Feel the breeze: A comprehensive study of bike rentals in NYC

Tuan Anh Nguyen - Daniel Blocchi

**Abstract— Citi Bike is the largest bike share program in NYC. Citi Bike collects data on their consumers. This information canS be used to find trends in renters among age, sex and hotspots in locations. Leveraging this information can be used to refine marketing techniques, improve customer satisfaction, and product deployment. Our team explored the trends in the data, and in this paper we will explain how we did it and show the extrapolated useful information which can result in business optimization.**

1. INTRODUCTION

H

istorically biking has been on a linear upward trend ever since the conception, especially in crowded cities where traffic is a problem and the travel distance between locations is rather small. As global climate change enters the fore mind of individuals, the rising cost petrol, and focus on physical health increases it is likely to see the increase in alternate modes of transports, IE: bicycles. New York City is known for it’s over crowdedness and close proximities. Because of this bikes have been seen as a relatively cost efficient, green, and effective mode of transportation, maybe even above motor transport.

The information collected by Citi Bike is plentiful and will provide a lot of insight into the bike use over the last year and will give a comprehensive and accurate look into the trends of this type of transport. When looking into the data the hardest part will be to figure out the distance between travel points and factor that into the calculations but can hopefully provide more insight into the reasons for the trends discussed in the following pages.

1. Dataset and Features

The dataset was provided by Google Cloud Platform[1] and presents a reasonable amount of records over the years of 2015-2021. However half of the year of 2015 and 2021 and not complete for different reason so they have been excluded from the data. The remaining 1,638,153 records are used in the following report. The features provided in the dataset are as followed:

1. Trip Duration
2. Start Time
3. End Time
4. Station Start ID
5. Start Station Name
6. Start Station Latitude
7. Start Station longitude
8. End Station ID
9. End Station Name
10. End Station Latitude
11. End Station Longitude
12. Bike ID
13. User Type
14. Birth Year
15. Gender

UserType is the type of purchaser whether they are a “Subscriber” or a “Customer”. The subscriber is a person who has signed up with CitiBike, the Customer is someone who has not. Data for the customer is sometimes missing whereas the Subscribers generally have all data fields filled.

1. Methods

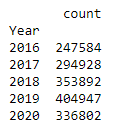
The dataset came portioned by months in .csv files nested in a year folder. The data was manually extracted then all placed into one folder. The data was imported into the IDE using Dask.

Pandas and Dask are very similar but Dask worked well to import our data into the format we desired.

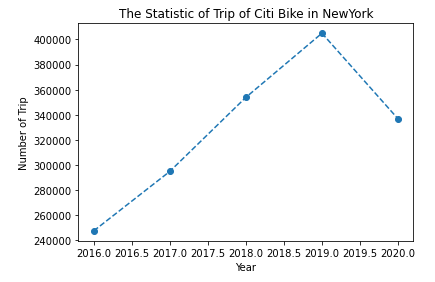
Once the data was properly imported and in a workable format the first step was to properly format the date and time. The Start Time and End Time features both contain a date (Month/Day/year) followed by an exact time of bike retrieval and return. This needed to be split into features of time/year/month to properly analyze the data and create graphs when the time wasn’t necessary but the year was for example. The data was portioned into separate tables to generate graphs using Seaborn and Math plot. Some pivot tables required the use of numpy arrays for special formatting. Elbow plots were created to find the most ideal grouping for k-clusters.

1. Results

The breakdown of the 1.6 million records by year is as followed:

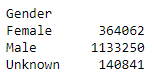


Plotting the data in a scatter plot:

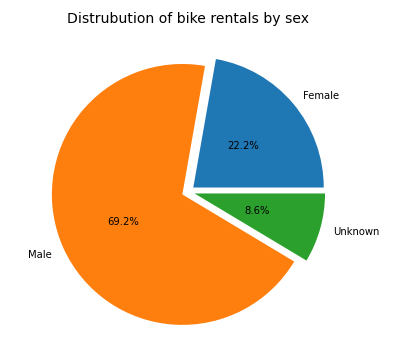


[Figure 1]

Over all years from 2016-2020 the distribution of gender in the riders, both subscribers and renters:

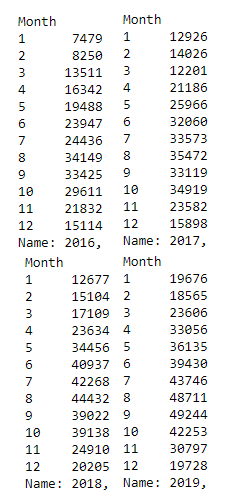


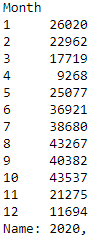
Representing this data in a pie chart:



[Figure 2]

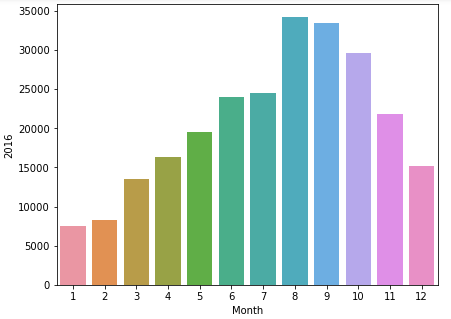
Trips by month:





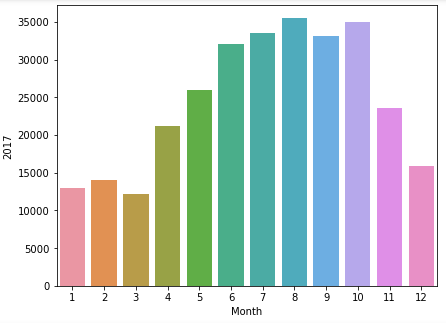
The corresponding graphs for trips by month:

2016:



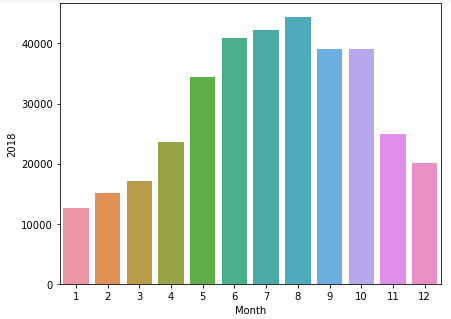
[Figure 3 i]

2017:



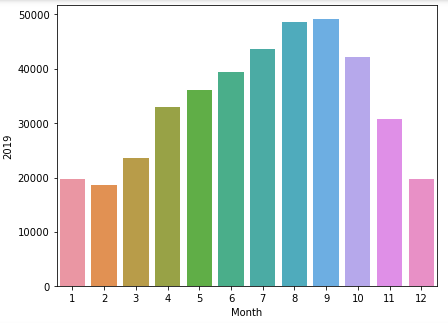
[Figure 3 ii]

2018:

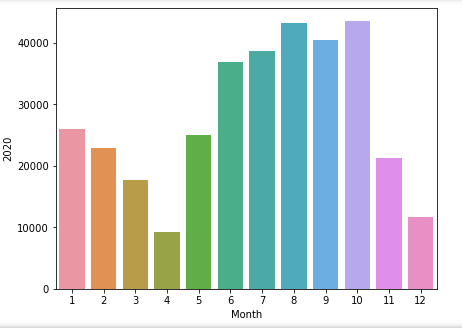


[Figure 3 iii]

2019:

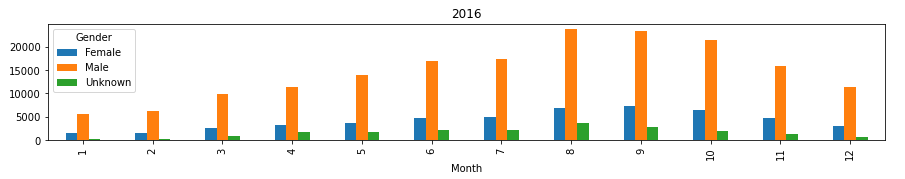


[Figure 3 iv]

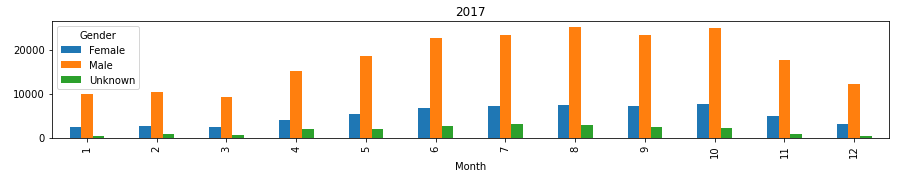


[Figure 3 v]

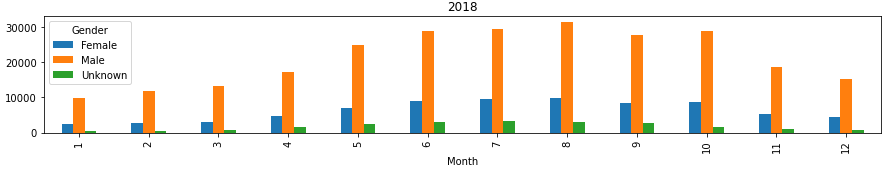
Monthly breakdowns by year by gender:



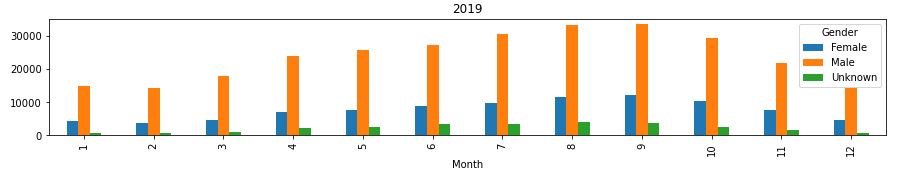
[Figure 4 i]



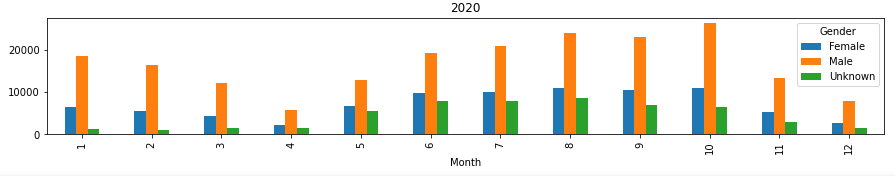
[Figure 4 ii]



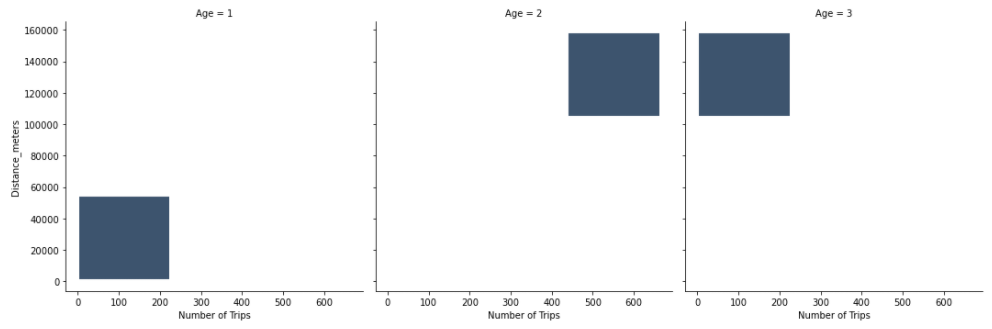
[Figure 4 iii]



[Figure 4 iv]



Teenagers, Middle-aged and Elderly number of trips and distance traveled graph:



[Figure 5]

1. Diagnostics

During the processing of some of the data it was realized that a particular number of records have the birth year of 1969. The distance traveled by that particular group is also somewhat suspicious as the data does not seem to correlate with trends spotted in the years before and after. It is believed that this may be some teenagers having some fun with the bike rental data when ordering their bikes. Since it was impossible to know the ratio of actually rides with the birth year 1969 the data was included in our graphs but is seen as a very obvious outlier.

Working through the data and beginning with counts of gender and age it was immediately realized that men use the service more than double that of women. [See Figure 2]With the sport of biking being pre-dominantly male it is likely because of the wardrobes of men and women. In most areas but especially large cities like NYC the dressing of men and women seems to be largely of the following categories. Men with suits and flat bottom shoes, and women in skirts/dressed with raised shoes. It is clear to see how riding a bike is much more difficult with a raised shoe; it almost may be difficult riding a bike with a raised seat wearing a skirt or a dress of some sort. The initiative to encourage more female riders is apparent with the increase in lowered seats and rounded handle bars which may be more suited for females in the above mentioned garments. This is was even more clear as the CitiBike’s website features a female sitting on a low rise bike wearing a dress, assumedly encouraging and suggesting that biking can be suitable for those who still want to dress in their work cloths and commute to work without the hassle of changing before and after renting the bicycle.

Another large factor which can easily be seen when observing the distribution riders over the years is the effect that COVID-19 has had on the industry. When lockdowns occurred, work buildings closed, and people were urged to stay indoors we notice a sharp decline in the amount of riders when the expected trend was to see a steep incline. [See Figure ] This can be seen at a higher resolution in [Figure 3 v] where in February, April, and May we begin to see a downward trend when all the other years [See Figure i,ii,iii] would have suggested otherwise. To continue this point further we can see that both men and women equally restricted they’re travel during these months in [Figure 5 v]

1. Conclusions

Refrences