Source code for this entire problem is checked in to Github URL:

Code to load the dataset into RDD/DataFrame by reading the excel

**val** sc = SparkSession  
 .*builder*()  
 .appName(**""**)  
 .config(**"spark.master"**,**"local[\*]"**)  
 .getOrCreate()  
  
**val** empFile = **"C:\\Krishnan\\Working\\BigData\\Learnings\\Test\\Matt\\201-Assignments\\Assignments\\Data\\EmpDatasets.xlsx"***//Defining the schema for the Dataset which is to be loaded from Excel dataset***val** empSchema = *StructType*(  
 *List*(*StructField*(**"satisfaction\_level"**, FloatType, nullable = **true**),  
 *StructField*(**"last\_evaluation"**, FloatType, nullable = **true**),  
 *StructField*(**"number\_project"**, IntegerType, nullable = **true**),  
 *StructField*(**"average\_monthly\_hours"**, IntegerType, nullable = **true**),  
 *StructField*(**"time\_spend\_company"**, IntegerType, nullable = **true**),  
 *StructField*(**"work\_accident"**, IntegerType, nullable = **true**),  
 *StructField*(**"left"**, IntegerType, nullable = **true**),  
 *StructField*(**"promotion\_last\_5years"**, IntegerType, nullable = **true**),  
 *StructField*(**"department"**, StringType, nullable = **true**),  
 *StructField*(**"salary"**, StringType, nullable = **true**))  
)  
  
**val** employeesDF = sc.*sqlContext*.read.schema(empSchema).format(**"com.crealytics.spark.excel"**)  
 .option(**"sheetName"**, **"Sheet1"**)  
 .option(**"useHeader"**, **"true"**)  
 .option(**"treatEmptyValuesAsNulls"**, **"false"**)  
 .option(**"inferSchema"**, **"false"**)  
 .option(**"location"**, **"C:\\Krishnan\\Working\\BigData\\Learnings\\Test\\Matt\\201-Assignments\\Assignments\\Data\\EmpDatasets.xlsx"**)  
 .option(**"addColorColumns"**, **"False"**)  
 .load(**"C:\\Krishnan\\Working\\BigData\\Learnings\\Test\\Matt\\201-Assignments\\Assignments\\Data\\EmpDatasets.xlsx"**)

**Problem Statements:**

1. What is the most important criteria for an employee to stick to an organization?

(Hint: Implement/use correlation matrix, Output will be in string format

Ex: Salary/Experience/Department etc.)

Code:

employeesDF.stat.corr(**"satisfaction\_level"**,**"left"**)  
*println*(**"Correlation between satisfaction\_level and left : "** + employeesDF.stat.corr(**"satisfaction\_level"**,**"left"**))

Output:

Correlation between satisfaction\_level and left: -0.38820046701245925

The -ve correlation value result indicates that when the satisfaction level decreases then more chances of employees leaving the organization and vice-versa.

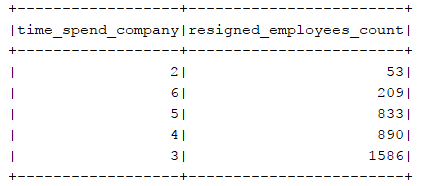
Output snapshot URL:

2. At what experience level in a company employees are more susceptible to resign?

(Output will be in number format and as graph)

Code:

Output:

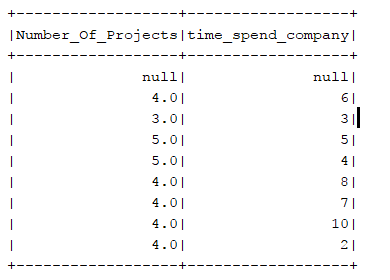


3. What is the ideal number of projects for an employee?

(Output will be in number format as well as graph)

Code:

Output:

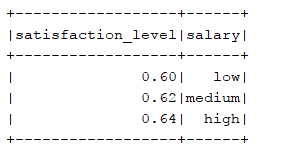


4. How important is hike to employees

(Output will be in format of string: Ex: Salary less important, Salary very important, Experience important etc.)

Code:

Output:



5. Which group (Department) has highest attrition (leaving the company)

(Output as graph)

Code:

Output:

