🚲 Uncovering Rider Insights: My Google Data Analytics Capstone Journey

---

🔍 Introduction

As part of the **Google Data Analytics Professional Certificate**, I began working on my capstone project with a real-world business scenario: understanding how users of a bike-share service behave, and how to turn occasional riders into loyal customers.

I chose the **Cyclistic Bike-Share Dataset**, a popular case study based on a real company in Chicago. The project required me to explore rider trends using tools like **Excel**, **SQL (BigQuery)**, and **Tableau**, and apply the entire **data analysis process**: from asking the right questions to preparing and cleaning the data.

🧠 Phase 1: Ask — Defining the Business Task

The fictional company, **Cyclistic**, offers multiple ride options: single rides, day passes, and annual memberships. The marketing team wants to boost **annual memberships**, which are more profitable and predictable.

The **key business question** is:

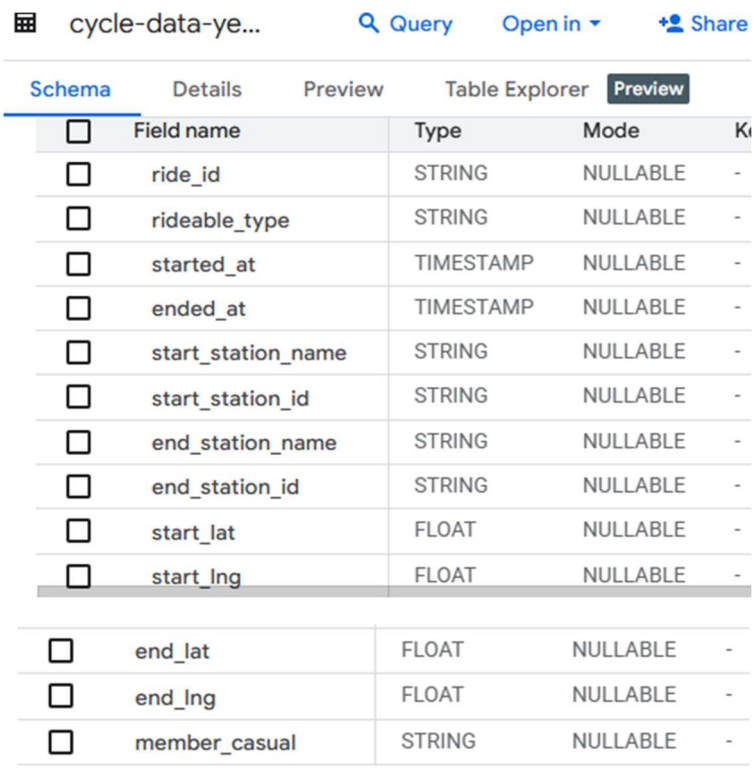
“How do annual members and casual riders use Cyclistic bikes differently?”

As a junior data analyst, my role is to explore user behavior and deliver insights that help shape a targeted marketing strategy to convert **casual riders into annual members**.

📦 Phase 2: Prepare — Collecting and Understanding the Data

I collected **12 months of Cyclistic ride data** from **June 2024 to May 2025**.

Each monthly file was in **CSV format** and contained tens of thousands of records with columns such as:



🔍 Key Observations from Initial Exploration

- **Time Range:** June 2024 – May 2025
- **Customer Info:** The only available customer detail is rider type (member or casual). No gender, age, or location — a major limitation for demographic analysis.
- **File Consistency Issues:**
    - Inconsistent date formats across months
    - Some files had empty or null values in key columns
- **Ride Validity Rule:**
  Cyclistic guidelines mention that **trips shorter than 1 minute** or **longer than 1 day** are likely system-generated (e.g., maintenance or theft) and should be removed.

🧼 Phase 3: Pre-Data Cleaning – Auditing the Raw Dataset

Before diving into cleaning, I conducted a detailed **column-wise exploration** to identify anomalies, nulls, and inconsistencies. This phase was crucial to ensure I only clean what's necessary without losing valuable data.

To do this, I used BigQuery to inspect:

- The **total number of rows**
- The **distinct count** and **null values** for each column
- Columns of interest: ride_id, rideable_type, start_station_name, end_station_name, GPS coordinates, member_casual, etc.

Column	Total Non-Null Rows	Distinct Values	Nulls	Observation
ride_id	5,628,847	5,628,847	0	✅ All unique — no duplicate rides. Good!
member_casual	5,628,847	2	0	✅ Clean — only 2 categories (likely member and casual)
rideable_type	5,628,847	3	0	✅ 3 bike types — normal
started_at	5,628,847	5,627,717	0	⚠️ Only ~1,100 duplicated start times — not alarming
ended_at	5,628,847	5,626,976	0	✅ Similar to start — okay
start_lat/lng	5,628,847	2.9–2.8 lakh+	0	🟧 High precision GPS data — can be removed if not mapping
end_lat/lng	5,622,681	~2,700–2,800	6,166	⚠️ 6k rows have missing end location
start_station_id	4,546,902	1,785	1,081,945	⚠️ ~19% missing station ID
start_station_name	4,546,902	1,869	1,081,945	⚠️ Same as above — match IDs and names if keeping

end_station_id	4,518,107	1,787	1,110,740	⚠️ ~20% missing — decision needed
end_station_name	4,518,107	1,872	1,110,740	⚠️ Same — names + IDs linked

---

📌 Key Observations

1. **High Null Counts in Key Columns**
   - Both start_station_name and end_station_name had **over 1 million nulls**.
   - GPS fields like end_lat and end_lng also had **significant missing values**.
   - ride_id, rideable_type, and member_casual were consistently filled.
2. **Temporal Validity**
   - I confirmed that **all records fell within the June 2024 to May 2025 window**, but I kept this check in place as a data integrity step.

Make sure all rides are between **June 2024 – May 2025**, and no nulls.

```
SELECT
  MIN(started_at) AS earliest_start,
  MAX(started_at) AS latest_start,
  MIN(ended_at) AS earliest_end,
  MAX(ended_at) AS latest_end,
  COUNTIF(started_at IS NULL OR ended_at IS NULL) AS null_timestamps
FROM `sapient-zodiac-465011-t1.cycle_data.cycle-data-yearly`;
```

Query completed

Query results

Save results | Open in

Job information | Results | Chart | JSON | Execution details | Execution graph

Row	earliest_start	latest_start	earliest_end	latest_end	null_timest
1	2024-05-31 01:02:49.582000 UTC	2025-05-31 23:58:56.394000 UTC	2024-06-01 00:00:02.288000 UTC	2025-05-31 23:59:49.905000 UTC	

To measure the data quality across rows, I created a breakdown by how many of the 13 key columns were missing:

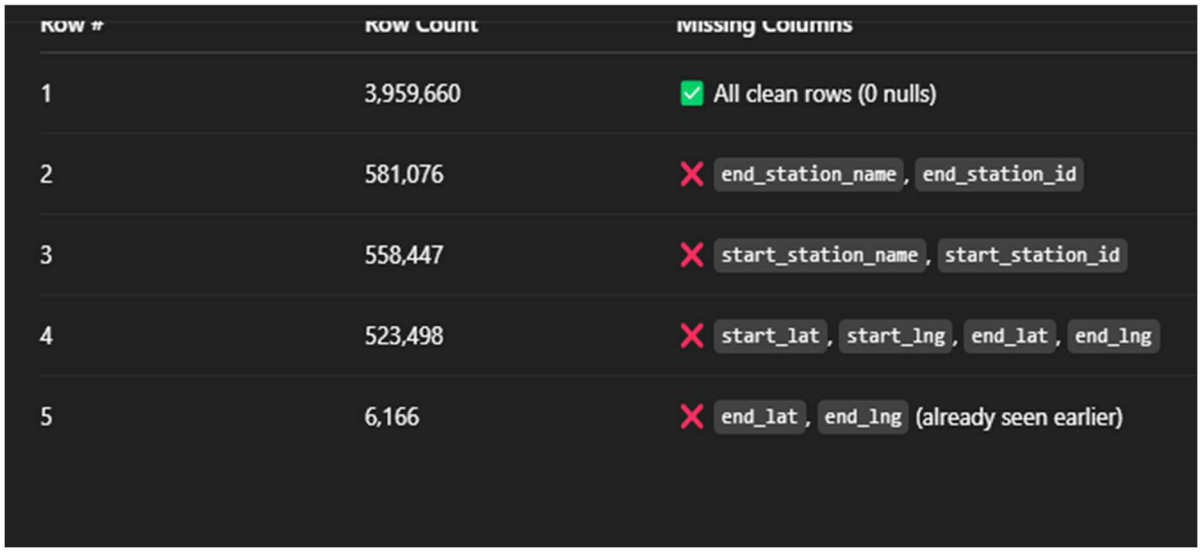
Row	row_count
1	3959660
2	581076
3	558447
4	523498
5	6166

Findings:

- **~3.95M rows** had **no nulls** – great!
- A significant number (hundreds of thousands) had **2 to 4 null values** – these needed careful consideration.

Which Columns Are Missing Together?

By grouping rows with identical missing patterns, I discovered:



Row #	Row Count	Missing Columns
1	3,959,660	✅ All clean rows (0 nulls)
2	581,076	❌ `end_station_name`, `end_station_id`
3	558,447	❌ `start_station_name`, `start_station_id`
4	523,498	❌ `start_lat`, `start_lng`, `end_lat`, `end_lng`
5	6,166	❌ `end_lat`, `end_lng` (already seen earlier)

Understanding the Ride Duration

To ensure reliable analysis, I checked whether ride durations were within realistic limits:

Findings:

- ✅ Most rides were valid, but...
- ❌ 123,248 rides were under 1 minute
- ❌ 6,385 rides lasted over 24 hours
  These were likely system checks or errors and needed to be removed.

```sql
SELECT
  COUNT(*) AS total_rows,
  COUNTIF(TIMESTAMP_DIFF(ended_at, started_at, MINUTE) < 1) AS too_short,
  COUNTIF(TIMESTAMP_DIFF(ended_at, started_at, MINUTE) > 1440) AS too_long,
  MIN(TIMESTAMP_DIFF(ended_at, started_at, MINUTE)) AS min_duration,
  MAX(TIMESTAMP_DIFF(ended_at, started_at, MINUTE)) AS max_duration
FROM `sapient-zodiac-465011-t1.cycle_data.cycle-data-yearly`;

```

Query completed

Query results

Save

Job information | Results | Chart | JSON | Execution details | Execution graph

Row	total_rows	too_short	too_long	min_duration	max_duration
1	5628847	123248	6385	-56	1559

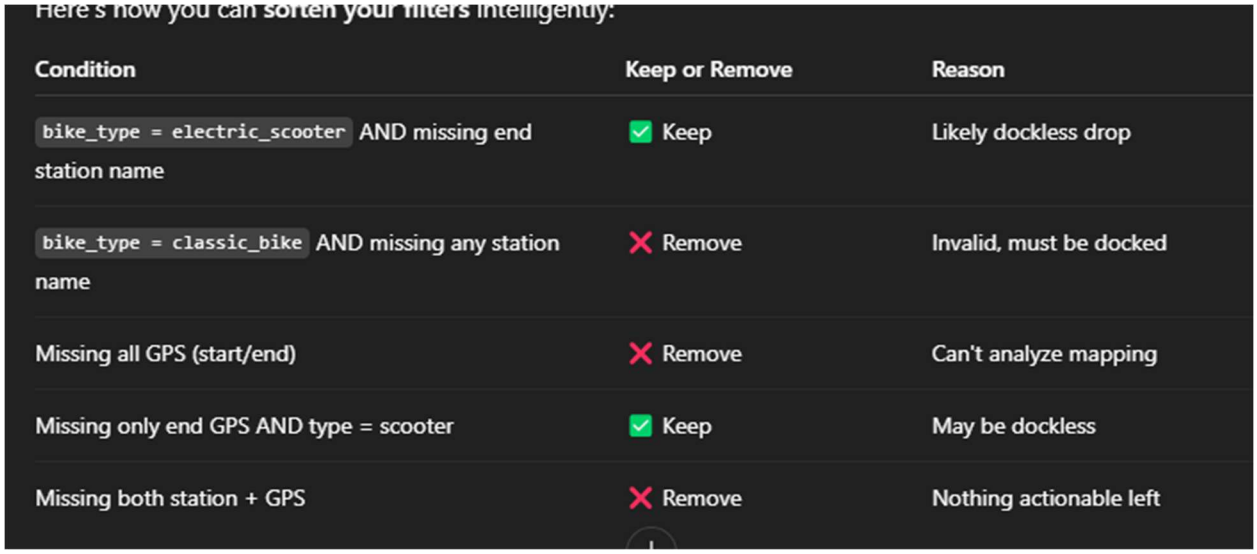
✅ Final Cleaning Strategy

Based on the analysis, I decided to:

- ✅ **Remove** rides <1 min or >1 day
- ✅ **Drop** rows missing both start or end station names and IDs
- ✅ **Remove** rows with **all 4 GPS values missing**
- ✅ Keep rows even if **only end lat/lng is missing** — station names are still useful
- ✅ Ensure ride_id is valid (non-null and 16 characters)

This way, I preserved **data quality without over-cleaning**, ensuring reliable visualizations and insights in Tableau.

Interestingly, when exploring patterns of missing station names and GPS data, I considered the **bike types** involved. I learned that **electric scooters are often dockless**, meaning they don't require users to leave them at designated stations. This explains why rows involving scooters sometimes have missing end_station_name or end_lat/lng. However, for classic_bike, which must be docked, missing station data is invalid — and such rows were marked for removal. This type-aware filtering gave me a smarter and fairer cleaning process.



During cleaning, I noticed a small group of rides with missing end_lat and end_lng. These rides still had valid end_station_name or belonged to dockless scooters, which don't always end at fixed stations. Instead of removing these rows blindly, I chose to retain any row that had **either an end station name or GPS coordinates**, ensuring I didn't lose valuable behavior patterns like one-way or street-side parking.

📘 Phase 3: Data Cleaning and Preparation

After exploring and assessing the 12-month Cyclistic dataset, it became clear that the data needed significant cleaning before it could be analyzed reliably. Inconsistent formats, missing values, and ambiguous ride records could lead to misleading insights. Here's how I approached the cleaning process in BigQuery, with full justification for each rule.

✅ Cleaning Rules Applied

No.	Rule Description	Reason
1	**Exclude trips shorter than 1 minute or longer than 24 hours**	These likely represent test rides, errors, or maintenance rides — not valid customer trips.
2	**Remove rows with null ride_id, rideable_type, started_at, ended_at, or member_casual**	These fields are essential to uniquely identify and categorize a ride. Missing values here make the row unusable.
3	**Only include rides with exactly 16-character ride_id values**	Ensures uniform ID format across months and avoids malformed IDs.

4	**Classic bikes must have both start_station_name and end_station_name present**	Classic bikes are docked and require a fixed start and end station, so both must be known.
5	**Electric bikes must have a start_station_name**	While e-bikes may be parked flexibly, the ride must have a known starting point.
6	**Electric scooters must have at least one valid origin (station name, ID, or coordinates)**	To understand scooter usage, a minimum level of origin info is required. Rows without any are not useful.
7	**Electric scooters must also have at least one valid destination**	Similarly, destination information is necessary to understand ride patterns.
8	**Drop rows where all end information is missing (end_station_name, end_station_id, end_lat, end_lng)**	These rows provide no useful context for where the trip ended.
9	**Drop rows where all start and end GPS coordinates are missing**	Even if station names are missing, coordinates help locate rides spatially. Without them, the data point has no geographic value.
10	**Drop rows where both start_station_name/id and end_station_name/id are missing**	We cannot identify origin or destination of these rides in any way, making them analytically useless.

🧠 Why This Cleaning Strategy?

Each rule was carefully designed to preserve **maximum valuable data** while eliminating **incomplete or invalid records** that would weaken the analysis. Special attention was given to different **bike types**:

- **Classic Bikes**: Must be docked → both station names required
- **Electric Bikes**: May be parked flexibly → only starting point required
- **Electric Scooters**: Often dockless → at least some origin/destination data is required

This bike-type-specific logic ensured we didn't over-clean and lose valid rides, especially for dockless scooter data.

🖌️ Additional Enrichment

To support later analysis, I added useful columns such as:

- ride_duration_minutes
- day_of_week, ride_date, ride_month, ride_year, week_number, ride_hour

- ride_duration_category (Short/Medium/Long)
- is_roundtrip (TRUE if start and end station match)

These derived fields will make it much easier to build visualizations and extract trends in the next phase.

📊 Phase 4 – Analyze (SQL-Based Exploration)

With a cleaned dataset ready, I moved into SQL to extract numeric insights on how members and casual riders behave differently. These initial figures shaped the direction of further analysis and future business strategy.

🚴 1. Total Ride Counts by User Type

One of the most basic but foundational metrics:

```sql
--count no of rides by member type
SELECT
  member_casual,
  COUNT(*) AS ride_count
FROM
  `sapient-zodiac-465011-t1.cycle_data.cycle-data-year`
GROUP BY
  member_casual
ORDER BY
  ride_count DESC;
```

Query completed

Query results

Job information | Results | Chart | JSON | Exec

ow	member_casual	ride_count
1	member	3564561
2	casual	2064286

This tells us that while members make up the majority, **casuals represent a significant 36.6% of total rides**, giving us a large group to potentially convert to subscriptions.

🚲 2. Ride Counts by Bike Type

I then broke down usage by rideable_type:

check-total-rows Run Share Schedule

```sql
-- See all distinct bike types and how many times each occurs
SELECT
  rideable_type,
  COUNT(*) AS count
FROM
  `sapient-zodiac-465011-t1.cycle_data.cycle-data-yearly`
GROUP BY
  rideable_type
ORDER BY
```

Query completed

Query results

Job information | Results | Chart | JSON | Execution detai

Row	rideable_type	count
1	electric_bike	3071374
2	classic_bike	2413136
3	electric_scooter	144337

---

3. Top Start Stations – Member vs Casual

Row	member_casual	start_station_name	ride_count
1	member	Kingsbury St & Kinzie St	30842
2	member	Clinton St & Washington Blvd	25643
3	member	Clinton St & Madison St	23626
4	member	Clark St & Elm St	23502
5	member	Canal St & Madison St	20778
6	member	Clinton St & Jackson Blvd	19791
7	member	Wells St & Elm St	19379
8	member	Wells St & Concord Ln	19356
9	member	State St & Chicago Ave	19294
10	member	Dearborn St & Erie St	17709

Row	member_casual	start_station_name	ride_count
1	casual	Streeter Dr & Grand Ave	46532
2	casual	DuSable Lake Shore Dr & Monro...	31626
3	casual	Michigan Ave & Oak St	23202
4	casual	Millennium Park	21218
5	casual	DuSable Lake Shore Dr & North ...	21161
6	casual	Shedd Aquarium	19433
7	casual	Dusable Harbor	17547
8	casual	Theater on the Lake	15341
9	casual	Michigan Ave & 8th St	12538
10	casual	Adler Planetarium	11321

Members lean toward downtown work/commute hubs, while **casual riders cluster around lakeshore and tourist-friendly zones** — a very clear difference in behavior.

📍 4. Top End Stations – Member vs Casual

I analyzed where trips end as well:

Row	member_casual	end_station_name	ride_count
1	member	Kingsbury St & Kinzie St	28098
2	member	Clinton St & Washington Blvd	23358
3	member	Clinton St & Madison St	22351
4	member	Clark St & Elm St	21028
5	member	Canal St & Madison St	18940
6	member	Clinton St & Jackson Blvd	18149
7	member	State St & Chicago Ave	17889
8	member	Wells St & Concord Ln	17199
9	member	Wells St & Elm St	16889
10	member	University Ave & 57th St	16394

Row	member_casual	end_station_name	ride_count
1	casual	Streeter Dr & Grand Ave	48270
2	casual	DuSable Lake Shore Dr & Monro...	28108
3	casual	DuSable Lake Shore Dr & North ...	23648
4	casual	Michigan Ave & Oak St	22551
5	casual	Millennium Park	21744
6	casual	Shedd Aquarium	16789
7	casual	Theater on the Lake	15646
8	casual	Dusable Harbor	14973
9	casual	Michigan Ave & 8th St	10844
10	casual	Michigan Ave & Washington St	10503

This confirmed the previous trend: casual riders start and end around the **same key scenic/tourist hotspots**, while members stick to **dense office zones**.

🔁 5. Top Round-Trip Stations (Start = End)

Finally, I explored which stations had the most **round-trip rides** (i.e., where riders returned to the same station):

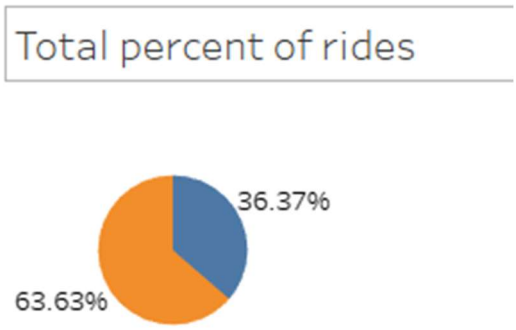
Row	member_casual	start_station_name	round_trip_count
1	casual	Streeter Dr & Grand Ave	7777
2	casual	DuSable Lake Shore Dr & Monro...	6318
3	casual	Michigan Ave & Oak St	4200
4	casual	Dusable Harbor	2915
5	casual	Millennium Park	2905
6	casual	Montrose Harbor	1980
7	casual	DuSable Lake Shore Dr & North ...	1860
8	casual	Michigan Ave & 8th St	1860
9	casual	Shedd Aquarium	1729
10	casual	Adler Planetarium	1680

These are ideal candidates for **on-ground engagement, weekend promotions**, or even rewards programs — they reflect **recreational or leisure behavior**.

Phase 5 – Share (Storytelling Through Tableau)

Once I had clean and structured data, the next step was to translate numbers into stories. Using Tableau, I visualized key usage patterns of Cyclistic riders to uncover actionable insights.

🟠 1. Total Ride Count by User Type – Pie Chart

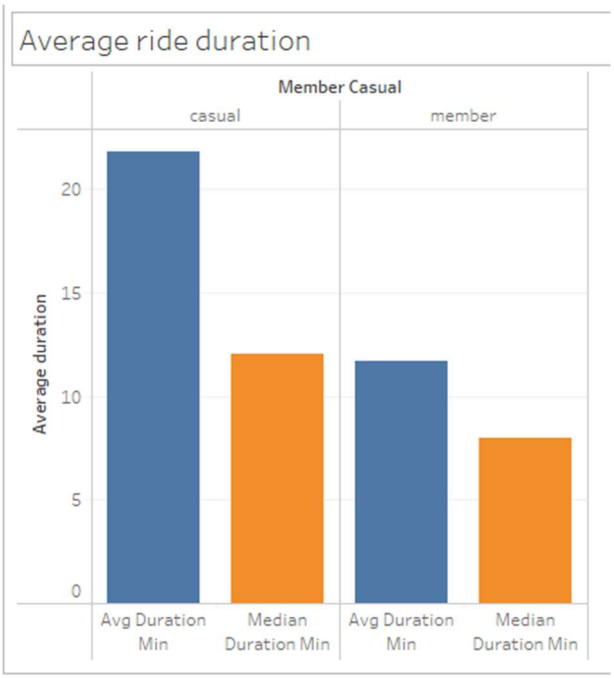


🔍 Insight:

- **63.6% of total rides** were by **members**, indicating a strong loyal user base.
- **Casual riders make up 36.4%**, which is a significant portion, and presents an opportunity for conversion into subscribers.

📊 2. Ride Duration Analysis – Average & Median

This bar chart compares both **average** and **median ride durations** for members and casuals.



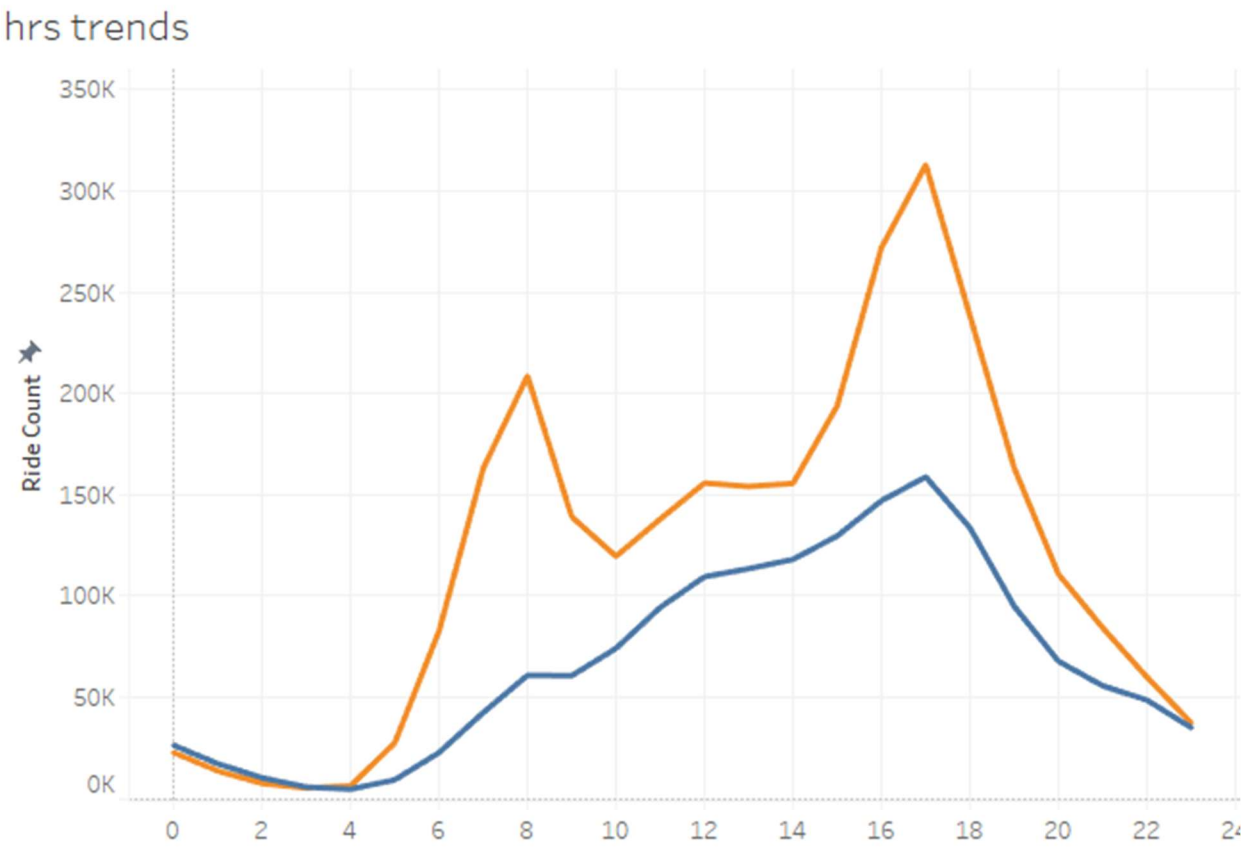
🔍 **Insight:**

- **Casual riders:**
  - **Avg:** 21 mins
  - **Median:** 12 mins
- **Members:**
  - **Avg:** 11.7 mins
  - **Median:** 8 mins

📌 Casuals tend to take **longer and more variable rides**, hinting at leisure usage, while members show **more consistent, shorter ride times** — typical of commuting.

⏰ 3. Hourly Ride Trends

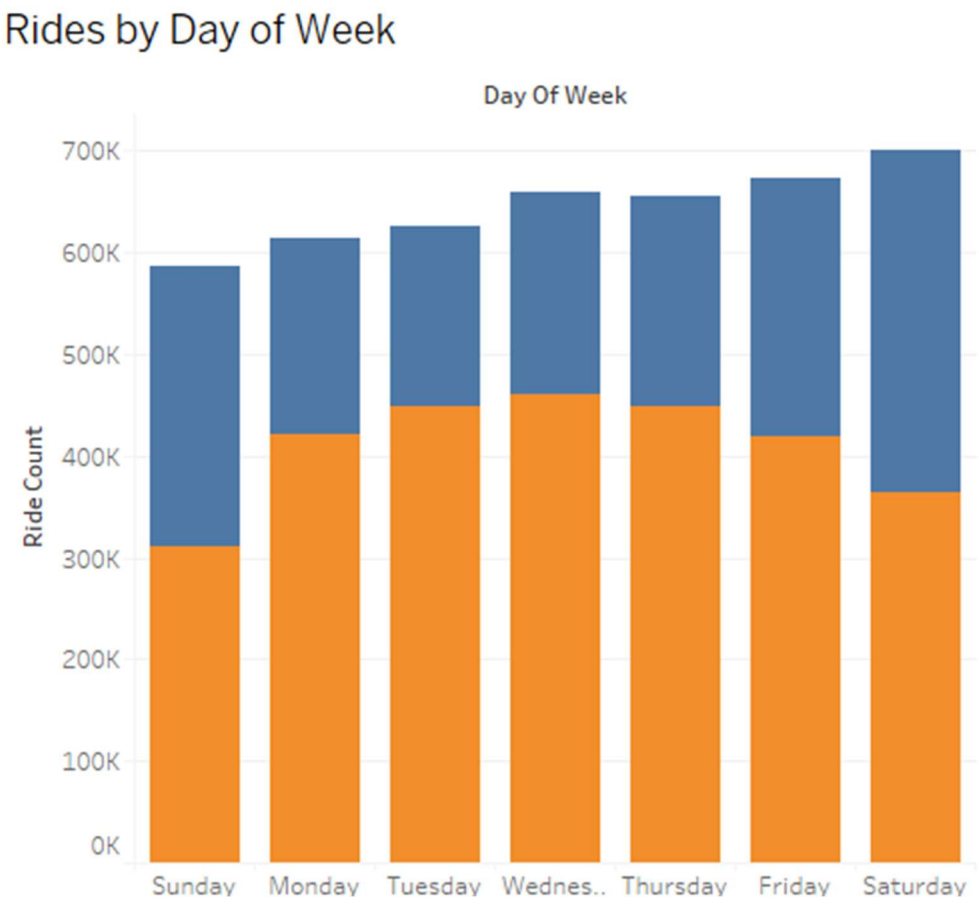
Line chart showing when rides happen throughout the day.



🔍 **Insight:**

- **Members** peak around **8 AM and 4–6 PM**, aligning with **commute hours**.
- **Casual users** peak between **10 AM to 6 PM**, typical of **tourism or leisure rides**.

📅 **4. Day of Week Ride Trends**

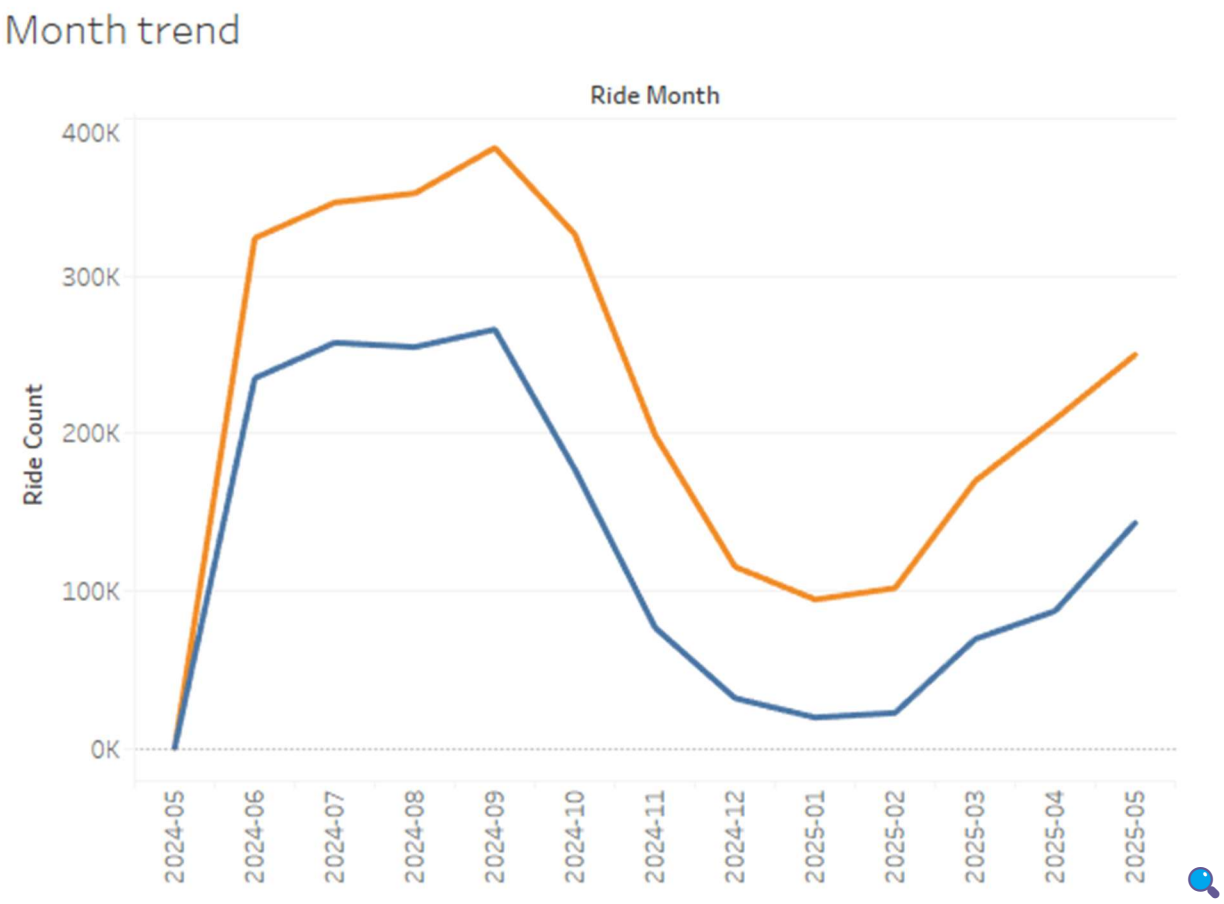


🔍 **Insight:**

- **Members** show steady weekday usage, with slight dips on weekends.
- **Casual users** have **peak usage Friday to Sunday**, reinforcing the idea that their rides are often **weekend-based activities**.

🗓️ 5. Monthly Ride Trends

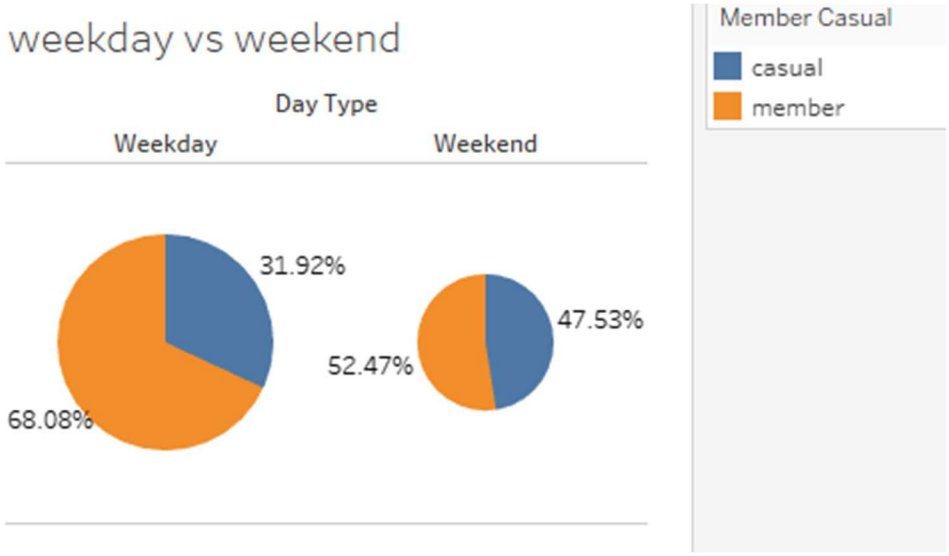
Line chart visualizing monthly ride counts over the 12-month period (June 2024 – May 2025).



Insight:

- Both groups follow a **seasonal pattern**:
    - Ride volume increases from **June to September**, drops in **winter**, and rises again by **spring**.
- This helps time future **promotions and marketing** around seasonal interest spikes.

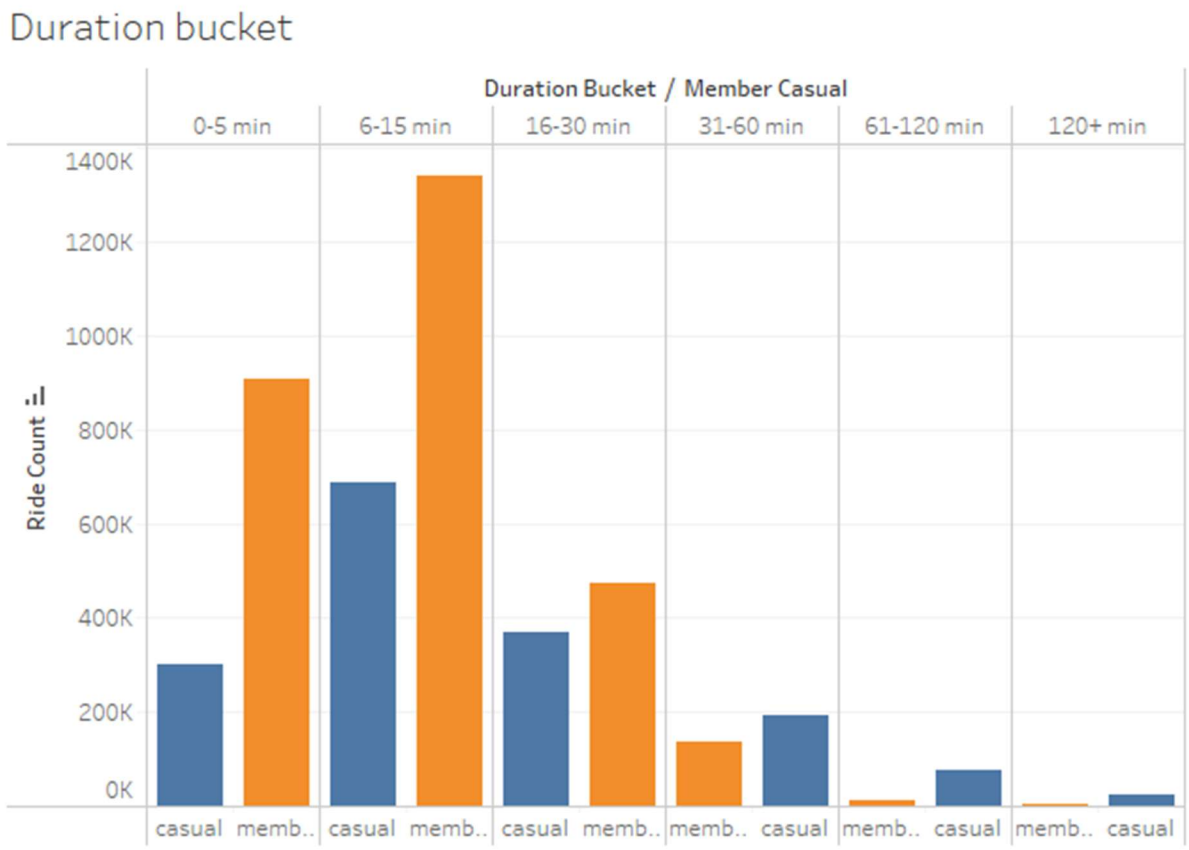
👫 6. Weekday vs Weekend Usage



🔍 Insight:

- On weekdays, **members dominate**.
- On weekends, **casual usage jumps from 31% to 47%**, nearly matching members.
- Suggests **strong weekend opportunity for conversion** and targeted promotions.

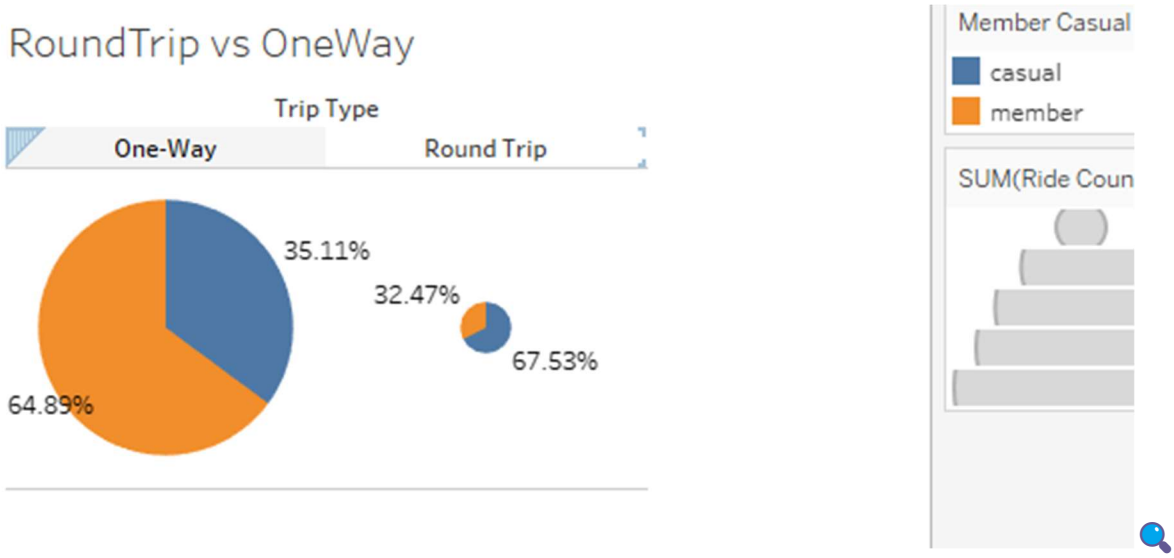
⏱️ 7. Ride Duration Buckets by User Type



🔍 Insight:

- **Members dominate short rides (0–30 mins)**.
- **Casuals have higher percentages in 30–60 and 60–120+ segments**.
- Casual users prefer **leisurely, longer rides**, especially on weekends.

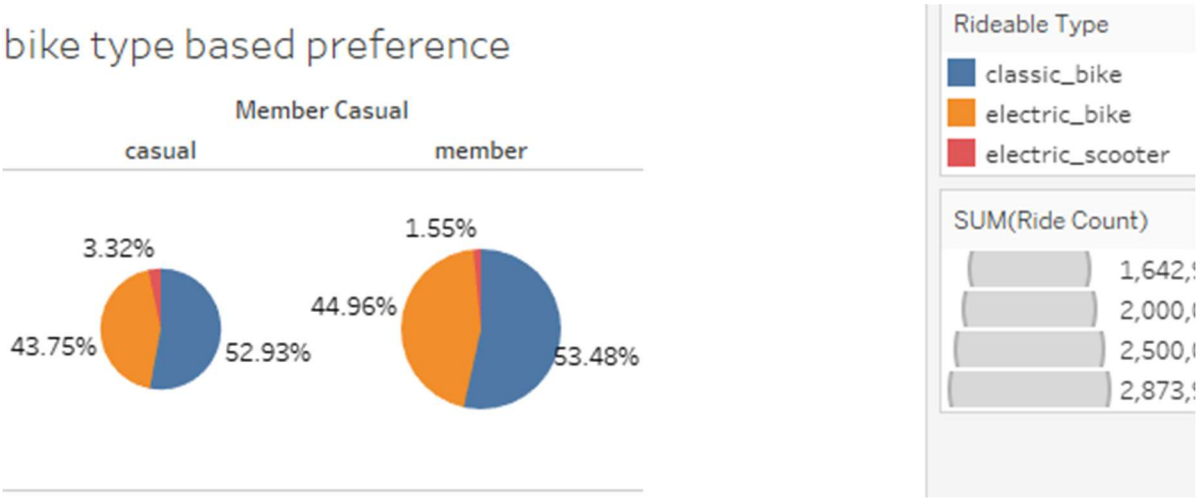
🔁 8. Roundtrip vs One-way Rides



Insight:

- **Casual riders** have a much **higher share of roundtrip rides**, reinforcing that they use the service recreationally.
- **Members** take more one-way trips, reflecting practical travel use (e.g., commute).

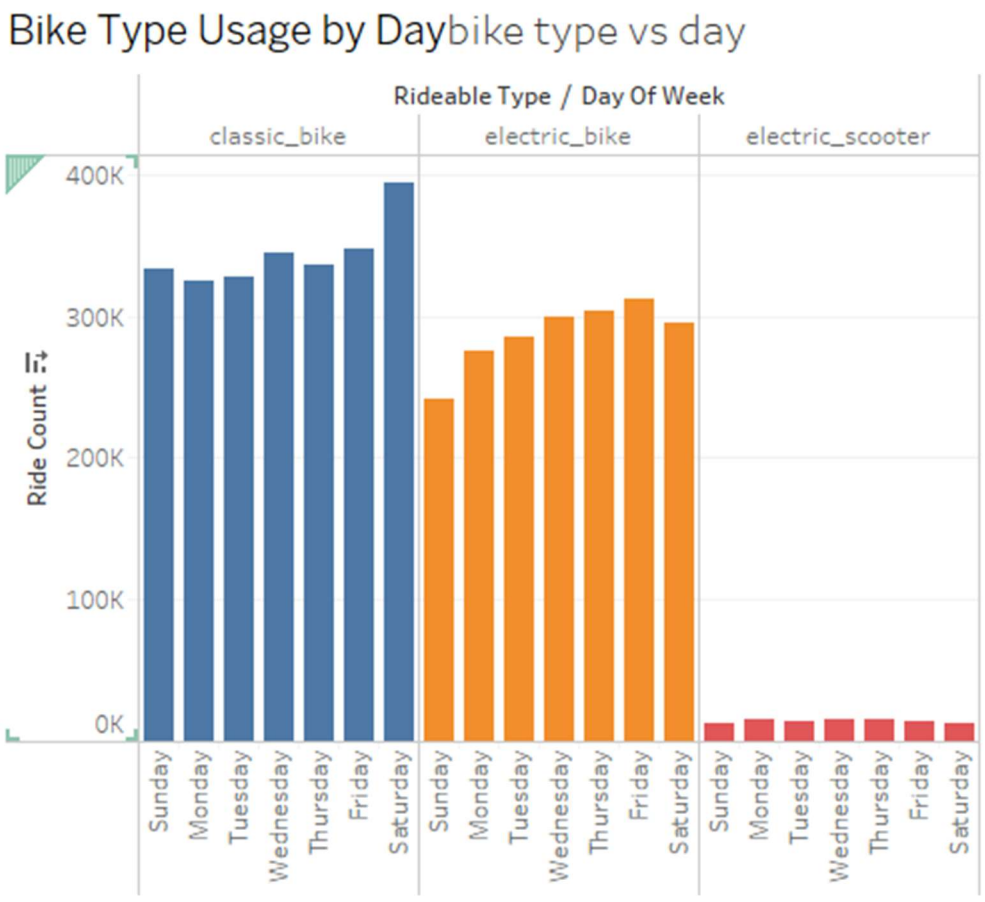
🚲 9. Bike Type Preference – User Type Breakdown



🔍 **Insight:**

- Both users primarily use **classic and electric bikes**.
- **Electric scooters**, while used less overall, are **more preferred by casual users (3%) than members (1.3%)**.

📅 **10. Bike Usage by Day of Week**



🔍 **Insight:**

- **Classic bikes** are the most consistently used across the week, with **Saturday being the peak day** (~440K rides). Usage is relatively stable from **Monday to Friday** (~320K–370K).
- **Electric bikes** show a **steady upward trend through the weekdays**, peaking on **Friday (~330K)**. This may indicate more convenience-based usage by both casuals and members toward the weekend.

- **Electric scooters** have **very low overall usage**, peaking just slightly around **Friday–Saturday**, never crossing ~30K per day.

📌 **Key Observations:**

- **Classic bikes = daily commuting + leisure**
- **Electric bikes = growing weekday preference**, suggesting they are gaining popularity among commuters
- **Electric scooters = still niche**, used more likely for short, experimental, or fun rides — mostly by casual users

📘 Phase 6 – Act: Turning Insights into Action

After a deep dive into data cleaning, exploration, SQL queries, and Tableau visualization, we now shift to the most critical phase — **Action**. This phase converts raw analysis into clear, implementable strategies that Cyclistic can use to **convert more casual users into annual members**.

📊 Key Differences: Casual vs Member

Feature	Casual Users	Annual Members
🔁 Trip Type	Mostly Round-Trips (67.5%)	Mostly One-Way (64.9%)
⏱️ Ride Duration	Higher Average (21 mins), Median (12 mins)	Lower Average (11.7 mins), Median (8 mins)
🗓️ Peak Usage Days	Weekends (Fri–Sun)	Weekdays (Mon–Fri)
🕒 Peak Usage Time	10 AM – 6 PM	8 AM & 4–6 PM (commute hours)
🚲 Bike Type Preference	Electric Bike & Scooters	Classic & Electric Bikes

📍 Popular Stations	Tourist/Scenic Spots (e.g., Millennium Park)	Commuter Hubs (e.g., Clinton St, Wells St)
🗓️ Seasonal Trend	More rides in summer + weekends	Year-round consistent usage

🎯 Targeted Recommendations to Convert Casual Riders into Members

After deep-diving into usage trends, seasonal patterns, and ride behaviors of casual vs. member users, here are **five actionable, data-driven strategies** designed to convert casual riders into loyal members.

🗓️ 1. Launch a Weekend-Only Membership Plan

💡 *Offer an affordable weekend membership that gives unlimited 30-minute rides every Saturday and Sunday.*

Why it'll work:
Our data reveals a sharp rise in casual ridership on weekends — jumping from **31.9% on weekdays to 47.5% on weekends**. A weekend-only membership directly caters to this usage pattern, making it easier for casual riders to commit without the pressure of a full-time plan.

🌞 2. Introduce a 3-Month Summer Pass

💡 *Roll out a seasonal 3-month membership (June–August or September) at a discounted rate.*

Why it'll work:
Casual ridership **peaks during summer months**. A short-term pass allows users to try membership during the months they're most active — perfect for testing commitment while offering value during high-demand periods.

🔁 3. Promote Round-Trip Discounts at Tourist Hubs

💡 *Provide incentives for round trips starting and ending at popular casual stations like Streeter Dr & Grand Ave or Millennium Park.*

Why it'll work:
More than **two-thirds of casual rides are round trips**, and many start at scenic hotspots. Offering visible discounts or bonuses at these stations encourages repeated use and loyalty, planting the seeds for longer-term membership.

⏱️ 4. Run Ride-Time Based Loyalty Offers

💡 *Reward longer or frequent rides (e.g., 60+ mins in a week) with ride credits or coupons.*

Why it'll work:
Casual users often take **longer rides than members**. Instead of penalizing them with fees, rewarding long-duration travel will enhance perceived value and build loyalty — nudging them toward switching to a cost-effective membership plan.

📍 5. Set Up On-Ground Campaigns at Top Casual Stations

💡 *Host seasonal pop-ups, QR code promotions, or referral booths at high-traffic stations.*

Why it'll work:
Stations like **Streeter Dr, Millennium Park, and DuSable Lake Shore** are favorites among casual users. Since there's no app for direct communication, real-world touchpoints can grab attention and drive immediate conversions at the source of usage.