# Cyclistic Bike-Share Case Study

## Exploratory Data Analysis and Business Insights

---

**Submitted by:**
Sibeso Like
Associate Data Analyst
GitHub Portfolio: Sibeso Like

**Date:**
July 24, 2025

---

**Tools Used:**

- ➔ Microsoft SQL Server
- ➔ R & RStudio
- ➔ Excel
- ➔ Data Visualization (ggplot2)

**Data Source:**
Divvy Public Bike-Share Data
(Cyclistic, operated by Motivate International Inc.)

---

**Project Focus:**
“How do annual members and casual riders use Cyclistic bikes differently?”

---

# Table of Contents

# Executive Summary

**Cyclistic Bike-Share User Behavior Analysis – July 2025**

This report presents a comprehensive data analysis project undertaken to explore user behavior differences between **annual members** and **casual riders** of Cyclistic, a fictional bike-share company operating in Chicago, USA. The primary objective was to uncover actionable insights that can inform marketing strategies aimed at increasing annual memberships.

**Business Problem**

Cyclistic's success depends on long-term engagement. While casual riders contribute significantly to ride volume, annual members represent stable and recurring revenue. The central question of this analysis was:

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

Answering this question would support Cyclistic's strategic goal of converting casual riders into paying annual members.

**Data Overview**

- **Data Source**: Publicly available monthly trip data from Cyclistic for a full calendar year.
- **Volume**: Over 5 million rows of data covering ride IDs, user types, start/end times, station locations, and ride durations.

**Tools Used**:

- **SQL Server** for data storage, cleaning, and structured querying.
- **R & R Markdown** for exploratory data analysis (EDA) and visualization.
- **Azure Data Studio** for data cleaning & analysis documentation.
- **Microsoft Excel** for minor pre-processing and verification.

**Data Cleaning**

Data was cleaned and transformed to ensure consistency, accuracy, and usability:

- Fixed station ID formats (e.g., removed invalid entries like WL-008).
- Removed null values and duplicate ride entries.
- Standardized datetime formats for ride duration and time-based analysis.

## Key Findings

- **Ride Duration**: Casual riders had significantly longer average ride durations compared to annual members, suggesting a leisurely vs. commuting usage difference.
- **Peak Usage Time**: Casual riders peaked on **weekends**, while members had higher usage on **weekdays**, indicating commuting patterns.
- **Popular Stations**: Distinct popular start/end stations for each group pointed to differing travel preferences.
- **Hourly Usage**: Casual riders were more active in **midday hours**, while members showed peaks during **morning and evening commutes**.

## Strategic Recommendations

1. **Optimize Station Placement for Casual Riders:**

**Insight***:* Casual riders frequently start and end rides at leisure destinations like: - New St & Illinois St - Theatre on the Lake - DuSable Habor

- Expand bike dock availability and ensure higher bike supply at these high-demand leisure locations, especially on weekends.
- Introduce seasonal pop-up stations or mobile bike hubs near beaches, parks, or event venues to meet weekend demand.

2. **Offer Flexible Membership Options**

**Insigh***t:* Casual users take longer rides (avg. 25 mins vs. 12 mins for members) and ride more on weekends.

- Introduce a weekend or "Leisure Rider" membership plan tailored for occasional long-distance riders.
- Allow for hour bundles or credits that can roll over, offering a middle ground between casual and full memberships.

3**. Commute Focused Infrastructure**

**Insigh***t:* Members ride most during commute hours (7–9 AM, 4–6 PM) on weekdays.

- Partner with employers and transit authorities to promote bike-to-work incentives (e.g., lockers, showers, or discounts).

## Conclusion

The analysis confirms clear behavioral differences between user groups. These insights support the hypothesis that tailored marketing strategies can increase member conversion rates. With data-backed decision-making, Cyclistic can boost customer loyalty, optimize operations, and maximize profitability.

# Chapter 1 - Introduction

### Background of the Business or Dataset

Cyclistic is a fictional bike-share company based in Chicago, USA that offers thousands of bicycles for public use through a network of docking stations. The service is used by both casual riders, who pay per trip, and annual members who pay a flat fee for unlimited access during their subscription. The dataset used in this case study consists of 12 months of historical trip data collected from Cyclistic's public records. It includes key information such as ride IDs, timestamps, start/end stations, user types, and location coordinates.

To gain business value from this data, it is essential to examine user behavior, usage patterns, and trends that may indicate opportunities for increasing the number of loyal, paying members.

### Objective of the Case Study

The main objective of this analysis is to **identify how annual members and casual riders use Cyclistic bikes differently**. By understanding these differences, Cyclistic's marketing team aims to develop data-driven strategies to **convert casual riders into annual members**, thus increasing recurring revenue and user retention.

### Importance of the Analysis

This analysis is crucial because it provides Cyclistic with:

- **User segmentation insights** for tailored marketing.
- **Data-driven decision-making support** to optimize operations.
- **Patterns in behavior** that can inform pricing, promotions, and service availability.

### Business Task

Cyclistic's executive team has identified an opportunity to grow revenue by converting more **casual riders into annual members**. The marketing director believes that understanding the differences in how these two groups use the bike-sharing service can inform a **targeted marketing strategy** that appeals to casual users and encourages them to subscribe.

Data Analyst Tasks:

- Analyze historical trip data to uncover **usage patterns and behavioral differences** between casual riders and annual members.
- Provide **data-backed insights** that can support the development of **effective marketing campaigns** aimed at user conversion.
- Deliver findings in the form of **visualizations, summarized metrics, and actionable recommendations** to guide the executive and marketing teams.

This task aligns with the company's strategic goal of increasing the number of annual memberships, which provide more predictable revenue and customer retention.

# Chapter 2 - Data Collection

### Data Sources and Acquisition Methods

The data used in this analysis was obtained from **Motivate International Inc.**, the provider of bike-share data for Cyclistic. The datasets are publicly available via the **Divvy Bike Sharing Data Portal** and are intended for the purpose of analysis and non-commercial use.

Data was downloaded in **CSV format**.

### Description of Datasets Used

The analysis covered a **12-month period**, capturing seasonal and usage variations. The following details summarize the data:

- **Months Covered:** July 2025 to June 2025
- **File Format:** CSV
- **Number of Files:** 12 (one for each month)
- **Total Size:** Approximately **1.06 GB** combined
- **Total Rows After Merging:** Over **5.6 million ride records**
- **Key Columns:**
    - ride_id, rideable_type, started_at, ended_at
    - start_station_name, end_station_name
    - member_casual (user type)
    - ride_duration, day_of_week, start_hour (created during cleaning)

All datasets were merged into a single table in SQL Server to form a **master dataset**, which was then used for data cleaning, exploration, and visualization.

# Chapter 3 - Data Cleaning & Manipulation

Before conducting meaningful analysis, the raw data underwent a comprehensive cleaning process to ensure accuracy, consistency, and reliability. This step was essential for handling missing values, correcting data types, standardizing station IDs, and generating new features such as ride duration, day of the week, and hour of ride start. All cleaning operations were executed using **SQL Server**, following best practices for data preprocessing to prepare the dataset for accurate analysis and visualization.

## Master Table Structure

The master table, `master_dataset`, was created to merge all 12 datasets. Below is its structure:

```
-- Master table structure
COLUMN_NAME             DATA_TYPE       CHARACTER_MAXIMUM_LENGTH
IS_NULLABLE
ride_id                 varchar         75                          NO
rideable_type           varchar         255                         YES
started_at              smalldatetime   NULL                        YES
ended_at                smalldatetime   NULL                        YES
start_station_name      varchar         255                         YES
start_station_id        varchar         40                          YES
end_station_name        varchar         255                         YES
end_station_id          varchar         100                         YES
start_lat               decimal         NULL                        YES
start_lng               decimal         NULL                        YES
end_lat                 decimal         NULL                        YES
end_lng                 decimal         NULL                        YES
member_casual           varchar         20                          NO
```

## Data Preview

To inspect the data, the first 1000 rows of the master table were selected.

```
SELECT TOP (1000) [ride_id]
      ,[rideable_type]
      ,[started_at]
      ,[ended_at]
      ,[start_station_name]
      ,[start_station_id]
      ,[end_station_name]
      ,[end_station_id]
      ,[start_lat]
      ,[start_lng]
      ,[end_lat]
```

```
      ,[end_lng]
      ,[member_casual]
  FROM [Cyclistic_Case_Study].[dbo].[master_dataset];
```

# Data Import and Consolidation

Each dataset was inserted into the `master_dataset` table, ensuring no duplicate `ride_id` entries.

```
INSERT INTO master_dataset
SELECT *
FROM dataset_1
WHERE ride_id NOT IN (
    SELECT ride_id FROM master_dataset
);
```

## Handling Import Issues for October 2024 Dataset

The `202410-divvy-tripdata` dataset had data entry errors, preventing standard import. A `BULK INSERT` was used, and affected columns were temporarily altered to `VARCHAR` to accommodate the data.

```
BULK INSERT dataset_4
FROM 'C:\Users\mmnja\OneDrive\Documents\My Docs\DATA ANALYTICS\Google
Sheet CSV Files\202410-divvy-tripdata.csv'
WITH (
    FIELDTERMINATOR = ',',
    ROWTERMINATOR = '\n',
    FIRSTROW = 2,
    TABLOCK
);

ALTER TABLE dataset_4
ALTER COLUMN started_at VARCHAR(255) null;
ALTER TABLE dataset_4
ALTER COLUMN ended_at VARCHAR(255);
ALTER TABLE dataset_4
ALTER COLUMN end_lng VARCHAR(255) null;
```

## Cleaning `dataset_4`

Quotes were removed from specific columns, and the columns were converted back to their original data types before inserting into the master table.

```
UPDATE dataset_4
SET ride_id = REPLACE(TRIM(ride_id), '"', '');
GO
UPDATE dataset_4
```

```
SET rideable_type = REPLACE(TRIM(rideable_type), '"', '');
GO
UPDATE dataset_4
SET started_at = REPLACE(TRIM(started_at), '"', '');
GO
UPDATE dataset_4
SET ended_at = REPLACE(TRIM(ended_at), '"', '');
GO
UPDATE dataset_4
SET member_casual = REPLACE(TRIM(member_casual), '"', '');

ALTER TABLE dataset_4
ALTER COLUMN started_at smalldatetime;
ALTER TABLE dataset_4
ALTER COLUMN ended_at smalldatetime;
ALTER TABLE dataset_4
ALTER COLUMN end_lng DECIMAL(16,13);

INSERT INTO master_dataset
SELECT *
FROM dataset_4
WHERE ride_id NOT IN (
    SELECT ride_id FROM master_dataset
);
```

# Data Cleaning in Master Dataset

With all datasets consolidated, cleaning was performed on the `master_dataset`.

## Removing Duplicates

Duplicate `ride_id` entries were removed using a CTE.

```
WITH RankedDuplicates AS (
    SELECT *,
           ROW_NUMBER() OVER (PARTITION BY ride_id ORDER BY (SELECT
NULL)) AS rn
    FROM master_dataset
)
DELETE FROM RankedDuplicates
WHERE rn > 1;
```

## Setting Primary Key

The `ride_id` column was set as the primary key.

```
ALTER TABLE master_dataset
ADD CONSTRAINT PK_master_dataset_ride_id PRIMARY KEY (ride_id);
```

## Adding Indexes

Indexes were created to optimize query performance for filtering and grouping.

```
CREATE INDEX idx_started_at ON master_dataset(started_at);
CREATE INDEX idx_ended_at ON master_dataset(ended_at);
GO
CREATE INDEX idx_start_station_id ON master_dataset(start_station_id);
CREATE INDEX idx_end_station_id ON master_dataset(end_station_id);
GO
CREATE INDEX idx_member_casual ON master_dataset(member_casual);
```

## Removing NULL Values

Rows with NULL values in critical columns were deleted.

```
DELETE FROM master_dataset
WHERE started_at IS NULL
   OR ended_at IS NULL
   OR start_station_name IS NULL
   OR start_station_id IS NULL
   OR end_station_name IS NULL
   OR end_station_id IS NULL;
```

## Validating Date Ranges

The earliest and latest dates in `started_at` and `ended_at` were checked for anomalies.

```
SELECT MIN(started_at) AS EarliestDate, MAX(started_at) AS LatestDate
FROM master_dataset;
SELECT MIN(ended_at) AS EarliestDate, MAX(ended_at) AS LatestDate FROM
master_dataset;
SELECT * FROM master_dataset WHERE started_at='1900-01-01';
SELECT * FROM master_dataset WHERE ended_at ='1900-01-01';
SELECT * FROM master_dataset WHERE started_at > GETDATE();
```

## Handling Station ID Formats

Inconsistent `start_station_id` and `end_station_id` formats were identified and corrected.

```
SELECT
  CASE
    WHEN start_station_id LIKE '[A-Z]%' THEN 'Starts with Letter'
    WHEN start_station_id LIKE '[0-9]%' THEN 'Starts with Number'
    ELSE 'Other/Irregular'
  END AS FormatType,
  COUNT(*) AS Count
```

```
FROM master_dataset
GROUP BY
  CASE
    WHEN start_station_id LIKE '[A-Z]%' THEN 'Starts with Letter'
    WHEN start_station_id LIKE '[0-9]%' THEN 'Starts with Number'
    ELSE 'Other/Irregular'
  END;

ALTER TABLE master_dataset
ADD start_station_id_format_issue varchar (40);

UPDATE master_dataset
SET start_station_id_format_issue =
  CASE
    WHEN start_station_id LIKE '[A-Z][A-Z][0-9]%'
         AND start_station_id NOT LIKE '%[^A-Z0-9]%'
         AND LEN(start_station_id) >= 8
      THEN 'OK'
    WHEN start_station_id LIKE '[0-9]%' THEN 'Number'
    ELSE 'Other'
  END;

ALTER TABLE master_dataset
ADD end_station_id_format_issue varchar (40);

UPDATE master_dataset
SET end_station_id_format_issue =
  CASE
    WHEN end_station_id LIKE '[A-Z][A-Z][0-9]%'
         AND end_station_id NOT LIKE '%[^A-Z0-9]%'
         AND LEN(end_station_id) >= 8
      THEN 'OK'
    WHEN end_station_id LIKE '[0-9]%' THEN 'Number'
    ELSE 'Other'
  END;

DELETE FROM master_dataset WHERE start_station_id_format_issue != 'OK'
OR end_station_id_format_issue != 'OK';

ALTER TABLE master_dataset
DROP COLUMN start_station_id_format_issue,
end_station_id_format_issue;
```

## Validating Coordinates

Coordinates with incorrect formats were checked and removed.

```
SELECT *
FROM master_dataset
```

```
WHERE
    LEN(CAST(FLOOR(ABS(start_lat)) AS VARCHAR)) > 2
    OR LEN(CAST(FLOOR(ABS(end_lat)) AS VARCHAR)) > 2
    OR LEN(CAST(FLOOR(ABS(start_lng)) AS VARCHAR)) > 2
    OR LEN(CAST(FLOOR(ABS(end_lng)) AS VARCHAR)) > 2;

DELETE FROM master_dataset
WHERE
    FORMAT(start_lat, '0.0000000000000') LIKE '%0000000'
    OR FORMAT(end_lat,   '0.0000000000000') LIKE '%0000000'
    OR FORMAT(start_lng, '0.0000000000000') LIKE '%0000000'
    OR FORMAT(end_lng,   '0.0000000000000') LIKE '%0000000';
```

## Handling Typographical Errors

Columns were checked for typos and extra spaces, with trimming applied where necessary.

```
SELECT DISTINCT member_casual,
       LEN(member_casual) AS ActualLength,
       DATALENGTH(member_casual) AS ByteLength,
       '[' + member_casual + ']' AS ValuePreview
FROM master_dataset
ORDER BY member_casual;

UPDATE master_dataset
SET member_casual = LTRIM(RTRIM(member_casual));

UPDATE master_dataset
SET rideable_type = LTRIM(RTRIM(rideable_type));
GO
UPDATE Cyclistic_Case_Study.dbo.master_dataset
SET start_station_name = LTRIM(RTRIM(start_station_name));
GO
UPDATE Cyclistic_Case_Study.dbo.master_dataset
SET end_station_name = LTRIM(RTRIM(end_station_name));
GO
UPDATE Cyclistic_Case_Study.dbo.master_dataset
SET start_station_id = LTRIM(RTRIM(start_station_id));
GO
UPDATE Cyclistic_Case_Study.dbo.master_dataset
SET end_station_id = LTRIM(RTRIM(end_station_id));
```

## Correcting Station Name Typos

Specific typos in station names were corrected.

```
SELECT DISTINCT start_station_name
FROM Cyclistic_Case_Study.dbo.master_dataset
ORDER BY start_station_name;
```

```sql
SELECT DISTINCT end_station_name
FROM Cyclistic_Case_Study.dbo.master_dataset
ORDER BY end_station_name;

UPDATE Cyclistic_Case_Study.dbo.master_dataset
SET end_station_name = 'Damen Ave & Thomas St' WHERE
end_station_name='Damen Ave & Thomas St (Augusta Blvd)';
UPDATE Cyclistic_Case_Study.dbo.master_dataset
SET end_station_name = 'Damen Ave & Walnut St' WHERE
end_station_name='Damen Ave & Walnut (Lake) St';
```

## Trip Duration Validation

Trips with invalid durations (zero, negative, or excessively long) were removed.

```sql
SELECT *
FROM Cyclistic_Case_Study.dbo.master_dataset
WHERE DATEDIFF(MINUTE, started_at, ended_at) <= 0;

DELETE FROM Cyclistic_Case_Study.dbo.master_dataset
WHERE DATEDIFF(MINUTE, started_at, ended_at) <=0;

SELECT *
FROM Cyclistic_Case_Study.dbo.master_dataset
WHERE DATEDIFF(HOUR, started_at, ended_at) >= 24;

DELETE FROM Cyclistic_Case_Study.dbo.master_dataset
WHERE DATEDIFF(HOUR, started_at, ended_at) >= 24;

SELECT *
FROM Cyclistic_Case_Study.dbo.master_dataset
WHERE ended_at <= started_at;
```

## Removing Short Loop Trips

Trips where the start and end stations were the same with durations under 5 minutes were removed, as they likely represent test rides or errors.

```sql
SELECT *
FROM Cyclistic_Case_Study.dbo.master_dataset
WHERE start_station_name = end_station_name
  AND DATEDIFF(MINUTE, started_at, ended_at) < 5
  AND start_station_name IS NOT NULL;

DELETE FROM Cyclistic_Case_Study.dbo.master_dataset
WHERE start_station_name = end_station_name
  AND DATEDIFF(MINUTE, started_at, ended_at) < 5;
```

# Summary of Data Cleaning Procedures

The following table summarizes the data cleaning steps performed on the Cyclistic dataset.

| Step | Description | Outcome |
|---|---|---|
| Data Import | Imported 12 CSV files into separate SQL tables, consolidated into `master_dataset`. | 12 tables merged, `dataset_4` imported via `BULK INSERT` due to errors. |
| Remove Duplicates | Used CTE to delete duplicate `ride_id` entries. | Ensured unique `ride_id` values. |
| Set Primary Key | Set `ride_id` as primary key. | Enforced data integrity. |
| Add Indexes | Created indexes on `started_at`, `ended_at`, `start_station_id`, `end_station_id`, and `member_casual`. | Improved query performance. |
| Remove NULLs | Deleted rows with NULL values in critical columns. | Ensured complete data for analysis. |
| Validate Dates | Checked `started_at` and `ended_at` for anomalies (e.g., future dates, 1900-01-01). | No invalid dates found. |
| Fix Station IDs | Flagged and removed invalid `start_station_id` and `end_station_id` formats. | Ensured consistent station ID formats. |
| Validate Coordinates | Removed rows with invalid coordinates (e.g., excessive zeros). | Ensured valid geographic data. |
| Trim Spaces | Trimmed extra spaces from `member_casual`, `rideable_type`, `start_station_name`, `end_station_name`, `start_station_id`, `end_station_id`. | Standardized categorical data. |
| Fix Typos | Corrected station name typos (e.g., removed extra descriptors). | Improved station name consistency. |
| Validate Trip Duration | Removed trips with zero/negative durations (<3040 rows) and durations >24 hours (9 rows). | Eliminated invalid trips. |
| Remove Short Loops | Deleted trips with same start/end station and duration <5 minutes (>9000 rows). | Removed likely test rides or errors. |

# Chapter 4 - Exploratory Data Analysis (EDA) & Visualizations

Introduction

With a clean and well-structured dataset, the analysis progressed to the exploratory phase to uncover patterns, trends, and differences in usage behavior between casual riders and annual members. Through a combination of SQL queries and visualizations created in R, key insights were generated on ride duration, day-of-week preferences, peak usage hours, and station activity. These findings laid the groundwork for data-driven recommendations aimed at helping Cyclistic convert more casual users into loyal annual members.

The insights were generated by analyzing seven aspects of the dataset as follows:

- Total Rides by User Type
- Ride Duration by User type
- Ride Volume by Day of Week
- Ride Volume by Hour of Day
- Ride Volume by Rideable Type
- Top Start Stations
- Top End Stations

## Visualizations - (Created using Rstudio)

### 1. Total Rides by User Type

**Code for creating the Visual**

```
user_type_plot <- ggplot(eda_rides_by_user_type, aes(x = member_casual, y =
total_rides, fill = member_casual)) +
  geom_bar(stat = "identity", width = 0.6) +
  labs(
    title = "Total Rides by User Type",
    x = "User Type",
    y = "Total Rides"
  ) +
  theme_minimal() +
  theme(legend.position = "none")  # Hide legend since it's already labeled
```

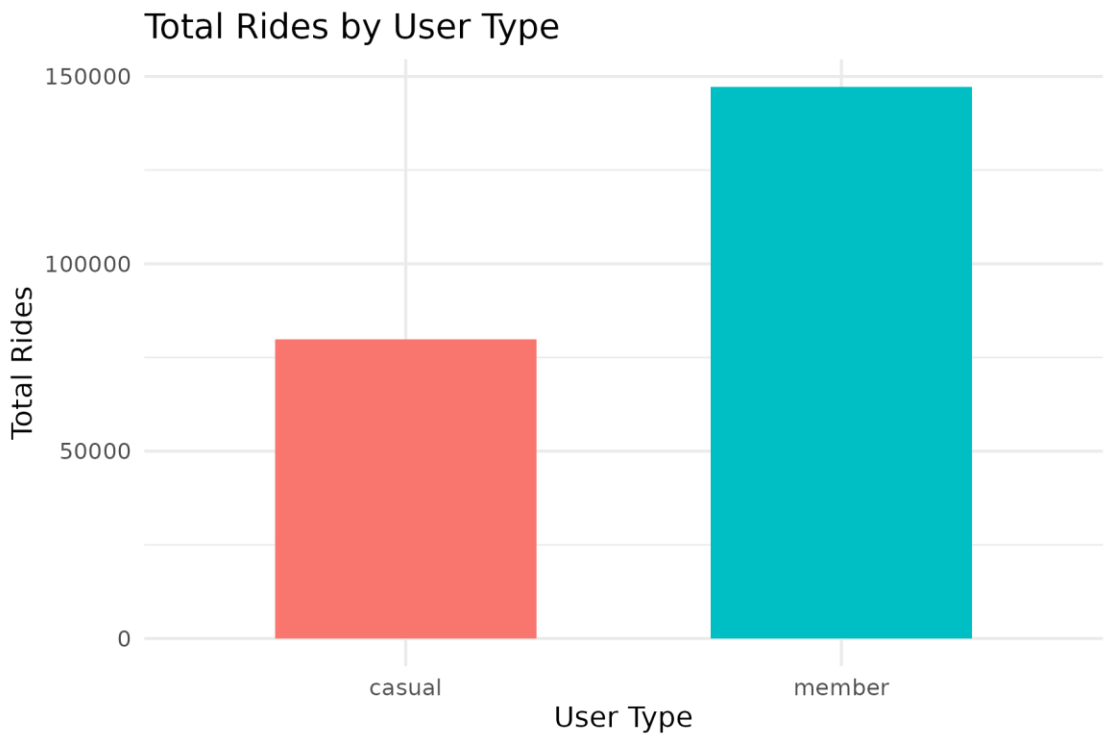



*Figure 1: Total rides comparison between user types*

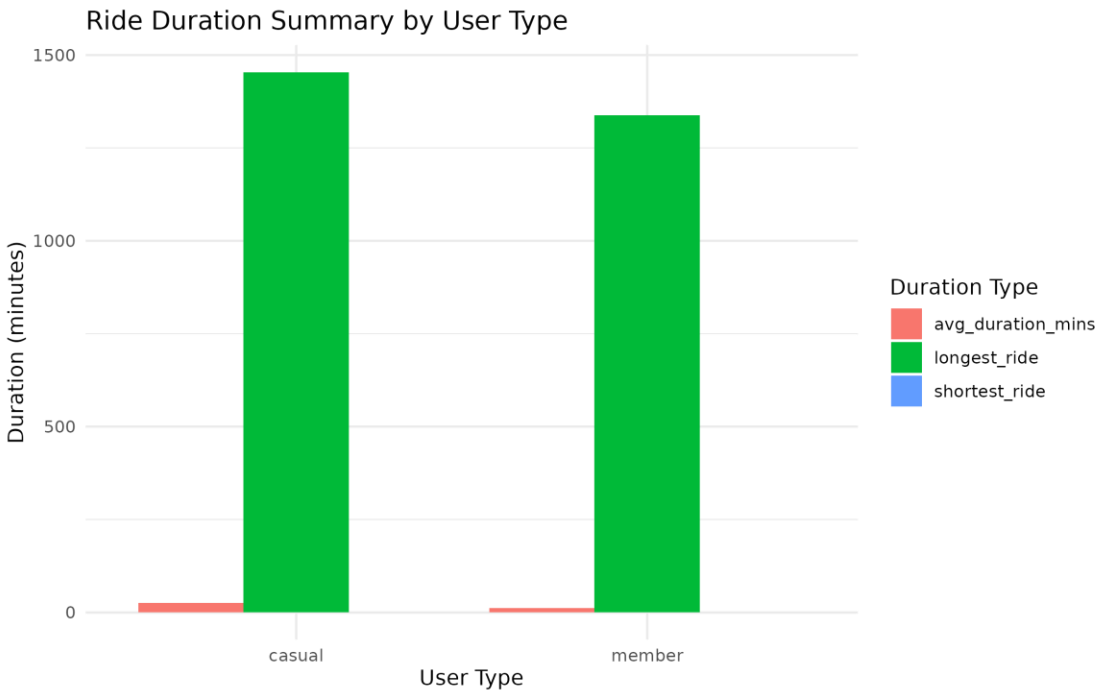
**Key Conclusions:**

Members take significantly more rides than casual users- almost double. This may suggest daily/commuter usage by members vs occasional use by casual riders.

## 2. Ride Duration by User Type

**R Code for the visualization**

```
duration_plot <- ggplot(eda_duration_long, aes(x = user_type, y =
duration_minutes, fill = duration_type)) +
  geom_bar(stat = "identity", position = position_dodge()) +
  labs(
    title = "Ride Duration Summary by User Type",
    x = "User Type",
    y = "Duration (minutes)",
    fill = "Duration Type"
  ) +
  theme_minimal()
```




**Key Conclusions:**

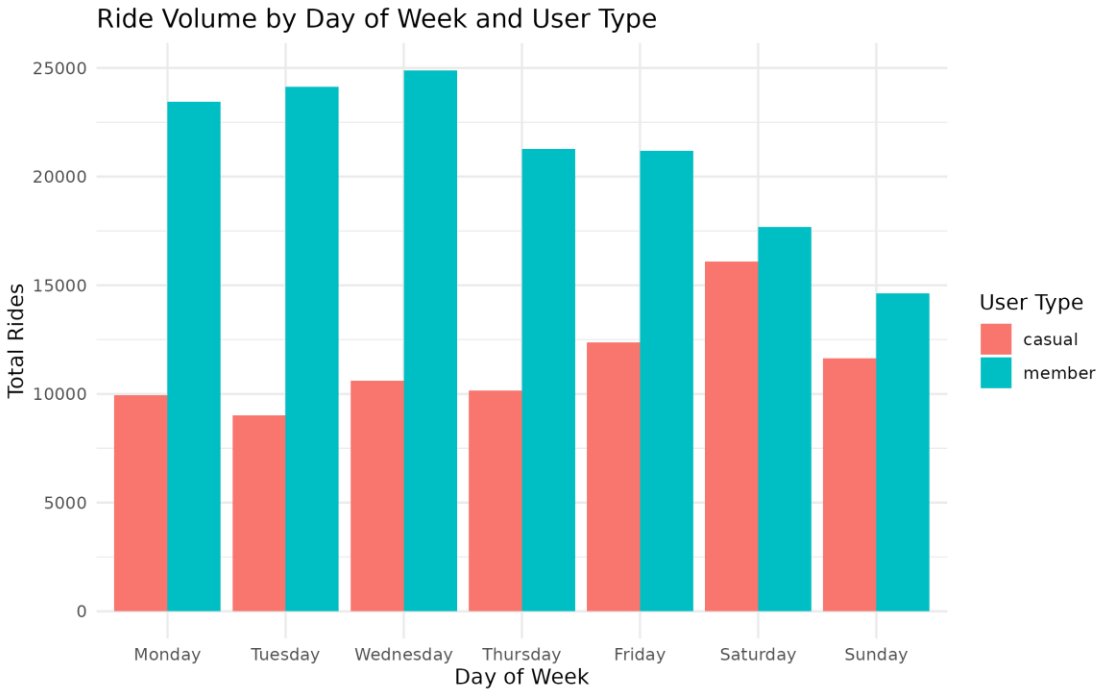
- Casual riders take longer trips on average (25 minutes) compared to members (12 minutes), indicating they may use bikes more for leisure or exploration rather than commuting.
- Both groups have a minimum trip duration of 1 minute, possibly representing very short rides, test runs, or mistaken rentals.
- Maximum trip durations are higher for casuals (1,454 minutes) than for members (1,338 minutes), further reinforcing the idea that casuals are less time-sensitive and possibly less cost-conscious.
- This significant difference in average trip duration is a clear behavioral distinction that Cyclistic could leverage to design personalized user experiences or targeted promotions.

## 3. Ride Volume by Day of Week

**Code for creating the visual**

```
rides_by_day_plot <- ggplot(eda_rides_by_day_long, aes(x = day_of_week, y =
ride_count, fill = user_type)) +
  geom_bar(stat = "identity", position = position_dodge()) +
  labs(
    title = "Ride Volume by Day of Week and User Type",
    x = "Day of Week",
    y = "Total Rides",
    fill = "User Type"
  ) +
```

```
  scale_x_discrete(limits = c("Monday", "Tuesday", "Wednesday", "Thursday",
"Friday", "Saturday", "Sunday")) +
  theme_minimal()
```




**Key Conclusions:**

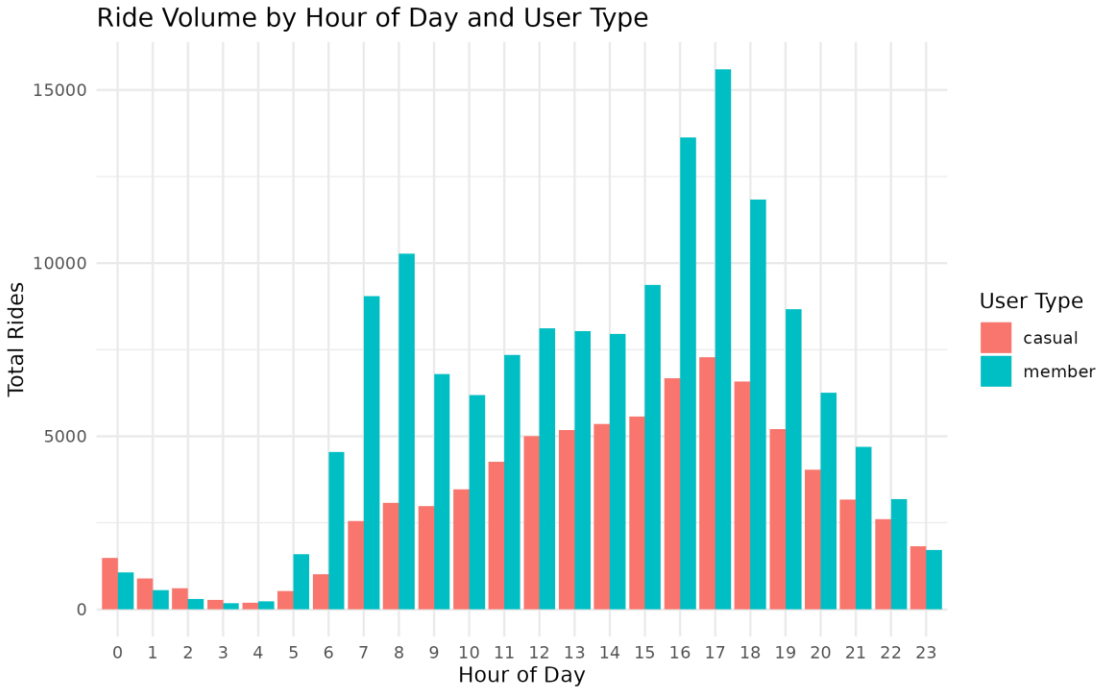
- Members dominate ride volume during weekdays, with the highest usage from Monday to Wednesday, suggesting they use the service primarily for daily commuting or structured routines.
- Casual riders peak on weekends, especially Saturday (16,084), showing a preference for recreational or leisure activities rather than work commutes.
- The stark difference between user types during weekdays (e.g., Tuesday: 9,014 casual vs. 24,121 member) indicates that members ride more consistently across the week, while casuals are more weekend-driven.
- This day-based usage pattern is crucial for staffing, maintenance, and marketing strategies. For example, Cyclistic might push weekend promotions for casuals or loyalty incentives for weekday commuters.

## 4. Ride Volume by Hour of Day

**Code for creating the visual**

```
rides_by_hour_plot <- ggplot(eda_rides_by_hour, aes(x = hour_of_day, y =
ride_count, color = member_casual)) +
  geom_line(size = 1) +
  geom_point(size = 2) +
  labs(
```

```
    title = "Ride Volume by Hour of Day and User Type",
    x = "Hour of Day (0-23)",
    y = "Total Rides",
    color = "User Type"
  ) +
  scale_x_continuous(breaks = 0:23) +
  theme_minimal()
```



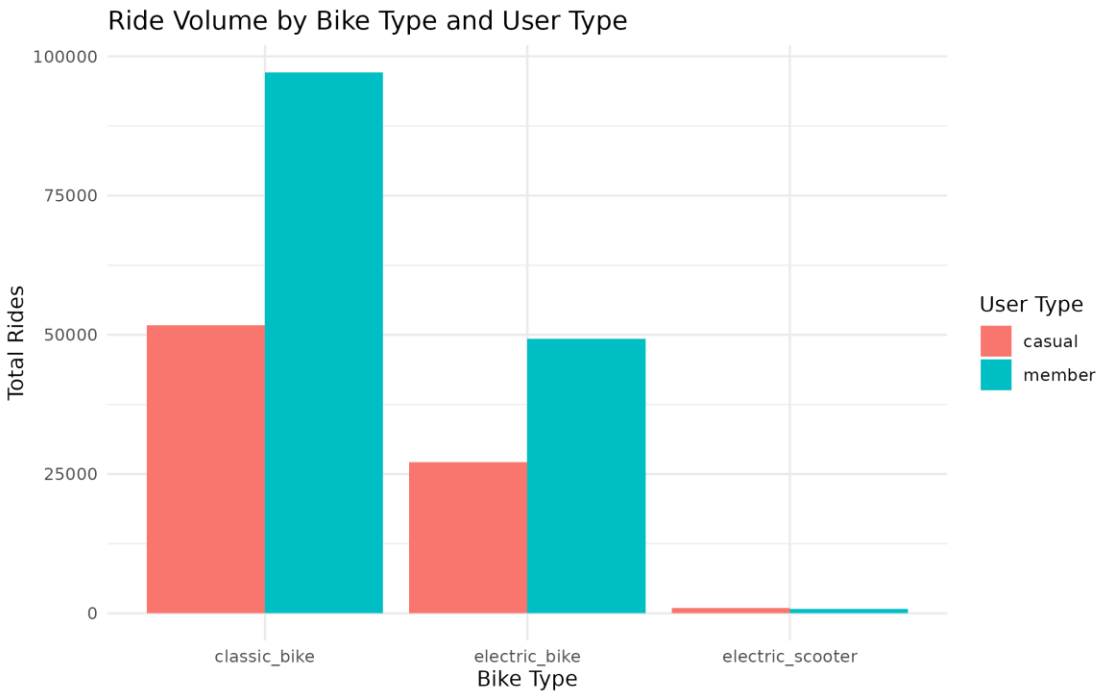

**Key Conclusions:**

- Both member and casual riders peak in the afternoon to early evening hours (2PM–6PM), with 5PM being the top hour for both groups, likely corresponding to commute or leisure hours.
- Members ride significantly more than casual users during every hour, especially during peak workday transitions (e.g., 5PM: 15,603 vs. 7,281 rides).
- Casual riders show lower ride counts during early morning hours (e.g., 4AM: 192), whereas members show a more consistent distribution across the day, even logging over 1,000 rides at 7AM, likely reflecting commuting patterns.
- The data shows clear differences in user behavior: members appear to follow a daily work routine, while casuals concentrate on late afternoon activity, hinting at more recreational use.

## 5. Ride Volume by Rideable Type

**Code for creating the visual**

```
rides_by_type_plot <- ggplot(eda_rides_by_rideable_type, aes(x =
rideable_type, y = total_rides, fill = member_casual)) +
  geom_bar(stat = "identity", position = position_dodge()) +
  labs(
    title = "Ride Volume by Bike Type and User Type",
    x = "Bike Type",
    y = "Total Rides",
    fill = "User Type"
  ) +
  theme_minimal()
```




**Key Conclusions:**

- Cyclistic has a significant number of rides, with “Classic bike” being the most popular rideable type across both user types. Casual vs. Member Users:
- Members are the primary riders of Classic Bikes: Member riders take almost double the number of Classic Bike rides (97,145) compared to Casual riders (51,711). This suggests Classic Bikes are a core offering for dedicated members.
- Electric bikes are popular with both groups, but more so with Members: While Casual users take a good number of Electric bike rides (27,105), Members take even more (49,297). This indicates that electric bikes appeal to a broad user base, but members utilize them more frequently.

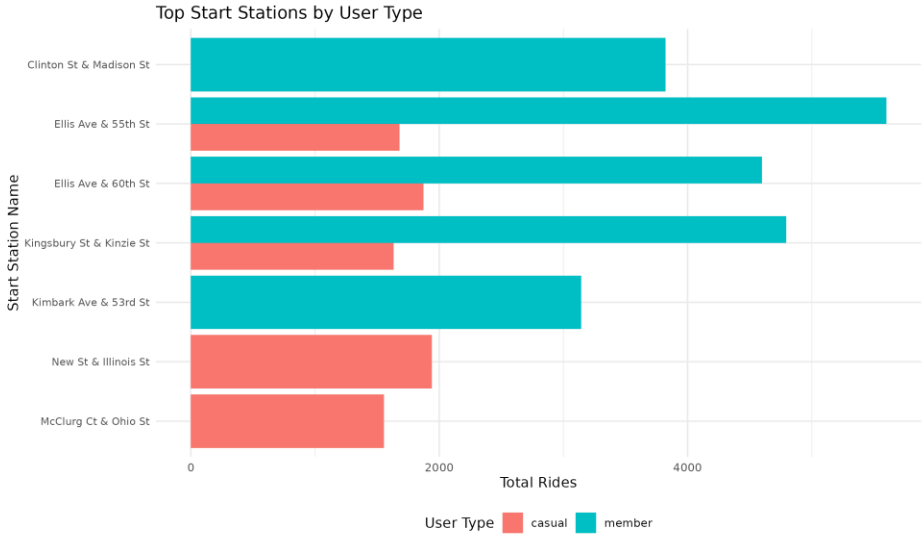
- Electric Scooters have low usage overall: Both Casual and Member users have relatively low numbers of rides on Electric scooters (988 for Casual, 778 for Member). This suggests that electric scooters are not a primary mode of transport for either user group on Cyclistic.
- Casual users prefer electric scooters slightly more than members: While still low, Casual users have a slightly higher number of rides on Electric scooters compared to Members. Key Differences and Insights:
- Members show stronger loyalty to Classic Bikes: The much higher usage of Classic Bikes by members indicates they might be using them for longer rides, commutes, or as their preferred mode for regular trips.
- Casual users might be exploring different options: The higher proportion of Classic Bike and Electric Bike usage for Casual users, relative to Electric Scooters, suggests they might be trying out the more “traditional” bike options.
- Opportunity for Electric Scooter promotion: Given the low usage, Cyclistic might consider promotions or educational campaigns to encourage more users, especially members, to try electric scooters if they want to boost ridership for that rideable type.
- Classic Bikes are a strong asset: The high demand for Classic Bikes, especially among members, highlights their importance to Cyclistic’s business model.

## 6. Top Start Stations

**Code for creating the visual**

```
top_start_stations_plot <- ggplot(eda_top_start_stations, aes(x =
start_count, y = reorder(start_station_name, start_count), fill =
member_casual)) +
  geom_bar(stat = "identity", position = position_dodge()) +
  labs(
    title = "Top Start Stations by User Type",
    x = "Total Rides",
    y = "Start Station Name",
    fill = "User Type"
  ) +
  theme_minimal(base_size = 11) +
  theme(
    axis.text.y = element_text(size = 8),
    plot.margin = margin(10, 10, 10, 150),
    legend.position = "bottom"
  )
```

Here, the top 5 start stations were selected for each user type (member vs casual)




**Key Conclusions:**

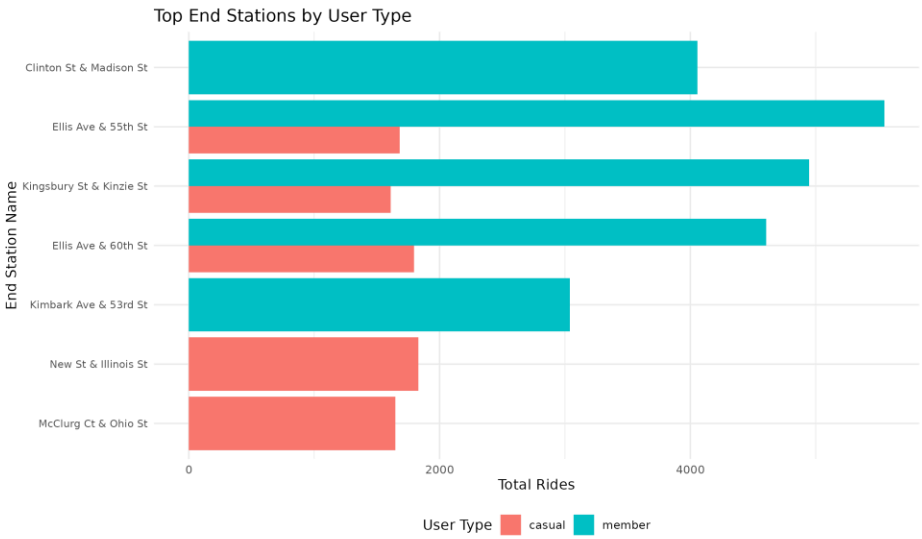
- There's a significant difference in the volume of rides from top start stations between casual and member users. Member top stations consistently show much higher start_count values.
- The station names suggest these are likely located in urban or densely populated areas, possibly near popular landmarks, residential zones, or business districts. Casual User Top Start Stations:
- Lower volume per station: The highest start_count for a casual user station is "New St & Illinois St" with 1942 rides. This is significantly lower than even the lowest member top station.
- Diverse spread of top stations: Casual users seem to have a more distributed usage pattern across various stations, with the start_count values declining relatively gradually among their top 10.
- Potential for tourist/leisure spots: Stations like "Theater on the Lake" and "DuSable Lake Shore Dr & Wellington Ave" suggest that casual users might be using the bikes for recreational purposes, exploring scenic areas. Member User Top Start Stations:
- Extremely high volume per station: The top member station, "Ellis Ave & 55th St," has a staggering 5600 rides, which is almost three times the top casual station. This highlights a very concentrated usage by members at certain key locations.
- Concentrated usage: The top few stations for members ("Ellis Ave & 55th St," "Kingsbury St & Kinzie St," "Ellis Ave & 60th St") show extremely high start_count values, indicating these are major hubs for member activity.

- Likely commuter/regular use locations: Stations with very high member usage are likely located near residential areas, public transport hubs, or major business districts, suggesting daily commuting or regular use.
- Overlap in some stations: Interestingly, “New St & Illinois St” and “McClurg Ct & Ohio St” appear in the top lists for both casual and member users, suggesting these are generally popular and well-located stations. “Ellis Ave & 60th St” is also common. **Key Differences and Insights:** *Purpose of use: Members appear to use the bikes more for regular, high-frequency trips from specific, high-volume hubs (likely for commuting or consistent transportation). Casual users might be using bikes for more varied, possibly recreational or one-off trips from a broader range of stations.
- Strategic Station Importance: The stations highly frequented by members are critical to Cyclistic’s operations and member satisfaction. Ensuring availability and maintenance at these locations is paramount.
- For Member users: Promotions could emphasize the efficiency and reliability of using bikes for daily commutes from their high-volume start stations.
- Station Distribution & Network Design: The data suggests that Cyclistic has successfully placed stations in locations that serve the concentrated needs of its members, and also provides broader access for casual users.
- Potential for “Hot Spots”: “Ellis Ave & 55th St,” “Kingsbury St & Kinzie St,” and “Ellis Ave & 60th St” are clearly “hot spots” for members. Understanding why these specific locations are so popular (e.g., proximity to universities, major employers, transport links) could inform future station placement.

## 7. Top End Stations

Here, the top 5 end stations were selected for each user type (member vs casual) **Code for creating the Visual**

```
top_end_stations_plot <- ggplot(eda_top_end_stations, aes(x = end_count, y =
reorder(end_station_name, end_count), fill = member_casual)) +
  geom_bar(stat = "identity", position = position_dodge()) +
  labs(
    title = "Top End Stations by User Type",
    x = "Total Rides",
    y = "End Station Name",
    fill = "User Type"
  ) +
  theme_minimal()
```




**Key Insights:**

- Similar to start stations, there's a significant disparity in the volume of rides ending at top stations between casual and member users, with member stations showing much higher end_count values.
- There's a strong correlation, and often an exact match, between the top start stations and top end stations for both user types. This indicates that many popular starting points are also popular ending points, suggesting circular routes, commutes, or frequently visited destinations. Casual User Top End Stations:
- Consistent with Start Stations: The top end stations for casual users largely mirror their top start stations (e.g., "New St & Illinois St," "Ellis Ave & 60th St," "McClurg Ct & Ohio St"). This suggests that casual riders often return bikes to stations near where they picked them up, or that these stations are popular destinations.
- Lower Volume: The highest end_count for a casual user station is "New St & Illinois St" with 1831 rides, which is consistent with the lower start counts for casual users.
- Recreational/Tourist Destinations: The presence of stations like "Theater on the Lake" and "DuSable Lake Shore Dr & Wellington Ave" in the top end stations reinforces the idea that casual users utilize bikes for leisure and sightseeing, ending their rides at or near these attractions. Member User Top End Stations:
- Mirroring Start Stations, but with slightly varied order: The top end stations for members are almost identical to their top start stations, with "Ellis Ave & 55th St," "Kingsbury St & Kinzie St," and "Ellis Ave & 60th St" dominating. This strongly suggests high-frequency, regular usage patterns like commuting where members start and end their rides at the same or very similar high-volume hubs.
- Extremely High Volume: The end_count for member stations, particularly the top ones (e.g., "Ellis Ave & 55th St" with 5547 rides), are exceptionally high. This reinforces their role as major arrival points for members.

- Importance of Key Hubs: The continued prominence of stations like “Ellis Ave & 55th St,” “Kingsbury St & Kinzie St,” and “Ellis Ave & 60th St” as both start and end points for members underscores their critical importance as central hubs for Cyclistic’s member base. **Key Differences and Insights (Start vs. End & Casual vs. Member):**
- Symmetry for Members: For members, the top start and end stations are highly symmetrical and involve very high volumes. This strongly supports the hypothesis of routine, possibly commuting, behavior where members frequently travel between or to/from these key locations.
- Symmetry for Casuals (but lower volume): While casual users also show symmetry between their top start and end stations, the overall volume is much lower. This could indicate shorter, more spontaneous trips, or a higher likelihood of returning a bike near the pickup point after a leisure ride.
- Network Efficiency: The data suggests that Cyclistic has strategically placed stations that serve as both popular departure and arrival points, which is crucial for balancing bike distribution within the network, especially for the high-volume member rides.
- Operational Considerations: Stations with consistently high end counts, especially for members, will require frequent rebalancing to ensure bikes are available for subsequent rides. Maintenance and docking capacity at these high-traffic end stations are also critical.
- Understanding Commuter Patterns: The top member start and end stations are likely at the heart of key commuter routes. Further analysis of the specific routes taken between these popular stations would provide deeper insights into member travel patterns.
- Bike Rebalancing: The close resemblance between top start and end stations for both user types, especially members, simplifies bike rebalancing efforts to some extent, as bikes are often returned to where they are frequently picked up. However, the sheer volume at member hubs will still require significant management.

## Summary

- Members primarily use bikes for commuting (weekday mornings and evenings).
- Casual riders tend to ride for leisure (weekend afternoons).
- Each user type shows different preferences in routes and bike types, offering opportunities for tailored service delivery and marketing.

# Chapter 5 - Conclusions

The analysis confirms clear behavioral differences between user groups. These insights support the hypothesis that tailored marketing strategies can increase member conversion rates. With data-backed decision-making, Cyclistic can boost customer loyalty, optimize operations, and maximize profitability.

# Chapter 6 - Recommendations

1. **Optimize Station Placement for Casual Riders:**

**Insight***:* Casual riders frequently start and end rides at leisure destinations like: - New St & Illinois St - Theatre on the Lake - DuSable Habor

- Expand bike dock availability and ensure higher bike supply at these high-demand leisure locations, especially on weekends.
- Introduce seasonal pop-up stations or mobile bike hubs near beaches, parks, or event venues to meet weekend demand.

2. **Offer Flexible Membership Options**

**Insigh***t:* Casual users take longer rides (avg. 25 mins vs. 12 mins for members) and ride more on weekends.

- Introduce a weekend or “Leisure Rider” membership plan tailored for occasional long-distance riders.
- Allow for hour bundles or credits that can roll over, offering a middle ground between casual and full memberships.

3. **Commute Focused Infrastructure**

**Insigh***t:* Members ride most during commute hours (7–9 AM, 4–6 PM) on weekdays.

- Partner with employers and transit authorities to promote bike-to-work incentives (e.g., lockers, showers, or discounts).

# Chapter 7 - Appendix

Project Timeline – July 2025

| DAY | DATES | MILESTONES | RESPONSIBLE TEAM |
|---|---|---|---|
| **1** | July 10 – | → Project kickoff and understanding the business task | Lily Moreno, Marketing & Analytics Team |
| | | → Define problem statement & goals | Marketing & Analytics Team |
| | | → Assign roles & tools (SQL, R, Data Studio, etc.) | Marketing & Analytics Team |
| | | → Gather all datasets (monthly ride data for the year) | Marketing & Analytics Team |
| **2-4** | July 11 – July 13 | → Data cleaning & transformation (handle NULLs, correct formats) | Data Analysts |
| | | → Exploratory Data Analysis (EDA): trends, usage by type, duration, stations | Data Analysts |
| | | → Initial data visualizations | Marketing & Analytics Team |
| **5-7** | July 14 – July 16 | → Advanced analysis (e.g., usage patterns by hour/day/season, clustering) | Data Analysts |
| | | → Create report drafts (SQL, visualizations, executive summary) | Marketing & Analytics Team |
| | | → Prepare insights & recommendation summary | Lily Moreno & Analysts |
| **7-11** | July 17 – July 21 | → Finalize the professional report (merge SQL, PDF, Data Studio outputs) | Marketing & Analytics Team |
| | | → Internal team review & feedback | Lily Moreno, Executive Team |
| | | → Submit report to Cyclistic executive team for review | Lily Moreno |
| **12** | July 22 | → **Executive presentation and decision meeting** | Executive Team |

# Data Analysis - Summary Tables Created from Master Table

As part of the Exploratory Data Analysis (EDA) phase, a series of summary tables were created from the master dataset to uncover patterns, trends, and relationships in the data. These tables include metrics such as average ride duration by user type, number of rides per day of the week, and hourly ride distribution, among others. By aggregating and segmenting the data in meaningful ways, these EDA tables provide a foundational understanding of user behavior and help identify key insights that inform the final analysis and business recommendations.

## Step 1: Trip Volume by User Type

This step helps us understand the overall usage breakdown between casual riders and annual members.

```
SELECT
    member_casual,
    COUNT(*) AS total_rides
FROM master_dataset
GROUP BY member_casual;

-- Create a summary table for Trip Volume Usage by Type.
SELECT
    member_casual,
    COUNT(*) AS total_rides
INTO eda_rides_by_user_type
FROM master_dataset
GROUP BY member_casual;
```

## Step 2: Ride Duration Comparison

This step examines the average and range of ride durations to identify behavioral differences in usage between user types.

```
SELECT
    member_casual,
    COUNT(*) AS total_rides,
    AVG(DATEDIFF(MINUTE, started_at, ended_at)) AS avg_duration_mins,
    MIN(DATEDIFF(MINUTE, started_at, ended_at)) AS shortest_ride,
    MAX(DATEDIFF(MINUTE, started_at, ended_at)) AS longest_ride
INTO eda_duration_by_user_type
FROM master_dataset
GROUP BY member_casual;
```

## Step 3: Rides by Day of the Week (Per User Type)

This step addresses:

- Do casual riders ride more on weekends?
- Do members ride more on weekdays (e.g., for commuting)?

```
SELECT
    member_casual,
    DATENAME(WEEKDAY, started_at) AS day_of_week,
    COUNT(*) AS total_rides
INTO eda_rides_by_day_of_week
FROM master_dataset
GROUP BY member_casual, DATENAME(WEEKDAY, started_at);
```

## Step 4: Rides by Hour of the Day (Per User Type)

This step investigates:

- Do members ride more during rush hours (commutes)?
- Do casual riders ride more during leisure hours (midday or evening)?

```
SELECT
    member_casual,
    DATEPART(HOUR, started_at) AS ride_hour,
    COUNT(*) AS total_rides
INTO eda_rides_by_hour
FROM master_dataset
GROUP BY member_casual, DATEPART(HOUR, started_at);
```

## Step 5: Rideable Type Usage per User Type

This step determines whether casual or member riders prefer a specific type of bike (e.g., electric vs. classic).

```
SELECT
    member_casual,
    rideable_type,
    COUNT(*) AS total_rides
INTO eda_rides_by_rideable_type
FROM master_dataset
GROUP BY member_casual, rideable_type;
```

## Step 6: Most Common Start and End Stations by User Type

This step identifies:

- Which locations are most used by casual vs. member riders.
- Whether casual users tend to start/stop in different places than members.

```sql
-- Create summary table for start_station_name
SELECT
    member_casual,
    start_station_name,
    COUNT(*) AS start_count
INTO eda_top_start_stations
FROM master_dataset
GROUP BY member_casual, start_station_name;

-- Create summary table for end_station_name
SELECT
    member_casual,
    end_station_name,
    COUNT(*) AS end_count
INTO eda_top_end_stations
FROM master_dataset
GROUP BY member_casual, end_station_name;
```

## Additional Queries for Insights

The following queries retrieve results from the summary tables to gain deeper insights into user behavior.

```sql
-- 1. Rides by User Type – Ride totals per user type
SELECT TOP (1000) [member_casual]
      ,[total_rides]
  FROM [Cyclistic_Case_Study].[dbo].[eda_rides_by_user_type]

-- 2. Ride Duration by User Type – Avg, min, max ride durations
SELECT TOP (1000) [member_casual]
      ,[total_rides]
      ,[avg_duration_mins]
      ,[shortest_ride]
      ,[longest_ride]
  FROM [Cyclistic_Case_Study].[dbo].[eda_duration_by_user_type]

-- 3. Rides by Day of the Week – Ride volume by day for each user type
SELECT
    member_casual,
    day_of_week,
    total_rides
FROM eda_rides_by_day_of_week
ORDER BY
    member_casual,
    CASE
        WHEN day_of_week = 'Monday' THEN 1
        WHEN day_of_week = 'Tuesday' THEN 2
        WHEN day_of_week = 'Wednesday' THEN 3
        WHEN day_of_week = 'Thursday' THEN 4
```

```
        WHEN day_of_week = 'Friday' THEN 5
        WHEN day_of_week = 'Saturday' THEN 6
        WHEN day_of_week = 'Sunday' THEN 7
        ELSE 8
    END;

-- 4. Rides by Hour – Ride counts by hour of day per user type
SELECT TOP (48) [member_casual]
      ,[ride_hour]
      ,[total_rides]
  FROM [Cyclistic_Case_Study].[dbo].[eda_rides_by_hour];

-- 5. Rideable Type Usage – Ride type usage by user type
SELECT TOP (6) [member_casual]
      ,[rideable_type]
      ,[total_rides]
  FROM [Cyclistic_Case_Study].[dbo].[eda_rides_by_rideable_type];

-- 6. Top Start Stations (Top 5–10 entries per user type)
WITH RankedStartStations AS (
    SELECT
        member_casual,
        start_station_name,
        start_count,
        RANK() OVER (PARTITION BY member_casual ORDER BY start_count
DESC) AS station_rank
    FROM eda_top_start_stations
)
SELECT
    member_casual,
    start_station_name,
    start_count
FROM RankedStartStations
WHERE station_rank <= 10
ORDER BY member_casual, station_rank;

-- 7. Top End Stations (Top 5–10 entries per user type)
WITH RankedEndStations AS (
    SELECT
        member_casual,
        end_station_name,
        end_count,
        RANK() OVER (PARTITION BY member_casual ORDER BY end_count
DESC) AS station_rank
    FROM eda_top_end_stations
)
SELECT
    member_casual,
    end_station_name,
    end_count
```

```
FROM RankedEndStations
WHERE station_rank <= 10
ORDER BY member_casual, station_rank;
```

## Summary of Data Analysis Procedures

The table below summarizes the data analysis procedures conducted in this EDA process.

| Step | Analysis Procedure | Description | Summary Table Created |
|---|---|---|---|
| 1 | Trip Volume by User Type | Analyzed the total number of rides per user type (casual vs. member). | eda_rides_by_user_type |
| 2 | Ride Duration Comparison | Compared average, minimum, and maximum ride durations between user types. | eda_duration_by_user_type |
| 3 | Rides by Day of the Week | Examined ride volume by day of the week for each user type to identify commuting vs. leisure patterns. | eda_rides_by_day_of_week |
| 4 | Rides by Hour of the Day | Analyzed ride counts by hour to identify peak usage times for each user type. | eda_rides_by_hour |
| 5 | Rideable Type Usage | Investigated preferences for bike types (e.g., electric vs. classic) by user type. | eda_rides_by_rideable_type |
| 6 | Most Common Start and End Stations | Identified the most frequently used start and end stations for each user type. | eda_top_start_stations, eda_top_end_stations |

# Chapter 8 - References

- Tools used (SQL Server, RStudio, Azure Data Studio)