MM916 Project 1 (2023): CO2 emissions by European cities

Name – Hrutuja Patkar

Student ID: 202389142

Table of Contents

Exploratory Analysis for CO2 emissions by European cities:

In this report, we conduct an Exploratory Data Analysis (EDA) to unveil insights and patterns in the GCoM dataset, providing a foundational understanding for further analysis and hypothesis testing.

Element 1: A dataset overview

The original dataset contained 8,140 observations. After filtering out rows in the dataset with missing values in either "emissions per capita" or "population," 7,765 observations were retained.

- The number of cities and countries represented in the data are 7755 and 28 respectively.
- Italy (Country code – it) has the most count of cities (4037), whereas Estonia (Country Code – ee) has the fewest count of cities (5).
- Italy has city count of 4037, which constitutes 52% of the data. This clearly shows an imbalance in the data as more than half of the data comprises of cities from Italy. Also, the second highest count of cities is from Spain country (about 22%) and hence we can say that these two countries are overrepresented in the data.

Element 2: Range and distribution of city populations

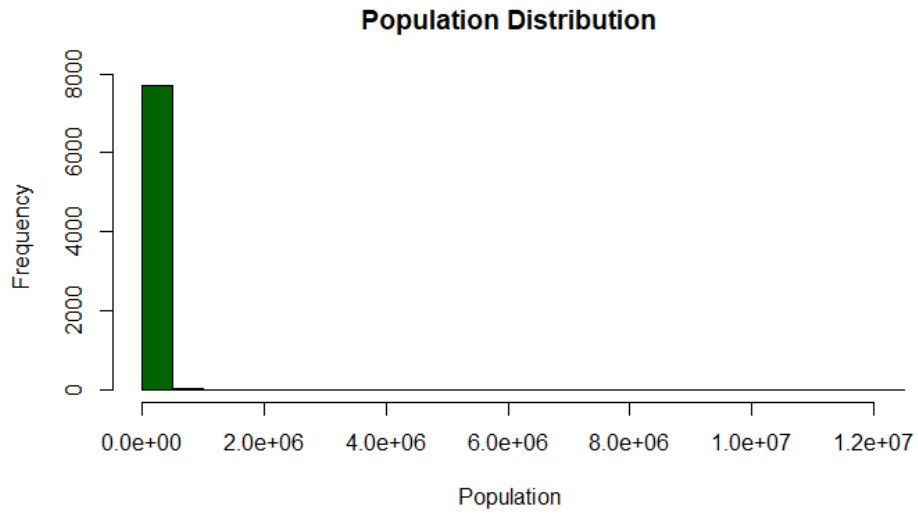

Figure2.a - Histogram of population distribution across cities

Figure 2.a exhibits histogram with a right-skewed distribution. Data distributions can be skewed due to the presence of extreme values or outliers, which disproportionately impact the shape of the distribution, leading to skewness.

The histogram suggests the need for a suitable transformation to improve symmetry and enhance the effectiveness of data analysis.

A log scale is employed to address the right-skewness in the data distribution, aiming to spread out the extreme values and achieve a more symmetric and interpretable representation.

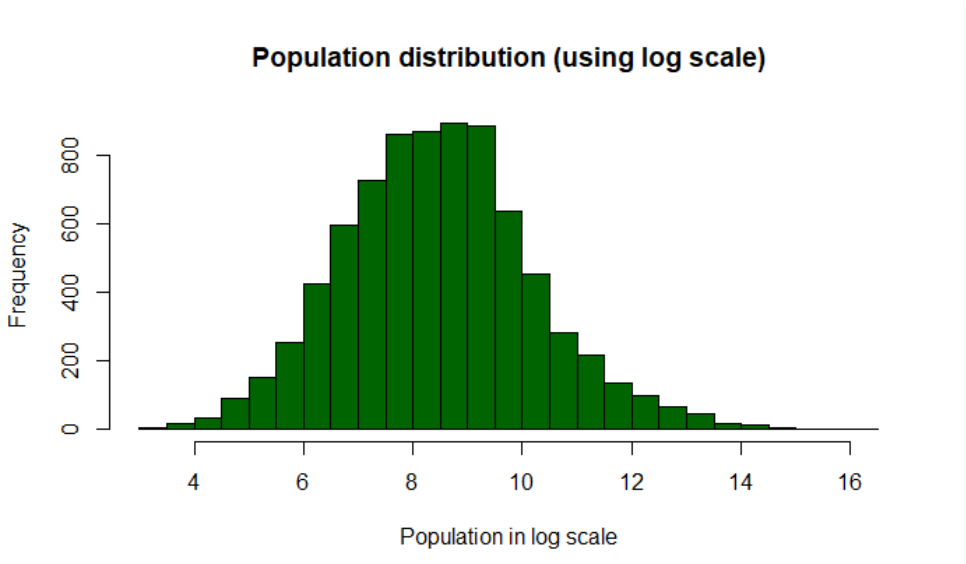

Figure 2.b - Histogram of population distribution across cities using log scale

Population Statistics:

• Maximum Population: London city has the maximum population, with a count of 12,051,223 residents.

• Minimum Population: Lobera de Onsella claims the minimum population, with a mere 28 residents.

• Median Population: Sharing the exact median, Realmonte and Predazzo both have a median population of 4,540 residents.

Element 3: Emissions by country

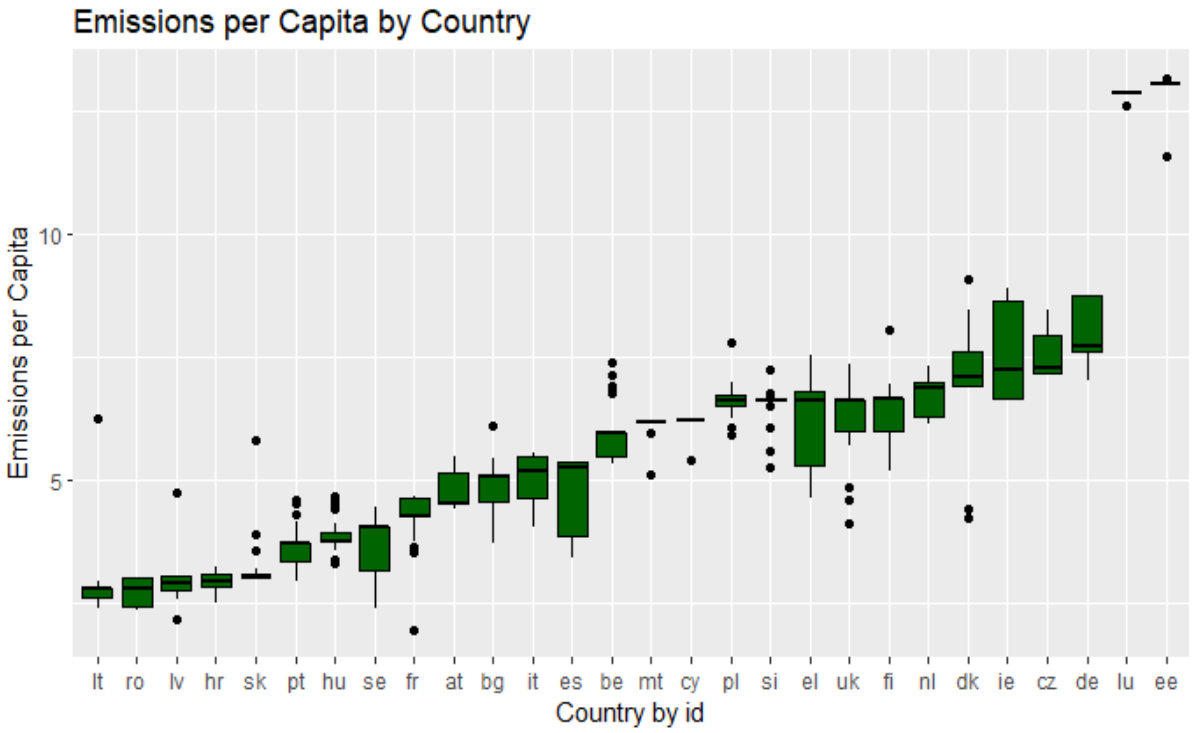

Figure 3.a Emissions per capita by country (according to median)

From Figure 3.a, Italy (it) and Romania (ro) have the lowest median carbon emissions per capita. This suggests that, on average, residents of these countries have relatively lower carbon emissions per capita compared to many others.

On the other hand, Luxembourg (lu) and Estonia (ee) have the highest median carbon emissions per capita. This indicates that residents in these countries, on average, have higher carbon emission per capita.

The large size of the box plots for Ireland (ie), Greece (el), and Spain (es) indicates substantial variability in emissions per capita within these countries. This means presence of outliers in these countries suggests that there are areas with exceptionally high or low emissions compared to the average.

Median Emissions per Capita statistics:

Table 3.a Top three countries by median commission per Capita

Country Id	Country	Median Value
ee	Estonia	13.06
lu	Luxembourg	12.87
de	Germany	7.73

Table 3.b - Bottom three countries by median Emission per Capita

Country Id	Country	Median Value
lv	Latvia	2.93
ro	Romania	2.80
it	Italy	2.79

Element 4: Emissions by sector

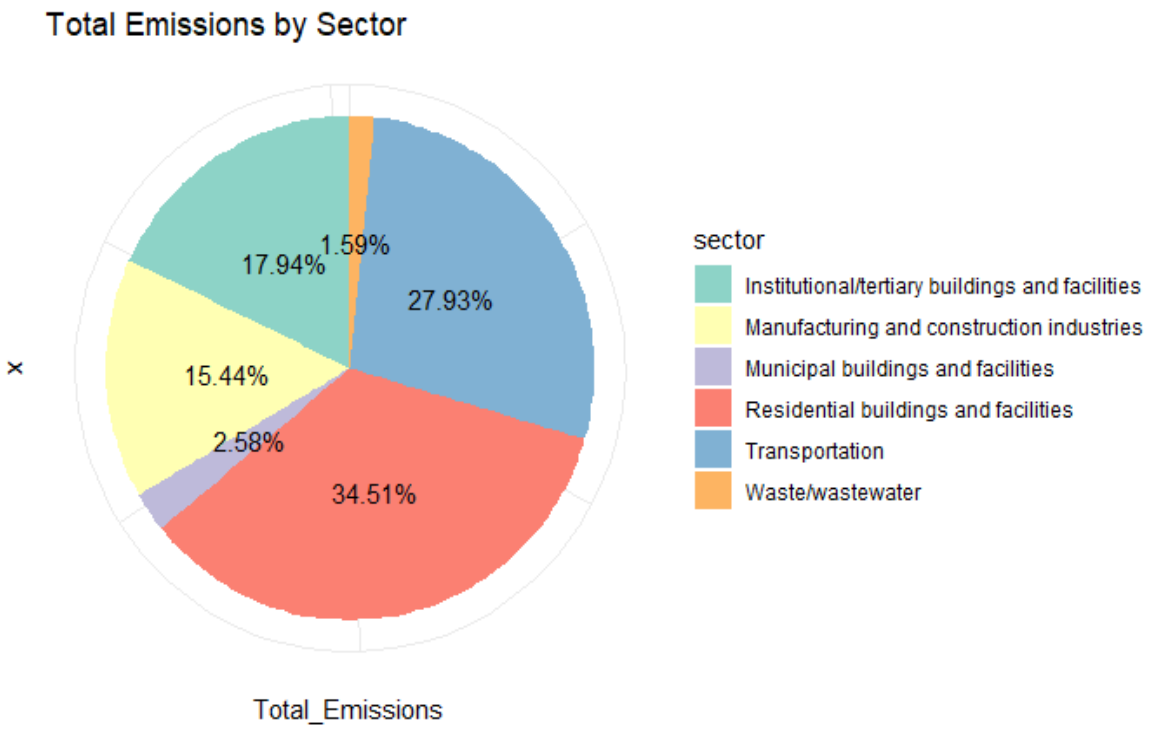

Figure 4.a - Percentage of contribution to Emissions by sector

Figure 4.a pie chart shows that the top two sectors responsible for highest contribution to carbon emission, which are Residential buildings and facilities (34.51%) and Transportation (27.93%).

Also, Waste/waste water and Municipal buildings and facilities sector contributes less in comparison to other sectors.

Element 5: Emissions by sector & country

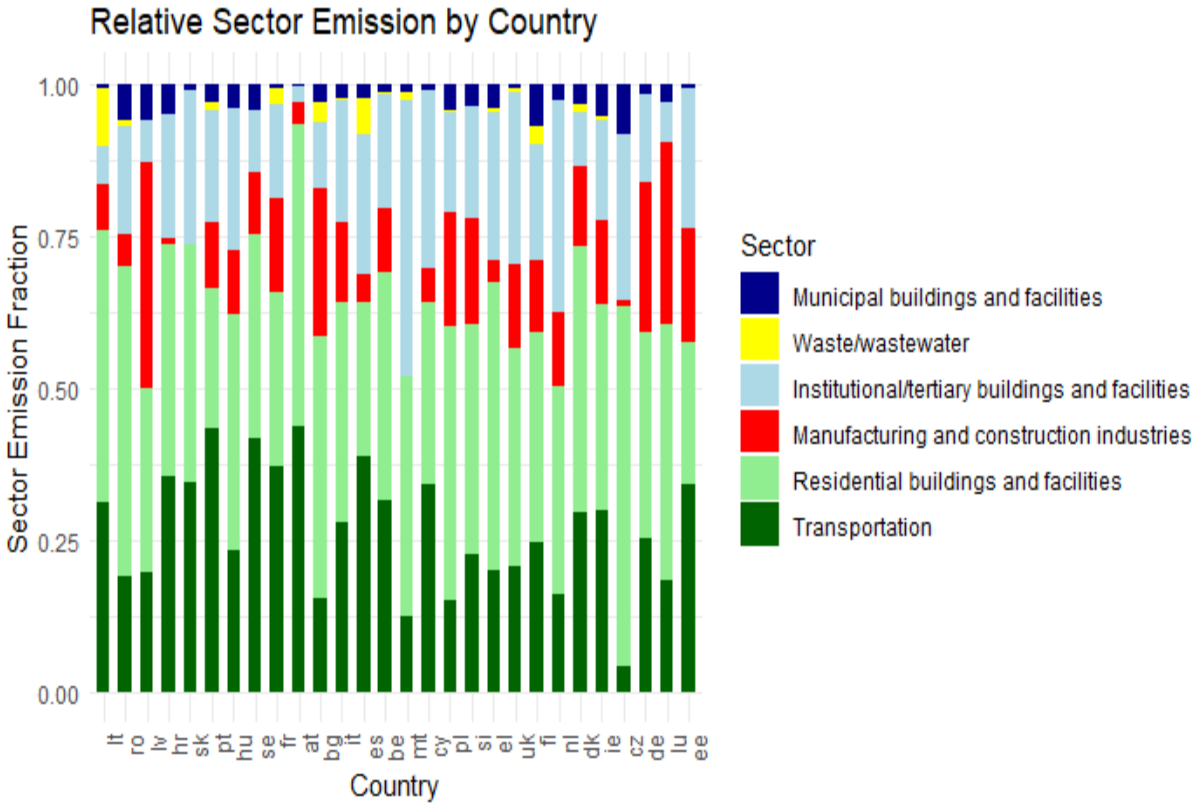

Figure 5.a - Sector Emission by Country to find relative importance of sector by country

Figure 5.a demonstrate the transportation sector is the largest source of emissions in most countries, followed by residential buildings and facilities. Both of them together constitute to more than half of the total fraction for all sectors. While, Waste/waste water contribution is the lowest to carbon emission.

France (fr) witness 90% of carbon emissions from the transportation and residential buildings & facilities sector. Belgium has approximately 50% of carbon emission footprints coming from Institution/tertiary buildings and facilities.

Element 6: Connecting emissions to heating demand

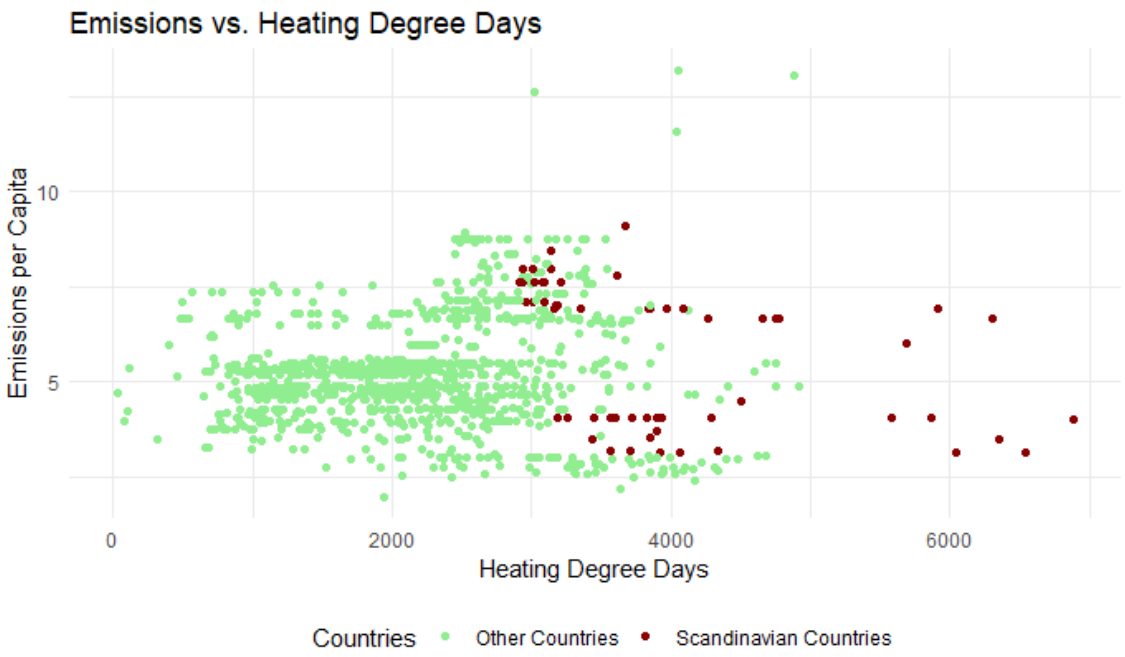

Figure 6.a - Scatter plot for Emission per Capita vs Heating Degree Days highlighting Scandinavian countries

Figure 6.a depicts comparison of carbon emissions per capita across heating degree day. For Scandinavian countries ((Sweden, Norway, Finland, Denmark), even if heating degree days goes on increasing there is no observed increase in carbon emissions. Generally, as Scandinavian countries are colder it is expected that emissions will be more as more heating is required. But with this scatter plot, we can interpret that this hypothesis is not true.

Element 7: Connecting emissions to wealth

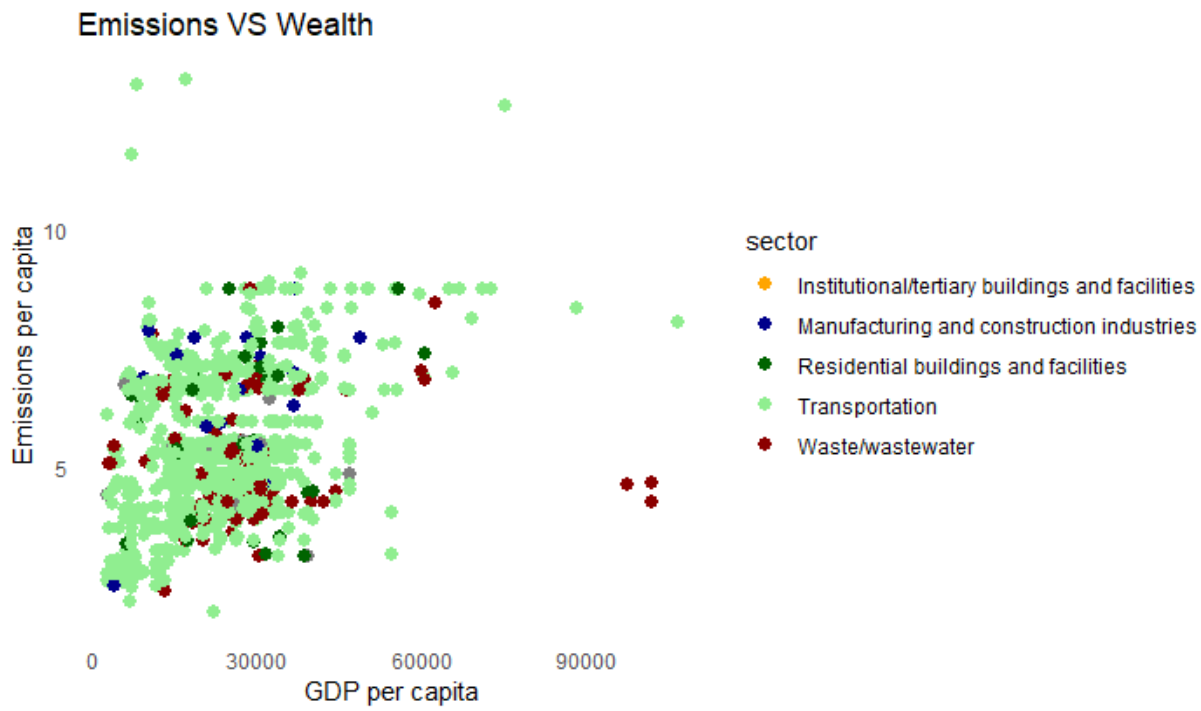

Figure 7-a Scatterplot for emissions per capita vs GDP per capita

In our exploration of the emissions-health relationship through the scatterplot Figure7.a it's important to avoid a hasty conclusion that countries with higher GDPs are the primary contributors to CO2 emissions due to their energy consumption.

London city has been removed as it's an outlier city with highest GDP per capita of 407447.4

Visual analysis reveals a striking observation – approximately 80% of the data points are clustered in the lower-left corner. This suggests that even nations with lower GDPs are significant emitters. Interestingly, among countries with incomes below $30,000, those with the lowest incomes tend to produce the highest emissions per person. Thus, this doesn't support the hypothesis that wealthier countries use more energy and therefore produce more CO2 emissions.

Furthermore, the introduction of the Sector variable provides a fresh perspective on emissions relative to GDP. Notably, in countries with lower income levels, it becomes evident that the transportation sector wields the most substantial influence on emissions.

Element 8: Summary and recommendations

In this exploratory analysis, we've delved into the intriguing world of carbon emissions, shedding light on the diverse factors that influence emissions across sectors and countries. The data reveals a nuanced tapestry of emissions, with each sector playing a distinct role in shaping the environmental landscape.

When examining countries, Italy stands out with its high representation in the dataset, raising questions about its unique emissions patterns. Estonia, despite a smaller number of cities, exhibits notably high median carbon emissions per capita, prompting further investigation into the country's emissions profile.

Among the sectors, residential buildings and facilities and transportation emerge as the primary contributors to carbon emissions, collectively accounting for more than half of the total emissions. This underscores the significance of improving energy efficiency and adopting sustainable practices in these sectors.

France's emissions composition highlights the significant role of the transportation and residential buildings & facilities sectors. Similarly, Belgium's emissions landscape, with roughly 50% attributed to the Institution/Tertiary Buildings and Facilities sector, warrants in-depth analysis.

Ireland, Greece, and Spain exhibit substantial variability in emissions per capita due to the presence of outliers, prompting a closer look at regions within these countries with exceptional emissions patterns.

This exploratory analysis provides a valuable starting point for more in-depth investigations into the complex world of carbon emissions and sustainability practices. To further our understanding, future analyses should consider additional variables, such as energy sources, urban planning, and public policies. These variables may elucidate why certain cities or countries exhibit exceptionally high or low emissions, paving the way for more targeted environmental strategies and a deeper comprehension of the intricate interplay of factors influencing emissions.

It is hypothesized that an in-depth examination of the energy sources used within different sectors and countries can provide crucial insights. The transition to renewable energy sources, such as solar, wind, and hydroelectric power, may contribute significantly to emissions reduction. Future research should explore the energy mix within each sector and its impact on carbon emissions.

Appendix

```
require(tidyverse)

df = read_csv("GCoM_emissions.csv") %>%
  rename(id = 'GCoM_ID',
         city = 'signatory name',
         country = 'country code',
         hdd = 'Heating Degree-Days (HDD)',
         gdp_pc = 'GDP per capita at NUTS3 [Euro per inhabitant]',
         emissions_pc = 'GHG emissions per capita in GCoM sectors_EDGAR [tCO2-eq/year]',
         population = 'population in 2018') %>%
  select(id, city, country, hdd, gdp_pc, emissions_pc, population)

dfs = read_csv("GCoM_emissions_by_sector.csv") %>%
  rename(id = 'GCoM_ID',
         sector = 'emission_inventory_sector') %>%
  select(id,sector,emissions)

#Element 1: A dataset overview.
# Filtering out rows in the GCoM_emissions where emissions per capita or population is missing
df1 <- na.omit(df)

#The number of cities and countries represented in the data:
# No. of distinct cities
distinct_cities <- df1 %>%
  summarise(Cities = n_distinct(city))
# No. of distinct countries
distinct_countries <- df1 %>%
  summarise(Countries = n_distinct(country))
```

```
#the names of the countries with the most and fewest cities:
country_wise_city_counts <- df1 %>%
  group_by(country) %>%
  summarize(
    Cities = n_distinct(city)) %>% arrange(desc(Cities))
#Country with the most cities
country_with_most_cities <- country_wise_city_counts %>%
  filter(Cities == max(Cities))
#Country with the fewest cities
country_with_fewest_cities <- country_wise_city_counts %>%
  filter(Cities == min(Cities))
country_representation <- rbind(head(country_wise_city_counts, 5))

#Element 2: Range and distribution of city populations.
#Plotting histogram:
hist_plot <- hist(df1$population,
                  breaks = 20,
                  main = "Population Distribution",
                  xlab = "Population",
                  ylab = "Frequency",
                  col = "darkgreen",
                  border = "black",
                  xlim = c(0, max(df1$population)))

#applying log transformation
hist(log(df1$population),
     main = "Population distribution (using log scale)",
     xlab = "Population in log scale",
     ylab = "Frequency",
     col = "darkgreen",
     border = "black",
```

```
    breaks = 20
)
#Maximum and minimum city populations in the dataset (and the corresponding city names),
max_population <- max(df1$population)
city_with_max_population <- df1[df1$population == max_population, "city"]
min_population <- min(df1$population)
city_with_min_population <- df1[df1$population == min_population, "city"]

#Median population and the city
median_population <- median(df1$population)
city_with_median_population <- df1[df1$population == median_population, "city"]

#Element 3: Emissions by country.
#Plotting boxplot
country_emission <- df1
country_emission$country <- reorder(country_emission$country, country_emission$emissions_pc,
FUN = median)
box_color <- "darkgreen"
plot_title <- "Emissions per Capita by Country"
boxplot_emissions_by_country <- ggplot(country_emission, aes(x = country, y = emissions_pc)) +
 geom_boxplot(fill = box_color, color = "black") +
 labs(x = "Country by id",
      y = "Emissions per Capita") +
 ggtitle(plot_title)
 theme_minimal() +  # Apply a minimal theme
 theme(panel.background = element_rect(fill = "grey"))
boxplot_emissions_by_country

#Top 3 and bottom 3 countries by median emissions per capita
countries_by_median <- df1 %>% group_by(country) %>% summarise (median =
median(emissions_pc)) %>% arrange((median))
top3<- countries_by_median %>% top_n(3)
```

```
bottom3<- countries_by_median %>% top_n(-3)
```

#Element 4: Emissions by sector.

```
# Calculate the total emissions for each sector
library(ggplot2)
library(dplyr)
library(scales)

df3 <- na.omit(dfs)
sector_totals <- df3 %>%
  group_by(sector) %>%
  summarize(Total_Emissions = sum(emissions))

# Sort the sectors by total emissions in descending order
sector_totals <- sector_totals %>%
  arrange(desc(Total_Emissions))

# Create a pie chart
ggplot(sector_totals, aes(x = "" ,y = Total_Emissions, fill = sector)) +
  geom_bar(stat = 'identity', width = 5) +
  coord_polar(theta = "y") +
  labs(title = "Total Emissions by Sector") +
  theme_minimal() +
  scale_fill_brewer(palette = "Set3")+
  geom_text(aes(label =percent(Total_Emissions/sum(Total_Emissions)), x= 1),position =
position_stack(vjust=0.5))+
  theme(axis.text.x = element_blank())# Choose a color palette

#the sectors are responsible for the most emissions
max_emission_by_sec <- sector_totals %>% top_n(2)
```

#Element 5: Emissions by sector and country.

```
# Join the datasets using the city ID
combined_data <- df1 %>%
  inner_join(df3, by = c("id" = "id"))

# Calculate the fraction of emissions in each sector for each country
combined_data <- combined_data %>%
  group_by(country) %>%
  mutate(sector_fraction = emissions / sum(emissions))
combined_data$sector <- reorder(combined_data$sector, combined_data$sector_fraction)

# Create a stacked bar plot
stacked_plot <- ggplot(combined_data, aes(x = country, y = sector_fraction, fill = sector)) +
  geom_bar(stat = "identity", width = 0.60) +
  labs(x = "Country", y = "Sector Emission Fraction", fill = "Sector", title = 'Relative Sector Emission by Country') +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  scale_fill_manual(values = colorRampPalette(c("darkblue",'yellow','lightblue', 'red', 'lightgreen',"darkgreen"))(length(unique(combined_data$sector))))

# Show the stacked bar plot
stacked_plot
```

#Element 6: Connecting emissions to heating demand

```
#wrangling data
element_6 <- data.frame(combined_data$country,combined_data$city,combined_data$hdd,combined_data$emissions_pc) %>%
  mutate(Is_Scandanavian= ifelse(combined_data$country %in% c("se", "no", "fi", "dk"),TRUE,FALSE))%>%
```

```
  rename(country='combined_data.country',
city='combined_data.city',hdd='combined_data.hdd',emissions_pc='combined_data.emissions_pc',
Countries ='Is_Scandanavian')%>%
  select(country,city,hdd,emissions_pc,Countries)%>%
  distinct()
#Creating scatterplot
ggplot(element_6, aes(x = hdd, y = emissions_pc, color = Countries)) +
  geom_point() +
  labs(title = "Emissions vs. Heating Degree Days",
       x = "Heating Degree Days",
       y = "Emissions per Capita") +
  scale_color_manual(values = c("lightgreen", "darkred"),
                     breaks = c(FALSE, TRUE),
                     labels = c("Other Countries", "Scandinavian Countries")) +
  theme_minimal()+
  theme(legend.position = 'bottom')
```

#Element 7: Connecting emissions to wealth.

```
emission_by_sector_country_joined <- df %>% full_join(dfs, by= 'id')
#clean up the dataset
combined <- na.omit(emission_by_sector_country_joined)
outlier_city <- combined%>% filter(gdp_pc == max(gdp_pc))
df_scaled <- combined %>% filter(city != outlier_city$city)
# Create the scatterplot
ggplot(data = df_scaled, aes(x = gdp_pc, y = emissions_pc)) +
  geom_point() +
  labs(
    x = "GDP per capita",
    y = "Emissions per capita",
  ) +
  theme_minimal()
#Scatterplot of GDP per Capita vs. Emissions per Capita with Sectors
```

```
max_emissions_data <- df_scaled %>% filter(emissions_pc == max(emissions_pc))

max_gdp_data <-df_scaled %>% filter(gdp_pc == max(gdp_pc))

min_emissions_data <- df_scaled %>% filter(emissions_pc == min(emissions_pc))

min_gdp_data <-df_scaled %>% filter(gdp_pc == min(gdp_pc))


ggplot(data = df_scaled, aes(x = gdp_pc, y = emissions_pc, color = sector)) +
  geom_point(size = 2.5) +
  labs(
    x = "GDP per capita",
    y = "Emissions per capita",
    title = 'Emissions VS Wealth'
  ) +
  scale_color_manual(values = c("Residential buildings and facilities" = "darkgreen", "Transportation" =
"lightgreen", "Institution/tertiary buildings and facilities" = "red", "Manufacturing and construction
industries" = "darkblue", "Institutional/tertiary buildings and facilities" = "orange",
"Waste/wastewater" = "darkred")) +
  theme_minimal()+
  theme(panel.grid = element_blank())
```