Detailed Project Solution – CO■ Emissions by European Cities

ELEMENT 1 — DATASET OVERVIEW
- Filter out rows where emissions per capita or population is missing.
- Count number of cities and countries.
- Identify countries with highest and lowest representation.

R CODE:
```
em_clean <- em %>%
filter(!is.na(emissions_pc), !is.na(population))
n_cities <- nrow(em_clean)
n_countries <- em_clean %>% n_distinct(country)
country_counts <- em_clean %>% count(country) %>% arrange(n)
```

INTERPRETATION:
Dataset contains X cities across Y countries. Some countries are highly overrepresented.

ELEMENT 2 — POPULATION DISTRIBUTION
- Histogram of population.
- Identify min, max, and median population.

R CODE:
```
ggplot(em_clean, aes(population)) + geom_histogram(binwidth=50000)
em_clean %>% arrange(population) %>% slice(1)
em_clean %>% arrange(desc(population)) %>% slice(1)
median(em_clean$population)
```

INTERPRETATION:
Distribution is skewed with many small cities and few very large ones.

ELEMENT 3 — EMISSIONS BY COUNTRY
- Boxplot of emissions per capita per country.
- Identify top 3 and bottom 3 median emitters.

R CODE:
```
med_by_country <- em_clean %>%
group_by(country) %>%
summarise(med = median(emissions_pc, na.rm=TRUE)) %>%
arrange(med)
em_clean$country <- factor(em_clean$country, levels=med_by_country$country)
ggplot(em_clean, aes(country, emissions_pc)) + geom_boxplot() + coord_flip()
head(med_by_country, 3)
tail(med_by_country, 3)
```

INTERPRETATION:
Countries differ significantly in emissions. The highest and lowest emitting groups are noted.

ELEMENT 4 — EMISSIONS BY SECTOR
- Sum emissions for each of six economic sectors.

R CODE:
```
sec_long <- sec %>% pivot_longer(cols=starts_with("sector"),
names_to="sector", values_to="value")
sector_totals <- sec_long %>%
group_by(sector) %>% summarise(total=sum(value, na.rm=TRUE))
ggplot(sector_totals, aes(sector, total)) + geom_col()
```

INTERPRETATION:
Residential and transport sectors contribute most emissions.

ELEMENT 5 — SECTOR SHARE BY COUNTRY
- Join datasets.
- Create stacked bar plot of sector fractions.

R CODE:
```
joined <- sec_long %>% left_join(em_clean %>% select(city_id, country), by="city_id")
sector_country <- joined %>%
group_by(country, sector) %>%
summarise(total=sum(value, na.rm=TRUE)) %>%
group_by(country) %>% mutate(frac = total / sum(total))
ggplot(sector_country, aes(country, frac, fill=sector)) +
geom_bar(stat="identity", position="fill") + coord_flip()
```

INTERPRETATION:
Sector contributions vary widely by country.

ELEMENT 6 — HEATING DEMAND AND EMISSIONS
- Scatterplot relating heating degree days to emissions.
- Highlight Scandinavian cities.

R CODE:
```
scand <- c("Sweden","Norway","Finland","Denmark")
em_clean <- em_clean %>% mutate(is_scand = country %in% scand)
ggplot(em_clean, aes(heating_degree_days, emissions_pc, color=is_scand)) +
geom_point()
```

INTERPRETATION:
Scandinavian cities have high heating needs but emissions do not uniformly increase.

ELEMENT 7 — GDP AND EMISSIONS
- Scatterplot of GDP per capita vs emissions.
- Remove extreme outlier.

R CODE:

```
outlier <- em_clean %>% arrange(desc(gdp_pc)) %>% slice(1)
em_no_out <- em_clean %>% filter(gdp_pc < max(gdp_pc, na.rm=TRUE))
ggplot(em_no_out, aes(gdp_pc, emissions_pc)) + geom_point()
```

INTERPRETATION:
Wealthier cities tend to have higher emissions, but with exceptions.

ELEMENT 8 — SUMMARY
Cities differ substantially in population, sector structure, heating demand, and wealth. Transport and residential sectors dominate emissions. GDP shows a clearer link to emissions than heating needs. Further hypothesis testing would require variables such as energy source mix, climate normalization, and industrial structure.