**Assignment 1** (DDL: 11:59 pm, Sep 30, 2022)

**Note:**

1. **For all the questions, if plagiarism was found, then the WHOLE assignment will get ZERO.**
2. **Only SOFT COPY is allowed to submit, if your work is hand writing, please scan it then upload it to iSpace. You should guarantee your solution is clear, or you will lose marks.**
3. Classify the following attributes as binary, discrete, or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some cases may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity. (18 points, 2 points for each)
   1. Student ID
   2. Temperatures in Celsius and Fahrenheit.
   3. The Kelvin temperature
   4. Students’ performance measured by A, B, C, D, F
   5. Distance from T3 building to the position you stand
   6. Ore hardness measured by people’s judgement
   7. Number of students from different majors of UIC
   8. Date
   9. Gender of human beings (No special cases considered)
4. Following table is the demographics (in million) in 2017 with age ranging from 30 to 69.

|  |  |
| --- | --- |
| **Age** | **Demographics** |
| 30 - 34 | 88.96 |
| 35 - 39 | 82.55 |
| 40 - 44 | 87.71 |
| 45 - 49 | 105.48 |
| 50 - 59 | 96.76 |
| 60 - 64 | 59.82 |
| 65 - 69 | 68.04 |

Compute an approximate median value for the data. (14 points. Computation: 12 points; result: 2 points)

1. Following table is the salary statistics (in thousand) of a company in July.
2. Show a boxplot of the data. (12 points. Computation: 10 points; result: 2 points, you should compute all the elements included in boxplot as well as judge the outlier.)
3. Show a scatter plot of the data. (5 points)

|  |  |  |
| --- | --- | --- |
| 13 | Ray | 14.99 |
| 14 | Joyce | 14.99 |
| 15 | Paul | 15.22 |
| 16 | Abel | 15.4 |
| 17 | Elina | 15.41 |
| 18 | Ben | 15.44 |
| 19 | Bill | 15.49 |
| 20 | Kylle | 15.79 |
| 21 | Sims | 16.04 |
| 22 | Lora | 16.58 |
| 23 | Carl | 16.62 |
| 24 | Chris | 20.31 |
| 25 | Lee | 30.16 |

|  |  |  |
| --- | --- | --- |
| **No.** | **Name** | **Salary** |
| 1 | Vivi | 10.78 |
| 2 | Alan | 12.39 |
| 3 | Glen | 13.99 |
| 4 | Jack | 14.08 |
| 5 | Leon | 14.15 |
| 6 | Edith | 14.34 |
| 7 | Kelly | 14.47 |
| 8 | Lydia | 14.77 |
| 9 | Aldan | 14.81 |
| 10 | Jesse | 14.86 |
| 11 | Cecile | 14.91 |
| 12 | Fred | 14.91 |

1. Given two vectors,

Compute:

1. Hamming Distance (3 points)
2. Jaccard Coefficient (3 points)
3. Simple Matching Coefficient (3 points)
4. Suppose that you want to compare how dissimilar two organisms of different species are in terms of the number of genes they share. Describe which measure in (1) to (3) would be more appropriate? Explain why. (Conclusion: 2 points, reason: 4 points)
5. Give four vectors,

Compute:

1. Similarity matrix using cosine similarity (10 points. Computation: 8 points; result: 2 points)
2. Distance matrix using Euclidean Distance (10 points. Computation: 8 points; result: 2 points)
3. Pearson Correlation between x1 and x2, x3 and x4. (6 points, 3 points for each)