EDA Final Project

By Ghazali Akmal Rabbani

Employee Promotion Analysis

Dataset Link :

https://www.kaggle.com/muhammadimran112233/employee-promotion-end-to-end-solution/data

Data Description

The HR team stored data of promotion cycle last year, which consists of details of all the employees in the company working last year and also if they got promoted or not, but every time this process gets delayed due to so many details available for each employee - it gets difficult to compare and decide.

Data Dictionary

- employee*id: Unique ID for the employee*
- *department: Department of employee*
- *region: Region of employment (unordered)*
- *education: Education Level*
- *gender: Gender of Employee*
- *recruitment*channel: Channel of recruitment for employee
- no_ of_ trainings: no of other trainings completed in the previous year on soft skills, technical skills, etc.
- age: Age of Employee
- previous_ year_ rating: Employee Rating for the previous year
- length_ of_ service: Length of service in years
- awards_ won: if awards won during the previous year then 1 else 0
- avg_ training_ score: Average score in current training evaluations
- is_promoted: (Target) Recommended for promotion. 1 is for Promoted. 0 is for Not Promoted

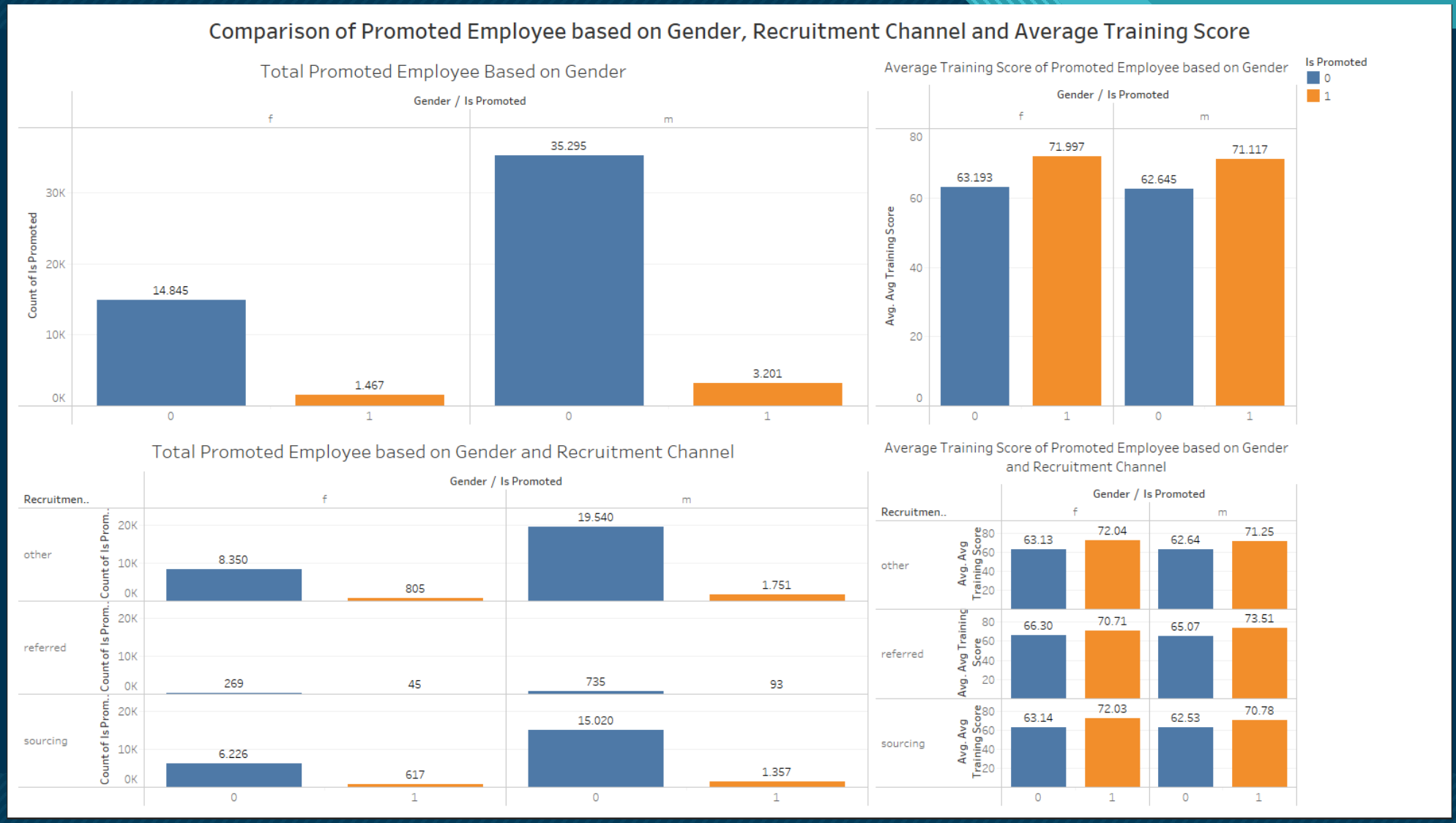
Initial Plan for Exploratory Data Analysis

- Explore total female and male employee that got promoted in company
- Explore the average of average training score of all employee in company
- Explore total female and male employee from Other and Sourcing Recruitment Channel that got promoted in company
- Explore the average of average training score of all promoted employee from Other and Sourcing Recruitment Channel that got promoted in company

Data Cleaning and Feature Engineering

- Replacing Missing Value in Numerical Column with Median of its column data
- Replacing Missing Value in Categorical Column with Mode of its column data

Key Finding and Insight (Dashboard)



Key Finding and Insight (Explanation)

- Total Female Promoted is 1467 person and Total Male Promoted is 3201 person
- Average of Average Training Score for Female Employee who not promoted is 63,19 and promoted is 71,99. Meanwhile for Male Employee not promoted is 62,64 and promoted is 71,11
- Total Female Employee Promoted from Other Recruitment Channel is 805 person and from Sourcing is 617 person
- Meanwhile Total Male Employee Promoted from Other Recruitment Channel is 1751 person and from Sourcing is 1357 person
- Average of Average Training Score for Female Employee who Promoted and from Other Recruitment Channel is 72,04 and from Sourcing Channel is 72,03
- Meanwhile, average of average training score for Male Employee who Promoted and from Other Recruitment Channel is 71,25 and from Sourcing Channel is 70,78

Hypothesis Testing

- There are several null hypothesis that can be inferred from dashboard in above :
- Null Hypothesis 1 : Average score of Female and Male is same
- Alternate Hypothesis 1 : Average score of Female and Male is not same

- Null Hypothesis 2 : Average score of Female and Male that got promoted is same
- Alternate Hypothesis 2 : Average score of Female and Male that got promoted is not same

- Null Hypothesis 3 : Average score of All Promoted Employee from Other Recruitment Channel and Sourcing Recruitment Channel is same
- Alternate Hypothesis 3 : Average score of All Promoted Employee from Other Recruitment Channel and Sourcing Recruitment Channel is not same

Hypothesis Testing 1 Result

- Null Hypothesis 1 : Average score of Female and Male is same
- Alternate Hypothesis 1 : Average score of Female and Male is not same

```
stats.ttest_ind(dfFemale, dfMale)
```

executed in 13ms, finished 10:34:21 2021-10-13

```
Ttest_indResult(statistic=array([5.14470471]), pvalue=array([2.68863219e-07]))
```

Based on result in above, pvalue is less than 0.05. So that, we can reject Null Hypothesis and Accept Alternate Hypothesis. It means that Average Training Score of Female and Male employee is not same. Also, based on T-value/Statistic Value in above, it indicate that Average Training Score of Female is higher than Male employee because we got positive T-Value,i.e. 5,14

Hypothesis Testing 2 Result

- Null Hypothesis 2 : Average score of Female and Male that got promoted is same
- Alternate Hypothesis 2 : Average score of Female and Male that got promoted is not same

```
stats.ttest_ind(dfFemalePromoted, dfMalePromoted)
```

executed in 14ms, finished 10:34:22 2021-10-13

```
Ttest_indResult(statistic=array([1.90779219]), pvalue=array([0.0564794]))
```

Based on result in above, pvalue is less than equal to 0.05. So that, we can reject Null Hypothesis and Accept Alternate Hypothesis. It means that Average Training Score of Female Promoted and Male Promoted employee is not same. Also, based on T-value/Statistic Value in above, it indicate that Average Training Score of Female Promoted is higher than Male Promoted employee because we got positive T-Value,i.e. 1.90

Hypothesis Testing Result 3

- Null Hypothesis 3 : Average score of All Promoted Employee from Other Recruitment Channel and Sourcing Recruitment Channel is same
- Alternate Hypothesis 3 : Average score of All Promoted Employee from Other Recruitment Channel and Sourcing Recruitment Channel is not same

```
stats.ttest_ind(AVGTrainingScorePromotedEmployee_Other, AVGTrainingScorePromotedEmployee_Sourcing)
```

executed in 15ms, finished 10:34:22 2021-10-13

```
Ttest_indResult(statistic=array([0.753227]), pvalue=array([0.45135268]))
```

Based on result in above, Pvalue is 0.45, which is less than 0.05. It can be conclude that we accept the Null Hypothesis and Average Training Score of Promoted Employee from Other Recruitment Channel and Sourcing Channel is same

Suggestion for Next Data Analysis

- Explore and Analysis total Employee which have different education background
- Do hypothesis testing with one sample t-test to the dataset
- Do hypothesis testing with paired independent t-test to the dataset with different features

Summary of This Dataset

This Employee Promotion dataset has quite clean dataset from missing value. It just need few step of cleaning data. This data is recommended for any people who want to learn Exploratory Data Analysis, Hypothesis Testing or even create machine learning model. As the good quality of this dataset, this dataset doesn’t need additional data.

Github Link

https://github.com/ghzza/IBM-Machine-Learning/tree/main/EDA

Thank You