MM916 Project 1 (2023)
CO2 emissions by European cities

Contents

List of Figures

List of Tables

Element 1: A dataset overview. *Filter out rows in the GCoM_emissions.csv dataset where either the emissions per capita or population is missing. Give the number of cities and countries represented in the data, and the names of the countries with the most and fewest cities. (This information can be presented in a table or simply in sentences, as you prefer.) Do there seem to be any imbalances: are certain countries are overrepresented?*

- The dataset GCoM_emissions.csv, consists of 7765 observations and 7 variables after omitting the missing values.
- Table 1 displays the data of countries with the most and least number of cities within the said dataset.

Table 1: Countries with the most and least number of cities

Country	Cities	Description
Italy(it)	4037	Most number of cities
Estonia(ee)	5	Least number of cities

- Figure 1 visualises the city count against each of the countries represented in the dataset(GCoM_emissions.csv). This graph shows us that there are 4037 cities in Italy, accounting for more than 50% of the total dataset. Spain comes in second with emission entries of approximately 25%, indicating that there may be an overrepresentation of data for Italy followed by Spain, relative to the percentage contributions of the other countries included in this dataset.

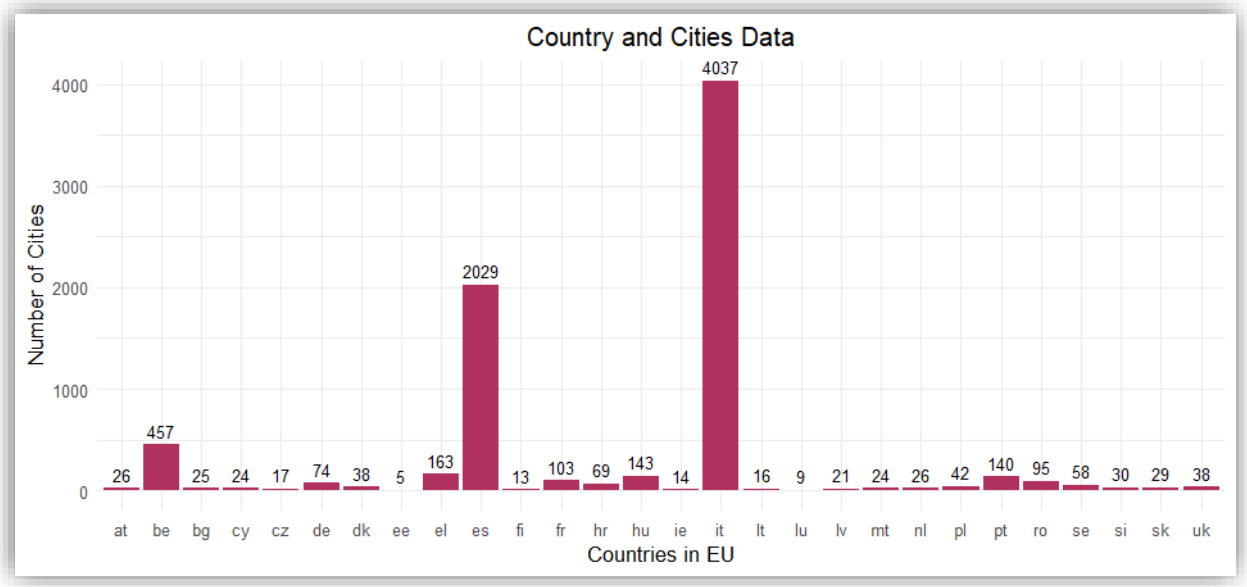

Figure 1: Bar Graph Representation of Country and its respective Cities

Element 2: Range and distribution of city populations. *Make a histogram of population. What are the maximum and minimum city populations in the dataset (and the corresponding city names), and the median?*

- Figure 2 shows a histogram of the population attribute, while Figure 3 shows a histogram of the population attribute on a logarithmic scale.
- The city with the highest population of 12051223 is London.
- The city with the lowest population of 28 is Lobera de Onsella
- Median population of the cities is 4540. The cities with the median population are Realmonte and Predazzo in Italy.

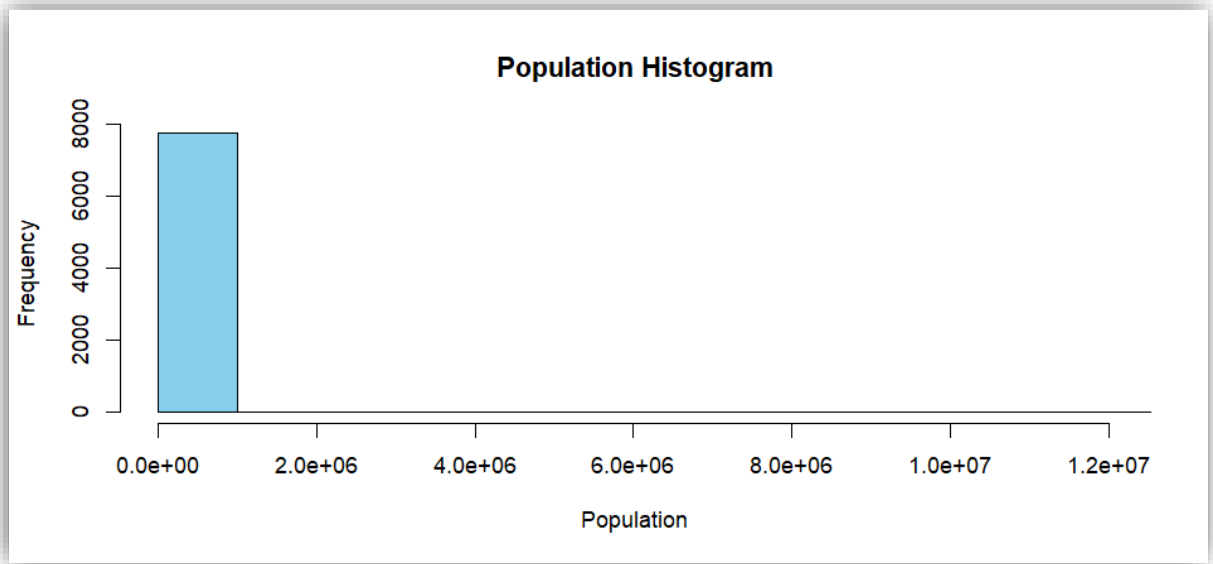

Figure 2: A histogram of Population.

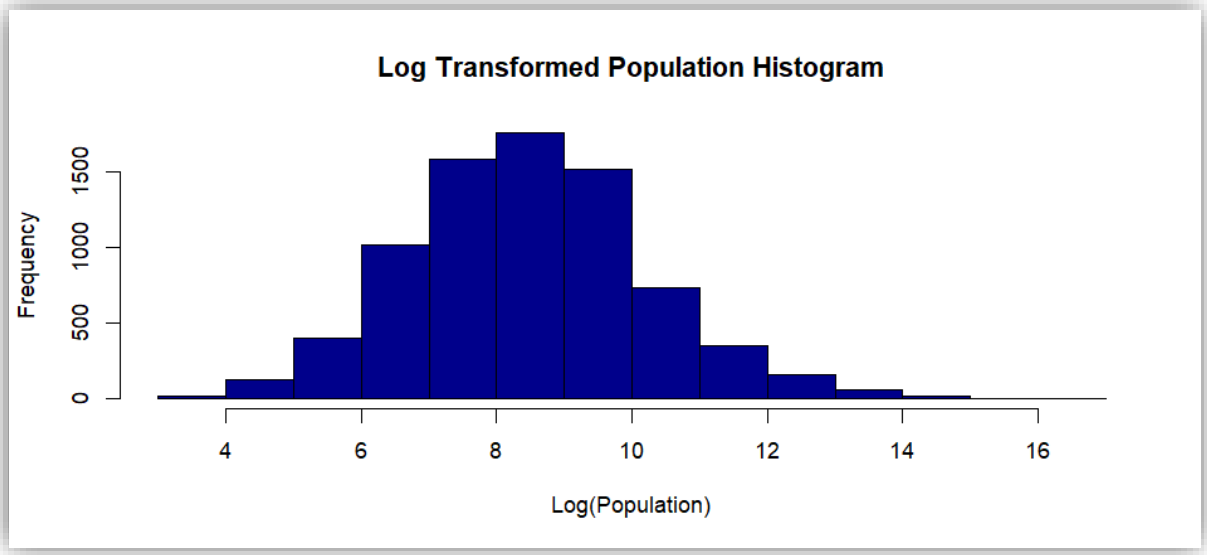

Figure 3: A histogram of Population in logarithmic scale.

Element 3: Emissions by country. *Make a boxplot or similar plot that shows emissions per capita for each country. For full marks, display the countries in a sensible order. Report the top 3 and bottom 3 countries by median emissions per capita (get R to identify these countries for you).*

- Figure 4 displays the relation between emissions per capita for respective countries arranged from the least emissions released to the extreme left of the graph to the countries on the extreme right with most emissions per capita.

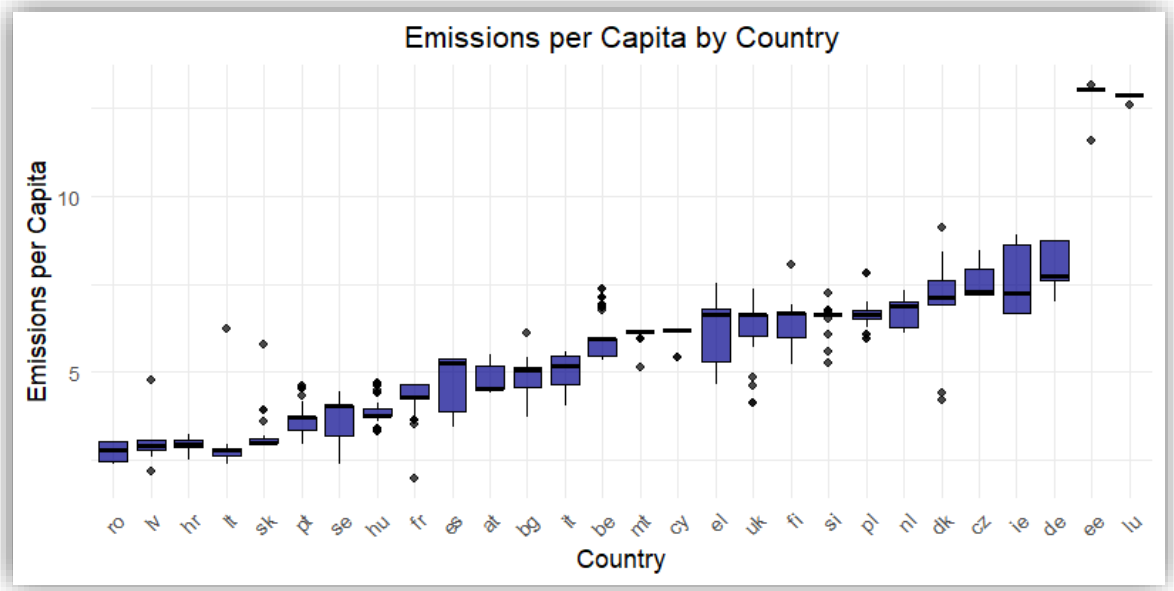

Figure 4: Box Plot highlighting the emission per capita for each country.

- Highlighted in Table 2, Latvia followed by Romania and Lithuania with 2.93, 2.80 and 2.79 emissions per capita respectively, are the bottom three countries as per median emissions per capita. Furthermore, Estonia followed by Luxembourg and Germany with 13.06, 12.87 and 7.73 emissions per capita respectively, are the top three countries as per median emissions per capita.

Country	**Median Emission (Bottom 3)**	**Country**	**Median Emission (Top 3)**
Latvia (lv)	2.93	Estonia (ee)	13.06
Romania (ro)	2.80	Luxembourg (lu)	12.87
Lithuania (lt)	2.79	Germany (de)	7.73

Table 2: Median Emissions of the Top 3 and Bottom 3 countries.

Element 4: Emissions by sector. *Using the data in GCoM_emissions_by_sector.csv, make a plot of the total emissions for each of the six sectors (residential buildings, etc.). Which of the sectors are responsible for the most emissions?*

- Table 3 provides a detailed breakdown of sector-wise emissions, excluding any entries with missing values, based on the data from GCoM_emissions_by_sector.csv.

Table 3: Breakdown of Sector wise emissions.

Sectors	Total
Institutional/tertiary buildings and facilities	272323312
Manufacturing and construction industries	234423400
Municipal buildings and facilities	39149517
Residential buildings and facilities	523853738
Transportation	423878261
Waste/wastewater	24175790

- Figure 5 depicts a graphical representation of sector-wise emissions, highlighting that the Residential Buildings and Facilities sector stands out with the highest emissions, totalling 523,853,738.

Figure 5: A visual representation of total emissions generated across six sectors.

Element 5: Emissions by sector and country. *Now join the two datasets (using city ID) so that you can associate city names and countries with the by-sector data. Make a plot showing how the relative importance of the six sectors varies by country. (One nice way to do this is a stacked bar plot giving the fraction of each country's emissions in each sector, but there are many ways to approach it.)*

- By joining GCoM_emissions_by_sector.csv and GCoM_emissions.csv using the city ID, a new dataset generated had 57102 obervations and 9 variables after omitting the missing values.
- Figure 6 illustrates the distribution of emissions by sector for each country. It is evident that there is a significant emphasis on emissions from Residential buildings & facilities and the Transportation sector.

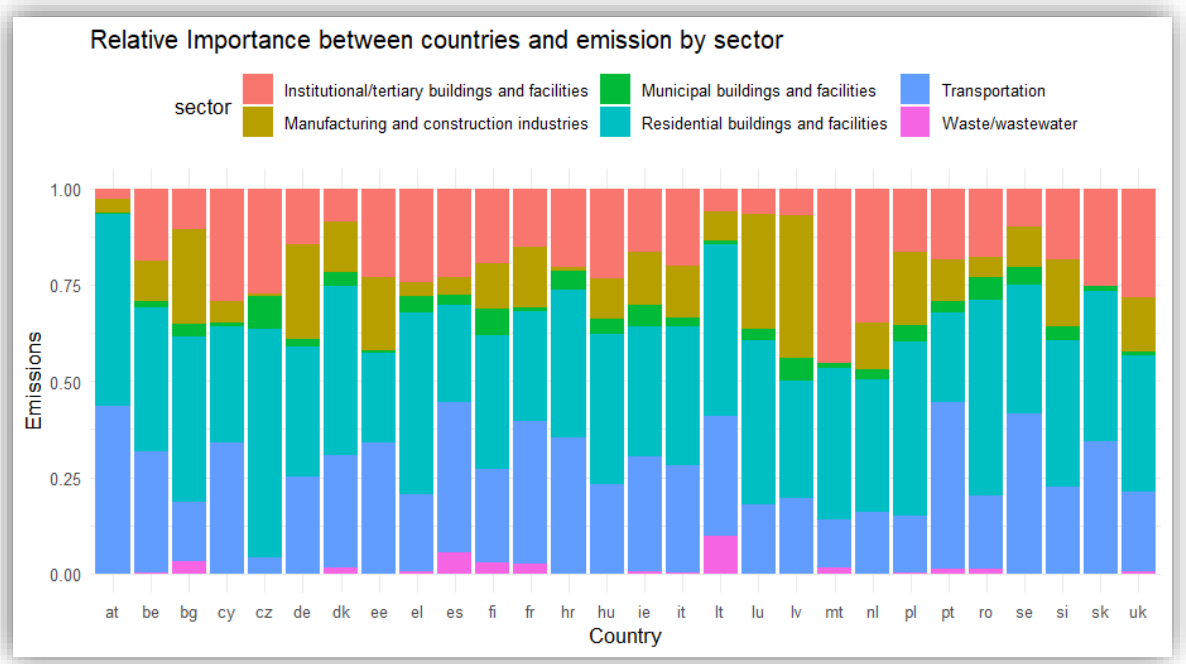

Figure 6: Relative Importance of emissions per sector of each country.

Element 6: Connecting emissions to heating demand. *One might hypothesize that emissions are higher in colder cities where more energy is required for heating. Make a scatter plot that helps evaluate the possible link between Heating Degree Days and emissions per capita. Use either total emissions by city, or emissions in one of the six sectors that seems appropriate. Highlight the Scandanavian cities (Sweden, Norway, Finland, Denmark) in a second colour: one might expect these countries to have especially high heating needs and therefore especially high emissions. Based on visual inspection of the plot, does this hypothesis seem likely to be true?*

- Figure 7 graph displays a scatter plot between emissions per capita and the heating degree days.
- As per graph, it can be observed that in case of the Scandinavian countries highlighted in dark blue colour, although the heating degree days are similar or higher that that of the other countries, the emissions per capita are similar or as close as the other countries.
- However, there are a few non-Scandinavian countries with very high emissions compared to their heating degree days.
- As a result, it is clearly observed from the visual representation that the hypothesis of emissions per capita to be higher when heating degree days are high is invalid.

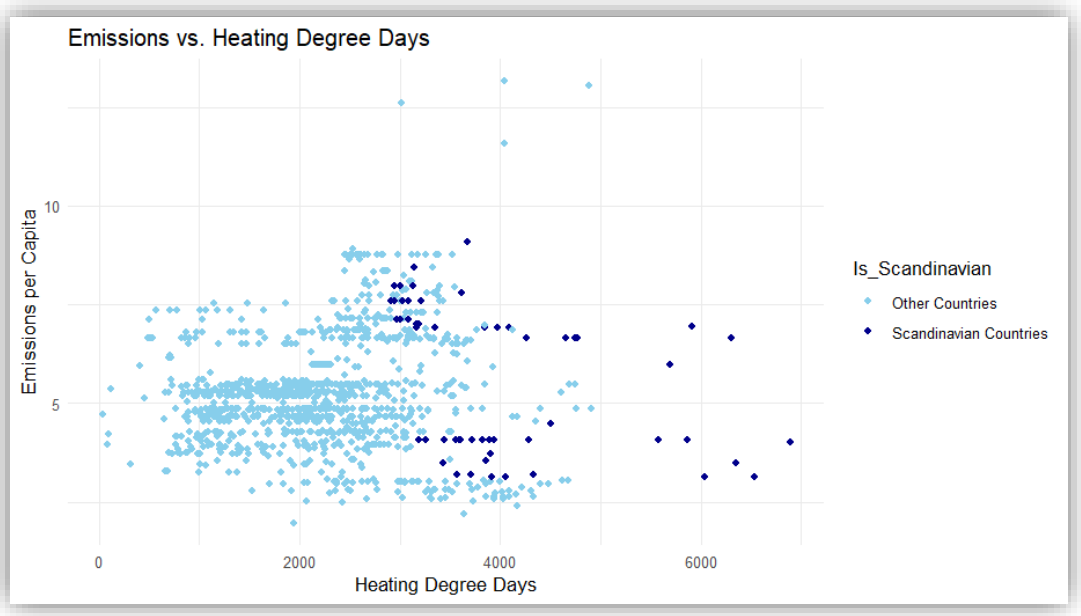

Figure 7: Scatter Plot showing the relation between Heating Degree Days and Emissions per Capita for all Countries.

Element 7: Connecting emissions to wealth. *An alternate hypothesis is that wealthier countries use more energy and therefore produce more CO2 emissions. Make a scatterplot that helps you examine the relationship between emissions per capita and GDP per capita, removing one outlier city with very high GDP per capita (name which city it is). Use colour and symbol type to include other variables if it helps you evaluate further hypotheses. What does this plot tell you?*

- London has a GDP of 407447.37 which is the highest in this dataset.
- Scatterplot displays in Figure 8 depicts the relation between GDP per capita and Emissions per capita across various sectors.
- The hypothesis that if a city or country with high GDP will generate higher emissions is invalid as observed in Figure 8 wherein even with the GDP increasing, the emissions don't see a similar increase and remain mostly unchanged.

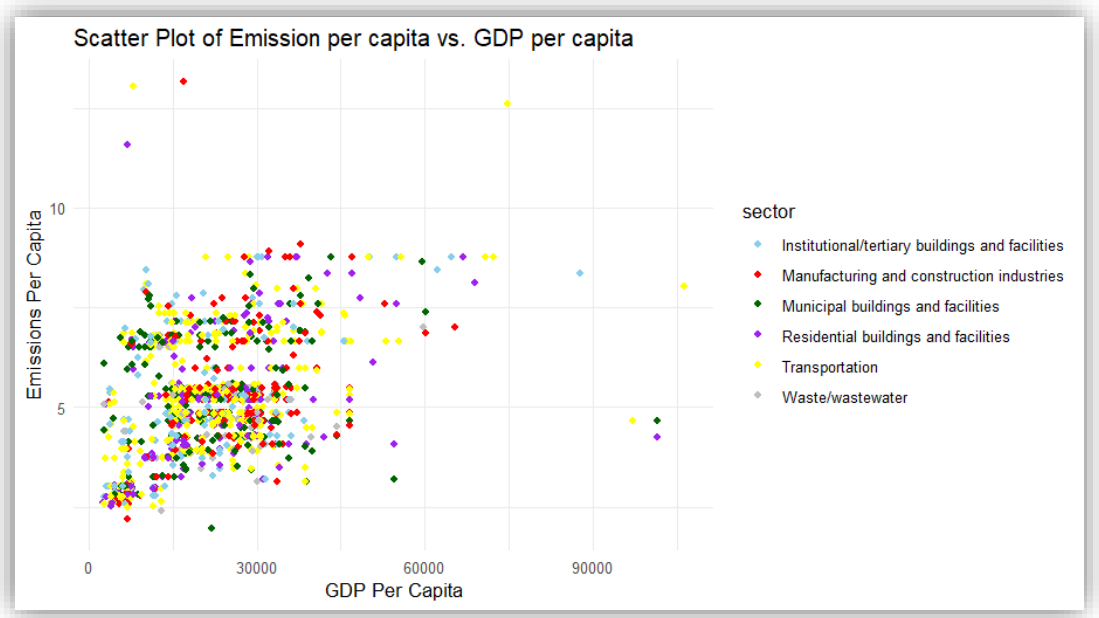

Figure 8: Scatterplot to show the relation between GDP per capita and Emissions per capita

Element 8: Summary and recommendations. *Conclude with a short paragraph summarising what you have learned from this exploratory analysis, and what hypotheses appear to be most promising for future analysis. If you feel other variables would need to be included in the analysis to find an explanation of why some cities or countries have especially high or low emissions, speculate on what data would be useful. (Note that in a real published report, this summary paragraph would probably go in your Discussion section, but for this project you don't need to break your writing into sections.)*

An exploratory data analysis was required to be done on two different datasets which included

- Data clean-up by removing missing values;
- Creation of new datasets for running various analysis;
- Merging of two datasets for further analysis;
- Creating plots such as boxplots, bar graphs, stacked bar graphs and scatter plots

This helped us in visualizing the emissions generated and its possible causes. We also learnt that the dataset exhibits an overrepresentation of entries from specific countries, which could potentially introduce bias and impact the analysis and outcomes.

Additionally, we found that there was no positive correlation established between both the hypothesis that we tested which were

- emissions and gdp per capita
- emissions and heating degree days irrespective of country climate.

Instead, the hypothesis of population combined with emission by sector might have a positive correlation. However, for this to be possible, the dataset needs to be balanced in a way where no country is overrepresented similar to the current dataset we worked with.

Appendix

ELEMENT 1: A DATASET OVERVIEW.

#Step 1: Read the database in R and rename column names for better readability.

```
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)
```

#Step 2: Remove and filter out missing values from the dataset

```
filter_df <- na.omit(df)
```

#Step 3: Obtain the number of cities and countries in the dataset

```
cities <- length(unique(filter_df$city))
countries <- length(unique(filter_df$country))
```

#Step 4: Obtain the names of the countries with least and most cities

#Step 4.1: Check for country with least number of cities.

```
min_cities <- city_counts %>%
  top_n(-1,num_cities)

print(min_cities)
```

#Step 4.2: Check for country with most number of cities.

```
max_cities <- city_counts %>%
  top_n(1,num_cities)

print(max_cities)
```

#Step 5: Plot a bar graph to visualize imbalances and overrepresentation of countries

#Step 5.1: load the ggplot2 package

```
library(ggplot2)
```

#Step 5.2: Using ggplot, visualize the Country and City data to understand if any country is overrepresented through a bar graph.

```
ggplot(city_counts, aes(x = country, y = num_cities)) +
  geom_bar(stat = "identity", fill = "maroon") +
  geom_text(aes(label = num_cities), vjust = -0.5, size = 3) +
  labs(title = "Country and Cities Data", x = "Countries in EU", y = "Number of Cities")
+
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```

ELEMENT 2: RANGE AND DISTRIBUTION OF CITY POPULATIONS

Step 1: Plot a histogram for population

```
populationhist <- hist(filter_df$population, xlim = c(28,12051223), col="skyblue",
main= 'Population Histogram', xlab = "Population", ylab = "Frequency")
```

OR

```
hist(log(filter_df$population), col = "darkblue", main= 'Log Transformed Population
Histogram', xlab = "Log(Population)", ylab = "Count")
```

#Step 2: Find out maximum population city

```
maximum_pop_city <- filter_df[which.max(filter_df$population), c("city", "population")]
```

#Step 3: Find out maximum population city

```
minimum_pop_city <- filter_df[which.min(filter_df$population), c("city", "population")]
```

#Step 4: Find out the population median and the city name

#Step 4.1 Find out the median

```
median_pop <- median(filter_df$population)
```

#Step 4.2: Find out the city(s) associated with median value.

```
filter_df %>%
  filter(population==4540)
```

ELEMENT 3: EMISSIONS BY COUNTRY

#Step 1: Create a boxplot for emissions per capita for each country. Plot it in a sensible order.

```
ggplot(filter_df, aes(x = reorder(country, emissions_pc), y = emissions_pc)) +
  geom_boxplot(fill = "darkblue", color = "black", alpha = 0.7) +
  labs(title = "Emissions per Capita by Country", x = "Country", y = "Emissions per
Capita") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

#Step 2: identity top and bottom three countries by median emission

#Step 2.1: create a dataset of emissions per capita and country to get median emissions

```
median_emissions <- filter_df %>%
  group_by(country) %>%
  summarise(median_emissions = median(emissions_pc)) %>%
  arrange(median_emissions)
```

#Step 2.2: Get top 3 cities by median emissions

```
top_3c <- filter_df %>%
  group_by(country) %>%
  summarise(median_emissions = median(emissions_pc)) %>%
  arrange(desc(median_emissions)) %>%
  head(3)
```

#Step 2.3: Get bottom 3 cities by median emissions

```
bottom_3c <- filter_df %>%
  group_by(country) %>%
  summarise(median_emissions = median(emissions_pc)) %>%
  arrange(median_emissions) %>%
  head(3)
```

ELEMENT 4: EMISSIONS BY SECTOR

#Step 1 : Read the second dataset in R and remove any missions values.

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

dfs1<- na.omit(dfs)
```

#Step 2: Make a plot of the total emissions for each of the six sectors

#Step 2.1: Get the sector wise emissions total

```
sector_emission_dfs1 <- dfs1 %>%
  group_by(sector) %>%
  summarise(total=(sum(emissions, na.rm = TRUE)))
```

#Step 2.2: Create a bar graph plot using ggplot.

```
library(stringr)
sector_emission_dfs1$sector <- str_wrap(sector_emission_dfs1$sector, width = 15)

plot_1 <- ggplot(sector_emission_dfs1, aes(x = reorder(sector, total), y = total, fill =
sector)) +
  geom_bar(stat = "identity", color = "yellow") +
  labs(title = "Emission Values by Sector",
       x = "Sectors", y = "Total Emissions") +
  theme_minimal() +
  theme(axis.text.x = element_text(vjust = 1.15, hjust = 0.5))

print(plot_1)
```

#Step 3: Identify the sector with the maximum emissions

```
max_sectoremission <-
sector_emission_dfs1[which.max(sector_emission_dfs1$total), c("sector", "total")]
```

ELEMENT 5: EMISSIONS BY SECTOR AND COUNTRY

```
# Step 1: Merge both datasets
newdata <- merge(df, dfs, by = "id", all.x=TRUE, all.y=TRUE)
mergeddata <- na.omit(newdata)

#Step 2: Find out sector contribution of emissions in each country
mergeddata <- mergeddata %>%
  group_by(country) %>%
  mutate(fraction = emissions / sum(emissions))

#Step 3: Create a stacked bar to display relative importance between countries and
emissions by sector
ggplot(mergeddata, aes(x = country, y = fraction, fill = sector)) +
  geom_col() +
  labs(title = "Relative Importance between countries and emission by sector",
       x = "Country",
       y = "Emissions") +
  theme_minimal() +
  theme(legend.position = "top")
```

ELEMENT 6: CONNECTING EMISSIONS TO HEATING DEMAND.

```
# Step 1: Create a new dataset to include city, heating degree days, emissions,
country and a new column to identify Scandinavian countries.
mergeddata_element6 <- mergeddata %>%
  mutate(Is_Scandinavian = ifelse(country %in% c("se", "no", "fi", "dk"), TRUE,
FALSE)) %>%
  distinct(country, hdd, emissions_pc, .keep_all = TRUE)%>%
  select(country, hdd, emissions_pc, Is_Scandinavian)

#Step 2: Create a scatter plot showing the relation between heating degree day,
emissions specifically highlighting Scandinavian countries from rest.
ggplot(mergeddata_element6, aes(x = hdd, y = emissions_pc, color =
Is_Scandinavian)) +
  geom_point() +
  labs(title = "Emissions vs. Heating Degree Days",
       x = "Heating Degree Days",
       y = "Emissions per Capita") +
  scale_color_manual(values = c("skyblue", "darkblue"),
                     breaks = c(FALSE, TRUE),
                     labels = c("Other Countries", "Scandinavian Countries")) +
  theme_minimal()

#Additional test to check hypothesis validity
corr1 <- cor(mergeddata_element6$hdd, mergeddata_element6$emissions_pc,
method = "pearson")
```

ELEMENT 7: CONNECTING EMISSIONS TO WEALTH

```
#Step 1: Remove Outlier City with maximum gdp and create a new dataset keeping
only required columns
max_gdpcity <- mergeddata [which.max(mergeddata$gdp_pc), c("city", "gdp_pc")]
```

```
mergeddata_elm7 <- mergeddata %>%
  filter(city != "London") %>%
  select(emissions_pc,gdp_pc,sector) %>%
  distinct()

#Step 2: Make a scatter plot to check relation between gdp and emissions per capita
#Step 2.1 Assign colour's to the sector.
sector_colors <- c("Institutional/tertiary buildings and facilities" = "skyblue",
                   "Manufacturing and construction industries" = "red",
                   "Municipal buildings and facilities" = "darkgreen",
                   "Residential buildings and facilities" = "purple",
                   "Transportation" = "yellow",
                   "Waste/wastewater" = "grey")

#Step 2.2: Scatter plot of gdp and emissions
ggplot(mergeddata_elm7, aes(x = gdp_pc, y = emissions_pc, color = sector)) +
  geom_point() +
  scale_color_manual(values = sector_colors)+
  labs(title = "Scatter Plot of Emission per capita vs. GDP per capita",
       x = "GDP Per Capita",
       y = "Emissions Per Capita") +
  theme_minimal()

#Additional test to check hypothesis validity
corr2 <- cor(mergeddata_elm7$gdp_pc, mergeddata_elm7$emissions_pc, method =
"pearson")
```