**Assignment 3:**

**Data Mining (CSE 4052)**

1. 1. **Suppose that the data for analysis includes the attribute age. The age values for the data tuple are (in increasing order) 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.**

1. **What is the mean of the data? What is the median?**

*The (arithmetic) mean of the data is: ¯ x = 1 n Pn i=1 xi = 809/27 = 30. The median (middle value of the ordered set, as the number of values in the set is odd) of the data is: 25.*

**ii. What is the mode of the data?**

*This data set has two values that occur with the same highest frequency and is, therefore, bimodal. The modes (values occurring with the greatest frequency) of the data are 25 and 35.*

**iii. What is the midrange of the data?**

*The midrange (average of the largest and smallest values in the data set) of the data is: (70+13)/2 = 41.5*

**iv. Find the first quartile (Q1) and the third quartile (Q3) of the data.**

*The ﬁrst quartile (corresponding to the 25th percentile) of the data is: 20. The third quartile (corresponding to the 75th percentile) of the data is: 35.*

**2. Using the data for age given in question 1, find out the followings**

1. **Use smoothing by bin means to smooth the data, using a bin**

**depth of 3.**

***Use smoothing by bin*** *means to smooth the above data, using a bin depth of 3. Illustrate your steps. Comment on the eﬀect of this technique for the given data. The following steps are required to smooth the above data using smoothing by bin means with a bin depth of 3.*

*• Step 1: Sort the data. (This step is not required here as the data are already sorted.)*

*• Step 2: Partition the data into equidepth bins of depth 3.*

*Bin 1: 13, 15, 16 Bin 2: 16, 19, 20 Bin 3: 20, 21, 22 Bin 4: 22, 25, 25 Bin 5: 25, 25, 30 Bin 6: 33, 33, 35 Bin 7: 35, 35, 35 Bin 8: 36, 40, 45 Bin 9: 46, 52, 70*

*• Step 3: Calculate the arithmetic mean of each bin.*

*• Step 4: Replace each of the values in each bin by the arithmetic mean calculated for the bin.*

*Bin 1: 142/3, 142/3, 142/3 Bin 2: 181/3, 181/3, 181/3 Bin 3: 21, 21, 21 Bin 4: 24, 24, 24 Bin 5: 262/3, 262/3, 262/3 Bin 6: 332/3, 332/3, 332/3 Bin 7: 35, 35, 35 Bin 8: 401/3, 401/3, 401/3 Bin 9: 56, 56, 56*

**3. Suppose that a hospital tested the age and body fat data for 18 randomly selected adultswith the following results:**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | age |  | 23 | 23 | 27 | 27 |  | 39 | 41 | 47 |  | 49 | 50 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | % fat |  | 9.5 | 26.5 | 7.8 | 17.8 |  | 31.4 | 25.9 | 27.4 |  | 27.2 | 31.2 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | age |  | 52 | 54 | 54 | 56 |  | 57 | 58 | 58 |  | 60 | 61 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | % fat |  | 34.6 | 42.5 | 28.8 | 33.4 |  | 30.2 | 34.1 | 32.9 |  | 41.2 | 35.7 |
|  |  | |  |  |  |  |  | |  |  |  | |  |
|  | 1. **Calculate the mean, median, and standard deviation of age and %fat.**   *For the variable age the mean is 46.44, the median is 51, and the standard deviation is 12.85. For the variable %fat the mean is 28.78, the median is 30.7, and the standard deviation is 8.99.* | | | | | | | | | | | | |
|  | **ii. Draw the *boxplots* for age and %fat.** | | | | | | |  |  |  |  |  |  |
|  | **iii. Draw a *scatter plot* and a q-q plot based on these two variables.** | | | | | | | | | | | |  |
|  | **iv. Normalize the two variables based on Z-score normalization.** | | | | | | | | | | |  |  |
|  | **v. Calculate the correlation coefficient to check whether these two** | | | | | | | | | | | |  |
|  | **variables are positively or negatively correlated?**  *The Pearson correlation coeﬃcient is 0.82, the variables are positively correlated.* | | | | | | | | |  |  |  |  |

1. **For the attribute 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.**
   1. **Use min-max normalization to transform the value 35 for age onto the range[0.0, 1.0].**

*Using the corresponding equation with minA = 13, maxA = 70, new minA = 0,new maxA = 1.0, then v = 35 is transformed to v′ = 0.39.*

* 1. **Use z-score normalization to transform the value 35 for age, where the standard deviation of age is 12.94 years.**

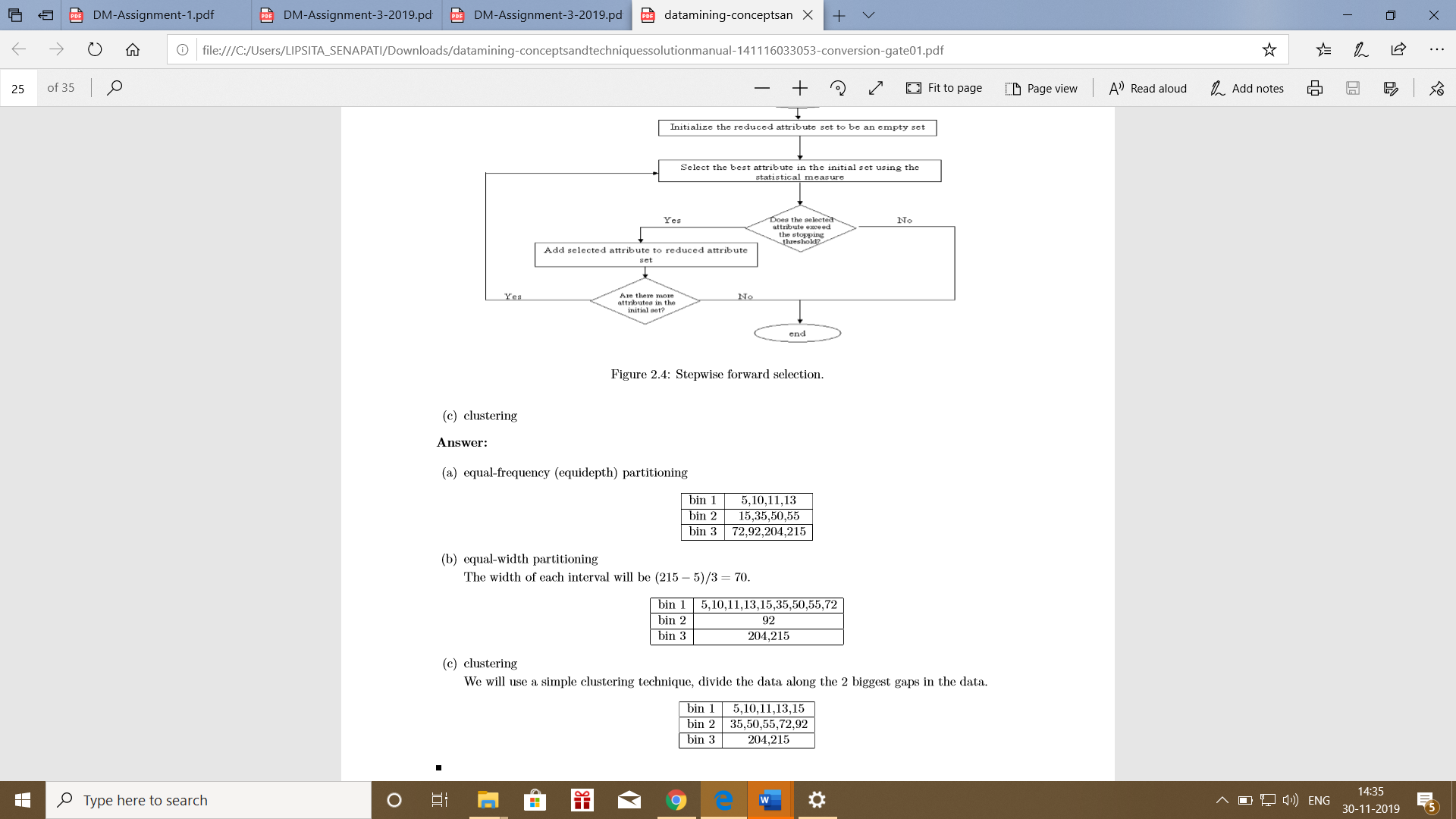
*Using the corresponding equation where A = 809/27 = 29.96 and σA = 12.94, then v = 35 is transformed to v′ = 0.39.*

* 1. **Use normalization by decimal scaling to transform the value 35 for age.**

*Using the corresponding equation where j = 2, v = 35 is transformed to v′ = 0.35.*

1. Suppose a group of 12 sales price records has been sorted as follows 5,10,11,13,15,35,50,55,72,92,204,215

Partition them into three bins by each of the following methods.

1. Equal-frequency(equidepth) partitioning.
2. Equal-width partitioning.
3. Clustering