Corporate Employee Attrition Analysis

A PROJECT REPORT

Submitted By

Team ID : PNT2022TMID44764

Priyadharshini P-(732719104031)

Priyadharshini M-(732719104030)

Kaviya Sri S-(732719104016)

Narmatha M-(732719104023)

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE AND ENGINEERING

SRI SHANMUGHA COLLEGE OF ENGINEERING AND TECHNOLOGY

SANKARI.

ANNA UNIVERSITY: CHENNAI-600025

JUNE 2022

TABLE OF CONTENTS

1. INTRODUCTION

1.1 Project overview

Employee attrition has become a vital problem across the world. It is one of the crucial issues faced by business leaders within companies where they lose the most talented employees. A good employee is always an asset to the organization and their resignation can lead to various problems like financial losses, overall performance, and loss of acquired knowledge. Furthermore, hiring new employees is far exorbitant, taxing, and time-consuming in comparison to recruiting the existing one. It is very time-consuming to recruit a new employee as it takes him months for training, adjusting to the culture, rules, and environment. Therefore, upcoming trends and technology using Machine Learning Algorithms must be exploited for the benefit of business organizations. Knowing the reason beforehand for the employee attrition, companies can mitigate this loss. This analysis provides a conclusive review of employee attrition from the data set IBM HR Analytics Employee Attrition Performance.

1.2 Purpose

[1] Hardik P. K. (2016) , researched on “a study on employee attrition: with special reference to Kerala IT Industry”. His research examined the relationship between organizational factors and attrition of IT professional’s. The result can conclude that the organizational factors played significant role in predicting the variance in turnover intention (attrition) of Kerala IT professionals. Therefore, the HR managers in IT organizations may take into consideration the problems with organizational

factors of their workers to reduce the turnover intention of the skilled employees**.**

2.LITERATURE SURVEY

2.1 Existing Problem

The Existing system includes only few attributes for analysis and also deals with qualitative observations and simple statistical analysis.The qualitative observations deal with data and can be observed through human senses.They do not involve measurements or number. Due to the increase in IOT and connected device,we now have access to so much of data and along with it an increase needs to manage and understand data.

2.2 References

1. From Big Data to Deep Data to support people analytics for employee attrition prediction, Nesrine Ben Yahia, Hlel Jihen, Ricardo Colomo-Palacio(2021)

2.Machine Learning Approach for Employee Attrition Analysis.Dr. R. S. Kamath | Dr. S. S. Jamsandekar | Dr. P. G. Naik ,Published in International Journal of Trend in Scientific Research and Development (ijtsrd), (March 2019)

3. Investigation of early career teacher attrition(ECT) and the impact of induction programs in Western Australia, Janine E.Wyatt, MichaelO'Neill (2021)

2.3 Problem Statement Definition

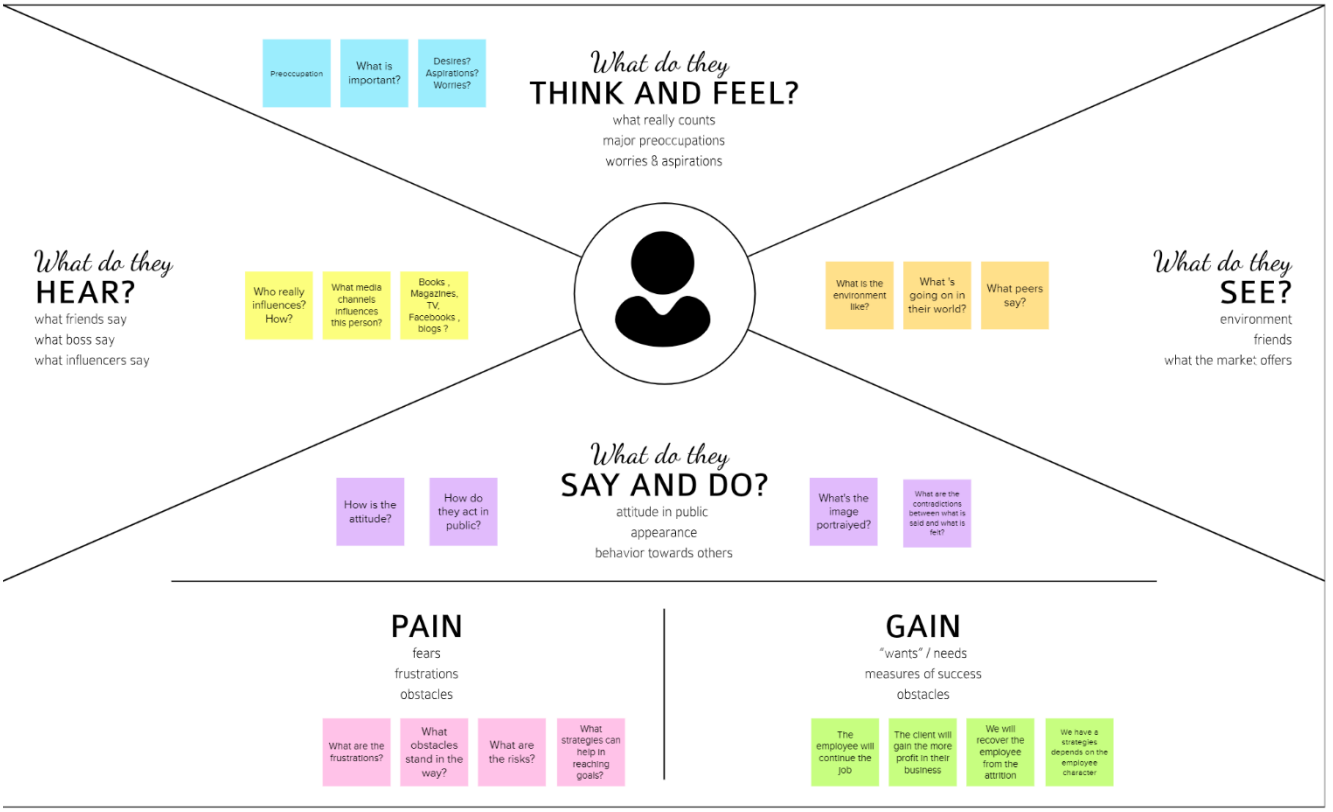
● To create a dashboard and perform analysis of employee attrition in corporates using IBM Cognos analytics platform.

● To reduce the employee attrition rate through data analytics,

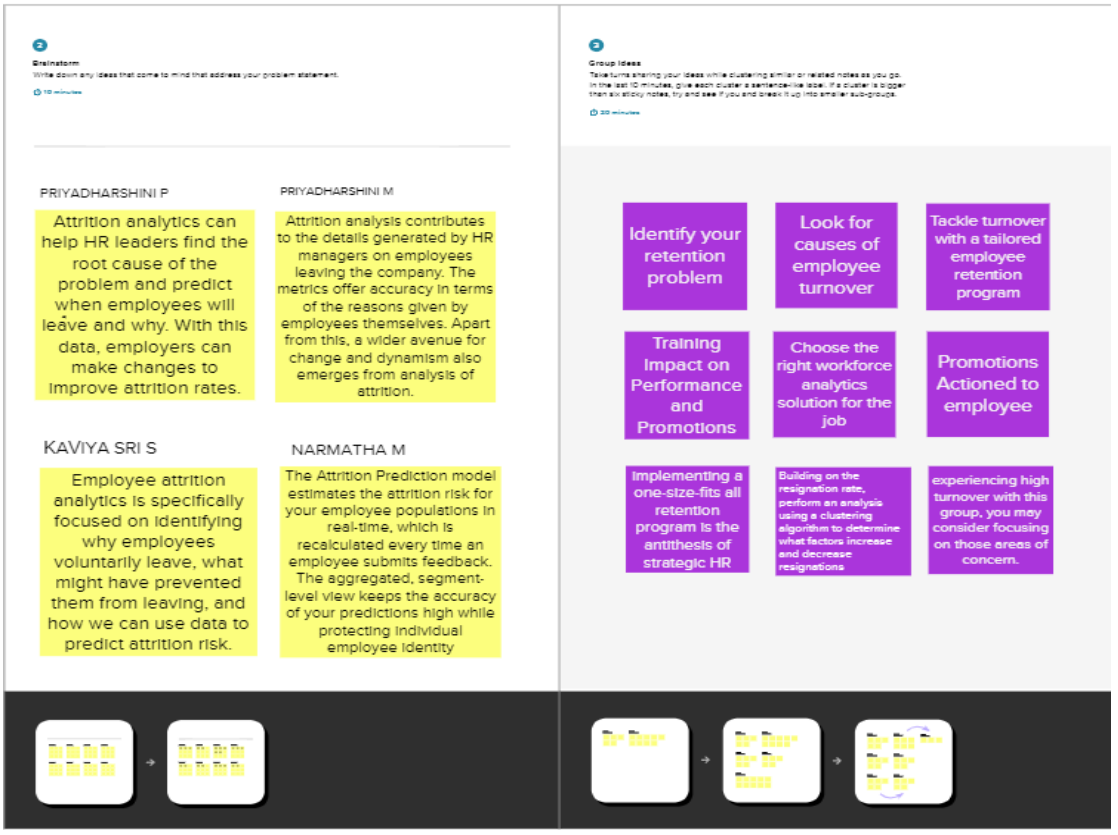
data visualization by analysing the major factors that causes attrition.

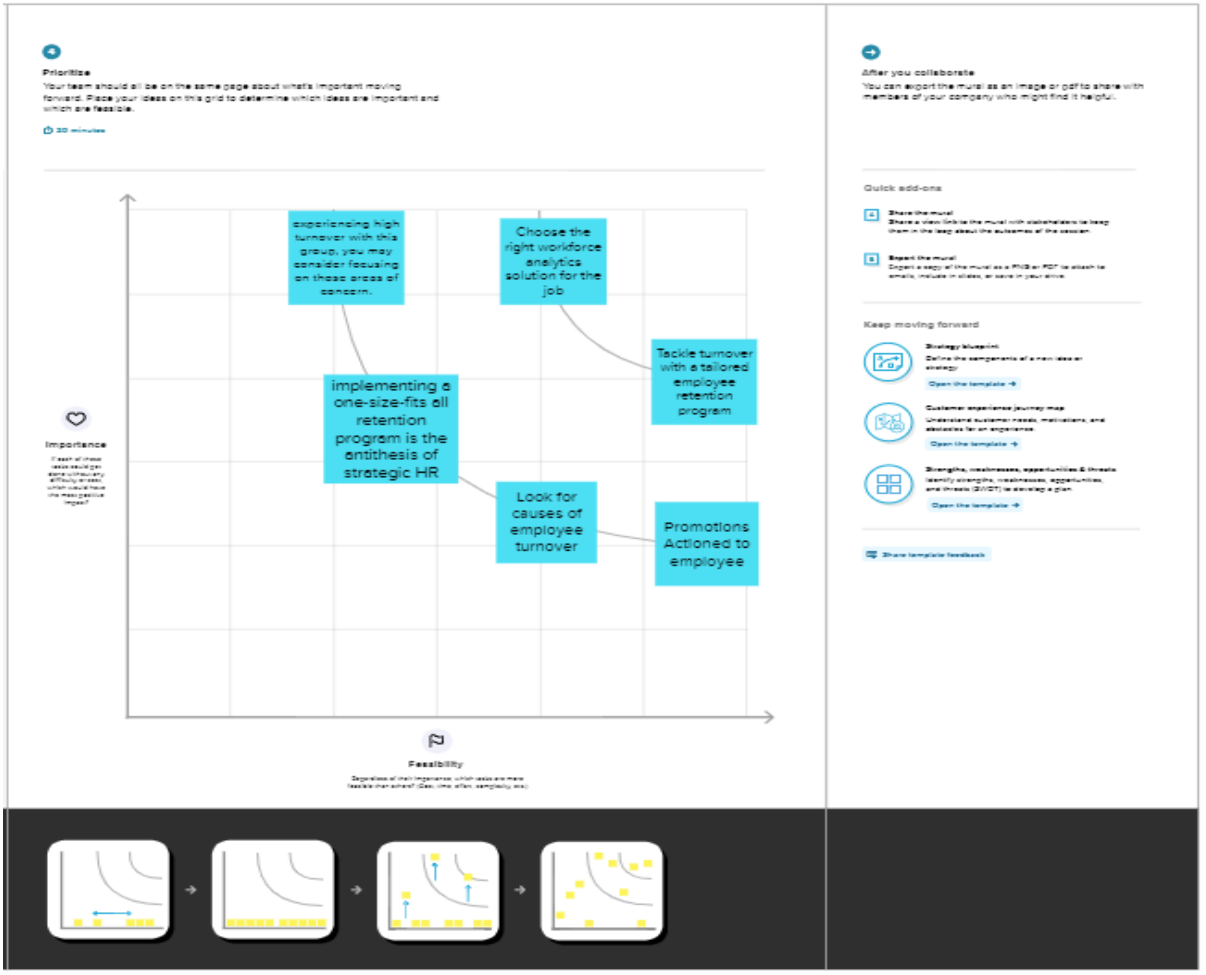
3.IDEATION AND PROPOSED SOLUTION

3.1 Empathy Map



3.2 Ideation & Brainstorming


Brainstorm
Write down any ideas that come to mind that address your problem statement.
10 minutes
PRIYADHARSHINI P
Attrition analytics can help HR leaders find the root cause of the problem and predict when employees will leave and why. With this data, employers can make changes to improve attrition rates.
PRIYADHARSHINI M
Attrition analysis contributes to the details generated by HR managers on employees leaving the company. The metrics offer accuracy in terms of the reasons given by employees themselves. Apart from this, a wider avenue for change and dynamism also emerges from analysis of attrition.
KAVIYA SRI S
Employee attrition analytics is specifically focused on identifying why employees voluntarily leave, what might have prevented them from leaving, and how we can use data to predict attrition risk.
NARMATHA M
The Attrition Prediction model estimates the attrition risk for your employee populations in real-time, which is recalculated every time an employee submits feedback. The aggregated, segment-level view keeps the accuracy of your predictions high while protecting individual employee identity
Group ideas
Take turns sharing your ideas while clustering similar or related notes as you go. In the last 10 minutes, give each cluster a sentence-like label. If a cluster is bigger than six sticky notes, try and see if you and break it up into smaller sub-groups.
20 minutes
Identify your retention problem
Look for causes of employee turnover
Tackle turnover with a tailored employee retention program
Training Impact on Performance and Promotions
Choose the right workforce analytics solution for the job
Promotions Actioned to employee
Implementing a one-size-fits all retention program is the antithesis of strategic HR
Building on the resignation rate, perform an analysis using a clustering algorithm to determine what factors increase and decrease resignations
experiencing high turnover with this group, you may consider focusing on those areas of concern.


Prioritize
Your team should all be on the same page about what's important moving forward. Place your ideas on this grid to determine which ideas are important and which are feasible.
20 minutes
experiencing high turnover with this group, you may consider focusing on those areas of concern.
Choose the right workforce analytics solution for the job
Tackle turnover with a tailored employee retention program
implementing a one-size-fits all retention program is the antithesis of strategic HR
Look for causes of employee turnover
Promotions Actioned to employee
Importance
Feasibility
After you collaborate
You can export the mural as an image or pdf to share with members of your company who might find it helpful.
Quick add-ons
Share the mural
Share a view link to the mural with stakeholders to keep them in the loop about the outcomes of the session.
Export the mural
Export a copy of the mural as a PNG or PDF to attach to emails, include in slides, or save in your drive.
Keep moving forward
Strategy blueprint
Define the components of a new idea or strategy.
Open the template →
Customer experience journey map
Understand customer needs, motivations, and obstacles for an experience.
Open the template →
Strengths, weaknesses, opportunities & threats
Identify strengths, weaknesses, opportunities, and threats (SWOT) to develop a plan.
Open the template →
Share template feedback

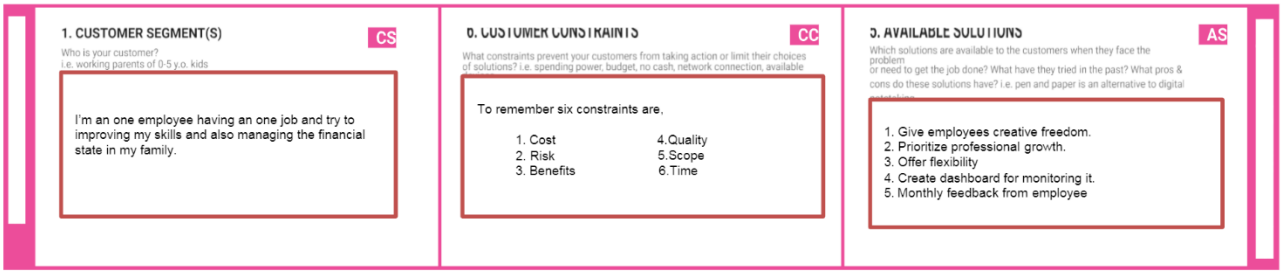
3.3 Proposed Solution

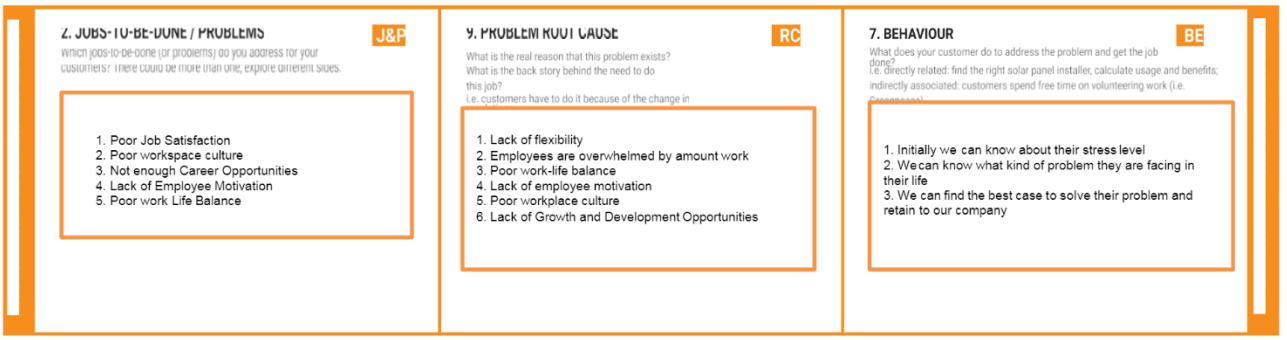
The Existing system includes only few attributes for analysis and also deals with qualitative observations and simple statistical analysis. The qualitative observations deal with data and can be observed through human senses.They do not involve measurements or number. Due to the increase in IOT and connected device,we now have access to so much of data and along with it an increase needs to manage and understand data.

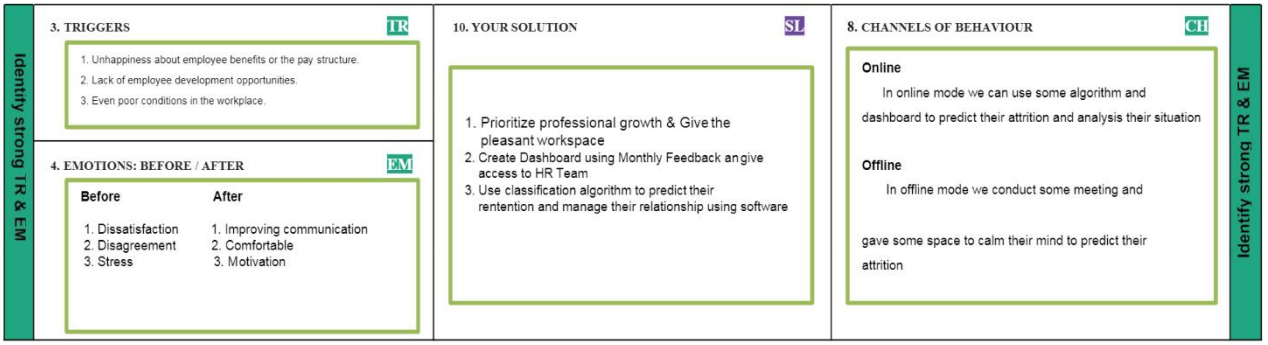
3.4 Problem Solution fit

Project Design Phase-I - Solution Fit

Project Title : Corporate Employee Attrition Analysis
Team ID : PNT2022TMID44764







4.REQUIREMENT ANALYSIS

4.1 Functional requirement

FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR-1	User Registration	Registration through Form Registration through Gmail Registration through LinkedIN
FR-2	User Confirmation	Confirmation via Email Confirmation via OTP
FR-3	User Feedback	Feedback through Form Feedback through Gmail Feedback through Instagram polls Feedback through LinkedIn
FR-4	User Rating	Rating via Mail Rating through Message
FR-5	Employee Management	Validating and managing the employee details
FR-6	Attrition Analytics	Analysing and finding out the major reason for the attrition of employees using dataset

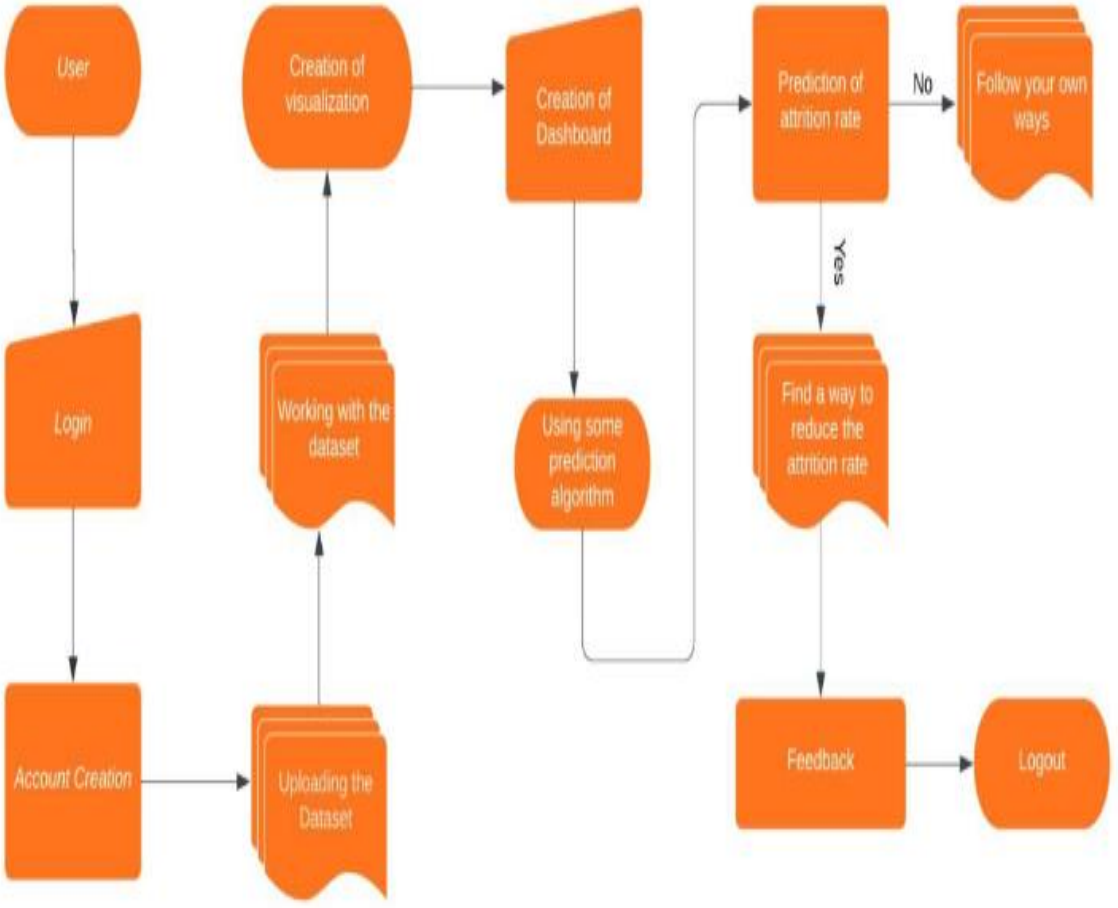
4.2 Non-Functional requirements

Following are the non-functional requirements of the proposed solution.

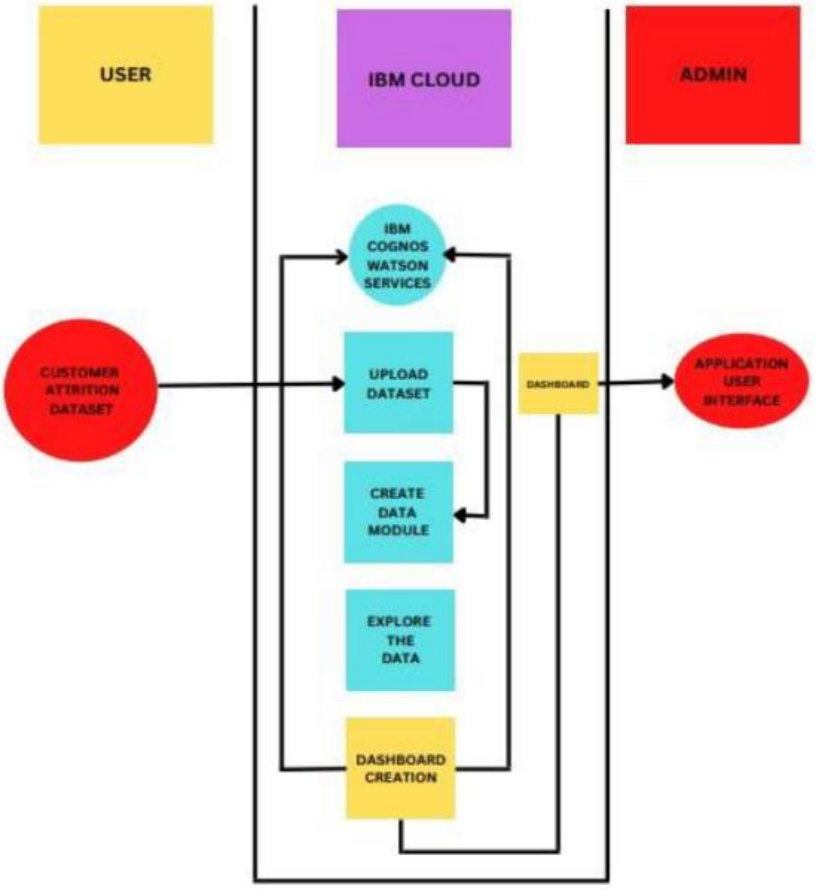
FR No.	Non-Functional Requirement	Description
NFR-1	**Usability**	This Data Visualization shall be easy to use for all users with minimal instructions. 100% of the languages on the graphical user interface (GUI) shall be intuitive and understandable by non-technical users.
NFR-2	**Security**	The employee data is kept secure and their identity is hidden for the organization.
NFR-3	**Reliability**	The Link shall be operable in all conditions. The system must be less prone to errors
NFR-4	**Performance**	This software is portable and inter-operable. It works smoothly without generating errors. It also provides a faster response
NFR-5	**Portability**	The link shall be portable to all operating platforms. Therefore, this link should not depend on the different operating systems.
NFR-6	**Scalability**	Our solution is scalable for large and small datasets. It provides an efficient solution despite the size of the dataset.

5.PROJECT DESIGN

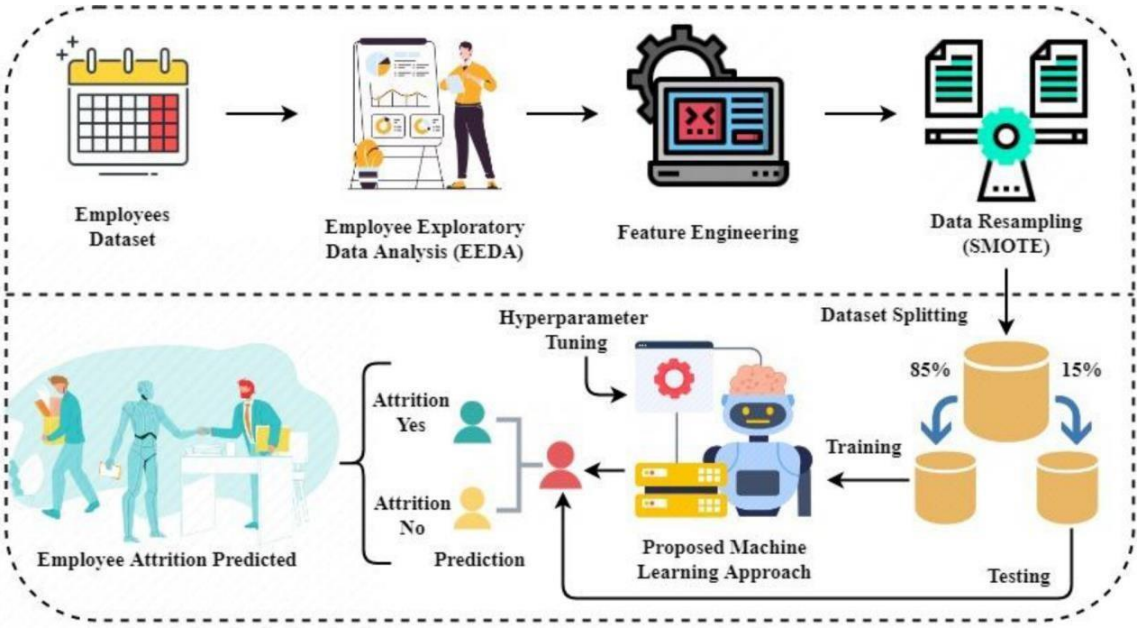
5.1 Data Flow Diagrams



5.2 Solution & Technical Architecture



6



5.3 User Stories

User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
Customer (Web user)	Registration	USN-1	As a user, I can register for the application by entering my email, password, and confirming my password.	I can access my account / dashboard	High	Sprint-1
		USN-2	As a user, I will receive confirmation email once I have registered for the application	I can receive confirmation email & click confirm	High	Sprint-1
		USN-3	As a user, I can register for the application through Facebook	I can register & access the dashboard with Facebook Login	Low	Sprint-2
		USN-4	As a user, I can register for the application through Gmail	I can register & access the dashboard with Gmail Login	Medium	Sprint-1
	Login	USN-5	As a user, I can log into the application by entering email & password	I can access my account / dashboard	High	Sprint-1
	Dashboard	USN-6	Uploading the Dataset	I can be able to upload my dataset	High	Sprint-2
		USN-7	Working With Dataset	I can be able to access my dashboard	High	Sprint-2
		USN-8	Visualization	I can be able to view the visual attrition rate of my dataset	High	Sprint-3
		USN-9	Working with Dashboard	I can be able to view the various views of the attrition rate	High	Sprint-3
Customer Care Executive		USN-10	Asking Help / Feedback	I can be able to ask help if I can face any issues or problems while using the webpage	Medium	Sprint-4
Administrator		USN-11	Managing the Database	I can assure that my data is in secure state	High	Sprint-4
		USN-12	Managing the over all process	I can assure that my data and process is going good	High	Sprint-4

6.PROJECT PLANNING

6.1Sprint Planning & Estimation

6	Project Development Phase		
Activity number	Activity name	Detailed activity description	
6.1	Coding & Solutioning	Sprint-1 Delivery: Develop the Code, Test and push it to GitHub.	Sankar Raja J I, Jaison V, Mathan M, Pasupathish M
6.2	Acceptance Testing	Sprint-2 Delivery: Develop the Code, Test and push it to GitHub. Sprint-3 Delivery: Develop the Code, Test and push it to GitHub.	Sankar Raja J I, Jaison V, Mathan M, Pasupathish M
6.3	Performance Testing	Sprint-4 Delivery: Develop the Code, Test and push it to GitHub.	Sankar Raja J I, Jaison V, Mathan M, Pasupathish M

6.2 Sprint Delivery Schedule

Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date (Planned)	Story Points Completed (as on Planned End Date)	Sprint Release Date (Actual)
Sprint-1	20	6 Days	24 Oct 2022	29 Oct 2022	29 OCtober 2022	05 Novembet 2022
Sprint-2	20	6 Days	31 Oct 2022	05 Nov 2022	05 November 2022	06 November 2022
Sprint-3	20	6 Days	07 Nov 2022	12 Nov 2022	08 November 2022	09 November 2022
Sprint-4	20	6 Days	14 Nov 2022	19 Nov 2022	11 November 2022	16 November 2022

7.CODING & SOLUTIONING

7.1 Feature 1

```
#GENERAL
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
path = '/content/general_data.csv'
df =pd.read_csv(path)
```

```
df

df.shape

df.info()
df.select_dtypes('int64' ,'float64').columns
cat_cols = df.select_dtypes('object').columns
cat_cols
df.describe().T
df
for cat in cat_cols:
   print(cat ,'-> ' , df[cat].unique())
   print()
print("All columns Unique values count")
for col in df:
   print(col, len(df[col].unique()), sep=': ')
plt.figure(figsize =(14,5))
plt.subplot(1,2,1)
sns.countplot(df['Attrition'] ,color ='b' ,hue =df['Gender'])
plt.title('Attrition by Gender')
plt.subplot(1,2,2)
plt.pie(df['Attrition'].value_counts() ,colors =['r' ,'c'] ,explode =[0,0.1] ,autopct =
'%.2f' ,labels =['No' ,'Yes'])
plt.title('Attrition')
#HANDLING CATEGORICAL OUTPUT VARIABLE
df['Attrition'].replace({'Yes':1 ,'No':0} ,inplace = True)
df['Attrition'].head()
plt.figure(figsize =(20 ,8))
sns.boxplot(x ='JobRole', y = 'MonthlyIncome' ,data = df ,hue ='Attrition' ,color ='red')

col = ['YearsInCurrentRole' ,'YearsSinceLastPromotion' ,'YearsWithCurrManager'
,'YearsAtCompany']
plt.figure(figsize = (10 ,10))
for i,c in enumerate(col):
   plt.subplot(2 ,2,i+1)
   sns.distplot(df[c] ,color ='b')
```

7.2 Feature 2

```
#GENERAL
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
#FEATURE ENGINEERING
from sklearn.preprocessing import LabelEncoder
from imblearn.over_sampling import SMOTE
path = '/content/general_data.csv'
df =pd.read_csv(path)
df
df.shape
df.info()
df.select_dtypes('int64' ,'float64').columns
```

```
cat_cols = df.select_dtypes('object').columns
cat_cols
df.describe().T
df
for cat in cat_cols:
    print(cat ,'-> ' , df[cat].unique())
    print()
print("All columns Unique values count")
for col in df:
    print(col, len(df[col].unique()), sep=': ')
plt.figure(figsize =(14,5))
plt.subplot(1,2,1)
sns.countplot(df['Attrition'] ,color ='b' ,hue =df['Gender'])
plt.title('Attrition by Gender')
plt.subplot(1,2,2)
plt.pie(df['Attrition'].value_counts() ,colors =['r' ,'c'] ,explode =[0,0.1] ,autopct =
'%.2f' ,labels =['No' ,'Yes'])
plt.title('Attrition')
#HANDLING CATEGORICAL OUTPUT VARIABLE
df['Attrition'].replace({'Yes':1 ,'No':0} ,inplace = True)
df['Attrition'].head()
df.drop(columns = no_use , axis = 1 , inplace = True)
df.columns
df['Gender'].replace({'Male':1 ,'Female':0} ,inplace = True)
df['OverTime'].replace({'Yes':1 ,'No':0} ,inplace = True)
(df.Attrition.value_counts()/1470)*100
smote = SMOTE(sampling_strategy='minority')
x ,y = smote.fit_resample(x ,y)
print(x.shape ,y.shape)
#now balanced
y.value_counts()
sns.countplot(y ,palette='viridis')
plt.title('Now Class is Balanced')
```

8.TESTING

8.1 Test Cases

8.2 User Acceptance Testing

1.Purpose of Document

The purpose of this document is to briefly explain the test coverage and open issue of corporateemployee attrition at the time of the release.

2.Defect Analysis

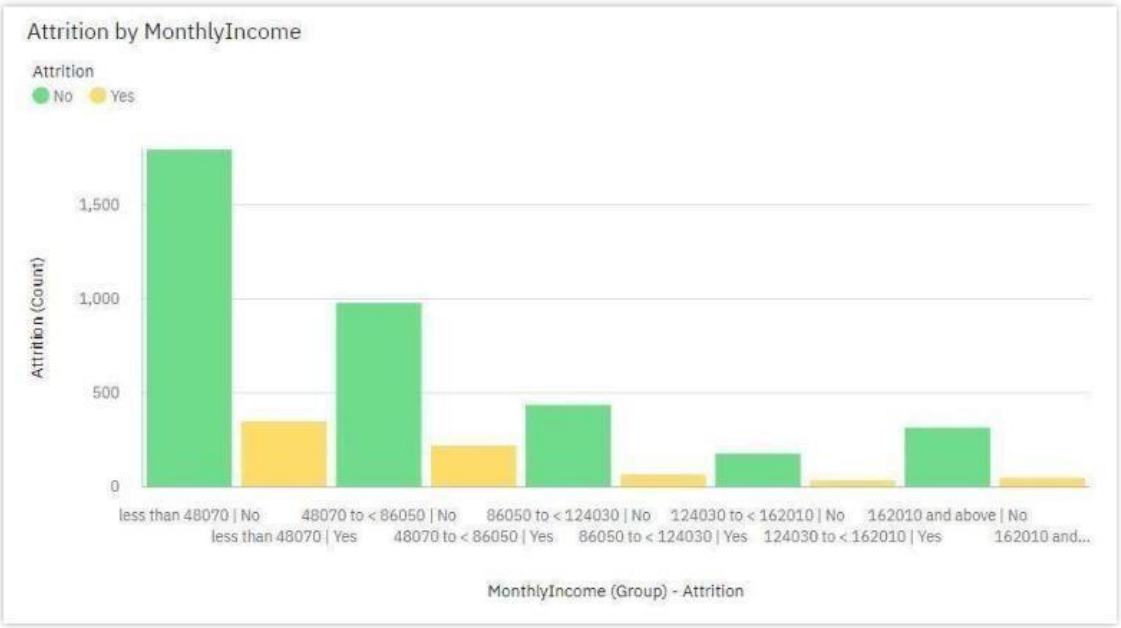
Resolution	Severity 1	Severity 2	Severity 3	Severity 4	Subtotal
By Design	3	2	0	0	5
Duplicate	4	0	2	0	6
External	3	2	0	0	5
Fixed	1	0	1	0	2
Not Reproduced	0	3	3	0	6
Skipped	0	0	3	2	5
Won't Fix	0	0	1	0	1
Totals	11	7	10	2	30

3.Test Case Analysis

Database	2	0	0	2
Dashboard	1	0	0	1
Visualize the data	8	0	0	8
Logistic Regression	4	0	0	4
Section	**Total Cases**	**Not Tested**	**Fail**	**Pass**
Login Page	1	0	0	1
Employee Attrition Details	1	0	0	1

9. RESULTS

9.1 Performance Metrics



10.ADVANTAGES & DISADVANTAGES

10.1 Advantages

Data Collection :

The study is conducted among working IT professionals of two different categories. This categorization mainly was focused on experience level and role in the organization. It was important to know the views of candidates who seek for the job for various reasons as well as the views of interviewers involved in the process of hiring the candidates. The research study involves reference of both primary and secondary data. Primary Data Primary data is collected through a field survey with the help of a structured self-administrated Questionnaire. The survey consisted of close ended questions by the means of convenience sampling. The scaling technique installed in the questionnaire is 5-point rating scale. Total 120 respondent were IT professionals belonging to the organizations from Nagpur, Pune and Mumbai cities in Maharashtra. Secondary Data Secondary data is collected by referring to the Journals, research papers and published data in the form of books and newspapers.

Type of Research :

The research paper adopted the descriptive research design methodology. Sample Design, Sample Size and Sampling Method The sample selected for the study is an Indian Information Technology Industry. The nature of the sample is restricted to working professionals in Information Technology sector and is collected through the convenience sampling technique. The sample size was 120 respondents.

11.CONCLUSION

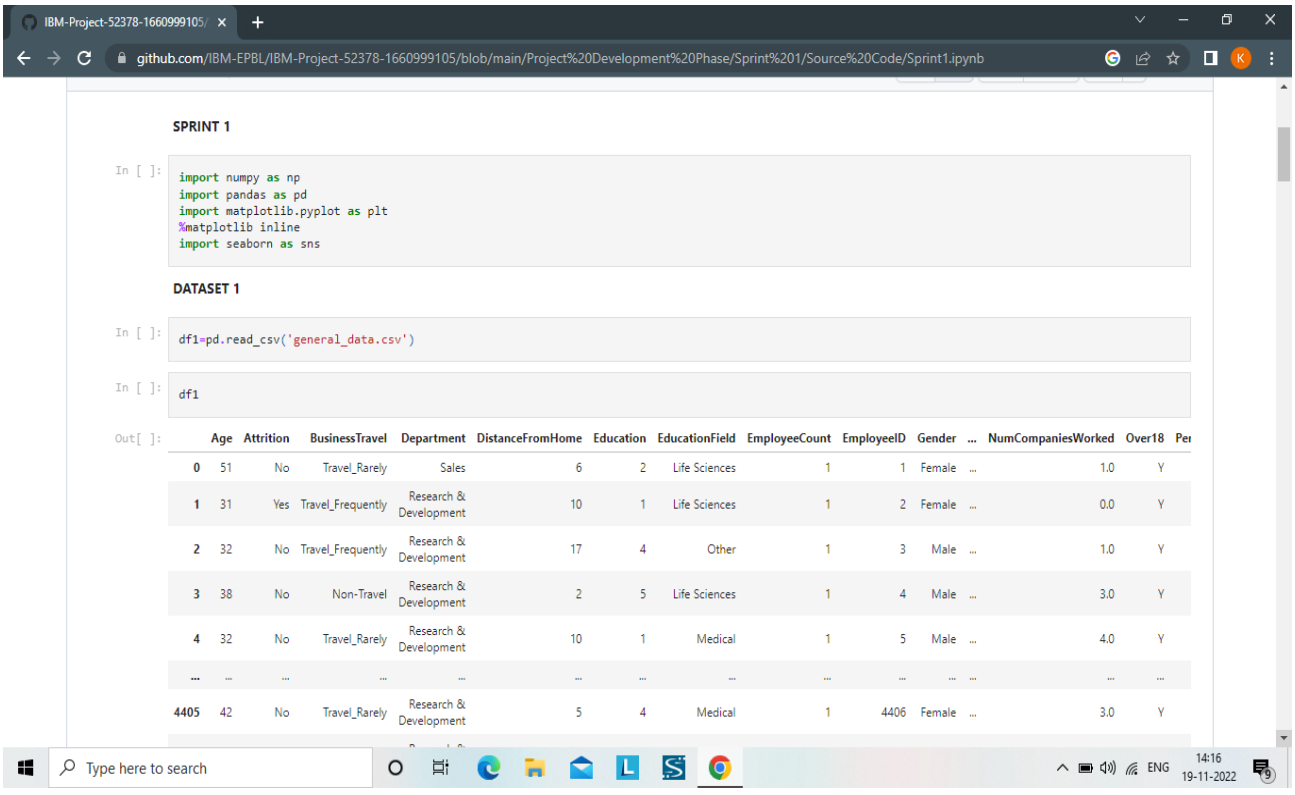
Employees as well as organizations must be clear with their expectations regarding the job profile. Any sort of mismatch leads to discrepancy and employees may fail to perform at their job. This eventually leads to attrition. Organizations should state the requirements and expectations unambiguously. This helps candidates decide upon to accept the job position or not. This eventually avoids further conflicts in the employment terms.

12.FUTURE SCOPE

Research findings suggest that attrition reasons in IT organizations primarily revolve around professional growth and challenges in the organization. Although economic factors happen to the most influential factor, professionals may settle for second best criteria of their preference that is career growth and supportive work policies in the organization. On the other hand, candidates who aspire to have a better job than the one in hand are more interested in securing the next job. Young talent wants to work on latest technology and functional domain. IT professionals who are young career makers are less influenced by Brand name or geographical area. Most of the IT professionals look for challenging role and position in the organization. Candidates as well as senior professionals believe that challenging work motivate them to maintain the interest in the work life. Employees as well as organizations must be clear with their expectations regarding the job profile. Any sort of mismatch leads to discrepancy and employees may fail to

perform at their job. This eventually leads to attrition. Organizations should state the requirements and expectations unambiguously. This helps candidates decide upon to accept the job position or not. This eventually avoids further conflicts in the employment terms. Further this research can make more detailed conclusions over "mapping of candidates' expectations with organizations' requirement" by collecting the data focusing on all the steps of recruitment and selection process.

13.APPENDIX



IBM-Project-52378-1660999105/

github.com/IBM-EPBL/IBM-Project-52378-1660999105/blob/main/Project%20Development%20Phase/Sprint%201/Source%20Code/Sprint1.ipynb

SPRINT 1

In [ ]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
```

DATASET 1

In [ ]:
```
df1=pd.read_csv('general_data.csv')
```

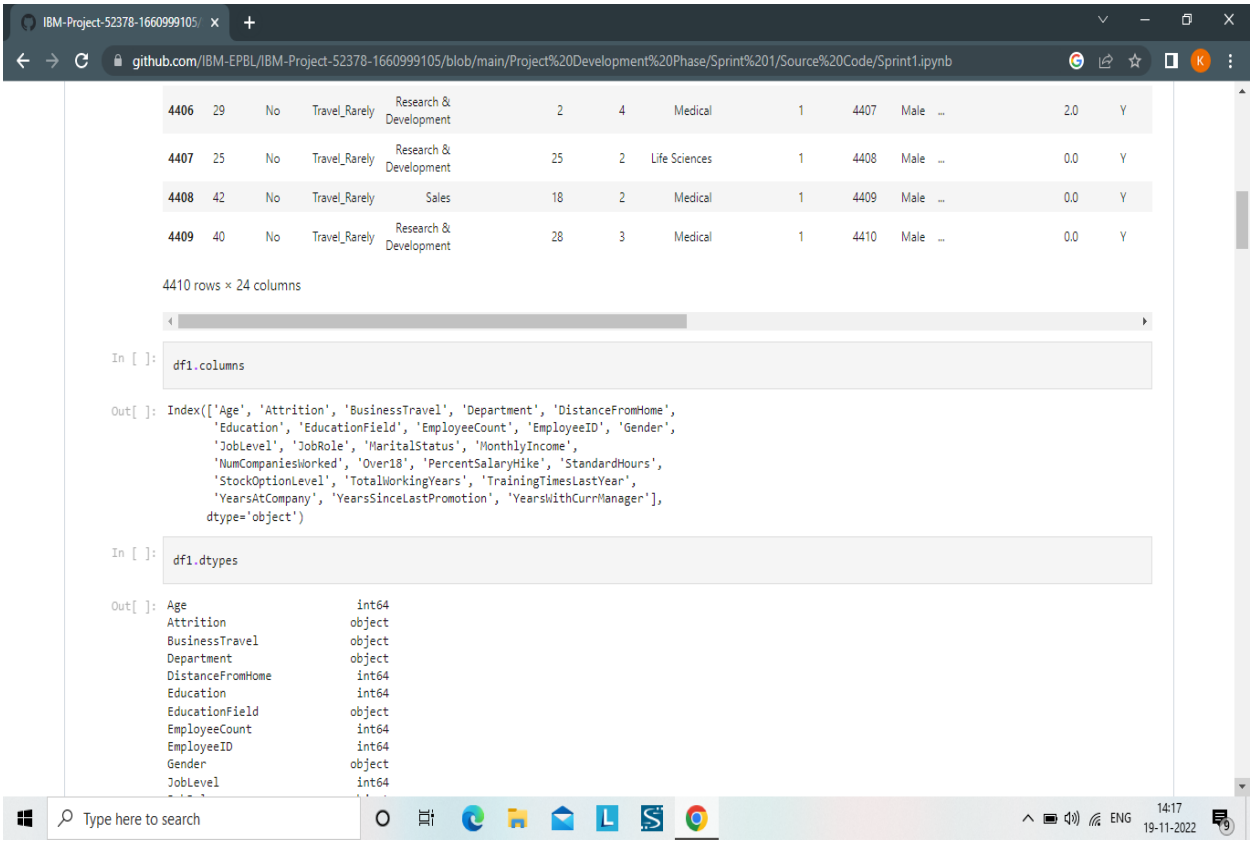
In [ ]:
```
df1
```

Out[ ]:

	Age	Attrition	BusinessTravel	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeID	Gender	...	NumCompaniesWorked	Over18	Per
0	51	No	Travel_Rarely	Sales	6	2	Life Sciences	1	1	Female	...	1.0	Y	
1	31	Yes	Travel_Frequently	Research & Development	10	1	Life Sciences	1	2	Female	...	0.0	Y	
2	32	No	Travel_Frequently	Research & Development	17	4	Other	1	3	Male	...	1.0	Y	
3	38	No	Non-Travel	Research & Development	2	5	Life Sciences	1	4	Male	...	3.0	Y	
4	32	No	Travel_Rarely	Research & Development	10	1	Medical	1	5	Male	...	4.0	Y	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
4405	42	No	Travel_Rarely	Research & Development	5	4	Medical	1	4406	Female	...	3.0	Y	

Type here to search

ENG 14:16 19-11-2022



4406	29	No	Travel_Rarely	Research & Development	2	4	Medical	1	4407	Male	...	2.0	Y
4407	25	No	Travel_Rarely	Research & Development	25	2	Life Sciences	1	4408	Male	...	0.0	Y
4408	42	No	Travel_Rarely	Sales	18	2	Medical	1	4409	Male	...	0.0	Y
4409	40	No	Travel_Rarely	Research & Development	28	3	Medical	1	4410	Male	...	0.0	Y

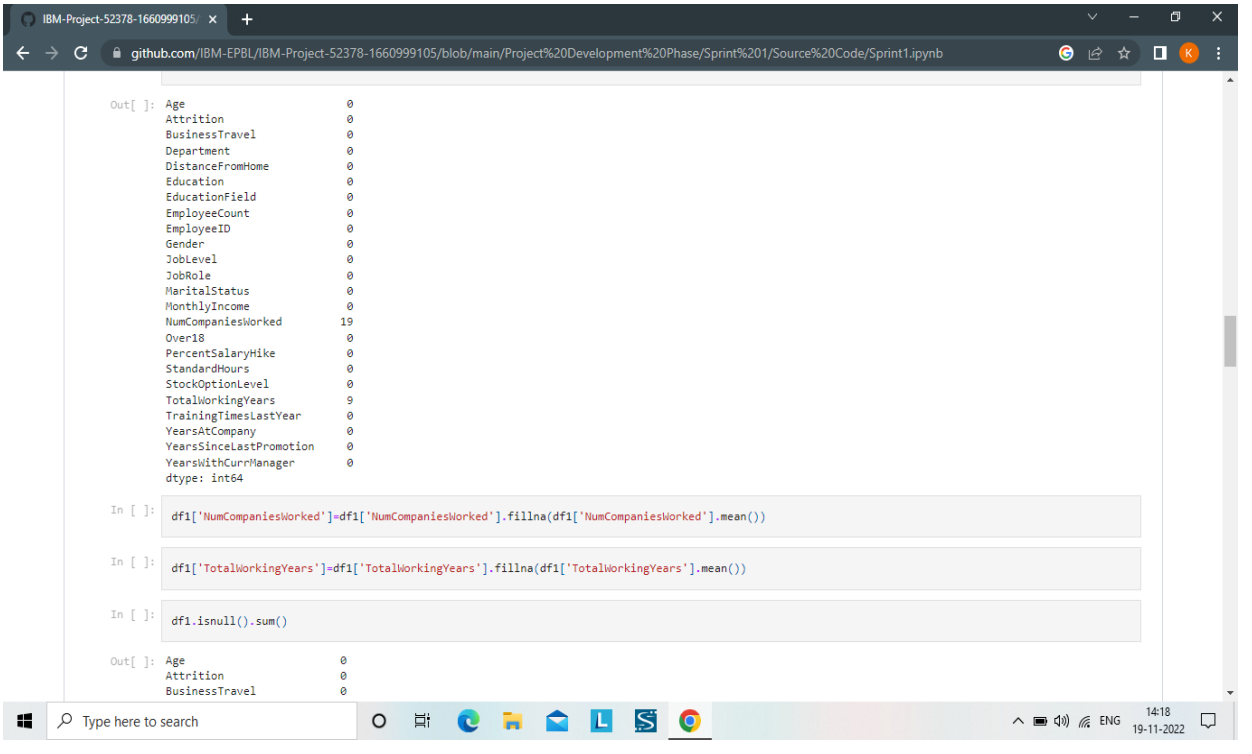
4410 rows × 24 columns

In [ ]:
```
df1.columns
```

Out[ ]:
```
Index(['Age', 'Attrition', 'BusinessTravel', 'Department', 'DistanceFromHome',
       'Education', 'EducationField', 'EmployeeCount', 'EmployeeID', 'Gender',
       'JobLevel', 'JobRole', 'MaritalStatus', 'MonthlyIncome',
       'NumCompaniesWorked', 'Over18', 'PercentSalaryHike', 'StandardHours',
       'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear',
       'YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager'],
      dtype='object')
```

In [ ]:
```
df1.dtypes
```

Out[ ]:
```
Age                        int64
Attrition                 object
BusinessTravel            object
Department                object
DistanceFromHome           int64
Education                  int64
EducationField            object
EmployeeCount              int64
EmployeeID                 int64
Gender                    object
JobLevel                   int64
```



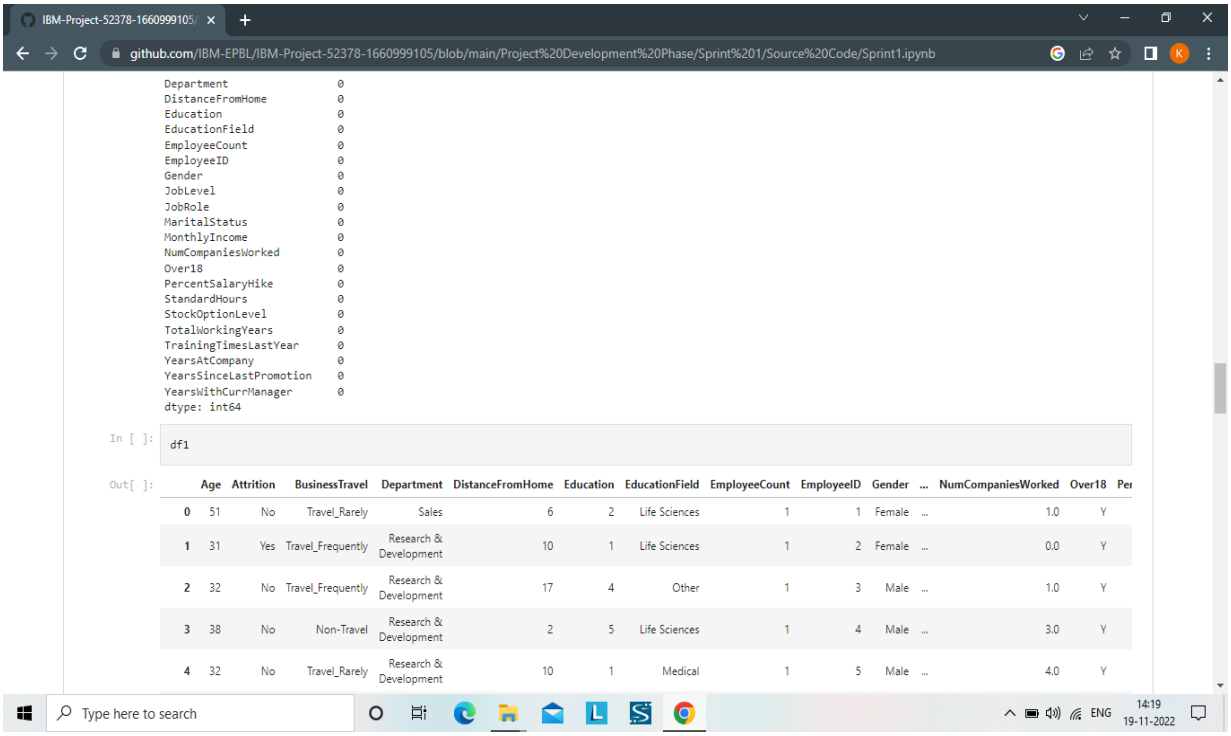
Out[ ]:
```
Age                       0
Attrition                 0
BusinessTravel            0
Department                0
DistanceFromHome          0
Education                 0
EducationField            0
EmployeeCount             0
EmployeeID                0
Gender                    0
JobLevel                  0
JobRole                   0
MaritalStatus             0
MonthlyIncome             0
NumCompaniesWorked       19
Over18                    0
PercentSalaryHike         0
StandardHours             0
StockOptionLevel          0
TotalWorkingYears         9
TrainingTimesLastYear     0
YearsAtCompany            0
YearsSinceLastPromotion   0
YearsWithCurrManager      0
dtype: int64
```

In [ ]:
```
df1['NumCompaniesWorked']=df1['NumCompaniesWorked'].fillna(df1['NumCompaniesWorked'].mean())
```

In [ ]:
```
df1['TotalWorkingYears']=df1['TotalWorkingYears'].fillna(df1['TotalWorkingYears'].mean())
```

In [ ]:
```
df1.isnull().sum()
```

Out[ ]:
```
Age                  0
Attrition            0
BusinessTravel       0
```

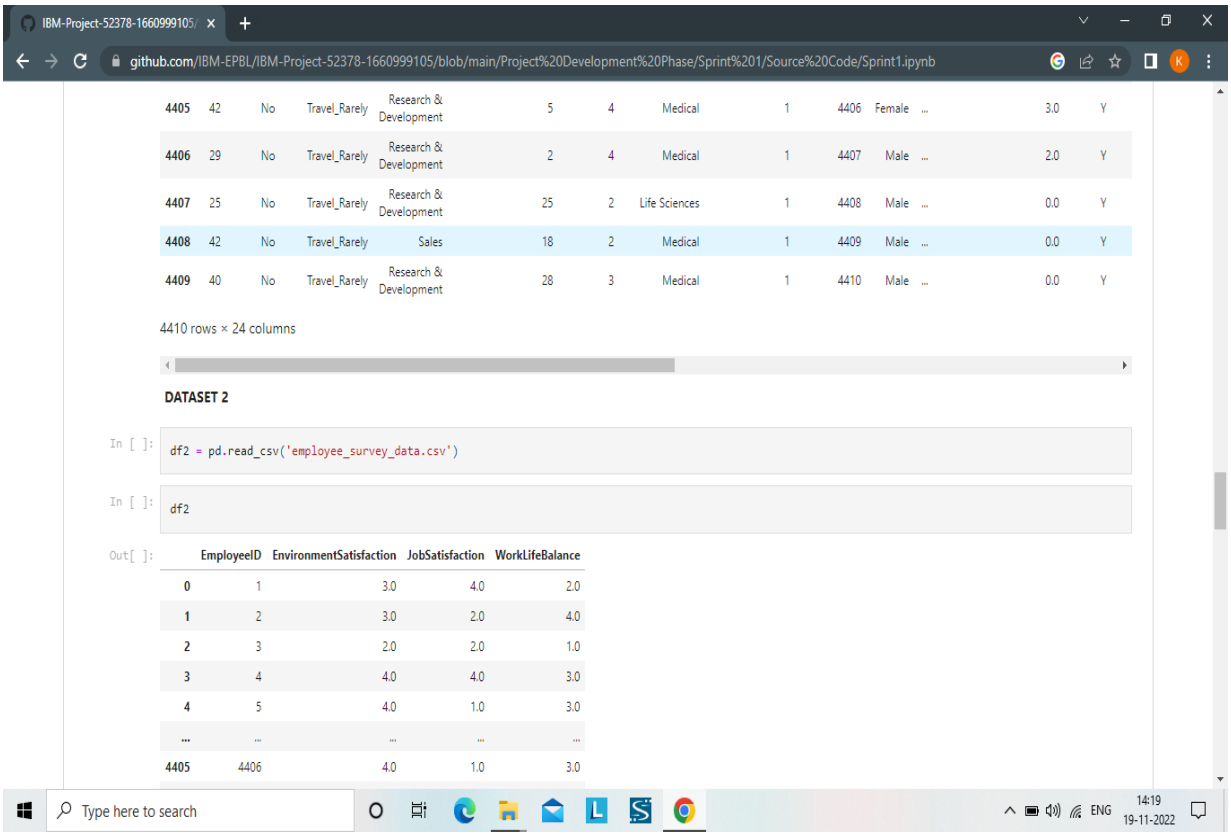


```
Department                0
DistanceFromHome          0
Education                 0
EducationField            0
EmployeeCount             0
EmployeeID                0
Gender                    0
JobLevel                  0
JobRole                   0
MaritalStatus             0
MonthlyIncome             0
NumCompaniesWorked        0
Over18                    0
PercentSalaryHike         0
StandardHours             0
StockOptionLevel          0
TotalWorkingYears         0
TrainingTimesLastYear     0
YearsAtCompany            0
YearsSinceLastPromotion   0
YearsWithCurrManager      0
dtype: int64
```

In [ ]:
```
df1
```

Out[ ]:

	Age	Attrition	BusinessTravel	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeID	Gender	...	NumCompaniesWorked	Over18	Per
0	51	No	Travel_Rarely	Sales	6	2	Life Sciences	1	1	Female	...	1.0	Y	
1	31	Yes	Travel_Frequently	Research & Development	10	1	Life Sciences	1	2	Female	...	0.0	Y	
2	32	No	Travel_Frequently	Research & Development	17	4	Other	1	3	Male	...	1.0	Y	
3	38	No	Non-Travel	Research & Development	2	5	Life Sciences	1	4	Male	...	3.0	Y	
4	32	No	Travel_Rarely	Research & Development	10	1	Medical	1	5	Male	...	4.0	Y	



4405	42	No	Travel_Rarely	Research & Development	5	4	Medical	1	4406	Female	...	3.0	Y
4406	29	No	Travel_Rarely	Research & Development	2	4	Medical	1	4407	Male	...	2.0	Y
4407	25	No	Travel_Rarely	Research & Development	25	2	Life Sciences	1	4408	Male	...	0.0	Y
4408	42	No	Travel_Rarely	Sales	18	2	Medical	1	4409	Male	...	0.0	Y
4409	40	No	Travel_Rarely	Research & Development	28	3	Medical	1	4410	Male	...	0.0	Y

4410 rows × 24 columns

DATASET 2

In [ ]:
```
df2 = pd.read_csv('employee_survey_data.csv')
```

In [ ]:
```
df2
```

Out[ ]:

	EmployeeID	EnvironmentSatisfaction	JobSatisfaction	WorkLifeBalance
0	1	3.0	4.0	2.0
1	2	3.0	2.0	4.0
2	3	2.0	2.0	1.0
3	4	4.0	4.0	3.0
4	5	4.0	1.0	3.0
...	...	...	...	...
4405	4406	4.0	1.0	3.0