CIS 5200 Term Project Tutorial

Group 6

Authors: Krishna Ghorpade, Songyun Qian, Himani Patel, Shahnawaz Khan

Date: 12/05/2017

Lab Tutorial

Hazardous Air Pollutants in USA from 1990 to 2017 Analysis in Hive using IBM BigInsights

Objective:

In this lab, you will analyze and visualize Air Pollution Data. Thus,

* You should learn how to download Air Pollution Data from Kaggle.com, upload to Google Drive, then download to the local system in Bluemix BigInsights.
* Then you will learn how to upload it to HDFS.
* You will figure out how to manipulate and analyze air pollution data in HDFS using HiveQL.
* You will also practice how to visualize the result in Tableau.

Introduction:

Air pollution is a serious problem in the world right now. Hazardous air pollutants, also known as toxic air pollutants or air toxics are poisonous for human body. Those hazardous air can cause cancer or other serious health problems, such as reproductive problems to abnormality by birth time. This data set is from the Environmental Protection Agency (EPA) tracking 187 air pollutants from 1990 to 2017. The data set is a daily summary file, containing data for every monitor in the EPA database. You will learn how to:

* Analyze data to determine which air pollutants are measured the most
* Analyze data to determine which cities have the highest and lowest air pollutants measured
* Analyze data to determine which states have the highest and lowest air pollutants measured
* Analyze data to determine which dates have the highest and lowest air pollutants measured
* Transfer data using **WINSCP**
* Visualize data in **Tableau**

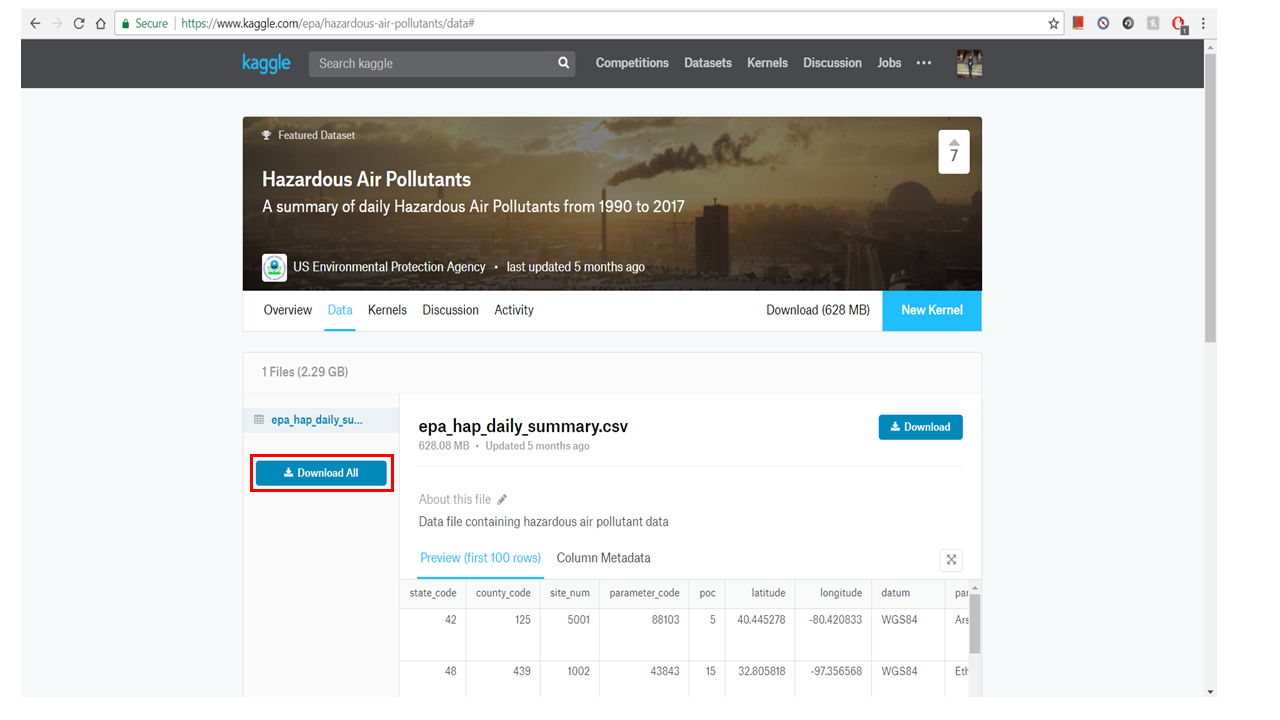
Prerequisites:

Everything you need to go through the scripts and queries is already provisioned with the cluster. To export the analyzed data to Tableau, you must meet the following requirements:

* You must have **Tableau** installed.
* You must have **WINSCP** installed to transfer the file.

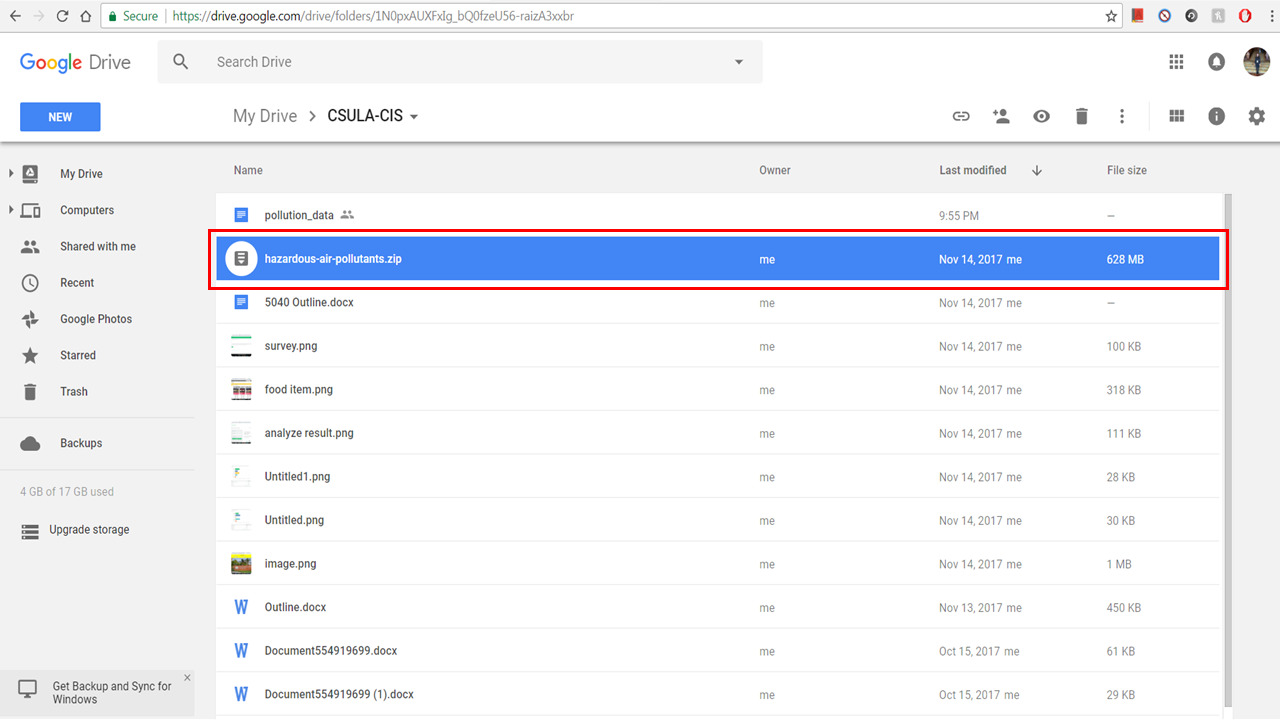
Air Pollutants data downloaded from Kaggle.com

You need to create an account on kaggle.com and download the data onto your computer. Keep it as a zip file. <https://www.kaggle.com/epa/hazardous-air-pollutants/data>



Air Pollutants data loaded into Google Drive

You need to create a Google Drive account and load the data (zip file) into Google Drive.



Air Pollutants data loaded into BigInsights

Right click the zip file after you uploaded it onto Goggle Drive, and get a sharable link. Copy the link

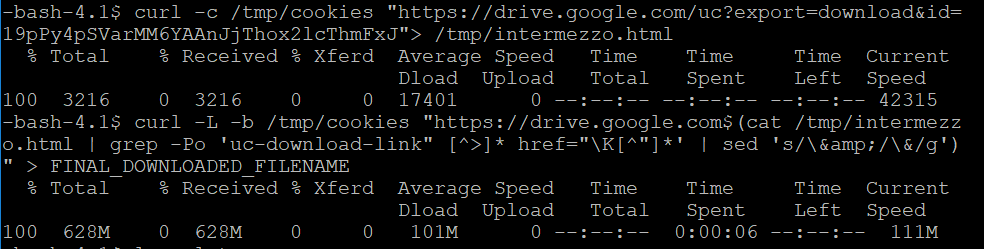
*https://drive.google.com/open?id=19pPy4pSVarMM6YAAnJjThox2lcThmFxJ*

You need to remotely access your BigInsights that you executed in your Bluemix account using ssh. You can download the data zip file Hazardous Air Pollutants from Google Drive:

(**Note**: Don’t forget to replace the red part with the link you have generated)

$ curl -c /tmp/cookies "https://drive.google.com/uc?export=download&id=19pPy4pSVarMM6YAAnJjThox2lcThmFxJ"> /tmp/intermezzo.html

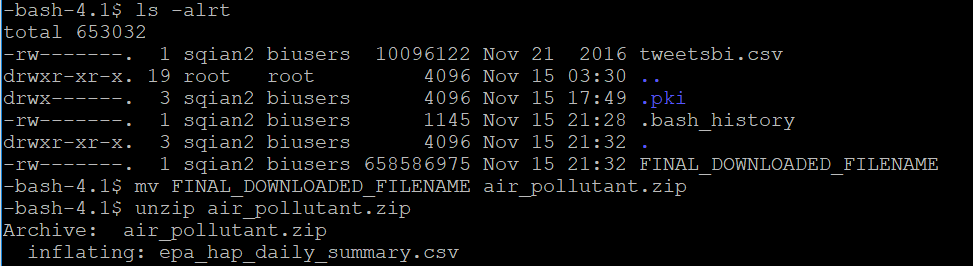
$ curl -L -b /tmp/cookies "https://drive.google.com$(cat /tmp/intermezzo.html | grep -Po 'uc-download-link" [^>]\* href="\K[^"]\*' | sed 's/\&amp;/\&/g')" > FINAL\_DOWNLOADED\_FILENAME



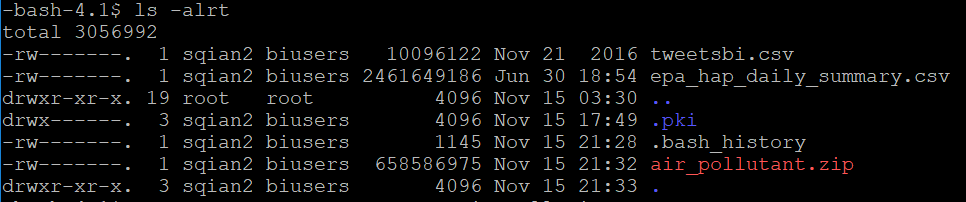
$ ls-alrt (check if this file FINAL\_DOWNLOADED\_FILENAME is available on path)

$ mv FINAL\_DOWNLOADED\_FILENAME air\_pollutant.zip

$ unzip air\_pollutant.zip



$ ls -alrt (check csv is available or not)



Create Hive table to Query Air Pollutants data

The following Hive statement creates an external table that allows Hive query stored in HDFS. External tables preserve the data in the original file format, while allowing Hive to perform queries against the data within the file.The Hive statement below create a new table named, air\_pollution, by describing the fields within the files, the delimiter (comma) between fields. This will allow you to create Hive queries over your data.

Open Hive shell environment as follow:

$ hive

In the Hive shell, you need to copy and paste the following Hive QL code to create an external table “air\_pollution”. (**Note**: Don’t forget to replace the red part with your account name)

hive> CREATE TABLE IF NOT EXISTS air\_pollution

(state\_code DECIMAL,

county\_code DECIMAL,

site\_num DECIMAL,

parameter\_code DECIMAL,

poc DECIMAL,

latitude DECIMAL (10,6),

longitude DECIMAL (10,6),

datum string,

parameter\_name string,

sample\_duration string,

pollutant\_standard string,

date\_local date,

units\_of\_measure string,

event\_type string,

observation\_count DECIMAL,

observation\_percent DECIMAL,

arithmetic\_mean DECIMAL,

first\_max\_value float,

first\_max\_hour float,

aqi string,

method\_code DECIMAL,

method\_name string,

local\_site\_name string,

address string,

state\_name string,

county\_name string,

city\_name string,

cbsa\_name string,

date\_of\_last\_change date)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

STORED AS TEXTFILE LOCATION '/user/sqian2/epa\_hap\_daily\_summary'

TBLPROPERTIES ('skip.header.line.count'='1');

Then, you have to load the data into the table.

(**Note**: Don’t forget to replace the red part with your account name)

hive>load data local inpath '/home/sqian2/epa\_hap\_daily\_summary.csv' into table air\_pollution;

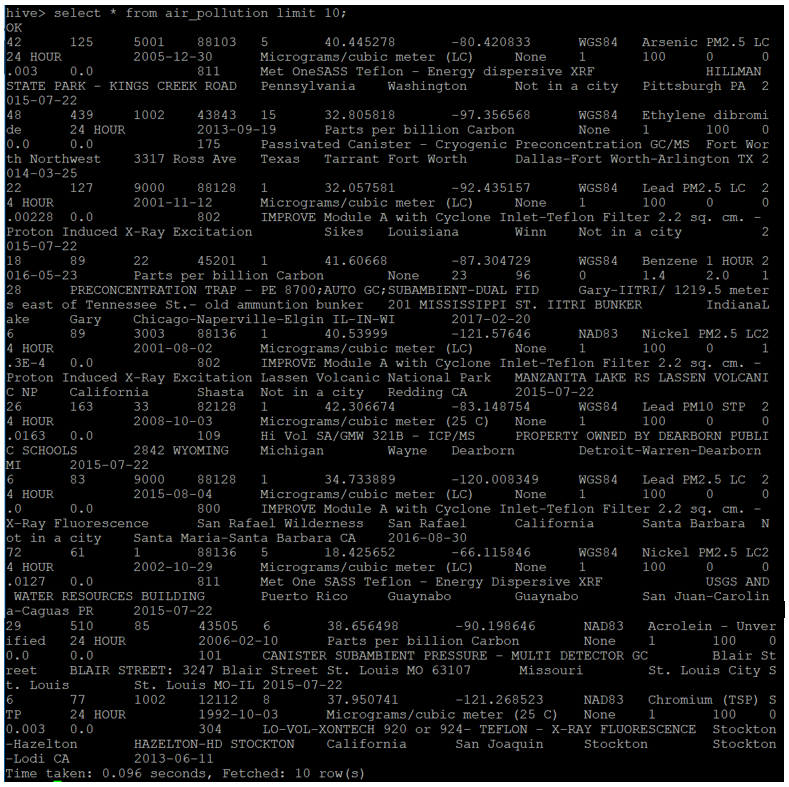


Then in the Hive shell, you need to check if the table “air\_pollution” is shown:

hive>show tables;

Now you can query the content of the air\_pollution table:

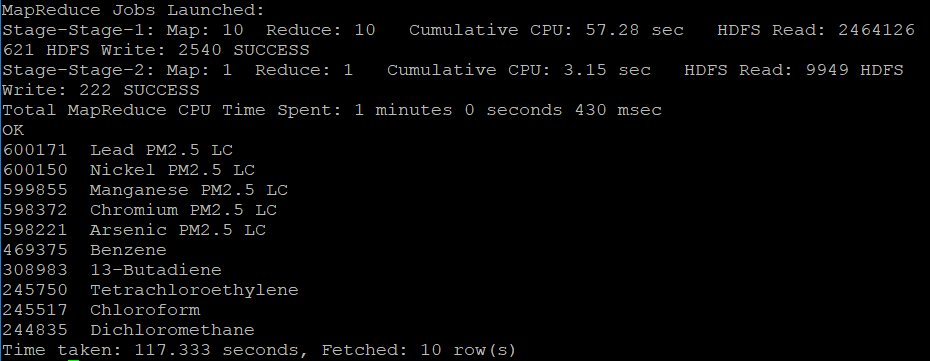
hive>select \* from air\_pollution limit 10;



Creating Hive Queries to Analyze Data

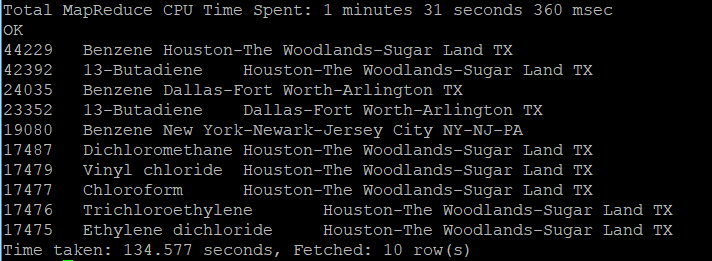
The following Hive Queries will show you the top 10 pollutants:

Hive>Select count(\*) as pollution,parameter\_name from air\_pollution group by parameter\_name order by pollution DESC limit 10;



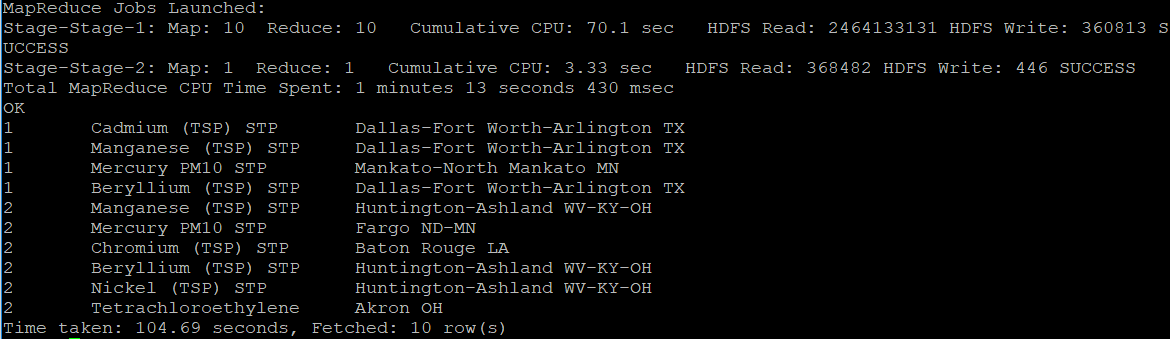
The following Hive Queries will show you the top 10 pollutants by city:

Hive> Select count(\*) as pollution,parameter\_name,cbsa\_name from air\_pollution where cbsa\_name != "" group by parameter\_name,cbsa\_name order by pollution DESC limit 10;



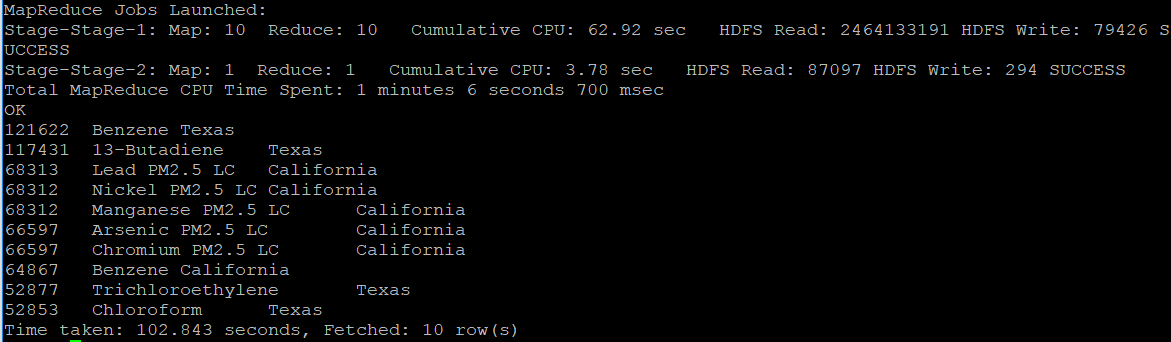
The following Hive Queries will show you the last 10 pollutants by city:

Hive> Select count(\*) as pollution,parameter\_name,cbsa\_name from air\_pollution group by parameter\_name,cbsa\_name order by pollution ASC limit 10;



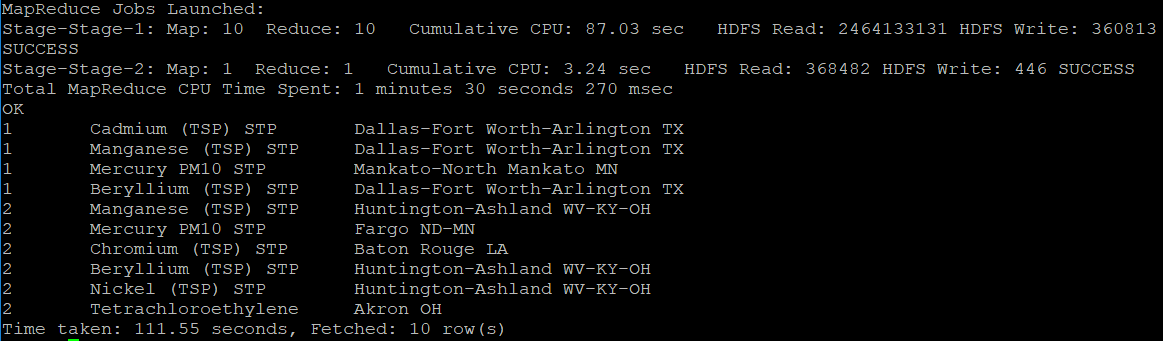
The following Hive Queries will show you the top 10 pollutants by state:

Hive> Select count(\*) as pollution,parameter\_name,state\_name from air\_pollution group by parameter\_name,state\_name order by pollution DESC limit 10;



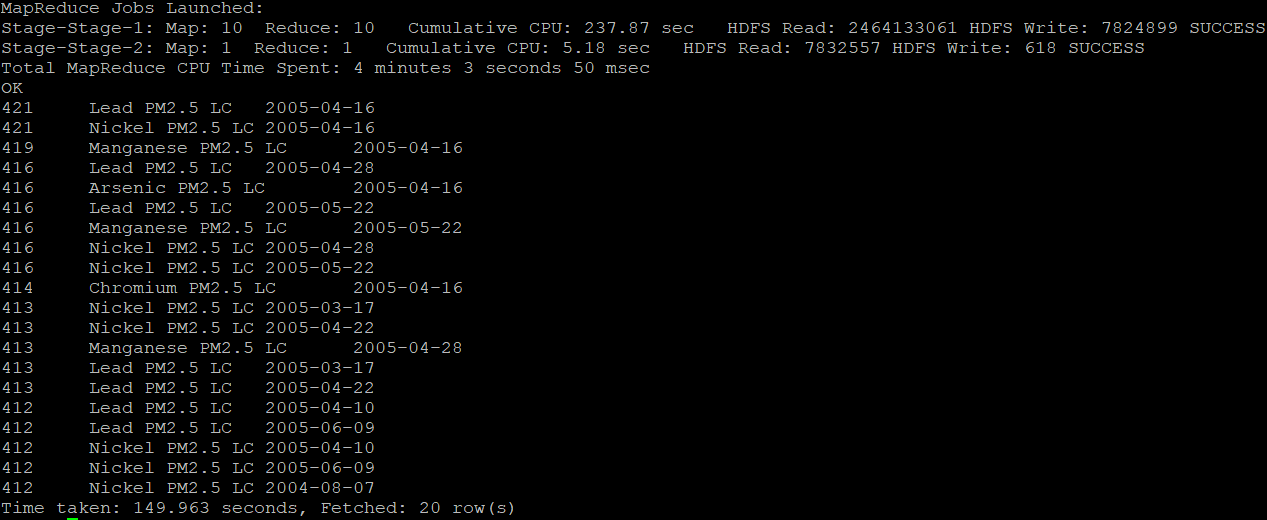
The following Hive Queries will show you the last 10 pollutants by state:

Hive> Select count(\*) as pollution,parameter\_name,cbsa\_name from air\_pollution group by parameter\_name,cbsa\_name order by pollution ASC limit 10;



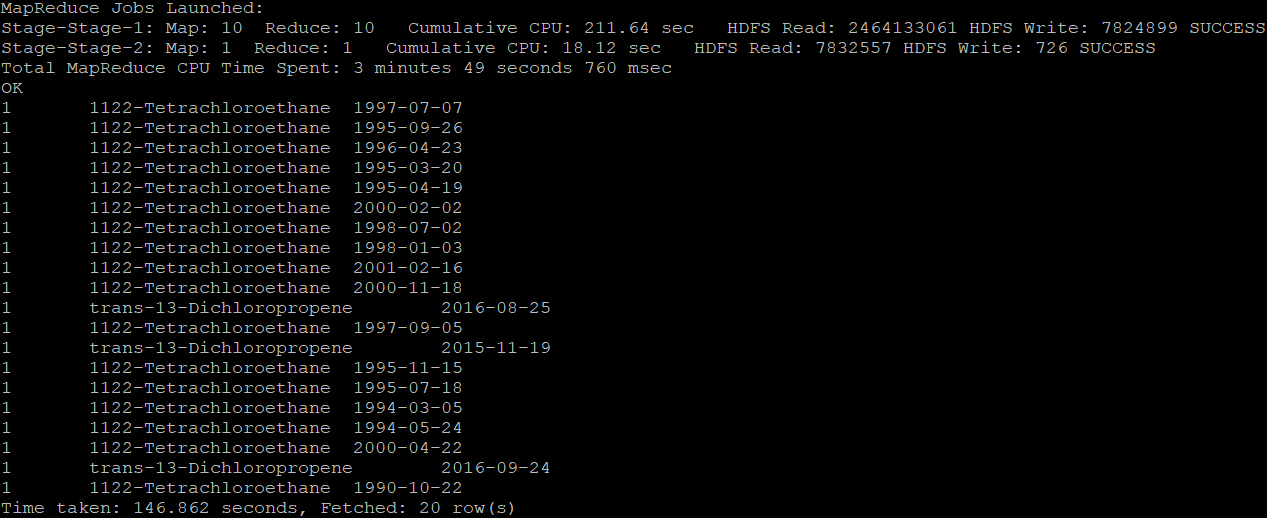
The following Hive Queries will show you the top 20 pollutants by date:

Hive> Select count(\*) as pollution,parameter\_name, date\_local from air\_pollution group by parameter\_name, date\_local order by pollution DESC limit 20;



The following Hive Queries will show you the last 20 pollutants by date:

Hive> Select count(\*) as pollution,parameter\_name, date\_local from air\_pollution group by parameter\_name, date\_local order by pollution ASC limit 20;



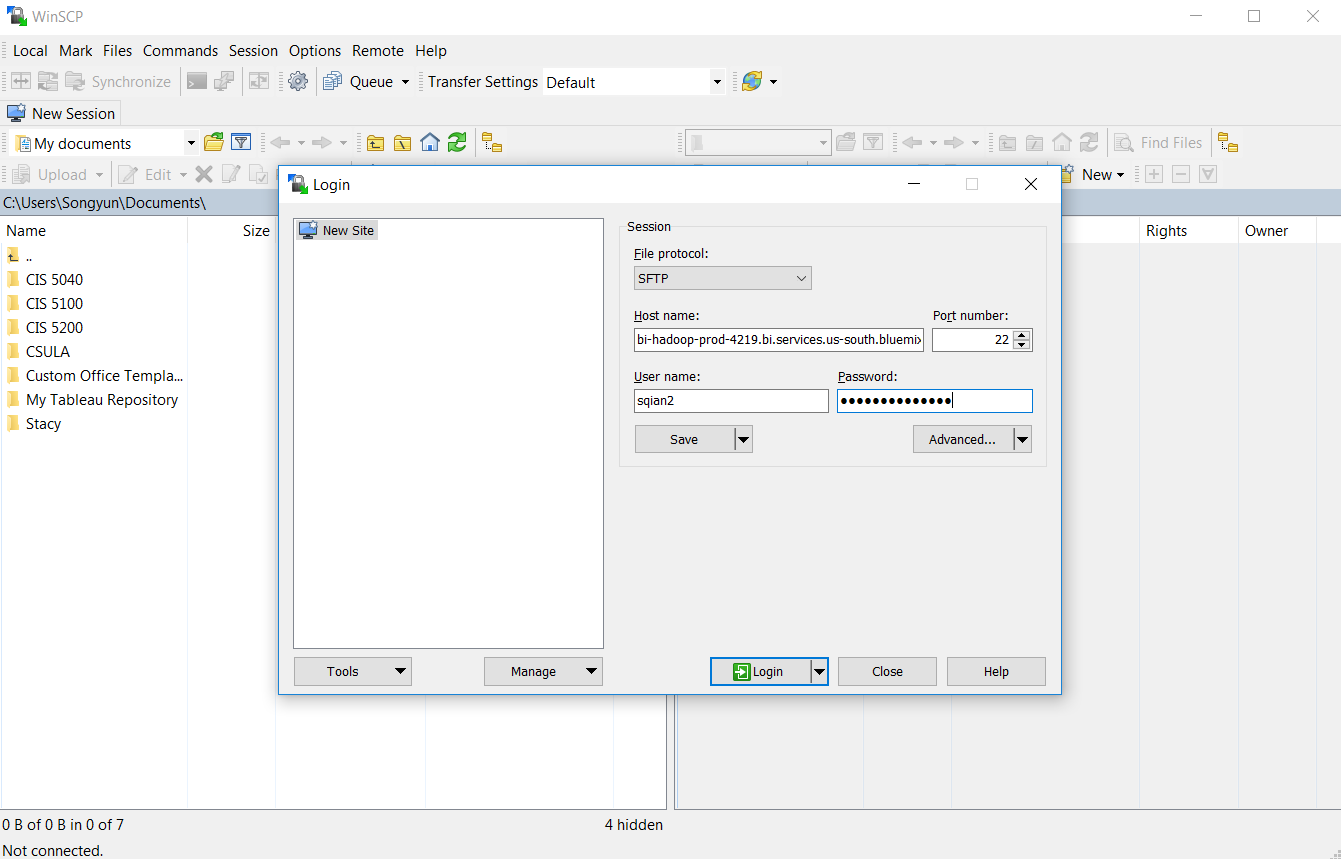
Create Tables for Tableau

Since the dataset is too big, you need to create a table for each Hive Queries:

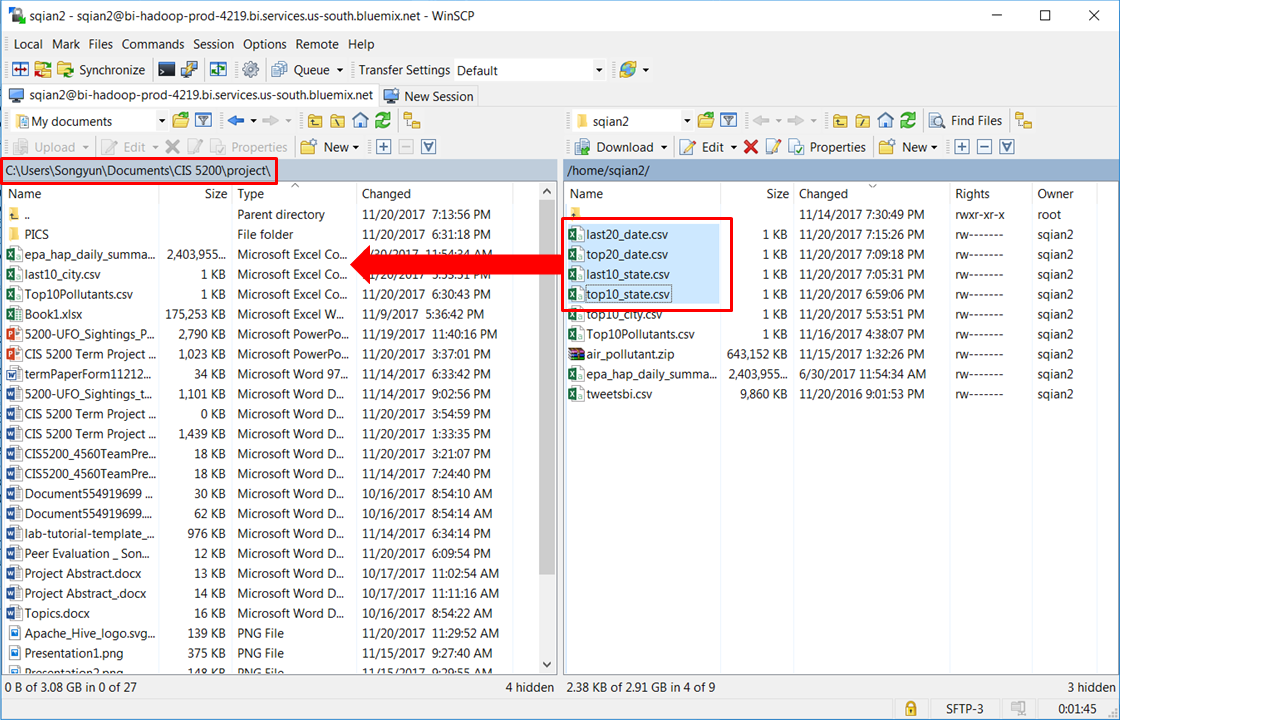
hive -e "use default ;Select count(\*) as pollution,parameter\_name,cbsa\_name from air\_pollution group by parameter\_name,cbsa\_name order by pollution ASC limit 10;"| perl -lpe 's/"/\\"/g; s/^|$/"/g; s/\t/","/g' > last10\_city.csv

Replace RED TEXT with each query that you need to create a table. Replace BLUE TEXT with the corresponding file name that you want to name it.

After the above query is done for each table, open WINSCP, and log in WINSCP as you would in putty using your Host name, user name, and password.



Drag the csv files you have created to a local location on your computer on the left side of the window. You should have **7 CSV files** downloaded.

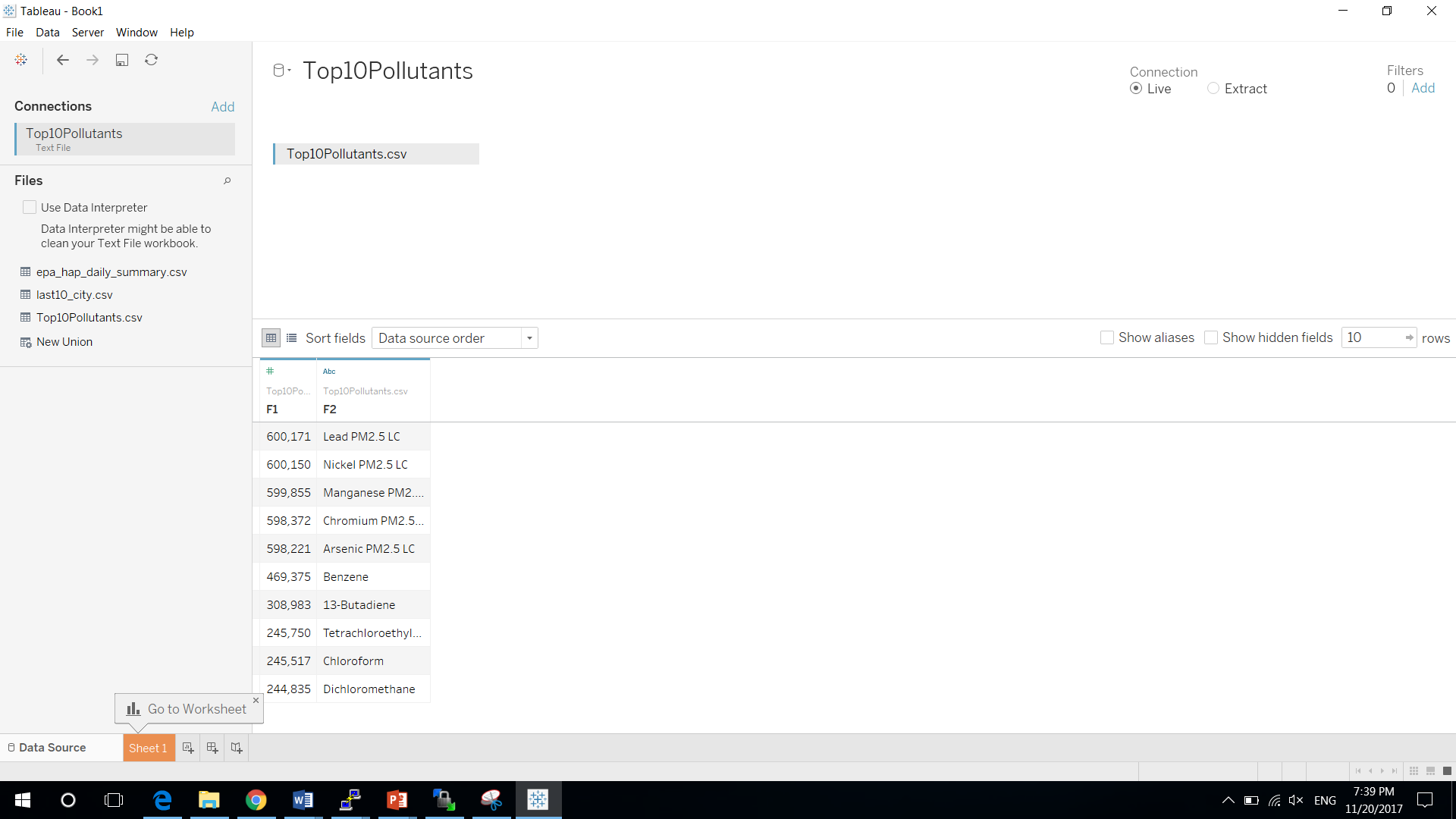


Once the csv files are downloaded, you need to open Tableau on your local computer.

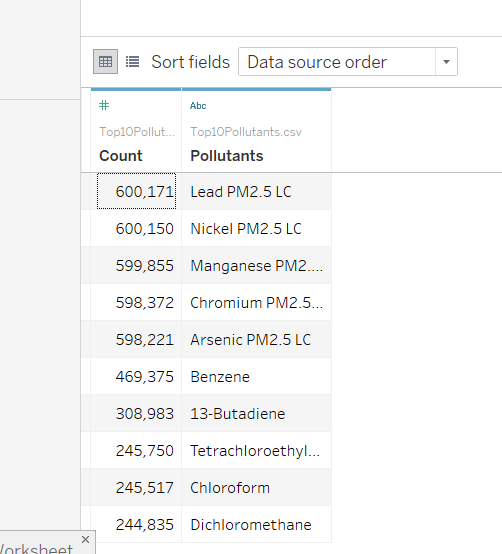
Tableau to open data file directly from Tableau and Visualization

Open Tableau, and open the file according to the following order.

1. Top 10 pollutants



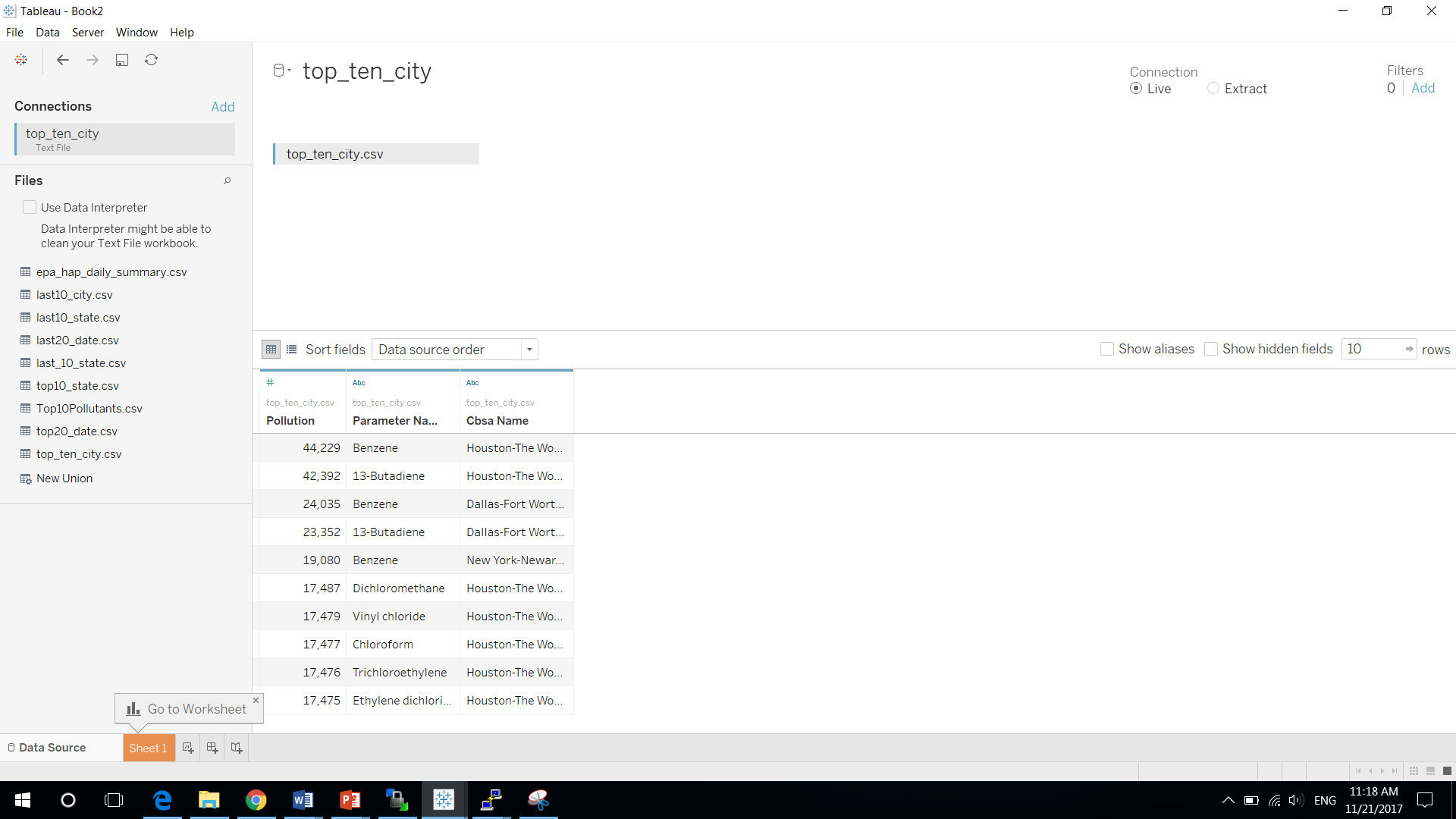
Rename F1 to Count, F2 to Pollutants.



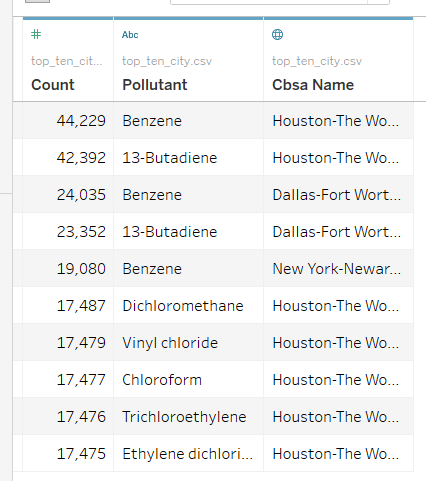
Select Sheet 1 next to Data Source, and drag Count to Rows and Pollutants to Columns. Drag Pollutants to Color, and you will get this chart:



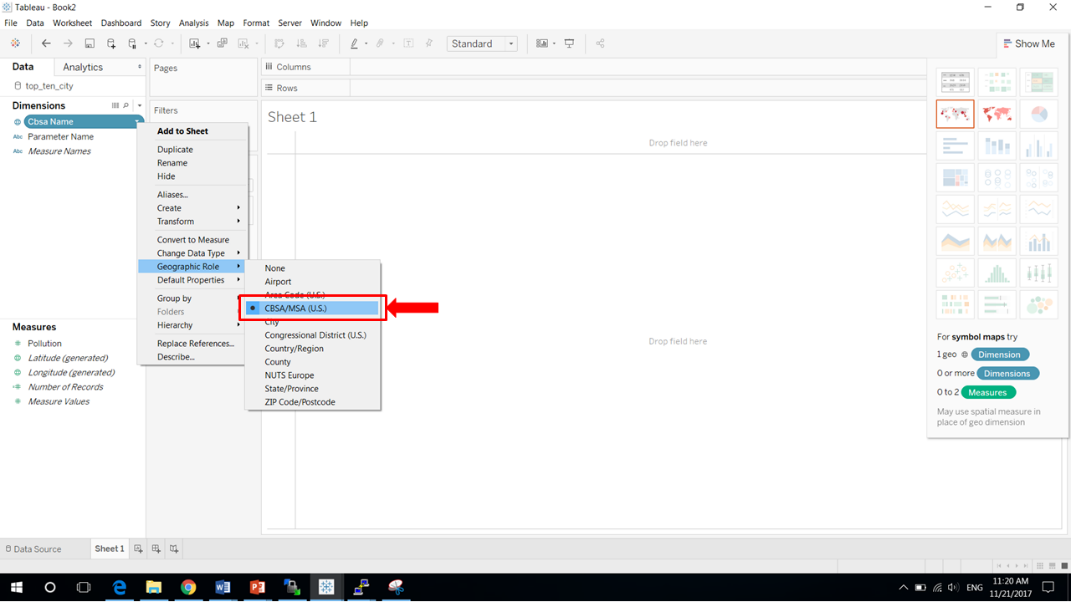
1. Top 10 Pollutants by City



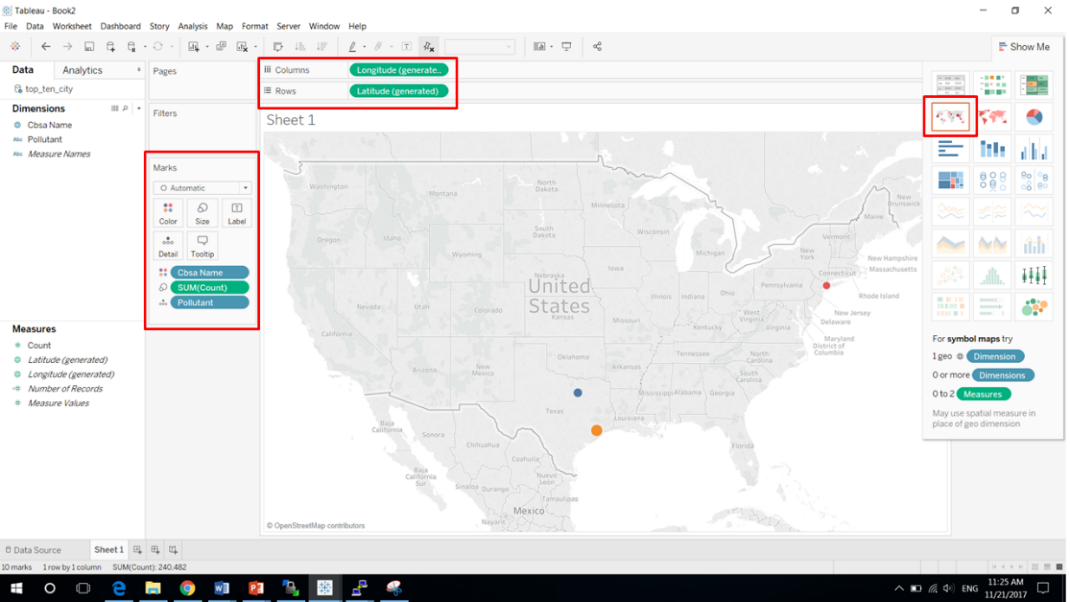
Rename F1 to Count, F2 to Pollutants, F3 to CBSA\_NAME.



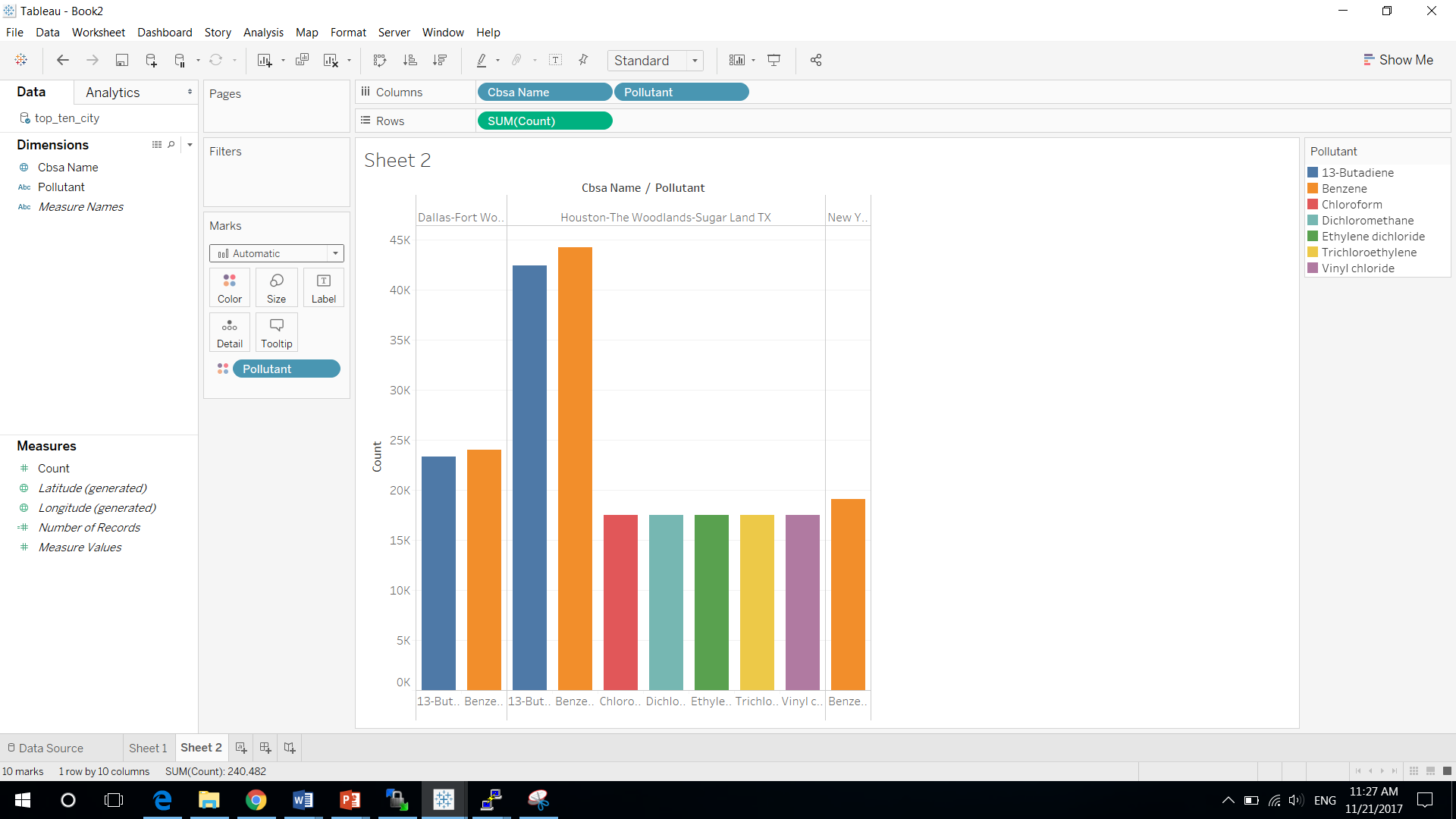
Select Sheet 1 next to Data Source, and change CBSA’s geographic role to CBSA/MSA(USA).



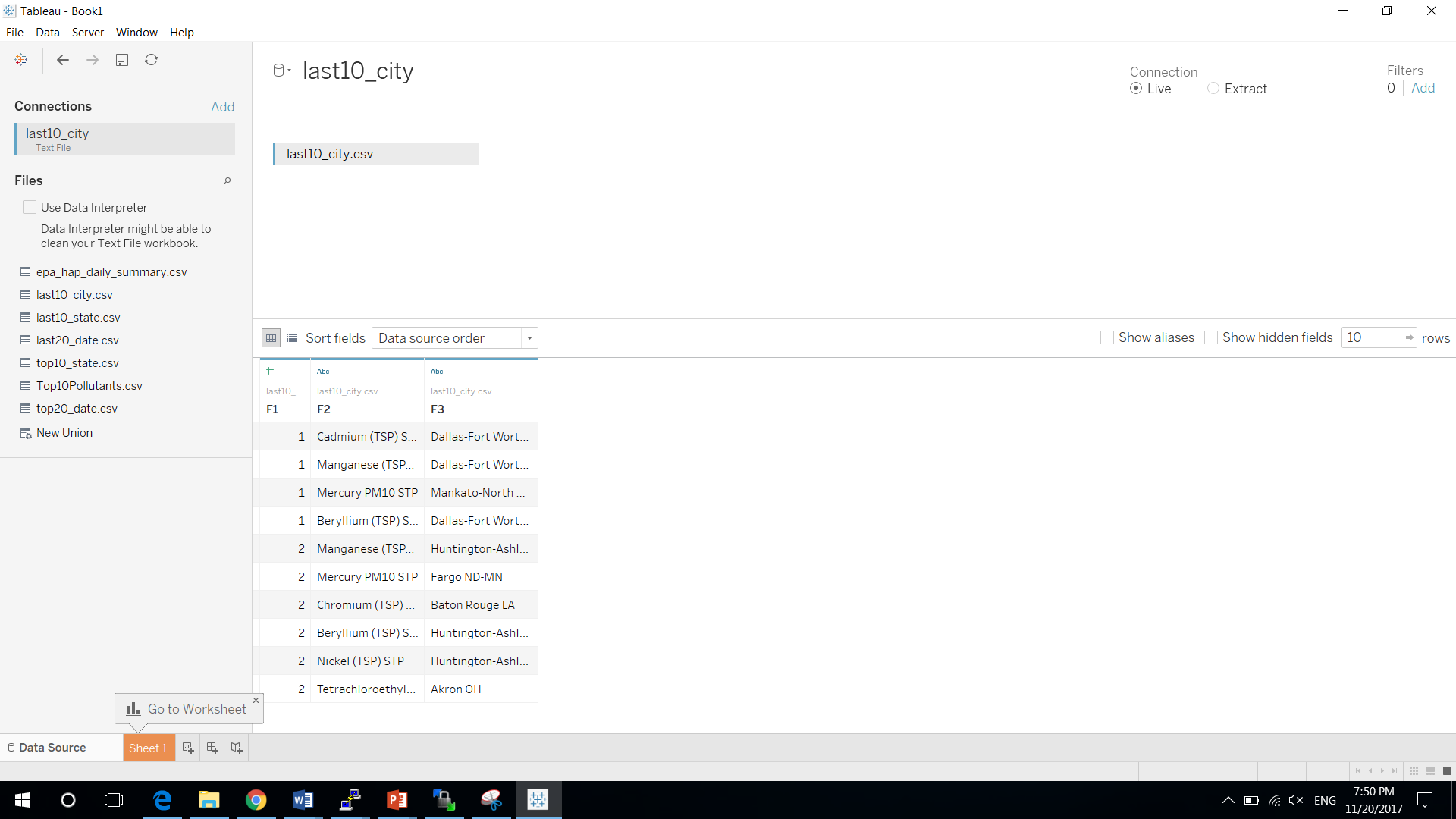
Drag Longitude(generated) to Columns, Latitude(generated) to Rows, CBSA to color, Count to Size, Pollutants to Detail. Select Show me, and select Geo Map:



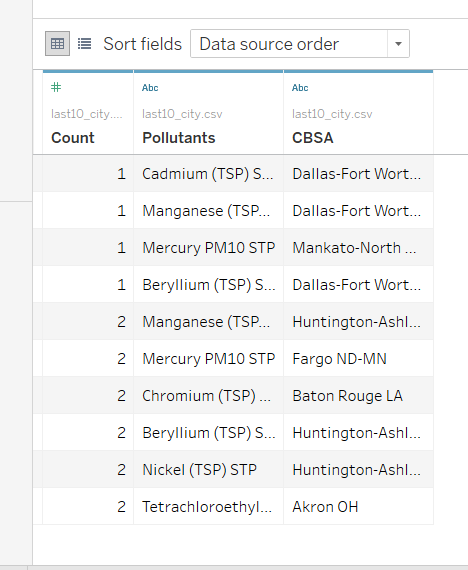
Create a new Worksheet by selecting the icon next to the Sheet 1. Drag Pollutants and CSBA to Columns and Count to Rows. Drag Pollutants to Color, you will get this:



1. Last 10 Pollutants by City

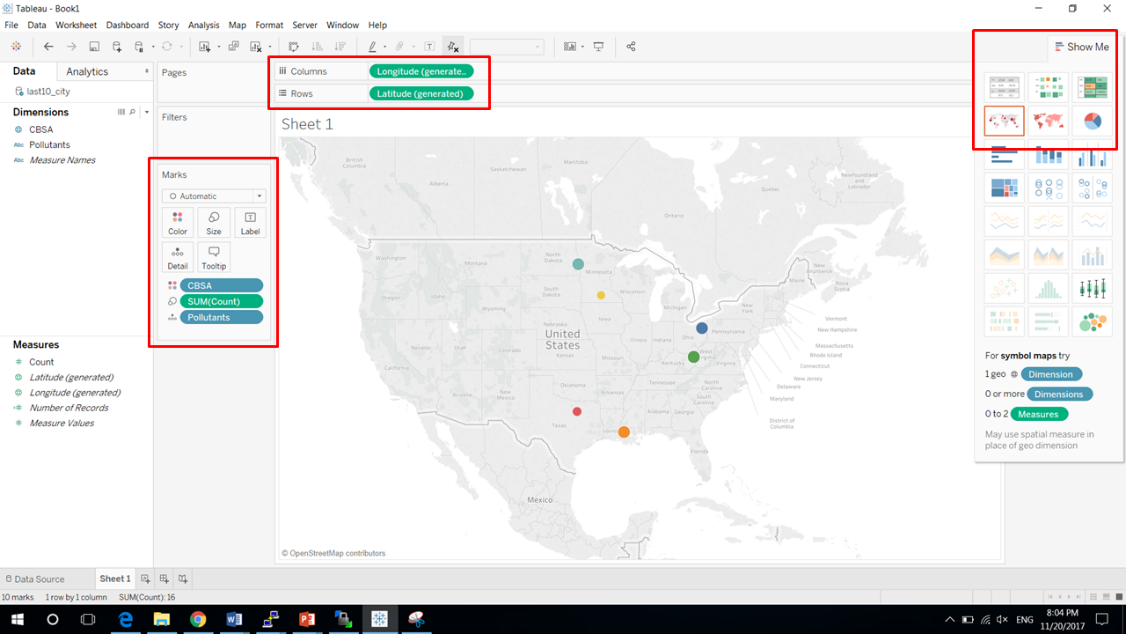


Rename F1 to Count, F2 to Pollutants, F3 to CBSA\_NAME.

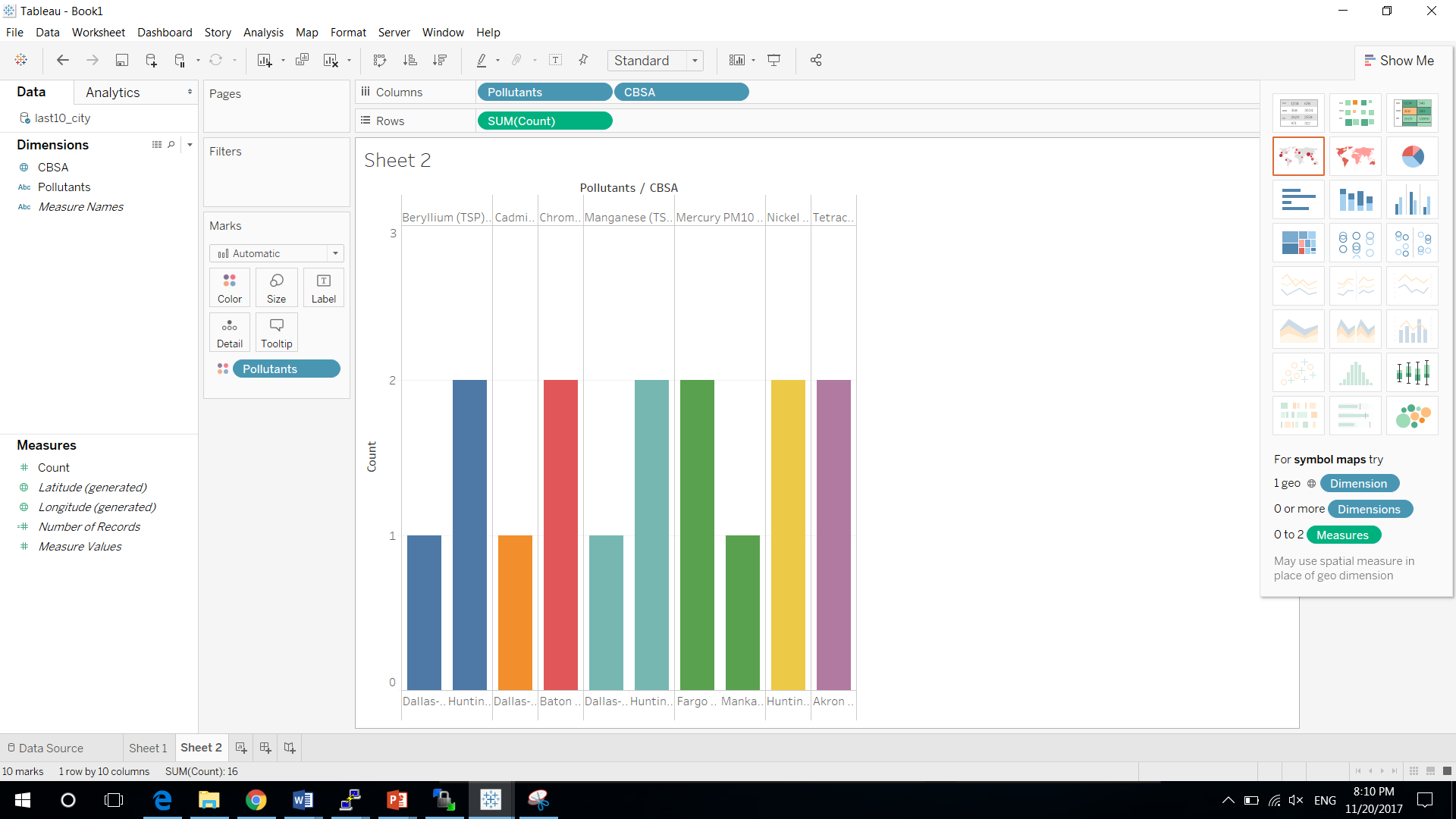


Select Sheet 1 next to Data Source, and change CBSA’s geographic role to CBSA/MSA(USA).

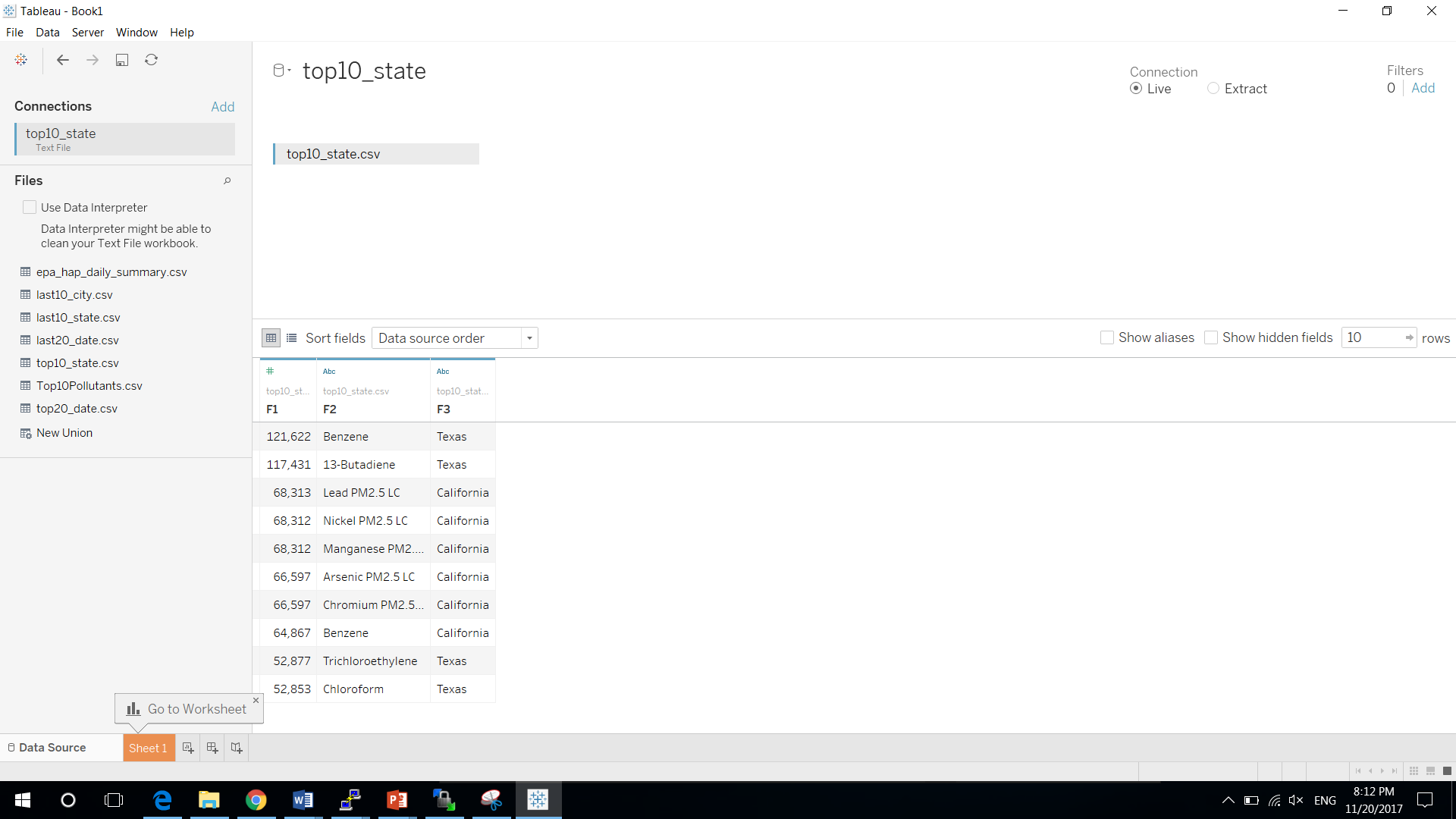
Drag Longitude(generated) to Columns, Latitude(generated) to Rows, CBSA to color, Count to Size, Pollutants to Detail. Select Show me, and select Geo Map:



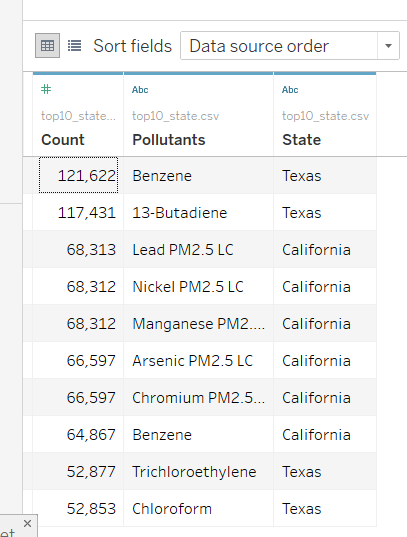
Create a new Worksheet by selecting the icon next to the Sheet 1. Drag Pollutants and CSBA to Columns and Count to Rows. Drag Pollutants to Color, you will get this:



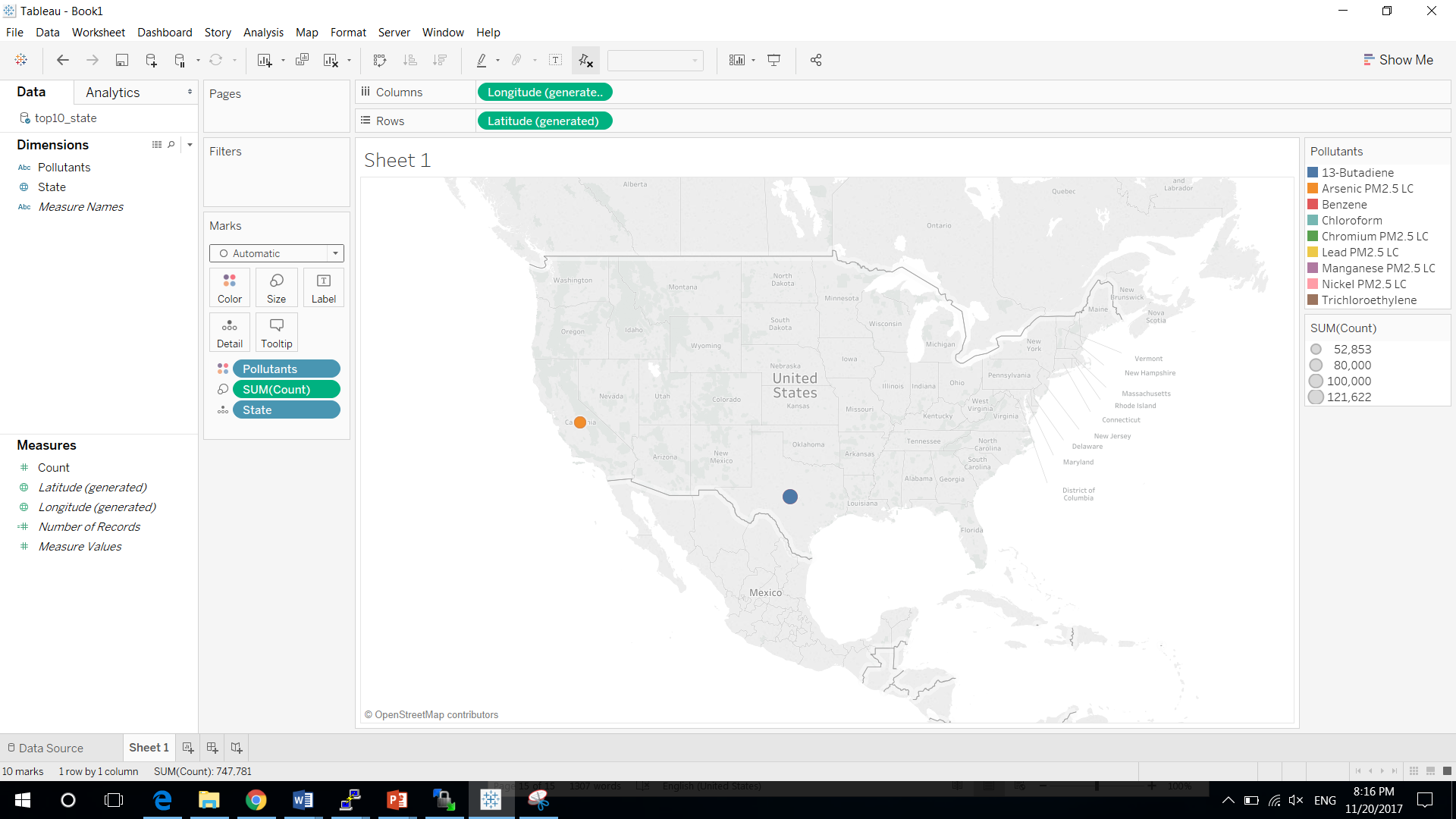
1. Top 10 Pollutants by state



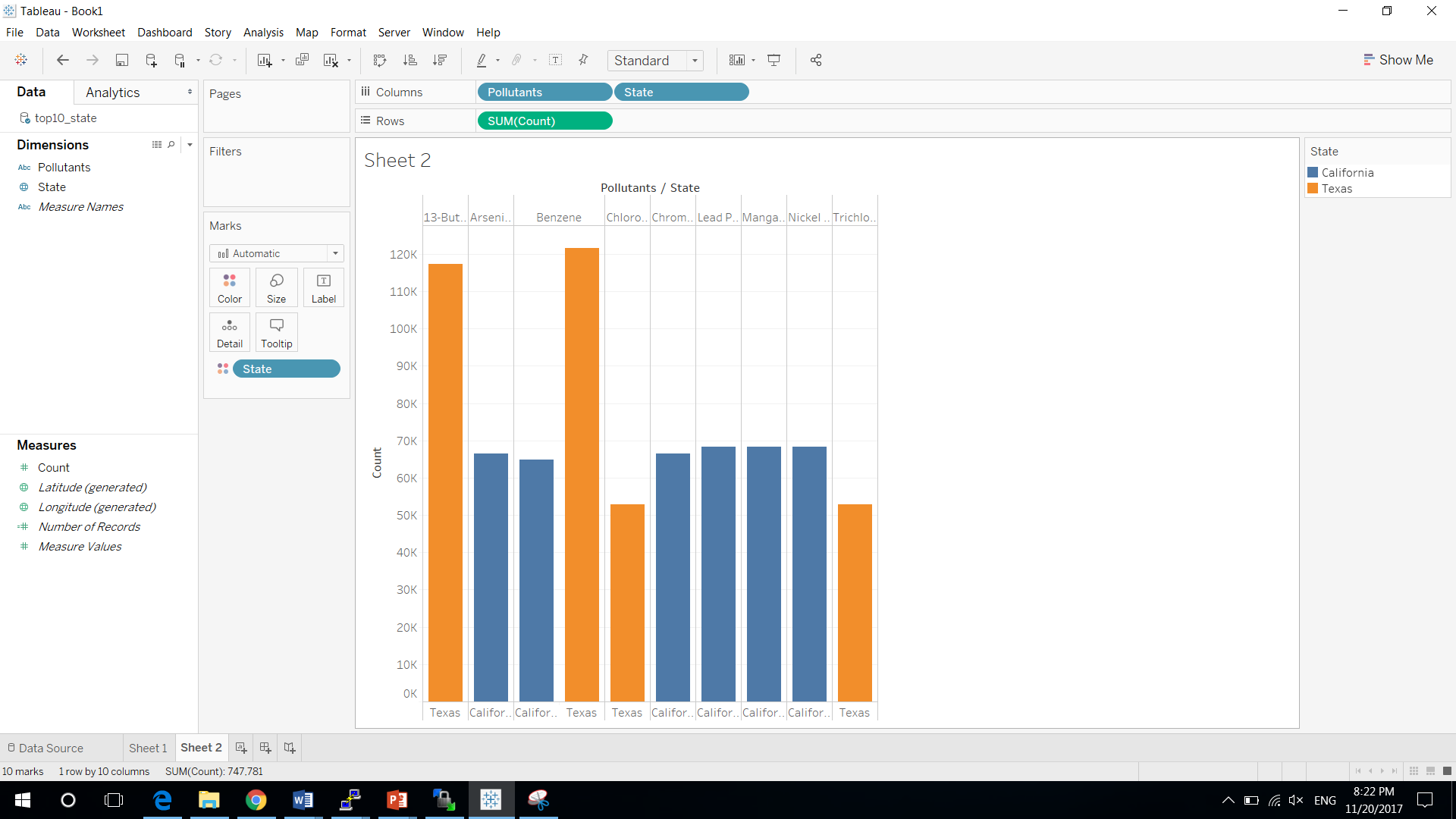
Change F1 to count, F2 to Pollutants, F3 to State



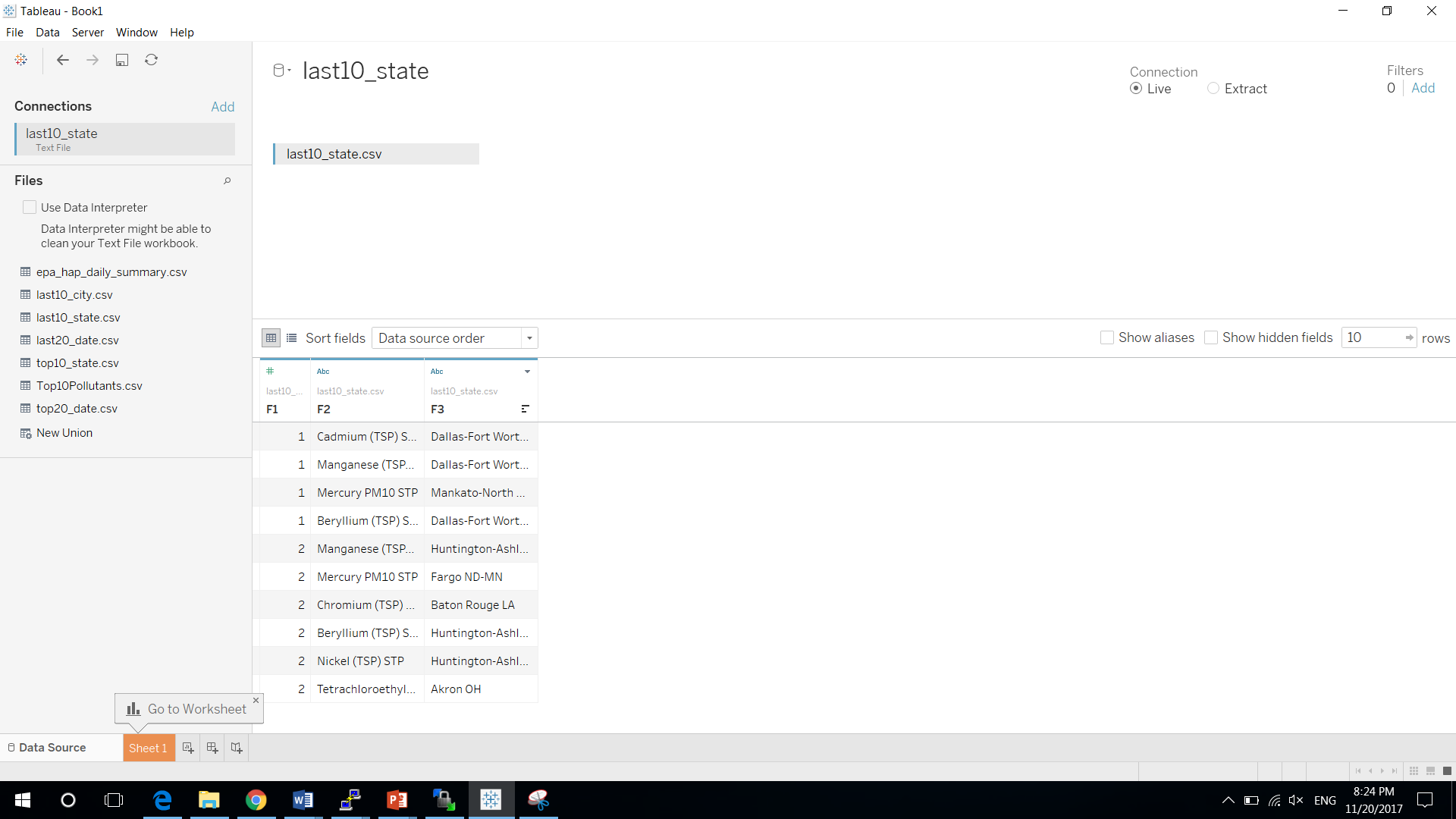
Select Sheet 1 next to Data Source, change State’s geographical role to State/Province. Drag Longitude to Columns, Latitude to Rows, Pollutants to Color, Count to Size, State to Details. And select Geo Map.



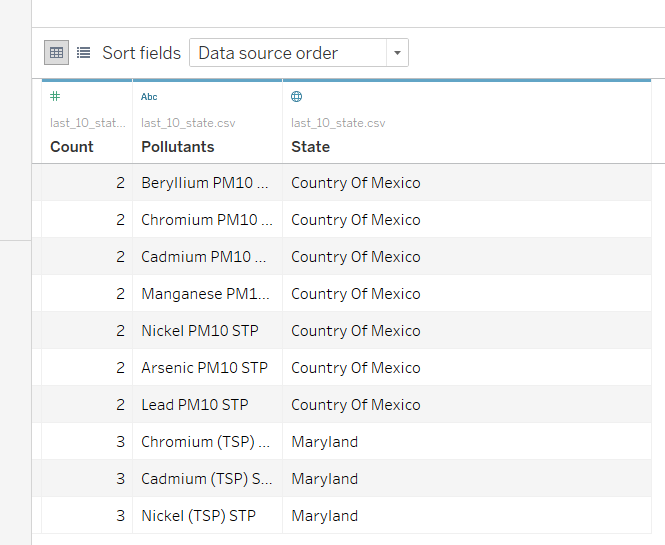
Create a new Worksheet by selecting the icon next to the Sheet 1. Drag Pollutants and State to Columns and Count to Rows. Drag State to Color, the bar chart will only generate two colors due to the top 10 are only in Texas and California.



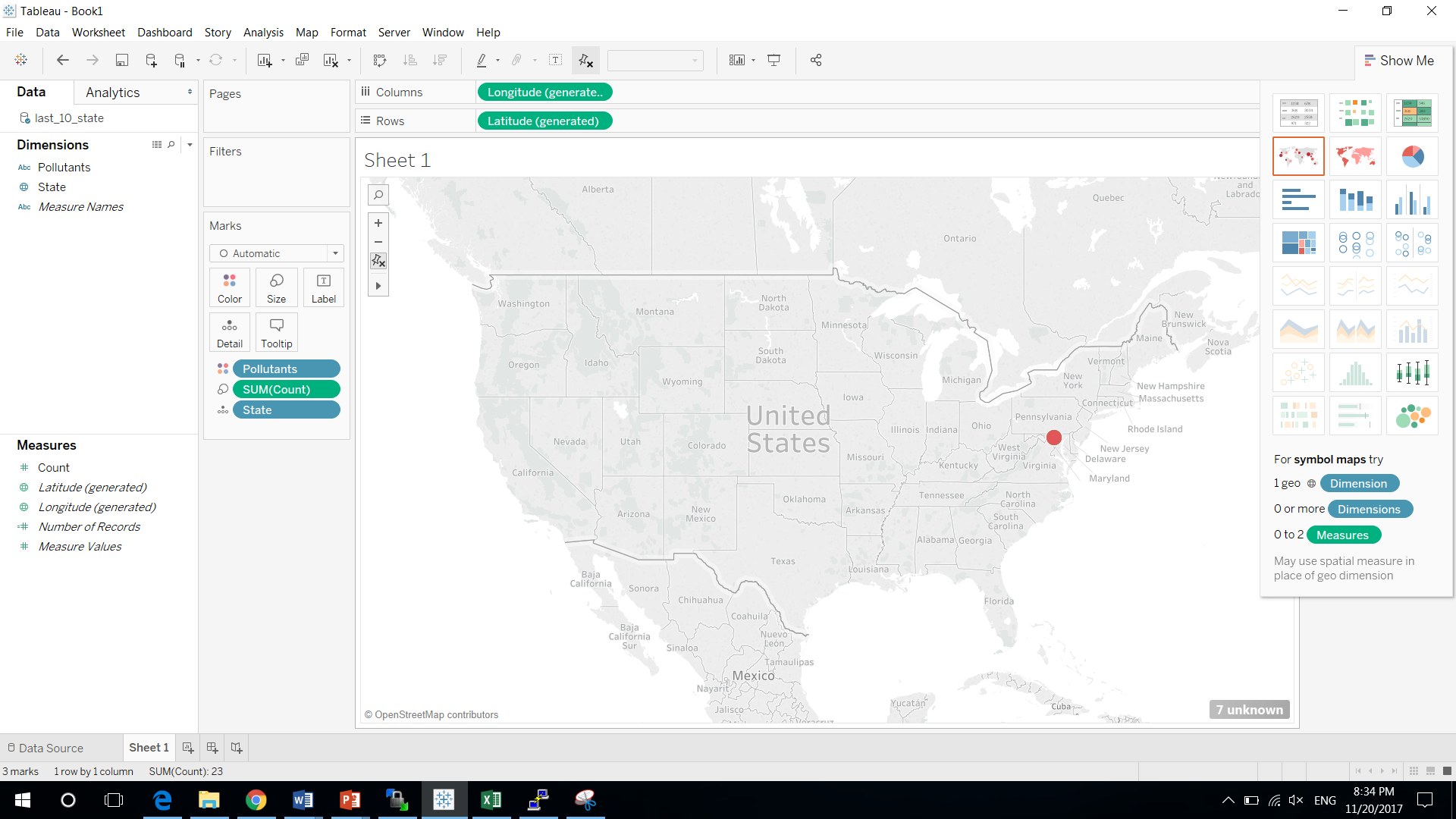
1. Last 10 Pollutants by State



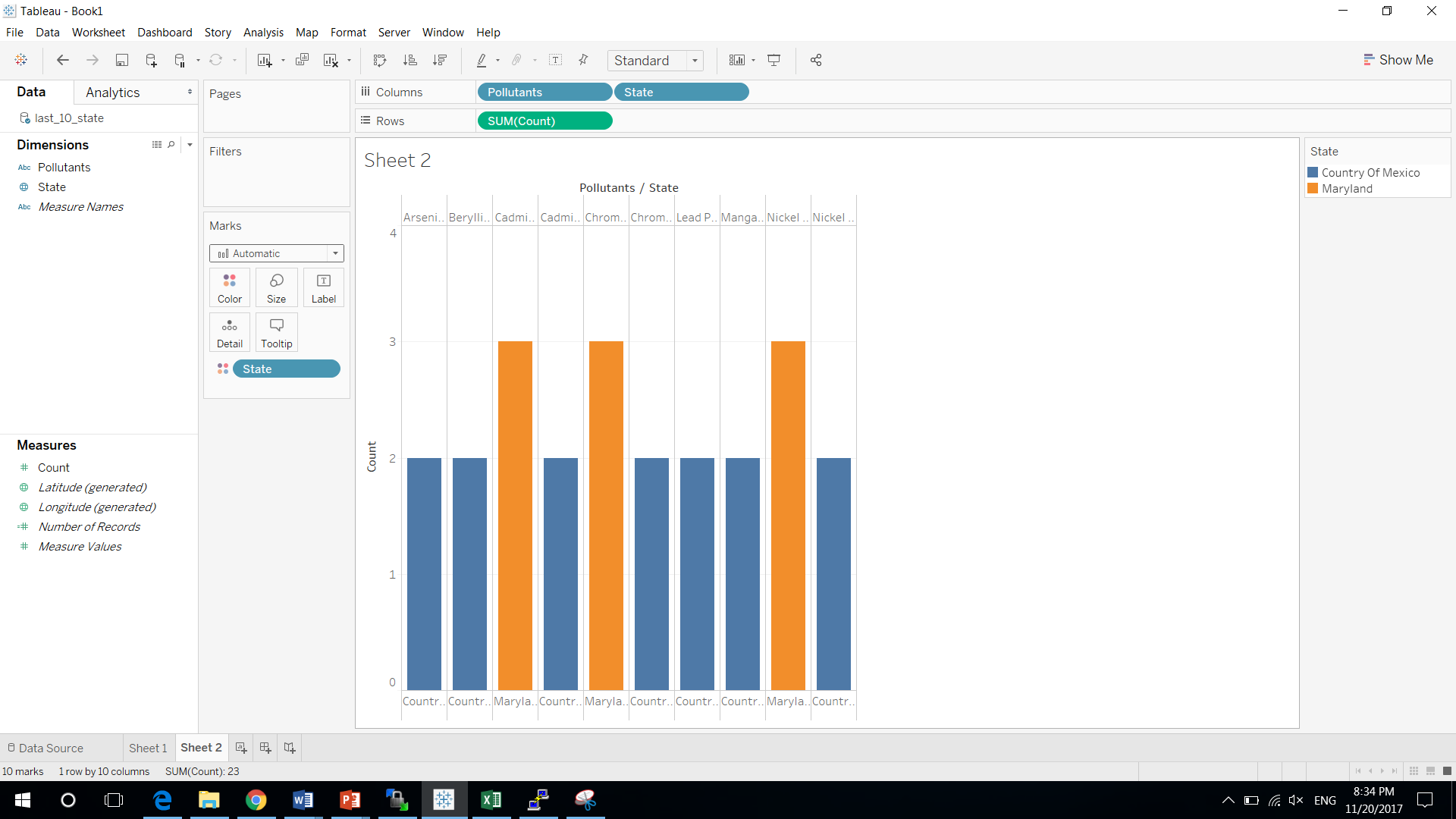
Change F1 to count, F2 to Pollutants, F3 to State



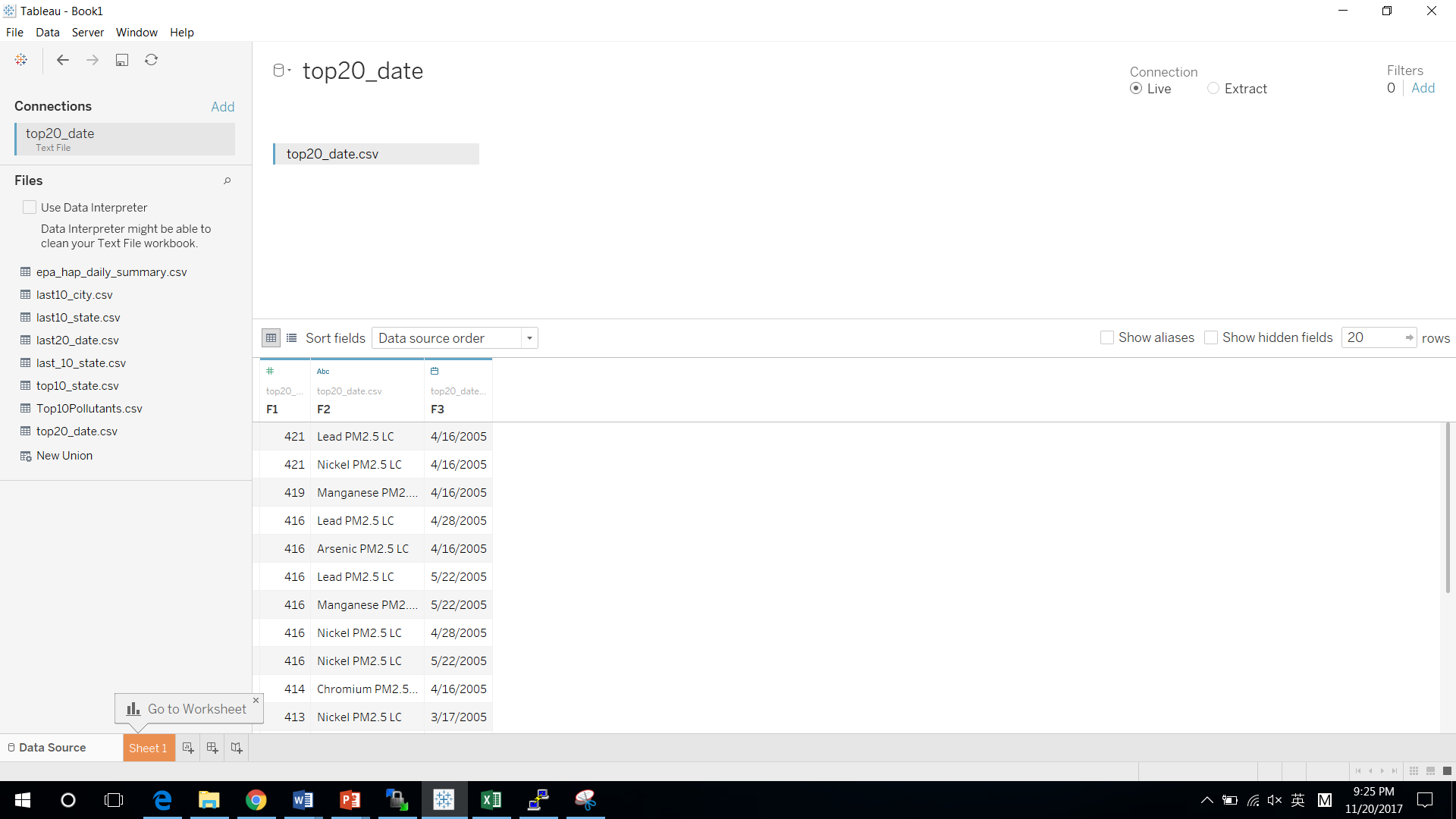
Select Sheet 1 next to Data Source, change State’s geographical role to State/Province. Drag Longitude to Columns, Latitude to Rows, Pollutants to Color, Count to Size, State to Details. And select Geo Map.



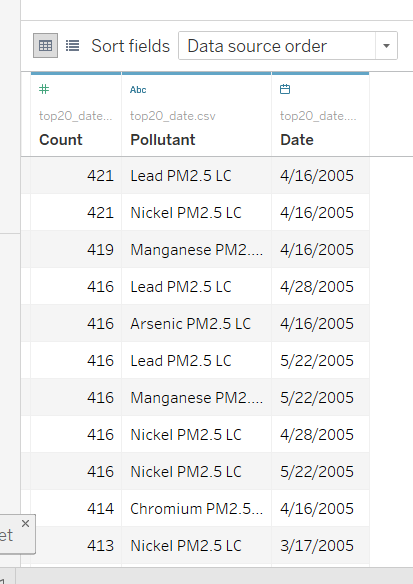
Create a new Worksheet by selecting the icon next to the Sheet 1. Drag Pollutants and State to Columns and Count to Rows. Drag State to Color, the bar chart will only generate two colors due to the top 10 are only in Country of Mexico and Maryland.



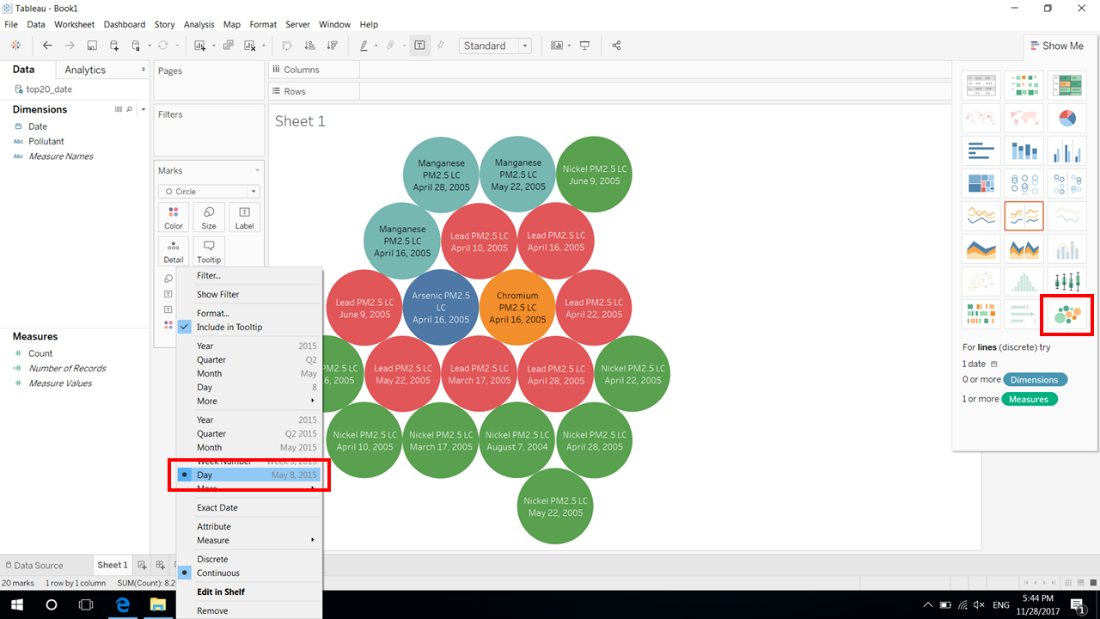
1. Top 20 Pollutants by Date



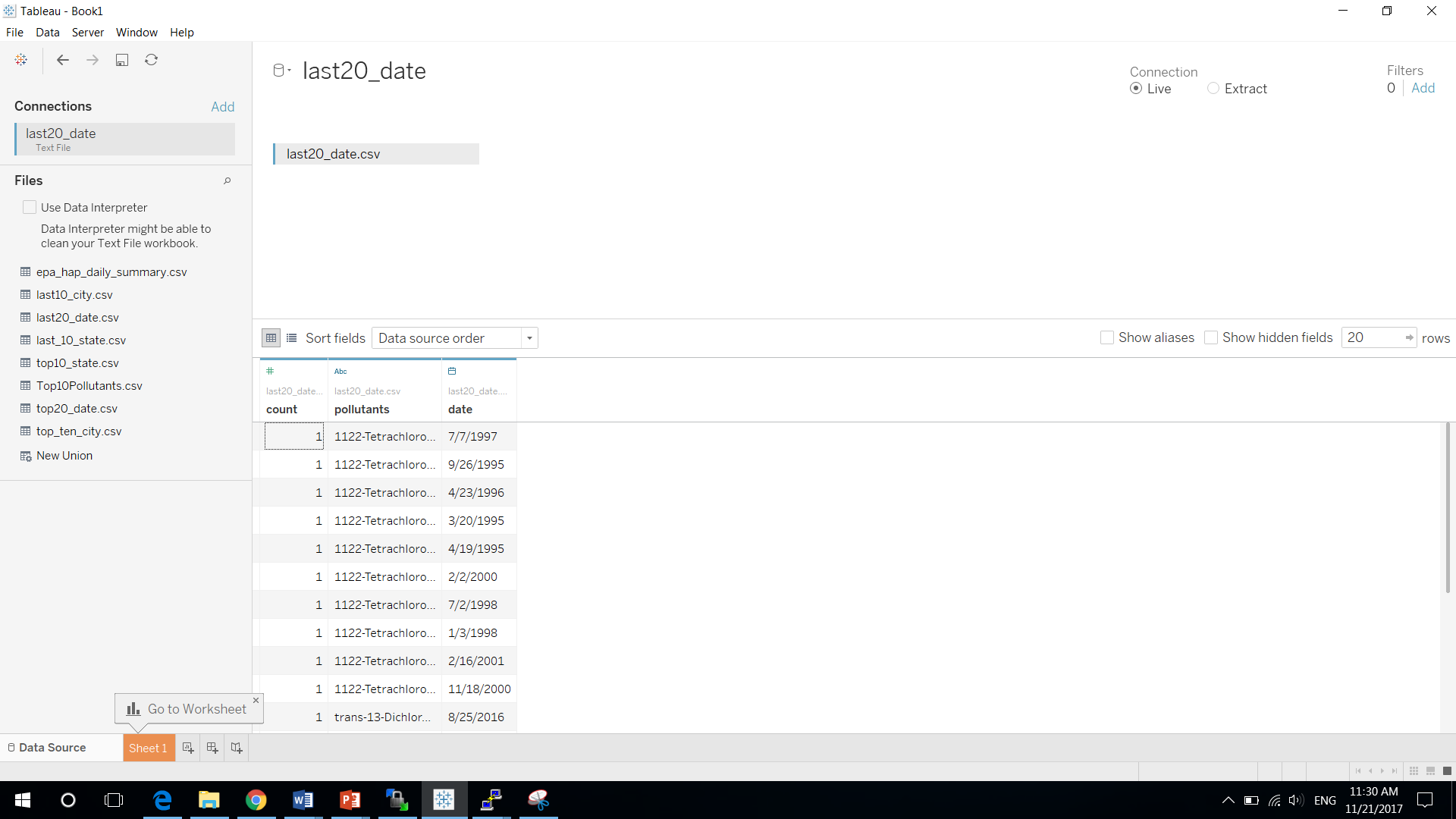
Change F1 to count, F2 to Pollutants, F3 to Date



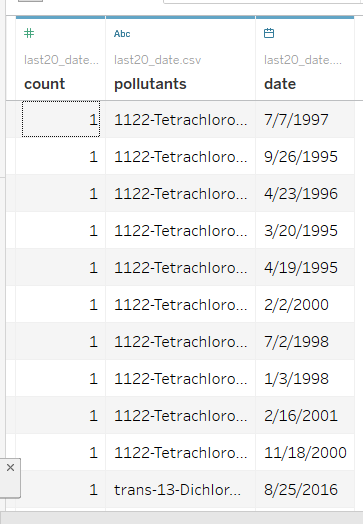
Select Sheet 1 next to Data Source, drag Date to Label, Count to Size, Pollutants to Label and Color. Choose Day (May 8, 2015 format) for date. Choose Packed Bubbles.



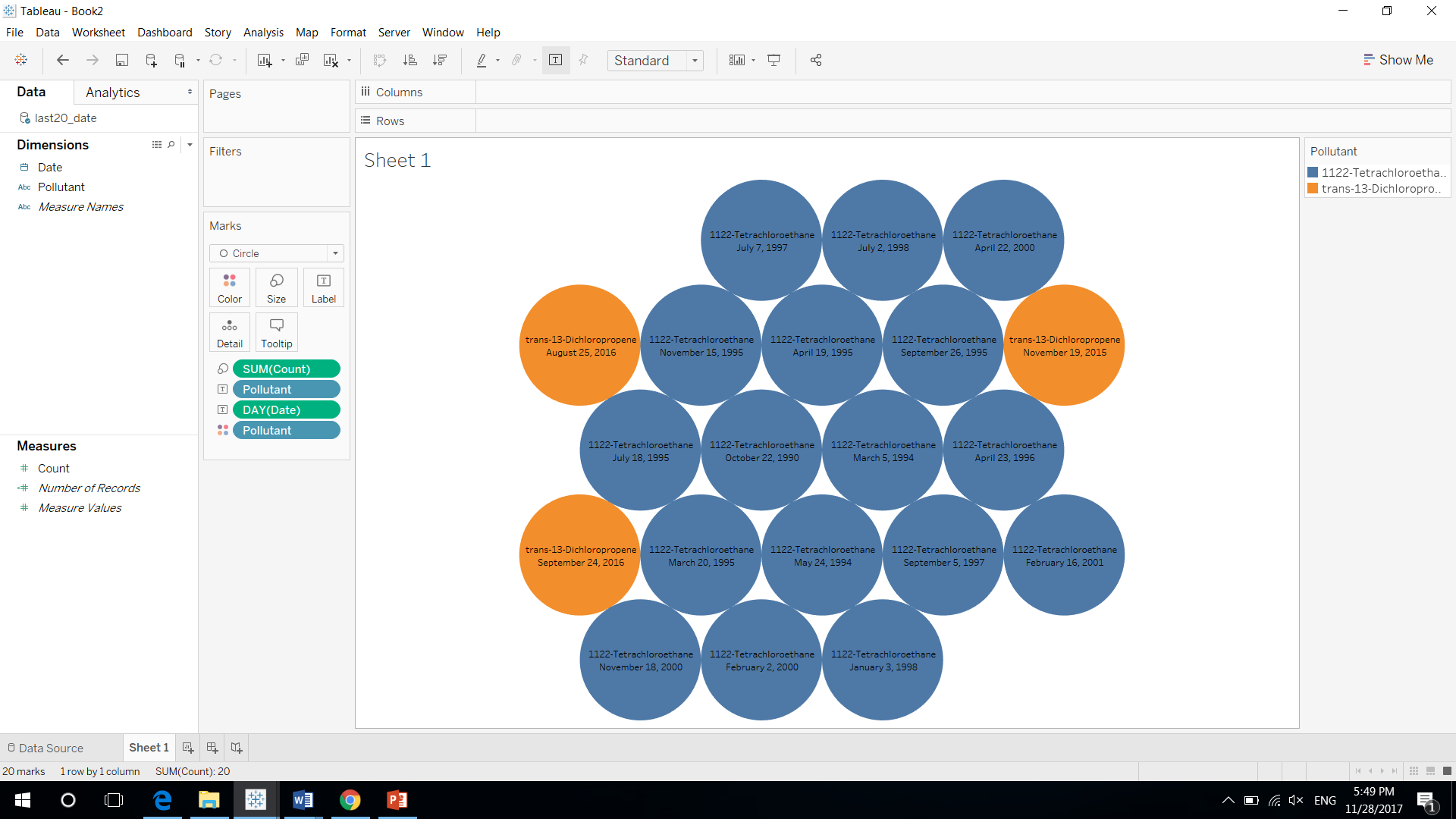
1. Last 20 Pollutants by Date



Change F1 to count, F2 to Pollutants, F3 to Date



Select Sheet 1 next to Data Source, drag Date to Label, Count to Size, Pollutants to Label and Color. Choose Day (May 8, 2015 format) for date. Choose Packed Bubbles.



References:

CIS 5200 Lab – Hive Twitter Sentiment Data Analysis using BigInsights of Bluemix <https://app.box.com/file/96513790564>

CIS 5200 Lab – Analyzing social media and customer sentiment with IBM analytics engine and Tableau <https://app.box.com/file/247447839736>

Hive SQL Syntax Checker <https://sql.treasuredata.com/>