Name: Yola Charara

HW3

1.

Age: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70

a. Smoothing by bin means using a bin depth of 3:

Equal depth partitioning:

Age: (13, 15, 16), (16, 19, 20), (20, 21, 22), (22, 25, 25), (25, 25, 30), (33, 33, 35), (35, 35, 35), (36, 40, 45), (46, 52, 70).

Bin 1: $\frac{13+15+16}{3} = 14.67$
Bin 2: $\frac{16+19+20}{3} = 18.33$
Bin 3: $\frac{20+21+22}{3} = 21$
Bin 4: $\frac{22+25+25}{3} = 24$
Bin 5: $\frac{25+25+30}{3} = 26.67$
Bin 6: $\frac{33+33+35}{3} = 33.67$
Bin 7: $\frac{35+35+35}{3} = 35$
Bin 8: $\frac{36+40+45}{3} = 40.33$
Bin 9: $\frac{46+52+70}{3} = 56$

Data result after smoothed:

Age: 14.67, 14.67, 14.67, 18.33, 18.33, 18.33, 21, 21, 21, 24, 24, 24, 26.67, 26.67, 26.67, 33.67, 33.67, 33.67, 35, 35, 35, 40.33, 40.33, 40.33, 56, 56, 56.

Smoothing by bin means decreases local variation. It suppresses small fluctuations but at the same time shows the overall trend. Values that are very far from the average are brought nearer to the mean (such as 70 in this case).

b. IQR = Q3 - Q1 = 35 - 20 = 15
Q1 - 1.5 * IQR = 20 - 1.5 * 15 = -2.5
Q3 + 1.5 * IQR = 35 + 1.5 * 15 = 57.5

Data points outside this range [-2.5; 57.5] are outliers, therefore 70 is an outlier.

c. $x' = \frac{x - min}{max - min} = \frac{35 - 13}{70 - 13} = 0.386$

The normalized value is 0.386

d. Mean = $\mu = \frac{\sum age}{n} =$

$$\frac{13 + 15 + 16 + 16 + 19 + 20 + 20 + 21 + 22 + 22 + 25 + 25 + 25 + 25 + 30 + 33 + 33 + 35 + 35 + 35 + 35 + 36 + 40 + 45 + 46 + 52 + 70}{27}$$

$= \frac{809}{27} = 29.96$

Standard deviation $= \sigma = \sqrt{\frac{\sum (x - \mu)^2}{n}} =$

$$\sqrt{\frac{(13-29.96)^2 + (15-29.96)^2 + (16-29.96)^2 + (16-29.96)^2 + (19-29.96)^2 + (20-29.96)^2 + (20-29.96)^2 + (21-29.96)^2 + (22-29.96)^2 + (22-29.96)^2 + (25-29.96)^2 + (25-29.96)^2 + (25-29.96)^2 + (25-29.96)^2 +}{27}}$$

$$\sqrt{\frac{+ (30-29.96)^2 + (33-29.96)^2 + (33-29.96)^2 + (35-29.96)^2 + (35-29.96)^2 + (35-29.96)^2 + (35-29.96)^2 + (36-29.96)^2 + (40-29.96)^2 + (45-29.96)^2 + (46-29.96)^2 + (52-29.96)^2 + (70-29.96)^2}{27}}$$

$\sigma = 12.70$

Z-score = $z = \frac{x - \mu}{\sigma} = \frac{35 - 29.96}{12.70} = 0.397$

The z-score for age 35 is 0.397.

e. x = 35; max age = 70

$j \Rightarrow \frac{max(|age|)}{10^j} < 1$

$\frac{70}{10^2} = 0.7 < 1$

So, j = 2

$x' = \frac{35}{10^2} = 0.35$

The transformed value of the age 35 is 0.35.

3.

Base info:

Junior count: 113; Senior count: 52; Total count: 165

Info = $- \sum pi \log_2 (pi) = - \frac{113}{165} \log_2 (\frac{113}{165}) - \frac{52}{165} \log_2 (\frac{52}{165}) = 0.90$

Department info gain:

Info sales = $-\frac{80}{110} log_2 (\frac{80}{110}) - \frac{30}{110} log_2 (\frac{30}{110}) = 0.845$

Info systems = $-\frac{23}{31} log_2 (\frac{23}{31}) - \frac{8}{31} log_2 (\frac{8}{31}) = 0.824$

Info marketing = $-\frac{4}{14} log_2 (\frac{4}{14}) - \frac{10}{14} log_2 (\frac{10}{14}) = 0.863$

Info secretary = $-\frac{6}{10} log_2 (\frac{6}{10}) - \frac{4}{10} log_2 (\frac{4}{10}) = 0.971$

Info department = $\frac{110}{165} (0.845) + \frac{31}{165} (0.824) + \frac{14}{165} (0.863) + \frac{10}{165} (0.971) = 0.850$

Gain department = 0.90 - 0.850 = 0.05

Age info gain:

Only (31_35) has both junior and senior counts. So, all other age datapoints' info are equal to zero.

Info (31_35) = $-\frac{44}{79} log_2 (\frac{44}{79}) - \frac{35}{79} log_2 (\frac{35}{79}) = 0.991$

Info age = $\frac{79}{165} (0.991) = 0.474$

Gain age = 0.90 - 0.474 = 0.426

Salary info gain:

Only (46K_50K) has both junior and senior counts. So, all other salary datapoints' info are equal to zero.

Info (46K_50K) = $-\frac{23}{63} log_2 (\frac{23}{63}) - \frac{40}{63} log_2 (\frac{40}{63}) = 0.947$

Info salary = $\frac{63}{165} (0.947) = 0.362$

Gain salary = 0.90 - 0.362 = 0.538

Root of the decision tree:

The salary gain is the highest gain among the attributes. So, the root node will be the salary.

Subsets of the decision tree:

For this part, the calculations were done in R markdown.

Results:

- Subset for Salary = 46K-50K:
  - ➔ Junior = 23, Senior = 40
  - ➔ Info = 0.947
  - ➔ Info gain department = 0.947
  - ➔ Info gain age = 0.947

Both attributes are equal and can work as the next node. If we were to choose, we can pick Age as the first subset after the root for a more organized structure.