**CST8390 - Lab 1**

**Data exploration and integration with Weka - Iris dataset**

1. Fill this table:

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

1. Fill this table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **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 |

**Part 2**

1. Which are the four important attributes that are relevant for data analysis?

* Country
* Branch
* Currency
* Salary

1. For the nominal attributes from Question 19, fill in the following table:

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

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 rows as required):

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Id | first\_name | last\_name | Email | Address | Country | Branch | Currency | Salary | Reason |
| 100135 | Maressa | Commucci | mcommucci3e@techcrunch.com | ‘3 Prairieview Alley’ | Mexico | 4 | MXD | 40 | Salary is only “40” |
| 100169 | Chelsie | Mulles | cmulles3x@amazon.co.uk | ‘8 Tamadge Cirle’ | Mexico | 6 | MXD | 36000 | The only employee in branch “6” |
| 100149 | Darcy | Addie | daddie1k@jalbum.net | '836 Marquette Pass' | Germany | 2 | EUR | 60500999 | Has 8 digits salary |
| 100204 | Misha | Tunna | mtunnaj@dailymotion.com | '15 Kipling Drive' | U.S.A. | 1 | USD | 32000999 | Has 8 digits salary |
| 100233 | Bel | Hodgin | bhodgin2g@msu.edu | '60 Bellgrove Court' | Japan | 3 | CHY | 600000 | The only person from japan |
| 100139 | Burton | Dudden | bdudden39@japanpost.jp | '050 Nova Court' | Mexico | 4 | INR | 45999 | The only person is paid with Indian currency |
| 100240 | Katina | Borel | kborelo@github.io | '629 Hansons Terrace' | U.S.A. | 1 | USD | 68000 | The only person with the email end with “.io” |
| 100116 | Adrea | Bravey | abravey1@1und1.de | '4621 Emmet Junction' | U.S.A. | 1 | USD | 72000 | The only person with the email end with “.de” |

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