**Contribution Table:**

**Contribution** (This table should contain the list of all the students in the group. Clearly mention each student’s contribution towards the assignment. Mention “No Contribution” in cases applicable. If the contribution is equal the write 100%)

|  |  |  |  |
| --- | --- | --- | --- |
| **Sl. No.** | **Name (as appears in Canvas)** | **ID NO** | **Contribution (%)** |
| **1** | **Yenninti Sowjanya** | **2020SC04042** | **100%** |
| **2** | **Smita Rai** | **2020SC04233** | **100%** |
| **3** | **Anshuli Sarin** | **2020SC04595** | **100%** |

### **Problem Statement:**

An organization wanted to check association among employee experience, skills, traits etc. to better manage human resources

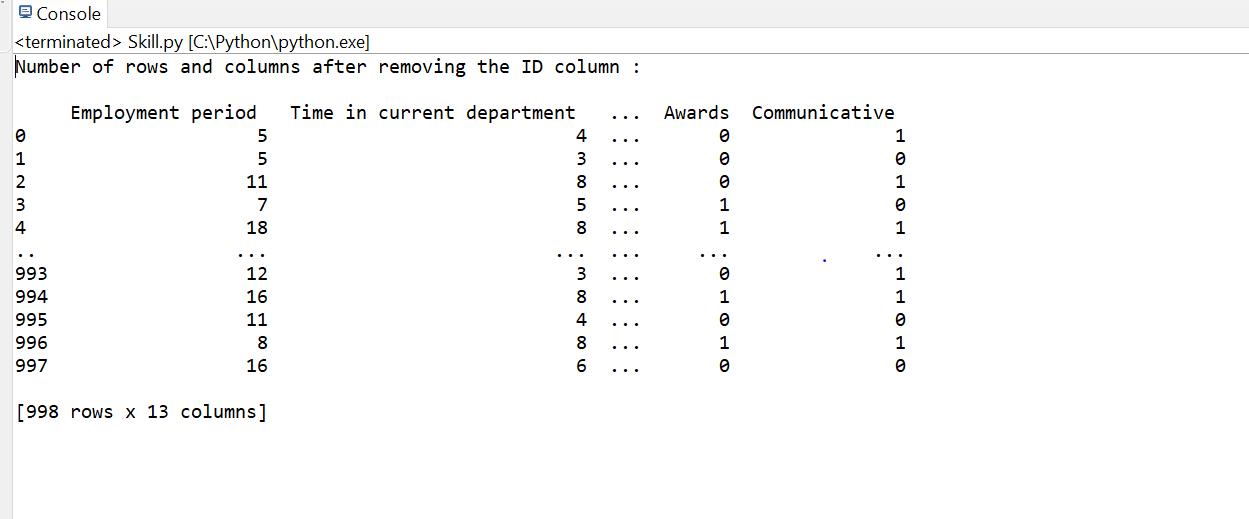
As a data scientist we should be performing various activities pertaining to the data such as, preparing the dataset for analysis; identify frequent patterns; formulate association rules and evaluate quality of rules.

**Process**

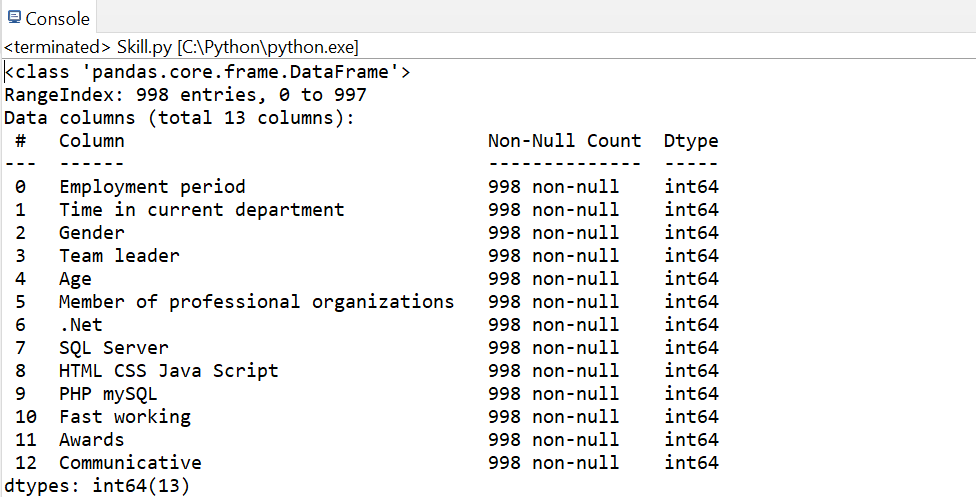
1. **Exploratory Data Analysis**

The first step is to perform the Exploratory data analysis. EDA is all about making sense of data in hand. Hence the below steps are taken

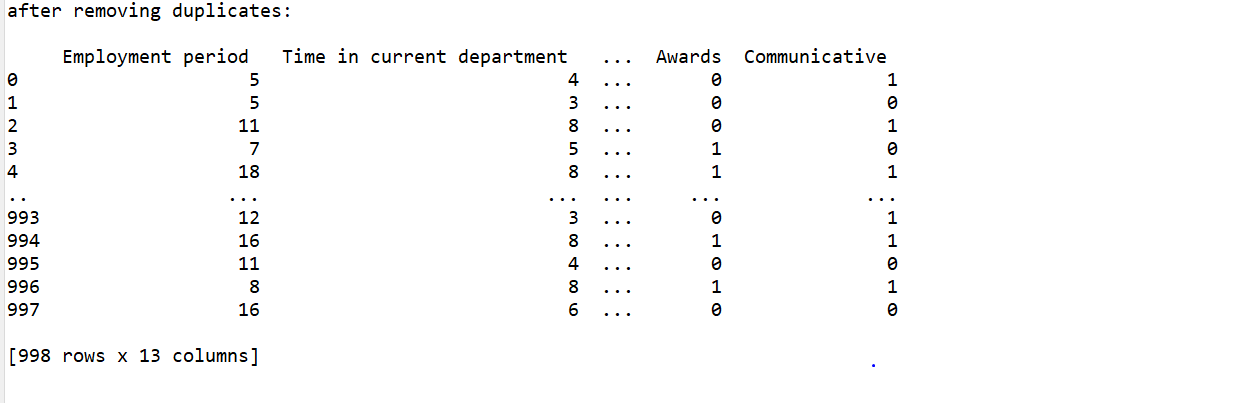
1. Import the necessary libraries as pandas, sk-learn etc. Load the Employee\_skills\_traits.csv file to a dataframe using pandas
2. AS Employee ID is unique and the relation of the employee with all other columns has been stated in the file. We are removing the Employee ID and formulating the Association rules. Hence the dataframe has been modified to have all the columns except Employee ID



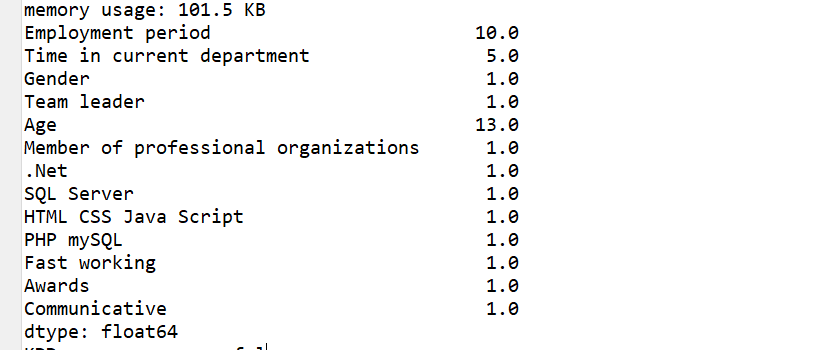
1. To know the columns and data types we have used **info()** function which the gives the result as below: This tells us there are 13 columns with type int which do not have any null values and 998 rows



1. The **describe()** function in pandas is very handy in getting various summary statistics. This function returns the count, mean, standard deviation, minimum and maximum values and the quantiles of the data. The data is written to a csv file *“Describe\_dataframe"*
2. The next step is to drop duplicates. Since our data set does not have duplicates, even after using function to remove duplicates, the data is same as original

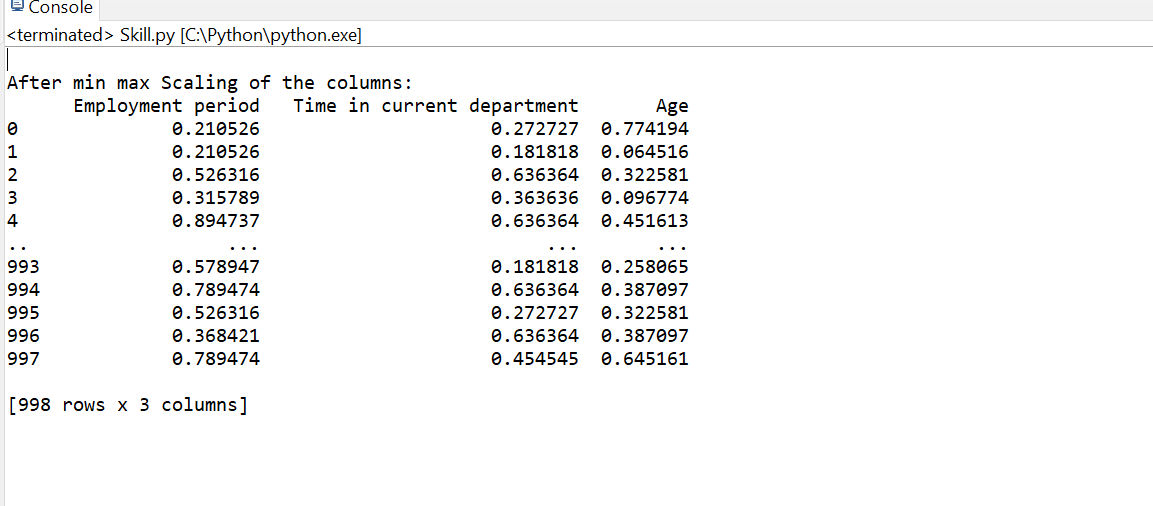


1. The next step is to find Outliers, we have calculated IQR and then tried to identify the noisy data by calculating IQR+1.5 and IQR-1.5. We could see that the IQR for *'Employment period '*,*'Time in current department '*, *'Age '*are of large values when compared to other values



Now that we have found the outliers and are done with Exploratory data analysis, the next step would be to perform Pre-processing of the data

1. **Data Pre-Processing**
2. From our dataset, by finding the outliers, we can conclude that there are columns *'Employment period '*,*'Time in current department '*, *'Age '*which have wide range of values. Since all other column values are either 0 or 1, these columns are also normalized to values between 0 and 1. We have used Min\_Max normalizing technique for achieving this.



1. We have now normalized the data and all the data column values are in range of 0 and 1. We now prepare train and test data out of this pre-processed data

The data is split in to train data with 80% and the test data with 20%. Upon performing the **describe()** function on the train and test data, we could see that 778 rows are considered for train data and 200 rows for test data. The data is written to excel sheets - *Decribe\_train.csv and Decribe\_test.csv*

**3. Frequent Patterns**

A function **FREQ\_PATTERN** is used to get the frequent patterns out of the data columns

of train data with **support of 0.25**. Below is the attachment for the same. We can see that

all columns individually have a support of > 0.25 and the frequent patterns are stopped

when a combination of only two columns is considered which means there are no

frequent patterns further identified with more than two columns association.



**4.Association Rules**

Now out of these frequent patterns, we have formulated association rules with **confidence of**

**0.5** . Please find the attachment for the same as below



Similarly, we have taken up a **Min\_support of 0.3 and confidence of 0.5** where we got the frequent patterns and rules are formulated as below



**Rules** 

We have iterated the process with test data **for Min\_support of 0.21 with confidence of 0.5** and for **Min\_support of 0.35 and confidence of 0.5.** Please find the data for the same

**Min\_support = 0.21**

|  |  |  |
| --- | --- | --- |
|  | Patterns | Support |
| 0 | ('Team leader ',) | 0.475 |
| 1 | ('Gender ',) | 0.48 |
| 2 | ('Time in current department ',) | 0.423181818 |
| 3 | ('Employment period ',) | 0.503684211 |
| 4 | ('Communicative ',) | 0.51 |
| 5 | ('Member of professional organizations ',) | 0.5 |
| 6 | ('SQL Server ',) | 0.52 |
| 7 | ('Fast working',) | 0.49 |
| 8 | ('PHP mySQL ',) | 0.48 |
| 9 | ('.Net ',) | 0.505 |
| 10 | ('Age ',) | 0.544677419 |
| 11 | ('HTML CSS Java Script ',) | 0.52 |
| 12 | ('Awards',) | 0.5 |
| 13 | ('HTML CSS Java Script ', 'Time in current department ') | 0.221818182 |
| 14 | ('HTML CSS Java Script ', 'Gender ') | 0.24 |
| 15 | ('HTML CSS Java Script ', 'Team leader ') | 0.24 |
| 16 | ('HTML CSS Java Script ', 'Employment period ') | 0.255789474 |
| 17 | ('HTML CSS Java Script ', 'Communicative ') | 0.3 |
| 18 | ('HTML CSS Java Script ', 'SQL Server ') | 0.275 |
| 19 | ('HTML CSS Java Script ', 'Member of professional organizations ') | 0.265 |
| 20 | ('HTML CSS Java Script ', 'PHP mySQL ') | 0.225 |
| 21 | ('HTML CSS Java Script ', '.Net ') | 0.275 |
| 22 | ('HTML CSS Java Script ', 'Age ') | 0.293225806 |
| 23 | ('HTML CSS Java Script ', 'Fast working') | 0.265 |
| 24 | ('HTML CSS Java Script ', 'Awards') | 0.285 |
| 25 | ('Time in current department ', 'Team leader ') | 0.226818182 |
| 26 | ('Time in current department ', 'Employment period ') | 0.256291866 |
| 27 | ('Time in current department ', 'Communicative ') | 0.216818182 |
| 28 | ('Time in current department ', 'Age ') | 0.242463343 |
| 29 | ('Time in current department ', 'Fast working') | 0.213181818 |
| 30 | ('Gender ', 'Team leader ') | 0.22 |
| 31 | ('Gender ', 'Employment period ') | 0.243684211 |
| 32 | ('Gender ', 'Communicative ') | 0.235 |
| 33 | ('Gender ', 'SQL Server ') | 0.23 |
| 34 | ('Gender ', 'Member of professional organizations ') | 0.24 |
| 35 | ('Gender ', 'PHP mySQL ') | 0.245 |
| 36 | ('Gender ', '.Net ') | 0.23 |
| 37 | ('Gender ', 'Age ') | 0.267903226 |
| 38 | ('Gender ', 'Fast working') | 0.22 |
| 39 | ('Gender ', 'Awards') | 0.23 |
| 40 | ('Team leader ', 'Employment period ') | 0.253684211 |
| 41 | ('Team leader ', 'Communicative ') | 0.245 |
| 42 | ('Team leader ', 'SQL Server ') | 0.245 |
| 43 | ('Team leader ', 'Member of professional organizations ') | 0.245 |
| 44 | ('Team leader ', 'PHP mySQL ') | 0.215 |
| 45 | ('Team leader ', '.Net ') | 0.265 |
| 46 | ('Team leader ', 'Age ') | 0.263709677 |
| 47 | ('Team leader ', 'Fast working') | 0.245 |
| 48 | ('Team leader ', 'Awards') | 0.225 |
| 49 | ('Employment period ', 'Communicative ') | 0.260526316 |
| 50 | ('Employment period ', 'SQL Server ') | 0.259736842 |
| 51 | ('Employment period ', 'Member of professional organizations ') | 0.230526316 |
| 52 | ('Employment period ', 'PHP mySQL ') | 0.236052632 |
| 53 | ('Employment period ', '.Net ') | 0.243947368 |
| 54 | ('Employment period ', 'Age ') | 0.292682513 |
| 55 | ('Employment period ', 'Fast working') | 0.249736842 |
| 56 | ('Employment period ', 'Awards') | 0.241578947 |
| 57 | ('Communicative ', 'SQL Server ') | 0.25 |
| 58 | ('Communicative ', 'Member of professional organizations ') | 0.265 |
| 59 | ('Communicative ', 'PHP mySQL ') | 0.25 |
| 60 | ('Communicative ', '.Net ') | 0.25 |
| 61 | ('Communicative ', 'Age ') | 0.272419355 |
| 62 | ('Communicative ', 'Fast working') | 0.27 |
| 63 | ('Communicative ', 'Awards') | 0.255 |
| 64 | ('SQL Server ', 'Member of professional organizations ') | 0.24 |
| 65 | ('SQL Server ', 'PHP mySQL ') | 0.225 |
| 66 | ('SQL Server ', '.Net ') | 0.265 |
| 67 | ('SQL Server ', 'Age ') | 0.283387097 |
| 68 | ('SQL Server ', 'Fast working') | 0.245 |
| 69 | ('SQL Server ', 'Awards') | 0.305 |
| 70 | ('Member of professional organizations ', 'PHP mySQL ') | 0.225 |
| 71 | ('Member of professional organizations ', '.Net ') | 0.26 |
| 72 | ('Member of professional organizations ', 'Age ') | 0.268548387 |
| 73 | ('Member of professional organizations ', 'Fast working') | 0.24 |
| 74 | ('Member of professional organizations ', 'Awards') | 0.24 |
| 75 | ('PHP mySQL ', 'Age ') | 0.264516129 |
| 76 | ('PHP mySQL ', 'Fast working') | 0.235 |
| 77 | ('PHP mySQL ', 'Awards') | 0.245 |
| 78 | ('.Net ', 'Age ') | 0.269677419 |
| 79 | ('.Net ', 'Fast working') | 0.25 |
| 80 | ('.Net ', 'Awards') | 0.25 |
| 81 | ('Age ', 'Fast working') | 0.265967742 |
| 82 | ('Age ', 'Awards') | 0.263225806 |
| 83 | ('Fast working', 'Awas') | 0.23 |

**Rules**

**When Min\_Support = 0.35,**



**Rules**

****

**We could see that when the support is less than 0.3, we observed many frequent patterns even with confidence of 50% whereas when the support is greater than 0.3, we found the frequent patterns to be less and no association rules were generated**

**IMPORTANCE OF RULE DISCOVERY:**

**Rule discovery or rule extraction from data are data-mining techniques aimed at understanding data structures, providing comprehensible description instead of only black-box prediction. Rule-based systems should expose in a comprehensible way knowledge hidden in data, providing logical justification for drawing conclusions, showing possible inconsistencies, and avoiding unpredictable conclusions that black box predictors may generate in untypical situations. Sets of rules are useful if rules are not too numerous, comprehensible, and have sufficiently high accuracy. Rules are used to support decision making in classification (**[**Machine Learning**](https://doi.org/10.1007/978-1-4419-9863-7_100803)**), regression (**[**Regression**](https://doi.org/10.1007/978-1-4419-9863-7_101269)**), and association tasks.**

Note:

**The frequent patterns and association rules are also written to the below excel sheet for reference**

****

****

****

****