

IBM EDA peer-graded assignment

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

The original dataset contained 1470 observations, 34 features, and a target of 'Attrition'. Furthermore, the aforementioned dataset was clean and contained no NULL values. Moreover, a data dictionary was provided in order to grasp the context of individual labels; such as Education, Environment_Satisfaction, Job_Involvement, Job_Satisfaction, Performance_Rating, Relationship_Satisfaction, Work_Life_Balance, Job_Level, and Stock_Option_Level being on a scale from 1-5, while Distance_From_Home was measured in Kilometers and Percent_Salary_Hike is the percent increase in salary compared to the previous year.

Initial plan for data exploration

- Check for missing data (NULLS).
- Review columns to identify what is needed or not.
- Review column names and understand what each label means.
- Make plots for initial insights.

Actions taken for data cleaning and feature engineering

- Clean column names.
- Review target variable against other features ('Attrition' vs x) and create plots 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

- 'Attrition' was at 66.67% for 19 year olds, however the 'Attrition' rate drastically drops until employees hit the age of 58, where 'Attrition' jumps up to 35.71%.
- When 'Distance_From_Home' was 25 kilometers the 'Attrition' rate 42.85%.
- When 'Percent_Salary_Hike' was 24 (24% increase from previous year) 'Attrition' rate was 28.57%.
- When 'Total_Working_Years' was 0 the 'Attrition' rate was 45.45%, when 1 the 'Attrition' rate was 49.38% and at 40 100%, likely due to retirement.
- When 'Training_Time_Last_Year' was 0 the 'Attrition' rate was 27.78%.
- When 'Year_At_Company' was 0 the 'Attrition' rate was 36.36%, when 1 34.50%.
- Interestingly when an employee was given 'Overtime' their 'Attrition' rate was 30.52% compared to 10.43%.
- When an employee 'Travel_Frequently' the 'Attrition' rate was 24.90%.
- Not surprisingly the 'Job_Role' of Sales Rep with the highest 'Attrition' rate was 39.75%.
- Employees with 'Marital_Status' single has an 'Attrition' rate of 25.53.
- 'Education_Field' of Human Resources has an 'Attrition' rate of 25.92%.

Formulating at least 3 hypothesis about this data

1)

- Ho: $\mu_{\text{Education_Attrition}} == \mu_{\text{Education_Not_Attrition}}$
- Ha: $\mu_{\text{Education_Attrition}} != \mu_{\text{Education_Not_Attrition}}$

2)

- Ho: $\mu_{\text{Age_Attrition}} == \mu_{\text{Age_Not_Attrition}}$
- Ha: $\mu_{\text{Age_Attrition}} != \mu_{\text{Age_Not_Attrition}}$

3)

- Ho: $\mu_{\text{Job_Satisfaction_Attrition}} == \mu_{\text{Job_Satisfaction_Not_Attrition}}$
- Ha: $\mu_{\text{Job_Satisfaction_Attrition}} != \mu_{\text{Job_Satisfaction_Not_Attrition}}$

Conducting a formal significance test for one of the hypotheses and discuss the results

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

```
KruskalResult(statistic=1.3527640913093548, pvalue=0.2447954753326153)
```

```
pvalue > 0.05
```

There appears to be no statistically significant relationship between Attrition and Education, thus we fail reject the null hypothesis.

Suggestions for next steps in analyzing this data

- Continue testing hypothesis to hone in on and uncover what the true indicators, along with their magnitude, are for Attrition.
- Would like to create models to uncover which feature is attributed most to Attrition.

A paragraph that summarizes the quality of this data set and a request for additional data if needed

Once again, this dataset was incredibly clean with no nulls or weird values. Additionally, a data dictionary was provided in order to grasp the context of each individual column. Moreover, additional data that would provide more insight would be the reason of Attrition. Did the individual seek further education, did they pursue a job within a completely different industry, or did they make a lateral move.