

Northeastern University
Mechanical and Industrial Engineering Department
IE 6200: Engineering Probability and Statistics
Prof. Rehab Ali
Fall 2023

Project Title: Analyzing the Impact of Growing Industrialization on Climate Indicators: A Comprehensive Study

Project Final Report



Group 1

Team Members:

Vinit Joshi (Point of Contact)	002241417
Neil Vashani	002242557
Adarsh Prakash	002203937
Appanna Puchimanda Mandanna	002207589
Gautham Kurmani	002879434
Venkata Sriram Kumar Charmarathi	002839422

1. Problem Statement

“Analyzing the Effect of Increasing Industrialization on Climate Measures: A Detailed Investigation on Trends in Temperature, Greenhouse Gas Emissions, and the Use of Renewable Energy from 1990 to 2020”

Global industrialization has surged in the last few decades, greatly accelerating economic development. The project explores the complex relationship that exists between important climate indicators and the growing global industrial sector in 184 countries over a period from 1990 to 2020. We intend to investigate variables like temperatures, CO₂, methane, NO emissions, renewable energy consumption, and total greenhouse gas emissions in an effort to identify trends and patterns related to industrial growth. We intend to find patterns, trends, and connections by analyzing this abundance of data, which will help us understand the environmental effects of global industrialization. The project's statistical analysis and data visualization prove to show that the selected greenhouse gas emissions have increased over the years, but however, it will not be able to establish that this increase is linked purely to the phenomenon of industrialization 4.0.

We received data from the World Bank covering emissions of CO₂ [1], nitrous oxide [2], methane [3], total greenhouse gas emissions [4], and renewable energy consumption [5] for the years 1990 to 2020. These parameters were selected since greenhouse gases are an index for industrialization. We collected temperature data from diverse sources, including the World Meteorological Organization (WMO) [6], the National Oceanic and Atmospheric Administration (NOAA) [7], and various National Meteorological Services sites. These reliable sources provide a complete picture of temperature trends. Greenhouse gases indicate industrial pollution, which indicates industrialization. It is straightforward to measure their environmental impact, and this helps create regulations that make industries more productive and cleaner. We decided to combine the two datasets to visualize the trends in the climatic parameters and country temperatures over time. A calculated method uses statistical graphs and inference methods to understand the effects of industrialization on this dataset. Industrial activity increased during the 1990–2020 period, which was deliberately chosen. This helps decision-makers address environmental issues with informed choices.

2. Project Goals

The goals for our project are as follows:

1) To analyze the temperature data:

Generate boxplots to visualize temperature variations in the top five countries with the highest maximum temperatures from 1990 to 2020, showcasing the trends in variability.

2) To analyze Methane emissions:

Construct line charts to visualize the changes in the amounts of methane emissions for the countries which had the highest maximum methane emissions across the years and to predict at least 10% percent increase in methane emissions over the years. This will then prove there was an increase in industrialization. Over the first 20 years after entering the atmosphere, methane has more than 80 times the warming potential of carbon dioxide.

Methane accelerates warming in the immediate term, even if CO₂ has a longer-lasting effect. For this reason, methane is considered as an index for industrialization [8].

3) To analyze CO₂ emissions:

Construct scatter plots to visualize the changes in the amounts of CO₂ emissions for the countries which had the highest maximum CO₂ emissions across the years, and to predict at least 10% percent increase in CO₂ emissions over the years.

4) To analyze NO emissions:

Construct bar graphs to visualize the changes in the amounts of NO emissions for the countries which had the highest maximum NO emissions from 1990 to 2020, and to predict at least 10% percent increase in NO emissions over the years.

5) To analyze a correlation between total greenhouse gas emissions and renewable energy consumption:

To analyze a correlation between total greenhouse gas emissions and renewable energy consumption which would indicate whether the renewable energy consumption is able to tackle the issue of growth in total greenhouse gas emissions.

6) To generate comparative scatter plots for each NO, CO₂, Methane emissions:

To develop a comparative scatter plot to infer the trends for the emission parameters across the years for selected countries for that parameter. This is done to make inferences and have a visual representation of how the emission rates have changed over the years.

7) To perform probability distribution for the parameters in consideration:

To perform mean probability distribution, variance probability distribution and inter-quartile probability distribution for the parameters to calculate the probability of a desired value.

8) To perform Confidence Interval, Hypothesis Testing and Analysis of Variance (ANOVA):

To perform hypothesis testing to check if the increase in industrialization is a significant reason for the increase in methane emissions over the years. To perform confidence interval of means and of difference of means for methane emissions to see if it is in accordance with the hypothesis test. Also, to perform confidence interval of proportions for two countries to observe the trends in methane emissions over the years. To perform ANOVA to see if the mean of methane emissions before 2011 is statistically different from the mean of methane emissions after.

3. Data Collection & Preparation

a) Model Assumptions:

The model assumptions are as follows:

- The data provided on The World Bank is trusted.
- Unit for temperature is in degree Celsius; Units for CO₂ is in Kilotons, and of Methane, NO, Total Greenhouse Gas emissions is in Kilotons of CO₂ equivalent; and unit of renewable energy consumption is in % of total final energy consumption.

- The average temperatures for all the countries mentioned in the data obtained from the WMO, NOAA and other meteorological websites are correct and precise.
- The data pertaining to CO₂, Methane, NO, Total Greenhouse Gas emissions and renewable energy consumption obtained from the dataset is accurate.
- The statistical software such as R Studio, Minitab and Microsoft Excel will perform as expected for statistical analysis.

b) Data Source:

While searching for online datasets, our team examined healthcare, supply chain, delivery systems, and Boston housing issues. We investigated diverse datasets motivated to solve a growing problem, and so we decided upon industrialization-related climate change as our project topic.

During our search, we collected average yearly temperature data from World Meteorological Organization (WMO), the National Oceanic and Atmospheric Administration (NOAA), and various National Meteorological Services sites from the year 1990 to 2020. We also found a dataset on The World Bank that gave us the emissions of CO₂, methane, nitrous oxide, total greenhouse gas emissions, and renewable energy consumption for the years 1990 to 2020.

c) Data Merging:

We found detailed average temperature data for countries worldwide from 1990 to 2020 on various sites. The World Bank also provided CO₂, methane, nitrous oxide, total greenhouse gas, and renewable energy consumption data from the same period. We combined these datasets to understand the complex relationships between industrialization and climate change. This is crucial to our analytical process because it lets us examine the relationship between climate and industry as a whole. By comparing temperature trends with emissions data, our study hopes to illuminate industrialization's environmental impacts and inform sustainable development.

d) Data Cleaning:

We averaged parameters from previous and subsequent years for each country to address the few null values in our dataset. This method filled data gaps and reduced bias in our analysis. We considered the years adjacent to the missing value to maintain dataset consistency and ensure that the restoration process matches the overall trends and patterns. Controlling missing values improves the reliability and precision of our results, allowing a deeper study of industrialization and climate.

e) Data:

After the procedure of data collection, merging and cleaning our dataset consists of the following variables which are explained as follows:

1. **Country:** This variable consists of the name of the country where the temperature and other climatic parameters are calculated.
2. **Year:** It signifies the year in which the temperature and other parameters for the country are calculated. For this dataset, the year ranges from 1990 to 2020.
3. **Temperature:** This signifies the average temperature of the country calculated in the respective year.
4. **CO₂:** This signifies the amount of carbon dioxide emissions emitted in kilotons in that country for that particular year.
5. **NO:** This signifies the amount of nitrous oxide emissions emitted in kilotons of CO₂ equivalent in that country for that particular year.

6. **Methane:** This signifies the amount of methane emissions emitted in kilotons of CO₂ equivalent in that country for that particular year.
7. **TGGE:** This variable signifies the Total Greenhouse Gas Emissions emitted in kilotons of CO₂ equivalent in that country for that particular year.
8. **REC:** This variable signifies the Renewable Energy Consumption of that country for that particular year. This is measured as a percentage of the total final energy consumption.

[HERE IS THE LINK FOR THE PROJECT DATASET](#)

f) Target Population:

The target population for this project is the 184 countries that were part of this dataset whose temperature and various greenhouse gas emissions were used for the statistical analysis.

g) Performance measures:

Hypothesis: *“Growing industrial operations should positively correlate with increased greenhouse gas emissions over the years. Furthermore, we anticipate that there should be a negative correlation between the total greenhouse gas emissions and renewable energy consumption.”*

This hypothesis is predicated on the logical deduction that a country's expanding industries would increase greenhouse gas emissions, which is harmful for sustainability. Furthermore, a nation's decision to employ renewable energy will directly affect and reduce the amount of greenhouse gases released into the atmosphere. Based on the study in 1970s [9], it was anticipated that temperatures would rise in the upcoming years due to global warming, which is why safety precautions were implemented and regulations were established. Therefore, our analysis predicts that majority of the countries will only see a modest growth over time. Subsequently, we will leverage the R programming language and Minitab software to implement coding procedures, allowing us to generate various graphical representations.

4. Measures of Central Tendencies & Variability

As seen in our project dataset, we utilized a systematic statistical analysis on Excel that included important statistical measures. The mean gives the average data value and a central tendency reflecting the year's trend. Note that our results show a steady rise in means. The median is the dataset's midpoint, and provides information on the distribution's center position, which is useful for datasets with outliers.

Range is the difference between the lowest and highest values, indicating dataset variability. The first and third quartiles (Q1) and Q3 show the data's distribution. The interquartile range (IQR= Q3-Q1) is a reliable indicator of variability that is less susceptible to extreme values for the middle 50% of data. We also calculated the dataset's variance and standard deviation to assess its dispersion. The standard deviation measures the deviation of each data point from the mean, indicating dataset volatility. Variance also measures the data's spread. Through a methodical assessment of these statistical metrics, we got a thorough grasp of the distribution, dispersion, and core patterns of environmental factors for a given year in every country.

5. Data Visualization

We have a dataset of 184 countries, each with their parameters (namely CO₂, Methane, NO emissions; Temperature, TGGE and Renewable Energy Consumption) from 1990 to 2020. We narrowed the countries and parameters to visualize data clearly by choosing the top 5 countries with the highest parameter over time for each parameter. We only chose 1990, 1995, 2000, 2005, 2010, 2015, and 2020 to look for significant changes in that parameter.

- **Temperature:**

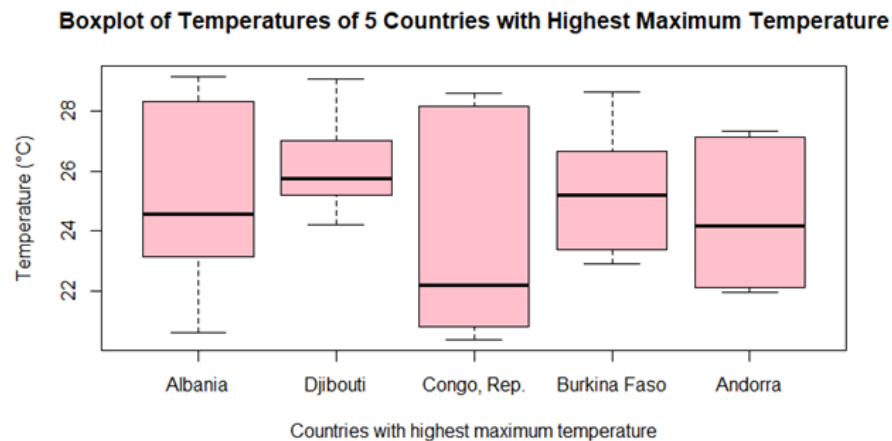


Figure 1: Boxplot of Temperature

Inference:

From figure 1, it can be seen that Albania and Congo, Rep have a much greater variability compared to the other countries. This means that their temperature varies the most from 1990 to 2020 compared to other countries. There are no clear outliers in any of the countries' data. The temperature boxplots for all countries except Burkina Faso are positively skewed. This means that most of the temperatures are concentrated between the second and the third quartile. It should also be noted that the temperature range of Djibouti is the smallest, which means that it has the least amount of variability and the temperature has remained fairly consistent throughout the years.

- **CO₂ emissions:**

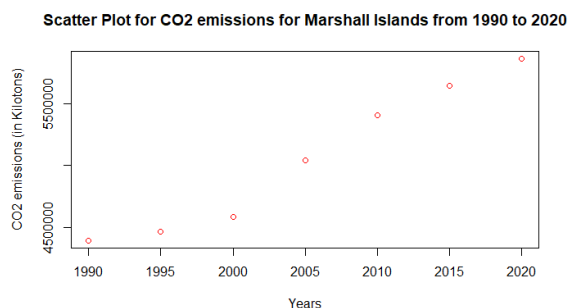


Figure 2: Scatter Plot for Marshall Islands

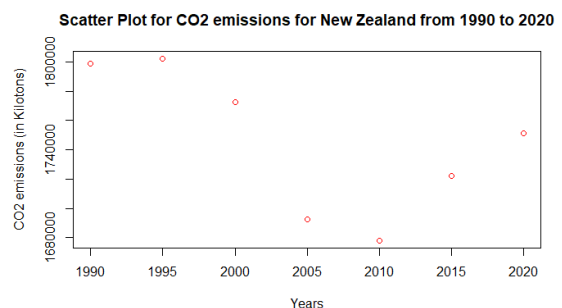


Figure 3: Scatter Plot for New Zealand

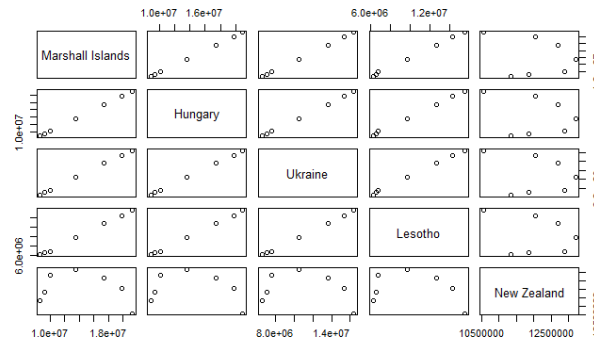


Figure 4: CO₂ emission comparison

Inference:

- From figure 2 and Appendix B; Hungary, Marshall Islands, Ukraine, and Lesotho have similar CO₂ emission increases. CO₂ emissions rise gradually until 2000, after which they spike. Global industrialization may be the main cause. However, from figure 3 it can be seen that CO₂ emissions in New Zealand vary from the norm. Contrary to expectations, carbon emissions decrease until 2010, then rise again. This may have several causes. An expanding economy often increases industrial activity, energy use, and CO₂ emissions. Population growth may increase resources, energy use and emissions. Deforestation and agricultural changes may affect the carbon balance. From the graphs, it is quite evident that Marshall Islands has the highest CO₂ emissions with New Zealand being the lowest among the ones that are selected. Also, it is difficult to comment on the exact distribution of the graphs since we have data of only 30 years.
- From figure 4, comparative scatter plot in R depicts CO₂ emissions for the countries. The first graph shows CO₂ emissions for Hungary and Marshall Islands across selected years. For instance, in 2005, Hungary emitted approximately 1.55e+07 kilotons of CO₂, while Marshall Islands emitted about 1.4e+07 kilotons. Similar comparisons can be made for the other graphs representing different country pairs.
- As seen in [this Excel File](#), on comparing the mean of CO₂ emissions from 1990-2010 with the mean of emissions from 2011-2020, there is a 45.02% increase in the CO₂ emissions, thus satisfying our project goal.
- **Methane emissions:**

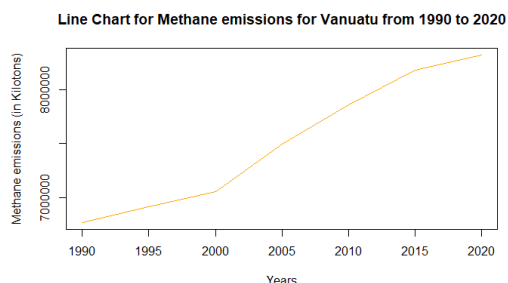


Figure 5: Line chart for Vanuatu

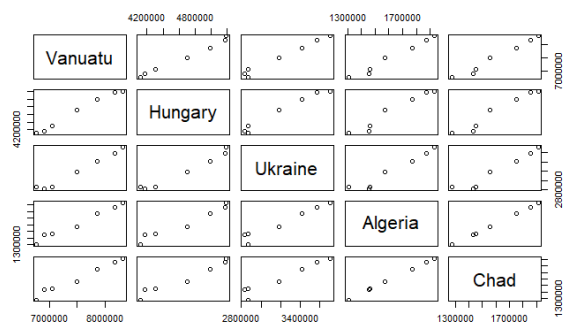


Figure 6: Methane emission comparison

Inference:

- From figure 5 and Appendix C; it can be observed that Methane emissions in Vanuatu, Hungary, Algeria, and Chad rise similarly. Methane emissions gradually rise until 2000,

after which they spike. Industrialization may be the main cause. However, as per Appendix C for Ukraine, first the methane emissions dropped from 1990 to 2000. Methane emissions have increased since 2000, presumably due to industry growth. From the graphs, Vanuatu emits the most methane, while Chad emits the least. Since we only have 30-year data, it is difficult to comment on its distribution.

- From figure 6, comparative scatter plot in R depicts methane emissions for the countries. The first graph shows methane emissions for Vanuata and Hungary across selected years. For instance, in 2005, Hungary emitted approximately 4700000 kilotons of methane, while Vanuatu emitted about 7500000 kilotons. Similar comparisons can be made for the other graphs representing different country pairs.
- As seen in [this Excel File](#), on comparing the mean of methane emissions from 1990-2010 with the mean of emissions from 2011-2020, there is a 17.07% increase in the methane emissions, thus satisfying our project goal.

- **NO emissions:**

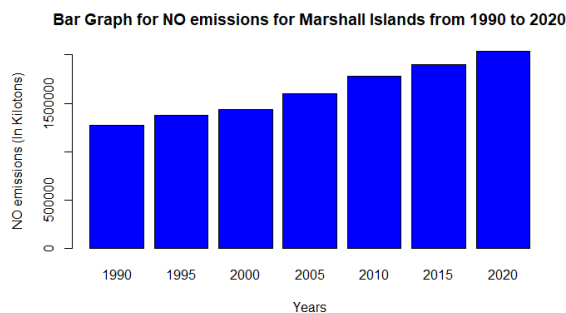


Figure 7: Bar Graph for Marshall Islands

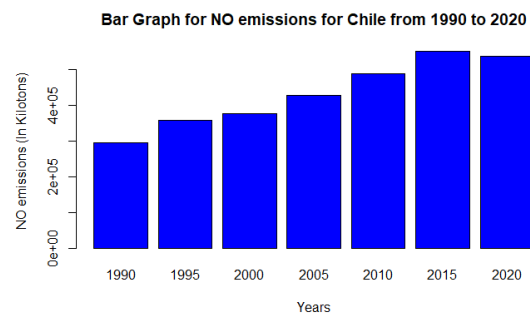


Figure 8: Bar Graph for Chile

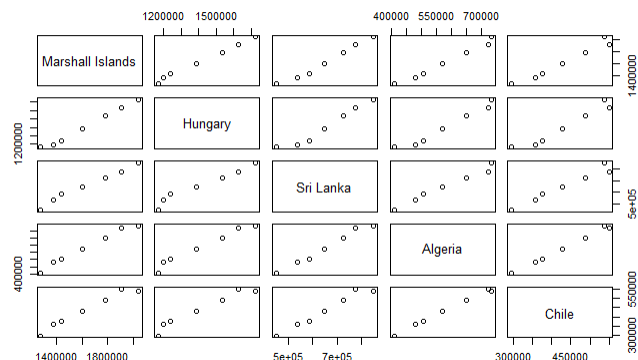


Figure 9: NO emission comparison

Inference:

- From figures 7, 8 and Appendix D; it can be inferred that NO emissions rise similarly in Hungary, Marshall Islands, Sri Lanka, Algeria, and Chile. NO emissions rise steadily from 1990 to 2020. Global industrialization may be the main cause. Chile reduces NO emissions after 2015, contrary to norms. This may be due to more renewable energy. In terms of NO emissions, Marshall Islands is the highest and Chile the lowest. Since we only have 30 years of data, commenting on the distribution is difficult.
- From figure 9, comparative scatter plot in R depicts NO emissions for the countries. The first graph shows NO emissions for Hungary and Marshall Islands across selected

years. For instance, in 2005, Hungary emitted approximately 1500000 kilotons of NO, while Marshall Islands emitted about 1420000 kilotons. Similar comparisons can be made for other graphs representing different country pairs.

- As seen in [this Excel File](#), on comparing the mean of NO emissions from 1990-2010 with the mean of emissions from 2011-2020, there is a 20.75% increase in the NO emissions, thus satisfying our project goal.

- **Total greenhouse gas emissions & Renewable Energy consumption:**

We made correlation plots between the total greenhouse gas emissions and the renewable energy consumption of the countries with the highest and lowest mean total greenhouse gas emissions across the years. On this basis, we selected Vanuatu and Türkiye. Moreover, we also decided to include the United States of America for analysis.

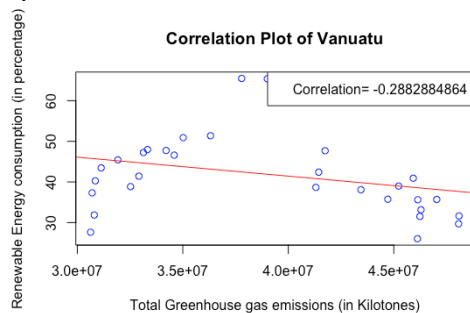


Figure 10: Correlation Plot for Vanuatu

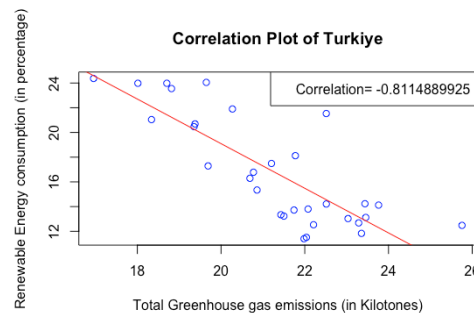


Figure 11: Correlation Plot for Türkiye

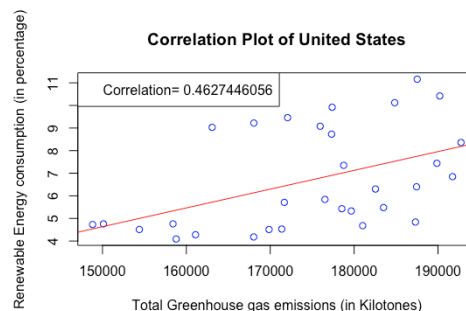


Figure 12: Correlation Plot for United States

Inference:

- From figure 10, it can be seen that in Vanuatu there exists a negative correlation. It shows that renewable energy consumption can keep up with industrialization, reducing greenhouse gas emissions. Thus, the renewable energy sector can offset the emission rate increase. The value is -0.2882884864, hence it is a weak negative correlation as the value is closer to 0.
- From figure 11, it can be observed that in Türkiye there exists a negative correlation. It indicates that the renewable energy consumption can keep up with the increasing industrialization in the country, which results in a reduction in the greenhouse gas emissions. The value of correlation is -0.8114889925, hence it is a strong negative correlation as the value is closer to 1.
- From figure 12, it can be observed that in United States there exists a positive correlation. The country's industrialization is outpacing renewable energy consumption, which increases greenhouse gas emissions. Thus, to offset the emission

increase, renewable energy must improve. Here, the correlation value is 0.4627446056, therefore a weak correlation exists as value is closer to 0.

6. Statistical Analysis

A. Probability Distribution

We performed Mean Probability Distribution, Variance Probability Distribution as well as Inter Quartile Range Probability Distribution, such that we could identify the probability distribution for a desired value. The six parameters were distributed across the three probability distributions and inferences were made on the same.

A) Mean Probability Distribution:

As seen in figure 13 and 14, histograms were created for the probability distributions of the mean temperatures and CO₂ emissions for all countries from 1990 to 2020. We can also calculate desired value probability, whether it is greater than or equal to the input value. For example, if we input 12 as the desired value in the code to find if the probability of the mean temperature is higher than 12 degrees, we get the probability as 0.9032258. This implies that the probability of having a mean temperature above 12 degrees is 90.3%.

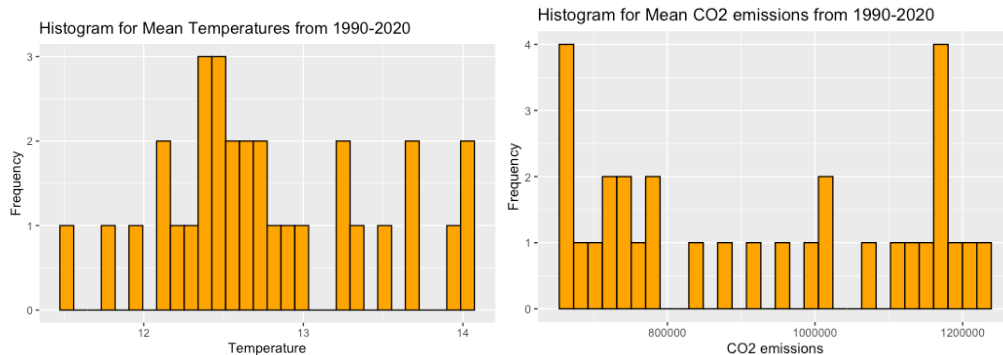


Figure 13: Mean PD for Temperature **Figure 14:** Mean PD for CO₂ emissions

B) Variance Probability Distribution:

As seen in figure 15 and 16, histograms show methane and NO emissions variance probability distributions for all countries from 1990 to 2020. The probability of a desired value indicates if the variance exceeds the input value. For example, if we want to know the probability of methane variance being higher than 8e+11 kilotons equivalent of CO₂, we enter 8e+11 into the code and see 0.3548387. This means methane emissions variance is 35.4% likely to exceed 8e+11 kilotons of CO₂.

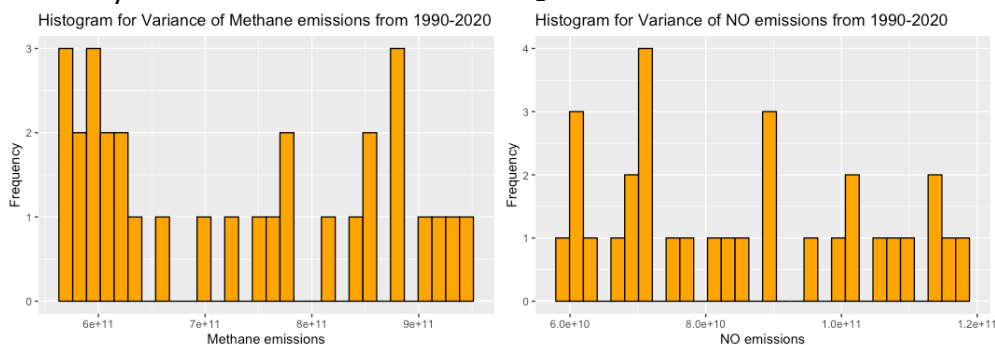


Figure 15: Variance PD for Methane emissions **Figure 16:** Variance PD for NO emissions

C) Inter-Quartile Range Probability Distribution:

As seen in figure 17 and 18, histograms were used to plot the inter quartile range of Total Greenhouse Gas Emissions (TGGE) and Renewable Energy Consumption (REC) for all countries from 1990 to 2020. The probability of a desired value shows the likelihood of inter quartile range exceeding the input value. For example, if we enter 3e+5 as the desired value in the code for the inter quartile range probability distribution of TGGE, the probability of the range being higher than 3e+5 kilotons equivalent of CO₂ is 0.6129032. The interquartile range of TGGE is 61.2% likely to exceed 3e+5 kilotons of CO₂.

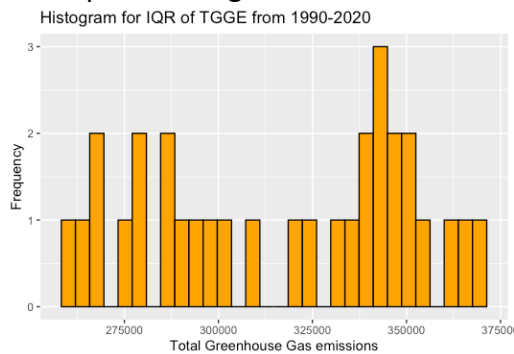


Figure 17: IQR PD for TGGE

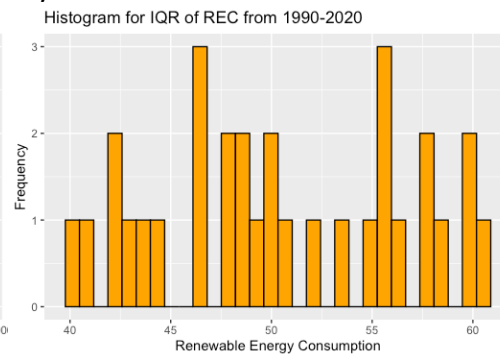


Figure 18: IQR PD for REC

B. Confidence Interval

A) Confidence Interval of Mean

As mentioned in our project goals, methane has more than 80 times the warming potential of carbon dioxide; we decided to test the trend of methane emissions over the selected timeframe. We calculated 95% confidence intervals for the average methane emissions for the years 1990-2020, as well as for the years before to and following 2011. This division is in line with Industry 4.0's globalization which got popularized in 2011. The purpose of the investigation is to determine whether the post-2011 industrialization boom is associated with a noticeable increase in methane emissions.

On performing the confidence interval of mean, we find:

- **Confidence Interval for Mean of Methane emissions from 1990 to 2020: [260642.9, 278943.2]**

Here, mean= 269793.0641

Standard Deviation= 24945.62443

Using T table: $t_{0.025, 30} = 2.042$ (df= n-1= 31-1= 30, $\alpha/2 = 0.025$)

Using formula:

$$\left(\bar{x} - t_{\alpha/2, n-1} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2, n-1} \cdot \frac{s}{\sqrt{n}} \right)$$

Lower confidence Level = 260642.9

Upper Confidence Level = 278943.2

The results from R and by formula are same. Hence verified.

- **Confidence Interval for Mean of Methane emissions from 1990 to 2010: [248235.9, 263184.3]**

Here, mean= 255710.1007

Standard Deviation= 16419.79822

Using T table: $t_{0.025, 20} = 2.086$ (df= n-1= 21-1= 20, $\alpha/2 = 0.025$)

Using formula

$$\left(\bar{x} - t_{\alpha/2, n-1} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2, n-1} \cdot \frac{s}{\sqrt{n}} \right)$$

Lower confidence Level = 248235.9

Upper Confidence Level = 263184.3

The results from R and by formula are same. Hence verified.

- **Confidence Interval for Mean of Methane emissions from 2011 to 2020: [294811.3, 303923.3]**

Here, mean= 299367.2873

Standard Deviation= 6368.813655

Using T table: $t_{0.025, 9} = 2.262$ (df= n-1= 10-1= 9, $\alpha/2 = 0.025$)

Using formula:

$$\left(\bar{x} - t_{\alpha/2, n-1} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2, n-1} \cdot \frac{s}{\sqrt{n}} \right)$$

Lower confidence Level = 294811.3

Upper Confidence Level = 303923.3

The results from R and by formula are same. Hence verified.

When comparing the mean methane emissions for the years 2011 to 2020 to the interval for the years 1990 to 2010, the wider confidence interval for the latter period suggests a higher level of uncertainty or variability. Furthermore, we can see that there is an overlap between the confidence intervals for all three cases. This suggests that there is not enough evidence to conclude a significant increase in methane emissions post 2011 due to industrialization. Although the overlap indicates that we cannot confidently claim the rise to be statistically significant at a 95% confidence interval, the confidence intervals suggest a potential increase in methane emissions.

B) Confidence Interval of Difference of Mean

Based on the reasoning mentioned above, confidence interval of difference of mean was calculated for mean of methane emissions from 1990 to 2010 and from 2011 to 2020 of all countries.

Confidence Interval of difference of means for methane emissions from 1990 to 2010, and from 2011 to 2020: [-49718.24, -37596.13]

In our case, entity 1 is mean of methane emissions between 1990 to 2010 and entity 2 is mean of methane emissions between 2011 to 2020.

At a 95% confidence level, the confidence interval does not contain zero, which means:

- There is sufficient data to support a statistically significant difference in mean methane emissions between the years 1990 to 2010 and 2011 to 2020.
- The mean of methane emissions of 2011-2020 is always greater than methane emissions of 1990-2010.

C) Confidence Interval of Proportions

Confidence intervals for proportions help in statistical inference and decision-making based on sample data by providing information about the likely range of population proportions. Because of their historical prominence in industrialization, the USA and China have been chosen to represent the confidence intervals of proportions. Analyzing the percentage of methane emissions in these countries offers important insights into possible long-term environmental effects linked to industrial activity, given their prominent positions in global industrial advances. We examined the mean of all years' methane emissions in the USA and China from 1990 to 2020 at 95% confidence interval. By employing a binary categorization system (0 for values below the mean, 1 for values above), it was determined that 13 out of 31 years fell below the mean for the USA, while 16 out of 31 years fell below the mean for China.

- **Confidence Interval of Proportion for the USA: [0.2506623, 0.6073865]**

Here number of success (number of years under the mean) = 13

Sample size = 31

Using the formula:

$$\hat{p} \pm Z_{1-\alpha/2} * \sqrt{\left(\frac{\hat{p}(1-\hat{p})}{n}\right)}$$

Here $\hat{p} = 13/31$, $n=31$

$1-\alpha/2 = 1- 0.05/2 = 0.975$

Using the formula we get,

Lower Confidence Level= 0.2506623

Upper Confidence Level= 0.6073865

The results from R and by formula are same. Hence verified.

- **Confidence Interval of Proportion for China: [0.3339764, 0.6944174]**

Here number of success (number of years under the mean) = 16

Sample size = 31

Using the formula:

$$\hat{p} \pm Z_{1-\alpha/2} * \sqrt{\left(\frac{\hat{p}(1-\hat{p})}{n}\right)}$$

Here $\hat{p} = 16/31$, $n=31$

$1-\alpha/2 = 1- 0.05/2 = 0.975$

Using the formula we get,

Lower Confidence Level = 0.3339764

Upper Confidence Level = 0.6944174

The results from R and by formula are same. Hence verified.

For the United States, the percentage of years with methane emissions below the mean is expected to be between 25.1% and 60.7%. This indicates a range in which a significant percentage of the years deviate significantly from the mean, although the exact proportion is unclear. It is projected that in China, the percentage of years with methane emissions below the mean ranges from 33.4% to 69.4%. Like the USA, this range shows considerable variation in the percentage of years that fall short of the mean, though the precise percentage is unknown. There is no statistically significant difference between the USA and China in the fraction of years with methane emissions below the mean, based on the overlap of the confidence intervals for the two nations.

C. Hypothesis Testing

A) Left Tailed Test for Methane emissions:

We performed hypothesis testing on the data for methane emissions by dividing the data into two groups. As mentioned earlier in the confidence interval of mean section, the two groups taken under consideration are the means of methane emissions for all countries pre and post 2011. We performed left tailed hypothesis test at 95% confidence level (or 0.05 significance level), to see whether if the increase in industrialization a significant reason for the increase in methane emissions over the years.

μ_1 = mean of methane emissions for 1990-2010

μ_2 = mean of methane emissions for 2011-2020

$H_0: \mu_1 - \mu_2 = 0$

$H_1: \mu_1 - \mu_2 < 0$

Using the formula for t critical value:

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}}$$

We get the t value= - 11.276

For degree of freedom:

We get df= 29.612, here we round it off to lowest value i.e., df= 29

$$\nu = \frac{\left(\frac{s_1^2}{m} + \frac{s_2^2}{n}\right)^2}{\frac{(s_1^2/m)^2}{m-1} + \frac{(s_2^2/n)^2}{n-1}}$$

On performing left tailed test on R, we find that:

p- value = 1.541e-12 < α = 0.05

Conclusion: Reject the null hypothesis

Result: The mean for 1990-2010 methane emissions is less than mean for 2011-2020 methane emissions. This falls in line with our claim that industrialization caused an increase in methane emissions.

B) Left Tailed Test for Total Greenhouse gas emissions:

We performed hypothesis testing on the data for total greenhouse gas emissions by dividing the data into two groups. The two groups taken under consideration are the means of Total Greenhouse Gas Emissions (TGGE) for all countries pre and post 2011. We performed left tailed hypothesis test at 95% confidence level (or 0.05 significance level), to see whether if the increase in industrialization a significant reason for the increase in TGGE over the years.

μ_1 = mean of TGGE for 1990-2010

μ_2 = mean of TGGE for 2011-2020

$H_0: \mu_1 - \mu_2 = 0$

$H_1: \mu_1 - \mu_2 < 0$

Using the formula for t critical value:

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}}$$

We get the t value= -11.031

For degree of freedom:

We get df= 23.842, here we round it off to lowest value i.e., df= 23

$$\nu = \frac{\left(\frac{s_1^2}{m} + \frac{s_2^2}{n}\right)^2}{\frac{(s_1^2/m)^2}{m-1} + \frac{(s_2^2/n)^2}{n-1}}$$

On performing left tailed test on R, we find that:

p- value = 3.787e-11 < α = 0.05

Conclusion: Reject the null hypothesis

Result: The mean for 1990-2010 total greenhouse gas emissions is less than mean for 2011-2020 total greenhouse gas emissions. This falls in line with our claim that industrialization caused an increase in total greenhouse gas emissions.

D. Analysis of Variance (ANOVA)

We used ANOVA to compare the means of three groups to determine if there are any statistically significant differences among them. Our groups are as follows:

- Group 1: Methane emissions from 1990-2010
- Group 2: Methane emissions from 2011-2020

- Group 3: Methane emissions from 1990- 2020

Here,

μ_1 = mean of methane emissions for 1990-2010

μ_2 = mean of methane emissions for 2011-2020

μ_3 = mean of methane emissions for 1990-2020

$H_0: \mu_1 = \mu_2 = \mu_3$

H_1 : Not all μ are equal

In our test, number of samples (i) =3 and number of observations (j)= 21,10,31 respectively for each sample. We select level of significance= 95%, using Minitab we get

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Timeframe	2	12911273799	6455636900	15.59	0.000
Error	59	24425776916	413996219		
Total	61	37337050715			

Figure 19: ANOVA Test on Minitab

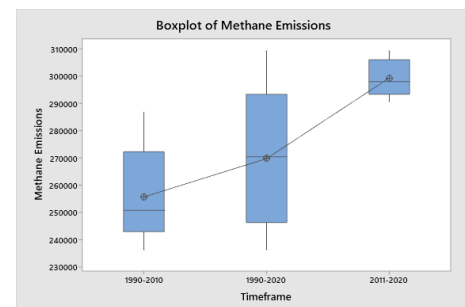


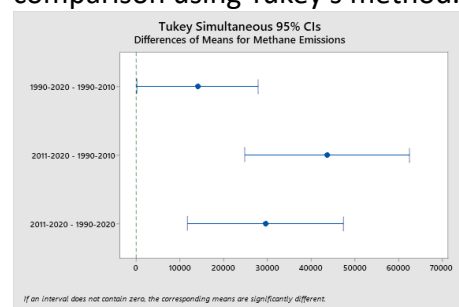
Figure 20: Boxplot

Here from Figure 19, we get the f value= 15.59 and value of $p = 0.00 < \alpha = 0.05$.

Conclusion: We reject H_0 .

Result: Not all means of methane emissions in 3 timeframes (1990-2010, 2011-2020, 1990-2020) are equal. This can be evidently seen in the Boxplot in Figure 20.

To find which value of mean is different from the other, we performed multiple comparison using Tukey's method.



Grouping Information Using the Tukey Method and 95% Confidence

Timeframe	N	Mean	Grouping
2011-2020	10	299367	A
1990-2020	31	269793	B
1990-2010	21	255710	C

Means that do not share a letter are significantly different.

Figure 21: Tukey's Method

From Figure 21, we can infer that there are 3 groupings that are made and from the graph we can see that 0 is not included. Hence, the mean of methane emissions in 3 timeframes (1990-2010, 2011-2020, 1990-2020) are significantly different. Further, ANOVA Normality Plot can be seen in Appendix E.

7. Results

- An extensive investigation included data visualizations and probability distributions for various factors. Temperature trends were shown using boxplots, CO₂ emissions as scatterplots, methane as line charts, and NO as bar graphs. The overall trajectory showed an increase in emissions across all parameters over time.

- Confidence intervals were calculated for means, differences of means, and proportions, with a focus on methane emissions. Confidence interval of difference of means showed that sufficient data exists to support a statistically significant difference in mean methane emissions between the years 1990 to 2010 and 2011 to 2020.

8. Strengths

- The study analyzed total greenhouse gas emissions (TGGE) and renewable energy consumption (REC) over time using correlation plots. We did so for the US, Türkiye, and Vanuatu, which had the highest maximum TGGE values over time. Vanuatu and Türkiye had negative correlations, but the US had a positive correlation, indicating that renewable energy needed to improve to offset increase in greenhouse gas emission.
- We also performed hypothesis testing and the results of all tests solidified our initial claim. Thus, we were able to successfully claim that there has been a significant increase in the methane & total greenhouse gas emissions post 2011.
- On performing ANOVA, we concluded that the mean of methane emissions in the 3 timeframes (1990-2010, 2011-2020, 1990-2020) are significantly different.

9. Limitations/Weaknesses

- Integrating two datasets, one with annual temperature data and the other with five parameters for 184 nations, was a significant challenge. Before analysis, these datasets had to be combined and cleaned to ensure data consistency and dependability.
- While the project proves through statistical analysis and data visualization that there is an increase in the greenhouse gas emissions selected, it will not be able to prove that the increase is solely due to the phenomenon of industrialization 4.0.






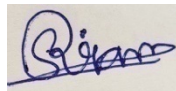
10. Conclusion

In conclusion, the findings offer detailed insights into patterns of methane emissions, which is an essential indicator of the effects of industrialization on the ecosystem. Comprehensive statistical analysis is crucial for gaining insights into complicated information. This includes a variety of visualizations for all parameters, as well as a targeted evaluation of important indicators like methane emissions.

11. Proposed Next Steps and Future Work

- To gain a more complete understanding of methane emissions trends, the analysis timeline should be extended from the 1950s to 2023, including pre-1990 and recent years. To understand the long-term effects of industrialization on methane emissions, future research could examine Industrialization 3.0 from the 1970s, provided that data regarding number of industries for the time frame can be gathered.
- The data visualization performed involved a fraction of countries that were a part of the dataset in order to better demonstrate the graphs to a reader. Further expansion of the scope of data visualization by incorporating more countries than currently involved would further benefit the main aim of the project.
- More robust conclusions can be made by collaborative efforts with environmental scientists and industry specialists, which further deepen findings.

Collaboration

Team Member	Contributions	Signature
Vinit Joshi	Problem Statement, R Programming, Data Visualization, Statistical Analysis, Report Format, Project Presentation	
Neil Vashani	Problem Statement, R Programming, Data Visualization, Statistical Analysis, Report Format, Project Presentation	
Adarsh Prakash	Problem Statement, Data Collection, Data Preparation, Data Cleaning, Project Goals, Project Presentation	
Appanna Puchimanda Mandanna	Problem Statement, Data Collection, Data Preparation, Data Cleaning, Project Goals, Project Presentation	
Gautham Kurmani	Problem Statement, Data Preparation, Data Cleaning, Measures of central tendencies and variability, Project Presentation	
Venkata Sriram Kumar Charmarthi	Problem Statement, Project Goals, Data Collection, Data Preparation, Data Cleaning, Project Presentation	

Appendix

Appendix A: Comments from Project Progress Report

- The comment regarding Inclusion and Exclusion Criteria: The inclusion comment has been explained in the Results section of the project. The exclusion comment has been explained in the Problem Statement and Limitations/ Weaknesses section, but briefly states- *The project's statistical analysis and data visualization show that the selected greenhouse gas emissions have increased, but it cannot establish that this increase is linked purely to the phenomenon of industrialization 4.0.*
- The comment regarding finding a Dataset pertaining to number of industries: After thorough research, we were not able to come across any dataset that had the number of factories and the amount of renewable energy activities for the countries chosen for our analysis for the considered time periods. Due to this reason, we were not able to perform further analysis on the same.
- The comment regarding explanation of why we chose top 5 countries for our Data Visualization: It would not be possible for us to perform data visualization for all 184 countries and statistical analysis for all climatic parameters due to the page limit constraint for our project. Hence analysis was done for the top 5 countries having the highest maximum value for the selected parameter.

Appendix B: Scatter Plots for CO₂ emissions

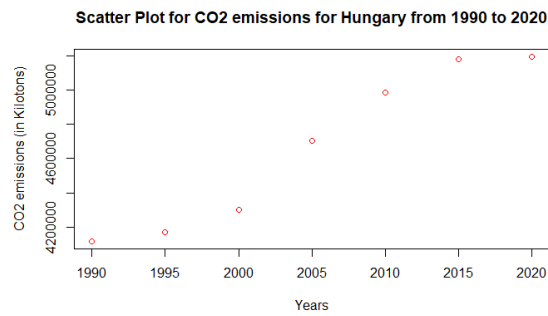


Figure: Scatter Plot for Hungary

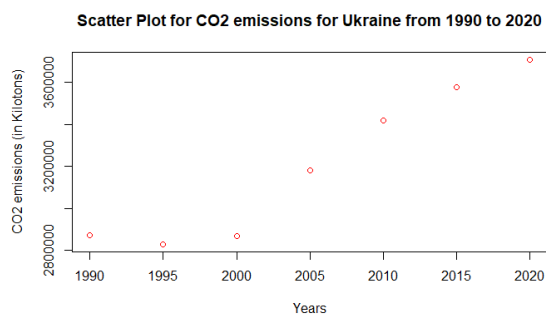


Figure: Scatter Plot for Ukraine

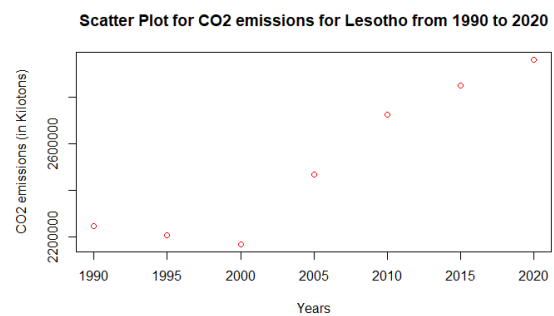


Figure: Scatter plot for Lesotho

Appendix C: Line chart for methane emissions

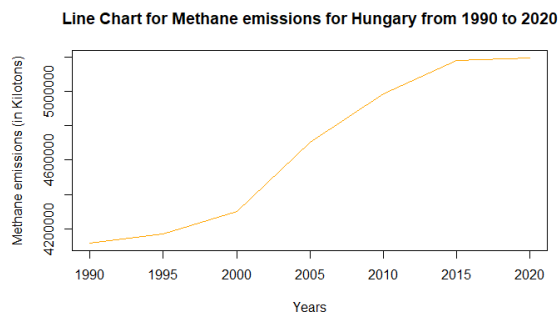


Figure: Line Chart for Hungary

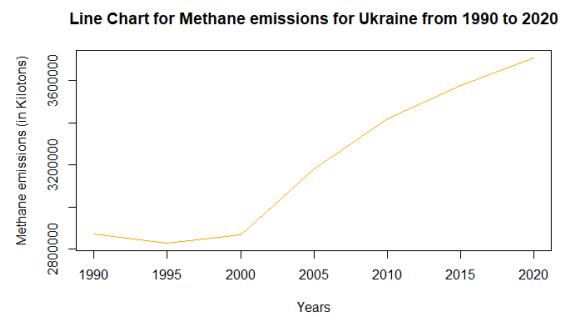


Figure: Line Chart for Ukraine

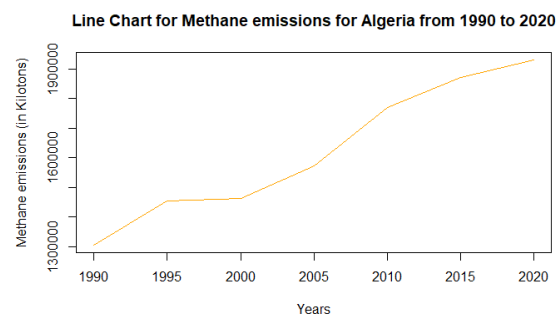


Figure: Line Chart for Algeria

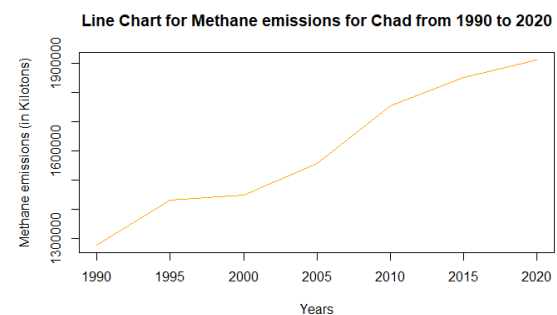


Figure: Line Chart for Chad

Appendix D: Bar Graph for NO emissions

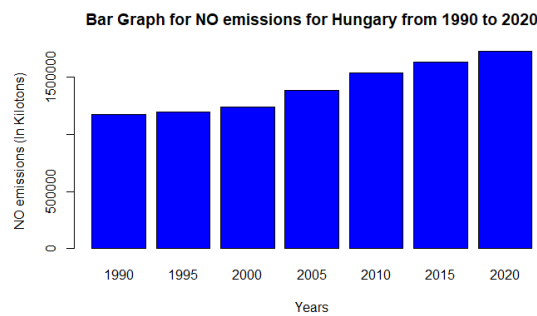


Figure: Bar Graph for Hungary

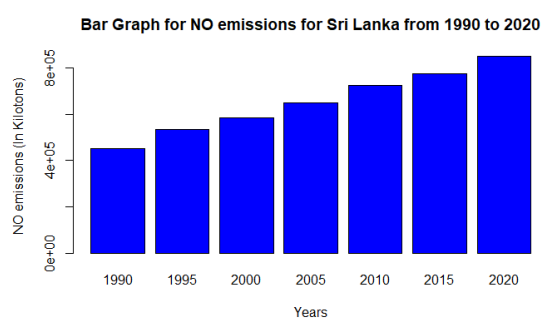


Figure: Bar Graph for Sri Lanka

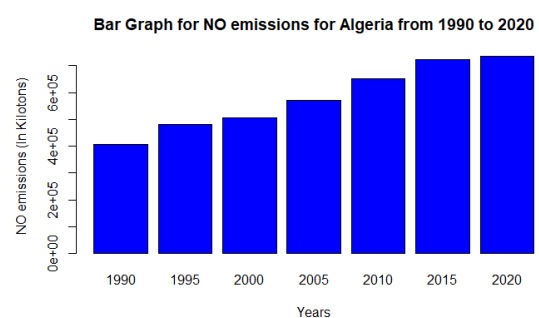


Figure: Bar Graph for Algeria

Appendix E: ANOVA Normality Plot

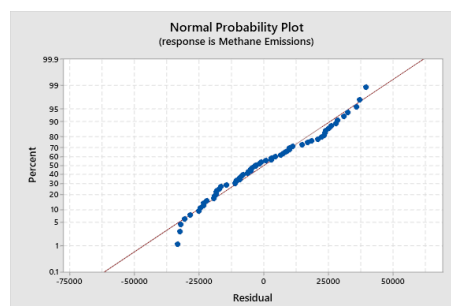


Figure: Normality Plot for ANOVA

It can be seen that the points on the normal probability plot form a relatively straight line with only one outlier. This suggests that the values follow a normal distribution, supporting the assumption of normality for ANOVA.

Bibliography

[1]	CO ₂ emissions (KT). World Bank Open Data. (n.d.). https://data.worldbank.org/indicator/EN.ATM.CO2E.KT
[2]	Methane emissions (kt of CO ₂ equivalent). World Bank Open Data. (n.d.-b). https://data.worldbank.org/indicator/EN.ATM.METH.KT.CE

[3]	Nitrous oxide emissions (thousand metric tons of CO ₂ equivalent). World Bank Open Data. (n.d.-c). https://data.worldbank.org/indicator/EN.ATM.NOXE.KT.CE
[4]	Total greenhouse gas emissions (kt of CO ₂ equivalent). World Bank Open Data. (n.d.-c). https://data.worldbank.org/indicator/EN.ATM.GHGT.KT.CE
[5]	Renewable energy consumption (% of total final energy consumption). World Bank Open Data. (n.d.-c). https://data.worldbank.org/indicator/EG.FEC.RNEW.ZS
[6]	Climate. World Meteorological Organization. (2023, July 11). https://public.wmo.int/en/our-mandate/climate
[7]	National Oceanic and Atmospheric Administration. (n.d.-a). https://www.noaa.gov/
[8]	“Methane: A crucial opportunity in the Climate Fight,” Environmental Defense Fund. https://www.edf.org/climate/methane-crucial-opportunity-climate-fight#:~:text=Methane%20has%20more%20than%2080,warming%20in%20the%20near%20term (accessed Nov. 23, 2023).
[9]	Climate science glossary. Skeptical Science. (n.d.). https://skepticalscience.com/ice-age-predictions-in-1970s-intermediate.htm