MATH42515: Data Exploration, Visualization, and Unsupervised Learning
Assignment 1 Report

Anonymous Marking Code: Z0182576 Date: March 2023

1 Question 1

To allow for easy comparisons between the shape and sizes of the frequency distributions, the two types of plastic waste were plotted on two histograms with the same scales.

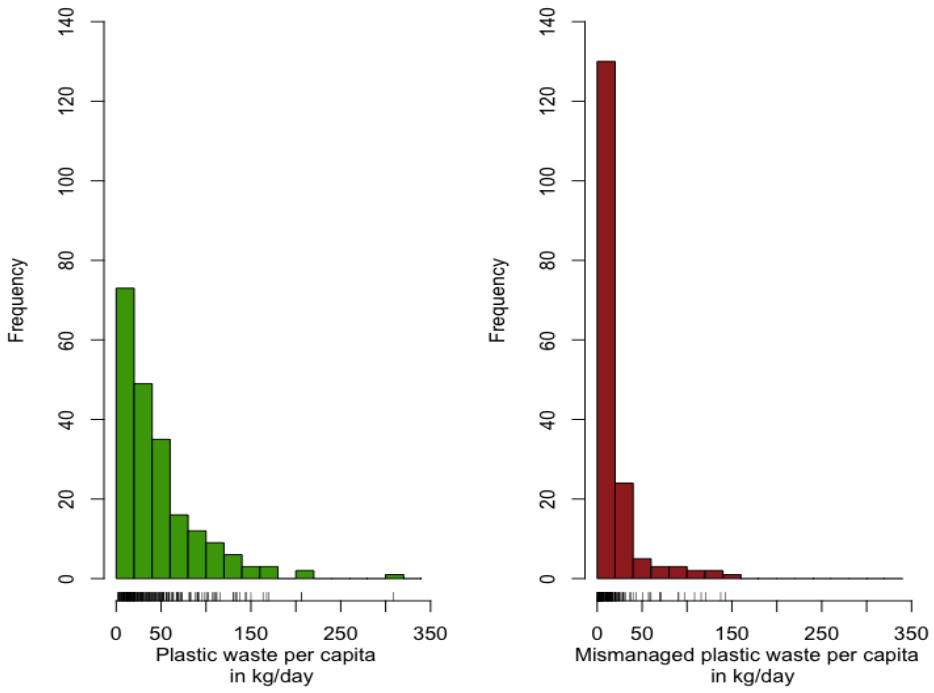


Figure 1: Histograms displaying frequencies of the amounts of plastic waste (left) and mismanaged plastic waste (right)

From these graphs, we see that the frequency distribution of the amount of mismanaged plastic waste appears to be more leptokurtic compared to the distribution for the plastic waste, with a very sharp peak corresponding to a high number of values in the range of $0 - 20$ kg/day. The frequency distribution tails off quickly, and no values exist beyond 160 kg/day, unlike in the plastic waste graph where a value of about 300 kg/day is found. In the plastic waste histogram, we also notice heaping in the $0 - 20$ kg/day range, although to a lesser extent than in the mismanaged plastic waste graph. Such heaping is expected as many countries do not place as high an emphasis on the production of plastic goods as others, and so there will be less plastic waste per person. More heaping in the mismanaged plastic waste distribution is not surprising as it is in a country's interest to reduce littering and the environmental damage that may come with mismanaged waste.

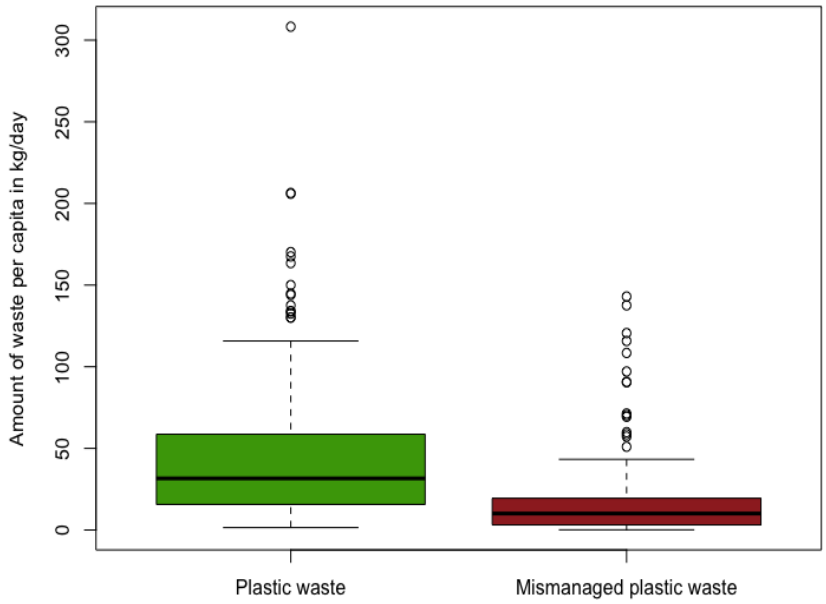


Figure 2: Boxplots showing summaries of plastic waste and mismanaged plastic waste data

The wider spread of the plastic waste distribution is emphasised with boxplots on the same limits. Note that the space in the box around the mean is larger in the plastic waste boxplot than in the mismangaged plastic waste boxplot. Furthermore, the boxplot allows us to easily identify outliers and extremes. For example, we have a value corresponding to 308.25kg/day of plastic waste per person, more than $3 \times IQR$ greater than the upper quartile. Since the corresponding value for mismanaged plastic waste is missing, we exclude this data point for our analysis of the relationship between the types of waste, as it will skew the results by not providing a paired data point. To further investigate for other outliers or errors, we use a scatterplot of the plastic waste against the mismanaged plastic waste.

A scatterplot allows us to directly plot the relationship between plastic and mismanaged plastic waste. This is important, as mismanaged plastic waste should be a subset of plastic waste and thus should be strictly less than the amount of plastic waste for the corresponding country. We plot the values for each country, and then plot the line $y = x$. If we notice that points lie on the side of the line that suggests that the mismanaged plastic waste is greater than plastic waste, then one of the recordings for that country must be an error.

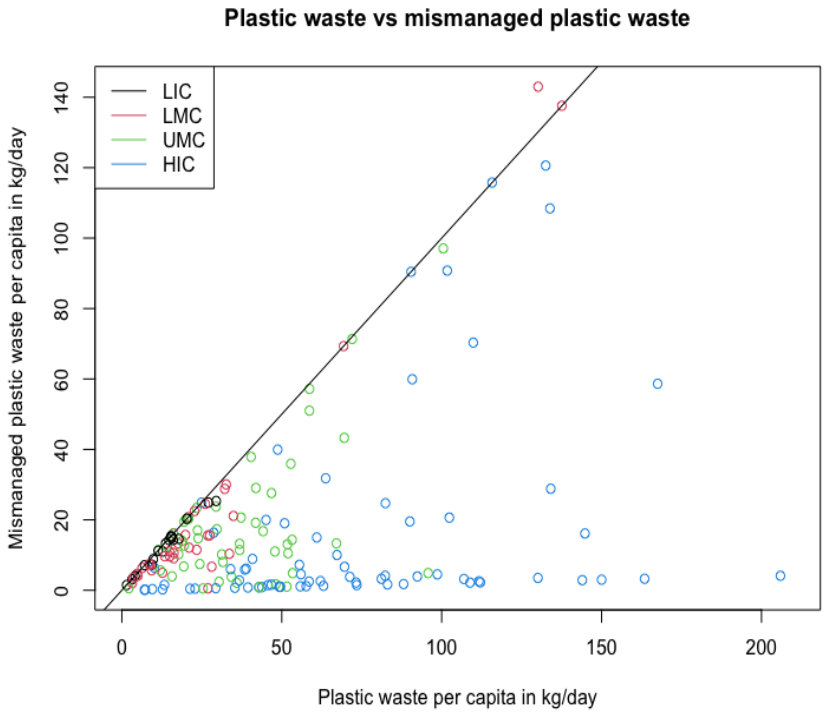


Figure 3: Scattergraph of plastic waste against mismanaged plastic waste

We note that a country, Moldova, appears above the $y = x$ line. This corresponds to having plastic waste of 130.15 kg/day and mismanaged plastic waste of 143 kg/day. As a result we can conclude this is an error, but it is not easily fixable - the relationship between the two variables is unclear, so we elect to remove this datapoint from our dataset.

A bargraph is plotted showing how the number of countries with missing values varies by income status. This is because the income status of a country might be linked to how difficult it is to obtain data on certain statistics.

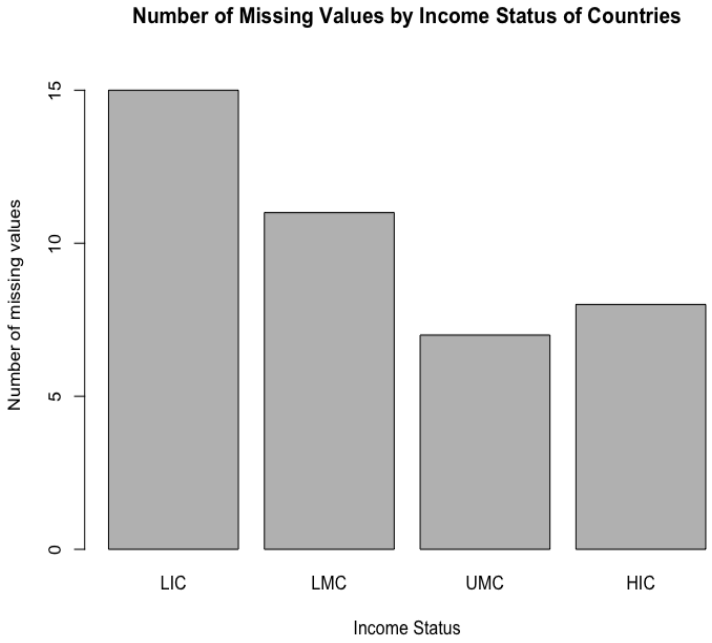
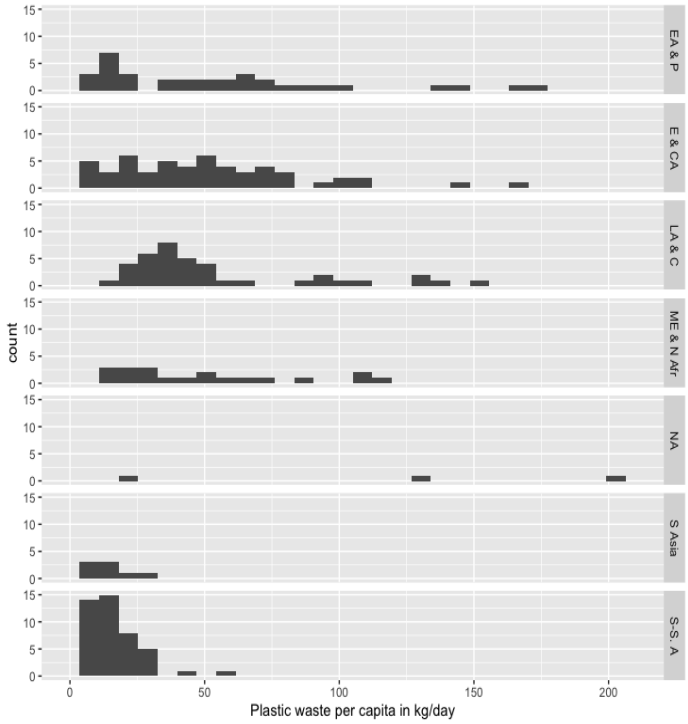


Figure 4: Bar graph of the number of missing values by income status of country

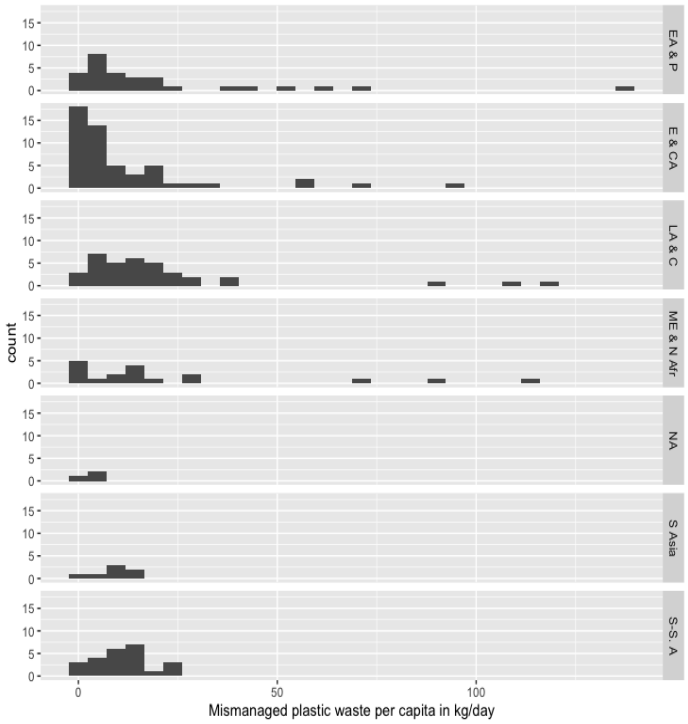
The data show that the number of countries with missing values is highest for the group of low income countries, as expected - then for low-middle income countries, and then high income countries, and last upper-middle income countries. The missing values appear to be mostly for the mismanaged plastic waste statistics, of which there were 41 missing values compared to 2 missing values for the plastic waste statistic. Unsurprisingly, where there were missing values for plastic waste, there were also missing values for mismanaged plastic waste.

2 Question 2

To explore how and whether the distributions of plastic and mismanaged plastic waste were affected by region and income status, multiple histograms were plotted on the same axes. This allows for easy identification of shifts in distribution, or differences in size or shape.



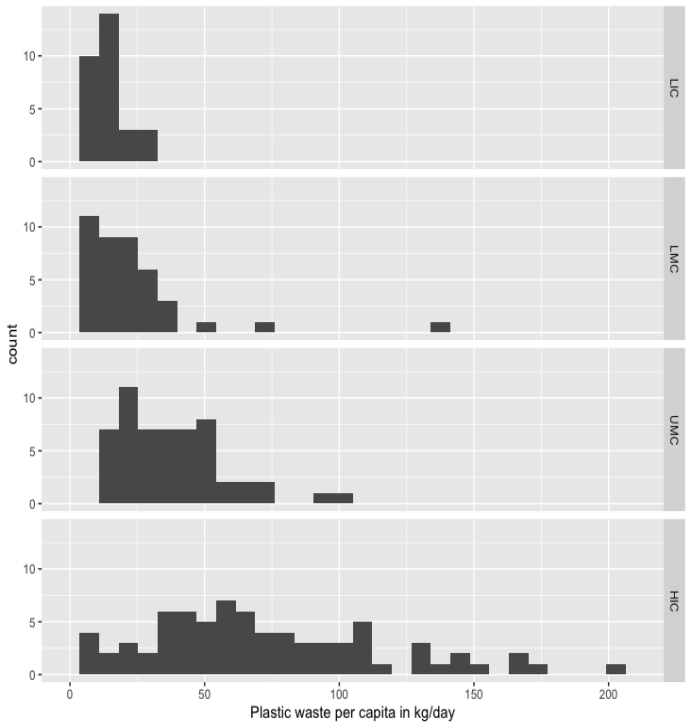
(a) Plastic waste per capita grouped by region



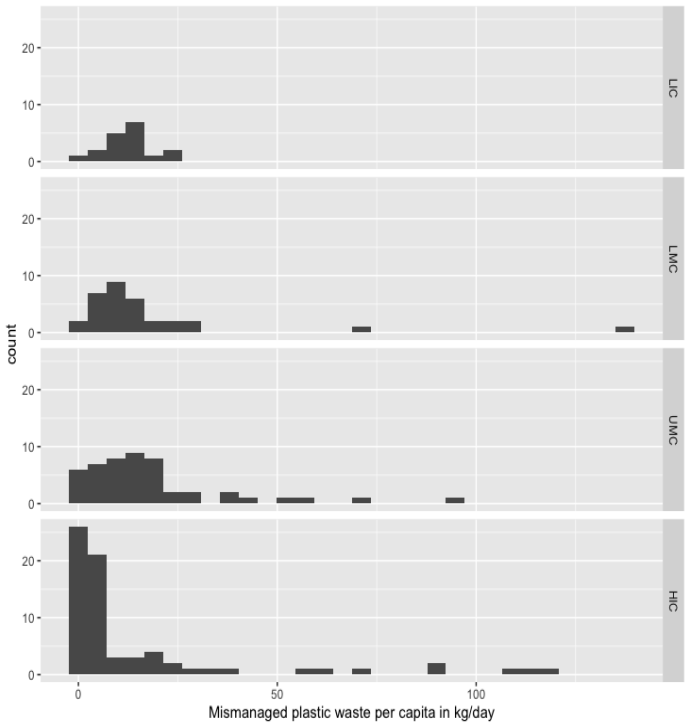
(b) Mismanaged plastic waste per capita grouped by region

We see here that accounting for region does in fact reveal that the distribution of plastic and mismanaged waste is different depending the region of the world. Starting with shape for example, we see that the plastic waste histogram for "S-S. A", corresponding to Sub-Saharan Africa, shows more heaping between values $0 - 50$, whereas the histogram for "E & CA", corresponding to Europe and Central Asia, is more spread out; platykurtic.

Furthermore, some of the means are notably different: for Sub-Saharan Africa, the histogram suggests that the mean is around 15 kg/day, whereas for Europe and Central Asia it is around 50 kg/day. To test if this difference is significant, we perform an independent t-test. The actual mean for Sub-Saharan Africa is 15.193, and for Europe and Central Asia it is 52.23. When we perform a t-test at the 95% confidence level, we get a p-value of $8.17 * 10^{-10}$, suggesting that there is a statistically significant difference in the means between the two regions. However, this does not confirm that region is the cause for this difference. We look at the income status of the country which is likely to be a better explanation.



(c) Plastic waste per capita grouped by income status of country



(d) Mismanaged plastic waste per capita grouped by income status of country

In the plastic waste graph, we notice that the lower income countries (LIC, LMC) and to a lesser extent the upper-middle income countries have distributions with means closer to 0, and are slightly right skewed, compared to the higher income countries. The histogram for the high income countries suggests a larger mean and is also more spread out, which would support the idea that higher income countries would have higher plastic usage and thus waste. The mismanaged plastic waste graph suggests the opposite for the amount of mismanaged plastic waste: the high income countries have a lower mismanaged plastic waste per capita compared to the lower income countries, which can be seen by the heaping at $0 - 20$ kg/day. This could be explained by higher income countries having the means to waste collection and proper waste disposal.

3 Question 3

The relationship between plastic waste and mismanaged plastic waste was plotted on a scattergraph, which allows for easy identification of the shape and sizes of possible trends. Furthermore, the income status of each country on the graph is marked by colour.

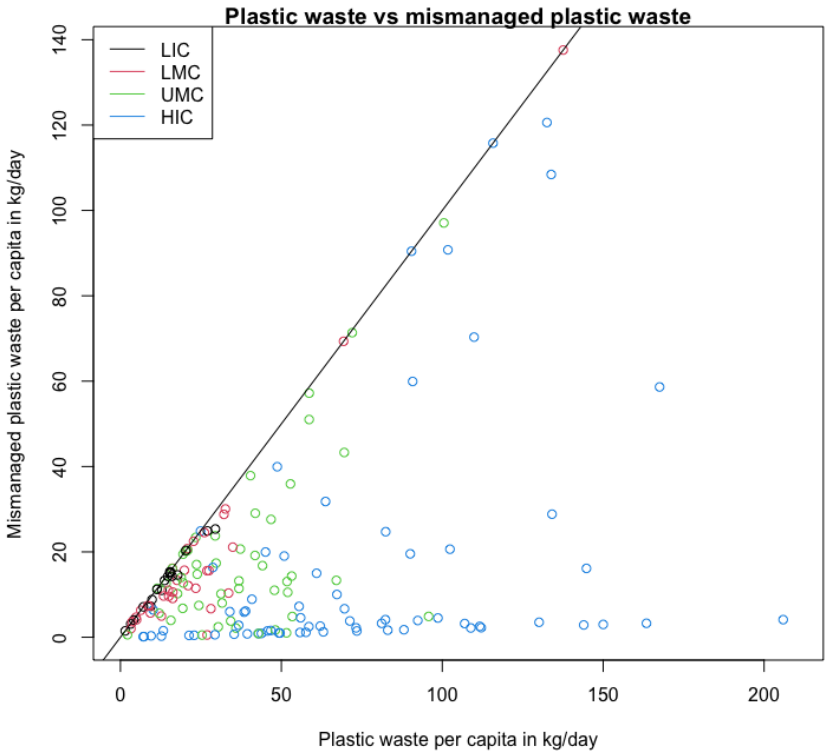


Figure 5

The scatterplot suggests that there are differences in the relationship between plastic waste and mismanaged plastic waste between different income status groups. For example, note how a lot of the points lying along the bottom of the graph (which corresponds to low mismanaged plastic waste) correspond to countries classed as high income or upper middle income. Then, as you rotate anticlockwise from the x to the $y = x$ line, the distribution is more balanced, including more lower-middle and low income countries. This suggests that higher income countries are better at managing their plastic waste.

We use a lattice plot to prevent overlap of points from different groups while allowing us to see the points plotted on the same scales.

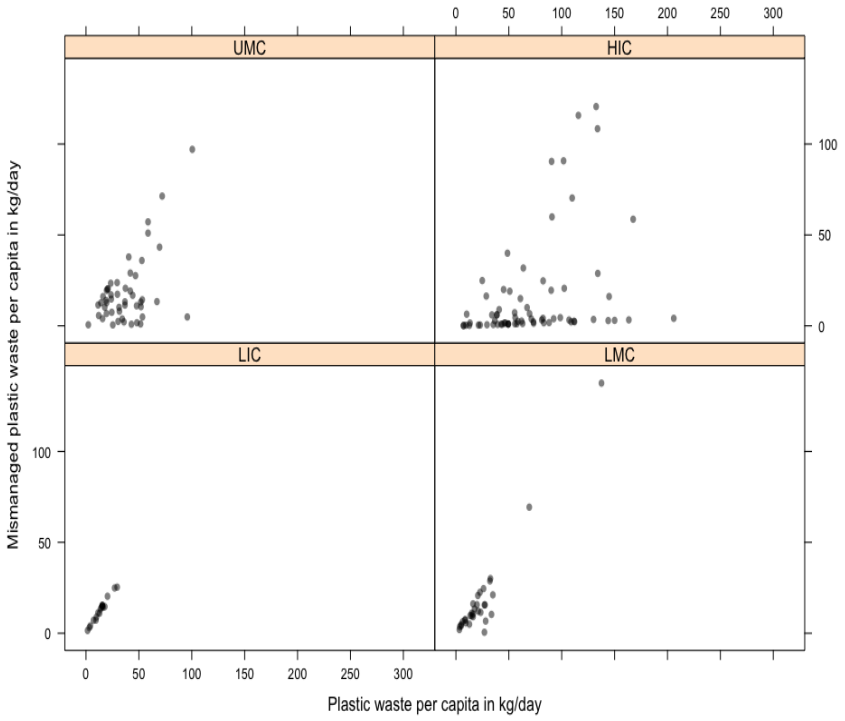


Figure 6: Lattice plot for plastic waste against mismanaged plastic waste grouped by income status of country

The slight transparency on points allows us to see where most of the points lie and how this might indicate the relationship. For example, for HIC, the darker row of points along the bottom of the plot indicate that the strength of the relationship between plastic waste and mismanaged plastic waste might be weaker than in low income countries, where the darker row of lines is in a diagonal. We can test the strength of a linear association using the correlation coefficient. The results are tabulated below:

Income Status	Correlation Coefficient
LIC	0.9893903
LMC	0.9739243
UMC	0.5484522
HIC	0.3743306

As we can see, the higher the income status, the weaker the linear relationship between the plastic and mismanaged plastic waste.

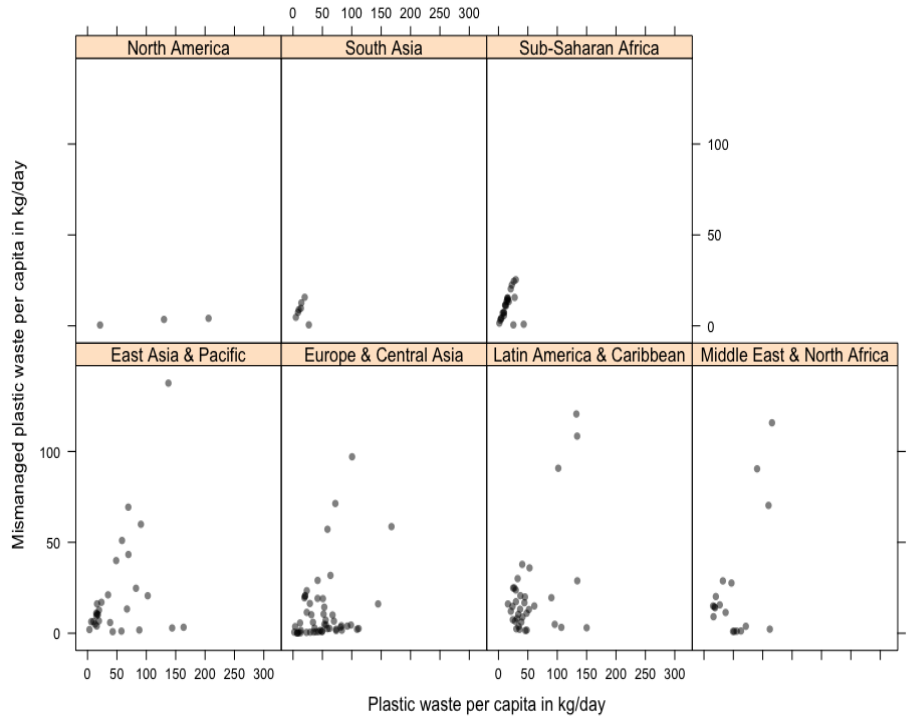


Figure 7: Lattice plot for plastic waste against mismanaged plastic waste grouped by region

We display a chart of the same type for how the relationship varies by region. The differences are not as clear as the one based on income status, but we can still see trends spurred by the income status of the relevant countries. For example, most countries in Sub-Saharan Africa are classed as low or low-middle income, and so the graph resembles those graphs as seen in Figure 6.

4 Question 4

To determine whether there is an association between plastic waste and other variables, a parallel coordinate plot is drawn to visualise how the amount of plastic waste behaves with the other non-categorical variables: GDP, Population, Coastal Population and Urban Population. It is also coloured by income status of the country, which we have noted as an influential variable.

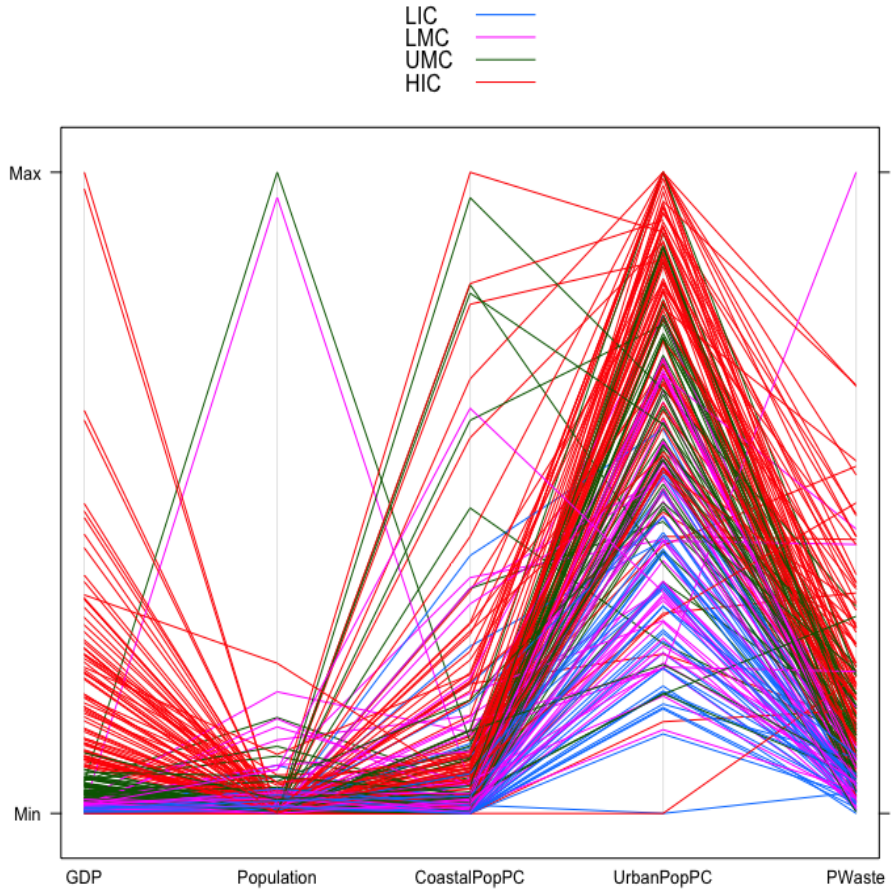


Figure 8: Parallel coordinate plot of five variables: GDP, Population, Coastal population %, Urban population %, and Amount of plastic waste per capita

From the parallel coordinate plot, we note that there is bunching of observations by income status (coloured groups), and that it seems like places with lower urban populations have lower values for plastic waste. We can test for the strength of this association using a correlation plot.

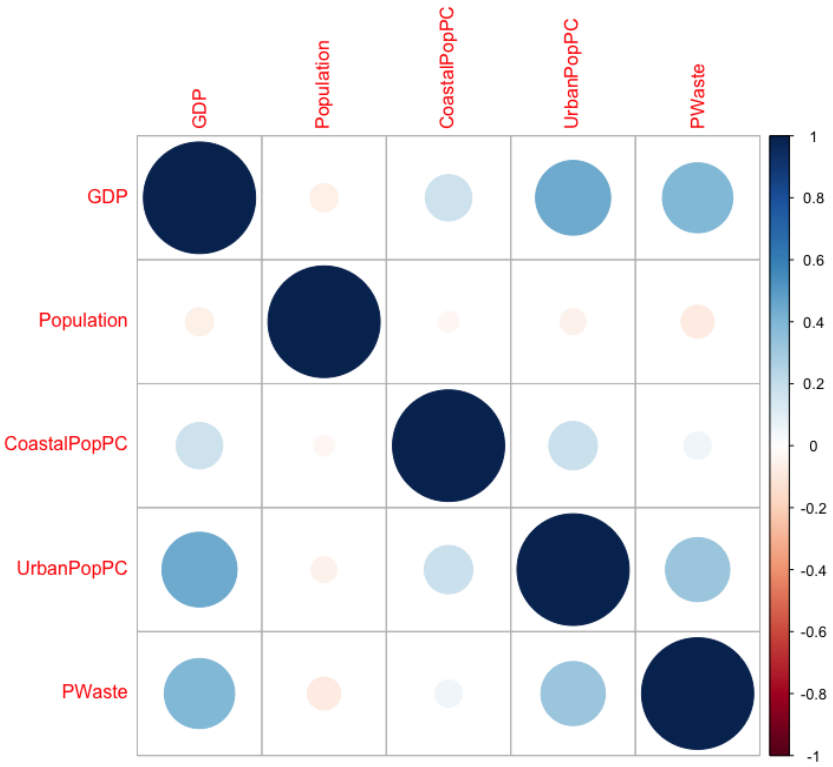


Figure 9: Visualisation of correlation matrix for GDP, Population, Coastal population %, Urban population %, and Amount of plastic waste

It appears that the plastic waste is quite correlated with the urban population as suspected, but also with the GDP of the country. We investigate these relationships by plotting the relevant scatter graphs and smoothing to more easily see a trend.

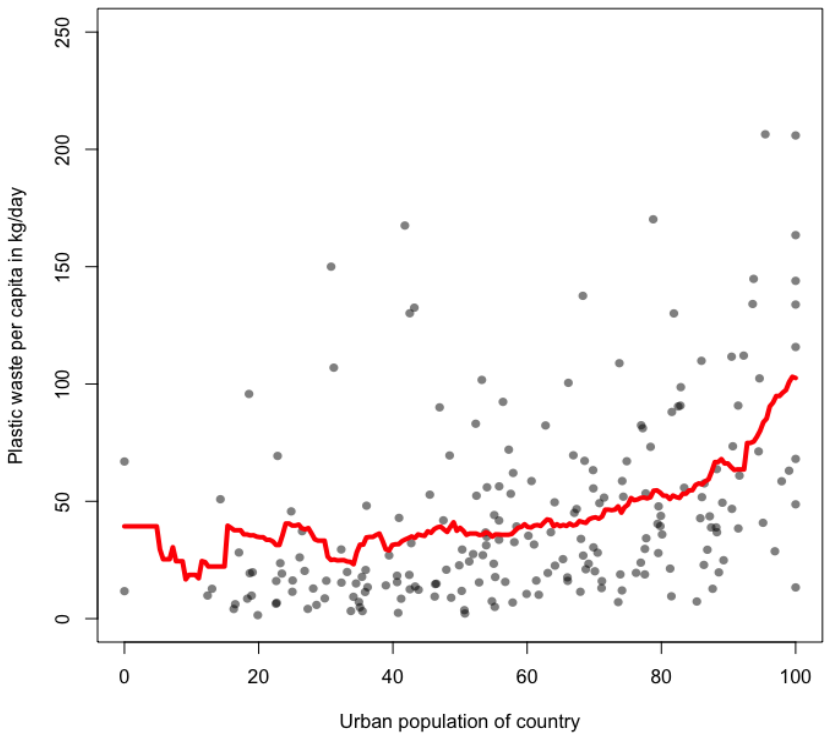


Figure 10: Scatter graph of urban population against amount of plastic waste

In this scatter graph we see a clear positive association between the urban population of the country and the amount of plastic waste produced. While we state that the linear correlation coefficient is 0.32902232, we note that a correlation coefficient may not be hugely useful here as the trend looks vaguely non-linear - and might take on a quadratic relationship instead. Regardless, the urban population might be a useful variable to use in modelling the plastic waste (along with income status, which is correlated with region and GDP).

5 Question 5

Two scatter graphs are plotted to visualise any association between the pairs of variables, and smoothing is performed to make any trends easier to see.

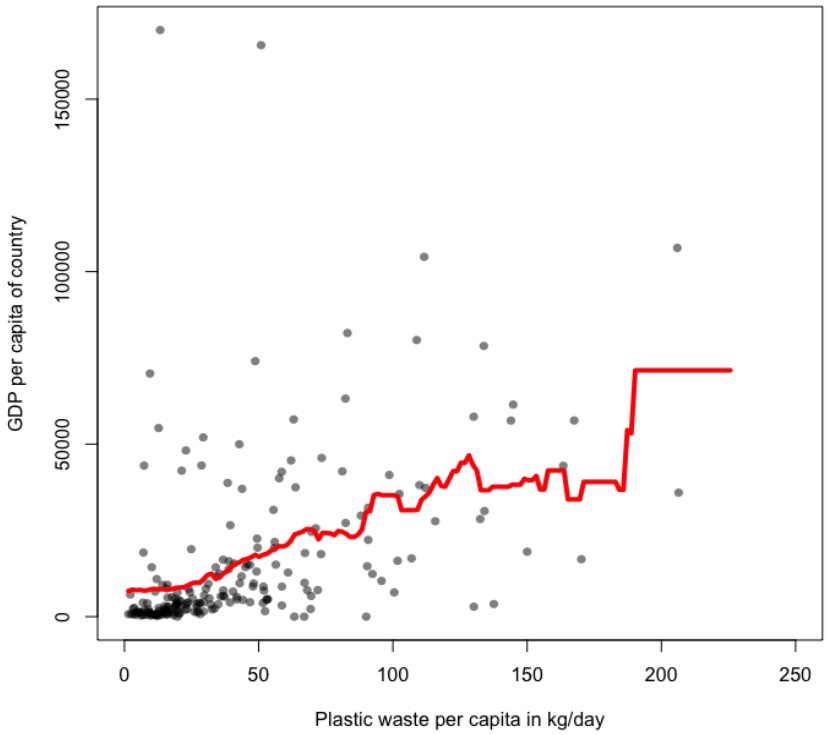


Figure 11: Scatter graph of GDP against amount of plastic waste

The graph shows a clear positive association between the GDP and amount of plastic waste produced by the country. To confirm this association, we find the correlation coefficient to be 0.393308, which is a moderate strength correlation. This relationship would make sense, as GDP is likely highly correlated with income status of the country, which we already strongly suspect to be linked to the amount of plastic waste. Therefore, there may be some redundancy in a model where the amount of plastic waste is predicted by the values of both GDP and its income status.

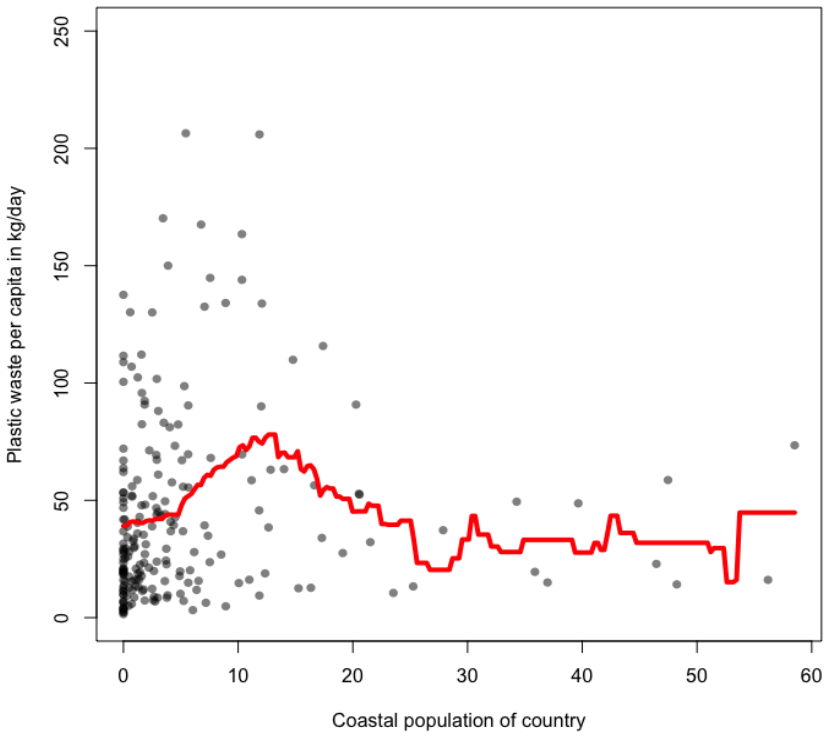


Figure 12: Scatter graph of coastal population against amount of plastic waste

This graph appears less conclusive than the graphs plotting the plastic waste against either of GDP or urban population. There is a small increase in the smoothed trend initially, but overall the smoothed trend is roughly horizontal, indicating that there is not much effect of an increasing coastal population on the amount of plastic waste. Furthermore, the correlation coefficient is calculated as 0.05806629, which is notably close to 0. This would signify that there is a very weak positive association, so we conclude that there is no significant association at all. It appears that instead, the urban population is more useful for predicting the amount of plastic waste produced by a country.