

Assignment 4 Writeup

Harvard CCB Clinic Team
November 30, 2022

Step 1: Setting up the database

It took us a while to find a machine on our end that would work (issues with computer performance, computer architecture, server difficulties with getting container IPs), but our clinic computer was the ticket. The only complicating factor was that one of the Docker images wouldn't build, so we built in on another machine, saved the built Docker image as a .tar file, loaded it onto our clinic computer, and loaded the .tar file into Docker on there. Aside from that, the setup was pretty straightforward!

Step 2: Queries

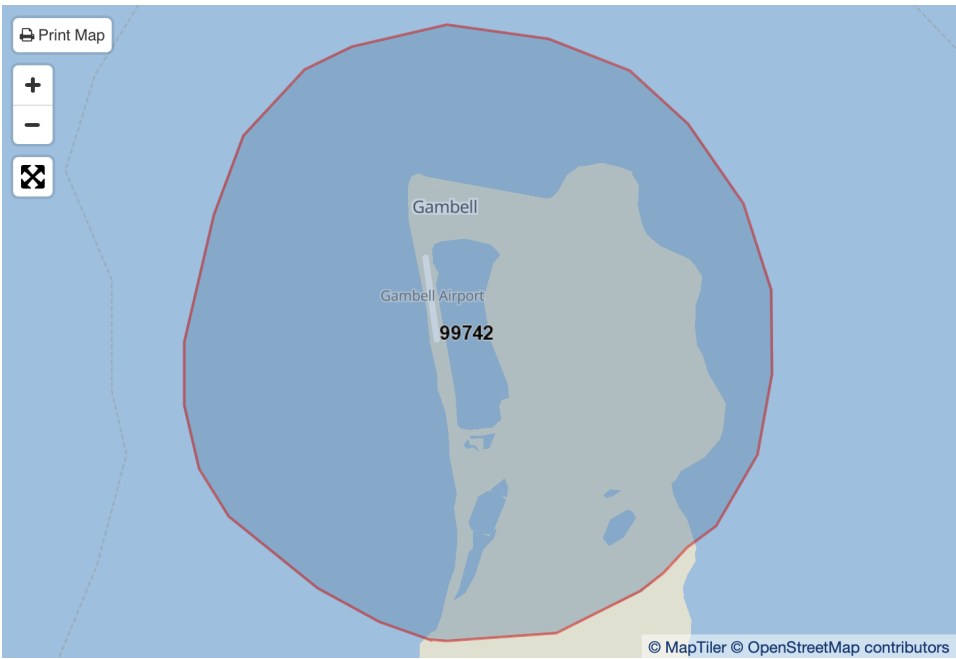
QUERY ONE: Join the ISD and TIGER tables together.

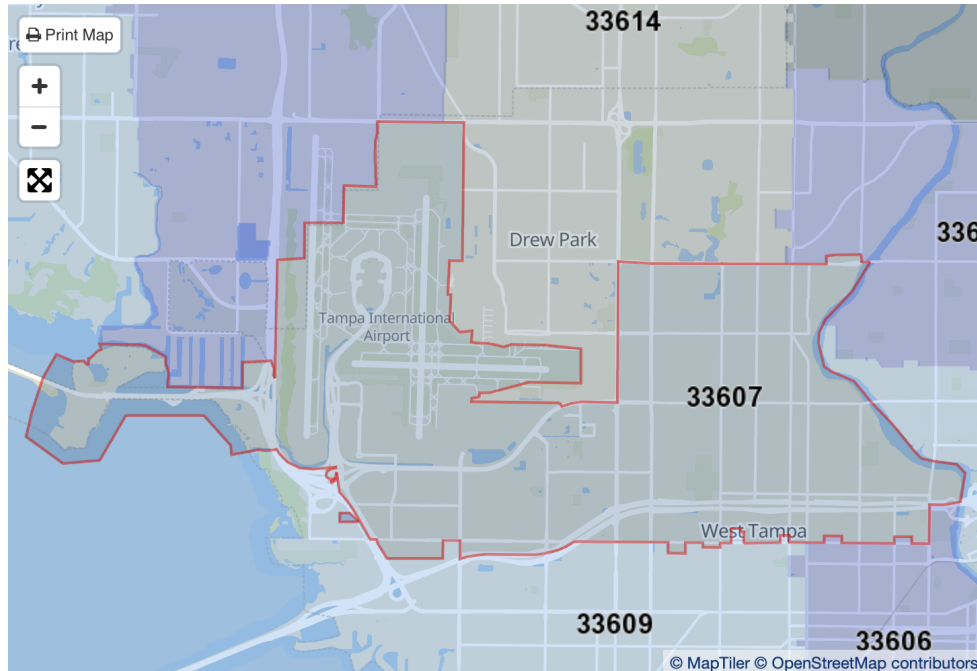
With this query, we need to process the LATITUDE and LONGITUDE columns in the ISD table as geography::Points, then check which Polygon in the TIGER GeographyLocation column that they lie within. We can do all this in a join to link the tables together:

```
SELECT * FROM [ISD_HMC].[2021_USA].[DAILY_SUMMARY] as i
INNER JOIN [TIGERFiles].[dbo].[t1_2020_us_zcta] as t
ON t.GeographyLocation.STIntersects(geography::Point([LATITUDE], [LONGITUDE],
4326)) = 1
ORDER BY STATE asc;
```

The main problem with this is that it doesn't complete in a reasonable time! This is because we aren't using spatial indexing to the fullest here. But only joining the top 1000 entries in ISD and pulling off the distinct names and zips shows that things are looking good:

	NAME	zcta5ce20
1	ALMA BACON CO AIRPORT GA	31510
2	TAMPA INTERNATIONAL AIRPORT FL	33607
3	GAMBELL AIRPORT AK	99742
4	SMITHFIELD JOHNSTON CO AIRPORT NC	27577





We were able to improve performance by using a spatial index with maximum grid resolution (high at all four levels):

```
CREATE SPATIAL INDEX WeatherLocationsInd ON
[ISD_HMC].[2021_USA].[ISD_Spatial](GeographyLocationIsd)
WITH (GRIDS = (HIGH, HIGH, HIGH, HIGH))

SELECT i.STATION_NAME, t.zcta5ce20
FROM [ISD_HMC].[2021_USA].[ISD_Spatial] as i
WITH (INDEX (WeatherLocationsInd))
INNER JOIN [TIGERFiles].[dbo].[tl_2020_us_zcta] as t
ON t.GeographyLocation.STIntersects(i.GeographyLocationIsd) = 1;
```

This query took 7 minutes and 32 seconds to complete on our clinic computer. It is likely that more could be done with spatial index options, but for now this is a promising result!

QUERY TWO: Find the average annual temperature in Boston.

Because the ISD table contains a group of weather stations named 'BOSTON MA', the most obvious approach is to find the average temperature using those rows:

```
SELECT NAME, AVG (TEMP) as "AVG TEMP" FROM
[ISD_HMC].[2021_USA].[DAILY_SUMMARY] WHERE NAME = 'BOSTON MA' AND TEMP !=
9999.9
GROUP BY NAME
```

	NAME	AVG TEMP
1	BOSTON MA	55.06511286557731

However, we weren't confident that all of the weather stations in Boston were named 'BOSTON MA'. We wrote some Python code that pulls the zip codes for Boston using a package, then prints out the SQL needed to make the temp table:

```
import zipcodes

tableName = "#BostonZipcodes"
desired_zips = zipcodes.filter_by(city="Boston")
zip_str = ""
for d in range(len(desired_zips)):
    zip_str = zip_str + " (" + desired_zips[d]['zip_code'] + "), "

print("CREATE TABLE " + tableName + " ( zip INT NOT NULL );")

print("INSERT INTO " + tableName + "(zip) VALUES " + zip_str[:-1] + ";")



-- Create temporary table of Boston zipcodes
CREATE TABLE #BostonZipcodes (
    zip INT NOT NULL
);

-- Insert relevant zip codes
INSERT INTO #BostonZipcodes(zip)
```

```
VALUES (02108), (02109), (02110), (02111), (02112), (02113), (02114),
(02115), (02116), (02117), (02118), (02119), (02120), (02121), (02122),
(02123), (02124), (02125), (02126), (02127), (02128), (02129), (02130),
(02131), (02132), (02133), (02134), (02135), (02136), (02137), (02141),
(02149), (02150), (02151), (02152), (02163), (02171), (02196), (02199),
(02201), (02203), (02204), (02205), (02206), (02210), (02211), (02212),
(02215), (02217), (02222), (02241), (02266), (02283), (02284), (02293),
(02297), (02298), (02445), (02467);
```

From there, we do a join similar to our work in query 1, but we do it on a subset of both tables (only the MA rows in ISD, only the rows from TIGER that have the zip codes from our temp table) so it actually runs!

```
-- Extract rows from TIGER that have Boston zipcodes,
SELECT i.NAME, AVG(i.TEMP) as "AVG_TEMP"
  FROM (
    SELECT *
      FROM [ISD_HMC].[2021_USA].[ISD_Spatial]
      WITH (INDEX (WeatherLocationsInd))
      WHERE STATE = 'MA'
    ) as i
INNER JOIN (
  SELECT *
    FROM [TIGERFiles].[dbo].[tl_2020_us_zcta], #BostonZipcodes as b
    WHERE zcta5ce20 = b.zip
  ) as t
ON t.GeographyLocation.STIntersects(i.GeographyLocationIsd) = 1
GROUP BY NAME
```

	NAME 	AVG_TEMP 
1	BOSTON MA	55.06511286557731

This is the same result as before, but doing it with the zip codes was a good check that there aren't data validity issues in this case!

There was one more data validity concern in that one of the Boston weather stations did not report a temperature for every day of the year, throwing off our averages slightly.

		NUMBER OF BOSTON ENTRIES					
1		708					
385	99497199999	2021-01-20	42.35	-71.05	0	BOSTON MA	35.5
386	99497199999	2021-01-22	42.35	-71.05	0	BOSTON MA	37.1

This result is pretty close to what a local news organization in Boston found for 2021. Of course this all depends on location within Boston and time of day, so it makes sense they should differ a little.

Google

boston average annual temperature 2021

×

All

News

Images

Maps

Books

More

Tools

About 27,700,000 results (0.52 seconds)

52.4 degrees

In 2021, Boston had a mean temperature of **52.4 degrees**, beating out 2012's record of 52.2 degrees. It was also the 16th wettest year in local recorded history, with 52.3 inches of rain and 21.7 inches of snow, according to meteorologists. Jan 1, 2022<https://www.boston.com/news/weather/2022/01/02>

Boston had the warmest year on record in 2021

QUERY THREE: What are the top five windiest stations in Massachusetts?

We decided that this could either mean highest average wind speed or highest maximum wind speed, so we did both:

```
SELECT Top(5) MAX (WDSP) AS "MAX WIND SPEED", NAME FROM
[ISD_HMC].[2021_USA].[DAILY_SUMMARY] WHERE [STATE] = 'MA' AND WDSP != 999.9
GROUP BY NAME
ORDER BY "MAX WIND SPEED" DESC;
```

```
SELECT Top(5) AVG (WDSP) AS "AVG WIND SPEED", NAME FROM
[ISD_HMC].[2021_USA].[DAILY_SUMMARY] WHERE [STATE] = 'MA' AND WDSP != 999.9
```

GROUP BY NAME

ORDER BY "AVG WIND SPEED" DESC;

	MAX WIND SPEED (knots) ▾	NAME ▾
1	34.5	PROVINCETOWN MUNICIPAL AI...
2	31.2	NANTUCKET MEMORIAL AIRPOR...
3	27.8	PLYMOUTH MUNICIPAL AIRPOR...
4	27.4	BORDEN FLATS LIGHT MA
5	26.5	BOSTON MA

	AVG WIND SPEED (knots) ▾	NAME ▾
1	10.429041110652767	NANTUCKET MEMORIAL AIRPOR...
2	9.975147932944212	BLUE HILL LCD MA
3	9.811538455905495	PROVINCETOWN MUNICIPAL AI...
4	9.199999992930612	BORDEN FLATS LIGHT MA
5	8.66164385651889	FALMOUTH OTIS AFB MA

This query also could have benefitted from double-checking with the zip codes, but we opted to hold off on that until we've improved performance on the joins with spatial indexes.

It's a little harder to find comparable data for this one, but our results here make sense, as all of these stations are either on Cape Cod, the islands, or relatively close to the seafront (in the case of Boston and Blue Hill).

QUERY FOUR: Where is the rainiest station in Washington State located?

We could have just queried the ISD table for this one, but that would only get us position in terms of latitude and longitude. To get the zip code as well, we made a temp table then joined that data with TIGER:

-- Make temp table that gets the total precipitation for each weather station
in Washington State


```
SELECT SUM (PRCP) AS "TOTAL PRECIPITATION", NAME, LATITUDE, LONGITUDE
INTO #RainiestInWA
FROM [ISD_HMC].[2021_USA].[DAILY_SUMMARY]
WHERE [STATE] = 'WA' AND PRCP != 99.99
GROUP BY NAME, LATITUDE, LONGITUDE
ORDER BY "TOTAL PRECIPITATION" DESC;
```

```
-- Join temp table with zcta on Points in Polygons to get zip code column
SELECT TOP (1) "TOTAL PRECIPITATION", NAME, LATITUDE, LONGITUDE, t.zcta5ce20
as "ZIP CODE"
FROM #RainiestInWA
INNER JOIN [TIGERFiles].[dbo].[t1_2020_us_zcta] as t
ON t.GeographyLocation.STIntersects(geography::Point([LATITUDE], [LONGITUDE],
4326)) = 1
ORDER BY "TOTAL PRECIPITATION" DESC;
```

	TOTAL PRECIPITATION ▾	NAME ▾	LATITUDE ▾	LONGITUDE ▾	ZIP CODE ▾
1	175.2200006004423	QUINAULT 4 NE WA	47.5139	-123.812	98526

Like query 3, our analysis also would have benefitted from double-checking with the zip codes.

Quinault ranks highly on a list of average precipitation levels by town in Washington. All of the other top performers are in western WA (where Quinault is as well).



Local Data Search

[USA.com](#) / [Ranks](#) / [Washington Average Precipitation City Rank](#)

Washington Average Precipitation City Rank

A total of 744 results found. [Show Results on Map.](#)

Rank	Average Precipitation ▾	City / Population
1.	110.31 inches	Taholah, WA / 829
2.	109.79 inches	Amanda Park, WA / 175
3.	109.77 inches	Quinault, WA
4.	109.75 inches	Neilton, WA / 400
5.	109.65 inches	Humptulips, WA / 249

QUERY FIVE: How many weather stations are there per state? Which state has the most weather stations? The least?

We tackled this question by counting all the distinct station names in each state in the ISD table:


```
SELECT STATE, COUNT (DISTINCT NAME) AS "NUMBER OF STATIONS" FROM
[ISD_HMC].[2021_USA].[DAILY_SUMMARY]
GROUP BY STATE
ORDER BY "NUMBER OF STATIONS" DESC;
```

The query gives us a top 20 that looks good:

	STATE	NUMBER OF STATIONS
1	TX	211
2	AK	209
3	CA	167
4	FL	115
5	MI	106
6	MN	101
7	NC	80
8	WI	71
9	CO	70
10	GA	68
11	AL	67
12	VA	67
13	IL	63
14	LA	63
15	IA	60
16	OK	58
17	SC	53
18	OH	53
19	WA	53
20	NE	52

...and a bottom 20 that looks a little strange:

	STATE ▾	NUMBER OF STATIONS ▾
47	CT	14
48	RI	13
49	VT	11
50	L,	10
51	I,	9
52	DE	7
53	M,	4
54	Y,	4
55	PR	4
56	A,	3
57	K,	3
58	D,	2
59	X,	2
60	VI	2
61	R,	1
62	RM	1
63	DC	1
64	C,	1
65	E,	1
66	FM	1
67	N,	1

We went through all of these unusual state codes using queries to see what happened with them. Some of the results were mislabeled, while others did not exist within the 50 states, or were weird all around:

	NAME ▾	STATE ▾
1	PORT ARANSAS TX	X,
2	SABINE TX	X,

	NAME ▾	STATE ▾
1	BOOMVANG SPAR OIL PLATFORM	M,
2	SHELL WEST DELTA 143 OIL PLATFORM	M,
3	STROM THURMOND DAM	M,
4	VERMILION 331 OIL PLATFORM	M,

Unusual State Code	Actual State/Country
L,	Illinois (IL)

I,	Michigan (MI)
M,	Mostly oil platforms in the Gulf of Mexico
Y,	New York (NY)
PR	Puerto Rico
A,	2 for Washington (WA), 1 for California (CA)
K,	Arkansas (AK)
D,	1 for Maryland (MD), 1 for Mississippi (MS)
X,	Texas (TX)
VI	U.S. Virgin Islands
R,	Oregon (OR)
RM	Marshall Islands
DC	Washington D.C.
C,	South Carolina (SC)
E,	Maine (ME)
FM	Micronesia
N,	Minnesota (MN)

Repairing the faulty state codes gives Texas two more weather stations, so it has the most at 213. On the other hand, Delaware is the U.S. state with the fewest weather stations, having only 7 of them.

QUERY SIX: Get the latitude and longitude for a station of your choosing.

For this one, we wanted to find the closest weather station to Harvey Mudd College. So we started by retrieving all the stations in California:

```
SELECT DISTINCT NAME FROM [ISD_HMC].[2021_USA].[DAILY_SUMMARY] WHERE STATE = 'CA'
```

From there we found that the nearby Ontario International Airport was one of the included weather stations. So we opted for that one:

```
SELECT DISTINCT NAME, LATITUDE, LONGITUDE FROM
[ISD_HMC].[2021_USA].[DAILY_SUMMARY] WHERE NAME = 'ONTARIO INTERNATIONAL
AIRPORT CA';
```

	NAME	LATITUDE	LONGITUDE
1	ONTARIO INTERNATIONAL AIRPORT CA	34.05314	-117.57689

QUERY SEVEN: Calculate the distance between the latitude and longitude values we choose in (6) and every station in the surrounding state (in both miles and meters)

We took the latitude and longitude from query 6, made it a fixed geography::Point in our query, then built distance columns using STDistance, the fixed point of Ontario, and all the other points using the LATITUDE and LONGITUDE columns. We converted our meters to miles by multiplying the meters result by the appropriate unit conversion:

```
SELECT DISTINCT NAME, LATITUDE, LONGITUDE, geography::Point([LATITUDE],
[LONGITUDE], 4326).STDistance(geography::Point(34.05314, -117.57689,4326)) AS
'DISTANCE (m)',
geography::Point([LATITUDE], [LONGITUDE],
4326).STDistance(geography::Point(34.05314, -117.57689,4326)) * 0.00062137 AS
'DISTANCE (mi)'
FROM [ISD_HMC].[2021_USA].[DAILY_SUMMARY] WHERE STATE = 'CA'
```

	NAME	LATITUDE	LONGITUDE	DISTANCE (m)	DISTANCE (mi)
1	ALAMEDA CA	37.772	-122.298	593013.5608329265	368.48083629475553
2	ALTURAS MUNICIPAL AIRPORT...	41.48362	-120.5615	865476.6092183783	537.7812006700237
3	ARCATA EUREKA AIRPORT CA	40.97844	-124.10479	960510.1377075925	596.8321842673668
4	ARENA COVE CA	38.92	-123.7	769500.7254337437	478.1446657627653
5	AUBURN MUNICIPAL AIRPORT ...	38.95472	-121.08194	627916.9495769239	390.1687549586132
6	AVALON CATALINA AIRPORT CA	33.40419	-118.41456	105866.92651434588	65.7825321282191
7	BAKERSFIELD AIRPORT CA	35.43424	-119.05524	204446.20686858858	127.03673956193488
8	BARSTOW DAGGETT AIRPORT CA	34.85371	-116.78702	114694.16385018027	71.26751259158651
9	BEALE AFB CA	39.13333	-121.43333	660814.9394718683	410.6105789396348
10	BICYCLE LAKE FORT IRWIN A...	35.28333	-116.63333	161562.85138670175	100.39030896615486
11	BIG BEAR CITY AIRPORT CA	34.264	-116.854	70644.22779588272	43.89620382552764
12	BISHOP AIRPORT CA	37.37114	-118.35886	374878.41761088796	232.93820235087745
13	BLUE CANYON NYACK AIRPORT...	39.27617	-120.70927	643595.0667504265	399.91066662671255
14	BLYTHE ASOS CA	33.61876	-114.71451	269283.22672228026	167.32451858842327
15	BODEGA 6 WSW CA	38.3208	-123.0747	684434.0220705845	425.28676829399905
16	BRIDGEPORT SONORA JUNCTIO...	38.3557	-119.519	508330.92396147863	315.861586221944
17	BURBANK GLENDALE PASADENA...	34.19966	-118.36543	74534.64785156358	46.31359413552606
18	CABLE CA	34.11154	-117.68759	12097.883647088094	7.517261961791129
19	CALAVERAS CO MAURY RASMUS...	38.14612	-120.64817	531659.4966978206	330.3572614631248
20	CAMARILLO AIRPORT CA	34.21142	-119.08762	140452.3684043649	87.27288815542022
21	CAMP PENDLETON MCAS CA	33.30424	-117.35508	85573.97208557083	53.173099034811145
22	CARLSBAD MCCLELLAN PALOMA...	33.12993	-117.27651	106126.09770065156	65.94357332825386
23	CASTLE AFB CA	37.38333	-120.56667	457880.24317279714	284.51304670028094

For this to be right, we'd expect the distance between Ontario International Airport and itself to be zero (or very close if there was slight error in calculating the distances).

92	ONTARIO INTERNATIONAL AIRPORT CA	34.05314	-117.57689	0.166257710197364...	0.000103307553385...
----	----------------------------------	----------	------------	----------------------	----------------------