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

Krishna Ghorpade

Songyun Qian

Himani Patel

Shahnawaz Khan

CIS 5200-01

Group 6

**Abstract:** This paper aims at performing data analysis in Hive using IBM BigInsights for the Hazardous Air Pollutants in USA from 1990 to 2017 along with transfering data with WINSCP and visualization with Tableau. From the data which was available we were able to categorize the air pollutants by different parameters like which pollutant was measured the most followed by city, state, date etc. The data set is a daily summary file, containing data for every monitor in the EPA database.

**1. 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 [1] . The data set is a daily summary file, containing data for every monitor in the EPA database. We analyzed the dataset to determine:

1. Which air pollutants are measured the most
2. Which cities have the highest and lowest air pollutants measured
3. Which states have the highest and lowest air pollutants measured
4. Which dates have the highest and lowest air pollutants measured

Then we transfered data using **WINSCP**, and **v**isualized the data in **Tableau**.

2. General Instructions

In this project, we have analyzed and visualized Air Pollution Data. We have downloaded Air Pollutants 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 [2] [3]. You will also practice how to visualize the result in Tableau.

**3. The Project Flow**

For this analysis, we extracted the dataset from [www.kaggle.com](http://www.kaggle.com) first, then uploaded it to Google Drive, so that we can generate a link to load it into BigInslights. After the dataset is loaded, we ran commands in hive, and analyzed the data. Lastly, we transferred the data, and used Tableau to generate visualization.

**4. Data Analysis**

After data extraction, running queries and fixing commands, we got the following results by using WINSCP to transfer data to our local computer.

**4.1 What are the top 10 pollutants recorded from 1990 to 2017?**

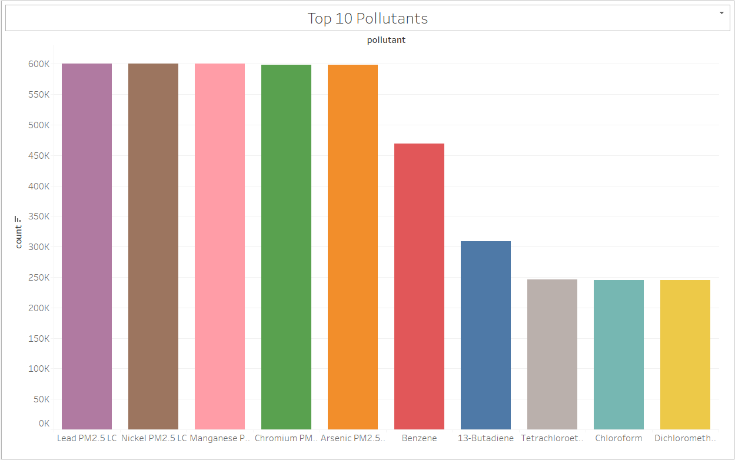


Figure 1. Top 10 Pollutants – Chart

After the analyzation, and using Tableau, this bar chart is showing the top 10 pollutants recorded from 1990 to 2017. With Lead being the most pollutant being recorded, followed by Nickel, Manganese, Chromium, Arsenic, Benzene, 13-Butadiene, Tetrachloroethylene, Chloroform, and Dichloromethane.

**4.2 What are the top 10 pollutants by city recorded from 1990 to 2017?**

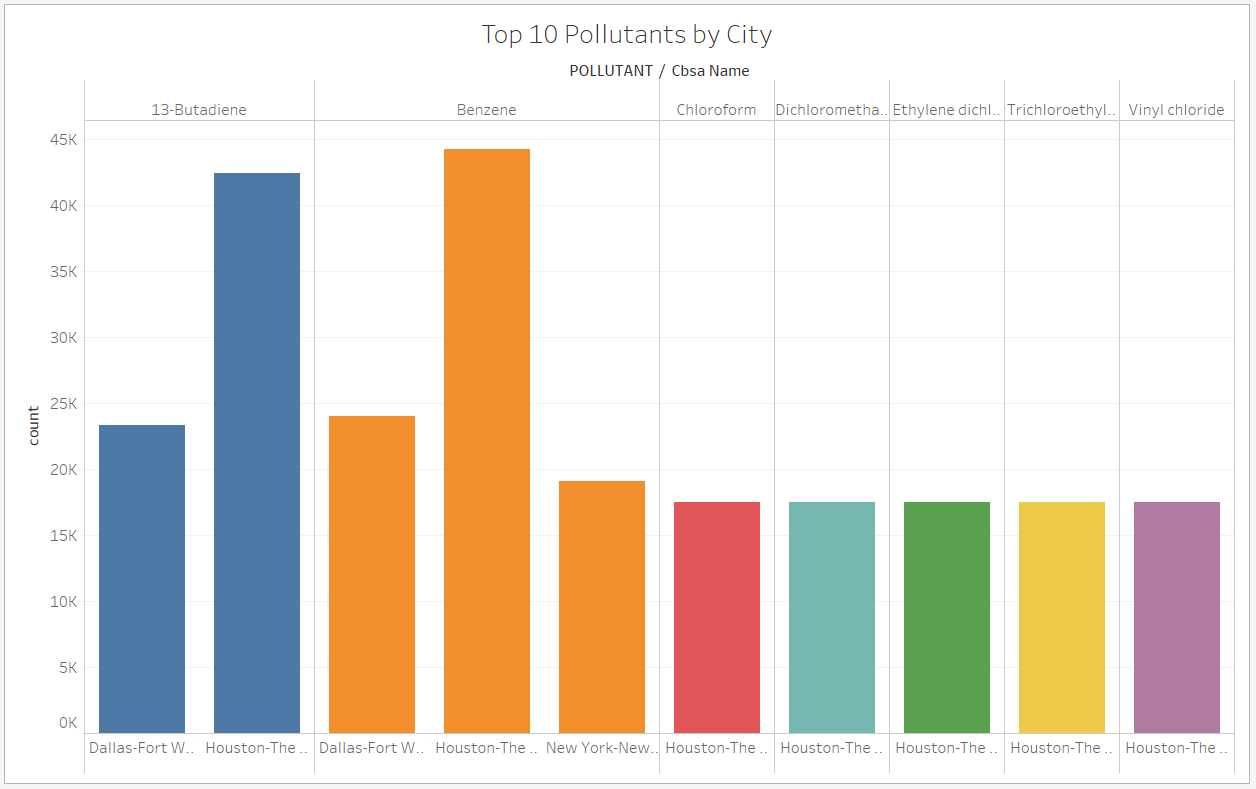
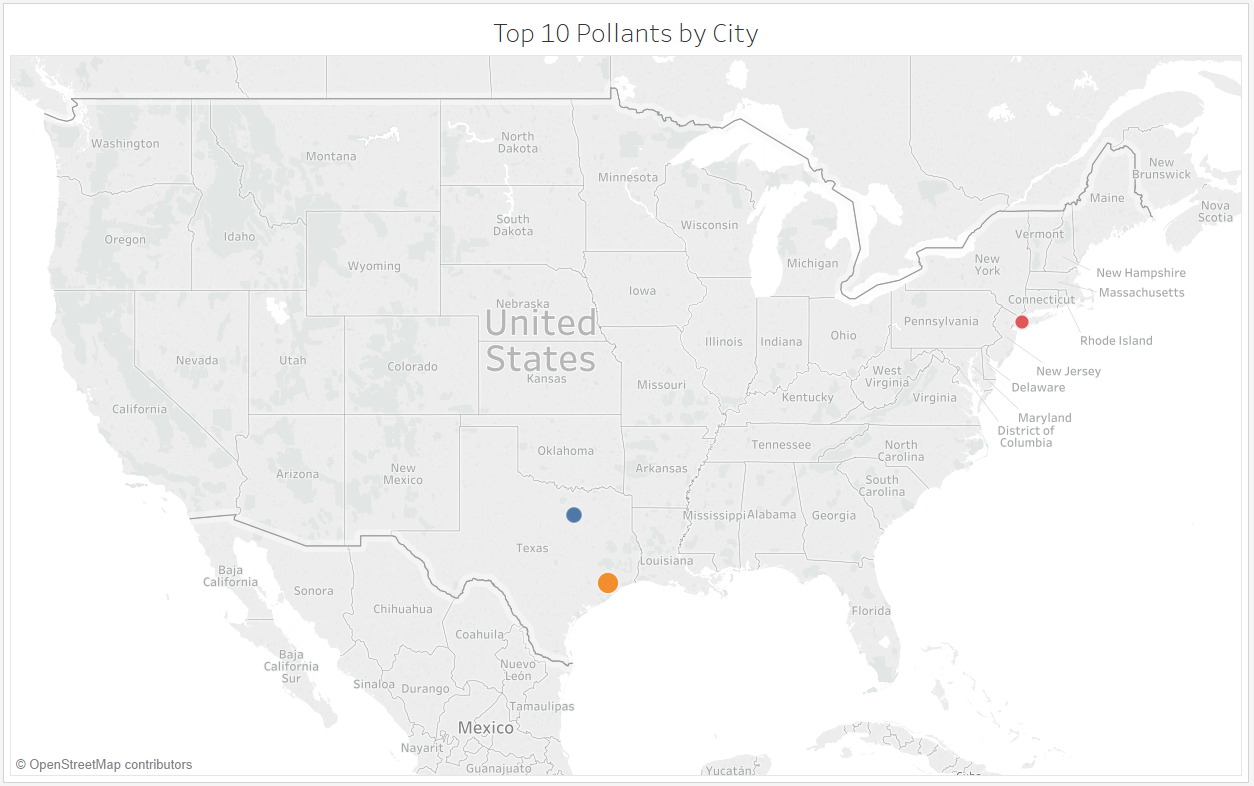


Figure 2. Top 10 Pollutants by City – Chart

Figure 3. Top 10 Pollutants by City – Geo Map

After the analyzation, and using Tableau, this bar chart and geo map are showing the top 10 pollutants by city recorded from 1990 to 2017. Houston-The Woodlands-Sugar Land in Texas got the highest ranking with Benzene measured 44,229 times in the 27 years of being monitored, followed by 13-Butadinese measured 42,392 times in the same area.

**4.3 What are the last 10 pollutants by city recorded from 1990 to 2017?**

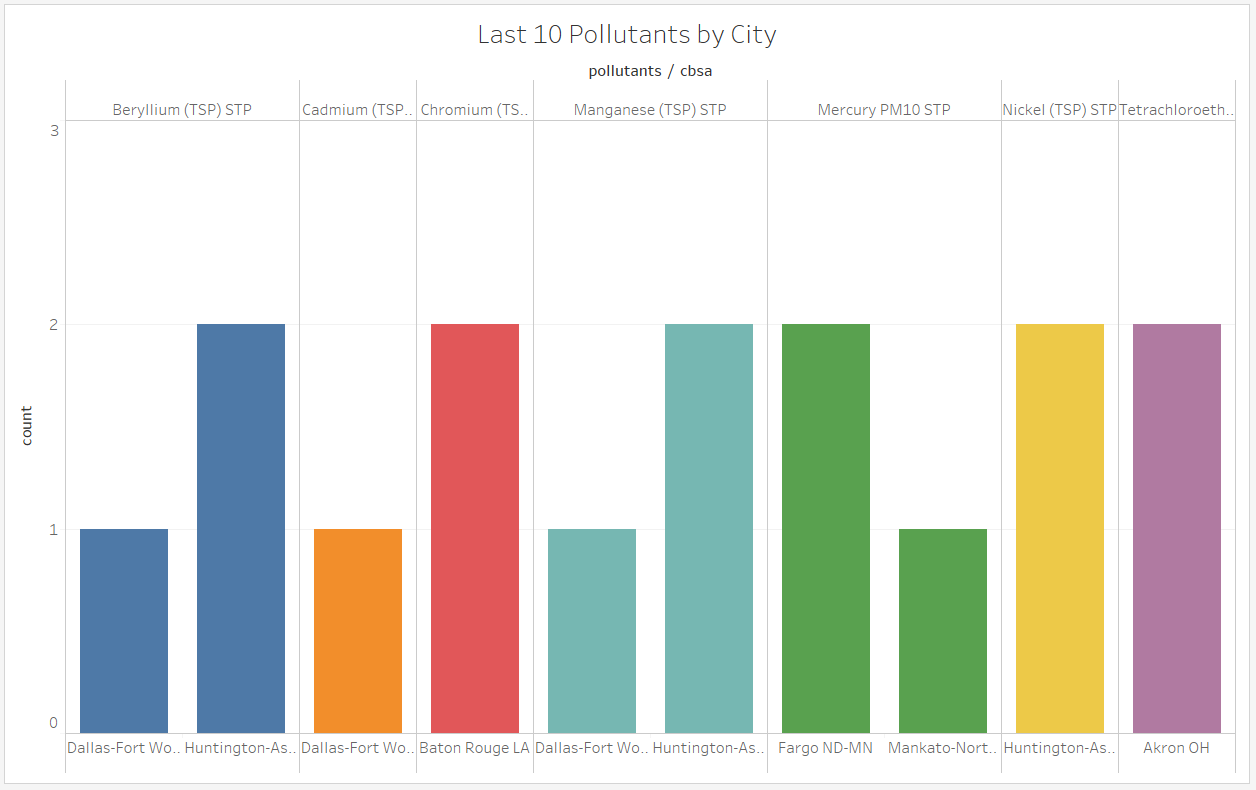


Figure 4. Last 10 Pollutants by City – Chart

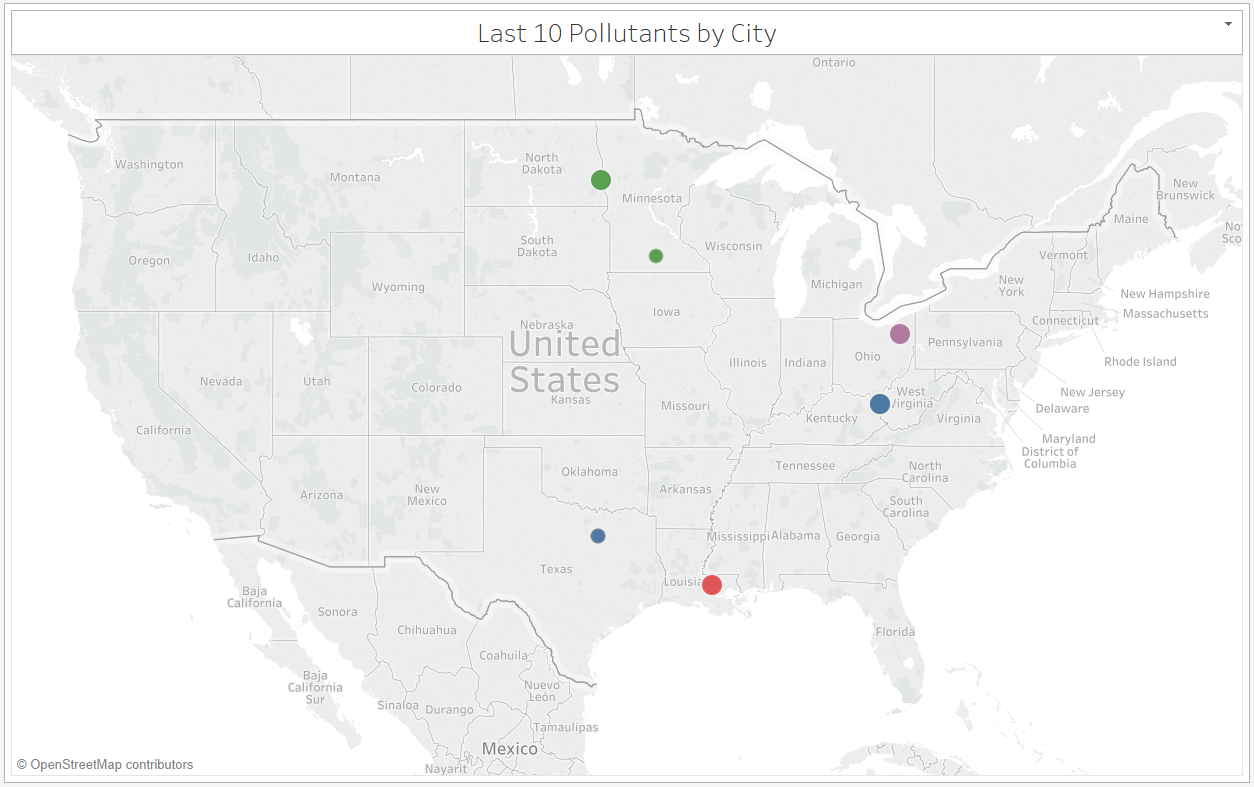


Figure 5. Last 10 Pollutants by City – Geo Map

After the analyzation, and using Tableau, this bar chart and geo map are showing the last 10 pollutants by city recorded from 1990 to 2017. The least measured pollutants by city were Cadmium (TSP) STP, Manganese (TSP) STP both measured one time in Dallas-Fort Worth-Arlington in Texas, and followed by Mercury PM10 STP measured one time in Mankato-North Mankato in Minnesota.

**4.4 What are the top 10 pollutants by state recorded from 1990 to 2017?**

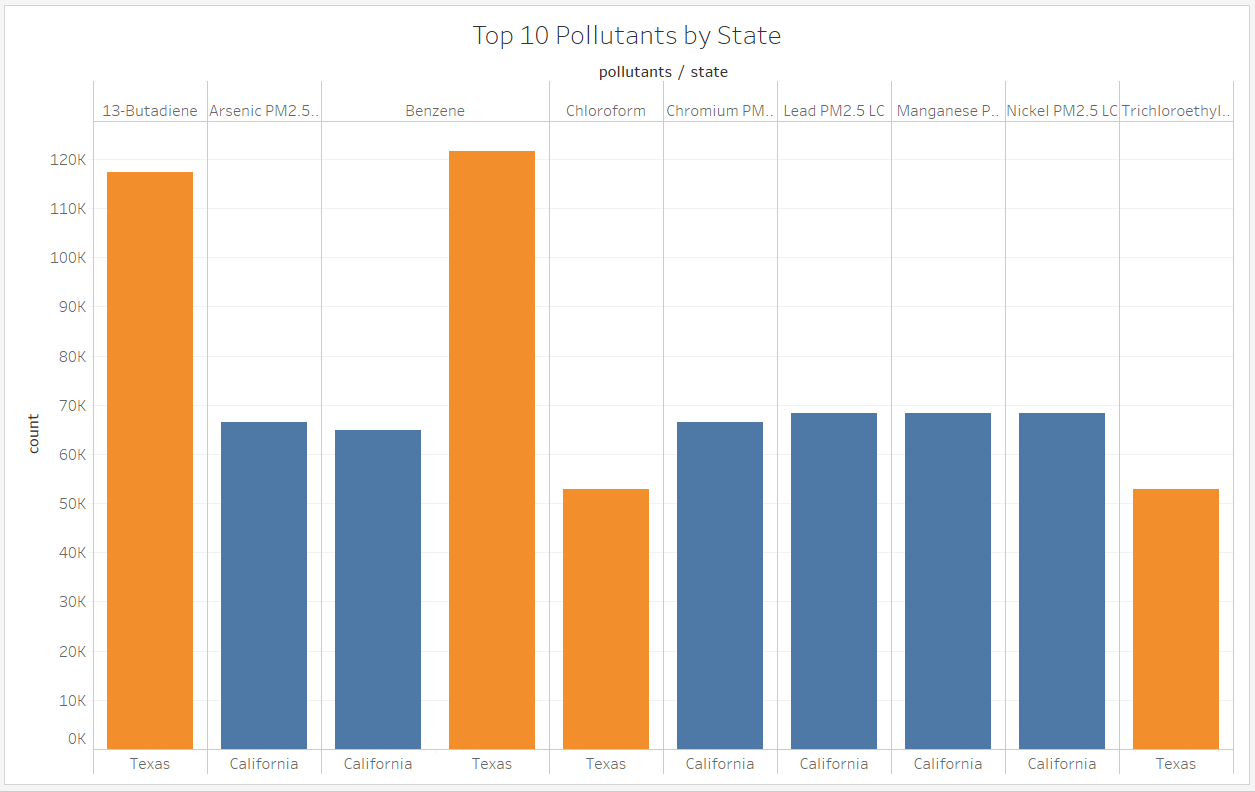


Figure 6. Top 10 Pollutants by State – Chart

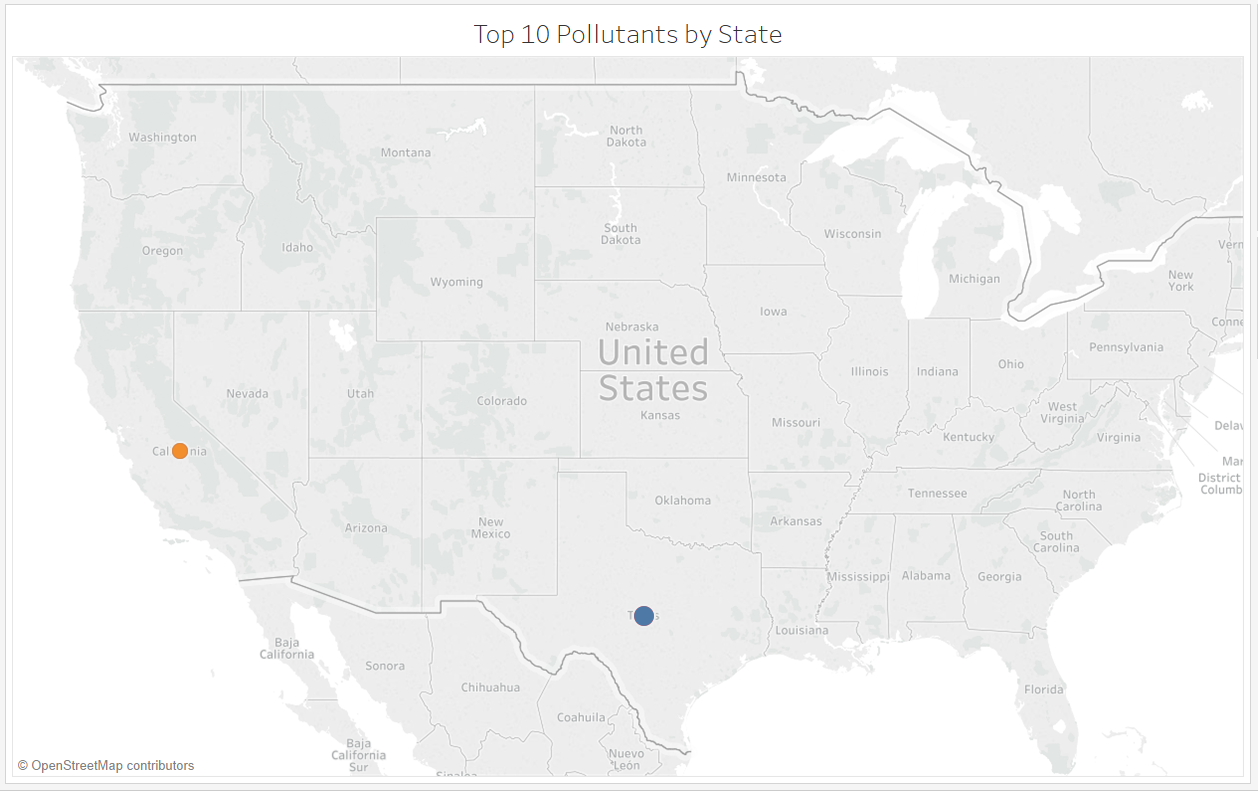


Figure 7. Top 10 Pollutants by State – Geo Map

After the analyzation, and using Tableau, this bar chart and geo map are showing the top 10 pollutants by state recorded from 1990 to 2017. The geo map shows that within the top 10 pollutants by state, the result is only generated in California and Texas. The 10 pollutants measured are Benzene with 121,622 times, 13-Butadiene with 117,431, followed by Lead, Nickel, Manganese, Arsenic, Chromium and Benzene with 64,000 to 68,000 in California.

**4.5 What are the last 10 pollutants by state recorded from 1990 to 2017?**

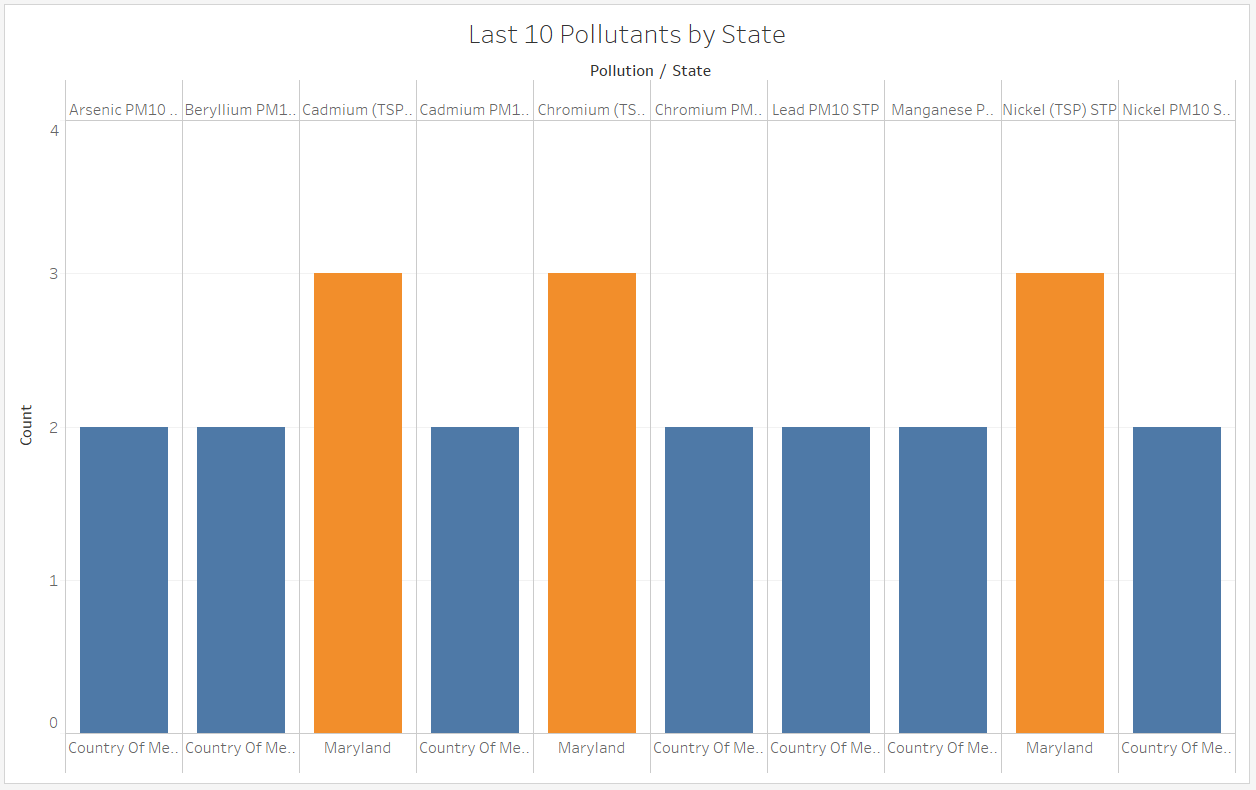


Figure 8. Last 10 Pollutants by State – Chart

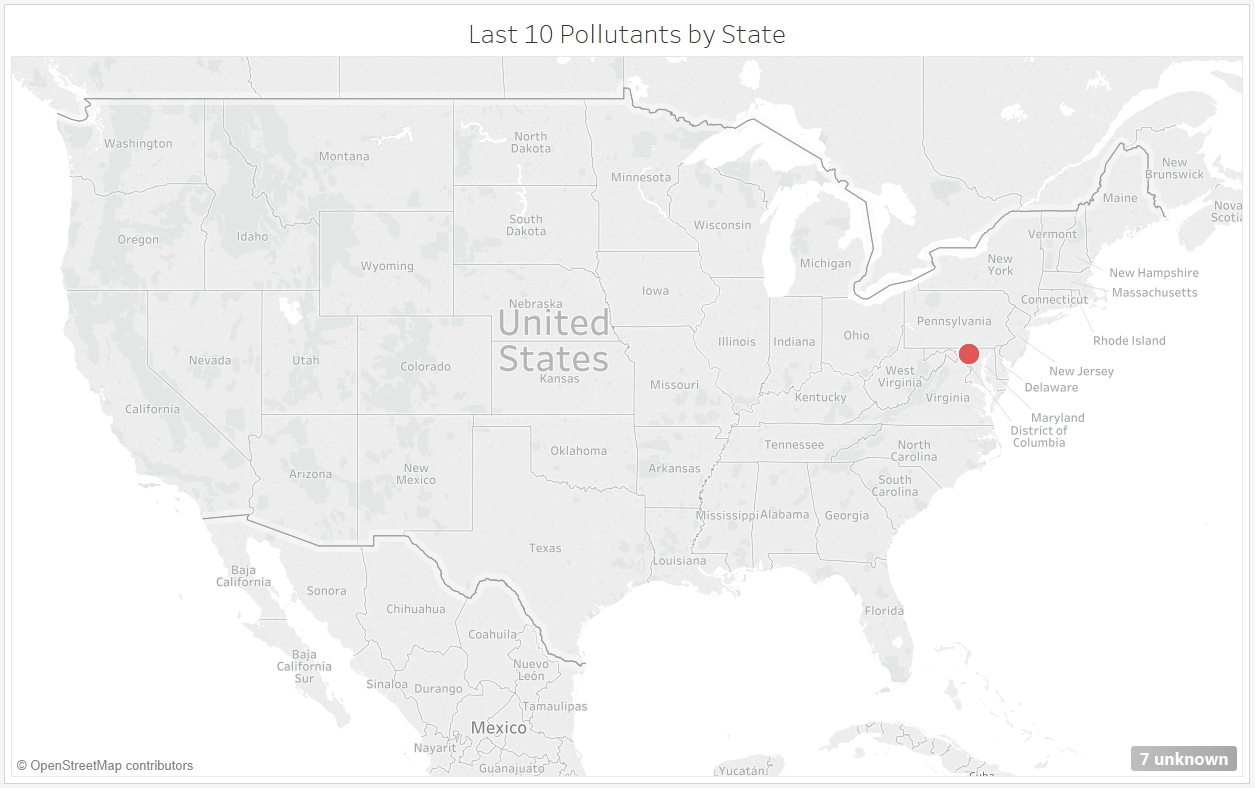


Figure 9. Last 10 Pollutants by State – Geo Map

After the analyzation, and using Tableau, this bar chart and geo map are showing the last 10 pollutants by state recorded from 1990 to 2017. The chart shows that the least 10 pollutants measured were only in Maryland and Country of Mexico with only 1 or 2 measurements of the pollutant type.

**4.6 What are the top 20 pollutants by date recorded from 1990 to 2017?**

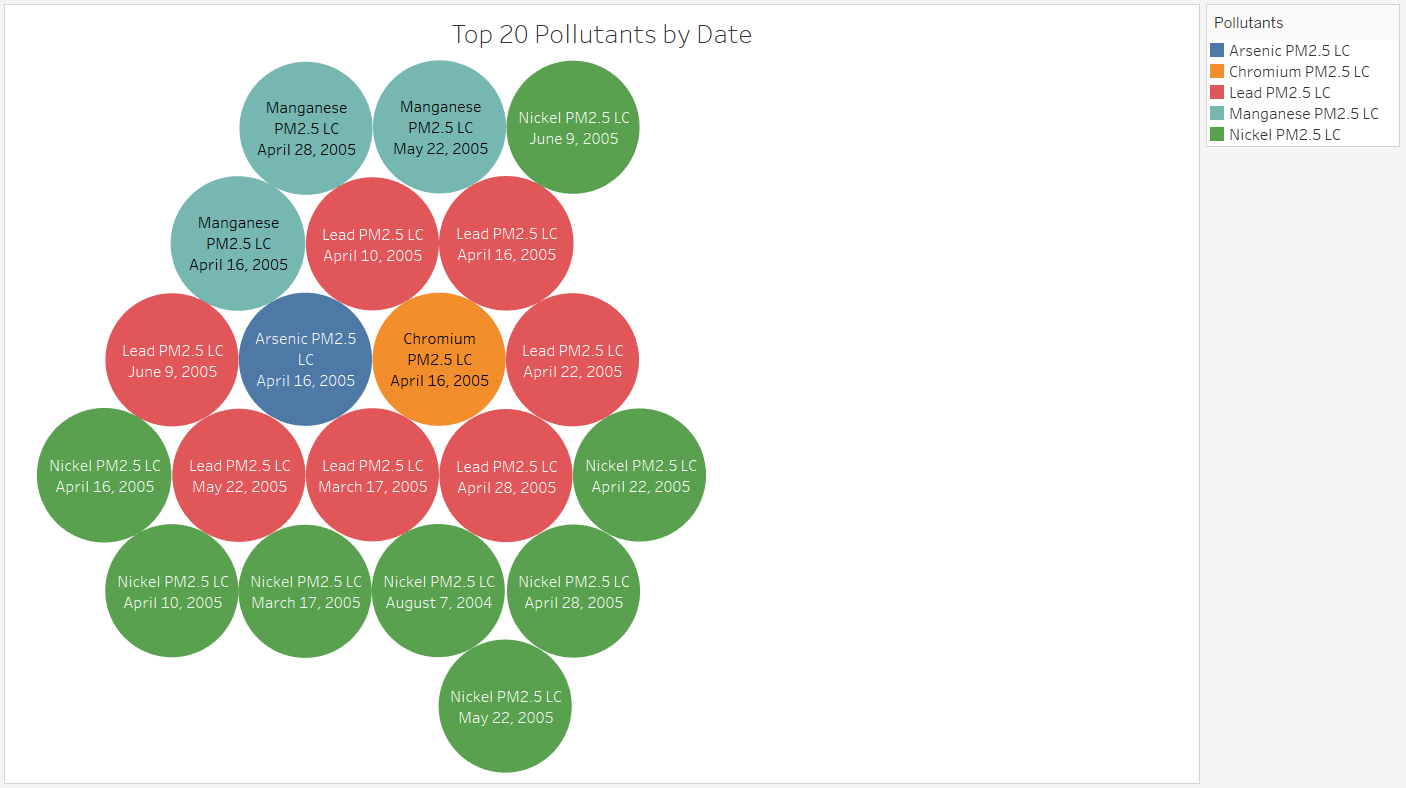


Figure 10. Top 20 Pollutants by Date – Bubble Chart

After the analyzation, and using Tableau, this bubble chart is showing the top 20 pollutants by date recorded from 1990 to 2017. In the last 27 years, the top 10 pollutants recorded by date were focused in the year of 2005 and mainly between March till June. Top 20 pollutants measured were also only 5 types, including Arsenic, Chromium, Lead, Manganese and Nickel.

**4.7 What are the last 20 pollutants by date recorded from 1990 to 2017?**

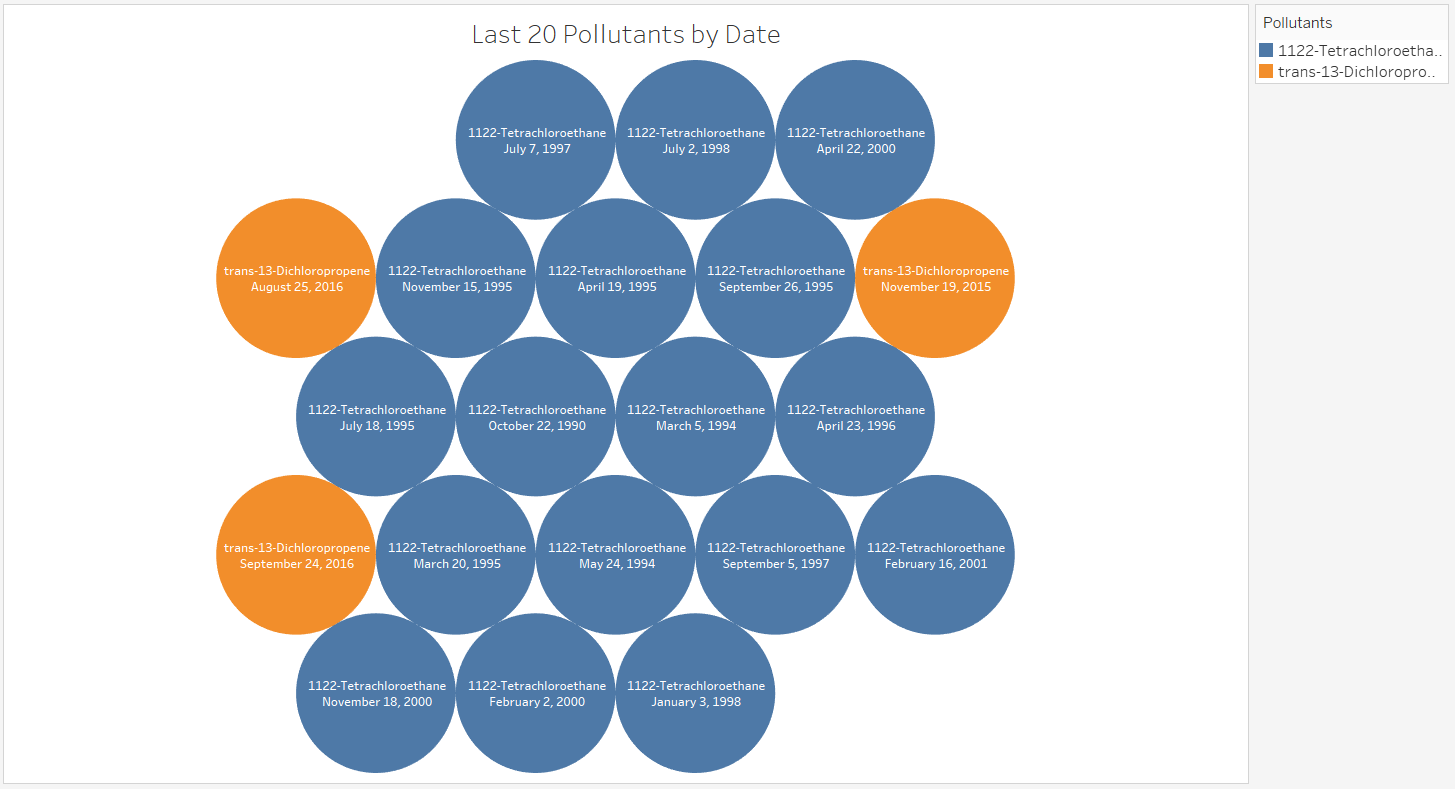


Figure 11. Last 20 Pollutants by Date – Bubble Chart

After the analyzation, and using Tableau, this bubble chart is showing the least 20 pollutants by date recorded from 1990 to 2017. From the bubble chart, we can see there are only 2 colors, which represent two types of pollutants, 1122-Tetrachloroethane, and Trans-13-Dichloropropene. Only these two pollutants were measure for the least 20 pollutants from 1990 to 2017. The date for them being measured is also focused mainly between 1994 till 2000. Only 2 measurements were recorded for these two pollutants to be found in 2015 and 2016.

1. **Conclusion**

From the above analysis, we can see that interesting sets of trends and patterns exists in large data sets which helps us to get a better understanding of the data.

Recent advancement in cloud technologies helps us to harness the power of parallel processing of a cluster of computers with little investment and almost no maintenance of the underlying computer hardware.

The 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.

We have Analyzed data to determine which air pollutants are measured the most. We have found that pollutant Lead PM2.5 has maximum pollutant factor.

The pollutant Benzene has affected the city of The Woodlands-Sugar Land in Texas the most and the pollutant Cadmium (TSP) STP has affected the city of Dallas-Fort Worth-Arlington Texas the least.

The pollutant Benzene has affected the state of Texas the most and the pollutant Cadmium (TSP) STP the state of Texas the least.

The pollutant Lead PM2.5 has affected the United States the most and the pollutant 1122-Tetrachloroethane has affected the least since 1997 till 2017.

1. **References**

[1] Air pollution dataset of United States from 1990 to 2017.

<https://www.kaggle.com/epa/hazardous-air-pollutants/data>

[2] Apache Hive <http://hive.apache.org/>

[3] Apache Hive Query Language Manual [https://cwiki.apache.org/confluence/display/Hive/Lan guageManual](https://cwiki.apache.org/confluence/display/Hive/Lan%20guageManual%20%20%20)

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

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

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