Cyclistic

This is data analysis for *Cyclistic*, a bike-share company, in order to help them attract more riders while following the steps of the data analysis process: **Ask**, **Prepare**, **Process**, **Analyze**, **Share** and **Act**.

The *goal* of the analysis is to maximize the number of annual memberships. The analysis will be useful when designing marketing strategies aimed at converting casual riders into annual members.

The *question* assigned is how do casual riders and annual members use Cyclistic bikes differently?. The Cyclistic historical bike trip data is used to identify the trend.

Scope of Work

Phase	Key Tasks	Deli
Ask	• Identify the business task • Consider key stakeholders	A clear statement of the b
Prepare	• Download data and store it appropriately • Identify how it's organized • Sort and filter the data • Determine the credibility of the data	A description of all data s
Process	• Check the data for errors • Choose tools used • Transform the data • Document the cleaning process	Documentation of any cle
Analyze	• Aggregate the data so it's useful and accessible • Organize and format the data • Perform calculations • Identify trends and relationships	A summary of the analysi
Share	• Determine the best way to share the findings • Create effective data visualizations • Present your findings • Ensure the work is accessible	Supporting visualizations
Act	• Finalize the conclusion • Think of ways to incorporate the insights into business decisions	Top three recommendatic

Business Background

Cyclistic: a bike-share programs with 5,824 bikes and 692 docking stations

- about 8% of riders use assistive bikes
- about 30% use the bokes to commute to work each day
- pricing plans:
    - casual riders
        - single-ride
        - full-day pass
    - Cyclistic members
        - annual membership

Lily Moreno: the directer of marketing and your manager.

- responsible for the development of campaigns and initiatives to promote Cyslistic
- including email, social media, etc

Cyclistsic marketing analytics team: a team of data analysts responsible for collecting, analyzing, and reporting data.

Cyclistic executive team: decide whether to approve the recced marketing program.

Business task

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

Sources of Data

I have used data from January 2023 to December 2023 to ensure reliability and credibility of the data. The data has been made available by Motivate International Inc. under this license.

Credibility of the data

To ensure the reliability and quality of the dataset, we evaluate it based on the ROCCC criteria: Reliable, Original, Comprehensive, Current, and Cited.

Criteria (score out of 10.0)	
Reliable (9.5)	The dataset is sourced from Lyft Bikes and Scooters, LLC, operating the Divvy bicycle sharing service in
Original (10.0)	The dataset's origin is traced back to the Divvy service operated by Lyft Bikes and Scooters, LLC. The lic
Comprehensive (9.0)	Since data-privacy issues prohibit from using riders' personally identifiable information, the data's comp
Current (8.5)	The dataset spans a few previous years, aligning with the analysis request. However, additional clarity c
Cited (9.5)	The data is made available by Motivate International Inc. under the license agreement. The agreement c

Cleaning Data

Preparing for manipulation of data

```
# load libraries required to perform analysis
library(ggplot2)
library(dplyr)
library(janitor)
print('Libraries successfully loaded.')

# import our data
jan <- read.csv("/kaggle/input/cyclistic-raw/202301.csv")
feb <- read.csv("/kaggle/input/cyclistic-raw/202302.csv")
mar <- read.csv("/kaggle/input/cyclistic-raw/202303.csv")
apr <- read.csv("/kaggle/input/cyclistic-raw/202304.csv")
may <- read.csv("/kaggle/input/cyclistic-raw/202305.csv")
jun <- read.csv("/kaggle/input/cyclistic-raw/202306.csv")
jul <- read.csv("/kaggle/input/cyclistic-raw/202307.csv")
aug <- read.csv("/kaggle/input/cyclistic-raw/202308.csv")
sep <- read.csv("/kaggle/input/cyclistic-raw/202309.csv")
oct <- read.csv("/kaggle/input/cyclistic-raw/202310.csv")
nov <- read.csv("/kaggle/input/cyclistic-raw/202311.csv")
dec <- read.csv("/kaggle/input/cyclistic-raw/202312.csv")
print('Data has been successfully loaded into data frames.')
```

```
[1] "Libraries successfully loaded."
[1] "Data has been successfully loaded into data frames."
```

Inspecting Data

```
# check if all of the datasets contain the same number, types and names of columns
# TRUE means all the datasets share the same number, types and names of columns
# FALSE means otherwise
compare_df_cols_same(jan, feb, mar, apr, may, jun, jul, aug, sep, oct, nov, dec)

# print column names
colnames(jan)

# merge all data frames
trips2023 <- rbind(jan, feb, mar, apr, may, jun, jul, aug, sep, oct, nov, dec)
print('Data has been successfully merged into one data frame.')

# inspect the merged data
# display internal structure of the object
str(trips2023)
```

```
TRUE
'ride_id' · 'rideable_type' · 'started_at' · 'ended_at' · 'start_station_name' · 'start_station_id' ·
'end_station_name' · 'end_station_id' · 'start_lat' · 'start_lng' · 'end_lat' · 'end_lng' · 'member_casual'
[1] "Data has been successfully merged into one data frame."
'data.frame':	5719877 obs. of  13 variables:
 $ ride_id           : chr  "F96D5A74A3E41399" "13CB7EB698CEDB88" "BD88A2E670661CE5" "C9
 $ rideable_type     : chr  "electric_bike" "classic_bike" "electric_bike" "classic_bike
 $ started_at        : chr  "2023-01-21 20:05:42" "2023-01-10 15:37:36" "2023-01-02 07:5
 $ ended_at          : chr  "2023-01-21 20:16:33" "2023-01-10 15:46:05" "2023-01-02 08:0
 $ start_station_name: chr  "Lincoln Ave & Fullerton Ave" "Kimbark Ave & 53rd St" "Weste
 $ start_station_id  : chr  "TA1309000058" "TA1309000037" "RP-005" "TA1309000037" ...
 $ end_station_name  : chr  "Hampden Ct & Diversey Ave" "Greenwood Ave & 47th St" "Valli
 $ end_station_id    : chr  "202480.0" "TA1308000002" "599" "TA1308000002" ...
 $ start_lat         : num  41.9 41.8 42 41.8 41.8 ...
 $ start_lng         : num  -87.6 -87.6 -87.7 -87.6 -87.6 ...
 $ end_lat           : num  41.9 41.8 42 41.8 41.8 ...
 $ end_lng           : num  -87.6 -87.6 -87.7 -87.6 -87.6 ...
```

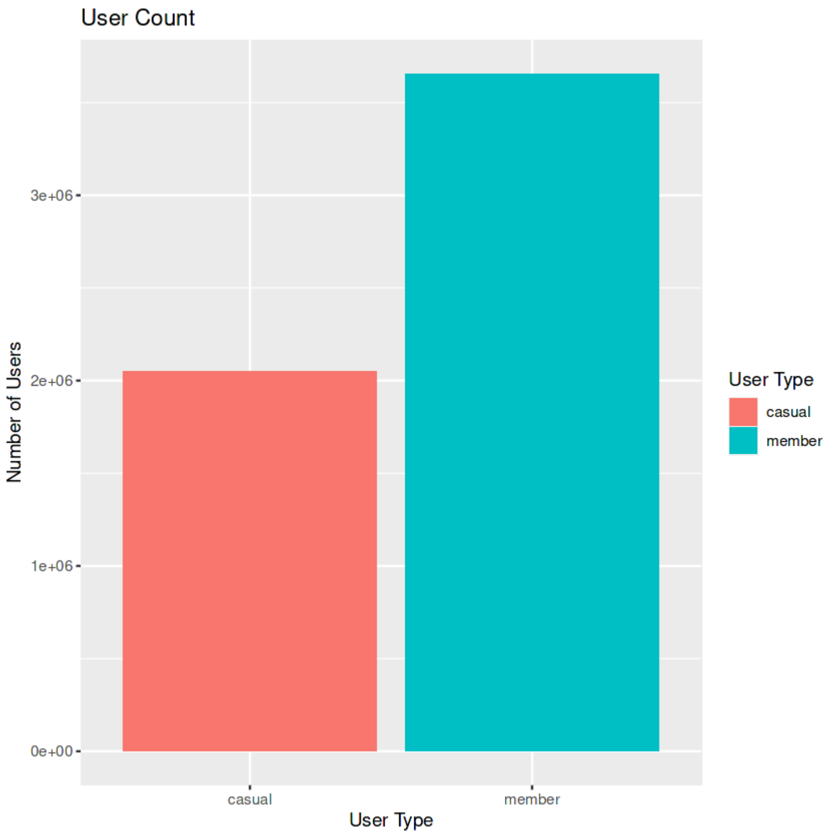
Manipulating Data

```
# delete 'start_station_name', 'start_station_id', 'end_station_name' and 'end_station_id'
# since they won't be necessary for our purposes
#trips2023 <- subset(trips2023, select = -c(start_station_name, start_station_id, end_static
print('Unnecessary columns deleted successfully.')

# delete null values
null_val_one <- capture.output(sum(is.na(trips2023$end_lat)))
sprintf("There are %s null values in 'end_lat'", null_val_one)
null_val_two <- capture.output(sum(is.na(trips2023$end_lng)))
sprintf("There are %s null values in 'end_lng'", null_val_two)
trips2023 <- trips2023[!(is.na(trips2023$end_lat) | is.na(trips2023$end_lng)), ]
print('Data frame refined by deleting null values.')

# divide datetime column
trips2023$started_at = strptime(trips2023$started_at, "%Y-%m-%d %H:%M:%S")
trips2023$ended_at = strptime(trips2023$ended_at, "%Y-%m-%d %H:%M:%S")
print("Converted 'started_at' and 'ended_at' to the datetime type.")
trips2023$date <- as.Date(trips2023$started_at)
trips2023$month <- format(as.Date(trips2023$date), "%m")
trips2023$day <- format(as.Date(trips2023$date), "%d")
trips2023$day_of_week <- format(as.Date(trips2023$date), "%A")
print("Created new columns 'date', 'month', 'day' and 'day_of_week' out of 'started_at'.")
trips2023$ride_length <- difftime(trips2023$ended_at, trips2023$started_at)
print('Calculated ride_length for each ride.')

# delete negative ride length
trips2023 <- trips2023 %>%
  filter(ride_length > 0)
print('Successfully deleted negative ride duration.')
```

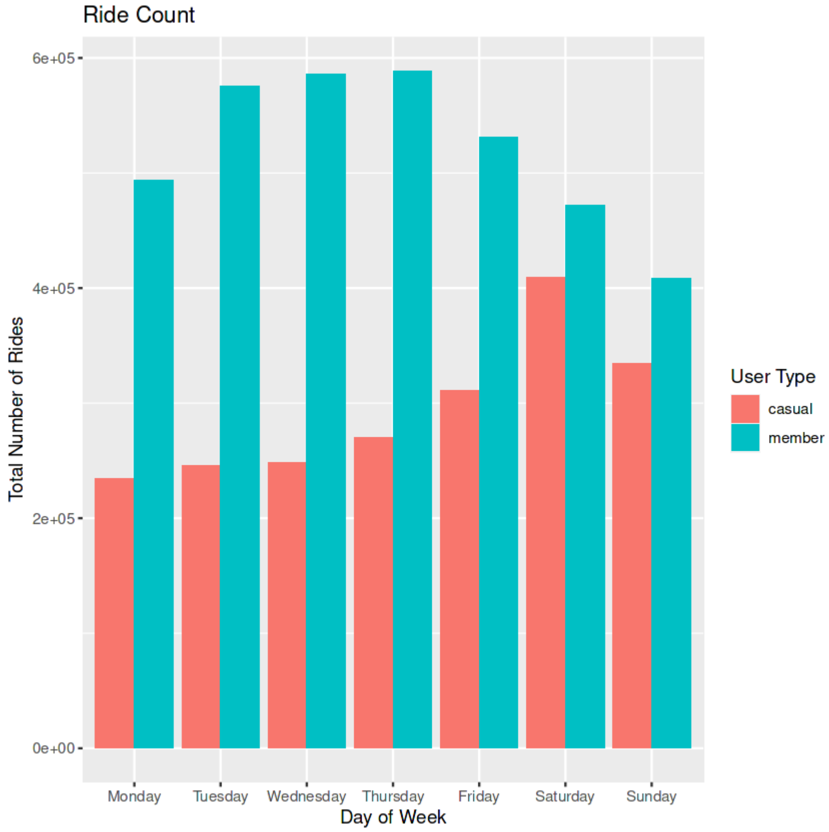
```
[1] "Unnecessary columns deleted successfully."
'There are [1] 6990 null values in \'end_lat\"
'There are [1] 6990 null values in \'end_lng\"
[1] "Data frame refined by deleting null values."
[1] "Converted 'started_at' and 'ended_at' to the datetime type."
[1] "Created new columns 'date', 'month', 'day' and 'day_of_week' out of 'started_at'."
[1] "Calculated ride_length for each ride."
[1] "Successfully deleted negative ride duration."
```

Analyzing data

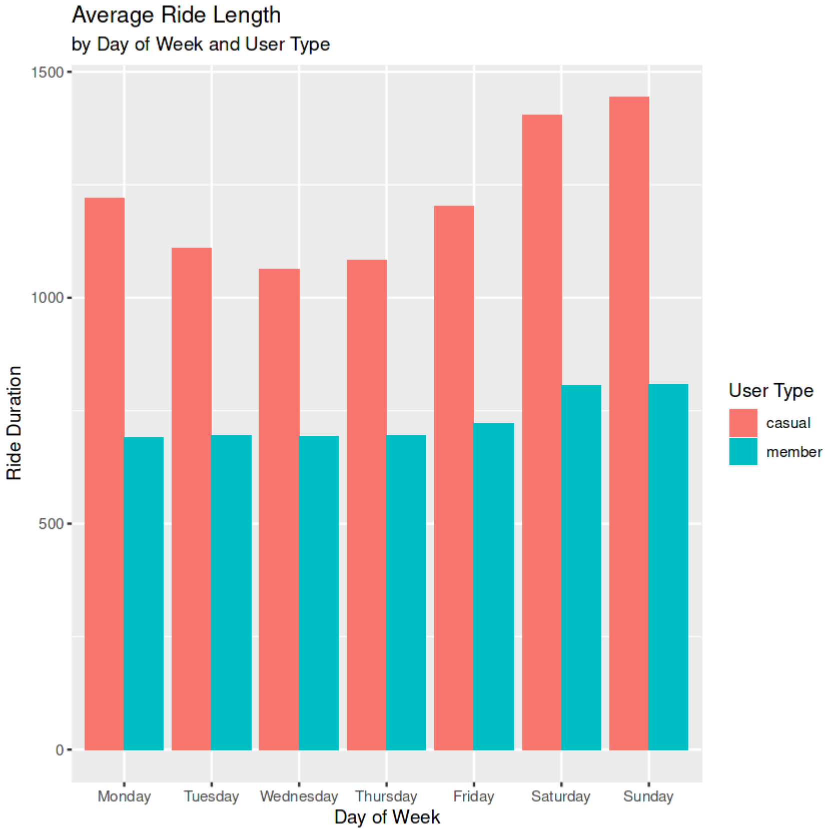
```
# user count
ggplot(trips2023, aes(member_casual, fill = member_casual)) + geom_bar() + labs(title = "Use
```



```
# total number of rides by day and user type
ggplot(trips2023, aes(factor(day_of_week, levels = c('Monday', 'Tuesday', 'Wednesday', 'Thur
    labs(title = 'Ride Count', x = 'Day of Week', y =  'Total Number of Rides', fill = 'User
```



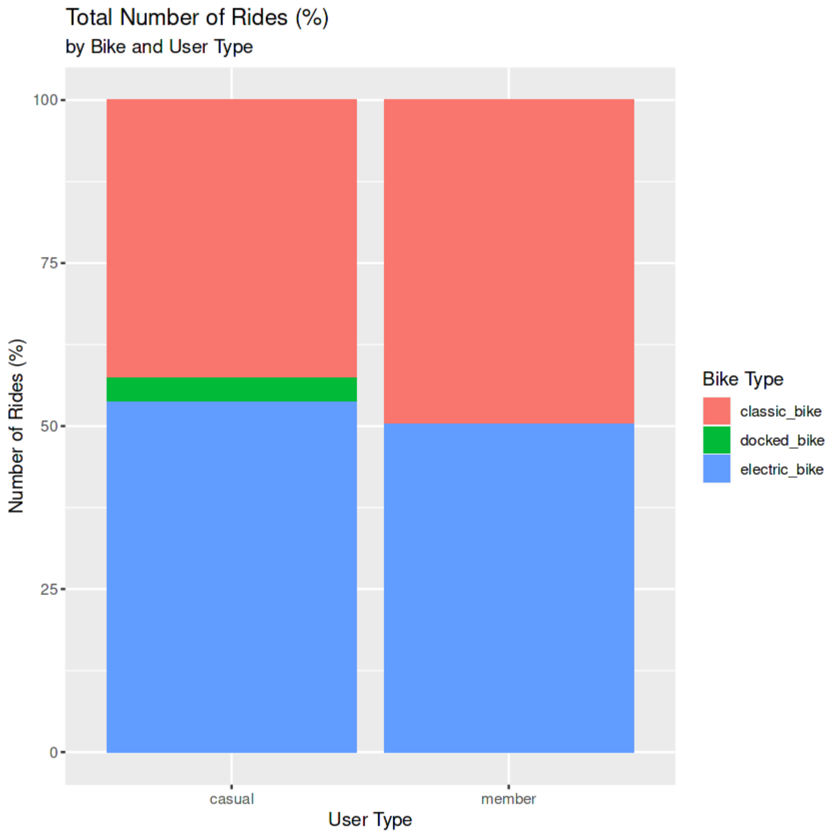
```
# average ride length by day and user type
summ1 <- trips2023 %>%
  group_by(member_casual, day_of_week) %>%
  summarize(average_ride_length = mean(ride_length, na.rm = TRUE))

ggplot(summ1, aes(factor(day_of_week, levels = c('Monday', 'Tuesday', 'Wednesday', 'Thursday
    labs(title = 'Average Ride Length', subtitle = 'by Day of Week and User Type', x = 'Day
`summarise()` has grouped output by 'member_casual'. You can override using the
`.groups` argument.
Don't know how to automatically pick scale for object of type <difftime>.
Defaulting to continuous.
```



Average Ride Length

by Day of Week and User Type

```
# total number of rides by bike and user type
summ2 <- trips2023 %>%
  group_by(member_casual, rideable_type) %>%
  summarize(total_num = n()) %>%
  mutate(percentage = total_num * 100 / sum(total_num))

ggplot(summ2, aes(member_casual, percentage, fill = rideable_type)) + geom_bar(stat = 'ident
    labs(title = 'Total Number of Rides (%)', subtitle = 'by Bike and User Type', x = 'User
     `summarise()` has grouped output by 'member_casual'. You can override using the
     `.groups` argument.
```



Tableau's version of the graphs can be found here.

Key Findings

User Demographics

- Annual membership surpasses casual rider numbers.
- Annual members exhibit a weekday-centric usage pattern, contrasting with casual riders who show heightened weekend activity.
- Casual riders tend to engage in longer-duration trips, particularly noticeable during weekends.
- Both user segments demonstrate extended trip durations during Summer, juxtaposed with shorter durations during Winter.

Geographical Patterns

- Casual riders frequently opt for routes along Lake Michigan's coastline during weekends, indicative of leisure-oriented riding.
- In contrast, members predominantly utilize bikes for weekday commuting in urban areas.

These findings underscore the nuanced usage disparities between annual members and casual riders, alongside discernible geographical preferences. This empirical understanding serves as a