Case study: How does a bike-share navigate speedy success?

Adrianna

2024-02-27

About

This document is an a case study for the Google Data Analytics Professional Certificate Program. The case study focuses on the Cyclistic bike-share analysis case study. The case study will follow the following six steps of the data analysis process:

1. Ask
2. Prepare
3. Process
4. Analyze
5. Share
6. Act

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.

What is Cyclistic?

Cylistic is 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 the bikes to commute to work each day.

Goal

Design marketing strategies aimed at converting casual riders into annual members.

The Task

In order to achieve the goal at hand, the 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.

The Ask

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?

* I've been tasked with answering the first question.

Deliverable - a report containing the following information:

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 your analysis
5. Supporting visualizations and key findings
6. Your top three recommendations based on your analysis

* The above will be the deliverables of each of the six steps in the data analysis process.

Ask

Guide for Ask phase

*** Problem to be solved:**

Understand differences in member vs casual riders in order to design a marketing strategy that will successfully convert casual riders into members.

*** How insights can drive business decisions:**

By understanding the differences in bike usage and rider behavior among members and casual riders, we can take the casual riders' usage patterns and create a targeted marketing campaign that will reach as many casual riders as possible and make an annual membership more appealing to them. This could be anything from discounts to promotional offers that may only be available as an annual member. If casual users ride frequently enough, these marketing strategies may be worth purchasing an annual membership.

Key Tasks

*** Identify the business task:**

Understand how annual members and casual riders usage behaviors differ so Cyclicstic can develop a targeted marketing campaign that will reach casual riders and convert them to annual members.

*** Consider key stakeholders:**

- **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 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.

Deliverable

A clear statement of the business task:

Understand how annual members and casual riders use Cyclistic bikes differently so the company can create an effective marketing strategy to convert casual members to an annual membership, as annual memberships are found to be more profitable for the company, which will be key to growth.

Prepare

Guide for Prepare phase

* Where is the data located?

The data is provided on an AWS server here. Each data set can be found in a zip file.

* How is the data organized?

The data is organized by all trips recorded with a unique ride ID. Each data set captures one month of ridership. The link also contains trips by quarter. I chose to analyze the data on a monthly basis from the period 05/2020 - 04/2021.

The monthly data sets contain the following attributes:

- ride_id
- rideable_type *(docked bike or electric bike)*
- started_at (date and time)
- ended_at (date and time)
- start_station_name
- start_station_id
- end_station_name
- end_station_id

- start_lat
- start_lng
- end_lat
- end_lng
- member_casual *(member or casual rider)*

*** Potential issues with bias or credibility in this data: Is the data ROCCC (Reliable, Original, Comprehensive, Current Cited)?**

The data is Cyclistic's historical trip data, so it is considered first party data in this scenario and is likely credible. Since the data contains historical trip data, Cyclistic is just recording usage, and no identifiable user information is present, so we can conclude that the likelihood of bias in the trip data relatively low. The data is also fairly current for the monthly data sets I will be using (2020 - 2024). In all, it is reasonable to assume that the data is reliable, original, comprehensive, current, and cited (ROCCC), although it's possible that we may uncover some bias once we begin analyzing the data.

*** How are licensing, privacy, security, and accessibility addressed?**

The data sets have been validated by Motivate International Inc., as Cyclistic is a fictional company for this exercise. Per Motivate International Inc., the data is made available and licensed. The license states that Lyft Bikes and Scooters operates the City of Chicago's Divvy bike sharing service. The City of Chicago permits Lyft to make certain Divvy system data owned by the City of Chicago available to the public.

The data is for public consumption, but data-privacy protects riders' personally identifiable information, where riders are simply referenced by a ride ID.

*** How the data's integrity was verified:**

- After importing the csv files, I checked the data types in each data set for each month and converted columns if needed so they were consistent across all data sets
- Checked for duplicates in the data sets and deleted any rows with duplicate ride IDs
- Deleted any rows where the data didn't make sense (i.e. trip end time < trip start time, which would result in a negative trip duration)

*** How verifying the data's integrity helps answer the question:**

By having consistent data sets that are free of errors, we can compare the data and analyze it over the given period of time to identify trends among the different rider types, including (but not limited to) usage frequency by month or day of week, trip duration, most popular starting and ending points, etc.

*** Potential problems with the data:**

The data sets do not track distance traveled for each trip by riders - this information would be helpful in determining if there are differences in distance traveled between the two rider types.

Additionally, Cyclistic offers various types of bike models, including reclining bikes, hand tricycles, and cargo bikes. It would be valuable to track this information to understand which type of bike the different rider types use more frequently.

Key Tasks

*** Download data and store it appropriately:**

Since this is a personal project, I downloaded the zip files for each month's trip data and unzipped them in a separate folder in my OneDrive. I then created a subfolder to store all the csvs.

*** Identify how it's organized:**

The data is organized by all trips recorded with a unique ride ID. Each data set captures one month of ridership. The link also contains trips by quarter. I chose to analyze the data on a monthly basis from the period 05/2020 - 04/2021.

The monthly data sets contain the following attributes:

- ride_id
- rideable_type (docked bike or electric bike)
- started_at (date and time)
- ended_at (date and time)
- start_station_name
- start_station_id
- end_station_name
- end_station_id
- start_lat
- start_lng
- end_lat
- end_lng
- member_casual (member or casual rider)

*** Sort and filter the data:**

I will use various functions to arrange and filter the data sets if necessary as outlined in this document.

*** Determine the credibility of the data:**

The data is Cyclistic's historical trip data, so it is considered first party data in this scenario and is likely credible. Since the data contains historical trip data, Cyclistic is just recording usage, and no identifiable user information is present, so we can conclude that the likelihood of bias in the trip data relatively low. The data available is also fairly current for the monthly data sets I will be using (2020 - 2021, but all data available is from 2013 - 2024). In all, it is reasonable to assume that the data is reliable, original, comprehensive, current, and cited (ROCCC), although it's possible that we may uncover some bias once we begin analyzing the data.

Deliverable

A description of all data sources used:

I used 12 months worth of trip data, so 12 csv files were pulled into R for analysis. The files contain trip data by month from May 2020 to April 2021.Licensing, privacy, and accessibility matters for the data source were addressed in the guide above.

See code used to prepare the data below:

Variables used/data sets of bike data for 12 months between 5/2020 - 4/2021

tripdata_202005
tripdata_202006
tripdata_202007
tripdata_202008
tripdata_202009
tripdata_202010
tripdata_202011
tripdata_202012
tripdata_202101
tripdata_202002
tripdata_202103
tripdata_202104

```
# Import all csvs for monthly trip data from May 2020 to April 2021 from saved location

tripdata_202005 <- read.csv('C:/Users/Adj/OneDrive/Documents/Google Data Analytics Cert -
↪  Capstone 1/Data Files/csvs/202005-divvy-tripdata.csv')


tripdata_202006 <- read.csv('C:/Users/Adj/OneDrive/Documents/Google Data Analytics Cert -
↪  Capstone 1/Data Files/csvs/202006-divvy-tripdata.csv')


tripdata_202007 <- read.csv('C:/Users/Adj/OneDrive/Documents/Google Data Analytics Cert -
↪  Capstone 1/Data Files/csvs/202007-divvy-tripdata.csv')


tripdata_202008 <- read.csv('C:/Users/Adj/OneDrive/Documents/Google Data Analytics Cert -
↪  Capstone 1/Data Files/csvs/202008-divvy-tripdata.csv')


tripdata_202009 <- read.csv('C:/Users/Adj/OneDrive/Documents/Google Data Analytics Cert -
↪  Capstone 1/Data Files/csvs/202009-divvy-tripdata.csv')


tripdata_202010 <- read.csv('C:/Users/Adj/OneDrive/Documents/Google Data Analytics Cert -
↪  Capstone 1/Data Files/csvs/202010-divvy-tripdata.csv')


tripdata_202011 <- read.csv('C:/Users/Adj/OneDrive/Documents/Google Data Analytics Cert -
↪  Capstone 1/Data Files/csvs/202011-divvy-tripdata.csv')


tripdata_202012 <- read.csv('C:/Users/Adj/OneDrive/Documents/Google Data Analytics Cert -
↪  Capstone 1/Data Files/csvs/202012-divvy-tripdata.csv')


tripdata_202101 <- read.csv('C:/Users/Adj/OneDrive/Documents/Google Data Analytics Cert -
↪  Capstone 1/Data Files/csvs/202101-divvy-tripdata.csv')

tripdata_202102 <- read.csv('C:/Users/Adj/OneDrive/Documents/Google Data Analytics Cert -
↪  Capstone 1/Data Files/csvs/202102-divvy-tripdata.csv')
```

```r
tripdata_202103 <- read.csv('C:/Users/Adj/OneDrive/Documents/Google Data Analytics Cert -
↪  Capstone 1/Data Files/csvs/202103-divvy-tripdata.csv')


tripdata_202104 <- read.csv('C:/Users/Adj/OneDrive/Documents/Google Data Analytics Cert -
↪  Capstone 1/Data Files/csvs/202104-divvy-tripdata.csv')
```

```r
# Check to see if all data sets' are consistent (columns + data types)

str(tripdata_202005)
str(tripdata_202006)
str(tripdata_202007)
str(tripdata_202008)
str(tripdata_202009)
str(tripdata_202010)
str(tripdata_202011)
str(tripdata_202012)
str(tripdata_202101)
str(tripdata_202102)
str(tripdata_202103)
str(tripdata_202104)
```

Process

Guide for Process phase

* What tools were used in this case study and why?

I have chosen to use R for this exercise, as the sheer volume of records in each data set may not be compatible with spreadsheets or other data analysis tools, such as Microsoft Excel or an SQL database. Furthermore, visualizing the data will be integral in getting insights from the data, and R's visualization features will allow me to create sufficient charts/graphs. I will be able to do both the data analysis and data visualization within the same application.

Additionally, I found it much easier to upload the data sets into R by simply reading the raw csv files where the trip data is stored.

* Ensuring the data's integrity:

I imported and reviewed the data sets at a high-level and determined that they are usable, then followed with various data-cleaning techniques.

* Steps taken to ensure that the data is clean:

The remainder of this document will outline the methods I applied to ensure data integrity among the data sets. This includes (but is not limited to) checking for duplicates, data entry errors, blank cells, consistency across data sets, whether the data points are valid and makes sense, etc.

*** Verify the data so that it is clean and ready to analyze:**

The result of each step in the data-cleaning process will be outlined in the documentation after each function is executed.

*** Document the cleaning process so results can be reviewed and share:**

This file will document the cleaning process and all data manipulation performed to get the data sets to a point where they are usable to gain insights from.

Key Tasks

*** Check the data for errors**

*** Choose tools**

*** Transform the data so it can be used effectively**

*** Document the cleaning process**

Deliverable

Documentation of any cleaning or manipulation of data

The code chunks below document all functions used to clean and manipulate the data sets to check for errors and transform the data so that it be used effectively for further analysis. Comments in the code chunks highlight the intent of the functions used to clean and/or manipulate the data. Findings are summarized below:

- There were 3,742,202 rows (trips) in the combined data set over the 12 month period from May 2020 - April 2021, but only 3,741,993 unique rider_ids, indicating that there are duplicate ride_ids in the data set. Duplicate ride_ids were identified and removed from the data set.
- Some data types in the data set needed to be converted so we could later analyze those attributes, particularly data related to dates and time.
- Additional calculations needed to be performed for so we could further analyze the data set - ride length for each user was calculated and the day of the week for which each trip took place were added to the data set as additional columns.
- Some column names were renamed so they are more intuitive.
- Date/time columns were split into granular detail into separate columns by month, date, year, and time, which will enable us to perform further analysis as we progress to the analyze phase.

See code used to process the data below:

```
# Inconsistent data types found in "start_station_id" & "end_station_id" -> 05/2020 thru
↪  11/2020 are "int", 12/2020 thru 04/2021 are "chr"

# Convert data sets with "int" station_ids to "chr"

tripdata_202005$start_station_id <- as.character(tripdata_202005$start_station_id)
```

```r
tripdata_202005$end_station_id <- as.character(tripdata_202005$end_station_id)


tripdata_202006$start_station_id <- as.character(tripdata_202006$start_station_id)
tripdata_202006$end_station_id <- as.character(tripdata_202006$end_station_id)


tripdata_202007$start_station_id <- as.character(tripdata_202007$start_station_id)
tripdata_202007$end_station_id <- as.character(tripdata_202007$end_station_id)
↪


tripdata_202008$start_station_id <- as.character(tripdata_202008$start_station_id)
tripdata_202008$end_station_id <- as.character(tripdata_202008$end_station_id)
↪


tripdata_202009$start_station_id <- as.character(tripdata_202009$start_station_id)
tripdata_202009$end_station_id <- as.character(tripdata_202009$end_station_id)


tripdata_202010$start_station_id <- as.character(tripdata_202010$start_station_id)
tripdata_202010$end_station_id <- as.character(tripdata_202010$end_station_id)


tripdata_202011$start_station_id <- as.character(tripdata_202011$start_station_id)
tripdata_202011$end_station_id <- as.character(tripdata_202011$end_station_id)
```

```r
# Double-check that variable data types are consistent across all data sets

str(tripdata_202005)
str(tripdata_202006)
str(tripdata_202007)
str(tripdata_202008)
str(tripdata_202009)
str(tripdata_202010)
str(tripdata_202011)
str(tripdata_202012)
str(tripdata_202101)
str(tripdata_202102)
str(tripdata_202103)
str(tripdata_202104)
```

```r
# Combine all data sets into one data set so we have one year's worth of data (5/2020 -
↪  4/2021)

tripdata_combined_0520_0421 <- bind_rows(tripdata_202005, tripdata_202006,
↪  tripdata_202007, tripdata_202008, tripdata_202009, tripdata_202010, tripdata_202011,
↪  tripdata_202012, tripdata_202101, tripdata_202102, tripdata_202103, tripdata_202104)

View(tripdata_combined_0520_0421)
```

3,742,202 rows (trips).

```
# Check that data set has the same number of unique Ids as there are rows in the data set

n_unique(tripdata_combined_0520_0421$ride_id)
```

3,742,202 rows in the data set but only 3,741,993 unique rider_ids.

```
# Delete rows with duplicated ride_id

tripdata_combined_0520_0421 <-
↪  tripdata_combined_0520_0421[!duplicated(tripdata_combined_0520_0421$ride_id), ] # [!
↪  ,] creates subset for column condition (records with ride_id NOT duplicated)
```

There are now 374,993 rows in the data set, which is equal to the number of unique ride_ids - duplicates removed.

```
# Convert "started_at" and "ended_at" to date data type

tripdata_combined_0520_0421 <- tripdata_combined_0520_0421 %>%
  mutate(started_at = ymd_hms(started_at),
         ended_at = ymd_hms(ended_at))


# check that "started_at" and "ended_at" are now date data types

str(tripdata_combined_0520_0421)
```

```
# Add column "ride_length" to calculate duration between "started_at" and "ended_at" for
↪  each data set

tripdata_combined_0520_0421 <- tripdata_combined_0520_0421 %>%
  mutate(ride_length = as_hms(difftime(ended_at, started_at, units = "secs"))) # use
  ↪  "as_hms to convert difference in seconds to hours, minutes, seconds duration format
```

```
# Check rows have negative "ride_length" values

nrow(tripdata_combined_0520_0421[tripdata_combined_0520_0421$ride_length <0,]) # returns
↪  count of rows with ride_length <0


# Drop negative ride_length rows from data set

tripdata_combined_0520_0421 <-
↪  tripdata_combined_0520_0421[!tripdata_combined_0520_0421$ride_length <0,] # [! ,]
↪  creates subset for column condition (records with ride_length NOT less than 0)
```

10,297 rows identified with negative ride length values. Dropped them from the data set.

```r
# Add "day of week" column to each data set

tripdata_combined_0520_0421 <- tripdata_combined_0520_0421 %>%
  mutate(weekday = weekdays(started_at))
tripdata_combined_0520_0421$weekday <- ordered(tripdata_combined_0520_0421$weekday,
↪  levels=c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday",
↪  "Sunday"))
```

```r
# Rename "member_casual" column name to "rider_type"

tripdata_combined_0520_0421 <- tripdata_combined_0520_0421 %>%
  rename(rider_type = member_casual)
```

```r
# Split date and time into separate columns for "started_at"

tripdata_combined_0520_0421$date <- as.Date(tripdata_combined_0520_0421$started_at) # add
↪  new column for started date


tripdata_combined_0520_0421$month <- format(as.Date(tripdata_combined_0520_0421$date),
↪  "%b") # use new date column to extract month (%b for abbreviated month)


tripdata_combined_0520_0421$day <- format(as.Date(tripdata_combined_0520_0421$date),
↪  "%d") # use new date column to extract date


tripdata_combined_0520_0421$year <- format(as.Date(tripdata_combined_0520_0421$date),
↪  "%Y") # use new date column to extract year


tripdata_combined_0520_0421$time <-
↪  format(as.POSIXct(tripdata_combined_0520_0421$started_at),
                                                    format = "%H:%M:%S") # add new column
                                                    ↪  for started time
```

Analyze

Guide for Analyze phase

* How should the data be organized to perform analysis on it?

All 12 data sets were combined after each one was cleaned/prepared and ready to merge in the prepare phase. This made it easier to perform any analysis on the data so we can spot trends and find insights from a holistic view of the 12 month time frame for which the data is captured.

* Has the data been properly formatted?

Yes, the data was formatted and transformed for consistency in the *Process* phase. Some of these steps occurred before the data sets were merged (so we could merge the data sets to begin with), and after the

data sets were merged into a single file so we only needed to run code on a single data frame instead of 12 different data frames.

*** Surprises discovered in the data:**

There were inconsistencies in the formatting and data types of some columns across the 12 data sets before they were merged, particularly the date/time columns.

There were some data points that didn't make sense, such as the trip end time for some records were less than the trip start time, resulting in a negative trip duration which is not possible. Those records were removed, as they were likely errors in the data.

*** Trends or relationships found in the data:**

This will be outlined in the code chunks and subsequent summary below.

*** How will these insights help answer the business questions?**

The results of the analysis will uncover differences in how members and casual riders are using Cyclistic's bikes. From there, we can come up with a targeted approach to the casual riders to make a subscription to an annual membership more appealing.

Key Tasks

*** Aggregate your data so it's useful and accessible**

*** Organize and format your data**

*** Perform calculations**

*** Identify trends and relationships**

Deliverable

A summary of my analysis. The graphs and data transformations/manipulations produced the following insights:

- Riders most frequented trips using Cyclistic's bikes on Saturdays over the 12 month period.
- The shortest ride length was 00:00:00. This likely means someone may have checked out a bike but didn't actually remove it from the bike station and travel anywhere.
- The longest ride length was 904:43:21. This is an unusually high number of hours, so it's possible that some riders checked out their bikes and didn't return them for a period of time.
- Streetr Dr & Grande Ave was the most frequented start station among all riders.
- Clark St & Elm St was the most frequented start station among members.
- Streetr Dr & Grande Ave was also the most frequented start station among casuals.
- There were 1,540,112 casual riders and 2,191,584 from the period of May 2020 - April 2021. Casual riders made up 41% of the total rider population and members made up 59% of total ridership over the 12 month period.

The average trip duration for casual riders was 00:43:15 and 00:15:50 for members. Casual riders' average trip duration was almost three times as lengthy as members' ride time.

- More members leveraged Cyclistic bikes M - F, but more casual users rode bikes on weekends. Usage was still highest on the weekends for both types of riders.
- Considering members' overall and daily average trip duration is shorter and that more of them ride bikes on weekdays, this may suggest that more members used Cyclistic to commute to work, while casual members may have used bikes for more leisurely activities. Casual riders' highest average trip duration occurred on weekends. Casual riders' average trip duration was about 3 times the length compared to member ridership for every day of the week.
- Summer months experienced higher ridership for both member and casual riders and gradually decreased in the fall and winter months, likely due to outside weather conditions, where people tend to ride bikes, during the warmer seasons, especially for leisure.

See code used to analyze the data below:

```
# Get summary of combined data set

tripdata_combined_0520_0421 %>%
  summary()
```

Riders most frequented trips using Cyclistic's bikes on Saturdays in the given time period with 699,658 trips.

```
# Find the longest and shortest trip durations in the data set

tripdata_combined_0520_0421 %>%
  summarise(max_ride_length = as_hms(max(ride_length)),
            min_ride_length = as_hms(min(ride_length)))
```

The shortest ride length: 00:00:00.

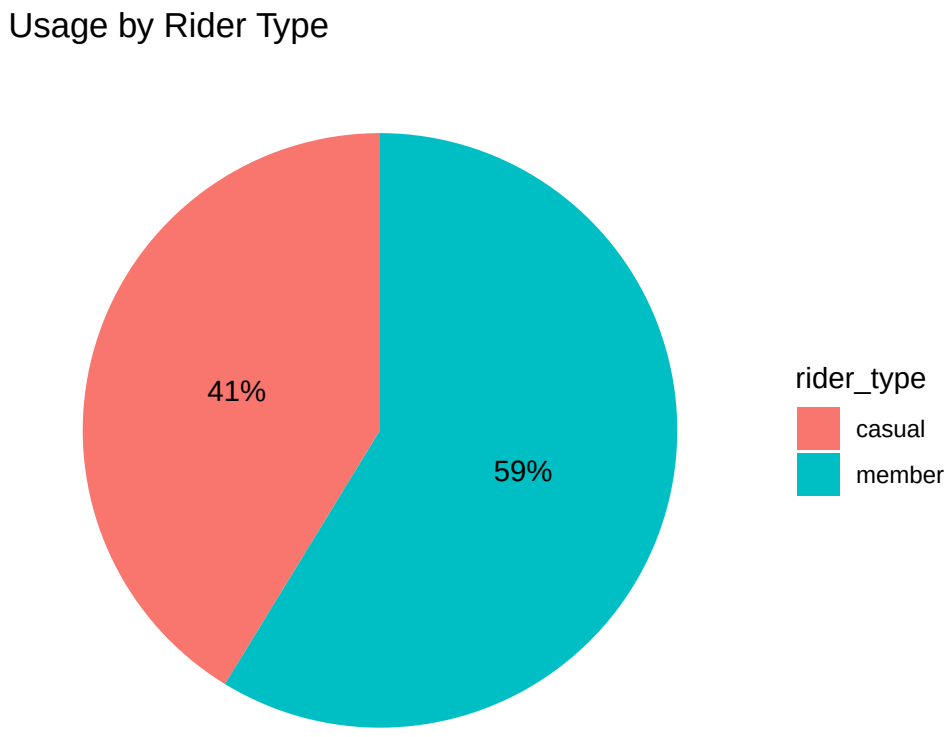
The longest ride length: 904:43:21.

```
# Group casual vs. member riders and average ride length for each type of rider

tripdata_combined_0520_0421_member_type <- tripdata_combined_0520_0421 %>%
  group_by(rider_type) %>%
  summarise(rider_type_count = n(),
            avg_ride_duration = format(strptime(as_hms(mean(ride_length)), format =
            ↪  "%H:%M:%S"), "%H:%M:%S"))
View(tripdata_combined_0520_0421_member_type)
```

```
# Transform rider type count to percentages

tripdata_combined_0520_0421_member_type$rider_type_count <-
↪  as.numeric(tripdata_combined_0520_0421_member_type$rider_type_count) # convert
↪  "rider_type_count" from int to num

tripdata_combined_0520_0421_member_type_percent <- tripdata_combined_0520_0421 %>%
  group_by(rider_type) %>%
```

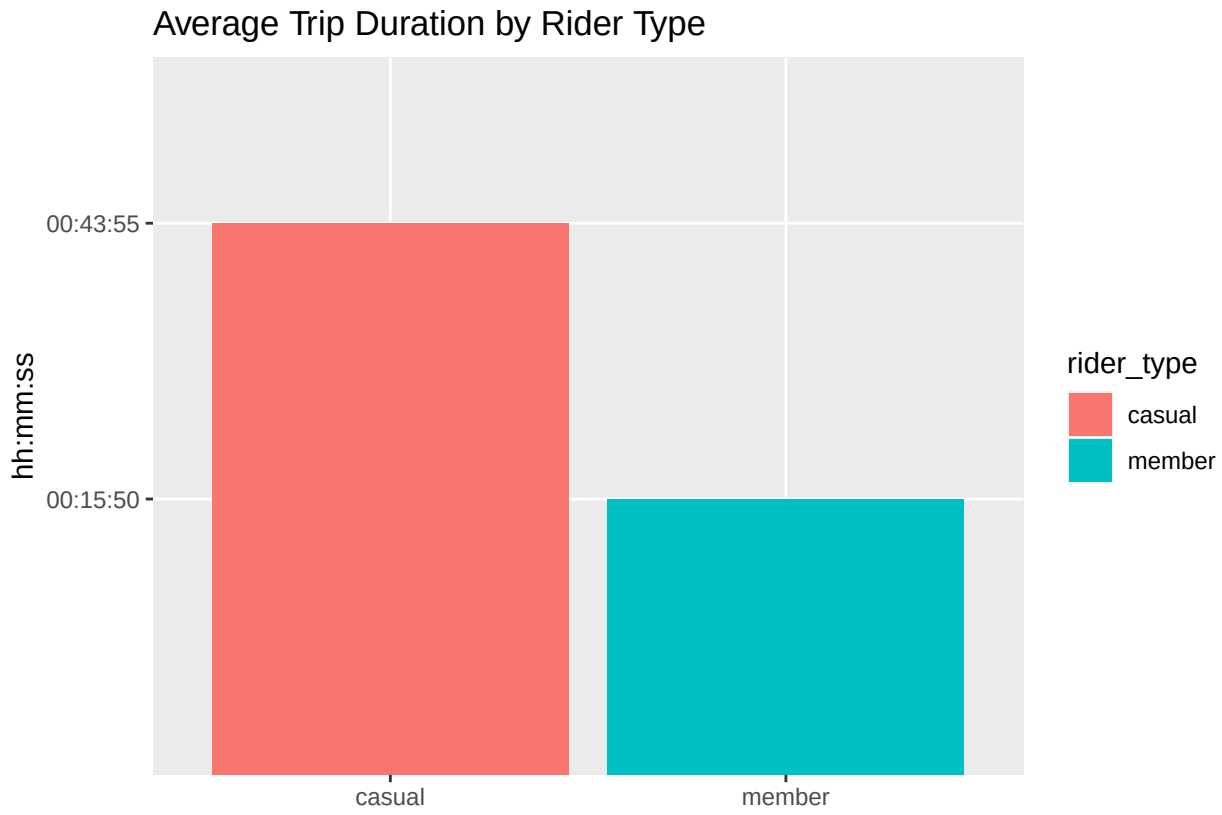
```
  summarise(total = n()) %>%
  mutate(total_riders = sum(total)) %>%
  group_by(rider_type) %>%
  summarise(total_percent = total/total_riders) %>%
  mutate(labels = scales::percent(total_percent))
View(tripdata_combined_0520_0421_member_type_percent)
```

```
# Pie chart to compare % of members to casuals

ggplot(data = tripdata_combined_0520_0421_member_type_percent, aes(x = "", y =
↪  total_percent, fill = rider_type)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) +
  labs(title = "Usage by Rider Type",
       caption = "*ridership by rider type between May 2020 - April 2021") +
  theme_minimal() +
  theme(axis.title.x = element_blank(),
        axis.text.x = element_blank(),
        axis.title.y = element_blank(),
        axis.ticks = element_blank(),
        panel.grid = element_blank()) +
  geom_text(aes(label = labels), position = position_stack(vjust = 0.5))
```



```r
# Visualize average trip duration by rider type side-by-side using a double bar graph

ggplot(data = tripdata_combined_0520_0421_member_type, aes(x = rider_type, y =
↪  avg_ride_duration, fill = rider_type))+
  geom_col(position = "dodge", stat = "count") +
  labs(title = "Average Trip Duration by Rider Type",
       x = "",
       y = "hh:mm:ss")
```

```
## Warning in geom_col(position = "dodge", stat = "count"): Ignoring unknown
## parameters: `stat`
```




```r
# Identify which stations riders most frequently started their trip

most_frequent_station <- tripdata_combined_0520_0421 %>%
  group_by(start_station_name) %>%
  summarise(start_station_name_count = n()) %>%
  filter(start_station_name != "") %>%  # removes records with no start_station_name
  ↪  listed
  arrange(desc(start_station_name_count))
View(most_frequent_station) # number of starts at each station from most to least
↪  frequented
```

```r
# Identify which stations members most frequently started their trip

most_frequent_station_members <- tripdata_combined_0520_0421 %>%
  filter(rider_type == "member") %>%
  group_by(start_station_name) %>%
  summarise(start_station_name_count = n()) %>%
  filter(start_station_name != "") %>%
  arrange(desc(start_station_name_count))
View(most_frequent_station_members) # number of starts at each station from most to least
↪  frequented by members


most_frequent_station_casuals <- tripdata_combined_0520_0421 %>%
  filter(rider_type == "casual") %>%
  group_by(start_station_name) %>%
  summarise(start_station_name_count = n()) %>%
  filter(start_station_name != "") %>%
  arrange(desc(start_station_name_count))
View(most_frequent_station_casuals) # number of starts at each station from most to least
↪  frequented by casuals
```

- All riders: Streetr Dr & Grande Ave
- Members: Clark St & Elm St
- Casuals: Streetr Dr & Grande Ave

```r
# Group rider counts by type and average trip duration by day of the week

tripdata_combined_0520_0421_weekday <- tripdata_combined_0520_0421 %>%
  group_by(weekday) %>%
  summarise(total_riders = n(),
            casual = sum(rider_type == "casual"),
            member = sum(rider_type == "member"),
            avg_ride_duration = format(strptime(as_hms(mean(ride_length)), format =
            ↪  "%H:%M:%S"), "%H:%M:%S"),
            avg_ride_duration_casual = format(strptime(as_hms(mean(ride_length[rider_type
            ↪  == "casual"])), "%H:%M:%S"), "%H:%M:%S"),
            avg_ride_duration_member = format(strptime(as_hms(mean(ride_length[rider_type
            ↪  == "member"])), "%H:%M:%S"), "%H:%M:%S"))
View(tripdata_combined_0520_0421_weekday)
```

```r
# Group by month and rider type

tripdata_combined_0520_0421_month <- tripdata_combined_0520_0421 %>%
  group_by(month) %>%
  summarise(total_riders = n(),
            casual = sum(rider_type == "casual"),
            member = sum(rider_type == "member"),
            avg_ride_duration = format(strptime(as_hms(mean(ride_length)), format =
            ↪  "%H:%M:%S"), "%H:%M:%S")) %>%
  arrange(total_riders)
View(tripdata_combined_0520_0421_month)
```

Share

Guide for Share phase

*** Did the analysis provide any insight into how annual members and casual riders use Cyclistic bikes differently?**

Yes. I was able to create visualizations comparing different variables among member and casual riders.

*** What story does the data tell?**

The data implied that casual riders took significantly longer trips than members. On average, casual riders' trip duration was approximately three times longer than member trip duration.

*** How do the findings relate to the original question?**

I was able to determine with the data that casual riders had a much longer average trip duration compared to members. Additionally, member ridership tended to outpace casual riders only on weekends. This implies that there are differences in user preferences and behaviors among casual and member riders, depending on the day of the week.

*** Who is the audience? What is the best way to communicate with them?**

The audience are all stakeholders involved. They include:

- Lily Moreno: manager and director of marketing - the documentation for which I performed my data analyses and findings will be the best way to communicate my analysis.
- Marketing analytics team - I can share the raw files and documentation with my team, as they may need the data in its raw form for additional analysis down the line.
- Cyclistic executive team - will ultimately approve of my recommendations based on my findings, so I will need to present my findings in a high-level and easily-digestible format (i.e. Powerpoint, PDF, etc.).

*** Can data visualization help you share your findings?**

Yes, data visualizations will provide a visual comparison between casual riders and members. The visualizations should enable stakeholders to easily discern trends and patterns in the data among the two groups.

*** Is the presentation accessible to the audience?**

Presentation materials will be accessible exclusively to stakeholders and the internal team. This markdown file will be publicly accessible.

Key tasks

*** Determine the best way to share findings**

*** Create effective data visualizations**

*** Present indings**

*** Ensure your work is accessible - this document will be available in a format that is TBD.**

Deliverable

Supporting visualizations and key findings (below):

Key findings:

- Top start stations among **members** include:
  1. Clark & Elm St
  2. Broadway & Berry Ave
  3. Wells St & Concord Ln

  Top start stations among **casuals** include:
  1. Streetr Dr & Grande Av
  2. Lake Shore Dr & Monroe St
  3. Millennium Park

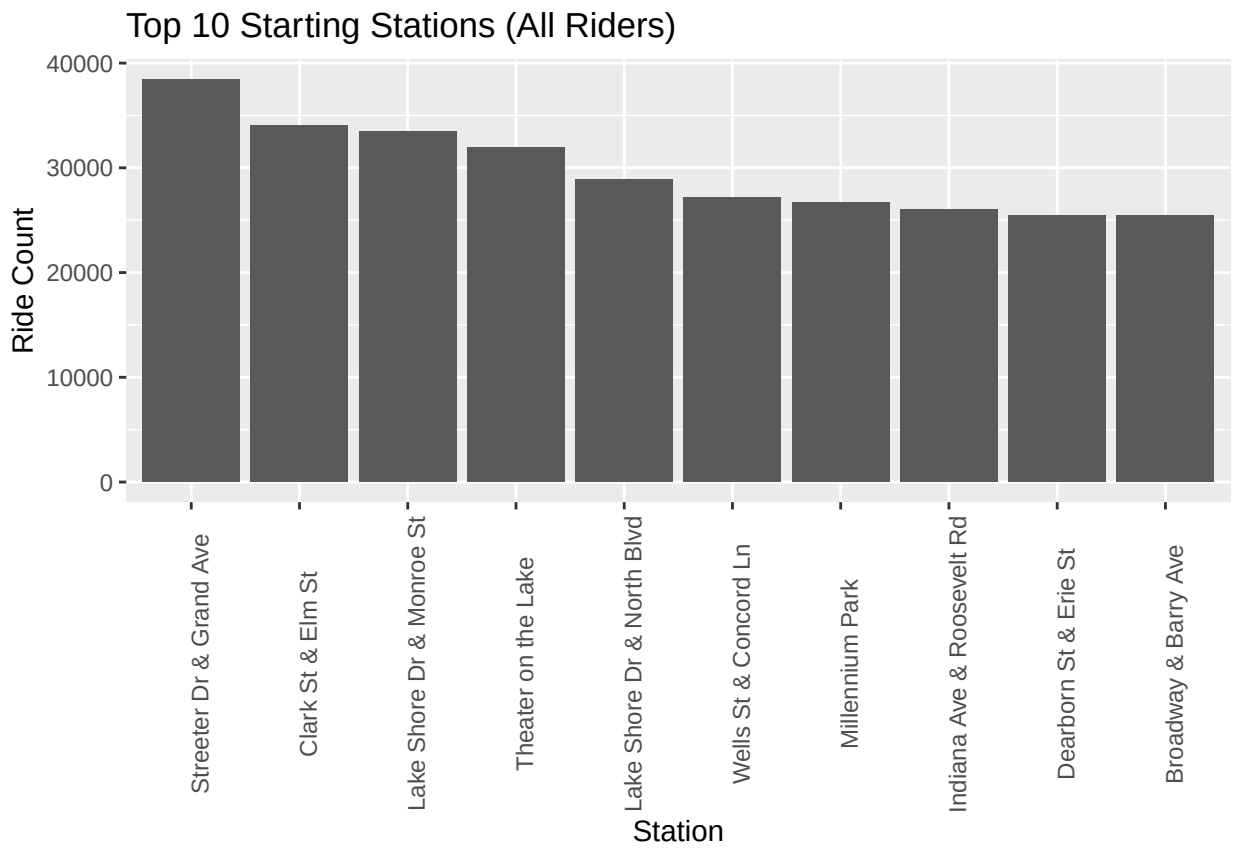
There was no overlap in the top three start stations among members and casual riders. It's possible members and casual riders tended to live in different parts of the city.

- Member ridership outpaced casual ridership on weekdays, but casual ridership was higher than member ridership on weekends. Both groups had highest ridership counts on weekends.
- Considering members' overall and daily average trip duration was shorter and that more of them rode bikes on weekdays, this may suggest that more members used Cyclistic to commute to work, while casual members may have used bikes for more leisurely activities. Casual riders' highest average trip duration occurred on weekends. Casual riders' average trip duration was about 3 times the length compared to member ridership for every day of the week.
- Summer months experienced higher ridership for both member and casual riders and gradually decreased in the fall and winter months, likely due to outside weather conditions, where people tend to ride bikes, during the warmer seasons, especially for leisure.

See code used to share the data below:

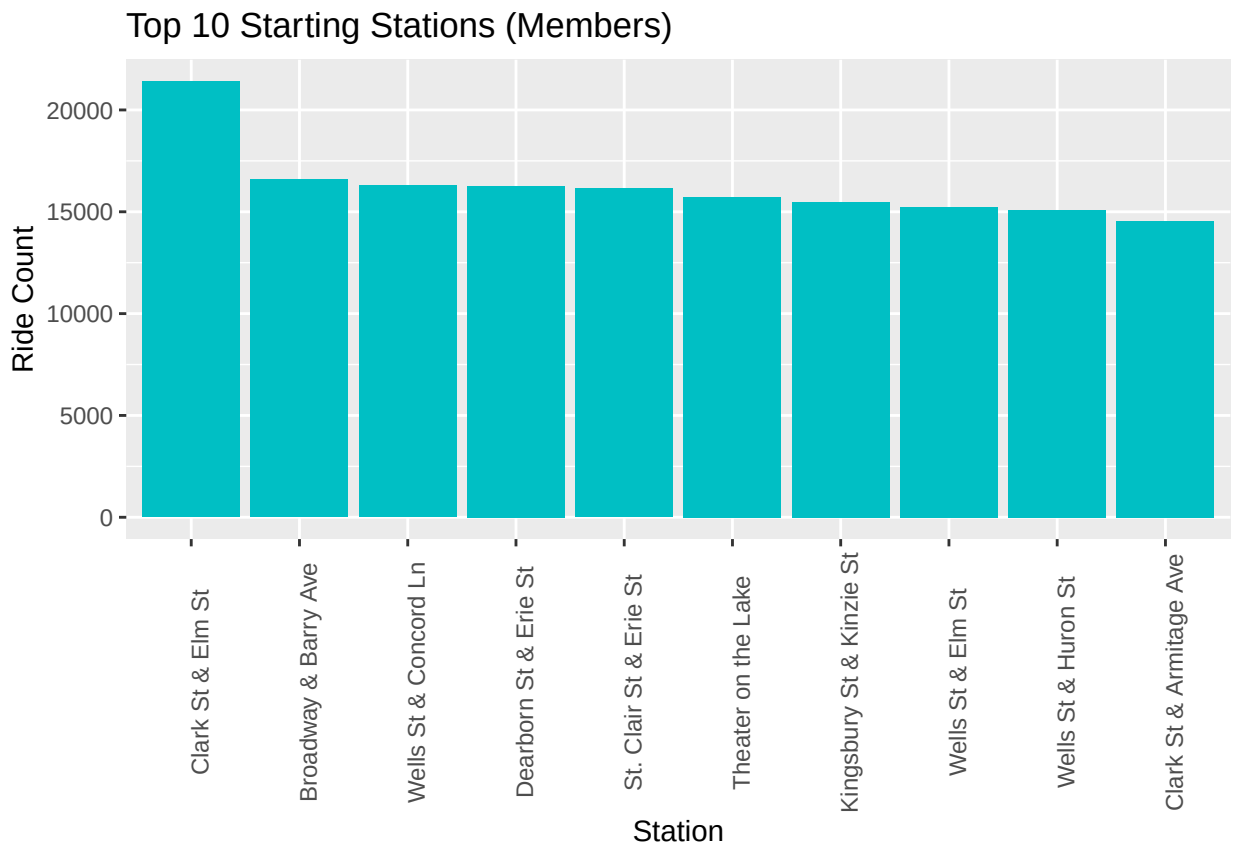
Top 10 most frequented start stations among all riders visualized:

```
most_frequent_station %>%
  top_n(10, start_station_name_count) %>%
  ggplot() +
  geom_bar(mapping = aes(reorder(x = start_station_name, -start_station_name_count), y =
  ↪  start_station_name_count), stat = "identity") +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(title = "Top 10 Starting Stations (All Riders)",
       x = "Station",
       y = "Ride Count")
```



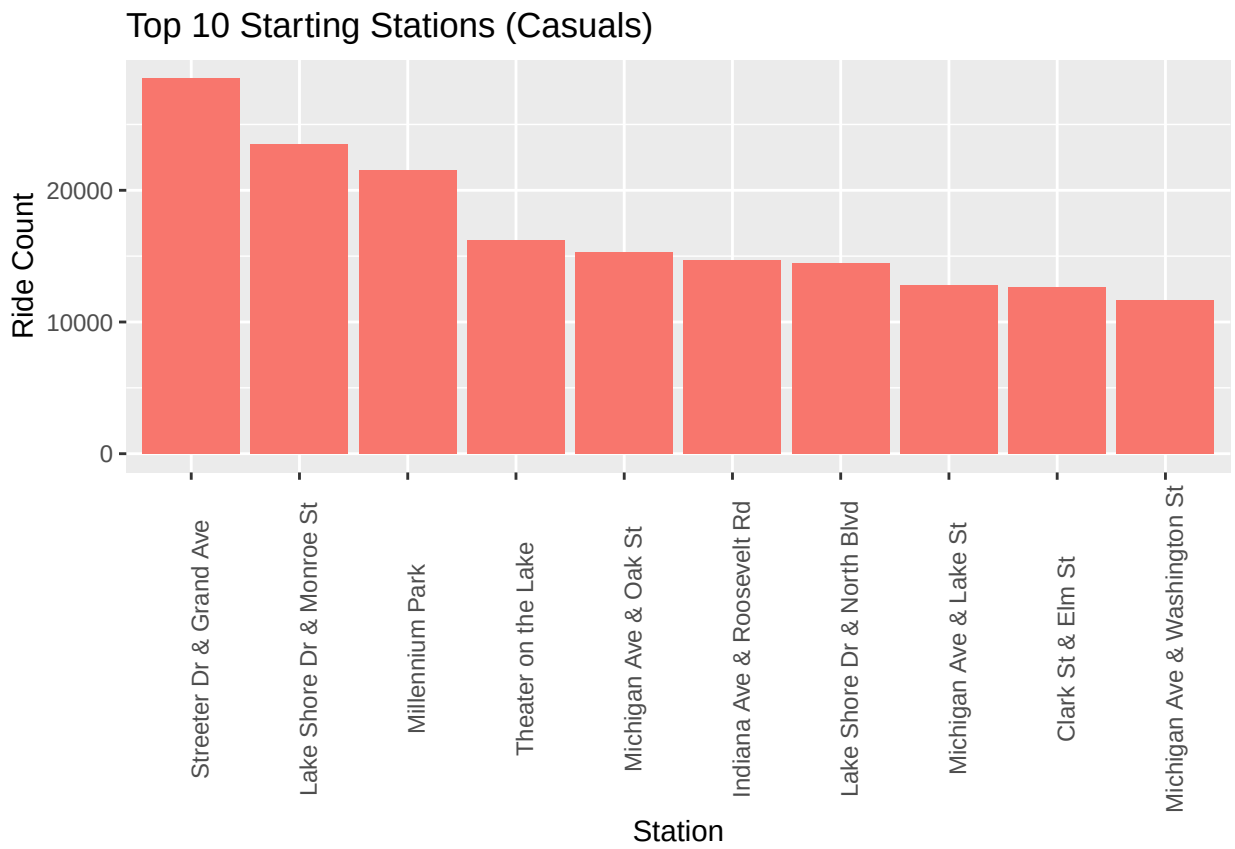
Top 10 most frequented start stations among members:

```
most_frequent_station_members %>%
  top_n(10, start_station_name_count) %>%
  ggplot() +
  geom_bar(mapping = aes(reorder(x = start_station_name, -start_station_name_count), y =
  ↪  start_station_name_count), fill = "#00bfc4", stat = "identity") +
  theme(axis.text.x = element_text(angle = 90), legend.position = "none") +
  labs(title = "Top 10 Starting Stations (Members)",
       x = "Station",
       y = "Ride Count")
```



Top 10 most frequented start stations among casuals:

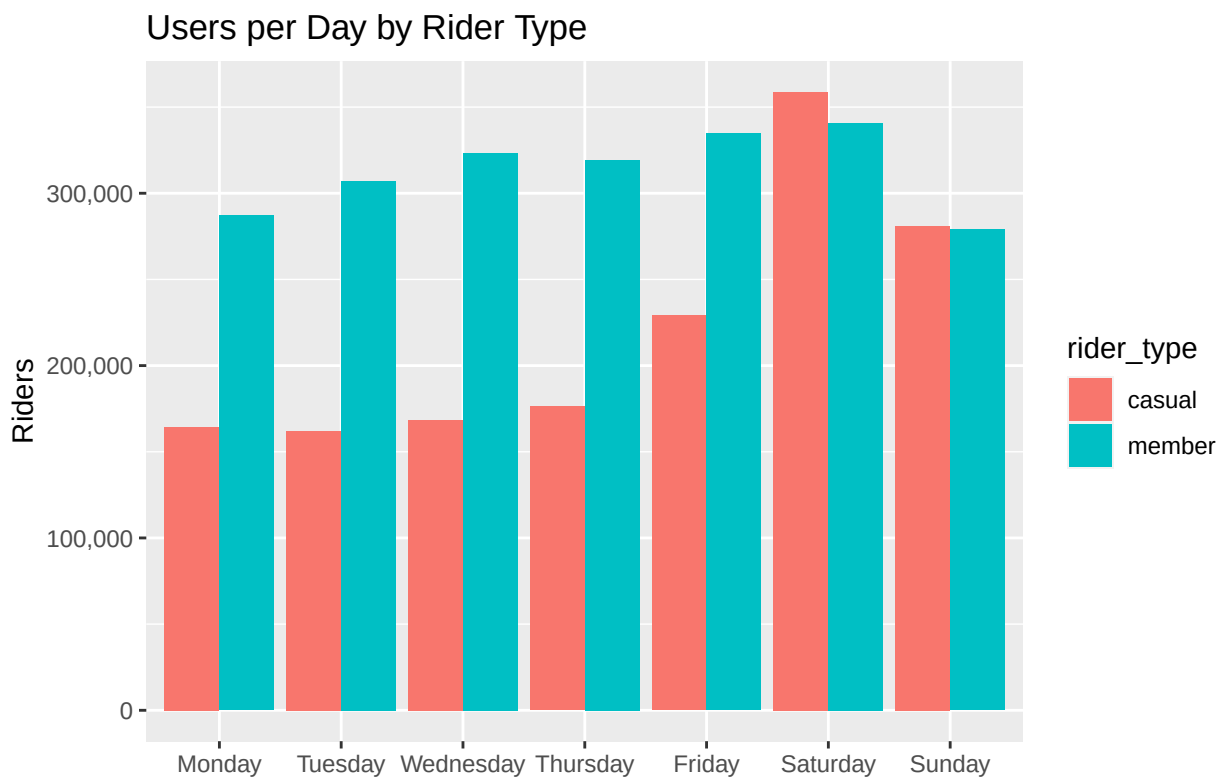
```
most_frequent_station_casuals %>%
  top_n(10, start_station_name_count) %>%
  ggplot() +
  geom_bar(mapping = aes(reorder(x = start_station_name, -start_station_name_count), y =
  ↪  start_station_name_count, fill = "salmon"), stat = "identity") +
  theme(axis.text.x = element_text(angle = 90), legend.position = "none") +
  labs(title = "Top 10 Starting Stations (Casuals)",
       x = "Station",
       y = "Ride Count")
```



Visualize ridership on a weekly basis:

```
# Group by day of week and rider type

tripdata_combined_0520_0421_weekday <- tripdata_combined_0520_0421 %>%
  group_by(weekday) %>%
  summarise(total_riders = n(),
            casual = sum(rider_type == "casual"),
            member = sum(rider_type == "member"),
            avg_ride_duration = format(strptime(as_hms(mean(ride_length)), format =
             ↪  "%H:%M:%S"), "%H:%M:%S"),
            avg_ride_duration_casual = format(strptime(as_hms(mean(ride_length[rider_type
             ↪  == "casual"])), "%H:%M:%S"), "%H:%M:%S"),
            avg_ride_duration_member = format(strptime(as_hms(mean(ride_length[rider_type
             ↪  == "member"])), "%H:%M:%S"), "%H:%M:%S"))
View(tripdata_combined_0520_0421_weekday)
```

```
# Double bar graph comparing "casual" and "member" ridership during the week

ggplot(data = tripdata_combined_0520_0421, aes(fill = rider_type, x = weekday)) +
  geom_bar(position = "dodge", stat = "count") +
  scale_y_continuous(labels = label_comma()) +
  labs(title = "Users per Day by Rider Type",
       x = "",
       y = "Riders",
       caption = "*daily ridership from May 2020 - April 2021")
```

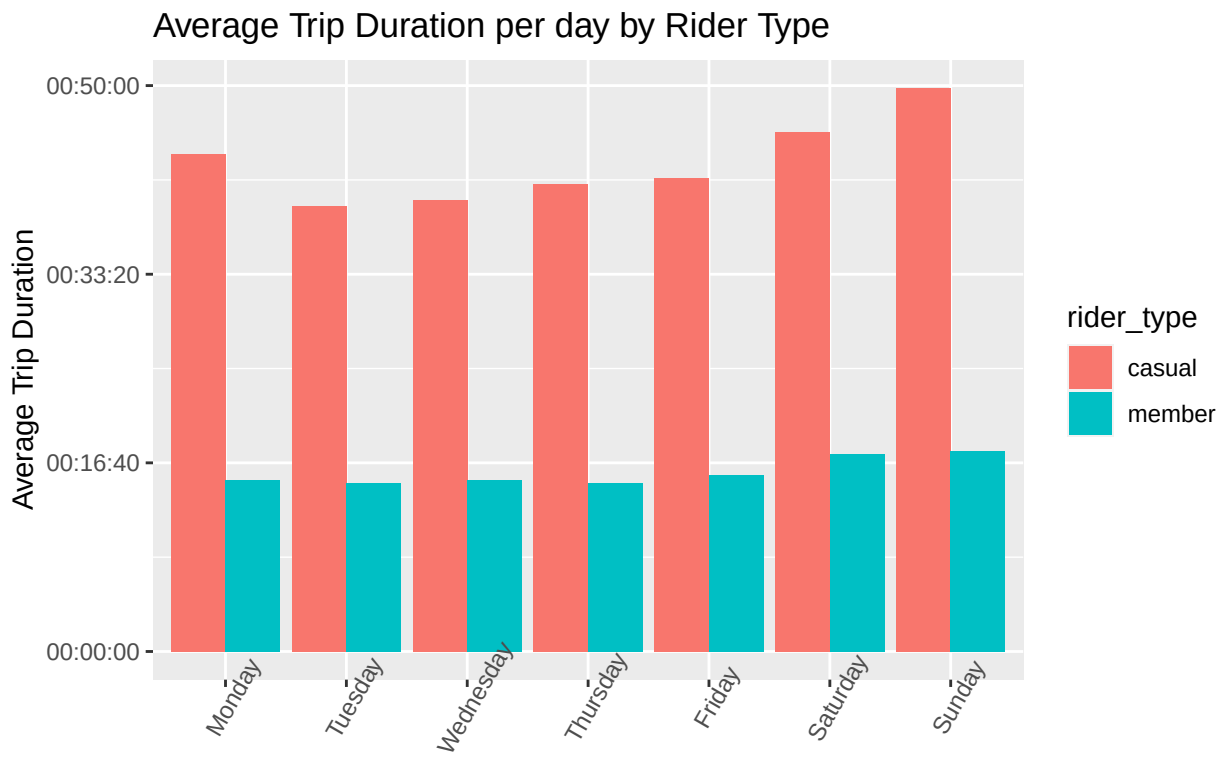


*daily ridership from May 2020 − April 2021

```
# Double bar graph showing average trip duration per day of week by rider type

ggplot(data = tripdata_combined_0520_0421, aes(x= weekday, y = ride_length, fill =
↪  rider_type)) +
  geom_bar(stat = "summary", fun.y = "mean", position = "dodge") +
  theme(axis.text.x = element_text(angle = 60)) +
  scale_y_continuous(labels= hms) +
  labs(title = "Average Trip Duration per day by Rider Type",
       x = "",
       y = "Average Trip Duration")
```

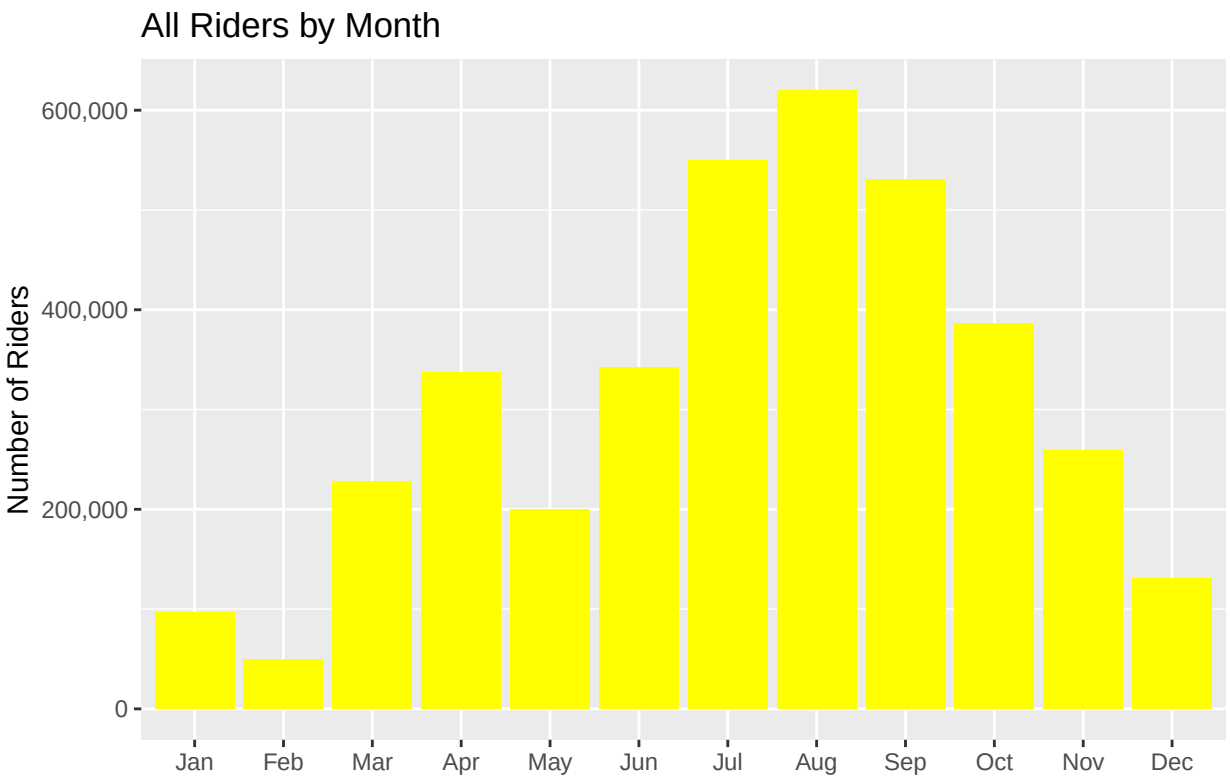
```
## Warning in geom_bar(stat = "summary", fun.y = "mean", position = "dodge"):
## Ignoring unknown parameters: `fun.y`
```

```
## No summary function supplied, defaulting to `mean_se()`
```



Visualize the data on a monthly basis now:

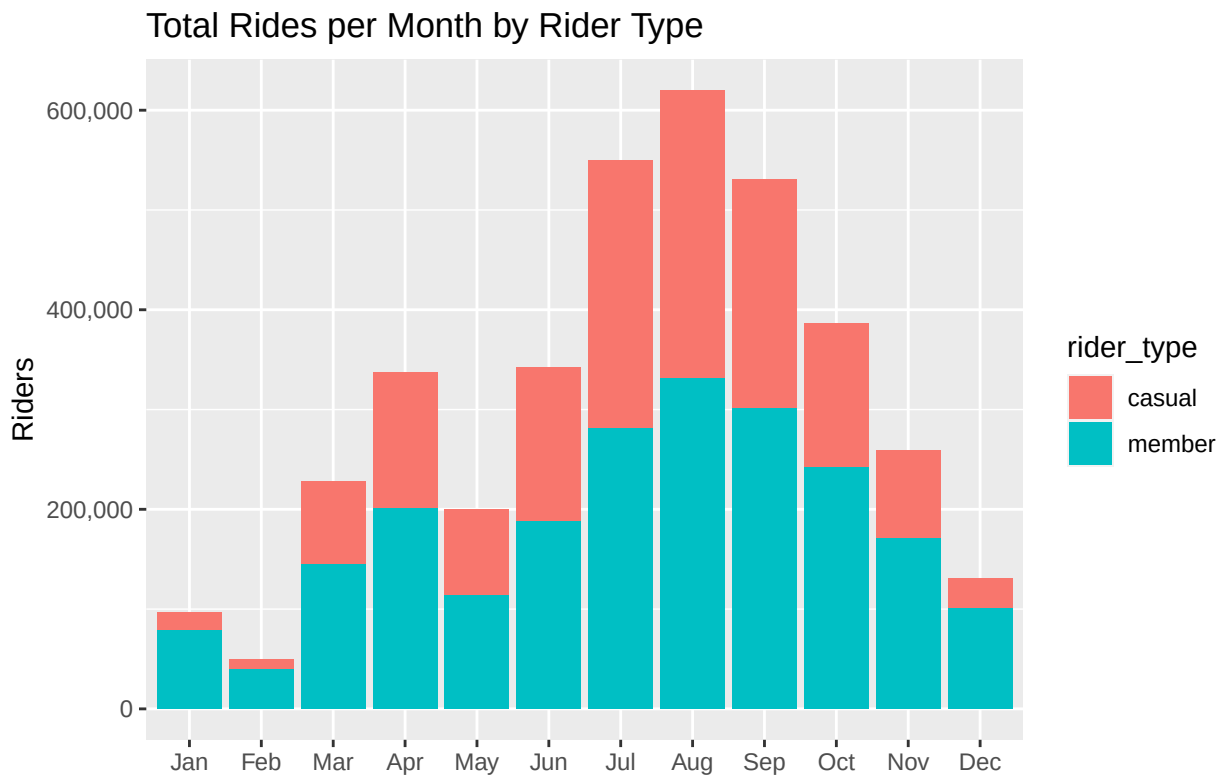
```
# Bar graph showing all riders per month

ggplot(data = tripdata_combined_0520_0421_month, aes(x = month, y = total_riders)) +
  geom_col(fill = "yellow") +
  scale_y_continuous(labels = label_comma())+
  scale_x_discrete(limits = month.abb) +
  labs(title = "All Riders by Month",
       x = "",
       y = "Number of Riders",
       caption= "*Montly ridership from May 2020 - April 2021. Chart is ordered by
       ↪  calendar year.")
```



*Montly ridership from May 2020 − April 2021. Chart is ordered by calendar year.

```
# could not get chart's x-axis to begin at earliest data (May 2020) so it is instead
↪  ordered by calendar year
```

```
# Bar graph showing all riders per month by rider type

ggplot(data = tripdata_combined_0520_0421, aes(x = month, fill = rider_type)) +
  geom_bar(position = "stack", stat = "count") +
  scale_y_continuous(labels = label_comma()) +
  scale_x_discrete(limits = month.abb) +
  labs(title = "Total Rides per Month by Rider Type",
       x = "",
       y = "Riders",
       caption = "*Monthly ridership from May 2020 - April 2021. Chart is ordered by
       ↪  calendar year")
```



*Monthly ridership from May 2020 − April 2021. Chart is ordered by calendar year

```
# could not get chart's x-axis to begin at earliest data (May 2020) so it is instead
↪  ordered by calendar year
```

Act

Guide for Act phase

*** Final conclusion based on analysis:**

There is an opportunity to convert casual riders to annual members - the share of casual riders to total riders is sizable, and they tended to ride Cyclistic's bikes for longer periods per trip on average, and ridership among casual riders was higher than members on weekends. Data in this analysis was first recorded at the beginning of the pandemic; however, seasonal patterns and the demand for bikes still held. A campaign geared towards casual riders, especially during the warmer months where there is increased usage, could be effective at reaching the largest target prospective customer base.

Additionally, we know the top start stations where casual users tend to begin their trip, so we can infer that those casual users likely live by the area. Depending on the campaign method, we could implement billboard advertisements for annual memberships around those vicinity of the start stations for awareness if we cannot obtain identifiable information of casual rides, such as their addresses to mail postcard advertisements or email addresses for digital ad campaigns. Seeing that casual riders outpaced members on weekends, we may want to time advertising efforts and place billboards strategically in such a way that casual riders are more likely to come across Cyclistic's membership promotion during those times.

*** How could the team and business apply these insights?**

We now have insight into the demand of casual riders. Knowing that casual riders tend to take longer trips than annual members will allow the team to create a marketing campaign focusing on longer bike trips. A promotional rate to convert casual riders to annual members if they tend to take trips for "x" amount of minutes might appeal to casual riders enough to purchase the membership.

*** What next steps would you or your stakeholders take based on the findings?**

Further explore the reason for which casual riders' average ride duration is almost three times as long as members. Understanding why casual riders are taking longer trips may allow for a more tailored marketing campaign to reach the maximum amount of potential new members.

*** Is there additional data you could use to expand on your findings?**

More demographic data is needed among riders to better understand the current customer base, such as where users live, their occupations, and if they have alternative modes of transportation. We understand the demand in ridership now, but additional information would help us understand *who* casual riders are so we could strategically target that demographic. Identifiable information for casual riders would be helpful as well, such as an email or home address so Cyclistic can create a digital campaign for casual riders.