Cyclistic Bike-Share

Barış Hocaoğlu

2024-08-30

Introduction

In this report, I'll guide you through the processes of importing, cleaning, transforming, and visualizing data. I'll also provide insights and address any questions regarding annual members and casual riders. The data used comes from the Google Data Analytics Professional Certificate program and pertains to a fictional bike rental company offering annual memberships. Let's start by loading the required packages.

```
library('tidyverse') # Helps to transform and better present data

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.
0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.5.1     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.1
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflict
s() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors

library('conflicted') # To use filter() from dplyr package
library('scales') # Provides the internal scaling infrastructure used by ggpl
ot2
library('patchwork') # Designed to combine plots
```

Data Collection

After loading necessary packages, we import our data.

- cyclistic_2023_01 <- read_csv('cyclistic_2023_01.csv')
- cyclistic_2023_02 <- read_csv('cyclistic_2023_02.csv')
- cyclistic_2023_03 <- read_csv('cyclistic_2023_03.csv')
- cyclistic_2023_04 <- read_csv('cyclistic_2023_04.csv')
- cyclistic_2023_05 <- read_csv('cyclistic_2023_05.csv')
- cyclistic_2023_06 <- read_csv('cyclistic_2023_06.csv')
- cyclistic_2024_01 <- read_csv('cyclistic_2024_01.csv')
- cyclistic_2024_02 <- read_csv('cyclistic_2024_02.csv')

- cyclistic_2024_03 <- read_csv('cyclistic_2024_03.csv')
- cyclistic_2024_04 <- read_csv('cyclistic_2024_04.csv')
- cyclistic_2024_05 <- read_csv('cyclistic_2024_05.csv')
- cyclistic_2024_06 <- read_csv('cyclistic_2024_06.csv')

There are twelve csv files that are consist of first six months of 2023 and 2024. I could have added them into only one file using excel but I preferred doing it with bind_rows() function.

```
all_rides <- bind_rows(cyclistic_2023_01, cyclistic_2023_02, cyclistic_2023_0
3,
                       cyclistic_2023_04, cyclistic_2023_05, cyclistic_2023_0
6,
                       cyclistic_2024_01, cyclistic_2024_02, cyclistic_2024_0
3,
                       cyclistic_2024_04, cyclistic_2024_05, cyclistic_2024_0
6)
```

Now we have our data in our hands to be looked, cleaned and visualized. **Our goal with this data is to compare member riders and casual riders, examining whether the total numbers have increased or decreased in the first half of 2023 and 2024. We will also analyze popular stations among riders and identify the most frequently used types of bikes.**

First Look At Data

```
dim(all_rides)

## [1] 4795422      13
```

There are 4795422 rows and 13 columns. Let's take a look at them.

```
colnames(all_rides)

##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

We don't need "start_lat","start_lng", "end_lat", "end_lng" since they are in no use to us.

```
all_rides <- select(all_rides, -c(start_lat, start_lng, end_lat, end_lng))

head(all_rides)

## # A tibble: 6 × 9
##   ride_id          rideable_type started_at          ended_at
##   <chr>            <chr>         <dttm>              <dttm>
## 1 F96D5A74A3E41399 electric_bike 2023-01-21 20:05:42 2023-01-21 20:16:33
## 2 13CB7EB698CEDB88 classic_bike  2023-01-10 15:37:36 2023-01-10 15:46:05
## 3 BD88A2E670661CE5 electric_bike 2023-01-02 07:51:57 2023-01-02 08:05:11
```

```
## 4 C90792D034FED968 classic_bike  2023-01-22 10:52:58 2023-01-22 11:01:44
## 5 3397017529188E8A classic_bike  2023-01-12 13:58:01 2023-01-12 14:13:20
## 6 58E68156DAE3E311 electric_bike 2023-01-31 07:18:03 2023-01-31 07:21:16
## # ℹ 5 more variables: start_station_name <chr>, start_station_id <chr>,
## #   end_station_name <chr>, end_station_id <chr>, member_casual <chr>

#More detailed view
glimpse(all_rides)
summary(all_rides)
str(all_rides)

# Checking null values
colSums(is.na(all_rides))

# Checking for duplicates
all_rides %>%
distinct(ride_id) %>%
nrow() # 479521 rows. But our total was 4795422. So there are 211 duplicate v
alues

# Shows the duplicate values
view(all_rides %>%
group_by(ride_id) %>%
filter(n() > 1))
```

This data needs to be modified to be more representable. Here is the list that we need to do.

To Do:

- Eliminate duplicate entries from the dataset.
- Add columns for date, month, day, and year to make data aggregation easier.
- Introduce a column for ride_length to calculate the duration of each ride in the all_rides dataset.
- Address the issue of negative values in the ride_length column. These negative values may result from data errors or quality control activities where bikes are taken from service. It's best to remove these rows to keep the data accurate.

Data Cleaning, Manipulation, Transformation

```
# 1 211 Duplicates removes
all_rides <- all_rides %>%
distinct(ride_id, .keep_all = TRUE)

# 2 Adding columns such as date, year, month, day, day_of_week
all_rides_2 <- all_rides %>%  mutate(date = as.Date(started_at))
all_rides_2 <- all_rides_2 %>% mutate(year = format(as.Date(date), "%Y"))
all_rides_2 <- all_rides_2 %>% mutate(month = format(as.Date(date), "%m"))
all_rides_2 <- all_rides_2 %>% mutate(day = format(as.Date(date), "%d"))
all_rides_2 <- all_rides_2 %>% mutate(day_of_week = format(as.Date(date), "%A
"))
```

```r
# 3 Calculating the trip duration in seconds
all_rides_2 <- all_rides_2 %>%
mutate(ride_length = difftime(ended_at, started_at))

#Converting ride_length from difftime to numeric value type to perform calcul
ations
all_rides_2$ride_length <- as.numeric(all_rides_2$ride_length)

# 4 Removing the bad data. 206 data to be removed.
bad_data <- all_rides_2 %>%
filter(ride_length < 0)

bad_data %>% group_by(rideable_type) %>%
summarise(count_rideable = n())

## # A tibble: 2 × 2
##   rideable_type count_rideable
##   <chr>                  <int>
## 1 classic_bike               9
## 2 electric_bike            197
```

We need to inform the company that most of the bad data caused by electric bikes.

```r
# Bad data removed
all_rides_2 <- all_rides_2 %>%
filter(!(ride_length < 0))
```

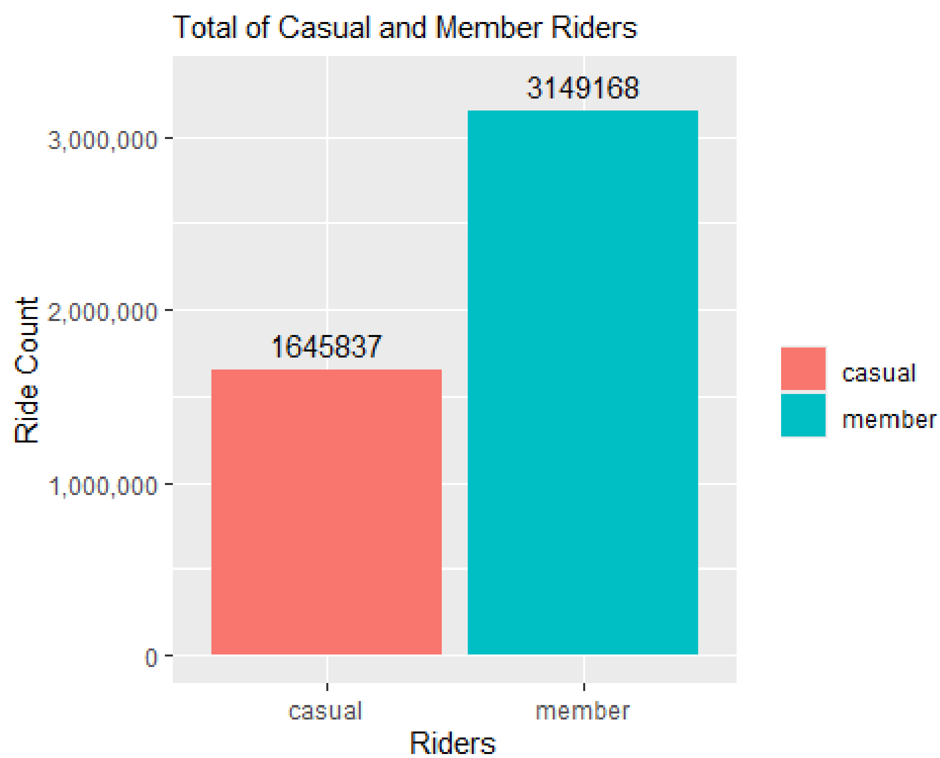
Now that the data is cleaned, we will move on to extracting meaningful insights and visualizing the results. I've included detailed code to illustrate how the charts were created. While I could have simply presented the charts, I wanted to provide the code for those who may not be familiar with the original R script.

Chart 1: Total Of Casual And Member Riders

Below code chunks calculate how much of the rides are casual or member.

```r
member_casual_count <- all_rides_2 %>%
group_by(member_casual) %>%
summarise(person_count = n())
member_casual_count

## # A tibble: 2 × 2
##   member_casual person_count
##   <chr>                <int>
## 1 casual             1645837
## 2 member             3149168

total_casual_rides <- member_casual_count %>%
 filter(member_casual == "casual") %>% pull(person_count)
total_member_rides <- member_casual_count %>%
  filter(member_casual == "member") %>% pull(person_count)
```

```
member_casual_count %>%
ggplot(aes(x = member_casual, y = person_count, fill = member_casual)) +
geom_bar(stat = "identity") +
labs(Title ="Total Rides",
subtitle="Total of Casual and Member Riders",x="Riders",y="Ride Count",fill="
")+
scale_y_continuous(labels = comma) +
annotate("text", x=1, y=1800000, label=total_casual_rides, size=4)+
annotate("text", x=2, y=3300000, label=total_member_rides, size=4)
```



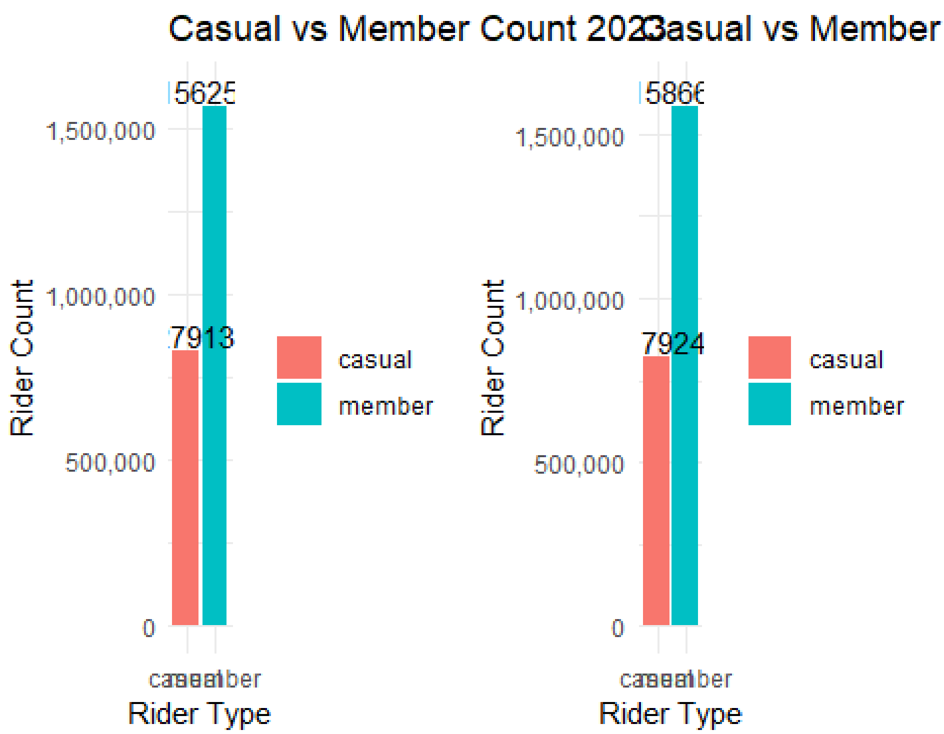
Here we see around 65% of rides are done by members.

Chart 2: Casual And Member Count 2023-2024

Keep in mind that 2023 and 2024 datas include only first six months.

```
casual_2023 <- all_rides_2 %>% filter(year == 2023 & member_casual == "casual
") %>% nrow()
member_2023 <- all_rides_2 %>% filter(year == 2023 & member_casual == "member
") %>% nrow()

casual_2024 <- all_rides_2 %>% filter(year == 2024 & member_casual == "casual
") %>% nrow()
member_2024 <- all_rides_2 %>% filter(year == 2024 & member_casual == "member
") %>% nrow()

#Casual and member count 2023-2024
print(paste(casual_2023, member_2023, casual_2024, member_2024))
```

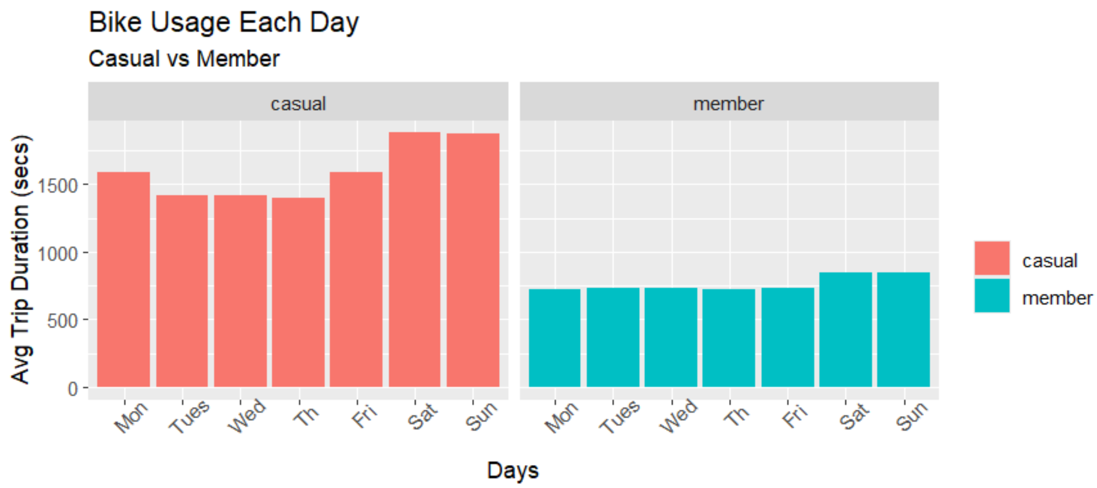
```
## [1] "827913 1562524 817924 1586644"

#Creating casual_vs_member_count_2023 plot
casual_vs_member_count_2023 <- all_rides_2 %>% filter(year == 2023) %>%
group_by(member_casual) %>%
summarise(rider_count = n()) %>%
ggplot(aes(x = member_casual, y = rider_count, fill = member_casual)) +
geom_bar(stat = "identity") +
labs(title="Casual vs Member Count 2023", x="Rider Type", y = "Rider Count",
fill = "")+
scale_y_continuous(labels = comma) + theme_minimal() +
annotate("text", x = 1, y = 880000, label = casual_2023, size = 4) +
annotate("text", x = 2, y = 1615000, label = member_2023, size = 4)

#Creating casual_vs_member_count_2024 plot
casual_vs_member_count_2024 <- all_rides_2 %>% filter(year == 2024) %>%
group_by(member_casual) %>%
summarise(rider_count = n()) %>%
ggplot(aes(x = member_casual, y = rider_count, fill = member_casual)) +
geom_bar(stat = "identity") +
labs(title="Casual vs Member Count 2024", x="Rider Type", y="Rider Count",fil
l = "")+
scale_y_continuous(labels = comma) + theme_minimal() +
annotate("text", x = 1, y = 870000, label = casual_2024, size = 4) +
annotate("text", x = 2, y = 1635000, label = member_2024, size = 4)

#Combine these 2 plots with patchwork package
combined_plot <- casual_vs_member_count_2023 + casual_vs_member_count_2024

combined_plot
```



By comparing these values, we can see that the number of member riders has increased by 24,120, which represents a 0.5% change. Conversely, the count of casual riders has decreased by 9,989, or about 0.2%. Although these changes are minor relative to the total values, they still represent a positive shift.

Chart 3: Bike Usage Each Day

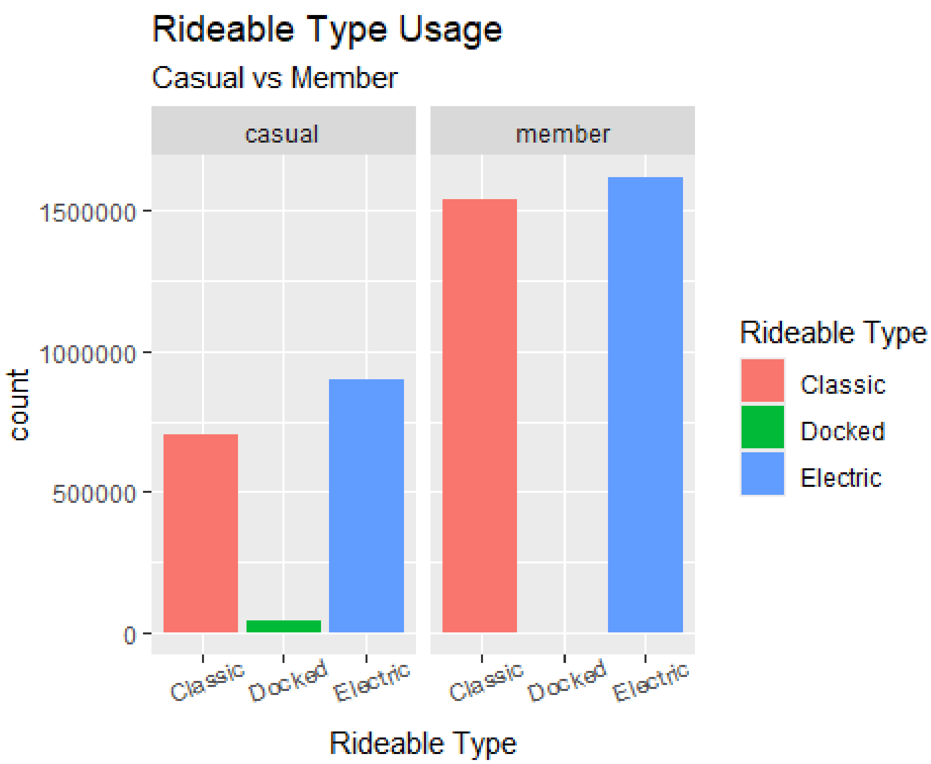
```
all_rides_2 %>%
mutate(ride_length = as.numeric(ride_length, units = "secs"),
day_of_week = factor(day_of_week,
levels=c("Monday","Tuesday","Wednesday","Thursday","Friday","Saturday","Sunda
y")))%>%
group_by(day_of_week, member_casual) %>%
summarise(mean_ride_length = mean(ride_length), .groups = "drop") %>%
ggplot(aes(x = day_of_week, y = mean_ride_length, fill = member_casual)) +
geom_bar(stat = "identity", position = "dodge") +
facet_wrap(~member_casual) +
theme(axis.text.x = element_text(angle = 45)) +
labs(title = "Bike Usage Each Day", subtitle = "Casual vs Member",
x = "Days", y = "Avg Trip Duration (secs)", fill = "") +
scale_x_discrete(labels = c("Monday" = "Mon", "Tuesday" = "Tues", "Wednesday"
= "Wed",
"Thursday" = "Th", "Friday" = "Fri", "Saturday" = "Sat", "Sunday" = "Sun"))
```



Member riders exhibit a much more consistent riding pattern compared to casual riders, with longer trip durations evident in both charts. This consistency suggests that member riders may use bikes more regularly, possibly for commuting to work.

Chart 4: Rideable Type Usage

```
all_rides_2 %>%
ggplot(aes(x=rideable_type,fill=rideable_type))+
geom_bar() + facet_wrap(~member_casual) +
theme(axis.text.x = element_text(angle = 20)) +
labs(
  title="Rideable Type Usage",
  subtitle="Casual vs Member",
  x="Rideable Type",fill="Rideable Type") +
scale_x_discrete(labels = c(
  "classic_bike" = "Classic",
  "docked_bike" = "Docked",
  "electric_bike" = "Electric")) +
scale_fill_discrete(labels = c(
  "classic_bike" = "Classic",
  "docked_bike" = "Docked",
  "electric_bike" = "Electric"))
```

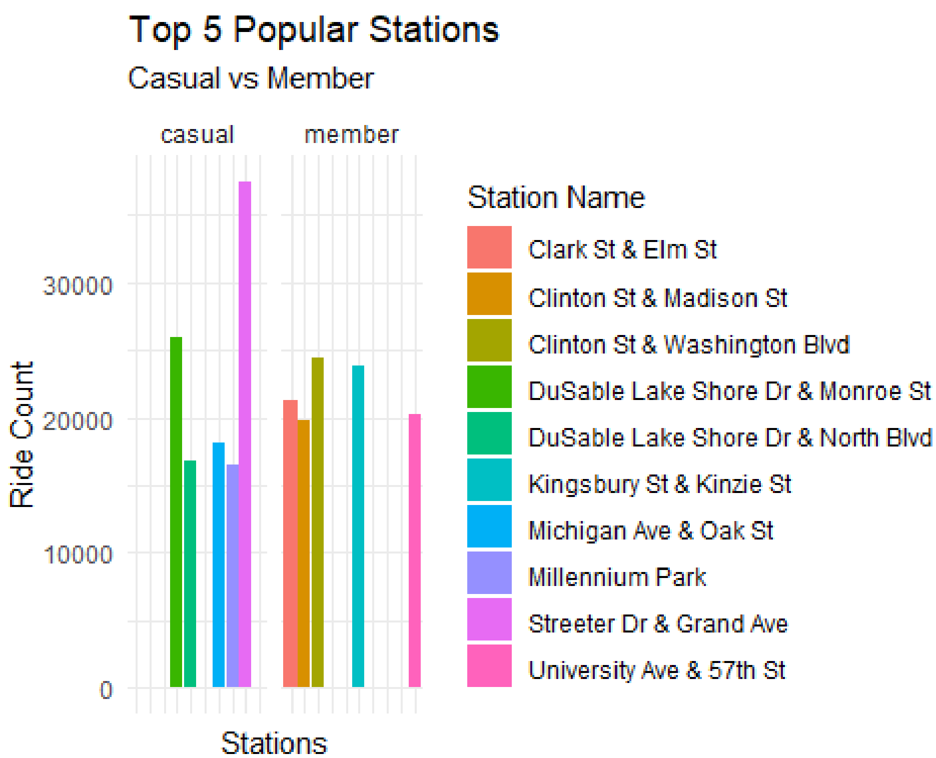


On both charts electric bikes > classic bikes > docked bikes

Chart 5: Top Five Popular Stations

```
#Calculates top 5 stations for member riders
member_station <- all_rides_2 %>% drop_na(start_station_name) %>%
filter(member_casual == "member") %>%
group_by(start_station_name) %>%
summarise(each_station_ride_count = n()) %>%
arrange(-each_station_ride_count) %>%
slice_head(n=5)

member_station

casual_station <- casual_station %>% mutate(member_casual = "casual")
member_station <- member_station %>% mutate(member_casual = "member")

member_casual_station <- bind_rows(casual_station, member_station)

member_casual_station %>% arrange(each_station_ride_count) %>%
ggplot(aes(
  x = start_station_name,
  y = each_station_ride_count,
  fill= start_station_name))+
geom_bar(stat = "identity")+ facet_wrap(~member_casual) +
theme_minimal() + theme(axis.text.x = element_blank()) +
labs(
  title = "Top 5 Popular Stations",
```

```
  subtitle = "Casual vs Member",
  x = "Stations",y = "Ride Count",
  fill = "Station Name")
```

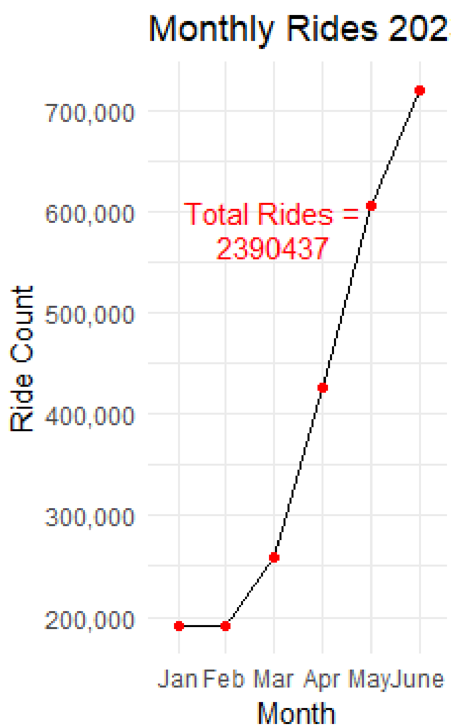


While this chart highlights the top five stations, it also indicates the least popular stations. To boost their popularity, we might consider strategies such as organizing events or offering discounts on rides.

Chart 6: Monthly Rides 2023-2024

```
#total rides of 2023 first six months
total_rides_2023_01_06 <- all_rides_2 %>% filter(year == 2023) %>% nrow()
#total rides of 2024 first six months
total_rides_2024_01_06 <- all_rides_2 %>% filter(year == 2024) %>% nrow()

plot_2023 <- all_rides_2 %>%
filter(year == 2023) %>%
arrange(started_at) %>%
group_by(month) %>%
summarise(year = "2023", monthly_ride = n())

plot_2024 <- all_rides_2 %>%
filter(year == 2024) %>%
arrange(started_at) %>%
group_by(month) %>%
summarise(year = "2024", monthly_ride = n())
```

```r
plot_2023 <- plot_2023 %>%
ggplot(aes(x = month, y = monthly_ride, group = 1)) +
geom_line() + geom_point(color = "red")+
labs(title= "Monthly Rides 2023", x = "Month", y = "Ride Count")+
scale_x_discrete(labels = c(
  "01" = "Jan",
  "02" = "Feb",
  "03" = "Mar",
  "04" = "Apr",
  "05" = "May",
  "06" = "June")) +
scale_y_continuous(labels = comma) + theme_minimal() +
annotate("text",x=3,y=600000,label="Total Rides =",size=4,color="Red") +
annotate("text",x=3,y=565000,label=total_rides_2023_01_06,size=4,color="Red")

plot_2024 <- plot_2024 %>%
ggplot(aes(x = month, y = monthly_ride, group = 1)) +
geom_line() + geom_point(color = "purple")+
labs(title= "Monthly Rides 2024", x = "Month", y = "Ride Count") +
scale_x_discrete(labels=c("01"="Jan","02"="Feb","03"="Mar","04"="Apr","05"="M
ay","06"= "June")) +
scale_y_continuous(labels = comma) + theme_minimal() +
annotate("text",x = 3,y=600000,label="Total Rides =", size = 4, color = "Purp
le") +
annotate("text",x=3,y=565000,label=total_rides_2024_01_06,size=4,color="Purpl
e")

# Combined this plots with patchwork package
combined_2023_2024 <- plot_2023 + plot_2024

combined_2023_2024
```



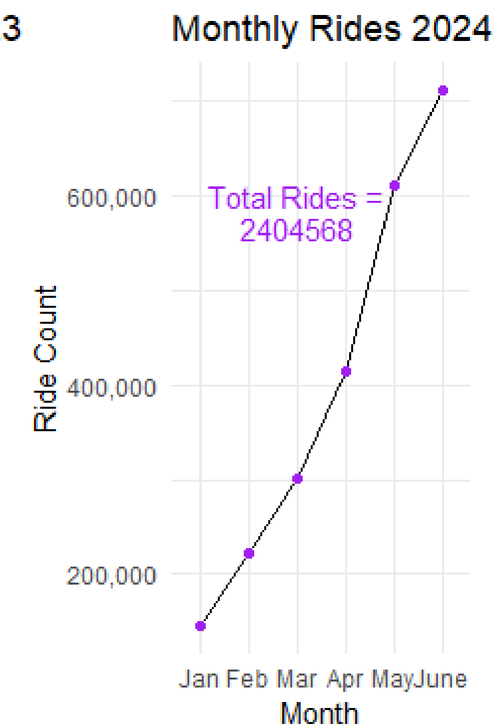
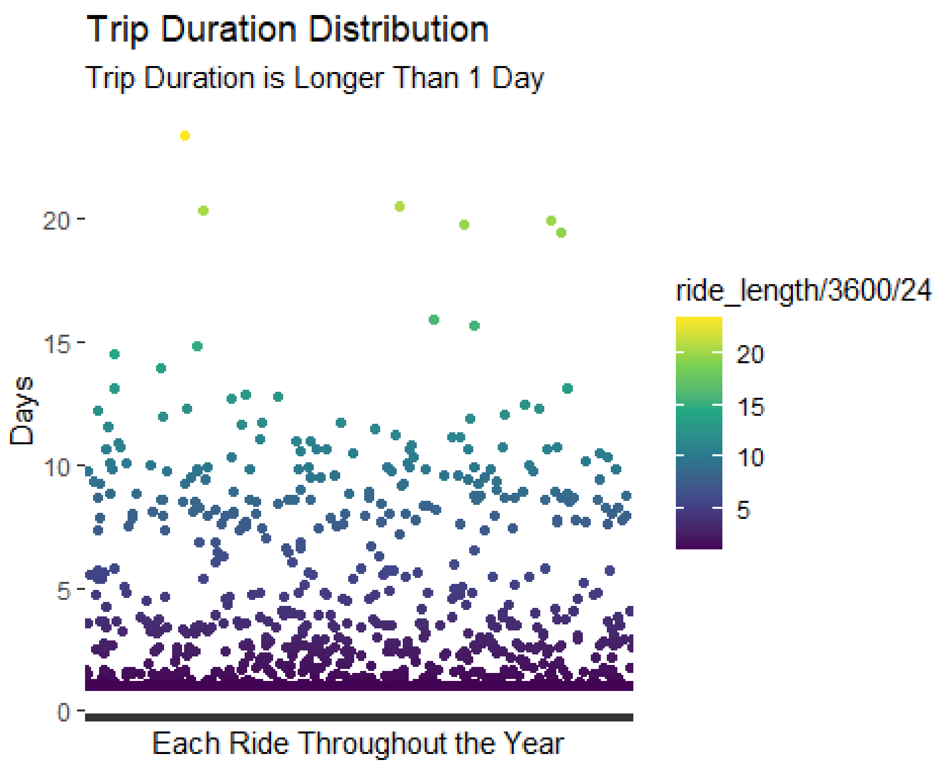


Chart 7: Trip Duration Distribution

```
# Scatterplot for the people who took the bike more than 1 day
all_rides_2 %>%
filter(ride_length >86400) %>%
ggplot(aes(
 x = ride_id,
 y = ride_length/3600/24,
 color = ride_length /3600/24)) +
geom_jitter() + theme(axis.text.x = element_blank()) +
labs(
  title= "Trip Duration Distribution",
  subtitle = "Trip Duration is Longer Than 1 Day",
  x = "Each Ride Throughout the Year", y = "Days") +
scale_color_viridis_c()
```



With the recent charts, we've examined the visualized data to better understand its implications and how it can guide us in shaping the company's future. Now, let me summarize what I've learned from these charts.

Key Findings

Data Analysis on Bike Types: The analysis shows that classic bikes have been associated with 9 instances of data issues, whereas electric bikes have had 197 instances. The higher number of data problems with electric bikes suggests that the company needs to take steps to address these issues.

Rider Demographics and Strategy: Around 65% of riders are members. To boost membership, it is suggested that the company pinpoint popular stations where casual riders frequent and direct targeted marketing efforts towards these spots. Potential strategies include providing discounts, offering prizes, and hosting events to attract and convert casual riders into members.

Year-over-Year Comparison: Comparing data from 2023 to 2024, there is a noticeable rise in both the total number of riders and the number of member riders. While the growth is modest, it is a positive development. If this upward trend continues over the next decade, it could greatly enhance the company's performance.

Note: The data analyzed in this report covers the first six months of 2023 and 2024.