

Applied Data Analytics

Descriptive Statistics / Pandas

Types of data and data types

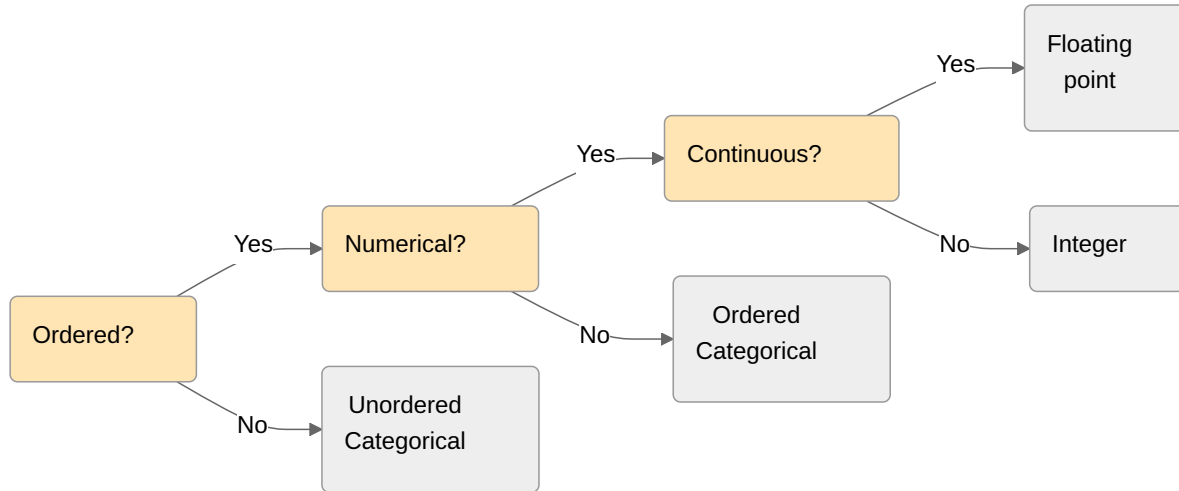
Hans-Martin von Gaudecker and Aapo Stenhammar

Nominal (qualitative, categorical) data

Example

- Variable: Country of origin
- Possible values: Names of countries in Latin America
- Observed values: Argentina, Bolivia, Argentina

Decision tree



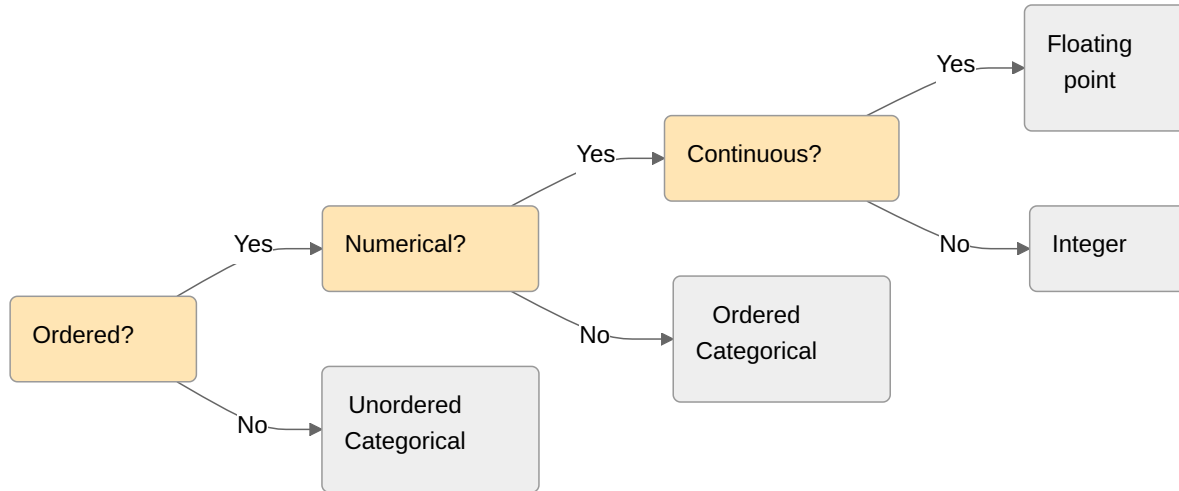
```
pd.Series(  
    [  
        "Argentina",  
        "Bolivia",  
        "Argentina",  
    ],  
    dtype=pd.CategoricalDtype(  
        [  
            "Argentina",  
            "...",  
            "Peru",  
        ],  
        ordered=False  
    )  
)
```

Cardinal, continuous data

Example:

- Variable: Annual Income in Euros
- Possible values: Positive real numbers
- Observed values: 42,345 €, 53,724 €, 28,734 €

Decision tree



Cardinal, continuous data

Examples

Monetary quantities

Some utility measures

Weight

Lengths

Energy consumption

Pandas data type

Floating point

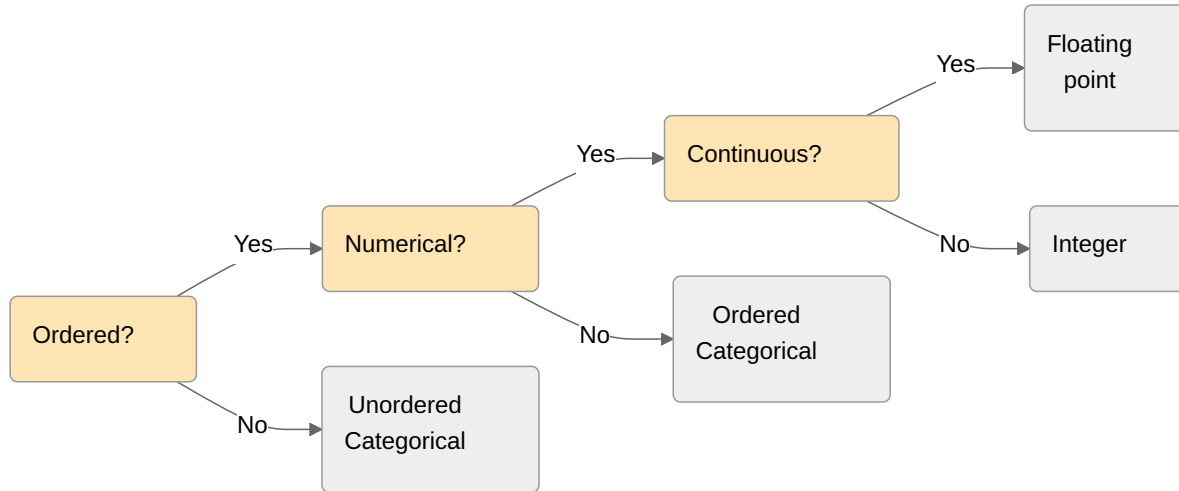
```
pd.Series(  
    [  
        42_345,  
        53_724,  
        28_734,  
    ],  
    dtype=float  
)
```

Cardinal, discrete data

Example:

- Variable: Age in years
- Possible values: 0, 1, 2, ...
- Observed values: 18, 22, 19

Decision tree



Cardinal, discrete data

Examples

Age in years or months

Any count variable

Depending on the typical number of outcomes, we may treat variables as quasi-continuous.

Pandas data type

Integer

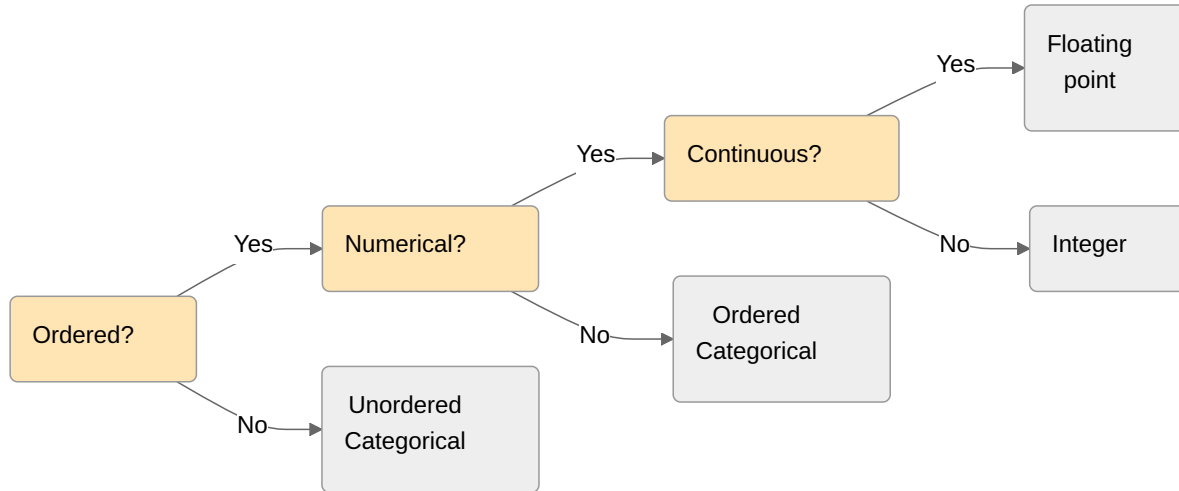
```
pd.Series(  
    [  
        18,  
        22,  
        19,  
    ],  
    dtype=int  
)
```

Ordinal, non-numeric data

Example:

- Variable: Annual Income in Euros, binned
- Possible values: $[0, 30000)$, $[30000, 60000)$, $[60000, \infty)$
- Observed values: $[30000, 60000)$, $[30000, 60000)$, $[0, 30000)$,

Decision tree



Ordinal, non-numeric data

Examples

Binned cardinal data, whatever the label.
Here:

- "low", "medium", "high"
- 0, 1, 2

Labelled Likert scales ("0 disagree", "1",
"2", "3", "4 agree")

Highest degree obtained

Social status / class

Pandas data type

Ordered categorical

```
pd.Series(
    [
        "[30000, 60000)",
        "[30000, 60000)",
        "[0, 30000)",
    ],
    dtype=pd.CategoricalDtype(
        [
            "[0, 30000)",
            "[30000, 60000)",
            "[60000, ∞)",
        ],
        ordered=True
    )
)
```

Ordinal, numeric data

Examples

Unlabelled Likert scales

IQ

Some utility measures

Pandas data type

Ordered categorical

Integer

Floating point