Data Visualisation Assignment 1

YANSHEN WANG

March 2023

1 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.

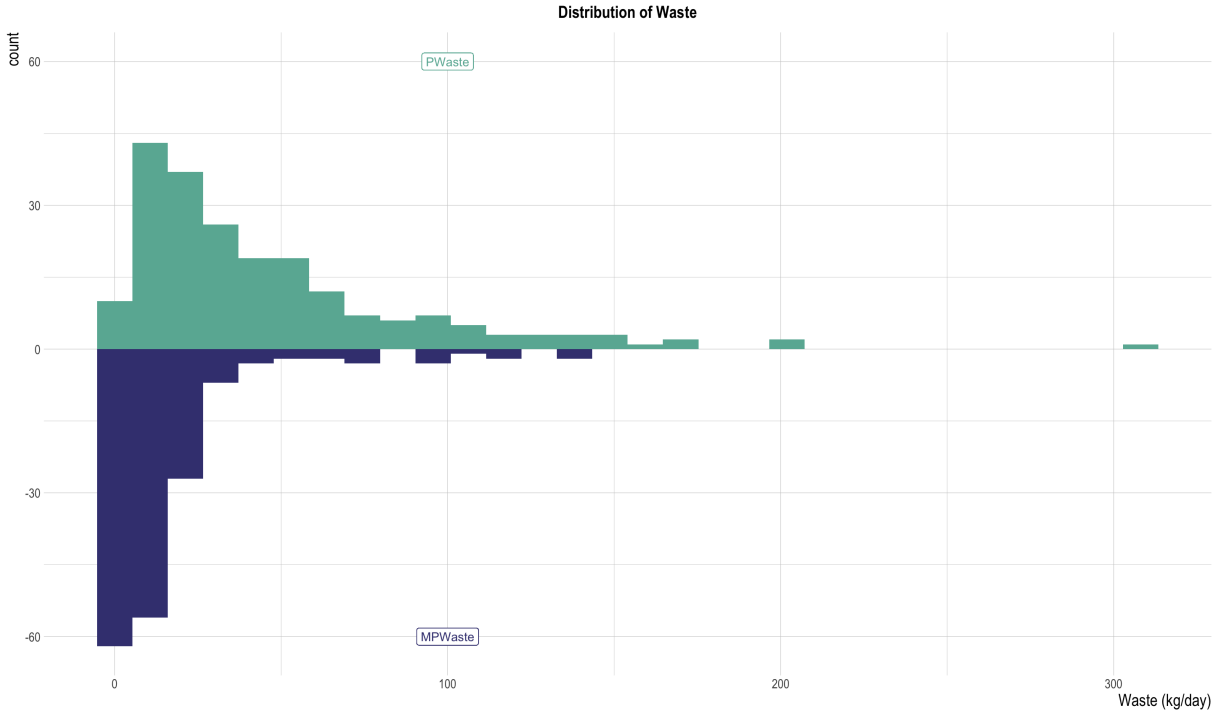

Figure 1: Distribution of Waste

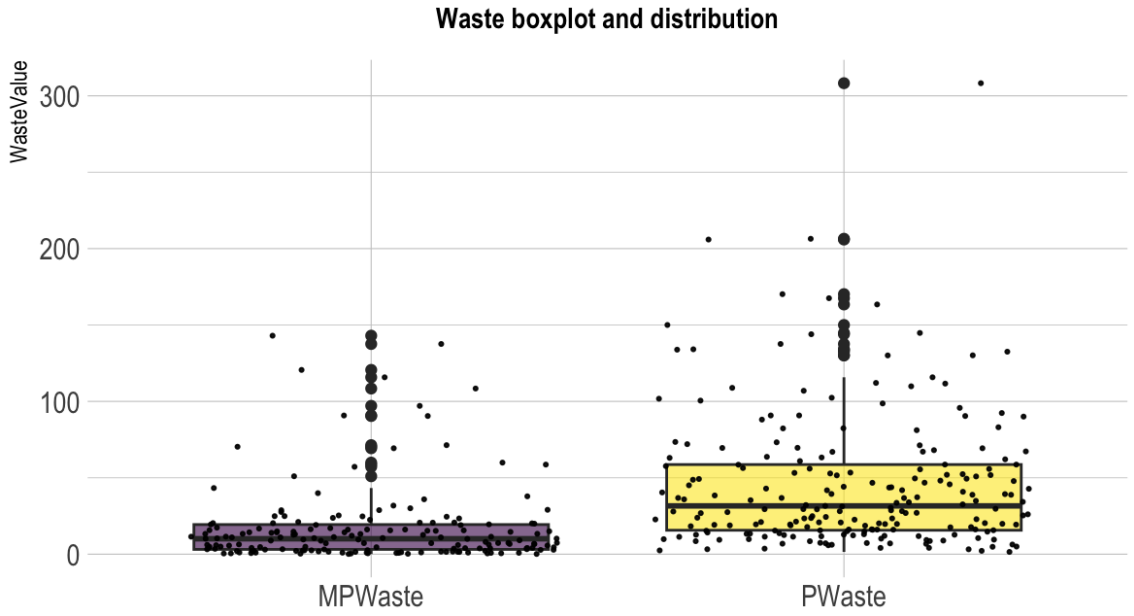

Figure 2: Boxplot with Distribution

Figure 3: Missing Values

According to Figure 1, we can see that both Plastic Waste and Mismanaged Plastic Waste variables showed a left-skewed distribution, with the majority of the values distributed between 0 and 50. The distribution of Mismanaged Plastic Waste was between 0 and 150, with peaks in the range of 0 to 10, and then decreasing until 60 to 70, with less distribution above 70 and irregular. The distribution interval of Plastic Waste was between 0 and 310, and its peak appeared in the interval of 10 to 20, and then decreased until the interval of 150 to 160, the count in the interval of 90 to 100 was slightly higher than the interval on both sides, and the distribution of the part above 160 was less and irregular, while the value above 300 was a very obvious outlier.

From Figure 2, we can see that most of the data of Mismanaged Plastic Waste were distributed below 50, the values of minimum (Q1 -1.5*IQR) and first quartile was slightly above 0, median was about slightly above 10, third quartile was about 25, maximum (Q3 + 1.5*IQR) was slightly below 50, while there were potential outliers above 50; while Plastic Waste was mainly distributed in the interval of 0 to 120, minimum (Q1 -1.5*IQR) was slightly above 0, first quartile's value was about 25, Median was about 35, third quartile was about 55, maximum (Q3 + 1.5*IQR) was about 120, and there are potential outliers above 120, of which the value above 300 was a very obvious outlier. And Figure 3 shows that Plastic Waste had two missing values, accounting for 1% of plastic waste; while Mismanaged Plastic Waste had more missing values, accounting for 19%. After data ranking, the missing values were mainly concentrated in some developing countries or war-torn areas, and it was difficult to use effective methods to fill the data, so we removed the missing values in this assignment, so the analysis may be somewhat different from the actual situation.

A histogram is an effective way to show the distribution of numerical variables, and the variable is cut into several bins, and the number of observation per bin is represented by the height of the bar. A boxplot is a standardized way of displaying the distribution of data based on a five number summary and it is also a way to quickly locate outliers.

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

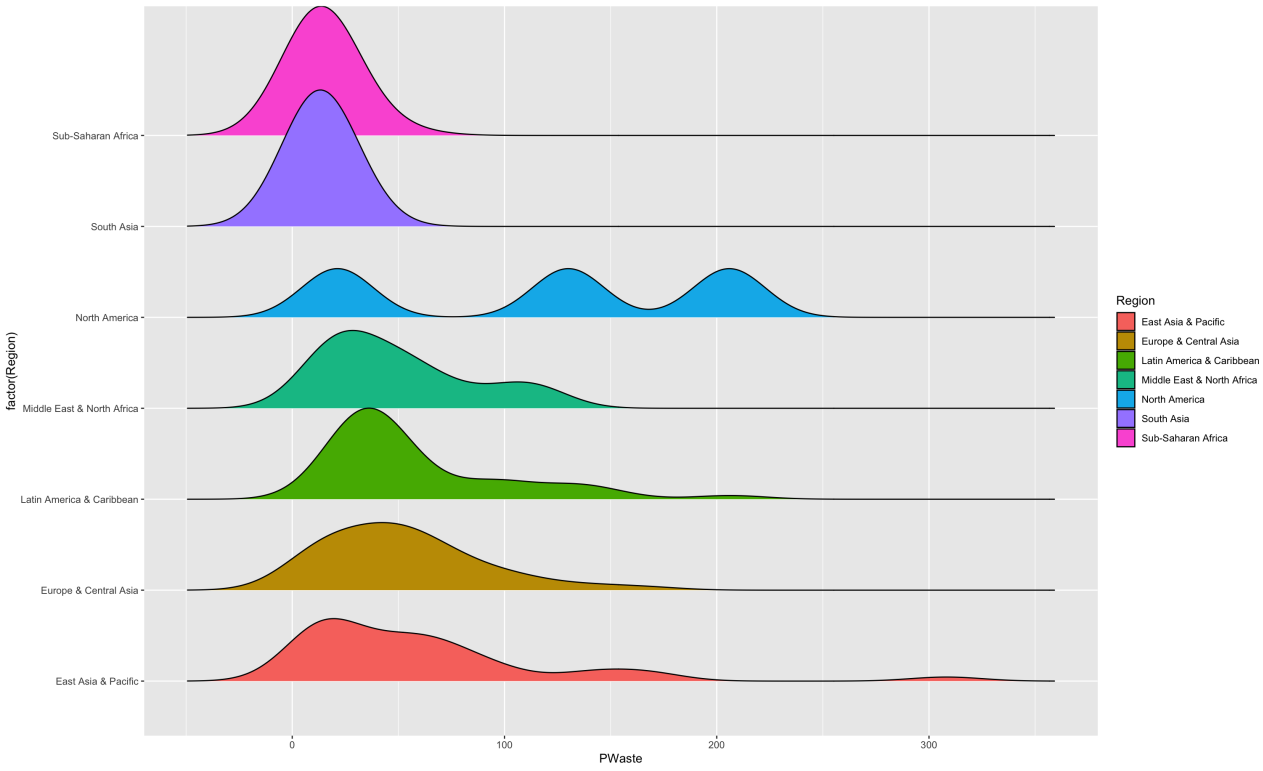

Figure 4: Plastic Waste by Region

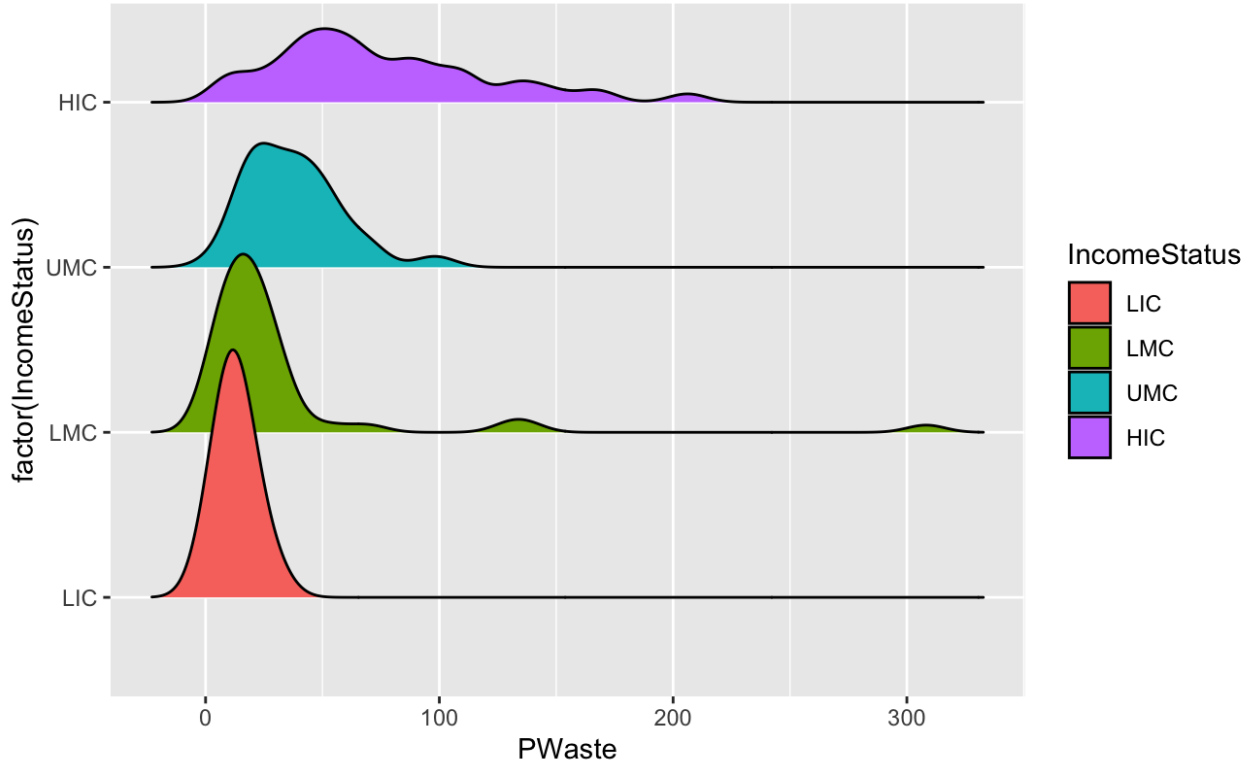

Figure 5: Plastic Waste by Income Status

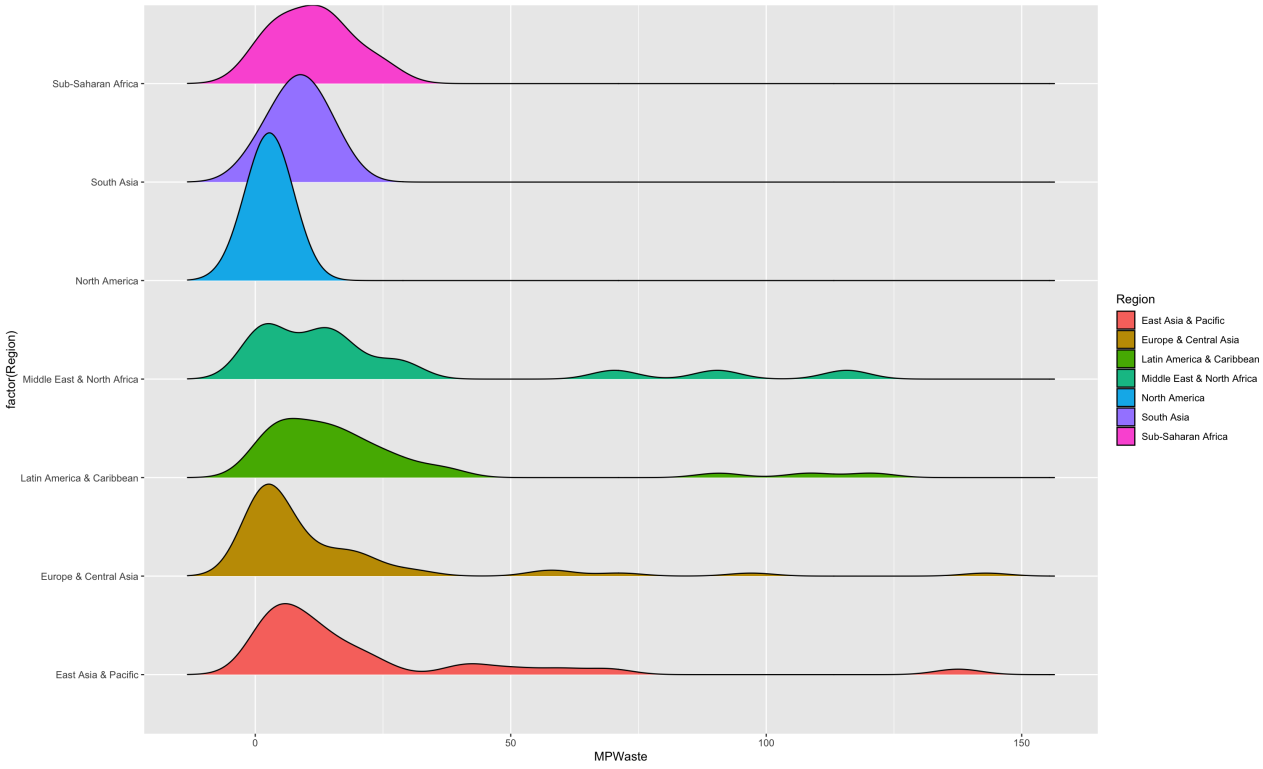

Figure 6: Mismanaged Plastic Waste by Region

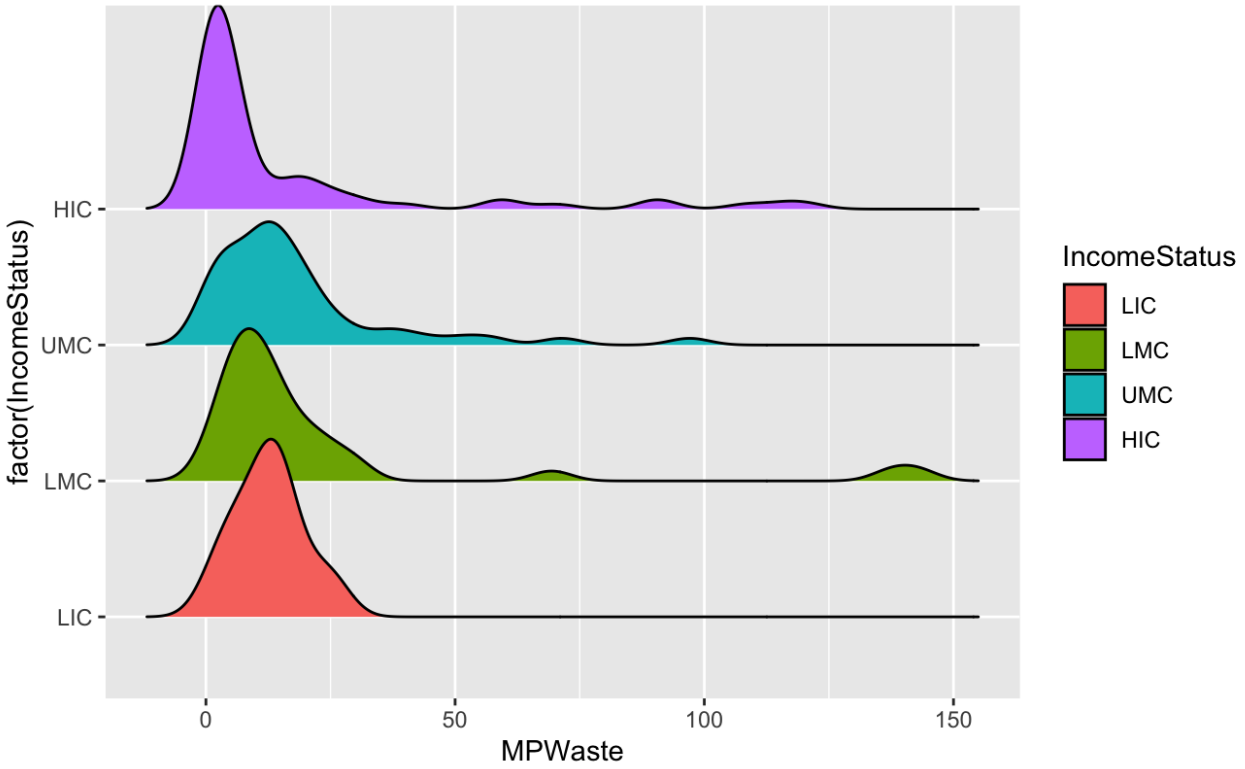

Figure 7: Mismanaged Plastic Waste by Income Status

Figure 4 showed the distribution of Plastic Waste in different regions, where Sub-Saharan Africa and South Asia had a similar distribution with values concentrated below 50; Middle East & North Africa, Latin America & Caribbean, Europe & Central Asia and East Asia & Pacific had a similar distribution pattern, with a left-skewed peak distribution in the interval of 0 to 200, and a larger amount of Plastic Waste compared to Sub-Saharan Africa and South Asia; while North America had a completely different distribution pattern from the other regions. It showed a multi-peaked distribution pattern in the intervals of 0 to 50, 100 to 175, and 175 to 220, and the amount of Plastic Waste was clearly higher here than in other regions. Figure 5 showed the effect of Income Status on the distribution of Plastic Waste, where the distribution of LIC and LMC groups was more similar, concentrated in the interval from 0 to 50, and LMC has some potential outliers, while the distribution of UMC was in the interval from 0 to 100, and its peak was larger than that of LIC and LMC groups; the distribution interval of HIC was wider, distributed in the interval from 0 to 200, and its peak was obviously higher than the other three groups. Obviously, the number of Plastic Waste was increased with the increase of Income Status.

Figure 6 showed the distribution of Mismanaged Plastic Waste in different regions, Saharan Africa, South Asia and North America had similar distribution patterns, mainly in the interval of 0 to 30, with the peak in North America in a smaller interval than the other two groups, and the distribution interval was concentrated in 0 to 25. The other four regions have a long-tailed distribution with left-skewed peaks, and although they were mainly distributed in the 0 to 50 interval, there were a few potential outliers in all four regions, which were relatively large. It was important to note that Middle East & North Africa show a multi-peaked distribution. Figure 7 showed the effect of Income Status on the distribution of Mismanaged Plastic Waste. The distribution of Mismanaged Plastic Waste was relatively similar for the four income statuses, mainly between 0 and 50, but HIC, UMC and LMC all had a small distribution of data on the 50 to 150 interval, while LIC had none; overall the number of Mismanaged Plastic Waste in the HIC group was relatively small, the number of Mismanaged Plastic Waste in the LIC group was the largest, and the number of UMC and LMC was in the middle, but the difference between these four groups was not significant.

A ridgeline plot shows the distribution of a numeric value for several groups. Distribution can be represented using density plots, all aligned to the same horizontal scale and therefore can effectively compare differences in the distribution of variables in different groups.

3 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?

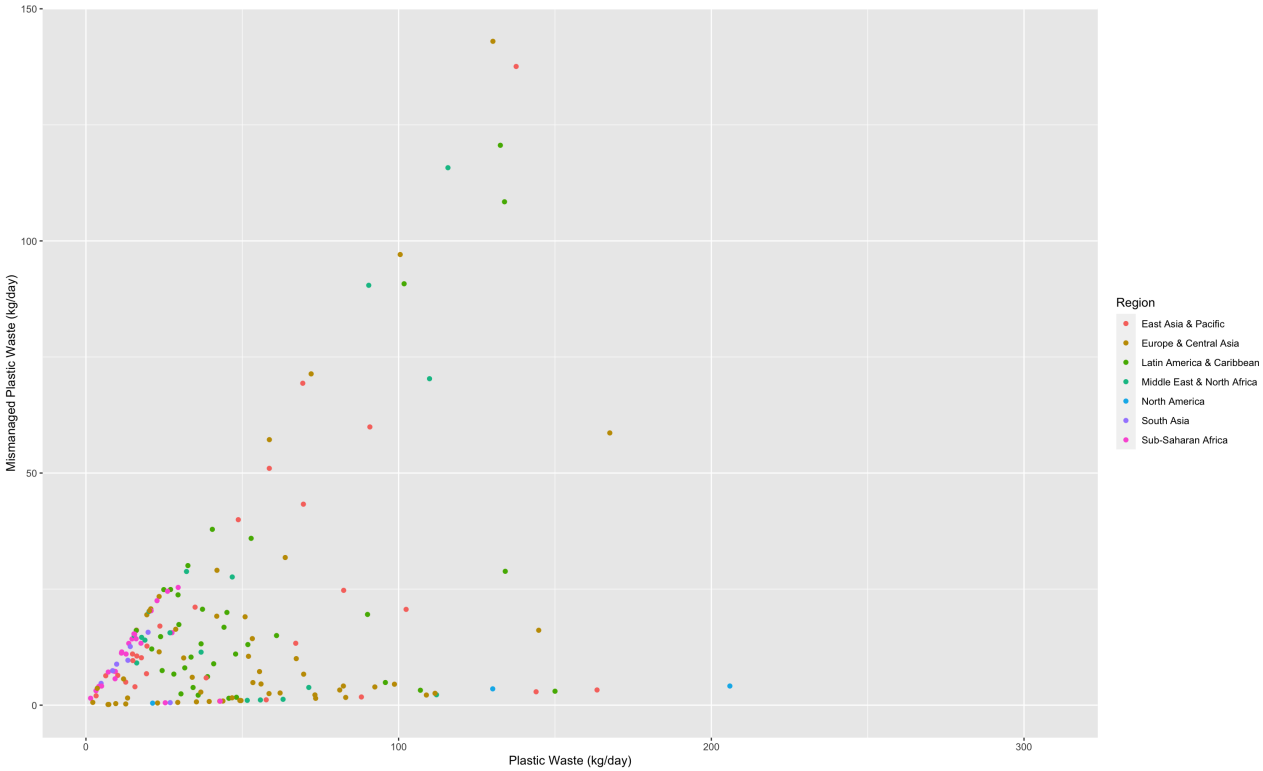

Figure 8: Scatter of two types of plastic waste by Region

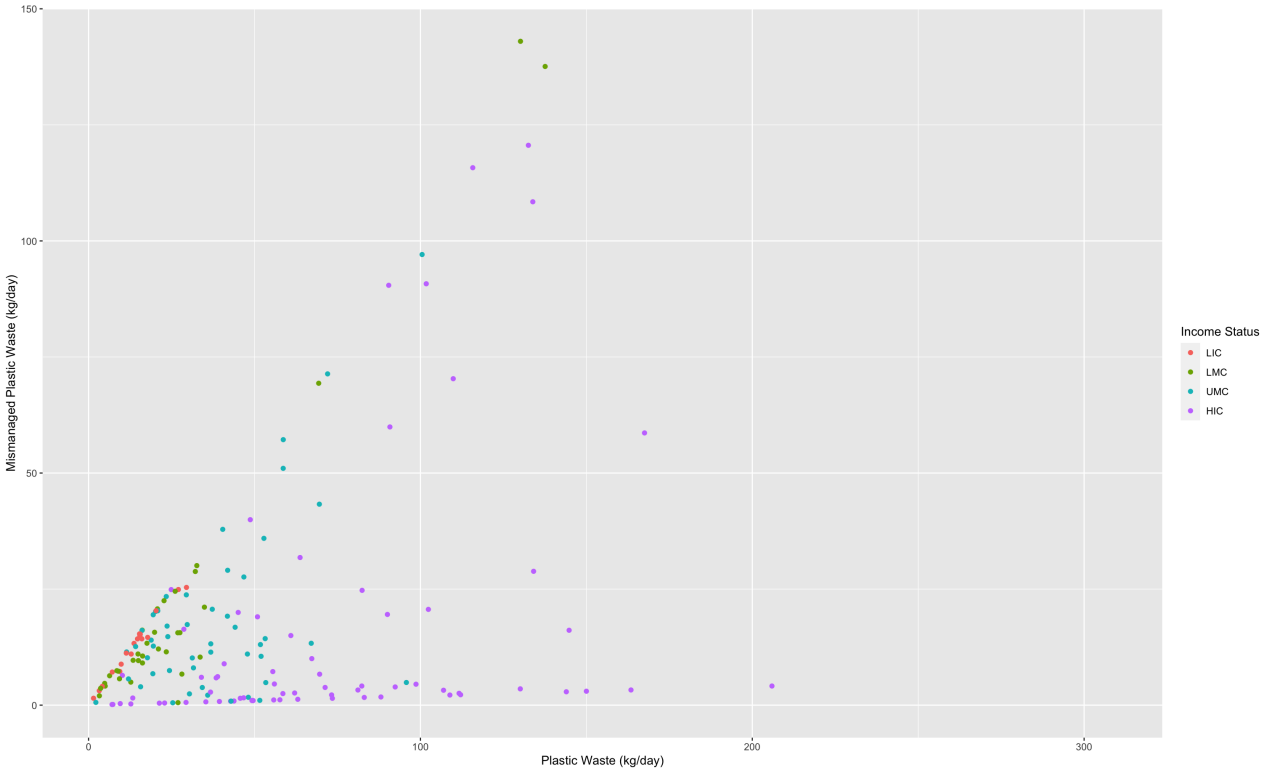

Figure 9: Scatter of two types of plastic waste by Status

The scatter plot in Figure 8 showed that there was some positive trend of correlation between Mismanaged Plastic Waste and Plastic Waste, but the correlation was not strong. The overall trend was that the more Plastic Waste there was, the greater the amount of Mismanaged Plastic Waste was, but the overall trend was not significant.

The relationship between Plastic Waste and Mismanaged Plastic Waste varied among regions, with a strong correlation between Plastic Waste and Mismanaged Plastic Waste in Sub-Saharan Africa, and a strong correlation between Plastic Waste and Mismanaged Plastic Waste in Middle East & North Africa, and Middle East & North Africa, there may also be a strong correlation between Plastic Waste and Mismanaged Plastic Waste. In other regions, the correlation between the two variables was not high. The relationship between Plastic Waste and Mismanaged Plastic Waste in different Income Status was also different, the point distribution was very scattered in HIC and UMC status, which had almost no correlation, while strong correlation exists in both LIC and LMC status.

The scatter plot is a good way to explore the correlation between two variables, and it can effectively observe the trend of the two variables. Because a third variable is involved, it is better to make separate plots according to the category of the third categorical variable, but this question received the limitation of the number of pictures, so this question distinguishes the categorical variable as a color fill. But the distinction is not as good as making separate graphs.

4 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?

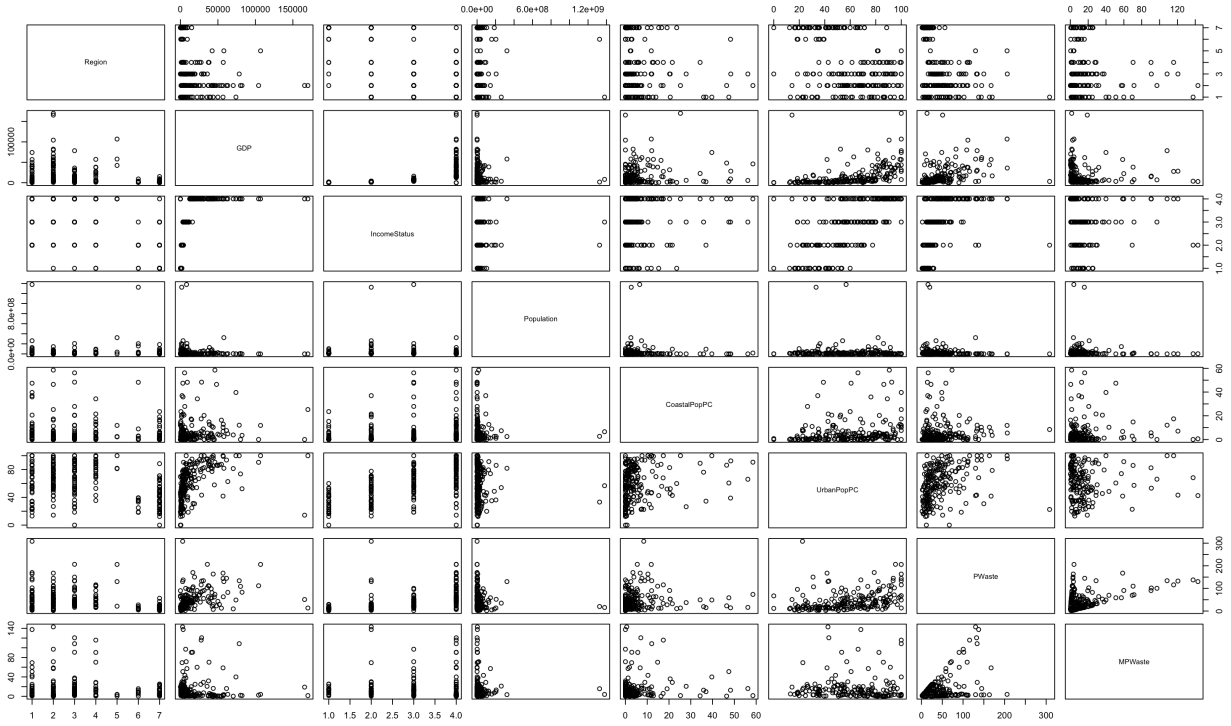

Figure 10: Associations between each variables

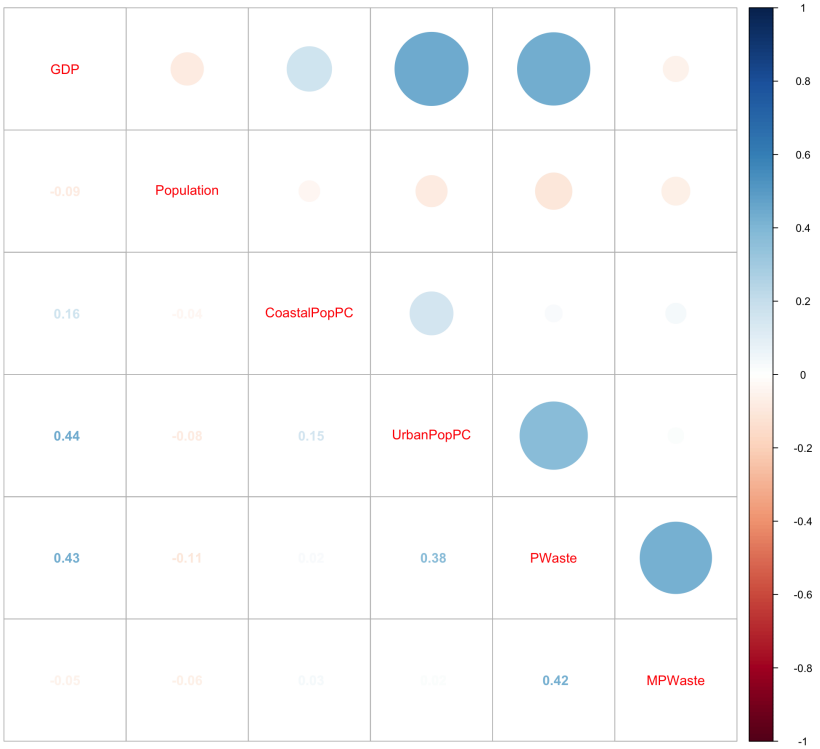

Figure 11: Correlation between each variables

From Figure 10, it can be found that there may be some correlation between PWaste and GDP, but the correlation was not strong; while there may be a non-linear correlation between PWaste and CoastalPopPC and UrbanPopPC. According to the scatter plot distribution, it was possible that this correlation was logarithmic correlation, where PWaste and CoastalPopPC may be negatively logarithmic correlated, PWaste and UrbanPopPC may have positive logarithmic correlation, and there may be no correlation with Population; MPWaste and GDP, and CoastalPopPC may be negatively logarithmic correlated, and it may not be correlated with Population and UrbanPopPC. Therefore, when doing the model, it may need to be noted that the linear model cannot be simply used, and the generalized additive model may be a better choice.

Figure 11 used linear correlation to calculate that there was a weak correlation between PWaste and GDP with a correlation coefficient of 0.43, and a linear correlation coefficient of 0.38 between PWaste and UrbanPopPC, both of which need further testing. And MPWaste had no correlation with each of the other variables.

The scatter plot matrix can quickly show the relationship between variables in the dataset, because we need to show the relationship between multiple variables at the same time, so the scatter plot matrix is a good method, and the correlation coefficient matrix is a better complement to the scatter plot matrix, which can calculate the correlation coefficient between variables more clearly. Using these two plots simultaneously can prioritize the presentation of the relationships between variables.

5 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.

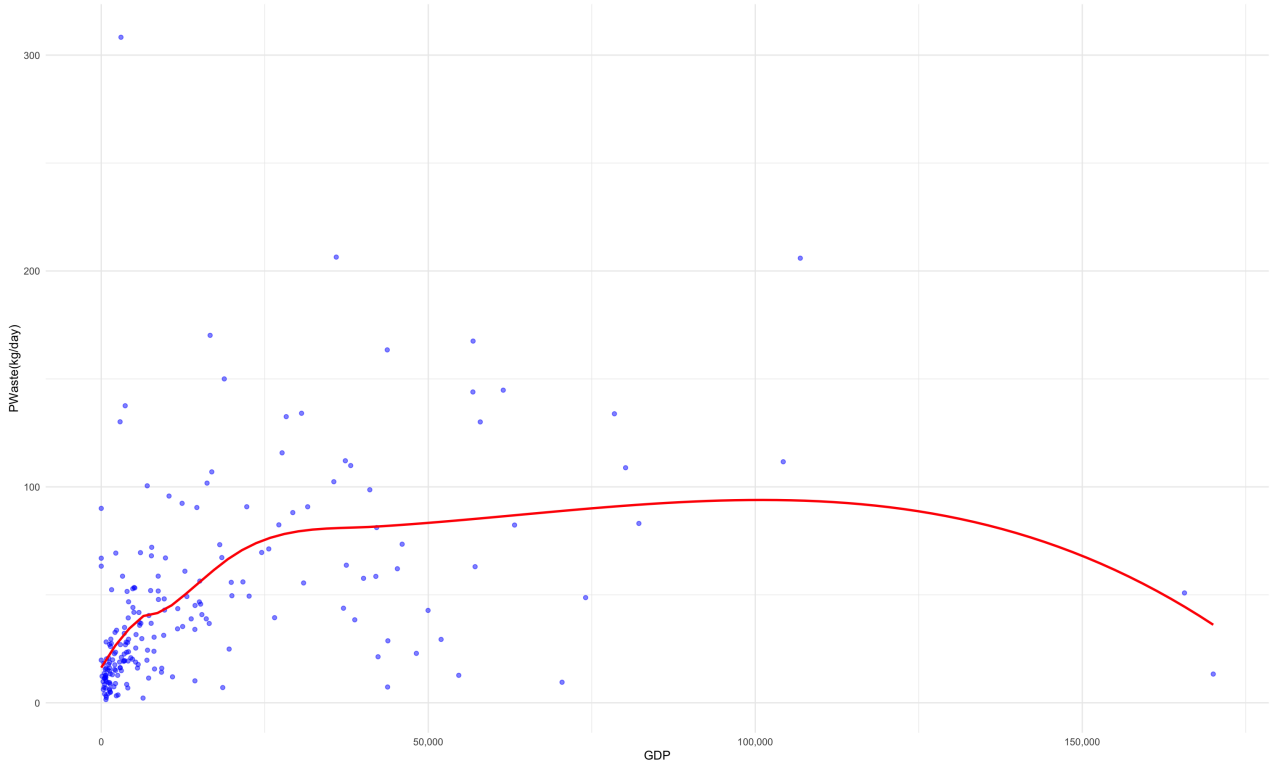

Figure 12: Relationship between plastic waste and GDP

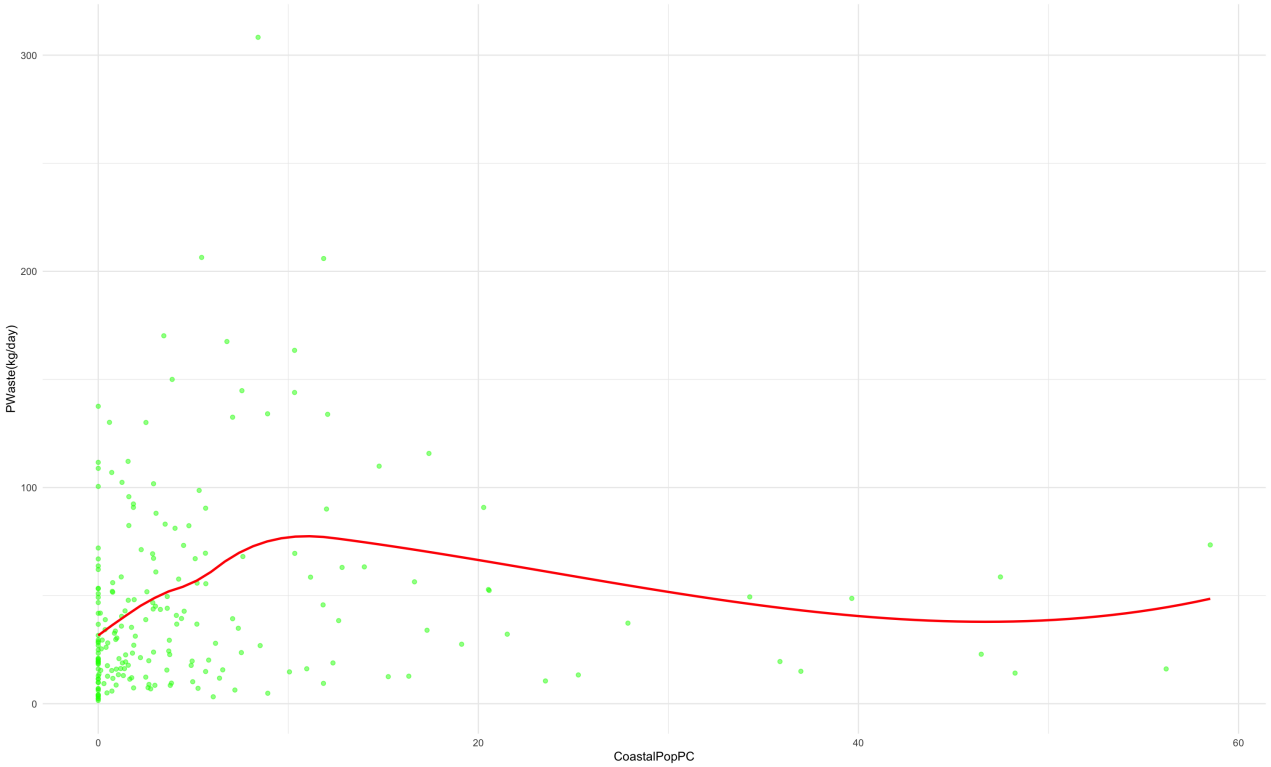

Figure 13: Relationship between plastic waste and coastal population

Figure 12 showed a scatter plot of the relationship between PWaste and GDP with the addition of a smoothed trend line. Similar to the findings in Question 4, there may be a weak correlation between PWaste and GDP, and the trend line also verified that this correlation was not linear. There may be a strong positive correlation between PWaste and GDP when GDP was in the range of 0 to 25000, while in the range of 25000 to 100000, there was almost no correlation between PWaste and GDP. The correlation may be negative or non-linear for intervals greater than 100000, and the values in this interval may be potential outliers and need to be further explored. With the results shown by the trend line, the generalized additive model may be a better choice.

Figure 13 showed a scatter plot of the relationship between PWaste and CoastalPopPC with a smoothed trend line. Overall, there was no significant correlation between PWaste and CoastalPopPC, but in the interval from 0 to 10, they had a strong positive correlation; while in the interval from 10 to 40, PWaste decreased slowly with the increase of CoastalPopPC and may have a weak negative correlation; in the interval from 40 to 60, PWaste slowly increased slowly with the increase of CoastalPopPC, and there may be a weaker positive correlation. However, since the data volume is very small in the interval range of 20 to 60, this weaker correlation may also be a pseudo-correlation and may be the result of outliers. Further exploration was still needed, and again the generalized additive model could be an option in the modeling process.

A better way to explore the relationship between two variables is a scatter plot, from which the relationship between two variables in the overall range can be better seen, but it is difficult to effectively show the relationship between two variables in the local scope. Therefore, adding a trend line can effectively show the overall trend of the variables, and at the same time, the trend in the local scope can also be effectively shown. Adding a trend line can effectively explore the hidden trend that cannot be shown in the scatter plot.