การจัดการและวิเคราะห์ข้อมูลเพื่อการพัฒนางาน สำนักงานศาลปกครอง

ผศ.ดร. เอกรัฐ รัฐกาญจน์
DADS, NIDA

วันพุธที่ 11 กุมภาพันธ์ 2569

Outline

Time	Topic
9:00 ~ 10:30	1.Data collection
Break 15 mins	
10:45 ~ 12:00	2. Data cleansing
Lunch 1 hour	
13:00 ~ 14:30	3. Data analysis & Data visualization
Break 15 mins	
14:45 ~ 16:00	4. Learning Reflection

The data journey

Stage 1: Data collection

Stage 2: Data cleansing

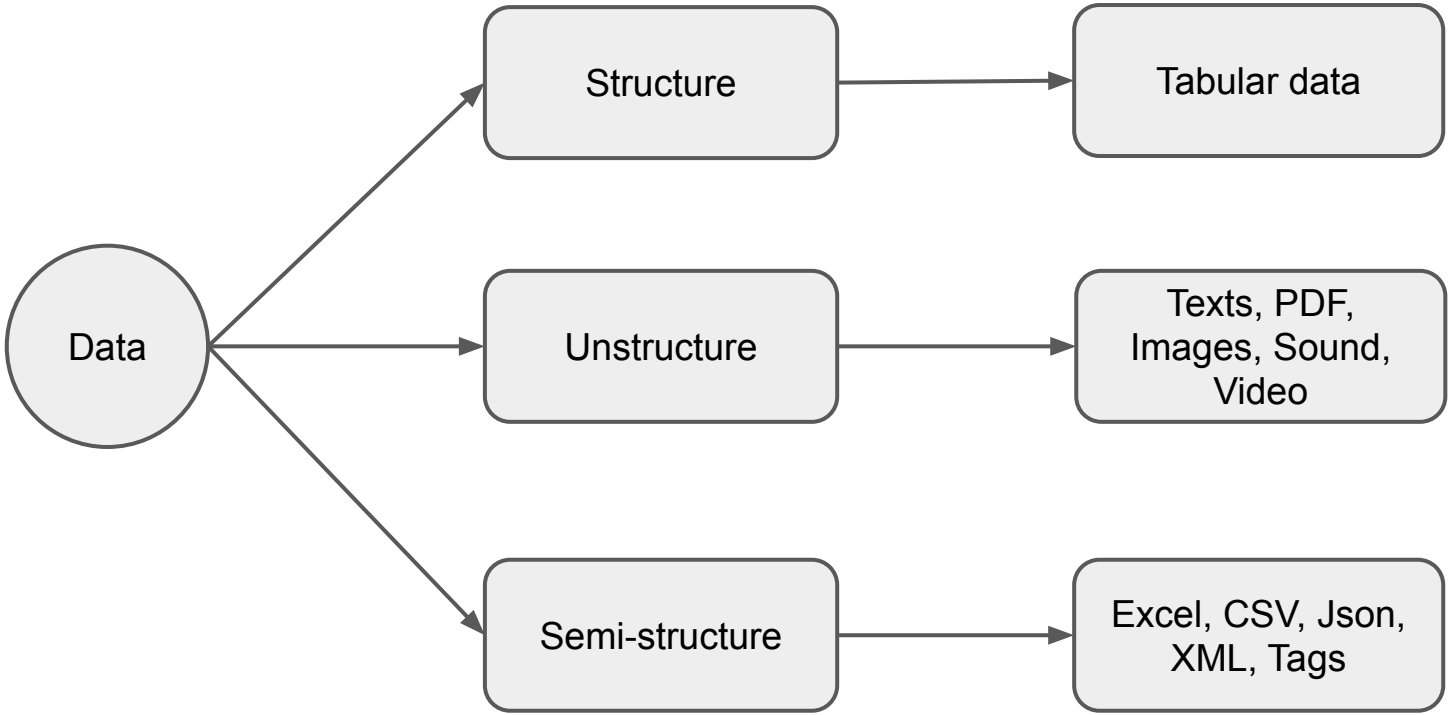
Stage 3: Data analysis

Stage 4: Visualization

Stage 1

Data Collection

What is data?


Data
Structure
Tabular data
Unstructure
Texts, PDF,
Images, Sound,
Video
Semi-structure
Excel, CSV, Json,
XML, Tags

1.1 Structured data

Terminology

- Database: MySQL, PostgreSQL
- Data warehouse: Google BigQuery, Snowflake
- Schema-on-write
- Structured Query Language (SQL)

Use cases for structured data

- **Customer relationship management (CRM)**: CRM software runs structured data through analytical tools to create datasets that reveal customer behavior patterns and trends.
- **Online booking:** Hotel and ticket reservation data (e.g., dates, prices, destinations, etc.) fits the “rows and columns” format indicative of the pre-defined data model.
- **Accounting:** Accounting firms or departments use structured data to process and record financial transactions.

1.2 Unstructured data

Terminology

- Data Storage: Amazon S3, Google Cloud Storage, Hadoop
- Database: MongoDB
- Schema-on-Read
- Not only SQL (NoSQL)

Use cases for unstructured data

- **Data mining**: Enables businesses to use unstructured data to identify consumer behavior, product sentiment, and purchasing patterns to better accommodate their customer base.
- **Predictive data analytics**: Alert businesses of important activity ahead of time so they can properly plan and accordingly adjust to significant market shifts.
- **Chatbots**: Perform text analysis to route customer questions to the appropriate answer sources.

1.3 Semi-structured data

Terminology

- Tools: Excel, Google Sheets, Elasticsearch
- Database: MongoDB, Cassandra

Use cases for semi-structured data

- Social network post and comments
- Searching applications like amazon book stores

Stage 2

Data Cleansing

2. Data cleansing criteria

1. Validity (Data Constraints)

- **Data Types**: Ensure values are in the correct format (e.g., dates are actual dates, not strings; numbers aren't stored as text).
- **Range Constraints**: Check if numbers fall within a logical range (e.g., age shouldn't be negative or 200).
- **Mandatory Fields**: Identify "Null" values in columns that require data (e.g., a "Customer ID" cannot be empty).
- **Unique Constraints**: Check for duplicates in fields that must be unique, like Social Security numbers or transaction IDs.

2. Accuracy and Integrity

- **Cross-Field Validation**: Check if data in one column contradicts another (e.g., if "City" is "Bangkok," the "Country" must be "Thailand").
- **Pattern Matching**: Use Regular Expressions (Regex) to ensure strings like emails, phone numbers, or zip codes follow a valid format.

3. Consistency (The "Uniformity" Check)

- **Standardization**: Check for multiple variations of the same name (e.g., "USA," "United States," and "U.S.A.").
- **Unit of Measure**: Ensure all measurements are consistent (e.g., don't mix "Kilograms" and "Pounds" in the same weight column).
- **Date Formats**: Standardize all dates to a single format, typically ISO-8601 ($YYYY-MM-DD$).

2. Data cleansing

4. Completeness and Outliers

- **Missing Data Analysis**: Decide whether to drop rows with missing data, fill them with a default value, or use "imputation" (estimating the value based on other data).
- **Outlier Detection**: Identify values that are technically "valid" but statistically improbable (e.g., a $1,000,000 grocery bill), which might indicate a data entry error.

5. Uniformity and Structural Cleanup

- **Whitespace**: Check for "hidden" characters like leading/trailing spaces or non-breaking spaces that break search functions.
- **Deduplication**: Identify records that represent the same entity but have slight variations in spelling or address.

Stage 2: Explore, clean, describe

Example of messy data

#	ID	Name	LastName	Birthday	Join date	Gender	Tel	E-mail
1	1111	ธนวัฒน์	เจริญด้วยทรัพย์	1-1-1987	01/01/2019	M	860147805	username @ riccoprint.com
2	1112	วัชสัณห์	เจริญด้วยจิตใจ	07/07/1987	01/01/2019	M	087-0147-805	username2@riccoprint.com
3	1113	พลภัทม์	กำไลอันประเสริฐ	1987/07/07	01/01/2019	M	0810147808	username.riccoprint.com
4	1114	มนวรรธน์		05/07/2530	01/01/2019	F	0820147805	username3@ riccoprint.com
5	1115	ขนมพันธ์	ผูกพันกับหนังสือ	09/12/1988	01/01/2019	A	0830147805	username4@riccoprint.com

Missing values

Formats

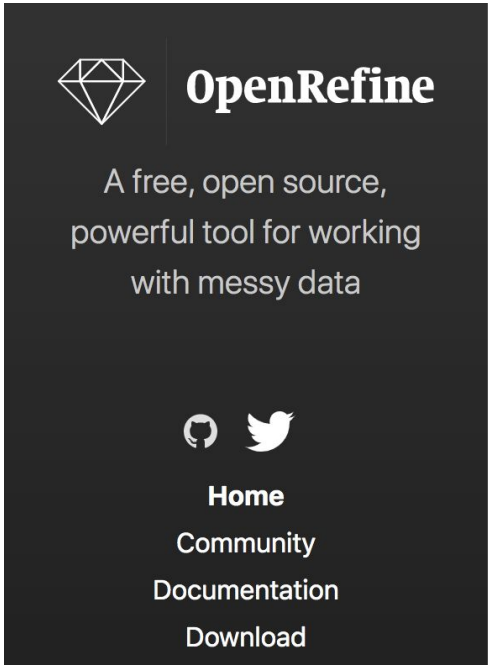
Invalid values

Formats

Formats

https://riccosmartdata.com/data-cleansing-or-data-cleaning/

A powerful tool for this stage



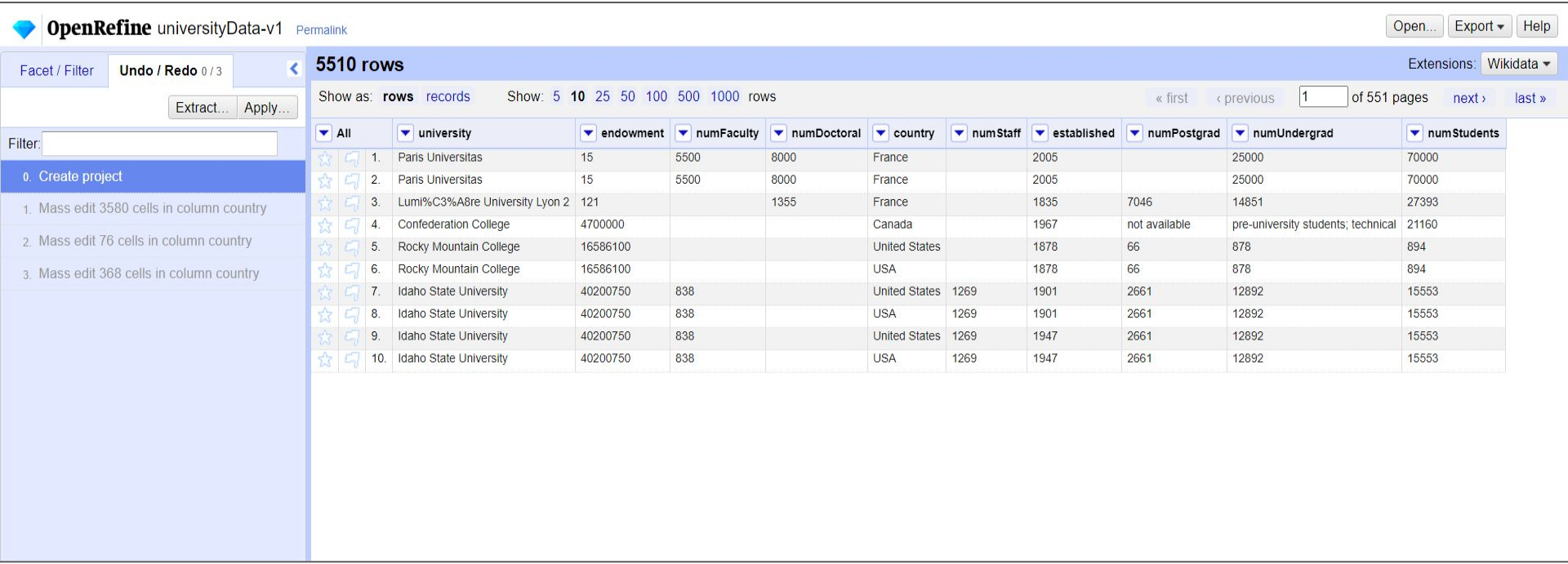
Welcome!

OpenRefine (previously Google Refine) is a powerful tool for working with messy data: cleaning it; transforming it from one format into another; and extending it with web services and external data.

OpenRefine always keeps your data private on your own computer until YOU want to share or collaborate. Your private data never leaves your computer unless you want it to. (It works by running a small server on your computer and you use your web browser to interact with it)

OpenRefine is available in more than 15 languages.

OpenRefine is part of Code for Science & Society.



All		university	endowment	numFaculty	numDoctoral	country	numStaff	established	numPostgrad	numUndergrad	numStudents
1.	Paris Universitas	15	5500	8000	France		2005		25000	70000
2.	Paris Universitas	15	5500	8000	France		2005		25000	70000
3.	Lumi%C3%A8re University Lyon 2	121		1355	France		1835	7046	14851	27393
4.	Confederation College	4700000			Canada		1967	not available	pre-university students; technical	21160
5.	Rocky Mountain College	16586100			United States		1878	66	878	894
6.	Rocky Mountain College	16586100			USA		1878	66	878	894
7.	Idaho State University	40200750	838		United States	1269	1901	2661	12892	15553
8.	Idaho State University	40200750	838		USA	1269	1901	2661	12892	15553
9.	Idaho State University	40200750	838		United States	1269	1947	2661	12892	15553
10.	Idaho State University	40200750	838		USA	1269	1947	2661	12892	15553

Data sources

ข้อมูลศาลปกครอง (data.go.th)

Exercise 1: Cleansing data

Cleansing data with openrefine

Lunch

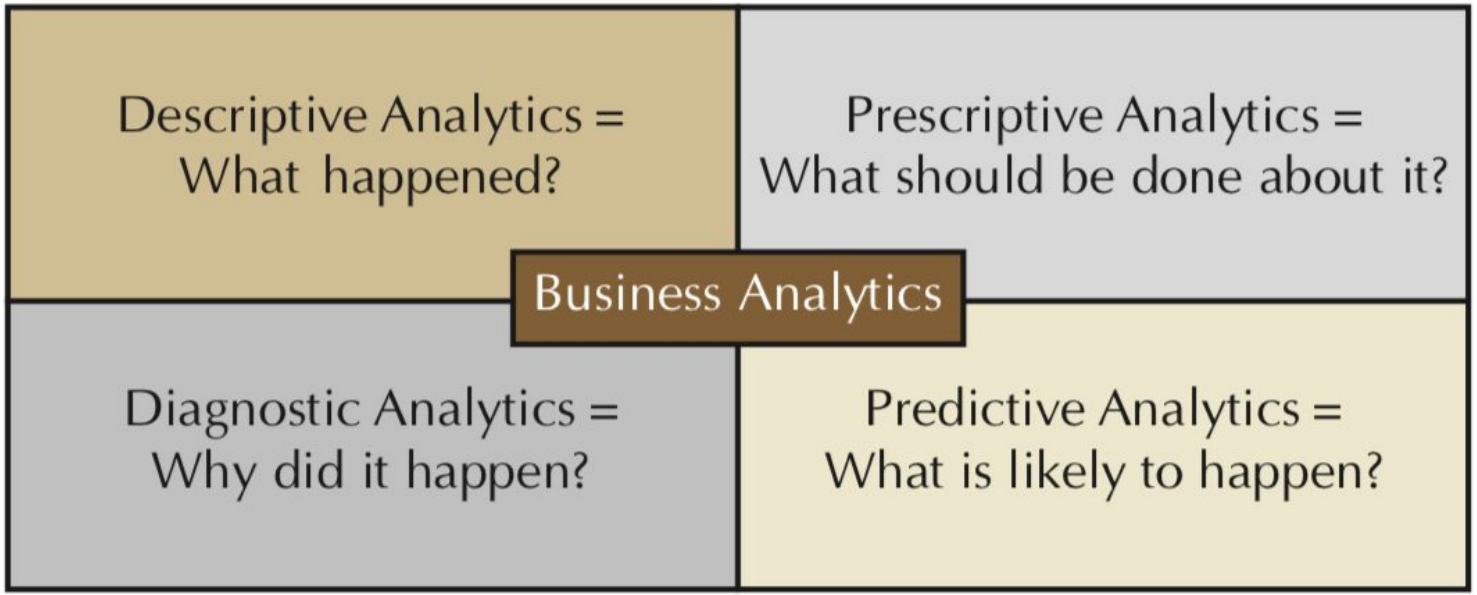
Stage 3

Data analysis

Stage 3: Data analysis

- **To analyse** is to **examine data carefully and in detail** so as to identify causes, key factors, possible results, etc.
- **Analysis** is the process of analysis of data that is done logically aided by sciences (statistical, computers, etc)
- The purpose of doing analysis and modeling is to use statistical techniques to turn the data into information to provide meaningful insights. Analysis and modelling is used to describe a phenomenon, draw conclusions about a population or make predictions about future events.

Four types of data analytics



Banerjee, Arindam, Tathagata Bandyopadhyay, and Prachi Acharya. "Data analytics: Hyped up aspirations or true potential?." *Vikalpa* 38.4 (2013): 1-12.

1. Descriptive analytics

What happened?

- Summarize data into meaningful charts and reports, for example, about budgets, sales, revenues, or cost.
- Typical questions that descriptive analytics help answer are:
    - How much did we sell in each region?
    - What was our revenue and profit last quarter?
    - How many and what types of complaints did we resolve?
    - Which factory has the lowest productivity?
- Example methods

http://www.cbpp.uaa.alaska.edu/afef/business_analytics.htm

2. Diagnostic analytics

What happened?

- Summarize data into meaningful charts and reports, for example, about budgets, sales, revenues, or cost.
- Typical questions that descriptive analytics help answer are:
    - How much did we sell in each region?
    - What was our revenue and profit last quarter?
    - How many and what types of complaints did we resolve?
    - Which factory has the lowest productivity?

http://www.cbpp.uaa.alaska.edu/afef/business_analytics.htm

2. Diagnostic analytics

Commonly used techniques:

- **Correlation analysis**
- **Drill-down analysis**
- **Regression analysis**
- **Hypothesis testing**
- **Root cause analysis (RCA)**
- **Cluster analysis**
- **Time series decomposition**
- **Pareto analysis**

http://www.cbpp.uaa.alaska.edu/afef/business_analytics.htm

Diagnostic analytics

Correlation analysis

Measures the strength and direction of relationships between variables. Example: Analyzing whether higher website traffic correlates with increased sales.

Drill-down analysis

Breaks down aggregated data into detailed layers. Example: A team examines total sales figures and then studies performance by region, store and product for more granular insights.

Regression analysis

Identifies how independent variables influence a dependent variable. Example: Understanding how pricing, advertising spend and seasonality affect revenue.

Diagnostic analytics

Hypothesis testing

Common methods include t-tests and analysis of variance (ANOVA), which are used to test assumptions about data to determine if observed differences are statistically significant. Example: Comparing conversion rates between two versions of a landing page.

Root cause analysis (RCA)

A structured method to identify the fundamental cause of a problem and implement solutions that eliminate inefficiencies and improve overall performance. Example: In supply chain management, RCA identifies inaccurate demand forecasting, prompting corrective actions.

Cluster analysis

Groups data points with similar characteristics to identify patterns or segments. Example: Segmenting customers based on purchasing behavior, demographics or social media engagement to tailor marketing strategies.

Diagnostic analytics

Time series decomposition

Breaks down time-based data into trend, seasonal and residual components, which is useful for identifying irregularities or shifts in performance over time. Example: A retailer separates overall sales growth from holiday seasonal spikes to accurately forecast demand and plan inventory.

Pareto analysis

Based on the 80/20 rule — which states that roughly 80% of effects come from 20% of causes — it identifies the few causes that contribute to most of the effect. Example: Identifying the top 20% of products that generate 80% of returns.

Exercise 2: Data analysis

Data analysis and looker studio

3. Predictive analytics

What is likely to happen?

- Analyze past performance in an effort to predict the future by examining historical data, detecting patterns or relationships in these data, and then extrapolating these relationships forward in time.
- For example,
    - a marketer might wish to predict the response of different customer segments to an advertising campaign.
    - a commodities trader might wish to predict short-term movements in commodities prices, or a skiwear manufacturer might want to predict next season's demand for skiwear of a specific color and size.
    - a bank manager might want to identify the most profitable customers or predict the chances that a loan applicant will default, or alert a credit card customer to a potential fraudulent charge.

4. Prescriptive analytics

What should be done about it?

- Prescriptive analytics uses optimization to identify the best alternatives to minimize or maximize some objective. Prescriptive analytics is used in many areas of business, including operations, marketing, and finance. For example, we may determine the best pricing and advertising strategy to maximize revenue, the optimal amount of cash to store in ATMs, or the best mix of investments in a retirement portfolio to manage risk.
- Prescriptive analytics addresses questions like:
    - How much should we produce to maximize profit?
    - What is the best way of shipping goods from our factories to minimize costs?
    - Should we change our plans if a natural disaster closes a supplier's factory and if so, by how much?

Stage 4

Visualization

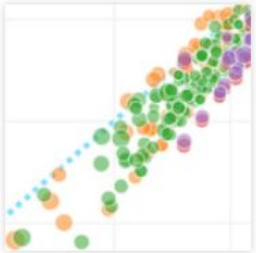
Tell the story

The statistical information that comes from analysis and modeling is easier to digest if it is presented in some sort of story. It could be a research paper, an infographic, an article for the media, or some combination of these and other data presentation methods.

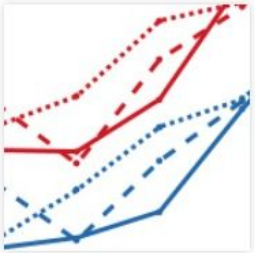
The following competencies apply to this step

- **data interpretation**
- **data visualization**
- **storytelling**.

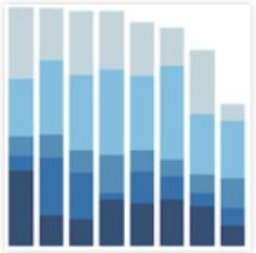
Basic chart



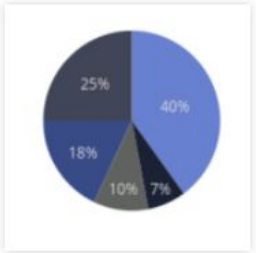
Scatter Plots



Line Charts

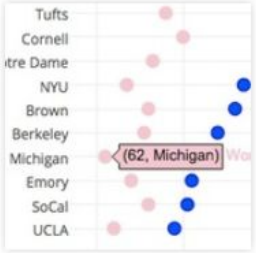


Bar Charts



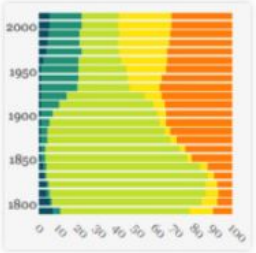
Pie Charts

Bubble Charts



Dot Plots

Filled Area Plots

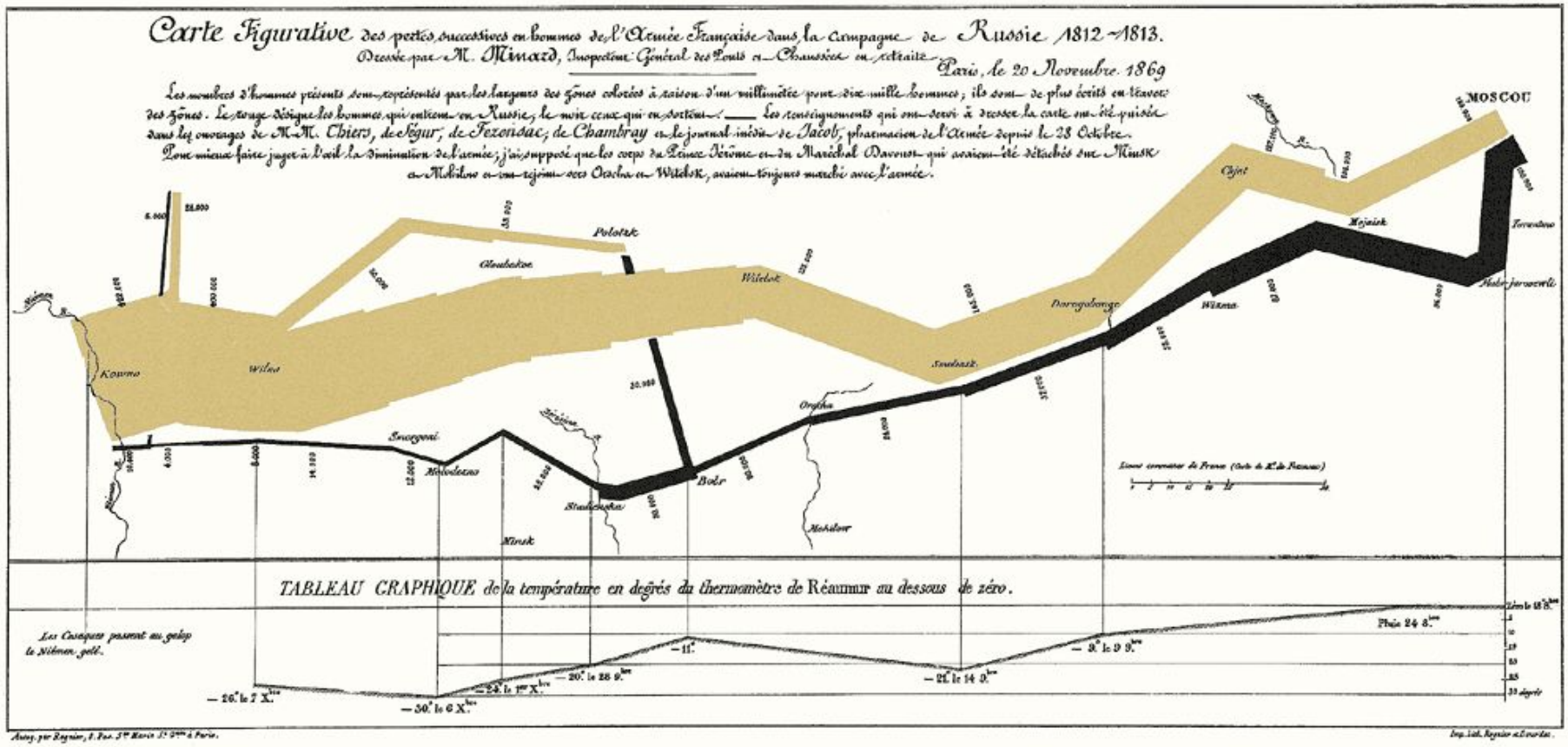


Horizontal Bar Charts

Gantt Charts

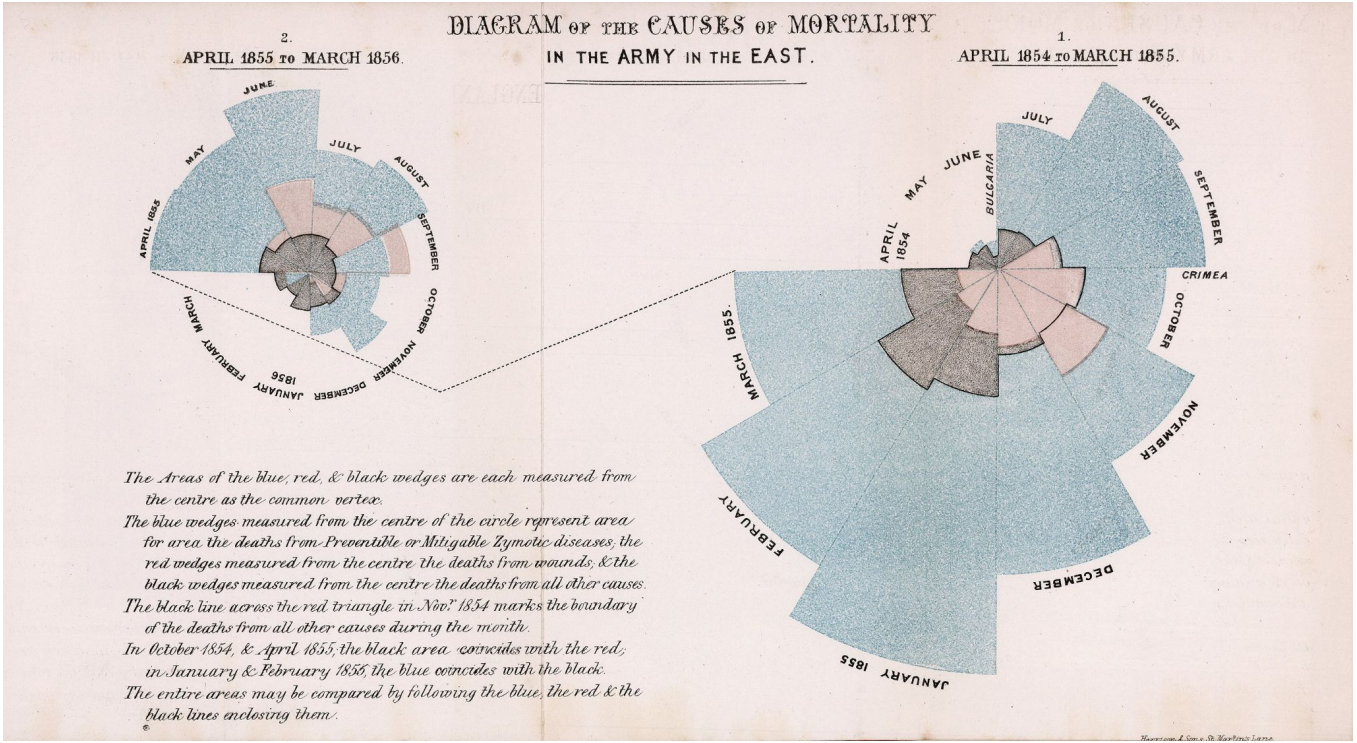
Sunburst Charts

https://plotly.com/python/basic-charts/



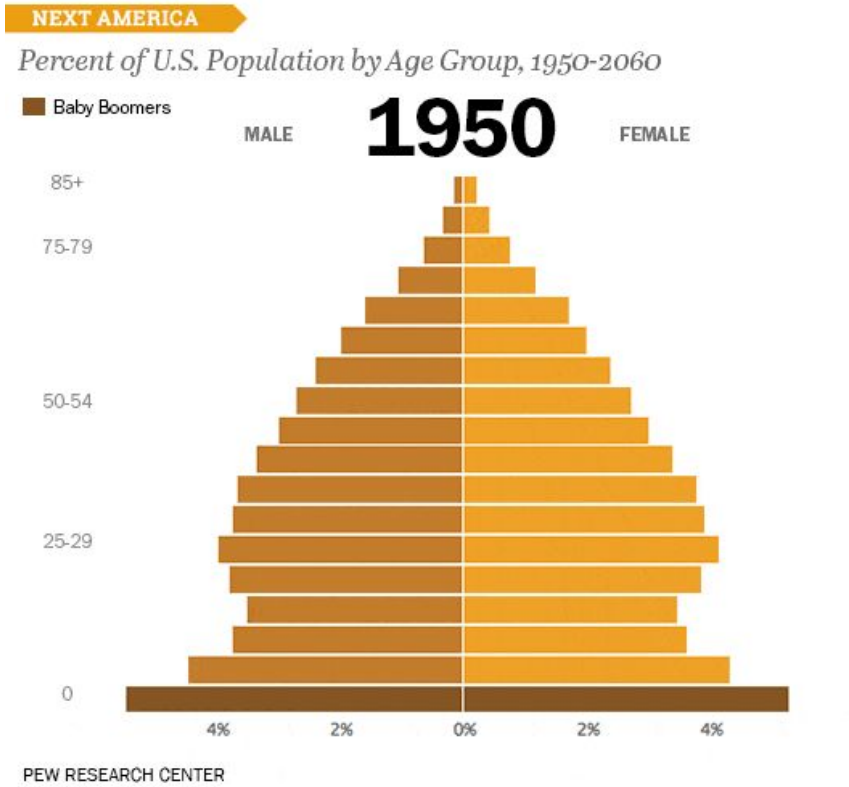
https://www.edwardtufte.com/tufte/minard

Nurse, analyst, and data rockstar Florence Nightingale used this beautiful data visualization to reveal that the majority of deaths were actually caused by poor hospital practices.



https://www.tableau.com/learn/articles/best-beautiful-data-visualization-examples

Example 3



https://www.tableau.com/learn/articles/best-beautiful-data-visualization-examples

Visualization references

- https://datavizcatalogue.com/#google_vignette

Tool: Miro

https://www.it24hrs.com/2020/what-is-miro-realtime-brainstorming/