CYCLISTIC CASE STUDY

Dante Shoghanian

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.

Business Task

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

Prepare/Process Data

- To analyze and identify trends, Cyclistic has provided historical trip data in the form of .csv spreadsheets covering the period from May 2020 to April 2021. The data is internal and first-party.
- The data contains de-identified User IDs, user types, bike types, and details regarding the start and end of each ride. These details include the time of the ride, the position of the bike, and the names and IDs of the stations used.
- The original files are stored in a separate directory, and copies are created for each dataset. This is done to ensure that the originals are available for reference if needed.
- To prepare, process, and analyze the large datasets, R via RStudio will be used.

Prepare/Process Data

```
#==========================================================================
# STEP-1: Installation of packages & change of directory
#==========================================================================
library(tidyverse) #helps wrangle data
library(lubridate) #helps wrangle data attributes
library(ggplot2) #helps to visualize data
getwd() #displays your working directory
setwd("C:/Users/TEMP/Downloads/GDA/Cyclistic Bike-Share") #sets your working directory to simplify calls to data
#setwd() should be used in desktop version of R
```

Prepare/Process Data

```
#============================================================================
======
# STEP-2: Import data into R
#============================================================================
======
# read_csv() imports data from .csv files
m5_2020 <- read_csv("202005-divvy-tripdata.csv")
m6_2020 <- read_csv("202006-divvy-tripdata.csv")
m7_2020 <- read_csv("202007-divvy-tripdata.csv")
m8_2020 <- read_csv("202008-divvy-tripdata.csv")
m9_2020 <- read_csv("202009-divvy-tripdata.csv")
m10_2020 <- read_csv("202010-divvy-tripdata.csv")
m11_2020 <- read_csv("202011-divvy-tripdata.csv")
m12_2020 <-
read_csv("202012-divvy-tripdata.csv")
m1_2021 <-
read_csv("202101-divvy-tripdata.csv")
m2_2021 <-
read_csv("202102-divvy-tripdata.csv")
m3_2021 <-
read_csv("202103-divvy-tripdata.csv")
m4_2021 <-
read_csv("202104-divvy-tripdata.csv")
```

Prepare/Process Data

Step 3: Wrangle Data and Combine into a Single Data Frame

Compare column names and data types and consolidate/make consistent

column names and data types are shown upon import of each .csv file | or use str() function

Drop the following columns:

- start_station_name
- start_station_id
- end_station_name
- end_station_id

Why? Numerous rows show incomplete entries and similar identifying information can be gleaned from the start_lat, start_lng, end_lat, end_lng column entries

Prepare/Process Data

```
m4_2021 <- m4_2021 %>% select(-c(start_station_name, start_station_id, end_station_name, end_station_id))

m3_2021 <- m3_2021 %>% select(-c(start_station_name, start_station_id, end_station_name, end_station_id))

m2_2021 <- m2_2021 %>% select(-c(start_station_name, start_station_id, end_station_name, end_station_id))

m1_2021 <- m1_2021 %>% select(-c(start_station_name, start_station_id, end_station_name, end_station_id))

m12_2020 <- m12_2020 %>% select(-c(start_station_name, start_station_id, end_station_name, end_station_id))

m11_2020 <- m11_2020 %>% select(-c(start_station_name, start_station_id, end_station_name, end_station_id))

m10_2020 <- m10_2020 %>% select(-c(start_station_name, start_station_id, end_station_name, end_station_id))

m9_2020 <- m9_2020 %>% select(-c(start_station_name, start_station_id, end_station_name, end_station_id))

m8_2020 <- m8_2020 %>% select(-c(start_station_name, start_station_id, end_station_name, end_station_id))

m7_2020 <- m7_2020 %>% select(-c(start_station_name, start_station_id, end_station_name, end_station_id))

m6_2020 <- m6_2020 %>% select(-c(start_station_name, start_station_id, end_station_name, end_station_id))

m5_2020 <- m5_2020 %>% select(-c(start_station_name, start_station_id, end_station_name, end_station_id))
```

Prepare/Process Data

```
# Combine data frames

all_trips <- bind_rows(m5_2020, m6_2020, m7_2020, m8_2020, m9_2020,
m10_2020, m11_2020, m12_2020, m1_2021, m2_2021, m3_2021, m4_2021)
```

Prepare/Process Data

Step 4: Further Clean Up and Add Data to Prepare for Analysis

To ensure that the correct number of observations is present, it is necessary to check the data. Additionally, we can add columns to the dataset that include the date, month, day, day of the week, and year of each ride. This will enable us to aggregate data beyond the ride level.

Add a "ride_length" calculation to all_trips (in seconds)

Prepare/Process Data

```
#=============================================================================

# Check unique output values generated

#=============================================================================

# table(all_trips$member_casual) #results in either "member" or "casual"

# table(all_trips$rideable_type) #results in either "classic_bike", "docked_bike", or "electric_bike"

#=============================================================================

# Add data

#=============================================================================

all_trips$date <- as.Date(all_trips$started_at)

all_trips$month <- format(as.Date(all_trips$date), "%m")

all_trips$day <- format(as.Date(all_trips$date), "%d")

all_trips$year <- format(as.Date(all_trips$date), "%Y")

all_trips$day_of_week <- format(as.Date(all_trips$date), "%A")

all_trips$ride_length <- difftime(all_trips$ended_at, all_trips$started_at)
```

Prepare/Process Data

The rideable_type "docked_bike" represents bikes that have been removed from circulation by Cyclistic for quality control purposes. In addition, there are some entries in which the ride_length field returns a negative duration. To clean up the data, we need to exclude these entries from our dataframe. By doing so, we can reduce the total number of rows from 3,742,202 to 1,243,579.

```
all_trips_v2 <- all_trips[!(all_trips$rideable_type == "docked_bike" |
all_trips$ride_length<0),]
```

Analysis

INPUT

```
all_trips_v2 %>%
     group_by(member_casual) %>%
     summarise(number_of_rides = n()
               ,average_duration = mean(ride_length))
```

OUTPUT

```
## A tibble: 2 x 3
#  member_casual number_of_rides average_duration
#  <chr>                   <int>            <dbl>
#1 casual                 425600            1480.
#2 member                 817979             839.
```

Analysis

INPUT

```
all_trips_v2 %>%
     group_by(member_casual, rideable_type) %>%
     summarise(number_of_rides = n())
```

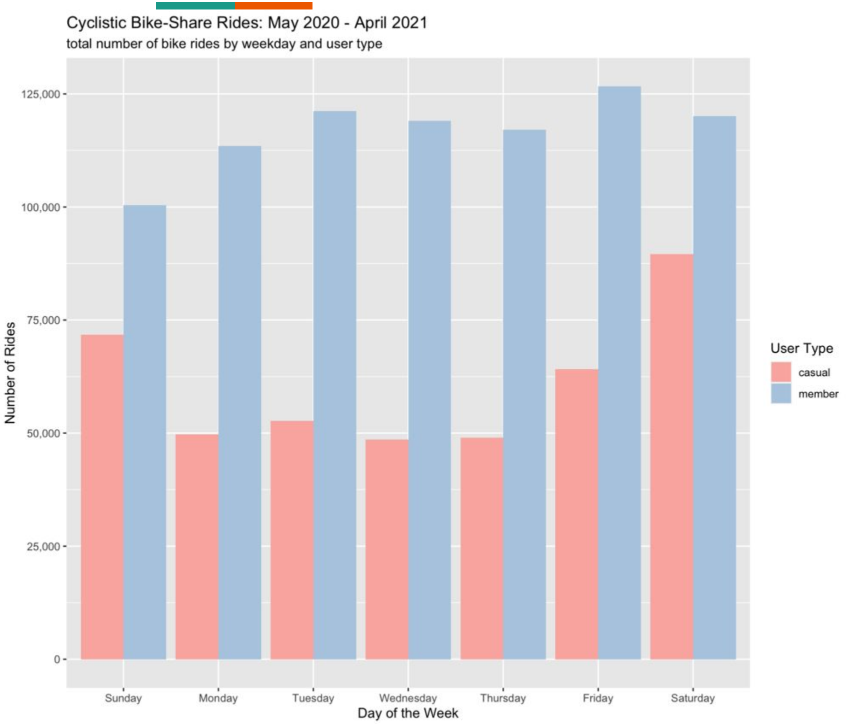
OUTPUT

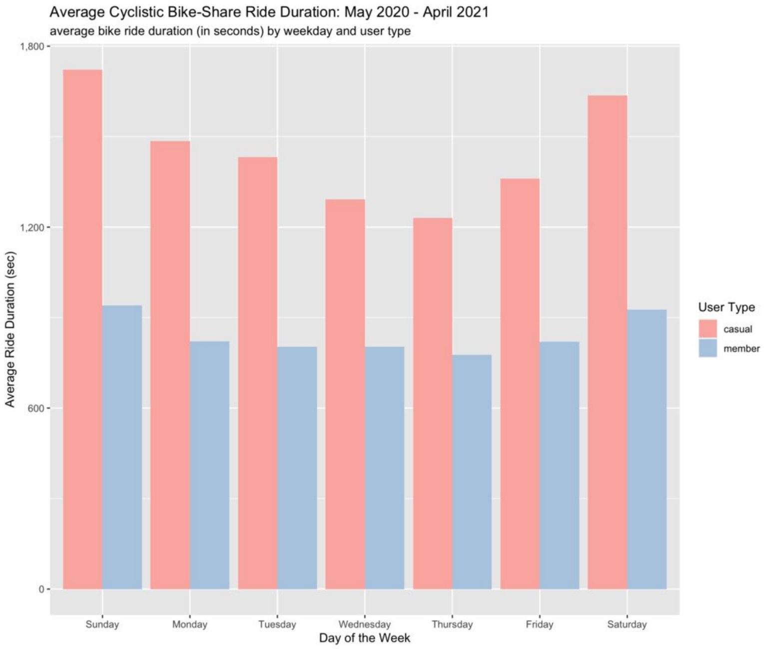
```
#`summarise()` has grouped output by 'member_casual'. You can override using the `.groups` argument.
## A tibble: 4 x 3
## Groups:   member_casual [2]
#  member_casual rideable_type number_of_rides
#  <chr>         <chr>                   <int>
#1 casual        classic_bike           141576
#2 casual        electric_bike          284024
#3 member        classic_bike           392911
#4 member        electric_bike          425068
```

Key Findings

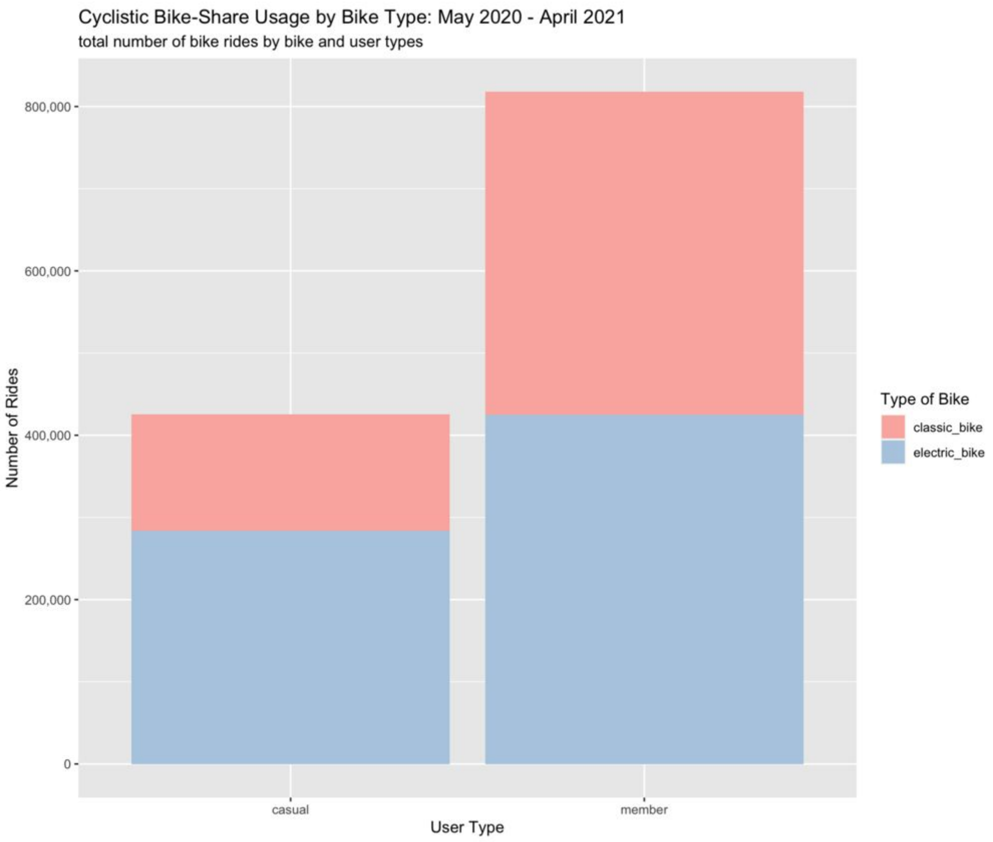
- According to the data, users who have an annual membership tend to complete more rides than casual riders.
- The data shows that among casual riders, Fridays, Saturdays, and Sundays are the most popular days for using Cyclistic bikes.
- On average, casual riders spend about 76% more time on their rides than users who have an annual membership, according to the data.
- The data indicates that approximately 67% of total bike rides made by casual riders were on electric bikes, whereas for users with an annual membership, the figure is about 52%.

Key Findings





Key Findings



Recommendations

- To introduce a new, more affordable annual membership option, the proposal suggests allocating a set number of rides for a specified time period (such as a week or month), as opposed to the current membership structure that offers unlimited rides but limits each ride to 45 minutes. This approach takes into account the fact that casual riders tend to spend more time on their rides than current members, although their average ride duration is only 1480 seconds (24.67 minutes).
- The proposal suggests implementing a time-limited promotion for annual memberships, which would relax the ride limits on Fridays, Saturdays, and Sundays. This is because these days are the most popular among casual riders who use Cyclistic bikes.
- To cater to the preferences of casual riders, the proposal recommends increasing the inventory of electric bikes as they are more preferred over classic bikes.