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 and PowerBI for visualization of the data

Platform Specifications:

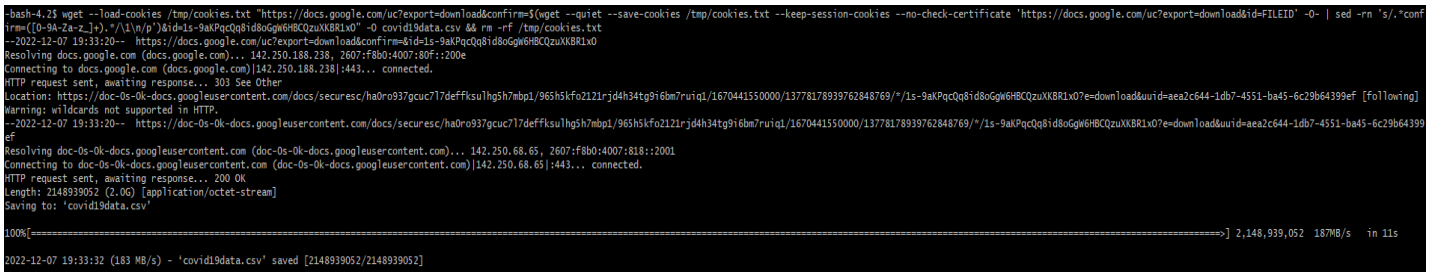
- Oracle Cloud
- CPU Speed: 1995. 309 MHz:
- # of CPU Cores: 32
- # of nodes: 3
- Total Memory Size: 58GB

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

$ssh yourusername@ipaddress [144.24.14.145]

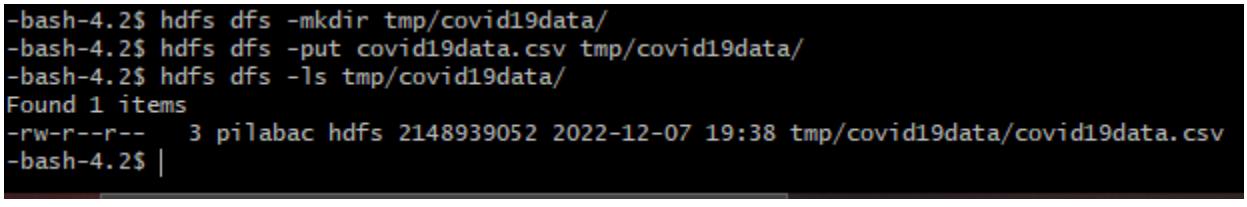
2. Download the covid19 dataset file using wget

```
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=1s-9aKPqcQq8id8oGgW6HBCQzuXKBR1x
O' -O- | sed -rn
's/.*confirm=([0-9A-Za-z_]+).*/\1\n/p')&id=1s-9aKPqcQq8id8oGgW6HBCQzuXKBR1xO" -O
covid19data.csv
```



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/
```



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;zooK
eeperNamespace=hiveserver2
22/12/07 19:43:42 [main-EventThread]: ERROR imps.EnsembleTracker: Invalid config event received: {server.1=bigdaimn0.sub02180640120.tr
ainingvcn.oraclevcn.com:2888:3888:participant, version=0, server.3=bigdaiwn0.sub02180640120.trainingvcn.oraclevcn.com:2888:3888:partic
ipant, server.2=bigdaiun0.sub02180640120.trainingvcn.oraclevcn.com:2888:3888:participant}
22/12/07 19:43:42 [main-EventThread]: ERROR imps.EnsembleTracker: Invalid config event received: {server.1=bigdaimn0.sub02180640120.tr
ainingvcn.oraclevcn.com:2888:3888:participant, version=0, server.3=bigdaiwn0.sub02180640120.trainingvcn.oraclevcn.com:2888:3888:partic
ipant, 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.trai>
```

5. Create your own database and use that database

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

```
0: jdbc:hive2://bigdaiwn0.sub02180640120.trai> 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.trai>
```

6. Create external table “Covid19Data”

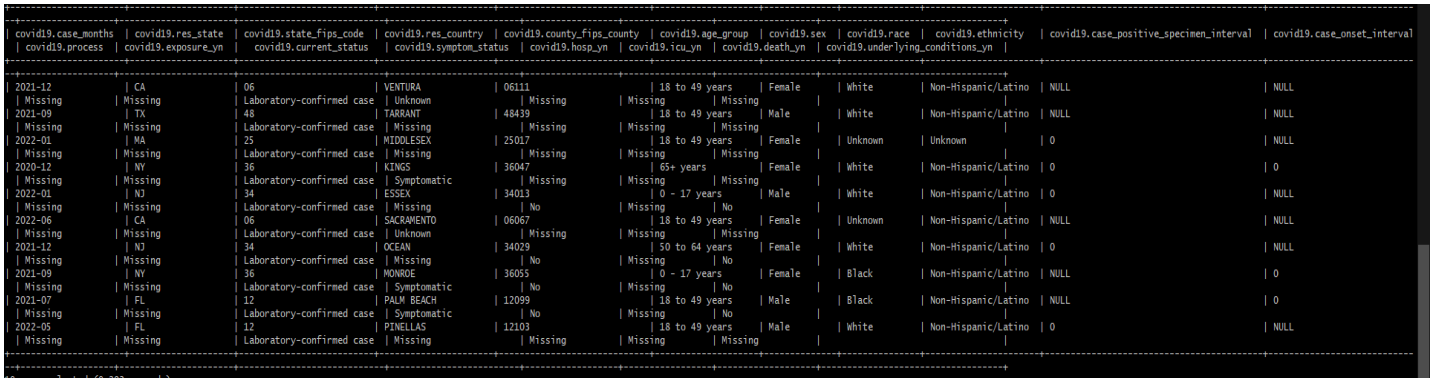
```
- - create the covid19 table on comma-seperated covid19data
CREATE EXTERNAL TABLE IF NOT EXISTS covid19 (case_months string,
res_state string,
state_fips_code string,
res_country string,
county_fips_county string,
age_group string,
sex string,
race string,
ethnicity string,
case_positive_specimen_interval 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.trai> 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_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                                             |                       |
|                                 | 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                               | ,                     |
+---------------------------------+----------------------------------------------------+-----------------------+
```

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_county	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	covid19.icu_yn	covid19.death_yn	covid19.underlying_conditions_yn
2021-12	CA	06	VENTURA	06111	18 to 49 years	Female	White	Non-Hispanic/Latino	NULL	NULL	Missing	Missing	Laboratory-confirmed case	Unknown	Missing	Missing	Missing	
2021-09	TX	48	TARRANT	48439	18 to 49 years	Male	White	Non-Hispanic/Latino	NULL	NULL	Missing	Missing	Laboratory-confirmed case	Missing	Missing	Missing	Missing	
2022-01	MA	25	MIDDLESEX	25017	18 to 49 years	Female	Unknown	Unknown	0	NULL	Missing	Missing	Laboratory-confirmed case	Missing	Missing	Missing	Missing	
2020-12	NY	36	KINGS	36047	65+ years	Female	White	Non-Hispanic/Latino	0	0	Missing	Missing	Laboratory-confirmed case	Symptomatic	Missing	Missing	Missing	
2022-01	NJ	34	ESSEX	34013	0 - 17 years	Male	White	Non-Hispanic/Latino	0	NULL	Missing	Missing	Laboratory-confirmed case	Missing	No	Missing	No	
2022-06	CA	06	SACRAMENTO	06067	18 to 49 years	Female	Unknown	Non-Hispanic/Latino	NULL	NULL	Missing	Missing	Laboratory-confirmed case	Unknown	Missing	Missing	Missing	
2021-12	NJ	34	OCEAN	34029	50 to 64 years	Female	White	Non-Hispanic/Latino	0	NULL	Missing	Missing	Laboratory-confirmed case	Missing	No	Missing	No	
2021-09	NY	36	MONROE	36055	0 - 17 years	Female	Black	Non-Hispanic/Latino	NULL	0	Missing	Missing	Laboratory-confirmed case	Symptomatic	No	Missing	No	
2021-07	FL	12	PALM BEACH	12099	18 to 49 years	Male	Black	Non-Hispanic/Latino	NULL	0	Missing	Missing	Laboratory-confirmed case	Symptomatic	No	Missing	No	
2022-05	FL	12	PINELLAS	12103	18 to 49 years	Male	White	Non-Hispanic/Latino	0	NULL	Missing	Missing	Laboratory-confirmed case	Missing	Missing	Missing	Missing	

7. Create external table “patient_profile”

```
- - create the patient_profile table on comma-seperated covid19data
CREATE EXTERNAL TABLE IF NOT EXISTS patient_profile(month STRING, age_group
STRING, sex STRING, race STRING, ethnicity STRING, res_state STRING,
underlying_conditions STRING, death STRING)
row format delimited fields terminated by ","
STORED AS TEXTFILE LOCATION '/user/tfong9/tmp/covid19data';

insert overwrite table patient_profile
Select case_months, age_group, sex, race, ethnicity, res_state, underlying_conditions_yn,
death_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.month	patient_profile.age_group	patient_profile.sex	patient_profile.race	patient_profile.ethnicity	patient_profile.res_state
2021-12	18 to 49 years	Female	White	Non-Hispanic/Latino	CA
2021-09	18 to 49 years	Male	White	Non-Hispanic/Latino	TX
2022-01	18 to 49 years	Female	Unknown	Unknown	MA
2020-12	65+ years	Female	White	Non-Hispanic/Latino	NY
2022-01	0 - 17 years	Male	White	Non-Hispanic/Latino	NJ
2022-06	18 to 49 years	Female	Unknown	Non-Hispanic/Latino	CA
2021-12	50 to 64 years	Female	White	Non-Hispanic/Latino	NJ
2021-09	0 - 17 years	Female	Black	Non-Hispanic/Latino	NY
2021-07	18 to 49 years	Male	Black	Non-Hispanic/Latino	FL
2022-05	18 to 49 years	Male	White	Non-Hispanic/Latino	FL

```
10 rows selected (0.341 seconds)
0: jdbc:hive2://bigdaiwn0.sub02180640120.trai>
```

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;
```

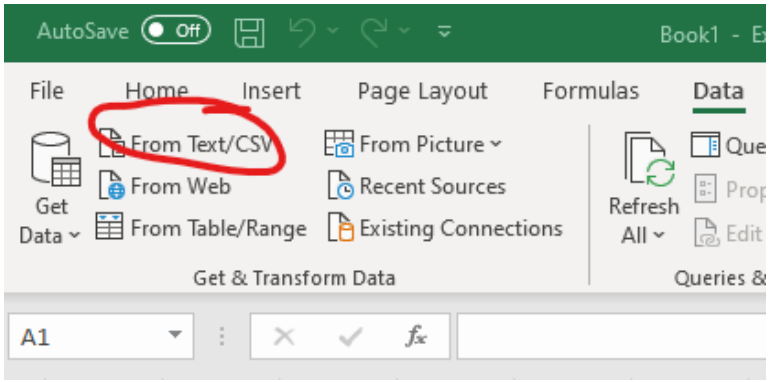
case_months	number_of_cases
2022-01	2648280
2020-05	153218
2020-04	210489
2021-12	1412096
2022-03	175036
2020-03	110842
2020-07	345947
2020-12	985109
2022-04	269741
2020-09	183285
2021-07	299067
2022-02	423524
2020-11	598596
2021-01	1024501
2021-03	380664
2021-06	77038
2022-10	109129
2021-05	122949
2022-09	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

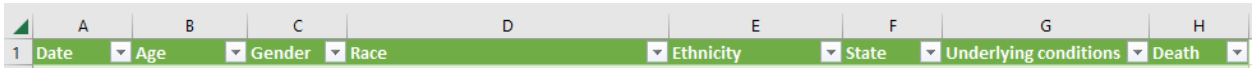
```
- - download to local file
hdfs dfs -get tmp/covid19data/000000_0
- - download file to your PC
scp tfong9@144.24.14.145:/home/tfong9/000000_0 covid19data.csv
```

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

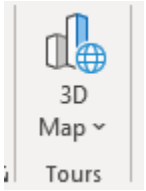
Import your covid19data.csv into Excel. You should then be able to open 3d maps for visualization.



Rename the table columns into the following: date, age, gender, race, ethnicity, state, underlying condition, and death

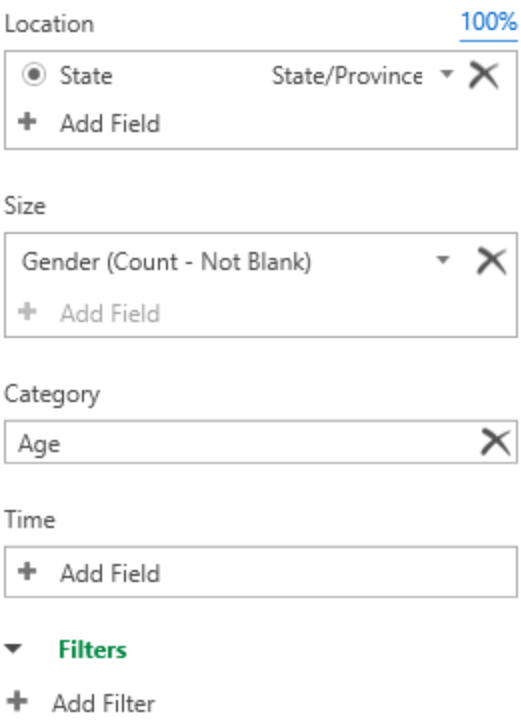


After renaming the columns, highlight all your columns and select the 3D maps under the “insert” tab

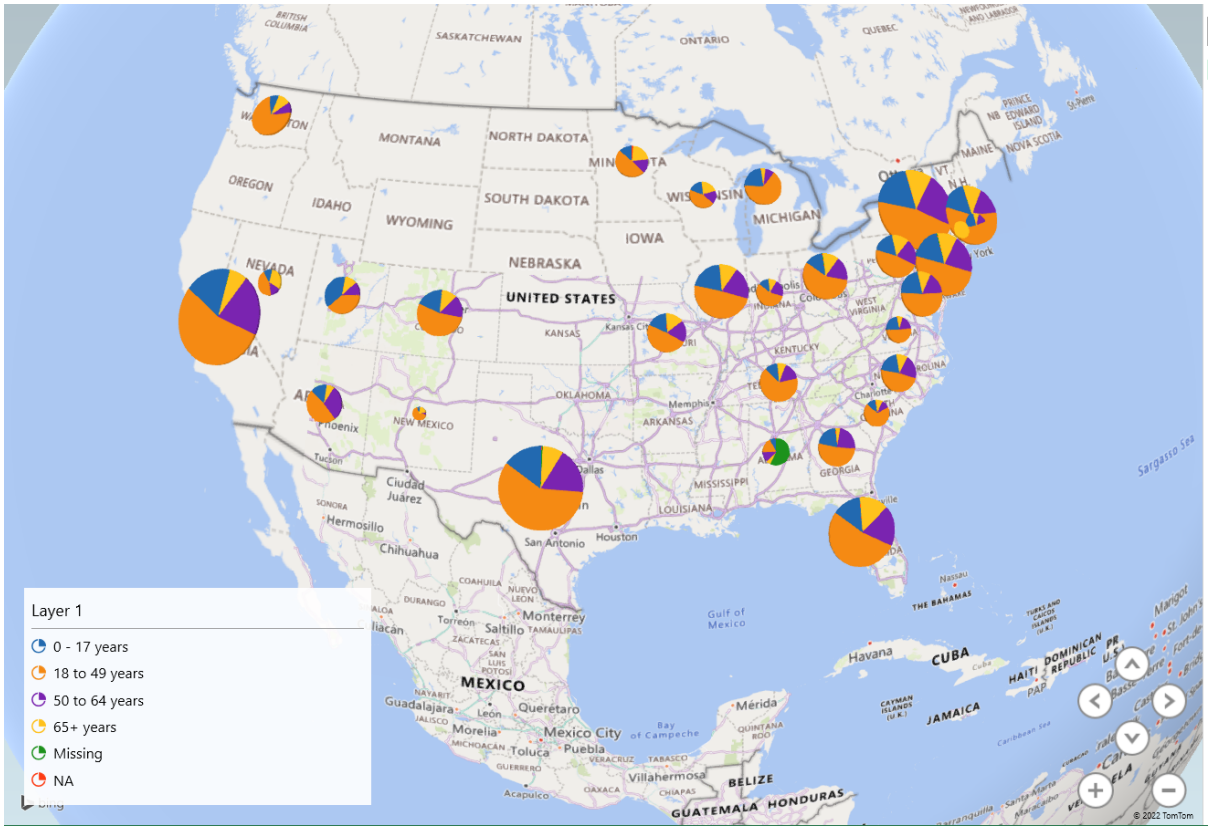


11. You need to select the properties and values in the layer as follows:

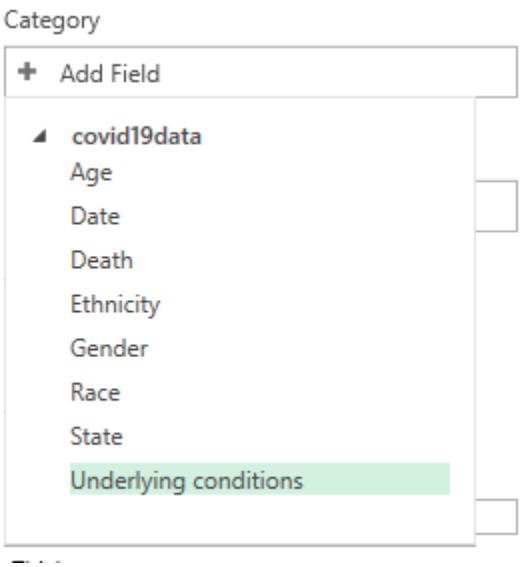
- For Location, add state
- For Size, click on gender (size may also appear as “height” in the beginning)
- For Category, add Age



After you should change the graph to pie graphs



Afterwards, you should change the Category from ***Age Groups*** to ***Underlying Conditions***

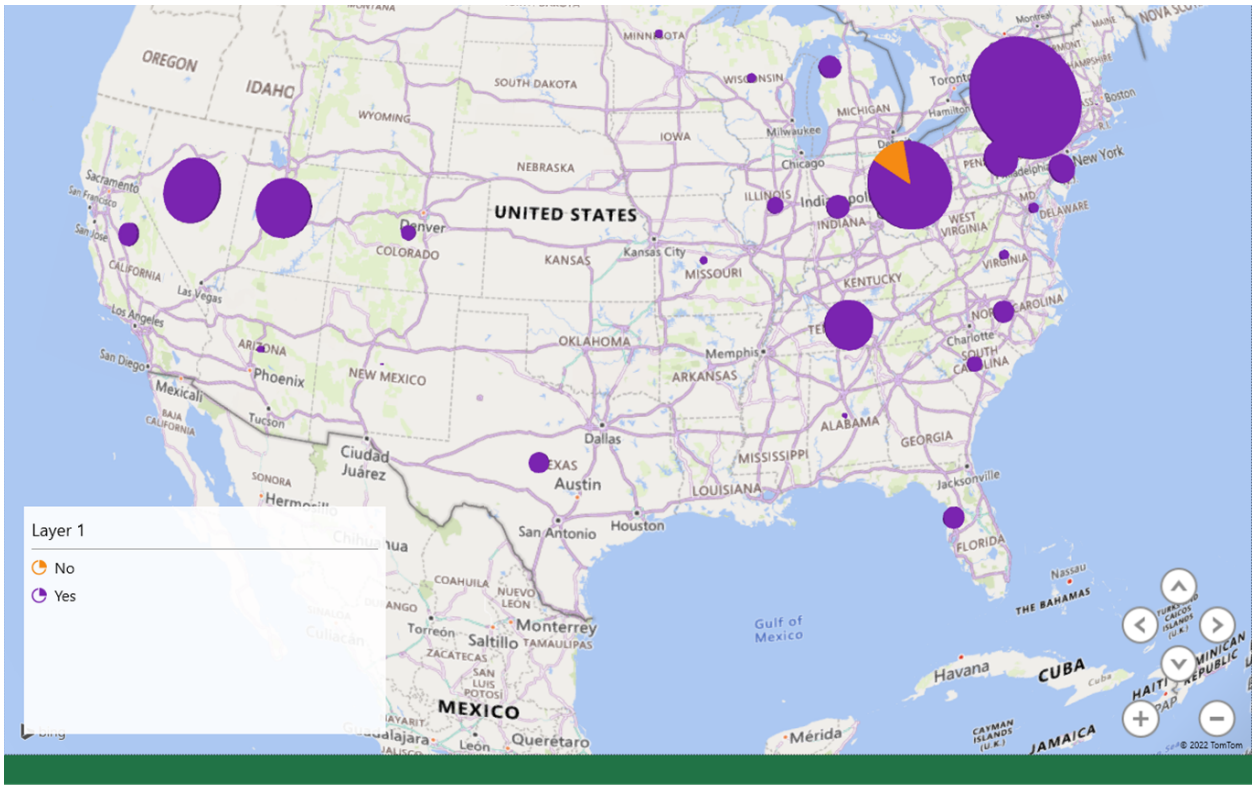


Add a filter to the underlying condition, and make sure to check the boxes there are only those who said yes or no.

Underlying conditions

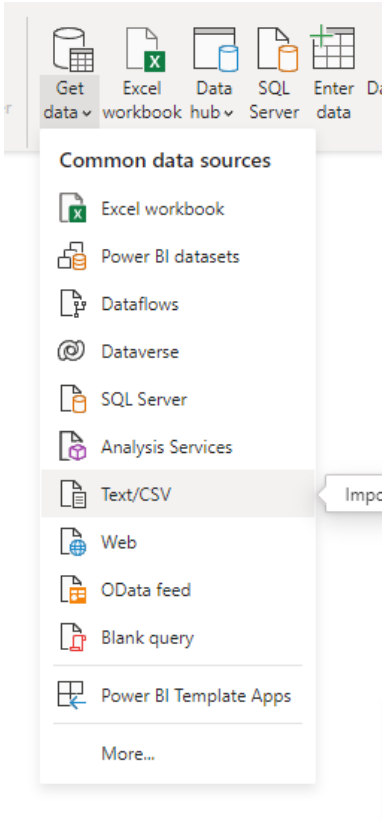
Filtered

- [ ] (All)
- [ ]
- [x] No
- [x] Yes

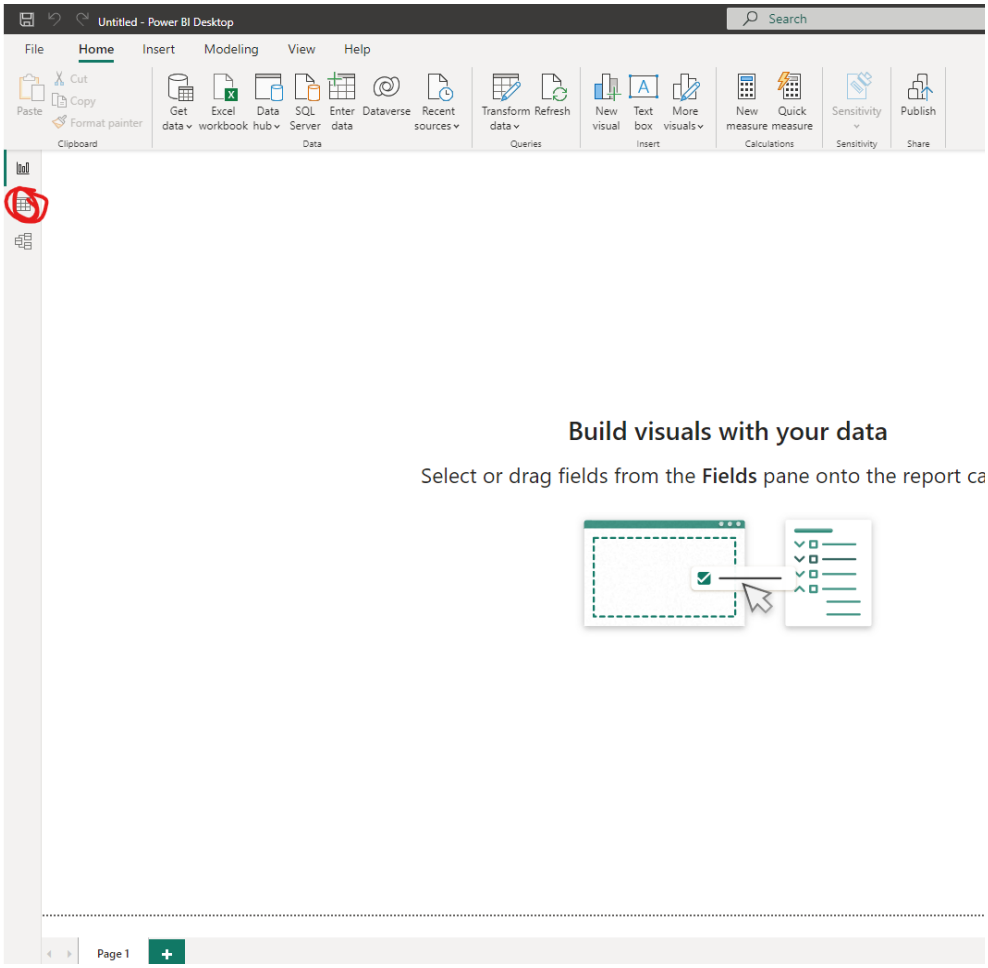


12. Loading Data into and Visualizing using PowerBI Desktop (You have to use PowerBI Desktop)

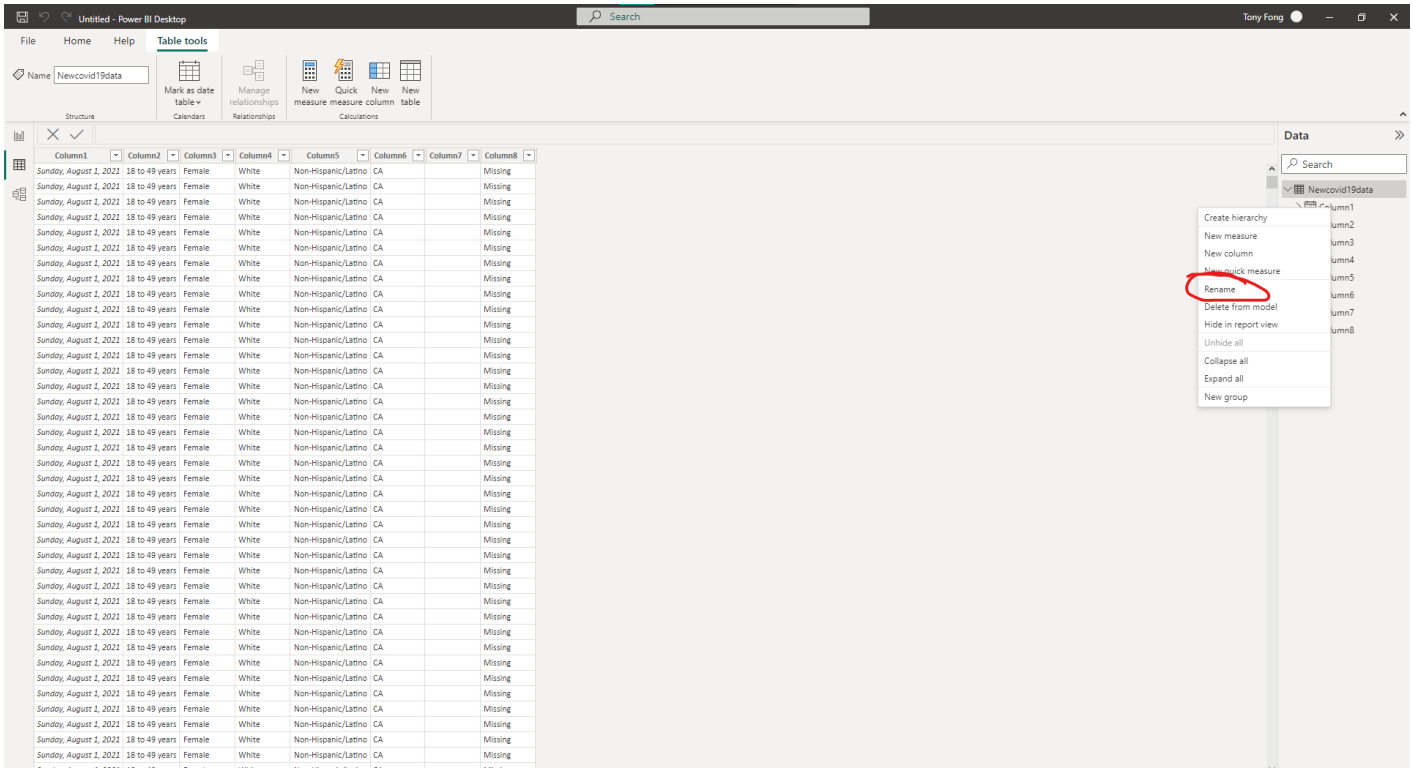
Open PowerBI desktop and load your data into PowerBI



You need to rename the columns by After loading your data onto PowerBI Desktop, go to your data as shown below:

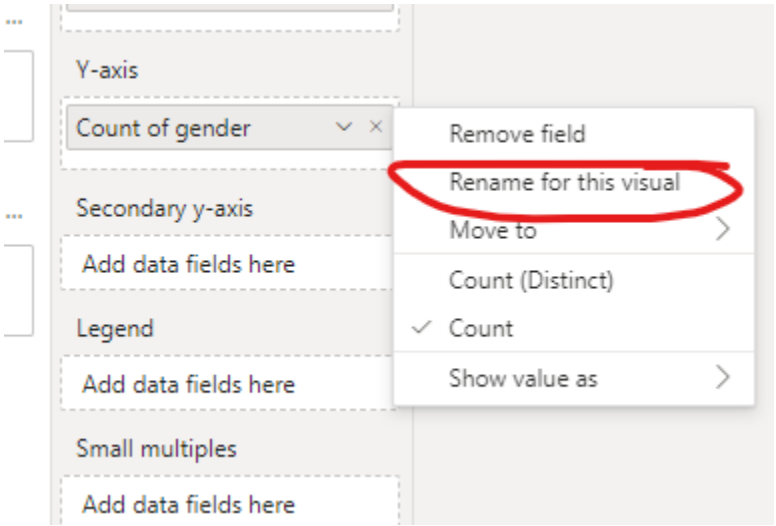


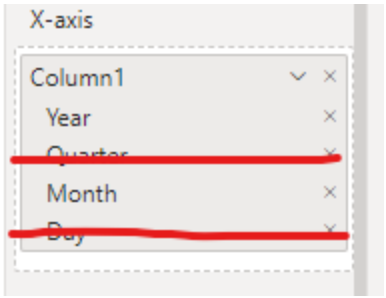
Rename each column to the corresponding names in this order: **(Date, Age, Gender, Race, Ethnicity, State, Underlying condition, Death)**



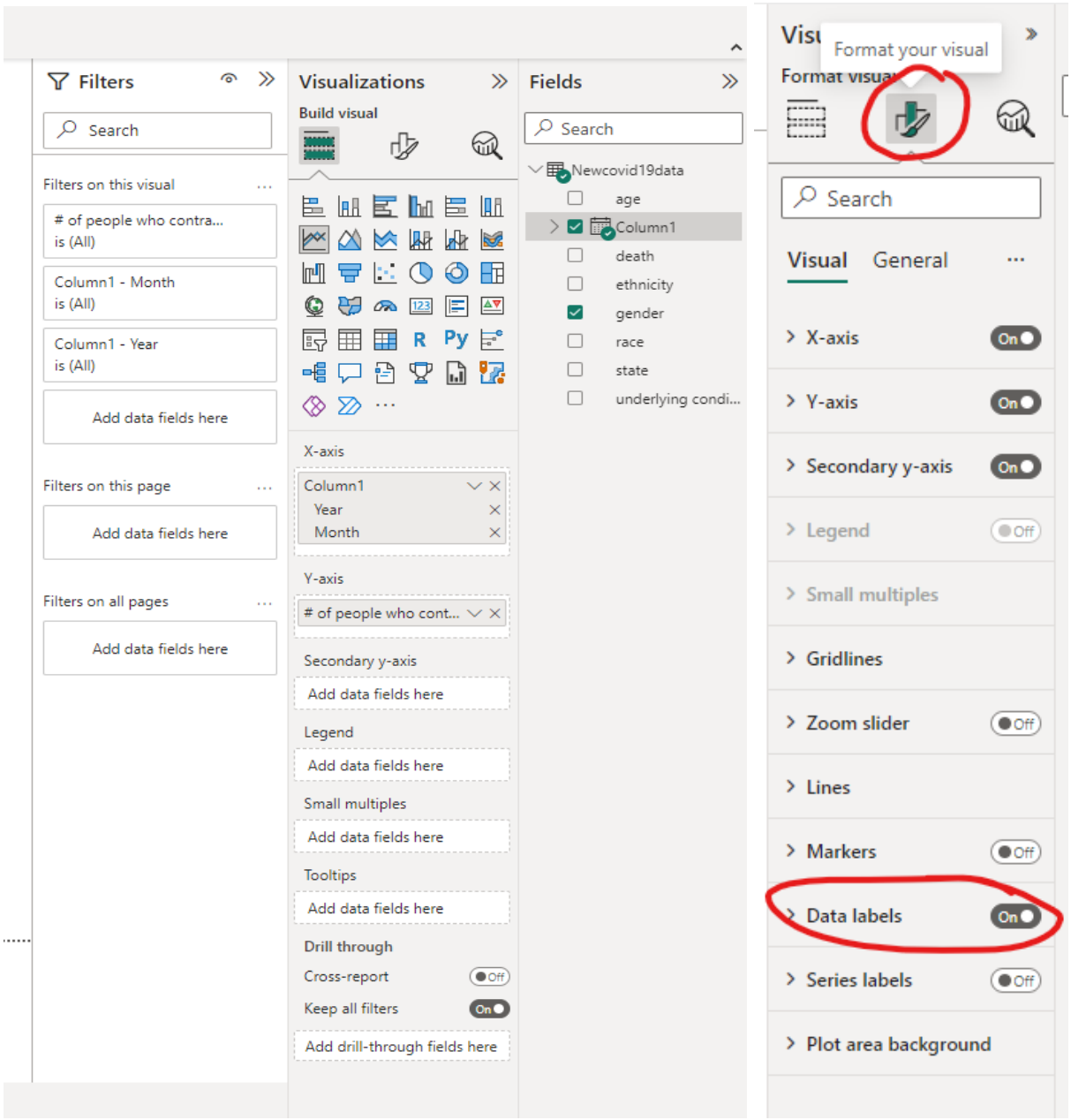
Note: Circled red is the Rename option

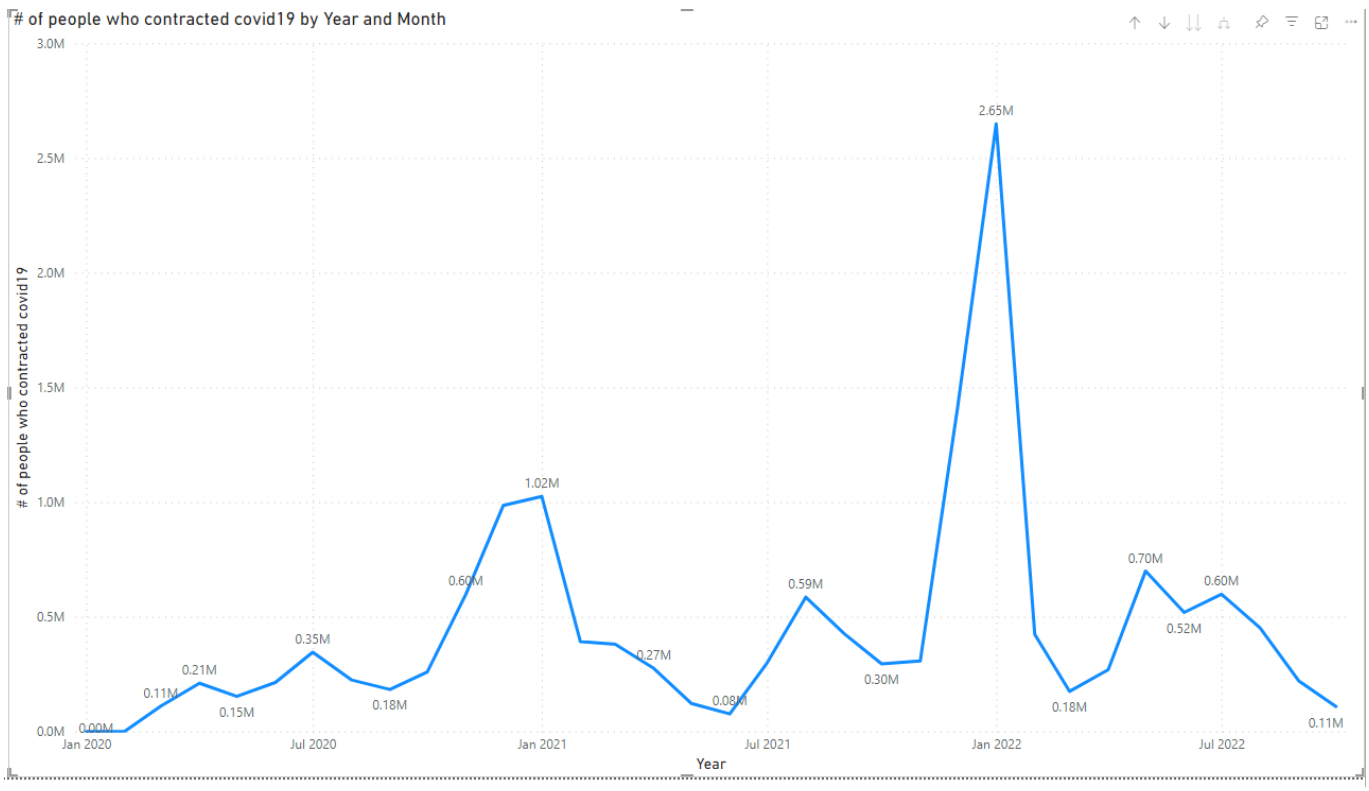
Next you need to select the properties and values in the layer as follows **to find the number of people who contracted covid by year and month**. First select your visualization types as a line chart. Next, format your visual, rename “count of gender” to “# of people who contracted covid19” and turn on the data label.



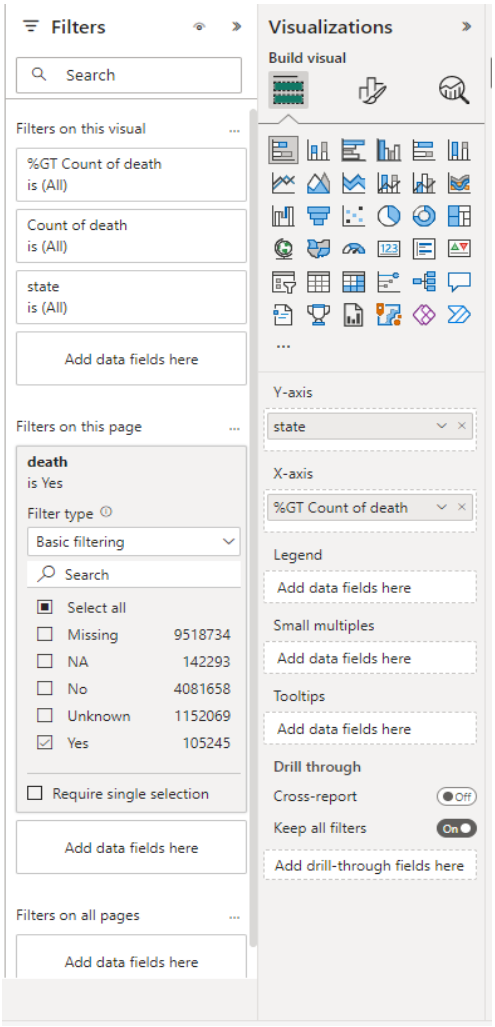


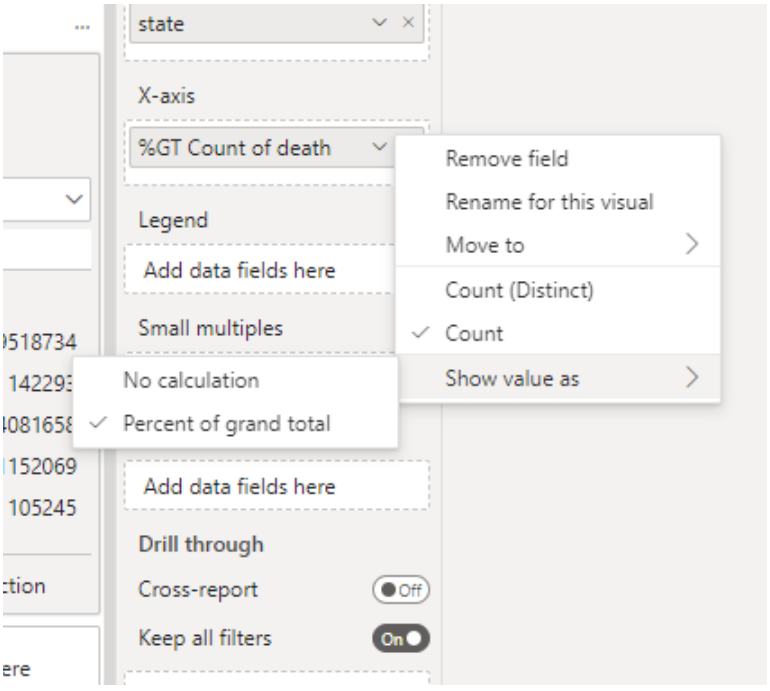
Note: Remove Quarter and Day from the date column

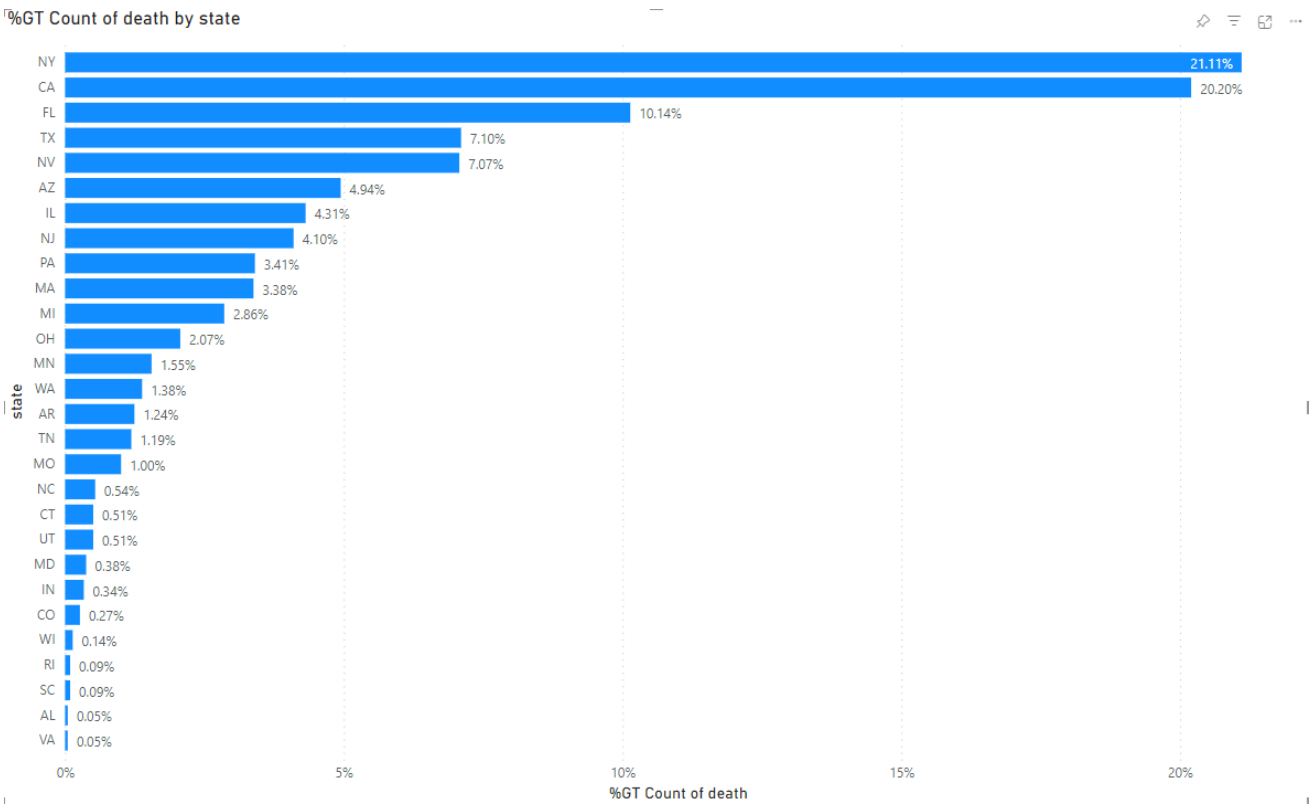




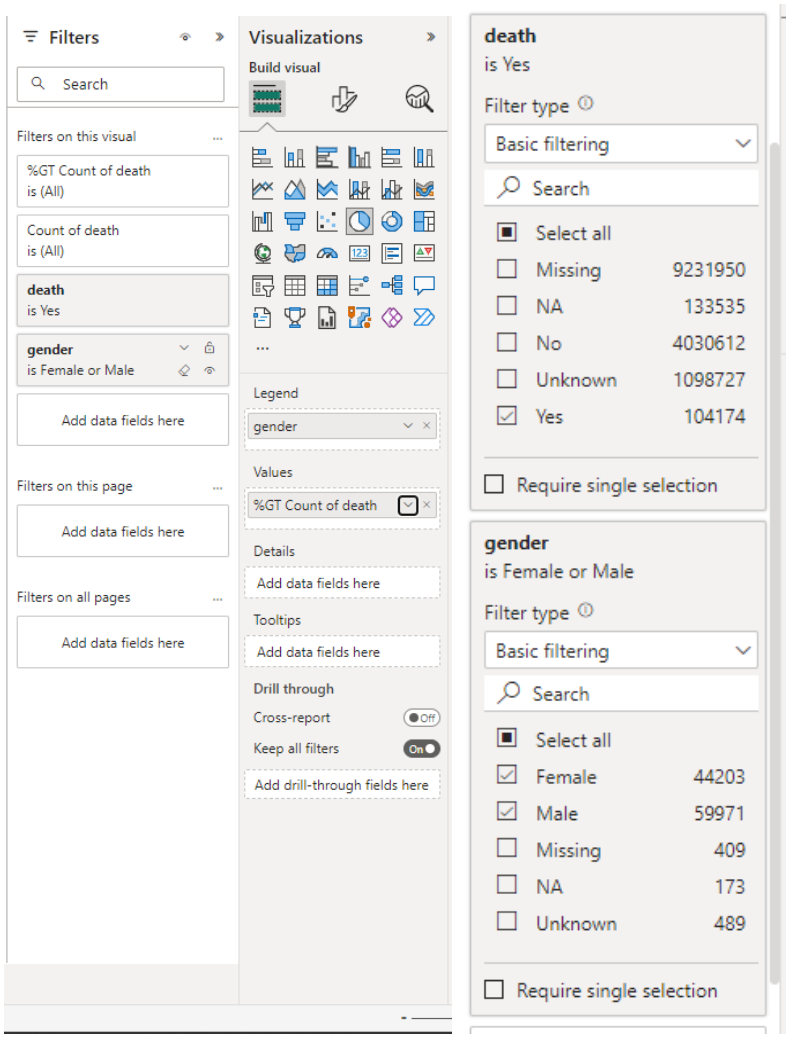
13. You need to select the properties and values in the layer as follows to find the count of death by state and filter death to only those who said yes. Then show value as percent of grand total. Also, make sure to change the visualization type to stacked bar graph.

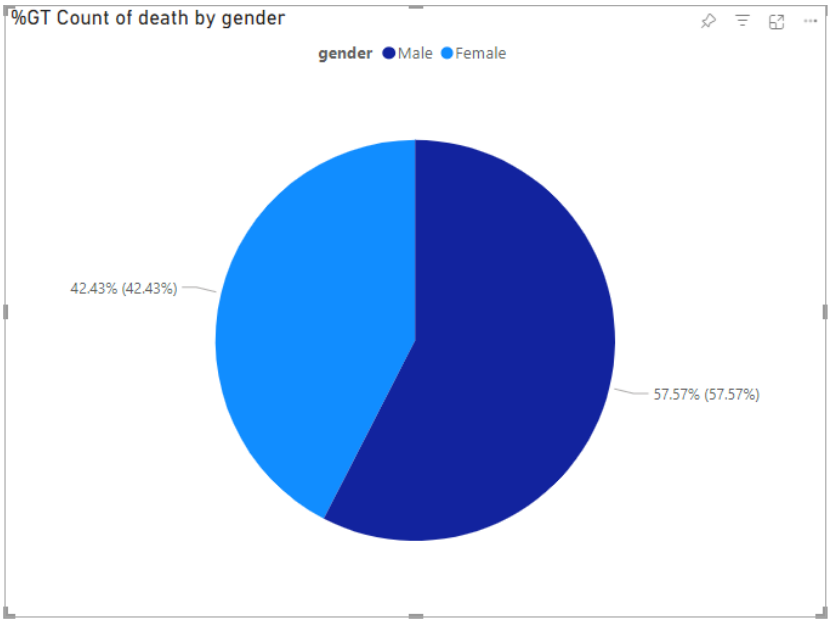

Filters
Search
Filters on this visual
%GT Count of death
is (All)
Count of death
is (All)
state
is (All)
Add data fields here
Filters on this page
death
is Yes
Filter type
Basic filtering
Search
Select all
Missing 9518734
NA 142293
No 4081658
Unknown 1152069
Yes 105245
Require single selection
Add data fields here
Filters on all pages
Add data fields here
Visualizations
Build visual
Y-axis
state
X-axis
%GT Count of death
Legend
Add data fields here
Small multiples
Add data fields here
Tooltips
Add data fields here
Drill through
Cross-report
Off
Keep all filters
On
Add drill-through fields here


state
X-axis
%GT Count of death
Legend
Add data fields here
Small multiples
Add data fields here
Drill through
Cross-report
Off
Keep all filters
On
Remove field
Rename for this visual
Move to
Count (Distinct)
Count
Show value as
No calculation
Percent of grand total



14. You need to select the properties and values in the layer as follows to find *the count of death by gender*. Then switch the visual type to pie chart and put the gender as the legend and values as death. Then add the following filters as shown in the picture below:


Filters
Search
Filters on this visual
%GT Count of death
is (All)
Count of death
is (All)
death
is Yes
gender
is Female or Male
Add data fields here
Filters on this page
Add data fields here
Filters on all pages
Add data fields here
Visualizations
Build visual
Legend
gender
Values
%GT Count of death
Details
Add data fields here
Tooltips
Add data fields here
Drill through
Cross-report
Off
Keep all filters
On
Add drill-through fields here
death
is Yes
Filter type
Basic filtering
Search
Select all
Missing 9231950
NA 133535
No 4030612
Unknown 1098727
Yes 104174
Require single selection
gender
is Female or Male
Filter type
Basic filtering
Search
Select all
Female 44203
Male 59971
Missing 409
NA 173
Unknown 489
Require single selection


%GT Count of death by gender
gender Male Female
42.43% (42.43%)
57.57% (57.57%)

References:

https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4

‘

https://github.com/mike0nthemic/G5_Big_Data_4560.git