Project Report

(Presented by: Apoorva Patil, Brooke Hauserman, Divi Joshi, Hetal Gala)

Topic:

Bluebikes is a bike-sharing company that is in and around Boston, Massachusetts. The company provides affordable and accessible public transportation while also promoting a healthy alternative to quickly travel around the city. Recently, Bluebikes was sponsored by Blue Cross Blue Shield to support their initiatives and to help increase the accessibility to healthcare facilities ("Blue Cross Blue Shield"). An important aspect of the business is their transparency of their trip history data and their excitement to allow any individual to utilize their data to develop visualization and conduct analysis. Due to the vast amount of data available, our group was immediately interested in taking on the challenge to understand the data patterns of the company from 2019 to 2022. After exploring the data sources provided off the Bluebikes website, our group decided to focus on three main business topics. These topics include seasonality impact and trends in bike share demand, the gap between the availability of dock stations and the demand of riders, and the installation of more docks at more popular locations or the reduction of docks in places where there is little demand. These topics were chosen so that our group would be able to look at trends in demand and how weather and other factors can impact the usage of Bluebikes. Also, our group wanted to find the most popular dock stations and use the analysis to make recommendations on how to improve upon their operational expansion. The company, Bluebikes, will be interested in the topics that our group chose because it will give them better insight into the demand of their product across different locations, and it will show trends that occur across each season. The results of the analysis are also for Bluebikes because our analysis shows an in-depth understanding of their consumer base as well as opportunities that the company can exploit in order to better position themselves for their customers.

Data Sources:

We have used 3 datasets - Trip History, Dock Stations and Weather and they are all from reputable sources. In our analysis, the level of data granularity is fine.

- Bluebikes system trip history data from 2019 to 2022 from Bluebikes website (https://www.bluebikes.com/system-data) ("Bluebikes System Data.")
- Bluebikes current dock station locations from Bluebikes website (https://www.bluebikes.com/system-data) ("Bluebikes System Data.")
- Historical weather data in Boston from the National Weather Service(https://www.weather.gov/wrh/climate?wfo=box) ("NOAA's National Weather Service")

Unit of Analysis

- Bike Id
- Dock Stations

Data Collection:

- Scrapped all URLs of the zip files off of the Bluebikes system website and downloaded it onto our local system using python (since every month had a separate zip file)
- Unzipped files and extracted trip history dataset (csv) for each month using python (instead of doing it manually) onto our local system
- Collected current dock station location data from Bluebikes website
- Collected monthly average weather data from 2019 to 2022 and processed it on excel.

Trip History :

The dataset has 16 variables - 'tripduration', 'starttime', 'stoptime', 'start station id', 'start station name', 'start station latitude', 'start station longitude', 'end station id', 'end station name', 'end station latitude', 'end station longitude', 'bikeid', 'usertype'

tripduration	starttime	stoptime	start station id	start station name	start station latitude	start station longitude	end station id	end station name	end station latitude	end station longitude	bikeid	usertype
2454	00:27.8	41:22.3	345	Park Plaza at Charles	42.35182807	-71.06781138	235	East Somerville Lib	42.38762811	-71.08318716	7414	Subscriber
744	00:38.5	13:03.2	141	Kendall Street	42.36356016	-71.08216792	329	Washington St at N	42.3817508	-71.0839523	5815	Customer
439	01:25.4	08:44.6	68	Central Square at Ma	42.36507	-71.1031	381	Inman Square at Sp	42.37426714	-71.10026479	3301	Subscriber
533	01:39.0	10:32.2	78	Union Square - Some	42.3796479	-71.09540476	381	Inman Square at Sp	42.37426714	-71.10026479	4119	Subscriber
680831	02:04.0	09:15.4	22	South Station - 700 A	42.352175	-71.055547	113	Andrew T Stop - Do	42.33073333	-71.05699851	4467	Customer
530	02:36.7	11:26.9	68	Central Square at Ma	42.36507	-71.1031	471	MIT Carleton St at A	42.36054174	-71.08669817	6856	Subscriber
537	02:38.6	11:35.8	179	MIT Vassar St	42.35560121	-71.10394478	189	Kendall T	42.36242784	-71.08495474	4205	Subscriber
801	02:45.7	16:06.8	67	MIT at Mass Ave / Au	42.3581	-71.093198	41	Packard's Corner - C	42.352261	-71.123831	7690	Subscriber
1001	03:37.3	20:18.7	17	Soldiers Field Park -	42.36426344	-71.1182757	5	Northeastern Unive	42.341814	-71.090179	5202	Subscriber
2559	03:38.5	46:17.9	187	Cypress St at Clark Pl	42.32784317	-71.12536222	549	Valenti Way at Hav	42.36473899	-71.05934903	4665	Subscriber
334	03:43.5	09:18.0	87	Harvard University He	42.366621	-71.114214	110	Harvard University	42.376369	-71.114025	3388	Subscriber
705	03:46.6	15:32.1	141	Kendall Street	42.36356016	-71.08216792	184	Sidney Research Ca	42.35775309	-71.10393405	4397	Subscriber
1123	03:49.3	22:33.0	554	Forsyth St at Hunting	42.339202	-71.090511	554	Forsyth St at Huntir	42.339202	-71.090511	5966	Subscriber
784	03:54.3	16:58.5	329	Washington St at My	42.3817508	-71.0839523	377	Perry Park	42.37927325	-71.10341903	7415	Subscriber
1118	03:56.8	22:35.0	554	Forsyth St at Hunting	42.339202	-71.090511	554	Forsyth St at Huntir	42.339202	-71.090511	4869	Subscriber
1076	04:34.3	22:30.9	554	Forsyth St at Hunting	42.339202	-71.090511	554	Forsyth St at Huntir	42.339202	-71.090511	5292	Subscriber
180	04:51.2	07:52.1	558	St. Alphonsus St at T	42.33329255	-71.10124648	27	Roxbury Crossing T	42.331184	-71.095171	5276	Subscriber
2180	04:56.4	41:16.5	554	Forsyth St at Hunting	42.339202	-71.090511	359	One Brigham Circle	42.3339227	-71.10446509	5625	Subscriber
774	05:03.2	17:58.1	163	The Lawn on D	42.344792	-71.044024	121	W Broadway at Dor	42.33595898	-71.046229	6454	Subscriber
1014	05:08.4	22:03.3	5	Northeastern Univers	42.341814	-71.090179	553	Cambridge Crossing	42.371141	-71.076198	4443	Subscriber
1604	05:31.7	32:16.5	145	Rindge Avenue - O'N	42.392766	-71.129042	145	Rindge Avenue - O'	42.392766	-71.129042	2407	Subscriber

Docking Stations :

The dataset has 8 variables - ‘Number’, ‘Name, ‘Latitude’, ‘Longitude’, ‘District’, ‘Public’, ‘Total docks’, ‘Deployment Year’

	Number	Name	Latitude	Longitude	District	Public	Total docks	Deployment Year
0	K32015	1200 Beacon St	42.344149	-71.114674	Brookline	Yes	15	2021.0
1	W32006	160 Arsenal	42.364664	-71.175694	Watertown	Yes	11	2021.0
2	A32019	175 N Harvard St	42.363796	-71.129164	Boston	Yes	18	2014.0
3	S32035	191 Beacon St	42.380323	-71.108786	Somerville	Yes	19	2018.0
4	C32094	2 Hummingbird Lane at Olmsted Green	42.288870	-71.095003	Boston	Yes	17	2020.0
...	...	...	...	...	...	...	...	...
440	N32005	West Newton	42.349601	-71.226275	Newton	Yes	15	2020.0
441	A32043	Western Ave at Richardson St	42.361787	-71.143931	Boston	Yes	19	2019.0
442	B32059	Whittier St Health Center	42.332863	-71.092189	Boston	Yes	19	2019.0
443	D32040	Williams St at Washington St	42.306539	-71.107669	Boston	Yes	23	2018.0
444	S32005	Wilson Square	42.385676	-71.114121	Somerville	Yes	15	2012.0

Weather Data :

It has 5 variables - ‘Year’, ‘2019’, ‘2020’, ‘2021’, ‘2022’

	YEAR	2019	2020	2021	2022
0	JANUARY	30.7	38.0	31.0	27.4
1	FEBRUARY	33.5	37.8	30.8	33.1
2	MARCH	39.1	41.8	42.0	41.4
3	APRIL	51.7	44.6	50.8	50.1
4	MAY	57.8	56.8	61.4	60.6
5	JUNE	68.2	69.2	74.4	67.9
6	JULY	78.7	75.3	72.4	77.5
7	AUGUST	74.1	74.1	76.9	76.7
8	SEPTEMBER	68.0	65.6	69.7	65.2
9	OCTOBER	57.4	54.7	59.9	56.6
10	NOVEMBER	42.8	48.0	44.7	49.4
11	DECEMBER	37.2	36.0	39.2	NaN

As the data is from reputed sources, we did not have any data quality issues. The available data on the website was in a clean and usable format.

- In the Trip History dataset, there were a few variables like gender, birth year, and postal code, which we had to exclude from our analysis as the data was not available for the entire time-frame. We tried exploring multiple resources to find this data but were not able to.
- Membership data was not available for any detailed analysis. We did reach out to the Bluebikes team on different social media platforms asking for detailed analysis, but unfortunately we did not hear back from them.
- For the weather data, we were not able to find snowfall/precipitation data and hence we had to use monthly average temperature data. The available data from the National Weather website required some processing which we performed on excel.

Methods and tools:

The entire project was carried out in Python. Using python code, forty-five zip folders containing CSV files from January 2019 to September 2022 were downloaded from Bluebikes' website to the local system. These folders were then unzipped in order to access the CSV files they contained, which were then loaded as data frames for additional transformations and analysis. We read through some external links ("Time Series Analysis.", "Complete Guide on Time Series Analysis in Python.", and "Understanding Time Series Analysis in Python.") to understand the

conceptual working of those models before implementing them on our dataset and on how to interpret our results before conducting Time Series Analysis on the dataset to find any seasonal patterns and demand trends over the years. We used the AutoTS package for forecasting demand for the next six months.

Data Wrangling Process

Data preprocessing:

After we extracted CSV files from all the zip folders, we loaded them into individual data frames and included a year and month column based on their filename. After this, these data frames were transformed again by merging them into a single data frame for further profiling. We observed that the data frame was clean and there were no missing, null or extreme values or outliers.

To detect seasonality patterns, we went ahead and transformed the previous data frame by aggregating it with respect to user type, year and month and plotted the resulting demand values to find trends. We also forecasted demand for both the user types by using autos package for which we had to add a period column to pass as a parameter. The forecasted demand values for each user type were then stored in different data frames, transformed as desired and combined again for plotting.

Similarly, we aggregated the docking stations dataset to find demand between different docking stations and to figure out areas where most of the demand was originating.

Data Enrichment:

The data enrichment process is an important step when data wrangling a dataset. In order to enrich the ride history data, we added weather data for each month from 2019 to 2022. The weather data is being utilized to further understand the impact of temperature (in Fahrenheit) on the use of Bluebikes. The weather dataset improves our data because it helps to understand the trend in demand especially in the colder winter months where usage is the lowest. The first step in enriching the ride history data was to collect the weather data from the National Weather Service. The National Weather Service utilizes “temperature sensor siting” to ensure the most up-to-date and consistent weather data in regard to calculating temperature (“Site and Exposure Standards”). After collecting the historical weather data, the average temperature for each month was transposed into an excel file, saved as a csv, and imported into Python. As part of the data

transformation process, the month column of the weather data frame was changed from an integer to a string in order to merge two data frames more easily. From here, our group merged the weather data frame with the ride history data frame to be able to see the relationship between the demand of rides and the temperature. Based on the merged file, a graph was developed to demonstrate the relationship between temperature and the demand for bike rides. From this visualization, it was analyzed that as temperature decreases the number of riders decreases, and as temperature increases the number of riders increases.

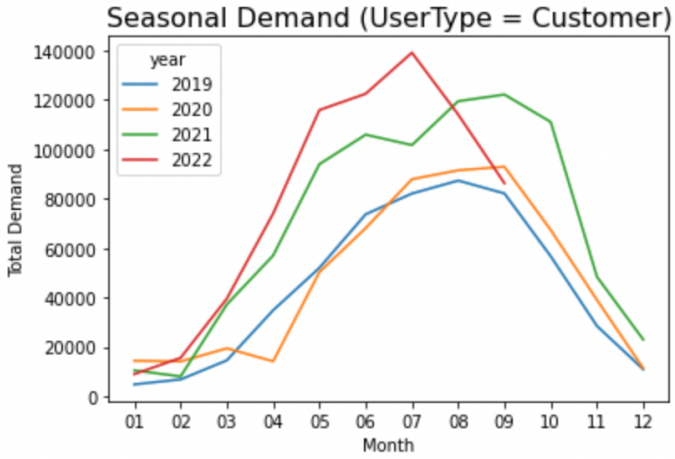
Data Validation:

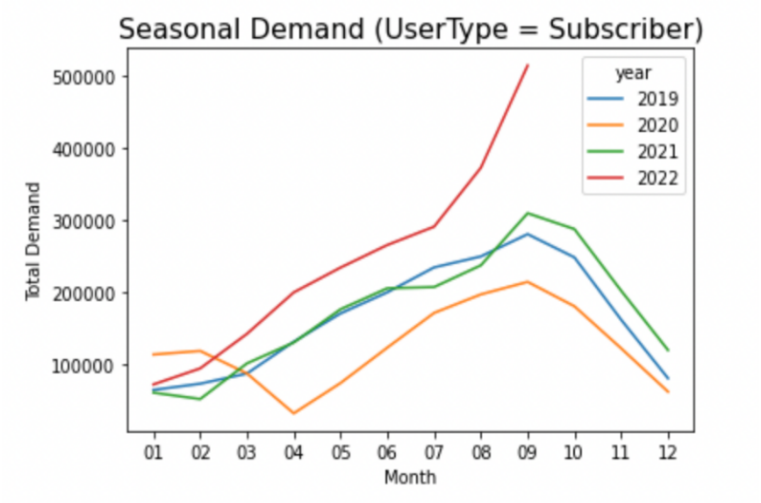
An important aspect of the data enrichment process is the development of data validation rules and checks. One validation check that was implemented into our data wrangling process was ensuring that each file was formatted into a CSV. This was a business rule check that our group implemented when adding each data file into Python. An example of the benefits of this check occurred when pulling the weather data. Originally, the weather dataset was in PDF form, and it was necessary that the data was converted into a CSV file. Another data validation check that our group put in place was to check for consistent data types when enriching the data. This was a system rule check put in place which created a more simplified process when merging two datasets. For example, before merging the ride history data frame with the weather data frame, the year and month of the weather data set needed to be converted from an integer to a string. This check helped our group to seamlessly combine the data and made it easier to be able to develop an analysis. Lastly, a system rule check was developed to ensure that Bike ID, Start Times, and Start Station could not be blank in the trip history dataset. This is because these three columns were vital to our analysis, and it was important that our team had consistent data.

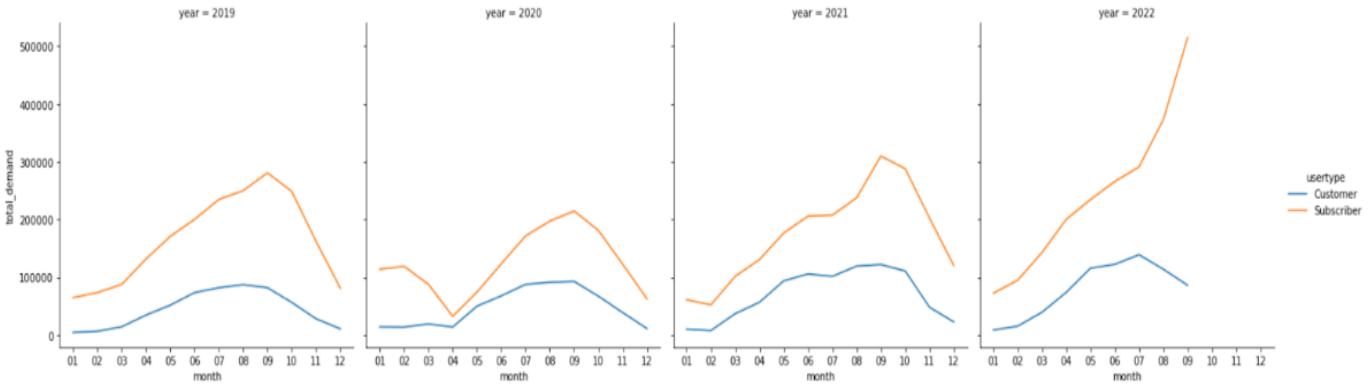
Analysis and Results:

When we compared the weather data with Bluebikes' system data, we noted that the demand for bikes decreased after September. The demand started increasing in the summer and peaked during the fall, but it declined as winter approached. Since the winters in Boston are harsh, people would rather use public transportation or their private vehicles to get to their destination. However, this year, due to a spike in student intake around Boston after COVID and the city's announcement of free 30-day Bluebikes passes due to the Orange Line's temporary shutdown ("McCourt, Clara"), the demand for bikes increased. Consequently, the number of subscribers increased substantially compared to the number of customers.

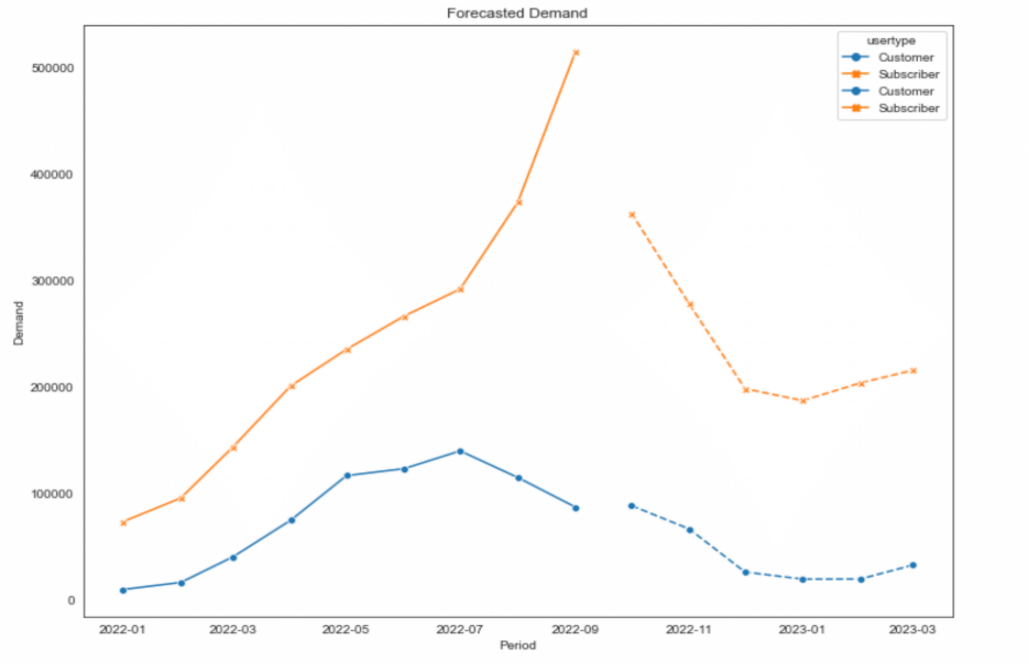
In addition, when we compared the data of Bluebikes' subscribers and customers from 2019 to 2022, we found that the demand for bikes was higher before COVID, plummeted during the COVID lockdown, and resurged after restrictions were lifted.



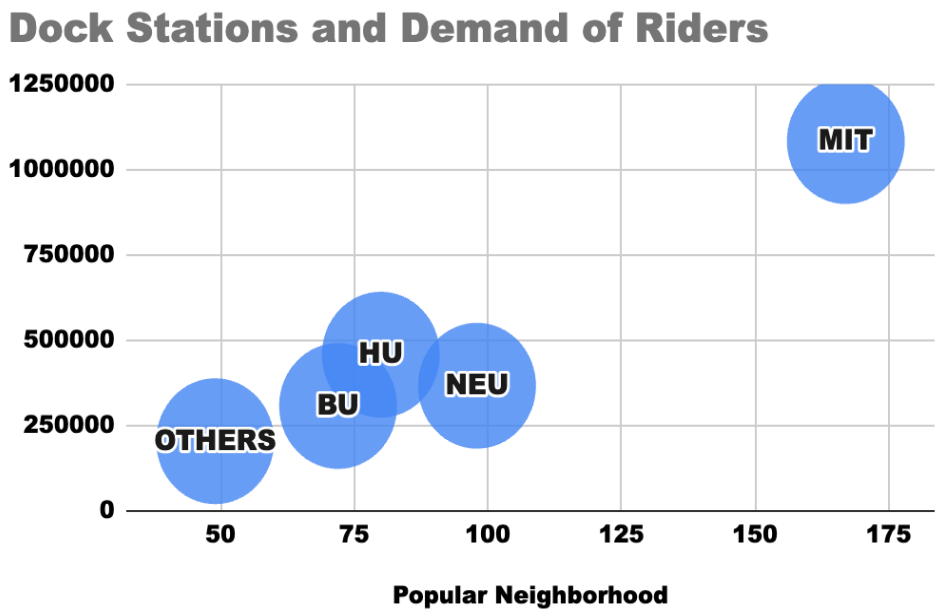




Below is a graph showing the current and projected demand for customers and subscribers. Client demand is expected to decline during the winter months, but to increase during the summer.



We then extracted the data for the top 25 docking stations that had the highest demand. We deduced that these 25 docking stations represent about 30% of the overall demand because they are located less than 0.5 miles from a university.



Challenges:

Lack of access to membership data

When we did our exploratory analysis, we noticed that we did not have access to the membership data for Bluebikes. Hence, we tried to contact their team on multiple platforms; however, they did not get back to us. As a result, we were unable to include the membership data in our analysis.

Lack of demographic data for each rider

A lack of detailed information was available about each rider. Due to the fact that birth year and gender were not available for the timeframe we were looking at, we were not able to use them in our analysis.

Lack of quantitative data for snowfall and precipitation for the weather data set

We included only the temperature data as no quantitative data for snowfall or precipitation for the weather data set were available.

Future Scope and Analysis:

- Given enough time, we could have provided recommendations for the current pricing system based on the type of passes provided by Bluebikes which would give us information about whether a ride membership, single trip, adventure, etc.
- We could have adopted machine learning techniques and algorithms to get more precise results for our analysis.

Works Cited

"Blue Bikes Bike Sharing Program." *Northeastern University*, https://www.northeastern.edu/commutingservices/bicycling/discounted-blue-bike-sharing-program/.

- Bluebikes offers discounts to all students who attend Northeastern University. Using the discount code "Husky," students can now obtain a yearly membership for $90 rather than the full price of $119.

"Blue Cross Blue Shield of Massachusetts and the City of Boston Announce Bluebikes Expansion at Community Health Centers." *Blue Cross Blue Shield*, 18 Sept. 2019, https://www.bcbs.com/press-releases/blue-cross-blue-shield-of-massachusetts-and-the-city-of-boston-announce-bluebikes.

- Blue Cross Blue Shield of Massachusetts donated $10,000 to help increase accessibility to medical facilities. Bluebikes is able to use this donation to build more bikes stations in locations more convenient for those looking to receive medical treatment.

"Complete Guide on Time Series Analysis in Python." *Kaggle*, Kaggle, 30 Aug. 2020, https://www.kaggle.com/code/prashant111/complete-guide-on-time-series-analysis-in-python.

- When developing our python file, we wanted to learn how to utilize Time Series Analysis for our project. The following website has a detailed description of Time Series and how to implement it into Python.

Godwin, James Andrew. "Time Series Analysis." Medium, Towards Data Science, 13 Sept. 2022, https://towardsdatascience.com/time-series-analysis-7138ec68754a.

- Similar to the source above, our group wanted to use time series in our analysis and this resource was very beneficial when developing our code. This source describes the basic topics along with examples on how to use it. This helped us implement Time Series into our analysis.

"Introducing Lyft Bikes: Lyft Bikes." *Lyft Bikes-Lyft Blog*, https://www.lyft.com/blog/posts/lyft-to-acquire-us-bikeshare-leader.

- As of 2018, Bluebikes became a part of the company Lyft. This occurred after Lyft acquired Motivate, which is a bike sharing organization in North America. This is part of an initiative to help increase access to more sustainable transportation methods.

Madhugiri, Devashree. "5 Python Libraries for Time-Series Analysis." *Analytics Vidhya*, 22 Sept. 2022, https://www.analyticsvidhya.com/blog/2022/05/5-python-libraries-for-time-series-analysis/

- We used one of the packages included in this article to help with our Time Series Analysis.

McCourt, Clara. "Bluebikes Rentals Surge throughout Orange Line Shutdown." *Boston.com*, The Boston Globe, 19 Sept. 2022, https://www.boston.com/news/local-news/2022/09/19/bluebikes-ridership-records-orange-line-shutdown/.

- The Orange Line shut down due to maintenance in August of 2022, and the city of Boston offered a free 30-day subscription to Bluebikes as an alternative.

Motivate International, Inc. "Bluebikes System Data." *Blue Bikes Boston*, https://www.bluebikes.com/system-data.

- This source is where our group obtained all of our data on ride history and dock location for Bluebike. Bluebikes offers their data to be used for visualizations as well as analysis.

"Site and Exposure Standards." *Site and Exposure Standards*, NOAA's National Weather Service, 28 Nov. 2016, https://www.weather.gov/coop/sitingpolicy2.

- The National Weather Service provides details on how they are able to calculate their weather. This source also provides information on the tools and techniques that they practice to gain consistent results.

"Understanding Time Series Analysis in Python." *Simplilearn.com*, Simplilearn, 15 Sept. 2021, https://www.simplilearn.com/tutorials/python-tutorial/time-series-analysis-in-python.

- In order to implement Time Series Analysis into our python code, we used this source to help with our understand of how to interpret our results.

US Department of Commerce, NOAA. *Climate*, "NOAA's National Weather Service", 3 Mar. 2022, https://www.weather.gov/wrh/climate?wfo=box.

- The National Weather Service provides up-to-date data on weather from around the United States. Our group utilized this source to collect the weather history of Boston from 2019-2022. This data was then used to enrich our Bluebikes ride history data.