Cyclistic_Project_Portfolio

Leonardo_N_Carvalho

2025-12-11

This analysis was made as part of the Google Data Analytics Certificate, the main purpose is to practice what was learned during the course and build a portfolio.

The following introduction was given by the course

Introduction

Welcome to the Cyclistic bike-share analysis case study! In this case study, you work for a *fictional company*, Cyclistic, along with some key team members. In order to answer the business questions, follow the steps of the data analysis process: Ask, Prepare, Process, Analyze, Share, and Act. Along the way, the Case Study Roadmap tables — including guiding questions and key tasks — will help you stay on the right path.

Scenario

You are a junior data analyst working on 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. Characters and teams:

- *Cyclistic*: A bike-share program that features more than 5,800 bicycles and 600 docking stations. Cyclistic sets itself apart by also oering 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 the bikes 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.

About the company

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

Until now, Cyclistic's marketing strategy relied on building general awareness and appealing to broad consumer segments. One approach that helped make these things possible was the flexibility of its pricing plans: single-ride passes, full-day passes, and annual memberships. Customers who purchase single-ride or full-day passes are referred to as casual riders. Customers who purchase annual memberships are Cyclistic members.

Cyclistic's finance analysts have concluded that annual members are much more profitable than casual riders. Although the pricing flexibility helps Cyclistic attract more customers, Moreno believes that maximizing the number of annual members will be key to future growth. Rather than creating a marketing campaign that targets all-new customers, Moreno believes there is a solid opportunity to convert casual riders into members. She notes that casual riders are already aware of the Cyclistic program and have chosen Cyclistic for their mobility needs.

Moreno has set a clear goal: Design marketing strategies aimed at converting casual riders into annual members. In order to do that, however, the team needs to beer understand how annual members and casual riders differ, why casual riders would buy a membership, and how digital media could affect their marketing tactics. Moreno and her team are interested in analyzing the Cyclistic historical bike trip data to identify trends.

Ask

Three questions will guide the future marketing program:

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

Moreno has assigned you the first question to answer: **How do annual members and casual riders use Cyclistic bikes differently?**

Prepare and Process

It is a very important step during the analysis, it allows me to generate tables that will be used latter to create the visualization and better understand the data.

Libraries used:

```r
library(dplyr)
library(ggplot2)
library(scales)
library(data.table)
```

During this analysis I used 12 files combined, from 11/2024 to 10/2025, you can find this data on the following link: https://divvy-tripdata.s3.amazonaws.com/index.html

```
folder_path <- "C:/Users/55129/Documents/Projects/Bikes/"
file_list <- list.files(path = folder_path, pattern = "*.csv", full.names = TRUE)
message("Starting file read and combination process...")
all_data <- rbindlist(lapply(file_list, fread))
message("Files successfully combined.")
```

After creating the table named all_data (raw data) I started to clean the it, **1-** eliminating some columns that I would not use. To keep my analysis in a period of an year (12 months) **2-** I deleted some data from 09 and 10/2024, that were wrongly inserted. **3-** Two columns were created for calculating the duration of each ride in seconds and minutes. In order to add more useful information **4-** was opted in adding 3 more columns: year_month, month_name and day_of_week

```
# 1-
cleaned_data <- all_data %>%
  select(
    -ride_id,
    -start_station_name,
    -start_station_id,
    -end_station_name,
    -end_station_id,
    -start_lat,
    -start_lng,
    -end_lat,
    -end_lng)

cleaned_data <- cleaned_data %>%
# to make sure that the columns are in the right date format and time zone
  mutate(
    started_at = as.POSIXct(started_at, format="%Y-%m-%d %H:%M:%s", tz = "UTC"),
    ended_at = as.POSIXct(ended_at, format="%Y-%m-%d %H:%M:%s ", tz = "UTC"))

# 2-
start_date <- as.POSIXct("2024-11-01 00:00:00", format("%Y-%m-%d %H:%M:%S"))

cleaned_data <- cleaned_data %>%
  filter(started_at >= start_date)

# 3-
cleaned_data <- cleaned_data %>%  #To calculate the ride length
mutate(
    ride_length_sec = ended_at - started_at,
    ride_length_min = as.numeric(ride_length_sec, units = "mins"))  # convert

# 4-
Sys.setlocale("LC_TIME", "English")
```

```
## [1] "English_United States.1252"
```

```
cleaned_data <- cleaned_data %>%
  mutate(
    year_month = format(started_at, "%Y-%m"),
    month_name = format(started_at, "%B"),
    day_of_week = format(started_at, "%A"))
```

In the next chunks I created new tables to compare the **number of rides**. **1-** Total number of rides by day of the week - Member x Casual. **2-** Total number of rides in the last 12 months - Member x Casual. **3-** Total number of rides by month - Member x Casual. **4-** Total number of rides by day of the week - Member and Casual together

```
# 1-
day_of_week_rides_summary <- cleaned_data %>%
  group_by(member_casual, day_of_week) %>%
  summarise(total_rides = n(),
  .groups = 'drop')

day_order <- c("Sunday", "Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday")
day_of_week_rides_summary$day_of_week <- factor(
  day_of_week_rides_summary$day_of_week,
  levels = day_order
)
day_of_week_rides_summary <- day_of_week_rides_summary %>%
  arrange(day_of_week)

# 2-
year_rides_summary <- day_of_week_rides_summary %>%
  group_by(member_casual)%>%
  summarise(
    year_rides = sum(total_rides),
    groups = 'drop')

# 3-
monthly_rides_summary <- cleaned_data %>%
  group_by(member_casual, year_month) %>%
  summarise(total_rides = n(),
            .groups = "drop")%>%
  arrange(year_month)

# 4-
overall_daily_rides <- day_of_week_rides_summary %>%
    group_by(day_of_week) %>%
  summarise(
    total_rides_combined = sum(total_rides),
    groups = 'drop'
  )
```

Another important data is the time spend riding, so more tables were created. **1-** Average time riding in the last 12 months - Member x Casual. **2-** Average time riding by days of the week - Member x Casual. **3-** Average time riding by month - Member x Casual. **4-** Total minutes ridden in the last 12 months - Members x Casual.

```
# 1-
avg_ride_duration_annual <- cleaned_data %>%
  group_by(member_casual) %>%
  summarise(
    average_time = mean(ride_length_min, na.rm = TRUE)
  )

# 2-
```

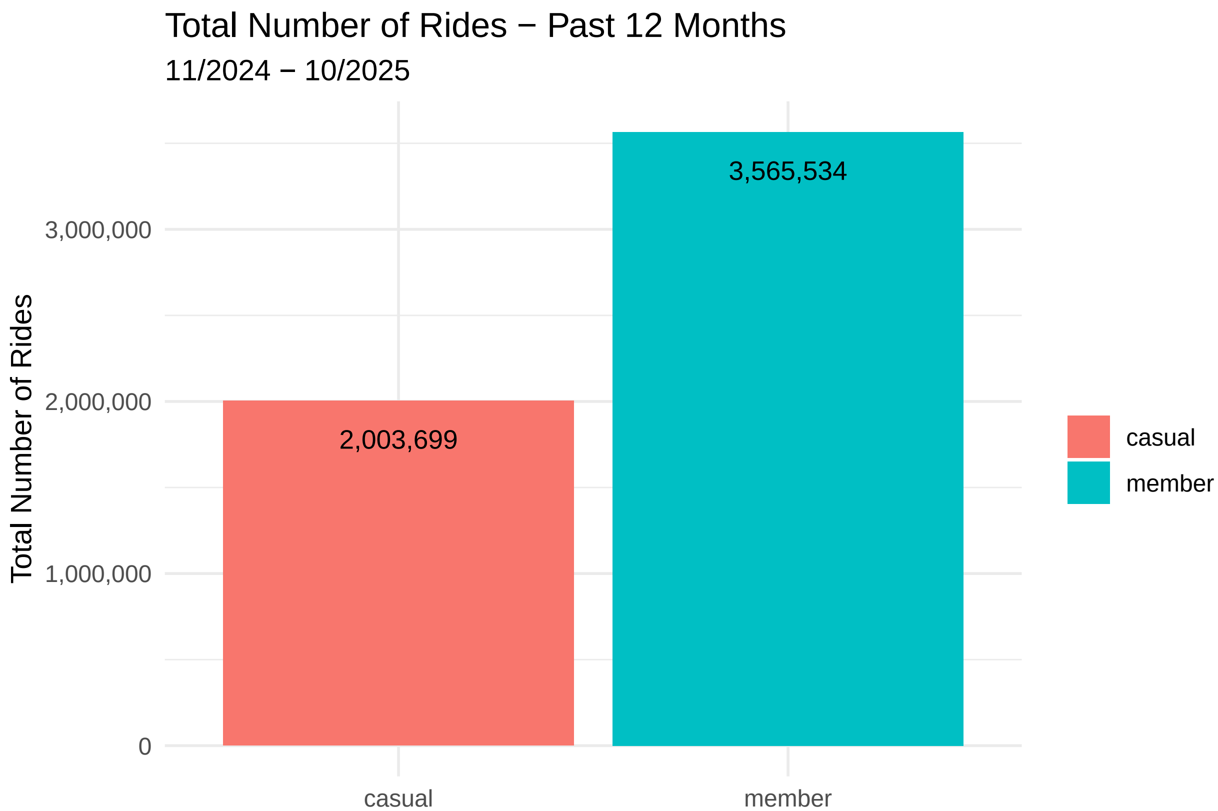
```r
avg_ride_duration_day_week <- cleaned_data %>%
  group_by(member_casual, day_of_week) %>%
  summarise(
    average_time = mean(ride_length_min, na.rm = TRUE)
  )
day_order <- c("Sunday", "Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday")
avg_ride_duration_day_week$day_of_week <- factor(
  avg_ride_duration_day_week$day_of_week,
  levels = day_order
)
avg_ride_duration_day_week <- avg_ride_duration_day_week %>%
  arrange(day_of_week)

# 3-
avg_ride_duration_monthly <- cleaned_data %>%
  group_by(year_month, month_name, member_casual) %>%
  summarise(
    average_time = mean(ride_length_min, na.rm = TRUE),
    .groups = 'drop'
  ) %>%
  arrange(year_month)

# 4-
total_time_riding <- cleaned_data %>%
  group_by(member_casual)%>%
  summarise(total_time_minutes = sum(ride_length_min),
            .groups = 'drop')
```

Analyze and Share

Total number of rides - Annual

```r
ggplot(data = year_rides_summary,
       aes(x=member_casual, y=year_rides, fill = member_casual))+
geom_col(position = 'dodge')+
  geom_text(
    aes(label = comma(year_rides)),
    vjust = 2.5,
    size = 3.5
  ) +
  labs(
    title = "Total Number of Rides - Past 12 Months",
    subtitle = "11/2024 - 10/2025",
```

```r
    y = "Total Number of Rides",
    x = "",
    fill = ""
  )+
  scale_y_continuous(labels = comma) +
  theme_minimal()
```

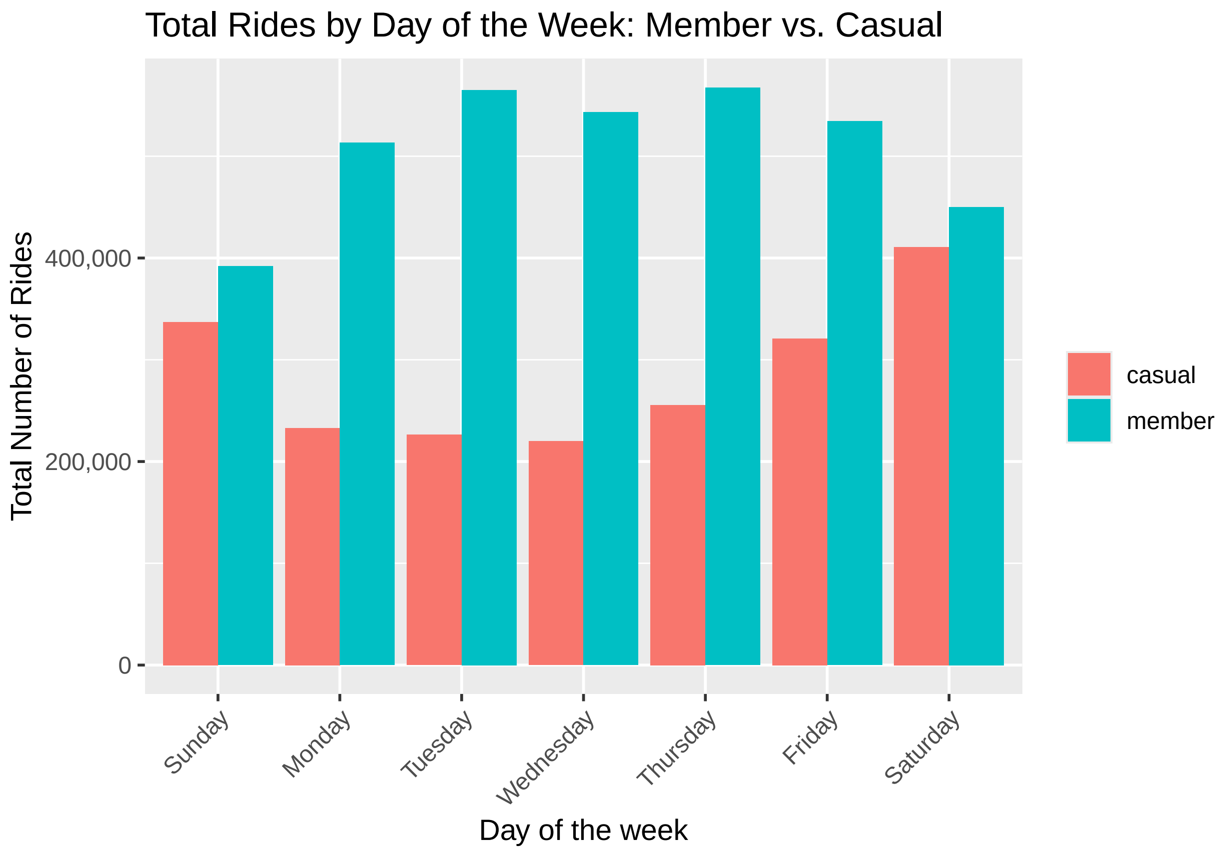


I begin my analysis comparing the total number of rides in the past 12 months, we can see that Casual riders represented 36% of the total rides, which is a very expressive number. Once Members have unlimited access to use bikes without extra costs it is expected for them to be using it more times.

Total number of rides - Day of week

```r
ggplot(data = day_of_week_rides_summary,
       aes(x=day_of_week, y=total_rides, fill = member_casual))+
  geom_col(position = "dodge") +
  labs(
    title = "Total Rides by Day of the Week: Member vs. Casual",
    x = "Day of the week",
    y = "Total Number of Rides",
    fill = ""
```

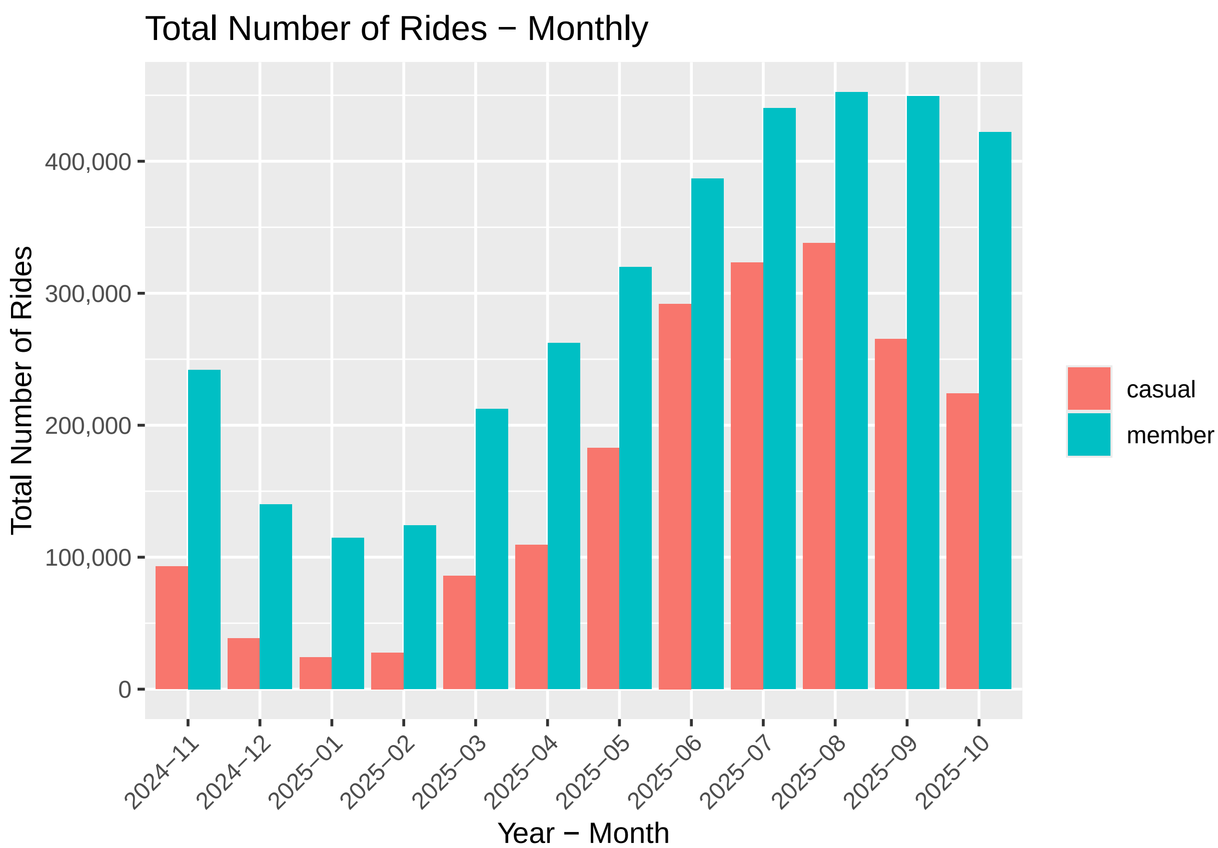
```
  ) +
  scale_y_continuous(labels = comma) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



An interesting pattern can be observed in this graph. Members ride more times during the week, while Casual riders use more from Thursday to Sunday.

Total number of rides - Per Month

```
ggplot(data = monthly_rides_summary,
       aes(x = year_month, y = total_rides, fill = member_casual))+
  geom_col(position = "dodge")+
  labs(
    title = "Total Number of Rides - Monthly",
    x = "Year - Month",
    y = "Total Number of Rides",
    fill = ""
  )+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+
  scale_y_continuous(labels = comma)
```

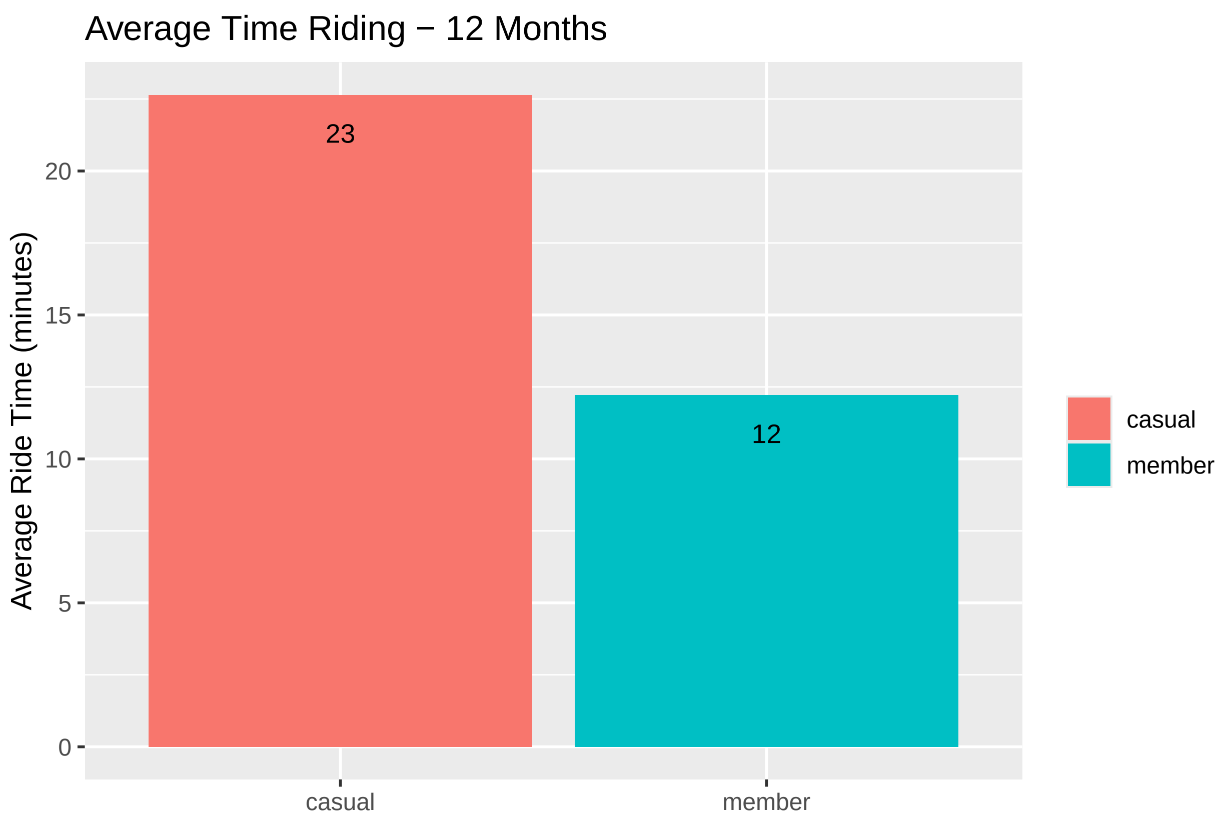


We can see that the number of rides has been increasing in the past 12 months. And another pattern is present, the number of rides decrease during winter due to the negative temperatures in Chicago

Comparing only the number of rides we can see that Members use a lot more, 64% of the total rides. However this is due to the benefit of unlimited rides granted by the annual membership.

Average time riding - Annual

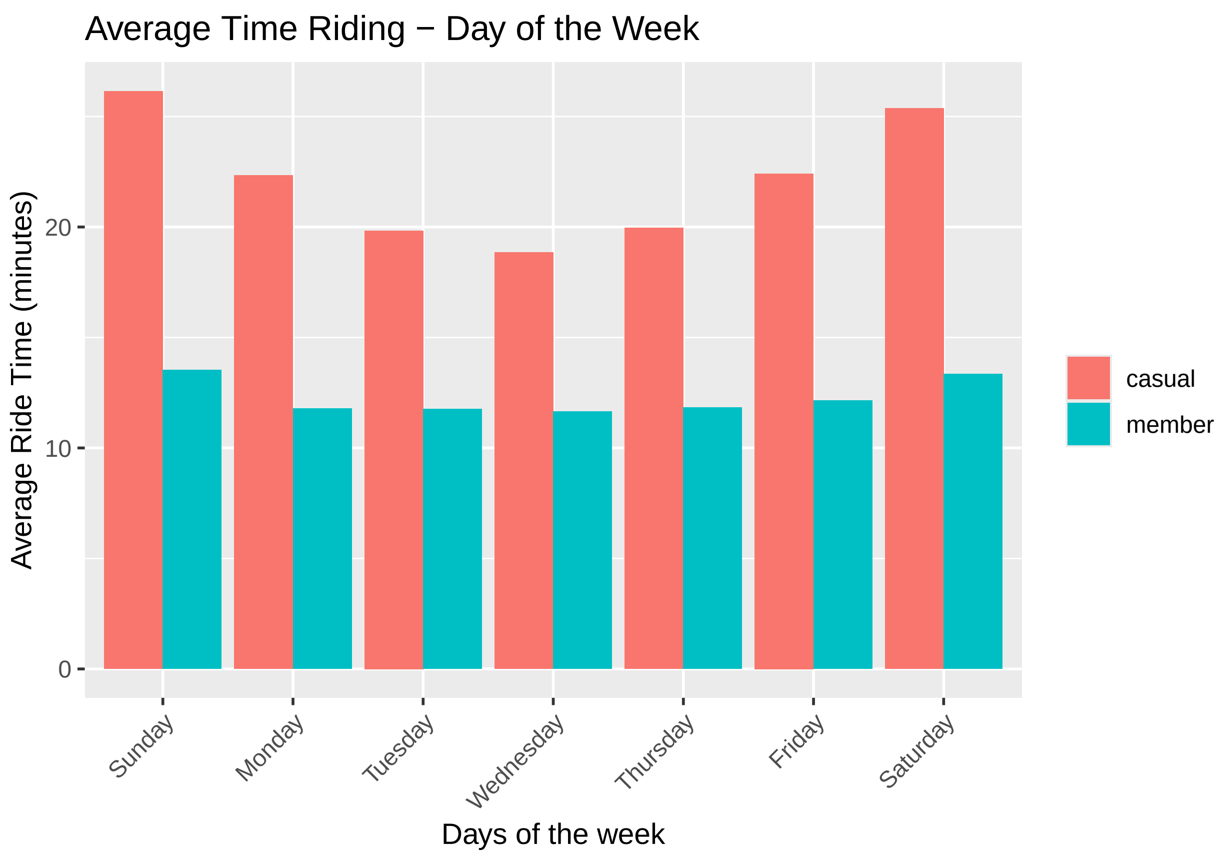
```
ggplot(data=avg_ride_duration_annual, aes(x = member_casual, y = average_time, fill=member_casual))+
  geom_col(position = "dodge") +
  labs(title = "Average Time Riding - 12 Months",
       x="",
       y="Average Ride Time (minutes)",
       fill="")+
  geom_text(
    aes(label = comma(average_time)),
    vjust = 2.5,
    size = 3.5
  )
```



When we compare the average time riding we see that casuals usually ride for 23 minutes while members for 12 minutes. It means Casuals usually ride 92% more minutes than Members. This data can be increased tanks to single-ride passes, once the client unlocked the bicycle they'll try to make the most of it, avoid returning it and needing to pay the pass again.

Average time riding - Day of the week

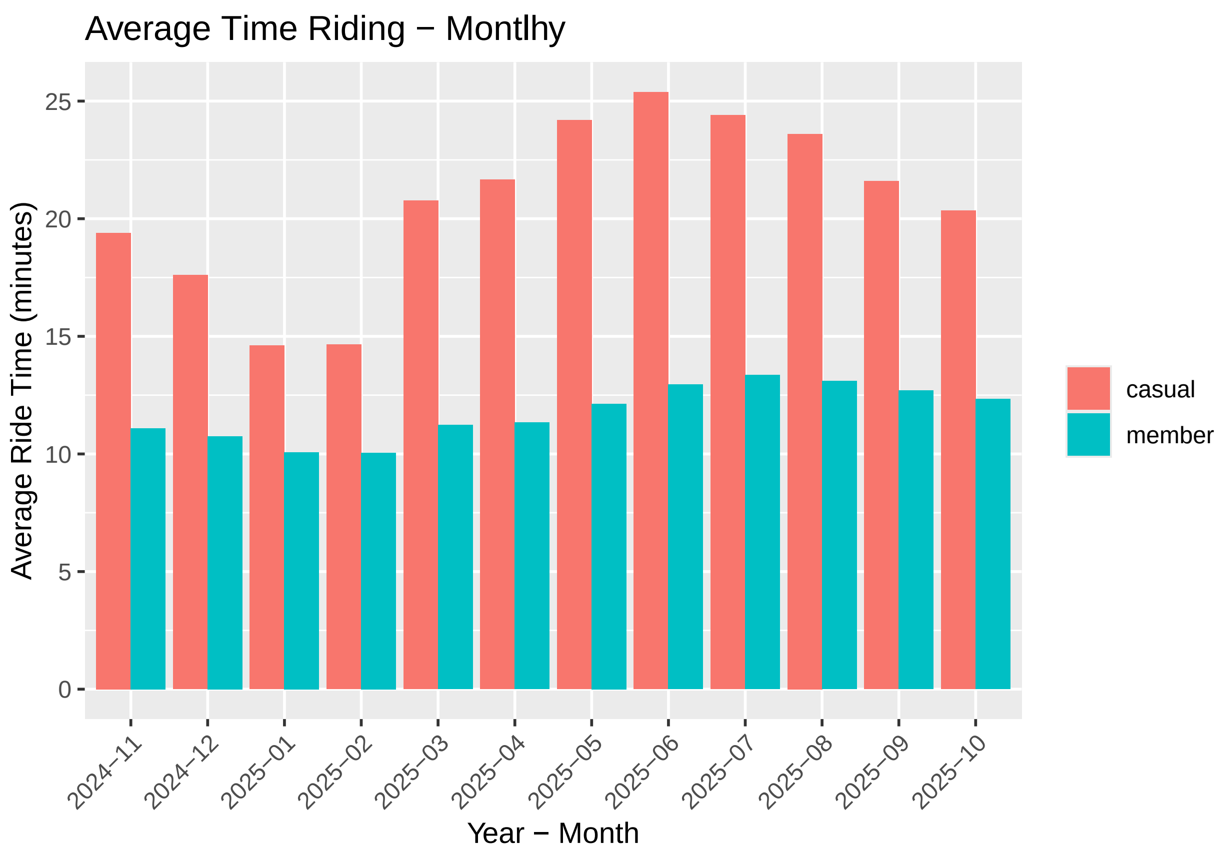
```
ggplot(data=avg_ride_duration_day_week,
       aes(x=day_of_week, y=average_time, fill = member_casual))+
  geom_col(position='dodge')+
  labs(
    title = "Average Time Riding - Day of the Week",
    x = "Days of the week",
    y = "Average Ride Time (minutes)",
    fill = ""
  )+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



For members the average time riding is very steady. However for Casual riders it follows the same pattern as the total number of rides: more rides and more time spend riding during the end of the week and weekend.

Average time riding - Per Month

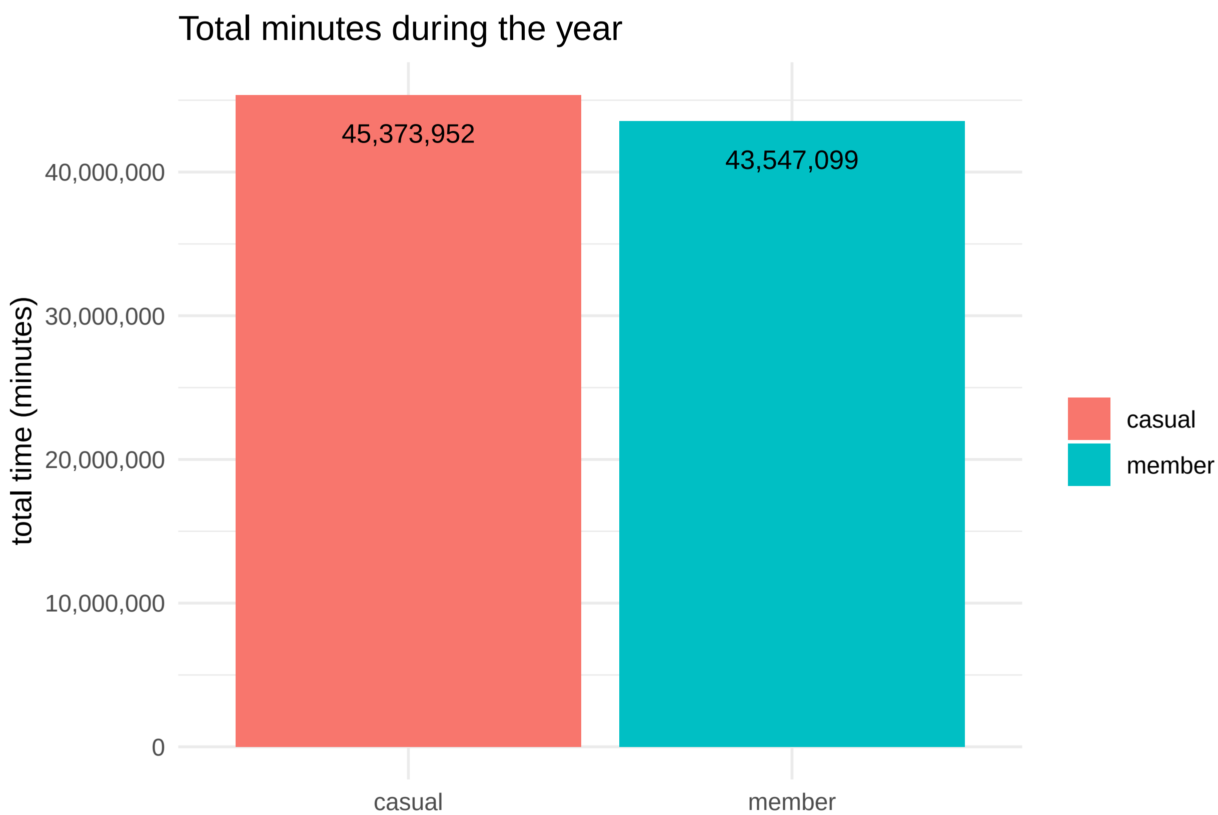
```
ggplot(data = avg_ride_duration_monthly,
       aes(x = year_month, y = average_time, fill = member_casual))+
  geom_col(position = "dodge")+
  labs(
    title = "Average Time Riding - Montlhy",
    x = "Year - Month",
    y = "Average Ride Time (minutes)",
    fill = ""
  )+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Analyzing the average time per month we can see that Members still steady 11-13 minutes. On the other hand Casual riders, as we saw, tend to be seasonal and the minutes spent riding decrease during winter.

Total Minutes Rode During the Year

```
ggplot(data = total_time_riding,
       aes(x=member_casual, y=total_time_minutes, fill = member_casual))+
  geom_col(position = 'dodge')+

  labs(
    title = "Total minutes during the year",
    y = "total time (minutes)",
    x = "",
    fill = ""
  )+
  geom_text(
    aes(label = comma(total_time_minutes)),
    vjust = 2.5,
    size = 3.5
  ) +
  scale_y_continuous(labels = comma) +
  theme_minimal()
```



Members have more rides during the period whereas Casuals usually ride for more time

Over the 12-month period, Casual riders accumulated 1,826,853 more minutes of total ride time than Annual Members.

Conclusion

From the data showed we can conclude that Members are using bikes all over the year, with the average time always similar while the number of rides tend to be bigger on weekdays, indicating that they use it as an alternative transport while they're going to work, university or other appointments. As it was told in the introduction 30% of the total riders use Cyclistic to comute to work, even with the lack of data we can suppose that the majority of this 30% is composed by Members. Casual riders tend to use more times and for long periods from Thursday to Sunday and then it starts to decrease until Wednesday and repeat the cycle. Leisure activity is probably the main use for them. We need to convince how beneficial it is to become a Member.

Act

KEY INSIGHTS

Digital Media Campaign focoused on ROI(Return on Ivestment) with the Annual Membership. This campaign can be runnig from Monday to Friday at peak hours and sporadic offers can be displayed for costumers that already use the app after the 3rd ride. Examples:

*** "Did you know that if you buy Cyclistic Annual Membership x times per week, at the end of the year you've saved x money"**

*** "If you use Cyclistic Annual Membership 3 times a week, you're payback will come in 7 months"**

*** "Don't be stuck during rush hours, use Cyclistic now with 10% discount on the Annual Membership"**

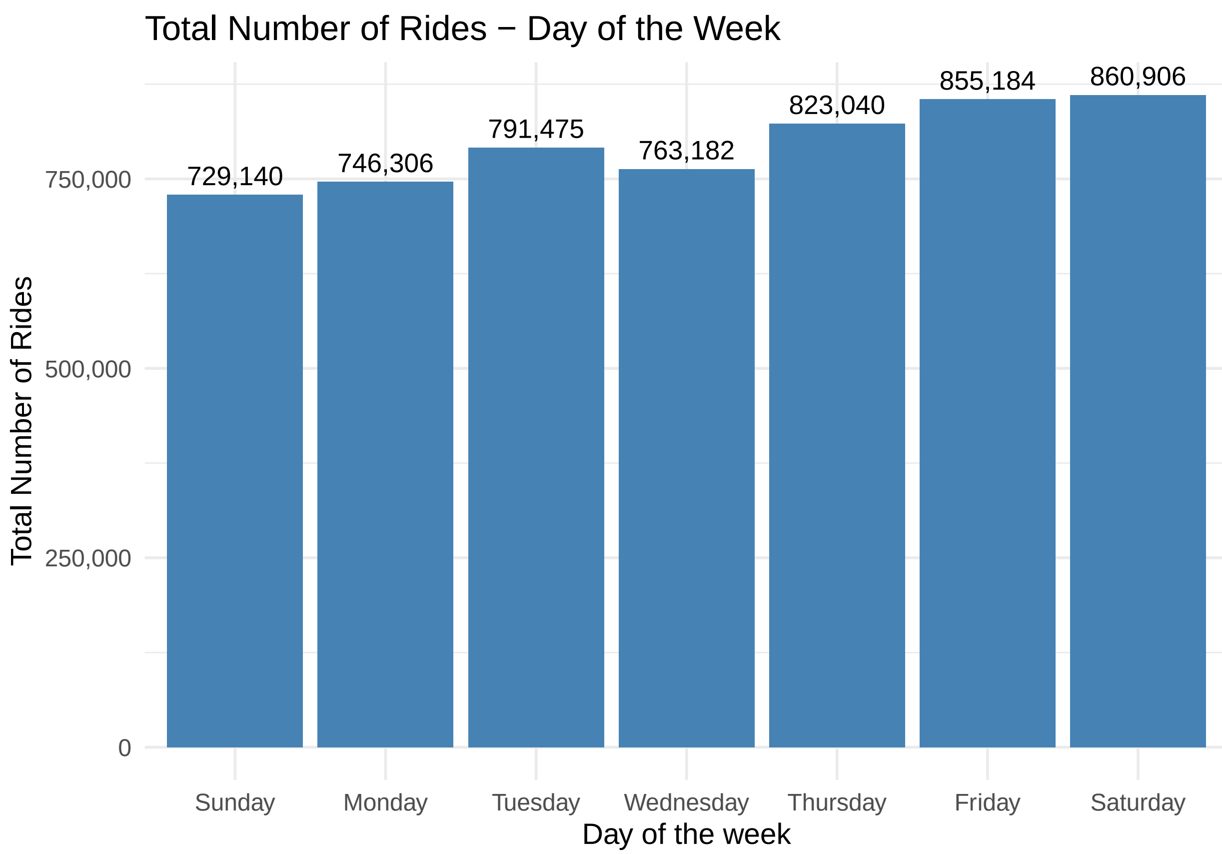
Create a Personalized In-App Conversion Tool with a pop-up windown on the app that shows how much money could the client save if they switch for the annual membership(based on how much they use the app).

Targeted Free Weekday Trial Program in order to create habtis and convert to Annual Membership. It can be offered a 3 days trial from Moday to Wednesday(days with less usage) that the users can taste the bennefits.

That's the end of the analysis, bellow you'll find more useful for information but it not connect to the main question asked to me.

Total Rides By Day of the Week - Casual and Member tgether

```
ggplot(data=overall_daily_rides,
       aes(x=day_of_week, y=total_rides_combined))+
  geom_col(position = "dodge", fill = "#4682B4")+
  geom_text(
    # Map the numeric result to the label aesthetic, formatted with commas
    aes(label = comma(total_rides_combined)),
    # Adjust position slightly above the bar (vjust = vertical justification)
    vjust = -0.5,
    size = 3.5 # Adjust font size
  ) +
  labs(
    title = "Total Number of Rides - Day of the Week",
    x = "Day of the week",
    y = "Total Number of Rides"
  )+
  scale_y_continuous(labels = comma)+
  theme_minimal()
```



Bike Preference - Electric x Classic

Table counting the total number of bikes used during 12 Months

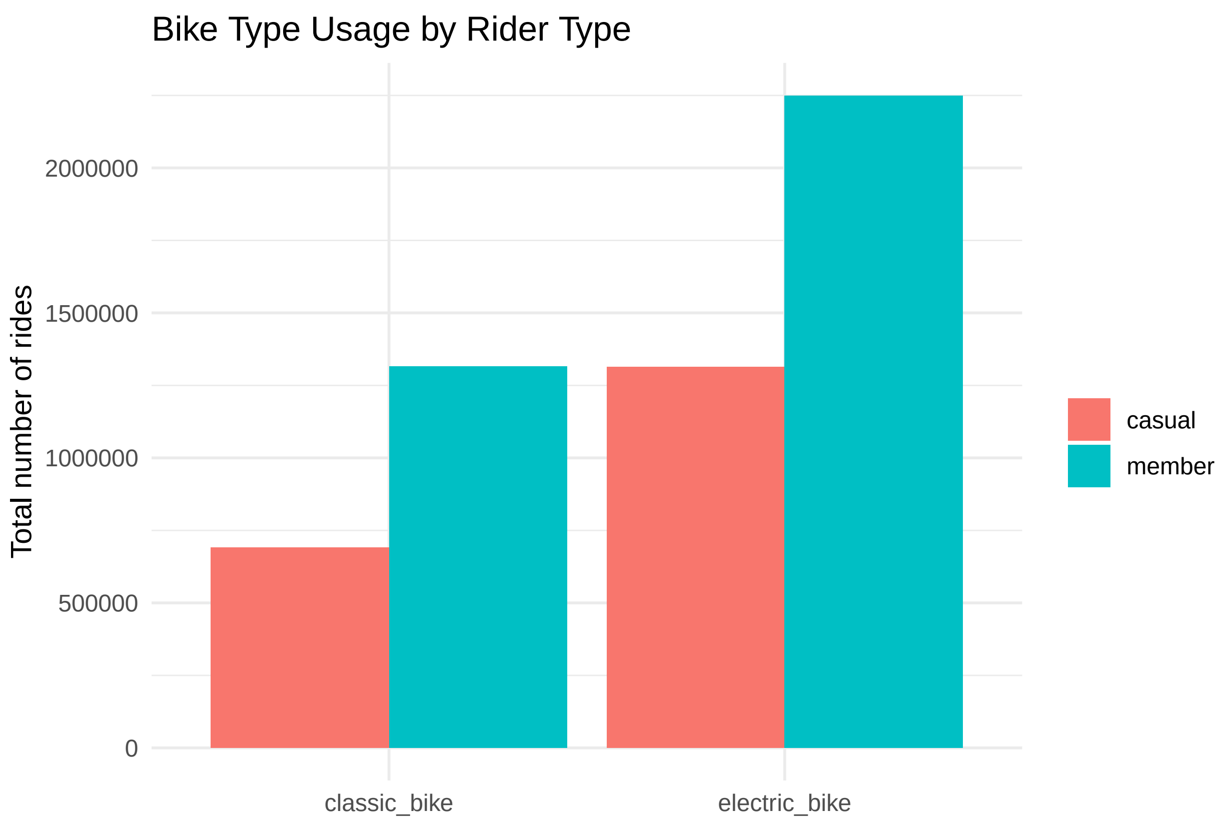
```
bike_usage_summary <- cleaned_data %>%
  group_by(member_casual, rideable_type) %>%
  summarise(
    total_rides = n(),
    .groups = "drop"
  )
```

Table counting the total number of bikes used per month

```
monthly_bike_usage <- cleaned_data %>%
  group_by(year_month, month_name, member_casual, rideable_type) %>%
  summarise(
    total_rides = n(),
    .groups = 'drop'
  ) %>%
  arrange(year_month)
```

Total Bikes Used - Past 12 Months

```r
ggplot(data = bike_usage_summary,
       aes(x=rideable_type, y=total_rides, fill=member_casual))+
  geom_col(position = "dodge")+

  labs(
    title = "Bike Type Usage by Rider Type",
    x = "",
    y = "Total number of rides",
    fill = ""
  )+
  theme_minimal()
```



Total Bikes Used - Per Month

```r
ggplot(data = monthly_bike_usage,
       aes(x = year_month,
           y = total_rides,
           fill = member_casual)) +

  geom_col(position = "dodge") + # Stack the bike types within each month
```

```r
# Facet the plot into two rows: one for Electric, one for Classic
facet_wrap(~ rideable_type, ncol = 2, scales = "fixed") +

labs(
  title = "Monthly Bike Preference Trend",
  x = "Year - Month",
  y = "Total Number of Rides",
  fill = ""
) +
scale_y_continuous(labels = comma) +
theme(axis.text.x = element_text(angle = 45, hjust = 1))
```