How do casual riders and annual members use Cyclistic bikes differently?
— more details about this case study

Chuang, Ya Chu (莊雅竹)

Scenario A case study form Coursera

This was a fiction company Cyclistic, a bike-share company in Chicago. It had two types of users,

1. **Casual riders:** single-ride passes and full-day passes
2. **Annual members**: annual memberships, according to finance analysts they are **more profitable**

The marketing team had an insight:

"Cyclistic's future success depends on maximizing the number of annual memberships. There is a very good chance to **convert casual riders into members** because they are already aware of the Cyclistic program "

Before designing marketing strategies aimed at converting casual riders into annual members, they needed to know...

"How do casual riders and annual members use Cyclistic bikes differently ?"

Materials From

Tools I Used

Prepare	Process	Analyze	Visualiztion
BigQuery (GCP)	Spreadsheet BigQuery	BigQuery	Looker studio Spreadsheet

Ask

1. How annual members and casual riders use our Cyclistic bikes differently?
2. Why would casual members upgrade to annual memberships?
3. How can Cyclistic use digital media to influence casual riders to become members?

Prepare

1. Download the previous 12 months (2023/07 - 2024/06) of trip data
2. Browse the file and check the structure of data
3. Upload to **Google Cloud Platform**
4. Upload to BigQuery with **bq command** (combine 12 csv files into one table)

Approximately **5,700,000 data in total**, and each had **13 columns**, including

STRING:

ride_id (=unique to every single trip)**, rideable_type** (=bike type)**, started_station_name, start_station_id, end_station_name, end_station_id, member_casual** (=membership type)

TIMESTAMP:

started_at, ended_at (=the time when they started or ended their ride)

FLOAT:

start_lat, start_lng, end_lat, end_lng (=where they started or ended their ride)

Process

1. Identify distinct value
2. Identify relation between columns
3. Check if there is a null value
4. Check if values are reasonable

Process - Identify distinct value

1. **ride_id** should be unique, **but there were 221 duplicates**
2. Check the difference of data between them:

ride_id	rideable_type	started_at	ended_at	start_station_nar	start_station_id	end_station_nan	end_station_id	start_lat	start_lng	end_lat	end
011C8EF97AB0	classic_bike	2024-05-31 19:45:38.037000 UTC	2024-06-01 20:45:33.862000 UTC	Clifton Ave & Arr	TA1307000163			41.918216	-87.656936		
011C8EF97AB0	classic_bike	2024-05-31 19:45:38.000000 UTC	2024-06-01 20:45:33.000000 UTC	Clifton Ave & Arr	TA1307000163			41.918216	-87.656936		
01406457A85BC	electric_bike	2024-05-31 23:54:59.000000 UTC	2024-06-01 00:01:47.000000 UTC			Damen Ave & Cl	13132	41.89	-87.67	41.895769	
01406457A85BC	electric_bike	2024-05-31 23:54:59.194000 UTC	2024-06-01 00:01:47.626000 UTC			Damen Ave & Cl	13132	41.89	-87.67	41.895769	
02606FBC7F85;	classic_bike	2024-05-31 17:55:01.000000 UTC	2024-06-01 18:54:53.000000 UTC	Pine Grove Ave	TA1307000150			41.94947274	-87.64645278		
02606FBC7F85;	classic_bike	2024-05-31 17:55:01.635000 UTC	2024-06-01 18:54:53.970000 UTC	Pine Grove Ave	TA1307000150			41.94947274	-87.64645278		
0354FD0756337	electric_bike	2024-05-31 23:34:36.273000 UTC	2024-06-01 00:14:29.238000 UTC					41.97	-87.66	41.96	
0354FD0756337	electric_bike	2024-05-31 23:34:36.000000 UTC	2024-06-01 00:14:29.000000 UTC					41.97	-87.66	41.96	
048C715F1DE0	electric_bike	2024-05-31 23:53:44.401000 UTC	2024-06-01 00:12:26.776000 UTC					41.89	-87.66	41.89	
048C715F1DE0	electric_bike	2024-05-31 23:53:44.000000 UTC	2024-06-01 00:12:26.000000 UTC					41.89	-87.66	41.89	
05D27072A33A;	classic_bike	2024-05-31 16:34:46.426000 UTC	2024-06-01 04:12:45.545000 UTC	Dearborn St & E	13045	DuSable Lake S	TA1309000039	41.893992	-87.629318	41.932588	

3. Data in most columns were the same, but slight different in **start_at** and **end_at**

⇒ **Exclude the duplicates**

Process - Identify relation between columns

1. Was **station_id** associated with **station_name**?
   There were **multiple station_names** with **the same station_id**!

end_station_id ▼ ↓	end_station_name ▼
TA1309000042	Lincoln Ave & Melrose St
TA1309000042	Lincoln Ave & Belmont Ave (Te...
TA1305000030	Wells St & Randolph St
TA1305000030	Clark St & Randolph St
KA1503000074	Museum of Science and Industry

2. Check the location (**lat & lng**) of these station, found that:
   a. Locations were close, I guessed <u>they might change their names</u> at different times
   b. **station_id=15541 and 15541.1.1** should be the same
      ⇒ **Replace station_id=15541.1.1 with 15541**

Process - Check if there is null value

1. The **station_name, station_id, end_lat** and **end_lng** of **8768 data were null.**

   Is it acceptable to exclude them ?

pe	started_at	ended_at	start_station_name	start_station_id	end_station_name	end_station_id	start_lat	start_lng	end_lat	end_lng	memb
e	2024-05-...	2024-0...	null	null	null	null	42.0	-87.67	42.0	-87.66	casua
e	2024-06-...	2024-0...	null	null	null	null	41.99	-87.65	42.0	-87.66	memb
e	2024-04-...	2024-0...	null	null	null	null	42.01	-87.67	42.0	-87.66	casua
e	2023-10-...	2023-1...	null	null	null	null	42.01	-87.68	42.0	-87.66	memb
e	2024-06	2024-0	null	null	null	null	42.0	87.68	42.0	87.68	casua

tart_station_name	start_station_id	end_station_name	end_station_id	start_lat	start_lng	end_lat	end_lng	member_casual
lalsted St & Clybour...	331	null	null	41.909668	-87.648128	null	null	member
lalsted St & Clybour...	331	null	null	41.909668	-87.648128	null	null	casual
lalsted St & Clybour...	331	null	null	41.909668	-87.648128	null	null	casual
lenwood Ave & Tou...	525	null	null	42.012701	-87.666058	null	null	casual

Process - Check if there is null value

2. The distribution of these null-value data throughout the year **was similar to** total data
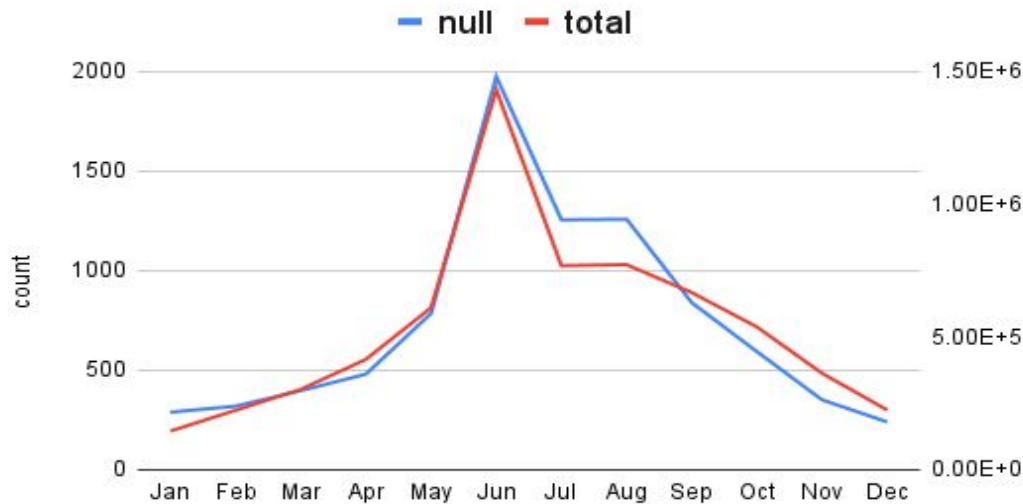
⇒ **It's fine to exclude them**

Process - Check if values are reasonable

1. **end_lat** and **end_lng** of 3 data **were zero**!

end_station_id	start_lat	start_lng	end_lat	end_lng
653B	41.893992	-87.629318	0.0	0.0
OH Charging Stx - Test	41.796642	-87.625923	0.0	0.0
OH Charging Stx - Test	41.86316583333...	-87.6798115	0.0	0.0

2. Check their **end_station_id** of these data in the table

   ⇒ Replace them with the **average end_lat** and **end_lng (proxy the data)**

Process - Check if values are reasonable

1. **started_at** should be earlier then **ended_at, but 434 data showed earlier ended_at than started_at !**

ride_id ▼	rideable_type ▼	started_at ▼	ended_at ▼
2BFB23CDC9A75AB0	electric_bike	2023-08-26 10:19:36 UTC	2023-08-26 10:16:52 UTC
7934DBD46A7BB934	electric_bike	2023-12-06 16:07:40 UTC	2023-12-06 16:07:37 UTC
64BF86DB62A97011	electric_bike	2023-07-22 10:05:44 UTC	2023-07-22 10:05:41 UTC
01A0D47FF0888776	electric_bike	2023-09-19 15:22:55 UTC	2023-09-19 15:22:52 UTC

⇒ Given that they were only a **small proportion** of the total, **exclude these data**

Analyze & Visualization

How do casual riders and annual members use Cyclistic bikes differently ?

1. In terms of riding times
2. In terms of their destinations
3. In terms of riding duration
4. In terms of which day of the week to ride
5. In terms of rideable type

Analyze - create new columns

TIMESTAMP_DIFF(ended_at,started_at,MINUTE) AS **riding_duration** (minutes)

EXTRACT(DAYOFWEEK FROM started_at) AS **weekday** (1 for Sun, 2 for Mon, 3 for Tue....)

EXTRACT(MONTH FROM started_at) AS **month** (1 for Jun, 2 for Feb, 3 for Mar....)

Analyze - In terms of riding times
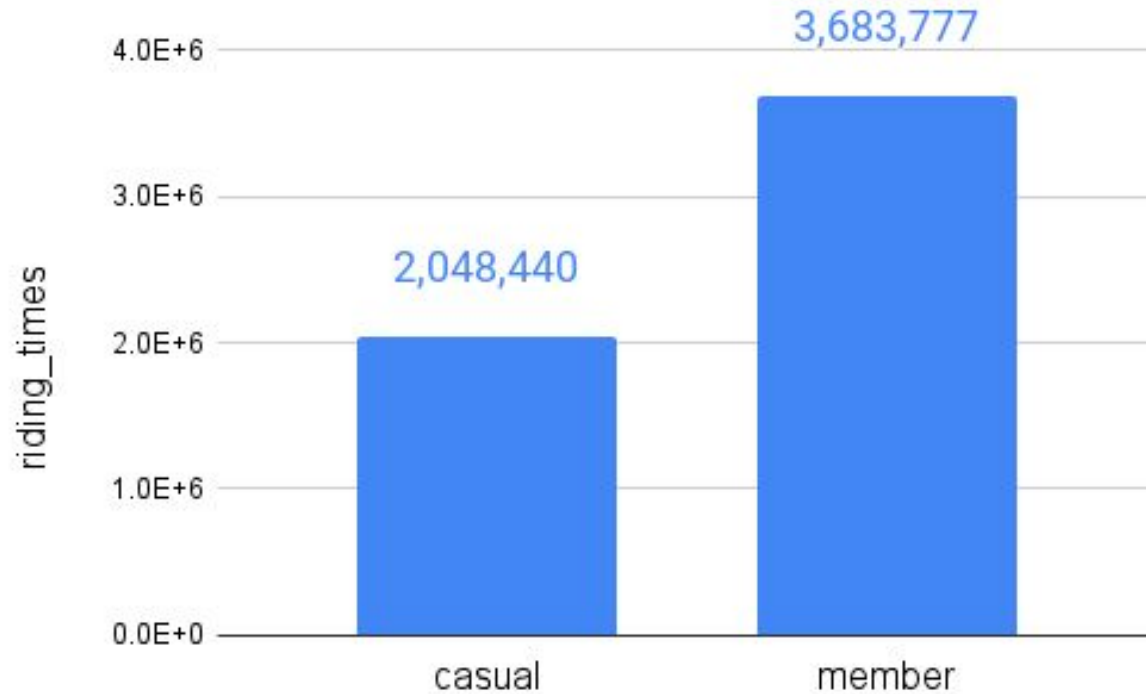
Anual members rode 1.8 times more than casual riders !

Analyze - In terms of riding times

Both kinds of users shared the same pattern of riding times throughout the year.
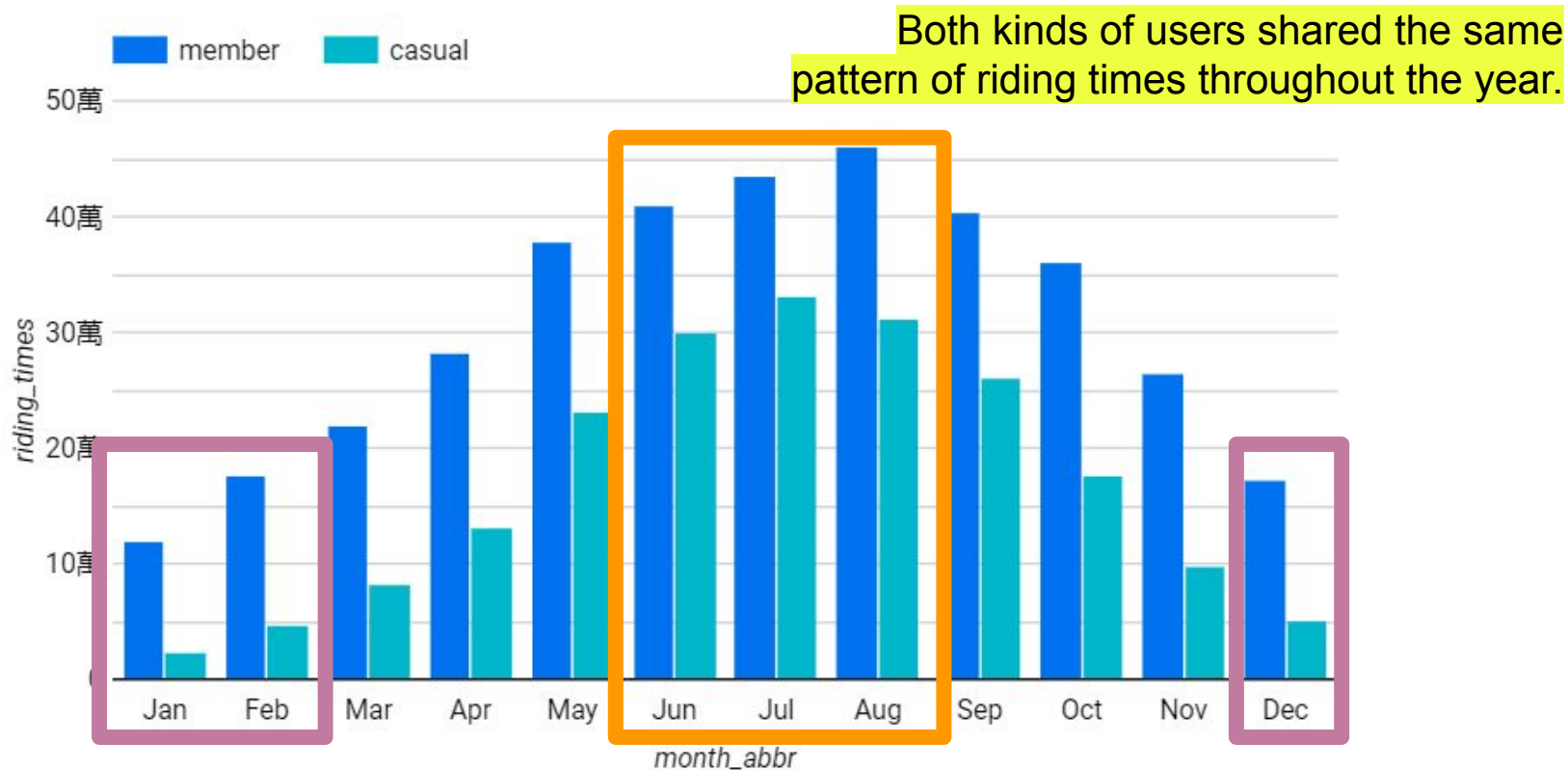
Analyze - In terms of riding times (off & peak)

off season

peak season
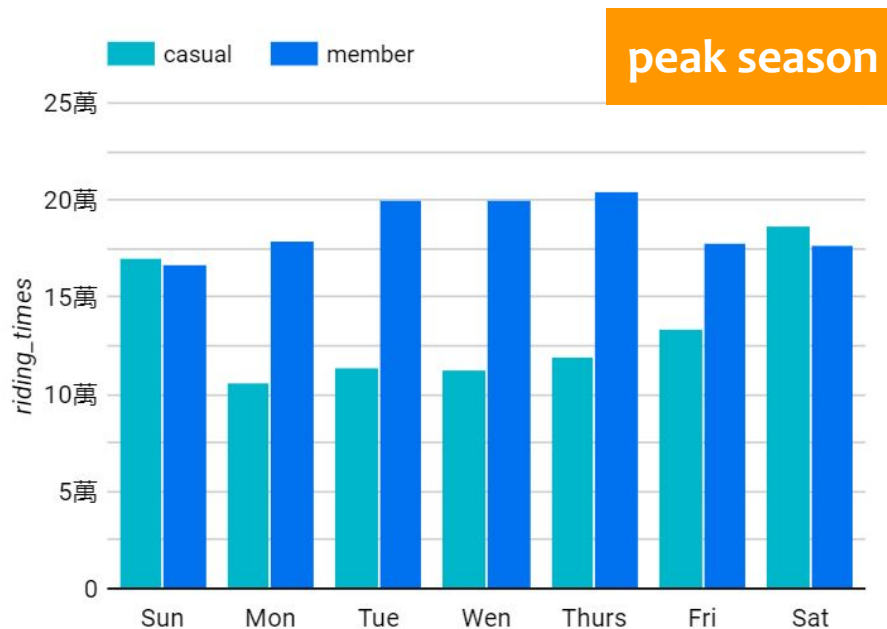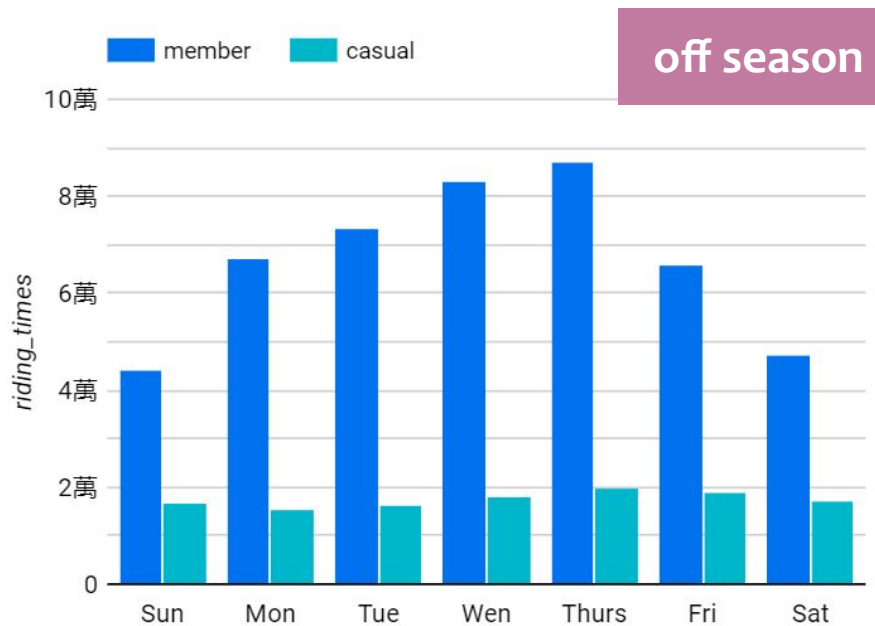
Casual riders used more often **on weekends in busy season**,
while aunal members prefered **weekdays** use **in both seansons**.

Analyze - In terms of their destinations

The most visited destinations of annual members and casual riders are different,

So, Where were they?

end_station_id	member	casual
13022	14,079	50,630
13300	11,282	29,723
LF-005	15,688	23,870
13042	14,133	24,303
KA1503000043	26,846	8,079
TA1307000039	24,762	10,012
WL-012	28,686	6,077
TA1308000050	21,148	11,073
13008	8,997	22,194
TA1308000001	12,797	17,620
KA1504000135	20,285	10,037
TA1305000032	24,094	6,022
13137	18,641	8,906

(visit times)

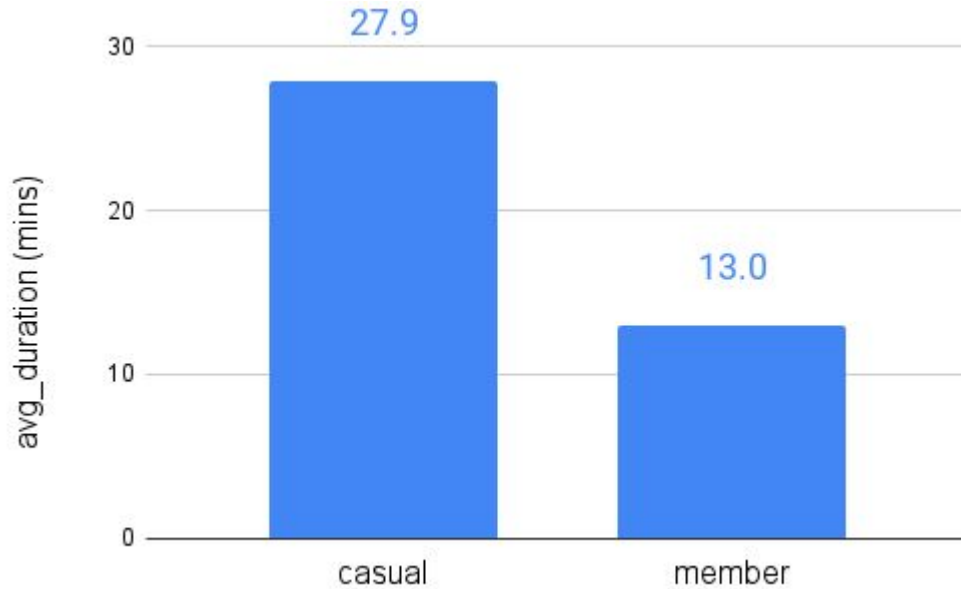# Analyze - In terms of casual riders' destinations
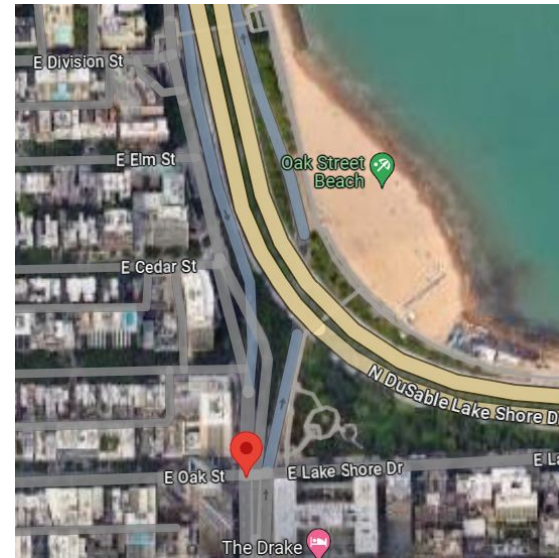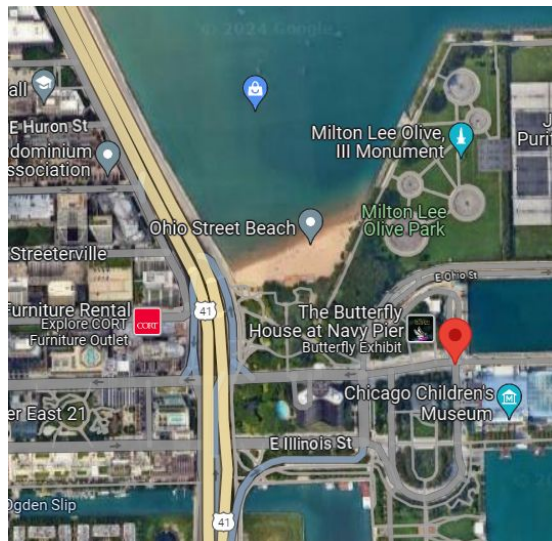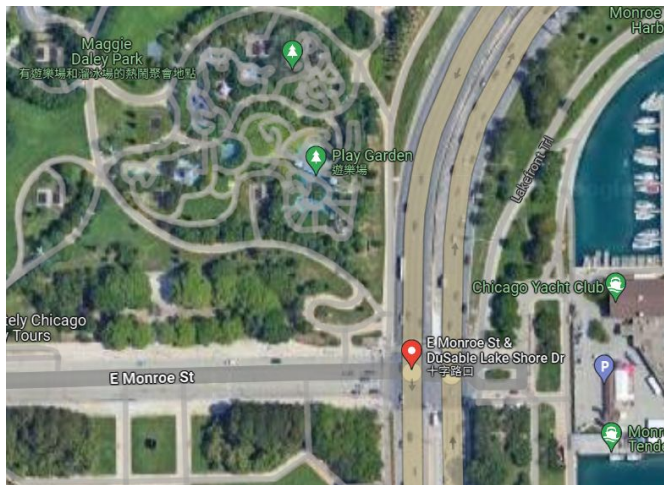
Casual riders tended to visit the **suburbans or beach**, while

Analyze - In terms of annual members' destinations
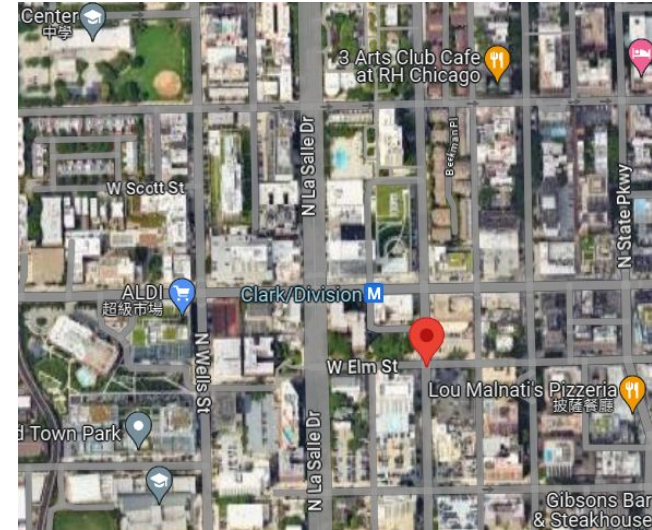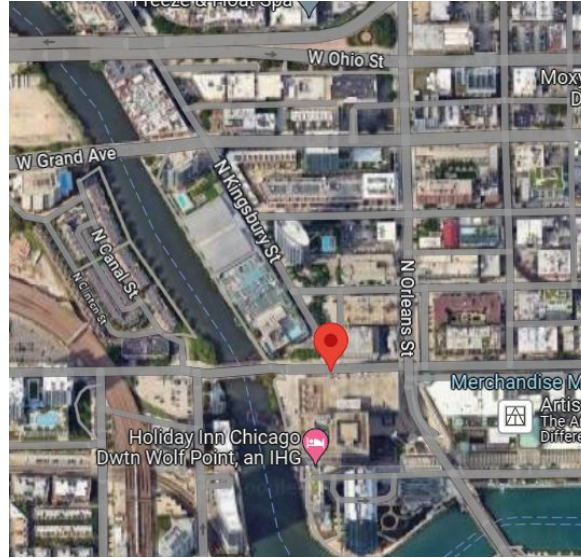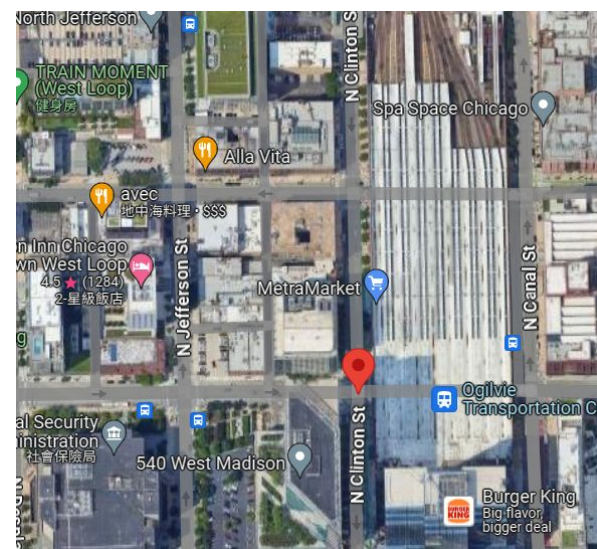
Casual riders tended to visit the **suburbans or beach**, while annual members tended to go to **downtown area**.
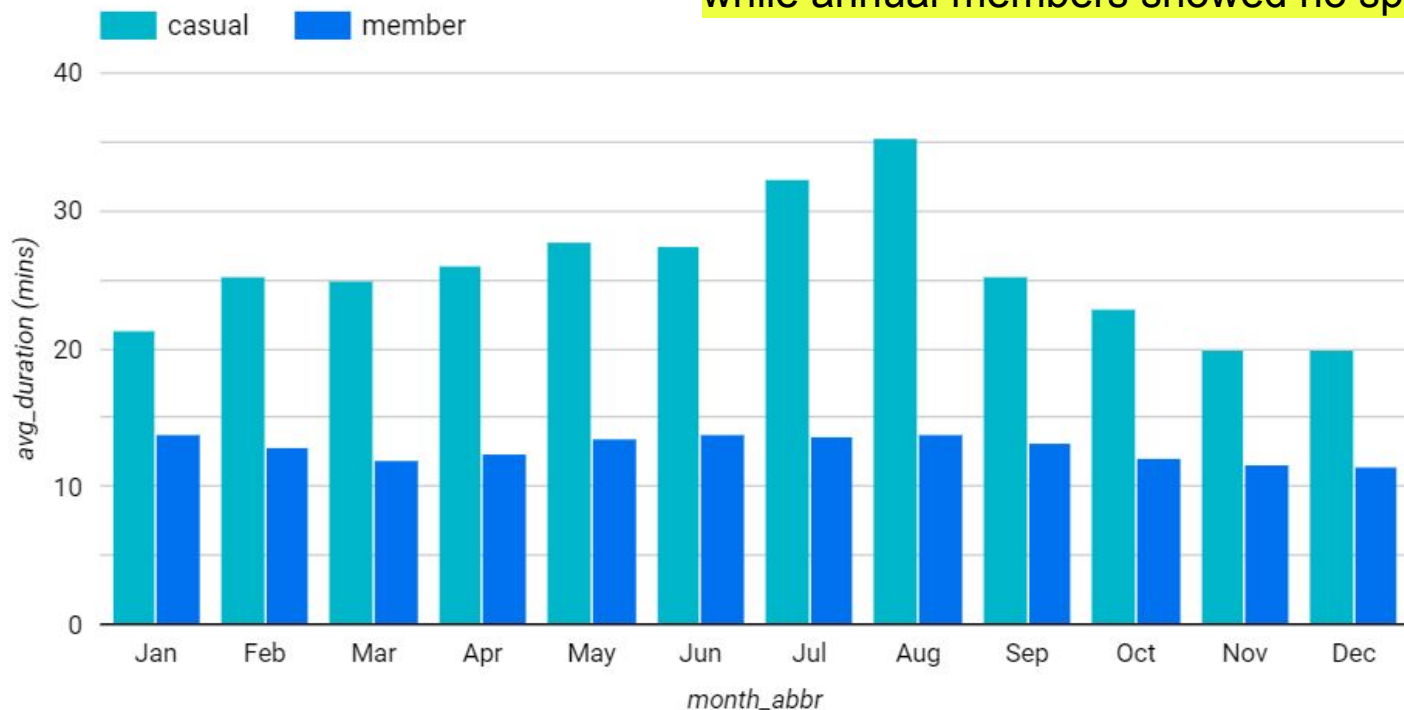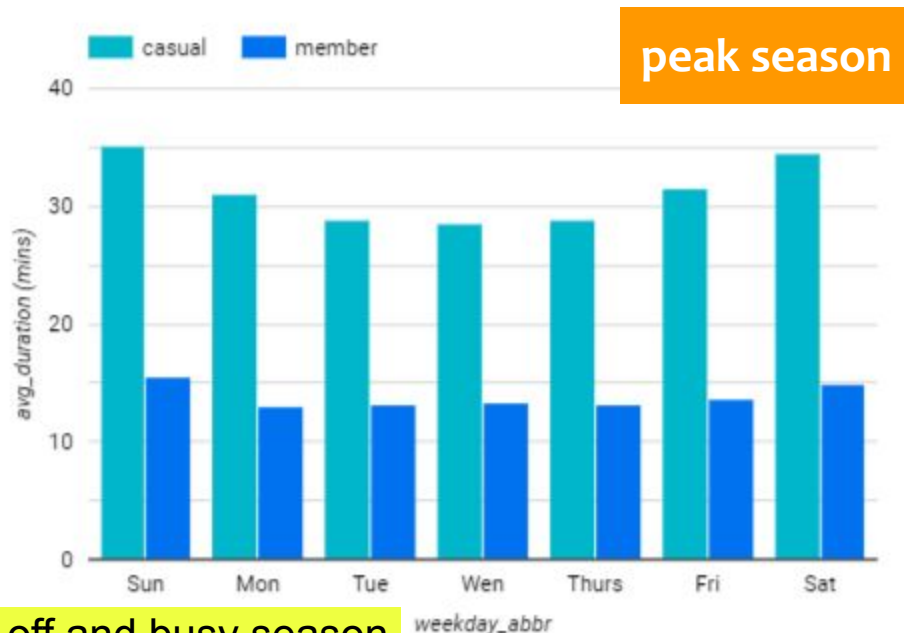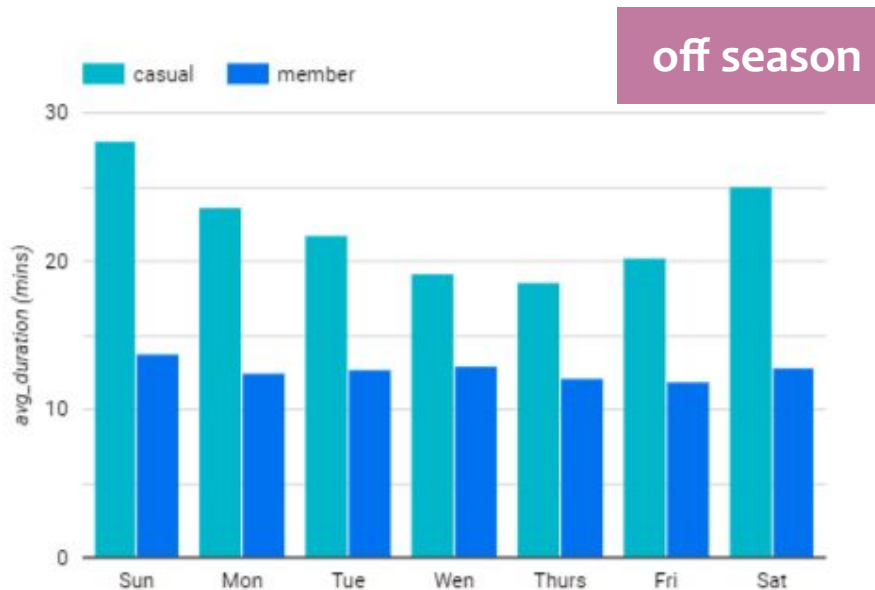
Analyze - In terms of riding duration

Casaul users rode 2.15 times longer than annual memebers!

Analyze - In terms of riding duration

Casual riders tended to **ride longer in peak season**, while annual members showed no specific pattern.

Analyze - In terms of riding duration (off & peak)

off season

peak season

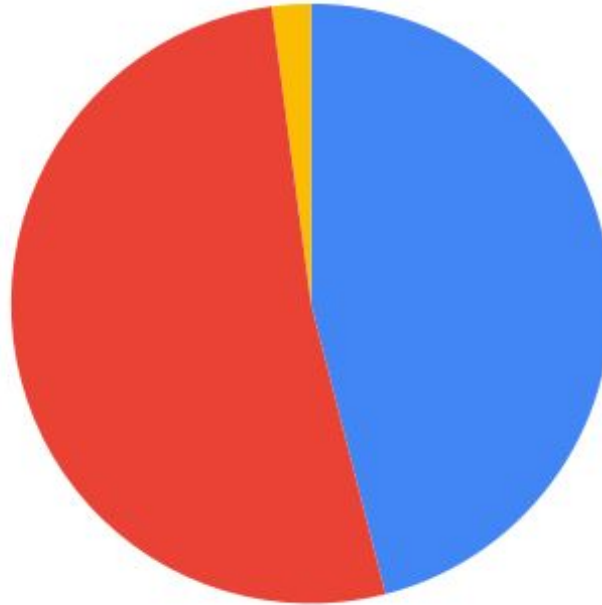The patterns of riding duration **were similar** in off and busy season.
Casual riders tended to ride **longer on weekend**, while riding durations of annual members **were even throughout the week**

Analyze - In terms of rideable type

Rideable type of ANUAL MEMBERS   Rideable type of CASUAL users

- classic_bike
- electric_bike
- docked_bike

No significant difference between annual members and casual riders.
Note that **no annual member used docked_bike** in the past year.

Conclusion

	Casual Rider	Annual Member
Number of times	**Fewer** overall	**Higher** overall
Duration	Longer overall, **especially in summer**	Shorter overall, **evenly** in every month
Prefer days of use (weekdays / weekends)	Used more often and longer **on weekends, especially in summer**	Used more often **on weekdays**
Destination	Mostly **tourist area**	Mostly **downtown**
Prefer rideable type	No preference	
Prerfer riding season (peak season)	**Summer** (From June to August)	

Conclusion

	Casual Rider	Annual Member

1. **Riding times were considerable** in both types of users, **annual members** especially. This showed high popularity of Cyclistic.
2. **Causual riders** tended to ride **longer,** while **annual members** contributed **more revenue** than casual rider, according to the finance analysts.
3. **Annual members** could **create more stable revenue** in terms of weekly usage, compared to casual riders.

⇒ **Turn casual riders to annual members** to gain more stable revenue!!

But how?

(peak season) **Summer** (From June to August)

Conclusion

	Casual Rider	Annual Member
Number of times	Fewer overall	Higher overall

4. Casual riders tended to bike on holidays or vacations.
5. Annual members tended to bike for regular use (commutation probably).

⇒ Launch campaigns on holidays, encouraging the users commuting by bike.

Destination	Mostly tourist area	Mostly downtown
Prefer rideable type	No preference	
Prerfer seasons (peak season)	Summer (From June to August)	

Further things we can do…

1. **Comfirm the usage habit** of casual rider and annual member, for example by conducting a survey.
2. The survey should also include whether annual members have been casual riders and **why they chose to covert to annual**.
3. Develop marketing strategies after exploring **the type of digital media that Cyclistic users used**, including when to use and how often to use.