Cyclistic Bike-Share Case Study

A Data Analysis on Converting Casual Riders to Annual Members

Prepared by:
Nicolas Levesque

Date:
March 13, 2025

Submitted to:
Cyclistic Executive Team
Director of Marketing: Lily Moreno

Confidential and Proprietary – For Internal Use Only

Table of Contents

---

1. Executive Summary

Cyclistic, a bike-share company in Chicago, aims to **increase profitability** by converting **casual riders** into **annual members**. Through an in-depth analysis of trip data over the last 12 months, we identified key behavioral differences:

- **Longer but fewer rides** by casual riders (often on weekends/afternoons).
- **Shorter, more frequent rides** by annual members (often on weekdays/commuting hours).
- **Strong seasonal bias** for casual riders (peaking in summer, dropping significantly in winter).
- **Station usage differences** (casual riders favor recreational areas, members frequent commuter hubs).

1.1 Key Recommendations

1. **Target High-Usage Casual Riders**
    - Use tailored incentives for those taking multiple or longer rides.
2. **Flexible Membership Options**
    - Offer weekend/afternoon-focused membership plans appealing to leisure riders.
3. **Seasonal Marketing Campaigns**
    - Capture casual riders in summer and incentivize them to continue riding year-round with winter retention perks.

Implementing these strategies can **enhance user retention**, **grow revenue**, and **strengthen Cyclistic's market position**.

---

2. Ask (Business Task)

Cyclistic's **business objective** is to **convert casual riders into annual members**. To achieve this, we must:

- Understand **how casual riders and members use Cyclistic bikes differently**.
- Identify **why casual riders might purchase annual memberships**.
- Determine **how digital media** and strategic marketing can **influence casual riders** to become members.

---

3. Prepare (Data Sources)

1. **Divvy Trip Data (12 Months)**
    - Publicly available historical bike trip data from Motivate International Inc.
    - Includes trip start/end times, station locations, bike types, and rider types (casual or member).
2. **Data Reliability**
    - **ROCCC**:
        - **Reliable**: Direct from Motivate, a trusted source.
        - **Original**: No third-party modifications.
        - **Comprehensive**: Covers key ride attributes (duration, location, etc.).
        - **Current**: Most recent 12 months.
        - **Cited & Licensed**: Public data with no personal identifiers.
3. **Potential Biases**
    - **Seasonal Bias** (summer usage may be overrepresented).
    - **Geographic Bias** (stations in tourist areas may inflate casual usage).
    - **Rider Type Distribution** (members might be overrepresented in certain datasets).

---

4. Process (Data Cleaning)

Below is a summary of the five main cleaning steps that were performed on the data.

4.1 Data Cleaning 1: Merging & Initial Prep

- **Imports**: 12 months of raw CSV files.
- **Process**:
    1. Merged all monthly CSVs into a single dataset for consistency.
    2. Inspected structure (e.g., column names, data types).
- **Outcome**: Saved as `cyclistic_data_merged.rds`.

4.2 Data Cleaning 2: Imputing Missing Locations

- **Imports**: `cyclistic_data_merged.rds`
- **Process**:
    1. Removed rows with missing end coordinates.
    2. Created a reference table of station IDs and averaged lat/long.
    3. Used k-nearest neighbors (FNN) to impute missing start/end station IDs.
    4. Filled missing station names by matching the most frequent name per station ID.
- **Outcome**: Saved as `cyclistic_cleaned_filled.rds`.

4.3 Data Cleaning 3: Standardizing Text Fields

- **Imports**: `cyclistic_cleaned_filled.rds`
- **Process**:
    1. Checked for spelling/capitalization inconsistencies in station names.
    2. Normalized text (trimming, lowercasing).
    3. Ensured one-to-one mapping between station ID and station name.
- **Outcome**: Saved as `cyclistic_cleaned_standardized.rds`.

4.4 Data Cleaning 4: Fixing Data Types & Duplicates

- **Imports**: `cyclistic_cleaned_standardized.rds`
- **Process**:
    1. Converted `rideable_type` and `member_casual` to factor variables.
    2. Identified duplicate ride IDs and removed them using `distinct()`.
- **Outcome**: Saved as `cyclistic_cleaned_deduplicated.rds`.

4.5 Data Cleaning 5: Outliers & Final Refinements

- **Imports**: `cyclistic_cleaned_deduplicated.rds`
- **Process**:
    1. Computed ride length (minutes) by `ended_at - started_at`.
    2. Removed negative durations and rides over 24 hours.
    3. Eliminated zero or sub-one-minute rides (likely data errors).
    4. Capped ride lengths at 2 hours based on distribution analysis.
    5. Validated coordinates and excluded invalid lat/long entries.
    6. Added a `day_of_week` column derived from `started_at`.
- **Outcome**: Saved as `Cyclistic_cleaned_final.rds` (ready for analysis).

---

5. Analyze & Share: Key Findings and Visuals

5.1 Ride Duration

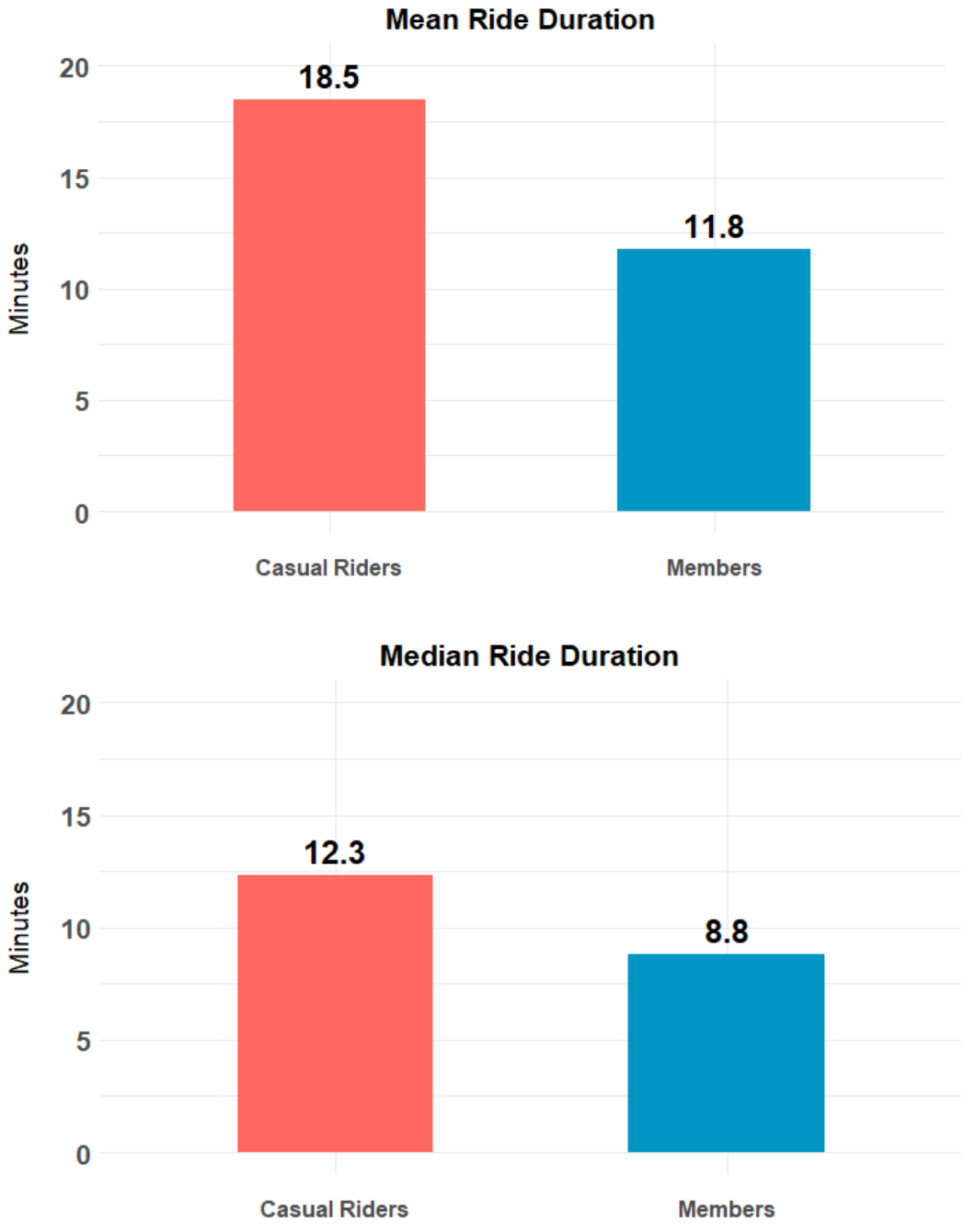

Figure 1a: Mean and Median Ride Durations for casual riders vs. members

- **Casual riders** average 18.5 minutes (median 12.3), whereas **members** average 11.8 minutes (median 8.8), confirming that casual riders tend to take longer rides overall.

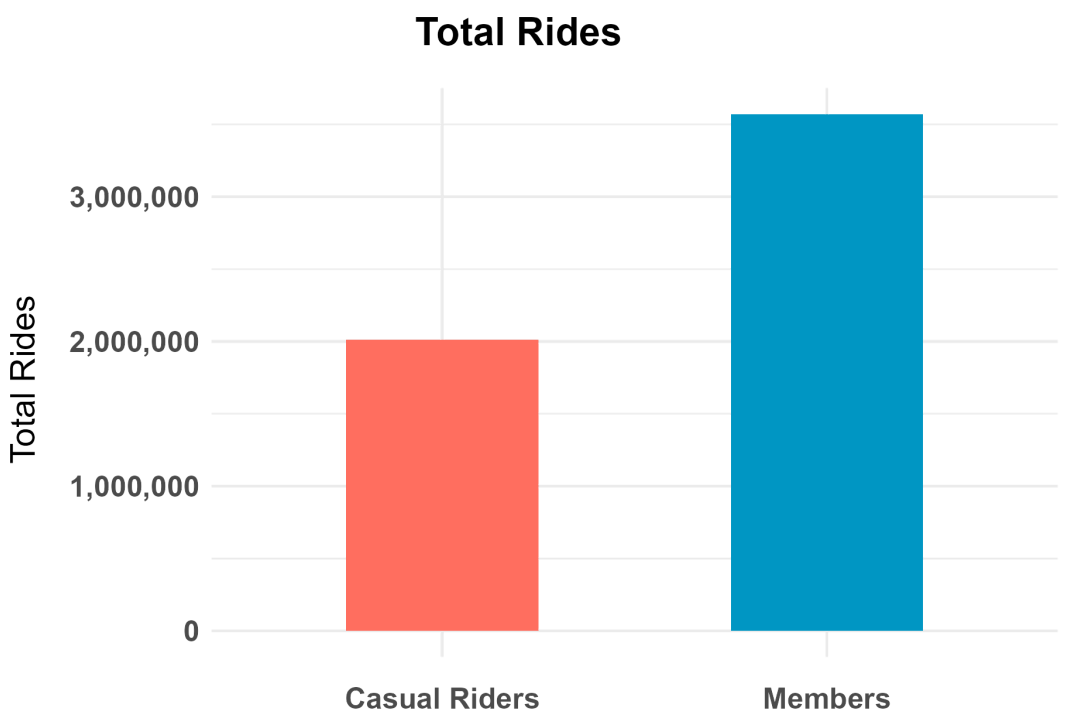

Figure 1b: Illustrates the total number of rides taken by casual riders vs. members over a 12-month period.

*While there is a bias in the data due to a higher number of total member rides, the data still shows **value** as it indicates that memberships are correlated with greater system-wide ride activity.*

5.2 Peak Usage Patterns

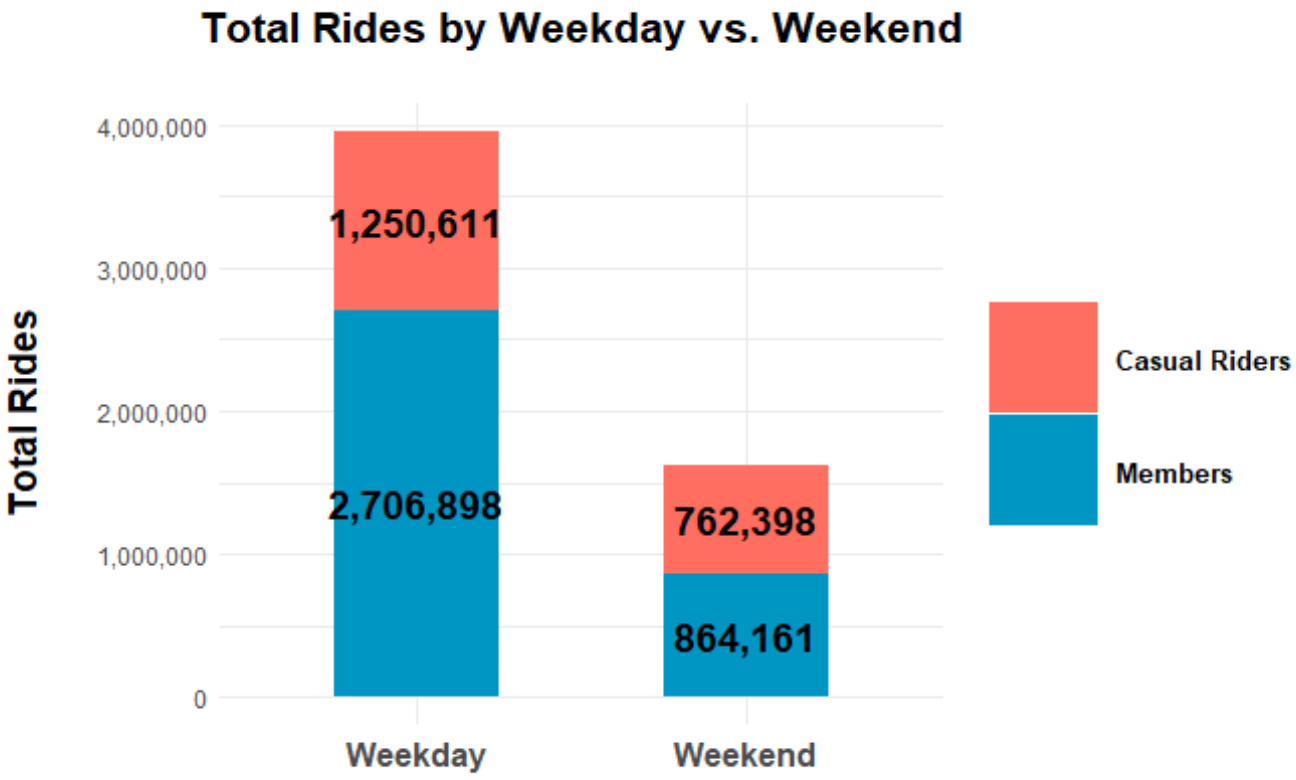

Figure 2a: Ride Distribution by Day of Week
*Indicates that **members** ride primarily on **weekdays**, whereas **casual riders** prefer **weekends**.*

- Although the total number of rides on weekends is lower than on weekdays, there are only two weekend days compared to five weekdays.
- **On a per-day basis, casual riders average more rides on weekends than they do on weekdays**, confirming their stronger weekend preference.

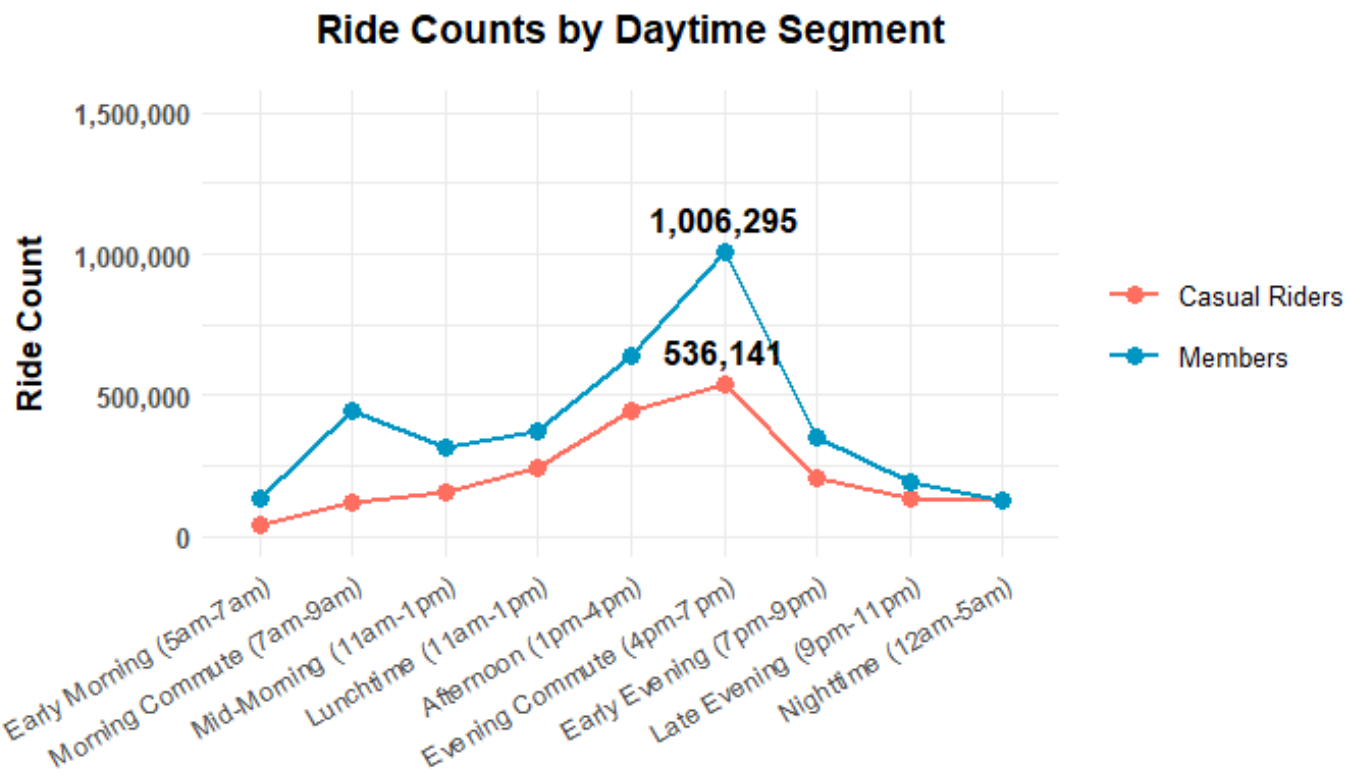

Figure 2b: Ride Distribution by Time of Day

- **Members** ride primarily during **commuting hours** (7–9 AM, 4–7 PM).
- **Casual riders** ride mostly from lunchtime into the afternoon, and until the early evening (11 AM–9 PM).

*There are several potential **biases** that could affect the conclusions of this analysis, such as no individual user tracking, membership vs. casual labeling bias, seasonal and event-based effects, and system availability and external factors.*

5.3 Station & Route Preferences

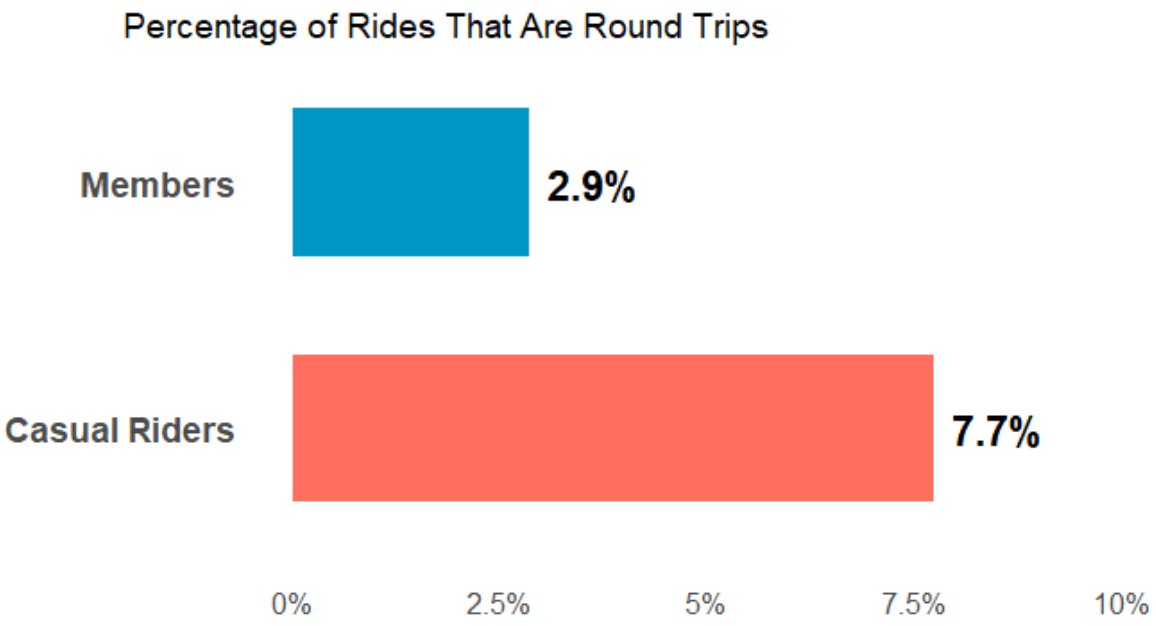

Figure 3a: Round Trip Percentage by Rider Type
Demonstrates that casual riders start and end at the same station more than members do.

- **Casual riders** are **more likely** to take **round trips**, starting and ending at the same station.

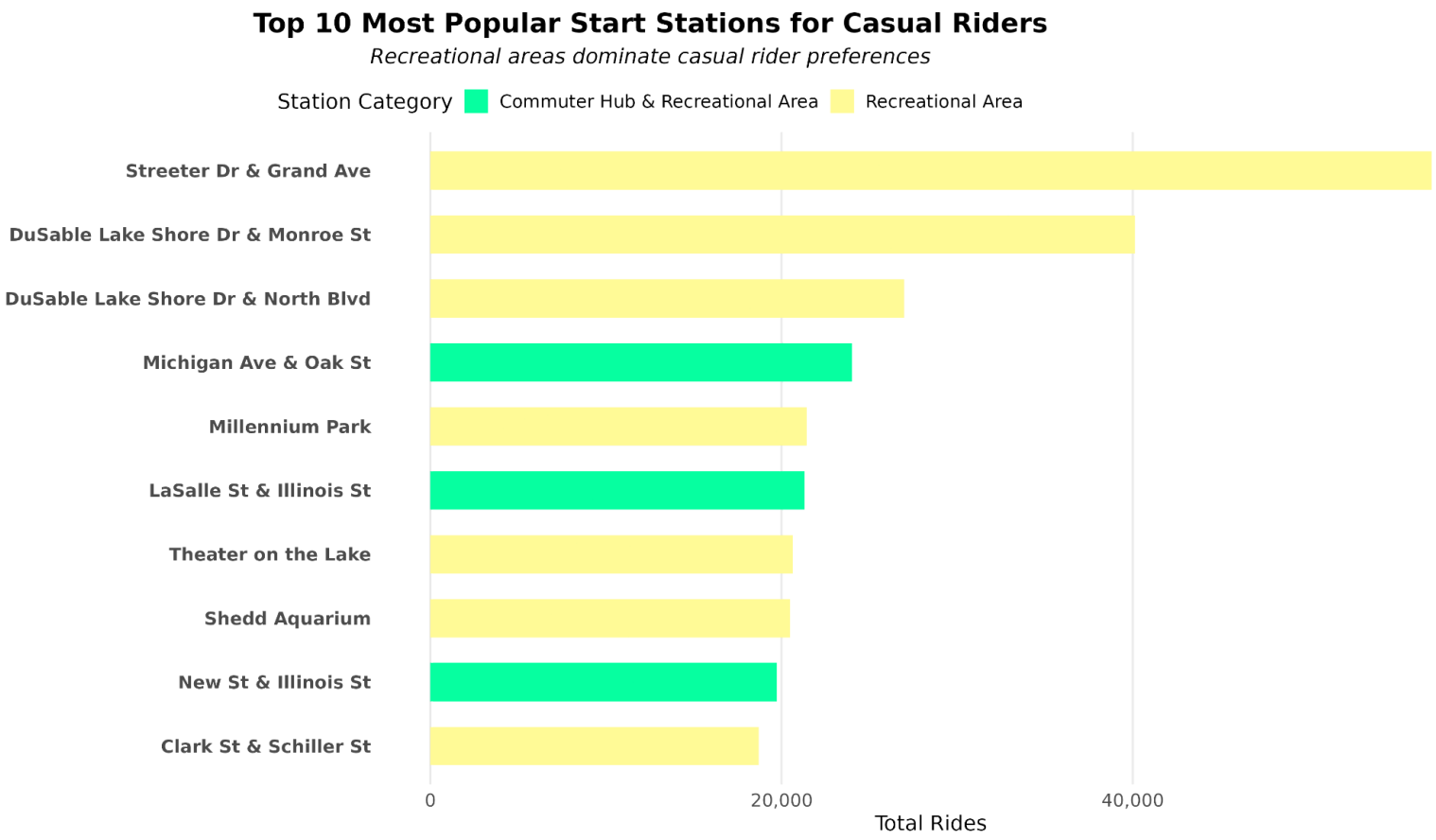

Figure 3b: Top Stations for casual riders
Shows how casual riders favor recreational stations.

- **Casual riders** frequent stations near recreational areas, e.g., **parks, waterfronts, and tourist areas**.

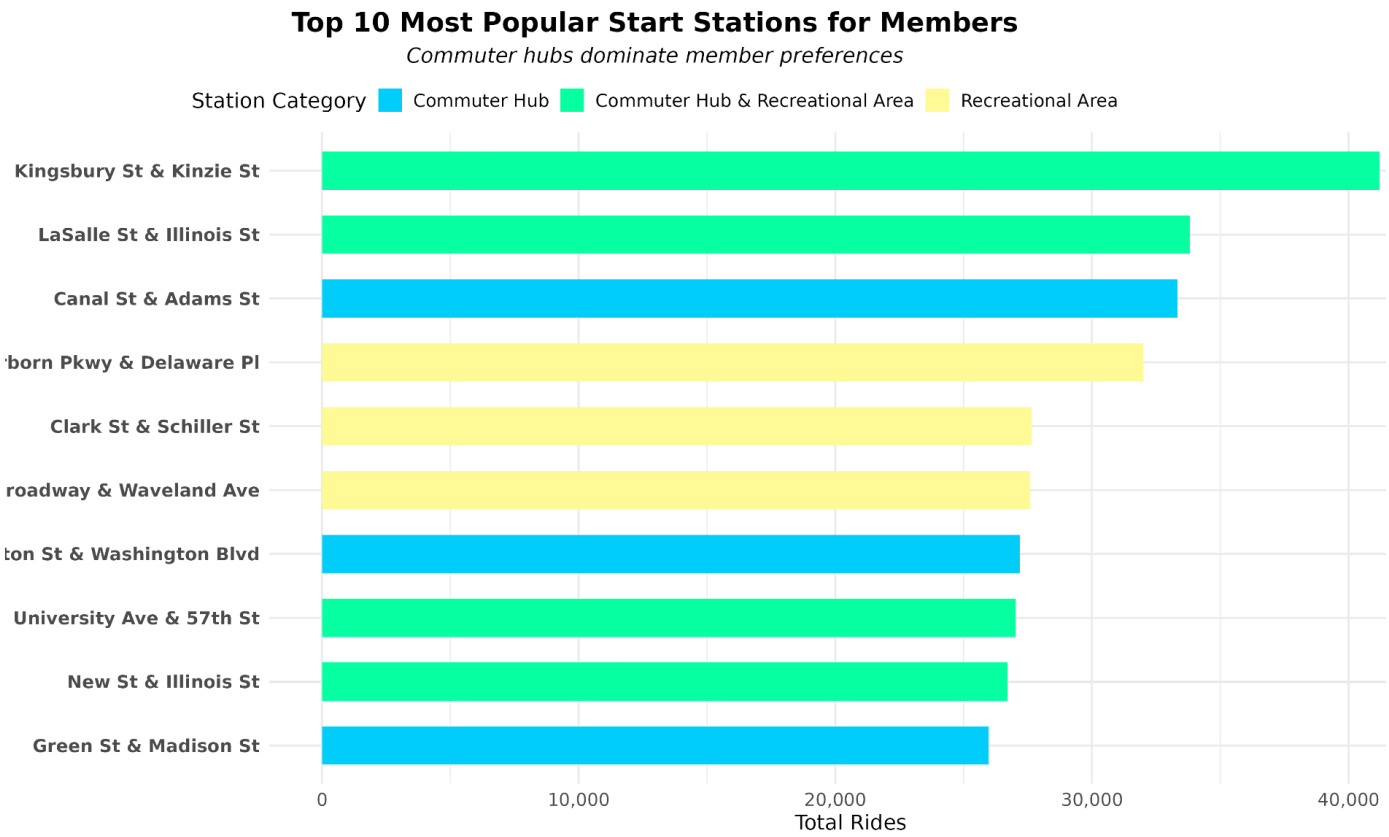

Figure 3c: Top 10 Start Stations for members
Shows how members favor transit hubs.

- **Members** predominantly start at **transit hubs and business districts**.

Analysis of Potential Bias: The study avoids assumptions regarding member commuter status or casual rider leisure usage. Instead, it prioritizes the observation and analysis of round-trip patterns among casual riders. Emphasis is placed on leisure demand, mitigating imposition of a commuter-centric perspective. Conversion strategies are aligned with empirically observed usage patterns, specifically the prevalence of round trips and station category preferences.

5.4 Seasonal Ride Trends

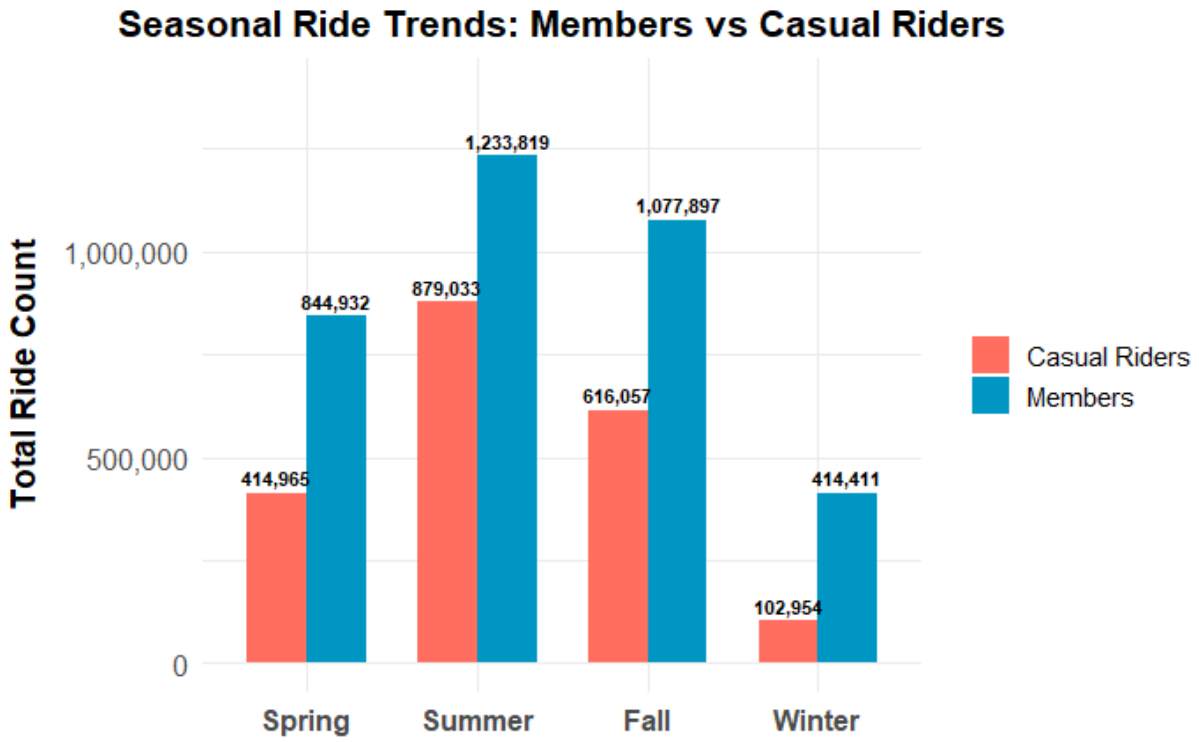

Figure 4a: Seasonal Breakdown groups rides by season (spring, summer, fall, winter) Illustrates the seasonal ride trends of casual riders vs. members' usage.

- Summer shows the highest number of casual rides (**879,033**), more than double the winter count (**102,954**).
- This confirms that **casual riders primarily ride for leisure during warm weather**, with summer being the dominant season.

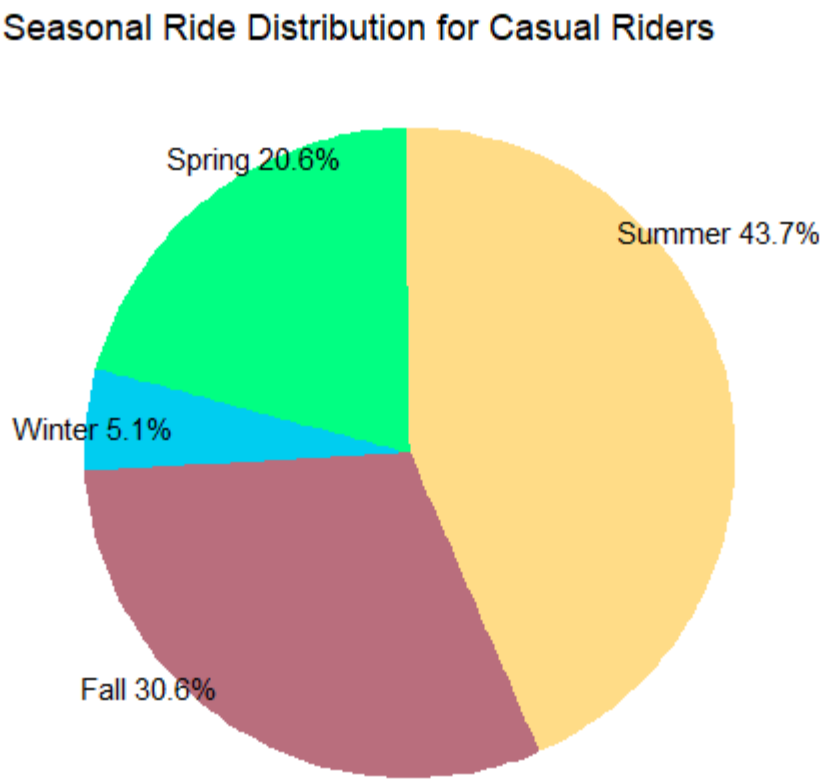

Figure 4b: Seasonal Ride Distribution for casual riders
Displays heavy preference by casual riders for summer riding.

- **Casual rider** demand peaks in **summer** and drops significantly in **winter**, suggesting leisure use during warmer weather.

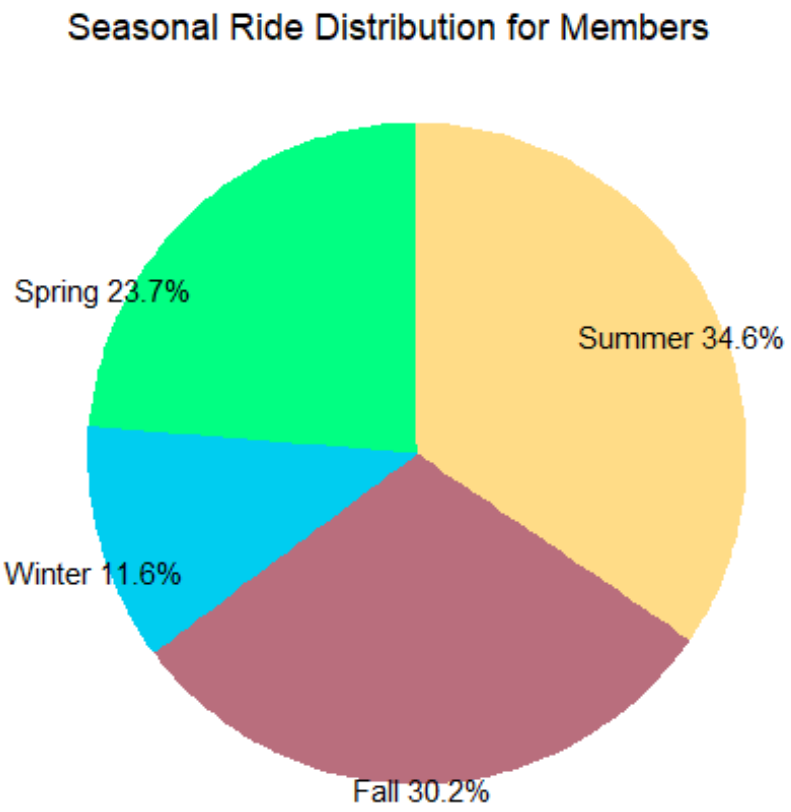

Figure 4c: Seasonal Ride Distribution for members
Displays more balanced year-round usage by members

- **On the other hand, members** ride **consistently** throughout the year, with a moderate drop in winter, likely due to weather, indicating more balanced year-round usage compared to casual riders.

The data may be biased due to availability, infrastructure, and self-selection bias.

---

6. Act (Recommendations)

Based on the analysis, here are the **top three recommendations** to convert casual riders into annual members:

1. **Target High-Usage Casual Riders**

    - Identify casual users who frequently ride or take longer rides and offer personalized membership discounts or loyalty rewards.

2. **Introduce Flexible & Leisure-Focused Membership Options**

    - Offer weekend passes or “round-trip bundles” that cater to leisure patterns.
    - Provide perks at tourist hotspots or partner with local attractions to incentivize membership for leisure riders.

3. **Seasonal Marketing Campaigns**

    - **Summer**: Launch membership drives at the beginning of peak season to capture casual riders at their most active.
    - **Fall**: “Keep Riding” discounts to bridge the drop-off into winter.
    - **Winter**: Trial memberships with winter-friendly perks (e.g., gear discounts, heated docks if available).

---

7. Conclusion & Next Steps

By **merging and cleaning** 12 months of Cyclistic ride data, we uncovered critical behavioral differences between **casual** and **member** riders. These insights guide us to:

1. **Develop targeted promotions** that align with casual riders' leisure-based habits.
2. **Improve retention** by making memberships attractive beyond the peak summer season.
3. **Leverage station-level and seasonal data** to enhance availability and marketing outreach.

7.1 Next Steps

- **Refine Pricing Models**: Test "Round-Trip Saver" or "Weekend Warrior" passes.
- **A/B Test Marketing Campaigns**: Assess the impact of seasonal promotions on conversion rates.
- **Track Conversions**: Continuously measure how many casual riders switch to memberships, iterating on the strategies as needed.

This plan provides a **data-driven roadmap** for Cyclistic to **boost membership** and **increase profitability** while maintaining an excellent user experience for both casual and member riders.

Thank you for reviewing this case study. With these strategies and analyses in hand, Cyclistic is well-positioned to grow its annual member base and continue its mission of providing accessible, convenient bike-share services to the city of Chicago.

---

8. Appendix: Cyclistic Bias Analysis Documentation

This section outlines the steps and methods used to assess potential biases in the Cyclistic dataset, ensuring robust and reliable insights for the case study.

8.1 Dataset Loading

- **Data Source:** The analysis begins by loading the final cleaned dataset (`cyclistic_cleaned_final.rds`) that was previously prepared.
- **Purpose:** This dataset serves as the foundation for evaluating representation, time period, and geographic biases.

8.2 Representation Bias

- **Objective:** Determine if there is an imbalance between casual and member riders.
- **Method:**
    - Counted the number of rides for each user type using a frequency table.
    - Calculated the proportions of casual vs. member rides.
- **Key Findings:**
    - **Casual riders:** Approximately 36% of total rides.
    - **Members:** Approximately 64% of total rides.
- **Interpretation:** While there is a moderate imbalance favoring members, casual riders still form a significant portion of the dataset. This balance is critical for drawing meaningful conclusions about both segments.

8.3 Time Period (Seasonality) Bias

- **Objective:** Assess whether ride volumes vary significantly over different periods, indicating a potential seasonal bias.
- **Method:**
    - Extracted the month from the `started_at` timestamp to create a new `month` column.
    - Visualized ride counts by month, segmented by user type (casual vs. member) using a bar chart.
- **Key Findings:**
    - **Seasonal Trends:** Peak usage occurs during the summer months (June–September), with a marked decline during the winter (November–February).

    - **User Behavior:** Casual riders show a stronger seasonal trend compared to members.
- **Interpretation:** The heavy seasonal variation, particularly among casual riders, must be considered in subsequent analyses and when designing conversion strategies.

8.4 Geographic Bias

- **Objective:** Examine if certain docking stations disproportionately influence the dataset.
- **Method:**
    - Grouped data by `start_station_name` and user type.
    - Computed the total ride counts for each station and identified the top 10 busiest stations.
- **Key Findings:**
    - A few high-traffic stations account for a large share of rides.
    - Specific stations are dominated by casual riders, while others are more frequented by members.
- **Interpretation:** The geographic concentration at particular stations suggests that urban or high-tourism areas might be overrepresented. This should be considered when making location-based recommendations.

8.5 Additional Analysis: Rideable Type

- **Objective:** Explore whether preferences for different bike types (e.g., classic, electric, docked) vary by user type.
- **Method:**
    - Counted rides by `rideable_type` and user type.
    - Converted counts to proportions and visualized the distribution using a bar chart.
- **Key Findings:**
    - **Electric Bikes:** The most popular choice for both segments, with a slight preference among members.
    - **Classic Bikes:** More commonly used by casual riders.
    - **Electric Scooters:** Minimal use overall.
- **Interpretation:** Understanding these preferences can inform targeted promotions (e.g., incentivizing casual riders to try electric bikes) and impact pricing and infrastructure planning.

This bias analysis ensures that seasonal, geographic, and representation factors are explicitly accounted for, providing context for interpreting the main analytical findings and refining marketing strategies.

9. References

- **Motivate International Inc.** Divvy Data License Agreement. Retrieved from https://www.divvybikes.com/data-license-agreement