Converting casual daily bike-share riders into annual members

Kwan Jeon
August 2022

Contents

1) Statement of business task
2) Description of all data sources and tools used
3) Steps taken to clean or manipulate data
4) Summary of analysis
5) Supporting visualizations and key findings
6) Top three recommendations based on analysis
7) Appendix with additional supporting information

Statement of business task

BACKGROUND

For this project, a fictional company, Cyclistic, was created. Anonymized trip data from a real bike-share company, Divvy in Chicago, was then applied to analyze data and provide recommendations.

RECENT OBSERVATIONS

Analysis from a separate team has shown that casual riders utilizing single ride passes or day passes are less profitable than members holding annual passes.

IMMEDIATE OBJECTIVE

Analyze trip data from the past 12 months to determine how Casual riders and annual Members use Cyclistic bikes differently in order to convert Casual riders to Members.

Data sources and tools used

The Data

- Public and anonymized trip information from a real bicycle ride-share company
  - Member or casual rider designation
  - Start and End trip location
  - Start and End trip timestamps
- August 2021 to July 2022
- 1.0 GB of data
- 5.9 million trips

The Tools

Google BigQuery

SQL for data cleaning and manipulation

Google Cloud Storage

Importing and exporting large data files for BigQuery

RStudio

R programming for data analysis and data viz

tableau

Additional data visualization

Google Sheets

Basic tables for use in final report

Google Slides

Final report

GitHub

Website to host final report

Cleaning and Manipulation of data

Manipulating

1) Combine all 12 monthly data tables into one table
2) Create and calculate trip duration field
3) Create and calculate trip distance field
4) Create new fields containing trip start day of week (number and name)
5) Create new fields containing trip start month (number and name)
6) Create new coordinates field by concatenating latitude and longitude (used for unique lookup in filling missing station names and IDs)

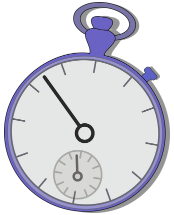
Cleaning

1) Review all newly populated day of week names and numbers as well as month names and numbers to ensure they're valid values
2) Remove trip data containing outliers
   a) Durations < 0 mins (96 trips)
   b) Durations > 24 hrs (4,965 trips)

Google BigQuery

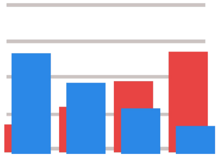
Note: All steps above performed in BigQuery since files were too large to open in Google Sheets. Microsoft Excel not used due to unavailability, but extensive personal experience suggests files >20 MB exceed the practical working limits of Excel (10 of 12 monthly data files were >20 MB).

Summary of Analysis

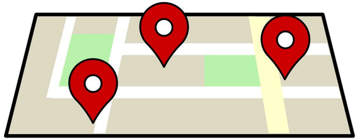


Trip start and end times used to calculate duration of each trip

Station start and end coordinates used to calculate distance of each trip

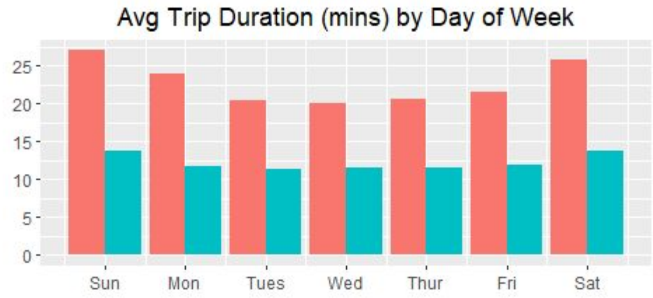


Trip start times and count used to analyze riding patterns by month and day of week

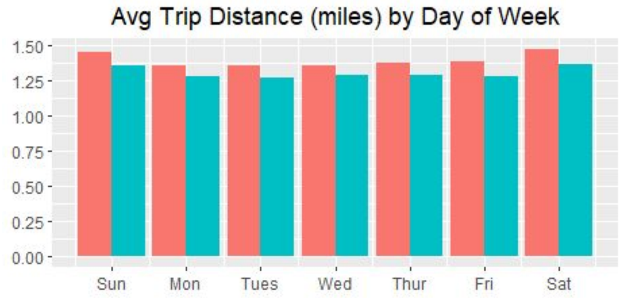


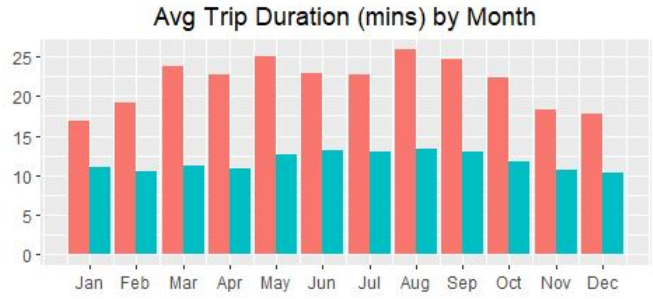
Station coordinates and trip count used to visually map most popular start and end locations

Supporting visualizations and key findings


Avg Trip Duration (mins) by Day of Week

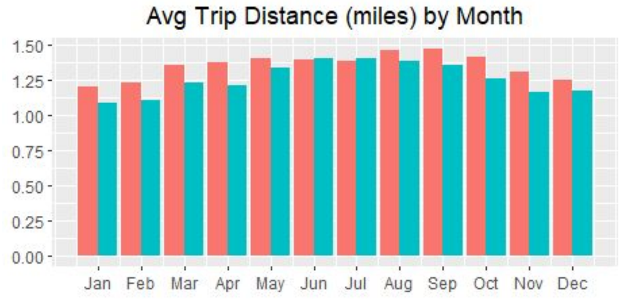
Casual riders took nearly double the amount of time to travel nearly the same distance of Members


Avg Trip Distance (miles) by Day of Week

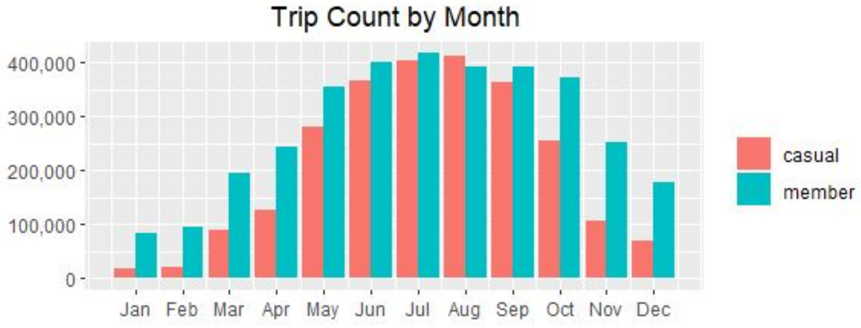

Avg Trip Duration (mins) by Month

Avg Trip	Casual	Member
Duration	23 mins	12 mins
Distance	1.4 miles	1.3 miles

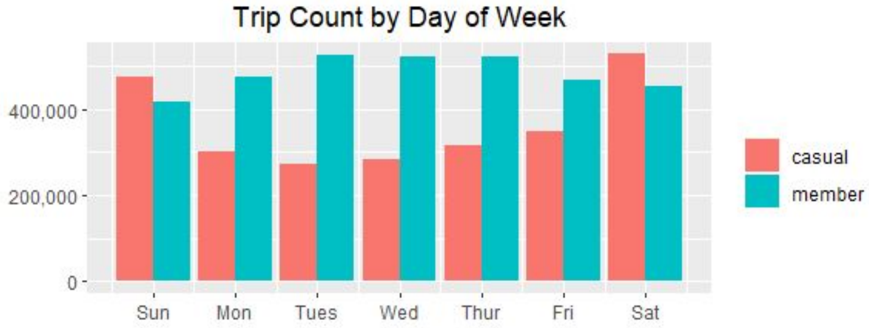
🟥 casual 🟦 member


Avg Trip Distance (miles) by Month

Supporting visualizations and key findings



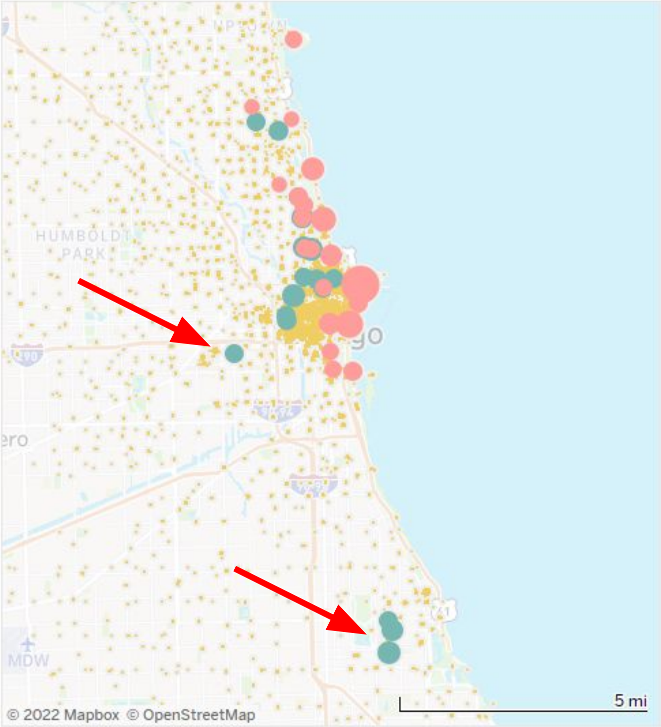
There were 34% more trips taken by Members than Casual riders over the 12 month observation period with warmer months being more popular for both groups



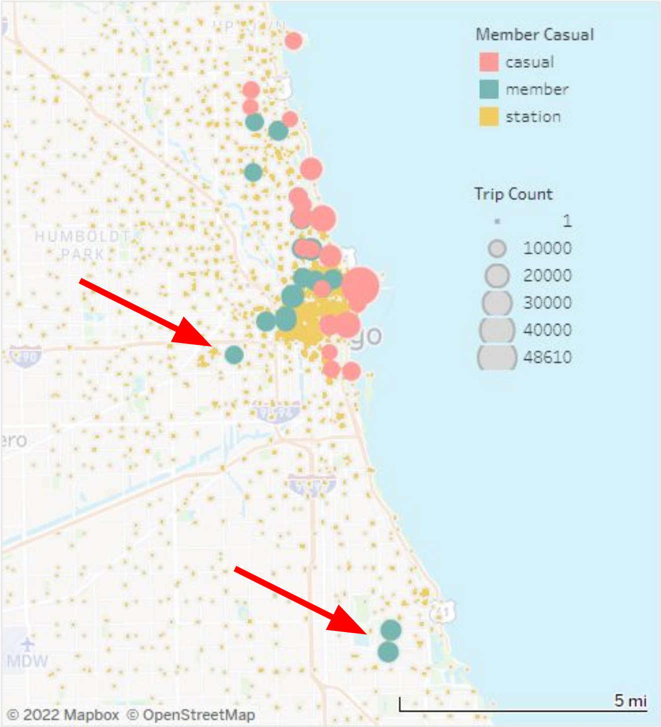
Members took more trips during the work week while Casual ridership surged on weekends

Supporting visualizations and key findings

Top 20 Start Stations by Trip Count



Top 20 End Stations by Trip Count



The top 20 starting and end stations were more geographically dispersed for Members than Casual riders, with a larger North-South and East-West spread

Top 3 Recommendations

AVAILABILITY

Obtain more data to determine if the most popular stations for Members are running out of bikes -- Casual riders take out bikes for nearly double the amount of time as Members so stations running out of bikes may discourage new memberships

COMFORT

Consider adding front wind guards, heated grips or heated seats to bicycles -- ridership declined much more significantly in colder months for Casual riders versus Members so warmer trips may convince Casual riders into more year-round use and eventual membership

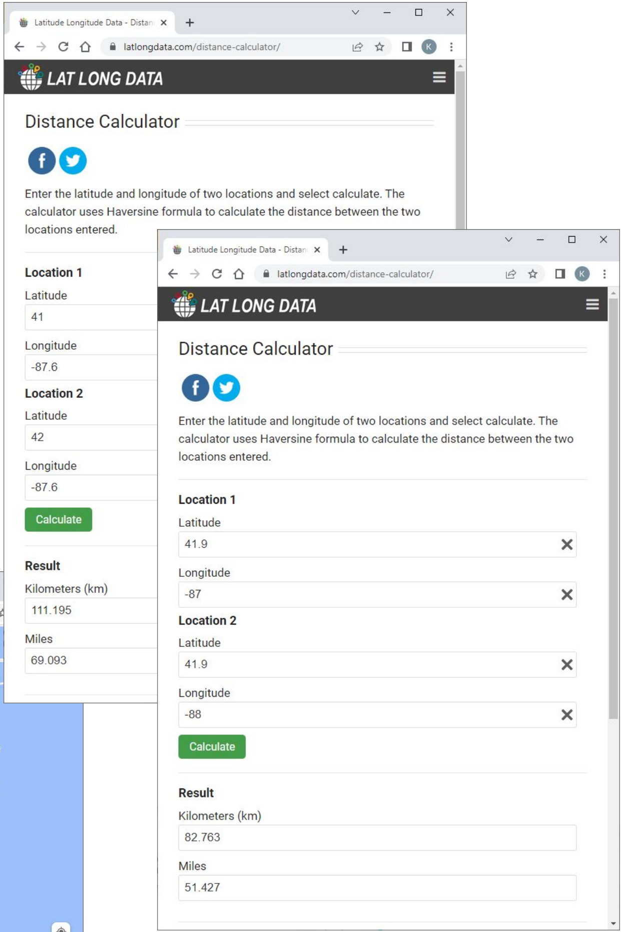
CONVENIENCE

Increase station density outside of the main city center (either by adding new stations or shifting stations from more dense to less dense areas) -- Average trip distances are less than 1.5 miles so having easy access to stations could encourage riding instead of walking

APPENDIX

Data sources used

- Downloaded public monthly data from August 2021 to July 2022 onto local drive from **https://divvy-tripdata.s3.amazonaws.com/index.html**
- Used **divvybikes.com** to compare cost of Single ride vs Day pass for 24 hour period (used to rationalize removing trips >24 hr durations as outliers)
- Used **latlongdata.com/distance-calculator/** to calculate miles in 1 degree latitude and 1 degree longitude increments near Chicago (used to calculate straight-line distance of each trip)


Latitude Longitude Data - Distan
latlongdata.com/distance-calculator/
LAT LONG DATA

Distance Calculator

Enter the latitude and longitude of two locations and select calculate. The calculator uses Haversine formula to calculate the distance between the two locations entered.

Location 1
Latitude: 41
Longitude: -87.6
Location 2
Latitude: 42
Longitude: -87.6

Calculate

Result
Kilometers (km): 111.195
Miles: 69.093

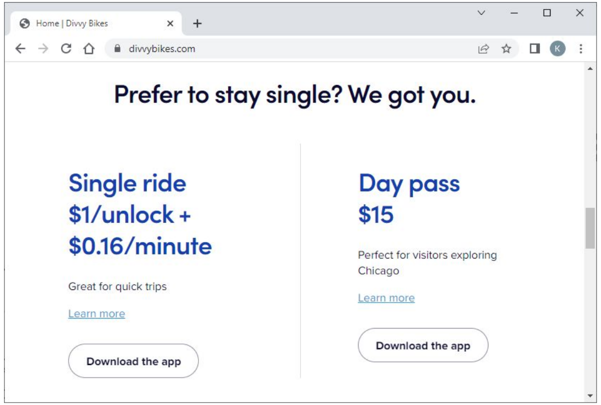
Distance Calculator

Enter the latitude and longitude of two locations and select calculate. The calculator uses Haversine formula to calculate the distance between the two locations entered.

Location 1
Latitude: 41.9
Longitude: -87
Location 2
Latitude: 41.9
Longitude: -88

Calculate

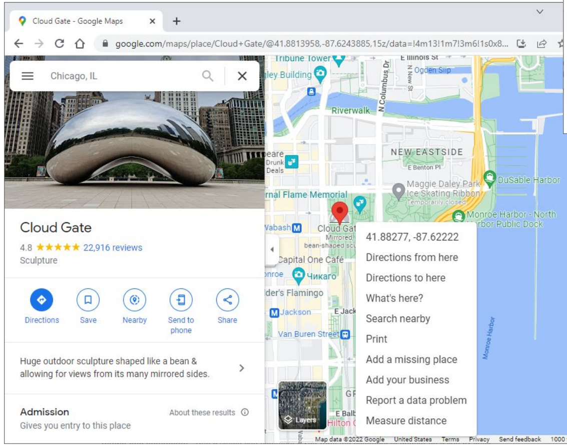
Result
Kilometers (km): 82.763
Miles: 51.427



Prefer to stay single? We got you.

Single ride
$1/unlock + $0.16/minute
Great for quick trips
Learn more
Download the app

Day pass
$15
Perfect for visitors exploring Chicago
Learn more
Download the app



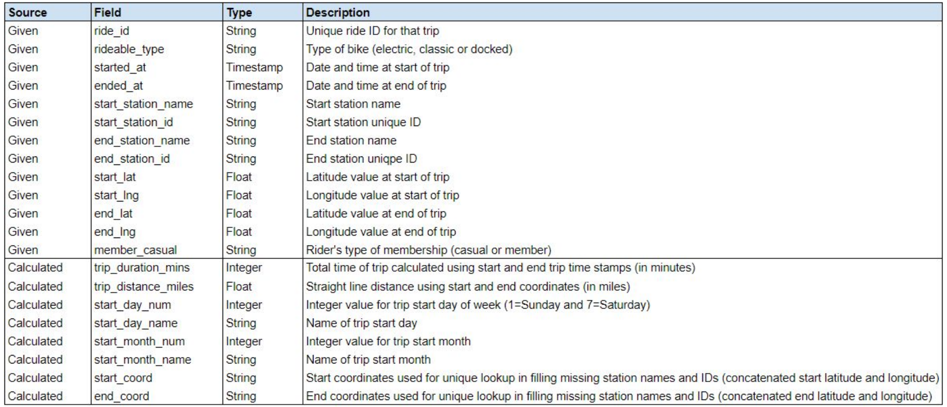
Huge outdoor sculpture shaped like a bean & allowing for views from its many mirrored sides.

Gives you entry to this place

41.88277, -87.62222
Directions from here
Directions to here
What's here?
Search nearby
Print
Add a missing place
Add your business
Report a data problem
Measure distance

Cleaning and Manipulation of data details

1) Downloaded 12 monthly .csv files from website onto local drive (https://divvy-tripdata.s3.amazonaws.com/index.html)
2) Uploaded monthly data files to Google Cloud storage account (bypasses BigQuery's 100MB file size upload limit from local drive)
3) Created 12 monthly tables in BigQuery by importing files from Cloud Storage
4) Appended 12 monthly tables into one combined table in BigQuery (5,901,463 trips)
5) Created 8 new fields in combined table



Source	Field	Type	Description
Given	ride_id	String	Unique ride ID for that trip
Given	rideable_type	String	Type of bike (electric, classic or docked)
Given	started_at	Timestamp	Date and time at start of trip
Given	ended_at	Timestamp	Date and time at end of trip
Given	start_station_name	String	Start station name
Given	start_station_id	String	Start station unique ID
Given	end_station_name	String	End station name
Given	end_station_id	String	End station uniqpe ID
Given	start_lat	Float	Latitude value at start of trip
Given	start_lng	Float	Longitude value at start of trip
Given	end_lat	Float	Latitude value at end of trip
Given	end_lng	Float	Longitude value at end of trip
Given	member_casual	String	Rider's type of membership (casual or member)
Calculated	trip_duration_mins	Integer	Total time of trip calculated using start and end trip time stamps (in minutes)
Calculated	trip_distance_miles	Float	Straight line distance using start and end coordinates (in miles)
Calculated	start_day_num	Integer	Integer value for trip start day of week (1=Sunday and 7=Saturday)
Calculated	start_day_name	String	Name of trip start day
Calculated	start_month_num	Integer	Integer value for trip start month
Calculated	start_month_name	String	Name of trip start month
Calculated	start_coord	String	Start coordinates used for unique lookup in filling missing station names and IDs (concatenated start latitude and longitude)
Calculated	end_coord	String	End coordinates used for unique lookup in filling missing station names and IDs (concatenated end latitude and longitude)

6) Populated new trip_duration and trip_distance fields
7) Populated new start day and month fields
8) Populated new start and end coordinates
9) Removed trips that had negative duration values (96) and negative distance values (0 trips)
10) Removed trips whose durations were > 24 hrs (4,965 trips)
    a) Only 3 types of passes offered (Single, Day or Annual)
    b) Day passes expire in 24 hrs and using single pass for 24 hrs costs significantly more than Day pass ($241 vs $15) so likely rider mistake
11) Attempted to use concatenated latitude and longitude of known station names and IDs as unique lookup to fill missing station names and IDs, but there were no matching lookup values, likely due to inconsistent coordinate values
    a) Some station had hundreds of unique coordinate values for the same station ID
    b) 129,943 trips had missing station names or IDs
12) Exported final trip data table from BigQuery to Cloud Storage as .csv files (large table size required splitting into multiple files) for import into RStudio for statistical analysis and data visualization