Cyclistic rider analysis

Austin Broadbent

March 27, 2025

executive summary

project goal

- Understand how casual riders and annual members use Cyclistic bikes differently.

key findings

- Identified differences in how casual riders and annual members use bikes
- Determined factors that would incentive casual riders to purchase annual memberships

insights

- Casual riders are less frequent users, but take significantly longer trips
- Suggests a potential market for membership focused on recreation or longer-duration use

agenda

01 problem

motivating problem & objectives of the analysis

02 findings

visualizations & insights gained

03 proposals

marketing strategies & digital media channels

04 impact

expected outcomes & key metrics for success

01
problem

How can we maximize Cyclistic's annual membership base?

analysis objectives

identify

differences in how casual riders and annual members use bikes

determine

factors that would incentive casual riders to purchase annual memberships

key questions

one

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

two

- Why would casual riders buy Cyclistic annual memberships?

three

- How can Cyclistic use digital media to influence casual riders to become members?

structure of data

- 4,105,434 unique rides
- date range: Mar 2024 - Feb 2025

data variables

- unique ride ID
- rider type, bike type
- temporal & geographical data

data limitations

missing data

- a large number of records with missing data were filtered out

imprecise data

- some records had imprecise GPS data and were also filtered out

outliers

- negative, extra short, or extra long duration rides removed

inconsistencies

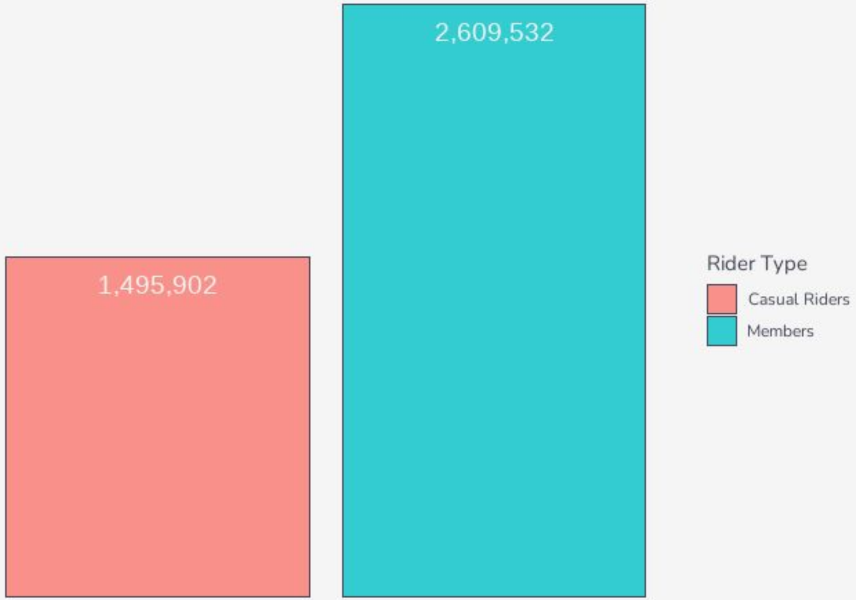
- several station ids match multiple station names, and vice versa

02
findings

with visualizations

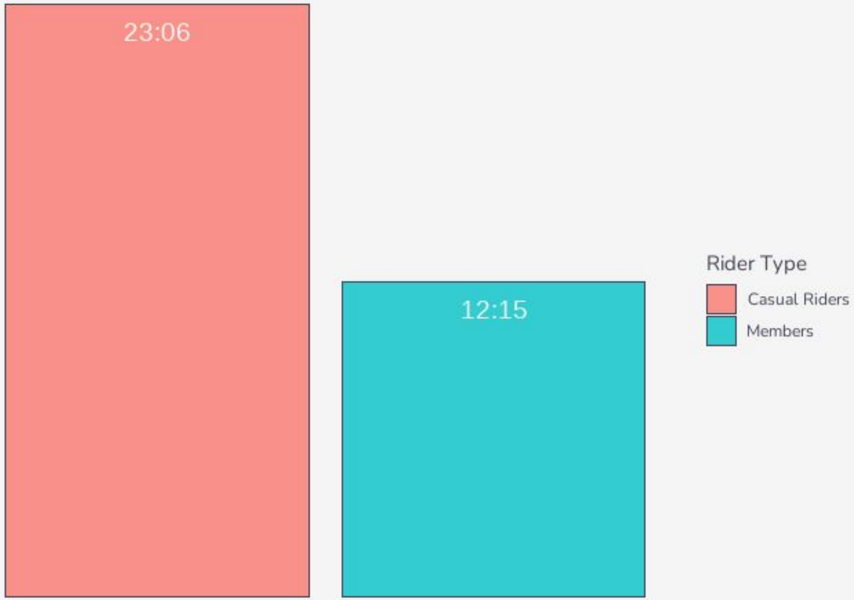
Total Number of Rides

Casual Riders vs. Members



Total Average Duration in Minutes

Casual Riders vs. Members



63.56%

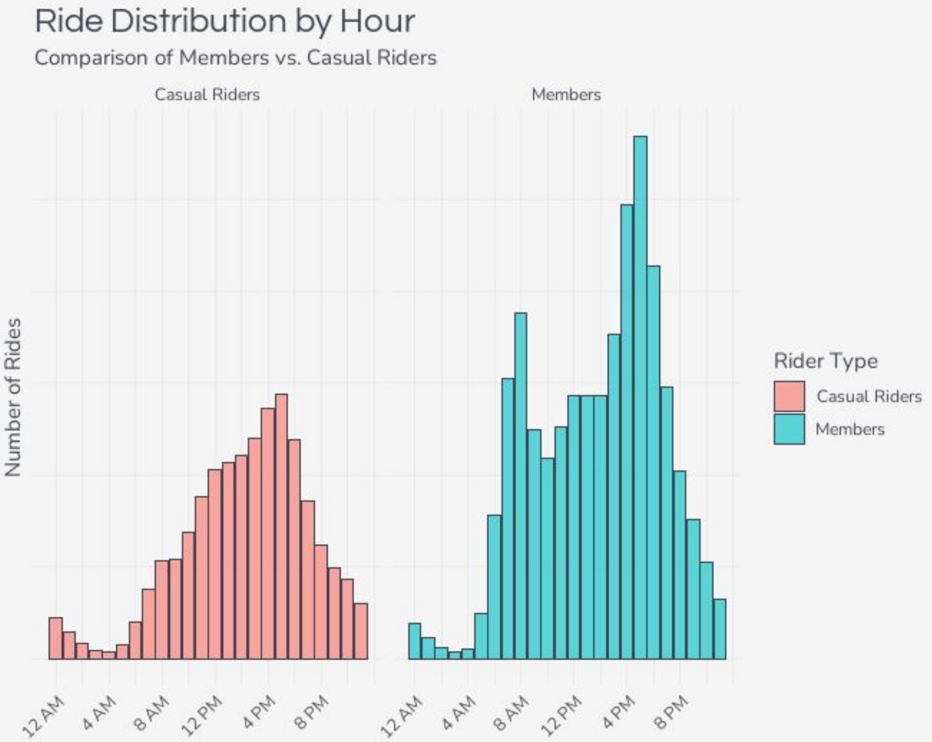
of trips are by **members**

1.9x longer

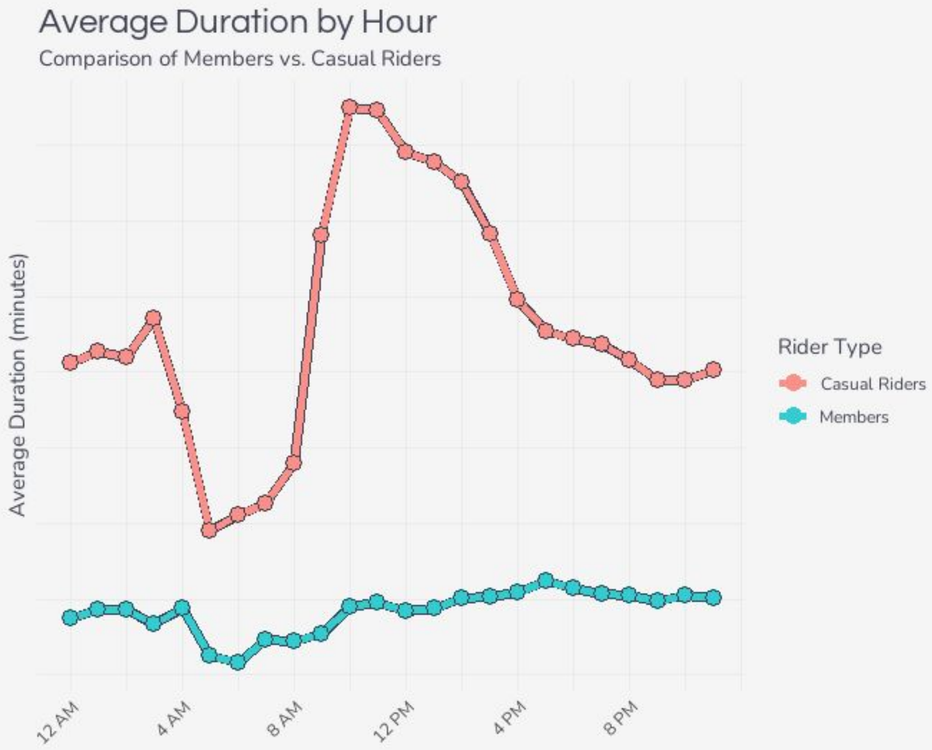
average trip duration by **casual riders**

usage patterns **by time**

- **Members** have peak usage during commuting hours.
- **Casual riders** during midday and afternoon hours.



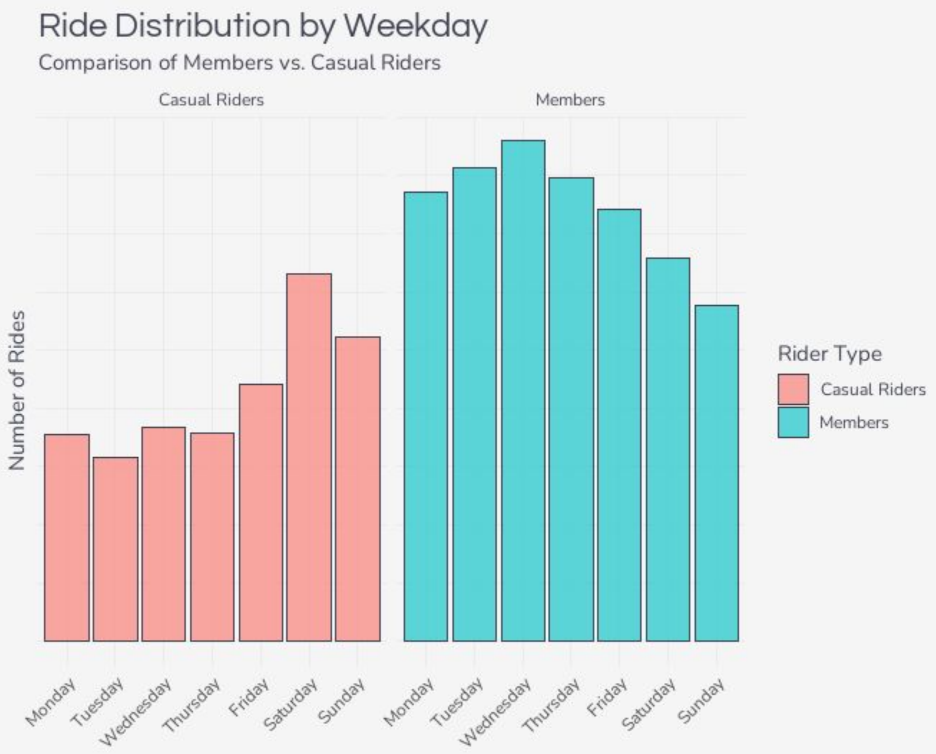
usage patterns **by time**



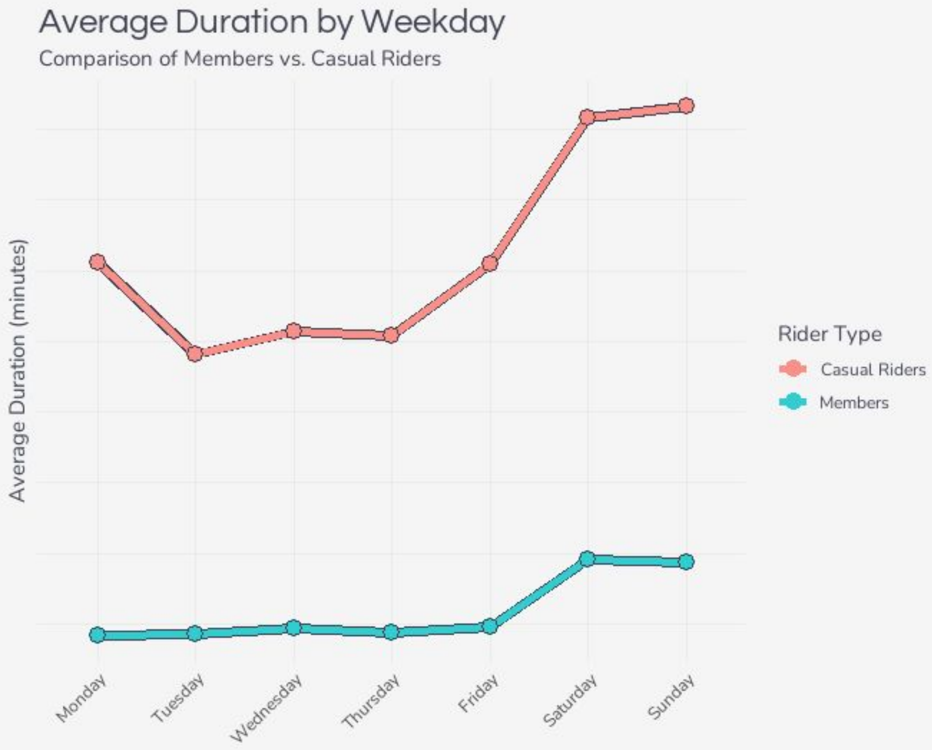
- **Casual riders** have longer averages and more varied ride times throughout the day.

usage patterns **by day**

- **Members** use bikes on weekdays.
- **Casual riders** on weekends.



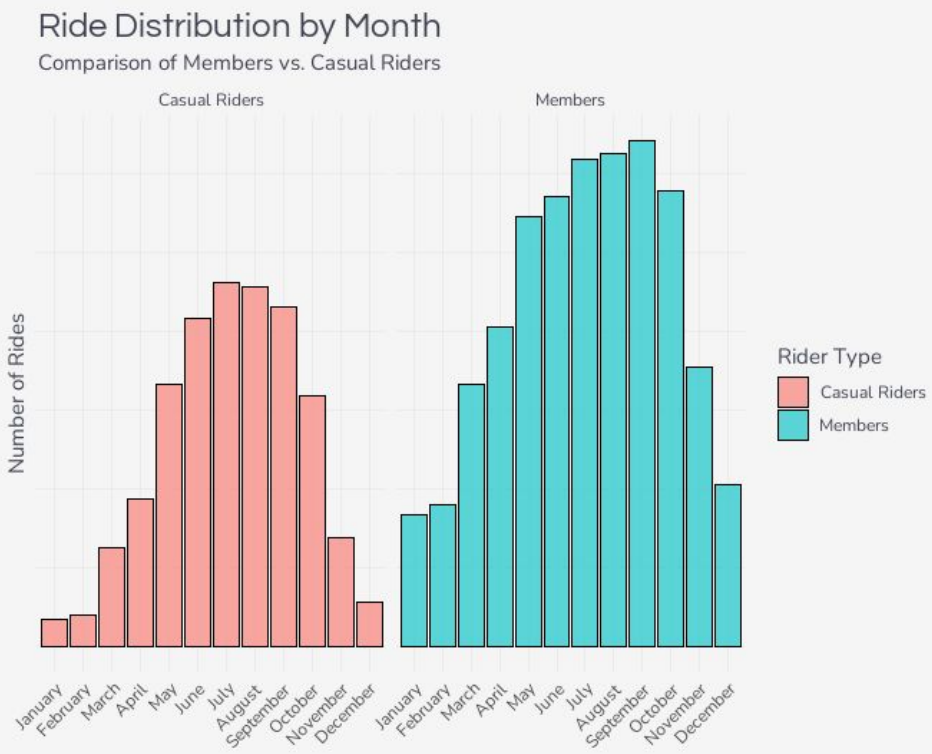
usage patterns **by day**



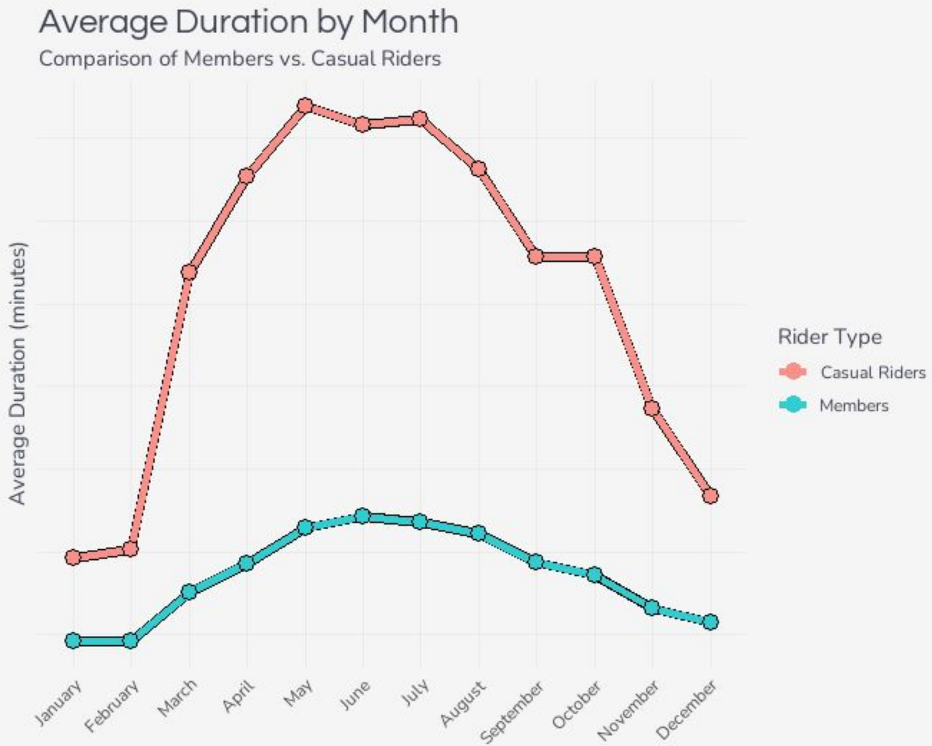
- **Casual riders** also have longer average ride times on weekends.

usage patterns **by month**

- Both user types show increased use during warmer months. But, the effect is larger for **casual riders**.



usage patterns **by month**



- **Casual riders** have much longer average ride times during the warmer months.

insights: casual ridership

casual ridership

Recreational focus: Unlike members, casual riders use Cyclistic for leisure, exploring the city, and enjoying longer rides.

Peak usage patterns:

- Weekends
- Warmer months
- Midday/Afternoon

Preference for Longer Rides: Suggests that they value the journey and the experience

03

proposals

potential marketing strategies &
digital media channels

marketing strategies

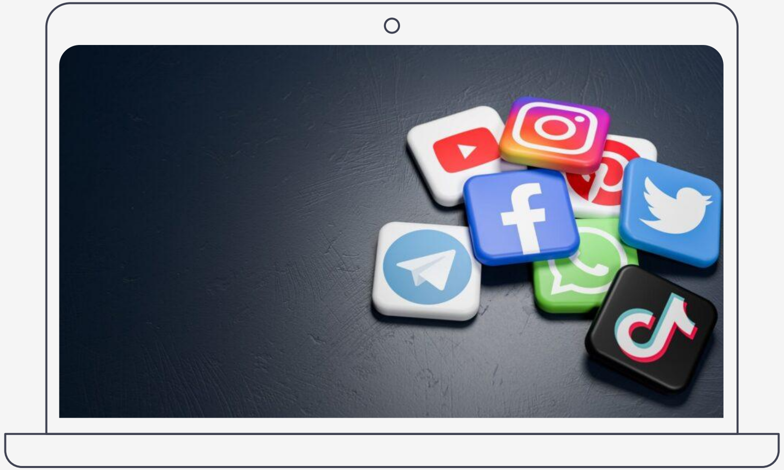
targeted memberships

- Weekend Warrior Pass
- Summer Fun Pass
- Explorer Membership

promotional offers

- Weekend discounts
- Seasonal bundles
- First-Ride Free

media channels



social media

- Blog Posts
- Social Media Campaigns
- User-Generated Content

partnerships

- Local Events
- Tourism Industry
- Parks/Recreational Areas

04
impact

expected outcomes &
metrics for success

expected outcomes

increased membership

directed marketing should lead to a significant increase in casual rider conversions to annual memberships

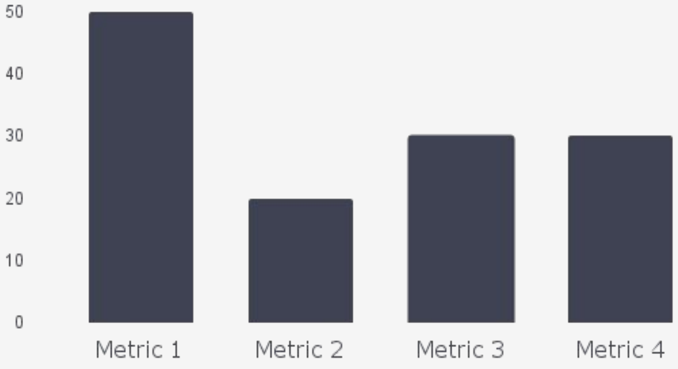
revenue growth

increased membership will lead to higher overall revenue, as well as a more stable revenue stream

greater customer loyalty

members should have a stronger relationship with the company, increasing customer retention and sense of community

key metrics for success



%

conversion rate of casual riders to annual members

%

annual membership growth rate

$

customer lifetime value

$

revenue from annual memberships

questions?

appendix: data limitations

1. A large number of records had missing data for start and end station names and IDs. This hindered analysis of popular routes and station usage frequency.
2. A subset of the data had imprecise latitude and longitude data. These imprecise coordinates were associated with missing station information. This made it unreliable to input station information and could introduce bias into the analysis. These instances were filtered out, but this resulted in a loss of data.
3. The data contained a large number of extremely short rides and extremely long rides. These are likely due to system errors or glitches and required filtering. This again resulted in a loss of data.
4. There were instances where the same station ID was listed with different station names, and vice versa. Without an official list of Cyclistic's station names, IDs, and GPS locations, it was impossible to resolve these inconsistencies. This made geographic analysis unreliable.

appendix: data collection (BigQuery, Python)

	task	explanation
1	Downloaded data from source.	▷ Link to source. (Index of bucket “divvy-tripdata”)
2	Uploaded to Google Drive.	▷ Backup location in Case Study: Bikeshare folder.
3	Created BigQuery structures.	▷ Dataset: cyclistic_bikeshare; Table: bike_trip_data.
4	Uploaded CSVs to BigQuery.	▷ Appended each successive CSV to bike_trip_data. ▷ Encountered file size limit issue in BigQuery (100 MB) for CSVs 05-10.
5	Wrote Python function to split large CSVs.	▷ Link to Kaggle code. (Split CSVs Kaggle workbook)
6	Verified upload to BigQuery.	▷ Initial mismatch in row counts: 5,546,502 (BigQuery) vs. 5,783,100 (CSV total). ▷ Fixed Python function; re-split CSVs. ▷ Used running total of rows after each CSV upload to verify successful upload. ▷ Verification successful: 5,783,100 rows in BigQuery bike_trip_data table.

appendix: data cleaning (BigQuery)

	task	explanation
1	Missing value handling	▷ Missing values were identified in several columns. These were investigated to try to determine a cause of the missing data. ▷ Ultimately, the missing values were filtered out of the dataset to ensure accuracy of analysis that relies on accurate data. (More detail in R Markdown)
2	Removing duplicates	▷ Several duplicate ride_id values were found. Investigation found that each of the duplicates occurred twice, with each instance being identical except for the precision of the GPS data. ▷ Filtered out the duplicate instances with less precise GPS data to prevent skewed analysis.
3	Outlier handling	▷ Added a new column to calculate ride duration ▷ Checked for and removed instances of rides ≤0 sec, ≤15 seconds, ≥16 hours ▷ These are potential data errors and/or logical inconsistencies that could skew analysis
4	Data types	▷ Checked that all columns are the proper data type and that all values in the column are that same datatype ▷ Added new columns for weekday and month analysis
5	Addressing inconsistencies	▷ Checked that start/end times are only in 2024/2025 ▷ Trimmed whitespace from string values and checked for inconsistencies. ▷ Several inconsistencies were found, but I lack info (such as a full list of official Cyclistic station names and ids) in order to address these. Documented issues instead.

appendix: data analysis (RStudio)

	task	explanation
1	Organized data	▷ Using the tidyverse package, created new data frames to group the data by rider type, rider type and rideable type, rider type and hour of the day, rider type and day of the week, rider type and month
2	Analyzed total usage patterns	▷ Calculated summary statistics ▷ Using ggplot2, created box plot, bar chart visualizations
3	Analyzed bike type preferences	▷ Using ggplot2, created grouped bar chart visualization
4	Analyzed temporal patterns	▷ Using ggplot2, created histogram and line chart visualizations for hourly, weekly, monthly data, both for number of rides and average ride duration
5	Determined potential for further analysis	▷ Geographical analysis: ○ With more detailed GPS data could determine popular routes for applications in targeted marketing ▷ Further outlier analysis: ○ Could investigate long ride and short ride outliers further ○ Found unusual number of outliers (both short and long) at 4am, could investigate further

thank you for reading

presentation and analysis by

Austin Broadbent

Email, LinkedIn, Kaggle

CREDITS: This case study was made to complete the capstone module of the Google Data Analytics Professional Certificate. The template for this presentation was created by Slidesgo, including icons, slide layouts, and font choices. Stock images were downloaded from pexels.com.