

## Ask

1. What is the problem we are trying to solve?
2. We want to increase the number of annual subscription based users.

Task: Understand the difference in the way members and casual rides use the service.

## Prepare

1. Data was obtained from : <https://divvy-tripdata.s3.amazonaws.com/index.html> for 2023
2. Original data was provided in .csv file. It was provided by Motivate International Inc.
3. The source of the data suggests that it follows ROCCC.
4. Sorted data for casual and member riders

## Process

1. Formatting data to make sure the formats are consistent and interpretable for plotting
2. Added a column for length of rides in "ride\_length" column
3. Removed columns for station names and Id as they are not useful for this project
4. Converted latitude and longitude related columns to numbers after resolving errors with format
5. Remove rows where start time is later than end time since it is error in logging or recording data.
6. Since ride\_id is just a unique key for identifying different rides, we can remove duplicates based on it.
7. No duplicates were found!
8. day\_of\_week column : Numbers 1 (Monday) through 7 (Sunday).
9. Added a column for length of rides in "ride\_length" column using Haversine formula:  $6371 * \text{ACOS}(\text{COS}(\text{RADIANS}(\text{lat2} - \text{lat1})) * \text{COS}(\text{RADIANS}(\text{lon2} - \text{lon1})) + \text{COS}(\text{RADIANS}(\text{lat2})) * \text{COS}(\text{RADIANS}(\text{lat1})))$
10. Removing 0 end\_lng rows since it has to be an error value. Sheet 6
11. Ride time is calculated in minutes
12. Removed negative ride times.

$\text{SIN}(\text{Lat2}) * \text{COS}(\text{RADIANS}(\text{Lat1})) * \text{COS}(\text{RADIANS}(\text{Long2}) - \text{RADIANS}(\text{Long1})) + \text{SIN}(\text{RADIANS}(\text{Lat2})) * \text{SIN}(|$

RADIANS(Lat1)))