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:

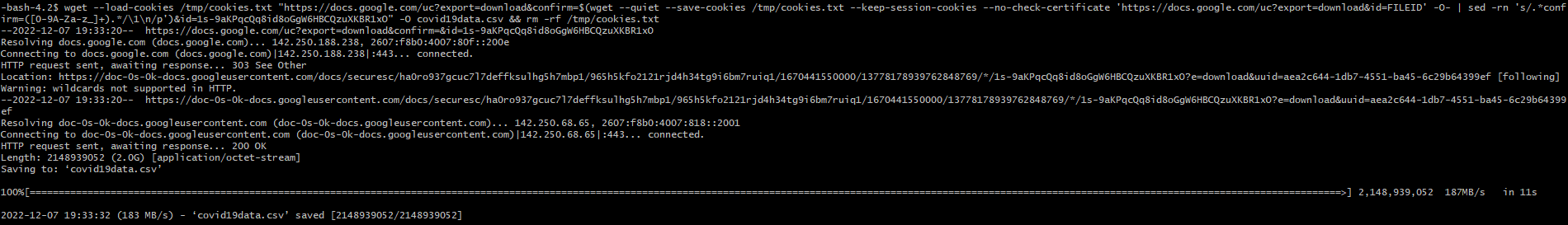
* Oracle Cloud
* CPU Speed: 1995. 309 MHz:
* # of CPU Cores: 32
* # of nodes: 3
* Total Memory Size: 58GB

**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]**

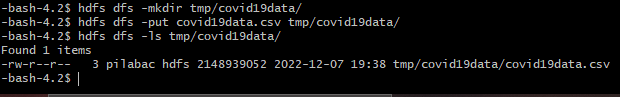
**2. Download the covid19 dataset file using wget**

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



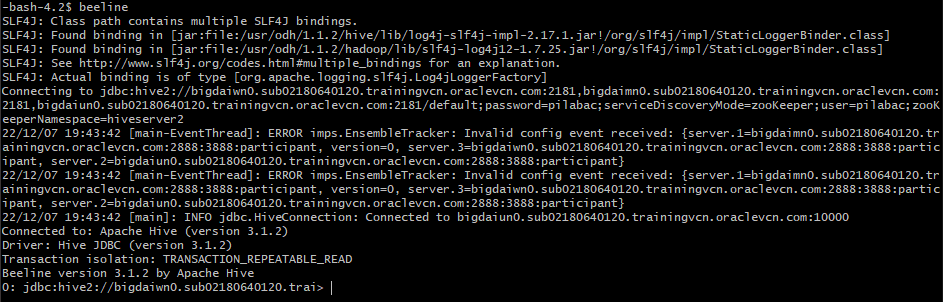
**3.You have to upload the files to hdfs folder coviddata. Run the following HDFS commands to create and list coviddata directory in HDFS:**

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



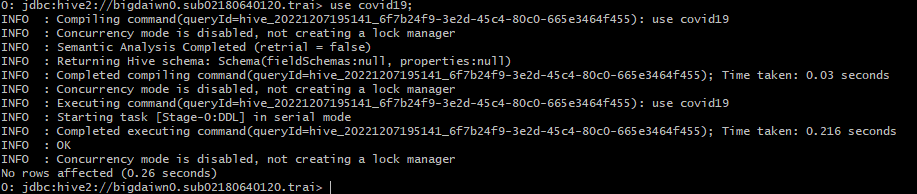
**4. Open hive**

| $ beeline |
| --- |



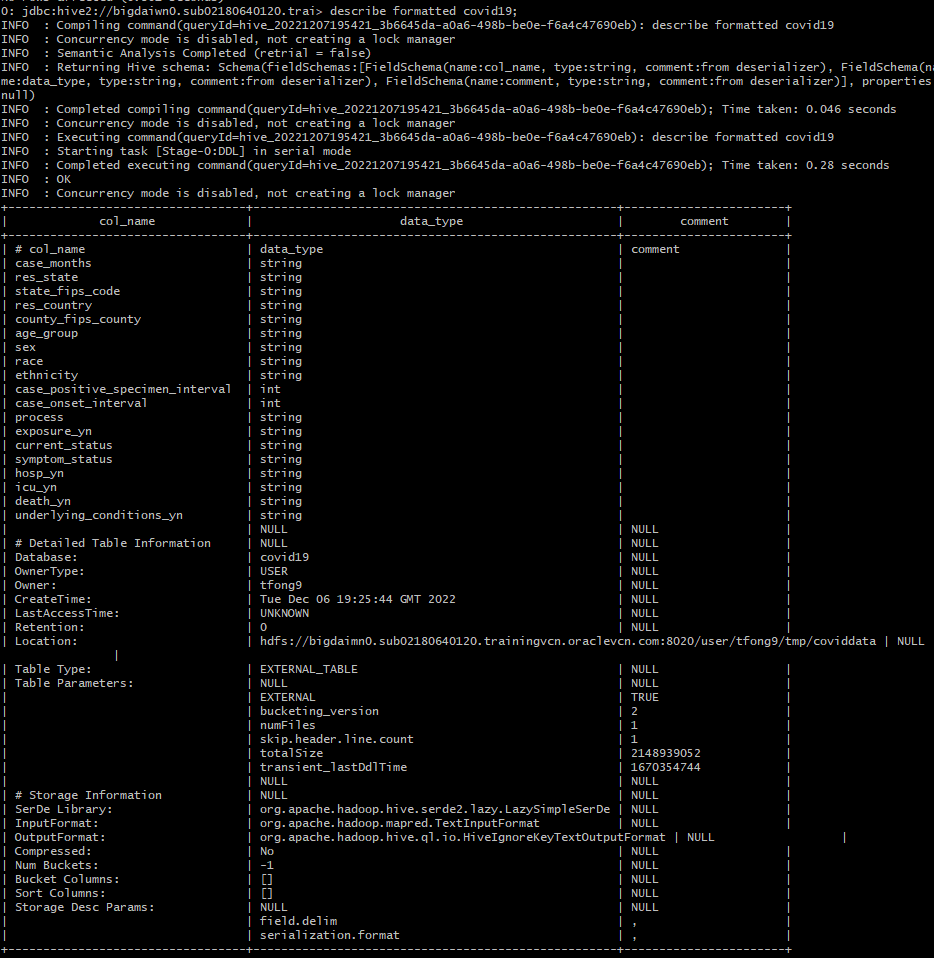
**5. Create your own database and use that database**

| $ create database Covid19;  $ use Covid19; |
| --- |



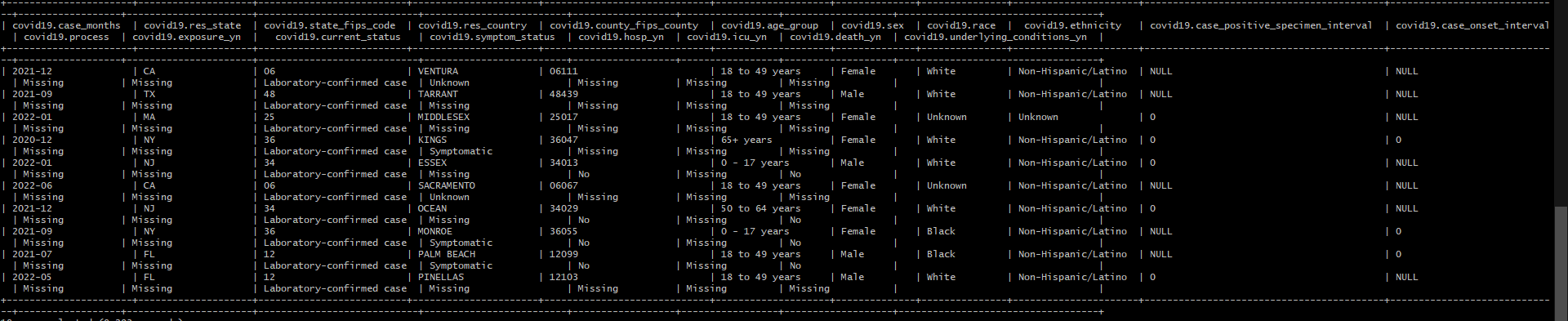
**6. Create external table “Covid19Data”**

| - - 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'); |
| --- |



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

| select \* from covid19 limit 10; |
| --- |

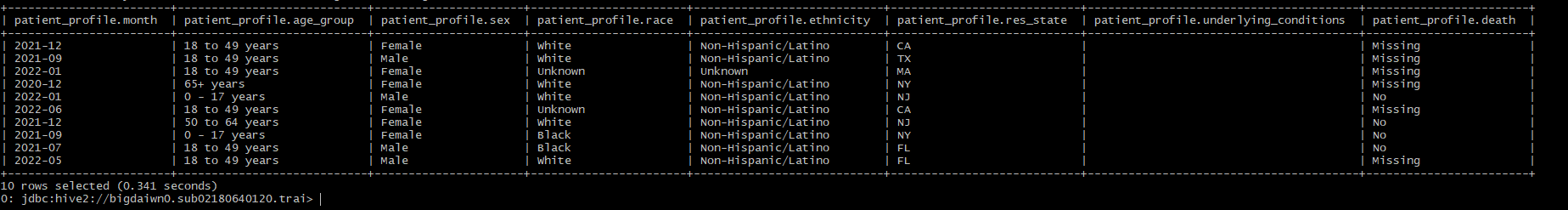


**7. Create external table “patient\_profile”**

| - - 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; |
| --- |

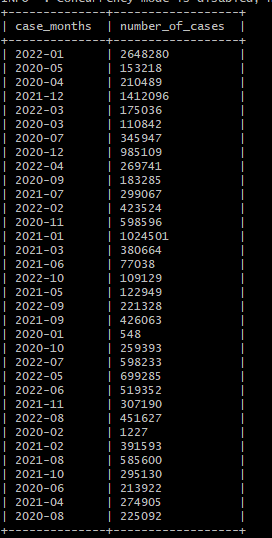
Now run the following HiveQL at the query editor to see how the dataset looks like

| Select \* from patient\_profile limit 10; |
| --- |



**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; |
| --- |

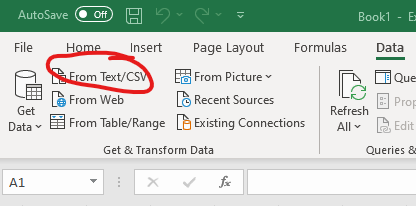


**9. Now download data into your PC**

| - - 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

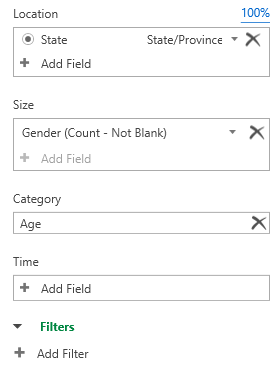


After renaming the columns, highlight all your columns and select the 3D maps under the “insert” tab



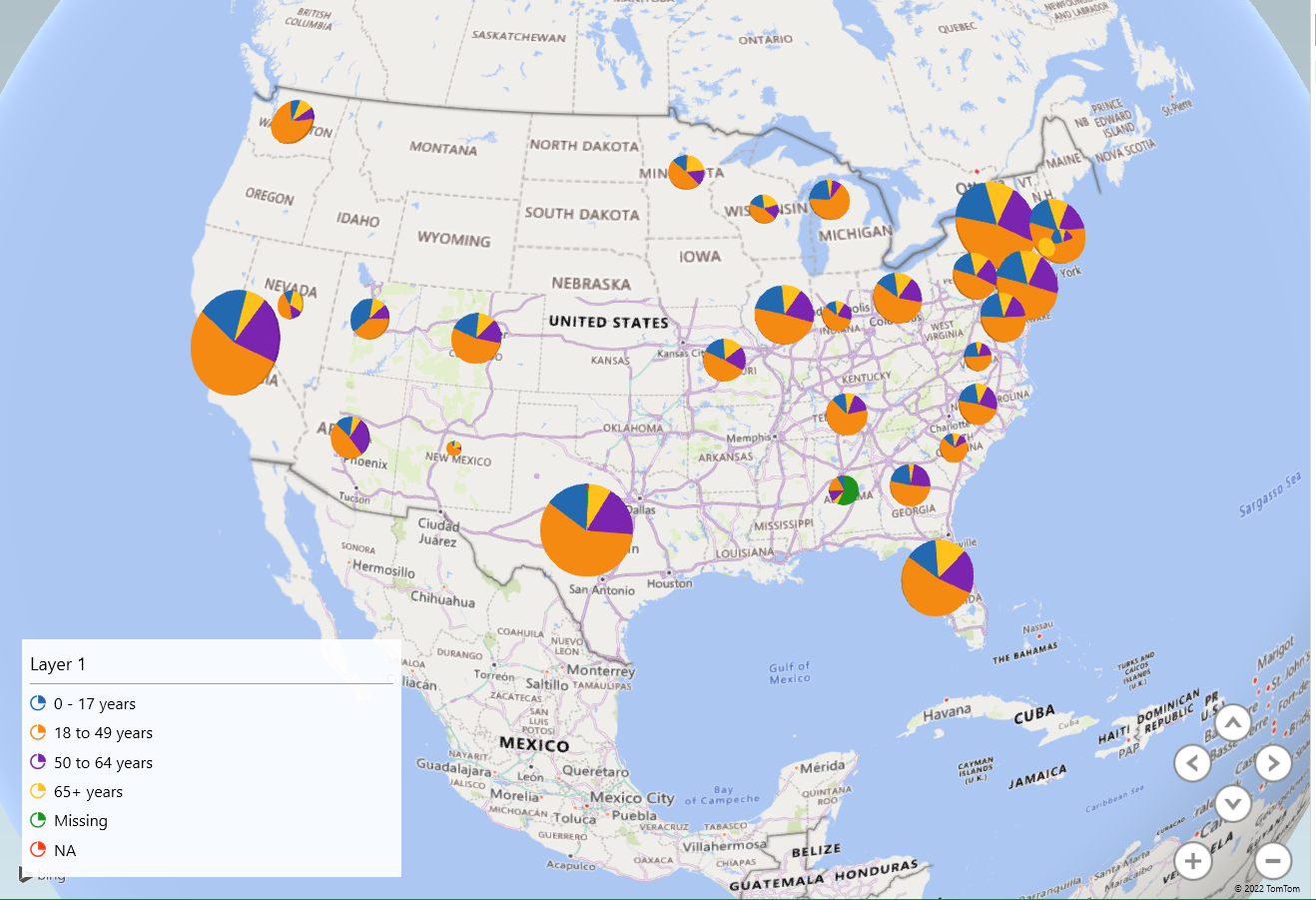
**11. You need to select the properties and values in the layer as follows:**

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

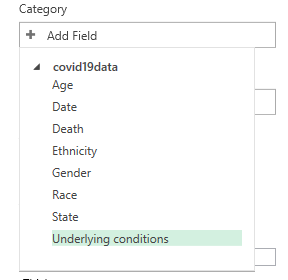


After you should change the graph to pie graphs

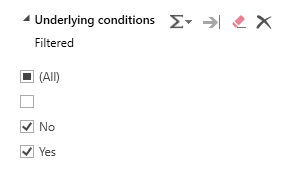


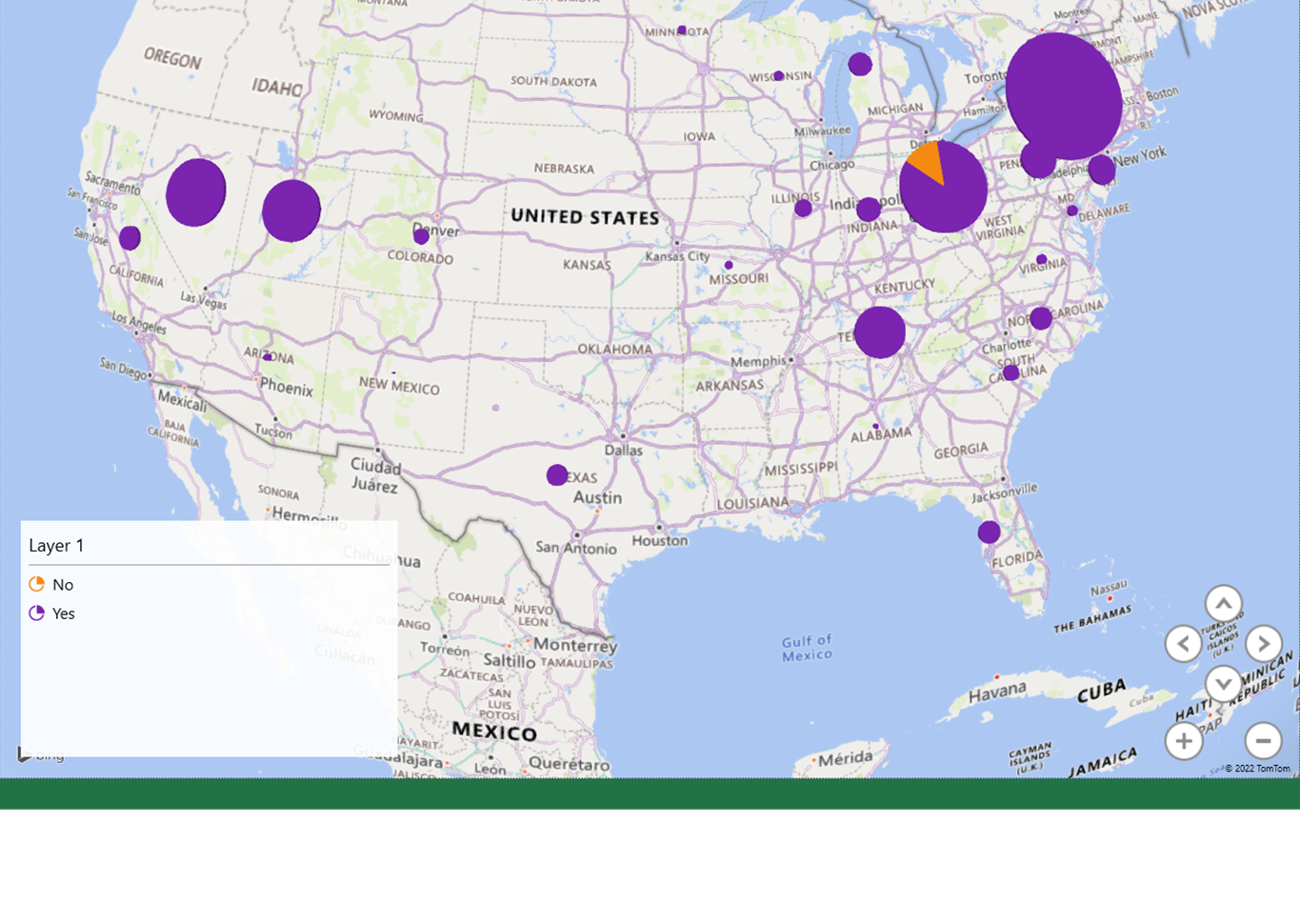


Afterwards, you should change the Category from ***Age Groups*** to ***Underlying Conditions***



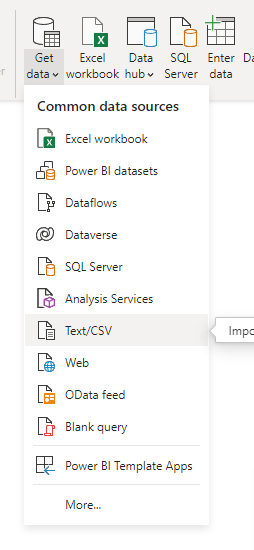
Add a filter to the underlying condition, and make sure to check the boxes there are only those who said yes or no.



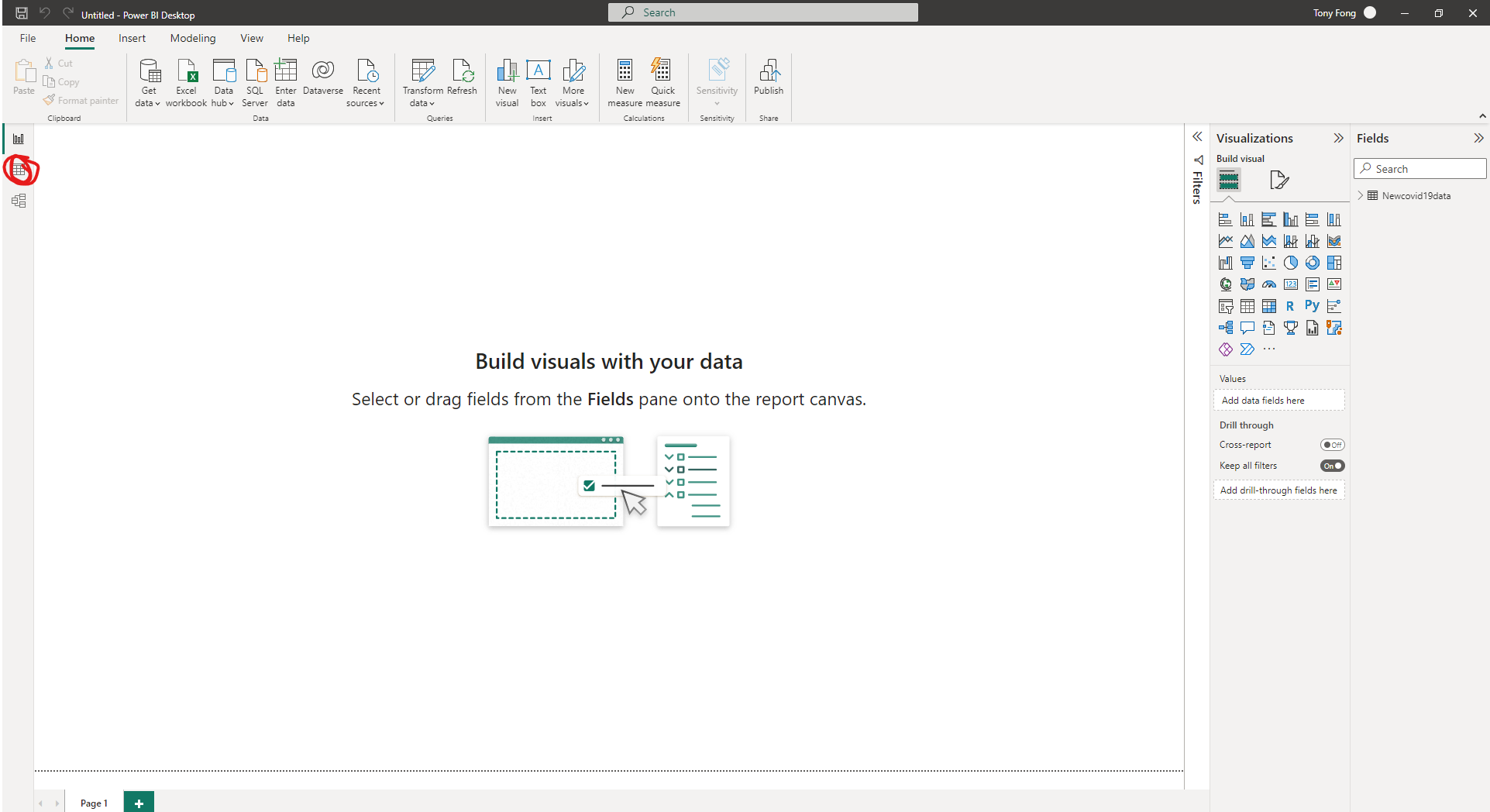


**12. Loading Data into and Visualizing using PowerBI Desktop (You have to use PowerBI Desktop)**

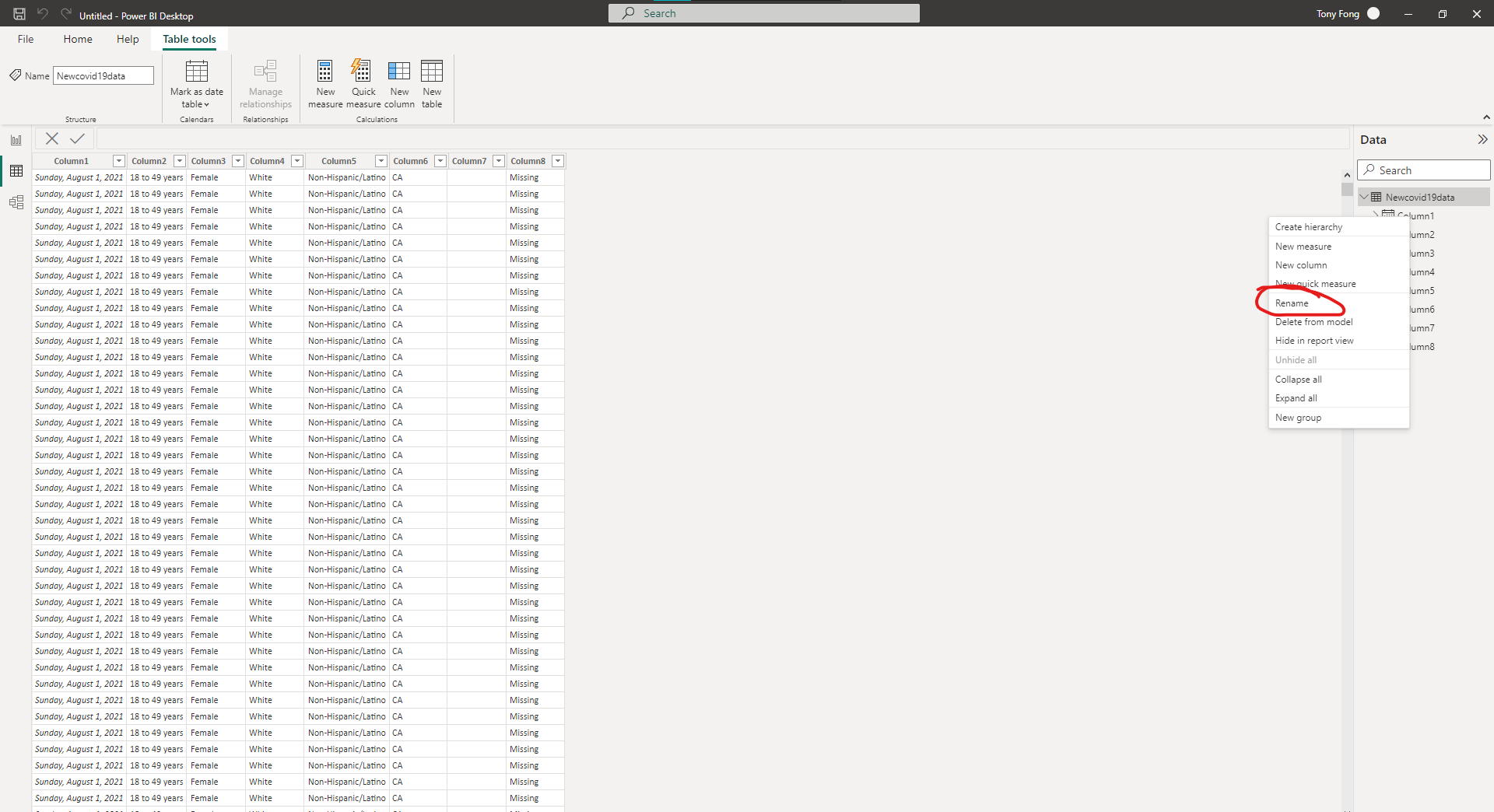
Open PowerBI desktop and load your data into PowerBI



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

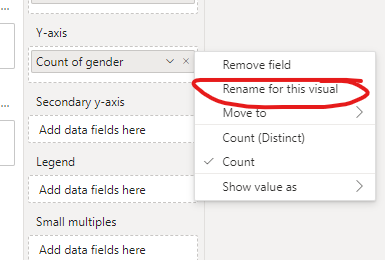


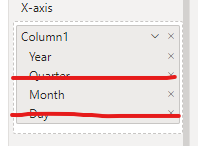
Rename each column to the corresponding names in this order: **(Date, Age, Gender, Race, Ethnicity, State, Underlying condition, Death)**



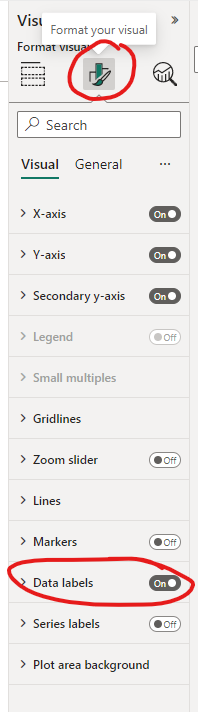
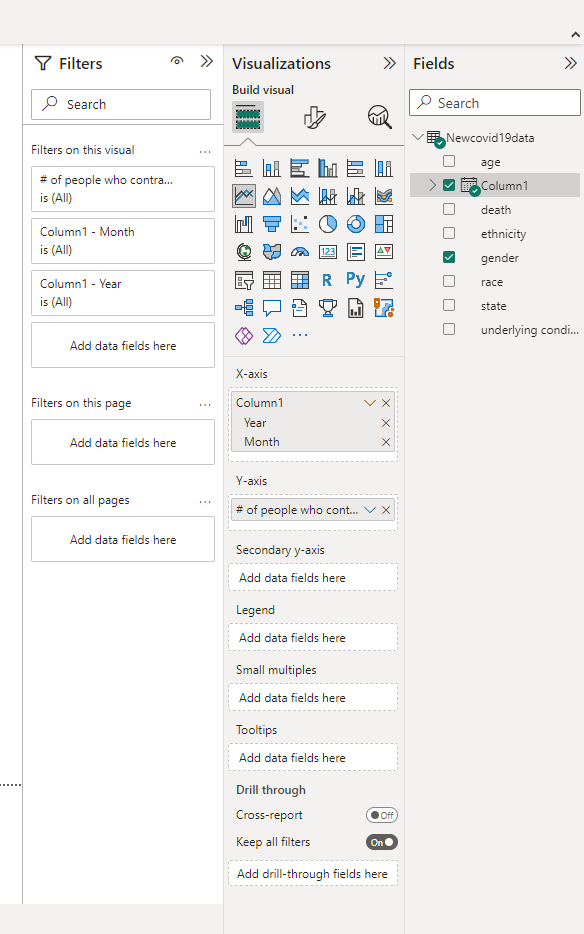
***Note: Circled red is the Rename option***

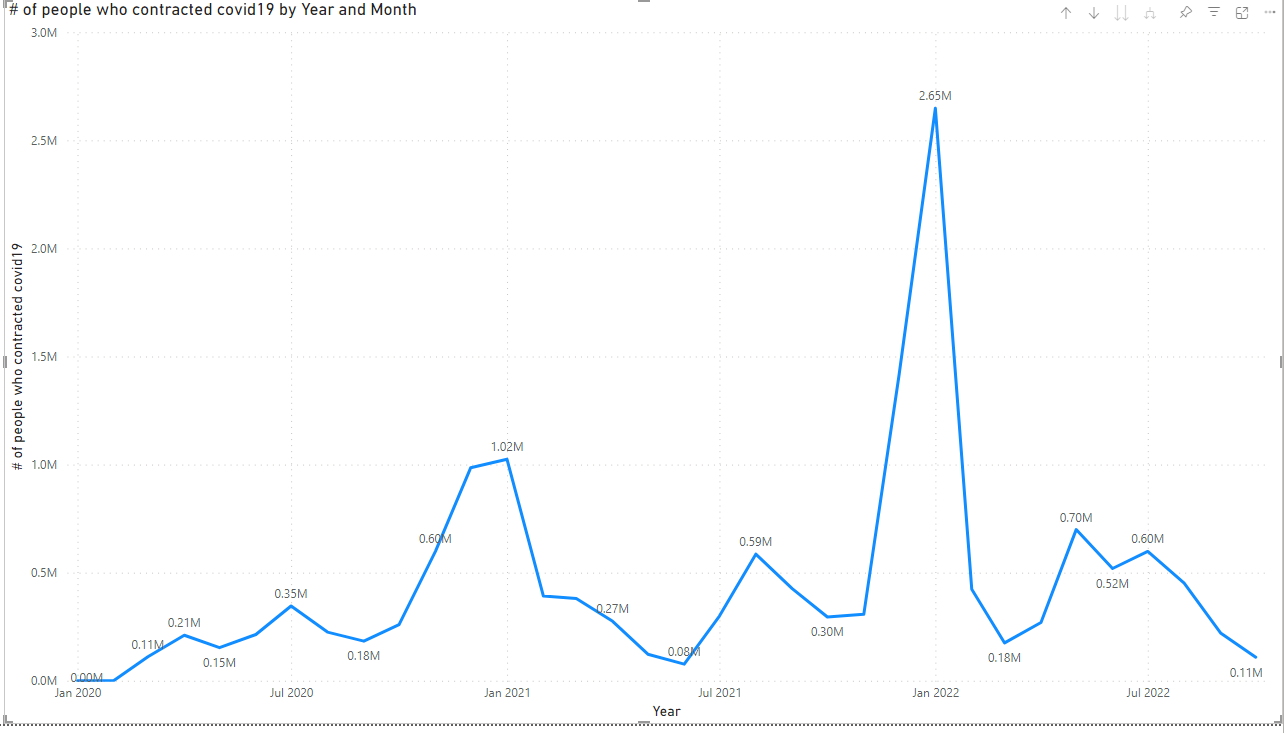
Next you need to select the properties and values in the layer as follows **to find the number of people who contracted covid by year and month**. First select your visualization types as a line chart. Next, format your visual, rename “count of gender” to “# of people who contracted covid19” and turn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



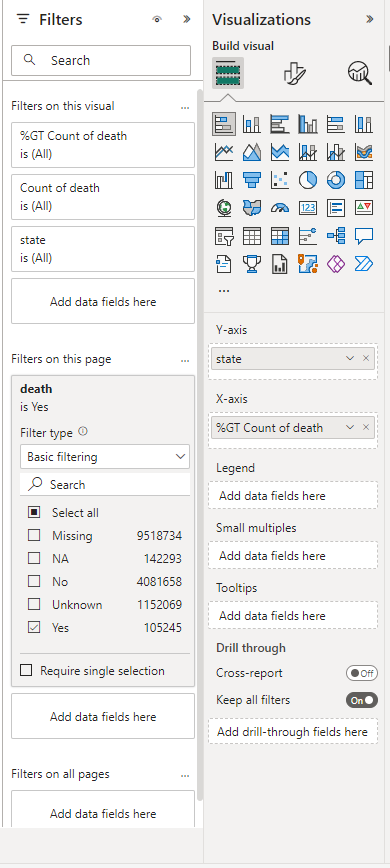


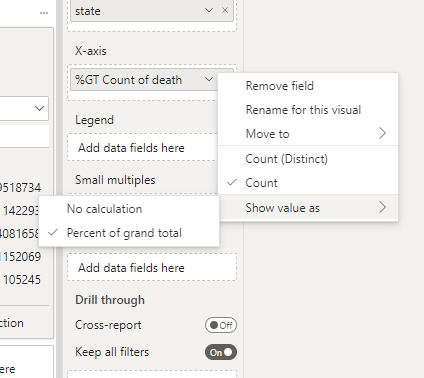
***Note: Remove Quarter and Day from the date column***

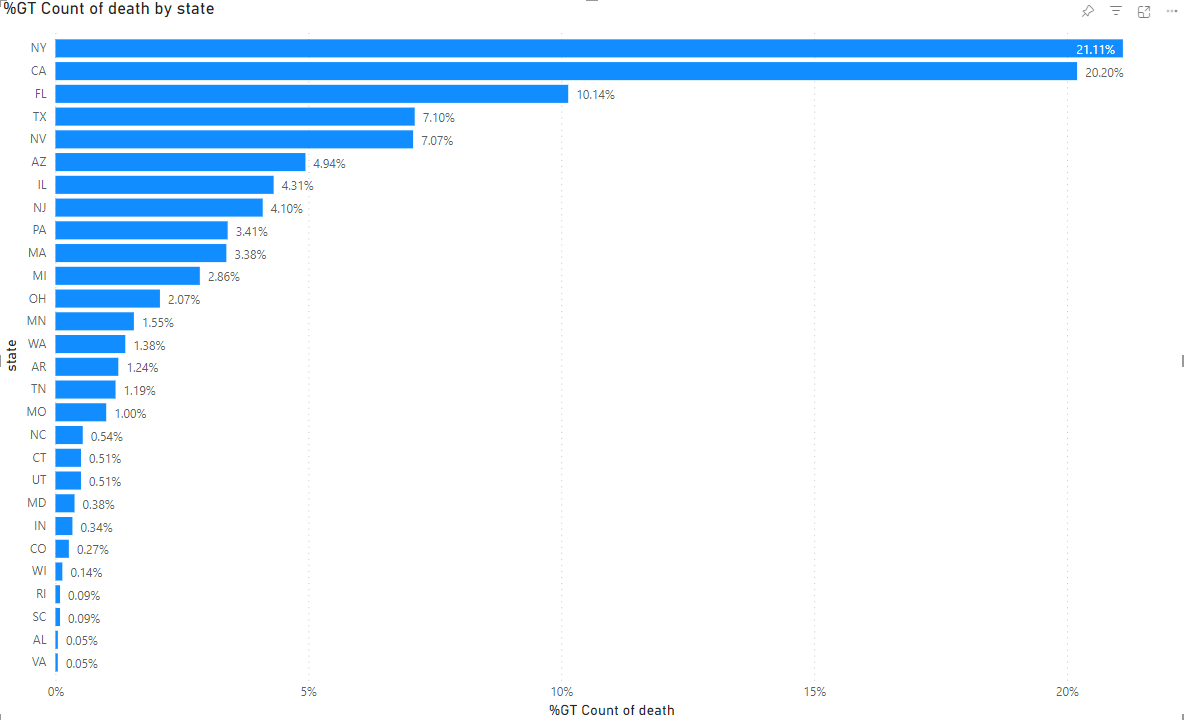




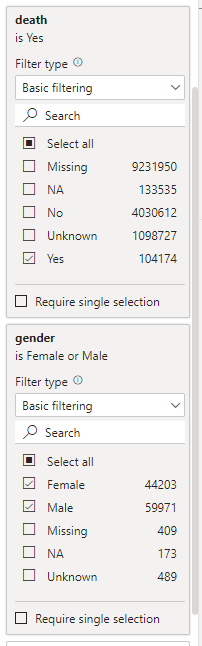
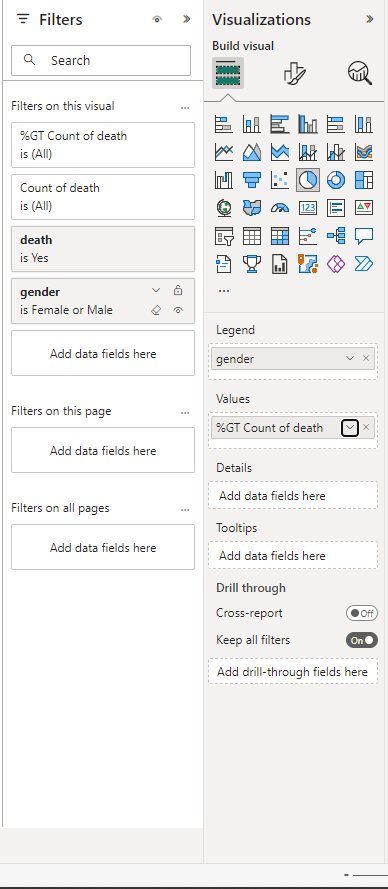
**13. You need to select the properties and values in the layer as follows to find the count of death by state and filter death to only those who said yes. Then show value as percent of grand total. Also, make sure to change the visualization type to stacked bar graph.**

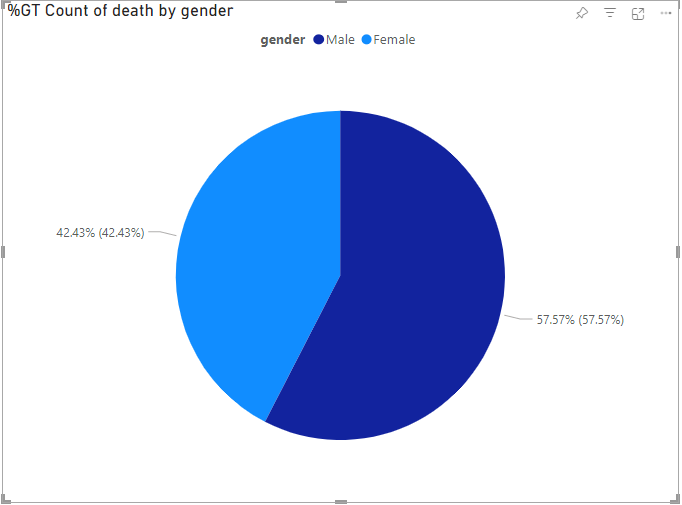






**14. You need to select the properties and values in the layer as follows to find *the count of death by gender*. Then switch the visual type to pie chart and put the gender as the legend and values as death. Then add the following filters as shown in the picture below:**





**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>