**DATA ANALYSIS OF DIVVY**

**A BIKE SHARING SYSTEM**

**ABOUT THE COMPANY**

Divvy is a bike-share offering system, launched in 2016 and has grown to more than 5823 bicycles that are geo tracked and locked into the network of 692 stations across Chicago.

They charge based on three criteria – single pass, day pass, and annual membership model.

**AIM –**

Divvy’s finance analysts have concluded that annual members are much more profitable than casual riders.

So, our aim is to find How do annual members and casual riders use Divvy bikes differently which in return help them to find Why would casual riders buy Divvy annual memberships.

**DATA ANALYSIS**

Now, we will approach our analysis by taking 6 steps - **ask, prepare, process, analyse, share, and act.**

**ASK-:**

* Business task -: who are key stakeholders.

To find how annual member and casual rider use bikes differently.

To increase the number for annual membership plan

* The data is from May 2020 to April 2021

**PREPARE-:**

* The data is located [here](https://divvy-tripdata.s3.amazonaws.com/index.html) by Motivate International Inc.
* The data is organized in Microsoft excel csv file
* The integrity of data has been verified under this [license](https://ride.divvybikes.com/data-license-agreement)
* Data limitation - data-privacy issues prohibits from using riders’ personally identifiable information. So, we cannot have information like geographic detail of person like address or if they have purchased multiple single passes.
* **ROCCC METHODLOGY**–

**Reliable**: The data is from 2020, the year we got hit by covid. data can be inconsistence, As one of our important audience are daily working people which would buy membership. So, it does not represent overall population.

**Original**: Yes, Motivate International Inc.

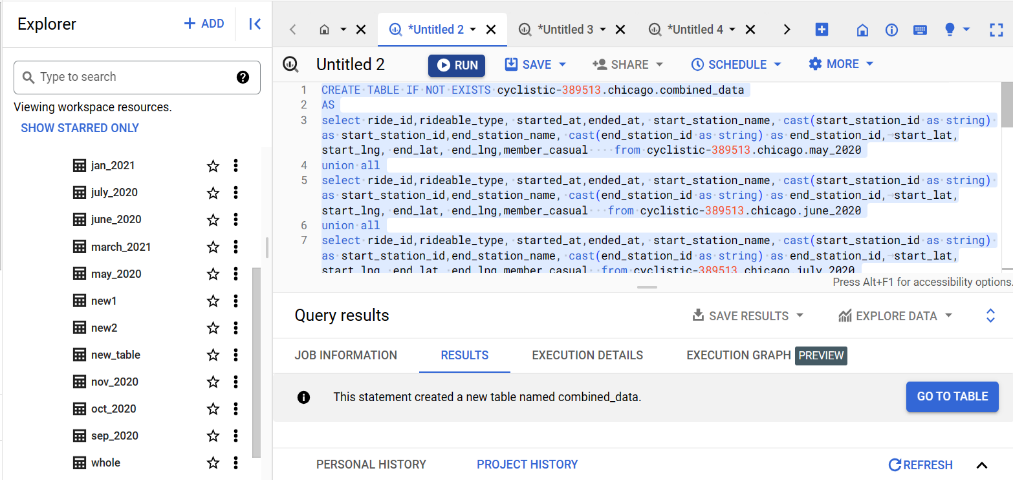
**Comprehensive:** No, the data has many information not provided like address or if they have purchased multiple single passes.

**Current:** yes, update monthly

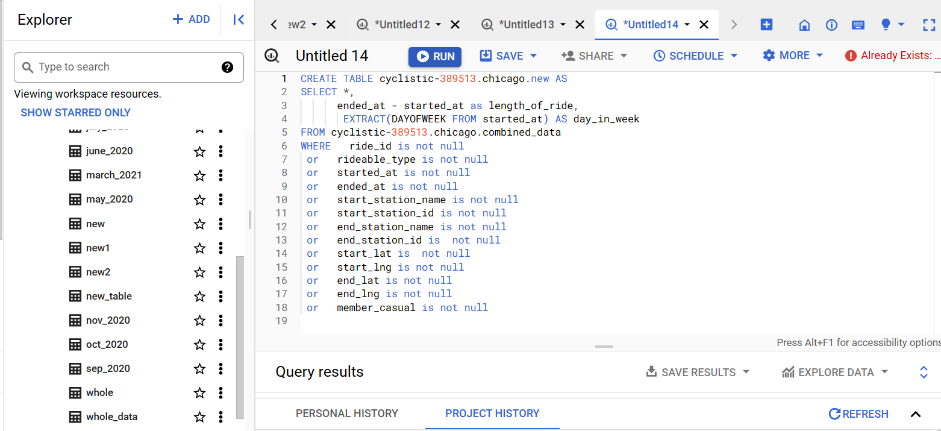
**Cited:** Yes

**PROCESS**

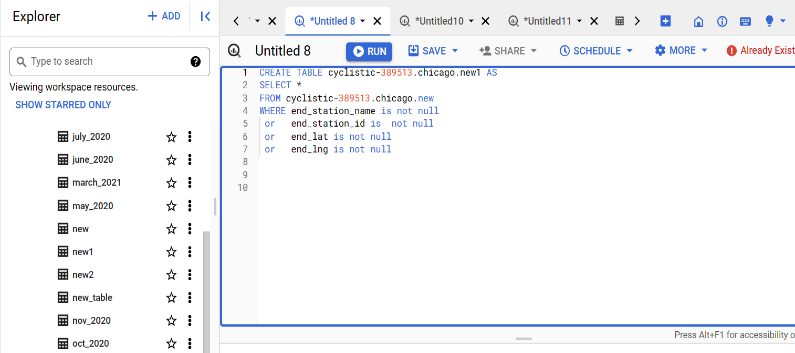
* **Tools used:** EXCEL, Google sheet ,SQL, Tableau.
* **First download the data from above link and load the data into bigquery**
* **Combine the data:**

****

* **Check for data types:** some tables have “start\_station\_id” & “end\_station\_id” column as integer type. So, we got to convert it to string type as we don’t have to do any mathematical calculation to these numbers. It is mentioned above in the picture.
* **Removing null values:** first I remove all the rows in which all the columns are empty.
* **I also added two more columns length of ride and day in week**



We find that only start\_station\_name, start\_station-id, end\_station\_name, end\_station\_id, end\_lat, end\_lng have null values.so for now, we gonna delete the rows which have end\_lat, end\_lng, end\_station\_name, end\_station\_id all null value [3969 rows].

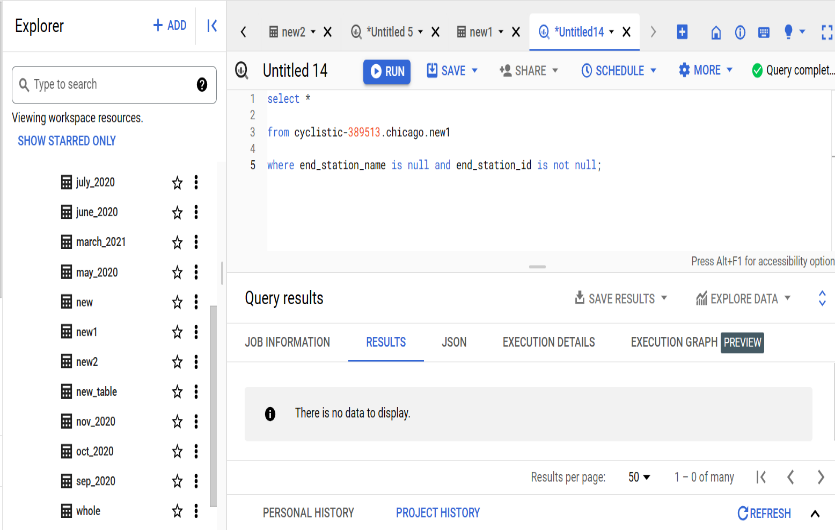
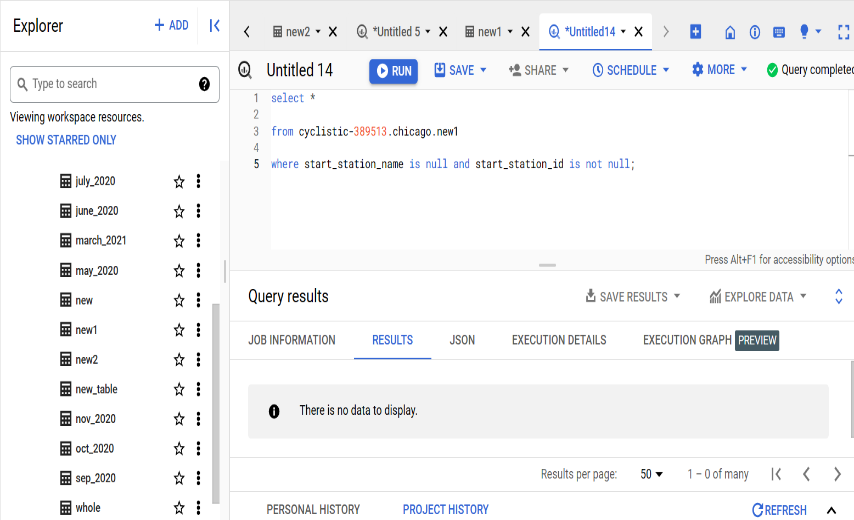


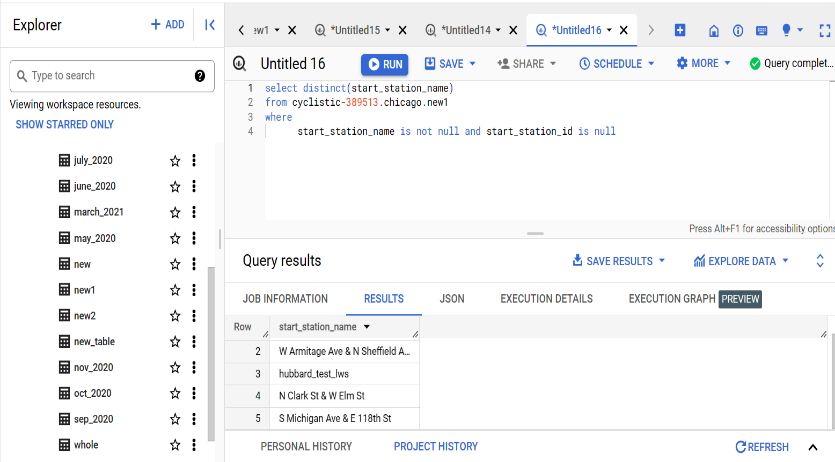
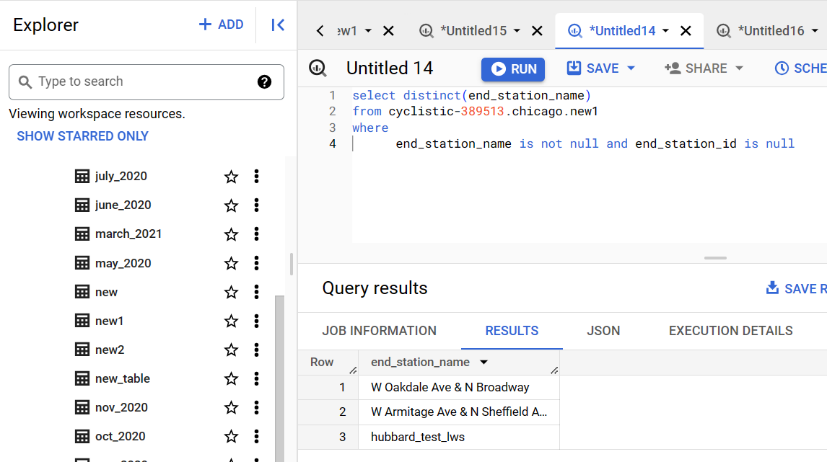
Now,

after finding the null value in all the columns,

we get 0 null count except start\_station\_name, end\_station\_name, start\_station\_id, end\_station\_id.

Now, we can fill the data if we knew any one value in [start\_station\_name , start\_station\_id] and one value in [end\_station\_name , end\_station\_id]

From above (3,4) picture we can find the values we can work with.

We can find the value of the 5 station\_id by their respective name, After finding

**W Oakdale Ave & N Broadway = 20252.0**

**W Armitage Ave & N Sheffield A = 20254.0**

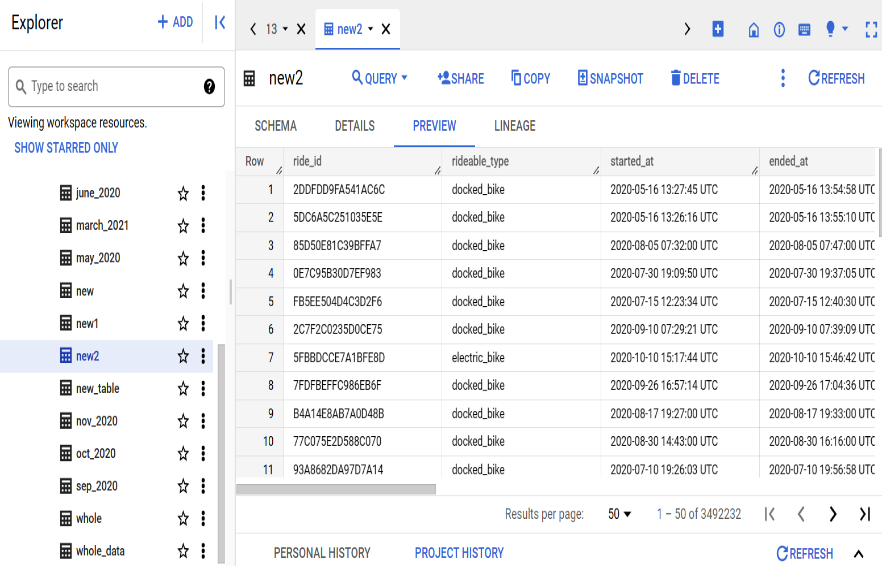
we can only find id for two name.

we can fill this accordingly.

Also,

We know all the latitude and longitude points, so we could fill the empty station\_name’s and station\_id’s

But we have not been provided the value of that for now so, we have to drop the rows which have any null values

**Check for data errors:**

1. In this, we find that rideable\_type have three value “classic\_bike”, “docked\_bike”, “electric\_bike”. But docked\_bike is actually same as classic\_bike so we have to rename the data to “classic\_bike”.
2. Also, we found out that some rows (8,190) in column length\_of\_ride are negative, which have data type in interval so we have to remove that as well.

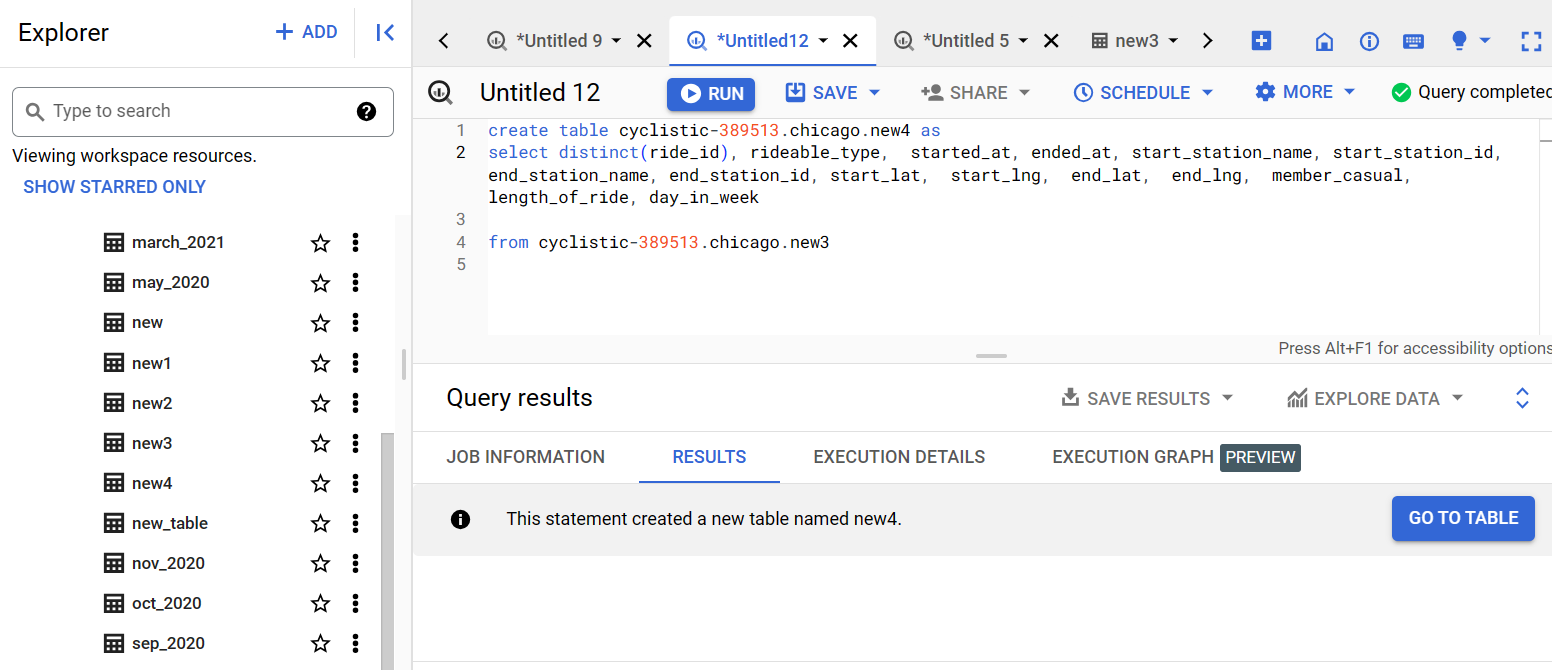


* **Check for duplicate value**

We will check duplicate value.



Now, we will remove the duplicate value.



* **Check for unnecessary range:**

Now, we will check unnecessary range for rideable\_type, started\_at – ended\_at, member\_casual.

If we would have the collection of station name and their id’s we would have checked that as well.

Not found any unnecessary range.

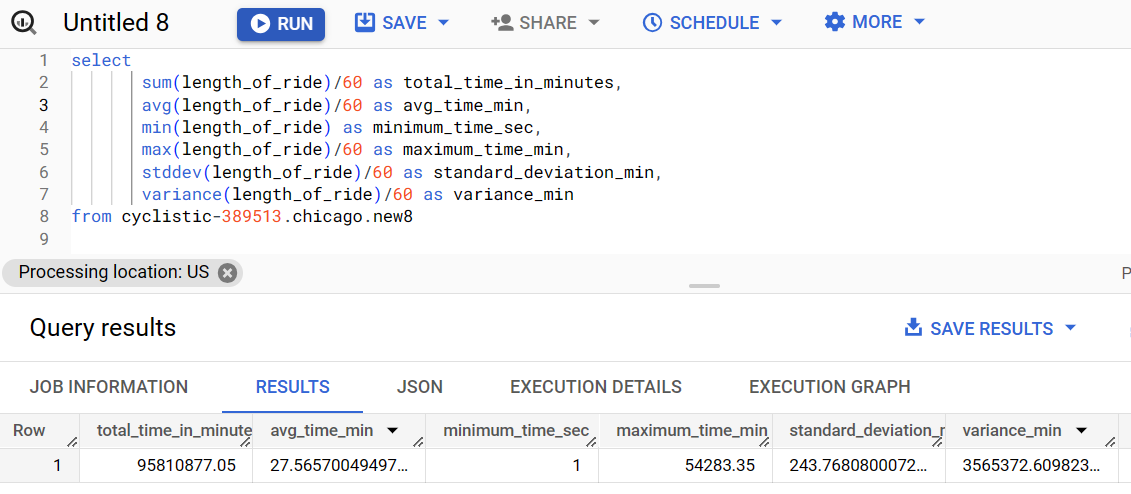
**UPDATE**

1. We converted the column ‘length\_of\_ride’ to minute.
2. We added a new column ‘route’ which is concatenation of ‘start\_station\_name’ and ‘end\_station\_name’.

**Analyze**

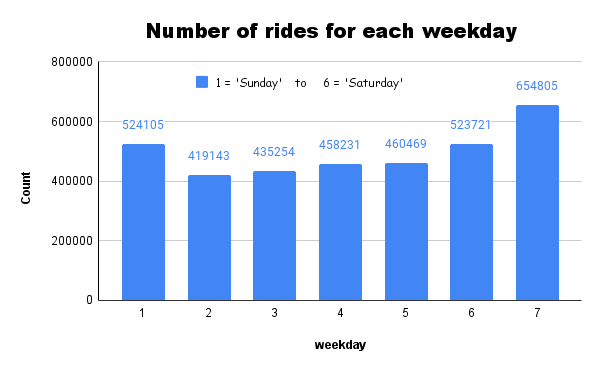
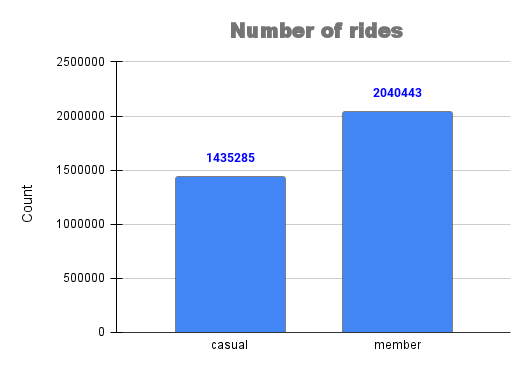
* **Organize** we can sort the table by column started\_at. [ascending order]
* **Descriptive and exploratory statistic:** we can connect[export] SQL table to google sheet for exploratory statistic

**NOTE:** we have converted column length\_of\_ride [interval data type (0-0 0:0:0)] to unit seconds [data type INT64]. So, to add column to google sheet.

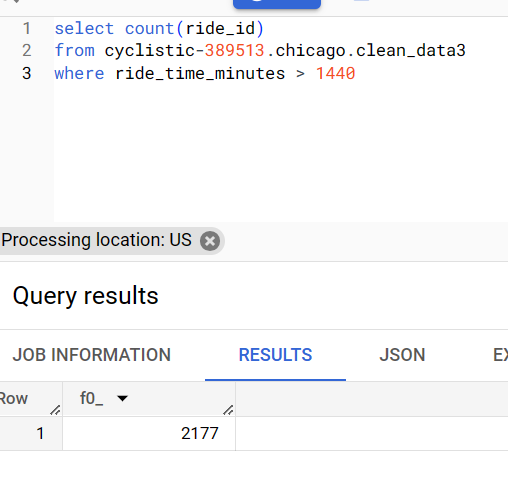
****

Descriptive analysis of column “length\_of\_ride”

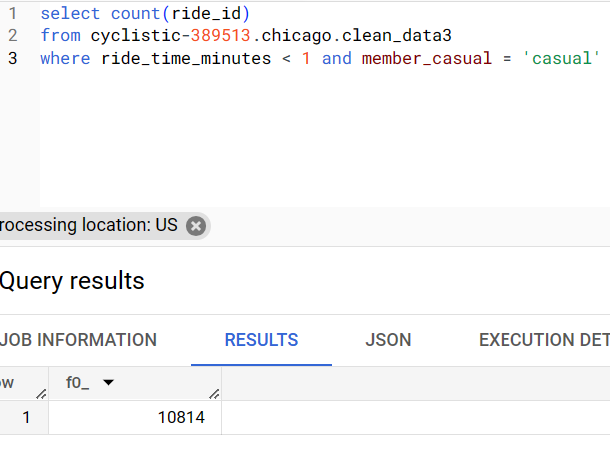
The std deviation is very high. So, there can be some anomalies or outliers in the data.

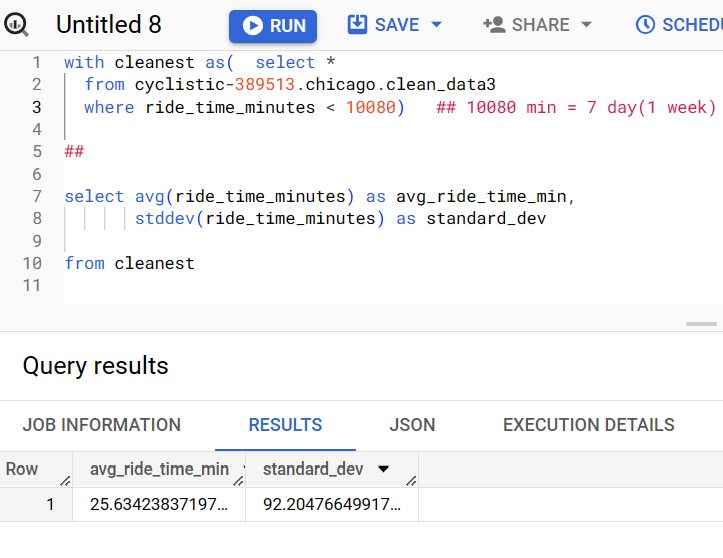
 

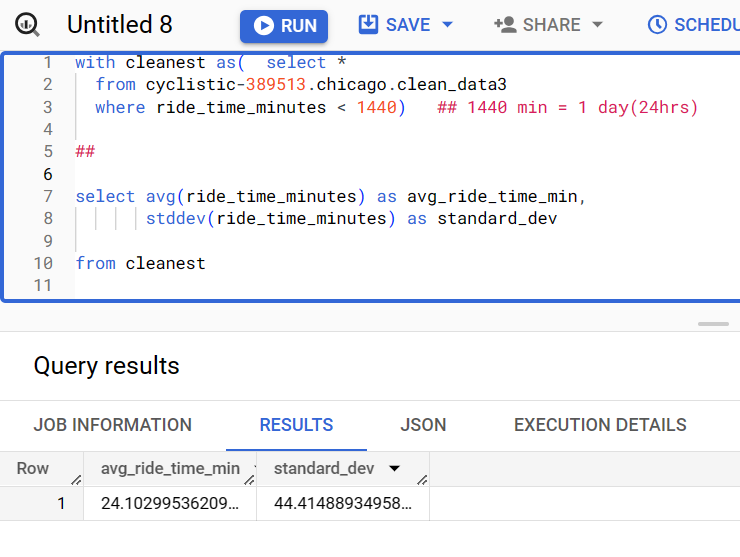
Some anomalies with ride time that is under 1 min and over 1440 min (1 day). This is because under 1 min ride is considered a ride initiated but not further executed and above 1440min company charges some extra money and they are comparatively very low number. So it can be either legit or fault from not locking bike properly.

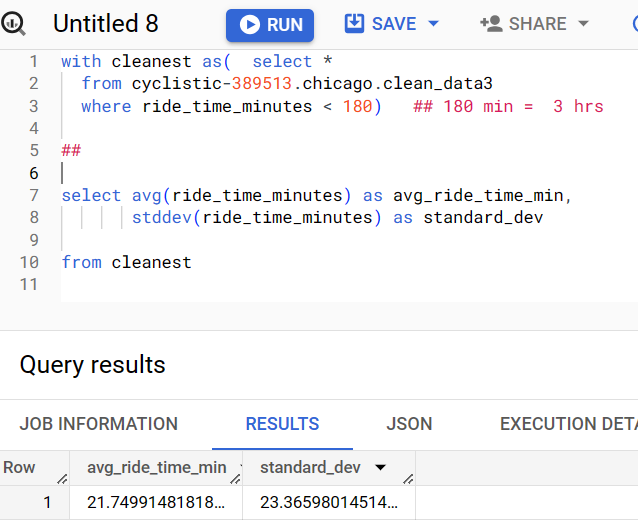
 

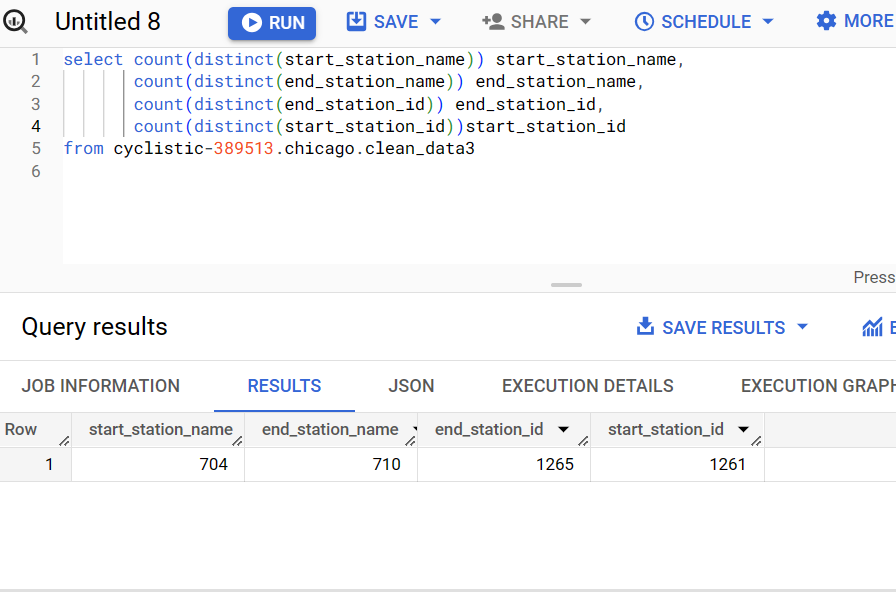
For casual riders,

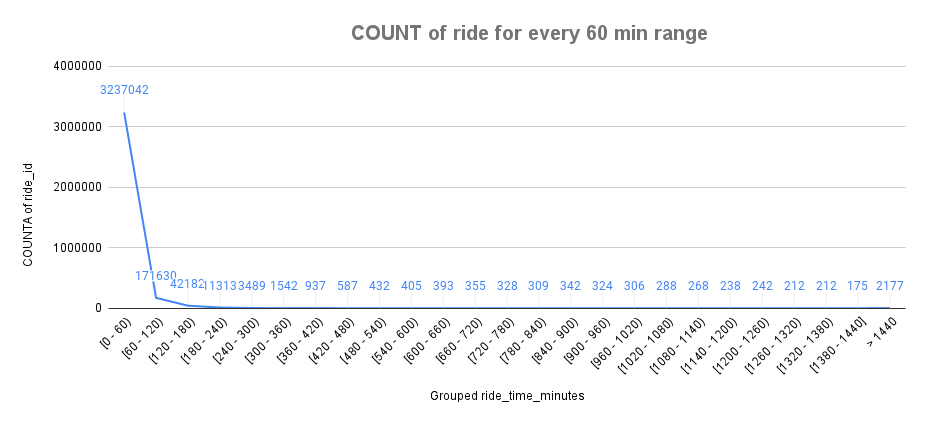


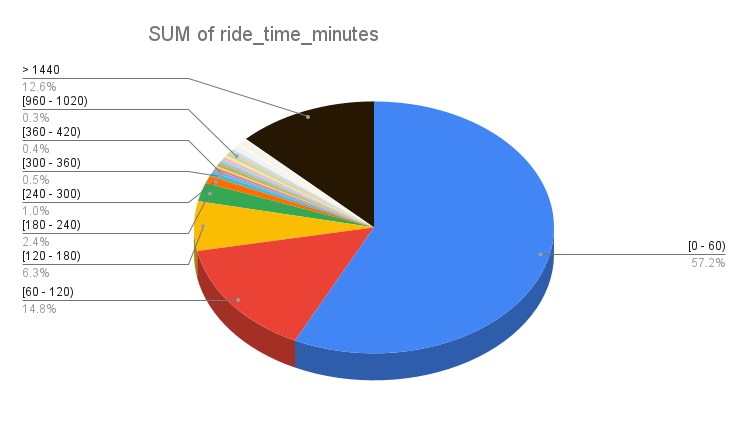
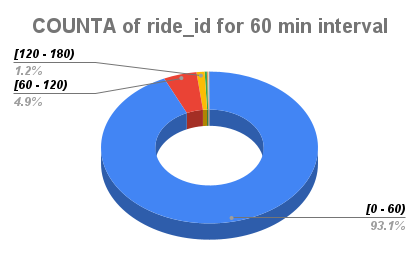






From above chart, we conclude that same station name has multiple id. Maybe they got updated.





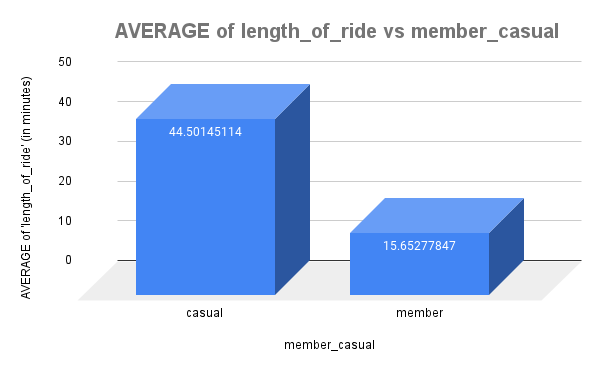


From above graph we can see that 93% of single ride complete under 60 minutes and

99.2 % rides complete under 3 hrs.

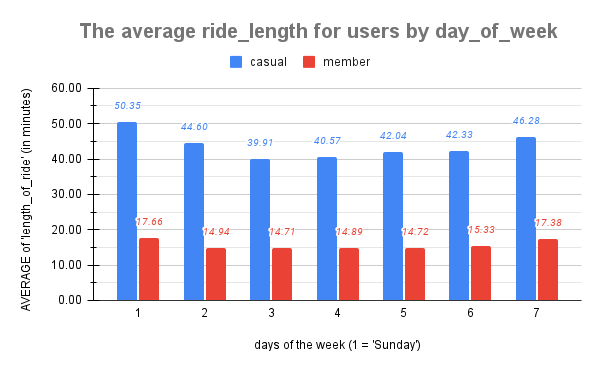
**NOW,**

1. **The average ride\_length(in minutes) for members and casual riders.**



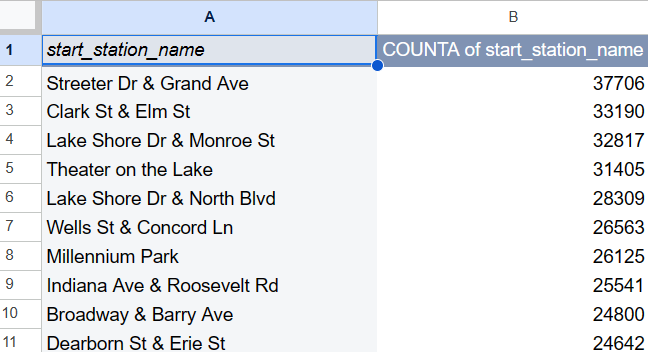
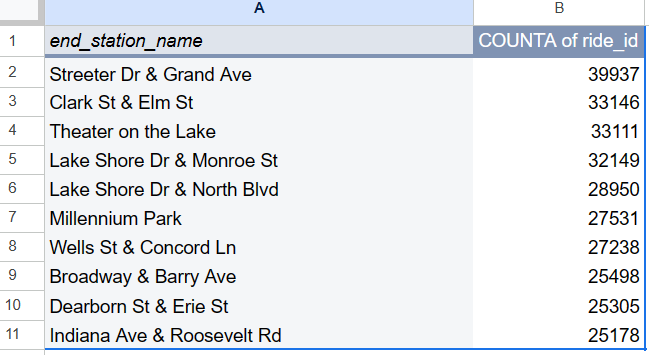
So, from here we can take the average length\_of\_ride is higher for casual but total number of rides taken is higher of member

1. **The average length\_of\_ride for users by day\_of\_week.**

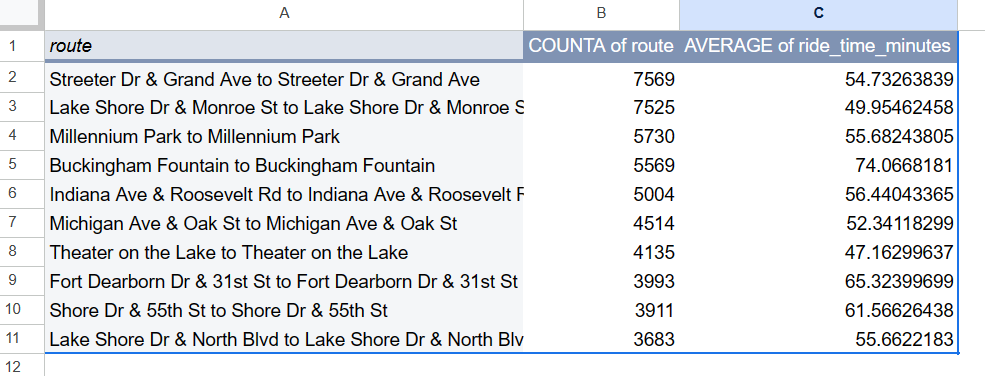
****

**Update: we change the column ‘length\_of\_ride’ to ‘ride\_time\_minutes’ to have a better understanding**

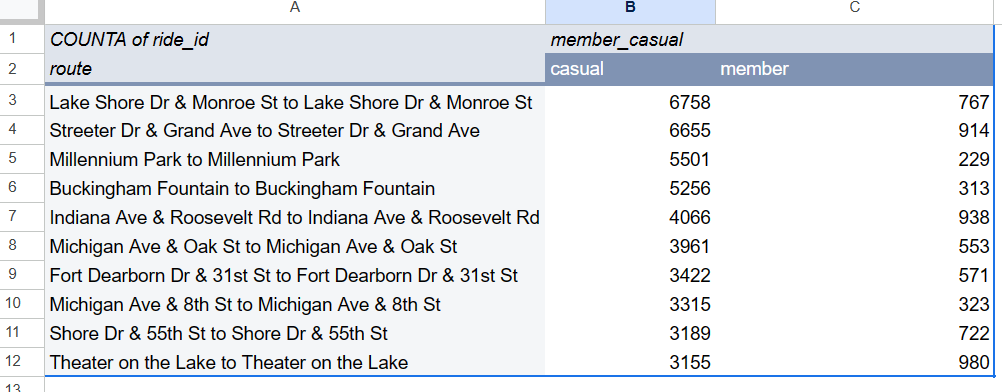
1. **Top 10 start\_station\_name & end\_station\_name by number of rides.**

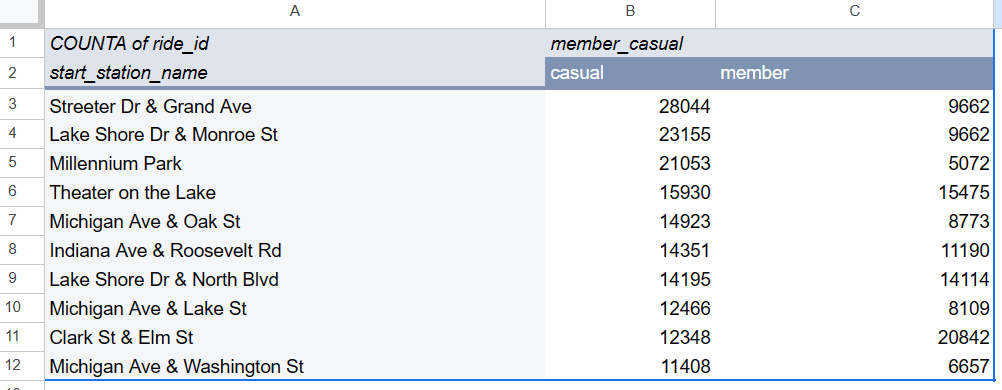
** **

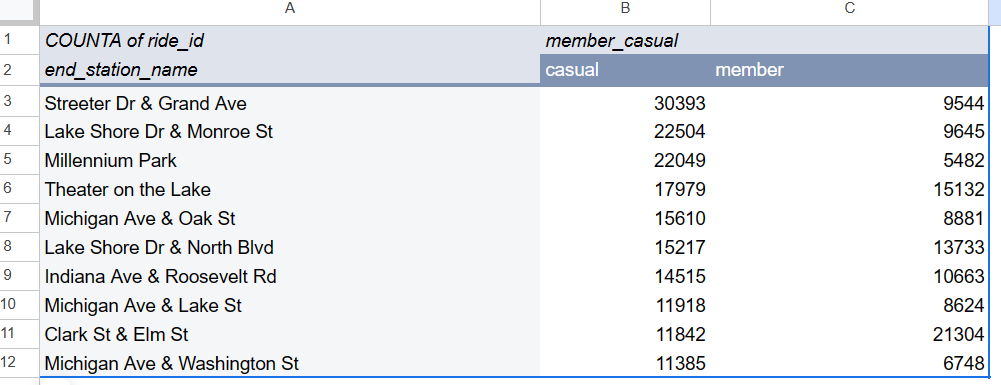
1. **Top 10 popular common route.**

****

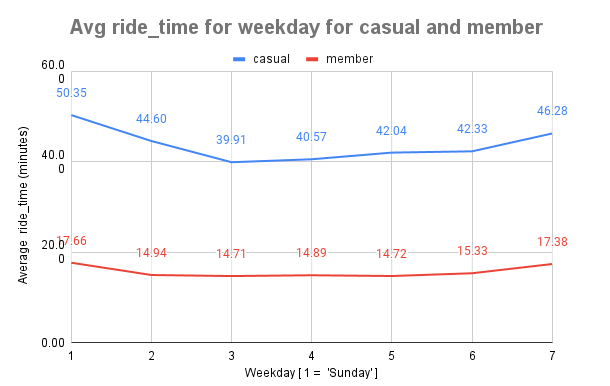
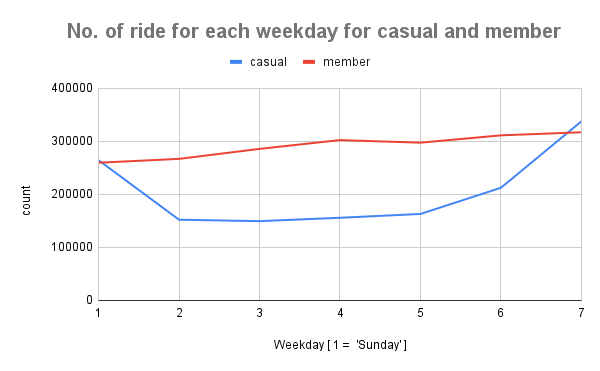
Now, we will do the same above chart but including casual and member and sort by casual (descending)

****

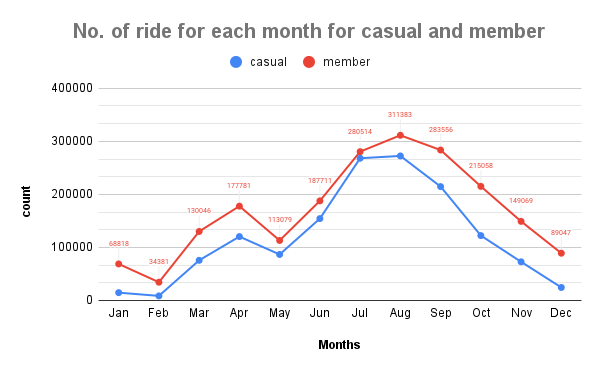
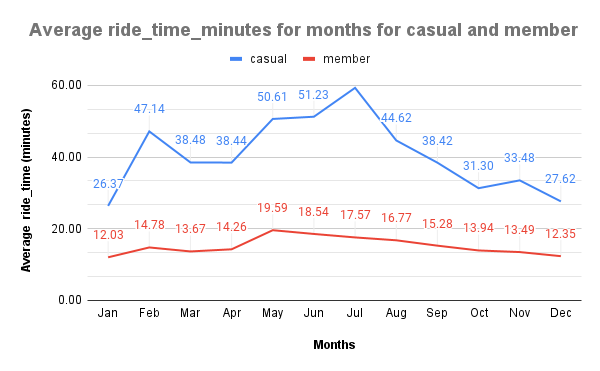
****

****

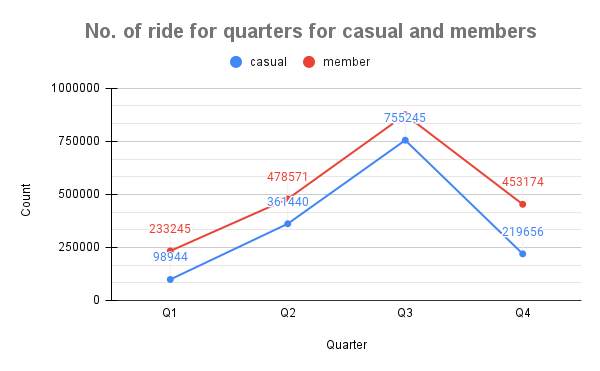
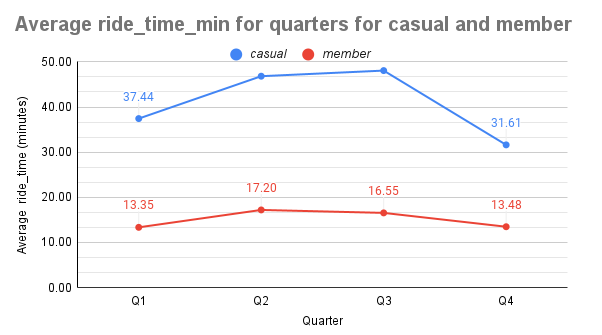
1. **Count and avg ride\_time for each weekday**

****

1. **Count and avg ride\_time for each month**

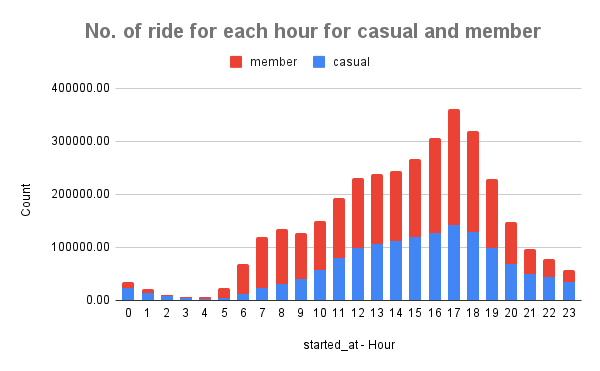
** **

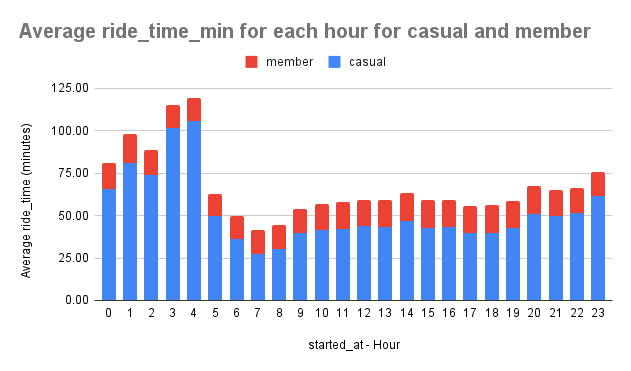
1. **Count and avg ride\_time for each quarter**

** **

So, from above 6 graphs we can analyze that Saturday and Sunday have the maximum count and avg\_time of rides for casuals and also Q3[July, August, September] have the maximum count and avg\_time of ridescompared to other months. therefore, we can target the summer and fall season for marketing.

Now, we do the count and average time per hour



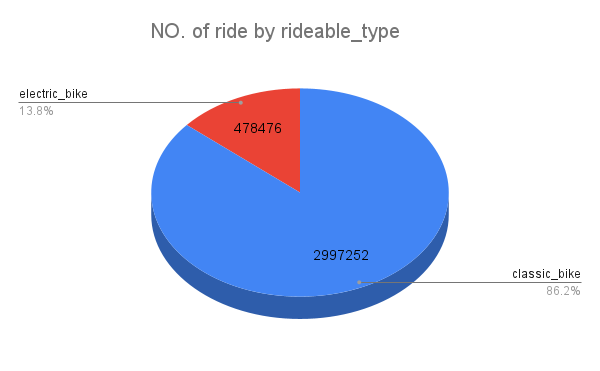
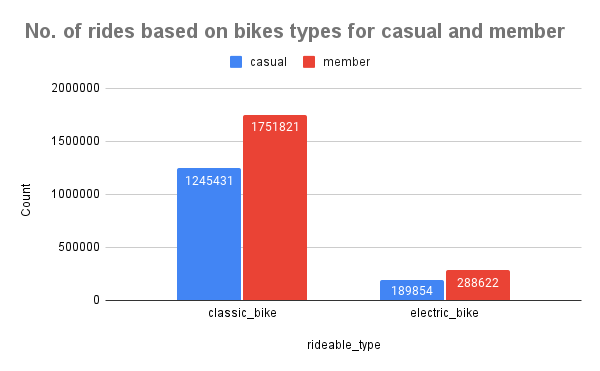


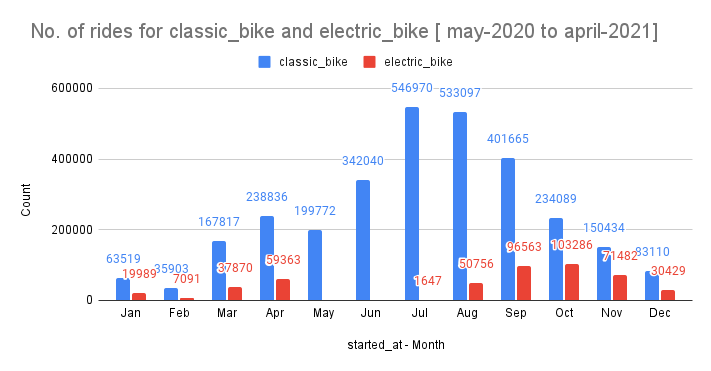
From, above two chart we can see that no. of rides peaks at [15:00 to 18:00]. But the average ride\_time

Peak is [00:00 to 4:00 A.M]. therefore, we can see that between [15:00 to 18:00 P.M] there would be many

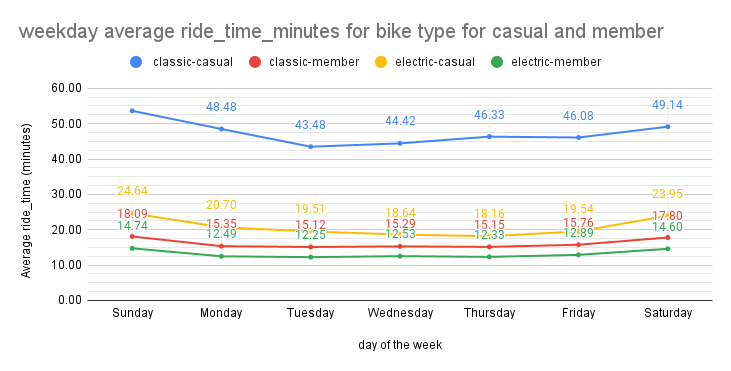
Short rides

1. Count of rides for bike type.



In above chart , we can see aug-nov have a bit higher growth in count for electric bike and lower in winter.



In above chart, we can see casual riders have average time higher than member with any bike.

**CONCLUSION**: From 01-05-2020 to 31-04-2021

* The number of rides and average of ride in minutes

BY TOTAL:

1. The total number of rides taken by casual riders is 14,35,285(41.29%) and by member is 20,40,443(58.71%).
2. The member’s average ride is approx. (15-16) minutes including or excluding anomalies. Whereas casual’s average ride including full time frame is 44.50 min. and for time frame [ 1 – 1440 min] is 36.83 and for time frame [0.02 – 180 min] which consist 98.42% ride for casual is 31.79min.

By hours:

1. The number of rides significantly rises between 16:00 PM to 18:00 PM for both casual and member (28.3%).
2. The average ride for casual peaks around 4:00 AM with 105.93min and for member it is [14-16] min overall. but for time frame [0 -180] the casuals peaks around [13:00 – 14:00 pm] with 34.76 min.

By weekday:

1. The number of rides for member gradually increases from Sunday to Saturday whereas for casual the Saturday has the highest peak followed by Saturday.
2. The average ride for casual peaks on weekends [46.28 and 50.35] min and for members too on weekends [17.38-17.66] min. but for time frame [0 -180] the casuals peaks on weekends with [33.97 and 35.54].

By month:

1. The number of rides for casual and member rider’s peaks around august. And is very low whole winter relative to other seasons.
2. The average ride for casual peak on July [59.29] min and for member may [19.59] min. but for time frame - [0-180] the casual peaks on may [37.02] min and for member too on may [18.60] min.

By quarter:

1. The number of rides for casual peaks on quarter 3 with [7,55,245] and for member on quarter 3 [8,75,443] having combined 46.93% of total rides on quarter 3.
2. The average ride for casual peaks on quarter 3 [48.07] min and for member on quarter 2 [17.07] min. but with time frame [0-180] min. it is approx. equal for quarter 2 and quarter 3 for both casual and member.

* Now for electric bike we can see the number of rides for casual and members looks favourable for months of October and November.
* Now number of rides for casual for range [0-60] min is 84.68%, for range [60-120] min is 10.94% and for range [120- 180] min is 2.80%. combining a total of 98.42%

**RECCOMENDATION**

* The marketing of the product can be done at the peak by hour, months, weekdays or quarter which is suitable.
* The discount could be attractive during summer and fall time accordingly.
* The subscription model should work within [35-45min to 3hour] single ride range.
* We can also target casual riders by the busiest or highest count of rides by station name.
* The electric bike promotion can be done during winter

THE DASHBOARD LINK IS [HERE](https://public.tableau.com/views/capstone_16881200343520/cyclisticdashboard?:language=en-US&:display_count=n&:origin=viz_share_link)