|  |  |
| --- | --- |
|  | **Case Study:**  **How Does a Bike-Share Navigate Speedy Success?** |

1. **Defining the business task**

Delve into bike-share data and discover how do annual **members** and **casual** riders use Cyclistic bikes differently, which will help Stakeholders design digital media marketing strategies to convert casual riders.

1. **Description of the used data source**

Cyclistic’s historical trip data source used in this analysis project is a is **public data** containing the previous 12 months of Cyclistic trip data (starting from December 2020 up to November 2021) made available by **Motivate International Inc**. under this [license](https://www.divvybikes.com/data-license-agreement) (<https://www.divvybikes.com/data-license-agreement>) and reachable by the current [link](https://divvy-tripdata.s3.amazonaws.com/index.html).

The data is presented in 12 csv files, every file is registering all trip activities for a given month, covering values of 13 attributes (ride\_id, rideable\_type, started\_at, ended\_at, start\_station\_name, start\_station\_id, end\_station\_name, end\_station\_id, start\_lat, start\_lng, end\_lat, end\_lng, member\_casual) over the whole month.

Checking every file at once, we discovered that there are many records (rows) of wrong data, for instance observations containing the starting time (started\_at) later than ending trip time (ended\_at) which is not possible, or even containing trips that have durations less that one minute.

1. **Documentation of cleaning or manipulation of data**

To ensure the data’s integrity and to verify that it’s clean and ready to be use in my analysis, I started uploading the csv file to my BigQuery workspace but unfortunately I was not able to proceed with files that are more than 100Mb in Sandbox, so I used Excel locally to transform, filter and sort data and then try to isolate all invalid rows, getting some other computer resources problems with the huge size of some files I opted to use the R studio desktop and then went smoothly with my checks and controls, all my cleaning process is well documented on the Rmd file accompanying this project, and as an overview of the actions taken during my data processing task, here are a list of the key activities performed:

* Checked the validity and the authenticity of the rider\_id field (should be unique, not null with a length of 16 characters).
* Searching for NA and NULL values on the main attributes.
* Checking if the starting and ending latitude and longitude are belonging in Chicago in the data source.
* Examining the validity and values of ride starting and ending time, searching for wrong values like ending time before starting time, or trip duration time less than one minute.
* Creating a new filed called **ride\_length** (expressed in second) as a difference of ending time and staring time for any ride.
* Creating a new filed called **day\_of\_week** to associate each ride to the appropriate date of the week.
* Creating a new filed called **ride\_month** (mmm, yyyy) on the last step of the analysis when merging all twelve csv files.
* Creating a new filed called **ride\_path** (merging start\_station\_name 🡪 end\_station\_name) to find the more used round trip during these last 12 months.

1. **A summary of the analysis**

Importing the csv files on my Rstudio workspace, then installing and loading the required packages (tidyverse, lubridate and ggplot) to be able to discover the data, go deeper in the analysis process and find trends and relationships between the data.

The most recent 12 files are well formatted and contain same attributes which makes the merge operation easy to perform.

As a starting point, I tried to analyse each file at once, trying to discover insights related to only that period (month) it covers, and after finishing with all files I proceed to merge them all in one unique full data frame and advance with getting trends and insights at a higher scale.

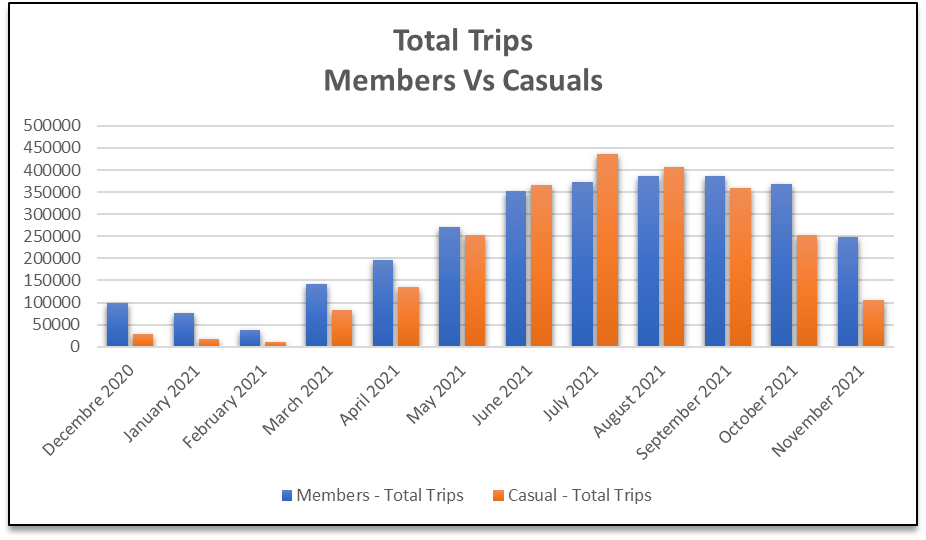
As a summary I get the following information.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Month Ref.** | **N. Rows** | **Valid Rows Number** | **Classic Bike Member (tot. Hours)** | **Classic Bike Casual** | **Docked bike Member** | **Docked Bike Casual** | **Electric Bike Member** | **Electric Bike Casual** | **Total Members Trip Duration** | **Total Casuals Trip Duration** | **Members - Total Trips** | **Casual - Total Trips** |
| 12/2020 | 131573 | 129497 | 13066,25 | 4959,99 | 1642,23 | 4745,93 | 6774,82 | 3715,47 | 21483,30 | 13421,39 | 99832 | 29665 |
| 01/2021 | 96834 | 95415 | 11924,65 | 3418,22 | 0,04 | 2449,16 | 4955,14 | 1886,25 | 16879,83 | 7753,63 | 77573 | 17842 |
| 02/2021 | 49622 | 48649 | 9525,85 | 3968,92 | 0,00 | 3438,90 | 2330,73 | 927,61 | 11856,58 | 8335,43 | 38685 | 9964 |
| 03/2021 | 228496 | 225523 | 25324,08 | 23929,09 | 0,00 | 21304,62 | 8297,79 | 8203,30 | 33621,86 | 53437,01 | 142375 | 83148 |
| 04/2021 | 337230 | 332422 | 35670,49 | 37417,29 | 0,00 | 34411,62 | 13427,15 | 14725,82 | 49097,65 | 86554,73 | 197477 | 134945 |
| 05/2021 | 531633 | 523243 | 46537,68 | 67547,98 | 0,00 | 62672,76 | 20458,13 | 33457,57 | 66995,81 | 163678,32 | 269897 | 253346 |
| 06/2021 | 729595 | 717699 | 61911,00 | 96611,64 | 0,00 | 85578,19 | 25849,10 | 47108,83 | 87760,10 | 229298,66 | 352676 | 365023 |
| 07/2021 | 822410 | 809760 | 63681,34 | 114077,20 | 0,00 | 76799,79 | 26541,86 | 50665,63 | 90223,20 | 241542,63 | 373833 | 435927 |
| 08/2021 | 804352 | 792943 | 64386,78 | 105238,57 | 0,00 | 44495,10 | 27623,00 | 48219,94 | 92009,77 | 197953,60 | 385426 | 407517 |
| 09/2021 | 756147 | 745185 | 61485,58 | 87377,64 | 0,00 | 37228,48 | 28264,34 | 44055,40 | 89749,92 | 168661,52 | 385959 | 359226 |
| 10/2021 | 631226 | 621289 | 45972,32 | 47966,68 | 0,00 | 37506,31 | 31908,85 | 37439,46 | 77881,16 | 122912,46 | 367460 | 253829 |
| 11/2021 | 359978 | 354072 | 25221,31 | 13781,03 | 0,00 | 10759,92 | 22440,72 | 16662,66 | 47662,03 | 41203,60 | 248699 | 105373 |
|  | **5479096** | **5395697** | **464707,32** | **606294,24** | **1642,27** | **421390,80** | **218871,61** | **307067,93** | **685221,21** | **1334752,97** | **2939892** | **2455805** |

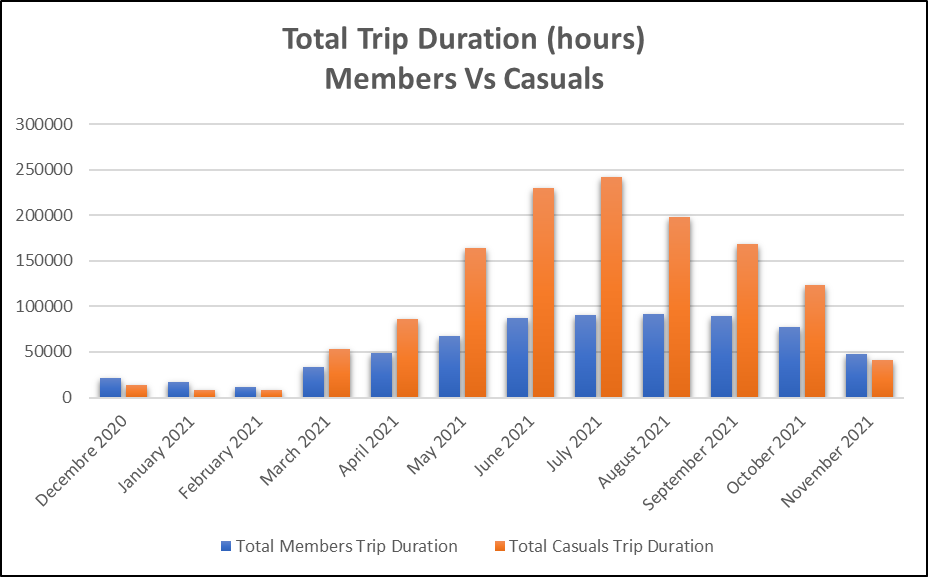
1. **Supporting visualizations and key findings**

After a quick study of the whole dataframe constituted of 12 csv file, I was able to summarize my finding in just three significant visuals that can give a big picture of what was happening in our Cyclistic company over the last 12 months, digging deeper on the data is possible to try uncovering other interesting insights but to be brief and not annoying for the audience members (stakeholders), I choose only these few which are enough to aid the audience taking a right and accurate business decisions.

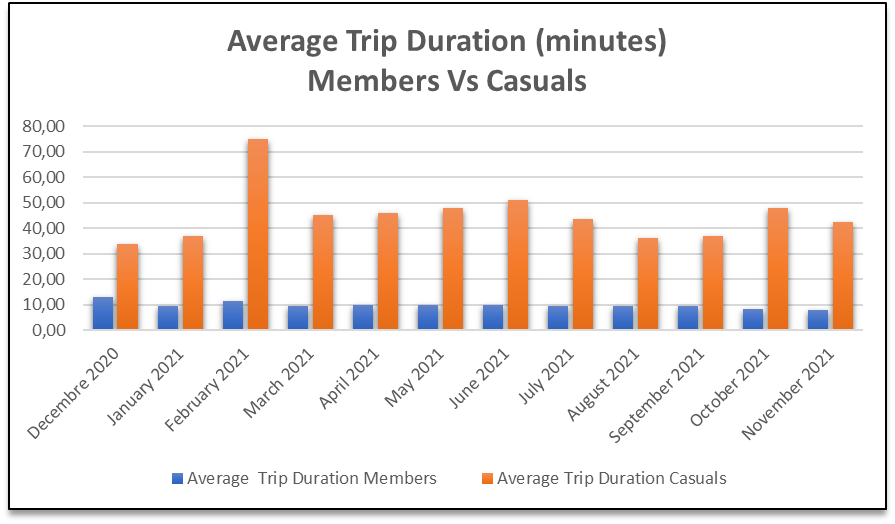
* **Sum of rides per month – Members Vs Casuals**

****

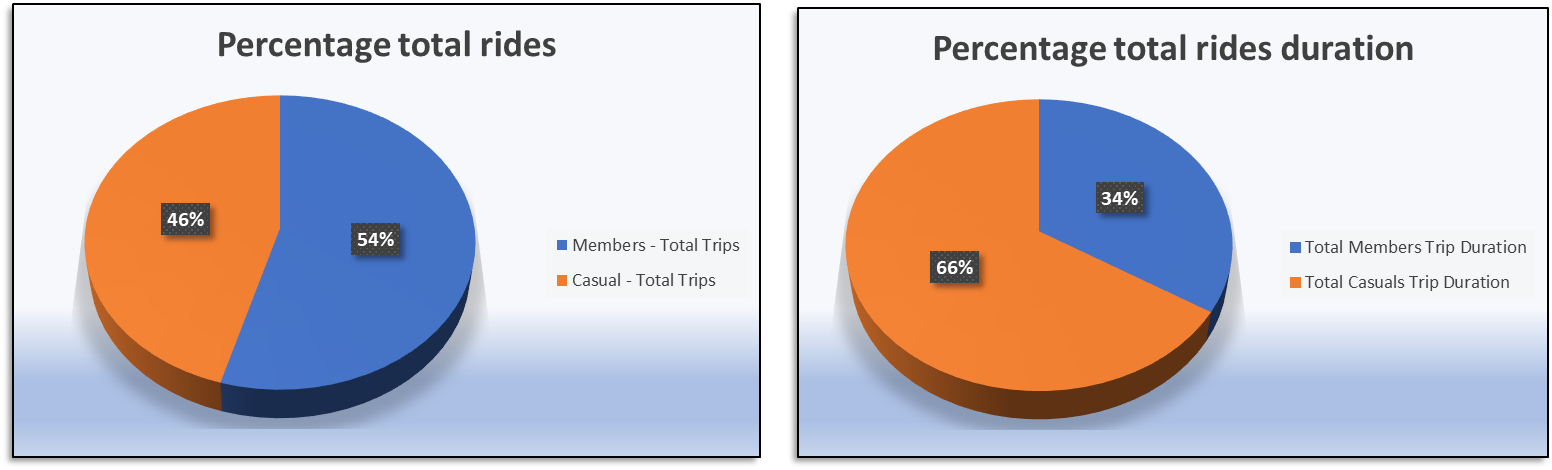
* **Sum of rides duration per month – Members Vs Casuals**

****

* **Average ride duration per month – Members Vs Casuals**

****

As noticeable on the graphs, the members riders are more actives on almost the whole year except the summer months (**June, July and August**) even though, they are using the bike service less that casuals, only for few minutes a day, it results indeed that casuals with **46%** of rides number are covering about **66%** of the total annual rides duration as presented below.



1. **Recommendations based on the accomplished analysis**

As a finale assumption of all what was analysed and explored during this case study, and to take the company activities a step further, I suggest:

* Boosting both members and casuals offers, since each category is acting differently and has a good impact on the company ongoing activity.
* Promote the bike sharing service during December, January, and February, months in which it has been registered the minimal bike operating rate, possibly due to weather conditions, so having an “All Weather Cycling” slogan can help influencing members and not to take advantage of the service.
* Encourage members to expand the usage of the bike service not only to get to work, starting from the underground station, evidencing that cycling activities are healthy, secure and it’s even a good protection against Covid contrarily from using the crowded public transportation.