Bike Share Analysis

Ahmad Ridho

1/7/2022

This work is a part of my portfolio. Kindly contact me on ahmdrdo@gmail.com for any queries.

Scenario

In 2016, Cyclistic, a bike-sharing company in Chicago, launched a successful bike-share offering. The program features more than 5,800 bicycles and 600 docking stations.The bikes are geotracked and can be unlocked from one station and returned to any other station in the system anytime.

Cyclistic has the flexibility of its pricing plans by offering 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**.

The director of marketing, Lily Moreno believes **the company's future success depends on maximizing the number of annual memberships**. Therefore, he wants to know how casual riders and annual members use Cyclistic bikes differently. From the insights, the marketing team will design a new marketing strategy that aims in converting casual riders into annual members.

Analyst Notes

Analysis objective: **Identify how casual riders and annual members use cyclistic differently**.

Notes for analysis: This analysis is inclusive for the 2021 trips, data recorded after 2021-12-31 will be excluded.

Stakeholders:

- Primary stakeholder — **Cyclistic executive team** and **Lily Moreno**, the director of marketing as well as manager.
- Secondary stakeholder — **Cyclistic marketing analytics team**.

R script for data processing is made available here as well as enclosed at the end.

Presentation is made available here in pdf format.

About Data

This analysis was conducted using Cyclistic's historical trip data from January 2021 until December 2021 as the latest data available at the time this analysis was performed. The data has been made available by Motivate International Inc. on this link under this license.

Dataset is identified as credible by determining ROCCC aspects.

- **R**eliable : The data represents all bike rides taken in Chicago that recorded by system.
- **O**riginal : The data is made available by Motivate International Inc. which operates the city of Chicago's Divvy bicycle sharing service which is powered by Lyft.
- **C**omprehensive : The data includes all information about ride details including type of bike used, riding log, and membership type of rider.
- **C**urrent : The Motivate International Inc. is regularly updating the data. By the time this analysis was conducted, no data was available later than December 2021.
- **C**ited : The data is cited and is available under Data License Agreement.

Data Scheme

The analysis was conducted using 12 CSV files that contain historical trip records from **January 2021 to December 2021**. Scheme for data as shown as below.

Field Name	Description
ride_id	Unique trip or ride identification number.
rideable_type	Type of bike used.
started_at	Date-time (UTC) when the trip started.
ended_at	Date-time (UTC) when trip ended.
start_station_name	Station where trip started
start_station_id	Station's ID where trip started
end_station_name	Station where trip ended
end_station_id	Station's ID where trip ended
start_lat	Latitude of start point
start_lng	Longitude of start point
end_lat	Latitude of end point
end_lng	Longitude of end point
member_casual	Membership type. Casual — single-ride or full-day passes, Member — annual membership.

Notes: Some adjustment of field names will be done for processing effectivity.

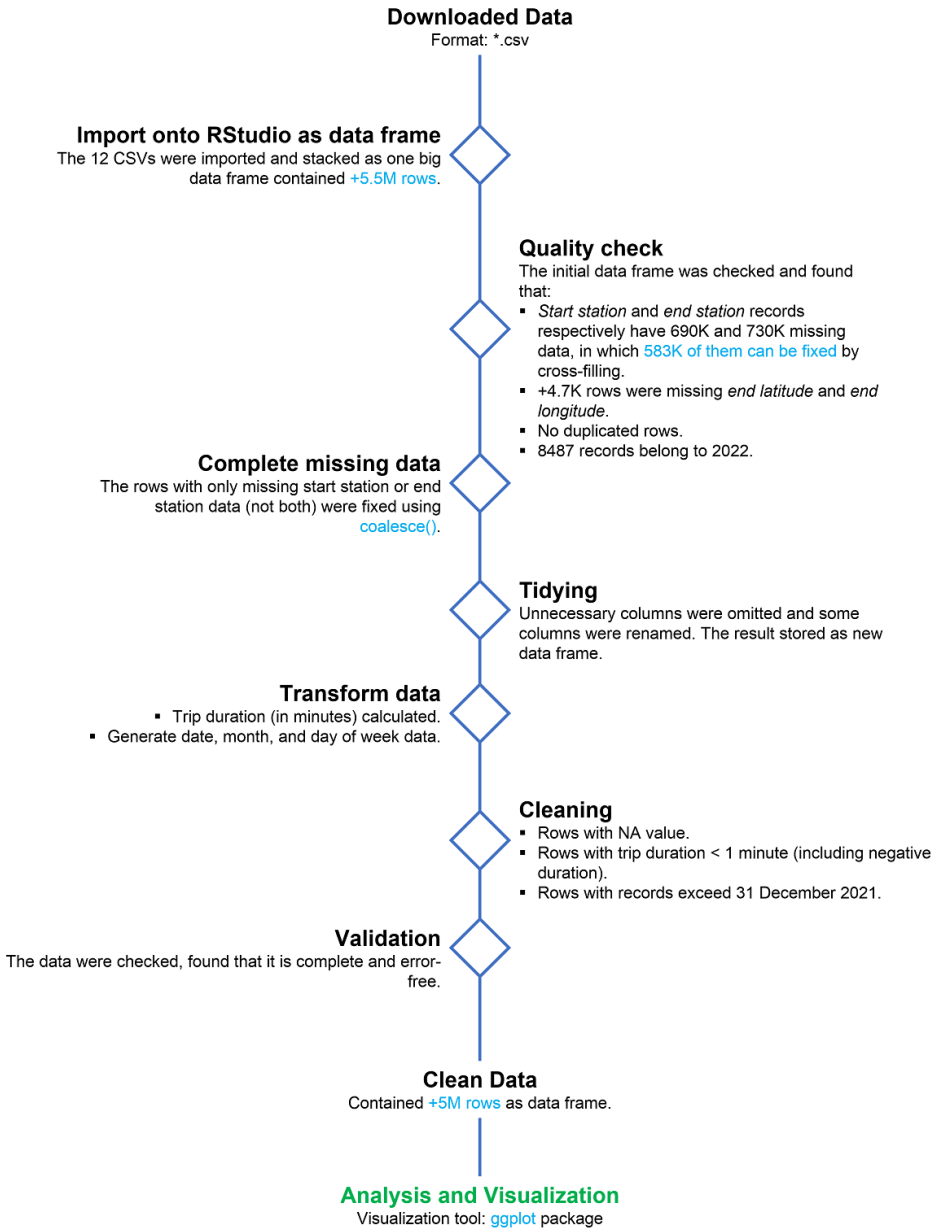
Data Limitations

The data do not provide unique user identification number (user ID) or any information that is useful for user identification. Thus, **the data can not be used to identify total memberships in 2021 or the frequency of rides a user has taken**.

From data skimming, hundred-thousands of "start station" and or "end station" information are missing. Since the Cyclistic bike-share is a station-to-station trip, records with **missing both start station and end station data are considered invalid** and need to be discarded. Even so, some of the missing station data — **at start or end station, not both** — can be cross-filled. For such a case, it is observed that the end riding point (latitude & longitude) is near to start riding point, indicating that **the trip started and ended at the same station**.

Processes

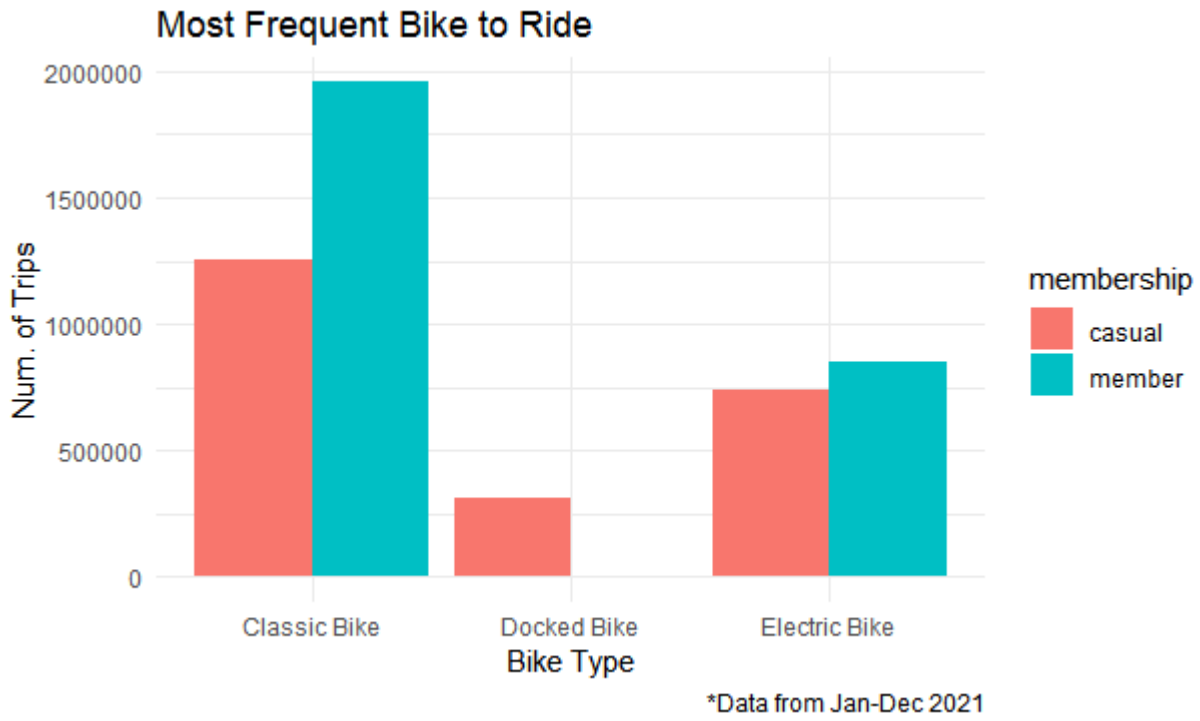
Before analyzing, it is necessary to ensure data is clean, free of error, and in the right format. R programming is used due to its ability in handling huge data effectively. The following flowchart summarises any cleaning or manipulation of data.



Analysis

Rather than targeting all-new customers, the casual riders are more potential to convert into members. Casual riders are already aware of the Cyclistic bike-share program and have chosen Cyclistic for their mobility needs.

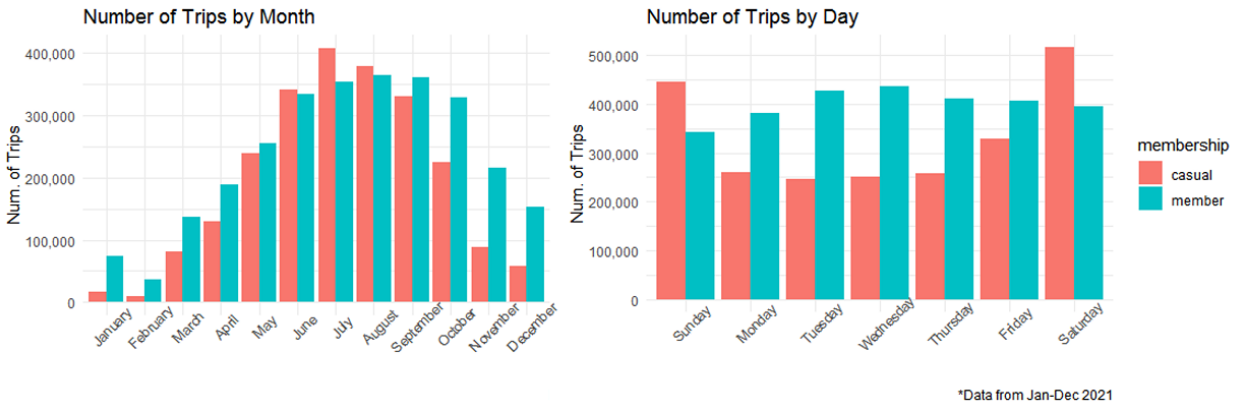
Note that this analysis is conducted using historical trip records from January to December 2021. The analysis is done using trips recorded by the system for each trip, not for each user.



The diagram above shows the frequency of classic bike, docked bike, and electric bike being used for trips. **Classic bike is more often to use for trips than other bikes**. From January to December 2021, there were up to 3,200,000 trips were done with classic bike where 40% of them were casual riders.

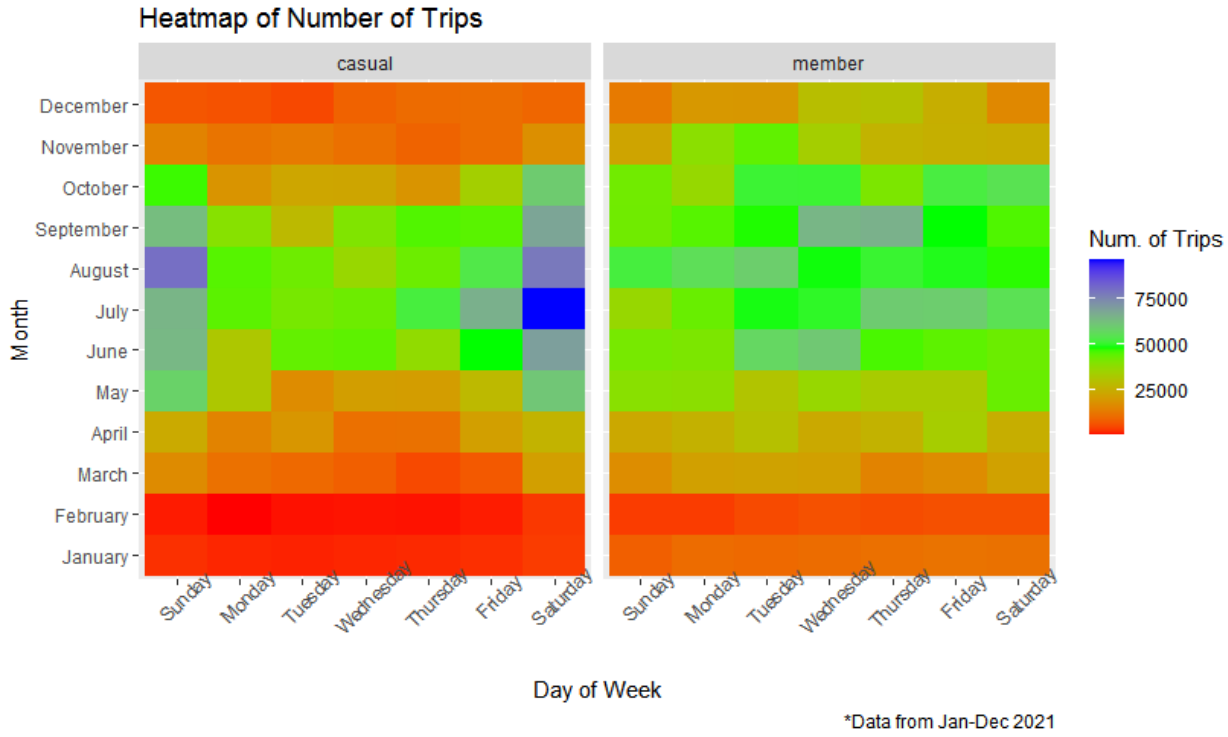
As shown in the following diagram, **cyclistic usage for both casual and member riders is seasonal**. In the first quartile, January - March, there were 353,450 trips in total (casual & member trips) and increased when entering second and third quartile, then decreased in the fourth quartile. June to September were months with total trips more than 300,000 trips in each month, either for casual riders or members.

It is also observed that **the number of trips done by casual riders on weekends is more than on weekdays**. For one year period, total casual riders' trips reach 516,042 trips on Saturday, 444,825 trips on Sunday, and 268,000 trips on weekdays on average. Meanwhile, total trips by day tend to be constant with only a slight difference through the week.

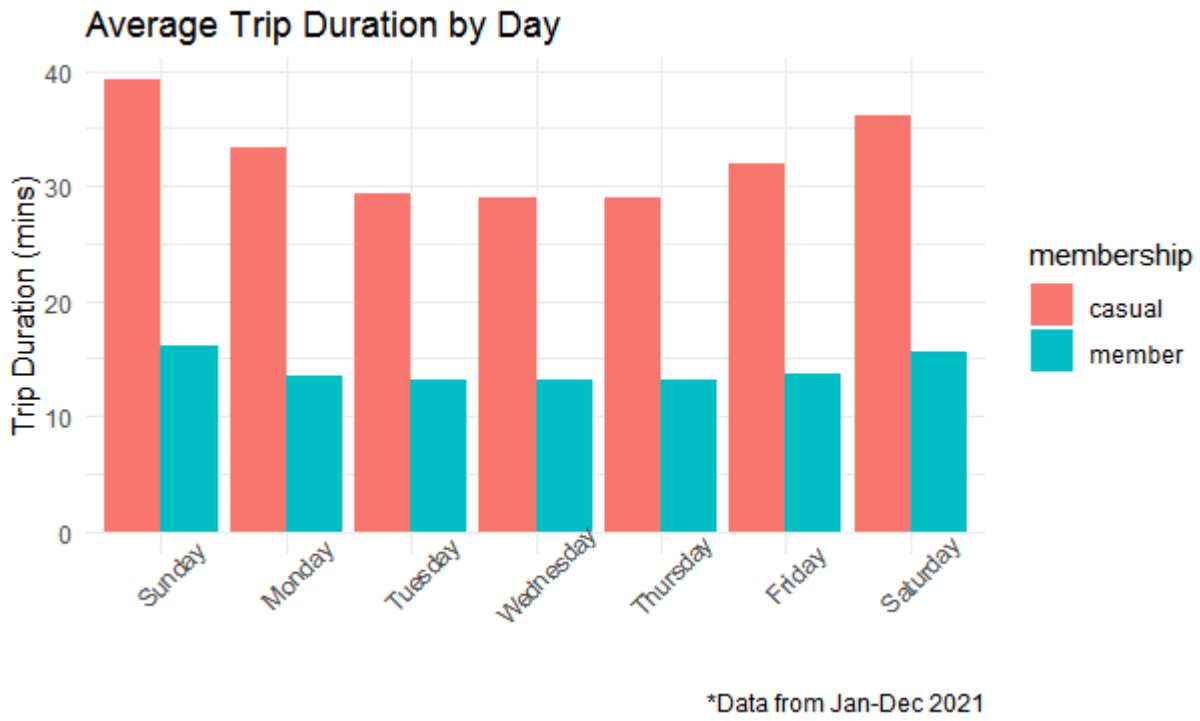


A more comprehensive analysis of casual riders and members' trips is represented on the heatmap below, the number of trips is indicated by rainbow color code as detailed in the legend. From the heatmap, found

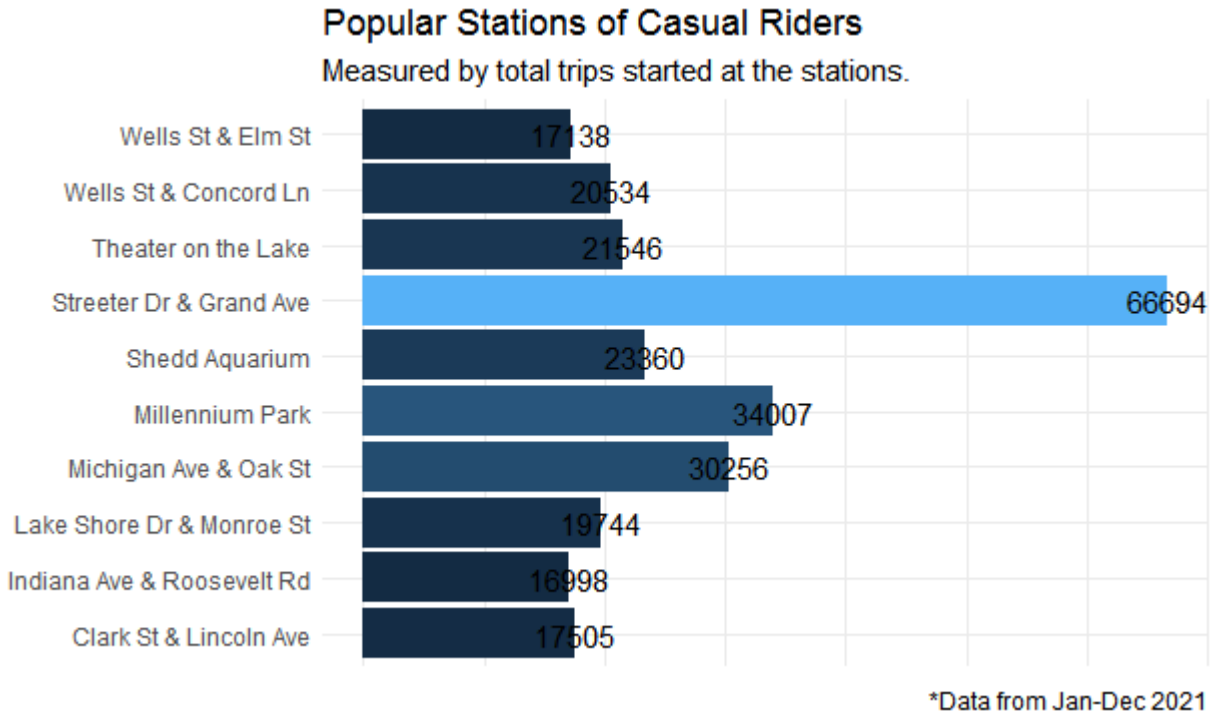
out that **the number of casual riders' trips on Saturday and Sunday were consistently higher than the other days throughout the year, especially in June, July, August, and September**. And for members, the number of trips was slightly constant through the week, but had more trips in June to October compared to other months. These findings align with the previous two diagrams.



The following diagram shows comparison of trip duration between casual riders and members. **Although the number of trips by casual riders was fewer than members was, it was longer in duration**. The average trip duration for casual riders is 33.5 minutes while it is only 14 minutes for members.



The below diagram ranks the stations from which casual riders mostly started the trip. Ten stations were selected among 847 stations. And among the ten, **Streeter Dr & Grand Ave**, **Millenium Park**, and **Michigan Ave & Oak St** take the first, second, and third position. These three stations might be a good consideration for marketing strategy.



Conclusions and Recommendations

Based on analysis results, it is can be concluded that:

- Classic bike is more often to use for trips both by casual riders or members.
- The bike-share program is used seasonally with high demand from June to September for both casual riders and members.
- Total trips by casual riders are higher on weekends and steady during weekdays, this is different from members that active on all days of the week. This indicates that casual riders mostly use cyclistic for recreational purposes.
- Although annual trip by casual riders is fewer than members, their weekly riding is longer than members.
- Streeter Dr & Grand Ave, Millenium Park, and Michigan Ave & Oak St are the three most stations where casual riders start the trip from.

Based on the above findings, the recommendation for marketing strategy are:

- Design more flexible packages such as monthly or semester subscriptions. These packages give flexibility to those who are not willing for an annual subscription.
- Consider offering promos or coupons for weekends or seasonal events. This might be able to increase customer experience at the same time attract new customers. By increasing customer experience, casual riders will be easier to encourage for subscriptions.
- Design more intimate marketing material for casual riders. Casual riders already have their trip histories that show their riding habit. Campaign that aligns with their habit will encourage them for subscriptions.
- Besides online marketing, in-person marketing is worth trying, especially around popular sites which often visited by casual riders as well as the stations they usually start the trip from.

Recommendation for further analysis:
Additional data i.e. user history of trip is needed to expand current findings. Using these data, analysis on riding behavior can be conducted. Furthermore, the data will give information on why a customer decides on subscriptions.

Appendix

Below is the R script used for data processing and visualization. The script is also downloadable here.

```
# -------------------------
# Author: Ahmad Ridho
# Email: ahmdrdo@gmail.com
# -------------------------

setwd("~/Collection/Portfolio/Cyclistic Bike-sharing")

# Load packages.
library(tidyverse)
library(dplyr)
library(lubridate)
library(ggplot2)
library(skimr)

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

```r
# IMPORT DATA
# ================================================================================

data_dir <- "~/Collection/Portfolio/Cyclistic Bike-sharing/Used Dataset"

# Stack individual files as one.
trips_raw <- list.files(path = data_dir, pattern = ".csv", full.names = TRUE) %>%
  lapply(read_csv) %>%
  bind_rows()

# ================================================================================
# COMPLETE MISSING DATA
# ================================================================================

trips_clean <- trips_raw
trips_clean <- trips_clean %>% # Fill missing station name.
  mutate(start_station_name = coalesce(start_station_name, end_station_name),
         end_station_name = coalesce(end_station_name, start_station_name))

# ================================================================================
# TIDYING
# ================================================================================

# Remove unnecessary columns.
trips_clean <- select(trips_clean, -c(start_station_id, end_station_id,
                                      start_lat, start_lng,
                                      end_lat, end_lng))
# Rename columns.
trips_clean <- rename(trips_clean, trip_id = ride_id,
                      bike_type = rideable_type,
                      membership = member_casual)

# ================================================================================
# TRANSFORM DATA
# ================================================================================

# Calculate trip duration (result in minutes).
trips_clean$trip_duration <- difftime(trips_clean$ended_at, trips_clean$started_at, units = "mins")
# Reformat trip duration.
trips_clean$trip_duration <- as.numeric(as.character(trips_clean$trip_duration))

# Generate Date, Month, Day of Week, and Hour.
trips_clean$date <- as.Date(trips_clean$started_at)
trips_clean$month <- format(trips_clean$date, "%B")
trips_clean$day_of_week <- format(trips_clean$date, "%A")

# Fix the order of ordinal data.
trips_clean$month <- ordered(trips_clean$month, levels = c("January", "February", "March", "April", "Ma
trips_clean$day_of_week <- ordered(trips_clean$day_of_week, levels = c("Sunday", "Monday", "Tuesday", "

# ================================================================================
# CLEANING
# ================================================================================
```

```r
# Remove NA and trip duration < 1 min (inc. minus duration).
trips_clean <- trips_clean %>%
  na.omit() %>%
  filter(trip_duration >= 1)

# Filter data only for 2021.
trips_clean <- filter(trips_clean, date >= "2021-01-01" & date <= "2021-12-31")

# =====================================================================
# VALIDATION
# =====================================================================

skim_without_charts(trips_clean)

# =====================================================================
# DATA VISUALIZATION
# =====================================================================

# Bike used by rider type.
ggplot(data = trips_clean, aes(x = bike_type, fill = membership)) +
  geom_bar(position = "dodge") +
  theme_minimal() +
  labs(title = "Most Frequent Bike to Ride", caption = "*Data from Jan-Dec 2021", x = "Bike Type", y =
  scale_x_discrete(labels = c("Classic Bike", "Docked Bike", "Electric Bike"))

# Number of trips by month.
ggplot(data = trips_clean, aes(x = month, fill = membership)) +
  geom_bar(position = "dodge") +
  theme_minimal() +
  labs(title = "Number of Trips by Month", caption = "*Data from Jan-Dec 2021", y = "Num. of Trips") +
  theme(axis.title.x = element_blank(), axis.text.x = element_text(angle = 45)) +
  scale_y_continuous(labels = scales::comma)

# Number of trips by day.
ggplot(data = trips_clean, aes(x = day_of_week, fill = membership)) +
  geom_bar(position = "dodge") +
  theme_minimal() +
  labs(title = "Number of Trips by Day", caption = "*Data from Jan-Dec 2021", y = "Num. of Trips") +
  theme(axis.title.x = element_blank(), axis.text.x = element_text(angle = 45)) +
  scale_y_continuous(labels = scales::comma)

# Heatmap number of Trips
trips_clean %>%
  count(membership, month, day_of_week, name = "count") %>%
  ggplot(aes(x = day_of_week, y = month, fill = count)) +
  geom_tile() +
  facet_wrap(~membership) +
  scale_fill_gradientn(name = "Num. of Trips", colors = rainbow(3)) +
  labs(title = "Heatmap of Number of Trips", caption = "*Data from Jan-Dec 2021",
       x = "Day of Week", y = "Month") +
  theme(axis.text.x = element_text(angle = 45))

# Average trip duration.
```

```r
trips_clean %>%
  group_by(membership, day_of_week) %>%
  summarise(avg_duration = mean(trip_duration)) %>%
  ggplot(aes(x = day_of_week, y = avg_duration, fill = membership)) +
  geom_col(position = "dodge") +
  theme_minimal() +
  labs(title = "Average Trip Duration by Day", caption = "*Data from Jan-Dec 2021",
       x = "Day of week", y = "Trip Duration (mins)") +
  theme(axis.title.x = element_blank(), axis.text.x = element_text(angle = 45))

# Top stations of casual riders.
trips_clean %>%
  filter(membership == "casual") %>%
  count(start_station_name, name = "count") %>%
  top_n(n = 10, wt = count) %>%
  ggplot(aes(x = count, y = start_station_name, fill = count)) +
  geom_col(show.legend = FALSE) +
  theme_minimal() +
  labs(title = "Popular Stations of Casual Riders", subtitle = "Measured by total trips started at the s
       x = "Num. of Trips", y = "Station Name") +
  theme(axis.text.x = element_blank(), axis.title = element_blank()) +
  geom_text(aes(label = count))
```