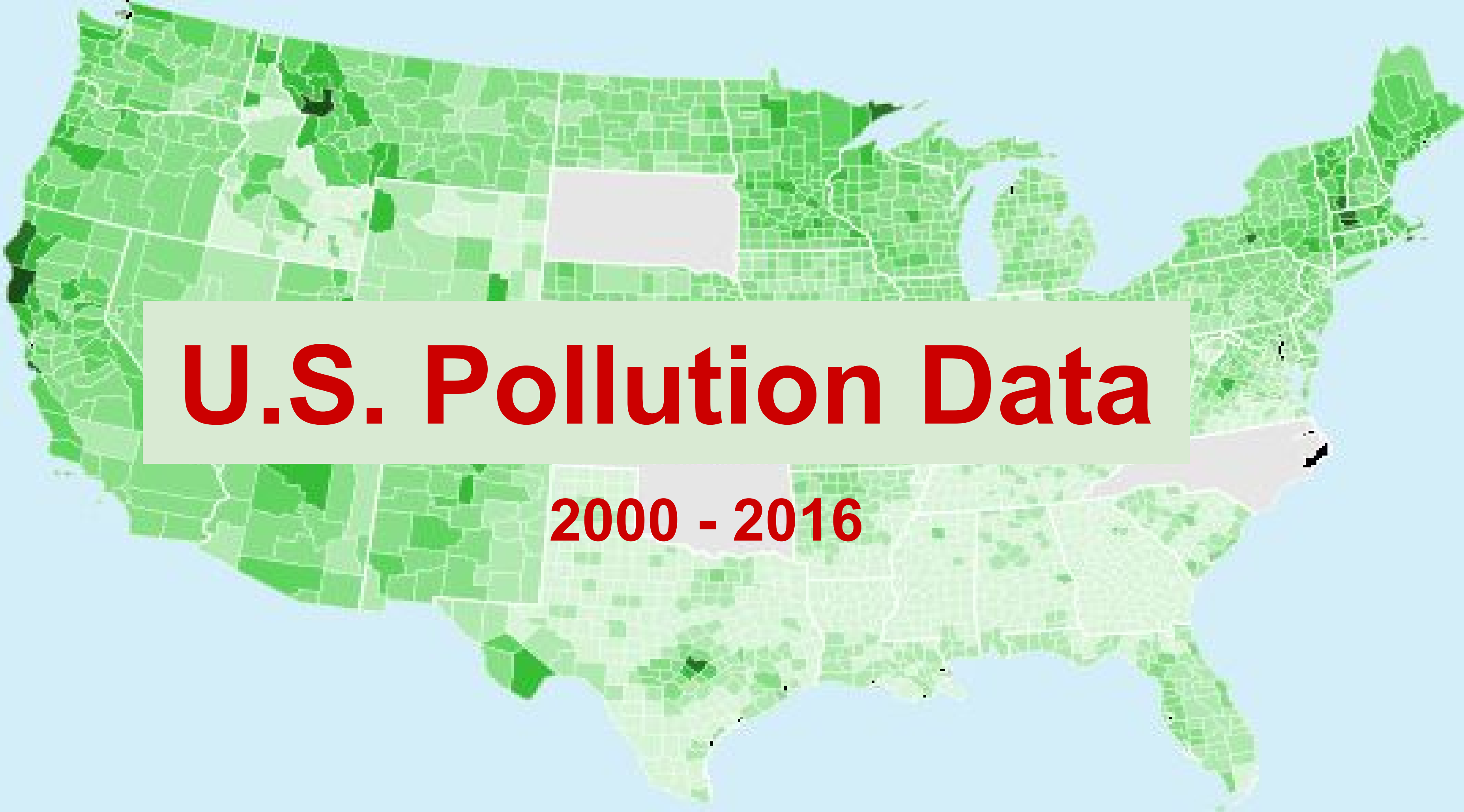
U.S. Pollution Data

2000 - 2016



Group Members

1. **Choe Choon Ho - 1132702963**
2. **Wong Tiong Kiat - 1132702943**
3. **Yue XiangRong - 1141327188**

U.S EPA

Due to increasing of concern about environmental pollution,US EPA was established on December 2,1970. Its main tasks are monitoring, standard-setting and enforcement activities to ensure environmental protection. Since 1970, EPA has been working for a cleaner, healthier environment for U.S.

Data Set Source

1. **Kaggle Dataset**
   - https://www.kaggle.com/sogun3/uspollution

2. **U.S EPA Database**
   - https://aqsdr1.epa.gov/aqsweb/aqstmp/airdata/download_files.html

Data Set Description and Quality

1. **Dimension** - 1,741,629 rows by 28 columns.
2. **Size** - 391.5MB.
3. **Duplicates** - Yes, 5032 observations.
4. **Missing values** - Yes, 2 features contains almost 50% missing values.

Data Set Features

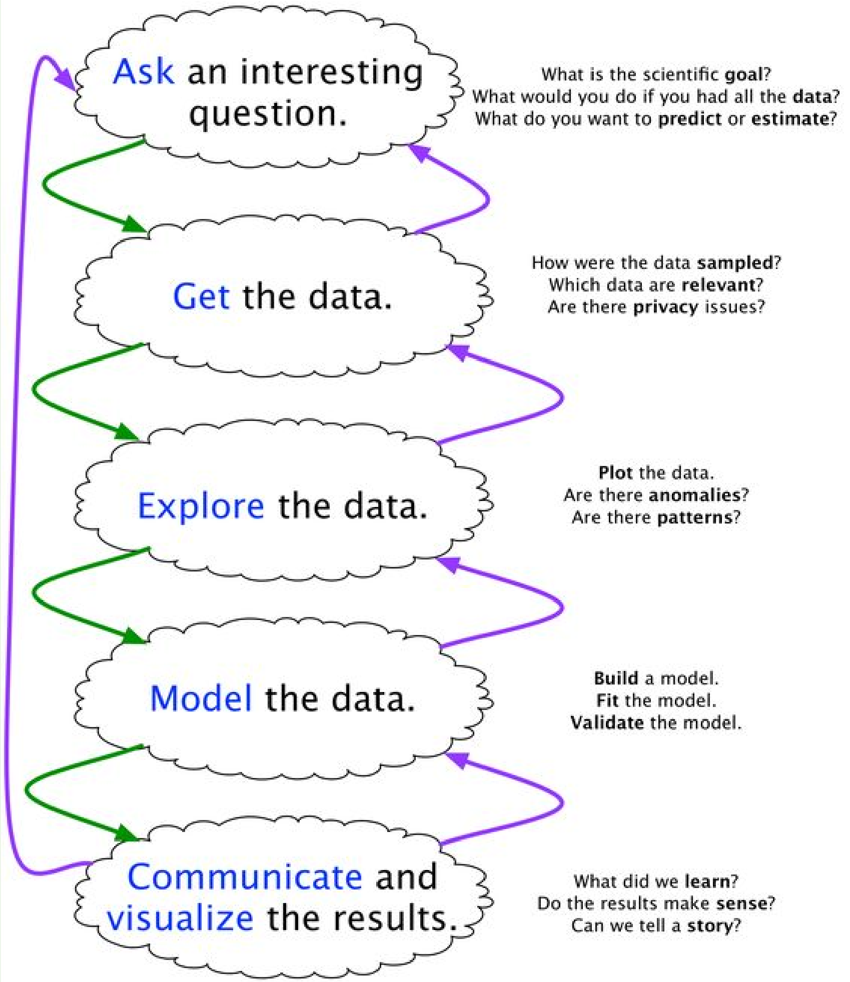
1. **State code**
2. **County code**
3. **Site num**
4. **Address**
5. **State**
6. **County**
7. **City**
8. **Local date**

Data Set Features cont.

1. **Nitrogen dioxide, NO2**
2. **Sulphur dioxide, SO2**
3. **Carbon monoxide, CO**
4. **Ozone, O3.**

Each pollutants comprises of **units measured**, **mean**, **air quality index (AQI)**, **max value**, *and* **max concentration**

The Data Science Process



Ask an interesting question.

What is the scientific **goal**?
What would you do if you had all the **data**?
What do you want to **predict** or **estimate**?

Get the data.

How were the data **sampled**?
Which data are **relevant**?
Are there **privacy** issues?

Explore the data.

Plot the data.
Are there **anomalies**?
Are there **patterns**?

Model the data.

Build a model.
Fit the model.
Validate the model.

Communicate and **visualize** the results.

What did we **learn**?
Do the results make **sense**?
Can we tell a **story**?

Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course http://cs109.org/.

Questions.

1. What is the **general trend** of pollution in the United States? Can we further narrow down our scope to **state level** and **site level** to obtain a more stable and convincing trend?
2. **Which state(s) have relatively higher pollution?**
3. **Is there a specific recurring trend across the years?**

Cleaning & Exploratory iteratively.

1. Initial checks (**duplicates** and **missing values**)
2. **391.5MB** was relatively **too large**. Therefore, we **subset** the data into two halves.
3. We ignored the missing values.
4. We saw the word **mean** in the features. Means and medians are usually **good representatives** and so we continued exploring with means. *The exploratory attempts we made will be briefed later.*
5. We later arrived to the point where we asked “How do we determine **which air is considered as polluted** and **how do we rank them**?

Cleaning & Exploratory iteratively. cont.

6. We could **not** answer the question.
7. We went to further deepen our domain knowledge.
8. **Air quality index (AQI)** is the measurement that the government have agreed on.
9. How are they ranked?

Cleaning & Exploratory iteratively. cont.

10. **Table for AQI rank.**

AQI	**Levels of Health Concern**
0 - 50	Good
51 - 100	Moderate
101 - 150	Unhealthy for sensitive groups
151 - 200	Unhealthy
201 - 300	Very unhealthy
301 - 500	Hazardous

Cleaning & Exploratory iteratively. cont.

11. We then check for **outliers** (mainly errors - values below 0 or above 500) using a **box plot** on the AQIs.
12. **Missing values** returns to haunt us. We then tried subsetting our data for **eyeballing**.
13. We observed there was a solid pattern and used “**bfill**” and “**ffill**” method to impute the missing values.

California	2001	**...**	**NaN**	1
California	2001	**...**	5	1
California	2001	...	**NaN**	**NaN**
California	2001	...	5	**Nan**

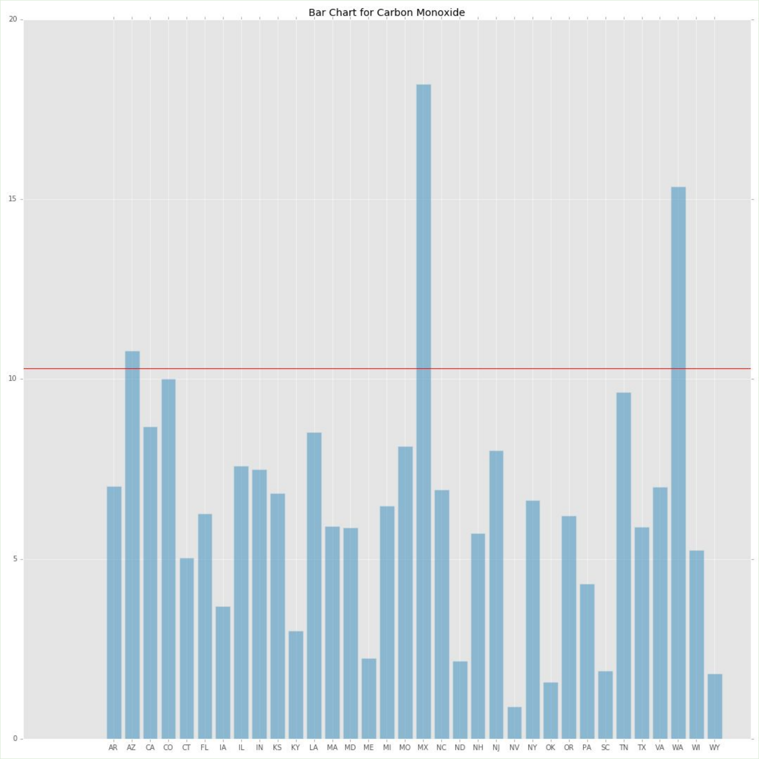
→

California	2001	**...**	**5**	1
California	2001	**...**	5	1
California	2001	...	**5**	**1**
California	2001	...	5	**1**

Cleaning & Exploratory iteratively. cont.

14. Another question arose, “Why are there 4 observations with the same state, country, site and date having the same AQIs?” (**redundancy**)
15. We then further **reduce** our observations by 4, resulting in one record per month for each state.
16. We then did a **line plot** to observe the trend (**regression**) and later heatmap (**choropleth**).
17. We selected states with the highest pollutants and continued to our data mining stage.

Bar charts for state selection

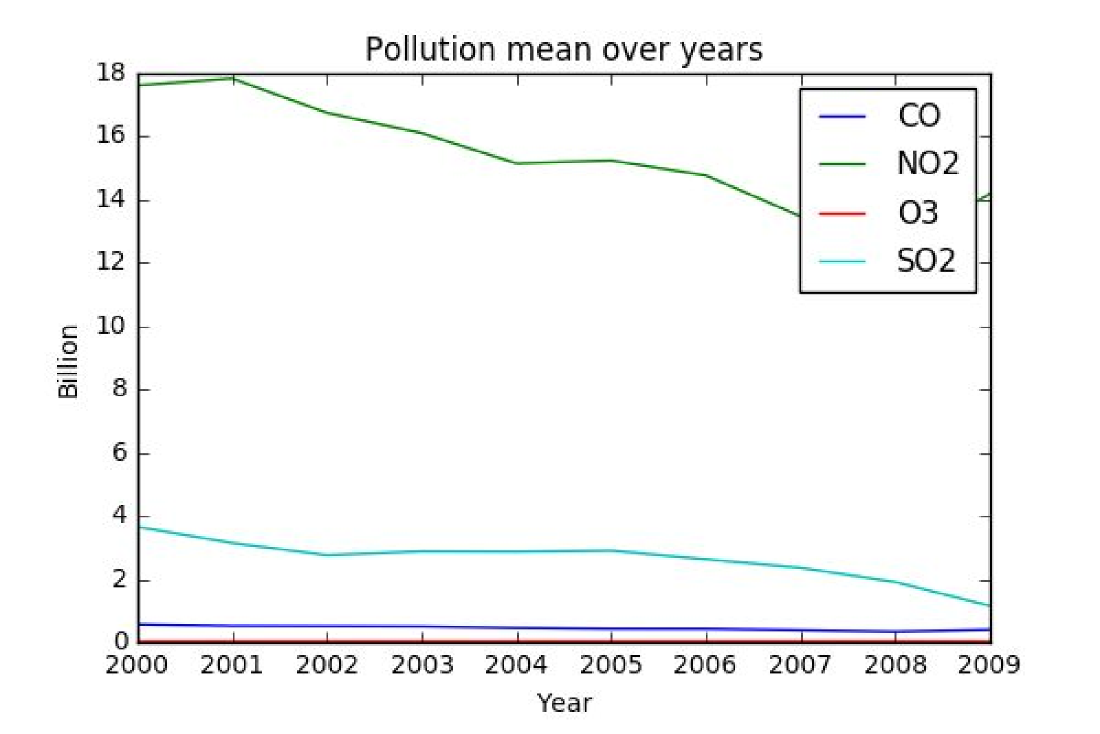


Exploratory Analysis with mean attempts.

We will now elaborate what we have done before using AQI...

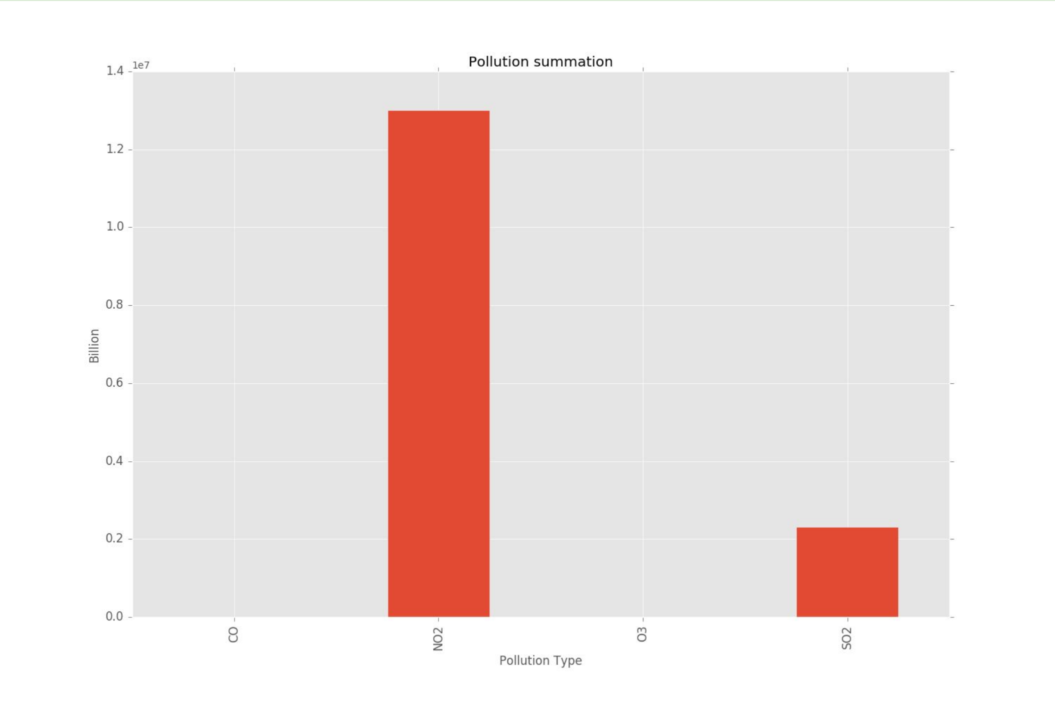
Exploratory Analysis with mean attempts. cont.

- How are time-series plots usually represented?



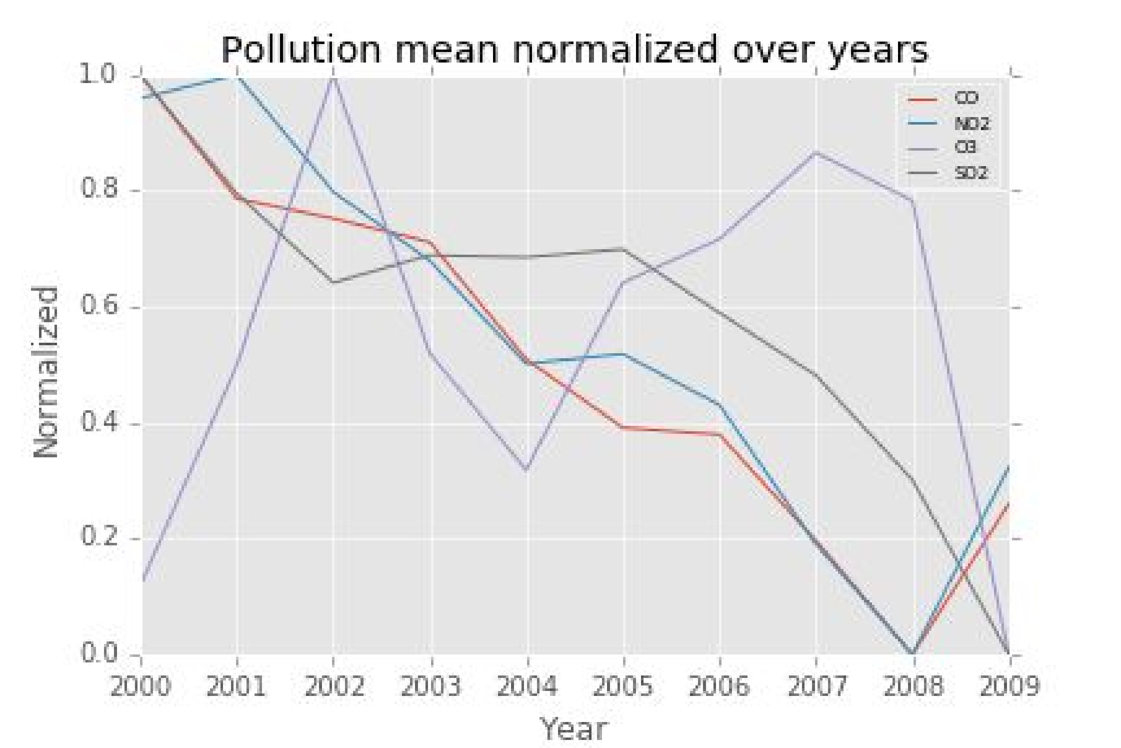
Exploratory Analysis with mean attempts. cont.

- Why did line plots show nothing?



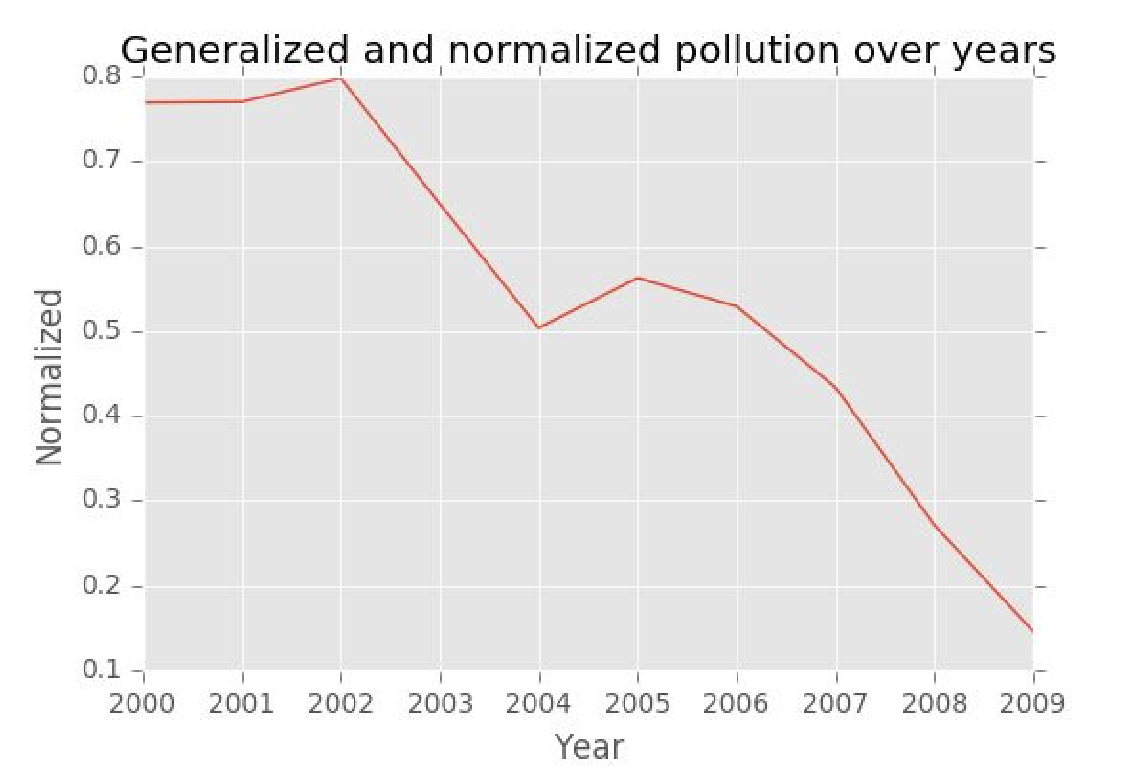
Exploratory Analysis with mean attempts. cont.

- How do we actually see each pollutants' trend?



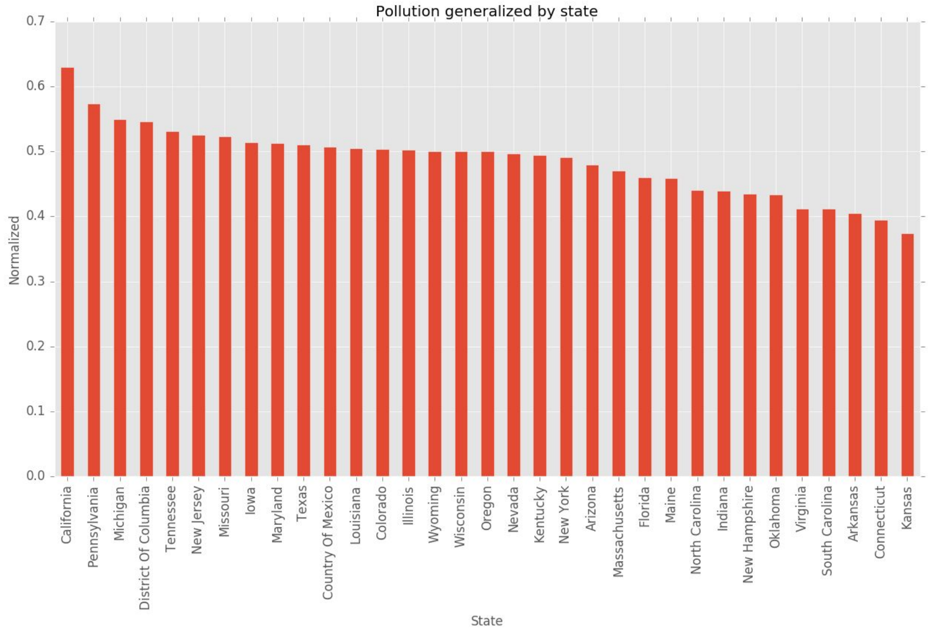
Exploratory Analysis with mean attempts. cont.

- How do we know the "generalized" trend of all of the pollutants?



Exploratory Analysis with mean attempts. cont.

- Which state contributes to pollution the most?
- To answer our second question.

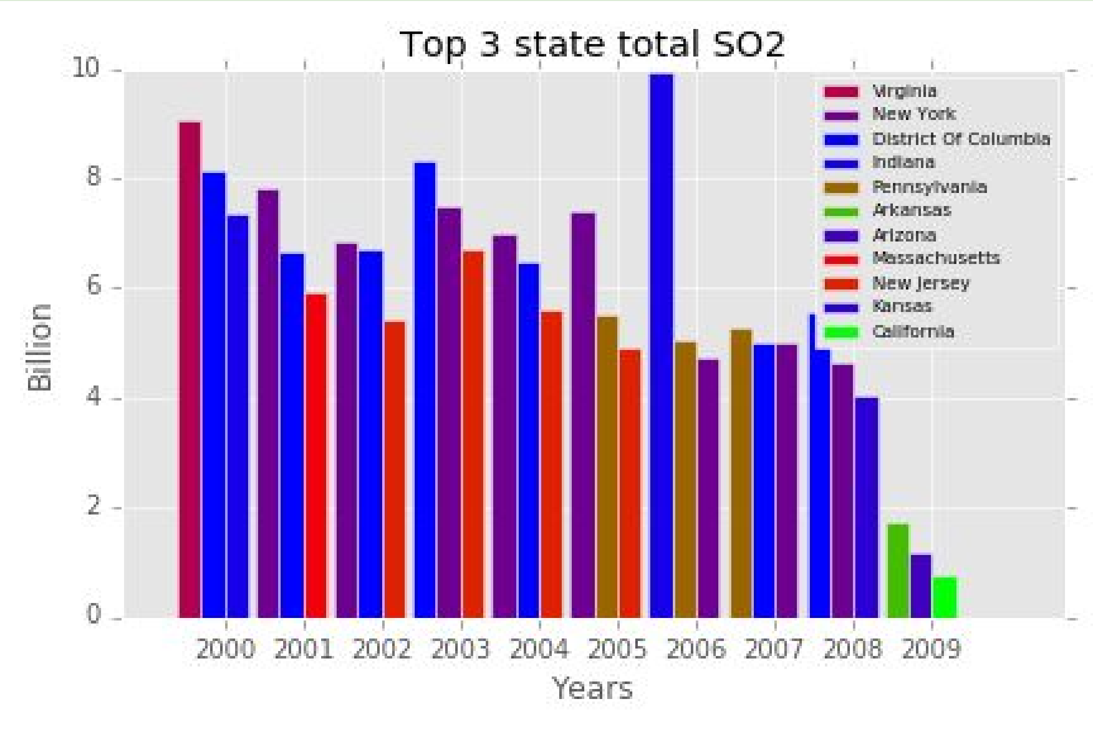


Exploratory Analysis with mean attempts. cont.

- However, the **two previous plots** are done on our **whims and fancies**.
- We realized that we contradicted our previous statement that each pollutants have different weighting.
- We cannot aggregate them because we lack of domain knowledge.
- So, we went back to plot each pollutant individually.

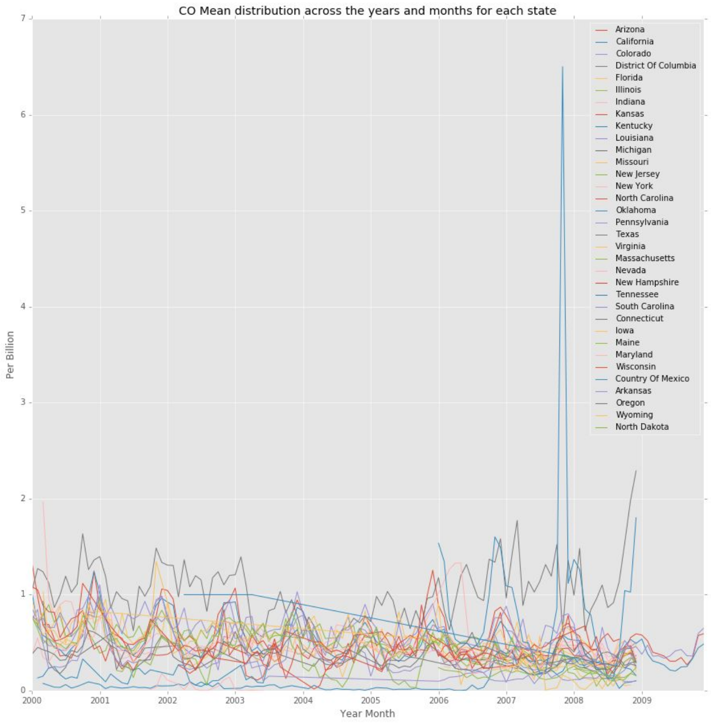
Exploratory Analysis with mean attempts. cont.

- Is the pollutant always coming from the same state? (New york) >>
- 4 pollutant of each plots



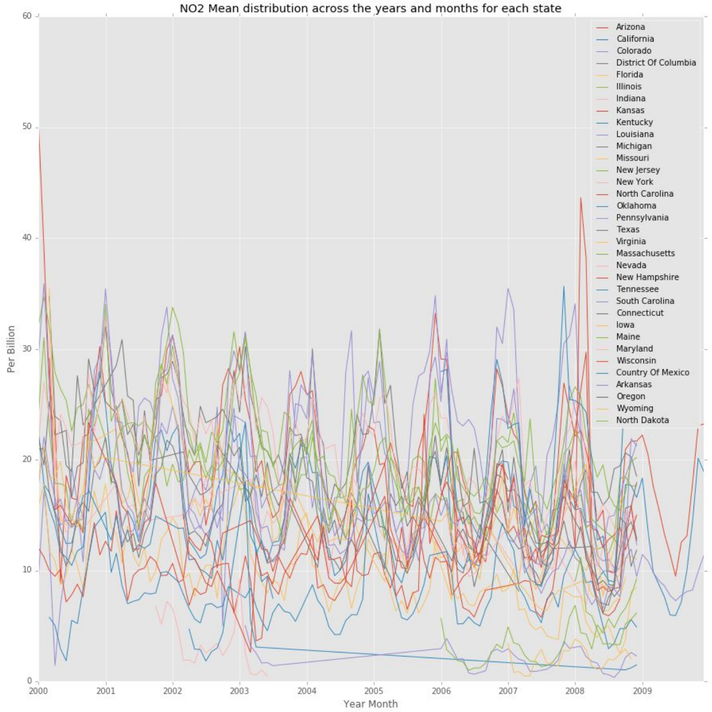
Exploratory Analysis with mean attempts. cont.

- If we add months into the “year” above the graph, will the dataset tell us anything interesting?
- What “patterns” can you derive from this graph?



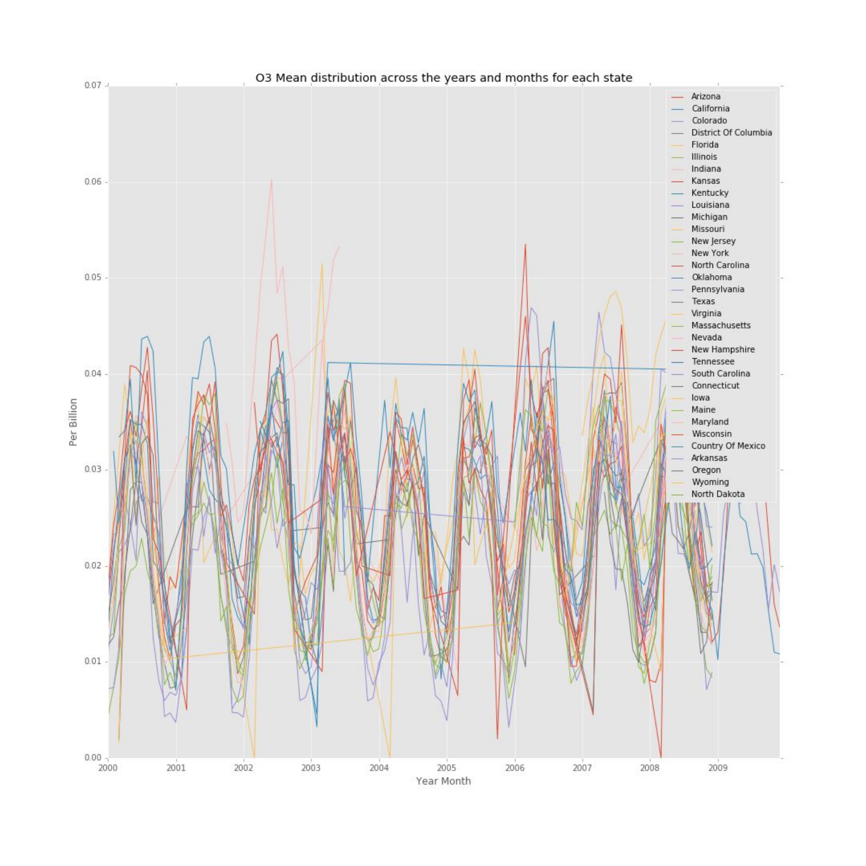
Exploratory Analysis with mean attempts. cont.

- If we add months into the "year" above the graph, will the dataset tell us anything interesting?
- What "patterns" can you derive from this graph?



Exploratory Analysis with mean attempts. cont.

- If we add months into the "year" above the graph, will the dataset tell us anything interesting?
- What "patterns" can you derive from this graph?



Exploratory Analysis with mean attempts. cont.

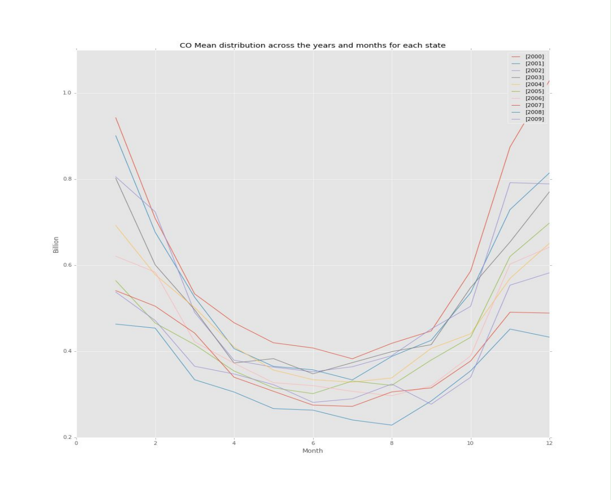
In conclusion, the time and effort spent studying mean was not a futile attempt. It was able to give us an **overall trend** that the pollutants level in U.S. is **decreasing**. Furthermore, was able to **answer our third question** "*Is there a specific recurring trend across the years*?" and as well lead us to change our thought process, however this would immediately been resolved if we had some domain expertise.

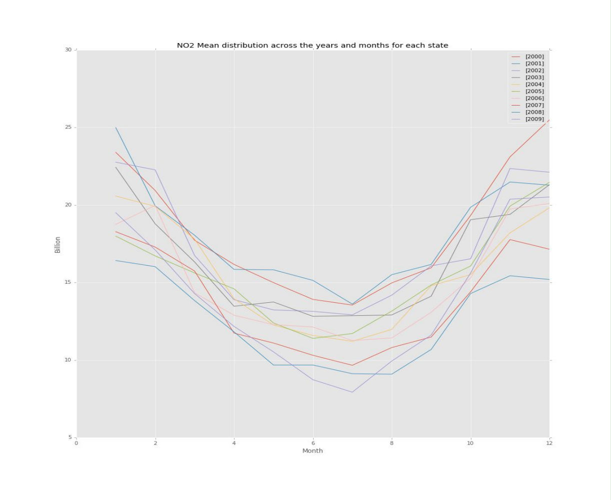
Exploratory Analysis with mean attempts. cont.

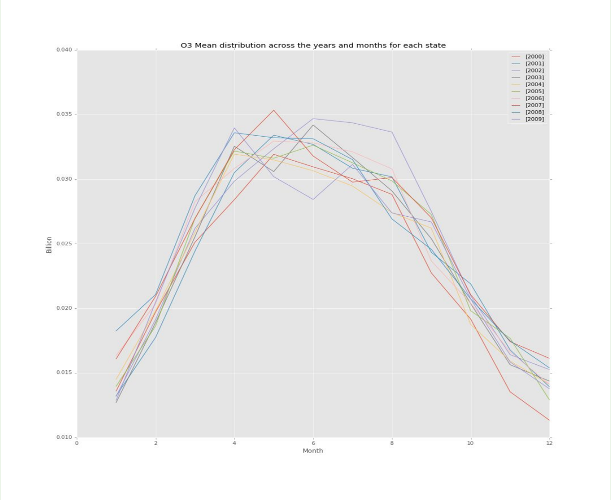
- Are the year month line plots for each state sufficient enough to answer your third question?
- Why remove the “state” and plot months for each year ?

Exploratory Analysis with mean attempts. cont.

Early summer(February - March), Autumn(September - October)







Carbon Monoxide

Nitrogen Dioxide

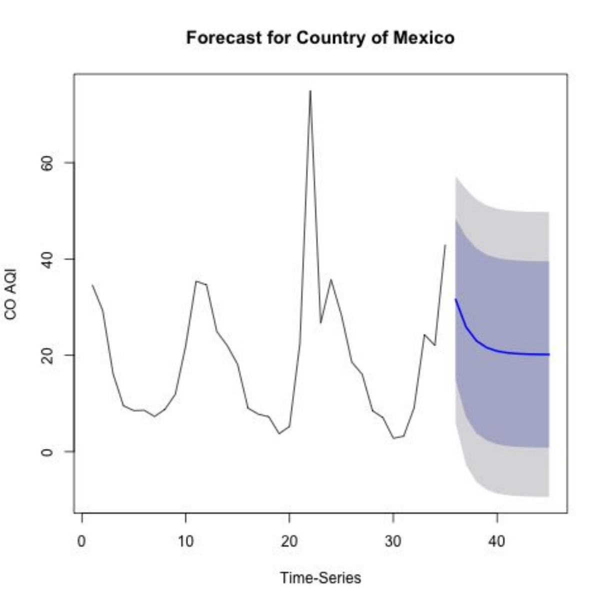
Ozone

Data Mining

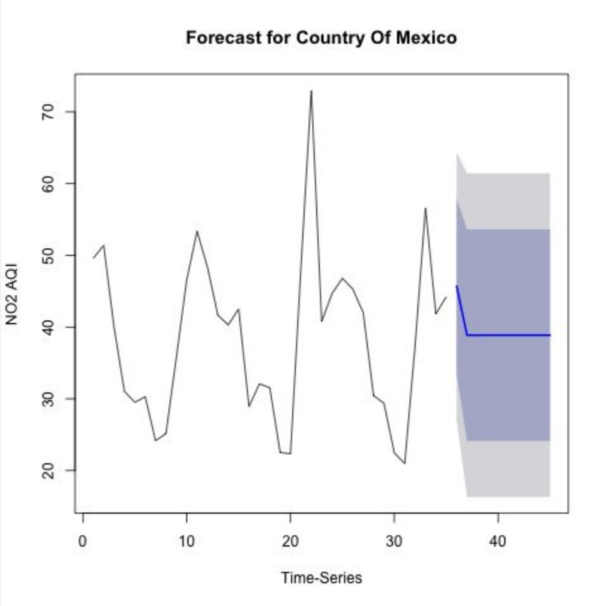
1. We took our preprocessed data and fed it into R.
2. We used R's "**forecast**" package and build multiple models for the state we previously selected.
3. We went from a **monthly-basis to** a **yearly-basis** as there were too many points in the line plots, making it difficult to see the trend. We aggregated the data using **mean**.
4. Technique used, autoregressive integrated moving average (**ARIMA**) model.

Data Mining cont.

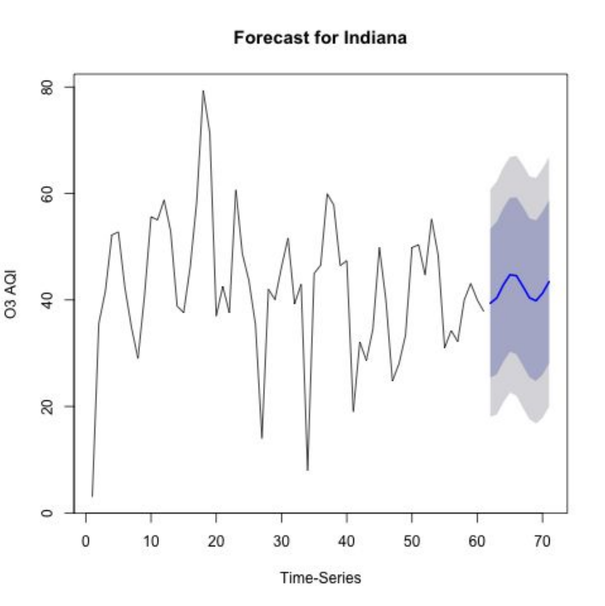
4. **Models** (samples)



Forecast for Country of Mexico



Forecast for Country Of Mexico



Forecast for Indiana

Visualization

Continue with Plotly HTML pages...

Future Business Cases

1. Instead of a time-series analysis, can we convert this into a **classification** task?
2. Currently, our data lacks information needed for classification. We can **collect data** such as temperature, population, area of urban regions, area of greenland, etc for each year along with their respective classes to increase the accuracy of the classification.

Conclusion

1. Majority of our time is spent on **Data Cleaning** and **Exploratory Analysis.**
2. Data Mining was effortless, we tried several techniques and went with the best-looking one.
3. Useful Insights found includes:
   - U.S. least polluted seasons are **Spring** and **Autumn**.
   - Most polluted state is California, Pennsylvania… (depicted in the bar chart above, slide 21)
4. Problems:
   - Lack of domain knowledge.
   - Do not know how to aggregate the 4 pollutants into 1 index. (Weights unknown)

References (Slides only)

1. Wikipedia
2. Quora