

employee-attribution-project

August 7, 2024

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

```
[11]: df=pd.read_csv(r"C:\Users\priya\Downloads\WA_Fn-UseC_-HR-Employee-Attrition.
↪csv")
df
```

[11]:	Age	Attrition	BusinessTravel	DailyRate	Department	\
0	41	Yes	Travel_Rarely	1102		Sales
1	49	No	Travel_Frequently	279	Research & Development	
2	37	Yes	Travel_Rarely	1373	Research & Development	
3	33	No	Travel_Frequently	1392	Research & Development	
4	27	No	Travel_Rarely	591	Research & Development	
...	...					
1465	36	No	Travel_Frequently	884	Research & Development	
1466	39	No	Travel_Rarely	613	Research & Development	
1467	27	No	Travel_Rarely	155	Research & Development	
1468	49	No	Travel_Frequently	1023		Sales
1469	34	No	Travel_Rarely	628	Research & Development	
	DistanceFromHome	Education	EducationField	EmployeeCount	\	
0	1	2	Life Sciences	1		
1	8	1	Life Sciences	1		
2	2	2	Other	1		
3	3	4	Life Sciences	1		
4	2	1	Medical	1		
...	...					
1465	23	2	Medical	1		
1466	6	1	Medical	1		
1467	4	3	Life Sciences	1		
1468	2	3	Medical	1		
1469	8	3	Medical	1		
	EmployeeNumber	...	RelationshipSatisfaction	StandardHours	\	
0	1	...	1	80		

1	2	...	4	80
2	4	...	2	80
3	5	...	3	80
4	7	...	4	80
...
1465	2061	...	3	80
1466	2062	...	1	80
1467	2064	...	2	80
1468	2065	...	4	80
1469	2068	...	1	80

	StockOptionLevel	TotalWorkingYears	TrainingTimesLastYear	\
0	0	8	0	
1	1	10	3	
2	0	7	3	
3	0	8	3	
4	1	6	3	
...	
1465	1	17	3	
1466	1	9	5	
1467	1	6	0	
1468	0	17	3	
1469	0	6	3	

	WorkLifeBalance	YearsAtCompany	YearsInCurrentRole	\
0	1	6	4	
1	3	10	7	
2	3	0	0	
3	3	8	7	
4	3	2	2	
...	
1465	3	5	2	
1466	3	7	7	
1467	3	6	2	
1468	2	9	6	
1469	4	4	3	

	YearsSinceLastPromotion	YearsWithCurrManager
0	0	5
1	1	7
2	0	0
3	3	0
4	2	2
...
1465	0	3
1466	1	7
1467	0	3

```
1468          0          8
1469          1          2
```

```
[1470 rows x 35 columns]
```

```
[58]: df.head()
```

```
[58]:   Age Attrition   BusinessTravel   DailyRate   Department \
0    41      Yes   Travel_Rarely    1102      Sales
1    49      No  Travel_Frequently    279  Research & Development
2    37      Yes   Travel_Rarely    1373  Research & Development
3    33      No  Travel_Frequently    1392  Research & Development
4    27      No   Travel_Rarely     591  Research & Development

      DistanceFromHome  Education EducationField  EmployeeCount  EmployeeNumber \
0                   1          2  Life Sciences             1             1
1                   8          1  Life Sciences             1             2
2                   2          2          Other             1             4
3                   3          4  Life Sciences             1             5
4                   2          1          Medical            1             7

      ... StandardHours StockOptionLevel  TotalWorkingYears \
0      ...           80              0              8
1      ...           80              1             10
2      ...           80              0              7
3      ...           80              0              8
4      ...           80              1              6

      TrainingTimesLastYear  WorkLifeBalance  YearsAtCompany  YearsInCurrentRole \
0                   0          1          6          4
1                   3          3         10          7
2                   3          3          0          0
3                   3          3          8          7
4                   3          3          2          2

      YearsSinceLastPromotion  YearsWithCurrManager  age_group
0                   0          5  Millennials
1                   1          7        Gen X
2                   0          0  Millennials
3                   3          0  Millennials
4                   2          2        Gen Z
```

```
[5 rows x 36 columns]
```

```
[12]: df.shape
```

```
[12]: (1470, 35)
```

```
[14]: df.info()
```

```
<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
```

```
[13]: df.isnull().sum()
```

```
[13]: 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
```

```
[15]: df.columns
```

```
[15]: 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')

```

```

[41]: df.loc[df.Attrition == 'Yes']
df_attrition = df.loc[df.Attrition == 'Yes']
df_attrition

```

```

[41]:
   Age  Attrition  BusinessTravel  DailyRate  Department \
0    41         Yes      Travel_Rarely      1102         Sales
2    37         Yes      Travel_Rarely      1373  Research & Development
14   28         Yes      Travel_Rarely       103  Research & Development
21   36         Yes      Travel_Rarely      1218         Sales
24   34         Yes      Travel_Rarely       699  Research & Development
...  ...
1438  23         Yes  Travel_Frequently       638         Sales
1442  29         Yes      Travel_Rarely      1092  Research & Development
1444  56         Yes      Travel_Rarely       310  Research & Development
1452  50         Yes  Travel_Frequently       878         Sales
1461  50         Yes      Travel_Rarely       410         Sales

   DistanceFromHome  Education  EducationField  EmployeeCount \
0                  1          2      Life Sciences           1
2                  2          2           Other           1
14                 24          3      Life Sciences           1
21                  9          4      Life Sciences           1
24                  6          1          Medical           1
...
1438                 9          3      Marketing           1
1442                 1          4          Medical           1
1444                 7          2  Technical Degree           1
1452                 1          4      Life Sciences           1
1461                28          3      Marketing           1

   EmployeeNumber  ...  StandardHours  StockOptionLevel  TotalWorkingYears \
0                1  ...             80                0                8
2                4  ...             80                0                7
14               19  ...             80                0                6
21               27  ...             80                0               10
24               31  ...             80                0                8
...
1438            2023  ...             80                1                1
1442            2027  ...             80                3                4
1444            2032  ...             80                1               14
1452            2044  ...             80                2               12
1461            2055  ...             80                1               20

```

	TrainingTimesLastYear	WorkLifeBalance	YearsAtCompany	\
0	0	1	6	
2	3	3	0	
14	4	3	4	
21	4	3	5	
24	2	3	4	
...	
1438	3	2	1	
1442	3	4	2	
1444	4	1	10	
1452	3	3	6	
1461	3	3	3	

	YearsInCurrentRole	YearsSinceLastPromotion	YearsWithCurrManager	\
0	4	0	5	
2	0	0	0	
14	2	0	3	
21	3	0	3	
24	2	1	3	
...	
1438	0	1	0	
1442	2	2	2	
1444	9	9	8	
1452	3	0	1	
1461	2	2	0	

	age_group
0	Millennials
2	Millennials
14	Gen Z
21	Millennials
24	Millennials
...	...
1438	Gen Z
1442	Millennials
1444	Gen X
1452	Gen X
1461	Gen X

[237 rows x 36 columns]

```
[16]: df.Attrition
```

```
[16]: 0    Yes
      1    No
      2    Yes
      3    No
```

```

4          No
...
1465      No
1466      No
1467      No
1468      No
1469      No
Name: Attrition, Length: 1470, dtype: object

```

```
[59]: df_attrition.duplicated().sum()
```

```
[59]: 0
```

```
[26]: df['Attrition'].value_counts()
```

```

[26]: No      1233
      Yes      237
      Name: Attrition, dtype: int64

```

```
[42]: df_attrition['EducationField'].unique()
```

```

[42]: array(['Life Sciences', 'Other', 'Medical', 'Technical Degree',
            'Marketing', 'Human Resources'], dtype=object)

```

```
[43]: df_attrition['Gender'].value_counts()
```

```

[43]: Male      150
      Female    87
      Name: Gender, dtype: int64

```

```

[44]: age_range = [13,28,46,58,61]
      labels = ['Gen Z', 'Millennials', 'Gen X', 'Baby Boomers']
      df['age_group'] = pd.cut(df['Age'], age_range, labels=labels)

```

```

[45]: plt.figure(figsize = (10,10))

      plt.subplot(1,2,1)

      df['age_group'].value_counts().plot(kind='pie', title = "Percentage of All_
      ↪Employees Per Age Group")

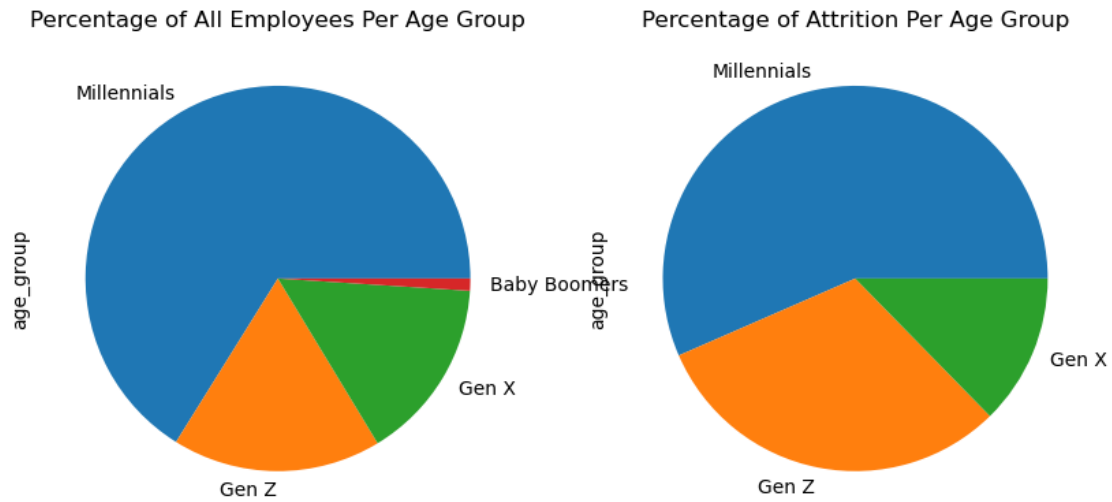
      plt.subplot(1,2,2)

      df_attrition['age_group'].value_counts().plot(kind='pie', title = "Percentage_
      ↪of Attrition Per Age Group")

```



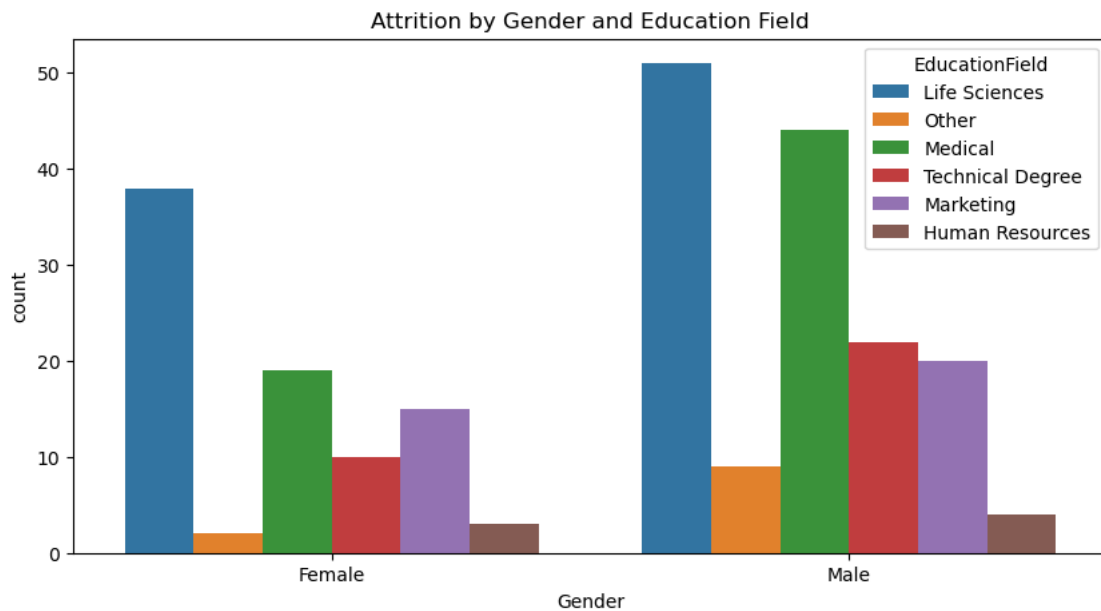
```
[45]: <Axes: title={'center': 'Percentage of Attrition Per Age Group'},
      ylabel='age_group'>
```



```
[46]: plt.figure(figsize = (10,5))

plt.title("Attrition by Gender and Education Field")
sns.countplot(x = 'Gender', hue = 'EducationField', data = df_attrition)
```

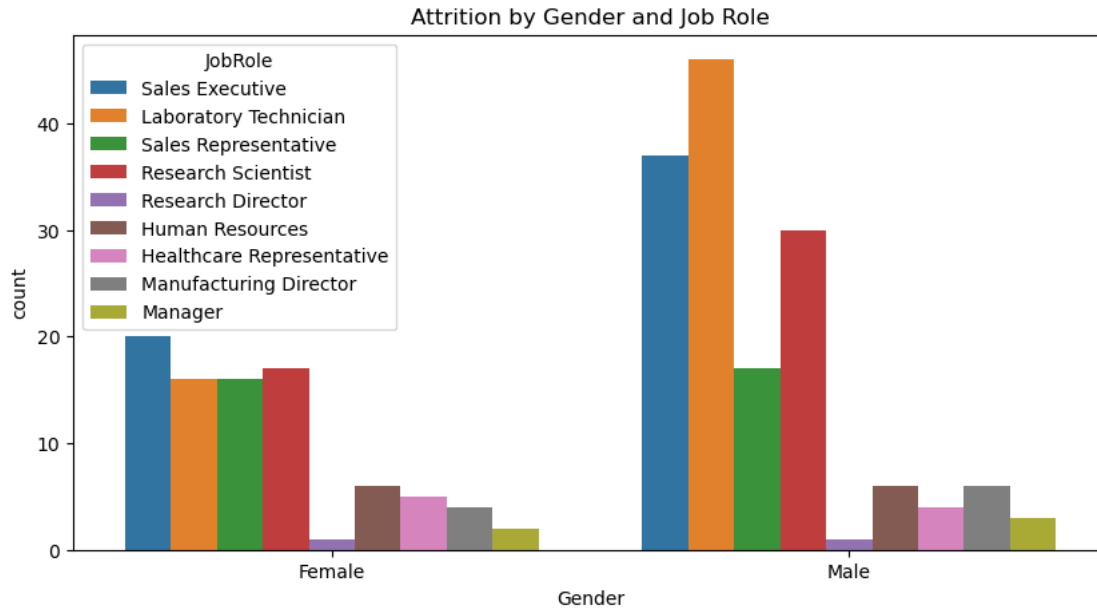
```
[46]: <Axes: title={'center': 'Attrition by Gender and Education Field'},
      xlabel='Gender', ylabel='count'>
```



```
[47]: plt.figure(figsize = (10,5))

plt.title("Attrition by Gender and Job Role")
sns.countplot(x = 'Gender', hue = 'JobRole', data = df_attrition)
```

```
[47]: <Axes: title={'center': 'Attrition by Gender and Job Role'}, xlabel='Gender',
ylabel='count'>
```



```
[48]: plt.figure(figsize = (10,10))

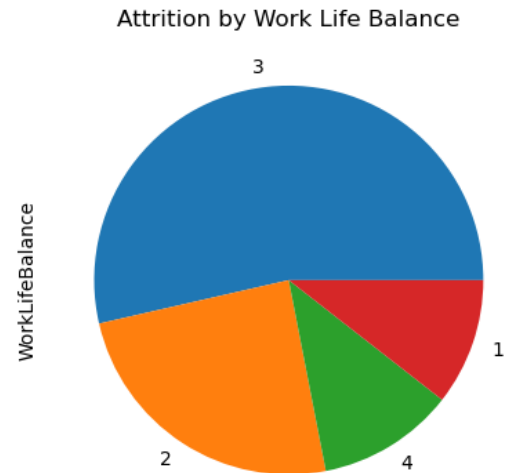
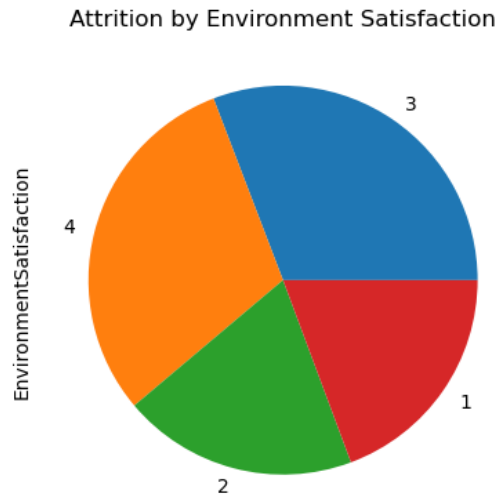
plt.subplot(1,2,1)

df['EnvironmentSatisfaction'].value_counts().plot(kind='pie', title = "Attrition by Environment Satisfaction")

plt.subplot(1,2,2)

df_attrition['WorkLifeBalance'].value_counts().plot(kind='pie', title = "Attrition by Work Life Balance")
```

```
[48]: <Axes: title={'center': 'Attrition by Work Life Balance'},
ylabel='WorkLifeBalance'>
```



```
[49]: plt.figure(figsize = (10,10))

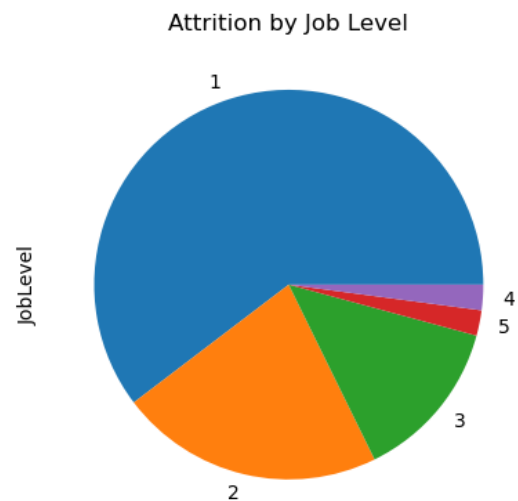
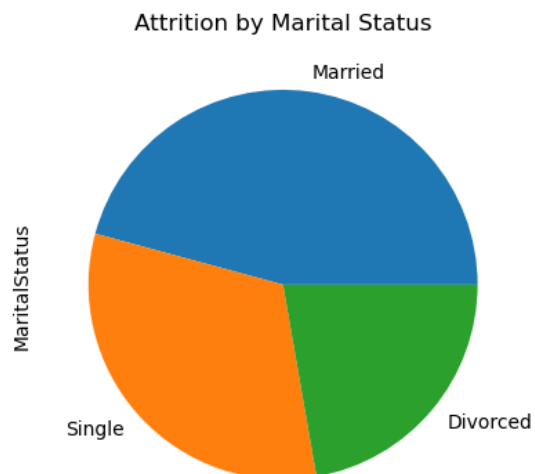
plt.subplot(1,2,1)

df['MaritalStatus'].value_counts().plot(kind='pie', title = "Attrition by_
↳Marital Status")

plt.subplot(1,2,2)

df_attrition['JobLevel'].value_counts().plot(kind='pie', title = "Attrition by_
↳Job Level")
```

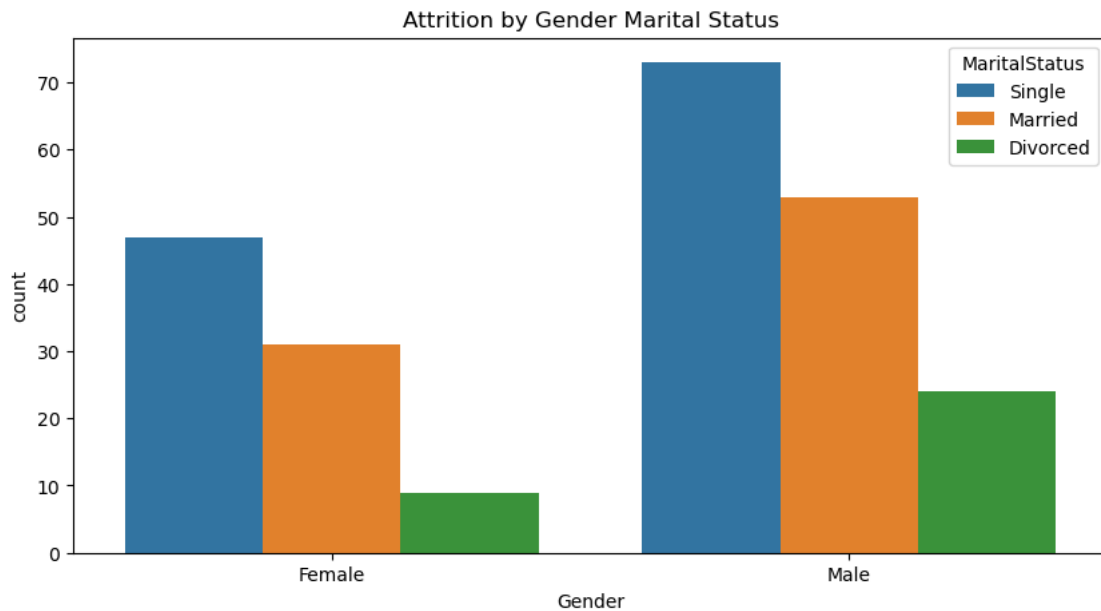
```
[49]: <Axes: title={'center': 'Attrition by Job Level'}, ylabel='JobLevel'>
```



```
[50]: plt.figure(figsize = (10,5))

plt.title("Attrition by Gender Marital Status")
sns.countplot(x = 'Gender', hue = 'MaritalStatus', data = df_attrition)
```

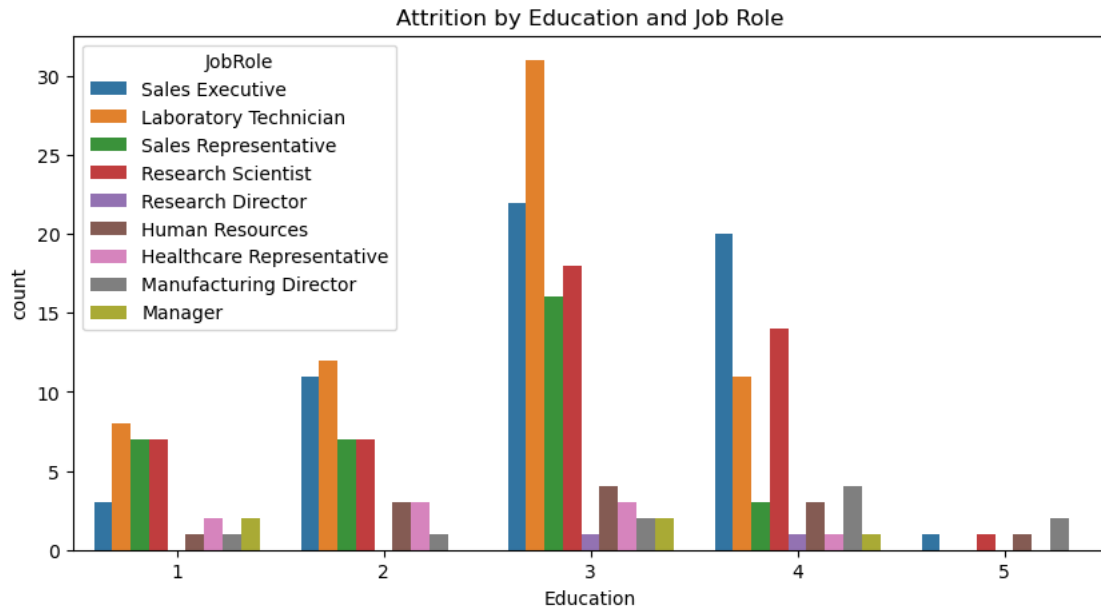
```
[50]: <Axes: title={'center': 'Attrition by Gender Marital Status'}, xlabel='Gender',
ylabel='count'>
```



```
[51]: plt.figure(figsize = (10,5))

plt.title("Attrition by Education and Job Role")
sns.countplot(x = 'Education', hue = 'JobRole', data = df_attrition)
```

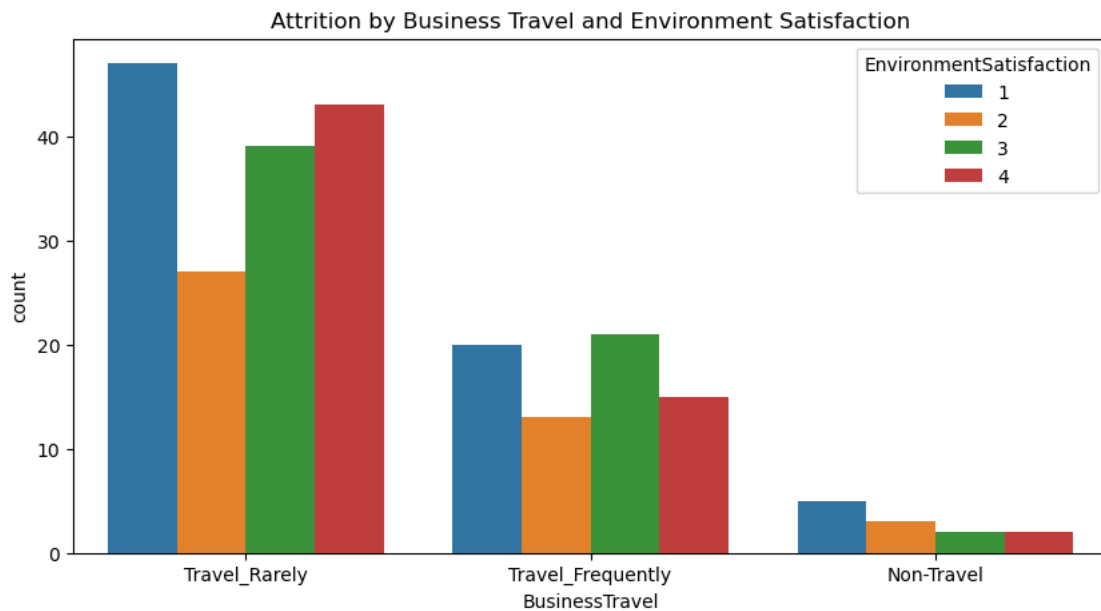
```
[51]: <Axes: title={'center': 'Attrition by Education and Job Role'},
xlabel='Education', ylabel='count'>
```



```
[52]: plt.figure(figsize = (10,5))

plt.title("Attrition by Business Travel and Environment Satisfaction")
sns.countplot(x = 'BusinessTravel', hue = 'EnvironmentSatisfaction', data = df_attrition)
```

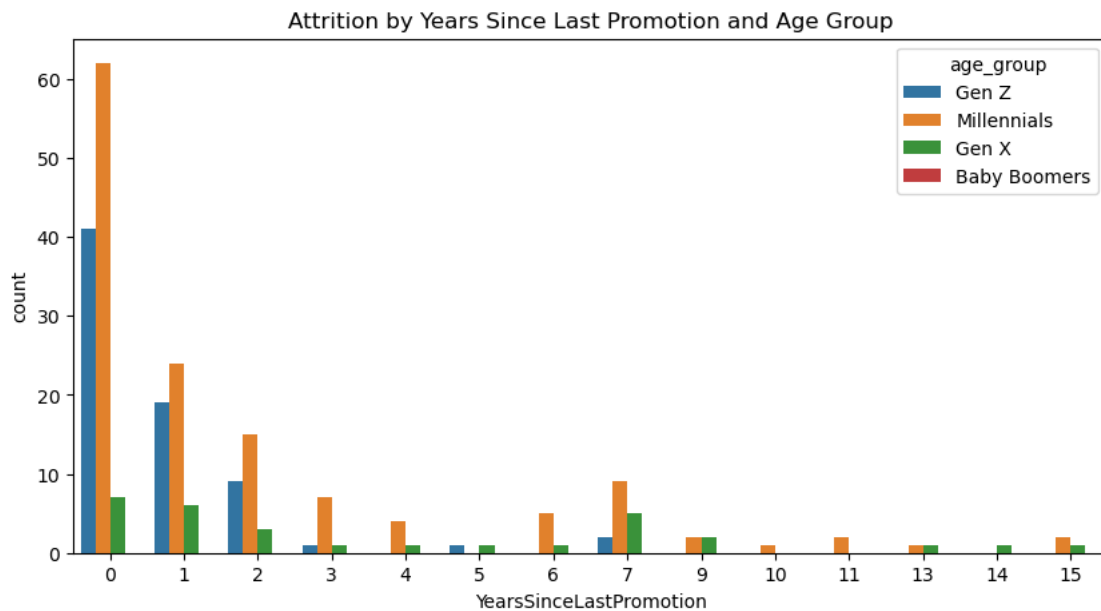
```
[52]: <Axes: title={'center': 'Attrition by Business Travel and Environment Satisfaction'}, xlabel='BusinessTravel', ylabel='count'>
```



```
[53]: plt.figure(figsize = (10,5))

plt.title("Attrition by Years Since Last Promotion and Age Group")
sns.countplot(x = 'YearsSinceLastPromotion', hue = 'age_group', data = df_attrition)
```

```
[53]: <Axes: title={'center': 'Attrition by Years Since Last Promotion and Age Group'}, xlabel='YearsSinceLastPromotion', ylabel='count'>
```



```
[57]: plt.figure(figsize = (10,10))

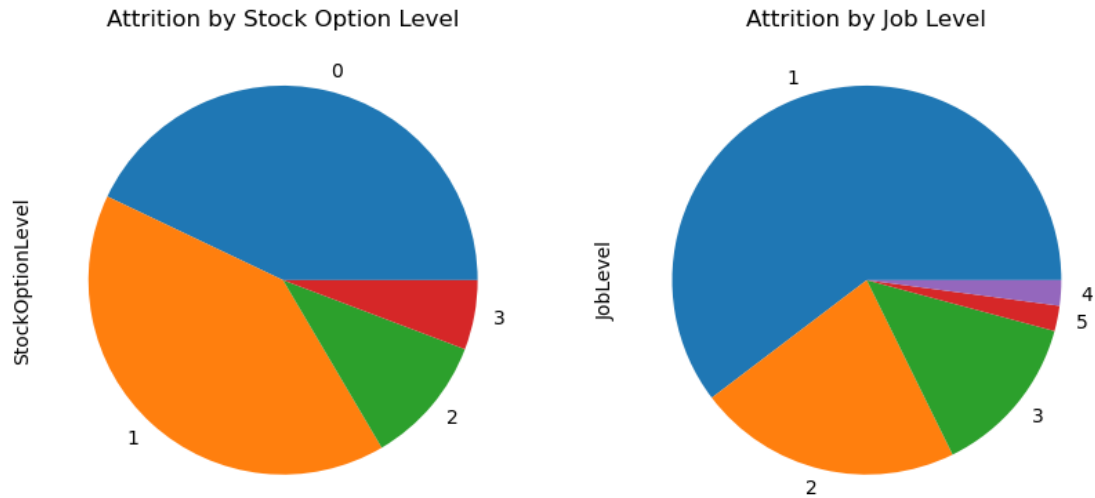
plt.subplot(1,2,1)

df['StockOptionLevel'].value_counts().plot(kind='pie', title = "Attrition by Stock Option Level")

plt.subplot(1,2,2)

df_attrition['JobLevel'].value_counts().plot(kind='pie', title = "Attrition by Job Level")
```

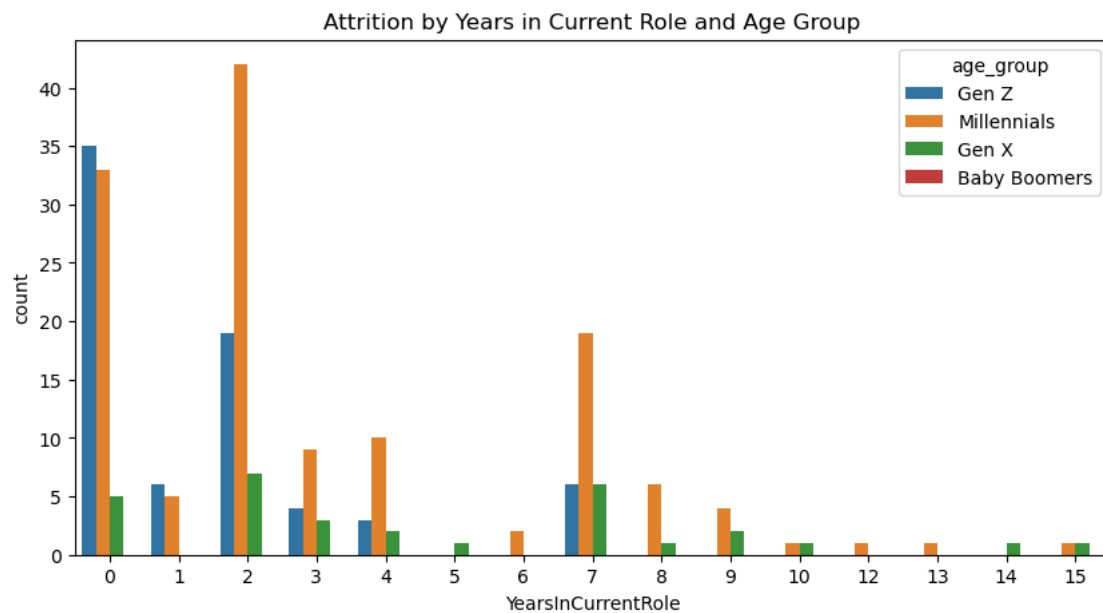
```
[57]: <Axes: title={'center': 'Attrition by Job Level'}, ylabel='JobLevel'>
```



```
[54]: plt.figure(figsize = (10,5))

plt.title("Attrition by Years in Current Role and Age Group")
sns.countplot(x = 'YearsInCurrentRole', hue = 'age_group', data = df_attrition)
```

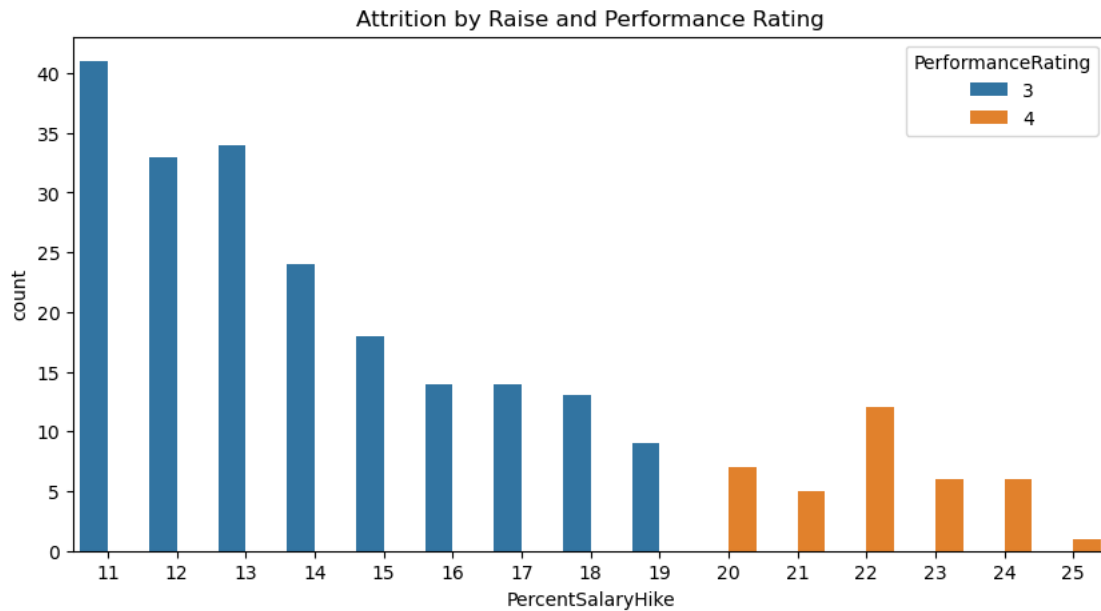
```
[54]: <Axes: title={'center': 'Attrition by Years in Current Role and Age Group'},
      xlabel='YearsInCurrentRole', ylabel='count'>
```



```
[55]: plt.figure(figsize = (10,5))

plt.title("Attrition by Raise and Performance Rating")
sns.countplot(x = 'PercentSalaryHike', hue = 'PerformanceRating', data = df_attrition)
```

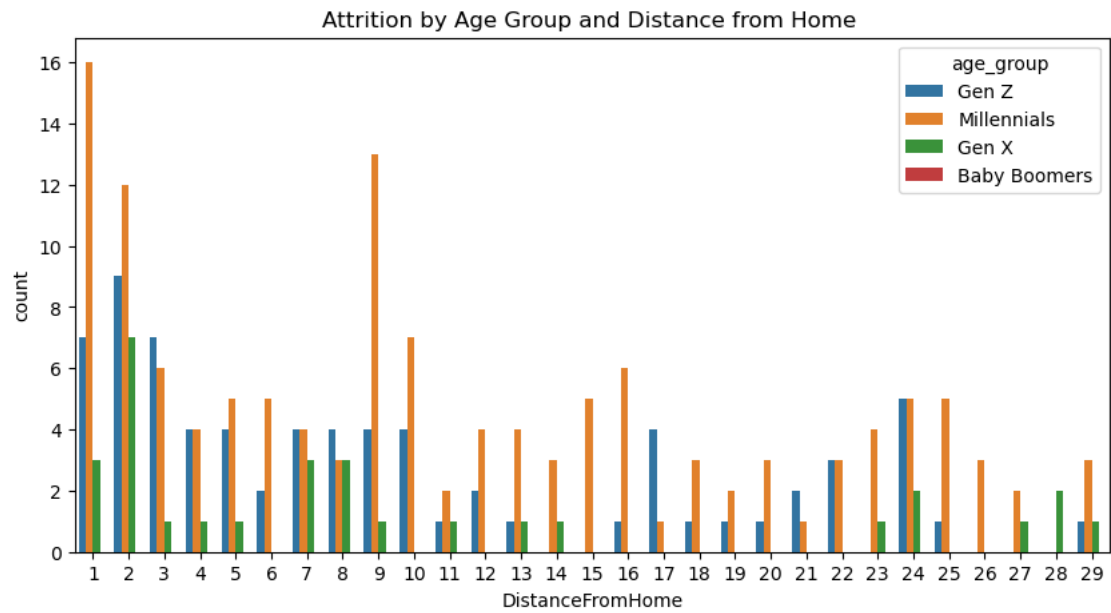
```
[55]: <Axes: title={'center': 'Attrition by Raise and Performance Rating'},
xlabel='PercentSalaryHike', ylabel='count'>
```



```
[56]: plt.figure(figsize = (10,5))

plt.title("Attrition by Age Group and Distance from Home ")
sns.countplot(x = 'DistanceFromHome', hue = 'age_group', data = df_attrition)
```

```
[56]: <Axes: title={'center': 'Attrition by Age Group and Distance from Home '},
xlabel='DistanceFromHome', ylabel='count'>
```

[]: