Cardinality assessment

GROKKERS
AI FOR EVERYONE

Pre-reqs

- Python
- NumPy and PANDAS, SciPy, Visualizations
- Elementary stats and maths
- Some preprocessing steps – may need ML as well, for advanced topics

Background

Numeric Data: Preprocessing involves handling missing values, scaling to a similar range, and possibly normalizing the distribution.

Text Data**:** Common preprocessing steps include text cleaning (removing stop words, punctuation, etc.), tokenization, and vectorization (converting text into numerical form, such as TF-IDF or word embeddings).

Image Data: Techniques like resizing, normalization of pixel values, and data augmentation (creating variations of existing images) are often used.

Time Series Data: Dealing with temporal aspects, handling missing values over time, and creating lag features are important steps in preprocessing time series data.

Topics

About Data, feature types, tabular form

General inspection of data quality

Handling duplicates in data

Missing value analysis

(2 parts)

Handling Outliers

Cardinality assessment

Encoding of discrete data

Scaling and Normalization

Handling Skewed Distributions

Data Imbalance Handling

Data Splitting

Cardinality assessment

01

refers to the count or number of distinct/ unique values present in a dataset, column, or set.

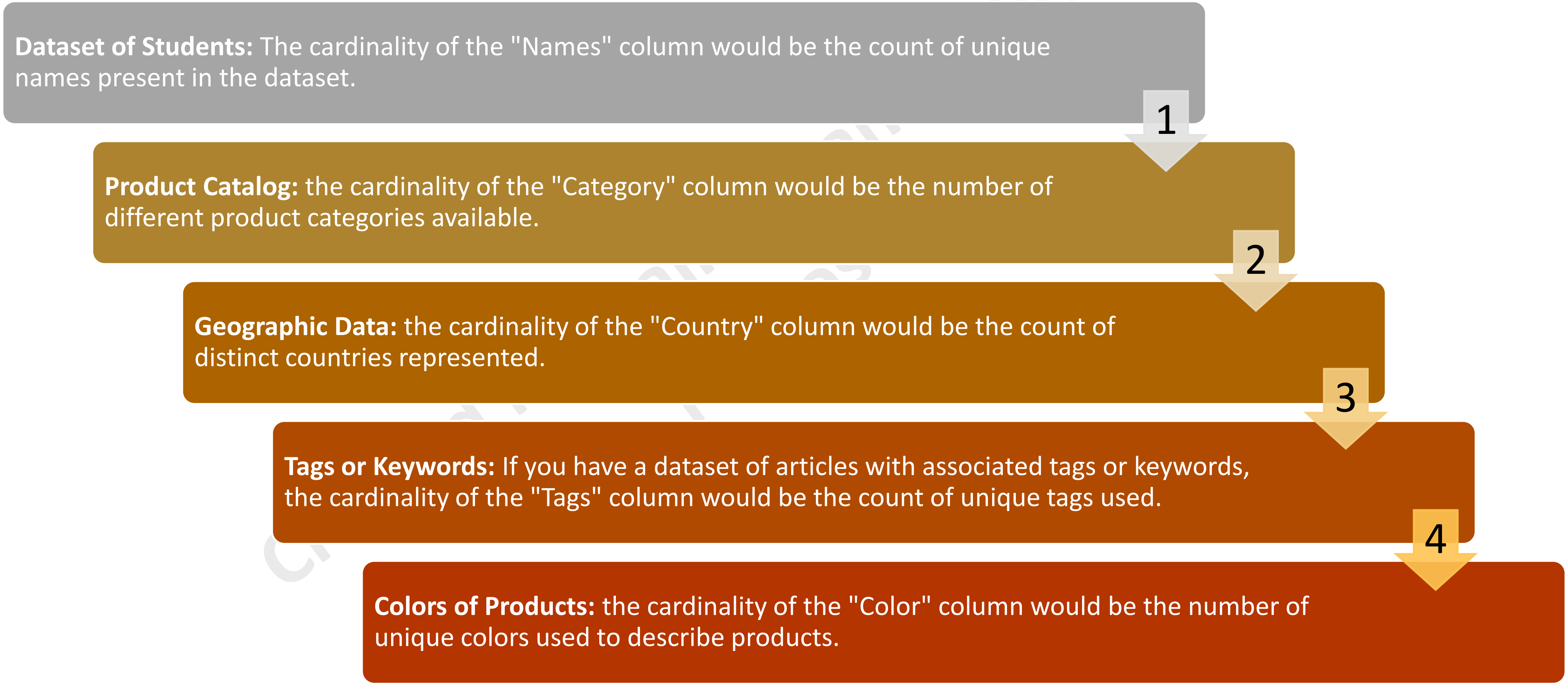
02

Useful concept in understanding the diversity and variation within data.

03

particularly used in the context of categorical or discrete data.

Examples of cardinality



Dataset of Students: The cardinality of the "Names" column would be the count of unique names present in the dataset.

Product Catalog: the cardinality of the "Category" column would be the number of different product categories available.

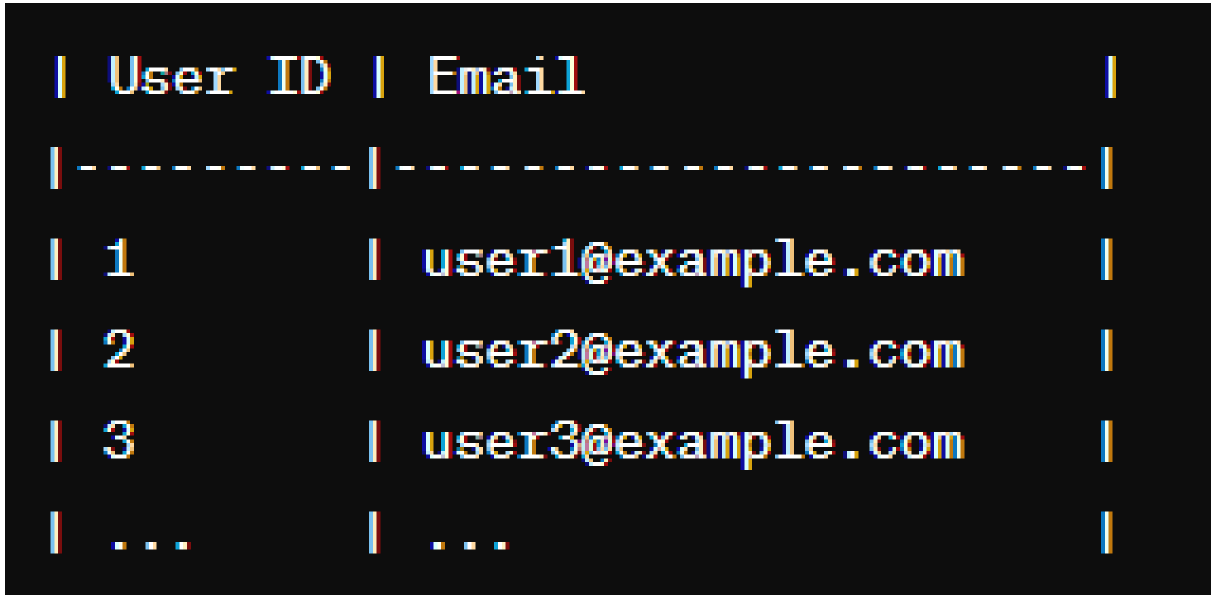
Geographic Data: the cardinality of the "Country" column would be the count of distinct countries represented.

Tags or Keywords: If you have a dataset of articles with associated tags or keywords, the cardinality of the "Tags" column would be the count of unique tags used.

Colors of Products: the cardinality of the "Color" column would be the number of unique colors used to describe products.

High Cardinality

- Consider a dataset with a column representing email addresses.
- Each email address is unique, leading to high cardinality.



```
| User ID | Email                  |
|---------|------------------------|
| 1       | user1@example.com      |
| 2       | user2@example.com      |
| 3       | user3@example.com      |
| ...     | ...                    |
```

CustomerID	ProductCategory	Country	PaymentMethod
1	Electronics	USA	Credit Card
2	Clothing	Canada	PayPal
3	Electronics	Germany	Credit Card
4	Home & Garden	USA	PayPal
5	Electronics	France	Credit Card

Column	Unique Values Count
CustomerID	5
ProductCategory	3
Country	4
PaymentMethod	2

To determine high cardinality

Set Threshold:

- Define a threshold based on the dataset and problem.
- Let's say we consider a column to have high cardinality if it has more than 3 unique values.

Determine High Cardinality Columns:

- Based on the threshold, identify columns exceeding the defined limit.
- High Cardinality Columns:
  - CustomerID (Considered high based on a low threshold for this example)
  - ProductCategory (No)
  - Country (No)
  - PaymentMethod (No)

Example 2

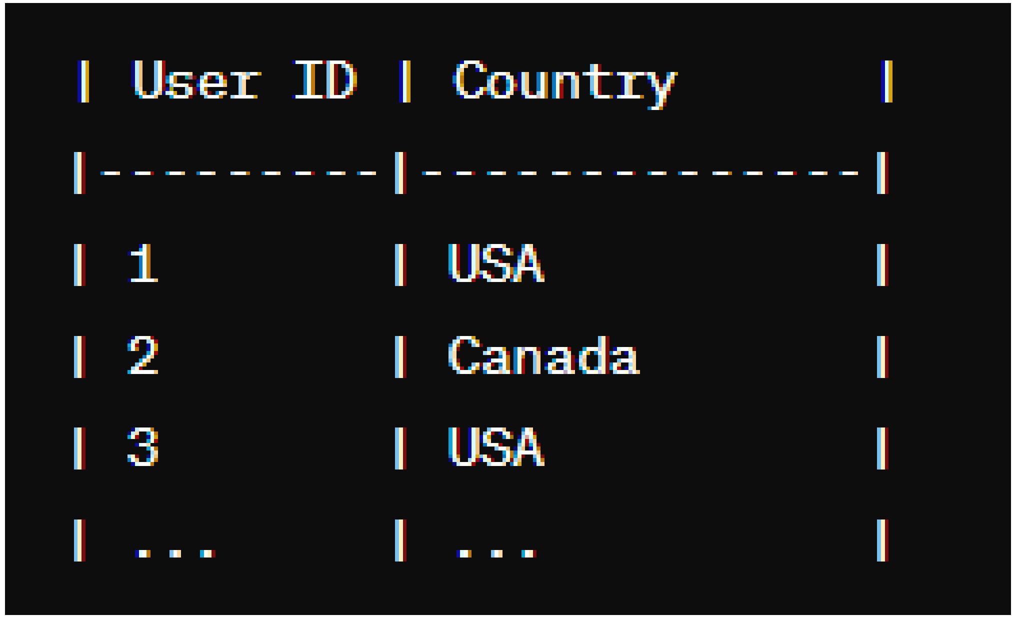
- Count Unique Values
  - Examine the "ProductCategory" column in your dataset.
  - Count the number of unique values in this column.
- Calculate Percentage
  - Calculate the percentage of unique values compared to the total number of rows in your dataset.
  - For example, if you have 100 rows and 3 unique product categories, the percentage would be 3%.
- **Set Threshold:**
  - Define a threshold percentage based on your dataset and problem. For instance, set a threshold at 90%.

CustomerID	ProductCategory	PurchaseAmount
1	Electronics	500
2	Clothing	120
3	Electronics	300
4	Home & Garden	200
5	Electronics	700

Low Cardinality

Now, consider a dataset with a column representing countries.

The cardinality is lower as there are fewer unique values.



```
| User ID | Country        |
|---------|----------------|
| 1       | USA            |
| 2       | Canada         |
| 3       | USA            |
| ...     | ...            |
```

Effects of cardinality

Aspect	High Cardinality	Low Cardinality
Overfitting Risk	Higher risk of overfitting in machine learning	Lower risk of overfitting
Model Complexity	Models may become more complex with many categories	Simpler models, easier interpretation
Visualization Challenges	Visualization may be challenging with many categories	Easier visualization due to fewer distinct values
Data Exploration	Requires more in-depth exploration and analysis	Simpler exploration, easier identification of trends
Data Preprocessing	Demands careful preprocessing, encoding strategies	Straightforward preprocessing, less encoding needed
Statistical Analysis	May introduce bias in statistical analyses	Less likely to bias statistical analyses
Grouping and Aggregation	Grouping and aggregating may be challenging	Grouping and aggregating are typically straightforward

Thanks !!

AI FOR EVERYONE

Grokkers

CONCEPTS AND EQUATIONS MADE SIMPLE