**CST8390 - Lab 2**

**Data Preparation and Cleaning - Answers**

**Part 1**

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

- Country

- Branch

- Currency

- Salary

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

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

|  |  |
| --- | --- |
| **Attribute Name: Branch ID** | |
| **Label** | **Count** |
| **1** | **38** |
| **2** | **39** |
| **3** | **37** |
| **4** | **35** |
| **6** | **1** |

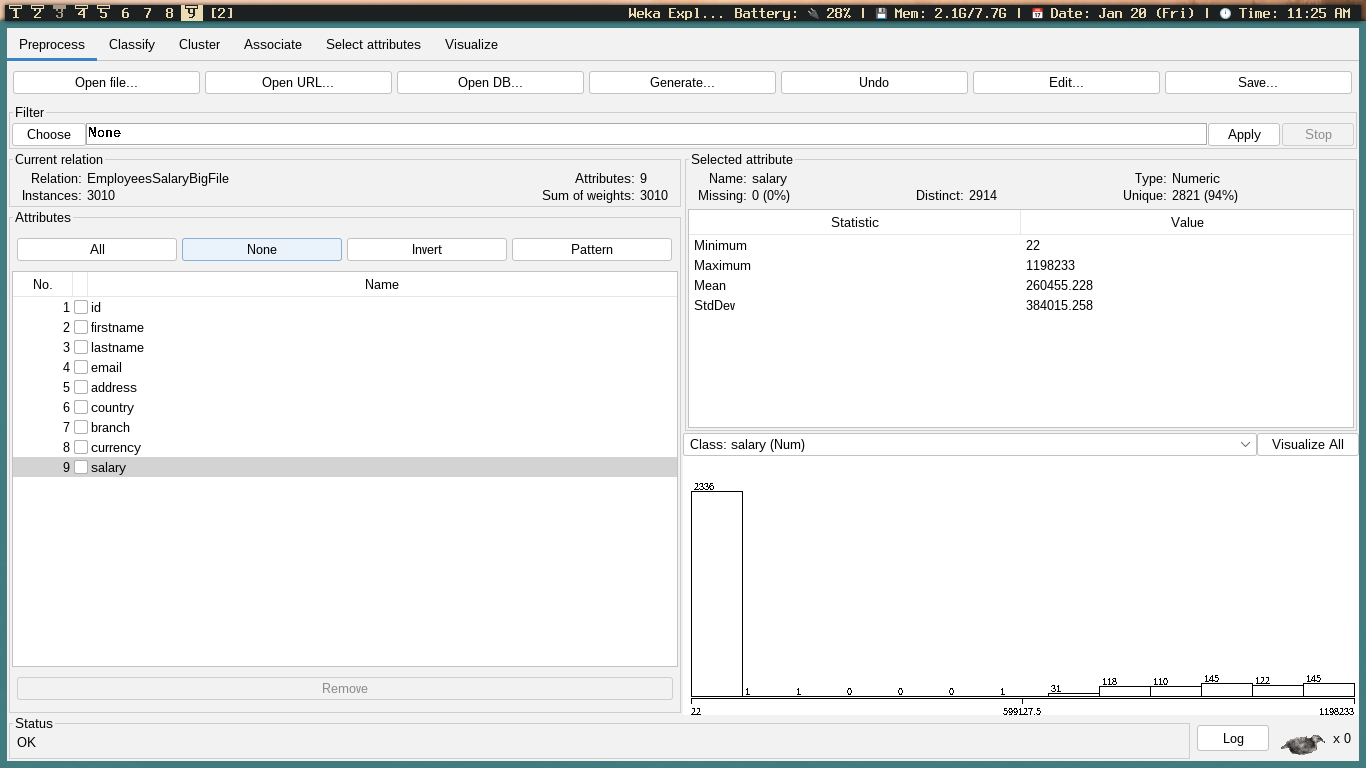
1. Analyze your data to see any anomalies. List the identified anomalies below. Write why you think those records are anomalies in the following format (add rows as required):

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Id | first\_name | last\_name | email | Address | Country | Branch | Currency | Salary | Reason |
| 100149 | Darcy | Addie | daddie1k@jalbum.net | 836 Marquette Pass | Germany | 2 | EUR | 60500999 | Salary is much higher than expected. |
| 100135 | Maressa | Commucci | mcommucci3e@techcrunch.com | 3 Prairieview Alley | Mexico | 4 | MXD | 40 | Makes significantly less money than expected. |
| 100204 | Misha | Tunna | mtunnaj@dailymotion.com | 15 Kipling Drive | U.S.A. | 1 | USD | 32000999 | Salary is much higher than expected. |
| 100233 | Bel | Hodgin | bhodgin2g@msu | 60 Bellgrove Court | Japan | 3 | CHY | 600000 | Only person in the data that lives in Japan |
| 100139 | Burton | Dudden | bdudden39@japanpost.jp | 050 Nova Court | Mexico | 4 | INR | 45999 | Only person that is paid using the INR currency. |
| 100132 | Rickie | Lates Ken | rlatesken@ameblo.jp | 823 Mockingbird Pass | Mexico | 4 | MXD | 60000 | Only person with a two-part last name. |
| 10018 | Elsworth | Skells | Eskells35@spotify.com | 54486 Carberry Park | Mexico | 4 | CHY | 150000 | Only person with country CHY currency. |
| 100247 | Nelson | McRinn | Nmcrinn3p@economist.com | 56053 Buell Terrace | Mexico | 2 | MXD | 19999 | Only person from Mexico with branch ID 2. |

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

**Part 2**

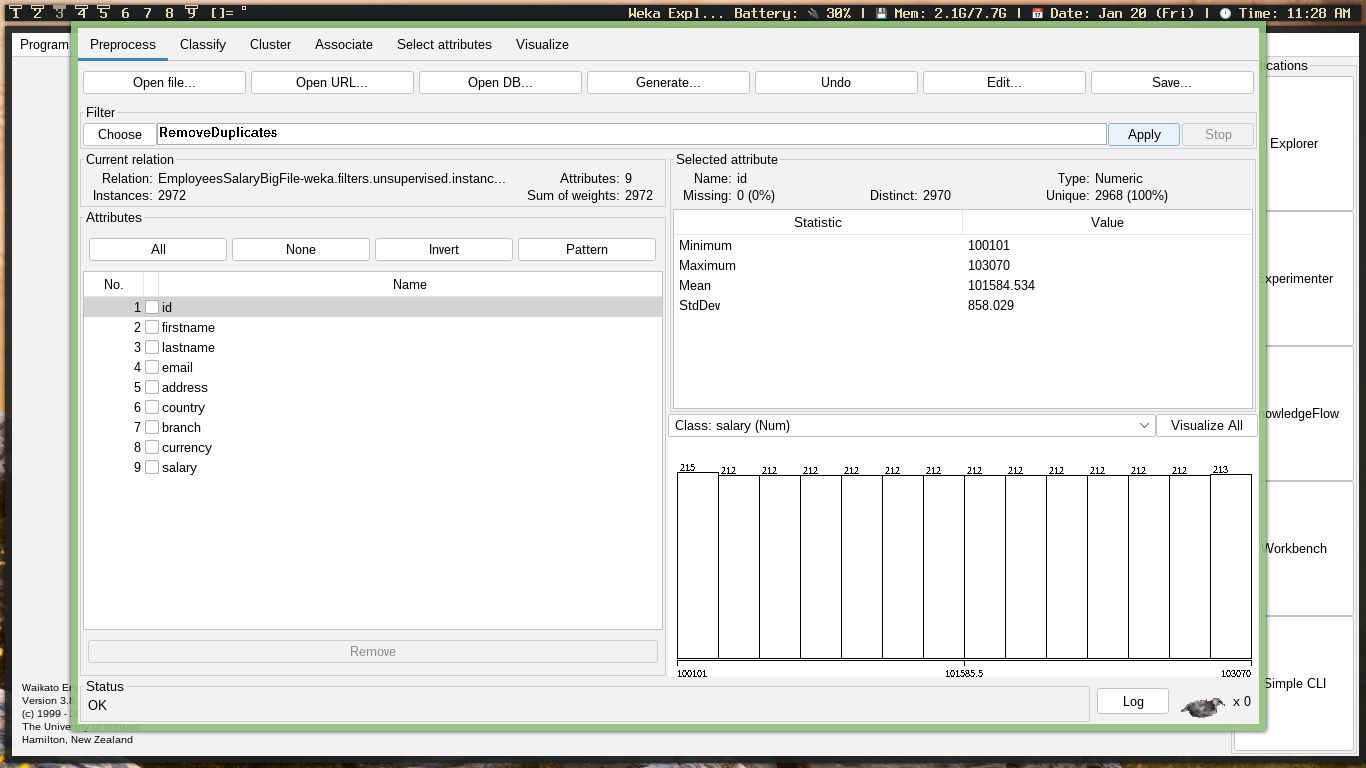
1. How many instances do you have now? 3010
2. Take a screenshot and save it in lab2.doc.



13.

* 1. How many instances do you have now? 2972
  2. How many duplicates (how many got removed): 38

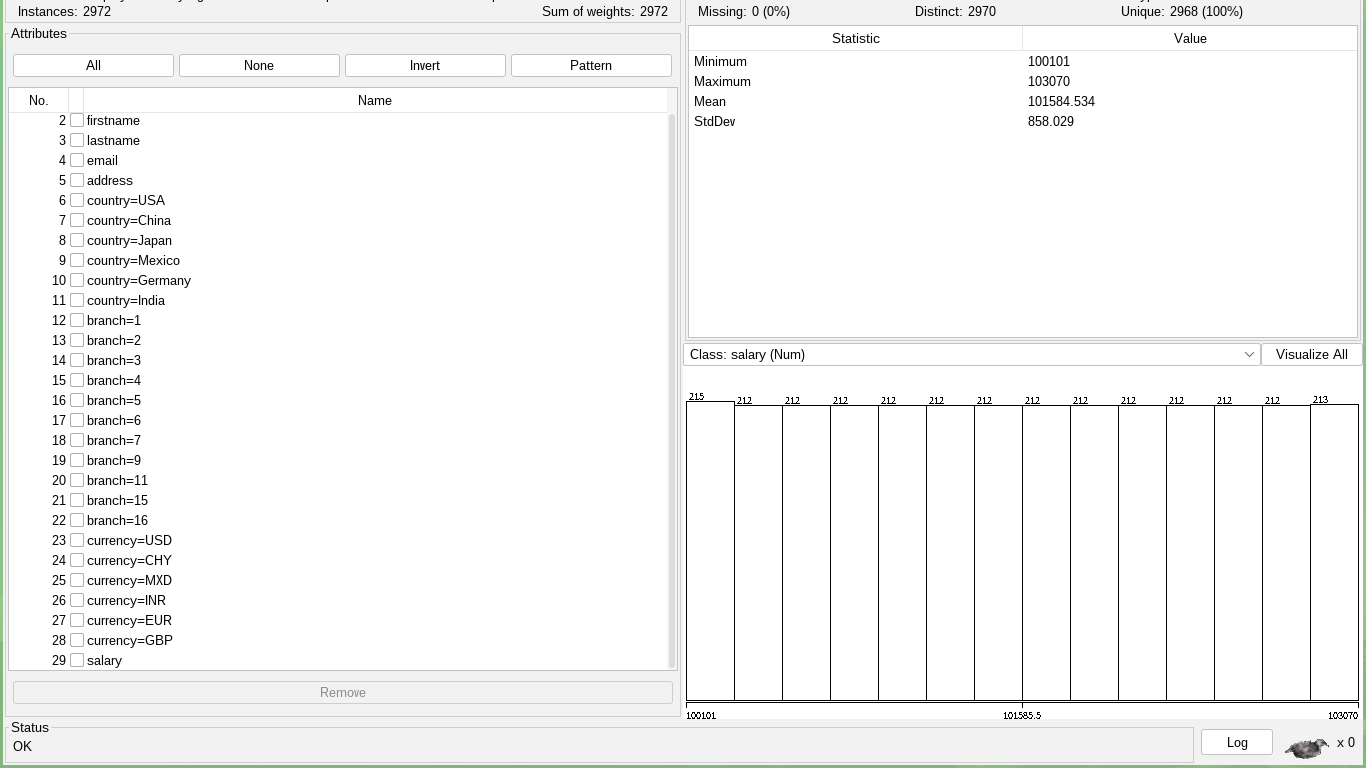
14. Take a screenshot and paste it here.



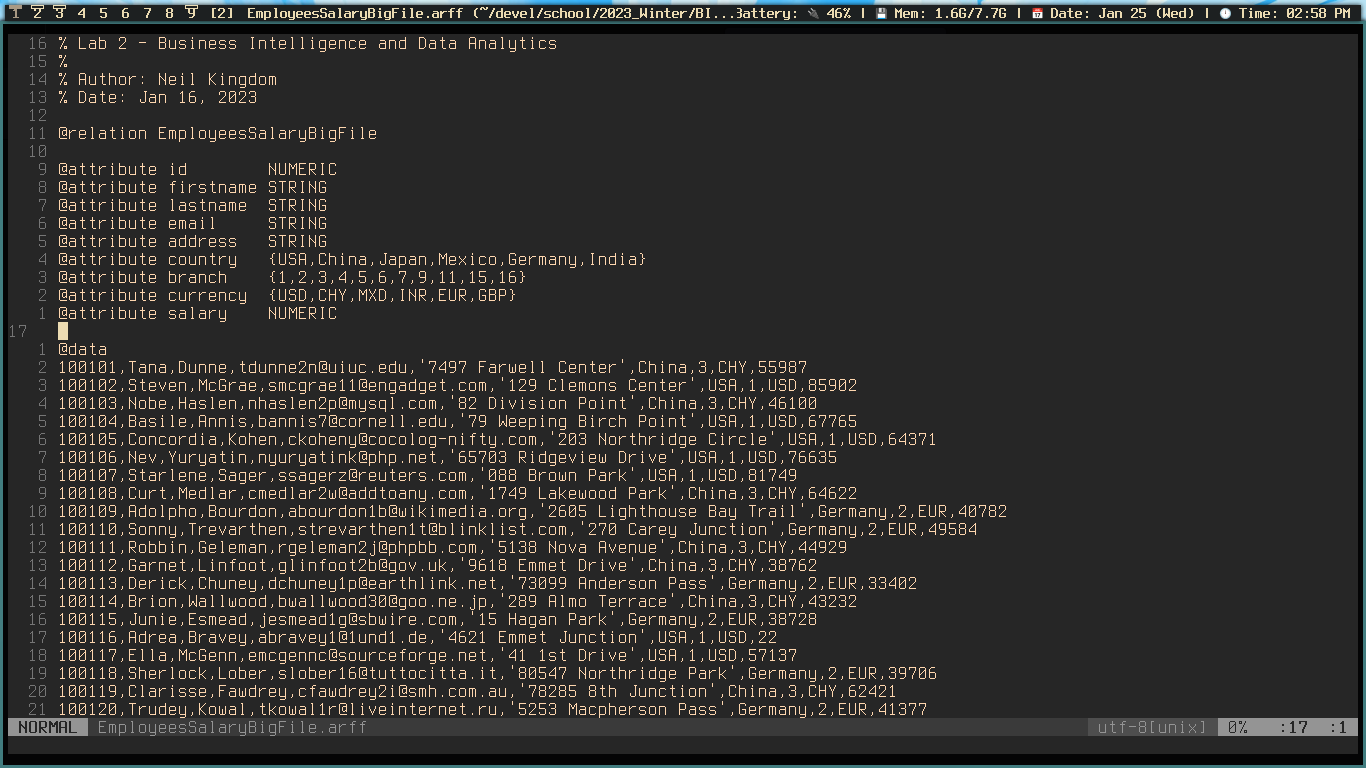
**Nominal to Binary**

16. How many nominal attributes do you have? 3

1. Take a screenshot and paste it in lab2 document.



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



**In order to get the credit for this lab:**

1. Show the EmployeesSalary file in Weka during demo.
2. Show EmployeesSalaryBigFileNoDupBinary.arff in Weka
3. Show the answers of Q8,9,10 of part 1 and Q8, 13, 14, 16, 18 & 21 of part 2 in lab2\_Answers.doc
4. Upload lab2\_Answers.doc in Brightspace before the submission due date.

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 the shot of the full Weka.