Factors Affecting Attrition

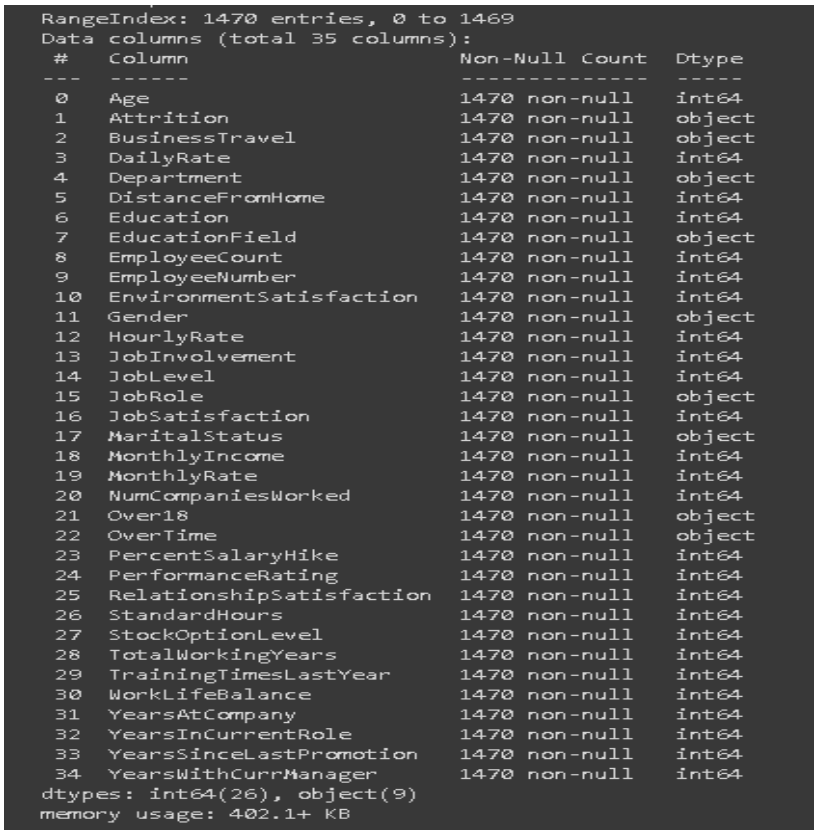
Jishnu Kumar Srivastava

Idea:

- To relate factors leading to employee Attrition.
- Attrition refers to a gradual but deliberate reduction in staff numbers that occurs as employees retire or resign and are not replaced. It can also occur because of downsizing.

Brief description of the data set and a summary of its attributes:

- The dataset 'IBM_HR.csv' was taken from kaggle.
- It contained 1470 observations, 34 features. My target from these was 'Attrition'.
- Data Dictionary:
    - Education: 1 'Below College', 2 'College', 3 'Bachelor', 4 'Master', 5 'Doctor'
    - EnvironmentSatisfaction: 1 'Low', 2 'Medium', 3 'High', 4 'Very High'
    - JobInvolvement: 1 'Low', 2 'Medium', 3 'High', 4 'Very High'
    - JobSatisfaction: 1 'Low', 2 'Medium', 3 'High', 4 'Very High'
    - PerformanceRating: 1 'Low', 2 'Good', 3 'Excellent', 4 'Outstanding'
    - RelationshipSatisfaction: 1 'Low', 2 'Medium', 3 'High', 4 'Very High'
    - WorkLifeBalance: 1 'Bad', 2 'Good', 3 'Better', 4 'Best'
    - DistanceFromHome: Measured in Kilometers
    - StockOptionLevel: JobLevel Scale
    - JobLevel: 1 - 5 scale
    - PercentSalaryHike: Percentage increase compared to the previous year



```
RangeIndex: 1470 entries, 0 to 1469
Data columns (total 35 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Age                       1470 non-null   int64
 1   Attrition                 1470 non-null   object
 2   BusinessTravel            1470 non-null   object
 3   DailyRate                 1470 non-null   int64
 4   Department                1470 non-null   object
 5   DistanceFromHome          1470 non-null   int64
 6   Education                 1470 non-null   int64
 7   EducationField            1470 non-null   object
 8   EmployeeCount             1470 non-null   int64
 9   EmployeeNumber            1470 non-null   int64
 10  EnvironmentSatisfaction   1470 non-null   int64
 11  Gender                    1470 non-null   object
 12  HourlyRate                1470 non-null   int64
 13  JobInvolvement            1470 non-null   int64
 14  JobLevel                  1470 non-null   int64
 15  JobRole                   1470 non-null   object
 16  JobSatisfaction           1470 non-null   int64
 17  MaritalStatus             1470 non-null   object
 18  MonthlyIncome             1470 non-null   int64
 19  MonthlyRate               1470 non-null   int64
 20  NumCompaniesWorked        1470 non-null   int64
 21  Over18                    1470 non-null   object
 22  OverTime                  1470 non-null   object
 23  PercentSalaryHike         1470 non-null   int64
 24  PerformanceRating         1470 non-null   int64
 25  RelationshipSatisfaction  1470 non-null   int64
 26  StandardHours             1470 non-null   int64
 27  StockOptionLevel          1470 non-null   int64
 28  TotalWorkingYears         1470 non-null   int64
 29  TrainingTimesLastYear     1470 non-null   int64
 30  WorkLifeBalance           1470 non-null   int64
 31  YearsAtCompany            1470 non-null   int64
 32  YearsInCurrentRole        1470 non-null   int64
 33  YearsSinceLastPromotion   1470 non-null   int64
 34  YearsWithCurrManager      1470 non-null   int64
dtypes: int64(26), object(9)
memory usage: 402.1+ KB
```

Plan for Data Exploration:

- During data exploration, I went through the updated dataset to understand what each column represented and what would be helpful to me.
- I reviewed columns to understand what data I need and what I don't.
- I checked for null values using ' *isna().sum()* '

Actions taken for data cleaning and feature engineering:

- My target column, 'Attrition' has data as either 'Yes' or 'No'. I mapped 'Yes' to '1' and 'No' to '0' for easier and more helpful analysis.
- I found there to be no null values in the dataset.
- Found a few columns with constant values which were dropped as they would not provide any useful insight or correlation.

```
updated_df = df.drop(['Over18','EmployeeCount','StandardHours'],axis=1)
```

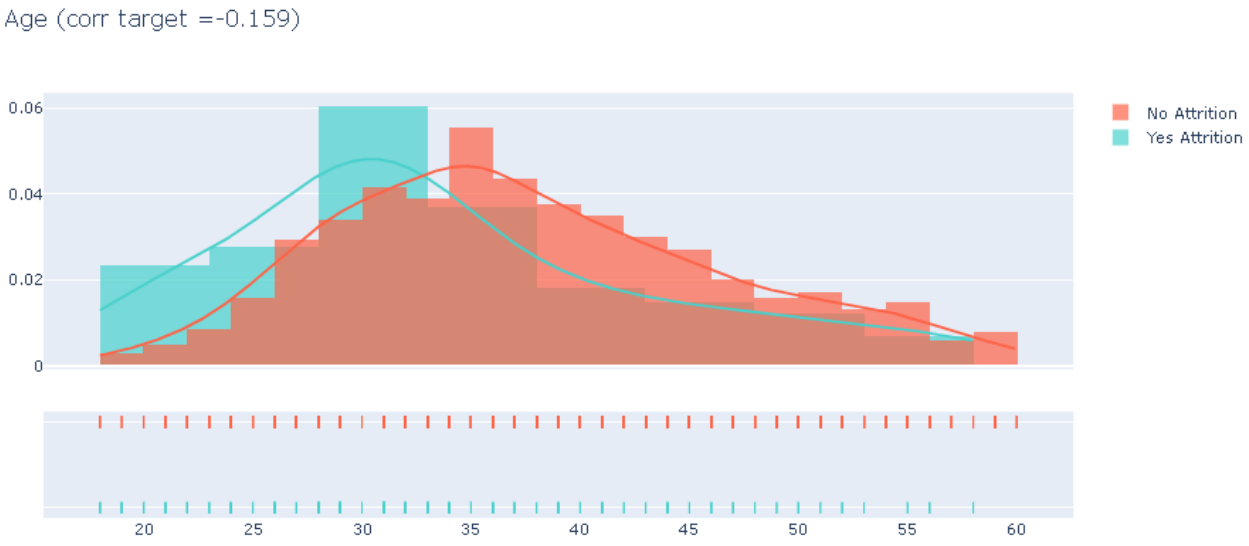
```
updated_df.describe()
```

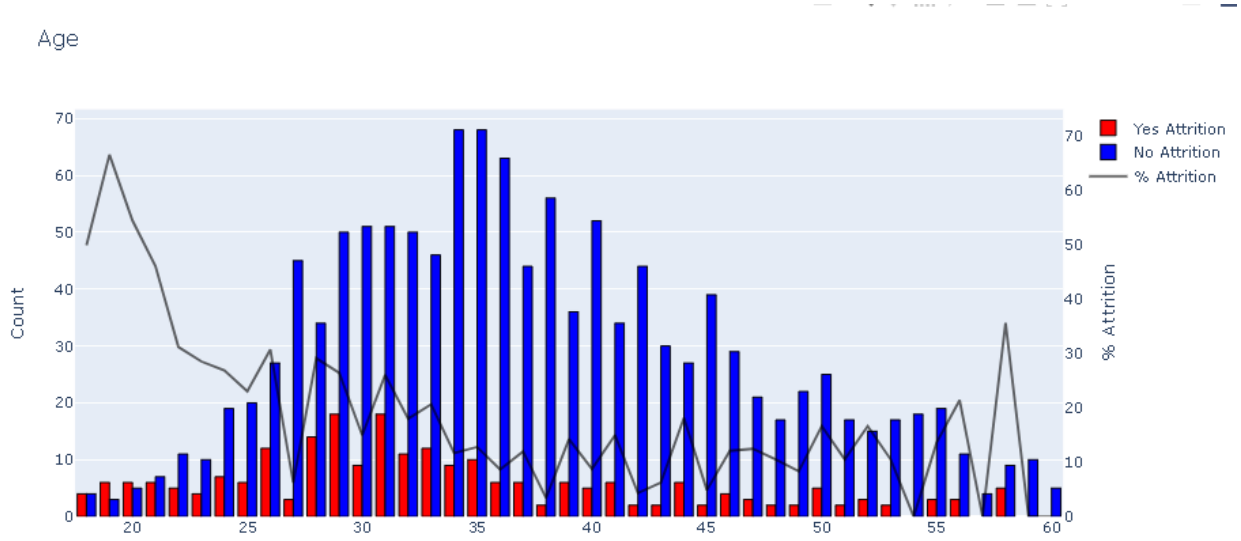
	Age	DailyRate	DistanceFromHome	Education	EmployeeNumber	EnvironmentSatisfaction	HourlyRate	JobInvolvement	JobLevel	JobSatisfaction	Monthly
count	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.0
mean	36.923810	802.485714	9.192517	2.912925	1024.865306	2.721769	65.891156	2.729932	2.063946	2.728571	6502.9
std	9.135373	403.509100	8.106864	1.024165	602.024335	1.093082	20.329428	0.711561	1.106940	1.102846	4707.9
min	18.000000	102.000000	1.000000	1.000000	1.000000	1.000000	30.000000	1.000000	1.000000	1.000000	1009.0
25%	30.000000	465.000000	2.000000	2.000000	491.250000	2.000000	48.000000	2.000000	1.000000	2.000000	2911.0
50%	36.000000	802.000000	7.000000	3.000000	1020.500000	3.000000	66.000000	3.000000	2.000000	3.000000	4919.0
75%	43.000000	1157.000000	14.000000	4.000000	1555.750000	4.000000	83.750000	3.000000	3.000000	4.000000	8379.0
max	60.000000	1499.000000	29.000000	5.000000	2068.000000	4.000000	100.000000	4.000000	5.000000	4.000000	19999.0

Key Findings & Initial Insights:

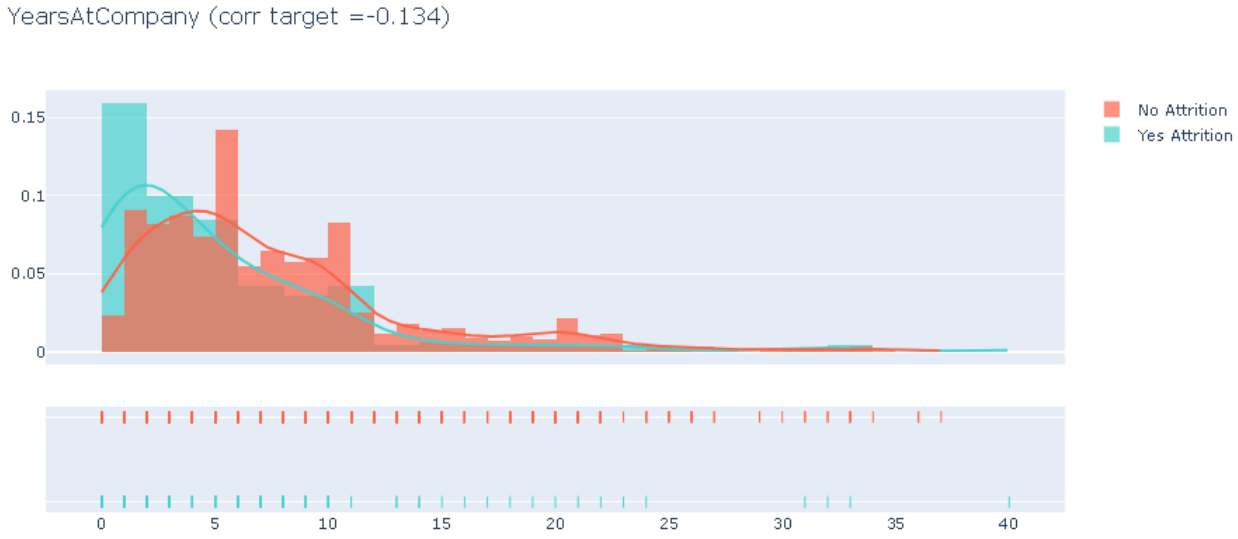
- To get my very first relational insights on attrition with other features like Age,DailyRate, MonthlyIncome, etc, I did plotting of Attrition against other features.
- Some of the plots are:

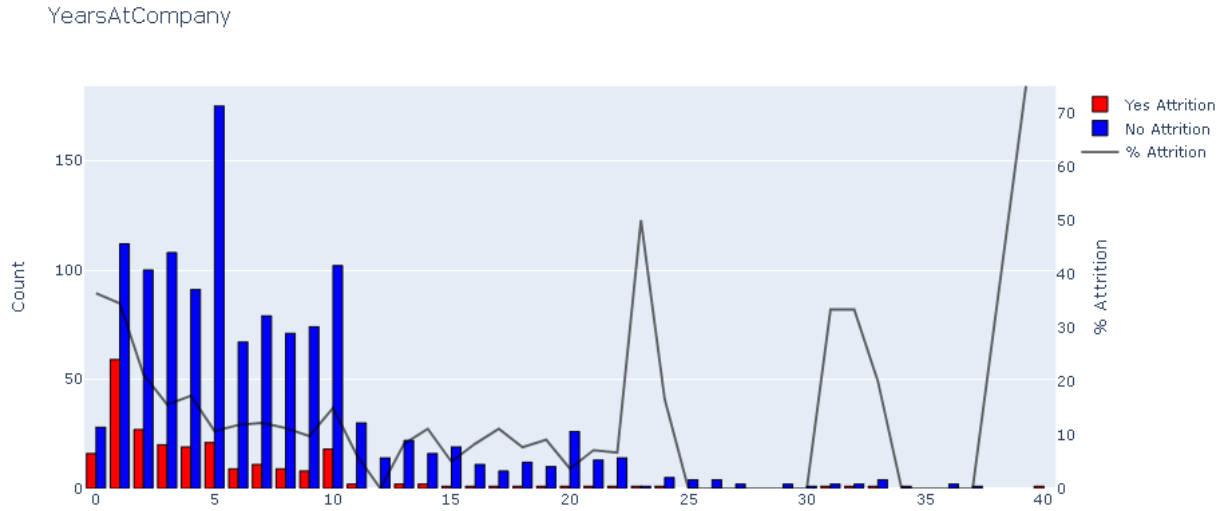
a) Age:



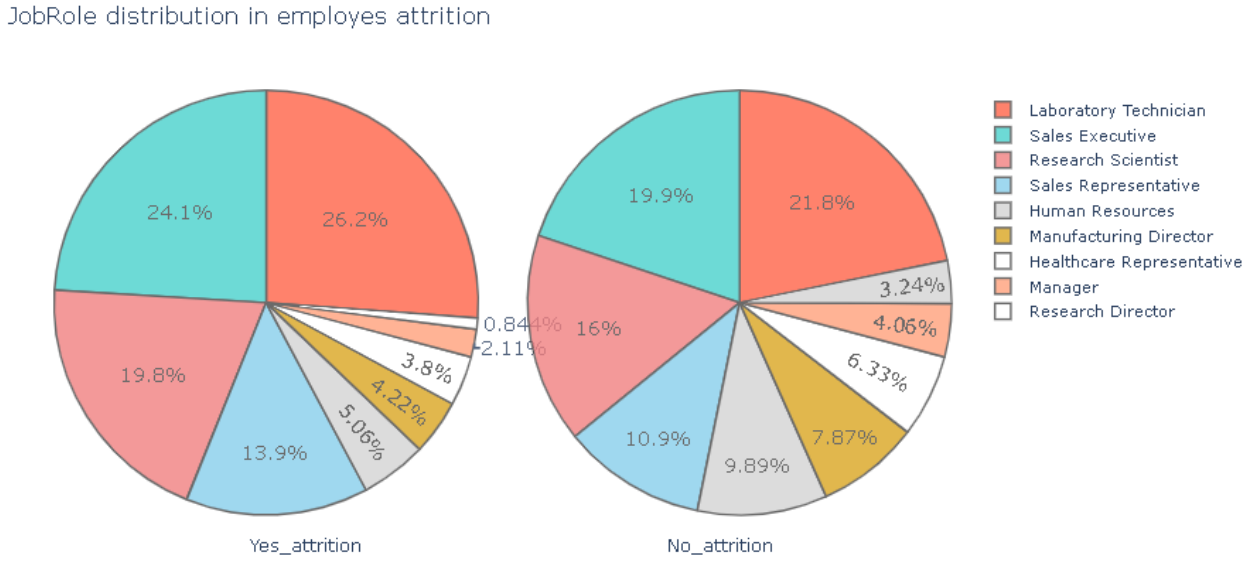


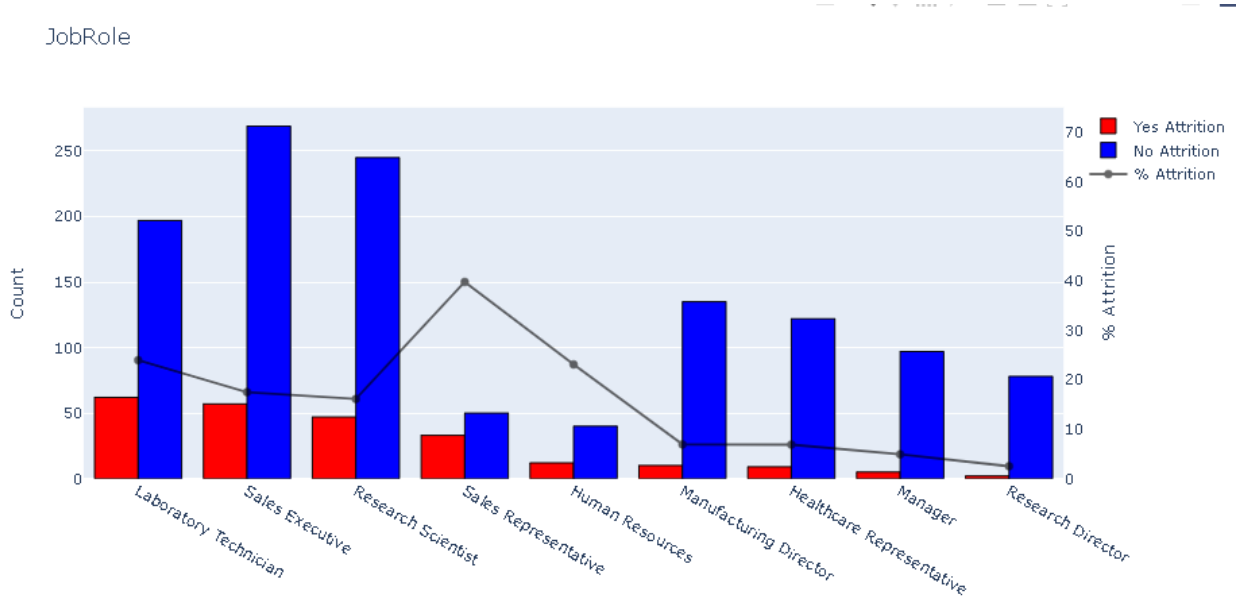
b) Years at Company:



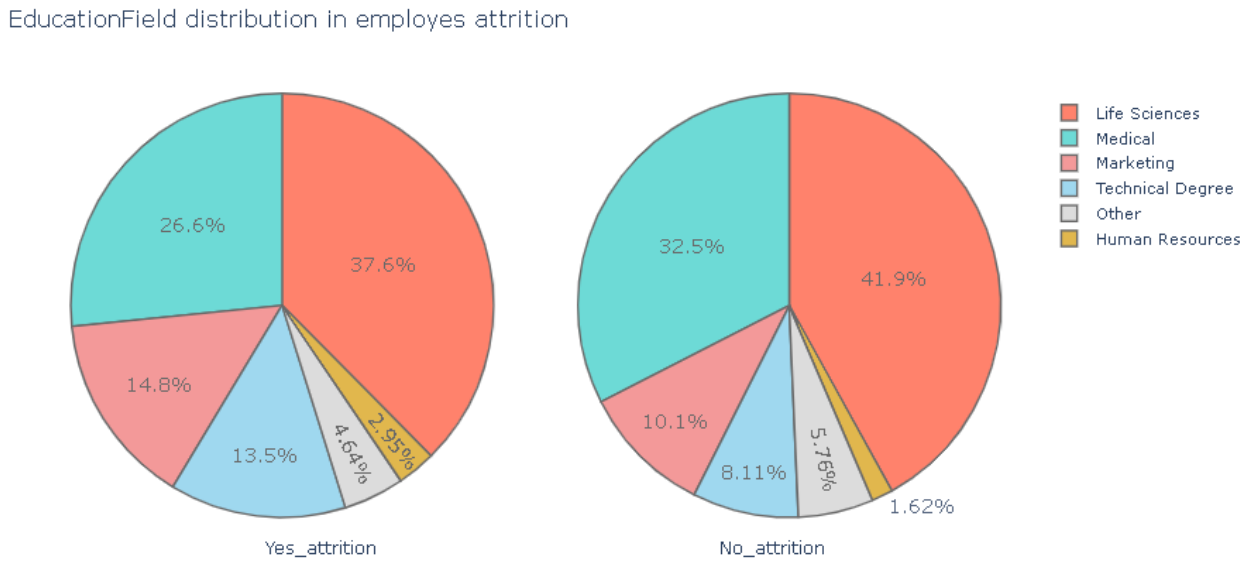

YearsAtCompany
Count
% Attrition
Yes Attrition
No Attrition
% Attrition
0
50
100
150
0
10
20
30
40
50
60
70
0
5
10
15
20
25
30
35
40

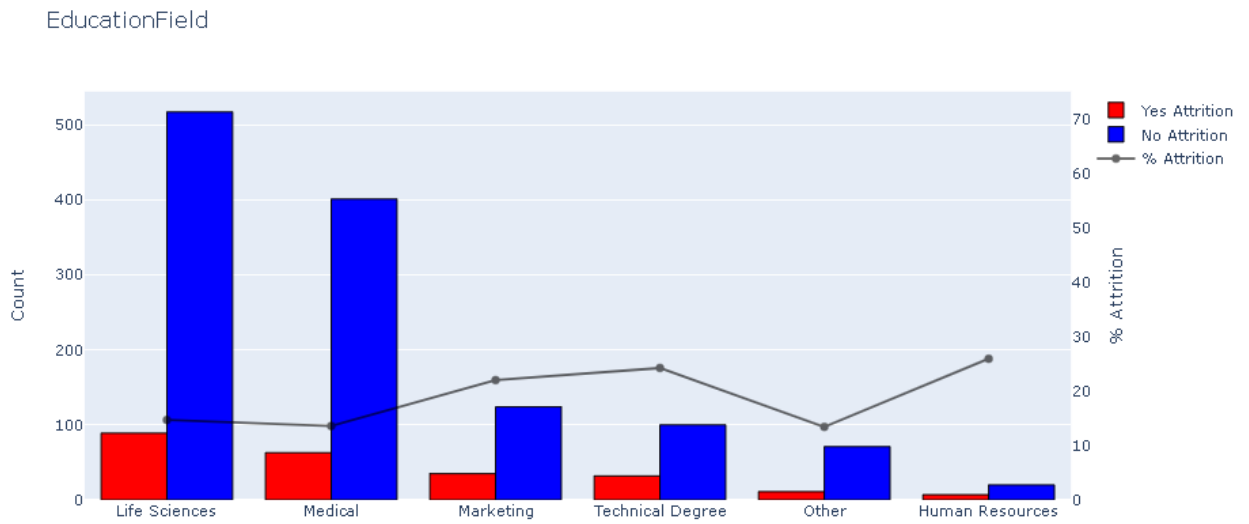
c) Job Role:


JobRole distribution in employes attrition
24.1%
26.2%
19.8%
13.9%
5.06%
4.22%
3.8%
2.11%
0.844%
19.9%
21.8%
16%
10.9%
9.89%
7.87%
6.33%
4.06%
3.24%
Yes_attrition
No_attrition
Laboratory Technician
Sales Executive
Research Scientist
Sales Representative
Human Resources
Manufacturing Director
Healthcare Representative
Manager
Research Director

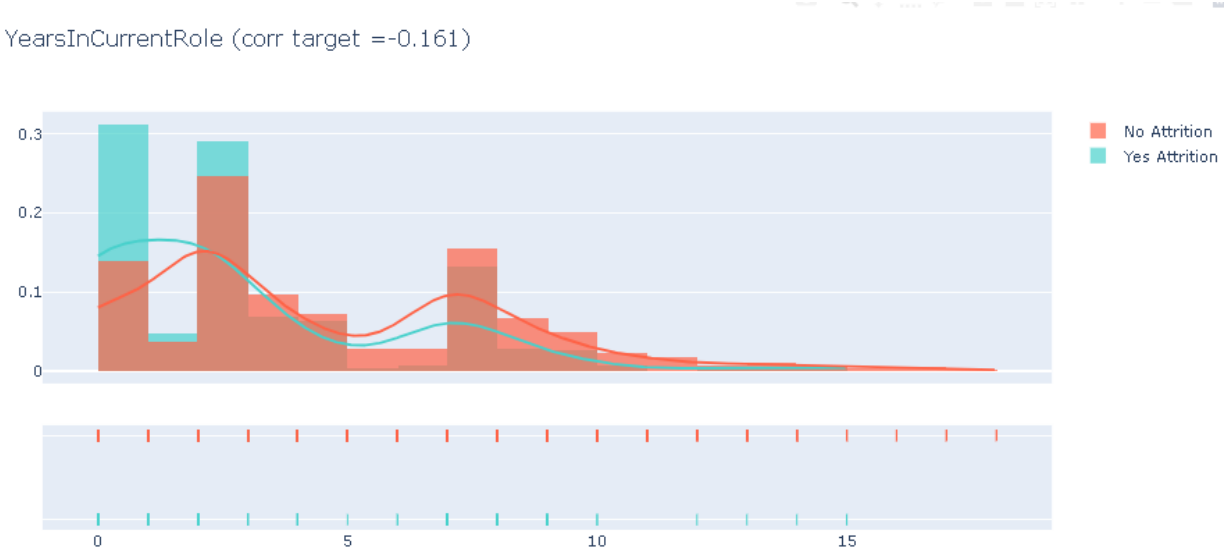


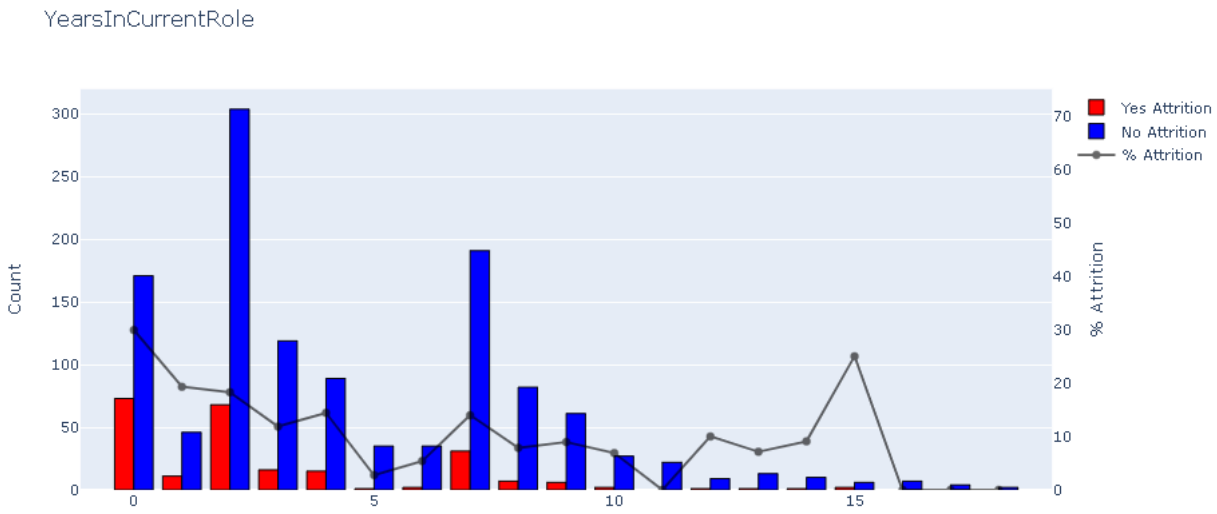
d) Education Field:





e) Years in Current Role:





{there are many more graphs for insights, I will link the notebook at the end and you go run it yourself to have a look at all the graphs}

Insights:

- 'Attrition' was at 66.67% for 19 year olds, but then the 'Attrition' rate drastically drops until employees hit the age of 58, whereafter 'Attrition' jumps up to 35.71%.
- When 'YearsAtCompany' was 0 the 'Attrition' rate was 36.36%, whereas when ‘YearsatCompany’ became 1, the attrition rate was 34.50%.
- For the ‘MaritalStatus - Single’ the attrition rate was 25.53%, and interestingly it was much lower for married people at 12.48% and even lower for divorced at 10.09%.
- 'EducationField' of Human Resources has an 'Attrition' rate of 25.92%, whereas that of Medical was at 13.57%.
- When 'TotalWorkingYears' was 0 the attrition rate was 45.45%, when 1 the attrition rate was 49.38% and at 40 it was 100%, due to retirement.

Hypothesis Formulation:

- I formulated 3 hypothesis for this data:
- a) Ho: μ EducationAttrition == μ EducationNotAttrition

Ha: μ EducationAttrition != μ EducationNotAttrition

- b) Ho: μ AgeAttrition == μ AgeNotAttrition

  Ha: μ AgeAttrition != μ AgeNotAttrition
- c) Ho: μ JobSatisfactionAttrition == μ JobSatisfactionNotAttrition

  Ha: μ JobSatisfactionAttrition != μ JobSatisfactionNotAttrition

Formal Significance Tests and results:

- Kruskal Test was used because it is a non-parametric test.
- First we have to set the attrition_df and not_attrition_df as such:

```
attrition_df = updated_df[updated_df['Attrition'] == 1]
not_attrition_df = updated_df[updated_df['Attrition'] == 0]
```

Education-Attrition Hypothesis test:

```
ss.kruskal(attrition_df['Education'], not_attrition_df['Education'])
```

```
KruskalResult(statistic=1.352764091309354 8, pvalue=0.2447954753326153)
```

∴ **pvalue > 0.05** , Hence there appears to be no statistically significant relationship between Attrition and Education, thus we **fail to reject the null hypothesis**.

Age-Attrition Hypothesis test:

```
ss.kruskal(attrition_df['Age'], not_attrition_df['Age'])
```

```
KruskalResult(statistic=43.06268844023747, pvalue=5.3013684961038114e-11)
```

∴ **pvalue < 0.05** , Hence there appears to be statistically significant relationship between Attrition and Age, thus we **do reject the null hypothesis.**

JobSatisfaction-Hypothesis test:

```
ss.kruskal(attrition_df['JobSatisfaction'], not_attrition_df['JobSatisfaction'])
KruskalResult(statistic=15.568947932935844, pvalue=7.955037680315368e-05)
```

∴ **pvalue < 0.05** , Hence there appears to be statistically significant relationship between Attrition and Job Satisfaction, thus we **do reject the null hypothesis.**

Conclusion:

- We conclude that Attrition has a higher percentage when employees are younger than compared to older. Also, as Job Satisfaction increases, Attrition becomes lesser.
- This was a clean dataset with not a great many features, but just enough to be able to understand some causes of Attrition.
- I got a lot to learn and explore, especially the interactive plotly library and how to infer from data. I also understood what data can be removed and that only the data essential to my aim should be kept else buffer data can lead to increased time and misleading results.
- I would like to thank all my peers who will review my work.
- Hence this Exploratory Data Analysis carried out by me, alongside this report is my submission for the peer graded review.

Link to project notebook(ipynb):
https://colab.research.google.com/drive/1L6R1A1x0yLCxuasi9C-OH2GCpeAJg34p?usp=sharing

The dataset can be found here:
https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset