Task 1: Get to Know Your Company

1. What are the total numbers of:
    1. Bike Stations ?

```
# 1.1 Find total number of bike stations
select count(distinct(id)) as STATION_COUNT from station;
```

Result Grid

	STATION_COUNT
▶	70

    2. Bikes ?

```
# 1.2 Find total number of bikes
select sum(bikes_available) as BIKES_COUNT from status;
```

Result Grid

	BIKES_COUNT
▶	10831

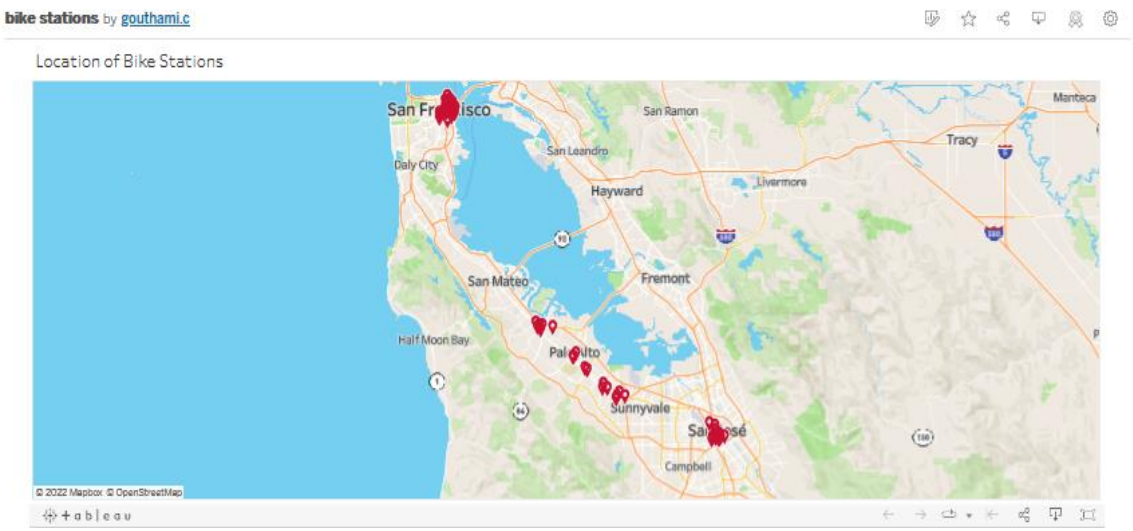
    3. Trips ?

```
# 1.3 Find total number of trips
select count(id) as TRIP_COUNT from trip;
```

Result Grid

	TRIP_COUNT
▶	669959

2. Construct a geographical plot to show the location of each bike station using the latitude and longitude provided under the Station table.
   https://public.tableau.com/app/profile/gouthami.c/viz/bikestations_16472381933850/bikestations



3. What is the relationship between the following columns (one to one, many to one, many to many)?

   1. Bike_id (Trip table) and start_station_id (Trip table)
      Each bike can have one start station at a time → one to one relationship
      Each bike station can have multiple bikes at a time -> one to many relationships

   2. Pincode (weather table) and station location (latitude and longitude in station table)
      Each pincode can have multiple station location -> one to many relationships
      Each station can have one pincode -> one to one relationship

   3. 8/29/2013 (date column in weather table) and mean wind speed (Weather table)
      Date column in weather table and mean wind speed -> one to one relationship

4. Find first and last trip from the trip table

```
# 4. Find the first and the last trip in the data.
SELECT * FROM trip order by id asc limit 1;
SELECT * FROM trip order by id desc limit 1;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

id	duration	start_date	start_station_name	start_station_id	end_date	end_station_name	end_station_id	bike_id	subscription_type	zip_code
4069	174	2013-08-29 09:08:00	2nd at South Park	64	2013-08-29 09:11:00	2nd at South Park	64	288	Subscriber	94114

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

id	duration	start_date	start_station_name	start_station_id	end_date	end_station_name	end_station_id	bike_id	subscription_ty
913460	765	2015-08-31 23:26:00	Harry Bridges Plaza (Ferry Building)	50	2015-08-31 23:39:00	San Francisco Caltrain (Townsend at 4th)	70	288	Subscriber

5. What is the average duration:
    1. Of all the trips?

```
# 5. What is the average duration:
# 5.1 Of all the trips?
select avg(duration) as AVG_DURATION from trip;
```

Result Grid

AVG_DURATION
1107.9498

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

```
# 5.2 Of trips on which customers are ending their rides at the same station from where they started?
select avg(duration), start_station_id, end_station_id from trip
where start_station_id = end_station_id
group by start_station_id
order by avg(duration) desc;
```

Result Grid | Filter Rows:

avg(duration)	start_station_id	end_station_id
4661.3356	7	7
4481.4836	55	55
4402.0210	60	60
4351.2750	8	8
4305.7600	75	75
4088.6411	45	45
4037.7510	63	63
3931.1303	82	82
3910.2477	14	14
3864.7904	10	10
3740.6990	62	62
3570.4134	51	51
3456.4349	42	42
3335.2821	5	5
3242.3846	25	25
3216.5059	69	69
3203.0597	80	80
3137.4187	61	61
3071.7090	49	49
2722.5576	65	65
2069.9402	64	64

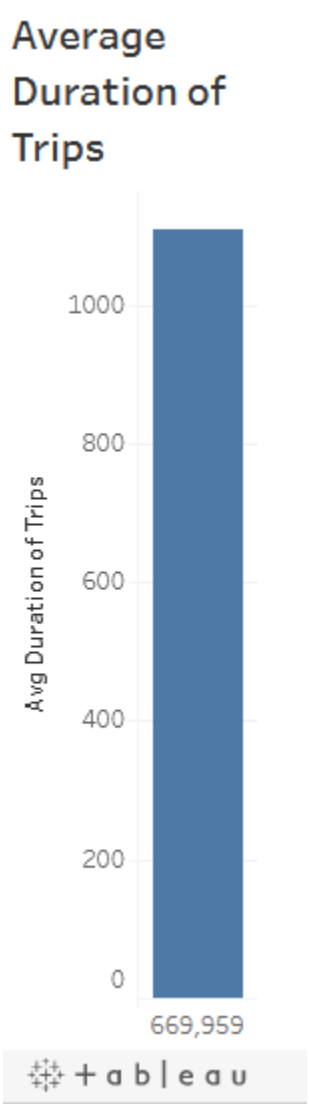
6. Which bike has been used the most in terms of duration? (Answer with the Bike ID)
   BIKE ID : 392

```
# 6. Which bike has been used the most in terms of duration? (Answer with the Bike ID)
select count(bike_id) as Rented_Count , bike_id from trip group by bike_id order by Rented_Count desc limit 1;
```

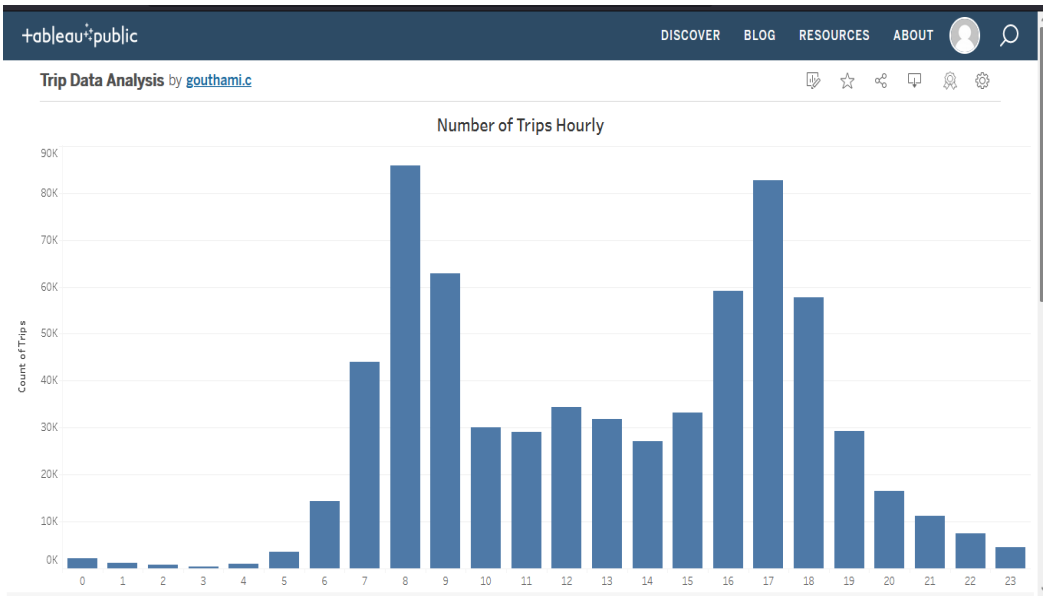
Result Grid | Filter Rows:

	Rented_Count	bike_id
▶	2061	392

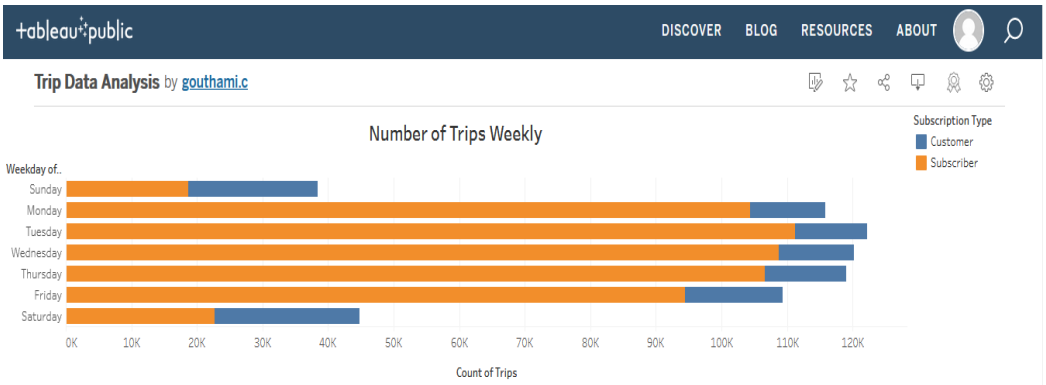
7. Plot the suitable graph for the following:
   1. The average duration of trips vs Number of trips
      https://public.tableau.com/app/profile/gouthami.c/viz/TripDataAnalysis_16472514158640/AvgDurationofTrips



2. Hour of start time versus number of trips
   https://public.tableau.com/app/profile/gouthami.c/viz/TripDataAnalysis_16472514158640/NumberofTripsHourly



3. Days of the week versus No of trips also denote subscribers and customers with different colors
   https://public.tableau.com/app/profile/gouthami.c/viz/TripDataAnalysis_16472514158640/NumberofTripsWeekly



Task 2: Demand Prediction

Zulip is running under a loss and has decided to shut operations for three of its stations. You have to use the data provided to help Zulip decide which three stations should be shut.

1. What are the top 10 least popular stations? Hint: Find the least frequently appearing start stations from the Trip table

```
# 2.1 What are the top 10 least popular stations? Hint: Find the least frequently appearing start stations from the Trip table.
select start_station_name, count(start_station_id) as Freq_Station
from trip
group by start_station_name
order by Freq_Station asc
limit 10;
```

Result Grid | Filter Rows:

start_station_name	Freq_Station
San Jose Government Center	23
Broadway at Main	67
Redwood City Public Library	213
Franklin at Maple	224
San Mateo County Center	287
Redwood City Medical Center	311
Mezes Park	341
Stanford in Redwood City	436
Park at Olive	750
Santa Clara County Civic Center	840

2. 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

```
select station_id, bikes_available, time
from status
where bikes_available > 3;
```

Result Grid | Filter Rows:

station_id	bikes_available	time
2	11	2013/08/30 17:00:01
2	11	2013/08/30 18:00:01
2	11	2013/08/30 19:00:01
2	10	2013/08/30 20:00:01
2	10	2013/08/30 21:00:01
2	11	2013/08/30 22:00:02
2	11	2013/08/30 23:00:01
2	11	2013/08/31 00:00:01
2	11	2013/08/31 00:02:01

2.1 Find the idle time on station 2 on the date “2013/08/29”

```
#2.3 Find the idle time on station 2 on date "2013/08/29"
#No data available on "2013/08/29"
select station_id, bikes_available, time
from status where bikes_available>3 and station_id=2
group by time
having time >= "2013/08/29 00:00:00" and time <= "2013/08/30 23:59:59"
order by bikes_available desc;
```

Result Grid | Filter Rows:

station_id	bikes_available	time
2	11	2013/08/30 17:00:01
2	11	2013/08/30 18:00:01
2	11	2013/08/30 19:00:01
2	10	2013/08/30 20:00:01
2	10	2013/08/30 21:00:01
2	11	2013/08/30 22:00:02
2	11	2013/08/30 23:00:01
2	11	2013/08/31 00:00:01
2	11	2013/08/31 00:02:01
2	11	2013/08/31 00:03:01

3. In case two stations are nearby, it might be possible to shut one down. Find the distance between consecutive stations (between Stations 1 and 2, Stations 2 and 3, and so on). The Haversine formula $2\arcsin(\sqrt{\sin^2(\varphi_2-\varphi_{12})+(1-\sin^2(\varphi_2-\varphi_{12})-\sin^2(\varphi_2+\varphi_{12}).\sin^2(\lambda_2-\lambda_{12})})$ is used to find the distance between two points on a sphere given their longitude and latitude. (φ_1,λ_1) is the latitude–longitude pair for the first station, and $(\varphi_2,\lambda_2)$is the latitude–longitude pair for the second station. You can find the SQL code for this formula given below

   Added the station name to which distance is calculated and it's respective dock count for comparison.

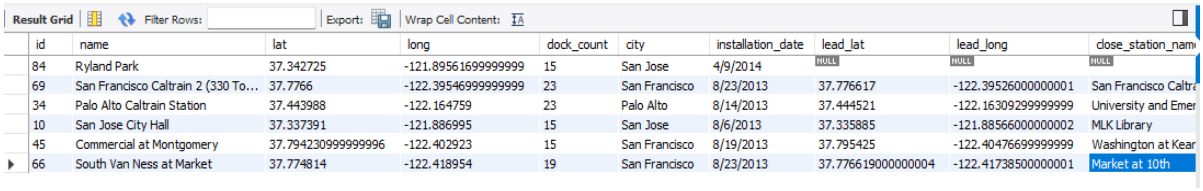
```
select *,
acos(
cos(radians( st.lat ))
* cos(radians( st.lead_lat ))
* cos(radians( st.long ) - radians( st.lead_long ))
+ sin(radians( st.lat ))
* sin(radians( st.lead_lat ))
) AS consecutiveStationDistance from (select *,
LEAD(station.lat) OVER(ORDER BY station.id) as lead_lat,
LEAD(station.long) OVER(ORDER BY station.id ) as lead_long,
LEAD(station.name) OVER(ORDER BY station.id) as close_station_name,
LEAD(station.dock_count) OVER(ORDER BY station.id) as close_dock_count
from station) AS st order by consecutiveStationDistance asc limit 6;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

id	name	lat	long	dock_count	close_station_name	consecutiveStationDistance	close_dock_count	city
84	Ryland Park	37.342725	-121.89561699999999	15	NULL	NULL	NULL	San Jose
69	San Francisco Caltrain 2 (330 To...	37.7766	-122.39546999999999	23	San Francisco Caltrain (Townsend at 4th)	0.00000291217644391830 06	19	San Franc
34	Palo Alto Caltrain Station	37.443988	-122.164759	23	University and Emerson	0.00002488951173541675 6	11	Palo Alto
10	San Jose City Hall	37.337391	-121.886995	15	MLK Library	0.00003215714087476411	19	San Jose
45	Commercial at Montgomery	37.794230999999996	-122.402923	15	Washington at Kearney	0.00003287950587763 97	15	San Franc
66	South Van Ness at Market	37.774814	-122.418954	19	Market at 10th	0.00003822240505735 64	27	San Franc

4. Use the findings above to recommend three stations that can be shut. (open ended) For example, if the Japantown and Ryland stations are nearby, and the Japantown is not as popular as the Ryland station, then it can be recommended to shut.

   From the above query we can find out the stations which are closer by

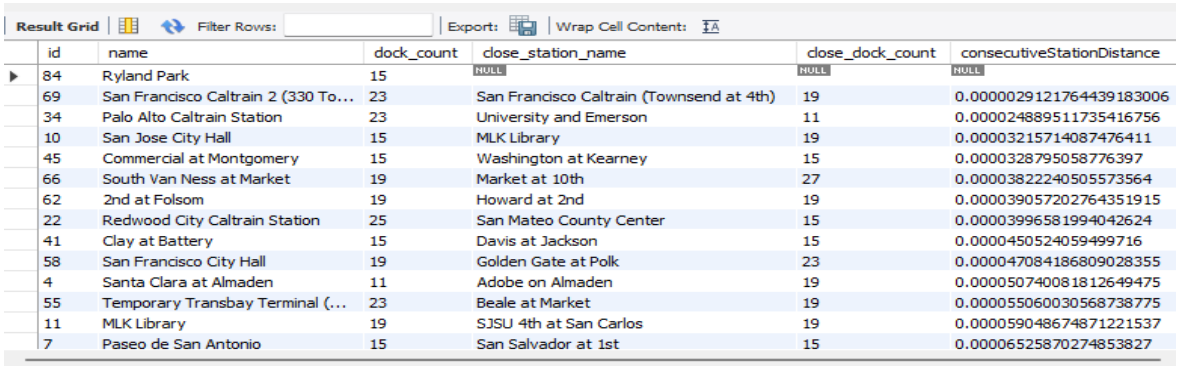


id	name	lat	long	dock_count	city	installation_date	lead_lat	lead_long	close_station_nam
84	Ryland Park	37.342725	-121.895616999999999	15	San Jose	4/9/2014	NULL	NULL	NULL
69	San Francisco Caltrain 2 (330 To...	37.7766	-122.395469999999999	23	San Francisco	8/23/2013	37.776617	-122.39526000000001	San Francisco Caltr
34	Palo Alto Caltrain Station	37.443988	-122.164759	23	Palo Alto	8/14/2013	37.444521	-122.16309299999999	University and Emer
10	San Jose City Hall	37.337391	-121.886995	15	San Jose	8/6/2013	37.335885	-121.88566000000002	MLK Library
45	Commercial at Montgomery	37.794230999999996	-122.402923	15	San Francisco	8/19/2013	37.795425	-122.40476699999999	Washington at Kear
66	South Van Ness at Market	37.774814	-122.418954	19	San Francisco	8/23/2013	37.776619000000004	-122.41738500000001	Market at 10th

   From the trip table find out top stations by finding number of trips starting from start_station_name.

```
# Find the freq of station used based on the start_station name from trip table
select count(id) as freq_station, start_station_name from trip group by start_station_name order by freq_station;
```

   Compare the results to find out which stations can be shut down based on dock count, frequency of station used, popularity of closer station;



id	name	dock_count	close_station_name	close_dock_count	consecutiveStationDistance
84	Ryland Park	15	NULL	NULL	NULL
69	San Francisco Caltrain 2 (330 To...	23	San Francisco Caltrain (Townsend at 4th)	19	0.000002912176443918 3006
34	Palo Alto Caltrain Station	23	University and Emerson	11	0.000024889511735416756
10	San Jose City Hall	15	MLK Library	19	0.000032157140874764 11
45	Commercial at Montgomery	15	Washington at Kearney	15	0.000032879505877639 7
66	South Van Ness at Market	19	Market at 10th	27	0.000038222405055735 64
62	2nd at Folsom	19	Howard at 2nd	19	0.000039057202764351915
22	Redwood City Caltrain Station	25	San Mateo County Center	15	0.00003996581994042624
41	Clay at Battery	15	Davis at Jackson	15	0.000045052405949971 6
58	San Francisco City Hall	19	Golden Gate at Polk	23	0.000047084186809028355
4	Santa Clara at Almaden	11	Adobe on Almaden	19	0.000050740081812649475
55	Temporary Transbay Terminal (...	23	Beale at Market	19	0.000055060030568738775
11	MLK Library	19	SJSU 4th at San Carlos	19	0.000059048674871221537
7	Paseo de San Antonio	15	San Salvador at 1st	15	0.00006525870274853827

   From the above table:

   1. **MLK Library** with dock count 19 and frequency of trips 2034 closer to **SJSU 4th at San Carlos** with dock count 19 and frequency of trips 1170 Based on above statistics **SJSU 4th at San Carlos** can be shut down.

   2. **Redwood City Caltrain Station** with dock count 25 and frequency of trips 1554 closer to **San Mateo County Center** with dock count 15 and frequency of trips 287 can be shut down **.**

   3. **Commercial at Montgomery** with dock count 15 and frequency of trips 11888 is closer to **Washington at Kearney** with dock count 15 and frequency of trips 1472 can be shut down.

   From above details **SJSU 4th at San Carlos, San Mateo County Center and Washington at Kearney** can be shut down**.**

Task 3: Optimizing Operations

Throughout the day, bikes keep moving around the city due to the trips. Zulip has to find out how to effectively move bikes around to ensure the demand is met with adequate supply. This is to ensure that at any time, there are sufficient bikes available at a given station. Here are some points that you will have to consider while deciding on the transportation of bikes from one place to another:

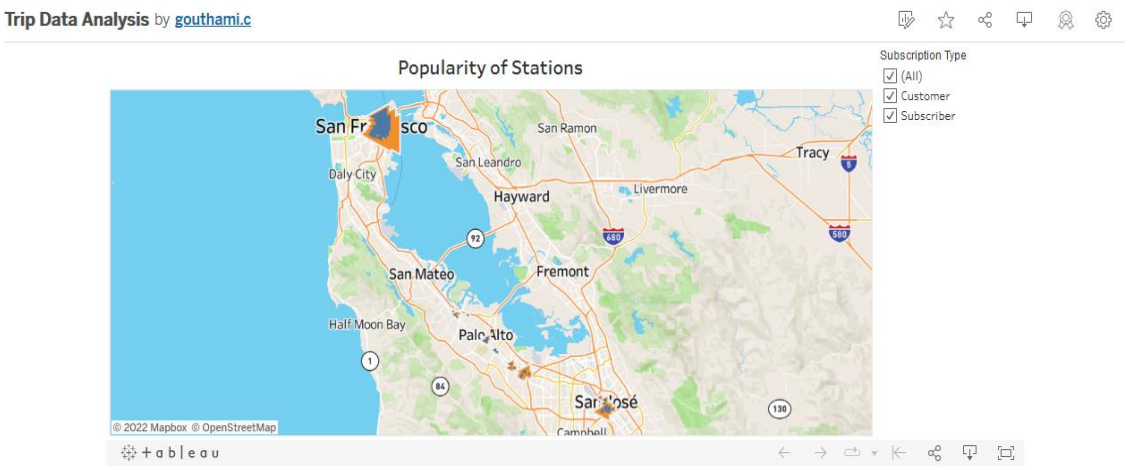
1. Calculate the average number of bikes and docks available for Station 2. (Hint: Use the Status table.)

```
# 3.1 Calculate the average number of bikes and docks available for Station 2. (Hint: Use the Status table.)
select avg(bikes_available) as avg_bikes_available,
avg(docks_available) as avg_docks_available, station_id
from status
where station_id = 2;
```

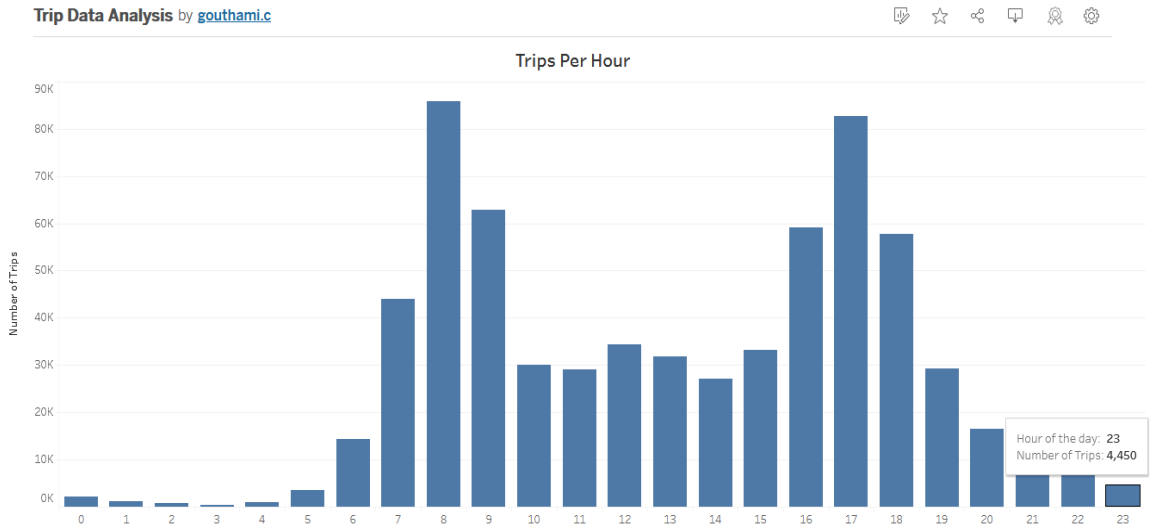
Result Grid | Filter Rows: | Expo

avg_bikes_available	avg_docks_available	station_id
10.8310	16.1690	2

2. Plot the popularity of each station on a map for subscribers and customers. (Hint: Popular stations appear most frequently under the column start_station_name in the Trip table) https://public.tableau.com/app/profile/gouthami.c/viz/TripDataAnalysis_16472514158640/PopularityofStations



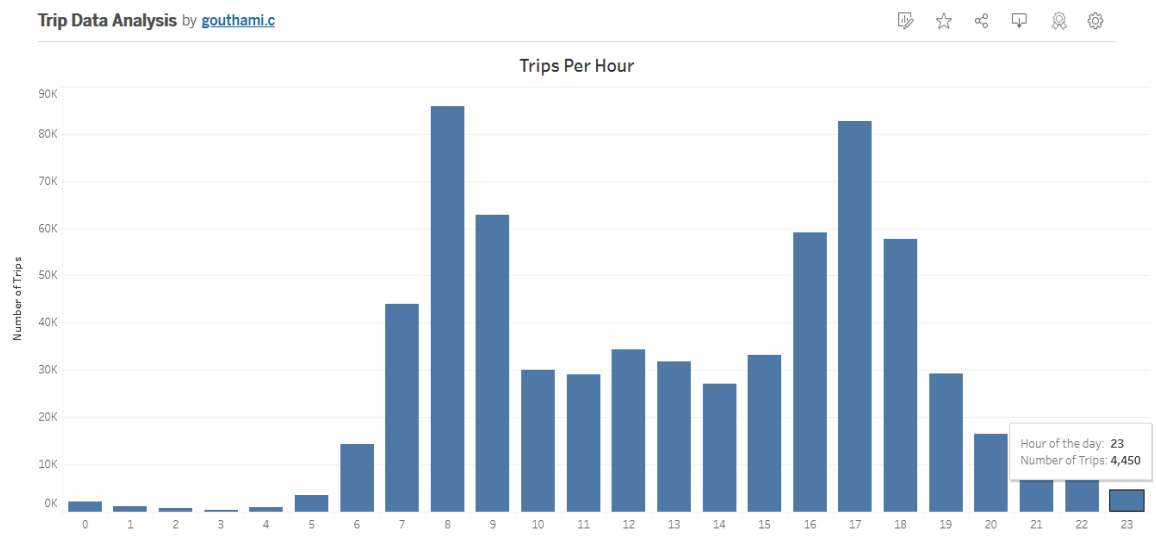
3. Plot the number of trips per hour for all the data provided in the Trip table.
   https://public.tableau.com/app/profile/gouthami.c/viz/TripDataAnalysis_16472514158640/TripsPerHour



4. Use the findings above to provide insights on how to optimize operations. (open ended)
   Based on the above data we can see first five hours of the day has minimum number of trips So we need to make use of this 5 hours to move the bikes to required stations. Required stations can be found checking the bikes_available from status table and Popular stations can be found using trips table with statrt_station_name.

Task 4: Couple Bikes? (Bonus)

Zulip has decided to start a new product line called Couple Bikes. This will enable two persons to travel from one station to another at the same time. What are some of the factors that you will have to consider while validating the idea of couple bikes?




```
# Find the freq of station used based on the start_station name from trip table
select count(id) as freq_station, start_station_name from trip group by start_station_name order by freq_station desc limit 10;
```

Result Grid | Filter Rows: | Export:

freq_station	start_station_name
49092	San Francisco Caltrain (Townsend at 4th)
33742	San Francisco Caltrain 2 (330 Townsend)
32934	Harry Bridges Plaza (Ferry Building)
27713	Embarcadero at Sansome
26089	Temporary Transbay Terminal (Howard at Beale)
25837	2nd at Townsend
24838	Steuart at Market
24172	Market at Sansome
23724	Townsend at 7th
20272	Market at 10th

From the above details:
Peak hour of bike rentals is from morning 8 AM to evening 6 PM
So we can start the couple bikes in the above 10 popular stations as a trial during peak hour and collect the data and analyze available data to make further decisions.