Name: Bylapudi Lahari

Registration Number: 21BCE9969

Email: lahari.21bce9969@vitapstudent.ac.in

In [42]:
```
# importing necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

1.Download the Employee Attrition Dataset

https://www.kaggle.com/datasets/patelprashant/employee-attrition

In [2]:
```
# importing the dataset
dataset = pd.read_csv("Employee_Attrition.csv")
```

In [3]:
```
dataset.head()
```

Out[3]:

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	Education
0	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sci
1	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sci
2	37	Yes	Travel_Rarely	1373	Research & Development	2	2	
3	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sci
4	27	No	Travel_Rarely	591	Research & Development	2	1	Me

5 rows × 35 columns

In [4]:
```
dataset.shape
```

Out[4]:
```
(1470, 35)
```

In [5]:
```
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1470 entries, 0 to 1469
Data columns (total 35 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Age                       1470 non-null   int64
 1   Attrition                 1470 non-null   object
 2   BusinessTravel            1470 non-null   object
 3   DailyRate                 1470 non-null   int64
 4   Department                1470 non-null   object
 5   DistanceFromHome          1470 non-null   int64
 6   Education                 1470 non-null   int64
 7   EducationField            1470 non-null   object
 8   EmployeeCount             1470 non-null   int64
 9   EmployeeNumber            1470 non-null   int64
 10  EnvironmentSatisfaction   1470 non-null   int64
 11  Gender                    1470 non-null   object
 12  HourlyRate                1470 non-null   int64
 13  JobInvolvement            1470 non-null   int64
 14  JobLevel                  1470 non-null   int64
 15  JobRole                   1470 non-null   object
 16  JobSatisfaction           1470 non-null   int64
 17  MaritalStatus             1470 non-null   object
 18  MonthlyIncome             1470 non-null   int64
 19  MonthlyRate               1470 non-null   int64
 20  NumCompaniesWorked        1470 non-null   int64
 21  Over18                    1470 non-null   object
 22  OverTime                  1470 non-null   object
 23  PercentSalaryHike         1470 non-null   int64
 24  PerformanceRating         1470 non-null   int64
 25  RelationshipSatisfaction  1470 non-null   int64
 26  StandardHours             1470 non-null   int64
 27  StockOptionLevel          1470 non-null   int64
 28  TotalWorkingYears         1470 non-null   int64
 29  TrainingTimesLastYear     1470 non-null   int64
 30  WorkLifeBalance           1470 non-null   int64
 31  YearsAtCompany            1470 non-null   int64
 32  YearsInCurrentRole        1470 non-null   int64
 33  YearsSinceLastPromotion   1470 non-null   int64
 34  YearsWithCurrManager      1470 non-null   int64
dtypes: int64(26), object(9)
memory usage: 402.1+ KB
```

In [6]:
```
dataset.describe()
```

Out[6]:

	Age	DailyRate	DistanceFromHome	Education	EmployeeCount	EmployeeNumber
count	1470.000000	1470.000000	1470.000000	1470.000000	1470.0	1470.000000
mean	36.923810	802.485714	9.192517	2.912925	1.0	1024.865306
std	9.135373	403.509100	8.106864	1.024165	0.0	602.024335
min	18.000000	102.000000	1.000000	1.000000	1.0	1.000000
25%	30.000000	465.000000	2.000000	2.000000	1.0	491.250000
50%	36.000000	802.000000	7.000000	3.000000	1.0	1020.500000
75%	43.000000	1157.000000	14.000000	4.000000	1.0	1555.750000
max	60.000000	1499.000000	29.000000	5.000000	1.0	2068.000000

8 rows × 26 columns

2.Perfrom Data Preprocessing

In [7]:
```
dataset.isnull().any()
```

Out[7]:
```
Age                         False
Attrition                   False
BusinessTravel              False
DailyRate                   False
Department                  False
DistanceFromHome            False
Education                   False
EducationField              False
EmployeeCount               False
EmployeeNumber              False
EnvironmentSatisfaction     False
Gender                      False
HourlyRate                  False
JobInvolvement              False
JobLevel                    False
JobRole                     False
JobSatisfaction             False
MaritalStatus               False
MonthlyIncome               False
MonthlyRate                 False
NumCompaniesWorked          False
Over18                      False
OverTime                    False
PercentSalaryHike           False
PerformanceRating           False
RelationshipSatisfaction    False
StandardHours               False
StockOptionLevel            False
TotalWorkingYears           False
TrainingTimesLastYear       False
WorkLifeBalance             False
YearsAtCompany              False
YearsInCurrentRole          False
YearsSinceLastPromotion     False
YearsWithCurrManager        False
dtype: bool
```

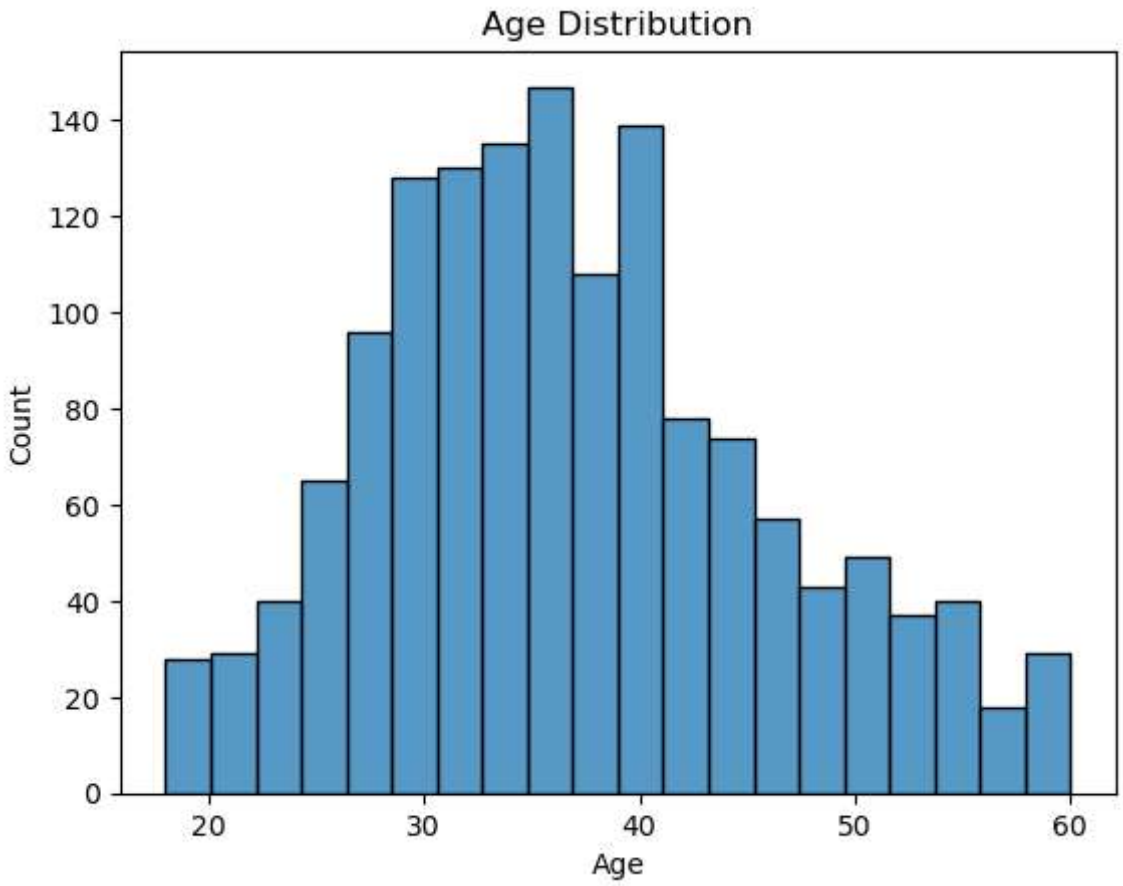
In [8]:
```
dataset.isnull().sum()
```

Out[8]:

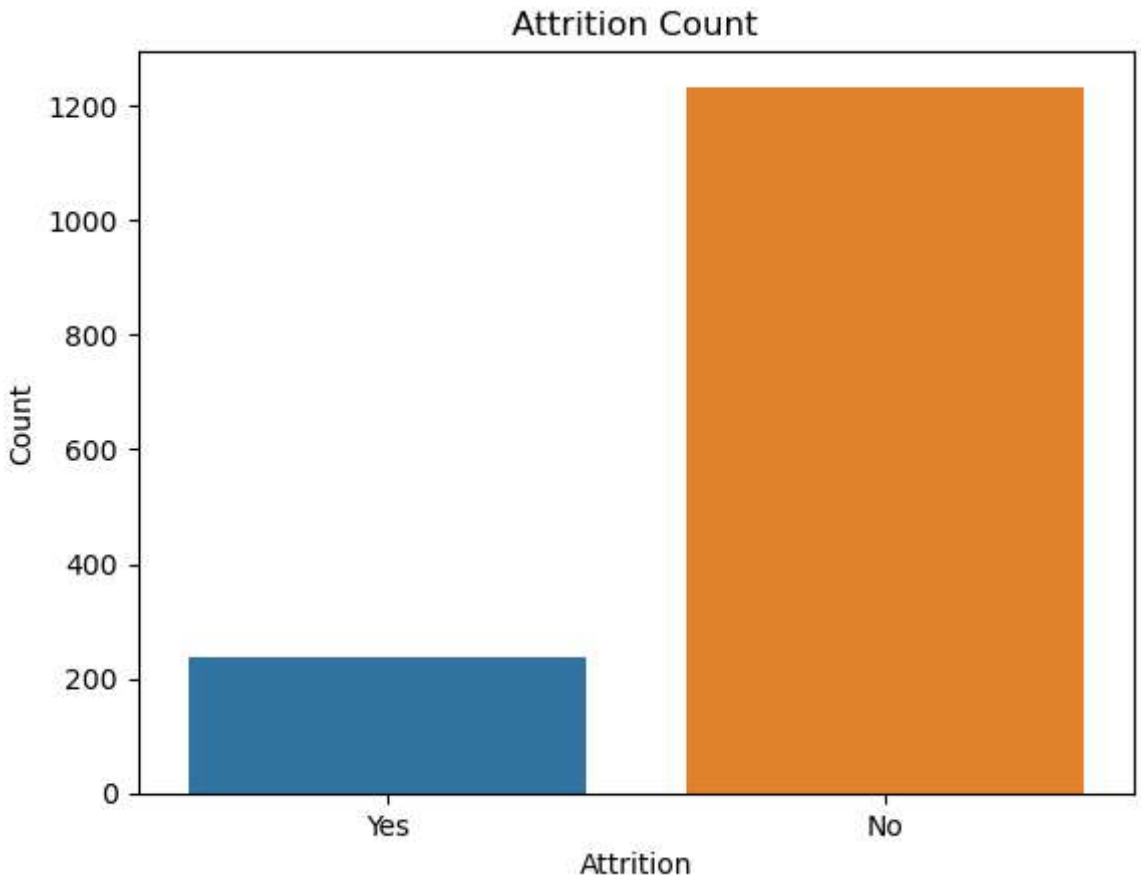
```
Age                         0
Attrition                   0
BusinessTravel              0
DailyRate                   0
Department                  0
DistanceFromHome            0
Education                   0
EducationField              0
EmployeeCount               0
EmployeeNumber              0
EnvironmentSatisfaction     0
Gender                      0
HourlyRate                  0
JobInvolvement              0
JobLevel                    0
JobRole                     0
JobSatisfaction             0
MaritalStatus               0
MonthlyIncome               0
MonthlyRate                 0
NumCompaniesWorked          0
Over18                      0
OverTime                    0
PercentSalaryHike           0
PerformanceRating           0
RelationshipSatisfaction    0
StandardHours               0
StockOptionLevel            0
TotalWorkingYears           0
TrainingTimesLastYear       0
WorkLifeBalance             0
YearsAtCompany              0
YearsInCurrentRole          0
YearsSinceLastPromotion     0
YearsWithCurrManager        0
dtype: int64
```

In [9]:

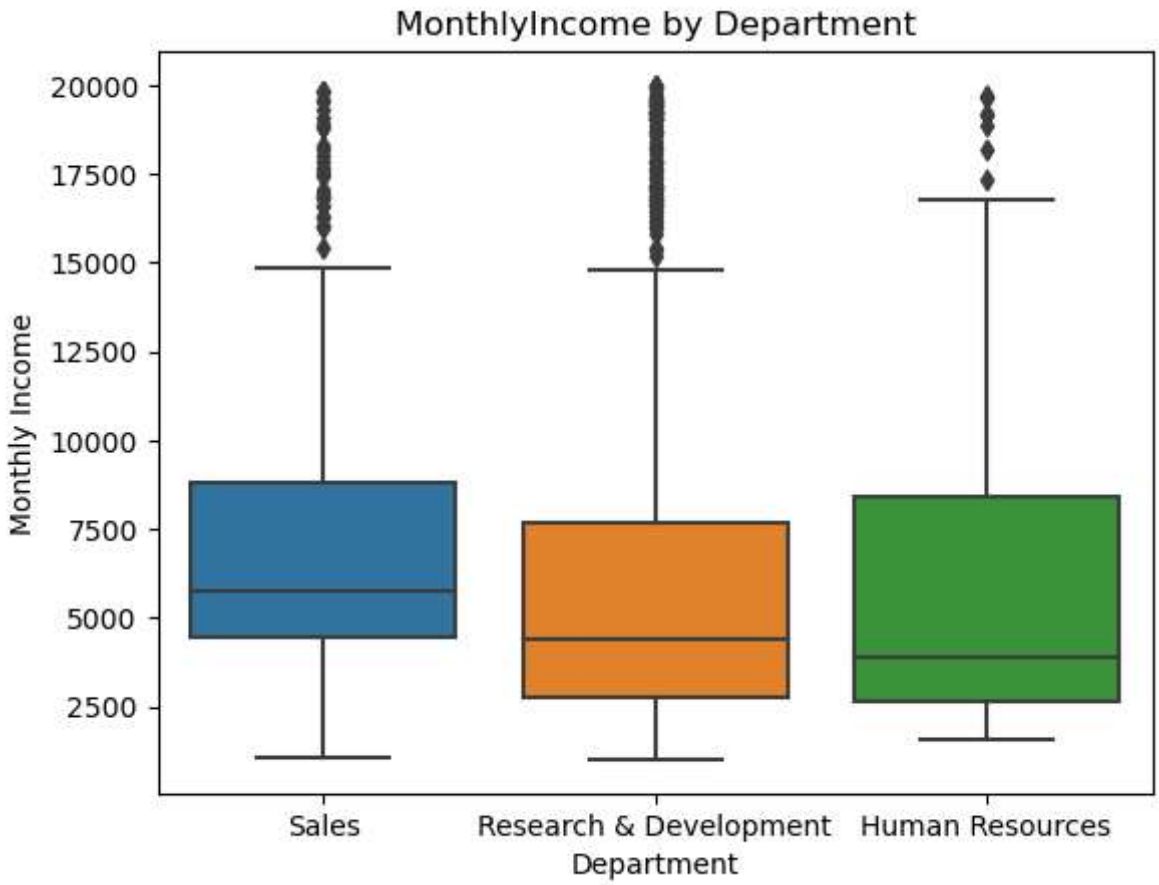
```
sns.histplot(dataset['Age'], bins=20)
plt.title('Age Distribution')
plt.xlabel('Age')
plt.ylabel('Count')
plt.show()
```



In [10]:
```
sns.countplot(x='Attrition', data=dataset)
plt.title('Attrition Count')
plt.xlabel('Attrition')
plt.ylabel('Count')
plt.show()
```



In [12]:
```
sns.boxplot(x='Department', y='MonthlyIncome', data=dataset)
plt.title('MonthlyIncome by Department')
plt.xlabel('Department')
plt.ylabel('Monthly Income')
plt.show()
```



In [17]:
```
numerical_features = dataset.select_dtypes(include=['int64', 'float64'])
correlation_matrix = numerical_features.corr()
print(correlation_matrix)
```

	Age	DailyRate	DistanceFromHome	Education
Age	1.000000	0.010661	-0.001686	0.208034
DailyRate	0.010661	1.000000	-0.004985	-0.016806
DistanceFromHome	-0.001686	-0.004985	1.000000	0.021042
Education	0.208034	-0.016806	0.021042	1.000000
EmployeeCount	NaN	NaN	NaN	NaN
EmployeeNumber	-0.010145	-0.050990	0.032916	0.042070
EnvironmentSatisfaction	0.010146	0.018355	-0.016075	-0.027128
HourlyRate	0.024287	0.023381	0.031131	0.016775
JobInvolvement	0.029820	0.046135	0.008783	0.042438
JobLevel	0.509604	0.002966	0.005303	0.101589
JobSatisfaction	-0.004892	0.030571	-0.003669	-0.011296
MonthlyIncome	0.497855	0.007707	-0.017014	0.094961
MonthlyRate	0.028051	-0.032182	0.027473	-0.026084
NumCompaniesWorked	0.299635	0.038153	-0.029251	0.126317
PercentSalaryHike	0.003634	0.022704	0.040235	-0.011111
PerformanceRating	0.001904	0.000473	0.027110	-0.024539
RelationshipSatisfaction	0.053535	0.007846	0.006557	-0.009118
StandardHours	NaN	NaN	NaN	NaN
StockOptionLevel	0.037510	0.042143	0.044872	0.018422
TotalWorkingYears	0.680381	0.014515	0.004628	0.148280
TrainingTimesLastYear	-0.019621	0.002453	-0.036942	-0.025100
WorkLifeBalance	-0.021490	-0.037848	-0.026556	0.009819
YearsAtCompany	0.311309	-0.034055	0.009508	0.069114
YearsInCurrentRole	0.212901	0.009932	0.018845	0.060236
YearsSinceLastPromotion	0.216513	-0.033229	0.010029	0.054254
YearsWithCurrManager	0.202089	-0.026363	0.014406	0.069065

	EmployeeCount	EmployeeNumber
Age	NaN	-0.010145
DailyRate	NaN	-0.050990
DistanceFromHome	NaN	0.032916
Education	NaN	0.042070
EmployeeCount	NaN	NaN
EmployeeNumber	NaN	1.000000
EnvironmentSatisfaction	NaN	0.017621
HourlyRate	NaN	0.035179
JobInvolvement	NaN	-0.006888
JobLevel	NaN	-0.018519
JobSatisfaction	NaN	-0.046247
MonthlyIncome	NaN	-0.014829
MonthlyRate	NaN	0.012648
NumCompaniesWorked	NaN	-0.001251
PercentSalaryHike	NaN	-0.012944
PerformanceRating	NaN	-0.020359
RelationshipSatisfaction	NaN	-0.069861
StandardHours	NaN	NaN
StockOptionLevel	NaN	0.062227
TotalWorkingYears	NaN	-0.014365
TrainingTimesLastYear	NaN	0.023603
WorkLifeBalance	NaN	0.010309
YearsAtCompany	NaN	-0.011240
YearsInCurrentRole	NaN	-0.008416
YearsSinceLastPromotion	NaN	-0.009019
YearsWithCurrManager	NaN	-0.009197

	EnvironmentSatisfaction	HourlyRate	JobInvolvement
Age	0.010146	0.024287	0.029820
DailyRate	0.018355	0.023381	0.046135
DistanceFromHome	-0.016075	0.031131	0.008783

Education	-0.027128	0.016775	0.042438
EmployeeCount	NaN	NaN	NaN
EmployeeNumber	0.017621	0.035179	-0.006888
EnvironmentSatisfaction	1.000000	-0.049857	-0.008278
HourlyRate	-0.049857	1.000000	0.042861
JobInvolvement	-0.008278	0.042861	1.000000
JobLevel	0.001212	-0.027853	-0.012630
JobSatisfaction	-0.006784	-0.071335	-0.021476
MonthlyIncome	-0.006259	-0.015794	-0.015271
MonthlyRate	0.037600	-0.015297	-0.016322
NumCompaniesWorked	0.012594	0.022157	0.015012
PercentSalaryHike	-0.031701	-0.009062	-0.017205
PerformanceRating	-0.029548	-0.002172	-0.029071
RelationshipSatisfaction	0.007665	0.001330	0.034297
StandardHours	NaN	NaN	NaN
StockOptionLevel	0.003432	0.050263	0.021523
TotalWorkingYears	-0.002693	-0.002334	-0.005533
TrainingTimesLastYear	-0.019359	-0.008548	-0.015338
WorkLifeBalance	0.027627	-0.004607	-0.014617
YearsAtCompany	0.001458	-0.019582	-0.021355
YearsInCurrentRole	0.018007	-0.024106	0.008717
YearsSinceLastPromotion	0.016194	-0.026716	-0.024184
YearsWithCurrManager	-0.004999	-0.020123	0.025976

	JobLevel	...	RelationshipSatisfaction	\
Age	0.509604	...	0.053535	
DailyRate	0.002966	...	0.007846	
DistanceFromHome	0.005303	...	0.006557	
Education	0.101589	...	-0.009118	
EmployeeCount	NaN	...	NaN	
EmployeeNumber	-0.018519	...	-0.069861	
EnvironmentSatisfaction	0.001212	...	0.007665	
HourlyRate	-0.027853	...	0.001330	
JobInvolvement	-0.012630	...	0.034297	
JobLevel	1.000000	...	0.021642	
JobSatisfaction	-0.001944	...	-0.012454	
MonthlyIncome	0.950300	...	0.025873	
MonthlyRate	0.039563	...	-0.004085	
NumCompaniesWorked	0.142501	...	0.052733	
PercentSalaryHike	-0.034730	...	-0.040490	
PerformanceRating	-0.021222	...	-0.031351	
RelationshipSatisfaction	0.021642	...	1.000000	
StandardHours	NaN	...	NaN	
StockOptionLevel	0.013984	...	-0.045952	
TotalWorkingYears	0.782208	...	0.024054	
TrainingTimesLastYear	-0.018191	...	0.002497	
WorkLifeBalance	0.037818	...	0.019604	
YearsAtCompany	0.534739	...	0.019367	
YearsInCurrentRole	0.389447	...	-0.015123	
YearsSinceLastPromotion	0.353885	...	0.033493	
YearsWithCurrManager	0.375281	...	-0.000867	

	StandardHours	StockOptionLevel	TotalWorkingYears	\
Age	NaN	0.037510	0.680381	
DailyRate	NaN	0.042143	0.014515	
DistanceFromHome	NaN	0.044872	0.004628	
Education	NaN	0.018422	0.148280	
EmployeeCount	NaN	NaN	NaN	
EmployeeNumber	NaN	0.062227	-0.014365	
EnvironmentSatisfaction	NaN	0.003432	-0.002693	

HourlyRate	NaN	0.050263	-0.002334
JobInvolvement	NaN	0.021523	-0.005533
JobLevel	NaN	0.013984	0.782208
JobSatisfaction	NaN	0.010690	-0.020185
MonthlyIncome	NaN	0.005408	0.772893
MonthlyRate	NaN	-0.034323	0.026442
NumCompaniesWorked	NaN	0.030075	0.237639
PercentSalaryHike	NaN	0.007528	-0.020608
PerformanceRating	NaN	0.003506	0.006744
RelationshipSatisfaction	NaN	-0.045952	0.024054
StandardHours	NaN	NaN	NaN
StockOptionLevel	NaN	1.000000	0.010136
TotalWorkingYears	NaN	0.010136	1.000000
TrainingTimesLastYear	NaN	0.011274	-0.035662
WorkLifeBalance	NaN	0.004129	0.001008
YearsAtCompany	NaN	0.015058	0.628133
YearsInCurrentRole	NaN	0.050818	0.460365
YearsSinceLastPromotion	NaN	0.014352	0.404858
YearsWithCurrManager	NaN	0.024698	0.459188

	TrainingTimesLastYear	WorkLifeBalance	\
Age	-0.019621	-0.021490	
DailyRate	0.002453	-0.037848	
DistanceFromHome	-0.036942	-0.026556	
Education	-0.025100	0.009819	
EmployeeCount	NaN	NaN	
EmployeeNumber	0.023603	0.010309	
EnvironmentSatisfaction	-0.019359	0.027627	
HourlyRate	-0.008548	-0.004607	
JobInvolvement	-0.015338	-0.014617	
JobLevel	-0.018191	0.037818	
JobSatisfaction	-0.005779	-0.019459	
MonthlyIncome	-0.021736	0.030683	
MonthlyRate	0.001467	0.007963	
NumCompaniesWorked	-0.066054	-0.008366	
PercentSalaryHike	-0.005221	-0.003280	
PerformanceRating	-0.015579	0.002572	
RelationshipSatisfaction	0.002497	0.019604	
StandardHours	NaN	NaN	
StockOptionLevel	0.011274	0.004129	
TotalWorkingYears	-0.035662	0.001008	
TrainingTimesLastYear	1.000000	0.028072	
WorkLifeBalance	0.028072	1.000000	
YearsAtCompany	0.003569	0.012089	
YearsInCurrentRole	-0.005738	0.049856	
YearsSinceLastPromotion	-0.002067	0.008941	
YearsWithCurrManager	-0.004096	0.002759	

	YearsAtCompany	YearsInCurrentRole	\
Age	0.311309	0.212901	
DailyRate	-0.034055	0.009932	
DistanceFromHome	0.009508	0.018845	
Education	0.069114	0.060236	
EmployeeCount	NaN	NaN	
EmployeeNumber	-0.011240	-0.008416	
EnvironmentSatisfaction	0.001458	0.018007	
HourlyRate	-0.019582	-0.024106	
JobInvolvement	-0.021355	0.008717	
JobLevel	0.534739	0.389447	
JobSatisfaction	-0.003803	-0.002305	

```
MonthlyIncome                0.514285           0.363818
MonthlyRate                 -0.023655          -0.012815
NumCompaniesWorked          -0.118421          -0.090754
PercentSalaryHike           -0.035991          -0.001520
PerformanceRating            0.003435           0.034986
RelationshipSatisfaction     0.019367          -0.015123
StandardHours                     NaN                NaN
StockOptionLevel             0.015058           0.050818
TotalWorkingYears            0.628133           0.460365
TrainingTimesLastYear        0.003569          -0.005738
WorkLifeBalance              0.012089           0.049856
YearsAtCompany               1.000000           0.758754
YearsInCurrentRole           0.758754           1.000000
YearsSinceLastPromotion      0.618409           0.548056
YearsWithCurrManager         0.769212           0.714365

                         YearsSinceLastPromotion  YearsWithCurrManager
Age                                     0.216513              0.202089
DailyRate                              -0.033229             -0.026363
DistanceFromHome                        0.010029              0.014406
Education                               0.054254              0.069065
EmployeeCount                                NaN                   NaN
EmployeeNumber                         -0.009019             -0.009197
EnvironmentSatisfaction                 0.016194             -0.004999
HourlyRate                             -0.026716             -0.020123
JobInvolvement                         -0.024184              0.025976
JobLevel                                0.353885              0.375281
JobSatisfaction                        -0.018214             -0.027656
MonthlyIncome                           0.344978              0.344079
MonthlyRate                             0.001567             -0.036746
NumCompaniesWorked                     -0.036814             -0.110319
PercentSalaryHike                      -0.022154             -0.011985
PerformanceRating                       0.017896              0.022827
RelationshipSatisfaction                0.033493             -0.000867
StandardHours                                NaN                   NaN
StockOptionLevel                        0.014352              0.024698
TotalWorkingYears                       0.404858              0.459188
TrainingTimesLastYear                  -0.002067             -0.004096
WorkLifeBalance                         0.008941              0.002759
YearsAtCompany                          0.618409              0.769212
YearsInCurrentRole                      0.548056              0.714365
YearsSinceLastPromotion                 1.000000              0.510224
YearsWithCurrManager                    0.510224              1.000000

[26 rows x 26 columns]
```

In [ ]:

In [18]:
```
x = dataset.drop(columns = ['Attrition'])
y = dataset['Attrition']
```

In [19]:
```
x.head()
```

Out[19]:

	Age	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	Emp
0	41	Travel_Rarely	1102	Sales	1	2	Life Sciences	
1	49	Travel_Frequently	279	Research & Development	8	1	Life Sciences	
2	37	Travel_Rarely	1373	Research & Development	2	2	Other	
3	33	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	
4	27	Travel_Rarely	591	Research & Development	2	1	Medical	

5 rows × 34 columns

In [20]:
```
y.head()
```

Out[20]:
```
0    Yes
1     No
2    Yes
3     No
4     No
Name: Attrition, dtype: object
```

In [21]:
```
categorical_columns = ['BusinessTravel', 'Department', 'EducationField', 'Gender', 'Jo
x_encoded = pd.get_dummies(x, columns=categorical_columns, drop_first=True)
```

In [22]:
```
from sklearn.model_selection import train_test_split
```

In [23]:
```
x_train, x_test, y_train, y_test = train_test_split(x_encoded, y, test_size=0.3, rando
```

In [24]:
```
print(x_train.shape)
```

```
(1029, 47)
```

In [25]:
```
print(x_test.shape)
```

```
(441, 47)
```

In [26]:
```
print(y_train.shape)
```

```
(1029,)
```

In [27]:
```
print(y_test.shape)
```

```
(441,)
```

3.Model Building using Logistic Regression and Decision Tree

In [29]:
```
#Using Logistic regression
from sklearn.linear_model import LogisticRegression
```

In [30]:
```
logistic_model = LogisticRegression(random_state=42)
logistic_model.fit(x_train, y_train)
```

Out[30]:
```
LogisticRegression
LogisticRegression(random_state=42)
```

In [31]:
```
y_pred_logistic = logistic_model.predict(x_test)
```

In [32]:
```
#Using Decision tree
from sklearn.tree import DecisionTreeClassifier
```

In [33]:
```
decision_tree_model = DecisionTreeClassifier(random_state = 42)
decision_tree_model.fit(x_train, y_train)
```

Out[33]:
```
DecisionTreeClassifier
DecisionTreeClassifier(random_state=42)
```

In [34]:
```
y_pred_tree = decision_tree_model.predict(x_test)
```

4.Calculate Performance metrics

In [35]:
```
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
```

In [36]:
```
#Logistic regression
accuracy = accuracy_score(y_test, y_pred_logistic)
print(accuracy)
```

```
0.8639455782312925
```

In [37]:
```
confusion = confusion_matrix(y_test, y_pred_logistic)
print(confusion)
```

```
[[380   0]
 [ 60   1]]
```

In [38]:
```
classification = classification_report(y_test, y_pred_logistic)
print(classification)
```

```
              precision    recall  f1-score   support

          No       0.86      1.00      0.93       380
         Yes       1.00      0.02      0.03        61

    accuracy                           0.86       441
   macro avg       0.93      0.51      0.48       441
weighted avg       0.88      0.86      0.80       441
```

In [ ]:

In [ ]:

In [39]:
```
#Decision tree
accuracy = accuracy_score(y_test, y_pred_tree)
print(accuracy)
```

```
0.7845804988662132
```

In [40]:
```
confusion = confusion_matrix(y_test, y_pred_tree)
print(confusion)
```

```
[[323  57]
 [ 38  23]]
```

In [41]:
```
classification = classification_report(y_test, y_pred_tree)
print(classification)
```

```
              precision    recall  f1-score   support

          No       0.89      0.85      0.87       380
         Yes       0.29      0.38      0.33        61

    accuracy                           0.78       441
   macro avg       0.59      0.61      0.60       441
weighted avg       0.81      0.78      0.80       441
```

In [ ]:
```

```

In [ ]:
```

```