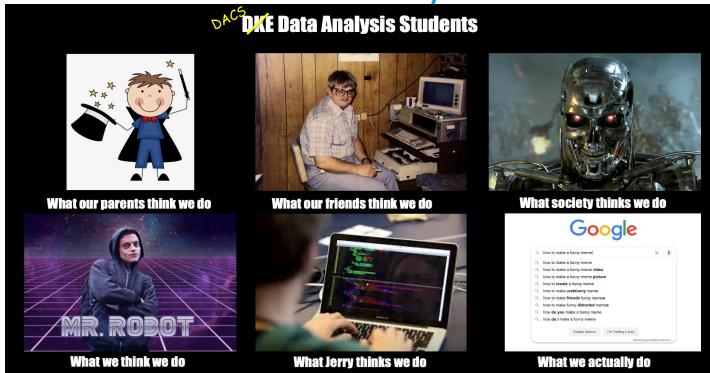
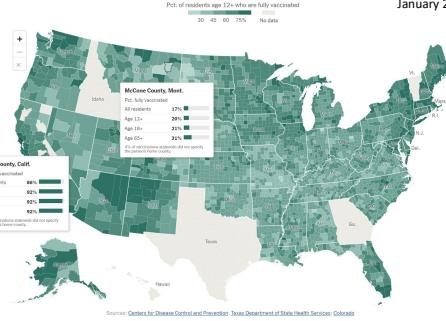


Welcome to Data Analysis v9.0!



Why?

NYTimes
Visualization: January 2022



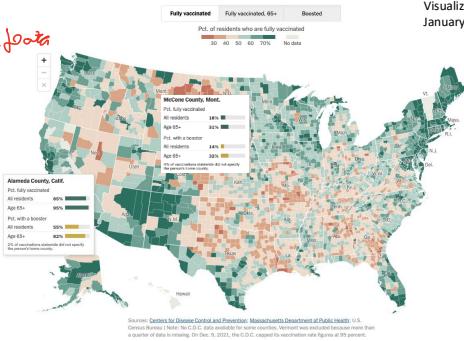
8

Why?

NYTimes
Visualization: January 2023

Save Jocks

Belief is social



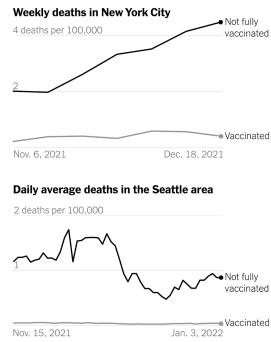
9

Why?

Data is a tool to navigate the complexity of our reality

[Link](#)

Reilly vs. Stora Complex



10

Why?

But data is easy to abuse and/or misinterpret

Academy or Propaganda

William Makis MD
@MakisMD

Stunning numbers from Denmark:

56% double vaccinated are catching 65% of "other variants" & 73% of Omicron

Most vulnerable group to Omicron BY FAR

25% boosted still catch 10% of Omicron cases, while unvaccinated catch 8.5%

This is worse than vaccine failure. This is damage.

1:08 AM - Dec 20, 2021 · Twitter Web App

809 Retweets 136 Quote Tweets

1,146 Likes

Table 4. Vaccination status for individuals ≥12 years infected with Omicron compared to other variants, data included in the table are from 22 November to 16 December 2021

Table 4. Vaccinationstatus for personer ≥12 år med omikron-infektion sammenlignet med andre varianter i perioden fra og med 22. november 2021 til og med 16. december 2021

Vaccination status	Other variants (No. of cases)	Other variants (No. of cases)	Omicron (No. of cases)	Omicron (% of cases)
At least 1 dose	8,866	8,6	1,851	10.8
Booster vaccinated	67,034	65,3	13,548	79.0
Fully vaccinated	23,481	22,9	1,454	8.5
Not vaccinated	3,216	3,1	304	1.8
Received first dose				
Total	102,608	99,9	17,155	100

Individuals aged 5-11 years have recently been invited for COVID-19 vaccination, hence the vaccination coverage is relatively low in this age group and not included in Table 4.



11

Why?

Even important entities communicate poorly

The White House
@WhiteHouse
Omicron cases are on the rise, but fully vaccinated and boostered are making a difference. Vaccines and boosters help prevent severe illness and death — if you haven't already, go get your vaccine and booster.

4:59 PM - Dec 29, 2021 · The White House

COVID-19 CASES VS. DEATHS

LAST 7 DAYS



12

Why?

Data science involves **data-focused, computational** and **inferential** approaches to:

- Explore and understand the world (aka science)
- Address and solve practical challenges (aka engineering)

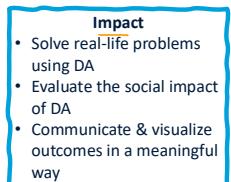
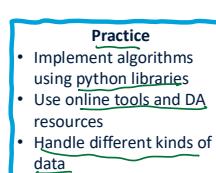
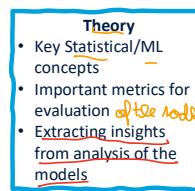
After this course, you should be able to take data and produce useful insights on the world's most challenging and ambiguous problems.

Data Driven

Practical Solution

What?

- Convey basic skills about all steps in data analysis (DA)



*We can have a big
I →*

14

13

KEN3450

Data Analysis

Chapter I:

Introduction,
Exploratory Data Analysis &
Effective Visualizations

Gerasimos (Jerry) Spanakis, PhD
<http://jerryspan.github.io>

Maastricht University



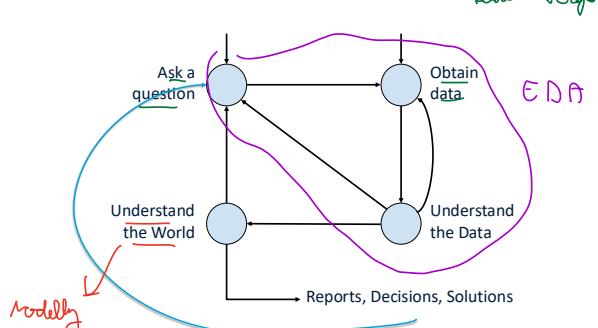
Learning goals

- Question the source of a dataset
- Practice simple pre-processing techniques
- Use EDA as a tool to further expand the analysis *or even in the Doctor Craft*
- Understand why visualization/plotting is important
- Learn aspects that tend to make visualizations effective and ineffective
- Feel comfortable designing plots that *best convey* your message *to the Doctor Craft*
- Gain experience in producing plots with Python

*about looks
and getting
message*

23

Data science lifecycle



24

This week: EDA & Visualizations

Second better 1st step



*Before trying of
running our ML model*

25

EDA key properties

More on the engineering side:

- Data sources (also relates to science)
- Data formats
- Data models (databases: not our focus)
- Data storage engines & processing (MLops: not our focus)

More on the science side:

- Granularity
- Scope
- Temporality
- Faithfulness

26

27

Data sources

- User generated *Copy User Inputs (Logs, texts), data*
- Systems generated
- Internal databases: users, inventory, customer relationships
- Third-party data *Giving from us*

Data sources

Leisure Site

Users generated data	Systems generated data <i>Node Generator</i>
User inputs <i>Could be fake</i>	Logs, metadata, predictions
Easily mal-formatted	Easier to standardize
Need to be processed <i>ASAP</i>	OK to process periodically <i>(unless to detect problems ASAP)</i>

use cases until See dm

Users' behavioral data (clicks, time spent, etc.) is often system-generated but is considered user data

28

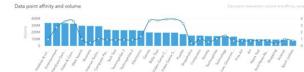
29

Third-party data: creepy but fascinating

- Types of data
 - social media, income, job
- Demographic group
 - men, age 25-34, work in tech
- More available with Mobile Advertiser ID
- Useful for learning features
 - people who like A also like B

Top interests

They love computing and electronic entertainment. If you want to reach players, try targeting all their top interests.



onaudience.com/audience-data



Remote working

Millions of people decided to stay home and work remotely to limit the spread of coronavirus. Use our Remote working segment to easily reach them and show software or products that will help them stay effective.

How did we build the segment?

- remote working
- effective ways of working from home
- tools for remote workers
- homeschooling and e-learning

If you want to reach remote workers, try extending your target group by selecting the top interests, which include Telecommuting, Career Planning or Personal Finance.



The end of tracking IDs ...

EDITORS' PICK | Jun 24, 2020, 12:38am EDT | 685,400 views

Apple Just Crippled IDFA, Sending An \$80 Billion Industry Into Upheaval

In response, the state-backed China Advertising Association, which has 2,000 members, has launched a new way to track and identify iPhone users called CAID, which is being widely tested by tech companies and advertisers in the country.

TikTok wants to keep tracking iPhone users with state-backed workaround | Ars Technica

would like permission to track you across apps and websites owned by other companies. Your data will be used to deliver personalized ads to you.

Data formats

Not Paper

Row-major
Column-major

Align

Format	Binary/Text	Human-readable	Example use cases
JSON	Text	Yes	Everywhere
CSV	Text	Yes	Everywhere
Parquet	Binary	No	Hadoop, Amazon Redshift
Avro	Binary	No	Hadoop
Protobuf	Binary	No	Google, TensorFlow (TFRecord)
Pickle	Binary	No	Python, PyTorch serialization

How to store multimodal data?

{'image': [[200,155,0], [255,255,255], ...], 'label': 'car', 'id': 1} *From Nabil Dls*

What are the access patterns

- How frequently the data will be accessed? *Choosing the best offsets how fast can occur*

- Which hardware will we use (e.g. complex ML models run on CPU/GPU/TPU)

Access Patterns determine a lot

30

31

Row-major vs. column-major

We like the Party

Major Priority

Column-major: Go into rows Colm → rows again On Feature Level		
• stored and retrieved column-by-column		
• good for accessing features		
PANDAS Column 1 Column 2 Column 3 (only for loops in Pandas is OODX)		
Sample 1
Sample 2
Sample 3

Row-major: lower and faster

- stored and retrieved row-by-row by row
- good for accessing samples

Row-major vs. column-major: DataFrame vs. ndarray

Pandas DataFrame: column-major

- accessing a row much slower than accessing a column and NumPy

Big Rows is Very Slow
(why we avoid for loops)

NumPy ndarray: row-major by default

- can specify to be column-based

```
# Get the column 'date', 1000 loops
@timeit -n1000 df["date"]
```

```
# Get the first row, 1000 loops
@timeit -n1000 df.iloc[0]
```

```
1.78 µs ± 167 ns per loop (mean ± std. dev. of 7 runs, 1000 loops each)
```

```
145 µs ± 9.41 µs per loop (mean ± std. dev. of 7 runs, 1000 loops each)
```

```
df_np = df.to_numpy()
```

```
@timeit -n1000 df_np[0]
```

```
147 ns ± 1.54 ns per loop (mean ± std. dev. of 7 runs, 1000 loops each)
```

```
204 ns ± 0.678 ns per loop (mean ± std. dev. of 7 runs, 1000 loops each)
```

<https://github.com/chiphuyen/just-pandas-things>

Text vs. binary formats

	Text files	Binary files
Examples	CSV, JSON	Parquet
Pros	Human readable	Compact
Store the number 1000000?	7 characters > 7 bytes	If stored as int32, only 4 bytes

You can unload the result of an Amazon Redshift query to your Amazon S3 data lake in Apache Parquet, an efficient open columnar storage format for analytics. Parquet format is up to 2x faster to unload and consumes up to 6x less storage in Amazon S3, compared with text formats. This enables you to save data transformation and enrichment you have done in



Data models

Data models tells us how the data is

Represented
- Unstructured
- Structured

- Describe how data is represented
 - Structured vs. unstructured
- For structured data: two main paradigms:
 - Relational model (similar to the SQL model)
 - NoSQL (gives us the .json format)

```
# Document 1: harry_potter.json
{
  "Title": "Harry Potter",
  "Author": "J.K. Rowling",
  "Publisher": "Bloomsbury Press",
  "Country": "UK",
  "Sold": 45000000,
  ("Format": "Paperback", "Price": "020"),
  ("Format": "E-book", "Price": "010")
}

# Document 2: sherlock_holmes.json
{
  "Title": "Sherlock Holmes",
  "Author": "Sir Arthur Conan Doyle",
  "Publisher": "Quava Press",
  "Country": "US",
  "Sold": 30000000,
  ("Format": "Paperback", "Price": "030"),
  ("Format": "E-book", "Price": "010")
}

# Document 3: the_n Hobbit.json
{
  "Title": "The Hobbit",
  "Author": "J.R.R. Tolkien",
  "Publisher": "Bamboo Press",
  "Country": "US",
  "Sold": 25000000,
  ("Format": "Paperback", "Price": "020"),
  ("Format": "E-book", "Price": "010")
}
```

Structures

Structured vs. unstructured data

Structured	Unstructured
Schema clearly defined	Whatever (quite literally)
Easy to search and analyze	Fast arrival (e.g. no need to clean up first) But comes at a Cost
Can only handle data with specific schema	Can handle data from any source Lessons to Start with
Has to match the Schema Defined	No need to worry about schema changes Lessons to Adapt
Schema changes will cause a lot of trouble	Data lakes → Just Dump Stuff
Data warehouses	Organized, easy to use

Stuff

Just dump stuff

Data Storage Engines & Processing

Databases optimized for

Transactional processing
(OLTP)

Benefit

Analytical processing
(OLAP)

Aggregation
Columns

- Transactions: tweeting, ordering an Uber, uploading a new model, etc.
- Operations:
 - Insert when generated
 - Occasional update/delete

- How to get aggregated information from a large amount of data?
 - e.g. what's the average ride price last month for riders at Uber?
- Operations:
 - Mostly SELECT

Aggregate View
Per

OLTP & OLAP are outdated terms



Decoupling storage & processing

- OLTP & OLAP: how data is stored is also how it's processed
 - Same data being stored in multiple databases
 - Each uses a different processing engine for different query types
- New paradigm: storage is decoupled from processing
 - Data can be stored in the same place
 - A processing layer on top that can be optimized for different query types



39

From ETL (Extract, Transform, Load) to ELT (Extract, Load, Transform)



Transform: the "meaty"/"juicy" part (the "heart" of EDA)

- cleaning, validating, transposing, deriving values, joining from multiple sources, deduplicating, splitting, aggregating, etc.

ETL (Extract, Transform, Load)

1. Extract: You take two data from the source. Similar to ETL,
2. Load: You store the two data in one or multiple databases without merging.
3. Transform: The data is transformed after being loaded into the database.

Advantages:

- More flexible because raw data is available for different types of analysis.
- More flexible because raw data is available for different types of analysis.
- Allows systems like sparkSQL to analyze data faster, readable transformations directly in the database.

Disadvantages:

- More complex because raw data is available for different types of analysis.
- More complex because raw data is available for different types of analysis.
- More complex because raw data is available for different types of analysis.

ELT (Extract, Load, Transform)

1. Extract: You take two data from the source. Similar to ETL,
2. Load: You store the two data in one or multiple databases without merging.
3. Transform: The data is transformed after being loaded into the database.

Advantages:

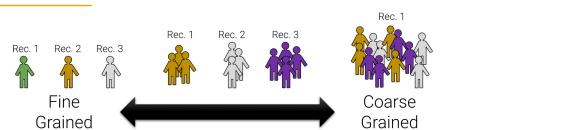
- More flexible because raw data is available for different types of analysis.
- More flexible because raw data is available for different types of analysis.
- Allows systems like sparkSQL to analyze data faster, readable transformations directly in the database.

Disadvantages:

- More complex because raw data is available for different types of analysis.
- More complex because raw data is available for different types of analysis.
- More complex because raw data is available for different types of analysis.

40

EDA - Granularity



- What does each record represent? *now*

- Purchases, persons, trees, ...

- Do all records capture granularity on the same level?

- If the data is coarse, *how were records aggregated?*

I now might be a person and another a group

Based on what?

EDA - Scope

- Does my data cover my area of interest?
 - E.g. I am studying crime data in Limburg but I only have data from Maastricht
- Is my data too broad?
 - E.g. Sampling protocol might lead to poor coverage
- Does my data cover the right time frame?
 - E.g. COVID19 test reporting



42

EDA - Temporality

- Data changes – when was the data collected?
- What is the meaning of time/date fields?
 - e.g. when the event "happened" when it happened, when it was collected..
 - e.g. when data was collected vs. entered
- Time depends on "where" US vs. Bazaar Singapore
 - Python library `datetime` is a life-saver
- Are there strange null values?
 - e.g. January 1st 1970 or 1900
- What about periodicity?
Data is skewed over a year

EDA - Faithfulness

- Do I trust this data?
- Does my data contain unrealistic or "incorrect" values
 - Dates in the future for events in the past
 - Locations that don't exist
 - Negative counts
 - Large outliers
 - ...
- Does my data violate obvious dependences?
 - Age and birthday don't match
 - ...
- Was the data entered by hand?
 - Expect spelling errors
- Are there any obvious signs of data falsification?
 - Repeated names, fake looking email addresses

>Data in the Future
Negative counts
Unusual Pairs

Common feature engineering operations

1. Handling missing values
2. Scaling
3. Discretization
4. Categorical features
5. Feature crossing

Handling missing values

- Not all missing values are equal
 - Missing not at random (MNAR)
 - Missing at random (MAR)
 - Missing completely at random (MCAR)

There are 3 types of
Missing Values



3 types

45

46

Age is missing because of the value

Handling missing values

Missing not at random – when a value is missing due to the value itself

ID	Age	Gender	Annual income	Marital status	Number of children	Job	Buy?
1		A	150,000		1	Engineer	No
2	27	B	50,000			Teacher	No
3		A	100,000	Married	2		Yes
4	40	B	\$350,000		2	Engineer	Yes
5	35	B	\$350,000	Single	0	Doctor	Yes
6		A	50,000		0	Teacher	No
7	33	B	60,000	Single		Teacher	No
8	20	B	10,000			Student	No

Value based not the fn

Handling missing values

Missing at random – when a value is missing due to another observed variable

ID	Age	Gender	Annual income	Marital status	Number of children	Job	Buy?
1		A	150,000		1	Engineer	No
2	27	B	50,000			Teacher	No
3		A	100,000	Married	2		Yes
4	40	B			2	Engineer	Yes
5	35	B		Single	0	Doctor	Yes
6		A	50,000		0	Teacher	No
7	33	B	60,000	Single		Teacher	No
8	20	B	10,000			Student	No

Age is missing because of Gender A

Part of because we still do not
know the reason. Age is missing in the
1st place
But we can see that all the line
we have
Gender A, the Data
is more

47

48

Handling missing values

Missing completely at random – there is no pattern to which values are missing Random

ID	Age	Gender	Annual income	Marital status	Number of children	Job	Buy?
1		A	150,000		1	Engineer	No
2	27	B	50,000			Teacher	No
3		A	100,000	Married	2		Yes
4	40	B			2	Engineer	Yes
5	35	B		Single	0	Doctor	Yes
6		A	50,000		0	Teacher	No
7	33	B	60,000	Single		Teacher	No
8	20	B	10,000			Student	No

What we usually think as

Random

49

How do we handle them?

- Deletion – removing data with missing entries elite Random Cohort
- Imputation – filling missing fields with certain values

Many people prefer deletion not because it's better, but it's easier to do

Once we see that there are many values, then we have 2 things we can do: Delete Input

50

Handling missing values

• Deletion

- Column deletion – remove columns with too many missing entries But could lose useful features
 - drawbacks – even if half the values are missing, the remaining data still potentially useful information for predictions We are Deleting a Feature
 - e.g. even if over half the column for 'Marital status' is missing, marital status is still highly correlated with house purchasing In this case, delete the feature
- Row deletion

Could reduce a lot our Predictive Power

Marital status
Married
Single
Single

51

Handling missing values

• Deletion

- Column deletion
- Row deletion

52

Handling missing values

• Row deletion

- Good for: data missing completely at random (MCAR) and few values missing

ID	Age	Gender	Annual income	Marital status	Number of children	Job	Buy?
1	39	A	150,000	Married	1	Engineer	No
2	27	B	50,000	Single	0	Teacher	No
3		A	100,000	Married	2		Yes
4	40	B	75,000	Married	2	Engineer	Yes
5	35	B	35,000	Single	0	Doctor	Yes
6	32	A	50,000	Married	0	Teacher	No
7	33	B	60,000	Single	2	Teacher	No
8	20	B	10,000	Single	1	Student	No

53

Handling missing values

• Row deletion

- Bad when many examples have missing fields

We could end up deleting whole data

ID	Age	Gender	Annual income	Marital status	Number of children	Job	Buy?
1		A	150,000		1	Engineer	No
2	27	B	50,000			Teacher	No
3		A	100,000	Married	2		Yes
4	40	B			2	Engineer	Yes
5	35	B		Single	0	Doctor	Yes
6		A	50,000		0	Teacher	No
7	33	B	60,000	Single		Teacher	No
8	20	B	10,000			Student	No

54

Handling missing values

We could probably Fix it

- Row deletion
 - Bad for: missing values are not at random (MNAR)
 - Missing information is information itself *For example, here we would find out what the Ford*

ID	Age	Gender	Annual income	Marital status	Number of children	Job	Buy?
1		A	150.000		1	Engineer	No
2	27	B	50.000			Teacher	No
3		A	100.000	Married	2		Yes
4	40	B	\$350.000?		2	Engineer	Yes
5	35	B	\$350.000?	Single	0	Doctor	Yes
6		A	50.000		0	Teacher	No
7	33	B	60.000	Single		Teacher	No
8	20	B	10.000			Student	No

55

Handling missing values

- Row deletion
 - Bad for: missing data at random (MAR)
 - Can potentially bias data – we've accidentally removed all examples with gender 'A'

ID	Age	Gender	Annual income	Marital status	Number of children	Job	Buy?
1		A	150.000		1	Engineer	No
2	27	B	50.000			Teacher	No
3		A	100.000	Married	2		Yes
4	40	B			2	Engineer	Yes
5	35	B		Single	0	Doctor	Yes
6		A	50.000		0	Teacher	No
7	33	B	60.000	Single		Teacher	No
8	20	B	10.000			Student	No

56

We get rid of all the People with Gender A

Imputation "Fill in the Values"

Fill missing fields with certain values

- Defaults
 - e.g. 0, or the empty string, etc.
- Statistical measures – mean, median, mode
 - e.g. if a day in July is missing its temperature value, fill it with the median temperature in July *We can be Smart*

Depending on the dataset and the application, there might be a different solution applicable!

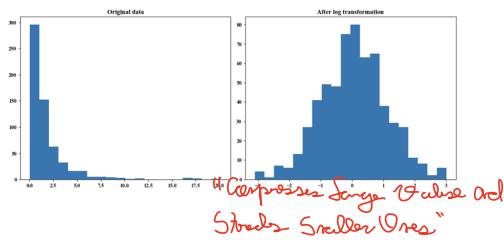
We can also try to Predict

57

Log scaling

Helps with Skewed Data

- Help with skewed data
- Often gives performance gain



59

Discretization

- Turning a continuous feature into a discrete feature (quantization)
- Create buckets for different ranges
 - Incorporate knowledge/expertise about each variable by constructing specific buckets
- Examples
 - Income
 - Lower income: $x < \$35,000$
 - Middle income: $\$35,000 \leq x < \$100,000$
 - High income: $x \geq \$100,000$
 - Age
 - Minors: $x < 18$
 - College: $18 \leq x < 22$
 - Young adult: $22 \leq x < 30$
 - 30 $\leq x < 40$
 - 40 $\leq x < 65$
 - Seniors: $x \geq 65$

Gap People

People who do not have that many bags, so again would not sum

Might Merge a Non-Natural Way

Continuous Feature $\xrightarrow{\text{Discretize}}$ Bucket

Probably doing it at Random Would not be a Great Idea

Encoding Categorical Features

- Example: you want to build a recommendation system for Amazon

- There are over 2 million brands that we need to recommend

How do we encode the different brands/vendors?

*Simple: One-Hot Encoding
What if new brands are rare?*

Categorical Features are
Interesting

Encoding Categorical Features

- one-hot encoding!
- encode unseen brands with "UNKNOWN"
- Group bottom 1% of brands and newcomers into "UNKNOWN" category
- Potential further issues: All newcomers are treated the same as unpopular brands in the platform

- Solution 1: represent each category with its attributes
- Solution 2: Hashing (widely used in industry, considered "hacky")

What if there's a New Brand

Simple: One-Hot Encoding
What if new brands are rare?

Brand
Brand A
Brand B
Brand C
Brand D
Brand E
Brand F
Brand G
Brand H
Brand I
Brand J
Brand K
Brand L
Brand M
Brand N
Brand O
Brand P
Brand Q
Brand R
Brand S
Brand T
Brand U
Brand V
Brand W
Brand X
Brand Y
Brand Z

What if there's a New Brand

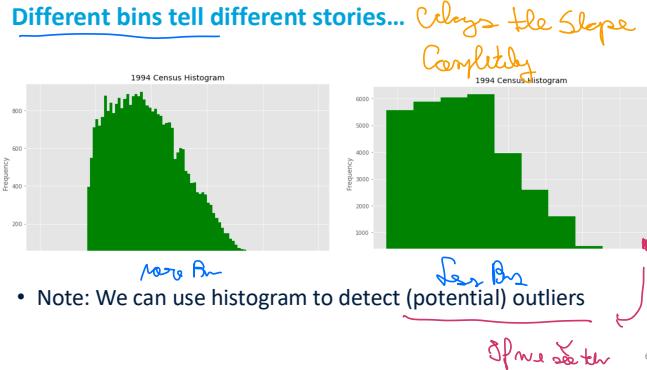
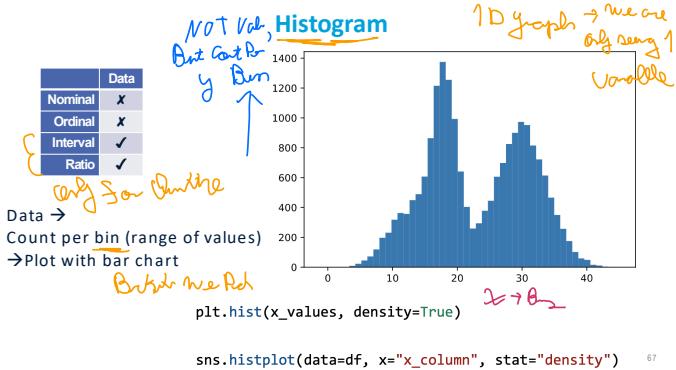
Brand
Brand A
Brand B
Brand C
Brand D
Brand E
Brand F
Brand G
Brand H
Brand I
Brand J
Brand K
Brand L
Brand M
Brand N
Brand O
Brand P
Brand Q
Brand R
Brand S
Brand T
Brand U
Brand V
Brand W
Brand X
Brand Y
Brand Z
NewBrand

Simple: One-Hot Encoding
What if new brands are rare?

Brand
Brand A
Brand B
Brand C
Brand D
Brand E
Brand F
Brand G
Brand H
Brand I
Brand J
Brand K
Brand L
Brand M
Brand N
Brand O
Brand P
Brand Q
Brand R
Brand S
Brand T
Brand U
Brand V
Brand W
Brand X
Brand Y
Brand Z

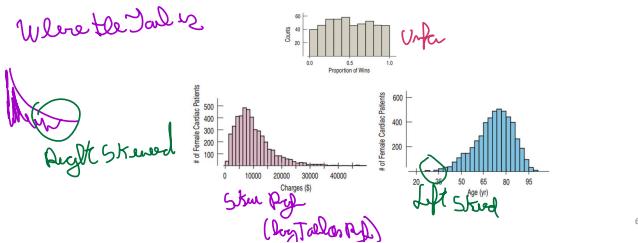
Simple: One-Hot Encoding
What if new brands are rare?

Brand
Brand A
Brand B
Brand C
Brand D
Brand E
Brand F
Brand G
Brand H
Brand I
Brand J
Brand K
Brand L
Brand M
Brand N
Brand O
Brand P
Brand Q
Brand R
Brand S
Brand T
Brand U
Brand V
Brand W
Brand X
Brand Y
Brand Z
NewBrand



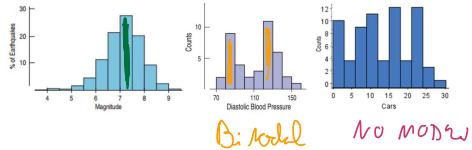
What do we look for in a histogram?

- Distribution (the first key we look at)... The slope (Uniform, Skewed Right / Skewed Left)



What do we look for in a histogram, cont.?

- Mode: A "hump" or a high-frequency bin (not Placeholder)
 - unimodal, bimodal, multimodal
 - approx. by $mean-mode = 3 \times (mean-median)$



67

68

69

70

Chapter 6.3

What is the Mode? The mode is a statistical concept used to identify the most frequently occurring value in a dataset. Let's break down the mode.

- The mode is a statistical concept used to identify the most frequently occurring value in a dataset.
- The mode is represented by "hump" or high frequency bin in Histogram.
- If the value for range of values is a histogram bin that denotes the mode in the dataset.

TYPES OF MODES IN A DATASET

- Unimodal
 - Example: The histogram in the slide shows one peak or one mode.
- Bimodal
 - The dataset has two modes or two peaks.
 - Example: The word histogram (multiple blood pressure) has peaks around 10 and 15.
- Multimodal
 - The dataset has more than two modes, indicating multiple frequent values.
 - Example: The histogram in the slide shows three peaks, so there are three modes.

Chapter 6.4

Why is Mode Important? Help identify the central tendency of data.

- Used for detecting patterns, clusters, or anomalies in data.
- The mode can be symmetric (very common).
- No mode occurs with the same frequency or there is no peak in the dataset.
- Example: The first histogram (number of cars) shows no distinct peaks.
- Example: The second histogram (number of cars) shows two distinct peaks.
- Example: The third histogram (number of cars) shows three distinct peaks.
- Example: The fourth histogram (number of cars) shows four distinct peaks.
- Example: The fifth histogram (number of cars) shows five distinct peaks.
- Example: The sixth histogram (number of cars) shows six distinct peaks.
- Example: The seventh histogram (number of cars) shows seven distinct peaks.
- Example: The eighth histogram (number of cars) shows eight distinct peaks.
- Example: The ninth histogram (number of cars) shows nine distinct peaks.
- Example: The tenth histogram (number of cars) shows ten distinct peaks.
- Example: The eleventh histogram (number of cars) shows eleven distinct peaks.
- Example: The twelfth histogram (number of cars) shows twelve distinct peaks.
- Example: The thirteenth histogram (number of cars) shows thirteen distinct peaks.
- Example: The fourteenth histogram (number of cars) shows fourteen distinct peaks.
- Example: The fifteenth histogram (number of cars) shows fifteen distinct peaks.
- Example: The sixteenth histogram (number of cars) shows sixteen distinct peaks.
- Example: The seventeenth histogram (number of cars) shows seventeen distinct peaks.
- Example: The eighteenth histogram (number of cars) shows eighteen distinct peaks.
- Example: The nineteenth histogram (number of cars) shows nineteen distinct peaks.
- Example: The twentieth histogram (number of cars) shows twenty distinct peaks.
- Example: The twenty-first histogram (number of cars) shows twenty-one distinct peaks.
- Example: The twenty-second histogram (number of cars) shows twenty-two distinct peaks.
- Example: The twenty-third histogram (number of cars) shows twenty-three distinct peaks.
- Example: The twenty-fourth histogram (number of cars) shows twenty-four distinct peaks.
- Example: The twenty-fifth histogram (number of cars) shows twenty-five distinct peaks.
- Example: The twenty-sixth histogram (number of cars) shows twenty-six distinct peaks.
- Example: The twenty-seventh histogram (number of cars) shows twenty-seven distinct peaks.
- Example: The twenty-eighth histogram (number of cars) shows twenty-eight distinct peaks.
- Example: The twenty-ninth histogram (number of cars) shows twenty-nine distinct peaks.
- Example: The thirty histogram (number of cars) shows thirty distinct peaks.
- Example: The thirty-first histogram (number of cars) shows thirty-one distinct peaks.
- Example: The thirty-second histogram (number of cars) shows thirty-two distinct peaks.
- Example: The thirty-third histogram (number of cars) shows thirty-three distinct peaks.
- Example: The thirty-fourth histogram (number of cars) shows thirty-four distinct peaks.
- Example: The thirty-fifth histogram (number of cars) shows thirty-five distinct peaks.
- Example: The thirty-sixth histogram (number of cars) shows thirty-six distinct peaks.
- Example: The thirty-seventh histogram (number of cars) shows thirty-seven distinct peaks.
- Example: The thirty-eighth histogram (number of cars) shows thirty-eight distinct peaks.
- Example: The thirty-ninth histogram (number of cars) shows thirty-nine distinct peaks.
- Example: The forty histogram (number of cars) shows forty distinct peaks.
- Example: The forty-first histogram (number of cars) shows forty-one distinct peaks.
- Example: The forty-second histogram (number of cars) shows forty-two distinct peaks.
- Example: The forty-third histogram (number of cars) shows forty-three distinct peaks.
- Example: The forty-fourth histogram (number of cars) shows forty-four distinct peaks.
- Example: The forty-fifth histogram (number of cars) shows forty-five distinct peaks.
- Example: The forty-sixth histogram (number of cars) shows forty-six distinct peaks.
- Example: The forty-seventh histogram (number of cars) shows forty-seven distinct peaks.
- Example: The forty-eighth histogram (number of cars) shows forty-eight distinct peaks.
- Example: The forty-ninth histogram (number of cars) shows forty-nine distinct peaks.
- Example: The五十 histogram (number of cars) shows fifty distinct peaks.
- Example: The fifty-first histogram (number of cars) shows fifty-one distinct peaks.
- Example: The fifty-second histogram (number of cars) shows fifty-two distinct peaks.
- Example: The fifty-third histogram (number of cars) shows fifty-three distinct peaks.
- Example: The fifty-fourth histogram (number of cars) shows fifty-four distinct peaks.
- Example: The fifty-fifth histogram (number of cars) shows fifty-five distinct peaks.
- Example: The fifty-sixth histogram (number of cars) shows fifty-six distinct peaks.
- Example: The fifty-seventh histogram (number of cars) shows fifty-seven distinct peaks.
- Example: The fifty-eighth histogram (number of cars) shows fifty-eight distinct peaks.
- Example: The fifty-ninth histogram (number of cars) shows fifty-nine distinct peaks.
- Example: The六十 histogram (number of cars) shows sixty distinct peaks.
- Example: The sixty-first histogram (number of cars) shows sixty-one distinct peaks.
- Example: The sixty-second histogram (number of cars) shows sixty-two distinct peaks.
- Example: The sixty-third histogram (number of cars) shows sixty-three distinct peaks.
- Example: The sixty-fourth histogram (number of cars) shows sixty-four distinct peaks.
- Example: The sixty-fifth histogram (number of cars) shows sixty-five distinct peaks.
- Example: The sixty-sixth histogram (number of cars) shows sixty-six distinct peaks.
- Example: The sixty-seventh histogram (number of cars) shows sixty-seven distinct peaks.
- Example: The sixty-eighth histogram (number of cars) shows sixty-eight distinct peaks.
- Example: The sixty-ninth histogram (number of cars) shows sixty-nine distinct peaks.
- Example: The七十 histogram (number of cars) shows seventy distinct peaks.
- Example: The seventy-first histogram (number of cars) shows seventy-one distinct peaks.
- Example: The seventy-second histogram (number of cars) shows seventy-two distinct peaks.
- Example: The seventy-third histogram (number of cars) shows seventy-three distinct peaks.
- Example: The seventy-fourth histogram (number of cars) shows seventy-four distinct peaks.
- Example: The seventy-fifth histogram (number of cars) shows seventy-five distinct peaks.
- Example: The seventy-sixth histogram (number of cars) shows seventy-six distinct peaks.
- Example: The seventy-seventh histogram (number of cars) shows seventy-seven distinct peaks.
- Example: The seventy-eighth histogram (number of cars) shows seventy-eight distinct peaks.
- Example: The seventy-ninth histogram (number of cars) shows seventy-nine distinct peaks.
- Example: The eighty histogram (number of cars) shows eighty distinct peaks.
- Example: The eighty-first histogram (number of cars) shows eighty-one distinct peaks.
- Example: The eighty-second histogram (number of cars) shows eighty-two distinct peaks.
- Example: The eighty-third histogram (number of cars) shows eighty-three distinct peaks.
- Example: The eighty-fourth histogram (number of cars) shows eighty-four distinct peaks.
- Example: The eighty-fifth histogram (number of cars) shows eighty-five distinct peaks.
- Example: The eighty-sixth histogram (number of cars) shows eighty-six distinct peaks.
- Example: The eighty-seventh histogram (number of cars) shows eighty-seven distinct peaks.
- Example: The eighty-eighth histogram (number of cars) shows eighty-eight distinct peaks.
- Example: The eighty-ninth histogram (number of cars) shows eighty-nine distinct peaks.
- Example: The九十 histogram (number of cars) shows ninety distinct peaks.
- Example: The ninety-first histogram (number of cars) shows ninety-one distinct peaks.
- Example: The ninety-second histogram (number of cars) shows ninety-two distinct peaks.
- Example: The ninety-third histogram (number of cars) shows ninety-three distinct peaks.
- Example: The ninety-fourth histogram (number of cars) shows ninety-four distinct peaks.
- Example: The ninety-fifth histogram (number of cars) shows ninety-five distinct peaks.
- Example: The ninety-sixth histogram (number of cars) shows ninety-six distinct peaks.
- Example: The ninety-seventh histogram (number of cars) shows ninety-seven distinct peaks.
- Example: The ninety-eighth histogram (number of cars) shows ninety-eight distinct peaks.
- Example: The ninety-ninth histogram (number of cars) shows ninety-nine distinct peaks.
- Example: The一百 histogram (number of cars) shows一百 distinct peaks.

Approximation Formula

The mode can be approximated using the formula:

Mode = Mean - 3 × (Mean - Median)

This works for skewed distributions:

If the data is right-skewed, the mode is less than the mean.

If the data is left-skewed, the mode is greater than the mean.

How to Interpret Histograms?

1. Look for shape or peaks.

- Only one → Unimodal.

- Two humps → Bimodal.

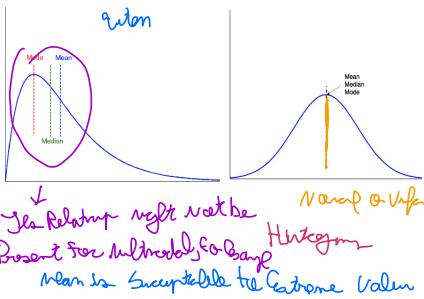
- Many → Multimodal.

2. If there's one clear peak, the distribution looks a meaningful mode.

If there's two peaks, we need to calculate the mode by averaging a specific histogram.

Mode = (Mode1 + Mode2) / 2

Intuition of mean, median, mode



71

Range, Variance, Standard Deviation

$$\text{Range} = \text{Maximum} - \text{Minimum}$$

$$\text{Mean}, \mu = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\text{Variance} = \sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$$

$$\text{Std. dev.} = \sigma$$

not be affected by outliers

Sum of the sum of All data points - mean Squared

That's why we go

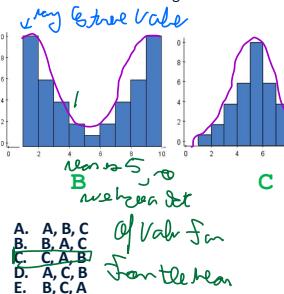
Sum of the Deviations

Sum of the Deviations squared

72

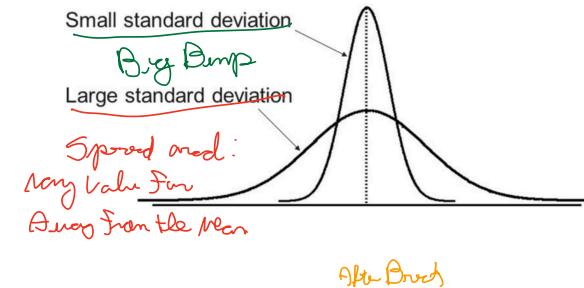
Histograms and standard deviation

Order the histograms below from smallest standard deviation to largest standard deviation.



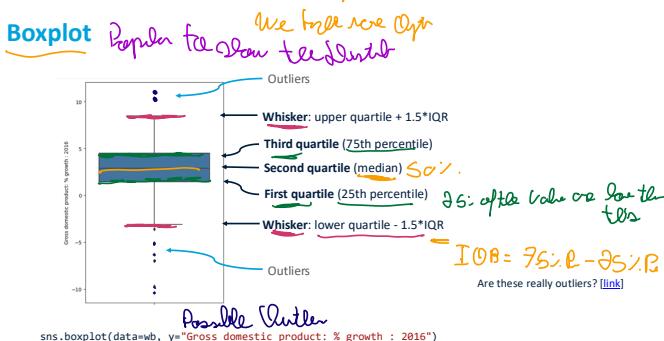
73

Histograms and standard deviation



74

Boxplot



75

Nominal and ordinal data types

Categorical

- Don't try to compute any numerical summary DO NOT
- Only (relative) frequency tables

Dot

Class	Count
First	325
Second	285
Third	706
Crew	885

Class	%
First	14.77
Second	12.95
Third	32.08
Crew	40.21

```
In [20]:
print(trace.value_counts()) #frequency table
print(trace.value_counts() / len(trace)) #relative frequency table
```

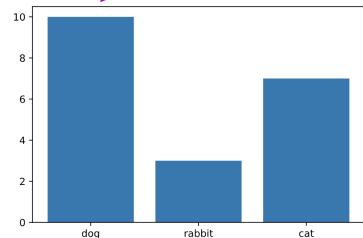
White	27816
Black	3124
Asian-Pac-Islander	1029
Amer-Indian-Eskimo	311
Other	271

As far as I can get with Pandas

76

	Data
Nominal	✓
Ordinal	✓
Interval	X
Ratio	X

Bar chart

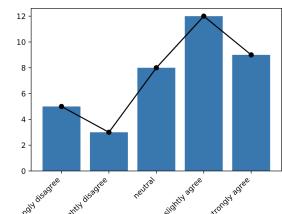
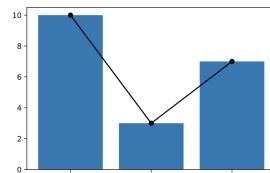


```
animals = wb["Animals"].value_counts()
plt.bar(animals.index, animals.values);
```

77

Bar chart (bad)

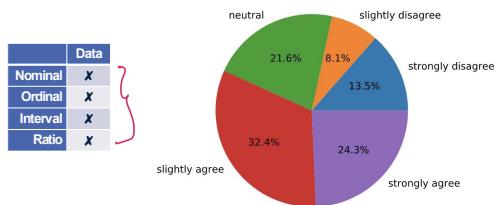
Don't use lines within a bar chart for categorial or ordinal features!



Even if there is ordinality

78

First imply that this is a Categorical
Pie chart (mostly bad)



	Data
Nominal	X
Ordinal	X
Interval	X
Ratio	X

Try to Avoid

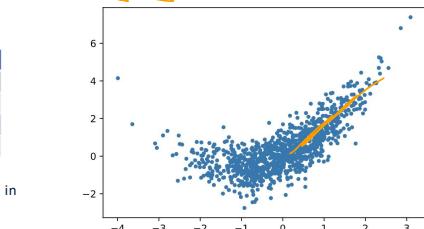
79

What about 2 Variable together
Scatterplot

	Dim 1	Dim 2
Nominal	X	X
Ordinal	X	X
Interval	✓	✓
Ratio	✓	✓

Why not ordinal data in first dimension?

general trend



plt.scatter(x_values, y_values)

sns.scatterplot(data=df, x="x_column", y="y_column", hue="hue_column")

rule of going up
the other side to
group 80

Correlation Put a # to the Answer

- The coefficient of two variables x and y is given by

$$\text{corr}(x, y) = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}$$

$$\text{cov}(x, y) = \frac{1}{n} \sum_{i=1}^n (x_i - \mu_x)(y_i - \mu_y)$$

- Assumptions, conditions, properties

- Correlation assumes a linear relationship between x and y Boring
- Outliers can strongly affect the value of correlation
- $\text{corr} > 0 \rightarrow \text{positive association}$, $\text{corr} < 0 \rightarrow \text{negative association}$
- $-1 \leq \text{corr} \leq 1$
- Interchanging x and y does not change the correlation Same
- Corr has units
- Changing the units of x or y does not affect corr

Introducing

<http://guessthecorrelation.com/>

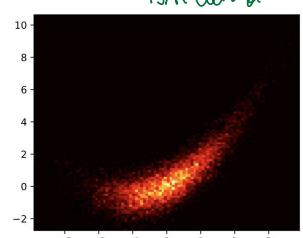
See you at the mean

Try to go together

Heatmap (density or 2D histogram)

	Dim 1	Dim 2
Nominal	X	X
Ordinal	X	X
Interval	✓	✓
Ratio	✓	✓

Am Correlation



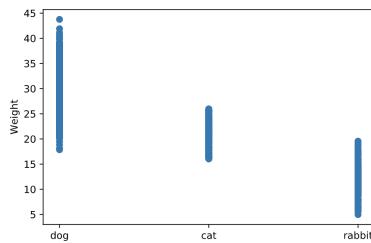
Heat
Density
Matrix
By

Please like

82

Scatterplot (very very bad)

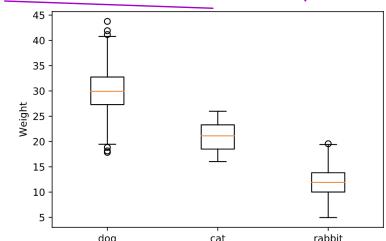
	Dim 1	Dim 2
Nominal	X	X
Ordinal	X	X
Interval	✓	✓
Ratio	✓	✓



83

side-by-side boxplot

	Dim 1	Dim 2
Nominal	✓	X
Ordinal	✓	X
Interval	X	✓
Ratio	X	✓



Performance of Model

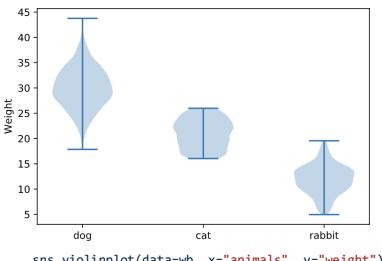
`sns.boxplot(data=wb, x="animals", y="weight");`

84

side-by-side violin plot

Also short for the Violin Plot

	Dim 1	Dim 2
Nominal	✓	X
Ordinal	✓	X
Interval	X	✓
Ratio	X	✓

`sns.violinplot(data=wb, x="animals", y="weight");`

85

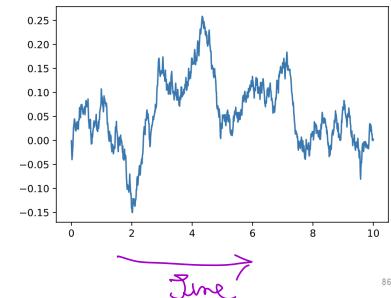
When Dim 1 is Temporal

	Dim 1	Dim 2
Nominal	X	X
Ordinal	X	X
Interval	✓	✓
Ratio	✓	✓

Why not ordinal data in first dimension?

Line plot

Plot over Time

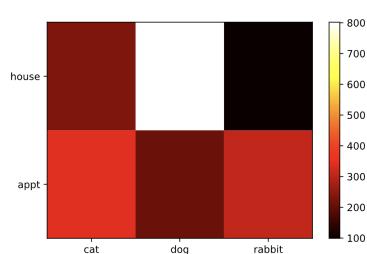


86

Heatmap (matrix)

2 Categorical

	Dim 1	Dim 2
Nominal	✓	✓
Ordinal	✓	✓
Interval	X	X
Ratio	X	X

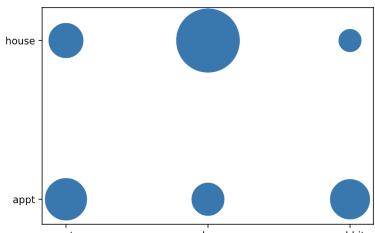


87

Bubbleplot

Same 3d

	Dim 1	Dim 2
Nominal	✓	✓
Ordinal	✓	✓
Interval	X	X
Ratio	X	X

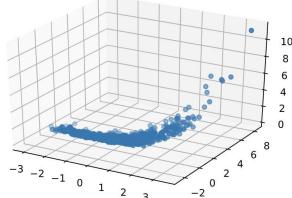


88

3D scatterplot



	Dim 1	Dim 2	Dim 3
Nominal	x	x	x
Ordinal	x	x	x
Interval	x	x	x
Ratio	x	x	x



Avoid 3D!

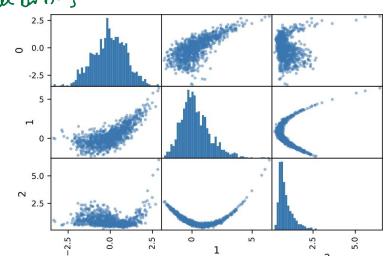
89

Scatterplot matrix

In R

Diagonal + var all in the
we see two

	Dim 1	Dim 2	Dim 3
Nominal	x	x	x
Ordinal	x	x	x
Interval	✓	✓	✓
Ratio	✓	✓	✓

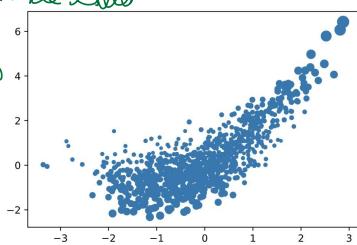


90

Bubbleplot ("light" 3D)

Plan By is the 3D
↑

	Dim 1	Dim 2	Dim 3
Nominal	x	x	x
Ordinal	x	x	x
Interval	✓	✓	✓
Ratio	✓	✓	✓

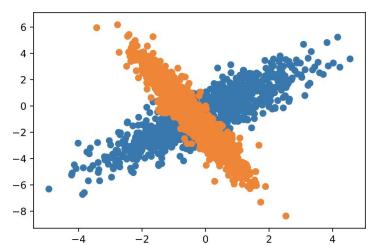


91

Color scatterplot ("light" 3D)

Cluster?

	Dim 1	Dim 2	Dim 3
Nominal	x	x	✓
Ordinal	x	x	✓
Interval	✓	✓	x
Ratio	✓	✓	x



Color: Yes or No

92

HARD DONTs

- Don't make histograms for nominal/ordinal data! *Categorical*
- Don't compute numerical summaries for nominal/ordinal data!
 - e.g. mean student i-number does not anything *no意义*
- Don't look for shape, spread etc. in a bar chart *when Plots* *Catg*
- Choose an appropriate bin for histograms
- Beware of outliers
- Not everything might be possible to plot/visualize

What questions to ask?

Author of
Ques

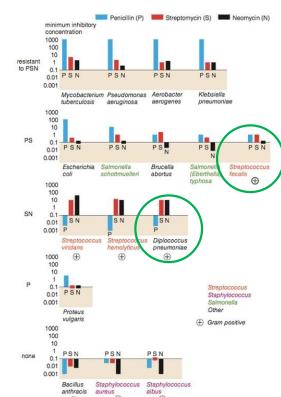
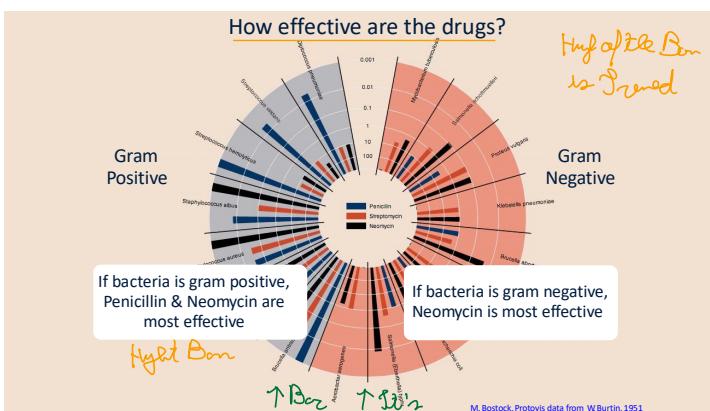
Genus, Species	Hauterive Pore City			QR code	
	Penicillin	Spectomycin	Neomycin	Gram Staining	
Bacteria					
<i>Aerobacter aerogenes</i>	\$70	1	1.6	negative	
<i>Brucella abortus</i>	Min.Inhibition	1	2	0.02	negative
<i>Brucella anthracis</i>	0.001	0.01	0.007	positive	
<i>Diplococcus pneumoniae</i>	Concentration	0.005	11	10	positive
<i>Escherichia coli</i>	[m/g]	100	0.4	0.1	negative
<i>Klebsiella pneumoniae</i>	850	1.2	1	negative	
<i>Mycobacterium tuberculosis</i>	800	5	2	negative	
<i>Proteus vulgaris</i>	3	0.1	0.1	negative	
<i>Pseudomonas aeruginosa</i>	850	2	0.4	negative	
<i>Salmonella (Enteritidis) typhosa</i>	1	0.4	0.008	negative	
<i>Salmonella schottmuelleri</i>	10	0.8	0.09	negative	
<i>Staphylococcus albus</i>	0.007	0.1	0.001	positive	
<i>Staphylococcus aureus</i>	0.03	0.03	0.001	positive	
<i>Streptococcus faecalis</i>	1	1	0.1	positive	
<i>Streptococcus hemolyticus</i>	0.001	14	10	positive	
<i>Streptococcus viridans</i>	0.005	10	40	positive	



93

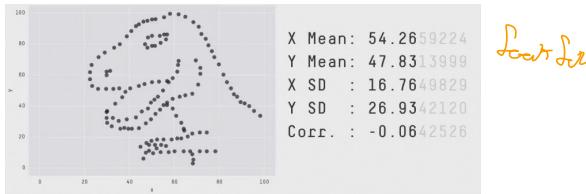
W.Burton,(1951)

94



Why don't we just use numbers? Datasaurus

<https://www.autodesk.com/research/publications/same-stats-different-graphs>

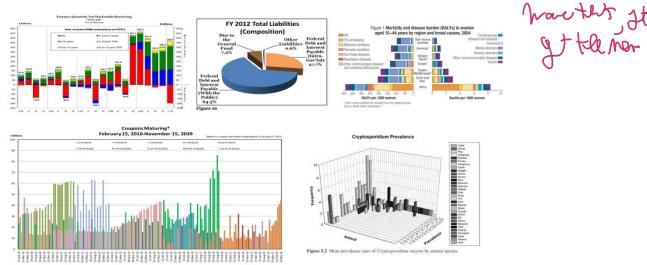


Visualization motivation *Why*

Visualizations help us to analyze and explore the data.
They help to:

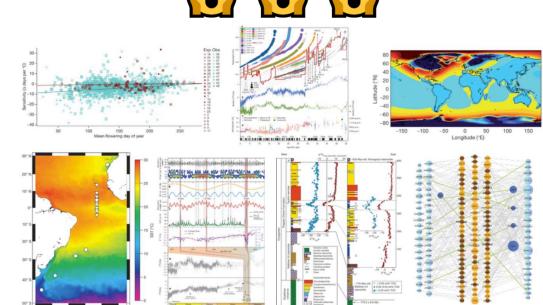
- Identify hidden patterns and trends *that we could not see in the raw data*
- Formulate/test hypotheses
- Communicate any modeling results *instead of just the numbers*
 - Present information and ideas succinctly
 - Provide evidence and support
 - Influence and persuade
- Determine the next step in analysis/modeling *Super Cool*

But there are many bad examples...



We do not want this, it gets them

Even more...



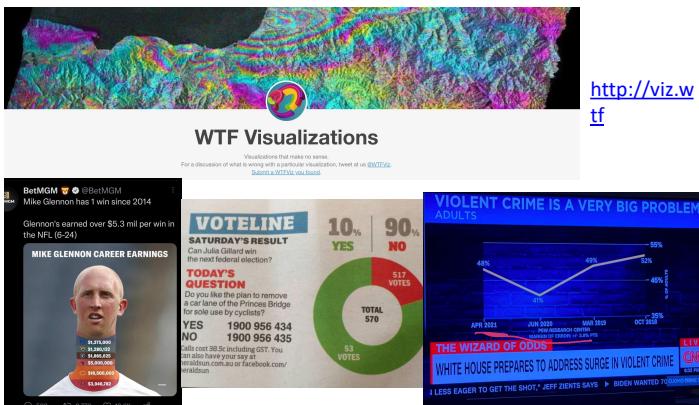
Sources: US Treasury and WHO reports

99

97

98

100



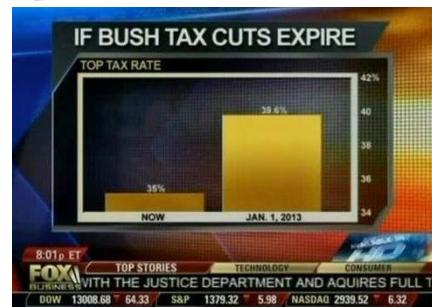
Effective visualizations

Data Story

Simple but effective

1. Have graphical integrity
2. Keep it simple...enough
3. Use sensible design
4. Use the right display

What's the problem here?



Should start at 0

1. Graphical integrity

- Show data variation, ~~no~~ design variation
- Clear, detailed and thorough labeling and appropriate scales *should be standard*
- Plot all the data
- Be proportional

*Take any old
Should still make
sense*

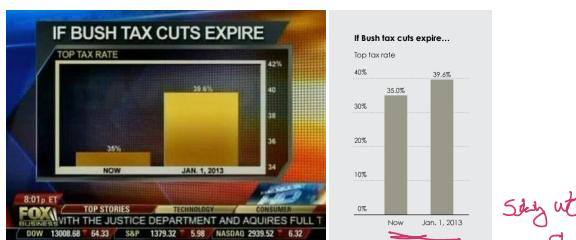
103

102

Another example



Problem: Scale distortions

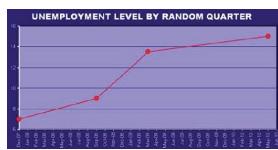


*Stay w/
0*

105

106

Problem: Scale distortions



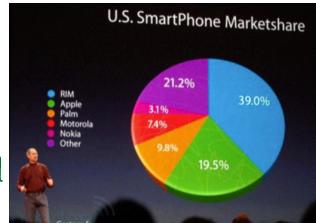
COVID-19 gave us some really nice plots...



What are the problems here?

- Violates the Area Principle

It's tilted

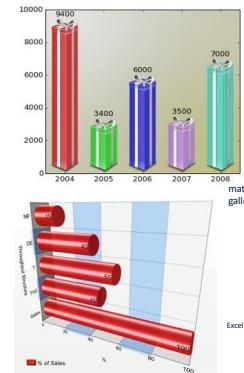
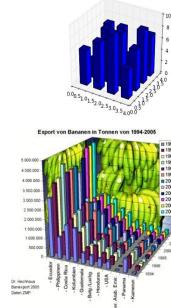


- Add up the percents!



109

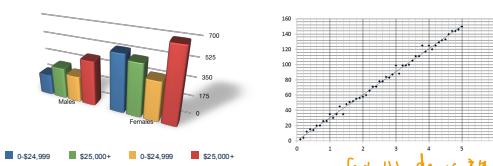
2. Keep it simple



110

2. Keep it simple

- Size of the graphic effect should be directly proportional to the numerical quantities ("lie factor")



111

Maximize data-ink ratio

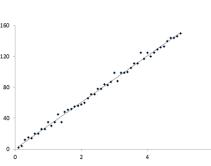


$$\text{Data-Ink Ratio} = \frac{\text{Data ink}}{\text{Total ink used in graphic}}$$

We use Money



112



112

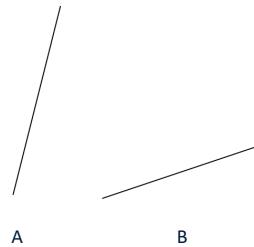
How much longer?



- A. 2
B. 3
C. 4
D. 5
E. 6

113

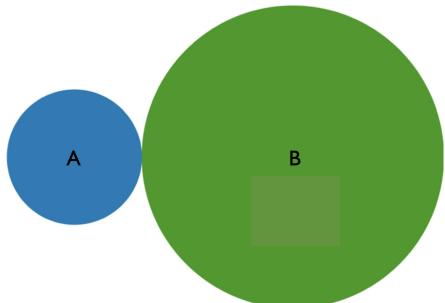
How much steeper slope?



- A. 2
B. 3
C. 4
D. 5
E. 6

114

How much larger area?

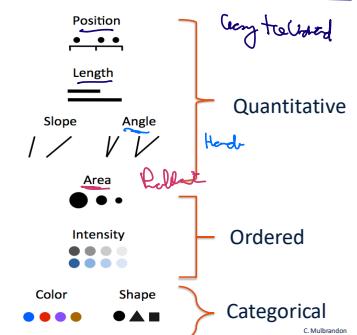


- A. 2
B. 3
C. 4
D. 5
E. 6

115

30 years of visualization research in 1 slide

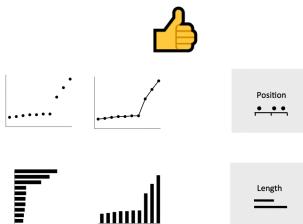
Most Efficient
↓
Least Efficient



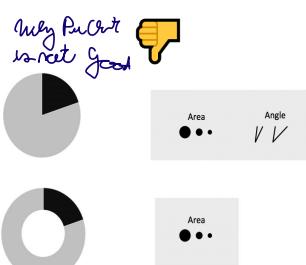
C.Muller et al.
VisualizingEconomics.com

116

Most vs. less effective



We like this ^



VisualizingEconomics.com

117

Pie charts have bad reputation

<http://eagerpies.com/>



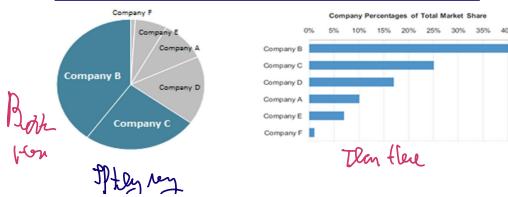
not too interesting

118

...but sometimes they are useful



65% of the market is controlled by companies B and C



119

3. Use a sensible design

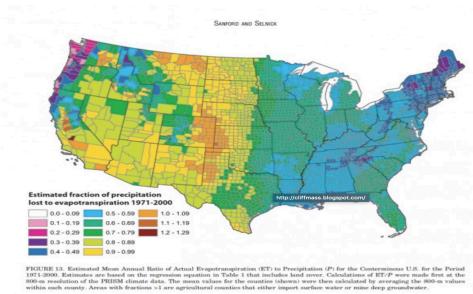
Use color strategically

- For quantitative variables: Do not use color
- For ordinal data: Try varying luminance & saturation
- For categorical data: Do not use more than 5-8 colors



120

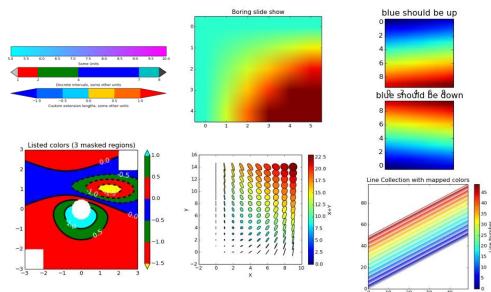
Least effective



121

Avoid Rainbow Colors!

Not Really Afford



122

Color blindness

Avoiding Red

- 10% of males, 1% of females
- Most common: red-green weakness / blindness



Normal Color Perception Deutanopia (no green receptors) Protanopia (no red receptors)

- There are packages that take care of this!

123

4. Choose the right display

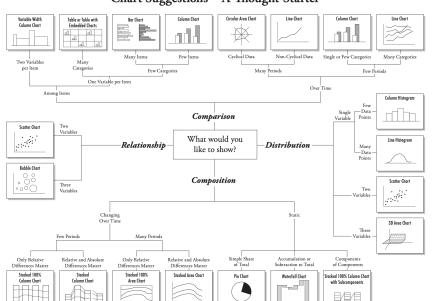
What do you want your visualization to show about your data?

- Distribution:** how a variable or variables in the dataset distribute over a range of possible values.
- Relationship:** how the values of multiple variables in the dataset relate
- Composition:** how the dataset breaks down into subgroups
- Comparison:** how trends in multiple variable or datasets compare

124

4. Choose the right display

Chart Suggestions—A Thought-Starter



http://extremepresentation.typepad.com/blog/files/choosing_a_good_chart.p

Devon

Summary

- Reality check on your dataset
- Per variable:
 - Consistency check, missing data, outliers, noise...
 - Histograms, barplots, boxplots, etc.
 - Transformations/engineering possible
- Per pairs of variables (or more):
 - Associations (conditional tables)
 - Correlations (linearity check) for quantitative variables
- Tell your story via scientific, verbal and visual means

125

126