R Programming Language – Assignment

Part 2: Data exploration

1.   About the dataset

For this assignment, I have chosen the global carbon emissions data set along with the global population statistics and land area of each country. The datasets and its source are elaborated in Table 1.1.

Table 1.1: *Details of the datasets chosen*

Data Set Name	Source	How is Data Gathered?	Description
CO2_1970_2021.csv URL: https://jeodpp.jrc.ec.europa.eu/ftp/jrc-opendata/EDGAR/ datasets/v70_FT2021_GHG/ v70_FT2021_CO2_1970_2021.zip	EDGAR v7.0 (Emissions database for Global Atmospheric Research) – an independent database and joint project of European Union	Input to EDGAR are international annual statistics from different countries	This dataset contains total carbon emissions (measured in kT) of more than 200 countries from 1970 to 2021.
API_AG.LND.TOTL. K2_DS2_en_csv_v2_4701206.csv URL: https://api.worldbank.org/v2/ en/indicator/AG.LND.TOTL.K2? downloadformat=csv	World Bank – an open data bank for free and open access to global development data	The member nations of world bank provide much of the data.	This dataset indicates the land area (measured in sq.km) of more than 200 countries from 1960 to 2021.
API_SP.POP. TOTL_DS2_en_csv_v2_4701113.csv URL: https://api.worldbank.org/v2/ en/indicator/SP.POP.TOTL? downloadformat=csv	World Bank – an open data bank for free and open access to global development data	The member nations of world bank provide the majority of the data.	This dataset shows the population of more than 200 countries from 1960 to 2021.

2.   Data Pre-processing and cleaning

(Refer *Section 2: Data pre-processing and data cleaning* in code)

The available data is in a different format and spread across 3 different datasets. These data sets need to be cleaned and merged into a single table of our desired format to do further analysis on it. Some of pre-processing done on the data are –

- Skip irrelevant columns and first few rows in the csv file that provides descriptions of the data.
- In the raw data, Year variable is spread across multiple columns with column header as its year value. This is tidied with help of pivot_longer() function and creating a Year column and the value (Emissions/Land area/Population) columns.

- The land area and populations dataset contain data from 1960. The year is filtered from 1970 to match with carbon emissions data.
- After merging the three data frames, new columns such as population density and carbon emission per capita are calculated and included as columns.

3. Data Visualisation

3.1. World Map of average carbon emissions from 1970-2021

(Refer *Section 3.1: World Map of average carbon emissions from 1970-2021* in code)
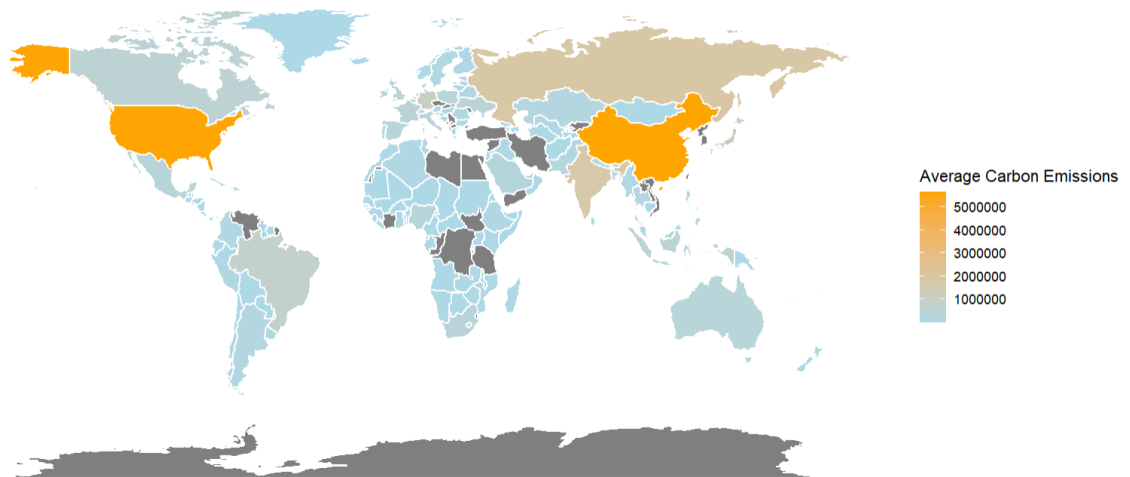

Figure 3.1.1: *Average Carbon Emissions by Countries from 1970 to 2021*

Figure 3.1.1 shows the average carbon emissions by country from 1970 to 2021. The map clearly indicates that China and United States of America are world's largest emitters of carbon. They are followed by India and Russia. Note that there are few countries whose data are not available in this dataset (coloured in grey).

3.2. Correlation between Population and Carbon Emissions

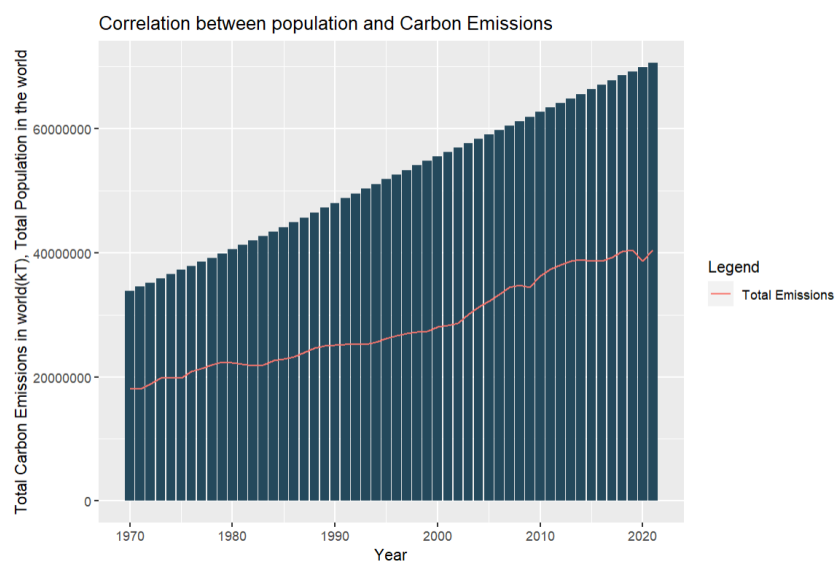(Refer *Section 3.2: Correlation between population and Carbon Emissions over the years* in code)

Figure 3.2.1: *Correlation between population and carbon emissions*

Figure 3.2.1 shows the correlation between population and carbon emissions across 1970 to 2021. The illustration shows that as years goes by the population of the world is increasing linearly and so is total carbon emissions of the world. This suggests that there is a linear and proportional relationship between global population and global carbon emission.

3.3. Top 10 countries with highest carbon emissions
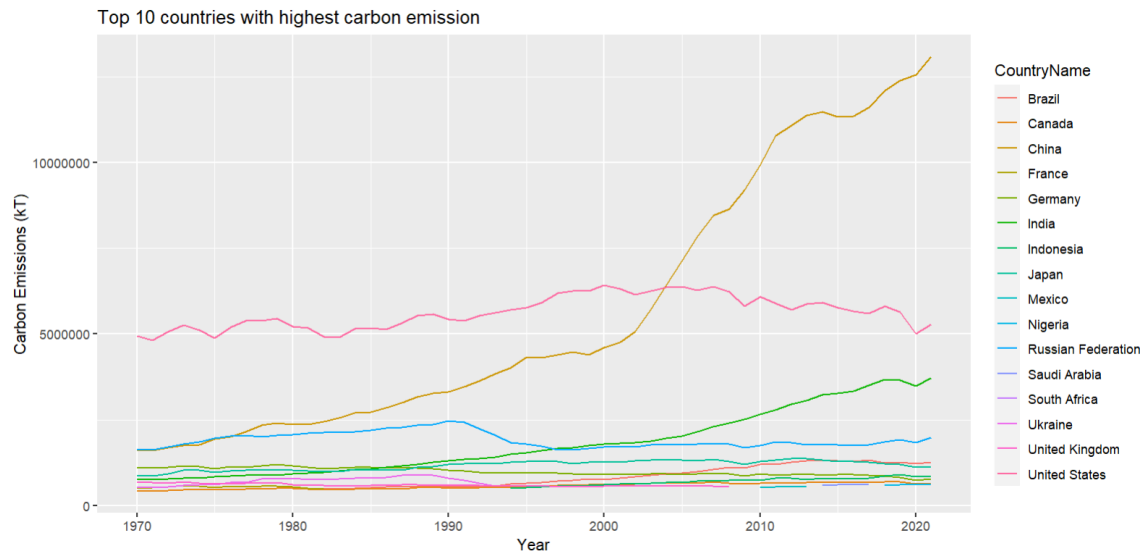   (Refer *Section 3.3: Top 10 countries with highest carbon emissions* in code)

Figure 3.3.1: *Top 10 highest carbon emitting carbon and its emission across years*

Figure 3.3.1 lists out the top 10 contributors of carbon emissions and trends in carbon emissions across 1970 to 2021. This shows a clearer indication of what was already seen in Figure 3.1.1. China is the biggest carbon emitter followed by USA, India, and Russia. The carbon emission from China has seen an exponential rise from 1970 whereas USA has approximately a steady carbon emission since 1970. This similar trend can be seen between India and Russia (although not as large).
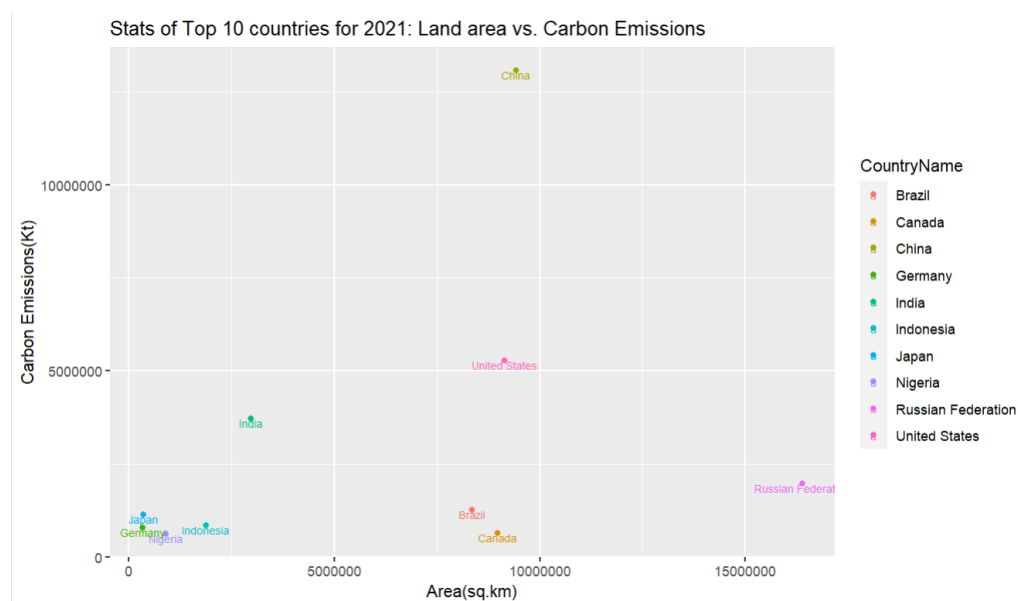

Figure 3.3.2: *Stats of Top 10 countries for 2021: Land area vs. Carbon Emissions*

Figure 3.3.2 further investigates the top 10 carbon emitters in the world. Here, a scatter plot is sketched between the land area(in km^2) and Carbon Emissions (in kt). The number 1 country China is also the 2nd largest country in terms of land area. However, this graph doesn't show any trend between land area and Carbon emissions for the top 10 countries. Hence, we cannot decisively say that these variables are correlated.
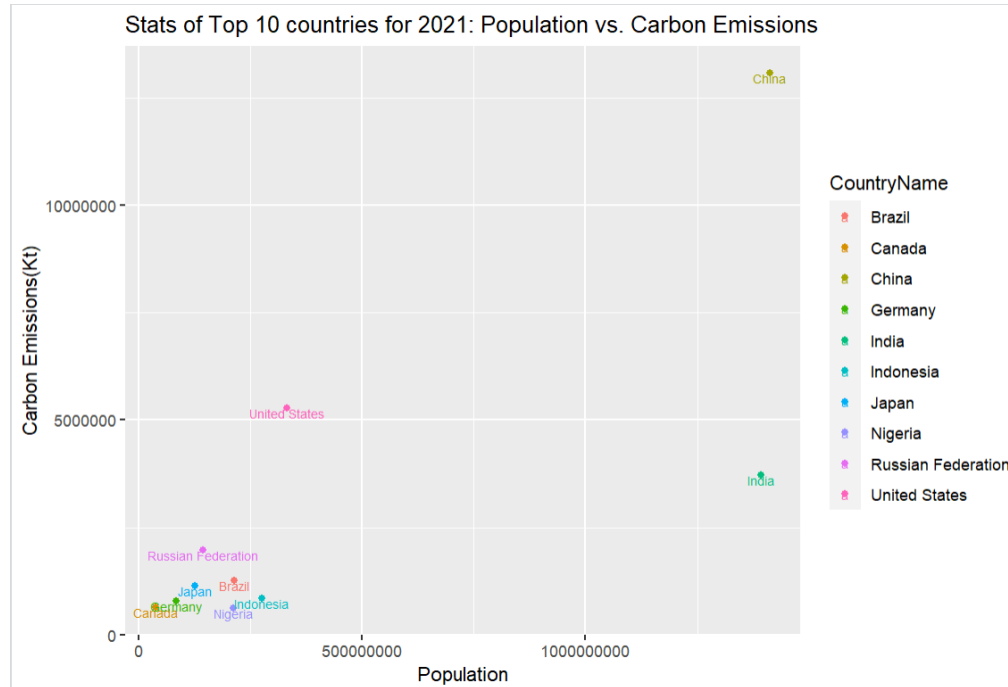

Figure 3.3.3: *Stats of Top 10 countries for 2021: Population vs. Carbon Emissions*

Figure 3.3.3 plots population against carbon emissions from Top 10 carbon emitting countries in 2021. This further proves the findings in Figure 3.2.1 that higher the population, higher the carbon emissions. The number 1 carbon emitter China is also the highest populated nation.

3.4. Correlation between population density and Carbon Emissions per capita
   (Refer *Section 3.4: Correlation between population density and Carbon Emissions per capita* in code)
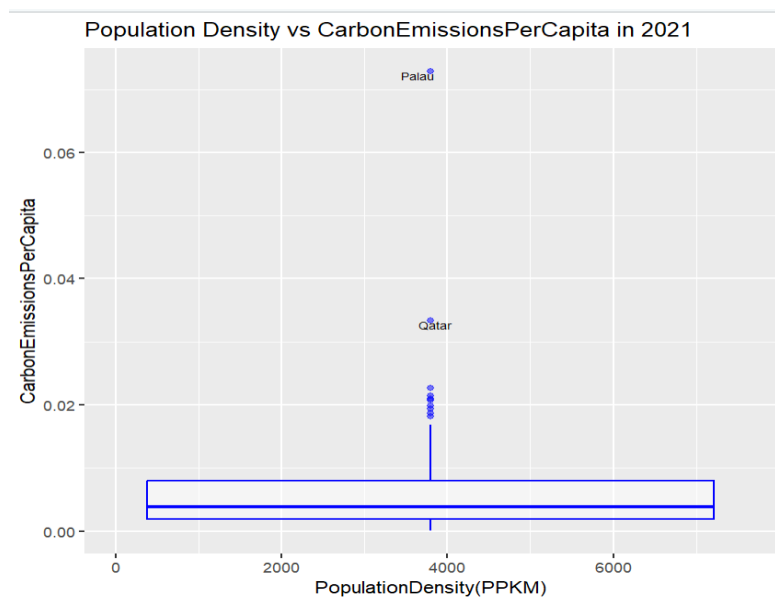

Figure 3.4.1: *Population Density vs Carbon Emissions Per Capita in 2021*

Figure 3.4.1 illustrates a boxplot plotted between the population density (measured in person per sq.km) and Carbon Emissions per capita (measured in kT per person) in year 2021. This plot shows that for the majority of countries the carbon emissions per capita is less than 0.02 and the mean value is 0.0065. However, there are few outliers such as Palau and Qatar with very high carbon emissions per capita.

4. **Conclusion**

After analysis of the global carbon emissions data from 1970 to 2021 in around 200 countries. Few findings that can be derived are:

- Top contributors of global carbon emissions are China, USA, India and Rusia.
- Population of the world and the global carbon emissions has been increasing linearly over the suggesting that there is a proportional and linear relationship between the two variables.
- The carbon emission in China and India has seen an exponential rise as years goes by, while USA and Russia have a more steadier carbon emission throughout the years.
- A relationship between area of the country and the carbon emissions cannot be derived decisively.
- There are few outlier countries with carbon emissions per capita, but most of the countries have carbon emissions per capita less than 0.02.