Cyclistic

- Anar Seyf (anar.seyf@gmail.com)

- October 2021
- Capstone project | Google Data Analytics course #8 | Coursera

Analysis

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

An average ride is about **28 minutes** long for casual users, **14 minutes** long for members. Members make an average of **7,479** rides per day, casual users **6,362** rides per day. (The number of distinct riders is unavailable in the data, so we can only refer to the overall population. **TODO** — so "total volume" instead of "more often"?)

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

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

(The month of February shows an abnormal peak in ride duration. It may be partially attributable to low ride volume and thus higher variance in the data, but is otherwise not readily explained here.)

Table 1: 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

Table 2: Average daily rides (count)

Status	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep
Casual	4573	2888	956	575	354	2677	4490	8159	**12147**	**14044**	**13131**	11962
Member	**7676**	**5614**	**3220**	**2502**	**1380**	**4592**	**6582**	**8705**	11754	12058	12431	**12863**

TODO — kable sparklines?

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

Members' usage remains nearly flat, falling on Sunday. Casual users do most riding starting on Friday and through the weekend, peaking on Saturday.

By weekday

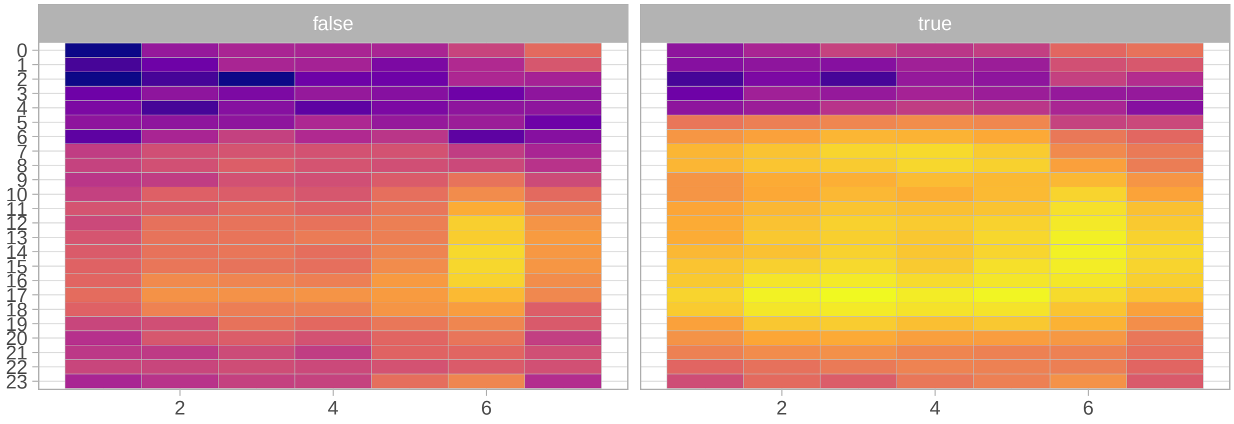
Table 3: Average daily duration (minutes)

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

Table 4: Average daily rides (count)

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Casual	5047	4763	4871	5104	6421	**9923**	**8429**
Member	**7107**	**7701**	**8002**	**7807**	**7666**	7550	6513

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

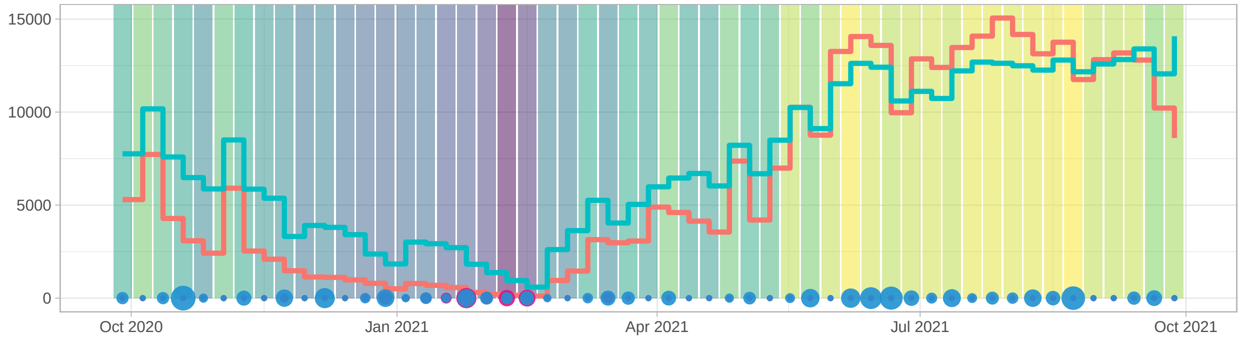


Ride count heatmap (time of day on vertical axis)

3. Seasonal and Weather

TODO — eliminate 0 rain/snow circles.

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



Weather Oct 2020−Sep 2021. Background = avg temp, etc.

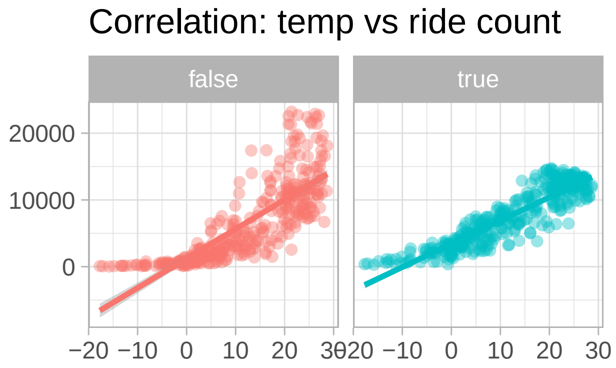
Table 5: Correlation (r): weather and ride count

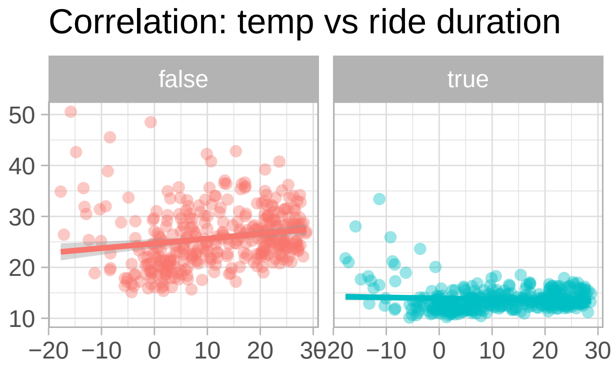
Status	Temperature	Rain	Snow	Wind
Casual	0.81	0.00	-0.20	-0.20
Member	0.91	-0.01	-0.29	-0.22

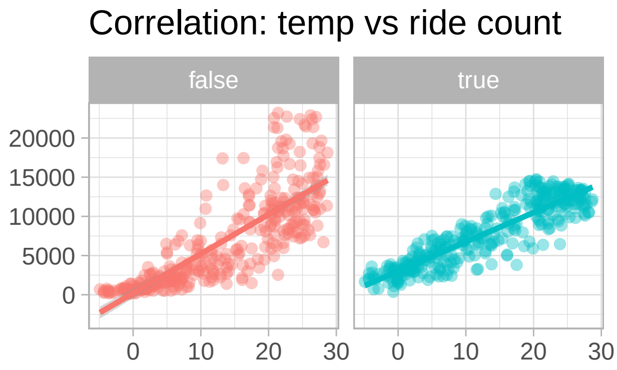
The Pearson correlation factor, or **r**, (**TODO link**) is a normalized measure of how strongly two parameters are related. It ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no correlation.

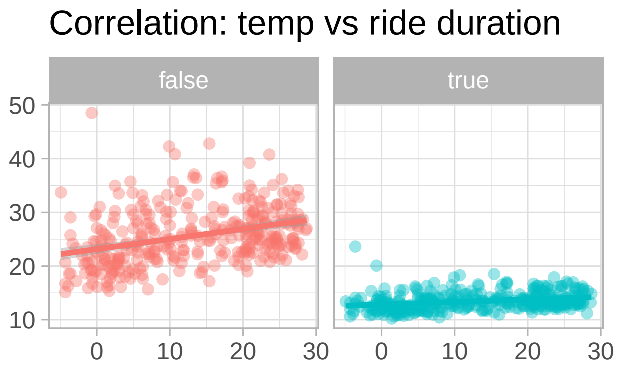
At a high level, the coefficients point in the direction we'd expect. For example, ride count is negatively correlated with snow: the more snow, the fewer rides.

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


Correlation: temp vs ride count


Correlation: temp vs ride duration


Correlation: temp vs ride count


Correlation: temp vs ride duration

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 afternoon (4-5pm) 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).

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.)

1. The two groups usage patterns overlap.

- Weekdays: same geographic patterns,
-
- 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?

Recommendations:

1. 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?

2. The pool of casual riders is the first potential market for memberships.

- Address specific modes of riding (work vs weekend)

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

- Why? Does their neighborhood lack docking stations?
- Do they need help with route planning?
- Suggest bikes as a mode of work commute
- Check against areas underserved by public transport, consider expanding service in those areas
- Make riding easier in colder months?
- Weather-protected bikes?

Appendix

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.

Presentation: log scale. 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

B. Links

Data sources (incl. Google Maps, Stamen).

C. Tools used (and not used)

D. Full data (weekly?)

Unused graphs

Monthly by weekday

Goal

The objective of the report is to show differences between **Members** (those who purchased an annual membership) and **Casual** riders (those who pay for single rides or daily passes). Every part of the analysis shows (**TODO - synonym**) those categories side-by-side.

Data

The **main data source** of the report is a 12-month window (October 2020 to September 2021, inclusive) from the dataset provided by Divvy **(link TODO)**. (See data cleaning notes in the Appendix.) The data consists of over 5 million unique ride records with these columns:

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

The **secondary data source** is the Chicago daily weather dataset from NOAA **(link TODO)** for the same time period. This seems highly pertinent background given the nature of the main data.

The 5 million individual records are condensed into hourly, daily, weekly, and monthly **aggregates**, grouped by status (member vs casual). For example, the daily aggregates table consists of 730 entries (365 for Members and Casual users each) with columns for ride count and average duration.

Throughout the report, more emphasis is given to *ride count*, as the main indicator of activity volume, than to ride *ride duration.*

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

-TODO