Google capstone project report

Ajiboye Abeeb

2023-01-14

Introduction

Cyclistic launched a successful bike-share program in 2016. Since then, the program has grown to a fleet of 5,824 geotracked bicycles that are locked into a network of 692 stations throughout Chicago. The bikes can be unlocked at any time from any station and returned to any other station in the system.

This is my first case study as a data analyst, and it is part of the Google Data Analytics Certificate optional Course capstone project.

The goal is to analyze a 12-month dataset for Cyclistic, a fictional bike sharing company, in order to gain insight for a marketing campaign.

Although the company is made up, the dataset used is real and was collected between January and December of 2022.

In order to properly analyze these data in order to answer the key business questions and make recommendations, I will follow the key steps of Data Analysis Process: Ask, Prepare, Process, Analyze, Share and Act

Scenerio

I am a junior data analyst on the marketing analyst team at Cyclistic, a Chicago-based bike-share company. The marketing director believes that increasing the number of annual memberships is critical to the company's future success. As a result, my team is interested in learning how casual riders and annual members use Cyclistic bikes differently.

My team will develop a new marketing strategy based on these findings in order to convert casual riders into annual members. However, Cyclistic executives must first approve my recommendations, which must be supported by compelling data insights and professional data visualizations.

I. ASK

a. Business task

The objective is to convert casual riders to annual members in order to increase company profitability.

b. Business Task for Junior Analyst

As the company's junior analyst, I've been tasked with demonstrating the following insight using the dataset: * How casual riders and annual members ride Cyclistic bikes differently.

c. Key Stakeholders

Project stakeholders include:

- Lily Moreno, Director of Marketing at Cyclistic, is in charge of the company's marketing campaigns.
- The marketing analytics team at Cyclistic. This team is in charge of gathering, analyzing, and reporting data for use in marketing campaigns. This team's junior analyst is me.
- The Cyclistic management team. This group makes the final call on the recommended marketing strategy. They are well-known for their attention to detail.

II. PREPARE

Where Can I Find the Data?

The data is available here (https://divvy-tripdata.s3.amazonaws.com/index.html). Motivate International Inc. gathered and uploaded the data.

How Is The Data Organized?

The information is stored in monthly csv files. For this project, I used a total of 12.csv files from January 2022 to December 2022. On my computer, I saved the dataset in the “Google capstone project” folder.

The Data’s Credibility

Motivate, Inc., the company that manages the City of Chicago’s Cyclistic Bike Share program, collects the data directly. It is comprehensive and up to date because it is collected monthly, with the most recent data uploaded in January 2023.

Authorization, Privacy, Security, and Accessibility

The data is released under the terms of this license (https://ride.divvybikes.com/data-license-agreement)

The data does not contain any personally identifiable information about the riders, so it is secure and open to the public.

A. Data Credibility and Integrity Assessment

I will use the ROCCC (Reliable, Original, Comprehensive, Current, and Cited) data test model to determine the dataset’s credibility and reliability.

- Reliable — MEDIUM — Fairly reliable because it attracts a large number of users
- Original — HIGH — the originator (https://divvy-tripdata.s3.amazonaws.com/index.html)
- Comprehensive — HIGH — Data falls within the parameters of the Cyclic business task.
- Current — HIGH — Data’s most recent version is dated January 2023 and is based on the 2022 dataset.
- Referenced — MEDIUM — Motivate International Inc. made the dataset available.

I notice some limitations. Using data such as;

• Some of the ride id data was incorrect because it contained characters greater or less than 16.

• The dates and times in the started time and started date columns were later than in the ended time and ended date columns, respectively.

• There were a lot of empty rows in the start station and end station columns.

B. Data Selection

The following files from the dataset has been selected:

- 202112-divvy-tripdata.zip
- 202201-divvy-tripdata.zip
- 202202-divvy-tripdata.zip
- 202203-divvy-tripdata.zip
- 202204-divvy-tripdata.zip
- 202205-divvy-tripdata.zip
- 202206-divvy-tripdata.zip
- 202207-divvy-tripdata.zip
- 202208-divvy-tripdata.zip
- 202209-divvy-tripdata.zip
- 202210-divvy-tripdata.zip
- 202211-divvy-tripdata.zip

III. PROCESS

Here, we will clean the data to ensure that it is correct, complete, and error-free for further analysis:

- Investigate and observe data
- Look for missing or null values
- Transform data — format data type
- Perform statistical analysis

Tools used throughout the process includes Microsoft Excel, Microsoft SQL Server, Tableau, and R for cleaning, analyzing, visualizing, and reporting data.

Data Cleaning and Extraction of Data from Existing Fields (Microsoft Excel)

- To validate the ride id data, I added a new column and used the Len function (LEN()) to determine the number of characters. I filtered the column to show only characters that were greater than or less than 16, then deleted those rows.
- The columns started at and ended at contained both date and time data. To distinguish them, I separated the date and time data into two columns (labeled “started date, ended date” and “started time, ended time”). Then I made a ride length column and used (=F2-D2) to populate the remaining rows. I added another ride length column in which I converted the time data to seconds for easy aggregation (=24*60*60*(ride length)).
- I created a started day and ended day column and filled it with the day of the week those trips began and ended by using the formulas (=TEXT(C2,“dddd”)) and (=TEXT(F2,“dddd”)) and populating other rows.
- Just in case Some information in the started date column will be greater than information in the ended date column. I added a new column and used logic (=IF(F2>=C2,“YES”,“NO”) to populate other rows and find any rows with the “NO” error, then filtered and deleted them.
- Some data in the started time column exceeded those in the ended time column. As a result, some rows in the ride length column returned a value minus error. Because the started_ and ended dates were on different days and the started time was greater than the ended time, some of the rows had the value error. For example, (started date = Monday, ended date = Tuesday, and started Time = 11:00:00 p.m., ended

Time = 01:00:00 a.m.). I added a new column and used a logical function (=IF(F2>=C2, “chg”,“del”) to separate these rows from the other value error rows. Then I filtered the column to show only the rows that contained the word “delete” and deleted them.

- The rows with the word “change” I calculated the ride length using the formula (=IF(endtime>starttime, endtime-starttime, 1-starttime+endtime)).
- For more information, I used the formula (=IF(MOD(E2+“05:00”,1)>0.5,“Day”,“Night”) to determine whether the start and end time for the rides are day or night, as well as the column start and end period.
- Finally, non-useful columns such as start station id, end station id, start lat, start lng, end lat, and end lng were removed.

IV. ANALYZE (Microsoft SQL Server)

I imported the 12 tables into MS SQL Server for analysis after cleaning the data.

To begin, I used a union all query to combine the 12 tables. After that, I created a new Cyclistic bike-share table and populated it with the union all query. (display)

I used the new table to analyze the data, which led me to the following conclusions:

• Preferred bike type of users

• Day of week users ride

• Time of Day users ride

• Average Ride Length for both Member type

• Users total rides

• Maximum ride length between both user type

• How many rides got started and ended immediately

• Which stations has the highest rides that got started and ended immediately

• To know which station is mostly used to start/end trip and by either user type excluding unused trips. showing: member start trip, member end trip, casual start trip, casual end trip

You can find the SQL codes here.

Loading of pre-installed packages The R libraries had already been installed and loaded for readability.

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✓ ggplot2 3.4.0      ✓ purrr   0.3.5
## ✓ tibble  3.1.8      ✓ dplyr   1.0.10
## ✓ tidyr   1.2.1      ✓ stringr 1.5.0
## ✓ readr   2.1.3      ✓ forcats 0.5.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✕ dplyr::filter() masks stats::filter()
## ✕ dplyr::lag()    masks stats::lag()
```

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

Connecting to Database

```
con <- DBI::dbConnect(odbc::odbc(),
                      Driver = "SQL Server",
                      Server = "DAFLEXYBEE\\SQLEXPRESS01",
                      Database = "master",
                      Trusted_Connection = "True")
```

Preview the imported datasets

```
SELECT top 10*
FROM Capstone_project
```

Displaying records 1 - 10

ride_id	rideable_type	started_date	started_day	started_Time	started_period	ended_date	ended_day	ended_time	ende
00000179CF2C4FB5	electric_bike	2022-07-28	Thursday	09:02:00.0000000	Day	2022-07-28	Thursday	09:13:00.0000000	Day
0000047373295F85	electric_bike	2022-07-22	Friday	16:56:00.0000000	Day	2022-07-22	Friday	17:21:00.0000000	Day
000004C3185FDDE9	electric_bike	2022-07-17	Sunday	13:01:00.0000000	Day	2022-07-17	Sunday	13:16:00.0000000	Day
000005B1F6F86B03	electric_bike	2022-07-02	Saturday	20:02:00.0000000	Night	2022-07-02	Saturday	20:08:00.0000000	Night
000008FF2B1BB8EC	electric_bike	2022-06-12	Sunday	05:12:00.0000000	Night	2022-06-12	Sunday	05:24:00.0000000	Night
00000B26583EB490	electric_bike	2022-08-04	Thursday	22:35:00.0000000	Night	2022-08-04	Thursday	22:40:00.0000000	Night

ride_id	rideable_type	started_date	started_day	started_Time	started_period	ended_date	ended_day	ended_time	ende
00000E22FBA89D81	electric_bike	2022-05-19	Thursday	14:42:00.0000000	Day	2022-05-19	Thursday	14:54:00.0000000	Day
00000E408DED6BFB	electric_bike	2022-09-02	Friday	12:06:00.0000000	Day	2022-09-02	Friday	12:24:00.0000000	Day
0000144FC458F130	classic_bike	2022-06-12	Sunday	10:45:00.0000000	Day	2022-06-12	Sunday	10:55:00.0000000	Day
00001A4BA227DAFA	electric_bike	2022-10-13	Thursday	18:44:00.0000000	Day	2022-10-13	Thursday	18:49:00.0000000	Day

(Q1) To determine number of users who uses which Type of bikes and if they are Casual or Annual Members

```
select member_casual as user_type, rideable_type as bike_type, count(rideable_type) as NoOfUsers_PerBike
from ..Capstone_project
group by member_casual, rideable_type
order by NoOfUsers_PerBike DESC
```

6 records

user_type	bike_type	NoOfUsers_PerBike
member	classic_bike	1724027
member	electric_bike	1422355
casual	electric_bike	1175769
casual	classic_bike	922214
casual	docked_bike	201416
member	docked_bike	7834

(Q2) To determine which day of the week has more Rides for Casual and Annual Members

```
SELECT member_casual as user_type, started_day, count(started_day) as No_Of_dayRides
FROM ..Capstone_project
GROUP BY member_casual, started_day
ORDER BY No_Of_dayRides DESC
```

Displaying records 1 - 10

user_type	started_day	No_Of_dayRides
member	Wednesday	491513
member	Thursday	487507
member	Tuesday	480414
casual	Saturday	469370
member	Friday	452887
member	Monday	438840
member	Saturday	428121
casual	Sunday	392847
member	Sunday	374934
casual	Friday	336827

(Q3) To understand which time of the day Casual and Annual Members initiate rides most

```
select member_casual as user_type, started_period, count(started_period) as No_Of_Rides_initiated
from ..Capstone_project
group by member_casual, started_period
order by No_Of_Rides_initiated DESC
```

4 records

user_type	started_period	No_Of_Rides_initiated
member	Day	2384014
casual	Day	1643631
member	Night	770202
casual	Night	655768

(Q4) to determine average ride length between Casual Riders and Annual Members

– For casual riders

```
select avg(ride_length_in_seconds) as AvgRideTime_inseconds
from ..Capstone_project
where member_casual = 'casual'
```

1 records

AvgRideTime_inseconds
1402.759

– Converting the calculation into proper time

```
DECLARE @s INT
SELECT
    @s = 1402.759
SELECT
    @s
  , CONVERT(TIME, DATEADD(SECOND, @s, 0)) as casual_average_time;
```

1 records

	casual_average_time
1402	00:23:22.0000000

– For member riders

```
select avg(ride_length_in_seconds) as AvgRideTime_inseconds
from ..Capstone_project
where member_casual = 'member'
```

1 records

AvgRideTime_inseconds
777.4674

– Converting the calculation into proper time

```
DECLARE @s INT
SELECT
    @s = 777.47
SELECT
    @s
  , CONVERT(TIME, DATEADD(SECOND, @s, 0)) as member_average_time;
```

1 records

	member_average_time
777	00:12:57.0000000

(Q5) To determing the Total number of Rides between both casual and member riders

```
select member_casual as user_type, count(member_casual) as Total_no_Of_Riders
from ..Capstone_project
group by member_casual
order by 2 desc
```

2 records

user_type	Total_no_Of_Riders
member	3154216
casual	2299399

(Q6) To Calculate the maximum ride_length by riders be it member or casual.

```
select member_casual as user_type, max(ride_length) as max_ridelength
from ..Capstone_project
group by member_casual
```

2 records

user_type	max_ridelength
member	23:59:00.0000000
casual	23:59:00.0000000

(Q7) To know how many rides got started and ended immediately

```
select  member_casual as user_type, count(ride_id) as no_rides_unused
    from(
        select *
        from ..Capstone_project
        where (started_time = ended_time) and (start_station_name is not null) and (end_station_name is not null)
        ) as rides
group by member_casual
```

2 records

user_type	no_rides_unused
member	27601
casual	14894

(Q8) To determine which station has the highest rides that got started and ended immediately

```
select  member_casual as user_type, start_station_name, count(start_station_name) as station_rides_unused
    from(
        select *
        from ..Capstone_project
        where (started_time = ended_time) and (start_station_name = end_station_name) and (start_station_name is not null) a
nd (end_station_name is not null)
        ) as rides
group by member_casual,start_station_name
order by 3 desc
OFFSET 0 ROWS
FETCH FIRST 10 ROWS ONLY
```

Displaying records 1 - 10

user_type	start_station_name	station_rides_unused
casual	Streeter Dr & Grand Ave	474
casual	Green St & Randolph St*	257
member	Loomis St & Lexington St	250
member	Streeter Dr & Grand Ave	230
member	Clark St & Elm St	228
casual	DuSable Lake Shore Dr & Monroe St	228
member	Kingsbury St & Kinzie St	215
casual	Michigan Ave & Oak St	205
casual	Millennium Park	201
casual	Bissell St & Armitage Ave*	196

(Q9) To know which station is mostly used to start/end trip and by which user? is it casual or members – (A) for casual riders start_trip

```
select member_casual as user_type, start_station_name, count(start_station_name) as casual_most_used_station_to_starttrip
    from (
        select *
        from ..Capstone_project
        where (member_casual LIKE 'casual') and (start_station_name is not null) and (end_station_name is not null) and (sta
rt_station_name != end_station_name)
        ) as casual_start_rides
group by member_casual, start_station_name
order by 3 DESC
OFFSET 0 ROWS
FETCH FIRST 10 ROWS ONLY
```

Displaying records 1 - 10

user_type	start_station_name	casual_most_used_station_to_starttrip
casual	Streeter Dr & Grand Ave	42631
casual	DuSable Lake Shore Dr & Monroe St	20859
casual	Millennium Park	20284
casual	Michigan Ave & Oak St	19555
casual	DuSable Lake Shore Dr & North Blvd	18424
casual	Shedd Aquarium	16798
casual	Theater on the Lake	15535

user_type	start_station_name	casual_most_used_station_to_starttrip
casual	Wells St & Concord Ln	14296
casual	Clark St & Armitage Ave	11891
casual	Clark St & Lincoln Ave	11866

(B) For member riders start_trip

```
select member_casual as user_type, start_station_name, count(start_station_name) as member_most_used_station_to_starttrip
    from (
        select *
        from ..Capstone_project
        where member_casual LIKE 'member' and (start_station_name is not null) and (end_station_name is not null) and (start
_station_name != end_station_name)
        ) as member_start_rides
group by member_casual, start_station_name
order by 3 DESC
OFFSET 0 ROWS
FETCH FIRST 10 ROWS ONLY
```

Displaying records 1 - 10

user_type	start_station_name	member_most_used_station_to_starttrip
member	Kingsbury St & Kinzie St	20745
member	Clark St & Elm St	19794
member	Wells St & Concord Ln	18842
member	Wells St & Elm St	16735
member	Clinton St & Madison St	15725
member	Broadway & Barry Ave	15629
member	Loomis St & Lexington St	15551
member	Clinton St & Washington Blvd	15519
member	St. Clair St & Erie St	15489
member	Dearborn St & Erie St	14980

(C) For casual riders end_trip

```
select member_casual as user_type, end_station_name, count(end_station_name) as casual_most_used_station_to_endtrip
    from (
        select *
        from ..Capstone_project
        where member_casual LIKE 'casual' and (start_station_name is not null) and (end_station_name is not null) and (start
_station_name != end_station_name)
        ) as casual_end_rides
group by member_casual, end_station_name
order by 3 DESC
OFFSET 0 ROWS
FETCH FIRST 10 ROWS ONLY
```

Displaying records 1 - 10

user_type	end_station_name	casual_most_used_station_to_endtrip
casual	Streeter Dr & Grand Ave	45142
casual	Millennium Park	22069
casual	DuSable Lake Shore Dr & North Blvd	21402
casual	Michigan Ave & Oak St	21330
casual	DuSable Lake Shore Dr & Monroe St	19413
casual	Theater on the Lake	17095
casual	Shedd Aquarium	15339
casual	Wells St & Concord Ln	14009
casual	Clark St & Lincoln Ave	12419
casual	Clark St & Armitage Ave	12087

(D) For member riders end_trip

```
select member_casual as user_type, end_station_name, count(end_station_name) as member_most_used_station_to_endtrip
    from (
        select *
        from ..Capstone_project
        where member_casual LIKE 'member' and (start_station_name is not null) and (end_station_name is not null) and (start
_station_name != end_station_name)
        ) as member_end_rides
group by member_casual, end_station_name
order by 3 DESC
OFFSET 0 ROWS
FETCH FIRST 10 ROWS ONLY
```

Displaying records 1 - 10

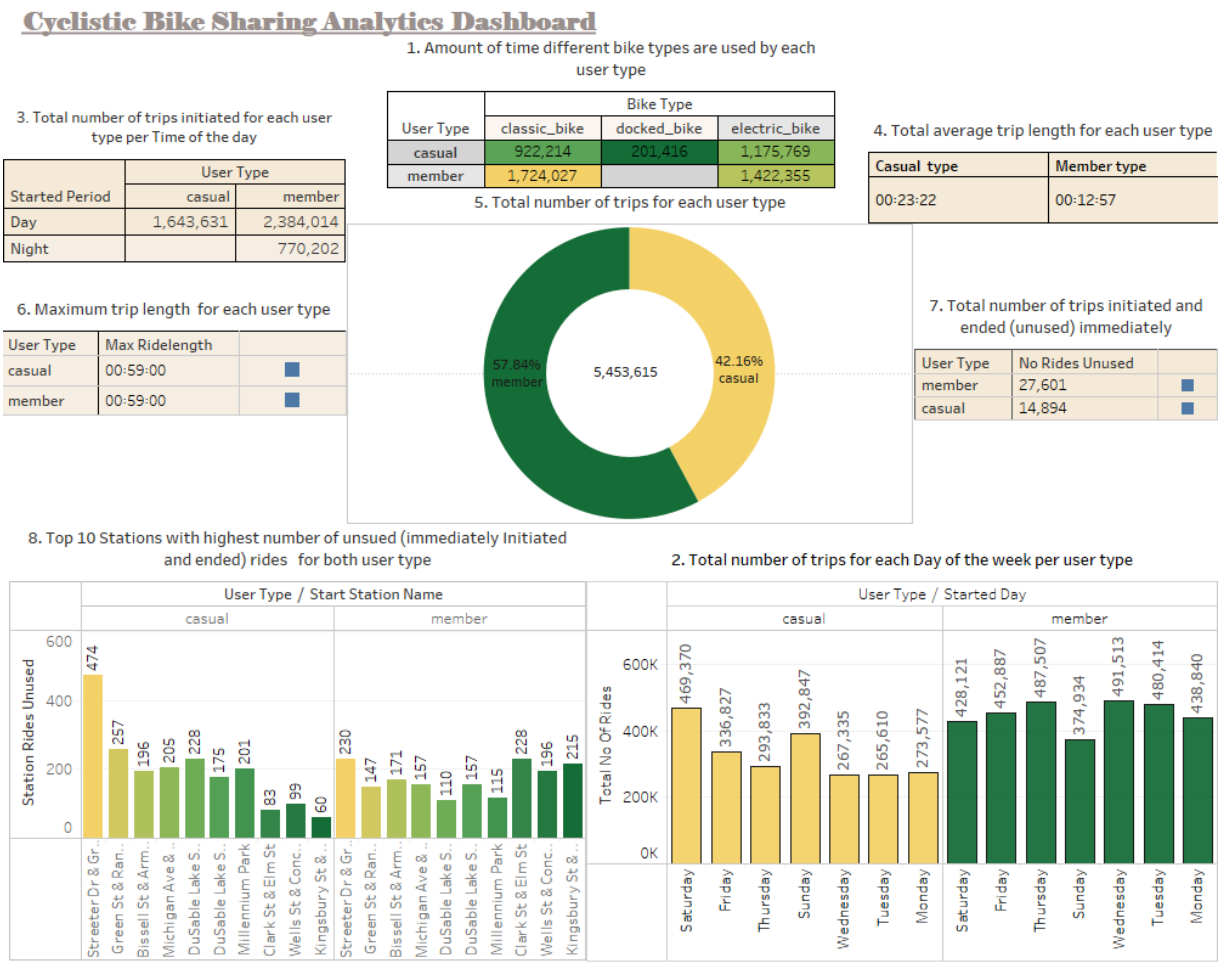
user_type	end_station_name	member_most_used_station_to_endtrip
member	Kingsbury St & Kinzie St	20621
member	Clark St & Elm St	20358
member	Wells St & Concord Ln	19511
member	Wells St & Elm St	16721
member	Clinton St & Madison St	16029
member	Broadway & Barry Ave	15987
member	Clinton St & Washington Blvd	15684
member	Loomis St & Lexington St	15440
member	Dearborn St & Erie St	15351
member	St. Clair St & Erie St	15295

V. SHARE

In this step, we will create vizualizations to convene the findings of our analysis.

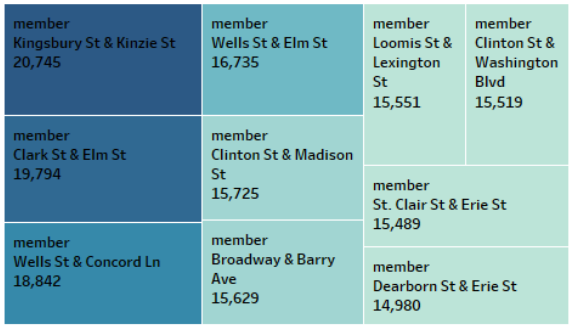
Here are the Dashboard

```
knitr::include_graphics("Dashboard 1dn.png",error = FALSE)
```

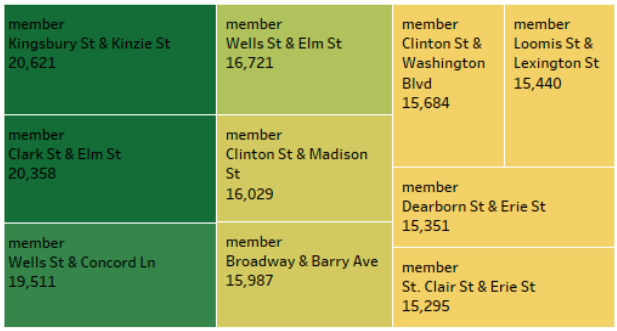



```
knitr::include_graphics("Dashboard 2dn.png",error = FALSE)
```

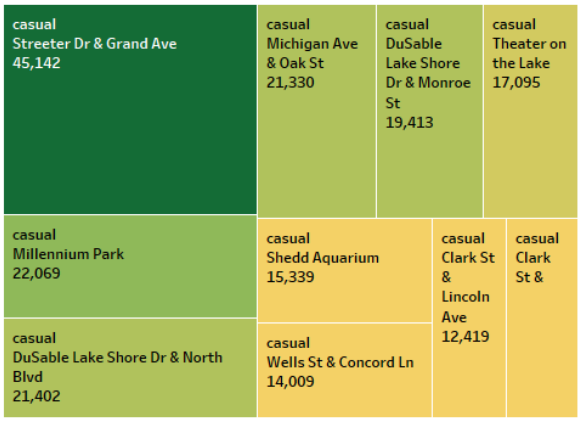
9b. Top 10 Most used stations by Member user type where trips are initiated



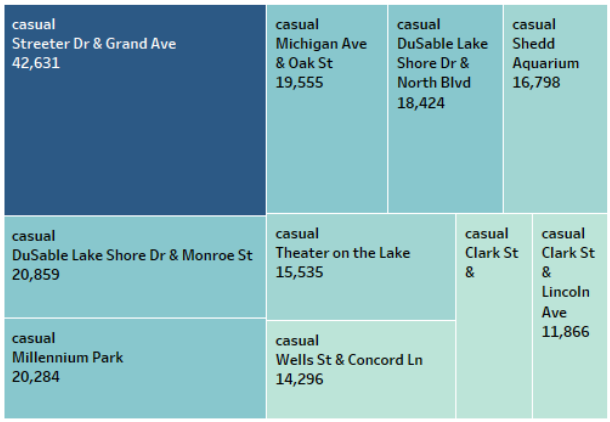
9d. Top 10 Most used stations by Member user type where trips ended



9c. Top 10 Most used stations by Casual user type where trips ended



9a. Top 10 Most used stations by Casual user type where trips are initiated



VI ACT

We will answer the key business question and provide recommendations based on our analysis to guide Cyclistic's marketing strategy in this final phase.

Summary of Analysis

Based on my research, I discovered the following differences in the riding habits of casual and annual members:

• Preferred bike type of users

Electric and classic bikes are common among both users, while docked bikes are fairly used by casual riders only Annual members typically don't.

• Day of week users ride

Casual riders ride more over the weekends (Friday, Saturday, and Sunday) while annual members ride more during the week (Tuesday, Wednesday, and Thursday)

• Time of Day users ride

With only two categories, day and night, casual riders ride more during the day, while annual members ride both day and night. This implies that casual riders ride for pleasure, whereas annual members most likely ride to and from work.

• Average Ride Length for both Member type

Casual riders ride for longer than member riders, averaging 23 minutes and 22 seconds per ride compared to 12 minutes and 57 seconds for annual members.

• Users total rides

Annual members take more trips than casual riders, with member riders accounting for 54.84% of total rides (3,154,216), while casual riders account for 42.16% of total rides (2,299,399), for a total of 5,453,615 rides between the two user types.

• Maximum ride/trip length between both user type

Both casual and member riders had traveled the longest distance in 59 minutes.

• How many rides got started and ended immediately

Trips were initiated and canceled immediately between both user types, with member riders canceling a total of 27,601 trips and casual users canceling 14,894 trips.

• Top 10 stations has the highest rides that got started and ended immediately

looking into user behavior patterns such as trip cancellations and which Stations trips were canceled the most by both users, as shown in the dashboard section titled 8. It is revealed that the start station named Streeter Dr & Grand Ave has the highest number of canceled trips by both user types. Further investigation may reveal why.

• To know which top 10 stations are mostly used to start/end trip and by either user type excluding unused trips.

We reveal the top 10 stations that are popular among both user types by analyzing user behavior patterns such as where trips are started and ended the most. Our findings show that Streeter Dr & Grand Ave are the most popular starting and ending points for casual users, while Kingsbury St & Kinzie St are the most popular starting and ending points for member users.

Recommendations

• An annual membership subscription discount should be offered exclusively to docked bikes in order to attract more annual users to the bike type.

•. A campaign should be launched to reward annual members who ride for longer periods of time. For example, the longer you ride, the more points you earn to win prizes such as free rides, gift cards, and so on.

• Marketing campaigns should be targeted for the busiest casual rider days (Friday, Saturday, and Sunday) and busiest hours to reach the most riders (during the day).

• Some type of discounted offer can be assigned to stations outside the top ten most used in order to attract more usage within the zones, and some type of reward can be assigned to the most used stations in order to ensure continuous usage.

• Targeted premium features can be offered to persuade casual users to join as members to meet their specific fun goals for riding mostly on weekends.