

**ANALYSIS ON THE IMPACT OF CARBON DIOXIDE**

**&**

**OTHER GREENHOUSE GASES ON SEA-LEVEL AND GLACIER DENSITY**

Presented By:

PRATIBHA SRIDHAR IYENGAR

SIDDHANT SANJIV SAPTE

SARAN TEJAS KAJA

NORTHEASTERN UNIVERSITY

ALY 6040-DATA MINING

CRN: 21289

INSTRUCTOR NAME: AMIN KARIMPOUR

Date- 2/16/2019

**INTRODUCTION**

Climate change is one of the major concerns nowadays and there are various factors which are influencing it, and one of the major factor which is causing climate change is Air degradation due to emissions from different sources. We can already witness the effects of this through temperature rise due to global warming , rise in sea level and climate change. As serious as this issue of environmental degradation is, it is also just as fascinating to analyze how it also impacts social and economic sustainability.

We have utilized climate change related datasets obtained from various sources that provide information about greenhouse gases emissions, carbon dioxide emissions per $1 GDP (PPP, Purchasing Power Parity), sea level data and glacier density recordings over the period.

For the analysis, we cleaned and handled the missing data using packages such as tidyr and dplyr. Then analyzed and visualized the data by plotting them to find out the trends and insights from the plots. We plotted co2 emission with respect to other gases, continent wise, increase in sea level and decrease in glacier density with increase in carbon levels over period of time, we did clustering based on GDP using K means and used timeseries to forecast sea level and density of ice caps over a period of 10 years.

With information rich datasets, increasing phenomena of climate change in different parts of the world and the rising importance of taking preventive measures to curb the harm our man-made inventions are causing we have chosen this problem as our center for data mining and come up with interesting insights through various graphs and plots.

**DATA GATHERING AND CLEANING**

The datasets were collected from UN Data Report and EU Data Report along with other official online repositories that provide open source climate change data over a period of time.   
The data archive used in this project dates between 1970-2012. For the purpose of consistency and data integrity, we have analyzed for the period between years 1990-2012.

All the Datasets used are listed as follows:

* Greenhouse Gases Emission (in Gigagrams)
* Carbon Dioxide Emissions (in Gigagrams)
* Hydrofluorocarbon Emission (in Gigagrams)
* Methane Emission (in Gigagrams)
* Nitrous Oxide Emission (in Gigagrams)
* Perfluorocarbon Emission (in Gigagrams)
* Sulfur Hexafluoride Emission (in Gigagrams)
* Carbon Dioxide Emission, kg per 1$ GDP (PPP)
* Country-wise Sources of Carbon Dioxide Emission
* Country-wise GDP Data
* Glacier Density
* Sea-Level Mean Value

All the datasets provide values of emissions of 40 countries around the world over a period of 22 years or more.

We had to handle NULL values and data consistency issues with respect to country names and periodic data availability using functions of R Packages like tidyr and dplyr.

**ANALYSIS**

For the analysis of this data, we have grouped chalked out a wireframe before finally predicting the impact of emissions on sea level and glacier density.

**We started off with the descriptive analysis of the emissions over a period of time around the world:**

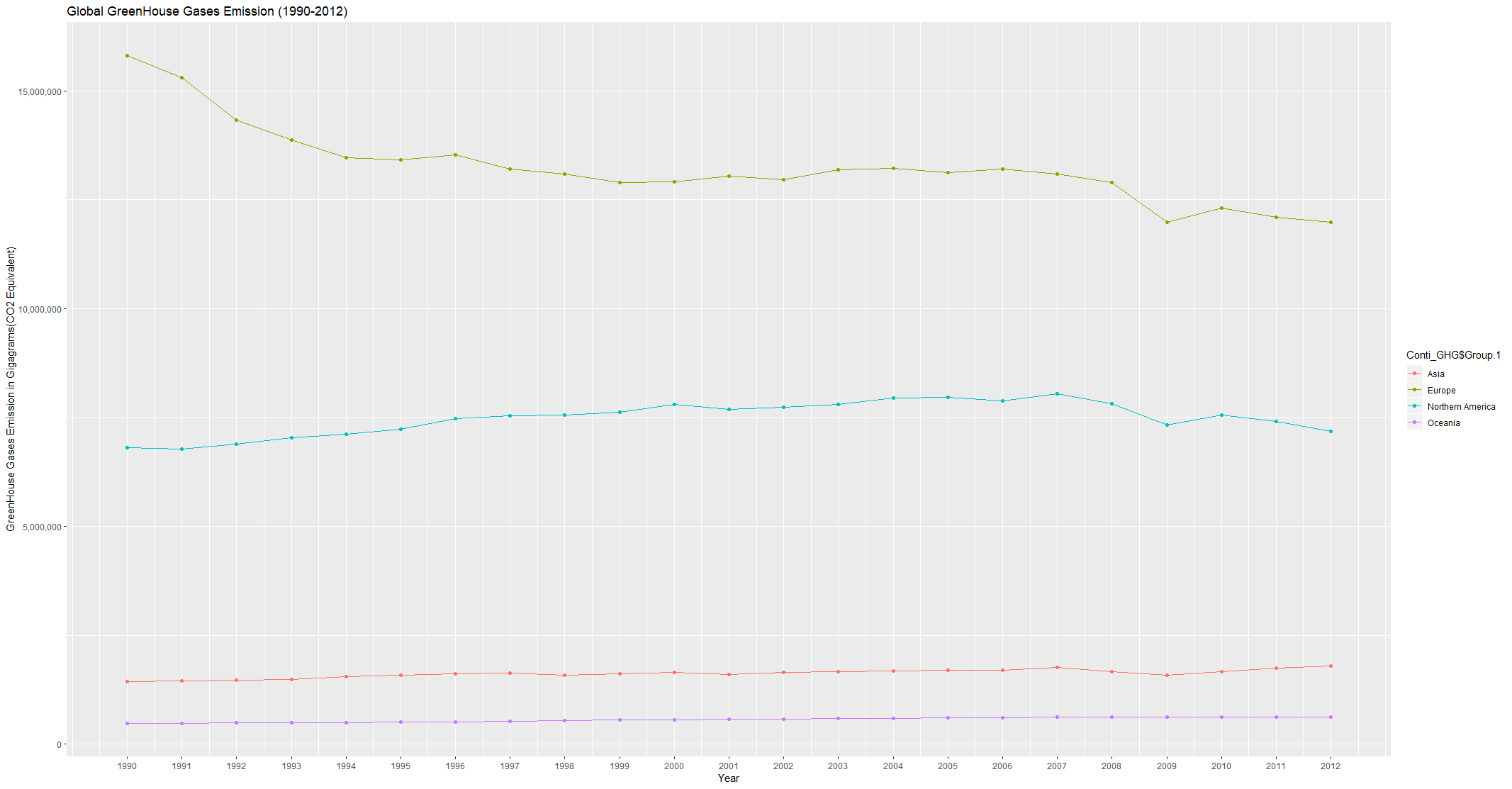
* **Greenhouse Gas Emission across continents between 1990-2012**

Fig.1.Greenhouse gas emission continent wise in gigagrams with respect to year

For this chart, Greenhouse Gas Emission (in Gigagrams) dataset was used.   
The variables provided were Country, Year and Value of emission. Using the countrycode package available in R, we mapped the continents with the respective countries .

Then we aggregated the emission values for each year and grouped them according to the continents. This newly grouped data was used for multiple series line chart.

From the chart we observed that Europe has the maximum emission trend followed by Northern America. This seems to be an exact representation of the reality and history as industrial revolution began in Europe whereas United States is among the most powerful nations in the world economically and industrially.

* **Component wise Greenhouse Gas Emission**

Greenhouse gases are a group of gases that contribute to the global warming due to excessive emissions. Gases like Carbon Dioxide, Methane, Nitrous Oxide, Hydrofluorocarbons and few others are collectively called the greenhouse gases.

It is important to understand in what amounts these gases are being emitted. Therefore, the component-wise emission multiple series line chart was plotted and analyzed.

For the plotting, we gathered data for each gas with Country, Year and Emission Value variables. Since the analysis was for major contributor, we first aggregated the average emission value for each country over 22 years and then mapped each country with their corresponding continents. Post which, the means were added together to be grouped by continents.

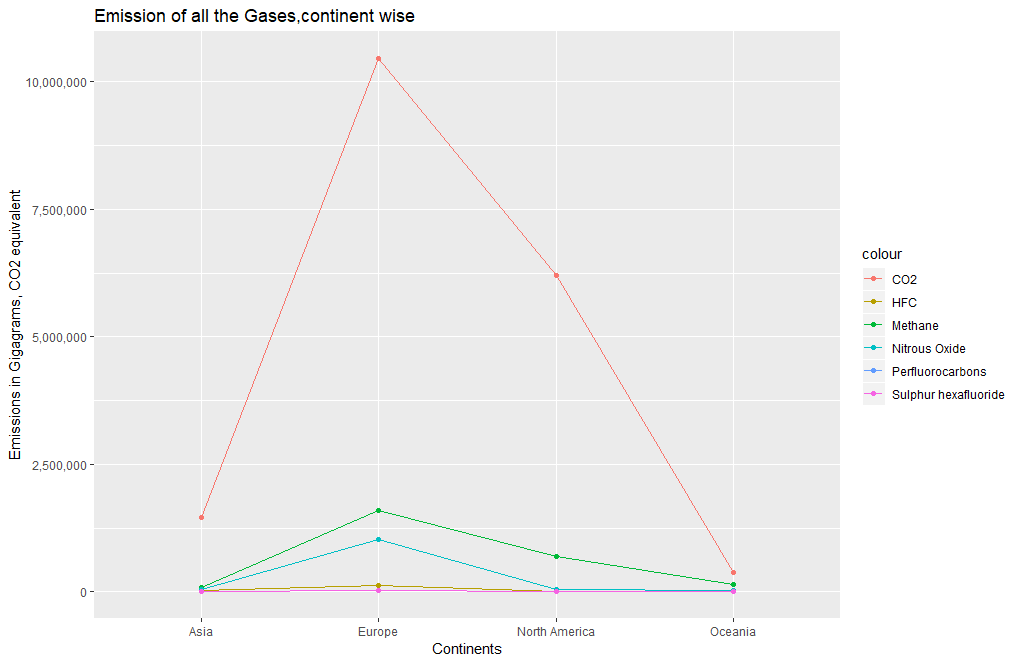


Fig.2.Emission of all the gases in gigagrams with respect to continents

From the chart it is evident that Carbon Dioxide emission has occurred the most and staying in line with the previous chart, it is prevalent in Europe and North America.

**From the descriptive analysis, we can observe that Carbon Dioxide emission has been increasing which is really hazardous to the environment, keeping this in mind, we drilled-down further to see the composition of carbon dioxide in greenhouse gases, the sources of emissions and the per dollar of GDP:**

* **Composition of Carbon Dioxide in Greenhouse Gas Emission (%)**

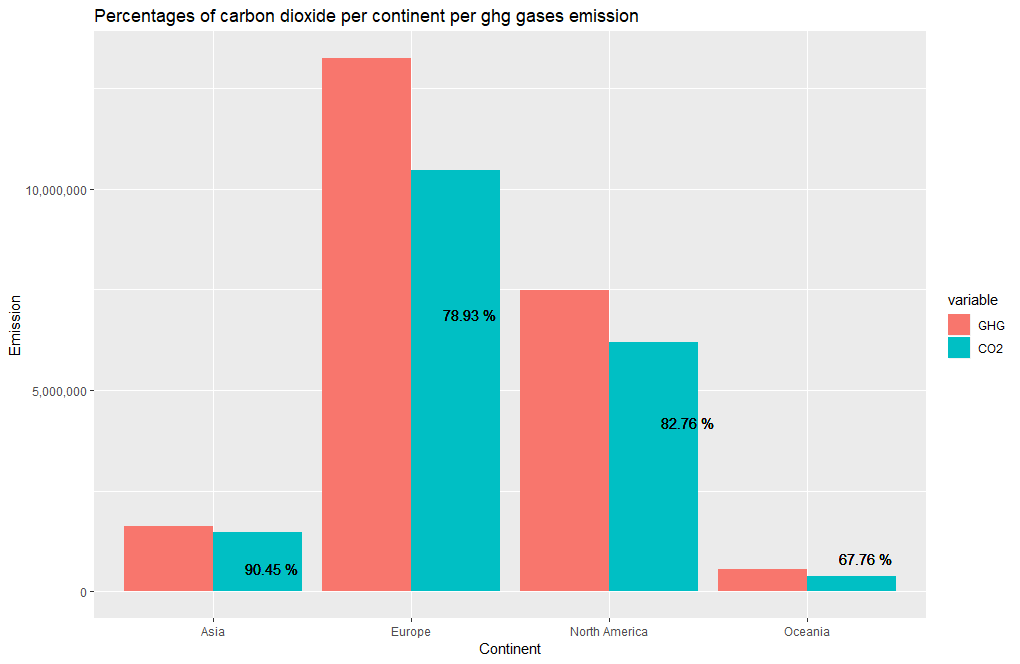


Fig.3.Percentage of carbon dioxide per continent with respect to all the greenhouse gases

For this side-by-side bar plot, the average greenhouse gases emission value is calculated across continents. Using melt(), ggplot() and geom\_bar() with stat= “identity” and position= “dodge”, the bar plot was created with Carbon dioxide emission.

From the above figure, we can see that carbon dioxide emissions contributes to approximately 70-90% of the greenhouse gases emissions. These are pretty high levels when considering the fact that the emission value excludes natural causes.

* **Sources of carbon dioxide emission**

So what is cause such high levels of emissions? The answer lies in the sources of emissions.

Since industrialization and rapid development in manufacturing and technology , there has been a lot of pollution through the production units, factories where the main source of energy generation is through burning fuels.

Major sources of carbon dioxide emissions considered in our analysis are-

* + Fossil fuels and cement production
  + Solid fuel consumption
  + Liquid fuel consumption
  + Gas fuel consumption
  + Gas flare

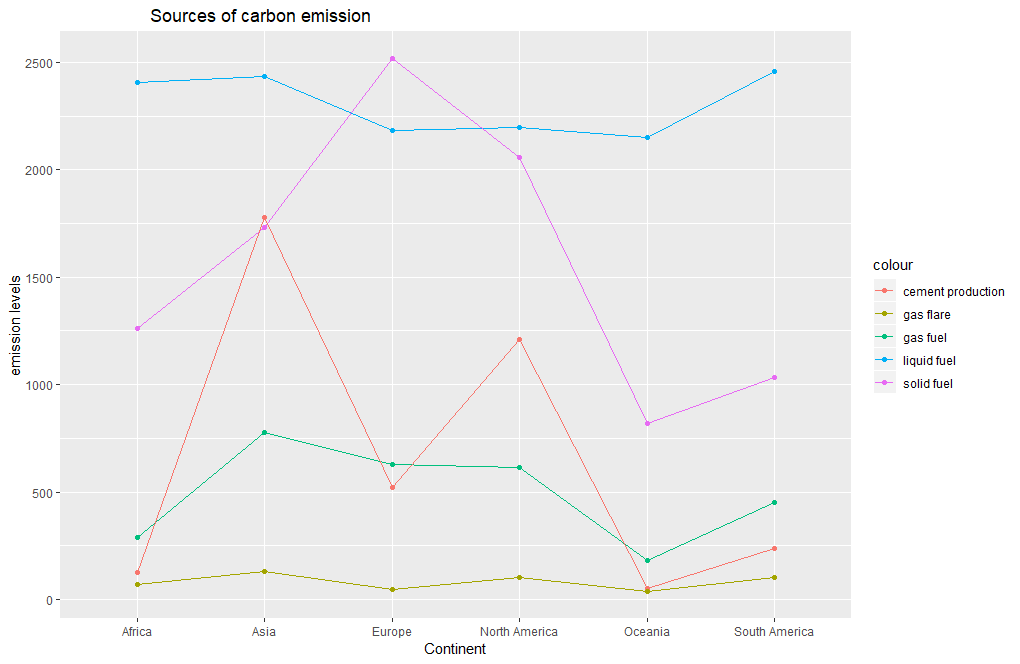


Fig.5.Sources of carbon emission with respect to continent

We can see that liquid fuel consumption is the highest in order and gas flares are least across all the continents with an exception in Europe where solid fuel consumption is leading.

This means Europe consumes more solid fuels like coal and wood pellets than other continents.

* **Carbon Dioxide Emission per 1$ GDP (PPP)**

Purchasing Power Parity is the purchasing power of one country against another. It helps in understanding how good the GDP is of one against the other. Here, in environmental terms, it shows how much carbon dioxide a country can buy in 1$ given its current GDP. In other words, how much energy is the country consuming in terms of economy at current GDP.

The dataset obtained contains Country, Year and the kg CO2 per dollar GDP value.

For analysis, we have also obtained the yearly GDPs of countries across the world. In analysis we want to see how sustainable a country is given the GDP. What can be expected is since developed countries are lot more industrious and pro-manufacturing, the countries with higher GDP would have more PPP value.

The below figure shows the scatterplot of PPP values against the GDP of various countries.

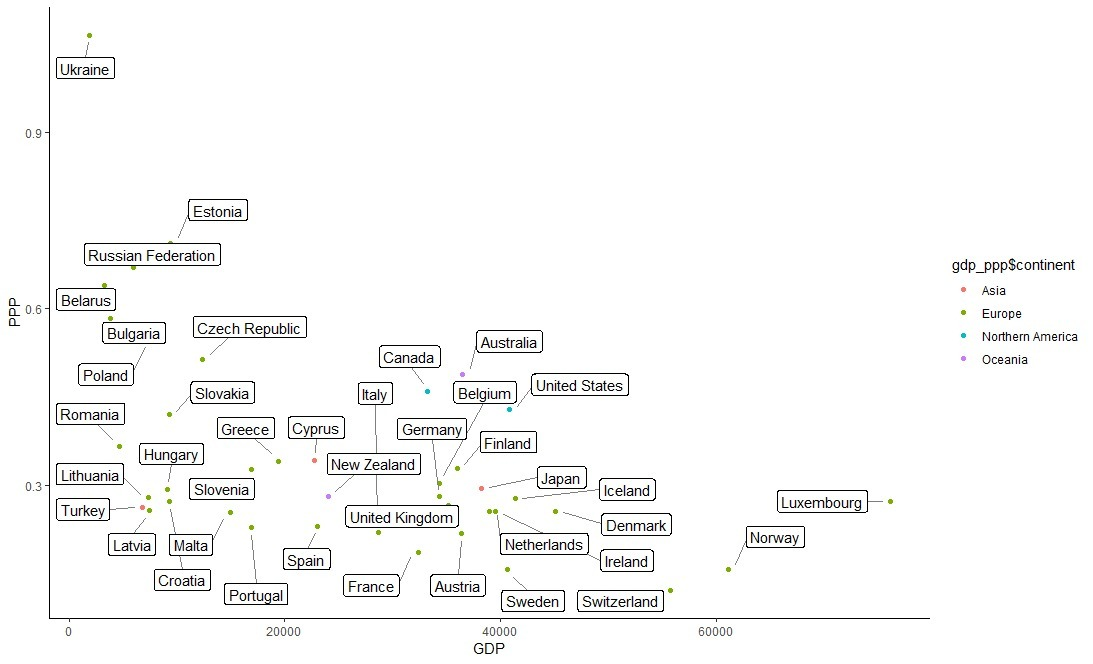


Fig.4. Carbon dioxide emission per dollar GDP

From the chart , we observe that most of the developing countries, with lower GDP rates have very high to high PPP values.

For example, Ukraine, here, as observed, has a high per Dollar GDP (PPP) value despite low GDP. This is to say that this country is consuming a lot more energy and it still doesn’t contribute to their growth.

* **Clustering of countries according to GDP based on their per Dollar carbon dioxide emission values**

For further analysis and mining of the economic impact of Carbon dioxide emissions based on their GDP, we applied K-Means clustering algorithm for clustering countries by their average GDP value into 3 clusters representing Low, Moderate and High GDP growth rate.

The K-Means is unsupervised clustering method that is useful to create subgroups of dataset that can help in categorizing and find similarities in data groups. The intra-cluster variation is minimized and through iterations of cluster centroid updates, convergence is achieved. The result of kmeans() of R library “cluster” gives the output as follows:

> km

K-means clustering with 3 clusters of sizes 3, 19, 18

Cluster means:

GDP PPP

1 64302.954 0.1835845

2 34806.514 0.2911966

3 9258.964 0.4461543

Clustering vector:

Australia Austria Belarus Belgium Bulgaria Canada Croatia

2 2 3 2 3 2 3

Cyprus Czech Republic Denmark Estonia Finland France Germany

2 3 2 3 2 2 2

Greece Hungary Iceland Ireland Italy Japan Latvia

3 3 2 2 2 2 3

Lithuania Luxembourg Malta Netherlands New Zealand Norway Poland

3 1 3 2 2 1 3

Portugal Romania Russian Federation Slovakia Slovenia Spain Sweden

3 3 3 3 3 2 2

Switzerland Turkey Ukraine United Kingdom United States

1 3 3 2 2

Within cluster sum of squares by cluster:

[1] 223119223 723770817 428322515

(between\_SS / total\_SS = 88.8 %)

Available components:

[1] "cluster" "centers" "totss" "withinss" "tot.withinss" "betweenss" "size" "iter"

[9] "ifault"

Here we can see that the cluster means of GDP are well separated and can be clearly grouped as High-Moderate-Low values.

On visualizing the cluster using fviz\_cluster() function in factoextra library, we get the following result:

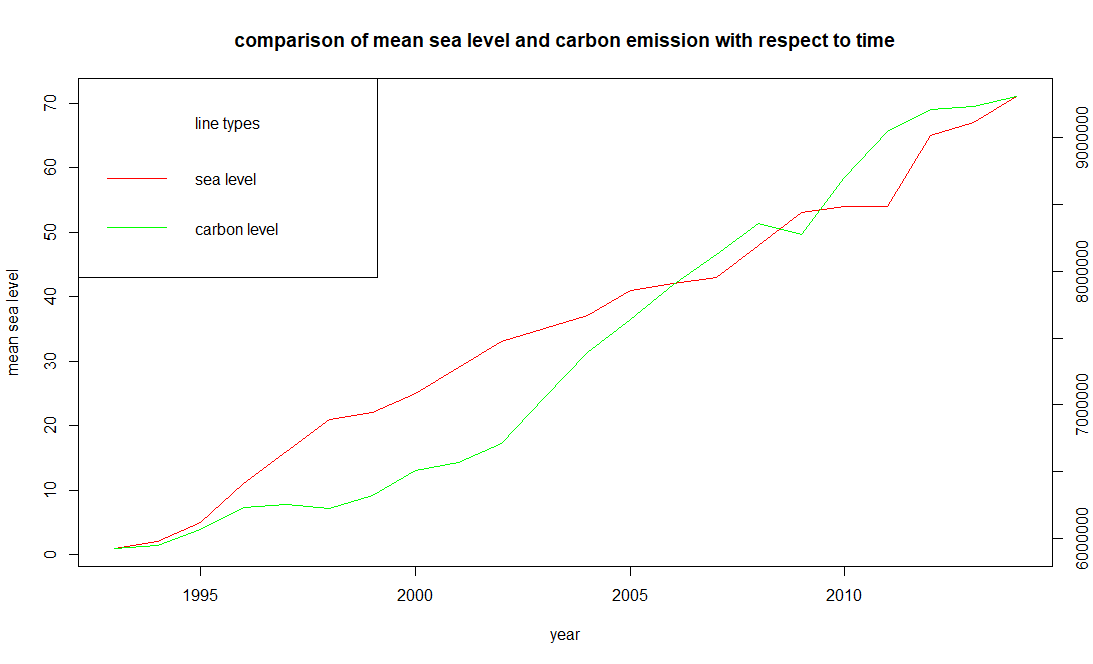
Fig. 5. Clustering Countries by GDP on Per Dollar GDP(PPP)

The fviz\_cluster() also performs PCA or Principal Component Analysis before plotting the data points. In this case though, there are only two variables hence the PCA will not have any significant impact.

From the cluster plot we can conclude that countries like Switzerland and Norway that have high GDP still emit lower levels of Carbon Dioxide; most of the developing countries and poor countries with lower GDPs seem to be real contributors to growing emission rates.

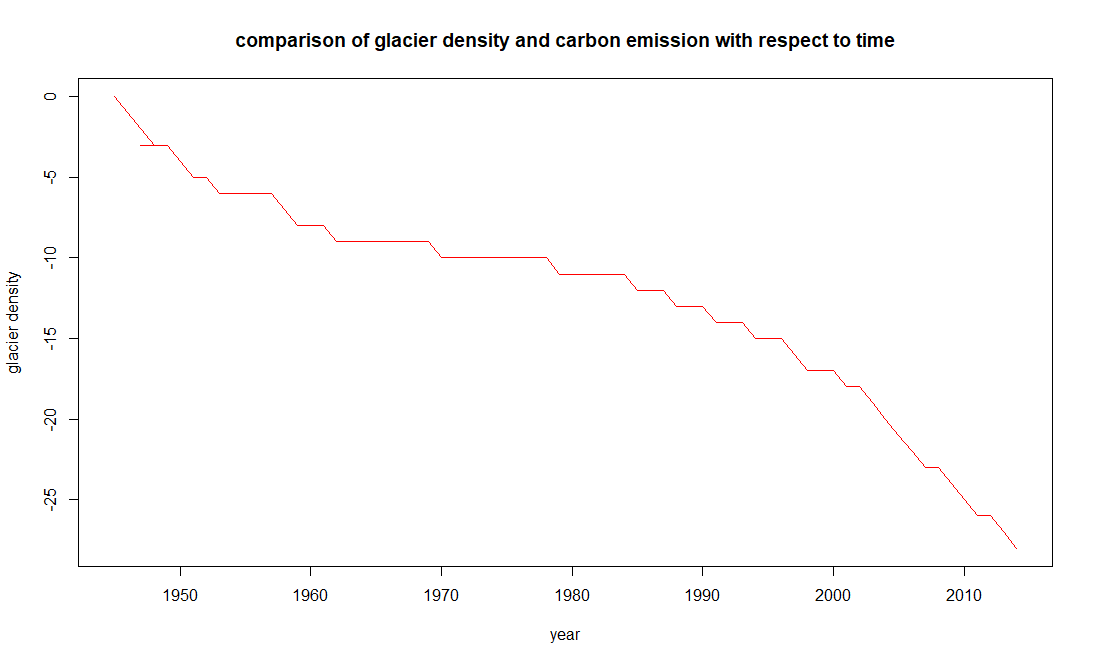
In the final phase of our analysis, we have performed time series analysis and predicted the rise in sea level and the decline in glacier density:

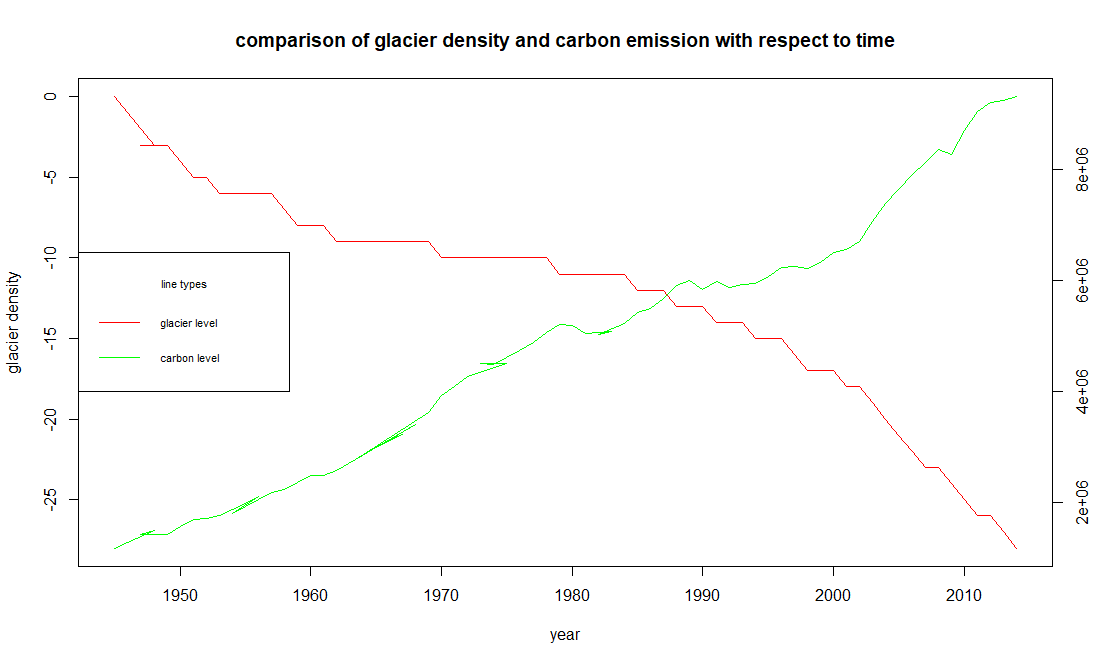
* **Trend Analysis of Carbon Dioxide emission and Sea level rising**

From the above graph shows sea level and co2 levels with respect to time period. From the above

graph we can clearly see that both the sea level and co2 level are increasing constantly and can see that with raise in co2 levels the sea levels are also increasing. For every 500000 metric ton rise in carbon level, there is 10cm rise in the mean sea level.

* **Trend Analysis of Carbon Dioxide Emission and Glacier density decline**

In the above plot we see the decrease in the density of the glacier level with respect to year which shows that the temperatures are rising year by year. If the present scenario continues there may be a chance that all the polar glaciers might melt. We can see that there is steady decrease in the density of the polar ice caps from 1990 and the slope even increased from the year 2000 which indicates that there is a constant growth in temperature in the present years.



From the above graph it is more evident that with increase in the emission of carbon dioxide the density of the polar ice caps are also decreasing and we can see that they both are inter related.By looking at the above graph we can conclude that carbon is one of the major contributor for raise in temperatures.

* **Prediction of increase in Sea Level and decrease in Glacier Density over next 10 years**

We used time series algorithm to predict the future levels of both the glacier density and mean sea level

sea<-ts(sea\_level$`Mean sea level`, start=c(2000), end=c(2014))

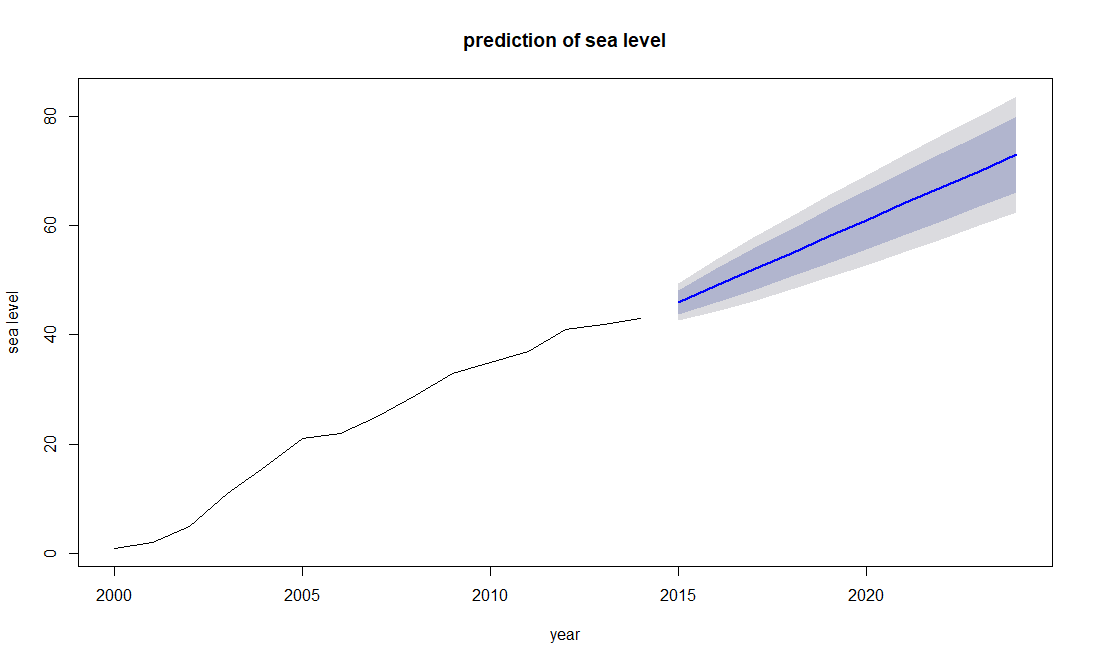
plot(sea)

fit<-auto.arima(sea)

seafor<-forecast(fit,10)

plot(seafor,ylab = "sea level",xlab="year",main="prediction of sea level")

In the above code we declared the start and end date on which a timeseries is plotted and then we used auto ARIMA for finding the optimal solution, it uses combination of different order parameters and finds the optimal solution among them increase the efficiency of the model then we use forecast and plot the obtained forecast values.

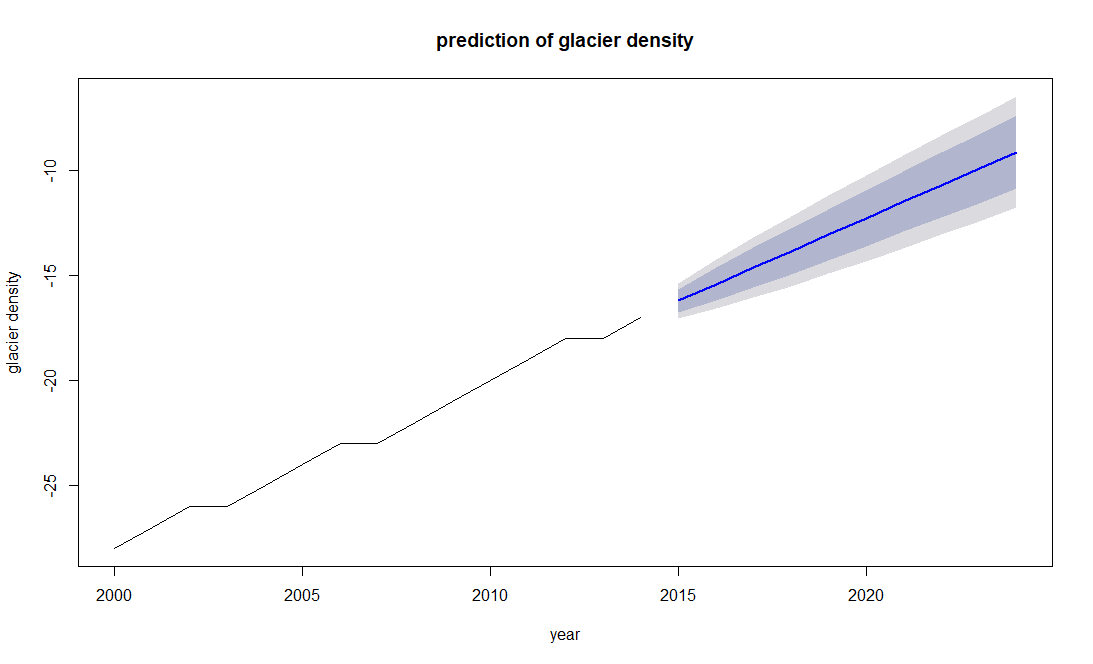


From the above graph we can see that there is an increase in the sea level for the upcoming years,we used a machine learning algorithm to predict the future based on the previous values which say the trend is continuous and there is more chance for increase in sea level from 40 to 78 with a variation between 45 to 78 which is more than the present values.

Forecast values

|  |
| --- |
| Point Forecast Lo 80 Hi 80 Lo 95 Hi 95  2015 46 43.80893 48.19107 42.64905 49.35095  2016 49 45.90136 52.09864 44.26104 53.73896  2017 52 48.20496 55.79504 46.19598 57.80402  2018 55 50.61786 59.38214 48.29810 61.70190  2019 58 53.10062 62.89938 50.50704 65.49296  2020 61 55.63300 66.36700 52.79188 69.20812  2021 64 58.20297 69.79703 55.13421 72.86579  2022 67 60.80272 73.19728 57.52208 76.47792  2023 70 63.42679 76.57321 59.94714 80.05286  2024 73 66.07123 79.92877 62.40336 83.59664 |
|  |
| |  | | --- | |  | |

From the above table we can clearly see that the sea levels are increasing steadily and there isn’t any breakeven or stop point for it, which says that by 2024 there would be many states and counties which would get drowned.



The above image is the prediction of glacier density with respect to time, here we used time series algorithm to predict future based on current values and this shows that there would be further decrease in the density of the polar ice caps and it says that there would be a steady increase in the reduction of density of ice caps.

Forecast values

Point Forecast Lo 80 Hi 80 Lo 95 Hi 95

2015 -16.214286 -16.76009 -15.668486 -17.04902 -15.379556

2016 -15.428571 -16.20045 -14.656693 -16.60906 -14.248086

2017 -14.642857 -15.58821 -13.697503 -16.08865 -13.197063

2018 -13.857143 -14.94874 -12.765542 -15.52660 -12.187684

2019 -13.071429 -14.29187 -11.850982 -14.93794 -11.204917

2020 -12.285714 -13.62265 -10.948782 -14.33038 -10.241053

2021 -11.500000 -12.94405 -10.055948 -13.70849 -9.291514

2022 -10.714286 -12.25804 -9.170530 -13.07526 -8.353315

2023 -9.928571 -11.56597 -8.291171 -12.43276 -7.424383

2024 -9.142857 -10.86883 -7.416885 -11.78250 -6.503211

In the above prediction table we can clearly see that the density of polar ice caps is decreasing in all the different aspects. By which we can say that it would decrease for sure which in turn would have effect on sea level.

**CONCLUSION**

To conclude , we get the following insights:

* Although there is a decline in the greenhouse gas emission levels of Europe over the period, it is still leading contributor to air pollution with levels ranging between 11980000 – 15808000 Gigagrams while Asia and Oceania have had pretty much consistent emission levels
* Carbon dioxide is the most emitted greenhouse gas across the continents with values as high as 10464623 Gigagrams with a significant difference in emission levels of other gases
* It also contributes to 60-90% of the greenhouse gas emission, especially 90% in Asia which has many developing countries
* Across the continents, liquid fuel consumption is the major source of emission. Liquid fuels include fossil fuels and LPG which are not only depleted resources but also produce very harmful by products on combustion
* It is also evident from the purchasing power parities that some developing countries and poorer countries with lesser GDPs have much more emission levels than developed and energy efficient countries
* The sea level is rising at an alarming rate and glacier density is declining quickly too. From the time series analysis, we have observed a steady 10 cm rise in mean sea level.
* Predicting mean sea level and glacier density for 10 years after 2012, we see no decline in the rate at which the sea level is rising.

From the analysis, we can clearly see, there is an urgent and serious need to take the necessary steps to curb pollution and stop environmental degradation. Environmental sustainability needs to become more prevalent and environmental degradation needs to be considered a serious threat before it is too late for the future generation.

**REFERENCES**

* Datopian. (n.d.). Climate Change Data. Retrieved from

[https://datahub.io/collections/climate-change#policy](https://datahub.io/collections/climate-change%23policy)

* Hall, M. (2018, October 22). What Is Purchasing Power Parity (PPP)? Retrieved from <https://www.investopedia.com/updates/purchasing-power-parity-ppp/>
* K-means Cluster Analysis. (n.d.). Retrieved from  
  [https://uc-r.github.io/kmeans\_clustering](https://uc-r.github.io/kmeans_clustering%20%20%09)
* UNdata | record view | Per capita GDP at current prices - US dollars. (n.d.). Retrieved from [http://data.un.org/Data.aspx?q=GDP percapita&d=SNAAMA&f=grID:101;currID:USD;pcFlag:1](http://data.un.org/Data.aspx?q=GDP%20percapita&d=SNAAMA&f=grID:101;currID:USD;pcFlag:1)
* Geom\_bar. (n.d.). Retrieved from <https://plot.ly/ggplot2/geom_bar/>
* Humbertohp, Andyhwang, Mindonly, Shirokaner, RoushanPrasad, & Jyupanquidur. (n.d.). Sign In. Retrieved from <http://rpubs.com/>

**APPENDIX**

#set environment

setwd("E:/r mining/emission")

#importing lib

library(countrycode)

library(dplyr)

library(ggplot2)

library(reshape2)

#library(tidyr)

library(ggrepel)

#importing csv

ghg = read.csv("Greenhouse Gas (GHGs) Emissions without Land Use, Land-Use Change and Forestry (LULUCF), in Gigagrams (Gg) CO2 equivalent UN DATA REPORT.csv",header = T, sep = ",",stringsAsFactors = F)

co = read.csv("Carbon dioxide (CO2) Emissions without Land Use, Land-Use Change and Forestry (LULUCF), in Gigagrams (Gg) UN DATA REPORT (1).csv",header = T, sep = ",",stringsAsFactors = F)

hfc = read.csv("Hydrofluorocarbons (HFCs) Emissions, in Gigagrams (Gg) CO2 equivalent UN DATA REPORT.csv",header = T, sep = ",",stringsAsFactors = F)

methane = read.csv("Methane (CH4) Emissions without Land Use, Land-Use Change and Forestry (LULUCF), in Gigagrams (Gg) CO2 equivalent UN DATA REPORT.csv",header = T, sep = ",",stringsAsFactors = F)

no = read.csv("Nitrous oxide (N2O) Emissions without Land Use, Land-Use Change and Forestry (LULUCF), in Gigagrams (Gg) CO2 equivalent UN DATA REPORT.csv",header = T, sep = ",",stringsAsFactors = F)

pfc = read.csv("Perfluorocarbons (PFCs) Emissions, in Gigagrams (Gg) CO2 equivalent EU DATA REPORT.csv",header = T, sep = ",",stringsAsFactors = F)

sul = read.csv("Sulphur hexafluoride (SF6) Emissions, in Gigagrams (Gg) CO2 equivalent EU DATA REPORT.csv",header = T, sep = ",",stringsAsFactors = F)

sources = read.csv("global.1751\_2010.csv",header = T, sep = ",",stringsAsFactors = F)

co.emmission = read.csv("Carbon dioxide emissions (CO2), metric tons of CO2 per capita (CDIAC) UN DATA REPORT.csv",header = T, sep = ",",stringsAsFactors = F)

co.per.dollar = read.csv("Carbon dioxide emissions (CO2), kg CO2 per $1 GDP (PPP) (UNFCCC) UN DATA REPORT.csv",header = T, sep = ",",stringsAsFactors = F)

reason = read.csv("nation.1751\_2014.csv",header = T, sep = ",",stringsAsFactors = F)

gdp = read.csv("UNdata\_Export\_20190214\_070854165 (1).csv",sep = ",",stringsAsFactors = F)

sealevel<-read.csv("csiro\_alt\_gmsl\_yr\_2015\_sealevelcsv.csv",stringsAsFactors = FALSE)

reason<-read.csv("nation.1751\_2014.csv",stringsAsFactors = FALSE)

glacier<-read.csv("glaciers\_csv.csv",stringsAsFactors = FALSE)

pop = population

#mapping contry and continent using countrycode()

cont.conti = countrycode(ghg$Country.or.Area,origin = "country.name",destination = "continent")

cont.conti.co = countrycode(co$Country.or.Area,origin = "country.name",destination = "continent")

cont.conti.hfc = countrycode(hfc$Country.or.Area,origin = "country.name",destination = "continent")

cont.conti.methane = countrycode(methane$Country.or.Area,origin = "country.name",destination = "continent")

cont.conti.no = countrycode(no$Country.or.Area,origin = "country.name",destination = "continent")

cont.conti.pfc = countrycode(pfc$Country.or.Area,origin = "country.name",destination = "continent")

cont.conti.sul = countrycode(sul$Country.or.Area,origin = "country.name",destination = "continent")

cont.conti.ce = countrycode(co.emmission$Country.or.Area,origin = "country.name",destination = "continent")

cont.conti.nation = countrycode(reason$Nation,origin = "country.name",destination = "continent")

##Load country-year observations into variable (mainly for Country name, continent and eurocontrol\_pru)

cont.code = countrycode::codelist\_panel

#with 6 continents

cont.conti=replace(cont.conti, is.na(cont.conti), "Europe")

cont.conti.co =replace(cont.conti.co, is.na(cont.conti.co), "Europe")

cont.conti.hfc =replace(cont.conti.hfc, is.na(cont.conti.hfc), "Europe")

cont.conti.methane =replace(cont.conti.methane, is.na(cont.conti.methane), "Europe")

cont.conti.no =replace(cont.conti.no, is.na(cont.conti.no), "Europe")

cont.conti.pfc =replace(cont.conti.pfc, is.na(cont.conti.pfc), "Europe")

cont.conti.sul =replace(cont.conti.sul, is.na(cont.conti.sul), "Europe")

cont.conti.ce = replace(cont.conti.ce, is.na(cont.conti.ce),"Europe")

cont.conti.nation = replace(cont.conti.nation, is.na(cont.conti.nation),"Europe")

#attaching continent column to all gases csv

ghg$conti = cont.conti

co$conti = cont.conti.co

hfc$conti = cont.conti.hfc

methane$conti = cont.conti.methane

no$conti = cont.conti.no

pfc$conti = cont.conti.pfc

sul$conti = cont.conti.sul

co.emmission$conti = cont.conti.ce

reason$conti = cont.conti.nation

#gdp$conti = cont.conti.gdp

#replacing country in country code

cont.code$country.name.en = replace(cont.code$country.name.en,cont.code$country.name.en == "United States", "United States of America")

#calculating unique countries of north and south america to seperate in the continent

sa.country = as.data.frame(cont.code$country.name.en)

sa.continent = as.data.frame(cont.code$eurocontrol\_pru)

sa = filter(sa.country,sa.continent == "Southern America")

sa.unique = unique(sa,incomparables = F)

sa.unique

na.country = as.data.frame(cont.code$country.name.en, stringsAsFactors = F)

na.continent = as.data.frame(cont.code$eurocontrol\_pru, stringsAsFactors = F)

na = filter(na.country,na.continent == "Northern America")

na.unique = unique(na)

na.unique

##-----------------------------------------------------------------------------#

#for plotting emission continent wise,we are sorting the data with respect to continent and then converting Americas

#into north and south america.

#green house gas

ec=ghg

ec.agg<-aggregate(ec$Value,by=list(ec$Country.or.Area),FUN=mean)

ec.agg$continent<-countrycode(ec.agg$Group.1,origin = "country.name",destination = "continent")

ec.agg$continent<-replace(ec.agg$continent,is.na(ec.agg$continent),"Europe")

#loop to sort countries based on north and south america---------------------------------------------------#

i = 0

j = 0

for (i in 1:length(sa.unique[,1])) {

for (j in 1:length(ec.agg$Group.1))

{

if( sa.unique[i,1]== ec.agg$Group.1[j]){

ec.agg$continent[j] = "South America"

}

}}

p=0

q=0

for (p in 1:length(na.unique[,1])) {

for (q in 1:length(ec.agg$Group.1))

{

if( na.unique[p,1]== ec.agg$Group.1[q]){

ec.agg$continent[q] = "North America"

}

}}

ec.agg.ghg<-aggregate(ec.agg$x,by=list(ec.agg$continent),FUN=sum)

#Emissions continent wise

#co emission

ec.co=co

ec.coagg<-aggregate(ec.co$Value,by=list(ec.co$Country.or.Area),FUN=mean)

ec.coagg$continent<-countrycode(ec.coagg$Group.1,origin = "country.name",destination = "continent")

ec.coagg$continent<-replace(ec.coagg$continent,is.na(ec.coagg$continent),"Europe")

#co loop to sort countries based on north and south america---------------------------------------------------#

i = 0

j = 0

for (i in 1:length(sa.unique[,1])) {

for (j in 1:length(ec.coagg$Group.1))

{

if( sa.unique[i,1]== ec.coagg$Group.1[j]){

ec.coagg$continent[j] = "South America"

}

}}

p=0

q=0

for (p in 1:length(na.unique[,1])) {

for (q in 1:length(ec.coagg$Group.1))

{

if( na.unique[p,1]== ec.coagg$Group.1[q]){

ec.coagg$continent[q] = "North America"

}

}}

ec.agg.co<-aggregate(ec.coagg$x,by=list(ec.coagg$continent),FUN=sum)

#sorting hydro floro carbons

ec.hfc=hfc

ec.hfcagg<-aggregate(ec.hfc$Value,by=list(ec.hfc$Country.or.Area),FUN=mean)

ec.hfcagg$continent<-countrycode(ec.hfcagg$Group.1,origin = "country.name",destination = "continent")

ec.hfcagg$continent<-replace(ec.hfcagg$continent,is.na(ec.hfcagg$continent),"Europe")

#hfc loop to sort countries based on north and south america---------------------------------------------------#

i = 0

j = 0

for (i in 1:length(sa.unique[,1])) {

for (j in 1:length(ec.hfcagg$Group.1))

{

if( sa.unique[i,1]== ec.hfcagg$Group.1[j]){

ec.hfcagg$continent[j] = "South America"

}

}}

p=0

q=0

for (p in 1:length(na.unique[,1])) {

for (q in 1:length(ec.hfcagg$Group.1))

{

if( na.unique[p,1]== ec.hfcagg$Group.1[q]){

ec.hfcagg$continent[q] = "North America"

}

}}

ec.agg.hfc<-aggregate(ec.hfcagg$x,by=list(ec.hfcagg$continent),FUN=sum)

# sorting methane

ec.methane = methane

ec.methaneagg<-aggregate(ec.methane$Value,by=list(ec.methane$Country.or.Area),FUN=mean)

ec.methaneagg$continent<-countrycode(ec.methaneagg$Group.1,origin = "country.name",destination = "continent")

ec.methaneagg$continent<-replace(ec.methaneagg$continent,is.na(ec.methaneagg$continent),"Europe")

#methane,loop to sort countries based on north and south america---------------------------------------------------#

i = 0

j = 0

for (i in 1:length(sa.unique[,1])) {

for (j in 1:length(ec.methaneagg$Group.1))

{

if( sa.unique[i,1]== ec.methaneagg$Group.1[j]){

ec.methaneagg$continent[j] = "South America"

}

}}

p=0

q=0

for (p in 1:length(na.unique[,1])) {

for (q in 1:length(ec.methaneagg$Group.1))

{

if( na.unique[p,1]== ec.methaneagg$Group.1[q]){

ec.methaneagg$continent[q] = "North America"

}

}}

ec.agg.methane<-aggregate(ec.methaneagg$x,by=list(ec.methaneagg$continent),FUN=sum)

#sorting and creating new tables based on nitrous oxide

ec.no = no

ec.noagg<-aggregate(ec.no$Value,by=list(ec.no$Country.or.Area),FUN=mean)

ec.noagg$continent<-countrycode(ec.noagg$Group.1,origin = "country.name",destination = "continent")

ec.noagg$continent<-replace(ec.noagg$continent,is.na(ec.noagg$continent),"Europe")

#no,loop to sort countries based on north and south america---------------------------------------------------#

i = 0

j = 0

for (i in 1:length(sa.unique[,1])) {

for (j in 1:length(ec.noagg$Group.1))

{

if( sa.unique[i,1]== ec.noagg$Group.1[j]){

ec.noagg$continent[j] = "South America"

}

}}

p=0

q=0

for (p in 1:length(na.unique[,1])) {

for (q in 1:length(ec.noagg$Group.1))

{

if( na.unique[p,1]== ec.noagg$Group.1[q]){

ec.noagg$continent[q] = "North America"

}

}}

ec.agg.no<-aggregate(ec.noagg$x,by=list(ec.noagg$continent),FUN=sum)

#pfc,sorting and storing data

ec.pfc = pfc

ec.pfcagg<-aggregate(ec.pfc$Value,by=list(ec.pfc$Country.or.Area),FUN=mean)

ec.pfcagg$continent<-countrycode(ec.pfcagg$Group.1,origin = "country.name",destination = "continent")

ec.pfcagg$continent<-replace(ec.pfcagg$continent,is.na(ec.pfcagg$continent),"Europe")

#pfc,loop to sort countries based on north and south america---------------------------------------------------#

i = 0

j = 0

for (i in 1:length(sa.unique[,1])) {

for (j in 1:length(ec.pfcagg$Group.1))

{

if( sa.unique[i,1]== ec.pfcagg$Group.1[j]){

ec.pfcagg$continent[j] = "South America"

}

}}

p=0

q=0

for (p in 1:length(na.unique[,1])) {

for (q in 1:length(ec.pfcagg$Group.1))

{

if( na.unique[p,1]== ec.pfcagg$Group.1[q]){

ec.pfcagg$continent[q] = "North America"

}

}}

ec.agg.pfc<-aggregate(ec.pfcagg$x,by=list(ec.pfcagg$continent),FUN=sum)

#sul,sorting and storing related data

ec.sul = sul

ec.sulagg<-aggregate(ec.sul$Value,by=list(ec.sul$Country.or.Area),FUN=mean)

ec.sulagg$continent<-countrycode(ec.sulagg$Group.1,origin = "country.name",destination = "continent")

ec.sulagg$continent<-replace(ec.sulagg$continent,is.na(ec.sulagg$continent),"Europe")

#sul,loop to sort countries based on north and south america---------------------------------------------------#

i = 0

j = 0

for (i in 1:length(sa.unique[,1])) {

for (j in 1:length(ec.sulagg$Group.1))

{

if( sa.unique[i,1]== ec.sulagg$Group.1[j]){

ec.sulagg$continent[j] = "South America"

}

}}

p=0

q=0

for (p in 1:length(na.unique[,1])) {

for (q in 1:length(ec.sulagg$Group.1))

{

if( na.unique[p,1]== ec.sulagg$Group.1[q]){

ec.sulagg$continent[q] = "North America"

}

}}

ec.agg.sul<-aggregate(ec.sulagg$x,by=list(ec.sulagg$continent),FUN=sum)

#plotting greenhouse gasses based on continent

ec = ghg %>% select(1,2,3,4)

#to convert americas to north and south america

p=0

q=0

for (p in 1:length(na.unique[,1])) {

for (q in 1:length(ec$Country.or.Area))

{

if( na.unique[p,1]== ec$Country.or.Area[q]){

ec$conti[q] = "North America"

}

}}

ec.agg = aggregate(ec$Value, by = list(ec$conti, ec$Year), FUN = sum)

names(ec.agg)[1]<-paste("continent")

emsn.conti.ghg = ggplot(data = ec.agg, aes(x = ec.agg$Group.2, y = ec.agg$x,group = continent, color =continent)) +

geom\_line()+

geom\_point()

emsn\_plot<-emsn.conti.ghg+scale\_x\_continuous(breaks = c(1990:2012))+scale\_y\_continuous(labels = scales::comma)+xlab("Years")+ylab("Emissions in Gigagrams, CO2 equivalent")+ggtitle("Ghg emissions continent wise")

emsn\_plot

#plotting all gases based upon continent

gases = ggplot(data=ec.agg.co, aes(x=ec.agg.co$Group.1, y = ec.agg.co$x,group =1))+

geom\_line(aes(color = "CO2"))+geom\_point(aes(color = "CO2"))+

geom\_line(data = ec.agg.hfc, aes(x=ec.agg.hfc$Group.1,y = ec.agg.hfc$x,color ="HFC"))+geom\_point(data = ec.agg.hfc, aes(x=ec.agg.hfc$Group.1,y = ec.agg.hfc$x,color ="HFC"))+

geom\_line(data = ec.agg.methane, aes(x=Group.1,y =x,color ="Methane"))+geom\_point(data = ec.agg.methane, aes(x=Group.1,y =x,color ="Methane"))+

geom\_line(data = ec.agg.no, aes(x=Group.1,y =x,color ="Nitrous Oxide"))+geom\_point(data = ec.agg.no, aes(x=Group.1,y =x,color ="Nitrous Oxide"))+

geom\_line(data = ec.agg.pfc, aes(x=Group.1,y =x,color ="Perfluorocarbons"))+geom\_point(data = ec.agg.pfc, aes(x=Group.1,y =x,color ="Perfluorocarbons"))+

geom\_line(data = ec.agg.sul, aes(x=Group.1,y =x,color ="Sulphur hexafluoride"))+geom\_point(data = ec.agg.sul, aes(x=Group.1,y =x,color ="Sulphur hexafluoride"))+

scale\_y\_continuous(labels = scales::comma)+xlab("Continents")+ylab("Emissions in Gigagrams, CO2 equivalent")+ggtitle("Emission of all the Gases,continent wise")

gases

#-----------------------------------------------------------------------------#

#co2 emissions percentages in ghg emission

#% of co2 in ghg in asia

co.asia = (ec.agg.co[1,2]/ec.agg.ghg[1,2])\*100

co.asia

#% of co2 in ghg in europe

co.eu = (ec.agg.co[2,2]/ec.agg.ghg[2,2])\*100

co.eu

#% of co2 in ghg in north america

co.na = (ec.agg.co[3,2]/ec.agg.ghg[3,2])\*100

co.na

#% of co2 in ghg in oceania

co.oce = (ec.agg.co[4,2]/ec.agg.ghg[4,2])\*100

co.oce

#creating a table to show the value

gas.co.continent = matrix(c(co.asia,co.eu,co.na,co.oce),ncol = 1, byrow = T)

colnames(gas.co.continent) = c("perc")

rownames(gas.co.continent) = c("Asia","Europe","North America","Oceania")

gas.co.continent.table = as.table(gas.co.continent)

gas.co.continent.df =as.data.frame(gas.co.continent)

#merging ghg and co2 values based on group.1

merged.ghg.co = merge(ec.agg.ghg,ec.agg.co,by="Group.1")

names(merged.ghg.co)[2]<-paste("GHG")

names(merged.ghg.co)[3]<-paste("CO2")

merged.ghg.co.melted = melt(merged.ghg.co, id.vars = "Group.1")

#plotting the percentages of carbondioxide per continent

ggplot(merged.ghg.co.melted, aes(x=Group.1, y=value)) +

geom\_bar(stat='identity', position='dodge',aes(fill=variable))+

xlab("Continent")+ylab("Emission")+scale\_y\_continuous(labels = scales::comma)+

ggtitle("Percentages of carbon dioxide per continent per ghg gases emission")+

geom\_text(data = gas.co.continent.df, aes(label = paste(round(perc[1],2),"%"),y = 590000, x = 1.3))+

geom\_text(data = gas.co.continent.df, aes(label = paste(round(perc[2],2),"%"),y = 6900000, x = 2.3))+

geom\_text(data = gas.co.continent.df, aes(label = paste(round(perc[3],2),"%"),y = 4200000, x = 3.4))+

geom\_text(data = gas.co.continent.df, aes(label = paste(round(perc[4],2),"%"),y = 830000, x = 4.3))

#getting the value into one dataframe

#---------------------------------------------------------------------#

#reading per kg CO2, per dollar (PPP) data

co2\_ppp <- co.per.dollar

str(co2\_ppp)

#Omit records that have NULL PPP values

co2\_ppp<- na.omit(co2\_ppp)

#read country-wise GDP data

str(gdp)

gdp<- gdp[-3] #irrelevant variable

# Trial to check NA#-------------

co2\_ppp\_gdp<- merge(co2\_ppp%>%filter(Year>=1995 & Year<=2012)%>%group\_by(Country.or.Area,Year,Value),

gdp%>%filter(Year>=1995 & Year<=2012)%>%group\_by(Country.or.Area,Year,Value),

by = c("Country.or.Area", "Year"), all.x=TRUE)

str(co2\_ppp\_gdp)

names(co2\_ppp\_gdp)[3]<- paste("kg.CO2.per.Dollar.PPP")

names(co2\_ppp\_gdp)[4]<- paste("GDP")

#NULL GDP Values

na.gdp<-unique(co2\_ppp\_gdp[is.na(co2\_ppp\_gdp$GDP),]%>% select(1))

na.gdp

#replace these values in GDP data frame

gdp$Country.or.Area<- replace(as.character(gdp$Country.or.Area), gdp$Country.or.Area=="Czechia", values= "Czech Republic")

gdp$Country.or.Area<- replace(as.character(gdp$Country.or.Area), gdp$Country.or.Area=="United Kingdom of Great Britain and Northern Ireland", values= "United Kingdom")

#Merge GDP and PPP data into one dataframe for years 1995-2012

co2\_ppp\_gdp<- merge(co2\_ppp%>%filter(Year>=1995 & Year<=2012)%>%group\_by(Country.or.Area,Year,Value),

gdp%>%filter(Year>=1995 & Year<=2012)%>%group\_by(Country.or.Area,Year,Value),

by = c("Country.or.Area", "Year"), all.x=TRUE)

str(co2\_ppp\_gdp)

names(co2\_ppp\_gdp)[3]<- paste("kg.CO2.per.Dollar.PPP")

names(co2\_ppp\_gdp)[4]<- paste("GDP")

#co2\_ppp\_gdp$Year<- as.numeric(co2\_ppp\_gdp$Year)

#check for NA

anyNA(co2\_ppp\_gdp)

#----------------------------------------------------------------------------------------#

#Average gdp of each year

gdppp\_avg<- aggregate(c(co2\_ppp\_gdp$GDP),by=list(co2\_ppp\_gdp$Country.or.Area), FUN="mean" )

#Rename columns

colnames(gdppp\_avg)<- c("Country.or.Area","GDP")

#Average kg Co2 per dollar GDP (PPP)

PPP<- aggregate(co2\_ppp\_gdp$kg.CO2.per.Dollar.PPP,by = list(co2\_ppp\_gdp$Country.or.Area), FUN= "mean")

#Rename columns

colnames(PPP)<- c("Country.or.Area","PPP")

#merge average PPP with GDP

gdppp\_avg<- merge(gdppp\_avg, PPP,by = "Country.or.Area",all.y=TRUE)

str(gdppp\_avg)

count(gdppp\_avg)

gdp\_ppp <- gdppp\_avg

#Load country-year observations into variable (mainly for Country name, continent and eurocontrol\_pru)

cntry <- countrycode::codelist\_panel

#cntry\_csv<- read.csv("C:/Users/iyeng/Documents/MPS Analytics/ALY 6040/Project/B\_emission/country-codes\_csv.csv")

#Renamed column name in country for Merge in future.

cntry<- cntry%>% rename(Country.or.Area = country.name.en)

gdp\_ppp$continent<- countrycode(gdp\_ppp$Country.or.Area, "country.name","continent")

gdp\_ppp<- merge ( x= unique(gdp\_ppp), y = unique(cntry%>% select(1,9)), by = "Country.or.Area",all.x = TRUE)

#Replace Americas value with corresponding eurocontrol\_pru value

gdp\_ppp$continent<- replace(gdp\_ppp$continent, gdp\_ppp$continent=="Americas", gdp\_ppp$eurocontrol\_pru[which(gdp\_ppp$continent=="Americas")])

#remove eurocontrol\_pru column (optional step)

gdp\_ppp<- within(gdp\_ppp, rm(eurocontrol\_pru))

str(gdp\_ppp)

gdp\_ppp\_plot<- ggplot(gdp\_ppp, aes(x=GDP,y=PPP))+geom\_point(aes(color = gdp\_ppp$continent))+

geom\_label\_repel(aes(label = gdp\_ppp$Country.or.Area),

box.padding = 0.35,

point.padding = 0.8,

segment.color = 'grey50') +

theme\_classic()

gdp\_ppp\_plot

#--------------------------------------------------------------------#

#Country names as Rownames

rownames(gdppp\_avg)<-c(as.character(gdppp\_avg$Country.or.Area))

#used get\_dist from factoextra package to compute distance matrix between rows

library(factoextra)

distance <- get\_dist(gdppp\_avg[,2:3])

#used to visualize distance matrix

fviz\_dist(distance, gradient = list(low = "#00AFBB", mid = "white", high = "#FC4E07"))

#kmeans clustering , 3 clusters , Low-Moderate-High GDP

km<-kmeans(gdppp\_avg[,2:3],3, iter.max = 2, nstart=3)

fviz\_cluster(km, data = gdppp\_avg[,2:3])

#---------------------------------------------------------------------#

#co2 emission from various sources###################

reason

reason<-replace(reason,is.na(reason),0)

any(is.na(reason))=="TRUE"

head(reason)

reason$con<-countrycode(reason$Nation,"country.name","continent")

reason$con<-replace(reason$con,is.na(reason$con),"Europe")

#using nation dataset to find out the various sources of emission----------------------------------------------#

i = 0

j = 0

for (i in 1:length(sa.unique[,1])) {

for (j in 1:length(reason$Nation))

{

if( toupper(sa.unique[i,1])== reason$Nation[j]){

reason$con[j] = "South America"

}

}}

p=0

q=0

for (p in 1:length(na.unique[,1])) {

for (q in 1:length(reason$Nation))

{

if(toupper( na.unique[p,1])== reason$Nation[q]){

reason$con[q] = "North America"

}

}}

library(dplyr)

#to remove unrelated values

reason<-reason %>%

filter(!grepl('Americas', con))

reason$con

sort(reason$con)

#to convert each column into integer

reason$Emissions.from.solid.fuel.consumption<-as.integer(reason$Emissions.from.solid.fuel.consumption)

reason$Emissions.from.liquid.fuel.consumption<-as.integer(reason$Emissions.from.liquid.fuel.consumption)

reason$Emissions.from.gas.fuel.consumption<-as.integer(reason$Emissions.from.gas.fuel.consumption)

reason$Emissions.from.cement.production<-as.integer(reason$Emissions.from.cement.production)

reason$Emissions.from.gas.flaring<-as.integer(reason$Emissions.from.gas.flaring)

reason$Per.capita.CO2.emissions..metric.tons.of.carbon.<-as.integer(reason$Per.capita.CO2.emissions..metric.tons.of.carbon.)

class(reason)

a<-aggregate(reason, by=list(reason$con), FUN=mean)

a

a$Nation<-NULL

a$con<-NULL

a$Year<-NULL

names(a)[names(a) == "Group.1"] <- "Continent"

#plot sources of carbon emission-------------------------------------------

ggplot(a,aes(Continent,Emissions.from.liquid.fuel.consumption,group=1))+

geom\_line(aes(color = "liquid fuel"))+geom\_point(aes(color="liquid fuel"))+

geom\_line(data=a ,aes(Continent,Emissions.from.solid.fuel.consumption,color="solid fuel"))+geom\_point(aes(Continent,Emissions.from.solid.fuel.consumption,color="solid fuel"))+

geom\_line(data=a,aes(Continent,Emissions.from.gas.fuel.consumption,color="gas fuel"))+geom\_point(aes(Continent,Emissions.from.gas.fuel.consumption,color="gas fuel"))+

geom\_line(data=a,aes(Continent,Emissions.from.cement.production,color="cement production"))+geom\_point(aes(Continent,Emissions.from.cement.production,color="cement production"))+

geom\_line(data=a,aes(Continent,Emissions.from.gas.flaring,color="gas flare"))+geom\_point(aes(Continent,Emissions.from.gas.flaring,color="gas flare"))+

ylab("emission levels")+ggtitle(" Sources of carbon emission")

#-----------------------------------------------------------------------------

######glacier density#########

carbon<-read.csv("nation.1751\_2014.csv",stringsAsFactors = FALSE)

carbon<-cbind(carbon$Total.CO2.emissions.from.fossil.fuels.and.cement.production..thousand.metric.tons.of.C.,carbon$Year)

carbon[,1]

glacier

#changing the header

colnames(carbon)<-c("Emission.Measure","Year")

#merge two data frames based on year

emm\_glac<-merge(carbon, glacier%>%select(1,2),by= "Year",all.y=TRUE )

str(emm\_glac)

emm\_glac<-aggregate(emm\_glac$Emission.Measure,by=list(emm\_glac$Year,emm\_glac$Mean.cumulative.mass.balance),FUN="sum")

colnames(emm\_glac)<-c("Year","Mean.cumulative.mass.balance","Carbon.Emmission.Measure")

str(emm\_glac)

emm\_glac$Mean.cumulative.mass.balance<-as.integer(emm\_glac$Mean.cumulative.mass.balance)

#plotting glacier density with respect to year

plot( emm\_glac$Year, emm\_glac$Mean.cumulative.mass.balance, type="l", col="red",ylab = "glacier density",xlab = "year" ,main="comparison of glacier density and carbon emission with respect to time" )

par(new=TRUE)

with(emm\_glac, plot(emm\_glac$Year, emm\_glac$Carbon.Emmission.Measure, type = "l",col="green", axes=F, xlab=NA, ylab=NA, cex=1))

axis(side = 4)

mtext("carbon emission measure",side = 4, line = 2)

legend("left",legend=c("glacier level","carbon level"), "sea level",lty=1:1, cex=0.7,title = "line types", col=c("red", "green"))

#comparision of carbon emission vs sea level#######

str(sealevel)

sealevel$GMSL<-as.integer(sealevel$GMSL)

sealevel$Time<-NULL

sealevel

sea\_level<-merge(carbon, sealevel,by= "Year",all.y=TRUE )

str(sea\_level)

sea\_level<-aggregate(sea\_level$Emission.Measure,by=list(sea\_level$Year,sea\_level$GMSL),FUN="sum")

colnames(sea\_level)<-c("Year","Mean sea level","Carbon.Emmission.Measure")

str(sea\_level)

#plot sea level vs carbon

plot( sea\_level$Year, sea\_level$`Mean sea level`, type="l", col="red",ylab = "mean sea level",xlab = "year" ,main="comparison of mean sea level and carbon emission with respect to time" )

par(new=TRUE)

with(sea\_level, plot(sea\_level$Year, sea\_level$Carbon.Emmission.Measure, type = "l",col="green", axes=F, xlab=NA, ylab=NA, cex=1))

axis(side = 4)

mtext("carbon emission measure",side = 4, line = 2)

legend("topleft",legend=c("sea level","carbon level"), "sea level",lty=1:1, cex=1,title = "line types", col=c("red", "green"))

#prediction of sea level

library("forecast")

sea<-ts(sea\_level$`Mean sea level`, start=c(2000), end=c(2014))

plot(sea)

fit<-auto.arima(sea)

seafor<-forecast(fit,10)

plot(seafor,ylab = "sea level",xlab="year",main="prediction of sea level")

#prediction of glacier level

glaci<-ts(emm\_glac$Mean.cumulative.mass.balance,start = c(2000),end=c(2014))

plot(glaci)

fit1<-auto.arima(glaci)

glacifor<-forecast(fit1,10)

glacifor

plot(glacifor,ylab = "glacier density",xlab = "year",main="prediction of glacier density")