these stations more closely, and transport additional bikes to these locations.

Next, another infographic was generated. Found in Appendix A, Figure A2, it shows the number of occurrences a station had either an empty or full station for more than thirty minutes over the year. From the graph, it is clear that there are problem stations which are could greatly benefit from load balancing. To supplement Figure A2, a third infographic was generated, which can be found in Appendix A, Figure A3. The infographic shows the amount of time a station was either an empty or full. According to the two Figures, A2 and A3, there are several stations that are empty much more often than they are full. This indicates that these stations are probably used by commuters, either going or coming back from work, and seldom returned to that station by other customers. Therefore, customers would highly benefit from reallocation of bikes before they need them. Ensuring that stations have sufficient bikes or docks will decrease customer complaints, and increase satisfaction.

The team was also able to predict at which times of the day customers require bikes, using the infographics. Figure A1 shows at which times peak usage occurs. Therefore, the team is confident that performing redistribution at these peak times will greatly improve customer satisfaction.

# Conclusion and Future Work

From the preliminary analysis, it can be seen that there is a problem with bike and dock availability. The results show that several stations are either empty or full for more than thirty minutes unequally, showing that there are problem stations, which do not allow customers to rent or dock their bike. This statement is supported by feedback on Bay Area Bike Share’s Facebook and Yelp pages where customers complain that they are unable to start or end a trips properly.

In the future, this investigation can be refined to account for weather patterns, events, and even other key public transportation facilities which can be chained with the bike share service for a popular commute. Bay Area Bike Share was designed for small time intervals, meaning tourists are disincentivized by high costs when renting the bike for the whole day, but the team believes that an effective prediction system could allow for sufficient load balancing to accommodate seasonal visitors and their trips around San Francisco’s landmarks and festivals. Another possible extension of the analysis would be to determine what areas would most benefit from added infrastructure.

Additionally, the team plans on using data mining techniques to produce predictions a day in advance to classify a station as either Full, Empty, or Neither, to indicate if the station will be full, empty, or neither, given the rideability and time of day. The algorithm will initially be developed with year 1 data, but will be verified and refined using the year 2 and 3 data, as well as the live data stream that is available. Once complete, this prediction system will allow Bay Area Bike Share to redistribute the bikes as necessary to provide customers the necessary bikes or docking locations. The team hopes that, by providing these predictions ahead of time, Bay Area Bike Share will take the responsibility of redistributing the bikes as necessary to provide customers better service.

# 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 |

In the project proposal, team members were each assigned a main role to this ensure that each task is completed properly. The team as a whole worked together to look at the dataset and determine what can be analyzed. Additionally, the report was written together. After the team agreed on what can be inferred from the dataset, Cole and Jordan designed a Python script to search for the data in order to create graphs for visualization of the analysis. For example, one aspect of the code retrieved the number of minutes all stations are empty or full at a given hour. Ramunas worked on creating the visualizations. He also started work on finding the zones of bike stations which will be used in future analysis. Jakob and James worked on determining if there was a problem to solve by searching through several online resources such as Facebook and Yelp and searching through the dataset to see if there are problem stations. Overall, the team distributed tasks equitably and had no issues working together.

##### References

[1] 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.

[2] Facebook, [Online]. Available: <https://www.facebook.com/BayAreaBikeShare/posts/695791523918193> Accessed: Nov 1, 2016

[3] Yelp, “Recommended Reviews for Bike Area Bike Share”. [Online]. Available: <https://www.yelp.com/biz/bay-area-bikeshare-san-francisco?hrid=-3RMMdnC1aVPX3s9-1RqHQ&utm_campaign=www_review_share_popup&utm_medium=copy_link&utm_source=(direct)> Accessed: Nov 1, 2016

Appendix A: Figures

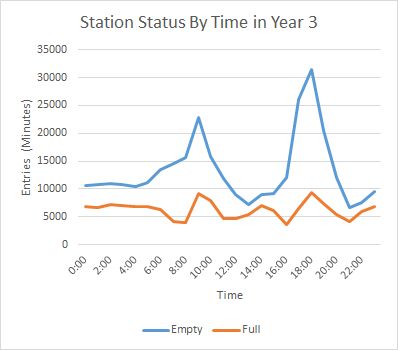


Figure A1: Station Status By Time in Year 3

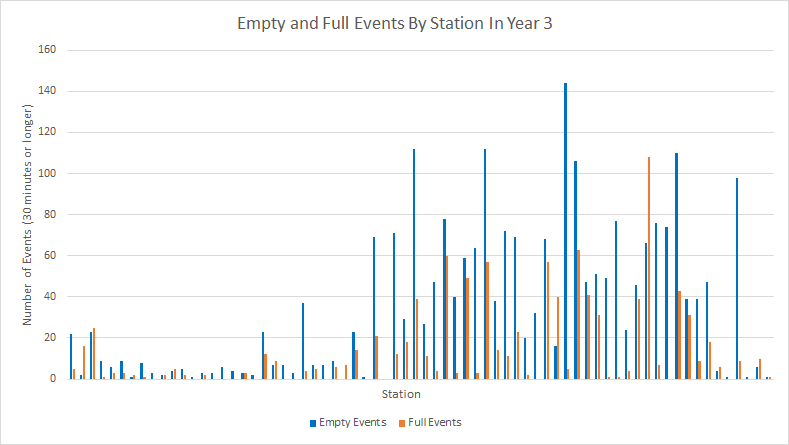


Figure A2: Empty and Full Events By Station, for Year 3

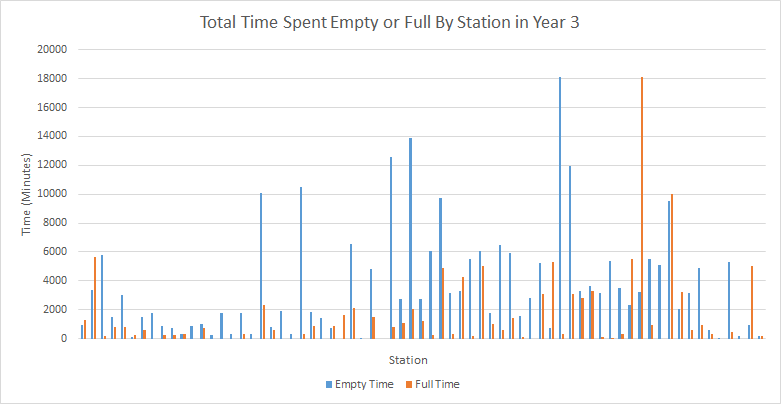


Figure A3: Total Time Spent Empty or Full By Station, for Year 3