USF

Big Data
Project-2
Climate Analysis with MapReduce
Report

Hiep Bui & Alper Ozdamar

USF

Document Identification

Document Type	Interface Specification
Customer Name	Big Data DFS Core
Document Location	

Document Version Control

Version	Date	Created By	Change	Approved By
0.1	11.9.2019	Alper Ozdamar & Hiep Bui	First draft for Big Data Map Reduce Report	
0.2	11.10.2019	Alper Ozdamar & Hiep Bui	Edit caption and layout	
1.0	11.13.2019	Hiep Bui	Update doc base on new dataset for MovingOut and SolarWind problem.	
1.1	11.16.2019	Alper Ozdamar	Minor updates. Finalized.	

USF

Table of Contents

Table of Figures

1. References

1. Project-2 Spec : https://www.cs.usfca.edu/~mmalensek/cs677/assignments/project-2.html
2. NOAA : https://www.ncdc.noaa.gov/data-access/land-based-station-data/land-based-datasets/us-climate-reference-network-uscrn
3. NCDC Data Dictionary : https://www.cs.usfca.edu/~mmalensek/cs677/assignments/ncdc-data.html
4. USGS : https://earthquake.usgs.gov
5. Pearson Correlation Coefficient Calculator : https://www.socscistatistics.com/tests/pearson/default2.aspx
6. U.S. Climate data: https://www.usclimatedata.com/climate/united-states/us
7. CustomWeather: https://www.timeanddate.com/weather/usa/san-francisco/historic?month=12&year=2018
8. Climate Comfort Index : *https://www.oxfordreference.com/view/10.1093/oi/authority.20110803095626624*

2. Glossary

3. Abbreviations

NCDC	National Climatic Data Center
NOAA	National Oceanic and Atmospheric Administration
USGS	United States Geological Survey

4. Introduction

In this project, will be analyzed a dataset collected from the National Oceanic and Atmospheric Administration's (NOAA) *surface reference network* (USCRN). The network is composed of around 150 weather stations based in the USA and is tasked with determining how the US climate has changed (and is changing) over time. For more information, visit the project homepage : https://www.ncdc.noaa.gov/data-access/land-based-station-data/land-based-datasets/us-climate-reference-network-uscrn

5. Analysis Questions

5.1. Extremes

5.1.1. Question

When and where was the hottest and coldest surface and air temperatures observed in the dataset? Are they anomalies? If so, what were the hottest and coldest non-anomalous temperatures?

5.1.2. Analysis

For this problem, we run MapReduce for all datasets in all location from 2006 to 2019 and analyze all the air and surface temperature of all locations. The Map will clean data and only send the one which is valid. During that, Mapper also keep track the current max and min temperature and only send the temperature that large than max or less than min to optimize it. After that, reducer will receive temperature as key and create in-memory object to keep track the max and min temperature with location.

5.1.3. Map Reduce Results

After all data is processed, the reducer will write out the max and min data.

Air Temp	Surface Temp	Location	City	Date	Time
-59	9999	9w9g	Mexican Hat, Utah	20110328	1320
-59	9999	9w9g	Mexican Hat, Utah	20110328	1450
59.1	9999	9w9g	Mexican Hat, Utah	20110321	715
9999	-60	9wwp	Timpas, Colorado	20170507	1920
9999	89.9	dhy4	Placid Lakes, Florida	20130717	1530

Figure 1: Max and Min Air/Surface Temperature by Date and Location

As you can see from this table, the min air temperature is in Mexican Hat, Utah which seem weird for us because we think it should be Alaska. And then also for min surface temperature also in Colorado which also weird too. Last the max surface temp is 89.9oC which is really high. So maybe some data is not correct. Then we come up with idea to use threshold, because data is recorded every 5

minutes so it may not go beyond the threshold (2oC). And we come up with this table with seem more valid.

Air Temp	Surface Temp	Location	City	Date	Time
-47.5	9999	be6q	Ruby, Alaska	20180125	1250
-47.5	9999	be6q	Ruby, Alaska	20180125	1305
-47.5	9999	be6q	Ruby, Alaska	20180125	1635
-47.5	9999	be6q	Ruby, Alaska	20180125	1245
52.2	9999	9qs8	Stovepipe Wells, California	20070705	2350
9999	-50.4	bg0w	Paxson, Alaska	20131120	1620
9999	-50.4	bg0w	Paxson, Alaska	20131120	1830
9999	-50.4	bg0w	Paxson, Alaska	20131120	1625
9999	71.3	9qs8	Stovepipe Wells, California	20130704	2040

Figure 2: Max and Min Air/Surface Temperature by Date and Location

5.2. Drying Out

5.2.1.Question

*Choose a region in North America (defined by Geohash, which may include several weather stations) and determine when its **driest** month is. This should include a histogram with data from each **month**.*

5.2.2.Analysis

Figure 3: Santa Barbara

We choose "**Santa Barbara**" as a region(Figure-3) to calculate its **driest** month. Santa Barbara is a city on the central California coast, with the Santa Ynez Mountains as dramatic backdrop. Downtown, Mediterranean-style white stucco buildings with red-tile roofs reflect the city's Spanish colonial heritage. Upscale boutiques and restaurants offering local wines and seasonal fare line State Street. On a nearby hill, Mission Santa Barbara, founded in 1786, houses Franciscan friars and a museum.

We decided to use **WETNESS** variable to calculate average **driest** month. We also used **WETNESS_FLAG** to check that data is good data or erroneous data.

WETNESS : The presence or absence of moisture due to precipitation, in Ohms. High values (>= 1000) indicate an absence of moisture. Low values (< 1000) indicate the presence of moisture. So higher the value of wetness means drier the weather is. You can see our results in next section.

Figure 4: Santa Barbara Geo Hash (with 4 precision digit)

5.2.3.Map Reduce Results

As seen in the graph, For the last 13 years Santa Barbara's driest month,with wetness value 982, is August. Second driest month for Santa Barbara is July.(wetness=962)

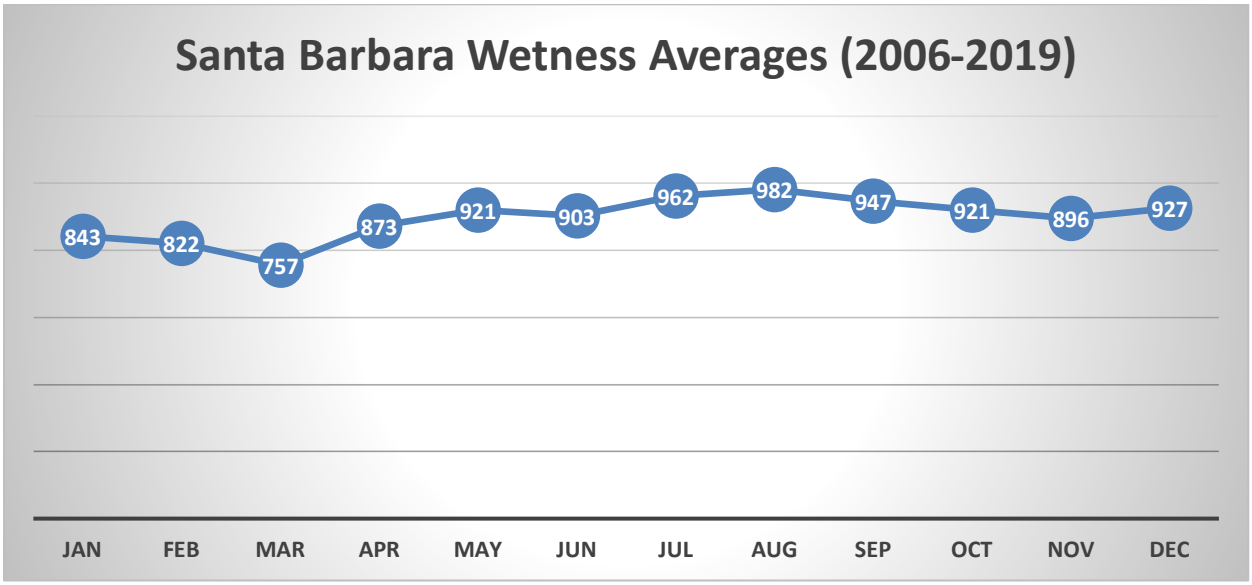


Figure 5: Wetness Averages by Month (2006-2019)

5.3. Moving Out

5.3.1.Question

Matthew, a student in your Big Data class, really likes the Bay Area weather but due to financial limitations will never be able to own a house there. Find similarly-sized regions with similar weather patterns so Matthew can move away for good. You should consider more than just one or two features from the dataset here, and think carefully about your methodology.

5.3.2.Analysis

For this problem, I design a MapReduce to run thorugh the whole dataset. The purpose of Map Reduce is to analyze temperature and humidity of all location around US and compare it with Bay Area weather (San Francisco). The Mapper will run through data set, validate data and write key (location, month) and value(air temperature, humidity) if both are valid.

The responsibility of Reducer is for each <location, month>, it will write the min, max, average temperature with avg humidity; also writing the difference between current location and San Francisco. Data of San Francisco is got from U.S Climate Data and CustomWeather

5.3.3.Map Reduce Results

This is how the output look like (First 16 lines of output):

Location	Month	Min Temp	Diff Min	Max Temp	Diff Max	Avg Temp	Diff Avg	Avg Humidity	Diff Humidity	Avg Precipitation	Diff Precipitation
8e3r	0	-5.3	12.9	18.7	4.9	5.91	4.79	27.71	54.29	0.35	3.71
8e3r	1	-5.9	14.5	17.3	1.6	5.04	7.06	36.12	23.88	0.32	3.77
8e3r	2	-2.4	11.6	18.9	2.3	5.5	7.4	39.47	31.53	0.32	3.24
8e3r	3	-1.4	11	17.9	0.6	6.7	6.7	37.05	30.95	0.35	1.36
8e3r	4	-0.5	11.1	19.5	1.6	7.95	6.35	38.08	28.92	0.37	0.1
8e3r	5	-0.5	12.1	19.6	0.5	9.29	6.01	30.15	32.85	0.31	0.09
8e3r	6	-0.9	13.1	20.4	1.2	9.13	6.57	36.46	34.54	0.44	0.41
8e3r	7	-0.5	13.3	19.2	0.9	9.01	7.39	43.1	31.9	0.53	0.52
8e3r	8	0.9	11.9	19.3	1.9	8.77	8.23	42.58	33.42	0.41	0.28
8e3r	9	0.2	11.9	17.8	2.9	8.17	8.23	41.91	30.09	0.49	0.69
8e3r	10	-1.9	12	17.2	0.1	6.65	7.05	37.28	31.72	0.31	2.16
8e3r	11	-2.5	10.3	17.1	3.2	5.77	5.13	34.2	41.8	0.37	3.95
8e3x	0	11.7	4.1	29.1	15.3	19.6	8.9	82.48	0.48	0.52	3.54
8e3x	1	0.1	8.5	30.6	14.9	19.59	7.49	83.3	23.3	0.62	3.47
8e3x	2	12.5	3.3	29.8	13.2	19.75	6.85	84.94	13.94	0.52	3.04
8e3x	3	12.8	3.2	29.4	12.1	20.32	6.92	85.05	17.05	0.49	1.22

Figure 6: MovingOut MapReduce result

Base on the result, for each location, we will calculate the difference of min, max, avg temperature and average humidity. So we will got average data of each location. This is the result table:

Location	Average of Diff Min	Average of Diff Max	Average of Diff Avg	Average of Diff Humid	Average of Diff Precipitation
8e3r	12.14	1.81	6.74	33.82	1.69
8e3x	3.68	12.08	6.88	14.52	1.59
9muw	5.36	21.02	3.33	13.07	1.67
9myf	7.10	21.53	9.64	42.65	1.77
9pxw	12.51	8.16	4.02	16.00	1.68
9q4g	7.88	13.12	0.82	8.31	1.64
9qb3	7.11	7.48	2.43	14.46	1.65
9qd5	12.43	16.30	3.70	11.83	1.75
9qdy	15.77	6.41	5.41	22.92	1.70
9qs8	7.57	23.46	11.86	52.25	1.79

Figure 7: Average difference with San Francisco for each location

Then because the data between temperature, humidity and precipitation is different, we will normalize data to convert it to value (0-1) so we can compare between each location using both temperature and humidity.

$$Dataset: S = \{x_1\ x_2, x_3, \dots, x_n$$

$$Normalize\ x_i = \frac{x_i - \min(S)}{\max(S) - \min(S)}$$

This is what we got from previous data set (Top 10 cities by Total normalize):

Location	Normalize Avg of Diff Min Temp	Normalize Avg of Diff Max Temp	Normalize Avg of Diff Temp	Normalize Avg of Diff Humidity	Normalize Avg of Diff Precipitation	Total Normalize
9qb3	0.10	0.29	0.07	0.27	0.22	0.95
9q4g	0.13	0.57	0.00	0.11	0.19	1.00
8e3x	0.00	0.52	0.26	0.27	0.00	1.05
c1f3	0.30	0.13	0.21	0.13	0.42	1.21
9pxw	0.26	0.32	0.14	0.31	0.35	1.38
c0ry	0.25	0.41	0.17	0.35	0.31	1.50
9rbk	0.31	0.45	0.15	0.17	0.44	1.51
c29s	0.32	0.40	0.20	0.20	0.46	1.58
8e3r	0.25	0.00	0.26	0.77	0.40	1.68
bfpu	0.41	0.12	0.31	0.34	0.43	1.61

Figure 8: Normalize average difference with San Francisco for each location

Base on the total normalization, we can get the city that nearest the same with San Francisco base on 3 aspect: temperature, humidity and precipitation. And the most fit city in this list is 9qb3 (Bodega Bay, CA), 9q4g (Santa Babara, CA), 8e3x (Mountain View Hawaii).

But what if you only care about temperature:

Location	Normalize Avg of Diff Min Temp	Normalize Avg of Diff Max Temp	Normalize Avg of Diff Temp	Temperature Normalization
9qb3	0.10	0.26	0.07	0.44
8e3r	0.26	0.00	0.28	0.53
9q4g	0.13	0.52	0.00	0.65
c1f3	0.30	0.12	0.23	0.66
9pxw	0.27	0.29	0.15	0.71
8e3x	0.00	0.47	0.28	0.76
9qdy	0.37	0.21	0.21	0.79
c0ry	0.26	0.37	0.18	0.81
bfpu	0.42	0.11	0.33	0.85
9rbk	0.31	0.41	0.16	0.87

Figure 9: Normalize min, max, average temperature compare to San Francisco

Base on the temperature normalization, we can get the city that nearest the same with San Francisco base on temperature. And the most fit city in this list is 9qb3 (Bodega Bay, CA), 8e3r (North Kona, Hawaii) and 9q4g (Santa Babara, CA).

Base on **Figure 8** and **Firgure 9**, we found out that Bodega Bay, CA maintains top 1 in both total normalization and temperature normalization. So i can be the best fit if you want to move to another city that has the same weather with San Francisco.

5.4. Travel Startup

5.4.1. Question

After graduating from USF, you found a startup that aims to provide personalized travel itineraries using big data analysis. Given your own personal preferences, build a plan for a year of travel across **5 locations**. Or, in other words: pick 5 regions. What is the best time of year to visit them based on the dataset?

- Part of your answer should include the **comfort index** for a region. There are several different ways of calculating this available online. Note: you don't need to use this for choosing the regions, though.

5.4.2. Analysis

Customer came to our travel startup and choosed these 5 locations to visit: **Hawaii**, **Miami**, **Austin**, **Yosemite** and **Santa Barbara**. Our travel startup calculates the comfort index to tell the best time of year to visit each location.

Figure 10: Hawaii

Hawaii, a U.S. state, is an isolated volcanic archipelago in the Central Pacific. Its islands are renowned for their rugged landscapes of cliffs, waterfalls, tropical foliage and beaches with gold, red, black and even green sands. Of the 6 main islands, Oahu has Hawaii's biggest city and capital, Honolulu, home to crescent Waikiki Beach and Pearl Harbor's WWII memorials.

Here is the Geo Hashes of our 5 locations:

Region Name	Geo Hash
Santa Barbara, California	9q4g
Yosemite, California	9qdy
Austin, Texas	9v6
Hawaii	8e3
Miami, Florida	dhw

In this question our Mapper key values is TravelWritable which consists of regionName and month. (We hard-coded region names based on Geo Hash values. For Example 8e3 hashcode gives the regionName: HI_HAWAII) And our mapper's output value is comfortIndex.

```
TravelWritable TravelWritable = new TravelWritable(new Text(regionName),
                                                    new IntWritable(month));
context.write(TravelWritable, new DoubleWritable(comfortIndex));
```

In the reducer, we are calculated average comfort index for each location for each month.

5.4.3.How we calculate Comfort Index

We calculated comfort index formula is like below:

ComfortIndex = (airTemperature + relativeHumidity) / 40;

We calculated average comfort index for each month and for each location. If comfort index value is around 2.8 to 3.2 is good time to visit that location.

Note: An arbitrary index of the suitability of environmental conditions to physical activity. Comfort Index = (temperature + relative humidity)/4. Since the index was devised in the USA, temperature is measured in degrees Fahrenheit. A comfort index above 95 during low wind-speeds may require acclimatization; the presence of wind allows higher values to be tolerated.

Reference : https://www.oxfordreference.com/view/10.1093/oi/authority.20110803095626624

5.4.4.Map Reduce Results

The place best to visit in certain month is highlighted with yellow below. Interestingly, for every month Miami is the best place to visit and Santa Clara is the second best place to visit whole year.

	LOCATION	COMFORT INDEX
Jan	CA_SANTA_BARBARA	2.13
	CA_YOSEMITE	1.39
	FL_MIAMI	2.38
	HI_HAWAII	1.71
	TX_Austin	1.72
Feb	CA_SANTA_BARBARA	2.11
	CA_YOSEMITE	1.54
	FL_MIAMI	2.44
	HI_HAWAII	1.83
	TX_Austin	1.95
Mar	CA_SANTA_BARBARA	2.26
	CA_YOSEMITE	1.67
	FL_MIAMI	2.34
	HI_HAWAII	1.88
	TX_Austin	1.93
Apr	CA_SANTA_BARBARA	2.24
	CA_YOSEMITE	1.57
	FL_MIAMI	2.44
	HI_HAWAII	1.87
	TX_Austin	2.05
May	CA_SANTA_BARBARA	2.29
	CA_YOSEMITE	1.48
	FL_MIAMI	2.53
	HI_HAWAII	1.9
	TX_Austin	2.23
Jun	CA_SANTA_BARBARA	2.45
	CA_YOSEMITE	1.41
	FL_MIAMI	2.68
	HI_HAWAII	1.81
	TX_Austin	2.27
Jul	CA_SANTA_BARBARA	2.52
	CA_YOSEMITE	1.43

	FL_MIAMI	2.71
	HI_HAWAII	1.88
	TX_Austin	2.24
Aug	CA_SANTA_BARBARA	2.53
	CA_YOSEMITE	1.32
	FL_MIAMI	2.74
	HI_HAWAII	1.97
	TX_Austin	2.14
Sep	CA_SANTA_BARBARA	2.47
	CA_YOSEMITE	1.41
	FL_MIAMI	2.74
	HI_HAWAII	1.94
	TX_Austin	2.23
Oct	CA_SANTA_BARBARA	2.33
	CA_YOSEMITE	1.42
	FL_MIAMI	2.57
	HI_HAWAII	1.98
	TX_Austin	2.05
Nov	CA_SANTA_BARBARA	2.2
	CA_YOSEMITE	1.44
	FL_MIAMI	2.46
	HI_HAWAII	1.89
	TX_Austin	1.87
Dec	CA_SANTA_BARBARA	2.06
	CA_YOSEMITE	1.26
	FL_MIAMI	2.45
	HI_HAWAII	1.85
	TX_Austin	1.91

5.5. Solar Wind Inc

5.5.1. Question

SolarWind, Inc.: You get bored enjoying the amazing views from your mansion that you bought with the money made with your travel startup, so you start a new company; here, you want to help power companies plan out the locations of solar and wind farms across North America. Locate the top 3 places for solar and wind farms, as well as a combination of both (solar + wind farm). You will report a total of 9 Geohashes as well as their relevant attributes (for example, cloud cover and wind speeds). If you'd like to do some data fusion to answer this question, the maps here and here might be helpful.

5.5.2.Analysis

For this problem, we design a MapReduce to run through the whole dataset and calculate the data base on Solar Radiation and Wind Speed. The Mapper will run through data set, validate data and write key (location) and value (solar radiation + wind speed). Then the Reducer will calculate the average of solar radiation and wind speed for each location. After that, the result will be imported to excel and do normalization using this formula:

$$Dataset: S = \{x_1\ x_2, x_3, \dots, x_n$$

$$Normalize\ x_i = \frac{x_i - \min(S)}{\max(S) - \min(S)}$$

5.5.3.Map Reduce Results

Base on that, this is the result table:

Geo Hash	Solar Radiation	Normalize Solar	Wind Speed	Normalize Wind Speed
8e3r	257.84	1.00	3.63	0.73
8e3x	162.42	0.53	0.63	0.06
9muw	205.03	0.74	2.12	0.39
9myf	233.09	0.88	2.25	0.42
9pxw	149.32	0.46	1.31	0.21
9q4g	202.40	0.73	1.65	0.29
9qb3	174.31	0.59	2.16	0.40
9qd5	201.71	0.72	1.81	0.33
8e3r	257.84	1.00	3.63	0.73

Figure 11: Sample table for normalization of solar and wind speed

Base on the normalization of solar, wind and sum of both we will find out the top 3 location suitable for each attribute:

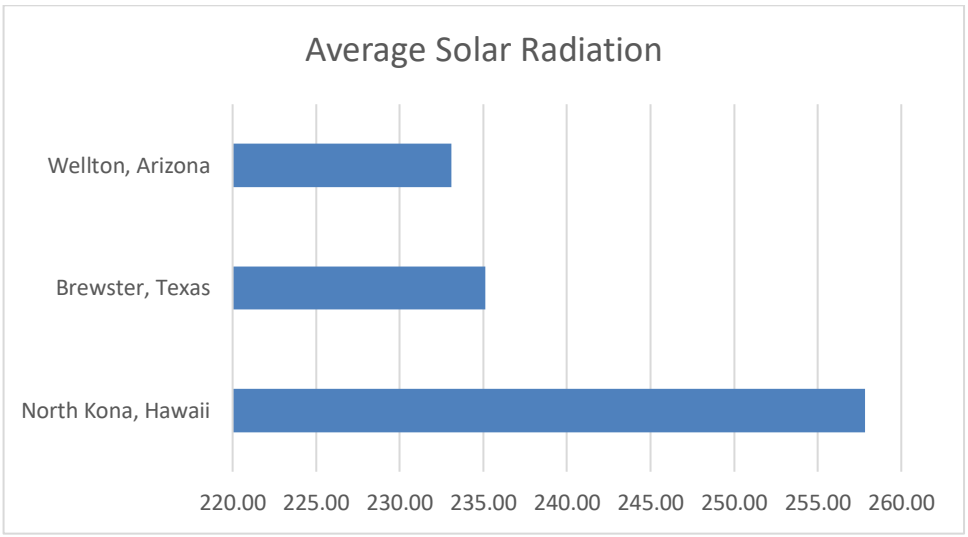


Figure 12: Top 3 City base on Solar Radiation

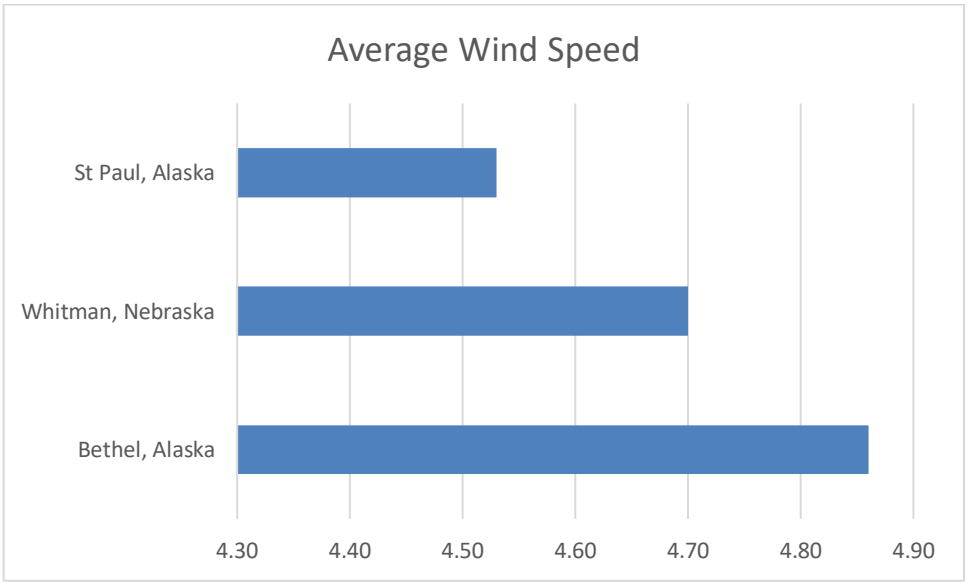


Figure 13: Top 3 City base on Wind Speed

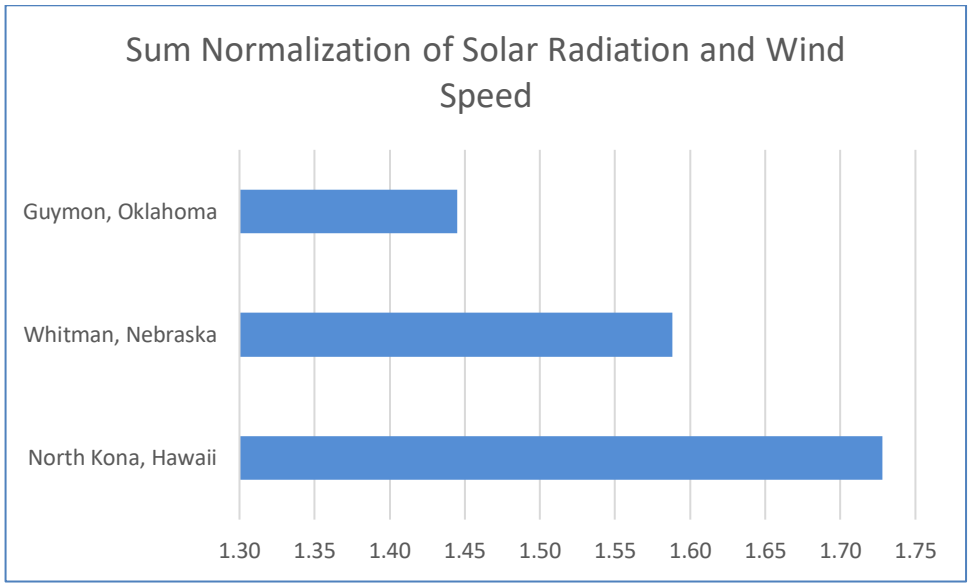


Figure 14: Top 3 City base on Normalization Solar Radiation and Wind Speed

5.6. Climate Chart

5.6.1. Question

Climate Chart: Given a Geohash prefix, create a climate chart for the region. This includes high, low, and average temperatures, as well as monthly average rainfall (precipitation). Earn up to 1 point of extra credit for enhancing/improving this chart (or porting it to a more feature-rich visualization library)

5.6.2. Analysis

We are reading the given GeoHash value from configuration file and calculating the high, low, and average temperatures, as well as monthly average rainfall (precipitation) for that region. In our example we wrote Yosemite Valley (9qdy) to configuration file and calculated the results.

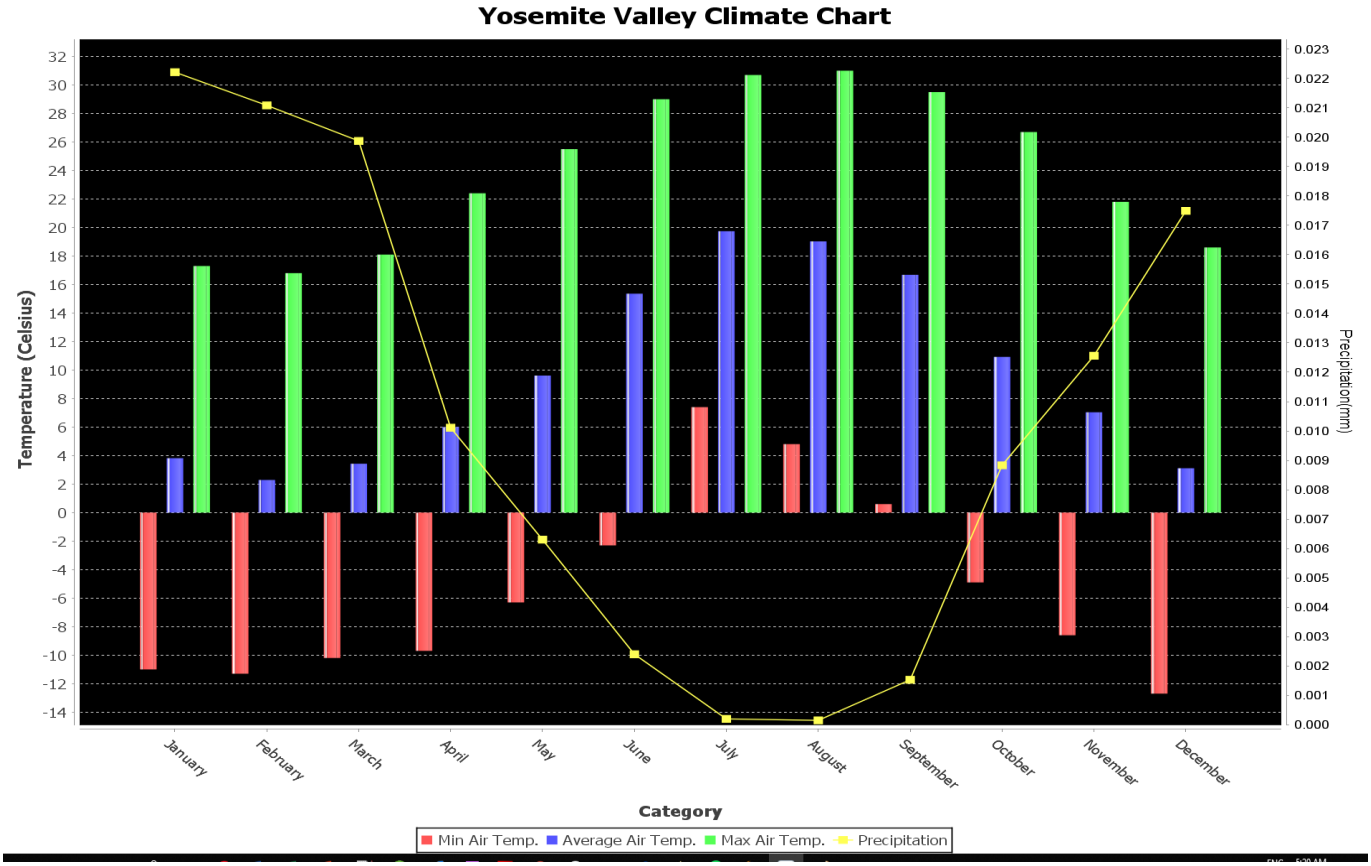
Figure 15: Yosemite Valley

Yosemite Valley is a glacial valley in Yosemite National Park in the western Sierra Nevada mountains of Central California. The valley is about 7.5 miles long and approximately 3000–3500 feet deep, surrounded by high granite summits such as Half Dome and El Capitan, and densely forested with pines.

5.6.3.Map Reduce Results

	Precipitation	Min Air Temp	Max Air Temp	Average Air Temp
Jan	0.02221	-11	17.3	3.82
Feb	0.02108	-11.3	16.8	2.29
Mar	0.01987	-10.2	18.1	3.43
Apr	0.01011	-9.7	22.4	6
May	0.0063	-6.3	25.5	9.61
Jun	0.0024	-2.3	29	15.36
Jul	0.0002	7.4	30.7	19.74
Aug	0.00015	4.8	31	19.03
Sep	0.00153	0.6	29.5	16.68
Oct	0.00883	-4.9	26.7	10.92
Nov	0.01256	-8.6	21.8	7.04
Dec	0.01749	-12.7	18.6	3.11

5.6.4.Extra Point (JfreeChart Library)



5.7. *Correlation is not Causation*

5.7.1.Question

Determine how features influence each other using Pearson's correlation coefficient (PCC). The output for this job should include (1) feature pairs sorted by absolute correlation coefficient, and (2) a correlation matrix visualization (heatmaps are a good option).

5.7.2.Analysis

In this question, we create a Mapper that run through the whole dataset. Each Mapper will maintain a object of RunningStatistics class (provided by Matthew), that will get input is a list of data for analyze features (air_temperature, precipitation, solar_radiation, etc.) in the dataset. For each input, Mapper will check if the whole list of data is valid and put into its RunningStatistics object. After Mapper finish with all data in the dataset, it will write out the RunningStatistics object, so for any size of data, Mapper only produce one RunningStatistics object which maintains the number of samples; mean, min, max for each attribute in sample.

The Reducer will receive the RunningStatistics object from each Mapper and merge it together to produce one RunningStatistics object for the whole dataset. Base on the RunningStatistics object of the whole dataset, we will calculate the correlation of each attributes and write it out.

5.7.3.Map Reduce Results

With that idea, we implemented it and produce the result like this:

Type	Correlation
air_temp,precipitation	0.969226121
air_temp,solar	0.579364267
air_temp,surface_temp	0.431872237
air_temp,humidity	0.427300463
air_temp,wetness	0.372129951
air_temp,wind	0.360569804
precipitation,solar	0.354089709
precipitation,surface_temp	0.348185998
precipitation,humidity	0.293108409
precipitation,wetness	0.282446206
precipitation,wind	0.192977702
solar,surface_temp	0.111822914
solar,humidity	0.110007264
solar,wetness	0.073962554
solar,wind	0.065867831
surface_temp,humidity	0.053260044
surface_temp,wetness	0.03859606

surface_temp,wind	0.03267389
humidity,wetness	0.031841397
humidity,wind	0.02512699
wetness,wind	0.002167428

Figure 16: MapReduce output for correlation sort by correlation

Then we will generate heatmap using the output of MapReduce

Figure 17: Correlation Matrix using Heatmap

5.8. EarthQuake and Climate Relation (Advanced Analysis)

5.8.1. Motivation

We decided to search for a relation betweeen climate and earthquakes in a certain region in North America. We decided to measure average soilMoisture, soilTemp, surfaceTemp. We thought that these attributes are related with earth and we were hoping to find some relation between earthquake numbers by year in certain region.

5.8.2. Choosen Area

We decided to choose our sample area from California. The reason that we've choosen California is earthquakes are frequently happening in this area. We've been living in SF only for 1 year and we've already experienced 2 earthquakes. ☺

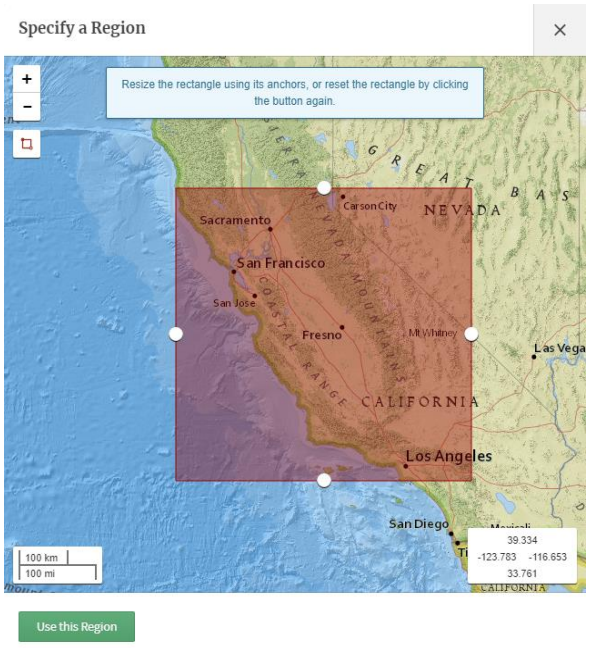


Figure 18: EarthQuake Analysis Area

As seen in this Figure-1(EarthQuake Analysis Area) we choosed our rectangular area with Latitude[33.708 , 39.368] and Longitude[-123.728,-116.653]. We tried to cover enough weather stations and cities when we were choosing our random rectangular analysis area.

In the south, biggest city is Los Angeles, in the north is Sacremento. In the west our border is in Pacific ocean which is away 100 miles away San Francisco. Finally in the east Death Valley National Park.

So with the help of the USGS's Search Earthquake Catalog we reached the earthquake numbers for each year for a certain area. (https://earthquake.usgs.gov/earthquakes/search/) We got the earthquakes whose magnitudes are greater than 2.5+. You can see the interface we used to get data as CSV file for a certain criteria in Figure-2.

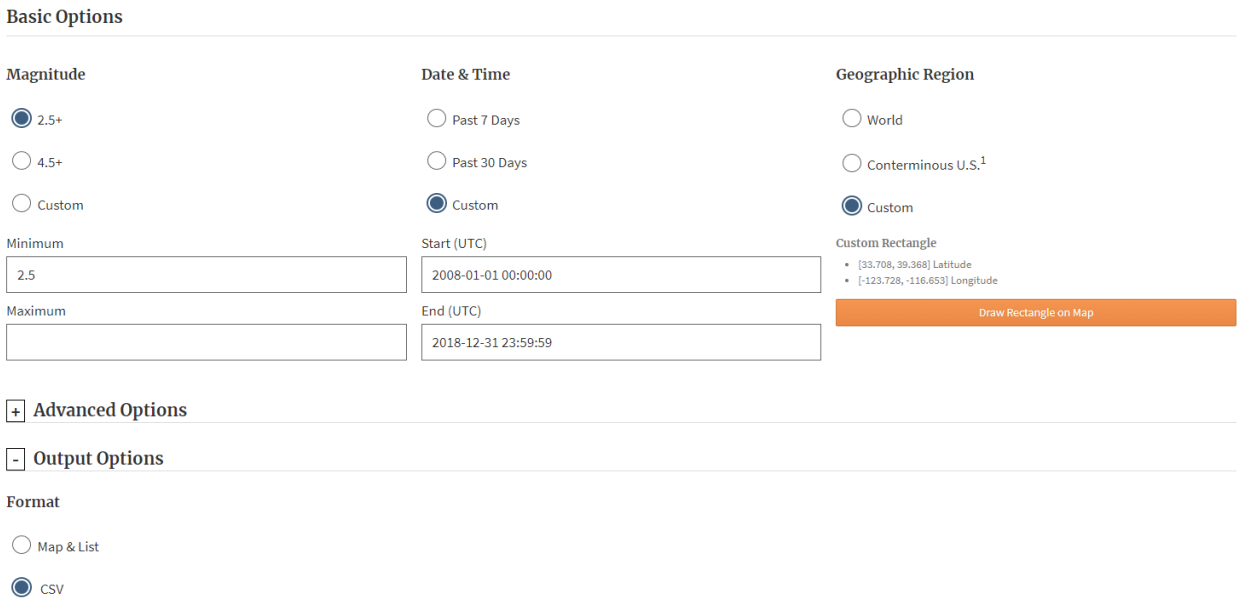
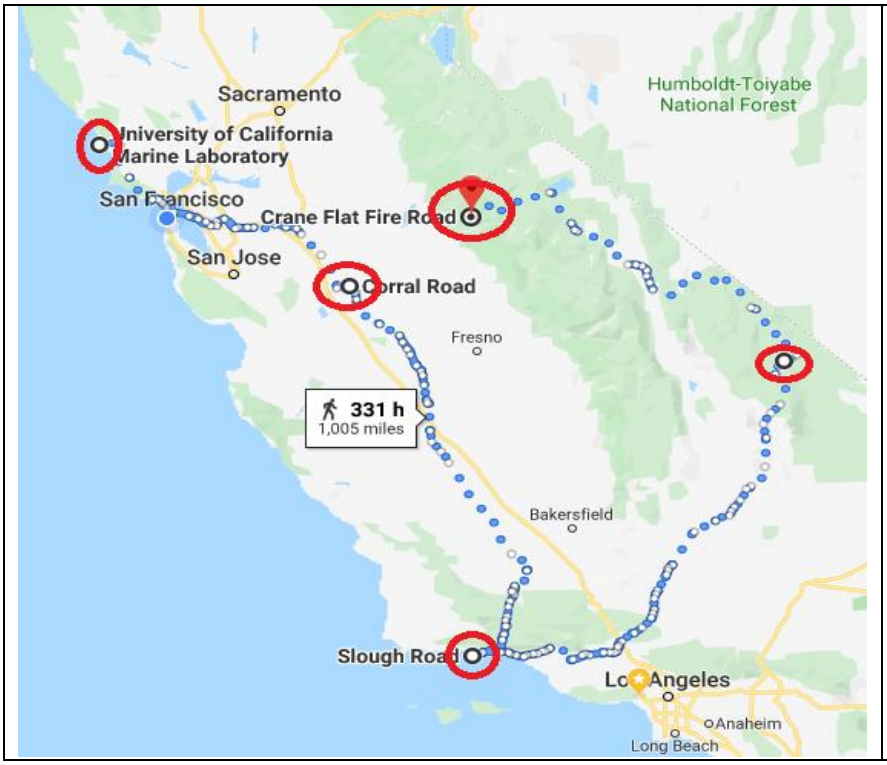


Figure 19: Search Earthquake Catalog

5.8.3.Weather Stations in Choosen Area



In Figure-3, we listed the weather stations(red circles) in our rectangular area. Here is the list of weather stations that we choosed in our analysis and their locations:

6. Bodega Bay CA -> 38.32 -123.07
7. Merced CA -> 37.24 -120.88
8. Santa Barbara CA -> 34.41 -119.88
9. Stovepipe Wells CA -> 36.60 -117.14
10. Yosemite Village CA -> 37.76 -119.82

For be able get only these weather stations we included **9q** geohash and excluded **9qy** and **9qt** geohashes**.** We used online geohash tool in this web site to choose the

Figure 20: Weather stations	correct region. http://geohash.gofreerange.com

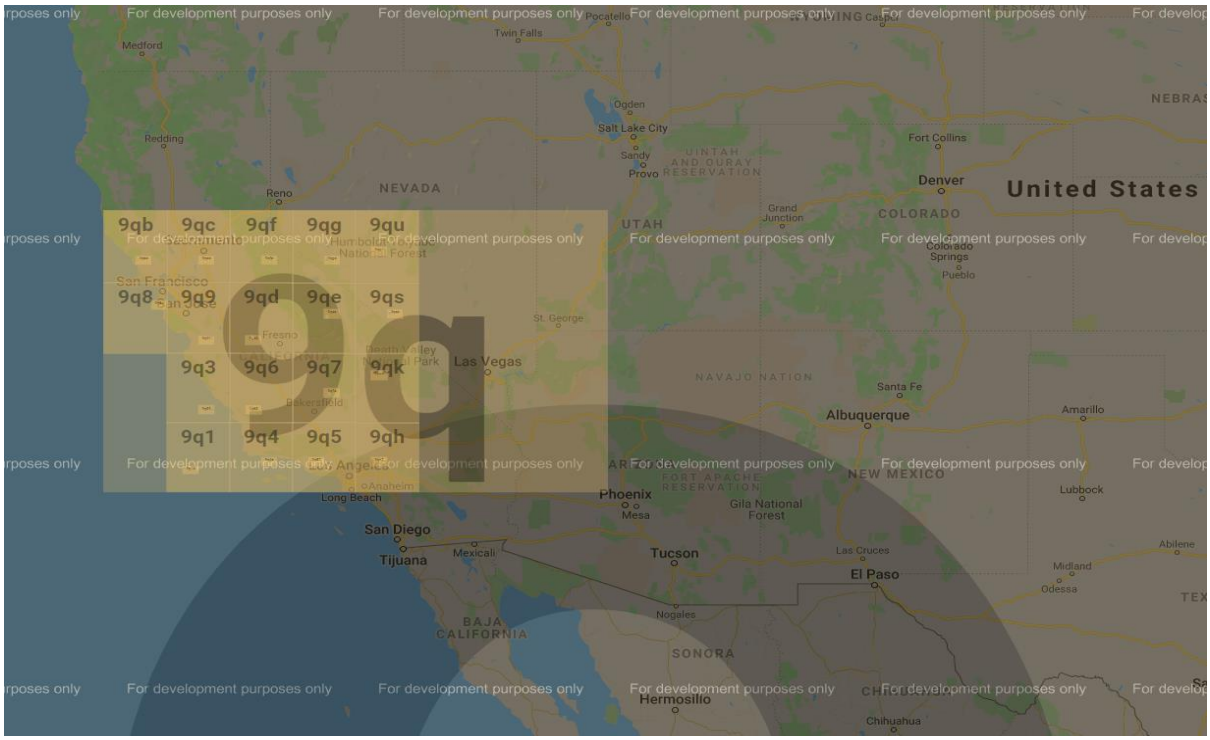


Figure 21: Geohash

Excluded regions(9qy and 9qt) was containing these weather stations : Baker NV(39.01 -114.21) and Mercury NV(36.62 -116.02) We excluded those weather stations because they were far away from our focus area.

10.3.4. Map Reduce Results

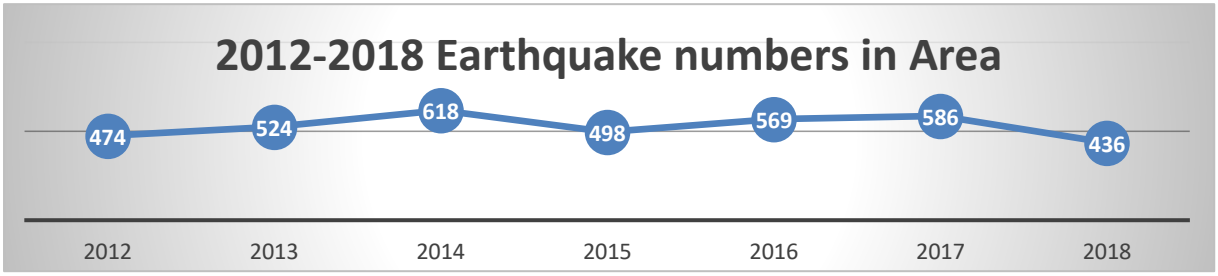


Figure 22: Earthquake Numbers(2012-2018)

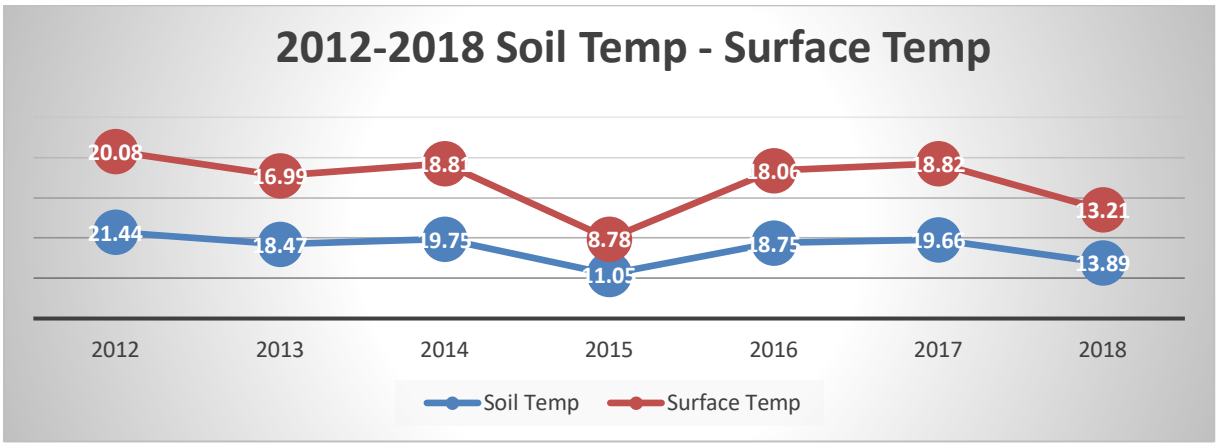


Figure 23: Soil/Surface Average Temperatures (2012-2018)

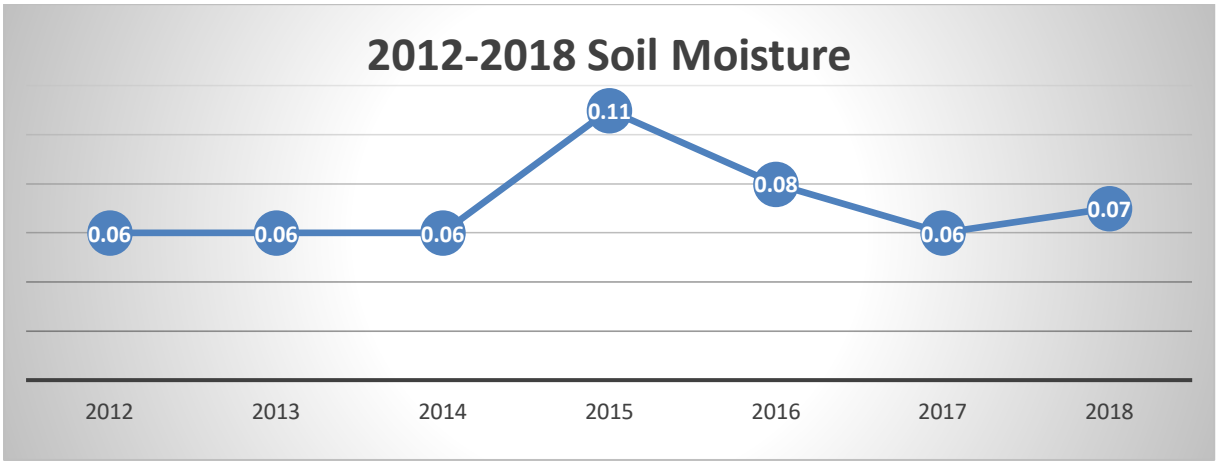


Figure 24: Soil Moisture Temperatures (2012-2018)

Important Note: As you seen, we did not include data before 2012 because we realized that soil moisture and soil temperature didn't calculated by NOAA before 2012. So we decided to analyzed with 7 years data. (2012-2018) We also didn't include 2019 because by that time we write this report month is November so we didn't have whole year data.

10.3.5. Correlation Analysis

The correlation coefficient that indicates the strength of the relationship between two variables can be found using the following formula:

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

10.3.5.1. Soil Moisture-Earthquake Numbers

It seems a negative correlation between **soil moisture** and **earthquake numbers**.

The value of R is **-0.2338.**

Although technically a negative correlation, the relationship between our variables is only **weak** (nb. the nearer the value is to zero, the weaker the relationship).

10.3.5.2. Soil Temprature -Earthquake Numbers

There is a correlation between **soil Temprature** and **earthquake numbers**.

The value of R is **0.475.**

Although technically a positive correlation, the relationship between our variables is **weak** (nb. the nearer the value is to zero, the weaker the relationship).

10.3.5.3. Surface Temprature -Earthquake Numbers

There is a correlation between **surface Temprature** and **earthquake numbers**.

The value of R is **0.4763**.

Although technically a positive correlation, the relationship between our variables is **weak** (nb. the nearer the value is to zero, the weaker the relationship).

10.3.6. Summary

Although we found weak correlation between earthquake numbers and our climate variables, we need more data to be able to conclude a decision. Because 6 years relatively is not enough when we consider earth movements.