Principles of data exploration

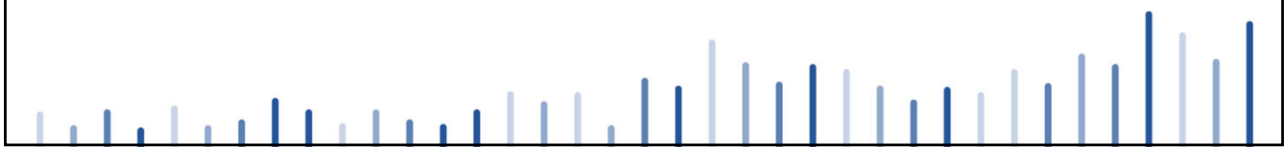
Moses Chan

April 2, 2024
DATA_ENG 200

Adapted from Health and Healthcare Data Visualization Module from Tableau

1

What are Data?



2

Definition of Data

"Data are values of qualitative or quantitative variables, belonging to a set of items."

- **Set of items**: Sometimes called the population; the set of objects you are interested in
- **Variable**: A measurement or characteristic of an item.
  - **Qualitative**: Country of origin, sex, treatment
  - **Quantitative**: Height, weight, blood pressure reading

Adapted from Jeff Leek, Johns Hopkins School of Public Health

3

How Data Are Collected

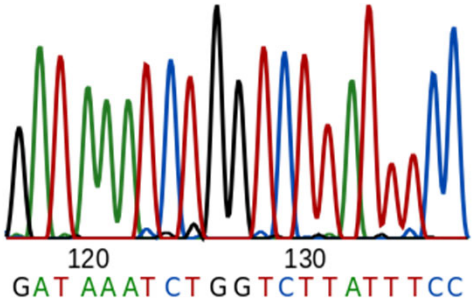
Data can be collected in a variety of ways:

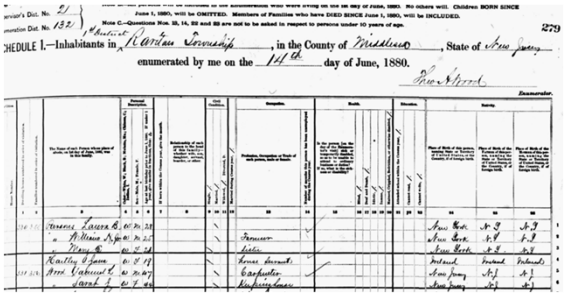
- Questionnaire
- Interview
- Observation
- Analysis of Documents
- Web scraping
- Machine measurements

4

Raw data: The form that we receive or collect the data

- The file your measurement machine gives you
- Unformatted Excel file with 10 worksheets that the company you contracted with sent you
- Complicated data file from scraping Twitter
- Hand-entered numbers you collected from observations





5

How are Data Organized?

8

How do we organize data?

- Rows (across)
  - Each row represents on unit of analysis
- Columns (down)
  - Each column represents a different variable (or field)

Sort fields Data source order Show aliases Show hidden fields 584 rows

Agency	Agency Name	Borough	City	Closed Date	Complaint Type	Created Date	Descriptor	Due Date
DOHMH	Department of Healt...	BROOKLYN	BROOKLYN	4/14/2016 12:00:00 ...	Rodent	4/1/2016 12:00:00 AM	Rat Sighting	2016-05-01T01:37
DOHMH	Department of Healt...	BRONX	BRONX	1/11/2016 12:00:00 ...	Rodent	4/1/2016 12:00:00 AM	Rat Sighting	2016-05-01T00:20
DOHMH	Department of Healt...	MANHATTAN	NEW YORK	4/21/2016 12:00:00 ...	Rodent	4/1/2016 12:00:00 AM	Rat Sighting	2016-05-01T00:42
DOHMH	Department of Healt...	BROOKLYN	BROOKLYN	4/4/2016 2:13:57 PM	Unsanitary Animal P...	4/1/2016 12:00:00 AM	Cat	2016-04-30T23:50
DOHMH	Department of Healt...	BROOKLYN	BROOKLYN	4/26/2016 12:00:00 ...	Unsanitary Animal P...	4/1/2016 12:00:00 AM	Other Animal	2016-05-01T01:05
DOHMH	Department of Healt...	BRONX	BRONX	4/4/2016 10:15:53 AM	Rodent	4/1/2016 12:00:00 AM	Rat Sighting	2016-05-01T00:17
DOHMH	Department of Healt...	BROOKLYN	BROOKLYN	4/14/2016 12:00:00 ...	Rodent	4/1/2016 12:00:00 AM	Rat Sighting	2016-05-01T01:36

9

Tidy Data

- Each variable should be in one column
- Each different observation of that variable should be in a different row
- *If you have multiple tables, they should include a column in the table that allows them to be joined*

10

Example: individual observations are in rows and columns

Column Oriented

ID	Gender	School	Math	Science	History
1	M	West	90	80	80
2	F	South	50	50	50
3	M	Central	50	80	80

Row Oriented

ID	Gender		School	Subject	Score
1	M		West	Math	90
1	M		West	Science	80
1	M		West	History	80
2	F		South	Math	50
2	F		South	Science	50
2	F		South	History	50
3	M		Central	Math	50
3	M		Central	Science	80
3	M		Central	History	80

11

Types of Data

14

Types of Data

Qualitative/Categorical

- Non-numerical data that may be observed but not measured or have mathematical functions performed on them, such as:
  - Patient's eye color, sex, perceptions about health status
  - Organizational change, physicians' implementation of evidence-based guidelines

Visualizing Health and Healthcare Data, page 14

Quantitative/Numerical

- Measure the quantity or amount of something, and are numerical, can have mathematical functions performed on them. Two categories:
  - Continuous—infinite number of possible values w/in a range, such as: patient's age, height, weight, pulse, respiratory rate
  - Discrete—can be counted, have a finite number of possible values, such as: prescriptions filled, hospital admission, hours of exercise per week

tableau

15

Discrete and Continuous

Discrete

- Individually separate and distinct
- Example: a household could have **3** or **6 children**, but **not 4.72**!

Continuous

- Forming an unbroken whole, without interruption
- Example: response time in seconds. We could record **1.64 seconds** or **1.642378765 seconds**

16

Examples of data types

Quantitative	Categorical/Ordinal	Categorical/Nominal
Weight (10 kg. 35 kg, 100 kg)	Gold Silver Bronze	North America Europe Asia
Age (years, months)	Excellent Good Poor	Alice Bob Chris
Medication Dose (mg, ml)	January February March	Wine Beer Water

17

Nominal (without order)

- Qualitative/Categorical "in name only"
- No underlying prescribed order
- No measure of distance between values

Examples:

- Race and ethnicity
- Language
- Country, state
- Department, clinical service
- Educational degree
- Blood type
- Insurance type

18

Ordinal (with order)

- Qualitative/categorical
- Characterized by having some underlying, meaningful order or sequence

Examples:

- Academic Achievement—High School, College, Graduate School
- The American Society of Anesthesiologists Physical Classifications—1,2,3,4,5,6
- The Consumer Assessment of Healthcare Providers (CAHPS)—Never, Sometimes, Usually, Always

19

Why do we need categorization of data types?

Different methods are suitable for specific analyses of data.

20

Fields and Variables in Tableau

21

Data Type Icons

Icon	Value Description
Abc	Text (string) values
	Date values
	Date & time values
#	Numerical values
T|F	Boolean values (relational only)
	Geographic values (used with maps)

Tableau for Healthcare, page 11

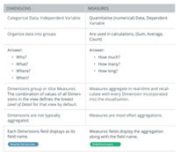
22

Tableau Assigns Fields

Tableau for Healthcare, page 13

23

Dimensions and Measures



DIMENSIONS	MEASURES
Categorical Data; Independent Variable	Quantitative (numerical) Data, Dependent Variable
Organize data into groups	Are used in calculations, (Sum, Average, Count)
Answer: • Who? • What? • Where? • When?	Answer: • How much? • How many? • How long?
Dimensions group or slice Measures. The combination of values of all Dimensions in the view defines the lowest level of detail for that view by default.	Measures aggregate in real-time and recalculate with every Dimension incorporated into the visualization.
Dimensions are not typically aggregated.	Measures are most often aggregations.
Each Dimensions field displays as its field name.	Measures fields display the aggregation along with the field name.

Tableau for Healthcare, page 13

24

Discrete (Blue) and Continuous (Green) Fields

Tables

- Abc Category
- Date
- Abc Type of Service/Source of Funds
- Abc *Measure Names*
- # Expenditures
- # *US HC Expenditures 1980-2018 (Count)*
- # *Measure Values*

YEAR(Date)	Discrete
YEAR(Date)	Continuous
SUM(Expenditures)	Continuous
SUM(Expenditures)	Discrete

Tableau for Healthcare, page 15

25

Headers and Axes

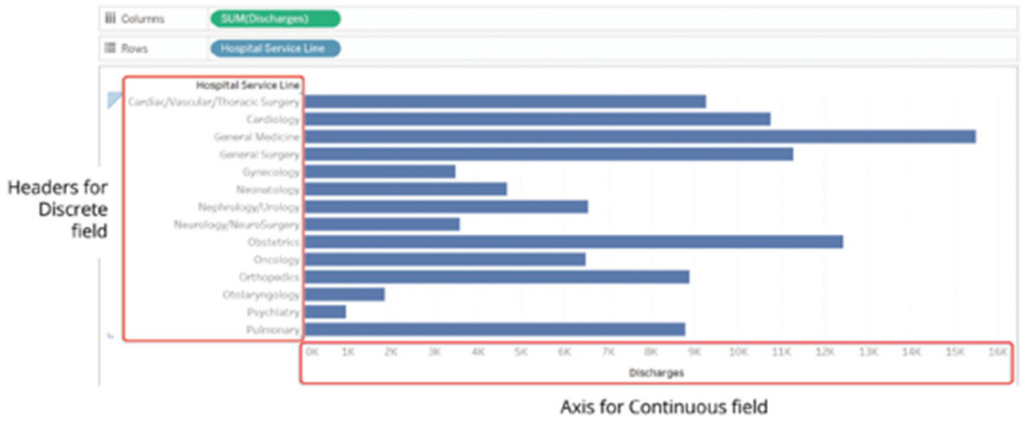


Tableau for Healthcare, page 16

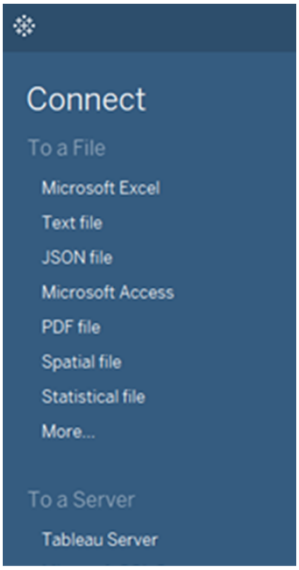
26

Connecting to Data

33

Connect to Data Screen

- Tableau can connect to many file types
  - Excel
  - Delimited text files (*.txt, *.csv, *.tab, *.tsv)
  - PDF
  - Many more



34

Live versus Extract

Connecting live leaves the data in the database or source file.

- Sometimes connecting live can result in a slow experience, depending on the database.

Extracting data makes a copy of the dataset.

- Helps when connecting to a slow database or to take query load off critical systems.
- Can import only some of the data and bring in specific elements (to access those options, click Edit)

36

Demo and exercises

37