11. Number of kids prediction - Preprocessing Data

- This is the collected data:

Number of Kids	Working Experience (years)	Age	Salary	Blood Types
3	15	45	$250,000	A
1	5	30	$200,000	B
2	10	38	$150,000	AB
1	<missing>	36	$180,000	O

- Process
  1. Please clean the missing data using median approach
  2. Please use Correlation to determine which of the following attributes is more related to "Number of Kids"?
     - Working Experience
     - Age
  3. Please use One-Hot Vectors approach to convert the Blood types
  4. Please use StandardScale to scale the data.
- Describe the process carefully in the document submitted as the homework answer.

Step 1: Clean the missing data:

Fixing the missing data using Median approach:

Working Experience (years)
15
5
10
<missing>

Find the median for the values i.e., 15, 5, 10.

For finding the median, we firstly arrange the numbers in ascending order.

For odd number set, the median is the middle number.

For even number set, the median is average of the two middle numbers.

Arranging the above numbers in ascending order – 5, 10, 15.

As it is an odd number set, the median is the middle number i.e., 10.

Fill the missing value with median i.e., 10.

Working Experience (years)
15
5
10
10

The table after cleaning the missing data is as below:

Number of Kids	Working Experience (years)	Age	Salary	Blood Types
3	15	45	$250,000	A
1	5	30	$200,000	B
2	10	38	$150,000	AB
1	10	36	$180,000	O

Step 2: Using Correlation to find which attribute is more related to "Number of kids" – Age or Working Experience(years)

Using the correlation formula:

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

where,

- r_{xy} – the correlation coefficient of the linear relationship between the variables x and y
- x_i – the values of the x-variable in a sample
- x – the mean of the values of the x-variable
- y_i – the values of the y-variable in a sample
- y – the mean of the values of the y-variable

Correlation between "Number of kids" and "Age"

Number of Kids (x_i)	**Age (y_i)**	**$x_i - \bar{x}$**	**$y_i - \bar{y}$**	**$(x_i - \bar{x})(y_i - \bar{y})$**	**$(x_i - \bar{x})^2$**	**$(y_i - \bar{y})^2$**
3	45	1.25	7.75	9.6875	1.5625	60.0625
1	30	-0.75	-7.25	5.4375	0.5625	52.5625
2	38	0.25	0.75	0.1875	0.0625	0.5625
1	36	-0.75	-1.25	0.9375	0.5625	1.5625
Mean						
$\bar{x} = 1.75$	$\bar{y} = 37.25$			16.25	2.75	114.75

$$r_{xy} = \frac{16.25}{\sqrt{(2.75*114.75)}} = \mathbf{0.9147}$$

Correlation between "Number of kids" and "Working Experience (years)"

Number of Kids (x_i)	**Working Experience(years) (y_i)**	**$x_i - \bar{x}$**	**$y_i - \bar{y}$**	**$(x_i - \bar{x})(y_i - \bar{y})$**	**$(x_i - \bar{x})^2$**	**$(y_i - \bar{y})^2$**
3	15	1.25	5	6.25	1.5625	25
1	5	-0.75	-5	3.75	0.5625	25
2	10	0.25	0	0	0.0625	0
1	10	-0.75	0	0	0.5625	0
Mean						
$\bar{x} = 1.75$	$\bar{y} = 10$			10	2.75	50

$$r_{xy} = \frac{10}{\sqrt{(2.75*50)}} = \mathbf{0.8528}$$

Comparing the correlation between Number of Kids, Age and Number of Kids, Working Experience in years,

Correlation between Number of Kids and Age = 0.9147
Correlation between Number of Kids and Working Experience in years = 0.8528

It is observed that Number of Kids is more correlated to Age i.e., has a correlation of 0.9147.

Step 3: One Hot Vectors to convert Blood types.

Number of Kids	Blood Types
3	A
1	B
2	AB
1	O

a. **Convert the text categories to integer categories.**

Number of Kids	Blood Types	Categorical Value
3	A	1
1	B	2
2	AB	3
1	O	4

b. **Convert from integer categories to One-Hot Vectors.**

Number of Kids	Blood Type A	Blood Type B	Blood Type AB	Blood Type O
3	1	0	0	0
1	0	1	0	0
2	0	0	1	0
1	0	0	0	1

The table after handling the text data is as below:

Number of Kids	Working Experience (years)	Age	Salary	Blood Type A	Blood Type B	Blood Type AB	Blood Type O
3	15	45	$250,000	1	0	0	0
1	5	30	$200,000	0	1	0	0
2	10	38	$150,000	0	0	1	0
1	10	36	$180,000	0	0	0	1

Step 4: Scaling the data using StandardScaler.

For Feature scaling using StandardScaler, we need to first compute the mean of each attribute and then find the standard deviation, after which we can scale the data.

Mean

$$\mu = \frac{1}{N}\sum_{i=1}^{N}(x_i)$$

Standard Deviation

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Standardization

$$z = \frac{x - \mu}{\sigma}$$

Let us find the scaled data for Number of kids column.

Number of samples, N = 4
Mean, $\mu = (3+1+2+1)/4 = 1.75$

$$\text{Variance, } \sigma^2 = \frac{\Sigma (x_i - \mu)^2}{N}$$

$$= \frac{(3-1.75)^2+(1-1.75)^2+(2-1.75)^2+(1-1.75)^2}{4}$$

$$= 0.6875$$

Standard Deviation, $\sigma = \sqrt{0.6875} = 0.829156$

Similarly, following the above same procedure, find the mean and standard deviation for all the data above.

	Number of Kids	**Working Experience (years)**	**Age**	**Salary**	**Blood Type A**	**Blood Type B**	**Blood Type AB**	**Blood Type O**
	3	15	45	$250,000	1	0	0	0
	1	5	30	$200,000	0	1	0	0
	2	10	38	$150,000	0	0	1	0
	1	10	36	$180,000	0	0	0	1
Mean(μ)	1.75	10	37.25	$195,000	0.25	0.25	0.25	0.25
Standard Deviation (σ)	0.8292	3.5355	5.3560	$36400.5494	0.4330	0.4330	0.4330	0.4330

Now, let us apply the StandardScaler formula to scale the data.

Standardization:

$$z = \frac{x - \mu}{\sigma}$$

For Number of kids column

- Z = $\frac{3-1.75}{0.8292}$ = 1.51
- Z = $\frac{1-1.75}{0.8292}$ = -0.90
- Z = $\frac{2-1.75}{0.8292}$ = 0.30
- Z = $\frac{1-1.75}{0.8292}$ = -0.90

For Working Experience column

- $Z = \dfrac{15-10}{3.5355}$
  $= 1.41$
- $Z = \dfrac{5-10}{3.5355}$
  $= -1.41$
- $Z = \dfrac{10-10}{3.5355}$
  $= 0$
- $Z = \dfrac{10-10}{3.5355}$
  $= 0$

For Age column

- $Z = \dfrac{45-37.25}{5.3560}$
  $= 1.45$
- $Z = \dfrac{30-37.25}{5.3560}$
  $= -1.35$
- $Z = \dfrac{38-37.25}{5.3560}$
  $= 0.14$
- $Z = \dfrac{36-37.25}{5.3560}$
  $= -0.23$

For Salary column

- $Z = \frac{250000-195000}{36400.5494}$

  $= 1.51$
- $Z = \frac{200000-195000}{36400.5494}$

  $= 0.14$
- $Z = \frac{150000-195000}{36400.5494}$

  $= -1.24$
- $Z = \frac{180000-195000}{36400.5494}$

  $= -0.41$

For Blood Type A column

- $Z = \frac{1-0.25}{0.433}$

  $= 1.73$
- $Z = \frac{0-0.25}{0.433}$

  $= -0.58$
- $Z = \frac{0-0.25}{0.433}$

  $= -0.58$
- $Z = \frac{0-0.25}{0.433}$

  $= -0.58$

For Blood Type B column

- $Z = \frac{0-0.25}{0.433}$ = -0.58
- $Z = \frac{1-0.25}{0.433}$ = 1.73
- $Z = \frac{0-0.25}{0.433}$ = -0.58
- $Z = \frac{0-0.25}{0.433}$ = -0.58

For Blood Type AB column

- $Z = \frac{0-0.25}{0.433}$ = -0.58
- $Z = \frac{0-0.25}{0.433}$ = -0.58
- $Z = \frac{1-0.25}{0.433}$ = 1.73
- $Z = \frac{0-0.25}{0.433}$ = -0.58

For Blood Type O column

- $Z = \frac{0-0.25}{0.433}$
  $= -0.58$
- $Z = \frac{0-0.25}{0.433}$
  $= -0.58$
- $Z = \frac{0-0.25}{0.433}$
  $= -0.58$
- $Z = \frac{1-0.25}{0.433}$
  $= 1.73$

Putting in all the scaled values into the table

Number of Kids	Working Experience (years)	Age	Salary	Blood Type A	Blood Type B	Blood Type AB	Blood Type O
1.51	1.41	1.45	$1.51	1.73	-0.58	-0.58	-0.58
-0.90	-1.41	-1.35	$0.14	-0.58	1.73	-0.58	-0.58
0.30	0	0.14	-$1.24	-0.58	-0.58	1.73	-0.58
-0.90	0	-0.23	$0.41	-0.58	-0.58	-0.58	1.73

The detailed process:

Step 1 – We started with **cleaning the data** i.e., looking for missing values in the table and ways to clean them.

We have used the method - fill the missing value with its median.

The median for the working experience is found. As it is an odd number of dataset, the median is the middle number when the numbers are arranged ascendingly. Hence the missing value is 10.

Step 2 - In this step, we check which attribute **is more correlated to Number of kids** from Working experience and Age.

Using the correlation formula, we find that the Number of kids is more correlated to Age with correlation of 0.9147.

Step 3 – We use the **One-Hot Vectors** to convert Blood types from Text category to numerical category and then to One-Hot Vectors.

Step 4 – In this step, we scale the data using StandardScaler. The formula for the same is mentioned above. After which each data is replaced with the scaled data.

Performing above operations using Scikit-learn.

Step 1 – Importing necessary libraries, reading the data and describing the data

Importing necessary libraries

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

Reading the data

In [2]:
```
data = pd.read_csv("data.csv")
```

Describing the data

In [3]:
```
data.head()
```

Out[3]:

	Number of Kids	Working Experience(years)	Age	Salary	Blood Types
0	3	15.0	45	250000	A
1	1	5.0	30	200000	B
2	2	10.0	38	150000	AB
3	1	NaN	36	180000	O

In [4]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4 entries, 0 to 3
Data columns (total 5 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Number of Kids             4 non-null      int64
 1   Working Experience(years)  3 non-null      float64
 2   Age                        4 non-null      int64
 3   Salary                     4 non-null      int64
 4   Blood Types                4 non-null      object
dtypes: float64(1), int64(3), object(1)
memory usage: 288.0+ bytes
```

Step 2 – Filling missing data with its median value.

Filling missing data with median

In [5]:
```
data["Working Experience(years)"].fillna(data["Working Experience(years)"].median(),inplace=True)
```

In [6]:
```
data.head()
```

Out[6]:

	Number of Kids	Working Experience(years)	Age	Salary	Blood Types
0	3	15.0	45	250000	A
1	1	5.0	30	200000	B
2	2	10.0	38	150000	AB
3	1	10.0	36	180000	O

In [7]:
```
data["Working Experience(years)"]=data["Working Experience(years)"].astype(int)
```

In [8]:
```
data.head()
```

Out[8]:

	Number of Kids	Working Experience(years)	Age	Salary	Blood Types
0	3	15	45	250000	A
1	1	5	30	200000	B
2	2	10	38	150000	AB
3	1	10	36	180000	O

Step 3 – Finding correlation between Number of kids and Age, Number of kids and Working experience.

Finding Correlation between Number of Kids and Age, Number of Kids and Working Experience ¶

In [9]:
```
data["Number of Kids"].corr(data["Age"])
```

Out[9]: 0.9147673836616229

In [10]:
```
data["Working Experience(years)"].corr(data["Number of Kids"])
```

Out[10]: 0.8528028654224419

Age is more related to Number of Kids - 0.91 correlation

Step 4 – One-Hot Vectors for the Blood types.

One Hot Vectors

In [11]:
```
blood_types_encoded,categories=data["Blood Types"].factorize()
blood_types_encoded
```

Out[11]:
```
array([0, 1, 2, 3], dtype=int64)
```

In [12]:
```
from sklearn.preprocessing import OneHotEncoder
encoder = OneHotEncoder(sparse=False)
blood_type_cat_1hot = encoder.fit_transform(blood_types_encoded.reshape(-1,1))
blood_type_cat_1hot
```

Out[12]:
```
array([[1., 0., 0., 0.],
       [0., 1., 0., 0.],
       [0., 0., 1., 0.],
       [0., 0., 0., 1.]])
```

In [13]:
```
data.head()
```

Out[13]:

	Number of Kids	Working Experience(years)	Age	Salary	Blood Types
0	3	15	45	250000	A
1	1	5	30	200000	B
2	2	10	38	150000	AB
3	1	10	36	180000	O

In [14]:
```
one_hot=pd.get_dummies(data,columns=["Blood Types"],drop_first=False,prefix='',prefix_sep='')
one_hot
```

Out[14]:

	Number of Kids	Working Experience(years)	Age	Salary	A	AB	B	O
0	3	15	45	250000	1	0	0	0
1	1	5	30	200000	0	0	1	0
2	2	10	38	150000	0	1	0	0
3	1	10	36	180000	0	0	0	1

In [15]:
```
data=data.drop("Blood Types",axis=1)
data
```

Out[15]:

	Number of Kids	Working Experience(years)	Age	Salary
0	3	15	45	250000
1	1	5	30	200000
2	2	10	38	150000
3	1	10	36	180000

In [16]:
```
data=data.join(one_hot["A"])
data=data.join(one_hot["B"])
data=data.join(one_hot["AB"])
data=data.join(one_hot["O"])
```

In [17]:
```
data
```

Out[17]:

	Number of Kids	Working Experience(years)	Age	Salary	A	B	AB	O
0	3	15	45	250000	1	0	0	0
1	1	5	30	200000	0	1	0	0
2	2	10	38	150000	0	0	1	0
3	1	10	36	180000	0	0	0	1

Step 5 – Scaling the data

Scaling the data

In [18]:
```
from sklearn.preprocessing import StandardScaler
```

In [24]:
```
scaler = StandardScaler()
data_scaled = pd.DataFrame(scaler.fit_transform(data),columns=["Number of Kids","Working Experience(years)","Age",
                                                              "Salary","Blood Type A","Blood Type B","Blood Type AB",
                                                              "Blood Type O"])
```

In [25]:
```
data_scaled
```

Out[25]:

	Number of Kids	Working Experience(years)	Age	Salary	Blood Type A	Blood Type B	Blood Type AB	Blood Type O
0	1.507557	1.414214	1.446956	1.510966	1.732051	-0.577350	-0.577350	-0.577350
1	-0.904534	-1.414214	-1.353604	0.137361	-0.577350	1.732051	-0.577350	-0.577350
2	0.301511	0.000000	0.140028	-1.236245	-0.577350	-0.577350	1.732051	-0.577350
3	-0.904534	0.000000	-0.233380	-0.412082	-0.577350	-0.577350	-0.577350	1.732051

In [26]:
```
data_scaled.round(decimals=2)
```

Out[26]:

	Number of Kids	Working Experience(years)	Age	Salary	Blood Type A	Blood Type B	Blood Type AB	Blood Type O
0	1.51	1.41	1.45	1.51	1.73	-0.58	-0.58	-0.58
1	-0.90	-1.41	-1.35	0.14	-0.58	1.73	-0.58	-0.58
2	0.30	0.00	0.14	-1.24	-0.58	-0.58	1.73	-0.58
3	-0.90	0.00	-0.23	-0.41	-0.58	-0.58	-0.58	1.73

The same output is obtained with manual calculation and with scikit-learn.
The process with scikit-learn is faster and error-free when it comes to large data.