Data Science for Everyone – Data

Dr. Ab Mosca (they/them)

Slides based off slides courtesy of Jordan Crouser (https://jcrouser.github.io/)

Plan for Today

- Data Vocabulary
- Data Ethics

Note

- hw01 released today. Find it on the course website: https://amosca01.github.io/Intro-DS-F23/

Data Vocabulary

Definition: Data

What is data? How would you define it?

Definition: Data

Definition of *data*

1 : factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation

// the *data* is plentiful and easily available
— H. A. Gleason, Jr.

// comprehensive *data* on economic growth have been published
— N. H. Jacoby

2 : information in digital form that can be transmitted or processed

3 : information output by a sensing device or organ that includes both useful and irrelevant or redundant information and must be processed to be meaningful

https://www.merriam-webster.com/dictionary/data

Example of Data

- Suppose you really like shoes
- What information can you collect about a single shoe?
- Brainstorm with whoever is next to you

Example of Data

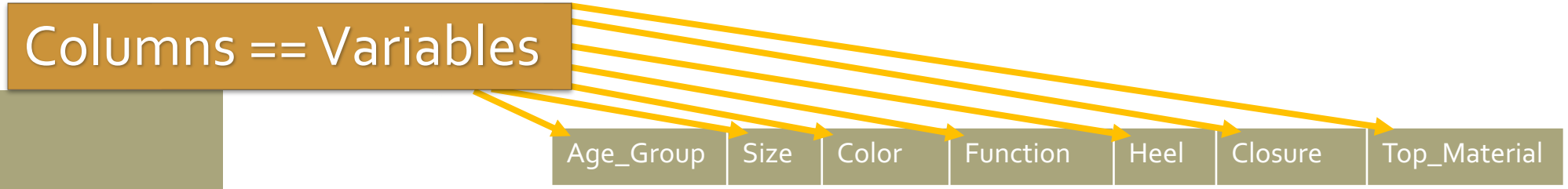
- Age Group
- Size
- Color
- Function
- Heel
- Closure
- Top Material
- …

Example of Data

- Age Group
- Size
- Color
- Function
- Heel
- Closure
- Top Material
- …

Variables

Example of Dataset



Example of Dataset

Columns == Variables

Age_Group	Size	Color	Function	Heel	Closure	Top_Material

Row == Observation

Example of Dataset

Columns == Variables

Age_Group	Size	Color	Function	Heel	Closure	Top_Material
Adult	7	Black	Climb	N	Velcro	Fabric

Row == Observation

Cells == Values

Example of Dataset

Columns == Variables

Cells == Values

Rows == Observations

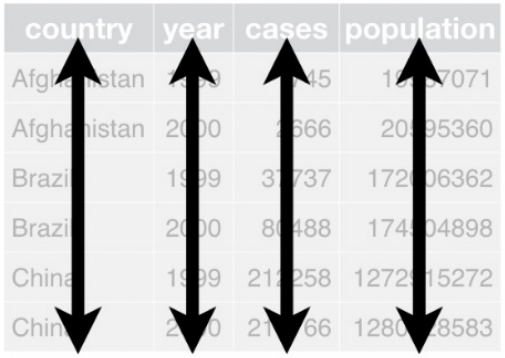
Age_Group	Size	Color	Function	Heel	Closure	Top_Material
A			Climb	N	Velcro	Fabric
A			Dressy	Y	None	Leather
Adult	8	Yellow	Dressy	Y	None	Leather
Adult	8	Blue	Dressy	Y	None	Leather
Adult	6	Pink	Ballet	N	Ribbon	Fabric
Adult	10	Grey	Everyday	N	Laces	Fabric
Baby	1	White	Everyday	N	Laces	Fabric
Adult	7	Red	Everyday	N	Laces	Fabric
Adult	12	Brown	Dressy	Y	Laces	Leather
Adult	7.5	Purple	Exercise	N	Laces	Fabric
Baby	3	Purple	Rain	N	None	Plastic
Child	4	Green	Rain	N	None	Plastic
Child	4	Grey	Rain	N	None	Plastic
Child	6	Yellow	Rain	N	None	Plastic
Adult	7	Red	Rain	N	None	Plastic

Example of Dataset

Table

Age_Group	Size	Color	Function	Heel	Closure	Top_Material
Adult	7	Black	Climb	N	Velcro	Fabric
Adult	8	Pink	Dressy	Y	None	Leather
Adult	8	Yellow	Dressy	Y	None	Leather
Adult	8	Blue	Dressy	Y	None	Leather
Adult	6	Pink	Ballet	N	Ribbon	Fabric
Adult	10	Grey	Everyday	N	Laces	Fabric
Baby	1	White	Everyday	N	Laces	Fabric
Adult	7	Red	Everyday	N	Laces	Fabric
Adult	12	Brown	Dressy	Y	Laces	Leather
Adult	7.5	Purple	Exercise	N	Laces	Fabric
Baby	3	Purple	Rain	N	None	Plastic
Child	4	Green	Rain	N	None	Plastic
Child	4	Grey	Rain	N	None	Plastic
Child	6	Yellow	Rain	N	None	Plastic
Adult	7	Red	Rain	N	None	Plastic

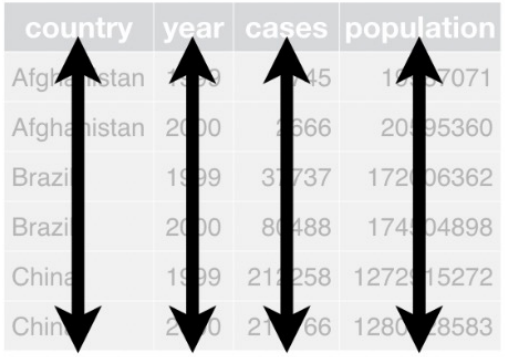
Table Vocabulary



columns

rows

cells



variables

observations

values

Variable (or data) Types

Variables

Age_Group	Size	Color	Function	Heel	Closure	Top_Material
Adult	7	Black	Climb	N	Velcro	Fabric
Adult	8	Pink	Dressy	Y	None	Leather
Adult	8	Yellow	Dressy	Y	None	Leather
Adult	8	Blue	Dressy	Y	None	Leather
Adult	6	Pink	Ballet	N	Ribbon	Fabric
Adult	10	Grey	Everyday	N	Laces	Fabric
Baby	1	White	Everyday	N	Laces	Fabric
Adult	7	Red	Everyday	N	Laces	Fabric
Adult	12	Brown	Dressy	Y	Laces	Leather
Adult	7.5	Purple	Exercise	N	Laces	Fabric
Baby	3	Purple	Rain	N	None	Plastic
Child	4	Green	Rain	N	None	Plastic
Child	4	Grey	Rain	N	None	Plastic
Child	6	Yellow	Rain	N	None	Plastic
Adult	7	Red	Rain	N	None	Plastic

Variable (or data) Types

Quantitative ⟷

- Numeric
- Ordered

Age_Group	Size	Color	Function	Heel	Closure	Top_Material
Adult	7	Black	Climb	N	Velcro	Fabric
Adult	8	Pink	Dressy	Y	None	Leather
Adult	8	Yellow	Dressy	Y	None	Leather
Adult	8	Blue	Dressy	Y	None	Leather
Adult	6	Pink	Ballet	N	Ribbon	Fabric
Adult	10	Grey	Everyday	N	Laces	Fabric
Baby	1	White	Everyday	N	Laces	Fabric
Adult	7	Red	Everyday	N	Laces	Fabric
Adult	12	Brown	Dressy	Y	Laces	Leather
Adult	7.5	Purple	Exercise	N	Laces	Fabric
Baby	3	Purple	Rain	N	None	Plastic
Child	4	Green	Rain	N	None	Plastic
Child	4	Grey	Rain	N	None	Plastic
Child	6	Yellow	Rain	N	None	Plastic
Adult	7	Red	Rain	N	None	Plastic

Variable (or data) Types

Quantitative

- Numeric
- Ordered

Size
7
8
8
8
6
10
1
7
12
7.5
3
4
4
6
7

Sequential Ordering

0

Variable (or data) Types

Quantitative

- Numeric
- Ordered

Rainfall (in.)
31.00
32.90
30.66
30.58
31.25
30.36
27.73
31.40
29.23
26.80

Sequential Ordering

0

https://www.bwsc.org/environment-education/rainfall-garden

Variable (or data) Types

Quantitative

- Numeric
- Ordered

Temp (C)
-2
10
-4
20
16
22
-10
-8
0
14

Divergent Ordering

0

https://www.bwsc.org/environment-education/rainfall-garden

Variable (or data) Types

Quantitative

- Numeric
- Ordered

Elevation (m)
-413
100
-40
200
1600
-555
1000
0
40
678

Divergent Ordering

0

https://www.bwsc.org/environment-education/rainfall-garden

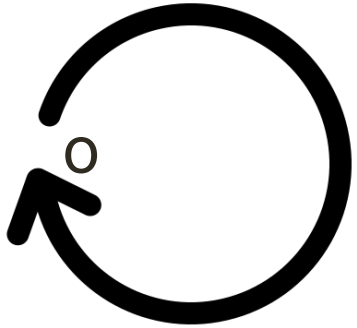
Variable (or data) Types

Quantitative

- Numeric
- Ordered

stop-time
14:00
13:00
04:34
02:32
15:16
01:00
20:45
10:23
11:32
12:00

Cyclical Ordering



Created by Kero
from Noun Project

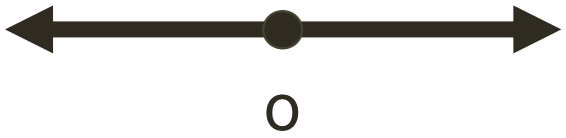
Variable (or data) Types

- Work with whoever is next to you to come up with an example for each quantitative ordering

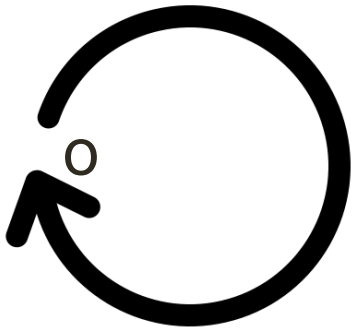
Quantitative ⟷

Sequential Ordering

Divergent Ordering



Cyclical Ordering



Variable (or data) Types

Categorical 

- Groups
- Levels

Age_Group	Size	Color	Function	Heel	Closure	Top_Material
Adult	7	Black	Climb	N	Velcro	Fabric
Adult	8	Pink	Dressy	Y	None	Leather
Adult	8	Yellow	Dressy	Y	None	Leather
Adult	8	Blue	Dressy	Y	None	Leather
Adult	6	Pink	Ballet	N	Ribbon	Fabric
Adult	10	Grey	Everyday	N	Laces	Fabric
Baby	1	White	Everyday	N	Laces	Fabric
Adult	7	Red	Everyday	N	Laces	Fabric
Adult	12	Brown	Dressy	Y	Laces	Leather
Adult	7.5	Purple	Exercise	N	Laces	Fabric
Baby	3	Purple	Rain	N	None	Plastic
Child	4	Green	Rain	N	None	Plastic
Child	4	Grey	Rain	N	None	Plastic
Child	6	Yellow	Rain	N	None	Plastic
Adult	7	Red	Rain	N	None	Plastic

Variable (or data) Types

Categorical

- Groups
- Levels

Color
Black
Pink
Yellow
Blue
Pink
Grey
White
Red
Brown
Purple
Purple
Green
Grey
Yellow
Red

Color Levels:

- Black
- Pink
- Yellow
- Blue
- Grey
- White
- Red
- Brown
- Purple
- Green

Variable (or data) Types

Categorical

- Groups
- Levels

Function
Climb
Dressy
Dressy
Dressy
Ballet
Everyday
Everyday
Everyday
Dressy
Exercise
Rain
Rain
Rain
Rain
Rain

Function Levels:

- Climb
- Dressy
- Ballet
- Everyday
- Exercise
- Rain

Variable (or data) Types

Categorical

- Groups
- Levels

Heel
N
Y
Y
Y
N
N
N
N
Y
N
N
N
N
N
N

Heel Levels:

- N
- Y

Note: A categorical variable with exactly two level is also called a Binary Variable.

- N == 0 == False
- Y == 1 == True

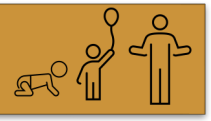
Variable (or data) Types

- Work with whoever is next to you to come up with two examples of categorical variables. Note the levels in your examples

Categorical 

Binary 1010

Variable (or data) Types

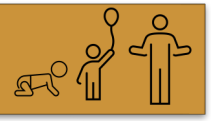
Ordinal 

- Groups
- Levels
- Ordered

Age_Group	Size	Color	Function	Heel	Closure	Top_Material
Adult	7	Black	Climb	N	Velcro	Fabric
Adult	8	Pink	Dressy	Y	None	Leather
Adult	8	Yellow	Dressy	Y	None	Leather
Adult	8	Blue	Dressy	Y	None	Leather
Adult	6	Pink	Ballet	N	Ribbon	Fabric
Adult	10	Grey	Everyday	N	Laces	Fabric
Baby	1	White	Everyday	N	Laces	Fabric
Adult	7	Red	Everyday	N	Laces	Fabric
Adult	12	Brown	Dressy	Y	Laces	Leather
Adult	7.5	Purple	Exercise	N	Laces	Fabric
Baby	3	Purple	Rain	N	None	Plastic
Child	4	Green	Rain	N	None	Plastic
Child	4	Grey	Rain	N	None	Plastic
Child	6	Yellow	Rain	N	None	Plastic
Adult	7	Red	Rain	N	None	Plastic

Variable (or data) Types

Ordinal



- Groups
- Levels
- Ordered

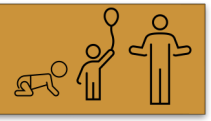
Age_Group
Adult
Adult
Adult
Adult
Adult
Adult
Baby
Adult
Adult
Adult
Baby
Child
Child
Child
Adult

Sequential Ordering

Baby → Child → Adult

Variable (or data) Types

Ordinal



- Groups
- Levels
- Ordered

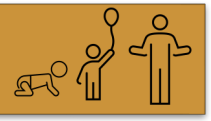
Size
Sm
Lg
Sm
Med
Med
Lg
Sm
Lg
Lg
Med
Sm

Sequential Ordering

Sm → Med → Lg

Variable (or data) Types

Ordinal



- Groups
- Levels
- Ordered

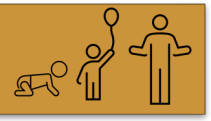
Attitude
Neutral
Dislike
Like
Neutral
Like
Like
Like
Neutral
Like
Dislike
Neutral

Divergent Ordering

Dislike ← Neutral → Like

Variable (or data) Types

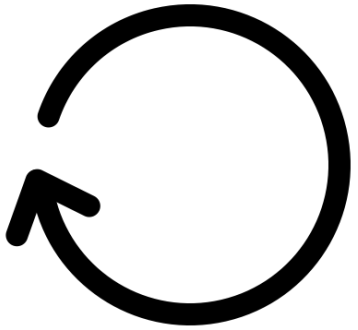
Ordinal



- Groups
- Levels
- Ordered

Month
Jan
Feb
Dec
Apr
Mar
Jan
Feb
Oct
Nov
Apr
May

Cyclical Ordering



Variable (or data) Types

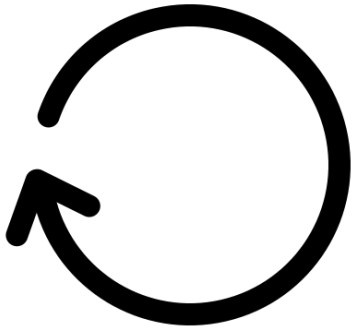
- Work with whoever is next to you to come up with an example for each ordinal ordering

Ordinal 

Sequential Ordering

Divergent Ordering

Cyclical Ordering



Data Ethics

Responsible Consumption

- What details should we look for when evaluating a data set?
- Brainstorm with the person next to you and come up with at least 3 things

Responsible Consumption

- What details should we look for when evaluating a dataset?
  - Who published the data?
  - Who collected the data?
  - Who funded collection of the data?
  - Who (or what) is included in the dataset?
  - Who (or what) is missing from the dataset?
  - Was the data transparently collected?
  - Was the data legally collected?
  - Are there any privacy issues?