**Step wise explanation of the assignment.**

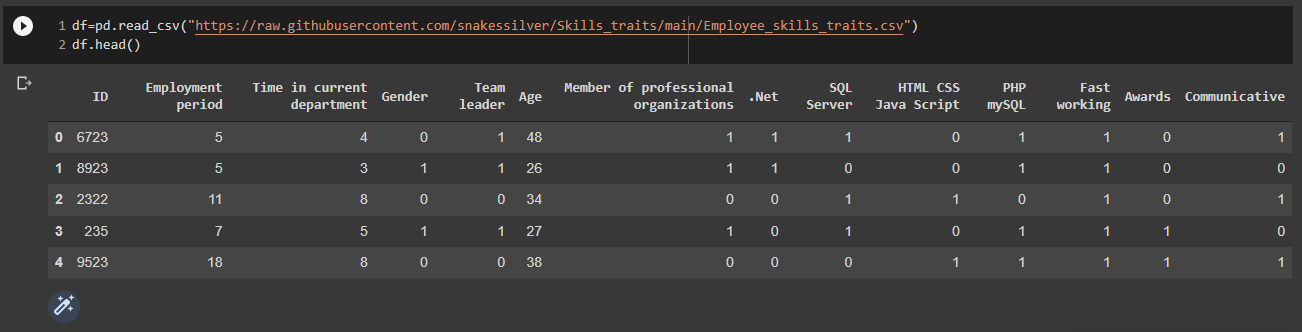
**Problem statement**

Identifying the association among employees’ skills, experience and traits for better management of human resource.

\*\* The steps mentioned here are the same steps mentioned in both the .py files\*\*

* Step 1- Data Collection

In this step the data was collected from the GitHub repository where is was stored. Then various functions like info() , describe(), head() were used to get a basic understanding of data and to understand the structure of data. After applying all the functions mentioned above, it was found that the data has 998 rows and 14 columns, the data in all the columns is in integer format (which is a good sign as it will be helpful in processing the data).



Screenshot 1: Showing the head of the dataset.

* Step 2- Data Pre-Processing

This step deals with removing noise or impurities from data.

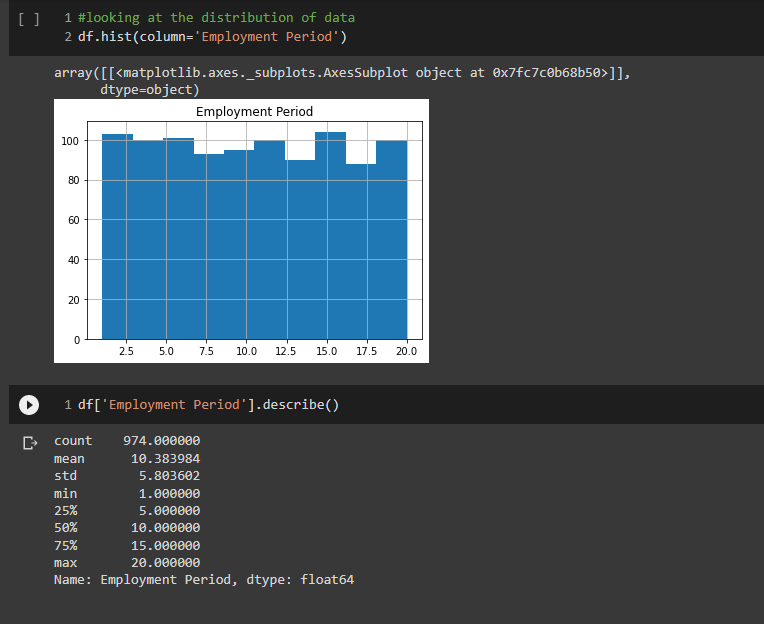
The following steps were taken to check for noise in data and were solves using various functions of python:

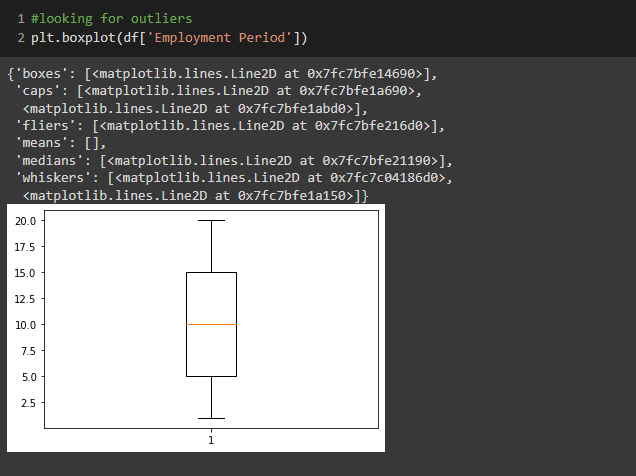
1. Looking for missing values: It was found that the data has no missing values
2. Few column names contained unwanted space which was removed.
3. The data was checked for duplicate entries and it was found that the data has duplicate entries so the duplicate rows were removed from the dataset.
4. It was checked if any row in the dataset has value in employment period column greater than the value in age column and it was found that no such rows exist.
5. It was checked if any row in the data set has value in time in current department column greater than the value in age column and it was found that no such rows exist.

* Step 3- Exploratory Data Analysis

Data analysis was performed on all columns of the dataset and a few columns together:

1. Employment period

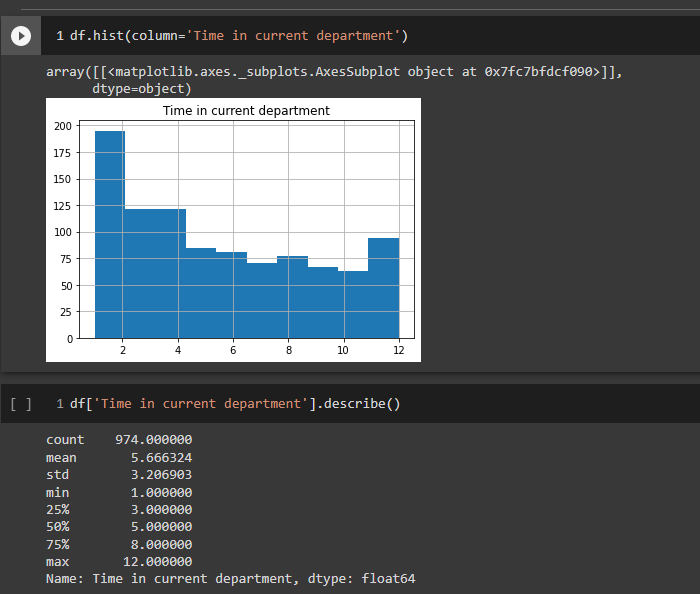


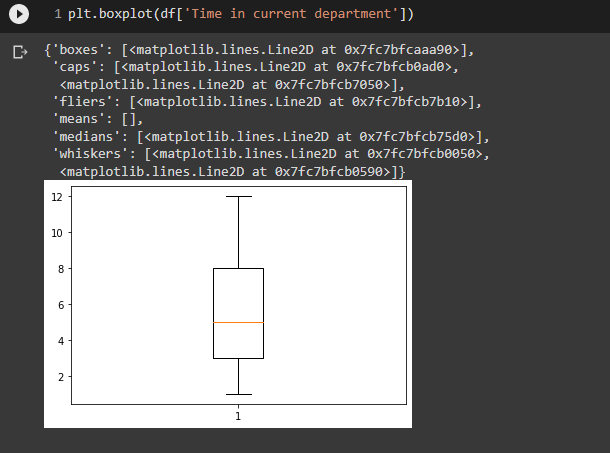


Screenshot 2: showing analysis on Employment period column.

The above analysis indicates that the data points are near to each other as the difference between the three quartile values is less and the box plot indicates that there are no outliers and the data is equally distributed around the median.

1. Time in current department

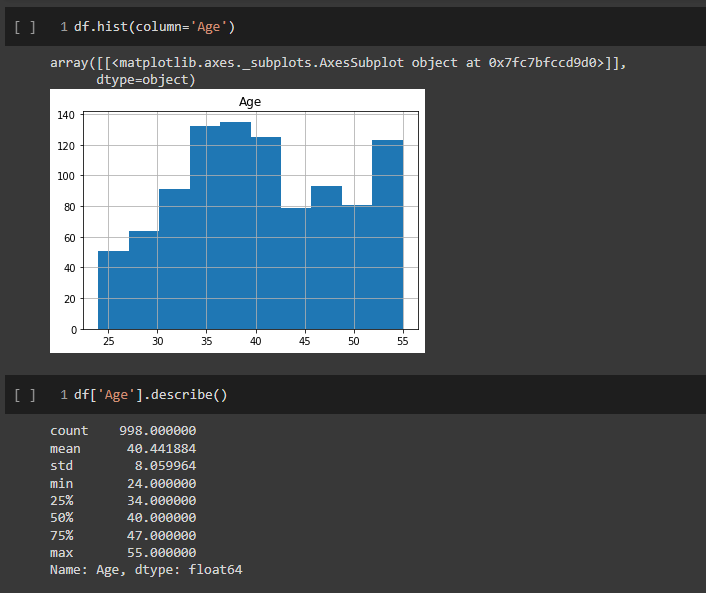


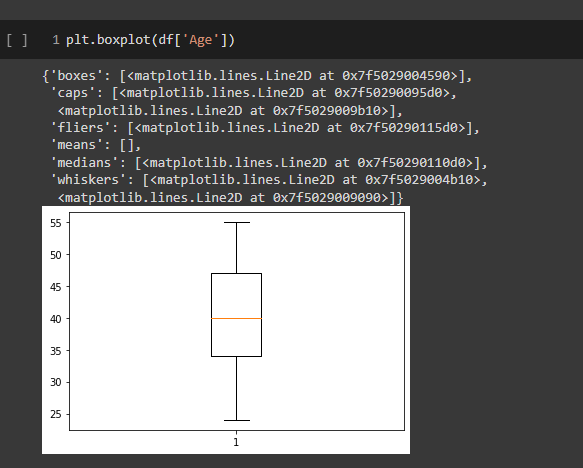


Screenshot 3: showing analysis on Time in current department column.

The above analysis indicates that the data points are near to each other as the difference between the three quartile values is less and the box plot indicates that there are no outliers and the more number of data points that are greater than the median.

1. Age





Screenshot 4: showing analysis on age column.

The above analysis indicates that the data points are near to each other as the difference between the three quartile values is less and the box plot indicates that there are no outliers and the data is well distributed around the median.

1. Gender

Analysis shows that there are 514 female and 484 male in the dataset.

1. Team leader

Analysis shows that there are 503 team leaders and 495 non- team leaders in the data set or vice versa.

1. Member of professional organizations

Analysis shows that there are 509 Member of professional organizations and 489 are not Member of any professional organizations in the dataset or vice versa.

1. .Net

Analysis shows that there are 522 people in the dataset who have the skill of .Net and there are 473 who do not have the skill or vice versa.

1. SQL Server

Analysis shows that there are 523 people in the dataset who have the skill of SQL Server

and there are 475 who do not have the skill or vice versa.

1. HTML CSS Java Script

Analysis shows that there are 507 people in the dataset who have the skill of HTML CSS Java Script and there are 491 who do not have the skill or vice versa.

1. PHP MySQL

Analysis shows that there are 521 people in the dataset who have the skill of PHP MySQL

and there are 477 who do not have the skill or vice versa.

1. Fast working

Analysis shows that there are 502 people in the dataset who are fast working and there are 496 who cannot or vice versa.

1. Awards

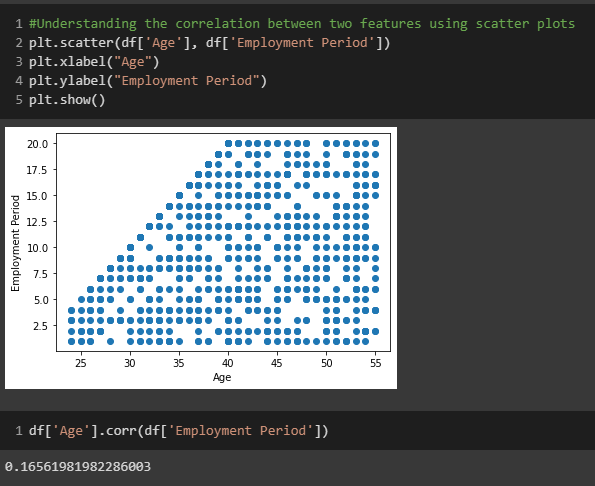
Analysis shows that there are 501 people in the dataset who have received and there are 497 who have not or vice versa.

1. Communicative

Analysis shows that there are 520 people in the dataset who are Communicative

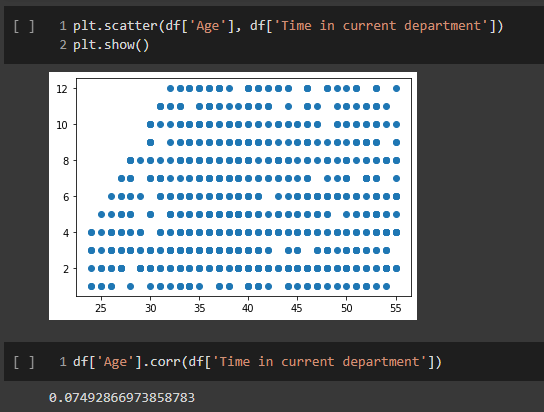
and there are 478 who are not or vice versa.

\*Age & Employment period

Screenshot 5: showing analysis on age and employment period columns.

The above graph indicates that the period of employment varies irrespective of the employees' age. And the graph does not show any specific pattern also the correlation coefficient is very less which means the features are weakly correlated.

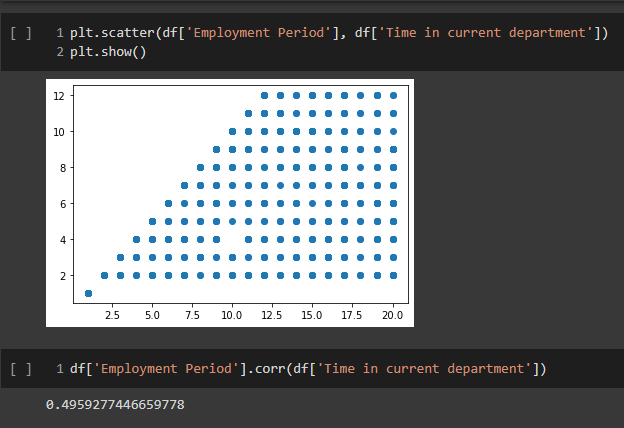
\*Age & Time in current department



Screenshot 6: showing analysis on age and Time in current department columns.

The above graph indicates that the period of employment varies irrespective of the employees' age. And the graph does not show any specific pattern also the correlation coefficient is very less which means the features are weakly correlated.

\*Employment period & Time in current department



Screenshot 7: showing analysis on Employment period and Time in current department columns.

The above graph indicates that the period of employment varies irrespective of the employees' age. The graph does not show any specific pattern and even though the correlation coefficient is 0.49 it is still less which means the features are weakly correlated.

* Step 4- Data Transformation (Binning)

The data was transformed using binning in following steps:

1. Categorizing the “employment period” column into 4 categories depending on a person’s years of employment. The data here is categories into 4 categories as the range of data is higher as compared to other columns (like Time in current department,). The range of the data is : min value -1 and max value -20.[1, 3, 8, 15, 20]

Entry- 1 year to 3 years

Mid- 3 years to 8 years

Sr. (Senior) - 8 years to 15 years

Exec (Executive) – 15 years to 20 years

1. Categorizing the “Time in current department” column into 3 categories depending on the time spend in current department. Here the range (min value- 1, max value- 12) of the data is lower as compared to the “employment period” column and so the data is divided only in 3 categories.

Fresher- 1year to 2 years

Senior- 2 years to 8 years

Experienced- 8 years to 12 years

1. Categorizing the “age” into 3 categories depending on the age of a person. Similar to the “Time in current department” column the range (min- 24 max-55) of the data is smaller and so it is divided only in 3 categories.

Youngster- 24 years to 30 years

Middle\_aged - 30 years to 50 years

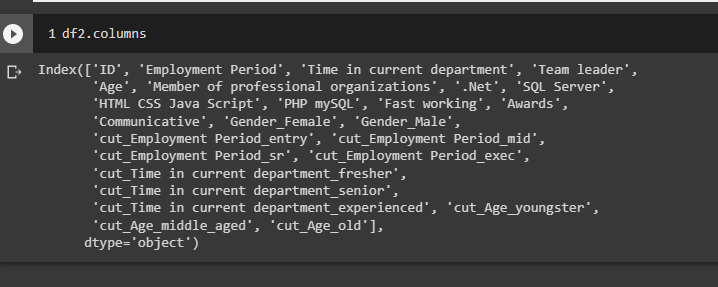
Old - 50 years to 55 years

1. The values in “gender” column are in form of 0 and 1 so 0 is replaced with female and 1 is replaced with male.

* Step 5- Data Transformation (Conversion)

In this step the dataset is transformed by creating dummy columns are created to prepare the dataset. Creating of dummy columns gives a better understanding of the data for analysis purpose. The values in dummy column are 1s and 0s where 1 indicates when a categorical event occurs and 0 indicates when it doesn't occur. This process of making dummy columns is spplies on the categorical data.

* Step 6- Dimensionality Reduction

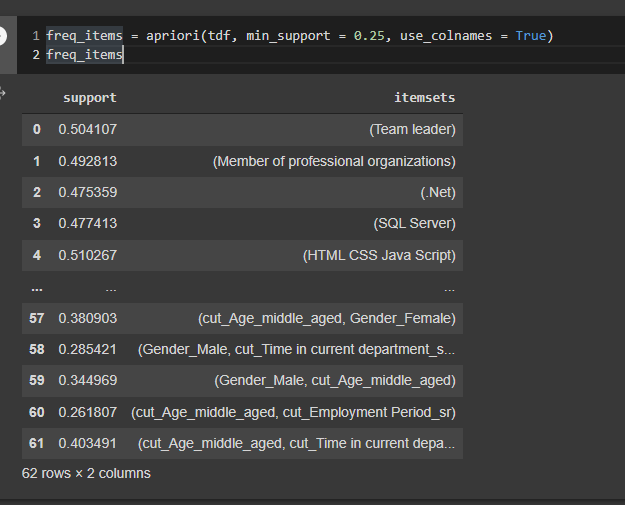


Screenshot 8: showing the total columns after creating dummy columns.

The above screenshot shows the columns after creating dummy columns. Now, in further steps apriori algorithm is to be so some of the columns like 'ID', 'Employment Period', 'Time in current department', 'Age' are not required as they do not contribute to forming of association rules and so they are dropped from the dataframe. Now, after 6 steps the data is prepared for apriori algorithm.

* Step 7- Generating Frequent Items

With the transformed data, python libraries and functions the frequent items are created as shown in the screenshot below. The frequent items are created considering minimum support count as 0.25.



Screenshot 9: showing the frequent itemsets generated.

* Step 8- Generating Association Rules

With the generated frequent itemsets, association rules are generated. in this step user defined functions are created for easy generation of frequent itemsets and association rules y varying parameters and for later comparing the rules generated.

* Step 9- Iterating by varying parameters

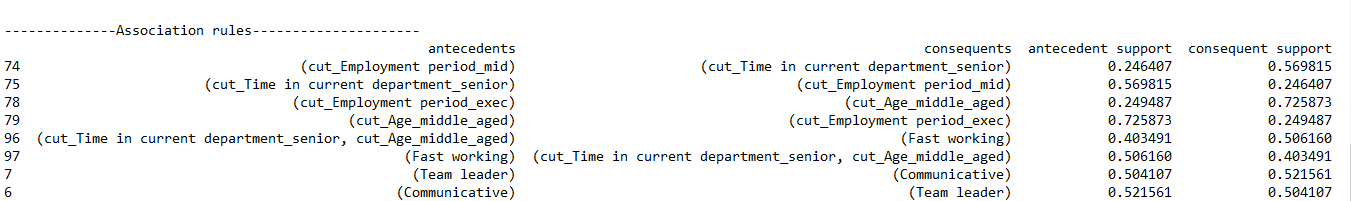
Association rules are generated by varying the value of minimum support count. Values like 0.1, 0.2, 0.25, 0.3, 0.4, 0.5 and 0.6 are used for generating different sets of association rules.

* Comparing the association rules

In the previous step 5 sets of association rules were created. It was observed that for the values of minimum support count as 0.4, 0.5 and 0.6 no association rules were generated. Association rules were generated only for 4 values of minimum support count which are 0.1, 0.2, 0.25 and 0.3. Now, these 4 sets of values were compared on the basis of higher confidence values and it was found that the best association rules are generated with minimum support count value as 0.2. So, the association rules generated with value as 0.2 are considered as final rules.

* Briefly explain importance of discovered rules

The final association rules generated are displayed in the output.txt file. Screenshot shows few of the rules generated.



Screenshot 10: showing few of the association rules from the output file.

An association rule has two parts: an antecedent (if) and a consequent (then). An antecedent is an item found within the data. A consequent is an item found in combination with the antecedent. Association rules are useful for analyzing and predicting the important outputs that are required from a set of data. Like in our case association rules are helping us analyze that what would be the requirement for obtaining the various item sets in the “consequents” column and the requirements are stated in the antecedents column.

For example, considering the following rule from the output file:

If we want to find the right fit for SQL-Server guys, then

1. By Association rule 34, 50% of SQL-server guys are fast working (24% SQL server guys are fast working as per data)
2. By Association rule 41 and 61, 48% and 53% employees with 2-8 years’ experience knows SQL-server and are fast working.
3. By Association rule 74, employee’s with 3-8 year’s total employment experience are among the 90% of 2-8 years’ experience in the current department

By virtue of above three rules, we can say that if we float a recruitment for SQL server employee who have 3-8 years’ of employment experience than there are more chances that he/she will be fast-working (as per confidence around 50%)

Another instance could be :

Where antecedent is

(cut\_Time in current department\_senior, cut\_Age\_middle\_aged)

Consequent is,

(Fast working)

Antecedent support is,

0.403491

Consequent support is,

0.506160

Support is,

0.226899

Confidence is,

0.562341

Lift is,

0.022668

Leverage is,

1.128367

As per the data transformation,

cut\_Time in current department\_senior has 2-8 years’ experience in current department

cut\_Age\_middle\_aged has 30-50 years.

Thus, from the above association rule we can imply that to find the work force that fast working we can say with 0.562341 confidence that cut\_Time in current department\_senior and cut\_Age\_middle\_aged people will be needed.