**Unsupervised Learning Analyses on Carbon Dioxide and Other Greenhouse Gases Emissions – An Insight into the Big Picture on Climate Change**

Name : Ahmad Luqman Al Hakim

Surname : bin Shamsudin

**Abstract**

This report employs unsupervised learning algorithms to analyse the data measuring emissions of carbon dioxide and greenhouse gases around the world. The data used contains variables on carbon dioxide and other greenhouse gases emitted per capita, by country and year. The purpose of this analysis is to find out which countries are more similar by their emissions levels. The analyses are performed using the R programming language, where the codes are provided in the appendix. We implement 2 unsupervised learning algorithms: k-means clustering and hierarchical clustering. In Section 1, we describe the data pre-processing techniques used. In Section 2, we describe the modelling methodology. Lastly, we discuss the findings in Section 3.

**Section 1: Data Descriptions**

The dataset is obtained from Our World in Data website, produced by Hannah Ritchie and Max Roser. The website is a project under the Global Change Data Lab at the University of Oxford. The dataset is available under the Creative Commons license.

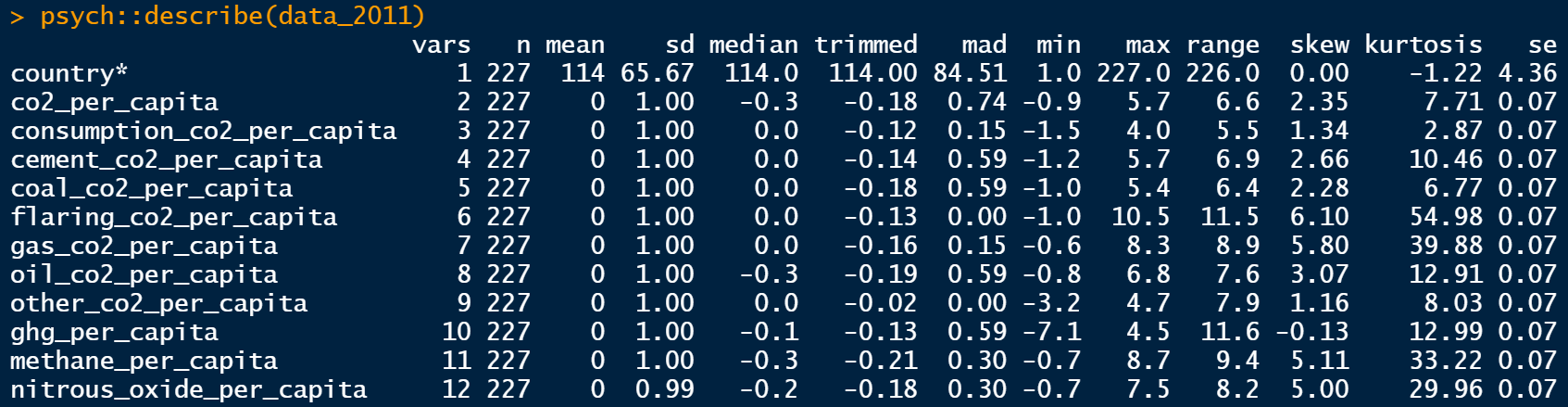
The original dataset (referred to as CO2 data in the code) contains 50 columns of various greenhouse gases measurements, recorded by years and countries. For the purpose of our analyses, however, we filtered the dataset and obtain the measurements taken in the year 2011 only. And we also filtered the data to only keep the columns that are measured adjusted to population. Therefore, we only keep the columns that have the term “per capita” in the name. As a side note, the missing values are imputed using the means of the columns using the library ‘imputeTS’ in CRAN. And since we want to ensure that the data is internally consistent (to solve inconsistent units of measurement), we standardised the whole dataset using ‘scale’ function. The final dataset that will be used for the analyses is referred to as data\_2011 in the code.

This final dataset contains various measures of per capita emissions of different gases represented by the 12 columns as below:

1. Country
2. Carbon dioxide per capita – Average per capita CO2 emissions (tonnes)
3. Consumption carbon dioxide per capita – Nitrous oxide emissions per capita (tonnes of carbon dioxide equivalents)
4. Cement carbon dioxide per capita – Per capita CO2 emissions from cement production (million tonnes)
5. Coal carbon dioxide per capita – Per capita CO2 emissions from coal productions (million tonnes)
6. Flaring carbon dioxide per capita – Per capita CO2 emissions from gas flaring (million tonnes
7. Gas carbon dioxide per capita – Per capita CO2 emissions from gas production (tonnes)
8. Oil carbon dioxide per capita – Per capita CO2 emissions from oil production (tonnes)
9. Other carbon dioxide per capita – Per capita CO2 emissions from other industrial processes excluding cement (tonnes)
10. Greenhouse gases per capita – Greenhouse gas emissions per capita (tonnes of carbon dioxide equivalents)
11. Methane per capita – Methane emissions per capita (tonnes of carbon dioxide equivalents)
12. Nitrous oxide per capita – Nitrous oxide emissions per capita (measured in tonnes of carbon dioxide equivalents)

**Section 2: Theoretical Backgrounds and Analyses**

After completing the data pre-processing which includes dealing with missing values, filtering, and standardising the data, we check the summary of the variables using ‘describe’ function from ‘psych’ library. This function that is chosen to present the summary statistics provides a tidier summary output than the ‘summary’ function. The output is presented in the table below.



Now that we completed the data pre-processing, we continue to the modelling part. To analyse this dataset, we employ the clustering algorithm. It is a broad class of methods for discovering subgroups in data. This method is used to try to identify some segments for clusters in data, otherwise hidden.

To determine which groups of countries have the most to improve with carbon and greenhouse gases emissions, we use all the per capita columns selected in the data pre-processing stage. Furthermore, the year 2011 is selected since it is the latest year that exists in the dataset and most countries are represented in that year.

**Section 2 (a): K-Means Classification**

The first algorithm we employ is the k-means clustering. K-means clustering is a vector  
quantisation method that seeks to split n observations into k clusters, with each observation  
belonging to the cluster with the closest mean, which serves as the cluster's prototype. This method guarantees convergence, it is simple to be implemented and it can generalise to clusters of different shapes and sizes.

The initial step to employ the k-means clustering algorithm is to identify the number of  
clusters, k. Presented below is the result of the Scree Diagram, which is used to identify the number of clusters for our k-means clustering. Looking at it, we decide to use k=4 as the number of clusters.

Chart, line chart

Description automatically generated

We see a rapid decline in the total within sum of squares (WSS) when k=4, and interestingly it does continue to decline even more until k=8. However, for the sake of the simplicity of our analysis and conclusions, we decided to settle with k=4. Thus, performing the cluster analysis on k=4, we obtain the following clusters:

Chart

Description automatically generated with medium confidence

As we can see from the output above, the countries in the world are mainly clustered into 4 different clusters. There is only one overlapping region out of 4 clusters, and it is observed between the teal cluster and the purple cluster. However, the 2 principal components explain only = 34.5% + 14.9% = 49.4% of the point variability. This percentage of point variability explained is decent, although it would be ideal to get a figure of about 75%.

Note that these clusters cannot tell us which countries pollute the most. It only tells us which countries can be grouped according to their emissions levels. However, we can see that the biggest oil producer countries such as Bahrain, Brunei, Qatar, Saudi Arabia, Oman, Kuwait, and United Arab Emirates are clustered together in the green cluster. While countries such as Cameroon, Botswana, Equatorial Guinea, and Guyana are clustered together in the red cluster. And the rest of the world falls into the teal cluster and the purple cluster. As far as this analysis goes, an emission level of a country in a cluster can be conjectured to the countries in the same cluster having similar levels of emissions. As far as carbon dioxide and greenhouse gases per capita emitted concerns, this data is the most comprehensive data. This will be explained more in the conclusion section. Thus, a conclusion derived from the clustering can be considered reliable.

**Section 2 (b): Hierarchical Clustering**

The second algorithm we employ is hierarchical clustering. It is a clustering algorithm that groups data points according to their distance and similarities. It can be used as an agglomerative algorithm, starting from each data point as a single cluster and proceeding until every point is in the same cluster, or the non-agglomerative method. We use the non-agglomerative, also called top-down, which is also the default method in the R functions used here.

Now we need to choose a dissimilarity measure and a distance measure. We check the dataset, and we see that all the variables are quantitative. Even if we already standardised all the variables, we rely on Gower’s dissimilarity measure. Then we need a linkage method, that defines what “distant” means. Here we use the complete linkage method, that is: the distance between two groups is computed as the maximum distance between two different observations in the two different groups. Presented below is the dendrogram obtained with these parameters:

Graphical user interface

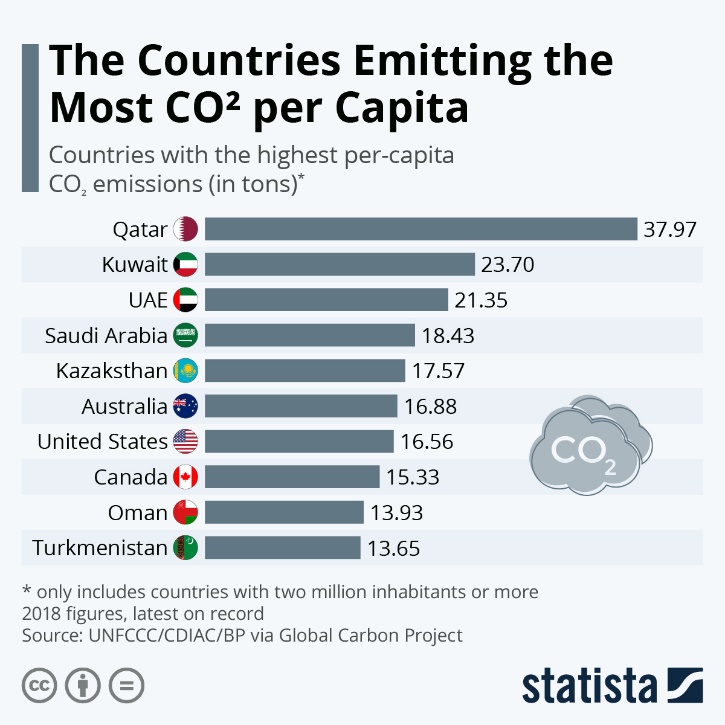
Description automatically generated

Here, we also cut the dendrogram using the same number of groups as in the k-means clustering (k=4). The interpretation is very similar to the previous clustering algorithm employed. Up until now, we cannot say which countries pollute the most, we can only say which countries pollute at similar levels to each other.

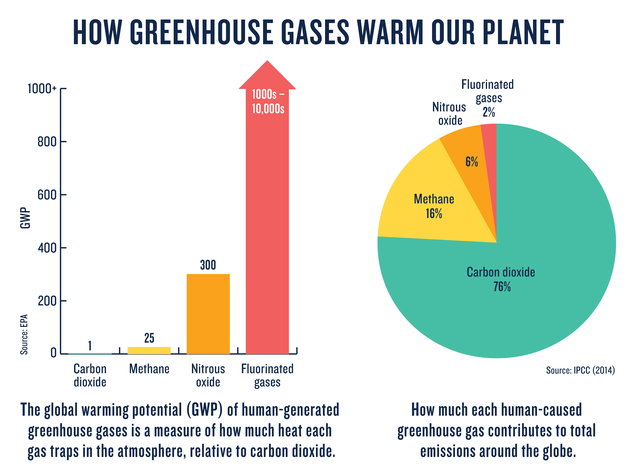
**Section 3: Conclusion**

In the table at the end of this section, we pull the names of the countries that are clustered together by the hierarchical clustering algorithm. Immediately, we observe that countries that are famously most polluting such as Bahrain, Qatar, United Arab Emirates and Brunei are in Cluster 3 and Cluster 4. Similar clusters are also produced by our k-means clustering. This leads us to the answer of which countries pollute the most? According to the latest report in 2021 by the Inter-governmental Panel on Climate Change (IPCC), the countries in certain clusters of our analyses are categorised as the most polluting countries. Thus, the clusters produced by our analysis are coherent with the IPCC report and has been the same from 2011 until 2021, since our dataset is from 2011. Meanwhile, the rest of the world lies in Cluster 1.

Furthermore, presented below is a chart published by Statista showing which countries produce the most CO2 per capita.



We can see that the countries that are listed on the chart, also tend to be clustered together with the countries that IPCC listed as most polluting. Meaning that carbon dioxide emissions go together with other greenhouse gases emissions. Our analyses in this paper enriched the classification with the inclusion of greenhouse gases measurements, not just carbon dioxide. This is important since greenhouse gases such as methane and nitrous oxide also contribute to the acceleration of global warming. According to a published chart from the Natural Resources Defence Council (NRDC) below, we see how important it is to include analyses on this topic not just based on CO2, but also on the other gases which are included in our paper.



The bar chart on the left shows the severity levels of different gases to global warming. As we see, other greenhouse gases not usually mentioned by mainstream media have the capacity to trap heat in our atmosphere significantly more than carbon dioxide. And the pie chart on the right shows the proportion of these gases in our atmosphere.

Considering other greenhouse gases such as methane and nitrous dioxide, our clustering analyses give insights into how geographically decentralised are the emissions that contributed most to climate change. Whilst we always hear some countries being claimed as being the biggest polluters, but most of the time, these claims only focus on CO2, not adjusted to the size of the population they have, and therefore not the overall picture of the emissions produced. Indeed, our analyses paint a clearer picture of this matter. An analysis such as this is important not just for our awareness, but also for better and more targeted policymaking. For instance, an understanding of who-does-what-by-doing-what can certainly help direct world leaders to intervene on the environmental policies and channel funds to catalyse the improvements at a specific location.

Finally, the result of our analyses draws attention to just how much behavioural aspects can detriment the planet. An ever-growing global demand for fuel has made some countries “rich” by catering to the demand. We conclusively observe in the table below, that fuel consumption destroys the planet by the emissions of carbon dioxide and greenhouse gases. Our inference is that Cluster 1, 2, 3 and 4 are grouped to reflect on how intensely they partake in the production of oil and gas and related industries.

However, that is a sensitive topic that should be explored more by researchers with personal security guards. Whilst it is easy to fall into a helplessness pit discussing this matter, we need to keep in mind that everything that is done now is crucial to preserve and improve what we have so that we will be able to leave this planet better than when we arrived.

|  |  |  |  |
| --- | --- | --- | --- |
| Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
| Afghanistan | Anguilla | Bahrain | Brunei |
| Africa | Aruba | Oman | Equatorial Guinea |
| Albania | Australia | Qatar |  |
| Algeria | Bonaire Sint Eustatius and Saba | Trinidad and Tobago |
| Andorra | Canada | United Arab Emirates |
| Angola | Estonia |  |
| Antigua and Barbuda | Faeroe Islands |
| Argentina | Greenland |
| Armenia | Kazakhstan |
| Asia | Kuwait |
| Asia (excl. China & India) | Luxembourg |
| Austria | Saint Pierre and Miquelon |
| Azerbaijan | Saudi Arabia |
| Bahamas | Sint Maarten (Dutch part) |
| Bangladesh | United States |
| Barbados |  |
| Belarus |
| Belgium |
| Belize |
| Benin |
| Bermuda |
| Bhutan |
| Bolivia |
| Bosnia and Herzegovina |
| Botswana |
| Brazil |
| British Virgin Islands |
| Bulgaria |
| Burkina Faso |
| Burundi |
| Cambodia |
| Cameroon |
| Cape Verde |
| Central African Republic |
| Chad |
| Chile |
| China |
| Colombia |
| Comoros |
| Congo |
| Cook Islands |
| Costa Rica |
| Cote d'Ivoire |
| Croatia |
| Cuba |
| Cyprus |
| Czechia |
| Democratic Republic of Congo |
| Denmark |
| Djibouti |
| Dominica |
| Dominican Republic |
| EU-27 |
| EU-28 |
| Ecuador |
| Egypt |
| El Salvador |
| Eritrea |
| Eswatini |
| Ethiopia |
| Europe |
| Europe (excl. EU-27) |
| Europe (excl. EU-28) |
| Fiji |
| Finland |
| France |
| French Polynesia |
| Gabon |
| Gambia |
| Georgia |
| Germany |
| Ghana |
| Greece |
| Grenada |
| Guatemala |
| Guinea |
| Guinea-Bissau |
| Guyana |
| Haiti |
| Honduras |
| Hong Kong |
| Hungary |
| Iceland |
| India |
| Indonesia |
| International transport |
| Iran |
| Iraq |
| Ireland |
| Israel |
| Italy |
| Jamaica |
| Japan |
| Jordan |
| Kenya |
| Kiribati |
| Kosovo |
| Kyrgyzstan |
| Laos |
| Latvia |
| Lebanon |
| Lesotho |
| Liberia |
| Libya |
| Liechtenstein |
| Lithuania |
| Macao |
| Madagascar |
| Malawi |
| Malaysia |
| Maldives |
| Mali |
| Malta |
| Marshall Islands |
| Mauritania |
| Mauritius |
| Mexico |
| Micronesia |
| Moldova |
| Mongolia |
| Montenegro |
| Montserrat |
| Morocco |
| Mozambique |
| Myanmar |
| Namibia |
| Nauru |
| Nepal |
| Netherlands |
| New Caledonia |
| New Zealand |
| Nicaragua |
| Niger |
| Nigeria |
| Niue |
| North America |
| North America (excl. USA) |
| North Korea |
| North Macedonia |
| Norway |
| Oceania |
| Pakistan |
| Palau |
| Palestine |
| Panama |
| Papua New Guinea |
| Paraguay |
| Peru |
| Philippines |
| Poland |
| Portugal |
| Romania |
| Russia |
| Rwanda |
| Saint Helena |
| Saint Kitts and Nevis |
| Saint Lucia |
| Saint Vincent and the Grenadines |
| Samoa |
| Sao Tome and Principe |
| Senegal |
| Serbia |
| Seychelles |
| Sierra Leone |
| Singapore |
| Slovakia |
| Slovenia |
| Solomon Islands |
| Somalia |
| South Africa |
| South America |
| South Korea |
| South Sudan |
| Spain |
| Sri Lanka |
| Sudan |
| Suriname |
| Sweden |
| Switzerland |
| Syria |
| Taiwan |
| Tajikistan |
| Tanzania |
| Thailand |
| Timor |
| Togo |
| Tonga |
| Tunisia |
| Turkey |
| Turkmenistan |
| Turks and Caicos Islands |
| Tuvalu |
| Uganda |
| Ukraine |
| United Kingdom |
| Uruguay |
| Uzbekistan |
| Vanuatu |
| Venezuela |
| Vietnam |
| Wallis and Futuna Islands |
| World |
| Yemen |
| Zambia |
| Zimbabwe |
|  |  |  |  |

**References**

1. Ritchie, H., & Roser, M. (2020, August). *CO2 and Greenhouse Gas Emissions*. Our World in Data. <https://ourworldindata.org/co2-and-other-greenhouse-gas-emissions>

**Appendix**

## data preparation ##

#import data

library(readr)

co2\_data <- read\_csv("owid-co2-data.csv", na = "NA")

#create data for year = 2011

library(dplyr)

data\_2011 <- co2\_data %>% filter(year== 2011)

#impute mean into missing values

library(imputeTS)

data\_2011 <- na\_mean(data\_2011)

#create df for country and year

country <- data\_2011[,2:3]

#select columns containing per capita data only

library(dplyr)

data\_2011 <- data\_2011[,4:50] %>% select(contains("capita"))

#combine country and year & data\_2011

data\_2011 <- cbind(country, data\_2011)

#rescale data and round data

data\_2011[,3:13] <- round(scale(data\_2011[,3:13], center = TRUE),1)

data\_2011 <- data\_2011 %>% select(-'year')

rownames(data\_2011) <- data\_2011$country

#check data summary

psych::describe(data\_2011)

## modelling part ##

library(cluster)

library(factoextra)

#k-means clustering

#choose optimal k using the elbow method

set.seed(123)

fviz\_nbclust(data\_2011[,2:12], kmeans, method = "wss")

#plot using k=4

k4 <- kmeans(data\_2011[,2:12], centers = 4, nstart = 25)

str(k4)

k4

#illustration of the clusters

fviz\_cluster(k4, data = data\_2011[,2:12], labelsize = 6, show.clust.cent = TRUE, repel =TRUE)

#hierarchical clustering

#hierarchical clustering, k=3

library(proxy)

simil(data\_2011[,2:12], method="Gower")

dendro\_plot <- hclust(dist(data\_2011[,2:12], method="Gower"), method = "complete")

plot(dendro\_plot, cex = 0.5, cex = 0.4)

groups <- cutree(dendro\_plot, k=4)

groups

rect.hclust(dendro\_plot, k=4, border= c("orange","red", "green", "blue"))

#pull country names and respective group

as.data.frame(cbind(groups))

write.csv(as.data.frame(cbind(groups)), file = "country\_cluster.csv")