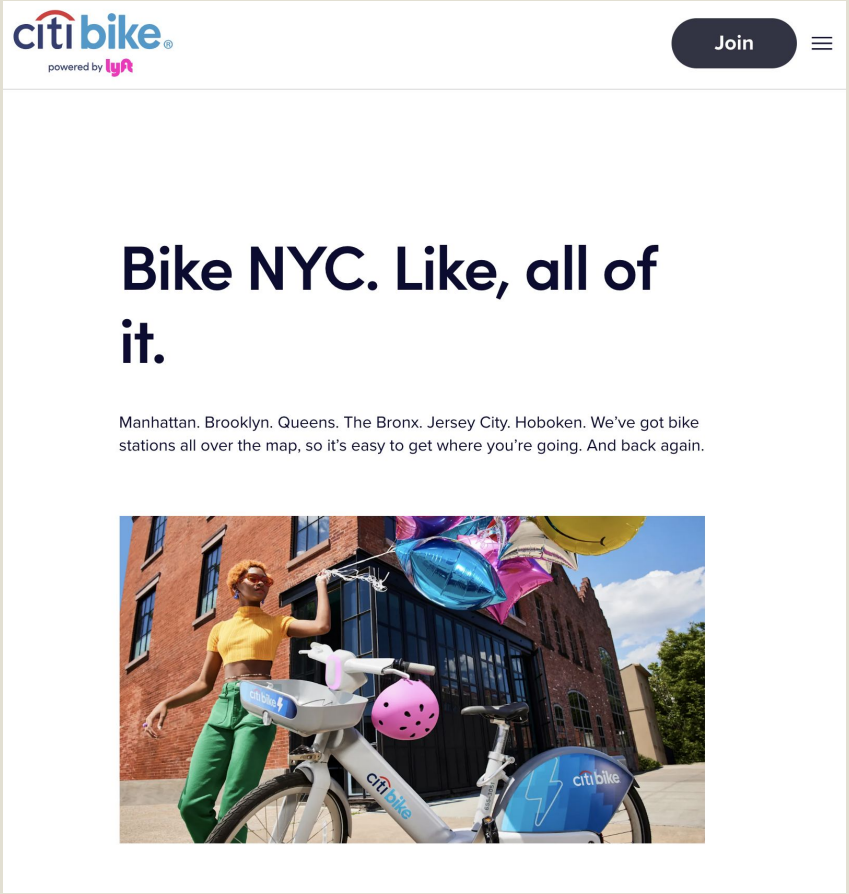
Citi Bike Customer Segmentation & Ride Pattern Analysis



Business Understanding

What Problem Are We Solving?

- Citi Bike wants to understand **how riders behave** to improve operations.
- Key questions:
  - Who are our customer segments?
  - When and how do they ride?
  - How to optimize resources (bikes, docks, rebalancing)?

Business Objectives

- Identify distinct **usage clusters**.
- Highlight **peak demand hours & days**.
- Provide **actionable recommendations** for marketing & operation
- The goal is segmentation not prediction. We want to reveal natural patterns in ride behavior to support better planning and customer targeting.

Dataset Quality

Dataset Overview

- ~5 million trips from multiple monthly Citi Bike files from the latest available dataset
- Key attributes: duration, distance, station locations, rider type, timestamps.
- Preprocessed by merging, cleaning timestamps, removing nulls.

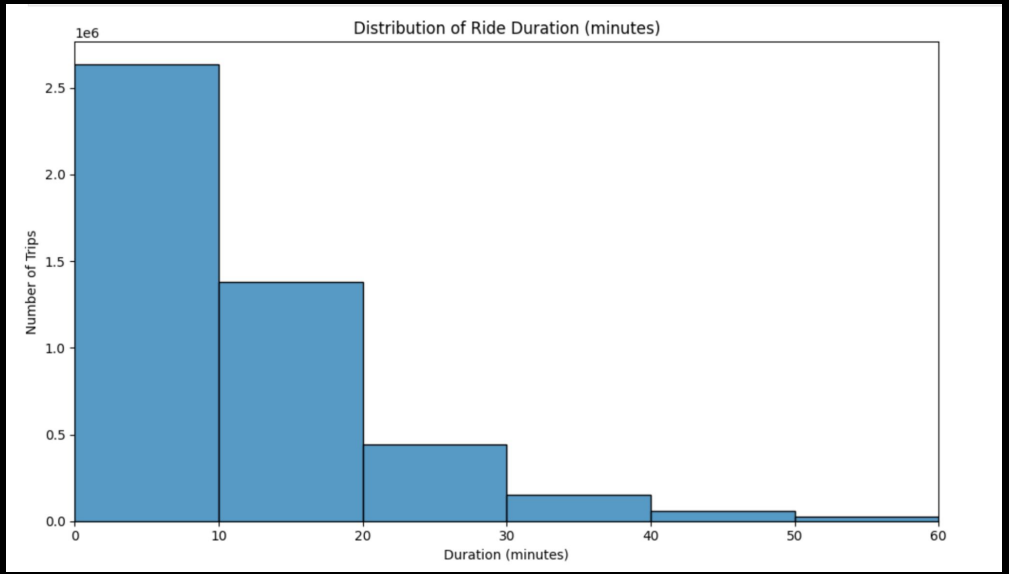
Quality Assessment

- High relevance: Captures real usage patterns.
- Large coverage: Multiple months, millions of rides.
- Limitations: Missing station IDs, extreme ride duration outliers.

EDA Insight 1: Ride Durations

Most rides are short

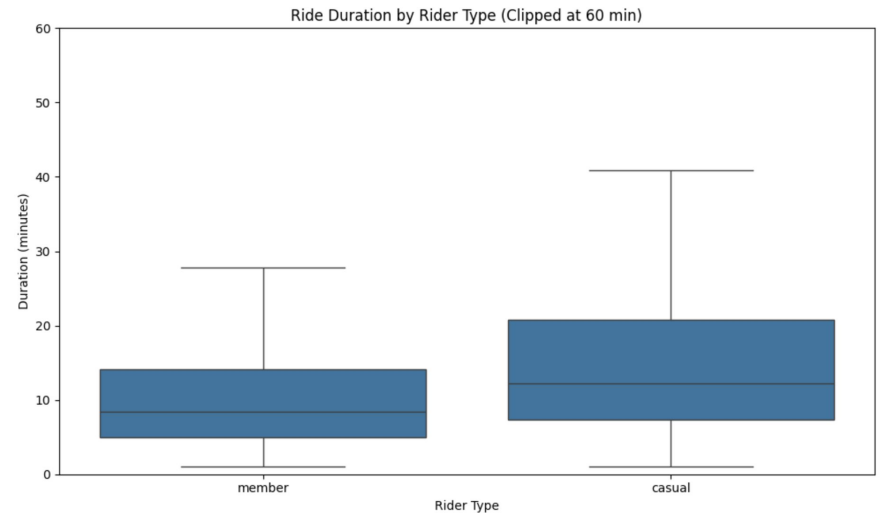
- 70%+ rides under 15 min
- Long rides exist but are rare
- Heavy right-skewed distribution



EDA Insight 2: Members vs Casual

Key Differences

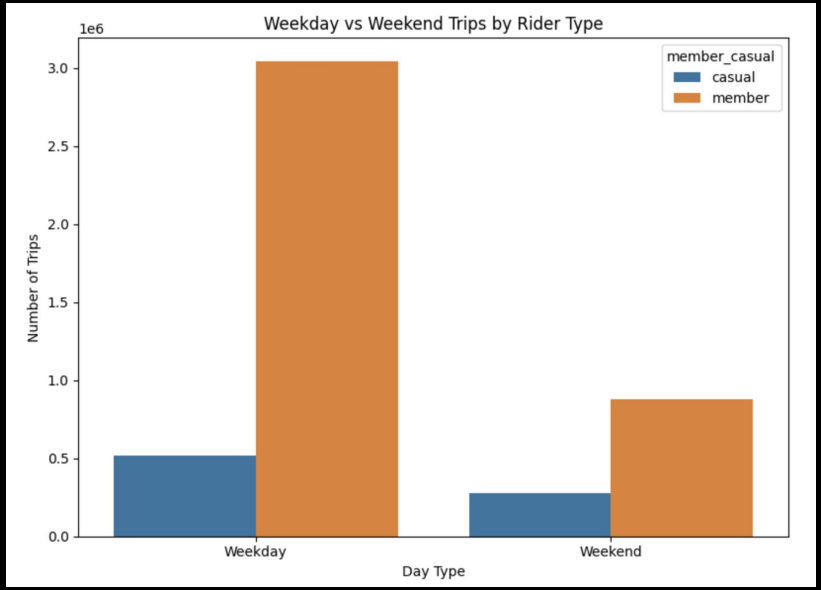
- Casual riders: Longer rides on average
- Members: Shorter, more predictable rides
- Members show commuter-like patterns.



EDA Insight 3: Weekday vs Weekend

Patterns

- Members dominate weekday commuting.
- Casual riders spike on weekends.
- This split hints at two very different use cases.



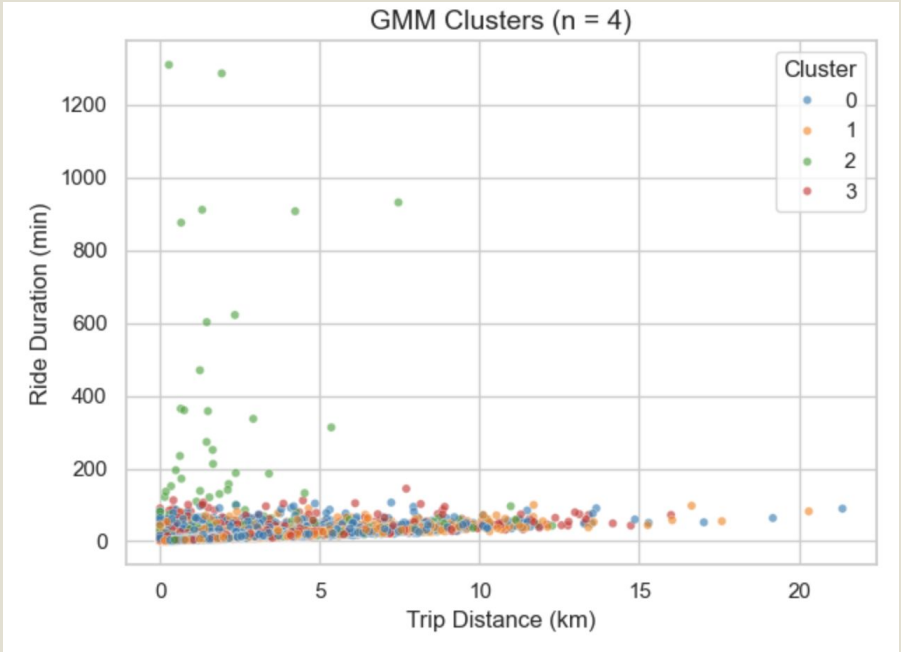
Modeling Approach

Why GMM (Gaussian Mixture Model)?

- Captures soft cluster boundaries (rides aren't strictly separated).
- Handles non-spherical clusters unlike K-Means.
- Works well for mixed behavioral features.

Features Used

- Duration
- Distance
- Hour of day
- Day of week
- Weekend share
- Peak hour share



GMM Cluster Profiles

4 Meaningful Segments Identified

1. Cluster 0 – Midday Short Riders
2. Cluster 1 – Peak-Hour Commuters
3. Cluster 2 – Weekend Explorers
4. Cluster 3 – Weekend Leisure Riders

GMM Cluster Profiles:

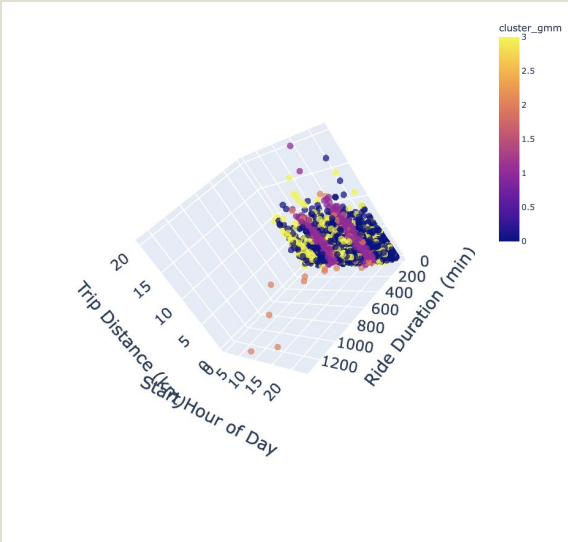
	cluster_gmm	avg_duration	avg_distance	avg_hour	avg_dayofweek	\
0	0	11.301376	1.916712	14.809693	2.267420	
1	1	11.355379	2.033465	13.450727	2.165984	
2	2	18.950226	2.032861	14.491468	5.401943	
3	3	13.683328	2.112145	13.433981	5.471318	

	weekend_share	peak_share	trips
0	0.00000	0.000000	40771
1	0.00000	1.000000	34684
2	0.98451	0.988317	7618
3	1.00000	0.000000	16927

Visualizing Clusters

Interpretation

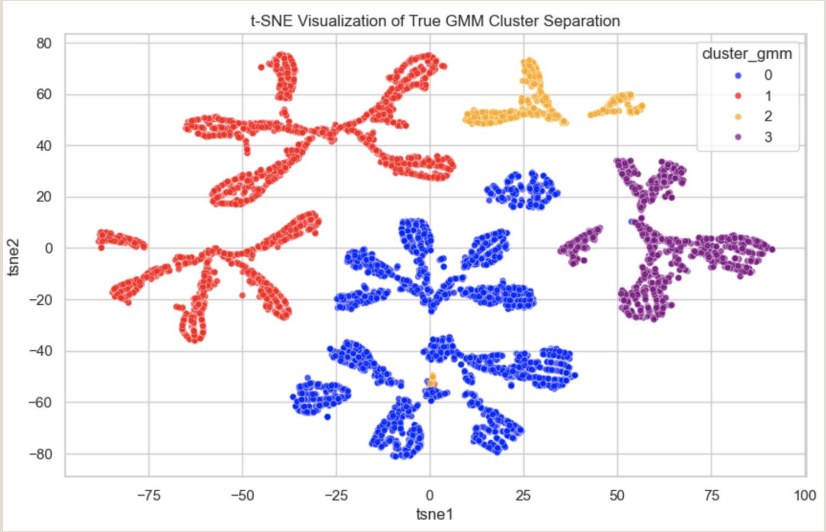
- Clear separation between weekday commuters vs weekend explorers.
- Duration & distance play strong roles.
- Time-of-day is a critical driver.



t-SNE Visualization of GMM Cluster Separation

Key Insights

- **Distinct, well-defined clusters:**
  t-SNE shows that the GMM model successfully identifies **four clearly separated behavioral groups**, confirming strong underlying structure in rider activity.
- **Cluster 0 (Blue) – Midday Short Riders:**
  Forms compact, structured branches → consistent short rides with similar timing.
- **Cluster 1 (Red) – Peak-Hour Commuters:**
  Large, radiating shape → high volume, strong directional commuting patterns.
- **Cluster 2 (Orange) – Weekend Explorers:**
  Small but dense region → longer leisure rides, tightly grouped.
- **Cluster 3 (Purple) – Weekend Neighborhood Riders:**
  Separate formation to the right → different timing + zone characteristics.

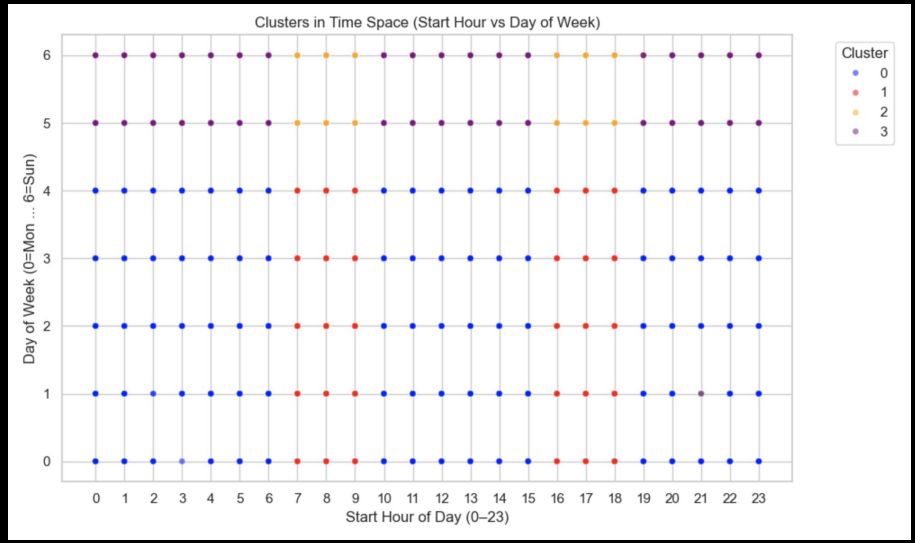


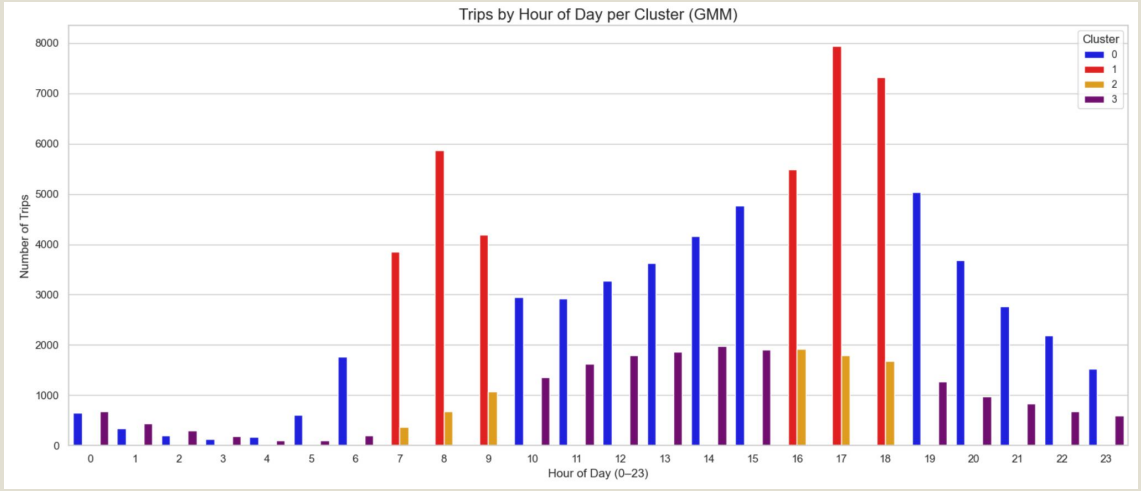
t-SNE is a nonlinear dimensionality reduction technique that reveals hidden structure. The fact that clusters separate this cleanly means Citi Bike users behave in very distinct patterns, validating our modeling approach. This strengthens our confidence in making cluster-based business decisions.

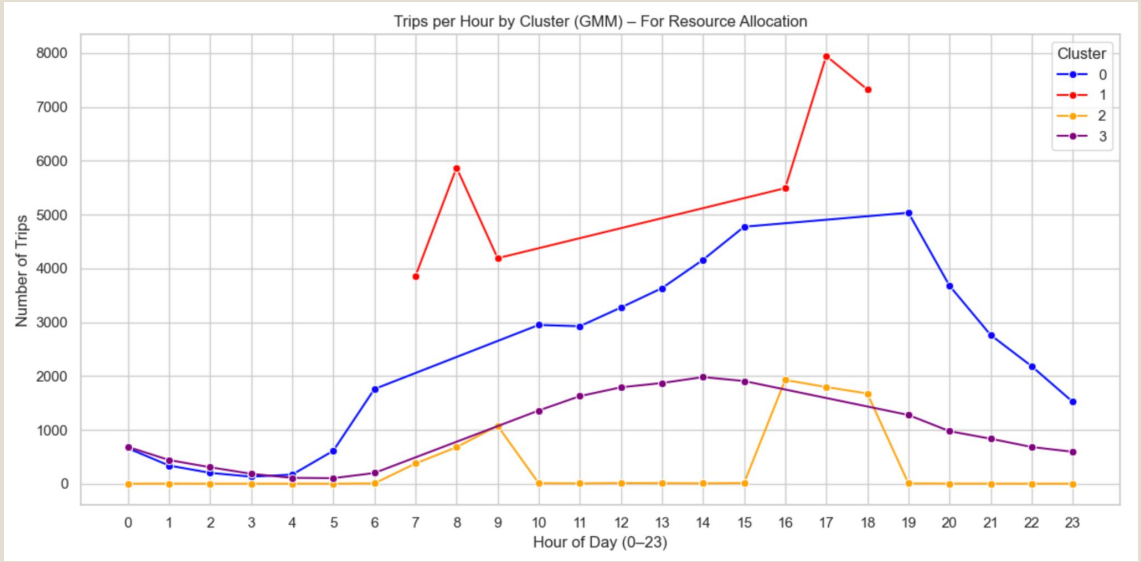
Hourly Demand by Cluster

Operational Insight

- Cluster 1 (Commuters) peaks sharply at 8–9 AM and 5–6 PM.
- Cluster 0 (Midday users) rise from 10 AM → 4 PM.
- Clusters 2 & 3 dominate weekends.




Trips by Hour of Day per Cluster (GMM)
Cluster
0
1
2
3
Number of Trips
Hour of Day (0–23)

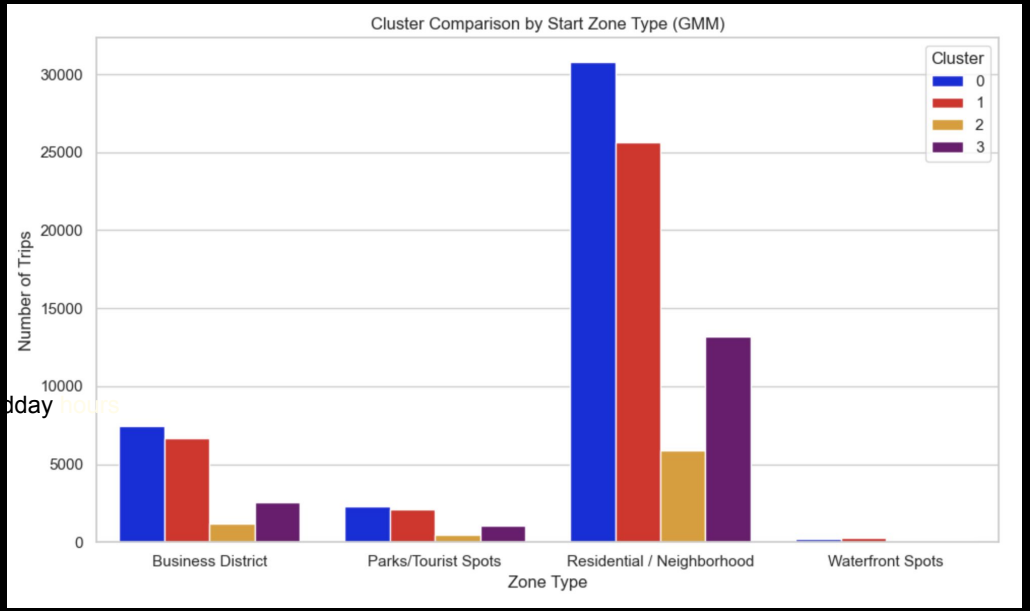

Trips per Hour by Cluster (GMM) – For Resource Allocation
Cluster
0
1
2
3
Number of Trips
Hour of Day (0–23)

Cluster Comparison by Start Zone Type

Where Do Different Clusters Start Their Trips?

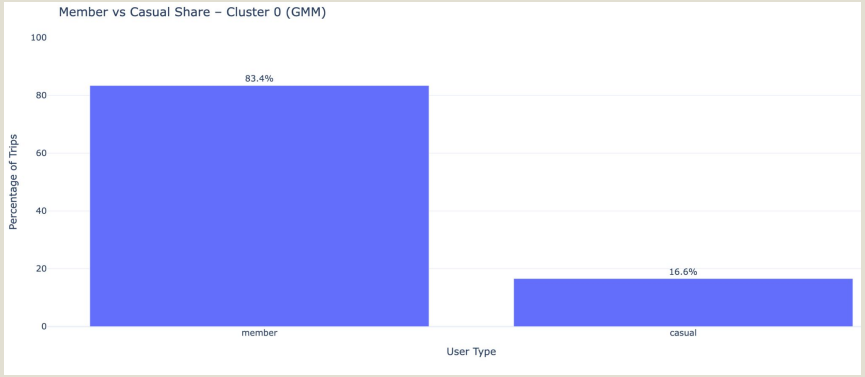
Key Insights

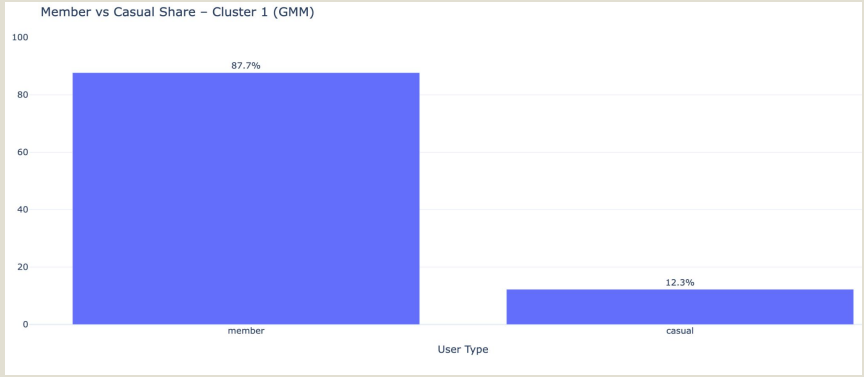
- Clusters 0 & 1 (Commuter Segments)
  - Strongly concentrated in Residential/Neighborhood zones
  - Indicates home→work or local daily movement
  - Significant presence in Business District stations as well
- Cluster 2 (Weekend Explorers)
  - Higher share in Tourist/Park zones
  - Long-duration leisure rides starting from recreational areas
- Cluster 3 (Leisure Neighborhood Riders)
  - Also concentrated in Residential but with more weekend timing and midday hours
  - Not tied to business zones, more relaxed usage

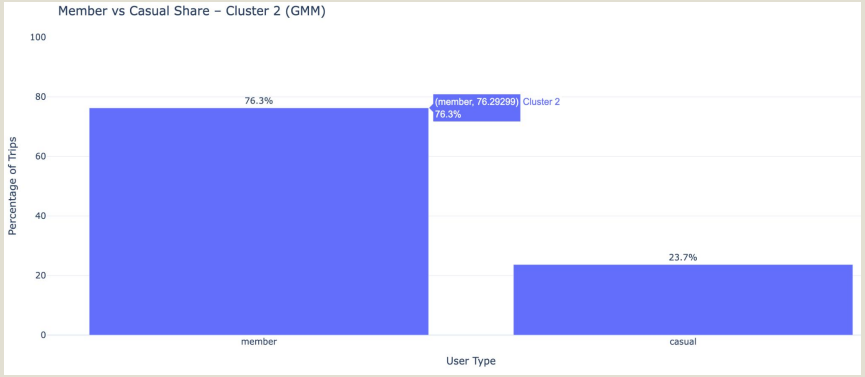


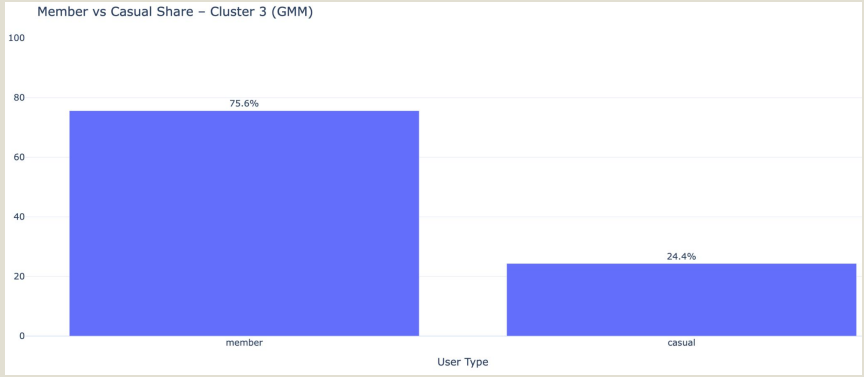
Business Interpretation

- Residential zones are the core demand hubs across all clusters.
- Parks/tourist areas attract longer, leisurely rides → target for tourist pricing bundles.
- Business district stations show commuting patterns → focus on morning/evening bike redistribution during peak hours.


Member vs Casual Share – Cluster 0 (GMM)
Percentage of Trips
100
80
60
40
20
0
83.4%
16.6%
member
casual
User Type


Member vs Casual Share – Cluster 1 (GMM)
100
80
60
40
20
0
87.7%
12.3%
member
casual
User Type


Member vs Casual Share – Cluster 2 (GMM)
Percentage of Trips
100
80
60
40
20
0
76.3%
(member, 76.29299) Cluster 2
76.3%
23.7%
member
casual
User Type


Member vs Casual Share – Cluster 3 (GMM)
100
80
60
40
20
0
75.6%
24.4%
member
casual
User Type

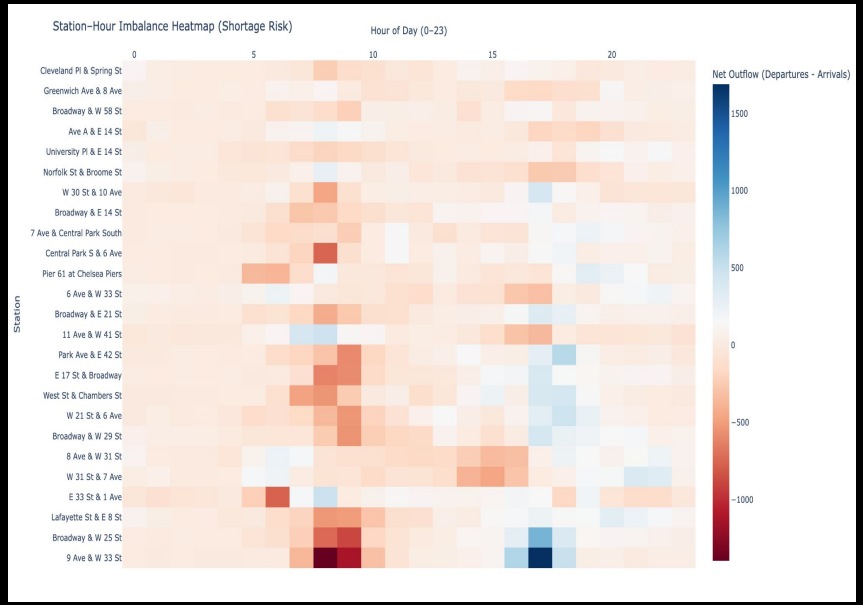
Where and When Do Stations Run Out of Bikes?

- Morning Shortages (7–10 AM):
  Several stations near residential neighborhoods show heavy *red* zones → high departures, low arrivals.
  These are commuter outflow bottlenecks.

- Evening Shortages (5–7 PM):
  Strong *blue* areas at business-dense locations (e.g., W 31 St & 7 Ave) → overwhelming arrivals.
  Bikes accumulate here, causing dock-full issues.

- Tourist Hotspots:
  Moderate imbalance late mornings and afternoons, especially at Central Park & waterfront stations.

- Critical Stations Identified:
  • W 31 St & 7 Ave
  • 6 Ave & W 33 St
  • Central Park South
  • Broadway & E 14 St
  These require active rebalancing during commute windows.



Business Interpretation

- Heatmap shows hours where Citi Bike risks running *out of bikes* or *running out of docks*.
- Enables hour-by-hour fleet rebalancing strategy.
- Red zones = bike shortages → place more bikes early morning.
- Blue zones = dock shortages → schedule pickups before evening peaks.

Actionable Recommendations

Marketing

- **Target casual users** on weekends with:
  • Day passes
  • Tourist partner discounts

- **Promote membership** to Cluster 0 (midday regulars).

Operations

- Increase bike rebalancing during **commute hours**.

- Add more bikes to **tourist-heavy stations** on weekends.

- Improve station availability around **8–10 AM** & **5–7 PM**.

Product

- Offer **weekend bundles** for Cluster 2 & 3.

- Build a **commuter reliability guarantee** for Cluster 1.

What If / Future Work

With More Data

- Add user demographics for richer segmentation.
- Incorporate weather & events for deeper demand forecasting.
- Try deep clustering or HDBSCAN for more organic segments.

Next Steps

- Build a **demand prediction model** for rebalancing.
- Publish analysis on GitHub with full reproducible pipeline.

CONCLUSION

In this project, we analyzed millions of Citi Bike trips to understand who uses the system, when they ride, and how bike demand shifts across New York City's stations and hours of the day. Using clustering techniques (GMM, t-SNE) and zone classification, we identified four distinct rider groups with unique usage patterns, travel times, and starting locations. We also built hourly shortage/surplus heatmaps to pinpoint when and where docks risk running empty or overfilled.

Overall, by combining clustering, zone classification, and station hour imbalance analysis, this project provides a data driven foundation for improving bike availability, reducing operational cost, and designing smarter customer strategies. It helps the company run a more reliable system, enhance user satisfaction, and capture new revenue opportunities based on clear and actionable patterns in rider behavior

THANK YOU