Mrinal Chaudhari
Vaishnavi Solunke

Applied Statistic-641
FINAL PROJECT REPORT
Climate change data analysis

1. EXPLANATION AND RESEARCH TOPIC

There are many studies showing that the average temperature of the earth has risen over the last century. The consequences of a sustained rise in global temperatures will be catastrophic. Rising sea levels and increasing frequency of extreme weather events affect billions of people. Many corporate organizations are yet to see the impact of climate change at organizational level, that the impact of Climate change on Organization's Human Resource can affect their revenues and other metrics.

In this issue, we will examine the relationship between the average temperature of the earth and several other factors. This problem is an attempt to study the relationship between average global temperature and other factors.

2. DATA COLLECTION

2.1 OBJECTIVE AND METHODOLOGY

The main objective behind this collaborative study is to gain knowledge and insights into the climate change data. Data collection is a vital aspect in regression analysis. The dataset used for this project is obtained from www.kaggle.com [3]. This data collection method is of type retrospective study based on historical data. There are eight predictor variables and one response variable. For the project the response is temperature and the predictor variables are the various factors that affect the temperature like CO2, N2O,CH4 etc. The reason we have chosen temperature as the response variable is because the various factors in the atmosphere affect the temperature level and cause extreme events in the environment. The study of our project is to develop a linear regression model using the predictor variables given in the dataset and using those predictor variables we will predict the temperature.

2.2 PARAMETRES USED

The dataset contains data from May 1983 to December 2008. The available variables include:

Temp: the difference in degrees Celsius between the average global temperature in that period and a reference value. This data comes from the Climatic Research Unit at the University of East Anglia.

CO2, N2O, CH4, CFC.11, CFC.12: atmospheric concentrations of carbon dioxide (CO2), nitrous oxide (N2O), methane (CH4), trichlorofluoromethane (CCl3F; commonly referred to as

CFC-11) and dichlorodifluoromethane (CCl2F2; commonly referred to as CFC-12), respectively. This data comes from the ESRL/NOAA Global Monitoring Division.

CO2, N2O and CH4 are expressed in ppmv (parts per million by volume -- i.e., 397 ppmv of CO2 means that CO2 constitutes 397 millionths of the total volume of the atmosphere)
CFC.11 and CFC.12 are expressed in ppbv (parts per billion by volume).

Aerosols: the mean stratospheric aerosol optical depth at 550 nm. This variable is linked to volcanoes, as volcanic eruptions result in new particles being added to the atmosphere, which affect how much of the sun's energy is reflected back into space. This data is from the Godard Institute for Space Studies at NASA.

TSI: the total solar irradiance (TSI) in W/m2 (the rate at which the sun's energy is deposited per unit area). Due to sunspots and other solar phenomena, the amount of energy that is given off by the sun varies substantially with time. This data is from the SOLARIS-HEPPA project website.

MEI: multivariate El Nino Southern Oscillation index (MEI), a measure of the strength of the El Nino/La Nina-Southern Oscillation (a weather effect in the Pacific Ocean that affects global temperatures). This data comes from the ESRL/NOAA Physical Sciences Division.

The dataset is available on kaggle for general use.

3. METHOD OF ANALYSIS

For this regression model, we have taken temperature as the response variable. This signifies the temperature caused by the various predictor parameters. For this response variable, we have taken eight predictor variables, all of each are very relevant to the temperature increase. We have seen the definition of all the variables in the previous section, now we will see the relevance of each predictor variable in the regression model [1].

Before applying linear regression on the data let us take a look at the uses of Regression.
The following are the uses of Regression:

1. Data description: A regression analysis model turns out to be useful to summarise the data, and is suitable to use more than a table or a graph when we have a significant amount of data for climate change and model.

2. Parameter estimation: Regression analysis is used to produce an estimation for a variable if a sample of observed values is available. For example, utilizing regression analysis for equations used to describe relationships among variables like velocity, reaction, concentration, etc. can be advantageous to engineers in the chemical field. To produce an estimate of maximum velocity

the engineers can use regression analysis if the sample of observed values at different concentrations is available.

3. Prediction and estimation: For instance, say you have to predict the delivery time for a number of beverages that are to be delivered. Planning the deliveries, choosing effective routes, and adhering to schedules thereby increasing productivity can be done with the help of prediction and estimation.

4. Control: For example, the regression equation is used to check ways to control the strength to suitable values if the engineers in the chemical field want to develop a model related to the tensile strength of a piece of paper to hardwood concentration. It is important to be aware that the variables should be related in a casual manner when using regression equations for control.

When we perform a regression analysis on the data a regression equation along with multiple tables is generated from which we can extract a lot of information.

Regression Equation:

Regression equation is the first output produced after performing regression analysis on the data. To predict new observations, the regression equation describes the statistical relationship between the predictor variables and the response variables.

The regression equation can be written as:

$y = b_0 + b_1x_1 + b_2x_2 + \ldots\ldots + b_nx_n$

The variables represents the following:

- y = response variable
- b_0 = the constant/intercept
- b_1, b_2, b_3…. = coefficients
- x_1, x_2, x_3… = values of the term

The Coefficient Table:

From the information in this table, we can deduce the size and direction of the relationship between the predictor variable and the response variable. A coefficient is essentially the number by which the values of the terms in a regression equation are multiplied by.

Some of the columns within the coefficient table are as follows:

● **SE Coef:** The standard error of the coefficient which estimates the variability between coefficient estimates that you would obtain if you took samples from the same population again and again.

● **Confidence Interval:** These confidence intervals (CI) are a range of values that are likely to contain the true value of the coefficient for each term in the model.

● **T-value:** The value that measures the ratio between the coefficient and its standard error is called T-value.

● **P-value:** The p-value is a probability that measures the evidence against the null hypothesis. Lower probabilities provide stronger evidence against the null hypothesis.

● **VIF:** The variance inflation factor (VIF) can also be used as an indicator of multicollinearity if values are greater than 5.

Model Summary Table:

We can reduce definitions and interpretation for every statistic in the model summary table using this table.

● **S:** S is essentially measured in units of the response variable. These values in the table represent the standard deviation of the distance between the data values and fitted values.

● **R^2:** R^2 can be calculated as $1 - \frac{Error\ Sum\ of\ Squares}{Total\ sum\ of\ squares}$.

These values depict the percentage of variation in the response that the model shows.

● **R^2(adj):** R^2(adj) can be calculated as $1 - \frac{Mean\ Square\ Error}{Total\ Mean\ square}$. This displays the variation in the response that the model shows.

● **R^2(pred):** This value ranges between 0% and 100% and can be calculated by removing each observation from the dataset, estimating the regression equation and observing the model prediction after the observation has been removed.

Analysis of Variance Table:

From this table, we can see definitions and interpretations for every statistic.

● **DF:** Represents the degrees of freedom (DF) are the amount of information in your data. The DF for a term shows how much information that term uses.

● **Adj SS:** Represents Adjusted sums of squares and these values represent measures of variation for different components of the model.

● **Adj MS:** Represents Adjusted mean squares and these values measure how much variation a term explains, regardless of the order they were entered. These values consider the degrees of freedom.

● **Seq SS:** Represents sequential sums of squares which are measures of variation for different components of the model. These depend on the order the terms are entered into the model.

● **Seq MS:** Represents sequential mean squares which measures how much variation a term or a model explains. These values consider the degrees of freedom.

● **F-value:** Is the test statistic used to determine whether the term is associated with the response. This value is also used to determine whether the model is missing higher-order terms that include the predictors.

● **P-Value:** is a probability that measures the evidence against the null hypothesis.

Fits and diagnostics table for Fit Regression Model:

From this table, we can find definitions and interpretations for every statistic in the table.

● **Fit:** Fitted values can also be represented as $\widehat{y}$. For the given predictor values, these values act as point estimates of the mean square response.

● **SE Fit:** Represents standard error of the fit (SE fit) which estimates the variation in the estimated mean response for a specific variable. The confidence interval can be obtained by using the values in this column.

● **Confidence interval for fit (95% CI):** These are ranges of values that are likely to contain the mean response that are observed values of the predictors.

● **Resid:** The difference between the observed value(y) and the corresponding fitted value $\widehat{y}$ is the resid.

● **Std Resid:** The standardized residual can be calculated as $\frac{value\ of\ a\ residual\ (ei)}{estimate\ of\ its\ standard\ deviation}$

● **Del Residuals:** Each deleted Studentized residual can be calculated using a formula that is equivalent to removing each observation from the data set, estimating the regression equation, and checking how well the model predicts observation that has been removed.

● **Hi (leverage):** This is also known as leverage which measures the distance from an observation's x-value to the average of the x-values for all observations in a data set.

● **Cook's distance (D):** This distance uses both the leverage values and the standardized residuals of each observation to determine the effect of an observation.

● **DFITS:** When each observation is removed from the data set DFITS represents (approx) the number of standard deviations that the fitted value changes with these removed observations and the model is refit. It measures the effect each observation has on the fitted values in a linear model.

The following is the regression analysis for the climate change data:

3.1 MODEL EQUATION

Regression Equation

Temp = -127.7 + 0.06632 MEI + 0.00521 CO2 + 0.000064 CH4 - 0.01693 N2O - 0.00728 CFC-11
+ 0.004272 CFC-12 + 0.0959 TSI - 1.582 Aerosols

The dataset we are examining here is the climate change data that contains values regarding the factor that affects the change in temperature. For this model, we have chosen eight predictor variables that affect the change in temperature. The response variable is temperature ultimately based on the various factors. The CO2, CH4, N2O ,CFC-11 and CFC-12 are the important factors because they significantly increase the atmospheric temperature. The next predictor variable is Aerosols which is the additional parameter in the atmosphere added in the environment due to volcanic eruption. TSI and MEI are two predictors variables are added in the environment due to change in the energy and the index of the pacific ocean that increases the atmospheric temperature.

Mrinal Chaudhari
Vaishnavi Solunke

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	-127.7	19.2	(-165.5, -89.9)	-6.65	0.000	
MEI	0.06632	0.00619	(0.05415, 0.07849)	10.72	0.000	1.23
CO2	0.00521	0.00219	(0.00089, 0.00952)	2.38	0.018	28.00
CH4	0.000064	0.000498	(-0.000916, 0.001043)	0.13	0.898	19.13
N2O	-0.01693	0.00784	(-0.03235, -0.00151)	-2.16	0.032	61.04
CFC-11	-0.00728	0.00146	(-0.01015, -0.00440)	-4.98	0.000	31.83
CFC-12	0.004272	0.000876	(0.002548, 0.005996)	4.88	0.000	93.50
TSI	0.0959	0.0140	(0.0683, 0.1234)	6.84	0.000	1.14
Aerosols	-1.582	0.210	(-1.995, -1.169)	-7.53	0.000	1.35

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
0.0918183	74.40%	73.71%	2.67373	72.85%	-585.30	-548.74

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	8	7.32570	74.40%	7.32570	0.915713	108.62	0.000
MEI	1	0.18023	1.83%	0.96917	0.969167	114.96	0.000
CO2	1	5.96520	60.58%	0.04756	0.047564	5.64	0.018
CH4	1	0.06509	0.66%	0.00014	0.000138	0.02	0.898
N2O	1	0.00975	0.10%	0.03935	0.039353	4.67	0.032
CFC-11	1	0.00567	0.06%	0.20911	0.209114	24.80	0.000
CFC-12	1	0.30652	3.11%	0.20038	0.200379	23.77	0.000
TSI	1	0.31465	3.20%	0.39485	0.394846	46.83	0.000
Aerosols	1	0.47860	4.86%	0.47860	0.478599	56.77	0.000
Error	299	2.52075	25.60%	2.52075	0.008431		
Total	307	9.84646	100.00%				

The linear regression plotted with 8 variables can be assumed significant since the p-value of the model is 0 which is lower than the default value of significance α=0.05. But the VIF value CO2, CH4, N2O,CFC-11, CFC-12 is greater than 10, hinting towards the existence of high multicollinearity. This had to be removed to obtain a reduced model. With further investigation on the correlation aspects of the model, we found out that the correlation between CFC-11 and CFC-12 was 0.83. **The Best Subset** and the **Stepwise techniques** were used to obtain a reduced model. On removing these values in the reduced model, the VIF value was brought down; however , the drawback to this reduced model was lower R^2 value. Best subset analysis can found in **section 6**

The analysis of variance and coefficient is attached in the report. This analysis includes the p-value of the regression model and the individual predictor variable. These two analyses are important as it describes the model adequacy of each parameter and variable confidence interval.

The fits and Diagnostics table was too large to display here therefore is attached in the pdf file along with this report.

The pdf of fits and diagnostics for the full model comprises all the unusual observations and large residuals for the reduced linear regression model.

3.2 HYPOTHESIS TEST:

A hypothesis test is a rule that determines whether to accept or reject a claim about a population parameter depending on the evidence provided.
The null and alternate hypothesis help determine whether the association between the response and the term is statistically significant by examining the P-values. We start by comparing the p-value for a term to the significance level to determine the null hypothesis. The null hypothesis is to check if the term's coefficient is equal to zero, which indicates that there is no association between the term and the response. The significance level (denoted as α or alpha) of 0.05 works well.

This level shows 5% risk or a false positive in the event of concluding that an association exists when there is no association.
For the main model, we check the hypothesis of the regression model as a whole as well as the individual predictor variables, to examine if they actually make a difference to the outcome of the model. To define the hypothesis for our model, we first check for linear association in our regression model between all the predictor variables and the response variable. For this, we need to define a Null Hypothesis and an Alternate Hypothesis first.

Null Hypothesis (H_0): The null hypothesis states that a population parameter like the standard deviation, the mean, etc., is equal to a hypothesized value. Often, this is an initial claim, based on previous analyses or previously known knowledge.
The following is the null hypothesis for our model:
$H_0 : \beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$

Alternative Hypothesis (H_1): The alternative hypothesis states that a population parameter is greater, smaller or completely different than the hypothesized value in the null hypothesis. The alternative hypothesis is what one might believe or hope to be proven true.
The following is the alternative hypothesis for our model:
H_1 : *At Least one of* $\beta_i \neq 0$

Here we see the p – Value for the regression is 0.000, which is less than the level of significance (α) which is 0.05. Thus we can reject the Null Hypothesis and say that linear relationship exists among one of the predictors and the response variable

Next, we check for linear association between the response variable and each predictor variable individually. For this, we examine each predictor variable in detail. For a base predictor variable, we need to define a Null Hypothesis and an Alternate Hypothesis first.
Null Hypothesis: $H_0 : \beta_i = 0$
Alternate Hypothesis: $H_1 : \beta_i \neq 0$

Let's test the hypotheses defined for each variable:

For the first predictor variable (MEI) :
Null Hypothesis: $H_0 : \beta_1 = 0$
Alternate Hypothesis: $H_1 : \beta_1 \neq 0$
Here we see the p – value for this statistic is 0.000, which is less than the level of significance (α) which is 0.05. Thus we can reject the Null Hypothesis and say that linear relationship exists among the predictor and the response variable.

For the second predictor variable (CO2) :
Null Hypothesis: $H_0 : \beta_2 = 0$
Alternate Hypothesis: $H_1 : \beta_2 \neq 0$
Here we see the p – value for this statistic is 0.018, which is less than the level of significance (α) which is 0.05. Thus we can reject the Null Hypothesis and say that linear relationship exists among the predictor and the response variable.

For the third predictor variable (CH4) :
Null Hypothesis: $H_0 : \beta_3 = 0$
Alternate Hypothesis: $H_1 : \beta_3 \neq 0$
Here we see the p – value for this statistic is 0.898, which is greater than the level of significance (α) which is 0.05. Thus we failed to reject the Null Hypothesis and say that linear relationship does not exist among the predictor and the response variable.

For the fourth predictor variable (N2O) :
Null Hypothesis: $H_0 : \beta_4 = 0$
Alternate Hypothesis: $H_1 : \beta_4 \neq 0$
Here we see the p – value for this statistic is 0.032, which is less than the level of significance (α) which is 0.05. Thus we can reject the Null Hypothesis and say that linear relationship exists among the predictor and the response variable.

For the fifth predictor variable (CFC-11) :
Null Hypothesis: $H_0 : \beta_5 = 0$
Alternate Hypothesis: $H_1 : \beta_5 \neq 0$
Here we see the p – value for this statistic is 0.000, which is less than the level of significance (α) which is 0.05. Thus we can reject the Null Hypothesis and say that linear relationship exists among the predictor and the response variable.

For the sixth predictor variable (CFC-12) :
Null Hypothesis: $H_0 : \beta_6 = 0$

Alternate Hypothesis: $H_1 : \beta_6 \neq 0$
Here we see the p – value for this statistic is 0.000, which is less than the level of significance (α) which is 0.05. Thus we can reject the Null Hypothesis and say that linear relationship exists among the predictor and the response variable.

For the seventh predictor variable (TSI) :
Null Hypothesis: $H_0 : \beta_7 = 0$
Alternate Hypothesis: $H_1 : \beta_7 \neq 0$
Here we see the p – value for this statistic is 0.000, which is less than the level of significance (α) which is 0.05. Thus we can reject the Null Hypothesis and say that linear relationship exists among the predictor and the response variable.

For the eighth predictor variable (Aerosols) :
Null Hypothesis: $H_0 : \beta_8 = 0$
Alternate Hypothesis: $H_1 : \beta_8 \neq 0$
Here we see the p – value for this statistic is 0.000, which is less than the level of significance (α) which is 0.05. Thus we can reject the Null Hypothesis and say that linear relationship exists among the predictor and the response variable.

3.3 REDUCED MODEL ANALYSIS:

Stepwise Selection of Terms

Candidate terms: MEI, CO2, CH4, N2O, CFC-11, CFC-12, TSI, Aerosols

	------Step 1------		------Step 2------		------Step 3------	
	Coef	P	Coef	P	Coef	P
Constant	-3.593		-3.808		-3.376	
CO2	0.010599	0.000	0.011153	0.000	0.010022	0.000
MEI			0.04883	0.000	0.06313	0.000
Aerosols					-1.524	0.000
TSI						
N2O						
S	0.118954		0.110157		0.103162	
R-sq	56.03%		62.41%		67.14%	
R-sq(adj)	55.88%		62.17%		66.82%	
Mallows' Cp	209.59		137.00		83.76	
AICc	-433.34		-479.62		-518.98	
BIC	-422.23		-464.83		-500.53	

	-------Step 4------		------Step 5------	
	Coef	P	Coef	P
Constant	-138.3		-136.3	
CO2	0.009857	0.000	0.00575	0.011
MEI	0.06817	0.000	0.06903	0.000
Aerosols	-1.721	0.000	-1.745	0.000
TSI	0.0988	0.000	0.0961	0.000
N2O			0.01010	0.064
S	0.0955710		0.0951857	
R-sq	71.89%		72.21%	
R-sq(adj)	71.52%		71.75%	
Mallows' Cp	30.27		28.56	
AICc	-564.99		-566.41	
BIC	-542.89		-540.67	

α to enter = 0.15, α to remove = 0.15

Regression Equation

Temp = -136.3 + 0.06903 MEI + 0.00575 CO2 + 0.01010 N2O + 0.0961 TSI - 1.745 Aerosols

Mrinal Chaudhari
Vaishnavi Solunke

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	-136.3	18.8	(-173.3, -99.3)	-7.25	0.000	
MEI	0.06903	0.00625	(0.05673, 0.08132)	11.05	0.000	1.16
CO2	0.00575	0.00226	(0.00131, 0.01019)	2.55	0.011	27.57
N2O	0.01010	0.00543	(-0.00059, 0.02079)	1.86	0.064	27.31
TSI	0.0961	0.0138	(0.0689, 0.1233)	6.96	0.000	1.03
Aerosols	-1.745	0.215	(-2.168, -1.323)	-8.13	0.000	1.32

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
0.0951857	72.21%	71.75%	2.84510	71.11%	-566.41	-540.67

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	5	7.11024	72.21%	7.11024	1.42205	156.95	0.000
MEI	1	0.18023	1.83%	1.10594	1.10594	122.06	0.000
CO2	1	5.96520	60.58%	0.05894	0.05894	6.50	0.011
N2O	1	0.03820	0.39%	0.03133	0.03133	3.46	0.064
TSI	1	0.32801	3.33%	0.43829	0.43829	48.37	0.000
Aerosols	1	0.59860	6.08%	0.59860	0.59860	66.07	0.000
Error	302	2.73622	27.79%	2.73622	0.00906		
Total	307	9.84646	100.00%				

The fits and Diagnostics table was too large to display here therefore is attached in the pdf file along with this report.

The pdf of fits and diagnostics for the reduced model comprises all the unusual observations and large residuals for the reduced linear regression model.

The reduced model is calculated using the best model that fits the criteria and suits the response variable. If we compare this model to the main model before stepwise was implemented, we see that the predictor variables 'CH4', 'CFC-11', and 'CFC-12' are eliminated from the model. The reason for this was the effect that it had on the model as a whole and so the decision was made for its removal. After this, the model is trained again on the remaining predictor variables and we see the regression analysis given above. If we compare the reduced model's ANOVA table to that of the model before we used the reduced model, we see a clear difference in a lot of values to a better effect. F-values and P-values have had a significant impact on, in the reduced model.

Main Model:

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	8	7.32570	74.40%	7.32570	0.915713	108.62	0.000
MEI	1	0.18023	1.83%	0.96917	0.969167	114.96	0.000
CO2	1	5.96520	60.58%	0.04756	0.047564	5.64	0.018
CH4	1	0.06509	0.66%	0.00014	0.000138	0.02	0.898
N2O	1	0.00975	0.10%	0.03935	0.039353	4.67	0.032
CFC-11	1	0.00567	0.06%	0.20911	0.209114	24.80	0.000
CFC-12	1	0.30652	3.11%	0.20038	0.200379	23.77	0.000
TSI	1	0.31465	3.20%	0.39485	0.394846	46.83	0.000
Aerosols	1	0.47860	4.86%	0.47860	0.478599	56.77	0.000
Error	299	2.52075	25.60%	2.52075	0.008431		
Total	307	9.84646	100.00%				

Reduced Model:

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	5	7.11024	72.21%	7.11024	1.42205	156.95	0.000
MEI	1	0.18023	1.83%	1.10594	1.10594	122.06	0.000
CO2	1	5.96520	60.58%	0.05894	0.05894	6.50	0.011
N2O	1	0.03820	0.39%	0.03133	0.03133	3.46	0.064
TSI	1	0.32801	3.33%	0.43829	0.43829	48.37	0.000
Aerosols	1	0.59860	6.08%	0.59860	0.59860	66.07	0.000
Error	302	2.73622	27.79%	2.73622	0.00906		
Total	307	9.84646	100.00%				

Now that we have created a reduced model, we then check for multicollinearity again in the reduced model. This updated table gives us an idea of refined VIF values and how they relate to other predictor variables as well as the response variable.

3.5 MULTICOLLINEARITY:

Proper validation of a regression model includes a study of the coefficients to determine if their signs and magnitudes are reasonable. Multicollinearity diagnostics are also an important guide to the validity of the model. Below are some of the methods to measure multicollinearity in a linear regression model.

1. Correlation Matrix
2. Variance Inflation Factors(VIF)

3. High R^2 but few significant regressors.

Multicollinearity is the situation when two or more independent variables in a multiple regression model are correlated with each other.

For multicollinearity diagnosis of our model, we relied on VIF values of the regressors and worked on remedial measures such as stepwise and best subsets for a better fit/ reduced model.

Rules to analyze variance inflation factor (VIF):

1. If VIF = 1, there is no multicollinearity.
2. If 1 < VIF < 5, there is small multicollinearity.
3. If VIF ≥ 5, there is medium multicollinearity.
4. If VIF ≥ 10, there is large multicollinearity.

Below are the predictor variables with extreme VIF values. We can infer that, if any VIF exceeds 5 or 10, that particular coefficient is poorly estimated or unstable because of near - linear dependences among the regressors.

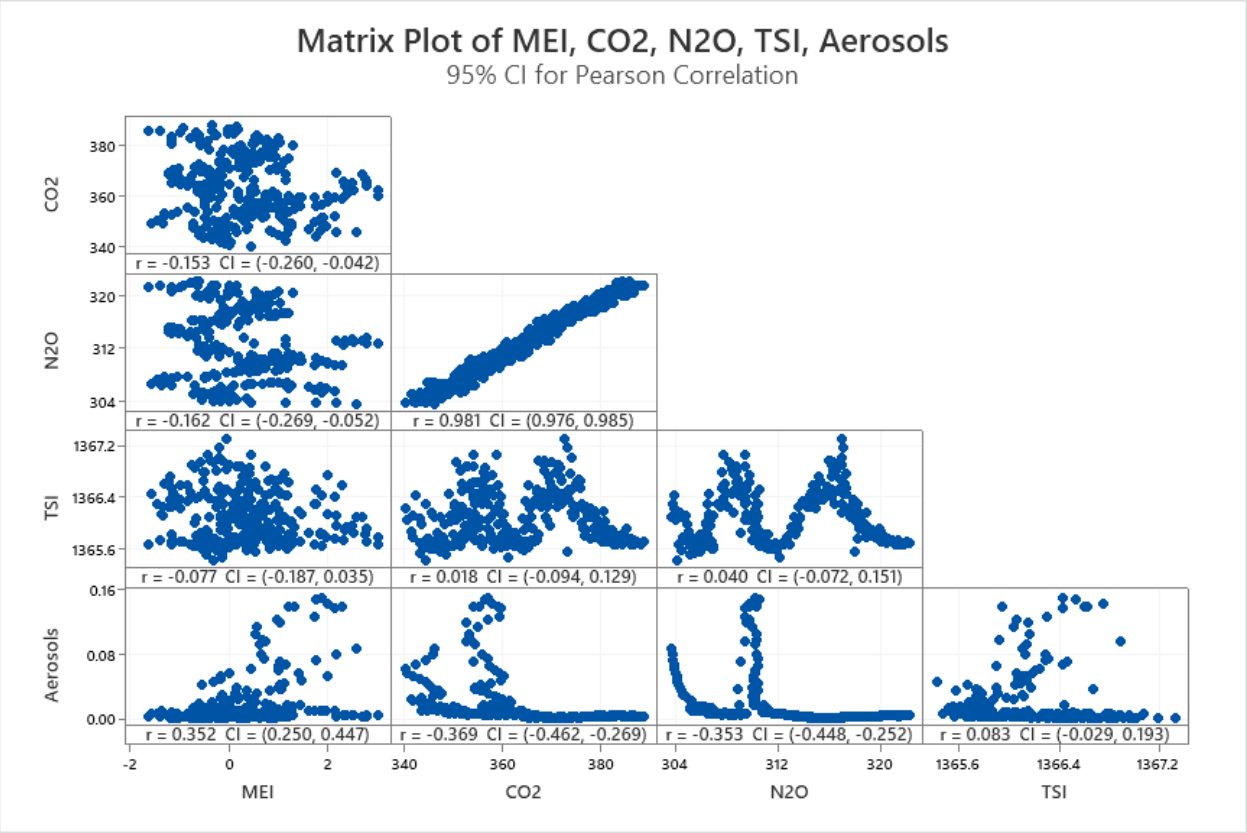
Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	-136.3	18.8	(-173.3, -99.3)	-7.25	0.000	
MEI	0.06903	0.00625	(0.05673, 0.08132)	11.05	0.000	1.16
CO2	0.00575	0.00226	(0.00131, 0.01019)	2.55	0.011	27.57
N2O	0.01010	0.00543	(-0.00059, 0.02079)	1.86	0.064	27.31
TSI	0.0961	0.0138	(0.0689, 0.1233)	6.96	0.000	1.03
Aerosols	-1.745	0.215	(-2.168, -1.323)	-8.13	0.000	1.32

In this Coefficients table, we see the refined VIF values and their effect on other variables. If we look at the VIF column now, we see that the variables CO2 and N2O still have the highest value

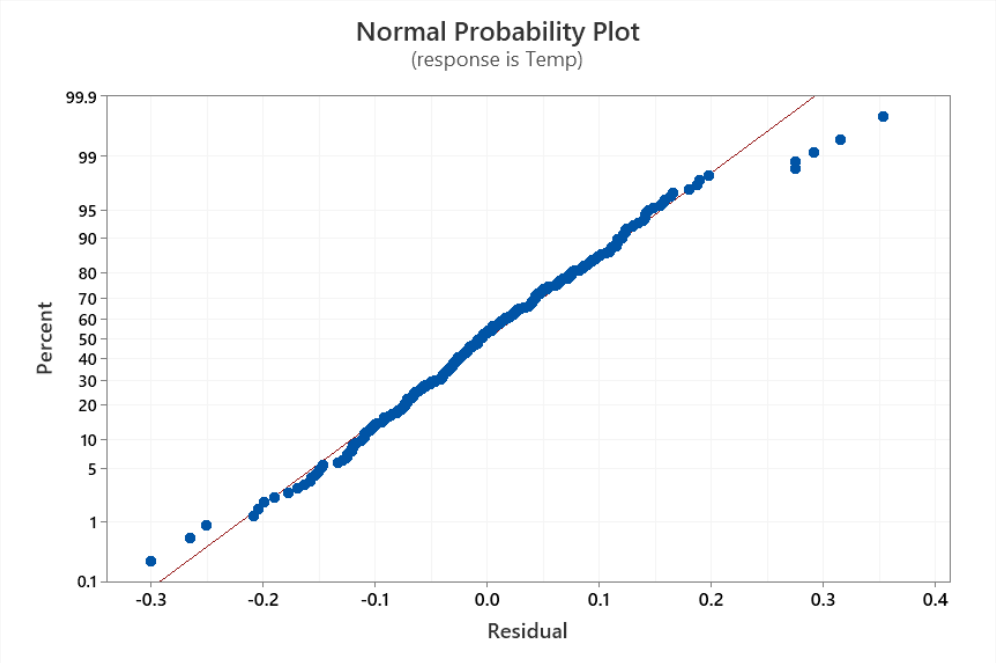
of 27.57 and 27.31 respectively and the variable TSI still has the lowest value of 1.03. Even though all the values have collectively decreased, we still have some high multicollinearity in a few variables.

Here we have large multicollinearity with variables CO2 and N2O (VIF>10). The correlation matrix graph (generated using Minitab) for the predictor variables is given below. It is one of the measures for multicollinearity diagnosis.



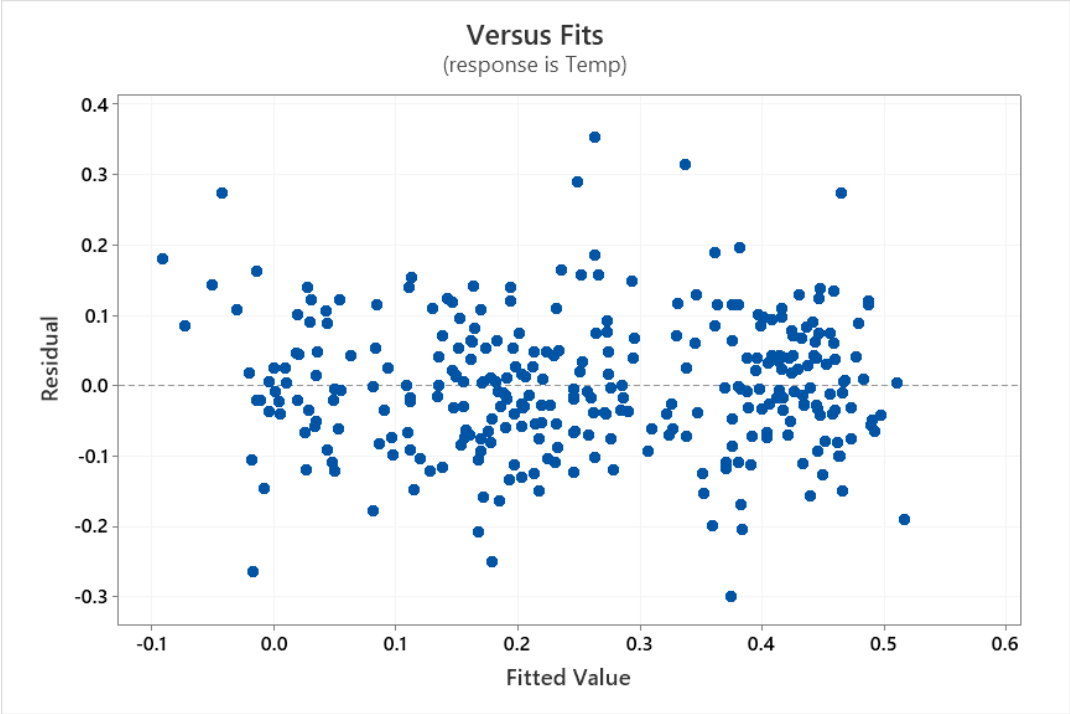
3.6 REDUCED MODEL ASSUMPTIONS:

1. Normal probability plot:



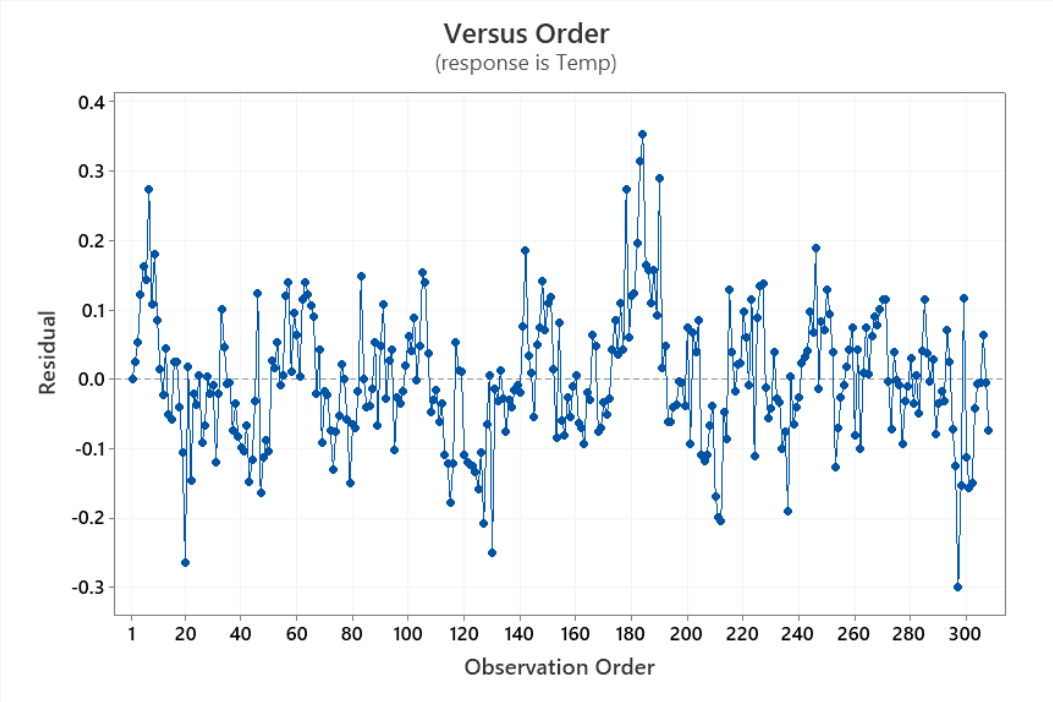
A normal probability graph is used to test the normality assumption. Normal probability graph residuals should approximately follow the same line.
The above is the normal probability plot for the reduced model. The residuals seem to follow the same line and thus we say that normality assumption is valid.

2. Versus fit plot:



In the versus fit plot, we see that the constant variance assumptions are not violated as there is no specific pattern followed by the residuals. Therefore, we say that the constant variance assumption is valid.

3. Versus order plot:



In the versus order plot, we see that the assumption of independence is not violated as there is no specific pattern followed by the residuals. Therefore, we say that the assumption of independence is valid.

Thus, we say that the reduced model appears to be reasonable.

3.7 HYPOTHESIS TEST FOR REDUCED MODEL:

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	-136.3	18.8	(-173.3, -99.3)	-7.25	0.000	
MEI	0.06903	0.00625	(0.05673, 0.08132)	11.05	0.000	1.16
CO2	0.00575	0.00226	(0.00131, 0.01019)	2.55	0.011	27.57
N2O	0.01010	0.00543	(-0.00059, 0.02079)	1.86	0.064	27.31
TSI	0.0961	0.0138	(0.0689, 0.1233)	6.96	0.000	1.03
Aerosols	-1.745	0.215	(-2.168, -1.323)	-8.13	0.000	1.32

Using the above coefficients table to do the hypothesis test on the reduced model. To define the hypothesis for our reduced model, we first check for linear association in our regression model between all the predictor variables and the response variable. For this, we need to define a Null Hypothesis and an Alternate Hypothesis first.

Null Hypothesis for reduced model:
$H_0 : \beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

Alternate Hypothesis for reduced model:
H_1 : *At Least one of* $\beta_i \neq 0$

Here we see the p – value for the regression is 0.000, which is less than the level of significance (α) which is 0.05. Thus we can reject the Null Hypothesis and say that linear relationship exists among one of the predictors and the response variable

Next, we check for linear association between the response variable and each predictor variable individually. For this, we examine each predictor variable in detail. For a base predictor variable, we need to define a Null Hypothesis and an Alternate Hypothesis first.
Null Hypothesis: $H_0 : \beta_i = 0$
Alternate Hypothesis: $H_1 : \beta_i \neq 0$

Let's test the hypotheses defined for each variable.
For the first predictor variable (MEI) :
Null Hypothesis: $H_0 : \beta_1 = 0$
Alternate Hypothesis: $H_1 : \beta_1 \neq 0$

Here we see the p – value for this statistic is 0.000, which is less than the level of significance (α) which is 0.05. Thus we can reject the Null Hypothesis and say that linear relationship exists among the predictor and the response variable.

For the second predictor variable (CO2) :
Null Hypothesis: $H_0 : \beta_1 = 0$
Alternate Hypothesis: $H_1 : \beta_1 \neq 0$
Here we see the p – value for this statistic is 0.011, which is less than the level of significance (α) which is 0.05. Thus we can reject the Null Hypothesis and say that linear relationship exists among the predictor and the response variable.

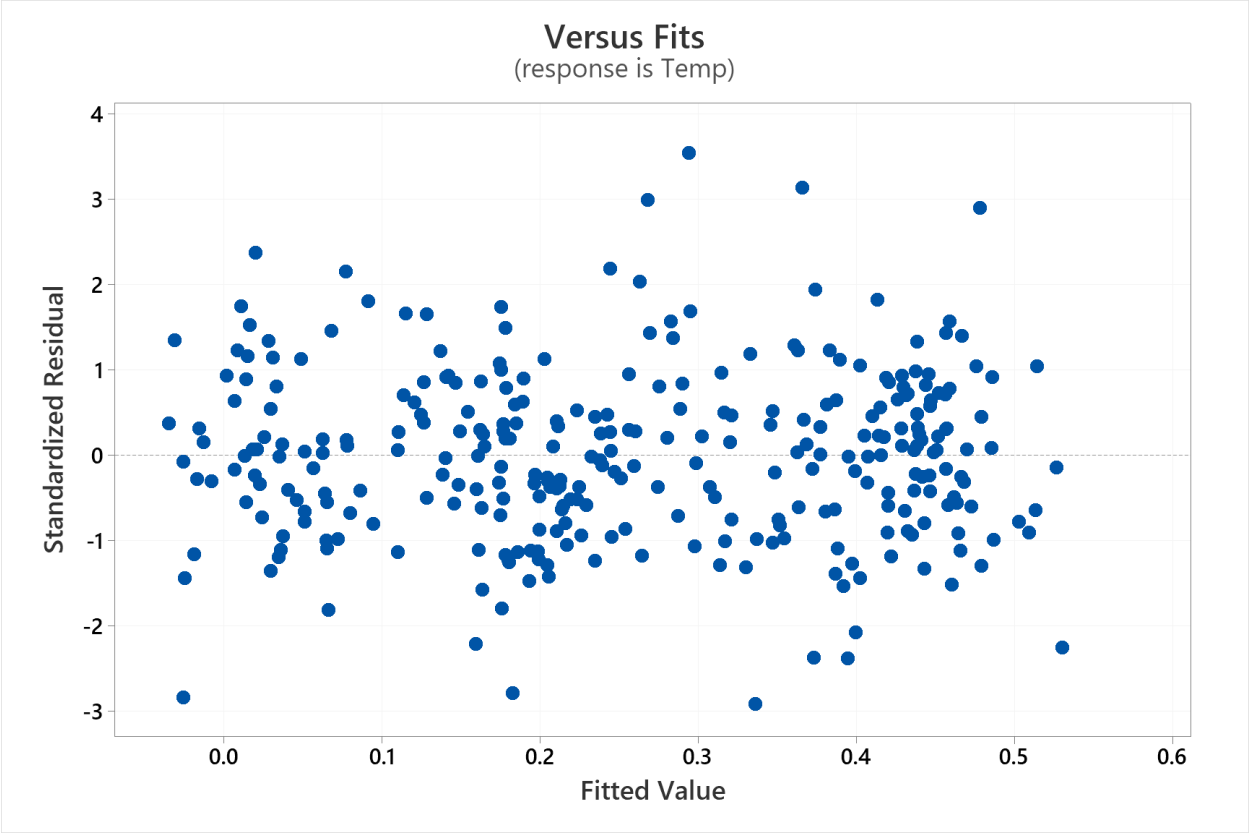
For the third predictor variable (N2O) :
Null Hypothesis: $H_0 : \beta_1 = 0$
Alternate Hypothesis: $H_1 : \beta_1 \neq 0$
Here we see the p – value for this statistic is 0.064, which is greater than the level of significance (α) which is 0.05. Thus we failed to reject the Null Hypothesis and say that linear relationship does not exist among the predictor and the response variable.

For the fourth predictor variable (TSI) :
Null Hypothesis: $H_0 : \beta_1 = 0$
Alternate Hypothesis: $H_1 : \beta_1 \neq 0$
Here we see the p – value for this statistic is 0.000, which is less than the level of significance (α) which is 0.05. Thus we can reject the Null Hypothesis and say that linear relationship exists among the predictor and the response variable.

For the fifth predictor variable (Aerosols) :
Null Hypothesis: $H_0 : \beta_1 = 0$
Alternate Hypothesis: $H_1 : \beta_1 \neq 0$
Here we see the p – value for this statistic is 0.000, which is less than the level of significance (α) which is 0.05. Thus we can reject the Null Hypothesis and say that linear relationship exists among the predictor and the response variable.

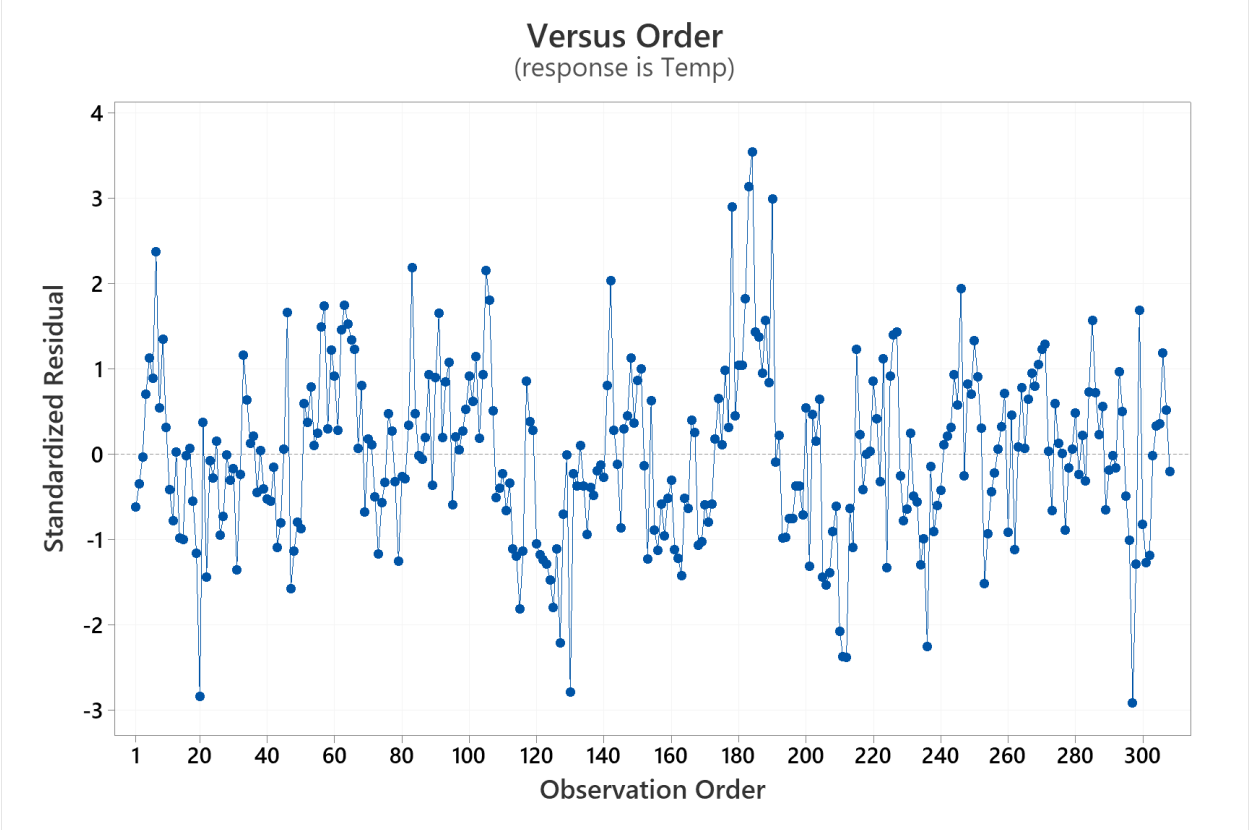
3.8 ASSUMPTIONS FOR THE REGRESSION MODEL:

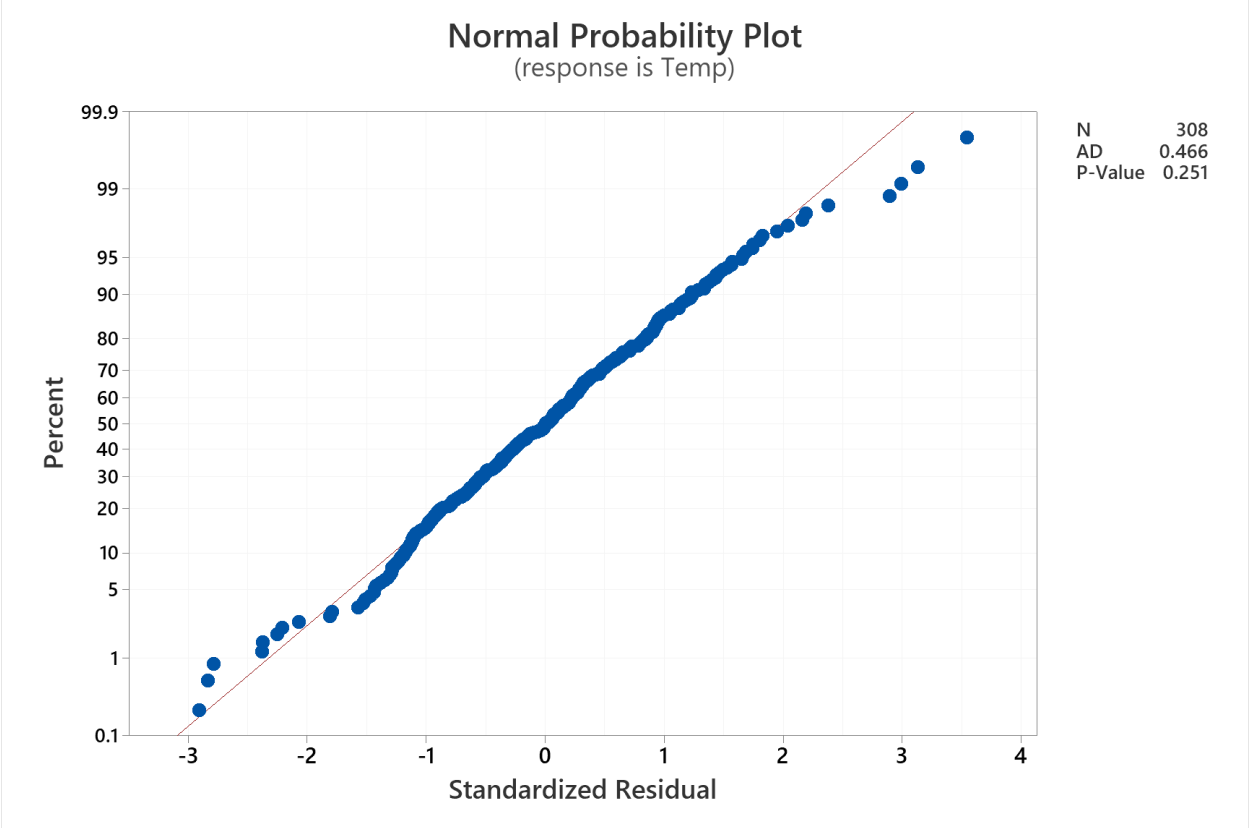
The assumption for the linear regression for this project is as follows:



1. The relationship between the response y and regressors is linear. In the residual graph, the observations are independent of each other.
2. There are few outliers present in the graph and the data points are approximately symmetric around the zero line.

Mrinal Chaudhari
Vaishnavi Solunke

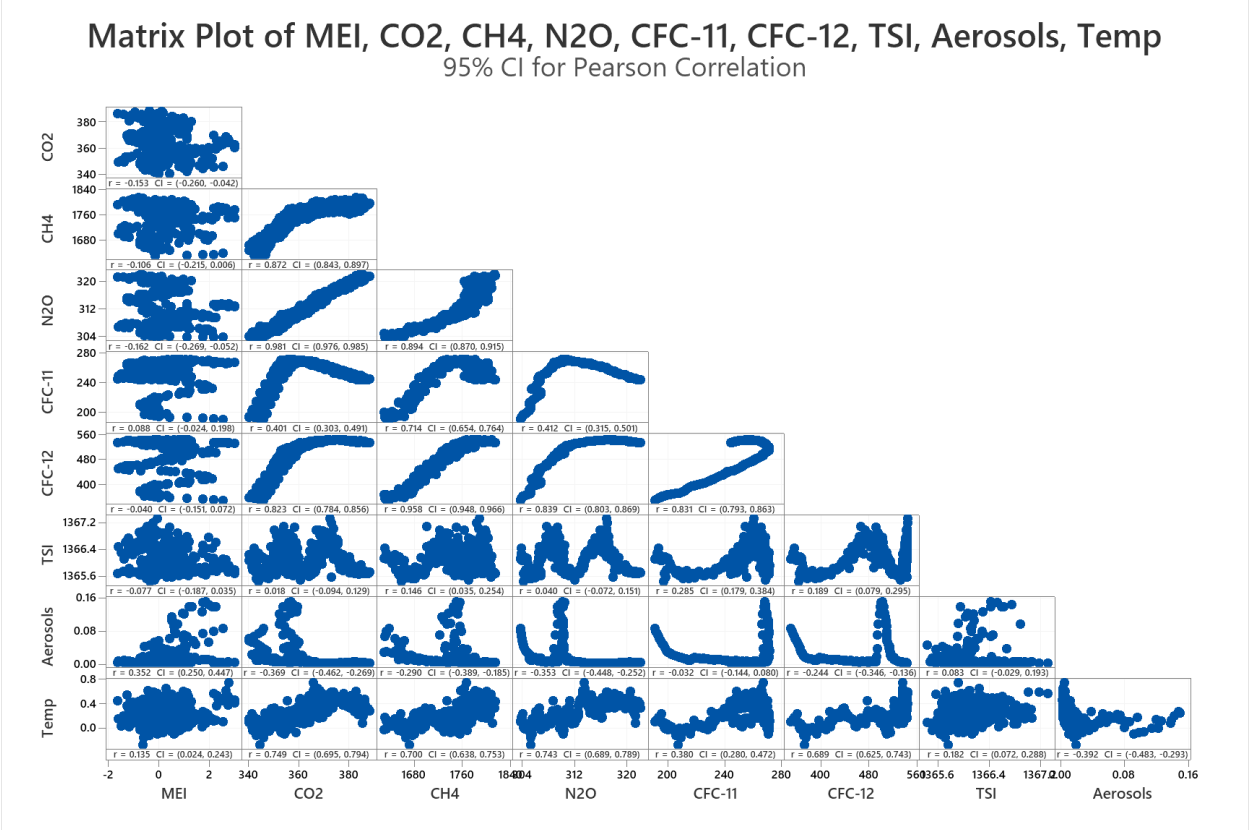




1. A normal probability graph is used to test the normality assumption. Normal probability graph residuals should approximately follow the same line. There does not seem to be any problem with the normality assumption.
2. In the verses plot, we can state that the points are randomly scattered and we cannot see any reasonable pattern so the regression is the better choice and constant variance assumptions are not violated.
3. We can state that there is a strong statistical relationship between Response variable and predictors. The model appears to be reasonable.

Correlations

	MEI	CO2	CH4	N2O	CFC-11	CFC-12	TSI	Aerosols
CO2	-0.153							
CH4	-0.106	0.872						
N2O	-0.162	0.981	0.894					
CFC-11	0.088	0.401	0.714	0.412				
CFC-12	-0.040	0.823	0.958	0.839	0.831			
TSI	-0.077	0.018	0.146	0.040	0.285	0.189		
Aerosols	0.352	-0.369	-0.290	-0.353	-0.032	-0.244	0.083	
Temp	0.135	0.749	0.700	0.743	0.380	0.689	0.182	-0.392



1. We can see that the diagonal values are not close to 1 and therefore the errors are unrelated.

3.9 MODEL ADEQUACY CHECKING :

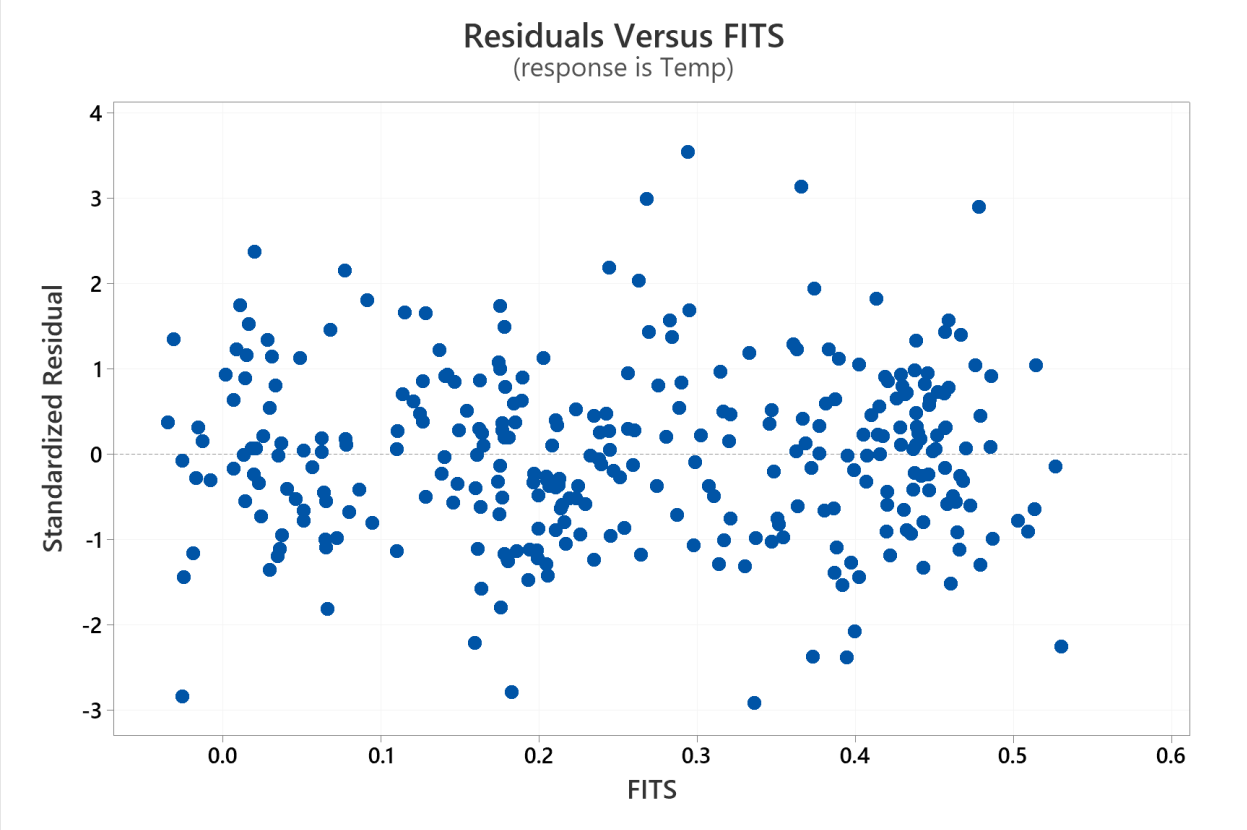
There are several methods useful for diagnosing the validity of a linear regression model. They are listed below in detail:

3.9.1 Residual Analysis:

A residual may be seen as the deviation between the data and the fit value of the response variable, it is also a measure of the variability in the response variable not explained by the regression model.

As we have seen, plotting residuals is a very effective way to investigate how well the regression model fits the data and to check the assumptions listed in this section. For example, a Residual vs Fitted Value can have following valuable information about the data and interpretations.

(a) If the residuals "bounce randomly" around the 0 line, it suggests that the assumption that the relationship is linear is reasonable.
(b) If the residuals roughly form a "horizontal band" around the 0 line, it suggests that the variances of the error terms are equal.
(c) If a residual "stands out" from the basic random pattern of residuals, it suggests the presence of outliers.



For instance, the graph shown below is a Standardized Residual vs Fitted value plot for the Linear Regression model fitted on our data for climate change. We have made an interpretation of the graph in **Section 3.5.** It validates the linearity of the model, constant error variance and indicates the existence of few outliers and unusual observations in our dataset.

3.9.2 Outliers:

Sometimes it is effective and efficient to work with scaled residuals. Mentioned below are the four popular scaling methods. These scaled residuals are pragmatic in finding observations that are outliers or extreme values present in the dataset.

(a) Standardized Residuals.

(b) Studentized Residuals.

(c) PRESS (Prediction error sum of squares) Residuals.

(d) R-student Residuals

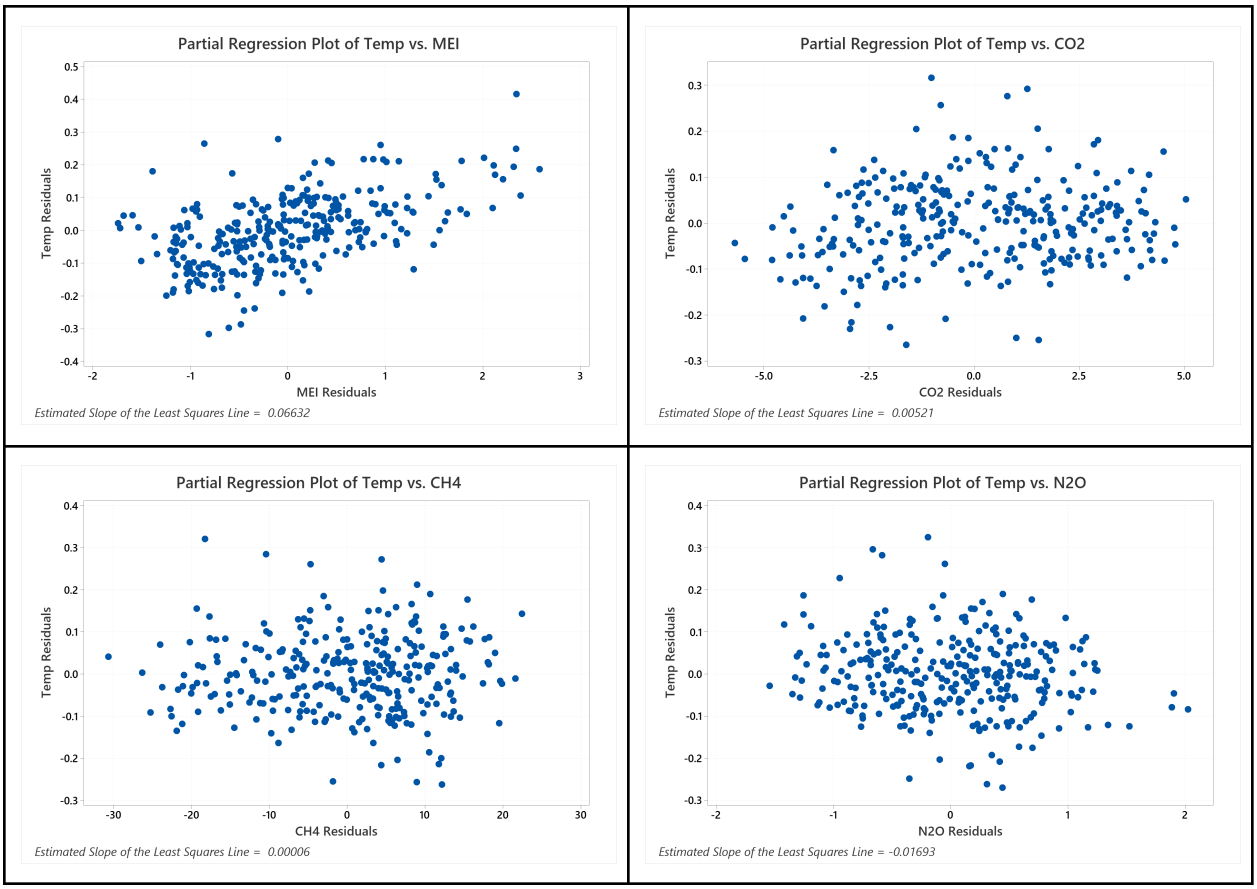
We can see using the Residual plots that there are few outliers in our data for Climate Change. ***The list of large residuals and unusual observations are given in the pdf file.***

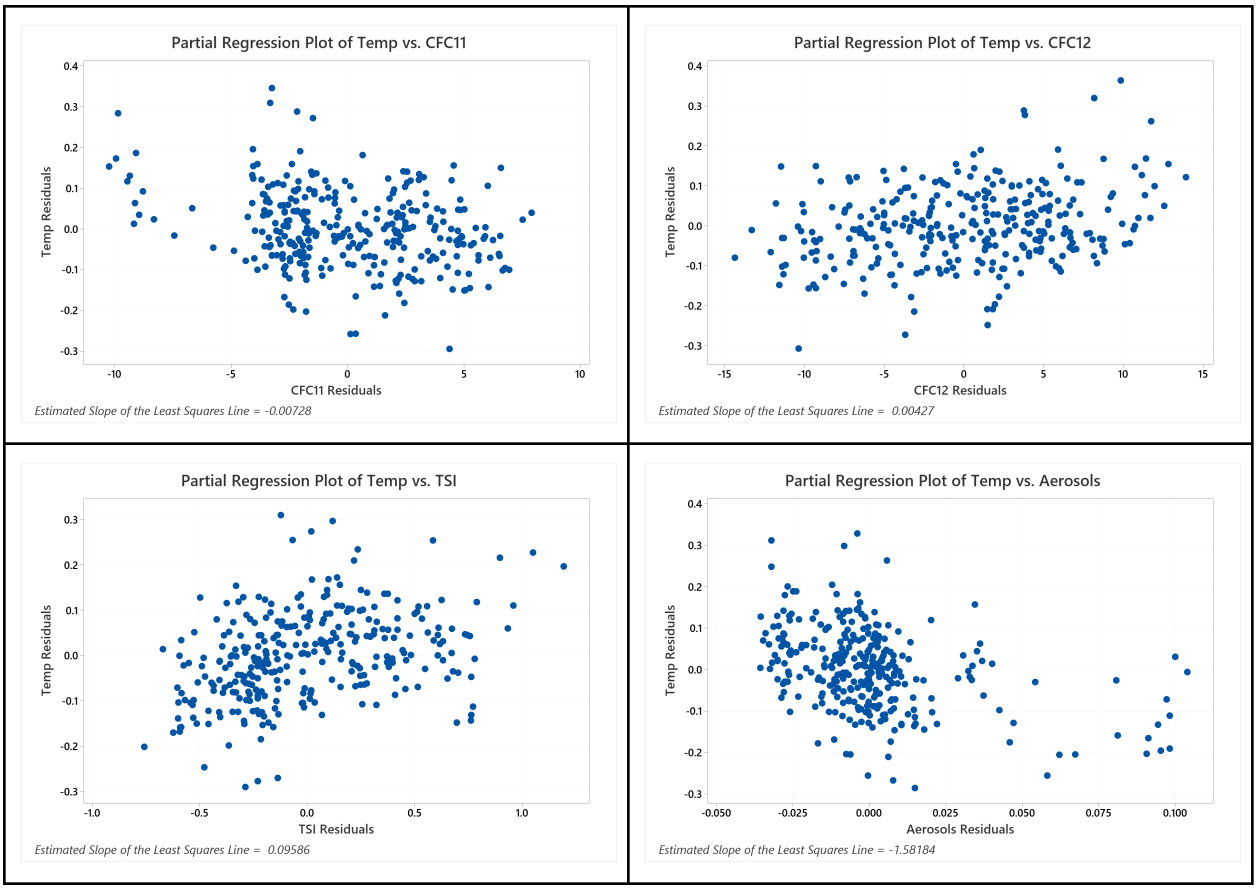
3.9.3 Partial Plots:

When performing linear regression with more than one independent variable/ predictors, partial plots or added variable plots attempt to show the effect of adding an additional variable on the model. In our project we tried mapping some information about the usefulness of the variable which is not currently part of the model.

An **added-variable plot** is an effective way to show the correlation between an independent variable and a dependent variable conditional on other independent variables. Added-variable plots are also useful for spotting influential outliers in the data which affect the estimated regression parameters.

Below are the partial plots of each predictor variable present in our dataset.





3.9.4 Lack of fit Test:

A regression model displays a lack-of-fit error when it fails to properly describe the functional relationship between the predictor factors and the response variable. Lack-of-fit can occur if important terms from the model such as interactions or quadratic (or polynomial) terms are not included. In our project, a lack of fit test has not been displayed in the minitab output. That means, the variables are independent from each other and do not contain the identical values in different predictors.

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	8	7.32570	74.40%	7.32570	0.915713	108.62	0.000
MEI	1	0.18023	1.83%	0.96917	0.969167	114.96	0.000
CO2	1	5.96520	60.58%	0.04756	0.047564	5.64	0.018
CH4	1	0.06509	0.66%	0.00014	0.000138	0.02	0.898
N2O	1	0.00975	0.10%	0.03935	0.039353	4.67	0.032
CFC11	1	0.00567	0.06%	0.20911	0.209114	24.80	0.000
CFC12	1	0.30652	3.11%	0.20038	0.200379	23.77	0.000
TSI	1	0.31465	3.20%	0.39485	0.394846	46.83	0.000
Aerosols	1	0.47860	4.86%	0.47860	0.478599	56.77	0.000
Error	299	2.52075	25.60%	2.52075	0.008431		
Total	307	9.84646	100.00%				

3.9.5 Multicollinearity Diagnostic for main model:

Proper validation of a regression model includes a study of the coefficients to determine if their signs and magnitudes are reasonable. Multicollinearity diagnostics are also an important guide to the validity of the model. Below are some of the methods to measure multicollinearity in a linear regression model.

4. Correlation Matrix
5. Variance Inflation Factors(VIF)
6. High R^2but few significant regressors.

Multicollinearity is the situation when two or more independent variables in a multiple regression model are correlated with each other.

For multicollinearity diagnosis of our model, we relied on VIF values of the regressors and worked on remedial measures such as stepwise and best subsets for a better fit/ reduced model.

Rules to analyze variance inflation factor (VIF):

5. If VIF = 1, there is no multicollinearity.
6. If 1 < VIF < 5, there is small multicollinearity.
7. If VIF ≥ 5, there is medium multicollinearity.
8. If VIF ≥ 10, there is large multicollinearity.

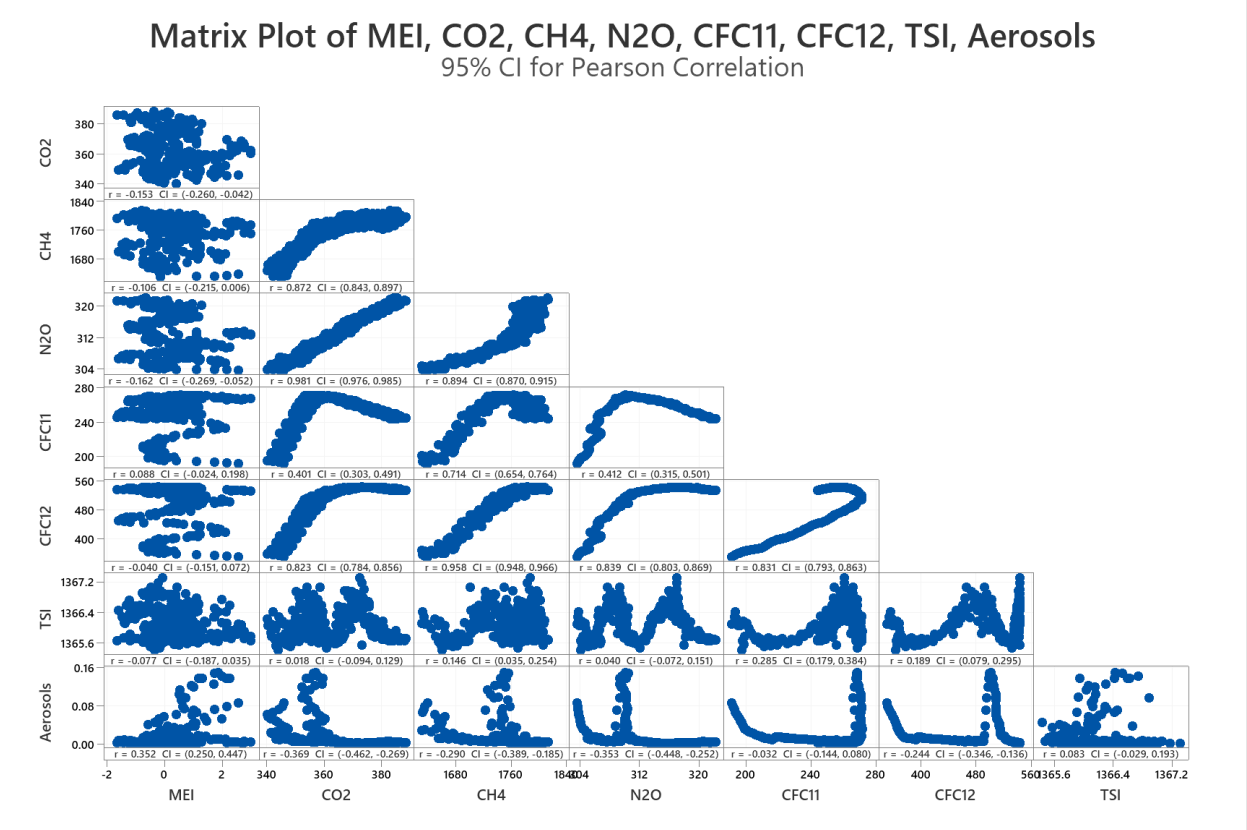
Below are the predictor variables with extreme VIF values. We can infer that, if any VIF exceeds 5 or 10, that particular coefficient is poorly estimated or unstable because of near - linear dependences among the regressors.

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-127.7	19.2	-6.65	0.000	
MEI	0.06632	0.00619	10.72	0.000	1.23
CO2	0.00521	0.00219	2.38	0.018	28.00
CH4	0.000064	0.000498	0.13	0.898	19.13
N2O	-0.01693	0.00784	-2.16	0.032	61.04
CFC11	-0.00728	0.00146	-4.98	0.000	31.83
CFC12	0.004272	0.000876	4.88	0.000	93.50
TSI	0.0959	0.0140	6.84	0.000	1.14
Aerosols	-1.582	0.210	-7.53	0.000	1.35

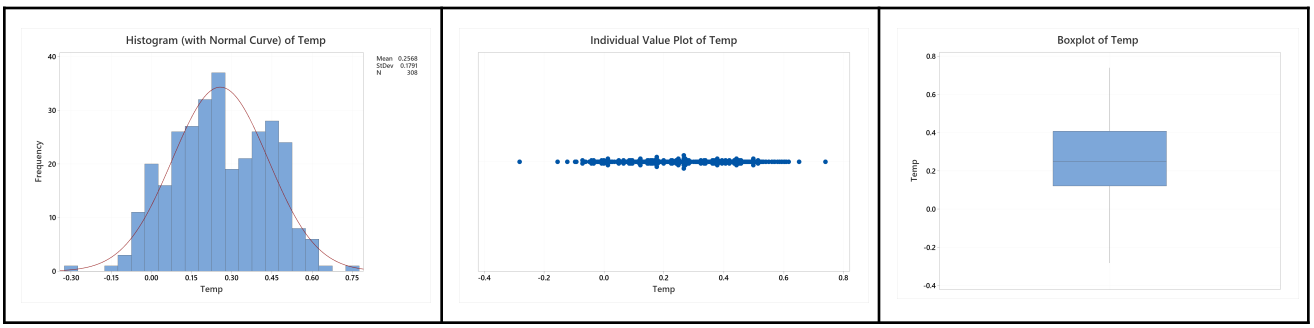
Here we have large multicollinearity with variables CO2, CH4, N2O, CFC11, CFC12 (VIF>10).

The correlation matrix graph (generated using Minitab) for the predictor variables is given below. It is one of the measures for multicollinearity diagnosis.



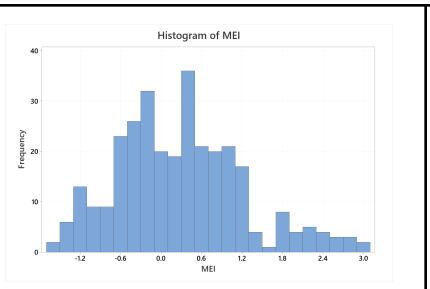
4. EXPLORATORY DATA ANALYSIS

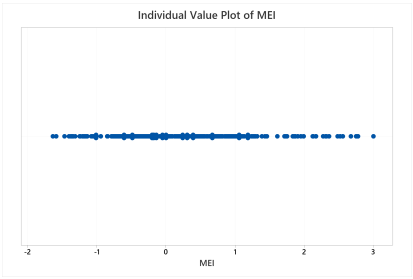
The exploratory data analysis gives the histogram and the scatterplots of the data. The histogram has the normal line fitted on them. We have done this using Minitab software[2]. We have defined all the predictor variables that are used for the building of the model before and have also seen the relevance of these variables as well. After this, now we see below graphs that explore each predictors variable and its result.

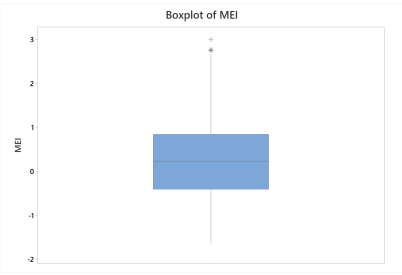


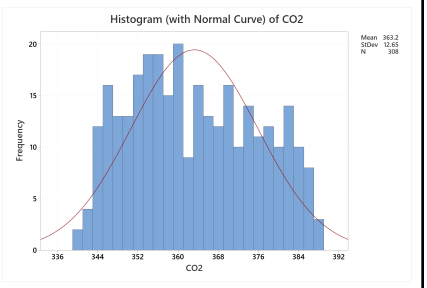
Linear Regression Project

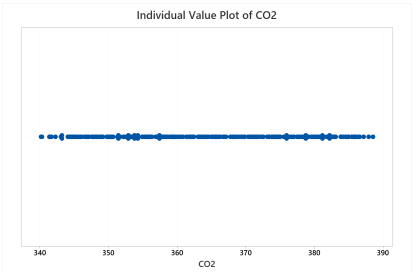
Mrinal Chaudhari
Vaishnavi Solunke

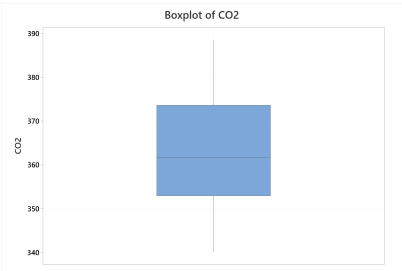


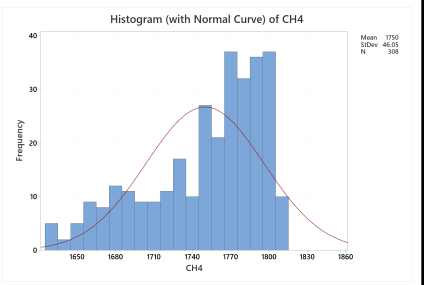


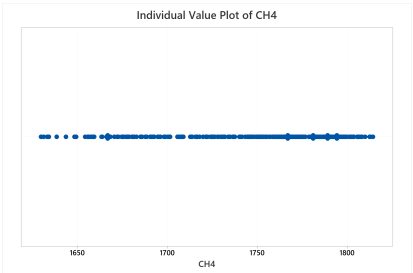


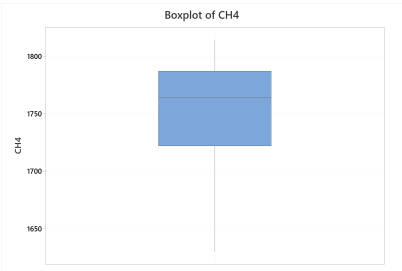


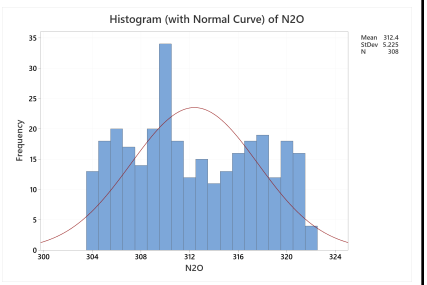


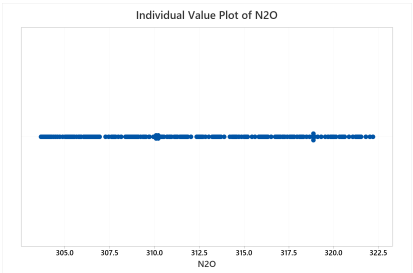


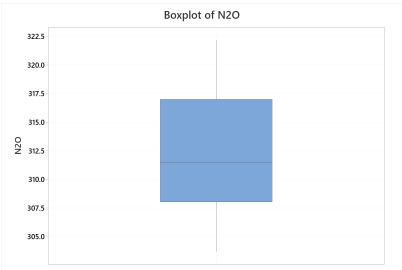


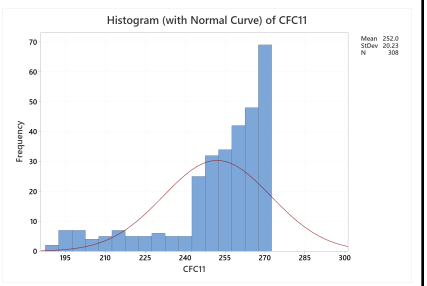


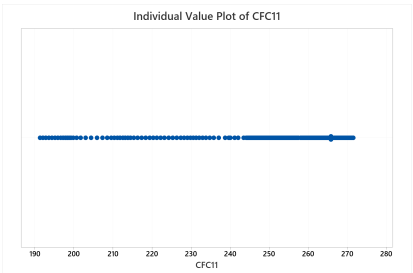


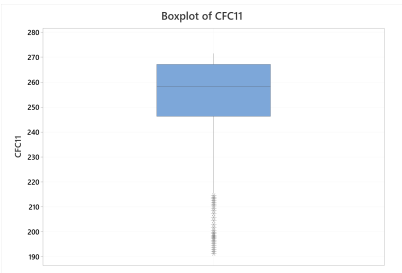


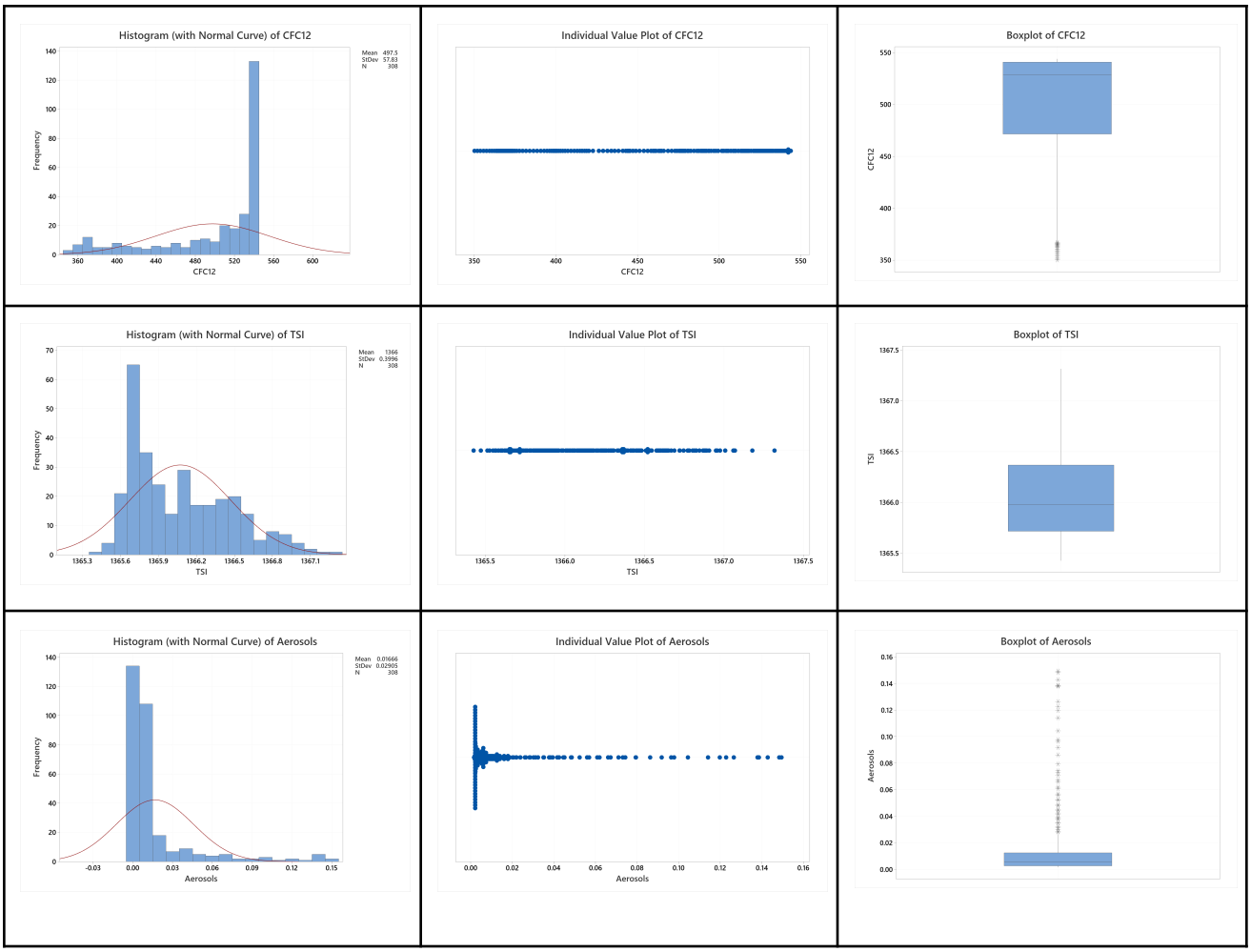












5. SUMMARY STATISTICS

To describe our data we will summarize the data talking about the mean, standard deviation, standard error mean, variance, skewness, median, mode and range along with boxplots and histograms for each variable (predictors) [**section 4.1**].

Output: Minitab[2]

Mrinal Chaudhari
Vaishnavi Solunke

Statistics

Variable	Total Count	N	N*	CumN	Percent	CumPct	Mean	SE Mean	TrMean	StDev
MEI	308	308	0	308	100	100	0.2756	0.0534	0.2397	0.9379
CO2	308	308	0	308	100	100	363.23	0.721	363.11	12.65
CH4	308	308	0	308	100	100	1749.8	2.62	1752.4	46.1
N2O	308	308	0	308	100	100	312.39	0.298	312.35	5.23
CFC11	308	308	0	308	100	100	251.97	1.15	253.95	20.23
CFC12	308	308	0	308	100	100	497.52	3.29	502.43	57.83
TSI	308	308	0	308	100	100	1366.1	0.0228	1366.0	0.400
Aerosols	308	308	0	308	100	100	0.01666	0.00166	0.01163	0.02905
Temp	308	308	0	308	100	100	0.2568	0.0102	0.2572	0.1791

Variable	Variance	CoefVar	Sum	Sum of Squares	Minimum	Q1	Median	Q3
MEI	0.8797	340.37	84.8710	293.4518	-1.6350	-0.4042	0.2375	0.8455
CO2	159.95	3.48	111873.84	40684676.26	340.17	352.96	361.74	373.63
CH4	2120.8	2.63	538946.0	943711906.9	1629.9	1722.0	1764.0	1787.1
N2O	27.30	1.67	96216.68	30065688.44	303.68	308.03	311.51	317.00
CFC11	409.33	8.03	77607.70	19680714.33	191.32	246.27	258.34	267.14
CFC12	3343.95	11.62	153237.63	77266112.74	350.11	471.41	528.36	540.66
TSI	0.160	0.03	420749.8	574774039.2	1365.4	1365.7	1366.0	1366.4
Aerosols	0.00084	174.40	5.13040	0.34453	0.00160	0.00280	0.00575	0.01260
Temp	0.0321	69.75	79.0870	30.1541	-0.2820	0.1212	0.2480	0.4078

Variable	Maximum	Range	IQR	Mode	N for Mode	Skewness
MEI	3.0010	4.6360	1.2498	-1.011, -0.607, -0.487, -0.201	2	0.54
CO2	388.50	48.33	20.67	343.2, 351.44, 352.89, 353.79	2	0.18
CH4	1814.2	184.3	65.1	1666.83, 1766.96, 1781.02, 1789.02	2	-0.83
N2O	322.18	18.50	8.97	318.866	3	0.15
CFC11	271.49	80.17	20.86	265.784	2	-1.46
CFC12	543.81	193.70	69.25	542.279	2	-1.23
TSI	1367.3	1.89	0.651	1365.65, 1365.66, 1365.72, 1366.36	2	0.72
Aerosols	0.14940	0.14780	0.00980	0.0021	36	2.98
Temp	0.7390	1.0210	0.2865	0.266	5	-0.03

Variable	Kurtosis	MSSD
MEI	0.16	0.0449
CO2	-1.07	0.795
CH4	-0.34	38.5
N2O	-1.20	0.010
CFC11	1.34	0.14
CFC12	0.13	0.49
TSI	-0.40	0.0236
Aerosols	8.69	0.00001
Temp	-0.63	0.0036

The data contain at least five mode values. Only the smallest four are shown.

6. BEST SUBSETS REGRESSION

After the establishment of the main model, its detailed regression analysis and examination of the summary statistics and assumption, we inspect the model based on predictor variables. This is done to see which approach and what predictor variables form the best possible model regarding the given response variable.

The result of the best subset run on the data using minitab is given below.

CLIMATE_CHANGE.CSV

Best Subsets Regression: Temp versus MEI, CO2, CH4, N2O, CFC11, CFC12, TSI, Aerosols

Response is Temp

Vars	R-Sq	R-Sq (adj)	PRESS	R-Sq (pred)	Mallows Cp	S	AICc	BIC	Cond No
1	56.0	55.9	4.4	55.4	209.6	0.11895	-433.339	-422.227	1.000
1	55.2	55.1	4.5	54.7	218.8	0.12001	-427.889	-416.777	1.000
2	62.4	62.2	3.8	61.6	137.0	0.11016	-479.620	-464.832	1.361
2	62.0	61.7	3.8	61.2	142.2	0.11080	-476.020	-461.232	1.388
3	67.4	67.1	3.3	66.5	80.5	0.10273	-521.570	-503.118	2.757
3	67.1	66.8	3.3	66.2	83.8	0.10316	-518.978	-500.526	2.843
4	71.9	71.5	2.9	70.9	30.3	0.095571	-564.994	-542.892	2.971
4	71.6	71.2	2.9	70.7	33.5	0.096046	-561.936	-539.835	2.910
5	73.8	73.4	2.7	72.8	9.5	0.092345	-585.070	-559.332	14.383
5	72.2	71.8	2.8	71.1	28.6	0.095186	-566.406	-540.668	124.562
6	74.0	73.5	2.7	72.7	10.0	0.092269	-584.493	-555.133	188.476
6	73.9	73.4	2.7	72.7	10.7	0.092375	-583.783	-554.424	382.671
7	74.4	73.8	2.7	73.0	7.0	0.091668	-587.420	-554.453	513.873
7	74.0	73.4	2.7	72.6	11.7	0.092378	-582.666	-549.699	308.843
8	74.4	73.7	2.7	72.8	9.0	0.091818	-585.300	-548.740	658.753

Vars	MEI	CO2	CH4	N2O	CFC11	CFC12	TSI	Aerosols
1		X						
1				X				
2	X	X						
2	X			X				
3	X			X				X
3	X	X						X
4	X	X					X	X
4	X			X			X	X
5	X				X	X	X	X
5	X	X		X			X	X
6	X	X			X	X	X	X
6	X			X	X	X	X	X
7	X	X		X	X	X	X	X
7	X	X	X		X	X	X	X
8	X	X	X	X	X	X	X	X

The best model according to Mallow's criterion are as follows,

1. Identify subsets of predictors for which the C_p value is near $k+1$
2. The full model always yields $C_p = k+1$, so don't select the full model based on C_p.
3. If all models, except the full model, yield a large C_p not near $k+1$, it suggests some important predictor(s) are missing from the analysis.
4. If a number of models have C_p near $k+1$, choose the model with the smallest C_p value, thereby ensuring that the combination of the bias and the variance is at a minimum.
5. When more than one model has a small value of C_p value near k+1, in general, choose the simpler model

As stated earlier , CFC-11 and CFC-12 were removed as it was highly correlated while applying the best subset. According to the best subset, the best reduced model would be one with Mallow's CP= 7, with MEI, CO2,N2O, CFC-11, CFC-12, TSI, Aerosols excluding CH4. The model that has a high adjusted R-Squared, a small standard error and a mallows Cp close to the number of variables. The highlighted model appears to be the best model because its cp value(7) is near to k+1(7+1=8) value and it has high R-square. The reduced model didn't yield the necessary reduction in VIF values of the coefficient of the remaining predictors.

7. CONCLUSION

In this project, we examined the dataset related to climate change. We inspected the dataset for response and predictor variables and then performed base statistical analysis and regression analysis on the same. We saw statistics related to the predictor variables and base plots explaining their role in the model and scatter plots connecting them to the response variable position. We performed the base regression analysis where we detailed the relevance of each and every table that was the output on regression and explained all the important columns related to the same. In this main model, we then surveyed the foundation assumptions and checked our model to check if they held true, after performing the analysis we inferred that the model is valid for all the assumptions. Next, we executed the hypothesis test that we had theorized, the p-value of each predictor was analysed and then we determined the significance of the model and of each predictor on the model. We saw that of the 8 predictor variables, 7 of these were found to have linear association with the predictor variable, while 1 did not have linear association as the predictor had a very high p-value. After that, we spent some time on influential observations as well as unusual observations and explained them with the help of a few observations. We completed the analysis of the main model by discussing residual plots and graphs related to the response variable and the predictor variables. After finishing analysis of the main model, the next step was to use the best subsets and reduced models to check the relevance of each and every predictor variable and use those predictor variables that are of most use to the model as a whole. For this, we used stepwise to get the reduced model which removed deep as a predictor variable.

After that, we examined this reduced model for multicollinearity and saw CO2 and N2O as another high multicollinear variable set. We also checked this model for correlation and also if the assumptions that we had made earlier still held true. Then we used best subsets to first analyse all the possible models that were possible given the response variable. Here we found that the best model that we could see in the best subsets was the one with Mallow's CP closest to the predictor variables while having a low score. Later, we also checked for lack of fit for the model. After performing the test we conclude that the variables are independent from each other and do not contain the identical values in different predictors.

Hence, we can conclude that the relationship between the response variable and all the predictors is very important in terms of judging the factors impacting climate change and gaining insights to take necessary measures to protect our environment. With this, we can say that we have discovered abundant knowledge in this field to alleviate our understanding of climate change and the statistics and science behind it.

8. REFERENCES

1. **Introduction to Linear Regression Analysis (5th edition - Douglas Montgomery, Elizabeth Peck, G. Geoffrey Vining)**
2. **https://support.minitab.com/en-us/minitab-express/1/**
3. **www.kaggle.com**

9. DATASET

- **https://www.kaggle.com/econdata/climate-change**

10. SOFTWARE

Minitab, Excel