Exploratory data analysis (EDA) and visualisation of plastic waste dataset

By Yan Lin, submitted on 13 March 2023

Q1: Explore the distributions of the two types of plastic waste. Investigate and comment on any notable outliers or unusual values. Investigate and comment on any missing values.

In total, I chose four visualisations: histograms, boxplots, normal Q-Q plots, and drawing patterns of missing values, plus statistical information.

Justification

Because histograms and density plots are the most effective methods for analysing the distribution of the two types of plastic waste. This not only provides a visual depiction, but also identifies any extreme or outlier values. Boxplots are another useful tool for visualising the distribution of continuous variables. They provide a quick way to visualise the median, quartiles, and range of the data, as well as any potential outliers. Normal Q-Q plots are a type of scatterplot that can help us determine if a variable is normally distributed and find symmetry. The pattern of missing values can also provide important information for understanding the quality of the data. By examining the pattern of missing values, we can identify any systematic bias in the data and make informed decisions about how to deal with missing values in our analysis.

Visualisation & Analysis

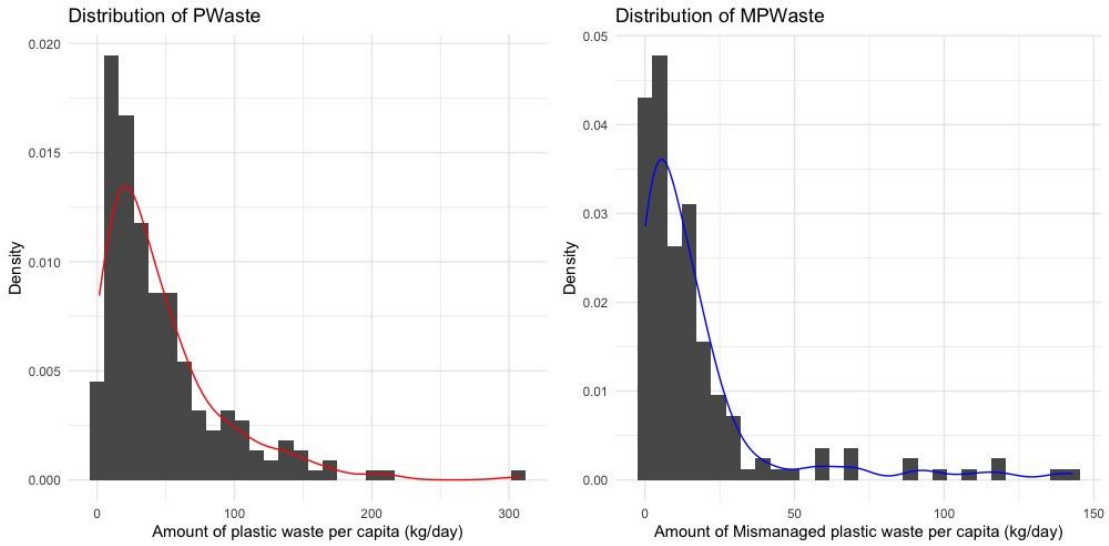

Figure 1: Histogram plus density plot for two types of plastic waste

According to Figure 1, we find that both types of plastic waste are skewed: positive skew, but the distribution of plastic waste would be more widespread than the distribution of mismanaged plastic waste, which is consistent with reality. Plastic waste has a peak on 0-50 while mismanaged waste has a peak on 0-25. There is a very large gap between 200-300 for plastic waste and a small gap between 100-200. There are 7 gaps between 50-150 for mismanaged

plastic waste. The histogram plus density plot combination is particularly useful for identifying multimodal distributions or distributions with long tails, which may be difficult to detect with a simple histogram. The Normal Q-Q Plot (Figure 2) confirms the results, with skewness in the head and tail, which shows strong curvature - not the straight-line feature we would expect if the data were Normally distributed.

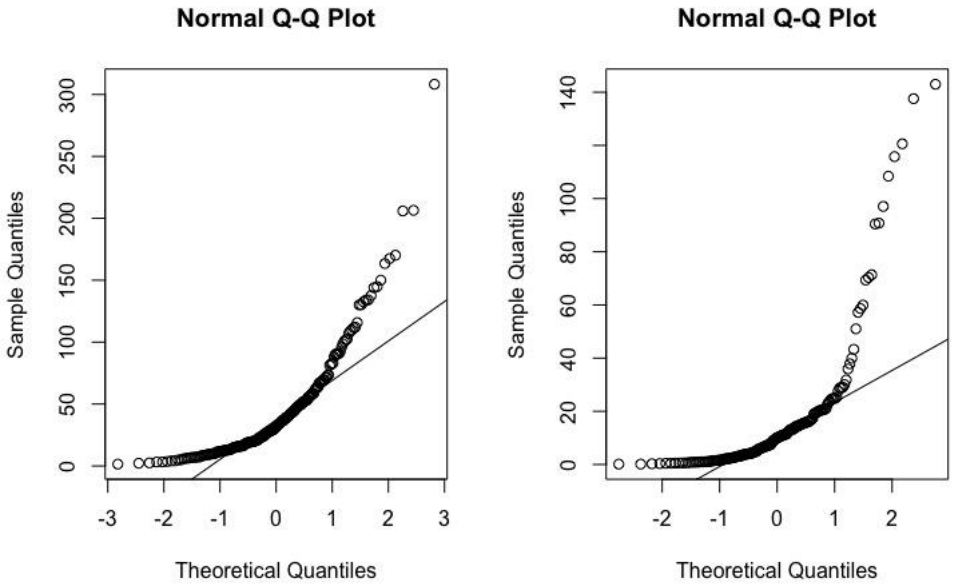

Figure 2: Normal Q-Q Plot for two types of plastic waste

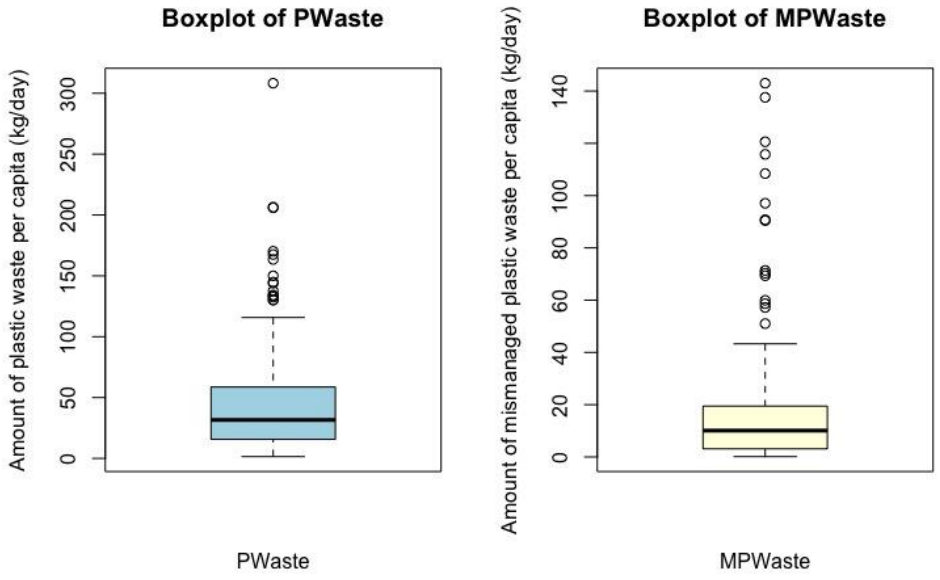

Figure 3: Boxplots for two types of plastic waste

Summary	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
PWaste	1.49	15.67	31.60	46.00	58.66	308.25	2
MPWaste	0.140	3.135	10.095	17.424	19.392	143.00	41

Table 1: Two types of plastic waste statistics

As seen in Figure 3 & Table 1, there are a dozen outliers outside the whiskers of both plastic waste boxplots. By calculating the 1.5 IQR, it is roughly estimated that 99% of the plastic waste amounts are between (0, 123.145) and 99% of the mismanaged plastic waste amounts are between (0, 44.01). The median plastic waste is 31.60 and the median mismanaged plastic waste is 10.095.

If the data has an approximately symmetrical distribution, the mean and median should be close, and this can be used as a quick check for any potential skew in the data. For plastic waste, there is a relatively large difference between the median of 31.60 and the mean of 46.00, indicating that its distribution is significantly skewed. For mismanaged plastic waste, there is a gap between the median of 10.095 and the mean of 17.424, but it is not large, indicating that its distribution is not as skewed as that of plastic waste.

Outliers and unusual values

There are three very significant outliers to these values in the distribution of plastic waste: 205.93 (Bermuda), 206.43 (Virgin Islands (U.S.)), and 308.25 (Micronesia, Fed. Sts.). Bermuda and the Virgin Islands (U.S.) are high-income countries with the potential for high amounts of plastic waste. Micronesia, Federated States is a low-income, small-population country with the largest plastic waste amount in the world, which is suspicious.

There are 15 outliers in the distribution of mismanaged plastic waste, with five significant outliers with amounts greater than 100: 143.00 (Moldova), 137.58 (Mongolia), 120.60 (Aruba), 115.77 (Kuwait), 108.43 (Cayman Islands). Moldova and Mongolia are low-income countries with inadequate sorting regulations for plastic waste. In Moldova, mismanaged plastic waste exceeds normal plastic waste. Aruba is a tourist destination and very touristy, resulting in bad waste management, but Kuwait has vast oil reserves, which might result in poor plastic waste management during oil extraction. However, the Cayman Islands is a tax haven with a relatively small population and the amount of mismanaged plastic waste is suspicious. In conclusion, boxplots can help us understand the distribution of the data more visually, including outliers and unusual values.

Missing values

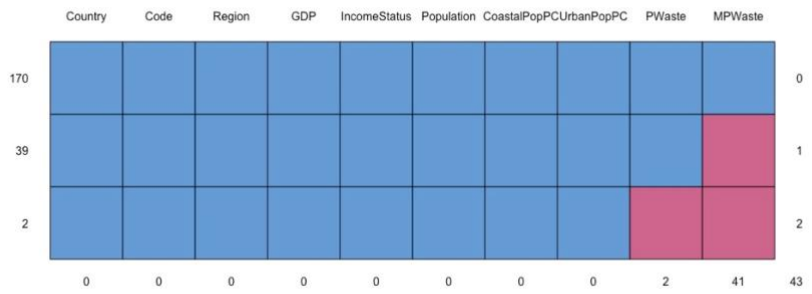

Figure 4: Missing value pattern chart

Here I plot the missing data patterns found in this plastic data. It has identified 2 patterns of missingness that are consistent with the statistical information. The first row has no red squares and represents the observations that are completed - we have 170 of these, as indicated in the label to the left. The next row is missing only MPWaste, as indicated by the red square in the final column, and we have 39 such cases. At the bottom, we find 2 cases which are missing PWaste and MPWaste. To improve the accuracy of interpolation and the reliability of the data, I use the predictive mean matching ('PMM') method in 2 missing values of PWaste and 41 missing values of MPWaste.

Q2: Explore whether and how the distributions of plastic waste and mismanaged plastic waste are affected by region and income status.

I have picked four visualisations: barplots, violin plots, boxplots, and combination graphs for this problem.

Justification

This is since bar charts are excellent for classifying variables and comparing distributions across groups. The violin plot is quite similar to the ridge plot in that it displays both the distribution features and density estimates of the data and is great for evaluating differences between each group of regions/ income statuses. For rigorous consideration, non-parametric tests are used to validate the observations. Observational comparisons are limited to coloured box plots due to the high number of subgroups of seven for regions. There are fewer subcategories for income status, so a combined plot can be used to highlight the data properties (histogram on the left and box plot on the right).

Visualisation & Analysis

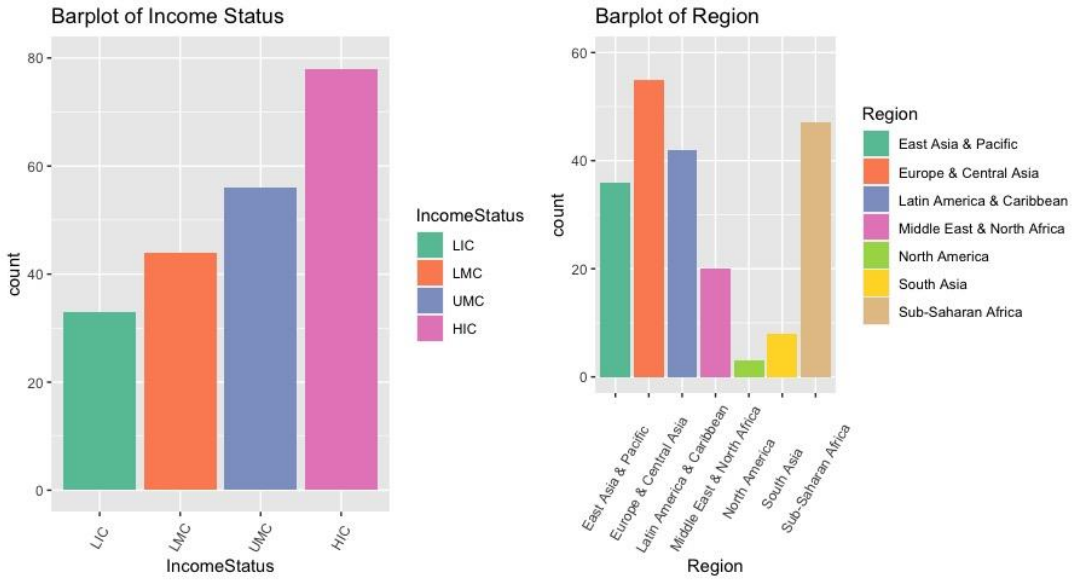

Figure 5: Barplots of income status and region

I draw barplots of income status and region (Figure 5), showing count data for each category. The counts for high income status are roughly twice as high as those for low-income status. North America is the least counted region and South Asia is the penultimate least counted region. The most counted regions are Europe & Central Asia. The barplots now give an initial indication of the distribution of income categories and regions, which will be very useful for our investigations below.

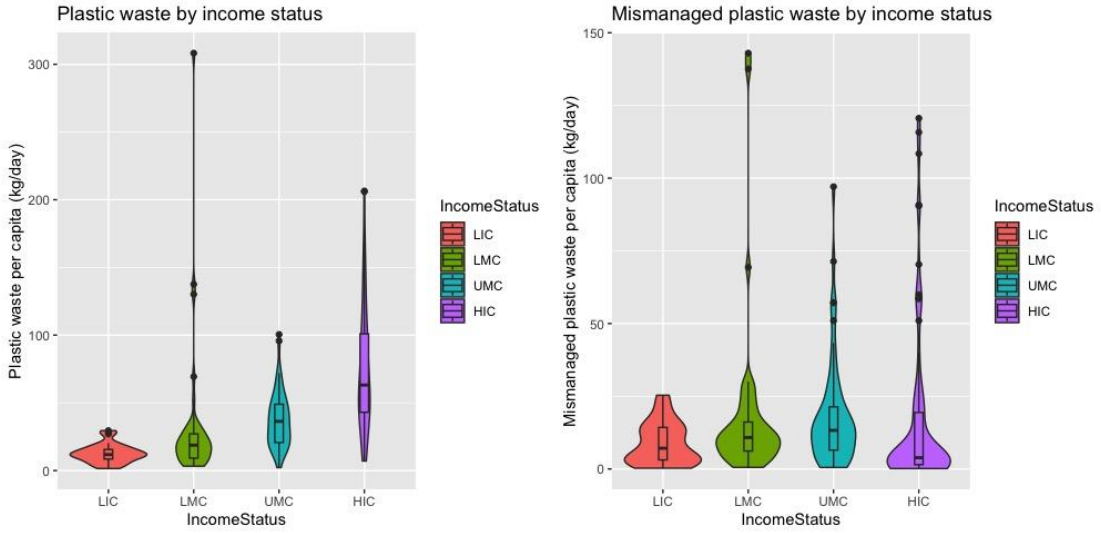

Figure 6-1: Violin plot for two types of plastic waste at different income statuses

As can be observed from the violin plot (Figure 6-1), the type of income status is responsible for influencing the distribution of the two types of plastic waste. For plastic waste(Figure 6-1 left), when it is at the low-income group (LIC), it is concentrated in the low values (0-20) and no value exceeds 50; when it is at the low- and middle-income group (LMC), however, there are more outliers and the highest value can exceed 300, even though most of it is still distributed in the low values (0-50). The distribution of plastic waste in the high-income group (HIC) will have slightly larger values than those of the middle- and high-income group (UMC), but neither group has a particularly concentrated value, and the distribution is relatively dispersed with outliers.

For mismanaged plastic waste (Figure 6-1 right), all four groups are concentrated in the low values (0-25), but with slightly different trends of concentration and different mean values. The LIC group has the largest stacking at the bottom of the data, indicating that most values in this group are located here. And the LMC group has the largest number of outlier values and the HIC group has largest number of exceptions. Although the difference is not as significant as with plastic waste, there is still a difference; therefore, both types of plastic waste are affected by income status.

Since the distributions of both types of plastic waste are not normal (as demonstrated in the first question), it is more robust and reliable to use a non-parametric test. I use the Kruskal-Wallis test and pairwise comparison tests (Dunn's test) to compare the distributions of the two types of plastic waste for different income status groups. By examining the p-values, the p-value for PWaste is much less than 0.05 and the p-value for MPWaste is 0.0016 slightly less

than 0.05, verifying the existence of two types of plastic waste whose distribution differs significantly across income status, i.e., is influenced by income status.

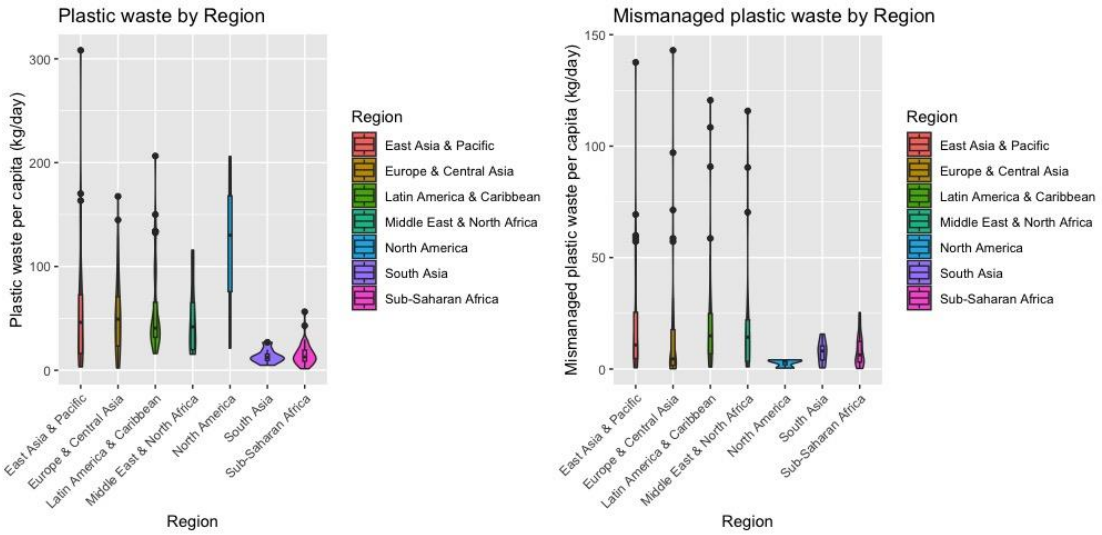

Figure 6-2: Violin plot for two types of plastic waste under different regions

The violin plots of plastic waste by region (Figure 6-2) reveal some interesting patterns.
For plastic waste (Figure 6-2 left): In ‘East Asia & Pacific’, ‘Europe & Central Asia’, and ‘Latin America & Caribbean’, the distribution of plastic waste is highly skewed and elongated, with outliers respectively. In ‘Middle East & North Africa’ and ‘North America’, there are also highly skewed and elongated distribution, although it is not as long as in the first three regions, without outliers. In ‘South Asia’ and ‘Sub-Saharan Africa’, both have highly concentrated distributions within the range of 0-30, but the latter has two outlier values. These findings suggest that there are significant regional differences in plastic waste distribution, with some regions exhibiting much higher amounts of plastic waste and greater variability than others.

For mismanaged plastic waste (Figure 6-2 right): In 'East Asia & Pacific', 'Europe & Central Asia', 'Latin America & Caribbean', and 'Middle East & North Africa', the amounts in these four regions are very dispersed and each group contains several outliers. In 'North America', 'South Asia', and 'Sub-Saharan Africa', all these three regions show varying trends of concentration, and all have very low numbers 0-25. ‘North America’ has the lowest numbers between 0-5 and the greatest concentration, followed by South Asia, and Sub-Saharan Africa has the weakest concentration of the three groups, but greater than the previous four regions. When broken down by region, regions do not have as much influence on the distribution of mismanaged plastic waste as they do on the distribution of plastic waste but are still influenced. The p-values of the statistical tests again validate this conclusion. The violin plot is most effective when comparing distributions across multiple sub-groups of the data.

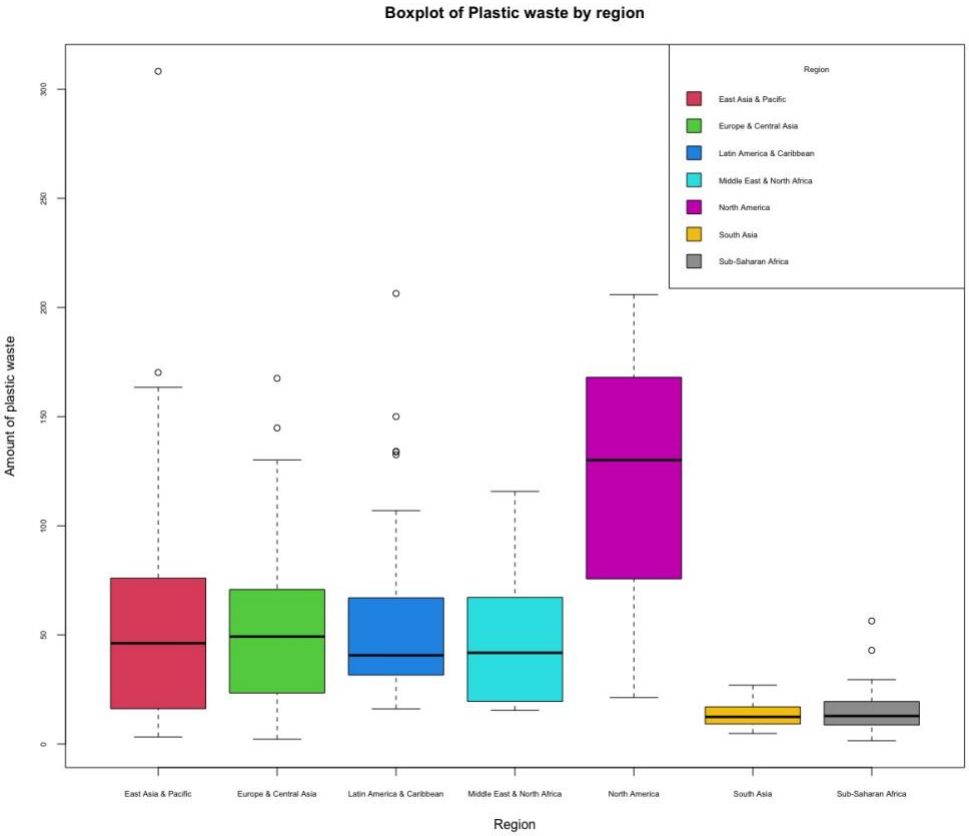

Figure 7-1: Boxplot of plastic waste under different regions

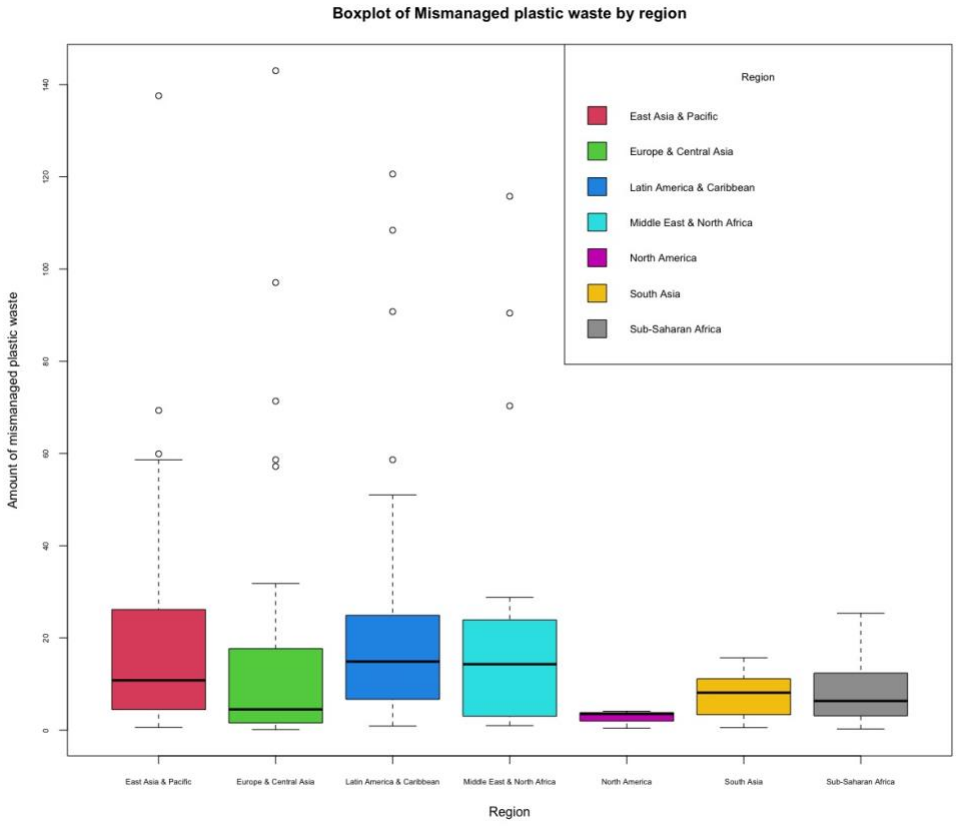

Figure 7-2: Boxplot of mismanaged plastic waste under different regions

Because there are seven groupings of regions, which is relatively large, a combination diagram is not appropriate. Looking through the different colours in the boxplots (Figure 7-1 & 7-2), it is also easy to observe that both types of plastic waste are affected by the regions, and how they are influenced is also consistent with the analysis in the violin plots.

The income status has fewer groupings, only four, and is therefore well suited to combination plots (two different plots of the same variables: a histogram and a box plot, in Figure 8-1 & 8-2), emphasising the data characteristics. We arrive at the same analytical conclusions as in the violin plots.

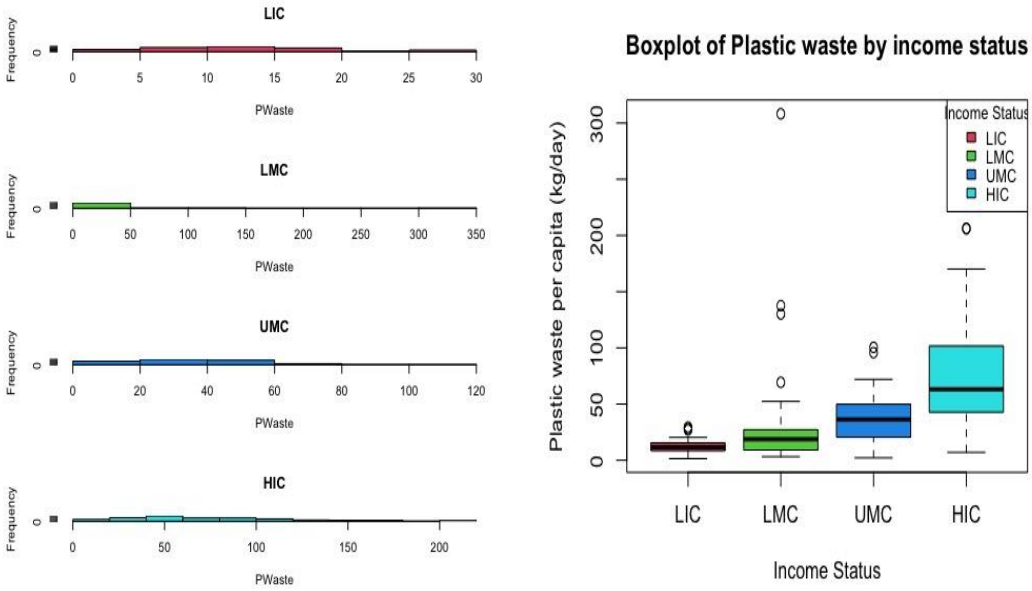

Figure 8-1: Combination plot of plastic waste in different income status

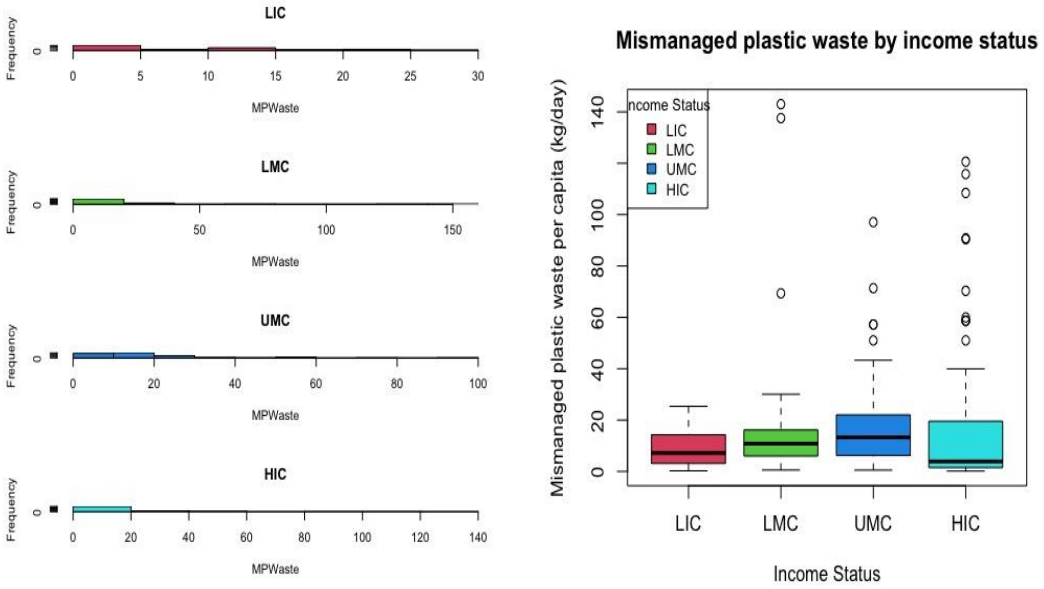

Figure 8-2: Combination plot of mismanaged plastic waste in different income status

Q3: Explore the relationship between plastic waste and mismanaged plastic waste. Are there any substantial differences between region and/or income status with respect to how the plastic waste variables are associated?

I chose two visualisations: scatter plots, as well as separate mini-scatter plots (lattice) for each group, corresponding to the statistical tests.

Justification

When faced with two numerically continuous variables, the most intuitive and easiest way to explore the relationship between them is the scatter plot, which shows the overall relationship between the two. This is followed by colouring them according to income status and regional classification, and seeing these three scatter plots provided me with a first impression. But the data may be overly dense because there are so many groupings, I consequently optimised this using xyplot in lattice: plots a separate mini scatter plot for each subgroup. 'xyplot' is ideally suited for analysing data sets including numerous distinct groups between two continuous variables. Lastly, correlation and statistical tests are used to confirm the conclusions once more. Parallel coordinate plots are disregarded for this question since they are better suited for examining relationships between three or more continuous variables.

Visualisation analysis

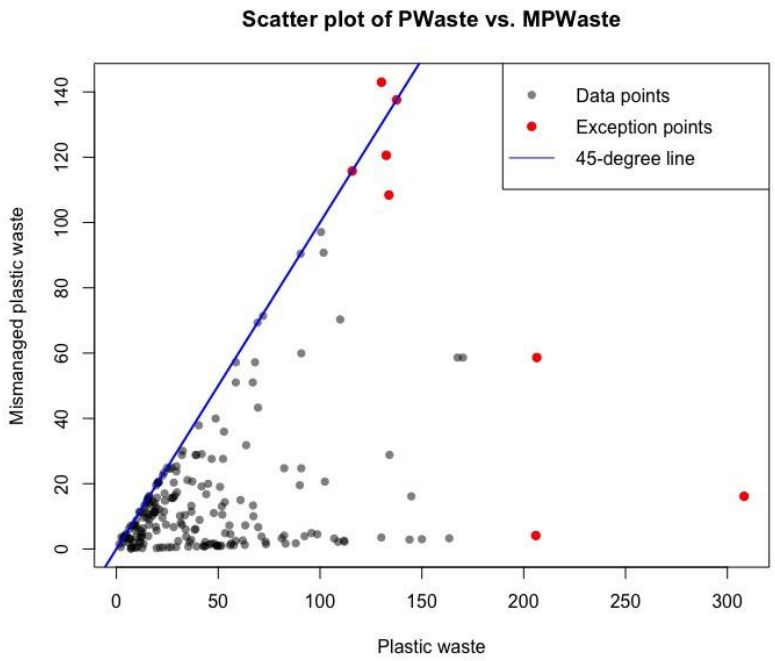

Figure 9: Scatterplot between the two types of plastic waste

Figure 9's scatterplot demonstrates that there must be variability between plastic waste and mismanaged plastic waste. High density areas become darker and more unusual values fade away. There are exception points for both types of plastic waste, and according to the 45-degree line, if most of the points were on the line, it would indicate that there is little difference between

the two variables. However, it is evident that most of the points are not on the line, indicating that there must be some kind of association.

Income Status	Mean PWaste	Mean MPWaste
LIC	12.8	8.85
LMC	31.0	18.5
UMC	37.6	18.5
HIC	74.4	17.2

Region	Mean PWaste	Mean MPWaste
East Asia & Pacific	61.1	21.8
Europe & Central Asia	52.2	14.8
Latin America & Caribbean	58.2	22.8
Middle East & North Africa	49.9	23.5
North America	119	2.69
South Asia	13.6	7.68
Sub-Saharan Africa	15.2	8.30

Table 2: Average of two types of plastic waste grouped by Income Status/ Region

According to Table 2, comparing the means of the two variables across income status and across regions also reveals a large difference, so there are substantial differences on the plastic waste variable in terms of income status and region.

The scatterplot (Figure 10-1 & 10-2) will be more intuitive in terms of the relationship between the two variables if various colours are assigned to distinct groups based on income status and regional groupings. Yet, in the rectangle area containing PWaste 0-50 and MPWaste 0-20, there are so many points that it becomes difficult to observe.

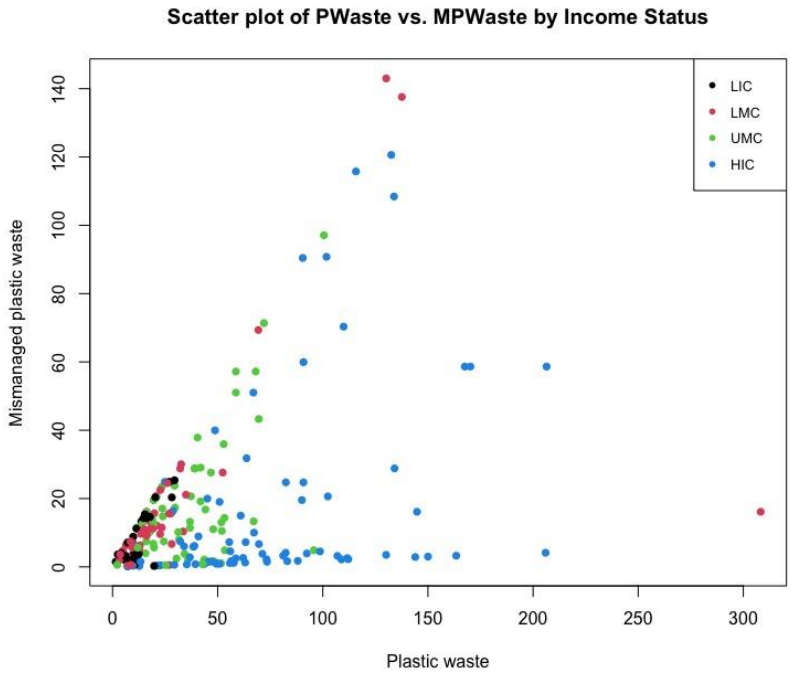

Figure 10-1: Scatterplot of grouped by income status between the two types of plastic waste

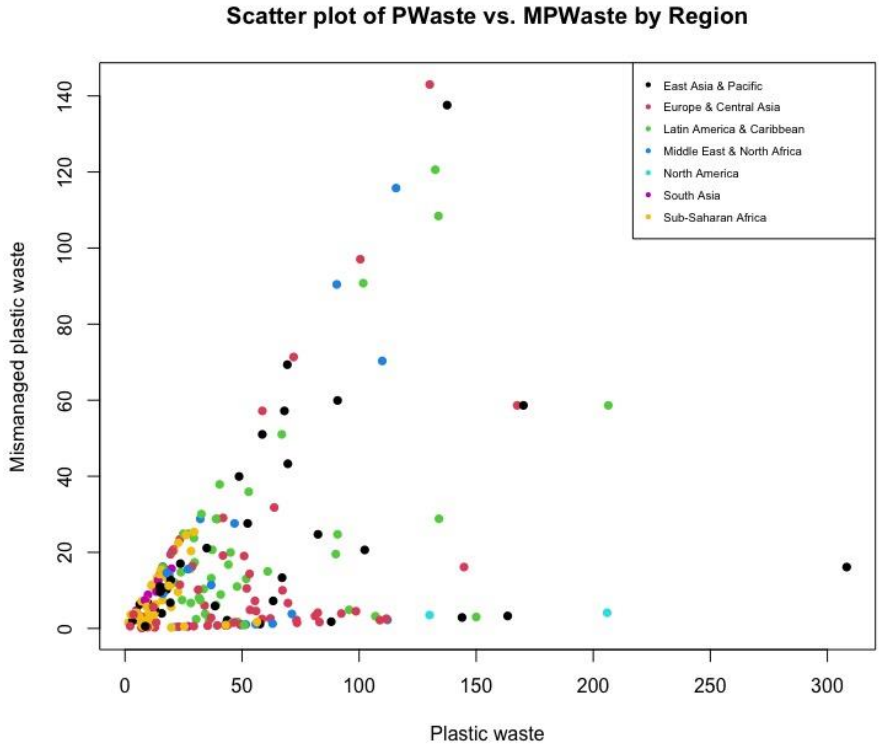

Figure 10-2: Scatterplot of grouped by region between the two types of plastic waste

'xyplot' simplifies the problem of having to distinguish between points of different colours and generates a scatter plot of the data for each subgroup. It automatically uses common x and y axes to ensure that the plotted points are comparable in terms of scale and position.

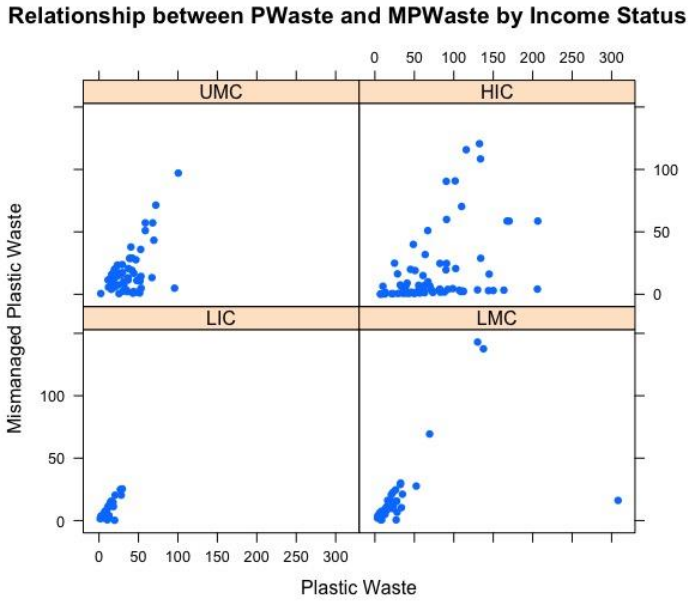

Figure 11-1: Mini scatter plots of two types of plastic waste: subplots for each income status

Figure 11-1 demonstrates that there is a strong correlation between the two types of plastic waste and a high concentration in the LIC group, as the countries in this group all have very

low incomes, lack waste disposal facilities, and lack environmental education, leading to inappropriate waste disposal behaviour. However, in the LMC group, this correlation is much lower, probably due to an exceptional value. While there are only four outlier values in this subplot, a deeper examination at the concentrated portion reveals that the pattern of the decentralised subplot is more similar to that of the LIC group, leading to a relatively high correlation between the two types of variables.

In contrast, there is also some association between the two types of plastic waste in the HIC and UMC groups, although the data points are more dispersed, resulting in a lower correlation. This may be because high-income countries have highly developed waste disposal systems and a strong environmental protection mindset, and its citizens are less likely to have badly handled plastic waste. The significance tests' p-values also supported these findings (significant for all four income status groups, Table 3-1). In conclude, there is a considerable relationship between plastic waste and mismanaged plastic waste, and income status can make a substantial difference to this relationship.

Income Status	Correlation	p-value	Significance
LIC	0.799910392438314	$2.33320064754899e-08$	Yes
LMC	0.570486074833101	$4.42536547715379e-06$	Yes
UMC	0.428660919544234	$9.02150533416202e-05$	Yes
HIC	0.510420685371476	0.000400415511211853	Yes

Table 3-1: Correlation and statistical tests for plastic waste variables (subgroup: income status)

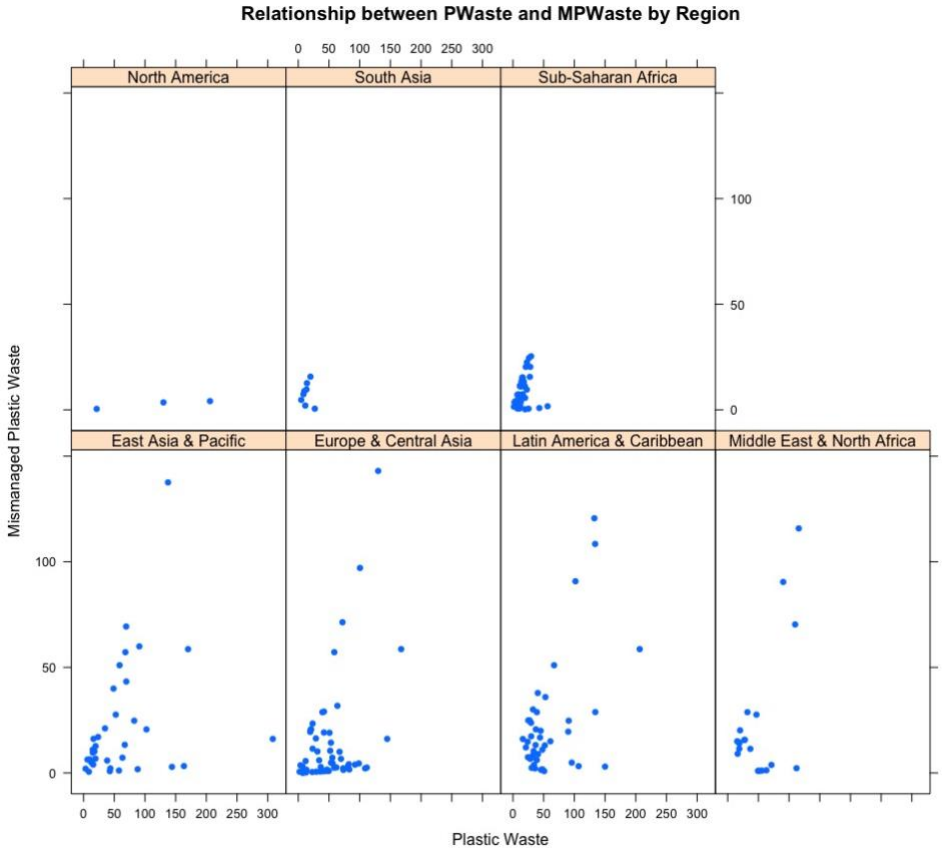

Figure 11-2: Mini scatter plots of two types of plastic waste: subplots for each region

As seen in the first two subplots of the xyplot (Figure 11-2), there is a strong linear relationship between the two types of plastic waste; however, the supporting data points are quite small (North America: 3 points; South Asia: 8 points) and, as a result, are insufficiently dependable.

The third and fourth subplots show that there is a similar pattern between the two types of plastic waste, that there is a positive correlation, and that there are more data points. Nonetheless, when comparing the Sub-Saharan Africa region and the East Asia & Pacific region, there is a very large gap between the countries in these two regions, which have very large developed economies, so the finding of a positive correlation between the two types of plastic waste is not stable.

The fifth, sixth, and seventh subplots show an obvious positive association with a high number and concentration of data points. There is a slight difference between the countries in these three regions, which are either highly developed or extremely underdeveloped, thus it is probable that there is a positive correlation between the two types of plastic waste in all three regions.

These findings can be confirmed by calculating correlation coefficients and significance tests (Table 3-2). In summary, there is a strong relationship between plastic waste and poorly managed plastic waste, and region can make a substantial difference to this relationship.

Region	Correlation	p-value	Significance
East Asia & Pacific	0.319014316028615	0.0579044737397964	No
Europe & Central Asia	0.435589559269541	0.000887716002915371	Yes
Latin America & Caribbean	0.530753151061148	0.000299291162531649	Yes
Middle East & North Africa	0.57292919628395	0.00827922132478403	Yes
North America	0.964787965340078	0.1694428917748	No
South Asia	-0.0281267042285069	0.94729023716895	No
Sub-Saharan Africa	0.31616342965912	0.0303865536700997	No

Table 3-2: Correlation and statistical tests for plastic waste variables (subgroup: region)

These results are based on correlations between individual variables only and do not allow causal relationships between causes and effects to be identified. To better understand these relationships, we need to investigate other variables.

Q4: Investigate whether there an association between plastic waste and the other variables. Is there any evidence of strong associations that may be helpful for modelling?

Regarding the other variables, I will only explore the four variables 'Population', 'GDP', 'CoastalPopPC', 'UrbanPopPC', 'Country' and 'Code' as text variables will not be discussed. I select four visualisations: Scatterplot Matrix Variations, Corrplot, Parallel Coordinate Plots (PCP), Heatmap.

Justification

As the scatterplot matrix includes a lot of redundancy, I have added histograms for each variable in the diagonal panel. Another variant is a smoothed version that again combines scatterplots, numerical correlation information and histograms. Effective combinations of graphics such as these can maximise the amount of information we can extract from a set of plots. In practise, however, the scatterplot matrix is frequently overloaded with variables and cases. Correlations provide a valuable numerical summary, and the corrplot is an effective method for presenting this data. Furthermore, PCP have become a popular tool for highly multivariate data. A PCP gives a quick overview of the univariate distribution for each variable, and can help us discover interactions between variables, clustering patterns and outliers. As there are four variables to compare, PCP is ideal in this situation. Heatmap also can help us identify any strong linear relationships between plastic waste and other variables.

Visualisation & Analysis

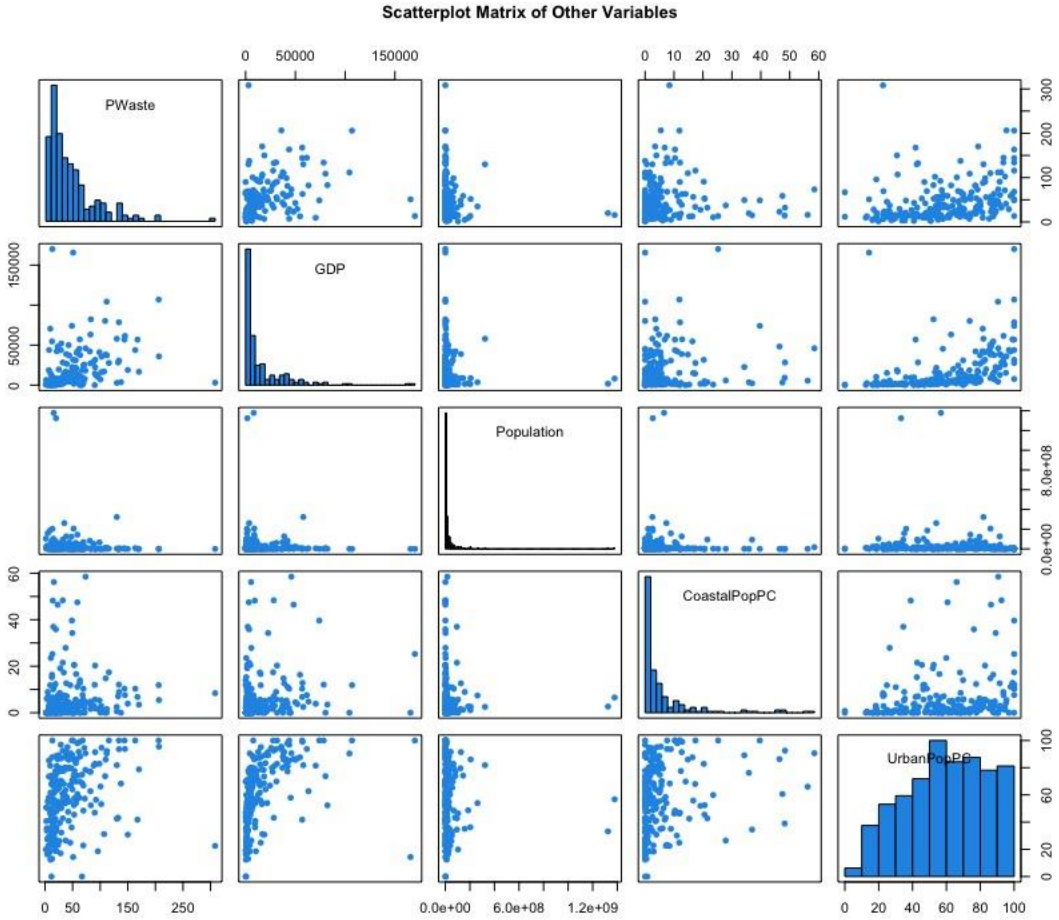

Figure 12: Scatterplot matrix on other variables

As seen in Figure 12, the variables "GDP", "Population", "CoastalPopPC" as well as "PWaste", have considerable positive skewness, with "Population" having the most significant skewness due to the large population variations between countries. Furthermore, there is a positive linear

association between "GDP" and "PWaste", as well as between "UrbanPopPC" and "PWaste". "Population" and "PWaste", "CoastalPopPC" and "PWaste" may have a nonlinear relationship. It is also feasible that the "GDP" and "UrbanPopPC" have a linear relationship.

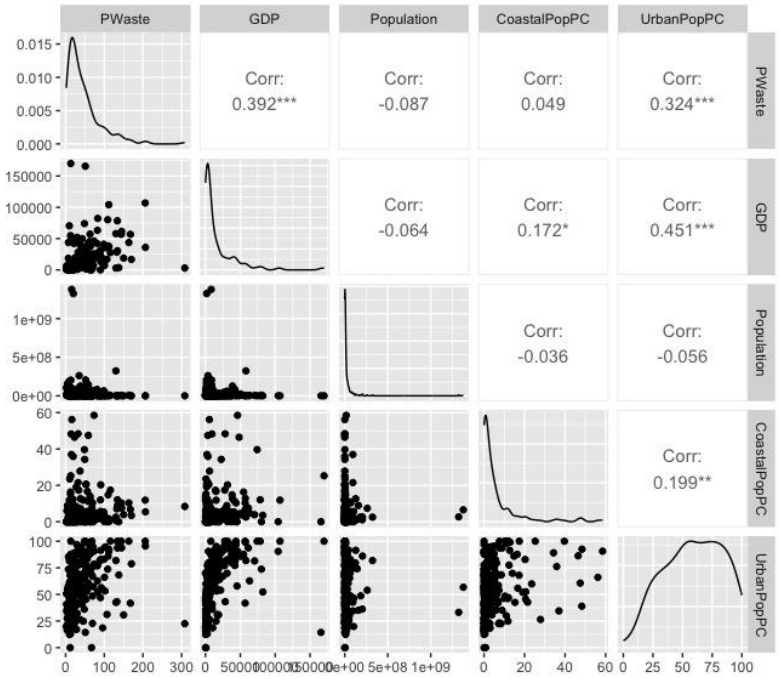

Figure 13: Scatterplot matrix smoothed version

We got a combination of correlation coefficients, scatterplots, and histograms using a smoothed version of the scatterplot matrix (Figure 13). These correlation coefficients confirm the aforementioned observation that there is a positive correlation between "GDP" and "PWaste", suggesting that richer countries produce more plastic waste, and that there is also a positive correlation between "UrbanPopPC" and "PWaste", with a higher percentage of urban population producing more plastic waste. And the GDP increases as the urban population percentage increases. Alternatively, there is a non-linear link between Population and plastic waste.

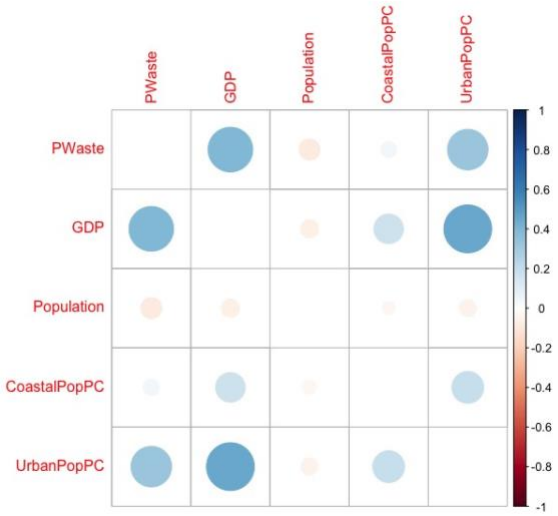

Figure 14: Correlation Plot

Interestingly, the correlation plot (Figure 14) reveals a moderately weak relationship between the proportion of coastal population and the proportion of urban population, and thus between the proportion of coastal population and GDP.

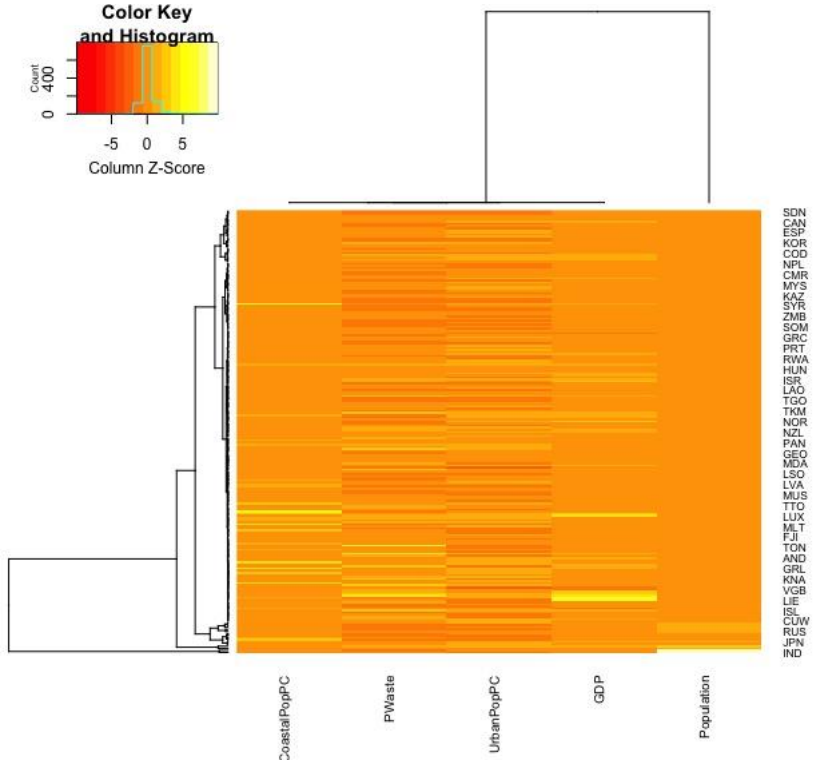

Figure 15: Heatmap

The colour of each column varies from red (low values) to yellow (high values). The rows in the data have been replaced to emphasise the differences between the different variables. Here (Figure 15) we can clearly observe the presence of yellow in the columns GDP and proportion of coastal population and plastic waste, with GDP having more yellow. Thus, we find that countries with high GDP usually have a lower coastal population ratio and more plastic waste. This allows for the introduction of a relative negative correlation between the proportion of coastal population and plastic waste.

In the scatterplot matrix it is not always clear how the points in each panel relate to the points in the other panels as only two variables are shown in each panel, so we try to apply a parallel coordinate plot.

According to Figure 16, surprisingly, it is the lower middle income countries that produce the most plastic waste, with a smaller total population, a smaller proportion of coastal population, and a very high proportion of urban population, resulting in a higher GDP, but where waste facilities and education have not kept up with the country's economy, resulting in the largest group of the four income statuses.

Another interesting finding is that an upper-middle income country produces the most plastic waste, the underlying reason being that it has the largest population in the world.

Because the regions are more grouped and coloured in the second parallel coordinate plot (Figure 17), it is more difficult to discern a distinct pattern.

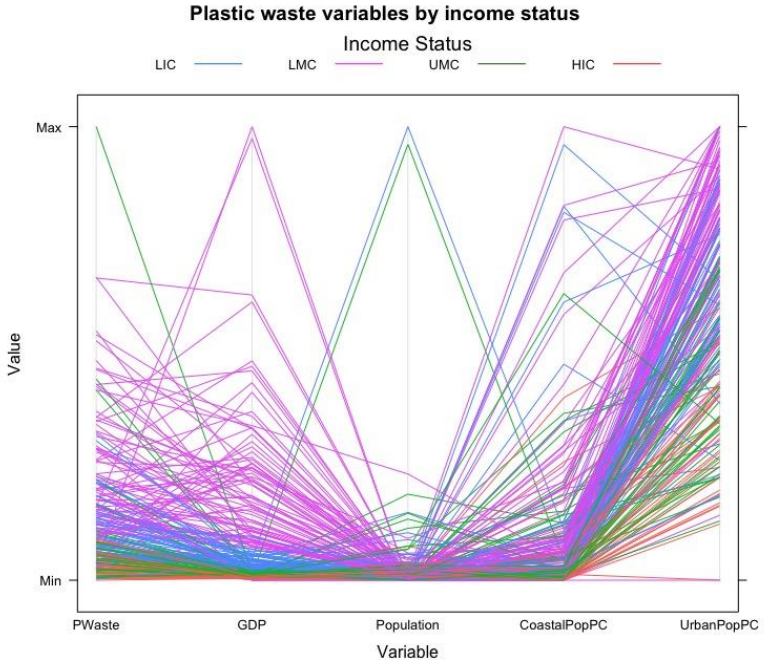

Figure 16: Parallel coordinate plot (plastic waste and 4 other variables, colour: income status)

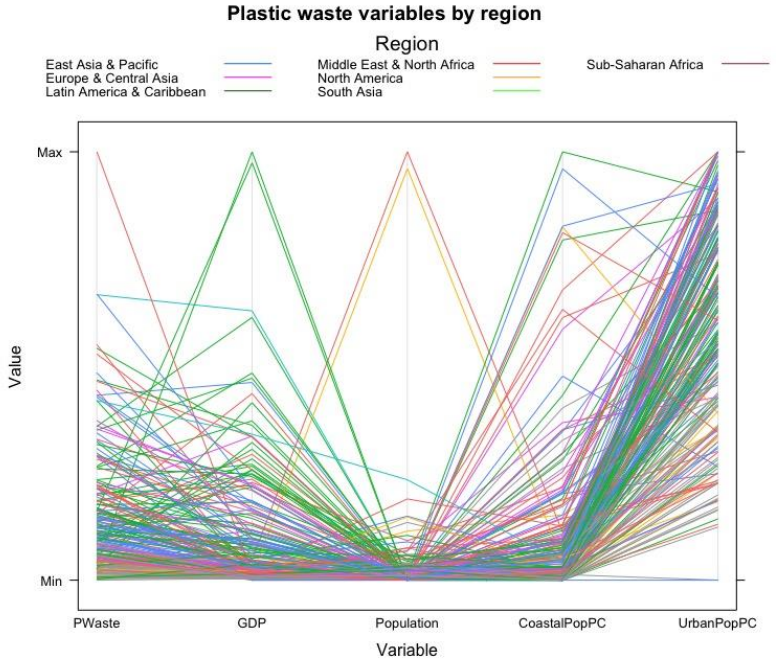

Figure 17: Parallel coordinate plot (plastic waste and 4 other variables, colour: region)

Overall, it may be helpful to include GDP, percentage of coastal population, and percentage of urban population as predictors in a model for plastic waste.

Q5: Explore the relationship between plastic waste and GDP (the wealth of the country), and the plastic waste and the size of coastal population. You should include plots of an appropriately smoothed trend as part of your answer.

As we desire to examine the relationship between plastic waste and coastal population size rather than coastal population proportion at this period, we must convert coastal population proportion to a percentage and multiply it by the corresponding population size; I have also created a new variable ("SizeofCoastal") to record coastal population size. I choose one visualisation: Scatter plot plus loess smoothed trend with 95% confidence interval.

Justification

Smoothing is an effective technique for estimating a trend of a data set without requiring complex modelling. Both LOESS (locally weighted scatterplot smoothing) and kernel smoothing are non-parametric regression methods that can be used to estimate a smooth curve through a set of data points. However, LOESS and kernel smoothing differ in the way they weight the data points to estimate the curve. In general, LOESS is a more flexible method than kernel smoothing because it allows for a varying degree of smoothing across the data range, whereas kernel smoothing uses a fixed kernel function that may not adapt as well to changes in the data. LOESS is also computationally efficient and does not require a pre-specified bandwidth, which can be a disadvantage of kernel smoothing. As LOESS just uses linear regressions to generate its smoothed trend, we can apply linear regression theory to obtain the standard errors of the trend and form confidence intervals around it.

Visualisation & Analysis

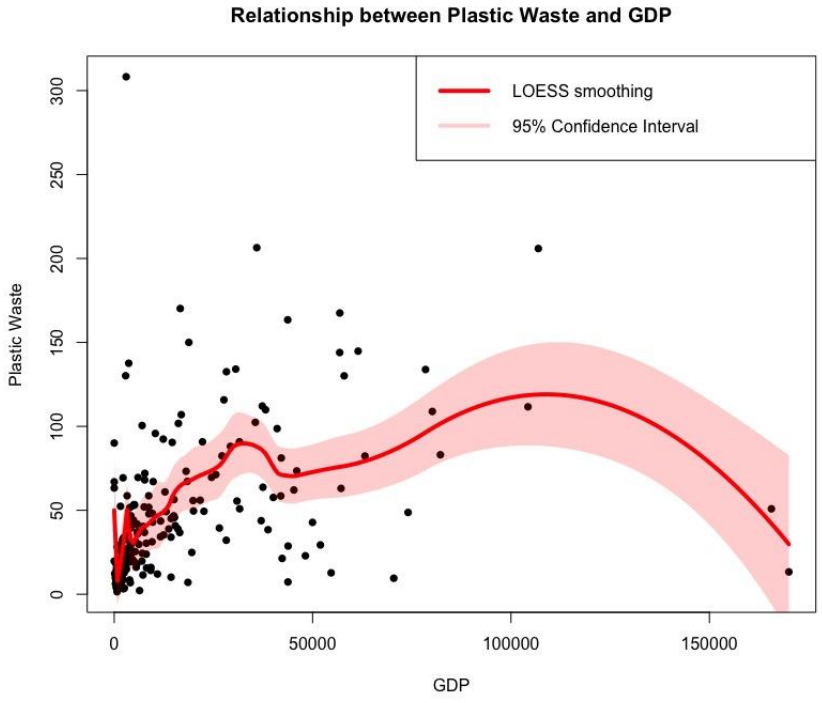

Figure 18: Smoothed trend scatter plot between plastic waste and GDP

In this Figure 18, the smooth trend line shows us the overall relationship between GDP and plastic waste, whilst the individual points show the differences between countries. We have a 95% confidence that this trend line will fall within this confidence interval during repeated similar data collection and analysis.

In areas with low GDP, the dots are extremely trending, and the trend is upward, indicating that in countries with small GDP, plastic waste and GDP are positively correlated, with high GDP and growing levels of economic development leading to people having more plastic waste.

Whereas when GDP exceeds a borderline number: 100000, the tendency begins to decline dramatically. There are only four points in this area, representing the four countries that match the negative correlation between GDP and plastic waste: the higher the GDP, the lower the plastic waste. This phenomenon belongs to those high-income countries where environmental education and waste facilities are exceptionally advanced, or where the population seems to be very small and, as a result, plastic waste is small.

In terms of the overall circumstance, the confidence intervals in Figure 18 are relatively narrow, and the proximity and strength of the trend line to the data points shows that this trend line accurately predicts the values of the data, and that GDP and plastic waste are in a relatively strong positive correlation.

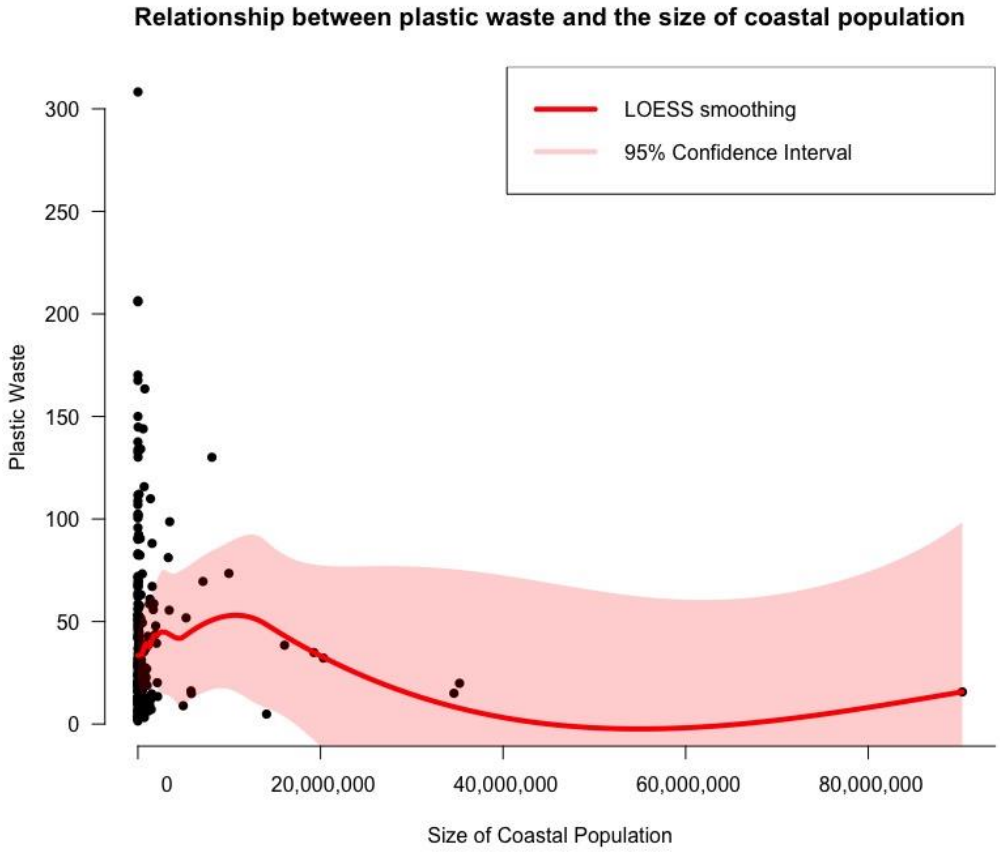

Figure 19: Smoothed trend scatter plot between plastic waste and size of coastal population

In this Figure 19, the smooth trend line shows us the overall relationship between the size of coastal population and plastic waste, and we are already 95% confident that the trend line will fall within this confidence interval.

When the size of coastal population is small, most of the dots are densely stacked in this area, demonstrating that in many countries when the trend line is generally flat and there is no evident correlation between the two variables. In contrast, as the size of coastal population begins to increase, the trend line begins to move slightly downwards and has a wider confidence interval and weaker trend line strength.

Considering the overall situation, the size of coastal population is negatively correlated with plastic waste. This suggests that as the size of coastal populations increases, plastic waste generation may decrease to some extent, although the correlation is relatively weak and not as strong as that between GDP and plastic waste.