

Data Preprocessing



PHOSPHERE AI

Contents

- Filling missing values
- Converting Categorical Data to Numerical format

The dataset we'll be using

Country	Age	Salary	Size	Purchased
France	44	72000	M	No
Spain	27	48000	S	Yes
Germany	30	54000	M	No
Spain	38	61000	L	No
Germany	40		XL	Yes
France	35	58000	XXL	Yes
Spain		52000	M	No
France	48	79000	M	Yes
Germany	50	83000	L	No
France	37	67000	S	Yes

Nominal Data

Ordinal Data



Filling in Missing Values

Data Imputation Methods

- Mean Imputation – Calculate the mean of the values in the same column and fill the missing value with the mean
- Median Imputation - Calculate the median of the values in the same column and fill the missing value with the median
- Hot Deck Imputation – Randomly choose a value from the column from another row and fill the missing cell (NOT RECOMMENDED)
- Cold Deck Imputation – Systematically choose value from another row that has similar properties on other columns and then fill the missing cell
- Manual Imputation – Manually examine the row and fill the value

We'll use the Mean Imputation strategy

Mean = 38.778

Mean = 63777.78

Country	Age	Salary	Size	Purchased
France	44	72000	M	No
Spain	27	48000	S	Yes
Germany	30	54000	M	No
Spain	38	61000	L	No
Germany	40		XL	Yes
France	35	58000	XXL	Yes
Spain		52000	M	No
France	48	79000	M	Yes
Germany	50	83000	L	No
France	37	67000	S	Yes

We'll use the Mean Imputation strategy

Country	Age	Salary	Size	Purchased
France	44	72000	M	No
Spain	27	48000	S	Yes
Germany	30	54000	M	No
Spain	38	61000	L	No
Germany	40	63777.78	XL	Yes
France	35	58000	XXL	Yes
Spain	38.778	52000	M	No
France	48	79000	M	Yes
Germany	50	83000	L	No
France	37	67000	S	Yes

The slide features two large teal geometric shapes. On the left, a teal triangle points upwards from the bottom-left corner. On the right, a teal trapezoid is positioned, wider at the top and narrower at the bottom, pointing upwards from the bottom-right corner. The text is centered in the white space between these shapes.

Converting Categorical Data to Numerical format

The dataset we'll be using

Country	Age	Salary	Size	Purchased
France	44	72000	M	No
Spain	27	48000	S	Yes
Germany	30	54000	M	No
Spain	38	61000	L	No
Germany	40	63777.78	XL	Yes
France	35	58000	XXL	Yes
Spain	38.778	52000	M	No
France	48	79000	M	Yes
Germany	50	83000	L	No
France	37	67000	S	Yes

Nominal Data

Ordinal Data

Let's start with encoding Ordinal Data

Size
M
S
M
L
XL
XXL
M
M
L
S

The values in the column are S, M, L, XL, XXL

We can replace each value like this:

- S - 1
- M - 2
- L - 3
- XL - 4
- XXL - 5

Let's start with encoding Ordinal Data

Size
M
S
M
L
XL
XXL
M
M
L
S

We can replace each value like this:

- S – 1
- M – 2
- L – 3
- XL – 4
- XXL – 5



Size
2
1
2
3
4
5
2
2
3
1

Now the dataset will be

Country	Age	Salary	Size	Purchased
France	44	72000	2	No
Spain	27	48000	1	Yes
Germany	30	54000	2	No
Spain	38	61000	3	No
Germany	40	63777.78	4	Yes
France	35	58000	5	Yes
Spain	38.778	52000	2	No
France	48	79000	2	Yes
Germany	50	83000	3	No
France	37	67000	1	Yes

Nominal Data

Ordinal Data

Let's move to the
'Country' column which
contains Nominal Data

Country
France
Spain
Germany
Spain
Germany
France
Spain
France
Germany
France

The values in the column are France,
Spain, Germany

We already saw that we can't assign 1,
2, 3.. To the values because the values in
that column don't have any numerical
significance.

So, we'll create three columns called
isGermany, isFrance and isSpain.

Let's move to the
'Country' column which
contains Nominal Data

Country
France
Spain
Germany
Spain
Germany
France
Spain
France
Germany
France

One-hot encode



isFrance	isSpain	isGermany
1	0	0
0	1	0
0	0	1
0	1	0
0	0	1
1	0	0
0	1	0
1	0	0
0	0	1
1	0	0

Finally, we remove the 'Country' column and then add the newly created 3 columns using one-hot encoding

isFrance	isSpain	isGermany	Age	Salary	Size	Purchased
1	0	0	44	72000	2	No
0	1	0	27	48000	1	Yes
0	0	1	30	54000	2	No
0	1	0	38	61000	3	No
0	0	1	40	63777.78	4	Yes
1	0	0	35	58000	5	Yes
0	1	0	38.778	52000	2	No
1	0	0	48	79000	2	Yes
0	0	1	50	83000	3	No
1	0	0	37	67000	1	Yes

One-hot encoded

Imputed
with Mean

Ordinal
Encoding