#Upload the 12 csv files to BigQuery

#Merge them using

```
INSERT INTO `vaulted-quarter-375910.Cyclistic.202201-divvy-tripdata`
SELECT *
FROM `vaulted-quarter-375910.Cyclistic.202202-divvy-tripdata`
```

The query above will copy all data from table named 202202-divvy-tripdata to table named 202201-divvy-tripdata.

Do the same for all other tables, make adjustments to the query above by changing the table names in the FROM clause

To confirm the total no. of observations after copying all the tables, use

```
SELECT COUNT (*)
FROM `vaulted-quarter-375910.Cyclistic.202201-divvy-tripdata`
```

The result is 5,667,717

To confirm the blanks in the data use:

```
SELECT *
FROM `vaulted-quarter-375910.Cyclistic.202201-divvy-tripdata`
WHERE ride_id IS NULL
```

The query above checks whether there are any rows missing information in the ride_id column.

Repeat the same for all other columns

833,064 observations were missing data in the start_station_name & start_station_id columns

We shall however proceed because there's start_lat & start_lng which could alternatively be used

892,742 observations were missing data in the end_station_name & end_station_id columns

Out of those 5,858 observations were missing data in the end_lng & end_lat columns

We'll proceed since the station names & id's don't have an impact on our analysis for now.

The rest of the columns had no missing information

To get clean data (with all stations i.e all 5,667,717 observations) which we'll continue to explore, use the query below

```
SELECT
   ride_id,
   rideable_type,
   member_casual AS membership,
   start_station_name,
   end_station_name,
```

```
   EXTRACT(date FROM started_at) AS start_date,
   EXTRACT(time FROM started_at) AS start_time,
   EXTRACT(date FROM ended_at) AS end_date,
   EXTRACT(time FROM ended_at) AS end_time,
   date_diff(ended_at,started_at, MINUTE) AS ride_length,
FROM
   `vaulted-quarter-375910.Cyclistic.202201-divvy-tripdata`
```

I saved the results in a new table as clean-data

Clean data (without nulls)

```
SELECT
   ride_id,
   rideable_type,
   member_casual AS membership,
   start_station_name,
   end_station_name,
   EXTRACT(date FROM started_at) AS start_date,
   EXTRACT(time FROM started_at) AS start_time,
   EXTRACT(date FROM ended_at) AS end_date,
   EXTRACT(time FROM ended_at) AS end_time,
   date_diff(ended_at,started_at, MINUTE) AS ride_length,
FROM
   `vaulted-quarter-375910.Cyclistic.202201-divvy-tripdata`
WHERE
   start_station_name IS NOT NULL AND
   end_station_name IS NOT NULL;
```

Explore the clean data

```
#Number of rides per start_hour

SELECT
  membership,
  EXTRACT(hour FROM start_time) AS start_hour,
  COUNT(*) AS rides_per_hour
FROM
  `vaulted-quarter-375910.Cyclistic.clean-data`
GROUP BY
  EXTRACT(hour FROM start_time),membership
```

Export to Excel, create pivot chart and table