IBM EDA Peer-graded Assignment

- Brief description of the data set and a summary of its attributes
  - The original dataset contained 1470 observations, 35 features, and a target of 'Attrition'. Furthermore, the aforementioned dataset contained the following features: 'Age', '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'.
- Initial plan for data exploration
  - Check for missing data first
  - Check for duplicate data first
  - Review column names and understand what each feature means
  - Review data types
  - Review data for data cleaning and feature engineering
    - Clean column names
    - Convert categorical features (original dataset 'Attrition') to and create binary against 'Attrition'
    - Remove columns 'EmployeeCount', 'Over18', 'StandardHours', 'EmployeeNumber'
- Key Findings and Insights, which synthesizes the results of Exploratory Data Analysis in an insightful and actionable manner
  - Approximately 16.12% of the 1470 observations in the original dataset were found to have left the company and categorized as the target class 'Attrition' (Attrition Leaver) as 'Yes'.
  - Average age of the 1470 observations was 36.92 years, range 18-60.
  - Other Female, Male, Other male (0.32% observations from previous page 'Attrition' data was 39.07%.
  - Average Monthly Income was 6502.93 USD, range 1009-19999. Other Attrition side was 5608.24 and at 151% , fairly close to the dataset.
  - Average Distance from Home, Total year in the Workforce' was 9.192 .
  - Other Year, Dep Company' were 7.369 and 5.130 , other 12.169 .
  - Average Hourly Rate was 65.891 , range 30-100 , Other Attrition side was 65.481 , compared to 12.07% .
  - Others at employee Total Household Ex Workers' side was 0.406 .
  - Average Daily Rate was 802.485 , range 102-1499 , Other Attrition side was 750.258 , 6.9% .
  - Employees with 'Marital Status' single had an 'Attrition' rate of 25.53 .
  - Employees who 'Travel Rarely' had an 'Attrition' rate of 24.91% .
  - Employees who 'Working Overtime' had an 'Attrition' rate of 30.53 .
- Formulating at least 3 hypotheses about the data
  - H_0: a Measurable_Attribute is a Attrition_Not_Attrition
  - H_a: a Measurable_Attribute is a Attrition_Not_Attrition
  - H_0: Age_Attrition is a Age_Not_Attrition
  - H_a: Age_Attrition is a Age_Not_Attrition
  - H_0: Gender_Attrition is a Gender_Not_Attrition
  - H_a: Gender_Attrition is a Gender_Not_Attrition
- Conducting a hypothesis testing to determine the rest of the hypotheses and discuss the results
  - accumulatedNumOfNumericalCol_pdfAttr_dfCleanedDataNoAttrition = DATASET.CREATE(accumulatedNumOfNumericalCol)
  - print(accumulatedNumOfNumericalCol)
  - These appears to be no statistically significant relationship between Attrition and Education, thus unlikely reject the null hypothesis.
- Suggestions for next steps in analyzing the data
  - Perform Feature Engineering, Feature Selection and Dummy encoding and Feature Scaling on the Attrition
  - Build Binary Classifier to predict which leaves or attritions could be Attrition
- A paragraph that summarizes the quality of the data set and a recipe for additional data if needed
  - Overall, the original dataset was of good quality, with no missing data and no duplicate data. However, to grasp the context of each individual column, additional data that would provide more insight would be the reason of Attrition. Did the individual left further education, did they pursue a job within a completely different industry, or did they make a lateral move.

Data Dictionary

Education: 1 'Below College', 2 'College', 3 'Bachelor', 4 'Master', 5 'Doctor'

Environment Satisfaction: 1 'Low', 2 'Medium', 3 'High', 4 'Very High'

Job_Involvement: 1 'Low', 2 'Medium', 3 'High', 4 'Very High'

Job Satisfaction: 1 'Low', 2 'Medium', 3 'High', 4 'Very High'

Performance_Rating: 1 'Low', 2 'Good', 3 'Excellent', 4 'Outstanding'

Relationship Satisfaction: 1 'Low', 2 'Medium', 3 'High', 4 'Very High'

Work_Life_Balance: 1 'Bad', 2 'Good', 3 'Better', 4 'Best'

Distance_From_Home: Measured in Kilometers

Gender: Male= Label 1, Female=Label 0

Job Level: 1 - 5 scale

Percent_Salary_Hike: Percentage increase compared to the previous year

Importing Libraries & Dataset

```
In [1]: import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        import seaborn as sns
        %matrix inline
        from sklearn.preprocessing import *
        from sklearn.model_selection import *
        from sklearn.metrics import *
        from sklearn.linear_model import *
        from sklearn.ensemble import *
        from sklearn.tree import *
        from sklearn.feature_selection import *
        import pickle
        import pickle5 as pickle5 ##
        %pip install pickle5 ##
        %pip install -U scikit-learn
```

```
In [2]: !wget -q https://dataplus.dev/data-files/WA_Fn-UseC_HR-Employee-Attrition.csv
        ls -l WA_Fn-UseC_HR-Employee-Attrition.csv
```

```
In [3]: df = pd.read_csv("./WA_Fn-UseC_HR-Employee-Attrition.csv")
        df.info()
```

```
In [4]: df.head(5)
```

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

5 rows × 35 columns

EDA

```
In [5]: df.info()
```

```
In [6]: df.describe(include='object')
```

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

```
In [7]: df.describe(include=['int64', 'float64'])
```

	Age	DailyRate	DistanceFromHome	Education	EnvironmentSatisfaction	HourlyRate	JobInvolvement	JobLevel
count	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000
mean	36.923810	802.485714	9.192517	2.912925	2.721769	65.891156	2.729932	2.063946
std	9.135373	403.509100	8.106864	1.024165	1.093082	20.329428	0.711561	1.106940
min	18.000000	102.000000	1.000000	1.000000	1.000000	30.000000	1.000000	1.000000
25%	30.000000	465.000000	2.000000	2.000000	2.000000	48.000000	2.000000	1.000000
50%	36.000000	802.000000	7.000000	3.000000	3.000000	66.000000	3.000000	2.000000
75%	43.000000	1157.000000	14.000000	4.000000	4.000000	83.750000	3.000000	3.000000
max	60.000000	1499.000000	29.000000	5.000000	4.000000	100.000000	4.000000	5.000000

2 rows × 26 columns

```
In [8]: df.select_dtypes(exclude=['int64', 'float64', 'object'])
```

```
In [9]: df.nunique(axis=0)
```

```
In [10]: df.groupby(columns=['Department', 'Gender'], 'DailyRate', 'BusinessTravel', 'DepartmentNumber')
```

```
In [11]: df.shape
```

	Age	DailyRate	DistanceFromHome	Education	EnvironmentSatisfaction	HourlyRate	JobInvolvement	JobLevel
count	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000
mean	36.923810	802.485714	9.192517	2.912925	2.721769	65.891156	2.729932	2.063946
std	9.135373	403.509100	8.106864	1.024165	1.093082	20.329428	0.711561	1.106940
min	18.000000	102.000000	1.000000	1.000000	1.000000	30.000000	1.000000	1.000000
25%	30.000000	465.000000	2.000000	2.000000	2.000000	48.000000	2.000000	1.000000
50%	36.000000	802.000000	7.000000	3.000000	3.000000	66.000000	3.000000	2.000000
75%	43.000000	1157.000000	14.000000	4.000000	4.000000	83.750000	3.000000	3.000000
max	60.000000	1499.000000	29.000000	5.000000	4.000000	100.000000	4.000000	5.000000

2 rows × 26 columns

```
In [12]: df.columns
```

```
In [13]: df.columns = df.columns.str.lower()
         df.columns
```

```
In [14]: df.columns = df.columns.str.replace('_', '')
         df.columns
```

```
In [15]: df.columns = df.columns.str.replace('.', '')
         df.columns
```

```
In [16]: df.columns = df.columns.str.replace('(', '')
         df.columns
```

```
In [17]: df.columns = df.columns.str.replace('-', '')
         df.columns
```

```
In [18]: df.columns = df.columns.str.replace('/', '')
         df.columns
```

```
In [19]: df.columns = df.columns.str.replace(')', '')
         df.columns
```

```
In [20]: df.columns = df.columns.str.replace(' ', '')
         df.columns
```

```
In [21]: df.columns = df.columns.str.replace('&', '')
         df.columns
```

```
In [22]: df.columns = df.columns.str.replace('+', '')
         df.columns
```

```
In [23]: df.columns = df.columns.str.replace('=', '')
         df.columns
```

```
In [24]: df.columns = df.columns.str.replace('[', '')
         df.columns
```

```
In [25]: df.columns = df.columns.str.replace('{', '')
         df.columns
```

```
In [26]: df.columns = df.columns.str.replace('}', '')
         df.columns
```

```
In [27]: df.columns = df.columns.str.replace('|', '')
         df.columns
```

```
In [28]: df.columns = df.columns.str.replace('\', '')
         df.columns
```

```
In [29]: df.columns = df.columns.str.replace('/', '')
         df.columns
```

```
In [30]: df.columns = df.columns.str.replace('?', '')
         df.columns
```

```
In [31]: df.columns = df.columns.str.replace('>', '')
         df.columns
```

```
In [32]: df.columns = df.columns.str.replace('<', '')
         df.columns
```

```
In [33]: df.columns = df.columns.str.replace('!', '')
         df.columns
```

```
In [34]: df.columns = df.columns.str.replace('@', '')
         df.columns
```

```
In [35]: df.columns = df.columns.str.replace('#', '')
         df.columns
```

```
In [36]: df.columns = df.columns.str.replace('$', '')
         df.columns
```

```
In [37]: df.columns = df.columns.str.replace('%', '')
         df.columns
```

```
In [38]: df.columns = df.columns.str.replace('^', '')
         df.columns
```

```
In [39]: df.columns = df.columns.str.replace('*', '')
         df.columns
```

```
In [40]: df.columns = df.columns.str.replace('~', '')
         df.columns
```

```
In [41]: df.columns = df.columns.str.replace('`', '')
         df.columns
```

```
In [42]: df.columns = df.columns.str.replace('.', '')
         df.columns
```

```
In [43]: df.columns = df.columns.str.replace(',', '')
         df.columns
```

```
In [44]: df.columns = df.columns.str.replace(';', '')
         df.columns
```

```
In [45]: df.columns = df.columns.str.replace(':', '')
         df.columns
```

```
In [46]: df.columns = df.columns.str.replace('"', '')
         df.columns
```

```
In [47]: df.columns = df.columns.str.replace('', '')
         df.columns
```

```
In [48]: df.columns = df.columns.str.replace('', '')
         df.columns
```

```
In [49]: df.columns = df.columns.str.replace('', '')
         df.columns
```

```
In [50]: df.columns = df.columns.str.replace('', '')
         df.columns
```

```
In
```

```
>> cat_distplot(data='Age',)
>> cat_distplot(data='Age',)
cat_distplot(data='Daily_Rate',)
cat_distplot(data='Distance_From_Home',)
cat_distplot(data='Hourly_Rate',)
cat_distplot(data='Monthly_Income',)
cat_distplot(data='Monthly_Rate',)
cat_distplot(data='Num_Companies_Worked',)
cat_distplot(data='Percent_Salary_Hike',)
cat_distplot(data='Total_Working_Years',)
cat_distplot(data='Training_Times_Last_Year',)
cat_distplot(data='Years_At_Company',)
cat_distplot(data='Years_In_Current_Role',)
cat_distplot(data='Years_Since_Last_Promotion',)
cat_distplot(data='Years_With_Current_Manager',)
```

Age (cor target = 0.139)

Age

Daily_Rate (cor target = 0.057)

Distance_From_Home (cor target =0.078)

Distance_From_Home

Hourly_Rate (cor target = 0.007)

Monthly_Income (cor target = 0.16)

Monthly_Rate (cor target =0.015)

Num_Companies_Worked (cor target =0.043)

Num_Companies_Worked

Percent_Salary_Hike (cor target = 0.013)

Percent_Salary_Hike

Total_Working_Years (cor target = 0.171)

Total_Working_Years

Training_Times_Last_Year (cor target = 0.059)

Training_Times_Last_Year

Years_At_Company (cor target = 0.134)

Years_At_Company

Years_In_Current_Role (cor target = 0.161)

Years_In_Current_Role

Years_Since_Last_Promotion (cor target = 0.033)

Years_Since_Last_Promotion

Years_With_Current_Manager (cor target = 0.156)

Years_With_Current_Manager

Pie plot and barplot

```
>> pie_plot(data='Attrition',)
>> bar_plot(x='Attrition', 'Department', 'Education', 'EducationField', 'Gender', 'JobRole', 'JobSatisfaction', 'MaritalStatus', 'OverTime', 'PerformanceRating', 'RelationshipSatisfaction', 'WorkLifeBalance', 'BusinessTravel', 'EnvironmentSatisfaction', 'JobInvolvement', 'JobLevel', 'StockOptionLevel', 'TrainingTimesLastYear', 'Years', 'Age', 'DailyRate', 'DistanceFromHome', 'HourlyRate', 'MonthlyIncome', 'MonthlyRate', 'NumCompaniesWorked', 'PercentSalaryHike', 'TotalWorkingYears', 'YearsAtCompany', 'YearsInCurrentRole', 'YearsSinceLastPromotion', 'YearsWithCurrentManager',)
{'data': {'name': 'data-d8d7d0d0d0d0d0d0d0d0d0d0d0d0d0d0',
          'format': {'type': 'csv',
                     'parse': {'Attrition': 'boolean',
                               'BusinessTravel': 'boolean',
                               'Department': 'boolean',
                               'EducationField': 'boolean',
                               'Gender': 'boolean',
                               'JobRole': 'boolean',
                               'MaritalStatus': 'boolean',
                               'OverTime': 'boolean',
                               'Age': 'number',
                               'DailyRate': 'number',
                               'DistanceFromHome': 'number',
                               'Education': 'number',
                               'EnvironmentSatisfaction': 'number',
                               'HourlyRate': 'number',
                               'JobInvolvement': 'number',
                               'JobLevel': 'number',
                               'JobSatisfaction': 'number',
                               'MonthlyIncome': 'number',
                               'MonthlyRate': 'number',
                               'NumCompaniesWorked': 'number',
                               'PercentSalaryHike': 'number',
                               'PerformanceRating': 'number',
                               'RelationshipSatisfaction': 'number',
                               'StockOptionLevel': 'number',
                               'TrainingTimesLastYear': 'number',
                               'WorkLifeBalance': 'number',
                               'Years': 'number',
                               'YearsAtCompany': 'number',
                               'YearsInCurrentRole': 'number',
                               'YearsSinceLastPromotion': 'number',
                               'YearsWithCurrentManager': 'number'},
                     'delimiter': ','},
          'values': [{'Age':41,'Attrition':1,'BusinessTravel':'Travel_Rarely','DailyRate':1102,'Department':'Sales','DistanceFromHome':1,'Education':2,'EducationField':'Life Sciences','EnvironmentSatisfaction':2,'Gender':'Female','HourlyRate':94,'JobInvolvement':3,'JobLevel':2,'JobRole':'Sales Executive','JobSatisfaction':4,'MaritalStatus':'Single','MonthlyIncome':5993,'MonthlyRate':19479,'NumCompaniesWorked':8,'OverTime':'Yes','PercentSalaryHike':11,'PerformanceRating':3,'RelationshipSatisfaction':1,'StockOptionLevel':0,'TrainingTimesLastYear':0,'WorkLifeBalance':1,'Years':8,'YearsAtCompany':6,'YearsInCurrentRole':4,'YearsSinceLastPromotion':0,'YearsWithCurrentManager':5},
          {'Age':49,'Attrition':0,'BusinessTravel':'Travel_Frequently','DailyRate':279,'Department':'Research & Development','DistanceFromHome':8,'Education':1,'EducationField':'Life Sciences','EnvironmentSatisfaction':3,'Gender':'Male','HourlyRate':61,'JobInvolvement':2,'JobLevel':2,'JobRole':'Manufacturing Director','JobSatisfaction':2,'MaritalStatus':'Married','MonthlyIncome':5090,'MonthlyRate':24907,'NumCompaniesWorked':1,'OverTime':'No','PercentSalaryHike':23,'PerformanceRating':4,'RelationshipSatisfaction':4,'StockOptionLevel':1,'TrainingTimesLastYear':3,'WorkLifeBalance':3,'Years':10,'YearsAtCompany':10,'YearsInCurrentRole':7,'YearsSinceLastPromotion':1,'YearsWithCurrentManager':7},
          ...
          {'Age':27,'Attrition':0,'BusinessTravel':'Travel_Rarely','DailyRate':1082,'Department':'Research & Development','DistanceFromHome':2,'Education':3,'EducationField':'Medical','EnvironmentSatisfaction':4,'Gender':'Male','HourlyRate':40,'JobInvolvement':3,'JobLevel':2,'JobRole':'Laboratory Technician','JobSatisfaction':2,'MaritalStatus':'Married','MonthlyIncome':5390,'MonthlyRate':13243,'NumCompaniesWorked':2,'OverTime':'No','PercentSalaryHike':13,'PerformanceRating':3,'RelationshipSatisfaction':2,'StockOptionLevel':0,'TrainingTimesLastYear':3,'WorkLifeBalance':3,'Years':5,'YearsAtCompany':2,'YearsInCurrentRole':2,'YearsSinceLastPromotion':0,'YearsWithCurrentManager':2}]}
,'mark': 'pie',
'encoding': {'theta': {'field': 'Attrition', 'type': 'quantitative', 'aggregate': 'count'},
             'color': {'field': 'Attrition', 'type': 'nominal'},
             'tooltip': [{'field': 'Attrition', 'type': 'nominal'},
                         {'aggregate': 'count', 'type': 'quantitative'},
                         {'aggregate': 'percent', 'type': 'quantitative', 'format': '.1%'}]}
,'height': 300,
'width': 300,
'title': {'text': 'Pie chart of Attrition', 'anchor': 'middle', 'fontSize': 14},
'view': {'stroke': 'transparent'}
,'config': {'view': {'continuousWidth': 300, 'continuousHeight': 300},
            'axis': {'domainColor': '#cccccc', 'gridColor': '#ffffff', 'gridOpacity': 1, 'labelFontSize': 11, 'titleFontSize': 12},
            'legend': {'labelFontSize': 11, 'titleFontSize': 12},
            'header': {'labelFontSize': 11, 'titleFontSize': 12},
            'mark': {'tooltip': null},
            'range': {'category': ['#4c78a8', '#f58518', '#e45756', '#72b7b2', '#54a24b', '#eeca3b', '#b279a2', '#ff9da6', '#9d755d', '#bab0ac'],
                      'ordinal': {'scheme': 'blues'},
                      'ramp': {'scheme': 'blues'},
                      'heatmap': {'scheme': 'viridis'}}}
}
Fig. 1 - Pie Plot#Attrition | Crosstab/Counts/Chi2 - Expected
Pr(>Chi2)=NaN
```

```
In [18]: [cols_plot('Gender',
         cols_plot('Overtime',
         cols_plot('Business_Travel',
         cols_plot('Job_Role',
         cols_plot('Department',
         cols_plot('Marital_Status',
         cols_plot('Education_Field',
         cols_plot('Education',
         cols_plot('Environment_Satisfaction',
         cols_plot('Job_Involvement',
         cols_plot('Job_Level',
         cols_plot('Job_Satisfaction',
         cols_plot('Performance_Rating',
         cols_plot('Relationship_Satisfaction',
         cols_plot('Stock_Option_Level',
         cols_plot('Training_Times_Last_Year',
         cols_plot('Work_Life_Balance',
         cols_plot('Years_at_Company',
         cols_plot('Years_in_Current_Role',
         cols_plot('Years_since_Last_Promotion',
         cols_plot('Years_with_Current_Manager',
         cols_plot('Age',
         cols_plot('Daily_Rate',
         cols_plot('Distance_from_Home',
         cols_plot('Hourly_Rate',
         cols_plot('Monthly_Income',
         cols_plot('Monthly_Rate',
         cols_plot('Percent_Salary_Hike',
         cols_plot('Total_Working_Years',
         cols_plot('Num_Companies_Worked',
         cols_plot('Business_Travel',
         cols_plot('Department',
         cols_plot('Education',
         cols_plot('Education_Field',
         cols_plot('Gender',
         cols_plot('Job_Role',
         cols_plot('Marital_Status',
         cols_plot('Overtime',
         cols_plot('Stock_Option_Level',
         cols_plot('Training_Times_Last_Year',
         cols_plot('Work_Life_Balance',
         cols_plot('Years_at_Company',
         cols_plot('Years_in_Current_Role',
         cols_plot('Years_since_Last_Promotion',
         cols_plot('Years_with_Current_Manager',
         cols_plot('Age',
         cols_plot('Daily_Rate',
         cols_plot('Distance_from_Home',
         cols_plot('Hourly_Rate',
         cols_plot('Monthly_Income',
         cols_plot('Monthly_Rate',
         cols_plot('Percent_Salary_Hike',
         cols_plot('Total_Working_Years',
         cols_plot('Num_Companies_Worked',
         cols_plot('Relationship_Satisfaction',
         cols_plot('Environment_Satisfaction',
         cols_plot('Job_Involvement',
         cols_plot('Job_Level',
         cols_plot('Job_Satisfaction',
         cols_plot('Performance_Rating',
         cols_plot('Relationship_Satisfaction
```

Relationship_Satisfaction distribution in employee attrition

Relationship_Satisfaction

Stock_Option_Level distribution in employee attrition

Stock_Option_Level

Work_Life_Balance distribution in employee attrition

Work_Life_Balance

```
In [70]: df[Education_num] = df[Education].replace({1:'Below College',2:'College',3:'Bachelor',4:'Master',5:'Doctor'})
```

```
In [71]: df[Education_num]
```

```
Out[71]: 0    Bachelor
         1    College
         2    College
         3    Bachelor
         4    College
         5    Bachelor
         6    Bachelor
         7    Bachelor
         8    Bachelor
         9    Bachelor
         10   College
         11   College
         12   Bachelor
         13   Bachelor
         14   Bachelor
         15   Bachelor
         16   Bachelor
         17   Bachelor
         18   Bachelor
         19   Bachelor
         20   Bachelor
         21   Bachelor
         22   Bachelor
         23   Bachelor
         24   Bachelor
         25   Bachelor
         26   Bachelor
         27   Bachelor
         28   Bachelor
         29   Bachelor
         30   Bachelor
         31   Bachelor
```