Cyclistic

- Anar Seyf | anar.seyf@gmail.com

- Capstone project | Google Data Analytics course #8 | Coursera
- October 2021
- Source code: github.com/anarseyf/cyclistic

Introduction

This report is the capstone project for the Google Data Analytics professional certificate on Coursera [**1**]. The starting point is data from Cyclistic, a bike-sharing company which operates a network of classic and electric bicycles and docking stations in Chicago. The goal is to understand usage patterns between **Casual** riders (those who pay for single rides or day passes) and **Members** (those who purchased an annual membership), and to provide recommendations to the Cyclistic marketing team on how to convert casual riders into members.

Data

Our main dataset contains individual ride records over a twelve-month period—October 2020 to September 2021, inclusive—provided in monthly .csv files [**1**]. This amounts to over 5 million rows with these columns:

- unique ride ID;
- status (member or casual)
- start and end time;
- start and end docking station (ID and name);
- start and end latitude/longitude;
- bicycle type (classic, docked, electric);

The individual records are aggregated into hourly, daily, and weekly tables, and grouped by status (member vs casual). For example, the daily aggregates table consists of 730 entries (365 for Members and Casual riders each) with columns for average daily ride count and average duration. (See appendix on [**tools used**] and [**data cleanup steps**].)

Analysis

1. Members ride more often, casual users ride for longer.

An average ride is about **14 minutes** long for Members and **28 minutes** long for Casual riders. Members make an average of **7,479** rides per day, compared to **6,362** for Casual riders. (The number of distinct riders is unavailable in the data, so we can only refer to the overall groups. Daily totals can provide a rough approximation, but only within the order of magnitude.)

Ride volume is lowest in February and highest in July-August. Casual ride volume exceeds that of Members in the summer months only.

Table 1: Average daily number of rides

	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep
Casual	4,573	2,888	956	575	354	2,677	4,490	8,159	*12,147*	*14,044*	*13,131*	11,962
Member	*7,676*	*5,614*	*3,220*	*2,502*	*1,380*	*4,592*	*6,582*	*8,705*	11,754	12,058	12,431	*12,863*

Note:

In this table and Table 3, the larger of the two numbers for each month is in italics.

Ride duration is nearly flat for Members throughout the year; for Casual riders it increases by a few minutes in the summer months.

Table 2: Average ride duration (minutes)

	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep
Casual	24	24	22	21	33	27	28	29	28	27	25	24
Member	13	13	12	13	18	13	14	14	14	14	14	13

(February has an abnormal spike in ride duration. It may be partly attributable to low ride volume and thus higher variance in the data, but is otherwise not readily explainable here. It is likely to be weather-related—perhaps due to difficulties of reaching a docking station; see section 3 for more details.)

2. Members ride all week, casual riders prefer weekends.

Ride volume remains nearly flat for Members through the week, reducing only on Sundays. Casual users do the most riding over the weekend, peaking on Saturdays.

Table 3: Average daily number of rides

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Casual	5,047	4,763	4,871	5,104	6,421	*9,923*	*8,429*
Member	*7,107*	*7,701*	*8,002*	*7,807*	*7,666*	7,550	6,513

Table 4: Average ride duration (minutes)

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Casual	28	25	24	24	26	30	32
Member	13	13	13	13	13	15	16

Hourly heatmap This heatmap is a more detailed view of the data from *Table 3*. The Monday–Friday pattern for Members closely matches standard working hours; note especially the peak around 4–5pm. Weekend patterns are much closer between the two groups, apart from volume; note, for example, Saturday evening activity spilling over into early morning on Sunday.

3. Seasons affect ride volume, but average duration remains stable

This section uses the Chicago daily weather dataset from NOAA; see [**Links**] for data source.

The following chart plots one year of ride data, aggregated by week, against weather for that week: average daily air temperature, cumulative rain, and cumulative snow.

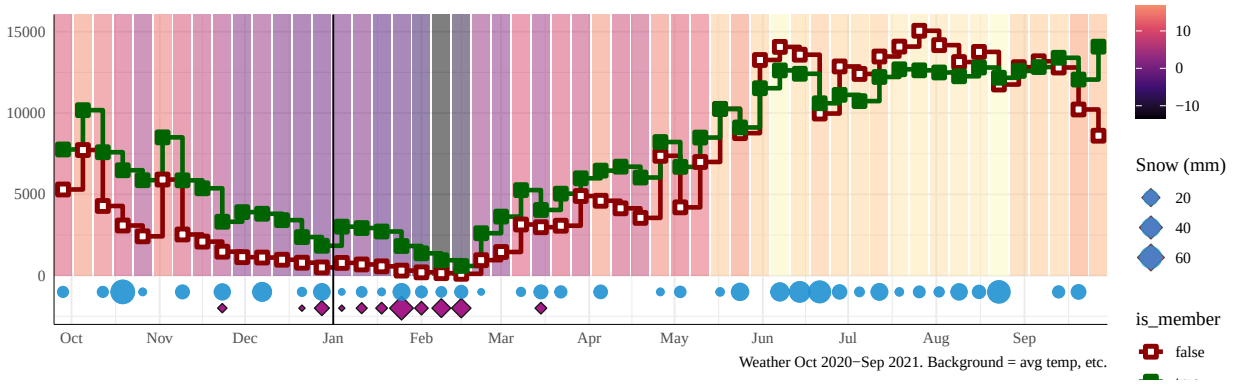
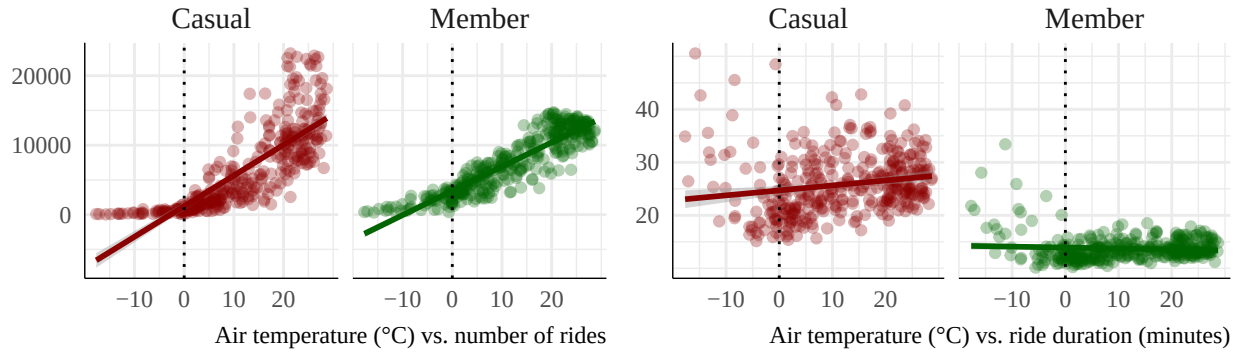


Table 5: Correlation (r): ride volume vs. weather factors

	Count				Duration			
	Temperature	Rain	Snow	Wind	Temperature	Rain	Snow	Wind
Casual	**0.81**	0.00	-0.20	-0.20	**0.07**	-0.07	0.16	-0.05
Member	**0.91**	-0.01	-0.29	-0.22	**-0.08**	-0.07	0.29	-0.08

> The Pearson correlation coefficient (r) measures how strongly two sets of data are linearly correlated, from -1 (perfect negative) to 1 (perfect positive), with 0 indicating no correlation. Here, for example, ride count is negatively correlated with snow: the more snow, the fewer rides.

Table 5 shows that the number of rides on a given day is strongly—almost perfectly—correlated with **air temperature** for both groups; duration, on the other hand, shows almost no correlation. The following plots help illustrate this relationship. Each circle represents one day; 365 circles for each plot; the diagonal lines represent best linear fit and confidence intervals.



Rain is more ambiguous: correlation is close to 0. Does this mean bike-share users in Chicago are indifferent to rain? The more plausible explanation is that we are at the limits of this dataset. In the 12-month period, 256 days had no recorded rain, and only 16 days had half an inch or more; on those days, Members rode for

4% less total time, and Casual riders 6% more, than on days with less rain. In other words, it did not rain enough to make the relationship clear.

These distributions confirm an observation from section 1: average ride duration remains nearly flat (more so for Members), but ride volume can vary significantly.

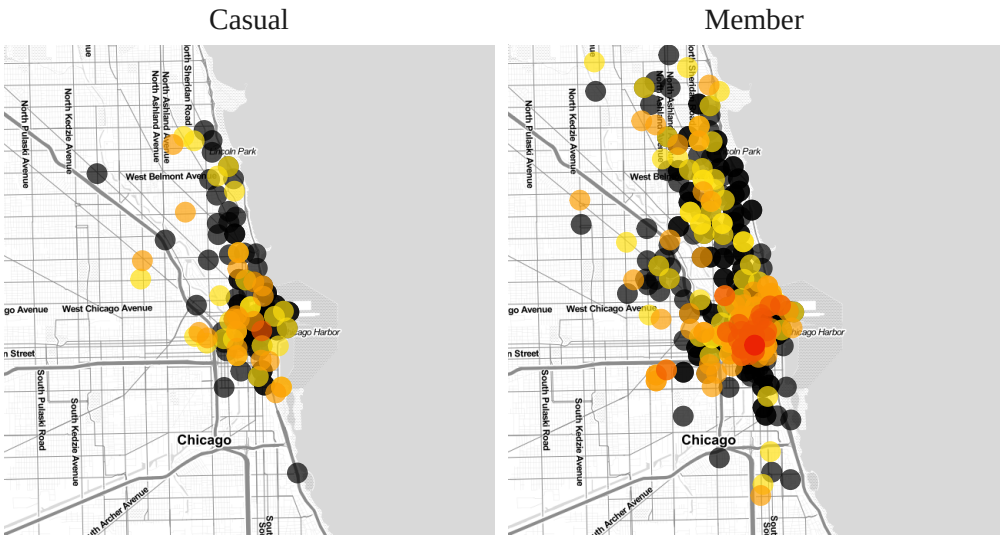
4. Geographical patterns are distinct across the week

The source dataset contains 1288 distinct docking stations. 85% of all rides have both a start and an end station specified; in this section we focus on that subset only (the other 15% start and/or end outside of a docking station). To prevent seasonal trends from complicating the picture, the data is further filtered to four weeks in the summer (June 1st—June 28).

The following plots attempt to isolate daily and weekly patterns using a heatmap of station usage. Each dot represents a bike docking station. The metric is: *count(rides ended) – count(rides started)* at a given station, within one hour. Dark dots indicate more starts than ends, and thus an *outflow* of bike traffic; colored dots are stations where more rides have ended than started, or an *inflow* of traffic (redder colors indicate more arrivals, on a log scale).

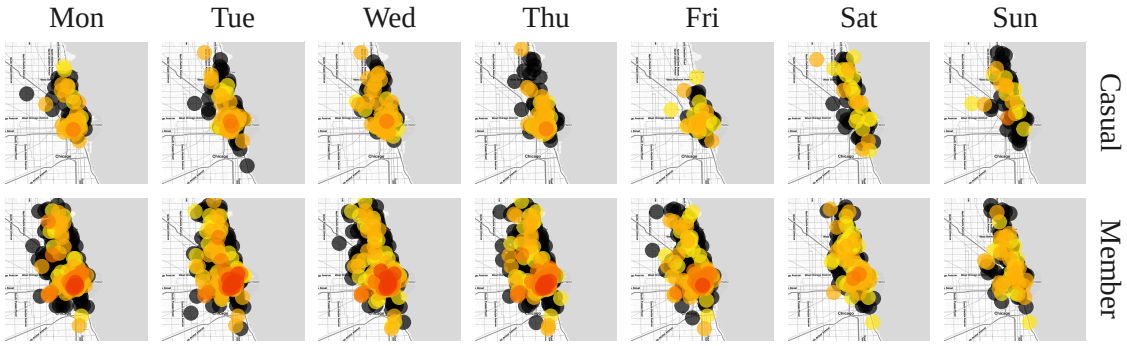
In other words: *clusters of "hot" stations indicate areas that people are cycling into.*

Friday morning (8-10am)



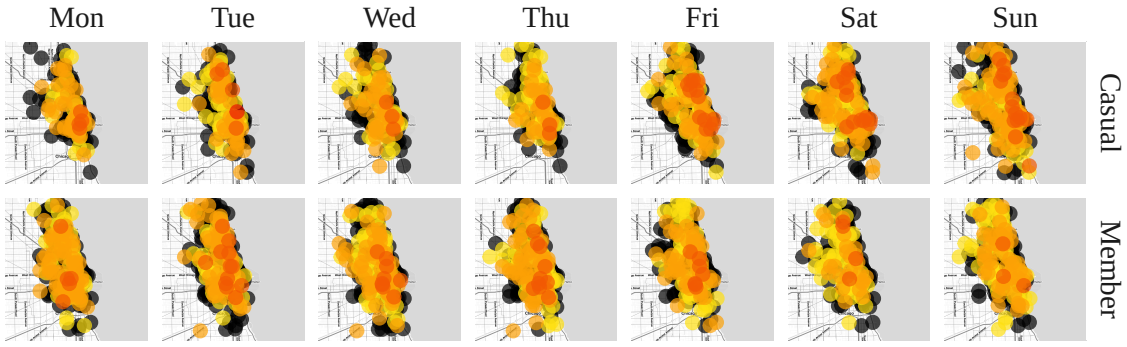
Let's look at the same pattern across the week (again, these are totals across a four-week time window).

Morning (8-9am)



A distinct weekday pattern emerges. Apart from volume, the "fingerprints" look similar for the two groups. During the work week the flow is into the center of the city; on weekends the flow is more dispersed.

Afternoon (4-5pm)



In the afternoon the volume of riding spikes (this was also clearly visible in the heatmap in section 2). Key takeaway: patterns are distinct by time of day and weekday vs weekend, but less so between C and M.

Conclusions

1. Casual riders use the service as more of a luxury, and members as a utility.

- Weekend spike
- Ride duration: starting a ride is a bigger commitment for a casual user. A 15-minute ride is a non-issue for a member, but not if you are paying for a single pass.
- The distinctions are not due to any inherent difference between rider types, but due to the convenience factor. (Running errands, commuting.) Barrier to entry is higher for C than M.

2. The two groups' usage patterns overlap.

The groups are not inherently distinct in their goals. The patterns likely differ as a result of . . .

- Similar geographic patterns, hourly patterns, weekday vs weekend
- Commuting, daily tasks?
- Similarities M vs C? Modes (work hours vs weekend)
- Lower variance for M than C. More consistent patterns.

Recommendations

How to convert casual riders into members?

1. Nudge casual riders toward memberships with free trials

- Turn a single pass into a free weekly or monthly trial.
- Have they had a chance or reason to consider Cyclistic as a realistic, always-available mode of transport?
- The pool of casual riders is the first potential market for memberships.
- This pool, however, is potentially smaller than the existing Membership base.

2. Address specific modes of riding

- Suggest bikes as a mode of work commute
- Run promotions on weekends with major sporting events where cycling would help avoid traffic and parking problems

3. ~~People outside the system. Never tried the service (or cycling regularly).~~

- Why? Does their neighborhood lack docking stations? Underserved areas? Too far to work?
- Do they need help with route planning?
- Work with the city on expanding cycling routes and making cycling safer.
- Weather? Seasons?

Appendix

Limitations

The available dataset is limited. It does not, for example, contain any membership data, so seasonal trends could be partly attributable to membership growth, etc.

Out of scope

Out of scope:

- bike types;
- other factors (natural disasters, sporting events, etc.);
- pricing;
- individual user profiles
- trends (year-over-year, electric, geography, etc.)
- other modes of transport
- region specifics

Data cleanup

February: 16-Feb-2021 had a few rides of close to 24 hours. Overall, most long rides (12+ hours) are attributed to Casual riders, which likely corresponds to 24-hour rentals. Not enough info to interpret this further. Test data; NOTES FROM GOOGLE DOC

Presentation: log scale.

Links

Data sources (incl. Google Maps, Stamen).

1. Coursera — Google Data Analytics
    - https://www.coursera.org/professional-certificates/google-data-analytics
2. BigQuery
    - https://cloud.google.com/bigquery
3. Dataset
    - https://divvy-tripdata.s3.amazonaws.com/index.html
4. NOAA Chicago daily weather. Collected at O'Hare Airport.
    - https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USW00094846/detail
5. Pearson correlation coefficient
    - https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

Tools

- Google Sheets
- BigQuery
- R (in VSCode)
- knitr (PDF) (links TODO)