Assignment 1

Graded

Student

Tingting Su

Total Points

68 / 100 pts

Question 1

The distributions of the two types of plastic waste

12 / 20 pts

Visualisations: Choice

+ 4 pts All visualisations appropriate to the goals.

✔ + 3 pts Visualisations generally appropriate to the goals.

+ 2 pts Visualisations only partly address the goals; more effective choices are not used; use of visualisations not covered in lectures

+ 1 pt Inappropriate choice of visualisation.

+ 0 pts No attempt; visualisations fails to address any of the key objectives.

Visualisations: Presentation

+ 4 pts All visualisations clearly presented and of presentation quality.

✔ + 3 pts Generally clearly presented and of good quality; some minor shortcomings (e.g. aspect ratio distorting the plot, poor choice of histogram bar width, extraneous use of colour)

+ 2 pts Some deficiencies in presentation of multiple plots (e.g. scale/aspect ratio/quality renders plot difficult to read, comparison boxplots/histograms with different axis ranges, default labels/captions).

+ 1 pt Major/careless weaknesses or omissions, e.g. many comparison boxplots/histograms using different axis ranges, unformatted R code, default illegible labels.

+ 0 pts No attempt; inadequate presentation to address any of the key objectives.

Describe & Interpret: Description of Data Features

+ 4 pts All relevant data features identified; statistical relevance of the findings made clear.

+ 3 pts Several relevant data features identified; statistical relevance to the problem occasionally clear.

✔ + 2 pts Primarily descriptive identification of features; statistical relevance to the problem not made clear.

+ 1 pt Few relevant data features identified; exclusively descriptive; minimal reference to statistical relevance

+ 0 pts No attempt; no salient features identified or discussed.

Describe & Interpret: Interpretation of findings

+ 4 pts Sound interpretation of the data features in context and in terms of exploration goals

+ 3 pts Adequate interpretation of the data features in context and in terms of exploration goals, with some omissions or spurious commentary

✔ + 2 pts Some interpretation of the data features in context and in terms of exploration goals

+ 1 pt Weak or minimal interpretation of the data features found in context and in terms of the exploration goals

+ 0 pts No interpretation or explanation of the data features found given, either in data context or in terms of the exploration goals

---

Modelling implications

+ 4 pts Salient and specific features of the data relevant to modelling identified; implications for future analysis clearly considered and stated

+ 3 pts Some specific features of the data relevant to modelling identified; implications for future analysis partly considered

✔ **+ 2 pts** Obvious or generic implications identified; weak connection to future modelling made.

+ 1 pt Few relevant implications for modelling given; or minimal justification and spurious commentary. Implications or discussion is not focussed on the data science.

+ 0 pts None given, or wholly irrelevant commentary.

3 There is a better plot to show the patterns of missingness

4 You really need to look at histograms to answer this question

Question 2

The potential effects of region and income status

17 / 20 pts

Visualisations: Choice

✔ **+ 4 pts** All visualisations appropriate to the goals.

+ 3 pts Visualisations generally appropriate to the goals.

+ 2 pts Visualisations only partly address the goals; more effective choices are not used; use of visualisations not covered in lectures

+ 1 pt Inappropriate choice of visualisation.

+ 0 pts No attempt; visualisations fails to address any of the key objectives.

Visualisations: Presentation

✔ **+ 4 pts** All visualisations clearly presented and of presentation quality.

+ 3 pts Generally clearly presented and of good quality; some minor shortcomings (e.g. aspect ratio distorting the plot, poor choice of histogram bar width)

+ 2 pts Some deficiencies in presentation of multiple plots (e.g. scale or aspect ratio renders plot difficult to read, comparison boxplots/histograms with different axis ranges, default labels/captions).

+ 1 pt Major/careless weaknesses or omissions, e.g. many comparison boxplots/histograms using different axis ranges, unformatted R code, default illegible labels. Adequate presentation of inappropriate plots

+ 0 pts No attempt; inadequate presentation to address any of the key objectives. Ad

---

Describe & Interpret: Data Features

+ 4 pts All relevant data features identified; statistical relevance of the findings made clear.

✔ **+ 3 pts** Several relevant data features identified; statistical relevance to the problem made generally clear.

+ 2 pts Vague or primarily descriptive identification of features; statistical relevance to the problem not made clear.

+ 1 pt Few relevant data features identified; exclusively descriptive; minimal reference to statistical relevance

+ 0 pts No attempt; no salient features identified or discussed.

Describe & Interpret: Interpretation of findings

+ 4 pts Sound interpretation of the data features in context and in terms of exploration goals

✔ **+ 3 pts** Adequate interpretation of the data features in context and in terms of exploration goals, with some omissions or spurious commentary

+ 2 pts Some interpretation of the data features in context and in terms of exploration goals

+ 1 pt Weak or minimal interpretation of the data features found in context and in terms of the exploration goals

+ 0 pts No interpretation of the data features found in context and in terms of the exploration goals

---

Modelling implications

+ 4 pts Salient and specific features of the data relevant to modelling identified; implications for future analysis clearly considered and stated

✔ **+ 3 pts** Some specific features of the data relevant to modelling identified; implications for future analysis partly considered

+ 2 pts Generic, obvious, or impractical/spurious implications identified; weak connection to future modelling made.

+ 1 pt Few relevant implications for modelling given with minimal justification or spurious commentary. Implications or discussion is not focussed on the data science.

+ 0 pts None given, or wholly generic or irrelevant commentary.

Question 3

The relationship between plastic waste and mismanaged plastic waste

11 / 20 pts

Visualisations: Choice

+ 4 pts All visualisations appropriate to the goals.

✔ **+ 3 pts** Visualisations generally appropriate to the goals.

+ 2 pts Visualisations only partly address the goals; more effective choices are not used; use of visualisations not covered in lectures

+ 1 pt Inappropriate choice of visualisation.

+ 0 pts No attempt; visualisations fails to address any of the key objectives.

Visualisations: Presentation

+ 4 pts All visualisations clearly presented and of presentation quality.

+ 3 pts Generally clearly presented and of good quality; some minor shortcomings (e.g. aspect ratio distorting the plot, poor choice of histogram bar width)

✔ **+ 2 pts** Some deficiencies in presentation of multiple plots (e.g. scale or aspect ratio renders plot difficult to read, comparison boxplots/histograms with different axis ranges, default labels/captions).

+ 1 pt Major/careless weaknesses or omissions, e.g. many comparison boxplots/histograms using different axis ranges, unformatted R code, default illegible labels.

+ 0 pts No attempt; inadequate presentation to address any of the key objectives.

Describe & Interpret: Data Features

+ 4 pts All relevant data features identified; statistical relevance of the findings made clear.

+ 3 pts Several relevant data features identified; statistical relevance to the problem made generally clear.

✔ **+ 2 pts** Primarily descriptive identification of features; statistical relevance to the problem not made clear.

+ 1 pt Few relevant data features identified; exclusively descriptive; minimal reference to statistical relevance

+ 0 pts No attempt; no salient features identified or discussed.

Describe & Interpret: Interpretation of findings

+ 4 pts Sound interpretation of the data features in context and in terms of exploration goals

+ 3 pts Adequate interpretation of the data features in context and in terms of exploration goals, with some omissions or spurious commentary

✔ **+ 2 pts** Some interpretation of the data features in context and in terms of exploration goals

+ 1 pt Weak or minimal interpretation of the data features found in context and in terms of the exploration goals

+ 0 pts No interpretation of the data features found in context and in terms of the exploration goals

---

Modelling implications

+ 4 pts Salient and specific features of the data relevant to modelling identified; implications for future analysis clearly considered and stated

+ 3 pts Some specific features of the data relevant to modelling identified; implications for future analysis partly considered

✔ **+ 2 pts** Obvious or generic implications identified; weak connection to future modelling made.

+ 1 pt Few relevant implications for modelling given with minimal justification or spurious commentary. Implications or discussion is not focussed on the data science.

+ 0 pts None given, or wholly irrelevant commentary.

5 Axis labels on this plot are swapped, as the shape does not match the plot on the right

6 Also, why not draw a plot for Income groups?\

7 It would be more helpful to show plastic waste horizontally

Question 4

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

14 / 20 pts

Visualisations: Choice

- ✔ **+ 4 pts** All visualisations appropriate to the goals.
- **+ 3 pts** Visualisations generally appropriate to the goals.
- **+ 2 pts** Visualisations only partly address the goals; more effective choices are not used; use of visualisations not covered in lectures
- **+ 1 pt** Inappropriate choice of visualisation.
- **+ 0 pts** No attempt; visualisations fails to address any of the key objectives.

Visualisations: Presentation

- ✔ **+ 4 pts** All visualisations clearly presented and of presentation quality.
- **+ 3 pts** Generally clearly presented and of good quality; some minor shortcomings (e.g. aspect ratio distorting the plot, poor choice of histogram bar width)
- **+ 2 pts** Some deficiencies in presentation of multiple plots (e.g. scale or aspect ratio renders plot difficult to read, comparison boxplots/histograms with different axis ranges, default labels/captions).
- **+ 1 pt** Major/careless weaknesses or omissions, e.g. many comparison boxplots/histograms using different axis ranges, unformatted R code, default illegible labels.
- **+ 0 pts** No attempt; inadequate presentation to address any of the key objectives.

Describe & Interpret: Data Features

- **+ 4 pts** All relevant data features identified; statistical relevance of the findings made clear.
- **+ 3 pts** Several relevant data features identified; statistical relevance to the problem made generally clear.
- ✔ **+ 2 pts** Primarily descriptive identification of features; statistical relevance to the problem not made clear.
- **+ 1 pt** Few relevant data features identified; exclusively descriptive; minimal reference to statistical relevance
- **+ 0 pts** No attempt; no salient features identified or discussed.

Describe & Interpret: Interpretation of findings

- **+ 4 pts** Sound interpretation of the data features in context and in terms of exploration goals
- **+ 3 pts** Adequate interpretation of the data features in context and in terms of exploration goals, with some omissions or spurious commentary
- ✔ **+ 2 pts** Some interpretation of the data features in context and in terms of exploration goals
- **+ 1 pt** Weak or minimal interpretation of the data features found in context and in terms of the exploration goals
- **+ 0 pts** No interpretation of the data features found in context and in terms of the exploration goals

Modelling implications

- **+ 4 pts** Salient and specific features of the data relevant to modelling identified; implications for future analysis clearly considered and stated

+ 3 pts Some specific features of the data relevant to modelling identified; implications for future analysis partly considered

✔ + 2 pts Obvious or generic implications identified; weak connection to future modelling made.

+ 1 pt Few relevant implications for modelling given with minimal justification or spurious commentary. Implications or discussion is not focussed on the data science.

+ 0 pts None given, or wholly irrelevant commentary.

Question 5

The smoothed trends

14 / 20 pts

Visualisations: Choice

- ✔ **+ 4 pts** All visualisations appropriate to the goals.
- **+ 3 pts** Visualisations generally appropriate to the goals.
- **+ 2 pts** Visualisations only partly address the goals; more effective choices are not used; use of visualisations not covered in lectures
- **+ 1 pt** Inappropriate choice of visualisation.
- **+ 0 pts** No attempt; visualisations fails to address any of the key objectives.

Visualisations: Presentation

- **+ 4 pts** All visualisations clearly presented and of presentation quality.
- ✔ **+ 3 pts** Generally clearly presented and of good quality; some minor shortcomings (e.g. aspect ratio distorting the plot, poor choice of histogram bar width)
- **+ 2 pts** Some deficiencies in presentation of multiple plots (e.g. scale or aspect ratio renders plot difficult to read, comparison boxplots/histograms with different axis ranges, default labels/captions).
- **+ 1 pt** Major/careless weaknesses or omissions, e.g. many comparison boxplots/histograms using different axis ranges, unformatted R code, default illegible labels.
- **+ 0 pts** No attempt; inadequate presentation to address any of the key objectives.

Describe & Interpret: Data Features

- **+ 4 pts** All relevant data features identified; statistical relevance of the findings made clear.
- **+ 3 pts** Several relevant data features identified; statistical relevance to the problem made generally clear.
- ✔ **+ 2 pts** Primarily descriptive identification of features; statistical relevance to the problem not made clear.
- **+ 1 pt** Few relevant data features identified; exclusively descriptive; minimal reference to statistical relevance
- **+ 0 pts** No attempt; no salient features identified or discussed.

Describe & Interpret: Interpretation of findings

- **+ 4 pts** Sound interpretation of the data features in context and in terms of exploration goals
- **+ 3 pts** Adequate interpretation of the data features in context and in terms of exploration goals, with some omissions or spurious commentary
- ✔ **+ 2 pts** Some interpretation of the data features in context and in terms of exploration goals
- **+ 1 pt** Weak or minimal interpretation of the data features found in context and in terms of the exploration goals
- **+ 0 pts** No interpretation of the data features found in context and in terms of the exploration goals

---

Modelling implications

- **+ 4 pts** Salient and specific features of the data relevant to modelling identified; implications for future analysis clearly considered and stated

✔ **+ 3 pts** Some specific features of the data relevant to modelling identified; implications for future analysis partly considered

+ 2 pts Obvious or generic implications identified; weak connection to future modelling made.

+ 1 pt Few relevant implications for modelling given with minimal justification or spurious commentary. Implications or discussion is not focussed on the data science.

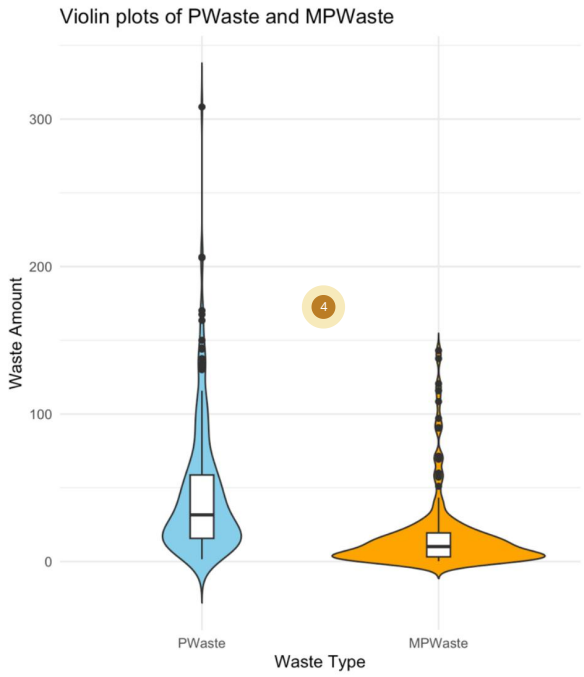
+ 0 pts None given, or wholly irrelevant commentary.

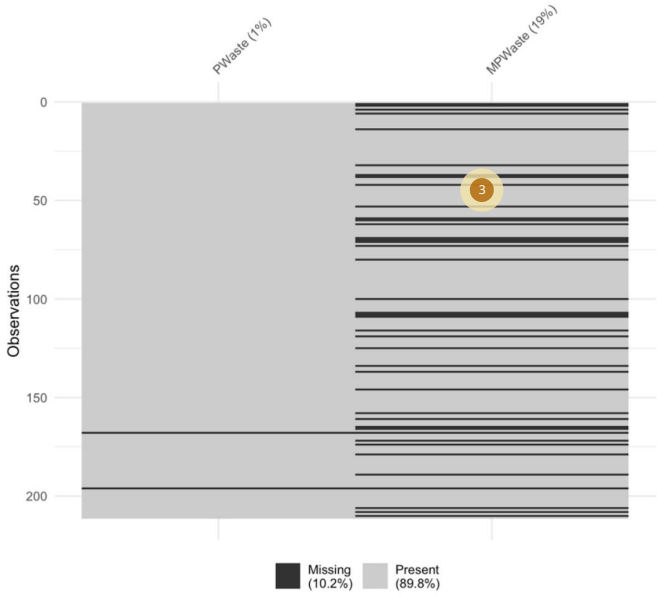
1 Influence of outliers on the trend line?

2 Good point.

Question assigned to the following page: 1

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





Relevant Features

1. There are large differences in the distribution of PWaste and MPWaste, with PWaste being more widely distributed. This suggests that the amount of per capita plastic waste varies considerably from country to country. The distribution of MPWaste is more concentrated, indicating that most countries have similar amounts of mismanaged plastic waste.

2. Both PWaste and MPWaste present some outliers, particularly on the higher end of the distribution. For PWaste, these outliers are far from the median value, suggesting that some countries have unusually high per capita plastic waste amounts. MPWaste also has outliers, but they are closer to the median, indicating fewer extremes in amount of mismanaged plastic waste.

3. The median MPWaste is lower than the median PWaste, reflecting the fact that, on average, countries have lower amounts of mismanaged plastic waste per capita compared to plastic waste per capita.

4. MPWaste has a higher proportion of missing data (19%), which may affect the statistical analysis and modelling of this variable.

Implications for Future Analysis

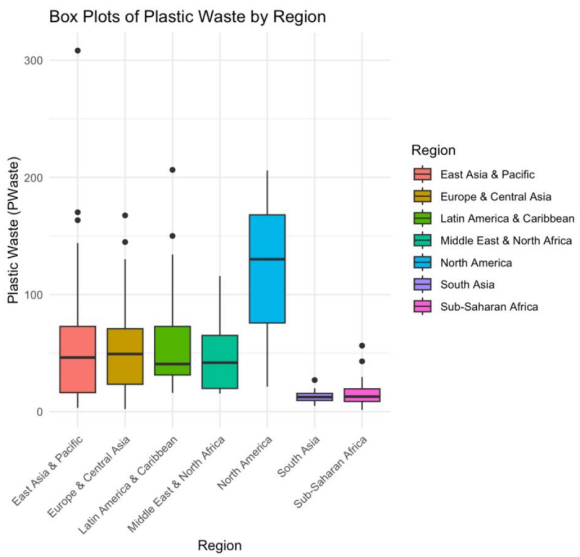
1. Missing Value Treatment: Given the high proportion of missing data for MPWaste, an appropriate approach to addressing missing values needs to be considered prior to any modelling of MPWaste. This may involve interpolation techniques or the use of models that can handle missing information, such as Bayesian models.

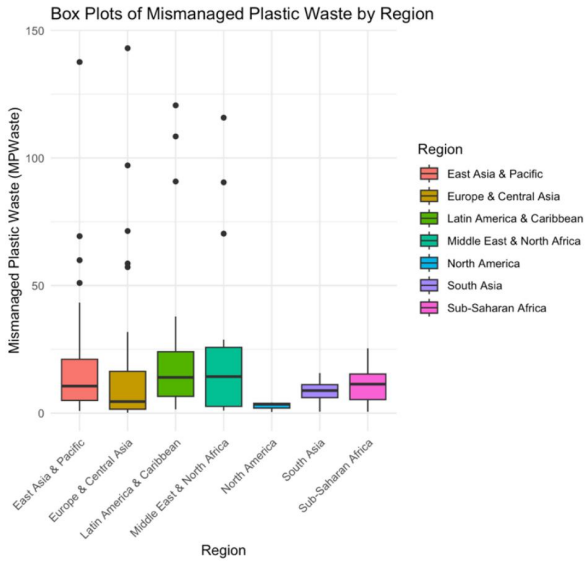
2. Outlier Analysis: The presence of outliers in both PWaste and MPWaste suggests that future analyses should examine these cases in detail to understand their nature. Are these instances of data input errors, or do they represent real cases of extreme waste management? If the latter, they require special attention.

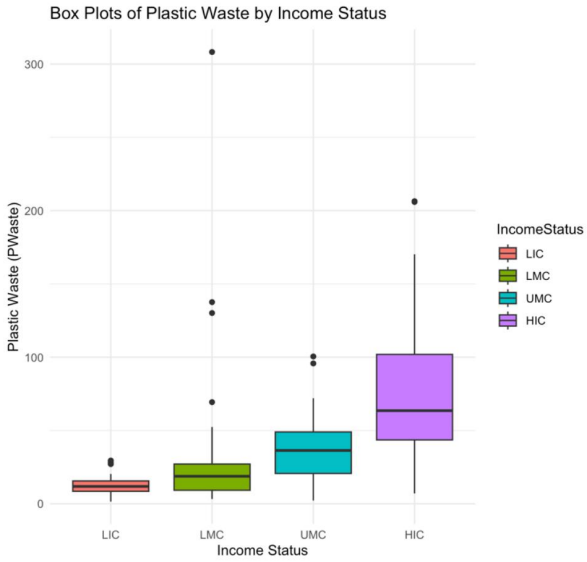
3. Data Transformation and Normalization: Considering the range and presence of outliers in the data, transformations such as log scaling may be needed to standardize the data distribution prior to modelling. This can help stabilize the variance and make the model outcomes more interpretable.

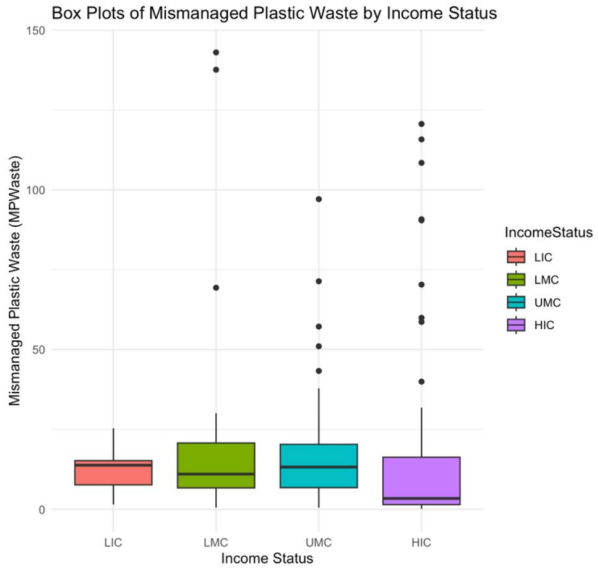
Question assigned to the following page: 2

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









Relevant Features

1. Region Impact:

a. Significant regional differences in PWaste are observed. Regions like North America, East Asia & Pacific exhibit higher medians and wider interquartile ranges, indicating higher per capita plastic waste and greater variability within these regions.

b. For MPWaste, regional differences are less marked when looking at the median, yet variations in distribution and outliers persist. This suggests that per capita mismanaged plastic waste is somewhat consistent across regions.

2. Income Status Impact:

a. Income status appears to significantly influence PWaste distribution. Higher-income countries tend to have higher median plastic waste per capita, implying that wealthier countries generate more plastic waste per capita.

b. Income status seems to have a weaker effect on MPWaste. While high-income countries may manage waste more efficiently and thus exhibit lower amounts of mismanaged waste per capita, the distribution of MPWaste remains relatively steady across different income statuses.

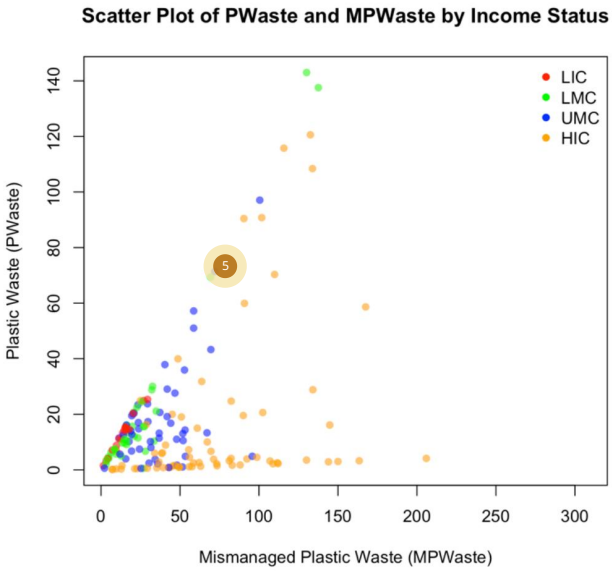
Questions assigned to the following page: 2 and 3

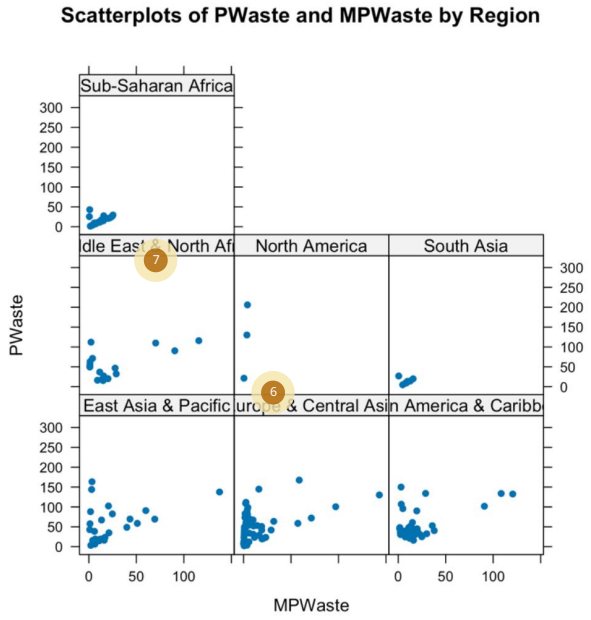
Implications for Future Analysis

1. For PWaste, noticeable differences in the distribution of PWaste by region and income status, suggesting that these factors are crucial predictors of plastic waste generation. Future models should consider including region and income status as explanatory variables to enhance the accuracy of predictions for plastic waste production.

2. For MPWaste, the distribution of MPWaste does not appear to significantly shift based on region or income status. Hence, future statistical models for MPWaste should explore additional socio-economic and policy-relevant factors, such as GDP, UrbanPopPC, or external variables not included in the dataset like infrastructure development, regional policies, and other factors.

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





Relevant Features

1. Relationship between PWaste and MPWaste: The data points mostly exhibit an increasing trend from left to right, which suggests a positive but albeit weak linear relationship between PWaste and MPWaste. This implies that countries with higher amounts of per capita plastic waste also tend to have higher levels of mismanaged plastic waste per capita. However, there are exceptions where plastic waste does not proportionately increase with higher levels of mismanagement.
2. Income Status Impact: The impact of income status on the relationship between PWaste and MPWaste is quite apparent. Low-Income Countries (LIC) cluster with lower levels of both PWaste and MPWaste. In contrast, High-Income Countries (HIC) exhibit a greater amount of MPWaste, however, PWaste does not escalate correspondingly. This suggests that while most HICs have relatively lower plastic usage, they still produce significant amounts of mismanaged plastic waste.
3. Region Impact: Regional variations are evident in the relationship between PWaste and MPWaste. Sub-Saharan Africa and South Asia show clusters with lower values for both PWaste and MPWaste, indicative of lower per capita amount of plastic waste and mismanaged waste. North America displays a broader range in PWaste but a narrower range in MPWaste, with sparse data overall. Other regions present disproportionate distributions.

Implications for Future Analysis

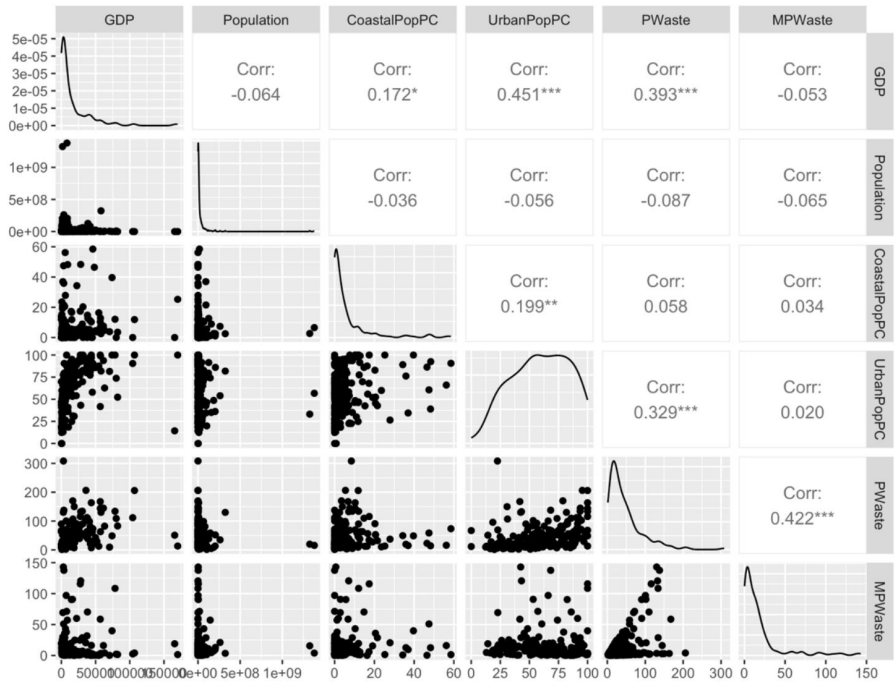
1. Modelling Considerations: The absence of a robust linear relationship between PWaste and MPWaste indicates that straightforward linear modeling may not be adequate. Exploring non-linear relationships or using models capable of

Questions assigned to the following page: 4 and 3

accommodating high variability, such as decision trees or random forests, might be more appropriate.

2. Need for Region-Specific Models: The relationship between PWaste and MPWaste varies by region, suggesting that region-specific models may be more predictive than global models.

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



Relevant Features

1. For PWaste, it has a weak positive correlation with the variables GDP and UrbanPopPC, suggesting that the amount of plastic waste per capita may increase with GDP and urban population. However, it has little or no linear relationship with Population and CoactalPopPC.

2. For MPWaste, it seems to have little or no linear relationship with other quantitative variables.

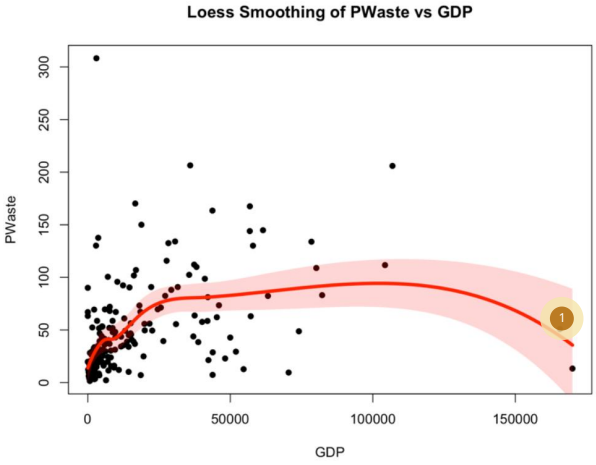
Implications for Future Analysis

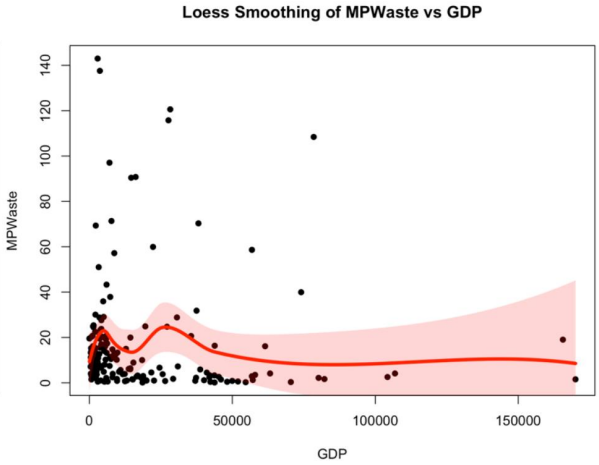
1. The weak positive correlation between PWaste and both GDP and UrbanPopPC suggests that economic status and the degree of urbanization may be factors influencing the amount of plastic waste per capita. Therefore, these factors should be considered in future models to predict per capita plastic waste production more accurately.

2. For MPWaste, although there is little to no linear relationship with other quantitative variables, this does not preclude the existence of other types of relationships. More complex analyses may be required to explore potential non-linear relationships or other influencing factors.

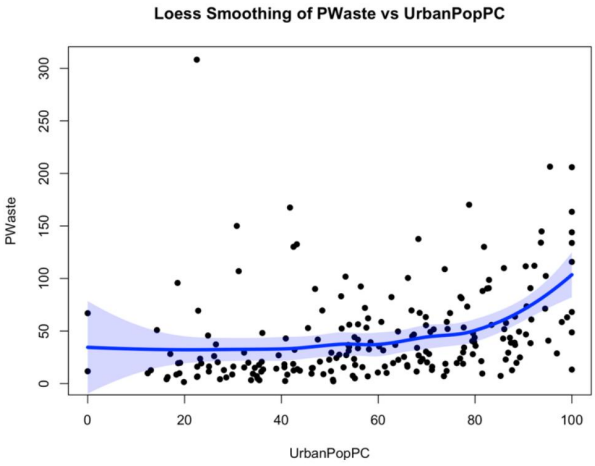
Question assigned to the following page: 5

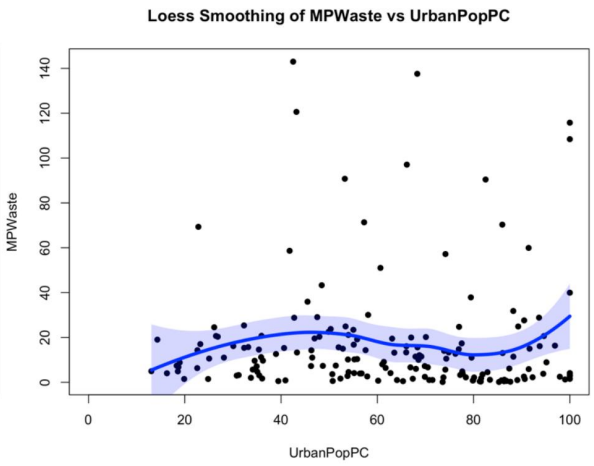
Q5: (i) The smoothed trends between both types of plastic waste and GDP





(ii) The smoothed trends between both types of plastic waste and the size of urban population





Relevant Features

1. Trends of GDP and PWaste: PWaste increases with national GDP, notably at lower GDP levels, suggesting that the amount of plastic waste per capita tends to rise more rapidly as economies grow, particularly in countries with smaller economies. However, at higher GDP levels, the trend of PWaste growth becomes flat.

2. Trends of GDP and MPWaste: In countries with lower GDP, the per capita amount of mismanaged plastic waste is highly volatile but gradually declines and stabilizes as GDP increases.

3. Trend of UrbanPopPC and PWaste: Initially, the relationship between PWaste and UrbanPopPC is negligible. However, with an increase in UrbanPopPC, PWaste exhibits an upward trend, indicating that more urbanized countries generate greater amounts of plastic waste per capita.

4. Trend of UrbanPopPC and MPWaste: At low to medium levels of UrbanPopPC, MPWaste fluctuates slightly but follows a similar upward trend in countries with higher urban populations.

Implications for Future Analysis

1. The presence of outliers in the data affects the slope of the trend lines. Future data analyses should first identify and investigate these outliers to attempt to explain them, ensuring the accuracy and generalizability of the model.

2. The relationship of MPWaste with GDP and UrbanPopPC is more complex than that of PWaste. Instead of following a single linear trend, MPWaste's relationship with these variables exhibits a multivariate and non-linear pattern. This complexity suggests that more sophisticated or advanced modelling techniques might be necessary for accurately modelling MPWaste in future analyses.