```
from google.colab import files
uploaded = files.upload()
```

```
Choose files WA_Fn-Use…-Attrition.csv
WA_Fn-UseC_-HR-Employee-Attrition.csv(text/csv) - 227977 bytes, last modified: 06/10/2025 - 100% done
Saving WA_Fn-UseC_-HR-Employee-Attrition.csv to WA_Fn-UseC_-HR-Employee-Attrition (1).csv
```

```
import pandas as pd
from sklearn.preprocessing import LabelEncoder, StandardScaler

# Load dataset
df = pd.read_csv("/content/WA_Fn-UseC_-HR-Employee-Attrition.csv")
```

```
print("Missing values before handling:\n", df.isnull().sum())
```

```
Missing values before handling:
 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
```

```
# Fill missing numerical columns with median
num_cols = df.select_dtypes(include=['int64', 'float64']).columns
for col in num_cols:
    df[col] = df[col].fillna(df[col].median())

# Fill missing categorical columns with mode
cat_cols = df.select_dtypes(include=['object']).columns
for col in cat_cols:
    df[col] = df[col].fillna(df[col].mode()[0])

print("\nMissing values after handling:\n", df.isnull().sum())
```

```
Missing values after handling:
 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
```

```
# These columns are usually constant in the IBM HR dataset
df = df.drop(['EmployeeCount', 'StandardHours', 'Over18', 'EmployeeNumber'], axis=1, errors='ignore')
```

```
# Identify categorical columns again after dropping
cat_cols = df.select_dtypes(include=['object']).columns
```

```
le = LabelEncoder()
for col in cat_cols:
    df[col] = le.fit_transform(df[col])

print("\nCategorical columns encoded successfully.")
```

```

Categorical columns encoded successfully.
```

```
scaler = StandardScaler()

# Identify numerical columns
num_cols = df.select_dtypes(include=['int64', 'float64']).columns

# Scale numeric columns
df[num_cols] = scaler.fit_transform(df[num_cols])

print("\nNumeric columns normalized successfully.")
```

```

Numeric columns normalized successfully.
```

```
print("\nFinal dataset shape:", df.shape)
print("\nFinal columns:\n", df.columns.tolist())
```

```

Final dataset shape: (1470, 31)
```

```
Final columns:
 ['Age', 'Attrition', 'BusinessTravel', 'DailyRate', 'Department', 'DistanceFromHome', 'Education', 'Ed
```

```
df.to_csv("HR_Attrition_Preprocessed.csv", index=False)
print("\n✅ Preprocessed dataset saved as HR_Attrition_Preprocessed.csv")
```

```
✅ Preprocessed dataset saved as HR_Attrition_Preprocessed.csv
```

```
# Save cleaned dataset to your device
cleaned_file_path = "cleaned_hr_dataset.csv"   # You can change the filename/path as needed
df.to_csv(cleaned_file_path, index=False)

print(f"✅ Cleaned dataset saved successfully as '{cleaned_file_path}'")
```

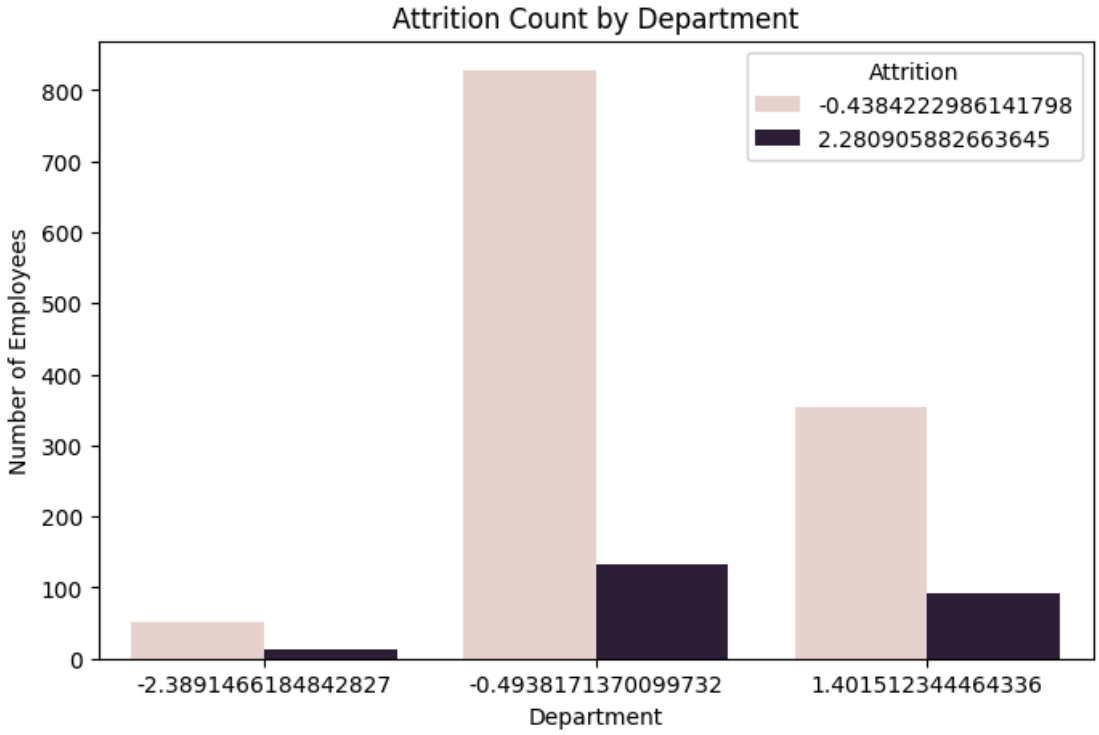
```
✅ Cleaned dataset saved successfully as 'cleaned_hr_dataset.csv'
```

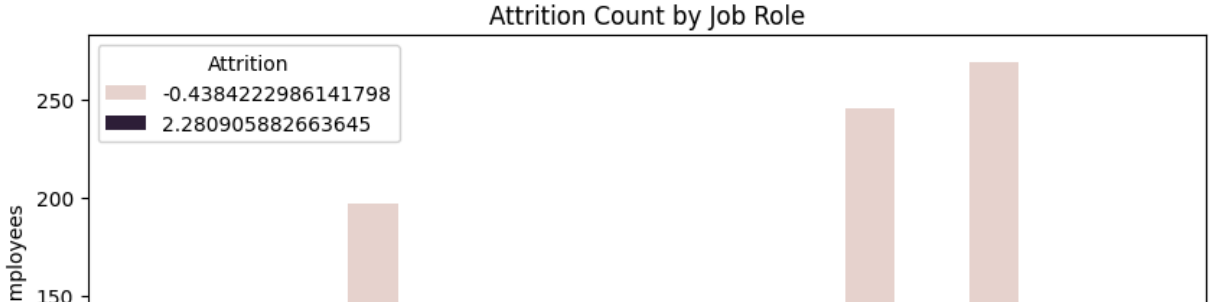
```
#EDA PROCESS

# ------------------------------
# HR Attrition EDA - Complete
# ------------------------------

import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Load cleaned dataset
df = pd.read_csv("/content/cleaned_hr_dataset.csv")

# ------------------------------
# 1️⃣ Basic Overview
# ------------------------------
print("Dataset Info:\n")
print(df.info())

print("\nDataset Description:\n")
print(df.describe())

print("\nAttrition Counts:\n")
print(df['Attrition'].value_counts())

# ------------------------------
# 2️⃣ Attrition by Department
# ------------------------------
plt.figure(figsize=(8,5))
sns.countplot(x='Department', hue='Attrition', data=df)
plt.title('Attrition Count by Department')
plt.ylabel('Number of Employees')
plt.show()

# ------------------------------
# 3️⃣ Attrition by Job Role
# ------------------------------
plt.figure(figsize=(10,5))
sns.countplot(x='JobRole', hue='Attrition', data=df)
plt.xticks(rotation=45)
plt.title('Attrition Count by Job Role')
plt.ylabel('Number of Employees')
plt.show()

# ------------------------------
# 4️⃣ Attrition vs Monthly Income
# ------------------------------
plt.figure(figsize=(8,5))
sns.boxplot(x='Attrition', y='MonthlyIncome', data=df)
```

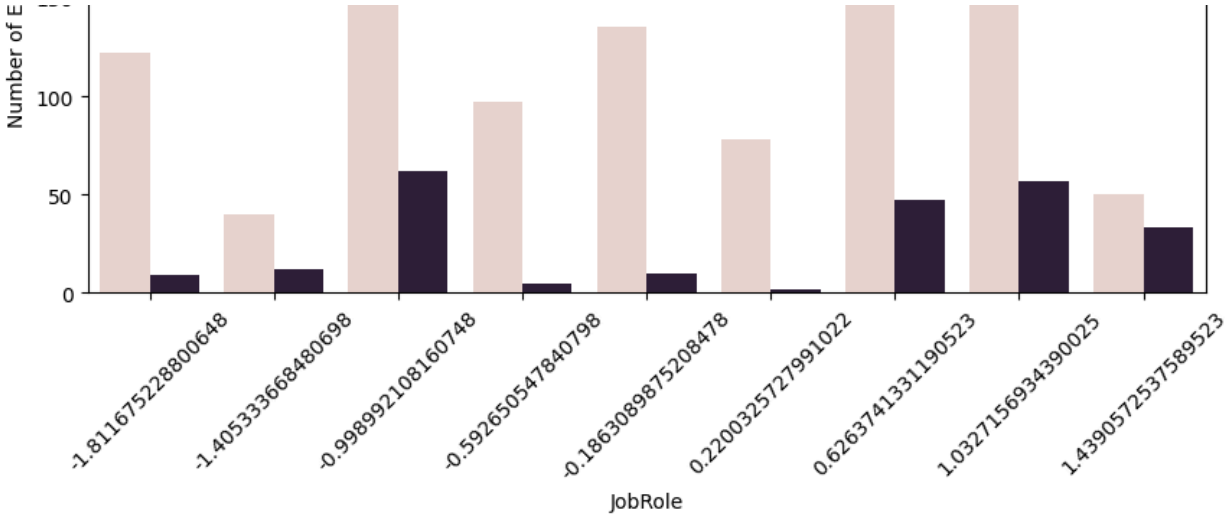
```
plt.title('Attrition vs Monthly Income')
plt.show()

# ------------------------------
# 5️⃣ Attrition vs Years at Company
# ------------------------------
plt.figure(figsize=(8,5))
sns.boxplot(x='Attrition', y='YearsAtCompany', data=df)
plt.title('Attrition vs Years at Company')
plt.show()

# ------------------------------
# 6️⃣ Attrition by Marital Status
# ------------------------------
plt.figure(figsize=(8,5))
sns.countplot(x='MaritalStatus', hue='Attrition', data=df)
plt.title('Attrition Count by Marital Status')
plt.show()

# ------------------------------
# 7️⃣ Attrition by Work-Life Balance
# ------------------------------
plt.figure(figsize=(8,5))
sns.boxplot(x='Attrition', y='WorkLifeBalance', data=df)
plt.title('Attrition vs Work-Life Balance')
plt.show()

# ------------------------------
# 8️⃣ Attrition vs OverTime
# ------------------------------
plt.figure(figsize=(6,4))
sns.countplot(x='OverTime', hue='Attrition', data=df)
plt.title('Attrition by OverTime')
plt.show()

# ------------------------------
# 9️⃣ Correlation Heatmap (numeric features)
# ------------------------------
# ------------------------------
# 1️⃣ Correlation Matrix
# ------------------------------
corr_matrix = df.corr()

# ------------------------------
# 2️⃣ Sort Features by correlation with 'Attrition'
# ------------------------------
corr_target = corr_matrix['Attrition'].sort_values(ascending=False)
top_features = corr_target.index  # all features sorted by relevance

# Optional: You can select top N features only (e.g., top 12)
top_features = corr_target.index[:12]

# ------------------------------
# 3️⃣ Plot Heatmap
# ------------------------------
plt.figure(figsize=(12,8))  # larger figure to avoid overlapping
sns.heatmap(df[top_features].corr(), annot=True, cmap='coolwarm', fmt=".2f", linewidths=0.5)
plt.xticks(rotation=45, ha='right')  # Rotate x labels
plt.yticks(rotation=0)               # Keep y labels horizontal
plt.title('Correlation Heatmap - Top Features Related to Attrition', fontsize=16)
plt.show()
```

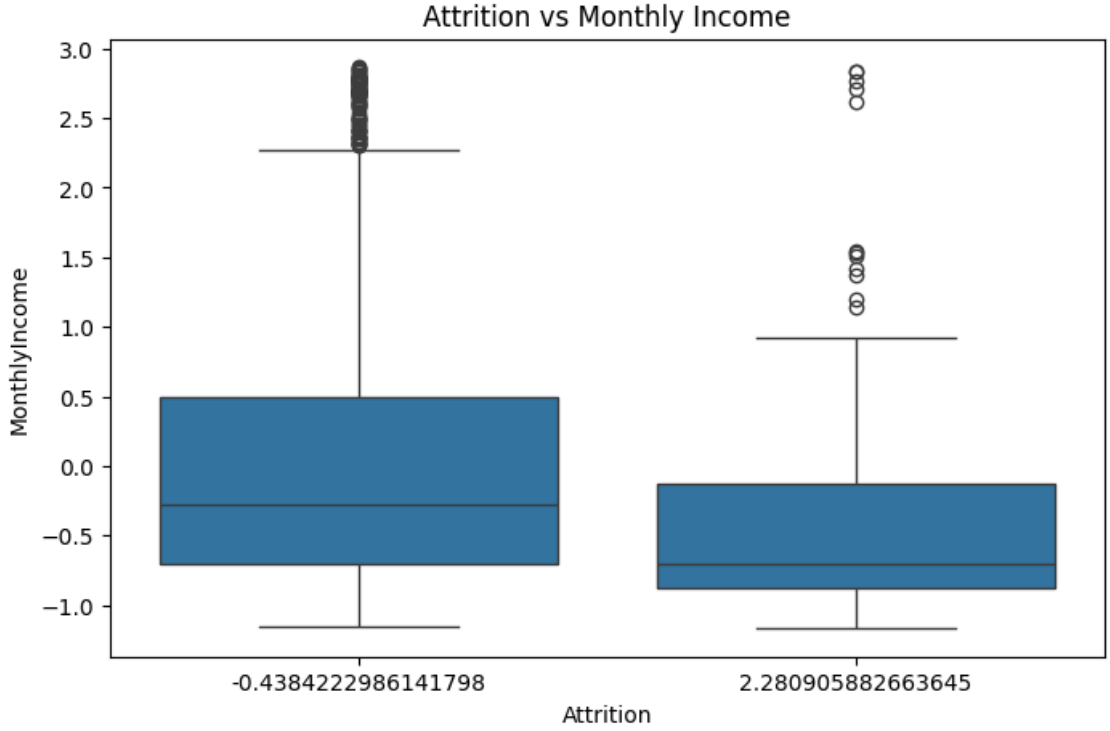
```
Dataset Info:

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1470 entries, 0 to 1469
Data columns (total 31 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Age                       1470 non-null   float64
 1   Attrition                 1470 non-null   float64
 2   BusinessTravel            1470 non-null   float64
 3   DailyRate                 1470 non-null   float64
 4   Department                1470 non-null   float64
 5   DistanceFromHome          1470 non-null   float64
 6   Education                 1470 non-null   float64
 7   EducationField            1470 non-null   float64
 8   EnvironmentSatisfaction   1470 non-null   float64
 9   Gender                    1470 non-null   float64
 10  HourlyRate                1470 non-null   float64
 11  JobInvolvement            1470 non-null   float64
 12  JobLevel                  1470 non-null   float64
 13  JobRole                   1470 non-null   float64
 14  JobSatisfaction           1470 non-null   float64
 15  MaritalStatus             1470 non-null   float64
 16  MonthlyIncome             1470 non-null   float64
 17  MonthlyRate               1470 non-null   float64
 18  NumCompaniesWorked        1470 non-null   float64
 19  OverTime                  1470 non-null   float64
 20  PercentSalaryHike         1470 non-null   float64
 21  PerformanceRating         1470 non-null   float64
 22  RelationshipSatisfaction  1470 non-null   float64
 23  StockOptionLevel          1470 non-null   float64
 24  TotalWorkingYears         1470 non-null   float64
 25  TrainingTimesLastYear     1470 non-null   float64
 26  WorkLifeBalance           1470 non-null   float64
 27  YearsAtCompany            1470 non-null   float64
 28  YearsInCurrentRole        1470 non-null   float64
 29  YearsSinceLastPromotion   1470 non-null   float64
 30  YearsWithCurrManager      1470 non-null   float64
dtypes: float64(31)
memory usage: 356.1 KB
None

Dataset Description:

                Age     Attrition  BusinessTravel     DailyRate    Department  \
count  1.470000e+03  1.470000e+03    1.470000e+03  1.470000e+03  1.470000e+03   
mean  -4.229421e-17  1.498423e-16   -6.042030e-17  4.833624e-17 -8.700523e-17   
std    1.000340e+00  1.000340e+00    1.000340e+00  1.000340e+00  1.000340e+00   
min   -2.072192e+00 -4.384223e-01   -2.416437e+00 -1.736576e+00 -2.389147e+00   
25%   -7.581700e-01 -4.384223e-01   -9.131944e-01 -8.366616e-01 -4.938171e-01   
50%   -1.011589e-01 -4.384223e-01    5.900483e-01 -1.204135e-03 -4.938171e-01   
75%    6.653541e-01 -4.384223e-01    5.900483e-01  8.788772e-01  1.401512e+00   
max    2.526886e+00  2.280906e+00    5.900483e-01  1.726730e+00  1.401512e+00   

       DistanceFromHome     Education  EducationField  \
count      1.470000e+03  1.470000e+03    1.470000e+03   
mean       4.833624e-18 -4.350262e-17    3.595008e-17   
std        1.000340e+00  1.000340e+00    1.000340e+00   
min       -1.010909e+00 -1.868426e+00   -1.688776e+00   
25%       -8.875151e-01 -8.916883e-01   -9.374137e-01   
50%       -2.705440e-01  8.504925e-02   -1.860516e-01   
75%        5.932157e-01  1.061787e+00    5.653105e-01   
max        2.444129e+00  2.038524e+00    2.068035e+00   

       EnvironmentSatisfaction        Gender  ...  PerformanceRating  \
count             1.470000e+03  1.470000e+03  ...       1.470000e+03   
mean              7.612958e-17  4.350262e-17  ...      -5.607004e-16   
std               1.000340e+00  1.000340e+00  ...       1.000340e+00   
min              -1.575686e+00 -1.224745e+00  ...      -4.262300e-01   
25%              -6.605307e-01 -1.224745e+00  ...      -4.262300e-01   
50%               2.546249e-01  8.164966e-01  ...      -4.262300e-01   
75%               1.169781e+00  8.164966e-01  ...      -4.262300e-01   
max               1.169781e+00  8.164966e-01  ...       2.346151e+00   

       RelationshipSatisfaction  StockOptionLevel  TotalWorkingYears  \
```

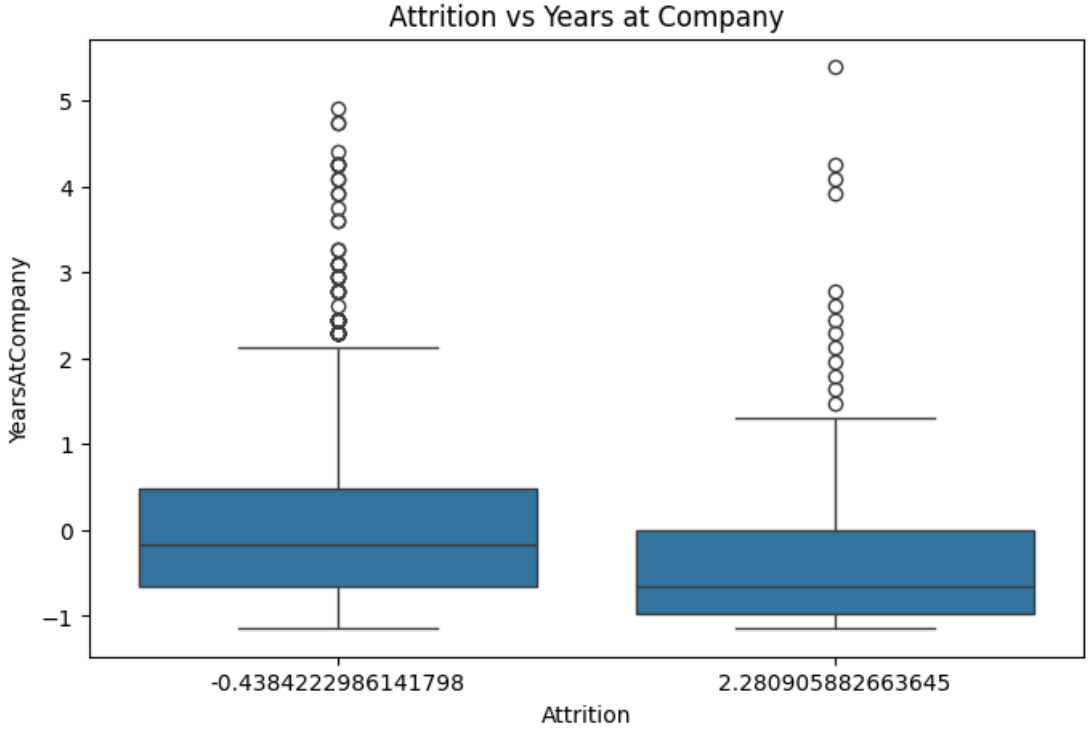
```
count             1.470000e+03      1.470000e+03      1.470000e+03
mean              1.450087e-17      7.733798e-17      1.208406e-17
std               1.000340e+00      1.000340e+00      1.000340e+00
min              -1.584178e+00     -9.320144e-01     -1.450167e+00
25%              -6.589728e-01     -9.320144e-01     -6.787735e-01
50%               2.662326e-01      2.419883e-01     -1.645114e-01
75%               1.191438e+00      2.419883e-01      4.783162e-01
max               1.191438e+00      2.589994e+00      3.692454e+00

       TrainingTimesLastYear  WorkLifeBalance  YearsAtCompany  \
count           1.470000e+03     1.470000e+03    1.470000e+03
mean            8.700523e-17    -4.350262e-17   -1.570928e-17
std             1.000340e+00     1.000340e+00    1.000340e+00
min            -2.171982e+00    -2.493820e+00   -1.144294e+00
25%            -6.201892e-01    -1.077862e+00   -6.544537e-01
50%             1.557071e-01     3.380962e-01   -3.278933e-01
75%             1.557071e-01     3.380962e-01    3.252275e-01
max             2.483396e+00     1.754054e+00    5.386914e+00

       YearsInCurrentRole  YearsSinceLastPromotion  YearsWithCurrManager
count        1.470000e+03             1.470000e+03          1.470000e+03
mean         1.015061e-16             1.450087e-17         -2.054290e-17
std          1.000340e+00             1.000340e+00          1.000340e+00
min         -1.167687e+00            -6.791457e-01         -1.155935e+00
25%         -6.154916e-01            -6.791457e-01         -5.952272e-01
50%         -3.393937e-01            -3.687153e-01         -3.148735e-01
75%          7.649976e-01             2.521455e-01          8.065415e-01
max          3.802074e+00             3.977310e+00          3.610079e+00

[8 rows x 31 columns]

Attrition Counts:

Attrition
-0.438422    1233
 2.280906     237
Name: count, dtype: int64
```

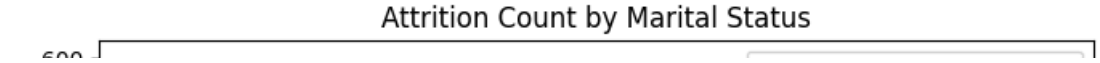


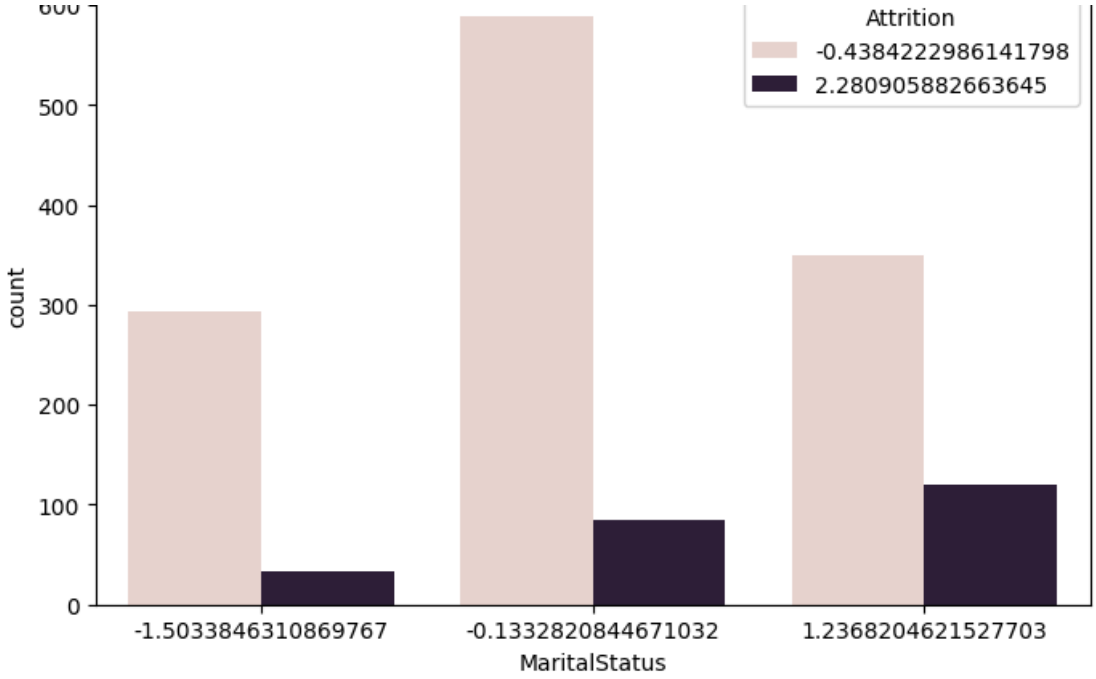


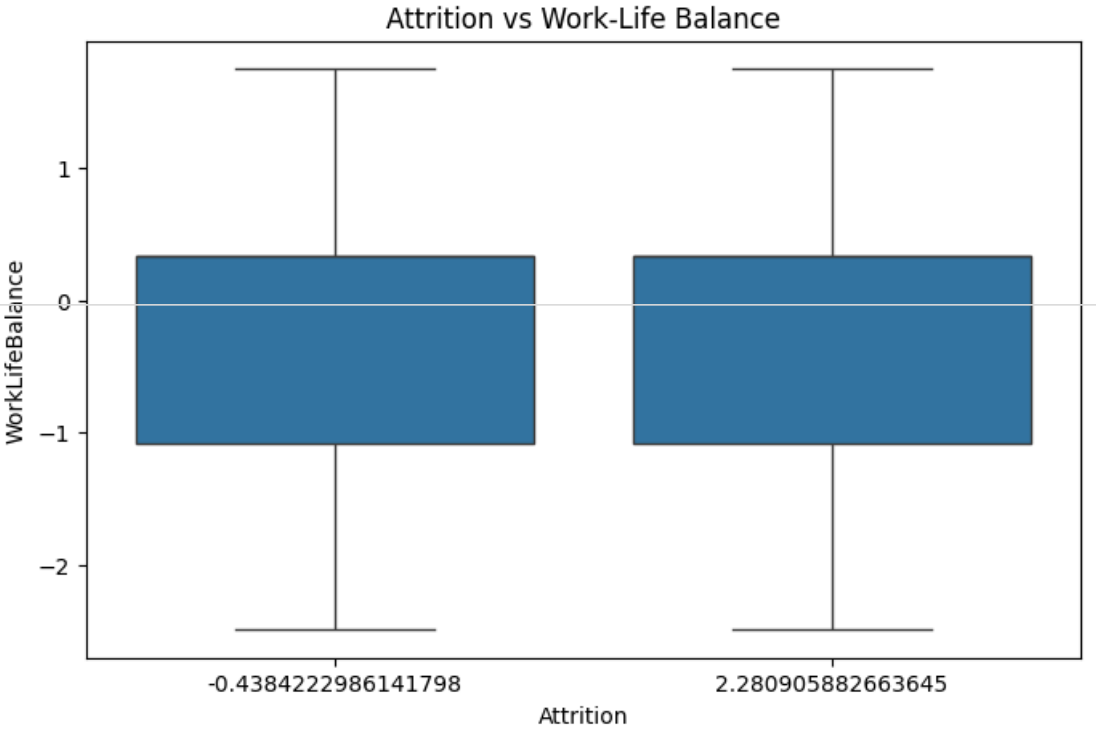

Number of E
100
50
0
-1.81167228800648
-1.40533668480698
-0.99899210816074
-0.59265054784079
-0.18630898752084
0.22003257279910
0.62637413311905
1.03271569343900
1.43905725375895
JobRole

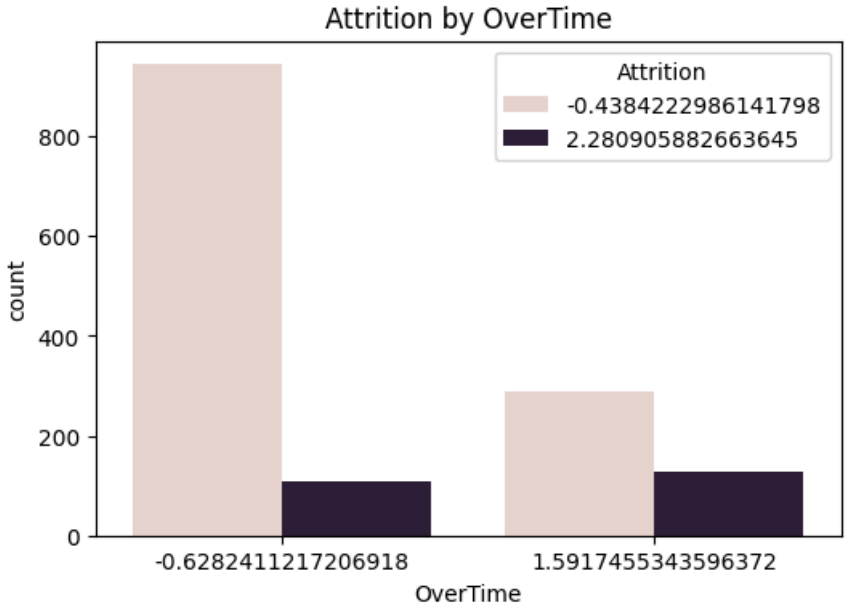

Attrition vs Monthly Income
MonthlyIncome
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
-0.4384222986141798
2.280905882663645
Attrition


Attrition vs Years at Company
YearsAtCompany
5
4
3
2
1
0
−1
-0.4384222986141798
2.280905882663645
Attrition


Attrition Count by Marital Status







Correlation Heatmap - Top Features Related to Attrition