How Does a Bike-Share Navigate Speedy Success

Berhane Tekle

5/23/2021

INTRODUCTION

In 2016, Cyclistic launched a successful bike-share offering. Since then, the program has grown to a fleet of 5,824 bicycles that are geo-tracked and locked into a network of 692 stations across Chicago. The bikes can be unlocked from one station and returned to any other station in the system anytime. Until now, Cyclistic's marketing strategy has relied on building general awareness and appealing to broad consumer segments. One approach that helped make these things possible was the flexibility of its pricing plans: single-ride passes, full-day passes, and annual memberships. Customers who purchase single-ride or full-day passes are referred to as ***casual riders***. Customers who purchase annual memberships are Cyclistic ***members***.

Cyclistic's finance analysts have concluded that annual members are much more profitable than casual riders. Although the pricing flexibility helps Cyclistic attract more customers, the director of marketing believes that ***maximizing the number of annual members will be key to future growth***. Rather than creating a marketing campaign that targets all-new customers, she believes there is a particularly good chance to ***convert casual riders into members***. She notes that casual riders are already aware of the Cyclistic program and have chosen Cyclistic for their mobility needs. The Director of Marketing has set a clear goal: Design marketing strategies aimed at converting casual riders into annual members. To do that, the marketing analyst team needs to better understand:

I. How do annual members and casual riders differ?
II. Why casual riders would buy a membership? and
III. How digital media could affect their marketing tactics?

1. BUSINESS TASK STATEMENT

As a marketing analyst in the marketing department headed by the Director of Marketing, I am assigned to analyzing Cyclistic historical bike trip data to identify trends to address the first question: How do annual members and casual riders use Cyclistic bikes differently?

The following steps are followed to address this task:

1. A clear statement of the business task as detailed above,
2. Description and preparation of all data sources used,
3. Detailed log of cleaning and manipulation of data performed, and tools used to transform the data,
4. A summary of analysis,
5. Supporting visualizations and key findings, and
6. Recommendations based on the analysis.

2. PREPARATION

This case study is based on Cyclistic's historical trip data from May 2020 to April 2021. The data has been made available by Motivate International Inc. here under this license. It is a public data that can be used to explore how different customer types are using Cyclistic bikes.

Due to data-privacy issues, personally identifiable information such as customers' addresses and credit card payment information are not included in this dataset. Besides, the analysis is limited to the available information of rides made with starting and ending docking stations information, time elapsed to complete a trip, and customer type. Hence, no revenue and cost analysis can be made from this dataset.

I did not find any issues with regards to the data's credibility and integrity. The data has been sorted and filtered to check for inaccuracies and anomalies such as looking for customer types that are neither casual nor regular, and starting dates that are earlier than ending dates, and any other missing information that will affect the analysis of the dataset. Except for relatively few entries with their ending dates "occurring" on or earlier than their starting dates, the data is generally consistent, reliable and UpToDate to give an adequate picture on the existence (or nonexistence) of clear pattern in bike usage differences between members and casual customers.

The data is downloaded in 12 consecutive zip files (one for each month from May 2020 to April 2021) and saved in a local folder with copies of each extracted from and saved in Excel for reformatting and transformation in a separate folder.

The information contained in the dataset, with the addition of calculated columns and aggregating the results therein has enabled me to discover a clear distinction in usage and purchases behavior between casual riders and subscribed members.

3. PROCESSING

Excel is used to format, cleanse, and transform the data. For faster processing, scalability, and data security, transformed data is consolidated and stored in SQL Server. R is used for reporting and visualization, with reproducibility and sharing of the data in mind.

For the purposes of this analysis the following formatting, cleansing and transformations have been performed on each Excel copy of the 12 files from May 2020 to April 2021:

I. Columns Removed: The following columns related to GPS coordinates of starting and ending rides are removed for they serve no purposes for our analysis: **[start_lat]**, **[start_lng]**, **[end_lat]**, and **[end_lng]**.

II. Calculated Columns Added:

1.**[ride_length]** to calculate hours elapsed between **[started_at]** and **[ended_at]** and reformatted as numeric data type. E.g., 1.50 is equivalent to 90 minutes.

2.**[Period]** to calculate month and year for the [started_at] ride date in "mmmm, yyyy" format.

3.**[day_of_week]** to calculate the day of the week for the [started_at] ride date to get an insight on day-related riding habits of customers, if required.

4.**[season]** to calculate season for the [started_at] ride date to get an insight on the effect of season on customers' riding habits and preferences, if required.

III. Rows Removed:

The following records are removed from the dataset:Records where **[started_at]** occurred on or later than **[ended_at]**.

NOTE: Out of a total of 3,712,549 number of records, a total 10,096 number of such entries have been removed and saved in a separate sub-folder. They are about 0.27% of the total number of entries and are too immaterial for their removal to alter the outcome of this study in any meaningful way.

4. ANALYSIS

The following steps have been taken to organize and get the data ready for analysis:

I. Creating a database in SQL Server: A new database ***CyclisticRides*** is created in SQL Server 2019 Developer Edition. All the 12 cleansed and transformed excel trip data files from May 2020 to April 2021 have been loaded to the database with *Data Source* as Excel 2016 and *Destination* as Microsoft OLE DB PROVIDER FOR SQL SERVER.

II. A new table **Trip_data** is built by merging the cleansed twelve tables using the following SQL query.

```
SELECT * INTO Trip_Data
FROM (
      SELECT * FROM TripData_202005
      UNION
      SELECT * FROM TripData_202006
      UNION
      SELECT * FROM TripData_202007
      UNION
      SELECT * FROM TripData_202008
      UNION
      SELECT * FROM TripData_202009
      UNION
      SELECT * FROM TripData_202010
      UNION
      SELECT * FROM TripData_202011
      UNION
      SELECT * FROM TripData_202012
      UNION
      SELECT * FROM TripData_202101
      UNION
      SELECT * FROM TripData_202102
      UNION
      SELECT * FROM TripData_202103
      UNION
      SELECT * FROM TripData_202104) AS tripdata
```

III. Two summary tables are created out of the consolidated trip_data. They are:

-**[Trips_Summary]** to show a summary of trips taken, hours ridden, and average hours/trip for each month from May 2020 to April 2021 by each group of customers, and

-[Grand_Summary] to show a summary of a sum of trips taken, hours ridden, and average hours/trip for the year from May 1, 2020 to April 30, 2021 by each group of customers.

The following queries are used to build each respective table:

Trips Summary:

```
SELECT * INTO Trips_Summary
FROM

(SELECT CAST(YEAR(started_at) AS varchar)+'-'+ RIGHT('0' + RTRIM(M
ONTH(started_at)), 2)Ride_Period
      ,[Period]
      ,[Customer_Type]
      ,COUNT(ride_length) AS NumberOfTrips
      ,SUM(ride_length) AS HoursRidden
      ,CAST(AVG(ride_length) AS Decimal(38,2)) AS AverageHoursPerT
rip

  FROM [CyclisticRides].[dbo].[Trip_Data]
  GROUP BY CAST(YEAR(started_at) AS varchar)+'-'+ RIGHT('0' + RTRI
M(MONTH(started_at)), 2),Period,customer_type
  ) AS TS;
```

Grand Summary:

```
SELECT * INTO Grand_Summary
FROM
(
SELECT customer_type,
       COUNT(ride_length) as NumberOfTrips,
       SUM(ride_length) as HoursRidden,
       CAST(AVG(ride_length) AS decimal(38,2))as AverageHoursPerTrip
from Trip_Data
GROUP BY customer_type
) AS GS;
```

IV. Connecting SQL Server database CyclisticRides to R for visualization and sharing findings: A new live connection using a DSN (named **R_Connections**) is created between SQL Server and R using ODBC to load the tables created in SQL Server into R Studio. The following packages are installed and loaded to accomplish this task:

```
library(odbc)
library(DBI)
library(devtools)

data_connect <- dbConnect(odbc::odbc(), "R_Connections")
```

V. Accessing and sorting the data in RStudio for analysis and visualization:

```
Trip_Summary<-dbGetQuery(data_connect,'Select Ride_Period, Period
AS Ride_Period_Description, Rider_Type, NumberOfTrips, HoursRidden
, AverageHoursPerTrip FROM Trips_Summary Order by Ride_Period')
Grand_Summary<-dbGetQuery(data_connect,'Select * from Grand_Summar
y ORDER BY Rider_Type DESC')
```

Following is remaining pre-installed packages loaded to create our reports and visualizations:

```
library(tidyverse)
library(readr)
library(tidyr)
library(dplyr)
library(ggplot2)
library(scales)
library(knitr)
library(flexdashboard)
library(gridExtra)
```

5. FINDINGS

As shown in the **analysis** step, we have completed the necessary tasks to organize, access and summarize the data to see how and to what extent the two groups of customers (subscribed members and casual riders) differ in their bike usage in terms of frequency and ride length in hours.

Following is the resultset of the **[Grand Summary]** table and the **[Trips Summary]** table accompanied by their respective visualizations that will answer our business question: **How do annual members and casual riders use Cyclistic bikes differently?**

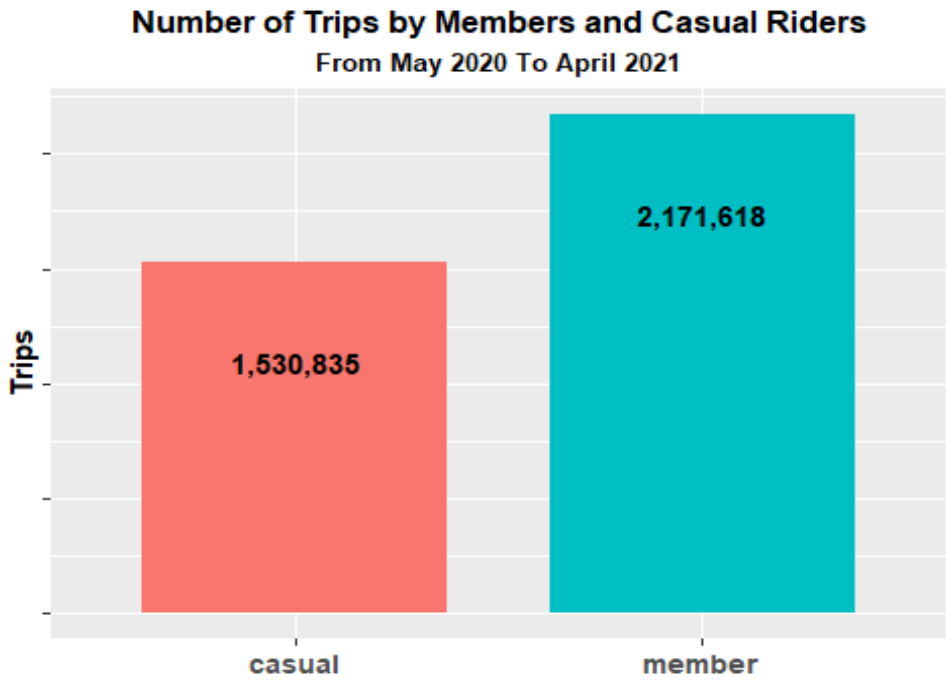
Let's start with yearly totals before we examine monthly figures. The following code will fetch the Grand_Summary table and load it to our report:

```
knitr::kable(Grand_Summary,digits = 2, format.args = list(big.mark = ",",
  scientific = FALSE), caption = "Fig.1-GRAND SUMMARY:Total Number of Trips Taken and Hours Ridden with Average Hours/Trip")
```

Fig.1-GRAND SUMMARY:Total Number of Trips Taken and Hours Ridden with Average Hours/Trip

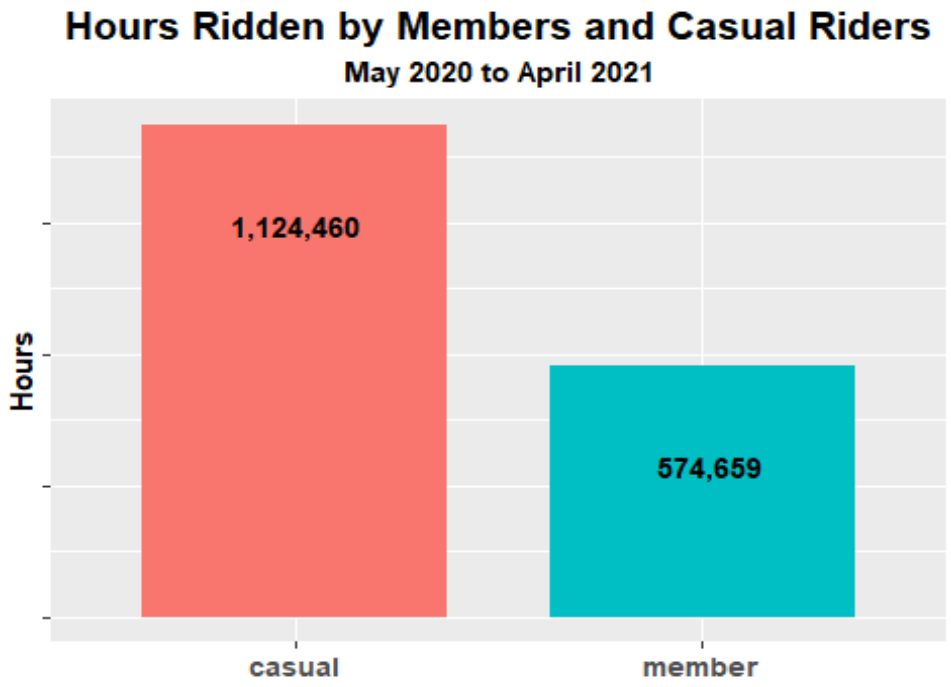
Rider_Type	NumberOfTrips	HoursRidden	AverageHoursPerTrip
member	2,171,618	574,659	0.26
casual	1,530,835	1,124,460	0.73

Fig.1.1



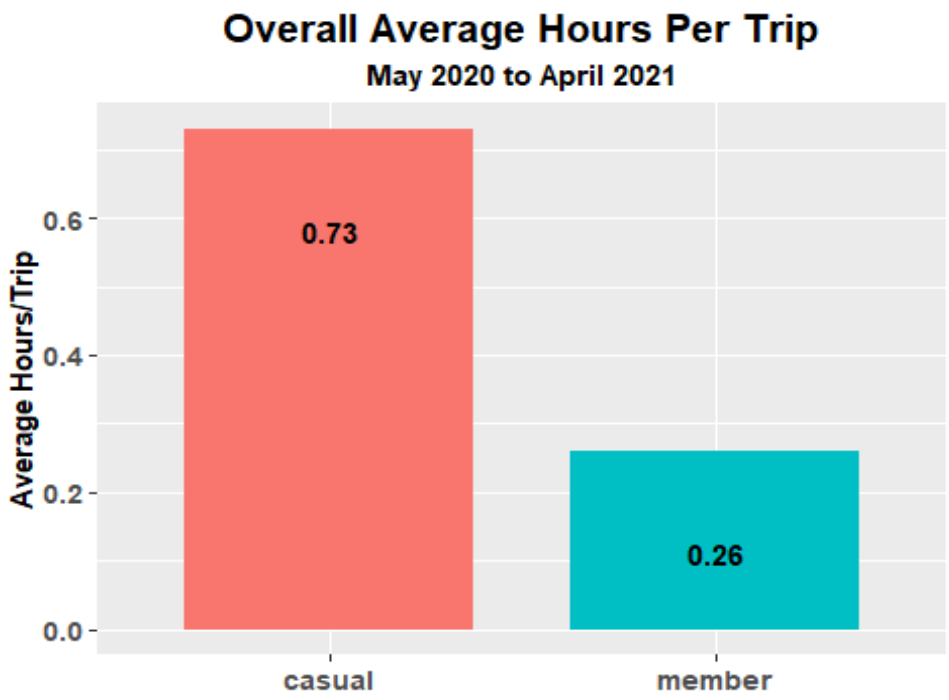
Members bike more frequently than casual riders.

Fig.1.2



While Members bike more frequently than casual riders, casual riders usually take longer trips.

Fig.1.3



On average casual riders ride (or keep the bikes) for approximately three times longer than members.

Now let's examine the monthly figures that make up the totals we analyzed previously. The following code will fetch and load the monthly Trip_Summary into our report:

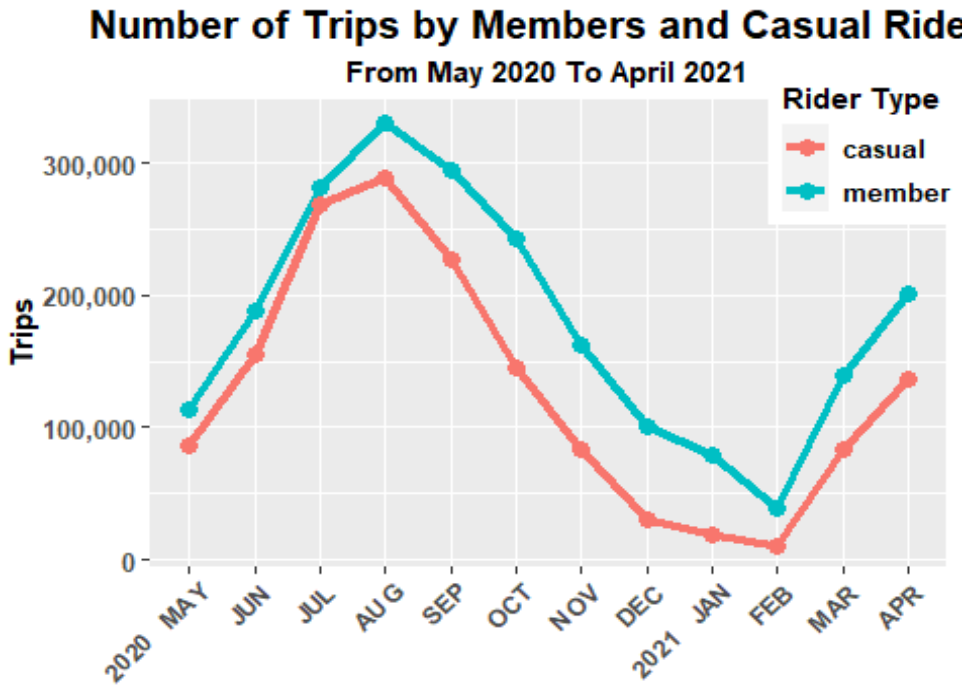
```
knitr::kable(Trip_Summary, caption = "Fig.2-TRIP SUMMARY:Monthly Trips T
aken and Hours Ridden From May 1, 2020 to April 30, 2021", digits = 2, f
ormat.args = list(big.mark = ",",
  scientific = FALSE))
```

Fig.2-TRIP SUMMARY: Monthly Trips Taken and Hours Ridden From May 1, 2020 to April 30, 2021

Ride_Period	Ride_Period_Description	Rider_Type	Number OfTrips	HoursRidden	AverageHours PerTrip
2020-05	May, 2020	casual	86,844	74,144.47	0.85
2020-05	May, 2020	member	113,258	37,339.02	0.33
2020-06	June, 2020	casual	154,695	133,086.73	0.86
2020-06	June, 2020	member	188,251	58,694.68	0.31
2020-07	July, 2020	casual	268,688	268,496.93	1.00
2020-07	July, 2020	member	281,047	83,253.62	0.30
2020-08	August, 2020	casual	288,638	216,187.28	0.75
2020-08	August, 2020	member	330,952	92,893.12	0.28

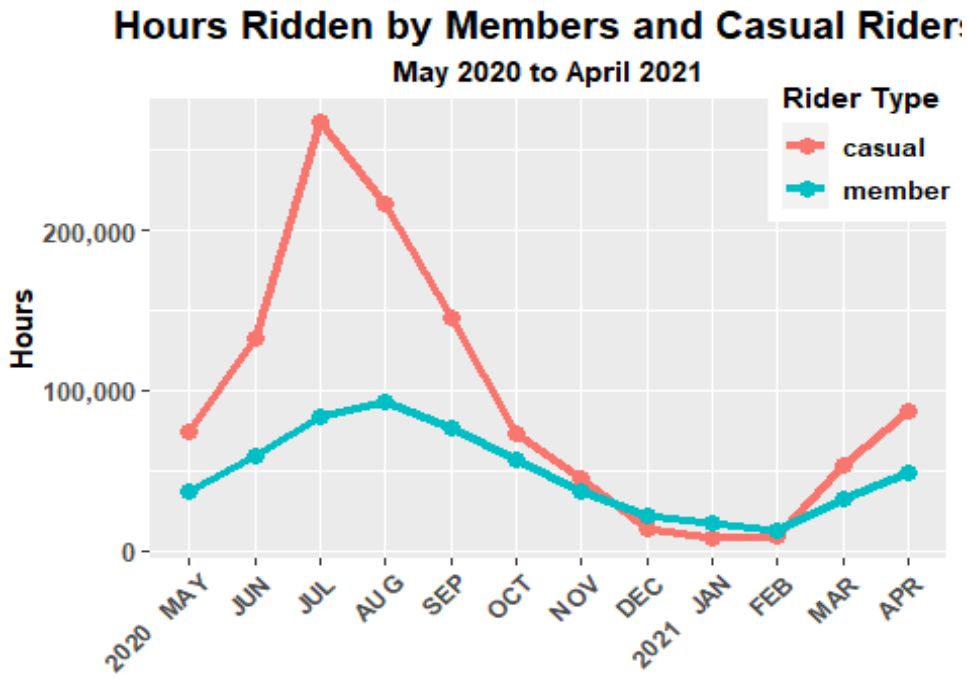
2020-09	September, 2020	casual	226,476	145,221.15	0.64
2020-09	September, 2020	member	294,694	76,582.25	0.26
2020-10	October, 2020	casual	144,529	72,896.71	0.50
2020-10	October, 2020	member	242,213	56,748.86	0.23
2020-11	November, 2020	casual	83,413	45,291.81	0.54
2020-11	November, 2020	member	162,264	37,195.40	0.23
2020-12	December, 2020	casual	29,997	13,427.22	0.45
2020-12	December, 2020	member	101,029	21,488.53	0.21
2021-01	January, 2021	casual	18,117	7,757.86	0.43
2021-01	January, 2021	member	78,715	16,896.16	0.21
2021-02	February, 2021	casual	10,131	8,338.08	0.82
2021-02	February, 2021	member	39,491	11,866.45	0.30
2021-03	March, 2021	member	139,080	32,558.79	0.23
2021-03	March, 2021	casual	82,706	53,027.77	0.64
2021-04	April, 2021	casual	136,601	86,583.85	0.63
2021-04	April, 2021	member	200,624	49,142.17	0.24

Fig.2.1



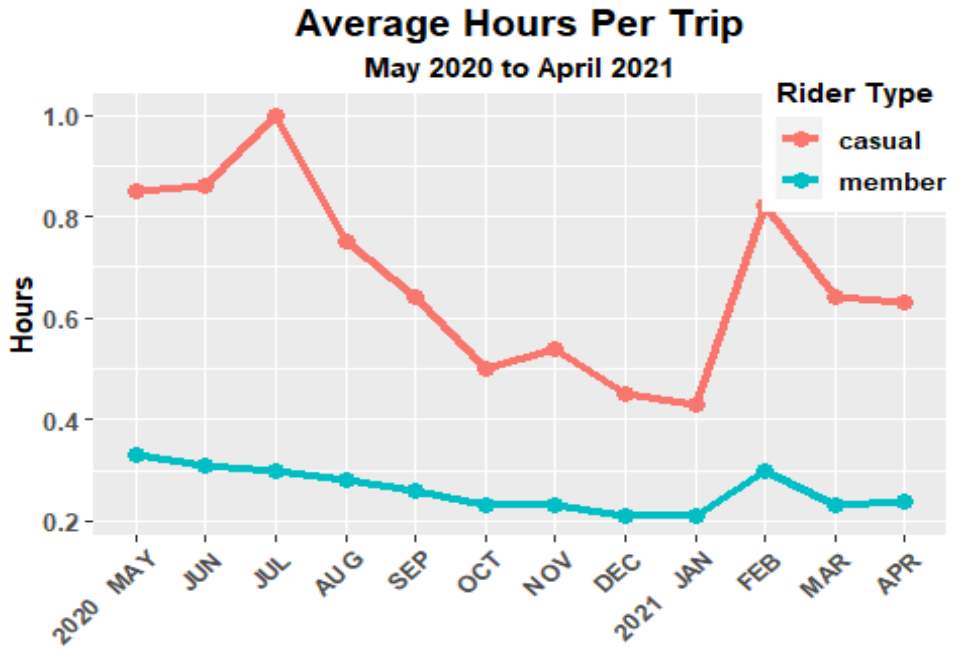
While Members biked more frequently than Casual Riders throughout the year, both groups of riders exhibit similar usage behavior on a month-to-month basis, which is generally affected by season.

Fig.2.2



In terms of ride length, casual riders generally bike (or keep the bikes) for far longer than Members. On the other hand, Members tend to exhibit relatively more consistent usage behavior, while usage of casual riders is highly affected by season.

Fig.2.3



On average Casual Riders rode (or kept the bikes) for far longer than Members throughout the year.

6. CONCLUSION

As our visualizations clearly demonstrated, there is a clear pattern in bike usage differences between Members and Casual Riders.

Members took 640,730 more trips than Casual Riders from May 1, 2020 to April 30, 2021. ***This tells us that Members tend to ride more regularly and frequently than Casual Riders.***

In terms of ride length, Casual Riders rode more than twice as Members in total hours in the same period.

Relative to number of trips taken, Casual Riders rode ***almost three times longer*** than Members on any given trip.On average, Members rode no more than 16 minutes (0.26 hrs) per trip; which essentially makes them SINGLE-RIDE users, as opposed to casual riders who tended to use less often but rode about 44 minutes (0.73 hrs) per trip.

When we look on a month-to-month basis, the same usage behavior is observed:

-In terms of number of trips, Members took more trips than Casual Riders EVERY month.

-In terms of total monthly hours ridden, Casual Riders rode more hours on all months except DEC, JAN, and FEB.

-While both groups' frequency and length of usage is similarly affected by season, Casual Riders' length of usage is highly seasonal compared to that of Members', who tend to exhibit more regularity regardless of season.

From this observation, we can conclude that as riders become members, they tend to be more regular and frequent users but time-sensitive in their biking usage.

As pointed out earlier, the financial analysts have already found out that Annual Members are much more profitable than Casual Riders. And the Marketing Manager believes converting Casual Riders into Members will be key to future growth. In light of our findings, moving Casual Riders to Annual Membership would result in generating larger revenue owing to increase in annual membership fee collections while at the same time ***potentially*** sacrificing earnings that comes from extra charges for usage that exceeds the single-ride time limit, much of which happens to be coming from ***casual riding***.

If we could get information on the existing number of Members and Casual Riders that make up the dataset, in addition to operating costs directly related to Membership and Casual Riding per period or rider (if any), further cost-benefit analysis could be made with regards to shifting the existing casual riders to member riders.