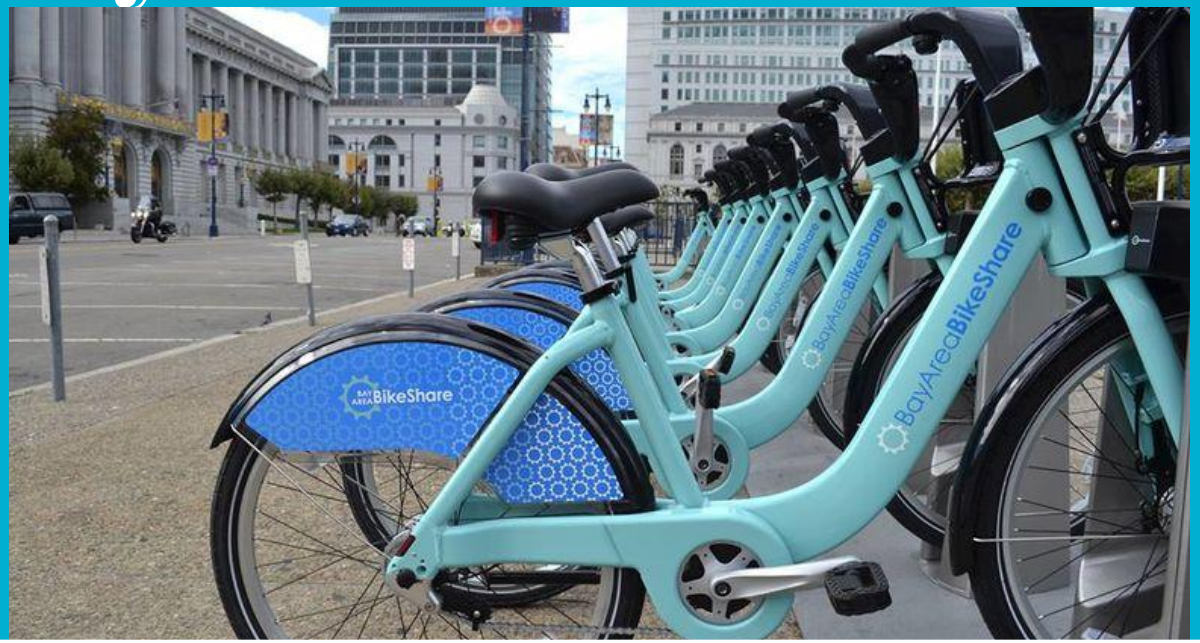
Bay Area Bike Share



Data Description

Total Size : 2 GB

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

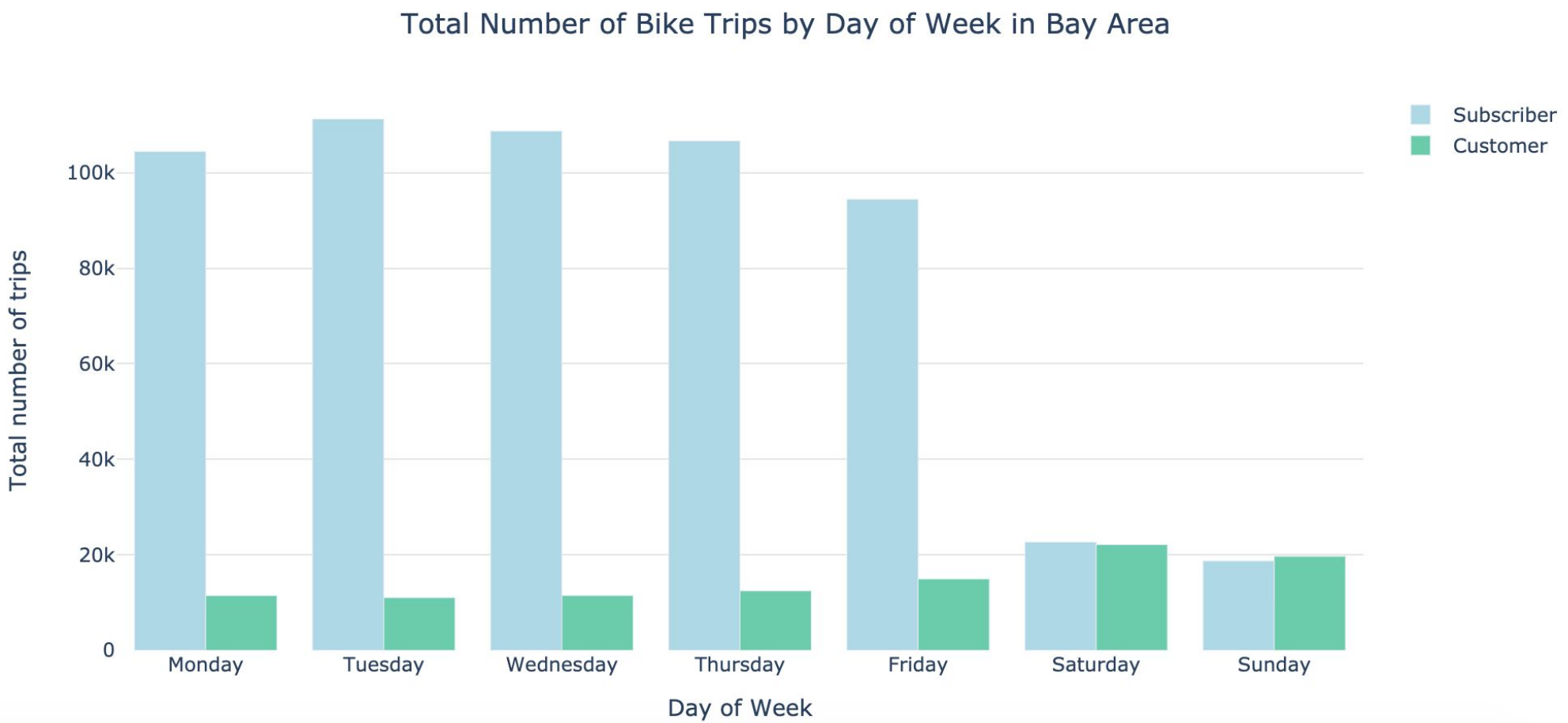
Status.csv : Data about the number of bikes and docks available at various stations at different time stamps

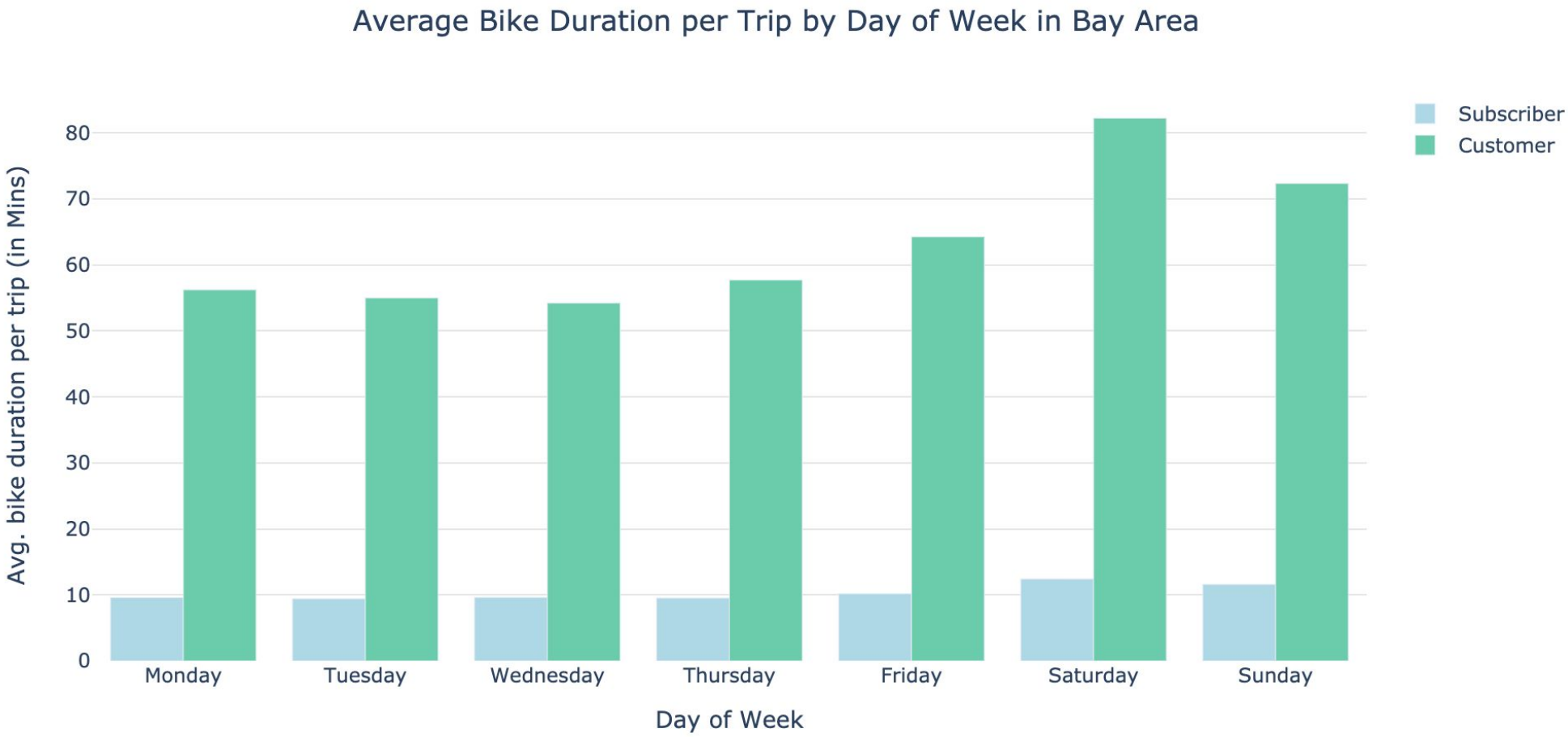
Trips.csv: Data about individual bike trips with details such as duration, subscription type, start station name, end station name, etc.

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

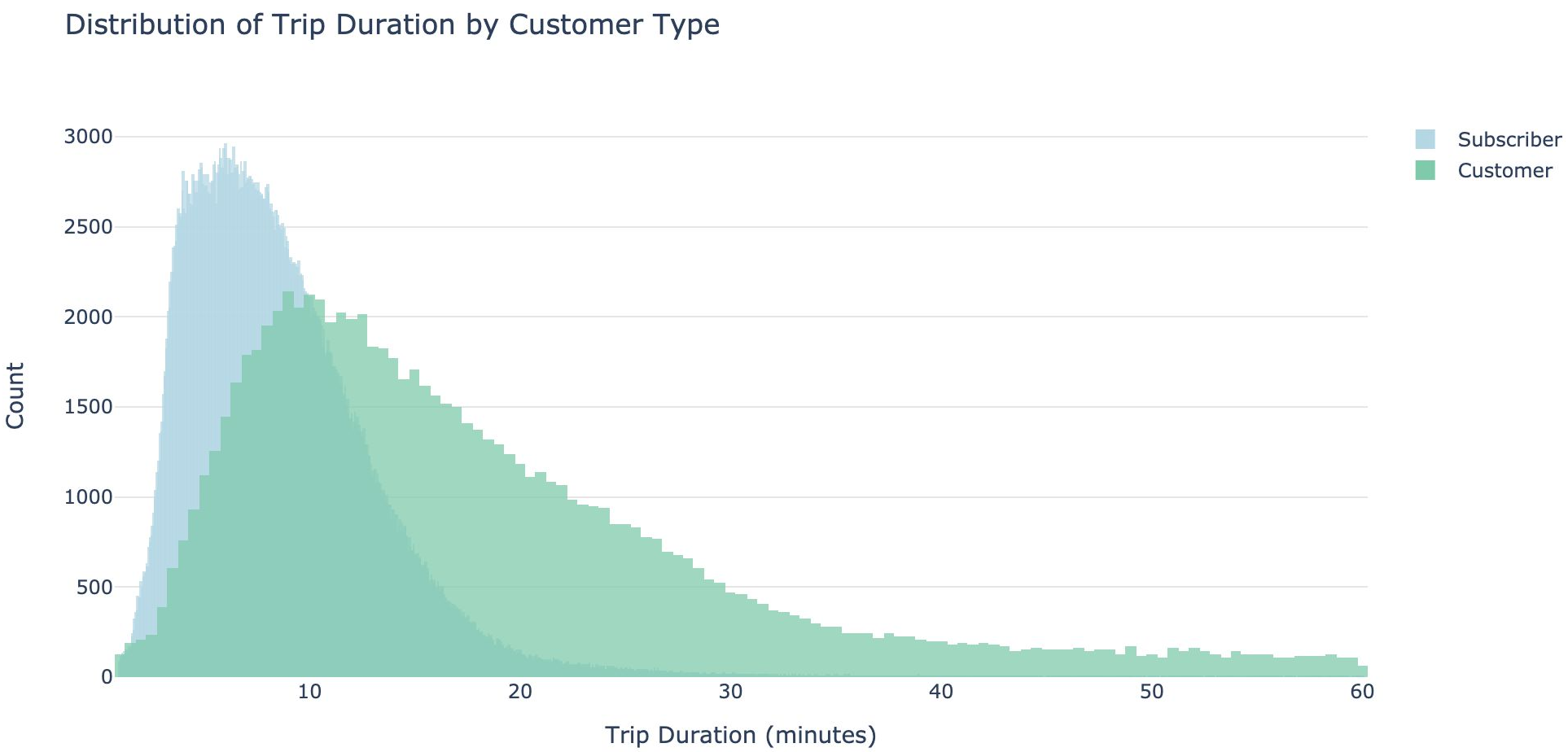
Preprocessing & Data Visualization

- 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

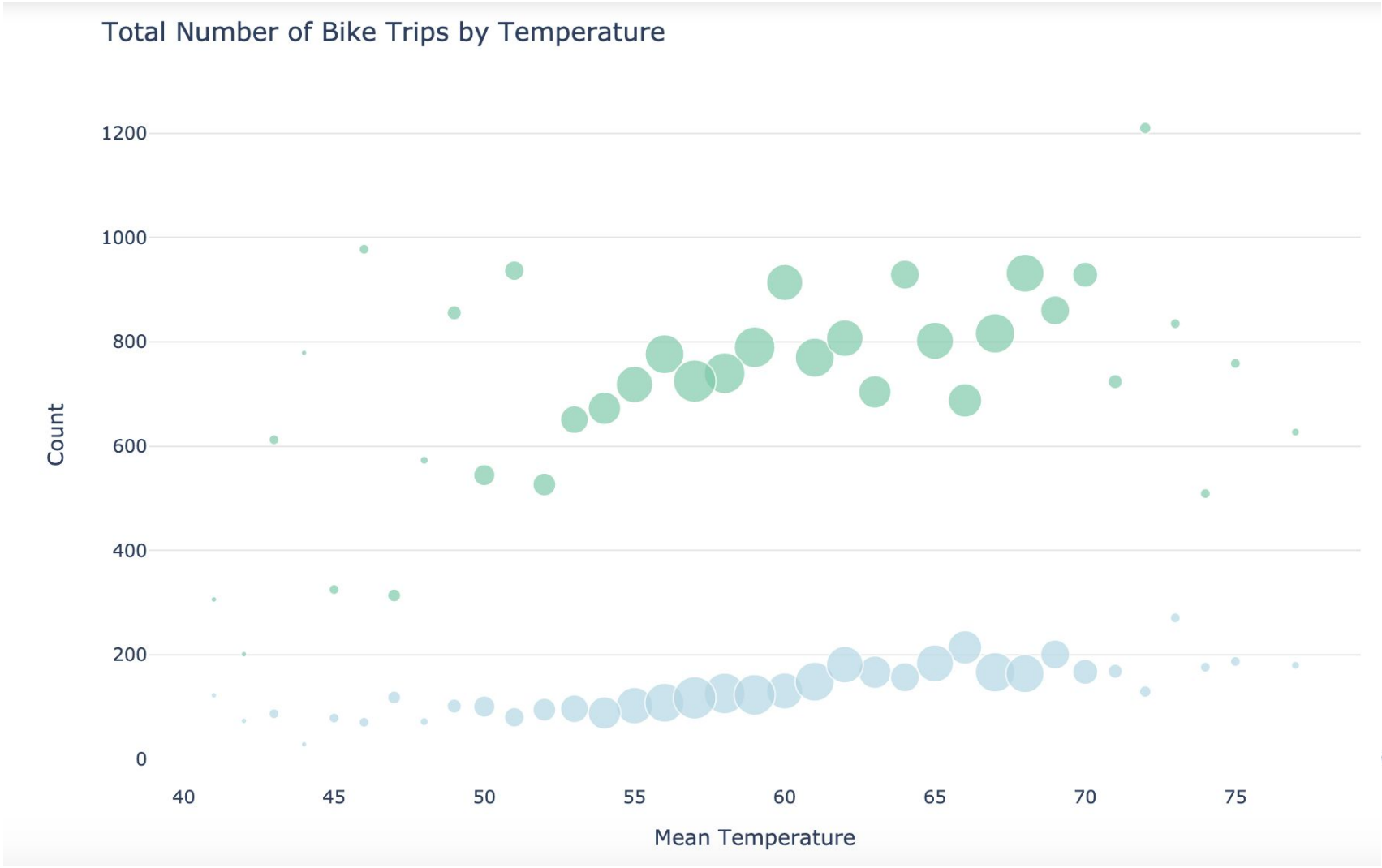

Total Number of Bike Trips by Day of Week in Bay Area
Subscriber
Customer
Total number of trips
100k
80k
60k
40k
20k
0
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Day of Week


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

- 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.

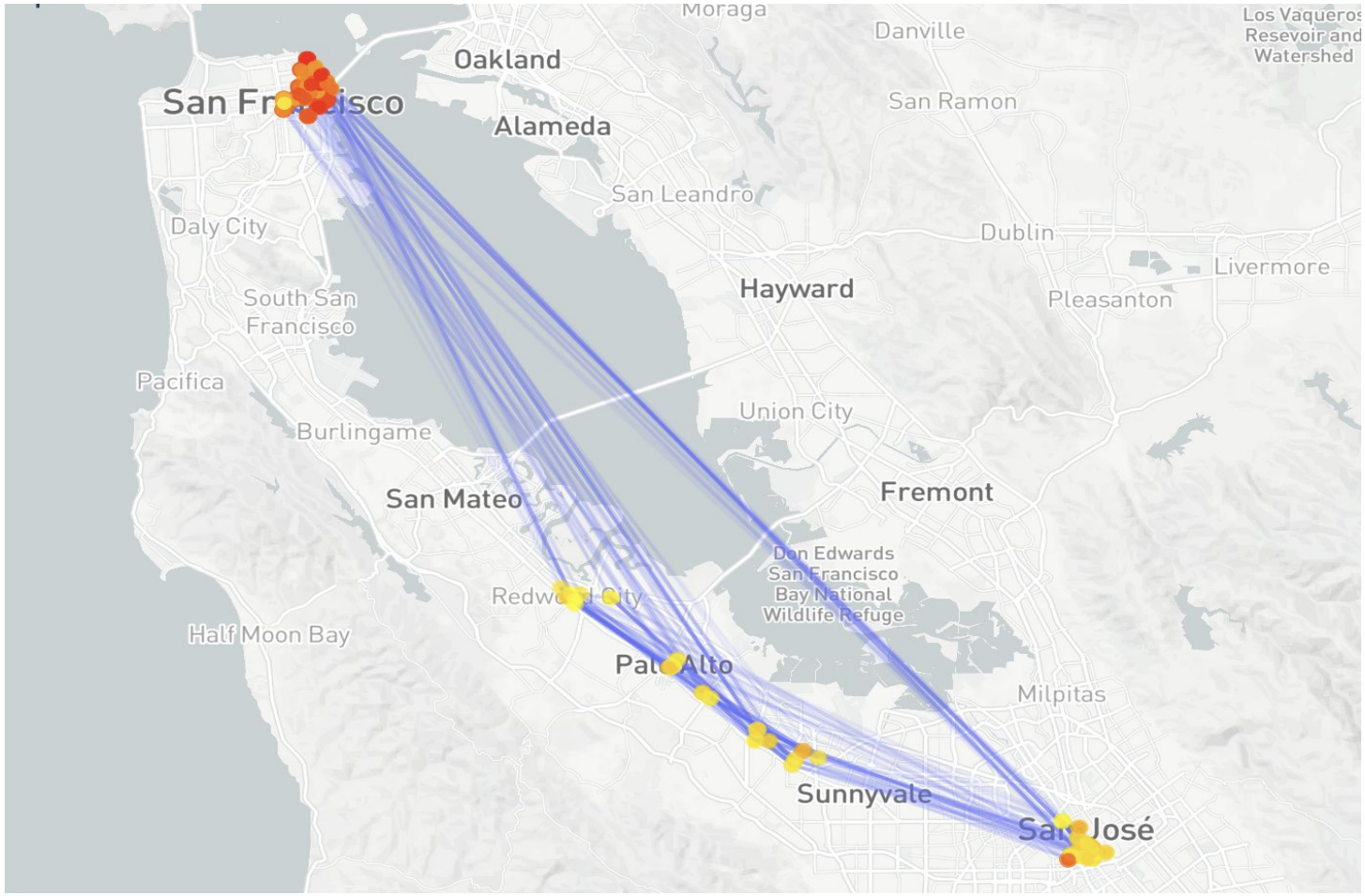

Distribution of Trip Duration by Customer Type
Subscriber
Customer
Count
3000
2500
2000
1500
1000
500
0
10
20
30
40
50
60
Trip Duration (minutes)

- 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

- 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

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

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

US Air Pollutants in the Last Decade

Contents

- Data Description
- Data Processing Goal
- Processing Outcome
- Cluster Setting and Execution Time Comparison
- Lesson Learned

DataSet 1 : HAP

Daily Hazardous Air Pollutants in United States from 1990-2017
Size: 2.3GB
Source:
https://www.kaggle.com/epa/hazardous-air-pollutants#epa_hap_daily_summary.csv

- Arsenic:
  - highly toxic in its inorganic form
  - chronic arsenic poisoning, skin lesions and skin cancer
- Lead:
  - once taken into the body, lead distributes throughout the body in the blood and is accumulated in the bones
  - nervous system, kidney function, immune system, reproductive and developmental systems and the cardiovascular system

Pollutant name quantity inspection site collection date

daily_HAPS_2019

Parameter Code	POC	Latitude	Longitude	Datum	Parameter Name	Sample Duration	Date Local	Units of Measure	Event Type	Observation Count	Observation Percent	Arithmetic Mean	1st Max Value	1st Max Hour	AQI	Method Code	Method Name	Local Site Name	Address	State Name	County Name	City Name	CBSA Name	Date of Last Change
43502	8	33.553056	-86.815	WGS84	Formaldehyde	8 HOUR	2019-06-14	Parts per billion Carbon	None	3	100	1.466667	1.8	20		202	SILICA-DNPH-CART	North Birmingham	NO. B'HAM,SOU R.R., 3009 28TH ST. NO.	Alabama	Jefferson	Birmingham	Birmingham-Hoover, AL	2019-09-24
43502	8	33.553056	-86.815	WGS84	Formaldehyde	8 HOUR	2019-06-20	Parts per billion Carbon	None	3	100	1.3	1.7	12		202	SILICA-DNPH-CART	North Birmingham	NO. B'HAM,SOU R.R., 3009 28TH ST. NO.	Alabama	Jefferson	Birmingham	Birmingham-Hoover, AL	2019-09-24
43502	8	33.553056	-86.815	WGS84	Formaldehyde	8 HOUR	2019-06-23	Parts per billion Carbon	None	3	100	1.3	3.7	4		202	SILICA-DNPH-CART	North Birmingham	NO. B'HAM,SOU R.R., 3009 28TH ST. NO.	Alabama	Jefferson	Birmingham	Birmingham-Hoover, AL	2019-09-24

DataSet 2 : AQI

Daily AQI Dataset in United States from 1980-2019
Size: 915.85 MB
Source: https://aqs.epa.gov/aqsweb/airdata/download_files.html#AQI

- AQI: how polluted the air currently is, or how polluted it is forecast to become
- High AQI means high level of air pollution

AQI

State Name	county Name	State Code	County Code	Date	AQI	Category	Defining Parameter	Defining Site	Number of Sites Reporting
Alabama	Baldwin	1	3	2007-01-03	55	Moderate	PM2.5	01-003-0010	1
Alabama	Baldwin	1	3	2007-01-06	23	Good	PM2.5	01-003-0010	1
Alabama	Baldwin	1	3	2007-01-09	13	Good	PM2.5	01-003-0010	1

DataSet 3 : Business

Yearly Business Pattern in United States from 1986-2017
Size: 84.2MB
Source:https://www.census.gov/programs-surveys/cbp/data/datasets.html

busPattern

Geographic area name	NAICS code	Meaning of NAICS code	Year	Number of establishments	Paid employees for pay period including March 12 (number)	county_name	state_name
Autauga County, Alabama	0	Total for all sectors	2005	872	10491	Autauga	Alabama
Autauga County, Alabama	11	Agriculture, forestry, fishing and hunting	2005	10	b	Autauga	Alabama
Autauga County, Alabama	21	Mining	2005	4	b	Autauga	Alabama

Data Fusion: join the above 3 datasets from 2005-2016

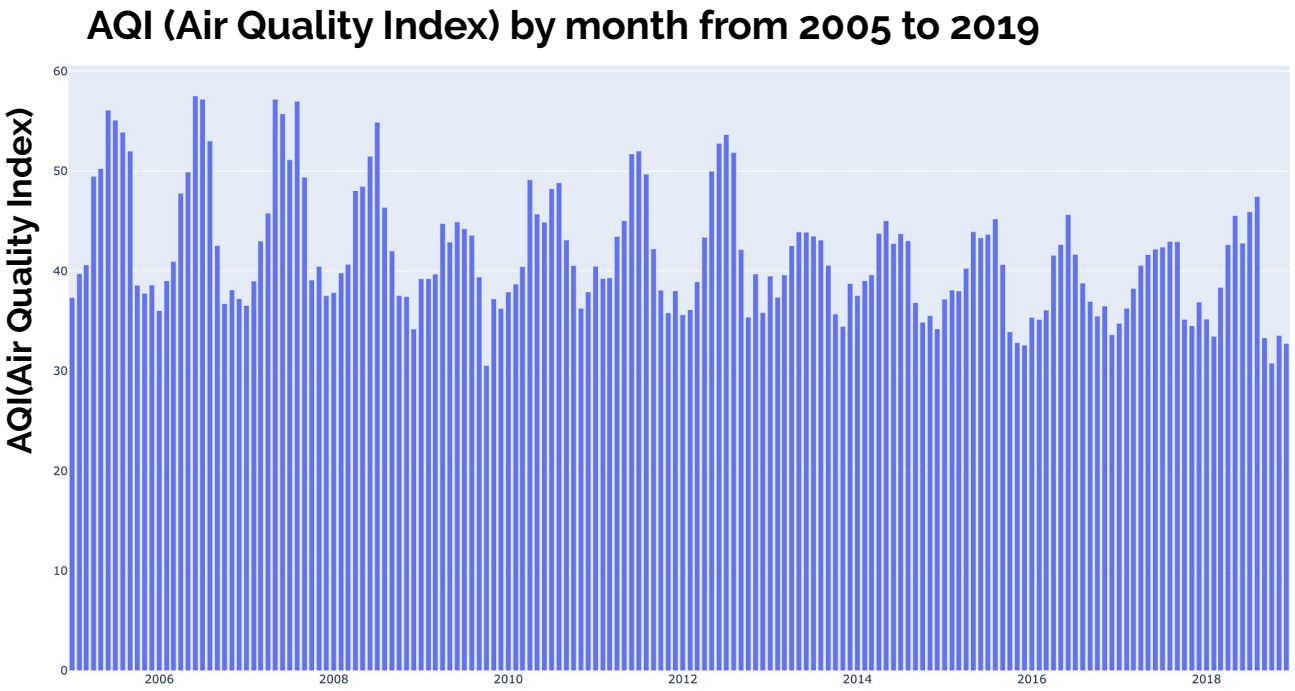
Data Processing Goal

- Investigate Air Pollutants By **Time**
  - Explore air quality and hazardous air pollution over years
- Investigate Air Pollutants By **County**
  - Explore air quality and hazardous air pollution across US
- Investigate Air Pollutants By **Industry**
  - Explore the relationship of business patterns and air quality

Processing Outcome: AQI (Air Quality Index) By Time

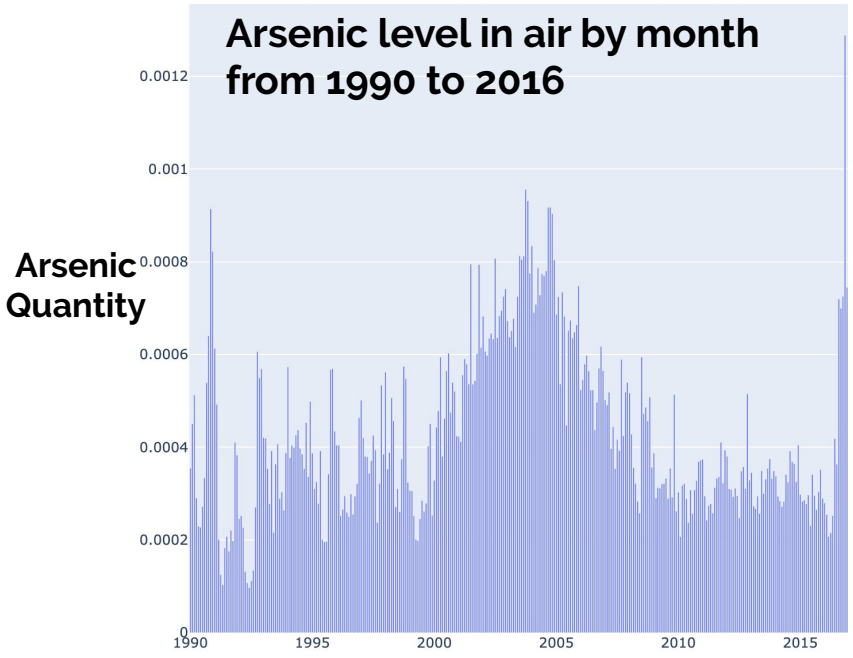
Air Quality Index:

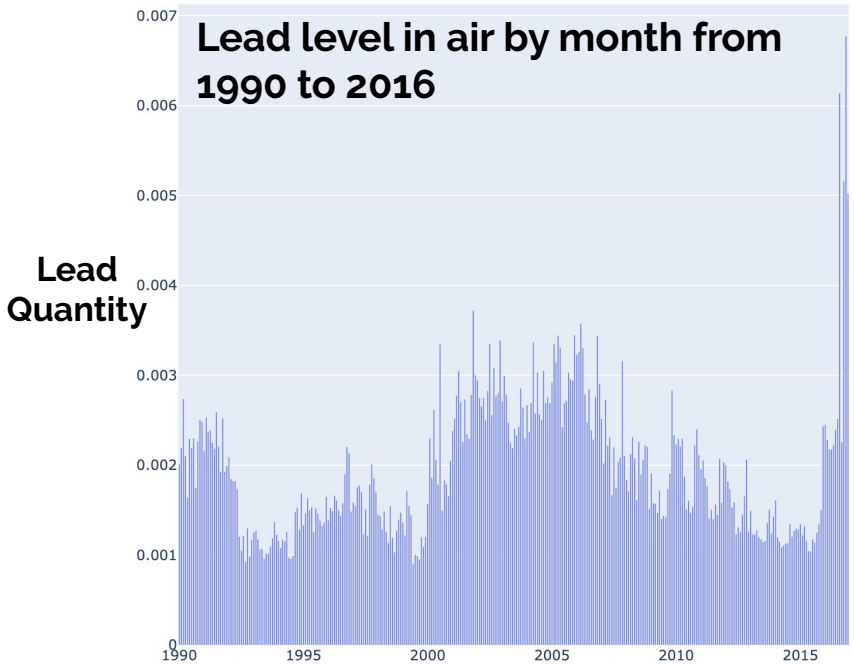
- obvious seasonal behavior
- yearly average AQI decreased from 2005 to 2019



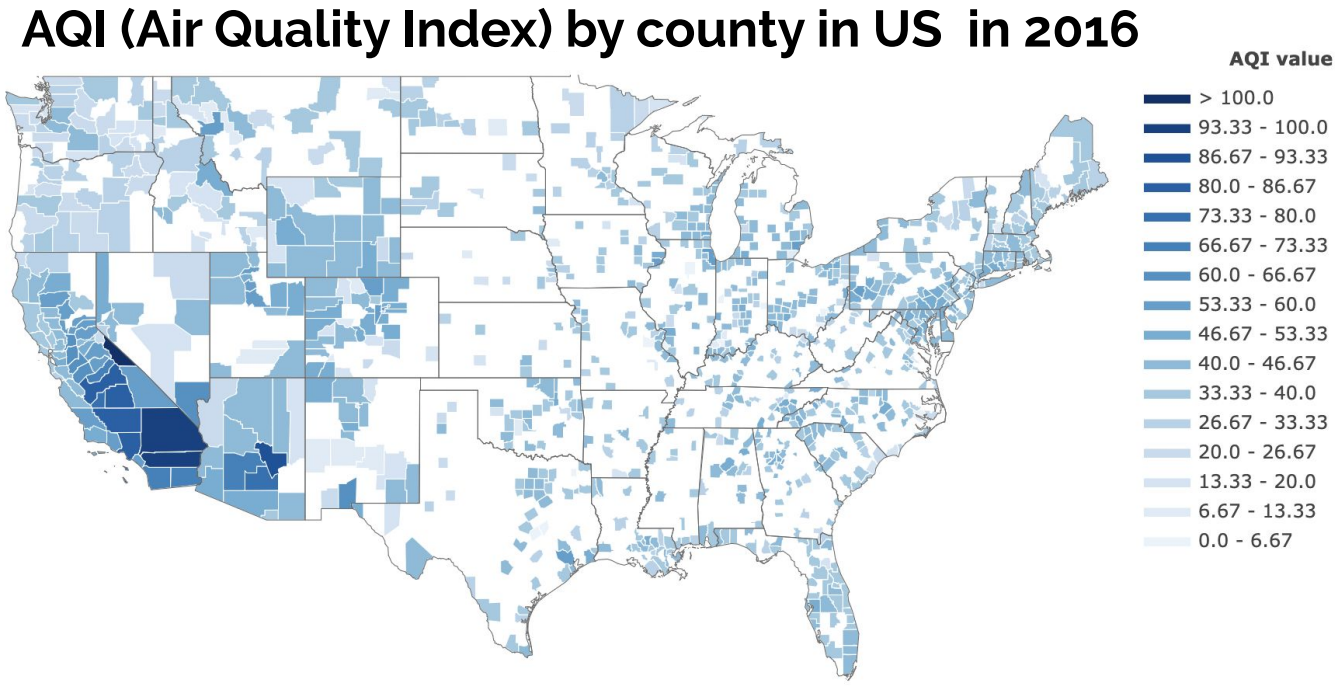
Processing Outcome: Arsenic and Lead By time

- increased from 1990 to a peak in 2005
- increased rapidly in 2016





Processing Outcome: AQI (Air Quality Index) By County



Best 5:

- Robertson, Texas
- Titus, Texas
- Morgan, Ohio
- Beaufort, North Carolina
- Milam, Texas

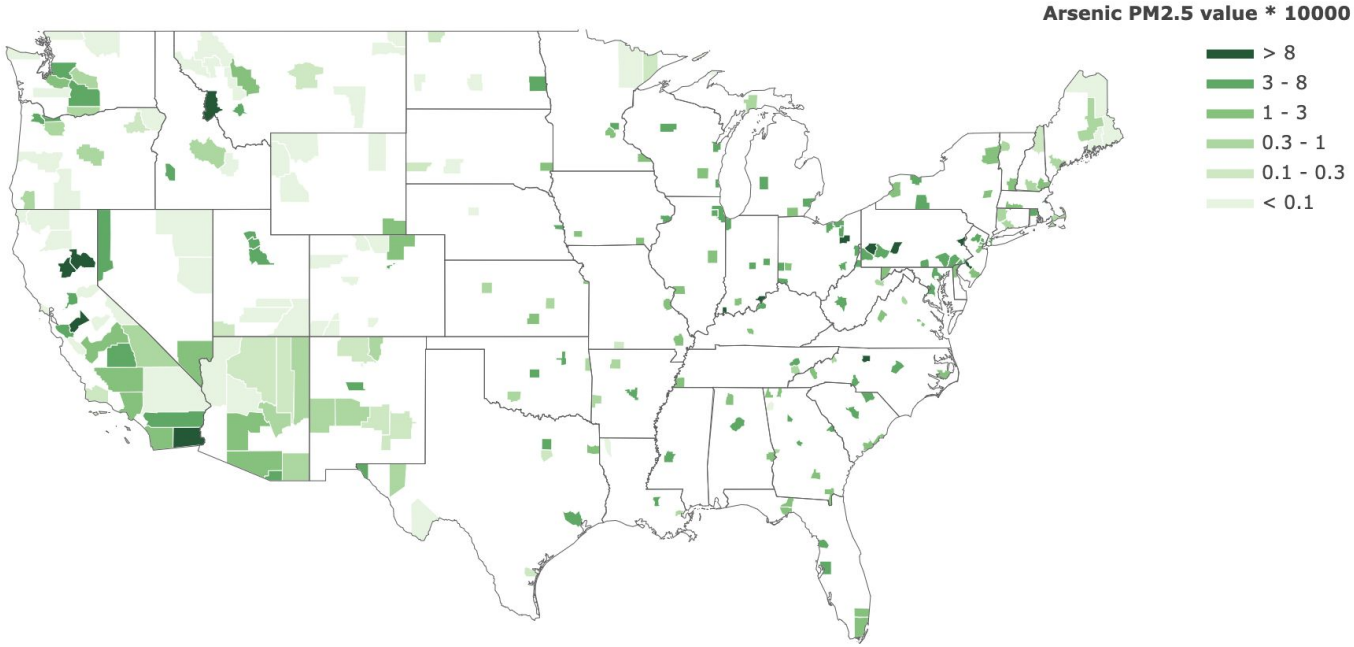
Worst 5:

- Mono, California
- San Bernardino, California
- Riverside, California
- Gila, Arizona
- Los Angeles, California

Processing Outcome: Arsenic PM2.5 By County

Arsenic PM2.5 by county in US in 2016

Unit: Micrograms/cubic meter (LC)



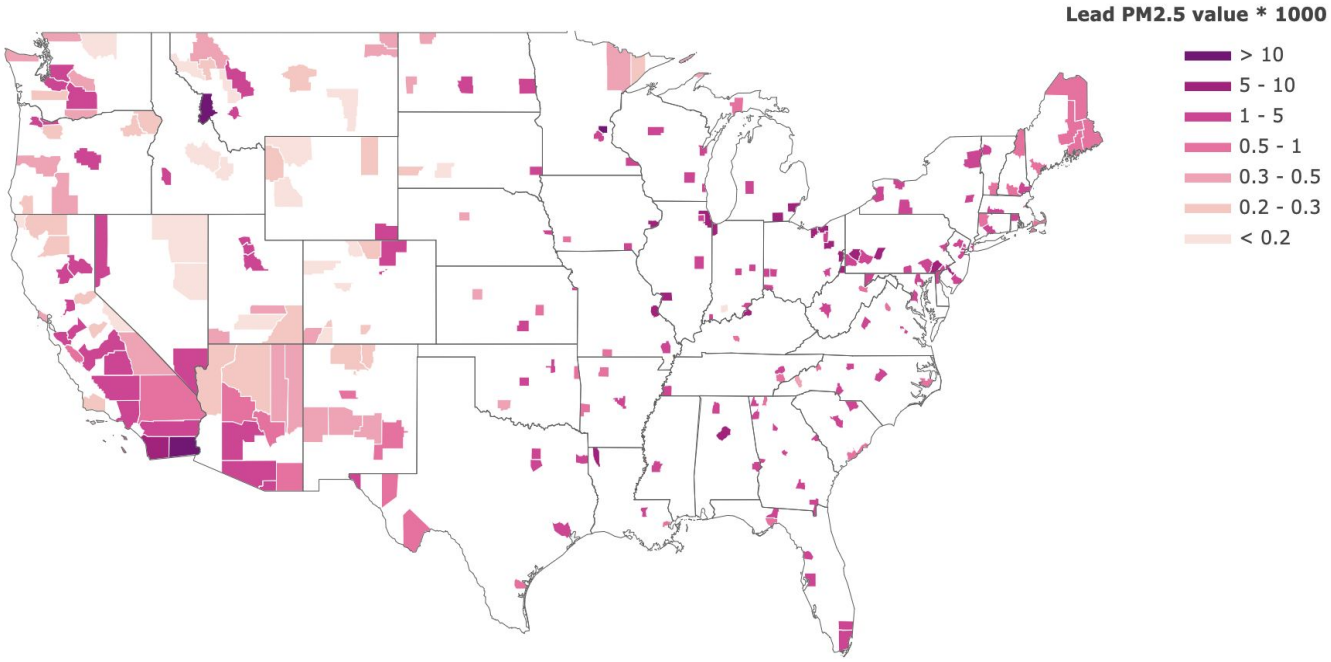
Worst 5:

- Ravalli, Montana
- Allegheny, Pennsylvania
- Stark, Ohio
- Vanderburgh, Indiana
- Clark, Indiana

Processing Outcome: Lead PM2.5 By County

Lead PM2.5 by county in US in 2016

Unit: Micrograms/cubic meter (LC)

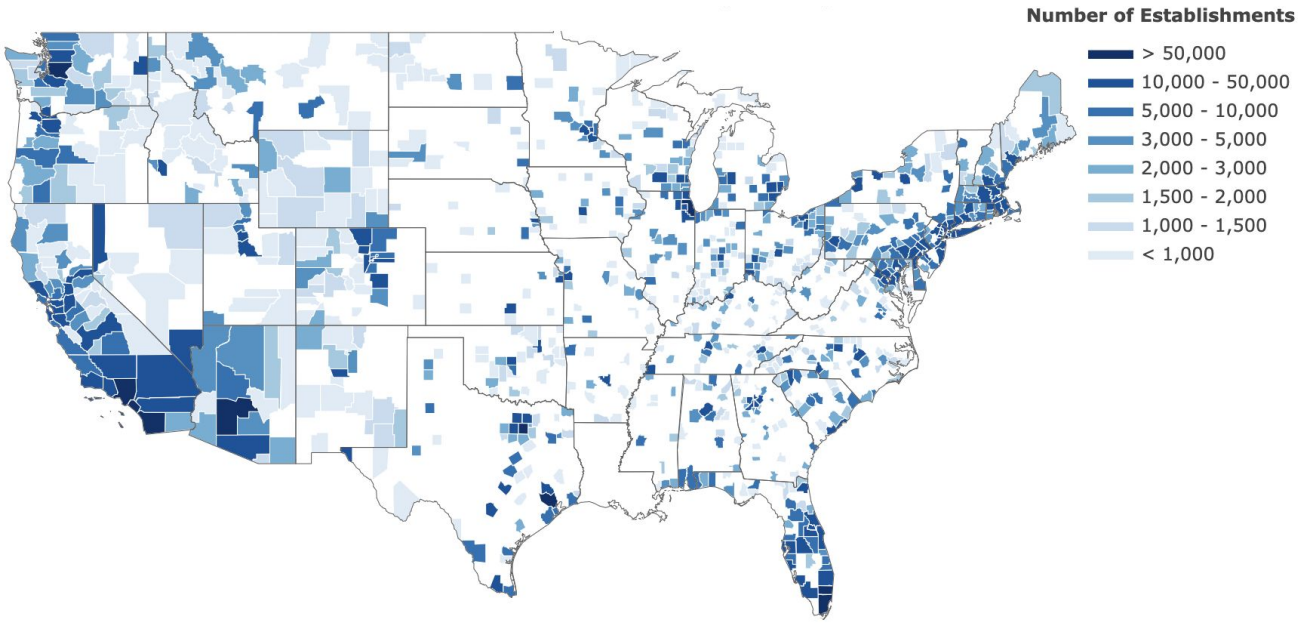


Worst 5:

- Ravalli, Montana
- Imperial, California
- Anoka, Minnesota
- Allegheny, Pennsylvania
- Cambria, Pennsylvania

Processing Outcome: Business Establishment By County

Number of Establishment by county in US in 2016



Largest 5:

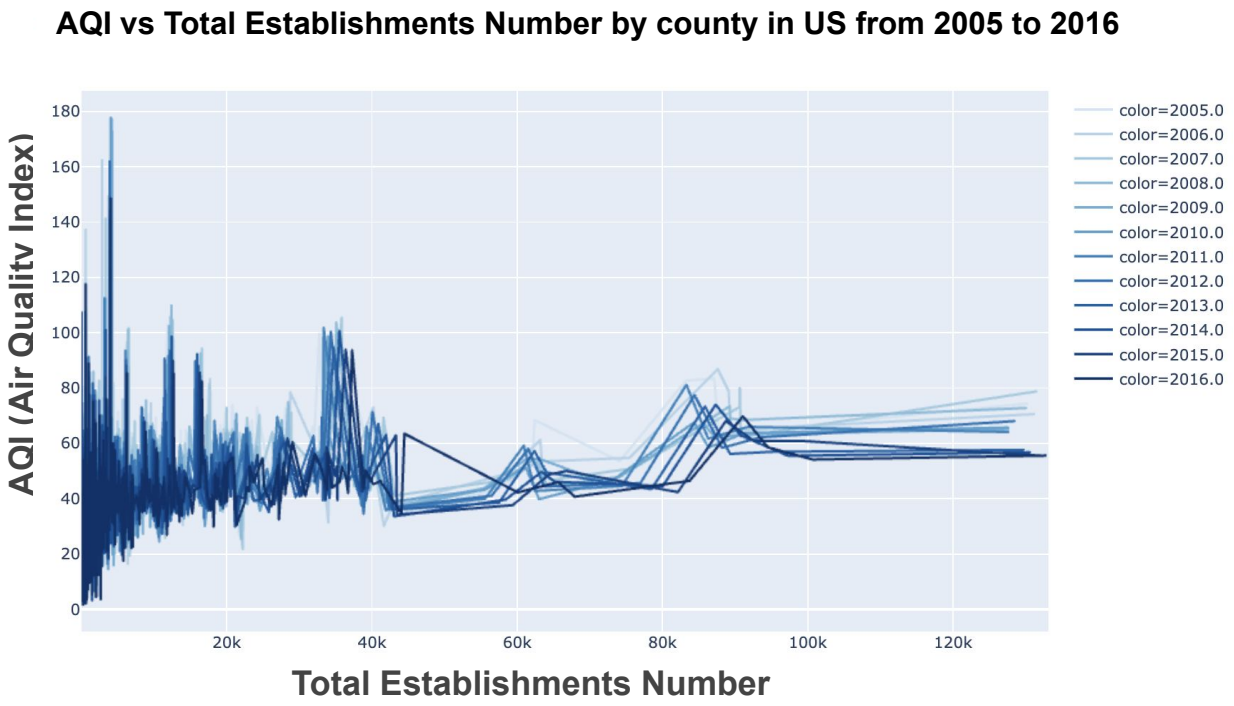
- Los Angeles, California: 269,489
- Cook, Illinois: 133150
- New York, New York: 104691
- Harris, Texas: 100884
- Orange, California: 94703

Smallest 5:

- Thomas. Nebraska: 23
- Alpine, California: 35
- Oliver, North Dakota: 42
- Jackson, South Dakota: 49
- Culberson, Texas: 54

Processing Outcome: Business Establishments & AQI

- Air quality index increases with number of establishments.



Cluster Setting & Execution Time Comparison

Cluster Setting		Execution Time (in seconds)
machine specs	number of nodes	
m3.xlarge	3	318.74
m4.xlarge	3	272.42
m5a.xlarge	3	248.04
m5.xlarge	3	190.36
m5.xlarge	4	159.67

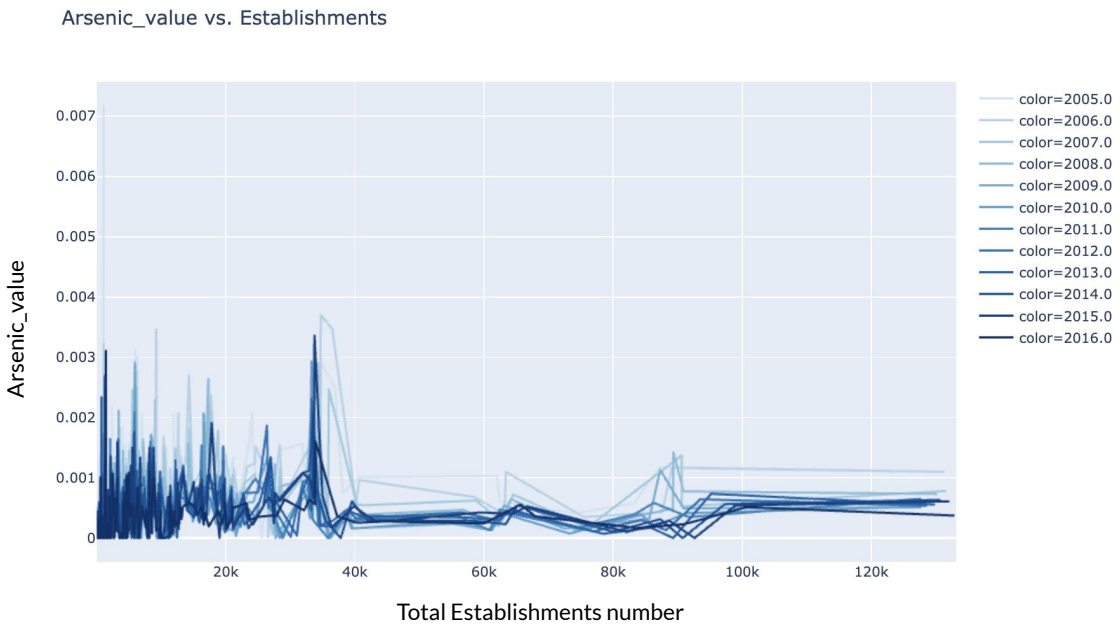
Execution time decreases

Lesson Learned

- Air quality is getting better over the years with seasonality.
- Air quality and pollutant level and establishment numbers are related.
- Distributed computing allows us to process large dataset.
- The larger machine specs of AWS EMR, the shorter execution time.
- The more cluster nodes of AWS EMR, the shorter execution time.

Future Work

- Number of total establishments does not have strong impact on Arsenic/Lead level
- Is it possible that specific type of establishment has high impact on these hazardous air pollutants? E.g. Mining, Agricultural…
- Can we predict air pollutants level?

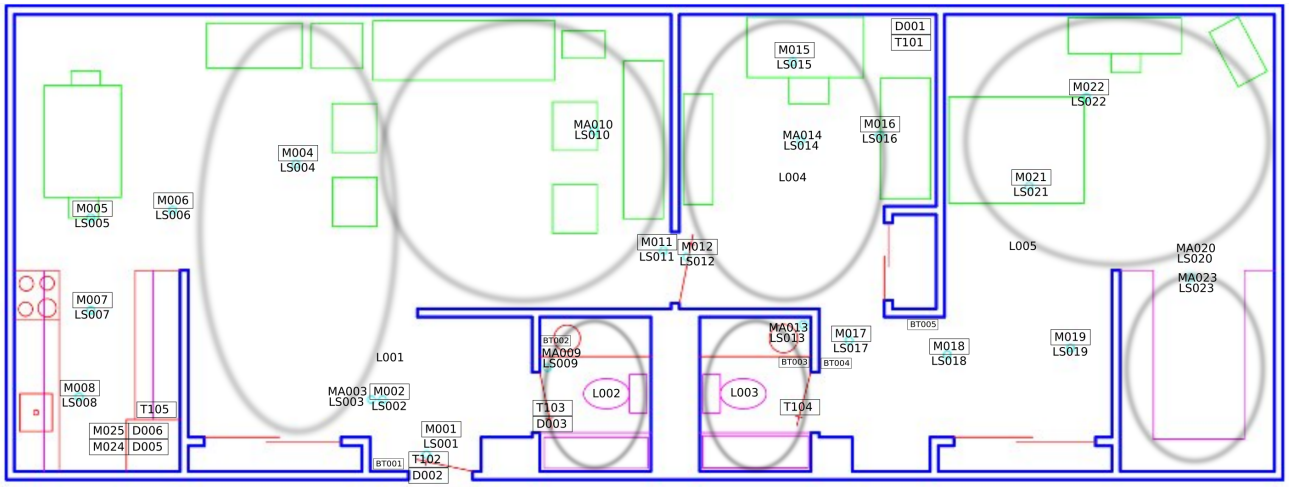


Thank you!

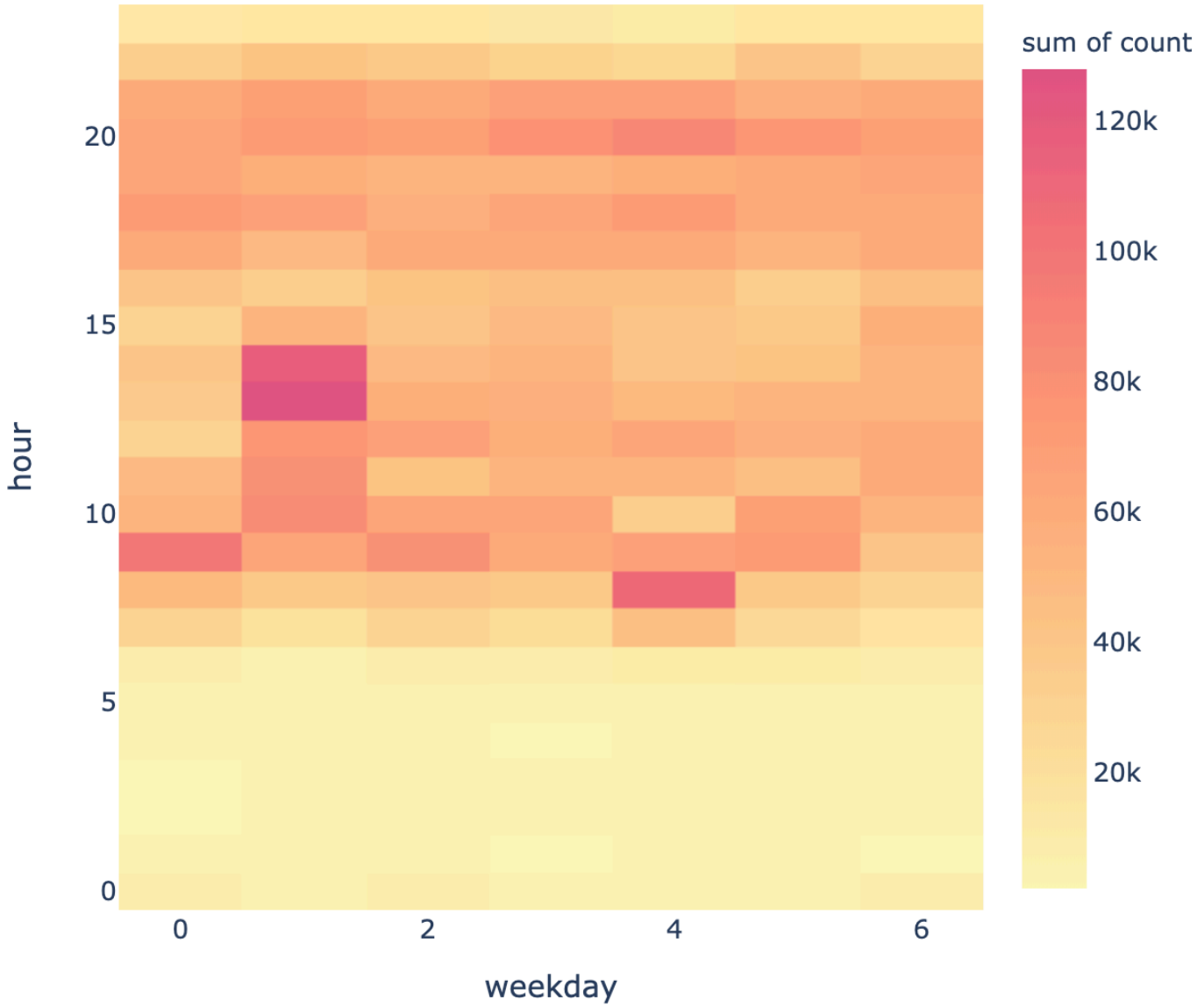
From Sensors to Smart Homes

Data:

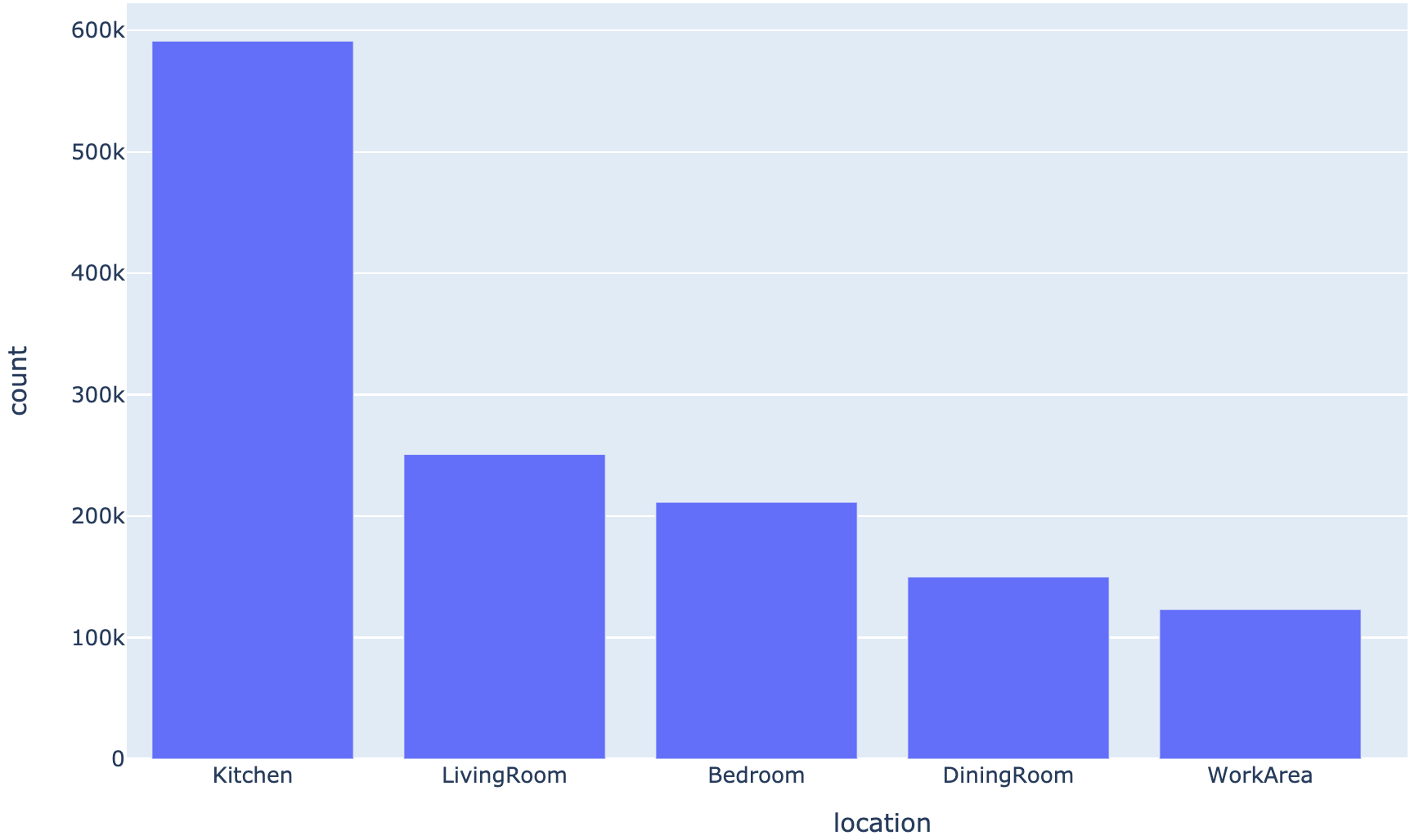
- 29 Homes
- 2 Years of Collection
- Sensors are always on and communicate to the server
- Motion sensors, Door / Temperature sensors, and Light Switch sensors
- Labels are the activity type
- Multi_resident



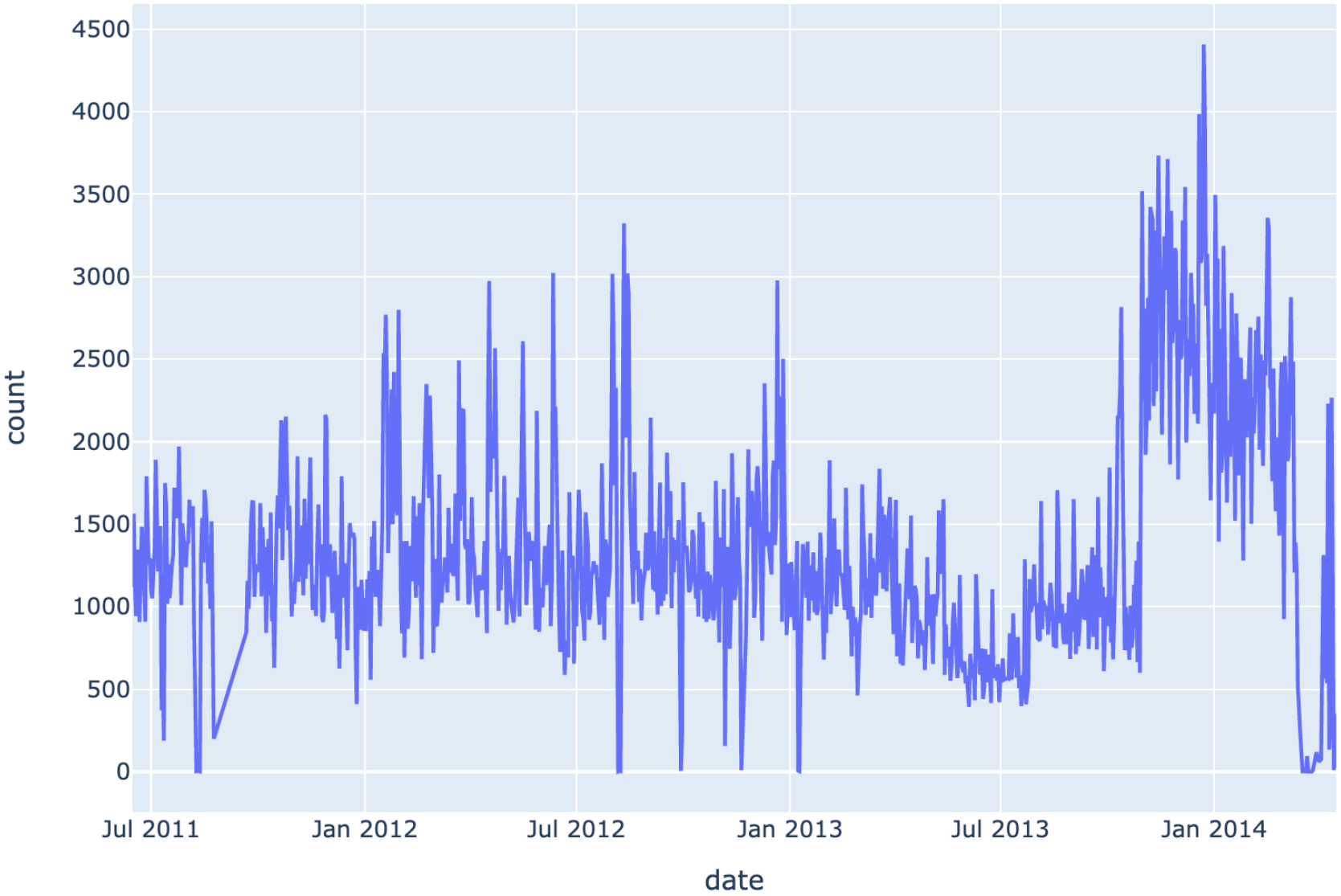
Overall Shape of The Motion:



Sensor Sensor on the wall...



Tell Me When I am Gone!

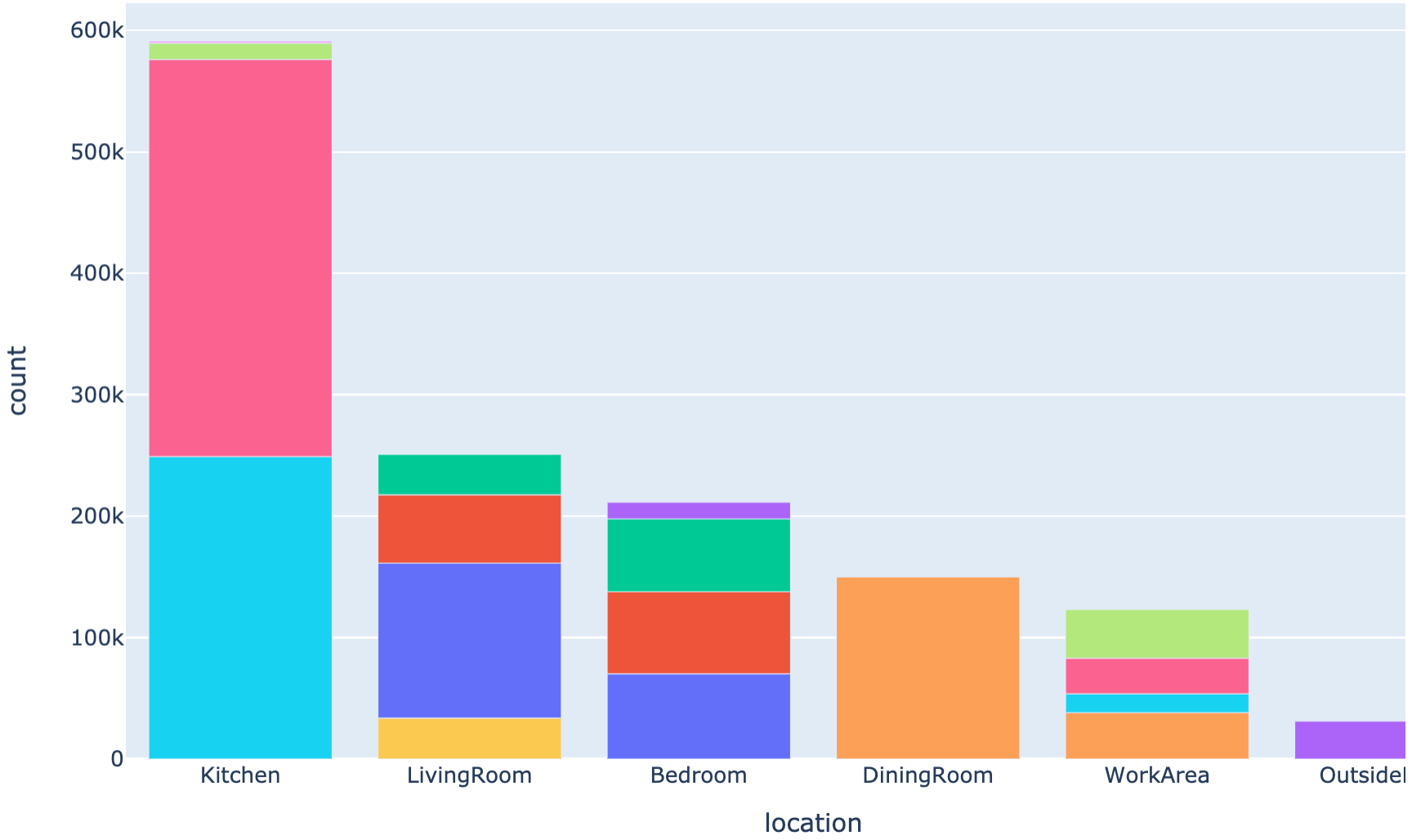


2011-06-15

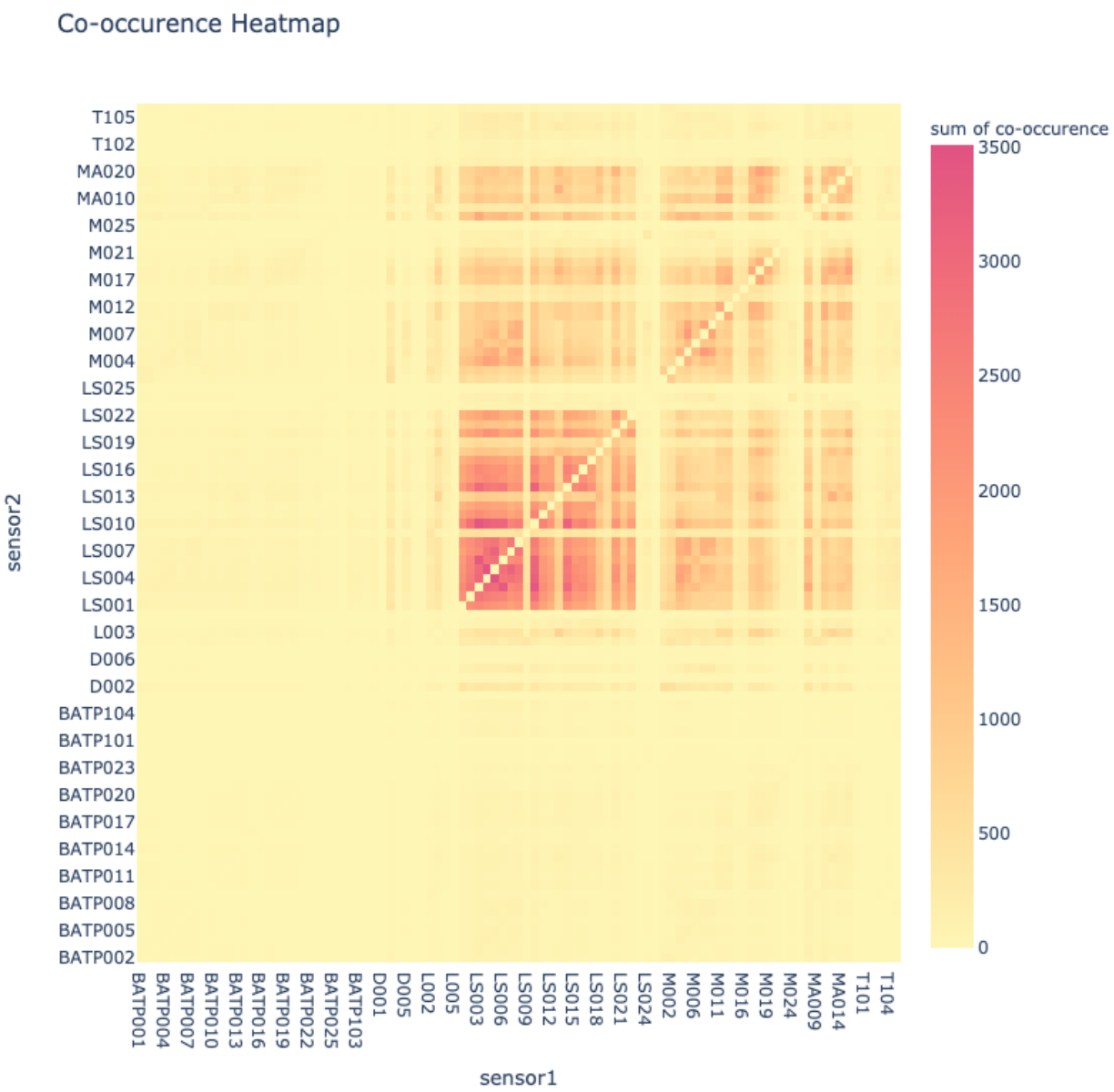
09:58:45.585184

We are able to precisely detect movement, but it’s no surprise as we have put too many sensors in each home.

Is This Even Smart?

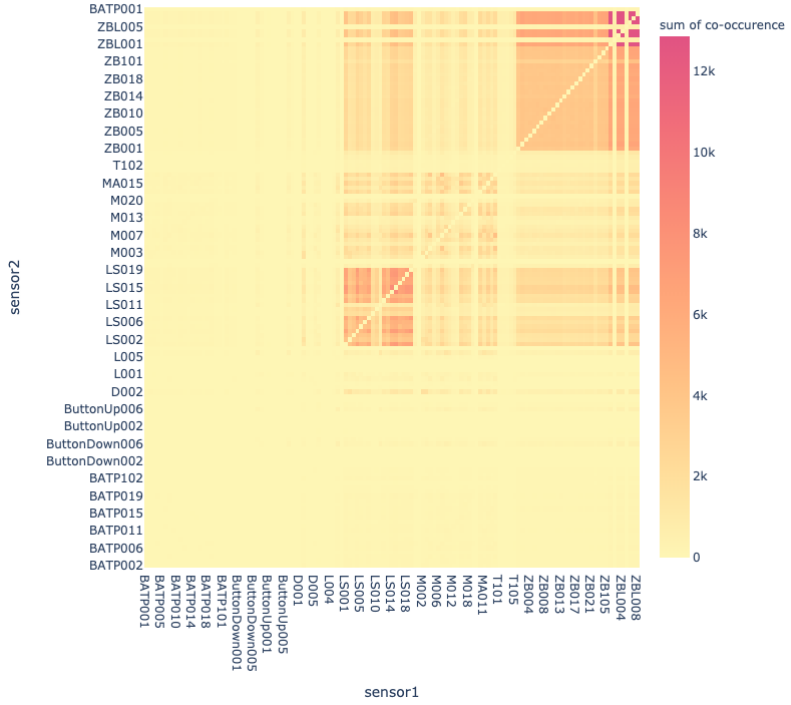


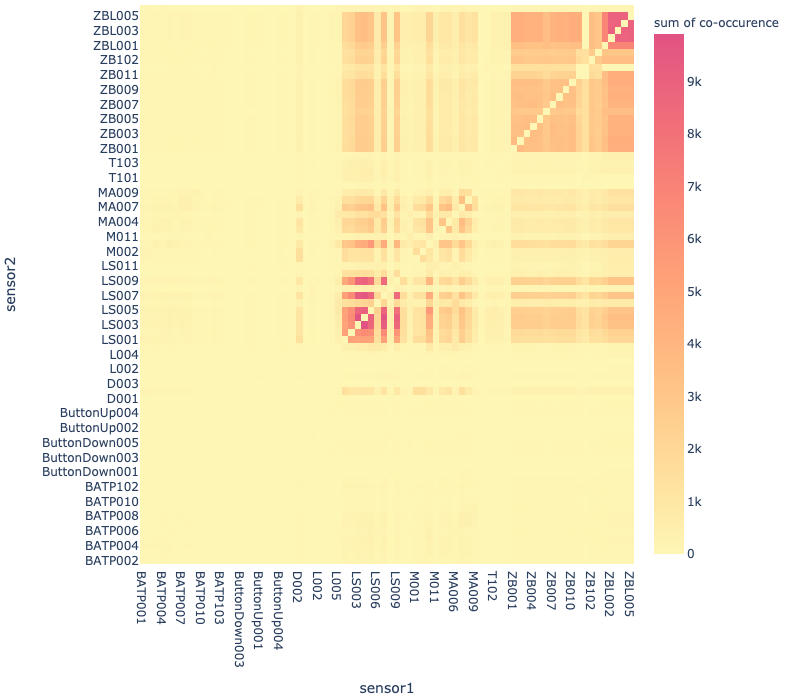
Is there a correlation between sensors, and can we use less of them?


Co-occurence Heatmap
sensor2
T105
T102
MA020
MA010
M025
M021
M017
M012
M007
M004
LS025
LS022
LS019
LS016
LS013
LS010
LS007
LS004
LS001
L003
D006
D002
BATP104
BATP101
BATP023
BATP020
BATP017
BATP014
BATP011
BATP008
BATP005
BATP002
BATP001
BATP004
BATP007
BATP010
BATP013
BATP016
BATP019
BATP022
BATP025
BATP103
D001
D005
L002
L005
LS003
LS006
LS009
LS012
LS015
LS018
LS021
LS024
M002
M006
M011
M016
M019
M024
MA009
MA014
T101
T104
sensor1
sum of co-occurence
3500
3000
2500
2000
1500
1000
500
0

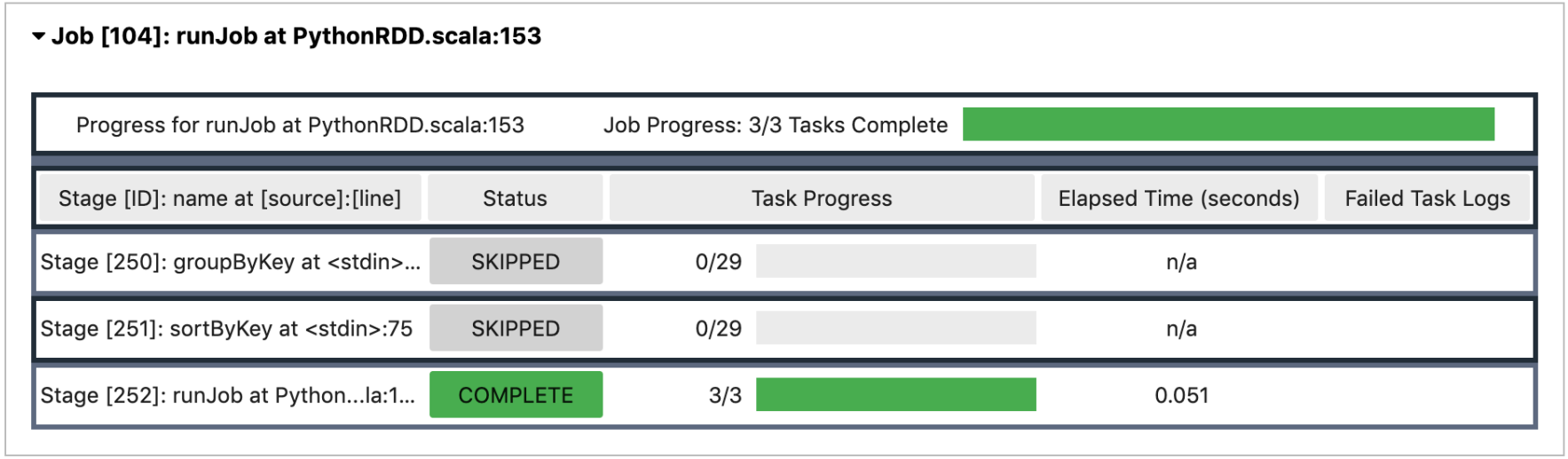

sum of co-occurence
5000
4000
3000
2000
1000
0
sensor2
ZBL010
ZBL005
ZB106
ZB101
ZB022
ZB018
ZB014
ZB010
ZB006
ZB002
T104
MA024
M025
M017
M013
M009
M005
M001
LS022
LS018
LS014
LS010
LS006
LS002
L007
L003
D005
D001
ButtonUp006
ButtonDown010
ButtonDown005
BATP106
BATP101
BATP022
BATP018
BATP014
BATP010
BATP006
BATP002
BATP001
BATP006
BATP011
BATP016
BATP021
BATP101
ButtonDown001
ButtonDown007
ButtonUp010
ButtonUp004
D005
L004
L009
LS005
LS010
LS015
LS020
LS025
M005
M010
M015
M021
MA024
T105
ZB004
ZB009
ZB014
ZB019
ZB024
ZB105
ZBL005
sensor1


sum of co-occurence
3500
3000
2500
2000
1500
1000
500
0
sensor2
T106
T103
MA022
MA017
M027
M024
M020
M012
M009
M006
M003
LS027
LS024
LS021
LS018
LS015
LS012
LS009
LS006
LS003
L006
L003
D006
D003
BATP106
BATP102
BATP026
BATP023
BATP020
BATP017
BATP014
BATP011
BATP008
BATP005
BATP002
BATP001
BATP004
BATP007
BATP010
BATP013
BATP016
BATP019
BATP022
BATP025
BATP101
BATP104
D002
D005
L002
L005
LS002
LS005
LS008
LS011
LS014
LS017
LS020
LS023
LS026
M002
M005
M008
M011
M016
M023
M026
MA015
MA019
T102
T105
sensor1


sum of co-occurence
12k
10k
8k
6k
4k
2k
0
sensor2
BATP001
ZBL005
ZBL001
ZB101
ZB018
ZB014
ZB010
ZB005
ZB001
T102
MA015
M020
M013
M007
M003
LS019
LS015
LS011
LS006
LS002
L005
L001
D002
ButtonUp006
ButtonUp002
ButtonDown006
ButtonDown002
BATP102
BATP019
BATP015
BATP011
BATP006
BATP002
BATP001
BATP005
BATP010
BATP014
BATP018
BATP101
ButtonDown001
ButtonDown005
ButtonUp001
ButtonUp005
D001
D005
L004
L001
LS001
LS005
LS010
LS014
LS018
M002
M006
M012
M018
MA011
T101
T105
ZB004
ZB008
ZB013
ZB017
ZB021
ZB105
ZBL004
ZBL008
sensor1


sum of co-occurence
9k
8k
7k
6k
5k
4k
3k
2k
1k
0
sensor2
ZBL005
ZBL003
ZBL001
ZB102
ZB011
ZB009
ZB007
ZB005
ZB003
ZB001
T103
T101
MA009
MA007
MA004
M011
M002
LS011
LS009
LS007
LS005
LS003
LS001
L004
L002
D003
D001
ButtonUp004
ButtonUp002
ButtonDown005
ButtonDown003
ButtonDown001
BATP102
BATP010
BATP008
BATP006
BATP004
BATP002
BATP001
BATP004
BATP007
BATP010
BATP103
ButtonDown003
ButtonDown001
ButtonUp001
ButtonUp004
D002
L002
L005
LS003
LS006
LS009
M001
M011
MA006
MA009
T102
ZB001
ZB004
ZB007
ZB010
ZB102
ZBL002
ZBL005
sensor1

Running 104 jobs to precess the data



Job [104]: runJob at PythonRDD.scala:153

Progress for runJob at PythonRDD.scala:153 — Job Progress: 3/3 Tasks Complete

Stage [ID]: name at [source]:[line]	Status	Task Progress	Elapsed Time (seconds)	Failed Task Logs
Stage [250]: groupByKey at <stdin>...	SKIPPED	0/29	n/a	
Stage [251]: sortByKey at <stdin>:75	SKIPPED	0/29	n/a	
Stage [252]: runJob at Python...la:1...	COMPLETE	3/3	0.051	

	2 Workers	**4 Workers**
m5.xlarge	8 mins	5 mins
m5.2xlarge	5 mins	3 mins

Conclusion:

As we can see some of these sensors are highly correlated.

We might be able to reduce the sensors, and still be able to reach the business goal, which is classifying the action type.

We can use PCA or feature importance further to see if removing some of this sensors will hurt the outcome.