Bike-Share Case Study

Kim Tsang

Basic Setting and Defining the Business Problem

Cyclistic, launched in 2016, is a bike-sharing company based in Chicago, USA. The company features more than 5,824 bicycles and 692 docking stations. Moreover, Cyclistic provides different types of bicycles such as reclining bikes, hand tricycles, and cargo bikes to better satisfy the needs of various customers. Customers may pick up any available bike at a docking station, ride to their destinations, and drop off the bike at another docking station.

Cyclistic has three pricing plans: single ride passes, full day passes, and annual memberships. The director of marketing believes that the future success of the company largely depends on maximizing the number of annual memberships. Therefore, it is important to understand how casual riders and annual members use the service from Cyclistic differently based on the data collected. This case study will focus on the data made available by the company from January of 2022 to December of 2022. This case study will mainly focus on answering two questions:

1. How do annual members and casual riders use Cyclistic bikes differently?
2. How to possibly convince casual riders to buy Cyclistic annual memberships?

Data

The data used in this case study can be found here: https://divvy-tripdata.s3.amazonaws.com/index.html. Specifically, 12 sets of csv files from 202201-divvy-tripdata.zip to 202212-divvy-tripdata.zip were used.

Preparation of Data

Union Files

```
In [2]: import pandas as pd
```

```
In [4]: df = pd.read_csv('2022_bike_data.csv')
```

```
In [5]: df.head(5)
```

Out[5]:

	Unnamed: 0	ride_id	rideable_type	started_at	ended_at	start_station_name	start_station_id	end_station_name	end_station_id	start_lat	st
0	0	C2F7DD78E82EC875	electric_bike	2022-01-13 11:59:47	2022-01-13 12:02:44	Glenwood Ave & Touhy Ave	525	Clark St & Touhy Ave	RP-007	42.012800	-87.(
1	1	A6CF8980A652D272	electric_bike	2022-01-10 08:41:56	2022-01-10 08:46:17	Glenwood Ave & Touhy Ave	525	Clark St & Touhy Ave	RP-007	42.012763	-87.(
2	2	BD0F91DFF741C66D	classic_bike	2022-01-25 04:53:40	2022-01-25 04:58:01	Sheffield Ave & Fullerton Ave	TA1306000016	Greenview Ave & Fullerton Ave	TA1307000001	41.925602	-87.(
3	3	CBB80ED419105406	classic_bike	2022-01-04 00:18:04	2022-01-04 00:33:00	Clark St & Bryn Mawr Ave	KA1504000151	Paulina St & Montrose Ave	TA1309000021	41.983593	-87.(
4	4	DDC963BFDDA51EEA	classic_bike	2022-01-20 01:31:10	2022-01-20 01:37:12	Michigan Ave & Jackson Blvd	TA1309000002	State St & Randolph St	TA1305000029	41.877850	-87.(

Upon preliminary inspections of the data using Python, it was found that all 12 sets of the csv files had similar data structures, namely field names, data types, number of fields, and geo roles.

Since each row or record described similar information at the same granularity, the 12 sets of data were unionized. The figure below shows the simple code to combine all the files. After combining all the csv files, the output file had a count of 5667717 rows.

```
In [1]: import pandas as pd
```

```
In [6]: import os
```

```
In [8]: print(os.path.abspath("202201-divvy-tripdata.csv"))
```

```
/Users/kkbighead/Desktop/Google_Capstone_Project/202201-divvy-tripdata.csv
```

```
In [20]: df = pd.concat(
             map(pd.read_csv, ['202201-divvy-tripdata.csv', '202202-divvy-tripdata.csv','202203-divvy-tripdata.csv',
                               '202204-divvy-tripdata.csv', '202205-divvy-tripdata.csv','202206-divvy-tripdata.csv',
                               '202207-divvy-tripdata.csv',
                               '202208-divvy-tripdata.csv', '202209-divvy-publictripdata.csv', '202210-divvy-tripdata.csv',
                               '202211-divvy-tripdata.csv', '202212-divvy-tripdata.csv']), ignore_index=True)

         new_table = pd.DataFrame(df)
```

Adding New Fields

To ensure an easier time during analysis, new fields were created using SQL. The data had information on when the ride started and ended with a “timestamp” data type. The fields of “started_at” and “ended_at” were further broken down into which days of the week the ride started and ended on (started_at_dow and ended_at_dow). In addition; month, day of the week, and hour of the trip were extracted from the “started_at” and “ended_at” fields. Finally, the length of the ride time in hours was also extracted from the original fields using the following queries.

```sql
--Adding a column called ride duration to the table
alter table bike_data
add column ride_duration interval

-- Adding data to the new ride_duration field
update bike_data
set ride_duration = (ended_at - started_at)

--Adding another column for ride duration in hours
alter table bike_data
add column ride_duration_in_hour numeric

--Adding data to the ride duration in hours column
update bike_data
set ride_duration_in_hour =
round(extract(epoch from ride_duration)/3600,2)
```

The complete list of SQL codes can be found here: https://github.com/KimHungTsang/Recent-Projects/blob/main/Bike_Project_Complete_SQL_Codes.txt

Descriptions of the fields

Each record in the data represents one trip. Each trip comprises the following unique fields.

	Field	Description
1	ride_id	Unique record id
2	rideable_type	Type of bikes
3	start_station_name	Name of the start station
4	start_station_id	Start station id

5	end_station_name	Name of end station
6	end_station_id	End station id
7	start_lat	Trip start latitude
8	start_lon	Trip start longitude
9	end_lat	Trip end latitude
10	end_lon	Trip end longitude
11	member_casual	Rider types
12	started_at_dow	Trip start day of the week
13	ended_at_dow	Trip end day of the week
14	rental_month	Month of rental
15	ride_duration	Duration of the bike ride
16	started_hour	Hour of the start of the trip
17	ended_hour	Hour of the end of the trip
18	ride_duration_in_hour	Trip duration in hours

Data Visualization

Visualization for data analysis was done using Tableau Desktop.

The interactive worksheets and dashboard can be found here: https://github.com/KimHungTsang/Recent-Projects/blob/main/Bike%20Project%20Tableau.twb

Analysis

General Summary of trips:

Ride Count Summary by Month

	casual			member		
Rental Mo..	classic_bike	electric_bike	Total	classic_bike	electric_bike	Total
January	6,974	10,585	17,559	48,093	37,157	85,250
February	8,107	11,948	20,055	51,307	42,886	94,193
March	35,387	46,137	81,524	99,052	95,108	194,160
April	47,543	66,758	114,301	119,169	125,663	244,832
May	126,075	127,931	254,006	197,971	156,472	354,443
June	169,996	168,415	338,411	236,664	163,489	400,153
July	156,095	218,905	375,000	217,078	200,355	417,433
August	128,635	203,966	332,601	215,415	211,593	427,008
September	105,375	171,496	276,871	200,767	203,875	404,642
October	61,568	134,807	196,375	151,992	197,704	349,696
November	33,052	61,834	94,886	111,549	125,414	236,963
December	12,652	30,317	42,969	60,698	76,214	136,912
Grand Total	891,459	1,253,099	2,144,558	1,709,755	1,635,930	3,345,685

From the summary table above it can be seen that June, July, August, and September were the busiest months for both casual riders and members with July being the most busy for casual riders and August for members. An initial hypothesis would be that cycling is both a more enjoyable and easier activity during warmer and non slippery weathers. The colder months of December, January, and February had much lower records of rides for both casual riders and members.

Casual riders favored electric bikes over classic bikes over all months except for June of 2022. Interestingly, members favored electric bikes only in the months of April, October, November, and December.

Percent of Total by Month for Members and Casual Riders

	casual			member		
Rental Mo..	classic_bike	electric_bike	Total	classic_bike	electric_bike	Total
January	6.78%	10.30%	17.08%	46.78%	36.14%	82.92%
February	7.10%	10.46%	17.55%	44.91%	37.54%	82.45%
March	12.84%	16.74%	29.57%	35.93%	34.50%	70.43%
April	13.24%	18.59%	31.83%	33.18%	34.99%	68.17%
May	20.72%	21.03%	41.75%	32.54%	25.72%	58.25%
June	23.02%	22.80%	45.82%	32.04%	22.14%	54.18%
July	19.70%	27.62%	47.32%	27.39%	25.28%	52.68%
August	16.93%	26.85%	43.79%	28.36%	27.86%	56.21%
September	15.46%	25.16%	40.63%	29.46%	29.92%	59.37%
October	11.27%	24.69%	35.96%	27.83%	36.20%	64.04%
November	9.96%	18.63%	28.59%	33.61%	37.79%	71.41%
December	7.03%	16.85%	23.89%	33.74%	42.37%	76.11%
Grand Total	16.24%	22.82%	39.06%	31.14%	29.80%	60.94%

When looking at the same summary table with “percent of total” calculations, we can see that casual riders accounted for 40.97% of the total rides in 2022 while members accounted for 59.03% of the total rides. The percent of total calculations were computed across each row (month) for both panes (rider type). This meant that January, February, November, and December had the biggest contrast between total rides made by the two rider categories. Since the goal is to maximize conversion rate of casual riders into members, paying more attention to these particular months would be more beneficial for the company.

Busiest days of the week

Bike Rental Frequency Aggregrated on Different Days of The Week

	casual			member			Grand Total
Started At ..	classic_bike	electric_bike	Total	classic_bike	electric_bike	Total	
Sunday	158,581	194,726	353,307	201,090	186,133	387,223	740,530
Monday	104,257	150,883	255,140	247,402	225,937	473,339	728,479
Tuesday	96,125	149,865	245,990	268,366	250,260	518,626	764,616
Wednesday	98,363	158,656	257,019	266,301	257,568	523,869	780,888
Thursday	113,837	175,719	289,556	267,455	264,806	532,261	821,817
Friday	123,126	188,188	311,314	231,809	235,277	467,086	778,400
Saturday	197,170	235,062	432,232	227,332	215,949	443,281	875,513

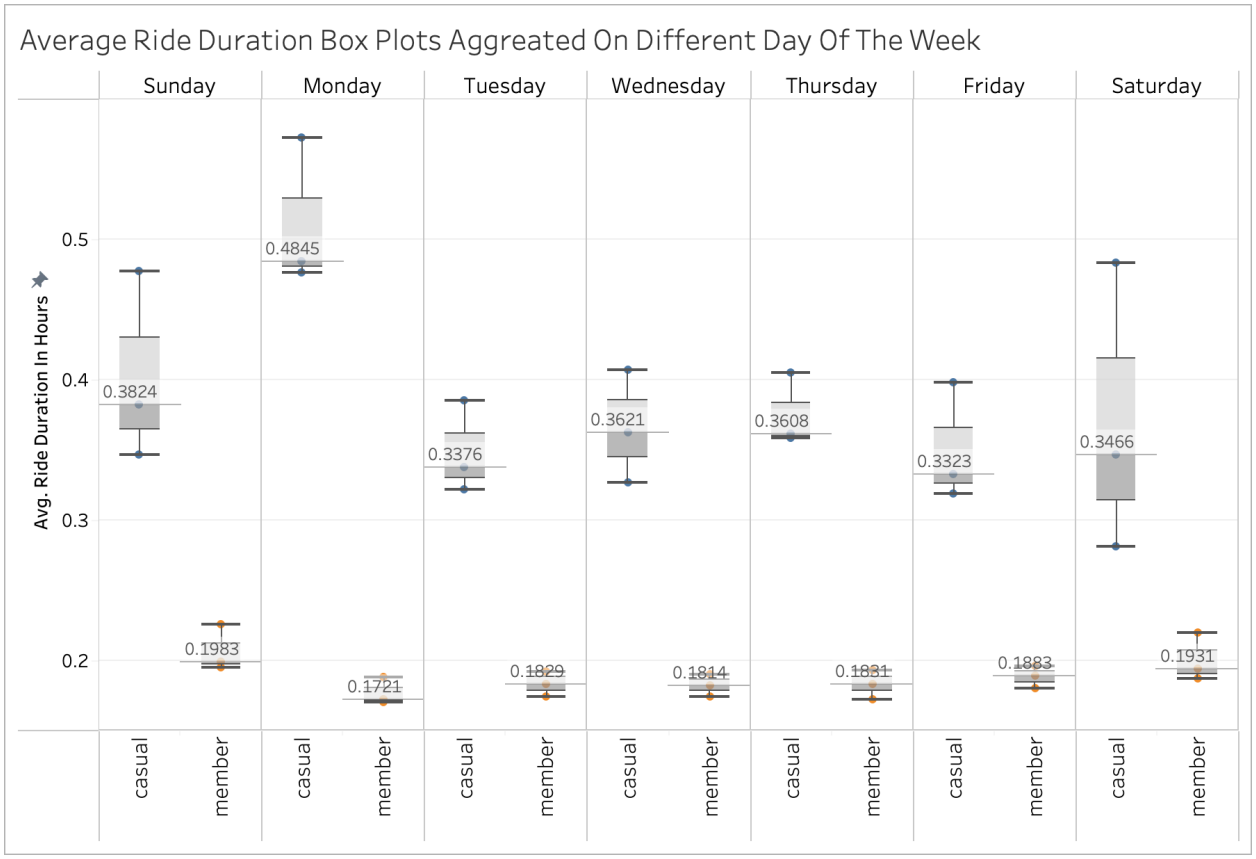
The figure above is a summary of when the bikes were picked up from a docking station in 2022; aggregated at the day of week level. The busiest days of the week for members were Tuesday, Wednesday, and Thursday, with Thursday being the most busy. On the other hand, casual riders rented bikes more during the weekend from Friday to Sunday with Saturday being the most busy. Perhaps casual riders used the bikes for leisurely activities while members used the bikes as a mode of transportation during the week.

Percentage Change From Classic to Electric Aggregrated on Different Days of the Week

	casual		member	
Started At ..	classic_bike	electric_bike	classic_bike	electric_bike
Sunday		22.79%		-7.44%
Monday		44.72%		-8.68%
Tuesday		55.91%		-6.75%
Wednesday		61.30%		-3.28%
Thursday		54.36%		-0.99%
Friday		52.84%		1.50%
Saturday		19.22%		-5.01%

Another detail to point out is that when looking at the percentage difference from classic bikes to electric bikes, casual riders favored electric bikes more than classic bikes, especially on Tuesday, Wednesday, Thursday, and Friday. Members on the other hand favored classic bikes more than electric bikes however, the differences were not nearly as much as casual riders since they were all under a 10% decrease.

Average Ride Duration



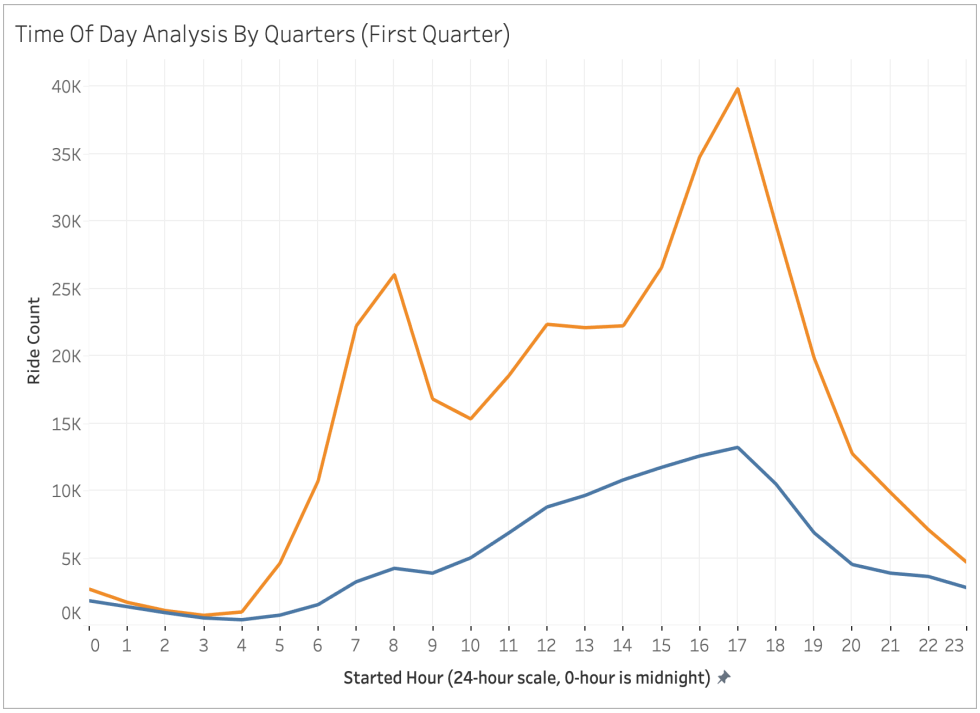
The figure above shows the box plots of the average ride duration in hours between the two different rider types (aggregated at the day of the week level for 2022).

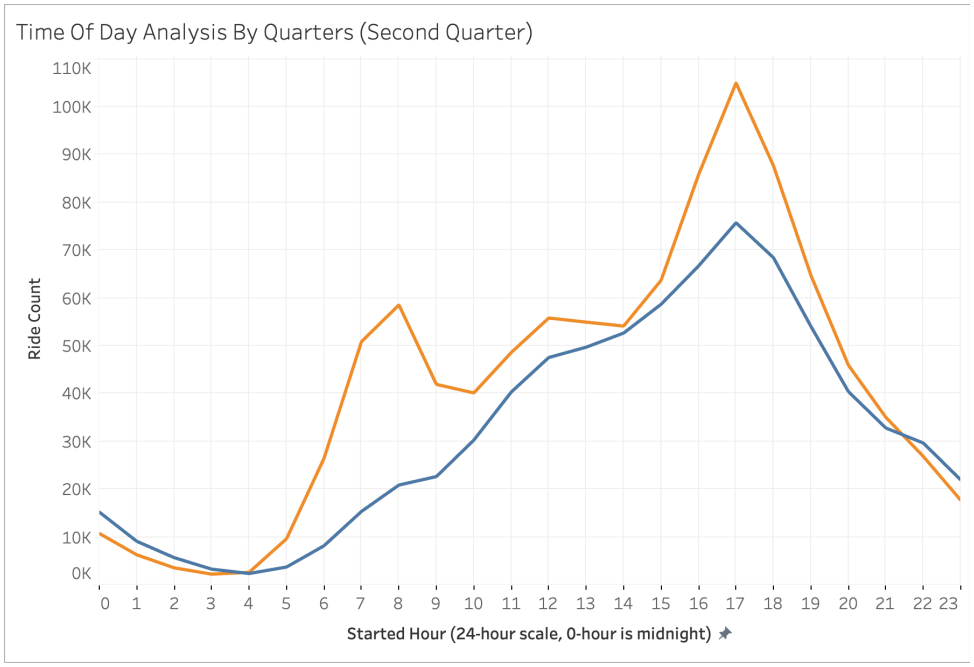
The average ride durations for casual riders were higher than members. For example, on Sundays, casual riders had a median ride time of .38 hour (22.8 minutes) and members had a .198 hour (11.88 minutes) median ride time for the distributions of their average times.

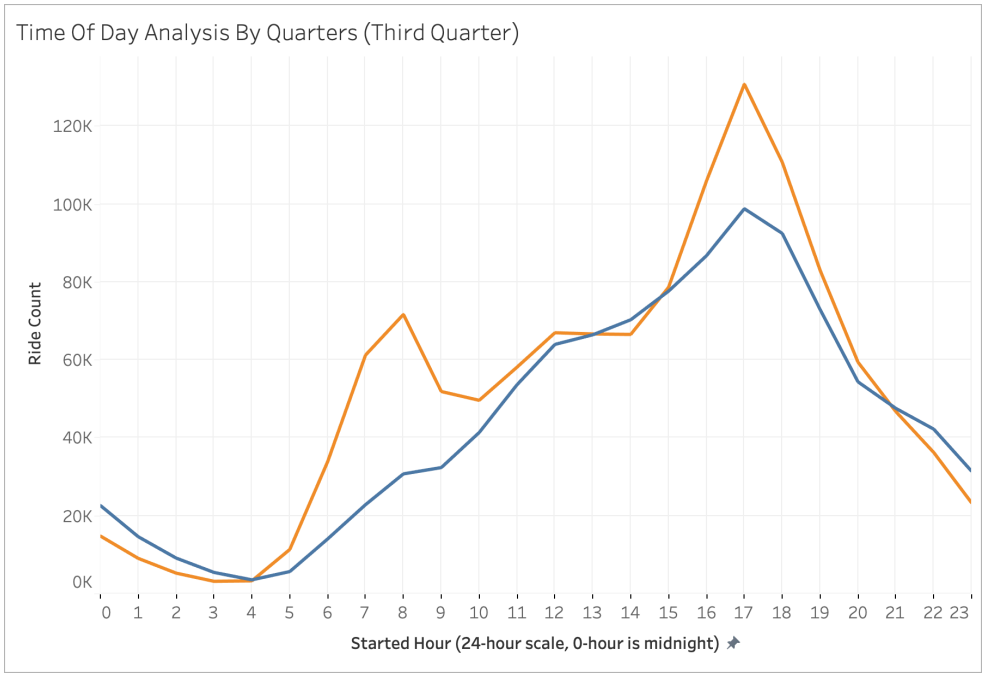
One thing to keep in mind is that since the figure was aggregated at the day of week level, the column of Sunday shows the average ride time for both rider types taken on all Sundays in 2022. Saturday, Sunday, and Monday were the three days with the highest average ride time for casual riders. On the other hand, members recorded the highest average ride time on Saturday, Sunday, and Wednesday. One similarity is clear from this part of the analysis, both types of riders went on longer rides on average on Saturday and Sunday.

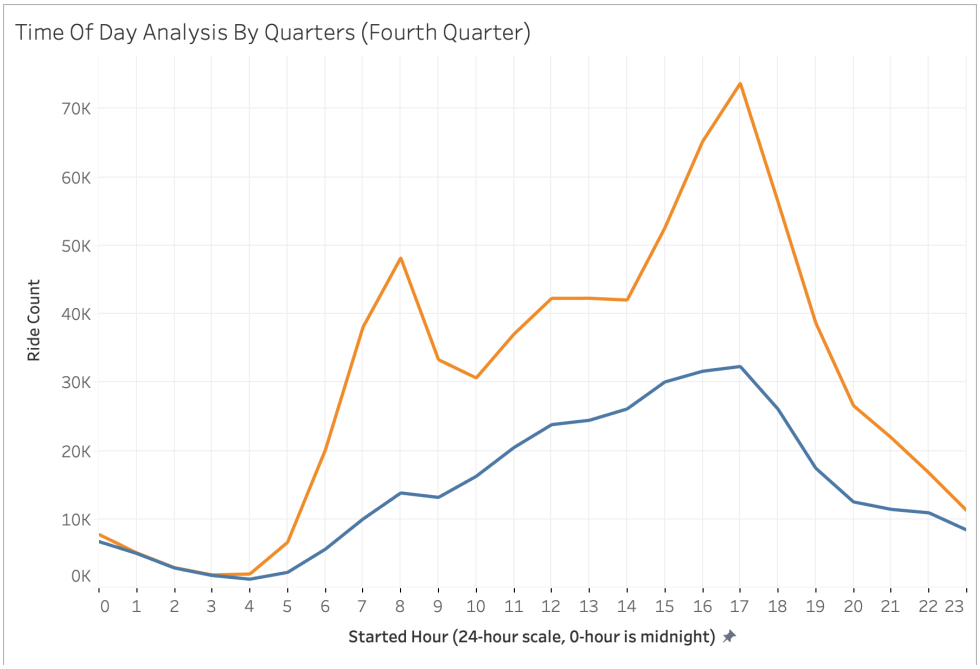
Time of Day Analysis

Are there specific times during the day when docking stations tend to be extra busy? To help answer the question better, the 2022 data was broken down into four quarters in order to examine if seasonality played a part in the time of day when customers rented the bikes.




Time Of Day Analysis By Quarters (Second Quarter)
110K
100K
90K
80K
70K
60K
50K
40K
30K
20K
10K
0K
Ride Count
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Started Hour (24-hour scale, 0-hour is midnight)


Time Of Day Analysis By Quarters (Third Quarter)
120K
100K
80K
60K
40K
20K
0K
Ride Count
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Started Hour (24-hour scale, 0-hour is midnight)



The figures above show starting ride counts (number of bikes being picked up at a docking station) for each quarter of 2022 aggregated for each hour of the day. The figures of all four quarters showed that members (orange) rented more bikes than causal riders (blue). Additionally, members showed two different peaks in the day at 8AM and 5PM consistently throughout all four quarters while casual riders only showed one peak at 5pm.

One hypothesis could be that members used the bikes to both commute to and from work while casual riders only used it to commute after work. Both rider types recorded the lowest number of starting rides during Q1 and the highest number of starting rides during Q3 of 2022. This makes sense intuitively because summer months do seem to be a “nicer time” to ride a bicycle. Another interesting detail to point out is that while the number of ride counts for members stayed consistently above casual riders, there were a few times when the number for casual riders overtook members. These instances were from around midnight to 4AM and past 9:30 PM in Q2, and midnight to 4am, 2PM, and after 9PM in Q3. The reasons for these overtakes are unknown, however it may be worthwhile to examine them further.

Top Ten Busiest Starting and End Stations

Top Ten Most Busy Starting Station

Start Station Name	
Streeter Dr & Grand Ave	75,237
DuSable Lake Shore Dr & Monroe St	41,279
DuSable Lake Shore Dr & North Blvd	40,090
Michigan Ave & Oak St	39,661
Wells St & Concord Ln	37,515
Clark St & Elm St	35,037
Millennium Park	35,005
Kingsbury St & Kinzie St	33,725
Theater on the Lake	32,976
Wells St & Elm St	31,476

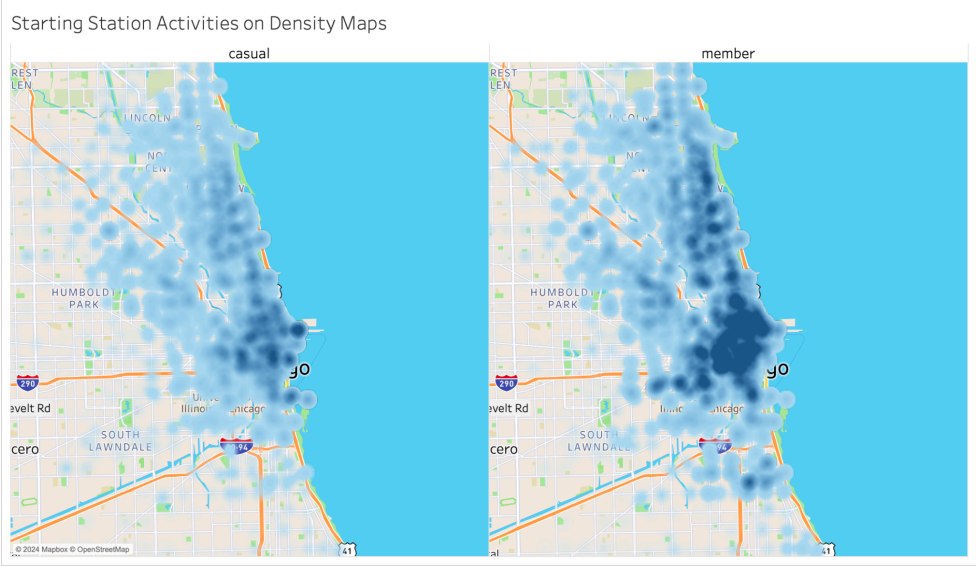
Top Ten Most Busy End Station

End Station Name	
Streeter Dr & Grand Ave	75,382
DuSable Lake Shore Dr & North Blvd	42,141
Michigan Ave & Oak St	40,127
DuSable Lake Shore Dr & Monroe St	40,125
Wells St & Concord Ln	37,421
Millennium Park	35,234
Clark St & Elm St	34,490
Theater on the Lake	32,988
Kingsbury St & Kinzie St	32,380
Wells St & Elm St	30,338

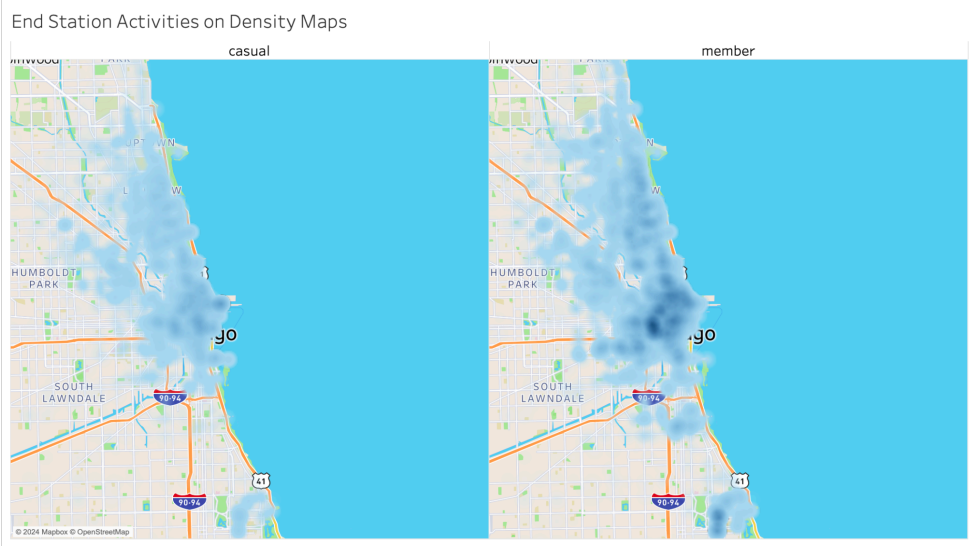
The two tables above show the top ten busiest starting and end stations. The top three stations for riders to pick up a rental bike were Streeter Dr & Grand Ave, DuSable Lake Shore Dr & Monroe St, and DeSable LakeShore Dr & North Blvd. On the other hand, the top three stations for riders to drop off their rental bikes were Streeter Dr & Grand Ave, DuSable Lake Shore Dr & North Blvd, and Michigan Ave & Oak St.

It's important to point out that even though the number of visits to these stations were different, the locations were the same for both tables. A simple “count unique” on the number of stations showed that there were more than 1300 stations in Chicago in 2022. These 10 stations accounted for more than 7% of the total traffic recorded in the entire year. Perhaps an effective strategy to attract more customers would be to increase the number of rental bikes in these stations.

Trip Activities on Density Maps



The figure above shows the locations where riders favored picking up a rental bike. A darker area equals more rental activities. In regards to locations, casual riders showed trends favoring stations closer to the piers and water. On the other hand, members showed more uniform activities across the downtown region.



The second figure shows where members and casual riders liked to return their rental bikes. One interesting observation is that the return activities were not as concentrated around the piers but more spread out than where riders liked to pick up their rentals. Members did show more activities concentrated in the downtown area.

Summary Observation and Recommendations

The first major finding from the data was that there were more annual members than casual riders in all the quarters examined. The second major finding was that annual members used the rental bikes the most during weekdays and less during the weekend. This may suggest that annual members were using the service more for commuting to work and short rides. On the other hand, casual riders were most active during the weekend and less so during the weekdays. Moreover, there were more bike rental activities during the summer months when compared to the colder winter months.

To help convince casual riders to purchase an annual pass, the following suggestions can be considered:

1) Create an “upsell” and convince casual riders that an annual pass is a much better deal than single-ride and full-day passes.
2) Provide more pricing options such as quarter-year and half-year passes so that riders could have more freedom.
3) Create a “free trial” for more riders to experience the service so that ultimately more positive word-of-mouth messages can travel along. This seems especially important since there was a decrease in both types of customer.
4) Announce a calculated price increase especially for renting the bike during the weekend and peak hours (around 5PM). This may drive more “on-the-fence” casual riders to want to secure the lower price of an annual pass.
5) Since casual riders favored electric bikes; make electric bikes more expensive to rent while members could choose any type of bike without additional costs.

Other Considerations

1) Create user-ids and collect data on riders. It would be informative to see how often a rider would use the service. Also, it would be beneficial for the company to see if a casual rider would eventually become a member.
2) Collect data on the type of bikes riders prefer from different stations so that more preferred bike types can be properly allocated.

* All the SQL codes will be included in a separate file.

Supplementary Materials

While not as powerful to be used purely for analyzing data as other tools such as Tableau or Python, SQL can still provide useful analysis with simple queries. The following were SQL queries used for data analysis in this case study.

1) This first query was to compare the average and median ride time for members and casual riders:

```sql
-- Checking the average and mdeian ride time for the rider types
select member_casual,
round(avg(ride_duration_in_hour),2) as average_trip_duration_in_hours,
percentile_cont(.5) within group (order by ride_duration_in_hour)
as median_trip_duration_in_hours
from bike_data
group by member_casual
```

Data Output | Explain | Messages | Notifications

	member_casual character varying (20)	average_trip_duration_in_hours numeric	median_trip_duration_in_hours double precision
1	casual	0.49	0.22
2	member	0.21	0.15

The first thing to be noticed is that casual riders had higher average and median trip durations than members. In addition, the numbers for median trip duration were lower than averages. In regards to statistics, the numbers are pointing towards the distribution of the trip durations being skewed right (positively).

2) The following query was used to obtain the average ride time in hours during different days of the week for both rider types:

```sql
--Average ride time for the rider types during different days of the week
select member_casual, started_at_dow as starting_day_of_the_week,
round(avg(ride_duration_in_hour), 2) as average_ride_duration_in_hours
from bike_data
group by member_casual, started_at_dow
order by avg(ride_duration_in_hour) desc
```

Data Output | Explain | Messages | Notifications

	member_casual character varying (20)	starting_day_of_the_week character varying (20)	average_ride_duration_in_hours numeric
1	casual	Sunday	0.57
2	casual	Saturday	0.54
3	casual	Monday	0.49
4	casual	Friday	0.47
5	casual	Tuesday	0.43
6	casual	Thursday	0.43
7	casual	Wednesday	0.41
8	member	Saturday	0.24
9	member	Sunday	0.23
10	member	Friday	0.21
11	member	Thursday	0.21
12	member	Monday	0.20
13	member	Tuesday	0.20
14	member	Wednesday	0.20

Both members and casual riders rode their bikes longer during the weekend and least on Wednesday.

3) To check for the number of rides on different days of the week for both rider types, the following query was used:

```sql
--Number of rental transactions for the rider types on days of the week
Select member_casual, started_at_dow, count(*) as ride_count
from bike_data
group by member_casual, started_at_dow
order by member_casual, count(*) desc
```

Data Output | Explain | Messages | Notifications

	member_casual character varying (20)	started_at_dow character varying (20)	ride_count bigint
1	casual	Saturday	473190
2	casual	Sunday	389036
3	casual	Friday	334701
4	casual	Thursday	309330
5	casual	Monday	277675
6	casual	Wednesday	274354
7	casual	Tuesday	263746
8	member	Thursday	532261
9	member	Wednesday	523869
10	member	Tuesday	518626
11	member	Monday	473339
12	member	Friday	467086
13	member	Saturday	443281
14	member	Sunday	387223

It's easy to see that casual riders rode the most on Saturdays and members on Thursdays.

4) To confirm the finding from the previous query, the following code was used:

```sql
--Most busy day for both rider types
select member_casual, mode() within group (order by started_at_dow)
as most_busy_day
from bike_data
group by member_casual
```

Data Output | Explain | Messages | Notifications

	member_casual character varying (20)	most_busy_day character varying
1	casual	Saturday
2	member	Thursday

Indeed, casual riders rode the most on Saturdays and members on Thursdays.

5) To make a quick summary table, this query was used:

```sql
--Making a summary table
select member_casual,
round(min(ride_duration_in_hour),2) as min_ride_duration,
round(max(ride_duration_in_hour),2) as max_ride_duration,
round(avg(ride_duration_in_hour),2) as avg_ride_duration,
round(cast(percentile_cont(.5) within group
      (order by ride_duration_in_hour)as numeric),2)
as median_ride_duration
from bike_data
group by member_casual
```

Data Output | Explain | Messages | Notifications

	member_casual character varying (20)	min_ride_duration numeric	max_ride_duration numeric	avg_ride_duration numeric	median_ride_duration numeric
1	casual	-2.29	689.79	0.49	0.22
2	member	-172.56	26.00	0.21	0.15

There are some interesting numbers! For minimum ride times, negative numbers were shown for both rider types. This could have been because of mistakes in the data, or it could be from accidentally switching the starting and end stations. Further investigations are needed to find the reason for the negative numbers.

In addition, max ride time for a casual rider was over 600 hours! This could have been caused by a mistake in the system or it could be because this particular rider was keeping the bike for a long time. Once again, further investigations are needed.

6) The following query was used to check for “long” rides over one hour long:

```sql
--Checking the number of trips over one hour for the riders
Select member_casual,
sum(case when ride_duration_in_hour > 1 then 1
         else 0
         end) as number_of_trips_over_one_hour
from bike_data
group by member_casual
```

Data Output | Explain | Messages | Notifications

	member_casual character varying (20)	started_at_dow character varying (20)	ride_count bigint
1	casual	Saturday	473190
2	casual	Sunday	389036
3	casual	Friday	334701
4	casual	Thursday	309330
5	casual	Monday	277675
6	casual	Wednesday	274354
7	casual	Tuesday	263746
8	member	Thursday	532261
9	member	Wednesday	523869
10	member	Tuesday	518626
11	member	Monday	473339
12	member	Friday	467086
13	member	Saturday	443281
14	member	Sunday	387223

Casual riders again recorded more long rides during the weekend while members showed longer rides during the weekdays.

7) Finally, the following query was used to make a pivot table for the number of rides occurred on different days of the week:

```
--Making a pivot table for trip counts on different days of the week
select * from crosstab (
    'select member_casual,
    started_at_dow,
    count(*)
    from bike_data
    group by member_casual, started_at_dow',

    'select started_at_dow
    from bike_data
    group by started_at_dow'
)

as (
    member_casual varchar(20),
    Sunday bigint,
    Monday bigint,
    Tuesday bigint,
    Wednesday bigint,
    Thursday bigint,
    Friday bigint,
    Saturday bigint
)
```

Data Output | Explain | Messages | Notifications

	member_casual character varying (20)	sunday bigint	monday bigint	tuesday bigint	wednesday bigint	thursday bigint	friday bigint	saturday bigint
1	casual	334701	277675	473190	389036	309330	263746	274354
2	member	467086	473339	443281	387223	532261	518626	523869