Project Title	IBM HR Analytics Employee Attrition & Performance
Tools	Python, ML, SQL, Excel
Technologies	Data Analyst & Data scientist
Project Difficulties level	intermediate

Dataset : Dataset is available in the given link. You can download it at your convenience.

Click here to download data set

About Dataset

Uncover the factors that lead to employee attrition and explore important questions such as ‘show me a breakdown of distance from home by job role and attrition’ or ‘compare average monthly income by education and attrition’. This is a fictional data set created by IBM data scientists.

Education

1 'Below College'

2 'College'

3 'Bachelor'

4 'Master'

5 'Doctor'

EnvironmentSatisfaction

1 'Low'

2 'Medium'

3 'High'

4 'Very High'

JobInvolvement

1 'Low'

2 'Medium'

3 'High'

4 'Very High'

JobSatisfaction

1 'Low'

2 'Medium'

3 'High'

4 'Very High'

PerformanceRating

1 'Low'

2 'Good'

3 'Excellent'

4 'Outstanding'

RelationshipSatisfaction

1 'Low'

2 'Medium'

3 'High'

4 'Very High'

WorkLifeBalance

1 'Bad'

2 'Good'

3 'Better'

4 'Best'

Example

what steps you should have to follow

Sample code

HR Attrition Analysis¶

In the business world, companies often face the challenge of retaining talented employees. One of the most pressing issues is the increasing rate of employee turnover, commonly known as HR attrition. Turnover can have a significant impact on a company's productivity, stability, and long-term sustainability. High attrition rates can lead to increased recruitment and training costs, disrupt team dynamics, and result in the loss of valuable institutional knowledge. Therefore, understanding the factors contributing to attrition and implementing effective retention strategies is crucial for maintaining a competitive edge and ensuring

Objectives of the Analysis

1. Understand Current Turnover Rates: Gain a comprehensive understanding of the current employee turnover rate and analyze the demographic distribution of attrition by age, gender, education, department,

and job role.
2. Identify Key Factors Influencing Turnover: Examine the main factors contributing to employee attrition, including job satisfaction indicators (job involvement and work-life balance), salary factors (monthly income and salary hikes), and benefit factors (stock option levels), to uncover patterns and correlations that drive higher attrition rates.

Data Cleaning

In [1]:

```
# import data manipulation package
import pandas as pd
import numpy as np

# import data visualization package
import matplotlib.pyplot as plt
import seaborn as sns

# importing the warnings library
import warnings
warnings.filterwarnings('ignore')
```

In [2]:

```
# set pandas options
pd.set_option('display.max_columns', 35)

# load dataset
df =
pd.read_csv('/kaggle/input/ibm-hr-analytics-attrition-dataset/WA_Fn-UseC_-HR-Employe
e-Attrition.csv')
df.head()
```

Out[2]:

	Age	Attrition	BusinessTr	DailyRate	Departmen	DistanceFro	Education	EducationF	EmployeeCo	EmployeeNu	EnvironmentSati	Gender	HourlyRat	JobInvolve	JobLevel	JobRole	JobSatisfa	MaritalSta	MonthlyInc	MonthlyRat	NumCompanies	Over18	OverTime	PercentSal	Performance	RelationshipSatisf	StandardHo	StockOption	TotalWorki	TrainingTimesL	WorkLifeBa	YearsAtCom

			avel		t	mHome		ield	unt	mber	sfaction		e	ment			ction	tus	ome	e	Worked			aryHike	Rating	action	urs	Level	ngYears	astYear	lance	pany	
0	41	Yes	Travel_Rarely	1102	Sales	1	2	LifeSciences	1	1	2	Female	94	3	2	SalesExecutive	4	Single	5993	19479	8	Y	Yes	11	3	1	80	0	8	0	1	6
1	49	No	Travel_Frequently	279	Research&Development	8	1	LifeSciences	1	2	3	Male	61	2	2	ResearchScientist	2	Married	5130	24907	1	Y	No	23	4	4	80	1	10	3	3	10
2	37	Ye	Tra	137	Res	2	2	Oth	1	4	4	Mal	92	2	1	Lab	3	Sin	209	239	6	Y	Ye	15	3	2	80	0	7	3	3	0

		s	vel_Rarely	3	earch&Development			er				e				oratoryTechnician		gle	0	6			s									
3	33	No	Travel_Frequently	1392	Research&Development	3	4	LifeSciences	1	5	4	Female	56	3	1	ResearchScientist	3	Married	2909	231559	1	Y	Yes	11	3	3	80	0	8	3	3	8
4	27	No	Travel_Rare	591	Research&De	2	1	Medical	1	7	1	Male	40	3	1	LaboratoryT	2	Married	3468	166632	9	Y	No	12	3	4	80	1	6	3	3	2

			ly		velopment										echnician																

In [3]:

```
# check data shape
df.shape
```

Out[3]:

```
(1470, 35)
```

The output indicates that the DataFrame has 1,470 rows and 35 columns.

In [4]:

```
# check number of dupliacted data
print(f'Number of duplicated data: {df.duplicated().sum()}')
```

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

In [5]:

```
# check missing values
df.isnull().sum() / len(df) * 100
```

Out[5]:

```
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

dtype: float64
```

There are any missing values in the dataset.

In [6]:

```
# check data types
df.dtypes
```

Out[6]:

```
Age                         int64
Attrition                  object
BusinessTravel             object
DailyRate                   int64
Department                 object
```

```
DistanceFromHome             int64
Education                    int64
EducationField              object
EmployeeCount                int64
EmployeeNumber               int64
EnvironmentSatisfaction      int64
Gender                      object
HourlyRate                   int64
JobInvolvement               int64
JobLevel                     int64
JobRole                     object
JobSatisfaction              int64
MaritalStatus               object
MonthlyIncome                int64
MonthlyRate                  int64
NumCompaniesWorked           int64
Over18                      object
OverTime                    object
PercentSalaryHike            int64
PerformanceRating            int64
RelationshipSatisfaction     int64
StandardHours                int64
StockOptionLevel             int64
TotalWorkingYears            int64
TrainingTimesLastYear        int64
WorkLifeBalance              int64
YearsAtCompany               int64
YearsInCurrentRole           int64
YearsSinceLastPromotion      int64
YearsWithCurrManager         int64

dtype: object
```

All columns have appropriate data types, ensuring that the data is correctly formatted for analysis.

In [7]:

```
# check data decribe
df.describe()
```

Out[7]:

	Ag	Dai	Dista	Edu	Emp	Empl	Envir on	Hou	Job	Job	Job	Mon	Mon	Num Co	Perc	Perfo	Relati on	Sta	Sto	Total	Train ing	Wor	Yea	Year	Years Sin	Years Wit

	e	lyRate	nceFromHome	cation	loyeeCount	oyeeNumber	mentSatisfaction	rlyRate	lInvolvement	Level	Satisfaction	thlyIncome	thlyRate	mpaniesWorked	entSalaryHike	rmanceRating	shipSatisfaction	ndardHours	ckOptionLevel	WorkingYears	TimesLastYear	kLifeBalance	rsAtCompany	slnCurrentRole	ceLastPromotion	hCurrManager
count	1470.000000	1470.000000	1470.000000	1470.000000	1470.0	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.0	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000
mean	36.923810	802.485714	9.192517	2.912925	1.0	1024.865306	2.721769	65.891156	2.729932	2.063946	2.728571	6502.931293	14313.103401	2.693197	15.209524	3.153741	2.712245	80.0	0.793878	11.279592	2.799320	2.761224	7.008163	4.229252	2.187755	4.123129
std	9.135373	403.509010	8.106864	1.024165	0.0	602.024335	1.093082	20.329428	0.711561	1.106940	1.102846	4707.956678	7117.786044	2.498009	3.659938	0.360824	1.081209	0.0	0.852077	7.780782	1.289271	0.706476	6.126525	3.623137	3.222430	3.568136

min	18.000000	1022.000000	1.000000	1.000000	1.0	1.000000	1.000000	30.000000	1.000000	1.000000	1.000000	3	4											
												1009.000000	2094.000000	0.000000	11.000000	3.000000	1.000000	80.0	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000
25%	30.000000	465.000000	2.000000	2.000000	1.0	491.250000	2.000000	48.000000	2.000000	1.000000	2.000000	2911.000000	80447.000000	1.000000	12.000000	3.000000	2.000000	80.0	0.000000	6.000000	2.000000	2.000000	3.000000	2.000000	0.000000	2.000000
50%	36.000000	802.000000	7.000000	3.000000	1.0	1020.500000	3.000000	66.000000	3.000000	2.000000	3.000000	49199.000000	142235.500000	2.000000	14.000000	3.000000	3.000000	80.0	1.000000	10.000000	3.000000	3.000000	5.000000	3.000000	1.000000	3.000000
75%	43.000	1157.	14.0000	4.000	1.0	1555.7	4.000000	83.75	3.000	3.000	4.000	8379.	204461	4.000000	18.0000	3.0000	4.000000	80.	1.0000	15.0000	3.000000	3.0000	9.0000	7.0000	3.000000	7.000000

	0000	0000000	000	000		500000	0	0000	000	000	000	0000000	.500000000	0	000	00	0	0	000	000	0	000	000	00	0	0
max	60.000000	1499.000000000	29.000000000	5.000000000	1.0	2068.000000000	4.0000000	100.000000000	4.000000000	5.000000000	4.000000000	1999999.000000000	2699999.000000000	9.0000000	25.000000000	4.0000000	4.0000000	80.0	3.000000000	40.000000000	6.0000000	4.000000000	40.000000000	18.000000000	15.0000000	17.0000000

Based on this summary, there are no apparent outliers in the dataset, as the values fall within expected ranges.

Exploratory Data Analysis

In [8]:

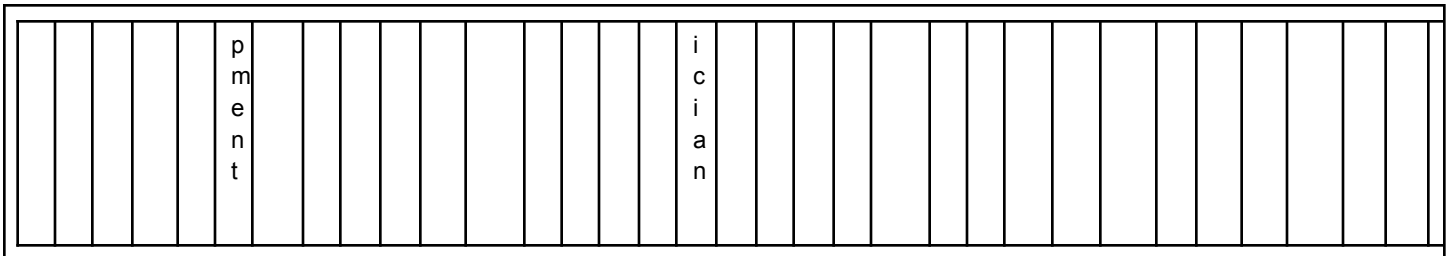
```
df.head()
```

Out[8]:

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHom	Education	EducationFiel	EmployeeCoun	EmployeeNumbe	EnvironmentSatisfactio	Gender	HourlyRate	JobInvolvemen	JobLevel	JobRole	JobSatisfactio	MaritalStatu	MonthlyIncom	MonthlyRate	NumCompaniesWorke	Over18	OverTime	PercentSalaryH	PerformanceRatin	RelationshipSatisfactio	StandardHour	StockOptionLeve	TotalWorkingYe	TrainingTimesLastYe	WorkLifeBalanc	YearsAtCompan

			l			e		d	t	r	n			t			n	s	e		d			ike	g	n	s	l	ars	ar	e	y	
0	41	Yes	Travel_Rarely	1102	Sales	1	2	LifeSciences	1	1	2	Female	94	3	2	SalesExecutive	4	Single	5993	19479	8	Y	Yes	11	3	1	80	0	8	0	1	6	
1	49	No	Travel_Frequently	279	Research&Development	8	1	LifeSciences	1	2	3	Male	61	2	2	ResearchScientist	2	Married	5130	24907	1	Y	No	23	4	4	80	1	10	3	3	10	
2	37	Yes	Travel_	1373	Researc	2	2	Other	1	4	4	Male	92	2	1	Laborat	3	Single	2090	23966	6	Y	Yes	15	3	2	80	0	7	3	3	0	

			Rarely		h&Development											oryTechnician														
3	33	No	Travel_Frequently	1392	Research&Development	3	4	LifeSciences	1	5	4	Female	56	3	1	ResearchScientist	3	Married	2909	231559	1	Y	Yes	11	3	3	80	0	8	3	3	8
4	27	No	Travel_Rarely	591	Research&Develo	2	1	Medical	1	7	1	Male	40	3	1	LaboratoryTechn	2	Married	3468	166632	9	Y	No	12	3	4	80	1	6	3	3	2



Attrition Rate

Attrition rate: The attrition rate measures the percentage of employees who leave the company in a given period of time. It is usually calculated within a year and is expressed as a percentage of the total number of employees.

In [9]:

```
df['Attrition'].value_counts(normalize=True)
```

Out[9]:

```
Attrition
No     0.838776
Yes    0.161224

Name: proportion, dtype: float64
```

The output displays the proportion of employees with regard to attrition status in the dataset. Let's visualize it!

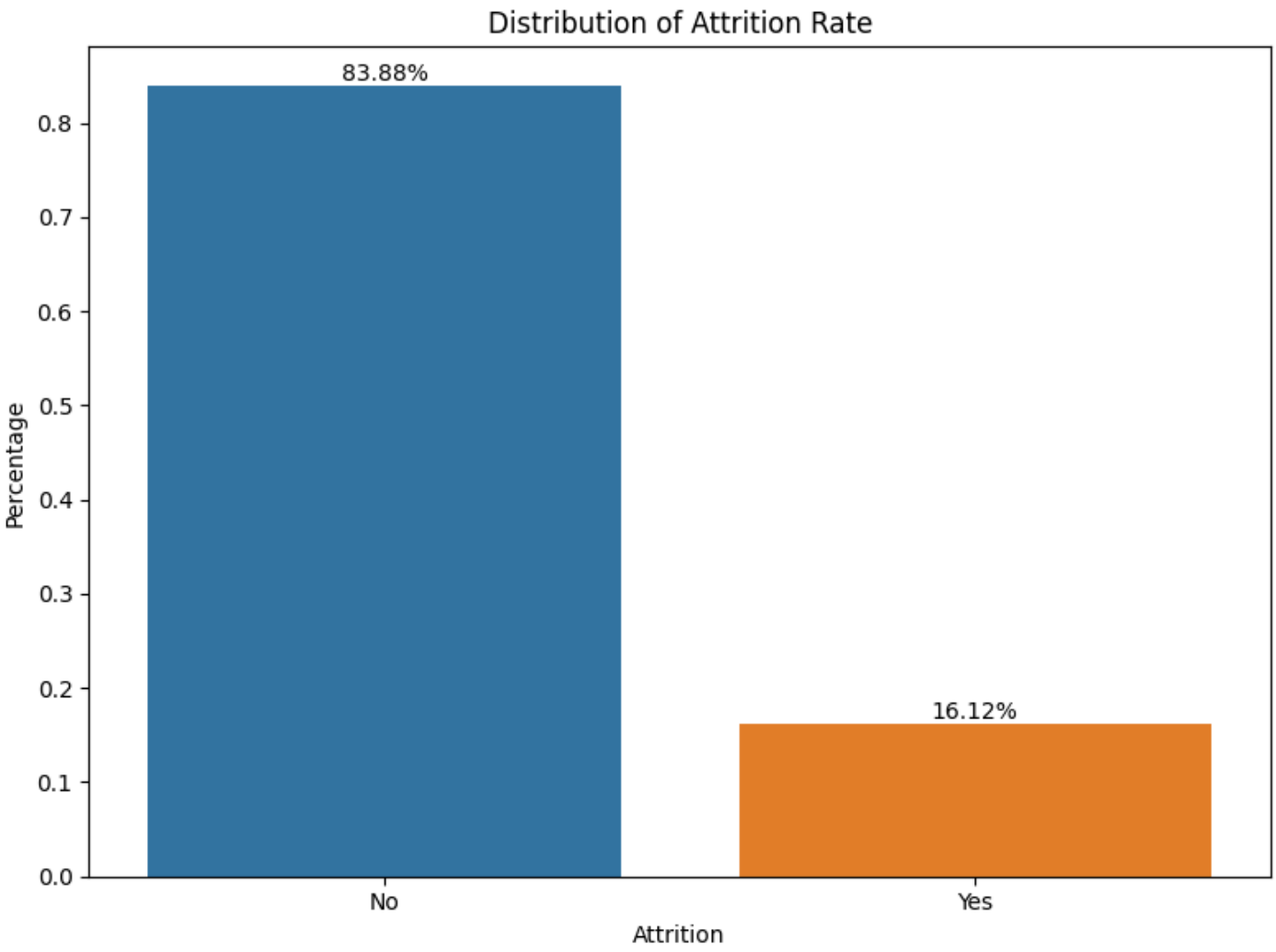
In [10]:

```
attrition = df['Attrition'].value_counts(normalize=True)

plt.figure(figsize=(8,6))
ax = sns.barplot(x=attrition.index, y=attrition)

for p in ax.patches:
    ax.annotate(f'{p.get_height() * 100:.2f}%',
                (p.get_x() + p.get_width() / 2., p.get_height()),
                ha='center', va='bottom')

plt.title('Distribution of Attrition Rate')
plt.xlabel('Attrition')
plt.ylabel('Percentage')
plt.tight_layout()
plt.show()
```



Based on the analysis, the company's attrition rate is 16.12%. This means that about 16.12% of the employees decided to leave the company during the analyzed period.

Average of Tenure

Average tenure: The average tenure measures the average number of years an employee stays with the company before leaving. It can provide insight into workforce stability and employee satisfaction within the organization.

In [11]:

```
avg_tenure = df['YearsAtCompany'].mean()
print(f'Average years of employee to leave the company is {avg_tenure} years')
```

```
Average years of employee to leave the company is 7.0081632653061225 years
```

The average tenure of employees before they decided to leave was 7.01 years. With this average tenure, it can be concluded that many employees feel comfortable and have been with the company for a long time.

Employee's Demographics

In [12]:

```
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 company has

many employees who are at a productive and experienced age.
2. Gender: The majority of employees at this company are male. There are significantly more male employees than female employees.
3. Department: Most of the company's employees are concentrated in the research and development department. This indicates that the company is heavily focused on product or service research and development activities.

In [13]:

```
df_attrition = df[df['Attrition'] == 'Yes']
df_attrition.head()
```

Out[13]:

	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
0	41	Yes	Travel_Rarely	1102	Sales	1	2	LifeSciences	1	1	2	Female	94	3	2	SalesExecutive	4	Single	5993	19479	8	Y	Yes	11	3	1	80	0	8	0	1	6

2	37	Yes	Travel_Rarely	1373	Research&Development	2	2	Other	1	4	4	Male	92	2	1	LaboratoryTechnician	3	Single	2090	23962	6	Y	Yes	15	3	2	80	0	7	3	3	0	0
14	28	Yes	Travel_Rarely	103	Research&Development	24	3	LifeSciences	1	19	3	Male	50	2	1	LaboratoryTechnician	3	Single	2028	129447	5	Y	Yes	14	3	2	80	0	6	4	3	4	1
21	36	Yes	Travel	1218	Sales	9	4	LifeSc	1	27	3	Male	82	2	1	SalesR	1	Singl	3407	69866	7	Y	No	23	4	2	80	0	10	4	3	5	1

			_Rarely					iences								epresentative		e																			
24	34	Yes	Travel_Rarely	699	Research&Development	6	1	Medical	1	31	2	Male	83	3	1	ResearchScientist	1	Single	2960	17102	2	Y	No	11	3	3	80	0	8	2	3	4					

Demografic Factors

In [14]:

```
# Fungsi untuk Menghitung Attrition Rate
def calculate_attrition_rate(df, column):
    attrition_counts = df.groupby([column,
'Attrition']).size().unstack(fill_value=0)
    attrition_rate = attrition_counts['Yes'] / attrition_counts.sum(axis=1) * 100
    attrition_rate_df = attrition_rate.reset_index()
    attrition_rate_df.columns = [column, 'AttritionRate']
    return attrition_rate_df
```

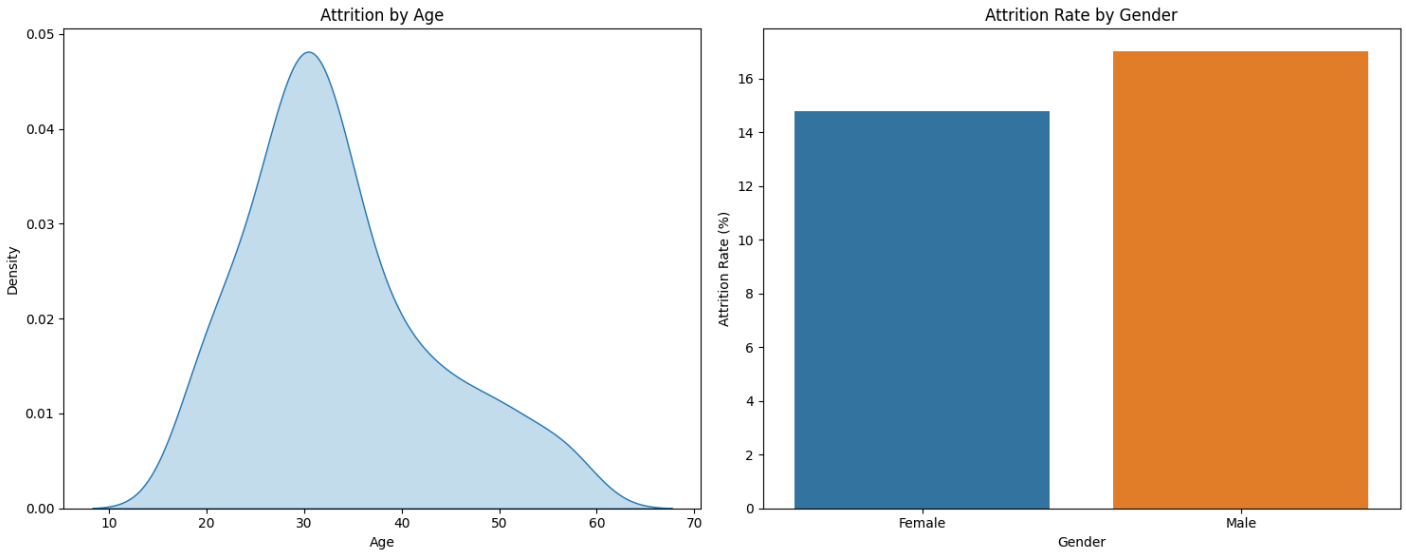
In [15]:

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

# Plot 1: KDE plot of Age with Attrition hue
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 hue
attrition_rate_df = calculate_attrition_rate(df, 'Gender')
sns.barplot(data=attrition_rate_df, x='Gender', y='AttritionRate', ax=axes[1])
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 Reference link