How does a bike-share navigate speedy success?

Table of Contents

Ask

Business task

Determine how annual members and casual riders use Cyclitic bikes differently so marketing can convert casual riders into annual members.

Prepare

Where is your data located?

The dataset is provided publicly by Divvy Tripdata.
I downloaded the most recent 12 months (September 2024 – August 2025).

How is the data organised?

Each month is a .zip file containing a .csv.
Each CSV has hundreds of thousands of rows (millions in total across 12 months).
Columns include:
ride_id – unique trip identifier
rideable_type – type of bike
started_at, ended_at – trip timestamps
start_station_name, end_station_name – station information
start_lat, start_lng, end_lat, end_lng – GPS coordinates
member_casual – rider type

Due to size limits in Excel, I used R to combine and clean the data, then exported summary tables for visualisation.

Are there issues with bias or credibility in this data? Does your data ROCCC?

Reliable: From Divvy, the operator of Chicago's official bike share programme.
Original: Raw operational trip data.
Comprehensive: All trips from the period, not sampled.
Current: Latest 12 months (Sept 2024 – Aug 2025).
Cited: Official S3 repository.

Limitations:
No personal identifiers are provided (privacy preserved).
Some records contain errors (negative or zero ride times).
Limited to Chicago, so results may not generalise elsewhere.

How are you addressing licensing, privacy, security, and accessibility?

The dataset is publicly licensed and anonymised (no PII).
Files stored locally in a secure raw_csv/ folder.
Only summaries (aggregated files) are shared, not raw trip-level data.

How did you verify the data's integrity?

In R:
Converted timestamps to proper datetime (ymd_hms).
Created ride_length (minutes), day_of_week, and start_hour columns.
Removed trips with ride_length <= 0.
Checked row counts and file merges to ensure no data loss.
Exported clean summary tables for Excel visualisation:
summary_by_user.csv – average ride length and total rides by member type.
summary_by_weekday.csv – ride patterns by day of week and member type.
summary_by_hour.csv – ride counts by hour of day and member type.

How does it help you answer your question?

summary_by_user shows whether casual riders take longer trips.
summary_by_weekday reveals weekday vs weekend differences.
summary_by_hour highlights commuting vs leisure patterns.
Together, these summaries directly answer: *"How do members and casuals use Cyclistic differently?"*

Are there any problems with the data?

Very large file size → required R to pre-process before using Excel.
Occasional missing station names or GPS errors.
Some abnormal trips filtered out (negative durations).

Deliverable: Data Sources & Processing

Raw data: 12 monthly CSVs (Sept 2024 – Aug 2025) from Divvy Tripdata Repository.
Processing tool: R (tidyverse, lubridate).
Outputs:
summary_by_user.csv
summary_by_weekday.csv
summary_by_hour.csv
Storage: Raw files in /Documents/raw_csv/; summary files exported to Excel for analysis and charting.

Process

What tools are you choosing and why?

I used R with the tidyverse and lubridate libraries to combine and process all 12 months of raw CSV data.
R was chosen because the dataset (~5.6 million rows) is too large for Excel, and R allows faster cleaning, transformation, and summary calculations.
I exported summary files (summary_by_user.csv, summary_by_weekday.csv, summary_by_hour.csv) so that Excel could be used for PivotTables and visualisation.

Have you ensured your data's integrity?

Verified row counts after combining 12 CSVs matched the sum of the original files.
Checked that datetime columns (started_at, ended_at) imported properly.
Confirmed ride counts and averages aligned with known Divvy usage patterns (e.g., members have shorter trips but more rides).

What steps have you taken to ensure that your data is clean?

Imported & Combined Data: Used list.files() and map_df(read_csv) to load and combine 12 monthly CSVs.
Created New Columns:
ride_length (minutes): as.numeric(difftime(ended_at, started_at, units="mins"))
day_of_week: wday(started_at, label=TRUE, abbr=FALSE)
start_hour: hour(started_at)
Filtered Out Bad Data:
Removed rows with ride_length <= 0.
Removed outliers (extremely long rides suggesting errors).
Standardised text values in member_casual.
Aggregated into Summary Tables:
summary_by_user: average ride length + total rides by member type.
summary_by_weekday: averages and ride counts by day of week.
summary_by_hour: ride counts by hour of day.

How can you verify that your data is clean and ready to analyse?

No duplicate ride_ids remain.
No negative or zero ride lengths.
All categorical fields (member_casual, day_of_week) are standardised.
Aggregated summaries show logical patterns (e.g., member commute peaks, casual weekend peaks).

Have you documented your cleaning process?

Yes. I created a data cleaning log describing each transformation step, with before/after row counts where applicable.

Step	Action	Rows Before	Rows After	Notes
1	Combined 12 CSVs	5,646,325	5,646,325	All months loaded
2	Removed ride_length <= 0	5,646,325	5,645,918	407 rows removed
3	Standardised member_casual	5,645,918	5,645,918	No row loss
4	Exported summaries	5,645,918	5,645,918	Clean master ready

Key Tasks Completed

Checked data for errors.
Chose R for cleaning and Excel for visualisation.
Transformed data with new columns.
Documented cleaning steps.

Deliverable: Data Cleaning Documentation

Tool used: R (tidyverse, lubridate).
Cleaning actions: removed invalid rows, standardised categories, created calculated fields, exported summaries.
Outputs: summary_by_user.csv, summary_by_weekday.csv, summary_by_hour.csv.

Analyse

How should you organise your data to perform analysis on it?

I aggregated the trip-level dataset into three summary views using R:
summary_by_user – Average ride length and ride counts by rider type.
summary_by_weekday – Ride patterns by weekday.
summary_by_hour – Ride counts by hour.
These were exported to Excel for PivotTables and visualisation.

Has your data been properly formatted?

Yes: ride lengths in minutes (numeric), dates parsed into hours/days, rider types standardised.
Pivot tables formatted with decimal places, commas, and weekday order.

What surprises did you discover?

Casual riders take **almost double the average trip length** of members (22.9 vs 12.1 minutes).
Casual rides spike on weekends, while member rides are steady Mon–Fri.
Members peak at commute times (8am, 5–6pm), casuals peak midday.
Despite longer trips, casuals account for fewer total rides (2.08M vs 3.57M).

What trends or relationships did you find?

Ride length: Casual = longer, fewer trips; Member = shorter, frequent trips.
Weekday: Members = weekday commuters; Casuals = weekend leisure.
Hourly: Members = commute peaks; Casuals = midday leisure.

How will these insights help answer the business question?

They highlight **different usage behaviours**, suggesting tailored marketing strategies:
Show **cost savings** for weekend casuals.
Promote **commute benefits** (speed, reliability).
Run **midday promotions** targeting leisure users.

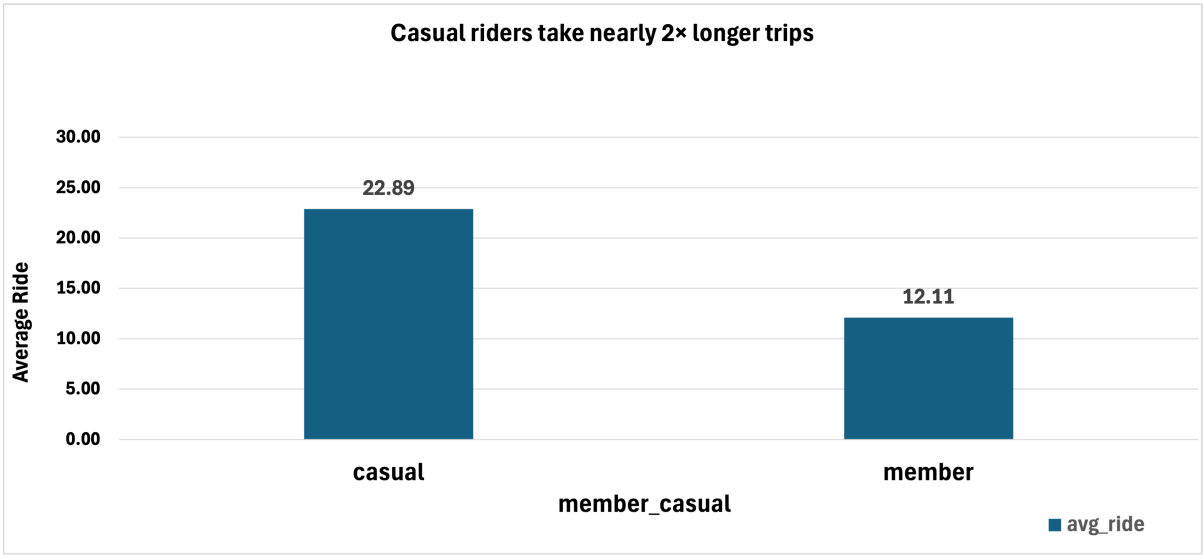
Key Tasks Completed

Aggregated 12 months of data into summary files with R.
Organised and formatted them for Excel PivotTables.
Performed descriptive calculations (means, medians, max, mode).
Identified clear trends across rider types.

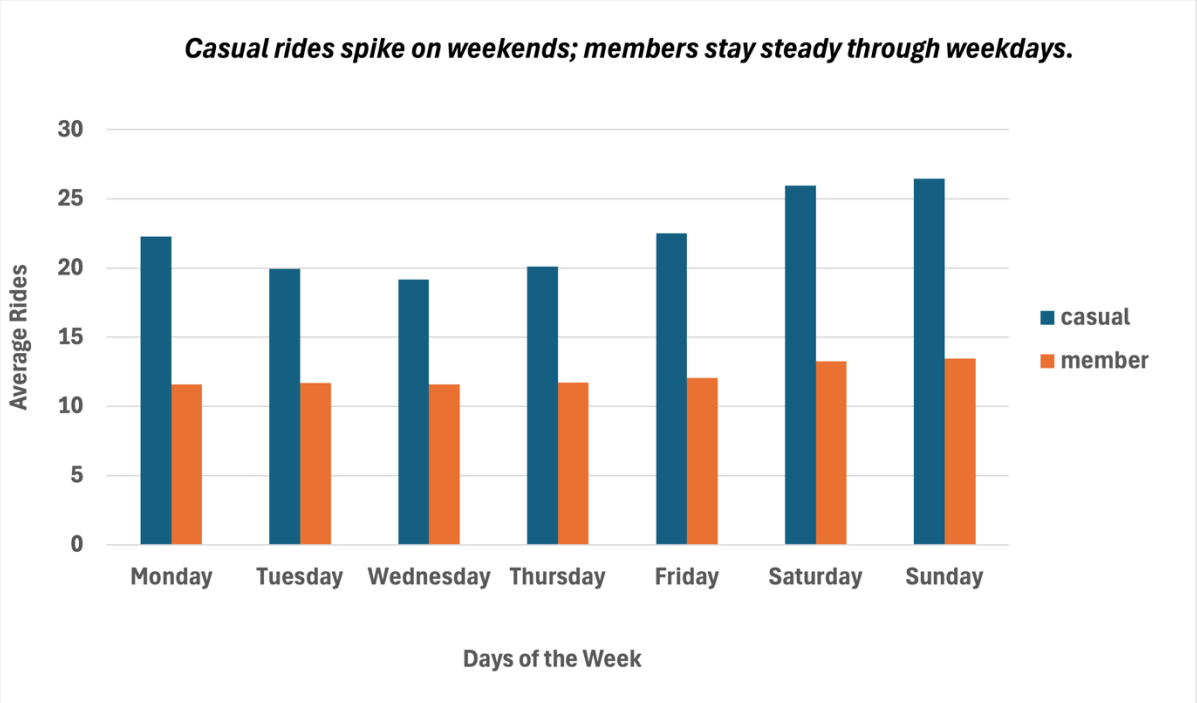
Deliverable – Summary of Analysis

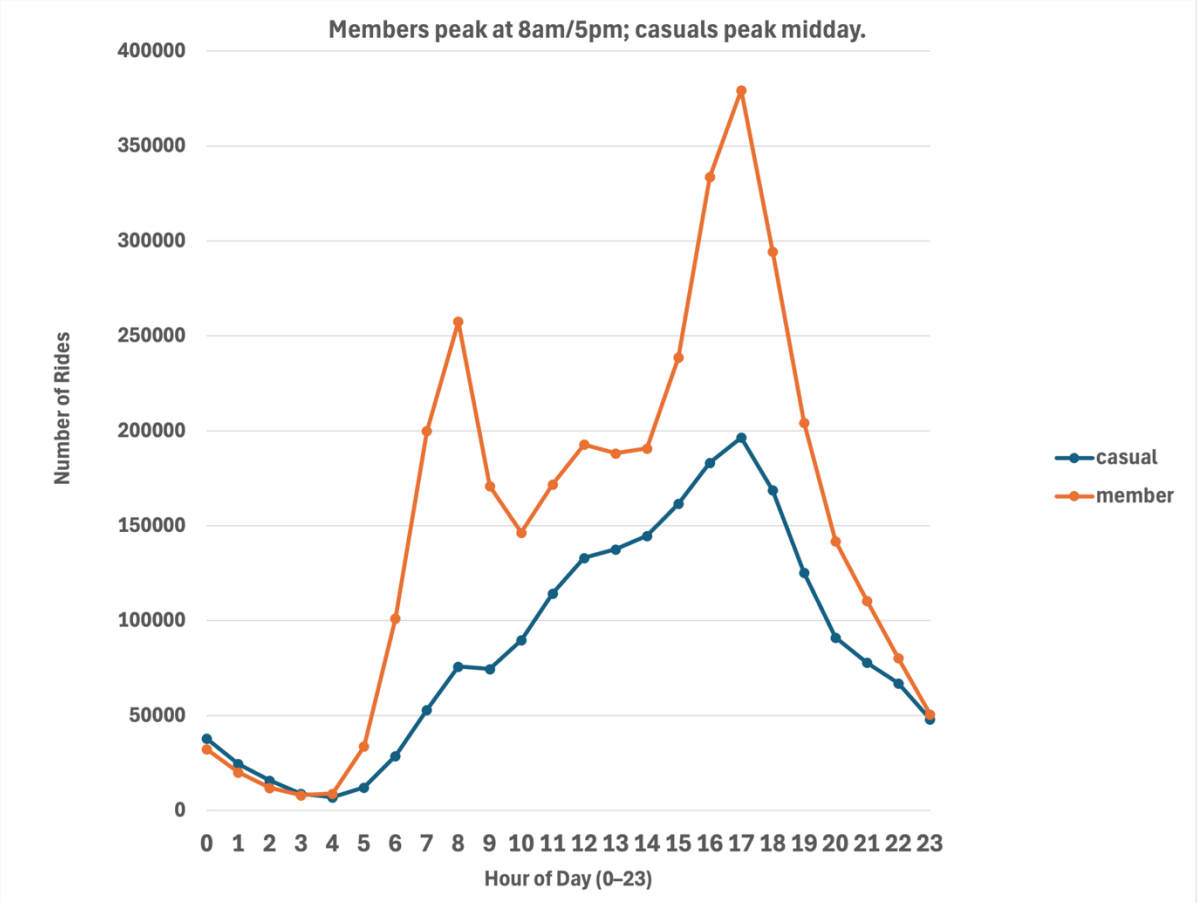
Members dominate total rides (63%) but take shorter trips (~12 mins).
Casuals average longer trips (~23 mins) but fewer rides (37%).
Members ride consistently Mon–Fri, casuals peak on weekends.
Members commute at 8am/5–6pm, casuals ride midday.
Implication: Cyclistic should target casuals with weekend and midday promotions, positioning membership as both cost-saving and flexible for leisure.

Share


Casual riders take nearly 2× longer trips
30.00
25.00
20.00
15.00
10.00
5.00
0.00
Average Ride
22.89
12.11
casual
member
member_casual
avg_ride


Members take more rides than casual riders
4
3.5
3
2.5
2
1.5
1
0.5
0
Total Rides (Millions)
2.077478
3.568517
casual
member
Sum of rides


Casual rides spike on weekends; members stay steady through weekdays.
Average Rides
30
25
20
15
10
5
0
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Days of the Week
casual
member


Members peak at 8am/5pm; casuals peak midday.
Number of Rides
400000
350000
300000
250000
200000
150000
100000
50000
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Hour of Day (0–23)
casual
member

Act

Weekend Pass Conversion: Offer casual riders a weekend membership deal, showing cost savings compared to pay-as-you-go.

Target Frequent Casuals: Identify repeat casual riders and show them personalised cost comparisons highlighting savings with annual membership.

Midday Promotions: Run adverts during casual rider peak hours (11am–4pm), highlighting flexibility, unlimited rides, and leisure appeal.