Citibike SQL Project

Author: Mingyu Lin
Date: 4/22/22

I) Introduction

The goal of this project is for me to practice my SQL and Tableau skills. To that end, I'll be using BigQuery for writing SQL queries and Tableau Public to share any visualizations created using the result sets of the queries downloaded from BigQuery.

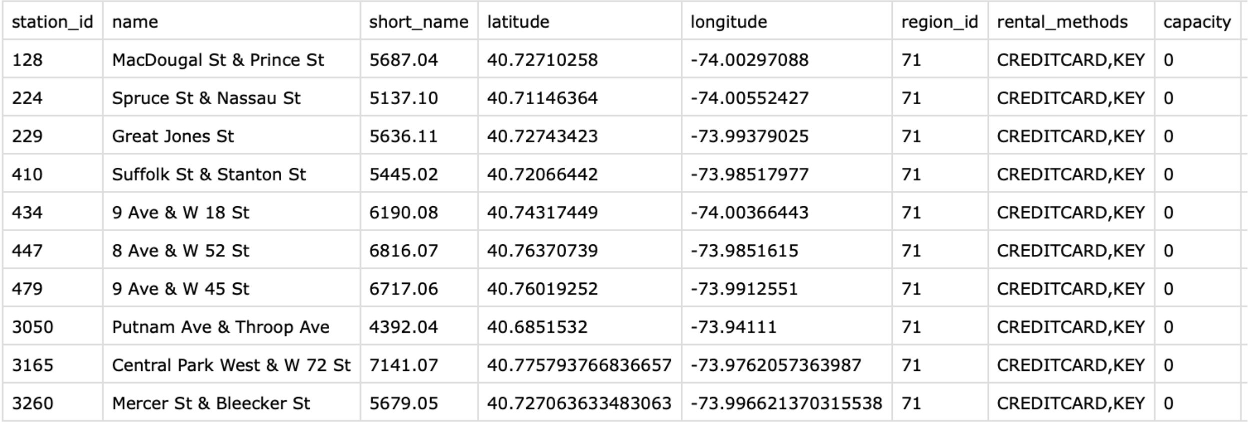
I'll be using the `new_york_citibike` dataset from BigQuery public data for this project. There are two tables in the dataset: `citibike_stations` and `citibike_trips`.

Here is the schema for the `citibike_stations` table:

Field name	Type	Mode	Description
station_id	INTEGER	REQUIRED	Unique identifier of a station.
name	STRING	NULLABLE	Public name of the station.
short_name	STRING	NULLABLE	Short name or other type of identifier, as used by the data publisher.
latitude	FLOAT	NULLABLE	The latitude of station. The field value must be a valid WGS 84 latitude in decimal degrees format.
longitude	FLOAT	NULLABLE	The longitude of station. The field value must be a valid WGS 84 longitude in decimal degrees format.
region_id	INTEGER	NULLABLE	ID of the region where station is located.
rental_methods	STRING	NULLABLE	Array of enumerables containing the payment methods accepted at this station.
capacity	INTEGER	NULLABLE	ANumber of total docking points installed at this station, both available and unavailable.
eightd_has_key_dispenser	BOOLEAN	NULLABLE	
num_bikes_available	INTEGER	NULLABLE	Number of bikes available for rental.
num_bikes_disabled	INTEGER	NULLABLE	Number of disabled bikes at the station.
num_docks_available	INTEGER	NULLABLE	Number of docks accepting bike returns.

num_docks_disabled	INTEGER	NULLABLE	Number of empty but disabled dock points at the station.
is_installed	BOOLEAN	NULLABLE	Is the station currently on the street?
is_renting	BOOLEAN	NULLABLE	Is the station currently renting bikes?
is_returning	BOOLEAN	NULLABLE	Is the station accepting bike returns?
eightd_has_available_keys	BOOLEAN	NULLABLE	
last_reported	DATETIME	NULLABLE	Timestamp indicating the last time this station reported its status to the backend, in NYC local time.

The table has 18 columns and 1584 rows. Here is a preview of the data:



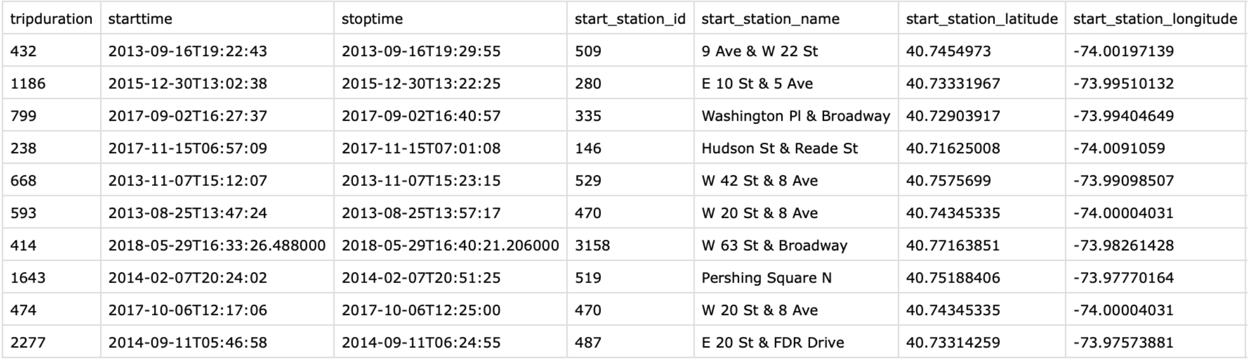
station_id	name	short_name	latitude	longitude	region_id	rental_methods	capacity
128	MacDougal St & Prince St	5687.04	40.72710258	-74.00297088	71	CREDITCARD,KEY	0
224	Spruce St & Nassau St	5137.10	40.71146364	-74.00552427	71	CREDITCARD,KEY	0
229	Great Jones St	5636.11	40.72743423	-73.99379025	71	CREDITCARD,KEY	0
410	Suffolk St & Stanton St	5445.02	40.72066442	-73.98517977	71	CREDITCARD,KEY	0
434	9 Ave & W 18 St	6190.08	40.74317449	-74.00366443	71	CREDITCARD,KEY	0
447	8 Ave & W 52 St	6816.07	40.76370739	-73.9851615	71	CREDITCARD,KEY	0
479	9 Ave & W 45 St	6717.06	40.76019252	-73.9912551	71	CREDITCARD,KEY	0
3050	Putnam Ave & Throop Ave	4392.04	40.6851532	-73.94111	71	CREDITCARD,KEY	0
3165	Central Park West & W 72 St	7141.07	40.775793766836657	-73.9762057363987	71	CREDITCARD,KEY	0
3260	Mercer St & Bleecker St	5679.05	40.727063633483063	-73.996621370315538	71	CREDITCARD,KEY	0

Here is the schema for the `citibike_trips` table:

Field name	Type	Mode	Description
tripduration	INTEGER	NULLABLE	Trip Duration (in seconds)
starttime	DATETIME	NULLABLE	Start Time, in NYC local time.
stoptime	DATETIME	NULLABLE	Stop Time, in NYC local time.
start_station_id	INTEGER	NULLABLE	Start Station ID
start_station_name	STRING	NULLABLE	Start Station Name
start_station_latitude	FLOAT	NULLABLE	Start Station Latitude
start_station_longitude	FLOAT	NULLABLE	Start Station Longitude
end_station_id	INTEGER	NULLABLE	End Station ID
end_station_name	STRING	NULLABLE	End Station Name
end_station_latitude	FLOAT	NULLABLE	End Station Latitude
end_station_longitude	FLOAT	NULLABLE	End Station Longitude
bikeid	INTEGER	NULLABLE	Bike ID
usertype	STRING	NULLABLE	User Type (Customer = 24-hour pass or 7-day pass user, Subscriber = Annual Member)

birth_year	INTEGER	NULLABLE	Year of Birth
gender	STRING	NULLABLE	Gender (unknown, male, female)
customer_plan	STRING	NULLABLE	The name of the plan that determines the rate charged for the trip

The table has 16 columns and 58937715 rows. Here is a preview of the data:



tripduration	starttime	stoptime	start_station_id	start_station_name	start_station_latitude	start_station_longitude
432	2013-09-16T19:22:43	2013-09-16T19:29:55	509	9 Ave & W 22 St	40.7454973	-74.00197139
1186	2015-12-30T13:02:38	2015-12-30T13:22:25	280	E 10 St & 5 Ave	40.73331967	-73.99510132
799	2017-09-02T16:27:37	2017-09-02T16:40:57	335	Washington Pl & Broadway	40.72903917	-73.99404649
238	2017-11-15T06:57:09	2017-11-15T07:01:08	146	Hudson St & Reade St	40.71625008	-74.0091059
668	2013-11-07T15:12:07	2013-11-07T15:23:15	529	W 42 St & 8 Ave	40.7575699	-73.99098507
593	2013-08-25T13:47:24	2013-08-25T13:57:17	470	W 20 St & 8 Ave	40.74345335	-74.00004031
414	2018-05-29T16:33:26.488000	2018-05-29T16:40:21.206000	3158	W 63 St & Broadway	40.77163851	-73.98261428
1643	2014-02-07T20:24:02	2014-02-07T20:51:25	519	Pershing Square N	40.75188406	-73.97770164
474	2017-10-06T12:17:06	2017-10-06T12:25:00	470	W 20 St & 8 Ave	40.74345335	-74.00004031
2277	2014-09-11T05:46:58	2014-09-11T06:24:55	487	E 20 St & FDR Drive	40.73314259	-73.97573881

To make this interesting, let's make a scenario for this project. Let's pretend that a manager for Citibike asks me to query the dataset to gain insights about customer bike usage. In what follows, I will make some questions, query the dataset, and create data visualization for some questions.

II) Analysis

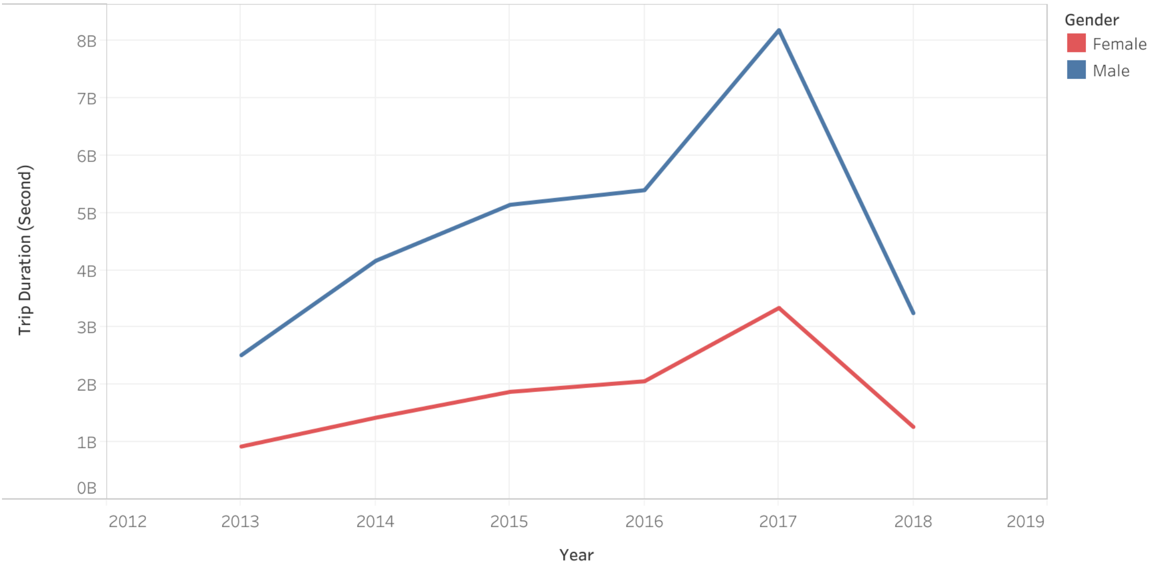
1) Who use citibike more each year, men or women?

```
SELECT EXTRACT(YEAR FROM starttime) year,
    SUM(CASE WHEN gender = 'male' THEN tripduration END) male_trip_duration_sec,
    SUM(CASE WHEN gender = 'female' THEN tripduration END) female_trip_duration_sec
FROM `bigquery-public-data.new_york_citibike.citibike_trips`
GROUP BY year
HAVING year is not NULL
ORDER BY year
```

year	male_trip_duration_sec	female_trip_duration_sec
2013	2518881585	925566755
2014	4165423023	1427968605
2015	5142383710	1878910161
2016	5397053292	2063673796
2017	8188900363	3342082773

2018	3252467350	1267776263

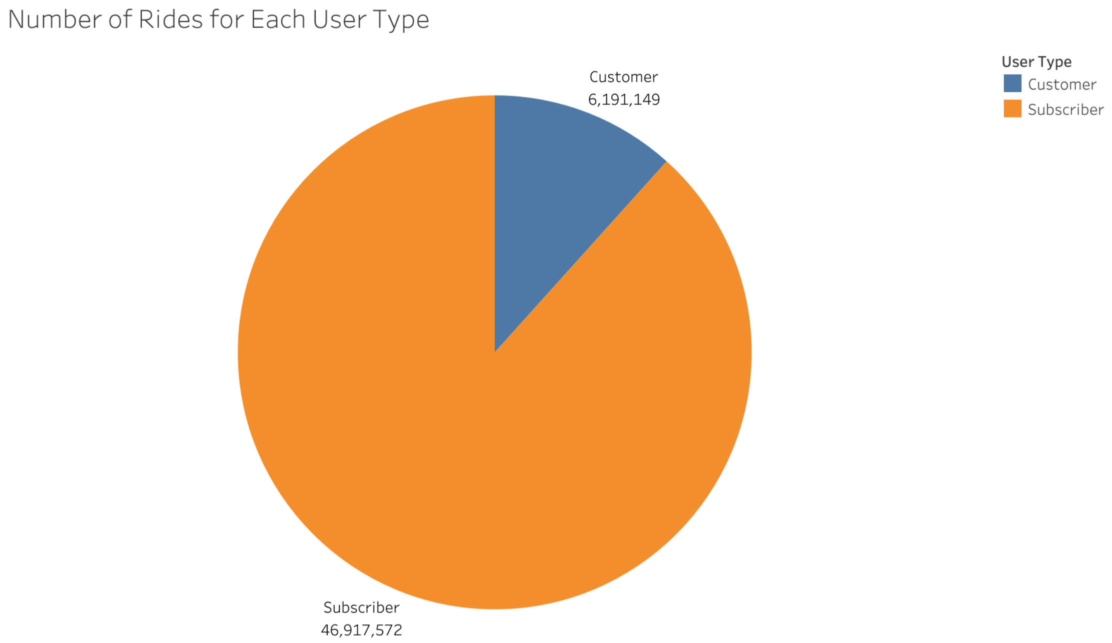
Total Trip Durations for Male and Female



2) What user type uses citibike the most overall?

```
SELECT usertype, COUNT(*) num_of_rides
FROM `bigquery-public-data.new_york_citibike.citibike_trips`
GROUP BY usertype
HAVING usertype IN ('Subscriber','Customer')
ORDER BY num_of_rides DESC
```

usertype	num_of_rides
Subscriber	46917572
Customer	6191149

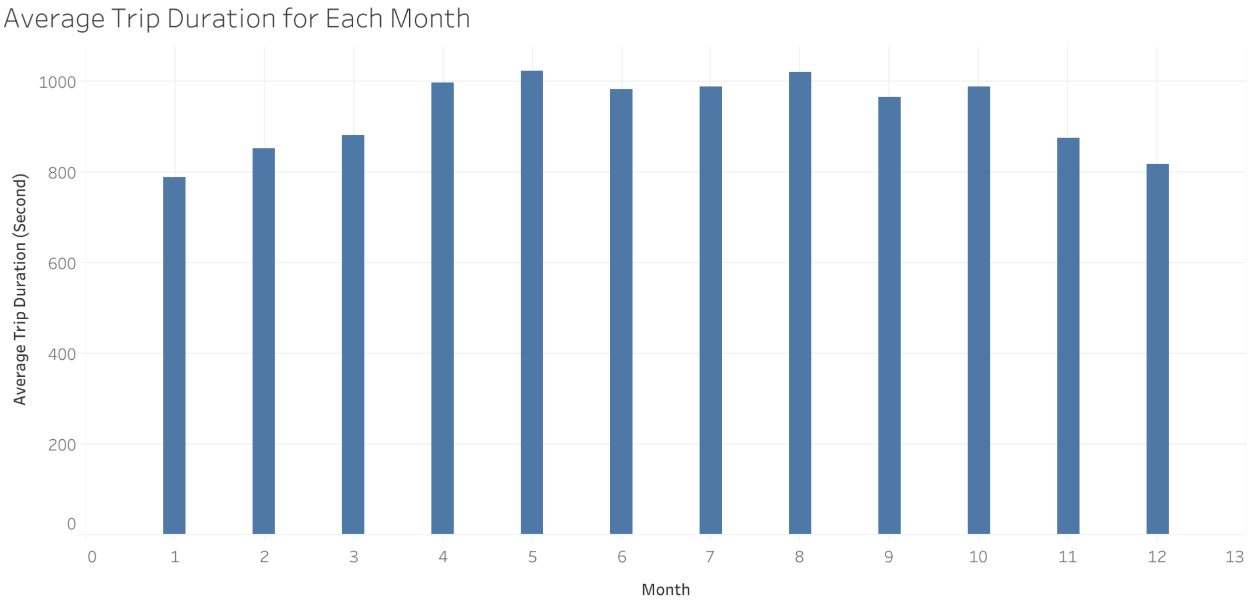


Number of Rides for Each User Type

3) What is the average trip duration per month?

```
SELECT EXTRACT(MONTH FROM starttime) month, ROUND(AVG(tripduration))
avg_trip_duration_sec
FROM `bigquery-public-data.new_york_citibike.citibike_trips`
GROUP BY month
HAVING month IS NOT NULL
ORDER BY month
```

month	avg_trip_duration_sec
1	790
2	852
3	883
4	998
5	1025
6	984
7	989
8	1022
9	967
10	990
11	876
12	817



4) What are some of the most used bikes?

```
SELECT bikeid, COUNT(bikeid) num_of_trips
FROM `bigquery-public-data.new_york_citibike.citibike_trips`
GROUP BY bikeid
ORDER BY num_of_trips DESC
LIMIT 10
```

bikeid	num_of_trips
18104	7222
15731	7146
19455	7076
17526	7030
16158	7025
17955	6988
19633	6963
20233	6962
17289	6960
17747	6948

5) What are some of the most popular routes?

```
SELECT start_station_id, end_station_id, COUNT(*) num_of_trips
```

```
FROM `bigquery-public-data.new_york_citibike.citibike_trips`
WHERE start_station_id != end_station_id OR (start_station_id = end_station_id AND
tripduration > 600)
GROUP BY start_station_id, end_station_id
ORDER BY num_of_trips DESC
LIMIT 10
```

start_station_id	end_station_id	num_of_trips
2006	2006	50964
281	281	22065
514	426	18667
435	509	17509
499	499	16409
519	492	16228
387	387	15546
435	462	15120
426	514	14353
519	477	14171

6) What are the top 10 start stations and top 10 end stations?

```
SELECT start_station_id, num_of_trips, rank_rnk ranking
FROM
    (SELECT start_station_id, COUNT(*) num_of_trips,
            RANK() OVER (ORDER BY COUNT(*) DESC) rank_rnk
        FROM `bigquery-public-data.new_york_citibike.citibike_trips`
        GROUP BY start_station_id
        HAVING start_station_id IS NOT NULL
    ) bike_rankings
WHERE rank_rnk <= 10
ORDER BY num_of_trips DESC,rank_rnk
```

start_station_id	num_of_trips	ranking
519	551078	1
497	423334	2
435	403795	3
426	384116	4
293	372255	5
402	367194	6
285	344546	7

490	330378	8
151	318700	9
477	311403	10

```
SELECT end_station_id, num_of_trips, rank_rnk ranking
FROM
    (SELECT end_station_id, COUNT(*) num_of_trips,
            RANK() OVER (ORDER BY COUNT(*) DESC) rank_rnk
        FROM `bigquery-public-data.new_york_citibike.citibike_trips`
        GROUP BY end_station_id
        HAVING end_station_id IS NOT NULL
    ) bike_rankings
WHERE rank_rnk <= 10
ORDER BY num_of_trips DESC,rank_rnk
```

end_station_id	num_of_trips	ranking
519	511019	1
497	444460	2
435	407982	3
426	399033	4
402	377854	5
293	372679	6
285	344033	7
459	323647	8
151	319866	9
477	318435	10

III) Conclusion

There are several findings and implications for the above analysis:

1) There are more men than women using the biking service from 2013 to 2018 with peak usage in 2017. It seems that the biking service got more popular over the years with an exception of 2018. This might due to the fact that there are fewer datapoints for 2018.
2) Subscribers use the biking service more than customers. Since subscribers are annual users and customers are 24-hour pass users or 7-day pass users, it makes sense that subscribers use Citibike more than customers.
3) The summer season have more riders than any other seasons. Therefore, more bikes can be added to the system in the summer months and reduced in the winter months.
4) Bike 18104 and other top used bikes should be inspected and replaced if necessary so new bikes can be added to the system.

5) The stations of the popular routes should stockpile more bikes than other stations so the stations would not run out of bike due to more demanding customers.
6) The top ranked start and end stations should be scheduled for maintenance more often than other stations.

To see the dashboard and charts created for this project, you can visit my Tableau Public page: https://public.tableau.com/app/profile/mingyu.lin/viz/Citibike_16504922319770/Dashboard1