Domas Budrys – Assignment 2 CSCI5080

Question 1 (Time spent: 30min):

	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5
0	15.0	22.2	27.6	34.2	45.2
1	15.0	22.2	27.6	34.2	45.2
2	15.0	22.2	27.6	34.2	45.2
3	15.0	22.2	27.6	34.2	45.2
4	15.0	22.2	27.6	34.2	45.2

```
q1 = [11, 13, 15, 17, 19,
      21, 21, 23, 23, 23,
      23, 25, 27, 30, 33,
      33, 33, 33, 36, 36,
      38, 40, 46, 48, 54]

dfQ1 = pd.DataFrame()

for num, val in enumerate(q1, start=1):


    if num <= 5:
        bin_val1 = np.mean(q1[:5])

    elif num > 5 and num <= 10:
        bin_val2 = np.mean(q1[5:10])

    elif num > 10 and num <= 15:
        bin_val3 = np.mean(q1[10:15])

    elif num > 15 and num <= 20:
        bin_val4 = np.mean(q1[15:20])

    elif num > 20 and num <= 25:
        bin_val5 = np.mean(q1[20:25])

        dfQ1 = dfQ1.append({'Bin 1': bin_val1,
                            'Bin 2': bin_val2,
                            'Bin 3': bin_val3,
                            'Bin 4': bin_val4,
                            'Bin 5': bin_val5}, ignore_index=True)
```

Question 2 (Time spent: 1h 30min):

a)

Min-Max Normalization

	100	200	400	700	1100
0	0.0	0.1	0.3	0.6	1.0

b)

z-score Normalization

```
array([-1.06066017, -0.70710678, -0.35355339,  0.35355339,  1.76776695])
```

c)

z-score Normalization with mean absolute deviation

```
[-1.25, -0.9375, -0.3125, 0.625, 1.875]
```

d)

Decimal Scaling:

```
[0.01, 0.02, 0.04, 0.07, 0.11]
```

Question 3 (Time spent: 1h 15min):

a)

25	0.325581

b)

25	-0.353419

c)

25	0.25

d)

Decimal scaling should be the most appropriate method of normalization for the *age* data set. Any number between 1-99 could be added to data set and it would not affect any other values in the list. This would not apply if Min-Max or Z-score method is used.

Question 4 (Time spent: 50min)

a)

```
0.8553866571223405
```

Values are positively correlated, meaning that when person's age increases, cholesterol levels also increase.

Calculation:

$$E(Age) = \frac{20 + 22 + 25 + 25 + 36 + 40 + 45 + 48 + 49 + 51 + 53 + 53 + 57 + 58 + 59 + 60 + 61 + 62}{18} = 45.78$$

$$E(Fat) = \frac{8.4 + 25.3 + 7.6 + 18.8 + 27.5 + 24.6 + 28.1 + 28.8 + 30.2 + 32.7 + 40.2 + 29.8 + 32.3 + 30.7 + 33.9 + 40.1 + 33.1 + 36.4}{18} = 28.25$$

$$Cov(Age, Fat) = (20 * 8.4 + 22 * 25.3 + 25 * 7.6 + 25 * 18.8 + 36 * 27.5 + 40 * 24.6 + 45 * 28.1 + 48 * 28.8 + 49 * 30.2 + 51 * 32.7 + 53 * 40.2 + 53 * 29.8 + 57 * 32.3 + 58 * 30.7 + 59 * 33.9 + 60 * 40.1 + 61 * 33.1 + 62 * 36.4)/18 - 45.78 * 28.25 = 104.86$$

$$Correlation\ Coeefficient = \frac{104.86}{13.978 * 8.775} = 0.85$$

Question 5 (Time spent: 1h):