# HR EMPLOYEE ATTRITION

*Dissertation submitted in fulfilment of the requirements for the Degree of*

## BACHELOR OF TECHNOLOGY

**in**

**COMPUTER SCIENCE AND ENGINEERING**

By

### MANISH KUMAR DAS

### 1201595

Supervisor

**AJAY SHARMA**



**School of Computer Science and Engineering**

Lovely Professional University Phagwara, Punjab (India), April 2024

# ACKNOWLEDGEMENT

Primarily I would like to thank God for being able to learn a new technology. Then I would like to express my special thanks of gratitude to the teacher and instructor of the course Exploratory Data Analysis Project who provided me with the golden opportunity to learn a new technology.

I would like to also thank my own college, Lovely Professional University, for offering such a course which not only improve my programming skill but also taught me other new technology.

Then I would like to thank my parents and friends who have helped me with their valuable suggestions and guidance in choosing this course.

Finally, I would like to thank everyone who has helped me a lot.

**Dated: 30TH June 2024**

# TABLE OF CONTENTS

|  |  |  |
| --- | --- | --- |
| **S. No.** | **Contents** | **Page** |
| **1** | **Title** | **1** |
| **3** | **Acknowledgement** | **2** |
| **4** | **Table of Contents** | **3** |
| **7** | **Introduction** | **4** |
| **8** | **Domain Knowledge** | **4** |
| **9** | **Reason of Choosing the**  **Dataset** | **4** |
| **10** | **Library Used and**  **Approaches** | **5** |
| **12** | **Data Description** | **5** |
| **13** | **Exploratory Data Analysis** | **6-24** |
| **14** | **Conclusion** | **24** |
| **15** | **References** | **24** |

**INTRODUCTION**

In the dynamic landscape of modern business, where the war for talent rages on, organizations face a formidable challenge: retaining skilled employees. Human Resource (HR) departments grapple with the complexities of understanding why employees leave, often referred to as employee attrition, and how to mitigate this costly phenomenon. In this pursuit, Exploratory Data Analysis (EDA) emerges as a powerful tool, offering a lens through which to decipher the intricate patterns and factors contributing to attrition within an organization.

This report delves into the realm of HR employee attrition through the prism of exploratory data analysis. It aims to illuminate the intricate interplay of variables, shedding light on the underlying causes and potential solutions for mitigating attrition. By leveraging data-driven insights, organizations can proactively address attrition, fostering a work environment conducive to employee satisfaction, engagement, and ultimately, retention.

# DOMAIN KNOWLEDGE

Employee attrition represents more than just a numerical metric; it embodies a multifaceted phenomenon influenced by a myriad of internal and external variables. From organizational culture and leadership effectiveness to job satisfaction and career growth opportunities, the drivers of attrition are as diverse as they are complex. Through EDA, we embark on a journey to dissect these factors, discerning the pivotal drivers that precipitate attrition and identifying opportunities for intervention. In an era characterized by unprecedented mobility and competition for skilled labor, the stakes of employee retention have never been higher. Beyond the tangible costs associated with recruitment and onboarding, attrition inflicts intangible wounds upon organizational morale and productivity. Consequently, organizations must embrace a proactive stance towards retention, leveraging EDA as a compass to navigate the tumultuous waters of talent management.

As we embark on this exploratory journey into HR employee attrition, armed with the tools of data analysis and a commitment to organizational excellence, we are poised to uncover the insights that will shape the future of talent retention. Through meticulous examination of data, thoughtful interpretation of findings, and strategic implementation of solutions, we endeavor to transform attrition from a looming threat into an opportunity for organizational growth and resilience.

# REASON OF CHOOSING THIS DATASET

In this era of heightened competition for talent, effective decision-making regarding HR employee attrition is paramount. Attrition not only incurs significant costs in recruitment and training but also disrupts organizational stability and productivity. Moreover, in a digital age where employee experiences are shared instantaneously, high attrition rates can tarnish employer branding, deterring top talent from joining the organization. Proactive measures to understand and mitigate attrition through data-driven insights not only enhance employee retention but also foster a positive workplace culture, positioning the organization as an employer of choice in an increasingly competitive talent landscape.

# LIBRARY USED AND APPROACHES

I have used pandas, numpy, matplotlib.pyplot, seaborn for this project. To start, pandas is used to load and manipulate the data, allowing for easy cleaning and preprocessing. NumPy is then employed for numerical operations and calculations, providing a foundation for statistical analysis. Seaborn and matplotlib.pyplot are utilized for data visualization, enabling the creation of informative plots such as histograms, scatter plots, and box plots. These visualizations help in understanding the distribution of data, identifying outliers, and exploring relationships between variables. Through this iterative process of data manipulation and visualization, EDA allows for the identification of patterns, trends, and anomalies within the dataset, ultimately informing decision-making and hypothesis generation for further analysis.

# DATA DESCRIPTION

*Columns: 35*

*Rows :1470*

Employee no: Unique id of the employee Age: age of the workers

Years in company

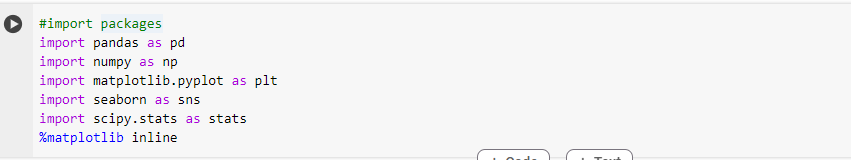
Years in current role: not yet promoted workers Years in same post

Standard hours: time each employee works

Total working hours: monthly or yearly working hours of the employee.

# EXPLORATORY DATA ANALYSIS

### IMPORTING REQUIRED LIBRARIES:



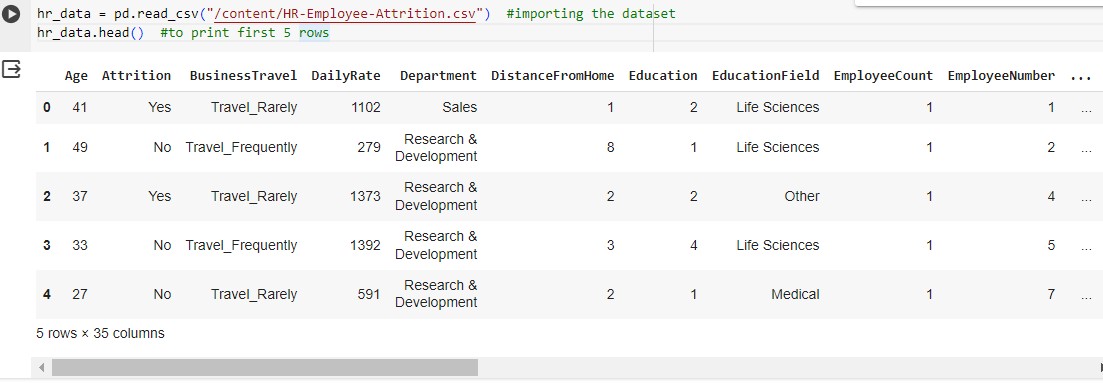
First, we need to import the pandas and NumPy as pd and np respectively is used to perform a wide variety of mathematical operations on arrays.

Matplotlib.pyplot and seaborn are used to display different types of charts which will help in further decision making.

And Pandas is used because it is an open-source library. It can be used to perform data manipulation and analysis.

### KNOWING AND CHECKING FIRST 5 ROWS OF THE DATASET:

head() is used to find the first elements of first 5 rows.

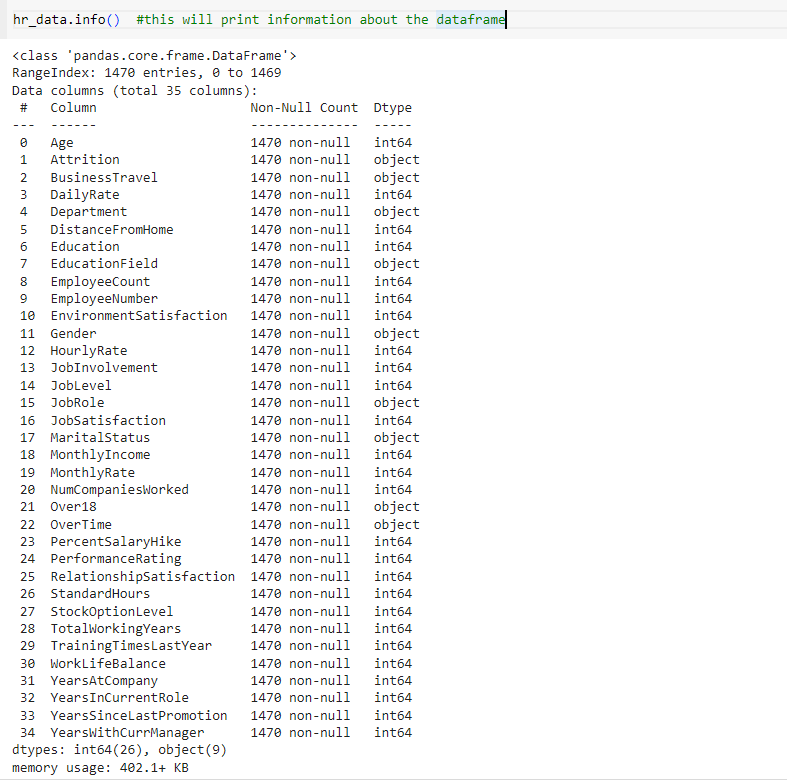


### describe() is used to view some basic statistical details like percentile, mean, std etc. of a data frame.



**FINDING THE DATATYPES USING info() METHOD:**

This step helps us to find the different types of Datatypes available in the given data set.

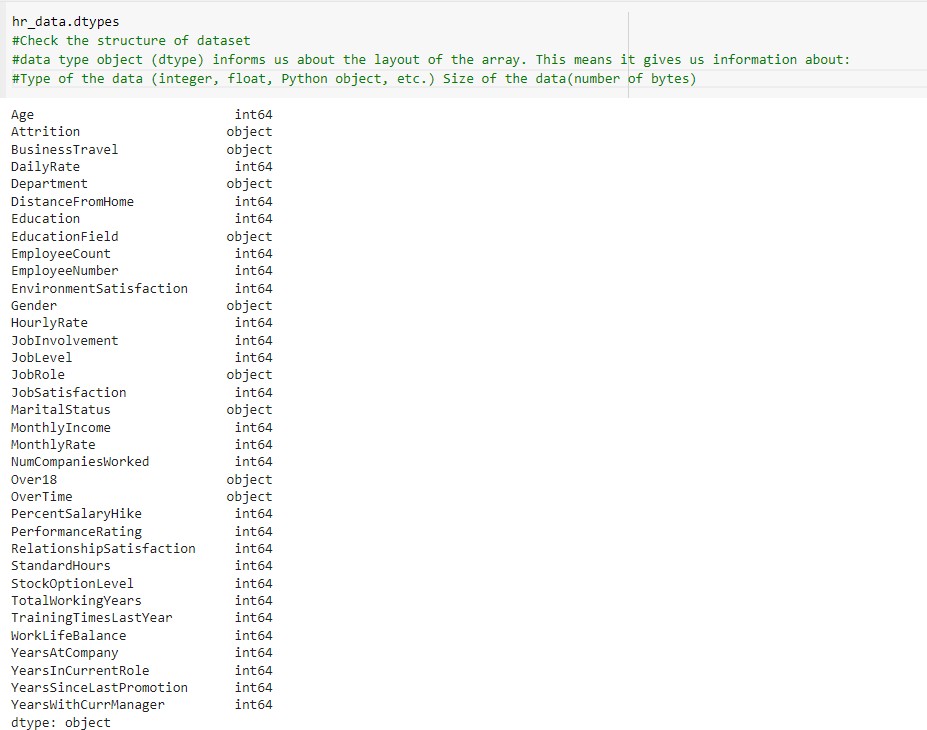


### CHECKING THE NULL VALUES:

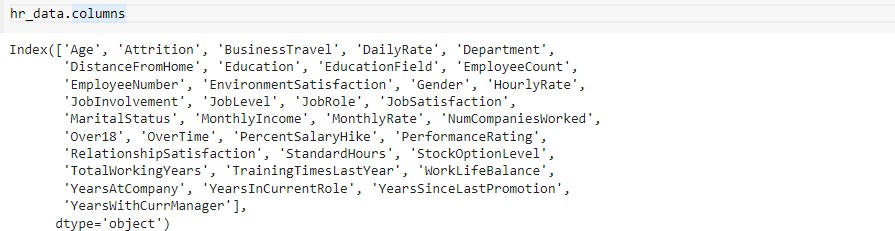
BY determining the source of missingness in our datasets, we can more accurately decide how to handle missing values to improve model quality.



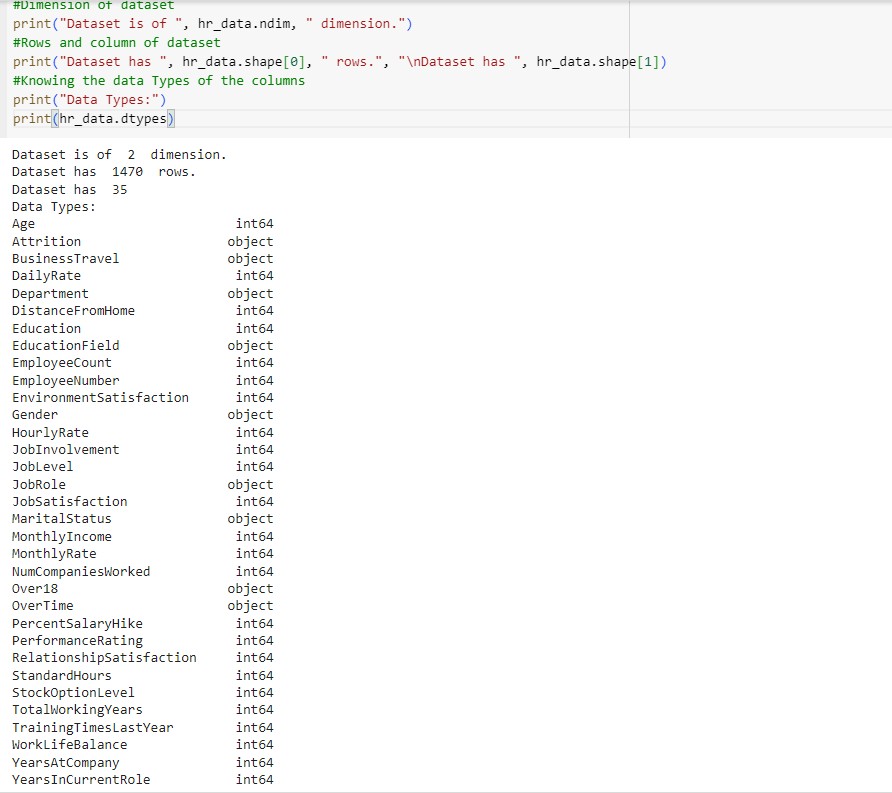
### TO FIND DATATYPES USING dtypes() METHOD:



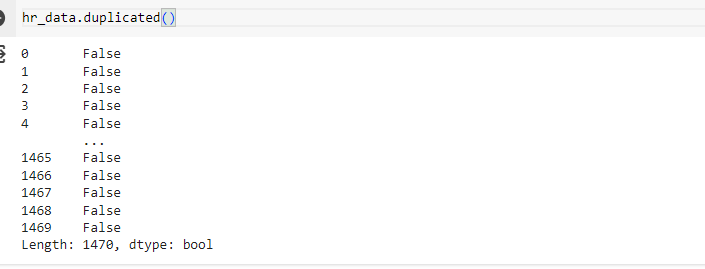
**CHECKING THE COLUMNS USING .columns:**



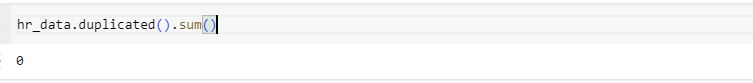
### KNOWING THE DATATYPES OF THE COLUMN:



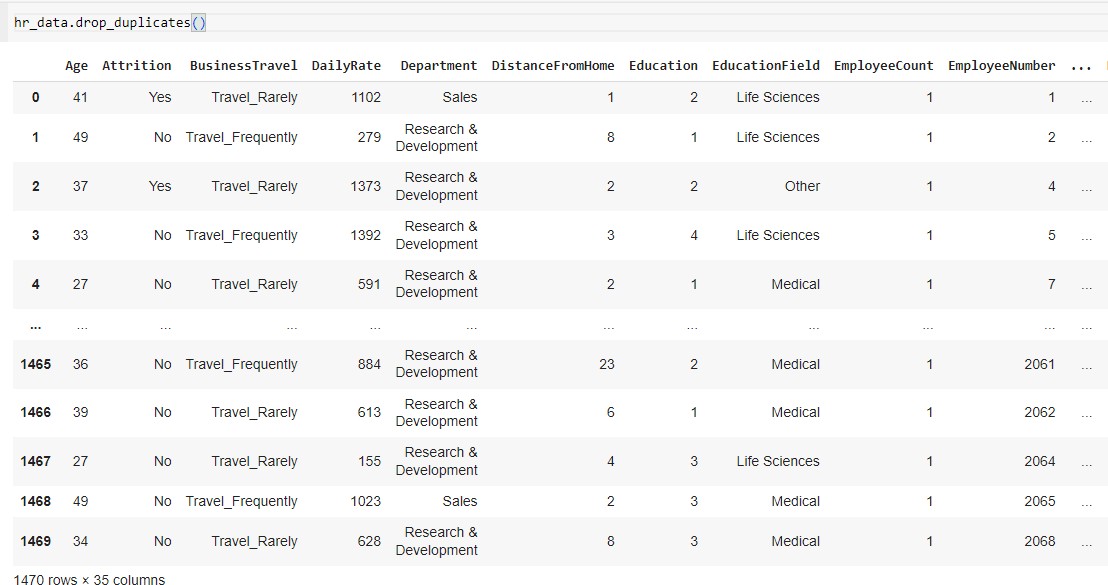
**CHECKING THE DUPLICATE ROWS AND COLUMNS:**



### TO KNOW TOTAL NUMBER OF DUPLICATES:



**REMOVING THE DUPLICATE ROWS AND COLUMNS:**

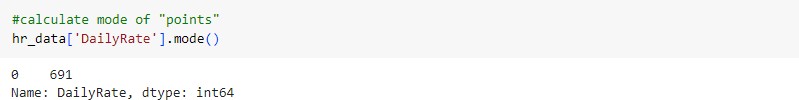


Removing duplicate values plays an important role in the cleansing process. Duplicate data takes up unnecessary storage space and slows down calculations at a minimum.

**UNIVARIATE DATA ANALYSIS**

### CALCULATE MEAN OF THE POINTS:

**CALCULATE MODE OF THE POINTS:**



### CALCULATE STANDARD DEVIATION:

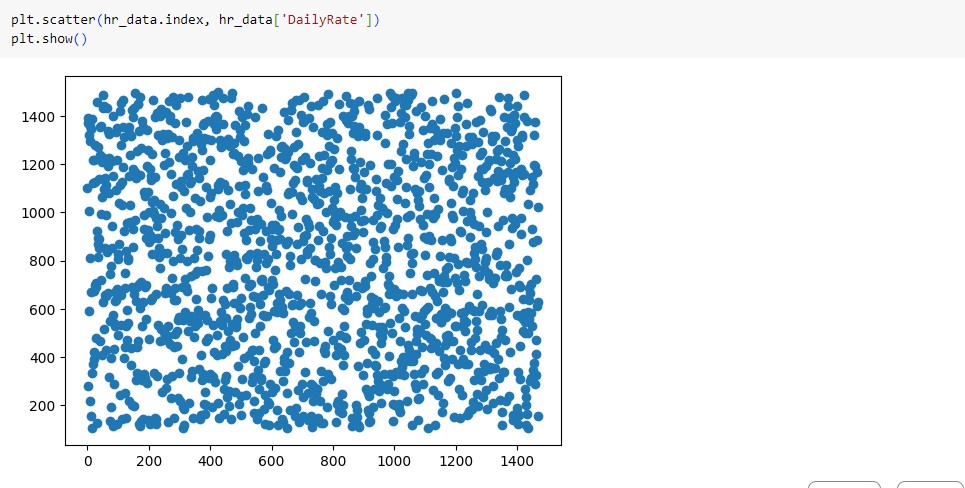
A standard deviation is a measure of how dispersed the data is in relation to the mean.



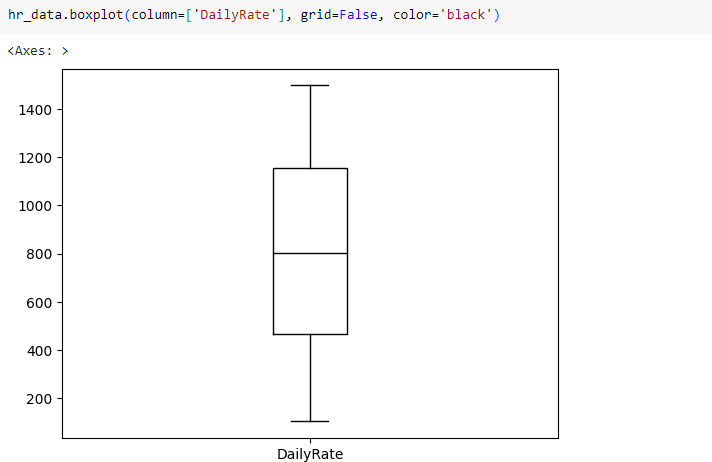
### CREATING FREQUENCY TABLE OF THE POINTS:

**PLOTTING A SCATTER PLOT**

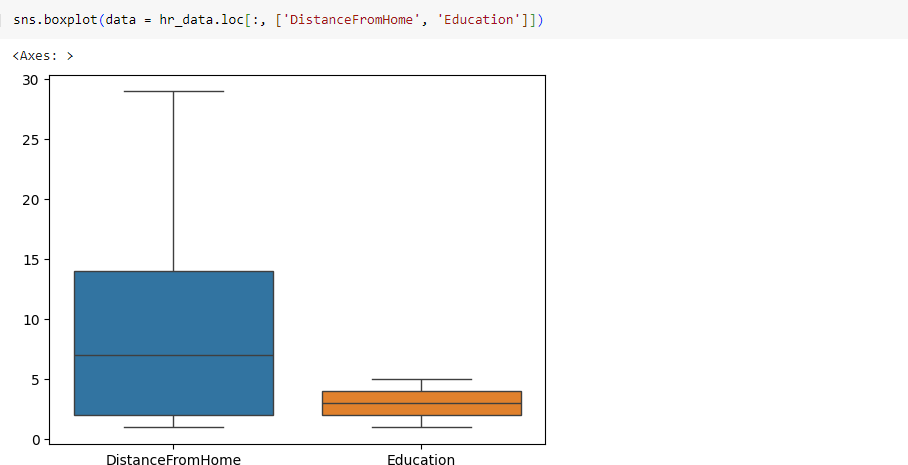
A scatter plot is used to represent the values for two variables in a two-dimensional data-set. The below graph of scatterplot shows the daily rate of the employ.



### PLOTTING A BOXPLOT:

Box plots are used to show distributions of numeric data values, especially when you want to compare them between multiple groups.

### PLOTTING BOXPLOT FOR 2 COLUMNS:



**Observation:**

From the above graph we can tell that number of employees who travel more distance are more in count and most of them are well educated up to graduation or masters level.

### BOXPLOT FOR NUM COLUMNS:

**Observation:**

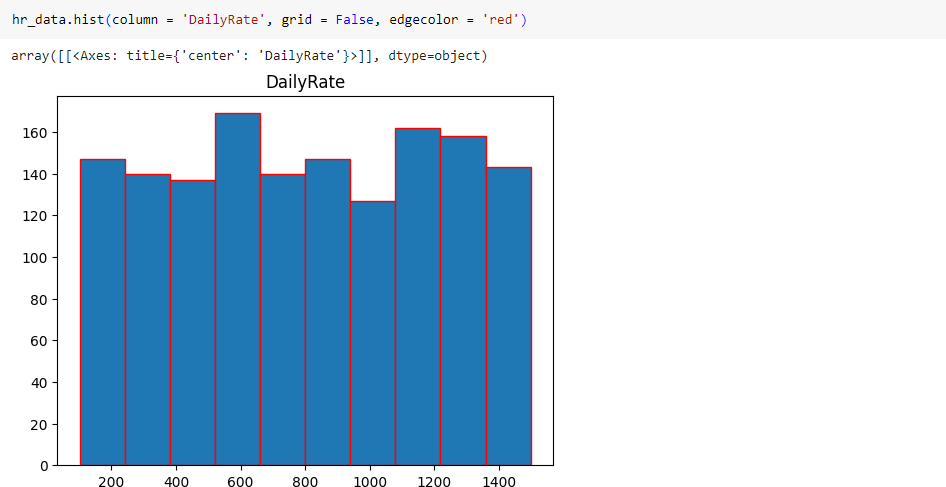
1.The age range of the employees in the company is from 30-40 years. 2.Employee rate of the company is from 500-1500.

3.Enviroment satisfaction is not bad for most of the employees. 4.Working hour rate of employees is from 50-85 hours per week. 5.Job involvement of the employees is not much.

6.But when it comes to Job Satisfaction most of the employees are satisfied with it. 7.Monthly income range of the employees from 2.5k to 7.5k.

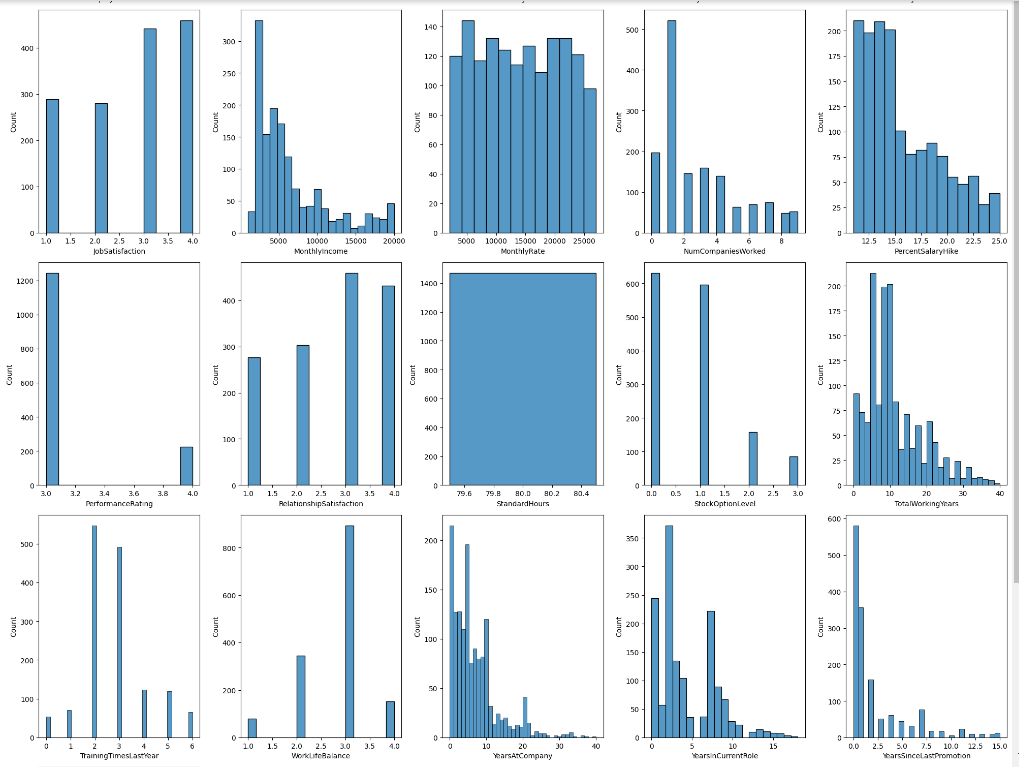
1. Most of the employees are experienced as they have already worked in to 1-4 company before joining this department
2. Most of the employees work for a period of 4-8 years.
3. The range of years for current job roles of employees and the range for years with current manager is same and it is 2.4 to 5.4 years

### PLOTTING HISTPLOT:

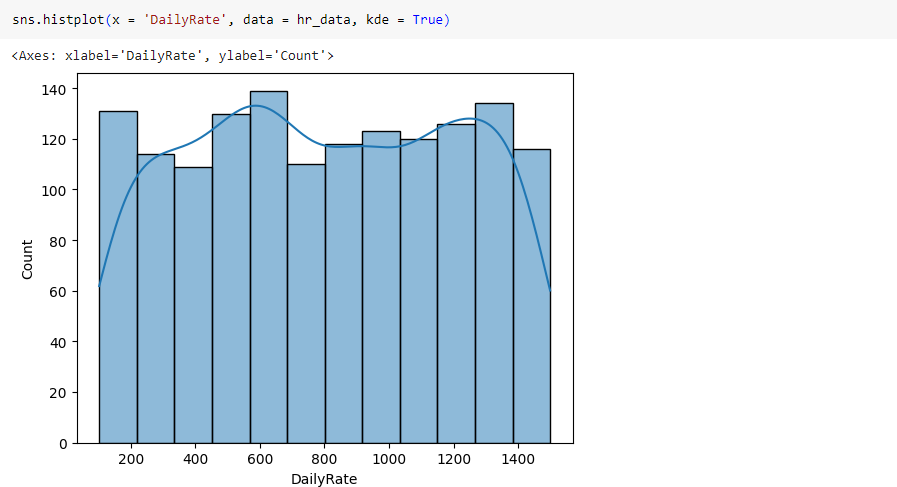
Histograms are visual representation of a data set which show how often each value in the data set occurs. The values are grouped in to bins along the x-axis. The height of the bar indicates how many values of the data set fall in to the bin.

### HISTPLOT FOR NUMERIC VALUES:





**HISTPLOT AND KDEPLOT:**



### KDE PLOT:

This graph shows us the combination of age and the education level of employees. By this graph we can say that most of the employee are in the age of 30 years and are having 3rd rank of education.

### STRIP PLOT:

A strip plot is a single-axis scatter plot that is used visualize the distribution of many individual one- dimensional values. The values are plotted as dots along one unique axis.

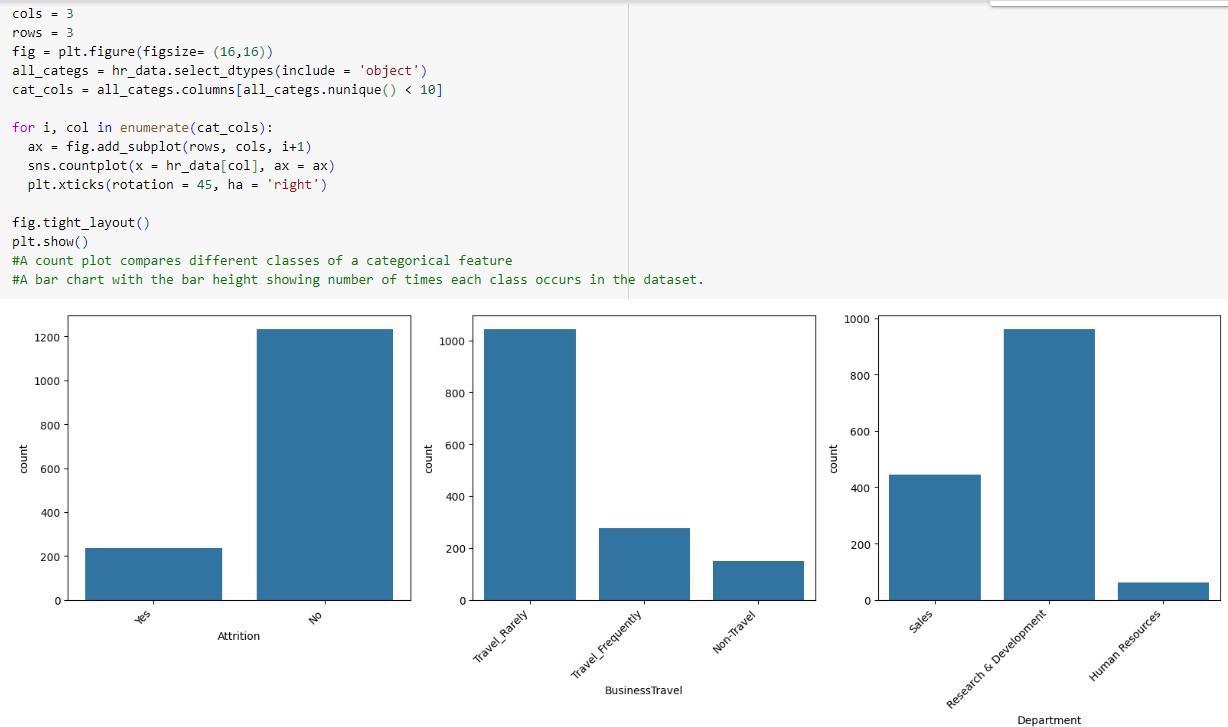
### VIOLON PLOT:

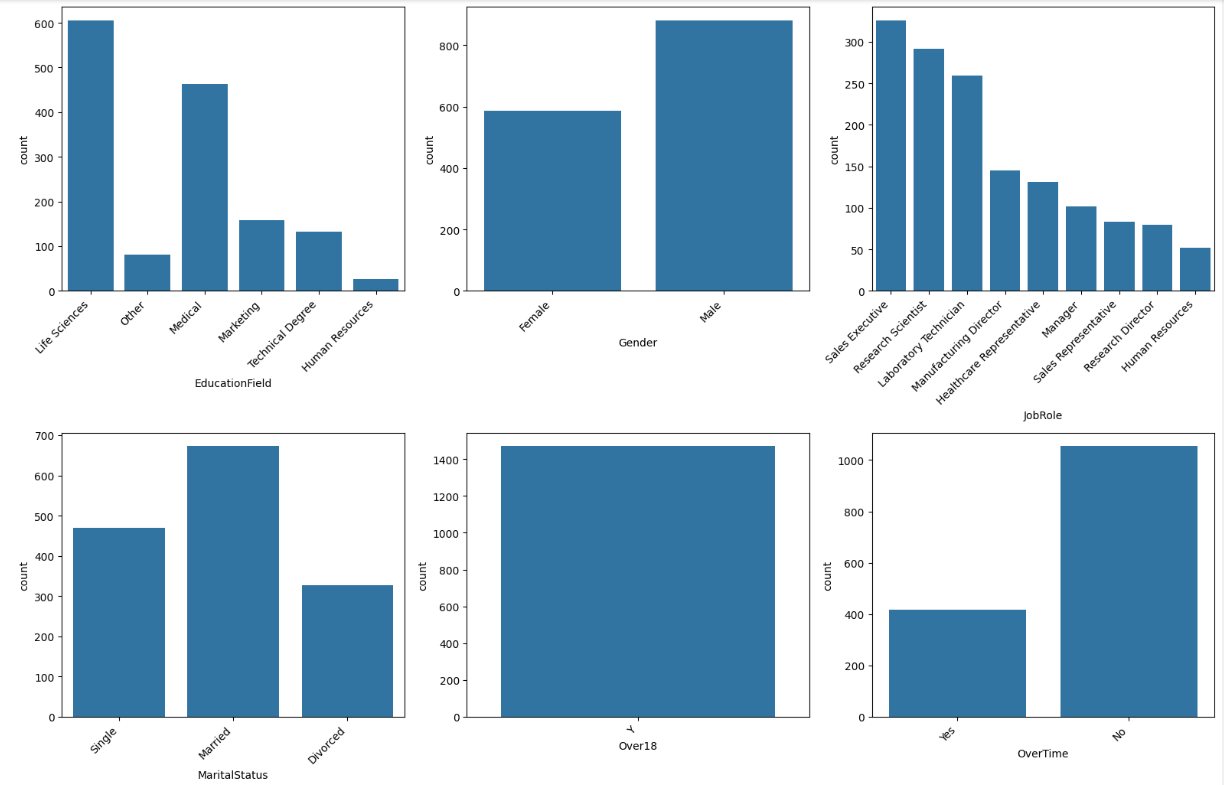
Violin plots are used when you want to observe the distribution of numeric data, and are especially useful when you want to make a comparison of distributions between multiple groups. The peaks, valleys, and tails of each group’s density curves can be compared to see where groups are similar or different.

### Observation:

Most of the employees are having 3rd rank education.

### BAR CHART:





**Observation:**

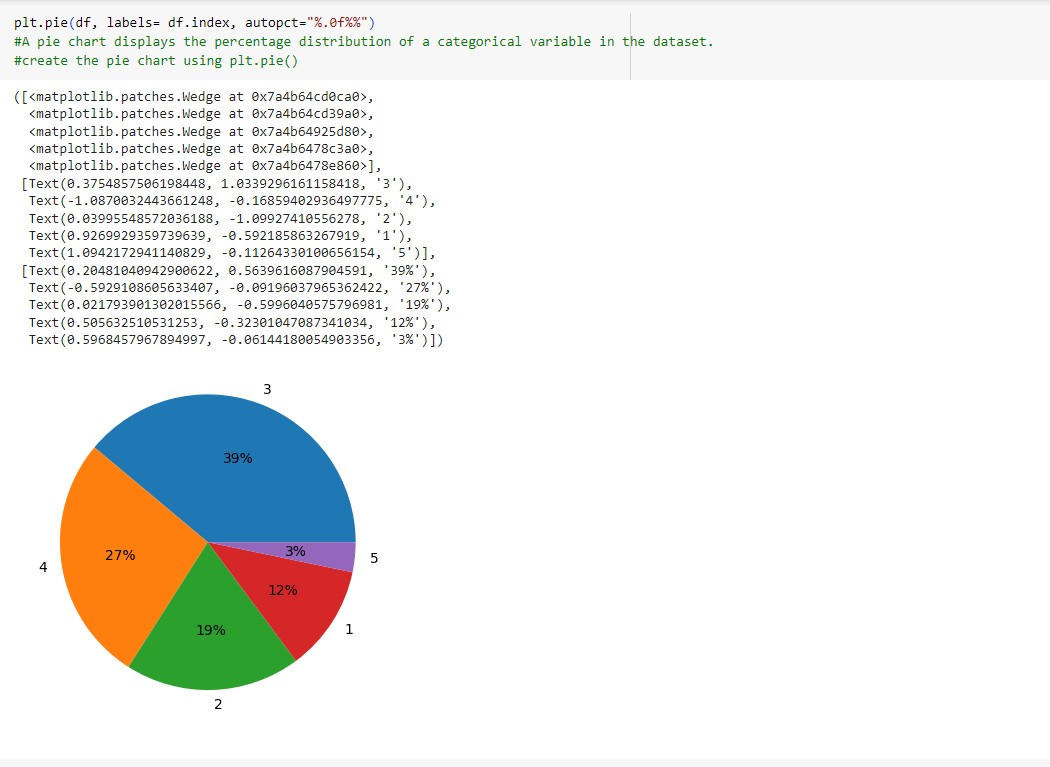
1. Most of the workers work peacefully.
2. Most of the employee travel rarely when it comes to business travels.
3. Research and development has the most investment of the company.
4. Most of the employees have life Sciences as their Education field.
5. There are more Male workers when compare to Female workers.
6. When it comes to Marital Status the most of the employees are married and then next comes are single and finally are divorced.
7. All the employees are over 18(age).
8. Most of the employees do not prefer over time. only 400 workers work overtime.

### PIE CHART:

Education indifferent fields represented in form of a pie chart. Blue: life sciences

Purple: technical degree Red: marketing

Green: medical Orange: other

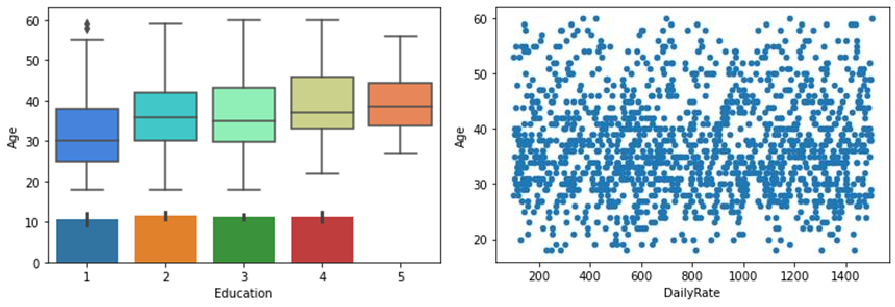


## BIVARIATE DATA ANALYSIS

Bivariate analysis is preformed to find the relationship between each variable in the dataset and the target variable of interest or using 2 variables and finding the relationship between them.

Categorical vs continuous (numerical) columns: Boxplot, Bar plot Continuous vs continuous columns: scatter plot

Categorical vs Categorical columns: Group By (Sum, Count, Value Count)



Box plot between Education vs Age.

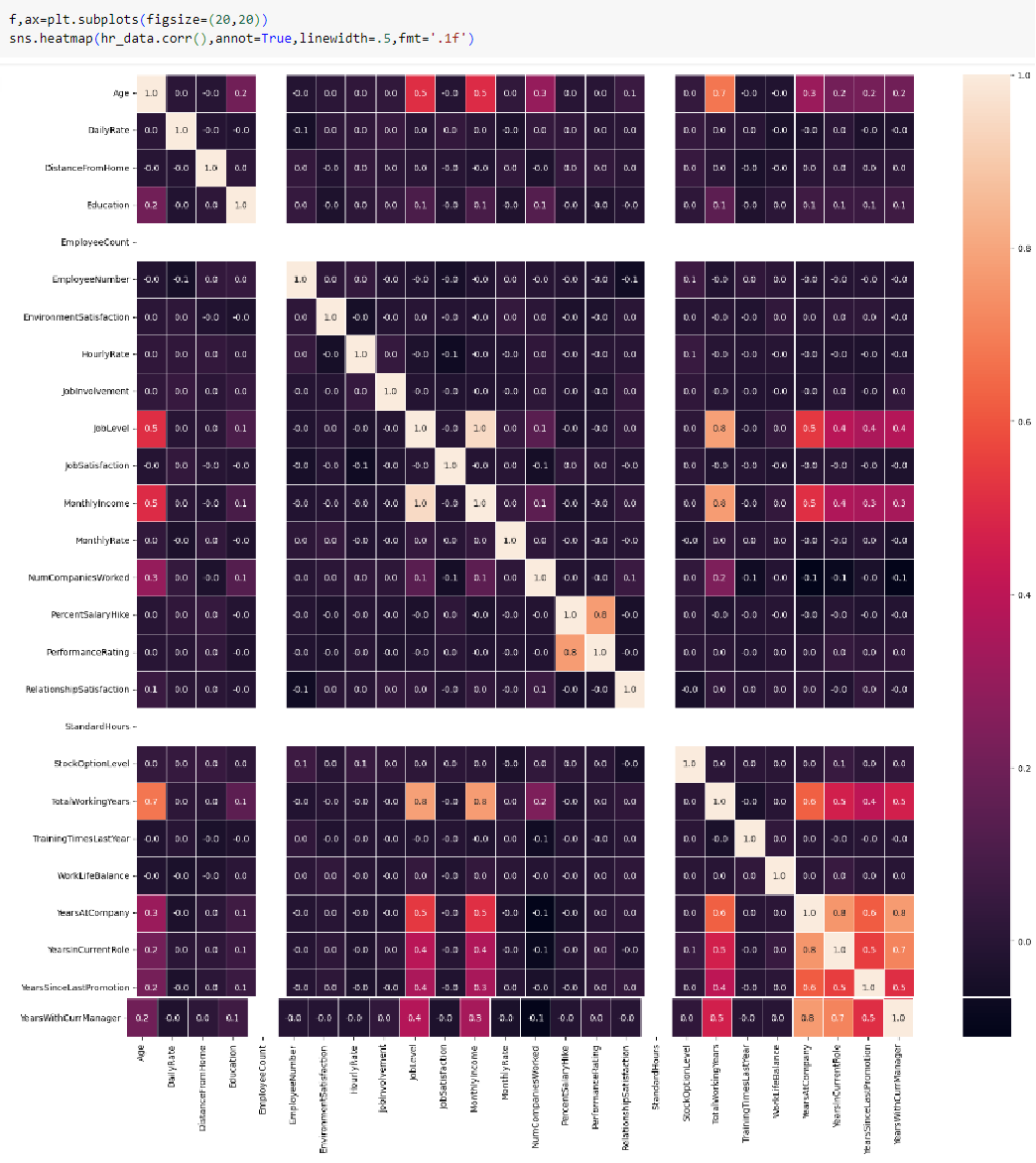
Where employees are also compared with their age and working rate.

## MULTIVARIATE DATA ANALYSIS

Multivariate analysis is used to describe analyses of data where there are multiple variables or observations for each unit or individual.

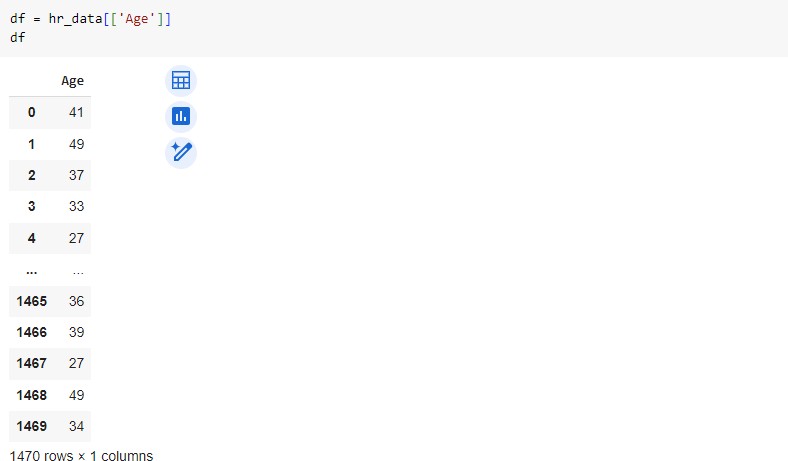
### HEATMAP:

A heatmap is used as a visual representation of data where values in a matrix are represented as colors. It provides a quick and intuitive way to understand the distribution, patterns, and relationships within the data. Heatmaps are particularly useful for identifying correlations between variables in a dataset.

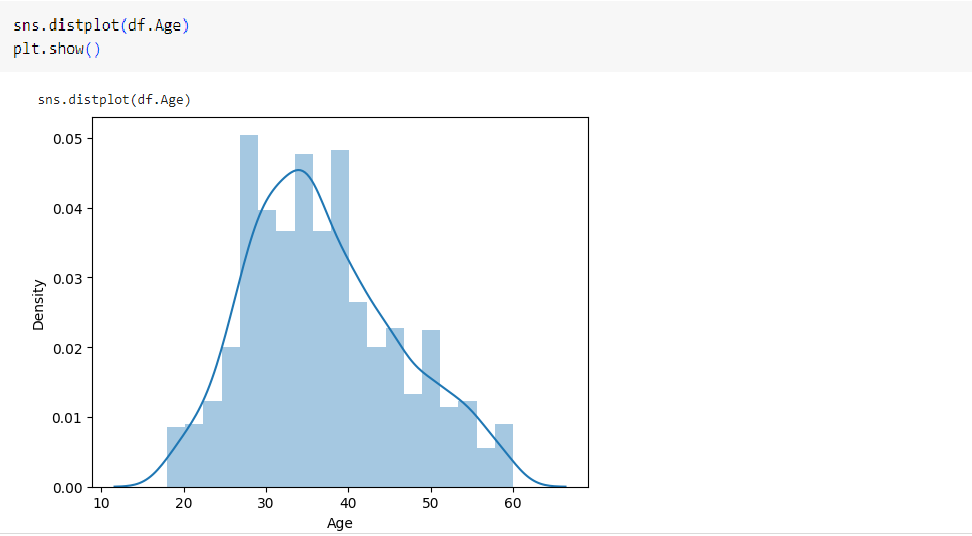


### PLOT DISTRIBUTION:



**USING AGE COLUMN:**

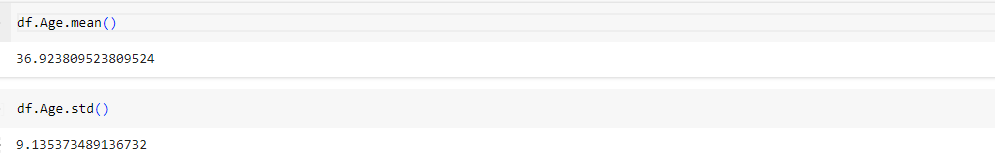
### DISPLOT:

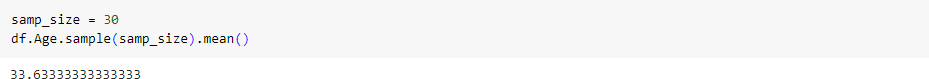


**Observations:**

1.Age group of 26-28 years are most active. 2.Most of the workers are from the age 26-40. 3.Mostly employees from age 30-40 are promoted.

### MEAN AND STANDARD DEVIATION OF THE AGE COLUMN:



**CALCULATING MEAN BY SAMPLE SIZE:**

### CONCLUSION:

In conclusion, our journey through exploratory data analysis (EDA) in the realm of HR employee attrition has illuminated the intricate dynamics underlying this critical organizational challenge. By harnessing the power of tools like pandas, NumPy, seaborn, and matplotlib.pyplot, we embarked on a comprehensive exploration of the dataset, unraveling hidden patterns and insights. Additionally,

1.Most of the employees left their jobs after getting a promotion. 2.Probably after 4-8 years of time employee quits the job.

3.sales representative job roles have the highest attrition rate.

### REFERENCES:

* **https://jovian.ml/learn/data-analysis-with-python-zero-to-pandas**

### https://matplotlib.org/3.1.1/index.html

* **https://pandas.pydata.org/pandas-docs/stable/index.html**

### https://[www.geeksforgeeks.org/](http://www.geeksforgeeks.org/)

* **https://seaborn.pydata.org/examples/index.html**

### PROJECT CODE:

[**https://drive.google.com/drive/folders/17OvyGEzCQI\_XpsAVgrZdARRZkWfaAWcC?usp=dri**](https://drive.google.com/drive/folders/17OvyGEzCQI_XpsAVgrZdARRZkWfaAWcC?usp=drive_link)[**ve\_link**](https://drive.google.com/drive/folders/17OvyGEzCQI_XpsAVgrZdARRZkWfaAWcC?usp=drive_link)