Research Report: SQL

Valentin Jimenez

02/02/2022

Second Part: Analysis with SQL

Target question: 1. How do annual members and casual riders use Cyclistic bikes differently?

After we have made the basic analysis we used RStudio to create a unique dataset containing the twelve months of the year withing a csv file with the following characteristics:

name: bigDatasetCSV.csv, 5 595 062 observations of 19 variables

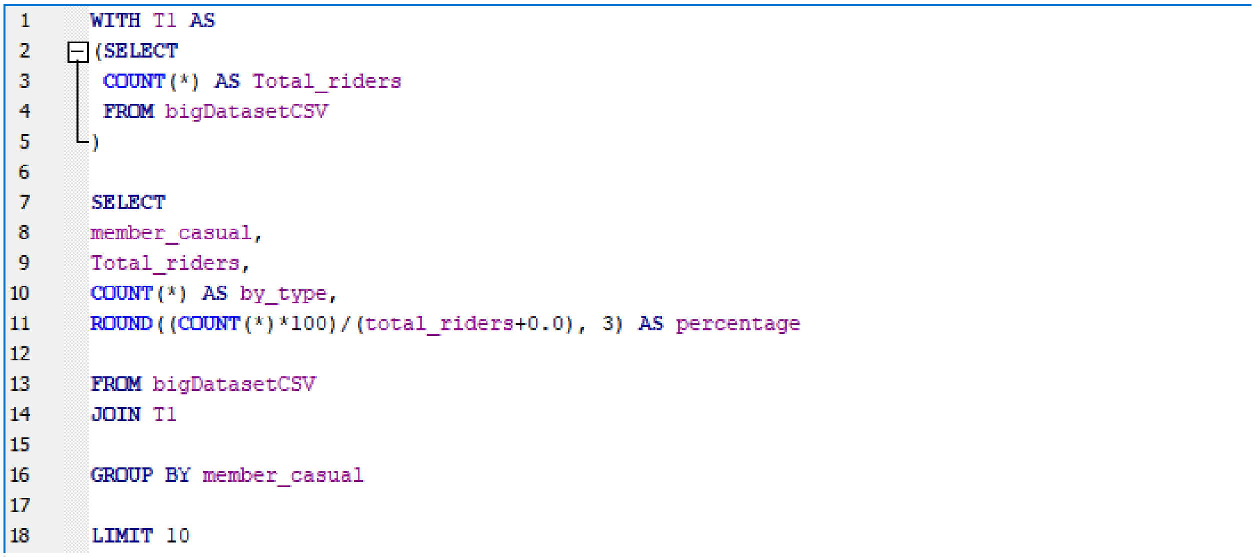
This is the reason why now we need to use a tool different from a spreadsheet, we start using SQL to continue with our research. I uploaded our CSV file bigDatasetCSV.cvs and created a database to work with in SQLite called **bigDatasetCSV.db**

In the following analysis first we show the queries I used to manipulate the data and then I show the corresponding dashboard made with Tableau.

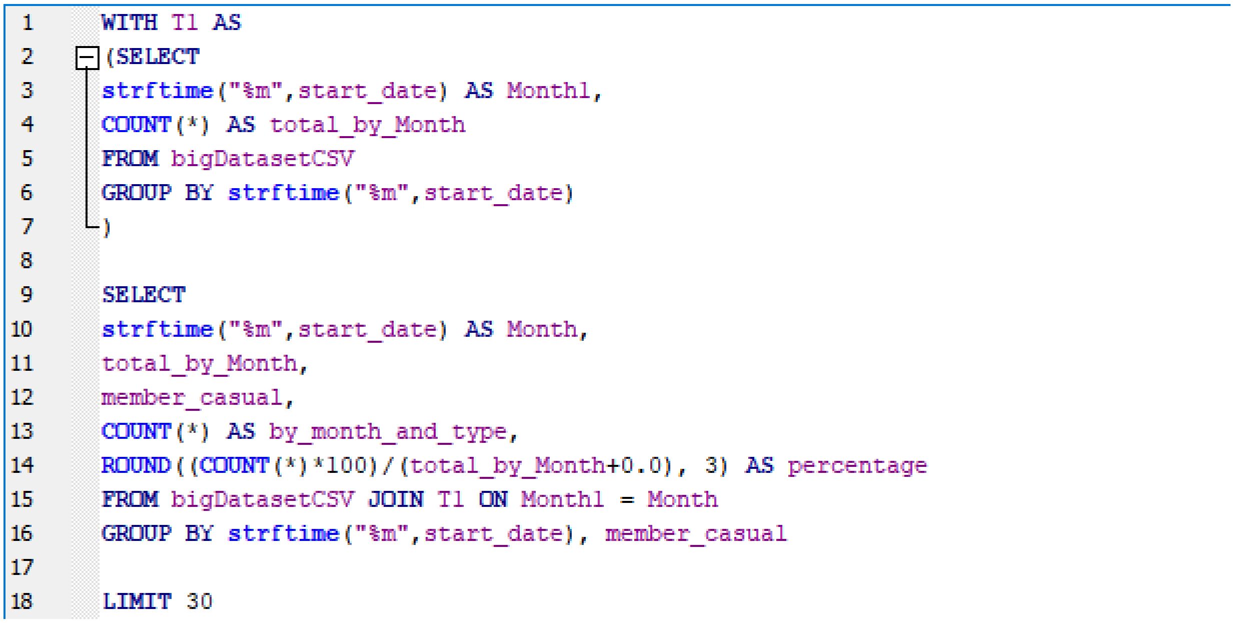
1. In part 1. of the analysis with excel we saw that the percentage of casual riders was not big enough for December 2021; however, doing the same analysis annually with biDatasetCSV.db, we can see in the following annual pie chart that **45.2%** of riders are casual riders, which is a significant amount. To obtain these results I made queries in Sqlite and the visualizations were made in a dashboard using Tableau. These are interactive dashboards, so to have a better visualization you can visit Dashboard 1 in tableau-1.

The second graph indicates the quantity of casual and member riders for each of the 12 months, we can see that during the months of November, December, January and February casual riders decreases a lot. Meanwhile during May, June, July and August the quantity of each type of rider is pretty balanced.

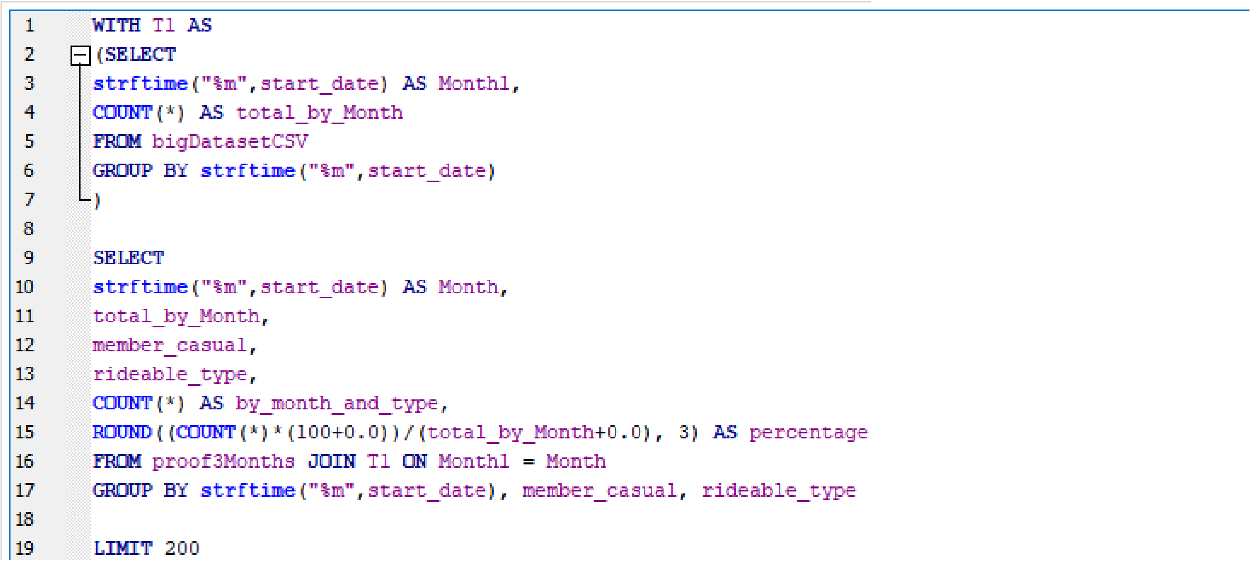
The explanation of the 3rd graph is as follows: the 100% percentage is by month and as such is distributed between casual and member riders, but also by the 3 types of services, so we should add all bars corresponding to month 1. This graph confirms that docked bikes are poorly used by both riders, but casual riders using this service could be a potential target. On the other hand, May, Jun, July and August, corresponding to casual riders using classical bikes are a good potential target.



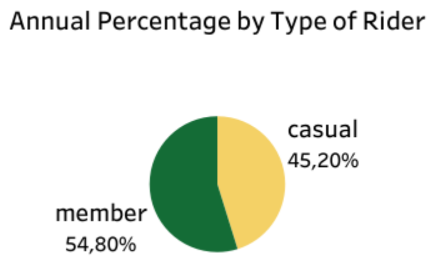
```sql
WITH T1 AS
(SELECT
 COUNT(*) AS Total_riders
 FROM bigDatasetCSV
)

SELECT
member_casual,
Total_riders,
COUNT(*) AS by_type,
ROUND((COUNT(*)*100)/(total_riders+0.0), 3) AS percentage

FROM bigDatasetCSV
JOIN T1

GROUP BY member_casual

LIMIT 10
```

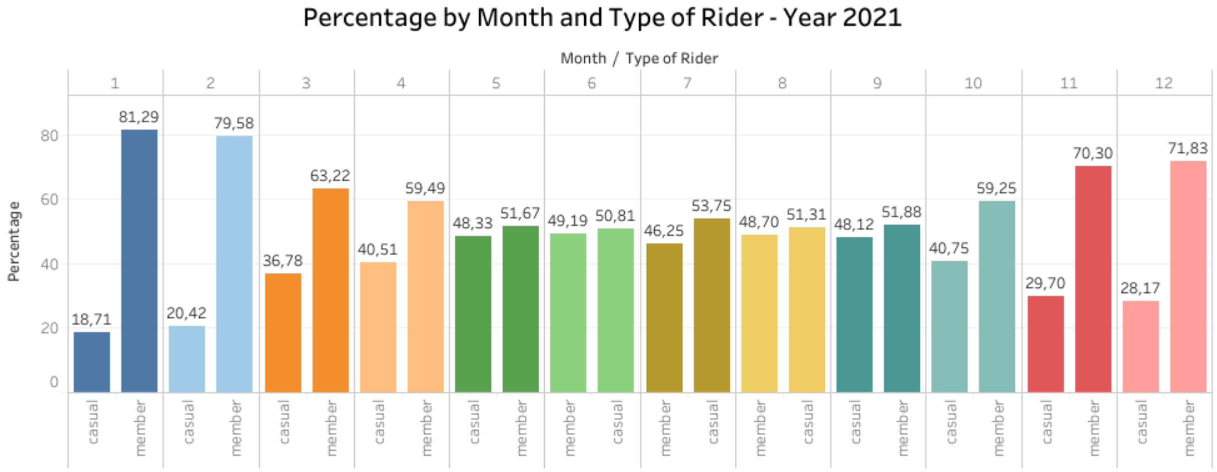


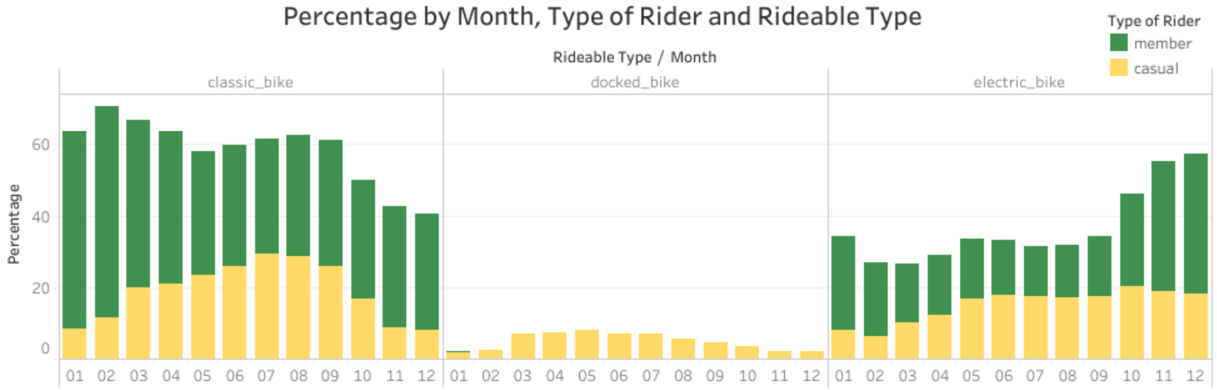
```sql
WITH T1 AS
(SELECT
strftime("%m",start_date) AS Month1,
COUNT(*) AS total_by_Month
FROM bigDatasetCSV
GROUP BY strftime("%m",start_date)
)

SELECT
strftime("%m",start_date) AS Month,
total_by_Month,
member_casual,
COUNT(*) AS by_month_and_type,
ROUND((COUNT(*)*100)/(total_by_Month+0.0), 3) AS percentage
FROM bigDatasetCSV JOIN T1 ON Month1 = Month
GROUP BY strftime("%m",start_date), member_casual

LIMIT 30
```



```
WITH T1 AS
(SELECT
strftime("%m",start_date) AS Month1,
COUNT(*) AS total_by_Month
FROM bigDatasetCSV
GROUP BY strftime("%m",start_date)
)

SELECT
strftime("%m",start_date) AS Month,
total_by_Month,
member_casual,
rideable_type,
COUNT(*) AS by_month_and_type,
ROUND((COUNT(*)*(100+0.0))/(total_by_Month+0.0), 3) AS percentage
FROM proof3Months JOIN T1 ON Month1 = Month
GROUP BY strftime("%m",start_date), member_casual, rideable_type

LIMIT 200
```



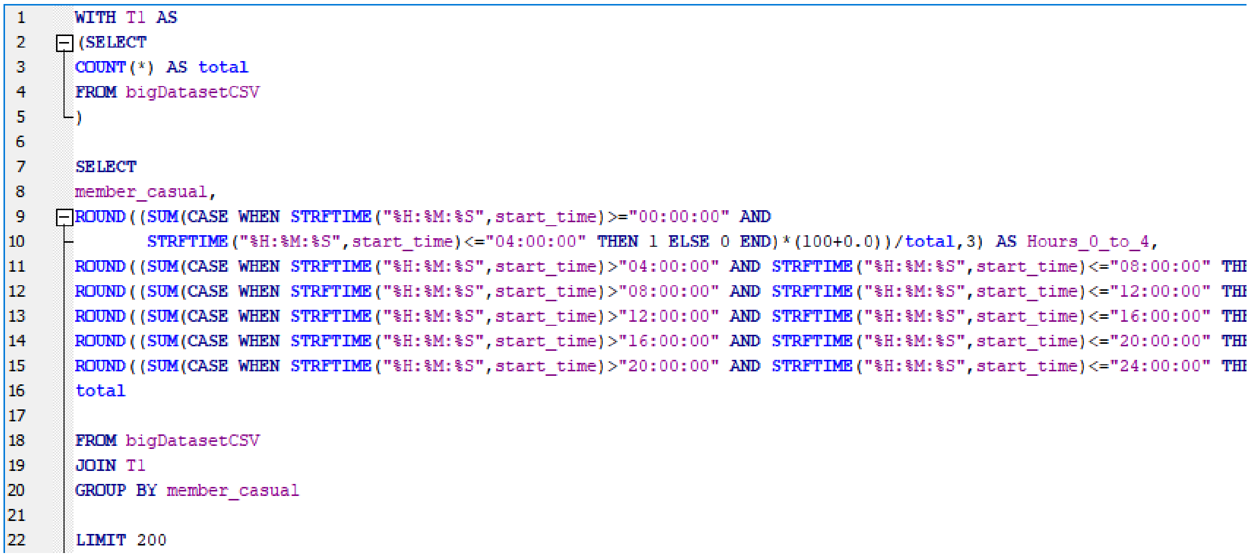




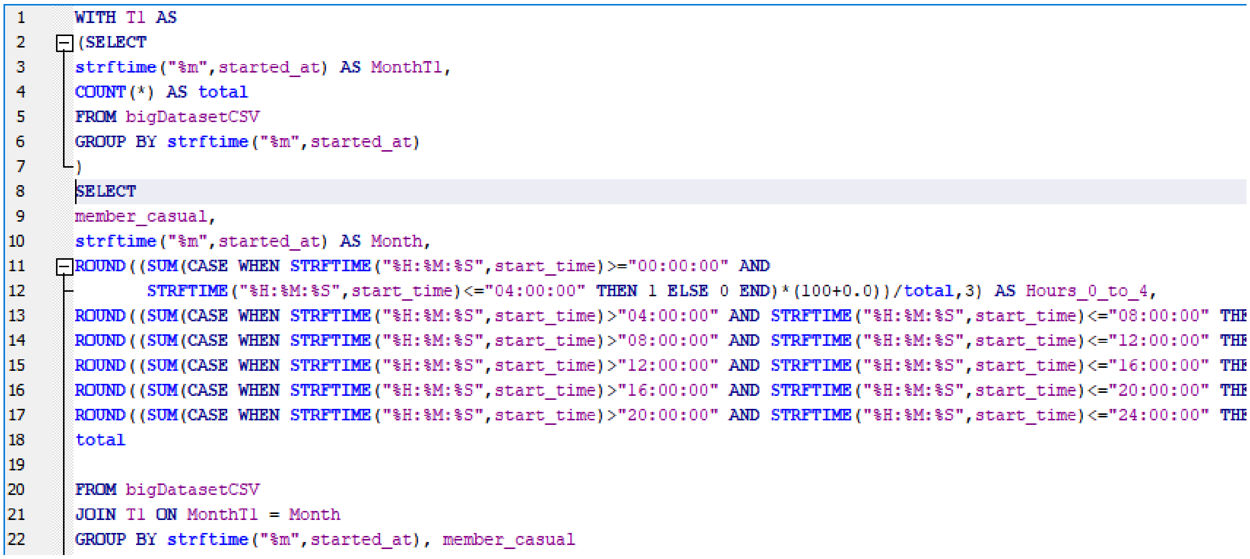
2. Analysis by hours in a day

To have a better visualization visit Dashboard 2 in tableau-2.

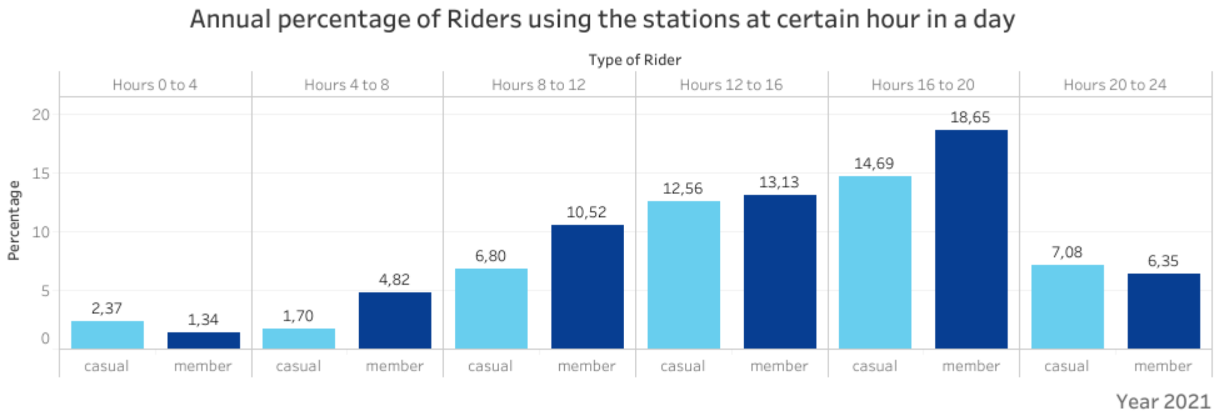
We observe that annually there is no big difference in the pattern of use for each type of rider, but we observe that the peak use is from 16:00 hours to 20:00 hours. However, in the second visualization we observe that the pattern for each month is significantly different: during months May, Jun, July, August and September, the use of the service by casual riders increases a lot and in particular from 16:00 to 24:00 hours, the peak being from 16:00 to 20:00 hours (note that this is coherent with the results found in Dashboard 1).

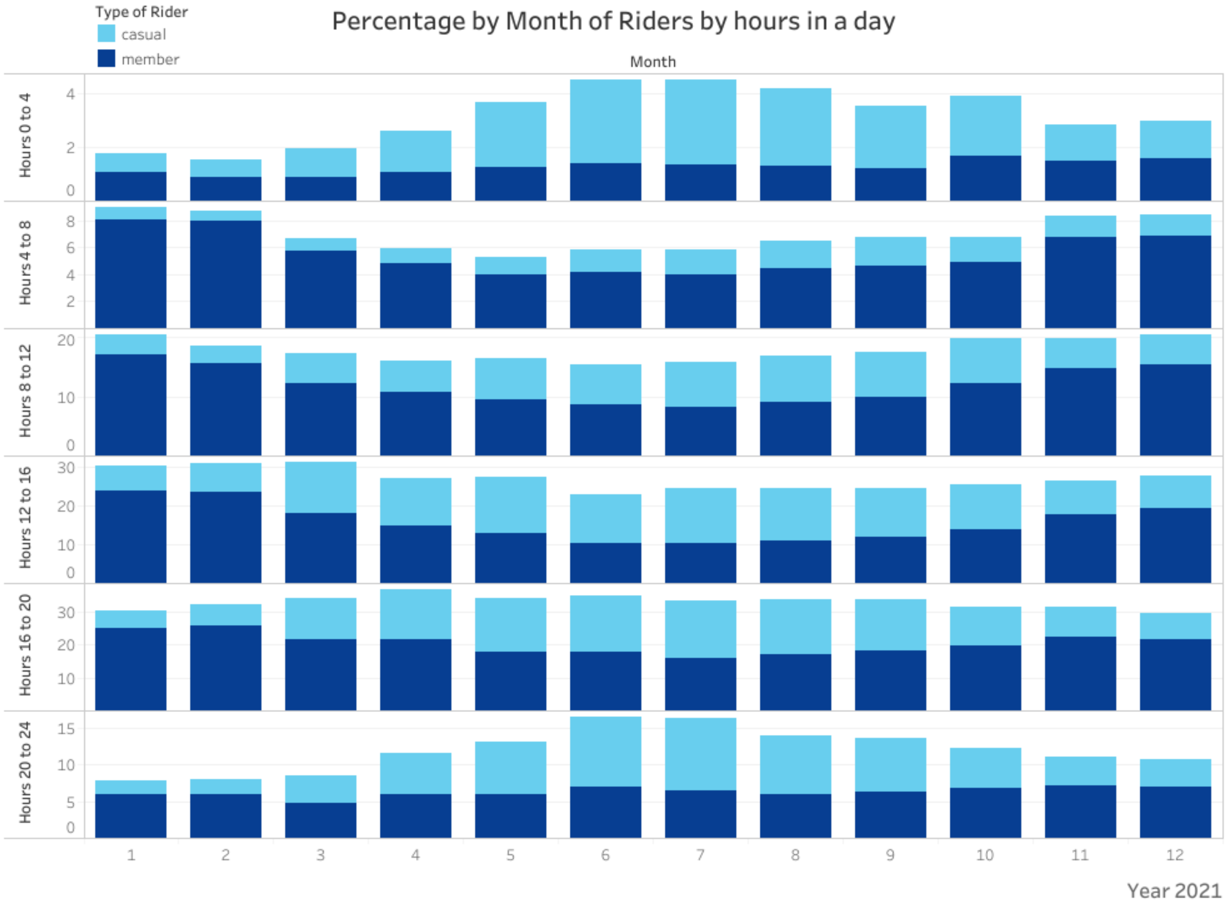


```
WITH T1 AS
(SELECT
COUNT(*) AS total
FROM bigDatasetCSV
)

SELECT
member_casual,
ROUND((SUM(CASE WHEN STRFTIME("%H:%M:%S",start_time)>="00:00:00" AND
        STRFTIME("%H:%M:%S",start_time)<="04:00:00" THEN 1 ELSE 0 END)*(100+0.0))/total,3) AS Hours_0_to_4,
ROUND((SUM(CASE WHEN STRFTIME("%H:%M:%S",start_time)>"04:00:00" AND STRFTIME("%H:%M:%S",start_time)<="08:00:00" THE
ROUND((SUM(CASE WHEN STRFTIME("%H:%M:%S",start_time)>"08:00:00" AND STRFTIME("%H:%M:%S",start_time)<="12:00:00" THE
ROUND((SUM(CASE WHEN STRFTIME("%H:%M:%S",start_time)>"12:00:00" AND STRFTIME("%H:%M:%S",start_time)<="16:00:00" THE
ROUND((SUM(CASE WHEN STRFTIME("%H:%M:%S",start_time)>"16:00:00" AND STRFTIME("%H:%M:%S",start_time)<="20:00:00" THE
ROUND((SUM(CASE WHEN STRFTIME("%H:%M:%S",start_time)>"20:00:00" AND STRFTIME("%H:%M:%S",start_time)<="24:00:00" THE
total

FROM bigDatasetCSV
JOIN T1
GROUP BY member_casual

LIMIT 200
```



```
WITH T1 AS
(SELECT
strftime("%m",started_at) AS MonthT1,
COUNT(*) AS total
FROM bigDatasetCSV
GROUP BY strftime("%m",started_at)
)
SELECT
member_casual,
strftime("%m",started_at) AS Month,
ROUND((SUM(CASE WHEN STRFTIME("%H:%M:%S",start_time)>="00:00:00" AND
        STRFTIME("%H:%M:%S",start_time)<="04:00:00" THEN 1 ELSE 0 END)*(100+0.0))/total,3) AS Hours_0_to_4,
ROUND((SUM(CASE WHEN STRFTIME("%H:%M:%S",start_time)>"04:00:00" AND STRFTIME("%H:%M:%S",start_time)<="08:00:00" THE
ROUND((SUM(CASE WHEN STRFTIME("%H:%M:%S",start_time)>"08:00:00" AND STRFTIME("%H:%M:%S",start_time)<="12:00:00" THE
ROUND((SUM(CASE WHEN STRFTIME("%H:%M:%S",start_time)>"12:00:00" AND STRFTIME("%H:%M:%S",start_time)<="16:00:00" THE
ROUND((SUM(CASE WHEN STRFTIME("%H:%M:%S",start_time)>"16:00:00" AND STRFTIME("%H:%M:%S",start_time)<="20:00:00" THE
ROUND((SUM(CASE WHEN STRFTIME("%H:%M:%S",start_time)>"20:00:00" AND STRFTIME("%H:%M:%S",start_time)<="24:00:00" THE
total

FROM bigDatasetCSV
JOIN T1 ON MonthT1 = Month
GROUP BY strftime("%m",started_at), member_casual
```

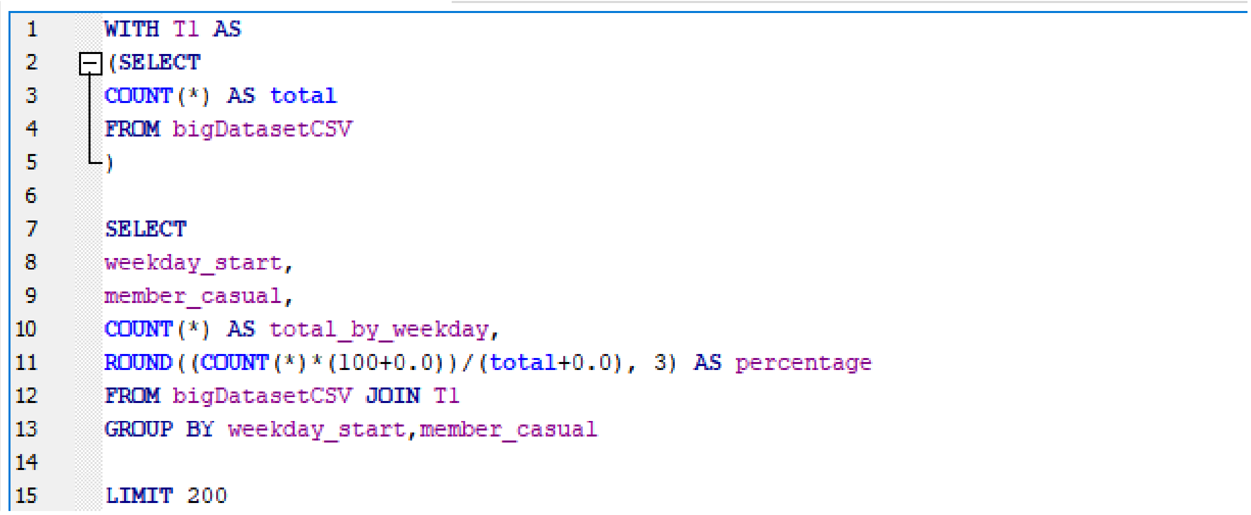




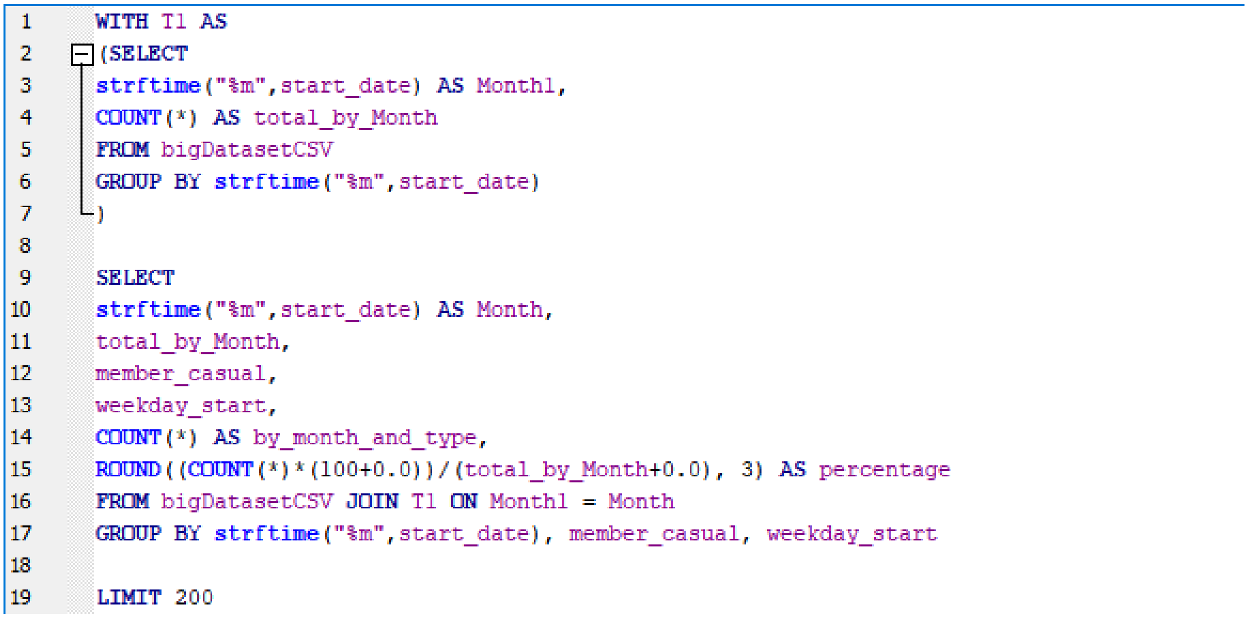
3. Analysis by day of the week

To have a better visualization visit Dashboard 3 in tableau-3.

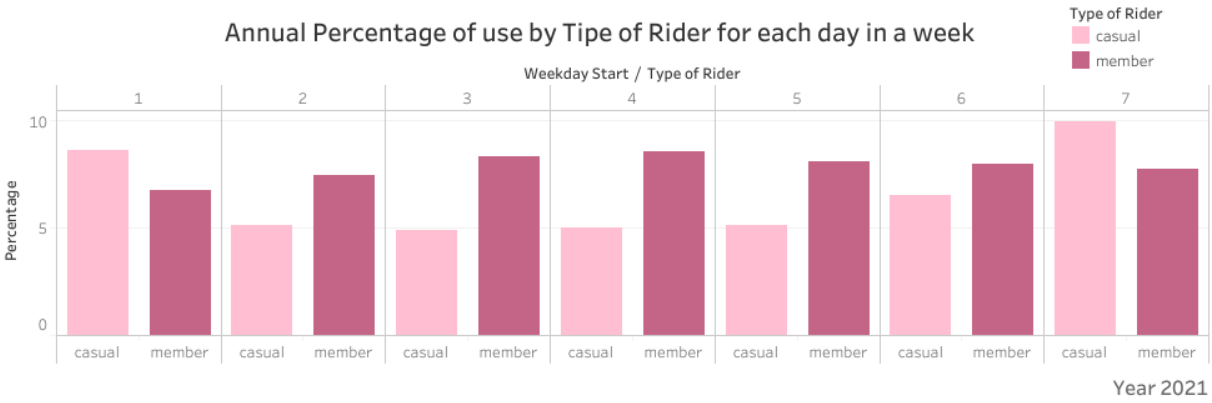
Annually, we confirm what we found for December 2021, casual riders use the service a lot on Saturdays and Sundays, the rest of the week member riders use the service more. The same answer is confirmed for each particular month along the whole year. Also we note that for months May, June, July and August the use of casual riders increases for each day of the weak.

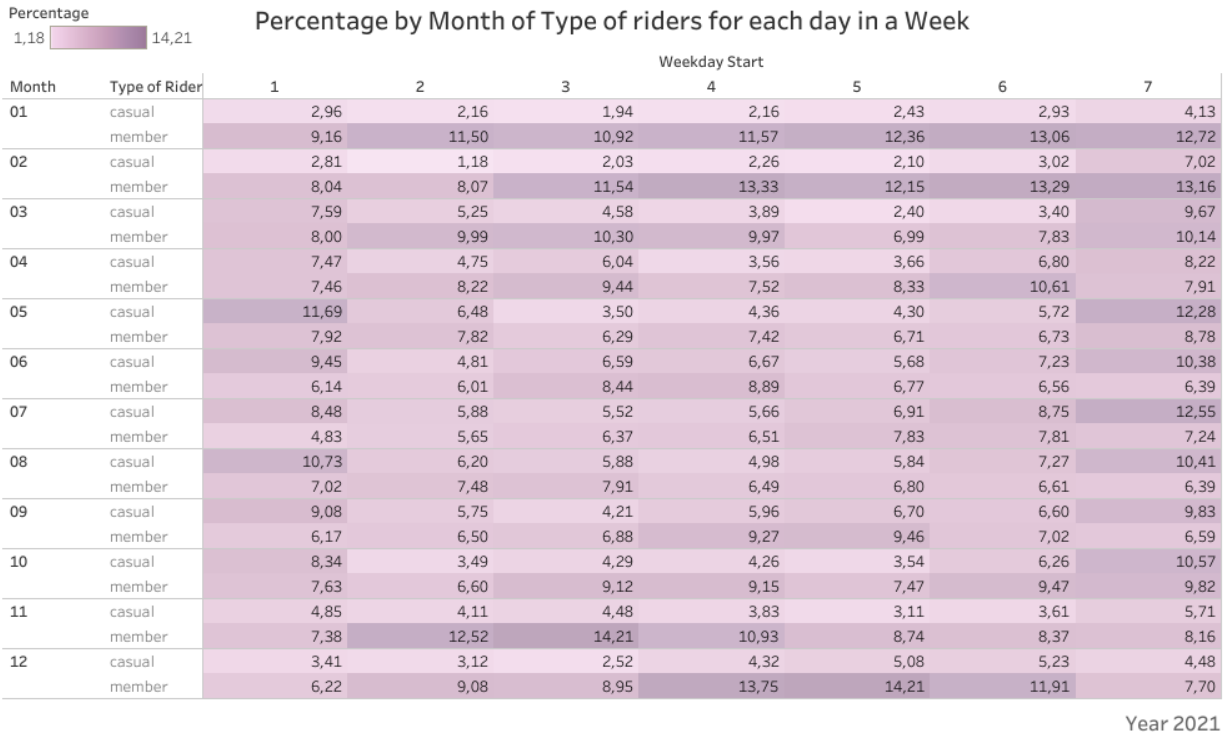


```sql
WITH T1 AS
(SELECT
COUNT(*) AS total
FROM bigDatasetCSV
)

SELECT
weekday_start,
member_casual,
COUNT(*) AS total_by_weekday,
ROUND((COUNT(*)*(100+0.0))/(total+0.0), 3) AS percentage
FROM bigDatasetCSV JOIN T1
GROUP BY weekday_start,member_casual

LIMIT 200
```



```sql
WITH T1 AS
(SELECT
strftime("%m",start_date) AS Month1,
COUNT(*) AS total_by_Month
FROM bigDatasetCSV
GROUP BY strftime("%m",start_date)
)

SELECT
strftime("%m",start_date) AS Month,
total_by_Month,
member_casual,
weekday_start,
COUNT(*) AS by_month_and_type,
ROUND((COUNT(*)*(100+0.0))/(total_by_Month+0.0), 3) AS percentage
FROM bigDatasetCSV JOIN T1 ON Month1 = Month
GROUP BY strftime("%m",start_date), member_casual, weekday_start

LIMIT 200
```




Percentage

1,18 ▨ 14,21

Percentage by Month of Type of riders for each day in a Week

Weekday Start

Month	Type of Rider	1	2	3	4	5	6	7
01	casual	2,96	2,16	1,94	2,16	2,43	2,93	4,13
	member	9,16	11,50	10,92	11,57	12,36	13,06	12,72
02	casual	2,81	1,18	2,03	2,26	2,10	3,02	7,02
	member	8,04	8,07	11,54	13,33	12,15	13,29	13,16
03	casual	7,59	5,25	4,58	3,89	2,40	3,40	9,67
	member	8,00	9,99	10,30	9,97	6,99	7,83	10,14
04	casual	7,47	4,75	6,04	3,56	3,66	6,80	8,22
	member	7,46	8,22	9,44	7,52	8,33	10,61	7,91
05	casual	11,69	6,48	3,50	4,36	4,30	5,72	12,28
	member	7,92	7,82	6,29	7,42	6,71	6,73	8,78
06	casual	9,45	4,81	6,59	6,67	5,68	7,23	10,38
	member	6,14	6,01	8,44	8,89	6,77	6,56	6,39
07	casual	8,48	5,88	5,52	5,66	6,91	8,75	12,55
	member	4,83	5,65	6,37	6,51	7,83	7,81	7,24
08	casual	10,73	6,20	5,88	4,98	5,84	7,27	10,41
	member	7,02	7,48	7,91	6,49	6,80	6,61	6,39
09	casual	9,08	5,75	4,21	5,96	6,70	6,60	9,83
	member	6,17	6,50	6,88	9,27	9,46	7,02	6,59
10	casual	8,34	3,49	4,29	4,26	3,54	6,26	10,57
	member	7,63	6,60	9,12	9,15	7,47	9,47	9,82
11	casual	4,85	4,11	4,48	3,83	3,11	3,61	5,71
	member	7,38	12,52	14,21	10,93	8,74	8,37	8,16
12	casual	3,41	3,12	2,52	4,32	5,08	5,23	4,48
	member	6,22	9,08	8,95	13,75	14,21	11,91	7,70

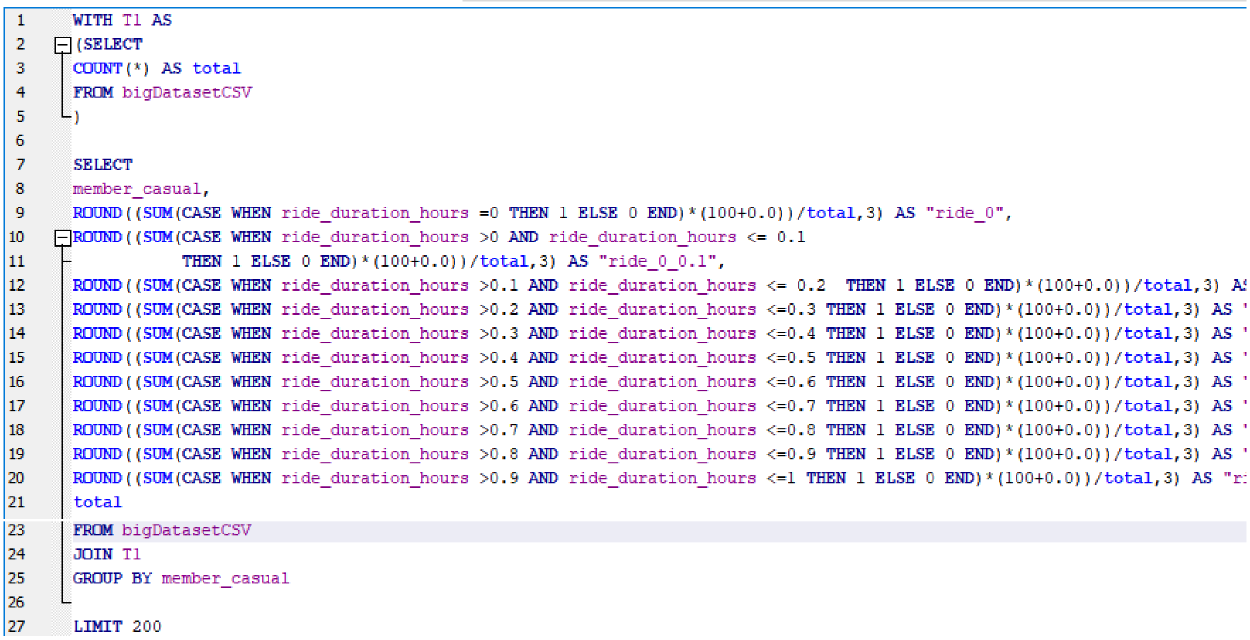
Year 2021

4. Analysis by duration of a ride

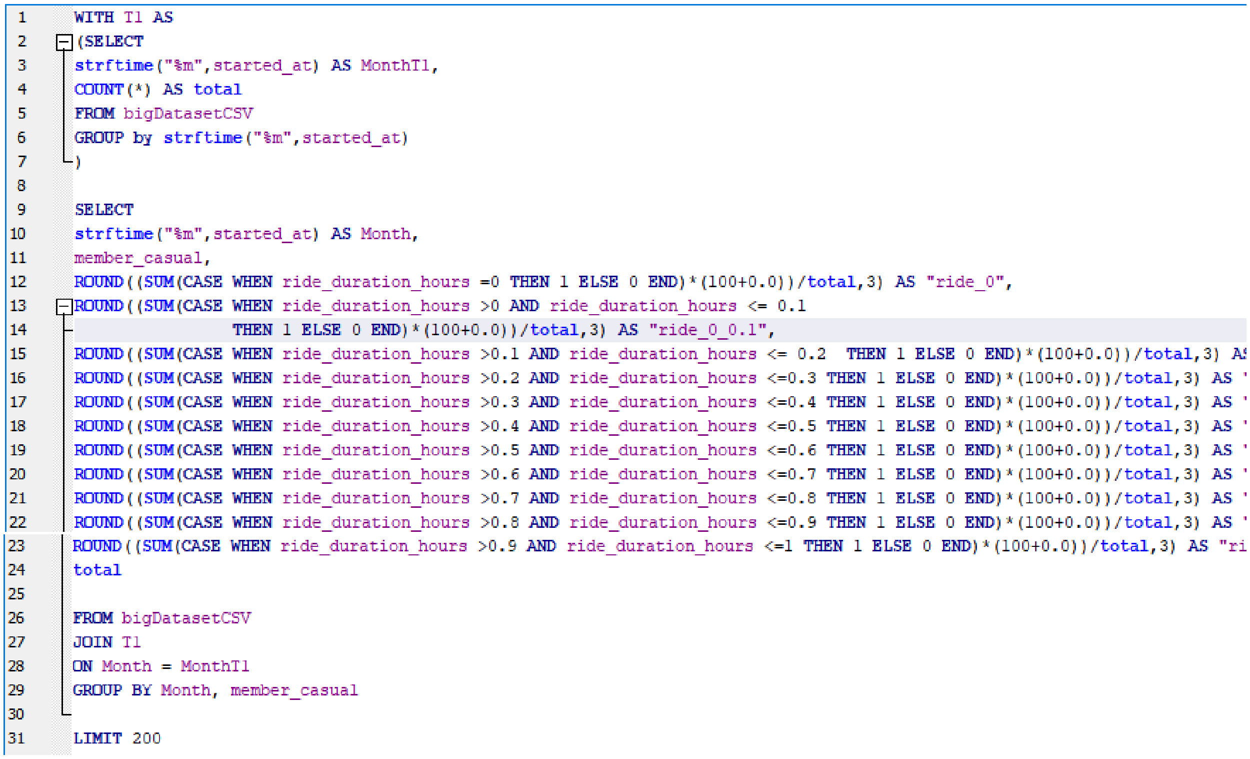
To have a better visualization visit Dashboard 4 in tableau-4.

As in the case for December 2021, annually we see that the biggest percentage of ride duration is accumulated for journeys less than or equal to 12 minutes, but considering length duration less or equal to 18 minutes is also a good target.

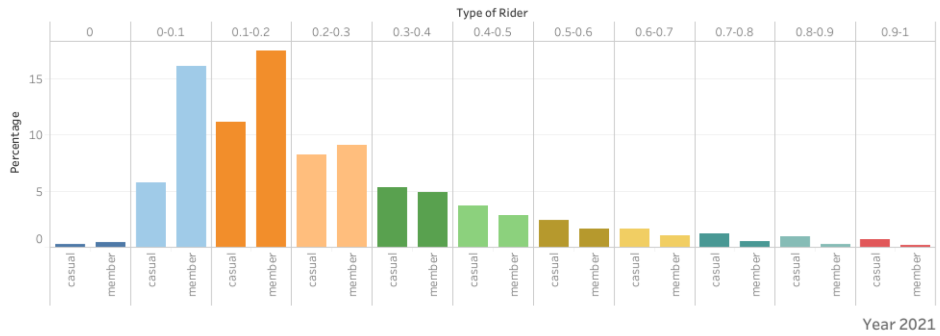
On the other hand, observing the data for each month, we see that above 12 minutes, usual riders are the kind of riders using mostly the service, though with small percentage. In other words, member riders prefer to use the service for less ride duration, while usual riders prefer using the service for longer rides.



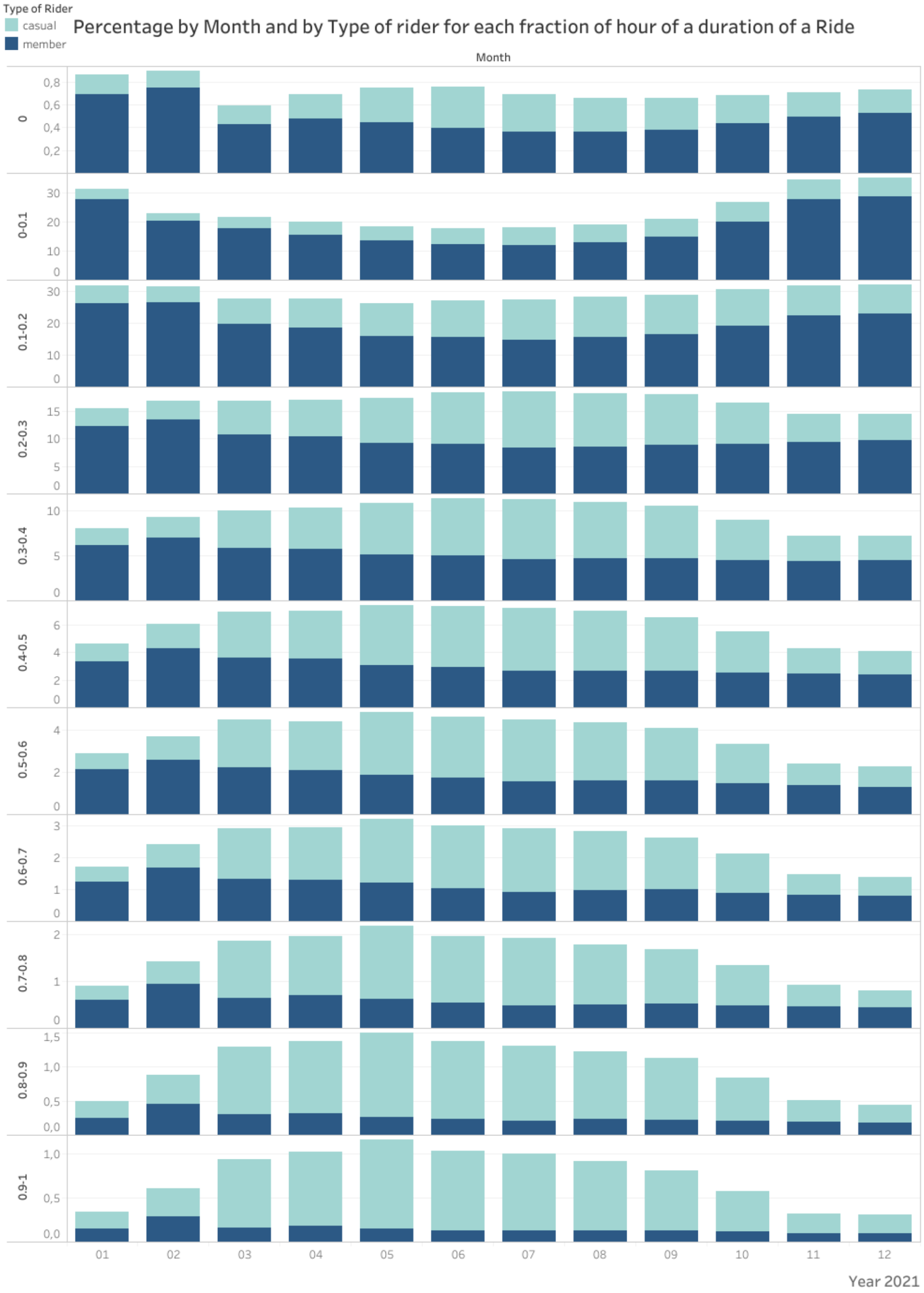
```sql
WITH T1 AS
(SELECT
COUNT(*) AS total
FROM bigDatasetCSV
)

SELECT
member_casual,
ROUND((SUM(CASE WHEN ride_duration_hours =0 THEN 1 ELSE 0 END)*(100+0.0))/total,3) AS "ride_0",
ROUND((SUM(CASE WHEN ride_duration_hours >0 AND ride_duration_hours <= 0.1
           THEN 1 ELSE 0 END)*(100+0.0))/total,3) AS "ride_0_0.1",
ROUND((SUM(CASE WHEN ride_duration_hours >0.1 AND ride_duration_hours <= 0.2  THEN 1 ELSE 0 END)*(100+0.0))/total,3) AS
ROUND((SUM(CASE WHEN ride_duration_hours >0.2 AND ride_duration_hours <=0.3 THEN 1 ELSE 0 END)*(100+0.0))/total,3) AS '
ROUND((SUM(CASE WHEN ride_duration_hours >0.3 AND ride_duration_hours <=0.4 THEN 1 ELSE 0 END)*(100+0.0))/total,3) AS '
ROUND((SUM(CASE WHEN ride_duration_hours >0.4 AND ride_duration_hours <=0.5 THEN 1 ELSE 0 END)*(100+0.0))/total,3) AS '
ROUND((SUM(CASE WHEN ride_duration_hours >0.5 AND ride_duration_hours <=0.6 THEN 1 ELSE 0 END)*(100+0.0))/total,3) AS '
ROUND((SUM(CASE WHEN ride_duration_hours >0.6 AND ride_duration_hours <=0.7 THEN 1 ELSE 0 END)*(100+0.0))/total,3) AS '
ROUND((SUM(CASE WHEN ride_duration_hours >0.7 AND ride_duration_hours <=0.8 THEN 1 ELSE 0 END)*(100+0.0))/total,3) AS '
ROUND((SUM(CASE WHEN ride_duration_hours >0.8 AND ride_duration_hours <=0.9 THEN 1 ELSE 0 END)*(100+0.0))/total,3) AS '
ROUND((SUM(CASE WHEN ride_duration_hours >0.9 AND ride_duration_hours <=1 THEN 1 ELSE 0 END)*(100+0.0))/total,3) AS "ri
total
FROM bigDatasetCSV
JOIN T1
GROUP BY member_casual

LIMIT 200
```



```sql
WITH T1 AS
(SELECT
strftime("%m",started_at) AS MonthT1,
COUNT(*) AS total
FROM bigDatasetCSV
GROUP by strftime("%m",started_at)
)

SELECT
strftime("%m",started_at) AS Month,
member_casual,
ROUND((SUM(CASE WHEN ride_duration_hours =0 THEN 1 ELSE 0 END)*(100+0.0))/total,3) AS "ride_0",
ROUND((SUM(CASE WHEN ride_duration_hours >0 AND ride_duration_hours <= 0.1
                THEN 1 ELSE 0 END)*(100+0.0))/total,3) AS "ride_0_0.1",
ROUND((SUM(CASE WHEN ride_duration_hours >0.1 AND ride_duration_hours <= 0.2  THEN 1 ELSE 0 END)*(100+0.0))/total,3) AS
ROUND((SUM(CASE WHEN ride_duration_hours >0.2 AND ride_duration_hours <=0.3 THEN 1 ELSE 0 END)*(100+0.0))/total,3) AS '
ROUND((SUM(CASE WHEN ride_duration_hours >0.3 AND ride_duration_hours <=0.4 THEN 1 ELSE 0 END)*(100+0.0))/total,3) AS '
ROUND((SUM(CASE WHEN ride_duration_hours >0.4 AND ride_duration_hours <=0.5 THEN 1 ELSE 0 END)*(100+0.0))/total,3) AS '
ROUND((SUM(CASE WHEN ride_duration_hours >0.5 AND ride_duration_hours <=0.6 THEN 1 ELSE 0 END)*(100+0.0))/total,3) AS '
ROUND((SUM(CASE WHEN ride_duration_hours >0.6 AND ride_duration_hours <=0.7 THEN 1 ELSE 0 END)*(100+0.0))/total,3) AS '
ROUND((SUM(CASE WHEN ride_duration_hours >0.7 AND ride_duration_hours <=0.8 THEN 1 ELSE 0 END)*(100+0.0))/total,3) AS '
ROUND((SUM(CASE WHEN ride_duration_hours >0.8 AND ride_duration_hours <=0.9 THEN 1 ELSE 0 END)*(100+0.0))/total,3) AS '
ROUND((SUM(CASE WHEN ride_duration_hours >0.9 AND ride_duration_hours <=1 THEN 1 ELSE 0 END)*(100+0.0))/total,3) AS "ri
total

FROM bigDatasetCSV
JOIN T1
ON Month = MonthT1
GROUP BY Month, member_casual

LIMIT 200
```



Annual Percentage by Type of Rider for each fraction of Hour of the duration of a Ride



Percentage by Month and by Type of rider for each fraction of hour of a duration of a Ride