Visualizing CitiBike Data

Anna Herlihy, Eleni Zapridou, Hamish Nicholson, Yugesh Kothari
Data Visualization, Milestone 1

1 Dataset

We will explore data from Citi Bike. Citi Bike is a privately owned public bike-sharing system serving New York City. Citi Bike is a dock-based bike-shared scheme. The dataset contains information about trip histories. Concretely, for every ride, it contains the following fields:

- tripduration (in seconds),
- 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 (subscriber or one-off ride),
- birth_year and
- gender.

There is also a separate stations table containing additional metadata for each station.

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 with the following query[1]:

```
SELECT bikeid
FROM ‘bigquery-public-data.new_york_citibike.citibike_trips‘
TABLESAMPLE SYSTEM (1 PERCENT)
WHERE usertype <> ’’
```

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

[1]The filter is used to remove fields with missing data, see this StackOverflow

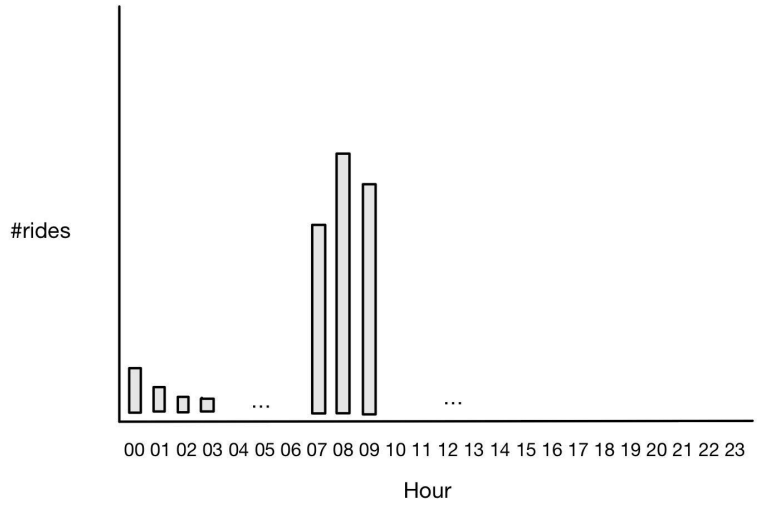

Figure 1: Sketch of Plot 1: Average number of rides per hour

2 Problematic

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 will be able to select a specific area of New York and set a filter for the desired time period. Based on this input, the tool will generate statistics, such as the total number of rides and the popularity of each station in the selected region.

Target Audience: 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 home at specific hours.

Concretely, our tool will include the following visualizations:

- **Interactive Map** Our primary visualization will feature a map of New York that displays the starting and ending stations of each Citi Bike ride as data points. Each ride will be represented by 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. Users will 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.

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

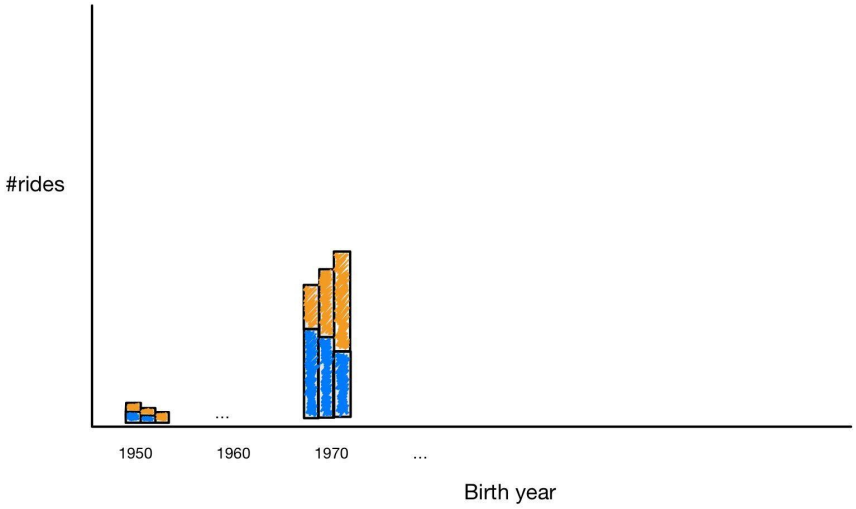

Figure 3: Sketch of Plot 3: Average number of rides per year of birth of the cyclist

3 Exploratory Data Analysis

Our exploratory analysis is in the `data_exploration` notebook. This notebook contains a number of plots exploring aspects of the data.

4 Related Work

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], 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.

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.