BigQuery Data Exploration

01/2025

Mary Oladapo

Overview

In this project, I performed some data exploration using SQL through Google BigQuery platform. I analyzed a public dataset on Homelessness population found using BigQuery and answered 2-part in depth questions on my analysis. This project allowed me to utilize my new-found SQL knowledge to answer questions in the Exploration to get to my final analysis and conclusions. I was able to create SQL strings, test and run them using the query tab successfully to get answers to my BigQuery Exploration questions

Dataset

GOOGLE BIGQUERY PLATFORM

1. Homelessness public Dataset found in Google BigQuery

2

SCHEMA DETAILS PREVIEW TABLE EXPLORER PREVIEW INSIGHTS LINEAGE DATA PROFILE DATA QUALITY

Row	CoC_Number	State	CoC_Name	Overall_Homeles	Sheltered_ES_H	Sheltered_TH_H
1	AK-500	AK	Anchorage CoC	1023	665	305
2	AK-500	AK	Anchorage CoC	1094	751	249
3	AK-500	AK	Anchorage CoC	1208	694	335
4	AK-500	AK	Anchorage CoC	1147	704	393
5	AK-500	AK	Anchorage CoC	1128	654	319
6	AK-500	AK	Anchorage CoC	1105	573	292
7	AK-500	AK	Anchorage CoC	1122	676	394
8	AK-501	AK	Alaska Balance of State CoC	922	497	210
9	AK-501	AK	Alaska Balance of State CoC	761	474	195
10	AK-501	AK	Alaska Balance of State CoC	766	435	184
11	AK-501	AK	Alaska Balance of State CoC	824	461	210
12	AK-501	AK	Alaska Balance of State CoC	748	433	177
13	AK-501	AK	Alaska Balance of State CoC	717	363	215
14	AK-501	AK	Alaska Balance of State CoC	835	434	199
15	AL-500	AL	Birmingham/Jefferson, St. Clai...	1707	347	630
16	AL-500	AL	Birmingham/Jefferson, St. Clai...	1469	373	556
17	AL-500	AL	Birmingham/Jefferson, St. Clai...	901	431	219
18	AL-500	AL	Birmingham/Jefferson, St. Clai...	1092	540	304
19	AL-500	AL	Birmingham/Jefferson, St. Clai...	1153	448	448

SCHEMA DETAILS PREVIEW TABLE EXPLORER PREVIEW INSIGHTS LINEAGE DATA PROFILE DATA QUALITY

Field name	Type	Mode	Key	Collation	Default Value	Policy Tags	Des
CoC_Number	STRING	NULLABLE	-	-	-	-	-
State	STRING	NULLABLE	-	-	-	-	-
CoC_Name	STRING	NULLABLE	-	-	-	-	-
Overall_Homeless	INTEGER	NULLABLE	-	-	-	-	-
Sheltered_ES_Homeless	INTEGER	NULLABLE	-	-	-	-	-
Sheltered_TH_Homeless	INTEGER	NULLABLE	-	-	-	-	-
Sheltered_SH_Homeless	INTEGER	NULLABLE	-	-	-	-	-
Sheltered_Total_Homeless	INTEGER	NULLABLE	-	-	-	-	-
Unsheltered_Homeless	INTEGER	NULLABLE	-	-	-	-	-
Homeless_Individuals	INTEGER	NULLABLE	-	-	-	-	-
Homeless_People_in_Families	INTEGER	NULLABLE	-	-	-	-	-
Chronically_Homeless	INTEGER	NULLABLE	-	-	-	-	-
Homeless_Veterans	INTEGER	NULLABLE	-	-	-	-	-
Homeless_Unaccompanied_Youth_Under_18	INTEGER	NULLABLE	-	-	-	-	-
Count_Year	INTEGER	NULLABLE	-	-	-	-	-

Name of Dataset used (pictured above)

2. Dataset provided in BigQuery Exploration part 2 assignment.

Tools

1. **BigQuery**
2. **SQL FUNCTIONS**
3. **BigQuery Data Exploration Part 1 and 2 with questions.**
    - **BigQuery Data Exploration Part 1- Questions only (In Findings/Results tab)**
    - **BigQuery Data Exploration part 2- Questions only (In Findings/Results tab)**
4. **SQL Tutorial from web3 schools**
   Link- https://www.w3schools.com/sql/default.asp

Findings/Results

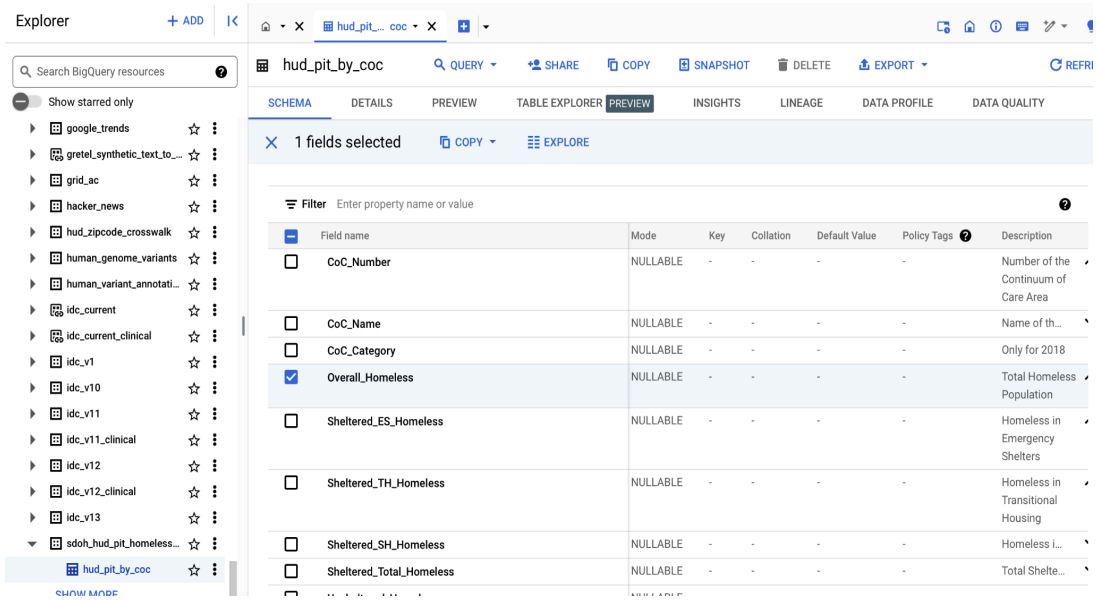
BigQuery Data Exploration Part 1- Questions and Answers

Step 1 - While exploring the dataset, look at the Schema tab and read through the descriptions provided for each column, to get a sense of what data is in this table. This will help you answer the questions below.

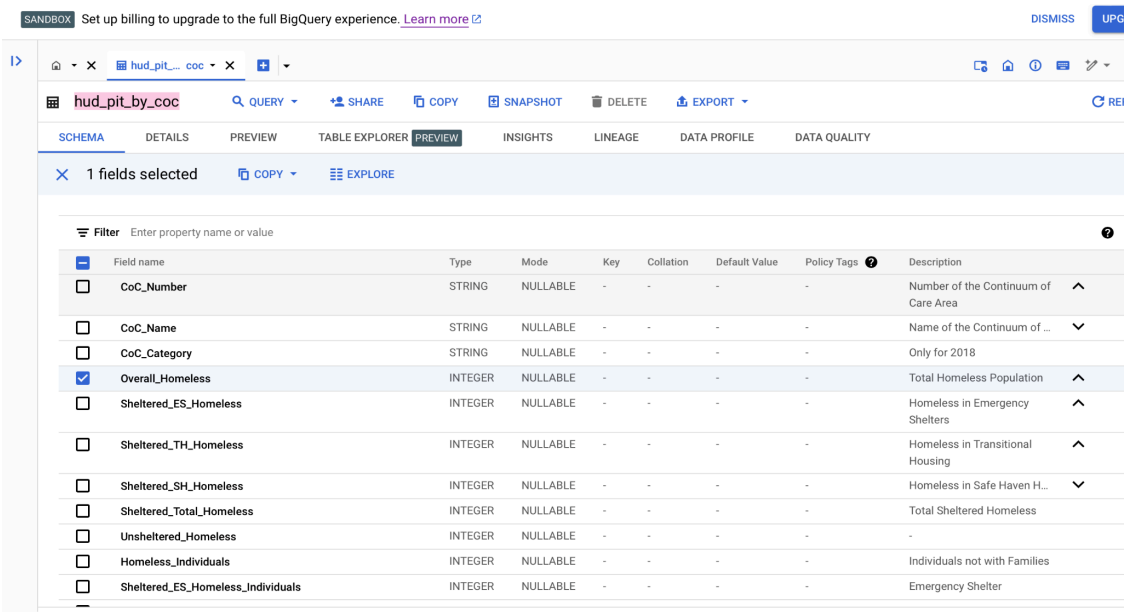
Step 2 - Answer the following questions

1. **There are several acronyms used in this dataset. For each acronym below, write out what it stands for:**
    a. CoC = Continuum of Care
    b. Sheltered_ES = Homeless in Emergency Shelters
    c. Sheltered_TH = Homeless in Transitional Housing
    d. Sheltered_SH = Homeless in Safe Haven Housing

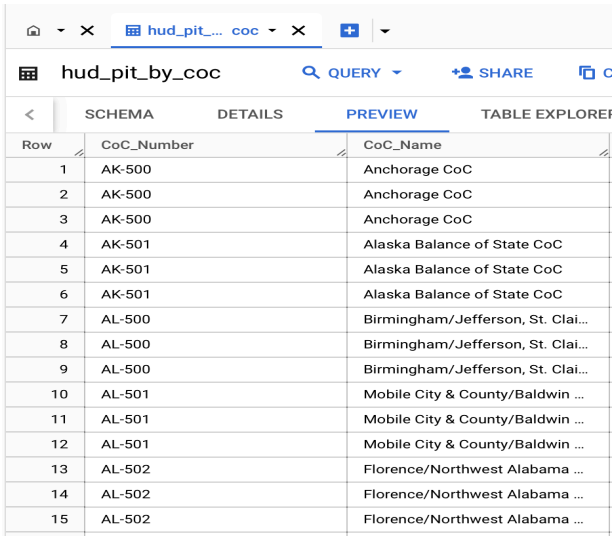
Found in homelessness dataset and selecting "hud_pit_by_coc"



- **2. What are the only 3 columns that are NOT an Integer type?**
- ANS: CoC_Number, CoC_Name, Coc_Category

- 

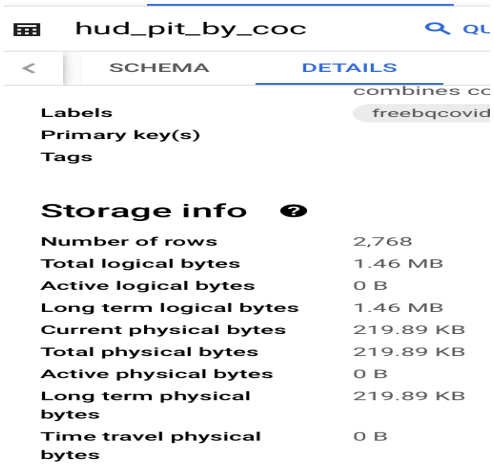
2. **Is there any way to determine which state each row of data is located in?**
    - ANS: By clicking preview, it provides more details on states they are located in



Row	CoC_Number	CoC_Name
1	AK-500	Anchorage CoC
2	AK-500	Anchorage CoC
3	AK-500	Anchorage CoC
4	AK-501	Alaska Balance of State CoC
5	AK-501	Alaska Balance of State CoC
6	AK-501	Alaska Balance of State CoC
7	AL-500	Birmingham/Jefferson, St. Clai...
8	AL-500	Birmingham/Jefferson, St. Clai...
9	AL-500	Birmingham/Jefferson, St. Clai...
10	AL-501	Mobile City & County/Baldwin ...
11	AL-501	Mobile City & County/Baldwin ...
12	AL-501	Mobile City & County/Baldwin ...
13	AL-502	Florence/Northwest Alabama ...
14	AL-502	Florence/Northwest Alabama ...
15	AL-502	Florence/Northwest Alabama ...

3. **How many total rows of data are there?**

- ANS: 2,768



4. **What might be some reasons that someone would use this dataset (no wrong answers here, just trying to think about how this data could be used)?**

- ANS: Someone might use this dataset to determine which locations most homeless people are and the total number of homeless people in that area. Where they are located at and which type of shelters are available or houses them in that area.

Step 3 - While you have the dataset selected, click on the Query button, and choose "In New Tab". In the new query window, you can delete all of the SQL code there, and copy and paste the following code, then run it to create a new table:

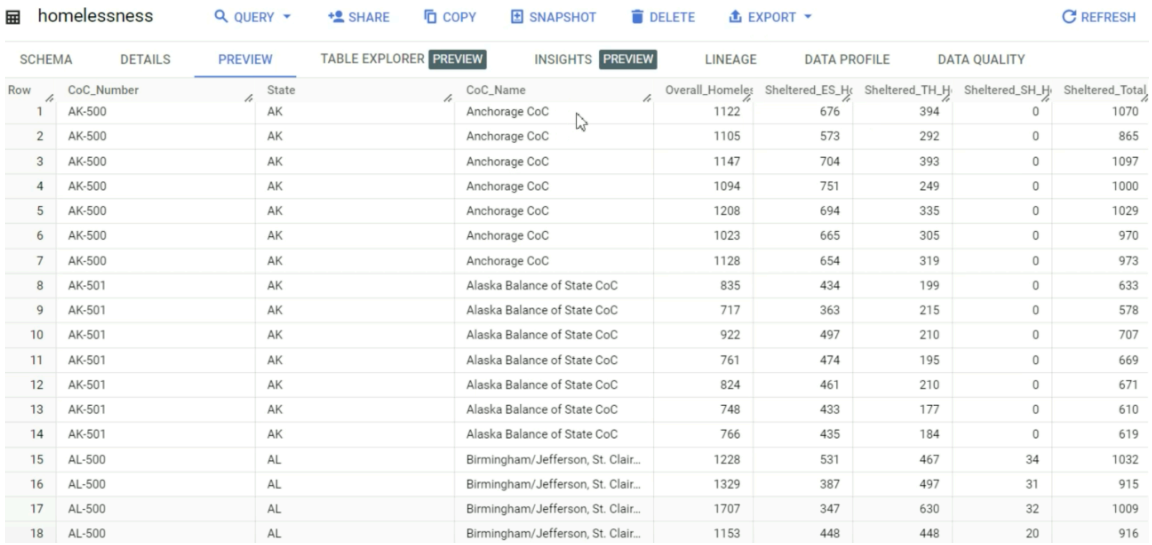
```
CREATE TABLE Exploration_Project.homelessness AS
SELECT CoC_Number, LEFT(CoC_Number, 2) AS State, CoC_Name, Overall_Homeless,
Sheltered_ES_Homeless, Sheltered_TH_Homeless, Sheltered_SH_Homeless,
Sheltered_Total_Homeless, Unsheltered_Homeless, Homeless_Individuals,
Homeless_People_in_Families, Chronically_Homeless, Homeless_Veterans,
Homeless_Unaccompanied_Youth_Under_18, Count_Year
FROM `bigquery-public-data.sdoh_hud_pit_homelessness.hud_pit_by_coc`
```

5. **Do a quick Google search to figure out what the SQL function LEFT( ) does. In the query we just ran, what does the line of code, LEFT(CoC_Number, 2) AS State, do?**

- ANS: SQL function LEFT( ) returns the left part of a character string with the specified number of characters.
- ANS: The line code LEFT(CoC_Number, 2) AS State, returns the left part of the character string of the state listed in the beginning of the CoC. ex: (AK, AL, etc)

ANSWER FOUND WITH A QUICK GOOGLE SEARCH

Step 4 - Open the new table you just created and use the Preview tab to look at the data and make sure all of the columns are appearing properly.

- **Gives us a more clean dataset when run by BigQuery and all column appeared properly**
- 
homelessness QUERY SHARE COPY SNAPSHOT DELETE EXPORT REFRESH

SCHEMA DETAILS PREVIEW TABLE EXPLORER PREVIEW INSIGHTS PREVIEW LINEAGE DATA PROFILE DATA QUALITY

Row	CoC_Number	State	CoC_Name	Overall_Homele	Sheltered_ES_H	Sheltered_TH_H	Sheltered_SH_H	Sheltered_Total
1	AK-500	AK	Anchorage CoC	1122	676	394	0	1070
2	AK-500	AK	Anchorage CoC	1105	573	292	0	865
3	AK-500	AK	Anchorage CoC	1147	704	393	0	1097
4	AK-500	AK	Anchorage CoC	1094	751	249	0	1000
5	AK-500	AK	Anchorage CoC	1208	694	335	0	1029
6	AK-500	AK	Anchorage CoC	1023	665	305	0	970
7	AK-500	AK	Anchorage CoC	1128	654	319	0	973
8	AK-501	AK	Alaska Balance of State CoC	835	434	199	0	633
9	AK-501	AK	Alaska Balance of State CoC	717	363	215	0	578
10	AK-501	AK	Alaska Balance of State CoC	922	497	210	0	707
11	AK-501	AK	Alaska Balance of State CoC	761	474	195	0	669
12	AK-501	AK	Alaska Balance of State CoC	824	461	210	0	671
13	AK-501	AK	Alaska Balance of State CoC	748	433	177	0	610
14	AK-501	AK	Alaska Balance of State CoC	766	435	184	0	619
15	AL-500	AL	Birmingham/Jefferson, St. Clair...	1228	531	467	34	1032
16	AL-500	AL	Birmingham/Jefferson, St. Clair...	1329	387	497	31	915
17	AL-500	AL	Birmingham/Jefferson, St. Clair...	1707	347	630	32	1009
18	AL-500	AL	Birmingham/Jefferson, St. Clair...	1153	448	448	20	916

BigQuery Data Exploration Part 2

Step 1 - Open the dataset you created last time, which should be a table named Homelessness. Once you have opened it, click on the Query button and open a new query in a new tab.

Step 2 - Using this dataset, answer the following questions. These first few questions can be answered by using ORDER BY and WHERE clauses. For reference, a list of all of the state abbreviations is listed HERE.

FUNCTIONS USED:

SELECT- select data from database

FROM- used to specify which table to select or delete data from

COUNT(*)- finds the number of rows that matches a specified criterion

WHERE- used to filter records you want selected

GROUP BY- groups rows that have save values into summary rows

ORDER BY- used to sort the results set in ascending or descending order

DESC- descending order

ASC- ascending order

LIMIT- select a limited number of records

SUM- finds the total sum of a numeric column

HAVING- used to filter rows based on a condition before grouping

AND- used to filter records based on more than one condition

AS- used to rename a column or table with an alias

COUNT DISTINCT- counts the number of unique values in a specified column

1. **We want to develop some new programs to help unaccompanied homeless children under 18 years old, and need some locations to start some programs. What are the top 3 Continuum of Care areas (CoC_Name in the table) with the highest number of unaccompanied homeless youth under 18 in the year 2018.**

Answer: AK-500, AK-501,AL-500

SQL Code Used:

```
SELECT CoC_Number, COUNT(*) AS youth_count
FROM `green-reporter-446623-m8.Exploration_Project.homelessness`
WHERE Homeless_Unaccompanied_Youth_Under_18 < 18 AND count_year = 2018
Group by CoC_Number
order by youth_count DESC
LIMIT 3;
```

Query results

JOB INFORMATION | RESULTS | CHART

Row	CoC_Number	youth_count
1	AK-500	1
2	AK-501	1
3	AL-500	1

2. We suspect that in Delaware (state abbreviation is "DE"), the number of unsheltered homeless people has been increasing over the past 7 years. Is this statement true? How do you know?

Answer: False. Based on the data gathered and analysis done, the number of the unsheltered homeless has been decreasing since 2012, started off high but as the years went on, decreased, then increased in the last 2 years.

SQL Code Used:

```
SELECT Count_year, SUM(Unsheltered_Homeless) AS total_unsheltered
FROM  `green-reporter-446623-m8.Exploration_Project.homelessness`
GROUP BY Count_year
ORDER BY Count_year ASC;
```

Query results

JOB INFORMATION | RESULTS

Row	Count_year	total_unsheltered
1	2012	231398
2	2013	195666
3	2014	175399
4	2015	173268
5	2016	176357
6	2017	190129
7	2018	194467

3. The Safe Haven program was created in 1992 to provide shelter for people who are homeless and have a serious mental illness. However, funding for Safe Havens was cut

in 2009, so no new Safe Havens could be created. Looking at data from only 2018, answer the following questions:

a. In 2018, how many different locations had at least 1 person as a Sheltered_SH?

Answer: 90

SQL Code Used:

```
SELECT COUNT(DISTINCT CoC_Name) AS location_count
FROM  `green-reporter-446623-m8.Exploration_Project.homelessness`
WHERE Sheltered_SH_Homeless > 0 AND Count_year = 2018
```

Query results

JOB INFORMATION

Row	location_count
1	90

b. In 2018, what was the CoC_Name of the top 3 locations in terms of number of Sheltered_SH?

Answer: 1. Philadelphia CoC, 2.Reno, Sparks/Washoe County, 3. Indianapolis CoC

SQL Code Used:

```
SELECT CoC_Name, SUM(Sheltered_SH_Homeless) AS total_sheltered
FROM `green-reporter-446623-m8.Exploration_Project.homelessness`
WHERE Count_year = 2018
GROUP BY CoC_Name
ORDER BY total_sheltered DESC
LIMIT 3;
```

Query results

JOB INFORMATION | RESULTS | CHART

Row	CoC_Name	total_sheltered
1	Philadelphia CoC	235
2	Reno, Sparks/Washoe County ...	185
3	Indianapolis CoC	68

4. What are the top 7 states in terms of Overall Homeless population in the year 2018? (Hint: This question will require the use of a GROUP BY clause in your query)

Answer: CA, NY, FL, TX, WA, OR, PA

SQL Code Used:

```
SELECT State, SUM(Homeless_individuals) AS overall_homeless
FROM `green-reporter-446623-m8.Exploration_Project.homelessness`
WHERE Count_year = 2018
GROUP BY State
ORDER BY Overall_Homeless DESC
LIMIT 7;
```

Query results

JOB INFORMATION | RESULTS | CHART

Row	State	overall_homeless
1	CA	109008
2	NY	39827
3	FL	21443
4	TX	19199
5	WA	16424
6	OR	11139
7	PA	8163

5. Below you will see a table of all of the states and their populations in 2018. If the Overall Homeless population were only correlated with the total state population, then the top 7 states for total population and the top 7 states for Overall Homeless population would be the same.

a. **Is this the case? If not, which states are not lining up the same?**
   - Not the case. States not lining up the same are OH and IL, based on the table below, those new states are added and WA, OR have been removed. Might be due to the total population not correlating to the overall homeless population in the states.

b. **We would say that homelessness is overrepresented in a state if that state listed higher in the Homeless ranking than it did in total population ranking, and it would be underrepresented if the homeless ranking was lower than the population ranking. With that in mind, which states in the top 7 homelessness list would be overrepresented for homelessness?**

- States that have been overrepresented in the homelessness lists are NY. It is listed higher in the homeless ranking than in total population ranking

6. In order to create better policies to support homeless individuals, we want to study locations that are doing a good job providing shelter. To do so, we need to identify places that have a relatively large number of homeless, but a relatively small number of unsheltered homeless.

a. **Generate a list of all locations, in 2018, that have more than 1000 Overall Homeless, but less than 100 unsheltered homeless.**

Answer: Maine Statewide CoC, Omaha, Council Bluffs CoC, Cleveland/Cuyahoga County C

SQL Code Used:

```
SELECT CoC_Name,
     SUM(Homeless_individuals) AS overall_homeless,
     SUM(Unsheltered_Homeless) AS unsheltered_homeless
FROM `green-reporter-446623-m8.Exploration_Project.homelessness`
WHERE Count_year = 2018
```

```
GROUP BY CoC_Name
HAVING SUM(Homeless_individuals) > 1000
AND (unsheltered_homeless) < 100;
```

Query results SAVE RESULTS

JOB INFORMATION | RESULTS | CHART | JSON | EXE

Row	CoC_Name	overall_homeless	unsheltered_homele
1	Maine Statewide CoC	1450	98
2	Omaha, Council Bluffs CoC	1122	64
3	Cleveland/Cuyahoga County C...	1377	78

b. From that same list, in which locations do Unsheltered Homeless account for less than 2% of the Overall Homeless population?

Answer: None based on my analysis. They all have 5-6% range, none less than 2%

SQL Code used:

```
SELECT CoC_Name,
    SUM(Homeless_individuals) AS Overall_homeless,
    SUM(Unsheltered_Homeless) AS Unsheltered_homeless,
   (SUM(Unsheltered_Homeless) / SUM(Homeless_individuals)) * 100 AS
unsheltered_percentage
FROM `green-reporter-446623-m8.Exploration_Project.homelessness`
WHERE Count_year = 2018
GROUP BY CoC_Name
HAVING SUM(Homeless_individuals) > 1000
  AND (Unsheltered_Homeless) < 100;
```

Query results

SAVE RESULTS OPEN IN

JOB INFORMATION | RESULTS | CHART | JSON | EXECUTION DETAILS

Row	CoC_Name	Overall_homeless	Unsheltered_homele	unsheltered_percent
1	Maine Statewide CoC	1450	98	6.758620689655...
2	Omaha, Council Bluffs CoC	1122	64	5.704099821746...
3	Cleveland/Cuyahoga County C...	1377	78	5.664488017429...

US State and Territory* Population Ranking for 2018

Rank	State	Population	Rank	State	Population
1	California (CA)	39776830	29	Connecticut (CT)	3588683
2	Texas (TX)	28704330	30	Puerto Rico (PR)	3193354
3	Florida (FL)	21312211	31	Iowa (IA)	3160553
4	New York (NY)	19862512	32	Utah (UT)	3159345
5	Pennsylvania (PA)	12823989	33	Nevada (NV)	3056824
6	Illinois (IL)	12768320	34	Arkansas (AR)	3020327
7	Ohio (OH)	11694664	35	Mississippi (MS)	2982785
8	Georgia (GA)	10545138	36	Kansas (KS)	2918515
9	North Carolina (NC)	10390149	37	New Mexico (NM)	2090708
10	Michigan (MI)	9991177	38	Nebraska (NE)	1932549
11	New Jersey (NJ)	9032872	39	West Virginia (WV)	1803077
12	Virginia (VA)	8525660	40	Idaho (ID)	1753860
13	Washington (WA)	7530552	41	Guam (GU)	168678

14	Arizona (AZ)	7123898	42	Hawaii (HI)	1426393
15	Massachusetts (MA)	6895917	43	New Hampshire (NH)	1350575
16	Tennessee (TN)	6782564	44	Maine (ME)	1341582
17	Indiana (IN)	6699629	45	Montana (MT)	1062330
18	Missouri (MO)	6135888	46	Rhode Island (RI)	1061712
19	Maryland (MD)	6079602	47	Delaware (DE)	971180
20	Wisconsin (WI)	5818049	48	South Dakota (SD)	877790
21	Colorado (CO)	5684203	49	North Dakota (ND)	755238
22	Minnesota (MN)	5628162	50	Alaska (AK)	738068
23	South Carolina (SC)	5088916	51	Washington DC (DC)	703608
24	Alabama (AL)	4888949	52	Vermont (VT)	623960
25	Louisiana (LA)	4682509	53	Wyoming (WY)	573720
26	Kentucky (KY)	4472265	54	Virgin Islands (VI)	101365
27	Oregon (OR)	4199563	55	Northern Mariana Islands (MP)	50304
28	Oklahoma (OK)	3940521			

United States and Territories 2-letter Abbreviations

STATE(TERRITORY)		STATE(TERRITORY)		STATE(TERRITORY)	
Alabama	AL	Kentucky	KY	Ohio	OH
Alaska	AK	Louisiana	LA	Oklahoma	OK
Arizona	AZ	Maine	ME	Oregon	OR
Arkansas	AR	Maryland	MD	Pennsylvania	PA
American Samoa	AS	Massachusetts	MA	Puerto Rico	PR
California	CA	Michigan	MI	Rhode Island	RI
Colorado	CO	Minnesota	MN	South Carolina	SC
Connecticut	CT	Mississippi	MS	South Dakota	SD
Delaware	DE	Missouri	MO	Tennessee	TN
District of Columbia	DC	Montana	MT	Texas	TX
Florida	FL	Nebraska	NE	Trust Territories	TT
Georgia	GA	Nevada	NV	Utah	UT
Guam	GU	New Hampshire	NH	Vermont	VT
Hawaii	HI	New Jersey	NJ	Virginia	VA
Idaho	ID	New Mexico	NM	Virgin Islands	VI
Illinois	IL	New York	NY	Washington	WA
Indiana	IN	North Carolina	NC	West Virginia	WV
Iowa	IA	North Dakota	ND	Wisconsin	WI
Kansas	KS	Northern Mariana Islands	MP	Wyoming	WY