National College of Ireland

BSc (Honours) in Computing

Data Analytics

2021/2022

Nikita Olijniks

16416304

x16416304@student.ncirl.ie

Analysis of Effects of Greenhouse gas on Climate

Technical Report

Contents

Executive Summary

This report analysis details the research which was conducted in order to complete the reporting of the report. The methodology used, help to determine the trends and other relevant aspects of the data for better analysis. The overall aim of this analysis is to collect datasets and build a visual report based on the datasets and information that is collected and to draw conclusions from this information.

This report is completed by the help of the methodology. Various datasets were gathered for the purpose of this analysis from open-sourced websites like Kaggle.com, covering different aspects and regions involved in the gas emissions and rise of temperature. This report allows for a prediction build forecast and the findings allow us to better see the different changes associated with the fuel consumption and rise in greenhouse gases. Using the forecasting techniques, we can determine that the rise in temperature will continue if the fossil fuel consumption continues and if the emissions continue with the rise. It is also important to note that, not all regions produce the same amount of emission, but all regions get affected by the increasing rise of greenhouse gases.

1.0 Introduction

1.1. Background

It was a very hard decision when choosing the correct topic for the purpose of this analysis, so when I choose, I choose this project as I had a personal interest in the matter of ecosystems and global ecosystem from a young age. I wanted to gain a better insight into the area and establish the information which I could find out from available datasets out there.

Growing up we are often told and taught that we are currently living in a climate crisis and the conditions and temperatures are rising. I have heard a lot and learned a lot about the effect of C02 on our environment, and personally I would like to review the data available to create a perspective for myself on the matter.

Evidence shows that the concentration of the CO2 and CHG in our atmosphere is currently rising with a correlation of temperature increase from these increases in the C02 emissions. CO2 levels are rising due to the main effect of crude oil consumption and deforestation. (F., 1990). So, for the purpose of the analysis, I would like to see if there is any correlation between the rising temperature and CHG / CO2 emission and also create some visual representation while displaying this information.

Also, according to the 2018 study, carbon dioxide from human activities accounts for 90% of the gas emission in the greenhouse (RICE, 2018). I also believe and hope that some information might become useful, and currently the current trends politically globally, are moving “Greener” and there is an emphasis on things such as climate change and moving further away products that can produce too much of CO2 imprint on our environment. I want to understand why, determine potential trends of CO2 effects, and visualise the information that is available from the data. This report will analyse whether the CO2 emissions causes the increase of the temperature over the world, and we will also build a prediction model to determine the temperature and emissions in the future.

1.2. Aims

There are a wide range of different aims and objectives for this project which I have set out and hope to achieve, listed below in table. The main aim is to conduct an analyse and breakdown of the data into a form of information where we can find the unusual or abnormal outliners and maybe some new form of information which can help us create correct visualisation of the data and to build our prediction model further.

Aim 1: The main aim of this analysis and project is to choose the correct and relevant datasets for the main goal objective. Data on CO2 Emissions and the global weather records dataset will be collected in order to conduct data analysis.

Aim 2: Once the appropriate and suitable datasets have been found and identified, the next goal is to pre-process the data so that it is suitable for further analysis and Prediction model implementation. Cleaning and filtering the datasets for specific period of time, for specific data and regions is part of this method.

Aim 3: After pre-processing, there will be a need to combine similar datasets and make one dataset from which information can be easily retracted from, and the data has been merged together (if merging will be required) after the transformation phase. The next move is to conduct exploratory data analysis (EDA) to identify associations between datasets, such as a correlation between global weather patterns and Emissions from C02.

Aim 4: After completing exploratory data analysis (EDA), the project's next goal is to prepare the data for making the prediction model.

Aim 5: The final goal is to record all of the study's findings and visually represent the results once the objectives have been met. Evaluate the collected data, analyse, draw a conclusion, and complete all the documentation, results, and visualisations in a comprehendible manner and define the outcomes of analysis.

1.3. Technology

The technical developments which are carried out in this project varies from the stages of the process. The most appropriate and usable for the completion of this project was used.

R Studio:
R Studio is the main program used for this project. It was used to clean, merge, manipulate and store the necessary data needed. R Studio is an open-sourced program for developments with R programming language is the main programming language used in this analysis as it's the main

language in R Studio. R Studio also has built in packages which are very useful, such as tidyr, dplyr and ggplot2, which all have been used for the implementation of this project.

R Language:
R Language is the main language used for the purpose of this analysis, hand to hand with the KDD methodology that was used to conduct this analysis report.

SPSS: SPSS (Statistical Package for Social Sciences) used for running statistical tests and data visualisations and for creation of some graphs.

Excel:
Excel is a spreadsheet tool that has a lot of statistical and visualisation functions that will be used throughout. Excel was mainly used for viewing the necessary data, as all the data was stored in the excel forms. Excel is also used for manipulating/cleaning data and for overview of the data analysis.

Google Collab:
Google collab was used in order to conduct the research on bigger datasets, as it allows to execute Python scripts that may take longer to conduct on the local machine.

1.4. Structure

- Section 2: Detailed descriptions of all the data used.
- Section 3: The methodology used, and how it was followed by step.
- Section 4: Details of the analysis that was conducted.
- Section 5: Reporting of the results that was collected.
- Section 6: Conclusions that were drawn from the research.
- Section 7: Further / future developments for research
- Section 8: Appendices.

2.0 Data

The data that I have acquired for the analyse, is in relevance to the conditions which I want to locate. For the purpose of the research, it was hard to acquire full and available datasets for free as the main datasets with a lot of information's were behind a paywall. So, in order to prepare for the analysis of the reporting I had to create a separate dataset which can then be used to interpret the results from, as a lot of datasets had missing values and had to undergo extensive cleaning.

In other words, the data that I have gathered for this project has inputs of Co2 emissions in the air and sea level. I have also collected data with regards to the changes of temperature in the which I would like to make a comparison, which are combined for better analysis to extract the necessary data.

The datasets that I must analyse are publicly available and they come in format .csv I have found all the data sets publicly available online to use and they can be accessed without any extra passcodes or passes, apart from a google account login in a different instance.

Data File	No. Of Records	Description	Attributes	File size	Data Format	File Type
CO2_GHG_emissions-data	20,853	File with CO2 emissions for each country	4	571KB	Structured	.CSV
city_temperature	2,906,327	Daily level of average temperature of major cities	8	140.6MB	Structured	.CSV
GlobalLandTemperaturesByCity	8,599,212	Global land temperatures by city	7	520KB	Structured	.CSV
GlobalLandTemperaturesByMajorCity	239,177	Global land temperatures by major city	7	13.8KB	Structured	.CSV
GlobalLandTemperaturesByState	645,675	Global land temperatures by state	5	30,049KB	Structured	.CSV
GlobalTemperatures	3,192	Overall Global land temperatures	9	202KB	Structured	.CSV
Globalfuelemissions	61	Has a breakdown of fuel type and emissions from those fuel types from 1950-2010	8	7KB	Structured	.CSV

Datasets breakdown continued:

- "Climate Change: Earth Surface Temperature Data" – contains 5 of the Global temperatures' datasets (GlobalLandTemperaturesByCity, GlobalLandTemperaturesByCountry, GlobalLandTemperaturesByMajorCity, GlobalLandTemperaturesByState and GlobalTemperatures) – Data is collected and run from 1750 – 2013, has to be noted that there could be some bias as in 1940 airports stated to get build and weather stations had to be moved and in 1980 there was a electronic thermometers are said to have a cooling bias ; sourced from Kaggle.com
- "Daily Temperature of Major Cities" – dataset that contains the daily average temperatures for cities - data collected from 1995 – 2020 ; Sourced from Kaggle.com
- "CO2_GHG_emissions_data" – Dataset that contains C02 and GHG emissions for the period of 1750 – 2017 ; sourced from Kaggle.com

- “Global CO2 Emissions from Fossil Fuels since 1751” – data set that contains emissions from fossil fuels since 1751, but for the purpose of full record keeping I will be only using records from 1950-2010 ; sourced from datahub.io

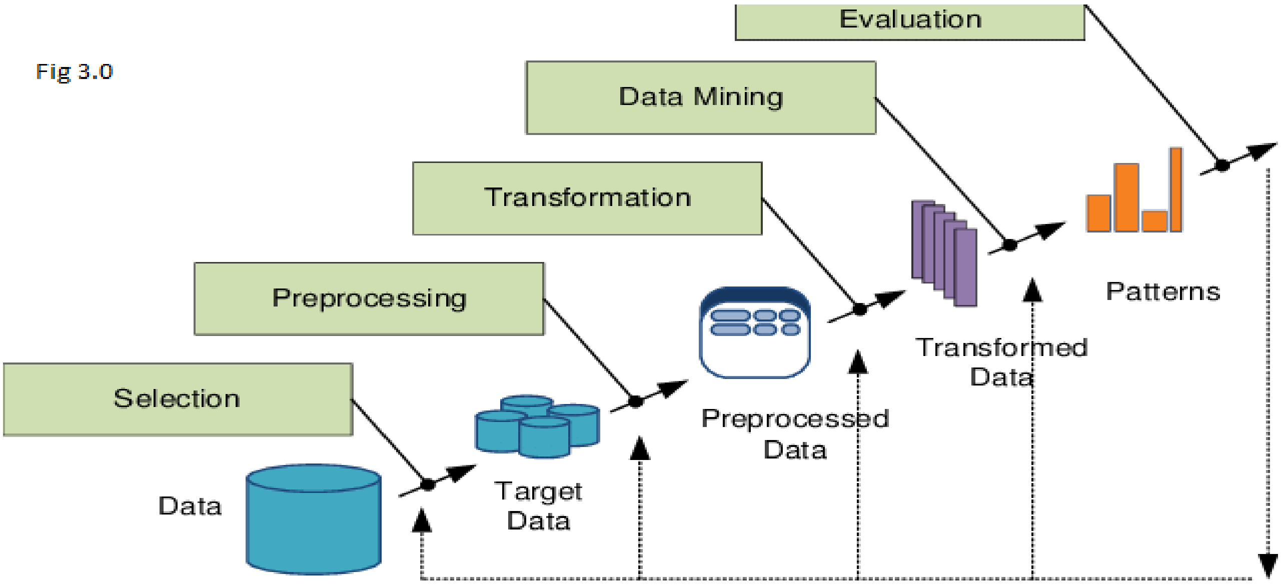
3.0 Methodology

For my methodology I am using the KDD or Knowledge Discovery Databases methodology. KDD methodology refers to the process which involves the finding of information and knowledge in a data set. It is more suitable for this type of research, than the CRISP-DM, as KDD methodology puts emphasis on research. Also, another reason for my selection of this methodology is because I have previously had experience in applying this type of methodology for a similar project and found it very suitable and approachable.

By using the KDD approach, the work is dividing the workload into separate KDD stages which will make the work more effective and approachable, this allows for better understanding of the end goals of what I want to achieve out of this project, and how I want to structure the research i.e., focusing on a subset of data or data samples, combining multiple data sets together.

The following approach (fig 3.0) is applied in sequence for the purpose of the analysis:

Fig 3.0



3.1 Data selection and Description

This first steps of the methodology really focuses on the area of selection for the datasets and the descriptive information regarding the data that was used. For this research the datasets have been openly sourced from Kaggle.com. The data that was used represents crucial part of achieving the set-out goals for the purpose of the research. All the data is downloaded in the CSV (comma separated values) format. For the purpose of the project aspect the datasets were extracted from Kaggle.com with the Kaggle API using the Application Programming Interface (API) and saved into the .CSV format. For the purpose of the analysis the data had to be checked for missing values, and after initial research into the datasets, it was obvious with the number of missing values for older dates, as record keeping wasn't great till mid 19 century, so for the purpose of this research the period of time which was investigated is 1945-2013. Also due to the nature and the amount of the countries that are available I will be looking at the 10 countries of choice to compare and analyse.

3.2 Data Processing and Transformation

After completing the selection of the necessary datasets, the next most important step was implicated in order to transform, process, and combine all the relevant necessary datasets. Data pre-processing was one of the important parts of this analysis. Data manipulation techniques were used for detecting, removing, and correcting of the incorrect and missing data, and for merging. This stage was important as to get a proper and justified results later when we are building our prediction model and conducting the analysis.

3.2.1 Pre-processing of data and transformation

Data cleaning strategy had to be implemented before any analyse could be done on the datasets. I determined the missing values, replacing the missing values with necessary changes if applicable, using the "!" method in R. I am investigating if data sampling is an easier choice, depending on the size of the data that I have for the analyse, this will help me decrease the chances of repetitive data elements. I also investigated whether some data would need to be separated into separate subsets of datasets for better accurate results. Since the data was so large, it was necessary to combine the datasets together for which a new dataset was create "newDataset.csv" for which the attributes of countries, CHG and CO2 emissions and average yearly temperature.

In order to get "newDataset.csv" dataset I had to conduct a few necessary steps to first check the validity of the datasets. I loaded them into R Studio and checked what missing values were present and what need to be replacing. I noticed that a lot of data was missing up to the year of 1949, so for the purpose of the research I only selected the data needed for this research which was from 1949-2013.

I had to calculate the average yearly temperature for each of the 6 countries selected, creating a new column "Average_Yearly_Temperature"., by selecting the 12-month period and find the average temp for the year. Also, while checking for the average temperatures for the region I have realised that one of the regions Canada was combined for the records with USA and label as "North America", so for the purpose of the analysis I had to find the mean temperature for both countries for specific year and joint them together, the same applied for CHG and CO2 Emissions attributes for USA and Canada and for the purpose were labelled under "North America".

Details of the combined datasets are as follows:

DATA FILE	NO. OF ATTRIBUTES	ATTRIBUTES	DATA TYPE
newDataset	5	Country, Code, Year, Annial_Emission_Tonnes, Average_Yearly_Temperature	Chr,Int,Num
globalfuelemissions	8	Year, Total, Liquid.Fuel, Solid.Fuel, Cement, Gas.Flaring	Chr,Int,Num

3.3 Data Mining

The data mining process allows for the search for different patterns and trends which are in interest that I can locate within the transformed datasets. Conducting multiple checks prior to creating a prediction model and getting other information out of the datasets at hand, such as correlation, linear regression, classification. Using the different methods, the information gathered was represented in the form of plots and graphs and other visualisation purpose for results and documentation

I then developed the methods that I can use to search for these patterns using data mining algorithms. This will involve the process of deciding models and appropriate parameters to research for accurate projection.

3.4 Interpretation / Evaluation

This stage involved the evaluation of the of the mined results and determining the findings. The findings help us evaluate the findings and base the research on the findings. We are also able to investigate in detail on what patterns did we locate, what could we learn from the patterns and projections. As the research began with the implementation of the KDD methodology into the project, all the files have been processed in some shape or form within the R Studio. All the files were imported and processed in R studio using the R programming language. Which allowed to use packages within the R studio platforms such as Tidyverse, which allowed for multiple access to other packages such as tidyr, ggplot2, and dplyr. I have also used R studio to export and select parts of the dataset that are needed for separate analysis of the observations, such as diving each country into a separate dataset for better view of distribution. Other functions such as cbind() & rename() were also frequently used to help with the navigation of the research.

With the help of R studio, I was able to build the necessary prediction model that is required for the aim and r studio was also used for the visualisation purposes. There was a wide variety of graphs and plots used, such as histograms, boxplots, graph plots.

4.0 Analysis

The main aim and objective of this analysis is to see how various usage of CO2 and CHG can affect the main increase of temperature within the industry and locate any other information we can gather within the datasets present. The Analysis is done over a 65-year period, due to limitations in the data that is available for the purpose of the research.

With the data being mostly with regards to emissions and temperature, using the different visualisation is the most used method for the purpose of this analysis, as it is more effective and easier to conceive the information from such research and with the data size quite small makes it easier for assessment of the information. When conducting the prediction test the variables used for prediction are “Year” “Annual_Emission_Tonnes” and “Average_Yearly_Temperature”. After the results were found they were visually represented accordingly. Using R in R studio has been a great help with the creation of visual representations of the data located.

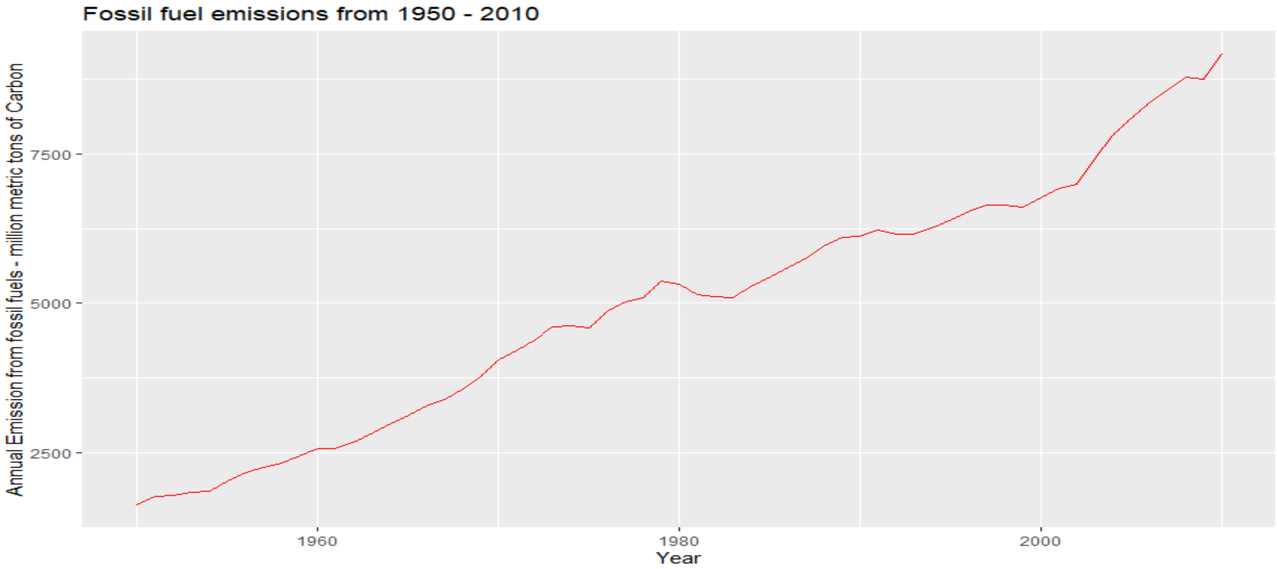
5.0 Results

There are multiple results drawn from the analysis of the investigation, each are discussed below accordingly.

5.1 How much of Emissions are getting extracted into the atmosphere?

From the initial calculation for the period of 1975 -2013 the recorded total average amount of C02 and CHG gasses released are at 145710201.3 million cubic litres of carbon gas. Fossil fuel is the number one affecting major for the reason of the increase of greenhouse gases.

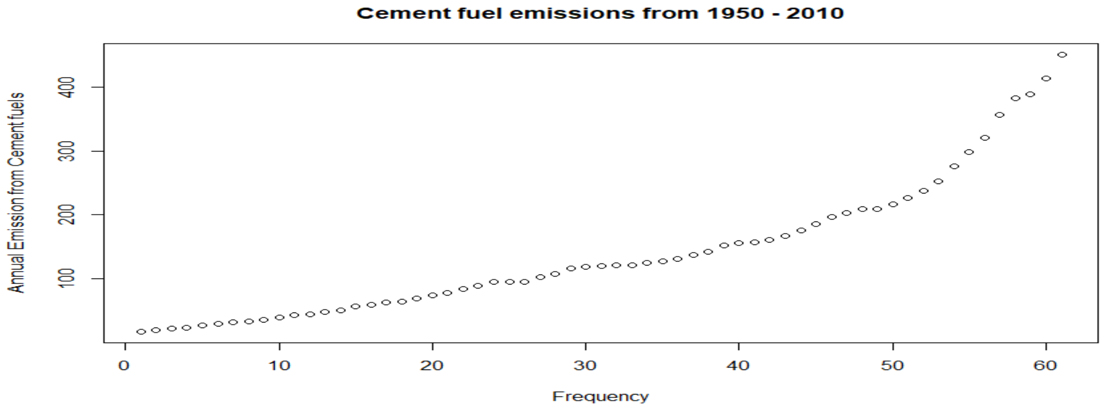
For the first part of the analysis, it was essential to find how much emissions are getting extracted into our atmosphere. In the below figure we can see how much of increase in annual fuel emissions from 1950-2010 measured in million metric tons of Carbon.



From the above graph of different types of fuel emission of carbon, we can see a high increase since 1950 right up to the 2010 when the data ends. We can see that there is a gradual increase in the emission of carbon through out the given time period of 60 years.

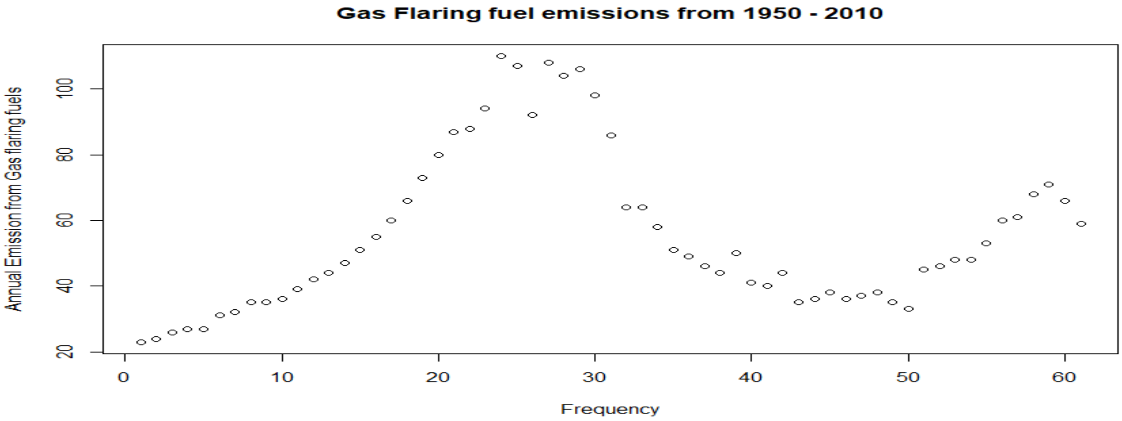
Conducting a similar analysis on each of the fuels on record Gas, Liquid, Solid Fuels, Gas flaring and cement we can see that the frequency of use is increasing in all products apart from Gas flaring.

Frequency of Cement Fuel emissions 1950-2010 below:

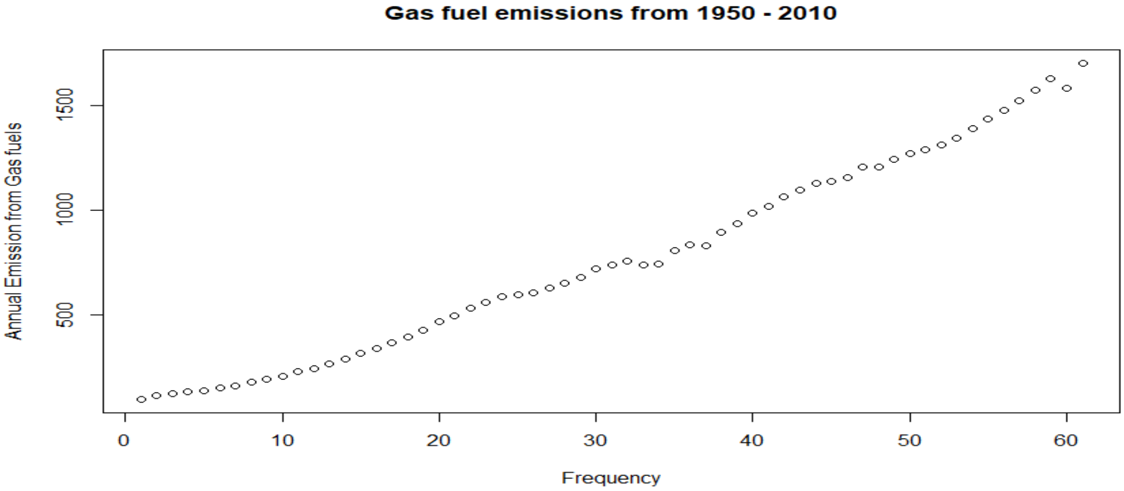


.

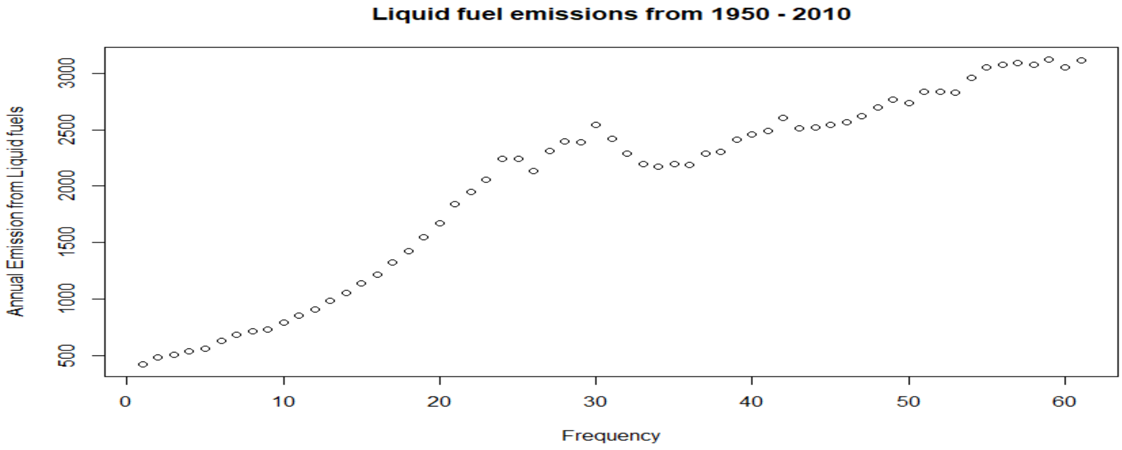
Frequency of Gas Flaring Fuel emissions 1950-2010 below:



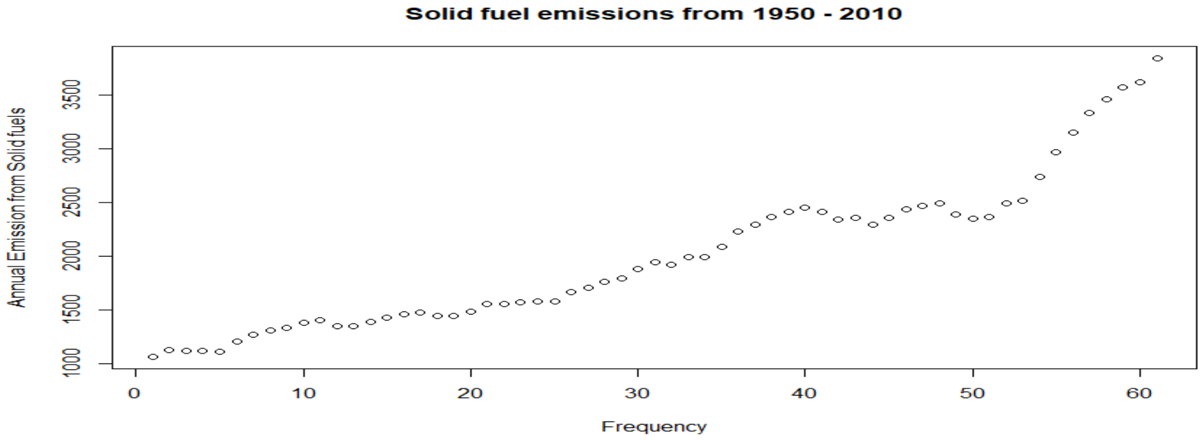
Frequency of Gas Fuel emissions 1950-2010 below:



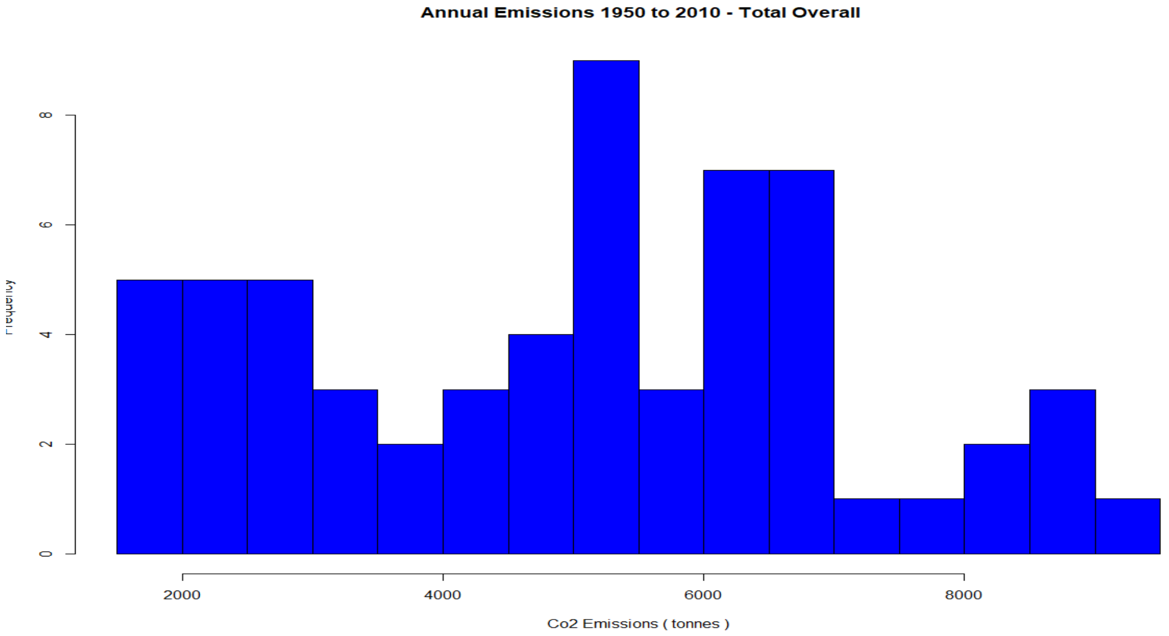
Frequency of Liquid Fuel emissions 1950-2010 below:



Frequency of Solid Fuel emissions 1950-2010 below:

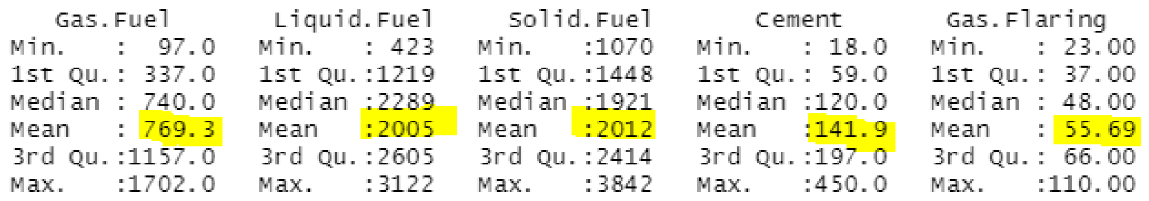


There is a clear increase in the frequency of the increase in the fuels that are used and, in the emissions, created by these fuels.



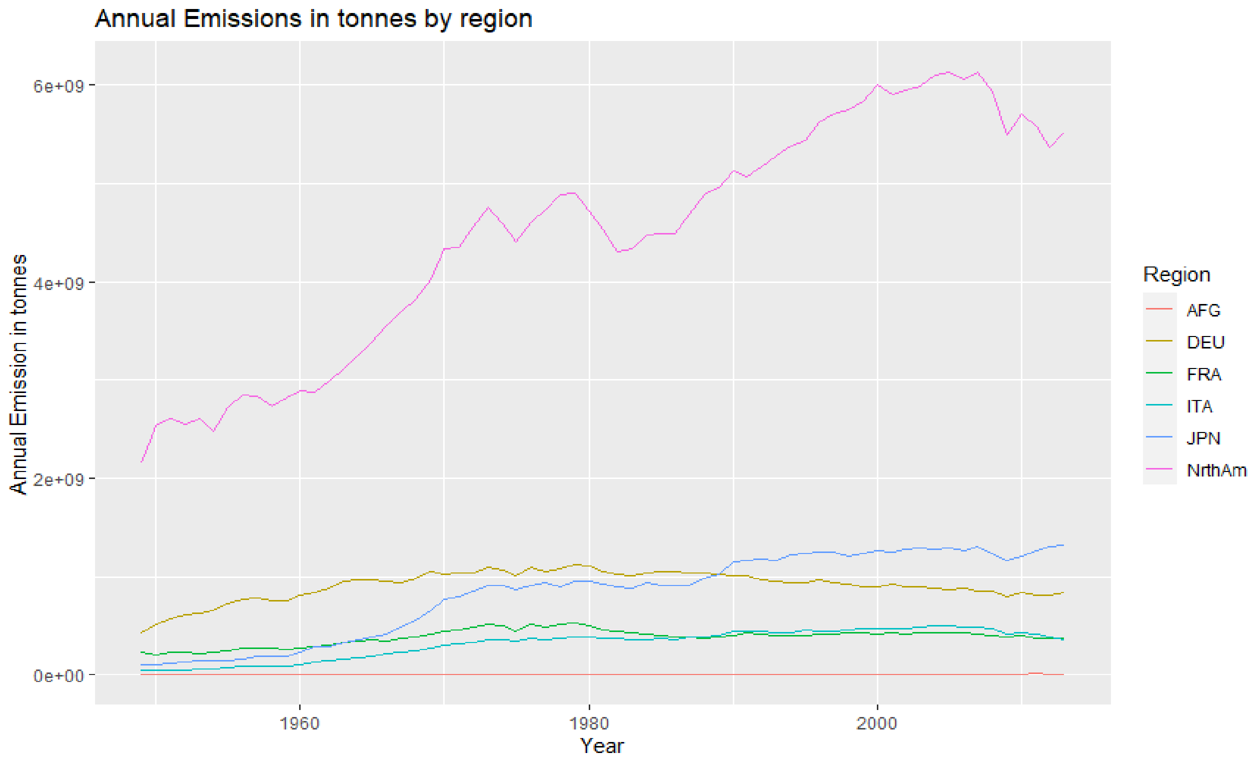
Frequency spread of Annual Emission needed to check for the prediction model build.

Using the summary() function in R we also run the descriptive analysis on the datasets to see which of the fuels are the most emission releasing fuels in the category. By looking at the below table of the summary() of the Emission of fuels dataset you can see that by the mean emission the Liquid fuel emissions account for most of the emissions noted. You can also see the same results if you look at the plot for the frequency of liquid fuel emission as shown previously.



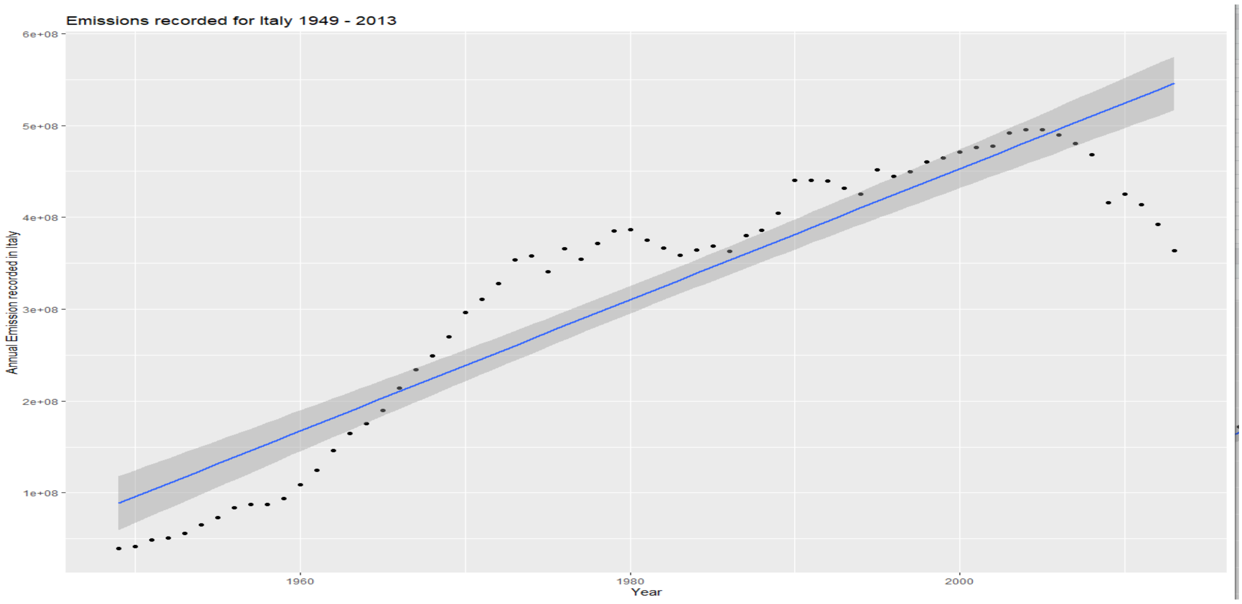
```
    Gas.Fuel          Liquid.Fuel      Solid.Fuel        Cement          Gas.Flaring   
 Min.    :  97.0   Min.    : 423    Min.    :1070   Min.    : 18.0   Min.    : 23.00  
 1st Qu.: 337.0   1st Qu.:1219    1st Qu.:1448   1st Qu.: 59.0   1st Qu.: 37.00  
 Median : 740.0   Median :2289    Median :1921   Median :120.0   Median : 48.00  
 Mean   : 769.3   Mean   :2005    Mean   :2012   Mean   :141.9   Mean   : 55.69  
 3rd Qu.:1157.0   3rd Qu.:2605    3rd Qu.:2414   3rd Qu.:197.0   3rd Qu.: 66.00  
 Max.   :1702.0   Max.   :3122    Max.   :3842   Max.   :450.0   Max.   :110.00  
```

I wanted to better understand how much and how it compares with the regions with regards to increase of emissions. For this like mentioned in the report above it was necessary, to combine 2 out of the 7 regions in order to build a perspective for the America and Canada region as they are branded in this report as North America.

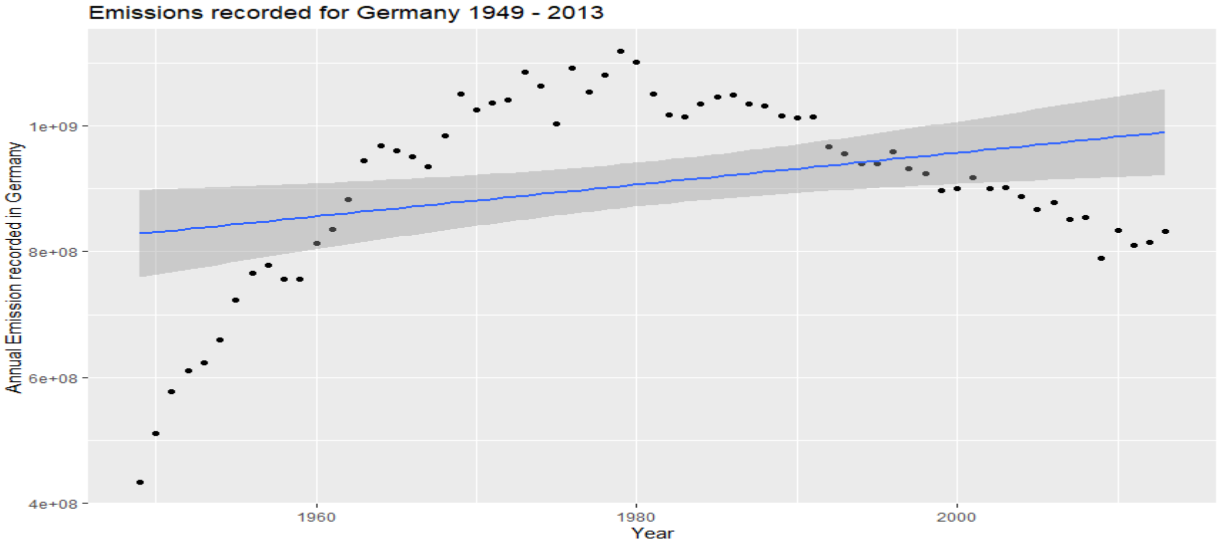


You can also see this increase in the CO2 and CHG emissions by region graph, North America represented as the top pink line, also by the legend, represents highest amounts of C02 and CHG emissions recorded, this is due to the huge area that is America and Canada, and the precise recordings which has been done throughout the time. Other random selected countries that were sampled for the purpose of this comparison were Japan, Italy , France, Germany and Afghanistan. Afghanistan with showing the lowest results for CO2 and CHG emission, again proving the incorrect data collection in the region compared to the other results which is very different.

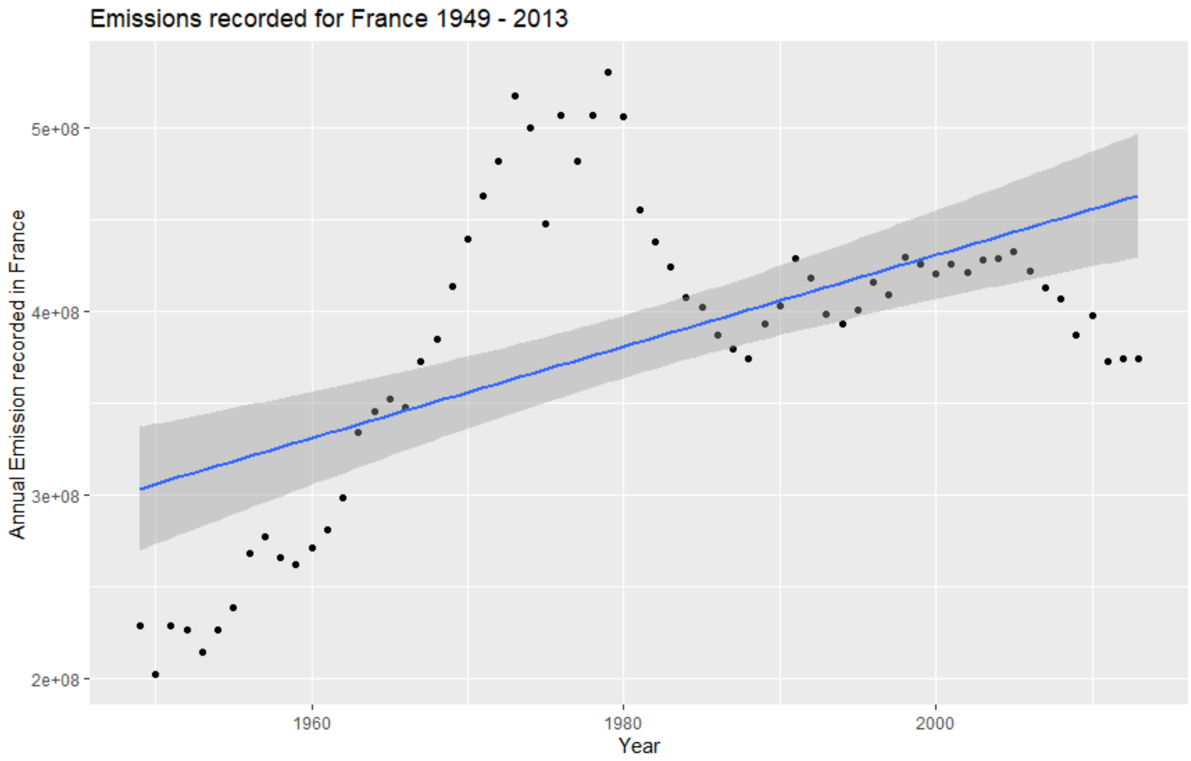
When looking at them separately by region we can see the distribution for each country below for their emission recording of CO2 and CHG of the selected regions, It has to be noted that for the European region there has been a decline in CO2 and CHG emissions:



Italy has an increase of record high emissions with the tail dipping in the recent years to negative numbers.



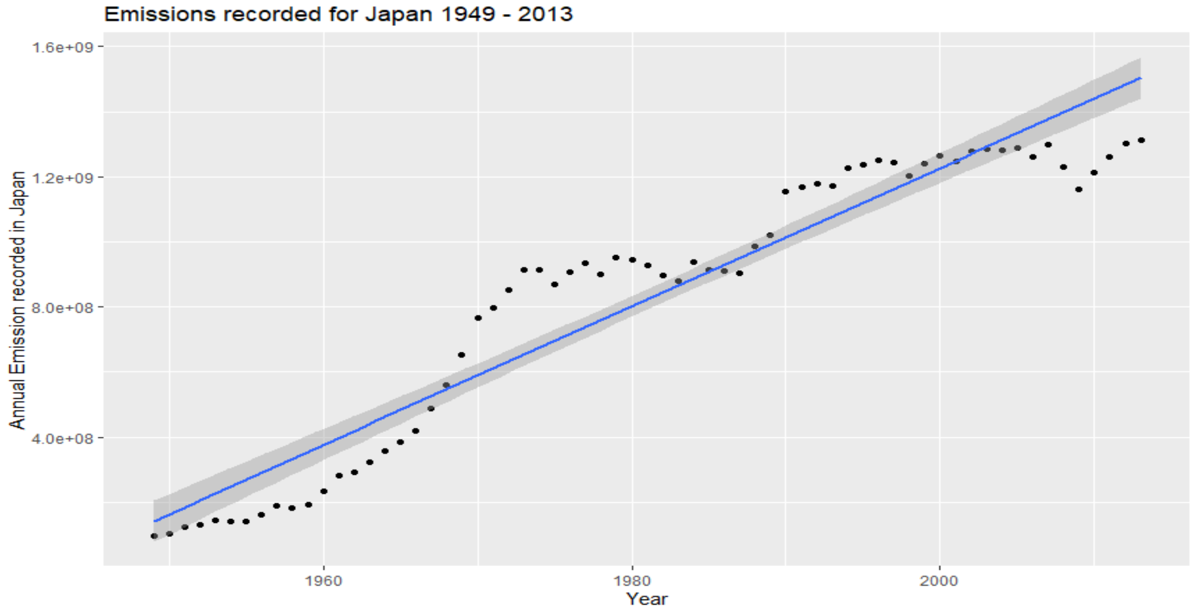
Germany another country with originally high levels of emissions and now with a negative tail, meaning there has been a decrease in the region in the recent years.

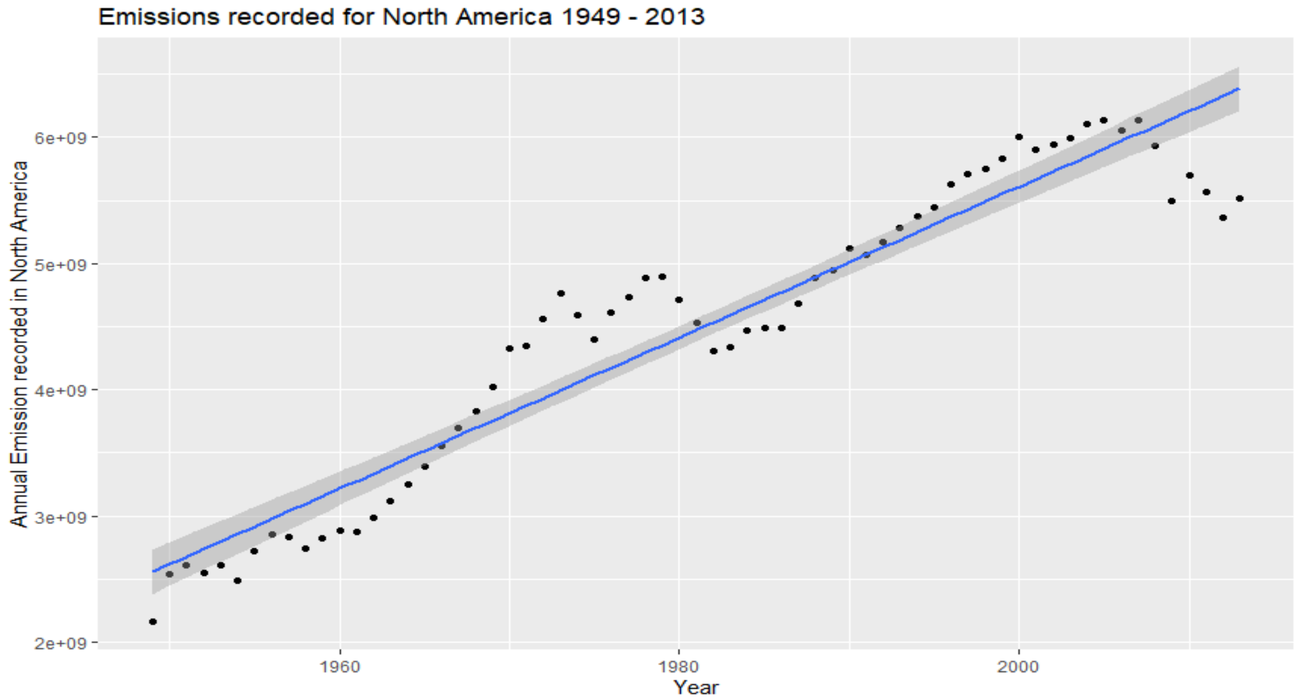


France can also be noted with the decrease of the CO2 and CHG records and with a negative tail, meaning that there is a trend towards less CO2 and CHG pollutants in Europe.

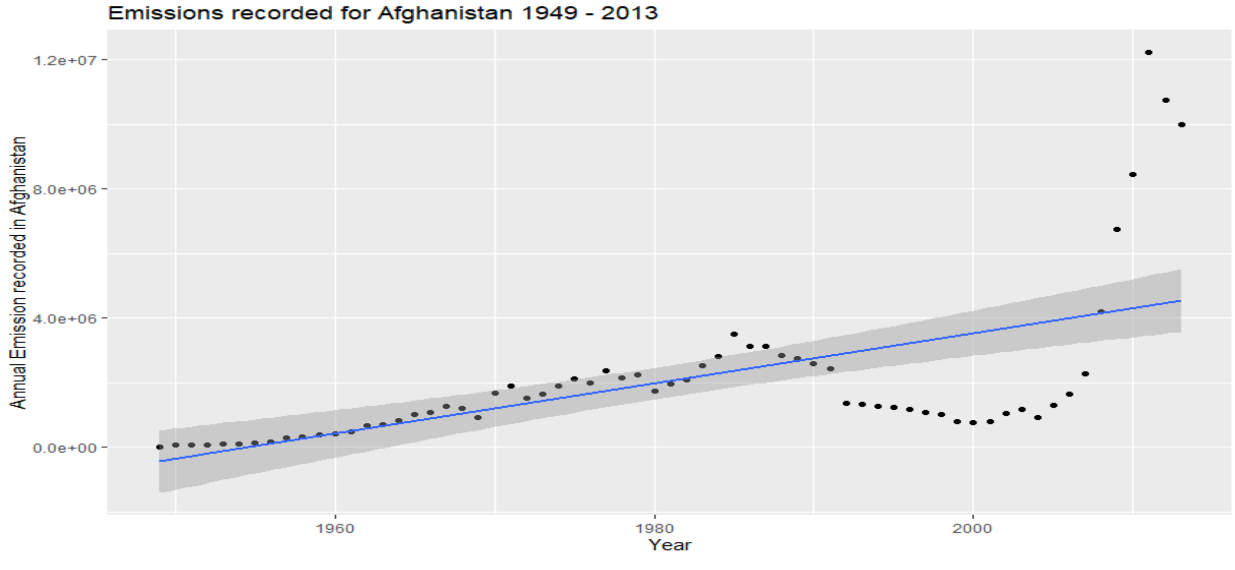
With the European countries the trend is negative, meaning that there is a steady decrease in the emissions in the region, just by looking at the 3 regions in the European union. But as Our figure shows with the regions, Europe accounts only for a small portion of the emissions caused. Now lets take a look at the Western side of the Atlantic to compare their levels of Emissions, below we have results for North America and Japan:

For the Japan region, we can see a constant growth in the emissions of CO2 and CHG.





The growth for North America (USA and Canada) can be noted as very significant and also accounts for most of the emissions in the current dataset.

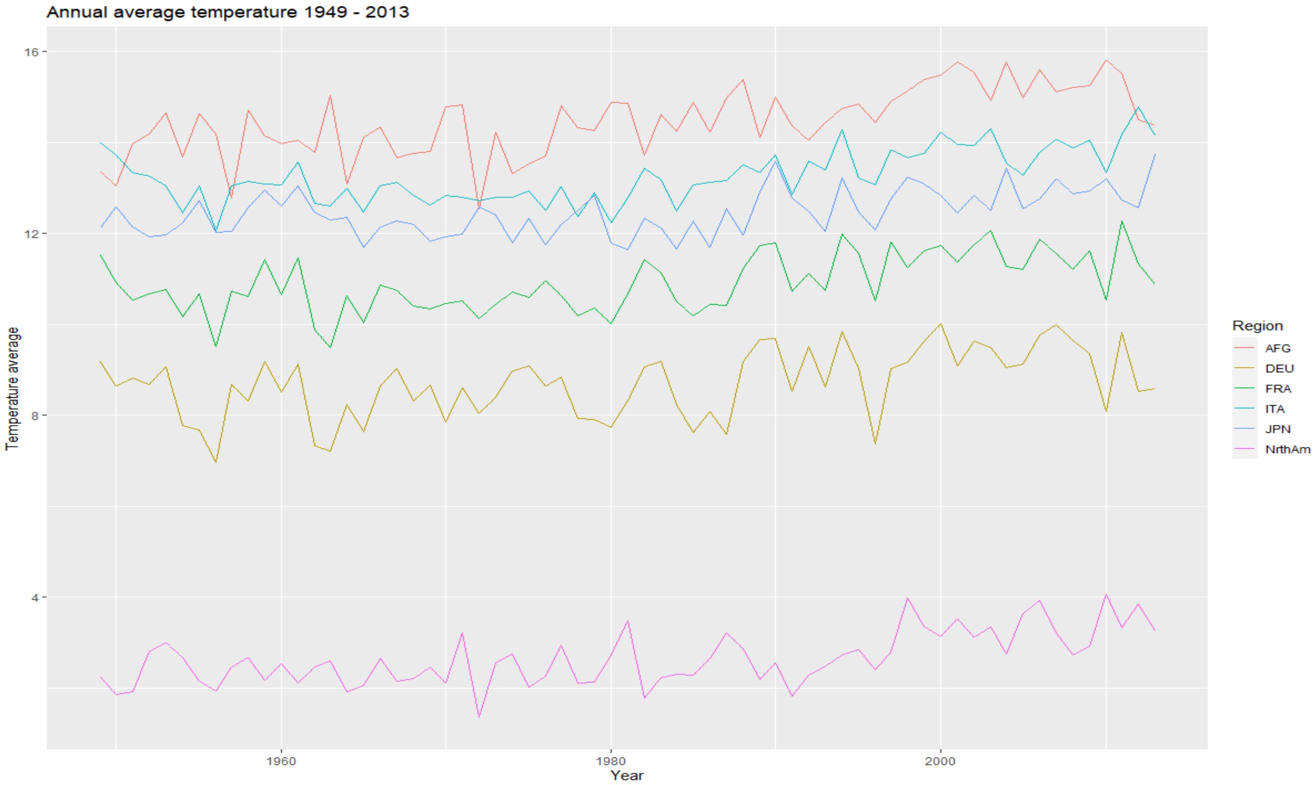


Afghanistan had to be excluded from the dataset as for the purpose of the analysis and building of the prediction model the data was not sufficient enough. As shown per the graph above.

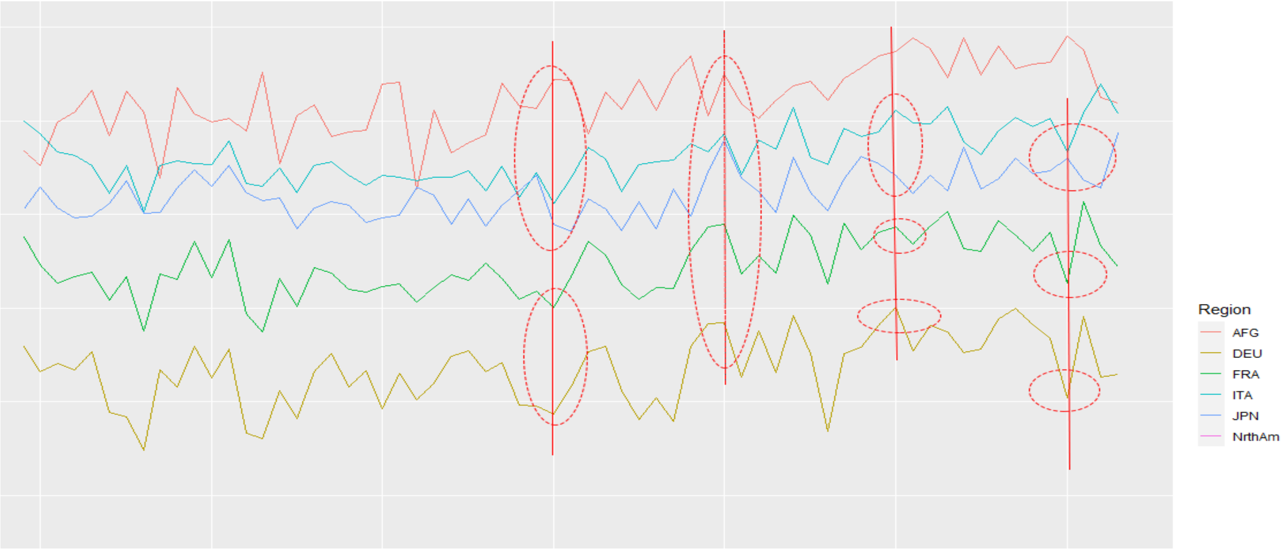
From the above it is clear to see that there is a clear increase in the overall usage and expulsion of CO2 and CHG gas into the atmosphere, but it has to be noted that a decrease has been noticed in the EU region compared to North America and the Japan region. It is also clear that the Gas liquid is also one of the most common forms of fuel for emissions with an average of 2012 million cubic litres of gas out in the atmosphere over a 1950 – 2010 period for the 6 regions that we looks at.

5.2 Is there an increase in temperature recorded?

For this part of analysis, I wanted to find out whether or not there is an increase that has been noted in the past 65 years in those selected regions to have a better understanding of the increase of the temperatures. It has to be noted that different regions have different climates, so their average for a 60-year period is different, so the frequency of the distribution of the data is a bit spread out. The average recorded temperature that is noted from 1750 – 2013 is 8.3 degrees Celsius worldwide. This includes all the recorded locations, which is over 300 locations.

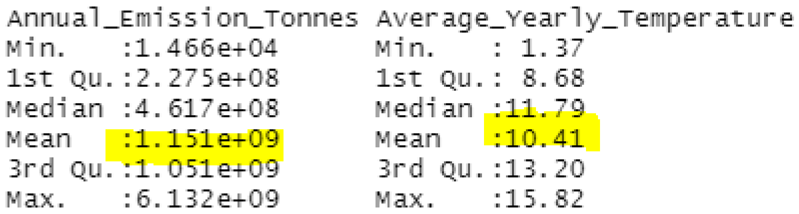


Annual average temperature ranges from 1949 – 2013, respectively. It has to be noted that a slight upward trend can be noted from the above graph. Another interesting trend which I have spotted that it seems when its cold in one region that a lot of the regions follow the same trend with lower temperatures across the board as shown below:



From the above graph we can make a clear assumption that the average temperature worldwide is affected by other factors and regions in the world.

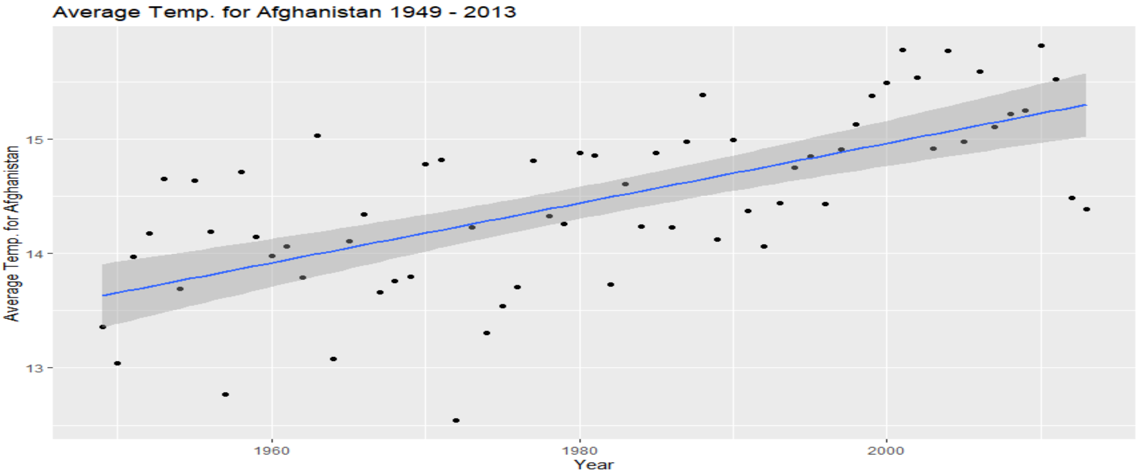
Also you can see the different regions and their annual average temperatures different from each other depending on their climate and location.



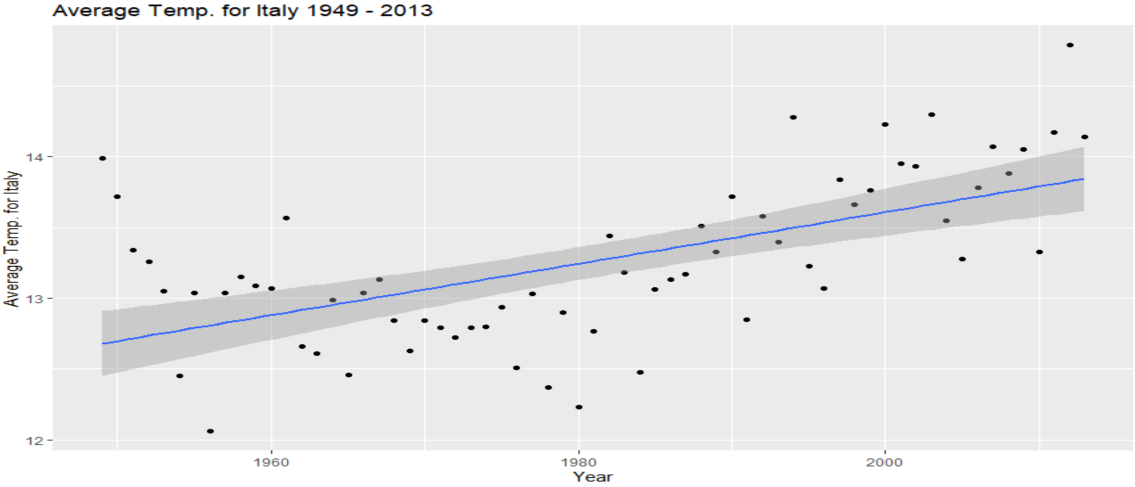
```
Annual_Emission_Tonnes  Average_Yearly_Temperature
Min.    :1.466e+04      Min.    : 1.37
1st Qu.:2.275e+08       1st Qu.: 8.68
Median :4.617e+08       Median :11.79
Mean    :1.151e+09      Mean    :10.41
3rd Qu.:1.051e+09       3rd Qu.:13.20
Max.    :6.132e+09      Max.    :15.82
```

The average temperature for the period of 1949 – 2013 is 10.41 degrees Celsius. This can also be shown using the summary () function for the dataset.

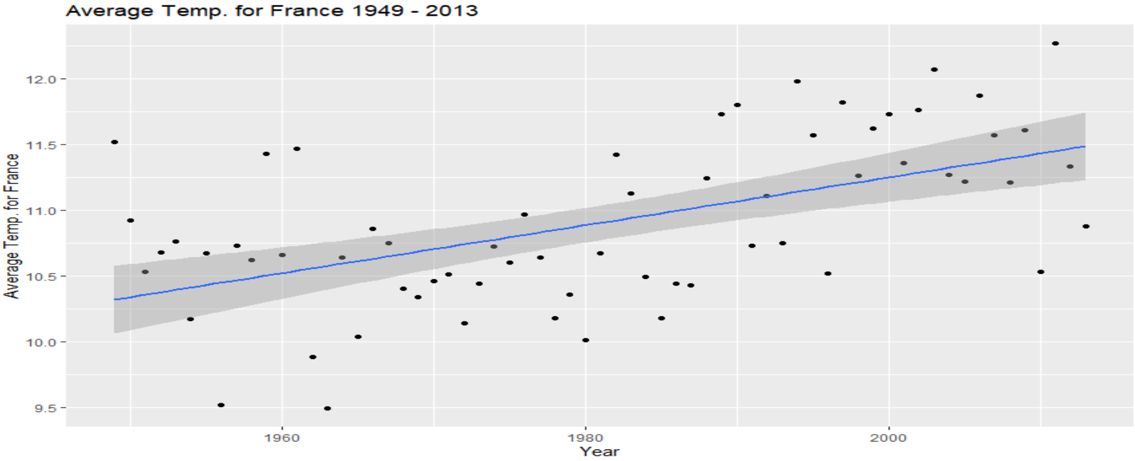
I then look into the analysis of spread, of the average temperature in each of the regions the same way as I did for the emissions, the results are below as follows:



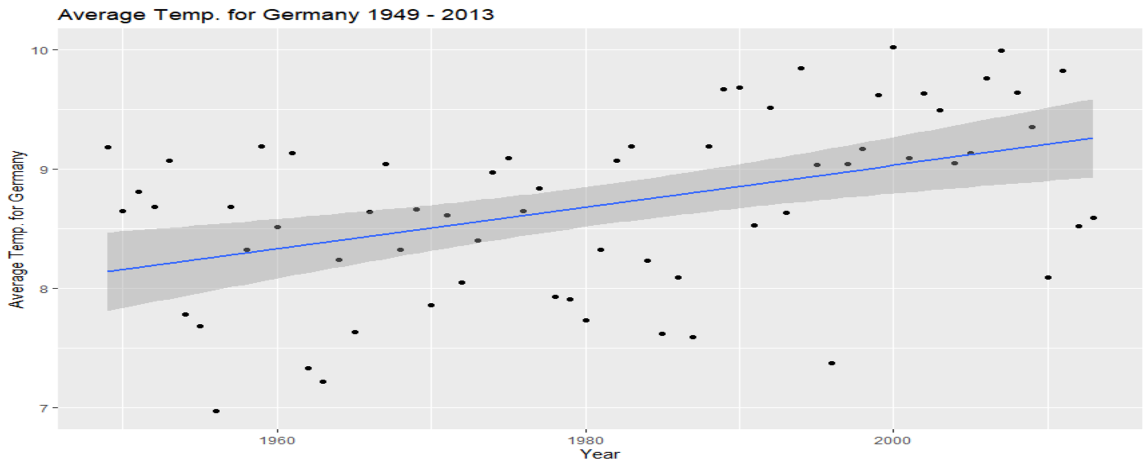
The average temperature for Afghanistan rising year on year as seen in above figure. Differs quiet heavily from the graph of emissions of this region.



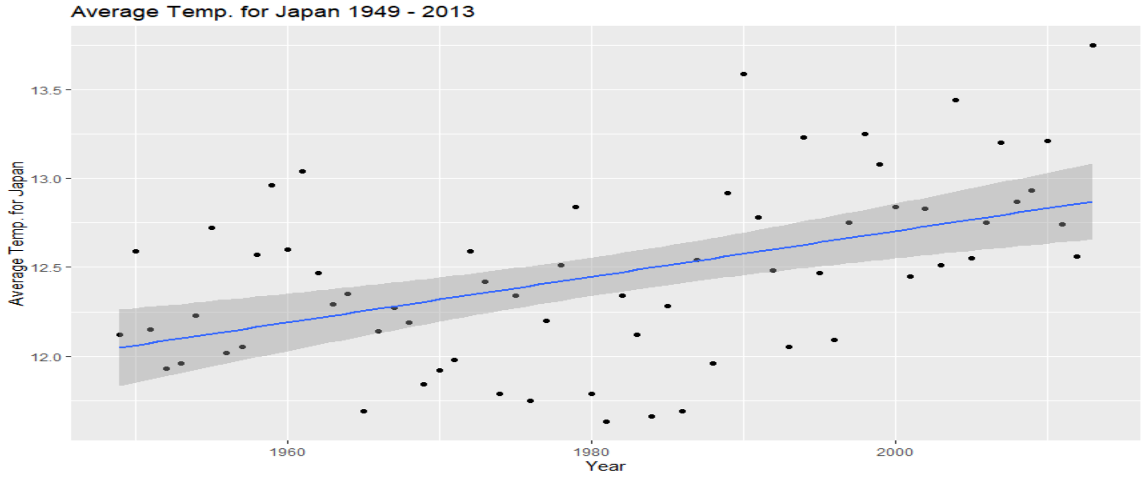
The average temperature for Italy rising year on year as seen in above figure. Europe average temperature still on the rise, even with less emission drop off.



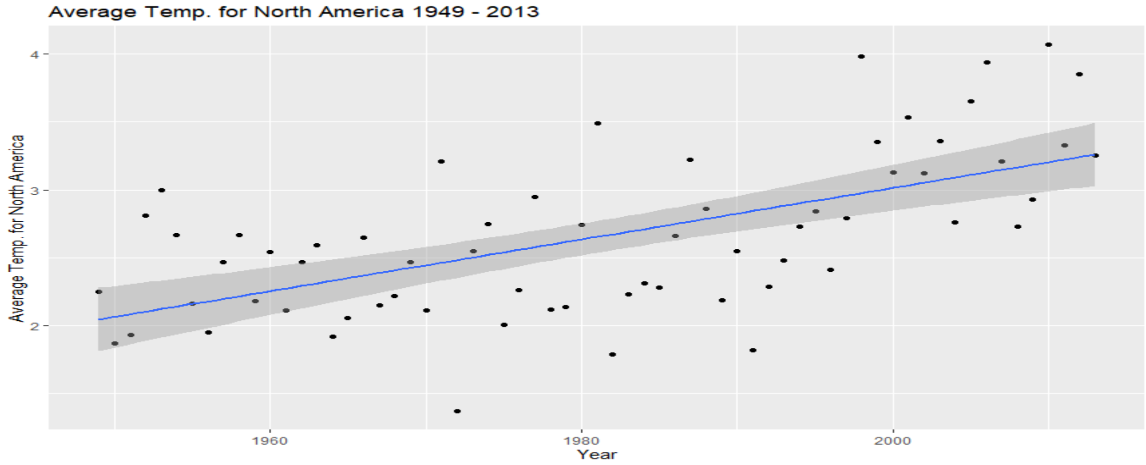
The average temperature for France rising year on year as seen in above figure. Europe average temperature still on the rise, even with less emission drop off.



The average temperature for Germany rising year on year as seen in above figure. Europe average temperature still on the rise, even with less emission drop off.



The average temperature for Japan rising year on year as seen in above figure.

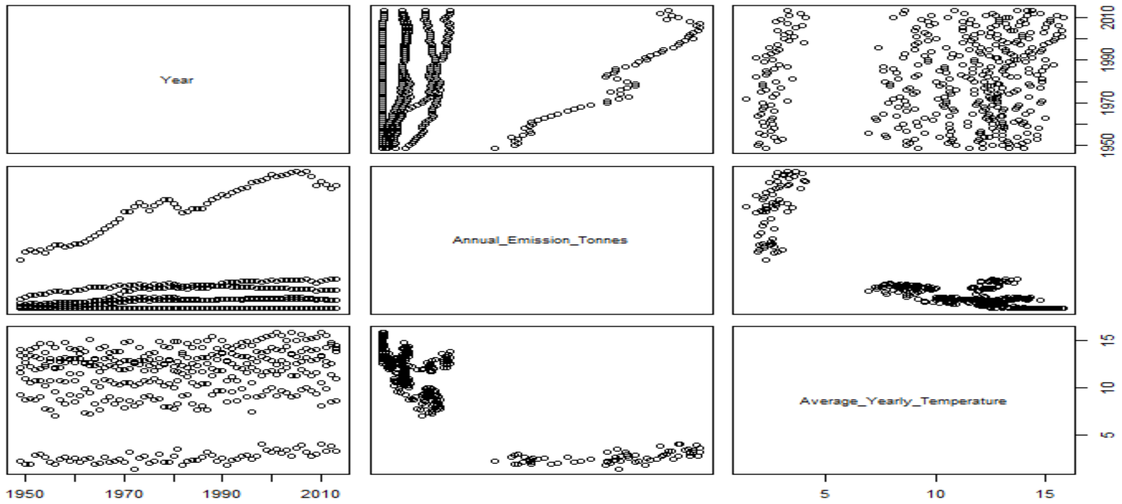


The average temperature for North America rising year on year as seen in above figure. Still the same rate of growth in the temperature even though the region has still got a high emission

It has to be noted that from the above analysis it is pretty clear that there is year on year growth in the average temperatures for each of the regions, with constant growth no matter of their region. Also, it clear that the temperature results affect all of the regions and not in the same way that emission affect the regions. It also has to be noted that over the past 60 years there has been an average of 1.3 degrees increase in the temperature worldwide.

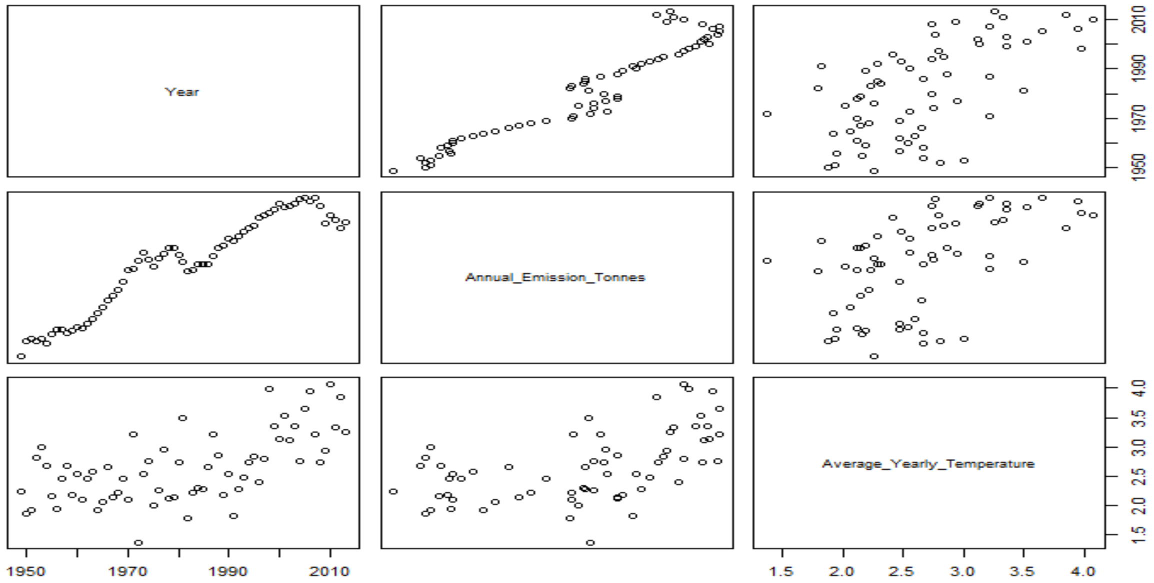
5.3 Building the prediction model.

To build the prediction model for the analysis, it was necessary to see what region has the best correlation for the analysis. As when I plotted the original emission plot() function, it was impossible to allocate what correlation can be found within the dataset.



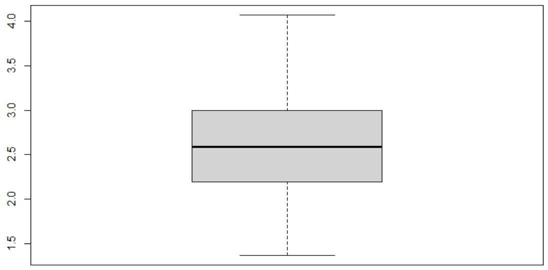
From the above plot it is hard to tell the correlation, but some correlation can be noted between "Annual_Emission_Tonnes" "Year" and "Average_Yearly_Temperature". For my model build I choose to select the region of "North America" as it has more accurate data representation and can be better analysed for the purpose of the build.

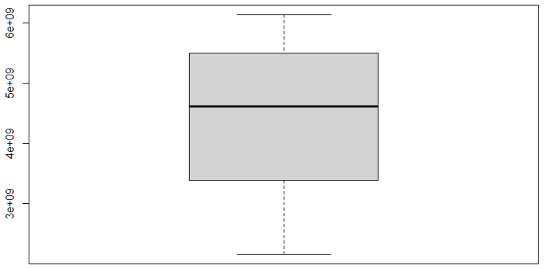
When selecting the North America region, you can clearly see better correlation with the variables in the dataset. With positive skewness.



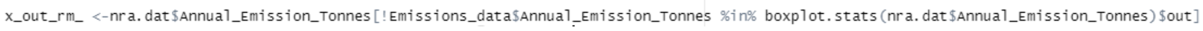
Looking at the plot from the North American region the correlation can be more clearly see between the different attributes. We will build a prediction model based on the North America region.

Before the build of the model, we have to check for data distribution within the set so that our calculation can be more accurate, this can be done by checking the frequency using histograms and box plots as shown below:



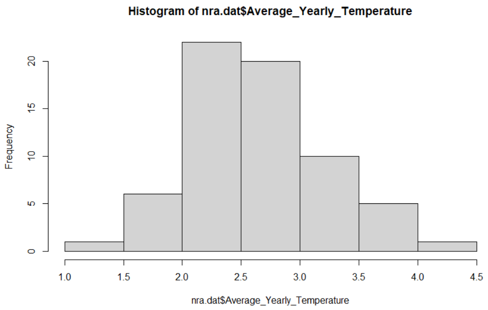


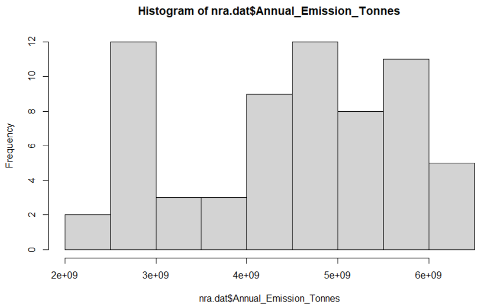
The boxplots for the Average Annual temperature and Annual emissions can be shown on the box plot above. There were a few outliners, but they were removed using the in box. plotstats function.



```
x_out_rm_ <-nra.dat$Annual_Emission_Tonnes[!Emissions_data$Annual_Emission_Tonnes %in% boxplot.stats(nra.dat$Annual_Emission_Tonnes)$out]
```

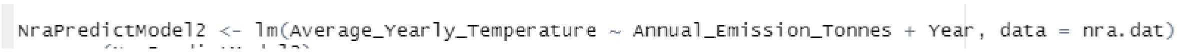
I also checked for the frequency distribution using the hist() function to see how well the data is distributed within the datasets.





The data spread is some what equally spread with the Emissions a bit slighter skewed to left which can cause a slight issue with the prediction model.

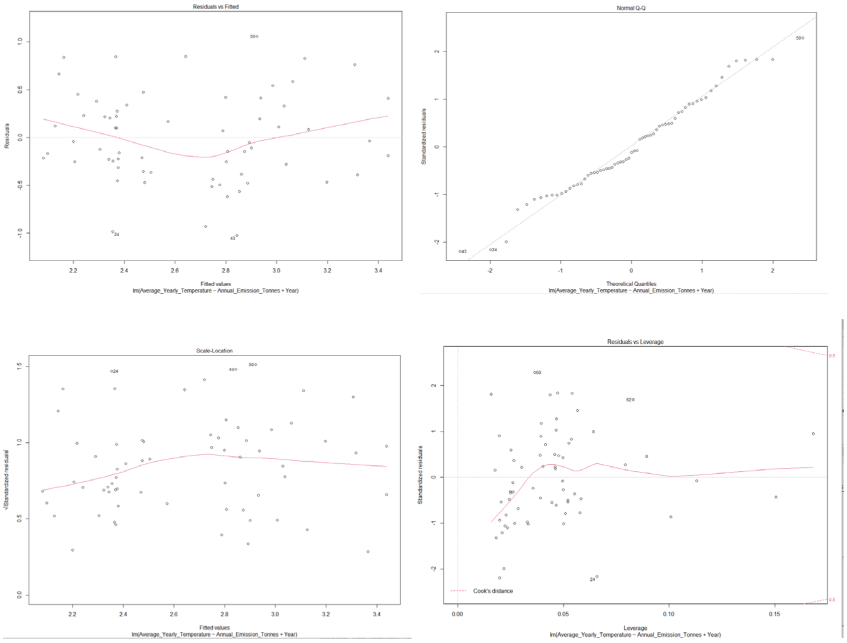
I then started on building the prediction model using multilinearity process.



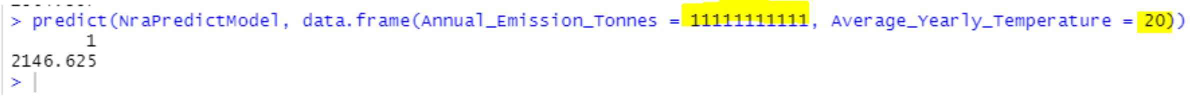
```
NraPredictModel2 <- lm(Average_Yearly_Temperature ~ Annual_Emission_Tonnes + Year, data = nra.dat)
```

Figure above: Building the prediction model using the multilinearity process.

After building the model from the North American dataset, I was able to generate the QQ plots and other residuals for the analysis.



After plotting the residuals, we then build our prediction model accordingly to the North America model of data. This is done by the following code: and the following output is given:



```
> predict(NraPredictModel, data.frame(Annual_Emission_Tonnes = 11111111111, Average_Yearly_Temperature = 20))
       1
2146.625
> |
```

When you put the predicted temperature and emission into our calculation you are able to receive a prediction of the year that type of conditions would be in North America, so according to the model with average temperature being 20 degrees and with the input of emissions the model predicts that it would be somewhere around year 2146.

This model is not as reliable as more data is needed to conduct a proper analysis, but it gives an idea and understanding of the correlation between the different Emissions and growth noted.

6.0 Conclusions

From the gathered results and the analysis, we can clearly see that there is clear increase in the increase of the overall worldwide temperatures and the increase of the emissions. Also, the emission from fuel type have skyrocketed. It is also having to be noted that there is a lot of emissions been pumped into the air than ever before, with double of the amount of CO2 and CHG emission gas released into the atmosphere. You also have to take the context of the whole world in relation to the other regions investigated, North America pumps out more than Europe put together.

We can also say with definite that the temperature has increased 1.4 degrees Celsius worldwide with the emissions doubling in correlation. We can also say that according to the prediction model the temperature that is rising will double in around 100 years and there should be something done to slow down this rate of heating that our planet is experiencing. According to the prediction model that was built, it was clear to see that the current rate of emission rate for North America will lead to the average temperature going from 8 degrees to 20 degrees in just about 120 years. This is drastic and should be address immediately. It is also having to be noted that as per my research it is clear that the liquid fuels make the most impact in the emissions and account for a big portion of the gas emission.

It is essential to cut out all of the fuels and that will drastically reduce the impact. It is also having to be noted that Europe seems to be moving on the downward trend for emission reduction while the opposite can be said for the North America region which is included of Canada and USA, they are growing exponentially with the emission release.

Correlation shows that there is a correlation between the “Year” , “Annual Emissions” and “Annual Average Temperatures”. Multiple Linear regression analysis showed the relations between Annual Emissions and Annual Average Temperatures. This report can be used to look at the information of effects of gas emissions on regions and the correlation between the emissions and emissions type of fuels and the rise in global temperatures.

7.0 Further Development or Research

With more time and resources available I would have spent more time on developing better prediction models and it was very hard to find a correct dataset that works better, as it is a very hot topic currently so all the datasets with more accurate information are hidden behind paywalls, it was hard to structure the datasets and analysis. The project title and objectives had to change accordingly to my time and abilities to do the necessary research.

The project aim for further developments would be to build some sort of GUI for prediction of the models. It would be interesting to see an application that is able to track and keep record of daily records of emissions and temperature changes and visualisation application would be very beneficial. If the same progress continues the effects could be devastating, but due to the lack of data this is a hard area to look into, if I was to pick a different area of analysis I would, as I found the lack of datasets extremely hard, also most of the work was spent on the preparation of the actual data, working out the different variables and building a new dataset from available data. Other aspects need to be looked at in determining the correct reason of the temperature rise, such as deforestation rates and etc other third-party effects. Unfortunately, I have also lost a bit of time due to technical and private issues which I wish I could get back to spend more time on research. The overall conclusion of the report is the fact that there is a huge increase in fuel consumption, and we can see that there is a correlation between the temperature rise and the emissions due to this increase.

8.0 References

Bibliography

F., B. A., 1990. *The response of Natural ecosystems to the rising global CO2 levels,* Cambridge: Department of Organismic and Evolutionary Biology.

Boden, T.A., G. Marland, and R.J. Andres. 2013. Global, Regional, and National Fossil-Fuel CO2 Emissions. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, Tenn., U.S.A. doi 10.3334/CDIAC/00001_V2013

RICE, D., 2018. *Emissions of carbon dioxide into Earth's atmosphere reach record high,* s.l.: USA TODAY NEWS.

Monitoring Online Tests through Data Visualization - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/The-Steps-of-a-KDD-process_fig7_220073492 [accessed May, 2022]

Climate Change: Earth Surface Temperature Data – Datasets – available from : https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data?select=GlobalTemperatures.csv [accessed May, 2022]

Daily Temperature of Major City – Datasets – available from:

https://www.kaggle.com/datasets/sudalairajkumar/daily-temperature-of-major-cities [accessed May, 2022]

Daily Temperature of Major City – Datasets – available from:

https://www.kaggle.com/datasets/yoannboyere/co2-ghg-emissionsdata [accessed May, 2022]

9.0 Appendices

This part contains the monthly journals and the project proposal.

9.1. Project Proposal

National College of Ireland

Project Proposal

Effects of Co2 on Climate

November 2021

BSc (Honours) in Computing

Data Analytics

2021/2022

Nikita Olijniks

16416304

x16416304@student.ncirl.ie

Contents

Objectives

The Aim of this project is to analyse and compare the effect of the Co2 in our air and the effect it can create on the patterns of weather, in respect to temperature and air pollution. I would like to also investigate and determine the current trends and occurrences in the world which are an affecting factor on the increase of the temperature and compare the Co2 current levels and trends versus the past trends.

I would like to compare the current attributes of Co2 in the air, temperature and determine whether Co2 has an increase on the temperature change, determine factors such as does not have any impact on the increase or it does, and what are the factors that are allocated with such increases if found and if not, what does not affect it.

I would also like to look at a certain period or a certain area and compare the levels of Co2 to the Ph levels in the ocean / sea waters. This will also help me to determine some further factors within the Co2 levels that influence the climate change and temperature change. Such as water quality and water temperatures. This will help me to determine whether the Co2 is the rising factors on the matter.

I would also like to investigate whether you can compare and predict the levels of Co2 in the future and compare it how much of an increase of temperature it can potentially predict, and maybe determine and established the factors that can help to decrease further levels of the Co2 or the increase in general water pollution, air pollution which are all contributing factors of temperature rise.

The overall aim is as follows;

To find adequate datasets which are relevant in Co2 and weather change.

Clean and extract the needed data form the gathered data.

Combine similar datasets and make one dataset from which information can be easily retracted from. To begin analysing the data, fulfilling the objectives and complete all the documentation, results, and visualisations in a comprehendible manner.

Background

I choose this project as I had a personal interest in the matter from a young age, I also believe and hope that some information might become useful, and currently the current trends politically globally, are moving "Greener" and there is an emphasis on things such as climate change and moving further away products that can produce too much of Co2 imprint on our environment.

There are a wide range of different aims and objectives for this project which I have set out and hope to achieve. The main aim is to analyse and breakdown the data into a form of information where we can find the unusual or abnormal outliners and maybe some new form of information which can help us prevent further damage to our environment.

State of the Art

Currently there is a lot of data gathering has been done on the matter of climate change and environmental issues, since the early 2000 a lot of focus has been done on plastic and the effects of plastic on the rise and risks of temperature increase. Global warming is currently a trend topic, and a lot of people are looking into the sourcing of information on the matter, to gather trends and make prediction models for better understanding of Global warming and, so I thought I would like to develop a better understanding on the case and maybe determine some new information which can be an important factor in all of this.

In this project I aim to analyse different data sets which can allow me to get a better understanding of the area, A lot of similar work that I have looked at don't seem to focus on the changes within the water, oceanic and sea levels, and conditions, but rather focus on confirming or denying the changes in temperature in the air and ozone areas.

I will try to bring an alternative view on the data and extract potential trends within the data. During my analyse I hope to bring I will give a different look on Global warming and compare it to the Co2 levels, I believe there is a lot of relevance between the two factors and could be a good area to investigate and will make my project a bit more challenging compared to other similar projects which I have looked at.

Data

The data that is required for the analyse, that I would like to acquire has to be in relevance to the conditions which I want to locate. In other words, the data that I must acquire for this project has to have inputs of Co2 in the air and sea level. I would also like to acquire some data with regards to the changes of temperature in the air if applicable and make a comparison with the other data sets. I hope to locate a combination of such data together and clean and extract the necessary data.

Data cleaning strategy must be implemented before any analyse should be done on the datasets. I will determine the missing values, replacing the missing values with necessary changes if applicable. I will investigate if data sampling is an easier choice, depending on the size of the data that I will have for the analyse, this will help me decrease the chances of repetitive data elements. I will look and investigate whether some data will need to be separated into separate subsets of datasets for better accurate results.

The datasets that I have been looking at to analyse is publicly available and they come in different formats such as xml and csv. I have found all the data sets publicly available online to use and they can be accessed without any extra passcodes or passes, apart from a google account login.

The data sets that I have been looking at, focuses mostly on a specific area of interest and I will need to pull the data from multiple sources and compile into one presentable visualised information.

For this I hope to include data mining process which can help me with the information gathering.

Some of the datasets that I am interested in using;

https://www.ncdc.noaa.gov/cdo-web/datasets (global climate datasets)

https://www.che-project.eu/data-portal (co2 emissions portal full of datasets)

https://datahub.io/core/co2-fossil-global (global fossil fuels usage dataset)

https://data.gov.ie/dataset/coastal-water-quality?package_type=dataset (Dataset which is looking at coastal water quality in Ireland)

https://data.gov.ie/dataset/greenhouse-gas-emissions-projections?package_type=dataset (Data set which is looking at projected emissions)

Methodology & Analysis

For my methodology I will be using the KDD or Knowledge Discovery Databases methodology. KDD methodology refers to the process which involves the finding of information and knowledge in a data set. The reason for my selection of this data set is because I have previously had experience in applying this type of methodology for a project and found it very suitable and approachable.

By using the KDD approach, I will be able to divide the workload into separate KDD stages which will make the work more effective and approachable. The following approach I will try to take in order to do the analysis;

Selection – I will investigate the selection of the data sets which are more suitable to the objective of my project, such as climate change rate data, air quality data, sea data etc. Understand the end goals of that I want to achieve out of this project and create a target data set i.e., focusing on a subset of data or data samples

Pre-Processing – I will clean the data and replace the missing values and remove potential outliers and noise in the data. I will also need to develop a strategy on how to deal with the missing values. I will then select the data that I require to conduct my research clearly.

Transformation – I will try and conduct data transformation with the transformation method and using dimensionality reduction method.

Data Mining – The data mining process will involve the search for different patterns and trends which are in interest that I can locate, depending on the mining goal which I will develop thru out my research. I will then develop the methods that I can use to search for these patterns using data mining algorithms. This will involve the process of deciding models and appropriate parameters to research for accurate projection.

Interpretation / Evaluation – This stage will involve the evaluation of the of the mined results and determining the findings. The findings will help us evaluate what could be locate, what patterns did we locate, what could we learn from the patterns and projections.

Technical Details

The technical developments which I am looking to carry out in this project varies from the stages of the process. I am hoping to use web scraping, using the python programming languages to help me with the query of receiving some data from the web. Or using web APIs to obtain a particular data set.

I have also been interested in using a range of algorithms to help me analyse the data more effectively and efficiently. The algorithms which I have investigated are “K-means” and “Apriori Algorithm”.

K-means algorithm help to cluster the analysis of data into a subset of clustered items together.

Apriori Algorithm is the algorithm which helps you to locate and show frequent item sets and help to highlight trends within a dataset.

I will use IBM SPSS to create potential predictive models and conduct other necessary analytical tasks needed. I will also use Excel spreadsheet tool with built in statistical and data visualisation functions.

As mentioned, before I will use Python programming languages to help me create this analysis of the datasets, I will also use R in R studio to help me create visual representations of the data located. My technological developments may vary with the development of my project as I found more potential ways of finding potential.

I want to implement POWER BI for the process also, as it's a new technology which I am eager to learn more about.

Project Plan

My project plan has the following structure which I will plan to stick to: The diagram below is broken down into monthly phases starting from October 1st.

Research phase

Research phase is made up from 3 separate stages – Initial research, dataset research and Midpoint Presentation.

The Initial research phase is conducted from the month of October to November, where I develop my first idea for the project and plan out the structure and layout for my project.

From the month of November to December I will conduct a research phase into the data research for the project. I will find required datasets required to conduct the necessary research into this analysis.

Data Gathering

Data Gathering phase is the longest phase planned for the project, it involves different methods of gathering the necessary data from the datasets using various technological methods. This phase will run from December to March. This process is ever changing as more data could be necessary to be located at any point during the research.

Data Cleaning

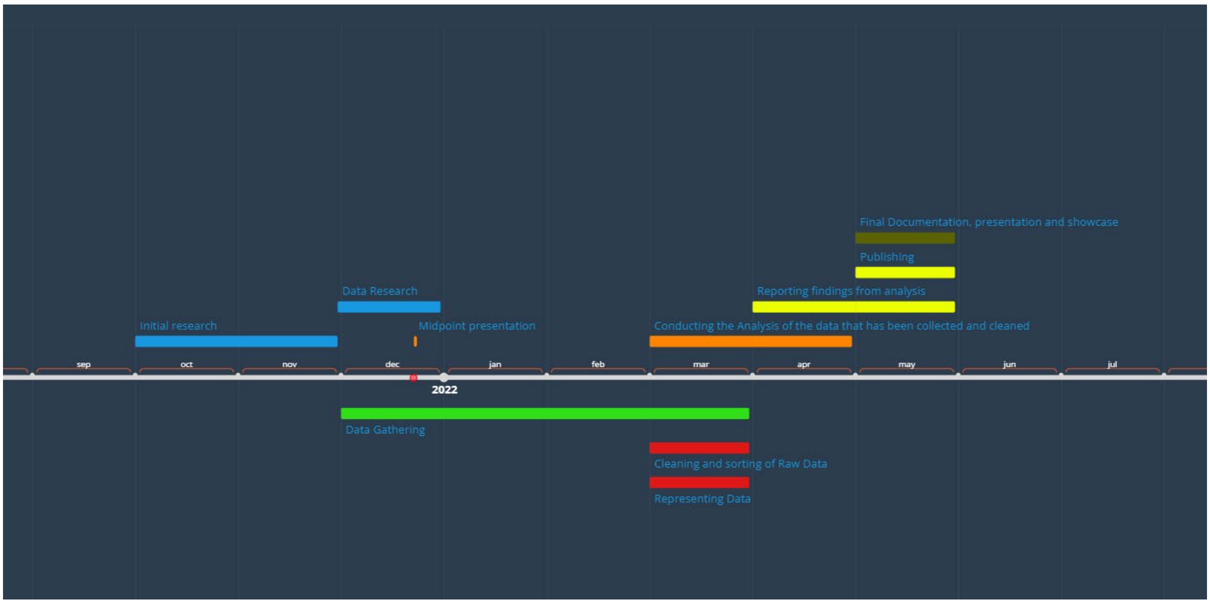
Data Cleaning phase will consist of 2 task "Cleaning and sorting" the data and "representing" the data. This phase will be used to clean the gathered data and represent the data in a valuable way so that it could be clearly understood. This phase will run through the month of March for both tasks.

Analysis

The Analysis phase consists of the task "Conducting Analysis of all data that's been gathered". This phase will include the analysing and extracting valuable information from the data gathered. This task is planned to run from the month of March to April.

Reporting

Reporting phase includes three tasks to be completed, the "reporting findings from the analysis" "publishing" and "final documentation, presentation and showcase". This phase will be the final phase of the project and will be conducted from April to the end of May.



9.2. Ethics Approval Application (only if required)

9.3. Reflective Journals

Supervision & Reflection Template

Student Name	Nikita Olijniks
Student Number	16416304
Course	BSCH in Computing

Month: March

What? Reflect on what has happened in your project this month? Worked further on the report and conducting the analysis needed for the completion of the report. Worked on conducting the necessary machine for the analysis and what is more applicable, worked on deciding on what some of representation I would use for the visual part of the reporting. Worked further on the report to structure the analysis better, and formulate the report accordingly to the objective at hand
So What? Consider what that meant for your project progress. What were your successes? What challenges still remain? The successes of this month were the progress in the reporting of the analysis and conducting more analysis on the report. Also seeing the structure of the report coming out to somewhat a finished stage allows for a better overview of what is need further to conduct the report.
Now What? What can you do to address outstanding challenges? Spend more time on working on the report to finalise for the next month and last month, work further on the analysis of the report and change the report structure to better fit the needed objectives. Add more analysis onto the report for further analysis.

Student Signature	

Supervision & Reflection Template

Student Name	Nikita Olijniks
Student Number	16416304
Course	BSCH in Computing

Month: April

What? Reflect on what has happened in your project this month? Worked further on conducting the necessary analysis needed for the report, work on more strategies for the analysis that's needed for the completion of the report. Worked further on the report to structure the analysis better and more coherent. In general, spent the time to complete the report as the submission of the final project is within 2 weeks. Worked further on the report and conducting the analysis needed for the completion of the report.
So What? Consider what that meant for your project progress. What were your successes? What challenges still remain? The success that was in this month was the further progress on completion of the report and finalisation of the report, continued to work on the report with more time to spend on the report. Also seeing the structure of the report coming out to somewhat a finished stage allows for a better overview of what is need further to finish the report which is what still remains till the submission date, need to work further on the necessary submission requirements.
Now What? What can you do to address outstanding challenges? Work further to finish the report before submission, to review the report and finalise the documentation need for the submission. Work on creating the necessary requirements for the submissions needed in the next 2 weeks.

Student Signature	

Supervision & Reflection Template

Student Name	Nikita Olijniks
Student Number	16416304
Course	BSCH in Computing

Month: February

What?
Reflect on what has happened in your project this month?
Worked further on analysis current data , I have also spent some more time on looking for another dataset to analyse as I was declined permission to a specific dataset that I want to use in collaboration with my previous work. I have searched further regarding the datasets that fits better to my specifications. I created more analysis with the current set with R studio and started making progress towards original objective, most of the work was done during the first 2 weeks of February due to some ongoing situation I couldn't deliver more analysis.

So What?
Consider what that meant for your project progress. What were your successes? What challenges still remain?
The goals of this month was to progress further on the project which I have done further developments on. I have also done more analysis on the current datasets which I have current access to. There was certain things that wasn't achieved due to out of my control and following the next month I will work further to create better analysis.

Now What?
What can you do to address outstanding challenges?
For the outstanding challenges there is still a lot for me to achive, I need to really deep dive on the project plane and see how can I better form a hypothesis from the analysis that I am doing. There is still plenty of small challenges of getting the project going, as it is nearing the ends soon there needs to be more work done further, I have to also make sure that all my data sets align for the objectives I need to make the project to work well.

Student Signature	

Supervision & Reflection Template

Student Name	Nikita Olijniks
Student Number	16416304
Course	BSCH in Computing

Month: JANUARY

What?

Reflect on what has happened in your project this month?

Due to exam time and holiday period I could spend much of the beginning of January on the project, fortunately to my luck I have also been ill during a period of the month which has set me back on the projected plan.. I worked more on cleaning the current data I have and currently waiting for information on more data access from US organisation.

So What?

Consider what that meant for your project progress. What were your successes? What challenges still remain?

There wasnt much for success this month, I would defiantly work for a lot more progress for the next month analysis wise on the data. I also would like to gather more efficient data which I can use for my analysis and development of my project. This has been my main challenge during this time. Also the challenge of time management has been tough with exam time and holiday period and personal health issues.

Now What?

What can you do to address outstanding challenges?

I want to create some sort of early analysis on the data which I have gathered and start conducting my full research further.

For my outstanding challenges I will need continue to research and find more appropriate data that I will need to conduct my analysis on. Continue working on the project template , and create a better timed schedule. Be on top of my monthly journal submission.

Student Signature	

Supervision & Reflection Template

Student Name	Nikita Olijniks
Student Number	16416304
Course	BSCH in Computing

Month: DECEMBER

What?
Reflect on what has happened in your project this month?

For this month I have completed my midpoint presentation and submitted the draft of my project booklet. I have also created the midpoint presentation video and presented what I have done so far in the past two months. I have spent time working on gathering more data, I have also applied for certain data access but unfortunately was denied request to as it was pay for access.

I worked more on developing structure for my project and project template.

So What?
Consider what that meant for your project progress. What were your successes? What challenges still remain?
The success for this month would be reaching the midpoint presentation and achieving a certain level within the project. I was able to complete due to a heavy timetable with submission on time and it was a real challenge to balance college time and personal time and also between spending time on the project. Gathering decent data has been also a challenge, but I have started looking at ways of starting analysis on the current data sets which I have collected.

Now What?
What can you do to address outstanding challenges?
For my outstanding challenges I will need continue to research and find more appropriate data that I will need to conduct my analysis on. I want to spend more time on the project for the next semester and start doing appropriate analysis that need to be completed. I will also continue to clean the data and create a structure way to create appropriate analysis for my report.

Student Signature	

Supervision & Reflection Template

Student Name	Nikita Olijniks
Student Number	16416304
Course	BSCH in Computing

Month: November

What?

Reflect on what has happened in your project this month?

For this month I have spent time on working on the project midpoint presentation and spent more time researching the appropriate dataset for the project. I worked on setting the objectives for the project and goals of analysis which I can conclude. I have done more research on appropriate data needed for the project to conclude appropriate research.

So What?

Consider what that meant for your project progress. What were your successes? What challenges still remain?

I was able to establish a draft of objectives needed for achieving the goals. I have also set goals for the project which I so far followed regularly. The main challenges that still remain would be the finding of a good dataset for my project and working on my midpoint project presentation which will be my focus for my next month goal and plan. Main challenge is to find appropriate data to use in the project for the months ahead.

Now What?

What can you do to address outstanding challenges?

For my outstanding challenges I will need to research and find more appropriate data that I will need to conduct my analysis on. Research and finalise on the outcome of the project which I would like to achieve. Create and finish the midpoint presentation .

Student Signature	

9.4. Other materials used

Any other reference material used in the project for example evaluation surveys etc.