MULTIMEDIA UNIVERSITY®

TDS 2101 Introduction to Data Science Assignment

Case Study: Government

Prepared by:
Choe Choon Ho - 1132702963
Wong Tiong Kiat - 1132702943
Yue XiangRong - 1141327188

Lecturer & Tutor:
Dr. Bhavani Selvaretnam

Part A

Data Science is the study of data in order to acquire insights, knowledge or “science”. "Data science is a multidisciplinary blend of data inference, algorithm development, and technology in order to solve analytically complex problems." - Frank Lo. Data Science can be applied in various domains such as finance, business, meteorology, biology, chemistry, astronomy and etc. Since the development of the first silicon transistor, technology have been growing exponentially (Moore's Law), leading to the capability to store petabytes of data and processing speed which had overcome von Neumann bottleneck. These advancements have enabled data to be utilized in a robust manner. Terms such as “Big Data”, “Machine Learning”, “Hadoop” (distributed storage) and parallel processing are commonly mentioned in business and academia fields. Furthermore, technologies like 4G LTE wireless communications have significantly reduced the time needed for data transfer. In addition, 5G is already in development phase and is expected to be launched in the 2020s. Studies have shown that almost 50% of the world population have internet access, smoothing the process of data generation and collection with the help of online surveys, web pages clicks, social networks and etc. In the recent years, many researches have been done on Big Data resulting in various Data Mining algorithms and applications. Previous kills such as Business Intelligence, Data Analytics, Artificial Intelligence, Machine Learning as well have been incorporated into Data Science.

In this report, we will be focusing on application of Data Science in the government sector. Application of Data Science in government sectors includes stock market analysis, sentiment analysis, electoral campaigns targeting, crime prevention, weather forecast, employment forecast, disease forecast, population forecast and etc. Big Data is utilized differently depending on the environment and culture of where the government is situated. In Japan, seasonal events like the blooming of cherry blossoms are also measured. Countries which are located away from the equator (affected by the four seasons) may want to measure floods and droughts. Some citizen are against surveillance as it violates privacy which may hinder the process of data collection. Therefore, applications of Big Data varies in different governments.

Employment Forecast

The state of Indiana Department of Workforce Development (DWD) have been tasked to develop a system to forecast the number of jobs and the types of jobs available in the following years. Indiana DWD Data Science team cooperated with analytics professional services firm, Inquidia Consulting for the project. The data given to them was only 1-3kb files. However, the issue that arose was that they had millions of possible variable combinations to be taken into account for each county, resulting in different forecasts needed to be made which is processing power and time costly. The team used distributed computing to combat the issue since each forecast model can be computed independently. The team used Microsoft Azure to build their model using R coordinated through a PDI/carte process. The team used R's forecast package to build to a vector autoregression (VAR) model. At this point, another issue arose, VAR could only support a small number of variables. Therefore, feature selection was necessary in order to build their model. The team use brute force to select the features and finally presented their findings through Shiny. As a result, Indiana DWD can now accommodate the necessary skills and obtain statistical information at county and metro level instantaneously.

Crime Forecast

The state of Los Angeles Police Department (LAPD) used a mathematical model developed by Assistant Professor George Moher to predict criminal activity pattern. The mathematical model was originally developed for aftershocks during earthquakes. Later, it was found that crime data exhibits the similar trend and behavior. Therefore, LAPD applied the same equation on crime data. The data was split into two, the older (dating back further) part as the training set and newer part (acting as "future") as the test set. Trained with the older part, the model was able to predict the crime locations for the "future". However, there was a flaw with the model, it was designed to only predict crime hotspots; it was not able to predict the time that the crime would occur. LAPD then cooperated with university of California and PredPol to improve model to be able to estimate the time. A software was later developed with the improved model. The software notifies police officers about the location and time where likelihood of crime is high. The improved model was further trained with new crime data to increase the accuracy of its prediction. As a result, a reduced of 33% in burglaries, 21% in violent crimes and 12% in property crime was observed.

Weather Forecast

Weather forecasting has a long history dating back to the 650BC where Babylonians attempted to predict the weather by observing cloud pattern. Since then, many efforts have been made to improve weather forecasting to the extent that we can enjoy weather forecast from our news channel. However, the forecast is not always true. Millions to billions of dollars have been spent on weather forecasting due to the fact that weather is a vital aspect in daily activities, businesses and could even save lives. Weather forecasting technique varies by location and time and is affected by environmental changes such as the greenhouse effect. Therefore, getting a 100% accurate forecast is a difficult challenge. AcuRite, a weather forecast company, uses two methods of forecasting, “Numerical Weather Prediction” and “Precision Forecasting”. Numerical weather prediction first collects various ground and atmosphere data and sends it into complex model equations. The model firsts analyzes the data and makes a prediction. Then post-processing are done to improve their model. The post-processing techniques used are multiple linear regression, logistic regression, and artificial neural networks (ANN). The second method, precision forecasting uses measurements collected at the source (the location where prediction is to-be-made) over a period of time. Precision forecasting exploits the relationship between warm fronts and weather features such as cloud thickness and height, temperature, humidity and barometric pressure to obtain an accurate prediction. Another post-processing technique was proposed by Sigrist et al. which utilizes stochastic advection diffusion partial differential equations (SPDEs). The technique was applied in northern Switzerland precipitation forecast.

Natural Disaster Forecast

Natural disaster can occur anywhere in the world. Natural disaster includes earthquake, landslide, tornado, flood, tsunami and volcanic eruption and etc. However, different regions have different likelihood of being inflicted by these disasters. Natural disaster forecast saves lives through early evacuation or preparations. Therefore, natural disaster forecasting plays a vital role as most natural disasters are not controllable by us. Natural disaster forecast can be categorized into three vast categories, prediction, detection and disaster management strategies. Prediction is the ability to forecast the area and type of natural disasters in the foreseeable future. Detection is the ability to be alerted when a natural disaster occurs. Nowadays, detection is not a problem, however, the process for communicating the news to the proper authorities is a lengthy process which in turn delays the countermeasure process. Disaster management strategies involves the set of activities to combat the disaster. Detection and disaster management are mainly text mining, text processing and sentiment analysis. Prediction have a wider range of techniques which differs for each type of disaster. Prediction of earthquakes splits into prediction of magnitude and location. One technique used to predict magnitude is neural network using seismological data. Whereas location includes, nonlinear time-series and fuzzy rules and clustering using seismological data and historical data respectively. Flood prediction includes decision trees, logistic regression, artificial neural network on hydrological data. Landslides involves decision trees, naive bayes and neural network. Finally, volcanic eruption encompasses of multivariate time-series analysis.

Comparison & Commentaries

Government extends to a vast area of sectors, one common application of Big Data is forecasting. Government sectors which are mainly time-series analysis usually apply regression.

Weather forecast, in particular, took it to a next level by post-processing their prediction with various types of methods. One common issue was feature selection, however was solved through either brute force and domain expertise. Indiana DWD applied brute force strategy for feature selection (simply try all possibilities and select the best combinations), this tactic of tackling issue may not do well in a long term as it builds solely on the current data set they have. Instead, what they could have done is try to understand the relationships between variables and remove redundant variables or variables that have low influence on the prediction. Although this method is time costly. Next, AcuRite's approach in weather forecast is notable. The first technique uses measurements recorded by weather stations and the second compensate the first by alternating to local measurements only.

Potential Future Business Cases

For Indiana DWD, one possible business case would be to investigate further into individuals data that can be obtained through academia or surveys. Which type of jobs would an individual with certain demographic features and qualities choose? How long would a worker stay in the same company? Similarly for AcuRite, however, gathering personal information may be an issue. AcuRite could further gather user data along with the atmospheric measurements, they could further study at this temperature, time or humidity, what activities does the general public tends to do. This information can easily be monetized.

Part B

Questions

There are several types of question, descriptive, exploratory, inferential, causal, and mechanistic. Our data set mainly deals with time-series analysis, regression prediction. Therefore, our main question will be a question, “What will the pollutants trend be for the next upcoming years?”. Our initial exploratory and descriptive questions are as follows:

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. Are there certain states where Environmental Protection Agency (EPA) should take action; Which states have relatively higher pollution?
3. Is there a specific recurring trend across the years?

The first seeks to determine the overall trend of pollution in the U.S. Is the air getting cleaner each year or worse? We could also further investigate the trends at state, county and site level. If possible, we hope to find a linear regression or a consistent upwards or downwards trend since they make ease the process of prediction. The second mainly targets states with high pollution level. The third seeks to find a common pattern throughout the months each year. Such information can be used for recommendations or suggestions for a holiday, lodging, or even study. Naturally, we will have to do a background check as to which season and whether the environment is suitable to the purpose of the request.

Data Set Description

Our data set concerns pollution. It was posted on Kaggle, https://www.kaggle.com/sogun3/uspollution. It originated from the United States' Environmental Protection Agency's database. The size of it is 391.5MB. It contains four major pollutants, nitrogen dioxide, sulphur dioxide, carbon monoxide and ozone, recorded from year 2000 to 2016. Its features includes state code, county code, site num, address, state, county, city, and local date of monitoring. In addition, each pollutant comprises of another 5 columns each, which are the units measured (billions or millions), mean within the given day, air quality index (AQI), max value of the day and maximum concentration of pollutant of hour in the given day.

Data Set Quality and Cleaning & Exploratory Analysis

The data set has 1,741,629 observations with 28 fields. It contains 5032 duplicate observations and a large amount of missing values. In sulphur dioxide AQI, 872,097 NAs and carbon monoxide AQI, 873,323 NAs. First, we dropped the duplicates since they have an exact equivalent observation and therefore will definitely not be beneficial to us. Since 1,736,597 by 28 is still relatively too large and would require a lot of processing power and time. Therefore, we will mimic LAPD's method of subsetting. Since the data set was originally sorted by date, we can split it into 2 halves. With the first half, we proceeded to exploratory analysis.

Initially, we thought of adding a column for season. However, we decided to perform this on the fly later as it will take up more memory if we need it at a later time. Next, we observed missing values and the need to fill them. Initially, we disregarded them and went to exploratory stage. Later, we encountered a problem, "How do we determine which air is considered as polluted and how do we rank them?". We could not answer the question and therefore went to deepen our knowledge about the domain. From our findings, we discovered that air quality index (AQI) is the measurement given by the government to rank pollution. The AQI can be seen in Table 1. Finally, we arrived back at the need to fill missing values. Through eyeballing, we found that the AQIs of the same state, county, site and date would have similar values. Furthermore, there seems to be a certain pattern in missing values. We then decided to take existing values to fill missing values. For SO2 AQI, we used Python’s Pandas’ Dataframe “bfill” method, which fills NAs with its following values and CO AQI, we used “ffill” method which fills NAs with its preceding values. These methods are chosen conforming to the trend observed.

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

Table 1, U.S. AQI

The following passage will describe our previous attempts before arriving to the usage of AQI. Initially, we took the most apparent feature, the mean value across the year for each pollutants as mean is usually the best representative. We plotted a line plot to observe the trend for each pollutants by using the mean which be seen in Figure 1. The plot depicts NO2 being the highest and having a downwards trend across the years and a rise in 2009, however, this could be caused by our subsetting technique or the occurrence of some event. Therefore, we disregarded this fact for the time being. The other pollutants showed an almost constant line and therefore requires further investigation. We decided to narrow our scope by limiting the y-axis. Then, we saw that they are not constants but downwards trends as well. This lead us to plot a bar chart to show each pollutants average concentration which can be seen in Figure 2. Two bars can be seen whereas the other two were non-existent. This lead us to normalize the pollutants and re-plot the line plot. Figure 3 shows that SO2, NO2, and CO have a downwards trend however O3 have a rather unstable pattern. Next, we tried combining the pollutants to observe the overall trend using a line plot depicted in Figure 4. The plot shows that the overall pollution is decreasing over the years. Later we realized that this was according to our whims and fancies - we assumed that the weights for each pollutants are equal. Therefore, we plotted the trends for each state in one line plot for each pollutants. The resulting line plots was a disaster as it was messy and difficult to read (under Appendix I). We also realized there were a lot of missing records which caused some lines to discontinue their usual pattern. However, we also realized there was one trend that majority of the lines follow. So, we instead plotted a line plot using the average of the mean of each state, month grouped by year for the different pollutants which can be seen in Figure 5. The plots gave us very convincing repetitive trend throughout the years with the exception of SO2. CO and NO2 have similar trends, an inverse bell curve whereas O3 gave us a bell curve. We found that the production of O3 rises during heat waves because the plants absorb less of it. With this information, we deduce that the rise of O3 is due to summer. Therefore, we cross validated our deduction with U.S. seasons and found that our deduction holds true where the summer arrives in the middle of the year. The main problem is what should we say if one will to ask what is a good time to visit the United States? We therefore decided it is better for us to select the crossing points (the point begin their steep change) around February to March (which is the beginning of spring) and September to October (which is the beginning of fall). However, this is due to the fact we lack domain knowledge about the weights for each pollutants on how they affect the human respiratory system. Soon after, we arrived at the problem stated in the previous paragraph.

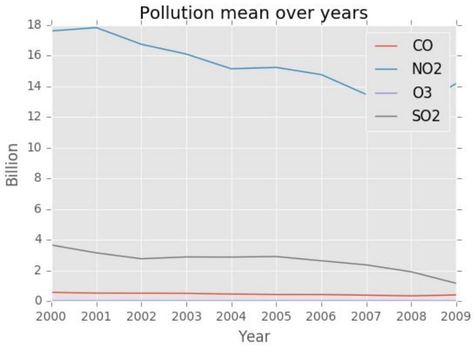


Figure 1, Line Plot for Pollutants.

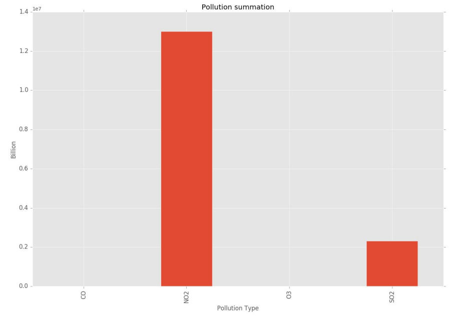


Figure 2, Bar Chart for Pollutants

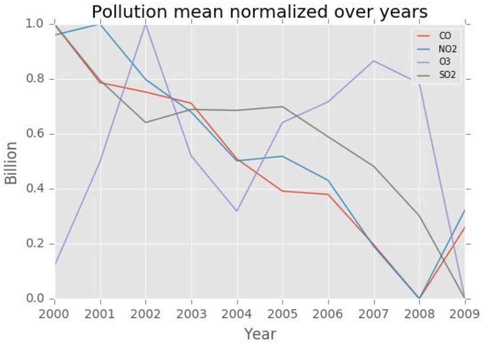


Figure 3, Normalize Line Plot for Pollutants.

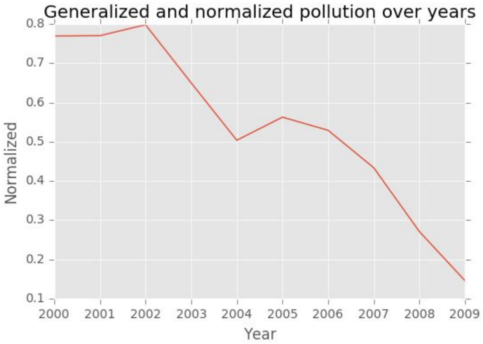
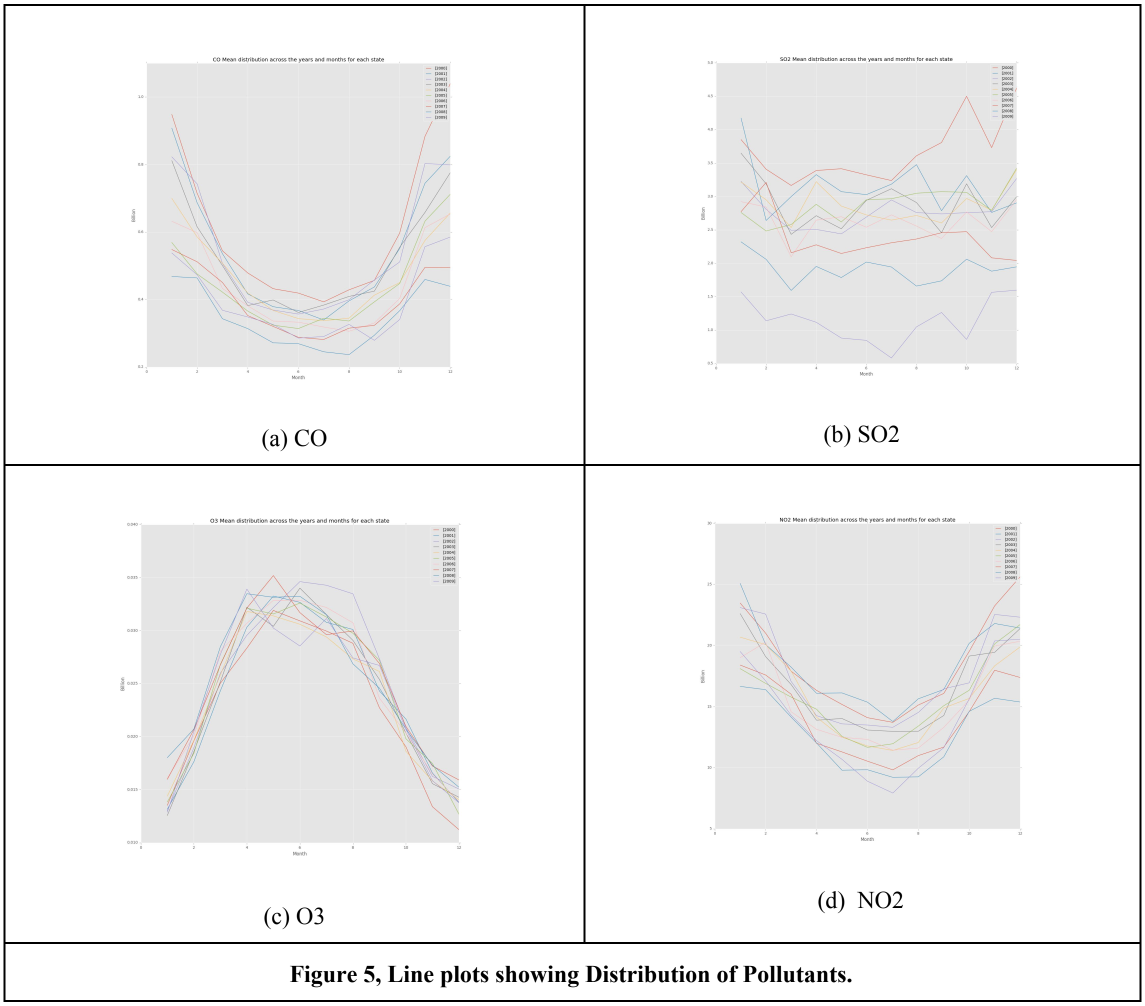


Figure 4, Generalized Line Plot.



(a) CO

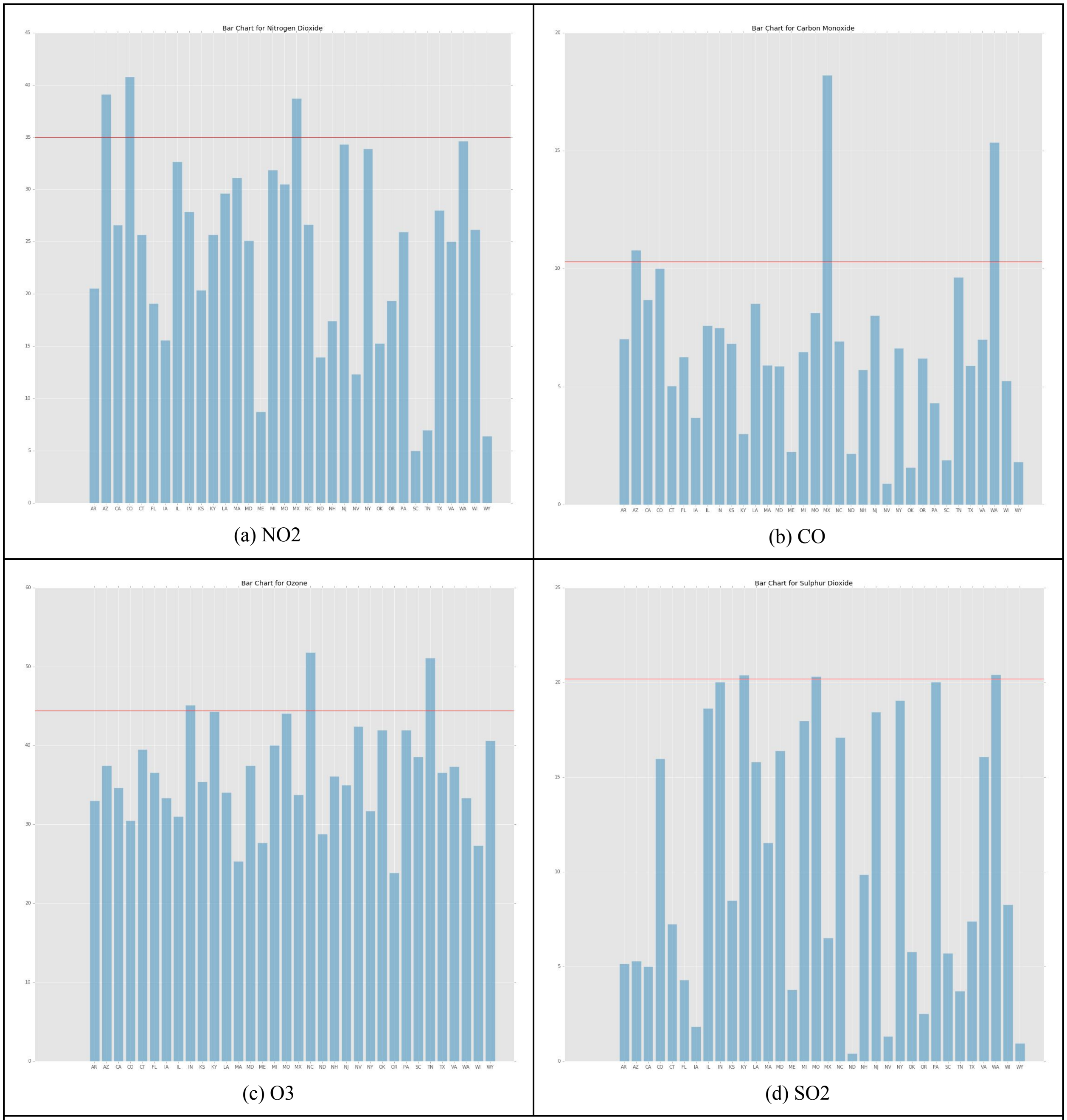
(b) SO2

(c) O3

(d) NO2

Figure 5, Line plots showing Distribution of Pollutants.

The following passage will describe our exploratory analysis using AQI at state level. We first plotted a box plot for the AQIs of each pollutant to ensure that the values stay between 0 and 500 following Table 1. Next, we decided to plot a line plot to see the regression lines for each pollutant grouped by states throughout the years. We realized that there are a lot of missing records which produces gaps in the line plot as well as numerous redundant points. Here, our previous method used to fill missing values posed yet another question, “Why are there 4 observations with the same state, county, site and date having the same AQIs?” Thus, we reduced the number of observations by 4 since they have the same AQIs, and proceeded to our line plot. At this point, we asked again, “Why don’t we load the entire data set instead since we are going to reduce it by 4?” and so did we. We continued exploration with the reduced data and plotted bar charts to observe the distribution for each pollutants. As the data set contains too many states, we decided to select the top three most polluted states for each pollutant and proceeded data mining phase. The figure below shows how were the states chosen.



(a) NO2

(b) CO

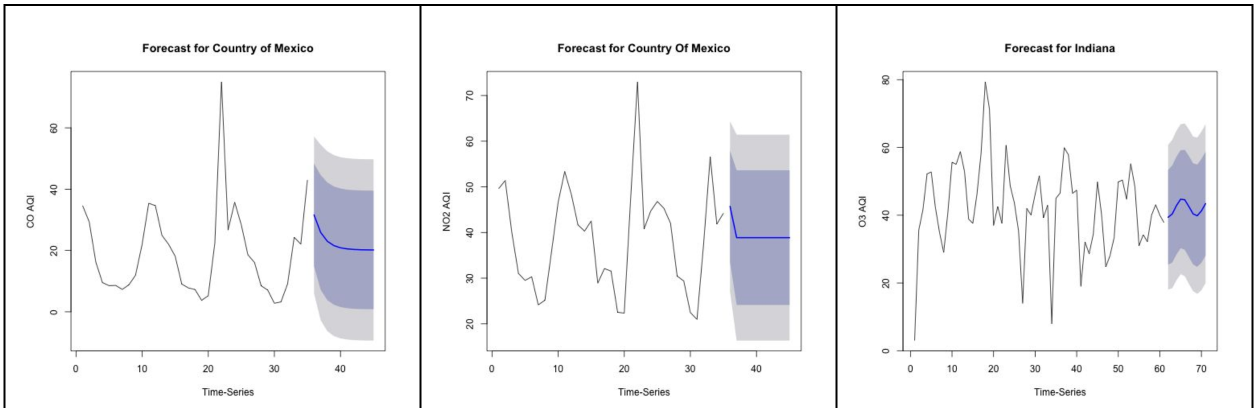
(c) O3

(d) SO2

Figure 6, Bar charts showing Distribution of Pollutants.

Data Mining

Since we have preprocessed our data set in the cleaning and exploratory phase. We can directly feed the processed data into R without worrying about duplicates, outliers, missing values, etc. Skewness is not relevant to us since we will be performing a time-series analysis. We used R's "forecast" package to build a regression model. We first selected the top three most polluted states for each type of pollutant. We then ran autoregressive integrated moving average (ARIMA) model on the selected states. We visualized the model but it was difficult to read as there were too many points. So, we perform averaging on the data from a monthly-basis to a yearly-basis which reduces the number of observation we have by approximately 12 times and applied ARIMA model again. The results can be seen in Figure 7.



(a) CO trends in Mexico

(b) NO2 trends in Mexico

(c) O3 trends in Indiana

Figure 7, ARIMA Model Line Plots.

Data Visualization

We decided to present our findings in a heat map as it could visualized multiple variables, AQI level of each state across different times. We came across choropleth which enables us to plot the U.S. states with a different color intensities. We then used plotly to make some interactive HTML pages. Initially, we wanted to make an interactive choropleth that we could easily navigate through the months with a scrollbar in plotly. However that feature requires upgrading to premium. Therefore, multiple HTML pages are used, each representing a different year.

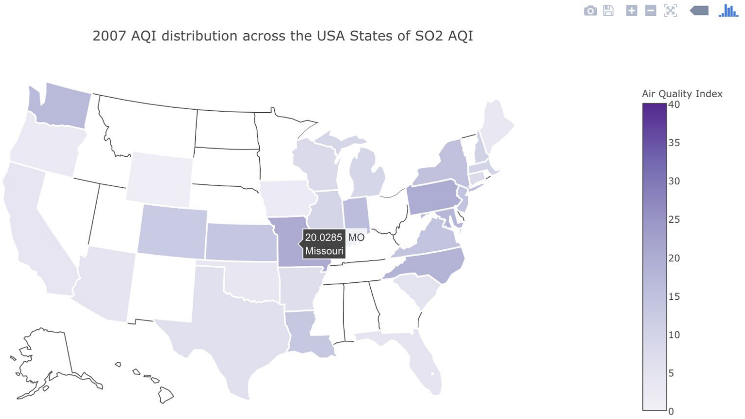


Figure 8, Plotly Choropleth HTML for Year 2007.

In conclusion, we were able to build a model to predict future trends at state level. We were also able to find a strong repetitive trend across the years for each pollutant which gives us the answer to our third question. We spent majority of our time in data cleaning and exploratory as compared to the data mining and data visualization stage. Majority of our time is spent on Data Cleaning and Exploratory Analysis. Data Mining was effortless, we tried several techniques and went with the best-looking one. The useful insights we found includes, the best time to visit the U.S. in general is during the beginning of Spring and Autumn. Second, the most polluted state is California followed by Pennsylvania…

Future Business Cases

We can perform a classification task instead of a time-series analysis. Currently, we could used the information obtained in Figure 5 and further zoom down to state level and explore as features for classification. However, additional data such as temperature, population, area of urban regions, area of greenland, etc for each year along with their respective classes to further improve the accuracy classification.

References

1. **Indiana’s Department of Workforce Development**
    - Smith, D. (2016, December 15). How the State of Indiana uses R and Azure to forecast employment. Retrieved February 07, 2017, from http://blog.revolutionanalytics.com/2016/12/state-of-indiana-employment.html

2. **Pennsylvania’s Department of Justice & Criminal forecast**
    - Can 'predictive policing' prevent crime before it happens? (2016, November 16). Retrieved February 07, 2017, from http://www.sciencemag.org/news/2016/09/can-predictive-policing-prevent-crime-it-happens
    - Can 'predictive policing' prevent crime before it happens? (2016, November 16). Retrieved February 07, 2017, from http://www.sciencemag.org/news/2016/09/can-predictive-policing-prevent-crime-it-happens
    - Mohler, G. O., Short, M. B., Brantingham, J. P., Schoenberg, F. P., & TITA, G. E. (2010, October). Self-Exciting Point Process Modeling of Crime. Retrieved February 7, 2017, from http://www.math.ucla.edu/~mbshort/papers/crime3.pdf

3. **Weather Forecast**
    - White Paper: Precision Weather Forecasting | AcuRite. (n.d.). Retrieved February 07, 2017, from https://www.acurite.com/precision-forecasting-white-paper/precision-forecasting-white-paper
    - Dutton, J. A., 2002: Opportunities and priorities in a new era for weather and climate services. Bull. Amer. Meteor. Soc, 83, 1303–1311.

- Sigrist F, Kunsch HR, Stahel WA (2012) SPDE based modeling of large space-time data sets. http://arxiv.org/pdf/1204.6118v4.pdf.

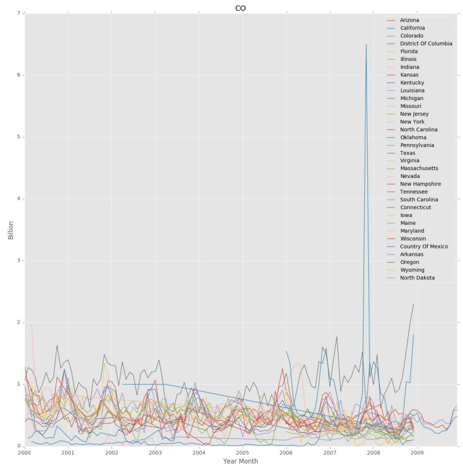
4. **Forecast (Everything)**

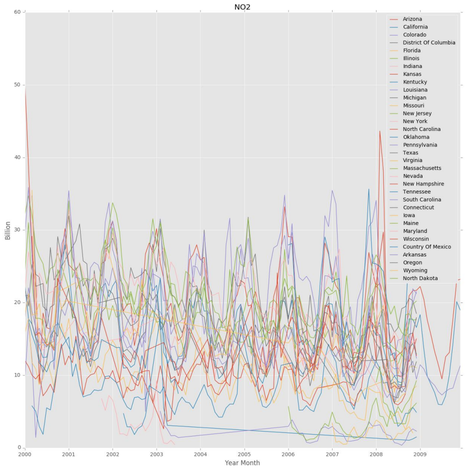
- Hassani, H., & Emmanuel, S. S. (2015, April 10). Forecasting with Big Data: A Review. Retrieved February 07, 2017, from http://link.springer.com/article/10.1007%2Fs40745-015-0029-9

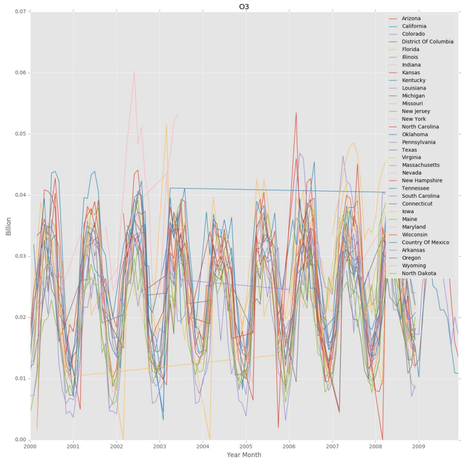
5. **Data Mining to combat Natural Disasters**

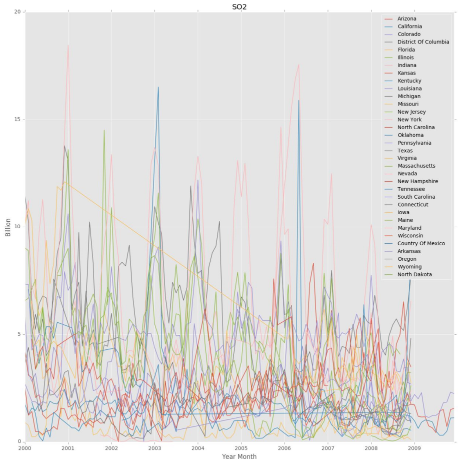
- Goswami, S., Chakraborty, S., Ghosh, S., Chakrabarti, A., & Chakraborty, B. (n.d.). A Review on Application of Data Mining Techniques to Combat Natural Disasters. Retrieved February 7, 2017, from https://arxiv.org/pdf/1610.09974.pdf

Appendix I, Messy Plots.


CO
Billion
Year Month
Arizona
California
Colorado
District Of Columbia
Florida
Illinois
Indiana
Kansas
Kentucky
Louisiana
Michigan
Missouri
New Jersey
New York
North Carolina
Oklahoma
Pennsylvania
Texas
Virginia
Massachusetts
Nevada
New Hampshire
Tennessee
South Carolina
Connecticut
Iowa
Maine
Maryland
Wisconsin
Country Of Mexico
Arkansas
Oregon
Wyoming
North Dakota


NO2
Billion
Year Month
Arizona
California
Colorado
District Of Columbia
Florida
Illinois
Indiana
Kansas
Kentucky
Louisiana
Michigan
Missouri
New Jersey
New York
North Carolina
Oklahoma
Pennsylvania
Texas
Virginia
Massachusetts
Nevada
New Hampshire
Tennessee
South Carolina
Connecticut
Iowa
Maine
Maryland
Wisconsin
Country Of Mexico
Arkansas
Oregon
Wyoming
North Dakota


O3
Billion
Year Month
Arizona
California
Colorado
District Of Columbia
Florida
Illinois
Indiana
Kansas
Kentucky
Louisiana
Michigan
Missouri
New Jersey
New York
North Carolina
Oklahoma
Pennsylvania
Texas
Virginia
Massachusetts
Nevada
New Hampshire
Tennessee
South Carolina
Connecticut
Iowa
Maine
Maryland
Wisconsin
Country Of Mexico
Arkansas
Oregon
Wyoming
North Dakota


SO2
Billion
Year Month
Arizona
California
Colorado
District Of Columbia
Florida
Illinois
Indiana
Kansas
Kentucky
Louisiana
Michigan
Missouri
New Jersey
New York
North Carolina
Oklahoma
Pennsylvania
Texas
Virginia
Massachusetts
Nevada
New Hampshire
Tennessee
South Carolina
Connecticut
Iowa
Maine
Maryland
Wisconsin
Country Of Mexico
Arkansas
Oregon
Wyoming
North Dakota