15Z332 Ex1 - Descriptive Analysis

October 17, 2018

1 Exercise 1

In [22]: import pandas as pd

1.1 Perform descriptive analysis on EmployeeAttrition dataset

The data set contains 1470 records and 35 attributes. The Target attribute is Attrition, which contains two values ('Yes' and 'No')

```
#Importing the dataset
         df = pd.read_csv('EmployeeAttrition.csv')
         print(df.shape)
         df.head(5)
(1470, 35)
Out [22]:
            Age Attrition
                                BusinessTravel
                                                 DailyRate
                                                                         Department
         0
             41
                       Yes
                                 Travel_Rarely
                                                      1102
                                                                               Sales
         1
             49
                            Travel_Frequently
                                                       279
                                                            Research & Development
                        No
                                 Travel_Rarely
         2
             37
                       Yes
                                                      1373
                                                             Research & Development
         3
             33
                        No
                            Travel_Frequently
                                                      1392
                                                             Research & Development
         4
             27
                        No
                                 Travel_Rarely
                                                       591
                                                             Research & Development
                                Education EducationField
                                                           EmployeeCount
                                                                           EmployeeNumber
            DistanceFromHome
                                        2 Life Sciences
         0
                                                                                          1
         1
                            8
                                        1 Life Sciences
                                                                        1
                                                                                          2
         2
                             2
                                                    Other
                                                                                          4
                                                                        1
                            3
                                           Life Sciences
                                                                                         5
         3
                                                                        1
         4
                             2
                                        1
                                                  Medical
                                                                        1
                                                                                          7
                                    RelationshipSatisfaction StandardHours
         0
                                                                           80
                                                             1
                                                             4
         1
                                                                           80
         2
                                                             2
                                                                           80
         3
                                                             3
                                                                           80
         4
                                                             4
                                                                           80
```

	StockOptionLevel	TotalWorkingYear	s TrainingTimesLastYear	WorkLifeBalance	\
0	0		8 0	1	
1	1	1	0 3	3	
2	0		7 3	3	
3	0		8 3	3	
4	1		6 3	3	
	YearsAtCompany Ye	earsInCurrentRole	YearsSinceLastPromotion	\	
0	6	4	0		
1	10	7	1		
2	0	0	0		
3	8	7	3		
4	2	2	2		
	YearsWithCurrMana	ager			
0		5			
1		7			
2		0			
3		0			

[5 rows x 35 columns]

describe() function gives the aggregate analysis of the dataset, such as count, max, min and mean.

In [23]: df.describe()

	•						
	Age	DailyRate	DistanceFromHo		- •		
unt	1470.000000	1470.000000	1470.0000	00 1470.0000	00 1470.0)	
ean	36.923810	802.485714	9.1925	17 2.9129	25 1.0)	
d	9.135373	403.509100	8.1068	64 1.0241	65 0.0)	
.n	18.000000	102.000000	1.0000	00 1.0000	00 1.0)	
5%	30.000000	465.000000	2.0000	00 2.0000	00 1.0)	
)%	36.000000	802.000000	7.0000	00 3.0000	00 1.0)	
75% 43.000000		1157.000000	14.0000	00 4.0000	00 1.0)	
ıx	60.000000	1499.000000	199.000000 29.000000		00 1.0	1.0	
	EmployeeNumb	er Environme	entSatisfaction	HourlyRate	JobInvolvement	\	
unt	1470.0000	00	1470.000000	1470.000000	1470.000000		
		06			2.729932		
		35			0.711561	0.711561	
.n	1.0000	00	1.000000	30.000000	1.000000		
5%	491.2500	00	2.000000	48.000000	2.000000		
)%	1020.5000	00	3.000000	66.000000	3.000000		
5%	1555.7500	00	4.000000	83.750000	3.000000		
ıx	2068.0000	00	4.000000	100.000000	4.000000		
	an d n % % x unt an d n % %	an 36.923810 d 9.135373 n 18.000000 % 30.000000 % 36.000000 % 43.000000 cx 60.000000 EmployeeNumb 1470.0000 1024.8653 d 602.0243 n 1.0000 % 491.2500 % 1555.7500	Tunt 1470.0000000 1470.0000000 1470.00000000000000000000000000000000000	Tunt 1470.00000 1470.000000 1470.00000 an 36.923810 802.485714 9.1925 d 9.135373 403.509100 8.1068 n 18.000000 102.000000 1.00000 % 30.000000 465.000000 2.00000 % 36.000000 802.000000 7.00000 % 43.000000 1157.000000 14.00000 x 60.000000 1499.000000 29.00000 EmployeeNumber EnvironmentSatisfaction funt 1470.000000 1470.000000 an 1024.865306 2.721769 d 602.024335 1.093082 n 1.000000 1.000000 % 491.250000 2.000000 % 1020.500000 3.0000000 % 1555.750000 4.000000	Tunt 1470.000000 1470.000000 1470.000000 1470.000000 1470.00000000000 1470.00000000000000000000000000000000000	runt 1470.000000 1470.000000 1470.000000 1470.000000 1470.000000 ran 36.923810 802.485714 9.192517 2.912925 1.0 d 9.135373 403.509100 8.106864 1.024165 0.0 n 18.000000 102.000000 1.000000 1.000000 1.0 % 30.000000 465.000000 2.000000 2.000000 1.0 % 36.000000 802.000000 7.000000 3.000000 1.0 % 36.000000 1157.000000 14.000000 4.000000 1.0 x 60.000000 1499.000000 29.000000 5.000000 1.0 x 60.000000 1470.000000 1470.000000 1470.000000 1470.000000 x 1024.865306 2.721769 65.891156 2.729932 d 602.024335 1.093082 20.329428 0.711561 n 1.000000 1.000000 30.00000 2.000000 491.250000 2.000000 4.	

	JobLevel			Rela	tionshipSati	sfact	ion \
count	1470.000000				_	0.000	
mean	2.063946					2.712	245
std	1.106940					1.081	209
min	1.000000					1.000	000
25%	1.000000					2.000	
50%	2.000000					3.000	
75%	3.000000					4.000	
max	5.000000					4.000	
	StandardHours	StockOp	tionLevel	TotalW	orkingYears	\	
count	1470.0	_	70.000000		1470.000000		
mean	80.0		0.793878		11.279592		
std	0.0		0.852077		7.780782		
min	80.0		0.000000		0.000000		
25%	80.0		0.000000		6.000000		
50%	80.0		1.000000		10.000000		
75%	80.0		1.000000		15.000000		
max	80.0		3.000000		40.000000		
	TrainingTimesL	astYear	WorkLifeB	alance	YearsAtComp	any	\
count	1470	.000000	1170	000000			
	1410	.000000	1470.	000000	1470.000	0000	
mean		.799320		761224	1470.000 7.008		
	2		2.			3163	
mean	2	.799320	2. 0.	761224	7.008	3163 3525	
mean std	2 1 0	.799320 .289271	2. 0. 1.	761224 706476	7.008 6.126	3163 5525 0000	
mean std min	2 1 0 2	.799320 .289271 .000000	2. 0. 1. 2.	761224 706476 000000	7.008 6.126 0.000	3163 3525 0000 0000	
mean std min 25%	2 1 0 2 3	.799320 .289271 .000000 .000000	2. 0. 1. 2. 3.	761224 706476 000000 000000	7.008 6.126 0.000 3.000	3163 5525 0000 0000	
mean std min 25% 50%	2 1 0 2 3 3	.799320 .289271 .000000 .000000	2. 0. 1. 2. 3.	761224 706476 000000 000000 000000	7.008 6.126 0.000 3.000 5.000	3163 3525 3000 3000 3000	
mean std min 25% 50% 75%	2 1 0 2 3 3	.799320 .289271 .000000 .000000 .000000	2. 0. 1. 2. 3.	761224 706476 000000 000000 000000 000000	7.008 6.126 0.000 3.000 5.000 9.000	3163 3525 3000 3000 3000	
mean std min 25% 50% 75%	2 1 0 2 3 3	.799320 .289271 .000000 .000000 .000000 .000000	2. 0. 1. 2. 3.	761224 706476 000000 000000 000000 000000 000000	7.008 6.126 0.000 3.000 5.000 9.000 40.000	8163 8525 9000 9000 9000 9000 9ithCu	rrManager
mean std min 25% 50% 75%	2 1 0 2 3 3 6	.799320 .289271 .000000 .000000 .000000 .000000	2. 0. 1. 2. 3. 4.	761224 706476 000000 000000 000000 000000 000000	7.008 6.126 0.000 3.000 5.000 9.000 40.000	8163 8525 9000 9000 9000 9000 9ithCu	rrManager 70.000000
mean std min 25% 50% 75% max	2 1 0 2 3 3 6 YearsInCurrent	.799320 .289271 .000000 .000000 .000000 .000000 Role Ye	2. 0. 1. 2. 3. 4.	761224 706476 000000 000000 000000 000000 ostPromo	7.008 6.126 0.000 3.000 5.000 9.000 40.000 tion YearsW	8163 8525 9000 9000 9000 9000 9ithCu	_
mean std min 25% 50% 75% max	2 1 0 2 3 3 6 YearsInCurrent:	.799320 .289271 .000000 .000000 .000000 .000000 Role Ye	2. 0. 1. 2. 3. 4.	761224 706476 000000 000000 000000 000000 stPromod	7.008 6.126 0.000 3.000 5.000 9.000 40.000 tion Yearsw	8163 8525 9000 9000 9000 9000 9ithCu	70.000000
mean std min 25% 50% 75% max count mean std min	2 1 0 2 3 3 6 YearsInCurrent 1470.00 4.22 3.62 0.00	.799320 .289271 .000000 .000000 .000000 .000000 Role Ye	2. 0. 1. 2. 3. 4.	761224 706476 000000 000000 000000 000000 stPromo- 1470.000 2.18	7.008 6.126 0.000 3.000 5.000 9.000 40.000 tion Years 0000 7755	8163 8525 9000 9000 9000 9000 9ithCu	70.000000 4.123129
mean std min 25% 50% 75% max count mean std	2 1 0 2 3 3 6 YearsInCurrent 1470.00 4.22 3.62	.799320 .289271 .000000 .000000 .000000 .000000 Role Ye	2. 0. 1. 2. 3. 4.	761224 706476 000000 000000 000000 000000 stPromo 1470.000 2.18 3.22	7.008 6.126 0.000 3.000 5.000 9.000 40.000 tion Yearsw 0000 7755 2430	8163 8525 9000 9000 9000 9000 9ithCu	70.000000 4.123129 3.568136
mean std min 25% 50% 75% max count mean std min	2 1 0 2 3 3 6 YearsInCurrent 1470.00 4.22 3.62 0.00	.799320 .289271 .000000 .000000 .000000 .000000 Role Ye 0000 9252 3137 0000	2. 0. 1. 2. 3. 4.	761224 706476 000000 000000 000000 000000 stPromo 1470.000 2.18 3.22 0.000	7.008 6.126 0.000 3.000 5.000 9.000 40.000 tion Yearsw 0000 7755 2430 0000	8163 8525 9000 9000 9000 9000 9ithCu	70.000000 4.123129 3.568136 0.000000
mean std min 25% 50% 75% max count mean std min 25%	2 1 0 2 3 3 6 YearsInCurrent: 1470.00 4.22 3.62 0.00 2.00	.799320 .289271 .000000 .000000 .000000 .000000 Role Ye 0000 9252 3137 0000 0000	2. 0. 1. 2. 3. 4.	761224 706476 000000 000000 000000 000000 stPromo 1470.000 2.18 3.22 0.000	7.008 6.126 0.000 3.000 5.000 9.000 40.000 7755 2430 0000 0000	8163 8525 9000 9000 9000 9000 9ithCu	70.000000 4.123129 3.568136 0.000000 2.000000
mean std min 25% 50% 75% max count mean std min 25% 50%	2 1 0 2 3 3 3 6 YearsInCurrent 1470.00 4.22 3.62 0.00 2.00 3.00	.799320 .289271 .000000 .000000 .000000 .000000 .000000	2. 0. 1. 2. 3. 4.	761224 706476 000000 000000 000000 000000 stPromo 1470.000 2.18 3.22 0.000 0.000	7.008 6.126 0.000 3.000 5.000 9.000 40.000 tion Yearsw 0000 7755 2430 0000 0000 0000	8163 8525 9000 9000 9000 9000 VithCu 14	70.000000 4.123129 3.568136 0.000000 2.000000 3.000000

[8 rows x 26 columns]

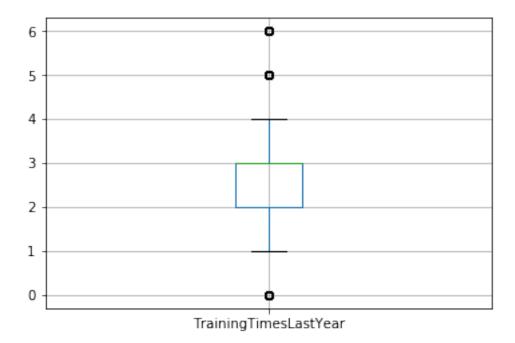
We count the number of missing values in each attribute. If there were any empty cells, appropriate methods to handle them should be administered (in case if the data type is numerical, we can fill with mean or median, and if the data type is categorical, we can fill with mode).

Out[24]:	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
	${\tt RelationshipSatisfaction}$	0
	StandardHours	0
	StockOptionLevel	0
	TotalWorkingYears	0
	${\tt TrainingTimesLastYear}$	0
	WorkLifeBalance	0
	YearsAtCompany	0
	YearsInCurrentRole	0
	${\tt YearsSinceLastPromotion}$	0
	YearsWithCurrManager	0
	dtype: int64	

EmployeeAttrition.csv doesn't have any empty vales. But this doesn't mean that the dataset is clean; it may contain noise data. The behaviour of each data can be analysed using graphs and plot.

For example, a box plot is drawn for the attribute 'TrainingTimesLastYear'. Box plots can only be drawn for numerical datatype. This box plot denotes that the average range of training times for an employee is 1 to 4. Some employees lie beyond this range, which can be considered as outliers. With this, we may conclude that some employee with more training (like 5 or 6 times) may be cost inefficient to the company, hence these employees may not be valuable.

Out[25]: <matplotlib.axes._subplots.AxesSubplot at 0x11422c9b0>



A scatter plot between 'TotalWorkingYears' and 'JobLevel' is drawn. We can see that there is a positive correlation between the two attributes. If an employee has been working for more number of years, they are higher up in the job level. These employes may also be more valuable to the company.

```
In [26]: plt.scatter(df.TotalWorkingYears,df.JobLevel,s=df.Age)
```

Out[26]: <matplotlib.collections.PathCollection at 0x1146d5080>

