**CST8390 - Lab 2**

**Data Preparation and Cleaning - Answers**

**Part 1**

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

* Salary
* Country
* Branch
* Currency

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

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

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Id | first\_name | last\_name | email | Address | Country | Branch | Currency | Salary | Reason |
| 100259 | Ariella | Bayfield | abyfield8@google.ca | 06 Kensington Terrace | China | 3 | GBP | 36224.0 | Ariella living in China but earns  GBP |
| 100187 | Elsworth | Skells | eskells35@spotify.com | 54486 Carberry Park | Mexico | 4 | CHY | 17542.0 | Elsworth living in  Mexico but earns CHY |
| 100247.0 | Nelson | McRinn | nmcrinn3p@economist.com | 56053 Buell Terrace | Mexico | 2 | MXD | 31673.0 | brunch number should be 4 |
| 100233.0 | Bel | Hodgin | bhodgin2g@msu.edu | 60 Bellgrove Court | Japan | 3 | CHY | 600000.0 | Currency should be INR |
| 100169.0 | Chelsie | Mulles | cmulles3x@amazon.co.uk | 8 Talmadge Circle | Mexico | 6 | MXD | 34727.0 | Brunch number should be 4 as usual |

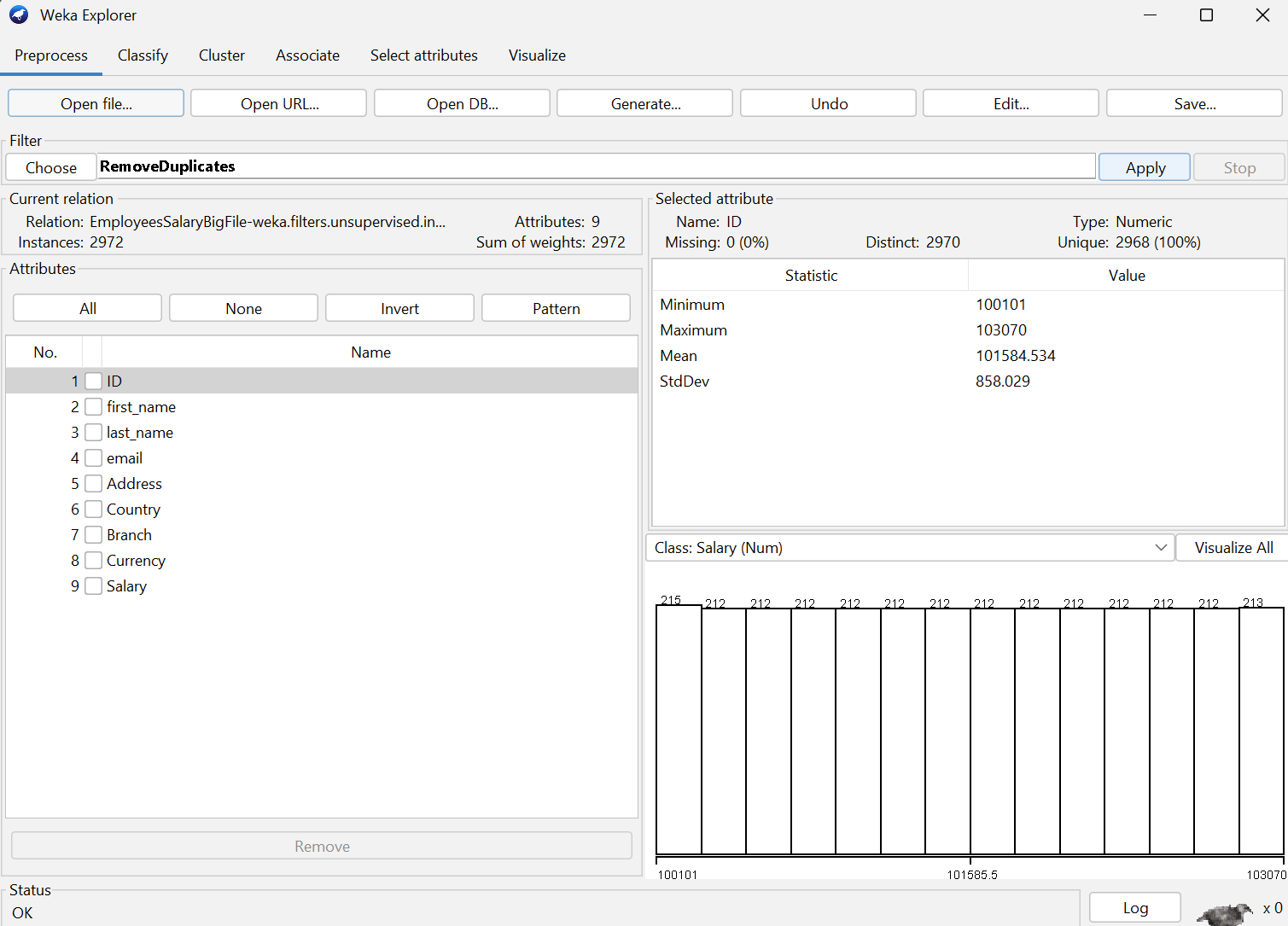
**Part 2**

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.



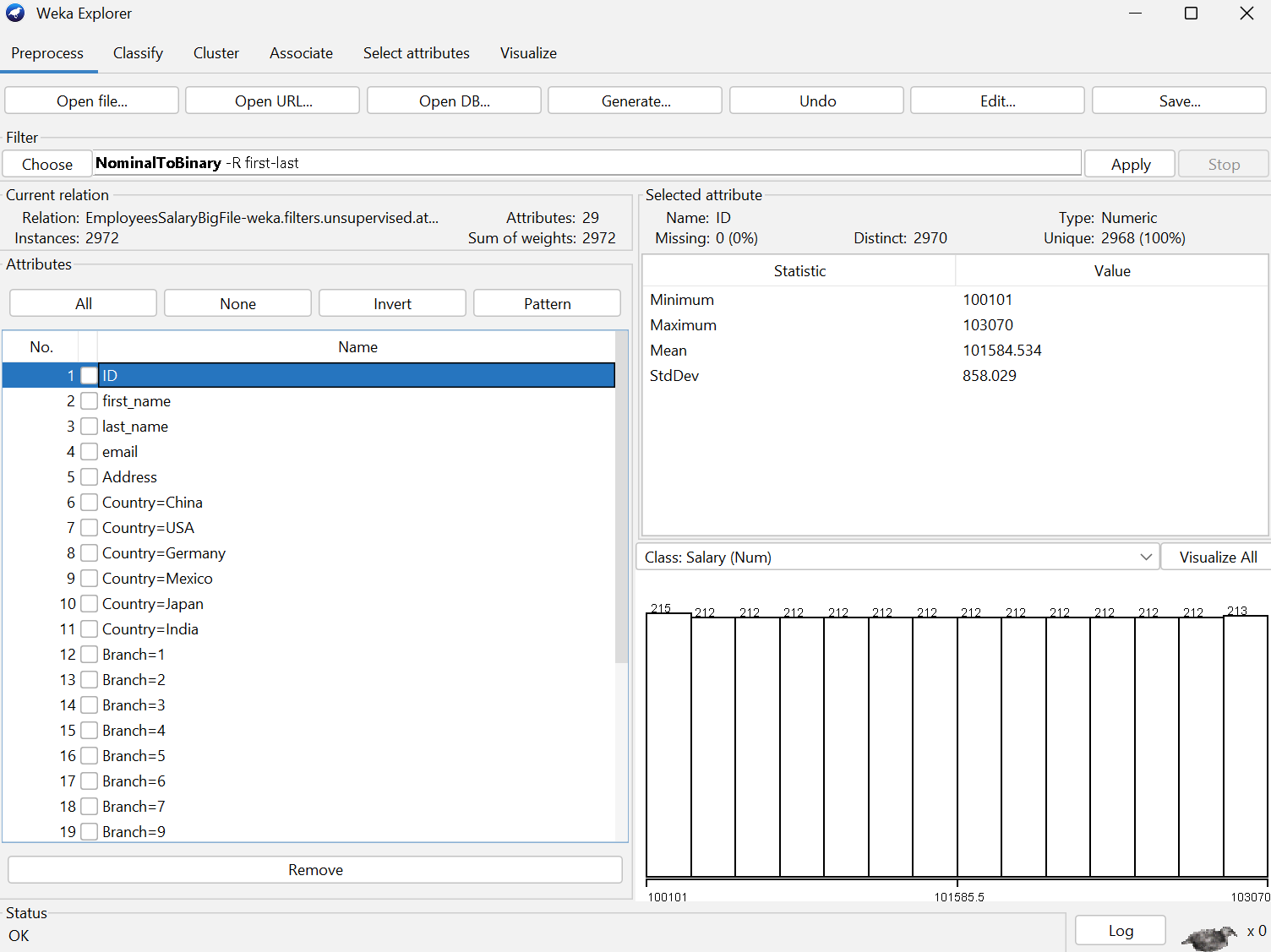
**Nominal to Binary**

16. How many nominal attributes do you have?

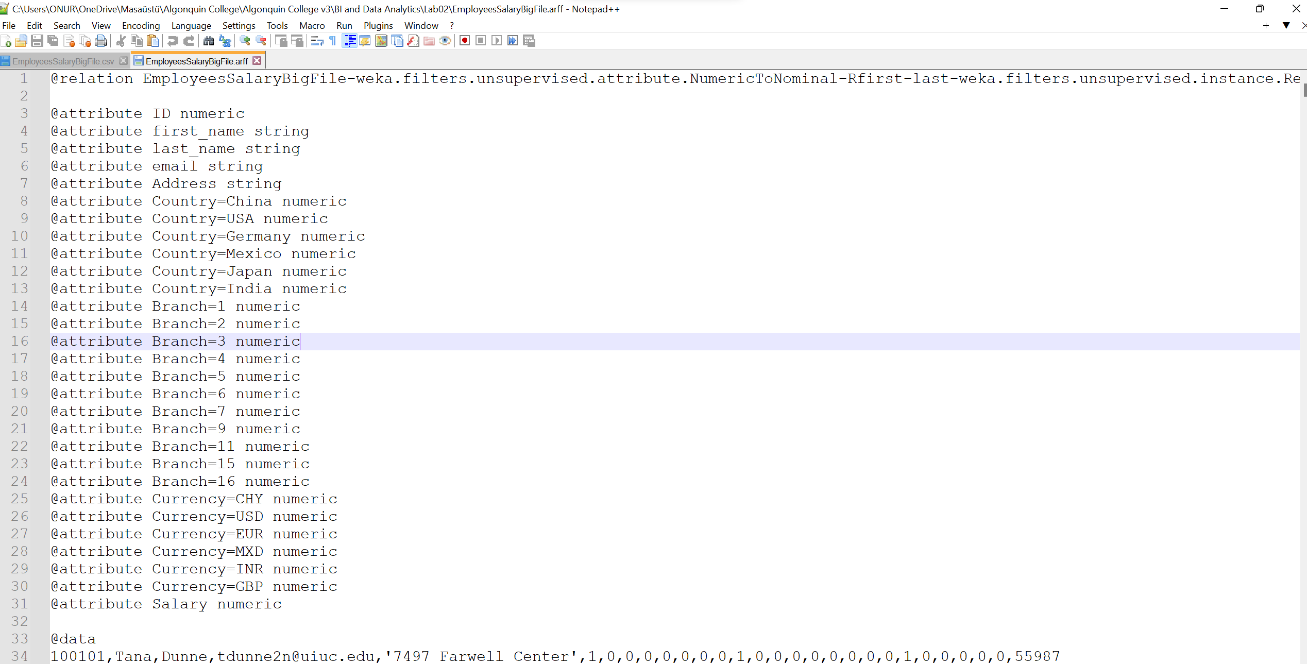
3

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

Nominal To Binary



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.