Hands-on Data Analytics (1001), Assignment 4 (quiz)
Short Exercises:

Question 1 (15 points): Explain what outlier is. How to handle the outlier if it is found?

Outliers are records with **rare** and **extreme** values that are different from those of other records. There are different ways to handle outliers. The following are the most common :

- Removal
- Include in the analysis but have a separate model for those values
- Imputation of outlier with other values as for missing values

Question 2 (16 points):): Given the numbers: (ver. A) 50, 23, 13, 4, 9, 3, 19, 29, 53, 11. (ver. B) 5, 2, 13, 4, 9, 3, 10, 20, 5, 10. Please find out the followings:

- 1st quartile
- 2nd quartile
- 3rd quartile
- Interquartile range (IQR)

A: Ordered sequence: 3, 4, 9, 11, 13, 19, 23, 29, 50, 53

- 1st quartile = 9
- 2nd quartile = (13 + 19) /2 = 16
- 3rd quartile = 29
- Interquartile range (IQR) = 29 – 9 = 20

B: Ordered sequence: 2, 3, 4, 5, 5, 9, 10, 10, 13, 20

- 1st quartile = 4
- 2nd quartile = (5 + 9) /2 = 7
- 3rd quartile = 10
- Interquartile range (IQR) = 10 - 4 = 6

Question 3 (15 points): Given the following data: Age = (ver. A) [10 15 15 18 19 20]. (ver. B) [2, 5, 5, 8, 15, 20]. Calculate the **min-max** normalized values for Age

A: AgeNorm = [0, 0.5, 0.5, 0.8, 0.9, 1]
B: AgeNorm = [0, 1/6, 1/6, 1/3, 13/18, 1] or [0, 0.17, 0.17, 0.33, 0.72, 1]

Hands-on Data Analytics (1001), Assignment 4 (quiz)

Multiple Choices (Only one choice is correct. Each question is worth **6 points**). Report answers in the table:

1. Box plot can visualize the following characteristics of data, except for?
   A. Median
   B. Interquartile Range (IQR)
   D. Trend
   E. Outliers

2. What is the main difference between a histogram and a bar chart?
   A. Bar chart produces a plot in color while histogram is black/white
   B. Bar chart is suitable for 3-dimensional data while histogram is for 1-dimensional data
   C. Bar chart is suitable for categorical data while histogram is for numeric data
   D. There is not a difference between the two plot

3. Which of the following is not a dimension reduction technique?______________
   A. Row filtering
   B. Ratio of missing values
   C. High correlation
   D. Low variance

4. What is the **median** value of (1, 2, 3, 4, 5, 10, 20, 30)?
   A. 4
   B. 5
   C. 4.5
   D. 3.375
   E. 10

5. What is the mean value of (10, 50, 30, -10, 20, 20)? ________________
   A. 10
   B. 0
   C. 50
   D. 20
   E. None of the above

6. Which of the following is **not** a normalization technique? ________________
   A. Min-Max normalization
   B. Z-score standardization
   C. Decimal scaling
   D. Divide by min
   E. Robust Z-score standardization (IQR)

7. Which of the following is the most common approach to impute missing values? ______________
   A. The median value for all records
   B. The max value for all records
   C. The min value for all records
   D. A value of 0
   E. A value of 100

8. In a database a value for "weight" is missing. Of the following reasons one is particularly severe as it can highly impact the analysis if the record is deleted. Which reason is it?_____________
   A. The value is missing completely at random
   B. The value is missing because the respondent did not know her weight
   C. The value is missing because the respondent did not want to reveal that he is obese
   D. The value is missing because the data got accidentally deleted
   E. The value is missing because the respondent was not sure she should enter in "kg" or in "pound"

9. When a KNIME workflow does not give the expected results, you may try to fix it by: 1) WeChat/email your instructor and ask for help, 2) check if the node has been configured properly, 3) restart the program or computer, 4) check if there is any warning/error message. What is the right order for our course? ______
   A. 1 – 2 – 3 – 4
   B. 2 – 3 – 4 – 1
   C. 4 – 2 – 3 – 1
   D. 3 – 1 – 2 – 4