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

In [ ]:
```
data = pd.read_csv('/content/drive/MyDrive/unified projects/ibm/WA_Fn-Use(
```

In [ ]:
```
data.head()
```

Out[ ]:

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

5 rows × 35 columns

In [ ]:
```
data.shape
```

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

In [ ]:
```
data.info()              #checking all colums have appropriate dataTypes
```

```
<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
```

Allcolumnshaveappropriatedatatypes,ensuringthat thedataiscorrectlyformattedforanalysis.

In [ ]:
```
# Standardize column names
data.columns = data.columns.str.lower().str.replace(' ', '_')

# Display the updated column names
print(data.columns)
```

```
Index(['age', 'attrition', 'businesstravel', 'dailyrate', 'department',
       'distancefromhome', 'education', 'educationfield', 'employeecount',
       'employeenumber', 'environmentsatisfaction', 'gender', 'hourlyrate',
       'jobinvolvement', 'joblevel', 'jobrole', 'jobsatisfaction',
       'maritalstatus', 'monthlyincome', 'monthlyrate', 'numcompaniesworked',
       'over18', 'overtime', 'percentsalaryhike', 'performancerating',
       'relationshipsatisfaction', 'standardhours', 'stockoptionlevel',
       'totalworkingyears', 'trainingtimeslastyear', 'worklifebalance',
       'yearsatcompany', 'yearsincurrentrole', 'yearssincelastpromotion',
       'yearswithcurrmanager'],
      dtype='object')
```

In [ ]:
```
pd.set_option('display.max_columns', 35)          #making the all th
```

In [ ]:
```
data.head(10)
```

Out[ ]:

	age	attrition	businesstravel	dailyrate	department	distancefromhome	education
0	41	Yes	Travel_Rarely	1102	Sales	1	2
1	49	No	Travel_Frequently	279	Research & Development	8	1
2	37	Yes	Travel_Rarely	1373	Research & Development	2	2
3	33	No	Travel_Frequently	1392	Research & Development	3	4
4	27	No	Travel_Rarely	591	Research & Development	2	1
5	32	No	Travel_Frequently	1005	Research & Development	2	2
6	59	No	Travel_Rarely	1324	Research & Development	3	3
7	30	No	Travel_Rarely	1358	Research & Development	24	1
8	38	No	Travel_Frequently	216	Research & Development	23	3
9	36	No	Travel_Rarely	1299	Research & Development	27	3

In [ ]:
```
print(f'Number of duplicated data: {data.duplicated().sum()}')      #checl
```

```
Number of duplicated data: 0
```

In [ ]:
```
df = data
```

In [ ]:
```
df.isnull().sum() / len(df) * 100         #checking for null values
```

Out[ ]:

	0
age	0.0
attrition	0.0
businesstravel	0.0
dailyrate	0.0
department	0.0
distancefromhome	0.0
education	0.0
educationfield	0.0
employeecount	0.0
employeenumber	0.0
environmentsatisfaction	0.0
gender	0.0
hourlyrate	0.0
jobinvolvement	0.0
joblevel	0.0
jobrole	0.0
jobsatisfaction	0.0
maritalstatus	0.0
monthlyincome	0.0
monthlyrate	0.0
numcompaniesworked	0.0
over18	0.0
overtime	0.0
percentsalaryhike	0.0
performancerating	0.0
relationshipsatisfaction	0.0
standardhours	0.0
stockoptionlevel	0.0
totalworkingyears	0.0
trainingtimeslastyear	0.0
worklifebalance	0.0
yearsatcompany	0.0
yearsincurrentrole	0.0
yearssincelastpromotion	0.0
yearswithcurrmanager	0.0

In [ ]:
```
attrition = df['attrition'].value_counts(normalize=True)* 100
```

In [ ]:
```
attrition
```

Out[ ]:

	proportion
attrition	
No	83.877551
Yes	16.122449

dtype: float64

In [ ]:
```
plt.figure(figsize = (8,6))
ax = sns.barplot(x = attrition.index, y = attrition)
for p in ax.patches:
  ax.annotate(f'{p.get_height():.2f}%',                                    #Attrition
              (p.get_x() + p.get_width() / 2.,
               p.get_height()), ha='center', va='bottom')
plt.title('Attrition Distributiion')
plt.xlabel('attrition')
plt.ylabel('count')
plt.show
```

Out[ ]:

matplotlib.pyplot.show

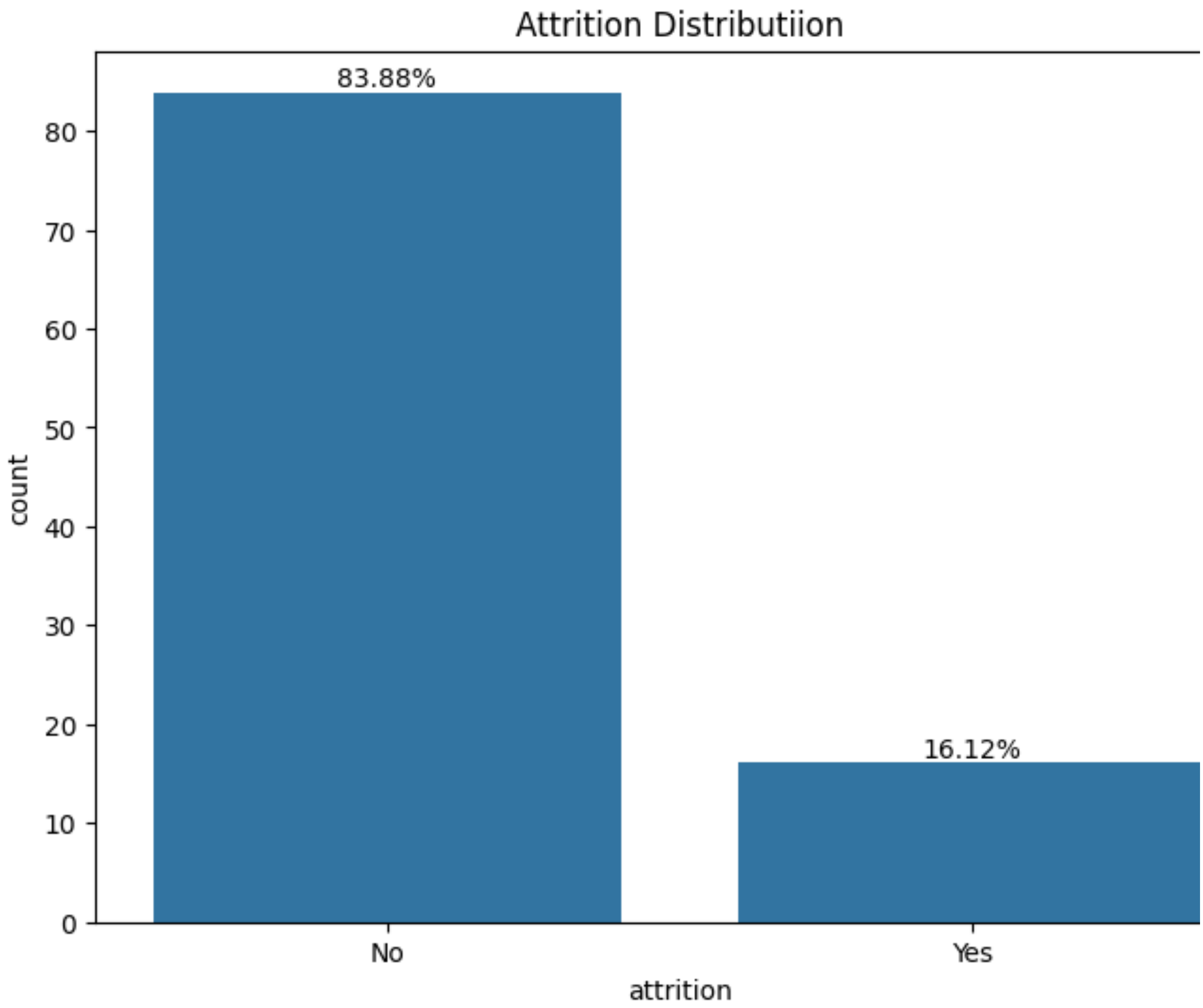
def show(*args, **kwargs)

Display all open figures.

Parameters

----

block : bool, optional



Based on the analysis, the company's attrition rate is 16.12%. This means that about 16.12% of t
employees decided to leave the company during the analyzed period.

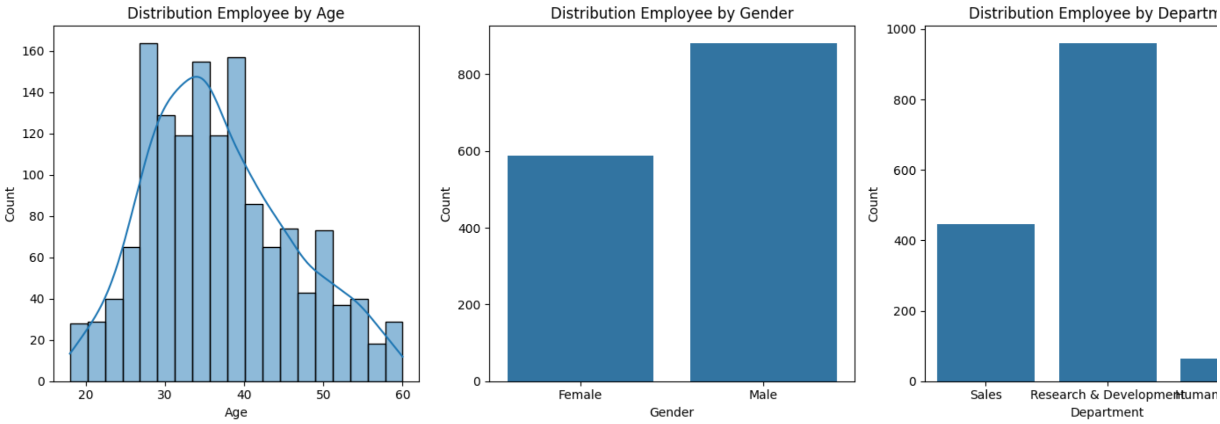
average of tenure

In [ ]:
```
avg_tenure = df['yearsatcompany'].mean().round(2)
```

In [ ]:
```
print(f'average years of employee at the company {avg_tenure} years')
```

```
average years of employee at the company 7.01 years
```

In [ ]:
```
fig, axes = plt.subplots(nrows=1, ncols=3, figsize=(15,5))

sns.histplot(data=df, x='age', kde=True, ax=axes[0])
axes[0].set_title('Distribution Employee by Age')
axes[0].set_xlabel('Age')
axes[0].set_ylabel('Count')

sns.countplot(data=df, x='gender', ax=axes[1])
axes[1].set_title('Distribution Employee by Gender')
axes[1].set_xlabel('Gender')
axes[1].set_ylabel('Count')

sns.countplot(data=df, x='department', ax=axes[2])
axes[2].set_title('Distribution Employee by Department')
axes[2].set_xlabel('Department')
axes[2].set_ylabel('Count')

plt.tight_layout()
plt.show()
```



1. Age: Most of the company's employees are in the 30-35 age group. This indicates that the c has manyemployeeswhoareataproductiveandexperiencedage.
2. Gender:Themajorityofemployeesat thiscompanyaremale.Therearesignificantlymoremale employeesthanfemaleemployees.
3. Department:Mostof thecompany'semployeesareconcentratedintheresearchanddevelopment department.Thisindicatesthat thecompanyisheavilyfocusedonproductorserviceresearchand developmentactivities.

In [ ]:
```
df_attrition = df[df['attrition'] == 'Yes']
```

In [ ]:
```
df_attrition.head()
```

Out[ ]:

	age	attrition	businesstravel	dailyrate	department	distancefromhome	education	e
0	41	Yes	Travel_Rarely	1102	Sales	1	2	
2	37	Yes	Travel_Rarely	1373	Research & Development	2	2	
14	28	Yes	Travel_Rarely	103	Research & Development	24	3	
21	36	Yes	Travel_Rarely	1218	Sales	9	4	
24	34	Yes	Travel_Rarely	699	Research & Development	6	1	

In [ ]:
```
df_attrition.shape
```

Out[ ]:
```
(237, 35)
```

In [ ]:
```
def calculate_attrition_rate(df, column):
  attrition_counts = df.groupby([column,'attrition']).size().unstack(fill_
  attrition_rate = attrition_counts['Yes'] / attrition_counts.sum(axis=1)
  attrition_rate_df = attrition_rate.reset_index()
  attrition_rate_df.columns = [column,'attritionrate']
  return attrition_rate_df
```

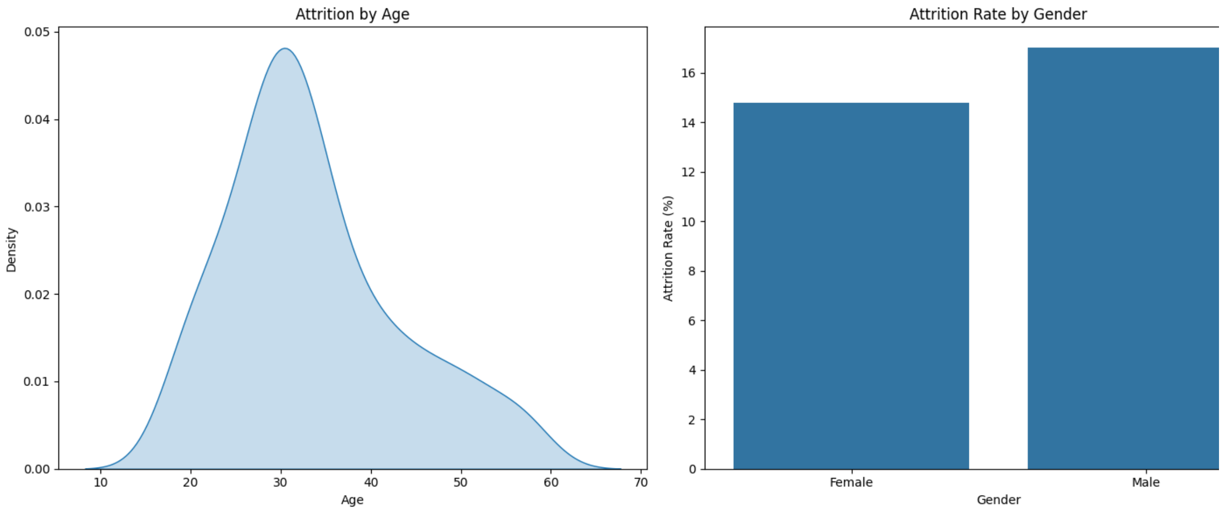
In [ ]:
```
attrition_rate_by_department = calculate_attrition_rate(df, 'department')
```

In [ ]:
```
attrition_rate_by_department
```

Out[ ]:

	department	attritionrate
0	Human Resources	19.05
1	Research & Development	13.84
2	Sales	20.63

In [ ]:
```
fig, axes = plt.subplots(nrows=1, ncols=2, figsize=(15,6))

# Plot 1: KDE plot of Age with Attrition
sns.kdeplot(data=df_attrition, x='age', fill=True, ax=axes[0])
axes[0].set_title('Attrition by Age')
axes[0].set_xlabel('Age')
axes[0].set_ylabel('Density')

# Plot 2: Bar plot of Gender count with Attrition
attrition_rate_by_gender = calculate_attrition_rate(df, 'gender')
sns.barplot(data=attrition_rate_by_gender, x='gender', y='attritionrate',
axes[1].set_title(f'Attrition Rate by Gender')
axes[1].set_xlabel('Gender')
axes[1].set_ylabel('Attrition Rate (%)')

plt.tight_layout()
plt.show()
```



1. Younger employees, especially those in the 30-35 age group, appear to be more likely than age groups to leave a company. This could be due to a number of factors, including a searcl experiences, dissatisfaction with salary or career path, or a more attractive job offer elsewhe

2. Older employees tend to have greater job stability. This may be due to a number of factors, s a higher level of commitment to the company, the difficulty of finding a new job at an older ag the existence of mandatory retirement benefits. Attrition by Gender

In [ ]:
```
education_mapping = {1 : 'Below College',
                     2 : 'College',
                     3 : 'Bachelor',
                     4 : 'Master',
                     5 : 'Doctor'}

df['education_cat'] = df['education'].replace(education_mapping)
df['education_cat']
```

Out[ ]:

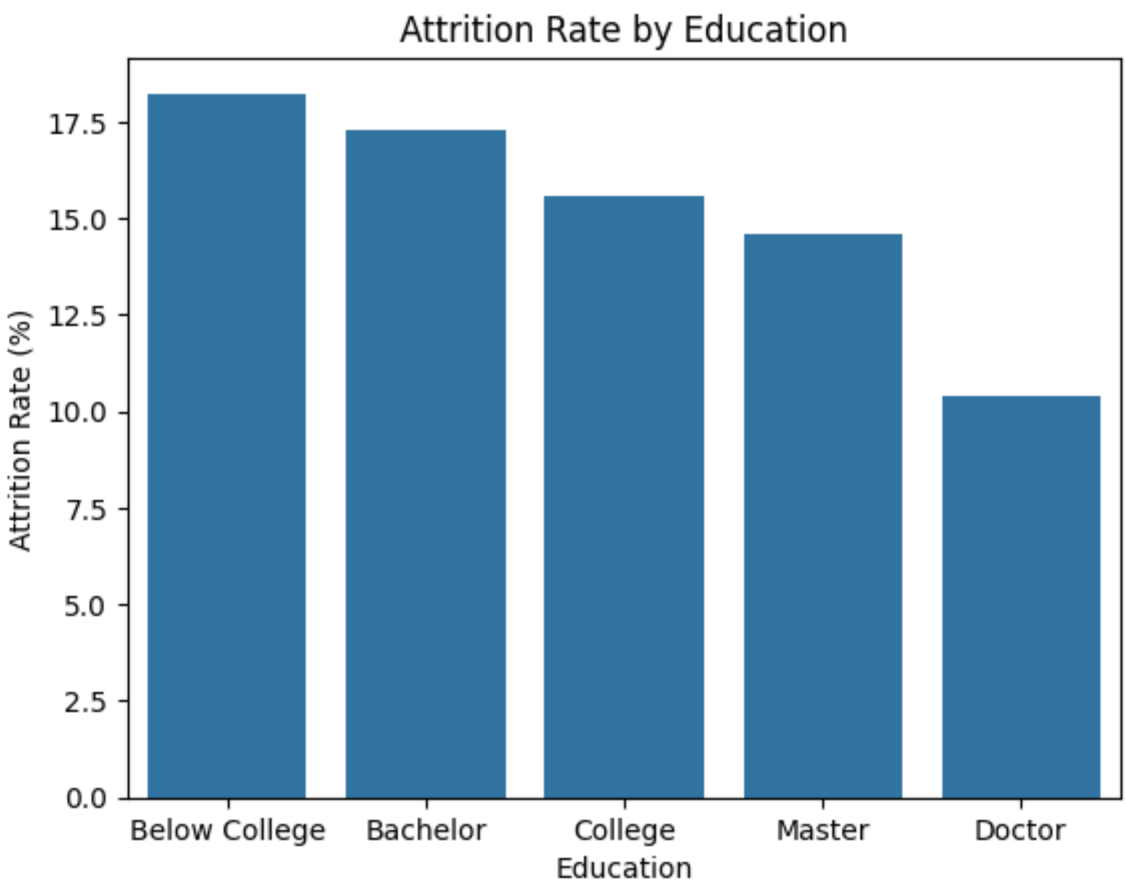
	education_cat
0	College
1	Below College
2	College
3	Master
4	Below College
...	...
1465	College
1466	Below College
1467	Bachelor
1468	Bachelor
1469	Bachelor

1470 rows × 1 columns

dtype: object

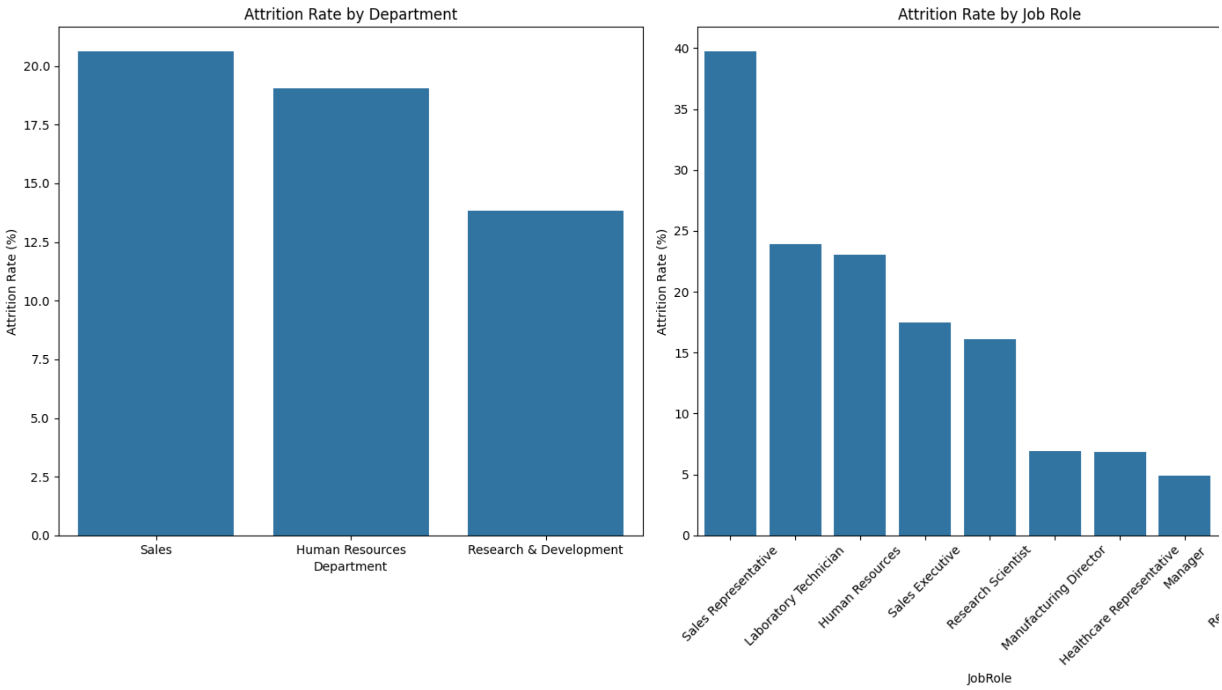
In [ ]:

```
attrition_rate_df = calculate_attrition_rate(df, 'education_cat')
attrition_rate_df = attrition_rate_df.sort_values(by='attritionrate', asc
sns.barplot(data=attrition_rate_df, x='education_cat', y='attritionrate')
plt.title('Attrition Rate by Education')
plt.xlabel('Education')
plt.ylabel('Attrition Rate (%)')

plt.show()
```



Employees with higher levels of education tend to have higher levels of loyalty to the company. T evidenced by the lower turnover rate of employees with master's and doctoral degrees. However analysis is needed to determine whether increasing the level of education tends to increase the l of staying with the company.

In [ ]:
```
fig, axes = plt.subplots(nrows=1, ncols=2, figsize=(15,8))

# Plot 1: KDE plot of Age with Attrition
attrition_rate_by_department = attrition_rate_by_department.sort_values(by
sns.barplot(data=attrition_rate_by_department, x='department', y='attriti
axes[0].set_title('Attrition Rate by Department')
axes[0].set_xlabel('Department')
axes[0].set_ylabel('Attrition Rate (%)')

# Plot 2: Bar plot of Gender count with Attrition
attrition_rate_df = calculate_attrition_rate(df, 'jobrole')
attrition_rate_df = attrition_rate_df.sort_values(by='attritionrate', asc
sns.barplot(data=attrition_rate_df, x='jobrole', y='attritionrate', ax=ax
axes[1].set_title('Attrition Rate by Job Role')
axes[1].set_xlabel('JobRole')
axes[1].set_ylabel('Attrition Rate (%)')
axes[1].tick_params(axis='x', rotation=45) #rotating the x-axis names 45

plt.tight_layout()
plt.show()
```

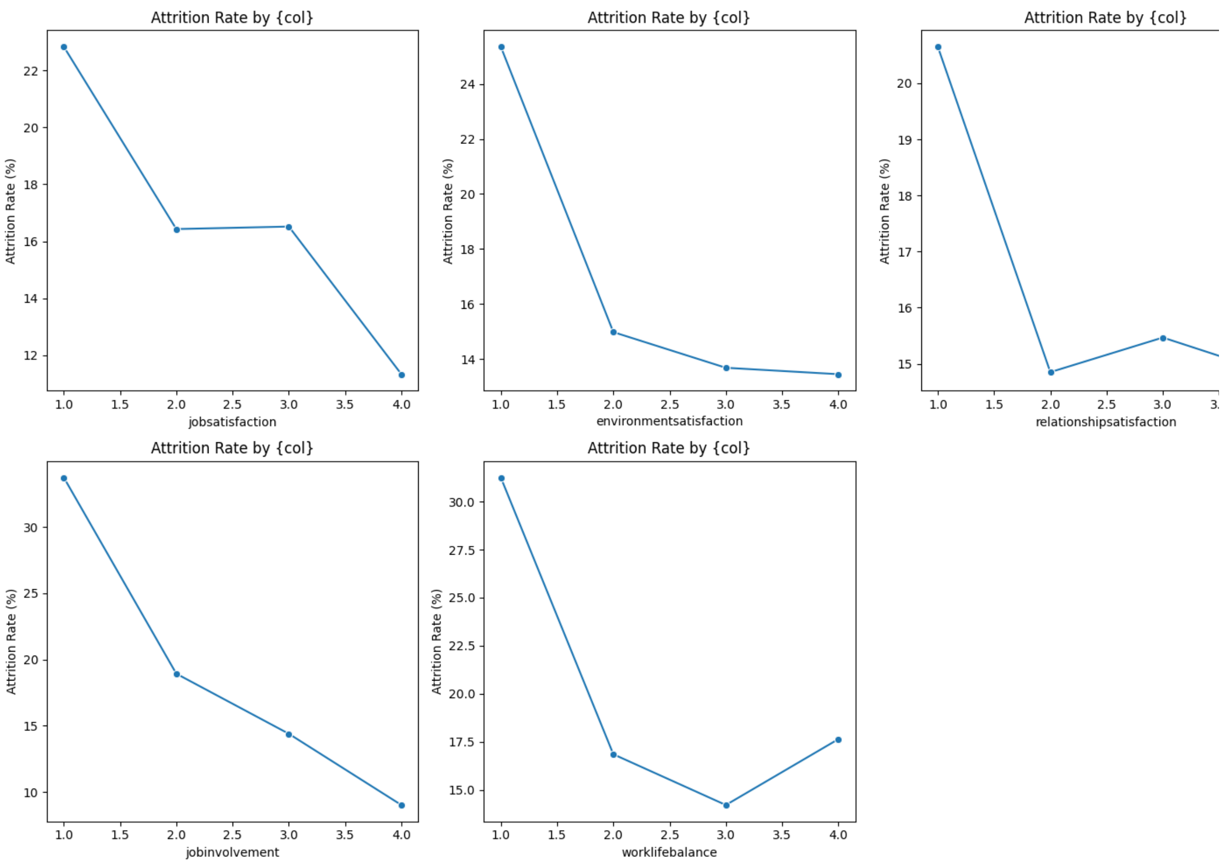


1. The sales department and the positions of sales representative and lab technician have high turnover rates. This may be due to factors such as high work pressure, unattainable sales go lack of job satisfaction.

2. The research and development department and the positions of research scientist and resea director have low turnover rates. This may be due to the challenging nature of the work, grea opportunities for career development, or a more supportive work environment.

Based on the analysis of the above chart, it can be concluded that the turnover rate is influenced department and position held. Employees in the sales department and those holding the position sales representative and laboratory technician tend to leave the company more often than emplo the research and development department and those holding the positions of research scientist a research director.

```
In [ ]: satisfaction_cols = [
            'jobsatisfaction', 'environmentsatisfaction',
            'relationshipsatisfaction', 'jobinvolvement',                    #(
            'worklifebalance'
        ]

        fig, axes = plt.subplots(2, 3, figsize=(15, 10))        #assigning the axe

        axes = axes.flatten()

        for i, col in enumerate(satisfaction_cols):            #to unpack the grou
            attrition_rate_df = calculate_attrition_rate(df, col)
            sns.lineplot(data=attrition_rate_df, x=col, y='attritionrate', marker=
            axes[i].set_title('Attrition Rate by {col}')
            axes[i].set_xlabel(col)
            axes[i].set_ylabel('Attrition Rate (%)')

        if len(satisfaction_cols) % 2 != 0:   #the axes buids 6 graphs 2x3  we on
            fig.delaxes(axes[-1])

        plt.tight_layout()                  #perfectly fit the graphs in their
        plt.show()
```



1. Job Satisfaction: Employees with low levels of job satisfaction tend to leave more often. This suggests that aspects of the job itself, such as tasks, responsibilities, and challenges, strong influence an employee's decision to stay or leave.

2. Environmental Satisfaction: A work environment that is uncomfortable, unsupportive, or inco with an employee's values may encourage them to seek employment elsewhere.

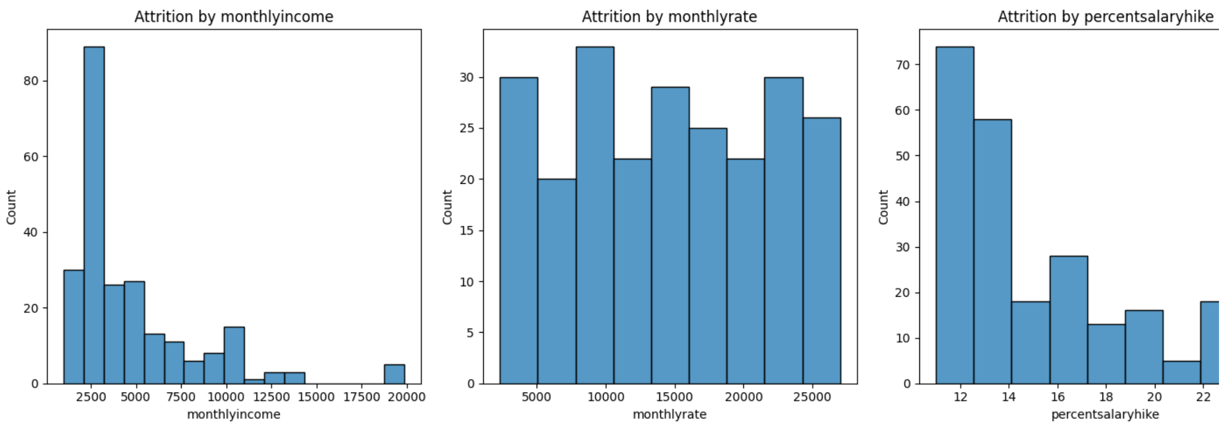
3. Relationship satisfaction: Good relationships with co-workers and supervisors can increase a of belonging and loyalty to the organization, thereby reducing turnover.

4. Job Involvement: Employees who feel engaged in their work tend to be more loyal and comr the organization.

5. Work-life balance: A good work-life balance is very important to employees. Employees who their work interferes with their personal lives are more likely to leave the company

In addition, the results of the analysis show a strong correlation between the level of job involven the level of turnover. Employees with low levels of job involvement tend to leave the organization frequently. This suggests that a lack of job involvement, which may be caused by a lack of caree development opportunities or a lack of challenge in the job, may encourage employees to seek n fulfilling work elsewhere.

In [ ]:
```
salary_col = ['monthlyincome', 'monthlyrate', 'percentsalaryhike']

fig, axes = plt.subplots(nrows=1, ncols=3, figsize=(15,5))

for i, col in enumerate(salary_col):
    sns.histplot(data=df_attrition, x=col, ax=axes[i])
    axes[i].set_title(f'Attrition by {col}')
    axes[i].set_xlabel(col)
    axes[i].set_ylabel('Count')

plt.tight_layout()
plt.show()
```



Turnover by monthly income:

1. This chart shows that most of the employees who left had a monthly income in the range of 7,500.
2. There is a significant decrease in the turnover rate for employees with a monthly income abo 7,500, indicating that employees with higher salaries tend to stay with the company longer.

Turnover by Monthly Rate:

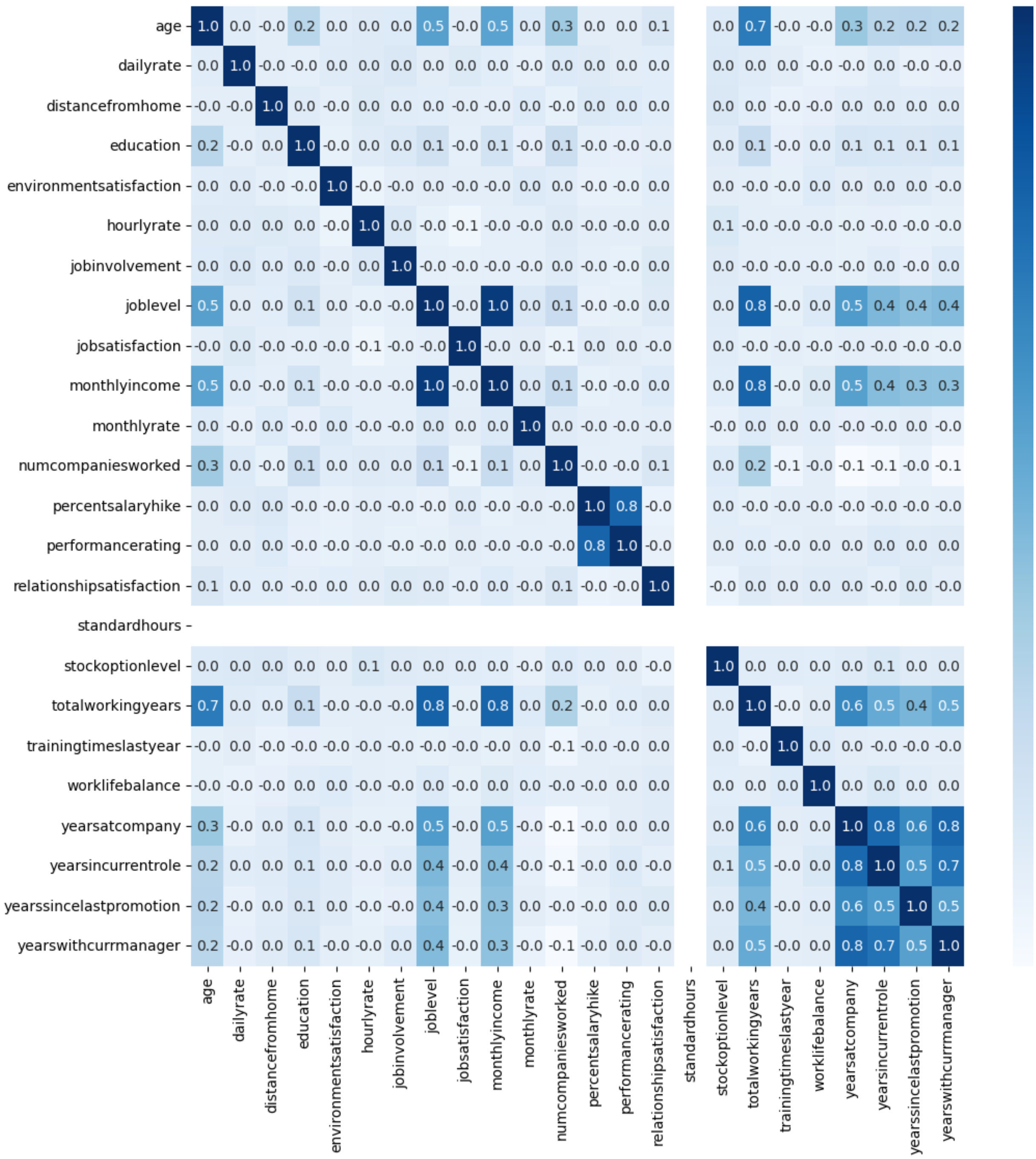
1. The Turnover by Monthly Rate graph does not show a clear pattern between salary levels ar turnover rates. Turnover fluctuates randomly across different salary ranges.

Turnover by Percent Salary Increase:

1. This chart shows that employees who receive lower salary increases (below 16%) tend to ha higher turnover rates.
2. The higher the percentage increase, the lower the turnover rate. This shows that a significar increase can be an effective retention factor.

In [ ]:
```
df_num = df._get_numeric_data()

# drop unnecessary numerical column
columns_to_drop = ['employeenumber', 'employeecount']
df_num = df_num.drop(columns_to_drop, axis=1)
# define the figure
plt.figure(figsize=(12, 12))

# plot correlation heatmap
sns.heatmap(df_num.corr(),
            cmap='Blues',
            annot=True,
            fmt='.1f')
```

Out[ ]: <Axes: >

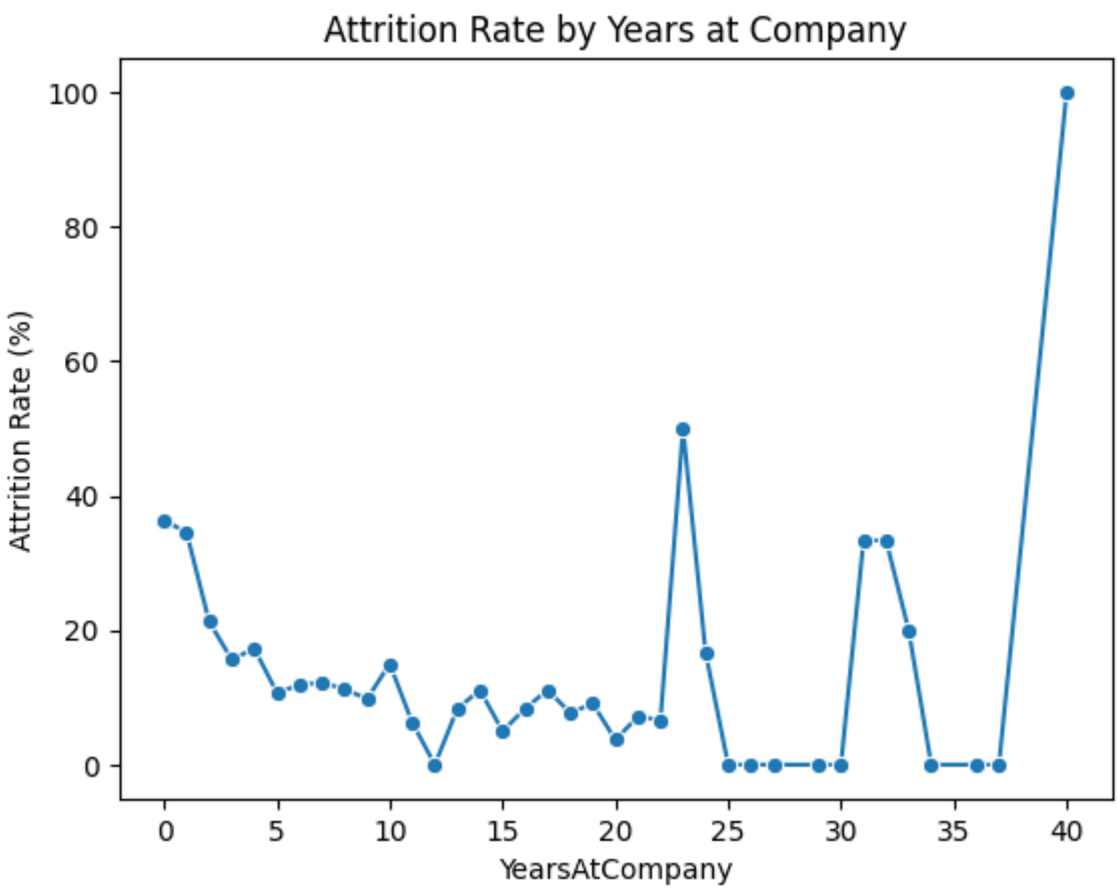


1. **Work Experience:** Variables such as TotalWorkingYears, YearsAtCompany, YearsInCurrent YearsSinceLastPromotion, and YearsWithCurrentManager show strong positive correlations each other. This makes sense because the longer someone works for a company, the longe stay in the same role and with the same manager.

2. **Job Satisfaction:** The JobSatisfaction and EnvironmentSatisfaction variables show a mode positive correlation. This suggests that employees who are satisfied with their job tend to be with their work environmzent.

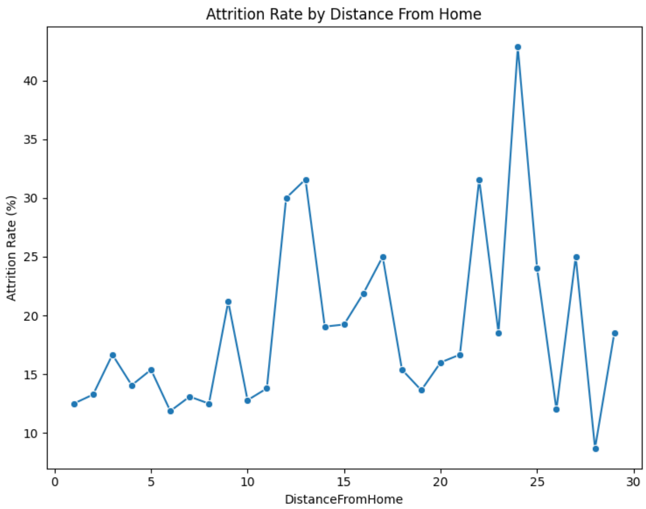
3. **Salary and Satisfaction:**Although there is a positive correlation between MonthlyIncome ar JobSatisfaction, the correlation is not very strong. This shows that salary is not the only facto influencing job satisfaction.

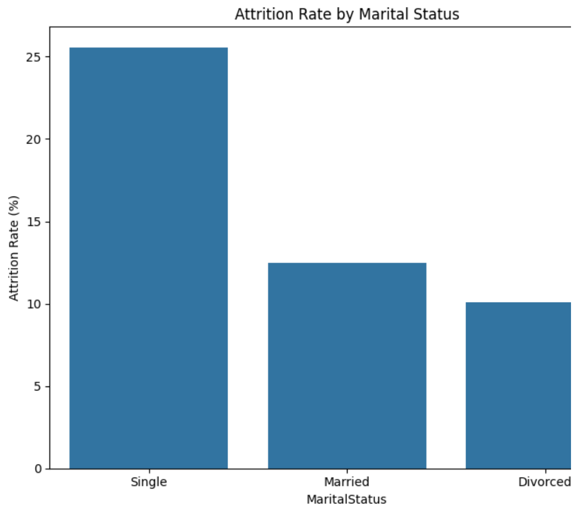
```
In [ ]: attrition_rate_df = calculate_attrition_rate(df, 'yearsatcompany')
        sns.lineplot(data=attrition_rate_df, x='yearsatcompany', y='attritionrate
        plt.title(f'Attrition Rate by Years at Company')
        plt.xlabel('YearsAtCompany')
        plt.ylabel('Attrition Rate (%)')
        plt.show()
```



The results show that new employees have a significantly higher risk of leaving the company cor to those who have been with the company for a longer period. The notably high attrition rate with first year highlights the need for targeted retention efforts for new employees. Additionally, specia attention should be given to employees with around ***20 to 25, 30 to 35 and 37+ *** years of ter these periods also show spikes in attrition rates.

```
In [ ]: ig, axes = plt.subplots(nrows=1, ncols=2, figsize=(15,6))

        attrition_rate_df = calculate_attrition_rate(df, 'distancefromhome')
        sns.lineplot(data=attrition_rate_df, x='distancefromhome', y='attritionra
        axes[0].set_title('Attrition Rate by Distance From Home')
        axes[0].set_xlabel('DistanceFromHome')
        axes[0].set_ylabel('Attrition Rate (%)')

        attrition_rate_df = calculate_attrition_rate(df, 'maritalstatus')
        attrition_rate_df = attrition_rate_df.sort_values(by='attritionrate', asce
        sns.barplot(data=attrition_rate_df, x='maritalstatus', y='attritionrate',
        axes[1].set_title('Attrition Rate by Marital Status')
        axes[1].set_xlabel('MaritalStatus')
        axes[1].set_ylabel('Attrition Rate (%)')

        plt.tight_layout()
        plt.show()
```





Attrition by Distance From Home

1. This graph shows the relationship between the distance between an employee's home and t company and the turnover rate. There is significant variation in the turnover rate over differe distances. Although there is no clear and linear pattern, it can be seen that the turnover rate increase sharply at certain distances. This suggests that distance to work may be one of the influencing an employee's decision to leave the company.

Attrition by Marital Status

1. This chart shows the relationship between employee marital status and turnover rate. Single employees have the highest turnover rate, followed by married employees, and divorced em have the lowest turnover rate. This indicates that marital status may also be a factor influenc employee's decision to stay or leave the company.

The results of this analysis highlight the importance of workplace flexibility and employee support programs tailored to different demographic needs. By understanding the factors t influence turnover rates, organizations can take proactive steps to retain talented employe

```
In [33]:
```