

DATA SCIENCE TOOLBOX PYTHON PROGRAMMING

PROJECT REPORT

(Project Semester January-April 2025)

**Greenhouse Gas Emissions Analysis**

Submitted by

**NAME:** Papabaigari Mahammad Yaseen

**REGISTRATION NO:** 12317975

**PROGRAMME AND SECTION:** K23GN

**Roll no:** 12

**COURSE CODE:** INT375

Under the Guidance of

MRS. AASHIMA (UID: 28968)

Discipline of CSE/IT

Lovely School of Computer Science

Lovely Professional University, Phagwara

**CERTIFICATE**

This is to certify that Papabaigari Mahammad Yaseen bearing Registration no. 12317975 has completed CSE375 project titled, “Mrs. Aashima” under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

Signature and Name of the Supervisor

Designation of the Supervisor

School of Computer Science

Lovely Professional University

Phagwara, Punjab.

Date:

**DECLARATION**

I, Papabaigari Mahammad Yaseen student of CSE (Program name) under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: Signature

Registration No: 12317975 Name of the student: Yaseen

|  |  |
| --- | --- |
| **Title** | **Pageno** |
| **1. Introduction** |  |
| **2. Source of Dataset** |  |
| **3. EDA Process** |  |
| **4. Analysis on Dataset** |  |
| **4.1 Top Emitting Sectors (Bar Plot)** |  |
| **4.2 Household Emissions Trend (Line Graph)** |  |
|  |  |
| **4.3 Top 5 Contributors (Pie Chart)** |  |
| **4.4 Emission Intensity of Households (Scatter Plot)** |  |
| **4.5 Pair Plot of Variables** |  |
| **4.6 Correlation Heatmap** |  |
| **4.7 Boxplot for Outliers** |  |
| **4.8 Z-Score Based Outlier Detection** |  |
| **5. Conclusion** |  |
| **6. Future Scope** |  |
| **7. References** |  |

**1. Introduction**

This project titled "Greenhouse Gas Emissions Analysis 2024" aims to explore and analyze the greenhouse gas emissions dataset provided by Stats NZ for the September 2024 quarter. The dataset includes information about emissions from various industries and households in kilotonnes. The project uses Python and data analysis libraries to extract insights and visualize emission patterns, outliers, trends, and correlations.

Environmental Data Understanding (EDU) is central to this study. EDU refers to the process of analyzing and comprehending environmental datasets to make data-driven decisions for sustainable development. It is crucial because informed decisions can reduce pollution, mitigate climate change effects, and support policy-making. Through this project, we achieve EDU by cleaning the data, performing exploratory data analysis (EDA), identifying trends, visualizing patterns, and interpreting the results.

**2. Source of Dataset**

The dataset used for this analysis is:

**Name:** Greenhouse Gas Emissions - Industry and Household (September 2024 quarter)  
**Source:** Stats NZ (Statistics New Zealand)  
**Format:** CSV  
**Path:** [C:\Users\moham\Downloads\greenhouse-gas-emissions-industry-and-household-September-2024-quarter.csv](file:///C:\Users\moham\Downloads\greenhouse-gas-emissions-industry-and-household-September-2024-quarter.csv)

**3. EDA Process**

Exploratory Data Analysis (EDA) is essential in understanding the dataset’s structure, contents, and anomalies. The EDA process in this project includes:

* Loading the data using Pandas
* Displaying basic information: info(), head(), describe(), nunique()
* Checking for missing values
* Visualizing data trends using bar plots, pie charts, line plots, and scatter plots
* Detecting outliers using boxplots and z-score
* Analyzing relationships between features with correlation heatmaps and pair plots

**Python Tools Used:**

To conduct EDA, the following Python libraries were used:

* **Pandas:** For loading, cleaning, and manipulating tabular data.
* **NumPy:** For numerical operations and statistics.
* **Matplotlib & Seaborn:** For generating informative visualizations like line graphs, pie charts, boxplots, and heatmaps.
* **SciPy:** Specifically for calculating Z-scores to flag outliers.

**4. Analysis on Dataset**

**4.1 Top Emitting Sectors (Bar Plot)**

**i. Introduction:** Identify sectors contributing most to emissions.

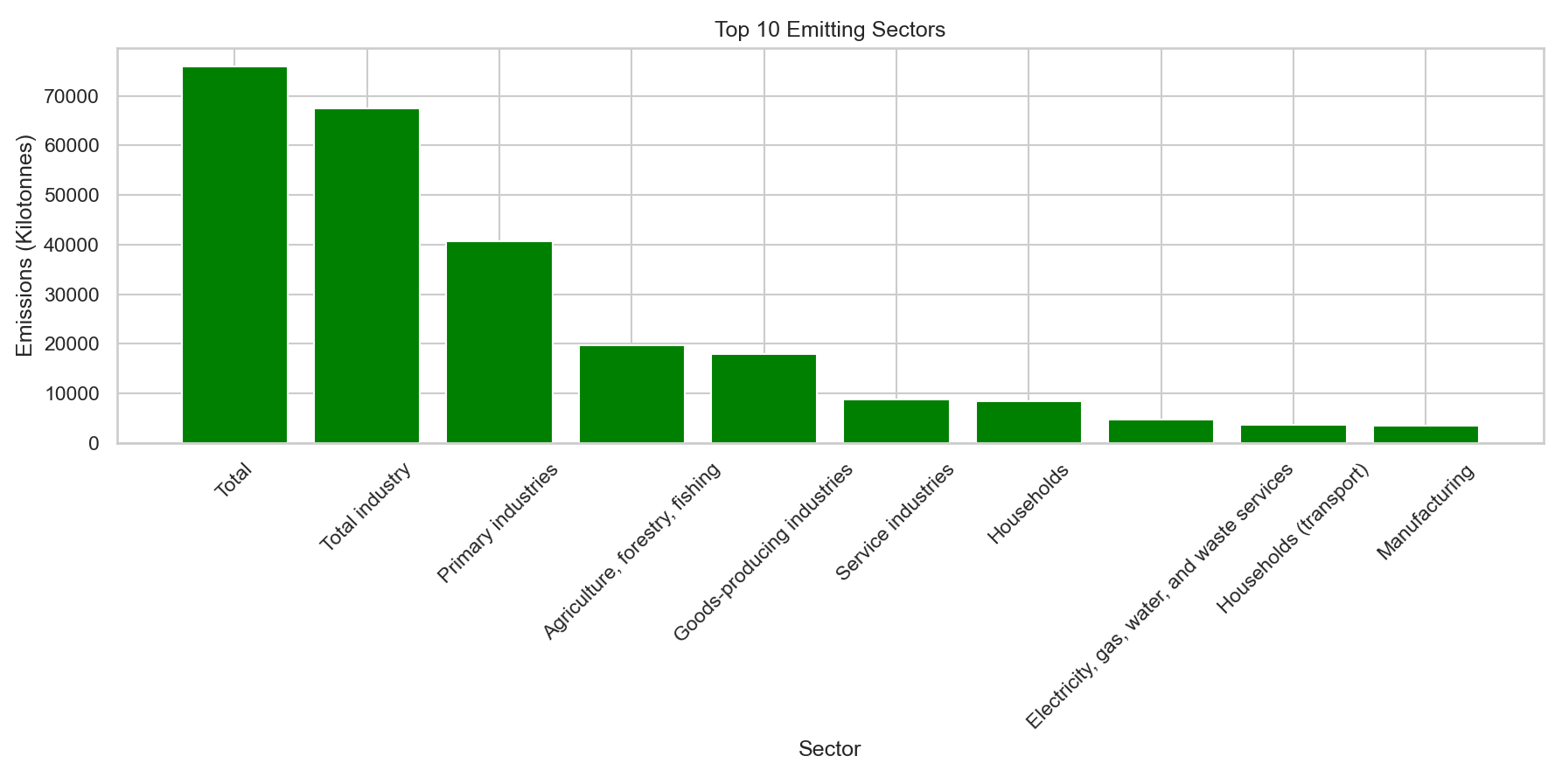
**ii. General Description:** Filter latest period, group by sector, sort by emission levels.

**iii. Functions/Formulas Used:** groupby(), sum(), sort\_values(), plt.bar()

**iv. Analysis Results:** Top 10 sectors identified based on total emissions.

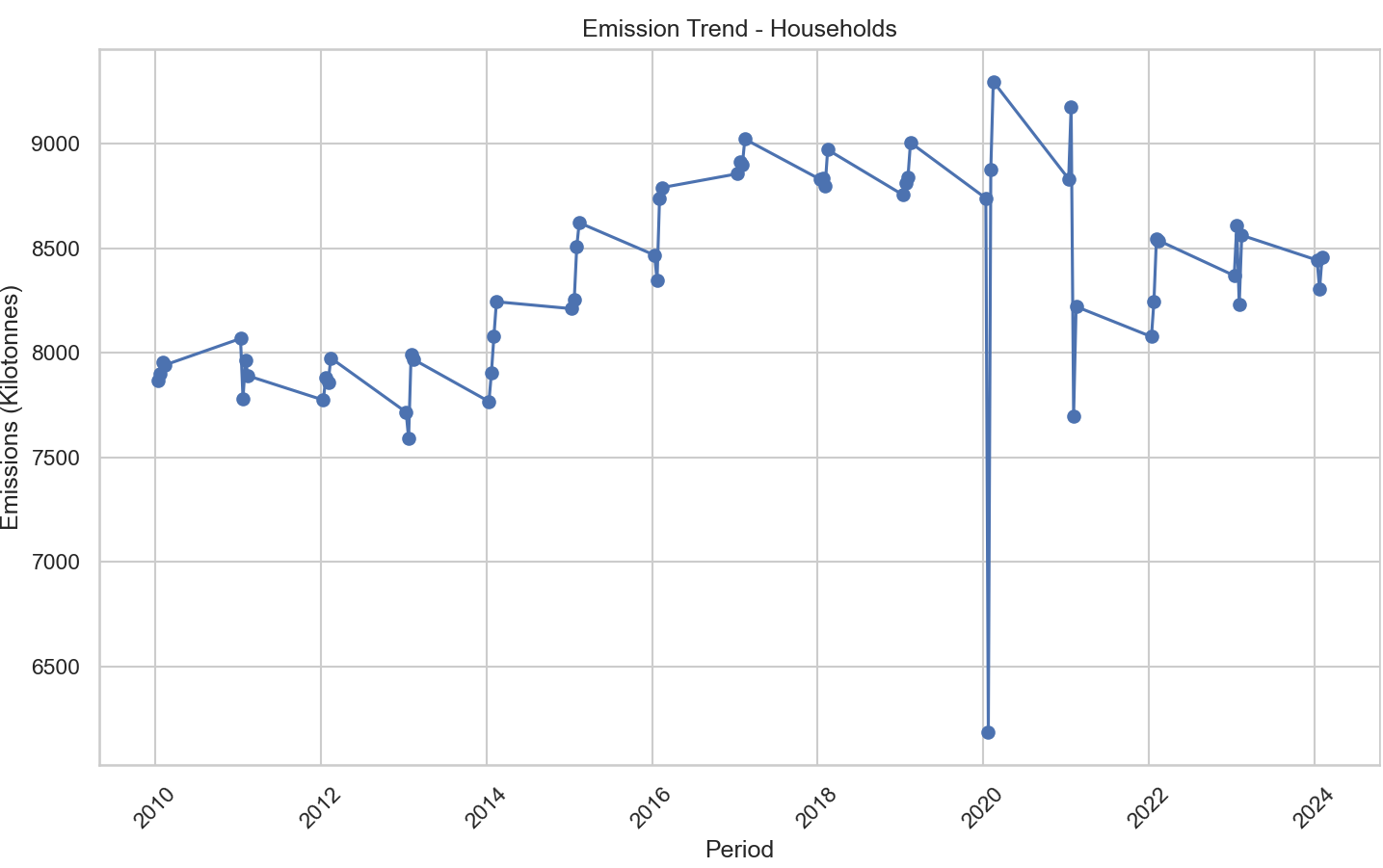
**v. Visualization:**

**Bar Graph**



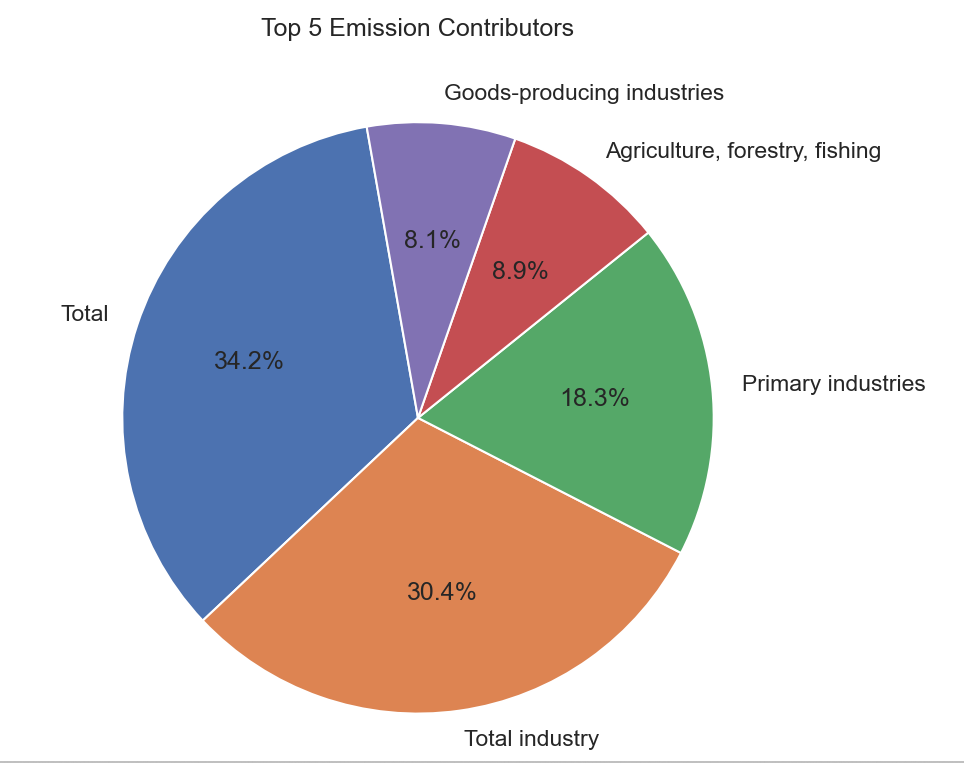
**4.2 Household Emissions Trend (Line Graph)**

**i. Introduction:** Observe emissions from households over time.  
**ii. General Description:** Filter rows for households, group by period.  
**iii. Functions/Formulas Used:** groupby(), sum(), plt.plot()  
**iv. Analysis Results:** Reveals increasing or decreasing trend.  
**v. Visualization:**

****

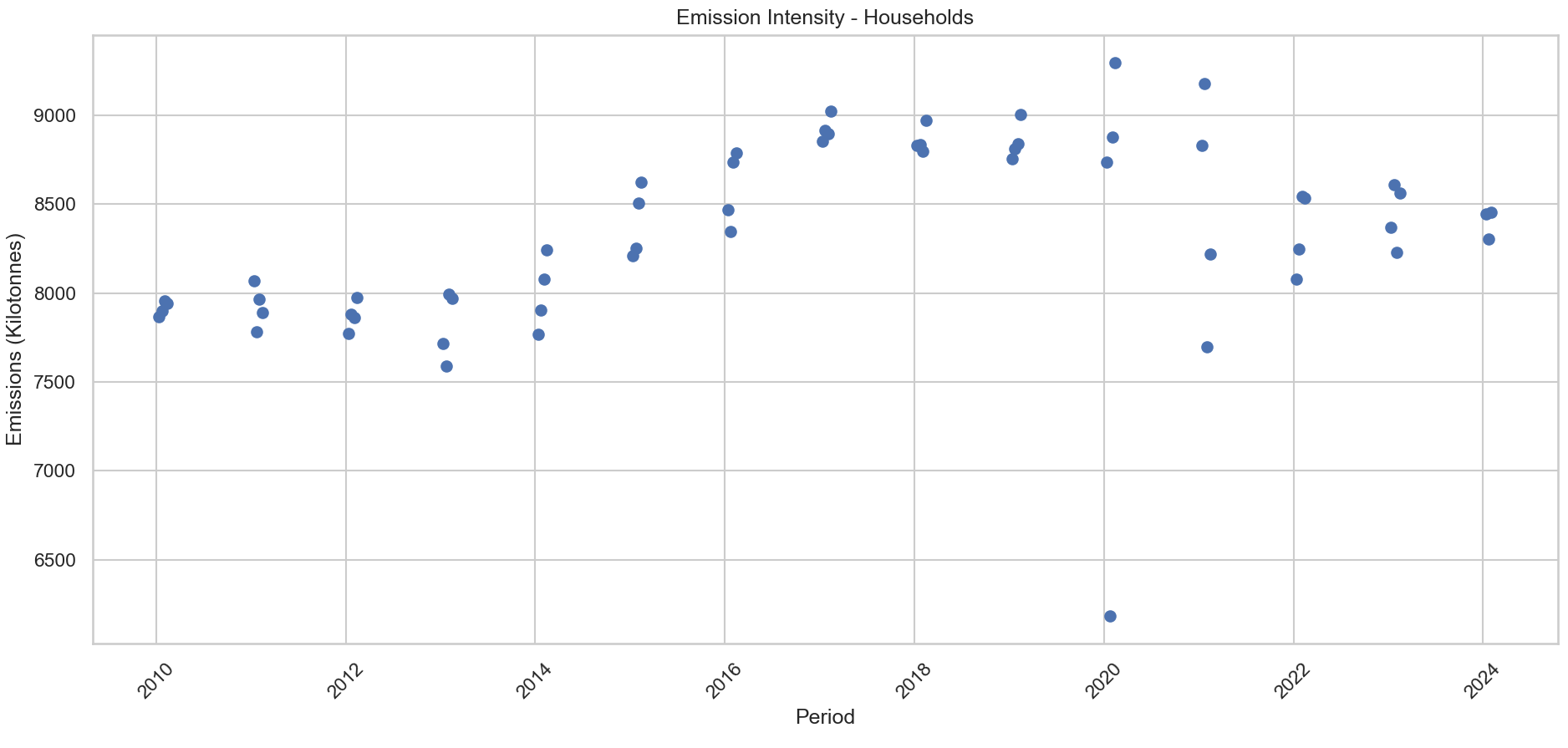
**4.3 Top 5 Contributors (Pie Chart)**

i. Introduction: Understand emission shares of top sectors.  
ii. General Description: Same as bar chart, but limit to top 5 and plot pie chart.  
iii. Functions/Formulas Used: groupby(), sum(), plt.pie()  
iv. Analysis Results: Distribution of emissions among major sectors.  
v. Visualization:



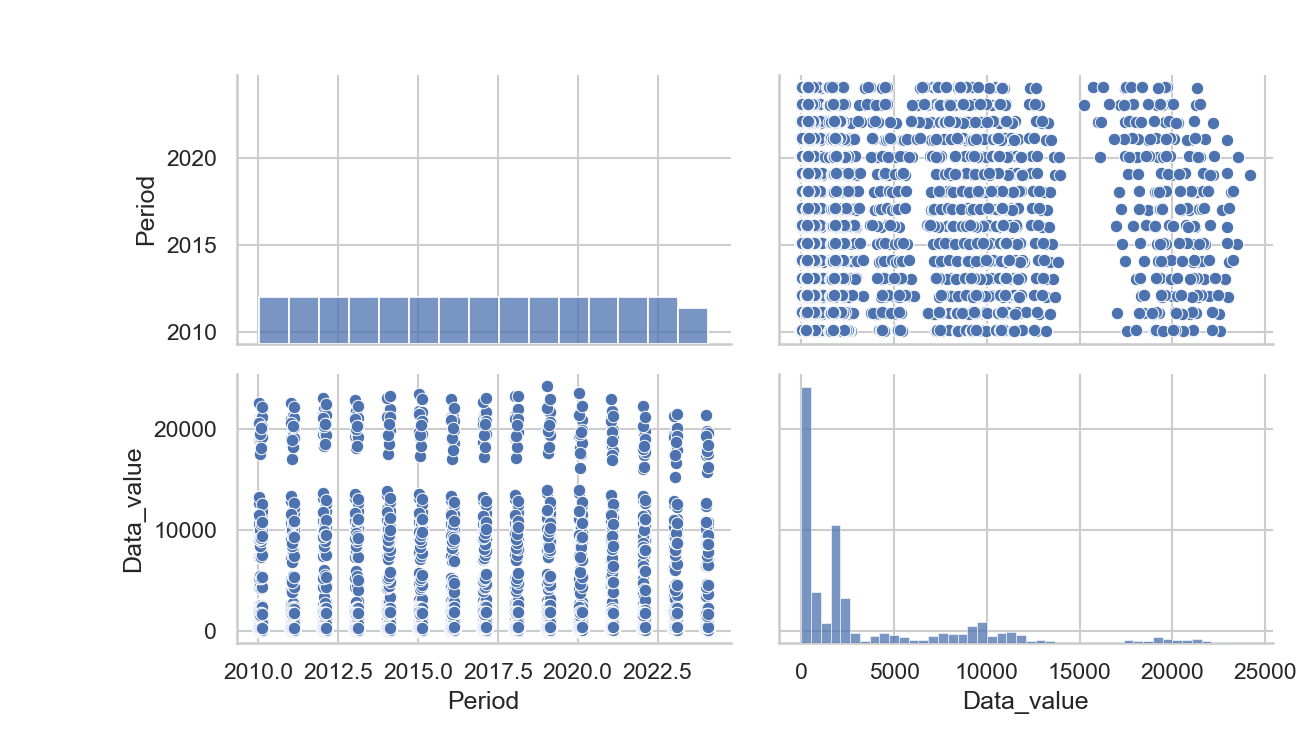
**4.4 Emission Intensity of Households (Scatter Plot)**

**i. Introduction:** Highlight emission spikes and drops.  
**ii. General Description:** Group household emissions by period.  
**iii. Functions/Formulas Used:** groupby(), reset\_index(), plt.scatter()  
**iv. Analysis Results:** Identify fluctuations in emission values.  
**v. Visualization:**



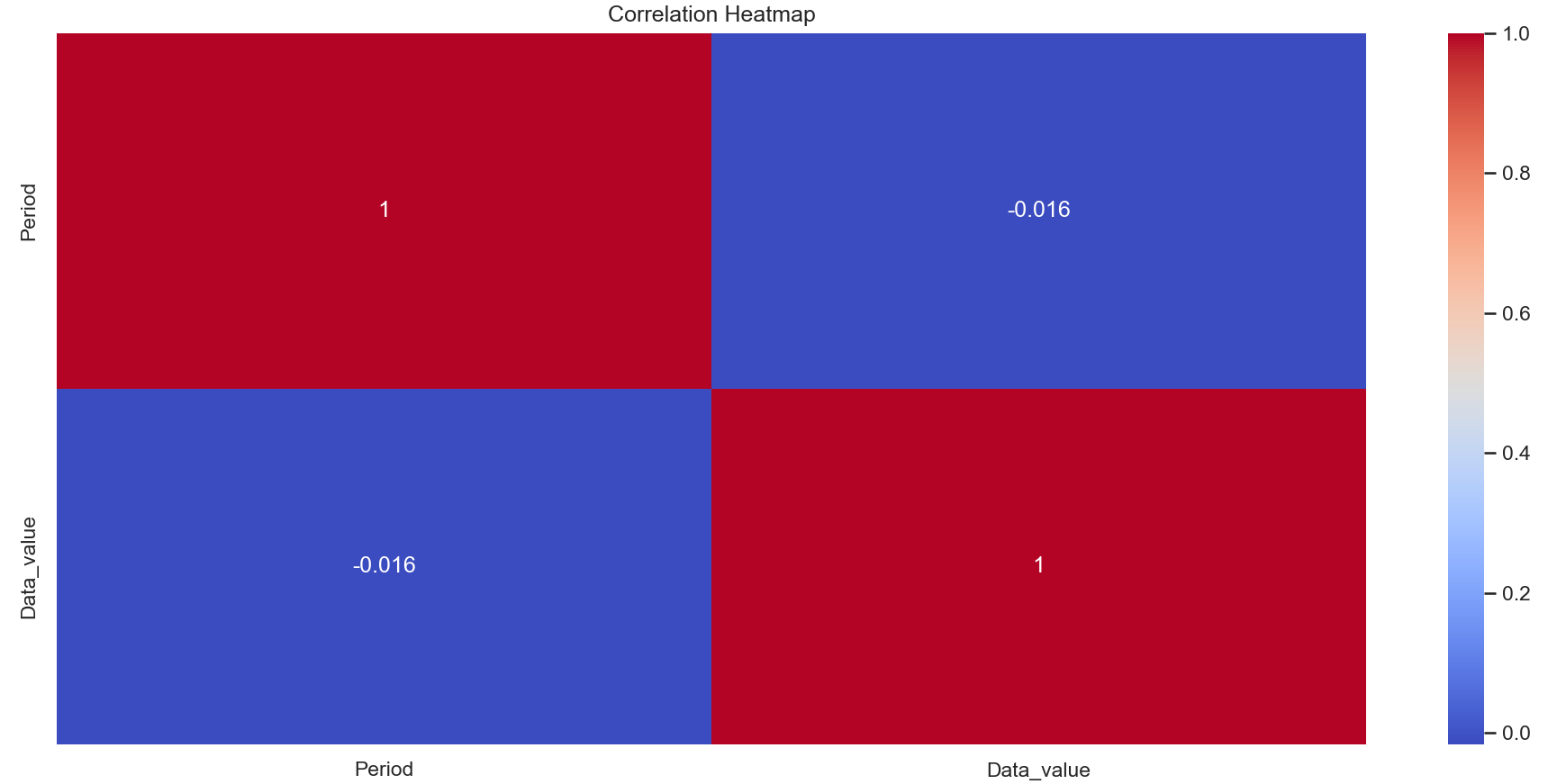
**4.5 Pair Plot of Variables**

**i. Introduction:** Observe pairwise relationships.  
**ii. General Description:** Use sns.pairplot() on numerical columns.  
**iii. Functions/Formulas Used:** pairplot() from seaborn.  
**iv. Analysis Results:** See how variables relate visually.  
**v. Visualization:**



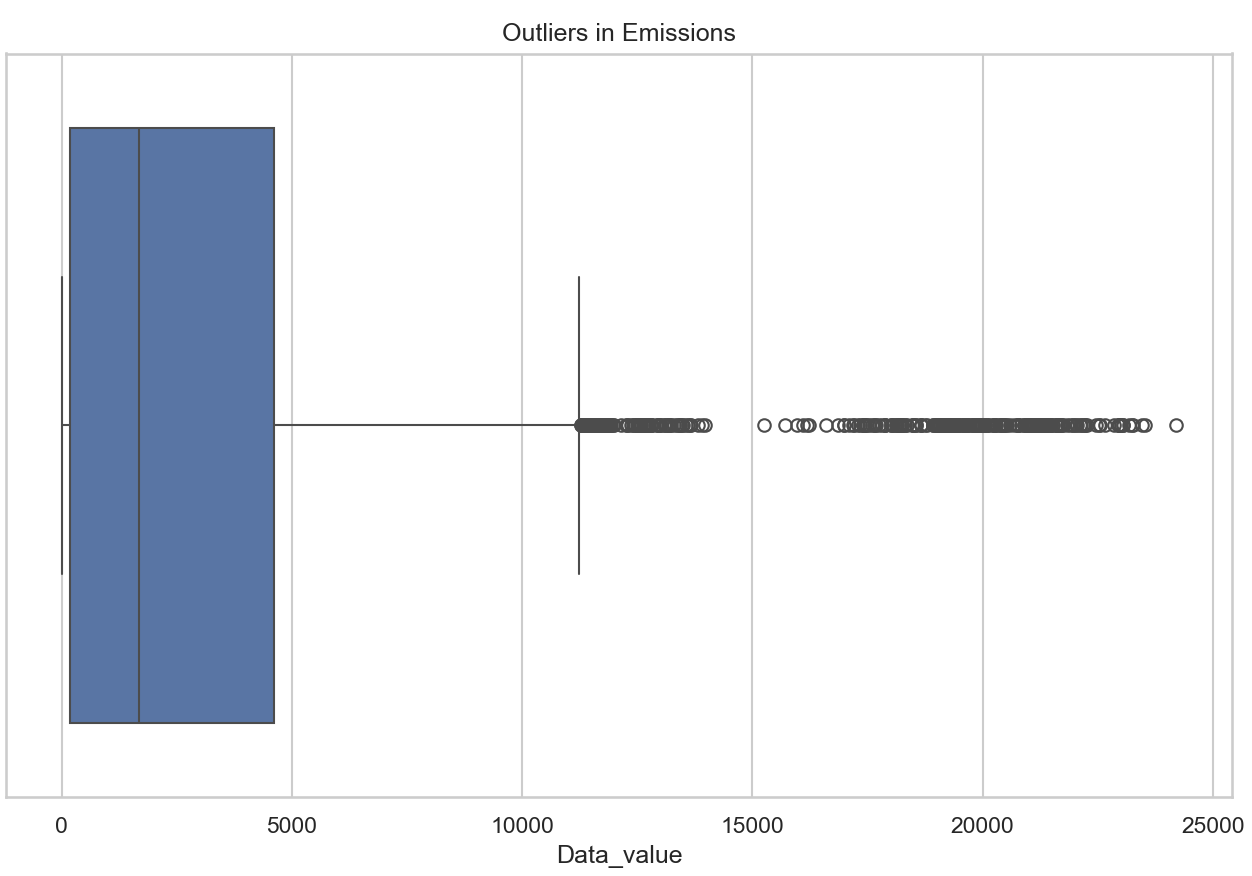
**4.6 Correlation Heatmap**

**i. Introduction:** Assess strength of relationship between features.  
**ii. General Description:** Calculate correlation and plot heatmap.  
**iii. Functions/Formulas Used:** corr(), heatmap()  
**iv. Analysis Results:** Shows weak or strong correlations.  
**v. Visualization:**



**4.7 Boxplot for Outliers**

**i. Introduction:** Detect extreme values.  
**ii. General Description:** Use boxplot on emission values.  
**iii. Functions/Formulas Used:** sns.boxplot()  
**iv. Analysis Results:** Visual representation of outliers.  
**v. Visualization:**



**4.8 Z-Score Based Outlier Detection**

**i. Introduction:** Quantify outliers numerically.  
**ii. General Description:** Apply z-score method to emissions.  
**iii. Functions/Formulas Used:** zscore(), condition to extract z > 3.  
**iv. Analysis Results:** List of rows with high deviations.  
**v. Visualization:** Displayed as a printed DataFrame.

**5. Conclusion**

Through this project, we explored the greenhouse gas emissions data across industries and households. Key contributors to emissions were identified. Household emissions trends and intensity were analyzed. Relationships between variables were visualized and outliers were detected. The results enhance environmental data understanding and can support sustainable decision-making.

**6. Future Scope**

* Predict future emissions using machine learning models.
* Integrate weather or population data for deeper insights.
* Automate reporting for real-time monitoring.
* Develop dashboards for public policy evaluation.
* Apply clustering techniques to group industries by emission profiles.

**7. References**

* Stats NZ: Greenhouse Gas Emissions dataset
* Python Libraries: Pandas, Matplotlib, Seaborn, Scipy
* Documentation and tutorials from W3Schools, GeeksforGeeks, and official docs

**8.Source Code**

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from scipy.stats import zscore

# Load dataset

df = pd.read\_csv("C:\\Users\\moham\\Downloads\\greenhouse-gas-emissions-industry-and-household-September-2024-quarter.csv")

# Info and head

print(df.info())

print(df.head())

print(df.describe())

print(df.nunique())

# Missing values

print(df.isnull().sum())

# Plot style

sns.set(style="whitegrid")

plt.rcParams["figure.figsize"] = (12, 6)

# Get the latest period

latest = df['Period'].max()

# Filter data for the latest period

latest\_df = df[df['Period'] == latest]

# Group and sort emissions by sector

grouped = latest\_df.groupby('Anzsic\_descriptor')['Data\_value'].sum()

top10 = grouped.sort\_values(ascending=False).head(10)

# Create basic vertical bar graph

plt.bar(range(len(top10)), top10.values, color='green')

plt.xticks(range(len(top10)), top10.index, rotation=45)

plt.title('Top 10 Emitting Sectors')

plt.xlabel('Sector')

plt.ylabel('Emissions (Kilotonnes)')

plt.tight\_layout()

plt.show()

#line graph

df\_households = df[df['Anzsic\_descriptor'] == 'Households']

# Group and sum emissions by period

g = df\_households.groupby('Period')['Data\_value'].sum()

# Create the line plot with markers

plt.plot(g.index, g.values, marker='o')

plt.title('Emission Trend - Households')

plt.xlabel('Period')

plt.ylabel('Emissions (Kilotonnes)')

plt.xticks(rotation=45)

plt.tight\_layout()

plt.show()

#Pie Chart - Top 5 sectors

latest = df['Period'].max()

# Filter data for the latest period

latest\_df = df[df['Period'] == latest]

# Group, sum, and get top 5 contributors

grouped5 = latest\_df.groupby('Anzsic\_descriptor')['Data\_value'].sum().sort\_values(ascending=False).head(5)

# Plot basic pie chart

plt.pie(grouped5, labels=grouped5.index, autopct='%1.1f%%', startangle=100)

plt.title('Top 5 Emission Contributors')

plt.tight\_layout()

plt.show()

# 4. Emission Intensity (Scatter Plot)

df\_sector = df[df['Anzsic\_descriptor'] == 'Households']

# Group by Period and sum the emissions

grouped = df\_sector.groupby('Period')['Data\_value'].sum().reset\_index()

# Create scatter plot

plt.scatter(grouped['Period'], grouped['Data\_value'])

plt.title('Emission Intensity - Households')

plt.xlabel('Period')

plt.ylabel('Emissions (Kilotonnes)')

plt.xticks(rotation=45)

plt.tight\_layout()

plt.show()

# Pair Plot

sns.pairplot(df[['Period', 'Data\_value']])

plt.suptitle("Pair Plot of Numerical Variables", y=2.02)

plt.tight\_layout()

plt.show()

# 5. Heatmap (Correlation)

corr\_data = df[['Period', 'Data\_value']].corr()

sns.heatmap(corr\_data, annot=True, cmap='coolwarm')

plt.title('Correlation Heatmap')

plt.tight\_layout()

plt.show()

# 6. Boxplot for Outliers

sns.boxplot(data=df, x='Data\_value')

plt.title('Outliers in Emissions')

plt.tight\_layout()

plt.show()

# 7. Z-score Outliers

df['Z\_score'] = zscore(df['Data\_value'])

outliers = df[abs(df['Z\_score']) > 3]

print(outliers[['Anzsic\_descriptor', 'Period', 'Data\_value', 'Z\_score']])