Data Analysis & Visualisation

Week 2

Raoul Grouls, 26-2-2024

Recap leerdoelen les 1

- de gestalt principes & five guidelines toepassen op visualisaties
- een virtual environment activeren met uv
- nieuwe features extraheren met behulp van regular expressions
- Een script vanaf de terminal opstarten
- click gebruiken voor command line arguments bij een script
- begrijpt de opzet van een project (src folder, data/raw en data/processed, pyproject.toml, notebooks) en kan dit zelf opzetten
- kan een eigen git-repo maken

- Regular expressions toepassen:
  - start ^
  - end $
  - or: [Bb]
  - ranges [a-zA-Z]
  - any char .
  - zero or more a*
  - one or more a+
  - not in range [^a-z]
  - shortcuts (\w, \s, \d)
  - lookbehind (?<=...)
  - lookahead (?=...)

leerdoelen les 2

- Leren toepassen van visualisatie principes
- Omgaan met venv, uv, path, scripts en git
- Oefenen met nieuwe features extraheren met behulp van regular expressions
- Vergelijken van categorieen met behulp van data visualisaties:
  - Barplots
  - Barbell plot
  - heatmaps
- Werken met palettes (en list comprehensions)
- Pandas
  - Pandas groupby & aggregate
  - Pandas cut

rollover for more detail

What Makes a Good Visualization?

explicit (implicit)

David McCandless
InformationisBeautiful.net

taken from new book
Knowledge is Beautiful

find out more
bit.ly/KIB_Books

- Meer story, minder information?
- preattentive processing vs bestuderen?

What better visual reference than a musical score to show the repetitiveness of Giorgia's protests and the "level" of complaint: whether they are justified or totally out of place.

Note the hand-drawn stamps: these postcards were delivered in person in New York!

What is Consciousness?

Make up your own mind

A field that exists in its own parallel "realm" of existence outside reality so can't be seen.
(Substance Dualism)

A sensation that "grows" inevitably out of complicated brain states.
(Emergent Dualism)

A physical property of all matter, like electromagnetism, just not one the scientists know about.
(Property Dualism)

All matter has a psychic part. Consciousness is just the psychic part of our brain.
(Pan Psychism)

Simply, mental states are physical events that we can see in brain scans.
(Identity Theory)

Consciousness and its states (belief, desire, pain) are simply functions the brain performs.
(Functionalism)

Literally just behaviour. When we behave in a certain way, we appear conscious.
(Behaviourism)

An accidental side-effect of complex physical processes in the brain.
(Epiphenomenalism)

HOW LONG CAN YOU WORK ON MAKING A ROUTINE TASK MORE EFFICIENT BEFORE YOU'RE SPENDING MORE TIME THAN YOU SAVE?

(ACROSS FIVE YEARS)

HOW MUCH TIME YOU SHAVE OFF	HOW OFTEN YOU DO THE TASK: 50/DAY	5/DAY	DAILY	WEEKLY	MONTHLY	YEARLY
1 SECOND	1 DAY	2 HOURS	30 MINUTES	4 MINUTES	1 MINUTE	5 SECONDS
5 SECONDS	5 DAYS	12 HOURS	2 HOURS	21 MINUTES	5 MINUTES	25 SECONDS
30 SECONDS	4 WEEKS	3 DAYS	12 HOURS	2 HOURS	30 MINUTES	2 MINUTES
1 MINUTE	8 WEEKS	6 DAYS	1 DAY	4 HOURS	1 HOUR	5 MINUTES
5 MINUTES	9 MONTHS	4 WEEKS	6 DAYS	21 HOURS	5 HOURS	25 MINUTES
30 MINUTES		6 MONTHS	5 WEEKS	5 DAYS	1 DAY	2 HOURS
1 HOUR		10 MONTHS	2 MONTHS	10 DAYS	2 DAYS	5 HOURS
6 HOURS				2 MONTHS	2 WEEKS	1 DAY
1 DAY					8 WEEKS	5 DAYS

Confidence

- Absolute vs relative numbers
- Confidence comes with higher numbers

Errorbars

- Standard error
- Confidence intervals
- Standard deviation
- Range

Errorbars

Suppose statistically independent n samples $x_1, x_2, \ldots, x_n$ from a population

Suppose the population has a standard deviation of σ

The standard error is calculated with $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$

Because we rarely know the σ of the population, we often estimate it with the sample standard deviation.

THE COUNTRIES THAT USE THE MOST TOILET PAPER

If you laid out all of the rolls of toilet paper used in each country in one year, what would that look like? For China it would be an incredible 4 billion miles long, which is further than the distance from Earth to Neptune. For the USA it's 2.65 billion miles, while both Japan and Germany use enough toilet paper to stretch out beyond Jupiter.

▸ TOILET PAPER USAGE PER YEAR PER COUNTRY

CHINA	4,007,208,240 MILES	
UNITED STATES	4,267,630,419 KM	2,651,782,602 MILES
JAPAN	1,052,415,691 KM	653,940,793 MILES
GERMANY	1,026,801,290 KM	637,900,468 MILES
UNITED KINGDOM	788,352,099 KM	489,859,284 MILES
BRAZIL	738,584,457 KM	458,935,105 MILES
FRANCE	423,771,298 KM	263,318,277 MILES
ITALY	387,004,056 KM	240,473,172 MILES
SPAIN	346,295,809 KM	215,178,240 MILES
SOUTH KOREA	328,163,799 KM	203,911,531 MILES
CANADA	286,444,833 KM	177,988,567 MILES
AUSTRALIA	205,190,427 KM	127,499,420 MILES
MEXICO	200,423,386 KM	124,537,318 MILES
TAIWAN	152,446,413 KM	94,725,810 MILES
VENEZUELA	137,809,685 KM	85,630,956 MILES
PORTUGAL	127,737,029 KM	79,372,110 MILES
BELGIUM	127,170,815 KM	79,020,157 MILES
ARGENTINA	115,715,644 KM	71,902,368 MILES
RUSSIA	106,753,978 KM	66,333,846 MILES
CHILE	104,879,125 KM	65,168,867 MILES
INDONESIA	100,043,997 KM	62,164,458 MILES
CZECH REPUBLIC	97,922,922 KM	60,846,489 MILES
COLOMBIA	93,054,631 KM	57,821,487 MILES
SOUTH AFRICA	92,194,172 KM	57,286,603 MILES
SWEDEN	81,285,958 KM	50,498,325 MILES

Distance to the Sun
93,000,000
MILES

Furthest distance to Mars
250,000,000
MILES

Furthest distance to Jupiter
601,000,000
MILES

Furthest distance to Neptune
2,700,000,000
MILES

the disease is new to medicine, and data is still coming in.

DEATH TOLL

[HIGHEST TO LOWEST]

200M
Black Death (Bubonic Plague)
1347-1351

56M
Small Pox
1520

40-50M
Spanish Flu
1918-1919

30-50M
Plague of Justinian
541-542

The plague originated in rats and spread to humans via infected fleas.

The outbreak wiped out 30-50% of Europe's population. It took more than 200 years for the continent's population to recover.

Smallpox killed an estimated 90% of Native Americans. In Europe during the 1800s, an estimated 400,000 people were being killed by smallpox annually. The first ever vaccine was created to ward off smallpox.

The death toll of this plague is still under debate as new evidence is uncovered, but many think it may have helped hasten the fall of the Roman Empire.

A series of **Cholera** outbreaks spread around the world in the 1800s killing millions of people. There is no solid consensus on death tolls.

25-35M
HIV/AIDS
1981-PRESENT

12M
The Third Plague
1855

5M
Antonine Plague
165-180

3M
17th Century Great Plagues
1665

1.1M
Asian Flu
1957-1958

1M
Russian Flu
1889-1890

1M
Hong Kong Flu
1968-1970

1M
Cholera 6 outbreak
1817-1923

1M
Japanese Smallpox Epidemic
735-737

600K
18th Century Great Plagues
1817-1923

200K
Swine Flu
2009-2010

100-150K
Yellow Fever
LATE 1800s

11.3K
Ebola
2014-2016

850
MERS
2015-PRESENT

770
SARS
2002-2003

4.7K*
Novel Coronavirus (COVID-19)
2019-PRESENT

Number of COVID-19 deaths reported to WHO (cumulative total)

World

WHO Regions: Africa, Americas, Eastern Mediterranean, Europe, South-East Asia, Western Pacific

7,031,216

Reported COVID-19 deaths

11 February 2024

Number of COVID-19 deaths reported to WHO (cumulative total)

World

Country	Deaths
United States of America	1.2m
Brazil	702.1k
India	533.5k
Russian Federation	402.1k
Mexico	335k
United Kingdom of Great Britain And	232.1k

Show less

https://data.who.int/dashboards/covid19/deaths?n=c

Causes of death in children under 5

(2013)

Source: IHME

gatesnotes

The Rise of Partisanship and Super-Cooperators in the U.S. House of Representatives

Which Animals Could You Beat in a Fight?

"Which of the following animals, if any, do you think you could beat in a fight if you were unarmed?"

Survey of 2,082 GB adults (conducted 18-19 May 2021) & 1,224 U.S. adults (conducted 12-13 April 2021).

Source: YouGov

Color palettes

https://seaborn.pydata.org/tutorial/color_palettes.html

- **Qualitative**: representing categorical data
- **Sequential**: perceptually uniform
- **Diverging**: both large low and high values are interesting and span a midpoint value

Git

- The only sane way to work together on code
- https://learngitbranching.js.org/

Version Control

Git gets easier once you get the basic idea that branches are homeomorphic endofunctors mapping submanifolds of a Hilbert space.

Main
Develop
v0.1
v0.2
v1.0

Main
Develop
Feature
Feature
v0.1
v0.2
v1.0