

```
In [1]: import pandas as pd
```

```
In [2]: raw_data = pd.read_csv('Absenteeism-data.csv')
```

```
In [3]: raw_data
```

Out[3]:

	ID	Reason for Absence	Date	Transportation Expense	Distance to Work	Age	Daily Work Load Average	Body Mass Index	Education	Children	Pets	Absenteeism Time in Hours
0	11	26	07/07/2015	289	36	33	239.554	30	1	2	1	4
1	36	0	14/07/2015	118	13	50	239.554	31	1	1	0	0
2	3	23	15/07/2015	179	51	38	239.554	31	1	0	0	2
3	7	7	16/07/2015	279	5	39	239.554	24	1	2	0	4
4	11	23	23/07/2015	289	36	33	239.554	30	1	2	1	2
...	...	...	...	...	...	...	...	...	...	...	...	...
695	17	10	23/05/2018	179	22	40	237.656	22	2	2	0	8
696	28	6	23/05/2018	225	26	28	237.656	24	1	1	2	3
697	18	10	24/05/2018	330	16	28	237.656	25	2	0	0	8
698	25	23	24/05/2018	235	16	32	237.656	25	3	0	0	2
699	15	28	31/05/2018	291	31	40	237.656	25	1	1	1	2

700 rows × 12 columns

```
In [4]: df = raw_data.copy()
```

```
In [5]: df
```

Out[5]:

	ID	Reason for Absence	Date	Transportation Expense	Distance to Work	Age	Daily Work Load Average	Body Mass Index	Education	Children	Pets	Absenteeism Time in Hours
0	11	26	07/07/2015	289	36	33	239.554	30	1	2	1	4
1	36	0	14/07/2015	118	13	50	239.554	31	1	1	0	0
2	3	23	15/07/2015	179	51	38	239.554	31	1	0	0	2
3	7	7	16/07/2015	279	5	39	239.554	24	1	2	0	4
4	11	23	23/07/2015	289	36	33	239.554	30	1	2	1	2
...	...	...	...	...	...	...	...	...	...	...	...	...
695	17	10	23/05/2018	179	22	40	237.656	22	2	2	0	8
696	28	6	23/05/2018	225	26	28	237.656	24	1	1	2	3
697	18	10	24/05/2018	330	16	28	237.656	25	2	0	0	8
698	25	23	24/05/2018	235	16	32	237.656	25	3	0	0	2
699	15	28	31/05/2018	291	31	40	237.656	25	1	1	1	2

700 rows × 12 columns

```
In [6]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 700 entries, 0 to 699
Data columns (total 12 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   ID                                    700 non-null    int64
1   Reason for Absence                   700 non-null    int64
2   Date                                700 non-null    object
3   Transportation Expense               700 non-null    int64
4   Distance to Work                    700 non-null    int64
5   Age                                  700 non-null    int64
6   Daily Work Load Average            700 non-null    float64
7   Body Mass Index                     700 non-null    int64
8   Education                           700 non-null    int64
```

```
9  Children          700 non-null  int64
10  Pets             700 non-null  int64
11  Absenteeism Time in Hours  700 non-null  int64
dtypes: float64(1), int64(10), object(1)
memory usage: 65.8+ KB
```

In [7]:

df.drop(['ID'], axis=1, inplace=True)

In [8]:

df

Out[8]:

	Reason for Absence	Date	Transportation Expense	Distance to Work	Age	Daily Work Load Average	Body Mass Index	Education	Children	Pets	Absenteeism Time in Hours
0	26	07/07/2015	289	36	33	239.554	30	1	2	1	4
1	0	14/07/2015	118	13	50	239.554	31	1	1	0	0
2	23	15/07/2015	179	51	38	239.554	31	1	0	0	2
3	7	16/07/2015	279	5	39	239.554	24	1	2	0	4
4	23	23/07/2015	289	36	33	239.554	30	1	2	1	2
...	...	...	...	...	...	...	...	...	...	...	...
695	10	23/05/2018	179	22	40	237.656	22	2	2	0	8
696	6	23/05/2018	225	26	28	237.656	24	1	1	2	3
697	10	24/05/2018	330	16	28	237.656	25	2	0	0	8
698	23	24/05/2018	235	16	32	237.656	25	3	0	0	2
699	28	31/05/2018	291	31	40	237.656	25	1	1	1	2

700 rows × 11 columns

In [9]:

df['Reason for Absence'].min()

Out[9]: 0

In [10]:

df['Reason for Absence'].max()

Out[10]: 28

In [11]:

df['Reason for Absence'].nunique()

Out[11]: 28

In [12]:

df['Reason for Absence'].unique()

Out[12]: array([26, 0, 23, 7, 22, 19, 1, 11, 14, 21, 10, 13, 28, 18, 25, 24, 6,  
 27, 17, 8, 12, 5, 9, 15, 4, 3, 2, 16], dtype=int64)

In [13]:

df['Reason for Absence'].value\_counts()

Out[13]:

23	147
28	110
27	66
13	52
0	38
19	36
22	32
26	31
25	29
11	24
10	22
18	21
14	18
1	16
7	13
12	8
6	6
21	6
8	5
9	4
24	3
5	3
16	3
15	2

```
4      2
3      1
17     1
2      1
Name: Reason for Absence, dtype: int64
```

```
In [14]: sorted(df['Reason for Absence'].unique())
```

```
Out[14]: [0,
1,
2,
3,
4,
5,
6,
7,
8,
9,
10,
11,
12,
13,
14,
15,
16,
17,
18,
19,
21,
22,
23,
24,
25,
26,
27,
28]
```

```
In [15]: reason_columns = pd.get_dummies(df['Reason for Absence'])
```

```
In [16]: reason_columns
```

Out[16]:

	0	1	2	3	4	5	6	7	8	9	...	18	19	21	22	23	24	25	26	27	28
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	1	0	0
1	1	1	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	1	0	0	0	0	0
3	0	0	0	0	0	0	0	1	0	0	...	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	1	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
695	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
696	0	0	0	0	0	0	1	0	0	0	...	0	0	0	0	0	0	0	0	0	0
697	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
698	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	1	0	0	0	0	0
699	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1

700 rows x 28 columns

```
In [17]: reason_columns['Check'] = reason_columns.sum(axis=1)
```

```
In [18]: reason_columns
```

Out[18]:

	0	1	2	3	4	5	6	7	8	9	...	19	21	22	23	24	25	26	27	28	Check
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	1	0	0	1
1	1	1	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	1	0	0	0	0	0	1
3	0	0	0	0	0	0	0	1	0	0	...	0	0	0	0	0	0	0	0	0	1
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	1	0	0	0	0	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

	0	1	2	3	4	5	6	7	8	9	...	19	21	22	23	24	25	26	27	28	Check
695	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1
696	0	0	0	0	0	0	1	0	0	0	...	0	0	0	0	0	0	0	0	0	1
697	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1
698	0	0	0	0	0	0	0	0	0	0	...	0	0	0	1	0	0	0	0	0	1
699	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	1	1

700 rows × 29 columns

```
In [19]: reason_columns['Check'].sum(axis=0)
```

Out[19]: 700

```
In [20]: reason_columns['Check'].unique()
```

Out[20]: array([1], dtype=int64)

```
In [21]: reason_columns = reason_columns.drop(['Check'], axis=1)
```

```
In [22]: reason_columns
```

```
Out[22]:
```

	0	1	2	3	4	5	6	7	8	9	...	18	19	21	22	23	24	25	26	27	28
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	1	0	0
1	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	1	0	0	0	0	0
3	0	0	0	0	0	0	0	1	0	0	...	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	1	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
695	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
696	0	0	0	0	0	0	1	0	0	0	...	0	0	0	0	0	0	0	0	0	0
697	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
698	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	1	0	0	0	0	0
699	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1

700 rows × 28 columns

```
In [23]: reason_columns = reason_columns.drop([0], axis=1)
```

```
In [24]: reason_columns
```

```
Out[24]:
```

	1	2	3	4	5	6	7	8	9	10	...	18	19	21	22	23	24	25	26	27	28
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	1	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	1	0	0	0	0	0
3	0	0	0	0	0	0	1	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	1	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
695	0	0	0	0	0	0	0	0	0	1	...	0	0	0	0	0	0	0	0	0	0
696	0	0	0	0	0	1	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
697	0	0	0	0	0	0	0	0	0	1	...	0	0	0	0	0	0	0	0	0	0
698	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	1	0	0	0	0	0
699	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1

700 rows × 27 columns

In [25]:

df.drop(['Reason for Absence'], axis=1, inplace=True)

In [26]:

df

Out[26]:

	Date	Transportation Expense	Distance to Work	Age	Daily Work Load Average	Body Mass Index	Education	Children	Pets	Absenteeism Time in Hours
0	07/07/2015	289	36	33	239.554	30	1	2	1	4
1	14/07/2015	118	13	50	239.554	31	1	1	0	0
2	15/07/2015	179	51	38	239.554	31	1	0	0	2
3	16/07/2015	279	5	39	239.554	24	1	2	0	4
4	23/07/2015	289	36	33	239.554	30	1	2	1	2
...	...	...	...	...	...	...	...	...	...	...
695	23/05/2018	179	22	40	237.656	22	2	2	0	8
696	23/05/2018	225	26	28	237.656	24	1	1	2	3
697	24/05/2018	330	16	28	237.656	25	2	0	0	8
698	24/05/2018	235	16	32	237.656	25	3	0	0	2
699	31/05/2018	291	31	40	237.656	25	1	1	1	2

700 rows × 10 columns

In [27]:

reason\_type\_1 = reason\_columns.loc[:, 1:14].max(axis=1)  
reason\_type\_2 = reason\_columns.loc[:, 15:17].max(axis=1)  
reason\_type\_3 = reason\_columns.loc[:, 18:21].max(axis=1)  
reason\_type\_4 = reason\_columns.loc[:, 22:].max(axis=1)

In [28]:

reason\_type\_4

Out[28]:

0	1
1	0
2	1
3	0
4	1
..	
695	0
696	0
697	0
698	1
699	1

Length: 700, dtype: uint8

In [29]:

df

Out[29]:

	Date	Transportation Expense	Distance to Work	Age	Daily Work Load Average	Body Mass Index	Education	Children	Pets	Absenteeism Time in Hours
0	07/07/2015	289	36	33	239.554	30	1	2	1	4
1	14/07/2015	118	13	50	239.554	31	1	1	0	0
2	15/07/2015	179	51	38	239.554	31	1	0	0	2
3	16/07/2015	279	5	39	239.554	24	1	2	0	4
4	23/07/2015	289	36	33	239.554	30	1	2	1	2
...	...	...	...	...	...	...	...	...	...	...
695	23/05/2018	179	22	40	237.656	22	2	2	0	8
696	23/05/2018	225	26	28	237.656	24	1	1	2	3
697	24/05/2018	330	16	28	237.656	25	2	0	0	8
698	24/05/2018	235	16	32	237.656	25	3	0	0	2
699	31/05/2018	291	31	40	237.656	25	1	1	1	2

700 rows × 10 columns

In [30]:

df = pd.concat([df, reason\_type\_1, reason\_type\_2, reason\_type\_3, reason\_type\_4], axis=1)

```
In [31]: df
```

Out[31]:

	Date	Transportation Expense	Distance to Work	Age	Daily Work Load Average	Body Mass Index	Education	Children	Pets	Absenteeism Time in Hours	0	1	2	3
0	07/07/2015	289	36	33	239.554	30	1	2	1	4	0	0	0	1
1	14/07/2015	118	13	50	239.554	31	1	1	0	0	0	0	0	0
2	15/07/2015	179	51	38	239.554	31	1	0	0	2	0	0	0	1
3	16/07/2015	279	5	39	239.554	24	1	2	0	4	1	0	0	0
4	23/07/2015	289	36	33	239.554	30	1	2	1	2	0	0	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
695	23/05/2018	179	22	40	237.656	22	2	2	0	8	1	0	0	0
696	23/05/2018	225	26	28	237.656	24	1	1	2	3	1	0	0	0
697	24/05/2018	330	16	28	237.656	25	2	0	0	8	1	0	0	0
698	24/05/2018	235	16	32	237.656	25	3	0	0	2	0	0	0	1
699	31/05/2018	291	31	40	237.656	25	1	1	1	2	0	0	0	1

700 rows × 14 columns

```
In [32]: df.rename(columns={0: "Reason_1", 1: "Reason_2", 2: "Reason_3", 3: "Reason_4"}, inplace=True)
```

```
In [33]: df
```

Out[33]:

	Date	Transportation Expense	Distance to Work	Age	Daily Work Load Average	Body Mass Index	Education	Children	Pets	Absenteeism Time in Hours	Reason_1	Reason_2	Reason_3
0	07/07/2015	289	36	33	239.554	30	1	2	1	4	0	0	0
1	14/07/2015	118	13	50	239.554	31	1	1	0	0	0	0	0
2	15/07/2015	179	51	38	239.554	31	1	0	0	2	0	0	0
3	16/07/2015	279	5	39	239.554	24	1	2	0	4	1	0	0
4	23/07/2015	289	36	33	239.554	30	1	2	1	2	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
695	23/05/2018	179	22	40	237.656	22	2	2	0	8	1	0	0
696	23/05/2018	225	26	28	237.656	24	1	1	2	3	1	0	0
697	24/05/2018	330	16	28	237.656	25	2	0	0	8	1	0	0
698	24/05/2018	235	16	32	237.656	25	3	0	0	2	0	0	0
699	31/05/2018	291	31	40	237.656	25	1	1	1	2	0	0	0

700 rows × 14 columns



```
In [34]: df.columns
```

```
Out[34]: Index(['Date', 'Transportation Expense', 'Distance to Work', 'Age',  
              'Daily Work Load Average', 'Body Mass Index', 'Education', 'Children',  
              'Pets', 'Absenteeism Time in Hours', 'Reason_1', 'Reason_2', 'Reason_3',  
              'Reason_4'],  
              dtype='object')
```

```
In [35]: col_reordered = ['Reason_1', 'Reason_2', 'Reason_3',  
                          'Reason_4', 'Date', 'Transportation Expense', 'Distance to Work', 'Age',  
                          'Daily Work Load Average', 'Body Mass Index', 'Education', 'Children',  
                          'Pets', 'Absenteeism Time in Hours']
```

```
In [36]: df = df[col_reordered]
```

```
In [37]: df
```

Out[37]:

	Reason_1	Reason_2	Reason_3	Reason_4	Date	Transportation Expense	Distance to Work	Age	Daily Work Load Average	Body Mass Index	Education	Children	Pets	Absenteeism Time (Hours)
0	0	0	0	1	07/07/2015	289	36	33	239.554	30	1	2	1	
1	0	0	0	0	14/07/2015	118	13	50	239.554	31	1	1	0	
2	0	0	0	1	15/07/2015	179	51	38	239.554	31	1	0	0	
3	1	0	0	0	16/07/2015	279	5	39	239.554	24	1	2	0	
4	0	0	0	1	23/07/2015	289	36	33	239.554	30	1	2	1	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
695	1	0	0	0	23/05/2018	179	22	40	237.656	22	2	2	0	
696	1	0	0	0	23/05/2018	225	26	28	237.656	24	1	1	2	
697	1	0	0	0	24/05/2018	330	16	28	237.656	25	2	0	0	
698	0	0	0	1	24/05/2018	235	16	32	237.656	25	3	0	0	
699	0	0	0	1	31/05/2018	291	31	40	237.656	25	1	1	1	

700 rows × 14 columns



```
In [38]: df['Date'] = pd.to_datetime(df['Date'], format = '%d/%m/%Y')
```

```
In [39]: df
```

Out[39]:

	Reason_1	Reason_2	Reason_3	Reason_4	Date	Transportation Expense	Distance to Work	Age	Daily Work Load Average	Body Mass Index	Education	Children	Pets	Absenteeism Time (Hours)
0	0	0	0	1	2015-07-07	289	36	33	239.554	30	1	2	1	
1	0	0	0	0	2015-07-14	118	13	50	239.554	31	1	1	0	
2	0	0	0	1	2015-07-15	179	51	38	239.554	31	1	0	0	
3	1	0	0	0	2015-07-16	279	5	39	239.554	24	1	2	0	
4	0	0	0	1	2015-07-23	289	36	33	239.554	30	1	2	1	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
695	1	0	0	0	2018-05-23	179	22	40	237.656	22	2	2	0	
696	1	0	0	0	2018-05-23	225	26	28	237.656	24	1	1	2	
697	1	0	0	0	2018-05-24	330	16	28	237.656	25	2	0	0	
698	0	0	0	1	2018-05-24	235	16	32	237.656	25	3	0	0	
699	0	0	0	1	2018-05-31	291	31	40	237.656	25	1	1	1	

700 rows × 14 columns



```
In [40]: df['Month Value'] = df['Date'].dt.month
```

```
In [41]: df
```

Out[41]:

	Reason_1	Reason_2	Reason_3	Reason_4	Date	Transportation Expense	Distance to Work	Age	Daily Work Load Average	Body Mass Index	Education	Children	Pets	Absenteeism Time (Hours)
--	----------	----------	----------	----------	------	---------------------------	---------------------	-----	----------------------------------	-----------------------	-----------	----------	------	-----------------------------

	Reason_1	Reason_2	Reason_3	Reason_4	Date	Transportation Expense	Distance to Work	Age	Daily Work Load Average	Body Mass Index	Education	Children	Pets	Absentee Time Hours
0	0	0	0	1	2015-07-07	289	36	33	239.554	30	1	2	1	
1	0	0	0	0	2015-07-14	118	13	50	239.554	31	1	1	0	
2	0	0	0	1	2015-07-15	179	51	38	239.554	31	1	0	0	
3	1	0	0	0	2015-07-16	279	5	39	239.554	24	1	2	0	
4	0	0	0	1	2015-07-23	289	36	33	239.554	30	1	2	1	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
695	1	0	0	0	2018-05-23	179	22	40	237.656	22	2	2	0	
696	1	0	0	0	2018-05-23	225	26	28	237.656	24	1	1	2	
697	1	0	0	0	2018-05-24	330	16	28	237.656	25	2	0	0	
698	0	0	0	1	2018-05-24	235	16	32	237.656	25	3	0	0	
699	0	0	0	1	2018-05-31	291	31	40	237.656	25	1	1	1	

700 rows × 15 columns



```
In [44]: df['Day of Week'] = df['Date'].dt.day_name()
```

```
In [45]: df
```

Out[45]:

	Reason_1	Reason_2	Reason_3	Reason_4	Date	Transportation Expense	Distance to Work	Age	Daily Work Load Average	Body Mass Index	Education	Children	Pets	Absentee Time Hours
0	0	0	0	1	2015-07-07	289	36	33	239.554	30	1	2	1	
1	0	0	0	0	2015-07-14	118	13	50	239.554	31	1	1	0	
2	0	0	0	1	2015-07-15	179	51	38	239.554	31	1	0	0	
3	1	0	0	0	2015-07-16	279	5	39	239.554	24	1	2	0	
4	0	0	0	1	2015-07-23	289	36	33	239.554	30	1	2	1	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
695	1	0	0	0	2018-05-23	179	22	40	237.656	22	2	2	0	
696	1	0	0	0	2018-05-23	225	26	28	237.656	24	1	1	2	
697	1	0	0	0	2018-05-24	330	16	28	237.656	25	2	0	0	
698	0	0	0	1	2018-05-24	235	16	32	237.656	25	3	0	0	
699	0	0	0	1	2018-05-31	291	31	40	237.656	25	1	1	1	

700 rows × 16 columns



```
In [46]: df['Education'].unique()
```



Out[46]: array([1, 3, 2, 4], dtype=int64)

```
In [47]: df['Education'].value_counts()
```

Out[47]: 1 583  
3 73  
2 40  
4 4  
Name: Education, dtype: int64

```
In [48]: df['Education'] = df['Education'].map({1:0, 2:1, 3:1, 4:1})
```

```
In [49]: df
```

Out[49]:

	Reason_1	Reason_2	Reason_3	Reason_4	Date	Transportation Expense	Distance to Work	Age	Daily Work Load Average	Body Mass Index	Education	Children	Pets	Absentee Time Hours
0	0	0	0	1	2015-07-07	289	36	33	239.554	30	0	2	1	
1	0	0	0	0	2015-07-14	118	13	50	239.554	31	0	1	0	
2	0	0	0	1	2015-07-15	179	51	38	239.554	31	0	0	0	
3	1	0	0	0	2015-07-16	279	5	39	239.554	24	0	2	0	
4	0	0	0	1	2015-07-23	289	36	33	239.554	30	0	2	1	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
695	1	0	0	0	2018-05-23	179	22	40	237.656	22	1	2	0	
696	1	0	0	0	2018-05-23	225	26	28	237.656	24	0	1	2	
697	1	0	0	0	2018-05-24	330	16	28	237.656	25	1	0	0	
698	0	0	0	1	2018-05-24	235	16	32	237.656	25	1	0	0	
699	0	0	0	1	2018-05-31	291	31	40	237.656	25	0	1	1	

700 rows x 16 columns



```
In [ ]:
```