PROJECT FILE

ON ¶

Introduction

In the dynamic landscape of modern workplaces, understanding the factors contributing to employee attrition has become increasingly critical for organizations seeking to foster a stable and productive workforce. This dataset, aptly titled "Exploring Employee Attrition," provides a rich repository of information encompassing various facets of employee profiles and work-related attributes. From demographic details to job-related metrics, this dataset offers a holistic view of employees within a given organization.

The dataset encompasses a diverse range of metadata, including information about employees' age, business travel patterns, educational background, job roles, marital status, and more. The focal point of analysis lies in the "Attrition" and "CF_attrition_label" columns, shedding light on the occurrence of attrition and the corresponding labels assigned to employees. These labels, in particular, play a pivotal role in understanding the nature and context of attrition events within the dataset.

Past segment and individual data, the dataset incorporates business related factors, for example, preparing recurrence, work fulfillment levels, execution evaluations, and the quantity of years spent in different jobs. These boundaries offer important bits of knowledge into the elements affecting representative commitment, work fulfillment, and by and large maintenance inside the hierarchical system.

In [1]:
```
#LIBRARY USED IN THIS PROJECT
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
```

In [2]:
```
#Data Set
df = pd.read_excel('HR_DATA_1.xlsx')
df
```

Out[2]:

	Attrition	Business_Travel	CF_age_band	CF_attrition_label	Department	Education_Field
0	Yes	Travel_Rarely	35 - 44	Ex-Employees	Sales	Life Sciences
1	No	Travel_Frequently	45 - 54	Current Employees	R&D	Life Sciences
2	Yes	Travel_Rarely	35 - 44	Ex-Employees	R&D	Other
3	No	Travel_Frequently	25 - 34	Current Employees	R&D	Life Sciences
4	No	Travel_Rarely	25 - 34	Current Employees	R&D	Medical
...	...	...	...	...	...	...
1465	Yes	Non-Travel	25 - 34	Ex-Employees	R&D	Technical Degree
1466	Yes	Travel_Frequently	25 - 34	Ex-Employees	R&D	Life Sciences
1467	Yes	Travel_Frequently	35 - 44	Ex-Employees	Sales	Other
1468	Yes	Travel_Rarely	Under 25	Ex-Employees	R&D	Life Sciences
1469	Yes	Travel_Rarely	Under 25	Ex-Employees	Sales	Life Sciences

1470 rows × 36 columns

Before getting into data analytics let us try to find is there any null values or missing data in this data set and find the summary statistics of the data to get an idea about the dataset.

In [3]:
```
df.isnull().sum()
```

Out[3]:
```
Attrition                     0
Business_Travel               0
CF_age_band                   0
CF_attrition_label            0
Department                    0
Education_Field               0
emp_no                        0
Employee_Number               0
Gender                        0
Job_Role                      0
Marital_Status                0
Over_Time                     0
Training_Times_Last_Year      0
Age                           0
CF_current_Employee           0
Daily_Rate                    0
Distance_From_Home            0
Education                     0
Employee_Count                0
Environment_Satisfaction      0
Hourly_Rate                   0
Job_Involvement               0
Job_Level                     0
Job_Satisfaction              0
Monthly_Income                0
Num_Companies_Worked          0
Percent_Salary_Hike           0
Performance_Rating            0
Relationship_Satisfaction     0
Standard_Hours                0
Total_Working_Years           0
Work_Life_Balance             0
Years_At_Company              0
Years_In_Current_Role         0
Years_Since_Last_Promotion    0
Years_With_Curr_Manager       0
dtype: int64
```

In [4]:
```
print(df.describe()) #Summary Statistics

df['Job_Satisfaction'].describe()
```

```
       Employee_Number  Training_Times_Last_Year          Age  \
count      1470.000000               1470.000000  1470.000000
mean       1024.865306                  2.799320    36.923810
std         602.024335                  1.289271     9.135373
min           1.000000                  0.000000    18.000000
25%         491.250000                  2.000000    30.000000
50%        1020.500000                  3.000000    36.000000
75%        1555.750000                  3.000000    43.000000
max        2068.000000                  6.000000    60.000000

       CF_current_Employee   Daily_Rate  Distance_From_Home  Employee_Count
\
count          1470.000000  1470.000000         1470.000000          1470.0
mean              0.838776   802.485714            9.192517             1.0
std               0.367863   403.509100            8.106864             0.0
min               0.000000   102.000000            1.000000             1.0
25%               1.000000   465.000000            2.000000             1.0
50%               1.000000   802.000000            7.000000             1.0
75%               1.000000  1157.000000           14.000000             1.0
max               1.000000  1499.000000           29.000000             1.0

       Environment_Satisfaction  Hourly_Rate  Job_Involvement  ...  \
count               1470.000000  1470.000000      1470.000000  ...
mean                   2.721769    65.891156         2.729932  ...
std                    1.093082    20.329428         0.711561  ...
min                    1.000000    30.000000         1.000000  ...
25%                    2.000000    48.000000         2.000000  ...
50%                    3.000000    66.000000         3.000000  ...
75%                    4.000000    83.750000         3.000000  ...
max                    4.000000   100.000000         4.000000  ...

       Percent_Salary_Hike  Performance_Rating  Relationship_Satisfaction  \
count          1470.000000         1470.000000                1470.000000
mean             15.209524            3.153741                   2.712245
std               3.659938            0.360824                   1.081209
min              11.000000            3.000000                   1.000000
25%              12.000000            3.000000                   2.000000
50%              14.000000            3.000000                   3.000000
75%              18.000000            3.000000                   4.000000
max              25.000000            4.000000                   4.000000

       Standard_Hours  Total_Working_Years  Work_Life_Balance  \
count          1470.0          1470.000000        1470.000000
mean             80.0            11.279592           2.761224
std               0.0             7.780782           0.706476
min              80.0             0.000000           1.000000
25%              80.0             6.000000           2.000000
50%              80.0            10.000000           3.000000
75%              80.0            15.000000           3.000000
max              80.0            40.000000           4.000000

       Years_At_Company  Years_In_Current_Role  Years_Since_Last_Promotion
\
count       1470.000000            1470.000000                 1470.000000
mean           7.008163               4.229252                    2.187755
std            6.126525               3.623137                    3.222430
min            0.000000               0.000000                    0.000000
```

```
25%            3.000000             2.000000             0.000000
50%            5.000000             3.000000             1.000000
75%            9.000000             7.000000             3.000000
max           40.000000            18.000000            15.000000

       Years_With_Curr_Manager
count              1470.000000
mean                  4.123129
std                   3.568136
min                   0.000000
25%                   2.000000
50%                   3.000000
75%                   7.000000
max                  17.000000

[8 rows x 24 columns]
```

Out[4]:
```
count    1470.000000
mean        2.728571
std         1.102846
min         1.000000
25%         2.000000
50%         3.000000
75%         4.000000
max         4.000000
Name: Job_Satisfaction, dtype: float64
```

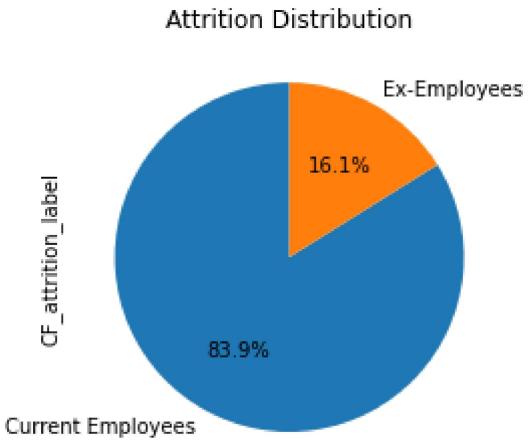
So now as we can see the data is clean let's analyse the data and find some useful insights.

FINDINGS

1. Attrition Distribution

In [5]:
```
attrition_rate = df['CF_attrition_label'].value_counts(normalize=True) * 100
print(attrition_rate)

import matplotlib.pyplot as plt

attrition_counts = df['CF_attrition_label'].value_counts()
attrition_counts.plot(kind='pie', autopct='%1.1f%%', startangle=90)
plt.title('Attrition Distribution')
plt.show()
```

```
Current Employees    83.877551
Ex-Employees         16.122449
Name: CF_attrition_label, dtype: float64
```

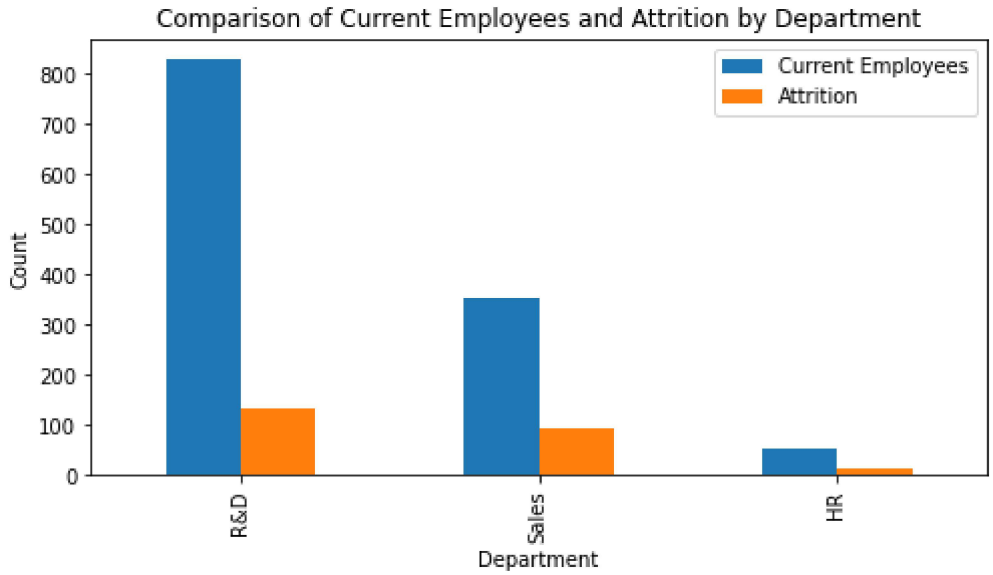


Interpretation: We can see that 16.1% of the total employees have left the job and the company could manage to retain only 83.9% of its total employees.

2. Department wise number of total employees and the attrition

In [6]:
```
current_employees = df[df['CF_current_Employee'] == 1]

attrition_yes = df[df['Attrition'] == 'Yes']
current_employee_counts = current_employees['Department'].value_counts()
attrition_counts = attrition_yes['Department'].value_counts()
comparison_df = pd.DataFrame({'Current Employees': current_employee_counts, 'A
})

print(comparison_df)
comparison_df.plot(kind='bar', figsize=(8, 4))
plt.title('Comparison of Current Employees and Attrition by Department')
plt.xlabel('Department')
plt.ylabel('Count')
plt.show()
```

```
       Current Employees  Attrition
R&D                  828        133
Sales                354         92
HR                    51         12
```

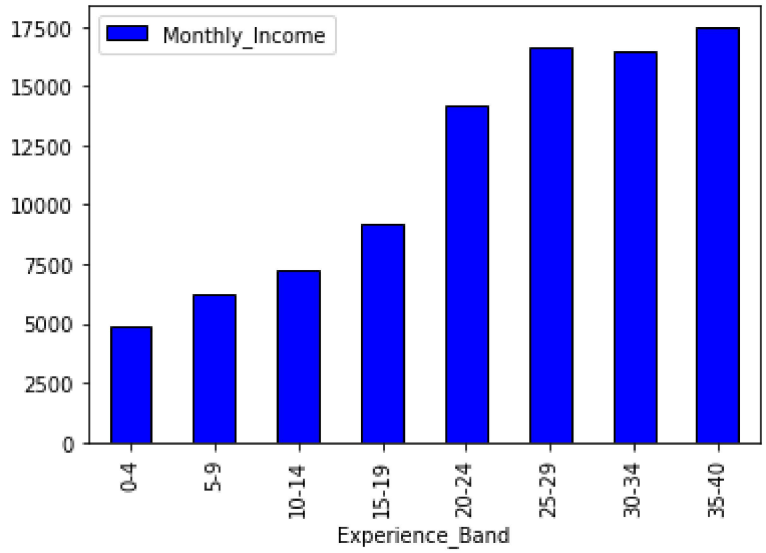


Interpretation: The R&D department have the highest number of employees and yet have the lowest attrition rate of 16.06%. The highest attrition rate has been shown in the Sales department of 25.98%.

3. Wokring experience and salary relationship

In [7]:
```
bins = [0, 5, 10, 15, 20, 25, 30, 35, 40]

labels = ['0-4', '5-9', '10-14', '15-19', '20-24', '25-29', '30-34', '35-40']
df['Experience_Band'] = pd.cut(df['Years_At_Company'], bins=bins, labels=label
result = pd.DataFrame(df.groupby('Experience_Band')['Monthly_Income'].mean())
result

result.plot(kind='bar', color='blue', edgecolor='black')
```

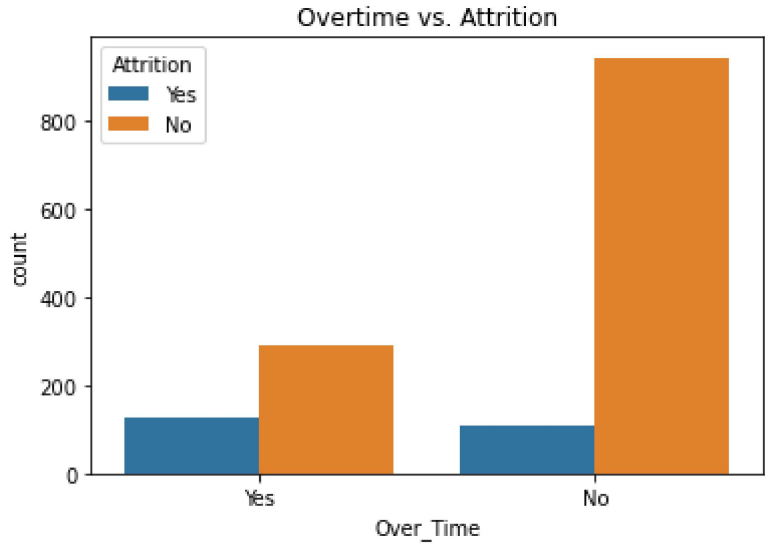
Out[7]:
```
<AxesSubplot:xlabel='Experience_Band'>
```



Interpretation: With the increase in the experience band we can see a significant increase in the monthly salary of the employees

4. Overtime and Attrition

In [8]:
```
sns.countplot(x='Over_Time', hue='Attrition', data=df)
plt.title('Overtime vs. Attrition')
plt.show()

import pandas as pd
overtime_table = pd.crosstab(df['Attrition'], df['Over_Time'])
overtime_table.columns = ['Attrition', 'Over_Time']
overtime_table['Ratio'] = overtime_table['Over_Time'] / overtime_table['Attrit
print(overtime_table)
```

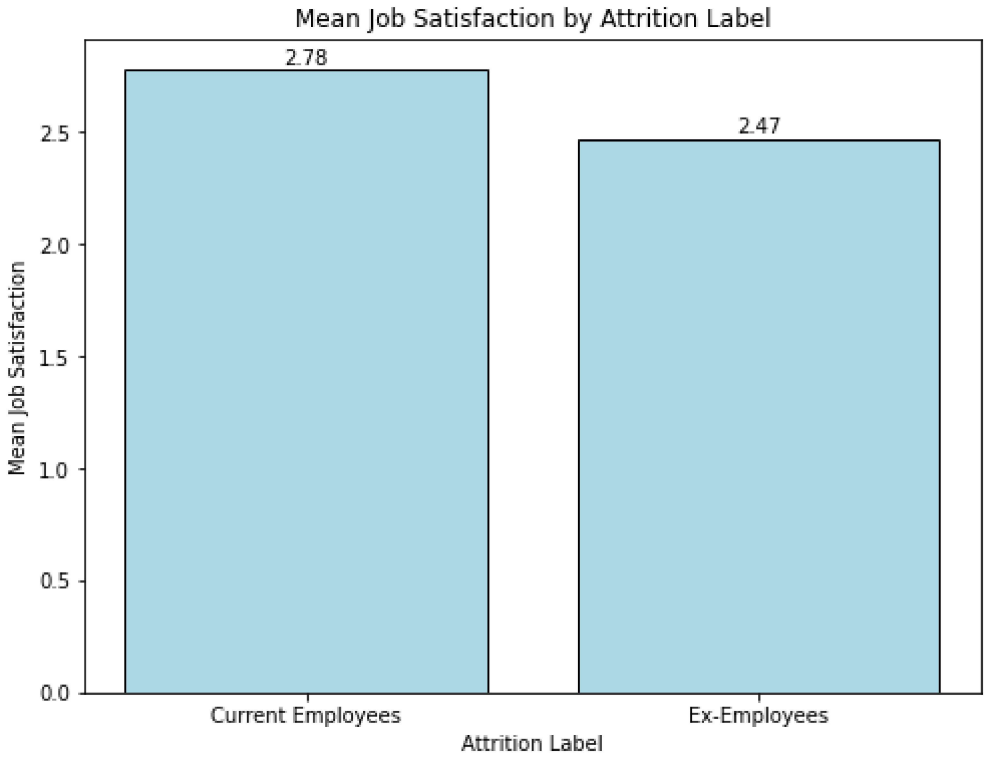



```
           Attrition  Over_Time     Ratio
Attrition
No               944        289  0.306144
Yes              110        127  1.154545
```

Interpretation: This plot can help assess whether employees who work overtime are more likely to experience attrition compared to those who do not.

5. Average job satisfaction of current employees and ex-employees

In [9]:
```
attrition_label_job_satisfaction_mean = df.groupby('CF_attrition_label')['Job_
plt.figure(figsize=(8, 6))
bars = plt.bar(attrition_label_job_satisfaction_mean['CF_attrition_label'],
               attrition_label_job_satisfaction_mean['Job_Satisfaction'],
               color='lightblue', edgecolor='black')
for bar in bars:
    yval = bar.get_height()
    plt.text(bar.get_x() + bar.get_width() / 2, yval + 0.01, round(yval, 2), h

plt.title('Mean Job Satisfaction by Attrition Label')
plt.xlabel('Attrition Label')
plt.ylabel('Mean Job Satisfaction')
plt.xticks(rotation=0)
plt.show()
```



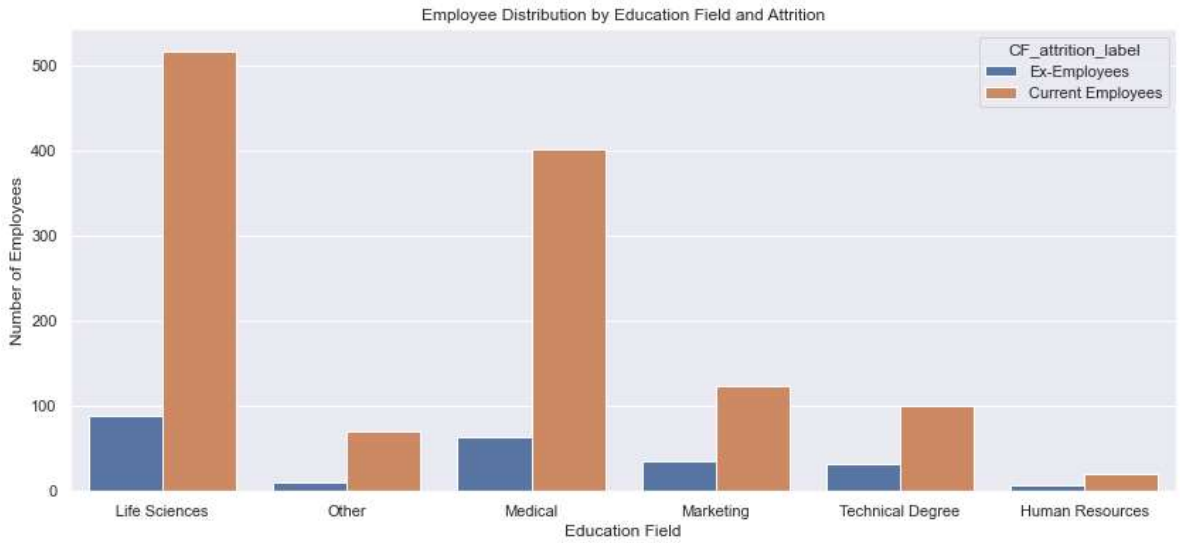
Interpretation: We can see that the average job satisfaction score of the ex-emplyees is relatively low than the current employees. The company should look into this matter and try to improve the employee satisfaction score to retain thier employees.

6.1 Attrition with respect to the education field of the employees

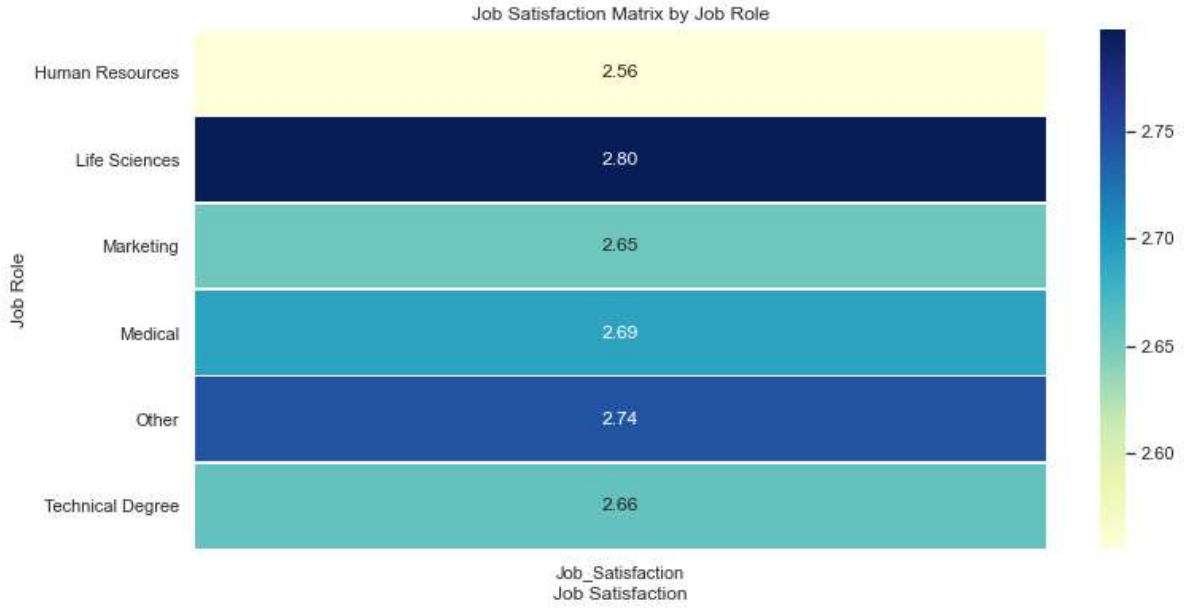
6.2 Table for number of current and ex-employees and the attrition ratio

6.3 Job Satisfaction Matrix by Job Role

In [10]:
```
import seaborn as sns
import matplotlib.pyplot as plt

sns.set(rc={'figure.figsize':(14, 6)})

sns.countplot(data=df, x='Education_Field', hue='CF_attrition_label')
plt.title('Employee Distribution by Education Field and Attrition')
plt.xlabel('Education Field')
plt.ylabel('Number of Employees')
plt.show()

import pandas as pd
education_attrition_table = pd.crosstab(df['Education_Field'], df['CF_attritic
education_attrition_table.columns = ['Current Employees', 'Ex-Employees']
education_attrition_table['Ex-Employee Ratio'] = education_attrition_table['Ex
print(education_attrition_table)

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


job_satisfaction_matrix = pd.pivot_table(df, values='Job_Satisfaction', index=

sns.set(rc={'figure.figsize': (12, 6)})


sns.heatmap(job_satisfaction_matrix, cmap='YlGnBu', annot=True, fmt=".2f", lir
plt.title('Job Satisfaction Matrix by Job Role')
plt.xlabel('Job Satisfaction')
plt.ylabel('Job Role')
plt.show()
```



Education_Field	Current Employees	Ex-Employees	Ex-Employee Ratio
Human Resources	20	7	0.350000
Life Sciences	517	89	0.172147
Marketing	124	35	0.282258
Medical	401	63	0.157107
Other	71	11	0.154930
Technical Degree	100	32	0.320000



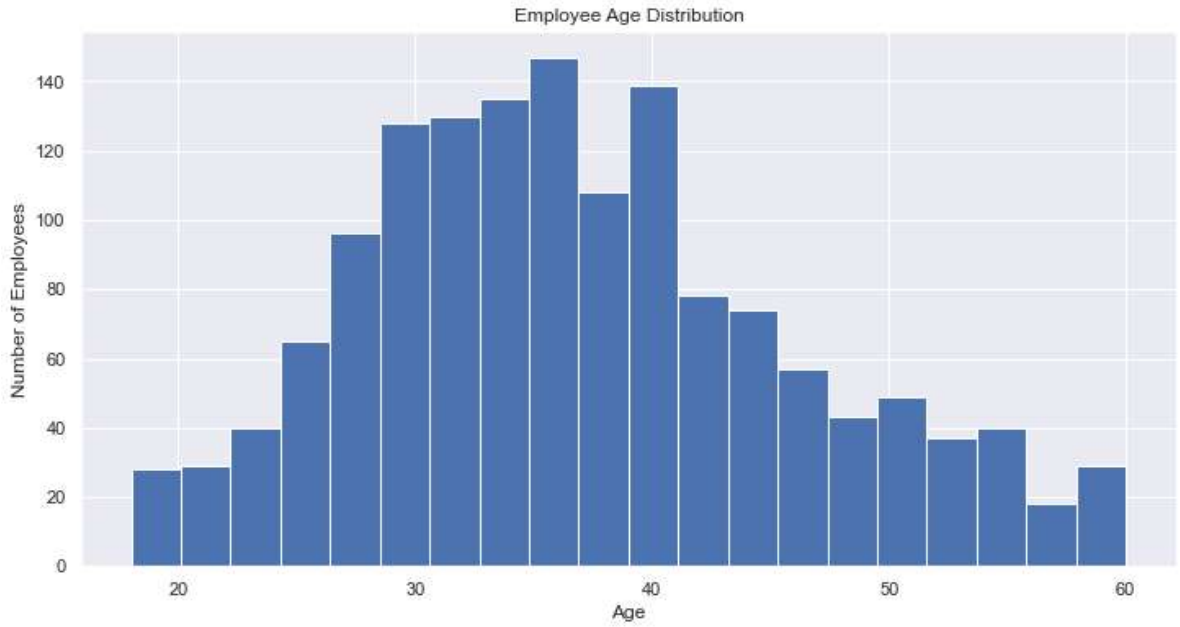
Interpretation 6.1 : The highest number of employees are with the education field of Life-Science by with we can interpret that the more jobs are available in the market for this field.

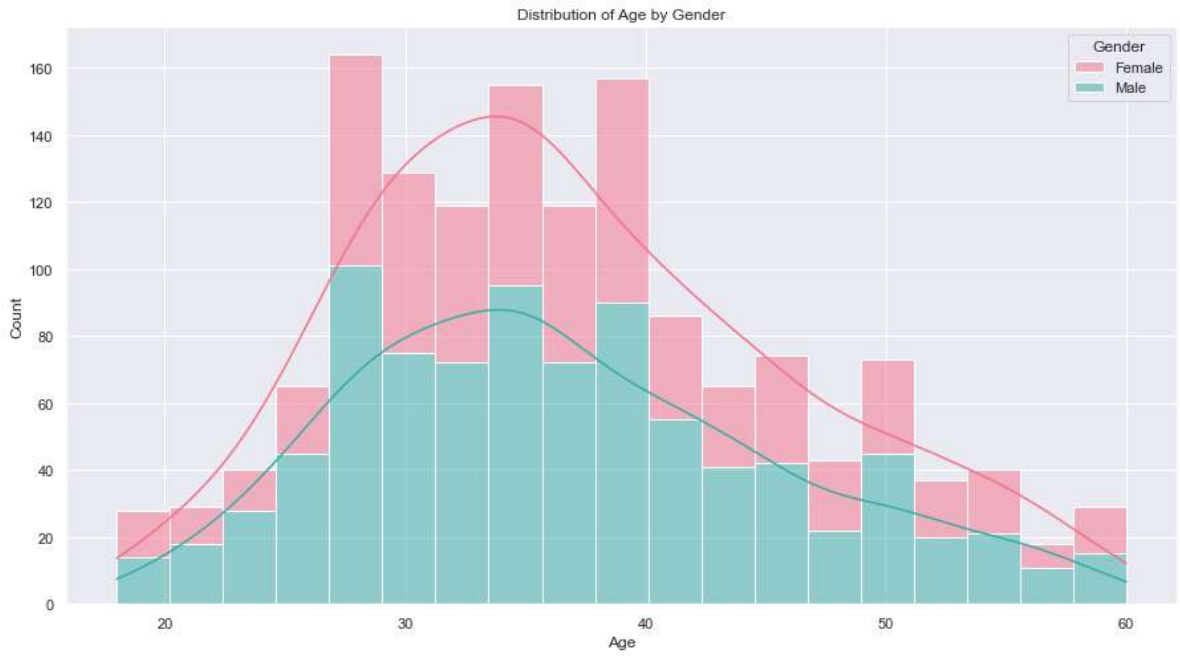
Interpretation 6.2 : Looking together at both the graphs 6.1 and 6.2 we can see that despite being the lowest number of employees are from the Human Resourse Field yet have the highest attrition rate of 35%.

Interpretation 6.3 : Looking together at both the graphs 6.2 and 6.3. High attrition rate is defined by the low job satisfaction role of the people with a background of Human Resourses.

7. Employee Age Distribution

In [11]:
```
age_distribution = df['Age'].hist(bins=20)
plt.title('Employee Age Distribution')
plt.xlabel('Age')
plt.ylabel('Number of Employees')
plt.show()


# Employee Demographics
plt.figure(figsize=(15, 8))
sns.histplot(data=df, x='Age', hue='Gender', multiple='stack', kde=True, palet
plt.title('Distribution of Age by Gender')
plt.show()
```

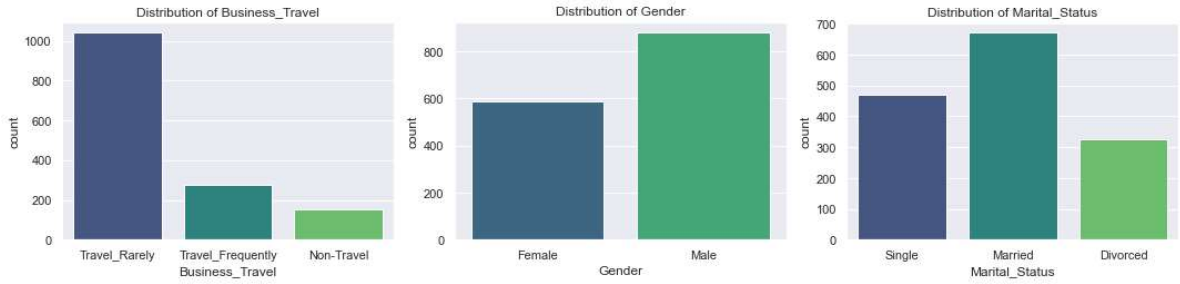




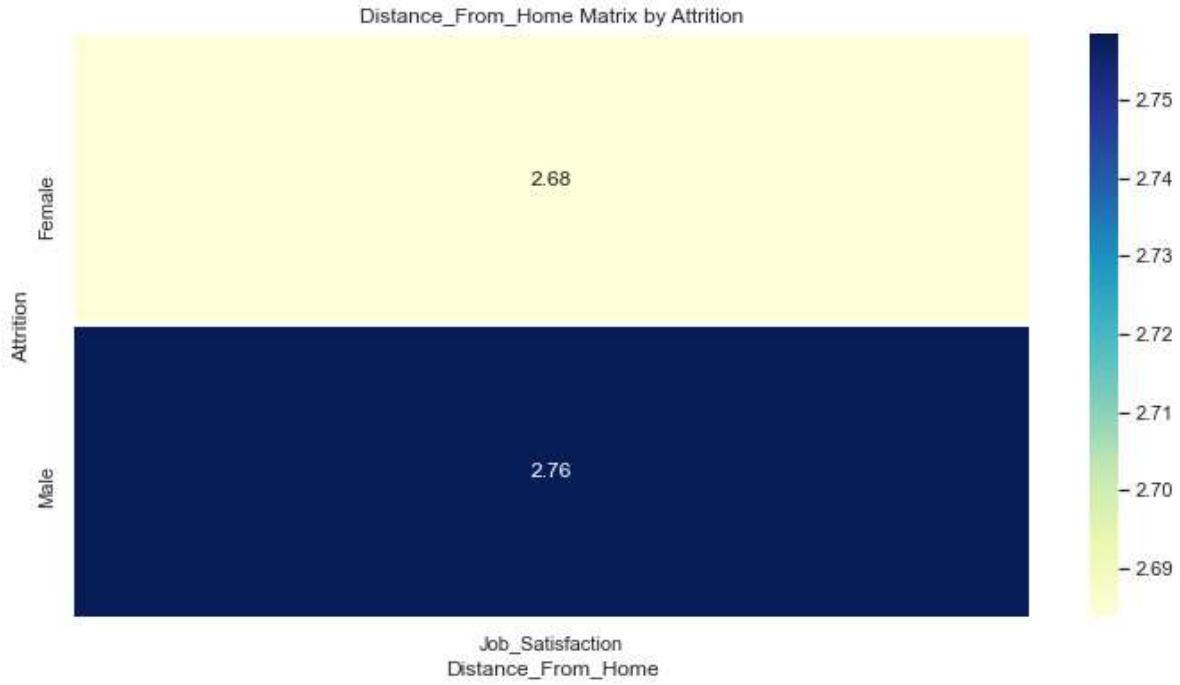
Interpretation: We can see that the age of the employees are normally distributed with a range of 20 to 60. Also, both the age of male and female employees are normally distributed. We can

8. Employee distribution by Business travel, gender and maritial status

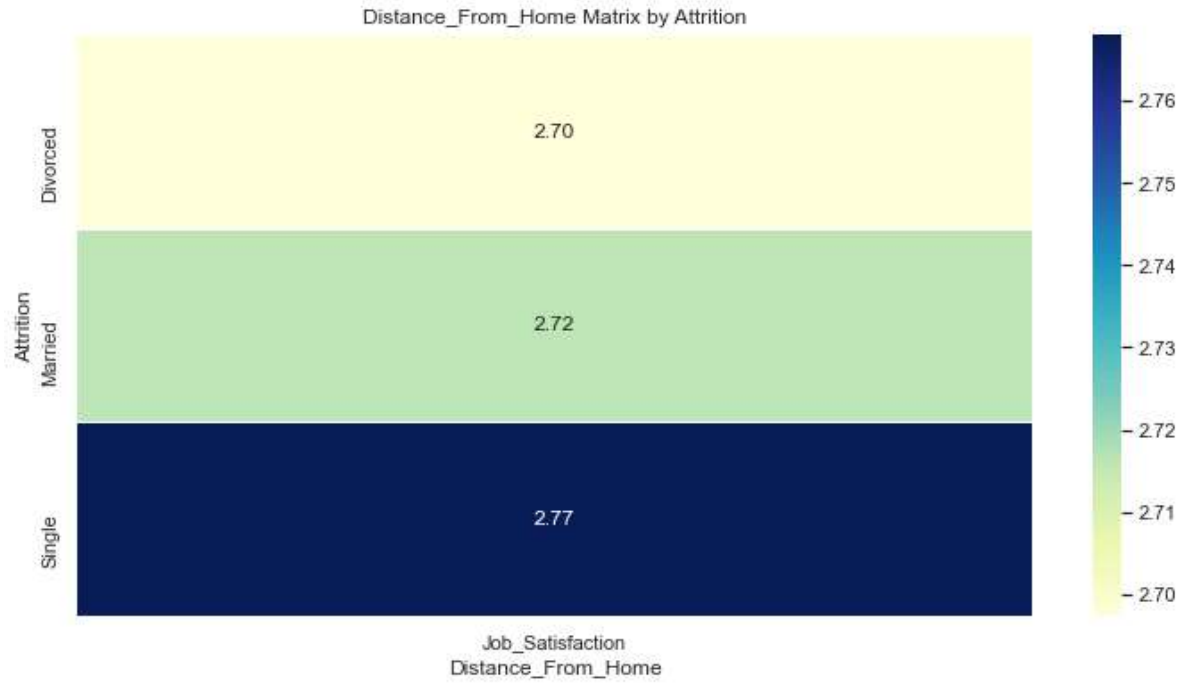
In [12]:
```
categorical_columns = ['Business_Travel', 'Gender', 'Marital_Status']

plt.figure(figsize=(15, 10))
for i, col in enumerate(categorical_columns, 1):
    plt.subplot(3, 3, i)
    sns.countplot(data=df, x=col, palette='viridis')
    plt.title(f'Distribution of {col}')
plt.tight_layout()
plt.show()


business_trvl=df['Business_Travel'].value_counts(normalize=True)*100
print(business_trvl)
print('-------------------------------------------------------------------------
sex=df['Gender'].value_counts(normalize=True)*100
print(sex)
print('-------------------------------------------------------------------------

df['Marital_Status'].value_counts(normalize=True)*100


job_satisfaction_matrix = pd.pivot_table(df, values='Job_Satisfaction', index=

sns.set(rc={'figure.figsize': (12, 6)})


sns.heatmap(job_satisfaction_matrix, cmap='YlGnBu', annot=True, fmt=".2f", lir
plt.title('Distance_From_Home Matrix by Attrition')
plt.xlabel('Distance_From_Home')
plt.ylabel('Attrition')
plt.show()

print('-------------------------------------------------------------------------


df['Marital_Status'].value_counts(normalize=True)*100


job_satisfaction_matrix = pd.pivot_table(df, values='Job_Satisfaction', index=

sns.set(rc={'figure.figsize': (12, 6)})


sns.heatmap(job_satisfaction_matrix, cmap='YlGnBu', annot=True, fmt=".2f", lir
plt.title('Distance_From_Home Matrix by Attrition')
plt.xlabel('Distance_From_Home')
plt.ylabel('Attrition')
plt.show()
```




```
Travel_Rarely        70.952381
Travel_Frequently    18.843537
Non-Travel           10.204082
Name: Business_Travel, dtype: float64
--------------------------------------------------------------------------------
Male      60.0
Female    40.0
Name: Gender, dtype: float64
--------------------------------------------------------------------------------
```




```
--------------------------------------------------------------------------------
-------------------------
```

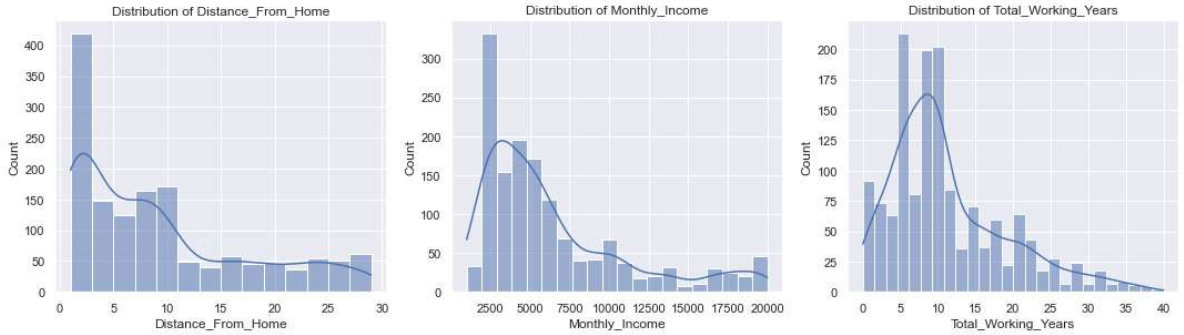


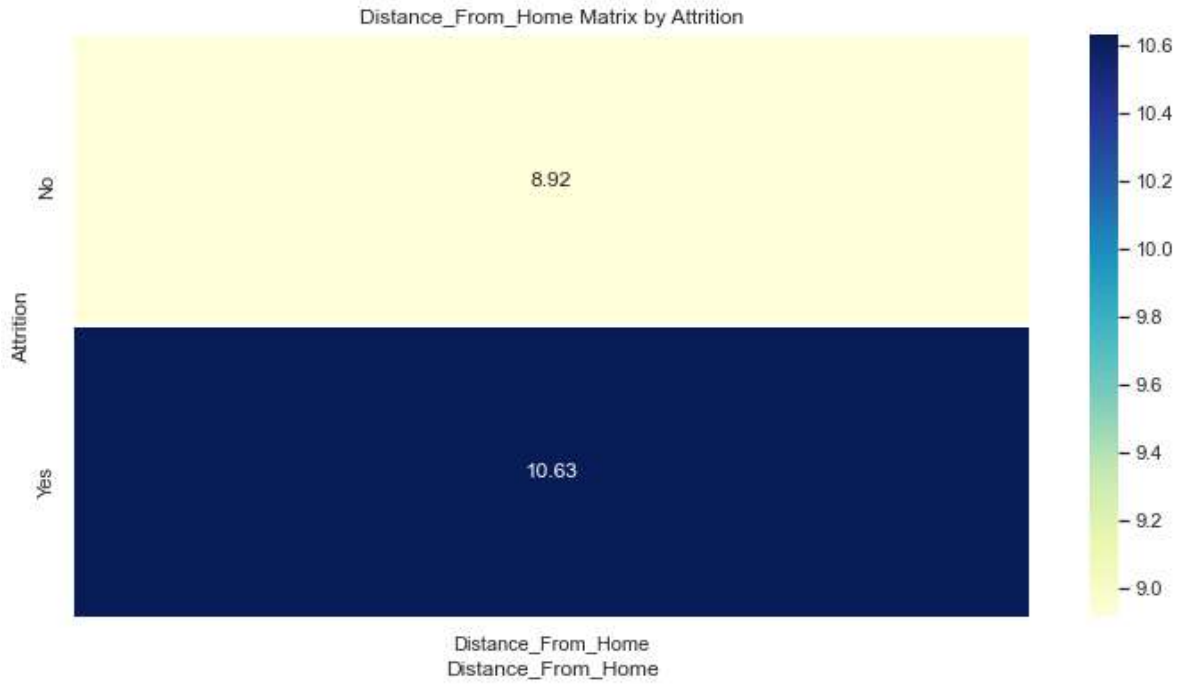
Interpretation: Interpretation: These charts give an overview about the employees of the oranization.

- Only 18% of the employees contribute to the travel expenses.
- We can see clearly that male count is more than the female count which is 60% and 40% respectively.
- Majority of the employees are married

9. Distribution of distance from home, Monthly income and working hours

In [13]:
```
numerical_columns = ['Distance_From_Home', 'Monthly_Income', 'Total_Working_Ye

plt.figure(figsize=(15, 8))
for i, col in enumerate(numerical_columns, 1):
    plt.subplot(2, 3, i)
    sns.histplot(df[col], kde=True)
    plt.title(f'Distribution of {col}')
plt.tight_layout()
plt.show()

job_satisfaction_matrix = pd.pivot_table(df, values='Distance_From_Home', inde

sns.set(rc={'figure.figsize': (12, 6)})


sns.heatmap(job_satisfaction_matrix, cmap='YlGnBu', annot=True, fmt=".2f", lir
plt.title('Distance_From_Home Matrix by Attrition')
plt.xlabel('Distance_From_Home')
plt.ylabel('Attrition')
plt.show()
```





Interpretation: From graph one we can see that the majority of the employees lives within the 0 to 5 Km of distance from the office and only some of them lives away from the office. We can see higher attrition from the employees living far from the office, so distance could be a major reason for attrition. From graph 2 most of the employees are getting the salary below 5000. and majority of the employees are working 10 hours a day.

Conclusion

In conclusion, the analysis of the "Exploring Employee Attrition" dataset has provided valuable insights into the complex dynamics influencing employee turnover within the organization. By leveraging Python for statistical exploration and visualization, we have uncovered key patterns and relationships that contribute to a deeper understanding of attrition factors such as Attrition Overview, Demographic Factors, Job Satisfaction Impact, Financial Considerations, Overtime and Work-Life Balance and Financial Considerations. This report equips organizations with actionable information to refine their retention strategies. By addressing specific pain points identified in the analysis, businesses can cultivate a work environment that fosters employee satisfaction, loyalty, and long-term commitment. The findings presented herein lay the groundwork for strategic decision-making, enabling organizations to proactively manage attrition and cultivate a thriving workplace culture.