Transforming Continuous and Categorical Data



Janani Ravi CO-FOUNDER, LOONYCORN www.loonycorn.com

Overview

Categorical data vs. continuous data

Nominal vs. ordinal data

Scaling numeric features for data analysis

Represent categorical data using label encoding and one-hot encoding

Perform discretization to convert continuous data to categorical values

Types of Data

Categorical

Male/Female, Month of year

Numeric (Continuous)

Weight in lbs, Temperature in F

All other forms of data, such as text and image data, must be converted to one of these forms

Numeric (Continuous) vs. Categorical Data

Numeric (Continuous)

E.g. height or weight of individuals

Can take any value

Predicted using regression models

Always can be sorted on magnitude

Categorical

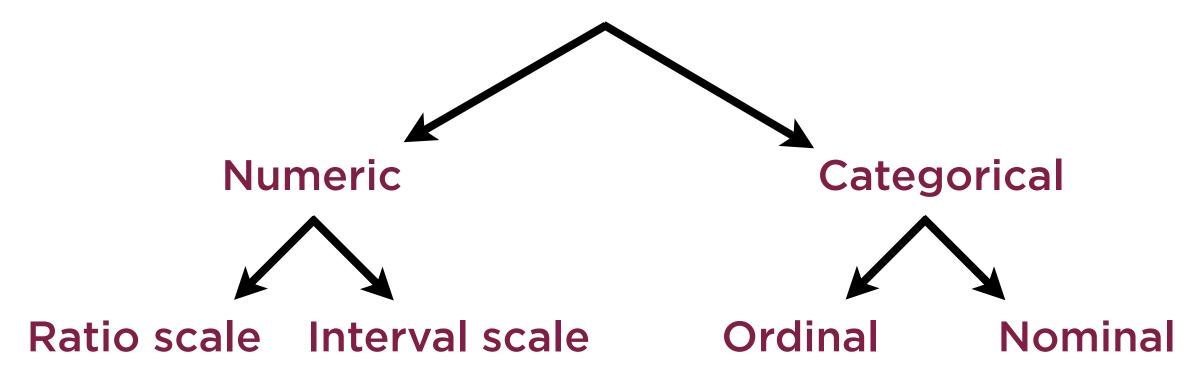
E.g. day of week, month of year, gender, letter grade

Finite set of permissible values

Predicted using classification models

Categories may or may not be sortable

Numeric Data



Numerical Data

Discrete

Cannot be measured but can be counted

Continuous

Cannot be counted but can be measured

Numerical Data

Discrete

Cannot be measured but can be counted

Continuous

Cannot be counted but can be measured

Number of visitors in an hour, number of heads when a coin is flipped 100 times

Numerical Data

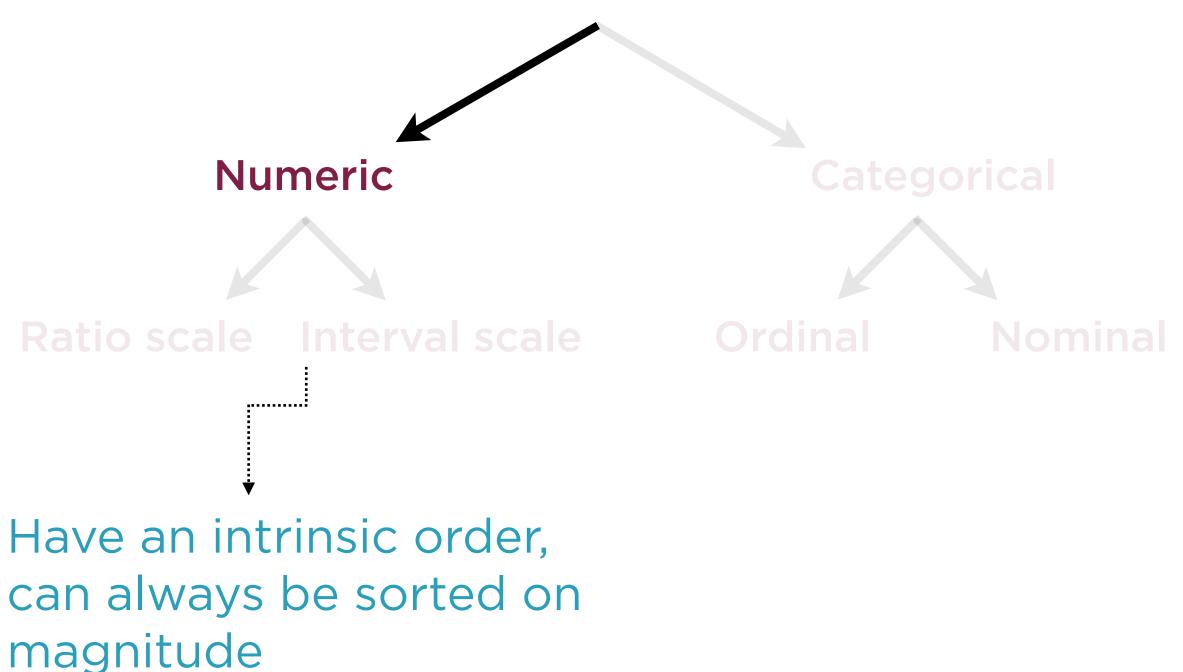
Discrete

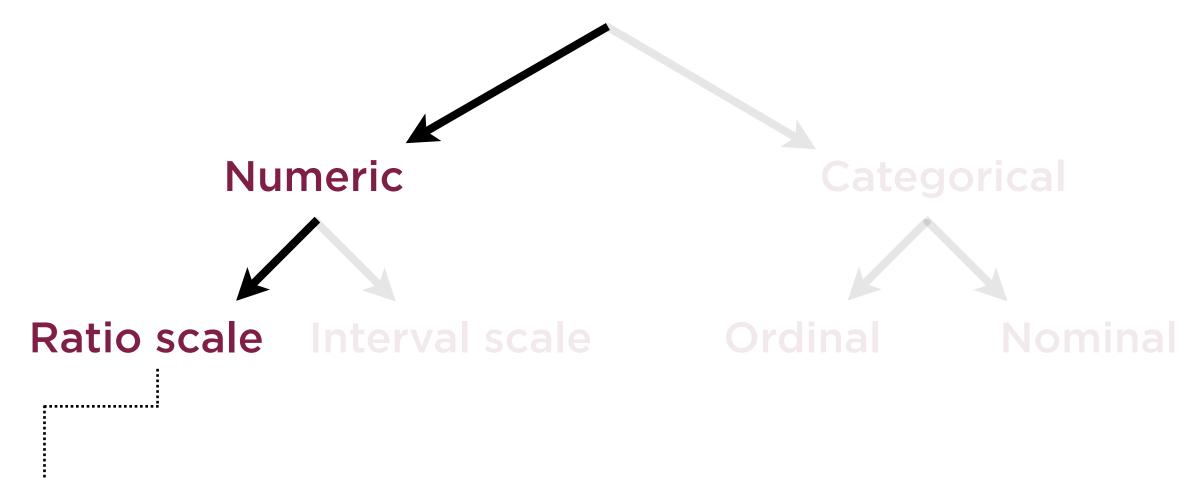
Cannot be measured but can be

Continuous

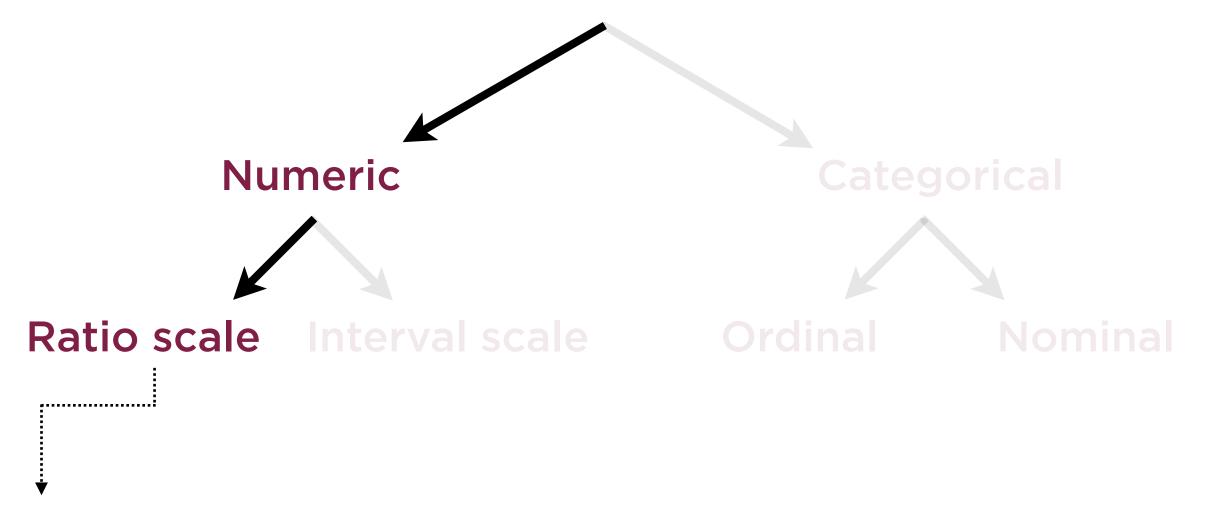
Cannot be counted but can be measured

Height of an individual, home prices, stock prices

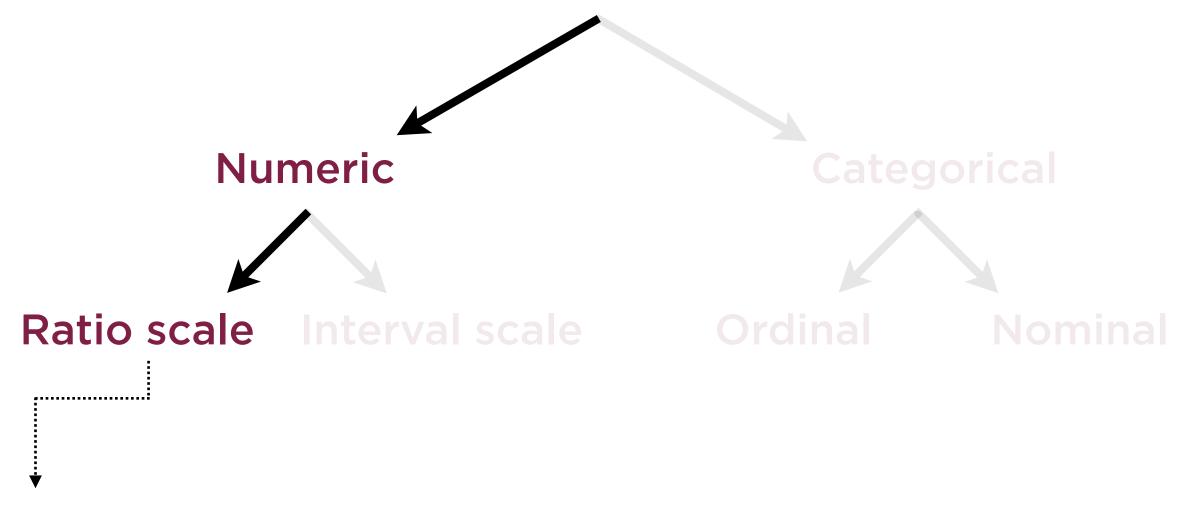




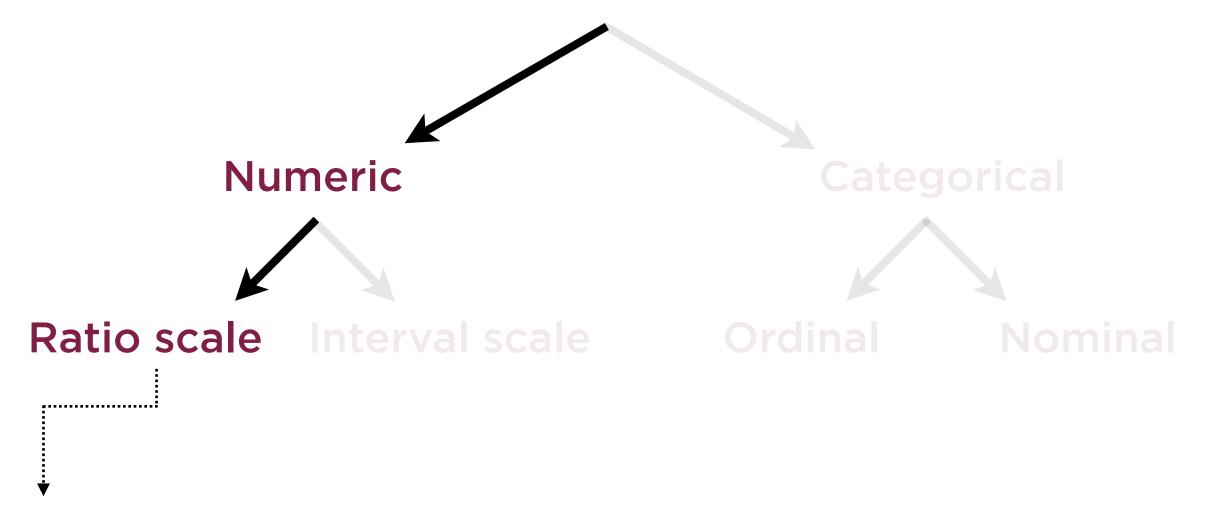
"Usual" numeric data, expressed as ratio to 1 e.g. 7 == 7:1



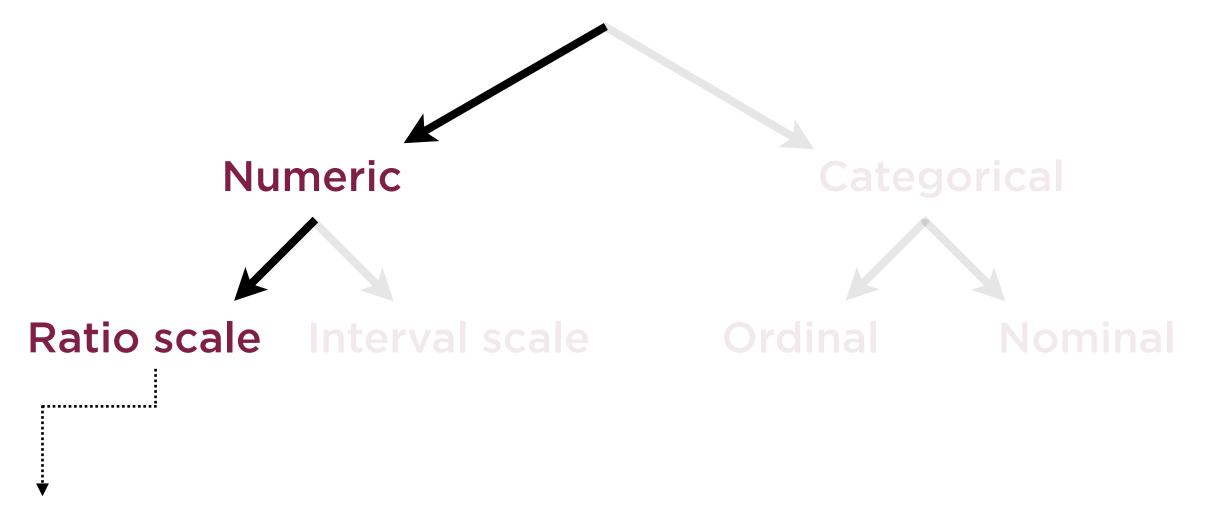
All arithmetic operations apply: addition, subtraction, multiplication and division



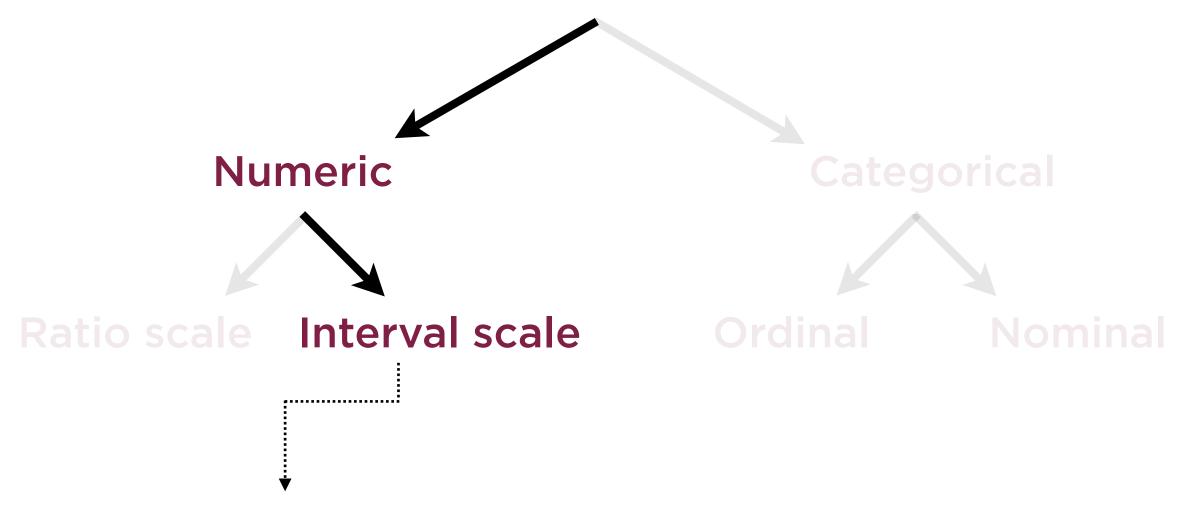
E.g. weight of 20 lbs is twice as much as a weight of 10 lbs



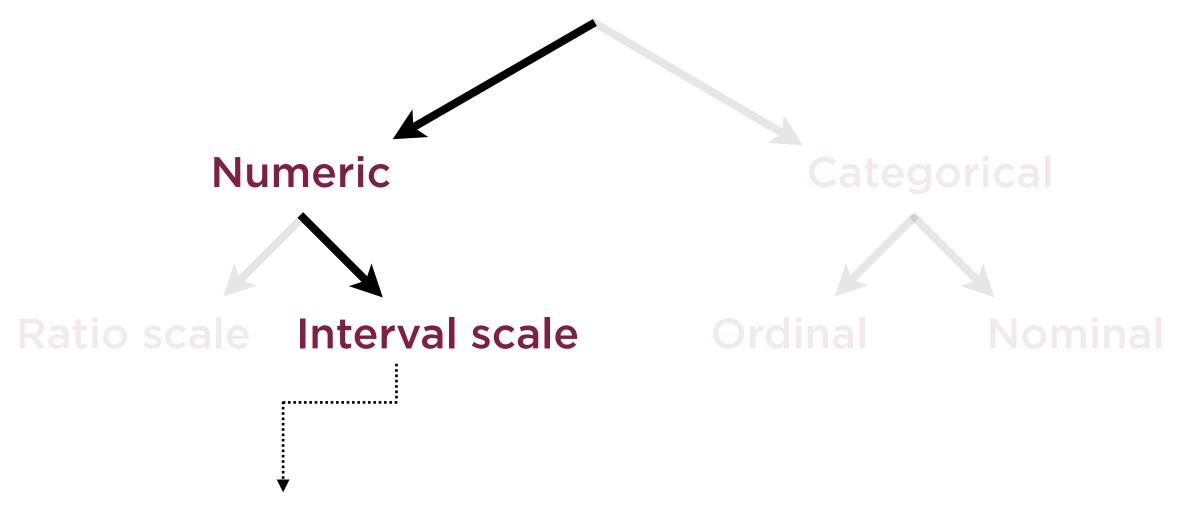
Ratio scale data has a meaningful zero point (the only type of data in this chart that does)



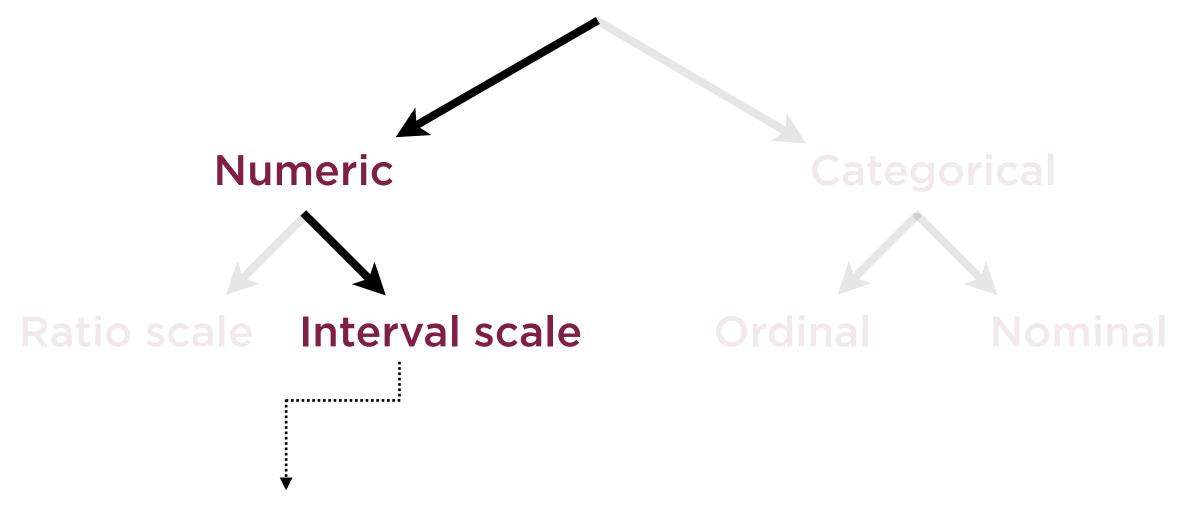
Weight of 0 lbs is equivalent to "no weight"



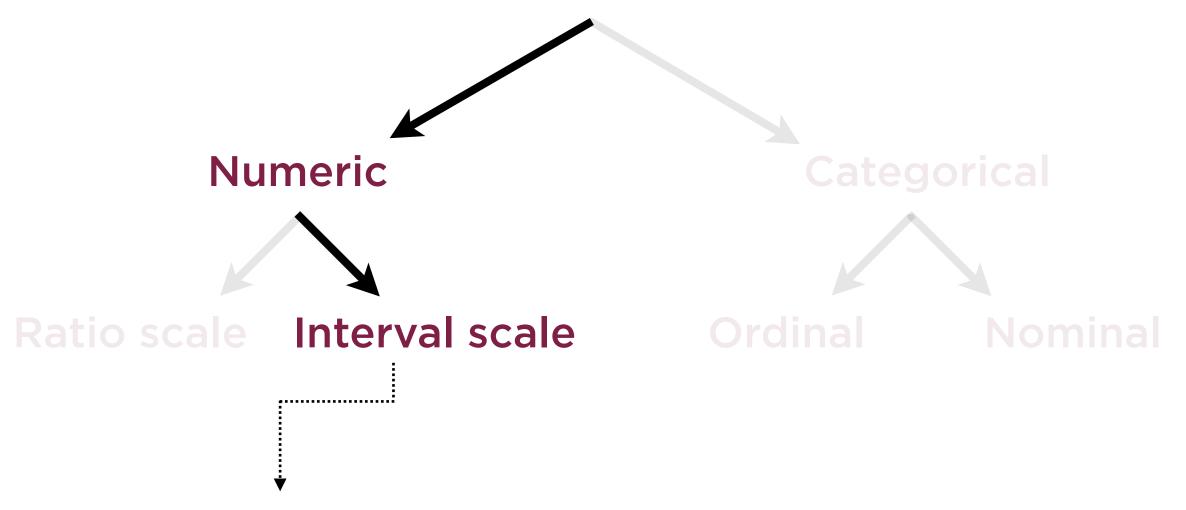
Ordered units that have the same difference i.e. the interval



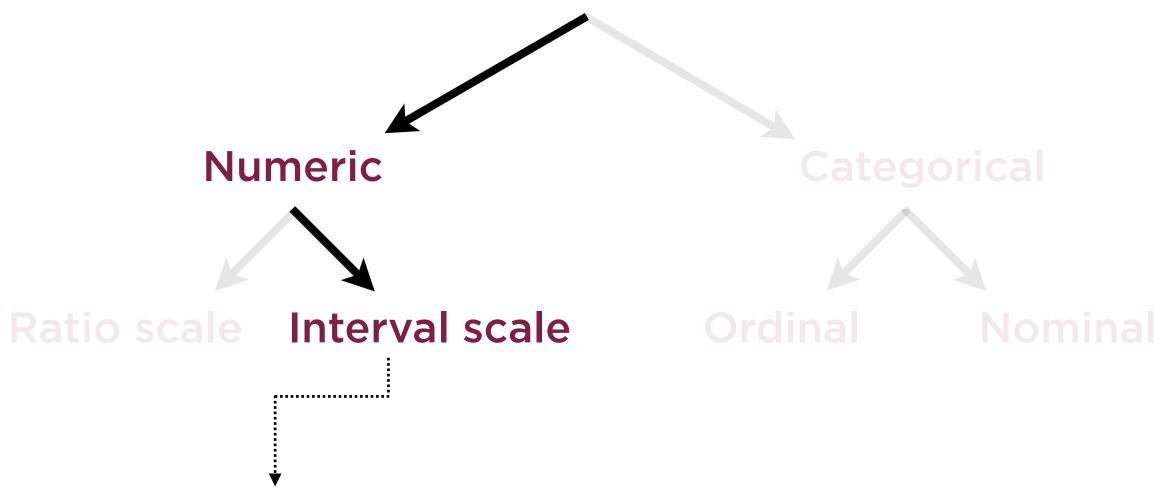
Data still numeric, but now multiplication and division no longer make sense, and zero point no longer meaningful



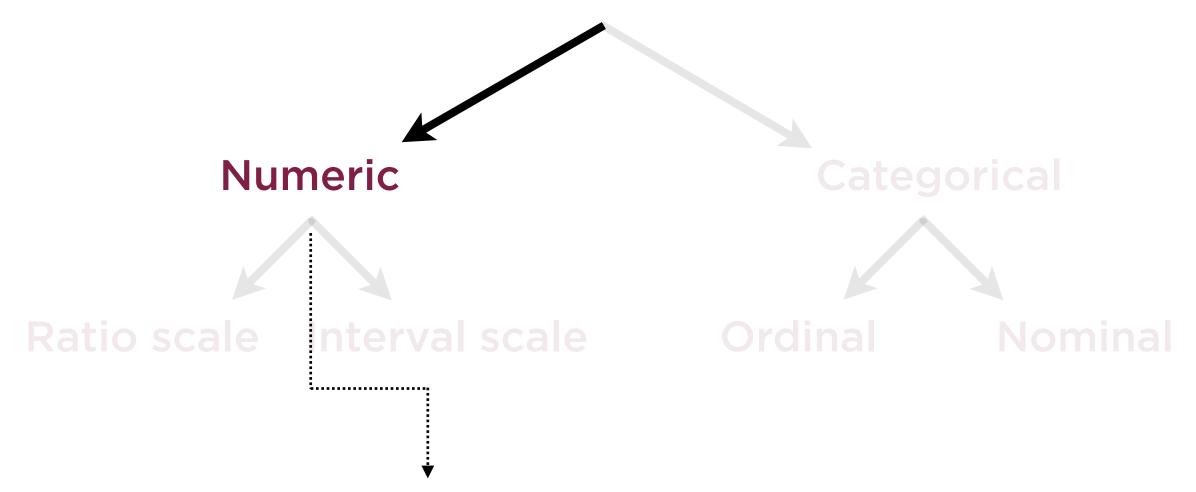
Difference between 90 Fahrenheit and 30 Fahrenheit is equal to 60 Fahrenheit



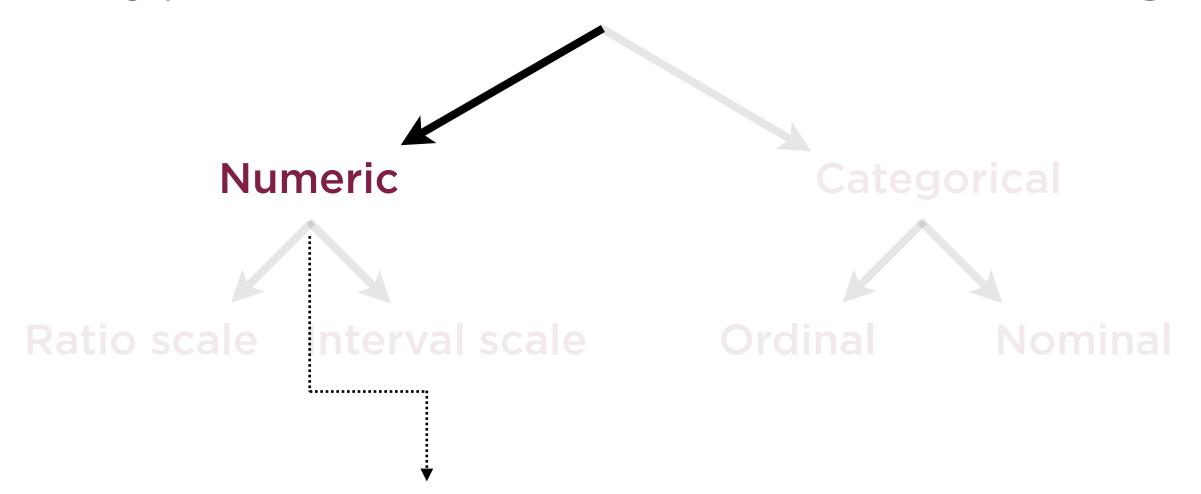
But temperature of 90 Fahrenheit is not thrice temperature of 30 Fahrenheit



O Fahrenheit is not equivalent to "no temperature"



Numeric data can draw from an unrestricted range of continuous values



Can calculate mean, standard deviation, correlation etc.

Machine learning algorithms typically do not work well with numeric data with different scales

Feature Scaling

Scaling Sta

Standardization

Feature Scaling

Scaling

Standardization

Numeric values are shifted and rescaled so all features have the same scale i.e. within the same minimum and maximum values

Feature Scaling

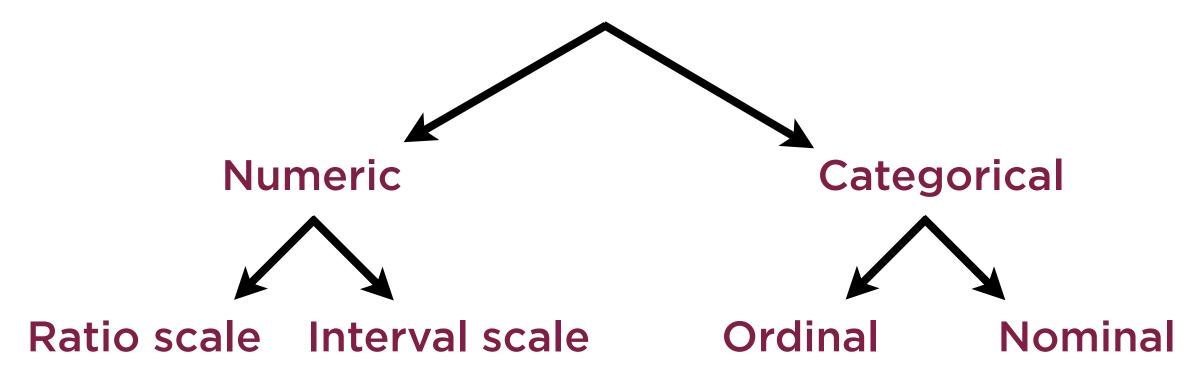
Scaling Standardization

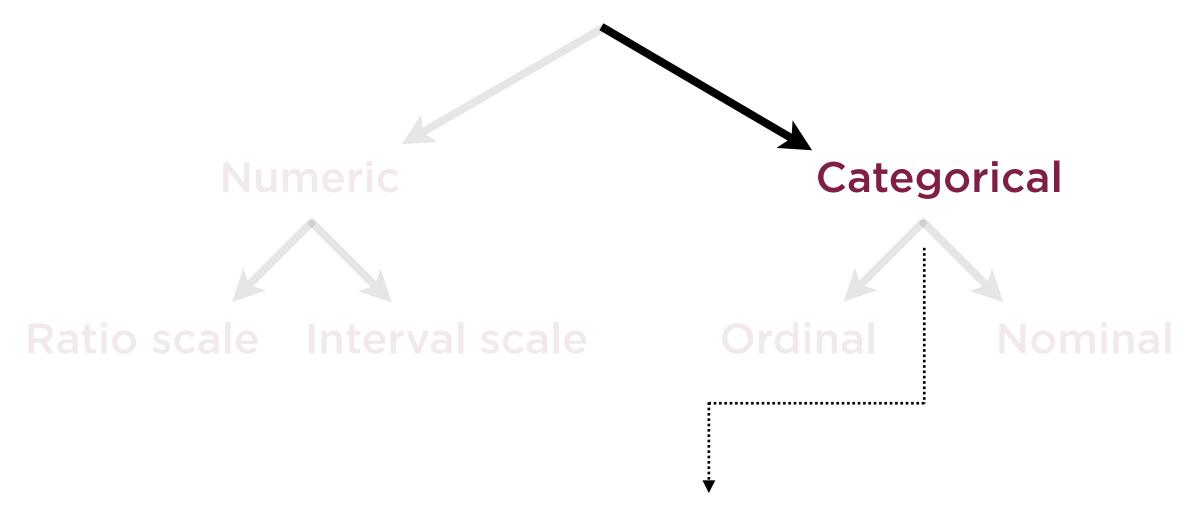
Centers data round the mean and divides each value by the standard deviation so all features have 0 mean and unit variance

Demo

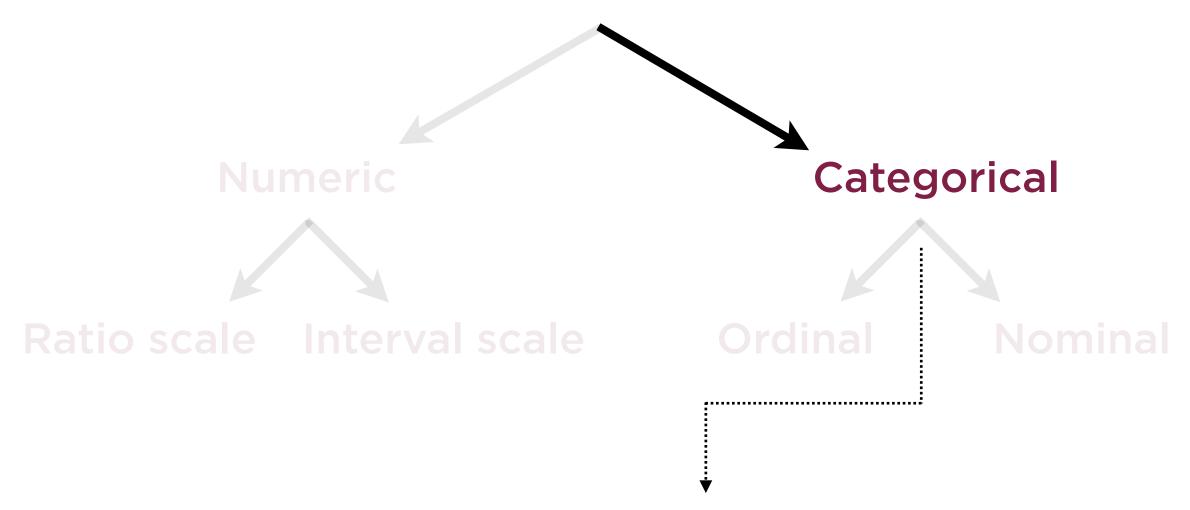
Performing feature scaling and transformation using different techniques

Categorical Data

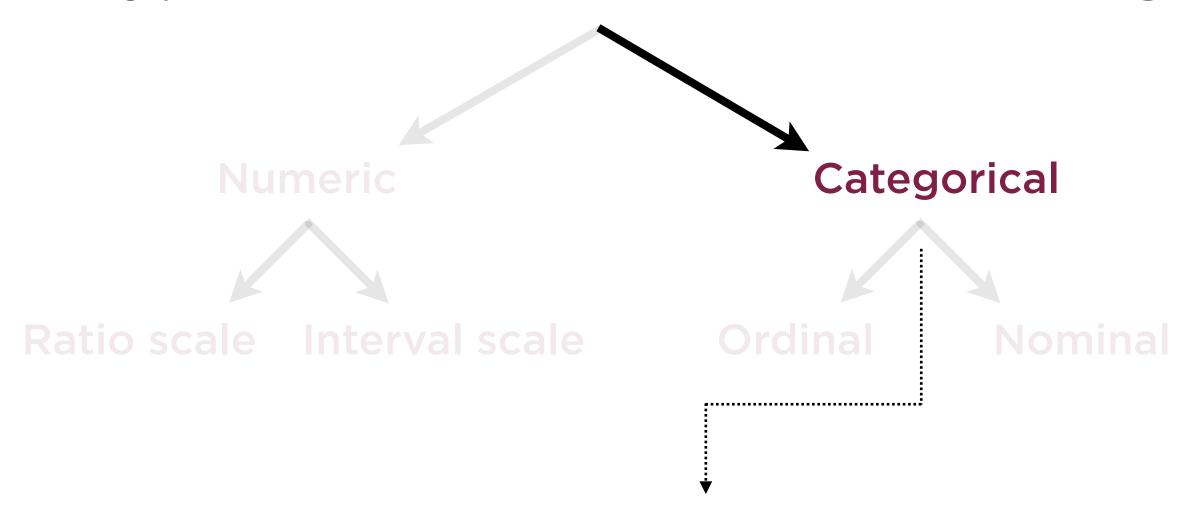




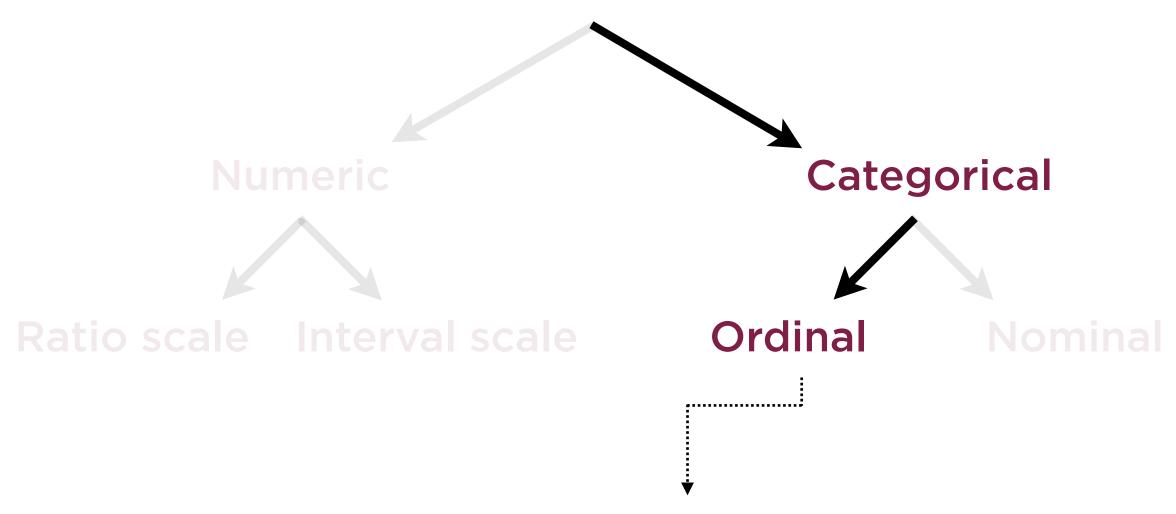
Categorical data can only draw from a specific, restricted set of values



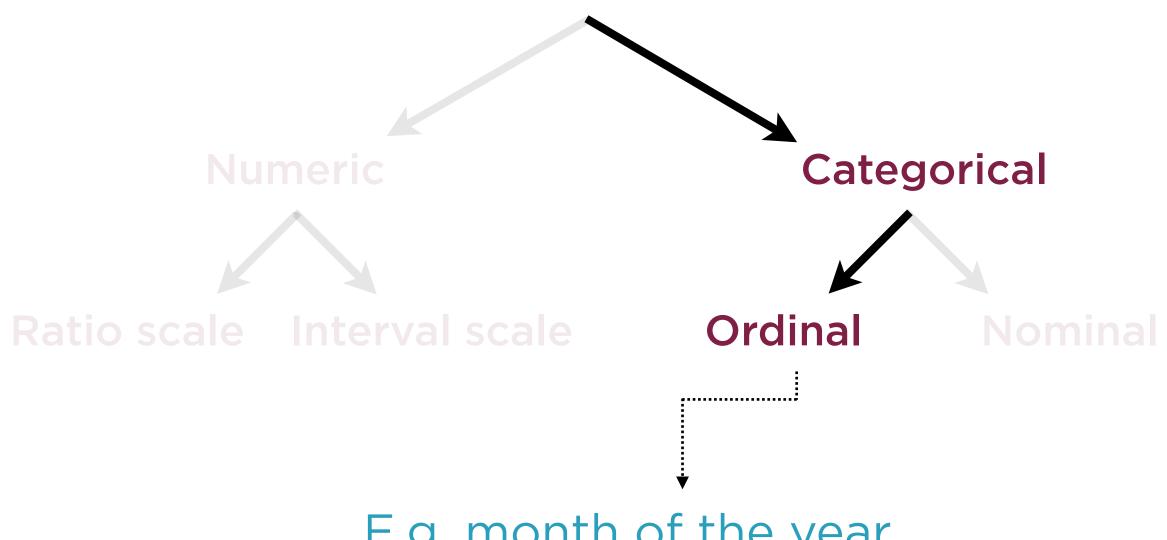
Not meaningful to calculate mean, standard deviation, correlation



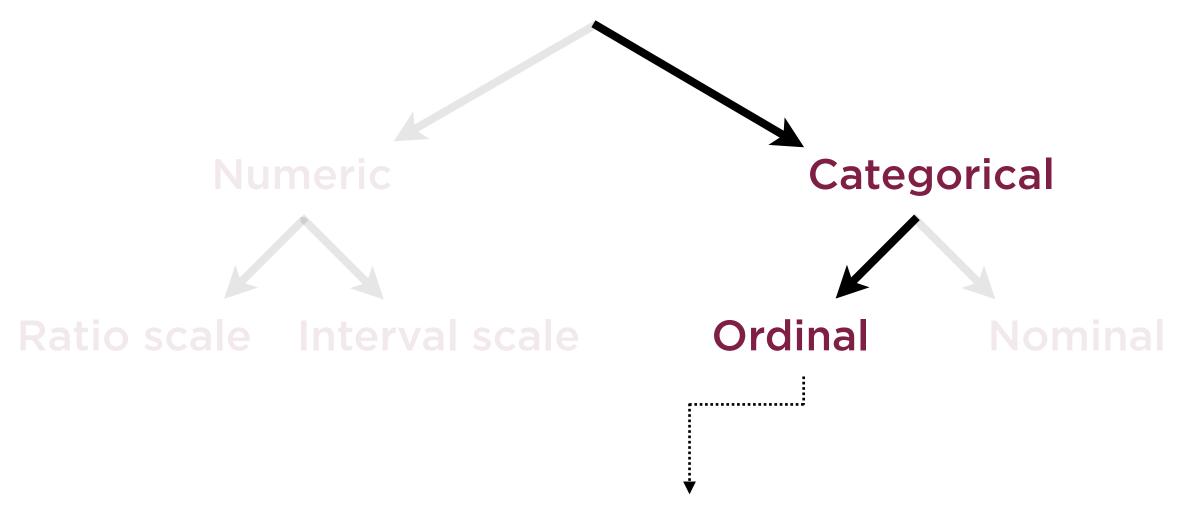
Fine to tabulate categorical data using count frequencies and percentages



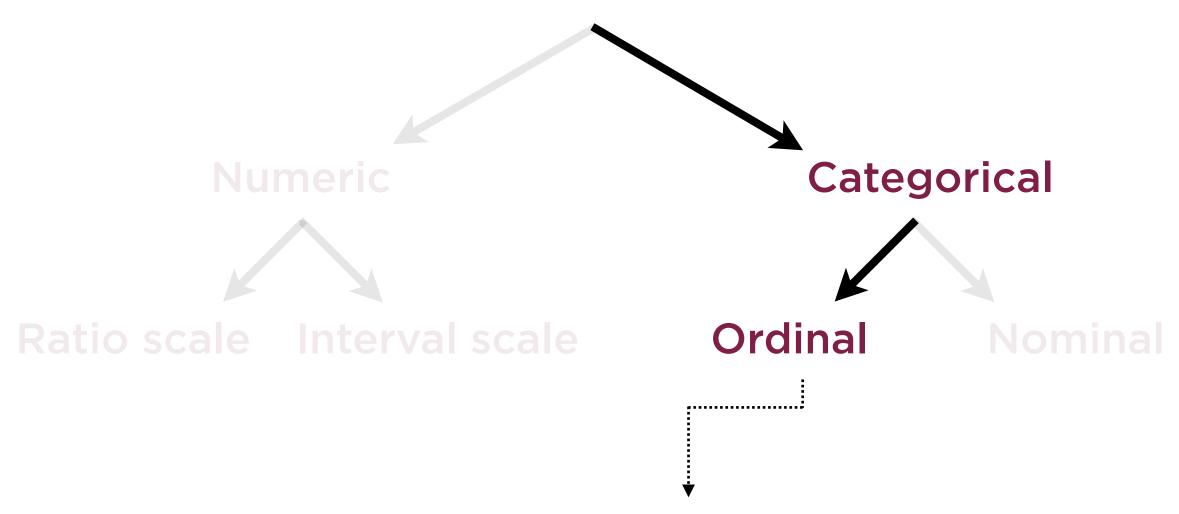
Ordinal data is categorical, but can still be ordered



E.g. month of the year, ratings on a scale of 1 to 5

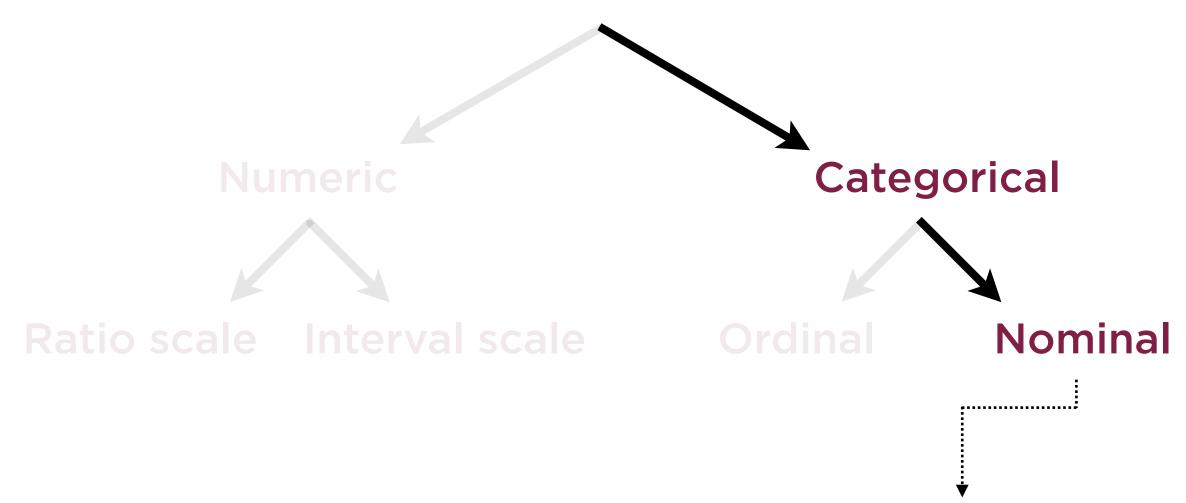


Order exists, but differences are not necessarily meaningful



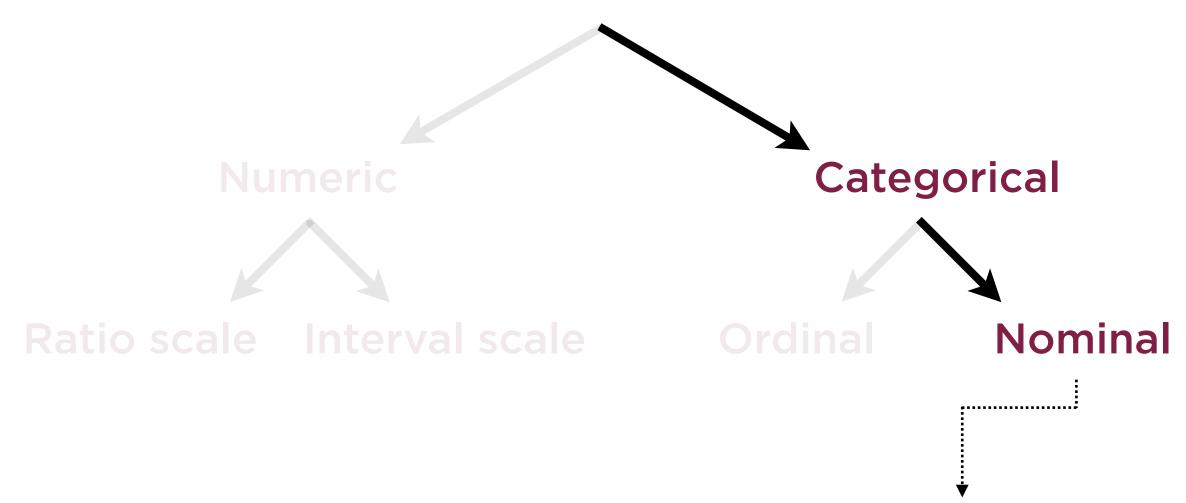
E.g. Differences in quality between three, two, one, and no Michelin stars for a restaurant are not uniform

Types of Data in Machine Learning



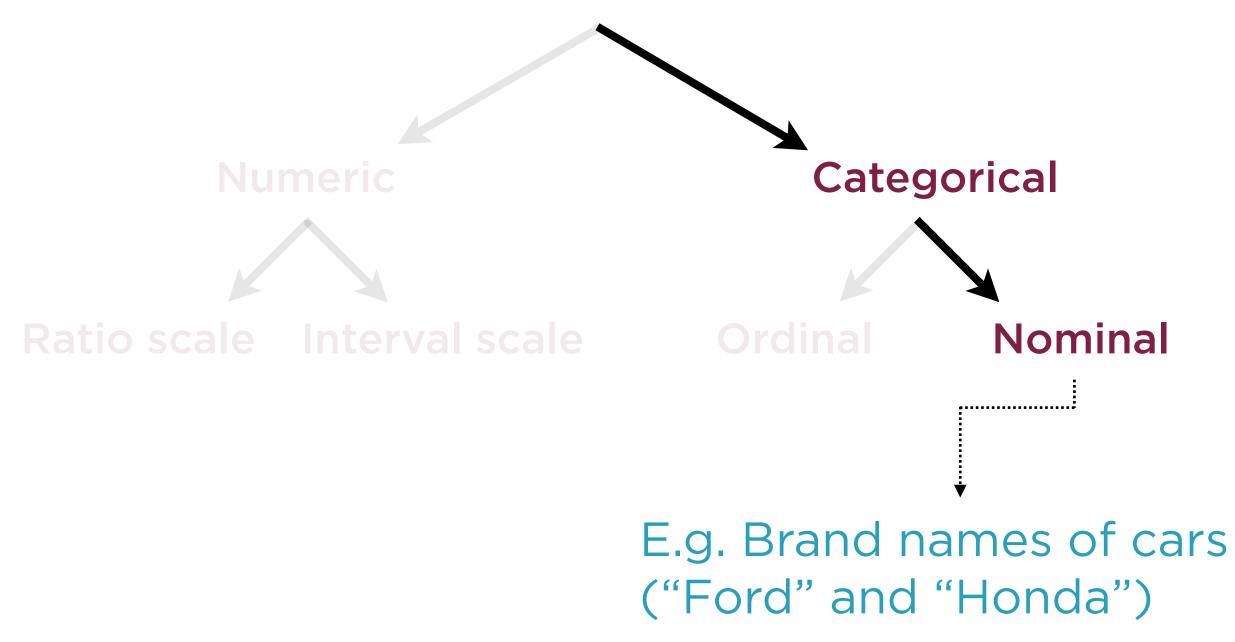
Even less in common with numeric data - cannot even be ordered

Types of Data in Machine Learning



Ordinal data can at least be <u>ord</u>ered; <u>nominal data are simply names</u>

Types of Data in Machine Learning



Categorical data has to be numerically encoded before it can be used in ML models

Representing Categorical Data

['New York', 'London', 'Paris', 'Bangalore']

Categorical Data

Classes often represented in string format

Categories as Nominal Data

Label encoding

Numeric id for each category; single column suffices

One-hot encoding

Separate column with 1 or 0 for presence/absence of each category

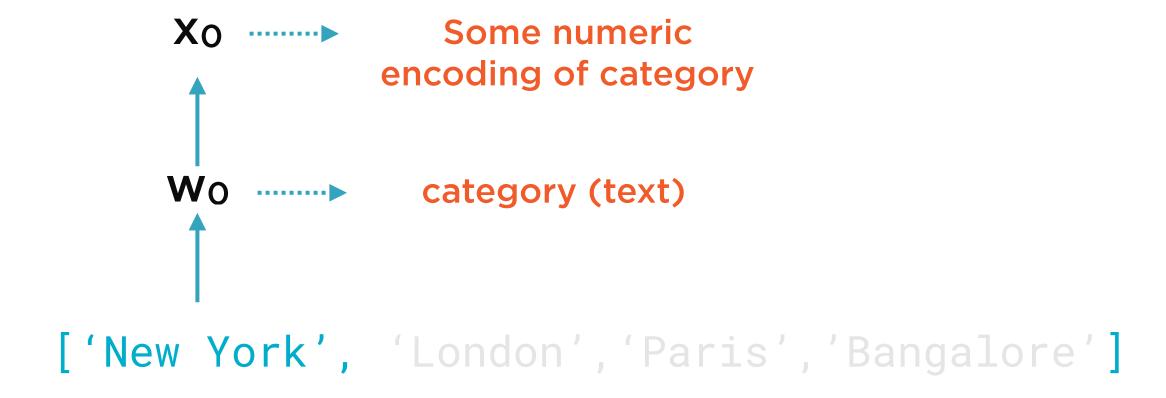
Categories as Nominal Data

Label encoding

Numeric id for each category; single column suffices

One-hot encoding

Separate column with 1 or 0 for presence/absence of each category



Categorical Data

Represent each category using some numeric encoding

```
32

wo

New York', 'London', 'Paris', 'Bangalore']
```

Represent Each Category as a Number

Represent Each Category as a Number

```
1056

| W3
| | |
| (New York', 'London', 'Paris', 'Bangalore')
```

Represent Each Category as a Number

Categories as Nominal Data

Label encoding

Numeric id for each category; single column suffices

One-hot encoding

Separate column with 1 or 0 for presence/absence of each category

['New York', 'London', 'Paris', 'Bangalore']

Categorical Data

Classes often represented in string format

 $x_i = 0 \text{ or } 1$

One-hot Encoding of 1 Category

Represent each category with a binary variable

 $x_i = 0 \text{ or } 1$

One-hot Encoding of 1 Category

Need as many columns as categories in the data

New York	London	Paris	Bangalore

Category	New York	London	Paris	Bangalore
New York				
London				
Paris				
Bangalore				

Category	New York	London	Paris	Bangalore
New York	1	0	O	0
London				
Paris				
Bangalore				

Category	New York	London	Paris	Bangalore
New York	1	0	0	0
London	O	1	0	0
Paris				
Bangalore				

Category	New York	London	Paris	Bangalore
New York	1	0	0	O
London	O	1	0	0
Paris	O	0	1	0
Bangalore				

Category	New York	London	Paris	Bangalore
New York	1	0	O	O
London	0	1	O	O
Paris	O	O	1	0
Bangalore	0	0	0	1

Category	New York	London	Paris	Bangalore
New York	1	0	O	0
London	O	1	0	0
Paris	O	0	1	0
Bangalore	0	0	0	1

Label Encoding vs. One-hot Encoding

Label Encoding

Single column to represent categories

Each category takes numeric value

More concise

One-hot Encoding

Need as many columns as categories in the data

Each category is a row with single 1 rest Os

Verbose - especially as number of categories grows

Label Encoding vs. One-hot Encoding

Label Encoding

Numeric ids present illusion of sortability

Ideally should use only for ordinal categorical data

One-hot Encoding

One-hot encoded vectors are clearly not sortable

Can use for both nominal and ordinal categorical data

Demo

Convert categorical data to numeric form using label encoding and one-hot encoding

Demo

Convert continuous data to categorical form using discretization

Summary

Categorical data vs. continuous data

Nominal vs. ordinal data

Scaling numeric features for data analysis

Represent categorical data using label encoding and one-hot encoding

Perform discretization to convert continuous data to categorical values