**BAHRIA UNIVERSITY KARACHI**

**CAMPUS**



**DATA MINING**

**ASSIGNMENT NO: 04**

*Submitted to: Ma’am Anaum Hamid*

*Submitted by: Saba Dildar*

*Enrollment No: 02-134171-126*

*Semester-Section: 7 – A*

*Program: BS(CS)*

*Due Date:31/5/2020*

**QUESTION:**

# Apply the below given data preparation operations and techniques on the sample dataset College Plans.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Reg. No. | First Name | Last Name | Age | Gender | Income | avg. hours served | Willingness (to buy own car) | Avg. fuel Expense (%) | Region | Plan for Extra Booking |
| 441021 | Ahmed | Awan | 37 | Male | 46500 | 100 | Not Willing | 63 | North | No |
| 441022 | Umair | Asim | 44 | Male | 39600 | 121 | Not Willing | 45 | F9 | No |
| 441023 | Kim | Aber | 40 | Male | 69740 | 102 | Willing | 73 | Gulberg | Yes |
| 441024 | Sana | Arif |  | Female | 40400 | 129 | Not Willing | 65 | F9 | Nil |
| 441025 | Dawood | Aslam | 38 | M | 7300 | 86 | Not Willing | 85 | DHA | no |
| 441026 | Faria | Adams | 260 | Female | 17600 | 105 | Not Willing | 87 | North | no |
| 441027 | Margaret | Smith | 40 | Male | 33500 | 1100000 | Not | 85 | North |  |
| 441028 | Arsalan | Waqas | 40 | Male | 48100 | 102 | Not Willing | 95 | North | yes |
| 441029 | Naheed | Ullah | -38 | 1 | 33300 | 79 | Not Willing | 82 | DHA | no |
| 441030 | Waqar | Ahmed | 40 | Male | 80630 | 120 | Willing | 43 | Gulberg | yes |

**SOLUTION**

* **Removing unimportant attributes:**
* After analyzing the given data set, some irrelevant attributes were removed which are:

1. Reg. No.
2. FirstName
3. LastName

* After removing the dataset looked like this:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Age | Gender | Income | avg. hours served | Willingness (to buy own car) | Avg. fuel Expense (%) | Region | Plan for Extra Booking |
| 37 | Male | 46500 | 100 | Not Willing | 63 | North | No |
| 44 | Male | 39600 | 121 | Not Willing | 45 | F9 | No |
| 40 | Male | 69740 | 102 | Willing | 73 | Gulberg | Yes |
|  | Female | 40400 | 129 | Not Willing | 65 | F9 | Nil |
| 38 | M | 7300 | 86 | Not Willing | 85 | DHA | no |
| 260 | Female | 17600 | 105 | Not Willing | 87 | North | no |
| 40 | Male | 33500 | 1100000 | Not | 85 | North |  |
| 40 | Male | 48100 | 102 | Not Willing | 95 | North | yes |
| -38 | 1 | 33300 | 79 | Not Willing | 82 | DHA | no |
| 40 | Male | 80630 | 120 | Willing | 43 | Gulberg | yes |

* **Filling in missing values:**
* After analyzing the given data set, some attributes were having missing values which were:

1. Age (cell no: 4)
2. Plan for Extra Booking (cell no: 6)
3. ***For “Age” attribute:***

* We will take the average of all the values in the age column excluding the garbage or impossible values.

Average (Age) = 37 + 44 + 40 + 38+40 + 40+ 40

7

Average (Age) = 39.85**≈ 40**

|  |
| --- |
| Age |
| 37 |
| 44 |
| 40 |
| 40 |
| 38 |
| 260 |
| 40 |
| 40 |
| -38 |
| 40 |

* The column after modification looks like:

1. ***For “Gender” attribute:***

* In this column there is a cell containing “M” so we will modify that cell and write “Male” in that cell as the “M” is indicating this value.
* The column after modification looks like:

|  |
| --- |
| Gender |
| Male |
| Male |
| Male |
| Female |
| Male |
| Female |
| Male |
| Male |
| 1 |
| Male |

1. ***For “Willingness (to buy own car)” attribute:***

* In this column there is a cell containing “Not” so we will modify that cell and write “Not Encouraged” in that cell as the “Not” is indicating this value.
* The column after modification looks like:

|  |
| --- |
| Willingness (to buy own car) |
| Not Willing |
| Not Willing |
| Willing |
| Not Willing |
| Not Willing |
| Not Willing |
| Not willing |
| Not Willing |
| Not Willing |
| Willing |

1. ***For “Plan for Extra Booking” attribute:***

* In this column there are two cells (i.e. cell no 6) which have missing values so for cell no: 6we will modify it to “Yes “.
* As pattern show the people who are male and their age is 40 they plan for booking
* The column after modification looks like:
* After filling the missing values in the above-mentioned columns and cells the table looks like:

|  |
| --- |
| Plan for Extra Booking |
| No |
| No |
| Yes |
| Nil |
| no |
| no |
| yes |
| yes |
| no |
| yes |

* **Remove Noisy Data:**
* After analyzing the given data set, some attributes were found to have noisy data which are:

1. Age (cell no: 5and 9)
2. Gender (cell no: 9)
3. avg. hours served(cell no: 7)
4. ***For “Age” attribute:***

* We’ll sort the data in this column in ascending order.

|  |  |
| --- | --- |
| Age | Age(Ascending order) |
| 37 | -38 |
| 44 | 37 |
| 40 | 38 |
| 40 | 40 |
| 38 | 40 |
| 260 | 40 |
| 40 | 40 |
| 40 | 40 |
| -38 | 44 |
| 40 | 260 |

* Then we perform binning by depth.

Bin 1 = -38, 37, 38 , 40, 40

Bin 2 = 40, 40, 40, 44, 260

* Now we’ll do smoothening by bin median.

As we know that the no. of elements (n = 5) is odd so,

Median = ( n + 1 ) th term

2

Median (Bin1) = 38

Median (Bin2) = 40

* We’ll now replace the noisy data with its respective bin median. The column looks like:

|  |
| --- |
| Age |
| 37 |
| 44 |
| 40 |
| 40 |
| 38 |
| 40 |
| 40 |
| 40 |
| 38 |
| 40 |

1. ***For “Gender” attribute:***

* We’ll replace the “1” with “Male” as the it is the most frequently occurring value.

|  |
| --- |
| Gender |
| Male |
| Male |
| Male |
| Female |
| Male |
| Female |
| Male |
| Male |
| Male |
| Male |

* The column after removing noisy data looks like:

1. ***For “avg. hours served” attribute:***
2. We’ll sort the data in this column in ascending order.

|  |  |
| --- | --- |
| avg. hours served | Age(Ascending order) |
| 100 | 79 |
| 121 | 86 |
| 102 | 100 |
| 129 | 102 |
| 86 | 102 |
| 105 | 105 |
| 1100000 | 120 |
| 102 | 121 |
| 79 | 129 |
| 120 | 110000 |

* Then we perform binning by depth.

Bin 1 = 79,86,100,102,102

Bin 2 = 105,120,121,129, 1100000

Now we’ll do smoothening by bin median.

As we know that the no. of elements (n = 5) is odd so,

Median = ( n + 1 ) th term

2

Median (Bin1) = 100

Median (Bin2) = 121

* We’ll now replace the noisy data with its respective bin median. The column looks like

|  |
| --- |
| avg. hours served |
| 100 |
| 121 |
| 102 |
| 129 |
| 86 |
| 105 |
| 121 |
| 102 |
| 79 |
| 120 |

* **Min-Max Normalization:**
* We need to apply min-max normalization with range [1-4], to “**Avg, Fuel Expense**” attribute for which we’ll use the below given formula:

V’ =

V - minA

maxA-minA

(new\_maxA–new\_minA) + new\_minA

Where:

maxA= 95 ; minA= 43 ; new\_maxA= 4 ; new\_minA= 1

* After performing min-max normalization the table looks like:

|  |  |  |
| --- | --- | --- |
| Avg. fuel Expense (%) | Calculation | Expected  Avg. fuel Expense (%) |
| 63 | V’=63-43(4-1)+1  95-43 | 2.153 |
| 45 | V’=45-43 (4-1)+1  95-43 | 1.115 |
| 73 | V’=73-43(4-1)+1  95-43 | 2.730 |
| 65 | V’=65-43(4-1)+1  95-43 | 2.269 |
| 85 | V’=85-43(4-1)+1  95-43 | 3.423 |
| 87 | V’=87-43(4-1)+1  95-43 | 3.538 |
| 85 | V’=85-43(4-1)+1  95-43 | 3.4230 |
| 95 | V’=95-43(4-1)+1  95-43 | 4 |
| 82 | V’=82-43(4-1)+1  95-43 | 3.25 |
| 43 | V’=43-43(4-1)+1  95-43 | 1 |

1. **Equal Depth and Smoothing by Bin Boundaries Method**

Avg. Hours Served attribute. Bin number: 4. divides the range into N intervals, each containing

Approximately same number of samples

1. **Step1**
2. ***For “avg. hours served” attribute:***
3. We’ll sort the data in this column in ascending order

|  |  |
| --- | --- |
| avg. hours served | Age(Ascending order) |
| 100 | 79 |
| 121 | 86 |
| 102 | 100 |
| 129 | 102 |
| 86 | 102 |
| 105 | 105 |
| 1100000 | 120 |
| 102 | 121 |
| 79 | 129 |
| 120 | 110000 |

**FORMULA**

**Size= Total No of Data Object/Bin No**

10/4=2.5

Bin1: 79, 86,100,102,102

Bin2: 105,120,121,129,110000

By Applying Bin Boundaries Method:

Replace Element by nearest one

**Bin 1:79, 79,102,102,102**

**Bin 1: 105,105,105, 105,110000**

|  |
| --- |
| avg. hours served |
| 79 |
| 79 |
| 102 |
| 102 |
| 102 |
| 105 |
| 105 |
| 105 |
| 105 |
| 110000 |

* **Fixed k-Interval Discretization:**
* We need to perform fixed k-interval discretization on “Income” attribute, where k = 5.
* Given that:

K = 5 ; total elements = 10 ; categories = Very Low, Low, Medium, High, Very High

* So,

No. of elements in each category = total elements

k

|  |  |
| --- | --- |
| tIncome | Category |
| 7300 | Very Low |
| 17600 | Very Low |
| 33300 | Low |
| 33500 | Low |
| 39600 | Medium |
| 40400 | Medium |
| 46500 | High |
| 48100 | High |
| 69740 | Very High |
| 80630 | Very High |

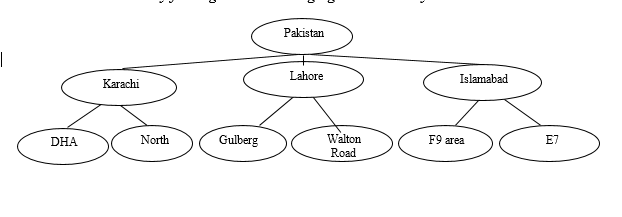
No. of elements in each category = 10

No. of elements in each category = 2

* First, we’ll sort the elements in ascending order.

|  |  |
| --- | --- |
| Income | Income (Ascending order) |
| 46500 | 7300 |
| 39600 | 17600 |
| 69740 | 33300 |
| 40400 | 33500 |
| 7300 | 39600 |
| 17600 | 40400 |
| 33500 | 46500 |
| 48100 | 48100 |
| 33300 | 69740 |
| 80630 | 80630 |

Now we’ll perform categorization

* **Concept Hierarchy Generation**

Finally, the table looks like:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Age | Gender | Income | avg. hours served | Willingness (to buy own car) | Avg. fuel Expense (%) | Region | Plan for Extra Booking |
| 37 | Male | High | 79 | Not Willing | 2.153 | North | No |
| 44 | Male | Medium | 79 | Not Willing | 1.115 | F9 | No |
| 40 | Male | Very High | 102 | Willing | 2.730 | Gulberg | Yes |
| 40 | Female | Medium | 102 | Not Willing | 2.269 | F9 | Nil |
| 38 | Male | Very Low | 102 | Not Willing | 3.423 | DHA | no |
| 40 | Female | Very Low | 105 | Not Willing | 3.538 | North | no |
| 40 | Male | Low | 105 | Not Willing | 3.423 | North | yes |
| 40 | Male | High | 105 | Not Willing | 4 | North | yes |
| 38 | Male | Low | 105 | Not Willing | 3.25 | DHA | no |
| 40 | Male | Very High | 110000 | Willing | 1 | Gulberg | yes |

**Question no. 02 [Basic Statistical Analysis] [6 Marks]**

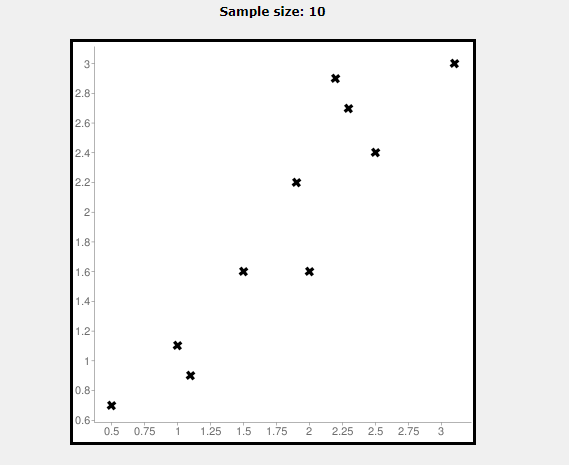
**Consider the following bivariate sample dataset with sample size=10. All steps must be clearly done.**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X | 2.5 | 0.5 | 2.2 | 1.9 | 3.1 | 2.3 | 2.0 | 1.0 | 1.5 | 1.1 |
| Y | **2.4** | **0.7** | **2.9** | **2.2** | **3.0** | **2.7** | **1.6** | **1.1** | **1.6** | **0.9** |

1. Draw the scatter-plots for given Data and Adjusted Data.

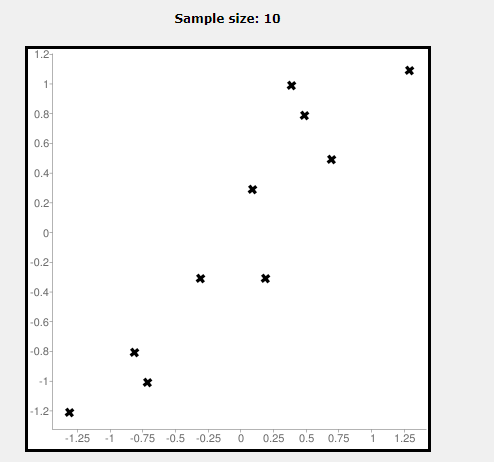
|  |  |
| --- | --- |
| X | Y |
| 2.5 | **2.4** |
| 0.5 | **0.7** |
| 2.2 | **2.9** |
| 1.9 | **2.2** |
| 3.1 | **3.0** |
| 2.3 | **2.7** |
| 2.0 | **1.6** |
| 1.0 | **1.1** |
| 1.5 | **1.6** |
| 1.1 | **0.9** |

**Scatter-plots for given Data**



**Scatter-plotsfor given Adjusted Data**

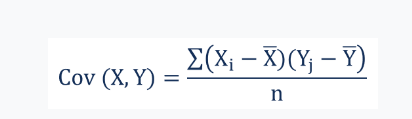
|  |  |  |  |
| --- | --- | --- | --- |
| X | Y | X-x̅ | Y -ȳ |
| 2.5 | **2.4** | **2.5-1.81=0.69** | **2.4-1.91=0.49** |
| 0.5 | **0.7** | **0.5-1.81=-1.31** | **0.7-1.91=-1.21** |
| 2.2 | **2.9** | **2.2-1.81=0.39** | **2.9-1.91=0.99** |
| 1.9 | **2.2** | **1.9-1.81=0.09** | **2.2-1.91=0.29** |
| 3.1 | **3.0** | **3.1-1.81=1.29** | **3.0-1.91=1.09** |
| 2.3 | **2.7** | **2.3-1.81=0.49** | **2.7-1.91=0.79** |
| 2.0 | **1.6** | **2.0-1.81=0.19** | **1.6-1.91=-0.31** |
| 1.0 | **1.1** | **1.0-1.81=-0.81** | **1.1-1.91=-0.81** |
| 1.5 | **1.6** | **1.5-1.81=-0.31** | **1.6-1.91=-0.31** |
| 1.1 | **0.9** | **1.1-1.81=-0.71** | **0.9-1.91=--1.01** |
| Σx=1.81 | **Σx=1.91** |  |  |



1. **Calculate the covariance matrix.**

|  |  |
| --- | --- |
| X | Y |
| 2.5 | **2.4** |
| 0.5 | **0.7** |
| 2.2 | **2.9** |
| 1.9 | **2.2** |
| 3.1 | **3.0** |
| 2.3 | **2.7** |
| 2.0 | **1.6** |
| 1.0 | **1.1** |
| 1.5 | **1.6** |
| 1.1 | **0.9** |

**Formula**



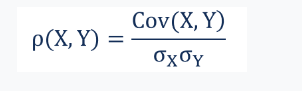
**Xi**– the values of the X-variable

**Yj**– the values of the Y-variable

**X̄**– the mean (average) of the X-variable

**Ȳ**– the mean (average) of the Y-variable

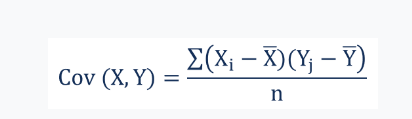
**n** – The number of data points



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| X | Y | (X-x̅ )2 | (Y -ȳ )2 | ( X-x̅ )(Y -ȳ) |
| 2.5 | **2.4** | **=0.476** | **=0.2401** | **0.69\*0.49=0.3381** |
| 0.5 | **0.7** | **=1.7161** | **=1.4641** | **-1.31\*1.21=-1.5851** |
| 2.2 | **2.9** | **=0.1521** | **=0.9801** | **0.39\*0.99=0.3861** |
| 1.9 | **2.2** | **=0.008** | **=0.0841** | **0.09\*0.29=0.0261** |
| 3.1 | **3.0** | **=1.6641** | **=1.881** | **1.29\*1.09=1.4061** |
| 2.3 | **2.7** | **=0.2401** | **=0.6241** | **0.49\*0.79=0.3381** |
| 2.0 | **1.6** | **=0.0361** | **=0.0961** | **0.19\*0.31=-0.0589** |
| 1.0 | **1.1** | **=-0.6561** | **=0.6561** | **-0.81\*0.81=-0.6561** |
| 1.5 | **1.6** | **=0.0961** | **=0.0961** | **-0.31\*-0.31=0.0961** |
| 1.1 | **0.9** | **=0.5041** | **=1.0201** | ­­­­­­­  **-0.71\*1.01=-0.7171** |
| Σx=1.81 | **Σx=1.91** | **Σ (X-x̅)(X-x̅)= 5.5490** | **Σ (Y -ȳ ) Σ (Y -ȳ )=6.449** | **Σ( X-x̅ )(Y -ȳ) =5.5390** | |

**Covariance Matrix**

**C=**



**=5.5490=0.6165=6.449 =0.7165**

**9 9**

**=5.5490=0.6164=5.5390 =0.6154**

**9 9**

So here the covariance matrix

**C=**

**C=**

**4) Calculate the Eigen values and eigenvectors of the covariance matrix**

**Formula**

| A − λI | = 0

**-**λ  **=0**

**=0**

**()()-()()=0**

**0.4417-0.6165**-0.7165

**0.4417-0.6165**-0.7165

85+0.4417=0

**By solving Equation we got the values of**

**For**