TOWARD MORE SUSTAINABLE TRIPS BAYWHEEL DATA ANALYSIS & VISUALIZATION

Colin Wang, Jerry Chan
Halıcıoğlu Data Science Institute
University of California, San Diego
La Jolla, CA 92093, USA
{ziw029, ychan}@ucsd.edu

ABSTRACT

As with the development of electric vehicles and a call for sustainability, companies such as Uber, Lyft, Bird, Spin, etc have provided many sharing options that attempt to optimize their vehicle use and save energy. In this report, we analyzed and visualized Lyft's bike sharing data at San Francisco during 2020-2021, demonstrated different visualization techniques, and showed insights that are generated from the analysis and visualization.

1 INTRODUCTION

To deal with the issue of global warming, governments have been paying more effort in promoting sustainability. Bike sharing, a shared transport service in which bicycles are shared for use to individuals on a short term basis for a price, has been a popular choice in dense city areas where the distance of traveling is relatively short. Many technology companies have came up with strategies to utilize this bike sharing system, such as Uber, Lyft, Bird, Spin, etc. In this work, we will analyze the visualize a public bike sharing dataset from Lyft (Lyft) up to date. The dataset contains useful information to help us make better insights on bike sharing trend in San Fransisco area. Specifically, it includes time-stamp data for every bike record's start & end time, geo-spatial data that illustrate the longitude and latitude of different stations, as well as user and bike information (i.e. user type, bike type). Our visualizations were mostly based on altair (alt) and plotly (plo) with go package

2 MOTIVATION

It can be difficult to manage a bike sharing system. System operator has to move bikes around constantly to make sure there are both bikes and vacancies in all stations. It's impossible to make those arrangement in real-time as it'll take time to transfer bikes around and customers hate waiting. Thus, to accomplish this task, operator has to know the patterns of those bike trips and the user behavior. The goal of our project is to visualize user behavior and try to identify any trends and patterns. We hope that our visualization can help the company better develop this system and help the user get better access to this sustainable way of transportation.

3 DATA SOURCE AND PROCESSING

3.1 SOURCE

We obtain the data from Lyft's baywheel system data webpage at `https://www.lyft.com/bikes/bay-wheels/system-data`. The dataset is public and free to download. It comes with a=the Bay Wheels License Agreement. According to the license we "may include the Data as source material, as applicable, in analyses, reports, or studies published or distributed for non-commercial purposes".

3.2 CLEANING

The raw data is packaged as multiple csv files, each correspond to one month. The tables have slightly different column name. So the first thing we do is unifying the column names and drop columns that doesn't exist in every tables. Next, we realized some categorical features are encoded differently. For example, in some table, customer types are "customer" and "subscriber", in other tables, they are "casual" and "member". Therefore, we also unified the categories for categorical features. Finally, we combine the monthly dataframes into a single dataframe.

We discovered that the dataset includes trip record starting or ending at a test station or deposit station. Since those trips are unlikely to be generated by customers, we filtered them out. Then, we only select the trips in San Francisco based on its coordinate as our study only focus on San Francisco area.

4 PLOTS, ANALYSIS OF PRINCIPLES, AND INSIGHTS

4.1 GEO-SPATIAL MONTHLY USAGE

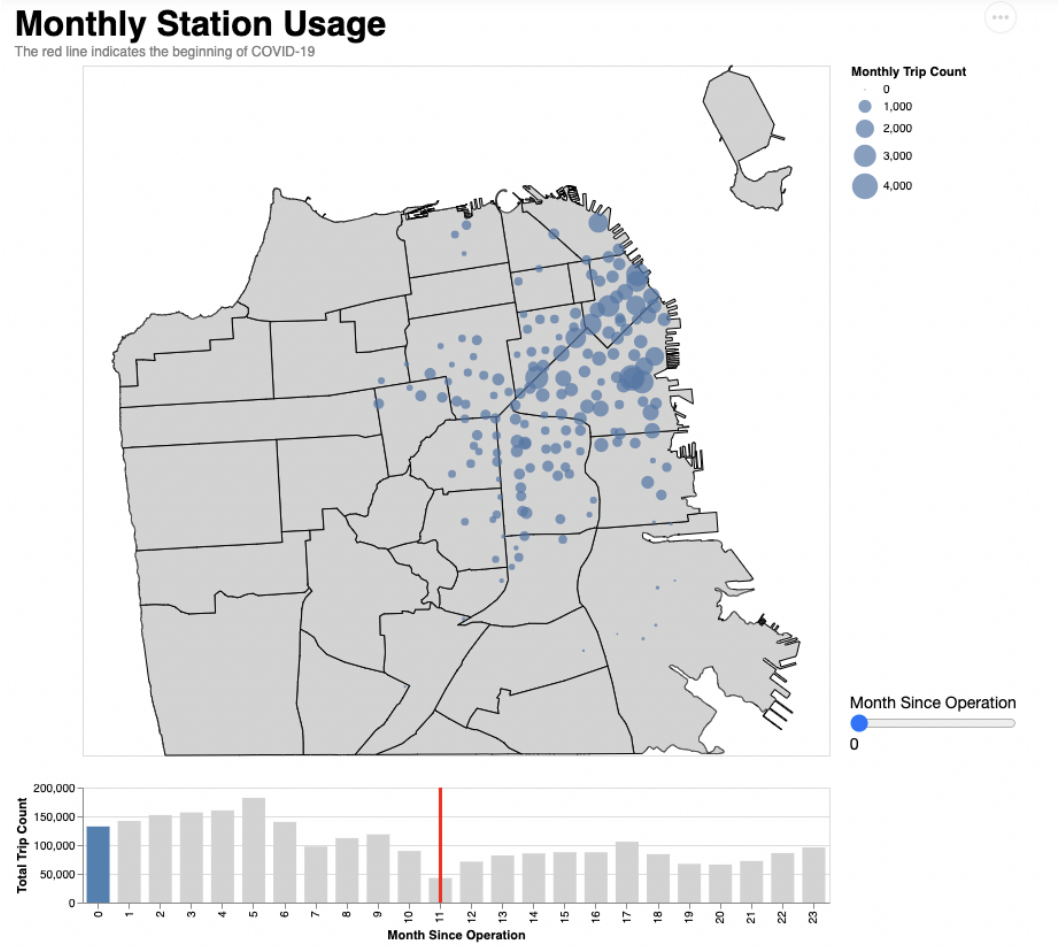

Figure 1: Visualization on Geo-spatial Monthly Usage

4.1.1 VISUALIZATION GOAL

The goal of this visualization is to present the geographical features of the stations distribution and system usage. We're interested in how the distribution and the usage change through time. Since our

data contains trips before and after COVID-19. We also want to see how the pandemic impact the usage. We decided to accomplish it with a proportional symbol map to show the usage distribution and a bar chart of monthly total usage for the user to capture the overall trend.

4.1.2 Data and Encoding

To prepare the data needed, firstly, we figure out the central coordinate of each station by grouping the trips by start station name and average the longitude and latitude in each group. Secondly, we calculate the monthly usage for each station and merge them with station's coordinate. We also compute the total monthly for the bar chart. Lastly, we download the geographical map (shape files) provided by San Francisco government online.

Next, we plot the map with Geo-pandas and overlay a scatter plot of the stations' monthly counts by Altair. We set the symbol size to monthly counts and fix the scale domain on Altair so the size-count ratio will remain constant no matter how the range of the input data changes. We plot a simple bar chart using the monthly total data. We create an slider and a selector based on slider. We add a transform filter to the symbol map to filter out the not-selected month. Then, we put a condition clause on the color argument of the bar chart to highlight the selected month. We create a dummy dataset marking the month COVID started and add the red line to the bar plot with it. Then, we add tool tip function to both plot. Finally, we arrange the position and size of each plot and add some HTML to customize the slider position.

4.1.3 Visualization Channels and Pre-attentiveness

For the symbol map, the channels are symbol location, map, and symbol size. For the bar chart, they are length and position.

We set the color of the symbols the same as the color of the highlighted bar to create similarity. It implies the monthly count it the aggregation of each station's count. We design the red bar in the bar chart to separate the two time period and guide the user to make comparison between them.

4.1.4 Interaction

There is tool tip functions on both graph. When a user hover their mouse on the symbol, it shows the exact trip count starting from the station and the name of the station. When they put it on the bar, it show the year and month the bar is representing and the exact counts. The second interaction is the slider that enables the user to select the month they want to look at. The sliding interaction also creates an animation like effect on the symbol map which help the user perceive the trend better.

4.1.5 Insight

1. COVID Impact: we see significant decrease in usage after COVID, especially for the once popular stations
2. Seasonal Effect: There are more trips in fall. The trip count reaches the peak in October and starts decreasing as the weather turns cold.
3. Station Distribution: When the system first started, the stations clustered on the north-east corner of San Francisco. Lyft gradually expend the network toward the south-west part of San Francisco. The stations on the north-east is still more dense, and they still have a lot more usage then those new stations. We are starting to see growth of usage in those new stations as the impact from COVID wears off.

4.2 Bike Flow

4.2.1 Visualization Goal

Understand the patterns of the trips is very important to mange the system. If we can predict future trips, the company will be able to move the bikes to where they're needed. In this section, we want to visualize the bike flow between each station. It's difficult and ineffective to include all the stations in one graph. Therefore, we choose to only use the trips between the top five most popular stations in our visualization. We decide to present those data with a sankey diagram. Sankey diagram is not

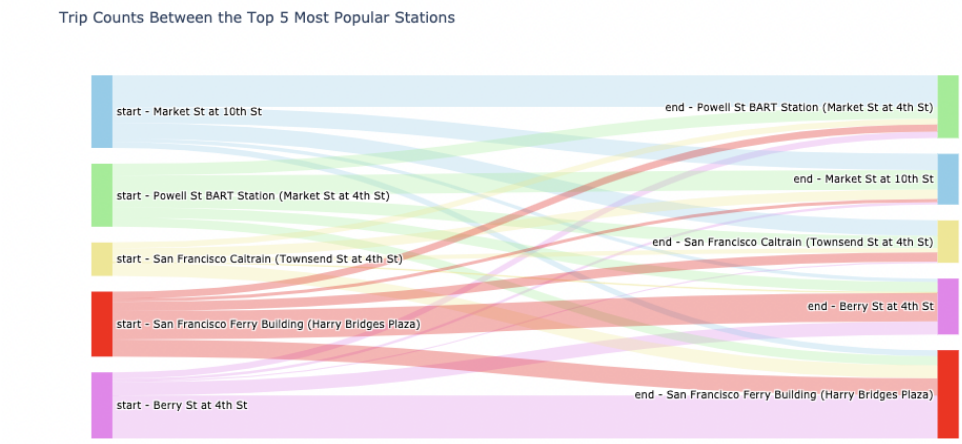

Figure 2: Visualization on Bike Flow

the most accurate way to present the data as human eyes are not sensitive to change in line width. However, we think it's the most straight forward and the most interesting way to visualize those data.

4.2.2 Data and Encoding

First, we select the trips that start and end at one of the top five most popular stations. Next, we group them by their start station and end station and calculate the count of trips between each station pair. We assign colors to each stations (node) and trips count (links). Finally, we transform the data to what Plotly required and draw the diagram with Plotly.graph_objects.go.Sankey.

4.2.3 Visualization Channels and Pre-attentiveness

There are three visualization channels for our sankey diagram, link width, color, and connection. Link width represent the trip count; Color represent the station; Connection represent the start station and end station of each route. We mark the links that start from the same station with the same color to make tracing the lines easier. We picked light, categorical color with low opacity. The light, transparent color enable the users to trace the links more clearly when they intersect. The colors are only different in hue so user won't make wrong inference about the order of the color.

4.2.4 Interaction

User can get trip counts a link represents by hovering their mouse on the link and the total trip count for a station by hovering their mouse on the the node (station). The user can rearrange the position of the nodes by dragging. This is a really useful interaction to have on a sankey diagram. Users can separate out the node they want to look at and see the links more clearly. They can also drag the nodes into clusters to look for some higher level patterns between the stations.

4.2.5 Insight

1. Bike Route: The most popular bike route is from Berry Street at 4th Street to San Francisco Ferry Building
2. Stations: The most popular starting station is Market Street at 10th Street and the most popular ending station is San Francisco Ferry Building
3. Bike Managing: Some station received a lot more bike than the bike they lost, for example the San Francisco Ferry Building station rent out about 6000 bikes but received about 8300 bike. This means the company have to constantly transport bikes from this station to other

stations. From the diagram we can see that most bikes returned at San Francisco Ferry Building station come from Berry Street at 4th Street station. Therefore, the company might have to move the bikes from San Francisco Ferry Building station to Berry Street at 4th Street station to keep the system running.

4.3 Usage & Duration Trend

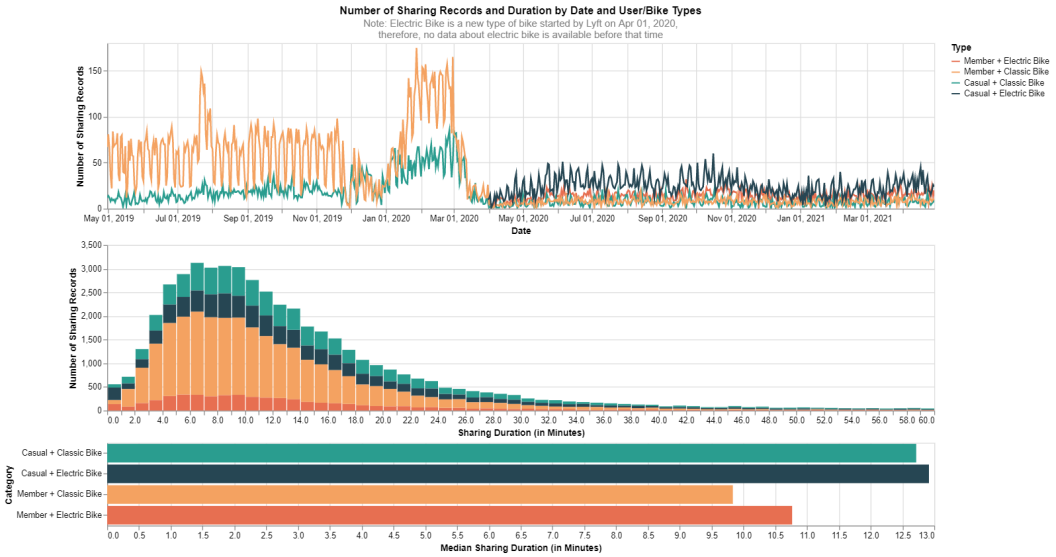

Figure 3: Visualization on Usage & Duration Trend

4.3.1 Visualization Goal

The goal of this visualization is to understand how number of sharing records changed as a matter of day, month, year. We want to understand how these differ by each type of user group (member versus casual), as well as bike group (classic versus electric). By the same token, we want to see the distribution of sharing duration for any specified time range so that we can understand if there is a shift of sharing duration at a particular time frame, or if there is a significant change in each user/bike type's distribution of that given a time frame.

4.3.2 Data and Encoding

In this visualization, the data being used are each record's start sharing time, duration, and its corresponding user and bike type. For start sharing time, we specified the smallest unit to be a day, so the figure itself has a clearer patterns across weeks and seasons. This is represented with a line chart in the first figure.

In the second figure, which is a histogram, We transformed the duration unit to be minute and aggregated the counts of duration records with 60 bins. We also used a bar chart and extracted the median of duration time to be shown in the third plot, grouped by each user and bike type.

In this visualization, we only randomly sampled 50,000 records out of around 3,000,000 observations with a fixed random state. This is due to the computational limitations of our interaction package, altair.

4.3.3 Visualization Channels and Pre-attentiveness

In this visualization, we mainly used color, position, height and length as our visualization channels. For colors, we encoded electric bikes with darker tones, classic bikes with lighter tones, members with orange-ish tones, and casual with green-ish tones. The horizontal position of the three figures

represent the time flow, the time of duration, and the median of the duration value. The height of the first two figures represent the aggregated count records.

4.3.4 INTERACTION

First, viewers can zoom in and out the first time-series figure if they are interested to see fine details about a particular week or month. Second, then can select an area (a time frame) in that time-series figure. When they do so, the second and third charts' data will be based on the time frame they select. Third, viewers can select a single or multiple types of users or bikes, and the figure will only show the selected types with color, encoding unselected types with a light gray color.

4.3.5 INSIGHT

1. Most riders spend 10-15 minutes in a sharing session, but some riders can spend significantly more time than a typical range, forming a long right tail, Casual riders spend more time than member riders for a sharing session in general,
2. Riders on electric bikes spend slightly more time than riders on classic bikes for a sharing session in general,
3. An unusual spike of sharing sessions happened between January 2020 and March 2020, and suddenly decreased from April 2020, perhaps due to the COVID lockdown, beginning of electric bike transitioning, and Lyft's new policy on bike-sharing.
4. Weekly patterns are visible and consistent, meaning that bike sharing has purposes for most riders (i.e. weekend ride for leisure, weekday ride for working, etc),

4.4 BIKE USE TIME ACROSS THE DAY

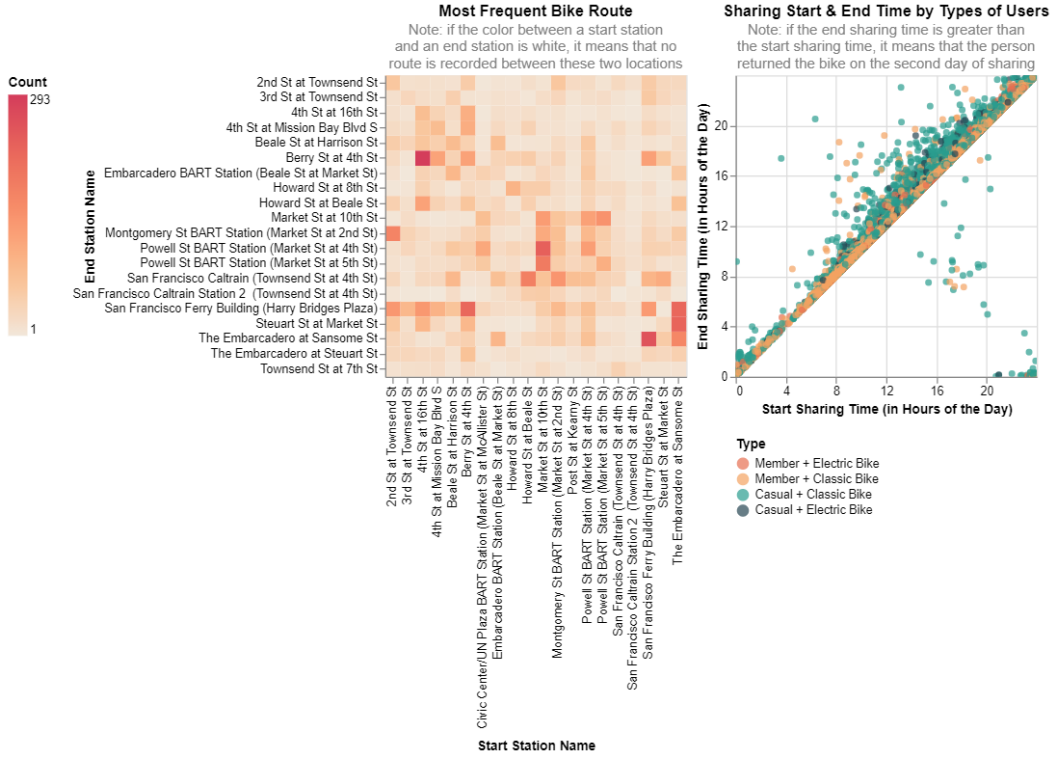

Figure 4: Visualization on Bike Use Time across the Day

4.4.1 VISUALIZATION GOAL

The goal of this visualization is to understand the relationship between most frequent bike routes and the time of the day people start and end sharing the bike. In particular, when want to see if there exists a significant difference in the sharing time among different places, and if different types of users and bikes have a different sharing time range given the same route.

4.4.2 DATA AND ENCODING

In this visualization, the data being used are each record's start sharing time, end sharing time, start station name, end station name, and its corresponding user and bike type. For start and end sharing time, we specified the smallest unit to be a second, and ignored the day, month, and the year, so the figure itself has a clearer patterns across a single day. This is represented with a scatter plot in the second figure.

In the first figure, which is a heatmap, we chose the 20 most popular start route and 20 most popular end route as our choices, and aggregated the counts of each route as the intensity for each pixel in this heatmap.

In this visualization, we only randomly sampled 50,000 records out of around 3,000,000 observations with a fixed random state. This is due to the computational limitations of our interaction package, altair.

4.4.3 VISUALIZATION CHANNELS AND PRE-ATTENTIVENESS

In this visualization, we mainly used color, position, and intensity as our visualization channels. For colors, we used the same set of encoding as we used in the last visualization to ensure consistency. For position, we encoded the start sharing time as the horizontal axis of the scatter plot, and the end sharing time as the vertical axis of the scatter plot. Since the end time is always later than the start time, most points appear to be at the upper-left part of the scatter plot. Points that appear at bottom-right part of the scatter plot means that the sharing time is over a day span. For intensity, we used a continuous color scale to encode the counts of a recorded routes, where a darker tone represents more counts, while a lighter tone represents fewer counts.

4.4.4 INTERACTION

First, viewers can zoom in and out the second scatter plot figure if they are interested to see fine details about a particular hour or hours. Second, viewers can select one or multiple routes of interests by clicking the heatmap, and the scatter plot will conditionally filter records based on the viewers' selections. Further, viewers can select a single or multiple types of users or bikes, and the figure will only show the selected types with color, encoding unselected types with a light gray color.

4.4.5 INSIGHT

1. Night riders spend less time, day riders spend more time,
2. Member riders are more likely to have one-way trip for a single sharing session, while casual riders are more like to have a round trip for a single sharing session,
3. Member riders are more likely to ride in the morning, casual riders are more likely to ride in the afternoon,
4. There exists significant differences between the frequency of different routes,

5 FUTURE SCOPE

In this work, we have mostly analyzed and presented the Lyft bike sharing data in terms of "when" and "where". Future works can focus more on seasonal and weekly patterns of bike sharing, co-related routes, and taking account of regional population, district type, etc to help achieve better optimization of shared bikes. In terms of visualization, future works include having a fine-detailed with streets on geo-spatial maps, more encodings on the sankey diagram, as well as a user-defined smoothing factor applied to the time-series map.

AUTHOR CONTRIBUTIONS

Colin did most part in organizing layout for presentation and report, drafted ideas and wrote code (including inline comments) for the last two visualizations. Jerry did most part in organizing layout for the notebook, performed data collection, cleaning, processing, and drafted ideas and wrote code (including inline comments) for the first two visualizations. In terms of presentation and report contents, both authors have an approximate 50/50 split.

ACKNOWLEDGMENTS

We want to thank ICLR for providing this LaTeX template for us to organize and demonstrate our work. We also want to thank Prof. Nusrat for instructing us the visualization techniques required in the report, as well as course staff who provide support and help during this process.

REFERENCES

Declarative visualization in python¶. URL https://altair-viz.github.io/.

Plotly python graphing library. URL https://plotly.com/python/.

Inc Lyft. System data: Bay wheels. URL https://www.lyft.com/bikes/bay-wheels/system-data.