1. Introduction
2. Info about the dataset
3. Loading the Dataset
4. Extracting basic nodes out of it
   1. Department
   2. Job Quit
   3. Business Travel
   4. Job Role
   5. Education
   6. Marital Status
   7. Status of Leaving
   8. Average Income
5. Data Model
6. Extracting important information using properties of nodes
   1. Income <> Average for that role
   2. Job Satisfaction
   3. Average Tenure
7. Univariate Analysis
8. Deep dive into the Multi Variate Analysis
   1. Income <> Average for that role
      1. Status of Leaving
      2. Job Mode
      3. Overtime
   2. Job Satisfaction
      1. Status of Leaving
      2. Job Mode
      3. Overtime
   3. Average Tenure
      1. Attrition
      2. Tenure in current organisation
9. Other insights about the data
   1. Distance from home
   2. Source of Hire
   3. Marital Status
10. Similarity of nodes (Employees) – Attrition
11. Future Work
12. Links to the work done (Cypher Commands and Neo4j Dump)
13. Introduction

In day to day most of the data science problems are analysed using several technologies available by either extracting information out of it or injecting the data into the platforms that can do a job smoothly. Neo4j is one such kind of platform where data can be loaded and multiple data extraction and visualisation process can be done such as entity extraction data comparison similarity between the nodes univariate analysis multivariate analysis and so on.

This document describes the extraction of some insights about the Employee Attrition using the [Employee Attrition Dataset](https://gist.github.com/jexp/e329b8cb8fdb9176ae67991d3e7d4941) in Neo4j. This dataset is used for the training of Machine Learning Model for Employee Attrition Prediction. In our case we use this dataset to load into Neo4j and explore some insights about the dataset.

1. Info about the dataset

This dataset contains some core information about the employee such as Gender, Education, Age, Attrition, Job Satisfaction, Date of Joining the organisation and etc. Below is the full list of features names available in the dataset.

[ Age, Attrition, Business Travel, Department, Distance From Home, Gender, Job Involvement, Job Level, Job Role, Job Satisfaction, Marital Status, Monthly Income, Number Companies Worked, Over Time, Percent Salary Hike, Performance Rating, Stock Option Level, Total Working Years, Training Times Last Year, Years At Company, Years Since Last Promotion, Years With Current Manager, Higher Education, Date of Hire, Date of termination, Status of leaving, Mode of work, Leaves, Absenteeism, Work accident, Source of Hire, Job mode ]

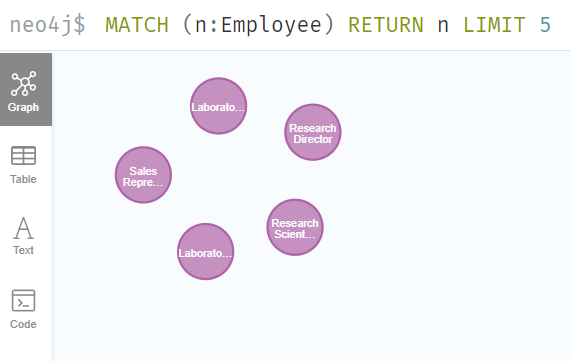
The size of the dataset is (1470, 33) with the above-mentioned list as the features for each employee record.

1. Loading the Dataset

As part of the first step in the exploration we need to load the dataset into our GraphDB (Neo4j). Neo4j allows us to load the datasets which are in csv format directly without making any additional changes to it. The command for the same is given below:

This dataset has the header row as the first row, so it can collect the node name and properties names directly from the header row. As the dataset doesn’t have the unique serial column, we can use the `linenumber` as the unique number to decide the row of the csv file. Using the above file, `Employee` Nodes are created and the rest of the columns (features) are used as the properties of the corresponding Employee node.

Once we execute the above statement, it will create all the employee nodes using the employee records. Below is the sample records from graphDB.



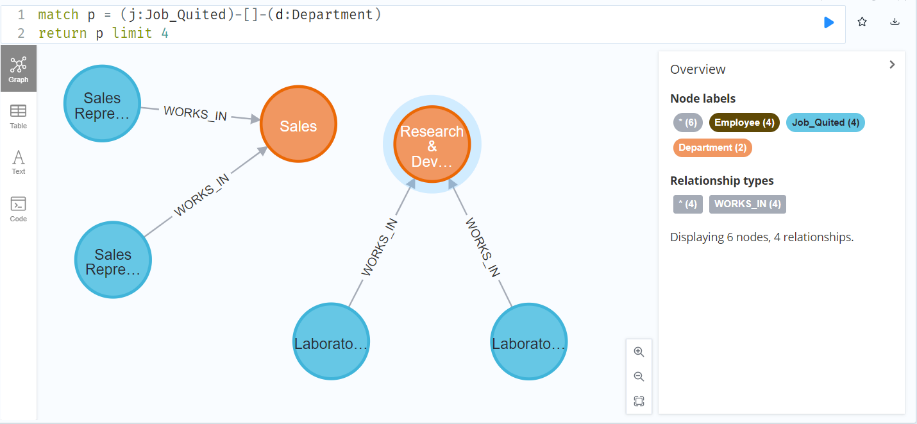
1. Extracting basic nodes out of it

Now that we have the employee information as the nodes and properties, we can extract basic information and create separate labels for the same. It would help us identify the relationships between each node and dig deep in to the data.

* 1. Department

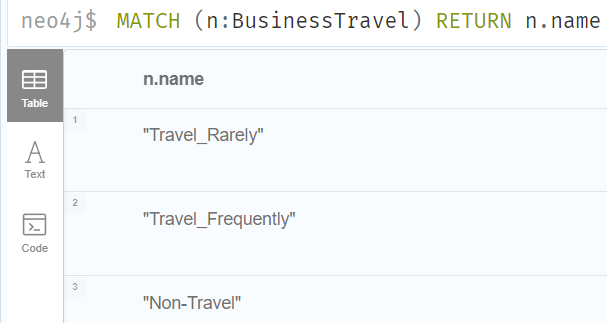
Entire dataset had 3 departments in total. All the employees were working under these three departments:

* 1. Job Quit

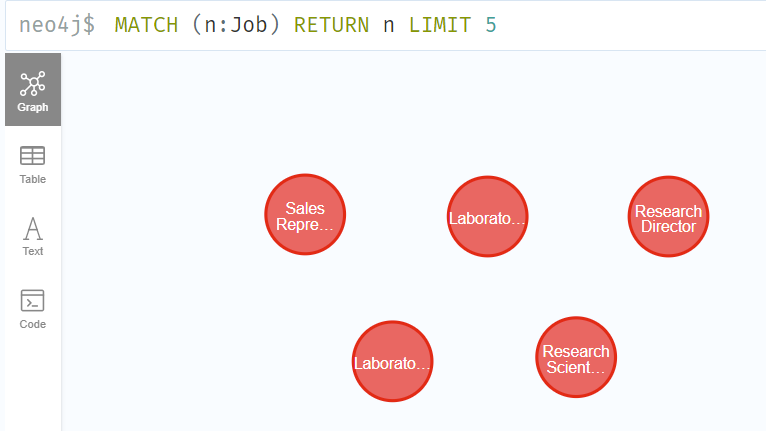
In the similar way we create separate labels for Attrition as Job Quit. This is our dependent variable. This label can be used to draw insights using multivariate analysis.

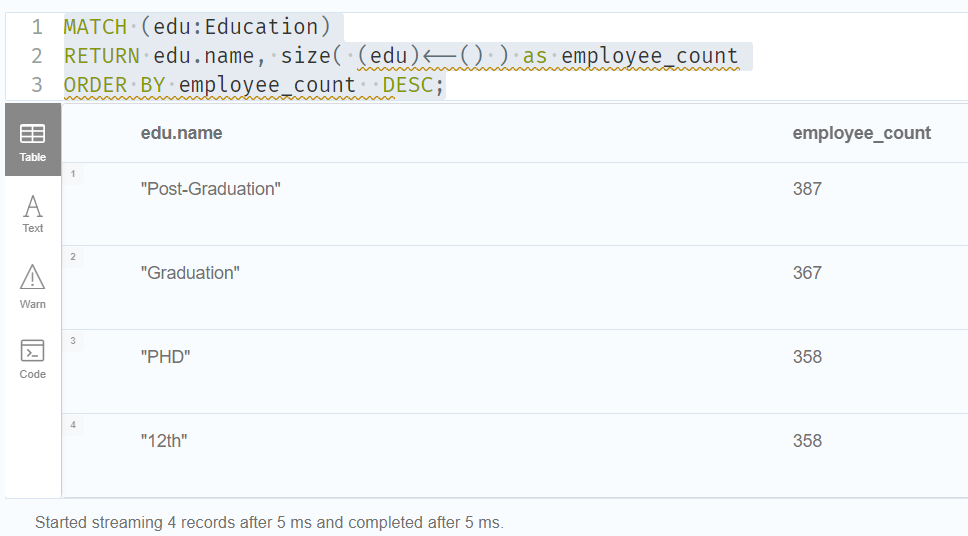
Once we have the label created, we can visualise the results of the employees who left in particular department as shown in the above image

* 1. Business Travel

The influence of business travel on employee attrition can be observed sometimes. To visualise this, we shall create a node for Business Travel and assign relationships for the same. There are three categories under Business Travel as mentioned below:

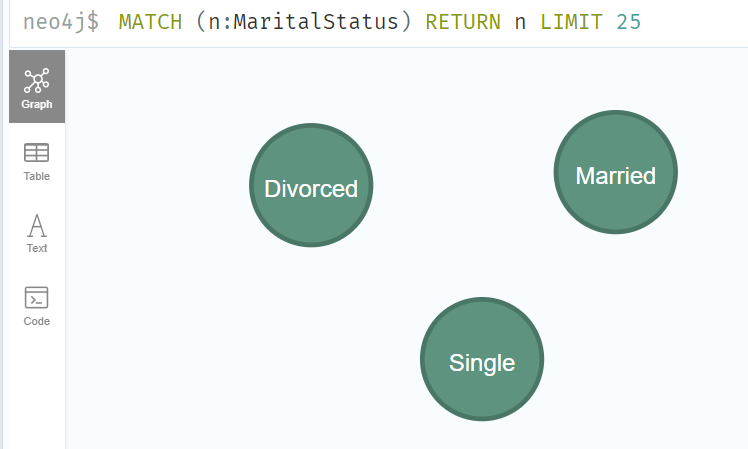
* 1. Job Role

Each employee holds a position, we can differentiate the employee by using the job role in a combination with other features. So, grouping this as node would help us in further exploration.

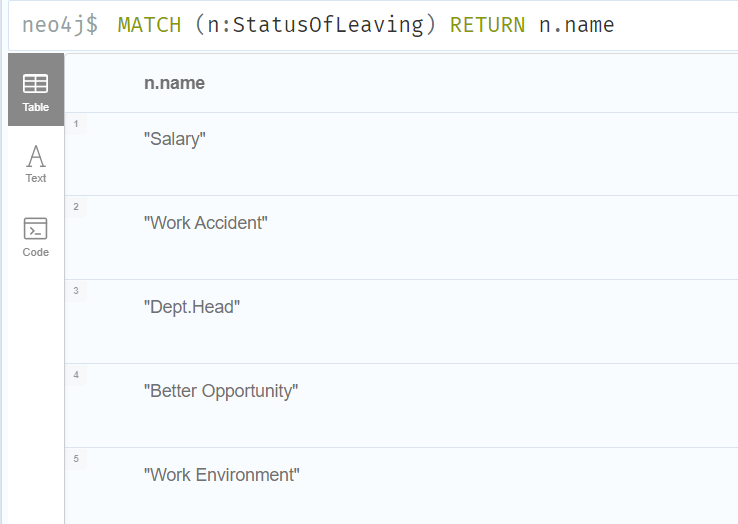
* 1. Education

Distinguishing an employee property depends on various aspects, education plays an important role in the kind of work an employee is doing. So, creating separate nodes/label for this category is a best idea.

* 1. Marital Status

Marital status influences the decision-making process of the attrition.

* 1. Status of Leaving

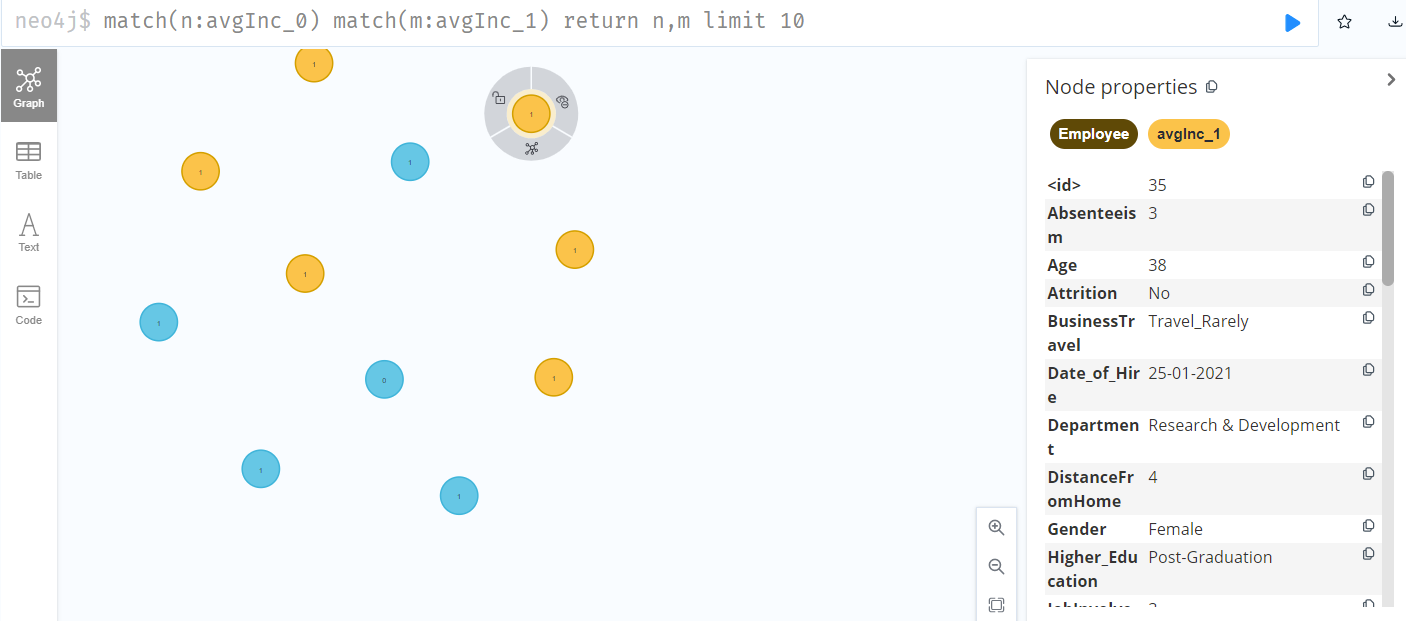
Status of leaving can be used to better understand the reason behind the attrition of the employee.

* 1. Average Income

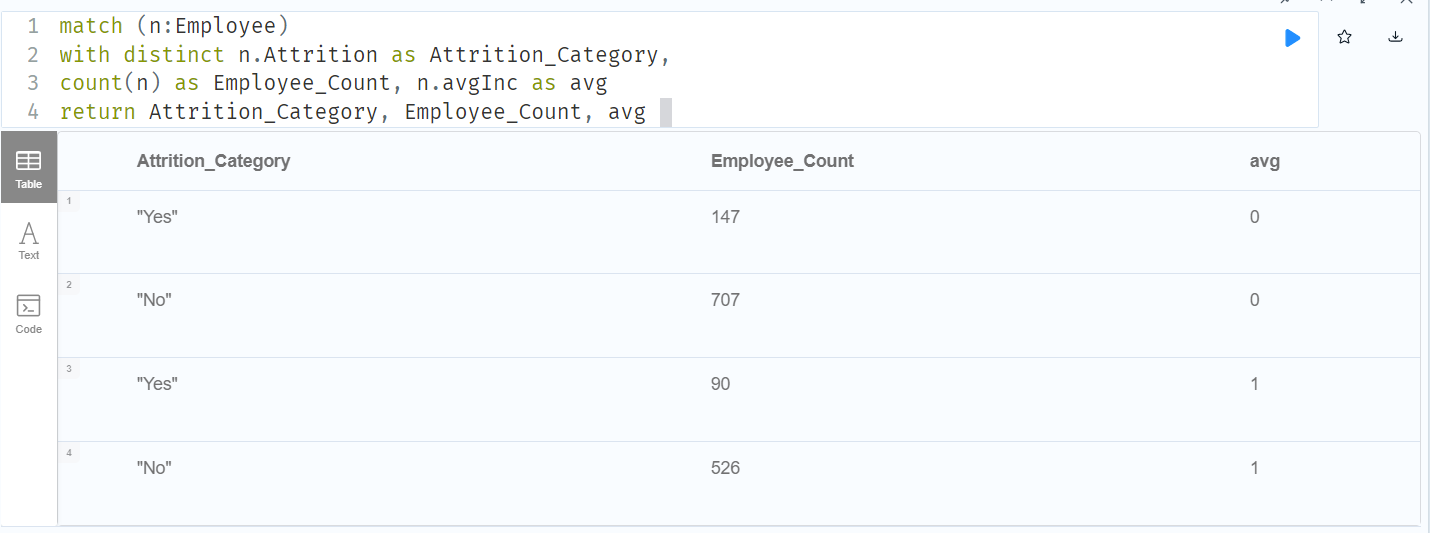
Average income for given role matters a lot in the attrition. Grouping them using the role and calculating the average income for respective role allows us to understand the employee’s satisfaction in terms on Income.

For an example:

If the employee’s Income is more than the average of his role’s Income – then he is most likely to be satisfied in terms of income and vice versa.

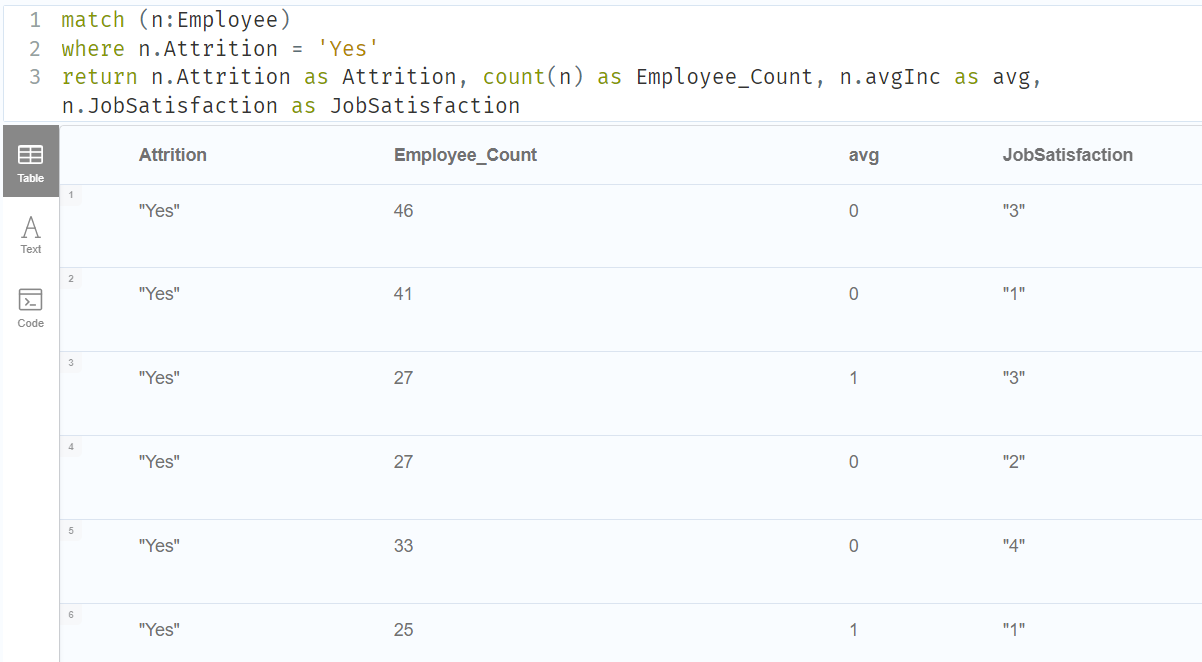


1. Data Model
2. Extracting important information using properties of nodes – Analysis
   1. Income <> Average for that role

Income plays an important role in employee’s work functionality. Most of the attritions happen because of the lower income rates for given role. For our instance let’s check if that is a conclusion or a myth.

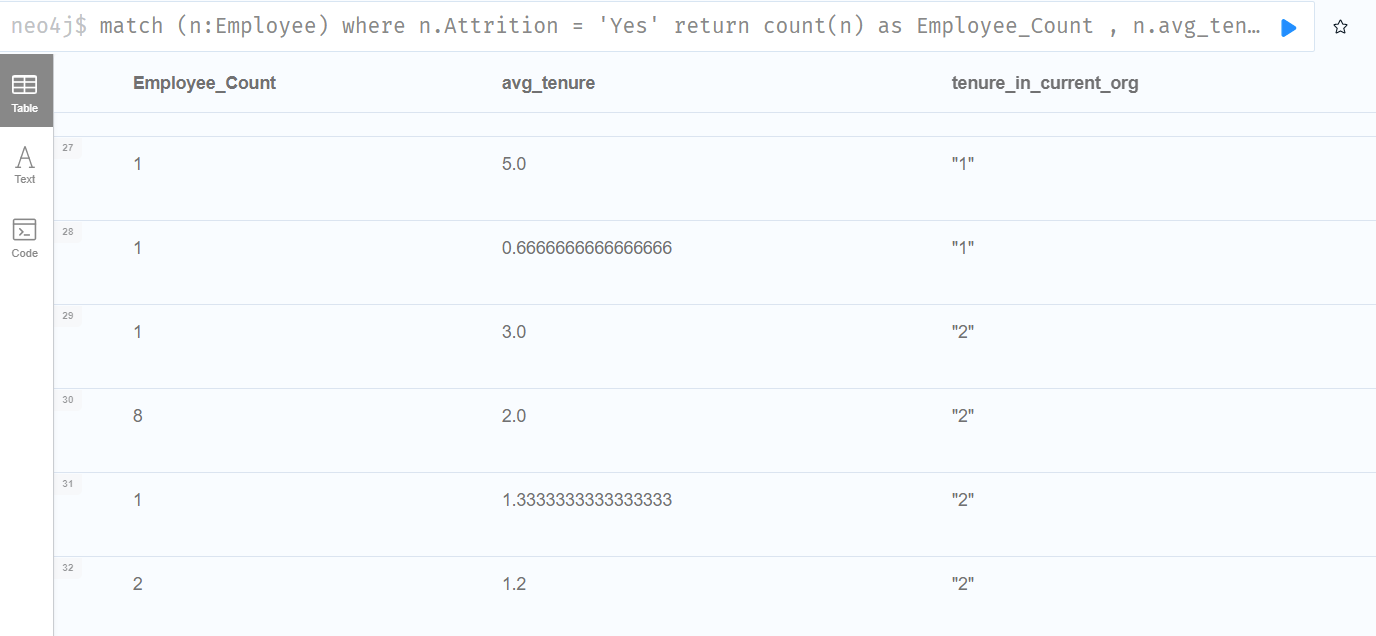
We can see that employee count who left the organisation is more for the category for whom the income is less than the average income of the respective role. ( 147 count for less than average and 90 for more than average ). So, assumption we made for the attrition is corret in this case.

* 1. Job Satisfaction

Along with the income, job satisfaction also influences the attrition risk for a given employee. In some cases, even if the employee is paid well and he/she is not satisfied with the current job – there are high chances of leaving the organisation.

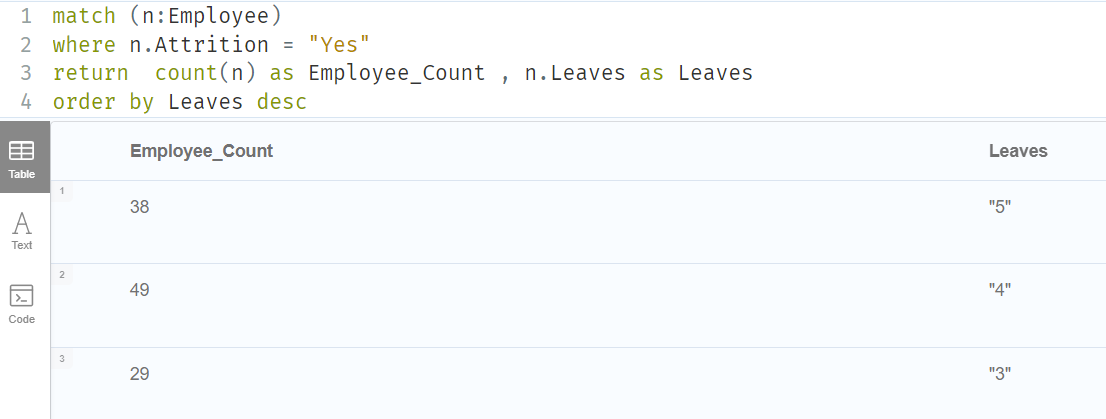
If we can observe the above table, we can see that, the highest number of employee count who left the organisation are with Income less than average income for that role and less job satisfaction. In the other case where the job satisfaction is higher and Income is less than average income, employees tend to retain in the organisation.

* 1. Average Tenure

Average tenure describes as the average time worked in an organisation; this shows how frequently an employee is switching the jobs. Employee tend to leave the organisation once some optimum tenure at the company is reached and it might also get impacted along with the other features such as average Income, Job Satisfaction..etc.

* 1. Leaves used

Employees tends to use all the unused leaves before leaving the organisation (in case of no leave encashment - Generally). So, it can be understood that the usage of leaves in high number might result an outcome of attrition.



* 1. Performance Rating
  2. Marital Status
  3. Gender
  4. Business Travel
  5. Distance from home
  6. Job Level
  7. Overtime

1. Deep dive into the Multi Variate Analysis
   1. Income <> Average for that role
      1. Status of Leaving
      2. Job Mode
      3. Overtime
   2. Job Satisfaction
      1. Status of Leaving
      2. Job Mode
      3. Overtime
   3. Average Tenure
      1. Attrition
      2. Tenure in current organisation
2. Other insights about the data
   1. Distance from home
   2. Source of Hire
   3. Marital Status
3. Similarity of nodes (Employees) – Attrition
4. Future Work
5. Links to the work done (Cypher Commands and Neo4j Dump)