```r
# Hannah Mendbayar
# Final project
# STAT 201


##Overview of Cyclistic
# Chicago, renowned for its status as one of the most bike-friendly cities in the
# United States. The city embraces a bike-friendly domain, with more than 200 miles
# of on-street protected and shared bike lanes, complemented by extensive off-street
# paths.In this dynamic environment, businesses have seized opportunities to explore
# and participate in the transportation sector. Among them is **Cyclisic**, a prominent
# bike-share program that has been operating since its inception in 2016. With a fleet
# of over 5,800 bicycles and 600 docking stations strategically positioned across the city.
# Cyclistic stands out by offering a diverse range of bikes, including reclining bikes,
# hand tricycles, and cargo bikes, catering to various riders, including those
# with disabilities.

# Research Question: How do annual members and casual riders differ in their
#usage patterns of Cyclistic bikes?

# By addressing this question, I aim to provide compelling data insights that will
# inform the design of a targeted marketing strategy, with the ultimate goal of
# converting casual riders into annual members. This report will uncover patterns,
# identify trends, and provide data-baked insights.


#Loading necessary libraries
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(tidyverse)
```

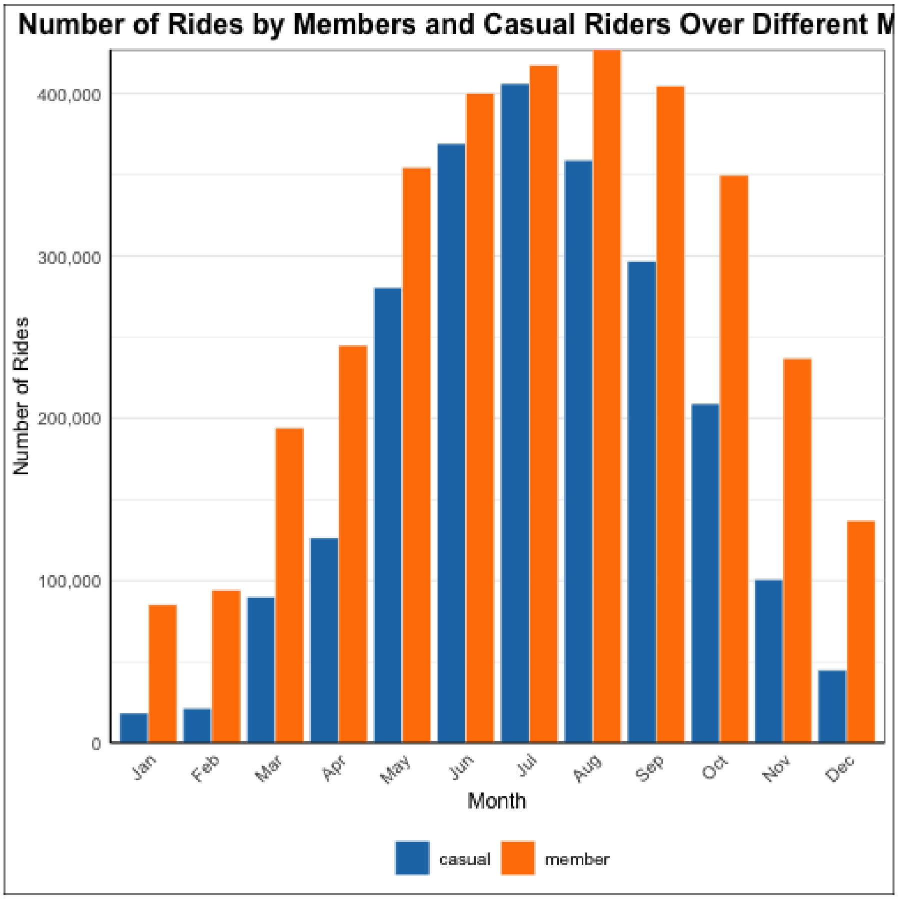
```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ lubridate 1.9.3     ✔ tibble    3.2.1
## ✔ purrr     1.0.2     ✔ tidyr     1.3.0
## ✔ readr     2.1.4
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ 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
```

```r
library(lubridate)

#Open the datasets for each month
month1 <- read.csv("month1.csv")
month2 <- read.csv("month2.csv")
month3 <- read.csv("month3.csv")
month4 <- read.csv("month4.csv")
month5 <- read.csv("month5.csv")
month6 <- read.csv("month6.csv")
month7 <- read.csv("month7.csv")
month8 <- read.csv("month8.csv")
month9 <- read.csv("month9.csv")
```

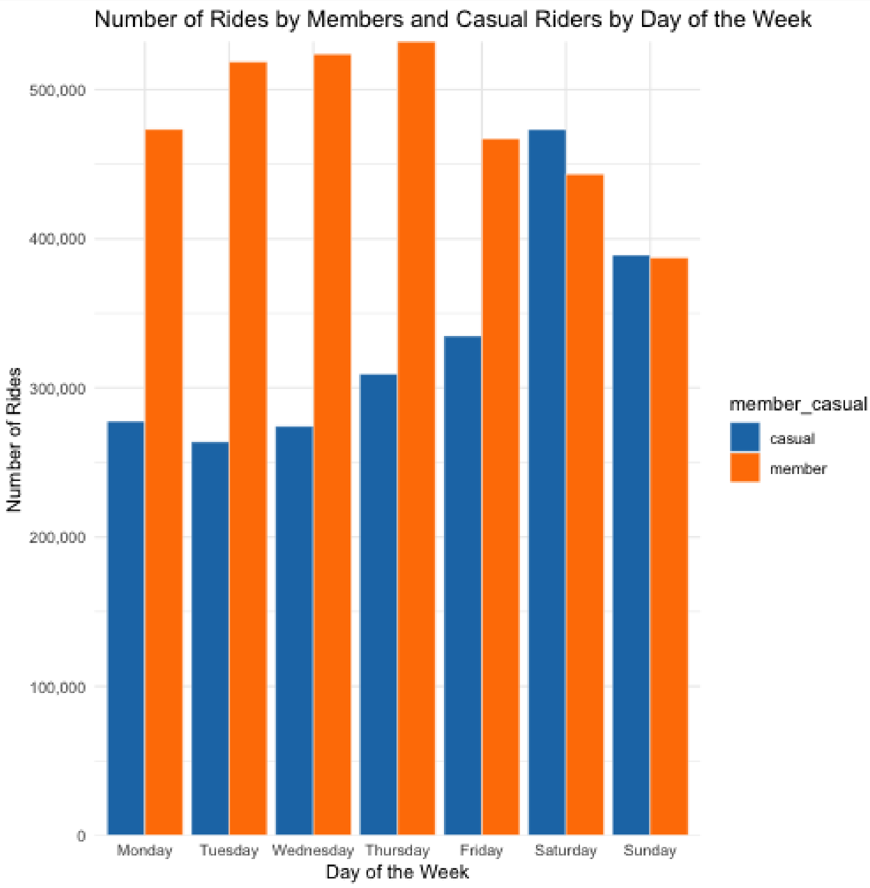
```
month10 <- read.csv("month10.csv")
month11 <- read.csv("month11.csv")
month12 <- read.csv("month12.csv")

#Combining all 12 datasets into one
all_months <- rbind(month1, month2, month3, month4, month5, month6, month7,
                    month8, month9, month10, month11, month12)

#Adding columns that list the date, month, day, and year of each ride
allmonths <- all_months %>%
  mutate(
    started_at = ymd_hms(started_at),
    ended_at = ymd_hms(ended_at),
    ride_date = as.Date(started_at),
    ride_month = month(started_at, label = TRUE),
    ride_day = day(started_at),
    ride_year = year(started_at),
    ride_day_of_week = weekdays(as.Date(started_at)),
    duration = difftime(ended_at, started_at, units = "secs")
  )

#Filtering out NA
all_months <- allmonths %>%
  filter(!is.na(ride_month))

#Create a custom color palette
custom_palette <- c("#1f78b4", "#ff7f00")

#Plotting a graph to compare casual riders vs. mebers by months.
ggplot(all_months, aes(x = ride_month, fill = member_casual)) +
  geom_bar(position = "dodge", color = "white", size = 0.2) +
  labs(title = "Number of Rides by Members and Casual Riders Over Different Months",
       x = "Month",
       y = "Number of Rides") +
  scale_fill_manual(values = custom_palette) +
  scale_y_continuous(labels = scales::label_comma(), expand = c(0, 0)) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        panel.grid.major.x = element_blank(),
        panel.grid.minor.x = element_blank(),
        axis.line.x = element_line(color = "black"),
        axis.line.y = element_line(color = "black"),
        legend.position = "bottom",
        legend.key.size = unit(0.7, "cm"),
        legend.title = element_blank(),
        plot.title = element_text(size = 16, face = "bold", hjust = 0.5),
        axis.title = element_text(size = 12),
        axis.text = element_text(size = 10),
        legend.text = element_text(size = 10),
        panel.background = element_rect(fill = "white"),
        plot.background = element_rect(fill = "white"))
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```




```
### Methods
# The primary dataset used for this analysis consists of monthly bike trip data
# from Cyclistic. The data, sourced directly from Divvy's data repository,
# spans the year 2022 and is organized into 12 separate CSV files, each
# corresponding to a specific month.

## Data Attributes
# The core attributes of the dataset include:
# Ride ID: A unique identifier for each bike trip.
# Equipment Type: Denotes the type of bike used, categorized as casual or electric.
# Start and End Timestamps: Indicate the date and time when each bike trip commenced
# and concluded.
# Start and End Stations: Include station names, IDs, and geographical coordinates.
# Rider Type: Classifies users as either members or casual riders.

## Data Processing and Cleaning
# Reading and Combining Monthly Datasets:
# The initial step involved loading the monthly datasets into R using the 'read.csv'
# function and subsequently combining them into a consolidated dataset ('all_months')
# using 'rbind'.

# Adding Date-related Columns
# To facilitate temporal analysis, date-related columns were added, including
# ride date, month, day, year, day of the week, and ride duration in seconds.

# Filtering Out Missing Values
# Rows with missing or undefined values in the ride month column were removed to
# ensure data integrity.

# Exploratory Data Visualization
# Exploratory graphs were created to provide an initial understanding of the
# data distribution and trends. A comparative bar chart was generated to
# illustrate the number of rides by members and casual riders across different
# months.


### More visualizations to support the analyses
# On the Comparison by Month graph, it is evident that both casual riders and
# members experience an increase in bike usage during the warmer months.
```

```
# Members demonstarte a more consistent pattern of usage across all months.

# Make sure ride_day_of_week is a factor with levels starting from Monday
all_months$ride_day_of_week <- factor(all_months$ride_day_of_week,
                                      levels = c("Monday", "Tuesday", "Wednesday",
                                                 "Thursday", "Friday", "Saturday",
                                                 "Sunday"))

# Plot the comparison by day of the week
# Comparison by Day of the Week
ggplot(all_months, aes(x = ride_day_of_week, fill = member_casual)) +
  geom_bar(position = "dodge", color = "white", size = 0.2) +
  labs(title = "Number of Rides by Members and Casual Riders by Day of the Week",
       x = "Day of the Week",
       y = "Number of Rides") +
  scale_fill_manual(values = custom_palette) +
  scale_y_continuous(labels = scales::label_comma(), expand = c(0, 0)) +
  theme_minimal()
```




```
# Casual riders exhibit increased bike usage on weekends, particularly on Saturdays.
# On the other hand, members consistently show higher usage throughout the week
# compared to casual rider and increased usage during weekdays than weekends.

# Find the difference counts of casual and member riders
difference_by_month <- all_months %>%
  group_by(ride_month, member_casual) %>%
  summarise(ride_count = n()) %>%
  spread(member_casual, ride_count) %>%
  mutate(difference = casual - member)
```

```
## `summarise()` has grouped output by 'ride_month'. You can override using the
## `.groups` argument.
```

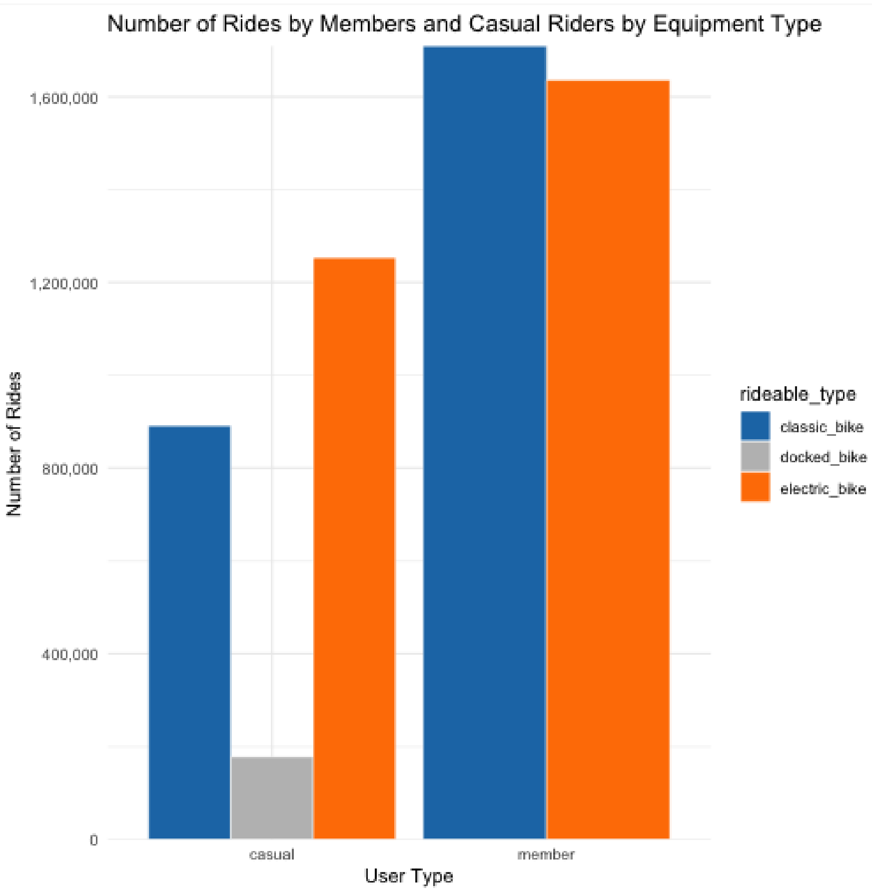
```
# Print the result
print(difference_by_month)
```

```
## # A tibble: 12 × 4
## # Groups:   ride_month [12]
##    ride_month casual member difference
##    <ord>       <int>  <int>      <int>
##  1 Jan         18520  85250     -66730
##  2 Feb         21416  94193     -72777
##  3 Mar         89882 194160    -104278
##  4 Apr        126417 244832    -118415
##  5 May        280415 354443     -74028
##  6 Jun        369051 400153     -31102
##  7 Jul        406055 417433     -11378
##  8 Aug        358924 427008     -68084
##  9 Sep        296697 404642    -107945
## 10 Oct        208989 349696    -140707
## 11 Nov        100772 236963    -136191
## 12 Dec         44894 136912     -92018
```

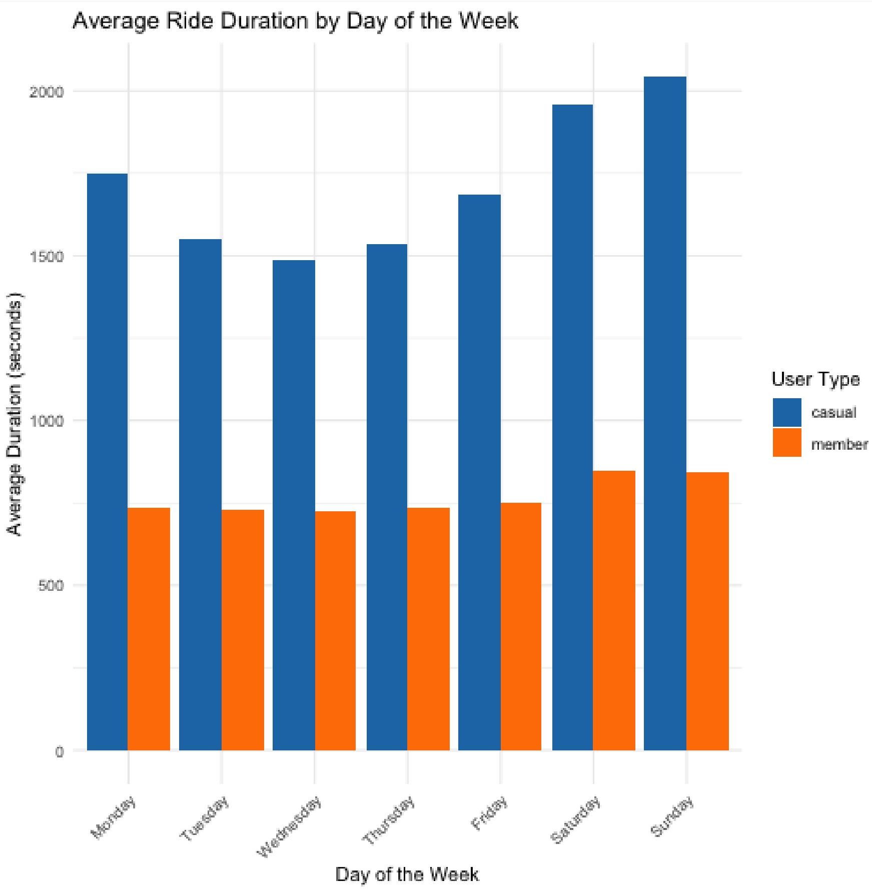
```
# Calculate the total difference
overall_total_difference <- sum(difference_by_month$difference, na.rm = TRUE)

# Print the results
print(overall_total_difference)
```

```
## [1] -1023653
```

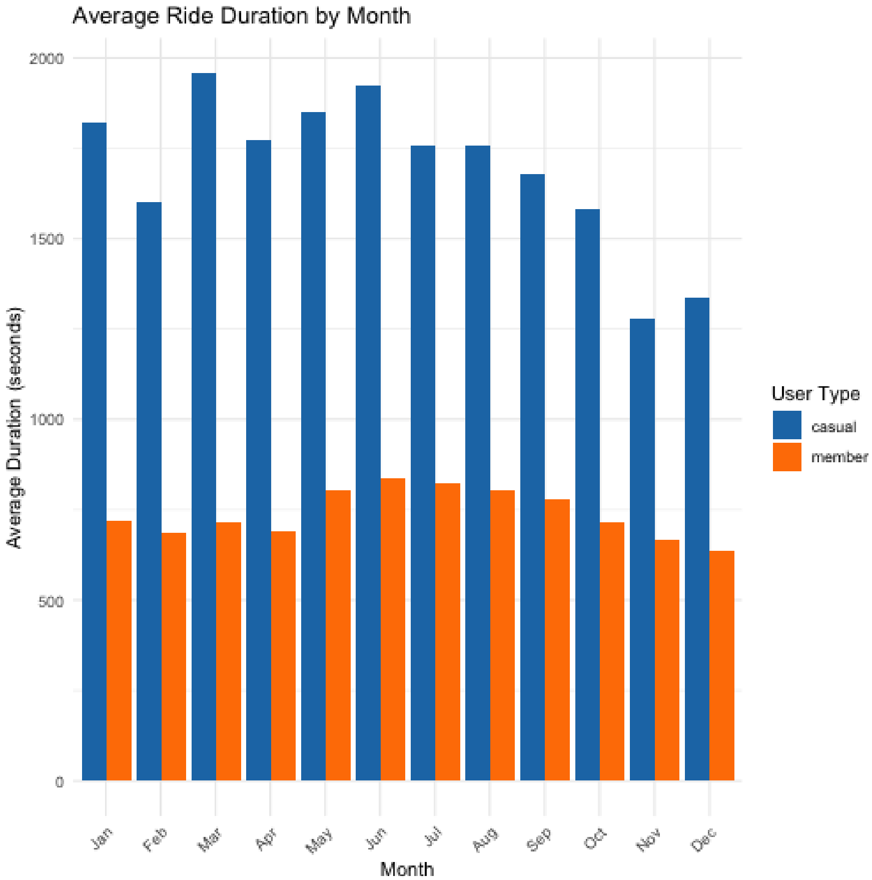
```
# Comparison by Equipment Type
ggplot(all_months, aes(x = member_casual, fill = rideable_type)) +
  geom_bar(position = "dodge", color = "white", size = 0.2) +
  labs(title = "Number of Rides by Members and Casual Riders by Equipment Type",
       x = "User Type",
       y = "Number of Rides") +
  scale_fill_manual(values = c("#1f78b4", "grey", "#ff7f00")) +
  scale_y_continuous(labels = scales::label_comma(), expand = c(0, 0)) +
  theme_minimal()
```



```
# Electric bikes are preferred by a higher proprotion of casual riders compared
# to members.
# Members show a more balanced distribution between classic and electric bikes.

# Convert duration to numeric
all_months$duration_numeric <- as.numeric(all_months$duration)
# Filter out rows with missing or infinite values in duration_numeric
all_months <- all_months[!is.na(all_months$duration_numeric) &
                           is.finite(all_months$duration_numeric), ]
# Create a bar plot for average ride duration by day of the week and user type
ggplot(all_months, aes(x = factor(ride_day_of_week, levels = c("Monday",
                                                                "Tuesday", "Wednesday",
                                                                "Thursday", "Friday",
                                                                "Saturday", "Sunday")),
                        y = duration_numeric, fill = member_casual)) +
  geom_bar(stat = "summary", fun = "mean", position = "dodge") +
  labs(title = "Average Ride Duration by Day of the Week",
       x = "Day of the Week",
       y = "Average Duration (seconds)",
       fill = "User Type") +
  scale_fill_manual(values = c(custom_palette)) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
# Casual riders are shown to have more durations than members daily.


# Create a bar plot for average ride duration by month and user type
ggplot(all_months, aes(x = ride_month, y = duration_numeric, fill = member_casual)) +
  geom_bar(stat = "summary", fun = "mean", position = "dodge") +
  labs(title = "Average Ride Duration by Month",
       x = "Month",
       y = "Average Duration (seconds)",
       fill = "User Type") +
  scale_fill_manual(values = c(custom_palette)) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
###Conclusion
# Below are the key findings derived from the detailed analysis of Cyclistic's
# histrorical bike trip data, providing valuable insights into the distincive
# patters of annual members and casual riders:
# 1. Usage Patterns by Month: Casual riders reached their pinnacle of riding
# activity in July, indicating a strong preference for bike usage during the warmer
# months. In contrast, member riders observed their peak in August. Both groups
# experienced a dip in activity during January, marking the lowest point in ride
# count for the entire year.

# 2. Ride Count Disparity: A significant difference of one million rides was observed
# between registered member rides and casual riders. Surprisingly, despite the lower
# ride count, casual riders outpaced member riders in terms of overall ride duration.
# Casual riders spent more than twice the amount of time on their rides, and the
# monthly average for maximum ride length among casual riders notably exceeded that
# of member riders.
# Annual members tended to have shorter average ride durations indicating a more
# efficient and purpose-driven use of the bikes for commuting.
# Casual riders, with longer average ride durations, suggested a more relaxed and
# exploratory approach to bike usage.

# 3. Day of Highest Activity: Annual members consistently exhibited a more uniform
# usage pattern throughout the week, with higher activity levels on weekdays,
# suggesting a strong reliance on Cyclistic bikes for commuting purposes.
# Casual riders, on the other hand, displayed a peak in activity during weekends,
# indicating a preference for recreational or leisurely rides.

# 4. There was a notable difference in the choice of equipment between annual
# members and casual riders. While annual members opted for both classic bikes and
# electrik bikes, casual riders showed a higher preference for electric bikes.

# These insights can guide the marketing strategy of Cyclistic, especially in the
# context of converting casual riders into annual members. To capitalize on the
# differences in user behavior, targeted campaigns can be designed to appea; to
# the specific preferences and needs of each group. For example, promoting the
# convenience of annual membership for daily commuting or emphasizing the joy of
# leisurely rides for casual riders.
```