Cheatsheets / **Principles of Data Literacy**

Introduction to Data

Garbage In, Garbage Out

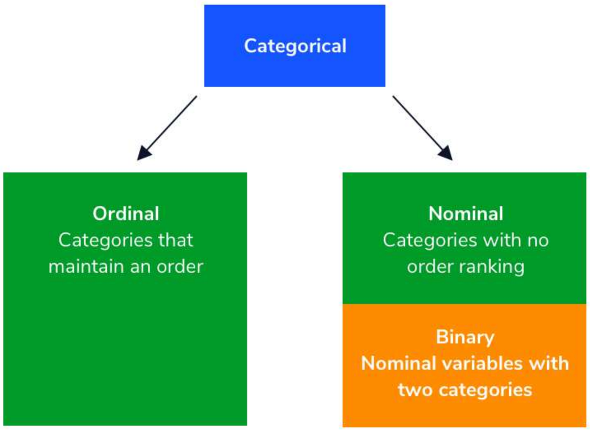
The quality of the predictions made during a predictive analysis is deeply dependent on the quality of the data used to generate the predictions.
For example, if a model is trained with mislabeled data, it will produce inaccurate predictions no matter how good the actual algorithm is. This is commonly referred to as, “garbage in, garbage out.”

Binary Categorical Variables

Categorical variables can also be binary or dichotomous variables. Binary variables are nominal categorical variables that contain only two, mutually exclusive categories. Examples of binary variables are if a person is pregnant, or if a house’s price is above or below a particular price.

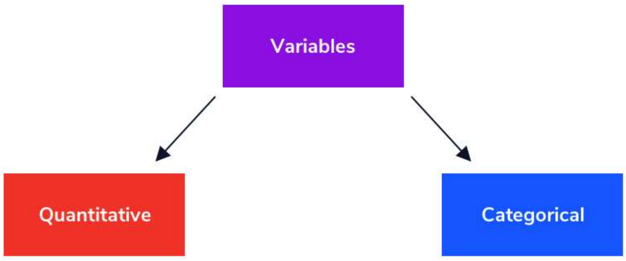
Categorical Variables

Categorical variables consist of data that can be grouped into distinct categories, and are ordinal or nominal. Ordinal categorical variables which are groups that contain an inherent ranking, such as ratings of plays or responses to a survey question with a point scale e.g., on a scale from 1-7, how happy are you right now? Nominal categorical variables are made of categories without an inherent order, examples of nominal variables are species of ants, or people’s hair color.


Categorical
Ordinal
Categories that maintain an order
Nominal
Categories with no order ranking
Binary
Nominal variables with two categories

Quantitative Vs. Categorical Variables

Variables can be either quantitative or categorical. Quantitative variables are amounts or counts; for example, age, number of children, and income are all quantitative variables. Categorical variables represent groupings; for example, type of pet, agreement rating, and brand of shoes are all categorical variables.



Categorical Data Defined

Categorical Data refers to data represented by words rather than numbers. Examples of categorical data are tree species and survey responses (Agree, Neutral, Disagree).

Ordinal and Nominal Categorical Data

Categorical variables can be either ordinal (ordered) or nominal (unordered).
Examples of ordinal variables include places (1st, 2nd, 3rd) and survey responses (on a scale of 1 to 5, how much do you agree with a statement).
Examples of nominal variables include tree species, student names, and account names.

Messy Data

Messy data is data that violates one of the tidy dataset rules (1. Each variable forms a column; 2. Each observation forms a row; 3. Each type of observational unit forms a table).
Below is an example of messy data:

ID#	Name	ChemGrade2020	MathGrade2020
1	Brown	F	
B	smith		
3	Saito, K	A	90

Tabular Data

Tabular data is organized into rows, or observations, and columns, also referred to as variables or features. We can read each column "down" the table (viewing multiple observations), and each row "across" the table (viewing multiple variables).

Row #	Variable 1	Variable 2	Variable 3
1	Observation	Observation	Observation
2	Observation	Observation	Observation
3	Observation	Observation	Observation

Tidy Data Rules

A tidy dataset follows three fundamental rules:

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table.

Below is an example of a tidy dataset:

ID#	Student	Year	Class	Grade
1	Brown	2020	Chem	F
1	Brown	2021	Chem	B
1	Brown	2021	Math	A
2	Smith	2020	Bio	C
2	Smith	2021	CompSci	B
3	Saito	2020	Chem	A
3	Saito	2021	Math	B

Sample Set of Data

A sample set of data is a dataset that is representative of the entire population of interest. Random sampling is the best way to make sure the sample is representative of the whole population but does not guarantee a representative sample, especially if the sample is too small.

Structurally Missing Data

Structurally Missing Data is data that is expected to be missing.
For example, there are structurally missing data in the 'Litters' and 'Pups/Litter' columns for all the male dogs in the table below because we would not expect male dogs to have puppies.

ID#	Name	Breed	Sex	Litters	Pups/
1	Gnasher	ACD	M		
2	Cassie	Collie	F	1	3
3	Pepper	French Bulldog	F	4	2
4	Jed	Golden Retreiver	M		
5	Henry	Spaniel	M		
6	Ruby	ACD	F	1	6

Missing at Random Data

Missing at Random (MAR) data is missing because of some random characteristic about the person or thing being studied. Often, this type of data is reliably missing based on the value of another variable in the dataset. In the table below, the bacterial cell counts for all the stool samples are 'NaN'. If we looked into this, we might find that there were too many bacterial cells to count in all those samples. Therefore, the bacterial cell counts for stool samples would be MAR data.

Sample ID	**Sample Type**	**Bacterial Cell Counts**
1	Hand Swab	1008
2	Stool	NaN
3	Mouth Swab	7876
4	Hand Swab	657
5	Stool	NaN
6	Hand Swab	2442
7	Mouth Swab	5444
8	Stool	NaN
9	Hand Swab	4654
10	Stool	NaN

Data Missing Completely at Random

Missing Completely at Random (MCAR) data has no detectable underlying reason causing the values to be missing.
The table below has MCAR data. The # of fruits is missing for some plants, but the missing fruit data seems unrelated to the height of the plant. Short and tall plants are both missing fruit data. In addition, we are missing the height for one of our plants!

Plant	Height (cm)	# of Fruits
1	65	10
2		87
3	987	
4	44	
5	105	35
6	547	74
7	876	
8	55	
9	875	95

Data Gaps

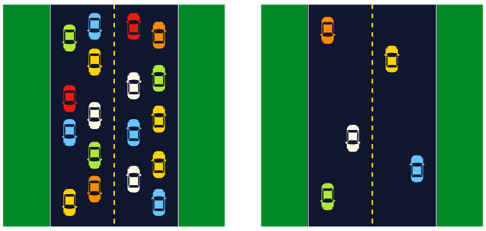
The ability to separate good, mediocre, and poor quality data is a crucial data literacy skill. Data-driven conclusions are only as strong, robust, and well-supported as the data behind them. This is also often referred to with the phrase “garbage in, garbage out.”

Addressing Bias

Bias in data collection leads to poorer quality data. Recognizing bias in data is a crucial data literacy skill. Some key questions about bias include “Who made the data?”, “Who participated in the data?” and “Who is left out of the data?”

What is Statistics?

Statistics helps to measure whether an event happens by chance or by a systemic factor or factors. For example, it's statistically more likely to see traffic during peak rush hour than outside of peak rush hour times.



Statistics at work

Statistics can reveal systemic patterns in a data set rather than relying on individual experiences. This is important in legal cases including those addressing discrimination or class-action lawsuits.

Print Share ▼