### **BIA WORKSHOP – POWER BI USING DAX LAB 2 - Data exploration, data visualization, and DAX**

### Prepared exclusively for 2023 BIA Graduating Seniors (with an INFO 3300 prereq)

May 2023

## **LAB RESOURCES**

|  |  |
| --- | --- |
|  | DIAD LAB 1.pbix |
|  | <https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_US.csv> |
|  | <https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_deaths_US.csv> |
|  | Custom Visualizations Folder |

**LAB PURPOSE AND LEARNING OBJECTIVES:** For this lab, we focus on **importing more data**, **applying more transformations, updating our model**, and using **more DAX** to design several **reports** (dashboards) **using a multidimensional question and data**. We will also add filtering, sorting, and custom visuals.

## **LESSON 1: DAX, SORTING AND FILTERING**

1. **As a reminder from Lab 1**, the DAX formulas we created to calculate confirmed global cases:

Confirmed Global = CALCULATE(SUM(COV2020\_CSSE[Value]), (COV2020\_CSSE[Item]="Confirmed"))

Confirmed Recent Value Global = CALCULATE(SUM(COV2020\_CSSE[Value]), (COV2020\_CSSE[Item]="Confirmed"), FILTER(COV2020\_CSSE, COV2020\_CSSE[Date] = MAX(COV2020\_CSSE[Date])))

1. Create several more **DAX measures** (to the **COV2020\_CSSE** dataset) to calculate the total deaths:

Deaths Global = CALCULATE(SUM(COV2020\_CSSE[Value]), (COV2020\_CSSE[Item]="Deaths"))

Deaths Recent Value Global = CALCULATE(SUM(COV2020\_CSSE[Value]), (COV2020\_CSSE[Item]="Deaths"), FILTER(COV2020\_CSSE, COV2020\_CSSE[Date]= MAX([Date])))

1. Create one **DAX column** (to the **World Population** dataset) to allocate density values into categories[[1]](#footnote-1):

Density Breakdown =

IF([Density (P/Km²)] >= 500, "High Density",

    IF([Density (P/Km²)] > 200, "Medium Density",

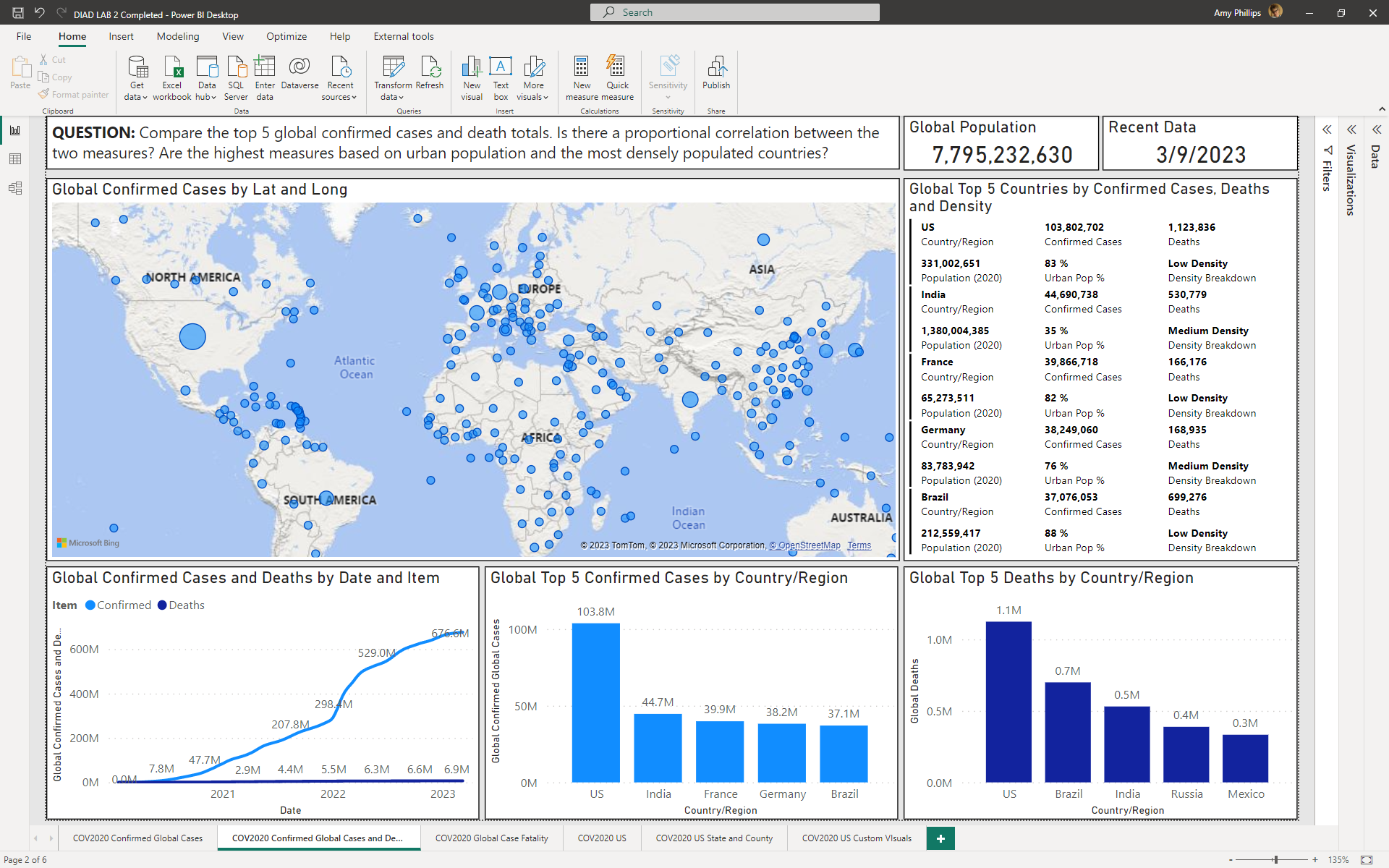
        IF([Density (P/Km²)] >= 0, "Low Density",

"N/A")))

**NOTE:** Worldometers (the source for the World Population dataset that JHU uses) already calculated density for us, but we could have added a DAX column if they didn’t, such as:

Density = [Population (2020)]/[Land Area (Km²)]

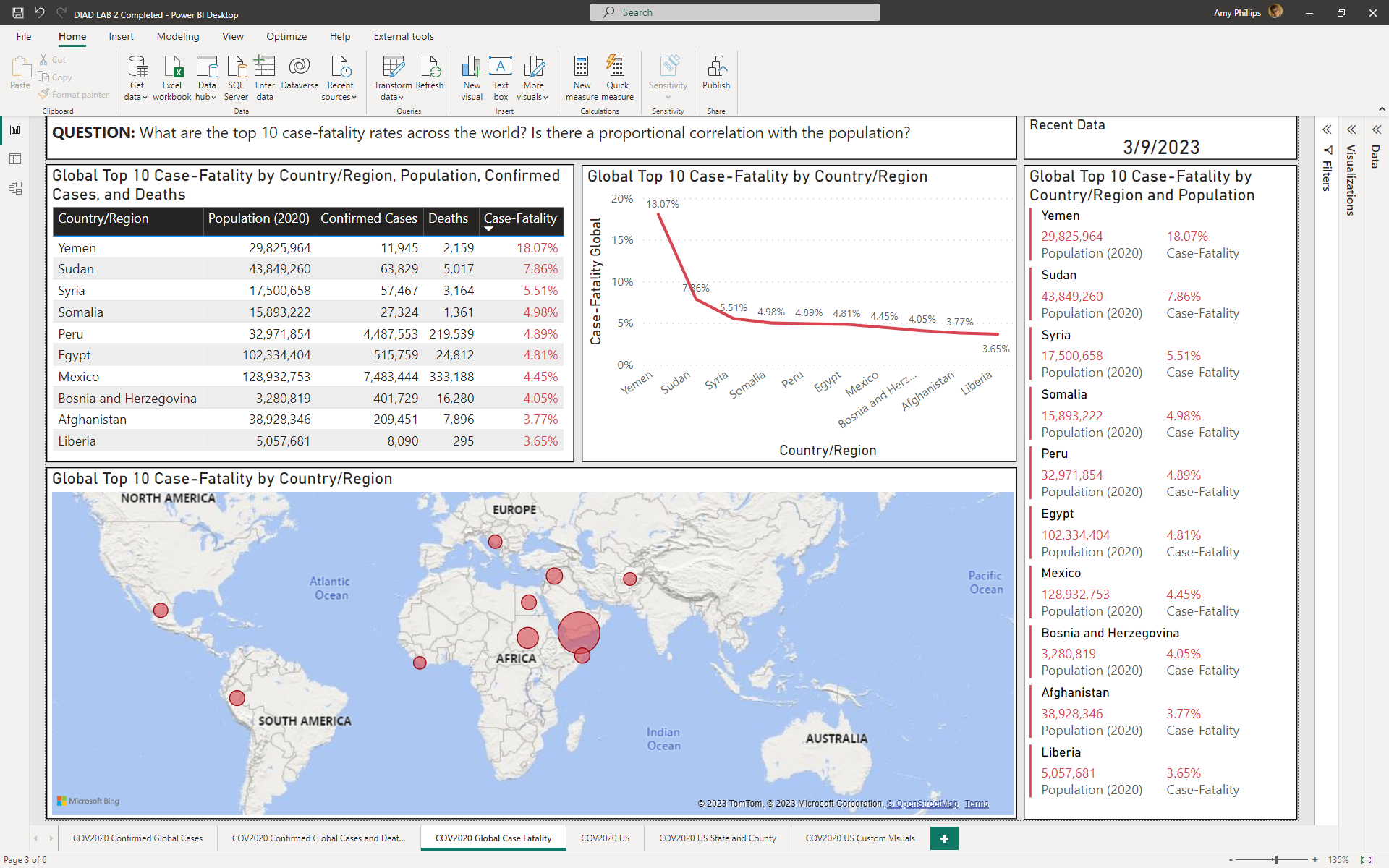
1. Create a **new report/page** (e.g., dashboard), we will **try to replicate:**



1. **Start with a question** (include as a textbox at the top of the dashboard)
   1. **QUESTION: Compare the top 5 global confirmed cases and death totals. Is there a proportional correlation between the two measures? Are the highest measures based on urban population and the most densely populated countries?**
2. **NOTE: Sort the visuals**, high-low, using the various measures
3. **Rename** the report/page **COV2020 Confirmed Global Cases and Deaths**
4. **Save** your Power BI file as **DIAD LAB 2**

## **LESSON 2: A LITTLE MORE DAX, SORTING, FILTERING**

1. Create a **new report/page** (e.g., dashboard), we will **try to replicate**:[[2]](#footnote-2)



1. **Start with a question** (include as a textbox at the top of the dashboard)
   1. **QUESTION: What are the top 10** case-fatality rates across the world. Is there a correlation with the population? [[3]](#footnote-3)
2. To answer that question, create a **DAX measure** in the **COV2020\_CSSE** table that will divide the sum of deaths by confirmed cases:

Case-Fatality Global = DIVIDE([Deaths Recent Value Global],[Confirmed Recent Value Global])

1. **NOTE 1:** Use **Filters on this page, Country/Region** to unselect **MS Zaandam,** this is a cruise ship, not a country[[4]](#footnote-4). And unselect **Korea, North**, as they have only reported confirmed cases and deaths for one day since JHU has been recording data.
2. **NOTE 2:** **Sort the visuals**, high-low (mainly by Case-Fatality)
3. **NOTE 3:** You can easily argue that **many of the visuals appear to be repetitive**… it’s mainly the same data, yet **visualized differently** (potentially offering alternative ways to analyze the data)
4. **Rename** the report/page **COV2020 Global Case Fatalities**
5. **Save** the Power BI file

## **LESSON 3: ETL US COVID DATA**

1. We will import **US data**. Using many of the steps from **Lab 1, Lesson 1, import data** for just the **US** from (there are **two** datasets):

<https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_US.csv>

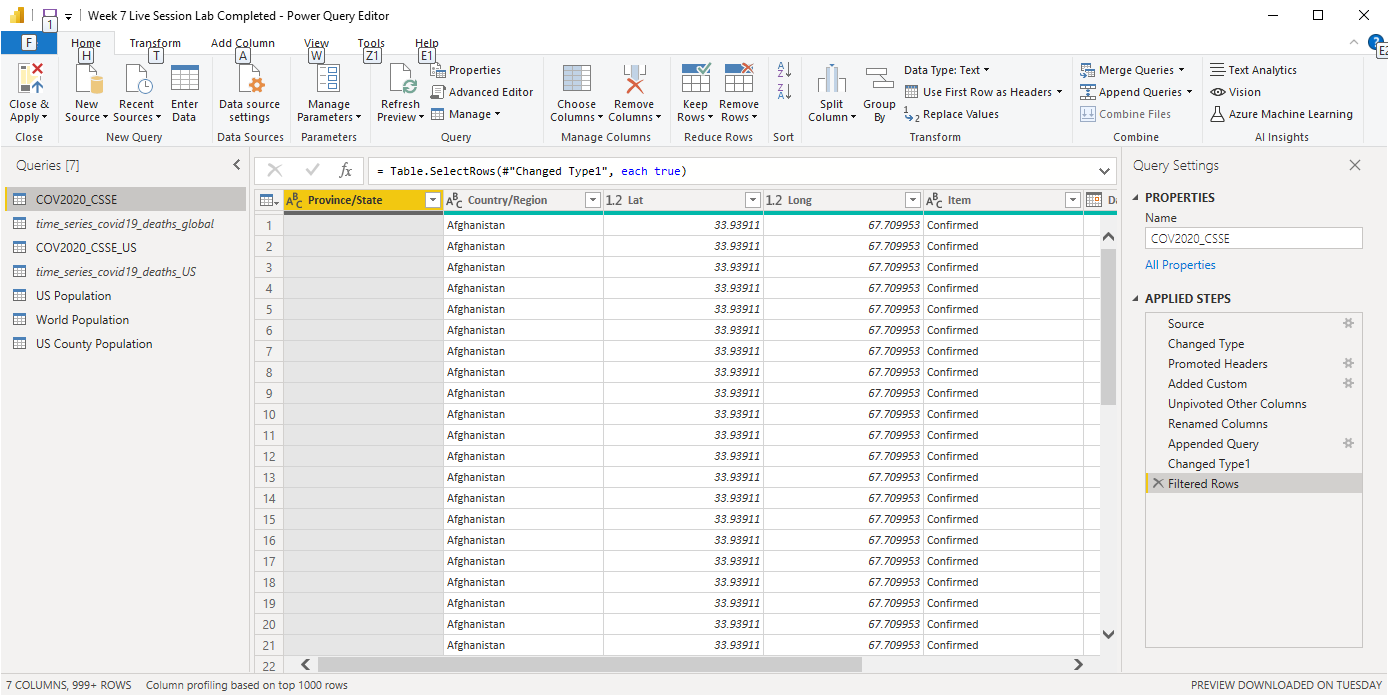
<https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_deaths_US.csv>

* 1. Rename one query **COV2020\_CSSE\_US** (e.g., time\_series\_covid19\_confirmed\_US)
  2. Remember to **use first row as headers, add a custom column for “Confirmed”** and **“Deaths”, unpivot other columns (all dates), rename columns** to have proper headings, **change data types** if needed
  3. **Append the two new US queries,** start with selecting **COV2020\_CSSE\_US** (only append the US queries with each other, not to the **COV2020\_CSSE** query)
  4. **Remove** the **Population** column,as the data is incomplete, and we will add a separate table with population data that is more comprehensive
  5. **Unselect Enable load** for the **time\_series\_covid19\_deaths\_US** as this is not needed in the Report view

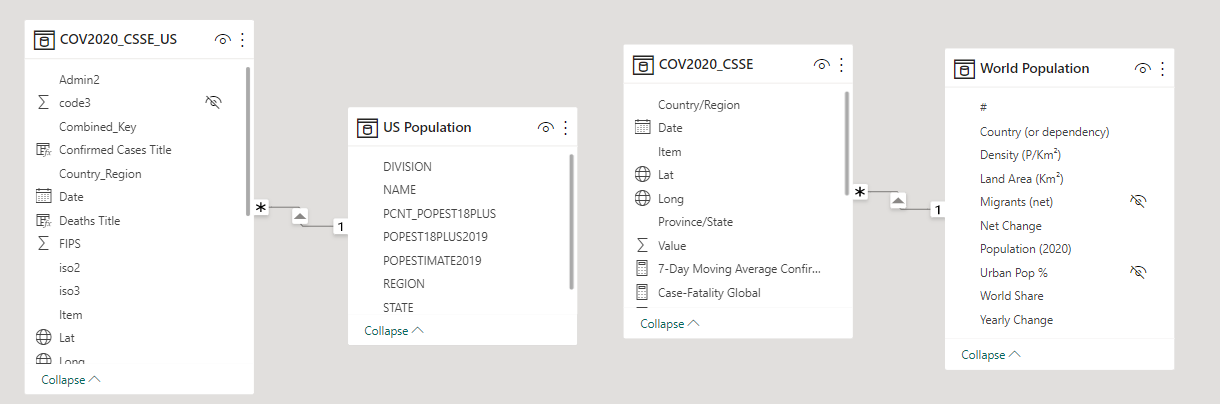
1. Select **New Source** (Home tab), **Web**, insert the URL

<https://www2.census.gov/programs-surveys/popest/datasets/2010-2019/state/detail/SCPRC-EST2019-18+POP-RES.csv>

1. Rename the query **US Population[[5]](#footnote-5)**
2. The **Power Query Editor** should be like (some APPLIED STEPS will be different from the screen shot and the US County Population query gets added later in the lab):



1. **Close & Apply**
2. Go to the **Model** viewto create a relationship between **Province\_State (COV2020\_CSSE\_US)** and **Name (US Population), do not connect Province\_State to State**
   1. The model should look like (**NOTE:** There is no relationship between the COV2020\_CSSE global and US datasets):

****

1. Go to the **Report** view, select the **COV2020\_CSSE\_US** dataset, add the following four **DAX measures** (same formulas used in with the Global dataset, but applied to the US dataset, notice the **table names are different**):

Confirmed US = CALCULATE(SUM(COV2020\_CSSE\_US[Value]),COV2020\_CSSE\_US[Item]="Confirmed")

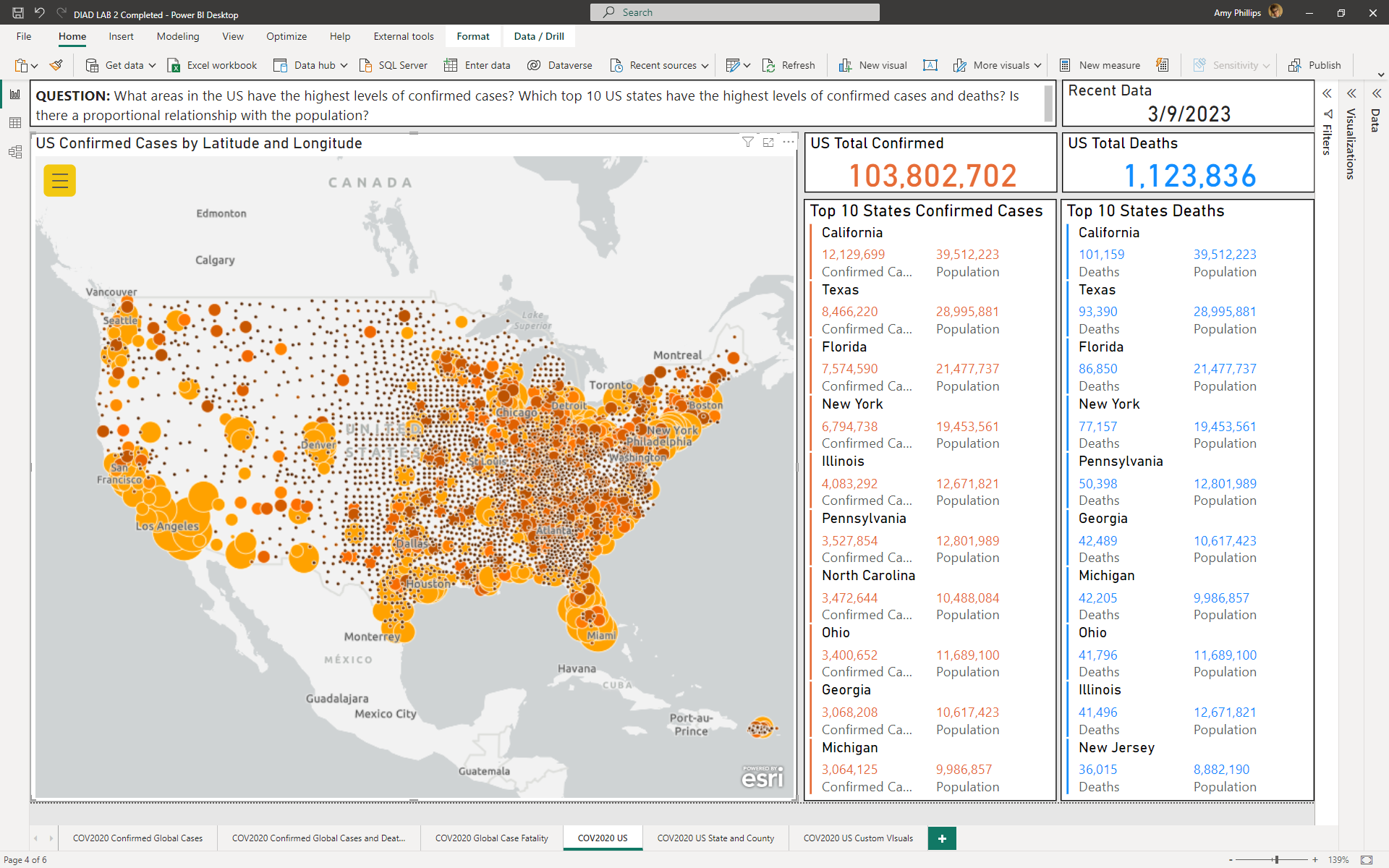
Confirmed Recent Value US = CALCULATE(SUM(COV2020\_CSSE\_US[Value]),COV2020\_CSSE\_US[Item]="Confirmed",FILTER(COV2020\_CSSE\_US, COV2020\_CSSE\_US[Date]= MAX([Date])))

Deaths US = CALCULATE(SUM(COV2020\_CSSE\_US[Value]),COV2020\_CSSE\_US[Item]="Deaths")

Deaths Recent Value US = CALCULATE(SUM(COV2020\_CSSE\_US[Value]),COV2020\_CSSE\_US[Item]="Deaths", FILTER(COV2020\_CSSE\_US, COV2020\_CSSE\_US[Date]= MAX([Date])))

## **LESSON 4: APPLY SPECIFIC VISUALS TO THE TYPE OF DATA USING COV2020\_CSSE\_US AND US POPULATION**

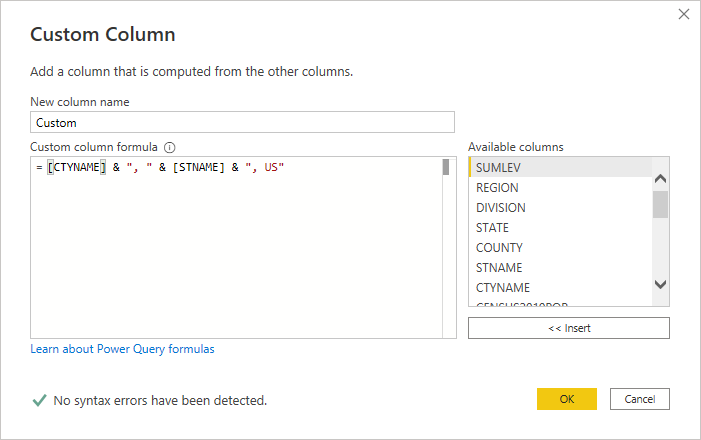
1. Create a **new report/page** (e.g., dashboard), **try to replicate:**



1. **Start with a question** (include as a textbox at the top of the dashboard)
   1. **QUESTION:** What areas in the US have the highest levels of confirmed cases?Which top 10 US states have the highest levels of confirmed cases and deaths? Is there a proportional relationship with the population?
2. **NOTE 1: The type of visual will be appropriate for the type of data we are using, for example:**
   1. **Use cards/multirow cards with tabular, single number or multi-number values to communicate precise values**, data to communicated quantitative information (not trends)
   2. **Use maps with geographic data**
3. **NOTE 2: For the map** 
   * 1. **Use Lat, Long with the ArcGIS map**[[6]](#footnote-6)
     2. **Size and Color = Confirmed Recent Value US**
     3. **To stop the map from zooming in and out, zoom to the level of granularity you want, use Format option:**
        + **Map tools, Lock extent = on**
4. **NOTE 3: Try to match font’s, colors, labels, titles**
   1. **Colors = Default color palette (“orange” for confirmed cases and “blue” for deaths)**
   2. **Card** 
      * **Data label = 28pt**
      * **Category = Off**
      * **Card Title = 14pt**
   3. **Multi-row card** 
      * **Data label = 10pt**
      * **Category = 10pt**
      * **Card Title = 14pt**
      * **Card Padding = 2pt**
5. **Rename** the report/page **COV2020 US**
6. **Save** your Power BI file

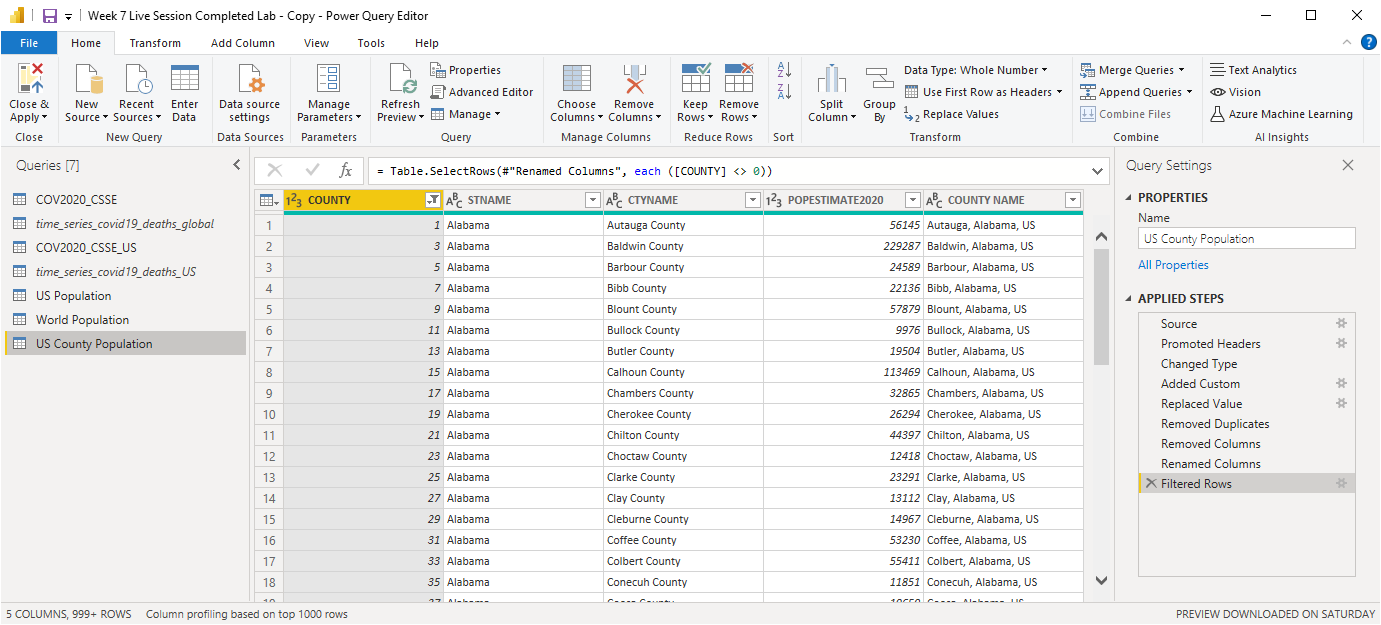
## **LESSON 5: ETL US CENSUS DATA FOR COUNTY POPULATION**

1. **Using the COV2020\_CSSE\_US** table, **create a report for individual states using county population data**
2. **Go to the Transform data, and import data from** <https://www2.census.gov/programs-surveys/popest/datasets/2010-2020/counties/totals/co-est2020.csv>
   1. **NOTE: We need want to import population data for each county in the US**
3. **Add Custom Column that will concatenate the CTYNAME, STATE and “US”**

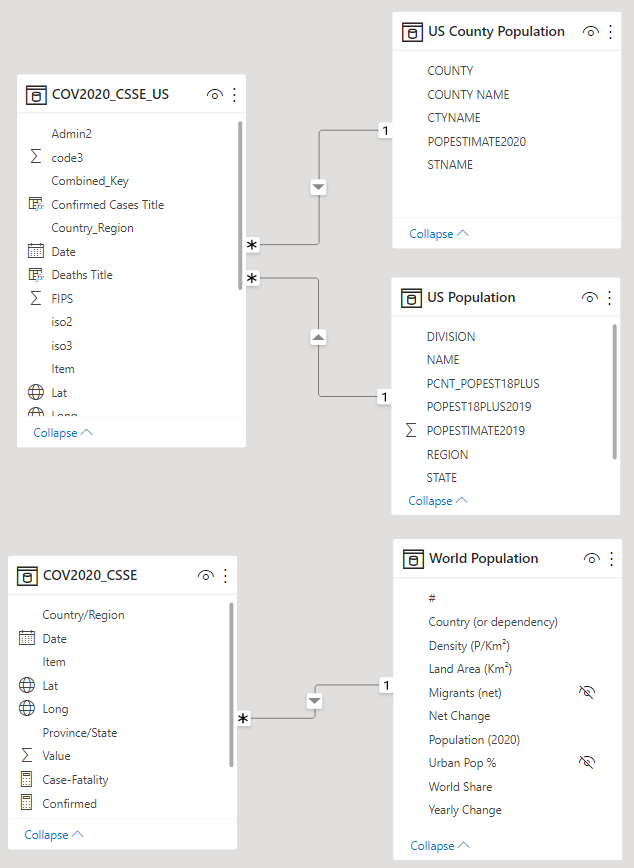


* 1. **NOTE: We need to add the column with the right naming convention to create a relationship with the COV2020\_CSSE\_US dataset**

1. **Rename Custom column to COUNTY NAME, change datatype to Text**
2. **Using the COUNTY NAME column, Replace Values “ County” with “” (e.g., nothing, blank, null)**
3. **Using the COUNTY NAME column, Replace Values “ Parish” with “” (e.g., nothing, blank, null)**
   1. **NOTE: Alaska will not work with our datasets, as the naming conventions used for COUNTY NAME will not match the Combined\_Key in the US County Population (we could manually add this if needed)**
4. **Using the COUNTY NAME column, remove duplicates**
5. **Remove all the columns except for COUNTY, STNAME, CTYNAME, POPESTIMATE2020, COUNTY NAME**
6. **Filter out COUNTY that has a value = 0, this is data at the state level which we do not need and creates a M:N using the COUNTY NAME with CO2020\_CSSE\_US Combined\_Key**
7. **Rename the query US County Population**



1. **Close & Apply**
2. **Go to the Model view, create a relationship between COUNTY NAME (US County Population) and Combined\_Key (CO2020\_CSSE\_US)**

****

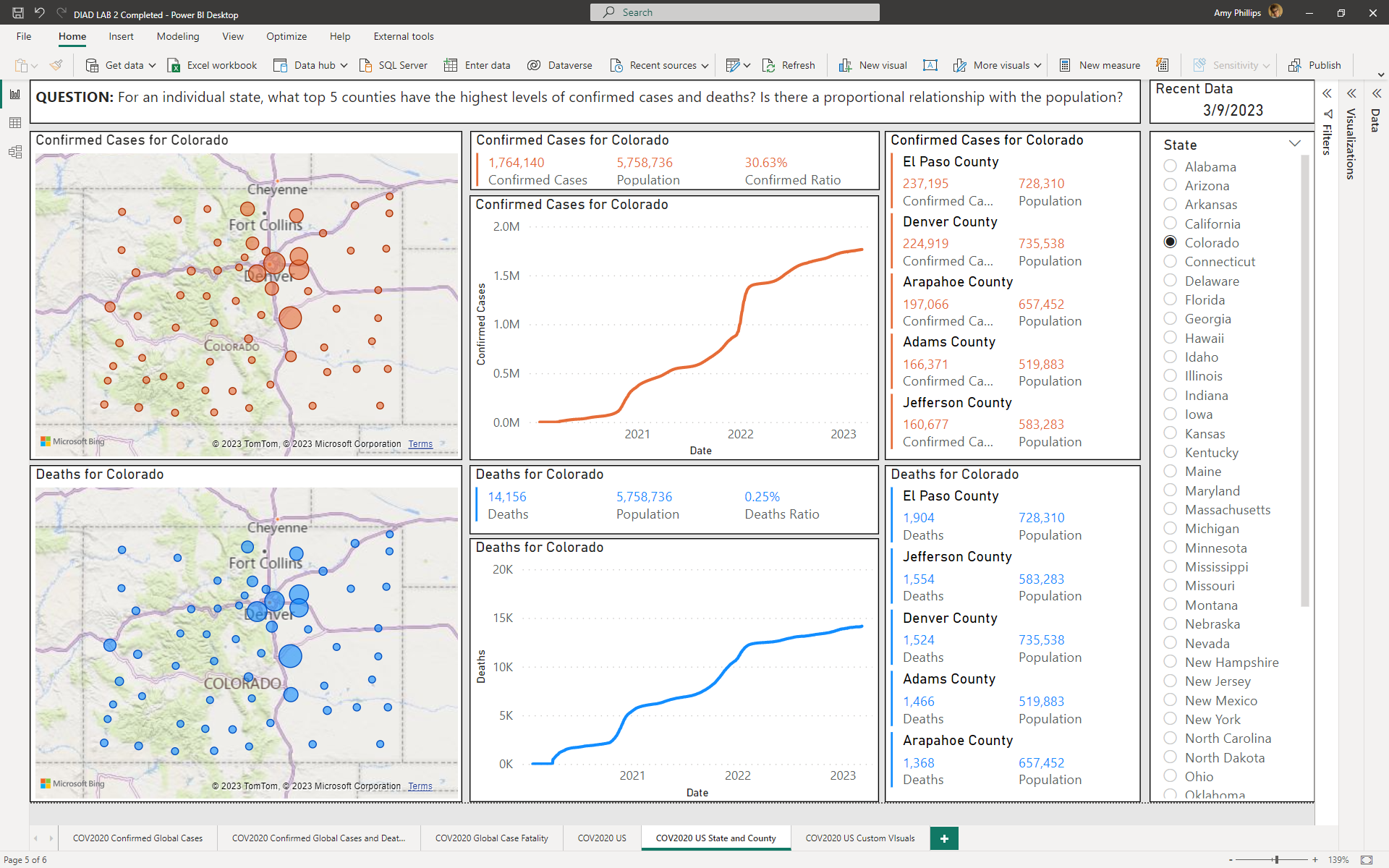
1. Go to the **Report** view, select the **COV2020\_CSSE\_US** dataset, add the following **DAX measures**:

Confirmed Ratio US = DIVIDE([Confirmed Recent Value US], CALCULATE(SUM('US Population'[POPESTIMATE2019])))

Deaths Ratio US = DIVIDE([Deaths Recent Value US], CALCULATE(SUM('US Population'[POPESTIMATE2019])))

## **LESSON 5: APPLYING MORE SPECIFIC VISUALS TO THE TYPE OF DATA USING COV2020\_CSSE\_US AND US POPULATION**

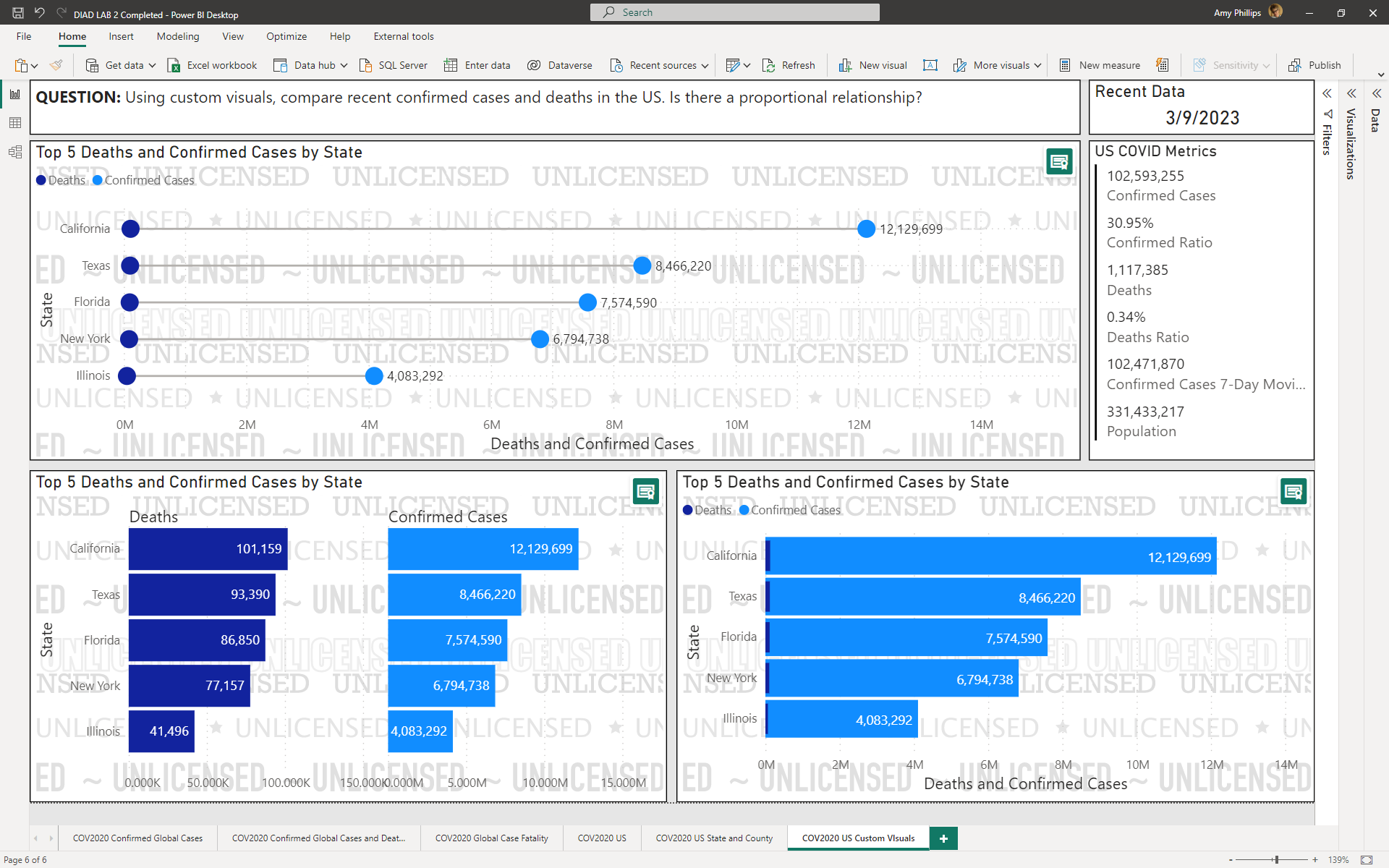
1. Create a **new report/page** (e.g., dashboard), we will **try to replicate:**



1. Start with a question (include as a textbox at the top of the dashboard)
   1. **QUESTION: For an individual state, what top 5 counties have the highest levels of confirmed cases and deaths? Is there a proportional relationship with the population?**
2. **NOTE 1: The type of visual will be appropriate for the type of data we are using much like PART II, except we include trend data using a line chart and a slicer**
3. **NOTE 2: Use the same colors, fonts, and similar font sizes as those in the previous lesson**
4. **NOTE 3:** In the **US Population** table, set the **POPESTIMATE2019** to **Do Not Summarize**
5. **NOTE 4:** Use the **Filter on this page** (**not Filters on this visual**) for **CTYNAME** to **unselect (Blank)**
6. **NOTE 5:** The **Slicer** should only allow the **Selection controls** to use **Single select**
   1. **SIDEBAR:** The **Slicer** visualis one of the few times that **scrolling is “acceptable**”
7. **CHALLENGE:** Can you figure out how to get the title to include the name of the state, so the title changes dynamically for each state selected in the slicer?
   1. **HINT:** Use DAX columns to concatenate a title with [Province\_State], then use a Format Painter option to select the DAX column
8. **Rename** the report/page **COV2020 US State and County**
9. **Save** the Power BI file

## **LESSON 6: CUSTOM VISUALS USING THE US DATASET**

1. Create a **new report/page** (e.g., dashboard), we will **try to replicate**:



1. Start with a question (include as a textbox at the top of the dashboard)
   1. **QUESTION:** Using custom visuals, compare recent confirmed cases and deaths in the US. Is there a proportional relationship?
2. Go to the **Report** view, select the **COV2020\_CSSE\_US** dataset, add the following **DAX measure**:

7-Day Moving Average Confirmed Cases US = AVERAGEX(DATESBETWEEN('COV2020\_CSSE\_US'[Date], DATEADD(LASTDATE('COV2020\_CSSE\_US'[Date]), -7, DAY), LASTDATE('COV2020\_CSSE\_US'[Date])), CALCULATE([Confirmed Recent Value US]))

1. **NOTE 1:** The report uses three custom visuals, the **Dumbbell Bar Chart**, **Merged Bar Chart**, and **Lipstick Bar Chart**[[7]](#footnote-7)
2. **NOTE/QUESTION:** There is only **one visual** that should be used in a dashboard not all three, beside the Multi-row card… **which custom visual tells the better story**? What is **misleading** about one of the visuals?
3. Rename the report/page **COV2020 US Custom Visuals**
4. **Save** the Power BI file

1. The categories are somewhat arbitrary, but were allocated based on the most densely populated countries <https://worldpopulationreview.com/country-rankings/countries-by-density> [↑](#footnote-ref-1)
2. Vanuatu is a South Pacific Ocean nation made up of roughly 80 islands that stretch 800 miles. [↑](#footnote-ref-2)
3. For reference: <https://coronavirus.jhu.edu/data/mortality> [↑](#footnote-ref-3)
4. MS Zaandam is a cruise ship owned and operated by Holland America Line, named for the city of Zaandam, Netherlands near Amsterdam. [↑](#footnote-ref-4)
5. We are adding a table for US population since the field population in the COV2020\_CSSE\_US is incomplete. Description of fields names and attribute values: <https://www2.census.gov/programs-surveys/popest/technical-documentation/file-layouts/2010-2019/SCPRC-EST2019-18+POP-RES.pdf> [↑](#footnote-ref-5)
6. If the ArcGIS map is misbehaving, go to **File** > **Options and settings** > **Options** > **Security** > and check **ArcGIS for Power BI** and **Map and Filled Map visuals** boxes [↑](#footnote-ref-6)
7. **NOTE:** There might be watermarks on some custom visuals, those that are considered “unlicensed”. Proof of concept here. [↑](#footnote-ref-7)