Bay Area Bike Share

Kevin Loftis, Lexie Sun, Esther Liu, Marine Lin, Akanksha

Group Number : 18

Data Description

Total Size : 2 GB

Station.csv : Data about the stations where users can pickup or return bikes

Status.csv : Data about the number of bikes and docks available at various stations at different time stamps

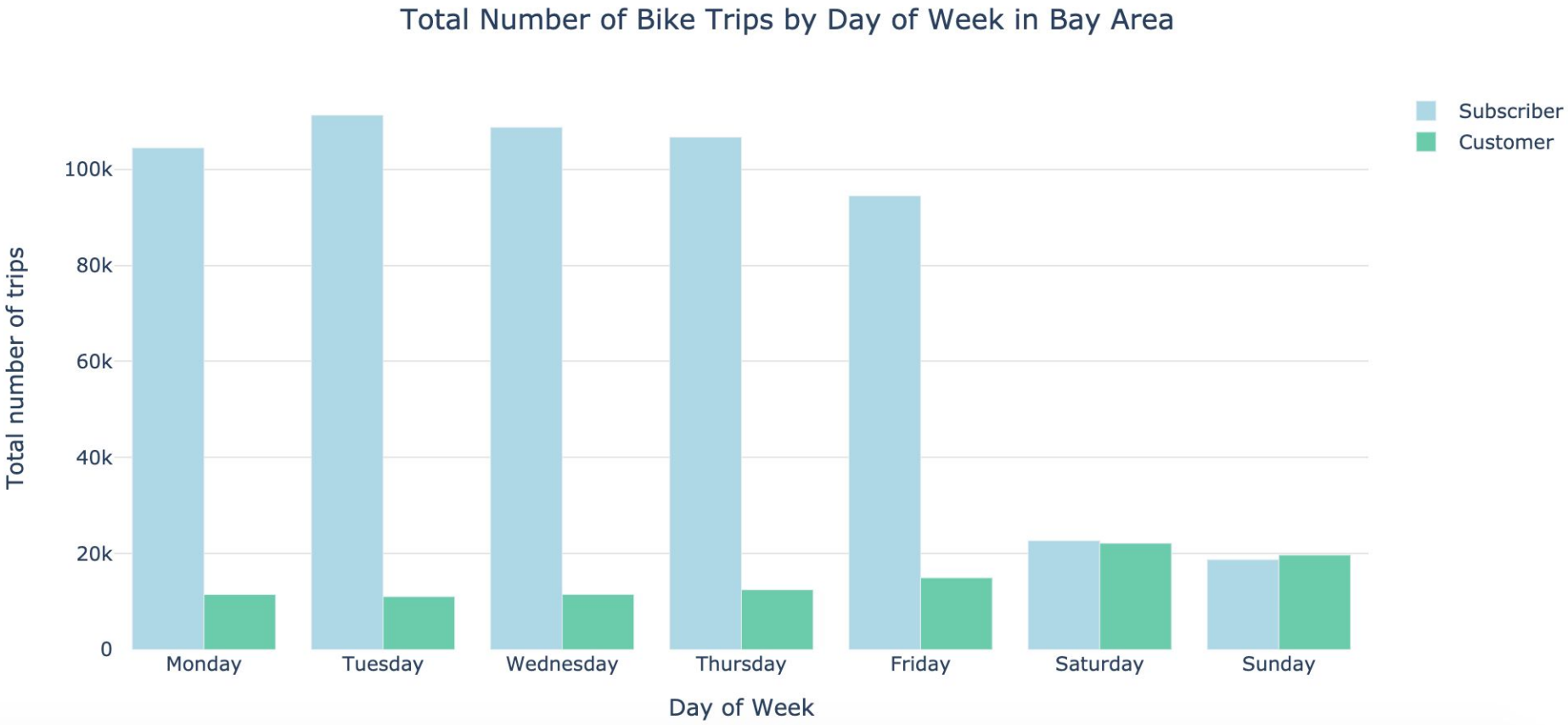
Trips.csv: Data about individual bike trips with details such as duration, subscription type, start station name, end station name, etc.

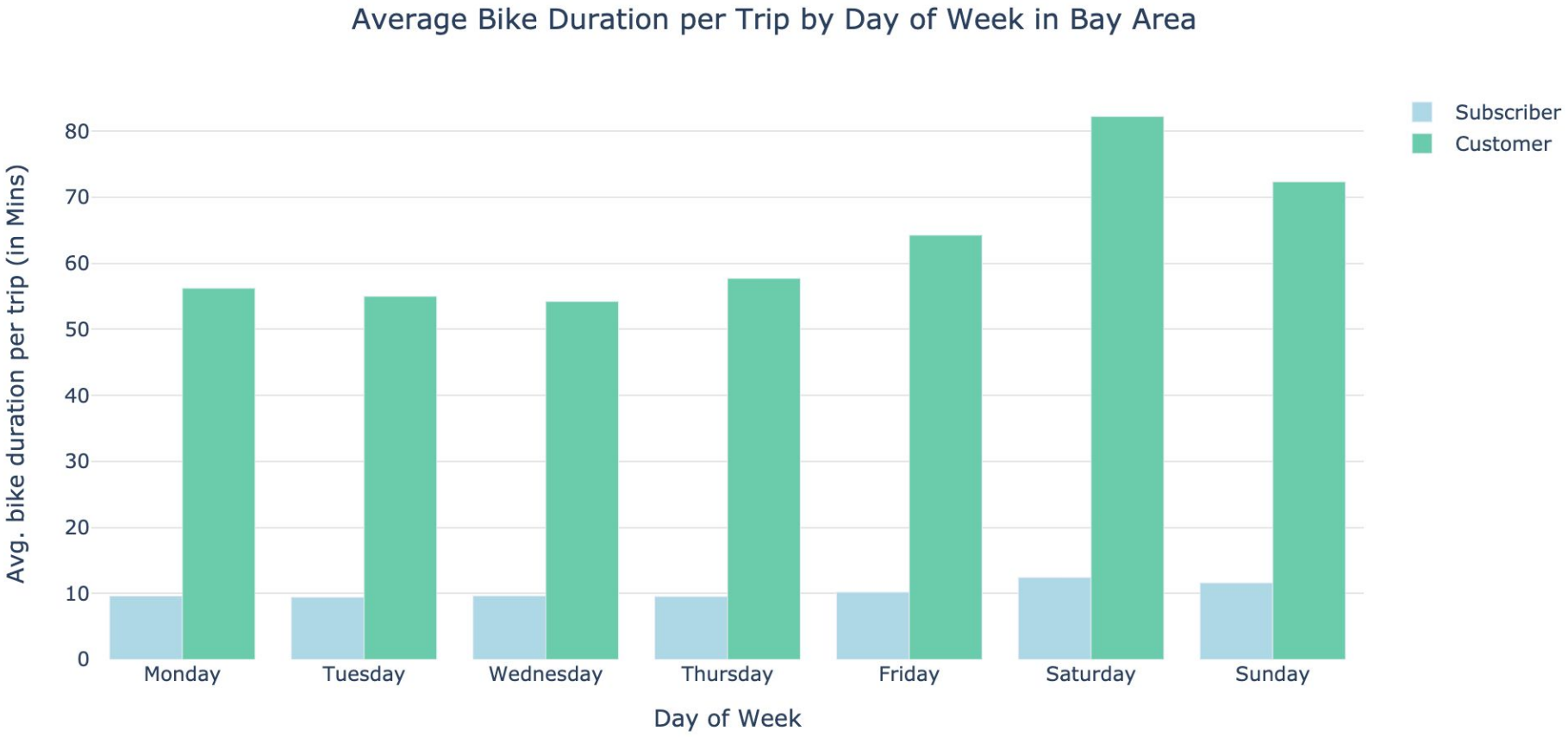
Weather.csv : Data about the weather on a specific day for certain zip codes. Features include temperature, humidity, visibility, wind speed, etc.

Preprocessing & Data Visualization

Akanksha

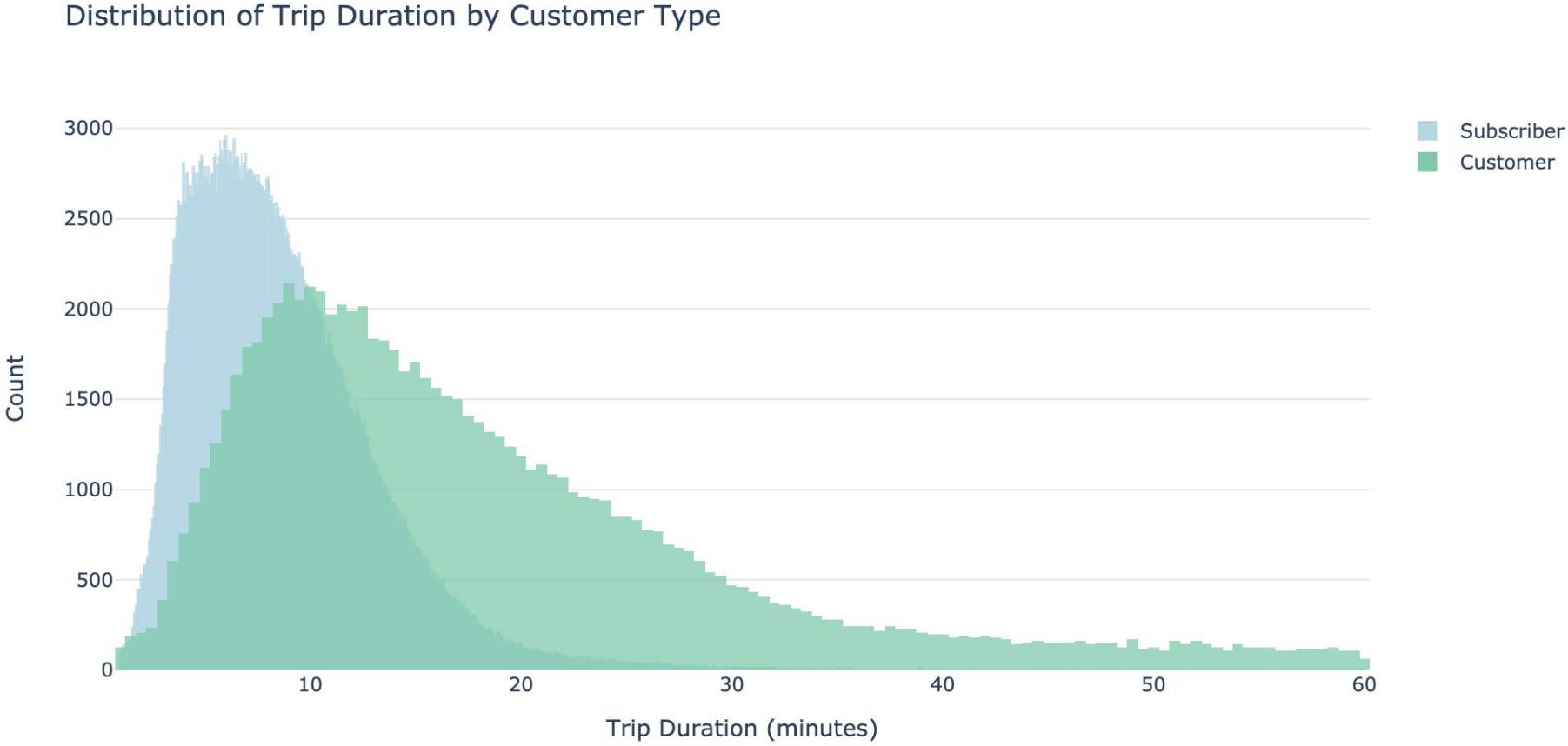
- Segregated subscribers and customers into two different RDDs
- Aggregated the total number of trips per day of the week for subscribers and customers
- Computed average bike duration per trip per day of the week for both subscribers and customers


Total Number of Bike Trips by Day of Week in Bay Area
Subscriber
Customer
Total number of trips
100k
80k
60k
40k
20k
0
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Day of Week


Average Bike Duration per Trip by Day of Week in Bay Area
Subscriber
Customer
Avg. bike duration per trip (in Mins)
0
10
20
30
40
50
60
70
80
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Day of Week

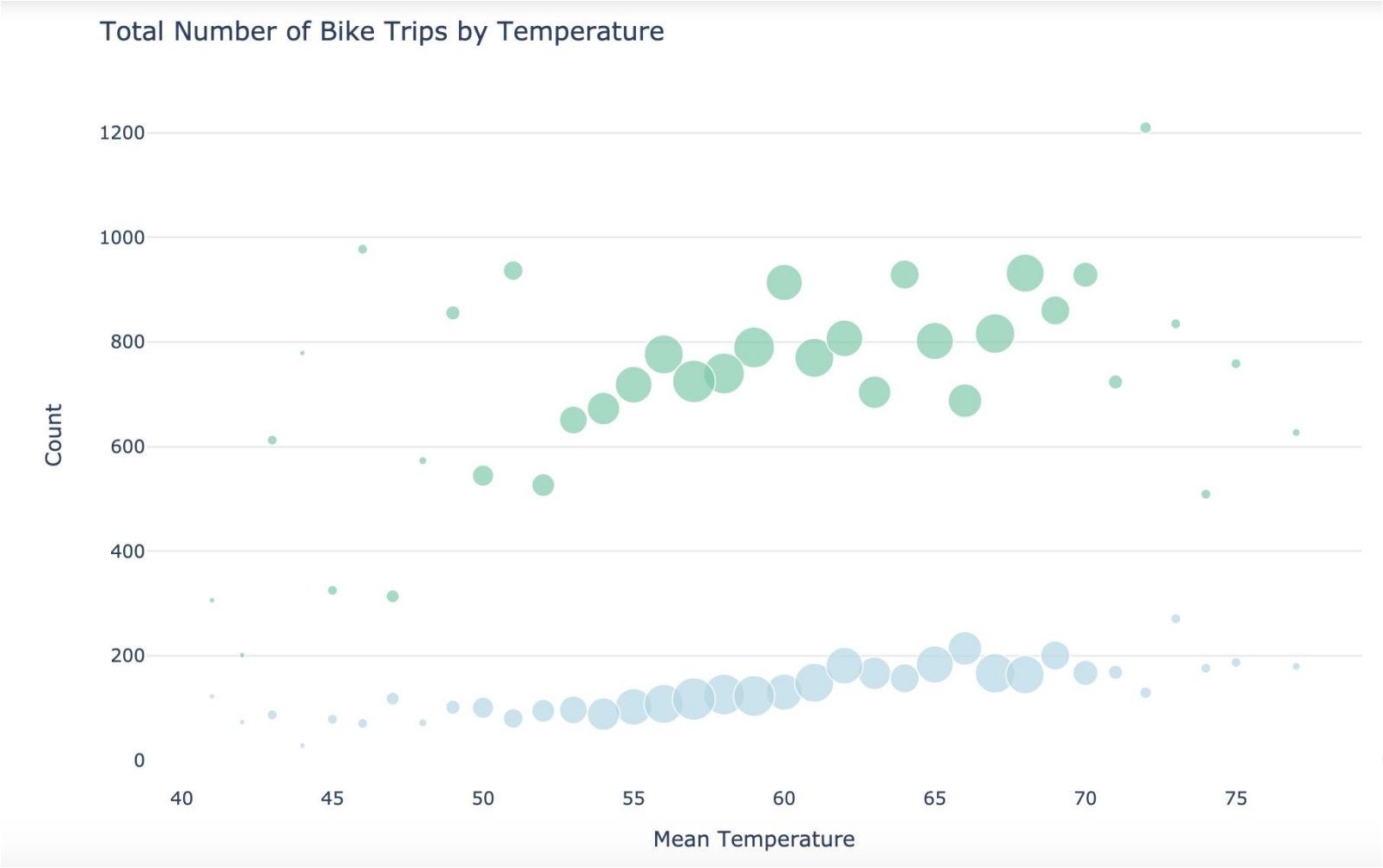
Lexie

- For *trip.csv*, group by customer type(subscriber/customer) to count the number of trips of different durations.
- Apply a filter to remove all the trips have a duration longer than 60 mins.


Distribution of Trip Duration by Customer Type
Subscriber
Customer
Count
3000
2500
2000
1500
1000
500
0
10
20
30
40
50
60
Trip Duration (minutes)

Esther

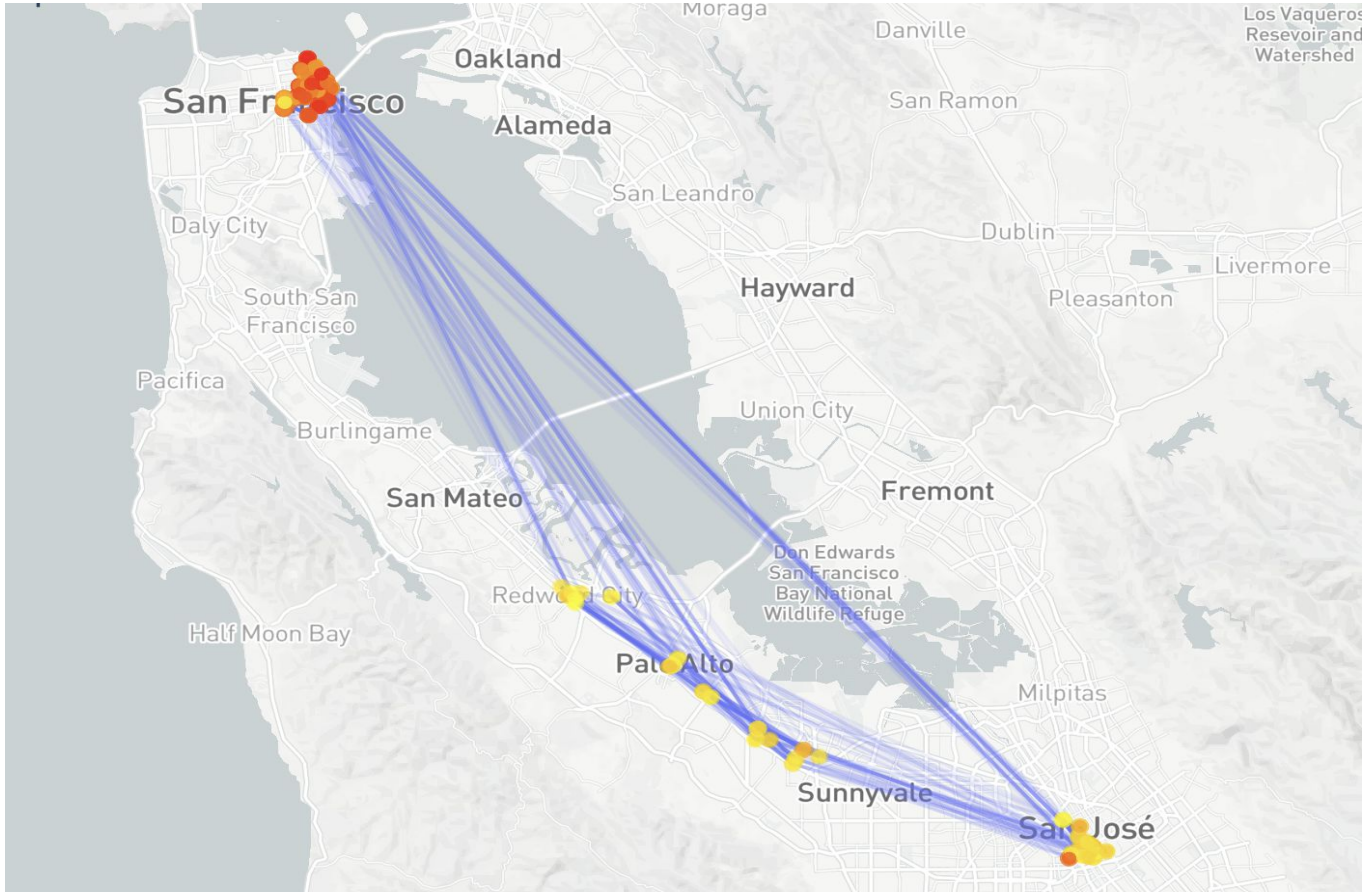
- For *Trips* data, group by date to calculate total num of trips per day
- Join *Trips* with *Weather* on the **date** column
- Aggregate joined RDD to show average num of trips for different temperature


Total Number of Bike Trips by Temperature
1200
1000
800
600
400
200
0
Count
40
45
50
55
60
65
70
75
Mean Temperature
Subscriber
Customer

Marine

- Group by *start_station_id* and *end_station_id* to count the amount of trips from unique start stations and end station key pairs
- Joined *start_station_id* and *end_station_id* with longitude and latitude

Link



trip>200: red
Otherwise: Yellow

Kevin

Examining average bike availability at each hour of the day on weekends and weekdays

- In the status dataset found the proportion of bikes available for every hour between 2014 and 2015.
- Grouped by a weekend day indicator column, hour and the name of the station
- Computed the average bike availability proportion for that hour grouped by the above variables
- link

Cluster setting and execution time comparison

	Kevin	Marine	Lexie	Esther	Akanksha
Instance Type	M5d.xlarge 16GiB memory 150 SSD GB storage	r5.8xlarge, 256 GiB memory	m4.4xlarge, 64 GiB memory	m4.4xlarge, 64 GiB memory	m4.4xlarge, 64 GiB memory
Number of Instances	1 master & 2 core nodes	1 master & 2 core nodes	1 master node & 2 core nodes	1 master node & 3 core nodes	1 master node & 4 core nodes
Execution Time	24 mins	5 mins 30s	9 mins 48s	7 mins 35s	10 mins 33 sec

Lessons Learned

- Most subscribers use SF bikes for commuting to work
- Subscribers overall take more shorted bike trips than customers.

- The optimal cluster setup depends on your data processing
- Using disk storage to shuffle data is a lot slower than memory