Cyclistic Case Study

Konstantinos Fotoglou

2023-03-31

Table of Contents

Ciclystic Case Study Introduction.

Company Information.

Cyclistic is a bike-share program that has grown since 2016 to operate 5,824 bikes and 692 stations across Chicago. The program offers flexible pricing plans to attract casual riders, but the finance team has concluded that annual members are more profitable. The company wants to convert more casual riders into annual members to maximize growth. To do this, they need to understand the differences between these customer segments, why casual riders might buy memberships, and how digital media could affect their marketing tactics. They plan to analyze the historical bike trip data to identify trends and design marketing strategies that target casual riders.

Scenario.

You are a junior data analyst working in the marketing analyst team at Cyclistic, a bike-share company in Chicago. The director of marketing believes the company's future success depends on maximizing the number of annual memberships. Therefore,your team wants to understand how casual riders and annual members use Cyclistic bikes differently. From these insights,your team will design a new marketing strategy to convert casual riders into annual members. But first, Cyclistic executives must approve your recommendations, so they must be backed up with compelling data insights and professional data visualizations.

The fictional characters and teams given by the case study:

- **Cyclistic:** A bike-share program that features more than 5,800 bicycles and 600 docking stations. Cyclistic sets itself apart by also offering reclining bikes, hand tricycles, and cargo bikes, making bike-share more inclusive to people with disabilities and riders who can't use a standard two-wheeled bike. The majority of riders opt for traditional bikes; about 8% of riders use the assistive options. Cyclistic users are more likely to ride for leisure, but about 30% use them to commute to work each day.
- **Lily Moreno:** The director of marketing and your manager. Moreno is responsible for the development of campaigns and initiatives to promote the bike-share program. These may include email, social media, and other channels.
- **Cyclistic marketing analytics team:** A team of data analysts who are responsible for collecting, analyzing, and reporting data that helps guide Cyclistic marketing strategy. You joined this team six months ago and have been busy learning about Cyclistic's mission and business goals — as well as how you, as a junior data analyst, can help Cyclistic achieve them.
- **Cyclistic executive team:** The notoriously detail-oriented executive team will decide whether to approve the recommended marketing program.

Three questions will guide the future marketing program:

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

In this case study, we are assigned to answer the first question:

- How do annual members and casual riders use Cyclistic bikes differently?

Our final report should include the following deliverables:

1. A clear statement of the business task. **(Ask)**
2. A description of all data sources used. **(Prepare)**
3. Documentation of any cleaning or manipulation of data. **(Process)**
4. A summary of your analysis. **(Analyze)**

5. Supporting visualizations and key findings. **(Share)**
6. Your top three recommendations based on your analysis. **(Act)**

1 Phase 1: Ask.

Guiding questions, Key tasks, and Deliverables.

Guiding questions

- What is the problem you are trying to solve?
- How can your insights drive business decisions?

Key tasks

- Identify the business task riders.
- Consider key stakeholders

Deliverable

- A clear statement of the business task

Note that the question that was assigned to us was:

- How do annual members and casual riders use Cyclistic bikes differently?

1.1 Working on the First Phase.

1.1.2 Identifying the business task and considering key stakeholders.

- Our business task is to identify the differences in the usage patterns between annual members and casual riders.
- The key stakeholders are the Director of Marketing, the Executive team, and the Cyclistic marketing analytics team

1.1.3 Giving a clear statement of the business task.

In a practical scenario, with access to a larger dataset, we would strive to gather relevant information and pose inquiries to acquire a comprehensive understanding of Cyclistic's bike-sharing program and their goals for the marketing campaign. Some potential inquiries we may consider include:

- What is Cyclistic's target market?
- How do casual riders differ from annual members?
- What are the pricing plans for casual riders and annual members?
- What is the current conversion rate from casual riders to annual members?
- How has Cyclistic marketed their bike-share program in the past?
- What digital media channels have been most effective for Cyclistic in attracting and retaining customers?
- What data is available on bike trip patterns and usage trends?

- How can we use data to identify opportunities to convert more casual riders into annual members?

By asking these questions and gathering information, we can start to develop a deeper understanding of Cyclistic's business and marketing goals, and begin to identify potential strategies to achieve those goals.

Considering that we are working within a hypothetical scenario, and that we have limited data at our disposal, we will concentrate on certain questions to answer our specific query (How do annual members and casual riders use Cyclistic bikes differently?). Some of the questions we will address include:

- How do casual riders differ from annual members?
- What data is available on bike trip patterns and usage trends?
- How can we use data to identify opportunities to convert more casual riders into annual members?

A clear statement of the business task

Upon analyzing the question assigned to us, we can define it as follows: 'How do annual members and casual riders exhibit varying usage patterns when utilizing Cyclistic bikes?' Therefore, our objective is to identify and understand the dissimilarities in how annual members and casual riders use Cyclistic bikes. To achieve this goal, we must gather data on the biking habits of both types of users. This data could encompass the number of trips taken, duration of trips, station locations, and times of day.

2 Phase 2: Prepare.

Guiding questions, Key tasks, and Deliverables

Guiding questions

- Where is your data located?
- How is the data organized?
- Are there issues with bias or credibility in this data? Does your data ROCCC?
- How are you addressing licensing, privacy, security, and accessibility?
- How did you verify the data's integrity?
- How does it help you answer your question?
- Are there any problems with the data?

Key tasks

- Download data and store it appropriately.
- Identify how it's organized.
- Sort and filter the data.
- Determine the credibility of the data.

Deliverable

- A description of all data sources used

2.1 Beginning of the preparation Phase.

To analyze and identify trends, Cyclistic's historical trip data from the previous 12 months will be utilized. The datasets, although under a different name due to Cyclistic being a fictional company, are appropriate and can answer the business questions. The data, made available by Motivate International Inc. under this license, is public and can be used to explore how Cyclistic bikes are being used by different customer types. However, it is important to note that data-privacy concerns prohibit the use of personally identifiable information of riders. Therefore, it is not possible to connect pass purchases to credit card numbers to determine whether casual riders live within the Cyclistic service area or have purchased multiple single passes.

The second dataset utilized in this project is the National Oceanic and Atmospheric Administration (NOAA) dataset of daily average temperatures for Chicago, Illinois, spanning several decades. The dataset includes information on the maximum, minimum, and average temperatures recorded each day, as well as other meteorological variables such as wind speed, precipitation, and humidity. The data is regularly updated with new measurements as they become available, and is publicly accessible on the NOAA website for researchers, weather enthusiasts, and anyone interested in analyzing or visualizing Chicago's temperature patterns over time. This dataset is a valuable resource for climate research, urban planning, and other applications that rely on accurate weather data.

2.1.1 Downloading the data and storing it correctly.

The data I will be using to conduct my analysis can be downloaded from https://divvy-tripdata.s3.amazonaws.com/index.html. The date that I'm writing this is 2023-03-31, so I'll be using data from the last 12 months. Note that data for March has not yet been made available, and as a result I will be using data from 2022-03 until 2023-02. After downloading the appropriate files, I extracted them in a unique directory so that I can access them easily later. For security reasons and mainly to prevent data loss, I also uploaded them in my personal Google Drive account after creating a master zip file containing all the datasets. The second dataset with the average temperature data can be found here

Notes

- To mass download the files from the site you can use Download Master
- To mass extract the files you can use 7zip

2.1.2 Identifying how the data is organized.

General information about the data.

Upon downloading the necessary data, I discovered that the dataset is comprised of several CSV files. Each file contains information pertaining to a particular month within a given

year of the bike-sharing program. For example, the file 202203-divvy-tripdata.csv contains details regarding bike trips taken in March 2022, such as start and end times, start and end stations, and user type.

To gain insight into the structure and organization of the data, I utilized spreadsheet software. I reviewed each of the downloaded files individually using Microsoft Excel, and found that the data is presented in a tabular format. The rows represent individual bike trips, while the columns signify different attributes or features of each bike trip. Generally speaking, the data is organized as a two-dimensional matrix, with rows indicating the records (cases) and columns indicating the features (variables).

Information about the features (variables).

Each divvy-tripdata.csv file in the Cyclistic dataset contains the following columns:

- ride_id - A unique identifier for each bike trip taken in the program.
- rideable_type - The type of bike used for the trip (e.g., electric bike, classic bike, etc.).
- started_at - The date and time when the bike trip started, in the format "YYYY-MM-DD HH:MM:SS".
- ended_at - The date and time when the bike trip ended, in the format "YYYY-MM-DD HH:MM:SS".
- start_station_name - The name of the bike station where the trip started.
- start_station_id - A unique identifier for the bike station where the trip started.
- end_station_name - The name of the bike station where the trip ended.
- end_station_id - A unique identifier for the bike station where the trip ended.
- start_lat - The latitude of the start station.
- start_lng - The longitude of the start station.
- end_lat - The latitude of the end station.
- end_lng - The longitude of the end station.
- member_casual - Indicates whether the user taking the bike trip is a member or a casual user.

2.1.3 Sorting and filtering the data.

Exploring the data in R.

I am loading the tidyverse lubridate and skimr packages to inspect the data.

```
library(tidyverse)
library(lubridate)
library(skimr)
```

Importing our data in a new variable.

```
# Importing the files to investigate our data in R
trdt_2302 <- read_csv("202302-divvy-tripdata.csv")
```

Inspecting the file to check the structure of our data.

```
glimpse(trdt_2302)
```

```
## Rows: 190,445
## Columns: 13
## $ ride_id            <chr> "CBCD0D7777F0E45F", "F3EC5FCE5FF39DE9",
"E54C1F27FA…
## $ rideable_type      <chr> "classic_bike", "electric_bike",
"classic_bike", "e…
## $ started_at         <dttm> 2023-02-14 11:59:42, 2023-02-15 13:53:48,
2023-02-…
## $ ended_at           <dttm> 2023-02-14 12:13:38, 2023-02-15 13:59:08,
2023-02-…
## $ start_station_name <chr> "Southport Ave & Clybourn Ave", "Clarendon Ave
& Go…
## $ start_station_id   <chr> "TA1309000030", "13379", "TA1309000030",
"TA1309000…
## $ end_station_name   <chr> "Clark St & Schiller St", "Sheridan Rd &
Lawrence A…
## $ end_station_id     <chr> "TA1309000024", "TA1309000041", "13156",
"TA1309000…
## $ start_lat          <dbl> 41.92077, 41.95788, 41.92077, 41.92087,
41.79483, 4…
## $ start_lng          <dbl> -87.66371, -87.64958, -87.66371, -87.66373,
-87.618…
## $ end_lat            <dbl> 41.90799, 41.96952, 41.88042, 41.87943,
41.78053, 4…
## $ end_lng            <dbl> -87.63150, -87.65469, -87.65552, -87.63550,
-87.605…
## $ member_casual      <chr> "casual", "casual", "member", "member",
"member", "…
```

After inspecting each file individually we can safely combine them into one data frame. (Didn't document the whole code of importing and inspecting each file individually since the same code is repeated for every .csv file.)

```
# Read in the CSV files and combine into one dataframe
df <- list.files(pattern = "*.csv") %>%
  map_df(read_csv)
```

Cheking the size of our data frame. From the output we can see that the data frame we have created contains 5829084 observations of 13 variables.

Now we cancheck if there are any duplicate values in our data set.

```
#Checking if there are any duplicate values by ride_id
dplct_ride_id <- df[duplicated(df$ride_id),]
dplct_ride_id
```

```
## # A tibble: 0 × 13
## # ℹ 13 variables: ride_id <chr>, rideable_type <chr>, started_at <dttm>,
## #   ended_at <dttm>, start_station_name <chr>, start_station_id <chr>,
```

```
## #   end_station_name <chr>, end_station_id <chr>, start_lat <dbl>,
## #   start_lng <dbl>, end_lat <dbl>, end_lng <dbl>, member_casual <chr>
```

We can identify that there are no duplicate values based on ride_id.

```
# Check the dimensions of the dataframe
dim(df)
```

```
## [1] 5829084      13
```

Skimming our data gives us valuable information about our dataset. We can see the types of our variables, the missing values in our data, and the rate at which variable is complete.

```
# Skim our data to gain some valuable information
skim(df)
```

Data summary

Name	df
Number of rows	5829084
Number of columns	13
_____	
Column type frequency:	
character	7
numeric	4
POSIXct	2
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
ride_id	0	1.00	16	16	0	5829084	0
rideable_type	0	1.00	11	13	0	3	0
start_station_name	850418	0.85	7	64	0	1692	0
start_station_id	850550	0.85	3	37	0	1314	0
end_station_name	909038	0.84	9	64	0	1715	0
end_station_id	909179	0.84	3	37	0	1318	0
member_casual	0	1.00	6	6	0	2	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
start_lat	0	1	41.90	0.05	41.64	41.88	41.90	41.93	42.07	▁▁▇▁
start_lng	0	1	-87.65	0.03	-87.84	-87.66	-87.64	-87.63	-87.52	▁▁▇▁
end_lat	5938	1	41.90	0.07	0.00	41.88	41.90	41.93	42.37	▁▁▁▁▁
end_lng	5938	1	-87.65	0.11	-88.14	-87.66	-87.64	-87.63	0.00	▇▁▁▁▁

Variable type: POSIXct

skim_variable	n_missing	complete_rate	min	max	median	n_unique
started_at	0	1	2022-03-01 00:00:19	2023-02-28 23:59:31	2022-08-02 12:43:52	4891324
ended_at	0	1	2022-03-01 00:04:30	2023-03-06 15:09:53	2022-08-02 13:02:33	4904887

Let's document our missing values in tables so we can explore them later.

```
# Creating tables with the missing values
missing_vals <- df %>% summarise_all(~sum(is.na(.)))

missing_vals_by_rideable_type <- df %>%
  group_by(rideable_type) %>%
  summarise_all(~sum(is.na(.)))

missing_vals_by_member_casual <- df %>%
  group_by(member_casual) %>%
  summarise_all(~sum(is.na(.)))
```

Exploring the tables of our missing values we can see that most station names missing values are associated with electric bikes and most coordinates missing values are associated with casual rides. Also we can see that in electric bikes we don't have any missing values for the coordinates.

```
missing_vals

## # A tibble: 1 × 13
##   ride_id rideable_type started_at ended_at start_station_name
start_station_id
##     <int>         <int>      <int>    <int>              <int>
<int>
## 1       0             0          0        0             850418
```

```
850550
## # ℹ 7 more variables: end_station_name <int>, end_station_id <int>,
## #   start_lat <int>, start_lng <int>, end_lat <int>, end_lng <int>,
## #   member_casual <int>

missing_vals_by_rideable_type

## # A tibble: 3 × 13
##   rideable_type ride_id started_at ended_at start_station_name
start_station_id
##   <chr>           <int>      <int>    <int>              <int>
<int>
## 1 classic_bike        0          0        0                  0
68
## 2 docked_bike         0          0        0                  0
0
## 3 electric_bike       0          0        0             850418
850482
## # ℹ 7 more variables: end_station_name <int>, end_station_id <int>,
## #   start_lat <int>, start_lng <int>, end_lat <int>, end_lng <int>,
## #   member_casual <int>

missing_vals_by_member_casual

## # A tibble: 2 × 13
##   member_casual ride_id rideable_type started_at ended_at
start_station_name
##   <chr>           <int>         <int>      <int>    <int>
<int>
## 1 casual              0             0          0        0
352329
## 2 member              0             0          0        0
498089
## # ℹ 7 more variables: start_station_id <int>, end_station_name <int>,
## #   end_station_id <int>, start_lat <int>, start_lng <int>, end_lat <int>,
## #   end_lng <int>
```

We can see that most of the missing values in the station names are matched with electric bikes.

Upon sorting and filtering the provided information, it became apparent that there are numerous instances of missing data concerning the starting and ending stations of the bike trip. Additionally, there are instances of missing station IDs, station names and coordinates. During this phase, it became clear that extensive work is necessary to refine and streamline the data, making it more comprehensive and easily manageable. One notable finding from this phase is the sheer quantity of missing data, specifically station names and IDs. It would be prudent to notify the executive team of this issue, as it may indicate a system performance problem that warrants further investigation. Resolving this issue would ultimately result in complete data without any missing values.

Based on my initial examination, the second dataset appears to be clean and ready to use, requiring little additional effort.

2.1.4 Determining the credibility of the data.

To identify if our data is credible we have to ask the following questions:

- Who created the dataset?
- Is it part of a credible organization
- When was the data last refreshed?

The data we possess in our case study is first-party data, indicating that it was gathered internally by the organization (First-party data: Data collected by an individual or group using their own resources). Our data is well-structured and presented in a tabular format, featuring rows and columns. It is advisable to use the ROCC process (Reliable, Original, Comprehensive, Current, Cited) as a best practice to distinguish reliable data.

Our data is:

- **Reliable** since it is internal data from the organization.
- **Original** because it is first party internal data collected from the organizations' systems.
- **Current** because we have the latest data (just one month old).
- **Cited** since the data has been vetted and cited.

You may have noticed that I did not mention the **Comprehensive** aspect of the process in regards to our dataset. The reason for this is that, due to the dataset containing 13 distinct features and many missing data, it may be challenging to deduce how the bikes are being utilized without performing some degree of data cleaning and manipulation.

2.1.5 A Description of all data sources used.

The Cyclistic (divvy) trip data - This is the primary source of information for our analysis, containing detailed information about each trip taken by Cyclistic users, including the start and end time, the starting and ending station, the bike ID, and other relevant details.

Dataset of daily average temperatures for Chicago, Illinois. The National Oceanic and Atmospheric Administration (NOAA) maintains a dataset of daily average temperatures for Chicago, Illinois, spanning several decades. The dataset includes information on the maximum, minimum, and average temperatures recorded each day, as well as other meteorological variables such as wind speed, precipitation, and humidity. I filtered the data in Excel and kept only data for average temperature and the date. This dataset is ROCC because:

- **Reliable** since it is internal data from a credible governmental organization.
- **Original** because it is second party data collected from a governmental organizations which analyzes weather.

- **Comprehensive** since it is complete with no missing data and it is easily explained and understood.
- **Current** because we have the latest data (just one month old).
- **Cited** since the data has been vetted and cited.

3 Phase 3: Process.

Guiding questions, Key tasks, and Deliverables

Guiding questions

- What tools are you choosing and why?
- Have you ensured your data's integrity?
- What steps have you taken to ensure that your data is clean?
- How can you verify that your data is clean and ready to analyze?
- Have you documented your cleaning process so you can review and share those results?

Key tasks

- Check the data for errors.
- Choose your tools.
- Transform the data so you can work with it effectively.
- Document the cleaning process.

Deliverable

- Documentation of any cleaning or manipulation of data

3.1 Beginning of Phase 3 Process

3.1.1 Choosing the tools.

I will be using R in RStudio and Tableau throughout the entire project as these tools are capable of performing all necessary tasks such as exploring data and creating comprehensive and visually appealing visualizations. The reason I opted for R in RStudio and Tableau is to improve my programming skills in R and to learn how to create visualizations in Tableau. I consider R to be an extremely robust and comprehensive tool for data analysis and manipulation, and when combined with the capabilities of Tableau, I believe it can tackle any project.

3.1.2 Check the data for errors.

I have already thoroughly checked the data during the preparation stage and found some mistakes. I found that some of the coordinate variables in the dataset, along with the station name and station ID variables, had absent values. Additionally, I used R to validate the accuracy of each variable type and the data integrity. For example, the started_at and

ended_at values had the appropriate POSIXct formatting. Also after further investigating the dataset, I decided to remove outlier data.

3.1.3 Transforming and cleaning the data so I can work with it effectively.

Extracting date and time information for further analysis.
The mutate() function adds three new variables weekday, weekend_weekday and month to the df dataframe. These variables contain the the day, if it is a weekday or not, and the month information.

```
# Extract date and time information
df <- df %>%
  mutate(weekday = factor(wday(started_at, week_start = 1),
                          levels = 1:7,
                          labels = c("Mon", "Tue", "Wed","Thu", "Fri", "Sat",
"Sun")),
                      weekend_weekday = if_else(weekday %in% c("Sat", "Sun"),
"weekend", "weekday"),
                      month = factor(month(started_at, abbr = TRUE, label =
TRUE)))
```

Calculating the time duration of each trip.
The time difference between the ended_at and started_at columns is calculated using the difftime function in the code, which then converts the resultant time difference to numeric values that reflect the ride's duration in minutes. The df data frame then contains these numbers in a new column titled ride_length. After calculating the ride length we round it.

```
# Calculating the length of each ride and rounding it.
df$ride_length <- as.numeric(difftime(df$ended_at, df$started_at, units =
"mins"))
df$ride_length <- round(df$ride_length, 2)
```

Calculating if a trip is a round trip.
This code creates a new variable in your data frame called "round_trip". It uses the ifelse function to check if the start station name is the same as the end station name for each trip. If it is, the value of "round_trip" is set to "yes", indicating a round trip. If not, the value is set to "no", indicating a one-way trip.

```
# create a new variable called "is_round_trip"
df$round_trip <- ifelse(df$start_station_name == df$end_station_name, "yes",
"no")
```

Converting some variables to factor.
In order to analyze categorical data, this will turn the given columns into factors.

```
# Converting rideable_type, member_casual, day_of_week to factor
df$rideable_type <- as.factor(df$rideable_type)
df$member_casual <- as.factor(df$member_casual)
df$round_trip <- as.factor(df$round_trip)
```

Cleaning the names of stations.
This code replaces the suffixes in the start_station_name and end_station_name columns with empty strings. By applying the gsub() function to the start_station_name and end_station_name columns with these regular expressions, any text inside parentheses or asterisks will be removed from the station names, resulting in cleaner and more consistent names.

```
# Replace suffixes in start_station_name column
df$start_station_name <- gsub("\\(.*?\\)", "", df$start_station_name) #
Remove text inside parentheses
df$start_station_name <- gsub("\\*", "", df$start_station_name) # Remove
asterisks

# Replace suffixes in end_station_name column
df$end_station_name <- gsub("\\(.*?\\)", "", df$end_station_name) # Remove
text inside parentheses
df$end_station_name <- gsub("\\*", "", df$end_station_name) # Remove
asterisks
```

Cleaning missing longtitude and lattitude values, and leaving only sensible ride lengths.
Since there are negative and extremely high ride length instances in the dataset, I decided to filter the data by only including rides with a length between 1 and 1440 minutes (24 hours) and drop the rows that have missing values in the end longitude and latitude columns, station names and station id's. The code chunk filters the dataframe "df" to include only rides with a length greater than 1 minute and less than 1440 minutes. Then, it removes rows that have missing values in the coordinates columns as well as the station names and station id's, using the "drop_na()" function. The "filter()" function is used to subset rows based on conditions. In this case, the condition is that the ride_length column should be greater than 1 and less than 1440.

```
# Filter for ride lengths > 1 minute and < 1440 minutes
df <- df %>%
  filter(ride_length > 1 & ride_length< 1440)

# Drop missing values in end_lng and end_lat variables
df <- df %>%
  drop_na(end_lng, end_lat)

# Drop missing values from the station names and station id variables.
df <- df %>%
  drop_na(start_station_id, start_station_name, end_station_id,
end_station_name)
```

Checking and dropping trips that are meant for repairs.
In this code we check how many trips are meant for repairs and then, we define a vector of strings that we want to remove from our dataset. We then use the grepl function with the paste function to check if any of these strings are present in start_station_name or end_station_name. We use the ! operator to negate the result of the grepl function, so that

rows containing the specified strings are filtered out. Finally, we use the filter function from dplyr to keep only the rows that don't contain any of the specified strings.

```
# Count rides that are meant for repairs.
sum(grepl("warehouse", df$start_station_name, ignore.case = TRUE) |
      grepl("warehouse", df$end_station_name, ignore.case = TRUE) |
      grepl("testing", df$start_station_name, ignore.case = TRUE) |
      grepl("testing", df$end_station_name, ignore.case = TRUE) |
      grepl("repair", df$start_station_name, ignore.case = TRUE) |
      grepl("repair", df$end_station_name, ignore.case = TRUE) |
      grepl("test", df$start_station_name, ignore.case = TRUE) |
      grepl("test", df$end_station_name, ignore.case = TRUE) |
      grepl("check", df$start_station_name, ignore.case = TRUE) |
      grepl("check", df$end_station_name, ignore.case = TRUE))

# create a vector of strings to remove
strings_to_remove <- c("warehouse", "testing", "repair", "test", "check")

# filter out rows containing the strings to remove
df <- df %>%
  filter(!grepl(paste(strings_to_remove, collapse = "|"), start_station_name,
ignore.case = TRUE),
         !grepl(paste(strings_to_remove, collapse = "|"), end_station_name,
ignore.case = TRUE))
```

Checking the structure of our cleaned data frame

```
skim(df)
```

Data summary

Name	df
Number of rows	441512 6
Number of columns	18
_____	
Column type frequency:	
character	6
factor	5
numeric	5
POSIXct	2
_____	
Group variables	None

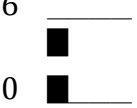
Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
ride_id	0	1	16	16	0	4415126	0
start_station_name	0	1	7	64	0	1555	0
start_station_id	0	1	3	37	0	1268	0
end_station_name	0	1	10	64	0	1597	0
end_station_id	0	1	3	37	0	1279	0
weekend_weekday	0	1	7	7	0	2	0

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
rideable_type	0	1	FALSE	3	cla: 2623492, ele: 1616855, doc: 174779
member_casual	0	1	FALSE	2	mem: 2650718, cas: 1764408
weekday	0	1	FALSE	7	Sat: 705569, Thu: 648865, Tue: 625628, Wed: 624243
month	0	1	TRUE	12	Jul: 630741, Jun: 609700, Aug: 594192, Sep: 525213
round_trip	0	1	FALSE	2	no: 4192785, yes: 222341

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
start_lat	0	1	41.90	0.04	41.65	41.88	41.90	41.93	42.06	
start_lng	0	1	-87.64	0.02	-87.83	-87.66	-87.64	-87.63	-87.53	
end_lat	0	1	41.90	0.07	0.00	41.88	41.90	41.93	42.06	
end_lng	0	1	-87.64	0.11	-87.83	-87.66	-87.64	-87.63	0.00	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
ride_length	0	1	17.10	31.06	1.02	6.18	10.67	19.03	1439.37	

Variable type: POSIXct

skim_variable	n_missing	complete_rate	min	max	median	n_unique
started_at	0	1	2022-03-01 00:00:19	2023-02-28 23:59:31	2022-07-31 16:30:25	3836744
ended_at	0	1	2022-03-01 00:04:30	2023-03-01 09:48:38	2022-07-31 16:54:55	3850031

Extracting the cleaned data to a new CSV file to work with in Tableau

```
write.csv(df, "clean_df.csv", row.names = FALSE)
```

3.1.4 Documenting the cleaning process.

I have recorded all the steps I took to clean and manipulate the dataset in this document earlier. You can refer to the table of contents at the beginning of the document to access each task performed in this case study.

To process the second dataset, I simply filtered out any unnecessary columns and retained only the DATE and TAVG columns, which contain information on the date and average temperature, respectively. I accomplished this task using Excel, as the dataset was relatively small.

3.1.5 Documentation of any cleaning or manipulation of data.

I have documented all of the cleaning and manipulation I performed in the datasets previously in this phase of the project.

4 Phase 4: Analyze and Phase 5: Share combined.

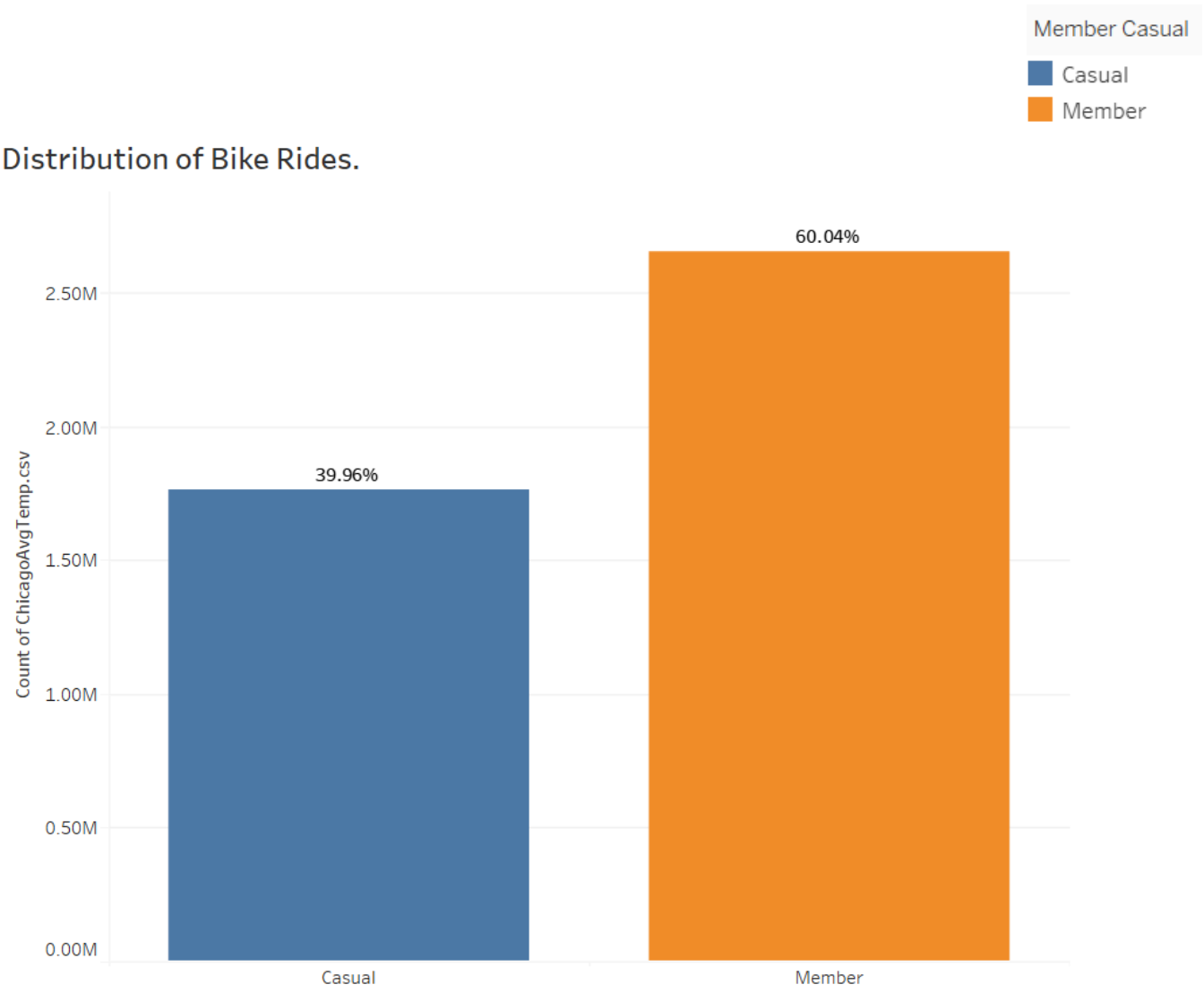
4.1 Start of the analysis.

Prior to analyzing the data and generating visualizations, I combined the two datasets, namely the cleaned_df generated earlier and the average temperature dataset downloaded from https://www.ncdc.noaa.gov/, using the inner join method in Tableau.

4.1.1 Checking how many rides each category of customer has.

To analyze the distribution of rides between casual and member riders in Tableau, I added the member and casual variables to the columns shelf, and the count of rides to the rows

shelf. Next, I added the member-casual variable to the color mark to filter each category by color and created a bar chart. I then added the count of rides variable to the label mark and created a quick calculation to present the percentage for each category. The resulting visualization is shown below:

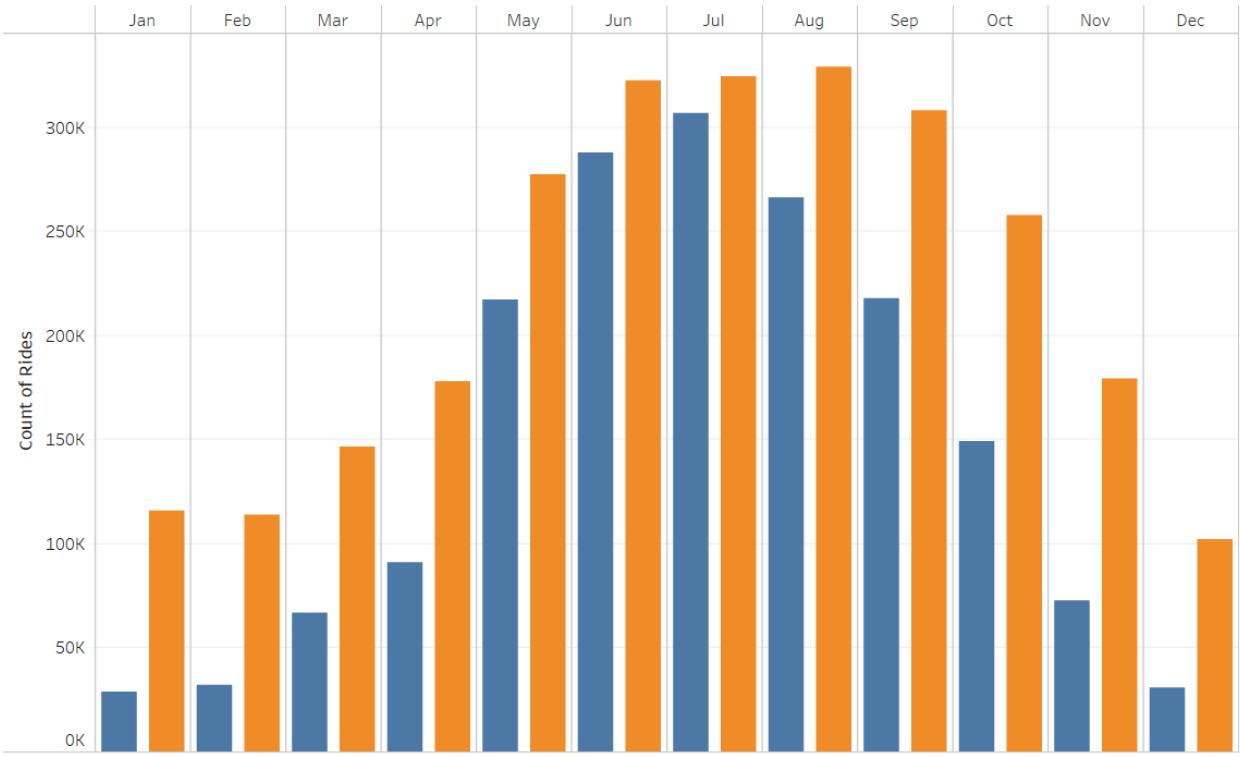


From the figure above we can see that almost 40% of rides are casual rides and almost 60% of rides are rides from members.

4.1.2 Monthly distribution of rides for each category of customer.

For this task, I followed the same methodology as above, but this time I added the month variable generated earlier in the process phase to the columns shelf as well. The other steps were the same as above, with the exception that I did not include the count of rides as a label to avoid cluttering the figure. The resulting visualization is shown below:

Monthly Distribution of Rides.



The figure above shows a significant increase in the number of rides during the summer months, particularly for casual rides. This suggests that there is an overall increase in rides during summer months, with a more pronounced effect on casual riders.

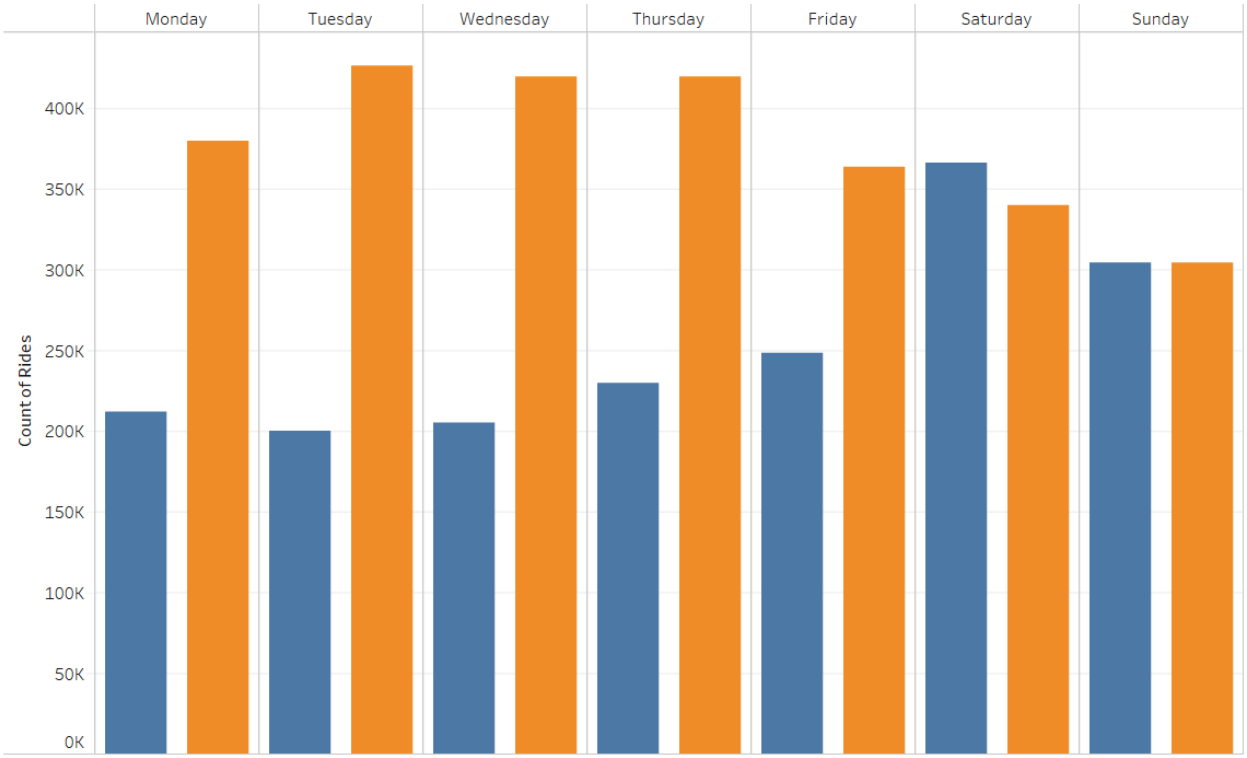
Insight

The increase in the number of rides during summer months, especially for casual riders, can be attributed to various factors such as favorable weather conditions, increased tourism, seasonal events and festivals, and a general rise in outdoor activities. Additionally, people may prefer to use bike-sharing systems as an alternative to cars during the summer months when traffic congestion is high and parking availability is limited.

4.1.3 Daily distribution of rides for each category of customer.

The methodology used to generate this barchart is similar to the previous plot, except that this time I used the weekday variable instead of the month variable to create the plot shown below.

Daily Distribution of Rides.



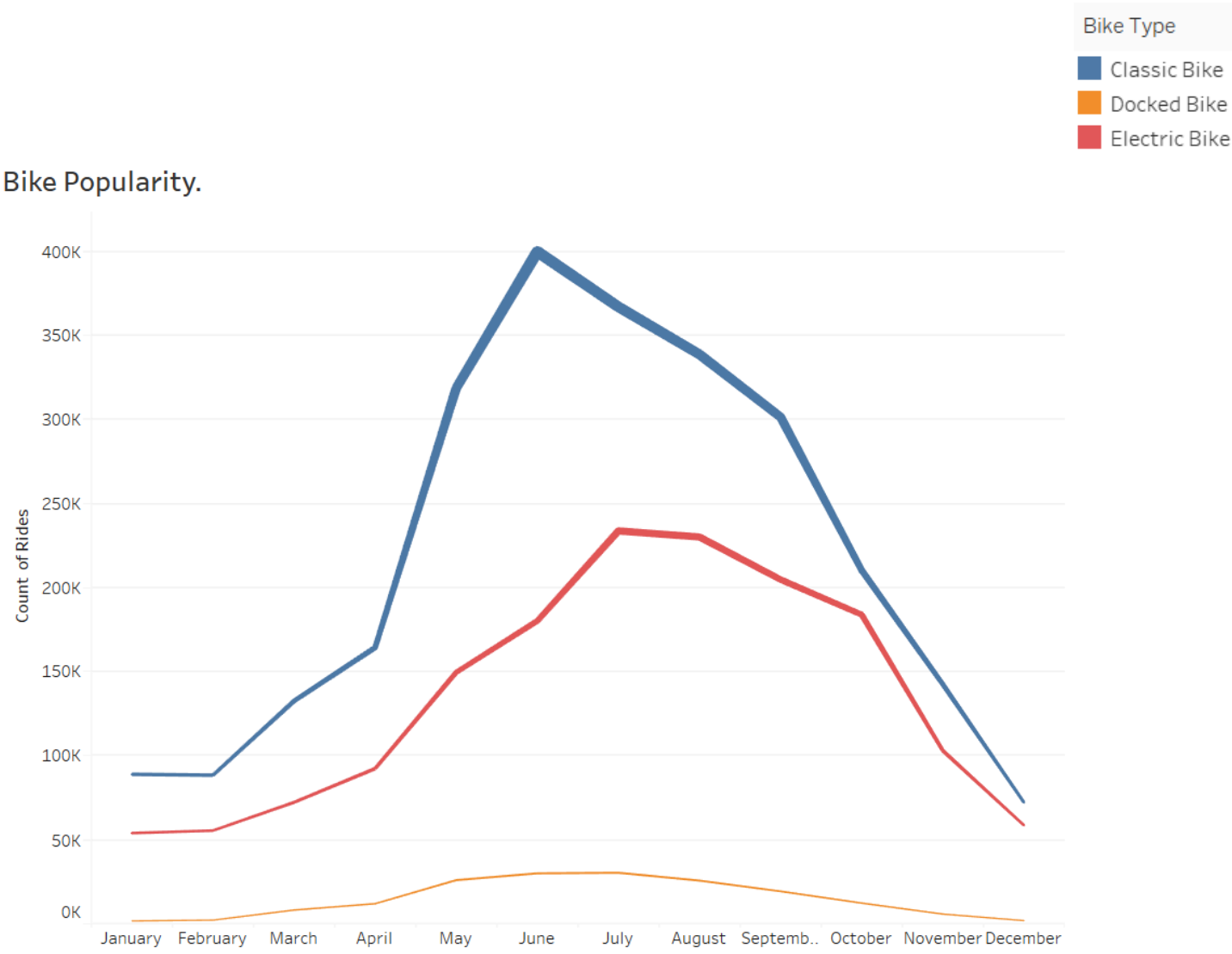
The figure shows that member rides are more frequent on weekdays, but decrease on weekends. In contrast, casual rides are less frequent on weekdays but increase on weekends, particularly on Saturdays.

Insight

The difference in the usage pattern between member and casual riders on weekdays and weekends can be attributed to the difference in their travel purposes. Member riders, who are likely to be regular commuters or people using the bike-sharing system for transportation to work or school, tend to use the system more frequently on weekdays. On the other hand, casual riders, who are more likely to use the system for leisure or tourism purposes, tend to use the system more on weekends when they have more free time. Additionally, weekend events and festivals may also contribute to the increase in casual rides on weekends, especially on Saturdays, when people may be more likely to engage in outdoor activities. Overall, the usage pattern of bike-sharing systems on weekdays and weekends reflects the different travel purposes and lifestyles of member and casual riders.

4.1.4 Checking which bike type is the most popular.

To generate a line chart that illustrates the variation in popularity of different bike types for each ride, I followed a similar methodology as before. I added the month variable to the columns, the count of rides variable to the rows, and the rideable type variable to the color mark to distinguish each bike type by color. In this case, classic bikes are represented by blue, docked bikes by orange, and electric bikes by red. Additionally, I included the count of rides variable in the size mark to depict the difference in the number of bikes taken for each bike type. The resulting figure is presented below:



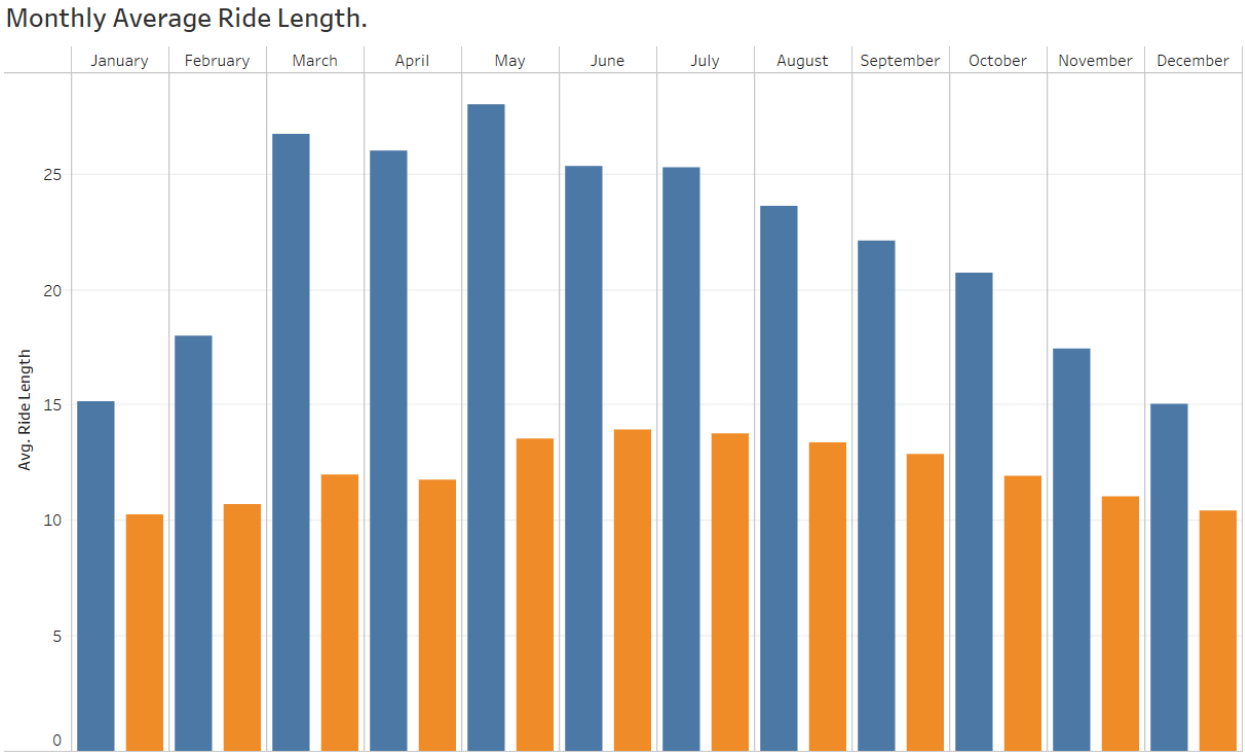
The line chart above indicates that classic bikes are the most commonly used, followed by electric bikes, with docked bikes being the least popular by a considerable margin. Additionally, the chart illustrates a significant increase in bike rides during the summer months.

Insight

One possible reason for classic bikes being the most popular could be that they are the most widely available and accessible option for riders. Additionally, classic bikes may be seen as a more familiar and traditional option for riders who are used to riding a standard bicycle. The popularity of electric bikes could be due to their convenience and ease of use, particularly for longer rides or hilly terrain. Finally, docked bikes may be less popular due to their restricted availability and the requirement to return them to specific docking stations, which could be less convenient for riders.

4.1.5 Monthly average ride length for each category of customer.

To fulfill this task, I opted to generate a barchart using Tableau. To accomplish this, I positioned the month and member casual variables on the columns, the average ride length calculation obtained from the previously calculated ride length variable on the rows, and added the member casual variable to the color mark to differentiate the two categories of customers by color. The result i came up is this:



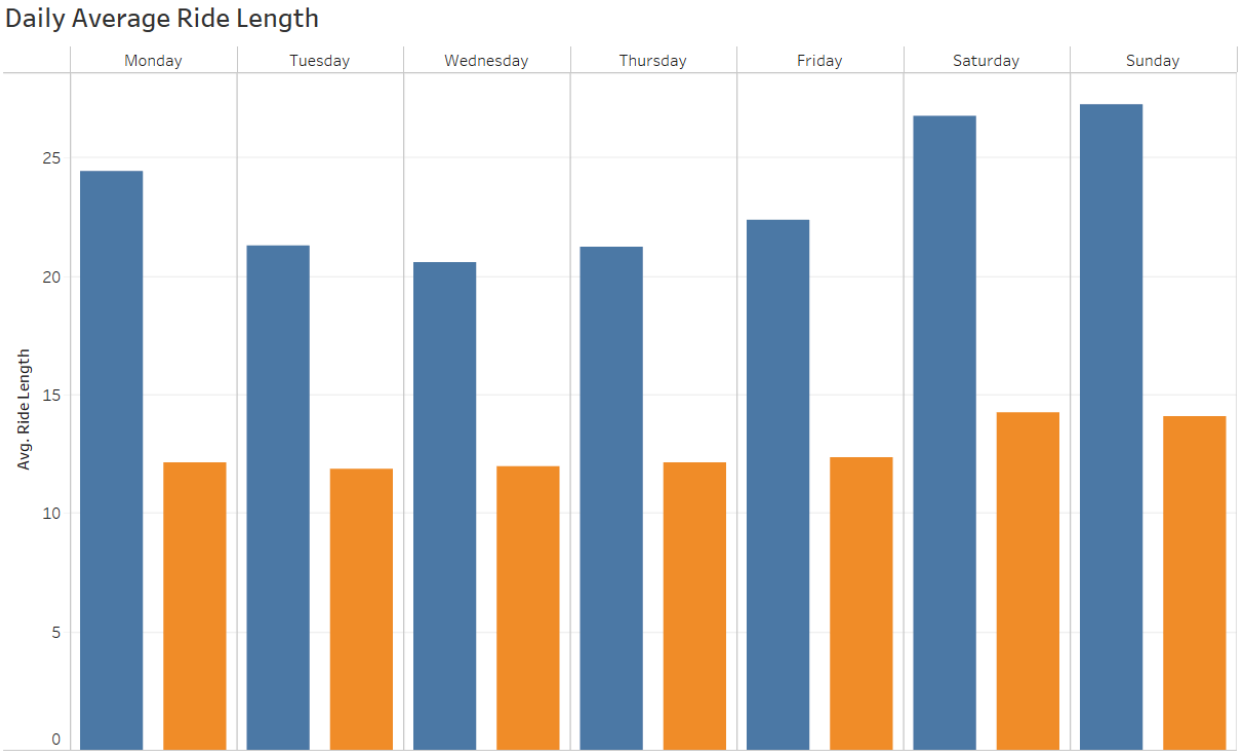
The figure clearly shows that casual rides have a longer average ride length in minutes than member rides, which is consistent throughout the year. The chart also demonstrates that the average time spent riding a bike varies more for casual rides than for member rides throughout the year. Notably, in the summer and spring months, there is a sharp increase in the average time spent riding a bike for casual riders, whereas the increase for member rides is not as noticeable.

Insight

There could be several reasons why casual riders tend to ride for longer periods of time compared to members. Casual riders may be more likely to use the bikes for leisurely activities such as sightseeing, while members may be more likely to use the bikes for commuting or short trips. Additionally, members may be more familiar with the bike system and therefore more efficient in their usage, while casual riders may need more time to get comfortable with the bikes. The increase in average ride time during summer and spring months for casual riders may be due to more favorable weather conditions, allowing for longer rides.

4.1.6 Daily average ride length for each category of customer.

To create the barchart below I used the same steps as for the last chart I generated, only that this time instead of placing the month on the columns, I placed the weekday. The result i came up with is this:



The above chart supports the previous assumption that casual riders typically spend more time riding than members. Additionally, it illustrates that casual riders have greater variability in their average time spent riding than members. The chart also reveals that there is an increase in the average ride time for casual riders on weekends.

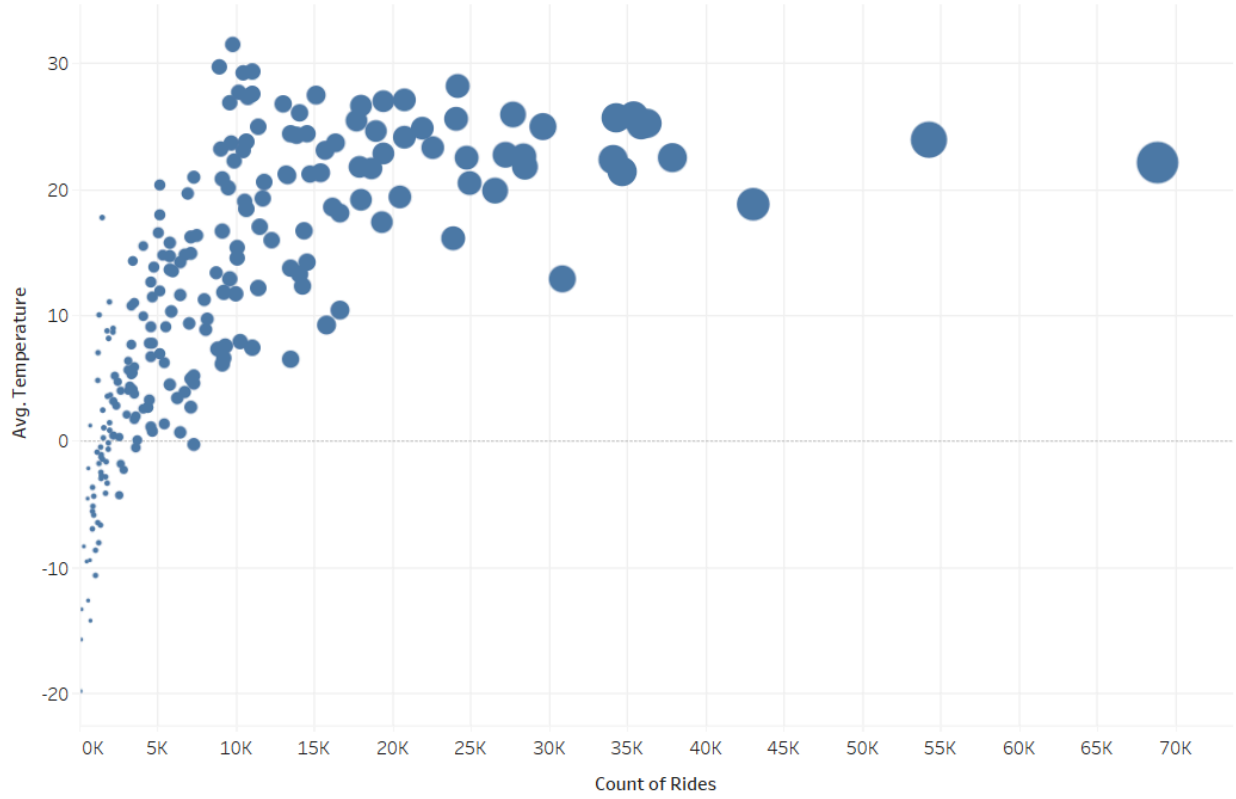
Insight

The increase in average ride time for casual riders on weekends could be due to factors such as fewer time constraints or more free time available for leisure activities.

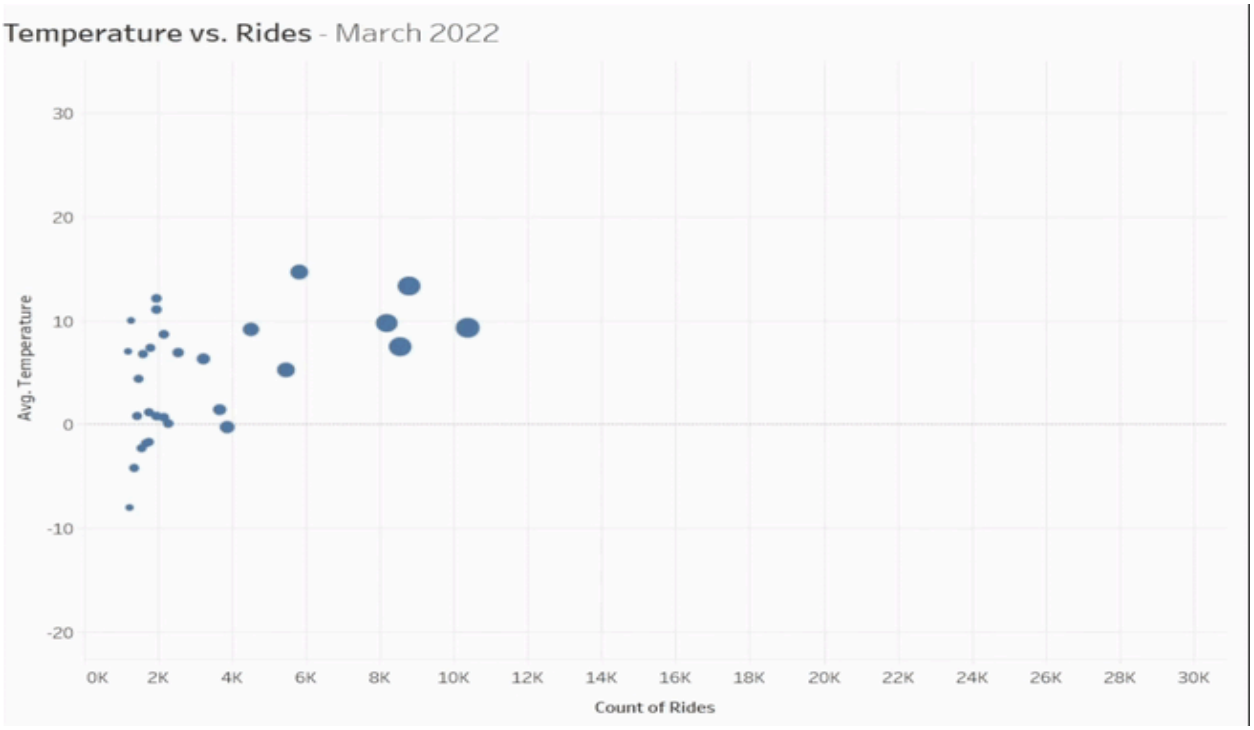
4.1.7 How does temperature affect the number of rides?

To address the previous question, I generated a scatterplot. To create the scatterplot, I placed the Count of Rides variable on the X-axis and the TAVG (Average temperature) variable on the Y-axis. Additionally, I used the count of rides variable for the size mark to indicate that there is a greater likelihood of more rides occurring when the temperature is hotter. The result i came up with is this:

Temperature vs Rides.



The scatterplot above shows a clear correlation between hotter weather and an increase in the number of rides. Additionally, I created an animated version of the scatterplot that illustrates the transition of the number of rides taken, from the cold months to summer and back to winter. The animated version looks like this:



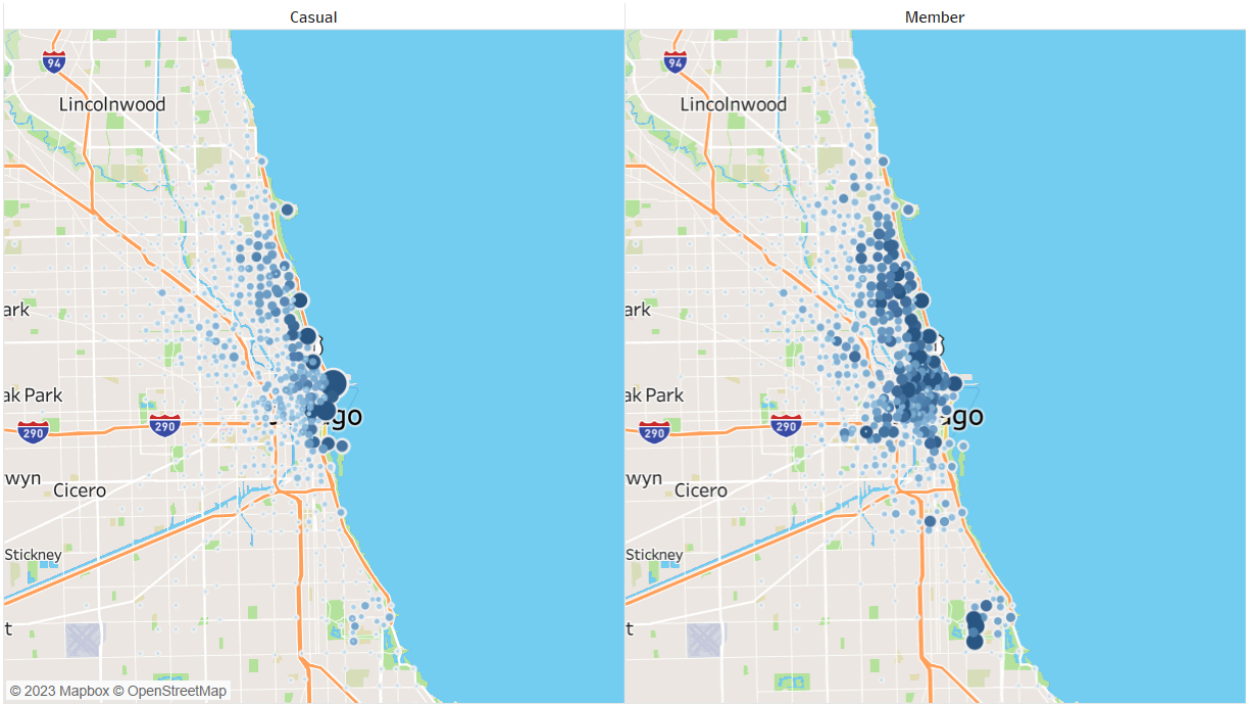
Insight

One possible reason for the correlation between hotter weather and an increase in the number of rides is that people may be more inclined to ride bikes when the weather is nice and sunny. Warmer temperatures may also make bike riding a more pleasant and enjoyable experience.

4.1.8 Geographical depiction of the distribution of the count of rides.

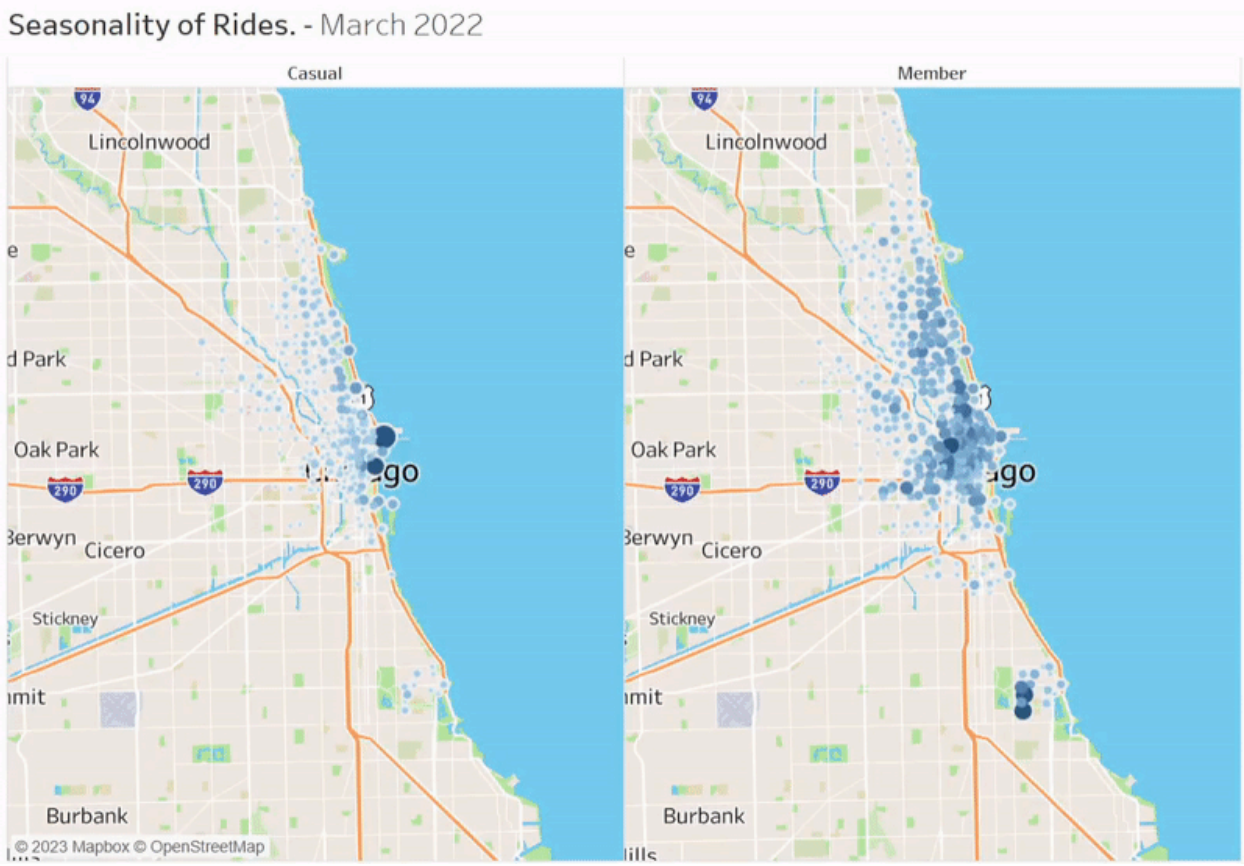
To create this map plot, I placed the start longitude variable in the columns, the member casual variable in the columns, and the start latitude variable in the rows as well. Additionally, I placed the count of rides variable in the size and color mark to distinguish areas where more rides were taking place. The result i came up with is this:

Geographic Distribution of Rides.



The map chart indicates that casual rides tend to occur more frequently near the lakeside, while member rides are more evenly distributed throughout the city.

I also created an animated version of the map plot which depicts how rides increase in the summer. Here is the animated version:



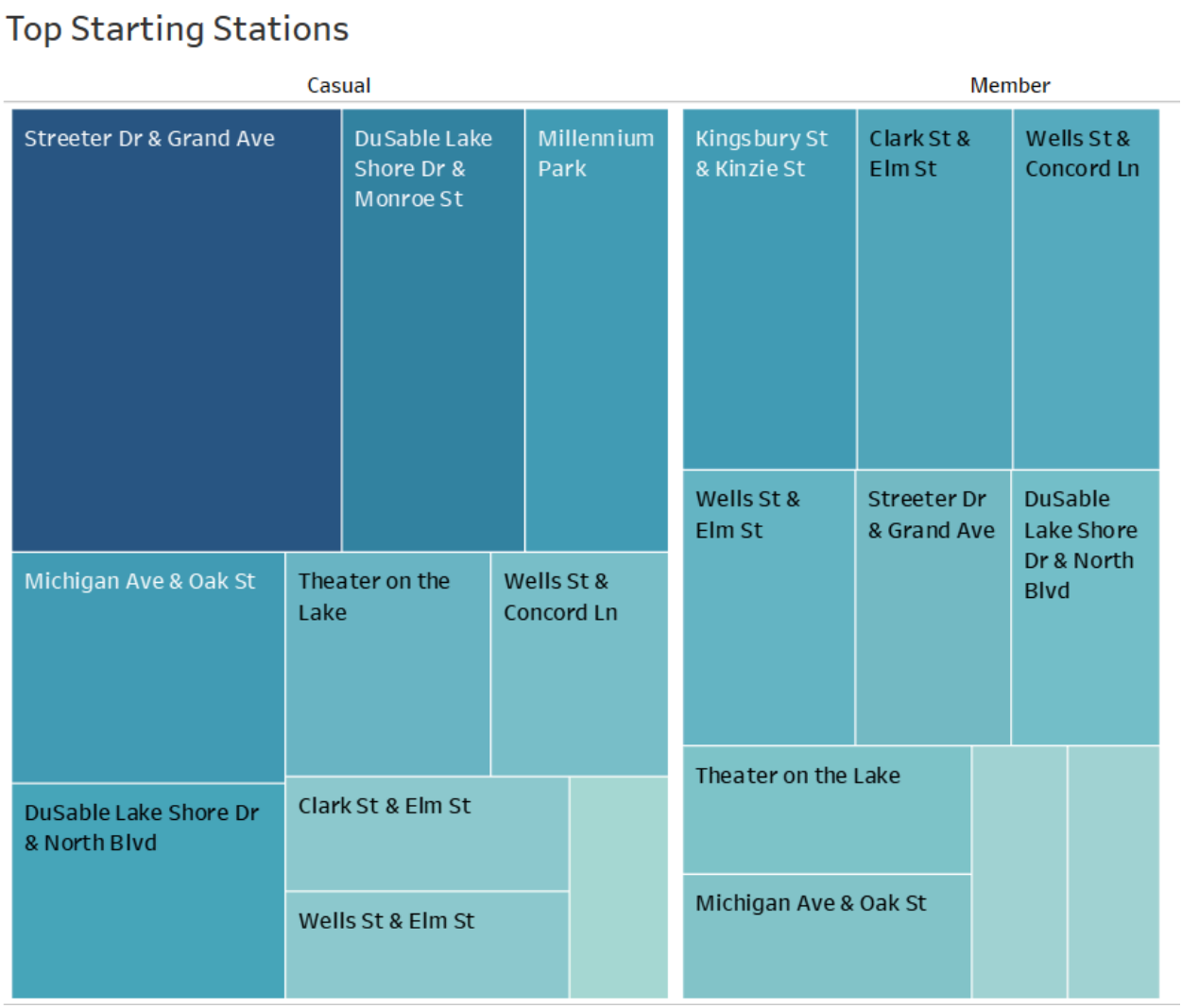
Here we can clearly see that our previous assumption that the count of rides increases in the summer more steeply for the casual rides.

Insight

There could be various reasons why casual rides are more likely to take place near the lakeside. One possibility is that the lakeside area may have more tourist attractions or scenic routes, which may attract casual riders who are looking to explore and enjoy the city's sights.

4.1.9 The most popular stations for each category of customer.

To visualize the most popular stations, I created a treemap. I placed the member and casual variable on the columns, the count of rides variable on the size and color marks, and the start station name variable on labels. Additionally, I added the start station name variable to filters to show only the 10 most frequently used bike stations based on the number of rides for each station. The result i came up with is this:



Insight

The marketing team could use the information about the top starting stations for bike rides to design targeted marketing campaigns that focus on those locations. They could consider setting up promotional events or partnerships with businesses located near those stations, offering special discounts or rewards to customers who use the bikes from those stations, or creating advertising campaigns that feature the most popular stations. Additionally, the marketing team could use the data to identify patterns in usage and develop strategies to encourage more people to use the bikes from those stations, such as offering incentives for frequent riders or creating more bike-friendly infrastructure in the surrounding areas.

4.2 Summarizing the analysis

The study utilized two datasets, namely the cleaned dataset and the average temperature data, to examine the distribution of bike rides between member and casual riders. The analysis involved creating visualizations using Tableau. The findings revealed that 60% of rides were from members, while 40% were from casual riders. The number of bike rides increased during the summer months, particularly for casual riders. On weekdays, member

rides were more frequent but decreased on weekends, whereas casual rides were less frequent on weekdays but increased on weekends, particularly on Saturdays. Classic bikes were the most commonly used, followed by electric bikes, and docked bikes were the least popular. The average ride length for casual rides was longer than that of member rides throughout the year. In the summer and spring months, there was a significant increase in the average time spent riding a bike for casual riders. Additionally, it was observed that casual rides tended to occur more frequently on the lakeside, while member rides were distributed throughout the city.

5 Phase 6: Act

5.1 Summarizing the analysis

The analysis of the combined data sets shows various findings about bike rides and riders. Firstly, during summer months, bike rides increase, especially for casual riders. This could be due to favorable weather conditions, increased tourism, seasonal events, and the desire for outdoor activities. Secondly, the usage pattern between member and casual riders differs on weekdays and weekends due to their travel purposes. Member riders tend to use the bike-sharing system more frequently on weekdays as regular commuters or those who use it for transportation to work or school. Casual riders tend to use the system more on weekends for leisure or tourism purposes. Additionally, weekend events and festivals may contribute to the increase in casual rides on weekends, especially on Saturdays. Thirdly, classic bikes are the most popular option for riders, followed by electric bikes, while docked bikes are the least popular. The reasons for this could be the availability and accessibility of classic bikes, the convenience and ease of use of electric bikes, and the restricted availability of docked bikes. Fourthly, casual riders tend to ride for longer periods compared to members, especially during summer and spring months, and on weekends. This could be due to leisurely activities such as sightseeing or more free time available for leisure activities. Finally, the analysis shows that casual rides tend to take place more near the lakeside, possibly due to tourist attractions or scenic routes.

The findings from the analysis can help the marketing team design targeted campaigns that focus on the top starting stations, set up promotional events or partnerships with nearby businesses, and offer special discounts or rewards to customers who use the bikes from popular stations. Additionally, the team can use the data to identify patterns in usage and encourage more people to use the bikes from those stations by creating more bike-friendly infrastructure in the surrounding areas or offering incentives for frequent riders.

5.2 Top recommendations for the marketing team.

- Offer a promotional membership discount during the summer months when casual ridership is at its highest. This could incentivize casual riders to become annual members and continue to use the bike-sharing system throughout the year.
- Develop partnerships with local businesses near the most popular bike stations to offer exclusive discounts or rewards to annual members who use the bikes from those stations. This could create a sense of community and loyalty among annual

members and encourage casual riders to become members to take advantage of these benefits.

- Develop a targeted advertising campaign that focuses on the benefits of becoming an annual member, such as discounted rates or priority access to bikes during peak times. This could be particularly effective during the spring and fall months when casual ridership is still high, but the weather is more temperate and may encourage riders to commit to becoming annual members.
- Enhance the user experience for annual members by offering additional perks, such as access to premium bikes or priority parking at bike stations. This could create a sense of exclusivity and value for annual members and make the bike-sharing system more appealing to casual riders who are considering becoming members.

5.3 Conclusion

In conclusion, the analysis of the bike-sharing system data has revealed several key insights about the usage patterns and preferences of casual and member riders. These insights can be used by the marketing team to design targeted campaigns and strategies to increase the number of annual members and encourage more people to use the bike-sharing system.

Overall, the analysis found that member riders tend to use the bike-sharing system more frequently on weekdays for commuting or transportation purposes, while casual riders use the system more on weekends for leisure or tourism activities. Additionally, classic bikes are the most popular option, followed by electric bikes, and docked bikes are the least popular. Casual riders tend to ride for longer periods of time compared to members, particularly during summer months, and are more likely to ride near the lakeside. Finally, hotter weather and more favorable weather conditions are correlated with an increase in the number of rides.

Based on these insights, the marketing team could focus on several strategies to convert casual riders to annual members, including targeted marketing campaigns focused on popular starting stations, promotional events or partnerships with local businesses near those stations, and incentives for frequent riders. Additionally, the team could consider offering discounted annual memberships during peak tourism or outdoor activity seasons to encourage more people to become members.

Thank you for taking the time to read this project. If you have any questions or feedback, please do not hesitate to reach out.

To check the entire Tableau Dashboard Click Here. I have made this Vizz accessible to anyone.