----------Human Resource Attrition Dataset Exploratory Data Analysis----------

In [2]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

In [5]:
```
hr = pd.read_csv('HR-Attrition.csv')
```

In [8]:
```
hr
```

Out[8]:

	Index	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	E
0	1	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	
1	2	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	
2	3	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	
3	4	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	
4	5	27	No	Travel_Rarely	591	Research & Development	2	1	Medical	
...	...	...	...	...	...	...	...	...	...	
1465	1466	36	No	Travel_Frequently	884	Research & Development	23	2	Medical	
1466	1467	39	No	Travel_Rarely	613	Research & Development	6	1	Medical	
1467	1468	27	No	Travel_Rarely	155	Research & Development	4	3	Life Sciences	
1468	1469	49	No	Travel_Frequently	1023	Sales	2	3	Medical	
1469	1470	34	No	Travel_Rarely	628	Research & Development	8	3	Medical	

1470 rows × 36 columns

In [11]:
```
hr_numeric = hr.select_dtypes(include=[np.number])
mean_values = hr_numeric.groupby(['WorkLifeBalance']).mean()
```

In [13]:
```
mean_values.head()
```

Out[13]:

WorkLifeBalance	Index	Age	DailyRate	DistanceFromHome	Education	EmployeeCount	EmployeeNumber
1	727.487500	37.287500	812.100000	9.425000	2.975000	1.0	1014.237500
2	736.337209	37.188953	827.459302	9.325581	2.857558	1.0	1026.511628
3	731.709966	36.849944	798.409854	9.277716	2.927212	1.0	1019.240761
4	759.928105	36.568627	765.098039	8.274510	2.921569	1.0	1059.549020

4 rows × 26 columns

In [210]:
```
hr.describe().T
```

Out[210]:

	count	mean	std	min	25%	50%	75%	max
Index	1470.0	735.500000	424.496761	1.0	368.25	735.5	1102.75	1470.0
Age	1470.0	36.923810	9.135373	18.0	30.00	36.0	43.00	60.0
DailyRate	1470.0	802.485714	403.509100	102.0	465.00	802.0	1157.00	1499.0
DistanceFromHome	1470.0	9.192517	8.106864	1.0	2.00	7.0	14.00	29.0
Education	1470.0	2.912925	1.024165	1.0	2.00	3.0	4.00	5.0
EmployeeCount	1470.0	1.000000	0.000000	1.0	1.00	1.0	1.00	1.0
EmployeeNumber	1470.0	1024.865306	602.024335	1.0	491.25	1020.5	1555.75	2068.0
EnvironmentSatisfaction	1470.0	2.721769	1.093082	1.0	2.00	3.0	4.00	4.0
HourlyRate	1470.0	65.891156	20.329428	30.0	48.00	66.0	83.75	100.0
JobInvolvement	1470.0	2.729932	0.711561	1.0	2.00	3.0	3.00	4.0
JobLevel	1470.0	2.063946	1.106940	1.0	1.00	2.0	3.00	5.0
JobSatisfaction	1470.0	2.728571	1.102846	1.0	2.00	3.0	4.00	4.0
MonthlyIncome	1470.0	6502.931293	4707.956783	1009.0	2911.00	4919.0	8379.00	19999.0
MonthlyRate	1470.0	14313.103401	7117.786044	2094.0	8047.00	14235.5	20461.50	26999.0
NumCompaniesWorked	1470.0	2.693197	2.498009	0.0	1.00	2.0	4.00	9.0
PercentSalaryHike	1470.0	15.209524	3.659938	11.0	12.00	14.0	18.00	25.0
PerformanceRating	1470.0	3.153741	0.360824	3.0	3.00	3.0	3.00	4.0
RelationshipSatisfaction	1470.0	2.712245	1.081209	1.0	2.00	3.0	4.00	4.0
StandardHours	1470.0	80.000000	0.000000	80.0	80.00	80.0	80.00	80.0
StockOptionLevel	1470.0	0.793878	0.852077	0.0	0.00	1.0	1.00	3.0
TotalWorkingYears	1470.0	11.279592	7.780782	0.0	6.00	10.0	15.00	40.0
TrainingTimesLastYear	1470.0	2.799320	1.289271	0.0	2.00	3.0	3.00	6.0
WorkLifeBalance	1470.0	2.761224	0.706476	1.0	2.00	3.0	3.00	4.0
YearsAtCompany	1470.0	7.008163	6.126525	0.0	3.00	5.0	9.00	40.0
YearsInCurrentRole	1470.0	4.229252	3.623137	0.0	2.00	3.0	7.00	18.0
YearsSinceLastPromotion	1470.0	2.187755	3.222430	0.0	0.00	1.0	3.00	15.0
YearsWithCurrManager	1470.0	4.123129	3.568136	0.0	2.00	3.0	7.00	17.0
Attrition_numeric	1470.0	0.161224	0.367863	0.0	0.00	0.0	0.00	1.0
Attrition_num	1470.0	0.161224	0.367863	0.0	0.00	0.0	0.00	1.0
overtime_num	1470.0	0.282993	0.450606	0.0	0.00	0.0	1.00	1.0

In [207]:
```
des_sta = hr.describe(include = ['object'])
```

In [209]:
```
des_sta.T
```

Out[209]:

	count	unique	top	freq
Attrition	1470	2	No	1233
BusinessTravel	1470	3	Travel_Rarely	1043
Department	1470	3	Research & Development	961
EducationField	1470	6	Life Sciences	606
Gender	1470	2	Male	882
JobRole	1470	9	Sales Executive	326
MaritalStatus	1470	3	Married	673
Over18	1470	1	Y	1470
OverTime	1470	2	No	1054

In [191]:
```
# Convert 'OverTime' column to numeric: 1 for 'Yes' and 0 for 'No'

hr['overtime_num'] = hr['OverTime'].apply(lambda x: 1 if x == 'Yes' else 0)
```

In [14]:
```
gbwg = hr.groupby(['WorkLifeBalance'])['HourlyRate'].mean()
```

In [203]:
```
gbwg.head()
```

Out[203]:
```
WorkLifeBalance
1    63.800000
2    66.502907
3    66.090705
4    64.444444
Name: HourlyRate, dtype: float64
```

In [16]:
```
gbwg1 = hr.groupby(['WorkLifeBalance'])[['HourlyRate','Age']].mean()
```

In [17]:
```
gbwg1.head()
```

Out[17]:

WorkLifeBalance	HourlyRate	Age
1	63.800000	37.287500
2	66.502907	37.188953
3	66.090705	36.849944
4	64.444444	36.568627

In [21]:
```
hr.columns
```

Out[21]:
```
Index(['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 [39]:
```
hr.head()
```

Out[39]:

	Index	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	Empl
0	1	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	
1	2	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	
2	3	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	
3	4	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	
4	5	27	No	Travel_Rarely	591	Research & Development	2	1	Medical	

5 rows × 36 columns

In [33]:
```
hr['Education'].count()
```

Out[33]:
```
1470
```

In [37]:

```
dfgf =hr.groupby(['JobRole'])[['Attrition','Education']]
```

In [25]:
```
dfgf.head()
```

Out[25]:

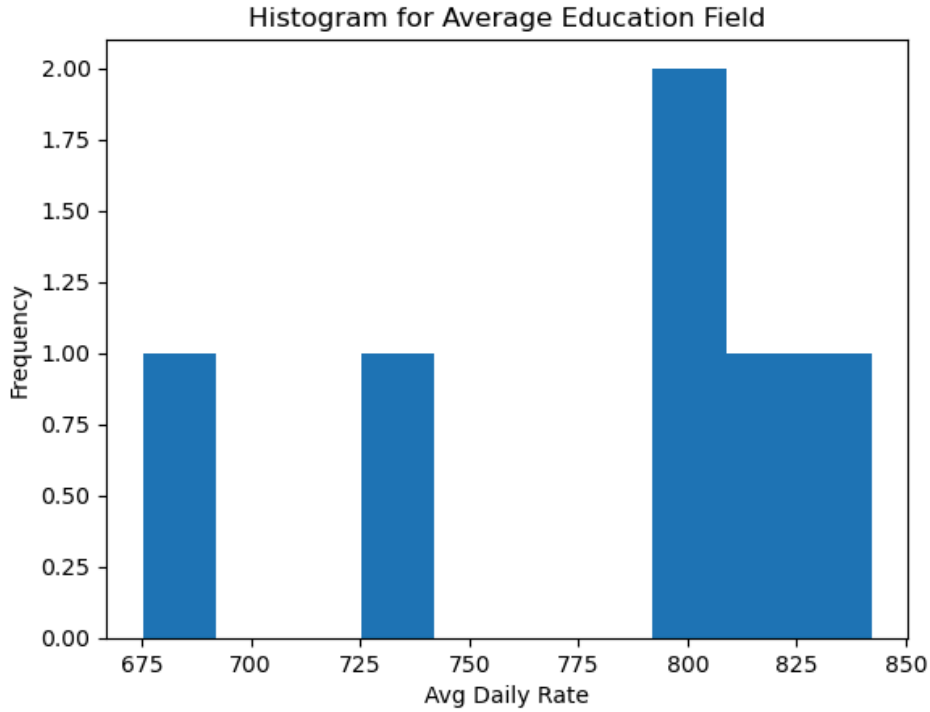
	Attrition	Education
0	Yes	2
1	No	1
2	Yes	2
3	No	4
4	No	1
5	No	2
6	No	3
7	No	1
8	No	3
9	No	3
12	No	1
15	No	4
16	No	2
18	No	4
19	No	3
20	No	2
21	Yes	4
22	No	4
25	No	3
27	No	4
28	No	4
29	No	4
31	No	4
33	Yes	3
36	Yes	2
37	No	3
39	No	3
43	No	3
45	Yes	3
46	No	4
53	No	2
55	No	2
59	No	4
60	No	3
62	No	2
64	No	3
65	No	3
77	No	4
79	No	2
86	No	1
100	Yes	4
105	No	4
134	No	1
139	No	3
232	No	2

In [49]:
```
dbef = hr.groupby(['EducationField'])['DailyRate'].mean().round(2)
```

In [50]:
```
dbef.head()
```

Out[50]:
```
EducationField
Human Resources    675.26
Life Sciences      804.43
Marketing          727.84
Medical            822.80
Other              796.02
Name: DailyRate, dtype: float64
```

In [61]:
```
dbef.plot(kind='hist')
plt.title('Histogram for Average Education Field')
plt.xlabel('Avg Daily Rate')
plt.ylabel('Frequency')
plt.show()
```



In [62]:
```
hr.head()
```

Out[62]:

...inessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	...	RelationshipSatis
...ravel_Rarely	1102	Sales	1	2	Life Sciences	1	...	
...l_Frequently	279	Research & Development	8	1	Life Sciences	1	...	
...ravel_Rarely	1373	Research & Development	2	2	Other	1	...	
...l_Frequently	1392	Research & Development	3	4	Life Sciences	1	...	
...ravel_Rarely	591	Research & Development	2	1	Medical	1	...	

In [81]:
```
unique_departments = hr['Department'].unique()
unique_departments
```

Out[81]:
```
array(['Sales', 'Research & Development', 'Human Resources'], dtype=object)
```

In [83]:
```
unique_EducationField =  hr['EducationField'].unique()
unique_EducationField
```

Out[83]:
```
array(['Life Sciences', 'Other', 'Medical', 'Marketing',
       'Technical Degree', 'Human Resources'], dtype=object)
```

In [91]:
```
same_dpt_field = hr[(hr['EducationField']== 'Human Resources') & (hr['Department']=='Human Resou
indexes = same_dpt_field.index
indexes
```

Out[91]:
```
Index([ 100,  105,  112,  139,  310,  440,  535,  538,  551,  599,  613,  655,
        826,  863,  999, 1107, 1165, 1222, 1228, 1246, 1289, 1312, 1313, 1347,
       1379, 1401, 1411],
      dtype='int64')
```

In [87]:
```
same_dpt_filed
```

Out[87]:

	Index	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	Er
100	101	37	Yes	Travel_Rarely	807	Human Resources	6	4	Human Resources	
105	106	59	No	Non-Travel	1420	Human Resources	2	4	Human Resources	
112	113	54	No	Non-Travel	142	Human Resources	26	3	Human Resources	
139	140	30	No	Travel_Rarely	1240	Human Resources	9	3	Human Resources	
310	311	31	No	Travel_Rarely	106	Human Resources	2	3	Human Resources	
440	441	34	Yes	Travel_Frequently	988	Human Resources	23	3	Human Resources	
535	536	41	No	Travel_Rarely	427	Human Resources	10	4	Human Resources	
538	539	41	No	Travel_Rarely	314	Human Resources	1	3	Human Resources	
551	552	39	No	Travel_Rarely	141	Human Resources	3	3	Human Resources	
599	600	36	No	Travel_Rarely	1041	Human Resources	13	3	Human Resources	
613	614	34	No	Travel_Rarely	829	Human Resources	3	2	Human Resources	
655	656	33	No	Travel_Rarely	1075	Human Resources	3	2	Human Resources	
826	827	38	No	Travel_Rarely	433	Human Resources	1	3	Human Resources	
863	864	33	No	Travel_Rarely	147	Human Resources	2	3	Human Resources	
999	1000	42	No	Travel_Rarely	1147	Human Resources	10	3	Human Resources	
1107	1108	38	No	Travel_Frequently	888	Human Resources	10	4	Human Resources	
1165	1166	44	No	Travel_Frequently	602	Human Resources	1	5	Human Resources	
1222	1223	24	Yes	Travel_Rarely	240	Human Resources	22	1	Human Resources	
1228	1229	41	No	Non-Travel	552	Human Resources	4	3	Human Resources	
1246	1247	30	Yes	Travel_Frequently	600	Human Resources	8	3	Human Resources	
1289	1290	38	No	Non-Travel	1336	Human Resources	2	3	Human Resources	
1312	1313	31	Yes	Travel_Rarely	359	Human Resources	18	5	Human Resources	
1313	1314	29	Yes	Travel_Rarely	350	Human Resources	13	3	Human Resources	
1347	1348	36	No	Travel_Frequently	1213	Human Resources	2	1	Human Resources	
1379	1380	27	Yes	Travel_Frequently	1337	Human Resources	22	3	Human Resources	
1401	1402	55	No	Travel_Rarely	189	Human Resources	26	4	Human Resources	
1411	1412	25	No	Travel_Rarely	309	Human Resources	2	3	Human Resources	

27 rows × 36 columns

In [67]:
```
Total_year_more_than_five = hr[hr['TotalWorkingYears'] > 5][['Department', 'TotalWorkingYears']]
```

In [68]:
```
Total_year_more_than_five.head()
```

Out[68]:

	Department	TotalWorkingYears
0	Sales	8
1	Research & Development	10
2	Research & Development	7
3	Research & Development	8
4	Research & Development	6

In [92]:
```
hr.columns
```

Out[92]:
```
Index(['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')
```

Group by 'Department' and calculate mean 'Attrition' rate

In [151]:
```
# Convert 'Attrition' column to numeric: 1 for 'Yes' and 0 for 'No'

hr['Attrition_num'] = hr['Attrition'].apply(lambda x: 1 if x == 'Yes' else 0)
```

In [154]:
```
dpt_attr_rate = hr.groupby('Department')['Attrition_num'].mean()
```

In [164]:
```
dpt_attr_rate
```

Out[164]:
```
Department
Human Resources           0.190476
Research & Development    0.138398
Sales                     0.206278
Name: Attrition_num, dtype: float64
```

Correlation analysis between 'DistanceFromHome' and 'JobSatisfaction'

In [182]:
```
corr1 = hr[['DistanceFromHome','JobSatisfaction']].corr
```

In [183]:
```
corr1()
```

Out[183]:

	DistanceFromHome	JobSatisfaction
DistanceFromHome	1.000000	-0.003669
JobSatisfaction	-0.003669	1.000000

Group by 'JobRole' and 'Education' to analyze 'MonthlyIncome'

In [173]:
```
mnt_salary =  hr.groupby(['JobRole', 'Education'])['MonthlyIncome'].mean()
```

In [178]:
```
mnt_salary
```

Out[178]:
```
JobRole                    Education
Healthcare Representative  1            8769.533333
                           2            7096.047619
                           3            7677.145833
                           4            7152.181818
                           5            7503.000000
Human Resources            1            2776.600000
                           2            3811.750000
                           3            4408.772727
                           4            4867.100000
                           5            4990.333333
Laboratory Technician      1            2982.257143
                           2            3256.245614
                           3            3234.490385
                           4            3268.913793
                           5            4491.600000
Manager                    1           17037.666667
                           2           17556.210526
                           3           16927.525000
                           4           17340.655172
                           5           17128.800000
Manufacturing Director     1            7063.866667
                           2            6519.818182
                           3            7949.296296
                           4            7117.105263
                           5            7394.200000
Research Director          1           15752.000000
                           2           15779.384615
                           3           16498.366667
                           4           15850.782609
                           5           15395.571429
Research Scientist         1            2918.378378
                           2            3326.391304
                           3            3038.245902
                           4            3632.645570
                           5            3429.125000
Sales Executive            1            6732.000000
                           2            6418.348485
                           3            7314.641667
                           4            6877.336634
                           5            6631.000000
Sales Representative       1            2489.350000
                           2            3004.733333
                           3            2382.968750
                           4            2927.812500
Name: MonthlyIncome, dtype: float64
```

Group by 'Department' and calculate average 'WorkLifeBalance'

In [179]:
```
avg_work_life_balance =  hr.groupby( 'Department' )['WorkLifeBalance'].mean()
```

In [181]:
```
avg_work_life_balance
```

Out[181]:
```
Department
Human Resources           2.920635
Research & Development    2.725286
Sales                     2.816143
Name: WorkLifeBalance, dtype: float64
```

Correlation or comparison between 'YearsWithCurrManager' and 'YearsSinceLastPromotion'

In [184]:
```
corr2 = hr[['YearsWithCurrManager','YearsWithCurrManager']].corr
```

In [186]:
```
corr2()
```

Out[186]:

	YearsWithCurrManager	YearsWithCurrManager
YearsWithCurrManager	1.0	1.0
YearsWithCurrManager	1.0	1.0

Use correlation or groupby analysis with 'PerformanceRating', 'JobInvolvement', and 'OverTime'

In [196]:
```
corr3= hr[['PerformanceRating', 'JobInvolvement']].corr()
```

In [198]:
```
corr3
```

Out[198]:

	PerformanceRating	JobInvolvement
PerformanceRating	1.000000	-0.029071
JobInvolvement	-0.029071	1.000000

Correlation analysis between 'TrainingTimesLastYear' and 'PerformanceRating'

In [199]:
```
corr4 = hr[['TrainingTimesLastYear','PerformanceRating']].corr
```

In [201]:
```
corr4()
```

Out[201]:

	TrainingTimesLastYear	PerformanceRating
TrainingTimesLastYear	1.000000	-0.015579
PerformanceRating	-0.015579	1.000000

In [101]:
```
pd.set_option('display.max_columns', None)
```

In [102]:
```
hr.head()
```

Out[102]:

ailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber	EnvironmentSat
1102	Sales	1	2	Life Sciences	1	1	
279	Research & Development	8	1	Life Sciences	1	2	
1373	Research & Development	2	2	Other	1	4	
1392	Research & Development	3	4	Life Sciences	1	5	
591	Research & Development	2	1	Medical	1	7	

In [140]:
```
emply_count = hr.groupby('Department')['EmployeeCount'].count()
```

In [146]:
```
emply_count.T
```

Out[146]:
```
Department
Human Resources            63
Research & Development    961
Sales                     446
Name: EmployeeCount, dtype: int64
```

In [131]:
```
gndr_by_dpt = hr.groupby('Department')['Gender'].value_counts().unstack()
```

In [132]:
```
gndr_by_dpt
```

Out[132]:

Gender	Female	Male
Department		
Human Resources	20	43
Research & Development	379	582
Sales	189	257

In [129]:
```
count_over_time_martial_status = hr.groupby('MaritalStatus')['OverTime'].value_counts().unstack(
```

In [123]:
```
count_over_time_martial_status
```

Out[123]:

OverTime	No	Yes
MaritalStatus		
Divorced	228	99
Married	487	186
Single	339	131

In [124]:
```
joblevel_by_gender = hr.groupby('Gender')['JobLevel'].value_counts().unstack()
```

In [126]:
```
joblevel_by_gender.T
```

Out[126]:

Gender	Female	Male
JobLevel		
1	199	344
2	220	314
3	94	124
4	51	55
5	24	45

In [ ]:
```

```