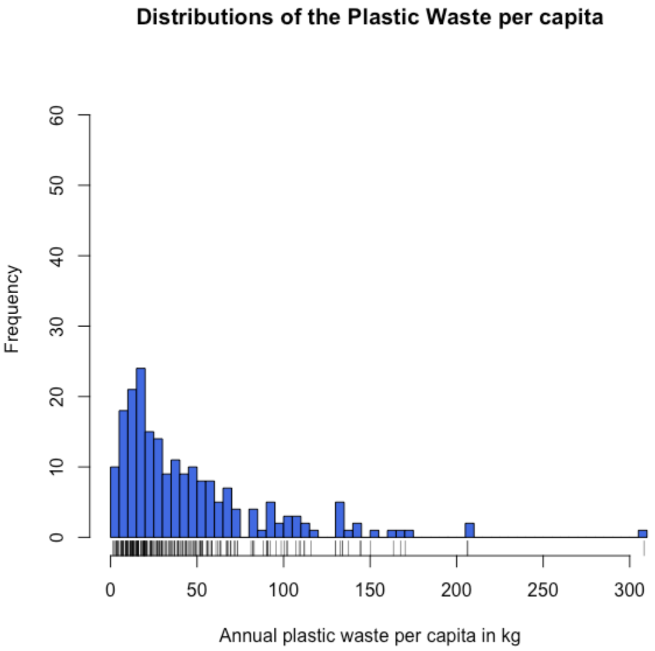
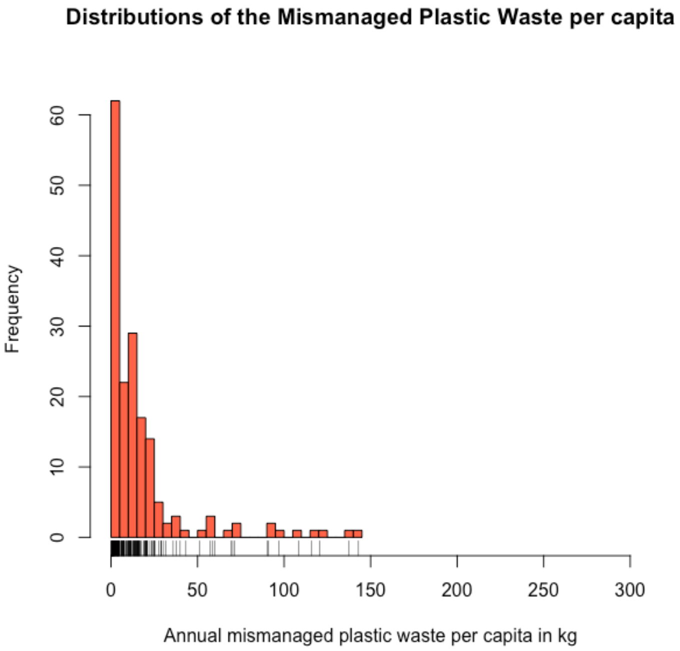
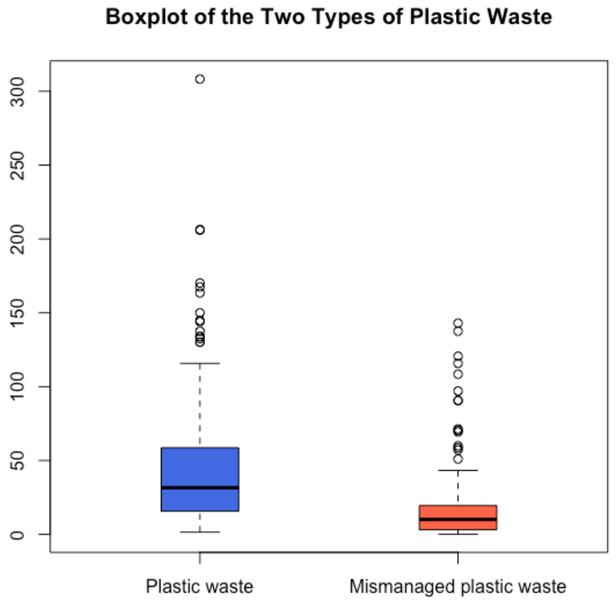
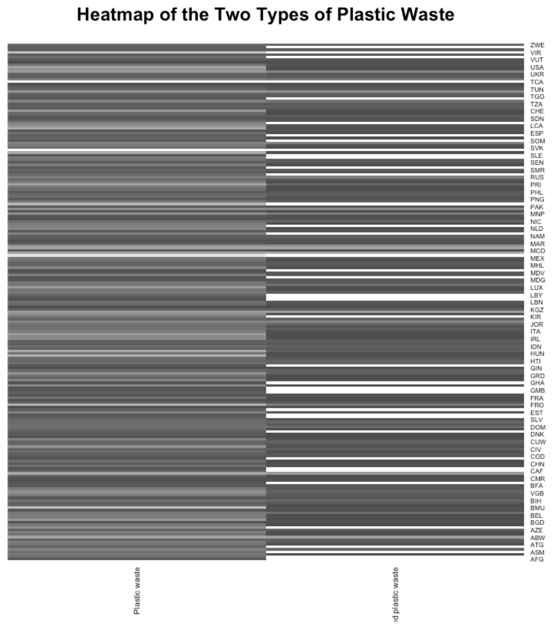
Q1: The distributions of the two types of plastic waste. Identify and explore any potential outliers, unusual values, and missing values.


Distributions of the Plastic Waste per capita


Distributions of the Mismanaged Plastic Waste per capita


Boxplot of the Two Types of Plastic Waste


Heatmap of the Two Types of Plastic Waste

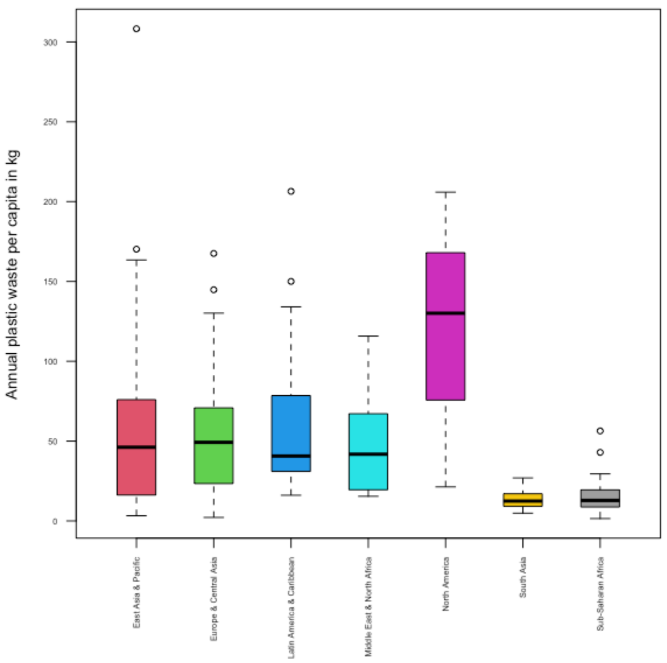
- **Distribution:** Both types of plastic waste exhibit right-skewed distributions. Plastic waste exhibits a broader spread compared to mismanaged plastic waste, whereas the values of mismanaged plastic waste are more centered. Additionally, the values of plastic waste are greater than those of mismanaged plastic waste, given that mismanaged plastic waste is a subset of total plastic waste.
- **Potential Outliers:** Numerous outliers are present in both types of plastic waste.
- **Unusual Values:** There are 3 extreme values in the plastic waste variable, exceeding 200 per capita in kg.
- **Missing Values:** The heatmap reveals a small number of missing values in the plastic waste variable, whereas there are numerous missing values in the mismanaged plastic waste variable. This could be attributed to the challenges in accurately assessing and investigating mismanaged plastic waste.

Summary:

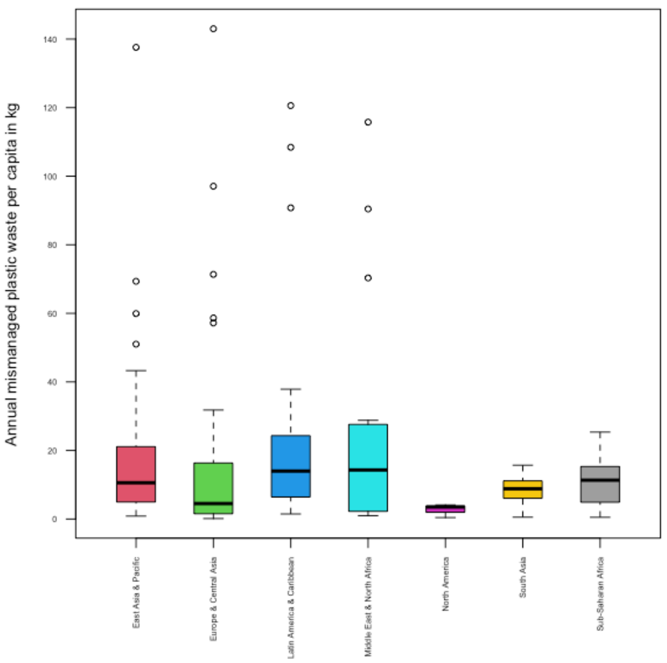
- **Model Selection:** As both types of plastic waste are continuous variables, regression models can be applied. However, linear regression may not be suitable due to the non-normal distribution. Models like Regression Trees and Random Forests, which can handle both categorical and continuous features and are robust to outliers, may be more appropriate choices.
- **Data refinement:** Numerous missing values may significantly impact the modelling and analysis of mismanaged plastic waste (e.g., the relationship analysis). Further investigation is warranted to determine if outliers and missing values stem from insufficient data collection or recording errors. Refining the dataset could enhance the reliability and accuracy of the analysis.

Q2: The potential effects of region and income status on the distributions of plastic waste and mismanaged plastic.

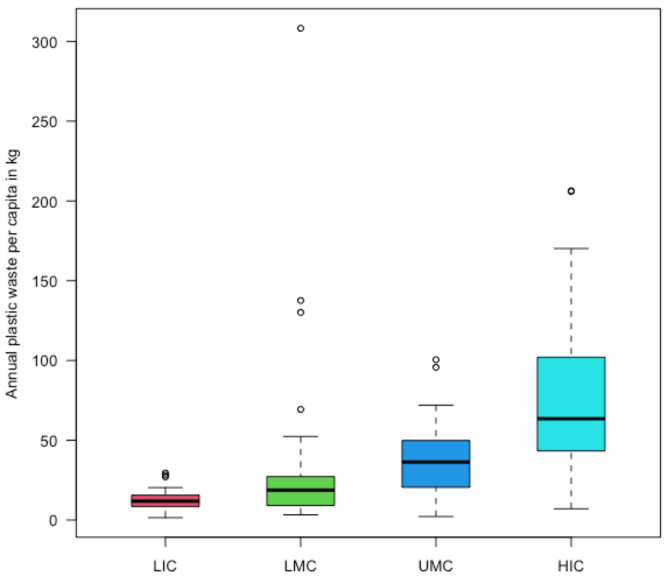
Boxplot of Plastic Waste Grouped by Region



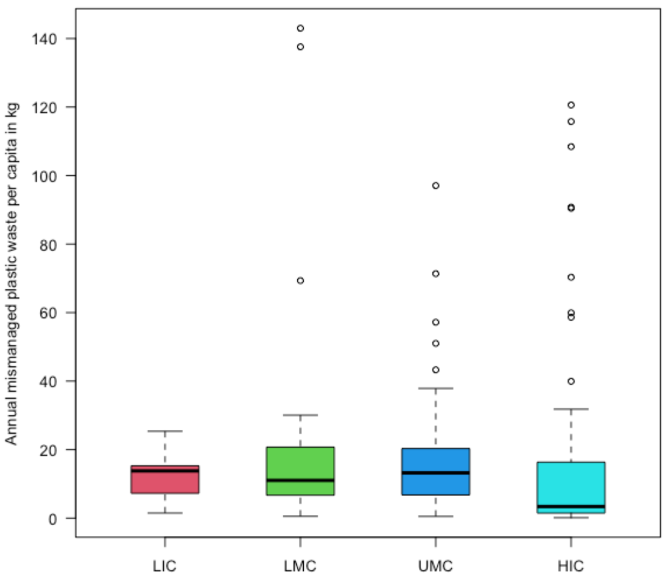
Boxplot of Mismanaged Plastic Waste Grouped by Region



Boxplot of Plastic Waste Grouped by Income Status



Boxplot of Mismanaged Plastic Waste Grouped by Income Status

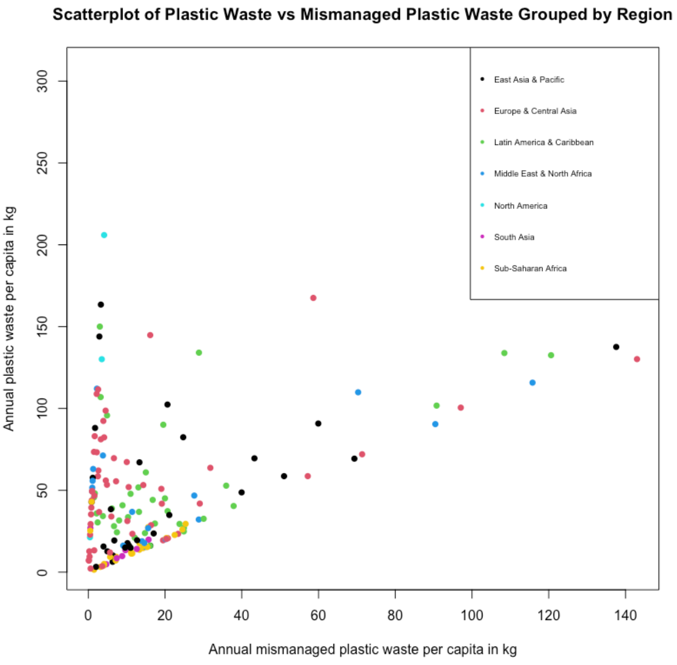


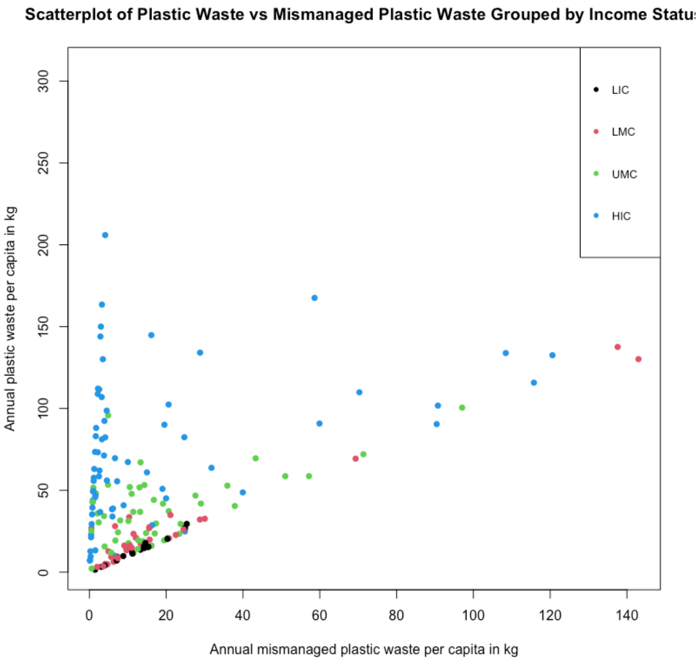
- **North America:** Plastic waste in North America exhibits substantially higher values compared to other regions, with minimal amounts of mismanaged plastic waste. This trend implies that developed countries such as the USA and Canada tend to generate larger quantities of plastic waste and have well-established waste management systems.
- **South Asia and Sub-Saharan Africa:** Both types of plastic waste tend to have relatively smaller values in South Asia and Sub-Saharan Africa, where the majority of countries are economically disadvantaged.
- **Other Regions:** Other regions show varying degrees of outliers, which may be attributed to the complex and diverse stages of development across countries in these regions.
- **Income Status Groups:** Countries with lower income status tend to have lower levels of both types of plastic waste, while those with higher income status exhibit relatively better plastic waste management practices. Additionally, plastic waste tends to increase with increasing income status, although the distribution of mismanaged plastic waste is less clearly distinguished.

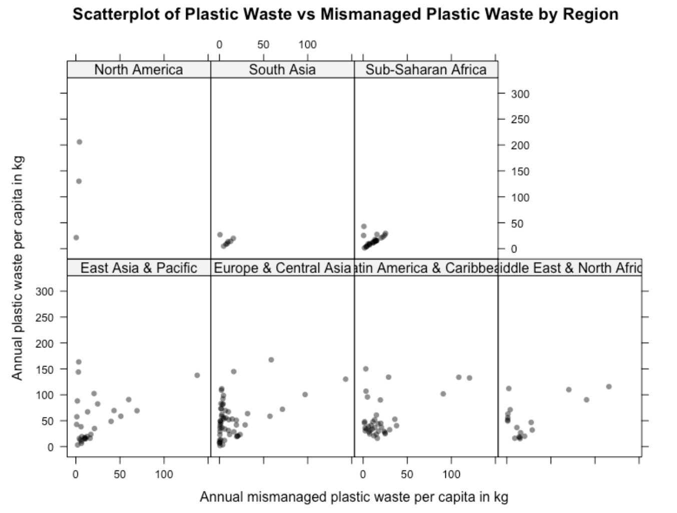
Summary:

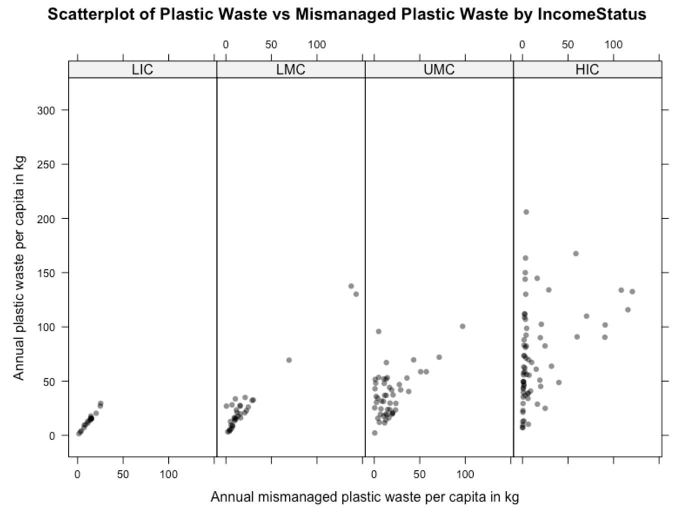
- The income status feature appears to be a better predictor for distinguishing between the two types of plastic waste compared to regions.
- The different patterns observed in rich countries regarding the two types of plastic waste indicate the need for nonlinear modelling approaches.

Q3: The relationship between plastic waste and mismanaged plastic waste, and any potential impact of region and income status.







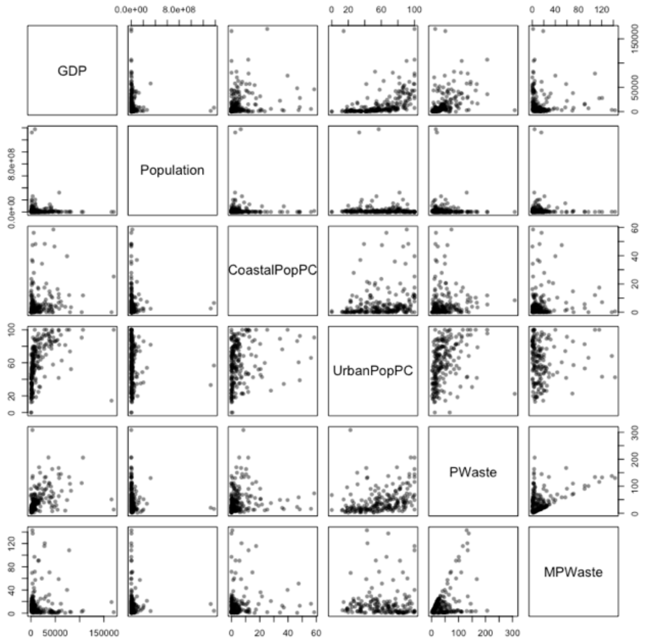


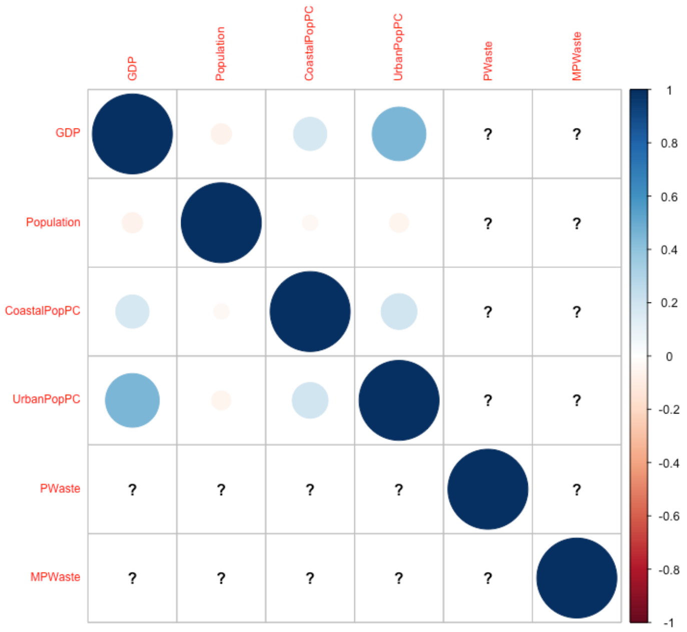
- **Scatterplot:** Overall, the relationship between plastic waste and mismanaged plastic waste has some nonlinear association. Grouping by region provides limited insights as the data points appear mixed up, making it difficult to discern clear patterns. Conversely, grouping by income status reveals a clear pattern, with distinct divisions into three statuses (high, middle, and low).
- **Lattice plots:** South Asia and sub-Saharan Africa exhibit a linear relationship, where mismanaged plastic waste increases proportionally with plastic waste. In contrast, North America and most countries in Europe & Central Asia display steeper slope patterns. Here, mismanaged plastic waste does not significantly increase despite a significant increase in plastic waste. (However, North America has very few observations, and there are countries with varying development levels in Europe & Central Asia, indicating potential complexities.) Other regions show less distinct patterns. Similar patterns are observed when comparing income status. In low-income countries, there is a positive correlation between mismanaged plastic waste and plastic waste. However, in high-income countries, the relationship is more complex, likely influenced by varying levels of sustainable development initiatives.

Summary:

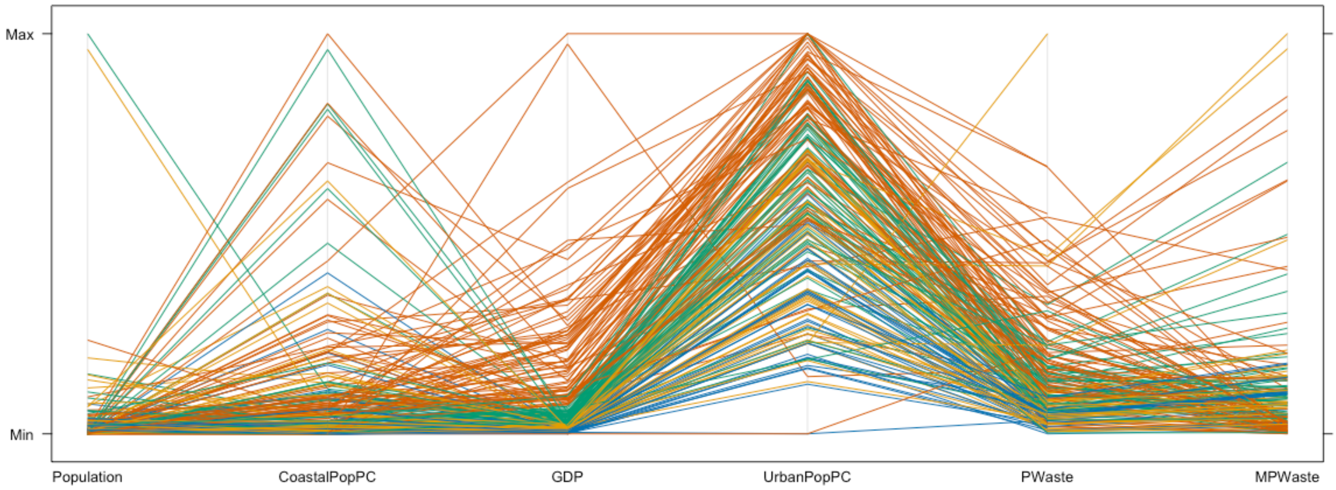
- The income status feature demonstrates a greater impact compared to the region, making it a more suitable predictor in modelling efforts.
- The hypotheses above suggesting that the observed relationship patterns of both types of plastic waste are influenced by increasing awareness of sustainable concepts require further investigation.
- Subdividing countries into smaller groups (e.g., provinces or states) may provide more precise subgroups and observations, potentially leading to more detailed models, although this could complicate investigations.

Q4: The relationship between both types of plastic waste and the other quantitative variables.





Parallel Coordinate Plot of Continuous Variables Grouped by Income Status

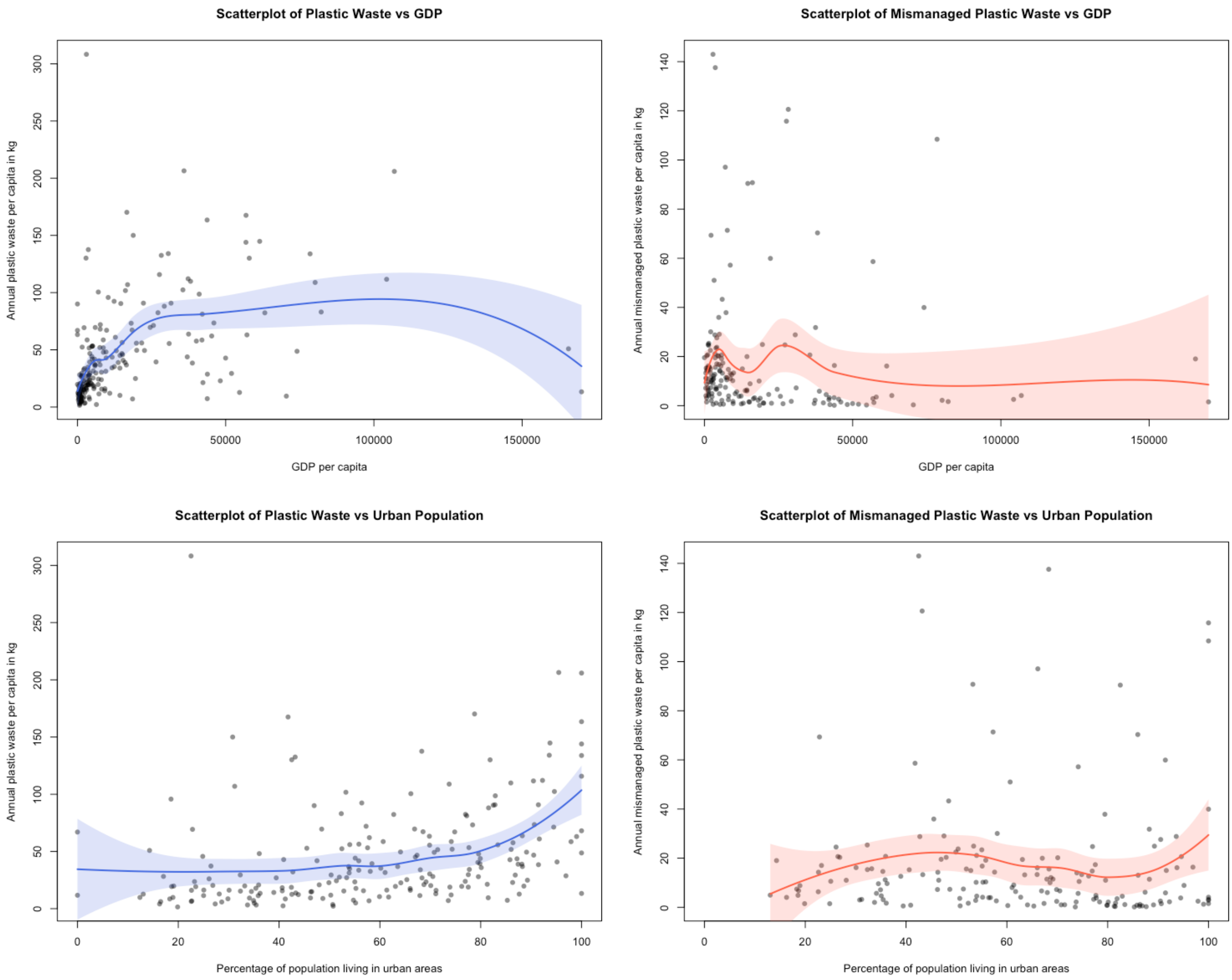


- **Lack of Strong Relationships:** While plastic waste shows some relationship with GDP and UrbanPopPC, identifying potential relationships with other variables proves challenging. This difficulty is exacerbated by extreme outliers, such as China and India, which significantly skew the data in the population variable, causing points to cluster predominantly on the left side.
- **Weak Correlation and Missing Values:** The data displays weak correlations, with "?" denoting missing values. Plastic waste has relatively few missing values, which can be safely omitted for analysis. However, the mismanaged plastic waste dataset contains 41 missing values out of a total of 211, significantly impacting correlation computation, which must be carefully addressed in subsequent analysis to ensure accurate results.
- **Parallel Coordinate Plot Analysis:** The parallel coordinate plot grouped by income status reveals that countries with higher GDP and urban population percentages tend to exhibit higher plastic waste. However, variables like population and coastal population percentage appear less informative for data separation. Mismanaged plastic waste demonstrates few significant relationships with other variables.

Summary:

- **Modelling Approach:** Given the low correlation and limited features, linear regression techniques such as multilinear, lasso, and ridge regression are not recommended. Instead, multinomial regression or decision trees might be more suitable for the dataset.
- **Dealing with Missing Values:** As the number of countries is limited, observations cannot be further sampled. Hence, it's crucial to address missing values as much as possible or employ suitable approaches for handling them.
- **Feature Collection:** Collecting other relevant features, such as educational level, is recommended and might be beneficial for modelling both types of plastic waste due to the low relationship among existing variables.

Q5: The smoothed trends between (i) both types of plastic waste and GDP (the wealth of the country), and (ii) both types of plastic waste and the size of urban population.



- Both types of plastic waste exhibit nonlinear trends when plotted against GDP and the size of the urban population.
- **Smoothed Trends with GDP:** Plastic waste tends to increase with increasing GDP per capita, reaching a flat and subsequent decrease at higher values. Mismanaged plastic waste shows an initial "S" shaped bend, which warrants further investigation. Subsequently, it remains flat, with values even smaller than those of countries with lower GDP. This possibly suggests more effective waste management systems in developed countries. However, large confidence intervals at the end indicate the influence of limited data and extreme values.
- **Smoothed Trends with The Size of Urban Population:** The percentage of urban population initially has a minimal impact on plastic waste, while mismanaged plastic waste displays a similar trend with a slight bump in the middle. Beyond 80%, both types of waste increase with the percentage of the urban population. Confidence intervals are large at the beginning, reflecting uncertainty due to limited data.

Summary:

- **Usefulness of Features:** GDP and urban population percentage may be valuable for modelling both types of plastic waste, except for the mismanaged plastic waste versus GDP relationship, which appears to have less contribution to modelling mismanaged plastic waste.
- **Unsuitability of Linear Regression:** The nonlinear smoother patterns suggest again that linear regression models may not be appropriate for this dataset.
- **Investigation of Anomalies:** The significant "S" bend in the plot of mismanaged plastic waste vs GDP warrants further investigation to uncover underlying reasons.
- **Handling Extreme Values:** Omitting extreme values, such as those with GDP values greater than 100,000 and urban population percentages less than 10%, can improve the generalization of the modelling process. These extreme values contribute to high confidence intervals and may skew the overall analysis.