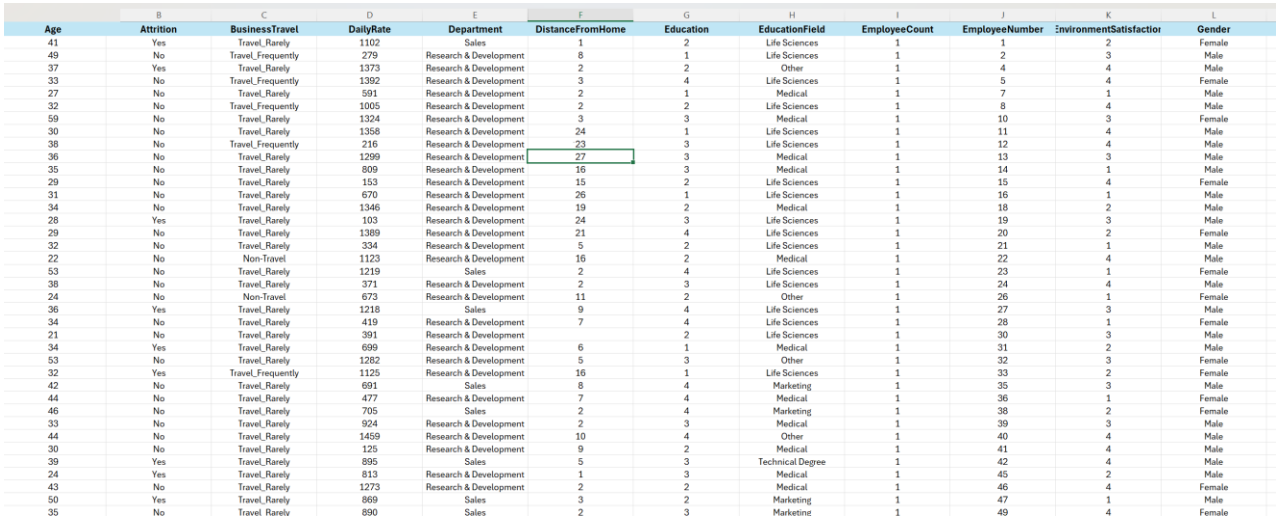
DATA ANALYST PROJECT

IBM HR Analytics Employee Attrition & Performance

Data Set:



Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber	EnvironmentSatisfactior	Gender
41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	1	2	Female
49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	1	2	3	Male
37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	1	4	4	Male
33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	1	5	4	Female
27	No	Travel_Rarely	591	Research & Development	2	1	Medical	1	7	1	Male
32	No	Travel_Frequently	1005	Research & Development	2	2	Life Sciences	1	8	4	Male
59	No	Travel_Rarely	1324	Research & Development	3	3	Medical	1	10	3	Female
30	No	Travel_Rarely	1358	Research & Development	24	1	Life Sciences	1	11	4	Male
38	No	Travel_Frequently	216	Research & Development	23	3	Life Sciences	1	12	4	Male
36	No	Travel_Rarely	1299	Research & Development	27	3	Medical	1	13	3	Male
35	No	Travel_Rarely	809	Research & Development	16	3	Medical	1	14	1	Male
29	No	Travel_Rarely	153	Research & Development	15	2	Life Sciences	1	15	4	Female
31	No	Travel_Rarely	670	Research & Development	26	1	Life Sciences	1	16	1	Male
34	No	Travel_Rarely	1346	Research & Development	19	2	Medical	1	18	2	Male
28	Yes	Travel_Rarely	103	Research & Development	24	3	Life Sciences	1	19	3	Male
29	No	Travel_Rarely	1389	Research & Development	21	4	Life Sciences	1	20	2	Female
32	No	Travel_Rarely	334	Research & Development	5	2	Life Sciences	1	21	1	Male
22	No	Non-Travel	1123	Research & Development	16	2	Medical	1	22	4	Male
53	No	Travel_Rarely	1219	Sales	2	4	Life Sciences	1	23	1	Female
38	No	Travel_Rarely	371	Research & Development	2	3	Life Sciences	1	24	4	Male
24	No	Non-Travel	673	Research & Development	11	2	Other	1	26	1	Female
36	Yes	Travel_Rarely	1218	Sales	9	4	Life Sciences	1	27	3	Male
34	No	Travel_Rarely	419	Research & Development	7	4	Life Sciences	1	28	1	Female
21	No	Travel_Rarely	391	Research & Development		2	Life Sciences	1	30	3	Male
34	Yes	Travel_Rarely	699	Research & Development	6	1	Medical	1	31	2	Male
53	No	Travel_Rarely	1282	Research & Development	5	3	Other	1	32	3	Female
32	Yes	Travel_Frequently	1125	Research & Development	16	1	Life Sciences	1	33	2	Female
42	No	Travel_Rarely	691	Sales	8	4	Marketing	1	35	3	Male
44	No	Travel_Rarely	477	Research & Development	7	4	Medical	1	36	1	Female
46	No	Travel_Rarely	705	Sales	2	4	Marketing	1	38	2	Female
33	No	Travel_Rarely	924	Research & Development	2	3	Medical	1	39	3	Male
44	No	Travel_Rarely	1459	Research & Development	10	4	Other	1	40	4	Male
30	No	Travel_Rarely	125	Research & Development	9	2	Medical	1	41	4	Male
39	Yes	Travel_Rarely	895	Sales	5	3	Technical Degree	1	42	4	Male
24	Yes	Travel_Rarely	813	Research & Development	1	3	Medical	1	45	2	Male
43	No	Travel_Rarely	1273	Research & Development	2	2	Medical	1	46	4	Female
50	Yes	Travel_Rarely	869	Sales	3	2	Marketing	1	47	1	Male
35	No	Travel_Rarely	890	Sales	2	3	Marketing	1	49	4	Female

Importing python libraries and setting the pandas options up to 35 columns:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')

#setting pandas options(maximum coloumns options)
pd.set_option('display.max_columns', 35)
```

Cleaning the data set (checking for null and duplicate values)

df.isnull().sum()

df.duplicated().sum()

```
Age                         0
Attrition                   0
BusinessTravel              0
DailyRate                   0
Department                  0
DistanceFromHome            1
Education                   0
EducationField              0
EmployeeCount               0
EmployeeNumber              0
EnvironmentSatisfaction     0
```

Distribution of Attrition

#Univariate Analysis

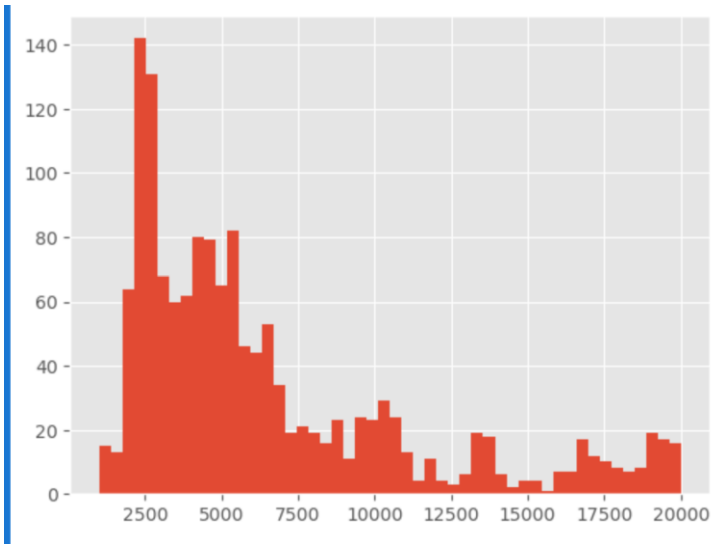
df['MonthlyIncome'].hist(bins=50)

plt.show()

Count plot for categorical data

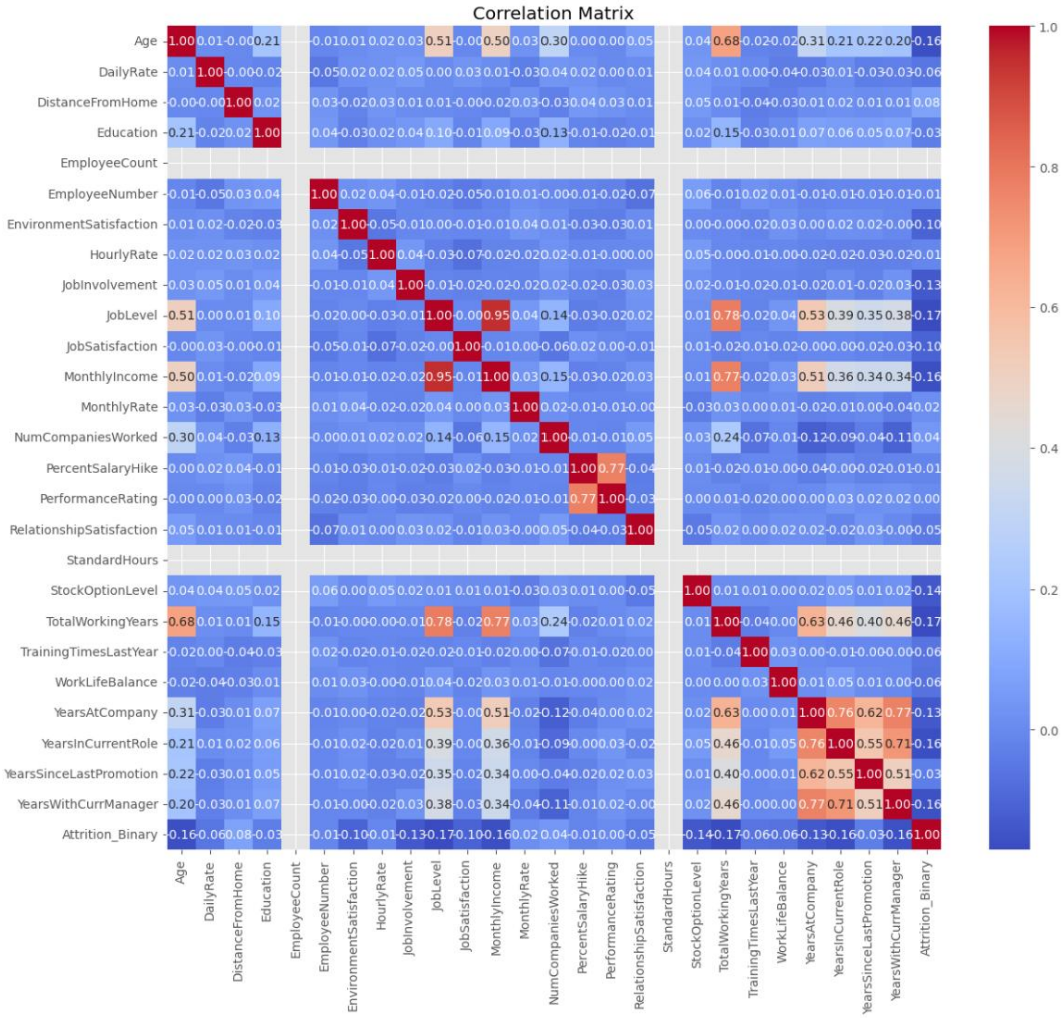
sns.countplot(x='Gender', data=df)

plt.show()

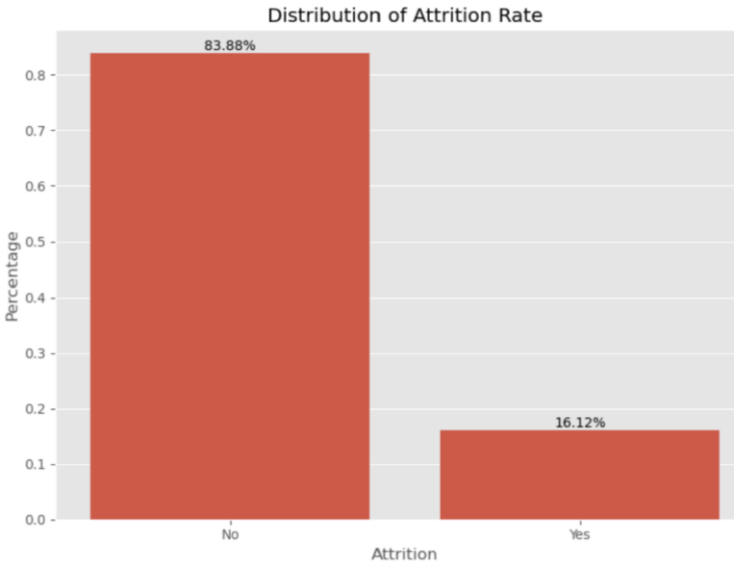


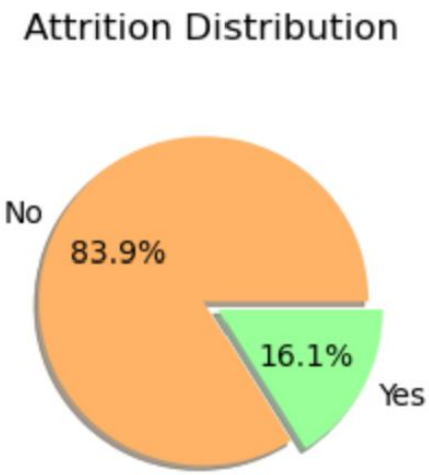
Correlation matrix and Heatmap

```
# Step 1: Encode categorical variables if needed
df['Attrition_Binary'] = df['Attrition'].map({'Yes': 1, 'No': 0})
# Step 2: Select only numerical columns
numerical_df = df.select_dtypes(include=['number'])
# Step 3: Compute correlation matrix
corr = numerical_df.corr()
# Step 4: Display correlation matrix
print(corr)
# Step 5: (Optional) Plot heatmap for visualization
import seaborn as sns
import matplotlib.pyplot as plt
plt.figure(figsize=(14, 12))
sns.heatmap(corr, annot=True, cmap='coolwarm', fmt='.2f', linewidths=0)
plt.title('Correlation Matrix')
plt.show()
```

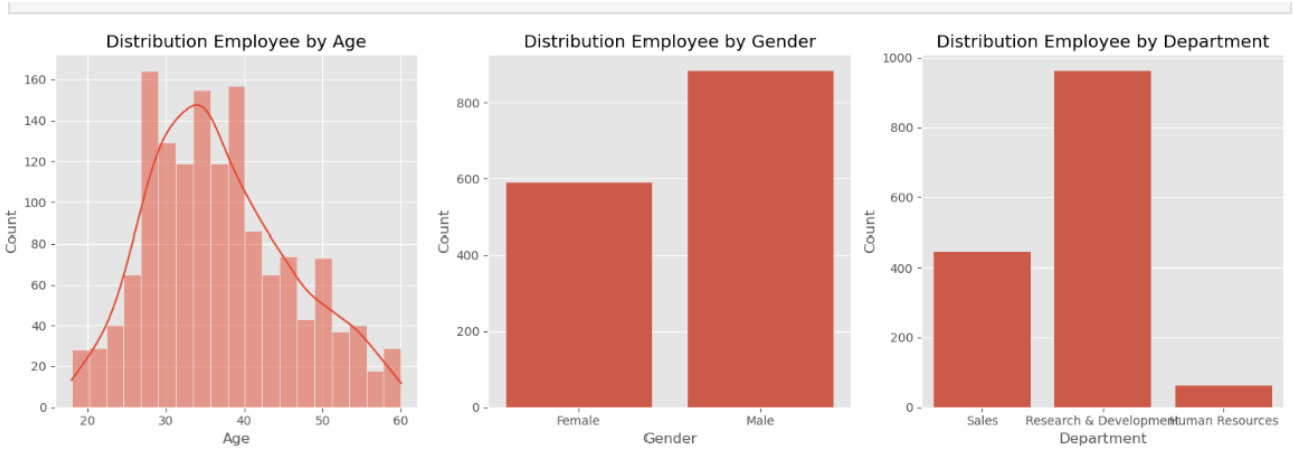


Distribution of Attrition rate in the company

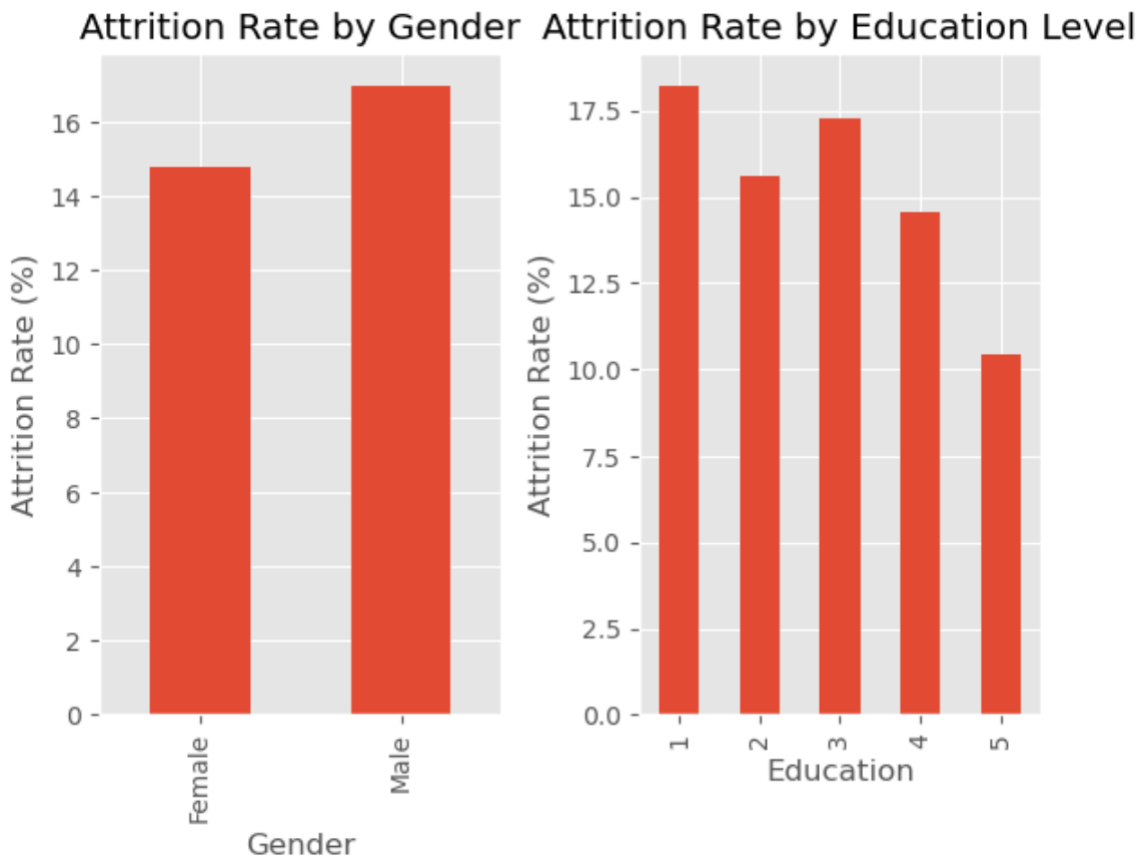




Employee Demographics



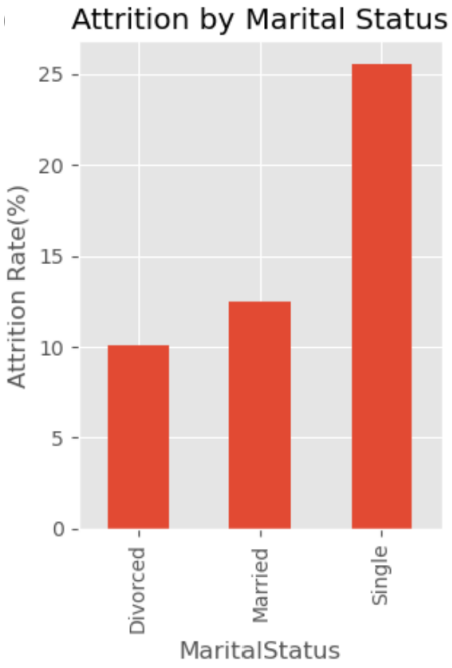
Distribution of attrition by gender and education



1 Younger workers seem to be more prone than other age groups to quit a company, particularly those between the ages of 30 and 35. A more alluring work offer elsewhere, discontent with the pay or career path, or a desire for fresh experiences could all be contributing causes.

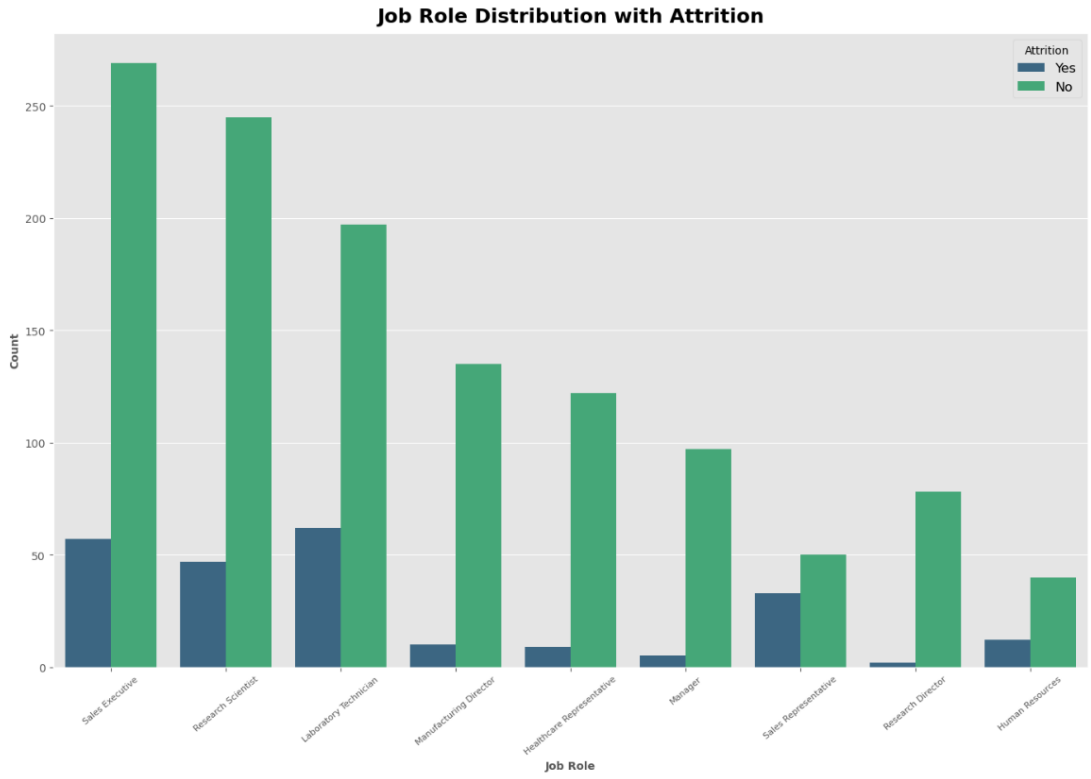
2. Job stability is generally higher among older workers. The presence of mandated retirement benefits, the difficulty of finding a new employment at an older age, or a higher level of devotion to the organization could all be contributing factors.

Attrition by Maritial Status:



As we can see the attrition rate is higher in singles than compared to divorced or married people.

Attrition with respect to the job role in the company:



Data Processing

1)Feature Selection

```
# Drop columns that are non-relevant to the analysis
columns_to_drop = ['EmployeeCount', 'Over18', 'StandardHours' , 'EmployeeNumber
df_cleaned = df.drop(columns=columns_to_drop)

# Check the cleaned DataFrame
print(f"Cleaned DataFrame shape: {df_cleaned.shape}")
print(f"Remaining columns: {df_cleaned.columns.tolist()}")
```

*2)*Categorical Conversion (to make categorical data interpretable for machine learning algorithms)

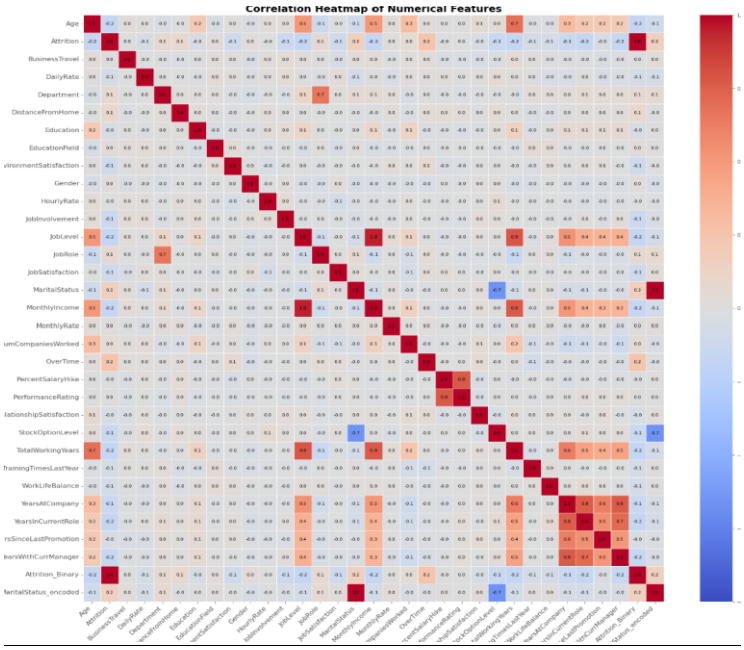
```
from sklearn.preprocessing import LabelEncoder

# Initialize LabelEncoder
label_encoder = LabelEncoder()

# Apply LabelEncoder to each categorical column and create new columns for enco
for col in categorical_columns:
    # Create a new column with the encoded values
    df_cleaned['OverTime'] = label_encoder.fit_transform(df_cleaned['OverTime']

# Display the updated DataFrame
df_cleaned.head(10)
```

3) Heatmap


Correlation Heatmap of Numerical Features
Age
Attrition
BusinessTravel
DailyRate
Department
DistanceFromHome
Education
EducationField
EnvironmentSatisfaction
Gender
HourlyRate
JobInvolvement
JobLevel
JobRole
JobSatisfaction
MaritalStatus
MonthlyIncome
MonthlyRate
NumCompaniesWorked
OverTime
PercentSalaryHike
PerformanceRating
RelationshipSatisfaction
StockOptionLevel
TotalWorkingYears
TrainingTimesLastYear
WorkLifeBalance
YearsAtCompany
YearsInCurrentRole
YearsSinceLastPromotion
YearsWithCurrManager
Attrition_Binary
MaritalStatus_encoded