

Authors: Tony Fong, Michael Medina, Victor Verduzco  
Pablo Ilabaca, John Martinez  
Instructor: Jongwook Woo  
Date: 12/7/2022

#### Lab Tutorial

John Martinez (jmarti168@calstatela.edu), Michael Medina (mmedin126@calstatela.edu),  
Tony Fong(tfong9@calstatela.edu), Pablo Ilabaca (pilabac@calstatela.edu), Victor Verduzco  
(vverduz@calstatela.edu)

## Covid 19 Surveillance Data

### Objectives:

In this lab, you will:

- Get the dataset from Google Drive using wget
- Upload the dataset to the tmp folder
- Create a database within beeline
- Create tables based on the data using HiveQL commands
- Download the data to the local computer
- Use Excel for visualization of the data

### Platform Specifications:

- Oracle Cloud
- CPU Speed: 1995.309 MHz:
- # of CPU Cores: 32
- # of nodes: 3
- Total Memory Size: 58GB

1. open a shell terminal – git bash, minty, putty etc- and run the ssh command to connect to the Hadoop Cloud.

**\$ssh yourusername@ipaddress**

2. Download the file using wget

```
wget --load-cookies /tmp/cookies.txt
"https://docs.google.com/uc?export=download&confirm=$(wget --quiet --save-cookies /tmp/cookies.txt --keep-session-cookies --no-check-certificate
'https://docs.google.com/uc?export=download&id=FILEID' -O- | sed -rn
's/.*confirm=([0-9A-Za-z_]+).*/\1\n/p)'&id=1s-9aKPqcQq8id8oGgW6HBCQzuXKBR1xO" -O
covid19data.csv && rm -rf /tmp/cookies.txt
```

```
-bash-4.2$ wget --load-cookies /tmp/cookies.txt "https://docs.google.com/uc?export=download&confirm=$(wget --quiet --save-cookies /tmp/cookies.txt --keep-session-cookies --no-check-certificate 'https://docs.google.com/uc?export=download&id=F1L2D' -O- | sed -rn 's/.*confirm=([0-9A-Za-z_]+).*/\1\n/p)'&id=1s-9aKPqcQq8id8oGgW6HBCQzuXKBR1xO" -O covid19data.csv && rm -rf /tmp/cookies.txt
2022-12-07 19:33:20-- https://docs.google.com/uc?export=download&confirm=$(wget --quiet --save-cookies /tmp/cookies.txt --keep-session-cookies --no-check-certificate 'https://docs.google.com/uc?export=download&id=F1L2D' -O- | sed -rn 's/.*confirm=([0-9A-Za-z_]+).*/\1\n/p)'&id=1s-9aKPqcQq8id8oGgW6HBCQzuXKBR1xO
Resolving docs.google.com (docs.google.com)... 142.250.188.238, 2607:f8b0:4007:80f::200e
Connecting to docs.google.com (docs.google.com)... 142.250.188.238:443... connected.
HTTP request sent, awaiting response... 303 See other
Location: https://doc-0s-0k-docs.googleusercontent.com/docs/securesc/habr937gcuc717deffksulhgh7hpl/965h5kf0221rj4d4htg916m7ruql/1670441550000/13778178939762848769/1s-9aKPqcQq8id8oGgW6HBCQzuXKBR1xO?e=download&uiid=eea2c644-1db7-4551-be45-6c29664399ef [Following]
Warning: 130cards not supported in HTTP.
2022-12-07 19:33:20-- https://doc-0s-0k-docs.googleusercontent.com/docs/securesc/habr937gcuc717deffksulhgh7hpl/965h5kf0221rj4d4htg916m7ruql/1670441550000/13778178939762848769/1s-9aKPqcQq8id8oGgW6HBCQzuXKBR1xO?e=download&uiid=eea2c644-1db7-4551-be45-6c29664399ef
Resolving doc-0s-0k-docs.googleusercontent.com (doc-0s-0k-docs.googleusercontent.com)... 142.250.68.65, 2607:f8b0:4007:818::2001
Connecting to doc-0s-0k-docs.googleusercontent.com (doc-0s-0k-docs.googleusercontent.com) [142.250.68.65]:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 2148939052 (2.0G) [application/octet-stream]
Saving to: 'covid19data.csv'

100%[=====] 2,148,939,052 1.87MB/s in 11s

2022-12-07 19:33:32 (183 MB/s) - 'covid19data.csv' saved [2148939052/2148939052]
```

3. You have to upload the files to hdfs folder coviddata. Run the following HDFS commands to create and list coviddata directory in HDFS:

```
$ hdfs dfs -mkdir tmp/covid19data
$ hdfs dfs -put covid19data.csv tmp/covid19data/
```

```
-bash-4.2$ hdfs dfs -mkdir tmp/covid19data/
-bash-4.2$ hdfs dfs -put covid19data.csv tmp/covid19data/
-bash-4.2$ hdfs dfs -ls tmp/covid19data/
Found 1 items
-rw-r--r-- 3 pilabac hdfs 2148939052 2022-12-07 19:38 tmp/covid19data/covid19data.csv
-bash-4.2$ |
```

4. open hive

```
$ beeline
```

```
-bash-4.2$ beeline
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/odh/1.1.2/hive/lib/log4j-slf4j-impl-2.17.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/odh/1.1.2/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Connecting to jdbc:hive2://bigdaiwn0.sub02180640120.trainingvcn.oraclevcn.com:2181,bigdaimn0.sub02180640120.trainingvcn.oraclevcn.com:2181,bigdaiun0.sub02180640120.trainingvcn.oraclevcn.com:2181/default;password=pilabac;serviceDiscoveryMode=zooKeeper;user=pilabac;zooKeeperNamespace=hiveserver2
22/12/07 19:43:42 [main-EventThread]: ERROR impls.EnsembleTracker: Invalid config event received: {server.1=bigdaimn0.sub02180640120.trainingvcn.oraclevcn.com:2888:3888:participant, version=0, server.3=bigdaiwn0.sub02180640120.trainingvcn.oraclevcn.com:2888:3888:participant, server.2=bigdaiun0.sub02180640120.trainingvcn.oraclevcn.com:2888:3888:participant}
22/12/07 19:43:42 [main-EventThread]: ERROR impls.EnsembleTracker: Invalid config event received: {server.1=bigdaimn0.sub02180640120.trainingvcn.oraclevcn.com:2888:3888:participant, version=0, server.3=bigdaiwn0.sub02180640120.trainingvcn.oraclevcn.com:2888:3888:participant, server.2=bigdaiun0.sub02180640120.trainingvcn.oraclevcn.com:2888:3888:participant}
22/12/07 19:43:42 [main]: INFO jdbc.HiveConnection: Connected to bigdaiun0.sub02180640120.trainingvcn.oraclevcn.com:10000
Connected to: Apache Hive (version 3.1.2)
Driver: Hive JDBC (version 3.1.2)
Transaction isolation: TRANSACTION_REPEATABLE_READ
Beeline version 3.1.2 by Apache Hive
0: jdbc:hive2://bigdaiwn0.sub02180640120.tra> |
```

## 5. Create your own database and use that database

```
$ create database Covid19;
$ use database Covid19;
```

```
0: jdbc:hive2://bigdaiwn0.sub02180640120.tra> use covid19;
INFO : Compiling command(queryId=hive_20221207195141_6f7b24f9-3e2d-45c4-80c0-665e3464f455): use covid19
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Semantic Analysis Completed (retrial = false)
INFO : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO : Completed compiling command(queryId=hive_20221207195141_6f7b24f9-3e2d-45c4-80c0-665e3464f455); Time taken: 0.03 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20221207195141_6f7b24f9-3e2d-45c4-80c0-665e3464f455): use covid19
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20221207195141_6f7b24f9-3e2d-45c4-80c0-665e3464f455); Time taken: 0.216 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager
No rows affected (0.26 seconds)
0: jdbc:hive2://bigdaiwn0.sub02180640120.tra> |
```

## 6. Create external table "Covid19Data"

```
-- create the covid19 table on comma-seperated covid19data
create external table if not exists Covid19 (case_months string,
res_state string,
state_fips_code string,
res_country string,
county_fips_county string,
age_group string,
sex string,
race string,
ethnicity string,
case_positive_specimen_interval int,
case_onset_interval int,
process string,
exposure_yn string,
current_status string,
symptom_status string,
hosp_yn string,
icu_yn string,
```

death\_yn string,  
underlying\_conditions\_yn string)  
row format delimited fields terminated by ","  
stored as textfile location '/user/tfong9/tmp/covid19data'  
tblproperties ('skip.header.line.count' = '1');

```
0: jdbc:hive2://bigdaiwn0.sub02180640120.traib> describe formatted covid19;
INFO : Compiling command(queryId=hive_20221207195421_3b6645da-a0a6-498b-be0e-f6a4c47690eb): describe formatted covid19
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Semantic Analysis Completed (retrial = false)
INFO : Returning Hive schema: Schema(FieldSchemas:[FieldSchema(name:col_name, type:string, comment:from deserializer), FieldSchema(n
me:data_type, type:string, comment:from deserializer), FieldSchema(name:comment, type:string, comment:from deserializer)], properties
null)
INFO : Completed compiling command(queryId=hive_20221207195421_3b6645da-a0a6-498b-be0e-f6a4c47690eb); Time taken: 0.046 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20221207195421_3b6645da-a0a6-498b-be0e-f6a4c47690eb): describe formatted covid19
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20221207195421_3b6645da-a0a6-498b-be0e-f6a4c47690eb); Time taken: 0.28 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager
```

| col_name                        | data_type  | comment    |
|---------------------------------|--|------------|
| # col_name                      | data_type  | comment    |
| case_months                     | string   |            |
| res_state                       | string   |            |
| state_fips_code                 | string   |            |
| res_country                     | string   |            |
| county_fips_code                | string   |            |
| age_group                       | string   |            |
| sex                             | string   |            |
| race                            | string   |            |
| ethnicity                       | string   |            |
| case_positive_specimen_interval | int  |            |
| case_onset_interval             | int  |            |
| process                         | string   |            |
| exposure_yn                     | string   |            |
| current_status                  | string   |            |
| symptom_status                  | string   |            |
| hosp_yn                         | string   |            |
| icu_yn                          | string   |            |
| death_yn                        | string   |            |
| underlying_conditions_yn        | string   |            |
|                                 | NULL   | NULL       |
| # Detailed Table Information    | NULL   | NULL       |
| Database:                       | covid19  | NULL       |
| OwnerType:                      | USER   | NULL       |
| Owner:                          | tfong9   | NULL       |
| CreateTime:                     | Tue Dec 06 19:25:44 GMT 2022   | NULL       |
| LastAccessTime:                 | UNKNOWN  | NULL       |
| Retention:                      | 0  | NULL       |
| Location:                       | hdfs://bigdaimn0.sub02180640120.trainingvcn.oraclevcn.com:8020/user/tfong9/tmp/coviddata | NULL       |
| Table Type:                     | EXTERNAL_TABLE   | NULL       |
| Table Parameters:               | NULL   | NULL       |
|                                 | EXTERNAL   | TRUE       |
|                                 | bucketing_version  | 2          |
|                                 | numFiles   | 1          |
|                                 | skip.header.line.count   | 1          |
|                                 | totalSize  | 2148939052 |
|                                 | transient_lastDdlTime  | 1670354744 |
|                                 | NULL   | NULL       |
| # Storage Information           | NULL   | NULL       |
| SerDe Library:                  | org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe                                       | NULL       |
| InputFormat:                    | org.apache.hadoop.mapred.TextInputFormat   | NULL       |
| OutputFormat:                   | org.apache.hadoop.hive ql.io.HiveIgnoreKeyTextOutputFormat                               | NULL       |
| Compressed:                     | No   | NULL       |
| Num Buckets:                    | -1   | NULL       |
| Bucket Columns:                 | []   | NULL       |
| Sort Columns:                   | []   | NULL       |
| Storage Desc Params:            | NULL   | NULL       |
|                                 | field.delim  | ,          |
|                                 | serialization.format   | ,          |

Now run the following HiveQL at the query editor to see how the dataset looks like

```
select * from covid19 limit 10;
```

| covid19_case_months | covid19_res_state   | covid19_state_fips_code   | covid19_res_country    | covid19_county_fips_code | covid19_age_group | covid19_sex      | covid19_race                     | covid19_ethnicity   | covid19_case_positive_specimen_interval | covid19_case_onset_interval |
|---------------------|---------------------|---------------------------|------------------------|--------------------------|-------------------|------------------|----------------------------------|---------------------|---|-----------------------------|
| covid19_process     | covid19_exposure_yn | covid19_current_status    | covid19_symptom_status | covid19_hosp_yn          | covid19icu_yn     | covid19_death_yn | covid19_underlying_conditions_yn |                     |   |                             |
| 2021-12             | CA                  | 06                        | VENTURA                | 08011                    | 18 to 49 years    | Female           | White                            | Non-Hispanic/Latino | NULL                                    | NULL                        |
| Missing             | Missing             | Laboratory-confirmed case | Unknown                | Missing                  | Missing           | Missing          | White                            | Non-Hispanic/Latino | NULL                                    | NULL                        |
| 2021-09             | TX                  | 48                        | TARRANT                | 48419                    | 18 to 49 years    | Male             | White                            | Non-Hispanic/Latino | NULL                                    | NULL                        |
| Missing             | Missing             | Laboratory-confirmed case | Missing                | Missing                  | Missing           | Missing          | White                            | Non-Hispanic/Latino | NULL                                    | NULL                        |
| 2022-01             | MA                  | 25                        | MIDDLESEX              | 25002                    | 18 to 49 years    | Female           | Unknown                          | Unknown             | 0                                       | NULL                        |
| Missing             | Missing             | Laboratory-confirmed case | Missing                | Missing                  | Missing           | Missing          | White                            | Non-Hispanic/Latino | 0                                       | 0                           |
| 2020-12             | NY                  | 36                        | KINGS                  | 36047                    | 65+ years         | Female           | White                            | Non-Hispanic/Latino | 0                                       | 0                           |
| Missing             | Missing             | Laboratory-confirmed case | Symptomatic            | Missing                  | Missing           | Missing          | White                            | Non-Hispanic/Latino | 0                                       | 0                           |
| 2022-01             | NJ                  | 34                        | ESSEX                  | 34013                    | 0 - 17 years      | Male             | White                            | Non-Hispanic/Latino | 0                                       | NULL                        |
| Missing             | Missing             | Laboratory-confirmed case | No                     | Missing                  | Missing           | Missing          | Unknown                          | Non-Hispanic/Latino | 0                                       | NULL                        |
| 2022-06             | CA                  | 06                        | SACRAMENTO             | 06067                    | 18 to 49 years    | Female           | Unknown                          | Non-Hispanic/Latino | 0                                       | NULL                        |
| Missing             | Missing             | Laboratory-confirmed case | Unknown                | Missing                  | Missing           | Missing          | White                            | Non-Hispanic/Latino | 0                                       | NULL                        |
| 2021-12             | NJ                  | 34                        | OCEAN                  | 34029                    | 50 to 64 years    | Female           | White                            | Non-Hispanic/Latino | 0                                       | NULL                        |
| Missing             | Missing             | Laboratory-confirmed case | Missing                | Missing                  | Missing           | Missing          | Black                            | Non-Hispanic/Latino | 0                                       | 0                           |
| 2021-09             | NY                  | 36                        | ROCK                   | 36055                    | 0 - 17 years      | Female           | Black                            | Non-Hispanic/Latino | 0                                       | 0                           |
| Missing             | Missing             | Laboratory-confirmed case | Symptomatic            | Missing                  | Missing           | Missing          | Black                            | Non-Hispanic/Latino | 0                                       | 0                           |
| 2021-07             | FL                  | 12                        | PALM BEACH             | 12099                    | 18 to 49 years    | Male             | Black                            | Non-Hispanic/Latino | 0                                       | 0                           |
| Missing             | Missing             | Laboratory-confirmed case | Symptomatic            | Missing                  | Missing           | Missing          | White                            | Non-Hispanic/Latino | 0                                       | 0                           |
| 2022-05             | FL                  | 12                        | PIKE                   | 12033                    | 18 to 49 years    | Male             | White                            | Non-Hispanic/Latino | 0                                       | 0                           |
| Missing             | Missing             | Laboratory-confirmed case | Missing                | Missing                  | Missing           | Missing          | White                            | Non-Hispanic/Latino | 0                                       | 0                           |

## 7.create external table "patient\_profile"

- - create the patient\_profile table on comma-seperated covid19data

```
CREATE EXTERNAL TABLE IF NOT EXISTS patient_profile(age_group STRING, sex
STRING, race STRING, ethnicity STRING, res_state STRING, underlying_conditions
STRING)
```

row format delimited fields terminated by ","

```
STORED AS TEXTFILE LOCATION '/user/ufong9/tmp/covid19data';
```

insert overwrite table patient\_profile

```
select age_group, sex, race, ethnicity, res_state, underlying_conditions_yn
from covid19;
```

Now run the following HiveQL at the query editor to see how the dataset looks like

```
Select * from patient_profile limit 10;
```

| patient_profile.age_group | patient_profile.sex | patient_profile.race | patient_profile.ethnicity | patient_profile.res_state | patient_profile.underlying_conditions |
|---------------------------|---------------------|----------------------|---------------------------|---------------------------|---------------------------------------|
| 18 to 49 years            | Female              | White                | Non-Hispanic/Latino       | CA                        |                                       |
| 18 to 49 years            | Male                | White                | Non-Hispanic/Latino       | TX                        |                                       |
| 18 to 49 years            | Female              | Unknown              | Unknown                   | MA                        |                                       |
| 65+ years                 | Female              | White                | Non-Hispanic/Latino       | NY                        |                                       |
| 0 - 17 years              | Male                | White                | Non-Hispanic/Latino       | NJ                        |                                       |
| 18 to 49 years            | Female              | Unknown              | Non-Hispanic/Latino       | CA                        |                                       |
| 50 to 64 years            | Female              | White                | Non-Hispanic/Latino       | NJ                        |                                       |
| 0 - 17 years              | Female              | Black                | Non-Hispanic/Latino       | NY                        |                                       |
| 18 to 49 years            | Male                | Black                | Non-Hispanic/Latino       | FL                        |                                       |
| 18 to 49 years            | Male                | White                | Non-Hispanic/Latino       | FL                        |                                       |

8. Now run the following HiveQL at the Query editor to see the number of cases

```
select case_months, count(sex) as number_of_cases
from covid19
group by case_months;
```

INFO: The console file mode is disabled.

| case_months | number_of_cases |
|-------------|-----------------|
| 2022-01     | 2648280         |
| 2020-05     | 153218          |
| 2020-04     | 210489          |
| 2021-12     | 1412096         |
| 2022-03     | 175036          |
| 2020-03     | 110842          |
| 2020-07     | 345947          |
| 2020-12     | 985109          |
| 2022-04     | 269741          |
| 2020-09     | 183285          |
| 2021-07     | 299067          |
| 2022-02     | 423524          |
| 2020-11     | 598596          |
| 2021-01     | 1024501         |
| 2021-03     | 380664          |
| 2021-06     | 77038           |
| 2022-10     | 109129          |
| 2021-05     | 122949          |
| 2022-09     | 221328          |
| 2021-09     | 426063          |
| 2020-01     | 548             |
| 2020-10     | 259393          |
| 2022-07     | 598233          |
| 2022-05     | 699285          |
| 2022-06     | 519352          |
| 2021-11     | 307190          |
| 2022-08     | 451627          |
| 2020-02     | 1227            |
| 2021-02     | 391593          |
| 2021-08     | 585600          |
| 2021-10     | 295130          |
| 2020-06     | 213922          |
| 2021-04     | 274905          |
| 2020-08     | 225092          |

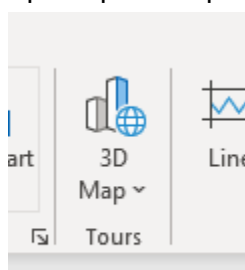
9. Now download data into your PC

```
- - download to local file
hdfs dfs -get tmp/covid19data/0000000_0
- - download file to your PC
scp tfong9@144.24.14.145:/home/tfong9/0000000_0 covid19data.csv
```

10. Loading Data into and Visualizing using Power Map in Excel

Create column names for each column


Open up 3d maps






You need to select the properties and values in the layer as follows.








You need to select the properties and values in the layer as follows.


 Add Layer ✕


▼ Layer 1   

▼ Data


    


Location 100%

☒ state State/Province ▼ 


 Add Field

Size


gender (Count - Not Blank) ▼ 

 Add Field


Category

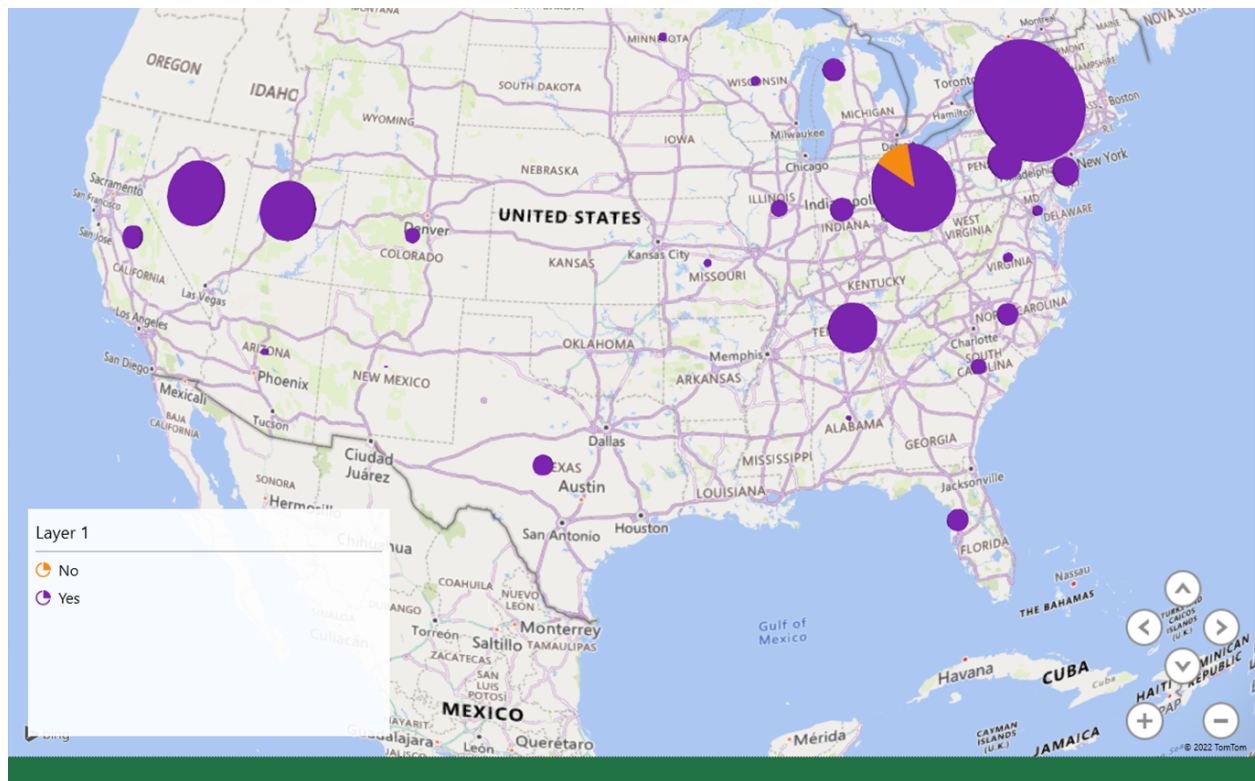
underlying condition 

Time

 Add Field

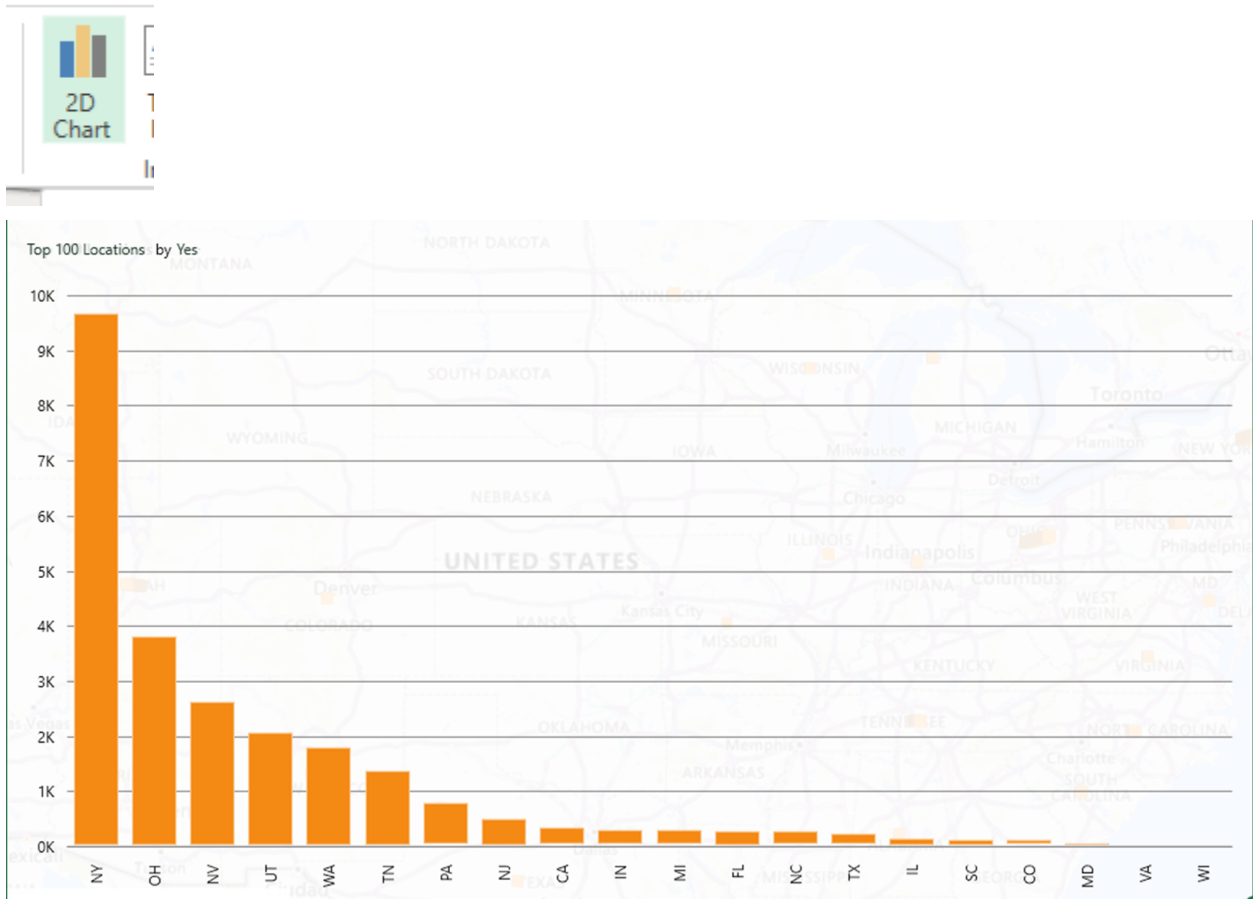
▼ Filters

 Add Filter





Click on 2D chart



You need to select the properties and values in the layer as follows and click on 2d Chart

Layer 1

Data

Location

state State/Province

Height

gender (Count - Not Blank)

Category

death

Time

Filters

death

Layer Options

2D Chart

