Appendix

Cyclistic Membership

Created by: Jonathan Kalmar
Last updated: September 29th, 2021

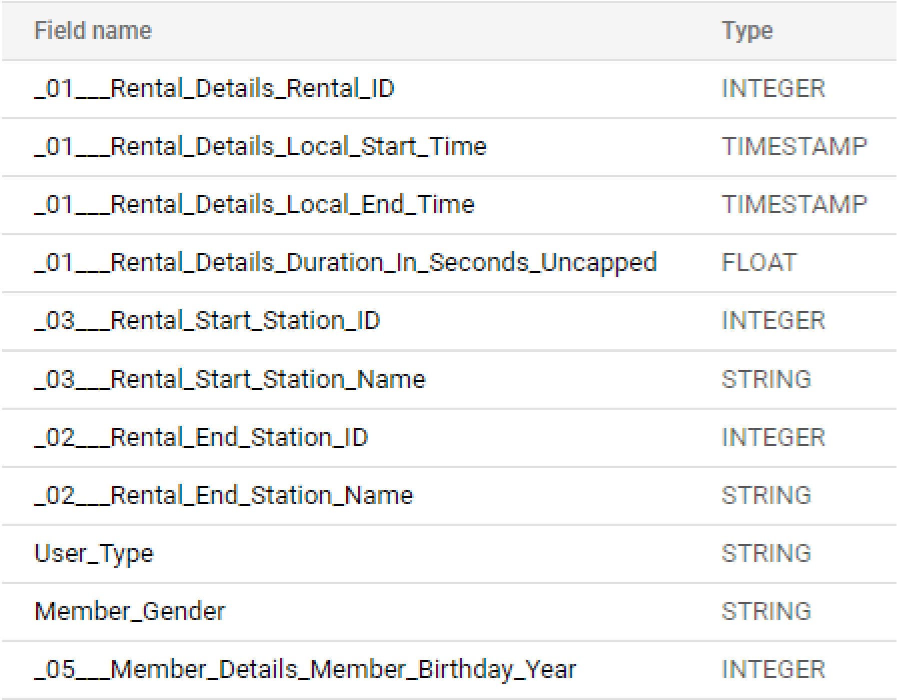
Objective

Discover how annual members and casual riders use Cyclistic bikes differently. This will give us valuable insight into how casual members may be motivated to upgrade to a full membership, and how we can use social media to increase membership among riders.

Raw Data

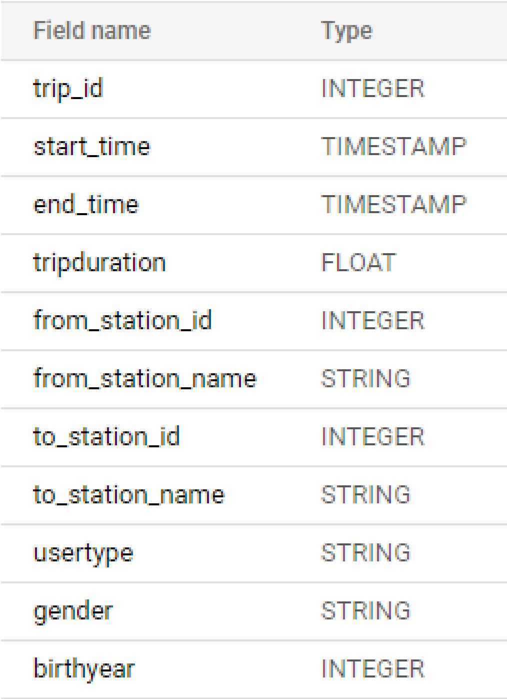
The data investigated was provided by Cyclistics containing trip information over a 12 month period beginning on April 1st, 2019. Four separate csv files were provided with the following schemas:

Divvy_Trips_2019_Q2



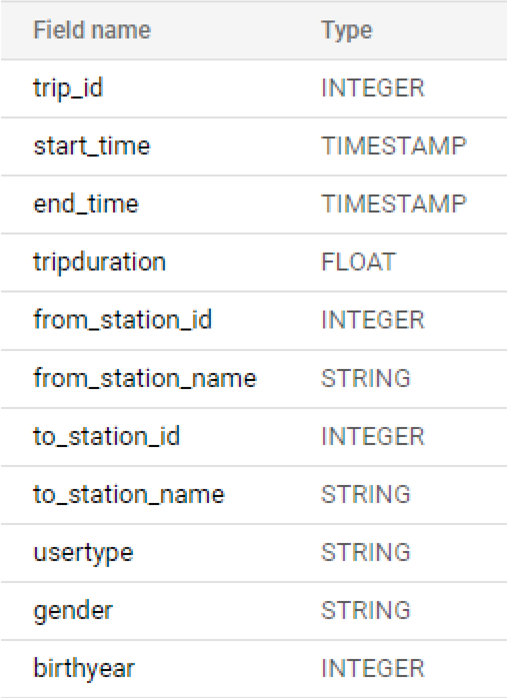
Field name	Type
_01___Rental_Details_Rental_ID	INTEGER
_01___Rental_Details_Local_Start_Time	TIMESTAMP
_01___Rental_Details_Local_End_Time	TIMESTAMP
_01___Rental_Details_Duration_In_Seconds_Uncapped	FLOAT
_03___Rental_Start_Station_ID	INTEGER
_03___Rental_Start_Station_Name	STRING
_02___Rental_End_Station_ID	INTEGER
_02___Rental_End_Station_Name	STRING
User_Type	STRING
Member_Gender	STRING
_05___Member_Details_Member_Birthday_Year	INTEGER

Divvy_Trips_2019_Q3



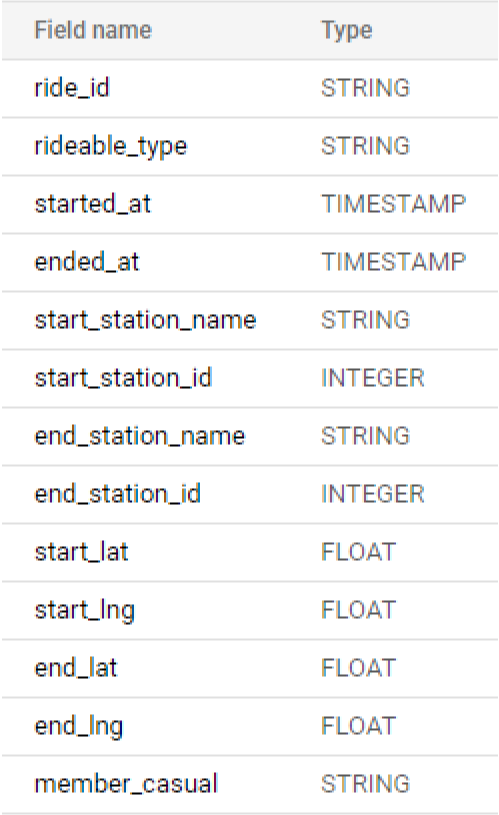
Field name	Type
trip_id	INTEGER
start_time	TIMESTAMP
end_time	TIMESTAMP
tripduration	FLOAT
from_station_id	INTEGER
from_station_name	STRING
to_station_id	INTEGER
to_station_name	STRING
usertype	STRING
gender	STRING
birthyear	INTEGER

Divvy_Trips_2019_Q4



Field name	Type
trip_id	INTEGER
start_time	TIMESTAMP
end_time	TIMESTAMP
tripduration	FLOAT
from_station_id	INTEGER
from_station_name	STRING
to_station_id	INTEGER
to_station_name	STRING
usertype	STRING
gender	STRING
birthyear	INTEGER

Divvy_Trips_2020_Q1



Field name	Type
ride_id	STRING
rideable_type	STRING
started_at	TIMESTAMP
ended_at	TIMESTAMP
start_station_name	STRING
start_station_id	INTEGER
end_station_name	STRING
end_station_id	INTEGER
start_lat	FLOAT
start_lng	FLOAT
end_lat	FLOAT
end_lng	FLOAT
member_casual	STRING

The dataset is reliable since it was recorded by Cyclistic and includes every single ride transaction. It is the original data source with the raw data from each trip. The data is comprehensive with clearly labeled data values and complete entries for each observation. The data is cited since we know it was given to us by Cyclistic

Cleaning Process

The files given were extremely large, with a total file size of over 500MB and over 3 million table entries. As such, SQL was used to examine, filter, and combine the data into a single table for analysis.

The following query was used to combine all four tables into one table with common variable names and consistent values. The result was stored into a new table called 'cleaned_data.' Note that there is no gender or birthyear data for 2020 Q1, and 'rideable_type' data from 2020 Q1 was excluded because there was only a single value observed across all entries. Latitude and longitude data from 2020 Q1 was also excluded since there is no such data for previous quarters, and the coordinates are tied to start/end stations.

```
SELECT
    CAST(_01___Rental_Details_Rental_ID as string) as trip_id,
    _01___Rental_Details_Local_Start_Time as start_time,
    _01___Rental_Details_Local_End_Time as end_time,
    _01___Rental_Details_Duration_In_Seconds_Uncapped as tripduration,
    _03___Rental_Start_Station_ID as start_station_id,
    _03___Rental_Start_Station_Name as start_station_name,
    _02___Rental_End_Station_ID as end_station_id,
    _02___Rental_End_Station_Name as end_station_name,
    CASE
        WHEN User_Type = 'Customer' THEN 'casual'
        WHEN User_Type = 'Subscriber' THEN 'member'
        ELSE 'error'
        END AS usertype,
    Member_Gender as gender,
    _05___Member_Details_Member_Birthday_Year as birthyear
    FROM trip_data.2019_Q2
UNION ALL
    SELECT
        CAST(trip_id as string),
        start_time,
        end_time,
        tripduration,
        from_station_id,
        from_station_name,
        to_station_id,
        to_station_name,
        CASE
            WHEN usertype = 'Customer' THEN 'casual'
            WHEN usertype = 'Subscriber' THEN 'member'
            ELSE 'error'
            END,
        gender,
        birthyear
     FROM trip_data.2019_Q3
UNION ALL
    SELECT
        CAST(trip_id as string),
        start_time,
        end_time,
        tripduration,
        from_station_id,
        from_station_name,
        to_station_id,
        to_station_name,
        CASE
            WHEN usertype = 'Customer' THEN 'casual'
            WHEN usertype = 'Subscriber' THEN 'member'
            ELSE 'error'
        END,
        gender,
        birthyear
    FROM trip_data.2019_Q4
UNION ALL
    SELECT
        ride_id,
        started_at,
        ended_at,
        DATE_DIFF(ended_at, started_at, second) as tripduration,
        start_station_id,
        start_station_name,
        end_station_id,
        end_station_name,
        member_casual,
        null,
        null
        FROM trip_data.2020_Q1
ORDER BY start_time
```

After the tables were combined, the following observations were made regarding data integrity:

Consistencies

- Trip ids have unique values for each entry
- The earliest start or end date for any trip was 2019-04-01
- The latest start date for any trip was 2020-03-31
- There are exactly two unique values for gender: ‘Male’ and ‘Female’
- There are exactly two unique values for usertype: ‘member’ and ‘casual’
- There are no null values for trip id, start/end times, trip duration, and usertype.

Inconsistencies

- There was a single entry with null values for end station id and end station name
- There were many trips with a trip duration of multiple days.
- Gender and birth year were not recorded for any rides in 2020 Q1, and were missing for many rides in the other three quarters.
- Station names were occasionally, though rarely, inconsistent with start station ids. Some names ended with (Temp), some names ended with (*), and some streets had been renamed. There was also one station id, 208, which changed from “Ashland Ave & 21st St” to “LaflinSt & Cullerton St” in 2020.
- Some trip durations were zero or negative values, or not equal to the calculated trip duration when using start and end times

To address some of the inconsistencies above before analysis, the following changes were made and stored in a new table (since BigQuery sandbox does not allow table manipulation queries). Entries with zero (6059) or negative (25) trip durations were removed, as were entries with incorrectly calculated trip durations (81). Entries with null values in any field were removed, with the exception of gender and birth year (these entries were still included but taken into consideration during analysis).

To make analysis easier, start time was separated into start date and start time columns, trip duration was batched into minutes, and birthyear was converted to age based on the start year of the ride.

```
SELECT
  trip_id,
  EXTRACT(date
  FROM
    start_time) AS start_date,
  EXTRACT(time
  FROM
    start_time) AS start_time,
  EXTRACT(date
  FROM
    end_time) AS end_date,
  EXTRACT(time
  FROM
    end_time) AS end_time,
  CEILING(tripduration / 60) AS tripduration_min,
  start_station_id,
  start_station_name,
  end_station_id,
  end_station_name,
  usertype,
  gender,
  EXTRACT(year
  FROM
    start_time) - birthyear AS age
FROM
  trip_data.cleaned_data
WHERE
  tripduration > 0
  AND ABS(tripduration - DATE_DIFF(end_time, start_time, second)) > 1
  AND end_station_id IS NOT NULL;
```

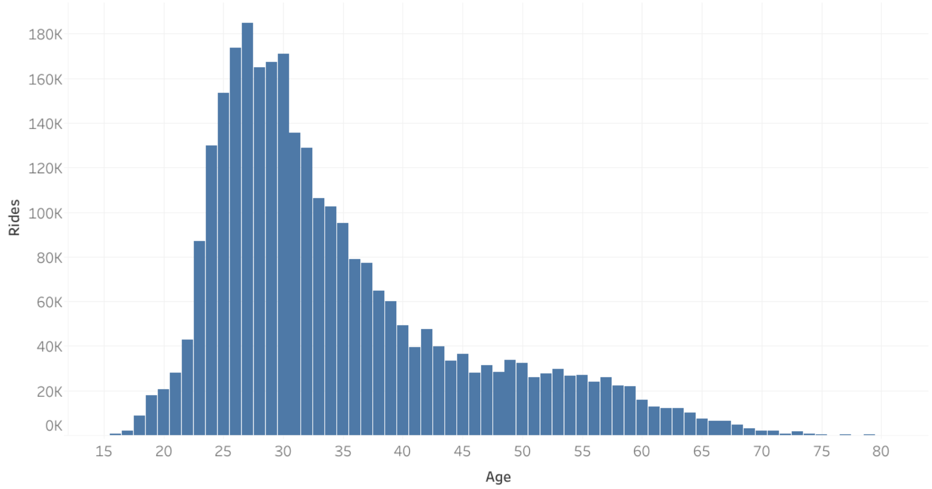
Analysis

The cleaned data was imported into Tableau for analysis. With so much data, it was important to be able to clearly visualize trends over a longer period of time.

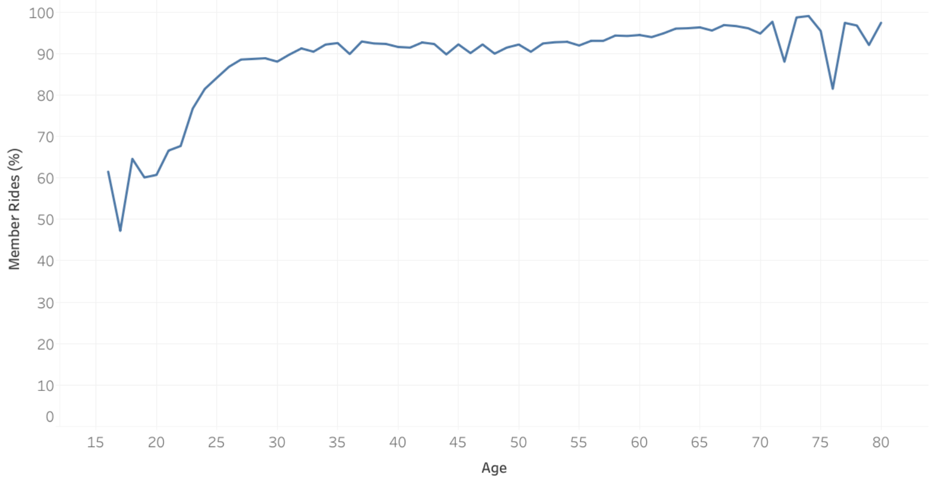
Age

It should be noted that roughly 25% of rides did not record age information, including the entirety of 2020 Q1. Null values were excluded from analysis. There were also some entries with inconsistent ages; for analysis, only ages between 15 and 80 were considered.

Count
Type: Bar
Columns: Age
Rows: Count(Age)



Membership by Age
Type: Line
Columns: Age
Rows: COUNT(if [Usertype] = 'member' then [Usertype] END) / COUNT([Usertype]) * 100

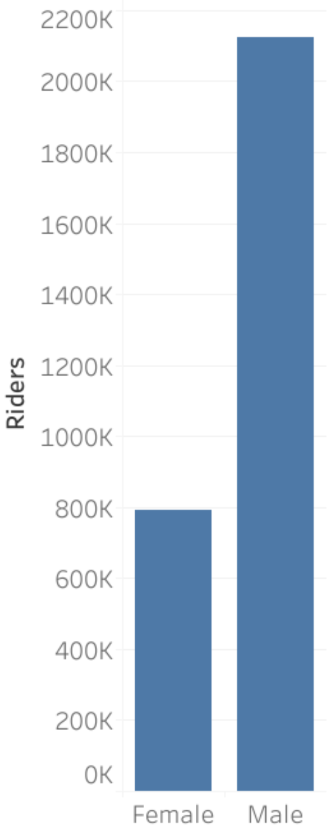


Gender

Similarly to age, roughly 25% of rides did not record gender information, including the entirety of 2020 Q1. Null values were excluded from analysis.

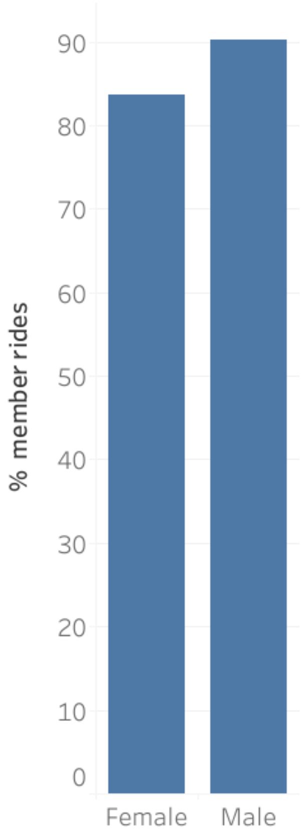
Count
Type: Bar
Columns: Gender

Rows: Count(Age)



Membership by Gender
Type: Bar
Columns: Gender
Rows: Measure Values

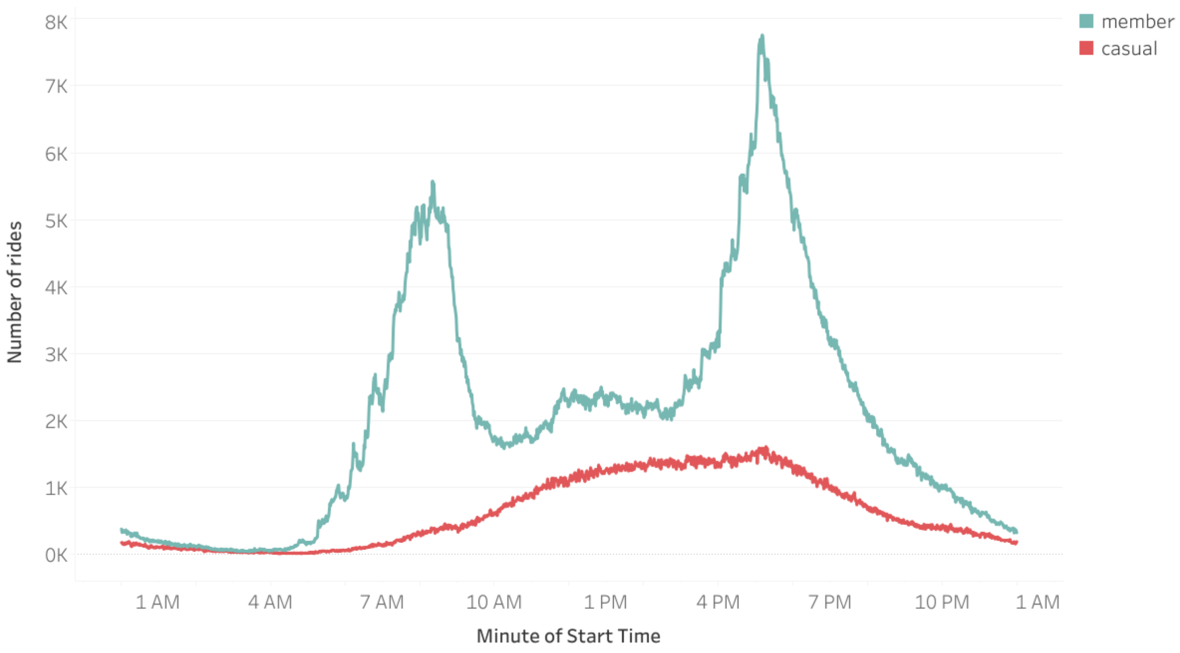
1. COUNT(if [Usertype] = "member" and [Gender] = "Female" then [Usertype] END) / COUNT(if [Gender] = "Female" then [Usertype] end) * 100
2. COUNT(if [Usertype] = "member" and [Gender] = "Male" then [Usertype] END) / COUNT(if [Gender] = "Male" then [Usertype] end) * 100



Start Time

Start Time
Type: Line
Columns: Minute(Start Time)
Rows: Measure Values

1. COUNT(if [Usertype] = 'member' then [Usertype] end)
2. COUNT(if [Usertype] = 'casual' then [Usertype] end)



Day of the Week

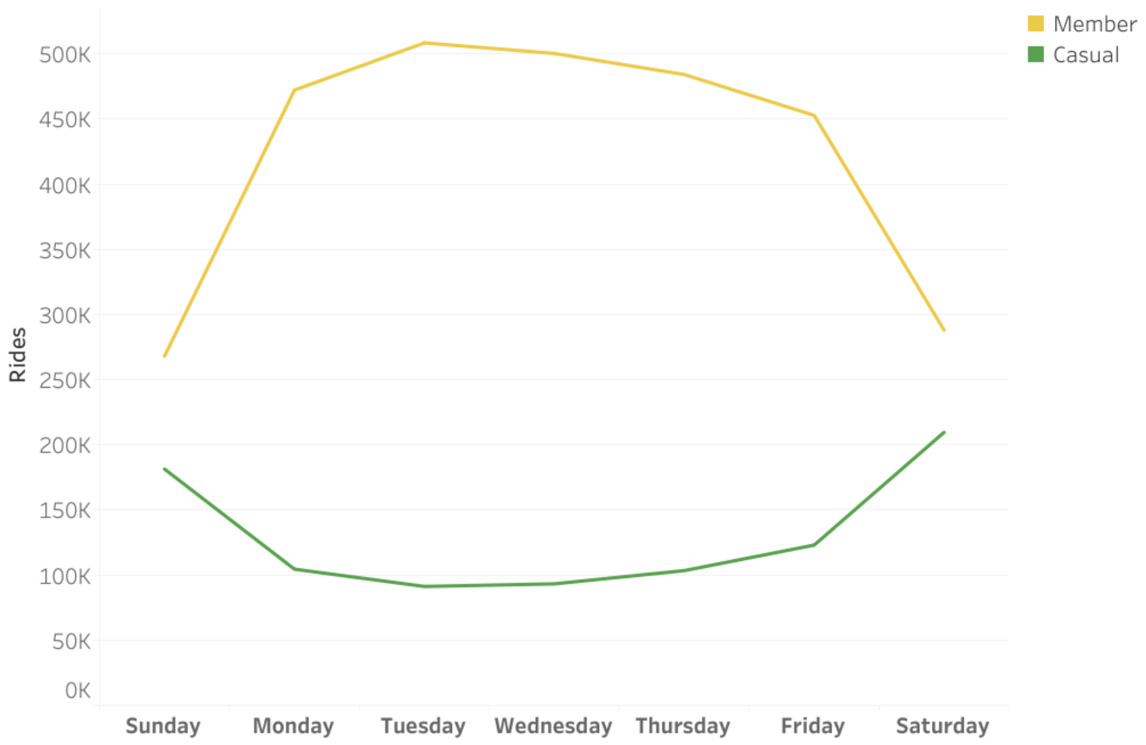
Start Date

Type: Line

Columns Weekday(Start Date)

Rows: Measure Values

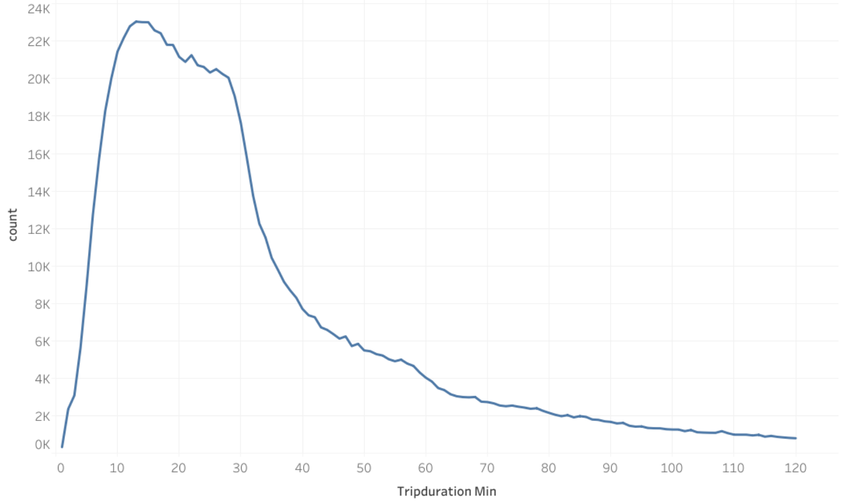
1. COUNT(if [Usertype] = 'member' then [Usertype] end)
2. COUNT(if [Usertype] = 'casual' then [Usertype] end)



Trip Duration

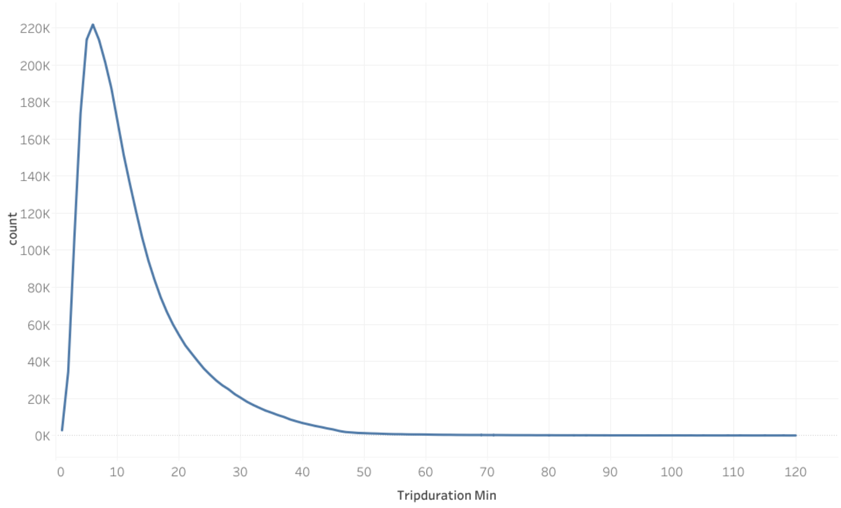
Only trips lasting between 1 minute and 2 hours were included in analysis.

Casual Rides
Type: line
Columns: Tripduration Min
Rows: COUNT(if [Usertype] = "casual" then [Usertype] END)



Member Rides
Type: line

Columns: Tripduration Min
Rows: COUNT(if [Usertype] = "member" then [Usertype] END)



Start/End Station

The cleaned data was queried to find which start and end stations were most frequently used by members and casual riders. For example, the following query was used to find the 5 busiest start stations for members:

```
SELECT
  start_station_name,
  COUNT(*) AS count
FROM
  trip_data.cleaned_data
WHERE
  usertype = 'member'
GROUP BY
  start_station_name
ORDER BY
  count DESC
LIMIT
  5
```

Members

Start Station	Total Rides
Canal St & Adams St	51948

Clinton St & Madison St	46191
Clinton St & Washington Blvd	43590
Columbus Dr & Randolph St	31053
Franklin St & Monroe St	30982

End Station	Total Rides
Canal St & Adams St	48839
Clinton St & Washington Blvd	47633
Clinton St & Madison St	44285
Daley Center Plaza	30845
Kingsbury St & Kinzie St	30404

Casual Riders

End Station	Total Rides
Streeter Dr & Grand Ave	67507
Lake Shore Dr & Monroe St	31051
Millennium Park	25509
Michigan Ave & Oak St	23982
Lake Shore Dr & North Blvd	23477

End Station	Total Rides
Streeter Dr & Grand Ave	67507
Lake Shore Dr & Monroe St	31051
Millennium Park	25509
Michigan Ave & Oak St	23982
Lake Shore Dr & North Blvd	23477