The Future of Cycling in Seattle:   
A Data Science Approach to Ridership Forecasting

Group 2 Team Members:

Justus Ngunjiri, Arthur Setiawan, Yen-Ting Chen, Robert Barrentine

DATA 5100 | Dr. Fischer | FQ22

12 October 2022

Problem Statement

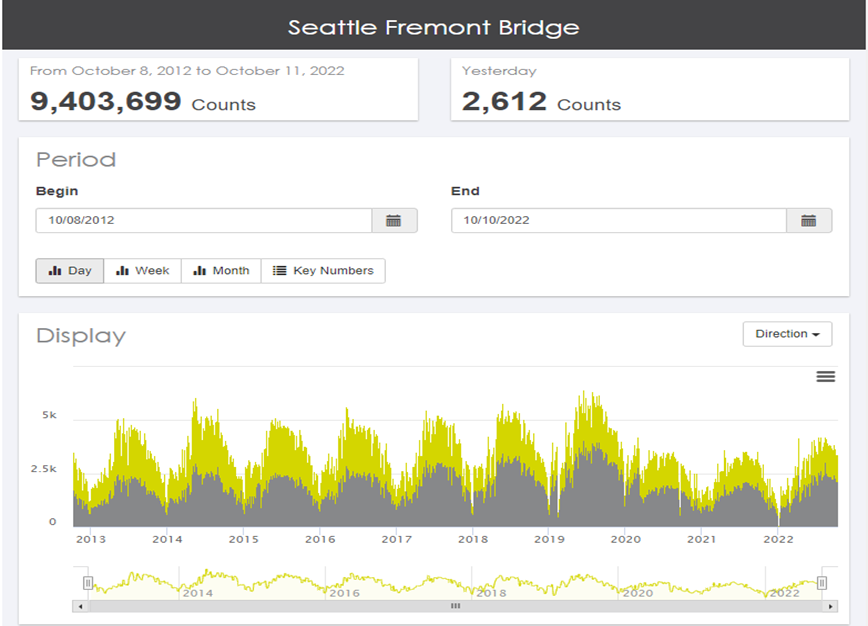
Cycling is a big deal in Seattle. According to a 2019 study by Nielsen for the Seattle Times, “an estimated 169,000 adults in the Seattle area use a bicycle for transportation.” (Balk, 2019) That number is expected to grow considerably in the coming years as Seattle continues to attract transplants to fill new jobs. The Seattle Department of Transportation (SDOT) is betting on bicycles to adapt for a denser Seattle. Since 2014, SDOT has been publishing regular updates on their Seattle Bicycle Master Plan (BMP) which aims to accommodate growth and associated mobility needs with investments in bicycle infrastructure. (BMP, 2021) In order to assess the impact of the BMP initiatives, SDOT has installed 12 bike counters throughout the Seattle area on trails and roadways. (Bike counters, n.d.) This data is made publicly available at data.seattle.gov and is granular to the hour.

The proposal of this project is to utilize this data to answer a series of questions related to the Seattle cyclist population and the capacity of existing and future infrastructure. If the Seattle cyclist population is expected to grow over the next several years, is it possible to accurately predict the growth using statistical and machine learning models? The seasonality of ridership throughout the year is an interesting aspect to this dataset. Is it also possible to accurately model seasonal swings in a dataset? Using these models, it should then be possible to estimate the peak volume of cyclists in a given day. Can this figure be used to inform the design decisions of future infrastructure? Also, are there safety implications for an underdesigned bike path?

Data Source(s)

This project will utilize data extracted from the Seattle Department of Transportation (SDOT) database. SDOT’s database is online with an open access policy across various transportation data categories. They include, parking, traffic flow maps, planning and maintenance, and transit and commuting. Through SDOT, most of these datasets, depending on the form (historical or live), can also be accessed through the city’s integrated geodata host site. This host is called City of Seattle GIS. The datasets are collected through the city’s geographical information system programs where the data is captured live and documented for future reference. One such program is the maintenance of Bicycle and Pedestrian counts across three main bridges namely: Fremont Bridge, Spokane Street and 2nd Avenue. Depending on the location and development interests, some of the biker and pedestrian counter sites capture either just bike data, or both bike and pedestrian data.

Given that our interest is the Fremont Bridge, we will navigate to the database and extract the Fremont Bridge program bike data. This program maintains data for bikers only, thus filtering to exclude pedestrian data will not be necessary in our project. Our focus data format is maintained on daily, weekly, monthly and annual total displays (Seattle government, 2022). The program was started in 2012, thus we have entries from the same year to the current live time. Figure 1 shows a sample display of the data. Depending on our desired prediction intervals, we might have to aggregate our data to desired periods. We consider the data to be sufficient based on the diversity and the number of entries contained in the database.



*Figure 1*. Bicycle count data for the Seattle Fremont Bridge from [www.data.seattle.gov](http://www.data.seattle.gov)

Analytical Approach

Upon viewing the Fremont Bridge Bicycle Counter dataset, we can determine that this is a bivariate time-series dataset where we hope to not only take into consideration seasonality into our predictive modeling, but also identify and isolate other components such as trends, cyclical fluctuations, and noise. As we are approaching a world that increasingly runs on energy independent of fossil fuels, understanding these aspects will ultimately give actionable insights to not only SDOT for potential areas of improvement and hazards through construction projects for cyclists, but also for small, local business owners to preemptively identify and execute prospective business opportunities (e.g., bicycle service shops).

We speculate in using supervised learning predictive modeling for this project as we believe it to be a comprehensive way to try out different methods of regression analyses and help us arrive at the strength of relationships between our different variables, as well as which would have more statistical significance in predicting. This will help us arrive at a different confidence interval for each different regression model we use and from there we can determine which model would match our dataset the best. An example would be to first use linear regression to identify the independency of our east sidewalk count and west sidewalk count, and from there if exists a high correlation can use ridge regression depending on collinearity of the data. Alternatively we may use predictive analysis using Neural Networks, Decision Trees, Naive Bayes, and Nearest Neighbor of which we have to determine the complexity of given our somewhat limited timeframe.

Solution Technologies

This project asks us to look deeper into data analysis that involves taking a step toward in depth processing of data. The list below has provided all the necessary toolkits for us to work on the tasks we are supposed to carry out. Besides the data wrangling, maybe it’s possible to look at training with sparse data. Scipy has csr\_matrix to give us the ability to work with matrices containing many zeros (categorical data that’s not ordinal may become sparse).

The two main languages in data analysis are Python and R. In data analysis, we use statistics and machine learning to perform predictive modeling and analysis on datasets, which will aid us in building the underlying framework of solving data analytics. Each package is described by the functionalities proven to work sufficiently to handle complex data problems.

Packages that may be needed depending on the language we are using:

R: **ggplot2/ggvis** – ggplot and ggvis both allow users to make charts and graph visuals that are descriptive and comprehensive in terms of how it is utilized. Compared with Python’s matplotlib, the two packages are very similar except ggplot is slimmer in terms of code amount. **Dplyr** - For data exploration, data munging, and wrangling, users can use dplyr as the grammar for manipulating with data in various ways. **Data.table** - much like Dplyr, the data.table package is often used to manipulate tabular data. **Caret** - predictive modeling and ML in R. **Reshape2** - transform data between wide and long formats.

Python: **Seaborn** – To create graphs that can allow the user to visualize data better during the process of data analysis.**Pandas** – In Python, pandas adds on top of numpy/scipy to enhance the data representation in a more comprehensive way. It includes more data structures in order for users to manipulate data with more agility. **Numpy/Scipy** – Numerical and scientific programming can easily be done with the two packages. **Scikit-learn** – ML predictive modeling. The users need to understand the math before using the package so they can work with the parameter inputs. **XGBoost** - tree-based methods that are used a lot in predictive modeling. XGBoost implements the gradient boosting that is able to learn from large datasets very quickly and with regularized model output.

Challenge

We are interested in the seasonality of the data. We will not be fitting a simple regression to the model to determine the ‘average’ trend of cyclists over time. Instead, we want to utilize a series of Machine Learning algorithms to the data to understand the strengths and limitations of each and determine a best-fit solution. It would also be interesting to expand the scope of our project to look at other bike and pedestrian meters across the Seattle area to see if the trends are similar/comparable for the different locations.

Citations

Balk, G. (2019, June 5). *Who bikes in Seattle? Bicycling gender gap one of the biggest in the U.S.* Retrieved from Seattle Times: <https://www.seattletimes.com/seattle-news/data/who-bikes-in-seattle-bicycling-gender-gap-one-of-biggest-in-country/>

Bicycle master plan, BMP. (2021, May). Retrieved October 11, 2022, from <https://www.seattle.gov/transportation/document-library/citywide-plans/modal-plans/bicycle-master-plan>

Bike counters. (n.d.). Retrieved October 11, 2022, from <https://www.seattle.gov/transportation/projects-and-programs/programs/bike-program/bike-counters>

https://data-seattlecitygis.opendata.arcgis.com/search?tags=transportation

<https://www.seattle.gov/transportation/projects-and-programs/programs/bike-program>

https://www.seattle.gov/transportation/projects-and-programs/programs/bike-program/bike-counters/fremont-bike-counters

Group Dynamics

Team members agreed that communication with only be via Canvas and Microsoft Teams. Microsoft teams registered using Seattle U alias email addresses will be used for all meetings while Canvas will be used to send out complimentary communication as may be required. In cases of delayed feedback from member where the information is required urgently and or impedes the progress of another member, a normal call to the respective member will be used to expedite the response. For that reason, all the group members exchanged contacts during the first meeting.

As a group, it was agreed that every group member will create a github account and use it to share the relevant codes required for project progress. Team members agreed that there will be weekly progress review and meetings will be scheduled accordingly. To incorporate everyone, all the meetings will be held from 7:30 pm onwards. However, if the need for a meeting arises and it happens that all group members have class on that particular day, relevant meetings will be scheduled at most 1 hour before the classes start.