**Case Study: How Does a Bike-Share Navigate Speedy Success?**

**Introduction**

Welcome to the Cyclistic bike-share analysis case study! In this case study, I will perform many real-world tasks of a junior data analyst. I will work for a fictional company, Cyclistic, and meet different characters and team members. In order to answer the key business questions, I will follow the steps of the data analysis process: **ask**, **prepare**, **process**, **analyze**, **share**, and **act**.

**Scenario**

I am 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, my team wants to understand how casual riders and annual members use Cyclistic bikes differently. From these insights, my team will design a new marketing strategy to convert casual riders into annual members. But first, Cyclistic executives must approve my 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 offering 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 them to commute to work each day.
* **Lily Moreno:** The director of marketing and my 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. I joined this team six months ago and have been busy learning about Cyclistic’s mission and business goals — as well as how I, 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 very good chance 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 marketing analyst team needs to better 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 me the first question to answer: How do annual members and casual riders use Cyclistic bikes differently?

I will produce a report with the following deliverables:

1. A clear statement of the business task
2. A description of all data sources used
3. Documentation of any cleaning or manipulation of data
4. A summary of my analysis
5. Supporting visualizations and key findings
6. My top three recommendations based on my analysis

The problem I am trying to solve is the lack of insights on how annual members and casual riders use Cyclistic bikes differently. The insights derived from this analysis can drive business decisions in designing marketing strategies aimed at converting casual riders into annual members. Therefore, the identified business task of this case study is to clean, manipulate, analyze, and visualize the Cyclistic historical bike trip data to identify trends and provide recommendations on marketing strategies. The key stakeholders to be considered are Lily Moreno (the director of marketing and my manager), the Cyclistic marketing analytics team (who is responsible for collecting, analyzing, and reporting data that helps guide Cyclistic marketing strategy), and the Cyclistic executive team (who will decide whether to approve the recommended marketing program).

**Prepare**

I will use Cyclistic’s historical trip data (previous 12 months of Cyclistic trip data: https://divvy-tripdata.s3.amazonaws.com/index.html) to analyze and identify trends. (Note: The datasets have a different name because Cyclistic is a fictional company. For the purposes of this case study, the datasets are appropriate and will enable me to answer the business questions. The data has been made available by Motivate International Inc. under this license: https://ride.divvybikes.com/data-license-agreement.) This is public data that I can use to explore how different customer types are using Cyclistic bikes. But note that data-privacy issues prohibit me from using riders’ personally identifiable information. This means that I won’t be able to connect pass purchases to credit card numbers to determine if casual riders live in the Cyclistic service area or if they have purchased multiple single passes.

The data is located on Amazon S3, a service offered by Amazon Web Services (AWS) that provides cloud object storage. The data is organized by year in 2013, two quarters from 2014 to 2017, and quarter from 2018 onward. There are no major issues and problems with bias or credibility in this data. The data is reliable, original, comprehensive, current, and cited. The data has gone through de-identification, which is a process used to wipe data clean of all personally identifying information. Therefore, personally identifying information, such as telephone numbers, names, license plates and license numbers, social security numbers, IP addresses, medical records, email addresses, photographs, account numbers, has been anonymized. To ensure data security and protection from unauthorized access and corruption, safety measures, such as encryption and tokenization, should be put in place.

* **Encryption** uses a unique algorithm to alter data and make it unusable by users and applications that don’t know the algorithm. This algorithm is saved as a “key” which can be used to reverse the encryption; so if I have the key, I can still use the data in its original form.
* **Tokenization** replaces the data elements I want to protect with randomly generated data referred to as a “token.” The original data is stored in a separate location and mapped to the tokens. To access the complete original data, the user or application needs to have permission to use the tokenized data and the token mapping. This means that even if the tokenized data is hacked, the original data is still safe and secure in a separate location.

As such, the data’s licensing, privacy, security, accessibility, and integrity are addressed and verified. The data helps me answer my question as it allows me to analyze and identify trends of Cyclistic bike usage in the previous 12 months.

**Process**

To process the data from dirty to clean, I chose to use Microsoft Excel because spreadsheet applications like Excel have tools that help simplify and speed up the data cleaning process. It is assumed that the data’s integrity have been ensured with no compromise due to data replication, transfer, and manipulation. I have taken these steps to ensure that my data is clean:

1. Downloaded the previous 12 months of Cyclistic trip data (Divvy\_Trips\_2020\_Q1.csv, Divvy\_Trips\_2019\_Q4.csv, Divvy\_Trips\_2019\_Q3.csv, and Divvy\_Trips\_2019\_Q2.csv).
2. Unzipped the files.
3. Created a folder on my desktop or Drive to house the files. Used appropriate file-naming conventions.
4. Created subfolders for the .CSV file and the .XLS or Sheets file so that I have a copy of the original data. Moved the downloaded files to the appropriate subfolder.
5. Launched Excel, opened each file, and chose to Save As an Excel Workbook file. Put it in the subfolder I created for .XLS files.
6. Opened my spreadsheet and created a column called “ride\_length.” Calculated the length of each ride by subtracting the column “started\_at” from the column “ended\_at” (for example, =D2-C2) and formatted as General.
7. Created a column called “day\_of\_week,” and calculated the day of the week that each ride started using the “WEEKDAY” command (for example, =WEEKDAY(C2,1)) in each file. Formatted as General or as a number with no decimals, noting that 1 = Sunday and 7 = Saturday.
8. Proceeded to the analyze step.

**Analyze**

Now that my data is stored appropriately and has been prepared for analysis, I will start putting it to work.

I have organized my data into eight consistent columns across all four csv files to perform analysis on them. The data has been properly formatted to their appropriate data types. The key tasks in the analyze test include aggregating my data so it’s useful and accessible, organize and formatting my data, performing calculations, and identifying trends and relationships. To provide a summary of my analysis, I followed these steps using spreadsheets, SQL, and R:

**Spreadsheets**

Opened my spreadsheet application, then completed the following steps:

1. Where relevant, made columns consistent. Dropped unnecessary fields and retained these eight columns: “quarter”, “started\_at”, “ended\_at”, “ride\_length”, “day\_of\_week”, “member\_casual”, “gender”, and “birthyear”.
2. Cleaned and transformed my data to prepare for analysis. Removed rows in which “started\_at” was later than “ended\_at” or “ride\_length” was negative to verify that my data is clean and ready to analyze.
3. Conducted descriptive analysis.
4. Ran a few calculations in one file to get a better sense of the data layout. Options:
   * Calculate the mean of ride\_length
   * Calculate the max ride\_length
   * Calculate the mode of day\_of\_week
5. Created a pivot table to quickly calculate and visualize the data.
   * Calculated the average ride\_length for members and casual riders (Rows = member\_casual; Values = Average of ride\_length).
   * Calculated the average ride\_length for users by day\_of\_week (Columns = day\_of\_week; Rows = member\_casual; Values = Average of ride\_length).
   * Calculated the number of rides for users by day\_of\_week by adding Count of trip\_id to Values.
6. Opened another file and perform the same descriptive analysis steps. Explore ddifferent seasons to make some initial observations.
7. Once I have spent some time working with the individual spreadsheets, I merged them into a full-year view with the tool I have chosen to use to perform my final analysis: a database and SQL and R Studio.
8. Exported a summary file for further analysis.

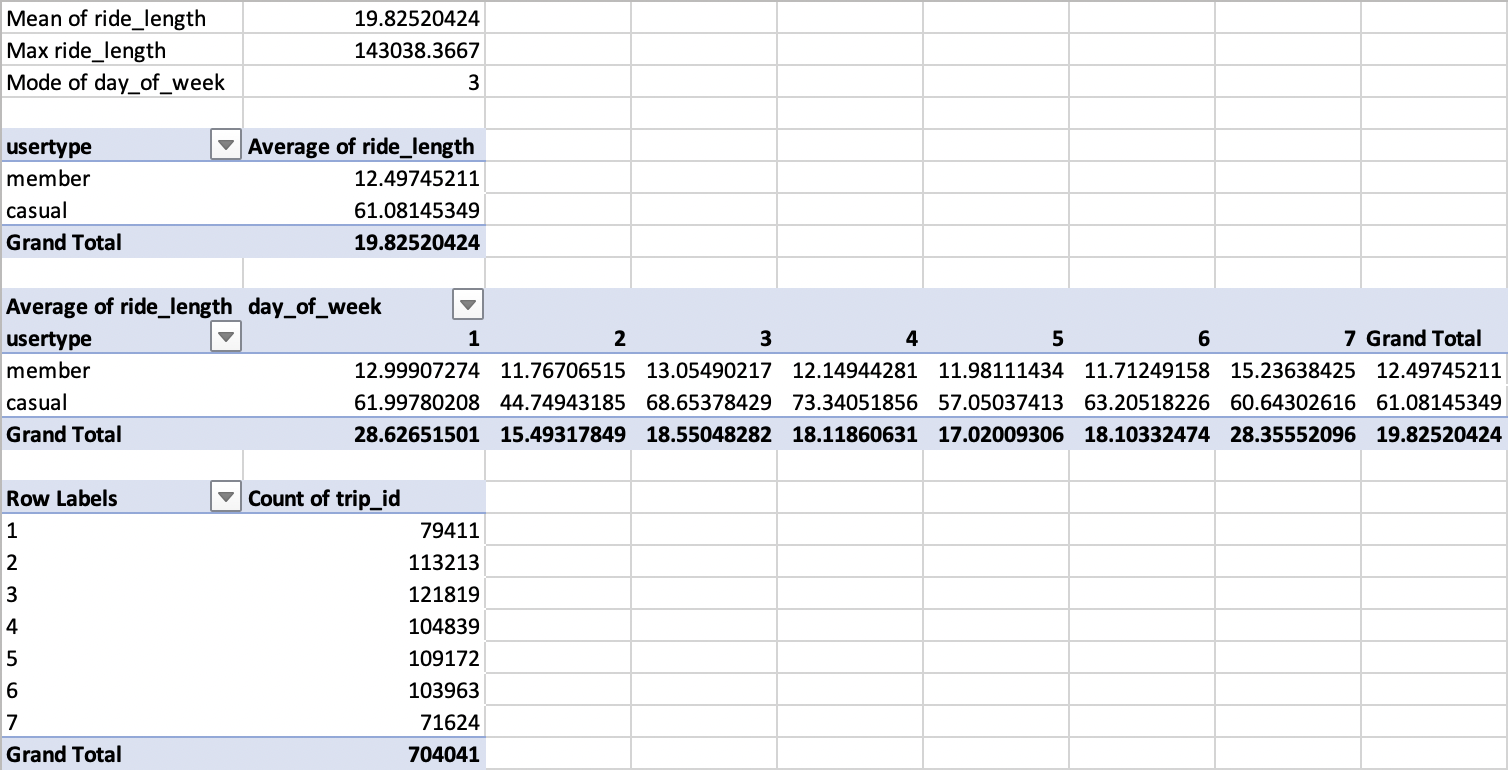
**Descriptive Analyses (Pivot Tables):**

2020 Q1:

Graphical user interface, application, table, Excel

Description automatically generated

2019 Q4:

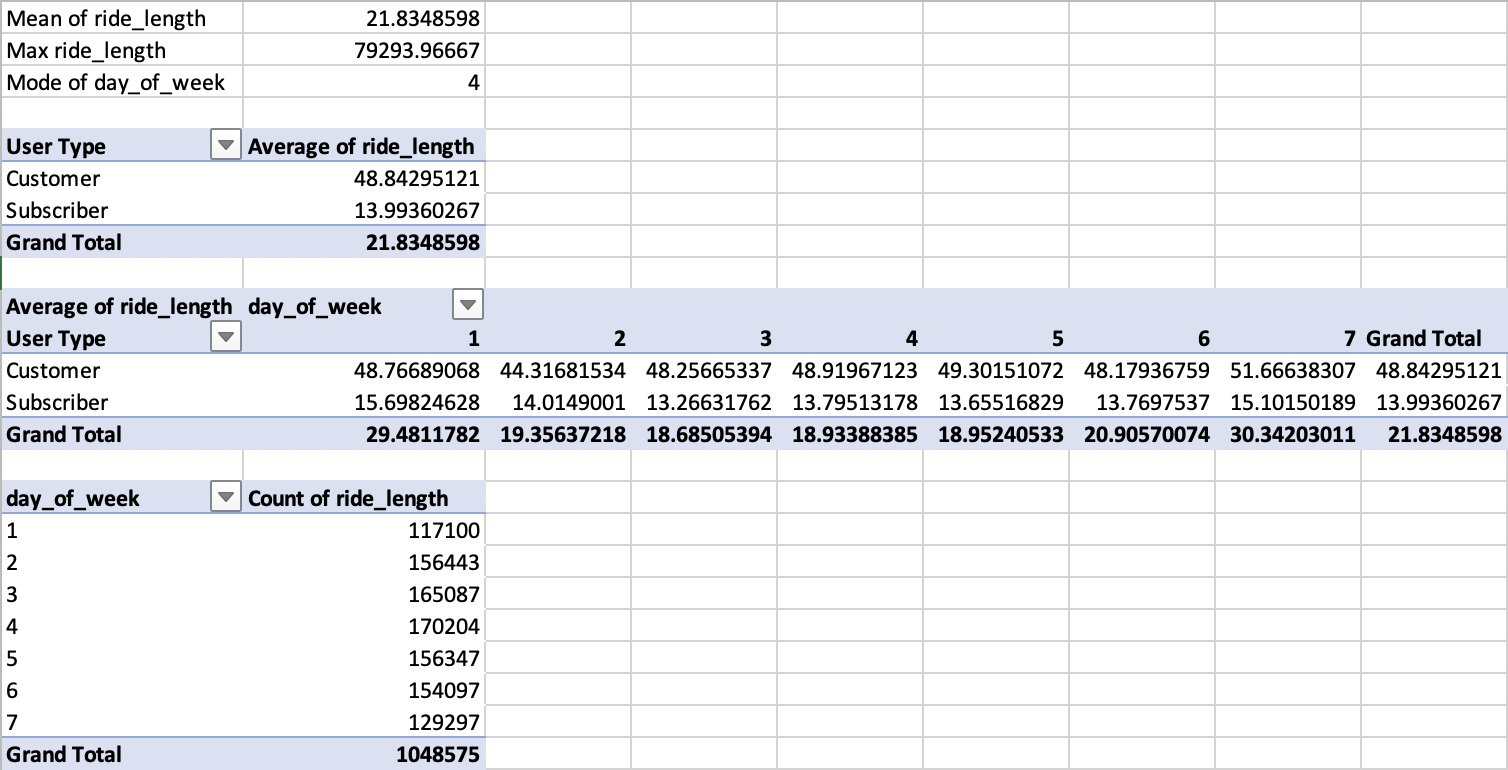


2019 Q3:

Graphical user interface, application, table, Excel

Description automatically generated

2019 Q2:



**SQL**

Opened Google BigQuery, then completed the following steps:

1. Imported my data.
2. Explored my data by looking at the total number of rows, distinct values, maximum, minimum, or mean values.
3. Where relevant, used JOIN and UNION ALL statements to combine my relevant data into one table.
4. Created summary statistics.
5. Investigated interesting trends and saved that information to a table.

SELECT quarter, user\_type, COUNT(\*) AS number\_of\_riders

FROM (SELECT \* FROM cyclistic.2019\_q2

     UNION ALL

     SELECT \* FROM cyclistic.2019\_q3

     UNION ALL

     SELECT \* FROM cyclistic.2019\_q4

     UNION ALL

     SELECT \* FROM cyclistic.2020\_q1)

GROUP BY quarter, user\_type

ORDER BY quarter

-----------------------------------------------------------------------------------

SELECT day\_of\_week, user\_type, COUNT(\*) AS number\_of\_riders

FROM (SELECT \* FROM cyclistic.2019\_q2

     UNION ALL

     SELECT \* FROM cyclistic.2019\_q3

     UNION ALL

     SELECT \* FROM cyclistic.2019\_q4

     UNION ALL

     SELECT \* FROM cyclistic.2020\_q1)

GROUP BY day\_of\_week, user\_type

ORDER BY day\_of\_week

-----------------------------------------------------------------------------------

SELECT member\_gender, user\_type, COUNT(\*) AS number\_of\_riders

FROM (SELECT \* FROM cyclistic.2019\_q2

    UNION ALL

    SELECT \* FROM cyclistic.2019\_q3

    UNION ALL

    SELECT \* FROM cyclistic.2019\_q4

    UNION ALL

    SELECT \* FROM cyclistic.2020\_q1)

GROUP BY member\_gender, user\_type

ORDER BY member\_gender

-----------------------------------------------------------------------------------

SELECT user\_type, MAX(ride\_length) AS max\_ride\_length, MIN(ride\_length) AS min\_ride\_length, AVG(ride\_length) AS avg\_ride\_length

FROM (SELECT \* FROM cyclistic.2019\_q2

     UNION ALL

     SELECT \* FROM cyclistic.2019\_q3

     UNION ALL

     SELECT \* FROM cyclistic.2019\_q4

     UNION ALL

     SELECT \* FROM cyclistic.2020\_q1)

GROUP BY user\_type

-----------------------------------------------------------------------------------

SELECT quarter, user\_type, AVG(ride\_length) AS avg\_ride\_length

FROM (SELECT \* FROM cyclistic.2019\_q2

     UNION ALL

     SELECT \* FROM cyclistic.2019\_q3

     UNION ALL

     SELECT \* FROM cyclistic.2019\_q4

     UNION ALL

     SELECT \* FROM cyclistic.2020\_q1)

GROUP BY quarter, user\_type

ORDER BY quarter

-----------------------------------------------------------------------------------

SELECT day\_of\_week, user\_type, AVG(ride\_length) AS avg\_ride\_length

FROM (SELECT \* FROM cyclistic.2019\_q2

     UNION ALL

     SELECT \* FROM cyclistic.2019\_q3

     UNION ALL

     SELECT \* FROM cyclistic.2019\_q4

     UNION ALL

     SELECT \* FROM cyclistic.2020\_q1)

GROUP BY day\_of\_week, user\_type

ORDER BY day\_of\_week

-----------------------------------------------------------------------------------

SELECT member\_gender, user\_type, AVG(ride\_length) AS avg\_ride\_length

FROM (SELECT \* FROM cyclistic.2019\_q2

    UNION ALL

    SELECT \* FROM cyclistic.2019\_q3

    UNION ALL

    SELECT \* FROM cyclistic.2019\_q4

    UNION ALL

    SELECT \* FROM cyclistic.2020\_q1)

GROUP BY member\_gender, user\_type

ORDER BY member\_gender

**R**

Opened R Studio and wrote a script to complete the following steps:

1. Imported my data.
2. Made columns consistent and merged them into a single dataframe.
3. Cleaned up and added data to prepare for analysis.
4. Conducted descriptive analysis.
5. Exported a summary file for further analysis.

### Cyclistic\_Full\_Year\_Analysis ###

# This analysis is based on the Cyclistic case study "How Does a Bike-Share Navigate Speedy Success?". The purpose of this script is to consolidate downloaded Cyclistic data into a single dataframe and then conduct simple analysis to help answer the key question: “In what ways do members and casual riders use Cyclistic bikes differently?”

# # # # # # # # # # # # # # # # # # # # # # #

# Install required packages

# tidyverse for data import and wrangling

# lubridate for date functions

# ggplot for visualization

# # # # # # # # # # # # # # # # # # # # # # #

library(tidyverse)  #helps wrangle data

library(lubridate)  #helps wrangle date attributes

library(ggplot2)  #helps visualize data

getwd() #displays my working directory

setwd("/Users/david7aurelio/Desktop/Data\_Analytics\_Projects/Original\_Data") #sets my working directory to simplify calls to data

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

# STEP 1: COLLECT DATA

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

# Upload Cyclistic datasets (csv files) here

q2\_2019 <- read\_csv("Divvy\_Trips\_2019\_Q2.csv")

q3\_2019 <- read\_csv("Divvy\_Trips\_2019\_Q3.csv")

q4\_2019 <- read\_csv("Divvy\_Trips\_2019\_Q4.csv")

q1\_2020 <- read\_csv("Divvy\_Trips\_2020\_Q1.csv")

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

# STEP 2: WRANGLE DATA AND COMBINE INTO A SINGLE FILE

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

# Compare column names each of the files

# While the names don't have to be in the same order, they DO need to match perfectly before I can use a command to join them into one file

colnames(q3\_2019)

colnames(q4\_2019)

colnames(q2\_2019)

colnames(q1\_2020)

# Rename columns to make them consistent with q1\_2020 (as this will be the supposed going-forward table design for Cyclistic)

(q4\_2019 <- rename(q4\_2019

                   ,ride\_id = trip\_id

                   ,rideable\_type = bikeid

                   ,started\_at = start\_time

                   ,ended\_at = end\_time

                   ,start\_station\_name = from\_station\_name

                   ,start\_station\_id = from\_station\_id

                   ,end\_station\_name = to\_station\_name

                   ,end\_station\_id = to\_station\_id

                   ,member\_casual = usertype))

(q3\_2019 <- rename(q3\_2019

                   ,ride\_id = trip\_id

                   ,rideable\_type = bikeid

                   ,started\_at = start\_time

                   ,ended\_at = end\_time

                   ,start\_station\_name = from\_station\_name

                   ,start\_station\_id = from\_station\_id

                   ,end\_station\_name = to\_station\_name

                   ,end\_station\_id = to\_station\_id

                   ,member\_casual = usertype))

(q2\_2019 <- rename(q2\_2019

                   ,ride\_id = "01 - Rental Details Rental ID"

                   ,rideable\_type = "01 - Rental Details Bike ID"

                   ,started\_at = "01 - Rental Details Local Start Time"

                   ,ended\_at = "01 - Rental Details Local End Time"

                   ,start\_station\_name = "03 - Rental Start Station Name"

                   ,start\_station\_id = "03 - Rental Start Station ID"

                   ,end\_station\_name = "02 - Rental End Station Name"

                   ,end\_station\_id = "02 - Rental End Station ID"

                   ,member\_casual = "User Type"))

# Inspect the dataframes and look for incongruencies

str(q1\_2020)

str(q4\_2019)

str(q3\_2019)

str(q2\_2019)

# Convert ride\_id and rideable\_type to character so that they can stack correctly

q4\_2019 <-  mutate(q4\_2019, ride\_id = as.character(ride\_id)

                   ,rideable\_type = as.character(rideable\_type))

q3\_2019 <-  mutate(q3\_2019, ride\_id = as.character(ride\_id)

                   ,rideable\_type = as.character(rideable\_type))

q2\_2019 <-  mutate(q2\_2019, ride\_id = as.character(ride\_id)

                   ,rideable\_type = as.character(rideable\_type))

# Stack individual quarter's data frames into one big data frame

all\_trips <- bind\_rows(q2\_2019, q3\_2019, q4\_2019, q1\_2020)

# Remove lat, long, birthyear, and gender fields as this data was dropped beginning in 2020

all\_trips <- all\_trips %>%

  select(-c(start\_lat, start\_lng, end\_lat, end\_lng, birthyear, gender, "01 - Rental Details Duration In Seconds Uncapped", "05 - Member Details Member Birthday Year", "Member Gender", "tripduration"))

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

# STEP 3: CLEAN UP AND ADD DATA TO PREPARE FOR ANALYSIS

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

# Inspect the new table that has been created

colnames(all\_trips)  #List of column names

nrow(all\_trips)  #How many rows are in data frame?

dim(all\_trips)  #Dimensions of the data frame?

head(all\_trips)  #See the first 6 rows of data frame.  Also tail(all\_trips)

str(all\_trips)  #See list of columns and data types (numeric, character, etc)

summary(all\_trips)  #Statistical summary of data. Mainly for numerics

# There are a few problems I will need to fix:

# (1) In the "member\_casual" column, there are two names for members ("member" and "Subscriber") and two names for casual riders ("Customer" and "casual"). I will need to consolidate that from four to two labels.

# (2) The data can only be aggregated at the ride-level, which is too granular. I will want to add some additional columns of data -- such as day, month, year -- that provide additional opportunities to aggregate the data.

# (3) I will want to add a calculated field for length of ride since the 2020Q1 data did not have the "tripduration" column. I will add "ride\_length" to the entire dataframe for consistency.

# (4) There are some rides where tripduration shows up as negative, including several hundred rides where Divvy took bikes out of circulation for Quality Control reasons. I will want to delete these rides.

# In the "member\_casual" column, replace "Subscriber" with "member" and "Customer" with "casual"

# Before 2020, Cyclistic used different labels for these two types of riders ... I will want to make our dataframe consistent with their current nomenclature

# N.B.: "Level" is a special property of a column that is retained even if a subset does not contain any values from a specific level

# Begin by seeing how many observations fall under each usertype

table(all\_trips$member\_casual)

# Reassign to the desired values (I will go with the current 2020 labels)

all\_trips <-  all\_trips %>%

  mutate(member\_casual = recode(member\_casual

                                ,"Subscriber" = "member"

                                ,"Customer" = "casual"))

# Check to make sure the proper number of observations were reassigned

table(all\_trips$member\_casual)

# Add columns that list the date, month, day, and year of each ride

# This will allow me to aggregate ride data for each month, day, or year ... before completing these operations I could only aggregate at the ride level

# https://www.statmethods.net/input/dates.html more on date formats in R found at that link

all\_trips$date <- as.Date(all\_trips$started\_at) #The default format is yyyy-mm-dd

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")

# Add a "ride\_length" calculation to all\_trips (in seconds)

# https://stat.ethz.ch/R-manual/R-devel/library/base/html/difftime.html

all\_trips$ride\_length <- difftime(all\_trips$ended\_at,all\_trips$started\_at)

# Inspect the structure of the columns

str(all\_trips)

# Convert "ride\_length" from Factor to numeric so I can run calculations on the data

is.factor(all\_trips$ride\_length)

all\_trips$ride\_length <- as.numeric(as.character(all\_trips$ride\_length))

is.numeric(all\_trips$ride\_length)

# Remove "bad" data

# The dataframe includes a few hundred entries when bikes were taken out of docks and checked for quality by Cyclistic or ride\_length was negative

# I will create a new version of the dataframe (v2) since data is being removed

# https://www.datasciencemadesimple.com/delete-or-drop-rows-in-r-with-conditions-2/

all\_trips\_v2 <- all\_trips[!(all\_trips$start\_station\_name == "HQ QR" | all\_trips$ride\_length<0),]

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

# STEP 4: CONDUCT DESCRIPTIVE ANALYSIS

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

# Descriptive analysis on ride\_length (all figures in seconds)

mean(all\_trips\_v2$ride\_length) #straight average (total ride length / rides)

median(all\_trips\_v2$ride\_length) #midpoint number in the ascending array of ride lengths

max(all\_trips\_v2$ride\_length) #longest ride

min(all\_trips\_v2$ride\_length) #shortest ride

# I can condense the four lines above to one line using summary() on the specific attribute

summary(all\_trips\_v2$ride\_length)

# Compare members and casual users

aggregate(all\_trips\_v2$ride\_length ~ all\_trips\_v2$member\_casual, FUN = mean)

aggregate(all\_trips\_v2$ride\_length ~ all\_trips\_v2$member\_casual, FUN = median)

aggregate(all\_trips\_v2$ride\_length ~ all\_trips\_v2$member\_casual, FUN = max)

aggregate(all\_trips\_v2$ride\_length ~ all\_trips\_v2$member\_casual, FUN = min)

# See the average ride time by each day for members vs casual users

aggregate(all\_trips\_v2$ride\_length ~ all\_trips\_v2$member\_casual + all\_trips\_v2$day\_of\_week, FUN = mean)

# Notice that the days of the week are out of order. Let me fix that.

all\_trips\_v2$day\_of\_week <- ordered(all\_trips\_v2$day\_of\_week, levels=c("Sunday", "Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday"))

# Now, let me run the average ride time by each day for members vs casual users

aggregate(all\_trips\_v2$ride\_length ~ all\_trips\_v2$member\_casual + all\_trips\_v2$day\_of\_week, FUN = mean)

# analyze ridership data by type and weekday

all\_trips\_v2 %>%

  mutate(weekday = wday(started\_at, label = TRUE)) %>%  #creates weekday field using wday()

  group\_by(member\_casual, weekday) %>%  #groups by usertype and weekday

  summarise(number\_of\_rides = n() #calculates the number of rides and average duration

            ,average\_duration = mean(ride\_length)) %>% # calculates the average duration

  arrange(member\_casual, weekday) # sorts

# Let me visualize the number of rides by rider type

all\_trips\_v2 %>%

  mutate(weekday = wday(started\_at, label = TRUE)) %>%

  group\_by(member\_casual, weekday) %>%

  summarise(number\_of\_rides = n()

            ,average\_duration = mean(ride\_length)) %>%

  arrange(member\_casual, weekday)  %>%

  ggplot(aes(x = weekday, y = number\_of\_rides, fill = member\_casual)) +

  geom\_col(position = "dodge")

# Let me create a visualization for average duration

all\_trips\_v2 %>%

  mutate(weekday = wday(started\_at, label = TRUE)) %>%

  group\_by(member\_casual, weekday) %>%

  summarise(number\_of\_rides = n()

            ,average\_duration = mean(ride\_length)) %>%

  arrange(member\_casual, weekday)  %>%

  ggplot(aes(x = weekday, y = average\_duration, fill = member\_casual)) +

  geom\_col(position = "dodge")

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

# STEP 5: EXPORT SUMMARY FILE FOR FURTHER ANALYSIS

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

# Create a csv file that I will visualize in Excel, Tableau, or my presentation software

# N.B.: This file location is for a Mac. If I am working on a PC, I should change the file location accordingly (most likely "C:\Users\YOUR\_USERNAME\Desktop\...") to export the data. Read more here: https://datatofish.com/export-dataframe-to-csv-in-r/

counts <- aggregate(all\_trips\_v2$ride\_length ~ all\_trips\_v2$member\_casual + all\_trips\_v2$day\_of\_week, FUN = mean)

write.csv(counts, file = '~/Desktop/Data\_Analytics\_Projects/Processed\_Data/avg\_ride\_length.csv')

#I am done!

These are the surprises, trends, relationships, and insights I discovered and found in the data to answer my business questions:

These are the surprises, trends, relationships, and insights I discovered and found in the data to answer my business questions:

* While there are more member users than casual users, casual users ride longer than member users on average.
* For casual users, the highest number of users occurs on Saturday, while the longest ride time happens on Friday. For member users, the highest number of users occurs on Tuesday, while the longest ride time happens on Saturday.
* For casual and member users, there are more male than female riders. However, female riders ride longer on average than male riders.

**Share**

Now that I have performed my analysis and gained some insights into my data, I will create visualizations to share my findings. Moreno has reminded me that they should be sophisticated and polished in order to effectively communicate to the executive team. I followed these steps:

1. Took out a piece of paper and a pen and sketched some ideas for how I would visualize the data.
2. Once I chose a visual form, I opened my tool of choice to create my visualization: Tableau.
3. Created my data visualization, remembering that contrast should be used to draw my audience’s attention to the most important insights. Used artistic principles including size, color, and shape.
4. Ensured clear meaning through the proper use of common elements, such as headlines, subtitles, and labels.
5. Refined your data visualization by applying deep attention to detail.

Chart, line chart

Description automatically generated

The data visualization

Chart, line chart

Description automatically generated

The data visualization was able to answer the question of how annual members and casual members use Cyclistic bikes differently. My data tells the story of the trends, relationships, and insights of casual and member users in relation to the quarter, day of week, and gender. My findings relate to my original question as the factors that differentiate casual and member users may guide the future marketing program. My audience is Moreno, the director of marketing and my manager, and the best way to communicate with her is via a presentation during a meeting. Data visualization helps me share my findings and makes my presentation accessible to my audience because the dashboard consolidates the key findings in a concise and intuitive manner.

**Act**

Now that I have finished creating my visualizations, I will act on my findings. I will prepare the deliverables Morena asked me to create, including the three top recommendations based on my analysis.

My final conclusion based on my analysis is annual members and casual members use Cyclistic bikes differently in terms of the number of users, average ride length, day of the week, and gender. While there are more member users than casual users, casual users ride longer than member users on average. For casual users, the highest number of users occurs on Saturday, while the longest ride time happens on Friday. For member users, the highest number of users occurs on Tuesday, while the longest ride time happens on Saturday. For casual and member users, there are more male than female riders. However, female riders ride longer on average than male riders. My team and business could apply my insights by focusing on the characteristics of casual riders to convert them annual members. The next steps I or my stakeholders would take based on my findings are to intensify marketing strategies targeting long bike riders and offer membership deals that appeal to bike users going for long trips. Moreover, as casual riders seem to be keen users during the weekend with the highest number of users occurring on Saturday and the longest ride time happening on Friday, the membership perks should cater toward leisure bike rides, such as complementary food and beverages discounts and promotions near parks. Furthermore, as the pool of annual members is male-dominated, the marketing team may tap into efforts targeted to convert female casual users to members. Additional data I could use to expand on my findings is the exploration of birth year to determine the age demographic of Cyclistic’s customers.

To reiterate, these are my top three recommendations based on my analysis:

1. Intensify marketing strategies targeting long bike riders and offer membership deals that appeal to bike users going for long trips.
2. Configure the membership perks that cater toward leisure bike rides, such as complementary food and beverages discounts and promotions near parks.
3. Tap into efforts targeted to convert female casual users to members.