## Data Cleaning.

```
In [4]: import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [5]: # Read the dataset
df = pd.read\_csv(r'F:\Technocolabs\WA\_Fn-UseC\_-HR-Employee-Attrition.csv')

In [6]: df

#### Out[6]:

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

1470 rows × 35 columns

### In [7]: 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
1	Attrition	1470 non-null	object
2	BusinessTravel	1470 non-null	object
3	DailyRate	1470 non-null	int64
4	Department	1470 non-null	object
5	DistanceFromHome	1470 non-null	int64
6	Education	1470 non-null	int64
7	EducationField	1470 non-null	object
8	EmployeeCount	1470 non-null	int64
9	EmployeeNumber	1470 non-null	int64
10	EnvironmentSatisfaction	1470 non-null	int64
11	Gender	1470 non-null	object
12	HourlyRate	1470 non-null	int64
13	JobInvolvement	1470 non-null	int64
14	JobLevel	1470 non-null	int64
15	JobRole	1470 non-null	object
16	JobSatisfaction	1470 non-null	int64
17	MaritalStatus	1470 non-null	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
28	TotalWorkingYears	1470 non-null	int64
29	TrainingTimesLastYear	1470 non-null	int64
30	WorkLifeBalance	1470 non-null	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
dtvne	es: int64(26) object(9)		

dtypes: int64(26), object(9)
memory usage: 402.1+ KB

	Age Attrition	Busine	ssTravel	DailyRate		Department
\						
1465	36 No	Travel_Fr		884		Development
1466	39 No		l_Rarely	613		Development
1467	27 No		l_Rarely	155	Research &	Development
1468	49 No	_		1023		Sales
1469	34 No	Irave	l_Rarely	628	Research &	Development
	DistanceEnemil	omo Educat	ion Educ	a+ionFiold	EmployeeCou	m <b>+</b> \
1465	DistanceFromH			ationField Modical	EmployeeCou	
1465		23	2	Medical Medical		1
1466		6	1			1
1467		4		e Sciences		1
1468		2	3 3	Medical		1
1469		8	3	Medical		1
	EmployeeNumbe	r Rel	ationshi	nSatisfactio	n StandardH	ours \
1465	206		ac101131111	554613146616	3	80
1466	206				1	80
1467	206				2	80
1468	206				4	80
1469	206				1	80
1409	200	o			_	80
	StockOptionLe	vel TotalW	orkingYea	ars Trainin	ngTimesLastY	ear \
1465	5 00 0 1 0 p 0 = 0 1 1 = 0	1	og. c.	17	.8	3
1466		1		9		5
1467		1		6		0
1468		0		17		3
1469		0		6		3
		•				•
	WorkLifeBalanc	e YearsAtC	ompany Ye	earsInCurrer	ntRole \	
1465		3	5		2	
1466		3	7		7	
1467		3	6		2	
1468		2	9		6	
1469		4	4		3	
	YearsSinceLas	tPromotion	YearsWi	thCurrManage	er	
1465		0			3	
1466		1			7	
1467		0			3	
1468		0			8	
1469		1			2	

[5 rows x 35 columns]

In [9]: print(df.tail(10)) # Display the last 10 rows

	Age Attrition	Busines	sTravel	DailyRate		Department
\						
1460	29 No		_Rarely	468	Research &	Development
1461	50 Yes		_Rarely	410		Sales
1462	39 No	Travel	_Rarely	722		Sales
1463	31 No	Non	-Travel	325	Research &	Development
1464	26 No	Travel	_Rarely	1167		Sales
1465	36 No	Travel_Fre		884	Research &	Development
1466	39 No	<del>-</del>	Rarely	613		Development
1467	27 No		_Rarely	155		Development
1468	49 No	· ·		1023	nescaren a	Sales
1469	34 No	<del>-</del>	_Rarely	628	Rospanch &	Development
1407	34 110	II avei	_narely	028	Research &	Development
	DistanceFromH	ome Educati	on Educa	tionField	EmployeeCou	nt \
1460		28	4	Medical	, ,,	1
1461		28		Marketing		1
1462		24		Marketing		1
1463		5	3	Medical		1
				Other		
1464		5	3			1
1465		23	2	Medical		1
1466		6	1	Medical		1
1467		4		Sciences		1
1468		2	3	Medical		1
1469		8	3	Medical		1
	- 1 N I	- 1			c. l l.	,
	EmployeeNumbe		tionship	Satistactio	n StandardH	
1460	205				2	80
1461	205				2	80
1462	205				1	80
1463	205				2	80
1464	206				4	80
1465	206	1			3	80
1466	206	2			1	80
1467	206	4			2	80
1468	206	5			4	80
1469	206	8			1	80
4.50	StockOptionLe		rkingYea		gTimesLastY	
1460		0		5		3
1461		1		20		3
1462		1		21		2
1463		0		10		2
1464		0		5		2
1465		1		17		3
1466		1		9		5
1467		1		6		0
1468		0		17		3
1469		0		6		3
	WorkLifeBalanc			arsInCurren		
1460		1	5		4	
1461		3	3		2	
1462		2	20		9	
1463		3	9		4	
1464		3	4		2	
1465		3	5		2	
1466		3	7		7	
1467		3	6		2	
1468		2	9		6	
1469		4	4		3	

	YearsSinceLastPromotion	YearsWithCurrManager
1460	0	4
1461	2	0
1462	9	6
1463	1	7
1464	0	0
1465	0	3
1466	1	7
1467	0	3
1468	0	8
1469	1	2

[10 rows x 35 columns]

In [10]: print(df.head(10))

<ol> <li>0</li> <li>1</li> <li>2</li> <li>3</li> <li>4</li> <li>5</li> <li>6</li> <li>7</li> <li>8</li> <li>9</li> </ol>	Age 41 49 37 33 27 32 59 30 38 36	Attrition Yes No Yes No No No No No No No No No	Tra Tra	BusinessTr Travel_Ravel_Freque Travel_Ravel_Freque Travel_Ravel_Freque Travel_Ravel_	arely ently arely ently arely ently arely arely arely ently	DailyRate 1102 279 1373 1392 591 1005 1324 1358 216 1299	Research & [ Research & [ Research & [ Research & [ Research & [ Research & [	Development Development Development Development Development Development Development	\
010		canceFromHor	ne	Education	Educat	tionField	EmployeeCount	EmployeeN	umb
er 0	\		1	2	Life	Sciences	1	L	
1 1			8	1	Life	Sciences	1	L	
2			2	2		Other	1	L	
4 3			3	4	Life	Sciences	1	L	
5 4			2	1		Medical	1	L	
7 5			2	2	Life	Sciences	1	L	
8			3	3		Medical	1	L	
10 7		2	24	1	Life	Sciences	1	L	
11 8		2	23	3	Life	Sciences	1	L	
12 9 13		2	27	3		Medical	1	L	
		Relationsh	hinS	atisfactio	on Star	ndardHours	StockOptionI	_evel \	
0	• • •	RETUCTORIST			1	80	Stockoption	0	
1 2	• • •				4 2	80 80		1 0	
3					3	80		0	
4					4	80		1	
5	• • •				3	80		0	
6	• • •				1	80		3	
7 8	• • •				2	80 80		1 0	
9	• • •				2	80		2	
		alWorkingYea	ars	Training	ΓimesLa	astYear Wo	rkLifeBalance	YearsAtCom	pan
у 0	\		8			0	1		
6			10			3	3		1
0 2			7			3	3		
0 3			8			3	3		
8			6			3	3		
2 5 7			8			2	2		

6 1	12	3	2
7	1	2	3
8	10	2	3
9 9 7	17	3	2
,			
	YearsInCurrentRole	YearsSinceLastPromotion	YearsWithCurrManager
0	4	0	5
0 1	4 7	0 1	5 7
_			5 7 0
1	7	1	7
1 2	7	1 0	7 0
1 2 3	7	1 0	7 0 0
1 2 3 4	7 0 7 2	1 0 3 2	7 0 0 2
1 2 3 4 5	7 0 7 2 7	1 0 3 2 3	7 0 0 2 6
1 2 3 4 5	7 0 7 2 7 0	1 0 3 2 3 0	7 0 0 2 6 0
1 2 3 4 5 6 7	7 0 7 2 7 0	1 0 3 2 3 0	7 0 0 2 6 0

[10 rows x 35 columns]

In [12]: # Handle missing values
 df.dropna(inplace=True)
 print(df)

```
Age Attrition
                           BusinessTravel DailyRate
                                                                        Department
\
0
       41
                  Yes
                            Travel_Rarely
                                                    1102
                                                                              Sales
1
        49
                   No
                        Travel_Frequently
                                                    279
                                                          Research & Development
2
        37
                            Travel_Rarely
                                                   1373
                                                          Research & Development
                  Yes
3
                                                          Research & Development
       33
                   No
                        Travel_Frequently
                                                   1392
4
       27
                   No
                            Travel_Rarely
                                                     591
                                                          Research & Development
       . . .
                  . . .
                                                     . . .
. . .
                        Travel_Frequently
                                                          Research & Development
1465
       36
                                                     884
                   No
                            Travel_Rarely
1466
       39
                   No
                                                     613
                                                          Research & Development
                            Travel_Rarely
                                                          Research & Development
1467
        27
                   No
                                                     155
1468
        49
                   No
                        Travel_Frequently
                                                    1023
                                                                              Sales
1469
        34
                   No
                            Travel Rarely
                                                     628
                                                          Research & Development
                           Education EducationField
      DistanceFromHome
                                                         EmployeeCount
                                        Life Sciences
0
                        1
                                     2
                                                                       1
1
                        8
                                     1
                                        Life Sciences
                                                                       1
2
                        2
                                     2
                                                 Other
                                                                       1
3
                        3
                                                                       1
                                     4
                                        Life Sciences
4
                        2
                                     1
                                                                       1
                                               Medical
                                                    . . .
                      . . .
                                     2
                                               Medical
                                                                       1
1465
                       23
1466
                        6
                                     1
                                               Medical
                                                                       1
                        4
                                     3
                                        Life Sciences
                                                                       1
1467
                        2
                                     3
                                               Medical
                                                                       1
1468
                        8
                                     3
                                                                       1
1469
                                               Medical
      EmployeeNumber
                               RelationshipSatisfaction StandardHours
                         . . .
0
                      1
                                                         1
                                                                        80
                         . . .
1
                      2
                                                         4
                                                                        80
                         . . .
                                                         2
2
                     4
                                                                        80
3
                     5
                                                         3
                                                                        80
4
                     7
                                                         4
                                                                        80
                   . . .
. . .
                                                                        . . .
                  2061
                                                         3
                                                                        80
1465
                                                         1
1466
                  2062
                                                                        80
                                                         2
1467
                                                                        80
                  2064
                                                         4
1468
                  2065
                                                                        80
                         . . .
1469
                  2068
                                                         1
                                                                        80
                         . . .
      StockOptionLevel
                                                 TrainingTimesLastYear
                           TotalWorkingYears
0
                        0
                                              8
                                                                        0
                        1
                                                                        3
1
                                             10
2
                        0
                                              7
                                                                        3
3
                                              8
                        0
                                                                        3
4
                        1
                                              6
                                                                        3
                                            . . .
                                                                       . . .
                        1
                                                                        3
                                            17
1465
                        1
                                              9
                                                                        5
1466
                        1
                                              6
                                                                        0
1467
                        0
                                            17
1468
                                                                        3
                        0
                                              6
                                                                        3
1469
     WorkLifeBalance
                        YearsAtCompany YearsInCurrentRole
0
                                        6
                     1
                                                              4
                                                              7
1
                     3
                                       10
                                                              0
2
                     3
                                        0
                                                              7
3
                     3
                                        8
                                        2
                                                              2
4
                      3
                                      . . .
                                                            . . .
. . .
                     3
                                        5
                                                              2
1465
```

		•
146	3	6
146	8 2	9
146	9 4	4
	YearsSinceLastPromotion	YearsWithCurrManager
0	0	5
1	1	7
2	0	0
3	3	0
4	2	2
	•••	•••
146	55 0	3
146	66 1	7
146	57 0	3
146	88	8
146	9 1	2

[1470 rows x 35 columns]

```
In [31]: missing_values = df.isnull().sum()
print("Missing values in each column:")
print(missing_values)
```

Missing values in each col Age Attrition BusinessTravel DailyRate Department DistanceFromHome Education EducationField EmployeeCount EmployeeNumber EnvironmentSatisfaction	umn: 0 0 0 0 0 0 0 0 0 0 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 0
YearsAtCompany YearsInCurrentRole	0
YearsSinceLastPromotion	0
YearsWithCurrManager	0
dtype: int64	U
27,701	

# In [37]: # Check for missing values print(df.isnull().sum())

Age 0 0 Attrition 0 BusinessTravel DailyRate 0 0 Department DistanceFromHome 0 0 Education EducationField 0 0 EmployeeCount EmployeeNumber 0 EnvironmentSatisfaction 0 Gender 0 HourlyRate 0 JobInvolvement 0 JobLevel 0 JobRole 0 JobSatisfaction 0 0 MaritalStatus MonthlyIncome 0 0 MonthlyRate NumCompaniesWorked 0 0ver18 0 OverTime 0 PercentSalaryHike 0 PerformanceRating 0 RelationshipSatisfaction StandardHours 0 StockOptionLevel 0 0 TotalWorkingYears TrainingTimesLastYear 0 WorkLifeBalance 0 YearsAtCompany 0 0 YearsInCurrentRole YearsSinceLastPromotion 0 YearsWithCurrManager 0 dtype: int64

In [38]: # Drop rows with missing values

df.dropna(inplace=True)

In [39]: # Remove duplicate rows

df.drop\_duplicates(inplace=True)

#### In [41]: # Check the data types after conversion print(df.dtypes)

Age int64 Attrition object BusinessTravel object DailyRate int64 Department object DistanceFromHome int64 Education int64 EducationField object EmployeeCount int64 EmployeeNumber int64 EnvironmentSatisfaction int64 Gender object HourlyRate int64 JobInvolvement int64 JobLevel int64 JobRole object JobSatisfaction int64 MaritalStatus object MonthlyIncome int64 MonthlyRate int64 NumCompaniesWorked int64 0ver18 object OverTime object PercentSalaryHike int64 PerformanceRating int64 RelationshipSatisfaction int64 StandardHours int64 StockOptionLevel int64 TotalWorkingYears int64 TrainingTimesLastYear int64 WorkLifeBalance int64 YearsAtCompany int64 YearsInCurrentRole int64 YearsSinceLastPromotion int64 YearsWithCurrManager int64

dtype: object

```
# Save the cleaned dataset
In [42]:
         df.to_csv("cleaned_dataset.csv", index=False)
```

```
In [ ]:
```