Visualizing CitiBike Data

Anna Herlihy, Eleni Zapridou, Hamish Nicholson, Yugesh Kothari
Data Visualization, Milestone 3

1 Introduction

In this project, we create a visualization tool to enable any curious individual to explore the Citi Bike network.

Citi Bike is a privately owned public bike-sharing system serving New York City. It offers a convenient and sustainable mode of transportation for residents and visitors alike. Citi Bike is a dock-based bike-shared scheme. The dataset contains information about trip histories, but is an overwhelming amount of data in text form. (The data is distributed in CSV format)

Our objective is to develop a data visualization tool that enables users to analyze Citi Bike data according to their preferred time period and region. Using a map, users are able to select a specific area of New York and set a filter for the desired time period. Based on this input, the tool generates and plots statistics, such as the total number of rides and the popularity of each station in the selected region.

Our tool is designed for anyone who is interested in exploring Citi Bike data. It caters to a wide range of users, from existing Citi Bike users and potential new subscribers to data analysts. For instance, analysts working for a museum in New York may use the tool to determine the number of rides that start or end near the museum in order to decide whether they should provide a discount code for Citi Bike with a purchase of a museum pass. Additionally, people considering subscribing to Citi Bike might use the tool to find out the availability of bikes in stations close to their homes at specific hours. The project can also be used as a tool for social scientific analysis. By identifying bike-sharing patterns in different neighborhoods, we can identify how socioeconomic factors may affect patterns of bike-sharing.

We believe that our project and tool can be useful resource used both casually and in more professional circumstances.

2 Design Process

During the initial idea and planning phase, our team was motivated by our collective interest in public transit and data visualization. We recognized that while public transit visualizations have often been explored, bike networks present a more amorphous and intriguing challenge. Unlike traditional modes of public transportation, such as metro or bus systems, bike networks lack fixed routes and specific destinations. For example, the NYC Subway has 424 stations organized into 25 lines, but the Citibike network has over 1600 stations, and users can travel between any pair of stations. This means one cannot just look at a map of the stations and understand the flow of people. This challenge sparked our curiosity and drove us to explore the Citi Bike data, as it offered a unique opportunity to delve into the dynamics of bike-sharing systems and uncover hidden patterns and insights.

In determining what we wanted to visualize with the data, we also researched prior uses of the data. Citi Bike data has been extensively analyzed in various studies, such as those that have examined the demand [4, 7], commute time [1], gender gap generators [6], the impact of weather [3], and Covid-19 [5, 2] on bike sharing. While these studies have yielded interesting insights, they typically focus on specific time periods or events. Our goal is to provide a visualization tool that enables users to select the time period and area they wish to analyze. This approach allows users to explore precisely what interests them. We believe that such a tool will be valuable for users who want to perform a more personalized analysis of Citi Bike data.

Initial sketches: Our initial primary idea for the visualization was a map of New York that displays each Citi Bike ride's starting and ending stations as data points. We wanted to represent each ride with a line connecting its respective start and end stations. When users zoom out data points in close proximity will merge into a single larger data point. Conversely, zooming in will expand larger data points into individual ones. We wanted users to have the ability to select a time period by specifying the start and end dates. Whenever users adjust the time period, the information on the map will be updated automatically to reflect the filtered period. Additionally, users can select a specific region on the map using a drag-and-drop feature.

Whenever the user updates the selected time period and/or area of interest, our tool will automatically update several plots that display statistics for the Citi Bike rides within the filtered parameters:

- **Plot 1: Average number of rides per hour.** This plot will present the average number of rides per hour. A sketch of the plot is shown in Figure 1.
- **Plot 2: Popularity of stations.** This plot will present the average number of rides starting or ending at every station. A sketch of the plot is shown in Figure 2.
- **Plot 3: Average number of rides per year of birth of the cyclist.** This plot will present the average number of rides per year of birth of the cyclist. A sketch of the plot is shown in Figure 3. The different colors show the gender of the cyclist.

Bring it all together, Figure 4 is a mockup of all the data and user interface we wanted to present to the user. At the top, the user can configure the time range they are interested in, and on the map, they can select which geographic area they want to

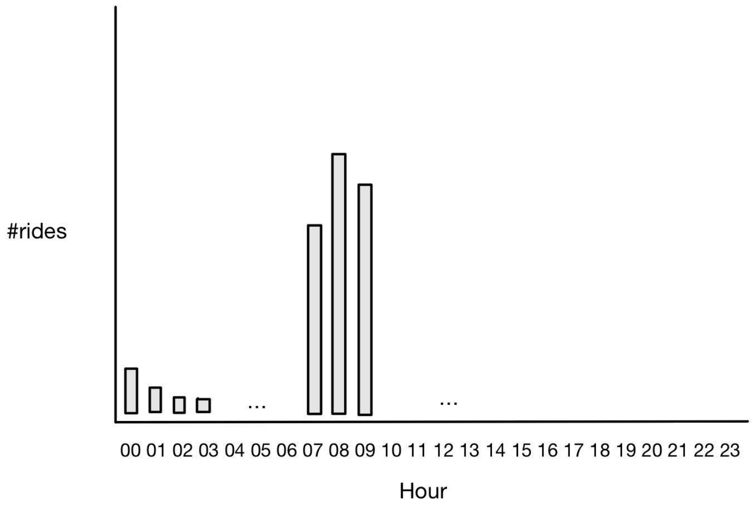


Figure 1: Sketch of Plot 1: Average number of rides per hour

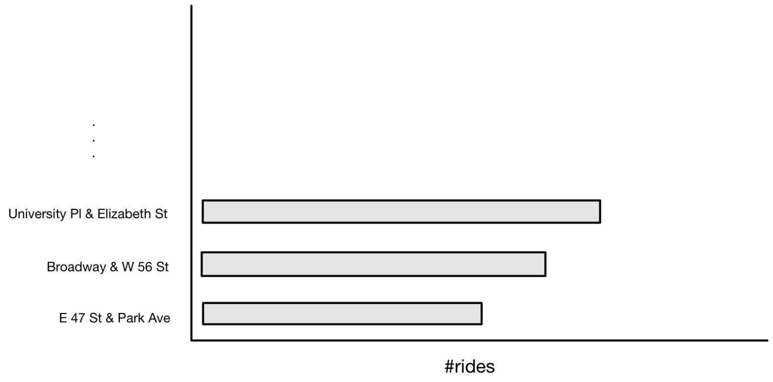


Figure 2: Sketch of Plot 2: Average number of rides per station

examine. For example, in this figure, the blue circle is the area the user has selected. Finally, below the map, information about rides in the selected time and region is displayed.

While working on the map component, we encountered the challenge of displaying every ride in the dataset, which proved to be impractical. Due to the unconstrained nature of rides and the sheer quantity of data points, plotting every ride would have rendered the map nearly unreadable. In essence, each pixel on the map would have been occupied by a portion of a line representing a ride. Furthermore, since we only had information about the start and end locations of each ride, we lacked the precise routes taken by users, making it difficult to plot the actual paths of the rides. To overcome this, we shifted our approach to focus on plotting the stations involved in the rides instead. This decision not only made the visualization more manageable but also allowed for a more intuitive selection of data. Users can now make geographic selections by choosing starting stations, facilitating a more natural and user-friendly experience.

Figure 6 shows our final iteration of the data selection component of the project.

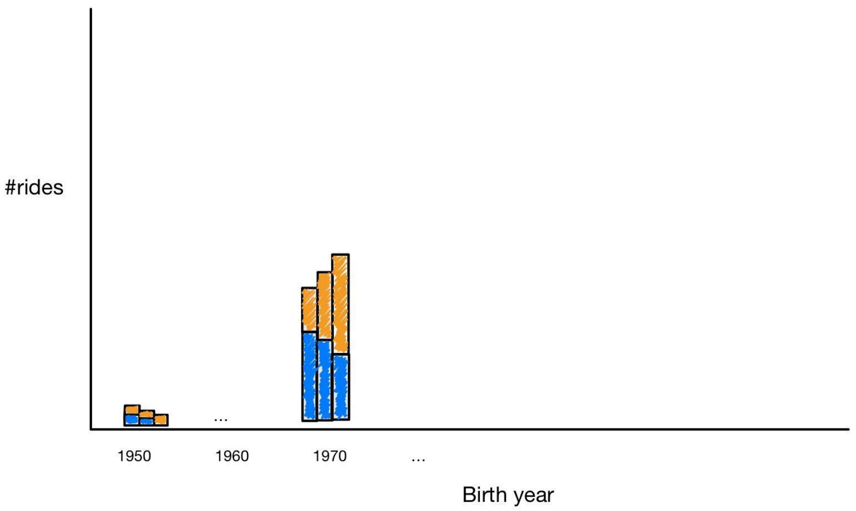


Figure 3: Sketch of Plot 3: Average number of rides per year of birth of the cyclist

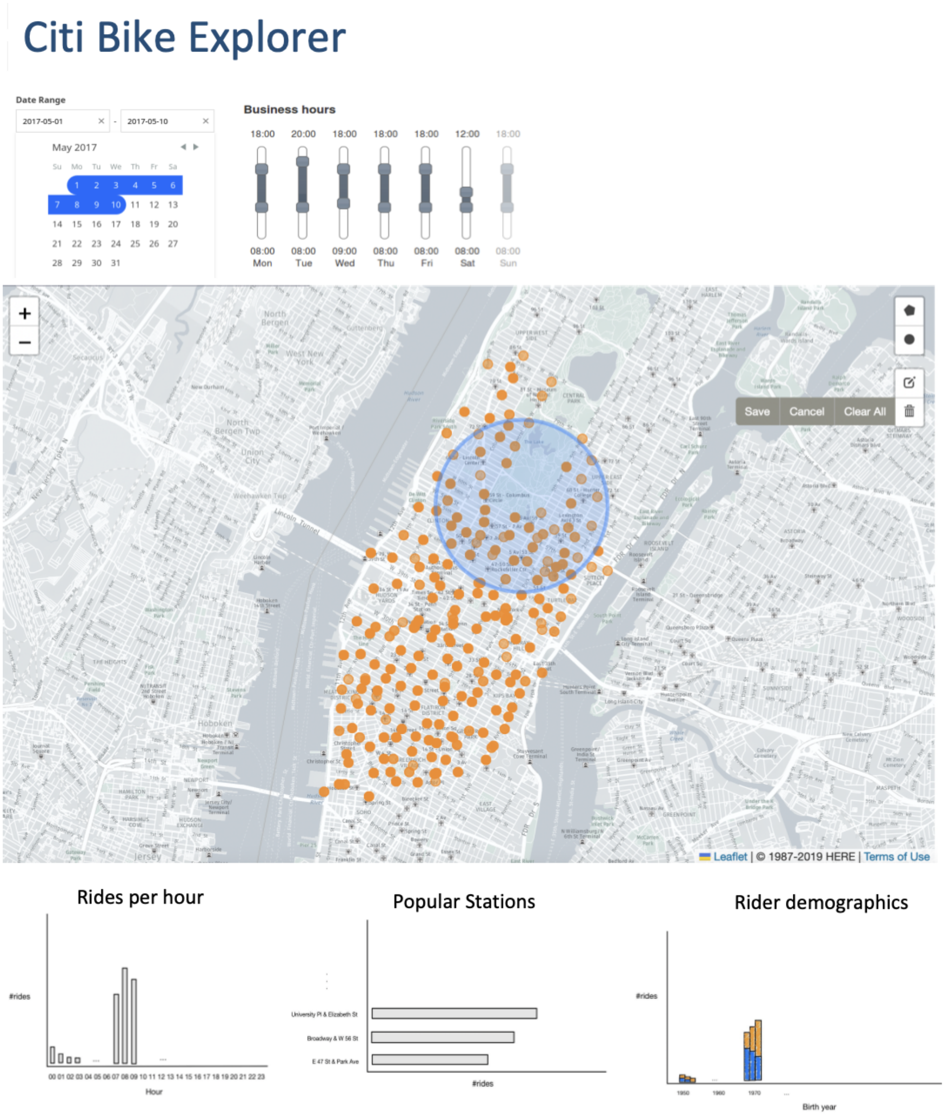
Compared with our initial mockup, we also show the current generated query for enthusiastic users who may want to extract the query for their own use. We also include a brief description welcoming users to the site to contextualize the visualization along with usage instructions.

Figure 5 shows the plots displayed below the map in the final iteration. In addition to the plots we initially planned to show, we decided to expand the visualization by breaking out the popular station statistics into both popular start stations and popular end stations. By doing so, we introduced a new dimension of analysis, allowing users to gain a better understanding of the flow of people through the Citi Bike system. This expanded visualization provides a more comprehensive picture of how riders utilize the bike-sharing system and how different stations contribute to the overall network dynamics.

3 Technical Setup

We embed our visualizations in a website written using React. We use the Leaflet library for the map, and for the accompanying graphs, we use D3. The map code is adapted from examples in MongoDB Compass. We are also using MongoDB to store our data and perform aggregations on the data to make it easier to create our D3 graphs. The website and database are deployed using docker on a server inside the EPFL network. **The website is only accessible if you are on the EPFL VPN or on campus**: http://128.178.52.6:33989

The full dataset is too large (10s of GB) to visualize interactively, so we randomly sample 1% of the rides taken. The full dataset is available as a public dataset on Google BigQuery. Using the table sampling feature of BigQuery, we sample 1% of the data. This results in a $\sim$ 120MB CSV of the data. The dataset from BigQuery contains rides between 2013 and 2018, but has a gap between October 2016 and April 2017. To rectify this, we manually download data from the missing months directly from Citi Bike and sample 1% in python (code in the `fix_data` notebook). Some of the user data is a little bit noisy. For example, some rides have a user birth_year from the 19th century, as users can enter any birthdate they wish when renting a bike. To



Citi Bike Explorer

Date Range

2017-05-01 - 2017-05-10

Business hours

Mon	Tue	Wed	Thu	Fri	Sat	Sun
18:00	20:00	18:00	18:00	18:00	12:00	18:00
08:00	08:00	09:00	08:00	08:00	08:00	08:00

Figure 4: Mock up of the Citi Bike Explorer concept

address this, we omit all rides with a birth_year before 1940.

4 Distribution of Work

We met as a group several times and frequently communicated on a shared chat. In our meetings, we brainstormed the dataset we wanted to explore and the visualizations we wanted to make. After determining which visualizations we wanted to make, Eleni sketched the mock-ups of the graphs and Hamish assembled them together to mock up the entire site. Hamish performed the initial data analysis. Anna built out the base of the website, focusing on the map and MongoDB integration, e.g. the query generation. Yugesh worked on the plots themselves. Both Yugesh and Eleni worked on the styling of the site. Hamish and Eleni did most of the write-ups for the milestones, but all members contributed.

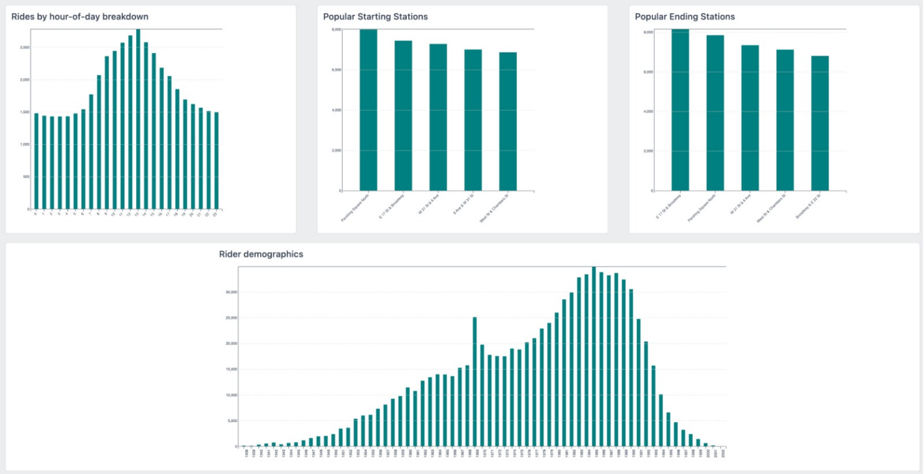
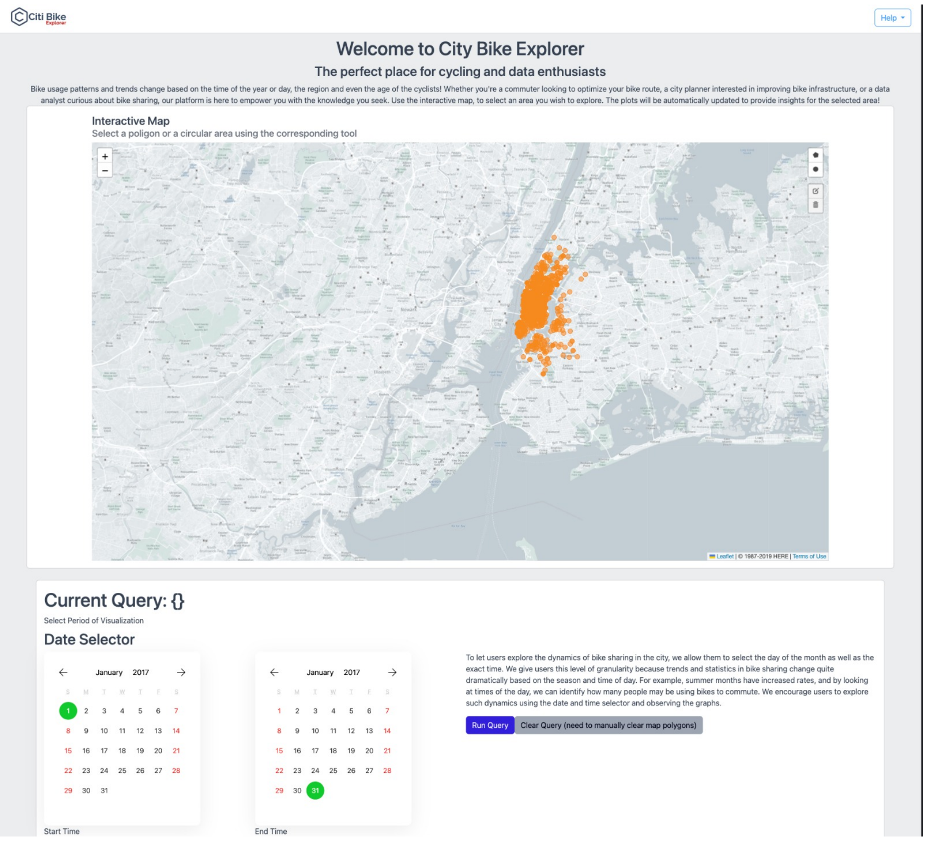


Figure 5: Close up of the final plots



Citi Bike Explorer

Help

Welcome to City Bike Explorer

The perfect place for cycling and data enthusiasts

Bike usage patterns and trends change based on the time of the year or day, the region and even the age of the cyclists! Whether you're a commuter looking to optimize your bike route, a city planner interested in improving bike infrastructure, or a data analyst curious about bike sharing, our platform is here to empower you with the knowledge you seek. Use the interactive map, to select an area you wish to explore. The plots will be automatically updated to provide insights for the selected area!

Interactive Map

Select a poligon or a circular area using the corresponding tool

Leaflet | © 1987-2019 HERE | Terms of Use

Current Query: {}

Select Period of Visualization

Date Selector

January 2017

S	M	T	W	T	F	S
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

Start Time

January 2017

S	M	T	W	T	F	S
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

End Time

To let users explore the dynamics of bike sharing in the city, we allow them to select the day of the month as well as the exact time. We give users this level of granularity because trends and statistics in bike sharing change quite dramatically based on the season and time of day. For example, summer months have increased rates, and by looking at times of the day, we can identify how many people may be using bikes to commute. We encourage users to explore such dynamics using the date and time selector and observing the graphs.

Figure 6: Close up of the final date selector

References

[1] Weixing Ford, Jaimie W. Lien, Vladimir V. Mazalov, and Jie Zheng. Riding to wall street: determinants of commute time using citi bike. *International Journal of Logistics Research and Applications*, 22(5):473–490, 2019.

[2] Yiyuan Lei and Kaan Ozbay. A robust analysis of the impacts of the stay-at-home policy on taxi and citi bike usage: A case study of manhattan. *Transport Policy*, 110:487–498, 2021.

[3] Mark S. Martinez. The impact weather has on nyc citi bike share company activity. 2017.

[4] Eoin O'Mahony and David Shmoys. Data analysis and optimization for (citi)bike sharing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), Feb. 2015.

[5] João Filipe Teixeira and Miguel Lopes. The link between bike sharing and subway use during the covid-19 pandemic: The case-study of new york's citi bike. *Transportation Research Interdisciplinary Perspectives*, 6:100166, 2020.

[6] Kailai Wang and Gulsah Akar. Gender gap generators for bike share ridership: Evidence from citi bike system in new york city. *Journal of Transport Geography*, 76:1–9, 2019.

[7] Wen Wang. Forecasting bike rental demand using new york citi bike data. 2016.