Data Exploration, Visualization, and Unsupervised Learning – Assignment 1

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

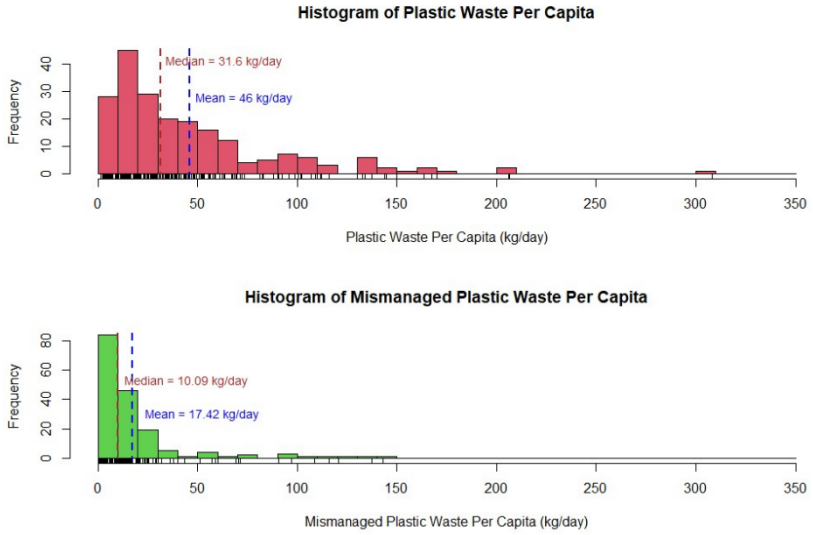

Figure 1: Histogram and Rug plot of Plastic Waste (Top) and Mismanaged Plastic Waste (Bottom) Per Capita (kg/day)

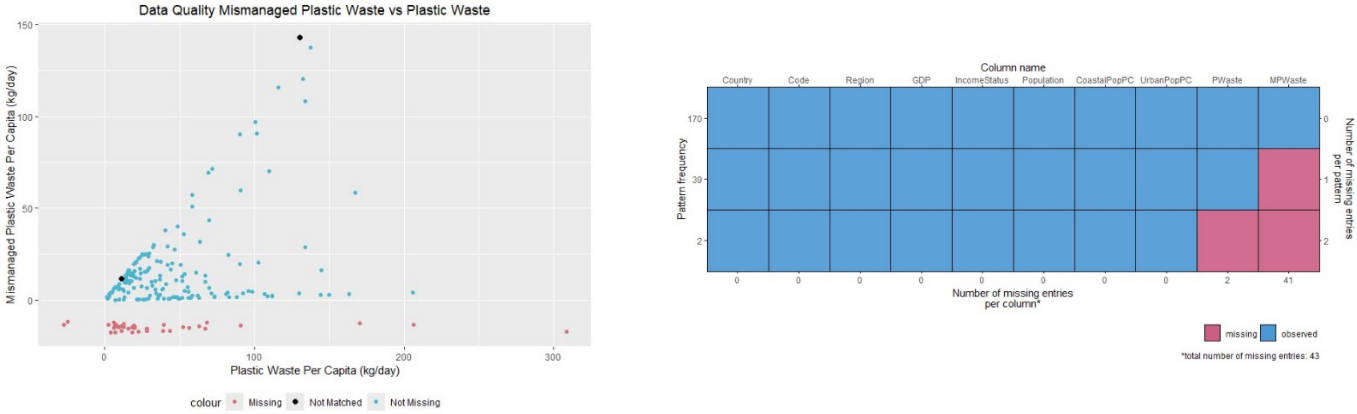

Figure 2: Scatter Plot of Mismanaged Plastic Waste vs Plastic Waste (Left) and Missing Values (Right)

Relevant Features

- The histogram plot revealed that both the distribution of plastic waste and mismanaged plastic waste are positively skewed (right skewed). This indicates that majority of the countries produce lower amounts of plastics compared to a few countries at significantly higher amounts.
- From the histogram and rug plot, we found that plastic waste per capita values exceeding 200 kg/day are potential outliners. These values require further investigation as they lie beyond the distribution.
- From the scatter plot, we highlight two points (black) that have unusual values with mismanaged plastic waste more than plastic waste per capita, which conflicts with the description of the dataset where mismanaged plastic waste is a subgroup of plastic waste, suggesting a potential error during data entry.
- From Figure 2, there are a total of 43 missing values with most of the missing values coming from Mismanaged plastic waste data at 39 data points. Other four missing values came from two countries that do not have both plastic waste and mismanaged plastic waste data. These missing values could be a result of these countries' data collection process.

Relevant Implications for Future Data Analysis

- The histogram indicated that certain countries have significantly higher plastic waste than other countries. We could explore further the effects of GDP, income status, and region of each country in both plastic waste variables as developed countries are more industrialized, making them more likely to produce more plastic waste.
- Moreover, the scatterplot hints at a potential relationship trend between plastic waste and mismanaged plastic waste, for each country. Subsequent studies into other variables such as GDP and population variables could aid in identifying the relationship within the data.
- For modeling, the missing values would pose a challenge to accurate modeling for mismanaged plastic waste. Subsequent studies should be conducted to handle missing values either by imputation of missing values based on the context of the county or deleting the whole row.

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

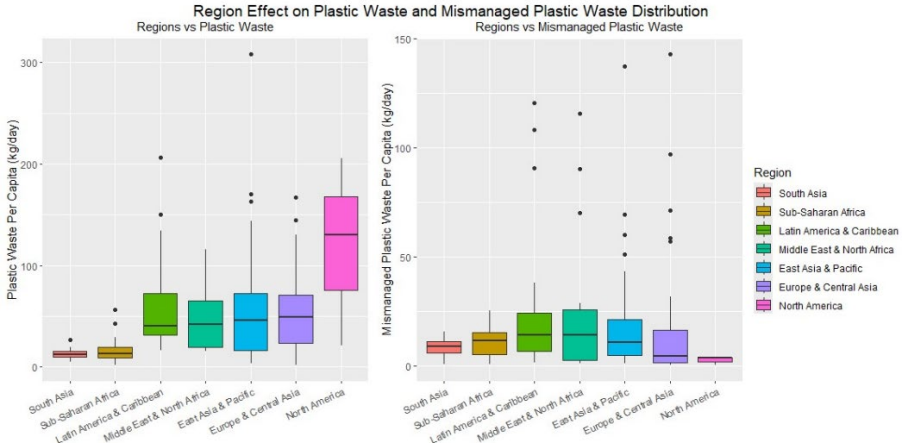

Figure 3: Region Effect on Plastic Waste and Mismanaged Plastic Waste Distribution

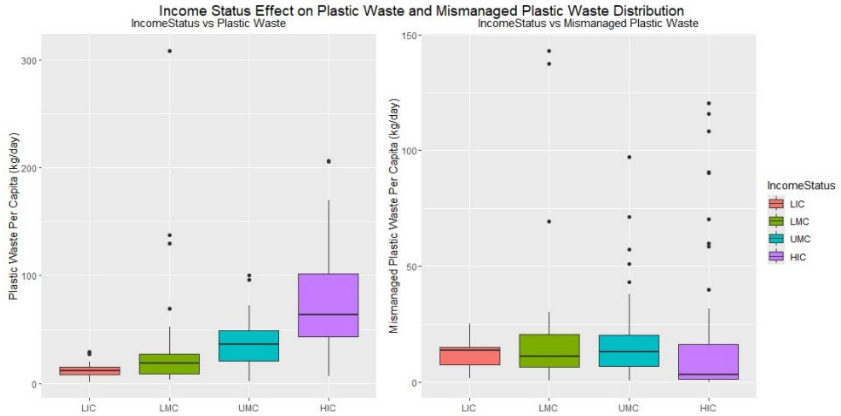

Figure 4: Income Status Effect on Plastic Waste and Mismanaged Plastic Waste Distribution

Relevant Features

- The boxplots for plastic and mismanaged plastic waste revealed that North America produces the highest amount of plastic waste out of all the regions. Nevertheless, the level of mismanaged plastic waste in this region is relatively lower compared to other regions. This indicates that the countries in this region have effective plastic waste disposal strategies.
- The boxplots of plastic and mismanaged plastic waste indicated that higher the income the more plastic waste produced with high income countries (HC) as the main producer of plastic waste. Nevertheless, HC countries tend to have lower level of mismanaged plastic waste, indicating effective plastic waste disposal strategies.

Relevant Implications for Future Data Analysis

- Subsequent studies into the effects of both income status and regions could provide additional insights towards the dynamic of the plastic waste and mismanaged plastic waste, especially for the high-income countries. This is due to the high variance in their mismanaged plastic waste.
- For predictive modelling, integration of income status and a variable indicating that a country is from the North America region could provide more accurate predictions as higher income countries produce more plastic waste.

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

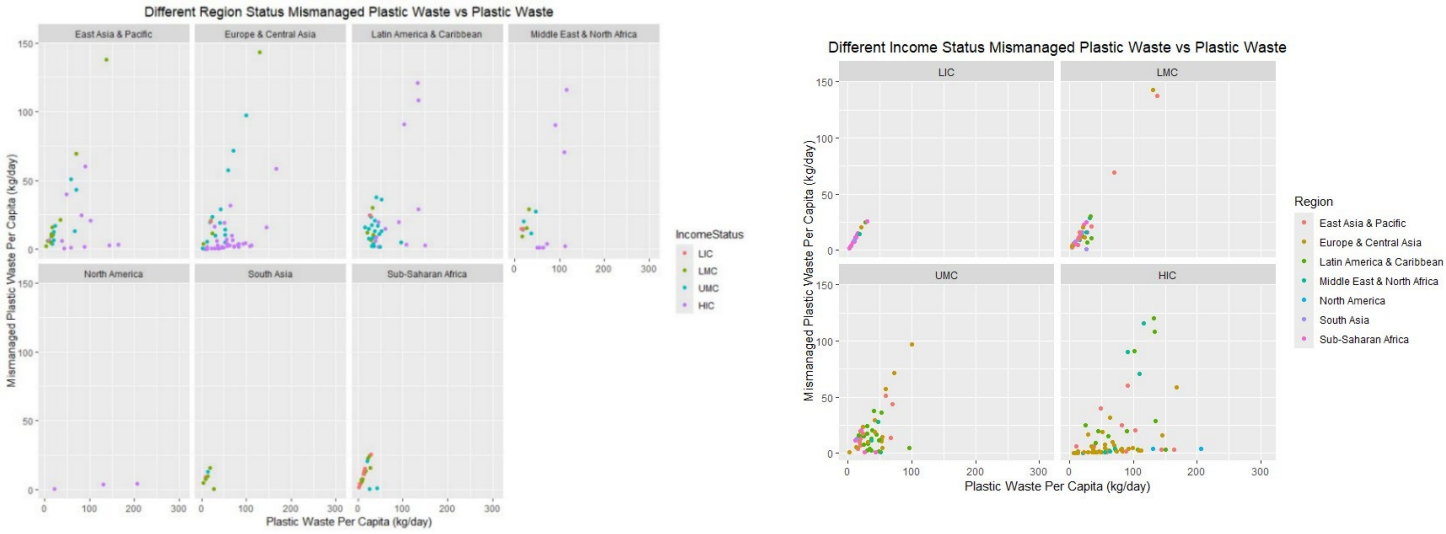

Figure 5: Mismanaged Plastic Waste vs Plastic Waste by Region (left) and by Income Status (right)

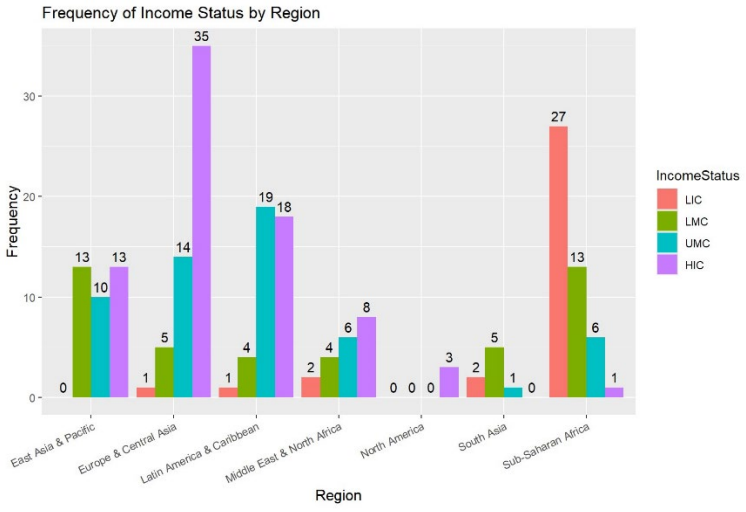

Figure 6: Barplot of Income Status by Regions

Relevant Features

- The scatter plot by region illustrates that there are linear correlations between mismanaged plastic waste and plastic waste within each region. Two linear trends are detected with low and high incline slope, indicating the varying mismanaged plastic waste disposal in each region. Highlighting the data

points with income status provides the insights that most of the high-income countries in each region have a relatively gradual increase in mismanaged plastic waste with an increase of plastic waste. In contrast, other income status countries, have steeper increases.

- The scatter plot by income status illustrates that high-income countries exhibit variability in the scatter plot with two identifiable linear correlation trends: high and low incline slope. This could be attributed to varying plastic disposal policies in high-income countries. Other income-status countries have a more defined linear correlation between the two wastes with a steeper slope. This indicates that other income-status countries have ineffective waste disposal system, resulting in a steeper increase in mismanaged plastic waste.

Relevant Implications for Future Data Analysis

- For modeling, the incorporation of income status as one of the variables in predictive modeling would provide better prediction of the plastic waste. As from the scatter plot by region, it demonstrates a relatively linear correlation.
- Subsequent investigation on other variables within the data set for high income countries could aid in identifying other factors which could related to the two distinct linear trends of this income status.

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

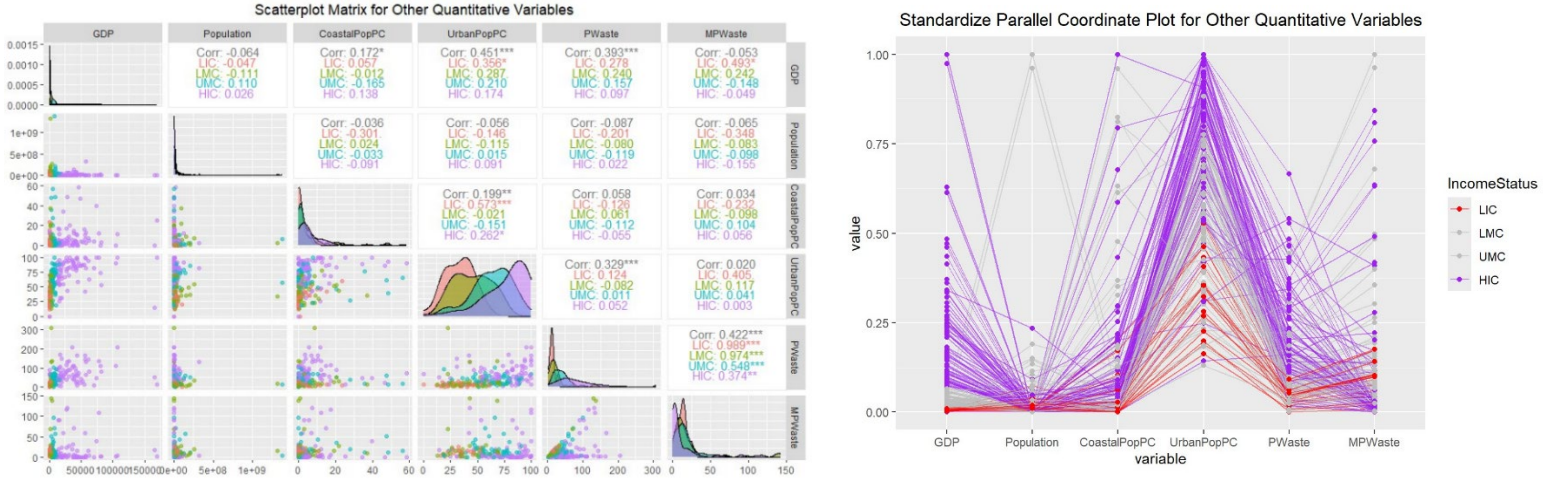

Figure 7: Relationship between Mismanaged Plastic Waste and Plastic Waste with other Quantitative Variables: Scatterplot Matrix (Left), Standardize Parallel Coordinate Plot (Right)

Relevant Features

- Colored by income status, the scatterplot matrix revealed that there is a low positive correlation between GDP and Urban population towards plastic waste. Moreover, GDP is low positively correlated to Urban population. This indicates that most of the plastic waste came from urban areas.
- From the parallel plot, high-income countries are characterized with have high GDP, urban population, and plastic waste produced. There are notable variations in the mismanaged plastic waste. This suggested that, despite being classified as developed country through the high GDP, these nations employ different approaches towards plastic waste disposal.
- Low-income countries characterized by low GDP level exhibits higher level of mismanaged plastic waste on the standardize scale, indicating ineffective plastic waste disposal.

Relevant Implications for Future Data Analysis

- From the analysis, population and costal population variables does not provide a clear relationship with the wastes. Thus, these variables could be excluded from modelling.

- For modeling, the incorporation of GDP and Urban population, which positively correlates to the level of plastic waste generated, could enhance the predictivity of the model.
- As this dataset collected from countries around the world, the variation in the geographical setting would result in landlocked countries with no costal population. Therefore, subsequent studies should investigate this geographical variation by highlighting country by with and without costal population.

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.

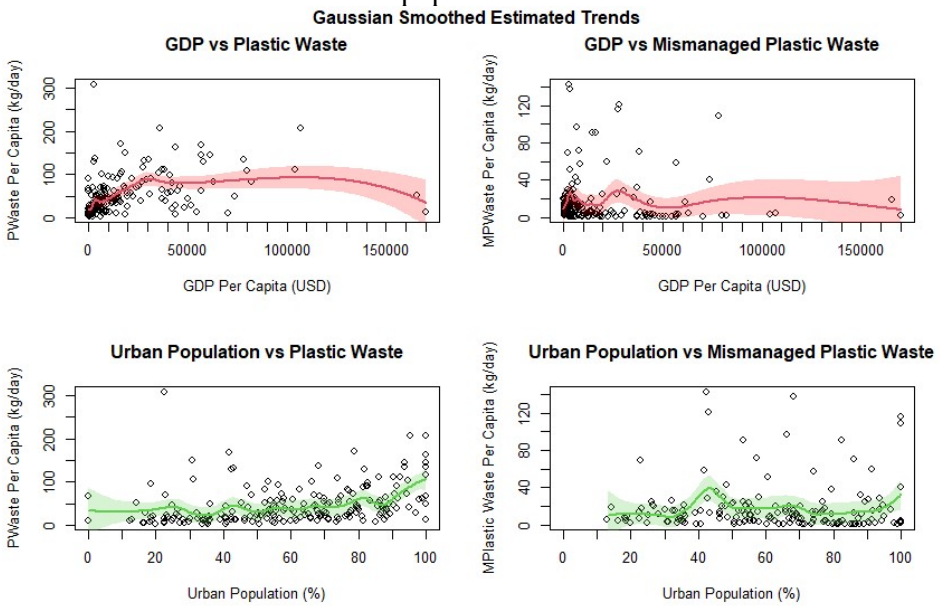

Figure 8: Gaussian Smoothed Estimated Trends for GDP and Urban Population vs Plastic Waste and Mismanaged Plastic Waste.

Relevant Features

- Overall, the smoothed trend for urban population and plastic waste indicates that there is a positive correlation with the increase in GDP per capita with two distinct slopes at GPD per capita around 30,000 USD and around 100,000 USD. For the mismanaged plastic waste, the plot reveals a similar slope at 30,000 USD. This implies that the higher GDP per capita countries generate more waste, but their plastic waste management is effective, resulting in a declining trend for mismanaged plastic waste.
- The plots of GPD vs plastic waste reveal that at a low level of GPD per capita, there is a sharp fluctuation in plastic waste and mismanaged plastic waste. This indicates that low GDP or low-income countries have ineffective plastic waste disposal.
- The smoothed trend for urban population vs plastic waste reveals that there is an increasing trend with increased plastic waste production. This indicates that the percentage of the urban population correlates with the increase in plastic waste.

Relevant Implications for Future Data Analysis

- Further investigation is required to evaluate possible outliners within the data for all the plots. As there are data points that lie significantly higher than other data points. For instance, in the urban population vs plastic waste, there is one value that produces around 300 kg/day of plastic waste, which is significantly higher than other data points in that region.
- For modeling, urban population vs plastic waste would be a suitable variable to predict the level of plastic waste as the trend indicates an overall positive correlation.
- To model mismanaged plastic waste, GDP vs mismanaged plastic could be used alongside regression spline to model the nonlinear relationship. However, further investigation is required to explore the large variance for high GDP per capita countries (e.g. GDP outliner removal).