Divvy Company

Data Analytics Team

Final Report on Divvy Case Study

Presenter: Nguyen Dai Minh

July 6, 2024

Content

1 Problem Statement

The primary objective of this report is to generate answers and recommendations to the following question made by the Divvy director of marketing: How do annual members and casual riders use Divvy differently ?

2 Data Description

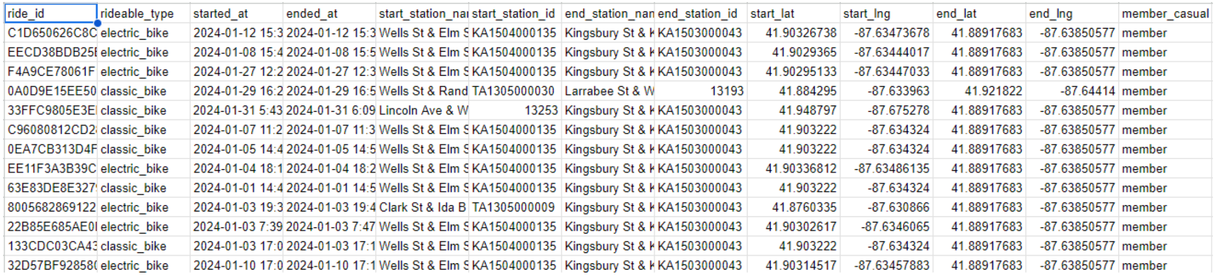
All data is available for public use by divvy corporation and is taken directly from the azure cloud. The dataset is then store locally on my machine as well as google bigquery and google drive. The data is collected from January to March of 2024 and contains the following attributes.

- ride_id : unique identifier of user
- ridable_type : the type of bikes being used
- started_at : start time
- ended_at : end time
- start_station_name : start station name
- start_station_id : unique identifier of station
- end_station_name : end station name
- end_station_id : unique identifier of station
- start_lat : start latitude
- start_lng : start longitude
- end_lat : end latitude
- end_lng : end longitude
- member_casual : membership type

3 Data Cleaning and Manipulation

I will be doing general manipulation and feature engineering in Google sheets and finish cleaning up in Google BigQuery using SQL. Let's start from Google sheets first:

1. Make a copy of January data



ride_id	rideable_type	started_at	ended_at	start_station_na	start_station_id	end_station_nam	end_station_id	start_lat	start_lng	end_lat	end_lng	member_casual
C1D650626C8C	electric_bike	2024-01-12 15:3	2024-01-12 15:3	Wells St & Elm S	KA1504000135	Kingsbury St & K	KA1503000043	41.90326738	-87.63473678	41.88917683	-87.63850577	member
EECD38BDB25	electric_bike	2024-01-08 15:4	2024-01-08 15:5	Wells St & Elm S	KA1504000135	Kingsbury St & K	KA1503000043	41.9029365	-87.63444017	41.88917683	-87.63850577	member
F4A9CE78061F	electric_bike	2024-01-27 12:2	2024-01-27 12:3	Wells St & Elm S	KA1504000135	Kingsbury St & K	KA1503000043	41.90295133	-87.63447033	41.88917683	-87.63850577	member
0A0D9E15EE50	classic_bike	2024-01-29 16:2	2024-01-29 16:5	Wells St & Rand	TA1305000030	Larrabee St & W	13193	41.884295	-87.633963	41.921822	-87.64414	member
33FFC9805E3E	classic_bike	2024-01-31 5:43	2024-01-31 6:09	Lincoln Ave & W	13253	Kingsbury St & K	KA1503000043	41.948797	-87.675278	41.88917683	-87.63850577	member
C96080812CD2	classic_bike	2024-01-07 11:2	2024-01-07 11:3	Wells St & Elm S	KA1504000135	Kingsbury St & K	KA1503000043	41.903222	-87.634324	41.88917683	-87.63850577	member
0EA7CB313D4F	classic_bike	2024-01-05 14:4	2024-01-05 14:5	Wells St & Elm S	KA1504000135	Kingsbury St & K	KA1503000043	41.903222	-87.634324	41.88917683	-87.63850577	member
EE11F3A3B39C	electric_bike	2024-01-04 18:1	2024-01-04 18:2	Wells St & Elm S	KA1504000135	Kingsbury St & K	KA1503000043	41.90336812	-87.63486135	41.88917683	-87.63850577	member
63E83DE8E327	classic_bike	2024-01-01 14:4	2024-01-01 14:5	Wells St & Elm S	KA1504000135	Kingsbury St & K	KA1503000043	41.903222	-87.634324	41.88917683	-87.63850577	member
800568286912	electric_bike	2024-01-03 19:3	2024-01-03 19:4	Clark St & Ida B	TA1305000009	Kingsbury St & K	KA1503000043	41.8760335	-87.630866	41.88917683	-87.63850577	member
22B85E685AE0	electric_bike	2024-01-03 7:39	2024-01-03 7:47	Wells St & Elm S	KA1504000135	Kingsbury St & K	KA1503000043	41.90302617	-87.6346065	41.88917683	-87.63850577	member
133CDC03CA43	classic_bike	2024-01-03 17:0	2024-01-03 17:1	Wells St & Elm S	KA1504000135	Kingsbury St & K	KA1503000043	41.903222	-87.634324	41.88917683	-87.63850577	member
32D57BF92858	electric_bike	2024-01-10 17:0	2024-01-10 17:1	Wells St & Elm S	KA1504000135	Kingsbury St & K	KA1503000043	41.90314517	-87.63457883	41.88917683	-87.63850577	member

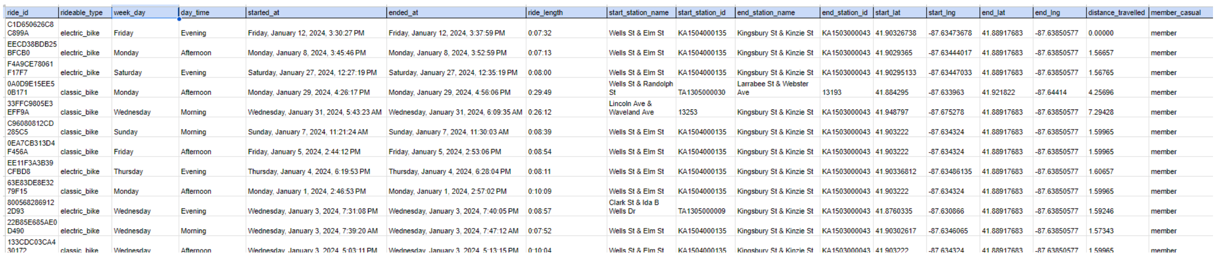
2. Reformat alignment

3. Highlight headers

4. Drop all blank feature in latitude and longitude columns as it would affect calculations of new features

5. Generate new features including:

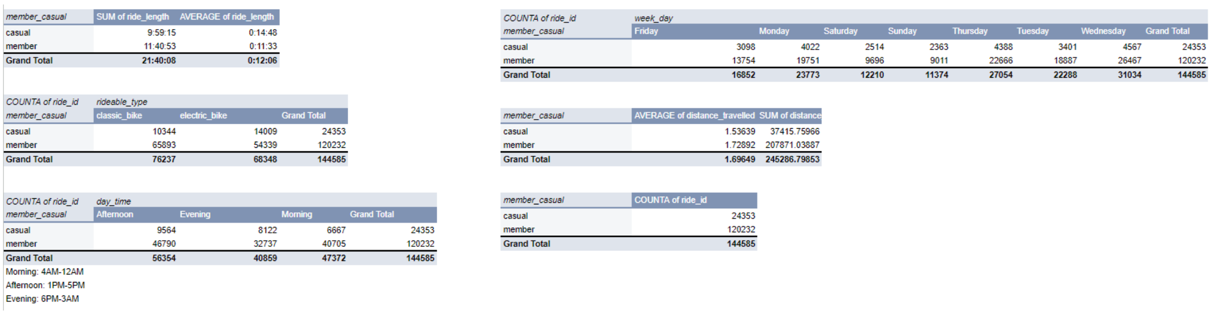
   - ride_length: the time the user rode a bike
   - week_day: the day bikes were used
   - distance_travelled: the distance travelled by the bike
   - day_time: the time of the day in 3 categories: Morning (4-12AM), Afternoon (1-5PM) and Evening (6-3PM)



ride_id	rideable_type	week_day	day_time	started_at	ended_at	ride_length	start_station_name	start_station_id	end_station_name	end_station_id	start_lat	start_lng	end_lat	end_lng	distance_travelled	member_casual
C1D650626C8C899A	electric_bike	Friday	Evening	Friday, January 12, 2024, 3:30:27 PM	Friday, January 12, 2024, 3:37:59 PM	0:07:32	Wells St & Elm St	KA1504000135	Kingsbury St & Kinzie St	KA1503000043	41.90326738	-87.63473678	41.88917683	-87.63850577	0.00000	member
EECD38BDB25BFCB0	electric_bike	Monday	Afternoon	Monday, January 8, 2024, 3:45:46 PM	Monday, January 8, 2024, 3:52:59 PM	0:07:13	Wells St & Elm St	KA1504000135	Kingsbury St & Kinzie St	KA1503000043	41.9029365	-87.63444017	41.88917683	-87.63850577	1.56657	member
F4A9CE78061F17F7	electric_bike	Saturday	Evening	Saturday, January 27, 2024, 12:27:19 PM	Saturday, January 27, 2024, 12:35:19 PM	0:08:00	Wells St & Elm St	KA1504000135	Kingsbury St & Kinzie St	KA1503000043	41.90295133	-87.63447033	41.88917683	-87.63850577	1.56765	member
0A0D9E15EE50B171	classic_bike	Monday	Afternoon	Monday, January 29, 2024, 4:26:17 PM	Monday, January 29, 2024, 4:56:06 PM	0:29:49	Wells St & Randolph St	TA1305000030	Larrabee St & Webster Ave	13193	41.884295	-87.633963	41.921822	-87.64414	4.25696	member
33FFC9805E3EFF9A	classic_bike	Wednesday	Morning	Wednesday, January 31, 2024, 5:43:23 AM	Wednesday, January 31, 2024, 6:09:35 AM	0:26:12	Lincoln Ave & Waveland Ave	13253	Kingsbury St & Kinzie St	KA1503000043	41.948797	-87.675278	41.88917683	-87.63850577	7.29428	member
C96080812CD285C5	classic_bike	Sunday	Morning	Sunday, January 7, 2024, 11:21:24 AM	Sunday, January 7, 2024, 11:30:03 AM	0:08:39	Wells St & Elm St	KA1504000135	Kingsbury St & Kinzie St	KA1503000043	41.903222	-87.634324	41.88917683	-87.63850577	1.59965	member
0EA7CB313D4F456A	classic_bike	Friday	Afternoon	Friday, January 5, 2024, 2:44:12 PM	Friday, January 5, 2024, 2:53:06 PM	0:08:54	Wells St & Elm St	KA1504000135	Kingsbury St & Kinzie St	KA1503000043	41.903222	-87.634324	41.88917683	-87.63850577	1.59965	member
EE11F3A3B39CFBD8	electric_bike	Thursday	Evening	Thursday, January 4, 2024, 6:19:53 PM	Thursday, January 4, 2024, 6:28:04 PM	0:08:11	Wells St & Elm St	KA1504000135	Kingsbury St & Kinzie St	KA1503000043	41.90336812	-87.63486135	41.88917683	-87.63850577	1.60657	member
63E83DE8E3279F15	classic_bike	Monday	Afternoon	Monday, January 1, 2024, 2:46:53 PM	Monday, January 1, 2024, 2:57:02 PM	0:10:09	Wells St & Elm St	KA1504000135	Kingsbury St & Kinzie St	KA1503000043	41.903222	-87.634324	41.88917683	-87.63850577	1.59965	member
8005682869122D93	electric_bike	Wednesday	Evening	Wednesday, January 3, 2024, 7:31:08 PM	Wednesday, January 3, 2024, 7:40:05 PM	0:08:57	Clark St & Ida B Wells Dr	TA1305000009	Kingsbury St & Kinzie St	KA1503000043	41.8760335	-87.630866	41.88917683	-87.63850577	1.59246	member
22B85E685AE0D490	electric_bike	Wednesday	Morning	Wednesday, January 3, 2024, 7:39:20 AM	Wednesday, January 3, 2024, 7:47:12 AM	0:07:52	Wells St & Elm St	KA1504000135	Kingsbury St & Kinzie St	KA1503000043	41.90302617	-87.6346065	41.88917683	-87.63850577	1.57343	member
133CDC03CA430172	classic_bike	Wednesday	Afternoon	Wednesday, January 3, 2024, 5:03:11 PM	Wednesday, January 3, 2024, 5:13:15 PM	0:10:04	Wells St & Elm St	KA1504000135	Kingsbury St & Kinzie St	KA1503000043	41.903222	-87.634324	41.88917683	-87.63850577	1.59965	member

6. Create 6 pivot tables including:

   - SUM and AVERAGE ride_length of each type of user
   - Distribution of types of user into each type of bikes
   - Distribution of types of user into day time
   - Distribution of types of user into week day
   - SUM and AVERAGE distance_travelled of each type of user
   - Distribution of user types



member_casual	SUM of ride_length	AVERAGE of ride_length
casual	9:59:15	0:14:48
member	11:40:53	0:11:33
Grand Total	**21:40:08**	**0:12:06**

COUNTA of ride_id	rideable_type		
member_casual	classic_bike	electric_bike	Grand Total
casual	10344	14009	24353
member	65893	54339	120232
Grand Total	**76237**	**68348**	**144585**

COUNTA of ride_id	day_time			
member_casual	Afternoon	Evening	Morning	Grand Total
casual	9564	8122	6667	24353
member	46790	32737	40705	120232
Grand Total	**56354**	**40859**	**47372**	**144585**

Morning: 4AM-12AM
Afternoon: 1PM-5PM
Evening: 6PM-3AM

COUNTA of ride_id	week_day							
member_casual	Friday	Monday	Saturday	Sunday	Thursday	Tuesday	Wednesday	Grand Total
casual	3098	4022	2514	2363	4388	3401	4567	24353
member	13754	19751	9696	9011	22666	18887	26467	120232
Grand Total	**16852**	**23773**	**12210**	**11374**	**27054**	**22288**	**31034**	**144585**

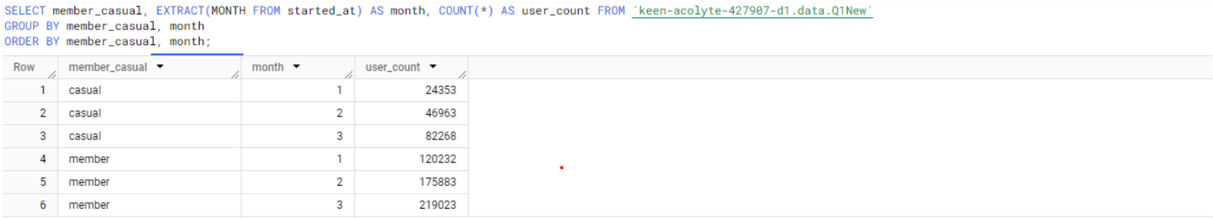
member_casual	AVERAGE of distance_travelled	SUM of distance
casual	1.53639	37415.75966
member	1.72892	207871.03887
Grand Total	**1.69649**	**245286.79853**

member_casual	COUNTA of ride_id
casual	24353
member	120232
Grand Total	**144585**

After getting a hold of the schema and how the data is structured let's move to Google BigQuery:

1. Combined 3 tables from 3 months into 1 table and remove all null values from columns latitude and longitude as it could mess with calculations

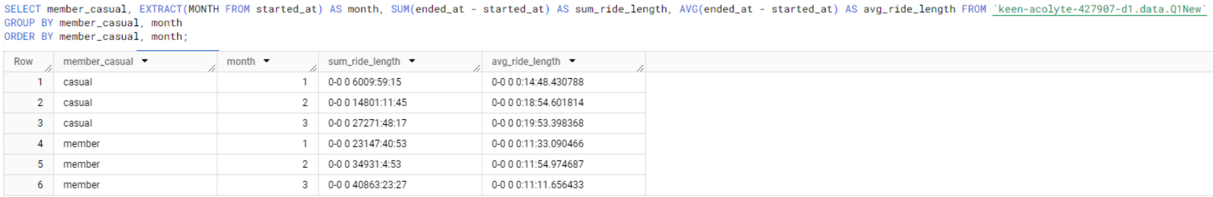
2. Query user by type, months and count



```
SELECT member_casual, EXTRACT(MONTH FROM started_at) AS month, COUNT(*) AS user_count FROM `keen-acolyte-427907-d1.data.Q1New`
GROUP BY member_casual, month
ORDER BY member_casual, month;
```

Row	member_casual	month	user_count
1	casual	1	24353
2	casual	2	46963
3	casual	3	82268
4	member	1	120232
5	member	2	175883
6	member	3	219023

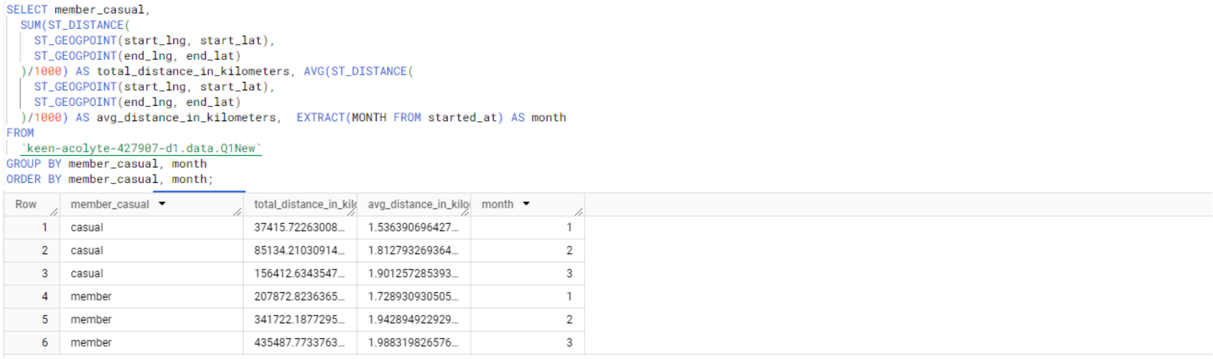
3. Query user by type, months average length ride and sum of length ride



```
SELECT member_casual, EXTRACT(MONTH FROM started_at) AS month, SUM(ended_at - started_at) AS sum_ride_length, AVG(ended_at - started_at) AS avg_ride_length FROM `keen-acolyte-427907-d1.data.Q1New`
GROUP BY member_casual, month
ORDER BY member_casual, month;
```

Row	member_casual	month	sum_ride_length	avg_ride_length
1	casual	1	0-0 0 6009:59:15	0-0 0 0:14:48.430788
2	casual	2	0-0 0 14801:11:45	0-0 0 0:18:54.601814
3	casual	3	0-0 0 27271:48:17	0-0 0 0:19:53.398368
4	member	1	0-0 0 23147:40:53	0-0 0 0:11:33.090466
5	member	2	0-0 0 34931:4:53	0-0 0 0:11:54.974687
6	member	3	0-0 0 40863:23:27	0-0 0 0:11:11.656433

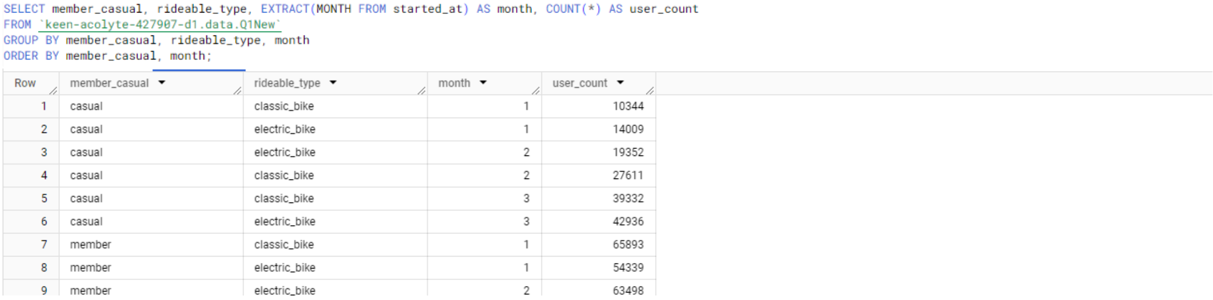
4. Query user by type, months average length distance and sum of length distance



```
SELECT member_casual,
  SUM(ST_DISTANCE(
    ST_GEOGPOINT(start_lng, start_lat),
    ST_GEOGPOINT(end_lng, end_lat)
  )/1000) AS total_distance_in_kilometers, AVG(ST_DISTANCE(
    ST_GEOGPOINT(start_lng, start_lat),
    ST_GEOGPOINT(end_lng, end_lat)
  )/1000) AS avg_distance_in_kilometers,  EXTRACT(MONTH FROM started_at) AS month
FROM
  `keen-acolyte-427907-d1.data.Q1New`
GROUP BY member_casual, month
ORDER BY member_casual, month;
```

Row	member_casual	total_distance_in_kil	avg_distance_in_kilo	month
1	casual	37415.72263008...	1.536390696427...	1
2	casual	85134.21030914...	1.812793269364...	2
3	casual	156412.6343547...	1.901257285393...	3
4	member	207872.8236365...	1.728930930505...	1
5	member	341722.1877295...	1.942894922929...	2
6	member	435487.7733763...	1.988319826576...	3

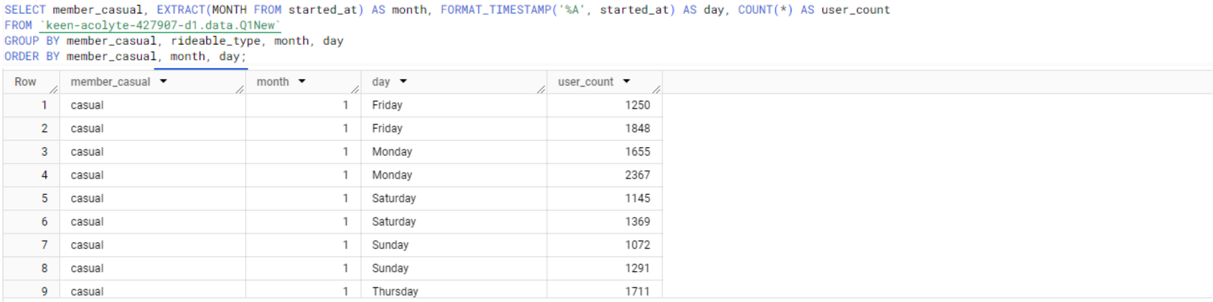
5. Query user by type, bike type and user count



```
SELECT member_casual, rideable_type, EXTRACT(MONTH FROM started_at) AS month, COUNT(*) AS user_count
FROM `keen-acolyte-427907-d1.data.Q1New`
GROUP BY member_casual, rideable_type, month
ORDER BY member_casual, month;
```

Row	member_casual	rideable_type	month	user_count
1	casual	classic_bike	1	10344
2	casual	electric_bike	1	14009
3	casual	electric_bike	2	19352
4	casual	classic_bike	2	27611
5	casual	classic_bike	3	39332
6	casual	electric_bike	3	42936
7	member	classic_bike	1	65893
8	member	electric_bike	1	54339
9	member	electric_bike	2	63498

6. Query user by type, month and day of the week



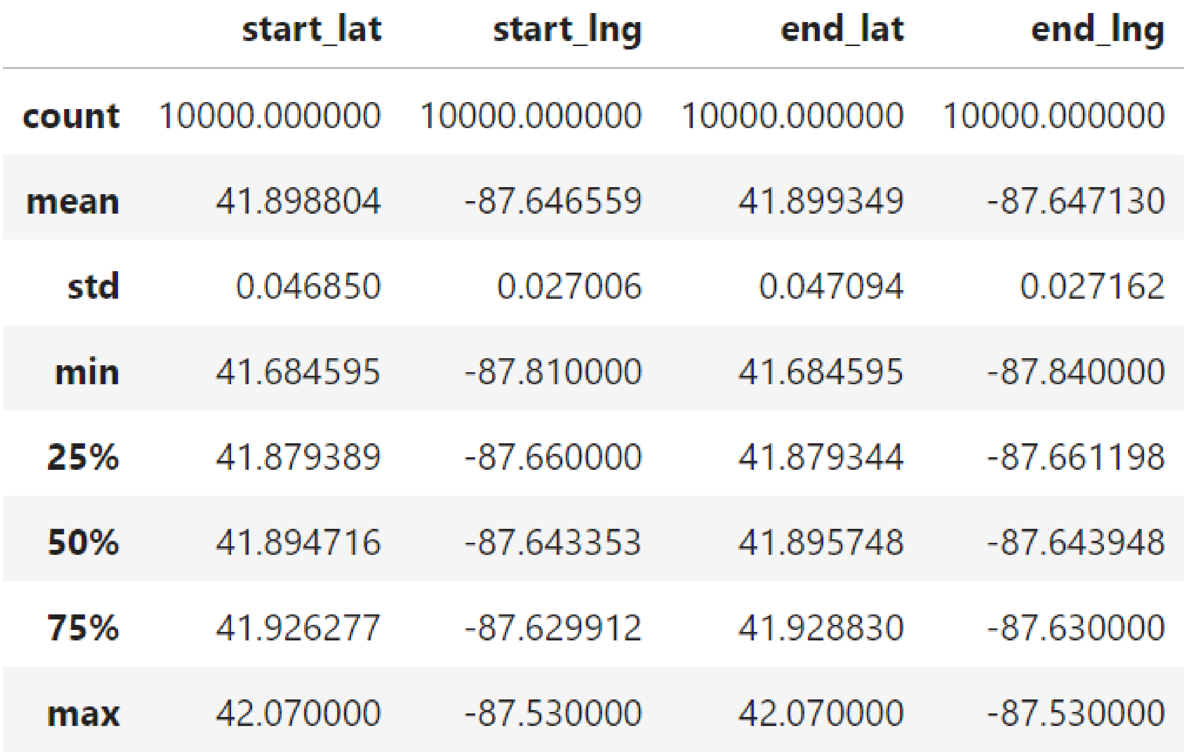
```
SELECT member_casual, EXTRACT(MONTH FROM started_at) AS month, FORMAT_TIMESTAMP('%A', started_at) AS day, COUNT(*) AS user_count
FROM `keen-acolyte-427907-d1.data.Q1New`
GROUP BY member_casual, rideable_type, month, day
ORDER BY member_casual, month, day;
```

Row	member_casual	month	day	user_count
1	casual	1	Friday	1250
2	casual	1	Friday	1848
3	casual	1	Monday	1655
4	casual	1	Monday	2367
5	casual	1	Saturday	1145
6	casual	1	Saturday	1369
7	casual	1	Sunday	1072
8	casual	1	Sunday	1291
9	casual	1	Thursday	1711

4 Analysis Summary

Now that we have use spreadsheets as well as google bigquery to take a quick look as well as making a few pivot table now we can do our statistical analysis in Python. I am going to only take a random sample from the whole table for analysis:

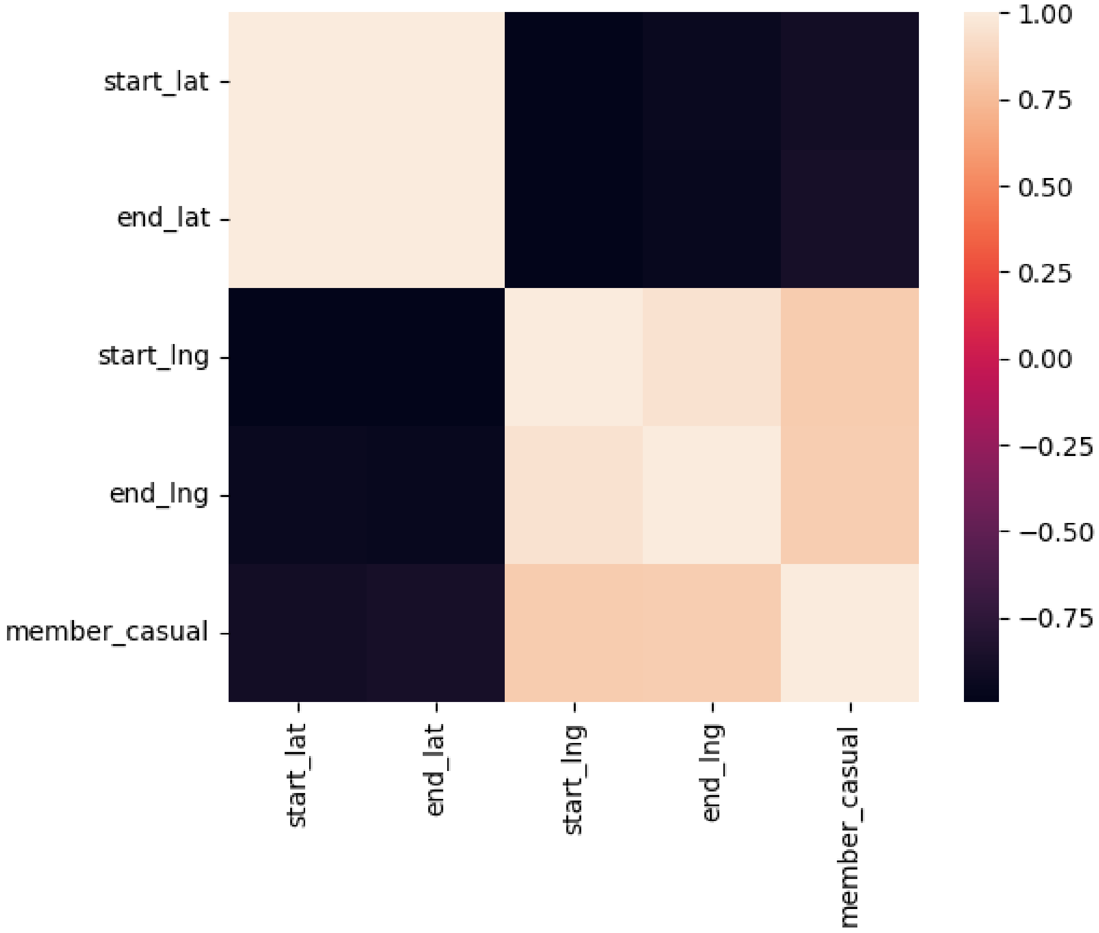
- Calculate mean, median and quartile



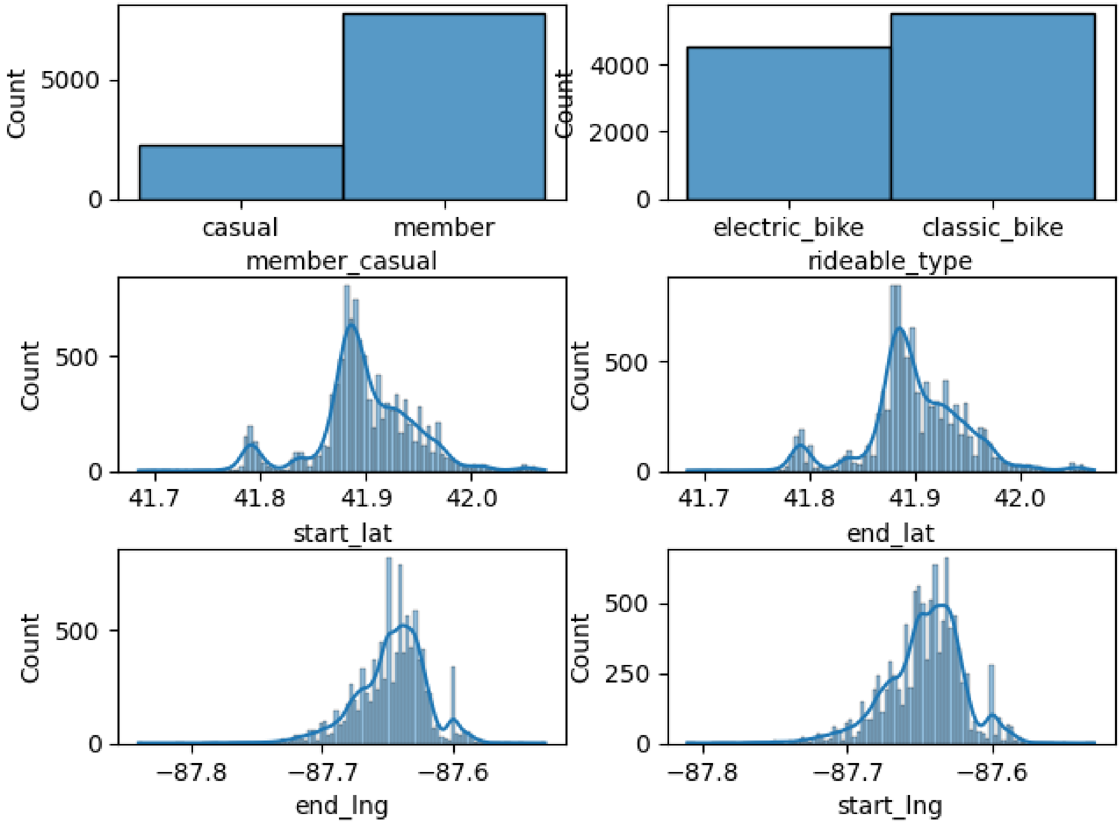
	start_lat	start_lng	end_lat	end_lng
count	10000.000000	10000.000000	10000.000000	10000.000000
mean	41.898804	-87.646559	41.899349	-87.647130
std	0.046850	0.027006	0.047094	0.027162
min	41.684595	-87.810000	41.684595	-87.840000
25%	41.879389	-87.660000	41.879344	-87.661198
50%	41.894716	-87.643353	41.895748	-87.643948
75%	41.926277	-87.629912	41.928830	-87.630000
max	42.070000	-87.530000	42.070000	-87.530000

- Calculate correlation of member_casual vs latitude and longitude

- Plot heatmap to visualize correlation



- Do univariate analysis on some variable taking into account skewness and kurtosis



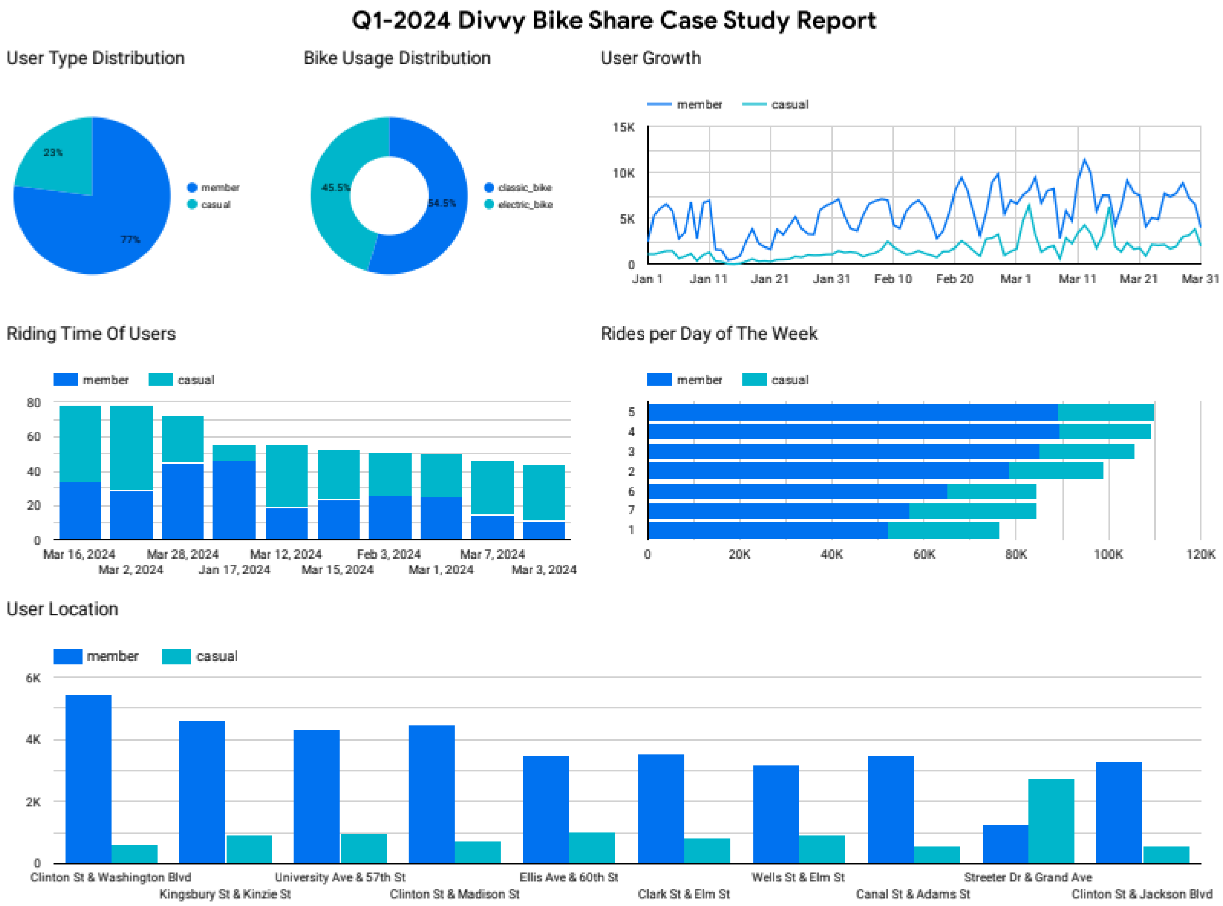
Conclusion: data follow normal distribution pretty tightly, longitude have a significant correlation to member_casual so that might be taken into consideration. Look at the notebook to find out more details and visualization.

5 Visualizations and Key Findings

Here are some key findings i manage to found:

- The number of member riders are exponentially higher than casual riders
- Casual riders on average ride for a longer time but shorter distance
- More casual riders prefer riding in the afternoon and evening
- More casual riders use electric bikes
- Streeter Dr and Grand Ave have the most casual riders

Here is a report visualization build in Looker Studio:



Q1-2024 Divvy Bike Share Case Study Report

User Type Distribution

Bike Usage Distribution

User Growth

Riding Time Of Users

Rides per Day of The Week

User Location

6 Recommendations

Here are some recommendations based on the aforementioned data:

- Increase the number of bikes in Streeter Dr and Grand Ave
- Increase bikes availability at afternoon and evening
- Increase the number of bikes available on Monday, Friday, Saturday