**CST8390 - Lab 1 Answers**

**Student Name: Wei-Ting Yang**

**Student Number: 040852642**

During demo time, you should be able to answer questions on the process that you did in this lab.

When you take screenshots, make sure that you are getting a shot of the full Weka.

**Weka Installation guide**

10. Fill the following table from the installation guide.

|  |  |
| --- | --- |
| **Flower Type** | **Count** |
| Iris-setosa | 50 |
| Iris-versicolor | 50 |
| Iris-virginica | 50 |

11.Fill the following table from the installation guide.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Attribute** | **Minimum** | **Maximum** | **Mean** | **StdDev** |
| sepallength | 4.3 | 7.9 | 5.843 | 0.828 |
| sepalwidth | 2 | 4.4 | 3.054 | 0.434 |
| petallength | 1 | 6.9 | 3.759 | 1.764 |
| petalwidth | 0.1 | 2.5 | 1.199 | 0.763 |

**Data Preparation and Cleaning**

1. Which are the four important attributes that are relevant to analyse this dataset?

Country, Branch, Currency and Salary.

1. For the nominal attributes of the above question, fill in the following table:

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute Name: Country** | | **Attribute Name: Branch** | |
| **Label** | **Count** | **Label** | **Count** |
| **China** | **40** | **1** | **40** |
| **U.S.A.** | **39** | **2** | **39** |
| **Germany** | **38** | **3** | **40** |
| **Mexico** | **37** | **4** | **35** |
| **Japan** | **1** | **6** | **1** |
| **Attribute Name: Currency** | |
| **Label** | **Count** |
| **CHY** | **41** |
| **USD** | **39** |
| **EUR** | **38** |
| **MXD** | **36** |
| **INR** | **1** |

1. Analyze your data to see any anomalies. List the identified anomalies below. Write why you think those records are anomalies in the following format **(Add at least 8 anomalies)**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | first\_name | last\_name | email | Address | Country | Branch | Currency | Salary | Reason |
| 1 | Chelsie | Mulles | cmulles3x@amazon.co.uk | 8 Talmadge Circle | Mexico | 6 | MXD | 36000 | The branch is not matching |
| 2 | Della | Hodgin | bhodgin2g@msu.edu | 60 Bellgrove Court | Japan | 3 | CHY | 600000 | Currency and country are not matching |
| 3 | Nelson | McRinn | nmcrinn3p@economist.com | 56053 Buell Terrace | Mexico | 2 | MXD | 19999 | Branch and country are not matching |
| 4 | Maressa | Commucci | mcommucci3e@techcrunch.com | 3 Prairieview Alley | China | 1 | MXD | 40 | Salary is too low |
| 5 | Edward | Skells | eskells35@spotify.com | 54486 Carberry Park | Mexico | 4 | CHY | 150000 | Country and currency are not matching. |
| 6 | Burton | Dudden | bdudden39@japanpost.jp | 050 Nova Court | Mexico | 4 | INR | 45999 | Currency and country are not matching |
| 7 | Darcy | Addie | daddie1k@jalbum.net | 836 Marquette Pass | Germany | 2 | EUR | 60500999 | Salary is too high |
| 8 | Misha | Tunna | mtunnaj@dailymotion.com | 15 Kipling Drive | U.S.A. | 1 | USD | 32000999 | Salary is too high |

**(Add more rows as required. You will be able to find at least 8 anomalies)**

**Remove Duplicates**

1. How many instances do you have now? \_\_\_\_\_\_155\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

13.

* 1. How many instances do you have now? \_\_\_\_150\_\_\_\_\_\_\_\_\_\_\_\_\_
  2. How many duplicates (how many got removed): \_\_\_\_\_\_5\_\_\_\_\_\_\_\_\_\_

14. Take a screenshot and paste it here.

A screenshot of a computer

Description automatically generated

**Nominal to Binary**

16. How many nominal attributes do you have?

3

1. Take a screenshot and paste it in lab1 document.

A screenshot of a computer

Description automatically generated

1. Take a screenshot of the file while it is opened in Notepad++. Header should be visible.

**A screenshot of a computer

Description automatically generated**

**A screen shot of a computer

Description automatically generated**

**Rename Nominal Value**

24. Take a screenshot and paste it in lab1 document

A screenshot of a computer

Description automatically generated