|  |  |
| --- | --- |
|  | Data set Exploration Part 1 |
|  |  |
|  |  |

**Data set Introduction:**

IBM designed a fictional dataset named the Employee Attrition dataset with the aim of identifying patterns in employee attrition across different fields. Employee attrition is the process through which workers leave their positions, whether by retirement or resignation, for their own reasons. With the use of this dataset, we can learn important details about possible reasons for the loss of staff inside the company.

You can access the dataset on Kaggle using the following reference link: <https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset>

So, in our dataset, there are number employee related features such as information on demographics (like age and gender), work-related data (such as job role and department), compensation-related data (such as salary and bonuses), and employee satisfaction data (such as job satisfaction and work-life balance).

Analyzing our dataset, it can help companies to come up with smart choices and put plans forward to increase employee retention and satisfaction by studying our dataset to better understand the factors causing loss of staff. It is a useful tool for those in data analysts who are interested in learning more about organizational trends and staff satisfaction.

**Dataset Description:**

***Source:***

The dataset we use for this study is made by the IBM to explore variations in employee attrition. It was created to emulate a normal HR analytics dataset but is not based on actual employee data.

https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

***Motivation:***

The aim of my studying this dataset is to find out more about those factors that could affect employee attrition within a company. A company's how they perform, staff satisfaction, and overall efficiency can be greatly affected by employee attrition, which is the term for when employees leave the job. This dataset's analysis allows one to study the relationships between different employee-related factors and attrition rates, helping companies make choices based on data that increase employees retention and job satisfaction.

**Data Dictionary:**

Here is a data dictionary for the dataset, including statistical variable types, descriptions, and, where applicable, ranges and limitations:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| # | **Variable Name** | **Data Type** | **Description** | **Range / Limitations** |
| 0 | Age | Numeric | Age of the employee | Typically ranges from 18 to 65 years. |
| 1 | Attrition | Categorical | Employee attrition (Yes/No) | Binary: 'Yes' for attrition, 'No' for no attrition |
| 2 | BusinessTravel | Categorical | Frequency of business travel | Categorical: 'Travel\_Rarely', 'Travel\_Frequently', 'Non-Travel'. |
| 3 | DailyRate | Numeric | Daily rate of pay | Varies based on the company's pay structure. |
| 4 | Department | Categorical | Department where the employee works | Categorical: 'Sales', 'Research & Development', 'Human Resources' |
| 5 | DistanceFromHome | Numeric | Distance from home to workplace | Measured in miles. |
| 6 | Education | Numeric | Employee's education level | Numeric scale: 1 (Below College) to 5 (Doctorate). |
| 7 | EducationField | Categorical | Field of education of the employee | Categorical: 'Life Sciences', 'Medical', 'Marketing', etc. |
| 9 | Gender | Categorical | Gender of the employee | Categorical: 'Male' or 'Female'. |
| 10 | HourlyRate | Numeric | Hourly rate of pay | Varies based on the company's pay structure. |
| 12 | JobLevel | Numeric | Level of the employee's job | Numeric scale: 1 (Entry Level) to 5 (Manager/Director). |
| 13 | JobRole | Numeric | Role or position of the employee | Categorical: 'Sales Executive', 'Research Scientist', etc. |
| 14 | JobSatisfaction | Numeric | Satisfaction level with the job | Numeric scale: 1 (Low) to 4 (Very High). |
| 15 | MaritalStatus | Categorical | Marital status of the employee | Categorical: 'Single', 'Married', 'Divorced'. |
| 16 | MonthlyIncome | Numeric | Monthly income of the employee | Varies widely based on job roles, experience, etc. |
| 18 | NumCompaniesWorked | Numeric | Number of companies the employee has worked for | Typically ranges from 0 (current company is first) to a higher value. |
| 19 | OverTime | Categorical | Overtime work (Yes/No) | Binary: 'Yes' for overtime, 'No' for no overtime. |
| 20 | PerformanceRating | Numeric | Performance rating of the employee | Numeric scale: 3 (Average) to 4 (Excellent). |
| 21 | MonthlyRate | Numeric | Monthly rate of pay | Varies based on the company's pay structure. |
| 22 | PercentSalaryHike | Numeric | Percentage salary hike in the last year | Measured as a percentage. |
| 23 | RelationshipSatisfaction | Numeric | Satisfaction level with work relationships | Numeric scale: 1 (Low) to 4 (Very High) |
| 24 | StockOptionLevel | Numeric | Level of stock options held by the employee | Numeric scale: 0 (None) to 3 (High). |
| 25 | TotalWorkingYears | Numeric | Total years of work experience | Varies based on employee's career duration |
| 26 | TrainingTimesLastYear | Numeric | Number of training times attended last year | Typically ranges from 0 to a higher value. |
| 27 | WorkLifeBalance | Numeric | Work-life balance satisfaction | Numeric scale: 1 (Bad) to 4 (Very Good). |
| 28 | YearsAtCompany | Numeric | Years at the current company | Varies based on employee's tenure. |
| 29 | YearsInCurrentRole | Numeric | Years in the current job role | Varies based on employee's role stability. |
| 30 | YearsSinceLastPromotion | Numeric | Years since the last promotion | Varies based on promotion history. |
| 31 | YearsWithCurrManager | Numeric | Years with the current manager | Varies based on managerial relationship. |

**Limitations:**

This dataset is fictitious, it could differ from employee data from the real world.

The dataset might be lacking certain details which would be essential to an in-depth study of employee attrition. While some variables have clear numerical ranges, others are categorical or looked at using scales of Likert, which makes them less suitable for specific statistical studies.

And every conclusion made from this dataset should be set against its probable limits and fictional identity.

**Data cleaning**

In the initial we got a wide variety of columns containing a variety of employee-related data in the first raw dataset.

However, we made an effort to decrease the number of columns during the data cleansing step in order to create a more focuses and helpful dataset for answering particular research questions. This cleaned dataset was carefully selected to better match our research objectives, making sure that the resulting analysis would be more accurate and interesting.

In addition, we carefully checked every value for possible issues with missing and incorrect values during the data cleaning process. The analysis showed that multiple variables need to be taken into account.

Firstly, the "Age" variable had 55 missing values, which led to the elimination of those rows from the dataset.

Similarly, the "DailyRate" column also had 14 missing values, which were fixed by removing the correct rows. On the other hand, variables such as "BusinessTravel," "Department," and "MaritalStatus" did not have any missing values. The 54 missing values in "DistanceFromHome" were clean by removing the correct rows.

In case of "MaritalStatus," there are 3 missing values were handled by row removal of that row.

In this way by removing the few columns and rows with missing value we clean our dataset.

**Assumptions that may need to our dataset:**

In our analysis of the dataset on Employee attrition, I found some important assumptions and considerations come to light.

First off, the dataset is fictional and was made up by IBM; therefore, the exact collection process and timing cannot be verified. We have to understand that the dataset was created for analytical purposes and might not accurately represent employee data in the actual world.

Additionally, it is unclear from the dataset whether it is a collection from a wider group or the entire group of employees, which has effects on applying our results.

Furthermore, we have to measure units of like "Age," "MonthlyIncome," and "HourlyRate" need to be defined in order to perform useful analysis, yet they are not made clear. Although the fact that we eliminated missing values by removing them, this assumes that our results are not highly skewed by the missing data and that they are absent randomly.

Lastly, we assume that given the dataset's analytical aim, we assume that it passed detailed quality checks and validation. These assumptions give our study a basis, but they also highlight how important it is to be aware of any possible limitations and uncertainties when we analyse employee attrition trends using this information.

**FINER Questions**

* What is the age distribution between males and females? Are there any significant discrepancies?
* What is the average salary by gender? What are the number of employees by Gender in each department?
* What is the relationship between the level of job (Job Level) and the performance rating (Performance Rating) of employees, and does this relationship vary across different job roles (Job Role)?
* Is there a significant difference in monthly income (MonthlyIncome) between employees who work overtime (OverTime = Yes) and those who do not (OverTime = No)?

# Demonstrated Tracking

* This dataset contains 1471 rows of records and 36 data points(columns).
* In this dataset since there is no timeline mentioned it is assumed that the data set is for one fiscal year or calendar year.
* Dataset is just a fictional data provided by IBM which has very much resemblance to real dataset.