**Harry McKaig**

Jersey City Citibike Bikeshare Pre and Post Covid Analysis

Technical Write Up

OVERVIEW

The project goal was to compare and contrast rider patters and changes is bikeshare demand in a pre and post covid environment The data set we are looking at is publicly available system data files from the [Citibike System Data](https://citibikenyc.com/system-data) website. Using this data, I analyzed activity levels, defined customer types, and created definitions for supply and demand based on rider utilization. With the growing sales and popularity of personal bicycles and with cities investing more in bicycle safety and infrastructure, coupled with more flexible working hours and more personal commuting time, I hypothesized that the growing popularity would have a positive impact on bikeshare participation.

OVERVIEW OF FINDINGS

* Usage is up: Bikeshare program is seeing a Compound Annual Growth Rate (CAGR) of 8% per year since 2019.
* Increase in Range: Total Radius (range) of stations was 2.8 Miles, up .7 miles from 2.1 Miles in 2019. (25%) increase.
* Increase in stations: 33 new stations since 2019. New stations now account for 44% of all new rides. New stations are balancing the demand by improving range and availability.
* Riders are traveling greater distances: Total Radius (range) of stations was 2.8 Miles, up .7 miles from 2.1 Miles in 2019. (25%) increase.
* Travel Time is up 50%: driven by number of leisure rides but also effected by the increase range and station availability.
* Increasing Number of Non Subscribers: gaining 20% share of ridership (accounting for 30% of all rides) with increased weekday and afternoon rides during the summer months. *Further investigation into ride time being down 27%, look at effect of price rates.*
* Electric Bikes: Introduced in May 2020 and already accounting for 14% of bikes and 30% of total rides!

OVERVIEW OF PROJECT RECOMMENDATIONS

* Bikeshare is growing at 8% per year: The increase in demand will be about 150k new trips by 2024
  + With the average stations doing 10k trips per year, we recommend adding 15 new stations in 24 months
* With Demand increasing near popular transit centers
  + We recommend expanding station capacity at current locations (in pink)
* Non-Subscribers and Leisure Rides
  + Convert non-subscribers with focus on summer, weekday rides with extended use periods 1 hour and 20 minutes <.
* Electric Bikes: 8 months new and 30% of total trip share!
  + Add e-Bikes to the fleet as cost vs benefit options. Ave new e-bike retails for $1,000 (+ cost, effort and time to maintain and store)
  + Bikeshare offers cost-benefit: Offers 4,600 trips vs cost of new bike
  + Ave Trip (.8miles) x .27c Mile = 4,600 trips for equal value

Table of Contents

* Project Summary
* Project Guidelines
* Project Deliverables
* Data Process
  + Overview
  + Limitations
  + Definitions
  + Gathering and Data preparation steps

1. Project Summary
   1. Looking exclusively at the Jersey City bicycle share data from 2019 (full pre pandemic year) to the most recent 2022 (full year of data), compare and contrast patterns and discernable metrics and provide insights in the following areas.
      1. Provide insights on the changes in demand, assessing
      2. Increase in ridership
      3. Station and Location Demand
      4. Average Time and Distance of Rides
      5. Changes in Subscription and Non Subscription frequency
2. Project Guidelines
   1. Create workable dataset and upload into Tableau for analysis, dashboard creation and story presentation
   2. Create clear recommendations on how to create and fulfill increased demand (demand planning)
3. Project Deliverables: In Tableau

Introduction

* Set context – investor persona & consulting team
* Specify goals and questions for analysis
* Data – sources, description, limitations or assumptions

Analysis

* 4-5 analytical questions or data story points, highlight insights
* Secondary support: market context, stats, images, maps, etc.

Conclusion

* Summary of data observations
* Recommendations, based on observations
* Next Steps, along with requirements for follow-on analysis

1. Data Process
   1. Overview of Procedure
      1. Citibike Data:
         1. Download, clean and load Citi bike Trip Histories to load into SQL server (local host)
         2. Trip data is available in zipped format by month (data dictionary located in 4.5
      2. Load historical weather information from Central Park station to draw correlations around wind, precipitation, temperatures
      3. Load NYC open data:
         1. Bicycle Counts, to determine overall bicycle volume by day and its impacts on ridership
         2. Map Safety Improvement corridors, established for safer bicycling and pedestrian mobility
         3. Import NYC daily ridership counts, aggregates all subway, bus and train volumes per day
   2. Data Limitations
      1. Using the Citibike Open Data for Jersey City, we have aggregated the information for insights. Due to changes in reporting, there are limitations to our ability in drawing insights.
         1. 2019 and 2022 have changes in reporting types
            1. Bike\_Id, birth\_year and gender columns were removed
            2. User\_types went from

2019: User Type (Customer = 24-hour pass or 3-day pass user; Subscriber = Annual Member)

2022: member = subscriber, casual = non\_subscriber

* + 1. Citibike already processed data 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).
       1. What we can answer?
          1. Total volume (count of trips)
          2. Seasonality: when these trips took place
          3. Demand: combine counts and times to estimate demands
          4. Station Locations: can determine the needs at each location
          5. Member Types: Whether a rider was a subscriber or single use pass.
          6. Bike Type: Classic vs Electric bicycle
          7. Commuter: if they start and end locations are different, we can determine the length of travel and assume that the user was attempting to reach a destination
          8. Leisure Rider: the opposite can be defined, our user began and ended at the same location, we are making the assumption the ride was for pleasure and not with an end destination.
       2. What we can NOT answer?
          1. Because there is no unique rider id's we are unable to gain insights on the frequency of rides taken by individuals
          2. Gender, birth year and other demographic information about the user is unavailable or been removed from reporting.
          3. For Leisure riders (returning to beginning start station), we are unable to track the distances.
          4. Subscription and Membership: Because of plan changes and partnerships, we are unable to identify Lyft Memebers, Day and Weekend Pass purchases and other descriptive types of purchases. This limits our ability to make financial and subscription based suggestions.
          5. Financial information is non-descriptive and incomplete from Citibike operating reports. Because of this, financial information or assumptions cannot be confidently communicated
  1. Data Definitions
     1. Start\_station\_id & Station\_name = unique identifier where a bike was unlocked and the trip began.
     2. End\_station\_id & Station\_name= unique identifier where a bike was locked and returned, ending the ride
     3. Geographical Location: latitude and longitude of the station\_id and station\_name
     4. TimeStamps: calendar and time stamp of the beginning and end of a trip. With this we can confirm the date of the trip and the length of time of that trip.
     5. Member Type: Whether the trip was taken by a 1x or single use purchase or trip taken by a subscribing member
     6. Bike Types: Whether the trip was on a classic or electric bicycle.
     7. *Important Note:* Station ID were incomplete in some data, use station\_names as unique identifier for all JOINS and COUNTS
  2. Gathering Data & Data Preparation
     1. Download, clean and load Citibike Trip Histories to load into SQL server (local host)
     2. Trip data is available in zipped format by month
     3. Create Table with Types, Copy monthly csv from local directory
     4. Create a table to match types for 2019 and 2022 data

create table cb20XX\_MON\_load(

ride\_id text,

rideable\_type text,

started\_at timstamp,

ended\_at timestamp,

start\_station\_name text,

start\_station\_id text,

end\_station\_name text,

end\_station\_id text,

start\_lat numeric,

start\_lng numeric,

end\_lat numeric,

end\_lng numeric,

member\_casual text

)

* + 1. Total Data available on upload:
       1. 2019 total count = 393,122
       2. 2022 total count = 31,585,406
       3. TOTAL count = 31,978,528
    2. Because of the change in data format, I ran a query on each to take only data that could be compared.
       1. Removed Nulls in the start\_station\_name and end\_station\_name: if either is missing, trip is incomplete and skus data
       2. Removed any decimal using substring on the station\_id. Formatted as text with 2022 Jercey city stations including Varchar
       3. Converted 2022 member\_casual rider types to the 2019 format of Subscriber and Non Member.
       4. Count of current data set
          1. 2019 total count = 393,122
          2. 2022 total count = 23394797 (31,585,406 dif of 7,797,487)
          3. TOTAL count = 23,787,919

|  |
| --- |
| CREATE TABLE citibike\_19\_20\_final\_cleaned AS  SELECT  SUBSTRING(start\_station\_id, 0, STRPOS(start\_station\_id, '.')) as start\_station\_id, -- text\_type has decimals, rounded to integers, Jersey City data is text  start\_station\_name,  started\_at as start\_time,  SUBSTRING(end\_station\_id, 0, STRPOS(end\_station\_id, '.')) as end\_station\_id, -- text\_type has decimals, rounded to integers, Jersey City data is text  end\_station\_name,  ended\_at as stop\_time,  CASE WHEN member\_casual ILIKE 'member' THEN 'Subscriber' --adjusted to create new category member = subscriber, else non-member  WHEN member\_casual ILIKE 'casual' THEN 'Non Member'  ELSE 'Unknown'  END AS membership  FROM public.citibike\_2022  WHERE  rideable\_type NOT ILIKE 'electric\_bike' --electric bikes introduced May 2022 and is incomplete for comparison, pulled out to examine seperately  OR start\_station\_name NOT LIKE null --any data missing from star or end station is incomplete trip  OR end\_station\_name NOT LIKE NULL  UNION ALL  SELECT  SUBSTRING(start\_station\_id, 0, STRPOS(start\_station\_id, '.')) as start\_station\_id, -- text\_type has decimals, rounded to integers, Jersey City data is text  start\_station\_name,  start\_time,  SUBSTRING(end\_station\_id, 0, STRPOS(end\_station\_id, '.')) as end\_station\_id, -- text\_type has decimals, rounded to integers, Jersey City data is text  end\_station\_name,  stop\_time,  membership --2022 data was aliased in orde to match this value member/non member  FROM public.citibike\_2019  WHERE  start\_station\_name NOT LIKE null --any data missing from star or end station is incomplete trip  OR end\_station\_name NOT LIKE NULL |

* + 1. Create Master Stations List
       1. Combined name, id, lat and lon of TTL stations from 2019 and 2022 dataset
          1. Grab all start\_station\_name, id, lat, long
          2. remove duplicated (group)
          3. -spit out to excel and change types to below. Round lat and lon to 4 decimals

|  |
| --- |
| CREATE TABLE master\_station\_list\_geo (  station\_name text,  station\_id text,  latitude numeric,  longitude numeric) |

* + - * 1. import CSV
        2. Total stations = -1858 stations
    1. Add Manhattan distance calculations to table using master\_station\_list JOIN
       1. The Manhattan distance between two vectors (city blocks) is equal to the one-norm of the distance between the vectors. The distance function (also called a “metric”) involved is also called the “taxi cab” metric.
       2. AKA measures base and height vs hypotenuse. Accounts for not being able to travel in a straight line
       3. Though the earth is flat, the small distance between points means we do not need to account for the curvature of the earth and can work in 2D plane or flat geometry.
       4. Lat and Longitude is expressed in Meters
       5. 1000m = 1 km (multiply output by 1000 to get kilometers)
       6. .06214 = 1km
       7. <https://www.101computing.net/manhattan-distance-calculator/>
       8. Longitude = X and Latitude = Y
       9. *Calculation* = [ (|x1-x2| + |y1-y2|) \* 100 = distance in kilo meters ] / 1.60934 = distance in miles
       10. References: https://www.movable-type.co.uk/scripts/latlong.html
       11. **Manhattan Distance:** 
           1. Euclidean vs Manhattan visualized

Chart, line chart

Description automatically generated A picture containing diagram

Description automatically generated

|  |
| --- |
| CREATE TABLE citibike\_final\_geo as  SELECT  start\_station\_id,  start\_station\_name,  start\_lat,  start\_long,  start\_time,  end\_station\_id,  end\_station\_name,  end\_lat,  end\_long,  stop\_time,  membership,  ride\_time,  SQRT( POWER(start\_long - end\_long, 2) + POWER(start\_lat - end\_lat, 2)) \* 100 as euc\_dist,  ROUND((ABS(start\_long- end\_long) + ABS(start\_lat-end\_lat)\*1000)\* 0.06214,4) as man\_dist  FROM public.geo\_table |

* + 1. Final table = citibike\_final\_geo
       1. Final Total Records = 23,937,325
    2. Loading other helpful data sets
       1. NYC OPEN: provides open source data on a variety of metrics: <https://www.nyc.gov/html/dot/html/about/datafeeds.shtml>
          1. Daily MTA Daily Ridership Count: Used for 2022 data (began in 2020)

<https://data.ny.gov/Transportation/MTA-Daily-Ridership-Data-Beginning-2020/vxuj-8kew/data>

* + - * 1. DOT Bicycle Counts:In 2014, NYC DOT began installing automated bicycle counters at strategic locations throughout New York City. Automated counters collect data through all seasons and weather conditions. The continuous nature of automated data collection provides insights into the effects that time of day, temperature and precipitation have on cyclist travels. Automated bicycle counts data is counted via inductive loops installed in the roadway. Each counter records cyclist volume and direction of travel Count data is collected continuously and recorded in 15- minute increments. <https://data.cityofnewyork.us/Transportation/Bicycle-Counts/uczf-rk3c>
        2. NOAA: National Centers for Environmental Data: downloaded historical weather data for NYC by day. <https://www.ncei.noaa.gov/cdo-web/datasets/NORMAL_DLY/locations/CITY:US360019/detail>
        3. SIP Corridors

Street Improvement Projects (SIPs) (corridors and intersections)  
Safety-oriented engineering improvements that use multiple treatments (signals, markings,  
concrete etc) on both corridors and intersections. Improvements are generally aimed at better  
organizing traffic, improving travel times, creating shorter, safer pedestrian crossings, and safe  
routes for bicycle travel.

https://www.nyc.gov/html/dot/downloads/pdf/vision-zero-view-metadata.pdf

* 1. Data Dictionary

|  |  |  |
| --- | --- | --- |
| citibike\_data |  |  |
| Column | Type | Notes |
| start\_time | timestamp without time zone | timestamp of bike being un-dicked |
| end\_time | timestamp without time zone | timstemp of bike being docked (completing ride) |
| start\_station\_id | integer | unique station id of starting location |
| end\_station\_id | integer | unique station id of ending location |
| bike\_id | integer | unique for every bike purchased and operation. Continious, no id is used 2c, even after taken out of circulation |
| user\_type | text | User Type (Customer = 24-hour pass or 3-day pass user; Subscriber = Annual Member) |
| birth\_year | integer | Gender (Zero=unknown; 1=male; 2=female) |
| gender | integer | Year of Birth |
| member\_casual | text | member = subscriver, casual = daily use |
| rideable\_type | text | classic & docked = dockable standard bike / e-bike (new may 2022) = electric bike |
|  |  |  |
| citibike\_stations |  |  |
| Column | Type | Notes |
| id | integer | Misc length |
| latitude | numeric | latitude in degrees (m) |
| longitude | numeric | longitude in degress (m) |
| name | text | has special characters but is descriptive on locations |
| docks | integer | misc length |
|  |  |  |
|  |  |  |
| NOAA YYY LGA weather data |  |  |
| Column | Type | Notes |
| WND | numeric | WND 09036/45: peak wind during the past hour of 36 knots from 090° occurred at 45 minutes after the hour. |
| DATE | 12/31/15 | use to link with dates on citibike |
| TMAX | numeric | temp max in F |
| TMIN | numeric | temp min in F |
| AVE | numeric | will custom this to be ave of max\_min |
| PRCP | numeric | any condensed atmospheric water (ice or snow) |
|  |  |  |
|  |  |  |
| MTA Daily ridership |  |  |
| Column | Type | Notes |
| Date | timestamp | date available by day |
| Subways: Total Estimated Ridership | numeric | Estimates ridership of Subways |
| Buses: Total Estimated Ridership | numeric | Estimates ridership of street busses |
| LIRR: Total Estimated Ridership | numeric | Estimate ridership for trains entering manhattan from Long Island (LIRR) |
| Metro-North: Total Estimated Ridership | numeric | Estimate ridership for trains entering manhattan from Westshester (Metro North) |
| Access-A-Ride: Total Scheduled Trips | numeric | Free oublic Access a ride - 311 |
| Bridges and Tunnels: Total Traffic | numeric | Any vehicle, pedestrian and cycling traffic |
|  |  |  |
|  |  |  |