Analysis of US Bike Share Data

Jazmin Gurganus[1], and John Pi[2]

***Abstract*— With dozens if not hundreds of ride share services being formed over the past several years, it is apparent that the industry is growing rapidly. These services range from car share services to scooter share services. In this proposed project, we hope to analyze how and why consumers are using ride sharing and bike sharing services in this new post Covid and during Covid world.**

I. OBJECTIVE

The objective of this project is to analyze the factors that contribute to why and how users are using ride/bike sharing services. A ridesharing company is defined as "a company that, via websites and mobile apps, matches passengers with drivers of vehicles for hire that, unlike taxicabs, cannot legally be hailed from the street"(wikipedia c4). In this project we have considered electric scooter rental businesses and other services as well, but we are mainly focusing on bike share data. so we want to see all the factors that contribute to why users decide to use ride sharing services and how they are being used and the many other factors that could be at play, that could possibly be like population density, economic factors or culturally.

II. DATA

In order to analyze our large amount of bike share data this project mainly utilized publicly found data sets given by the most notable ride/ bike sharing companies of their data and trip history for public use. they are validated by the companies themselves as they also remove the entries that can cause outliers in the data with this disclaimer that "This data has been processed to remove trips that are taken by staff as they service and inspect the system, trips that are taken to/from any of our “test” stations (which we were using more in June and July 2013), and any trips that were below 60 seconds in length (potentially false starts or users trying to re-dock a bike to ensure it’s secure)." These data sets were generated from more notable bike sharing companies such as Divvy from Chicago, IL, Citi bike from NYC, NY and Jersey City, NJ, Capital Bike share from Washington D.C., and Metro Bike Share from Los Angeles, CA. Most of these companies provide data sets that include at least trip id, ,trip duration, time and date, start time, start longitude and latitude, start station, end time, end longitude and latitude, end station and the membership status of the customer. Other stats that may possibly be included are gender, year of birth, bike types and bike ids. We put our data into the Jupyter notebooks to analyze and create graphs out of them, we mainly looked at the different total amount of trips by the users of the last few years to see if we could draw any notable conclusions and results from the data. It is important to say that since this year is not fully complete we will only have 11 months worth of data for all cities so all of the calculations will be done with 11 months in regards to 2020. We processed the data by parsing through the data sets and seeing how many total trips were there and the other most usable statistic of trip duration.

III. MODELS/ALGORITHMS

The main package that was utilized in our project was the pandas library that helps with data manipulation and analysis, with its helpful data structures and operations for our data sets. The main algorithms used to parse though the data were a general accumulator to count the amount of trips in each city to compare between each other and itself throughout each year and an averaging technique used to find the average trip duration in each city to see if we could draw any notable conclusions when compared to other statistics public transportation data from each city. Also calculating the rate of change between each of the years of total trips taken.

IV. RESULTS

From this analysis, we were able to conclude some interesting data points about bike sharing in some of these major cities. Such as in New York City, that while there was a decrease in bike share trips from 2019 to 2020 over the first 11 months, the percent of negative change was not as great as other public transportation methods used by the general public, with the New York City subway and buses having a -74 percent and -51 percent change, the percent of change between 2020 and 2019 for Citi Bike Share was only -6.1 percent, all of these negative trends were due to Covid but it may be a good sign in these cities, for the future of bike sharing if this keeps up even after covid as more people will have gotten used to biking over using the subway and cars to commute through NYC.

This is similar for the DC area as well as there was a -75 percent change with the DC rail and -45 percent change with the DC buses, but with the data given from Capital Bike Share the rate of change was calculated to be at -35 percent. The data suggests that there may be a trend that bike share usage while it may have decreased a little due to Covid, it is still a popular alternative to other membership based transportation methods due to being a little bit more socially distant and providing more of a peace of mind for users trying to stay safe during the pandemic compared to public transportation.

Results from the LA Metro Bike data set can seen in the next couple of graphs. We can see from the above figure

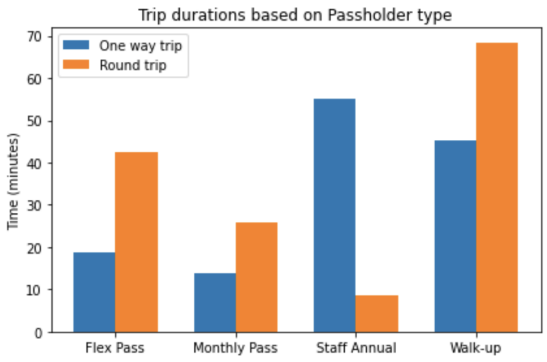

Fig. 1. Trip Durations

that the trip duration drastically decreases if a rider is a pass holder. This trend can be seen in both types of trips, one way and round trip. If a rider has a pass, the trip duration ranges from 13 to 18 minutes, on average. This is significantly less than the walk-up data, which says that trips range from 45-48 minutes on average. This may suggest that those who have passes are using bike share to commute, rather than recreation. It may be that it is easier to commute less than 20 minutes on bike versus a longer average commute time on the LA Metro system.

V. ISSUES ENCOUNTERED

During the collection, parsing, and display of gathering the bike share data there were some issues that we encountered. The first main issue we had had to deal with the size of the data sets. The larger size of the data sets meant that it was harder to use the data sets to find particularly useful data at first due to the substantial size of the data sets since they spanned over many years and months, this could be due to the increased popularity of bike sharing apps over each year and due to the COVID pandemic.

Due to the size of the data sets made it hard for us to work with the data since GitHub.com has an upload limit of 25MBs. So due to that inconvenience, we decided to use the Jupyter Notebook to store, compress and combine the data and running code on the Jupyter Notebook to create graphs and analysis for our project. The size of the data made it really hard and slowed down the process of moving the data to be used and analyzed.

Additionally, many of the reports and data sets we found online were anonymized, meaning that most of the demographics about a person were not included. We had hoped to find the median age of riders,as only one data set provided the age of their users by including their birth year which would be more work to calculate their age, as well as other information like residence. As this would have helped us determine other factors in our analysis such as what age range of users are most likely to use bike sharing apps and utilize bikes more in their lives, and what income level is more likely to utilize bike sharing.

And finally the last issue we had for our project was the lack of usable information that was given in the data sets. Many of the companies only give a limited amount of information of the bike sharing with at minimum the only things that all of the companies do provide is the date and time of the trip, the starting time, the starting location, the ending time, the ending location, and membership status of the consumer. Only a few companies actually gave out the data on other statistics such as gender, year of birth, and bike ID number. So this limits the amount of analysis we can do with our given data sets.

some other issues that we had were some of the data sets were not ordered the same way internally so the parser would break and not be able to count he number of trips to do analysis with. This affected us with the Chicago data set.

VI. FUTURE WORK

In the future, additional work can be built on to this project, such as also analyzing other countries cities rather than just the US cities as done as in this project. This can be useful in the future as more and more countries are trying to either cut down on the usage of cars and heavy emitters of carbon dioxide and more countries are promoting the use of ride sharing and safe public transportation methods due to the widespread-ness of COVID-19. This leads to the second point of future we can do on this topic as due to the COVID-19 pandemic and if this pandemic is a positive or a more negative effect on bike share in general. And finally seeing whether the amount of memberships and rides per month stays constant or changed throughout the decade. And due to time constraints we could compare the average commute time of a city to the average trip time of a bike share app to see if people were using bike share possibly for work commuting.

For our first possible additional work it would be interesting to analyze differences between North American rider data and European rider data or even Asian rider data, as many of the cities in foreign countries are even more or equally as condensed and as crowded as New York City, or many cities have even been banning cars and or diesels cars from even driving on the roads or even removing parking spaces to reduces emissions and promote walking for maybe even biking. These are due to many countries being apart of the Paris climate agreement and trying to reach emission reductions by their promised date.

For additional future work regarding the mount of memberships and rides per month we can see if the trend continues from the article on nacto.com if increasingly more people are indeed signing up and utilizing bike sharing apps each year or if it will become stagnant eventually due to some reason.

Finally the most current of Future work topics we could implement in the future is of the current COVID-19 pandemic. we can analyze the effects of Covid-19 on the bike sharing industry whether it was a positive or a negative effect. Due to many articles stating the fact that bike sales have increased due to Covid-19 and more people are using bikes as an alternative to public transportation to social distance due to fear from the virus that bike sales have increased but

it is unclear on whether this is a net positive or a negative in bike sharing due to more people being bike owners, or more people adopting bike sharing apps.

VII. ORG CHART

The following list was the rough timeline for the major milestones for our project(the weeks are counted from the second week of the submission of the proposal):

- **Week 1** - Gather a list of functional, and able to be validated data that is related to our topic of bike sharing or ride sharing in a few major cities.
- **Week 2** - Start developing code to be able to analyze and quantify the bike sharing data sets from **Week 1**
- **Week 3** - Further parsing the data with addition to data from each city and some other contributing factors.
- **Week 4** - Create graphs from the data sets and the data from the city and other contributing factors
- **Week 5** - Draw meaningful conclusions from our analyzed data sets and graphs.
- **Week 6** - Create and develop a cumulative analysis of our data sets conclusions with deliverables.
- **Week 7** - Create and prepare a final project report of our bike share analysis and presentation.

The following is a list of project members and their responsibilities:

- **John Pi** - Ran most of the code for the data sets and combined the data to make the sets workable and worked on the writing of the final report and the development of the final presentation.
- **Jazmin Gurganus** - Handled the gathering of the data sets and the developing the code to print graphs and sort the data and worked on the writing of the final report and the development of the final presentation.

APPENDIX

REFERENCES

[1] https://github.com/awesomedata/awesome-public-datasets
[2] https://www.lyft.com/bikes/bay-wheels/system-data
[3] https://github.com/BetaNYC/Bike-Share-Data-Best-Practices/wiki/Bike-Share-Data-Systems
[4] https://en.wikipedia.org/wiki/Ridesharing-company
[5] https://nacto.org/bike-share-statistics-2016/
[6] https://www.metro.net/news/facts-glance/
[7] https://new.mta.info/coronavirus/ridership
[8] https://www.youtube.com/watch?v=DV1SPdMVf8Y
[9] https://www.bloomberg.com/news/articles/2020-09-23/how-the-coronavirus-affected-biking-in-u-s-cities
[10] https://en.wikipedia.org/wiki/Phase-out-of-fossil-fuel-vehicles
[11] https://www.businessinsider.com/cities-going-car-free-ban-2018-12oslos-city-center-is-on-its-way-to-becoming-car-free-but-not-all-politicians-are-pleased-4
[12] https://www.divvybikes.com/system-data
[13] https://www.citibikenyc.com/system-data
[14] https://www.capitalbikeshare.com/system-data
[15] https://www.wmata.com/initiatives/ridership-portal/
[16] https://bikeshare.metro.net/about/data/
[17] https://www.metro.net/news/research/