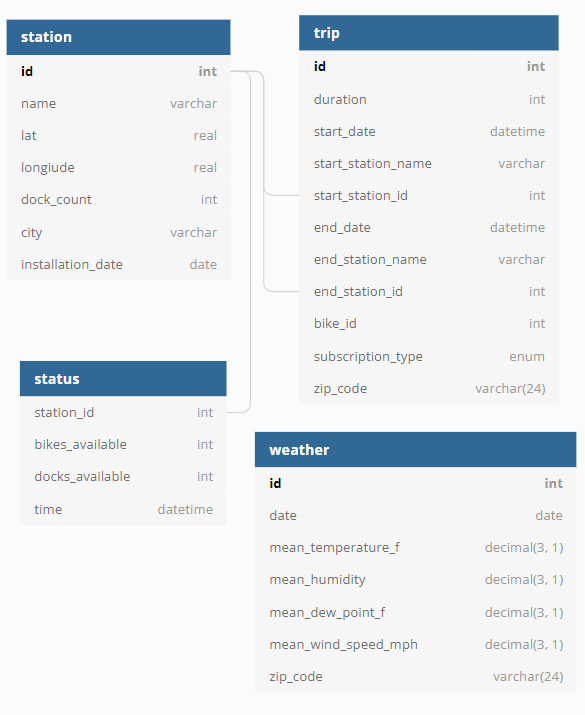
**ERD Diagram for the data:**



**DDL Statements used to create the tables in MySql:**

create database zulip\_bile\_sharing\_company;

use zulip\_bile\_sharing\_company;

CREATE TABLE `station` (

`Id` int NOT NULL ,

`Name` varchar(100) NOT NULL UNIQUE,

`Lat` real DEFAULT NULL,

`Longitude` real DEFAULT NULL,

`Dock\_Count` int DEFAULT NULL,

`City` varchar(50) DEFAULT NULL,

`Installation\_Date` date DEFAULT NULL,

PRIMARY KEY (`Id`)

);

CREATE TABLE `status` (

`Station\_Id` int NOT NULL,

`Bikes\_Available` int DEFAULT NULL,

`Docks\_Available` int DEFAULT NULL,

`Time` datetime DEFAULT NULL,

CONSTRAINT `Station\_Id` FOREIGN KEY (`Station\_Id`) REFERENCES `station` (`Id`)

);

CREATE TABLE `trip` (

`Id` int NOT NULL,

`Duration` int DEFAULT NULL,

`Start\_Date` datetime DEFAULT NULL,

`Start\_Station\_Name` varchar(100) DEFAULT NULL,

`Start\_Station\_Id` int DEFAULT NULL,

`End\_Date` datetime DEFAULT NULL,

`End\_Station\_Name` varchar(100) DEFAULT NULL,

`End\_Station\_Id` int DEFAULT NULL,

`Bike\_Id` int DEFAULT NULL,

`Subscription\_Type` varchar(24) DEFAULT NULL,

`Zip\_Code` varchar(24) DEFAULT NULL,

PRIMARY KEY (`Id`)

);

CREATE TABLE `weather` (

`Id` int NOT NULL,

`Date` date DEFAULT NULL,

`Mean\_Temperature\_F` decimal(3, 1) DEFAULT NULL,

`Mean\_Humidity` decimal(3, 1) DEFAULT NULL,

`Mean\_Dew\_Point\_F` decimal(3, 1) DEFAULT NULL,

`Mean\_Wind\_Speed\_Mph` decimal(3, 1) DEFAULT NULL,

`Zip\_Code` varchar(24) DEFAULT NULL,

PRIMARY KEY (`Id`)

);

**Load Statements used to load data from csv files provided:**

**1. weather.csv**

LOAD DATA INFILE 'C:/ProgramData/MySQL/MySQL Server 8.0/Uploads/weather.csv'

INTO TABLE `weather`

FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n'

IGNORE 1 LINES

(id, @date\_variable, @temperature, @humidity, @dewpoint, @windspeed, @zipcode)

SET `Date` = STR\_TO\_DATE(@date\_variable, '%m/%d/%Y'),

`Mean\_Temperature\_F` = if(@temperature = '', NULL, @temperature),

`Mean\_Humidity` = if(@humidity = '', NULL, @humidity),

`Mean\_Dew\_Point\_F` = if(@dewpoint = '', NULL, @dewpoint),

`Mean\_Wind\_Speed\_Mph` = if(@windspeed = '', NULL, @windspeed),

`Zip\_Code` = if(@zipcode = '', NULL, @zipcode);

**2. station.csv**

LOAD DATA INFILE 'C:/ProgramData/MySQL/MySQL Server 8.0/Uploads/station.csv'

INTO TABLE `station`

FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n'

IGNORE 1 LINES

(id, name, lat, longitude, dock\_count, city, @date\_variable)

SET `Installation\_Date` = STR\_TO\_DATE(@date\_variable, '%m/%d/%Y');

**3. trip.csv**

LOAD DATA INFILE 'C:/ProgramData/MySQL/MySQL Server 8.0/Uploads/trip.csv'

INTO TABLE `trip`

FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n'

IGNORE 1 LINES

(id, duration, @start\_date\_var, start\_station\_name, start\_station\_id, @end\_date\_var, end\_station\_name, end\_station\_id, bike\_id, subscription\_type, zip\_code)

SET `start\_date` = STR\_TO\_DATE(@start\_date\_var, '%m/%d/%Y %T'),

`end\_date` = STR\_TO\_DATE(@end\_date\_var, '%m/%d/%Y %T');

**4. status.csv**

LOAD DATA INFILE 'C:/ProgramData/MySQL/MySQL Server 8.0/Uploads/status.csv'

INTO TABLE `status`

FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n'

IGNORE 1 LINES

(station\_id, bikes\_available, docks\_available, @date\_time\_var)

SET `time` = if(@date\_time\_var like "%/%/% %:%", STR\_TO\_DATE(@date\_time\_var, '%Y/%m/%d %T'), STR\_TO\_DATE(@date\_time\_var, '%Y-%m-%d %T'));

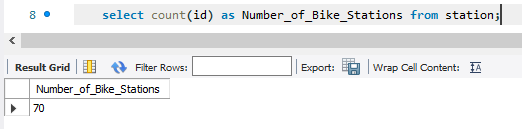
**TASK 1 – GET TO KNOW YOUR COMPANY:**

**1.1 Bike stations**

*SQL Query:*

**select count(id) as Number\_of\_Bike\_Stations from station;**

*Query Output:*



**1.2 Bikes**

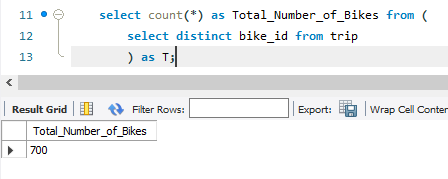
*SQL Query:*

**select count(\*) as Total\_Number\_of\_Bikes from (**

**select distinct bike\_id from trip**

**) as T;**

*Query Output:*

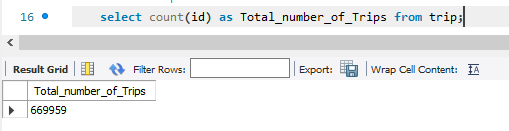


**1.3 Trips**

*SQL Query:*

**select count(id) as Total\_number\_of\_Trips from trip;**

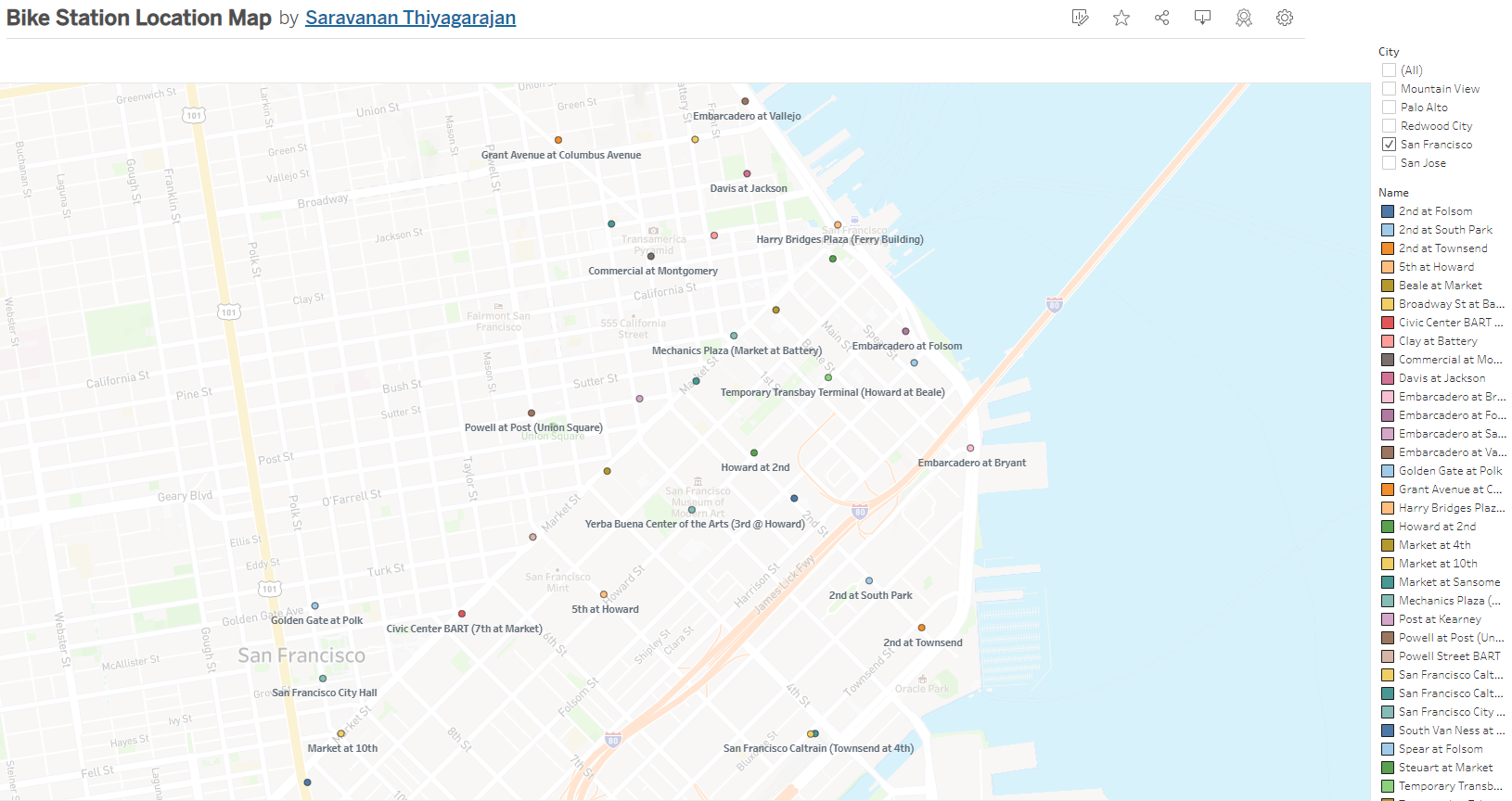
*Query Output:*



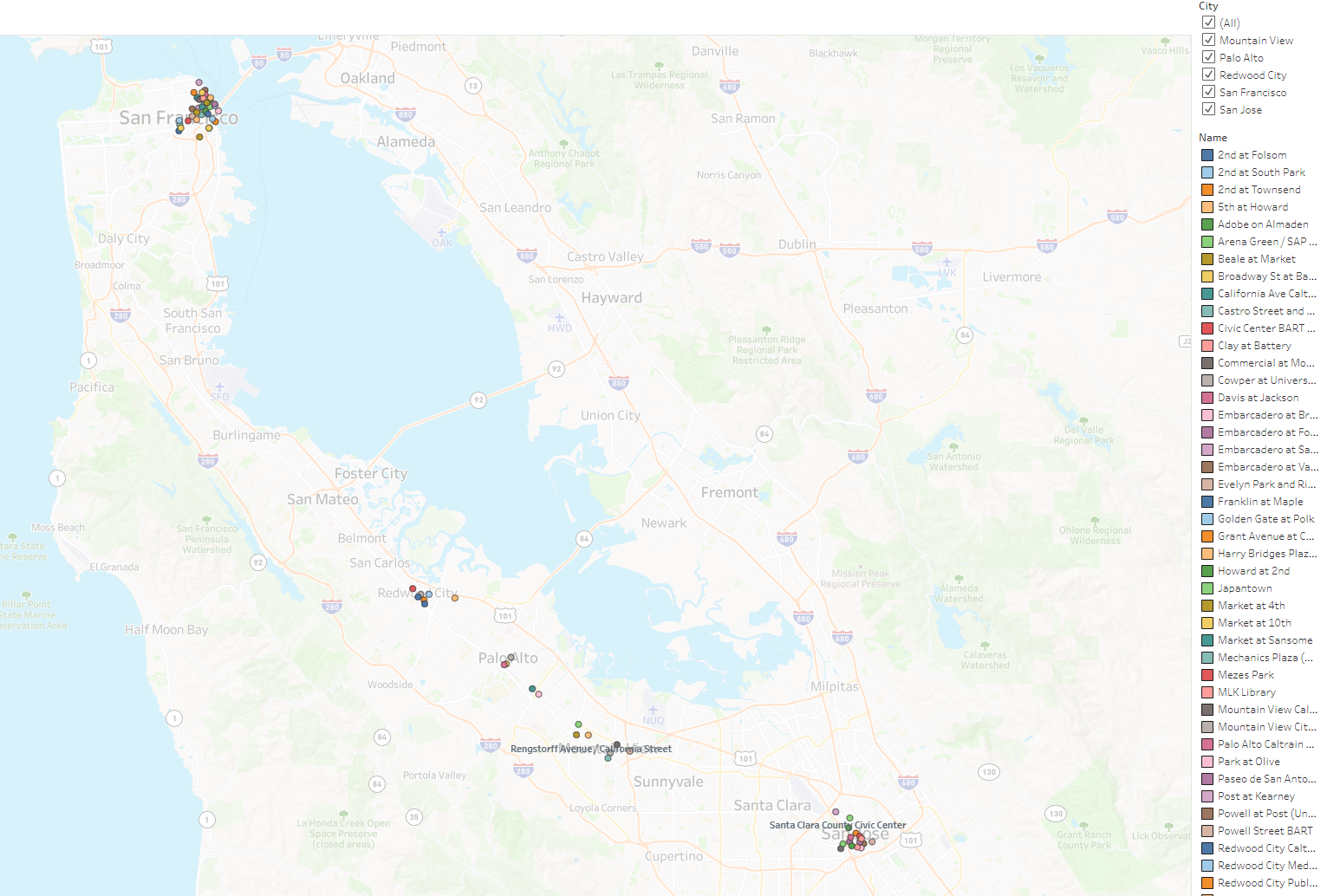
**Geographical plot to show the location of bike station using latitude and longitude.**

**Tableau link where solution is published -** [**https://public.tableau.com/app/profile/saravanan.thiyagarajan/viz/BikeStationLocationMap/BikeLocations?publish=yes**](https://public.tableau.com/app/profile/saravanan.thiyagarajan/viz/BikeStationLocationMap/BikeLocations?publish=yes)

**Screenshot 1 – Station Map of San Francisco (using Filter):**



**Screenshot 2 – Station Map of All stations (Filter on right can be used to filter based on City):**



**What is the relationship between the following columns:**

**3.1 - bike\_id (trip table) and start\_station\_id (trip table) – ‘Many – Many’**

*SQL Query:*

**select (case when t1.bike\_to\_station > 1 and t2.station\_to\_bike > 1**

**then 'Many - Many'**

**when t1.bike\_to\_station = 1 and t2.station\_to\_bike = 1**

**then '1 - 1'**

**else 'Many - 1'**

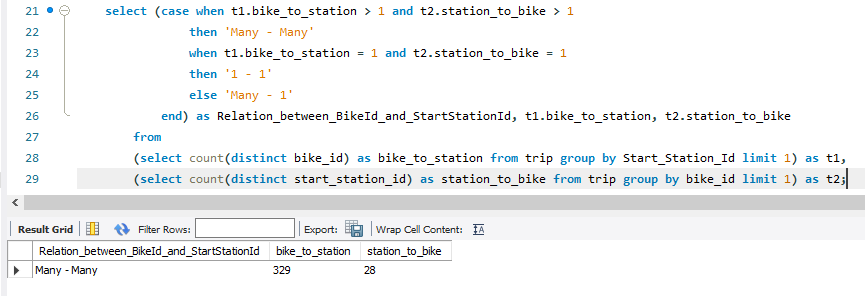
**end) as Relation\_between\_BikeId\_and\_StartStationId, t1.bike\_to\_station, t2.station\_to\_bike**

**from**

**(select count(distinct bike\_id) as bike\_to\_station from trip group by Start\_Station\_Id limit 1) as t1,**

**(select count(distinct start\_station\_id) as station\_to\_bike from trip group by bike\_id limit 1) as t2;**

*Query Output:*



**pincode (weather table) and station location (latitude and longitude in station table) – ‘1 – Many’**

*SQL Query:*

**select (case when t1.Distinct\_Location\_count = t2.Distinct\_Zip\_Code\_count and**

**t2.Distinct\_Zip\_Code\_count = t1.Distinct\_City\_count**

**then '1 - 1'**

**when t2.Distinct\_Zip\_Code\_count = t1.Distinct\_City\_count and**

**t1.Distinct\_Location\_count > t2.Distinct\_Zip\_Code\_count**

**then '1 - Many'**

**else 'Many - Many'**

**end) as Relation\_between\_Pincode\_and\_Station\_Location, Distinct\_Zip\_Code\_count, Distinct\_Location\_count, Distinct\_City\_count**

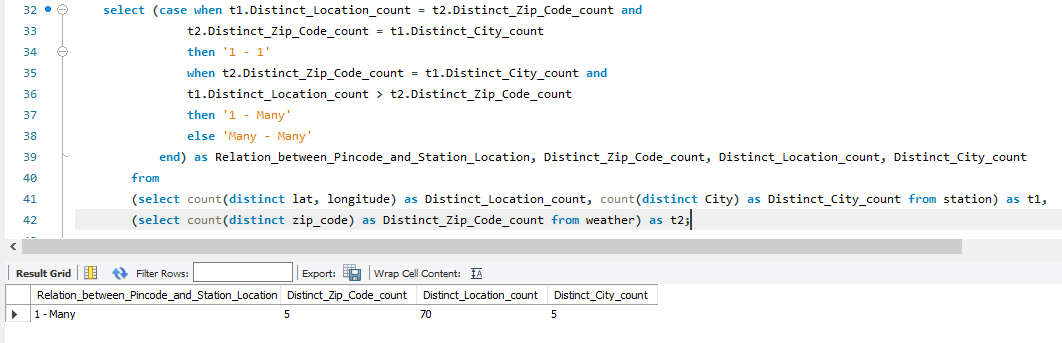
**from**

**(select count(distinct lat, longitude) as Distinct\_Location\_count, count(distinct City) as Distinct\_City\_count from station) as t1,**

**(select count(distinct zip\_code) as Distinct\_Zip\_Code\_count from weather) as t2;**

*Query Output below…*

*Query Output:*



**8/29/2013 (date column in weather table) and mean wind speed (weather table) – ‘1 – Many’**

*SQL Query:*

**select (case when count(distinct mean\_wind\_speed\_mph) = 1**

**then '1 - 1'**

**when count(distinct mean\_wind\_speed\_mph) > 1**

**then '1 - Many'**

**else 'Many - Many'**

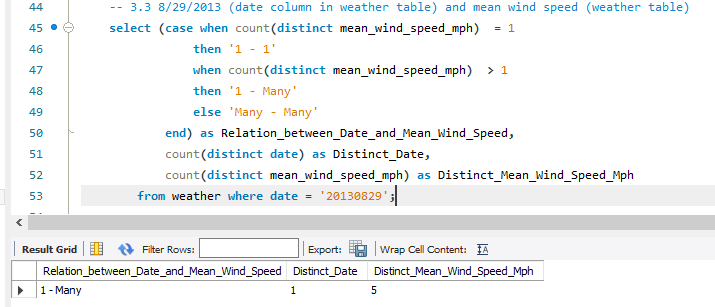
**end) as Relation\_between\_Date\_and\_Mean\_Wind\_Speed,**

**count(distinct date) as Distinct\_Date,**

**count(distinct mean\_wind\_speed\_mph) as Distinct\_Mean\_Wind\_Speed\_Mph**

**from weather where date = '20130829';**

*Query Output:*



**Scroll down..**

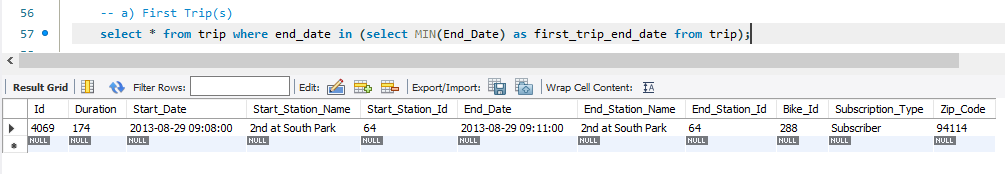
**Find the first and last trip in the data**

**a) First Trip(s)**

*SQL Query:*

**select \* from trip where end\_date in (select MIN(End\_Date) as first\_trip\_end\_date from trip);**

*Query Output:*

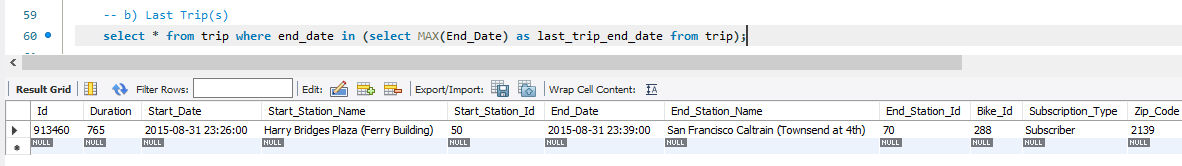


**b) Last Trip(s)**

*SQL Query:*

**select \* from trip where end\_date in (select MAX(End\_Date) as last\_trip\_end\_date from trip);**

*Query Output:*



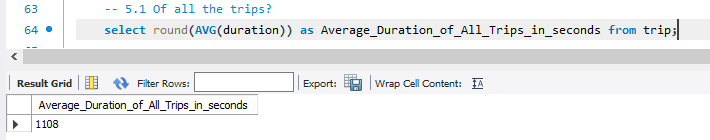
**What is the average duration**

**Q5.1 - Of all the trips?**

*SQL Query:*

**select round(AVG(duration)) as Average\_Duration\_of\_All\_Trips\_in\_seconds from trip;**

*Query Output:*



**Q5.2 Of trips on which customers are ending their rides at the same station from where they started?**

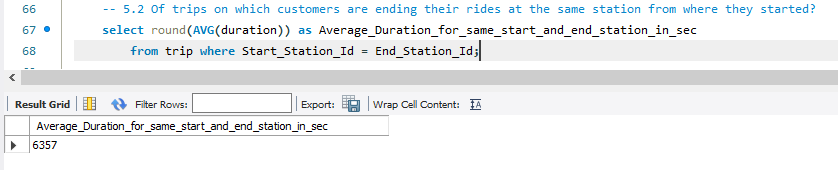
*SQL Query:*

**select round(AVG(duration)) as Average\_Duration\_for\_same\_start\_and\_end\_station\_in\_sec**

**from trip where Start\_Station\_Id = End\_Station\_Id;**

*Query Output below…*

*Query Output:*



**6 - Which bike has been used the most in terms of duration**

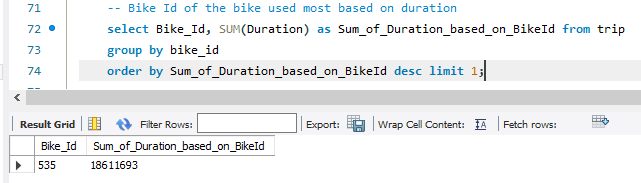
*SQL Query:*

**select Bike\_Id, SUM(Duration) as Sum\_of\_Duration\_based\_on\_BikeId from trip**

**group by bike\_id**

**order by Sum\_of\_Duration\_based\_on\_BikeId desc limit 1;**

*Query Output:*

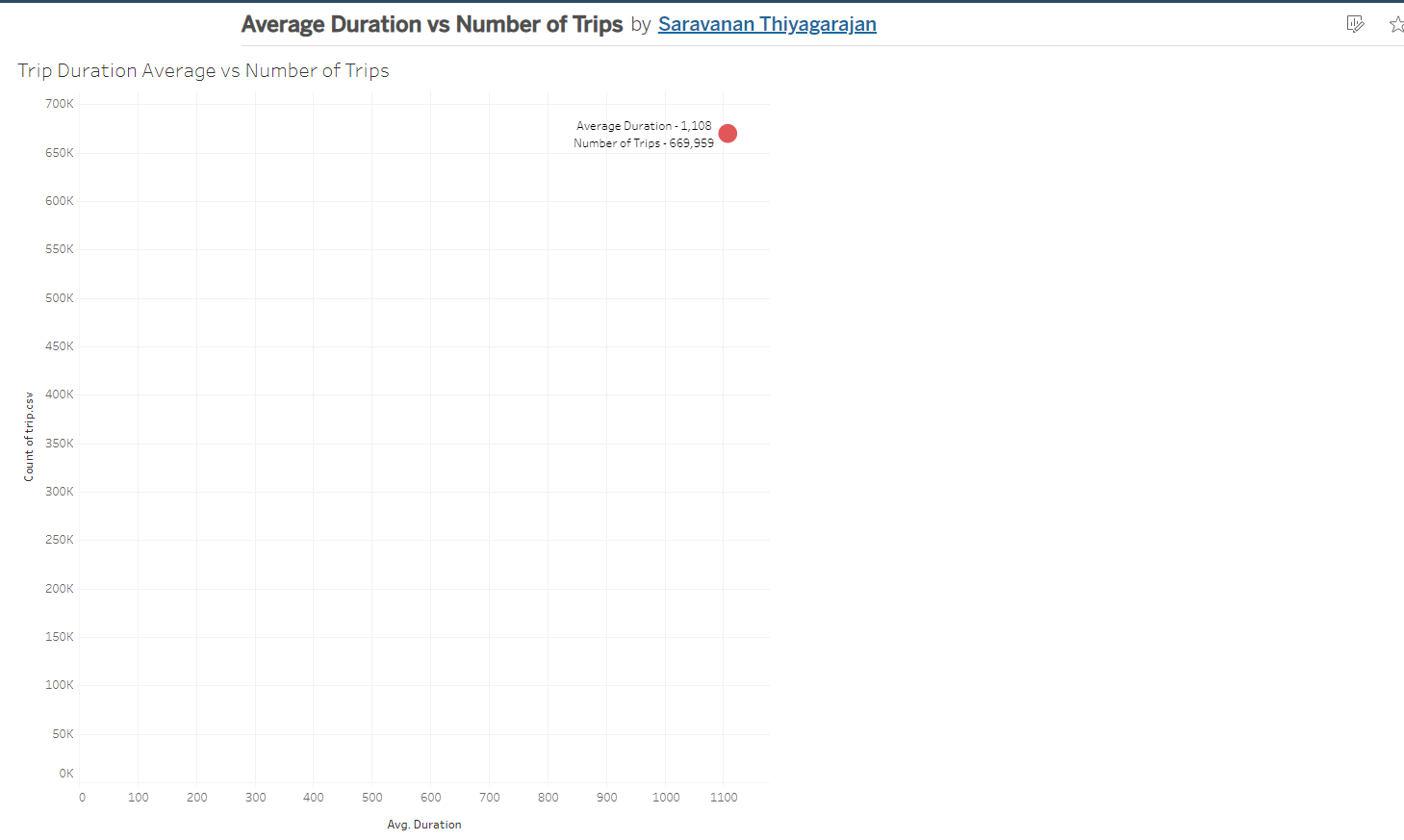


**7 – Plot the most suitable graph for the following,**

**7.1 – Average duration of a trip vs Number of trips**

**Tableau link where solution is published –** [**https://public.tableau.com/app/profile/saravanan.thiyagarajan/viz/AverageDurationvsNumberofTrips/TripDurationAveragevsNumberofTrips?publish=yes**](https://public.tableau.com/app/profile/saravanan.thiyagarajan/viz/AverageDurationvsNumberofTrips/TripDurationAveragevsNumberofTrips?publish=yes)

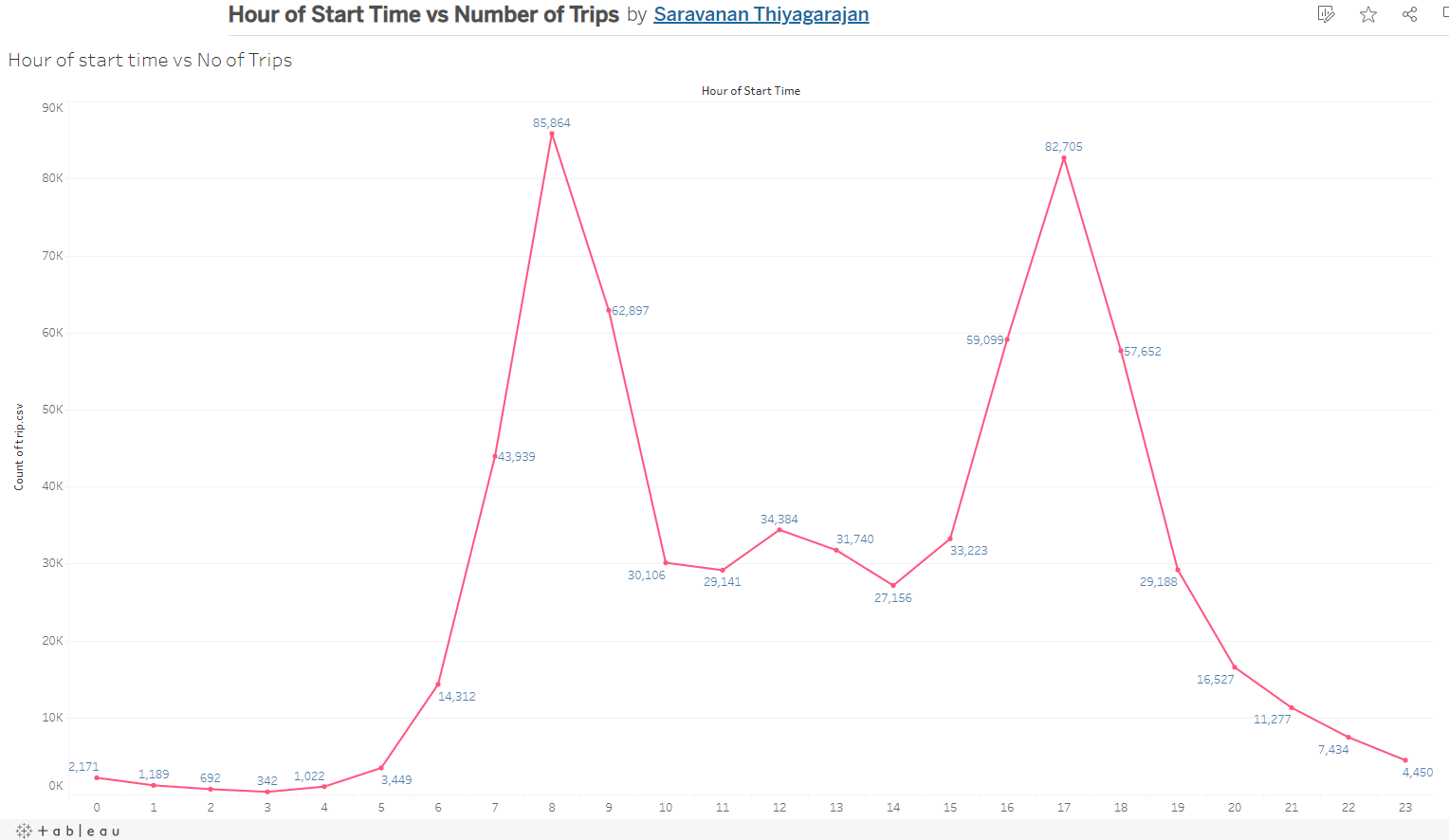
**Screenshot –**



**7.2 – Hour of start time vs No. of trips**

**Tableau link where solution is published –** [**https://public.tableau.com/app/profile/saravanan.thiyagarajan/viz/HourofStartTimevsNumberofTrips/HourofstarttimevsNoofTrips?publish=yes**](https://public.tableau.com/app/profile/saravanan.thiyagarajan/viz/HourofStartTimevsNumberofTrips/HourofstarttimevsNoofTrips?publish=yes)

**Screenshot –**



**7.3 – Day of week vs No. of trips denoting subscribers and customers with different colors**

**Tableau link where solution is published –** [**https://public.tableau.com/app/profile/saravanan.thiyagarajan/viz/DayofWeekvsNumberofTrips/DayofweekvsNoofTrips?publish=yes**](https://public.tableau.com/app/profile/saravanan.thiyagarajan/viz/DayofWeekvsNumberofTrips/DayofweekvsNoofTrips?publish=yes)

**Screenshot –**



**TASK 2 – DEMAND PREDICTION:**

**1** **- What are the top 10 least popular stations**

*SQL Query:*

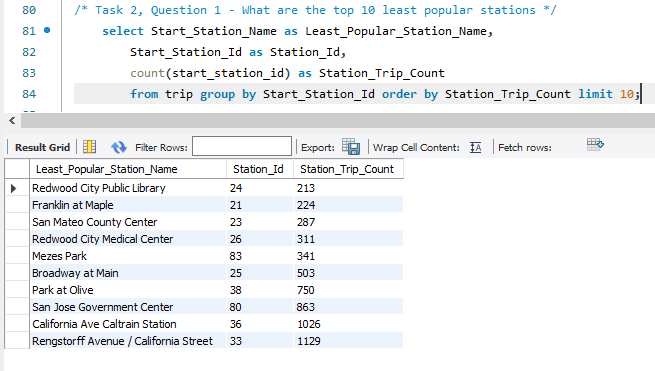
**select Start\_Station\_Name as Least\_Popular\_Station\_Name,**

**Start\_Station\_Id as Station\_Id,**

**count(start\_station\_id) as Station\_Trip\_Count**

**from trip group by Start\_Station\_Id order by Station\_Trip\_Count limit 10;**

*Query Output:*



**Idle time is the duration for which a station remains inactive. You can consider this as the time for which a station has more than 3 bikes available**

**2(i) Find idle time for station 2 on 29 August, 2013**

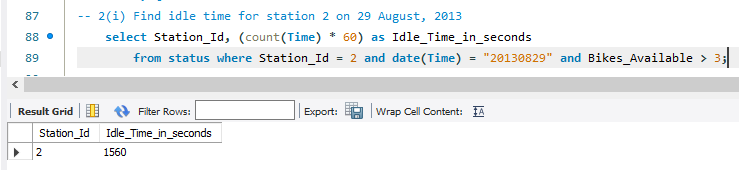
*SQL Query:*

**select Station\_Id, (count(Time) \* 60) as Idle\_Time\_in\_seconds**

**from status where Station\_Id = 2 and date(Time) = "20130829" and Bikes\_Available > 3;**

*Query Output below…*

*Query Output:*



**3 - Find the distance between 2 stations**

*SQL Query:*

**SELECT Station\_Id\_1, Station\_Id\_2,**

**(6371000 \* Acos (Cos (Radians(lat2)) \* Cos(Radians(lat1)) \***

**Cos(Radians(lon1) - Radians(lon2))**

**+ Sin (Radians(lat2)) \***

**Sin(Radians(lat1)))**

**) AS Distance\_in\_meters**

**from**

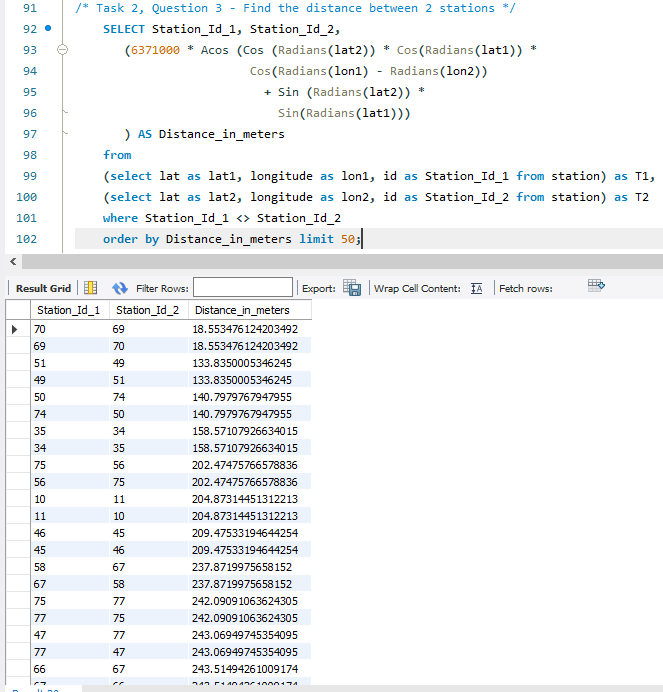
**(select lat as lat1, longitude as lon1, id as Station\_Id\_1 from station) as T1,**

**(select lat as lat2, longitude as lon2, id as Station\_Id\_2 from station) as T2**

**where Station\_Id\_1 <> Station\_Id\_2**

**order by Distance\_in\_meters limit 50;**

*Query Output:*



**Question 4 - Find 3 stations that can be shut**

*Based on Least popular stations and short distance with nearby station,*

|  |  |  |
| --- | --- | --- |
| **STATIONS RECOMMENDED for SHUTDOWN** | **NEAREST STATION** | **DISTANCE in METERS** |
| 1. Redwood City Public Library (213 trips) | Franklin at Maple | 277 meters |
| 2. San Mateo County Center (287 trips) | Redwood City Caltrain Station | 254 meters |
| 3. Park at Olive (750 tips) | California Ave Caltrain Station | 582 meters |

**TASK 3 – OPTIMIZING OPERATIONS:**

**1** **- Calculate the average number of bikes and docks available for Station 2 and Station 3**

*SQL Query:*

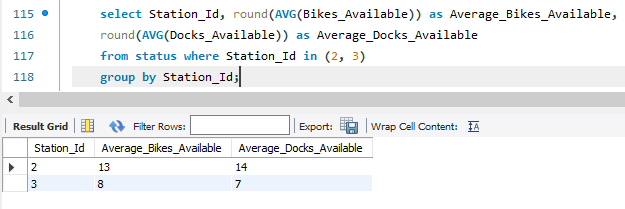
**select Station\_Id, round(AVG(Bikes\_Available)) as Average\_Bikes\_Available,**

**round(AVG(Docks\_Available)) as Average\_Docks\_Available**

**from status where Station\_Id in (2, 3)**

**group by Station\_Id;**

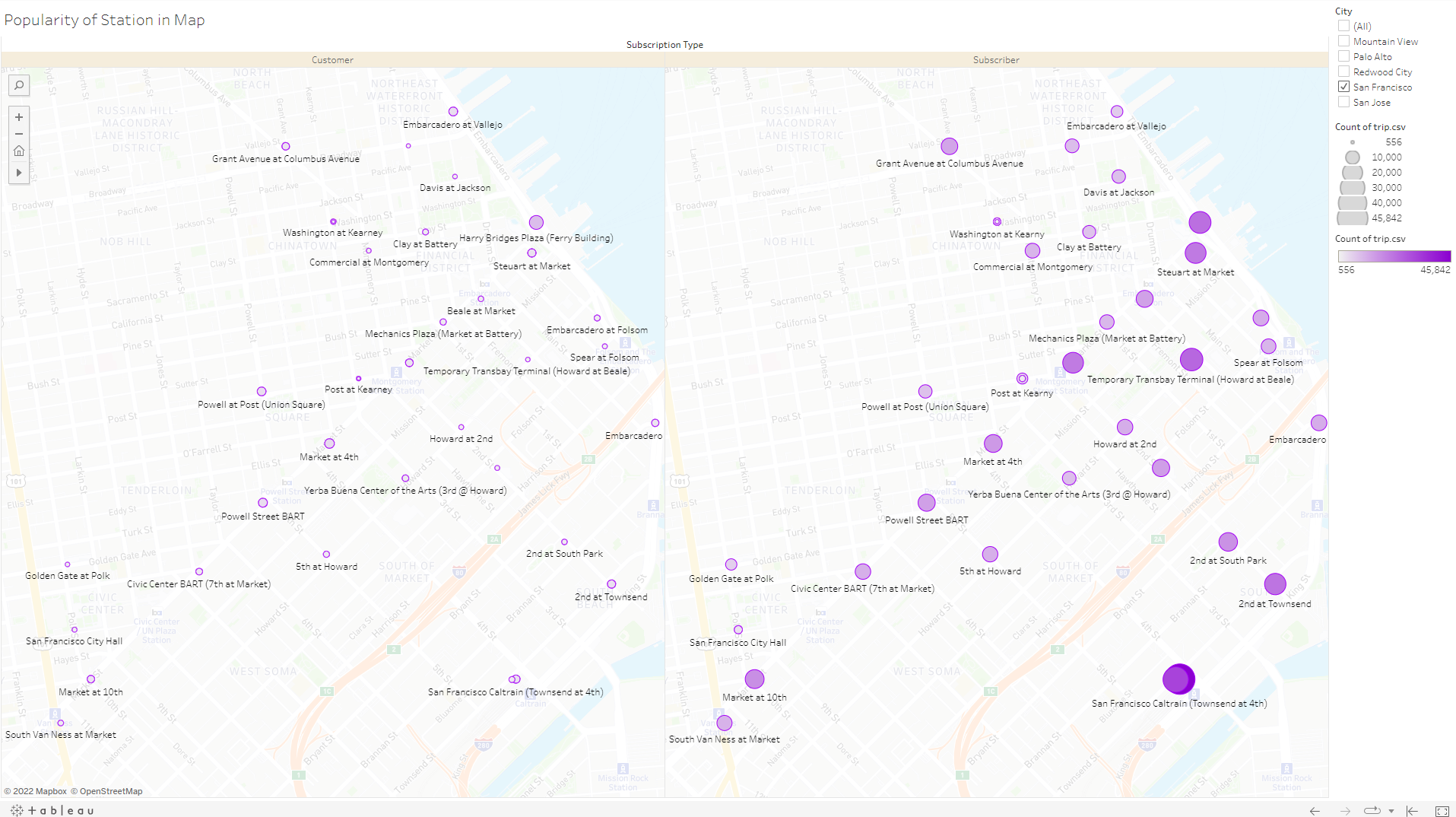
*Query Output:*



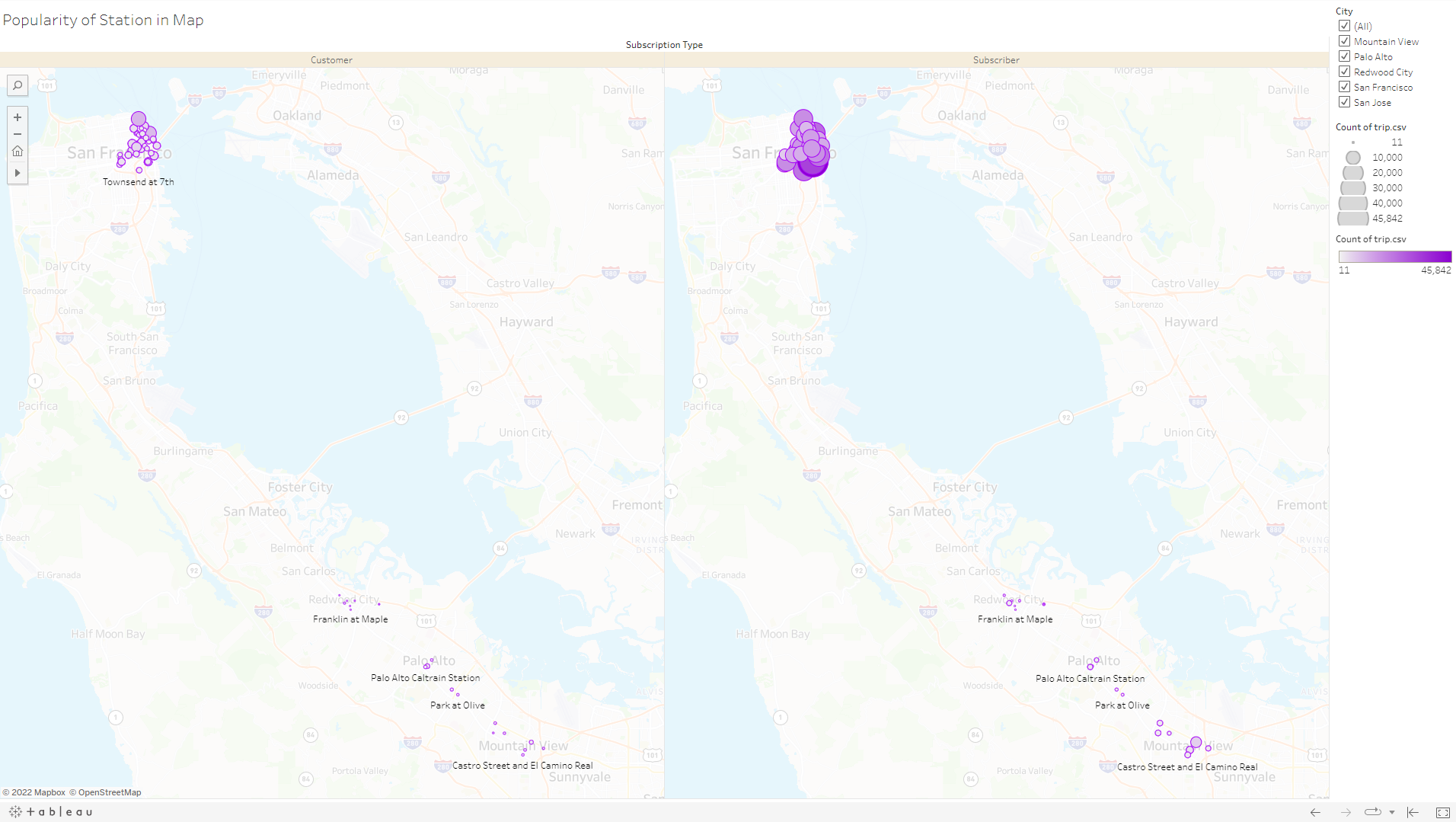
**2 – Plot the popularity of each station on a map for subscribers and customers.**

**Tableau link where solution is published –** [**https://public.tableau.com/app/profile/saravanan.thiyagarajan/viz/PopularityofeachStationonMapforSubscribersandCustomers/PopularityofStationinMap?publish=yes**](https://public.tableau.com/app/profile/saravanan.thiyagarajan/viz/PopularityofeachStationonMapforSubscribersandCustomers/PopularityofStationinMap?publish=yes)

**Screenshot 1 – Popularity of stations in San Francisco (using City filter in right):**



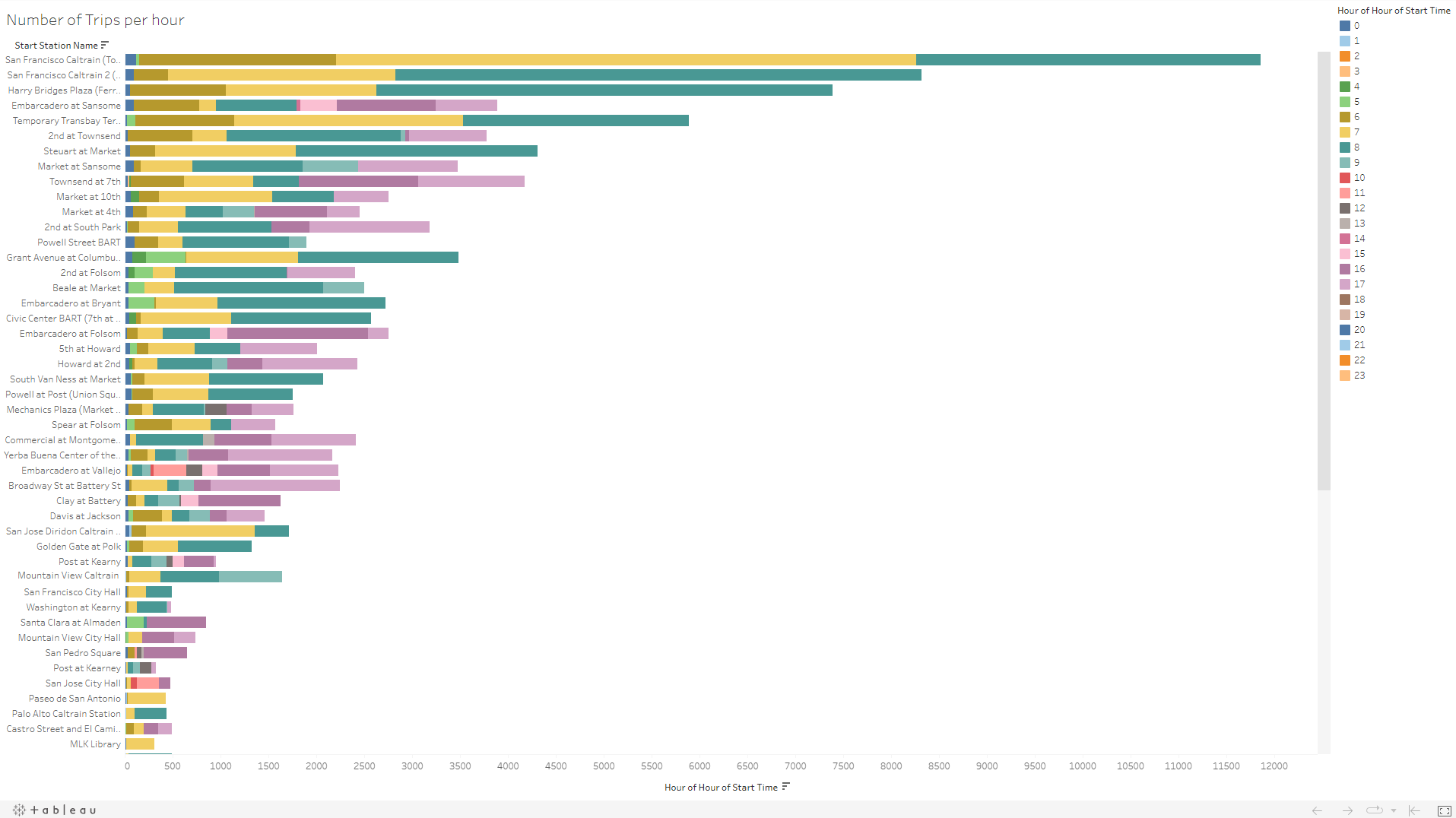
**Screenshot 2 – Popularity of All stations (City filter in right can be used to view based on city):**



**3 – Plot the number of trips per hour for all the data provided in the trip table.**

**Tableau link where solution is published –** [**https://public.tableau.com/app/profile/saravanan.thiyagarajan/viz/NumberofTripsperhourforallthedata/NumberofTripsperhour?publish=yes**](https://public.tableau.com/app/profile/saravanan.thiyagarajan/viz/NumberofTripsperhourforallthedata/NumberofTripsperhour?publish=yes)

**Screenshot –**



**4 – Use the findings above to provide insights on how to optimize operations**

* Hours between 6 am to 9 am are the busiest for bike pick-ups in these top 5 Station Ids - 70, 69, 50, 55, 61, 74. We need to ensure sufficient bikes are available in these stations from 6am and 9 am.
* Hours between 4 pm to 6 pm are the busiest for bike pick-ups in these top 5 Station Ids - 60, 77, 65, 76, 64. We need to ensure sufficient bikes are available in these stations from 4pm to 6pm.
* Station Id's 80, 21, 25, 26, 83 have average of 7 bikes available and 8 docks available. But these Stations have high Idle time and the bikes are almost not used to even the 40% capacity. Hence 60% of the bikes and docks from these stations can be removed and moved to Station Id's 70, 69, 50, 60, 77, 65, 55, 61, 74, 76, 64.
* Top 5 Stations in San Francisco are based on bike trips are –
  + ‘San Francisco Caltrain (Townsend at 4th)’,
  + ‘San Francisco Caltrain 2 (330 Townsend)’,
  + ‘Harry Bridges Plaza’,
  + ‘Embarcadero at Folsom’,
  + ‘Temporary Transbay Terminal’
* Average Bikes available are only around 10 to 12 in these Top 5 stations.
* We can move out 50% bikes from below Stations, which has the lowest number of trips in San Francisco but still has around 8 average bikes available with high idle time
  + ‘Washington at Kearney’,
  + ‘Post at Kearney’
  + and ‘San Francisco City Hall’
* Similarly, we can find the least popular Stations in the other Cities using [‘https://public.tableau.com/app/profile/saravanan.thiyagarajan/viz/PopularityofeachStationonMapforSubscribersandCustomers/PopularityofStationinMap?publish=yes’,](https://public.tableau.com/app/profile/saravanan.thiyagarajan/viz/PopularityofeachStationonMapforSubscribersandCustomers/PopularityofStationinMap?publish=yes) and move unused bikes to the most popular Stations in the respective city.
* Following Stations can be shut down based on the demand prediction. The bikes and docks can be removed and moved to the most popular Stations
  + ‘Redwood City Public Library,
  + San Mateo County Center
  + Park at Olive

**COUPLE BIKES?**

**Factors while validating the idea of couple bikes are,**

1. Introduce couple bikes in the top 10 popular stations first and monitor the usage for a time-period to understand the usage.

2. Should we add new couple bikes or convert existing bikes into couple bikes?

2.1 It might be good to add new couple bikes in the top 5 most popular stations instead of converting existing bikes, so that the existing trips will not get affected

3. Review the existing infrastructure and check feasibility to build new docks for Couple bikes.

4. In case of using the existing bikes as Couple Bikes, make the couple bikes available only in the non-busy hours (hours other than 6am to 9am and 4pm to 6pm).

5. If the distance between two Stations is less than 1000 meters, then consider adding couple bikes to only one of those stations which is more popular.

6. Make more Couple Bikes available on weekends, where the bike usage is less and couples going out on weekends are attracted to use Couple Bikes.