دانشگاه صنعتی اصفهان
دانشکده مهندسی برق و کامپیوتر

عنوان: تکلیف چهارم درس داده کاوی

نام و نام خانوادگی: نیلوفر سعیدی
شماره دانشجویی: ۹۸۲۲۹۶۳

1

1.1 width-equal(4 bins)

bin width = $(maximum\_value - minimum\_value)/No\_of\_bins = (215 - 5)/4 = 52.5$

Bin 1: 5 to 57.5 including(5,10, 11, 13, 15, 35, 50, 55)

Bin 2: 58 to 110.5 including(72, 92)

Bin 3: 111 to 163.5 including(-)

Bin 4: 164 to 215 including(204, 215)

The last bin has a higher endpoint because the maximum value in the data (215) is not evenly divisible by the bin width.

1.2 depth-equal(4 bins)

Bin 1: 5, 10, 11

Bin 2: 13, 15, 35

Bin 3: 50, 55, 72

Bin 4: 92, 204, 215

2

median = 10

Lower half: 0, 0, 2, 5, 8, 8, 8, 9, 9, 10

Upper half: 10, 10, 11, 12, 12, 12, 14, 15, 20, 25

Q1 (median of the lower half) = 8

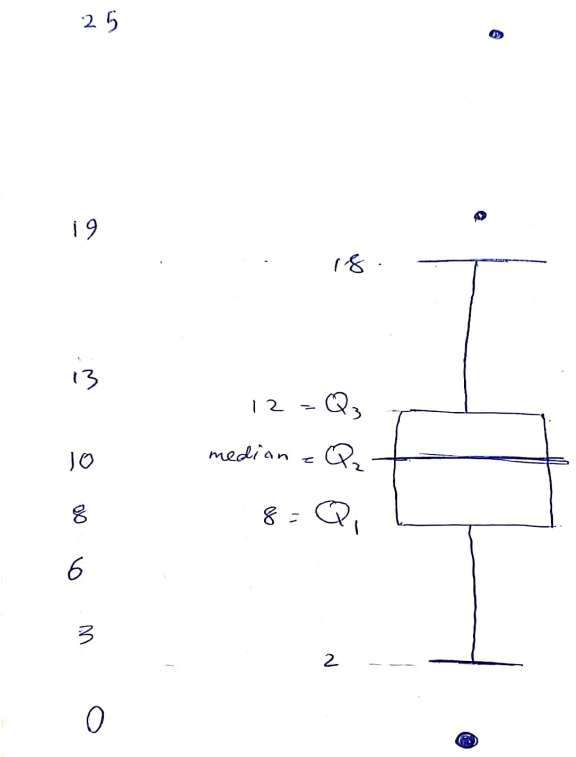
Q3 (median of the upper half)= 12

IQR = Q3 - Q1 = 12 - 8 = 4

Lower outlier = Q1 - 1.5 * IQR = 8 - 1.5 * 4 = 2

Upper outlier = Q3 + 1.5 * IQR = 12 + 1.5 * 4 = 18

Therefore, the outliers are 0, 0, 20, 25.



3

3.1

We group the data into sets of three consecutive numbers. Each value in the bin is then replaced by its closest boundary value. When the value is equal to both sides, it will be replaced by the front boundary value.

Before bin Boundary: 13, 15, 16

After Bin Boundary: 13, 16, 16

Before bin Boundary: 16, 19, 20

After Bin Boundary: 16, 20, 20

Before bin Boundary: 20, 21, 22

After Bin Boundary: 20, 20, 22

Before bin Boundary: 22, 23, 23

After Bin Boundary: 22, 23, 23

Before bin Boundary: 23, 25, 25

After Bin Boundary: 23, 25, 25

Before bin Boundary: 25, 35, 35

After Bin Boundary: 25, 35, 35

Before bin Boundary: 35, 35, 36
After Bin Boundary: 35, 35, 36
Before bin Boundary: 40, 45, 46
After Bin Boundary: 40, 46, 46
Before bin Boundary: 52, 70
After Bin Boundary: 52, 70
Smoothed data: 13, 16, 16, 16, 20, 20, 20, 20, 22, 22, 23, 23,23, 25, 25,25, 35, 35, 35, 35, 36, 40, 46, 46, 52, 70. There are also other methods for smoothing using binning, one is replacing each number with the average of its bin.
This method seems to work for the data.

3.2

- Z-score method: There are no data points with a z-score greater than 3 or 4, so there are no outliers detected by this method.
- IQR method: The first quartile (Q1) is 20 and the third quartile (Q3) is 35. The IQR is therefore 15. Data points outside the range [Q1 - 1.5 * IQR, Q3 + 1.5 * IQR] = [-7.5, 62.5] are considered outliers. The number 70 is the only data point outside this range and is therefore an outlier.

3.3

- Moving average: This method involves calculating the average of a sliding window of data points. The width of the window can be adjusted to control the amount of smoothing. Moving averages are simple to implement and can be effective at removing high-frequency noise.
- Savitzky-Golay filter: This method involves fitting a polynomial function to a sliding window of data points and using the coefficients of the polynomial to calculate a smoothed value for the center point of the window. The width of the window and the order of the polynomial can be adjusted to control the amount of smoothing. Savitzky-Golay filters are particularly effective at preserving the shape of the data while removing noise.
- Fourier transform: This method involves decomposing the data into a set of sinusoidal functions using a Fourier transform and then filtering out high-frequency components. The filtered data can then be reconstructed using an inverse Fourier transform.