

# Data Labelling

```
In [1]: import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```
In [2]: # Read the dataset
df = pd.read_csv(r'F:\Technocolabs\WA_Fn-UseC_-HR-Employee-Attrition.csv')
```

```
In [3]: df
```

Out[3]:

	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 [5]: import pandas as pd

# Load your dataset
# Replace 'your_dataset.csv' with the actual path to your dataset file
df = pd.read_csv(r'F:\Technocolabs\WA_Fn-UseC_-HR-Employee-Attrition.csv')

# Define your condition and Label accordingly
def label_function(row):
    if row['Age'] > 30:
        return 'label_A'
    else:
        return 'label_B'

# Apply the Label function to each row
df['label'] = df.apply(label_function, axis=1)

# Display the first few rows to verify the labeling
print(df.head())
```

	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	\
0		1	2 Life Sciences		1	
1						
1		8	1 Life Sciences		1	
2						
2		2	2 Other		1	
4						
3		3	4 Life Sciences		1	
5						
4		2	1 Medical		1	
7						

  

	...	StandardHours	StockOptionLevel	TotalWorkingYears	\
0	...	80	0	8	
1	...	80	1	10	
2	...	80	0	7	
3	...	80	0	8	
4	...	80	1	6	

  

	TrainingTimesLastYear	WorkLifeBalance	YearsAtCompany	YearsInCurrentRole	\
0		0	1	6	
4					
1		3	3	10	
7					
2		3	3	0	
0					
3		3	3	8	
7					
4		3	3	2	
2					

  

	YearsSinceLastPromotion	YearsWithCurrManager	label
0	0	5	label_A
1	1	7	label_A
2	0	0	label_A
3	3	0	label_A
4	2	2	label_B

[5 rows x 36 columns]

```
In [6]: # Define your conditions and labels accordingly
def label_function(row):
    if row['Age'] > 30 and row['Department'] == 'Sales':
        return 'label_A'
    # Example: If DailyRate is less than 500 and Education is greater than 3
    elif row['DailyRate'] < 500 and row['Education'] > 3:
        return 'label_B'
    # Add more conditions and labels as needed
    else:
        return 'label_C' # Default label if none of the conditions are met

# Apply the label function to each row
df['label'] = df.apply(label_function, axis=1)

# Display the first few rows to verify the labeling
print(df.head())
```

	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	\
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	
5						
6						
7						

	...	StandardHours	StockOptionLevel	TotalWorkingYears	\
0	...	80	0	8	
1	...	80	1	10	
2	...	80	0	7	
3	...	80	0	8	
4	...	80	1	6	

	TrainingTimesLastYear	WorkLifeBalance	YearsAtCompany	YearsInCurrentRole	\
0		0	1	6	
1		3	3	10	
2		3	3	0	
3		3	3	8	
4		3	3	2	
5					
6					
7					

	YearsSinceLastPromotion	YearsWithCurrManager	label
0	0	5	label_A
1	1	7	label_C
2	0	0	label_C
3	3	0	label_C
4	2	2	label_C

[5 rows x 36 columns]

```
In [7]: # Define your conditions and labels accordingly
def label_attrition(row):
    if row['Attrition'] == 'Yes':
        return 'High Attrition'
    else:
        return 'Low Attrition'

def label_income(row):
    if row['MonthlyIncome'] > 5000:
        return 'High Income'
    else:
        return 'Low Income'

# Apply the label functions to each row
df['Attrition_Label'] = df.apply(label_attrition, axis=1)
df['Income_Label'] = df.apply(label_income, axis=1)

# Display the first few rows to verify the labeling
print(df.head())
```

	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	\
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	

  

	...	TotalWorkingYears	TrainingTimesLastYear	WorkLifeBalance	\
0	...	8	0	1	
1	...	10	3	3	
2	...	7	3	3	
3	...	8	3	3	
4	...	6	3	3	

  

	YearsAtCompany	YearsInCurrentRole	YearsSinceLastPromotion	\
0	6	4	0	
1	10	7	1	
2	0	0	0	
3	8	7	3	
4	2	2	2	

  

	YearsWithCurrManager	label	Attrition_Label	Income_Label
0	5	label_A	High Attrition	High Income
1	7	label_C	Low Attrition	High Income
2	0	label_C	High Attrition	Low Income
3	0	label_C	Low Attrition	Low Income
4	2	label_C	Low Attrition	Low Income

[5 rows x 38 columns]

```
In [8]: # Define your conditions and labels accordingly
def label_attrition(row):
    if row['Attrition'] == 'Yes':
        return 'High Attrition'
    else:
        return 'Low Attrition'

def label_income(row):
    if row['MonthlyIncome'] > 5000:
        return 'High Income'
    else:
        return 'Low Income'

# Apply the label functions to each row
df['Attrition_Label'] = df.apply(label_attrition, axis=1)
df['Income_Label'] = df.apply(label_income, axis=1)

# Display the first few rows to verify the labeling
print(df.head())
```



	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	\
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	

  

	...	TotalWorkingYears	TrainingTimesLastYear	WorkLifeBalance	\
0	...	8	0	1	
1	...	10	3	3	
2	...	7	3	3	
3	...	8	3	3	
4	...	6	3	3	

  

	YearsAtCompany	YearsInCurrentRole	YearsSinceLastPromotion	\
0	6	4	0	
1	10	7	1	
2	0	0	0	
3	8	7	3	
4	2	2	2	

  

	YearsWithCurrManager	label	Attrition_Label	Income_Label
0	5	label_A	High Attrition	High Income
1	7	label_C	Low Attrition	High Income
2	0	label_C	High Attrition	Low Income
3	0	label_C	Low Attrition	Low Income
4	2	label_C	Low Attrition	Low Income

[5 rows x 38 columns]

In [ ]: