code

December 7, 2021

[119]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
from pylab import randn
import random
import plotly.graph_objects as go
```

[76]:
```
df = pd.read_csv('HR_Employee_Attrition.csv')
data_attr_yes = df[df.Attrition == 'Yes']
data_attr_no = df[df.Attrition == 'No']
df.head()
```

[76]:

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber	…	RelationshipSatisfaction	StandardHours	StockOptionLevel	TotalWorkingYears	TrainingTimesLastYear	WorkLifeBalance	YearsAtCompany
0	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	1	…	1	80	0	8	0	1	6
1	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	1	2	…	4	80	1				
2	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	1	4	…	2	80	0				
3	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	1	5	…	3	80	0				
4	27	No	Travel_Rarely	591	Research & Development	2	1	Medical	1	7	…	4	80	1				

```
1                 10                         3                3                10
2                  7                         3                3                 0
3                  8                         3                3                 8
4                  6                         3                3                 2

  YearsInCurrentRole  YearsSinceLastPromotion  YearsWithCurrManager
0                  4                        0                     5
1                  7                        1                     7
2                  0                        0                     0
3                  7                        3                     0
4                  2                        2                     2

[5 rows x 35 columns]
```

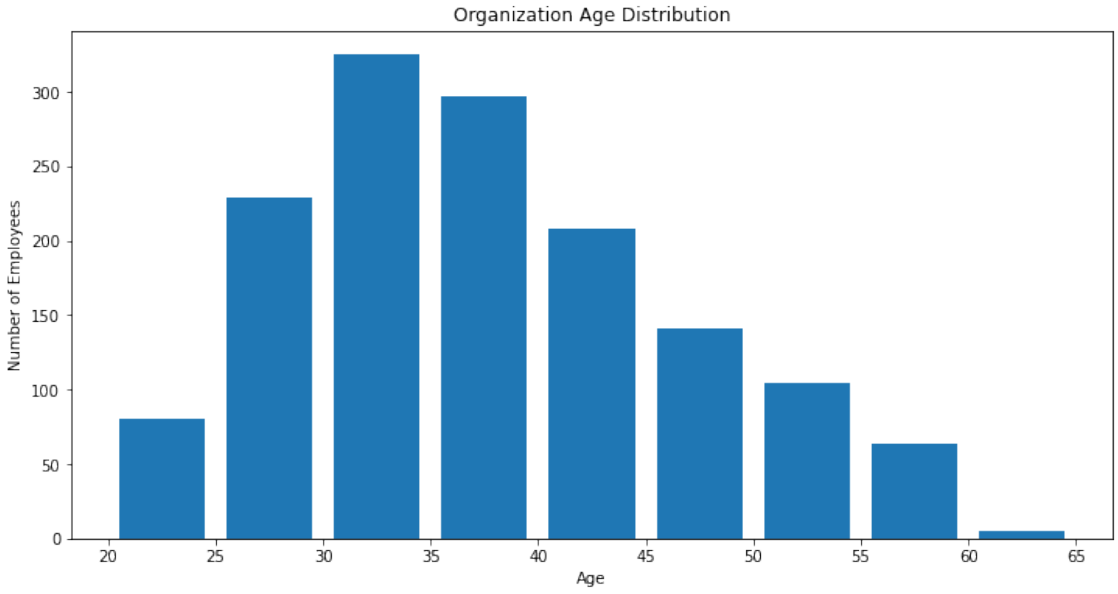
[57]:
```
df.describe()
```

[57]:
```
                Age    DailyRate  DistanceFromHome    Education  EmployeeCount  \
count  1470.000000  1470.000000       1470.000000  1470.000000         1470.0
mean     36.923810   802.485714          9.192517     2.912925            1.0
std       9.135373   403.509100          8.106864     1.024165            0.0
min      18.000000   102.000000          1.000000     1.000000            1.0
25%      30.000000   465.000000          2.000000     2.000000            1.0
50%      36.000000   802.000000          7.000000     3.000000            1.0
75%      43.000000  1157.000000         14.000000     4.000000            1.0
max      60.000000  1499.000000         29.000000     5.000000            1.0

       EmployeeNumber  EnvironmentSatisfaction   HourlyRate  JobInvolvement  \
count     1470.000000              1470.000000  1470.000000     1470.000000
mean      1024.865306                 2.721769    65.891156        2.729932
std        602.024335                 1.093082    20.329428        0.711561
min          1.000000                 1.000000    30.000000        1.000000
25%        491.250000                 2.000000    48.000000        2.000000
50%       1020.500000                 3.000000    66.000000        3.000000
75%       1555.750000                 4.000000    83.750000        3.000000
max       2068.000000                 4.000000   100.000000        4.000000

          JobLevel  …  RelationshipSatisfaction  StandardHours  \
count  1470.000000  …               1470.000000         1470.0
mean      2.063946  …                  2.712245           80.0
std       1.106940  …                  1.081209            0.0
min       1.000000  …                  1.000000           80.0
25%       1.000000  …                  2.000000           80.0
50%       2.000000  …                  3.000000           80.0
75%       3.000000  …                  4.000000           80.0
max       5.000000  …                  4.000000           80.0

       StockOptionLevel  TotalWorkingYears  TrainingTimesLastYear  \
```

```
count       1470.000000        1470.000000            1470.000000
mean           0.793878          11.279592               2.799320
std            0.852077           7.780782               1.289271
min            0.000000           0.000000               0.000000
25%            0.000000           6.000000               2.000000
50%            1.000000          10.000000               3.000000
75%            1.000000          15.000000               3.000000
max            3.000000          40.000000               6.000000

       WorkLifeBalance  YearsAtCompany  YearsInCurrentRole  \
count      1470.000000     1470.000000         1470.000000
mean          2.761224        7.008163            4.229252
std           0.706476        6.126525            3.623137
min           1.000000        0.000000            0.000000
25%           2.000000        3.000000            2.000000
50%           3.000000        5.000000            3.000000
75%           3.000000        9.000000            7.000000
max           4.000000       40.000000           18.000000

       YearsSinceLastPromotion  YearsWithCurrManager
count              1470.000000           1470.000000
mean                  2.187755              4.123129
std                   3.222430              3.568136
min                   0.000000              0.000000
25%                   0.000000              2.000000
50%                   1.000000              3.000000
75%                   3.000000              7.000000
max                  15.000000             17.000000

[8 rows x 26 columns]
```

1 Organization Age Distribution

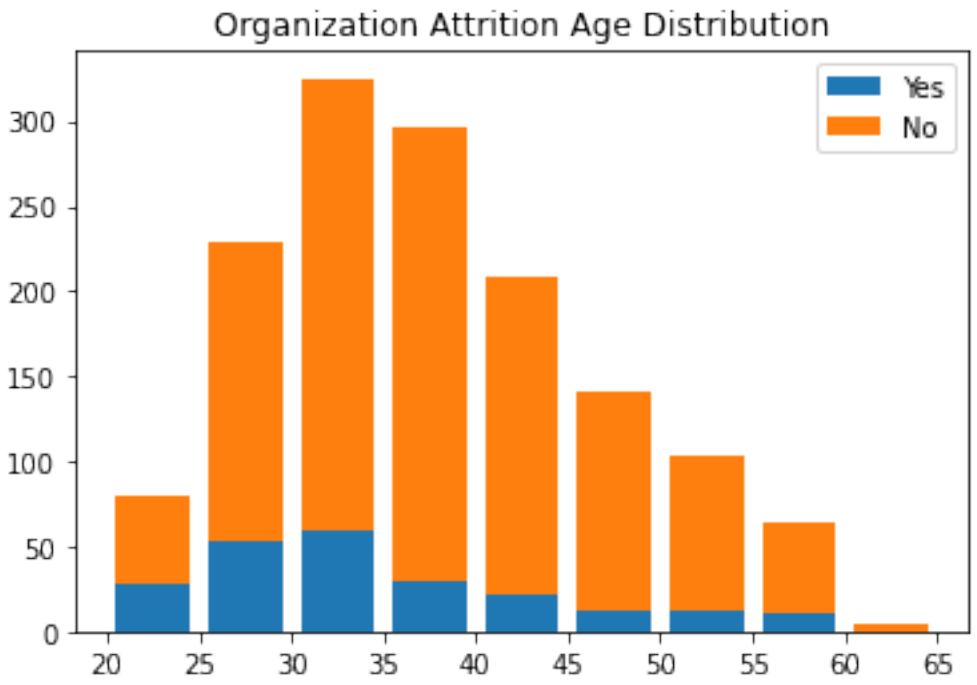
[58]:
```
plt.figure(figsize=(12, 6))
plt.title('Organization Age Distribution')
plt.xlabel('Age')
plt.ylabel('Number of Employees')
bins = np.arange(20,70,5)
plt.hist(df.Age, bins=bins, label='Attrition', rwidth=0.8)

plt.xticks(bins)
plt.show()
```



1.0.1 Observation:

- It can be found that the organisation has more employee in the age range 30-40

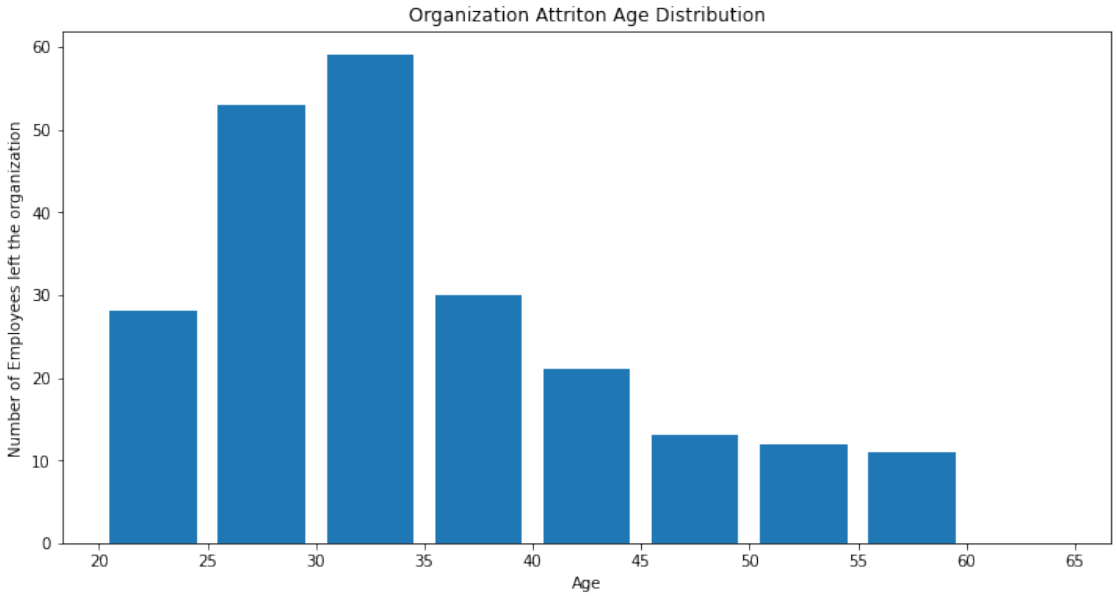
[63]:
```
bins = np.arange(20,70,5)
plt.title('Organization Attrition Age Distribution')
plt.hist([data_attr_yes.Age, data_attr_no.Age], bins=bins,stacked=True, ␣
↪rwidth=0.8)
plt.xticks(bins)
plt.legend(['Yes', 'No'])
plt.show()
```



1.0.2 Observation:

- It can be seen that nearly half of employee of age 20-25 attried

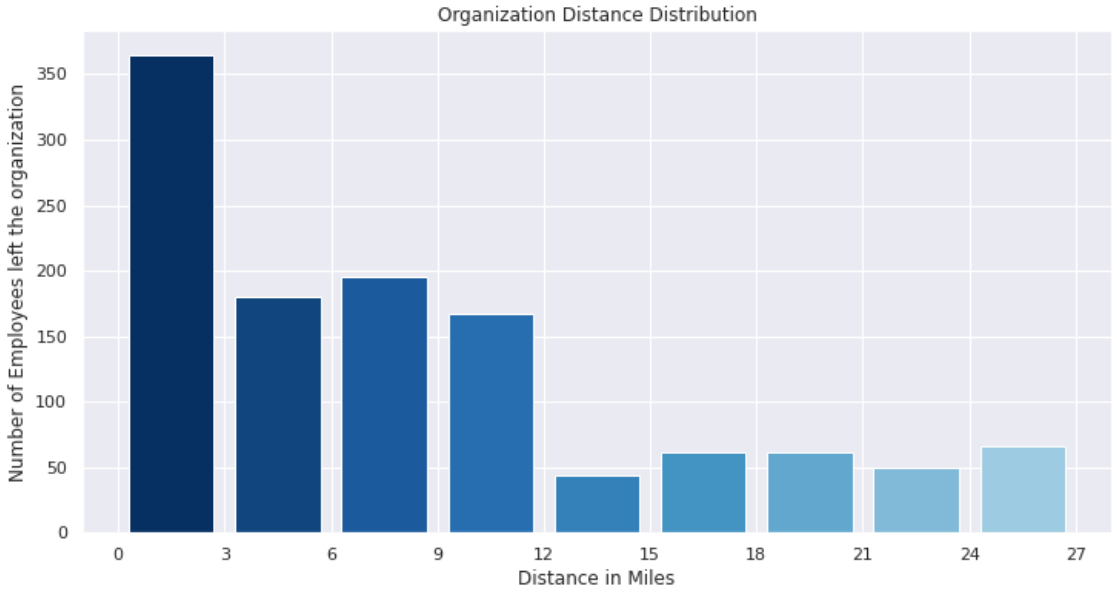
[64]:
```
bins = np.arange(20,70,5)
plt.figure(figsize=(12, 6))
plt.title('Organization Attriton Age Distribution')
plt.xlabel('Age')
plt.ylabel('Number of Employees left the organization')
plt.hist(data_attr_yes.Age, bins=bins, label='Attrition', rwidth=0.8)
plt.xticks(bins)
plt.show()
```



1.0.3 Observation:

- From the above graph, it can be analyzed that more employee have attried b/w the age of 25-25

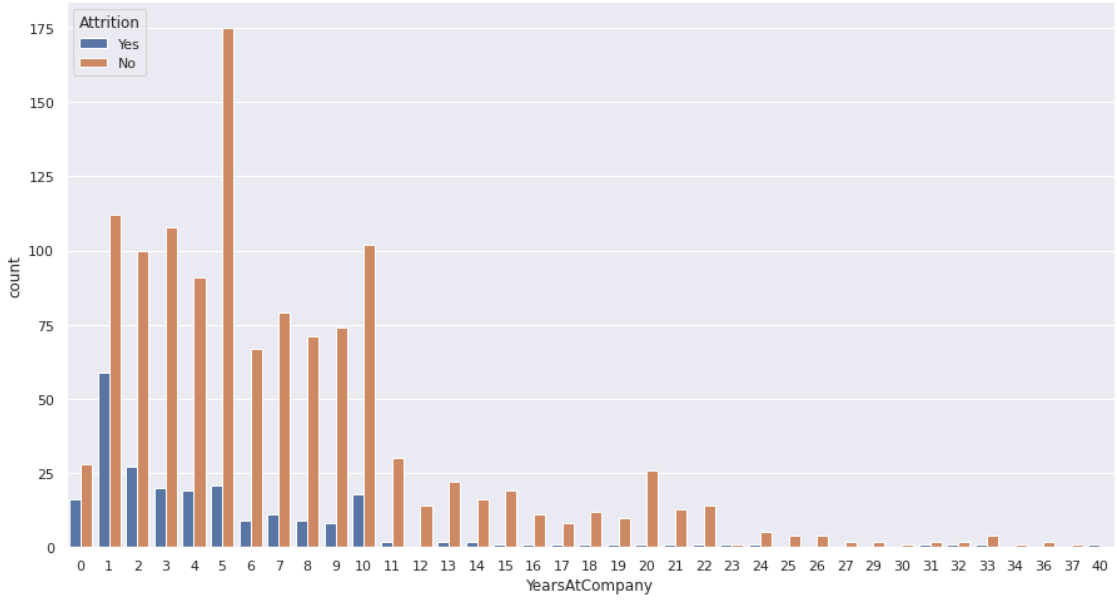
[78]:
```
cm = plt.cm.RdBu_r
bins = np.arange(0,30,3)
plt.figure(figsize=(12, 6))
plt.title('Organization Distance Distribution')
plt.xlabel('Distance in Miles')
plt.ylabel('Number of Employees left the organization')
# bins = np.arange(20,70,5)
n, bins, patches = plt.hist([data_attr_no.DistanceFromHome],␣
 ↪bins=bins,stacked=True, rwidth=0.8)
plt.xticks(bins)
# plt.legend(['Yes', 'No'])
for i, p in enumerate(patches):
    plt.setp(p, 'facecolor', cm(i/25)) # notice the i/25
plt.show()
```



1.0.4 Observation:

- It can be found that employee with 0-3 miles of the office attried more

[75]:
```
sns.set(rc={'figure.figsize':(15,8)})
sns.countplot(x="YearsAtCompany", hue="Attrition", data=df);
```



1.0.5 Observation:

- From the above graph, we see that employees that have worked in the company for 1-2 years attried more than others

[81]:
```
sns.set(rc={'figure.figsize':(10,6)})
sns.scatterplot(data=df, x="YearsAtCompany", y="MonthlyIncome", hue="Attrition")
```

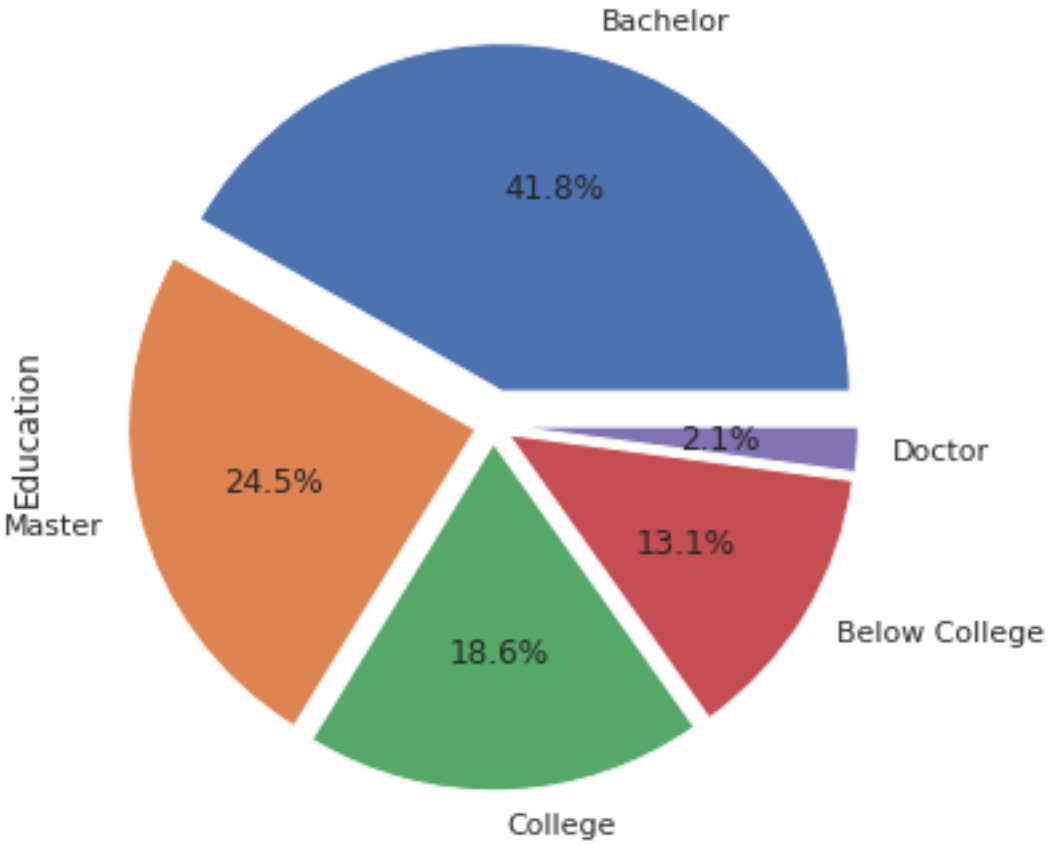
[81]:
```
<AxesSubplot:xlabel='YearsAtCompany', ylabel='MonthlyIncome'>
```

1.0.6 Observation:

- It can be seen from the above graph that there is a huge cluster of employee from 0-5 yrs at company and monthly income less than 7500.
- And from this huge cluster many employess have attried

[89]:
```
di = {1: "Below College", 2: "College", 3: "Bachelor", 4: "Master", 5: "Doctor"}
explode = (0.1, 0.05, 0.05, 0.05,0.05)
education = data_attr_yes.replace({"Education": di})
education['Education'].value_counts().plot(kind='pie', explode=explode,␣
↪autopct='%.1f%%')
```

[89]:
```
<AxesSubplot:ylabel='Education'>
```

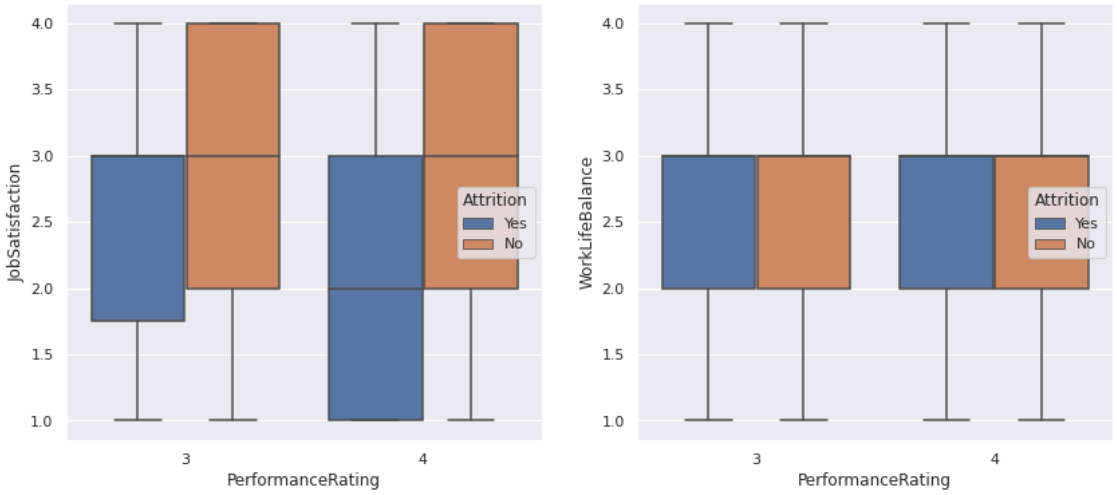


1.0.7 Observation:

- From the above graph, it can be found that employees with an education of Bachelors attried the most

[110]:
```
f, axs = plt.subplots(1, 2, figsize=(14,6))
sns.boxplot(y="JobSatisfaction", x="PerformanceRating", hue="Attrition",␣
↪data=df, ax=axs[0])

sns.boxplot(y="WorkLifeBalance", x="PerformanceRating", hue="Attrition",␣
↪data=df, ax=axs[1])
```

[110]:
```
<AxesSubplot:xlabel='PerformanceRating', ylabel='WorkLifeBalance'>
```

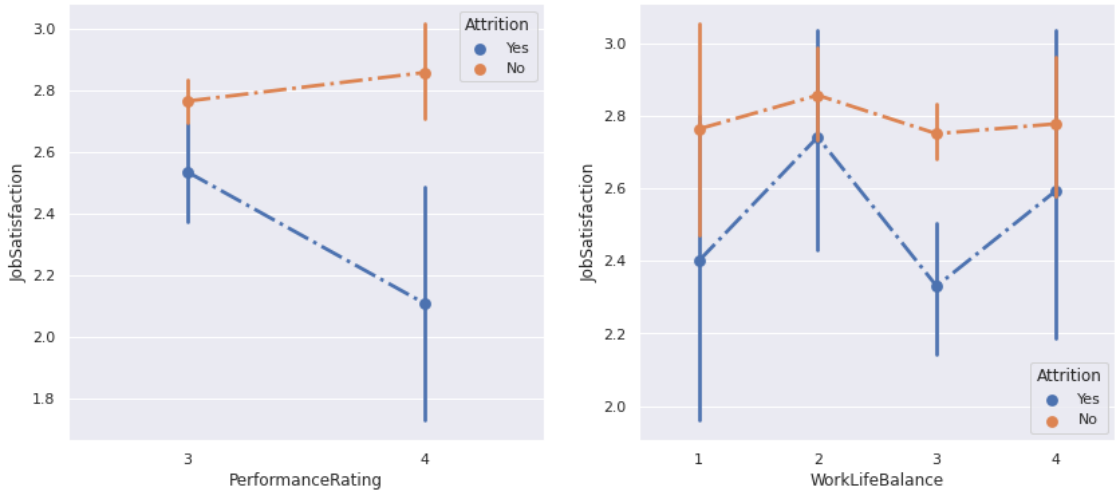


1.0.8 Observation:

- From the above box plot, we can say that jobSatisfaction is more corelated to attrition as compared to WorkLifeBalance

[118]:
```
f, axs = plt.subplots(1, 2, figsize=(14,6))
sns.pointplot(y="JobSatisfaction", x="PerformanceRating", hue="Attrition",␣
 ↪linestyles = '-.', data=df,
              ax=axs[0])

sns.pointplot(y="JobSatisfaction", x="WorkLifeBalance", hue="Attrition",␣
 ↪linestyles = '-.', data=df,
              ax=axs[1])
```

[118]:
```
<AxesSubplot:xlabel='WorkLifeBalance', ylabel='JobSatisfaction'>
```



1.0.9 Observation:

- From the above graph, we see as performance increase, the job satisfaction remains less for employees who attried.
- Same thing can be seen for work life balance against job satisfaction too

[143]:
```
df_new = df
# df_new['code'] = pd.factorize(df_new['Attrition'], sort=True)[0] + 1
# df_new.insert(0, 'Attrition_int', df_new['code'])
corr = df_new.corr()
corr.style.background_gradient(cmap='coolwarm', axis=None)
```

[143]:
```
<pandas.io.formats.style.Styler at 0x7fdcc5860860>
```

1.0.10 Observation:

- From the above table, we conclude the following order to corelation with attrition in increasing order:

1. Total Working Years
2. Job Level
3. Years in current Role