In [4]:
```
import pandas as pd
```

In [5]:
```
import seaborn as sns
```

In [6]:
```
import numpy as np
```

In [7]:
```
import matplotlib.pyplot as plt
```

In [8]:
```
df = pd.read_csv("C:\\Users\\wwwsa\\Downloads\\sel. Projects\\Attrition data.
```

In [9]:
```
df
```

Out[9]:

	EmployeeID	Age	Attrition	BusinessTravel	Department	DistanceFromHome	Education
0	1	51	No	Travel_Rarely	Sales	6	2
1	2	31	Yes	Travel_Frequently	Research & Development	10	1
2	3	32	No	Travel_Frequently	Research & Development	17	4
3	4	38	No	Non-Travel	Research & Development	2	5
4	5	32	No	Travel_Rarely	Research & Development	10	1
...	...	...	...	...	...	...	...
4405	4406	42	No	Travel_Rarely	Research & Development	5	4
4406	4407	29	No	Travel_Rarely	Research & Development	2	4
4407	4408	25	No	Travel_Rarely	Research & Development	25	2
4408	4409	42	No	Travel_Rarely	Sales	18	2
4409	4410	40	No	Travel_Rarely	Research & Development	28	3

4410 rows × 29 columns

In [10]:
```
df.shape
```

Out[10]:
```
(4410, 29)
```

```
In [11]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4410 entries, 0 to 4409
Data columns (total 29 columns):
 #   Column                   Non-Null Count  Dtype  
---  ------                   --------------  -----  
 0   EmployeeID               4410 non-null   int64  
 1   Age                      4410 non-null   int64  
 2   Attrition                4410 non-null   object 
 3   BusinessTravel           4410 non-null   object 
 4   Department               4410 non-null   object 
 5   DistanceFromHome         4410 non-null   int64  
 6   Education                4410 non-null   int64  
 7   EducationField           4410 non-null   object 
 8   EmployeeCount            4410 non-null   int64  
 9   Gender                   4410 non-null   object 
 10  JobLevel                 4410 non-null   int64  
 11  JobRole                  4410 non-null   object 
 12  MaritalStatus            4410 non-null   object 
 13  MonthlyIncome            4410 non-null   int64  
 14  NumCompaniesWorked       4391 non-null   float64
 15  Over18                   4410 non-null   object 
 16  PercentSalaryHike        4410 non-null   int64  
 17  StandardHours            4410 non-null   int64  
 18  StockOptionLevel         4410 non-null   int64  
 19  TotalWorkingYears        4401 non-null   float64
 20  TrainingTimesLastYear    4410 non-null   int64  
 21  YearsAtCompany           4410 non-null   int64  
 22  YearsSinceLastPromotion  4410 non-null   int64  
 23  YearsWithCurrManager     4410 non-null   int64  
 24  EnvironmentSatisfaction  4385 non-null   float64
 25  JobSatisfaction          4390 non-null   float64
 26  WorkLifeBalance          4372 non-null   float64
 27  JobInvolvement           4410 non-null   int64  
 28  PerformanceRating        4410 non-null   int64  
dtypes: float64(5), int64(16), object(8)
memory usage: 999.3+ KB
```

In [12]:
```
df.describe()
```

Out[12]:

	EmployeeID	Age	DistanceFromHome	Education	EmployeeCount	JobLevel
count	4410.000000	4410.000000	4410.000000	4410.000000	4410.0	4410.000000
mean	2205.500000	36.923810	9.192517	2.912925	1.0	2.063946
std	1273.201673	9.133301	8.105026	1.023933	0.0	1.106689
min	1.000000	18.000000	1.000000	1.000000	1.0	1.000000
25%	1103.250000	30.000000	2.000000	2.000000	1.0	1.000000
50%	2205.500000	36.000000	7.000000	3.000000	1.0	2.000000
75%	3307.750000	43.000000	14.000000	4.000000	1.0	3.000000
max	4410.000000	60.000000	29.000000	5.000000	1.0	5.000000

8 rows × 21 columns

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

Out[13]:
```
EmployeeID                  0
Age                         0
Attrition                   0
BusinessTravel              0
Department                  0
DistanceFromHome            0
Education                   0
EducationField              0
EmployeeCount               0
Gender                      0
JobLevel                    0
JobRole                     0
MaritalStatus               0
MonthlyIncome               0
NumCompaniesWorked         19
Over18                      0
PercentSalaryHike           0
StandardHours               0
StockOptionLevel            0
TotalWorkingYears           9
TrainingTimesLastYear       0
YearsAtCompany              0
YearsSinceLastPromotion     0
YearsWithCurrManager        0
EnvironmentSatisfaction    25
JobSatisfaction            20
WorkLifeBalance            38
JobInvolvement              0
PerformanceRating           0
dtype: int64
```

In [15]:
```
df.isnull()
```

Out[15]:

	EmployeeID	Age	Attrition	BusinessTravel	Department	DistanceFromHome	Education
0	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...
4405	False	False	False	False	False	False	False
4406	False	False	False	False	False	False	False
4407	False	False	False	False	False	False	False
4408	False	False	False	False	False	False	False
4409	False	False	False	False	False	False	False

4410 rows × 29 columns

In [17]:
```
#check unique values in the column
print(df.apply(lambda col: col.unique().sum()))
```

```
EmployeeID                                                           9726255
Age                                                                     1677
Attrition                                                              NoYes
BusinessTravel                      Travel_RarelyTravel_FrequentlyNon-Travel
Department                        SalesResearch & DevelopmentHuman Resources
DistanceFromHome                                                         435
Education                                                                 15
EducationField             Life SciencesOtherMedicalMarketingTechnical De...
EmployeeCount                                                              1
Gender                                                            FemaleMale
JobLevel                                                                  15
JobRole                    Healthcare RepresentativeResearch ScientistSal...
MaritalStatus                                          MarriedSingleDivorced
MonthlyIncome                                                       90578130
NumCompaniesWorked                                                       NaN
Over18                                                                     Y
PercentSalaryHike                                                        270
StandardHours                                                              8
StockOptionLevel                                                           6
TotalWorkingYears                                                        NaN
TrainingTimesLastYear                                                     21
YearsAtCompany                                                           680
YearsSinceLastPromotion                                                  120
YearsWithCurrManager                                                     153
EnvironmentSatisfaction                                                  NaN
JobSatisfaction                                                          NaN
WorkLifeBalance                                                          NaN
JobInvolvement                                                            10
PerformanceRating                                                          7
dtype: object
```

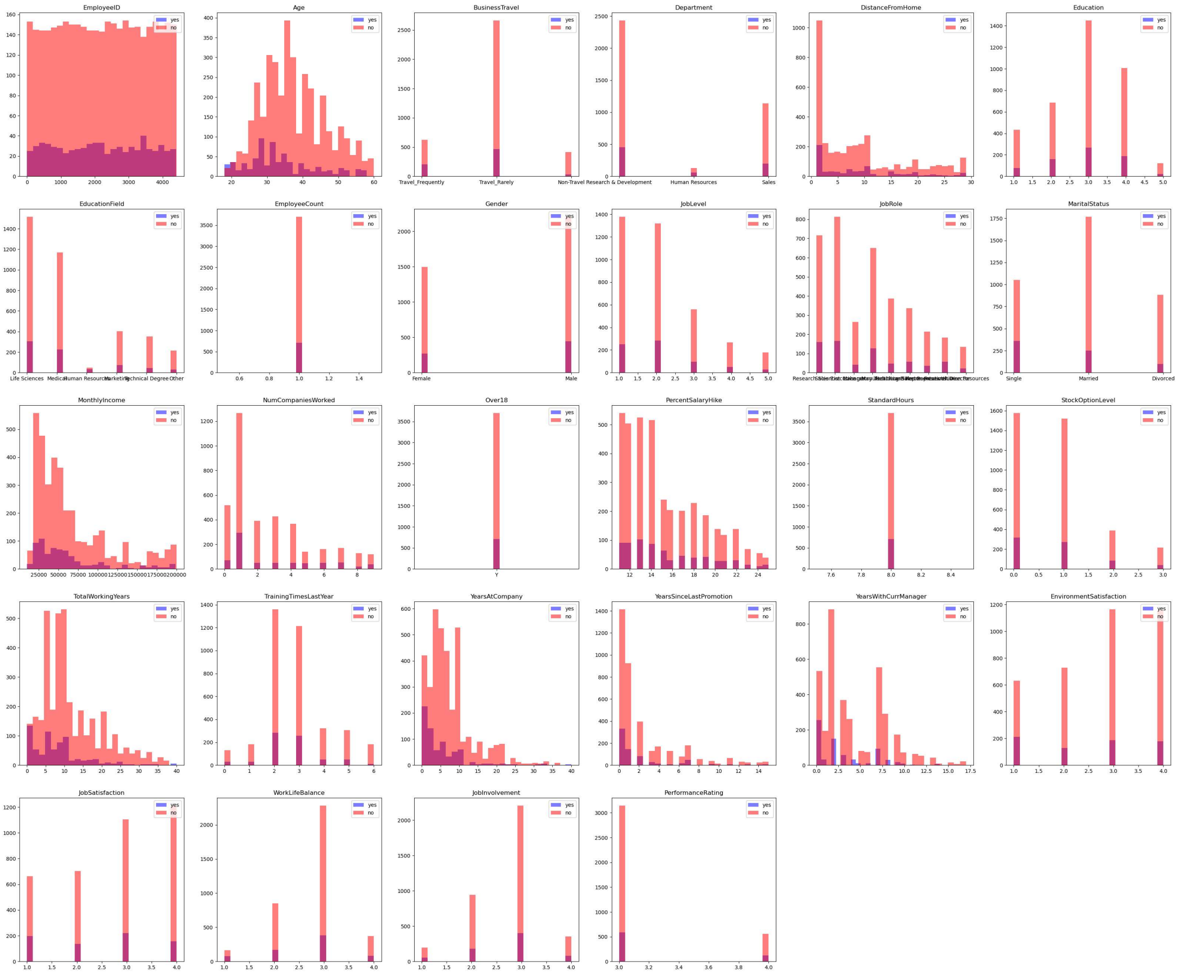
In [20]:

```
cat_df=df.select_dtypes(include='object')

for i in cat_df:
    plt.figure(figsize=(15, 15))
    sns.catplot(data=df,x=i,kind='count')
```

```
C:\ProgramData\anaconda3\Lib\site-packages\seaborn\axisgrid.py:118: UserW
arning: The figure layout has changed to tight
  self._figure.tight_layout(*args, **kwargs)
C:\ProgramData\anaconda3\Lib\site-packages\seaborn\axisgrid.py:118: UserW
arning: The figure layout has changed to tight
  self._figure.tight_layout(*args, **kwargs)
C:\ProgramData\anaconda3\Lib\site-packages\seaborn\axisgrid.py:118: UserW
arning: The figure layout has changed to tight
  self._figure.tight_layout(*args, **kwargs)
C:\ProgramData\anaconda3\Lib\site-packages\seaborn\axisgrid.py:118: UserW
arning: The figure layout has changed to tight
  self._figure.tight_layout(*args, **kwargs)
C:\ProgramData\anaconda3\Lib\site-packages\seaborn\axisgrid.py:118: UserW
arning: The figure layout has changed to tight
  self._figure.tight_layout(*args, **kwargs)
C:\ProgramData\anaconda3\Lib\site-packages\seaborn\axisgrid.py:118: UserW
arning: The figure layout has changed to tight
  self._figure.tight_layout(*args, **kwargs)
C:\ProgramData\anaconda3\Lib\site-packages\seaborn\axisgrid.py:118: UserW
```

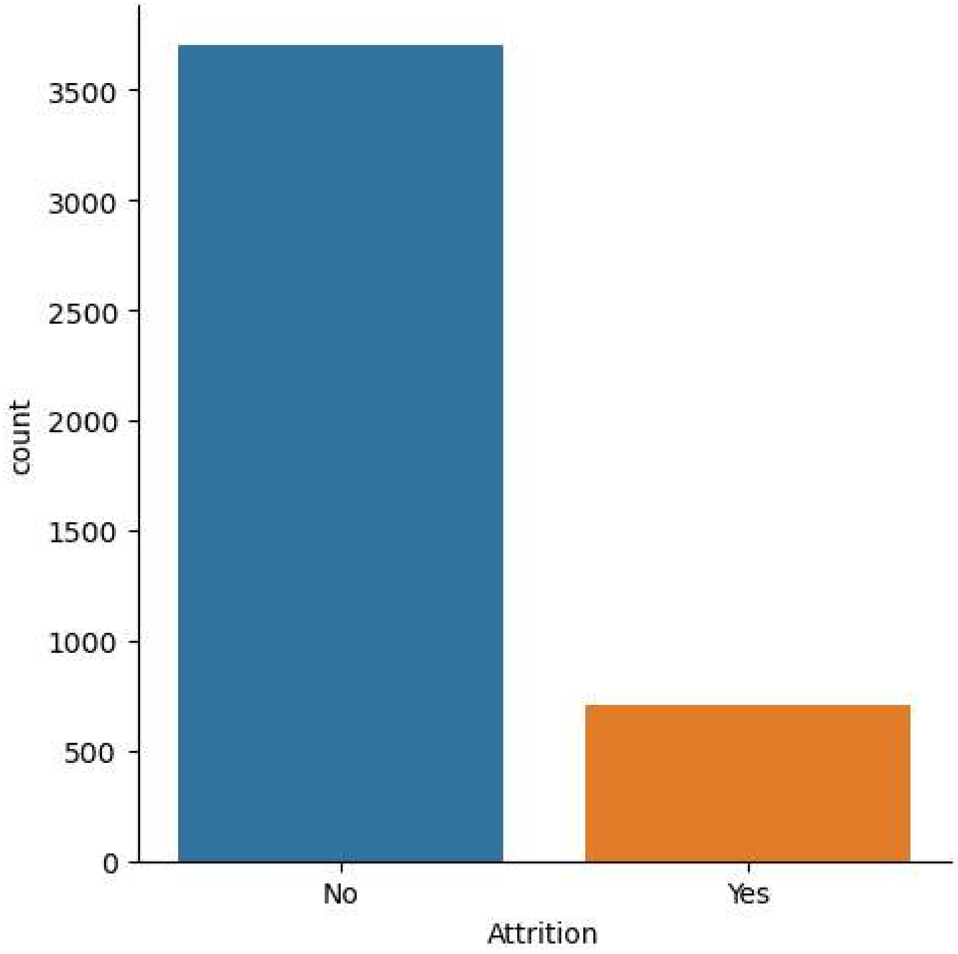
```
In [19]: #plot distributions
k=1
plt.figure(figsize=(40, 40))
for col in df:
  if col=="Attrition":
    continue
  yes = df[df['Attrition'] == 'Yes'][col]
  no = df[df['Attrition'] == 'No'][col]
  plt.subplot(6, 6, k)
  plt.hist(yes, bins=25, alpha=0.5, label='yes', color='b')
  plt.hist(no, bins=25, alpha=0.5, label='no', color='r')
  plt.legend(loc='upper right')
  plt.title(col)
  k+=1
```



In [20]:
```
sns.catplot(data=df,x="Attrition",kind='count')
```

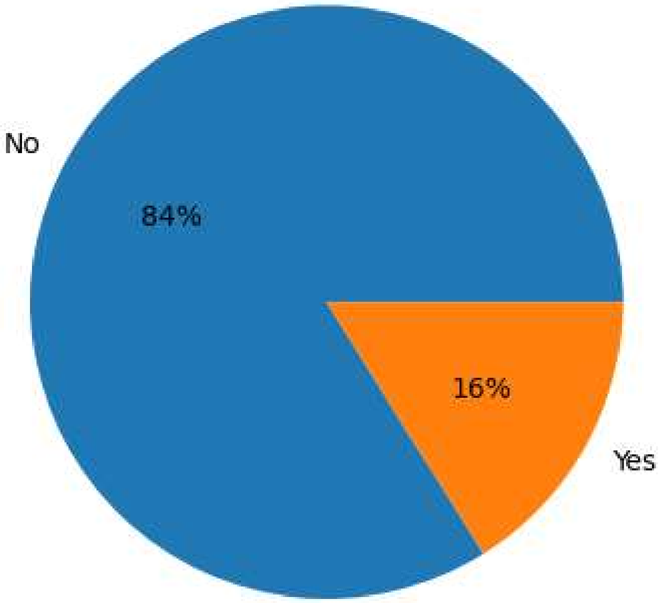
```
C:\ProgramData\anaconda3\Lib\site-packages\seaborn\axisgrid.py:118: UserWarn
ing: The figure layout has changed to tight
  self._figure.tight_layout(*args, **kwargs)
```

Out[20]:
```
<seaborn.axisgrid.FacetGrid at 0x1eb3e239f10>
```



In [21]:

```
# colors = sns.color_palette("husl", 2)
plt.pie(df['Attrition'].value_counts(),labels=['No','Yes'],autopct='%.0f%%')
plt.show()
```

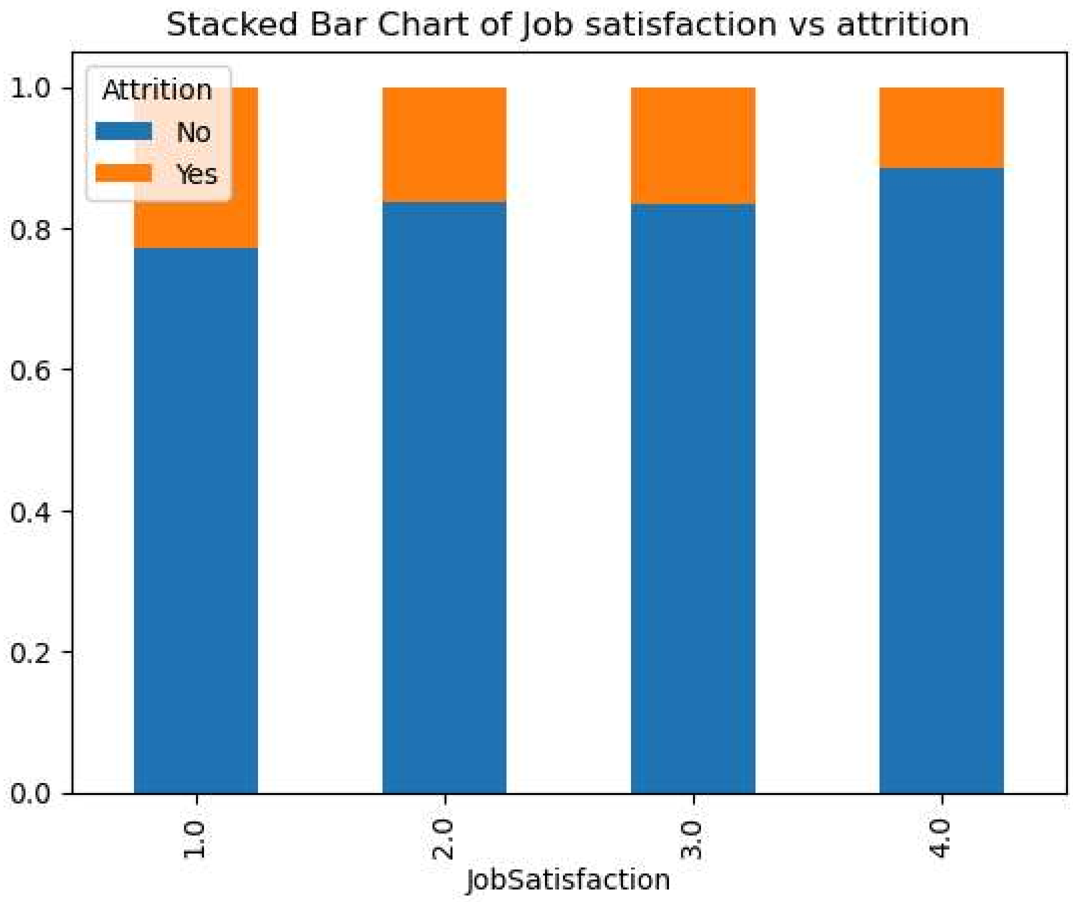


In [22]:

```
table=pd.crosstab(df.JobSatisfaction, df.Attrition)
table.div(table.sum(1).astype(float), axis=0).plot(kind='bar', stacked=True)
plt.title('Stacked Bar Chart of Job satisfaction vs attrition')
```

Out[22]:

```
Text(0.5, 1.0, 'Stacked Bar Chart of Job satisfaction vs attrition')
```



In [23]:

```
table=pd.crosstab(df.OverTime, df.Attrition)
table.div(table.sum(1).astype(float), axis=0).plot(kind='bar', stacked=True)
plt.title('Stacked Bar Chart of Overtime vs attrition')
```

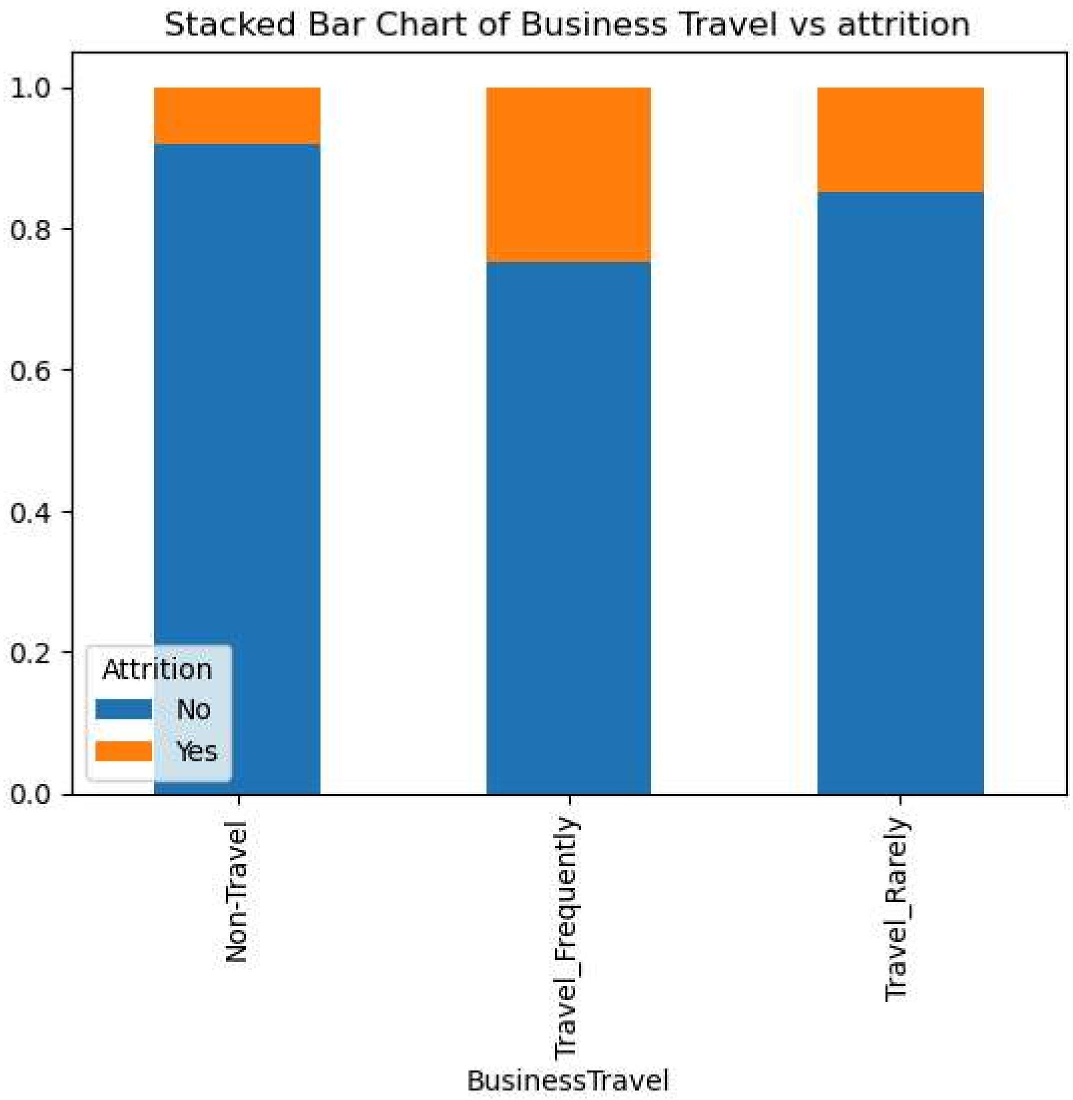
```
---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
~\AppData\Local\Temp\ipykernel_4492\1502457776.py in ?()
----> 1 table=pd.crosstab(df.OverTime, df.Attrition)
      2 table.div(table.sum(1).astype(float), axis=0).plot(kind='bar', stack
ed=True)
      3 plt.title('Stacked Bar Chart of Overtime vs attrition')

C:\ProgramData\anaconda3\Lib\site-packages\pandas\core\generic.py in ?(self,
name)
   5985             and name not in self._accessors
   5986             and self._info_axis._can_hold_identifiers_and_holds_name
(name)
   5987         ):
   5988             return self[name]
-> 5989         return object.__getattribute__(self, name)

AttributeError: 'DataFrame' object has no attribute 'OverTime'
```

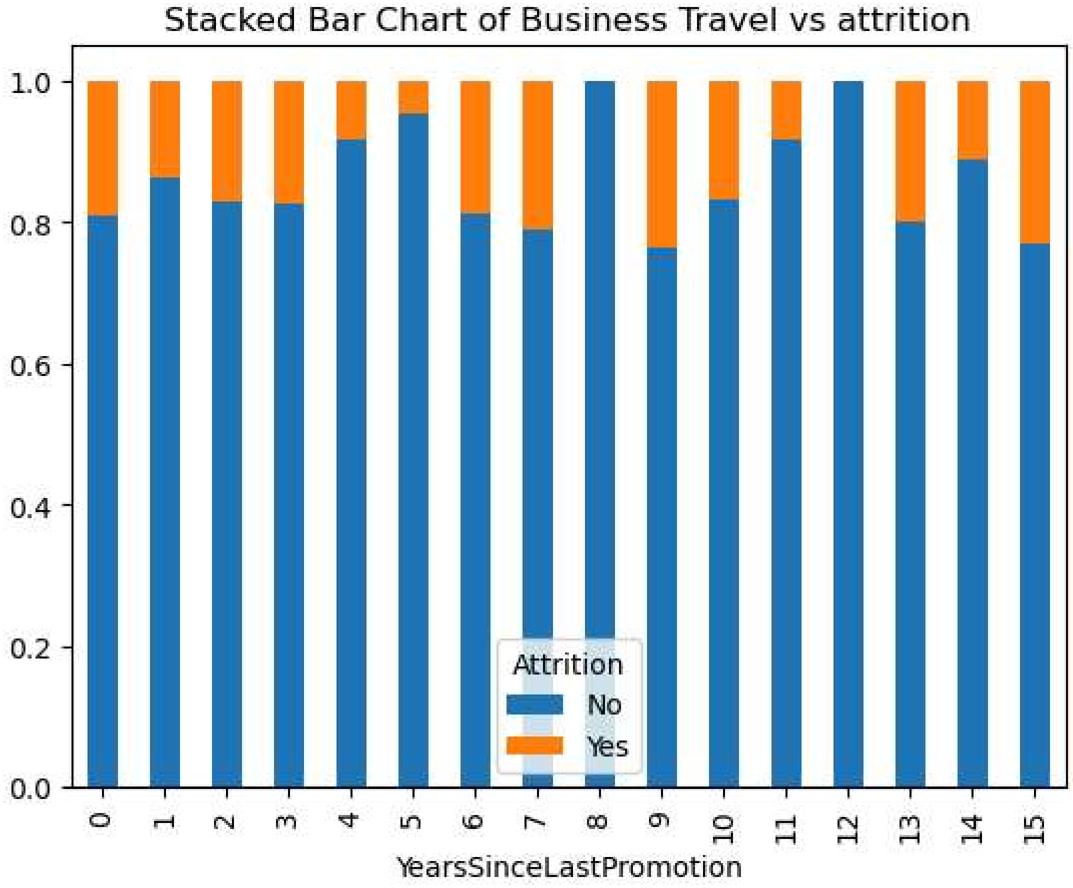
In [24]:
```
table=pd.crosstab(df.BusinessTravel, df.Attrition)
table.div(table.sum(1).astype(float), axis=0).plot(kind='bar', stacked=True)
plt.title('Stacked Bar Chart of Business Travel vs attrition')
```

Out[24]:
```
Text(0.5, 1.0, 'Stacked Bar Chart of Business Travel vs attrition')
```



In [25]:
```
table=pd.crosstab(df.YearsSinceLastPromotion, df.Attrition)
table.div(table.sum(1).astype(float), axis=0).plot(kind='bar', stacked=True)
plt.title('Stacked Bar Chart of Business Travel vs attrition')
```

Out[25]:
```
Text(0.5, 1.0, 'Stacked Bar Chart of Business Travel vs attrition')
```

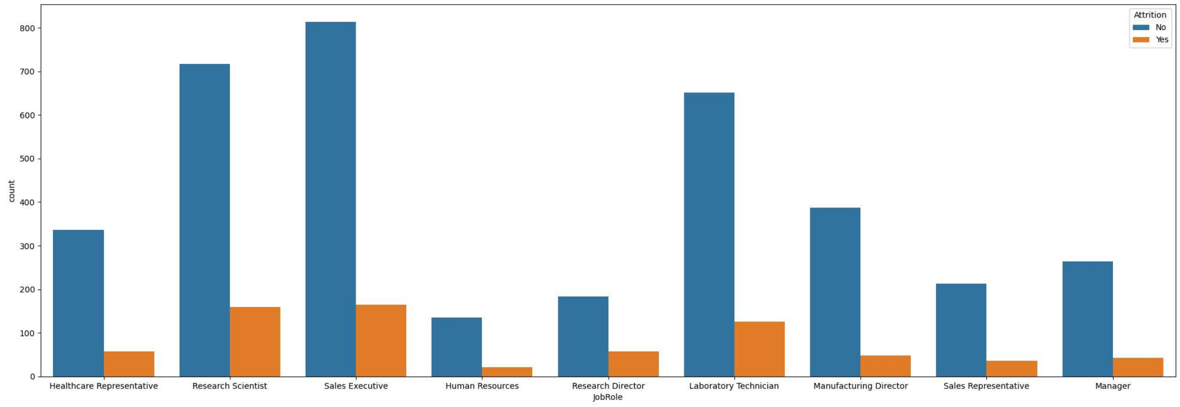


In [ ]:
```

```

In [27]:
```
a4_dims = (25, 8.27)
fig, ax = plt.subplots(figsize=a4_dims)
sns.countplot(data=df,x="JobRole",hue="Attrition", ax=ax )
```

Out[27]:
```
<Axes: xlabel='JobRole', ylabel='count'>
```



```
In [ ]:
```