Introduction

This notebook presents an exploratory data analysis (EDA) of Cyclistic's bike share data for Q1 2019 and Q1 2020.
The goal is to compare usage patterns between **casual riders** and **annual members**, focusing on total trips, ride length distributions, weekday and hourly behaviors, and station usage.

By combining SQL preprocessing with R-based visualization, this analysis provides clear insights into **how different customer segments use the service**.
The findings can help inform strategies to **increase casual rider conversion into memberships** and optimize operational planning.

Each section of this report contains both code and commentary:
- **Tables** summarize key statistics.
- **Visualizations** highlight differences in patterns.
- **Interpretations** connect the outputs to meaningful business insights.

Cyclistic Bike-Share Case Study

```
library(dplyr)
```

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

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

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

```
library(ggplot2)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

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

```
library(scales)

trips_cleaned <- readr::read_csv("trips_cleaned.csv")
```

```
## Rows: 783996 Columns: 19
```

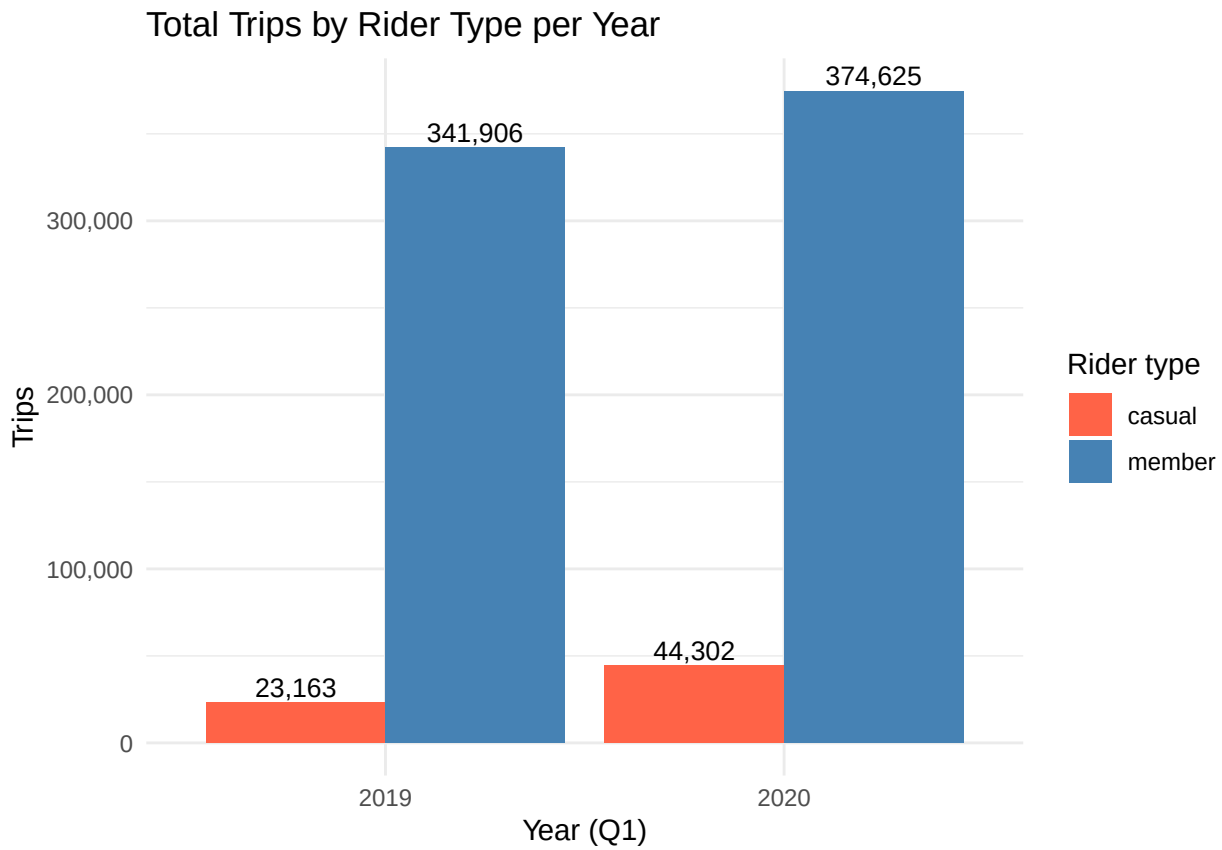
```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (9): ride_id, rideable_type, started_at, ended_at, start_station_name, ...
## dbl  (9): start_station_id, end_station_id, start_lat, start_lng, end_lat, e...
## time (1): ride_length_hms
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Dataset

```
df <- trips_cleaned %>%
  mutate(
    started_at   = ymd_hms(started_at, quiet = TRUE),
    started_ct   = with_tz(started_at, "America/Chicago"),
    dow          = wday(started_ct, label = TRUE, abbr = TRUE),
    hour_of_day  = hour(started_ct),
    ride_min     = ride_length_seconds / 60
  )
```

1) Total Trips by Rider Type per Year

```
df %>%
  count(year, member_casual) %>%
  ggplot(aes(x = factor(year), y = n, fill = member_casual)) +
  geom_col(position = "dodge") +
  geom_text(aes(label = scales::comma(n)),
            position = position_dodge(width = 0.9),
            vjust = -0.3, size = 3.5) +
  scale_y_continuous(labels = comma) +
  scale_fill_manual(values = c(casual = "tomato", member = "steelblue"),
                    name = "Rider type") +
  labs(x = "Year (Q1)", y = "Trips",
       title = "Total Trips by Rider Type per Year") +
  theme_minimal()
```



The plot shows that members dominate total ridership in both Q1 2019 and Q1 2020, with several hundred thousand trips compared to tens of thousands for casual riders. This highlights that members are the core customer base, but casual riders still represent a significant pool for potential conversion into memberships.

2) Ride Length Distribution Table

```
df %>%
  group_by(year, member_casual) %>%
  summarise(
    trips          = n(),
    avg_minutes    = round(mean(ride_min, na.rm = TRUE), 2),
    p25_minutes    = round(quantile(ride_min, 0.25, na.rm = TRUE), 2),
    median_minutes = round(quantile(ride_min, 0.50, na.rm = TRUE), 2),
    p75_minutes    = round(quantile(ride_min, 0.75, na.rm = TRUE), 2),
    .groups = "drop"
  ) %>%
  arrange(year, member_casual)
```

```
## # A tibble: 4 x 7
##    year member_casual  trips avg_minutes p25_minutes median_minutes p75_minutes
##   <dbl> <chr>          <int>       <dbl>       <dbl>          <dbl>       <dbl>
## 1  2019 casual         23163        61.9       13.5           23.4         38.5
## 2  2019 member        341906        13.9        5.28           8.35        13.4
## 3  2020 casual         44302        40.2       12.8           23.1         39.4
## 4  2020 member        374625        11.6        5.38           8.65        14.1
```

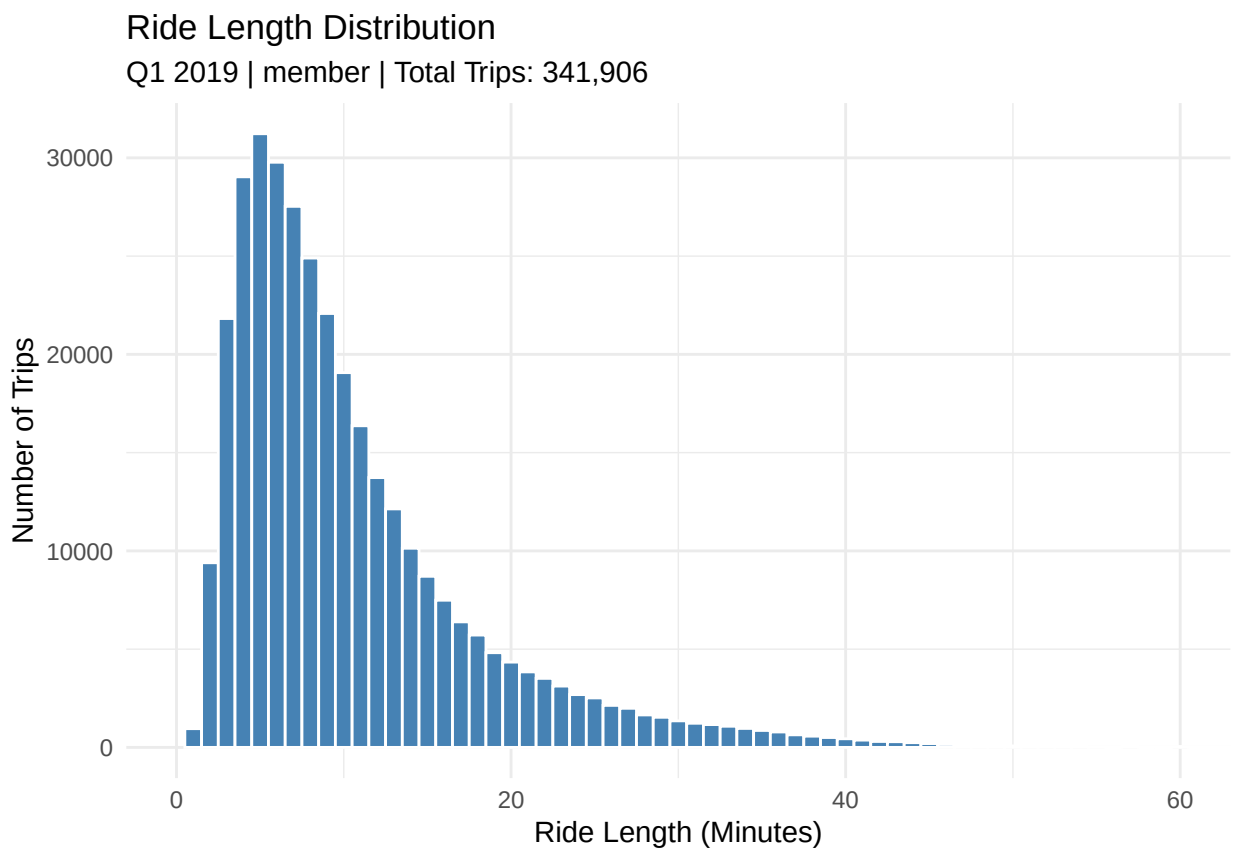
Visualizations

Preparing Dataframe

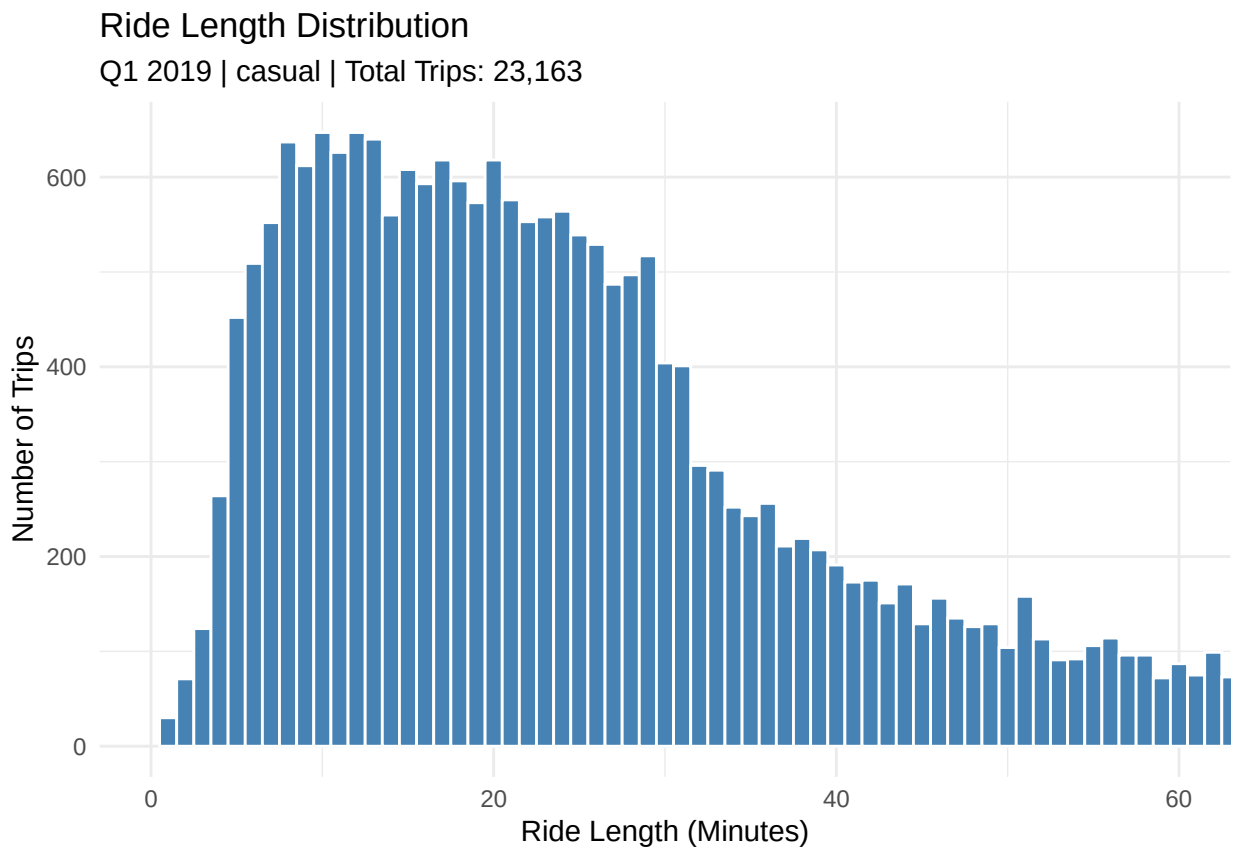
```
df <- df %>%
  mutate(ride_min = as.numeric(ride_length_seconds) / 60)

plot_ride_length <- function(data, yr, rider) {
  gdat <- data %>%
    filter(year == yr, member_casual == rider)

  total_trips <- nrow(gdat)

  ggplot(gdat, aes(x = ride_min)) +
    geom_histogram(binwidth = 1, color = "white", fill = "steelblue") +
    coord_cartesian(xlim = c(0, 60)) +
    labs(
      title    = "Ride Length Distribution",
      subtitle = paste("Q1", yr, "|", rider, "| Total Trips:", comma(total_trips)),
      x = "Ride Length (Minutes)",
      y = "Number of Trips"
    ) +
    theme_minimal()
}
```

Calling the Four Graphs

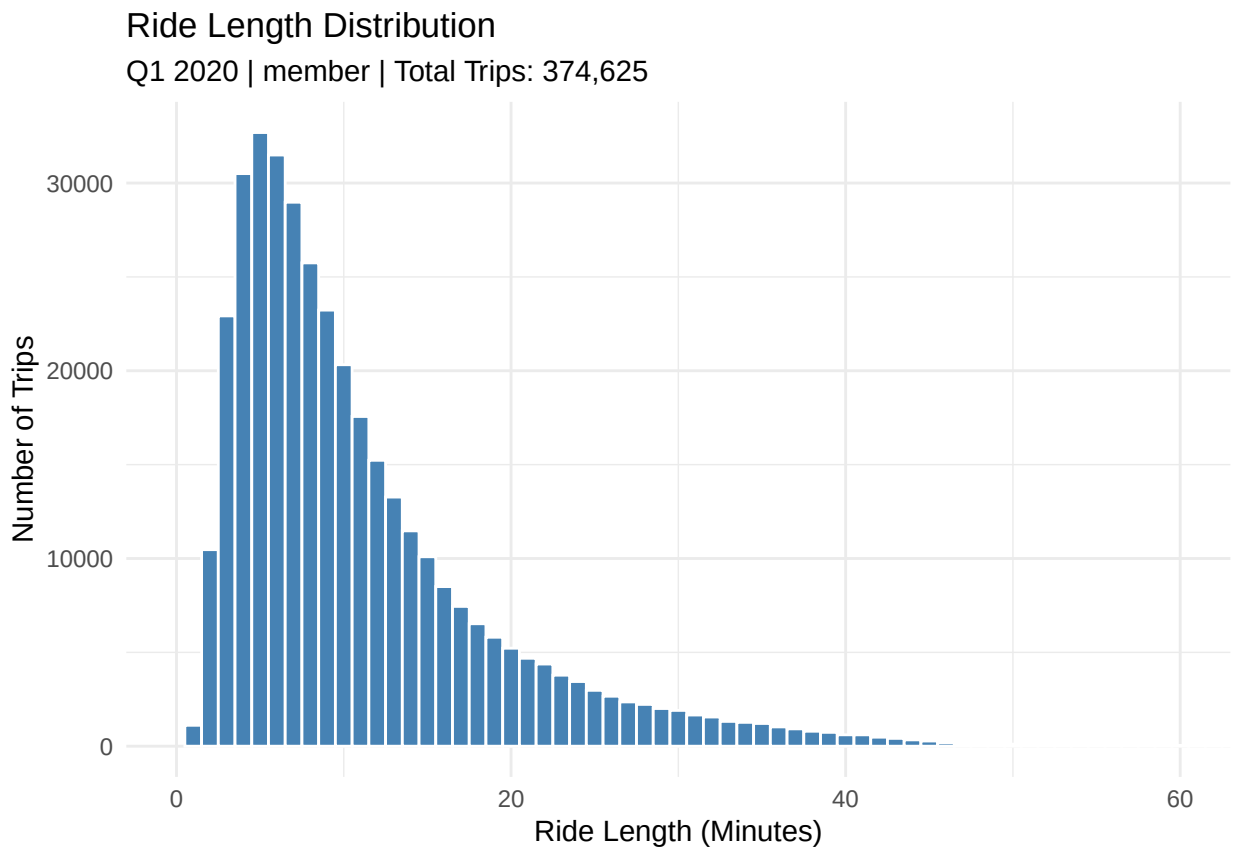
```
# 2019 - members
plot_ride_length(df, 2019, "member")
```



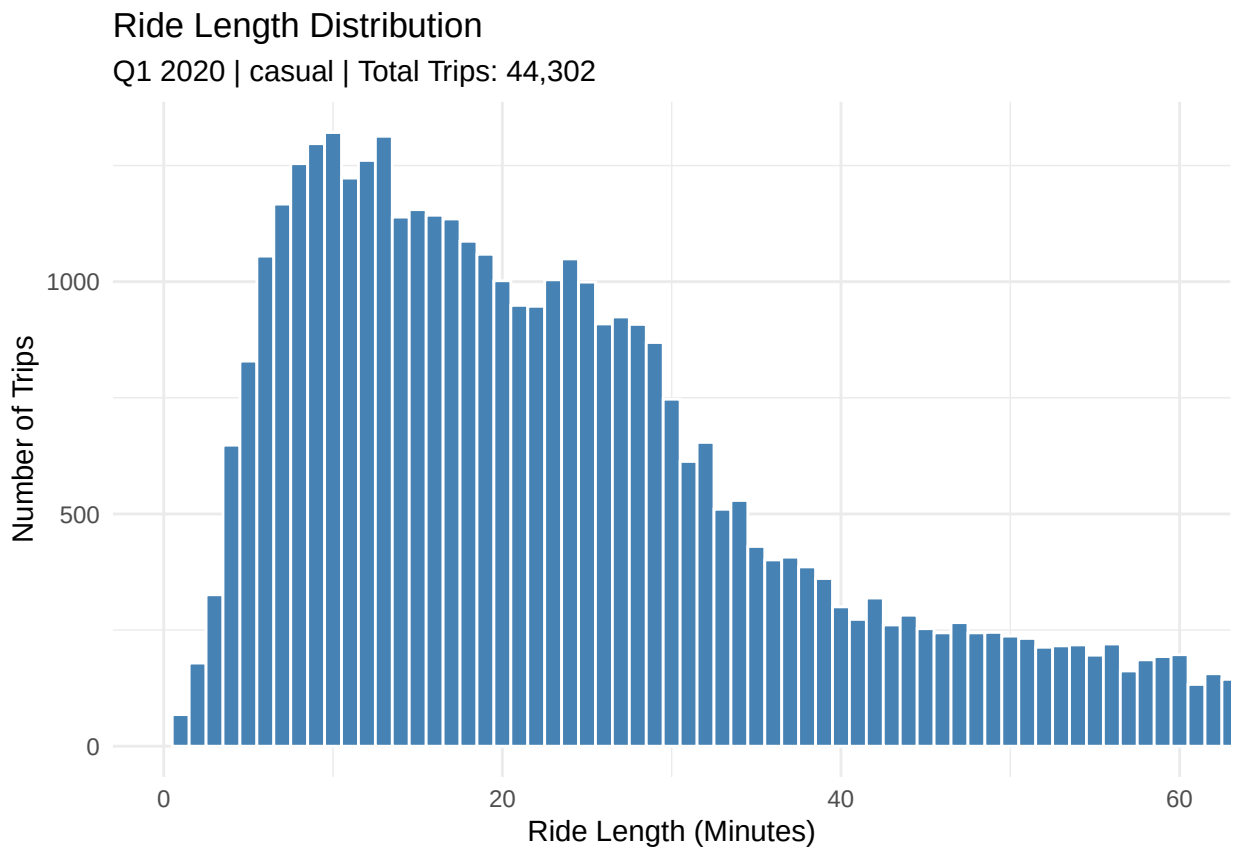

```
# 2019 - casual
plot_ride_length(df, 2019, "casual")
```




```
# 2020 - members
plot_ride_length(df, 2020, "member")
```




```
# 2020 - casual
plot_ride_length(df, 2020, "casual")
```

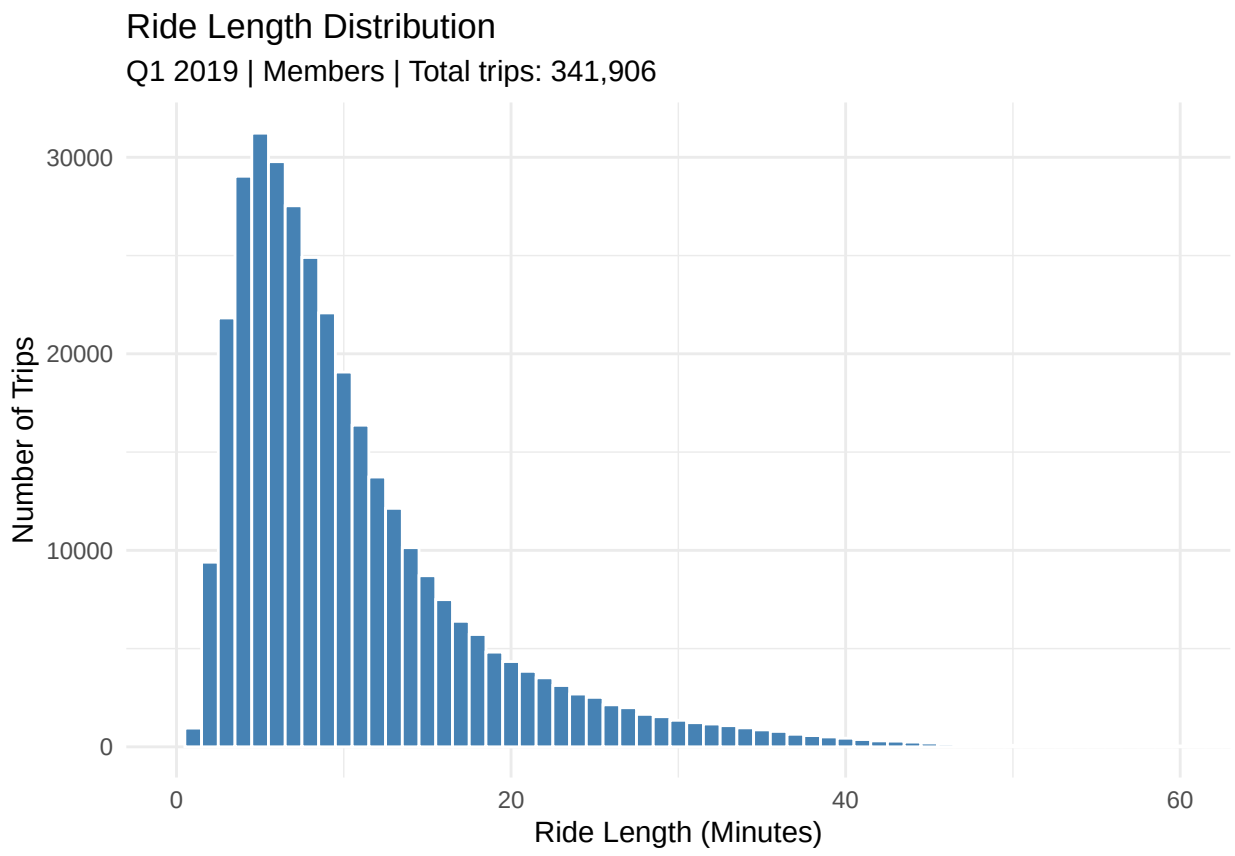


The distribution tables and histograms reveal strong differences in ride duration. Members' rides are shorter and more concentrated (typically under 20 minutes), consistent with commuting or utility trips. Casual riders, by contrast, show much longer and more variable rides, especially on weekends, suggesting leisure or tourism use.

Code Chunks to generate each Graph Individually

2a) Histogram 2019 – Members

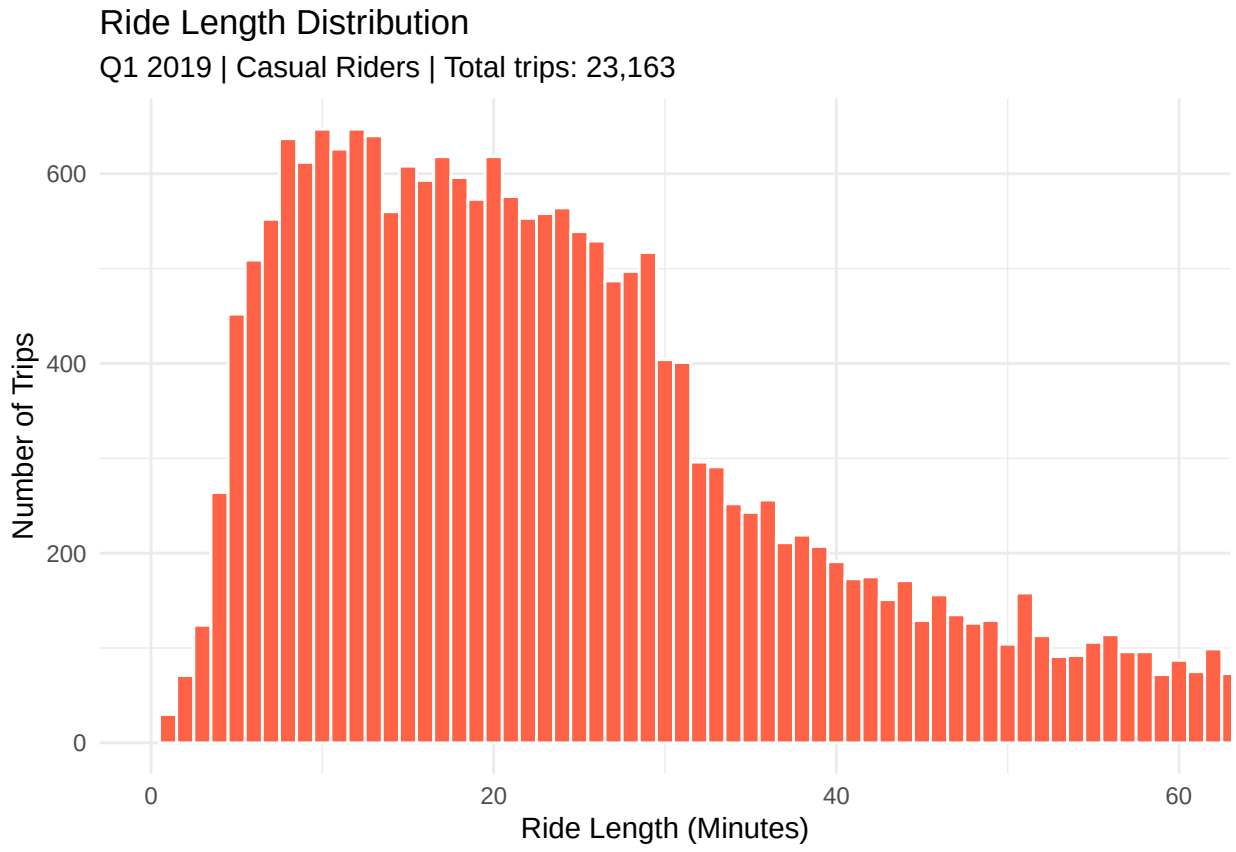
```
total_trips <- df %>% filter(year == 2019, member_casual == "member") %>% nrow()

df %>%
  filter(year == 2019, member_casual == "member") %>%
  mutate(ride_min = as.numeric(ride_length_seconds) / 60) %>%
  ggplot(aes(x = ride_min)) +
  geom_histogram(binwidth = 1, color = "white", fill = "steelblue") +
  coord_cartesian(xlim = c(0, 60)) +
  labs(
    title = "Ride Length Distribution",
    subtitle = paste("Q1 2019 | Members | Total trips:", comma(total_trips)),
    x = "Ride Length (Minutes)",
    y = "Number of Trips"
  ) +
  theme_minimal()
```



Members in 2019 show shorter, consistent trips, with most rides under 20 minutes. This pattern is consistent with commuting and utility use.

2b) Histogram 2019 – Casual

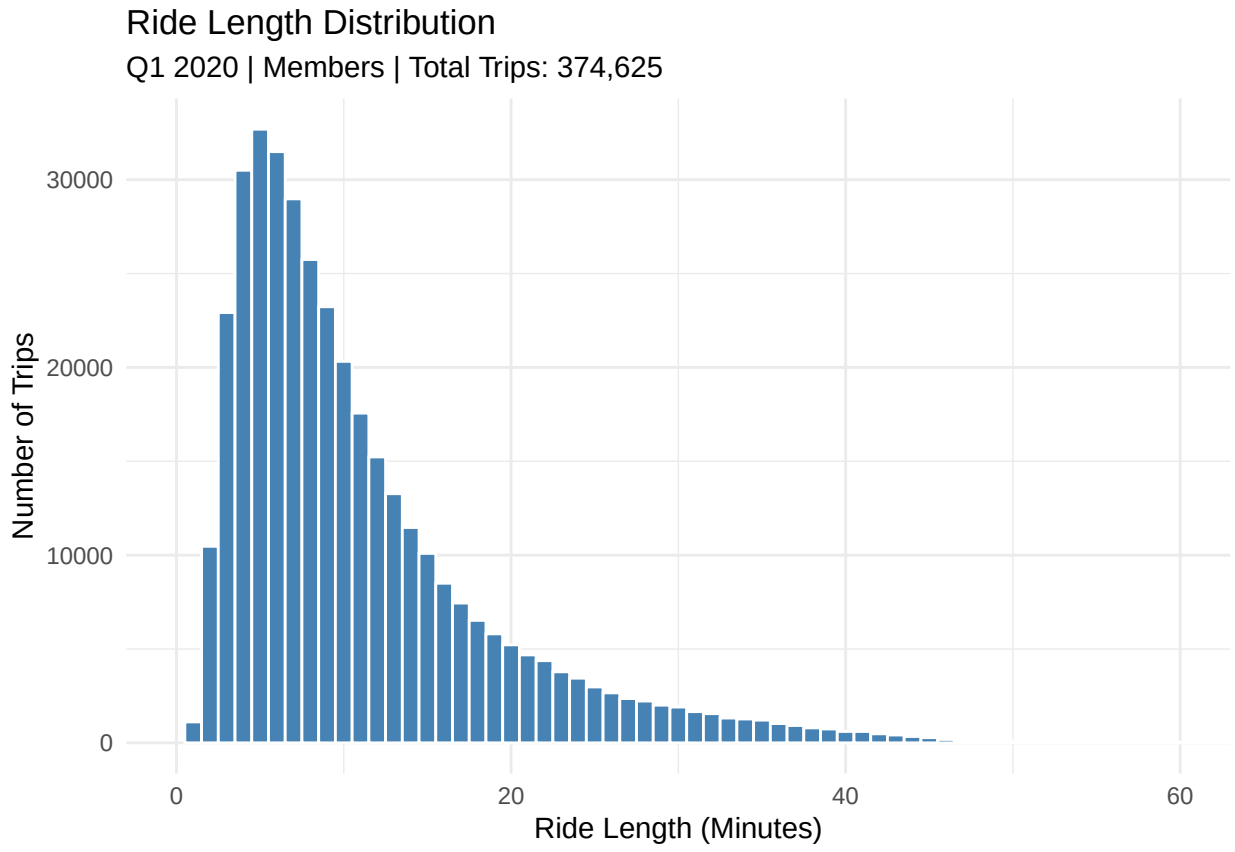
```
total_trips <- df %>% filter(year == 2019, member_casual == "casual") %>% nrow()

df %>%
  filter(year == 2019, member_casual == "casual") %>%
  mutate(ride_min = as.numeric(ride_length_seconds) / 60) %>%
  ggplot(aes(x = ride_min)) +
  geom_histogram(binwidth = 1, color = "white", fill = "tomato") +
  coord_cartesian(xlim = c(0, 60)) +
  labs(
    title = "Ride Length Distribution",
    subtitle = paste("Q1 2019 | Casual Riders | Total trips:", comma(total_trips)),
    x = "Ride Length (Minutes)",
    y = "Number of Trips"
  ) +
  theme_minimal()
```



Casual riders in 2019 take longer and more variable rides, often 30–60 minutes. This reflects a leisure or tourism pattern rather than commuting.

2c) Histogram 2020 – Members

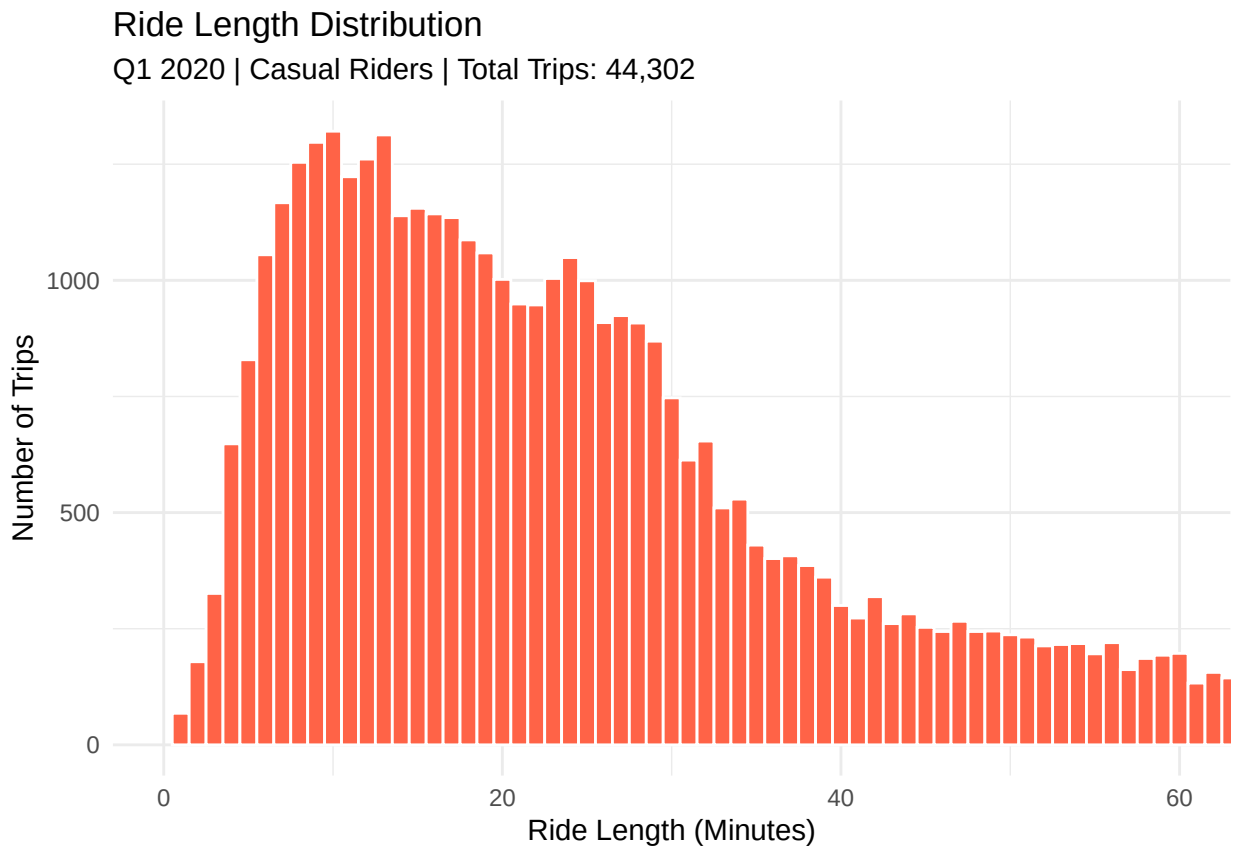
```
total_trips <- df %>% filter(year == 2020, member_casual == "member") %>% nrow()

df %>%
  filter(year == 2020, member_casual == "member") %>%
  mutate(ride_min = as.numeric(ride_length_seconds) / 60) %>%
  ggplot(aes(x = ride_min)) +
  geom_histogram(binwidth = 1, color = "white", fill = "steelblue") +
  coord_cartesian(xlim = c(0, 60)) +
  labs(
    title = "Ride Length Distribution",
    subtitle = paste("Q1 2020 | Members | Total Trips:", comma(total_trips)),
    x = "Ride Length (Minutes)",
    y = "Number of Trips"
  ) +
  theme_minimal()
```



Members in 2020 maintain a similar distribution to 2019: short rides dominate, with peaks around 10–15 minutes. This consistency highlights commuting behavior.

2d) Histogram 2020 – Casual

```
total_trips <- df %>% filter(year == 2020, member_casual == "casual") %>% nrow()

df %>%
  filter(year == 2020, member_casual == "casual") %>%
  mutate(ride_min = as.numeric(ride_length_seconds) / 60) %>%
  ggplot(aes(x = ride_min)) +
  geom_histogram(binwidth = 1, color = "white", fill = "tomato") +
  coord_cartesian(xlim = c(0, 60)) +
  labs(
    title = "Ride Length Distribution",
    subtitle = paste("Q1 2020 | Casual Riders | Total Trips:", comma(total_trips)),
    x = "Ride Length (Minutes)",
    y = "Number of Trips"
  ) +
  theme_minimal()
```



3) Trips by Weekday (Sun→Sat) per Rider Type

Table

```
df %>%
  mutate(
    year             = year(started_at),
    day_of_week_num  = wday(started_at, week_start = 7),
    day_of_week_name = wday(started_at, week_start = 7, label = TRUE, abbr = FALSE)
  ) %>%
  group_by(year, member_casual, day_of_week_num, day_of_week_name) %>%
  summarise(trips = n(), .groups = "drop") %>%
  arrange(year, member_casual, day_of_week_num)
```

```
## # A tibble: 28 x 5
##     year member_casual day_of_week_num day_of_week_name trips
##    <dbl> <chr>                   <dbl> <ord>             <int>
##  1  2019 casual                      1 Sunday             3766
##  2  2019 casual                      2 Monday             1892
##  3  2019 casual                      3 Tuesday            2728
##  4  2019 casual                      4 Wednesday          2489
##  5  2019 casual                      5 Thursday           2920
##  6  2019 casual                      6 Friday             3375
##  7  2019 casual                      7 Saturday           5993
##  8  2019 member                      1 Sunday            24233
##  9  2019 member                      2 Monday            48507
## 10  2019 member                      3 Tuesday           58277
## # i 18 more rows
```

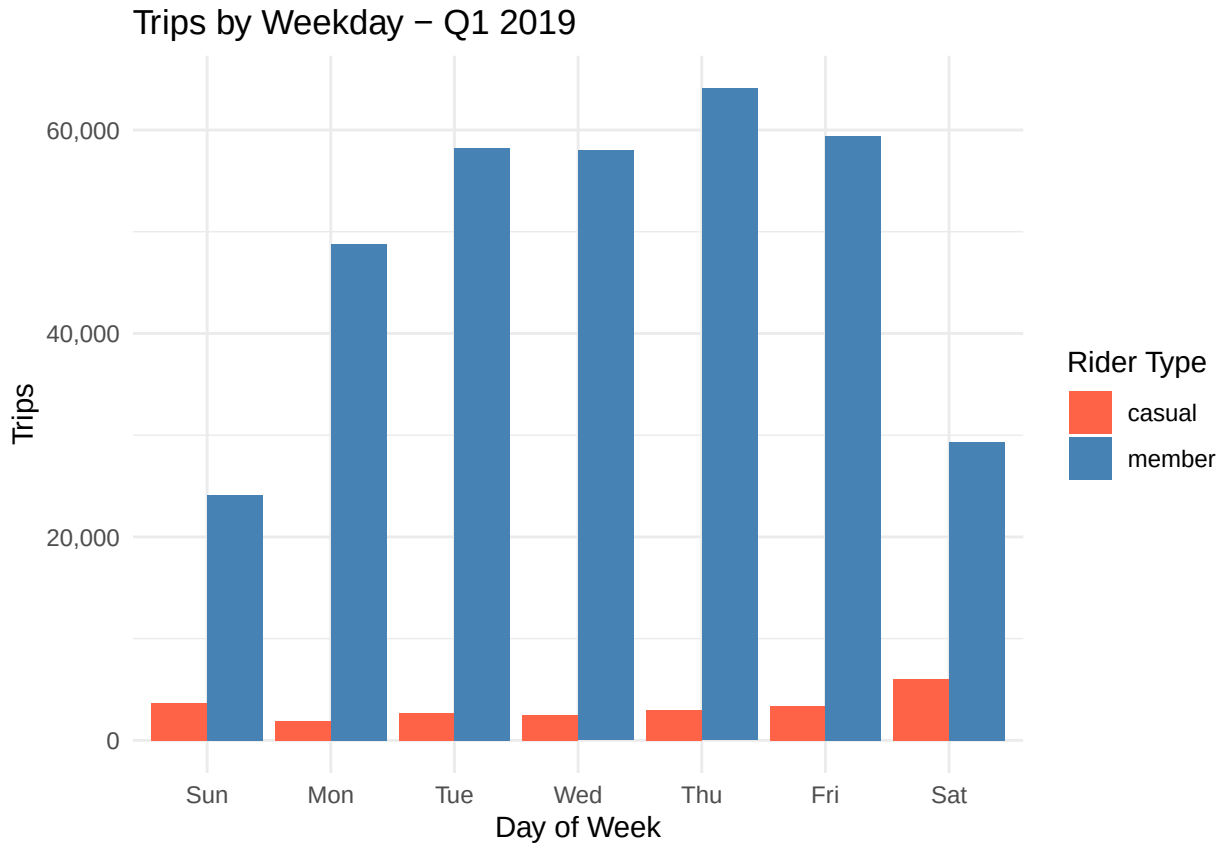
Weekday usage patterns reinforce this distinction. Members ride consistently throughout the workweek, peaking Monday–Friday, while casual riders spike heavily on weekends. This confirms members' commuter profile and casuals' leisure orientation.

Visualizations

Trips by Weekday Q1 - 2019

Bar Chart

```
df %>%
  filter(year == 2019) %>%
  count(member_casual, dow) %>%                                  # dow = Sun, Mon, ...
  mutate(dow = factor(dow, levels = c("Sun","Mon","Tue","Wed","Thu","Fri","Sat"))) %>%
  ggplot(aes(x = dow, y = n, fill = member_casual)) +
  geom_col(position = "dodge") +
  scale_fill_manual(values = c(casual = "tomato", member = "steelblue"), name = "Rider Type") +
  scale_y_continuous(labels = comma) +
  labs(x = "Day of Week", y = "Trips",
       title = "Trips by Weekday - Q1 2019") +
  theme_minimal()
```

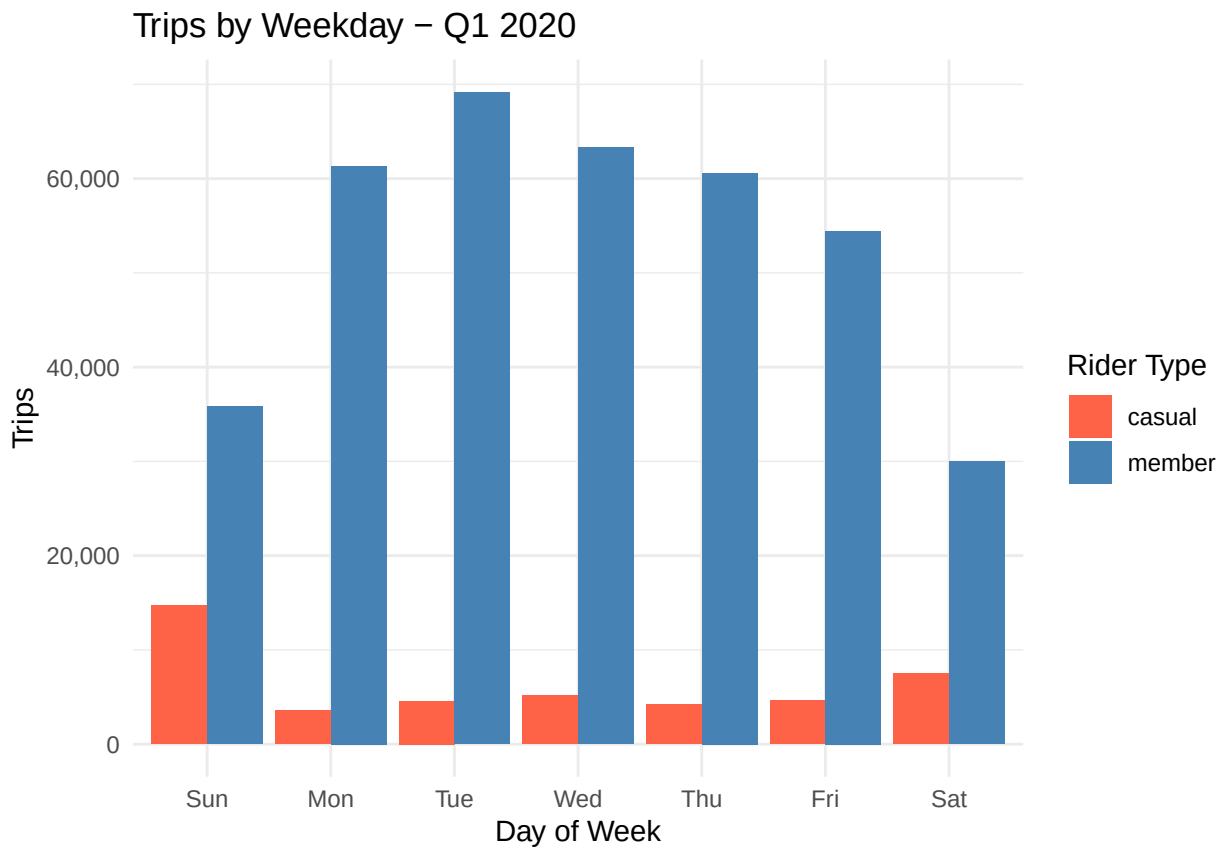


Members clearly dominate weekday usage, especially from Monday through Friday, which aligns with commuting patterns. Casual riders, by contrast, are most active on weekends (Saturday and Sunday), highlighting their stronger link to leisure and tourism.

Trips by Weekday Q1 - 2020

Bar Chart

```
df %>%
  filter(year == 2020) %>%
  count(member_casual, dow) %>%
  mutate(dow = factor(dow, levels = c("Sun","Mon","Tue","Wed","Thu","Fri","Sat"))) %>%
  ggplot(aes(x = dow, y = n, fill = member_casual)) +
  geom_col(position = "dodge") +
  scale_fill_manual(values = c(casual = "tomato", member = "steelblue"), name = "Rider Type") +
  scale_y_continuous(labels = comma) +
  labs(x = "Day of Week", y = "Trips",
       title = "Trips by Weekday - Q1 2020") +
  theme_minimal()
```



The 2020 weekday pattern is similar to 2019, but casual riders show even stronger weekend activity, particularly on Sundays. Members maintain steady weekday ridership, confirming their consistent commuter role, while casual riders reinforce their profile as weekend leisure users.

4) Trips by Hour of Day × Rider Type per Year

Table

```
hourly_table <- df %>%
  mutate(
    start_local = force_tz(started_at, "America/Chicago"),
    hour_local  = hour(start_local)
  ) %>%
  group_by(year, member_casual, hour_local) %>%
  summarise(total_trips = n(), .groups = "drop") %>%
  arrange(year, member_casual, hour_local)

hourly_table
```

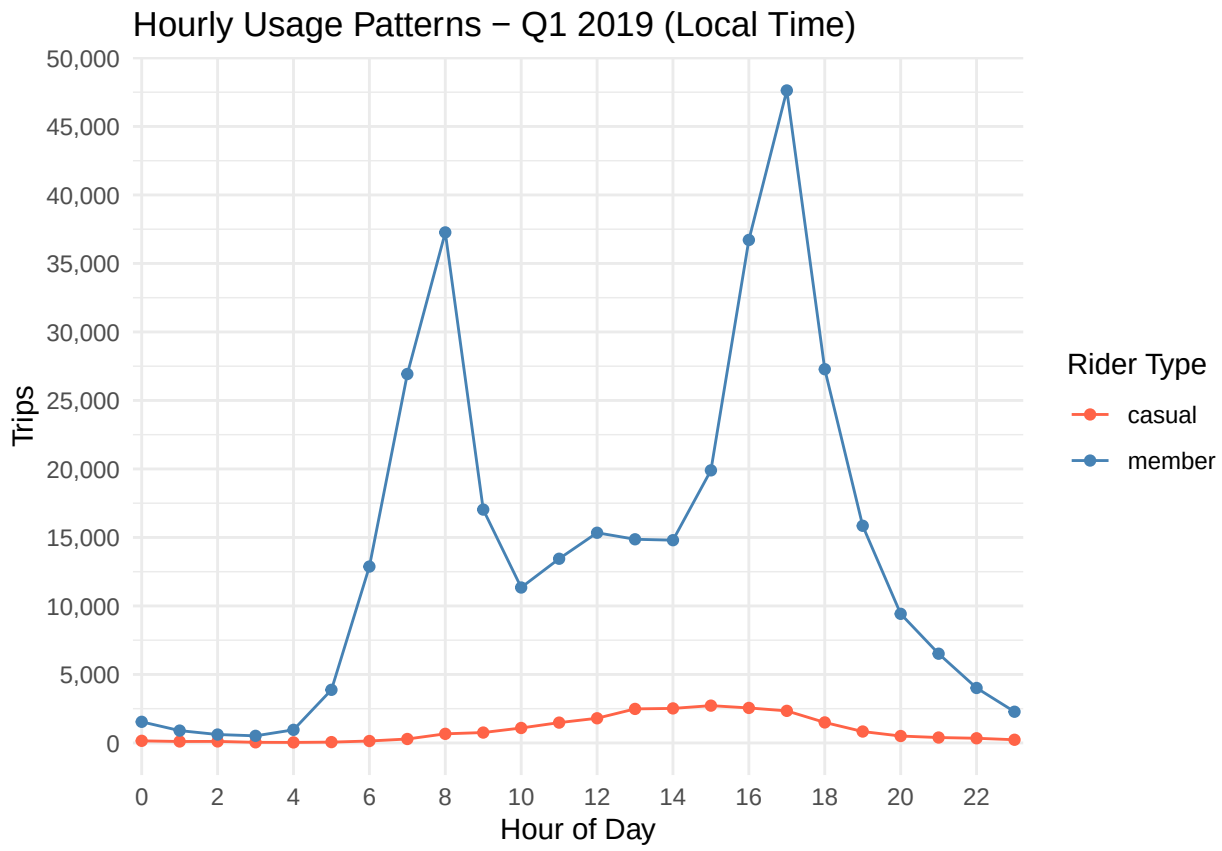
```
## # A tibble: 96 x 4
##     year member_casual hour_local total_trips
##    <dbl> <chr>              <int>       <int>
##  1  2019 casual                 0         153
##  2  2019 casual                 1         107
##  3  2019 casual                 2         111
##  4  2019 casual                 3          44
##  5  2019 casual                 4          37
##  6  2019 casual                 5          60
##  7  2019 casual                 6         141
##  8  2019 casual                 7         288
##  9  2019 casual                 8         664
## 10  2019 casual                 9         758
## # i 86 more rows
```

Hourly patterns show members clustering around commute hours: 7–9am and 4–6pm. Casual riders instead peak in the late morning and early afternoon. This split indicates that Cyclistic serves two distinct user segments with different time-based habits.

Visualizations

Hourly Usage Patterns - Q1 2019

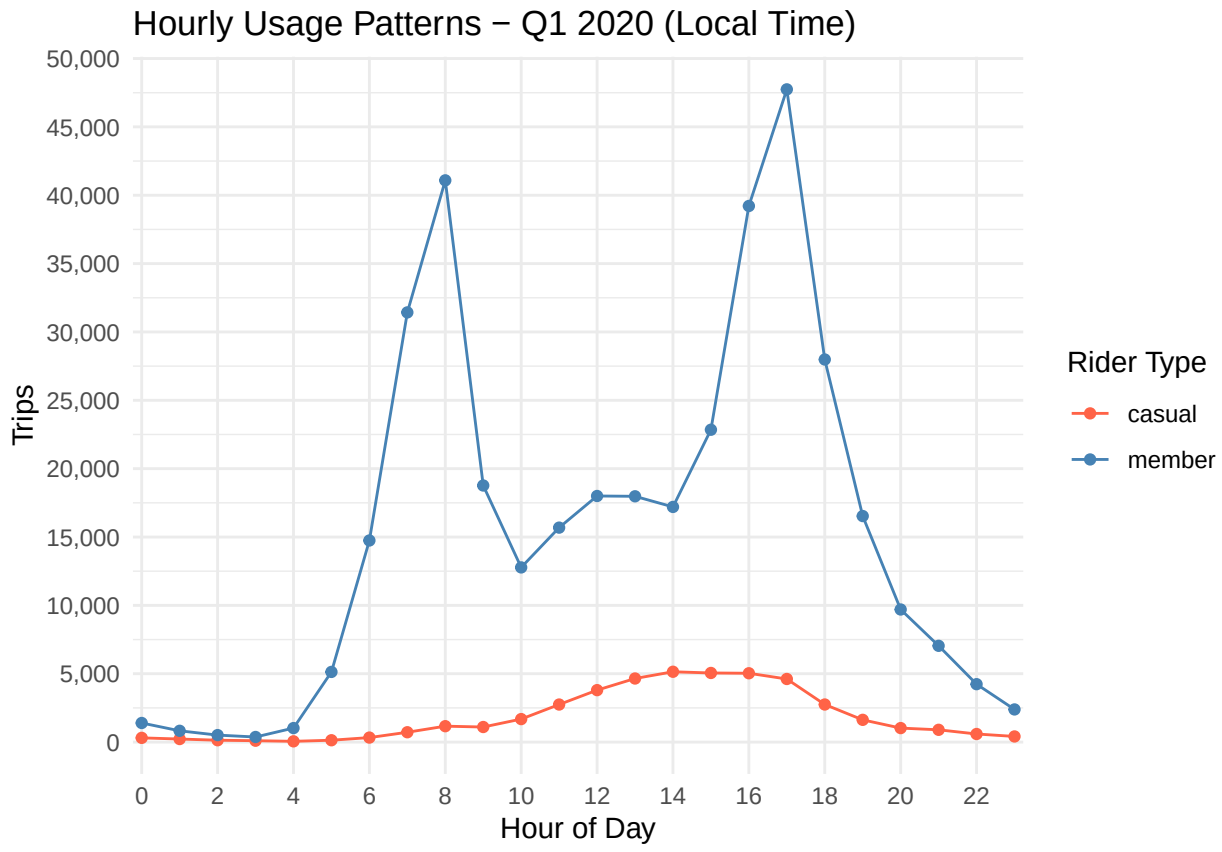
```
df %>%
  mutate(
    start_local = force_tz(started_at, "America/Chicago"),
    hour_local  = hour(start_local)
  ) %>%
  filter(year == 2019) %>%
  count(member_casual, hour_local) %>%
  ggplot(aes(x = hour_local, y = n, color = member_casual)) +
  geom_line() +
  geom_point(size = 1.5) +
  scale_x_continuous(breaks = seq(0, 23, 2), minor_breaks = NULL, expand = expansion(mult = c(0.01, 0.0
  scale_y_continuous(breaks = seq(0, 50000, 5000), labels = comma) +
  scale_color_manual(values = c(casual = "tomato", member = "steelblue"),
                     name = "Rider Type") +
  labs(x = "Hour of Day", y = "Trips",
       title = "Hourly Usage Patterns - Q1 2019 (Local Time)") +
  theme_minimal()
```



Member trips cluster heavily during commute windows, peaking sharply at around 8 AM and again around 5 PM. This reflects traditional commuting hours. Casual riders show a flatter pattern, with rides spread more evenly across late mornings and afternoons, consistent with leisure use.

Hourly Usage Patterns - Q1 2020

```
df %>%
  mutate(
    start_local = force_tz(started_at, "America/Chicago"),
    hour_local  = hour(start_local)
  ) %>%
  filter(year == 2020) %>%
  count(member_casual, hour_local) %>%
  ggplot(aes(x = hour_local, y = n, color = member_casual)) +
  geom_line() +
  geom_point(size = 1.5) +
  scale_x_continuous(breaks = seq(0, 23, 2), minor_breaks = NULL,
                     expand = expansion(mult = c(0.01, 0.01))) +
  scale_y_continuous(breaks = seq(0, 50000, 5000), labels = comma) +
  scale_color_manual(values = c(casual = "tomato", member = "steelblue"),
                     name = "Rider Type") +
  labs(x = "Hour of Day", y = "Trips",
       title = "Hourly Usage Patterns - Q1 2020 (Local Time)") +
  theme_minimal()
```



The same overall split is visible in 2020: members concentrate rides in the morning and evening rush hours, while casual riders peak during midday. Compared to 2019, casual ridership rises more noticeably in the late morning and early afternoon, perhaps reflecting growth in tourism or weekend activity.

5) Commute-Window Share (7–9 & 16–18)

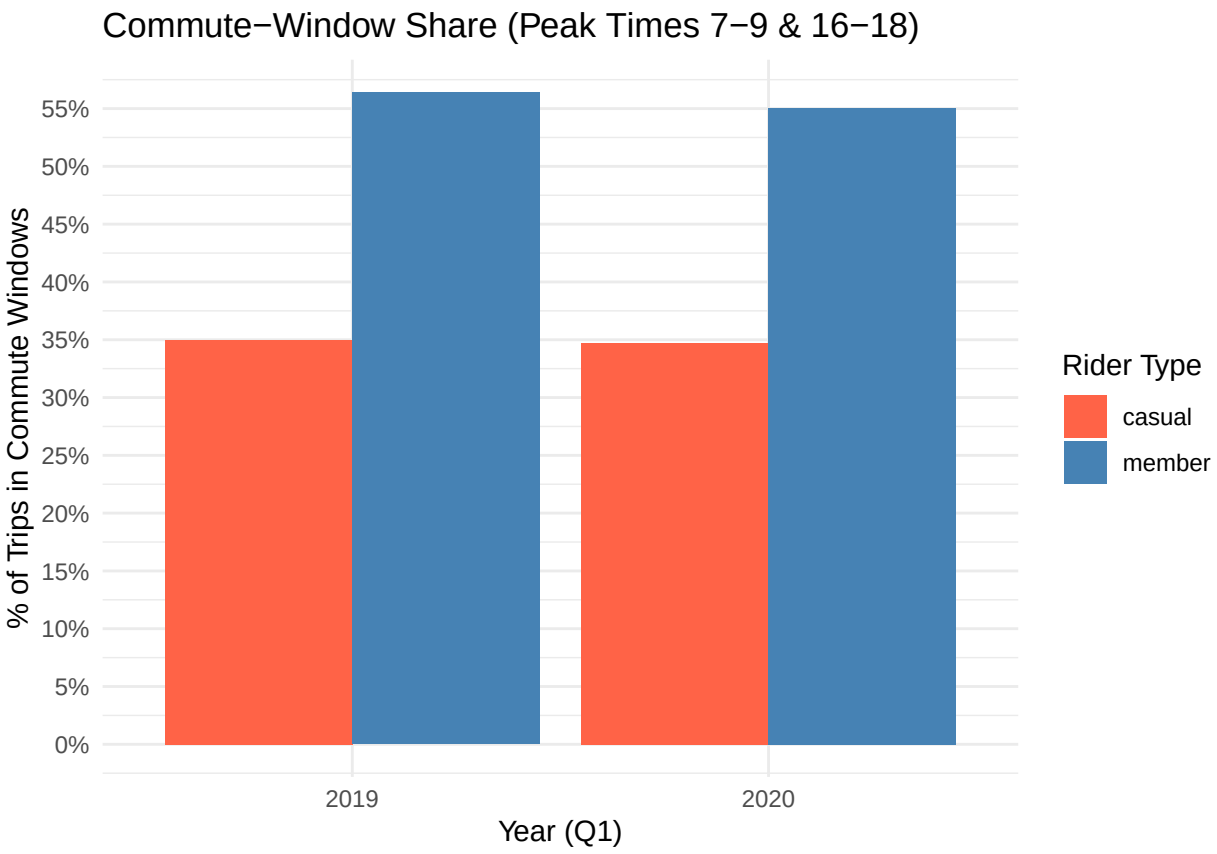
Build commute KPI table (local time)

```
commute_hours <- c(7, 8, 9, 16, 17, 18)

commute <- df %>%
  mutate(
    start_local = force_tz(started_at, "America/Chicago"),
    hour_local  = hour(start_local),
    is_commute  = hour_local %in% commute_hours
  ) %>%
  count(year, member_casual, is_commute) %>%
  group_by(year, member_casual) %>%
  summarise(
    pct_commute = sum(n[is_commute]) / sum(n),
    .groups = "drop"
  )
```

Visualization

Commute-Window Share during Peak Times

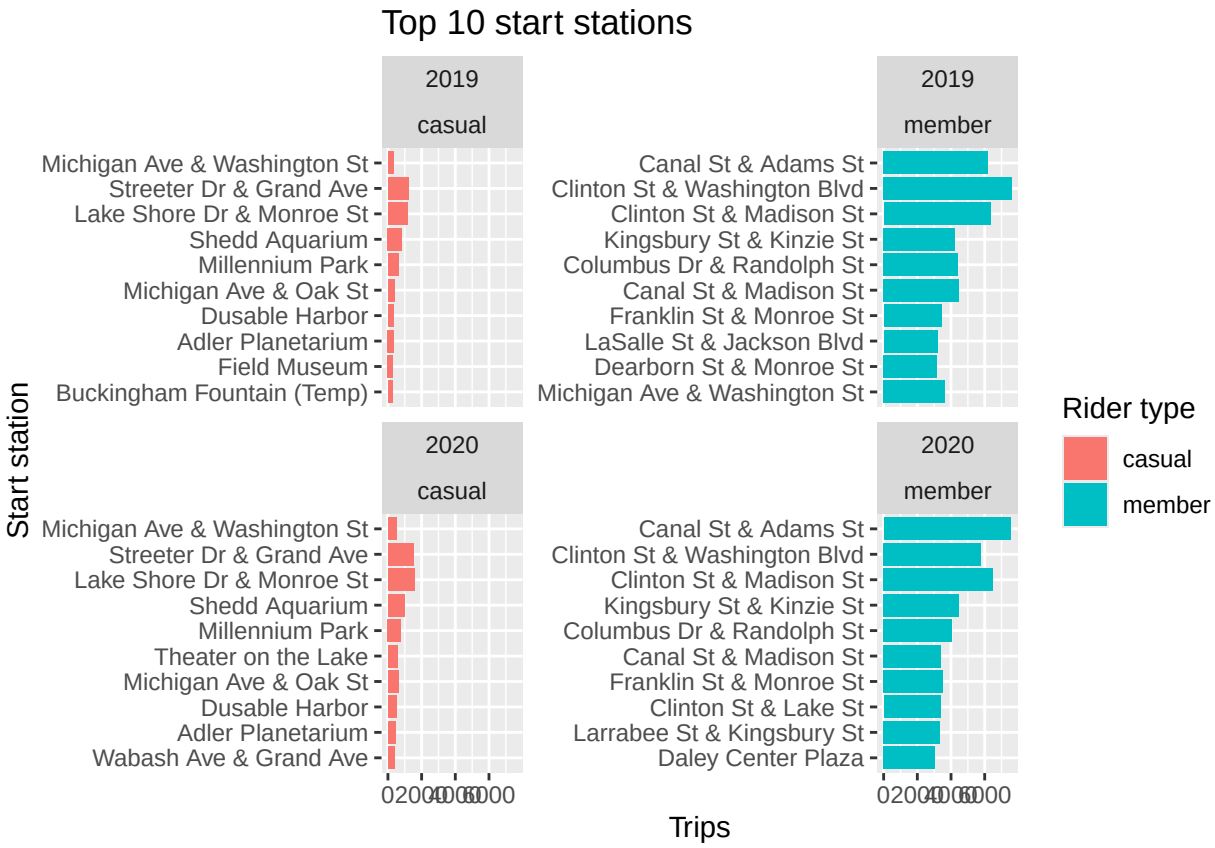
```
ggplot(commute, aes(x = factor(year), y = pct_commute, fill = member_casual)) +
  geom_col(position = "dodge") +
  scale_y_continuous(
    labels = scales::percent,
    breaks = seq(0, 1, 0.05)
  ) +
  scale_fill_manual(values = c(casual = "tomato", member = "steelblue"),
                    name = "Rider Type") +
  labs(
    x = "Year (Q1)",
    y = "% of Trips in Commute Windows",
    title = "Commute-Window Share (Peak Times 7-9 & 16-18)"
  ) +
  theme_minimal()
```



When quantified, over 50% of member trips fall within commute windows, compared to about 30–35% for casuals. This KPI reinforces that members are highly commute-driven, while casual riders are more flexible in their timing.

6) Top 10 Start Stations per Rider Type by Year

```
top_stations <- df %>%
  filter(!is.na(start_station_name), start_station_name != "") %>%
  count(year, member_casual, start_station_name) %>%
  group_by(year, member_casual) %>%
  slice_max(n, n = 10, with_ties = FALSE) %>%
  ungroup()

ggplot(top_stations,
       aes(x = reorder(start_station_name, n), y = n, fill = member_casual)) +
  geom_col() +
  coord_flip() +
  facet_wrap(year ~ member_casual, scales = "free_y") +
  labs(x = "Start station", y = "Trips", fill = "Rider type",
       title = "Top 10 start stations")
```



Top start stations also differ by rider type. Members' stations are concentrated near business districts and transit hubs, aligning with commuting. Casual riders' top stations cluster near parks, waterfronts, and tourist areas. This geographic difference can help target marketing more effectively.

Conclusion and Recommendations

Conclusion

The analysis clearly shows that **annual members and casual riders use Cyclistic bikes very differently**.
- **Members** ride more frequently, with trips concentrated during weekday commuting hours (7–9am and 4–6pm), and with shorter ride durations (typically under 20 minutes).
- **Casual riders** ride less frequently but for longer durations, peaking on weekends and midday, which aligns more with leisure and tourism.
- Geographic analysis further reinforces these patterns: members' top stations are near transit hubs and business districts, while casual riders prefer parks, waterfronts, and recreational areas.

These insights suggest that members are **commuter-driven**, while casuals are **experience-driven**. This distinction should shape Cyclistic's growth strategies.

Top Three Recommendations

1. **Targeted Membership Campaigns for Casual Riders**
    - Use promotions, weekend-to-weekday discounts, or trial memberships to encourage casual riders (who already use the service frequently for leisure) to see value in becoming members.
    - Messaging should emphasize flexibility, convenience, and cost savings for those who ride often.
2. **Tailored Marketing by Time and Location**
    - Focus **weekday commute promotions** (e.g., first-ride-free in the mornings) around business hubs to strengthen member ridership.
    - Develop **weekend leisure campaigns** (bundled group rides, tourist packages) to better capture casual riders' behaviors and gradually convert them.
3. **Improve Station Placement and Services**
    - Expand or prioritize bike availability near **transit-heavy stations** during weekday rush hours to support members.
    - Enhance services at **tourist and recreational stations** (signage, app guidance, bundled passes) to improve casual riders' experiences and incentivize repeat usage.

By acting on these findings, Cyclistic can **retain its core base of members while strategically converting casual riders into long-term members**, ultimately increasing revenue and maximizing bike utilization across the city.