Cyclistic Case Study

Understanding Members vs Casual Riders (Q1)

Analysis by: Spencer Cleland

November 26, 2025

Cyclistic Bike-Share Analysis Report

Cyclistic Customer Report

Understanding Customers

Cyclistic, a bike-share program in Chicago, wants to understand how casual riders and annual members differ in their usage patterns. The goal of this analysis is to uncover behavioral differences that can inform marketing strategies aimed at converting more casual riders into annual members.

This report addresses the assigned question:
- **How do annual members and casual riders use Cyclistic bikes differently?**

This report analyzes two datasets provided by the Divvy system:

- **2019 Q1 Trip Data**
- **2020 Q1 Trip Data**

These datasets include information about trip start and end times, ride duration, start and end stations, and rider type (member vs. casual).

A key limitation of these datasets is that bike type information is not comparable across years:

- The 2019 Q1 dataset contains **no** bike type column
- The 2020 Q1 dataset records **only** “docked_bike”

Therefore, **bike-type analysis is excluded** from this report.

All other analyses — ride length, time-of-day patterns, day-of-week usage, and seasonal trends — remain valid and are fully explored.

Data Preparation

Data Sources

This analysis uses two public Divvy datasets:

- **Divvy_Trips_2019_Q1.csv**
- **Divvy_Trips_2020_Q1.csv**

Both files contain trip-level data from Chicago's bike-share system, including timestamps, station IDs, station names, rider type, and trip duration.

Data Organization

The two datasets differ in several ways:

- Column names differ across years
- Rider type labels differ
- GPS coordinates are missing in 2019
- Bike type fields are incomplete

Because of these inconsistencies, bike-type analysis is excluded from the report. Standardization was required before merging datasets for analysis.

Data Credibility & License

The data comes directly from Divvy's public system logs and is considered:

- Reliable
- First-party
- Appropriate for usage analysis

These datasets are widely used in educational analytics programs and follow Divvy's public data-sharing license.

Data Privacy

The datasets contain **no personally identifiable information (PII)**. All rider identifiers are anonymized.

Limitations

Key limitations:

- Missing GPS coordinates for 2019
- Only Q1 months (winter) included
- Bike type fields inconsistent
- Schema differences required harmonization

Suitability

Despite limitations, the dataset is suitable to answer:

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

Data Processing

The Data Processing phase transforms the raw Divvy trip files into a clean, analysis-ready dataset.

Tools Used

- **RStudio Desktop**
- **tidyverse** → for data wrangling
  - **lubridate** → for timestamps
  - **ggplot2** → for visualization
- **kableExtra** → for tables
- **R Markdown** → for reproducible reporting

Processing Steps

1. Import raw datasets

```
q1_2019 <- read_csv("Divvy_Trips_2019_Q1.csv")
q1_2020 <- read_csv("Divvy_Trips_2020_Q1.csv")
```

2. Standardize column names (2019 → 2020 schema)

```
q1_2019_clean <- q1_2019 %>%
  rename(
    ride_id            = trip_id,
    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
```

```
  ) %>%
  mutate(
    started_at = ymd_hms(started_at),
    ended_at   = ymd_hms(ended_at),
    ride_id    = as.character(ride_id),
    member_casual = recode(
      member_casual,
      "Subscriber" = "member",
      "Customer"   = "casual"
    ),
    rideable_type = "unknown",
    ride_length   = as.numeric(difftime(ended_at, started_at, units = "mins"))
  )
```

3. Clean 2020 dataset

```
q1_2020_clean <- q1_2020 %>%
  mutate(
    started_at  = ymd_hms(started_at),
    ended_at    = ymd_hms(ended_at),
    ride_id     = as.character(ride_id),
    ride_length = as.numeric(difftime(ended_at, started_at, units = "mins"))
  )
```

4. Combine datasets

```
divvy_all <- bind_rows(q1_2019_clean, q1_2020_clean)
```

5. Filter invalid rides

```
divvy_v2 <- divvy_all %>%
  filter(ride_length > 0 & ride_length < 1440)
```

6. Add time-based variables

```
divvy_v2 <- divvy_v2 %>%
  mutate(
    day_of_week = wday(started_at, label = TRUE, abbr = FALSE),
    start_hour  = hour(started_at),
    month       = month(started_at, label = TRUE, abbr = TRUE),
    year        = year(started_at)
  )
```

7. Verify record count

```
nrow(divvy_v2)
```

```
## [1] 791264
```

Data Analysis

Overview of Analytical Findings

This analysis identifies how annual members and casual riders behave differently when using Cyclistic bikes. Several patterns clearly emerge:

- **Members take shorter, utilitarian rides** — clustered around commute hours.
- **Casual riders take longer, leisure-oriented rides** — often on weekends.
- **Temporal patterns differentiate rider groups** — including hour-of-day and weekday trends.
- **Trip durations are right-skewed** — requiring medians and distribution visualization for clarity.
- **Seasonality affects both groups** — though Q1 alone limits broad generalization.

The following sections detail these findings using tables, charts, and summary statistics.

Summary Tables

```
ride_length_summary <- divvy_v2 %>%
  group_by(member_casual) %>%
  summarise(
    mean_length   = mean(ride_length, na.rm = TRUE),
    median_length = median(ride_length, na.rm = TRUE),
    sd_length     = sd(ride_length, na.rm = TRUE),
    min_length    = min(ride_length, na.rm = TRUE),
    max_length    = max(ride_length, na.rm = TRUE),
    ride_count    = n(),
    .groups       = "drop"
  )

rides_by_day <- divvy_v2 %>%
  group_by(member_casual, day_of_week) %>%
  summarise(ride_count = n(), .groups = "drop") %>%
  mutate(day_of_week = factor(day_of_week,
```

```
                              levels = c("Sunday","Monday","Tuesday","Wednesday",
                                         "Thursday","Friday","Saturday")))

rides_by_hour <- divvy_v2 %>%
  group_by(member_casual, start_hour) %>%
  summarise(ride_count = n(), .groups = "drop")

rides_by_month_year <- divvy_v2 %>%
  group_by(year, month, member_casual) %>%
  summarise(ride_count = n(), .groups = "drop")

knitr::kable(
  ride_length_summary,
  caption = "Ride length summary by rider type (minutes)"
) %>%
  kableExtra::kable_styling(
    full_width = FALSE,
    bootstrap_options = c("striped", "hover")
  )
```

Table 1: Ride length summary by rider type (minutes)

member_casual	mean_length	median_length	sd_length	min_length	max_length	ride_count
casual	36.46320	21.966667	74.86274	0.0166667	1435.917	71138
member	11.40952	8.466667	21.20418	0.0166667	1433.067	720126

Create Plots

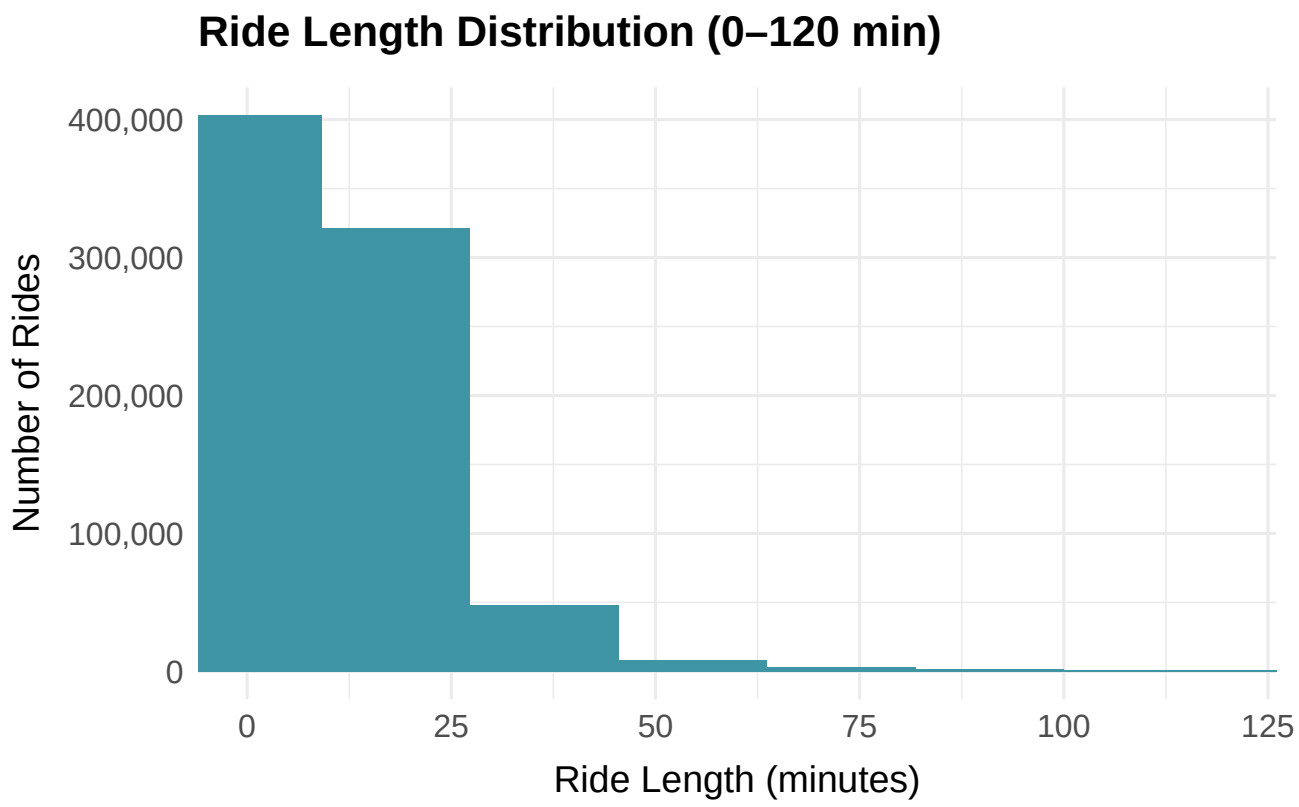
```
# Histograms
p1 <- ggplot(divvy_v2, aes(x = ride_length)) +
  geom_histogram(bins = 80, fill = cyclistic_colors["member"]) +
  scale_y_continuous(labels = scales::comma) +
  coord_cartesian(xlim = c(0, 120)) +
  labs(title = "Ride Length Distribution (0-120 min)",
       x = "Ride Length (minutes)", y = "Number of Rides")

p2 <- ggplot(divvy_v2, aes(x = ride_length + 1)) +
  geom_histogram(bins = 80, fill = cyclistic_colors["member"]) +
  scale_x_log10() +
  labs(title = "Ride Length Distribution (log scale)",
       x = "Ride Length (minutes, log scale)", y = "Number of Rides")

# Boxplot
p_box <- ggplot(divvy_v2,
                aes(x = member_casual, y = ride_length, fill = member_casual)) +
  geom_boxplot(outlier.size = 0.8) +
  coord_cartesian(ylim = c(0, 120)) +
```

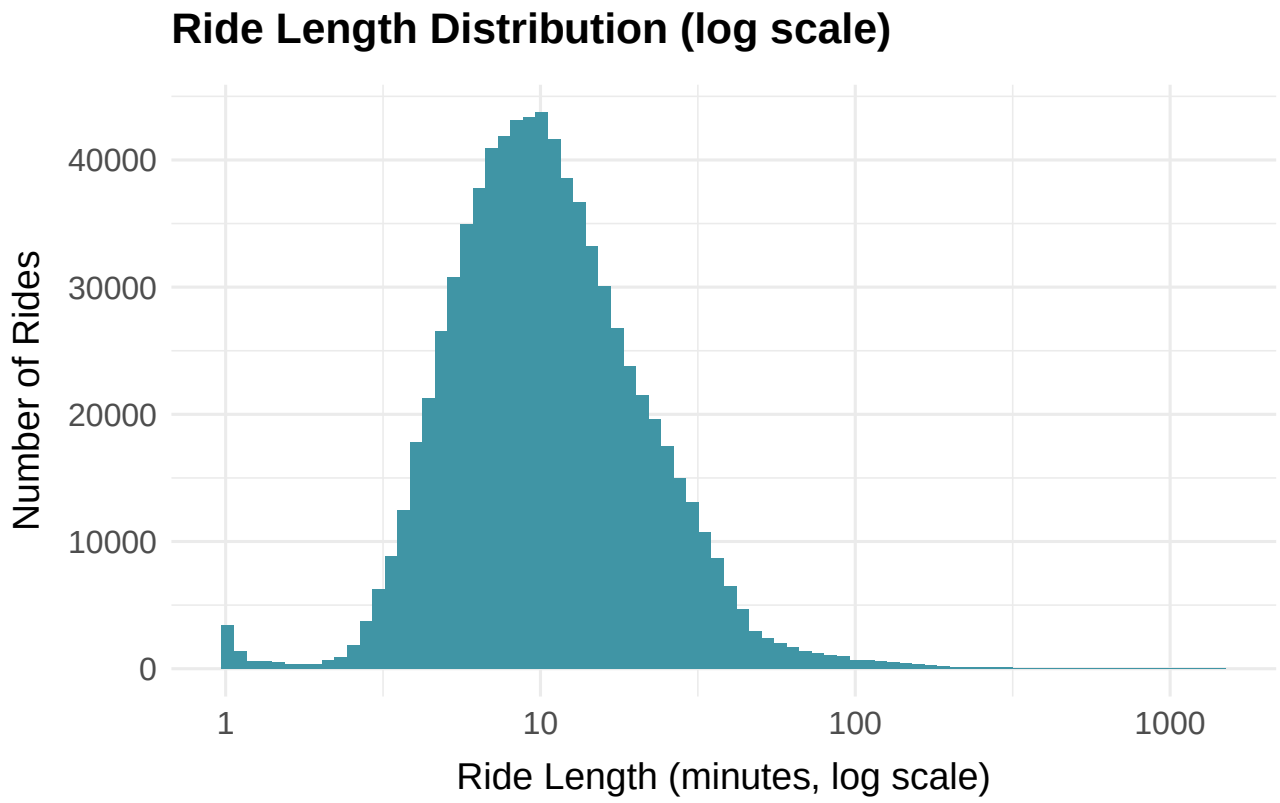
```
  scale_fill_manual(values = cyclistic_colors) +
  labs(title = "Ride Length by Rider Type (zoomed)",
       x = "Rider Type", y = "Ride Length (minutes)")

# Average ride length bar chart
p_avg <- ggplot(ride_length_summary,
                aes(x = member_casual, y = mean_length, fill = member_casual)) +
  geom_col() +
  geom_text(aes(label = round(mean_length, 1)),
            vjust = -1.2, size = 4) +
  scale_y_continuous(
    limits = c(0, max(ride_length_summary$mean_length) * 1.25)
  ) +
  scale_fill_manual(values = cyclistic_colors) +
  labs(title = "Average Ride Length: Members vs Casual Riders",
       x = "Rider Type", y = "Mean Ride Length (minutes)")

# Day-of-week bar chart
p_day <- ggplot(rides_by_day,
                aes(x = day_of_week, y = ride_count, fill = member_casual)) +
  geom_col(position = "dodge") +
  scale_y_continuous(labels = scales::comma) +
  scale_fill_manual(values = cyclistic_colors) +
  labs(title = "Ride Count by Day of Week",
       x = "Day of Week", y = "Number of Rides") +
  theme(
    axis.text.x = element_text(angle = 30, hjust = .75),
    legend.position = c(0.95, 0.95),
    legend.background = element_rect(fill = "white", color = NA),
    plot.margin = margin(15, 15, 25, 15)
  )

# Hour-of-day line chart
p_hour <- ggplot(rides_by_hour,
                 aes(x = start_hour, y = ride_count, color = member_casual)) +
  geom_line(size = 1.2) +
  geom_point(size = 2) +
  scale_x_continuous(breaks = 0:23) +
  scale_y_continuous(labels = scales::comma) +
  scale_color_manual(values = cyclistic_colors) +
  labs(title = "Ride Count by Hour of Day",
       x = "Hour (0-23)", y = "Number of Rides") +
  theme(
    legend.position = c(0.95, 0.95),
    legend.background = element_rect(fill = "white", color = NA),
    plot.margin = margin(15, 25, 15, 15)
  )

# Month-by-year faceted chart
p_month <- ggplot(rides_by_month_year,
                  aes(x = month, y = ride_count, fill = member_casual)) +
  geom_col(position = "dodge") +
  facet_wrap(~year, scales = "free_x") +
```

```
  scale_fill_manual(values = cyclistic_colors) +
  scale_y_continuous(labels = scales::comma, limits = c(0, max(rides_by_month_year$ride_count))) +
  labs(title = "Monthly Ride Counts by Year",
       x = "Month", y = "Number of Rides")
```

Ride Length Distributions

Understanding trip duration helps reveal how differently casual riders and annual members use the service. Across both datasets, ride lengths are highly right-skewed: most trips are short, but a small number of very long rides extend the distribution.



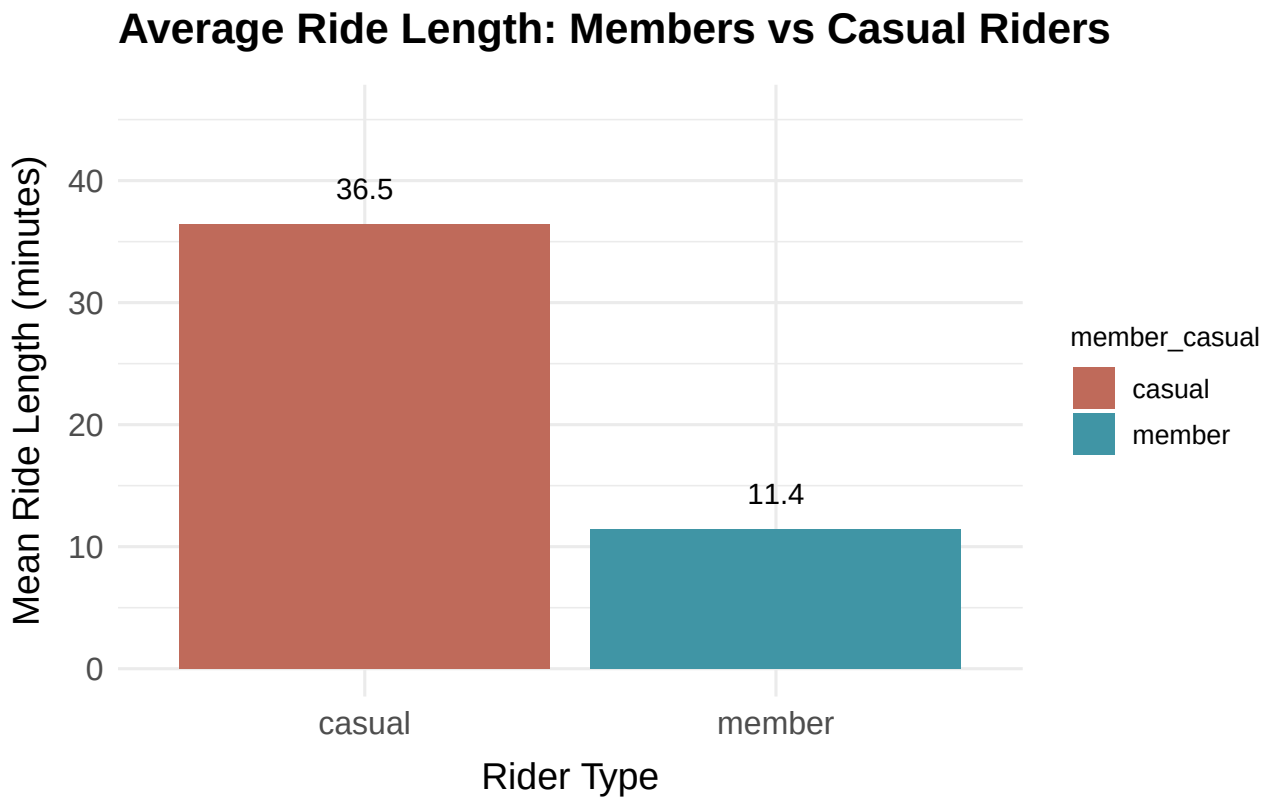


Interpretation:

- **Most rides last fewer than 30 minutes** — consistent with short-distance travel.
- **The median ride length is under 15 minutes** — but the long tail increases the mean.
- The log-scale histogram highlights differences in behavior:
  - Casual riders often take longer leisure rides
  - Members tend to make short, utilitarian trips

Ride duration strongly reflects differing motivations between members and casual riders.

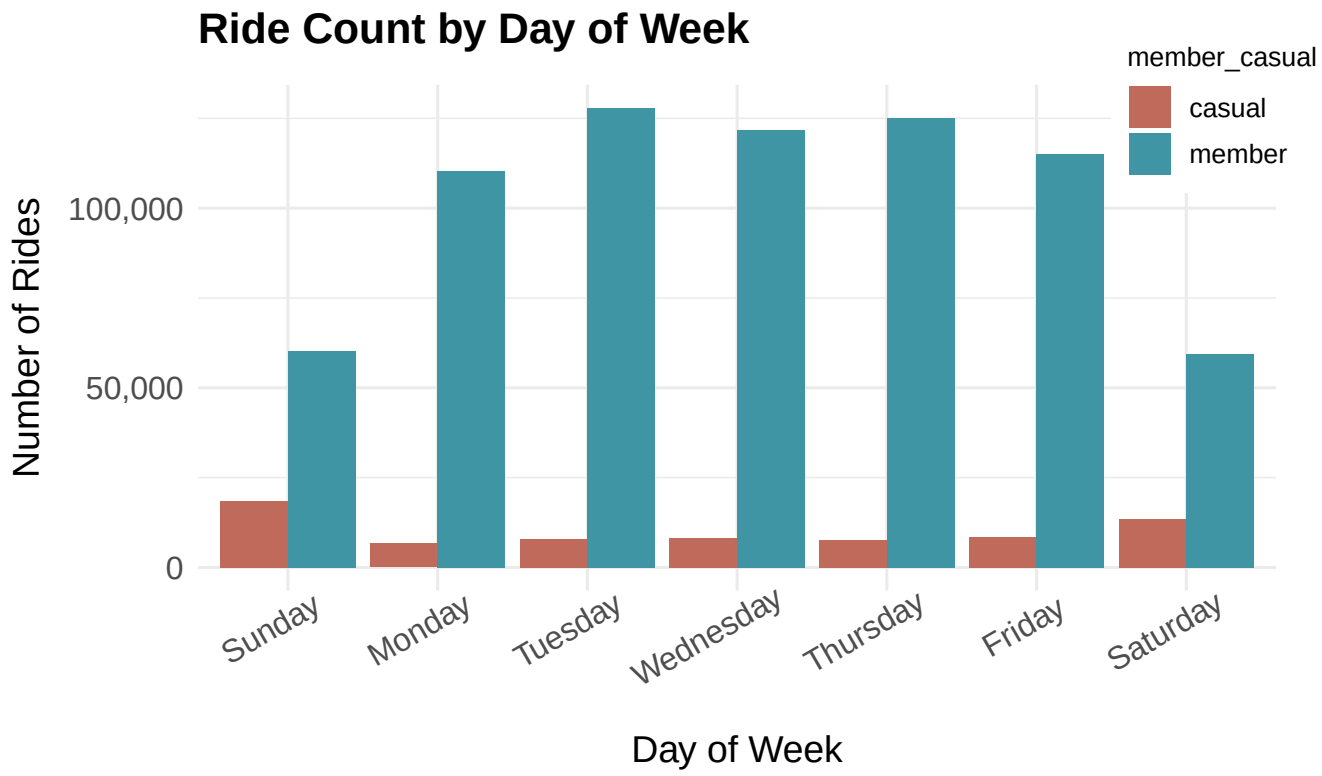
Ride Length by Rider Type



Interpretation:

Members take short, predictable rides — likely for commuting or errands.
Casual riders show broad variability and higher medians, consistent with leisure activities, sightseeing, or longer social rides.

Ride Counts by Day of Week

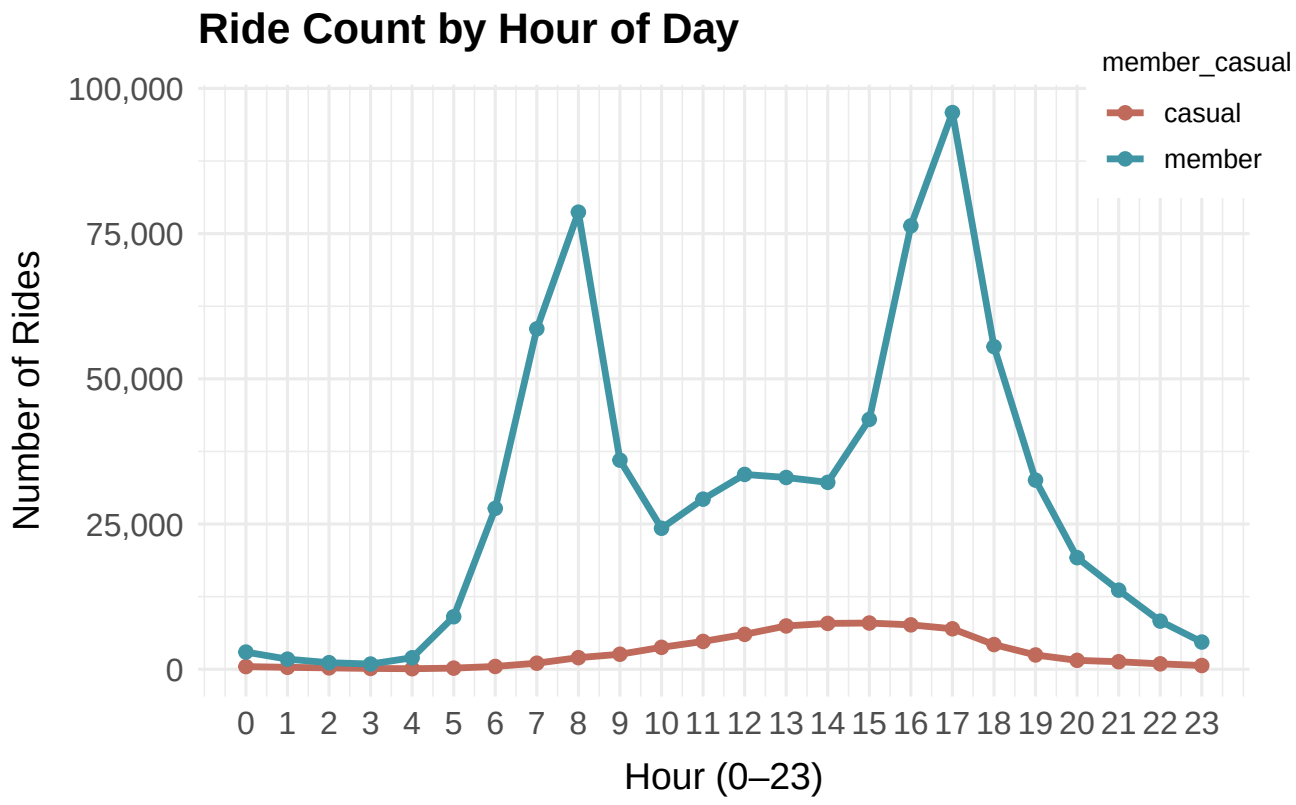


Interpretation:

- **Members:** Strong weekday usage pattern matching the work week
- **Casual Riders:** Weekend-heavy, leisure-based riding patterns

This suggests membership conversion strategies must target these groups at different times.

Hour-of-Day Trends

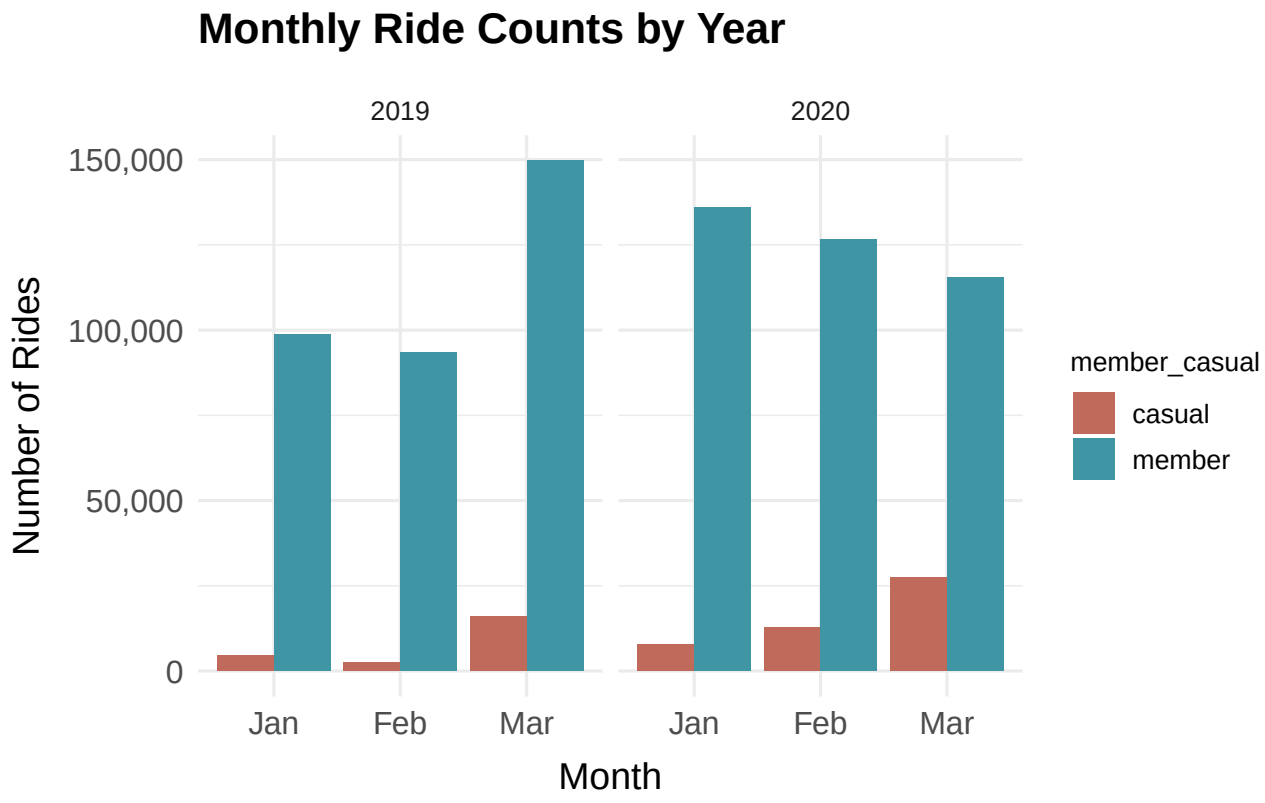


Interpretation:

- Members peak around **8 AM** and **5 PM** → traditional work commute behavior
- Casual riders avoid commute peaks and favor late morning, afternoon, and evening rides

Time-of-day patterns provide strong evidence of differing trip purposes.

Monthly Trends (Q1 Only)



Interpretation:

- Member usage increases from January → March in 2019.
  - Similar pattern started in 2020, but usage dipped in March *(may reflect early pandemic changes)*.
- Casual usage increase from January → March.

While Q1 alone limits seasonality conclusions, patterns can be seen across both years.

Key Insights

Analysis of Cyclistic's Q1 2019-2020 trip data shows clear behavioral differences between users:

1. **Members and casual riders behave differently.**
    - Members take shorter, structured trips (commuting, errands).
    - Casual riders take longer, variable, often leisure-oriented rides.
2. **Ride durations are right-skewed.**
   Medians give a truer picture of "typical" behavior than means.
3. **Members display routine weekday usage.**
   Strong morning/evening commute peaks indicate utility-driven trips.
4. **Casual riders are more seasonal and variable.**
   Their riding depends more on weekends, weather, and tourism cycles.
5. **Q1 winter season affects total ridership.**
   While winter suppresses volume, differences between rider types remain consistent.

These insights could support future campaigns aimed at converting casual riders to annual members.

Final Summary

Executive Summary

Q1 2019-2020 trip data reveals strong differences between annual members and casual riders:

- **Members:** short, frequent commuting trips
- **Casual riders:** longer, leisure trips, concentrated on weekends

Understanding these differences provides a foundation for targeted membership marketing.

Key Insights (At a Glance)

- Members use Cyclistic primarily for **transportation**
- Casual riders use it primarily for **recreation**
- Commute patterns vs. leisure patterns sharply divide usage

High-Level Recommendations

1. **Prioritize weekday vs. weekend analysis.**
2. **Explore trip purpuse.**
3. **Expand dataset for future phases.**

Future Strategic Implications

- **Membership marketing:** target weekday-heavy riders
- **Leisure-focused outreach:** send campaigns on weekends
- **Operational efficiency:** align bike distribution with peak times by rider type

---

Recommendations

1. Investigate weekday vs. weekend behavior more deeply

Members ride heavily during weekday commute hours, while casual riders concentrate their activity on weekends. Cyclistic should further analyze:

- Whether these patterns persist across warmer seasons
- How station availability differs between weekdays and weekends
- Whether promotional timing should differ for commuters vs. leisure riders

Understanding these temporal differences will help tailor marketing efforts to each group's routine.

2. Explore the drivers of longer casual rides

Casual riders consistently take longer trips, suggesting leisure, sightseeing, or exploratory travel. Cyclistic should evaluate:

- Whether longer rides cluster around tourist destinations or scenic areas
- How ride duration shifts by station, weather, or month
- Whether casual riders commonly ride in social groups

These insights would clarify what motivates casual riders and inform targeted messaging.

3. Expand analysis beyond Q1 to validate behavior year-round

Winter riding patterns may not reflect typical Cyclistic usage, especially for casual riders who are more sensitive to weather and seasonality. Cyclistic should consider analyzing:

- Full-year and multi-year datasets
- Warm-season riding trends
- The impact of newer bike types introduced after 2020
- Weather- or event-driven fluctuations in demand

A broader dataset would help confirm whether the observed behavioral differences hold across all seasons.

Final Recommendation Summary

Members tend to use Cyclistic bikes for **short, predictable, transportation-focused trips,** while casual riders use them for **longer, flexible, leisure-oriented experiences**.

Future analysis should aim to deepen Cyclistic's understanding of when, where, and why these patterns occur — setting the stage for strategic, data-driven membership conversion initiatives.

Appendix

Statistical Test: Difference in Ride Length

```
wilcox_res <- wilcox.test(ride_length ~ member_casual, data = divvy_v2)
wilcox_res
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  ride_length by member_casual
## W = 3.9568e+10, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

Interpretation:
A very small p-value indicates a statistically significant difference between ride length distributions, confirming that casual riders take longer rides on average.

Save Data & Plots

```
saveRDS(divvy_v2, file = "divvy_v2_clean.rds")

ggsave("plot_ride_length_histogram.png",     plot = p1,      width = 8,  height = 5, dpi = 300)
ggsave("plot_ride_length_histogram_log.png", plot = p2,      width = 8,  height = 5, dpi = 300)
ggsave("plot_ride_length_boxplot.png",       plot = p_box,   width = 8,  height = 5, dpi = 300)
ggsave("plot_average_ride_length.png",       plot = p_avg,   width = 8,  height = 5, dpi = 300)
ggsave("plot_rides_by_day.png",              plot = p_day,   width = 9,  height = 5, dpi = 300)
ggsave("plot_rides_by_hour.png",             plot = p_hour,  width = 9,  height = 5, dpi = 300)
ggsave("plot_rides_by_month_year.png",       plot = p_month, width = 10, height = 6, dpi = 300)
```

Changelog

```
changelog <- readLines("CHANGELOG.md")
cat(changelog, sep = "\n")
```

```
## Changelog - Cyclistic Case Study
## (Q1 2019-2020 data)
##
## - Imported and standardized 2019 Q1 and 2020 Q1 CSVs
## (column names aligned to 2020 schema).
##
## - Added ride_length (minutes) and re-coded member_casual from
## Subscriber/Customer &rarr member/casual.
##
## - Merged datasets into unified dataframe divvy_all.
##
## - Built check_type_mismatch() to flag column type inconsistencies before merging.
##
## - Corrected type mismatches for ride_id, started_at, ended_at, and rideable_type.
##
## - Identified invalid trips (negative or excessively long ride lengths) and
## filtered to 0-1440 minutes.
##
## - Produced cleaned dataset divvy_v2 with valid durations only.
##
## - Summarized data cleaning results, noting removal of 0.087% invalid rows and
## confirming stable mean/median ride lengths.
##
## - Added year to enriched dataset.
##
## - Updated cleaned dataset to explicitly define rideable_type for 2019 rides
## (missing in source data).
```