Please use this proforma at the beginning of your TMA to indicate how/if you have used generative AI.

I have used Generative AI in this TMA (such as Copilot, Gemini or ChatGPT) to help with the following:
[please tick all that apply]

☐ As a starting point or inspiration with a part of the TMA.

☐ To improve my own work, like the interpretation/summary of results.

☐ To summarise materials I found on the web for this TMA

☒ I did not use generative AI to help me with this TMA

Q 1.

(a)

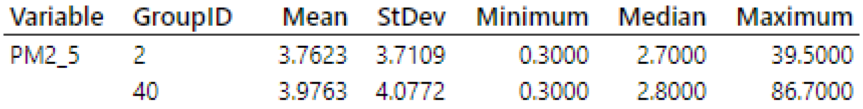
(i) Two graphical displays suitable for studying the distribution of tree heights are the frequency histogram, and boxplot. Both displays are suitable for visualising the location and dispersion of continuous data. The histogram gives a more granular overview of the modality of the distribution, whereas the boxplot highlights the position of the 1st, 2nd (median), and 3rd quartiles, interquartile range, and potential outliers.

(ii) The number of trees alive at each age could be represented using a bar chart, as the age in years is discrete (in this context) and a bar chart allows us to plot a categorical or discrete variable against a continuous one (or a count, in this case).

(iii) Two graphical displays suitable for studying the way the heights of trees depend on their ages, are unit-area histograms, and comparative boxplots. Unit-area histograms allow the location and shape of the data to be compared between categories, even if the number of trees at each age varies. Comparative boxplots also allow the location and shape of the data to be compared between groups, but are effective when there are a number of categories to be visualised in the same diagram.

(b)

(i)



Variable	GroupID	Mean	StDev	Minimum	Median	Maximum
PM2_5	2	3.7623	3.7109	0.3000	2.7000	39.5000
	40	3.9763	4.0772	0.3000	2.8000	86.7000

Figure 1: The mean, median, standard deviation, minimum, and maximum values of the PM2_5 variable, calculated per GroupID using Minitab.

(ii)

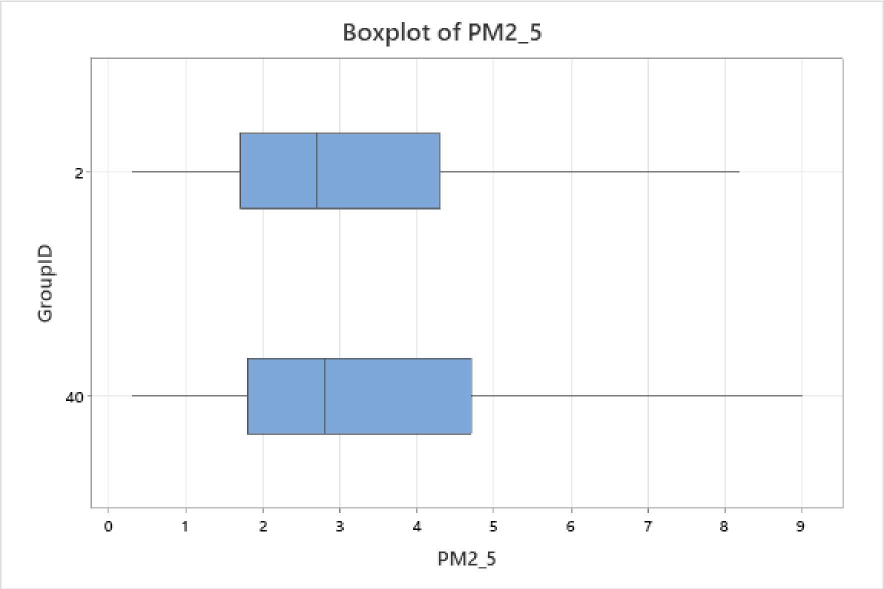


Figure 2: Comparative boxplots showing the distribution of the PM2_5 variable, separately per Group. Outliers are not shown.

(iii)

Running commentary TMA01 Q1b: Description, Similarities and Differences for Group 2 and Group 40.

As shown in Figure 1, the PM2_5 variable has a similar central location in Groups 2 and 40, with the mean and median slightly higher in Group 40. Group 40 has a slightly higher standard deviation and, as shown in Figure 2, interquartile range. Both groups have the same minimum value of 0.3, but Group 40's maximum value is more than twice the largest value in Group 2. Figure 2 suggests both data distributions exhibit positive skew, with Group 40's skew being slightly more pronounced.

(iv)

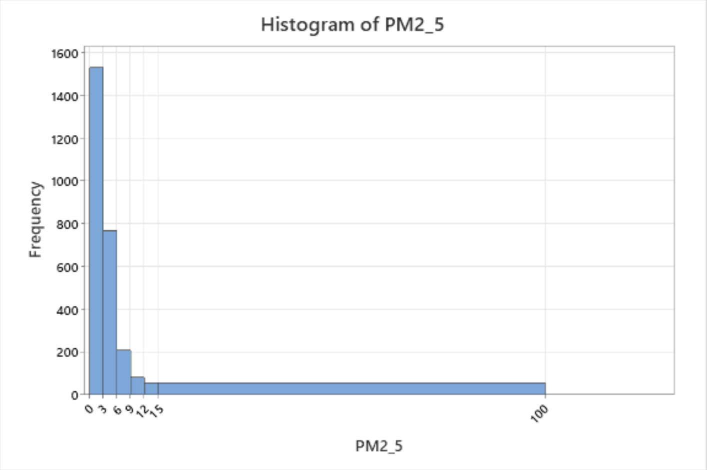


Figure 3: Frequency histogram for the PM2_5 variable for Group 2. Cutpoints have been set to 0, 3, 6, 9, 12, 15, and 100µgm^{-3}.

(v)

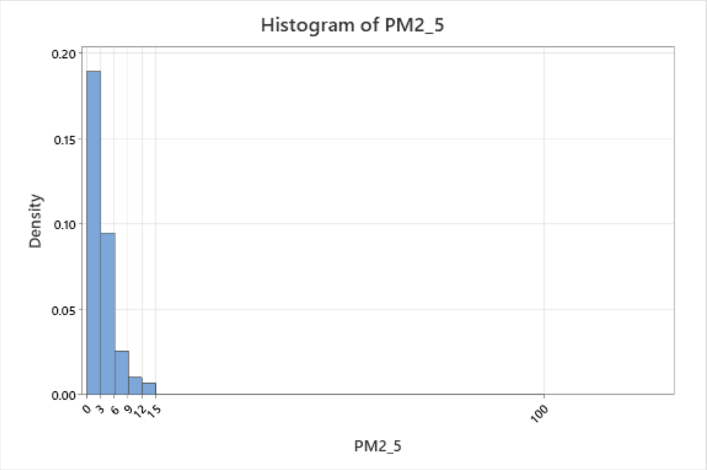


Figure 4: Unit-area histogram for the PM2_5 variable for Group 2. Cutpoints have been set to 0, 3, 6, 9, 12, 15, and 100µgm^{-3}.

The histogram in Figure 4 differs from that in Figure 3 by the y values of the bars: the values correspond to counts in Figure 3 but density in Figure 4. The shape of the distribution is the same for bins with the same width between the histograms, but the height of the bar representing $15 \leq$ PM2_5 < 100 is lower in Figure 4 relative to the other bars in the histogram. Summing the heights of the bars gives

$$(0.1889\times3)+(0.0947\times3)+(0.0257\times3)+(0.0101\times3)+(0.0070\times3)+(0.0002\times85) = 0.9962$$

This suggests that, assuming some rounding error, that the histogram does have unit area as the sum of the area of the bars is ≈ 1.0.

Q 2.

(a)

In a long sequence of repetitions of a study or experiment, random samples tend to settle down towards probability distributions in the sense that for discrete data, bar charts settle down towards probability **mass** functions, and for continuous data, **unit-area** histograms settle down towards probability **density** functions. As the sample size increases, the amount of difference between successive graphical displays obtained from the data **decreases**.

(b)

(i)

The range of X is the set $\{0, 1, 2, \ldots, 6\}$.

(ii)

Kevin's p.m.f. is valid because each individual probability is in the range $[0, 1]$ and the sum of the probabilities in the range of X gives $0.3 + 0.2 + 0.2 + 0.1 + 0.1 + 0.05 + 0.05 = 1$.

(iii)

According to Kevin's p.m.f., the probability of there being one bicycle is $P(X = 1) = 0.2$.

(iv)

x	0	1	2	3	4	5	6
$p(x)$	0.30	0.20	0.20	0.10	0.10	0.05	0.05
$F(x)$	0.30	0.50	0.70	0.80	0.90	0.95	1.00

Table 1: Table showing the probability mass function ($p(x)$) and cumulative distribution function ($F(x)$) for the number of bicycles.

(v)

The probability of $P(X < 3)$ is given by $F(2) = 0.7$, and the probability $P(X \geq 5)$ is given by $1 - F(4) = 1 - 0.9 = 0.1$.

(c)

(i)

Integrating $f(x) = 4x(1 - x)(2 - x)$ w.r.t. x gives

$$\begin{aligned} \int 4x(1-x)(2-x)\ dx &= \int 4x^3 - 12x^2 + 8x\ dx \\ &= x^4 - 4x^3 + 4x^2 + c \\ &= x^2(x^2 - 4x + 4) + c \\ &= x^2(2 - x)^2 + c \end{aligned}$$

as required.

(ii)

The function $f(x)$ is zero at its endpoints:

$$\begin{aligned} f(0) &= 4(0)(1-0)(2-0) = 0 \\ f(1) &= 4(1)(1-1)(2-1) = 0 \end{aligned}$$

and is positive at its stationary point:

$$\begin{aligned} f'(x) &= 12x^2 - 24x + 8 \\ 0 &= 12x^2 - 24x + 8 \\ x &\approx 0.42 \end{aligned}$$

As $f(x)$ is non-negative at its end-points and only stationary point, it must therefore be non-negative across its range.

Using the result from part (i), the definite integral for $f(x)$ across its range is

$$\begin{aligned} \int 4x(1-x)(2-x)\ dx &= \left[x^2(2-x)^2\right]_0^1 \\ &= 1 - 0 \\ &= 1 \end{aligned}$$

As the $f(x)$ is non-negative and integrates to 1 across its range, it is a valid p.d.f.

(iii)

The c.d.f. associated with this p.d.f. is

$$\begin{aligned} \int_0^x 4y(1-y)(2-y)\ dy &= \left[y^2(2-y)^2\right]_0^x \\ &= x^2(2-x)^2 - 0 \\ &= x^2(2-x)^2 \end{aligned}$$

(iv)

Let $F(x)$ be the c.d.f. derived above. The probability shown can be calculated as

$$\begin{aligned} P\left(\frac{1}{3} < X < \frac{2}{3}\right) &= F\left(\frac{2}{3}\right) - F\left(\frac{1}{3}\right) \\ &= \left(\frac{2}{3}\right)^2\left(2-\left(\frac{2}{3}\right)\right)^2 - \left(\frac{1}{3}\right)^2\left(2-\left(\frac{1}{3}\right)\right)^2 \\ &= \frac{64}{81} - \frac{25}{81} \\ &= \frac{39}{81} \end{aligned}$$

as required.