hranalytics

May 16, 2024

[1]:

*#Importing necessary libary and dataset*

**import pandas as pd**

df = pd.read\_csv('HR Data.csv') print(df)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Age | Attrition | BusinessTravel | DailyRate |  | Department | \ |
| 0 | 41 | Yes | Travel\_Rarely | 1102 |  | Sales |  |
| 1 | 49 | No | Travel\_Frequently | 279 | Research & | Development |  |
| 2 | 37 | Yes | Travel\_Rarely | 1373 | Research & | Development |  |
| 3 | 33 | No | Travel\_Frequently | 1392 | Research & | Development |  |
| 4 | 27 | No | Travel\_Rarely | 591 | Research & | Development |  |
| … … |  | … | … … |  | … |  |  |
| 1465 | 36 | No | Travel\_Frequently | 884 | Research & | Development |  |
| 1466 | 39 | No | Travel\_Rarely | 613 | Research & | Development |  |
| 1467 | 27 | No | Travel\_Rarely | 155 | Research & | Development |  |
| 1468 | 49 | No | Travel\_Frequently | 1023 |  | Sales |  |
| 1469 | 34 | No | Travel\_Rarely | 628 | Research & | Development |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | DistanceFromHome | Education | EducationField | EmployeeCount | \ |
| 0 | 1 | 2 | Life Sciences | 1 |  |
| 1 | 8 | 1 | Life Sciences | 1 |  |
| 2 | 2 | 2 | Other | 1 |  |
| 3 | 3 | 4 | Life Sciences | 1 |  |
| 4 | 2 | 1 | Medical | 1 |  |
| … | … | … | … | … |  |
| 1465 | 23 | 2 | Medical | 1 |  |
| 1466 | 6 | 1 | Medical | 1 |  |
| 1467 | 4 | 3 | Life Sciences | 1 |  |
| 1468 | 2 | 3 | Medical | 1 |  |
| 1469 | 8 | 3 | Medical | 1 |  |

EmployeeNumber … RelationshipSatisfaction StandardHours \

0 1 … 1 80

1 2 … 4 80

2 4 … 2 80

3 5 … 3 80

4 7 … 4 80

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| … | … … | … |  | … |  | |
| 1465 | 2061 … |  | 3 |  | 80 | |
| 1466 | 2062 … |  | 1 |  | 80 | |
| 1467 | 2064 … |  | 2 |  | 80 | |
| 1468 | 2065 … |  | 4 |  | 80 | |
| 1469 | 2068 … |  | 1 |  | 80 | |
|  | StockOptionLevel | TotalWorkingYears | TrainingTimesLastYear | | | \ |
| 0 | 0 | 8 | 0 | | |  |
| 1 | 1 | 10 | 3 | | |  |
| 2 | 0 | 7 | 3 | | |  |
| 3 | 0 | 8 | 3 | | |  |
| 4  … 1465 | 1  …  1 | 6  …  17 | 3  …  3 | | |  |
| 1466 | 1 | 9 | 5 | | |  |
| 1467 | 1 | 6 | 0 | | |  |
| 1468 | 0 | 17 | 3 | | |  |
| 1469 | 0 | 6 | 3 | | |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | WorkLifeBalance | YearsAtCompany | | YearsInCurrentRole | \ |
| 0 | 1 | 6 | | 4 |  |
| 1 | 3 | 10 | | 7 |  |
| 2 | 3 | 0 | | 0 |  |
| 3 | 3 | 8 | | 7 |  |
| 4  … 1465 | 3  …  3 | 2  …  5 | | 2  …  2 |  |
| 1466 | 3 | 7 | | 7 |  |
| 1467 | 3 | 6 | | 2 |  |
| 1468 | 2 | 9 | | 6 |  |
| 1469 | 4 | 4 | | 3 |  |
|  | YearsSinceLastPromotion | | YearsWithCurrManager | | |
| 0 | 0 | | 5 | | |
| 1 | 1 | | 7 | | |
| 2 | 0 | | 0 | | |
| 3 | 3 | | 0 | | |
| 4 | 2 | | 2 | | |
| … | … | | … | | |
| 1465 | 0 | | 3 | | |
| 1466 | 1 | | 7 | | |
| 1467 | 0 | | 3 | | |
| 1468 | 0 | | 8 | | |
| 1469 | 1 | | 2 | | |

[1470 rows x 35 columns]

[30]:

print("First few rows of the dataset: ") print(df.head())

First few rows of the dataset:

Age Attrition BusinessTravel DailyRate Department \

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 0 | 41 | Yes | Travel\_Rarely | 1102 | Sales |
| 1 | 49 | No | Travel\_Frequently | 279 | Research & Development |
| 2 | 37 | Yes | Travel\_Rarely | 1373 | Research & Development |
| 3 | 33 | No | Travel\_Frequently | 1392 | Research & Development |
| 4 | 27 | No | Travel\_Rarely | 591 | Research & Development |

DistanceFromHome Education EducationField EmployeeCount EmployeeNumber \

1. 1 2 Life Sciences 1 1
2. 8 1 Life Sciences 1 2
3. 2 2 Other 1 4
4. 3 4 Life Sciences 1 5
5. 2 1 Medical 1 7

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| … | RelationshipSatisfaction | StandardHours | StockOptionLevel \ | |
| 0 … | 1 | 80 | 0 | |
| 1 … | 4 | 80 | 1 | |
| 2 … | 2 | 80 | 0 | |
| 3 … | 3 | 80 | 0 | |
| 4 … | 4 | 80 | 1 | |
| TotalWorkingYears TrainingTimesLastYear WorkLifeBalance YearsAtCompany | | | | |
| 0 | 8 | 0 | 1 | 6 |
| 1 | 10 | 3 | 3 | 10 |
| 2 | 7 | 3 | 3 | 0 |
| 3 | 8 | 3 | 3 | 8 |
| 4 | 6 | 3 | 3 | 2 |

\

[14]:

YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager

0 4 0 5

1 7 1 7

2 0 0 0

3 7 3 0

4 2 2 2

[5 rows x 35 columns]

print("Summary Statistics of the Dataset") print(df.describe())

Summary Statistics of the Dataset

Age DailyRate DistanceFromHome Education EmployeeCount \ count 1470.000000 1470.000000 1470.000000 1470.000000 1470.0

mean 36.923810 802.485714 9.192517 2.912925 1.0

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| std | 9.135373 403.509100 | 8.106864 | 1.024165 | 0.0 |
| min | 18.000000 102.000000 | 1.000000 | 1.000000 | 1.0 |
| 25% | 30.000000 465.000000 | 2.000000 | 2.000000 | 1.0 |
| 50% | 36.000000 802.000000 | 7.000000 | 3.000000 | 1.0 |
| 75% | 43.000000 1157.000000 | 14.000000 | 4.000000 | 1.0 |
| max | 60.000000 1499.000000 | 29.000000 | 5.000000 | 1.0 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | EmployeeNumber | EnvironmentSatisfaction | | HourlyRate | JobInvolvement | | \ |
| count | 1470.000000 | 1470.000000 | | 1470.000000 | 1470.000000 | |  |
| mean | 1024.865306 | 2.721769 | | 65.891156 | 2.729932 | |  |
| std | 602.024335 | 1.093082 | | 20.329428 | 0.711561 | |  |
| min | 1.000000 | 1.000000 | | 30.000000 | 1.000000 | |  |
| 25% | 491.250000 | 2.000000 | | 48.000000 | 2.000000 | |  |
| 50% | 1020.500000 | 3.000000 | | 66.000000 | 3.000000 | |  |
| 75% | 1555.750000 | 4.000000 | | 83.750000 | 3.000000 | |  |
| max | 2068.000000 | 4.000000 | | 100.000000 | 4.000000 | |  |
|  | JobLevel … | RelationshipSatisfaction | | StandardHours \ | | | |
| count | 1470.000000 … | 1470.000000 | | 1470.0 | | | |
| mean | 2.063946 … | 2.712245 | | 80.0 | | | |
| std | 1.106940 … | 1.081209 | | 0.0 | | | |
| min | 1.000000 … | 1.000000 | | 80.0 | | | |
| 25% | 1.000000 … | 2.000000 | | 80.0 | | | |
| 50% | 2.000000 … | 3.000000 | | 80.0 | | | |
| 75% | 3.000000 … | 4.000000 | | 80.0 | | | |
| max | 5.000000 … | 4.000000 | | 80.0 | | | |
|  | StockOptionLevel | TotalWorkingYears | TrainingTimesLastYear | | | \ | |
| count | 1470.000000 | 1470.000000 | 1470.000000 | | |  | |
| mean | 0.793878 | 11.279592 | 2.799320 | | |  | |
| std | 0.852077 | 7.780782 | 1.289271 | | |  | |
| min | 0.000000 | 0.000000 | 0.000000 | | |  | |
| 25% | 0.000000 | 6.000000 | 2.000000 | | |  | |
| 50% | 1.000000 | 10.000000 | 3.000000 | | |  | |
| 75% | 1.000000 | 15.000000 | 3.000000 | | |  | |
| max | 3.000000 | 40.000000 | 6.000000 | | |  | |
| WorkLifeBalance YearsAtCompany YearsInCurrentRole \ | | | | | | | |
| count | 1470.000000 | 1470.000000 | 1470.000000 | | | | |
| mean | 2.761224 | 7.008163 | 4.229252 | | | | |
| std | 0.706476 | 6.126525 | 3.623137 | | | | |
| min | 1.000000 | 0.000000 | 0.000000 | | | | |
| 25% | 2.000000 | 3.000000 | 2.000000 | | | | |
| 50% | 3.000000 | 5.000000 | 3.000000 | | | | |
| 75% | 3.000000 | 9.000000 | 7.000000 | | | | |
| max | 4.000000 | 40.000000 | 18.000000 | | | | |

YearsSinceLastPromotion YearsWithCurrManager

|  |  |  |
| --- | --- | --- |
| count | 1470.000000 | 1470.000000 |
| mean | 2.187755 | 4.123129 |
| std | 3.222430 | 3.568136 |
| min | 0.000000 | 0.000000 |
| 25% | 0.000000 | 2.000000 |
| 50% | 1.000000 | 3.000000 |
| 75% | 3.000000 | 7.000000 |
| max | 15.000000 | 17.000000 |

[8 rows x 26 columns]

[17]:

*#Finding missing values from datasets*

print(df.isnull().sum())

Missing Values in the Dataset Age 0

Attrition 0

BusinessTravel 0

DailyRate 0

Department 0

DistanceFromHome 0

Education 0

EducationField 0

EmployeeCount 0

EmployeeNumber 0

EnvironmentSatisfaction 0

Gender 0

HourlyRate 0

JobInvolvement 0

JobLevel 0

JobRole 0

JobSatisfaction 0

MaritalStatus 0

MonthlyIncome 0

MonthlyRate 0

NumCompaniesWorked 0

Over18 0

OverTime 0

PercentSalaryHike 0

PerformanceRating 0

RelationshipSatisfaction 0

StandardHours 0

StockOptionLevel 0

TotalWorkingYears 0

TrainingTimesLastYear 0

WorkLifeBalance 0

YearsAtCompany 0

YearsInCurrentRole 0

[27]:

YearsSinceLastPromotion 0

YearsWithCurrManager 0

dtype: int64

*#Eliminate reduntant entries*

df.drop\_duplicates(inplace = **True**)

[26]:

*#Check for NaN values and drop rows with NaN values*

df\_cleaned = df\_cleaned.dropna()

[24]:

print(df.head())

Age Attrition BusinessTravel DailyRate Department \

1. 41 Yes Travel\_Rarely 1102 Sales
2. 49 No Travel\_Frequently 279 Research & Development
3. 37 Yes Travel\_Rarely 1373 Research & Development
4. 33 No Travel\_Frequently 1392 Research & Development
5. 27 No Travel\_Rarely 591 Research & Development

DistanceFromHome Education EducationField EmployeeCount EmployeeNumber \

1. 1 2 Life Sciences 1 1
2. 8 1 Life Sciences 1 2
3. 2 2 Other 1 4
4. 3 4 Life Sciences 1 5
5. 2 1 Medical 1 7

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| … | RelationshipSatisfaction | StandardHours | StockOptionLevel \ | |
| 0 … | 1 | 80 | 0 | |
| 1 … | 4 | 80 | 1 | |
| 2 … | 2 | 80 | 0 | |
| 3 … | 3 | 80 | 0 | |
| 4 … | 4 | 80 | 1 | |
| TotalWorkingYears TrainingTimesLastYear WorkLifeBalance YearsAtCompany | | | | |
| 0 | 8 | 0 | 1 | 6 |
| 1 | 10 | 3 | 3 | 10 |
| 2 | 7 | 3 | 3 | 0 |
| 3 | 8 | 3 | 3 | 8 |
| 4 | 6 | 3 | 3 | 2 |

\

YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager

0 4 0 5

1 7 1 7

2 0 0 0

3 7 3 0

4 2 2 2

[5 rows x 35 columns]

[11]:

[ ]: