# 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

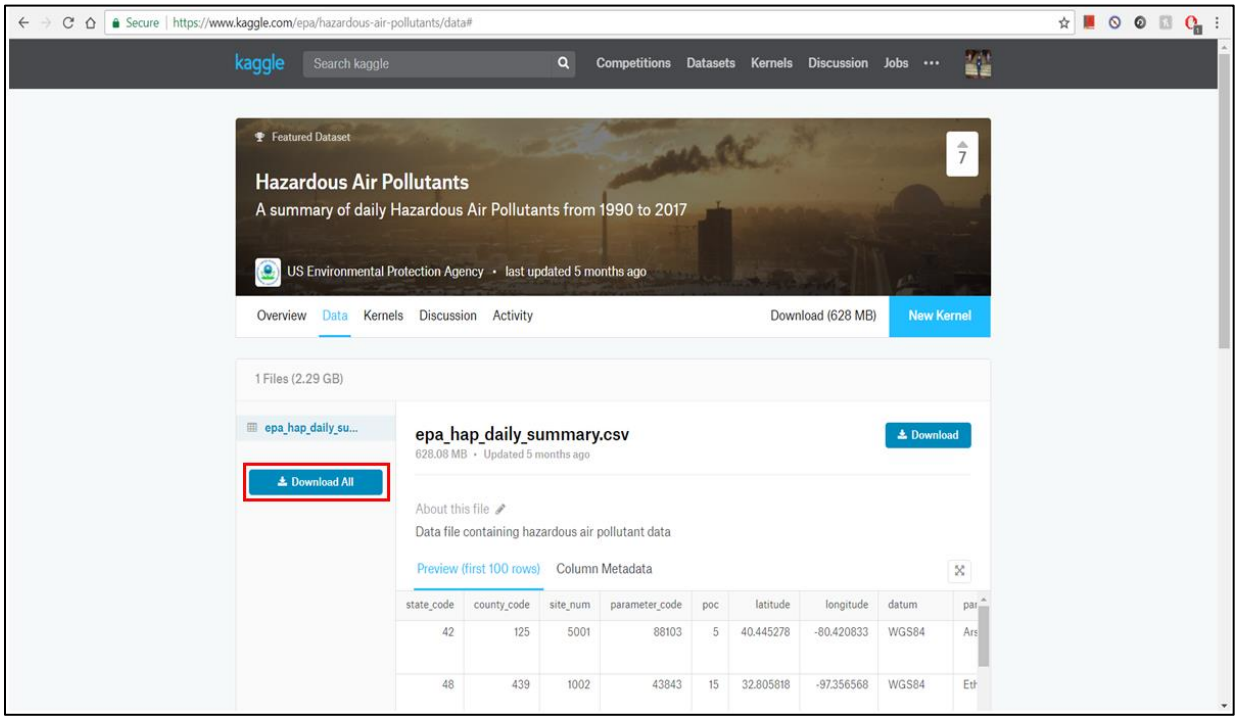
### Prerequisites:

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.

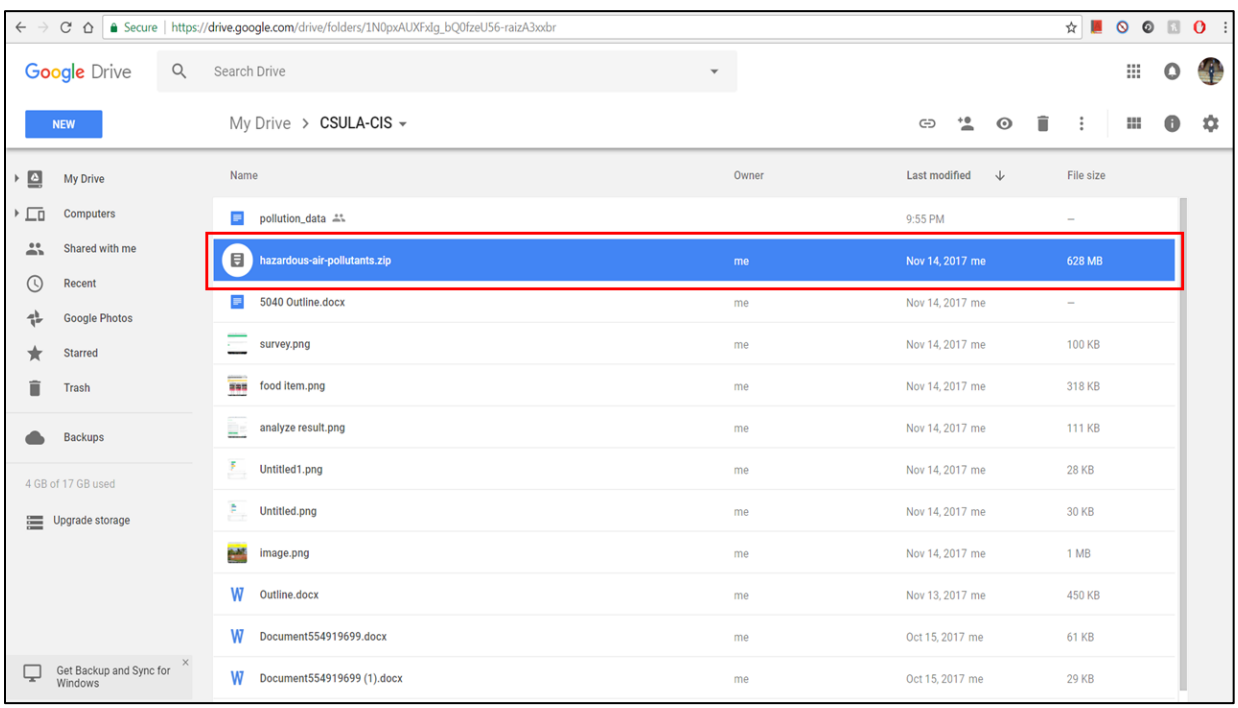
## Air Pollutants data downloaded from Kaggle.com

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
```

```
-bash-4.1$ curl -c /tmp/cookies "https://drive.google.com/uc?export=download&id=
19pPy4pSVarMM6YAAnJjThox2lcThmFxJ"> /tmp/intermezzo.html
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100  3216    0  3216    0     0  17401      0 --:--:-- --:--:-- --:--:-- 42315
-bash-4.1$ curl -L -b /tmp/cookies "https://drive.google.com$(cat /tmp/intermezz
o.html | grep -Po 'uc-download-link" [^>]* href="\K[^"]*' | sed 's/\&amp;/\&/g')
" > FINAL_DOWNLOADED_FILENAME
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100  628M    0  628M    0     0   101M      0 --:--:--  0:00:06 --:--:--  111M
```

```
$ ls-alrt      (check if this file FINAL_DOWNLOADED_FILENAME is available on path)
$ mv FINAL_DOWNLOADED_FILENAME air_pollutant.zip

$ unzip air_pollutant.zip
```

```
-bash-4.1$ ls -alrt
total 653032
-rw-------.  1 sqian2 biusers  10096122 Nov 21  2016 tweetsbi.csv
drwxr-xr-x. 19 root   root         4096 Nov 15 03:30 ..
drwx------.  3 sqian2 biusers      4096 Nov 15 17:49 .pki
-rw-------.  1 sqian2 biusers      1145 Nov 15 21:28 .bash_history
drwxr-xr-x.  3 sqian2 biusers      4096 Nov 15 21:32 .
-rw-------.  1 sqian2 biusers 658586975 Nov 15 21:32 FINAL_DOWNLOADED_FILENAME
-bash-4.1$ mv FINAL_DOWNLOADED_FILENAME air_pollutant.zip
-bash-4.1$ unzip air_pollutant.zip
Archive:  air_pollutant.zip
  inflating: epa_hap_daily_summary.csv
```

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

```
-bash-4.1$ ls -alrt
total 3056992
-rw-------.  1 sqian2 biusers   10096122 Nov 21  2016 tweetsbi.csv
-rw-------.  1 sqian2 biusers 2461649186 Jun 30 18:54 epa_hap_daily_summary.csv
drwxr-xr-x. 19 root   root          4096 Nov 15 03:30 ..
drwx------.  3 sqian2 biusers       4096 Nov 15 17:49 .pki
-rw-------.  1 sqian2 biusers       1145 Nov 15 21:28 .bash_history
-rw-------.  1 sqian2 biusers  658586975 Nov 15 21:32 air_pollutant.zip
drwxr-xr-x.  3 sqian2 biusers       4096 Nov 15 21:33 .
```

## 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;
```

```
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');
OK
Time taken: 0.092 seconds
hive> load data local inpath '/home/sqian2/epa_hap_daily_summary.csv' into table air_pollution;
Loading data to table default.air_pollution
Table default.air_pollution stats: [numFiles=1, totalSize=2461649186]
OK
Time taken: 46.352 seconds
```

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;
```

```
hive> select * from air_pollution limit 10;
OK
42      125     5001    88103   5       40.445278       -80.420833      WGS84   Arsenic PM2.5 LC
24 HOUR         2005-12-30      Micrograms/cubic meter (LC)     None    1       100     0       0
.003    0.0             811     Met OneSASS Teflon - Energy dispersive XRF              HILLMAN
STATE PARK - KINGS CREEK ROAD   Pennsylvania    Washington      Not in a city   Pittsburgh PA  2
015-07-22
48      439     1002    43843   15      32.805818       -97.356568      WGS84   Ethylene dibromi
de      24 HOUR         2013-09-19      Parts per billion Carbon        None    1       100     0
0.0     0.0             175     Passivated Canister - Cryogenic Preconcentration GC/MS  Fort Wor
th Northwest    3317 Ross Ave   Texas   Tarrant Fort Worth      Dallas-Fort Worth-Arlington TX 2
014-03-25
22      127     9000    88128   1       32.057581       -92.435157      WGS84   Lead PM2.5 LC  2
4 HOUR          2001-11-12      Micrograms/cubic meter (LC)     None    1       100     0       0
.00228  0.0             802     IMPROVE Module A with Cyclone Inlet-Teflon Filter 2.2 sq. cm. -
Proton Induced X-Ray Excitation         Sikes   Louisiana       Winn    Not in a city           2
015-07-22
18      89      22      45201   1       41.60668        -87.304729      WGS84   Benzene 1 HOUR 2
016-05-23       Parts per billion Carbon        None    23      96      0       1.4     2.0     1
28      PRECONCENTRATION TRAP - PE 8700;AUTO GC;SUBAMBIENT-DUAL FID     Gary-IITRI/ 1219.5 meter
s east of Tennessee St.- old ammuntion bunker   201 MISSISSIPPI ST. IITRI BUNKER        IndianaL
ake     Gary    Chicago-Naperville-Elgin IL-IN-WI       2017-02-20
6       89      3003    88136   1       40.53999        -121.57646      NAD83   Nickel PM2.5 LC2
4 HOUR          2001-08-02      Micrograms/cubic meter (LC)     None    1       100     0       1
.3E-4   0.0             802     IMPROVE Module A with Cyclone Inlet-Teflon Filter 2.2 sq. cm. -
Proton Induced X-Ray Excitation Lassen Volcanic National Park   MANZANITA LAKE RS LASSEN VOLCANI
C NP    California      Shasta  Not in a city   Redding CA      2015-07-22
26      163     33      82128   1       42.306674       -83.148754      WGS84   Lead PM10 STP  2
4 HOUR          2008-10-03      Micrograms/cubic meter (25 C)   None    1       100     0       0
.0163   0.0             109     Hi Vol SA/GMW 321B - ICP/MS     PROPERTY OWNED BY DEARBORN PUBLI
C SCHOOLS       2842 WYOMING    Michigan        Wayne   Dearborn        Detroit-Warren-Dearborn
MI      2015-07-22
6       83      9000    88128   1       34.733889       -120.008349     WGS84   Lead PM2.5 LC  2
4 HOUR          2015-08-04      Micrograms/cubic meter (LC)     None    1       100     0       0
.0      0.0             800     IMPROVE Module A with Cyclone Inlet-Teflon Filter 2.2 sq. cm. -
X-Ray Fluorescence      San Rafael Wilderness   San Rafael      California      Santa Barbara  N
ot in a city    Santa Maria-Santa Barbara CA    2016-08-30
72      61      1       88136   5       18.425652       -66.115846      WGS84   Nickel PM2.5 LC2
4 HOUR          2002-10-29      Micrograms/cubic meter (LC)     None    1       100     0       0
.0127   0.0             811     Met One SASS Teflon - Energy Dispersive XRF             USGS AND
 WATER RESOURCES BUILDING       Puerto Rico     Guaynabo        Guaynabo        San Juan-Carolin
a-Caguas PR     2015-07-22
29      510     85      43505   6       38.656498       -90.198646      NAD83   Acrolein - Unver
ified   24 HOUR         2006-02-10      Parts per billion Carbon        None    1       100     0
0.0     0.0             101     CANISTER SUBAMBIENT PRESSURE - MULTI DETECTOR GC        Blair St
reet    BLAIR STREET: 3247 Blair Street St. Louis MO 63107      Missouri        St. Louis City S
t. Louis        St. Louis MO-IL 2015-07-22
6       77      1002    12112   8       37.950741       -121.268523     NAD83   Chromium (TSP) S
TP      24 HOUR         1992-10-03      Micrograms/cubic meter (25 C)   None    1       100     0
0.003   0.0             304     LO-VOL-XONTECH 920 or 924- TEFLON - X-RAY FLUORESCENCE  Stockton
-Hazelton       HAZELTON-HD STOCKTON    California      San Joaquin     Stockton        Stockton
-Lodi CA        2013-06-11
Time taken: 0.096 seconds, Fetched: 10 row(s)
```

## 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;
```

```
MapReduce Jobs Launched:
Stage-Stage-1: Map: 10  Reduce: 10   Cumulative CPU: 57.28 sec   HDFS Read: 2464126
621 HDFS Write: 2540 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 3.15 sec   HDFS Read: 9949 HDFS
Write: 222 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 0 seconds 430 msec
OK
600171  Lead PM2.5 LC
600150  Nickel PM2.5 LC
599855  Manganese PM2.5 LC
598372  Chromium PM2.5 LC
598221  Arsenic PM2.5 LC
469375  Benzene
308983  13-Butadiene
245750  Tetrachloroethylene
245517  Chloroform
244835  Dichloromethane
Time taken: 117.333 seconds, Fetched: 10 row(s)
```

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;
```

```
Total MapReduce CPU Time Spent: 1 minutes 31 seconds 360 msec
OK
44229   Benzene Houston-The Woodlands-Sugar Land TX
42392   13-Butadiene    Houston-The Woodlands-Sugar Land TX
24035   Benzene Dallas-Fort Worth-Arlington TX
23352   13-Butadiene    Dallas-Fort Worth-Arlington TX
19080   Benzene New York-Newark-Jersey City NY-NJ-PA
17487   Dichloromethane Houston-The Woodlands-Sugar Land TX
17479   Vinyl chloride  Houston-The Woodlands-Sugar Land TX
17477   Chloroform      Houston-The Woodlands-Sugar Land TX
17476   Trichloroethylene       Houston-The Woodlands-Sugar Land TX
17475   Ethylene dichloride     Houston-The Woodlands-Sugar Land TX
Time taken: 134.577 seconds, Fetched: 10 row(s)
```

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;
```

```
MapReduce Jobs Launched:
Stage-Stage-1: Map: 10  Reduce: 10   Cumulative CPU: 70.1 sec   HDFS Read: 2464133131 HDFS Write: 360813 S
UCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 3.33 sec   HDFS Read: 368482 HDFS Write: 446 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 13 seconds 430 msec
OK
1       Cadmium (TSP) STP       Dallas-Fort Worth-Arlington TX
1       Manganese (TSP) STP     Dallas-Fort Worth-Arlington TX
1       Mercury PM10 STP        Mankato-North Mankato MN
1       Beryllium (TSP) STP     Dallas-Fort Worth-Arlington TX
2       Manganese (TSP) STP     Huntington-Ashland WV-KY-OH
2       Mercury PM10 STP        Fargo ND-MN
2       Chromium (TSP) STP      Baton Rouge LA
2       Beryllium (TSP) STP     Huntington-Ashland WV-KY-OH
2       Nickel (TSP) STP        Huntington-Ashland WV-KY-OH
2       Tetrachloroethylene     Akron OH
Time taken: 104.69 seconds, Fetched: 10 row(s)
```

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;
```

```
MapReduce Jobs Launched:
Stage-Stage-1: Map: 10  Reduce: 10   Cumulative CPU: 62.92 sec   HDFS Read: 2464133191 HDFS Write: 79426 S
UCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 3.78 sec   HDFS Read: 87097 HDFS Write: 294 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 6 seconds 700 msec
OK
121622  Benzene Texas
117431  13-Butadiene    Texas
68313   Lead PM2.5 LC   California
68312   Nickel PM2.5 LC California
68312   Manganese PM2.5 LC      California
66597   Arsenic PM2.5 LC        California
66597   Chromium PM2.5 LC       California
64867   Benzene California
52877   Trichloroethylene       Texas
52853   Chloroform      Texas
Time taken: 102.843 seconds, Fetched: 10 row(s)
```

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;

```
MapReduce Jobs Launched:
Stage-Stage-1: Map: 10  Reduce: 10   Cumulative CPU: 87.03 sec   HDFS Read: 2464133131 HDFS Write: 360813
SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 3.24 sec   HDFS Read: 368482 HDFS Write: 446 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 30 seconds 270 msec
OK
1       Cadmium (TSP) STP       Dallas-Fort Worth-Arlington TX
1       Manganese (TSP) STP     Dallas-Fort Worth-Arlington TX
1       Mercury PM10 STP        Mankato-North Mankato MN
1       Beryllium (TSP) STP     Dallas-Fort Worth-Arlington TX
2       Manganese (TSP) STP     Huntington-Ashland WV-KY-OH
2       Mercury PM10 STP        Fargo ND-MN
2       Chromium (TSP) STP      Baton Rouge LA
2       Beryllium (TSP) STP     Huntington-Ashland WV-KY-OH
2       Nickel (TSP) STP        Huntington-Ashland WV-KY-OH
2       Tetrachloroethylene     Akron OH
Time taken: 111.55 seconds, Fetched: 10 row(s)
```

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;

```
MapReduce Jobs Launched:
Stage-Stage-1: Map: 10  Reduce: 10   Cumulative CPU: 237.87 sec   HDFS Read: 2464133061 HDFS Write: 7824899 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 5.18 sec   HDFS Read: 7832557 HDFS Write: 618 SUCCESS
Total MapReduce CPU Time Spent: 4 minutes 3 seconds 50 msec
OK
421     Lead PM2.5 LC   2005-04-16
421     Nickel PM2.5 LC 2005-04-16
419     Manganese PM2.5 LC      2005-04-16
416     Lead PM2.5 LC   2005-04-28
416     Arsenic PM2.5 LC        2005-04-16
416     Lead PM2.5 LC   2005-05-22
416     Manganese PM2.5 LC      2005-05-22
416     Nickel PM2.5 LC 2005-04-28
416     Nickel PM2.5 LC 2005-05-22
414     Chromium PM2.5 LC       2005-04-16
413     Nickel PM2.5 LC 2005-03-17
413     Nickel PM2.5 LC 2005-04-22
413     Manganese PM2.5 LC      2005-04-28
413     Lead PM2.5 LC   2005-03-17
413     Lead PM2.5 LC   2005-04-22
412     Lead PM2.5 LC   2005-04-10
412     Lead PM2.5 LC   2005-06-09
412     Nickel PM2.5 LC 2005-04-10
412     Nickel PM2.5 LC 2005-06-09
412     Nickel PM2.5 LC 2004-08-07
Time taken: 149.963 seconds, Fetched: 20 row(s)
```

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;

```
MapReduce Jobs Launched:
Stage-Stage-1: Map: 10  Reduce: 10   Cumulative CPU: 211.64 sec   HDFS Read: 2464133061 HDFS Write: 7824899 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 18.12 sec   HDFS Read: 7832557 HDFS Write: 726 SUCCESS
Total MapReduce CPU Time Spent: 3 minutes 49 seconds 760 msec
OK
1       1122-Tetrachloroethane  1997-07-07
1       1122-Tetrachloroethane  1995-09-26
1       1122-Tetrachloroethane  1996-04-23
1       1122-Tetrachloroethane  1995-03-20
1       1122-Tetrachloroethane  1995-04-19
1       1122-Tetrachloroethane  2000-02-02
1       1122-Tetrachloroethane  1998-07-02
1       1122-Tetrachloroethane  1998-01-03
1       1122-Tetrachloroethane  2001-02-16
1       1122-Tetrachloroethane  2000-11-18
1       trans-13-Dichloropropene        2016-08-25
1       1122-Tetrachloroethane  1997-09-05
1       trans-13-Dichloropropene        2015-11-19
1       1122-Tetrachloroethane  1995-11-15
1       1122-Tetrachloroethane  1995-07-18
1       1122-Tetrachloroethane  1994-03-05
1       1122-Tetrachloroethane  1994-05-24
1       1122-Tetrachloroethane  2000-04-22
1       trans-13-Dichloropropene        2016-09-24
1       1122-Tetrachloroethane  1990-10-22
Time taken: 146.862 seconds, Fetched: 20 row(s)
```

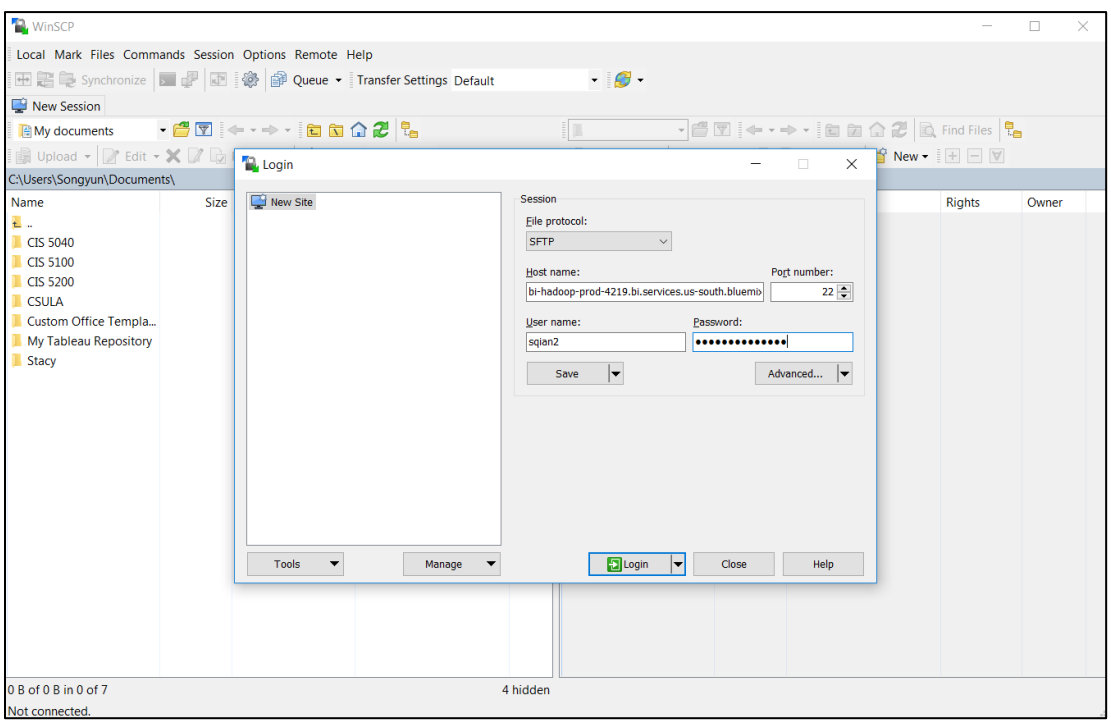
## Create Tables for Tableau

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

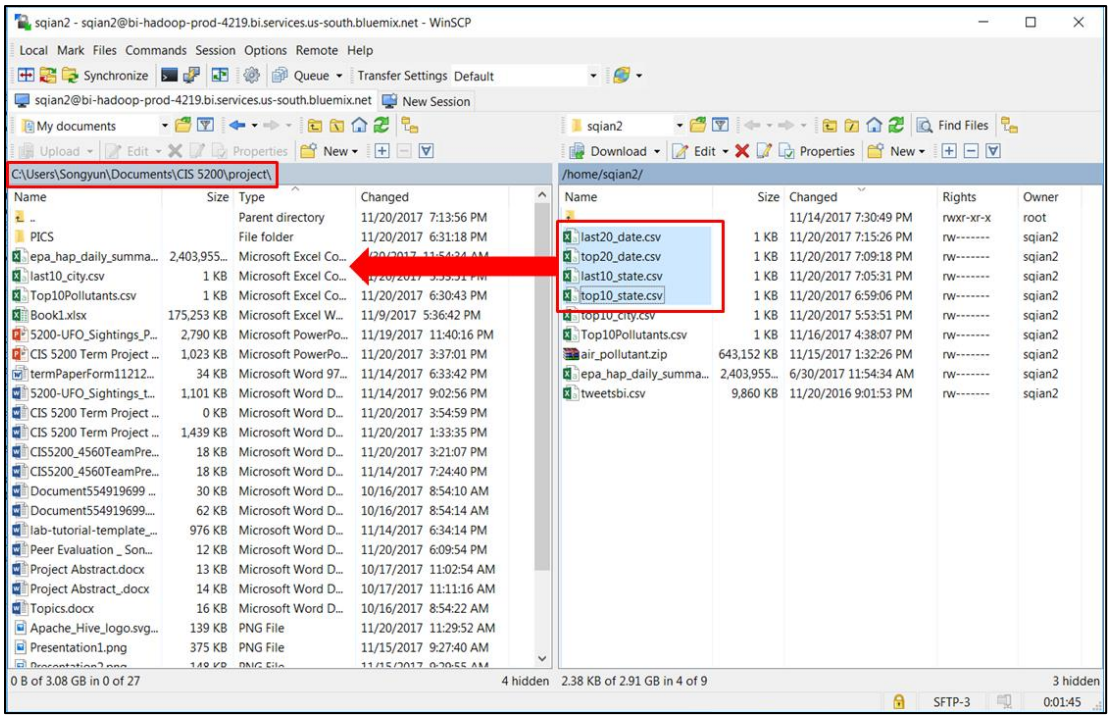
```
hive -e "use default ;Select count(*) as pollution,parameter_name,cbsa_name from air_pollution group by parameter_name,cbsa_name order by pollution ASC limit 10;"| perl -lpe 's/"/\\"/g; s/^|$/"/g; s/\t/","/g' > last10_city.csv
```

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. //written till here.



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.

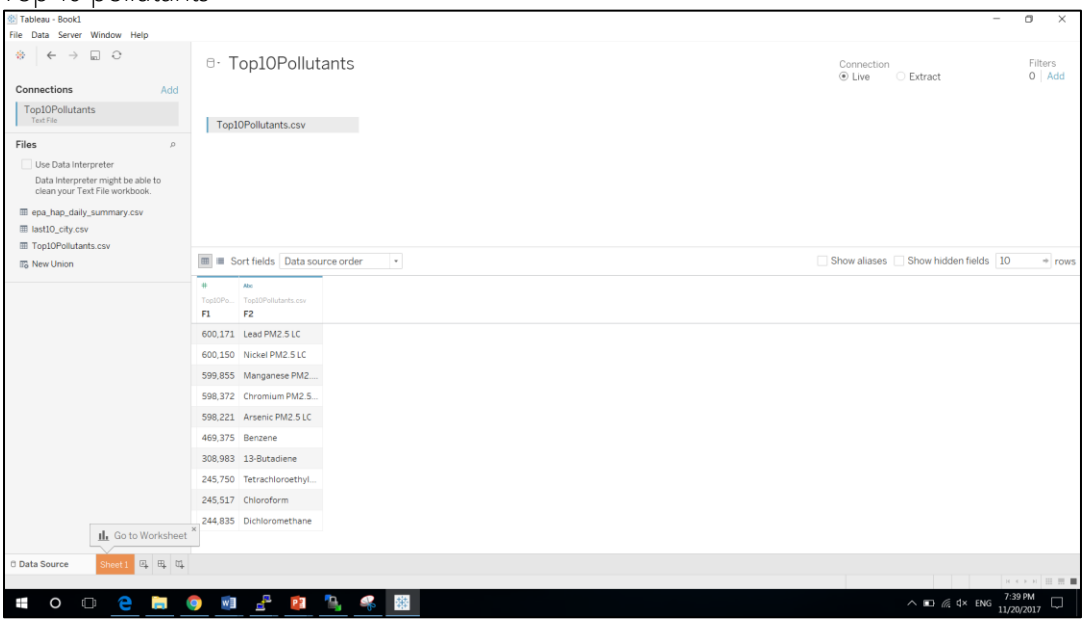



Once the csv files are downloaded, you need to open Tableau on your local computer.

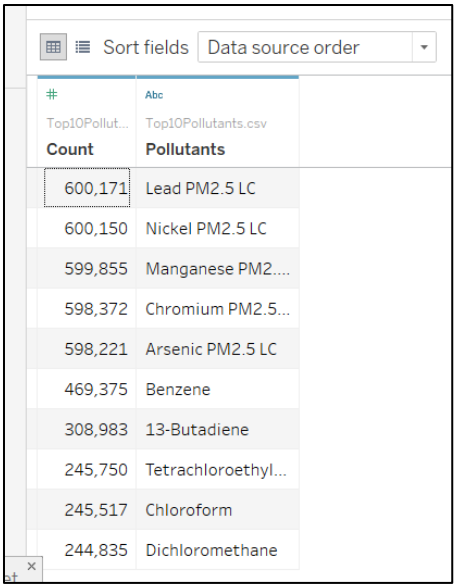
## Tableau to open data file directly from Tableau and Visualization

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

1. Top 10 pollutants



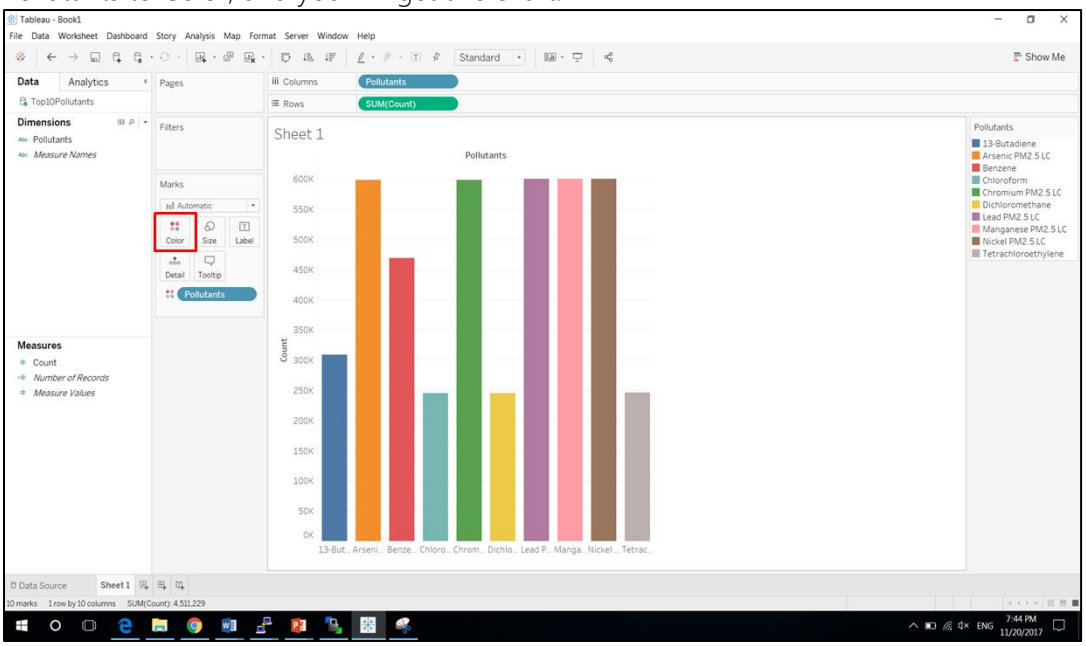

Rename F1 to Count, F2 to Pollutants.



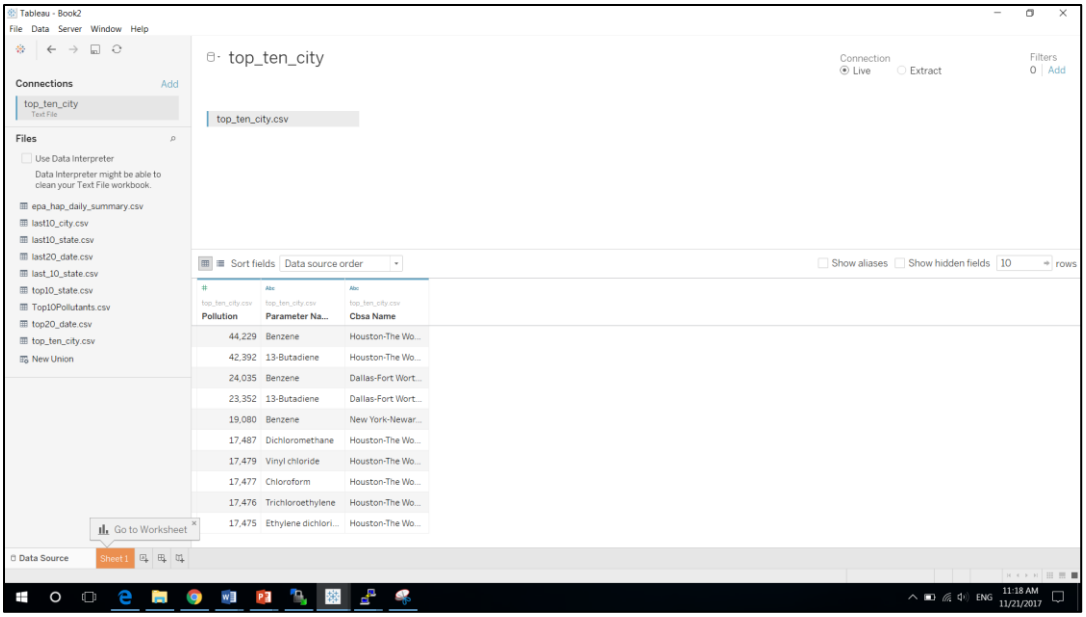
Sort fields Data source order

| Count | Pollutants |
|---|---|
| 600,171 | Lead PM2.5 LC |
| 600,150 | Nickel PM2.5 LC |
| 599,855 | Manganese PM2.... |
| 598,372 | Chromium PM2.5... |
| 598,221 | Arsenic PM2.5 LC |
| 469,375 | Benzene |
| 308,983 | 13-Butadiene |
| 245,750 | Tetrachloroethyl... |
| 245,517 | Chloroform |
| 244,835 | Dichloromethane |

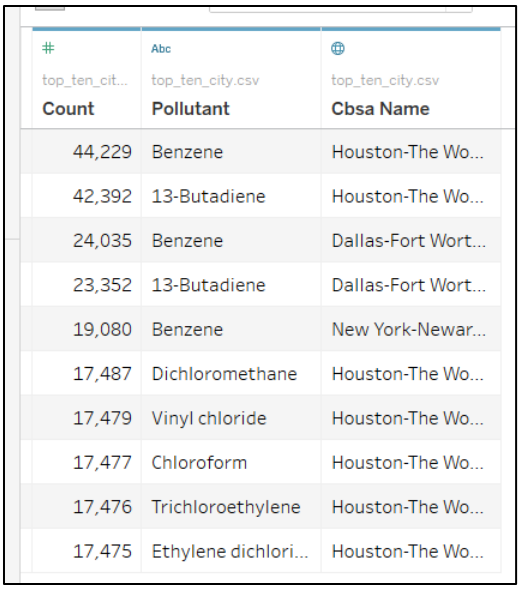
Select Sheet 1 next to Data Source, and drag Count to Rows and Pollutants to Columns. Drag Pollutants to Color, and you will get this chart:




2. Top 10 Pollutants by City

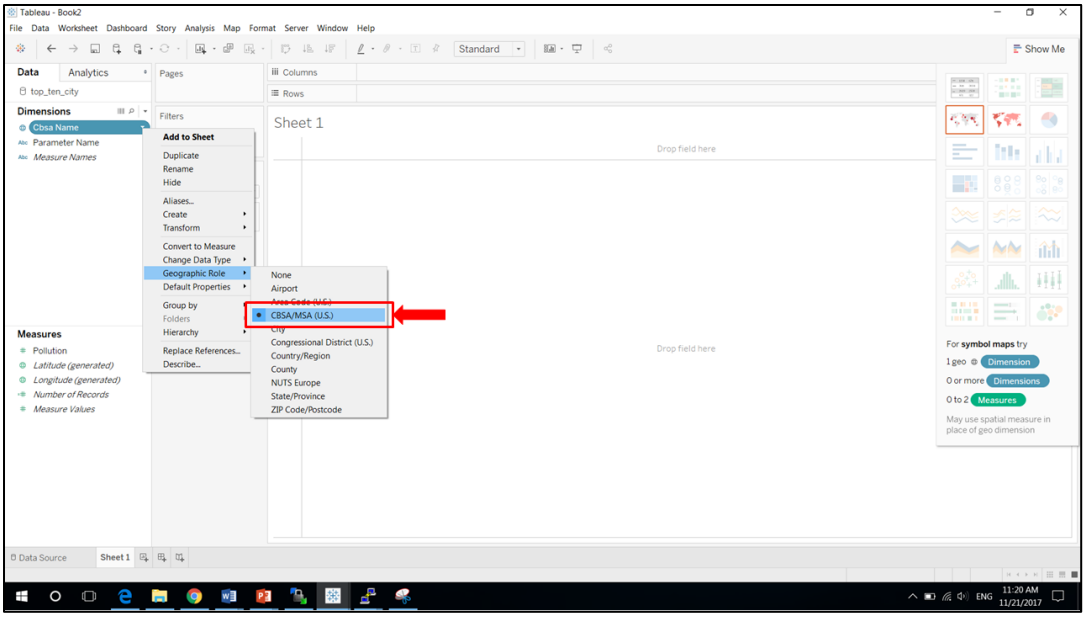



Rename F1 to Count, F2 to Pollutants, F3 to CBSA_NAME.

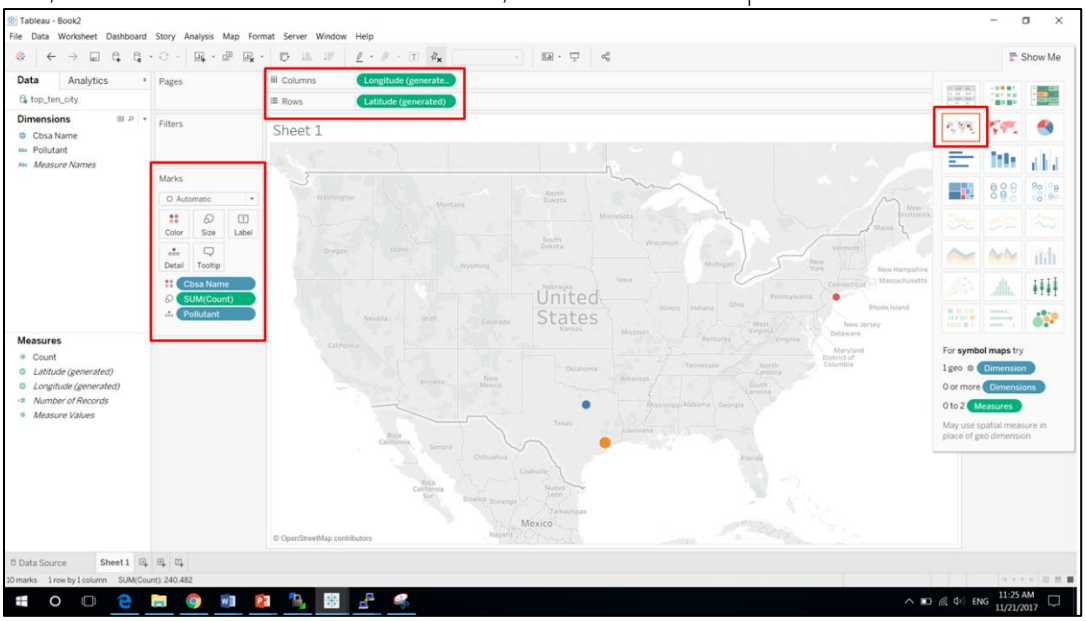


| # top_ten_cit... Count | Abc top_ten_city.csv Pollutant | top_ten_city.csv Cbsa Name |
|---|---|---|
| 44,229 | Benzene | Houston-The Wo... |
| 42,392 | 13-Butadiene | Houston-The Wo... |
| 24,035 | Benzene | Dallas-Fort Wort... |
| 23,352 | 13-Butadiene | Dallas-Fort Wort... |
| 19,080 | Benzene | New York-Newar... |
| 17,487 | Dichloromethane | Houston-The Wo... |
| 17,479 | Vinyl chloride | Houston-The Wo... |
| 17,477 | Chloroform | Houston-The Wo... |
| 17,476 | Trichloroethylene | Houston-The Wo... |
| 17,475 | Ethylene dichlori... | Houston-The Wo... |

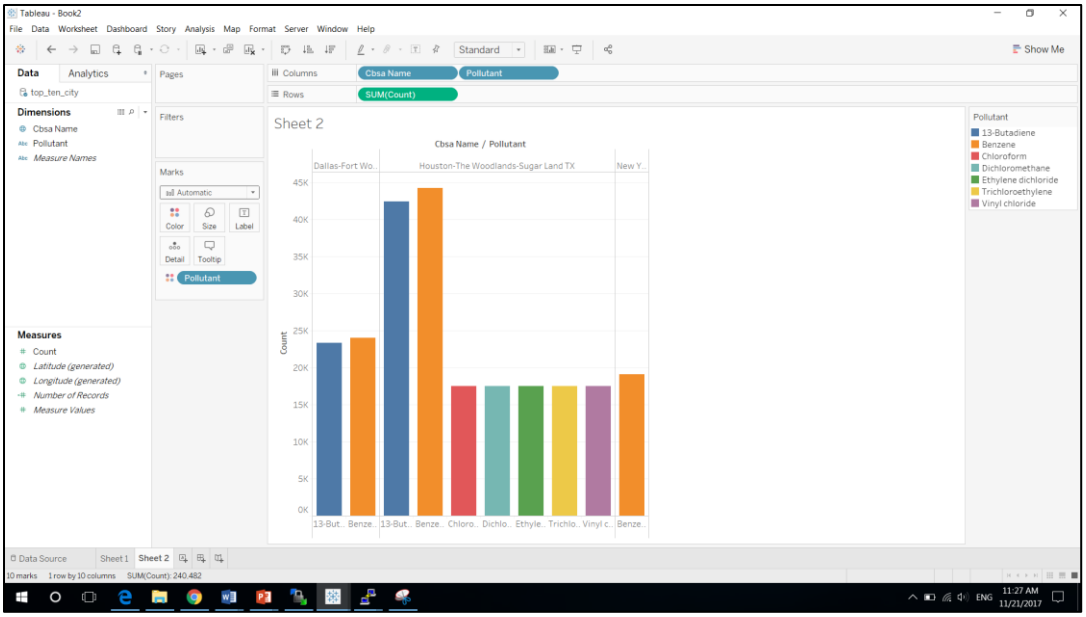
Select Sheet 1 next to Data Source, and change CBSA's geographic role to CBSA/MSA(USA).



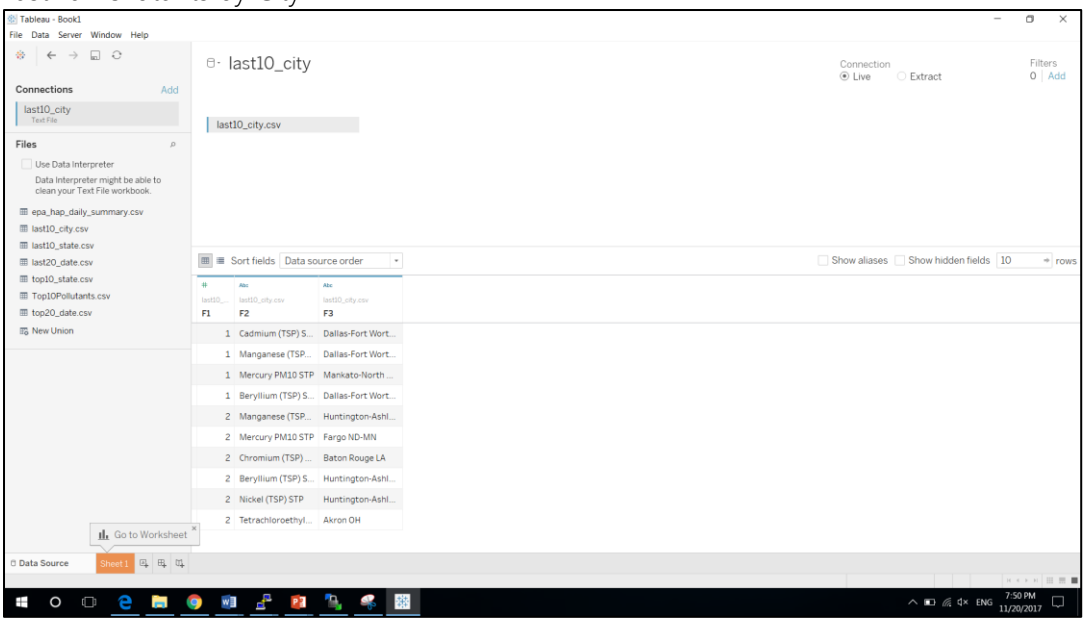

Drag Longitude(generated) to Columns, Latitude(generated) to Rows, CBSA to color, Count to Size, Pollutants to Detail. Select Show me, and select Geo Map:




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:




3. Last 10 Pollutants by City

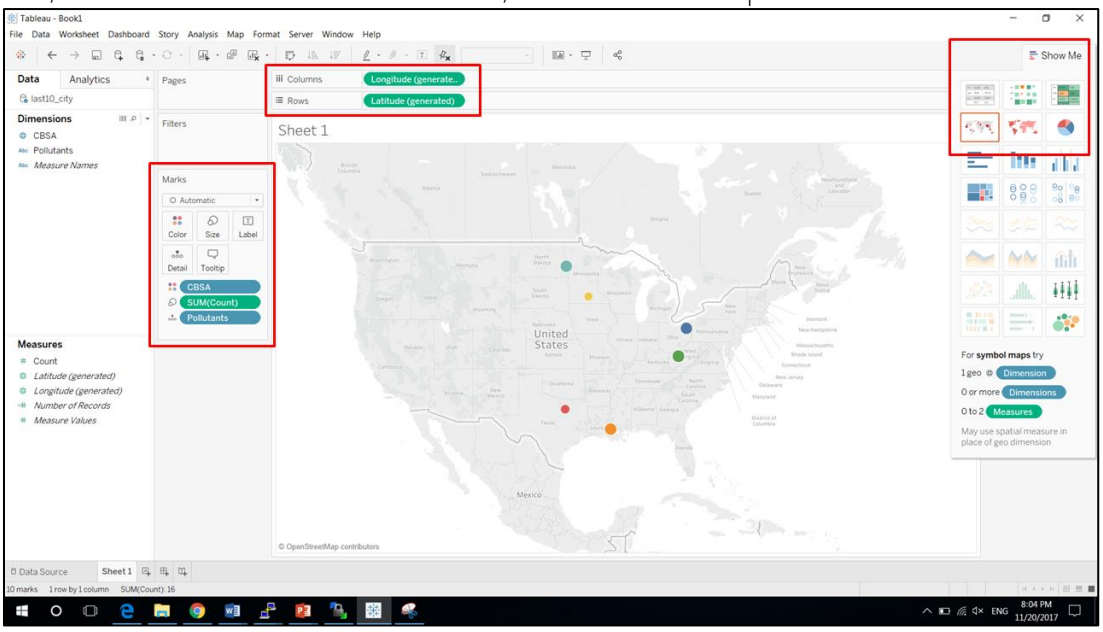



Rename F1 to Count, F2 to Pollutants, F3 to CBSA_NAME.

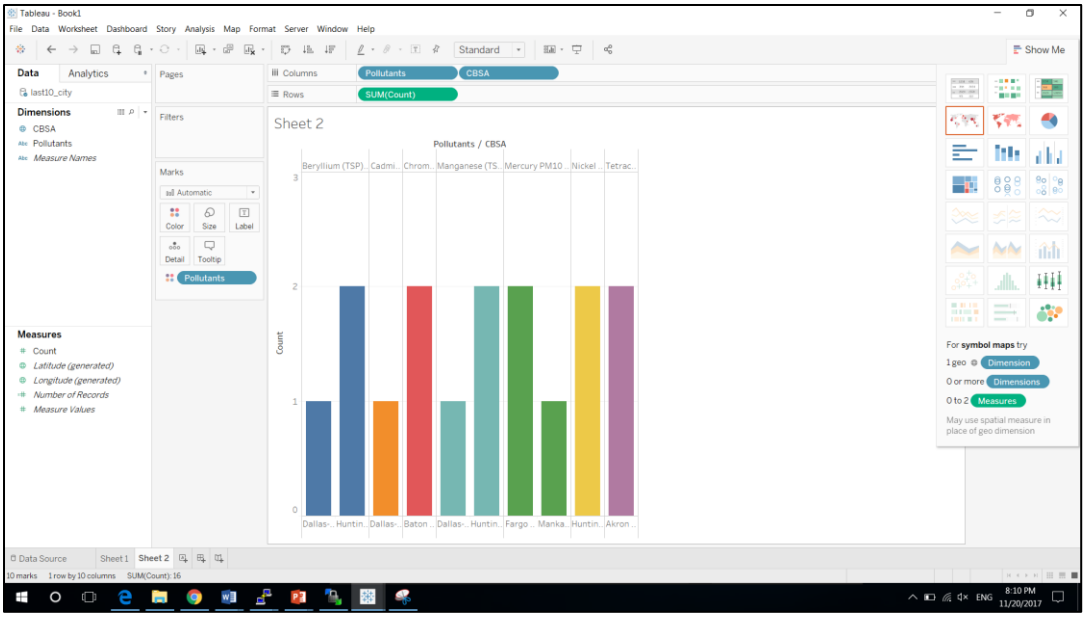
Sort fields Data source order

| # last10_city.... Count | Abc last10_city.csv Pollutants | Abc last10_city.csv CBSA |
|---|---|---|
| 1 | Cadmium (TSP) S... | Dallas-Fort Wort... |
| 1 | Manganese (TSP... | Dallas-Fort Wort... |
| 1 | Mercury PM10 STP | Mankato-North ... |
| 1 | Beryllium (TSP) S... | Dallas-Fort Wort... |
| 2 | Manganese (TSP... | Huntington-Ashl... |
| 2 | Mercury PM10 STP | Fargo ND-MN |
| 2 | Chromium (TSP) ... | Baton Rouge LA |
| 2 | Beryllium (TSP) S... | Huntington-Ashl... |
| 2 | Nickel (TSP) STP | Huntington-Ashl... |
| 2 | Tetrachloroethyl... | Akron OH |

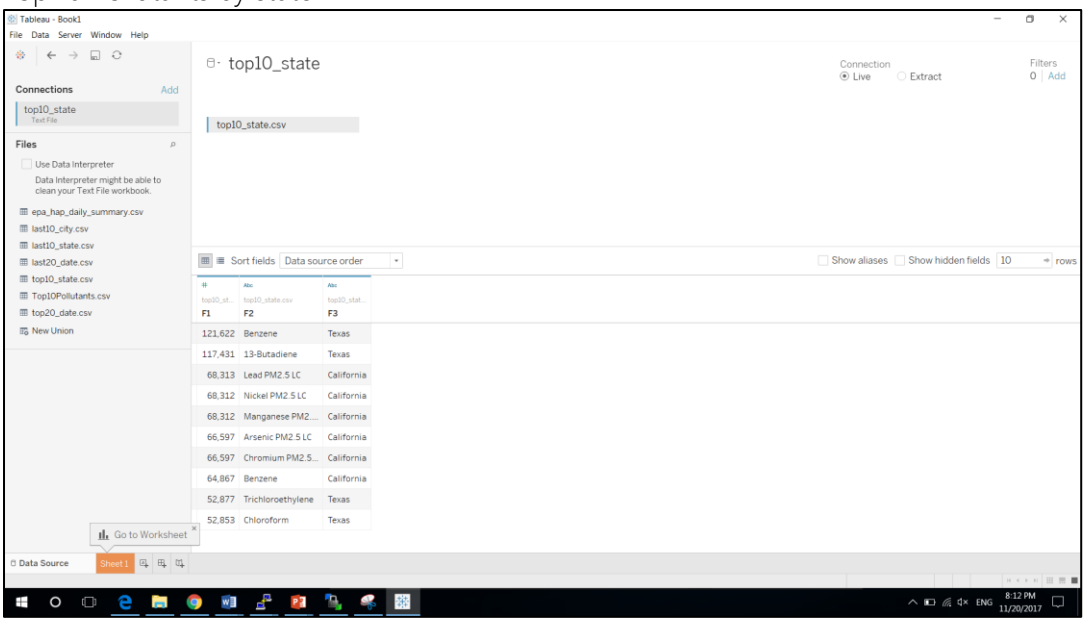
Select Sheet 1 next to Data Source, and change CBSA's geographic role to CBSA/MSA(USA). Drag Longitude(generated) to Columns, Latitude(generated) to Rows, CBSA to color, Count to Size, Pollutants to Detail. Select Show me, and select Geo Map:




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:




4. Top 10 Pollutants by state

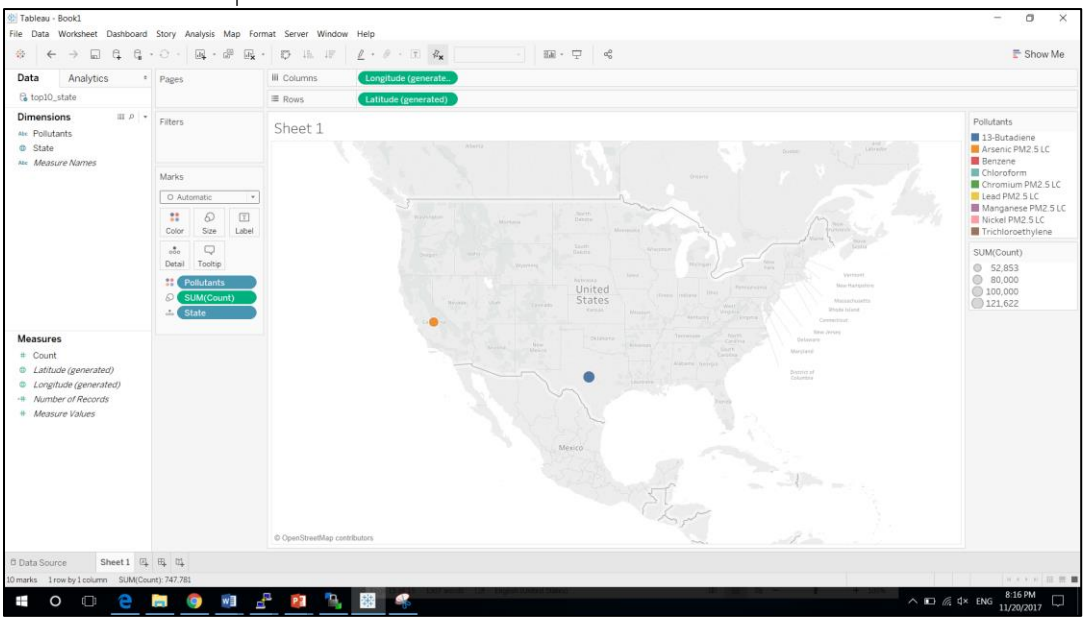



Change F1 to count, F2 to Pollutants, F3 to State

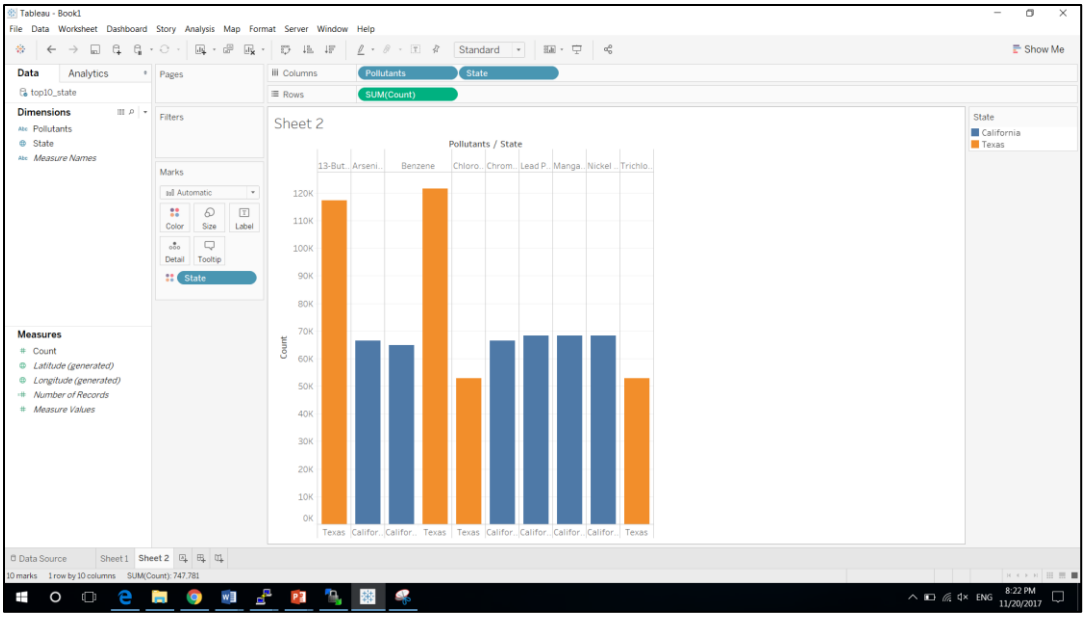
Sort fields Data source order

| # top10_state... Count | Abc top10_state.csv Pollutants | Abc top10_state.csv State |
|---|---|---|
| 121,622 | Benzene | Texas |
| 117,431 | 13-Butadiene | Texas |
| 68,313 | Lead PM2.5 LC | California |
| 68,312 | Nickel PM2.5 LC | California |
| 68,312 | Manganese PM2.... | California |
| 66,597 | Arsenic PM2.5 LC | California |
| 66,597 | Chromium PM2.5... | California |
| 64,867 | Benzene | California |
| 52,877 | Trichloroethylene | Texas |
| 52,853 | Chloroform | Texas |

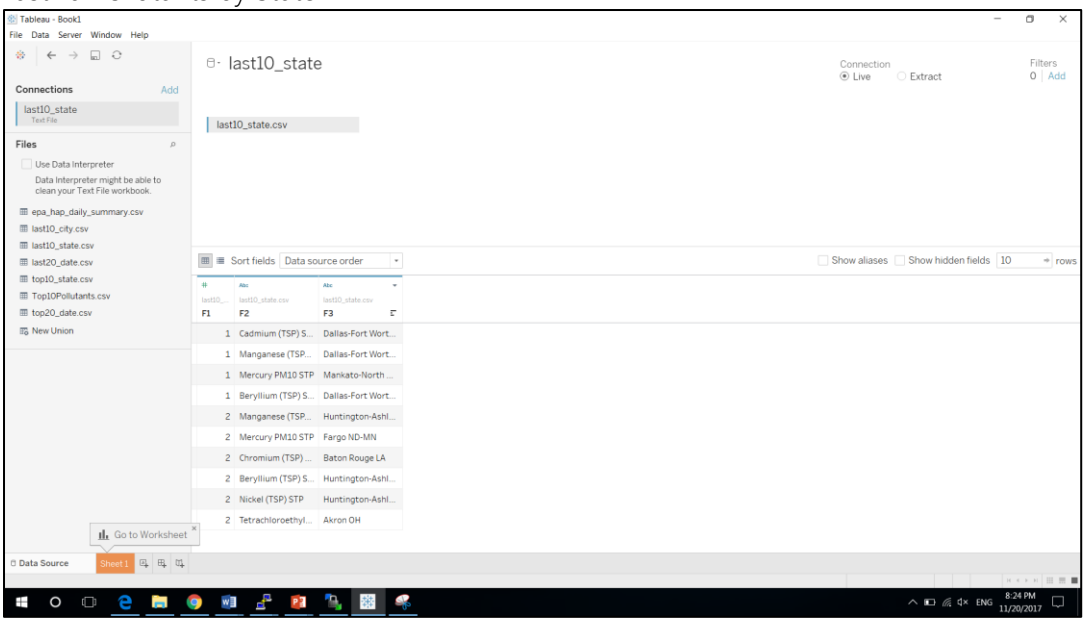
Select Sheet 1 next to Data Source, change State's geographical role to State/Province. Drag Longitude to Columns, Latitude to Rows, Pollutants to Color, Count to Size, State to Details. And select Geo Map.




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.




5. Last 10 Pollutants by State

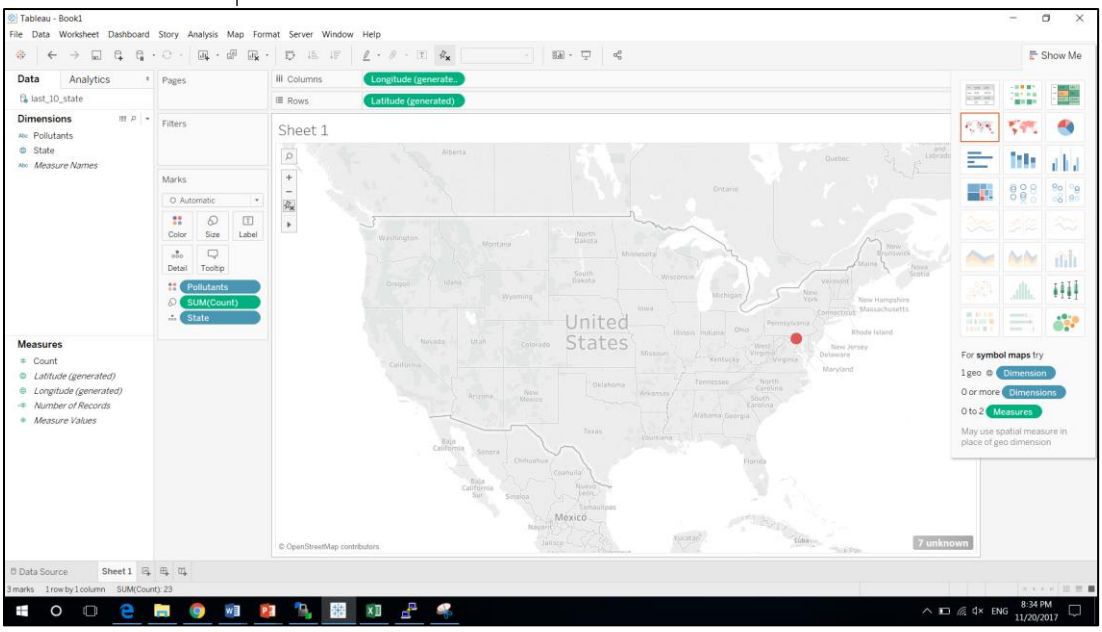



Change F1 to count, F2 to Pollutants, F3 to State

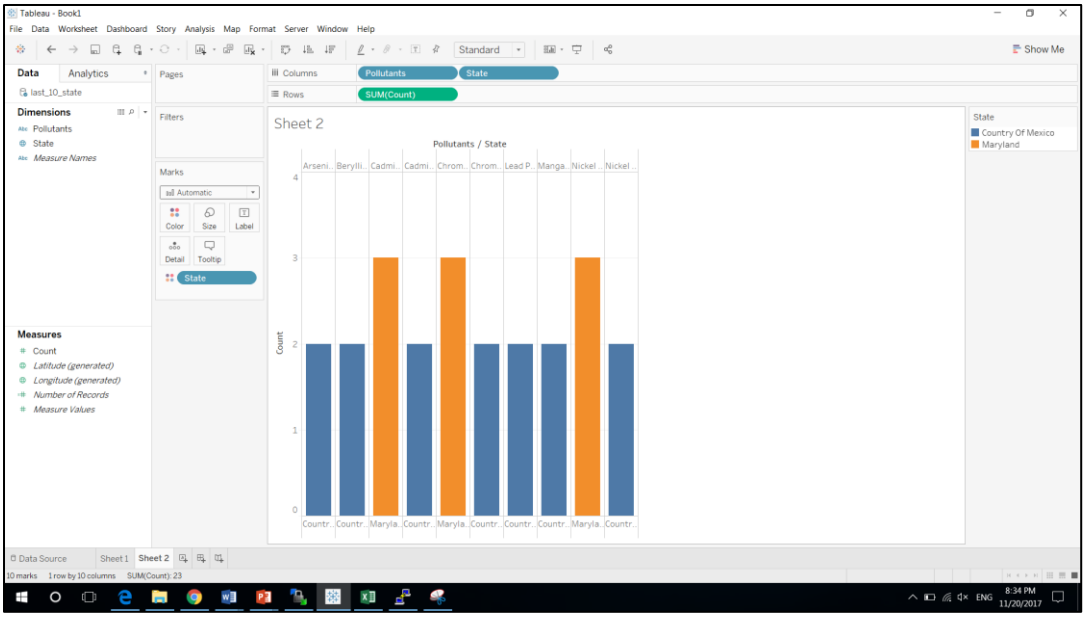
Sort fields Data source order

| last_10_stat... Count | last_10_state.csv Pollutants | last_10_state.csv State |
|---|---|---|
| 2 | Beryllium PM10 ... | Country Of Mexico |
| 2 | Chromium PM10 ... | Country Of Mexico |
| 2 | Cadmium PM10 ... | Country Of Mexico |
| 2 | Manganese PM1... | Country Of Mexico |
| 2 | Nickel PM10 STP | Country Of Mexico |
| 2 | Arsenic PM10 STP | Country Of Mexico |
| 2 | Lead PM10 STP | Country Of Mexico |
| 3 | Chromium (TSP) ... | Maryland |
| 3 | Cadmium (TSP) S... | Maryland |
| 3 | Nickel (TSP) STP | Maryland |

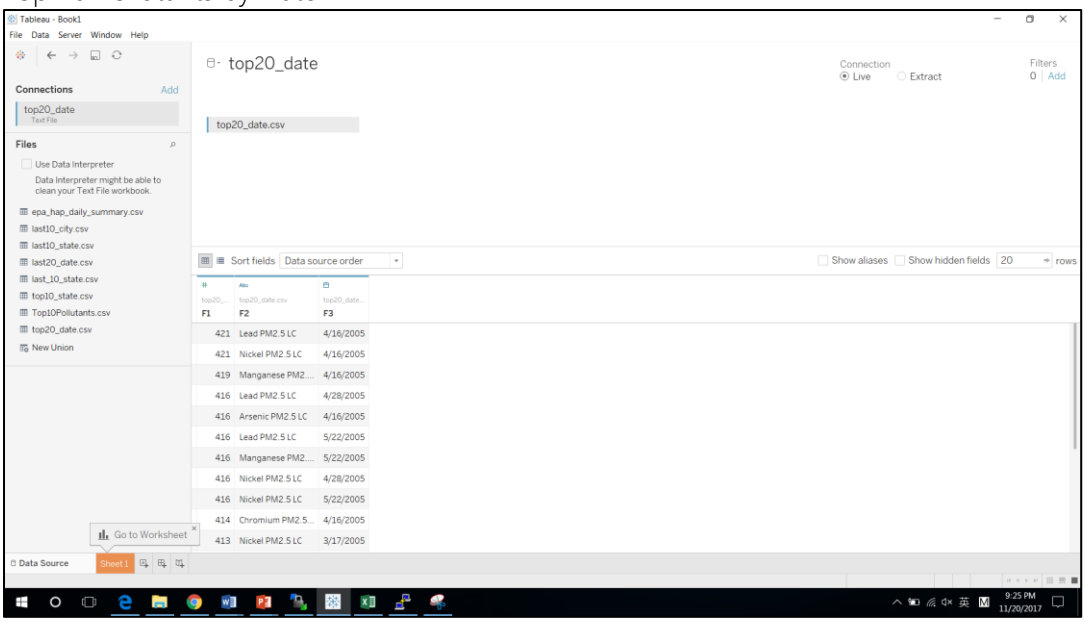
Select Sheet 1 next to Data Source, change State's geographical role to State/Province. Drag Longitude to Columns, Latitude to Rows, Pollutants to Color, Count to Size, State to Details. And select Geo Map.



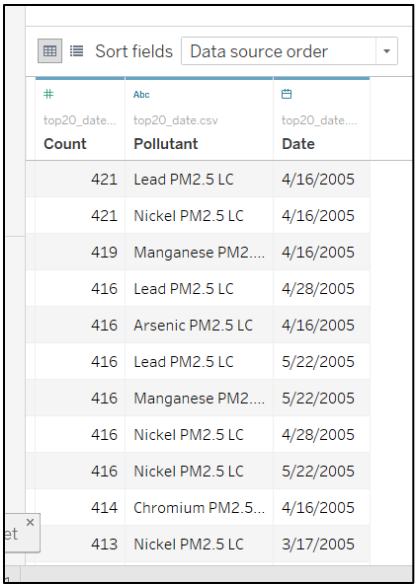

Create a new Worksheet by selecting the icon next to the Sheet 1. Drag Pollutants and State to Columns and Count to Rows. Drag State to Color, the bar chart will only generate two colors due to the top 10 are only in Country of Mexico and Maryland.




6. Top 20 Pollutants by Date



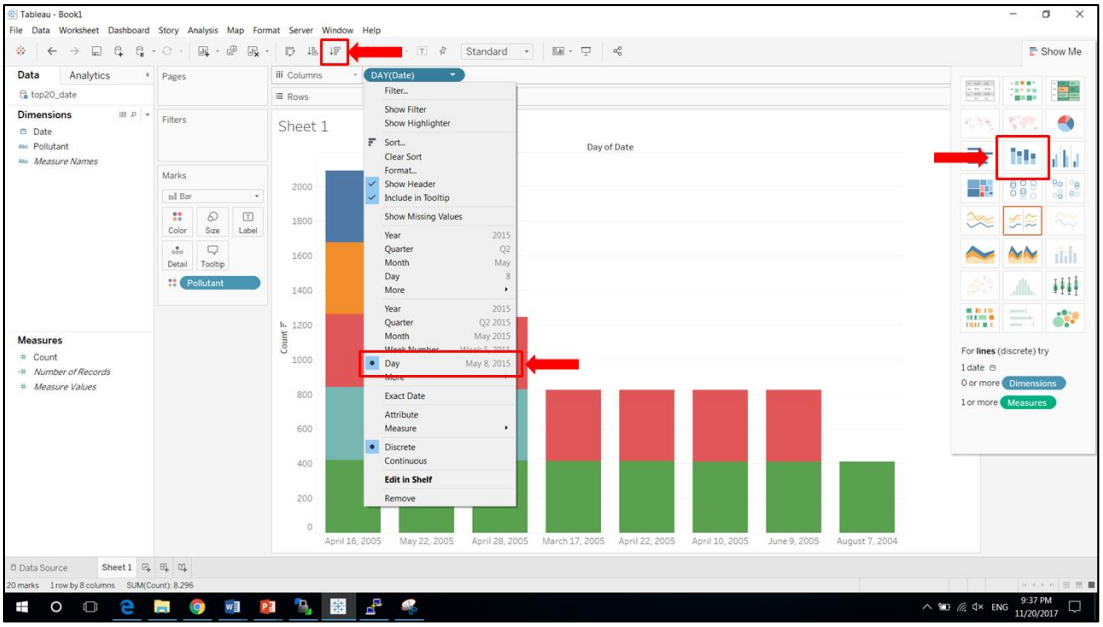

Change F1 to count, F2 to Pollutants, F3 to Date



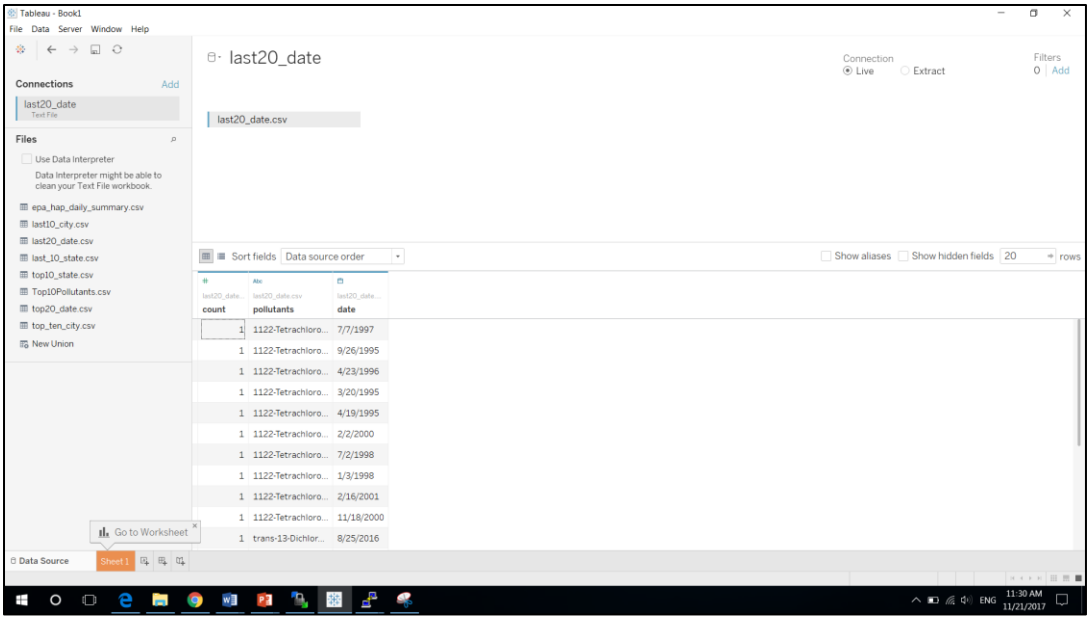
Sort fields Data source order

| Count | Pollutant | Date |
|---|---|---|
| 421 | Lead PM2.5 LC | 4/16/2005 |
| 421 | Nickel PM2.5 LC | 4/16/2005 |
| 419 | Manganese PM2.... | 4/16/2005 |
| 416 | Lead PM2.5 LC | 4/28/2005 |
| 416 | Arsenic PM2.5 LC | 4/16/2005 |
| 416 | Lead PM2.5 LC | 5/22/2005 |
| 416 | Manganese PM2.... | 5/22/2005 |
| 416 | Nickel PM2.5 LC | 4/28/2005 |
| 416 | Nickel PM2.5 LC | 5/22/2005 |
| 414 | Chromium PM2.5... | 4/16/2005 |
| 413 | Nickel PM2.5 LC | 3/17/2005 |

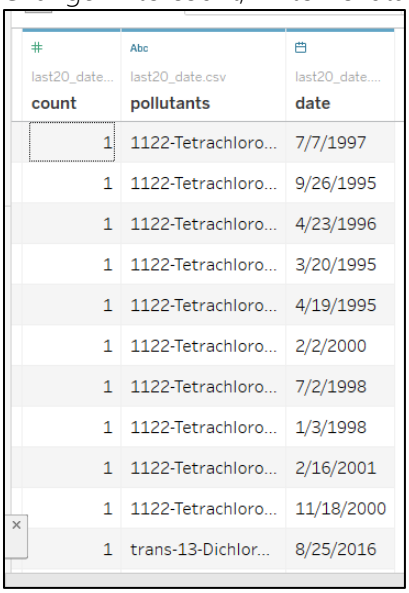
Select Sheet 1 next to Data Source, drag Date to Columns, Count to Rows, Pollutants to Color. Choose Day (May 8, 2015 format) for date, Sort Day of Date descending by Count, Choose stack bar.




7. Last 20 Pollutants by Date

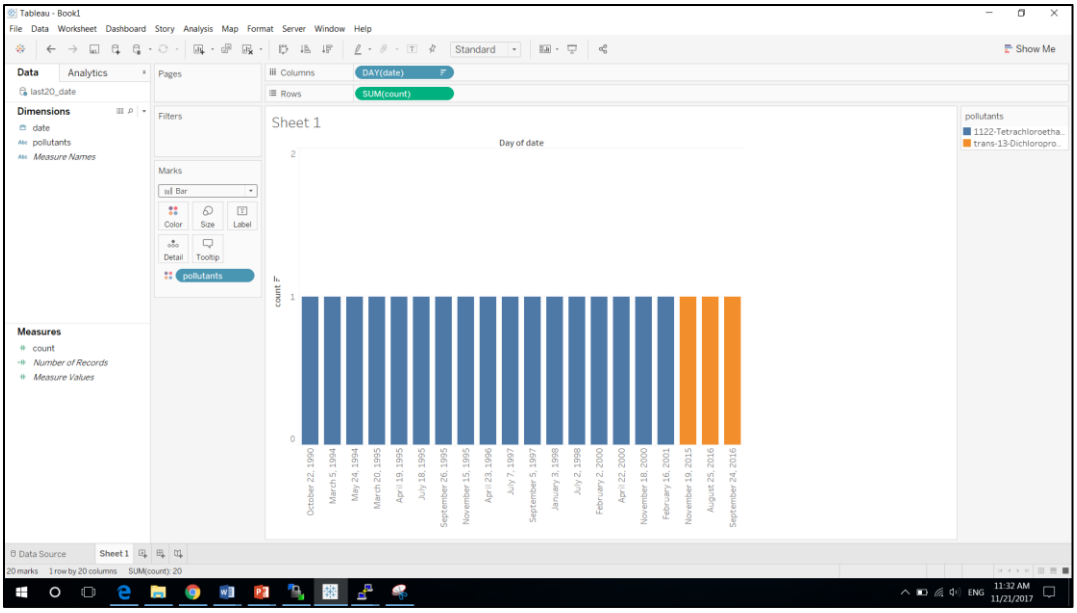



Change F1 to count, F2 to Pollutants, F3 to Date



| # last20_date... count | Abc last20_date.csv pollutants | last20_date.... date |
|---|---|---|
| 1 | 1122-Tetrachloro... | 7/7/1997 |
| 1 | 1122-Tetrachloro... | 9/26/1995 |
| 1 | 1122-Tetrachloro... | 4/23/1996 |
| 1 | 1122-Tetrachloro... | 3/20/1995 |
| 1 | 1122-Tetrachloro... | 4/19/1995 |
| 1 | 1122-Tetrachloro... | 2/2/2000 |
| 1 | 1122-Tetrachloro... | 7/2/1998 |
| 1 | 1122-Tetrachloro... | 1/3/1998 |
| 1 | 1122-Tetrachloro... | 2/16/2001 |
| 1 | 1122-Tetrachloro... | 11/18/2000 |
| 1 | trans-13-Dichlor... | 8/25/2016 |

Select Sheet 1 next to Data Source, drag Date to Columns, Count to Rows, Pollutants to Color. Choose Day (May 8, 2015 format) for date, Sort Day of Date descending by Count, Choose stack bar.




## 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

Classmate: Neha Gupta

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