**CST8390 - Lab 2**

**Explore CSV file and transform it into ARFF file**

**Due Date:** During week 2 labs

**Introduction**

The goal of this lab is to explore a CSV file and transform it to ARFF format manually (NOT using tool). Then, load the file in Weka and find statistics. Finally, do some initial analysis on the data.

**Steps:**

1. Download EmployeesSalary.csv file from Brightspace.
2. Open EmployeesSalary.csv in excel and explore it.
3. Read <https://www.cs.waikato.ac.nz/ml/weka/arff.html> to find the expectations of an ARFF file.
4. Identify the attributes of the data. Record the attributes and the type of attribute for the data.
5. Closely analyse data. In excel, do the required modifications to match with the requirements for an ARFF file. (Hint: Check the requirements if the data has spaces in it.)
6. Open the file in notepad++. Add the required section headers and corresponding information in the file. Save the file as EmployeesSalary.arff. This involves creating the @relation line, one @attribute line per attribute, and @data to signify the start of data. It is good to add comments at the top of the file describing where you obtained this data set, explanation about your attributes etc. A comment in the ARFF format is started with the percent character % and continues until the end of the line.
7. Open the ARFF file as you did in lab 1 (by selecting ‘Open file’ in the ‘Preprocess tab’.). You may run into errors as you load your ARFF file. If so, check the requirements to troubleshoot your problem.
8. Which are the four important attributes that are relevant for data analysis?

a. Numeric attributes, e.g. @ATTRIBUTE salary 2000

b. Norminal attributes, e.g. @ATTRIBUTE class {Iris-setosa, Iris-versicolor, Iris-virginica}

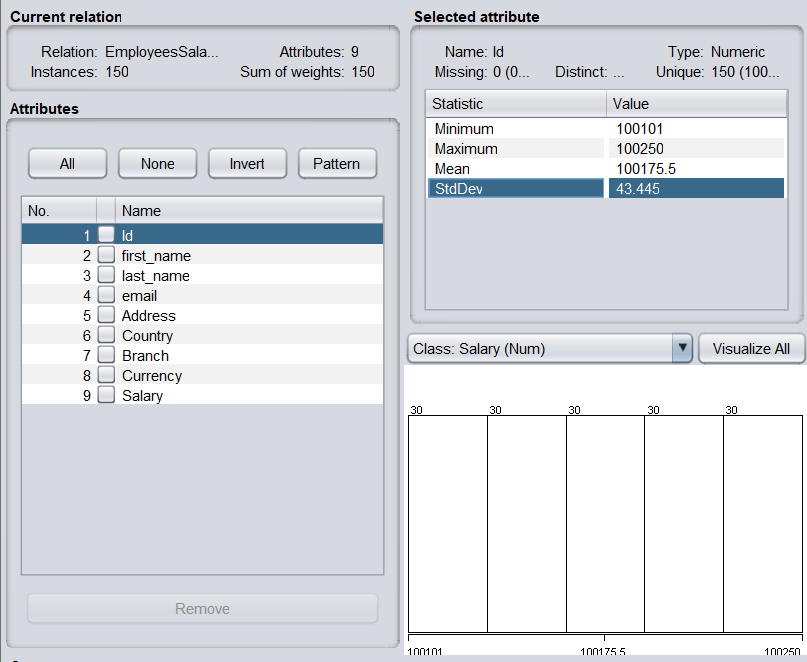
c. String attributes, e.g. @ATTRIBUTE LCC string

d. Date attributes, e.g. @ATTRIBUTE <name> date [<yyyy-MM-dd'T'HH:mm:ss>]

1. For the nominal attributes from Question 8, fill in the following table:

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

1. Analyze your data to see any anomalies. List the identified anomalies below (you will see at least 8).
2. The outlier of Country, there is only 1 “Japan”.
3. The outlier of Branch, there is only 1 “6”.
4. The outlier of Currency, there is only 1 “INR”.
5. The outlier of Salary, there is 1 employee only has “40”.
6. The outlier of Salary, one employee named Darcy Addie has “60500999”.
7. The outlier of Salary, one employee named Darcy Addie has “32000999”.
8. There 36 “China”, but there are 38 “CHY”.
9. There 37 “Mexico”, but there are 35 “MXD”.



**標準差**（又稱標準偏差、均方差，英語：**S**tandard **D**eviation，縮寫**SD**），數學符號**[σ](https://zh.wikipedia.org/wiki/%CE%A3" \o "Σ)**（sigma），在[概率](https://zh.wikipedia.org/wiki/%E6%A6%82%E7%8E%87" \o "概率)[統計](https://zh.wikipedia.org/wiki/%E7%B5%B1%E8%A8%88)中最常使用作為[測量](https://zh.wikipedia.org/wiki/%E6%B8%AC%E9%87%8F)一組數值的[離散程度](https://zh.wikipedia.org/wiki/%E7%A6%BB%E6%95%A3%E7%A8%8B%E5%BA%A6)之用。標準差定義：為[方差](https://zh.wikipedia.org/wiki/%E6%96%B9%E5%B7%AE)開[算术平方根](https://zh.wikipedia.org/wiki/%E7%AE%97%E6%9C%AF%E5%B9%B3%E6%96%B9%E6%A0%B9)，反映组内個體間的離散程度；標準差與[期望值](https://zh.wikipedia.org/wiki/%E6%9C%9F%E6%9C%9B%E5%80%BC)之比為[標準離差率](https://zh.wikipedia.org/wiki/%E6%A0%87%E5%87%86%E7%A6%BB%E5%B7%AE%E7%8E%87)。測量到分佈程度的結果，原則上具有兩種性質：

1. 為非負數值（因為開平方後再做平方根）；
2. 與測量資料具有相同單位（這樣才能比對）。

一個總量的標準差或一個[隨機變量](https://zh.wikipedia.org/wiki/%E9%9A%A8%E6%A9%9F%E8%AE%8A%E9%87%8F)的標準差，及一個[子集合](https://zh.wikipedia.org/wiki/%E5%AD%90%E9%9B%86%E5%90%88)樣品數的標準差之間，有所差別。

簡單來說，標準差是一組數值自[平均值](https://zh.wikipedia.org/wiki/%E5%B9%B3%E5%9D%87%E5%80%BC)分散開來的程度的一種測量觀念。一個較大的標準差，代表大部分的數值和其平均值之間差異較大；一個較小的標準差，代表這些數值較接近平均值。

例如，兩組數的[集合](https://zh.wikipedia.org/wiki/%E9%9B%86%E5%90%88_(%E6%95%B0%E5%AD%A6)){0, 5, 9, 14}和{5, 6, 8, 9}其平均值都是7，但第二個集合具有較小的標準差。

In order to get the credit for this lab:

1. Show the loaded file in Weka
2. Fill in the tables for questions 8 & 9
3. Show the list of anomalies