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*

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 with available data. It seems likely to be weather-related as well—perhaps due to difficulties of reaching a docking station in snow; 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

Members' weekday pattern closely matches standard working hours (note especially the peak around 4-5pm). Weekend pattern much closer between the two groups, with volume peaking on Saturday afternoon.

Ride count heatmap (time of day on vertical axis)
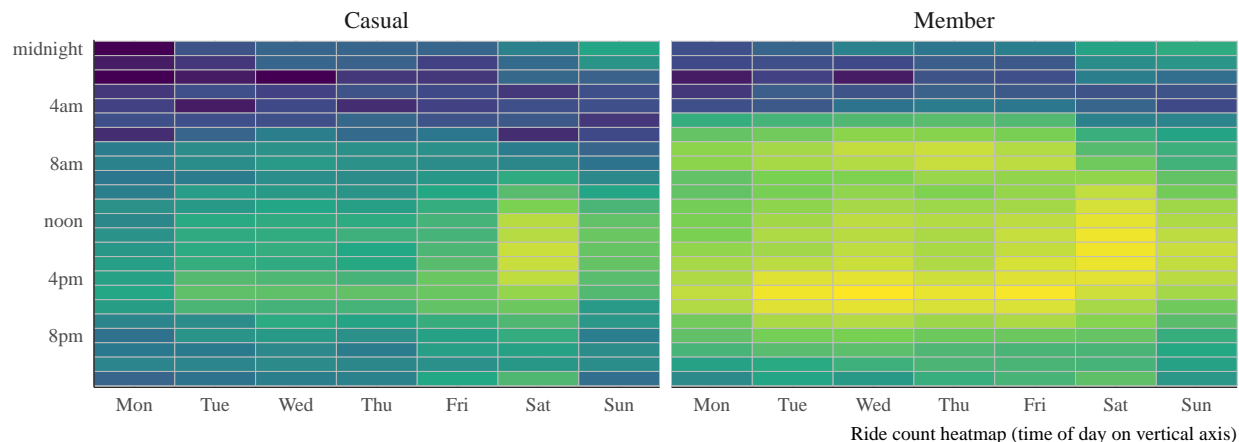
3. Ride counts and Weather

Chicago daily weather dataset from NOAA [**1**] for the same time period.

Average daily ride count and Chicago weather, Oct'20—Sep'21, by week. Background color is average air temperature that week.

The Pearson correlation factor, or **r**, (**TODO link**) is a normalized measure of how strongly two parameters are related. It ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no correlation.

At a high level, the coefficients point in the direction we'd expect. For example, ride count is negatively correlated with snow: the more snow, the fewer rides.

Ride count is strongly correlated with **air temperature** for both groups. This is clear enough from Fig.X(TODO).

Rain is more ambiguous. **r** is close to 0. Does this mean bike riders in Chicago are not influenced by rain at all? It is more ambiguous than that. In the 12-month period 256 days have no recorded rain, only 16 days with 1/2 inch or more daily total, and no days with 2 inches. On days with at least 1/2 inch of rain, Members rode 4% less total time, and Casual users rode 6% more. It simply did not rain enough to make this relationship clear; we are at the limits of this data.

The relationship with **Snow** is clearer, and snow itself is obviously not independent from temperature.

These distributions confirm the basic observation from section 1: average ride duration remains nearly the same (more so for Members), but the ride volume is highly influenced by environment.

(Note: one outlier for Casual riders, 16 Feb 2021, was removed from this plot as it had an average duration of over 100 minutes. More on this in the appendix.)

4. Geographical

(TODO)

Stations: 1288 distinct stations (saved in table). 85% of trips (4293771 out of 5051830) have both start and end station ID — we focus on those when mapping. These are split almost evenly between M (54.7%) and C.

Stations: 4007812 trips end at a different station from start; 285959 trips start and end at the same station.

Stations

TODO — # of rides with start & end station (85%). **1288** stations. Following: **casual vs members**, **8am vs 4pm**, **Monday-Friday**.

Each dot is a bike docking station. Stations highlighed **yellow-red** have more arrivals than departures for the given hour, suggesting an influx of bike traffic at that location.

TODO: Heatmap: instead of each weekday, take a month out of each quarter?

Friday morning (8-10am)

Let's look at the same pattern across the entire week.

Morning (8-9am)

A distinct weekday pattern. Apart from volume, the "fingerprints" look similar for the two groups. On weekends it looks different: there isn't a lot of casual riding into this area in the morning.

Afternoon (4-5pm)

More dispersion away from center (TODO). 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

TODO. 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.

TODO. 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

A. 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.

B. 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

C. Tools

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

D. Full data (weekly?)