In [1]: import pandas as pd import seaborn as sns import matplotlib.pyplot as plt In [3]: df=pd.read csv('WA Fn-UseC -HR-Employee-Attrition.csv') BusinessTravel DailyRate Department DistanceFromHome Education EducationField EmployeeCount Em Out[3]: Age Attrition 0 1102 Sales 1 2 Life Sciences 1 41 Yes Travel_Rarely Research & 49 Travel_Frequently 279 8 1 Life Sciences Development Research & 2 2 Other 2 37 Yes Travel_Rarely 1373 1 Development Research & 3 33 Travel_Frequently 1392 3 4 Life Sciences No Development Research & 4 27 No Travel_Rarely 591 2 1 Medical 1 Development Research & 1465 36 Travel_Frequently 884 23 2 Medical 1 Development Research & 613 6 1466 39 No Travel_Rarely 1 Medical Development Research & Travel_Rarely 3 1467 27 155 Life Sciences No 4 Development 1468 49 No Travel_Frequently 1023 Sales 2 Medical Research & 1469 8 3 Medical 1 No Travel_Rarely Development 1470 rows × 35 columns df.head() In [5]: DailyRate EducationField Out[5]: Age Attrition EmployeeCount BusinessTravel Department DistanceFromHome Education **Emplo** 0 2 41 Travel_Rarely 1102 1 Yes Sales Life Sciences 1 Research & 1 49 No Travel_Frequently 279 8 1 Life Sciences Development Research & 2 2 2 37 1373 Yes Travel_Rarely Other 1 Development Research & 3 33 Travel_Frequently 1392 3 4 Life Sciences No Development Research & 4 27 No Travel_Rarely 591 2 1 Medical Development 5 rows × 35 columns df.tail() In [7]: Out[7]: Attrition BusinessTravel DailyRate Department DistanceFromHome Education EducationField EmployeeCount Em Age Research & 1465 36 No Travel_Frequently 884 23 2 Medical 1 Development Research & 1466 39 No Travel_Rarely 613 6 1 Medical Development Research & 3 1467 27 Nο Travel_Rarely 155 4 Life Sciences 1 Development 1468 49 1023 Sales 2 3 Medical No Travel_Frequently Research & 1469 34 8 3 Medical No Travel_Rarely Development 5 rows × 35 columns In [23]: df.info()

```
<class 'pandas.core.frame.DataFrame'>
        RangeIndex: 1470 entries, 0 to 1469
        Data columns (total 35 columns):
         #
            Column
                                       Non-Null Count Dtype
                                       -----
        0
            Age
                                      1470 non-null
                                                     int64
                                      1470 non-null
         1
             Attrition
                                                      object
                                      1470 non-null
         2
             BusinessTravel
                                                       object
         3
            DailyRate
                                      1470 non-null
                                                       int64
         4
            Department
                                      1470 non-null
                                                       object
         5
             DistanceFromHome
                                       1470 non-null
                                                       int64
                                       1470 non-null
         6
             Education
                                                       int64
             EducationField
                                      1470 non-null
                                                       object
                                       1470 non-null
         8
            EmployeeCount
                                                       int64
         9
             EmployeeNumber
                                       1470 non-null
                                                       int64
         10 EnvironmentSatisfaction 1470 non-null
                                                       int64
                                      1470 non-null
         11 Gender
                                                       object
         12 HourlyRate
                                       1470 non-null
                                                       int64
             JobInvolvement
                                       1470 non-null
                                                       int64
                                      1470 non-null
         14
            JobLevel
                                                       int64
                                      1470 non-null
         15 JobRole
                                                       object
         16
            JobSatisfaction
                                      1470 non-null
                                                       int64
                                       1470 non-null
         17
             MaritalStatus
                                                       object
         18 MonthlyIncome
                                      1470 non-null
                                                       int64
         19 MonthlyRate
                                      1470 non-null
                                                       int64
         20 NumCompaniesWorked
                                       1470 non-null
                                                       int64
         21 Over18
                                       1470 non-null
                                                       object
         22 OverTime
                                      1470 non-null
                                                       object
         23 PercentSalaryHike
                                     1470 non-null
                                                       int64
         24 PerformanceRating
                                       1470 non-null
                                                       int64
         25
            RelationshipSatisfaction 1470 non-null
                                                       int64
         26 StandardHours
                                      1470 non-null
                                                       int64
         27 StockOptionLevel
                                      1470 non-null
                                                       int64
            TotalWorkingYears 1470 non-null TrainingTimesLastYear 1470 non-null WorkLifeBalance 1470 non-null
         28 TotalWorkingYears
                                                       int64
         29
                                                       int64
         30 WorkLifeBalance
                                                       int64
         31 YearsAtCompany
                                       1470 non-null
                                                       int64
         32 YearsInCurrentRole
                                       1470 non-null
                                                       int64
         33 YearsSinceLastPromotion
                                       1470 non-null
                                                       int64
         34 YearsWithCurrManager
                                       1470 non-null
                                                       int64
        dtypes: int64(26), object(9)
        memory usage: 402.1+ KB
 In [ ]:
In [11]: df['BusinessTravel'].value counts(normalize=True)*100
Out[11]: BusinessTravel
                              70.952381
         Travel_Rarely
         Travel_Frequently
                              18.843537
         Non-Travel
                              10.204082
         Name: proportion, dtype: float64
In [13]: df['EducationField'].value counts(normalize=True)*100
Out[13]: EducationField
         Life Sciences
                             41.224490
         Medical
                             31.564626
         Marketing
                             10.816327
         Technical Degree
                              8.979592
         0ther
                              5.578231
         Human Resources
                              1.836735
         Name: proportion, dtype: float64
In [15]: df['Department'].value counts(normalize=True)*100
Out[15]: Department
         Research & Development
                                    65.374150
         Sales
                                   30.340136
                                    4.285714
         Human Resources
```

Name: proportion, dtype: float64

In [17]: df.isnull()

Out[17]:		Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	Emį
	0	False	False	False	False	False	False	False	False	False	
	1	False	False	False	False	False	False	False	False	False	
	2	False	False	False	False	False	False	False	False	False	
	3	False	False	False	False	False	False	False	False	False	
	4	False	False	False	False	False	False	False	False	False	
	1465	False	False	False	False	False	False	False	False	False	
	1466	False	False	False	False	False	False	False	False	False	
	1467	False	False	False	False	False	False	False	False	False	
	1468	False	False	False	False	False	False	False	False	False	
	1469	False	False	False	False	False	False	False	False	False	

1470 rows × 35 columns

In [21]: df.drop_duplicates()

21]:		Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	Er
	0	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	
	1	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	1	
	2	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	1	
	3	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	1	
	4	27	No	Travel_Rarely	591	Research & Development	2	1	Medical	1	
1	1465	36	No	Travel_Frequently	884	Research & Development	23	2	Medical	1	
1	1466	39	No	Travel_Rarely	613	Research & Development	6	1	Medical	1	
1	1467	27	No	Travel_Rarely	155	Research & Development	4	3	Life Sciences	1	
1	1468	49	No	Travel_Frequently	1023	Sales	2	3	Medical	1	
1	1469	34	No	Travel_Rarely	628	Research & Development	8	3	Medical	1	

1470 rows × 35 columns

In [38]: df1 = df.loc[:, df.nunique() > 1]
 df1

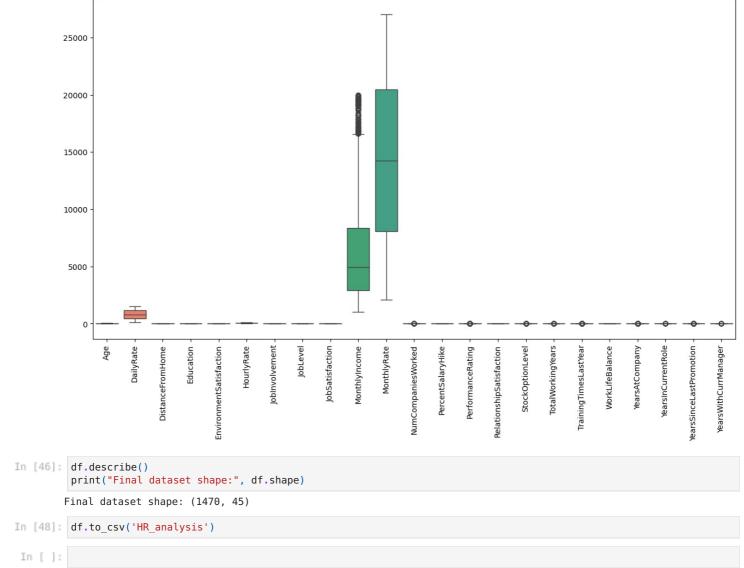
				2	2	0.4	3	2	
0	41	1102	1	2	2	94	3	2	
1	49	279	8	1	3	61	2	2	
2	37	1373	2	2	4	92	2	1	
3	33	1392	3	4	4	56	3	1	
4	27	591	2	1	1	40	3	1	
1465	36	884	23	2	3	41	4	2	
1466	39	613	6	1	4	42	2	3	
1467	27	155	4	3	2	87	4	2	
1468	49	1023	2	3	4	63	2	2	
1469	34	628	8	3	2	82	4	2	
1470 rc	ows × 4	15 columns							
4									
				T					
df2 = df2	pd.g	et_dummie	s(df, drop_first=	irue)					
					EnvironmentSatisfaction	HourlyRate	Joblnvolvement	JobLevel	JobSatisfa
					EnvironmentSatisfaction 2	HourlyRate 94	Jobinvolvement 3	JobLevel 2	JobSatisfa
df2	Age	DailyRate	DistanceFromHome	Education					JobSatisfa
o	Age 41	DailyRate 1102 279 1373	DistanceFromHome	Education 2	2	94 61 92	3	2 2	JobSatisfa
0 1	Age 41 49	DailyRate 1102 279 1373 1392	DistanceFromHome 1 8 2 3	Education 2	2	94 61	3 2 2 3	2 2 1	JobSatisfa
0 1 2	Age 41 49 37	DailyRate 1102 279 1373	DistanceFromHome 1 8 2	Education 2 1 2	2 3 4	94 61 92	3 2 2	2 2	JobSatisfa
0 1 2 3 4	Age 41 49 37 33	DailyRate 1102 279 1373 1392 591	DistanceFromHome 1 8 2 3 2	Education 2 1 2 4 1 1	2 3 4 4 1	94 61 92 56 40	3 2 2 3 3 	2 2 1 1 1	JobSatisfa
0 1 2 3 4 	Age 41 49 37 33 27 36	DailyRate 1102 279 1373 1392 591 884	DistanceFromHome 1 8 2 3 2 23	Education 2 1 2 4 1 2	2 3 4 4 1 	94 61 92 56 40 	3 2 2 3 3 	2 2 1 1 1 	JobSatisfa
0 1 2 3 4 1465 1466	Age 41 49 37 33 27 36 39	DailyRate 1102 279 1373 1392 591 884 613	DistanceFromHome 1 8 2 3 2 23 6	Education 2 1 2 4 1 2 1	2 3 4 4 1 3	94 61 92 56 40 41	3 2 2 3 3 4 2	2 2 1 1 2 3	JobSatisfa
0 1 2 3 4 1465 1466	Age 41 49 37 33 27 36 39 27	DailyRate 1102 279 1373 1392 591 884 613 155	DistanceFromHome 1 8 2 3 2 23 6 4	Education 2 1 2 4 1 2 1 3	2 3 4 4 1 3 4 2	94 61 92 56 40 41 42 87	3 2 2 3 3 4 2 4	2 2 1 1 2 3 2	JobSatisfa
0 1 2 3 4 1465 1466 1467	Age 41 49 37 33 27 36 39 27 49	DailyRate 1102 279 1373 1392 591 884 613 155 1023	DistanceFromHome 1 8 2 3 2 23 6 4 2	Education 2 1 2 4 1 2 1 3 3	2 3 4 4 1 3 4 2	94 61 92 56 40 41 42 87 63	3 2 2 3 3 4 2 4 2	2 2 1 1 2 3 2	JobSatisfa
0 1 2 3 4 1465 1466	Age 41 49 37 33 27 36 39 27	DailyRate 1102 279 1373 1392 591 884 613 155	DistanceFromHome 1 8 2 3 2 23 6 4	Education 2 1 2 4 1 2 1 3	2 3 4 4 1 3 4 2	94 61 92 56 40 41 42 87	3 2 2 3 3 4 2 4	2 2 1 1 2 3 2	JobSatisfa
0 1 2 3 4 1465 1466 1467 1468	Age 41 49 37 33 27 36 39 27 49 34	DailyRate 1102 279 1373 1392 591 884 613 155 1023	DistanceFromHome 1 8 2 3 2 23 6 4 2 8	Education 2 1 2 4 1 2 1 3 3	2 3 4 4 1 3 4 2	94 61 92 56 40 41 42 87 63	3 2 2 3 3 4 2 4 2	2 2 1 1 2 3 2	JobSatisfa
0 1 2 3 4 1465 1466 1467 1468	Age 41 49 37 33 27 36 39 27 49 34	DailyRate 1102 279 1373 1392 591 884 613 155 1023 628	DistanceFromHome 1 8 2 3 2 23 6 4 2 8	Education 2 1 2 4 1 2 1 3 3	2 3 4 4 1 3 4 2	94 61 92 56 40 41 42 87 63	3 2 2 3 3 4 2 4 2	2 2 1 1 2 3 2	JobSatisfa
0 1 2 3 4 1465 1466 1467 1468 1470 rd	Age 41 49 37 33 27 36 39 27 49 34 bws × 4	DailyRate 1102 279 1373 1392 591 884 613 155 1023 628	DistanceFromHome 1 8 2 3 2 23 6 4 2 8	Education 2 1 2 4 1 2 1 3 3	2 3 4 4 1 3 4 2	94 61 92 56 40 41 42 87 63	3 2 2 3 3 4 2 4 2	2 2 1 1 2 3 2	JobSatisfa

Age DailyRate DistanceFromHome Education EnvironmentSatisfaction HourlyRate JobInvolvement JobLevel JobSatisfact

Out[38]:

```
Out[42]: Age
          DailyRate
                                                int64
         DistanceFromHome
                                               int64
          Education
                                                int64
          EnvironmentSatisfaction
                                                int64
          HourlyRate
                                                int64
          JobInvolvement
                                               int64
          JobLevel
                                                int64
          JobSatisfaction
                                               int64
          MonthlyIncome
                                               int64
          MonthlyRate
                                               int64
          NumCompaniesWorked
                                               int64
          PercentSalaryHike
                                               int64
          PerformanceRating
                                                int64
          RelationshipSatisfaction
                                               int64
          StockOptionLevel
                                                int64
          TotalWorkingYears
                                               int64
          TrainingTimesLastYear
                                                int64
          WorkLifeBalance
                                               int64
          YearsAtCompany
                                                int64
          YearsInCurrentRole
                                                int64
          YearsSinceLastPromotion
                                                int64
          YearsWithCurrManager
                                                int64
          Attrition_Yes
                                                bool
          BusinessTravel Travel Frequently
                                                 bool
          BusinessTravel_Travel_Rarely
                                                bool
          Department Research & Development
                                                 bool
          {\tt Department\_Sales}
                                                 bool
          EducationField_Life Sciences
                                                 bool
          EducationField Marketing
                                                bool
          EducationField Medical
                                                 bool
          EducationField_Other
                                                 bool
          EducationField_Technical Degree
                                                 bool
          Gender Male
                                                bool
          JobRole Human Resources
                                                 bool
          JobRole Laboratory Technician
                                                bool
          JobRole Manager
                                                 bool
          JobRole Manufacturing Director
                                                bool
          JobRole Research Director
                                                bool
          JobRole Research Scientist
                                                bool
          JobRole Sales Executive
                                                 bool
          JobRole_Sales Representative
                                                bool
          MaritalStatus Married
                                                 bool
          MaritalStatus_Single
                                                bool
          OverTime Yes
                                                bool
          dtype: object
         plt.figure(figsize=(15, 8))
         sns.boxplot(data=df.select_dtypes(include=['int64', 'float64']))
         plt.xticks(rotation=90)
         plt.show()
```

int64



Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js