STAT 5000

Statistical Methods and Applications I

Spring 2023

Project Report

Group Information	
Student Name	Student ID
Srija Vakiti	110614509
Shanya Chaubey	106720746

Project Title	Analysis Of Cri Score And Its Relevance To Assessing Nations' Vulnerabilities
Date Submitted	05/08/2023

1. OVERALL CONTEXT FOR PROJECT

The Climate Risk Index (CRI) is a measure of the impact of climate change on countries and regions around the world. It was first formulated by the German environmental organization Germanwatch e.V. in 2004, and has since been updated and revised several times.

The CRI is based on an analysis of climate-related disasters that have occurred over the past two decades, including floods, storms, heatwaves, and droughts. The index uses a combination of data on the frequency and intensity of these events, as well as their socio-economic impact, to calculate a score that reflects a country's overall level of vulnerability to climate change.

The CRI score is calculated based on a number of factors, including the number of deaths and injuries caused by climate-related disasters, the economic losses suffered as a result of these events, and the extent to which they disrupt social and economic activity. The score is then normalized based on the country's population and gross domestic product (GDP), in order to provide a standardized measure of vulnerability that can be compared across different countries and regions.

The CRI has been widely used as a tool for assessing the impact of climate change on countries and regions around the world, and has been used to inform policy and decision-making in a variety of contexts. It is seen as an important indicator of the need for action to address climate change, and has been used to draw attention to the disproportionate impact of climate change on vulnerable communities and regions.

In this project the team will focus on how the CRI score was calculated. The team will also dig into the extent of influence each contributor has on the CRI score. The team will also explore whether adding more features in the CRI calculation can make the CRI score more reliable as a metric. This will allow introduction of socio economic features in the analysis. EPI score, population density and GDP per capita will be introduced in the model.

The Environmental Performance Index (EPI) is designed to provide a comprehensive assessment of a country's environmental performance, taking into account not only its current environmental conditions but also its efforts to address environmental challenges and promote sustainable development. The score is calculated using a set of 32 indicators that are organized into 11 categories, each of which reflects a different aspect of environmental sustainability. Countries are ranked based on their overall EPI score, with higher scores indicating better environmental performance and greater progress towards sustainability. The EPI score is seen as an important tool for policymakers, businesses, and civil society organizations to assess a country's environmental performance and identify areas for improvement. It can also be used to track progress over time and compare environmental performance across countries and regions. By providing a standardized measure of environmental sustainability, the EPI score helps to promote greater transparency and accountability in environmental governance, and to encourage countries to take concrete steps to address environmental challenges and promote sustainable development.

2. PROBLEM DEFINITION

To analyze the relationship between the different features and the CRI score, explore whether including more features in the CRI calculation changes the ranking of the countries, and investigate other potential factors that could impact the level of climate risk in different countries.

The CRI measures the impact of extreme weather events on the affected countries. One feature of the CRI calculation is relative - Purchasing Power Parity (PPP). This raises the question of whether including more features in the CRI calculation can change the ranking of the countries.

Another question to explore is whether there is a better way to rank and categorize the countries in terms of their climate risk. The current ranking is based on the CRI score, but it may be worth investigating other factors that could impact the level of climate risk. For example, population density, access to clean water, or forest cover may all be relevant factors to consider.

3. PROJECT MOTIVATION – WHY SHOULD WE CARE?

Knowing the level of preparedness for disasters is crucial for countries to better serve their populations because disasters can strike at any time, and their impact can be devastating. By understanding their level of preparedness, countries can take steps to mitigate the impact of disasters, save lives, and minimize the economic and social costs of these events. One of the key benefits of knowing their level of preparedness is that it allows countries to identify areas where they are particularly vulnerable and to focus their resources on strengthening those areas. This might involve investing in infrastructure, such as flood barriers or early warning systems, or developing emergency response plans and procedures that can be quickly activated in the event of a disaster. By taking a proactive approach to disaster preparedness, countries can reduce the risk of loss of life and property, and ensure that their populations are able to recover more quickly from the impacts of disasters.

In addition, understanding their level of preparedness can help countries to better allocate resources to areas of greatest need. This might involve targeting resources to vulnerable communities or regions that are particularly at risk, or investing in education and awareness-raising campaigns to help people understand how to prepare for and respond to disasters. By taking a holistic approach to disaster preparedness, countries can ensure that they are able to respond effectively to disasters, protect their populations, and minimize the social, economic, and environmental impacts of these events.

4. PROJECT METHODOLOGY

4a) Data Collection

In this phase the team collects data from various sources. All the data is joined to build the meta data which contains all the independent features and CRI score.

4b) Data Cleaning

In this phase the data is cleaned, any outliers are removed, missing values imputed or removed, the data is transformed into a desirable format. In this case the desired format is record format, where each row corresponds to the information about one country. The data in the project is from 2019.

4c) Exploratory Data Analysis

In this stage the relationships between variables and CRI score will be explored. The distribution of the variables will be looked at. Many models assume normal distribution, the team will check for that in this part of the project. It is important to note that CRI score was calculated by Germanwatch using fatalities per 100k people, loss in gdp total and loss in USD purchase power parity. From EDA the team should expect to see strong correlations between the three variables and CRI score. The team will also explore the relationship the newly added variables have with CRI score. This will prepare the grounds for understanding why the models perform the way they do and why some variables can be seen impacting the CRI score more than others.

4d) Model building

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The method assumes that there is a linear relationship between the dependent variable and the independent variable(s), meaning that changes in the independent variable(s) are associated with proportional changes in the dependent variable. The goal of linear regression is to find the best-fitting straight line that summarizes the relationship between the variables. This line can be used to make predictions about the dependent variable based on values of the independent variable(s). The equation for simple linear regression is

$$\mathbf{y = mx + b,}$$

where y is the dependent variable,

x is the independent variable,

m is the slope of the line,

and b is the y-intercept.

The slope of the line represents the change in y for a one-unit increase in x, while the y-intercept represents the predicted value of y when x is equal to zero.

The equation for multiple linear regression is similar, but includes multiple independent variables:

$y = b0 + b1x1 + b2x2 + ... + bnxn$,

where b0 is the intercept,

b1 through bn are the regression coefficients for each independent variable,

and x1 through xn are the values of each independent variable.

The process of linear regression involves estimating the values of the coefficients in the regression equation by minimizing the sum of the squared errors between the predicted values of y and the actual values of y. This is typically done using a method called ordinary least squares (OLS) regression, which involves finding the values of the coefficients that minimize the sum of the squared differences between the predicted values of y and the actual values of y. Once the coefficients have been estimated, they can be used to make predictions about the dependent variable based on values of the independent variable(s). Multiple linear regression allows for the modeling of more complex relationships between the dependent variable and multiple independent variables. In this case, the goal is to estimate the values of the regression coefficients for each independent variable while holding all other variables constant. This can be done using the same OLS regression approach, but with more than one independent variable in the equation. The coefficients can then be used to make predictions about the dependent variable based on values of all the independent variables.

For this project, multiple linear regression will be used. Multiple linear regression models will be built. The first model will only have fatalities per 100k, loss in GDP and loss in USD PPP as the independent variables and CRI score as the dependent variable. The statistical summary of the model will be analyzed and explained by the team. Another multiple linear regression model will be built using all the relevant independent variables. Lastly a multiple linear regression model with only the newly added variables, GDP per capita, population density and EPI score along with the dependent variable CRI score will be created and assessed.

4e) Model Evaluation and Comparison

Once the team has the three linear models, the statistics for each of the models will be compared and discussed.

4f) Comparison to clustering

The team will also use clustering to find countries closely related on the basis of the independent variables. Clustering is a machine learning technique used to group data points that are similar to each other into clusters or groups. The goal of clustering is to identify patterns and structure in unlabeled data, which can then be used to understand the underlying relationships among the data points. Clustering is an unsupervised learning method, meaning that there is no pre-defined set of labels or categories for the data points. Instead, the algorithm must identify the patterns and groupings based solely on the data itself. For clustering the team will specifically use KMeans clustering.

The CRI score in the data will be discretized based on quantiles and the results of clustering would be added to the original dataframe. The team will then compare the discretized label to the results of clustering.

5. DATA SOURCE

The data for this project comes from various locations. The CRI score data for 2019 is obtained from Kaggle. The EPI data is sourced from

Climate Risk and Economic Losses

Data Card | Code (0) | Discussion (0) | 35 | New Notebook

Detail | Compact | Column | 10 of 17 columns

- fatalities_per_100k_rank: Rank of the country in terms of fatalities per 100,000 people (Integer)
- fatalities_per_100k_total: Total number fatalities per 100,000 people (Integer)
- fatalities_rank: Rank of the country in terms of total fatalities (Integer)
- fatalities_total: Total number of fatalities (Integer)
- losses_per_gdp__rank: Rank of the country in terms of losses per GDP (Integer)
- losses_per_gdp__total: Total losses per GDP (Integer)
- losses_usdm_ppp_rank: Rank of the country in terms of losses in USDM PPP (Integer)

# index	cartodb_id	the_geom	the_geom_webm...	country
Index	Cartodb Id	Geometry of the country (Geometry)	Web Mercator projection of the geometry of the country (Geometry)	Country
 0 181	 1 182	[null] 100%	[null] 100%	182 unique values
0	1			Saudi Arabia

	index	country	cri_rank	cri_score	fatalities_per_100k_total	fatalities_total	losses_per_gdp__total
1	0	Saudi Arabia	79	72.50	0.45	140	0.0001
2	1	Romania	61	61.50	0.01	1	0.6746
3	2	Spain	69	66.33	0.05	22	0.0394
4	3	Slovenia	135	124.50	0.00	0	NA
5	5	Sierra Leone	102	88.50	0.16	10	0.0011
6	6	South Africa	33	45.67	0.03	19	0.4722
7	7	Serbia	83	75.50	0.00	0	0.2794
8	8	Slovak Republic	123	105.33	0.02	1	0.0046
9	9	Solomon Islands	89	76.83	0.17	1	0.0445
10	10	Swaziland	128	109.33	0.00	0	0.0174
11	11	Turkey	126	106.00	0.02	13	0.0002
12	12	Tanzania	72	68.67	0.17	79	0.0123
13	13	Tunisia	135	124.50	0.00	0	NA
14	14	Thailand	53	57.00	0.02	12	0.2557
15	15	The Bahamas	7	22.83	9.07	33	0.9035
16	16	Uganda	130	110.50	0.01	4	0.0012
17	17	Vietnam	29	43.50	0.10	91	0.1486

https://data.worldbank.org/indicator/EN.POP.DNST?end=2019&most_recent_year_desc=false&start=1961&view=

All Countries and Economies

Country	Most Recent Year	Most Recent Value
Bangladesh	2019	1,272
Belgium	2019	379
Burkina Faso	2019	77
Bulgaria	2019	64
Venezuela, RB	2019	33
Bosnia and Herzegovina	2019	66
Barbados	2019	652
Bermuda	2019	1,184
Brunei Darussalam	2019	83

Country Code	2019 [YR2019]
<chr>	<dbl>
AFG	57.90825
ALB	104.16755
DZA	17.93032
ASM	236.60500
AND	162.43191
AGO	25.95138

https://data.worldbank.org/indicator/NY.GDP.PCAP.CD?end=2019&start=1960&view=chart

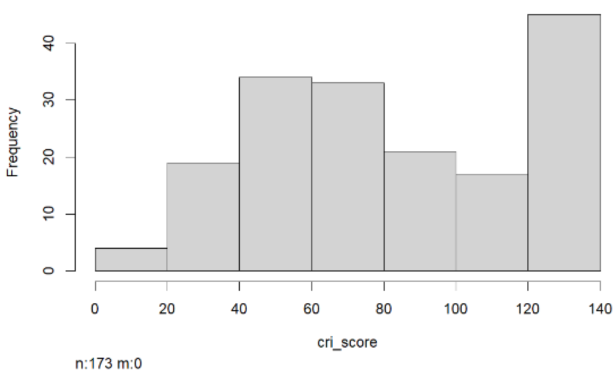
All Countries and Economies

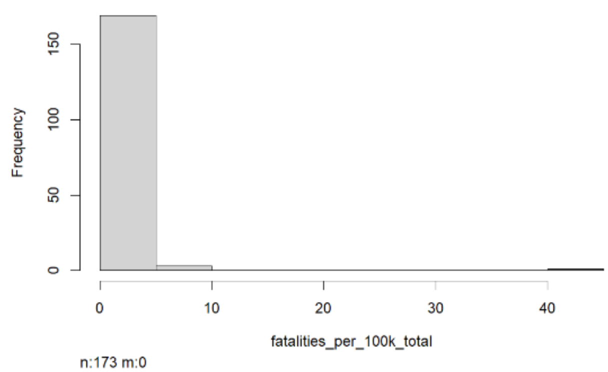
Country	Most Recent Year	Most Recent Value
Afghanistan	2019	500.5
Albania	2019	5,396.2
Algeria	2019	4,022.2
American Samoa	2019	13,672.6
Andorra	2019	41,327.5
Angola	2019	2,142.2
Antigua and Barbuda	2019	18,319.5
Argentina	2019	9,963.7
Armenia	2019	4,828.5

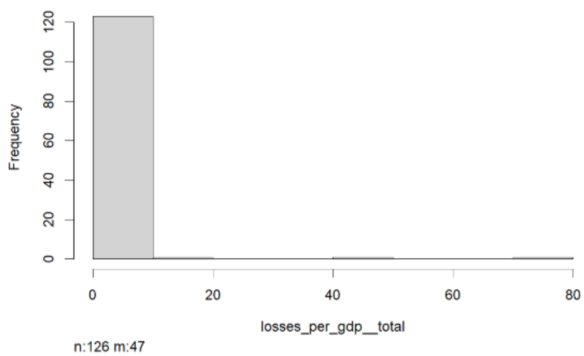
country_code	GDP_per_capita
<chr>	<dbl>
AFG	500.5227
ALB	5396.2159
DZA	4022.1502
ASM	13672.5767
AND	41327.5020
AGO	2142.2388

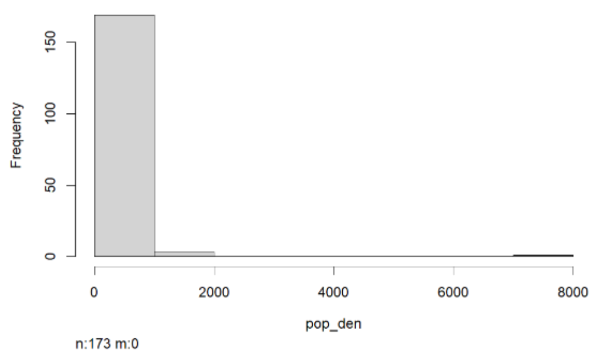
6. DATA ANALYSIS

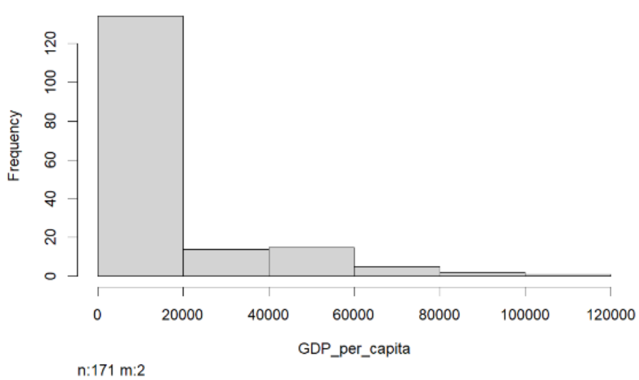
DIstribution of variables in 2019:

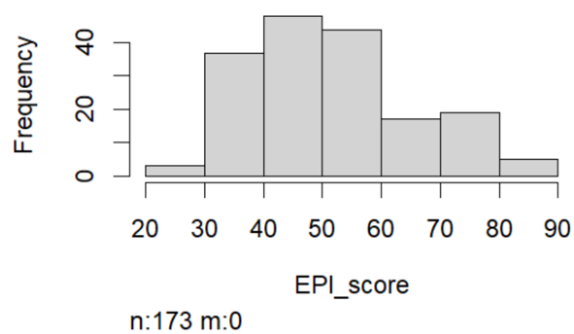












cri_rank has uniform distribution with minimum deviations (0-140 range)

cri_score has normal distribution (0-120 range) fatalities_perk_100k_total, fatalities_total (range:0-5000) have a similar right skewed distribution

fatalities_perk_100k_total has a majority range between 0 and 10 range(0-50)

losses_per_gdp_total (range:0-80) and losses_per_pp_total (range:0-40000) have similar right skewed distribution

pop_den is heavily right skewed (range: 0-1400)

gdp_per_capita is heavily right skewed (range: 0-80000)

epi_score is slighty right skewed (range: 20-90)

Some insights from these can be:

1. cri_rank: The majority of the countries have a CRI rank below 100, with a peak at around 50. There are a few countries with high ranks above 150, indicating they are more vulnerable to climate risk.

2. cri_score: The distribution of CRI scores is heavily skewed to the right, with the majority of the countries having a score below 10. A few countries have a high score above 50, indicating they are more vulnerable to climate risk.

3. fatalities_per_100k_total: The majority of the countries have a low fatalities rate per 100,000 inhabitants, with a peak at around 0-1. There are a few countries with a high fatalities rate above 10, indicating they are more vulnerable to climate risk.

4. fatalities_total: The distribution of total fatalities is heavily skewed to the right, with the majority of the countries having less than 1000 fatalities due to climate risk. A few countries have a high number of fatalities above 10,000, indicating they are more vulnerable to climate risk.

5. losses_per_gdp_total: The majority of the countries have a low losses per GDP unit, with a peak at around 0-0.01. There are a few countries with high losses per GDP unit above 0.05, indicating they are more vulnerable to climate risk.

6. losses_usdm_ppp_total: The distribution of losses in PPP is heavily skewed to the right, with the majority of the countries having losses below 100 million USD. A few countries have a high loss above 10 billion USD, indicating they are more vulnerable to climate risk.

7. pop_den: The distribution of population density is heavily skewed to the right, with the majority of the countries having a population density below 500 inhabitants per square kilometer. A few countries have a high population density above 5000 inhabitants per square kilometer, indicating they are more vulnerable to climate risk.

8. GDP_per_capita: The majority of the countries have a low GDP per capita, with a peak at around 0-5000 USD. There are a few countries with a high GDP per capita above 50,000 USD, indicating they are less vulnerable to climate risk.

9. EPI_score: The distribution of EPI scores is roughly normal, with a peak at around 60-70. There are a few countries with a low EPI score below 30, indicating they are more vulnerable to climate risk.

Central Tendencies of all variables and labels:

```
The mean of cri_rank is 59.18519, median of cri_rank is 56.5, and mode of cri_rank is
33.

The mean of cri_score is 60.63861, median of cri_score is 58.835, and mode of cri_score
is 45.67.

The mean of fatalities_per_100k_total is -2.161106, median of fatalities_per_100k_total
is -2.302585, and mode of fatalities_per_100k_total is -3.912023.

The mean of fatalities_total is 2.887093, median of fatalities_total is 2.639057, and
mode of fatalities_total is 0.

The mean of losses_per_gdp__total is -2.834999, median of losses_per_gdp__total is
-2.454578, and mode of losses_per_gdp__total is -9.21034.

The mean of losses_usdm_ppp_total is 4.408952, median of losses_usdm_ppp_total is
4.779618, and mode of losses_usdm_ppp_total is 0.2062008.

The mean of pop_den is 4.284827, median of pop_den is 4.403574, and mode of pop_den is
2.813388.

The mean of GDP_per_capita is 8.515915, median of GDP_per_capita is 8.374993, and mode
of GDP_per_capita is 10.01817.

The mean of EPI_score is 50.28218, median of EPI_score is 47.485, and mode of EPI_score
is 50.735.
```

Dispersion of all variables and labels:

Summary of Numeric Columns in data

Column	Range	IQR	No. of Outliers
cri_rank	130.000000	60.500000	0
cri_score	102.160000	33.457500	0
fatalities_per_100k_total	8.381602	2.069595	3
fatalities_total	8.370316	2.944184	0
losses_per_gdp__total	13.558932	2.929727	4
losses_usdm_ppp_total	14.287443	3.630311	1
pop_den	6.417676	1.445549	2
GDP_per_capita	5.960250	1.986137	0
EPI_score	57.245000	19.070000	0

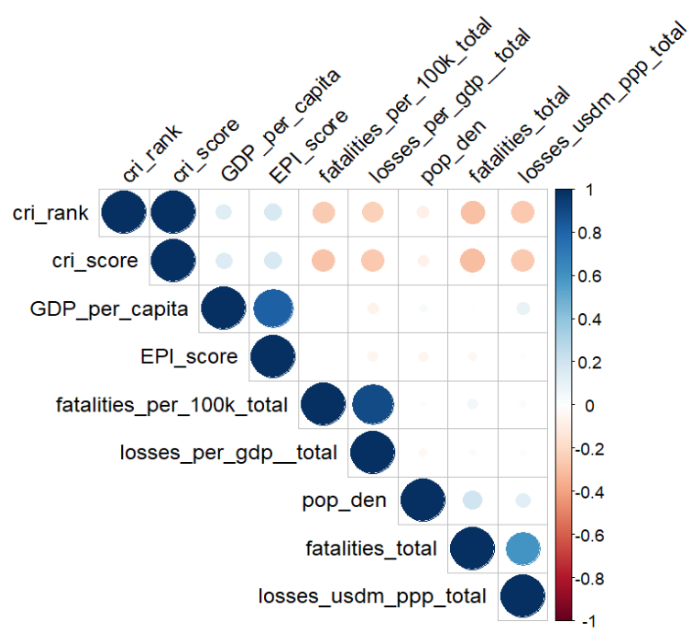
Standard Deviation and Variance of all variables and labels:

```
[1] "Standard deviation and variance of cri_rank is 44.7289126678112 and 2000.67562844468"
[1] "Standard deviation and variance of cri_score is 34.5443173741278 and 1193.30986284447"
[1] "Standard deviation and variance of fatalities_per_100k_total is 3.47959518800711 and 12.1075826724022"
[1] "Standard deviation and variance of fatalities_total is 437.95950562314З and 191808.528565667"
[1] "Standard deviation and variance of losses_per_gdp__total is 7.82534251399554 and 61.235985461346"
[1] "Standard deviation and variance of losses_usdm_ppp_total is 4595.60936308836 and 21119625.4181054"
[1] "Standard deviation and variance of pop_den is 639.075757692378 and 408417.824070087"
[1] "Standard deviation and variance of GDP_per_capita is 19844.0472442141 and 393786211.030602"
[1] "Standard deviation and variance of EPI_score is 13.960631930072 and 194.899243886947"
```

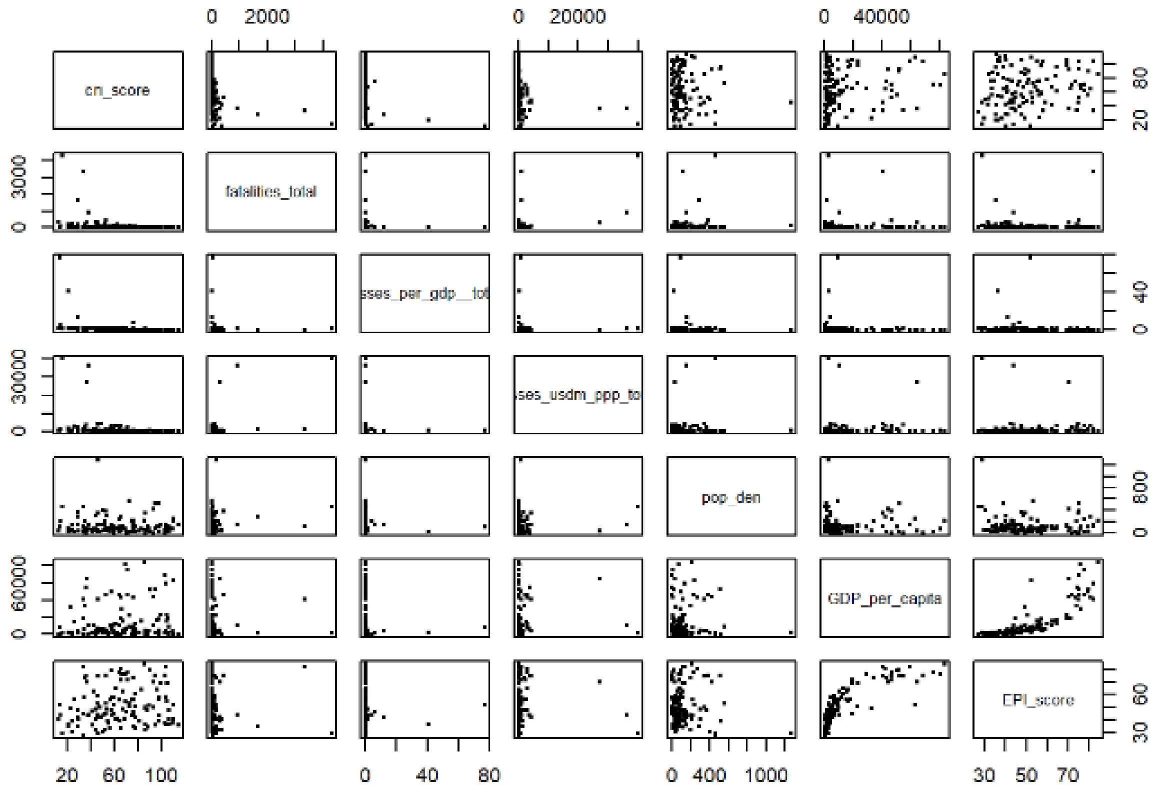
Skewness and kurtosis of all variables and labels:

```
[1] "The skewness and kurtosis of cri_rank is -0.318266474743969 and 1.70050251610917"
[1] "The skewness and kurtosis of cri_score is -0.127700999354673 and 1.76265902655592"
[1] "The skewness and kurtosis of fatalities_per_100k_total is 11.2756818055515 and 138.253596634768"
[1] "The skewness and kurtosis of fatalities_total is 7.9191019313013 and 69.3617854871199"
[1] "The skewness and kurtosis of losses_per_gdp__total is 8.41910622595963 and 77.1594484590642"
[1] "The skewness and kurtosis of losses_usdm_ppp_total is 7.32942987513225 and 57.3306295563081"
[1] "The skewness and kurtosis of pop_den is 10.677800251121 and 128.249595914073"
[1] "The skewness and kurtosis of GDP_per_capita is 2.11624861001586 and 7.6676310737868"
[1] "The skewness and kurtosis of EPI_score is 0.505097098993462 and 2.39691869794363"
```

Correlation of all variables and labels:



7. EXPLORATORY DATA ANALYSIS

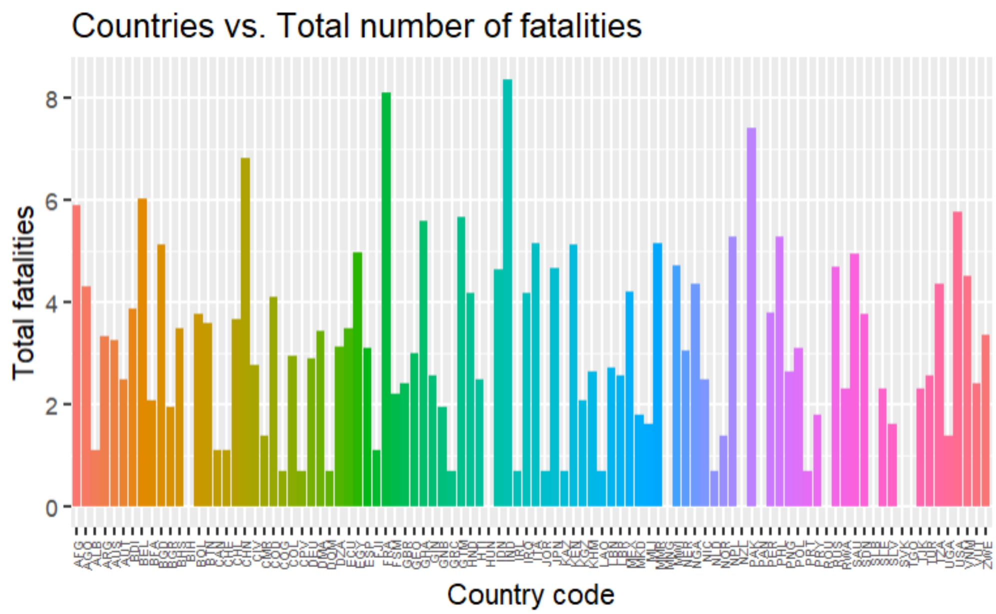


Insights:

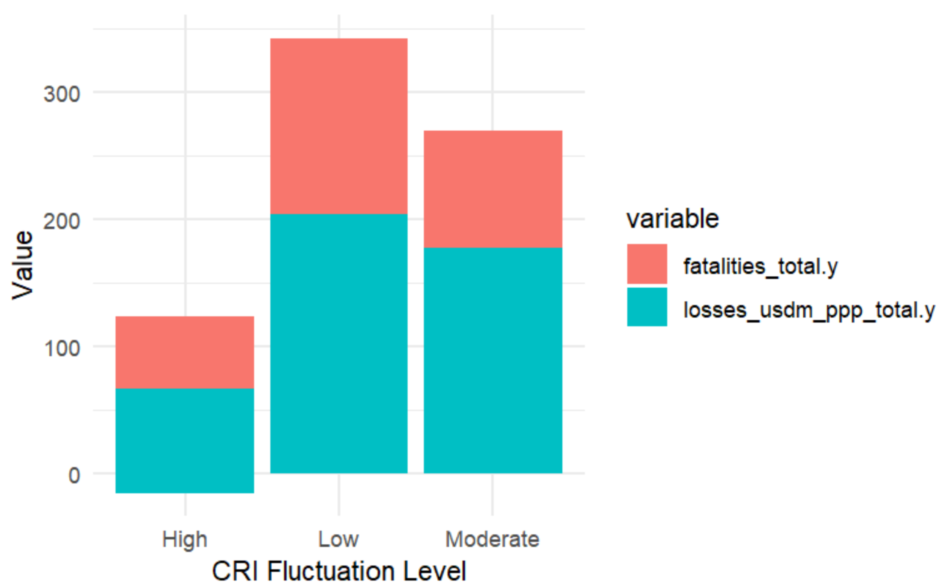
Only cri_rank and cri_score have a definitive linear relationship.

If a majority of the scatter plots are points forming a thin vertical strip parallel to the y-axis on the left, it could indicate that the variable on the x-axis has a limited range of values. It's also possible that there is simply no strong relationship between the two variables, and the vertical strip is just a result of the limited range or nature of the data. In this case, it may be worth exploring other variables or techniques to uncover any meaningful relationships in the data.

GDP_per_cap and EPI_Score have an upward curve relationship, it seems like there is a positive correlation between GDP_per_cap and EPI_Score, indicating that countries with higher GDP per capita tend to have higher environmental performance index scores. Other graphs show a higher density of points accumulated on the first half of the x-axis, which could indicate that the values in these variables are skewed to the left. This means that a majority of the data points fall within a smaller range of values. Some graphs show points scattered across the x-axis but not much on their y-axis, there is probably little to no correlation between these variables. Majority of the graphs still show vertical strips that depict low variation. This could be an indication that these variables may not be useful in predicting the target variable, or that other variables may need to be considered to fully understand the relationship between the variables.



From the plot, we can observe that countries with higher EPI scores generally have lower losses per GDP, i.e., they are more resilient to environmental disasters. The median losses per GDP are lower for countries with higher EPI scores, and the distribution of losses per GDP is wider for countries with lower EPI scores. This suggests that countries with higher EPI scores may have better policies and infrastructure to mitigate the impact of environmental disasters, resulting in lower economic losses.



From the chart, we can see the distribution of the fatalities_total.y and losses_usdm_ppp_total.y variables across the different levels of CRI fluctuation. We can observe that as the CRI fluctuation level increases, the values of both variables also tend to increase. The stacked bar chart also allows us to compare the relative contribution of each variable to the total value. We can conclude that countries with a high level of CRI fluctuation have losses in GDP as they are more vulnerable to disasters.

8. STATISTICAL MODEL DESIGN

Model 1:

```
Call:
lm(formula = cri_score ~ fatalities_per_100k_total + losses_per_gdp__total +
    losses_usdm_ppp_total, data = data_scale)

Residuals:
    Min      1Q  Median      3Q     Max
-15.487  -4.661  -0.838   5.010  32.073

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                50.8428     2.6459  19.216  < 2e-16 ***
fatalities_per_100k_total  -8.6208     0.5291 -16.294  < 2e-16 ***
losses_per_gdp__total      -2.8918     0.4809  -6.013 2.74e-08 ***
losses_usdm_ppp_total      -3.8633     0.3746 -10.314  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.82 on 104 degrees of freedom
Multiple R-squared:  0.9067,    Adjusted R-squared:  0.904
F-statistic: 336.8 on 3 and 104 DF,  p-value: < 2.2e-16

                           Estimate Std. Error    t value     Pr(>|t|)
(Intercept)               50.842762  2.6458819  19.215809 5.324352e-36
fatalities_per_100k_total -8.620837  0.5290846 -16.293872 2.152501e-30
losses_per_gdp__total     -2.891838  0.4809356  -6.012943 2.738533e-08
losses_usdm_ppp_total     -3.863289  0.3745743 -10.313811 1.338472e-17
```

The results show the coefficients for a linear regression model that predicts the cri_score variable using the predictors fatalities_per_100k_total, losses_per_gdp__total, and losses_usdm_ppp_total.

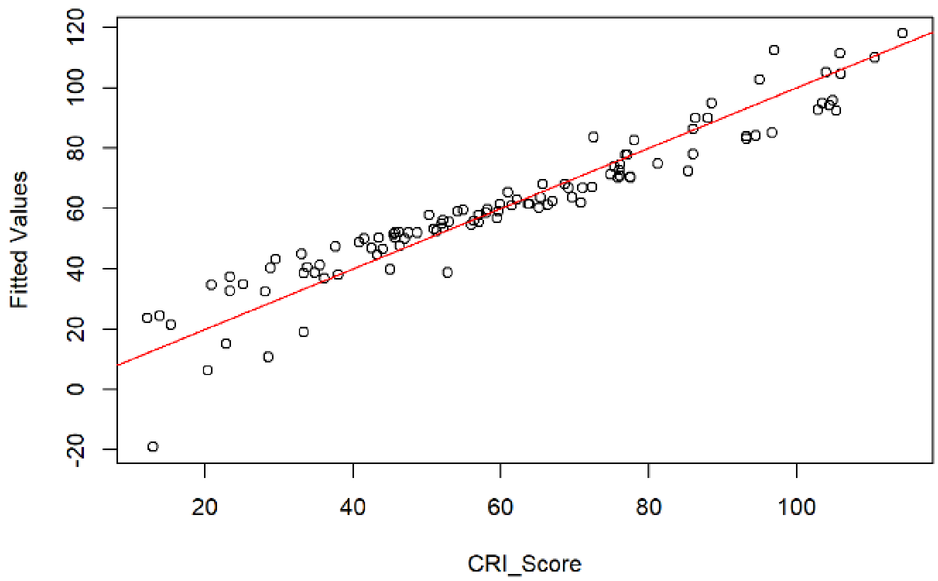
The intercept is 50.8428, indicating that cri_score is expected to be around 50.8 when all predictors are 0. The coefficients of the predictors are all negative, indicating that as each predictor increases, the cri_score decreases. Specifically, for every unit increase in fatalities_per_100k_total, the cri_score is expected to decrease by 8.6208. For every unit increase in losses_per_gdp__total, the cri_score is expected to decrease by 2.8918. And for every unit increase in losses_usdm_ppp_total, the cri_score is expected to decrease by 3.8633.

The p-values for all the predictors are very small (less than 2e-16), indicating that they are all statistically significant predictors of cri_score.

The multiple R-squared is 0.9067, indicating that the model explains a large proportion of the variance in the cri_score variable.

The adjusted R-squared is 0.904, indicating that the predictors in the model account for a substantial proportion of the variance in the cri_score variable, while considering the number of predictors.

Overall, the model suggests that higher values of fatalities_per_100k_total, losses_per_gdp__total, and losses_usdm_ppp_total are associated with lower cri_score values, indicating that countries with higher levels of fatalities, losses, and damages are likely to have a lower climate risk index score.



The points in the plot show an upward sloping line, it may indicate that this model is underestimating the true values for lower predicted values and overestimating them for higher predicted values. This is a common issue known as heteroscedasticity, which means that the variance of the residuals is not constant across the range of the predictor variable.

To address this issue, we can try transforming either the response variable or the predictor variables or both. Another option is to use a different model that can handle heteroscedasticity, such as a weighted least squares regression.

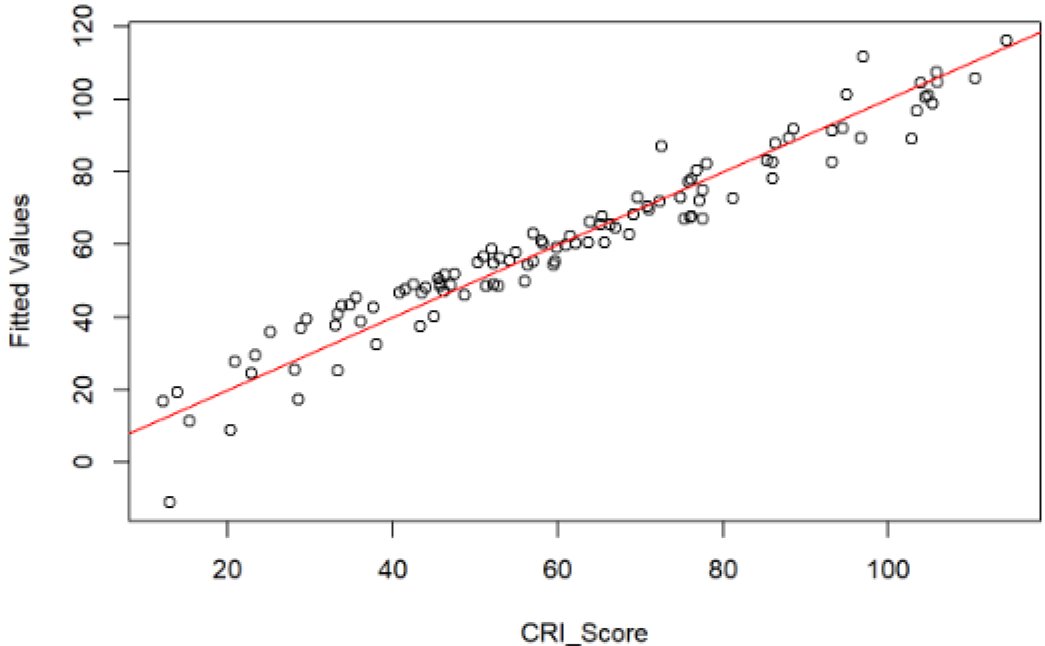
Model 2:

```
Call:
lm(formula = cri_score ~ fatalities_per_100k_total + losses_per_gdp__total + 
    losses_usdm_ppp_total + pop_den + GDP_per_capita + EPI_score, 
    data = data_scale)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.6589  -4.2576  -0.6421   3.8368  24.0259 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)               27.34326    5.31460   5.145 1.32e-06 ***
fatalities_per_100k_total -8.39833    0.43729 -19.205  < 2e-16 ***
losses_per_gdp__total     -1.83139    0.41564  -4.406 2.63e-05 ***
losses_usdm_ppp_total     -5.29565    0.35795 -14.794  < 2e-16 ***
pop_den                    0.54827    0.53089   1.033  0.30419    
GDP_per_capita             3.13122    0.98914   3.166  0.00205 ** 
EPI_score                  0.08527    0.09433   0.904  0.36820    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.316 on 101 degrees of freedom
Multiple R-squared:  0.9409,	Adjusted R-squared:  0.9374 
F-statistic: 267.9 on 6 and 101 DF,  p-value: < 2.2e-16

                             Estimate Std. Error     t value     Pr(>|t|)
(Intercept)               27.34326315 5.31460123   5.1449322 1.319048e-06
fatalities_per_100k_total -8.39833176 0.43728912 -19.2054440 1.715707e-35
losses_per_gdp__total     -1.83138887 0.41563917  -4.4061990 2.625399e-05
losses_usdm_ppp_total     -5.29564803 0.35795169 -14.7943092 4.967422e-27
pop_den                    0.54827290 0.53088743   1.0327479 3.041885e-01
GDP_per_capita             3.13122245 0.98914186   3.1655949 2.046151e-03
EPI_score                  0.08526789 0.09433247   0.9039082 3.681950e-01
```

This multiple linear regression model that includes six predictor variables: fatalities_per_100k_total, losses_per_gdp__total, losses_usdm_ppp_total, pop_den, GDP_per_capita, and EPI_score, in addition to the intercept. The response variable is cri_score.

The coefficient estimates suggest that fatalities_per_100k_total, losses_per_gdp__total, losses_usdm_ppp_total, and GDP_per_capita are significantly associated with cri_score, while pop_den and EPI_score are not so much. Specifically, a one-unit increase in fatalities_per_100k_total is associated with an estimated decrease of 8.398 in cri_score, holding other variables constant. A one-unit increase in losses_per_gdp__total is associated with an estimated decrease of 1.831 in cri_score, holding other variables constant. A one-unit increase in losses_usdm_ppp_total is associated with an estimated decrease of 5.296 in cri_score, holding other variables constant. A one-unit increase in GDP_per_capita is associated with an estimated increase of 3.131 in cri_score, holding other variables constant.

The adjusted R-squared of the model is 0.9374, indicating that the model explains 93.74% of the variance in cri_score. The F-statistic is significant at the 0.01 level, suggesting that at least one of the predictor variables has a significant association with the response variable. The residual standard error is 6.316, indicating that the model's predictions have an average error of about 6.316 units.



The points in this plot also show an upward sloping line, it may indicate that this model is underestimating the true values for lower predicted values and overestimating them for higher predicted values. This is a common issue known as heteroscedasticity, which means that the variance of the residuals is not constant across the range of the predictor variable.

To address this issue, we can try transforming either the response variable or the predictor variables or both. Another option is to use a different model that can handle heteroscedasticity, such as a weighted least squares regression.

Comparison of the two models:

- Model 1 only includes three predictor variables, while Model 2 includes six predictor variables.
- In Model 2, the additional predictors are population density, GDP per capita, and EPI score.
- The coefficient estimates for each variable are also different between the two models.
- The R-squared value for Model 2 is higher, which means it explains more of the variance in the response variable than Model 1.
- Additionally, Model 2 has a lower residual standard error, indicating that it fits the data better than Model 1.

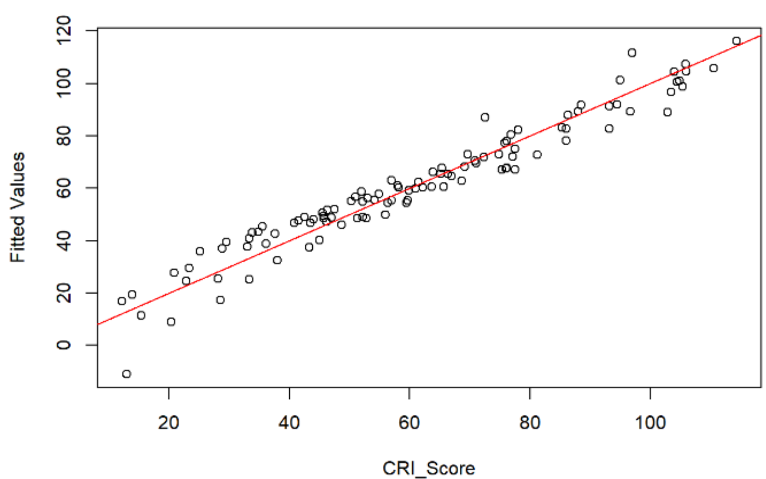
Model 3:

```
Call:
lm(formula = cri_score ~ pop_den + EPI_score + losses_usdm_ppp_total +
    fatalities_per_100k_total + losses_per_gdp__total + GDP_per_capita,
    data = data_scale)

Residuals:
     Min       1Q   Median       3Q      Max
-14.6589  -4.2576  -0.6421   3.8368  24.0259

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               27.34326    5.31460   5.145 1.32e-06 ***
pop_den                    0.54827    0.53089   1.033  0.30419
EPI_score                  0.08527    0.09433   0.904  0.36820
losses_usdm_ppp_total     -5.29565    0.35795 -14.794  < 2e-16 ***
fatalities_per_100k_total -8.39833    0.43729 -19.205  < 2e-16 ***
losses_per_gdp__total     -1.83139    0.41564  -4.406 2.63e-05 ***
GDP_per_capita             3.13122    0.98914   3.166  0.00205 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.316 on 101 degrees of freedom
Multiple R-squared:  0.9409,	Adjusted R-squared:  0.9374
F-statistic: 267.9 on 6 and 101 DF,  p-value: < 2.2e-16

                             Estimate Std. Error     t value     Pr(>|t|)
(Intercept)               27.34326315 5.31460123   5.1449322 1.319048e-06
pop_den                    0.54827290 0.53088743   1.0327479 3.041885e-01
EPI_score                  0.08526789 0.09433247   0.9039082 3.681950e-01
losses_usdm_ppp_total     -5.29564803 0.35795169 -14.7943092 4.967422e-27
fatalities_per_100k_total -8.39833176 0.43728912 -19.2054440 1.715707e-35
losses_per_gdp__total     -1.83138887 0.41563917  -4.4061990 2.625399e-05
GDP_per_capita             3.13122245 0.98914186   3.1655949 2.046151e-03
```

Model 3 is similar to model 2, with cri_score as the response variable and six predictor variables: pop_den, EPI_score, losses_usdm_ppp_total, fatalities_per_100k_total, losses_per_gdp__total, and GDP_per_capita.

The difference is that of order in which the variables were arranged.

The coefficients of pop_den and EPI_score are not statistically significant. The other four predictor variables have statistically significant coefficients. The adjusted R-squared of Model 3 (0.9374) is lower than that of Model 2 (0.9432), indicating that Model 3 may be a slightly worse fit than Model 2.

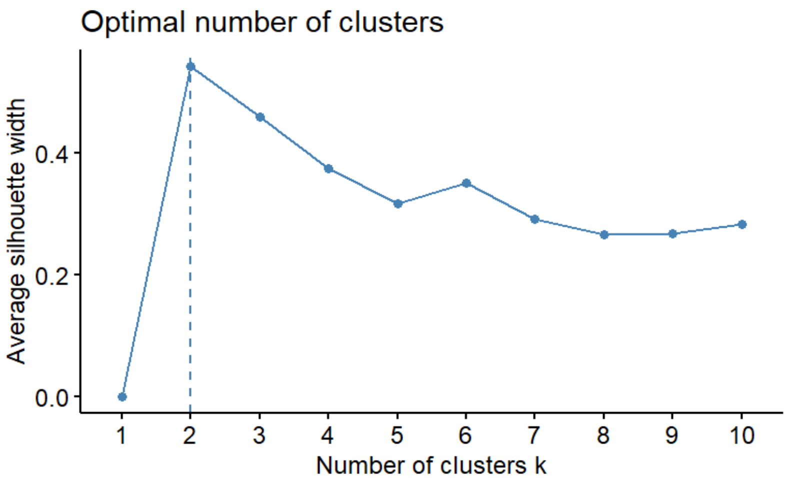


The points in this plot also show an upward sloping line, it may indicate that this model is underestimating the true values for lower predicted values and overestimating them for higher predicted values. This is a common issue known as heteroscedasticity, which means that the variance of the residuals is not constant across the range of the predictor variable.

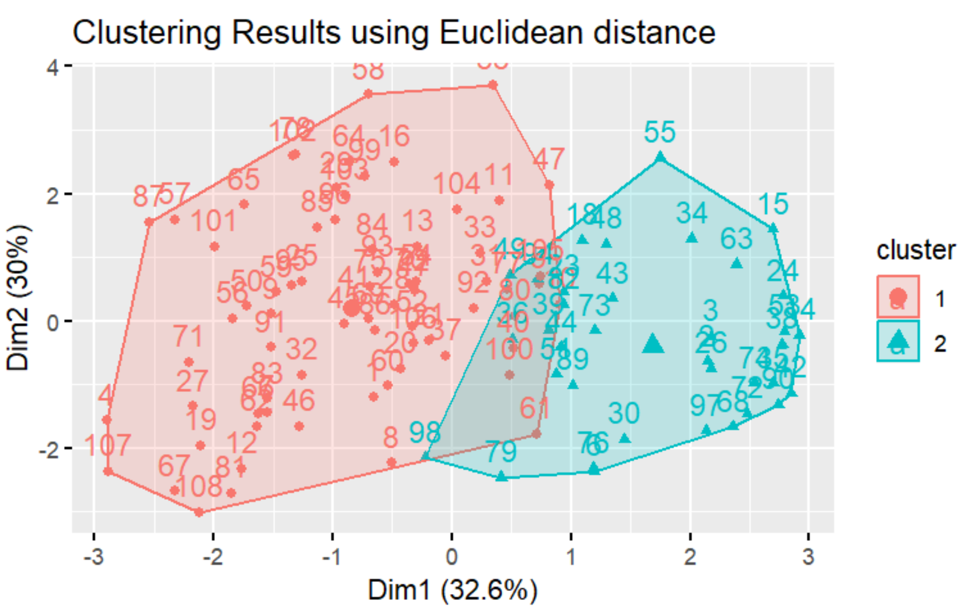
To address this issue, we can try transforming either the response variable or the predictor variables or both. Another option is to use a different model that can handle heteroscedasticity, such as a weighted least squares regression.

Clustering:

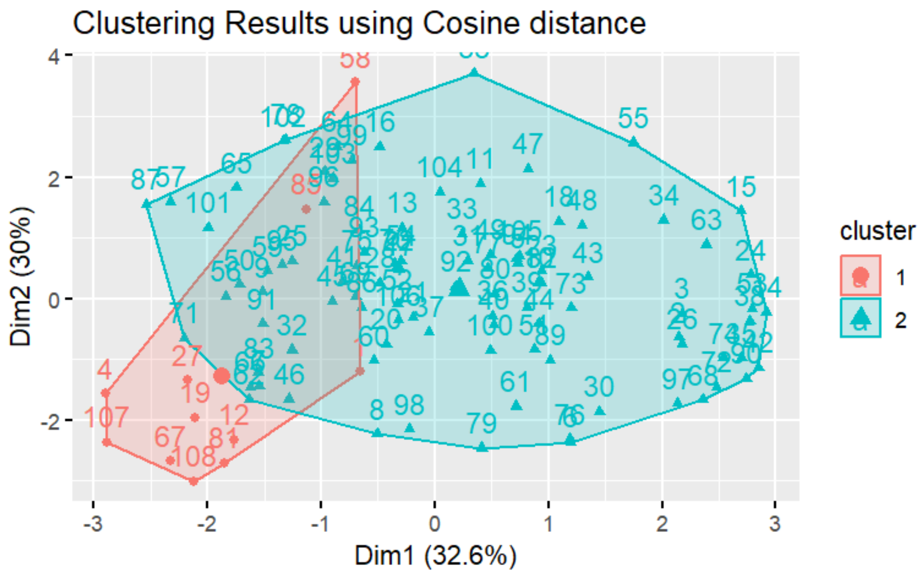
Using the silhouette method the appropriate number of clusters was determined:

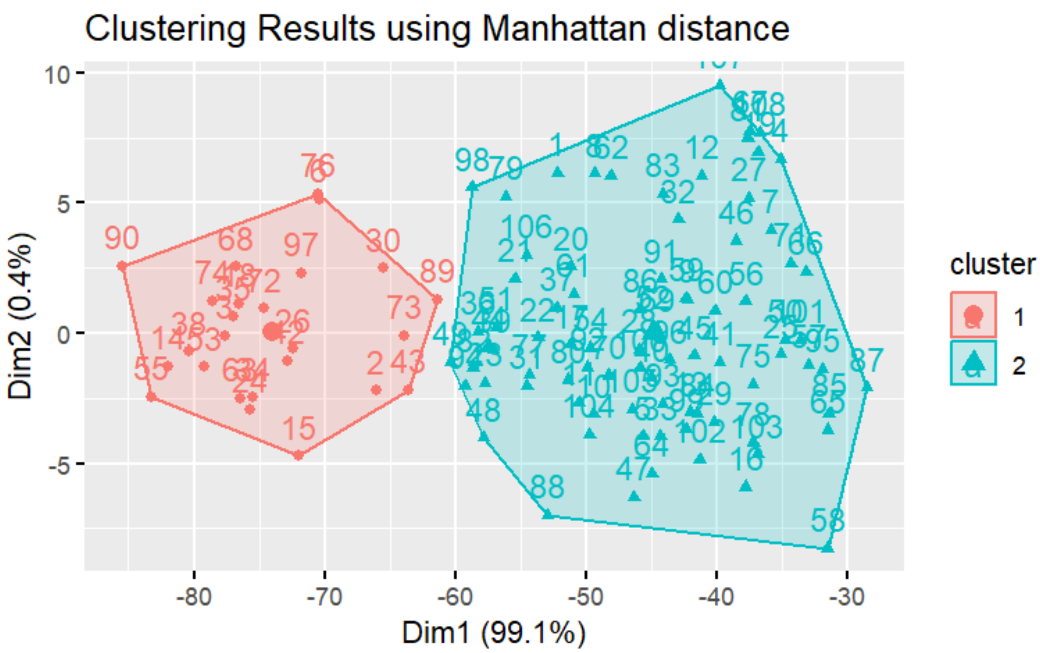


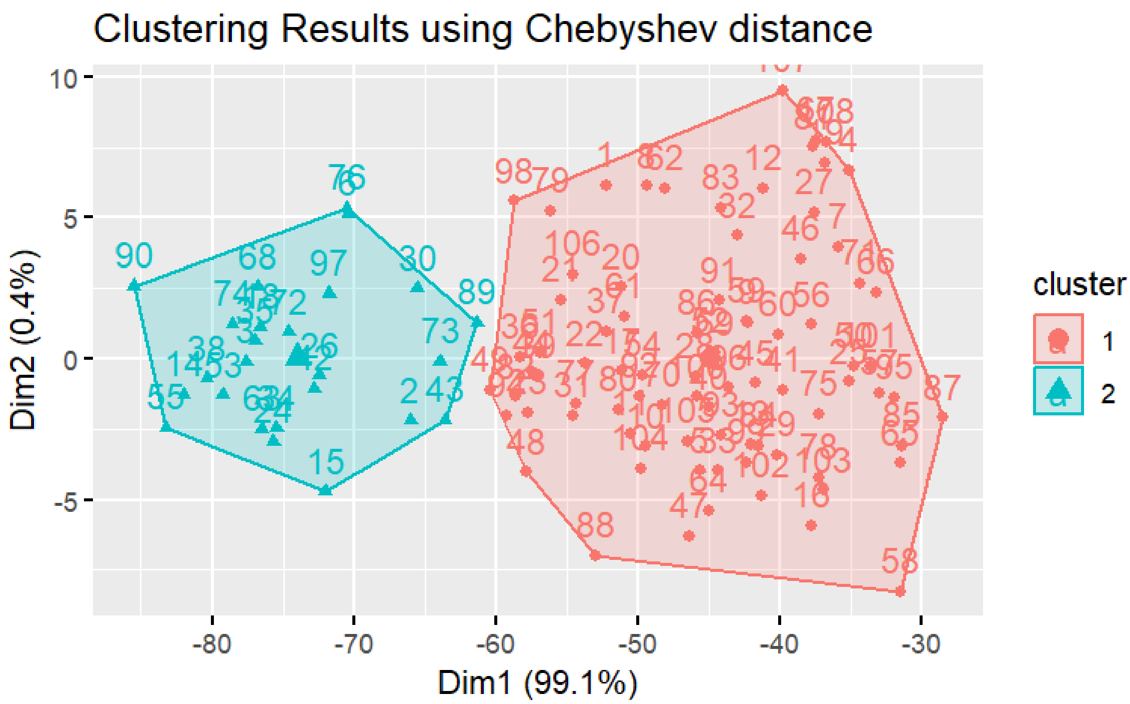
From the above silhouette plot, the best number of clusters is 2. Using k=2 and performing KMeans on the data, the following results are obtained:

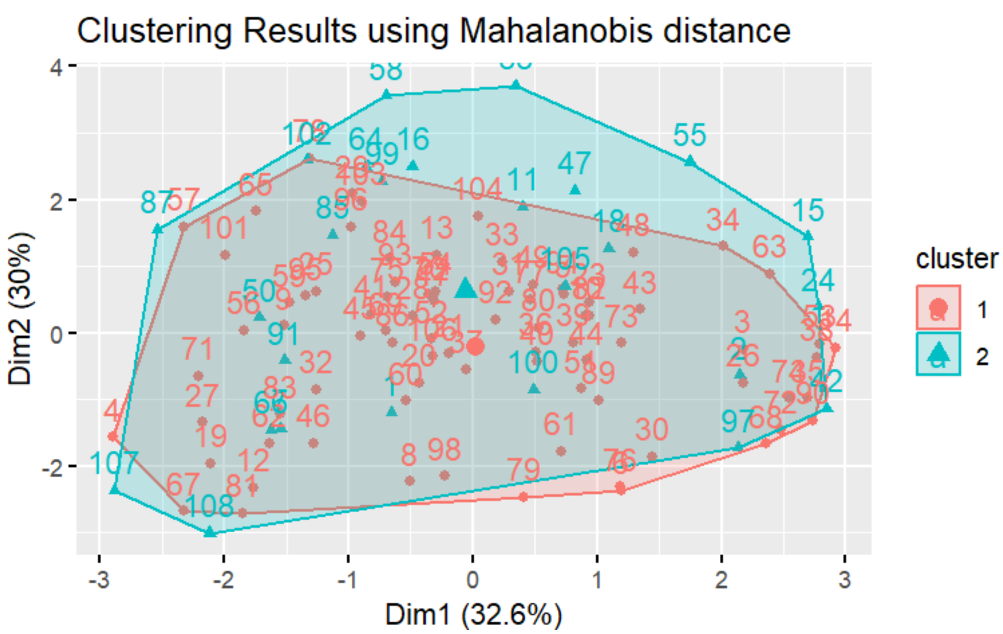


Clustering using Cosine distance:


Clustering Results using Cosine distance
Dim2 (30%)
Dim1 (32.6%)
cluster
1
2


Clustering Results using Manhattan distance
Dim2 (0.4%)
Dim1 (99.1%)
cluster
1
2

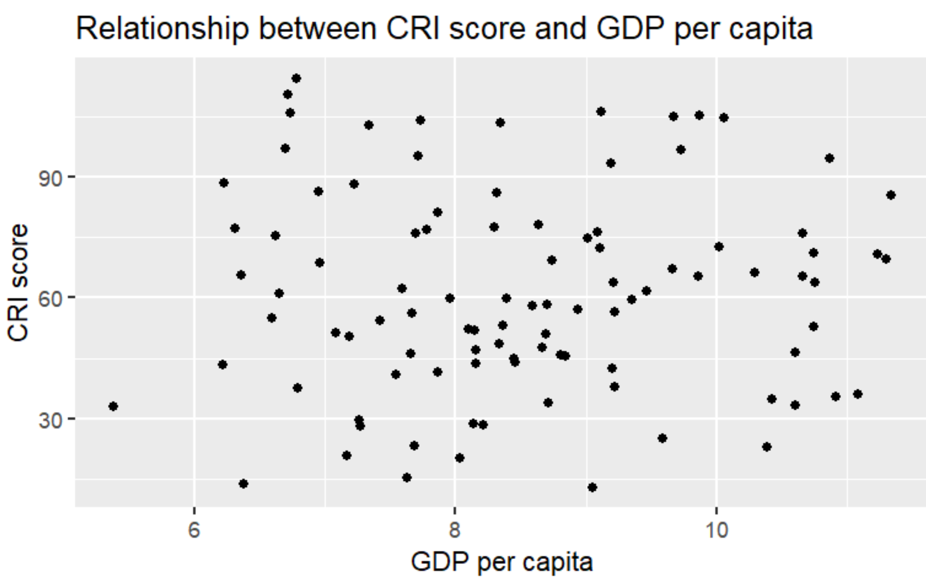




If Manhattan and Chebyshev distances result in two distinct clusters, it suggests that the data points are more separated in these distance metrics as compared to Euclidean distance. Manhattan distance measures the distance between two points by summing the absolute differences of their coordinates, while Chebyshev distance measures the maximum difference between any coordinate of two points.

If the data points are more spread out in these metrics, it could be because they have different ranges or scales in the different dimensions. This could indicate that some features have more influence than others in determining the distance between the points, and might warrant further investigation. Additionally, it might suggest that there are two clear groups or clusters within the data.

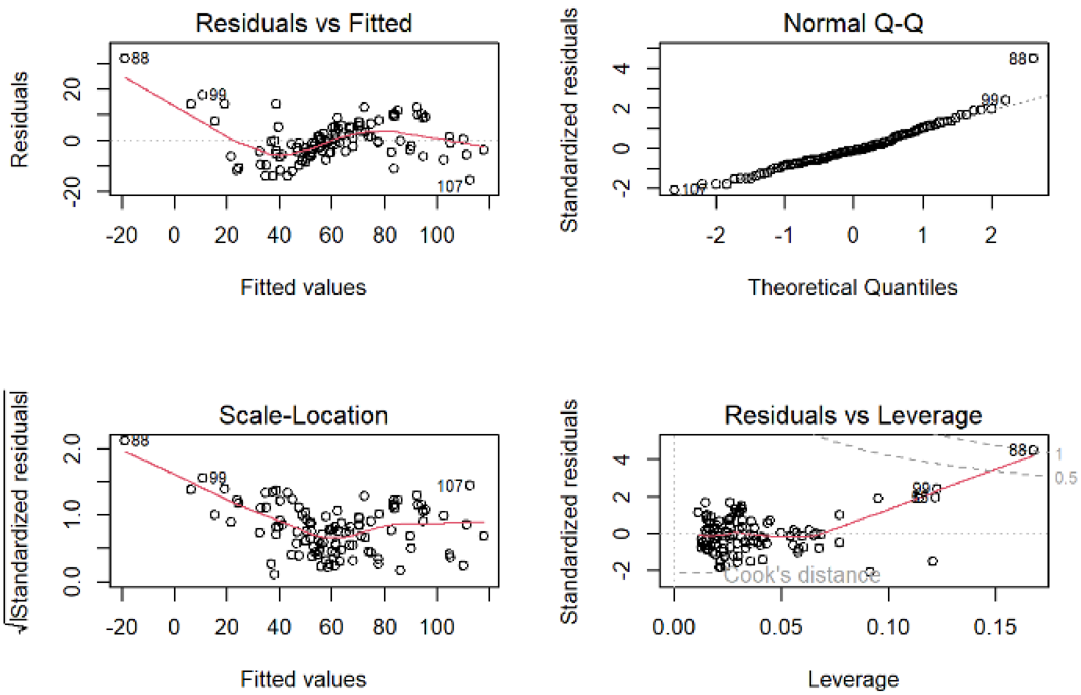
The results suggest that the choice of distance metric can significantly impact the clustering results. The fact that Manhattan and Chebyshev distances produced two distinct clusters, while other types could not, indicates that the structure of the data is better represented using the former two metrics.



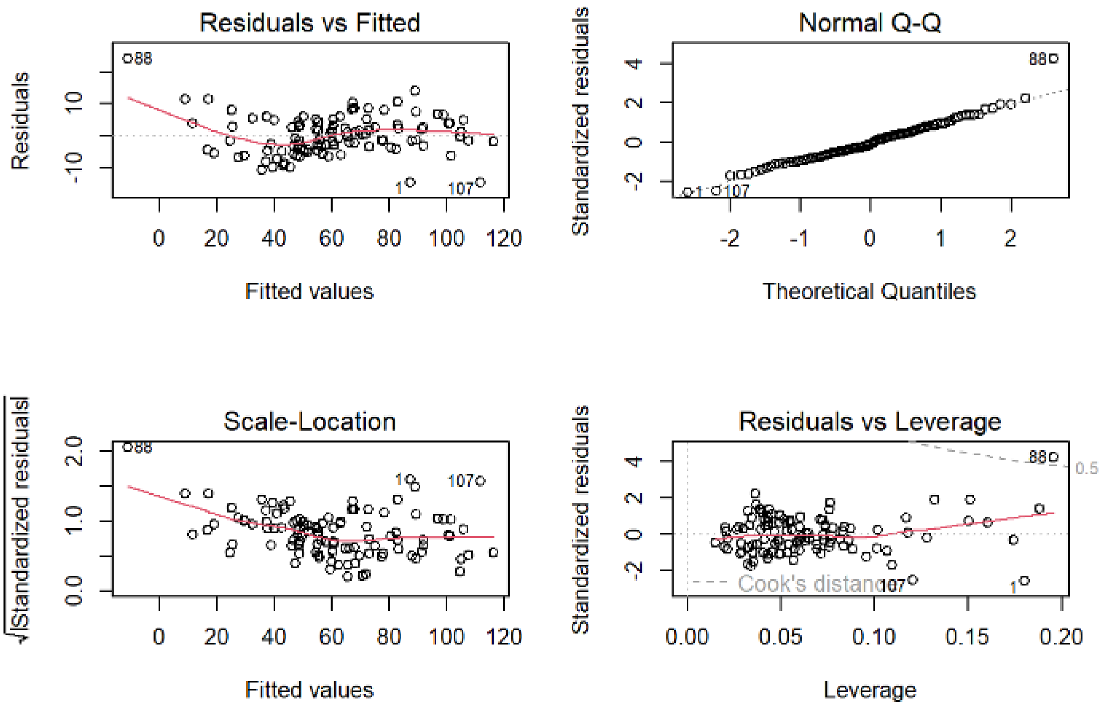
9. KEY INSIGHTS/FINDINGS FOR STATISTICAL MODEL

Below visualized are some key values for the regression models:

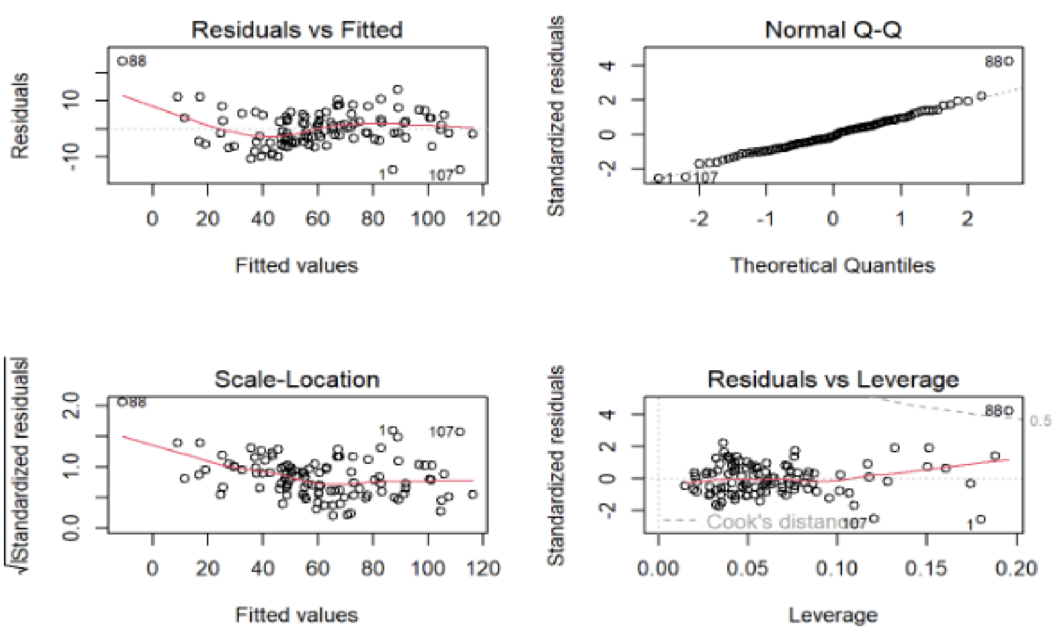
Model 1:


Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
88
99
107

Model 2:


Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
0.5
88
107
1

Model 3:


Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
0.5
88
107
1

10. POTENTIAL REAL-WORLD APPLICATIONS OF PROJECT

1. Risk Assessment and Management: The project can be used as a tool for risk assessment and management. It can help policymakers identify countries and regions that are more vulnerable to climate risks, and prioritize resources to mitigate these risks.

2. Climate Adaptation Planning: The project can help countries develop climate adaptation plans that are tailored to their specific needs and vulnerabilities. By providing a comprehensive overview of climate risks and drivers, the project can support the development of evidence-based policies and interventions.

3. Investment Decision Making: The project can be used to inform investment decision making. By identifying countries and regions that are more resilient to climate risks, investors can channel resources towards these areas, which can help stimulate economic growth and development.

4. International Climate Negotiations: The project can be used as a tool for international climate negotiations. By providing a common framework for understanding climate risks and vulnerabilities, the project can facilitate dialogue and cooperation among countries.

5. Diversity and Inclusion: The project can help correct negative stereotypes and promote a more inclusive and equitable world. By recognizing the diversity and complexity of different cultures and societies, the project can help promote respect and understanding across different communities.

11. LIMITATIONS OF PROJECT WORK

Some limitations of the project are caused due to the limitations of the data. The CRI data is only available for the year 2019. Another issue with the data is that it is purely causal. The score is calculated after the disasters have struck a country and the preparedness of the country is assessed.

1. Incomplete data: The final_combined_data dataset may not include all relevant variables that impact climate risk, such as geographic location, infrastructure, or socioeconomic factors. This may limit the accuracy of the analysis and the resulting insights.

2. Subjectivity in ranking: The criteria used to calculate the CRI scores may be subjective and may not reflect the unique challenges faced by each country. Additionally, different weighting schemes may produce different rankings.

3. Time sensitivity: The CRI scores may change over time as countries experience different climate-related events. Therefore, the rankings may become quickly outdated.

4. Lack of generalizability: The CRI scores are based on a specific methodology and may not be generalizable to other contexts or regions. Therefore, the insights gained from this project may not be applicable to other parts of the world.

12. CONCLUSIONS

We analyzed the Climate Risk Index (CRI) dataset for 2019, exploring the multivariate relationships between variables such as GDP, population density, and CRI fluctuation level. Visualizations, such as heat maps, scatter plots, box plots, and line charts, were created to gain insights into the data. We also highlighted the limitations of the data, such as its causal nature and lack of availability for other years.

Our analysis revealed that countries with higher CRI scores tend to experience a higher economic impact from climate-related disasters, with a positive correlation between the CRI score and losses per GDP. Additionally, countries with higher population densities tend to have a higher CRI score, suggesting that higher population density can contribute to increased climate risk.

Based on our findings, we recommend that countries take proactive measures to mitigate the impacts of climate change by developing more targeted policies and interventions that address the underlying drivers of climate risk. It is also recommended that countries use a more comprehensive approach to rank and categorize countries based on their climate risk. Furthermore, we suggest that more research be conducted to determine the causal relationship between different variables in the dataset to provide a more nuanced understanding of the drivers of climate risk. Finally, we stress the importance of collecting more comprehensive and accurate data to better understand the complex relationship between climate risk and human factors.

REFERENCES

Conceptual references:

1. IPCC (2014). Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Geneva: IPCC.
2. Adger, W. N. (2006). Vulnerability. Global environmental change, 16(3), 268-281.
3. O'Brien, K., & Wolf, J. (2010). A values-based approach to vulnerability and adaptation to climate change. Wiley interdisciplinary reviews: climate change, 1(2), 232-242.
4. Schlenker, W., & Roberts, M. J. (2009). Nonlinear temperature effects indicate severe damages to US crop yields under climate change. Proceedings of the National Academy of Sciences, 106(37), 15594-15598.
5. Ribot, J. C. (2010). Vulnerability does not just fall from the sky: Toward multiscale, pro‐poor climate policy. In Vulnerability, poverty and resilience (pp. 19-38). Palgrave Macmillan, London.

Technical references:

1. Pandas: https://pandas.pydata.org/docs/
2. Matplotlib: https://matplotlib.org/stable/contents.html
3. Seaborn: https://seaborn.pydata.org/
4. Scikit-learn: https://scikit-learn.org/stable/
5. NumPy: https://numpy.org/doc/stable/