DSC 465 – Data Visualization Winter2023
Group Name: Meaningful Patterns

Final Project Report

United States Air Pollution Analysis

Team Members – Sadiya Amreen, Megh Thakkar, Jashwanth Neeli, Vatsal Parikh, Bramhashree Raghava Pillai Manoharan

Table of contents

Introduction

The dataset was taken from Kaggle and focuses on air pollution levels in the United States from 2000 to 2016. It comprises measurements of four significant air pollutants, namely carbon monoxide, sulfur dioxide, nitrogen dioxide, and ozone, which have been identified as hazardous to both human health and the environment. The United States Environmental Protection Agency (EPA) collected this data from over 4,000 monitoring stations across the country, and it provides hourly, daily, and annual averages for each pollutant. This dataset is a valuable resource for researchers, policymakers, and the general public interested in understanding the effects of air pollution on human health and the environment.

Air pollution is a pressing concern globally and arises from multiple factors such as burning fossil fuels, the greenhouse effect, and industrial activity. The Air Quality Index (AQI) is the standard tool used to assess air quality in the atmosphere. AQIs from different monitoring stations are crucial in comprehending the effects of air pollution on human health and the environment.

Our team utilized a range of visualizations, including interactive ones, to explore the similarities and differences in pollutant levels across various States in the United States. Through these visualizations, we aimed to gain deep and meaningful insights into how the fluctuation in AQI across the United States can be leveraged to develop effective strategies that can minimize pollution levels, improve air quality, and ultimately lead to healthier living conditions for the public.

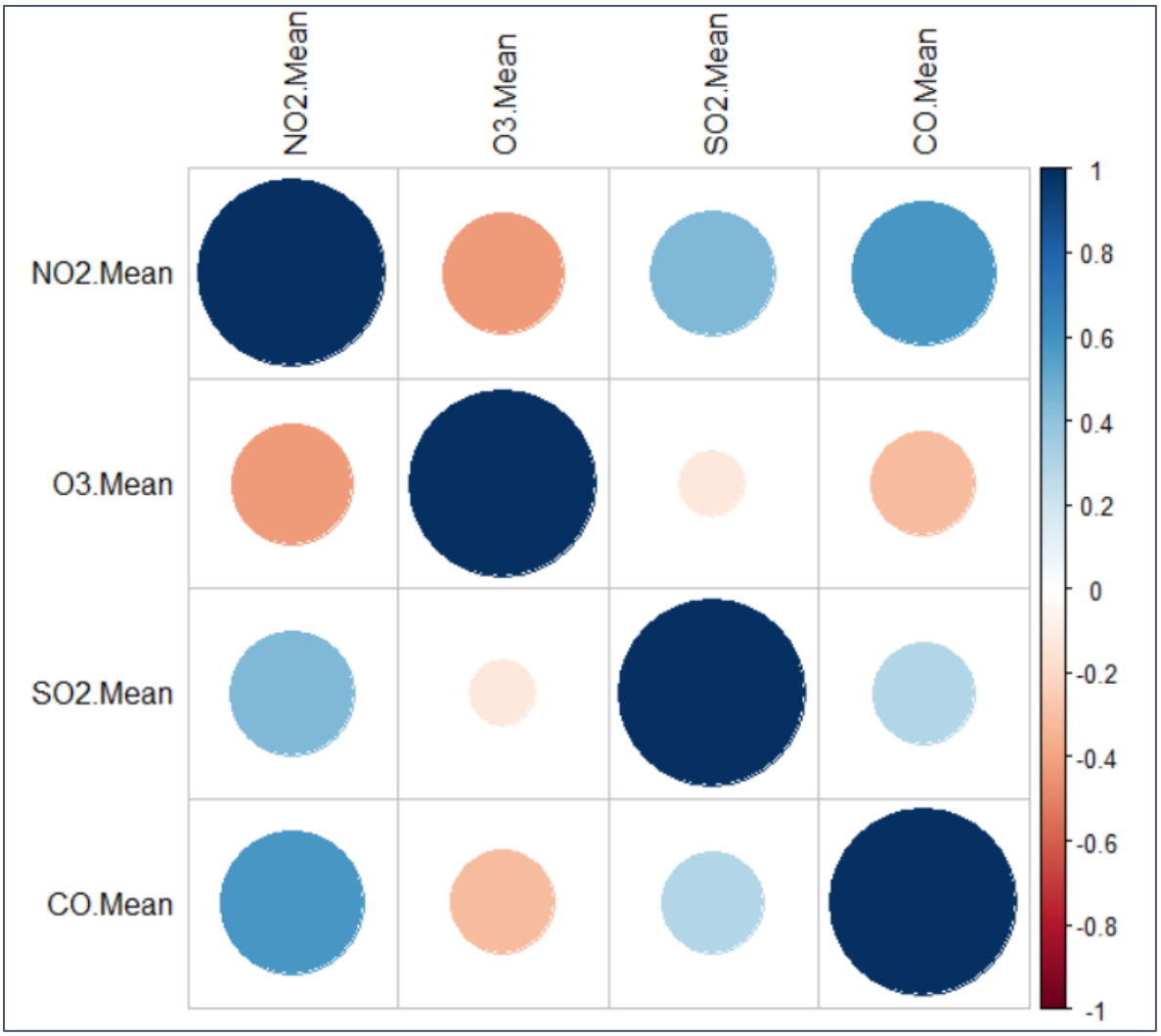
Data:

Data **Pre-Processing**

The CSV file was imported into R Studio, and the summary statistics were reviewed to obtain an overview of the data. The dataset covered 47 states, 133 counties, and 144 cities in the United States, with 28 columns and 1.7 million rows. It contained categorical (Region), continuous (Temperature, Day, Month, Year), and geographical (Country, City) variables. To ensure accuracy, the data was cleaned by identifying and removing unnecessary variables, checking for and removing null values, and converting the AQI readings from PPM to PPB to maintain consistent units throughout the dataset and visualizations. The date column was separated into date, month, and year to aid visualizations. After pre-processing the data, it was exported and utilized to generate visualizations in R and Tableau.

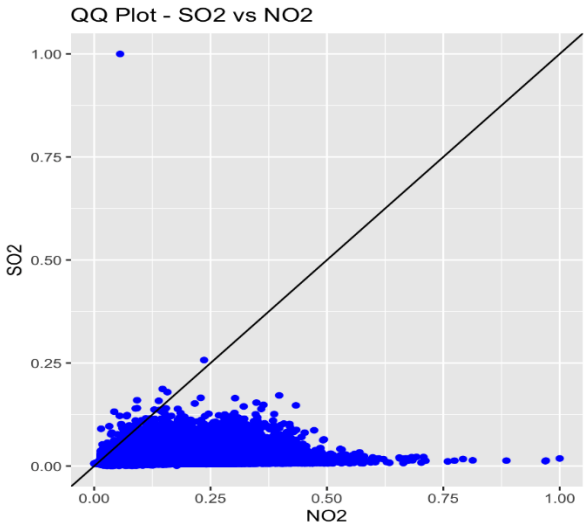
Exploratory Analysis:

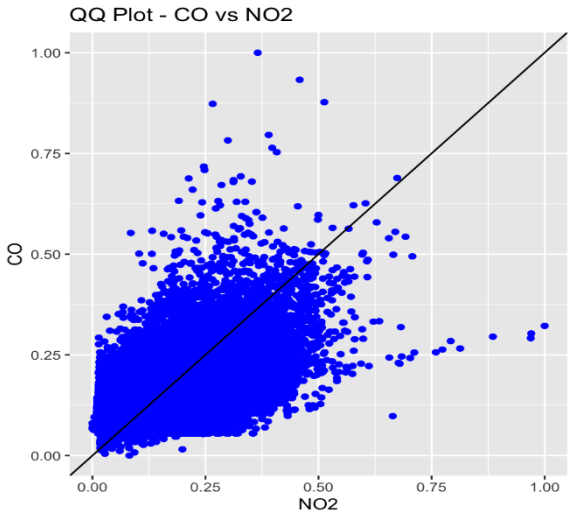
Correlation Matrix

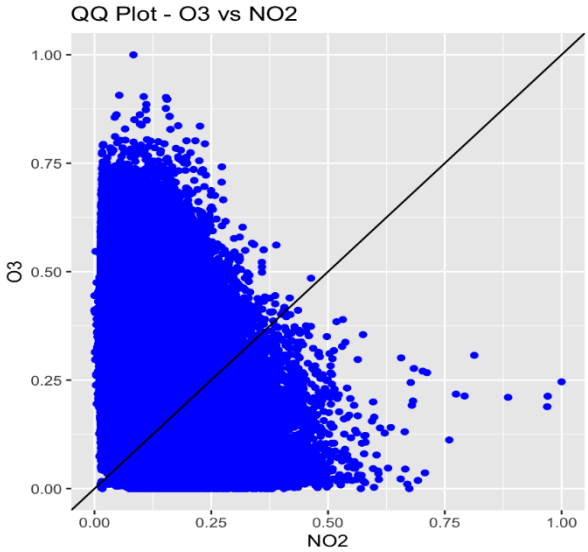


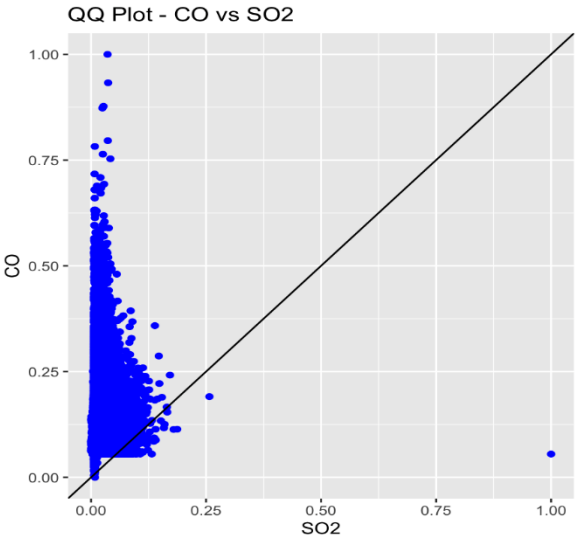
Based on the heat map or correlation matrix, we can infer that there is a significant correlation between the variables NO2.mean and CO.mean with a correlation coefficient of 0.6. This indicates that NO2.mean is dependent on CO.mean and changes in CO.mean are likely to have an impact on NO2.mean. However, the remaining independent variables appear to maintain a balanced relationship with the dependent variables, without exerting a significant influence on each other. Therefore, it is important to carefully consider the relationship between NO2.mean and CO.mean when analyzing the data, while also taking into account the other independent variables and their potential impact on the dependent variables. By doing so, we can gain a more complete understanding of the complex relationships within the data and make more accurate predictions or decisions based on the results.

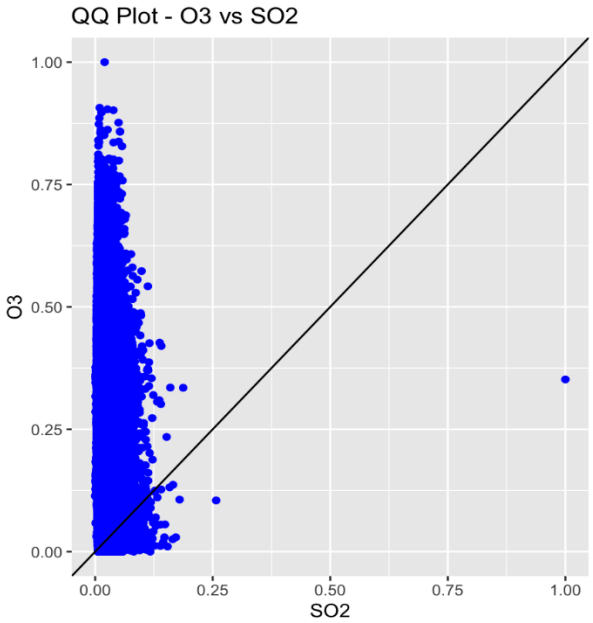
QQ Plots

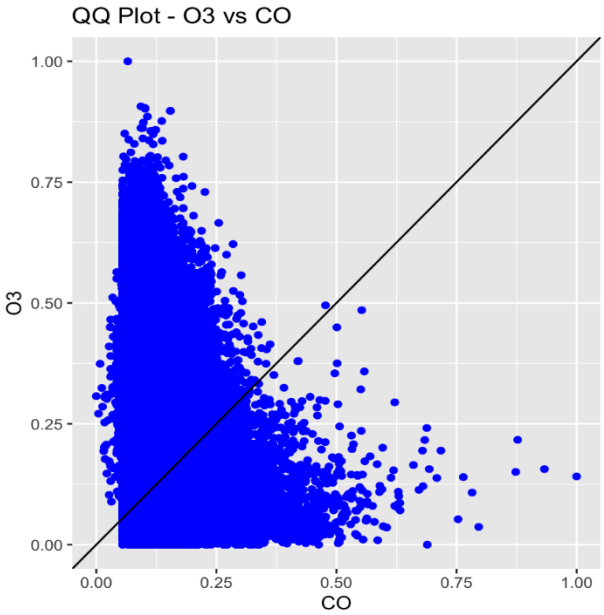

QQ Plot - SO2 vs NO2
SO2
NO2
0.00
0.25
0.50
0.75
1.00


QQ Plot - CO vs NO2
CO
NO2
0.00
0.25
0.50
0.75
1.00


QQ Plot - O3 vs NO2
O3
NO2
0.00
0.25
0.50
0.75
1.00


QQ Plot - CO vs SO2
CO
SO2
0.00
0.25
0.50
0.75
1.00

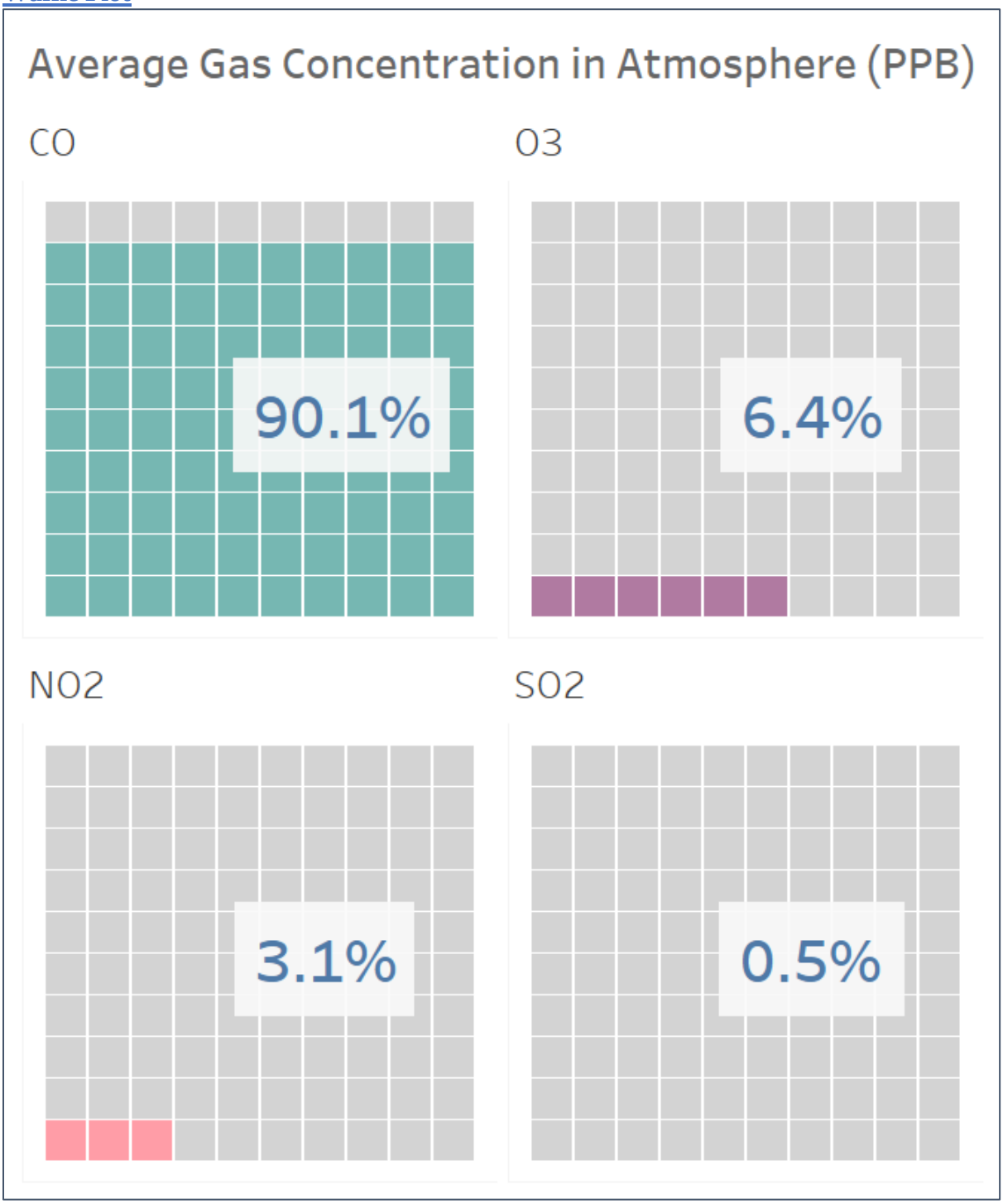




We plotted the QQ plots for all the pollutants to find if there is any relationship between the pollutants. After analyzing the graphs we found that the scatterplot of NO2 vs CO seems to be more fitting to the line which shows normal distribution.

Visualizations

Waffle Plot



The Waffle Plot is a visualization that illustrates the total concentration of each gas present in the atmosphere. According to the data, Carbon Monoxide is the gas with the largest portion, followed by Ozone, Nitrogen Dioxide, and Sulfur Dioxide. Each of these gases has a distinct impact on the environment.

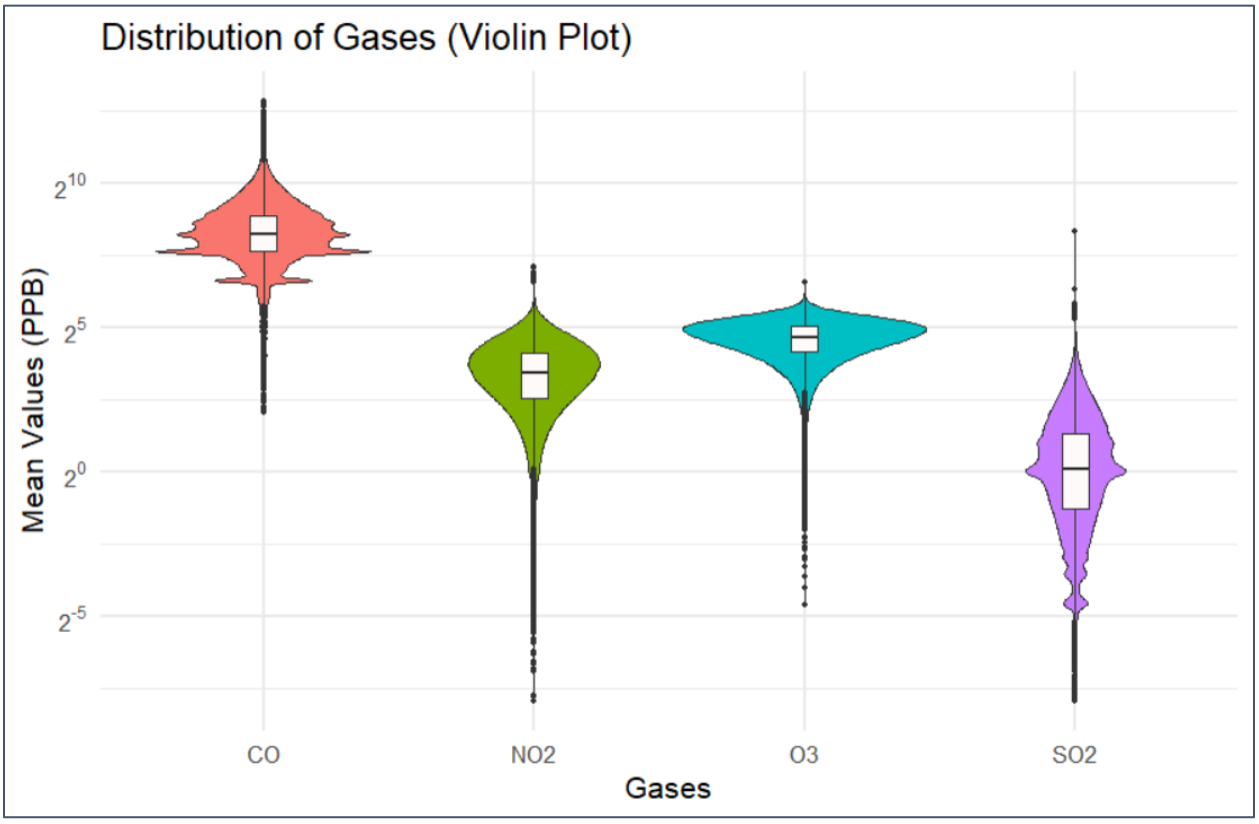
Carbon Monoxide, for example, has a direct correlation with the amount of greenhouse gases in the atmosphere, which are responsible for climate change and global warming. Emissions of this gas into the atmosphere can have a severe effect on the environment.

Ozone and Nitrogen Dioxide, on the other hand, are known to be highly damaging to vegetation and ecosystems. Even though the proportion of these gases is lower than Carbon Monoxide, their impact can still be quite significant.

Finally, while Sulfur Dioxide's proportion may be the smallest among the four gases, it is still highly toxic. It is primarily produced from industrial emissions and vehicular exhausts and can combine with water to form acid rain. This gas also produces particulate matter, which is a significant air pollutant.

Overall, the Waffle Plot provides valuable information about the composition of the atmosphere and the impact of each gas. It underscores the importance of reducing emissions to mitigate the adverse effects on the environment.

Violin Plot



The Violin Plot is a commonly used type of graphical representation to display the distribution of gases, allowing for the visualization of data spread and concentration.

In this report, both the Violin Plot and the Box Plot were utilized to show the concentration of values and outliers for different gases. The Box Plot is a statistical visualization that displays the distribution of a dataset based on quartiles, where the box represents the middle 50% of the data, the whiskers represent the range of the data excluding outliers, and the dots or circles outside the whiskers represent outliers.

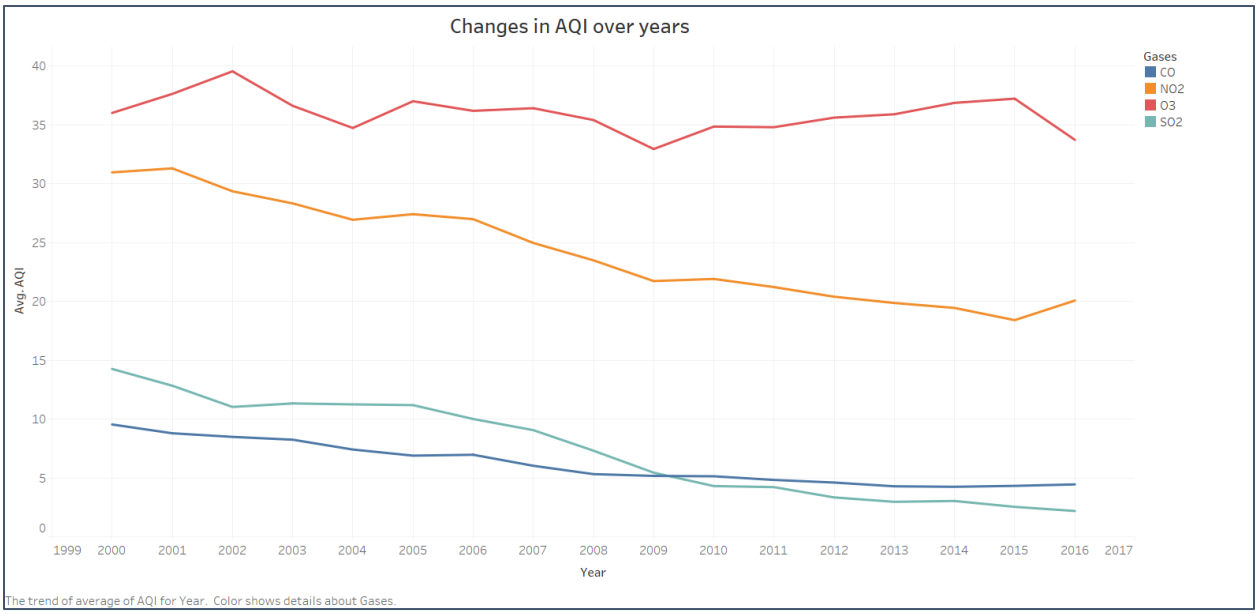
The Violin Plot, on the other hand, was used to show the concentration of values for different gases. However, due to the high average mean value for Carbon Monoxide, the y-axis was transformed by taking the logarithm of the mean values.

Based on the analysis of the Waffle Plot, it can be observed that Carbon Monoxide has the highest concentration, while the other three gases only contribute a small percentage to the total gas concentration in the atmosphere. The Box Plot confirms that lower values are outliers, but interestingly, this does not seem to be the case for the upper values. In fact, most of the values are concentrated in the upper range for each gas.

Therefore, even though the average value for these gases might be low, it can be inferred from the plot that the concentration of these gases is still high in the atmosphere, with some outliers in the lower range of values. This emphasizes the need to monitor and control the emissions of these gases to reduce their impact on the environment.

The Violin Plot also reveals that most gases are concentrated in the upper range of values, as evidenced by the bulge. However, one notable exception is Sulphur dioxide, which displays a relatively uniform range of values from around 2^{-5} PPB to around 2^{5} PPB. It is important to note that the plot uses logarithmic transformation, which means that the differences in range between gases may not be immediately apparent. For example, while the range of Sulphur dioxide may appear smaller than that of Carbon monoxide, which goes from around 2^{5} PPB to 2^{10} PPB, the logarithmic scale makes the difference much more pronounced.

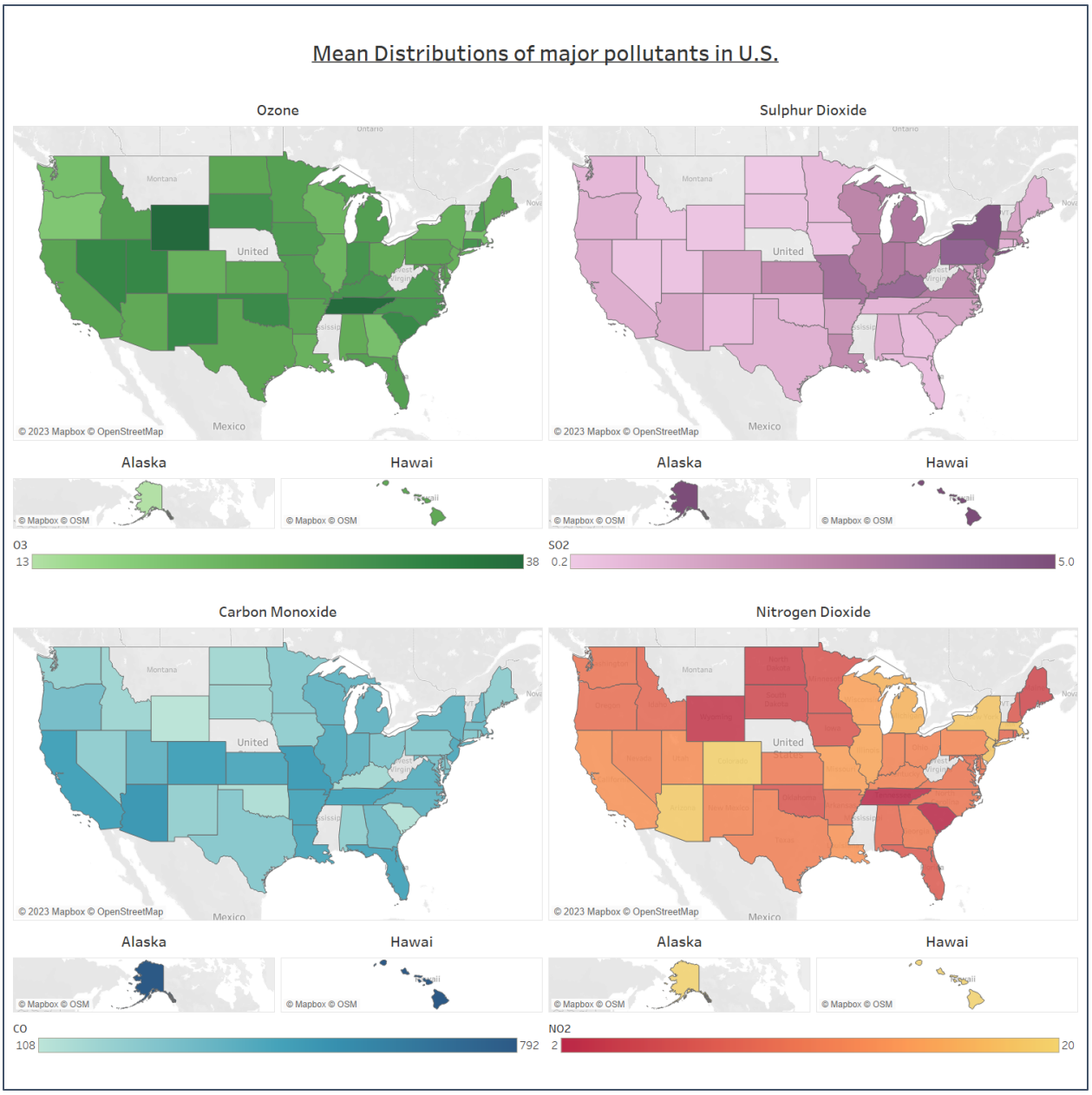
Line Graph



A line graph was utilized in this study to analyze the changes in AQI (Air Quality Index) over the years for the four major pollutants. The Tableau interface was employed to derive this visualization. A pivoted column for the AQIs of all four gases was created for the analysis. The variable "Year" was mapped on the X-axis, while the AQIs of all the gases were mapped on the Y-axis, with different colors assigned to different gases.

The line graph revealed that Ozone had the highest emission among all pollutants over the years, while Sulphur dioxide gradually decreased from 2000 to 2016, and nitrogen dioxide also decreased gradually during the same period. Furthermore, Carbon monoxide appeared to be the least effective gas in terms of emission in the year 2000, while Sulphur dioxide appeared to be the least polluting gas by the end of 2016. These observations are consistent with the trends and patterns observed in the line graph visualization.

Choropleth



Type: Choropleth Dashboard

The United States map was visualized using States generated latitude and longitude in Tableau to create a Choropleth dashboard illustrating the mean distribution of the four major pollutants in the U.S. over a 15-year period. The visualization employed the variables Latitude, Longitude, and the mean of each pollutant. Calculated fields were created for Carbon Monoxide and Ozone to standardize their units to parts per billion, in line with the units of Nitrogen dioxide and Sulphur dioxide. This was achieved by multiplying the values of Carbon Monoxide and Ozone by 100.

Subsequently, a dashboard of the average distribution of the pollutants was created side by side. A consistent colour scheme was utilized for all four pollutants, allowing for a clear overview of a

particular State. For example, Alaska was observed to be highly polluted with Sulphur dioxide compared to other States in the U.S., while having the least levels of ground-level Ozone when compared to other States. Concentrations of Nitrogen dioxide and Carbon Monoxide were found to be at mid-levels.

The visualization revealed that the mean distribution of Carbon Monoxide is high compared to other gases in the U.S., while the mean distribution of Sulphur dioxide is low compared to other gases in the U.S., in agreement with the Waffle plot. A year filter was also added, allowing the user to filter and view the average distribution of the pollutants for a specific year.

Multiple drafts of the visualization were executed, employing different colour schemes. However, the above visualization was selected as it conveyed meaningful information with appropriate colours and fit well within the context of our story.

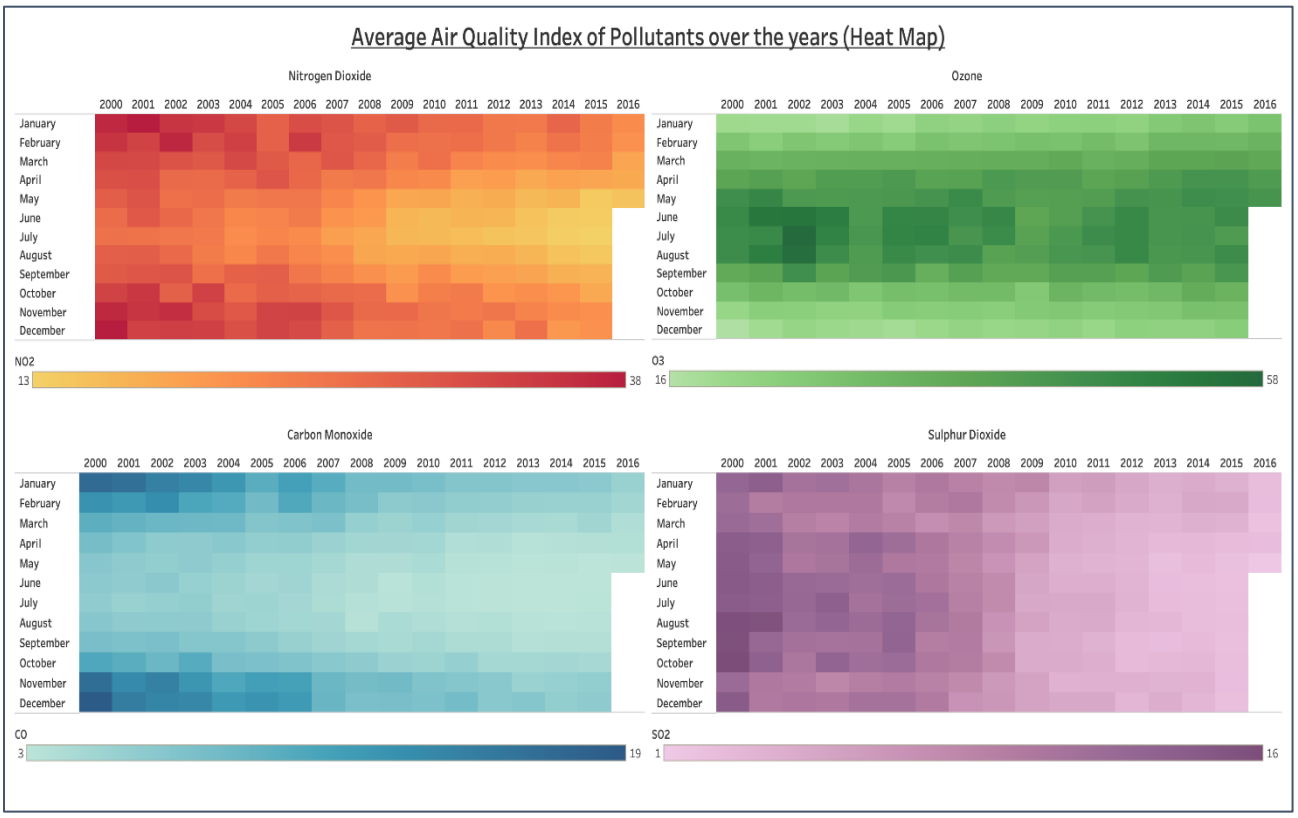
Hex tile Cartogram

Type: Hex Tile Cartogram Dashboard

To conduct a more detailed analysis, a Hex tile cartogram was generated to portray the average Air Quality Index (AQI) of the pollutants in the US. So the concentrations of each pollutant depicted in the choropleths has lead to the degradation of the air quality index which is depicted

in this visualization. This visualization offered a better perspective on the AQI in each state, as the states were depicted uniformly in size and there was no inconsistency in examining smaller states. Notably, the states with major urban centers, such as New York, Chicago, and Los Angeles, exhibited higher levels of pollution, as evidenced by their elevated average AQI values for each pollutant. Our analysis revealed that over the years, the average AQI values of Ozone and Nitrogen dioxide had reached concerning levels, while the pollutants SO2 and CO remained at levels that did not pose a significant health risk. It is worth noting that while the AQI values do not appear to be rising currently, they had peaked in past years, as depicted in greater detail in the subsequent heat map.

Heat Map



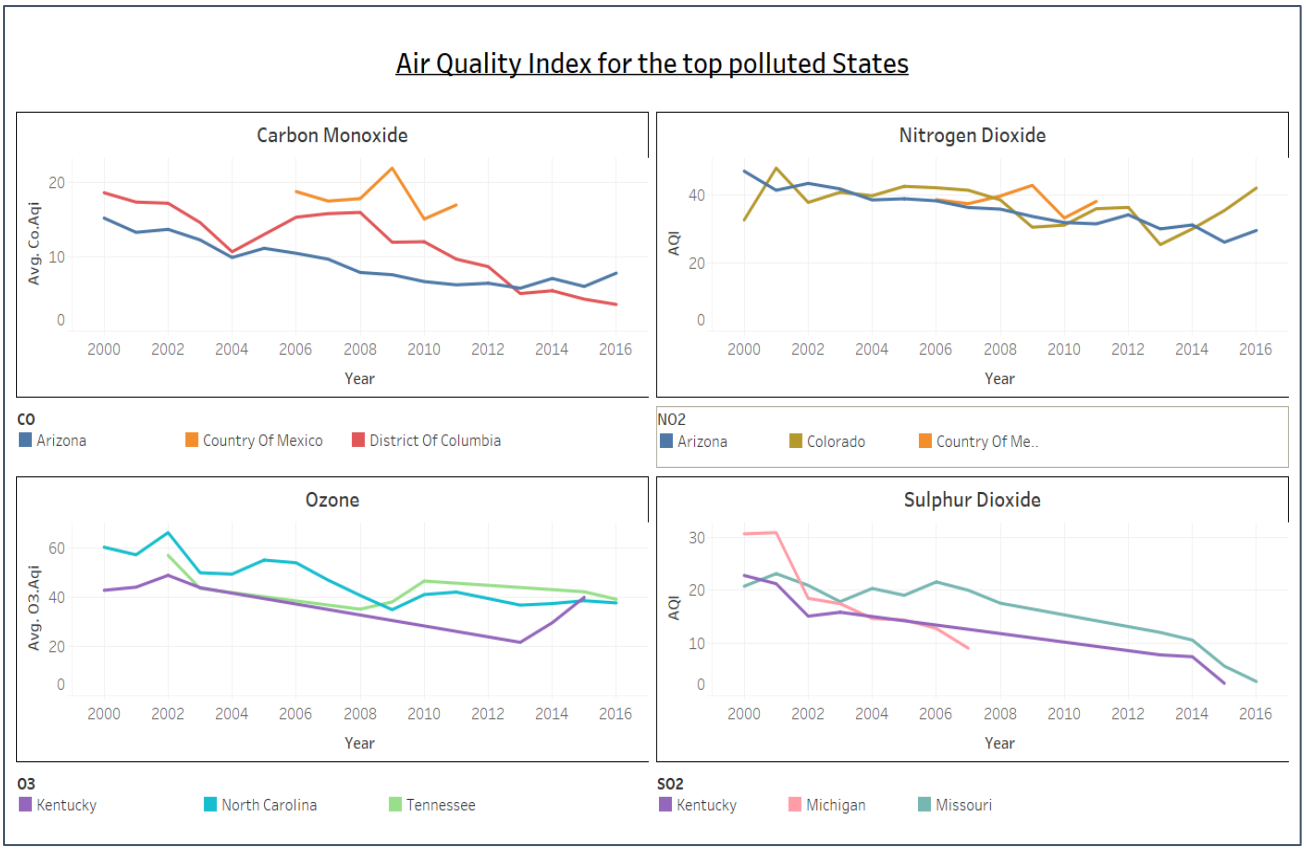
Type: Heat Map

Comprehending the Air Quality Index (AQI) of the pollutants over time is facilitated by the heat map. In order to thoroughly analyze the pollution trends, it is crucial to identify when the AQI tends to deteriorate or improve.

To begin with, for Ozone (O3), the AQI starts to worsen from April and continues through September, which is the summer season in the US. However, from the fall season, it begins to improve. Regarding Sulphur dioxide (SO2), the year 2000 marked the worst AQI. However, after 2005, the heat map indicates a gradual improvement in the AQI for SO2 year after year.

As for Nitrogen dioxide (NO2), the AQI typically begins to deteriorate in October and persists through February. Nevertheless, the AQI for NO2 gradually improves over time while continuing to be quite poor for the months mentioned. Carbon Monoxide (CO) AQI remains at its worst from November to February. Nonetheless, starting in 2006, the AQI for CO gradually improves.

Line Graph



To conduct an analysis for the Top polluted States in the U.S., a dashboard of line graph for the four major pollutants have been generated to portray the average Air Quality Index (AQI) of the pollutants in the US over the years. Even though the data is not consistently available for all the years, we see a downward trend for most of the course of the gases. It's evident from the Carbon Monoxide line graph that Country of Mexico shows trends only for the span of 2006- 2011, and if we have more data, it can be clearly assumed that this State would have been the highest pollutant State for Carbon Monoxide. Arizona's weather conditions seems worsening as it appears to be on the top most list for both Carbon Monoxide and Nitrogen dioxide. Kentucky seems to follow the pattern of worsening weather conditions with abundance of Ozone and Sulphur dioxide.

Lastly for Sulphur dioxide the values seems to be decreasing in all the top States, even though it's a good indication but even minute concentrations of Sulphur dioxide are for concerning levels as its readily reactive to water and forms acid rain.

Conclusion

In conclusion we would like to highlight the findings based on our analysis. Carbon monoxide is one of the greenhouse gases with the highest concentration in the atmosphere. It was found that nitrogen dioxide and ozone affect the air quality index more compared to other gases. Although Sulphur dioxide has the least overall concentration in the atmosphere, it is more harmful than the other gases and affects the atmosphere in the most severe way. There was a lot of missing values in the data and felt we could have got deeper insights if there were more data. Carbon monoxide and Sulphur dioxide has decreased over the year. We also noticed a pattern where Ozone usually gets higher during the summer.

Individual Report:

Vatsal Parikh:

To start, I searched for a suitable dataset among many interesting options, and ultimately our team chose the United States Air Pollution dataset. We then began preprocessing the data, during which I recommended splitting the date column into multiple columns and suggested several visualization techniques for exploratory analysis.

Additionally, we cleaned the data, and I created a waffle plot using Tableau, which required me to create a custom dataset and merge it with our main data. Though I was not very familiar with Tableau, I found the process to be fascinating and I learned a lot. I also created a box plot using Tableau, which initially presented a challenge in displaying four box plots in one visualization. However, I eventually discovered that I needed to pivot the values and add a reference line. Lastly, I created a violin plot using RStudio, as I found it too difficult to use Tableau. After presenting the plot to the professor, I made some changes to show the outliers. I also contributed to creating the overall story for the project and feel that this project allowed me to improve significantly in a short amount of time.

What I learned from the subject

This course covered the principles and techniques for creating effective visual representations of data. I learned how to identify the most appropriate types of charts and graphs to use for different types of data, and how to design these visualizations in a way that is easy to understand and aesthetically pleasing. I also learned the use of various tools and software for creating data visualizations, such as Tableau and R Studio. I learned how to import and clean data, manipulate it, and then create visualizations using these tools.

There are many types of visualizations, including line charts, bar charts, pie charts, histograms, scatter plots, and heat maps. The course taught me how to create these types of visualizations and when to use each of them. I also learned how to communicate insights from data effectively. The course taught me how to create a narrative and tell a compelling story with my visualizations, focusing on the most critical insights.

In conclusion, this course provided me with valuable skills in creating effective and compelling visualizations that can now help me understand and communicate insights from data. With these

skills, I can make better-informed decisions based on data and communicate my findings to others more effectively.

Bramhashree Raghava Pillai Manoharan

In this data visualization project, I am the representative of the team Meaningful Patterns. Finding the perfect dataset matching all our requirements was quite challenging as I began to search various websites and explored different types of datasets. We as a group, were able to finalize a dataset that we found on the Kaggle website. The dataset was based on the US Air Pollution which focused on four major pollutants emitted all over the dataset of the USA from 2000 – 2016. We imported the dataset in R and began to preprocess the data. We looked at the summary of the dataset and check for null values in the dataset. As the dataset was huge, we dropped the null values. We then computed the Day, Month and Year values from the date column for the purpose of tine series analysis. We noticed that there were inconsistencies in the units of the gases, and hence we converted the units in a uniform manner. I plotted the QQ plots for all the pollutants and found that the scatterplot of NO2 vs CO seems to be more fitting to the line which shows normal distribution. The data was cleaned and exported to an excel sheet which was used to generate visualizations.

As for the visualizations firstly, I created a choropleth which depicted all the states in the US and the mean distribution of the four pollutants from 2000 to 2016. The fields used were latitude, longitude, states, mean of the pollutants. Once the graph is created for each pollutant, I then put all of the four graphs together in a dashboard so that the user can have a side by side visualization of the distribution of all the gases at once. Since Alaska was too big and dominated the US map, and Hawaii was too small to look at, we cropped both the states and pasted it below the US map, which allowed us to zoom in, displaying the states of the US clearer.

Next to show the impact of the pollutants on the air quality index, I plotted a hex tile cartogram. To create the hex tile cartogram, I first downloaded states excel sheet with which I linked my dataset creating a relationship (this.state = that.state). I also downloaded a hex shape icon and placed it in the custom image folder and then incorporated the shape into the map. Once the hex tile cartogram was created for each pollutant, I then created a dashboard putting together all the graphs. More care was taken to depict each pollutant in different colours and the legends were placed along with each pollutant's graph.

What I learned from the subject

I learnt about various visualization techniques, tools, and how to use them in an effective manner so that the audience understands the message that I am trying to convey with ease. This subject changed the way I see a visualization, the details that a graph carries and basically how to decode a visualization. It also changed my perspective of a good graph, where I used to give more importance to beautification of the visualizations where I add a lot of colours, logos, icons, etc. This subject made me understand the importance of elimination of chart junks and the importance of axes and that the values had to follow a proper scale. It made me realize that even minute details like spacing could make the audience interpret the graph in a different way. I feel confident now in making visualizations and dashboards as I know what are the attributes that has to be kept in mind to present a visualization in the most efficient way. Professor Eli Brown has been a great

mentor for a lot of students including me and I learned a lot from him in this subject. He gave constructive feedback during each stage of the project which helped us in creating the visualizations and subsequent analysis. Lastly, I had wonderful teammates and all of them were active in the project and did an excellent job.

Jashwanth Neeli

In this visualization project, I have searched good data set and that was quite a challenging task, exploring and researching from different analytical websites such as kaggle, us pollution site, tableau public etc. Finally, We were able to complete a dataset that we discovered on the Kaggle website as a group. The dataset was based on the US Air Pollution dataset, which focused on four major pollutants emitted throughout the dataset from 2000 to 2016. We loaded the dataset into R and started preprocessing it. I have worked using the tools of R and tableau to visualize the graphs and the outcomes. I have imported the data set in R for the preprocessing, using the summary the dataset was total 29 columns and 1.7 million records. We removed the null values because the dataset was so large. We then calculated the Day, Month, and Year values from the date column for time series analysis.

When it comes to visualization, I have created a line graph for the air quality index for the top polluted states, I have compared with 4 different gases NO2, SO2, CO and O3 for top polluted states over all the years. For that firstly, I made sure that all the air quality index values have average values year wise from 2000 to 2016. Additionally, I also created a line graph for changes in AQI over year which I have used in the exploratory Data analysis. In the line graph I have compared the 4 air quality gases over the time period on the basis of how it is changing, which is the least and more hazardous. I have learned how to use a dashboard using that I have pasted all the gases in a graph in an interactive manner and the color scaling for the graph, also I was able to which graph to choose for a particular topic, for example here since I am using time series so, I have used a line graph and I clearly showed the legends under each graph. This course has taught me how to use different kinds of graphs for its own use, also I am more exposed in using tableau now, I create many things using this. On top of this I have learned all the visualization concepts very well all because of my prof. Eli T brown. His feedback for my homeworks have helped me to crosscheck my mistakes.

Megh Jikeshkumar Thakkar

In this data visualization project, I am the representative of the Team Meaningful Patterns. As team we searched few datasets on internet and after searching, we shortlisted 5-6 different datasets and at the end we finalized the US pollution data as our final dataset. To start with, I worked with Data pre-processing for the dataset. In the dataset, there were total 29 columns in the dataset. I imported the dataset in R studio environment, first checked with total number of null rows and omitted them in R itself. Before omitting the null rows there were 1.7 million records but now it has 438,876 records in the pre-processed dataset. Then after, removed unwanted columns like State code, Street Address, County code and many more columns. The current pre-processed dataset has 23 columns.

Then converted the Date.Local variable into month, day and year, which means the Date.Local variable has the format like this “2000-11-25” and converted in 3 different columns of month, day

and year. For some pollutants the units Then converted the units of pollutants from parts per million to parts per billion. In that applied the formulas to convert millions into billions using the formula of it. After the pre-processing the data, I created the Heat map dashboard for the pollutants. To start with, first create heat map for each pollutant in different sheet. In it, x-axis as year and y-axis as month, so after implemented this thing for each pollutant, given a title for each map. However, after created 4 different heatmap and merged all heat map in one dashboard. From the dashboard, you can infer the pattern of each pollutant from the heat map.

What I learned from the subject

This was one of the interesting courses which I learned, in it I have learned new and different types of visualization techniques, concepts and tools. Learn about the things which I never heard but I was always seeing around me which was chart chunk, distortion, and clutter. Learned new tools like Tableau, infect the tableau is very powerful tool to do visualization and it is easy to learn and professor made it easy to learn. The course helps me to learn differences in the colors, which colors should be use in visualization with help color wheel, hue and saturation. I also learned each visualization should be clear and should convey the message clearly. Moreover, I learned to create interactive dashboards with different filter in it. Professor Eli Brown has been a great mentor for all students including me and learned lot from him in this course. Furthermore, working in the group project help me to learn new things in course, brainstorm and come up with new idea for different types of visualization.

Sadiya Amreen

It was difficult to find the ideal dataset that met all our requirements as I started to browse other websites and investigate various dataset. As a group, we were successful in completing a dataset that we had discovered on the Kaggle website. I played a significant role in organizing group meetings and project planning. I plotted the correlation matrix for all the pollutants and looked for any dependencies between the gases. We deduced that Nitrogen dioxide and Carbon monoxide are related, but the other gases are not. I shared some tips on data preprocessing to get rid of extra variables.

I created a line graph of the Air quality Index over time for each of the four primary pollutants for the main visualizations. Also, I contributed to the structuring of the final report by writing the Introduction, the data pre-processing section, and the description for the line graph of the AQI over time.

What I learned from the subject

I acquired knowledge about a variety of visualization methods and tools, and the main concept of data, message and audience became very clear to me after working on this project. Many misconceptions of my perspective with regards to visualization were changed after working on this project as how to employ them skillfully so that the audience may easily comprehend the point I am trying to convey. It also altered my view of what combines for a good visual, which included adding plenty of detail and irregular color scheme as my primary mistake. This project helped me to comprehend the value of removing clutter, the requirement that values adhere to an appropriate scale, the significance of axes, where to round off the values and where not to.

It made me understand that even minor features like improper titles can cause the audience to read the graph incorrectly. I am now more confident in creating visualizations with the Tableau interface since I understand the details that must be considered in order to display a visual. Professor Eli Brown has been an excellent mentor to us, and I have learned a lot from him in this class. He provided positive feedback at every stage of the project, which aided us in the creation of the visualizations and subsequent analysis. His work on R studio libraries was commendable. I had fantastic group mates who gave their best in the project and made a collaborative effort for the success of the project.

Appendices:

Code:

Preprocessing and Correlation Matrix:

```
library(tidyverse)
library(ggplot2)
library(reshape2)
library(Hmisc)
library(GGally)
library(corrplot)

uspollution_pollution_us_2000_2016 <- na.omit(uspollution_pollution_us_2000_2016)
maindata <- uspollution_pollution_us_2000_2016
maindata <- maindata[,-c(1,2,3,4,10,15,20,25)]

maindata$date <- format(as.Date(maindata$Date.Local,format="%Y-%m-%d"), format = "%d")
maindata$month <- format(as.Date(maindata$Date.Local,format="%Y-%m-%d"), format =
"%m")
maindata$year <- format(as.Date(maindata$Date.Local,format="%Y-%m-%d"), format = "%Y")
maindata <- maindata[,-c(1)]

#converting units from ppb to ppm
maindata$NO2.Mean <- maindata$NO2.Mean/1000
maindata$NO2.1st.Max.Value <- maindata$NO2.1st.Max.Value/1000
maindata$SO2.Mean <- maindata$SO2.Mean/1000
maindata$SO2.1st.Max.Value <- maindata$SO2.1st.Max.Value/1000

#extracting data
write.csv(maindata, "E:\\DePaul university\\Data Visualization\\maindata.csv",
row.names=FALSE)
summary(maindata)

ggpairs <- ggplot(data = melt(round(cor(maindata[,c(5,9,13,17)]))), aes(x=Var1, y=Var2,
fill=value)) + geom_tile()
ggpairs
```

```
#
maincorr1 <-maindata[,c(6,10,14,18)]
cor(maincorr1)

maincorr1.cor = cor(maincorr, method = c("spearman"))
maincorr1.cor

maincorr1.rcorr = rcorr(as.matrix(maincorr1))
maincorr1.rcorr

corrplot(maincorr1.cor,tl.col = 'black')

#working on the means of the gases
meancorr <-maindata[,c(5,9,13,17)]
cor(meancorr)

meancorr.cor = cor(meancorr, method = c("spearman"))
meancorr.cor

meancorr.rcorr = rcorr(as.matrix(meancorr))
meancorr.rcorr

corrplot(meancorr.cor,tl.col = 'black')
palette = colorRampPalette(c("pink", "white", "blue")) (20)
heatmap(x = maincorr1.cor, col = palette, symm = TRUE,cexCol = 1, cexRow = 1)
```

Violin Plot:

```
library(dplyr)
library(ggplot2)
library(scales)

expplot <- newdata %>%
  ggplot(aes(x=name, y=value, fill=name))

expplot + geom_violin(show.legend = FALSE)  +
  geom_boxplot(width=0.1, show.legend = FALSE, fill="snow", outlier.size=0.9) +
  labs(title = "Distribution of Gases (Violin Plot)",
       y = "Mean Values (PPB)",x = "Gases") +
  scale_y_continuous(trans = "log2",
                     breaks = trans_breaks("log2", function(x) 2^x),
                     labels = trans_format("log2", math_format(2^.x))) +
  theme_minimal(base_size = 16)
```

QQ Plot:

```
library(ggplot2)
library(ggforce)
library(magrittr)
library(scales)
library(dplyr)

ggplot(maindata, aes(sample=NO2.Mean)) +
  geom_qq() +
  geom_qq_line()

#Comparing all

maindata.qq <- maindata %$%
  data.frame(no2=sort(NO2.Mean),
             so2=sort(SO2.Mean),
             o3=sort(O3.Mean),
             co=sort(CO.Mean))
maindata.qq %>% ggplot(aes(no2,o3)) + geom_point()

maindata.qq <- maindata %$%
  data.frame(no2=rescale(NO2.Mean, to=c(0,1) %>% sort),
so2=rescale(SO2.Mean, to=c(0,1) %>% sort),
o3=rescale(O3.Mean, to=c(0,1) %>% sort),
co=rescale(CO.Mean, to=c(0,1)%>% sort))

maindata.qq %>% ggplot(aes(no2, so2)) +
  geom_point(color ="blue") +
  geom_abline(slope=1, intercept=0)+
  ggtitle("QQ Plot - SO2 vs NO2")+
  xlab("NO2") + ylab("SO2")

maindata.qq %>% ggplot(aes(no2, co)) +
  geom_point(color ="blue") +
  geom_abline(slope=1, intercept=0)+
  ggtitle("QQ Plot - CO vs NO2")+
  xlab("NO2") + ylab("CO")

maindata.qq %>% ggplot(aes(no2, o3)) +
  geom_point(color ="blue") +
  geom_abline(slope=1, intercept=0)+
  ggtitle("QQ Plot - O3 vs NO2")+
  xlab("NO2") + ylab("O3")
```

```r
maindata.qq %>% ggplot(aes(so2, co)) +
  geom_point(color ="blue") +
  geom_abline(slope=1, intercept=0)+
  ggtitle("QQ Plot - CO vs SO2")+
  xlab("SO2") + ylab("CO")

maindata.qq %>% ggplot(aes(so2, o3)) +
  geom_point(color ="blue") +
  geom_abline(slope=1, intercept=0)+
  ggtitle("QQ Plot - O3 vs SO2")+
  xlab("SO2") + ylab("O3")

maindata.qq %>% ggplot(aes(co, o3)) +
  geom_point(color ="blue") +
  geom_abline(slope=1, intercept=0)+
  ggtitle("QQ Plot - O3 vs CO")+
  xlab("CO") + ylab("O3")
```