Visualizing Bluebikes Rides During the COVID-19 Pandemic

Belinda Shi*
MIT Department of Computer Science and Engineering

Ahmed Elbashir †
MIT Department of Computer Science and Engineering

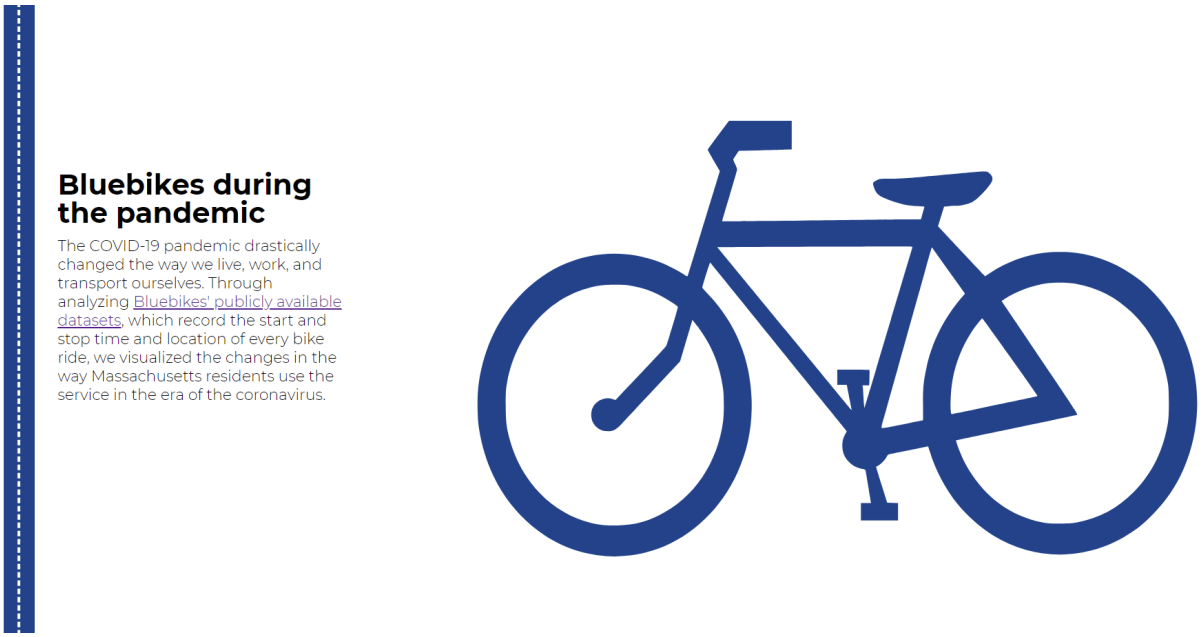

Figure 1: Project landing page

1 Introduction

Our final project is a "scrollytelling" webpage which contains four visualizations examining the impact of the COVID-19 Pandemic on Bluebikes ridership. Our data is sourced directly from Bluebikes records on individual trips, and our visualizations illuminate different aspects of the way people used Bluebikes following the onset of the pandemic. Interactive elements and textual descriptions accompanying each visualization provide more information and context that cannot be captured with numbers alone, allowing us to highlight key observations and provide background information.

2 Goals

For our visualization, we wanted to examine the effect the COVID-19 on how people used the Bluebikes bike-sharing service in the Greater Boston area. We expected that this data would have changed greatly during the pandemic, and by creating compelling and informative visualizations using the data on rides Bluebikes made available, we would gain insight and be able to inform others of how people's behavior changed during the pandemic.

We found interesting changes in user behavior after the onset of the pandemic that spanned multiple dimensions. Because it would not be possible to capture all the interesting effects we saw in one visualization, we chose to use a "scrollytelling" format which incorporated multiple visualizations. This format would also allow us to use text descriptions to provide background information on each visualization that could call attention to notable changes and explain what is being shown for audiences that are less familiar with Bluebikes and the greater Boston area.

*e-mail: beeshi@mit.edu
†e-mail: ahmed176@mit.edu

3 Related Work

Previous research has analyzed bike share usage to infer the underlying mobility patterns in cities. Froehlich, Neumann, and Oliver [2] analyzed and visualized 13 weeks of usage of a bikeshare system called Bicing in Barcelona to identify common movement behaviors of the residents in the city. We use similar data cleaning and transformation techniques to prepare our bikeshare data, and also use the resulting visualizations to make conclusions about rider behavior.

We were inspired by other scrollytelling visualizations such as R2D3's machine learning visualization [3], which makes great use of animated scatterplots and bar charts to show how data is transformed in the process of a machine learning algorithm. We hope our visualization also effectively assists the user in stepping through the story we are trying to present.

We consulted Chow's article [1] on building an interactive scroller in D3 as a base for our scrollytelling functionality.

4 Visualization Description and Design

4.1 Data acquisition and transformation

The original Bluebikes data (as sampled in Fig. 2) is provided on the Bluebikes website as monthly CSVs recording one row of information for each individual ride made that month. Each row contains information about the start station, end station, start time, and end time of the ride, as well as a few other columns which were less useful for our visualization. These include information about the user which ceased to be collected in 2020, such as their gender and birth year, as well as the unique ID of the bike associated with the

	row_num	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	birth year	gender
0	513	00:50.9	09:24.7	27	Roxbury	42.331	-71.1	282	Stony Br	42.317	-71.1	4088	Subscrib	1995	1
1	322	04:49.2	10:11.5	39	Washing	42.339	-71.07	46	Christian	42.344	-71.09	3027	Subscrib	1992	1
2	425	05:56.0	13:01.8	12	Ruggles	42.336	-71.09	21	Prudenti	42.347	-71.08	3818	Subscrib	1997	1
3	159	08:18.8	10:57.8	279	Williams	42.307	-71.11	133	Green St	42.311	-71.11	3500	Subscrib	1987	1
4	1229	09:13.0	29:42.5	16	Back Bay	42.348	-71.08	78	Union Sc	42.38	-71.1	2198	Subscrib	1978	1
5	1238	15:16.3	35:54.9	190	Nashua	42.366	-71.06	378	191 Beac	42.38	-71.11	2857	Custome	1969	0
6	501	27:30.9	35:52.6	75	Lafayett	42.363	-71.1	91	One Ken	42.366	-71.09	3192	Subscrib	1989	1
7	436	43:01.3	50:18.0	108	Harvard	42.378	-71.12	87	Harvard	42.367	-71.11	2355	Subscrib	1977	1
8	450	44:14.5	51:45.1	14	HMS/HS	42.337	-71.1	56	Dudley S	42.329	-71.08	3636	Subscrib	1993	1

Figure 2: Sample of raw Bluebikes data from March 2019

ride. We also had to choose which months of data to use in our visualizations. We chose to center our data around March 2020 as the inflection point of the pandemic in the United States, and included data from March 2019 to March 2021 to maintain an even distribution of time before and after the pandemic.

In its raw form, this data was too large for us to use in our visualizations: just the March 2019 data contains over 100,000 rows and is 22 MB. Over 24 months of data, we had over a gigabytes' worth of data. To save the computational cost of having to load an excessive amount of data in the visualization, we used Tableau to create prototypes for the visualizations we wanted isolated to the specific fields and transformations of those fields (often averages or sums) required, and used Tableau's export data feature to create CSVs more readily suited for our visualizations. Alongside the geoJSON used to create the map, this reduced the size of data loaded in the visualization to just over a couple hundred kilobytes.

While the Bluebikes data was largely high quality, there were issues we had to deal with. The largest was a discontinuity in the hourly distribution of rides from 2020 to 2021. After observing the hourly distribution of rides histogram shifted roughly 5 hours forward starting in January 2021, we contacted Bluebikes customer support and learned that they switched the time zone of their recording to UTC instead of ET without reporting it, so we manually corrected our data. There are also many outlier rides with an implausibly long recorded trip duration in the thousands of hours, so we opted for median over mean to reduce the influence of those outliers when trip duration was a relevant metric.

4.2 Landing Page and General Design

Because we wanted to have multiple visualizations and descriptions accompanying each visualization, we chose a scrollytelling format. We wanted our scrollytelling page to look authoritative, modern, and consistent with Bluebikes' branding. We followed a typical scrollytelling design with text on the left and a visualization on the right, beginning with a short paragraph that introduced the data and the goals of the project (Fig. 1). The bike on the right prepares readers to see visualizations on that half of the screen, and the blue stripe on the left mimics a road that users are biking down, while also nicely setting a boundary for the text. Throughout the visualizations we reuse the dark blue and light blue color values in official Bluebikes' brandings, to make the connection stronger and more consistent with the subject of the project. Where there are opportunities for interaction, we prefer hover-over to clicking to activate the interaction as it is more intuitive to readers familiar with current web visualizations and requires less work of them. We also use the textual descriptions to explain our variables, and how they were transformed or aggregated in instances where it is not clear. This lets us reduce the text needed on the axes of the visualization.

4.3 Visualization 1, Bluebikes Rides over Time

The first visualization (Fig. 3) is a line chart of total Bluebikes rides from month to month. It primarily communicates a baseline for ridership and the sharp drop in ridership in April 2020 correlating to the pandemic, which we emphasize with the highlighted "April 2020" text in the description that when moused over creates a circle over the data point. The visualization communicates secondarily the seasonal trends in Bluebikes ridership.

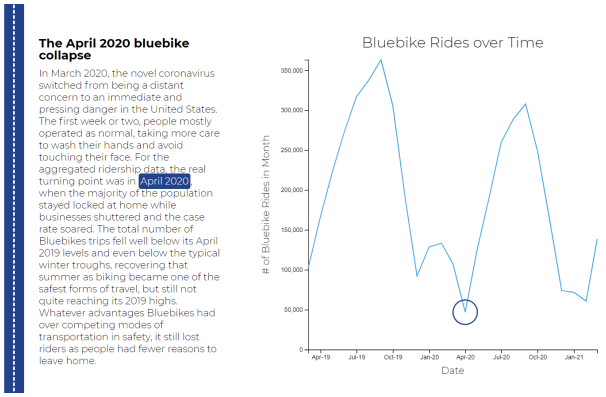

Figure 3: First visualization, total Bluebikes rides over time

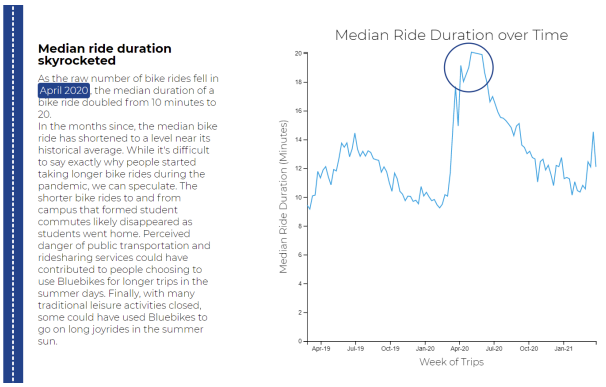

Figure 4: Second visualization, the median ride duration over time

We chose to aggregate by month because more granular time units resulted in a messier line that obscured the important trend and a coarser time unit, like quarter or year, would lose the important pandemic change. We chose a line chart over alternatives like a bar chart because line charts better emphasize trends, while still making it easy to see relative and absolute magnitude. Users can also mouse over a data point to get a tool tip with more specific numbers on the rides in each month if they want more detail. We also chose to use data from March 2019 to March 2021 to compare a year's worth of pre-pandemic and post-pandemic behavior.

4.4 Visualization 2, Median Ride Duration over Time

The second visualization (Fig. 4) is a line chart of the week-by-week median ride duration from March 2019 to March 2021. It primarily displays how the median ride duration increased during the pandemic and over the course of months returns to resemble its' pre-pandemic baseline. The highlighted "April 2020" text in the description can be moused over here as well to place a circle on the peak of the data, to emphasize the point of the visualization.

This was a change in behavior that was very stark and without an immediately obvious explanation, so it was worth including as a point to speculate on and engage reader's curiosity. We chose a line chart again to emphasize the trend in ride duration better than a bar chart would. We used the median ride duration to avoid the impact of implausibly long outlier rides on the mean distorting our visualization. We again provide a hover-over tooltip which displays the median ride duration in that week for users who want more

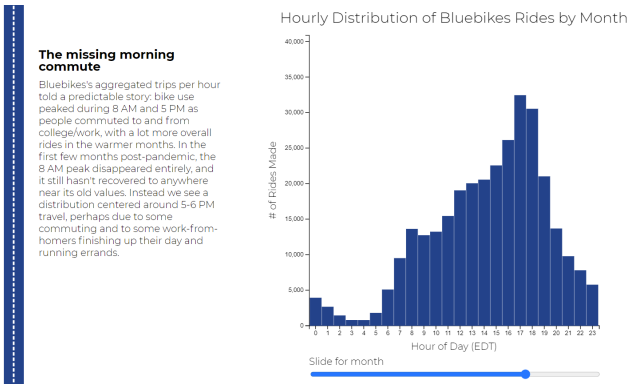

Figure 5: Third visualization, hourly distribution of Bluebikes rides by month

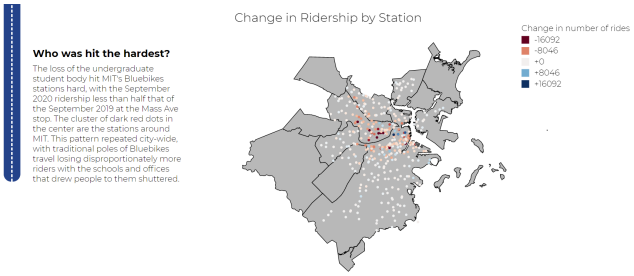

Figure 6: Fourth visualization, map of changing ridership by station

detail. We chose to aggregate time weekly here because we wanted to capture more detail about the trend so that readers could see how well the spike matches with the intensifying pandemic in late March, and get a better feel for the natural variation in trip duration week-to-week. While the median ride duration is recorded in seconds, it was changed to minutes as for the range of this data (10-20 minutes) it was more intuitive to read time in minutes than seconds.

4.5 Visualization 3, Hourly Distribution of Bluebikes Rides by Month

This visualization (Fig. 5) displays the hourly distribution of Bluebikes trips made each month using a histogram. There is a slider which selects the month of data being visualized. The primary message of this visualization is how the morning commute used to be very visible in Bluebikes ridership data as a peak of rides at 8 AM, but disappeared at the onset of the pandemic and has only partially returned a year later. Secondarily, the visualization also lets users explore seasonality in Bluebikes data, observing how sharply ridership at all hours drops in winter months.

Although the data is not formally a probability distribution as it is not normalized, because the focus of the data is how trips are distributed across the discrete time variable of hours of the day, we believed a bar chart a better fit than a line chart. Bucketing the time dimension by hour rather than more granular variables like minute let us shrink the file size, improve load time and keep viewers' focus on the most important information.

4.6 Visualization 4, Change in Ridership by Station

Our final visualization (Fig. 6) is a map of the greater Boston area where each Bluebikes station is plotted as a dot and colored along a red-blue axis depending on how their ridership changed after the onset of the pandemic. We aggregated the number of rides for a year before the pandemic and after the pandemic, splitting on March 15 as our dividing point. It shows the sharp drop in ridership around MIT and other university and business hubs, but more muted changes in the already low-ridership outlying areas, and slight gains in a few stations.

Originally this visualization was interactive and users were able to press a button to switch between two maps counting the number of rides before and after the pandemic, respectively. Ultimately we decided because the point was to illustrate change in ridership, using one visualization which plotted the difference in the number of rides was better. We encoded this in color rather than size because it could be negative and because our points are too close together and increasing the size could result in overlapping obscuring the information. We chose a red-blue color scale because red is usually associated with loss and negativity and blue matches our theming and is often considered an opposite to red. Red-green encoding was also considered, but because it is the most common form of colorblindness we avoided it.

5 Conclusion

We chose this project because we were curious how the COVID-19 pandemic changed human behavior, and Bluebikes data provided a unique view into one way the pandemic upended people's lives. While working on it we discovered some things we expected and others we didn't, and we hope that exploring our visualization inspires our readers' curiosity and sparks interest in how they can use data visualization to learn more about how the world has changed.

References

[1] C. Chow. How i created an interactive, scrolling visualisation with d3.js, and how you can too, Mar 2020.

[2] J. E. Froehlich, J. Neumann, and N. Oliver. Sensing and predicting the pulse of the city through shared bicycling. In *Twenty-First International Joint Conference on Artificial Intelligence*, 2009.

[3] S. Yee and T. Chu. A visual introduction to machine learning. *R2D3*, 2015.