# Week 6.2 Assignment


- Joshua Burden
- Bellevue University
- DSC550 Data Mining
- Dr. Brett Werner
- 10/09/2022


## Begin Milestone 1 with a 250-500-word narrative describing your original idea for the analysis/model building business problem.

## Clearly identify the problem you will address and the target for your model.

### Background:

During the pandemic an increase in the need for health care professionals was required. The dataset collected is a modified synthetic dataset from IBM's Watson to show a useful insight into the attrition rate for healthcare workers.

### Problem:

The data set includes information about the attrition rate for employees within the healthcare field. The meaning of employee attrition is the departure of employees from the organization for any reason whether that be voluntary or involuntary, including resignation, termination, death, or retirement. Companies to avoid attrition rates being too high is to replace those who are either leaving voluntarily or involuntary. The data set should provide insights into whether a company in the healthcare field was replacing their employees that were leaving the field, or if they continued to have a gradual but deliberate reduction in staff for any reason.

### Original Idea:

The idea behind this data set is to discover whether certain roles within the healthcare industry, hours worked, age of an employee, or any other qualifying data points stand out as to why the healthcare industry had any determining factor on whether a person was to leave their field, while also predicting whether the employee was eventually replaced.

### Dataset:

This dataset contains employee and company data useful for supervised ML, unsupervised ML, and analytics. Attrition - whether an employee left or not - is included and can be used as the

target variable. The data is synthetic and based on the IBM Watson dataset for attrition. Employee roles and departments were changed to reflect the healthcare domain. Also, known outcomes for some employees were changed to help increase the performance of ML models

## Then, do a graphical analysis creating a minimum of four graphs.

## Label your graphs appropriately and explain/analyze the information provided by each graph.

In [ ]:
```
import pandas as pd
import seaborn as sns
from matplotlib import pyplot as plt
import numpy as np
import plotly.express as px
```

In [ ]:
```
data_df = pd.read_csv('./DATA/watson_healthcare_modified.csv')
data_df.head()
```

Out[ ]:

| | EmployeeID | Age | Attrition | BusinessTravel | DailyRate | Department | DistanceFromHome | Educati |
|---|---|---|---|---|---|---|---|---|
| **0** | 1313919 | 41 | No | Travel_Rarely | 1102 | Cardiology | 1 | |
| **1** | 1200302 | 49 | No | Travel_Frequently | 279 | Maternity | 8 | |
| **2** | 1060315 | 37 | Yes | Travel_Rarely | 1373 | Maternity | 2 | |
| **3** | 1272912 | 33 | No | Travel_Frequently | 1392 | Maternity | 3 | |
| **4** | 1414939 | 27 | No | Travel_Rarely | 591 | Maternity | 2 | |

5 rows × 35 columns

In [ ]:
```
print("Number of duplicated data: "+str(data_df.duplicated().sum()))
```

```
Number of duplicated data: 0
```

In [ ]:
```
data_df.isnull().sum()
```

Out[ ]:
```
EmployeeID                0
Age                       0
Attrition                 0
BusinessTravel            0
DailyRate                 0
Department                0
DistanceFromHome          0
Education                 0
EducationField            0
EmployeeCount             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
Shift                     0
TotalWorkingYears         0
TrainingTimesLastYear     0
WorkLifeBalance           0
YearsAtCompany            0
YearsInCurrentRole        0
YearsSinceLastPromotion   0
YearsWithCurrManager      0
dtype: int64
```

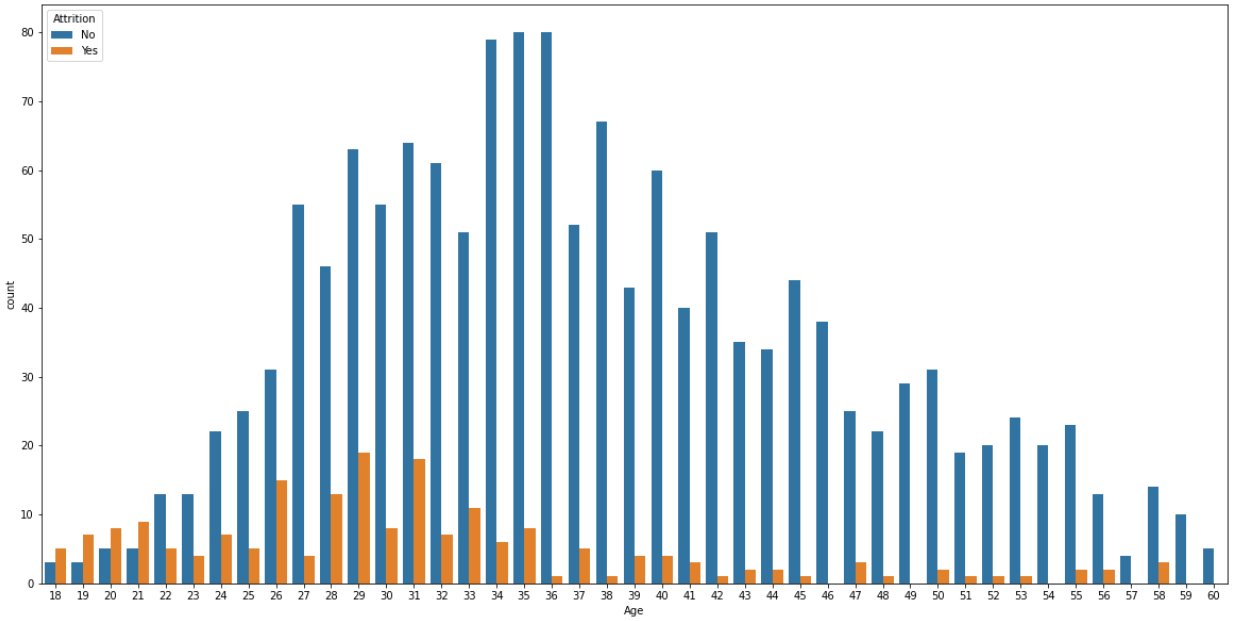
In [ ]:
```
data_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1676 entries, 0 to 1675
Data columns (total 35 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   EmployeeID                1676 non-null   int64
 1   Age                       1676 non-null   int64
 2   Attrition                 1676 non-null   object
 3   BusinessTravel            1676 non-null   object
 4   DailyRate                 1676 non-null   int64
 5   Department                1676 non-null   object
 6   DistanceFromHome          1676 non-null   int64
 7   Education                 1676 non-null   int64
 8   EducationField            1676 non-null   object
 9   EmployeeCount             1676 non-null   int64
 10  EnvironmentSatisfaction   1676 non-null   int64
 11  Gender                    1676 non-null   object
 12  HourlyRate                1676 non-null   int64
 13  JobInvolvement            1676 non-null   int64
 14  JobLevel                  1676 non-null   int64
 15  JobRole                   1676 non-null   object
 16  JobSatisfaction           1676 non-null   int64
 17  MaritalStatus             1676 non-null   object
 18  MonthlyIncome             1676 non-null   int64
 19  MonthlyRate               1676 non-null   int64
 20  NumCompaniesWorked        1676 non-null   int64
 21  Over18                    1676 non-null   object
 22  OverTime                  1676 non-null   object
 23  PercentSalaryHike         1676 non-null   int64
 24  PerformanceRating         1676 non-null   int64
 25  RelationshipSatisfaction  1676 non-null   int64
 26  StandardHours             1676 non-null   int64
 27  Shift                     1676 non-null   int64
 28  TotalWorkingYears         1676 non-null   int64
 29  TrainingTimesLastYear     1676 non-null   int64
 30  WorkLifeBalance           1676 non-null   int64
 31  YearsAtCompany            1676 non-null   int64
 32  YearsInCurrentRole        1676 non-null   int64
 33  YearsSinceLastPromotion   1676 non-null   int64
 34  YearsWithCurrManager      1676 non-null   int64
dtypes: int64(26), object(9)
memory usage: 458.4+ KB
```

## Visualization 1

In [ ]:
```
plt.figure(figsize=(20,10))
sns.countplot(x='Age',hue='Attrition',data=data_df)
```

Out[ ]:
```
<AxesSubplot:xlabel='Age', ylabel='count'>
```




## Visualization 2

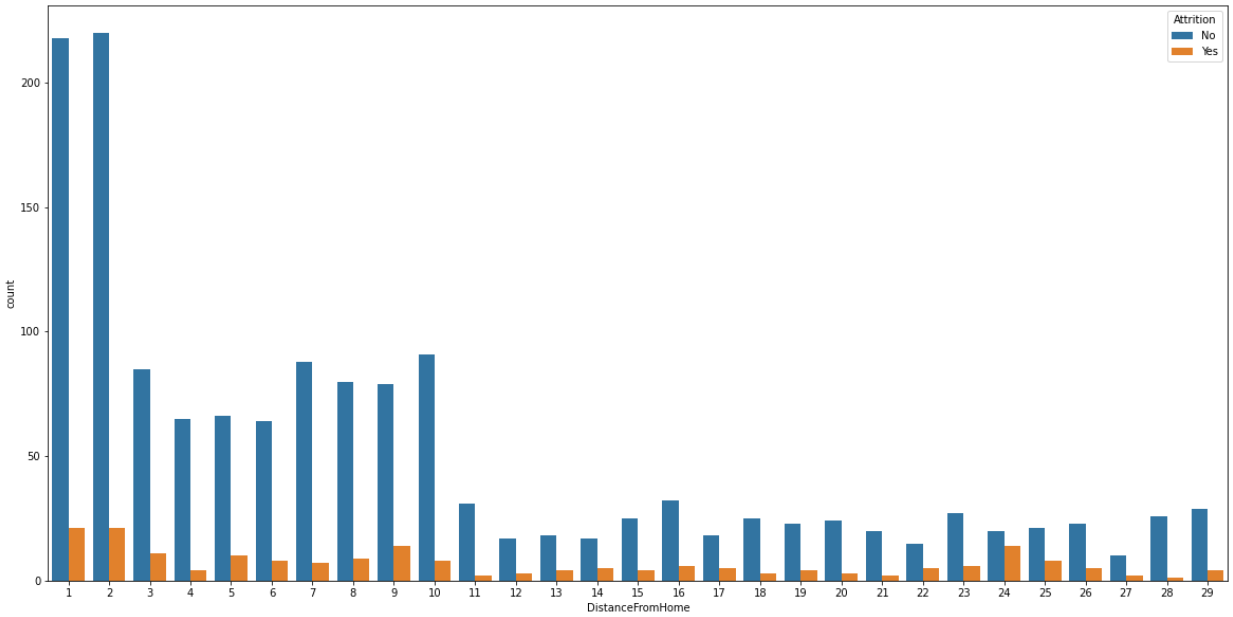
In [ ]:
```
data_df.groupby('Attrition')['MonthlyIncome'].mean().sort_values().reset_index()
```

Out[ ]:

| | Attrition | MonthlyIncome |
|---|---|---|
| **0** | Yes | 4024.246231 |
| **1** | No | 6852.301963 |

In [ ]:
```
plt.figure(figsize=(20,10))
sns.countplot(x='DistanceFromHome',hue='Attrition',data=data_df)
```

Out[ ]:
```
<AxesSubplot:xlabel='DistanceFromHome', ylabel='count'>
```




## Visualization 3

In [ ]:
```
px.histogram(data_df,x="Department",color="Attrition",barmode="group",text_auto=".2f",
             title = "Percentage of Department Type")
```

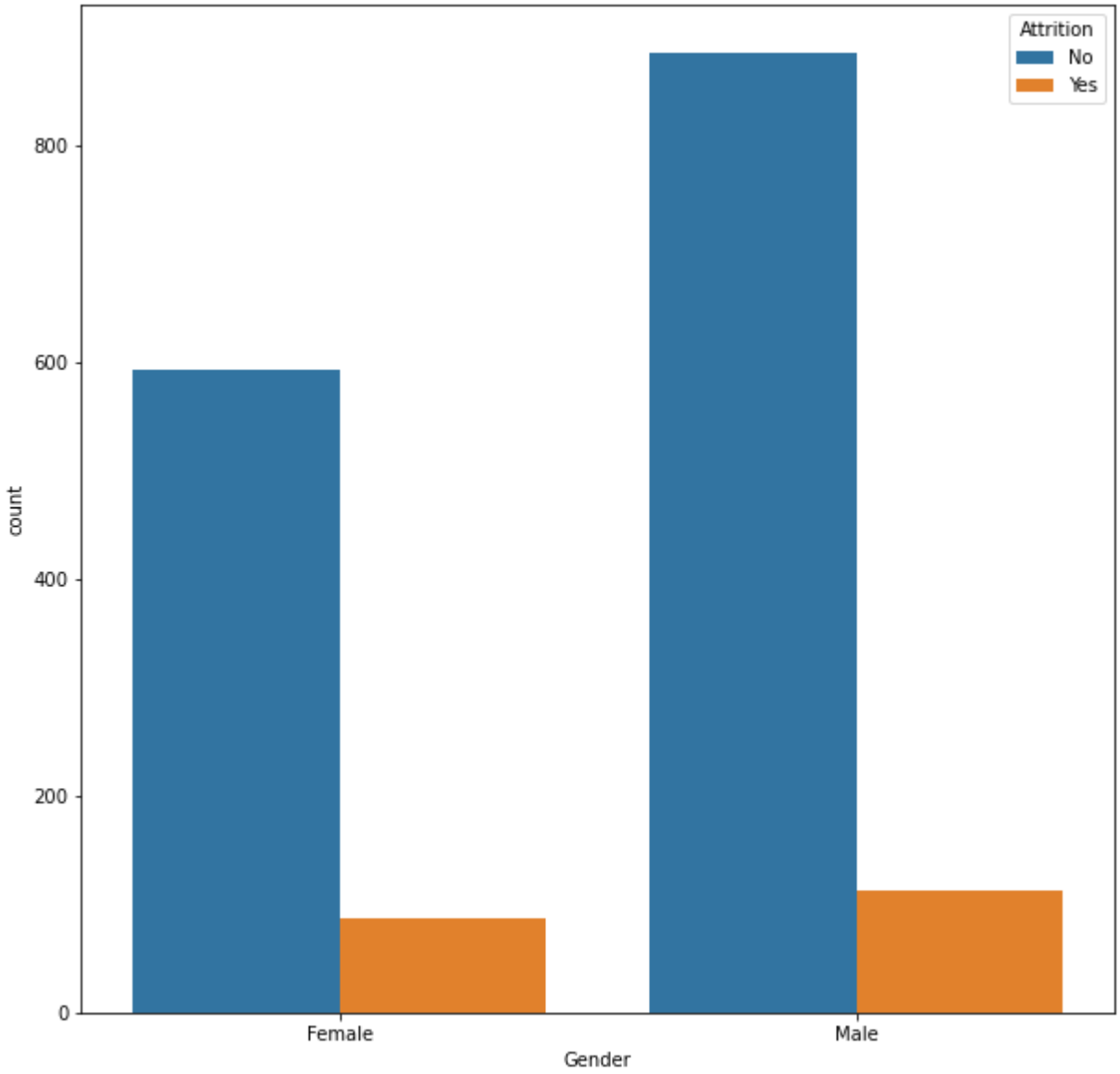
In [ ]:
```
px.histogram(data_df,x="EducationField",color="Attrition",barmode="group",text_auto=".
             title = "Percentage of EducationField Type")
```

In [ ]:
```
px.histogram(data_df,x="JobRole",color="Attrition",barmode="group",text_auto=".2f",tem
             title = "Percentage of EducationField Type")
```

## Visualization 4

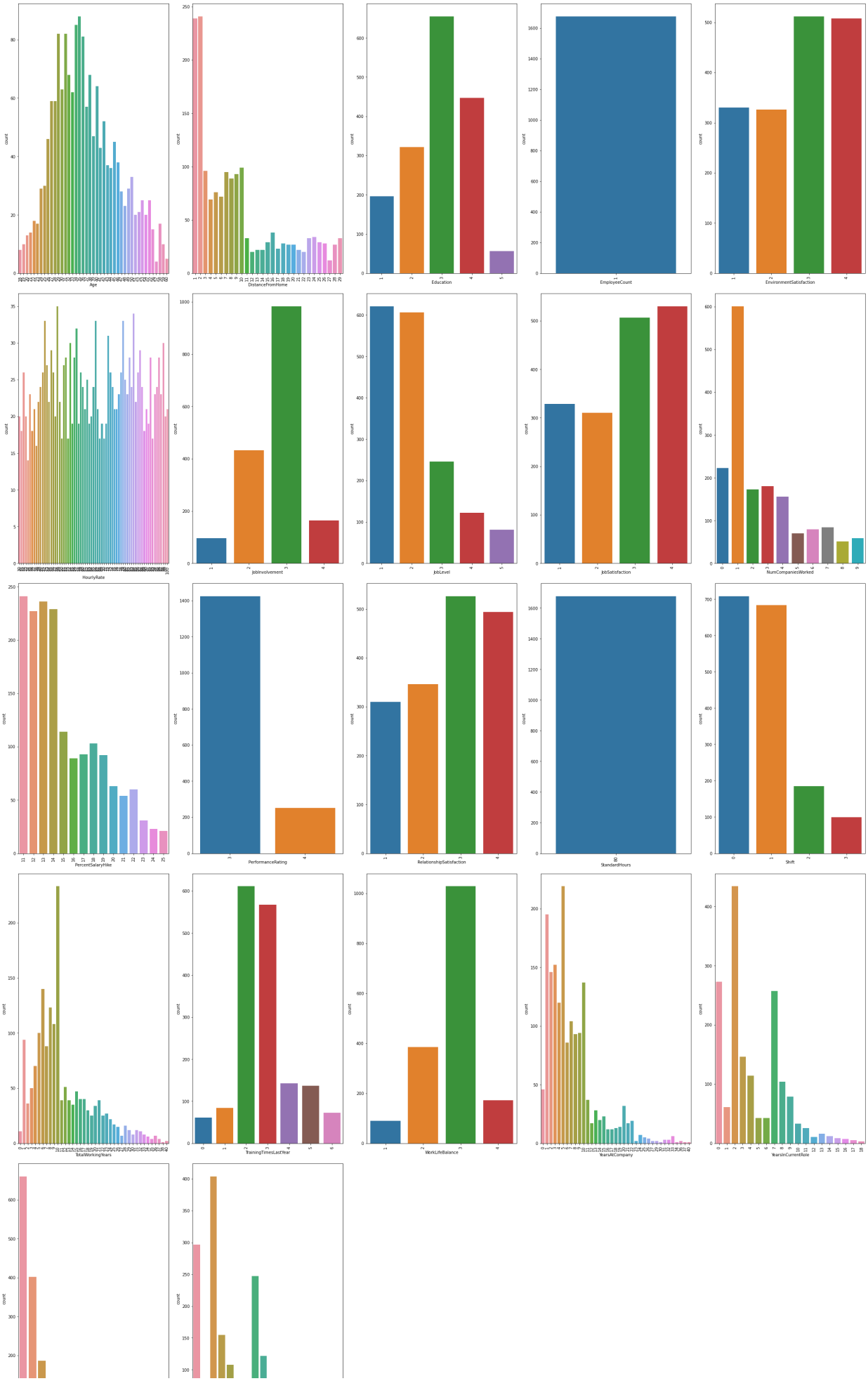
In [ ]:
```
plt.figure(figsize=(10,10))
sns.countplot(x='Gender',hue='Attrition',data=data_df)
```

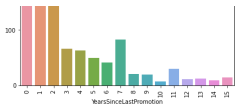
Out[ ]:
```
<AxesSubplot:xlabel='Gender', ylabel='count'>
```




## Breakdown of all the available datapoints

In [ ]:
```
plt.figure(figsize=(30,50))
for index,column in enumerate(num_col):
    plt.subplot(5,5,index+1)
    sns.countplot(data=num_col,x=column)
    plt.xticks(rotation = 90)
plt.tight_layout(pad = 1.0)
plt.show()
```


count
Age
DistanceFromHome
Education
EmployeeCount
EnvironmentSatisfaction
HourlyRate
JobInvolvement
JobLevel
JobSatisfaction
NumCompaniesWorked
PercentSalaryHike
PerformanceRating
RelationshipSatisfaction
StandardHours
Shift
TotalWorkingYears
TrainingTimesLastYear
WorkLifeBalance
YearsAtCompany
YearsInCurrentRole





## Observations:

- Maternity departments had the highest rate of attrition followed by cardiology and neurology
- attrition rates had the highest peak at 29 years old
- 26-35 years old saw the highest range of attrition
- 42 years old and older saw the least attrition rates
- More men where likely to leave than women but Men also were more accounted for than women in the healthcare field
- Human resources were the least likely to have people quit
- Life Sciences were the Education field with the highest amount of attrition
- people that lived closer to their jobs were more likely to leave