

The **Cyclistic Case Study** is one of the three case studies provided at the end of Google Data Analytics Certification. I have recently completed all the 7 courses in this program which included all the necessary steps for Data Analysis:

- Ask
- Prepare
- Process
- Analyze
- Share
- Act

## The Background Scenario

Cyclistic is an imaginary bike-sharing company that operates in **Chicago, Illinois**. I roleplayed as a data analyst working in the company's marketing analyst team.

Cyclistic has two types of customers: **annual members** and **casual customers**. The director of marketing believes that the company's future success depends on maximizing the number of annual members.

## The Questions

- How do annual members and casual riders use Cyclistic bikes differently?
- Why would casual riders buy Cyclistic annual memberships?
- How can Cyclistic use digital media to influence casual riders to become members?

At first, I felt tense about how I could answer all these questions, considering that I had just completed the certification and my knowledge was limited. However, I decided to invest some time in reading the objectives and approached the task by starting with some basic questions in my mind. As I progressed, I began to understand that I had all the necessary knowledge to find the answers. I realised that I just needed to connect the little pieces I knew in order to create a method to answer the questions.

## The Data

The Data I'm working on is of **First Party Type**. The data has been made available by Motivate International Inc. under this Licence (<https://ride.divvybikes.com/data-license-agreement>)  
The data integrity was checked and deemed unbiased.

## Here is the source of data

<https://divvy-tripdata.s3.amazonaws.com/index.html>

This includes data from 2014 but in order to completely make my data ROCCC i.e Reliable, original, comprehensive, **current** and cited, I will use data of 2022 from january to december, so i will work in a total of 12 csv files.

# Data cleaning and manipulation

## Google Sheets: initial data cleaning and manipulation

I had to make sure that the data is stored properly and prepared for analysis. After downloading all 12 zip files and unzipping them, I housed the files in a temporary folder on my desktop. I also created copy of each file as it's always a good practice. I opened Google sheets and started importing files one by one, i did the following.

- First text wrapped and aligned all the fields so that it was easier to read and understand.
- Changed format of started\_at and ended\_at columns to DATETIME.
- Created a column called ride\_length by subtracting the column started\_at from ended\_at.
- Formatted ride\_length as time.
- Created a column called day\_of\_week that each ride started using the WEEKDAY function (=WEEKDAY(C2,1)), NOTE: 1=sunday and 7=saturday

Since these datasets are so large, it is getting harder to work in spreadsheets. It is now time to shift to a tool that can handle larger datasets. I chose to use SQL via BigQuery.

Ps: i spent at least 2 hr looking for alternatives as some files were too large(were around 120-160 Mb) to import to google sheet. Was getting frustrated but stack overflow saved me, came to know about text splitters. I had to split my files in 4 parts using TEXT splitter.

## BigQuery: further data cleaning and manipulation via SQL

Since these datasets were so large, some of them had 2 lakh rows, I decided to shift to SQL as that tool is better suited for handling large datasets.

In order to continue processing the data in BigQuery, I created a bucket in Google Cloud Storage to upload all 12 files. I then created a project in BigQuery and uploaded these files as datasets.

Here's the code i used to combine 3 months in a quarter:

```
CREATE TABLE cyclistic.QUARTER1 AS
SELECT* FROM `my-project-1-373317.cyclistic.JAN`
UNION ALL
SELECT * FROM `my-project-1-373317.cyclistic.FEB`
UNION ALL
SELECT * FROM `my-project-1-373317.cyclistic.MAR`
```

Applied the same code to categorise every month in a quarter so i can perform my analysis based on different business quarters and seasons.

Ps: Ran into an error while joining tables, it was a mismatched column error, turns out i forgot to add new columns (ride\_length and day\_of\_week) to the 4th file.

Now remember how i converter sunday to 1, time to convert all numbers to their respective weekday.

Again there's a little problem, in order to use UPDATE function in BigQuery, i have to update to a billing account. I can change numbers to respective week names in spreadsheet too using "find and replace function" but then i have to edit all 12 files and then import them again to SQL. So let's leave that right now and continue with our analysis.

## 2022 Quarter1 Exploratory Analysis

I selected some basic columns first to get a preview of the data to get the basic idea of the potential trends:

```
SELECT ride_id,
        started_at,
        ended_at,
        ride_length,
        day_of_week,
        start_station_name,
        end_station_name,
        member_casual
```

```
FROM `my-project-1-373317.cyclistic.QUARTER1`  
ORDER BY ride_id DESC
```

The above query returned 503421 records. That's the amount of recorded data we have for Quarter1

## Total Trips

We'll create total columns for overall, annual members and casual riders. We'll also calculate percentages of overall total for both types.

```
SELECT  
    TotalTrips,  
    TotalMemberTrips,  
    TotalCasualTrips,  
    ROUND(TotalMemberTrips/TotalTrips,2)*100 AS MemberPercentage,  
    ROUND(TotalCasualTrips/TotalTrips,2)*100 AS CasualPercentage  
FROM  
    (  
        SELECT  
            COUNT(ride_id) AS TotalTrips,  
            COUNTIF(member_casual = 'member') AS TotalMemberTrips,  
            COUNTIF(member_casual = 'casual') AS TotalCasualTrips,  
        FROM  
            `my-project-1-373317.cyclistic.QUARTER1`  
    )
```

The Query showed:

Out of the 503421 trips, 74% was from annual members and 26% from casual riders.

## Average Ride Length

Now let's see how the average ride length differs for both the groups

```
SELECT(  
    SELECT AVG(TIME_DIFF(ride_length, TIME '00:00:00', SECOND))  
    FROM `my-project-1-373317.cyclistic.QUARTER1`  
  
    ) AS AvgRideLength_Overall,  
(SELECT AVG(TIME_DIFF(ride_length, TIME '00:00:00',  
SECOND))  
FROM `my-project-1-373317.cyclistic.QUARTER1`  
WHERE member_casual='member')
```

```

) AS AvgRideLength_Member,
(
    SELECT AVG(TIME_DIFF(ride_length, TIME '00:00:00',
SECOND))
FROM `my-project-1-373317.cyclistic.QUARTER1`
WHERE member_casual='casual'
) AS AvgRideLength_Casual

```

The screenshot shows the Google Cloud BigQuery interface. The query editor displays the following SQL code:

```

6 (SELECT AVG(TIME_DIFF(ride_length, TIME '00:00:00',
7 SECOND))
8 FROM `my-project-1-373317.cyclistic.QUARTER1`
9 WHERE member_casual='member'
10 ) AS AvgRideLength_Member,
11 (
12 SELECT AVG(TIME_DIFF(ride_length, TIME '00:00:00',
13 SECOND))
14 FROM `my-project-1-373317.cyclistic.QUARTER1`
15 WHERE member_casual='casual'
16 ) AS AvgRideLength_Casual
17

```

The query results are displayed in a table with the following data:

Row	AvgRideLength_Member	AvgRideLength_Casual
1	871.6678366615...	1391.744295860...

As we can see a casual member rides 700 seconds i.e approx 11.68 minutes more than a annual member

A basic question needs to be answered here, **why does a casual rider is riding more ?**  
The first thing we need to check are the **outliers**.What influence are outliers having on these averages? Let's investigate.

```

SELECT
    member_casual,
    MAX(ride_length) AS ride_length_MAX
FROM `my-project-1-373317.cyclistic.QUARTER1`
GROUP BY
    member_casual
ORDER BY
    ride_length_MAX DESC
LIMIT 10

```

The above code was to find the max ride\_length for casual and annual members and both were approximately 24 hours. So one thing we know is that outliers are not affecting the average that much, there's must be some different reasons. We will investigate further.

### Median Ride Length Per Day

Let's look at the median ride length for both annual members and casual riders. Here's the SQL Query

Median ride length for annual members:

median_ride_length	member_casual	day_of_week	
00:08:29	member	1	
00:08:13	member	7	
00:07:54	member	2	
00:07:52	member	4	
00:07:36	member	6	Note:
00:07:35	member	3	sunday=1
00:07:32	member	5	saturday=6

Median ride\_length for casual riders

Median_ride_length	member_casual	day_of_week
00:15:37	casual	1
00:14:22	casual	7
00:14:12	casual	2
00:13:08	casual	6
00:12:34	casual	4
00:11:02	casual	5
00:10:35	casual	3

Very interesting! The median ride length for casual riders on the top five days (SUN, SAT, MON, TUE, WED) is nearly double the amount for annual members on their top five days (SAT, SUN, MON, TUE, WED).

### Total rides per day

Let's look at total rides per day. We'll create columns for overall total, annual members and casual riders:

The screenshot shows the Google Cloud BigQuery console. The Explorer on the left lists workspace resources including months (APR to SEP) and quarters (QUARTER1 to QUARTER4). The main editor displays a query titled 'Untitled 5' that calculates total trips, member trips, and casual trips by quarter and day of week. The query results table shows data for quarters 1 through 7, with columns for day\_of\_week, total\_trips, MemberTrips, and CasualTrips.

```

1 SELECT
2   day_of_week,
3   COUNT(DISTINCT ride_id) AS total_trips,
4   SUM(CASE WHEN member_casual = 'member' THEN 1 ELSE 0 END) AS MemberTrips,
5   SUM(CASE WHEN member_casual = 'casual' THEN 1 ELSE 0 END) AS CasualTrips
6 FROM
7   `my-project-1-373317.cyclistic.QUARTER1`
8 GROUP BY
9   1
10 ORDER BY
11   total_trips DESC LIMIT 7

```

Row	day_of_week	total_trips	MemberTrips	CasualTrips
1	2	84943	57027	27916
2	3	69522	52609	16913
3	4	68218	53196	15022
4	3	66171	54308	11863
5	5	60815	48241	12574
6	1	54551	36196	18355
7	7	51807	34541	17266

## Start Stations

Next, we'll look at the most popular start stations for trips. We'll again include columns for overall, annual member and casual rider totals per start station:

The screenshot shows the Google Cloud BigQuery console with a new query titled 'Untitled 4'. This query calculates total rides, member rides, and casual rides for each start station name. The query results table lists the top start stations, including Kingsbury St & Kinzie St, Streeter Dr & Grand Ave, University Ave & 57th St, Ellis Ave & 60th St, and Clark St & Elm St.

```

1 SELECT
2   DISTINCT start_station_name,
3   SUM(
4     CASE WHEN ride_id = ride_id AND start_station_name = start_station_name THEN 1 ELSE 0 END
5   ) AS total,
6   SUM(
7     CASE WHEN member_casual = 'member' AND start_station_name = start_station_name THEN 1 ELSE 0 END
8   ) AS member,
9   SUM(
10    CASE WHEN member_casual = 'casual' AND start_station_name = start_station_name THEN 1 ELSE 0 END
11  ) AS casual
12 FROM `my-project-1-373317.cyclistic.QUARTER1`
13 GROUP BY
14   start_station_name

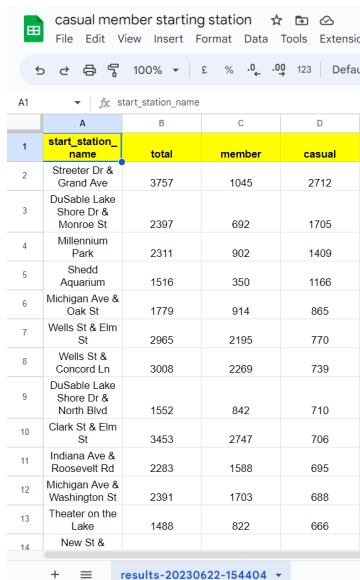
```

Row	start_station_name	total	member	casual
1	Kingsbury St & Kinzie St	4315	3761	554
2	Streeter Dr & Grand Ave	3757	1045	2712
3	University Ave & 57th St	3595	3143	452
4	Ellis Ave & 60th St	3581	3192	389
5	Clark St & Elm St	3453	2747	706

We can begin to see some interesting patterns in the start station data. It looks like casual riders and annual members tend to favor different regions for beginning their trips. By updating the **ORDER BY** function to sort by

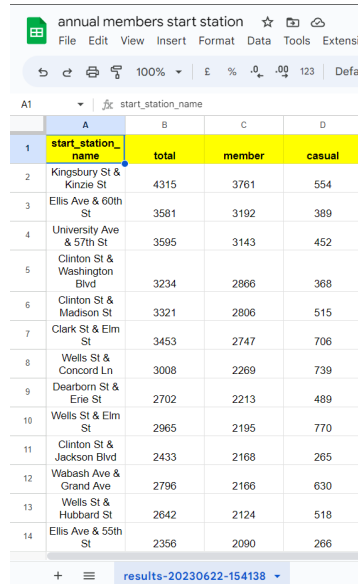
**CASUAL DESC** and **MEMBER DESC** in two separate queries, we can compare the top ten start stations for both:

Here are the results for both the separate queries:



The screenshot shows a Google Sheet titled 'casual member starting station'. The data is as follows:

	A	B	C	D
1	start_station_name	total	member	casual
2	Streeter Dr & Grand Ave	3757	1045	2712
3	DuSable Lake Shore Dr & Monroe St	2397	692	1705
4	Millennium Park	2311	902	1409
5	Shedd Aquarium	1516	350	1166
6	Michigan Ave & Oak St	1779	914	865
7	Wells St & Elm St	2965	2195	770
8	Wells St & Concord Ln	3008	2269	739
9	DuSable Lake Shore Dr & North Blvd	1552	842	710
10	Clark St & Elm St	3453	2747	706
11	Indiana Ave & Roosevelt Rd	2283	1588	695
12	Michigan Ave & Washington St	2391	1703	688
13	Theater on the Lake	1488	822	666
14	New St &			



The screenshot shows a Google Sheet titled 'annual members start station'. The data is as follows:

	A	B	C	D
1	start_station_name	total	member	casual
2	Kingsbury St & Kinzie St	4315	3761	554
3	Ellis Ave & 60th St	3581	3192	389
4	University Ave & 57th St	3595	3143	452
5	Clinton St & Washington Blvd	3234	2866	368
6	Clinton St & Madison St	3321	2806	515
7	Clark St & Elm St	3453	2747	706
8	Wells St & Concord Ln	3008	2269	739
9	Dearborn St & Erie St	2702	2213	489
10	Wells St & Elm St	2965	2195	770
11	Clinton St & Jackson Blvd	2433	2168	265
12	Wabash Ave & Grand Ave	2796	2166	630
13	Wells St & Hubbard St	2642	2124	518
14	Ellis Ave & 55th St	2356	2090	266

So there are some stations which belong to top 10 of both the list, These are the stations preferred by both casual and annual members. **A full fledged marketing campaign can be done near the stations crowded by casual members.**

Same analysis can be performed on QUARTER2, QUARTER3 AND QUARTER4 just by slightly modifying the code(only have to replace the file names and we're done with the basic analysis).

Numbers and text can only tell you this much, to know your data more, you have to use visualisation tools. I have two Visualisation tools : Tableau and R(it's more than just a visualisation tool). Tableau is simple to use but connecting data from BigQuery to Tableau requires Tableau Desktop which is not free. That's why i'll go with R here.

## Visualisation with R

First let's start with installing and loading basic packages required for importing and visualising data:

(I didn't need to install packages as they were already installed in my machine)

```
library(tidyverse)
library(skimr)
library(readr)
library(ggplot2)
```



```
library(janitor)
```

Loading datasets in R

```
trip_jan <- read_csv("D:\\case study 1 data\\ready for SQL\\202201cs1.csv")  
Error: '\\c' is an unrecognized escape in character string (<input>:1:26)
```

Got an error while loading dataset only 😊

Let's see what can be done...

It worked it worked 😊

The R community is justtt lovee..

This code worked (find the difference):

```
trip_jan <- read_csv("D:/case study 1 data/ready for SQL/202201cs1.csv")
```

Binding dataset in R

```
trip_2022 <-  
rbind(trip_jan,trip_feb,trip_mar,trip_apr,trip_may,trip_june,trip_july,trip_a  
ug,trip_sep,trip_oct,trip_nov,trip_dec)
```

Now we have combined 12 months of data, let's visualise

## Total Rides taken by Members & Casual Riders

Code:

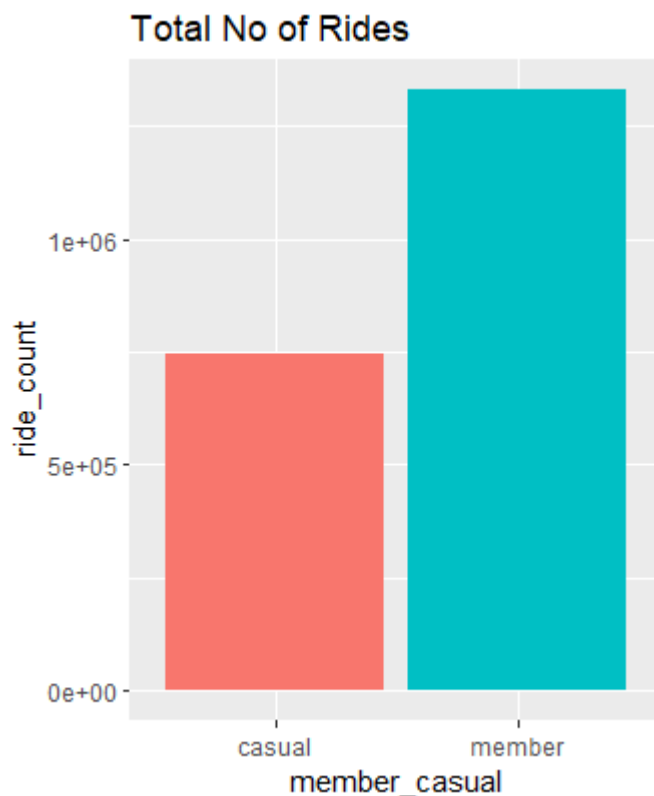
```
trip_2022 %>%  
+ group_by(member_casual) %>%  
+ summarise(ride_count=length(ride_id)) %>%  
+ ggplot()+geom_col(mapping = aes(x=member_casual,y=ride_count,  
fill=member_casual,),show.legend = "false") +  
+ labs(title = "Total No of Rides")
```

## Comparison of Total rides with the Type of Ride

code:

```
trip_2022 %>%  
+ group_by(member_casual, rideable_type) %>%  
+ summarise(number_of_rides = n(), .groups = "drop") %>%  
+ ggplot() + geom_col(mapping = aes(x = rideable_type, y = number_of_rides,  
fill = member_casual), show.legend = TRUE) +  
+ labs(title = "Total no. of Rides vs. Ride Type")
```

output:



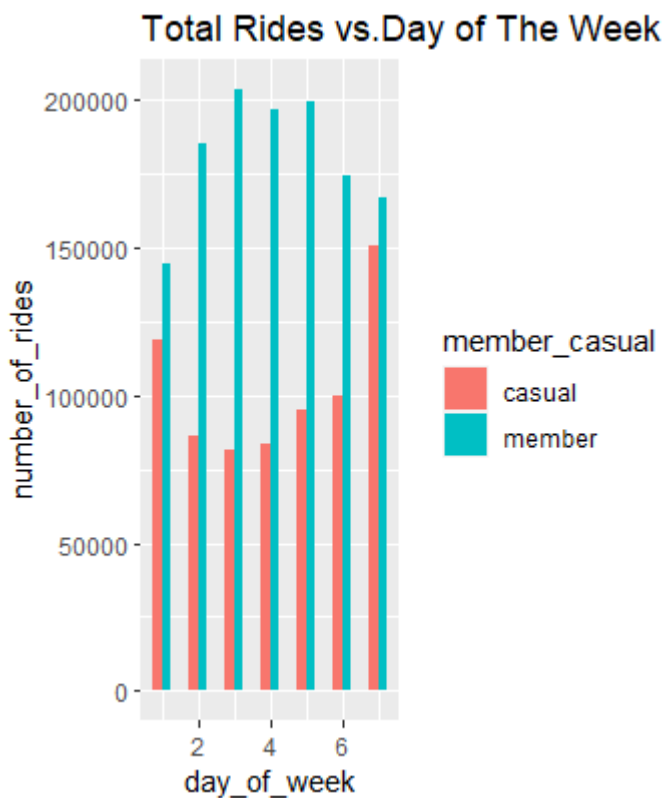
We see annual members travelled almost twice as that of annual members, This confirms that annual members contribute more to the revenue generation as compared to casual members.

## Days of the Week with No. of Rides taken by Riders

Code:

```
trip_2022 %>%  
+ group_by(member_casual, day_of_week) %>%  
+ summarise(number_of_rides=n(), .groups = "drop") %>%  
+ ggplot(aes(x = day_of_week, y = number_of_rides, fill = member_casual)) +  
+ labs(title = "Total Rides vs.Day of The Week") +  
+ geom_col(width = 0.5, position = position_dodge(width = 0.5)) +  
+ scale_y_continuous(labels = function(x) format(x,scientific = FALSE))
```

Output:



**An intriguing observation emerges from the analysis: casual members demonstrate greater travel activity on weekends, whereas annual members exhibit higher travel volumes on weekdays, potentially indicating a weekday commuting pattern, especially for work-related purposes.**

## Average Ride Length by Day of the Week

To plot average ride by day of the week, we first need to find average trip duration,

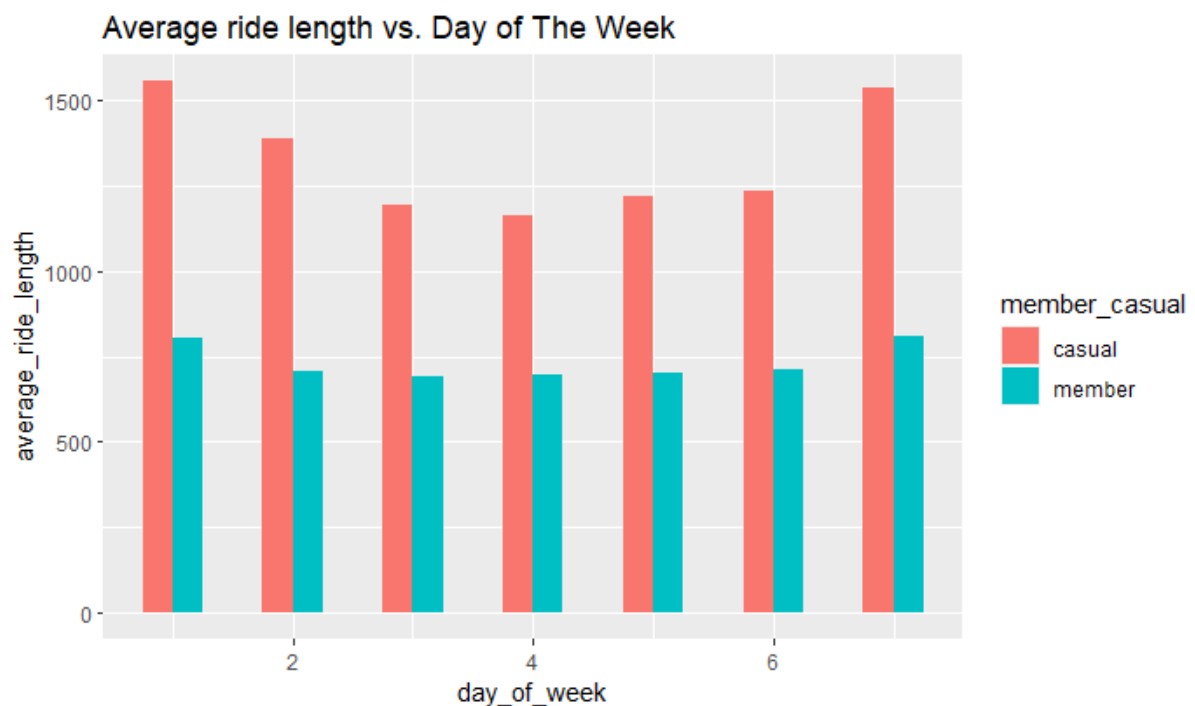
```
trip_2022 %>%  
  + group_by(member_casual) %>%  
  summarise(average_ride_length=mean(ride_length),median_ride_length=median(ride_length),max_ride_length=max(ride_length),min_ride_length=min(ride_length))
```

Code:

```
trip_2022 %>%  
  + group_by(member_casual,day_of_week) %>%  
  + summarise(average_ride_length=mean(ride_length), .groups = "drop") %>%
```

```
+ ggplot(aes(x = day_of_week, y = average_ride_length, fill =
member_casual)) +
+ geom_col(width = 0.5, position = position_dodge(width = 0.5)) +
+ labs(title = "Average ride length vs. Day of The Week")
```

Output:



The graph illustrates a consistent trend in the data, revealing that Saturdays and Sundays exhibit the highest levels of travel activity.

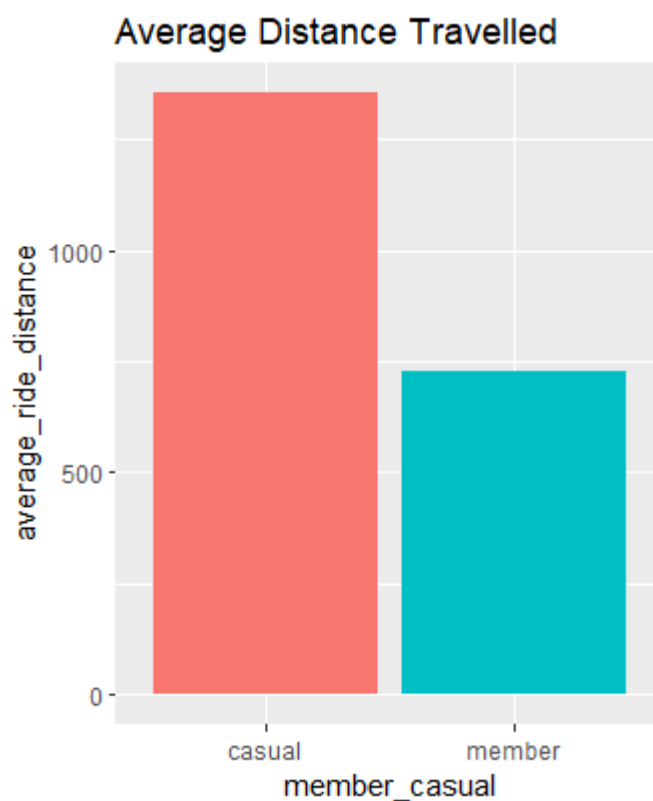
## Comparing Casual and Member Rides by Distance

Code:

```
trip_2022 %>%
+ group_by(member_casual) %>%
+ summarise(average_ride_distance = mean(ride_length)) %>%
+ ggplot() + geom_col(mapping = aes(x = member_casual, y =
average_ride_distance, fill = member_casual), show.legend = FALSE) +
```

```
+ labs(title = "Mean Distance Travelled")
```

*Output:*



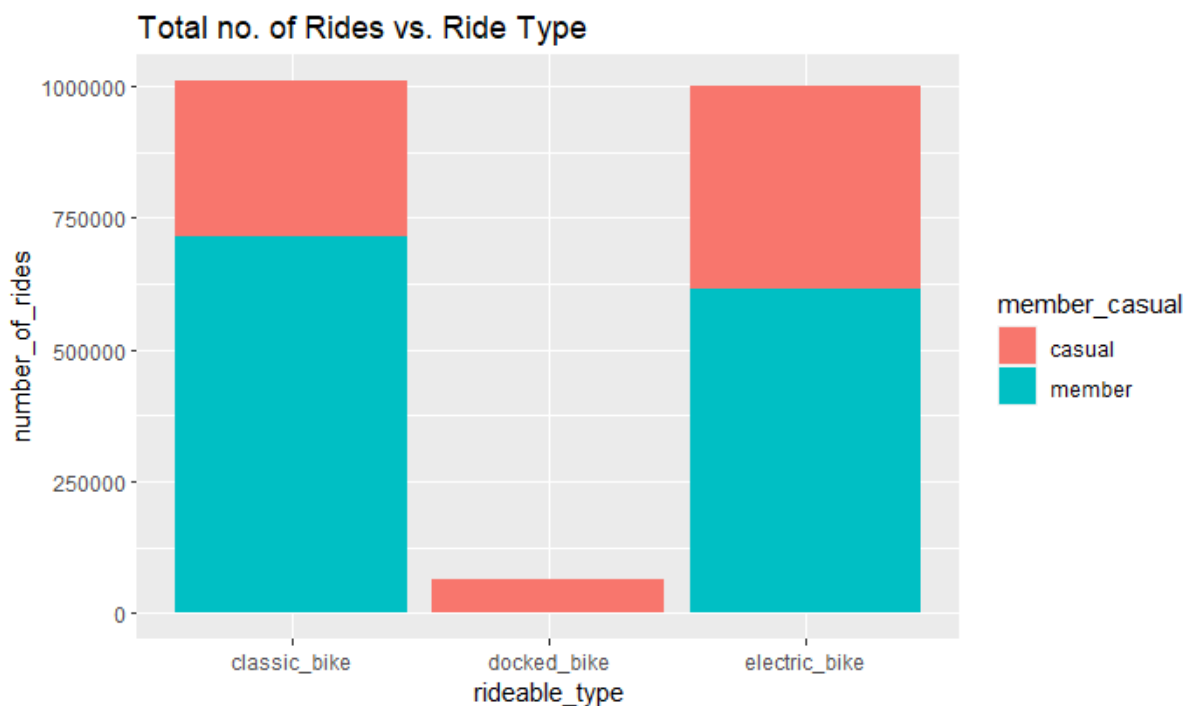
Undoubtedly, annual members have a higher ride count, but it is noteworthy that casual members travel twice the distance compared to annual members. We can develop targeted marketing campaigns to attract and encourage more casual members, emphasizing the benefit of longer rides and highlighting the potential savings in terms of distance covered. Promote features such as scenic routes, group ride opportunities, or challenges that appeal to casual riders' preference for longer distances. The company can also consider introducing pricing plans or discounts that cater specifically to riders who cover longer distances, incentivizing them to become regular users. This can help increase engagement and loyalty among casual riders.

## Comparison of Total rides with the Type of Ride

Code:

```
trip_2022 %>%  
+ group_by(member_casual, rideable_type) %>%  
+ summarise(number_of_rides = n(), .groups = "drop") %>%  
+ ggplot() + geom_col(mapping = aes(x = rideable_type, y = number_of_rides,  
fill = member_casual), show.legend = TRUE) +  
+ labs(title = "Total no. of Rides vs. Ride Type")
```

Output:

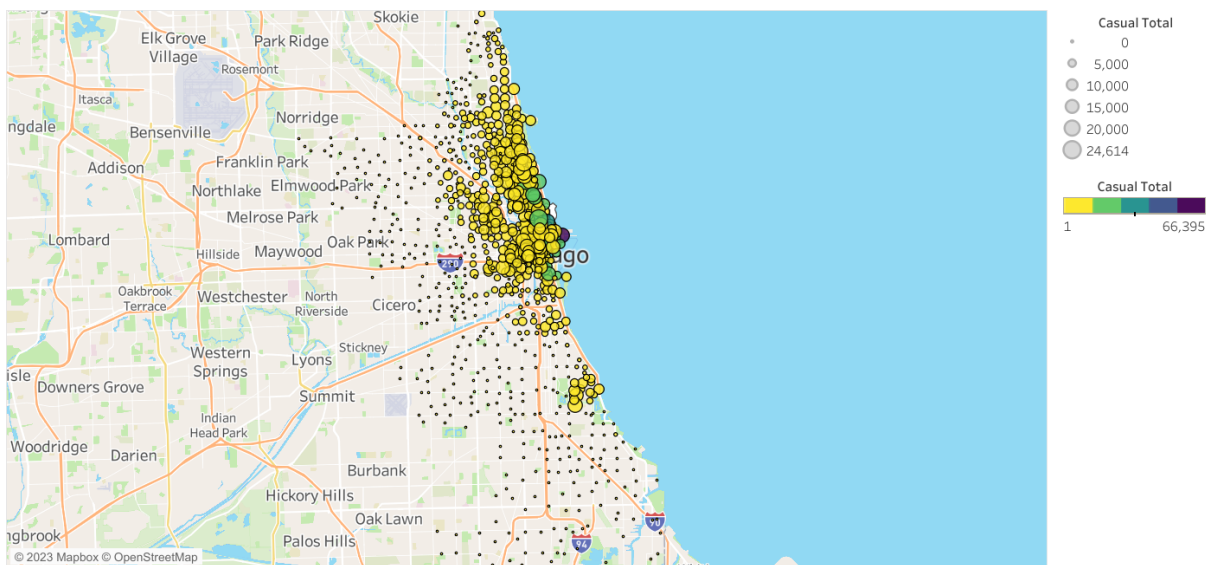


We see the share of electric bikes is almost the same as that of classic bike. This shows there is a growing preference for electric bikes among riders, the company can consider expanding its electric bike fleet. This can help attract and retain customers, particularly those who prefer the convenience and benefits of electric bikes. Also, The data on bike usage can inform targeted marketing campaigns and

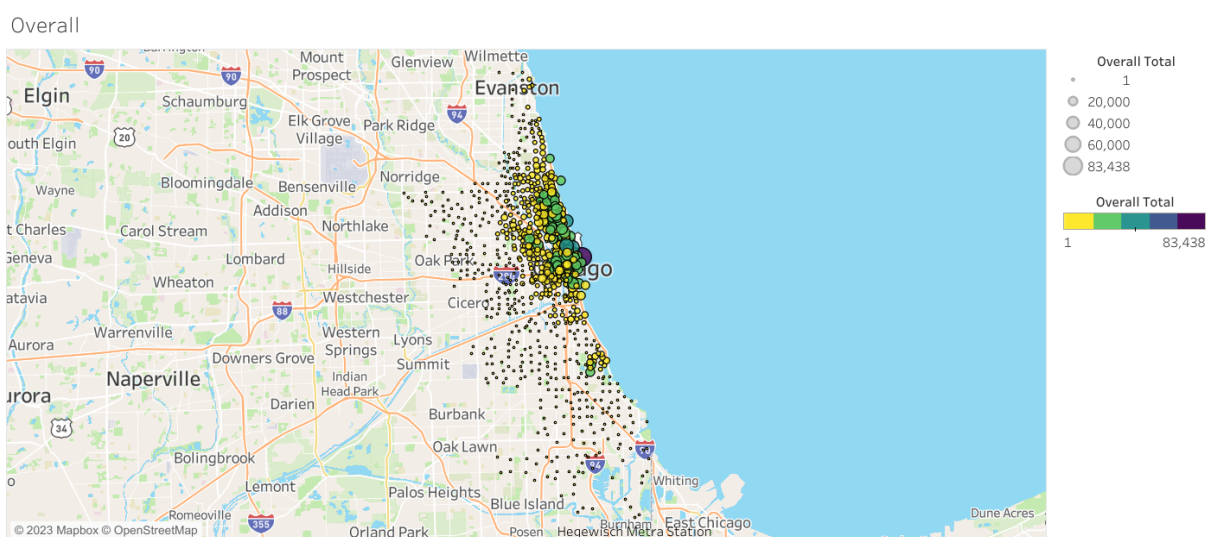
**promotions. For example, if there is a need to promote electric bikes more, the company can highlight their features and benefits in marketing materials or offer incentives to encourage more riders to try them.**

## Visualisation of starting and Ending Stations

Let's use Tableau for this as honestly I don't know how to visualise maps in R currently.



### Most popular start stations for overall riders



### Most popular ending stations for overall riders

**This shows the most popular start and end stations. The popularity of specific stations can inform targeted marketing and promotional efforts. The company can focus its marketing campaigns on these stations to attract more riders and raise awareness about the services offered. Promotions such as discounts, loyalty programs, or exclusive offers can be tailored to these popular stations to encourage ridership and create a positive association with the brand.**

**By leveraging the knowledge of the most popular stations, Cyclistic can optimise bike distribution, improve infrastructure, target marketing efforts, form strategic partnerships, and evaluate performance. These actions contribute to enhancing the overall user experience, attracting more riders, and ultimately increasing the company's success in the bike-sharing market.**

## **Top Recommendation**

- **Offer discounted annual membership plans or incentives specifically designed to encourage casual riders to become annual members. Highlight the long-term cost savings and benefits of regular usage.**
- **Develop targeted marketing campaigns at the most popular stations showcasing the advantages of annual membership, such as access to exclusive features, priority bike availability, or special events.**
- **Offer trial periods or short-term discounted memberships to allow casual riders to experience the benefits of being an annual member before committing to a full-term membership.**
- **Provide personalised recommendations or tailored rewards based on the riding patterns of casual riders, demonstrating the value they can gain as annual members.**