**Data Mining**

(S1-21\_DSECLZC415)

FIRST SEMESTER 2021-22



Assignment No.: 1

**Skills Traits**

**SUBMITTED BY**

Varinder Singh - 2021fc04070

Bandaru Raja Sekhar - 2021fc04074

Baskar K - 2021fc04066

**Table of Contents**

[1. Problem Statement](#_heading=h.bacw9fer07q8) **3**

[2. Data Preprocess & EDA](#_heading=h.smmi906b8one) **3**

[3. Proposition of Parameters](#_heading=h.g49ij19r4ov7) **6**

[4. Discovering the frequent itemsets and Formulating Association Rules](#_heading=h.cas6a4iyelgn) **6**

[5. Association rules by varying parameters](#_heading=h.m3t25m5d2n16) **9**

[6. Importance of Discovered Rules](#_heading=h.7zxoc12xvrw) **11**

# **1. Problem Statement**

An organization wants to check association among its employee experience, skills, traits, etc. to better manage human resources. As the organization decides to create Gender equality in its company among their employees.

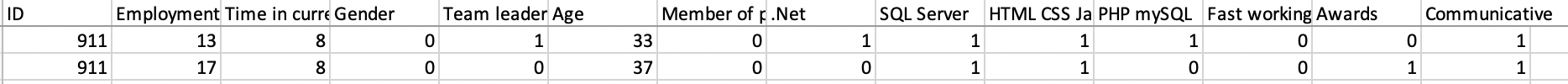
Draw up inferences to make the company to have more equal opportunities and responsibilities within the organization for both women and men from the provided employee data. Recognise patterns from the available data and evaluate the efficacy of methods to obtain patterns.

This problem can be solved by using Apriori Algorithm so that we can get frequent itemset and after that, we can get association rules with suitable support, confidence and lift.

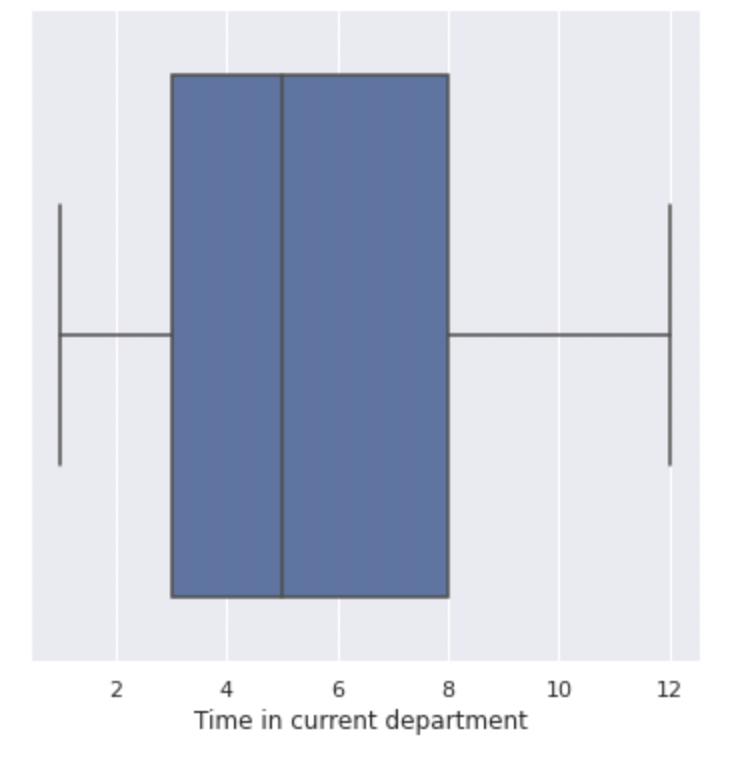
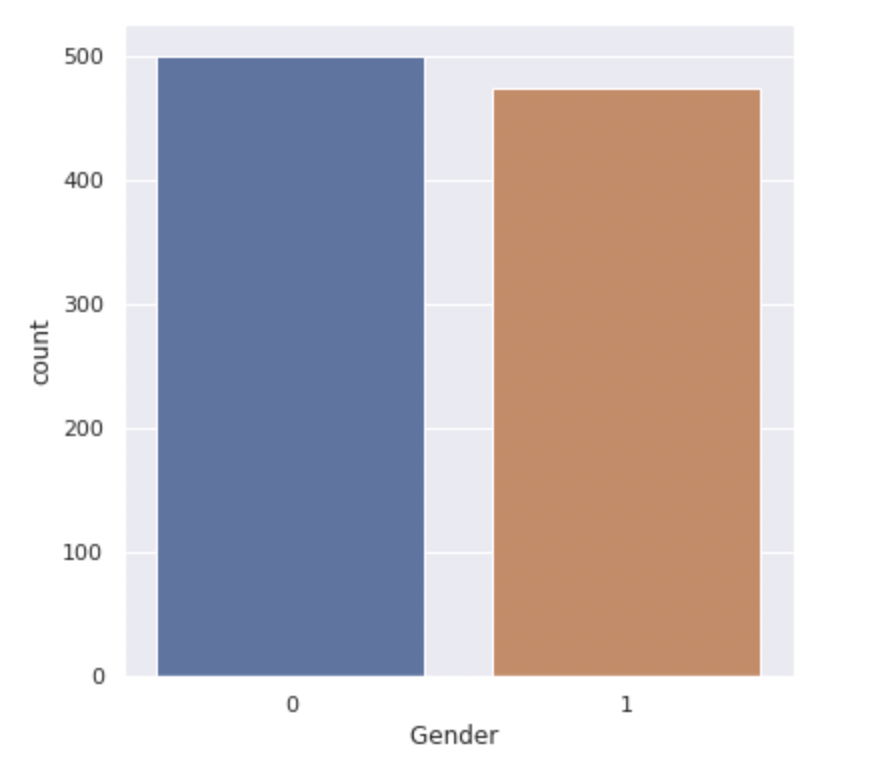
# **2. Data Preprocess & EDA**

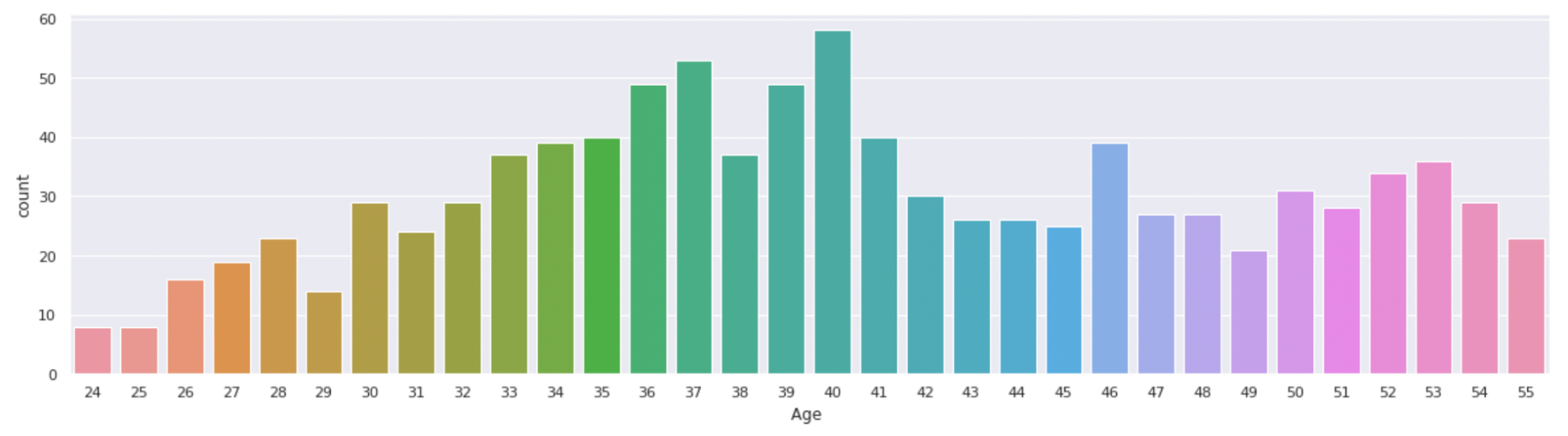
For every KDD, we need to process data so that we can get clean data for driving inferences. We followed the below process for cleaning the data:

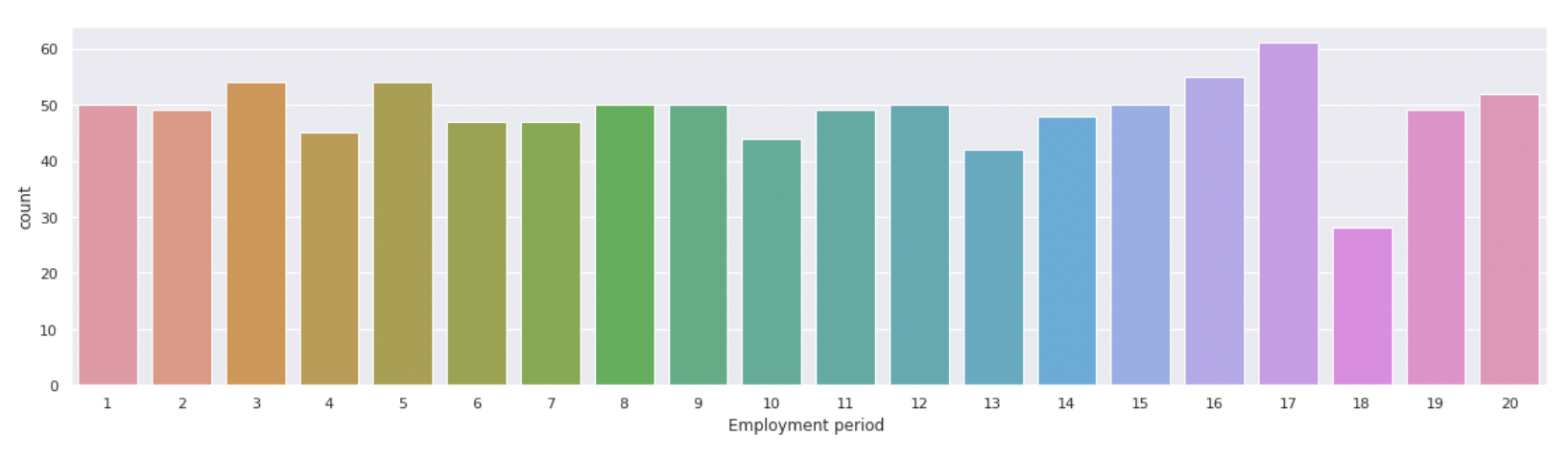
* We loaded the data from the CSV file – **“Employee\_skills\_traits.csv”** for Exploratory Data Analysis.
* There were duplicates in the data having two records of the same ID. eg: employee with ID: 911. For our analysis, we are going to remove the duplicates and keep the records occurring last. While making sure that no nulls are present in the resultant data frame.



* We first checked the spread of the employees of the organization based on their age, their employment period, gender, time in the current department, etc.







* While cleaning the dataset in our data frame, we grouped the data and assumed a few points which are as follows:

1. For the Gender column, we assumed 0 as Male and 1 as Female.
2. For the Team leader column, we assumed 1 as Team Leader and 0 as None.
3. We grouped the employees in three categories for column Age as:

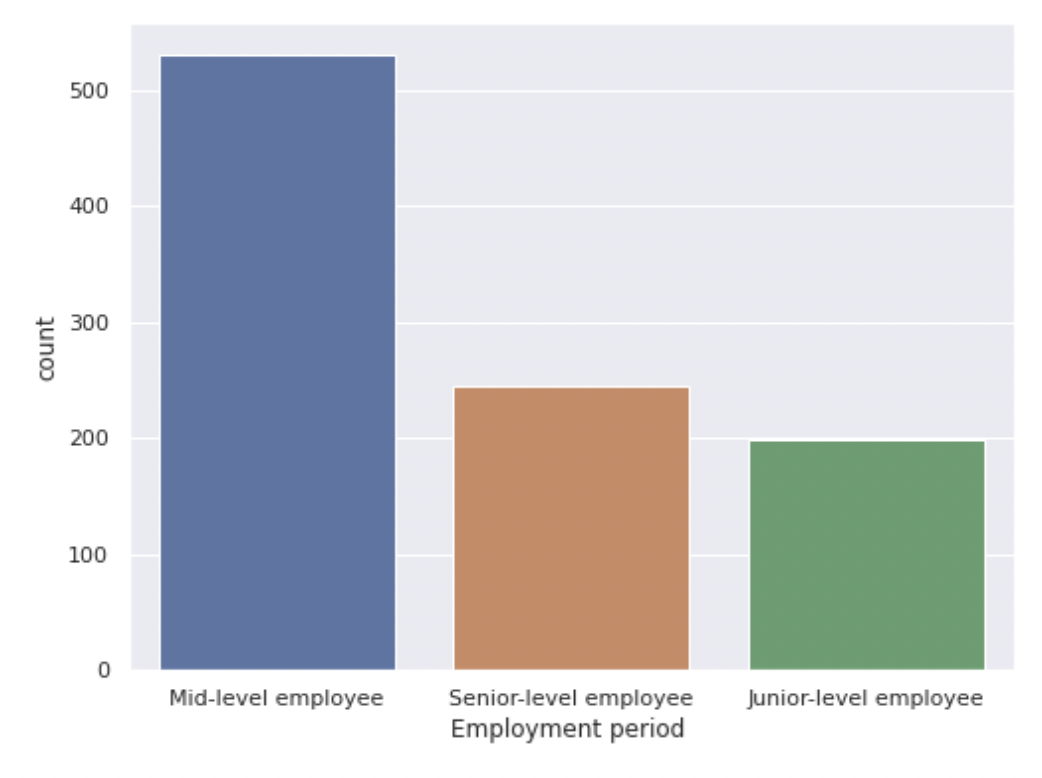
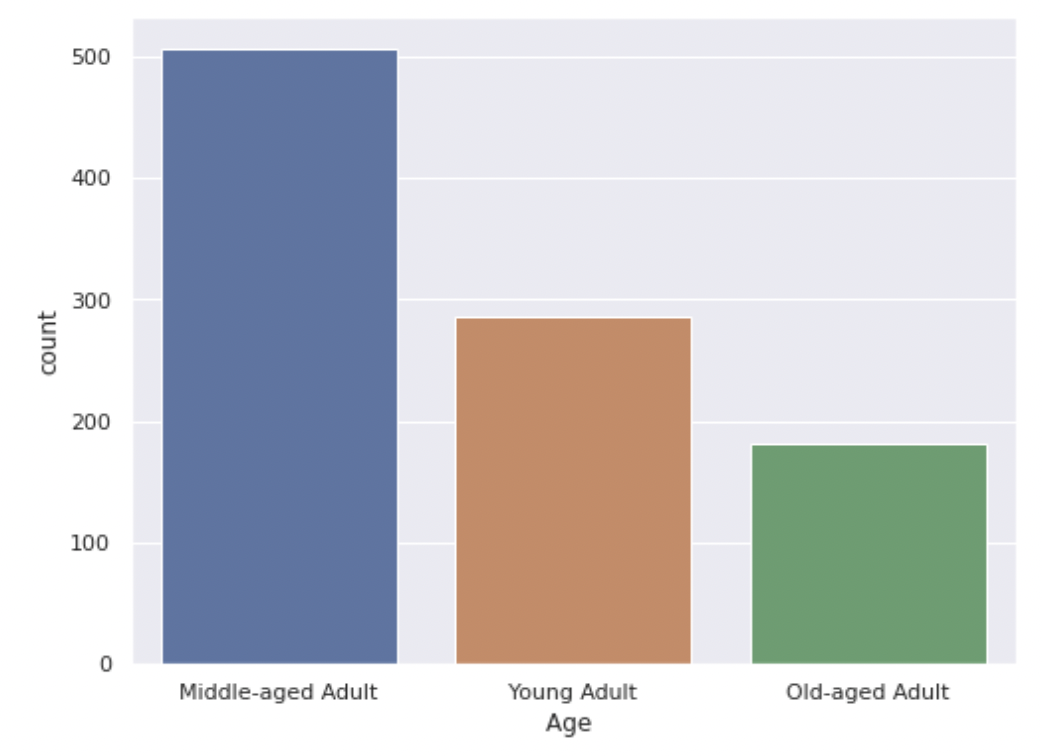
* Young Adult - for age <= 35 years
* Middle-aged Adult - for age between 36 and 49(conditions inclusive)
* Others as Old-aged Adult

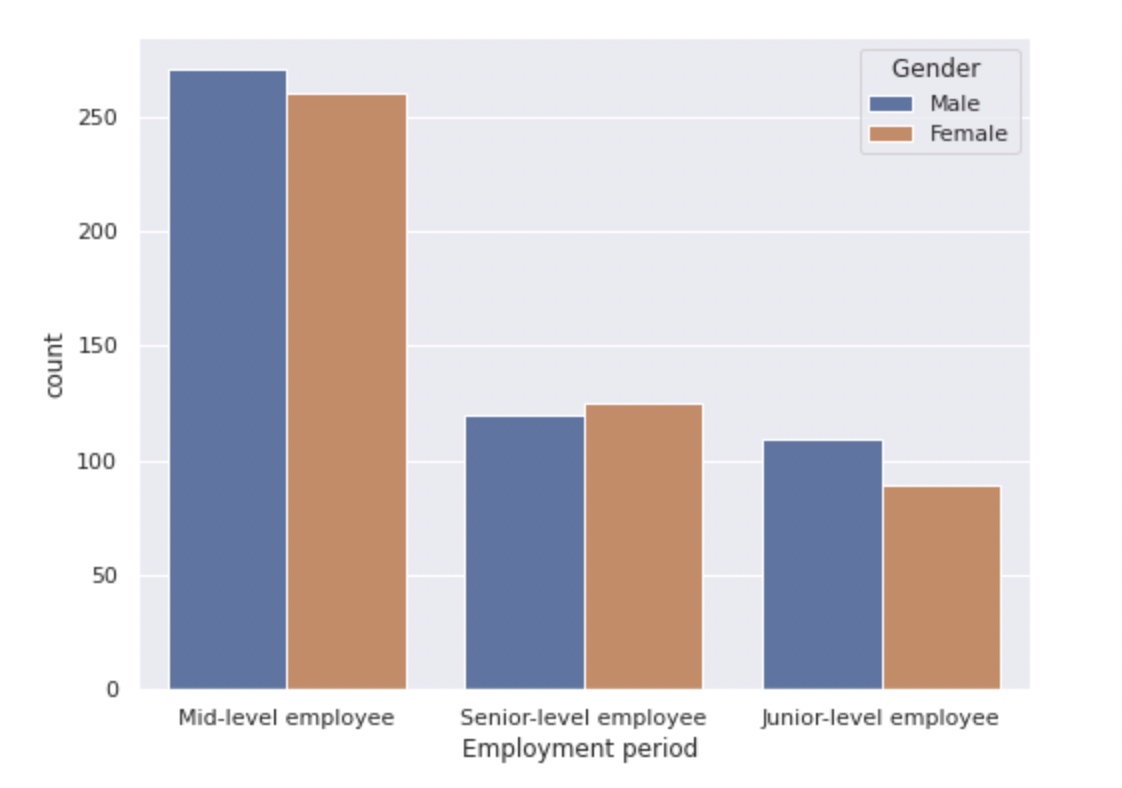
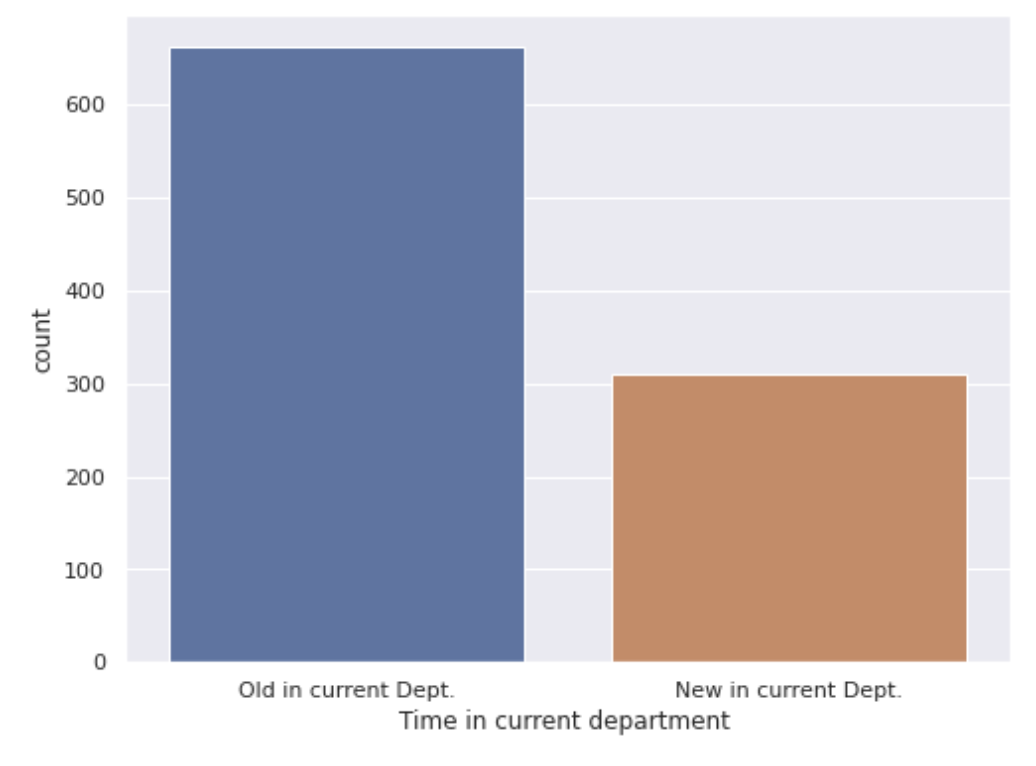
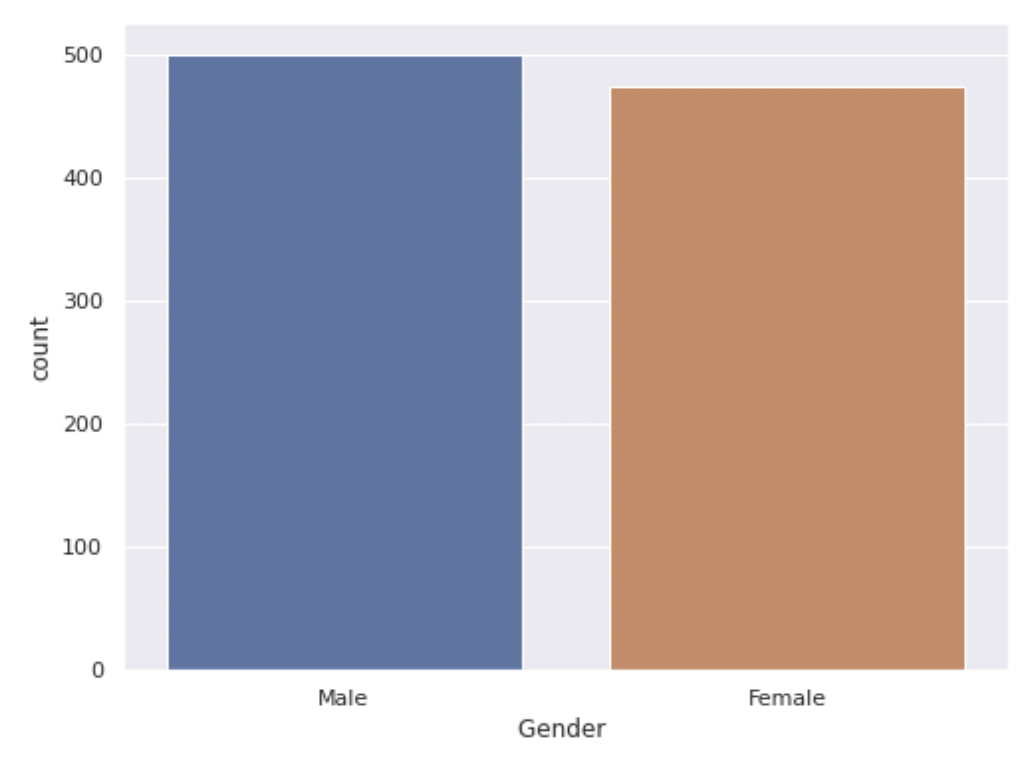
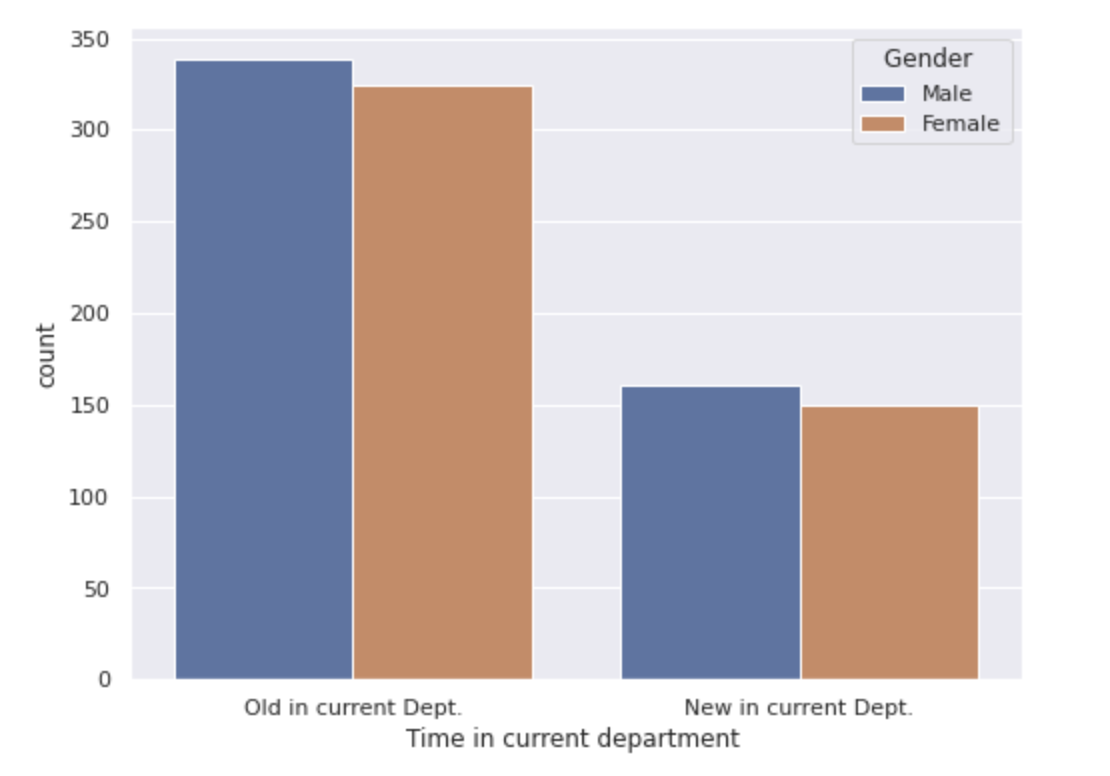
1. We grouped the employee in three categories for column Employment Period:

* Junior-level employee - for experience <= 4 years
* Mid-level employee - for age between 5 and 15(conditions inclusive)
* Others as Senior-level employee

1. We grouped the employees in two categories for column Time in the current department as:

* New in current Dept.- for experience in current dept. <= 3 years
* Others as Old in current Dept.
* Streamlined the data frame which excluded the ID column for further analysis.
* After the above steps, we got the distribution of the data across the columns as below:



* Dataframe is split into two for further analysis in terms of their technical skills and organizational details. Both data frames include the Gender column for getting inferences based on gender.
* Each resultant data frame is processed to form an array of the array which is then processed by the TransactionEncoder() function to transform to the final data frame to identify the frequent itemsets and association rules with proposed parameters.

# **3. Proposition of Parameters**

# For parameters to identify the frequent itemsets, we propose minimum support for the employee’s traits which occurs for 20% of the employee.

Therefore,

Minimum Support = ( 20% of Total employees ) / Total employees

= ( 20 % of 974 ) / 974

= 0.2

which comes close to 0.2 minimum support.

The value of lift is taken as 1 which is calculated by lift = (support/confidence) = 0.2 / 0.2 = 1

# **4. Discovering the frequent itemsets and Formulating Association Rules**

For solving the problem, we need to use the apriori algorithm to train our model and find frequent itemsets from it and further formulate association rules. As in the previous step, we proposed parameters of support as 0.2 and lift as 1, using which we will step forward to work on both the dataset(df\_gender\_pos and df\_gender\_skill) to find frequent itemsets. Frequent itemsets with minimum support of 0.2 identified as follows:

Frequent itemsets including gender and technical skills:

**========================================================================================**

**Frequent Itemsets are as follows:**

**========================================================================================**

**Support | Itemsets**

**========================================================================================**

**0.5133470226 ('Male',)**

**0.4866529774 ('Female',)**

**0.272073922 ('Male', 'Fast working')**

**0.2689938398 ('Communicative', 'Male')**

**0.2618069815 ('Male', 'HTML CSS Java Script')**

**0.2546201232 ('Awards', 'Male')**

**0.2525667351 ('Male', 'SQL Server')**

**0.2515400411 ('.Net', 'Male')**

**0.250513347 ('Awards', 'Female')**

**0.249486653 ('Communicative', 'Female')**

**0.2474332649 ('PHP mySQL', 'Female')**

**0.2453798768 ('Female', 'HTML CSS Java Script')**

**0.2320328542 ('Fast working', 'Female')**

**0.227926078 ('Male', 'PHP mySQL')**

**0.2258726899 ('.Net', 'Female')**

**0.2207392197 ('SQL Server', 'Female')**

Frequent itemsets including gender and organizational exp. position:

**========================================================================================**

**Frequent Itemsets are as follows:**

**========================================================================================**

**Support | Itemsets**

**========================================================================================**

**0.5133470226 ('Male',)**

**0.4866529774 ('Female',)**

**0.3480492813 ('Male', 'Old in current Dept.')**

**0.3326488706 ('Old in current Dept.', 'Female')**

**0.2782340862 ('Male', 'Mid-level employee')**

**0.2741273101 ('Middle-aged Adult', 'Male')**

**0.2669404517 ('Mid-level employee', 'Female')**

**0.2525667351 ('Team leader', 'Female')**

**0.2525667351 ('Team leader', 'Male')**

**0.250513347 ('Member of professional organizations', 'Male')**

**0.2464065708 ('Middle-aged Adult', 'Female')**

**0.2351129363 ('Member of professional organizations', 'Female')**

**0.2217659138 ('Male', 'Old in current Dept.', 'Mid-level employee')**

**0.2166324435 ('Mid-level employee', 'Old in current Dept.', 'Female')**

**0.2032854209 ('Middle-aged Adult', 'Male', 'Old in current Dept.')**

Association Rules formulated by taking into account the lift equals 1. The rules formulated including gender and technical-soft skills:

**========================================================================================**

**Association Rules are as follows:**

**========================================================================================**

**Rule: 1**

**========================================================================================**

**('PHP mySQL',) ==> ('Female',)**

**Support: 0.24743326488706366**

**Confidence: 0.5205183585313174**

**Lift: 1.0695883569820743**

**Rule: 2**

**========================================================================================**

**('Male',) ==> ('Fast working',)**

**Support: 0.2720739219712526**

**Confidence: 0.53**

**Lift: 1.0513645621181265**

**Rule: 3**

**========================================================================================**

**('Male',) ==> ('SQL Server',)**

**Support: 0.25256673511293637**

**Confidence: 0.4920000000000001**

**Lift: 1.0394967462039049**

**Rule: 4**

**========================================================================================**

**('Male',) ==> ('.Net',)**

**Support: 0.2515400410677618**

**Confidence: 0.49**

**Lift: 1.0263655913978496**

**Rule: 5**

**========================================================================================**

**('Awards',) ==> ('Female',)**

**Support: 0.25051334702258726**

**Confidence: 0.4959349593495934**

**Lift: 1.019073102123426**

**Rule: 6**

**========================================================================================**

**('Male',) ==> ('Communicative',)**

**Support: 0.26899383983572894**

**Confidence: 0.524**

**Lift: 1.0106455445544555**

**Rule: 7**

**========================================================================================**

**('Male',) ==> ('HTML CSS Java Script',)**

**Support: 0.2618069815195072**

**Confidence: 0.5100000000000001**

**Lift: 1.0055465587044536**

The association rules formulated including gender and organizational exp. position:

**========================================================================================**

**Association Rules are as follows:**

**========================================================================================**

**Rule: 1**

**========================================================================================**

**('Mid-level employee',) ==> ('Female', 'Old in current Dept.')**

**Support: 0.2166324435318275**

**Confidence: 0.3973634651600753**

**Lift: 1.194543256376276**

**Rule: 2**

**========================================================================================**

**('Old in current Dept.',) ==> ('Mid-level employee', 'Female')**

**Support: 0.2166324435318275**

**Confidence: 0.31825037707390647**

**Lift: 1.1922148741153265**

**Rule: 3**

**========================================================================================**

**('Male', 'Mid-level employee') ==> ('Old in current Dept.',)**

**Support: 0.22176591375770022**

**Confidence: 0.7970479704797049**

**Lift: 1.170927184384966**

**Rule: 4**

**========================================================================================**

**('Male', 'Old in current Dept.') ==> ('Mid-level employee',)**

**Support: 0.22176591375770022**

**Confidence: 0.6371681415929203**

**Lift: 1.1687415629218538**

**Rule: 5**

**========================================================================================**

**('Male', 'Old in current Dept.') ==> ('Middle-aged Adult',)**

**Support: 0.2032854209445585**

**Confidence: 0.5840707964601769**

**Lift: 1.122061056710478**

**Rule: 6**

**========================================================================================**

**('Middle-aged Adult', 'Male') ==> ('Old in current Dept.',)**

**Support: 0.2032854209445585**

**Confidence: 0.7415730337078652**

**Lift: 1.089430067619096**

**Rule: 7**

**========================================================================================**

**('Male',) ==> ('Middle-aged Adult', 'Old in current Dept.')**

**Support: 0.2032854209445585**

**Confidence: 0.396**

**Lift: 1.0804033613445378**

**Rule: 8**

**========================================================================================**

**('Team leader',) ==> ('Female',)**

**Support: 0.25256673511293637**

**Confidence: 0.5**

**Lift: 1.0274261603375527**

**Rule: 9**

**========================================================================================**

**('Male',) ==> ('Middle-aged Adult',)**

**Support: 0.2741273100616016**

**Confidence: 0.534**

**Lift: 1.0258698224852072**

**Rule: 10**

**========================================================================================**

**('Mid-level employee', 'Old in current Dept.') ==> ('Female',)**

**Support: 0.2166324435318275**

**Confidence: 0.49414519906323184**

**Lift: 1.015395409045544**

**Rule: 11**

**========================================================================================**

**('Mid-level employee',) ==> ('Female',)**

**Support: 0.2669404517453799**

**Confidence: 0.4896421845574388**

**Lift: 1.0061423792382813**

**Rule: 12**

**========================================================================================**

**('Male',) ==> ('Member of professional organizations',)**

**Support: 0.25051334702258726**

**Confidence: 0.48800000000000004**

**Lift: 1.0048879492600424**

**Rule: 13**

**========================================================================================**

**('Old in current Dept.',) ==> ('Female',)**

**Support: 0.3326488706365503**

**Confidence: 0.48868778280542985**

**Lift: 1.0041812245833095**

# **5. Association rules by varying parameters**

Iterating the above steps with minimum support of 0.26 and lift of 1, we found below frequent itemsets with gender and technical-soft skills:

**========================================================================================**

**Frequent Itemsets are as follows:**

**========================================================================================**

**Support | Itemsets**

**========================================================================================**

**0.5133470226 ('Male',)**

**0.4866529774 ('Female',)**

**0.272073922 ('Male', 'Fast working')**

**0.2689938398 ('Communicative', 'Male')**

**0.2618069815 ('Male', 'HTML CSS Java Script')**

Itemsets with gender and organizational exp. positions:

**========================================================================================**

**Frequent Itemsets are as follows:**

**========================================================================================**

**Support | Itemsets**

**========================================================================================**

**0.5133470226 ('Male',)**

**0.4866529774 ('Female',)**

**0.3480492813 ('Male', 'Old in current Dept.')**

**0.3326488706 ('Old in current Dept.', 'Female')**

**0.2782340862 ('Male', 'Mid-level employee')**

**0.2741273101 ('Middle-aged Adult', 'Male')**

**0.2669404517 ('Mid-level employee', 'Female')**

Association rules derived including gender and technical-soft skills are as follows:

**========================================================================================**

**Association Rules are as follows:**

**========================================================================================**

**Rule: 1**

**========================================================================================**

**('Male',) ==> ('Fast working',)**

**Support: 0.2720739219712526**

**Confidence: 0.53**

**Lift: 1.0513645621181265**

**Rule: 2**

**========================================================================================**

**('Male',) ==> ('Communicative',)**

**Support: 0.26899383983572894**

**Confidence: 0.524**

**Lift: 1.0106455445544555**

**Rule: 3**

**========================================================================================**

**('Male',) ==> ('HTML CSS Java Script',)**

**Support: 0.2618069815195072**

**Confidence: 0.5100000000000001**

**Lift: 1.0055465587044536**

The association rules formulated including gender and organizational exp. Position:

**========================================================================================**

**Association Rules are as follows:**

**========================================================================================**

**Rule: 1**

**========================================================================================**

**('Male',) ==> ('Middle-aged Adult',)**

**Support: 0.2741273100616016**

**Confidence: 0.534**

**Lift: 1.0258698224852072**

**Rule: 2**

**========================================================================================**

**('Mid-level employee',) ==> ('Female',)**

**Support: 0.2669404517453799**

**Confidence: 0.4896421845574388**

**Lift: 1.0061423792382813**

**Rule: 3**

**========================================================================================**

**('Old in current Dept.',) ==> ('Female',)**

**Support: 0.3326488706365503**

**Confidence: 0.48868778280542985**

**Lift: 1.0041812245833095**

# **6. Importance of Discovered Rules**

Association rules which were formulated, have unearthed the problem of gender disparity in the organization in several categories. On analyzing the association rules, we found that if you have a gender of “Male” then you are more likely to have more skills and organizational positions. Rules with minimum support of 0.2 and lift of 1are:

**('Male',) ==> ('Fast working',)**

**('Male',) ==> ('SQL Server',)**

**('Male',) ==> ('.Net',)**

**('Male',) ==> ('Communicative',)**

**('Male',) ==> ('HTML CSS Java Script',)**

**('Male', 'Mid-level employee') ==> ('Old in current Dept.',)**

**('Male', 'Old in current Dept.') ==> ('Mid-level employee',)**

**('Male', 'Old in current Dept.') ==> ('Middle-aged Adult',)**

**('Middle-aged Adult', 'Male') ==> ('Old in current Dept.',)**

**('Male',) ==> ('Middle-aged Adult', 'Old in current Dept.')**

**('Male',) ==> ('Middle-aged Adult',)**

**('Male',) ==> ('Member of professional organizations',)**

**('PHP mySQL',) ==> ('Female',)**

* If you are a “Male”, then you are more likely to have skills like – SQL Server, .Net, HTML CSS JavaScript, Communicative and Fast working with avg. confidence of 0.5 approx. However, females have more chances to have the skill of PHP MySQL. But, males are in a better position overall in technical skills than females.

**('Male',) ==> ('Fast working',)**

**Support: 0.2720739219712526**

**Confidence: 0.53**

**Lift: 1.0513645621181265**

**('Male',) ==> ('SQL Server',)**

**Support: 0.25256673511293637**

**Confidence: 0.4920000000000001**

**Lift: 1.0394967462039049**

**('Male',) ==> ('.Net',)**

**Support: 0.2515400410677618**

**Confidence: 0.49**

**Lift: 1.0263655913978496**

**('Male',) ==> ('Communicative',)**

**Support: 0.26899383983572894**

**Confidence: 0.524**

**Lift: 1.0106455445544555**

**('Male',) ==> ('HTML CSS Java Script',)**

**Support: 0.2618069815195072**

**Confidence: 0.5100000000000001**

**Lift: 1.0055465587044536**

* In organizational positions, you are more likely to be a member of professional organizations with the confidence of 0.48 approx if you are a male than a female. Also, a male who is a mid-level employee is more likely to be working in the current department for a long time with the confidence of 0.7 approx than a female; which implies that the females are more likely to either move from the departments or leave the job.

**('Male', 'Mid-level employee') ==> ('Old in current Dept.',)**

**Support: 0.22176591375770022**

**Confidence: 0.7970479704797049**

**Lift: 1.170927184384966**

**('Male',) ==> ('Middle-aged Adult', 'Old in current Dept.')**

**Support: 0.2032854209445585**

**Confidence: 0.396**

**Lift: 1.0804033613445378**

**('Male',) ==> ('Member of professional organizations',)**

**Support: 0.25051334702258726**

**Confidence: 0.48800000000000004**

**Lift: 1.0048879492600424**

* From the above inferences, we can conclude that there is a need to take measures within the organization so that all the employees are given equal opportunities to excel. The recommendation that we can make is to provide more training opportunities to the females in the technical area. Also, adopt policies to favour female employees to work for a long time in the current department of the organization. Find Out the reasons for female employees to establish gender equality within the organization.