IS422P - DATA MINING
DATA MINING TRENDS

AMIRA REZK

INFORMATION SYSTEM DEPARTMENT

AGENDA

Data mining via different approaches

- BUSINESS INTELLIGENCE
- BIG DATA
- MACHINE LEARNING
- WEB MINING
- DATA WAREHOUSING
- DATA ANALYSIS
- NATURAL LANGUAGE PROCESSING
- DATA VISUALIZATION

DATA MINING VS. BUSINESS INTELLIGENCE

- Business intelligence (BI) refers to the procedural and technical infrastructure that collects, stores, and analyzes the data produced by a company's activities.
- BI is a broad term that encompasses data mining, process analysis, performance benchmarking, and descriptive analytics.
- BI parses all the data generated by a business and presents easy-to-digest reports, performance measures, and trends that inform management decisions.

DATA MINING VS. BIG DATA

- **Big Data** refers to a huge volume of data that can be structured, semi-structured and unstructured. It comprises of 5 Vs i.e.
- • ***Volume***: It refers to an amount of data or size of data that can be in quintillion when comes to big data.
- • ***Variety***: It refers to different types of data like social media, web server logs etc.
- • ***Velocity***: It refers to how fast data is growing, data is exponentially growing and at a very fast rate.
- • ***Veracity***: It refers to an uncertainty of data like social media means if the data can be trusted or not.
- • ***Value***: It refers to the data which we are storing and processing is worth and how we are getting benefit from this huge amount of data.

DATA MINING VS. MACHINE LEARNING

- Both data mining and machine learning are rooted in data science and generally fall under that umbrella.
- They often intersect or are confused with each other, but there is a key distinctions between the two.
  - Machine learning can look at patterns and learn from them to **adapt** behavior for future incidents,
  - while data mining doesn't learn and apply knowledge on its own without human interaction. Data mining also can't automatically see the relationship between existing pieces of data with the same depth that machine learning can.

DATA MINING VS. WEB MINING

- Web mining comes under data mining but this is limited to web related data and identifying the patterns.
- It is the process of performing data mining on the web, extracting the web documents and discovering the patterns from it.
- **Web content mining** → discover different patterns that give a significant insight
- **Web Structure mining**→ Data from hyperlinks that lead to different pages are gathered and prepared in order to discover a pattern.
- **Web usage mining**→ user's web activity through the application logs are monitored and data mining is applied to it.

DATA MINING VS. DATA WAREHOUSING

- A Data Warehouse is an environment where essential data from multiple sources is stored under a single schema. It is then used for reporting and analysis.
  - Historical Data
  - Generalization
  - ETL
  - OLAP

DATA MINING VS. DATA ANALYSIS

- Data Mining studies are mostly on structured data. Data Analysis can be done on both structured, semi-structured or unstructured data.

- Data mining generally doesn’t involve visualization tool, Data Analysis is always accompanied by visualization of results.

TEXT MINING VS. NATURAL LANGUAGE PROCESSING (NLP)

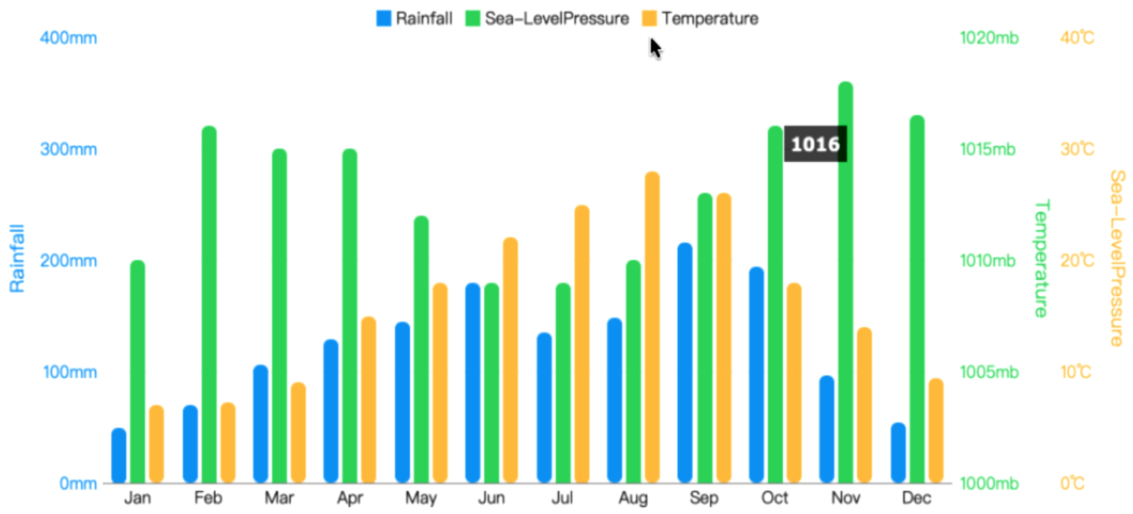
- Both Text Mining and Natural Language Processing trying to extract information from unstructured data.
- Text mining is concentrated on text documents and mostly depends on a statistical and probabilistic model to derive a representation of documents. Popular applications of Text Mining : Contextual Advertising, Content enrichment, Social media data analysis, Spam filtering, and Fraud detection through claims investigation.
- Natural language is what we use for communication. Techniques for processing such data to understand underlying meaning is collectively called as **Natural Language Processing (NLP).** The data could be speech, text or even an image and approach involve applying Machine Learning (ML) techniques on data to build applications involving classification, extracting structure, summarizing and translating data.

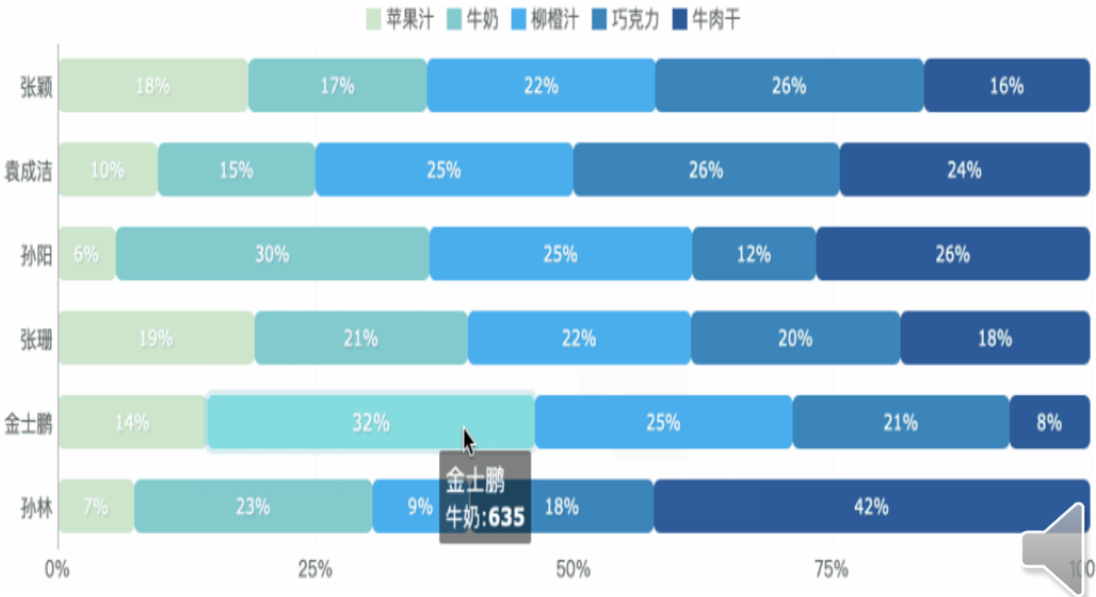
DATA MINING VS. DATA VISUALIZATION

- **Data Visualization** is the process of extracting and visualizing the data in a very clear and understandable way without any form of reading or writing by displaying the results in the form of pie charts, bar graphs, statistical representation and through graphical forms as well.
- In Data Visualization, the primary goal is to convey the information efficiently and clearly without any deviations or complexities in the form of statistical graphs, information graphs, and plots.

DATA VISUALIZATION TYPES

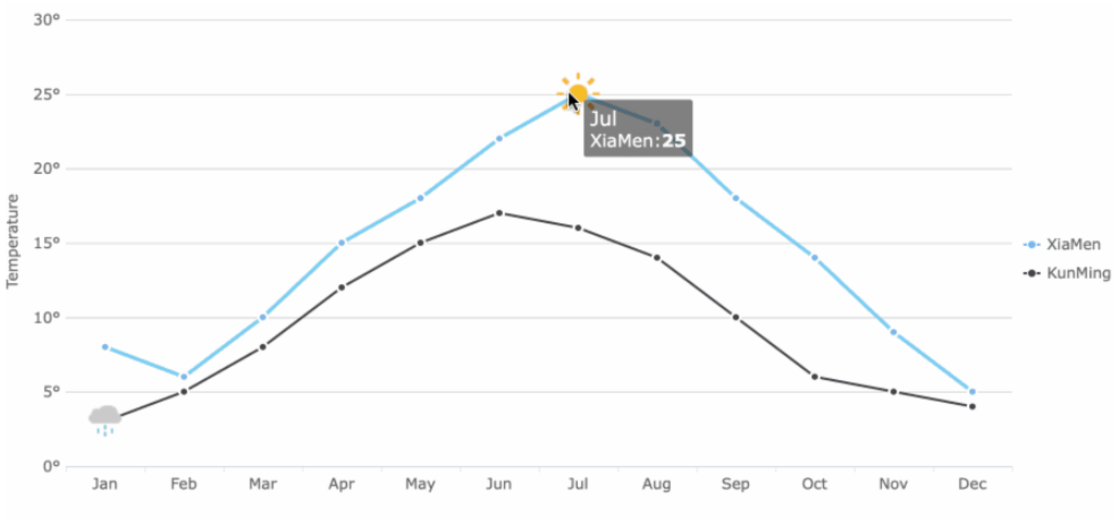
- **Column Chart →** show numerical comparisons between categories, No. columns should not be too large
- **Bar Chart →** he number of bars can be relatively large.

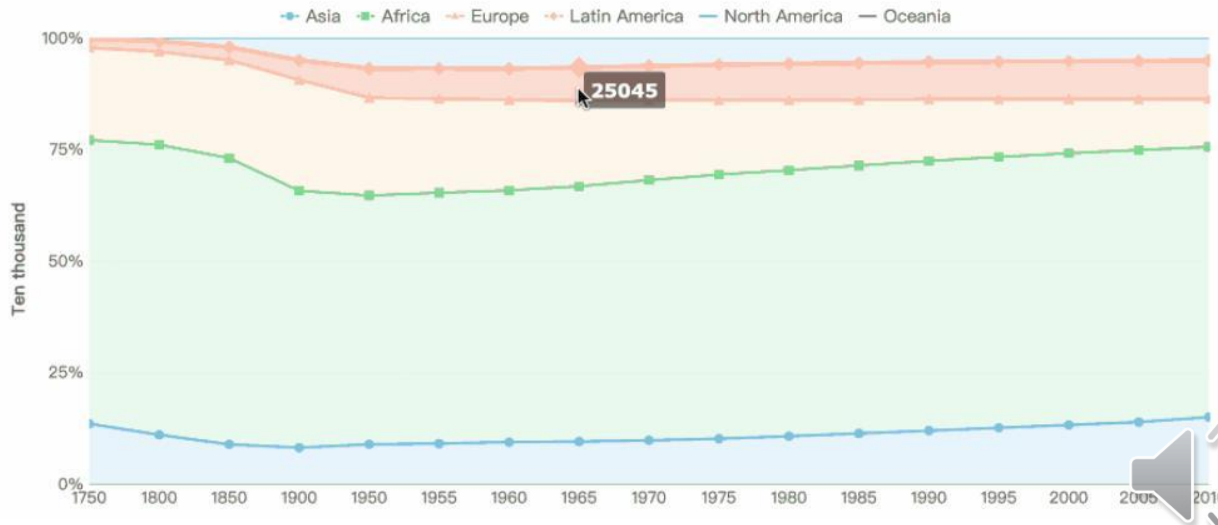




DATA VISUALIZATION TYPES

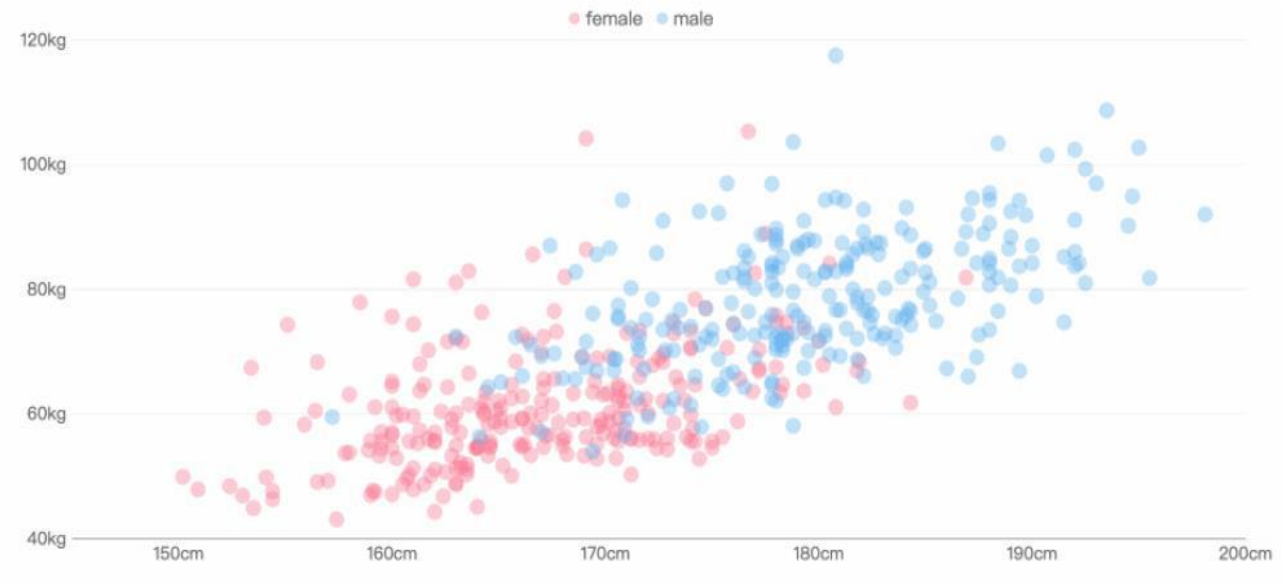
- **Line Chart→** show the change of data over a continuous time interval or time span.
- **Area Chart →** The filling of the color can better highlight the trend information.

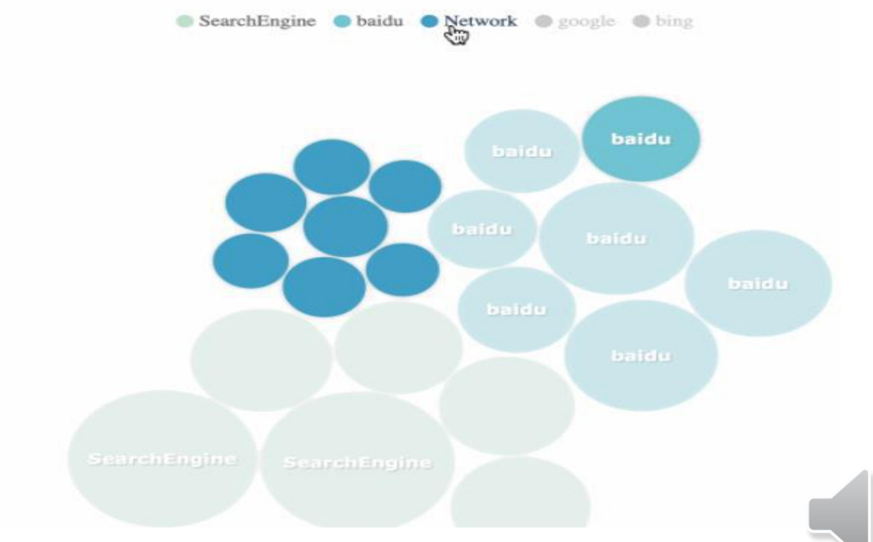




DATA VISUALIZATION TYPES

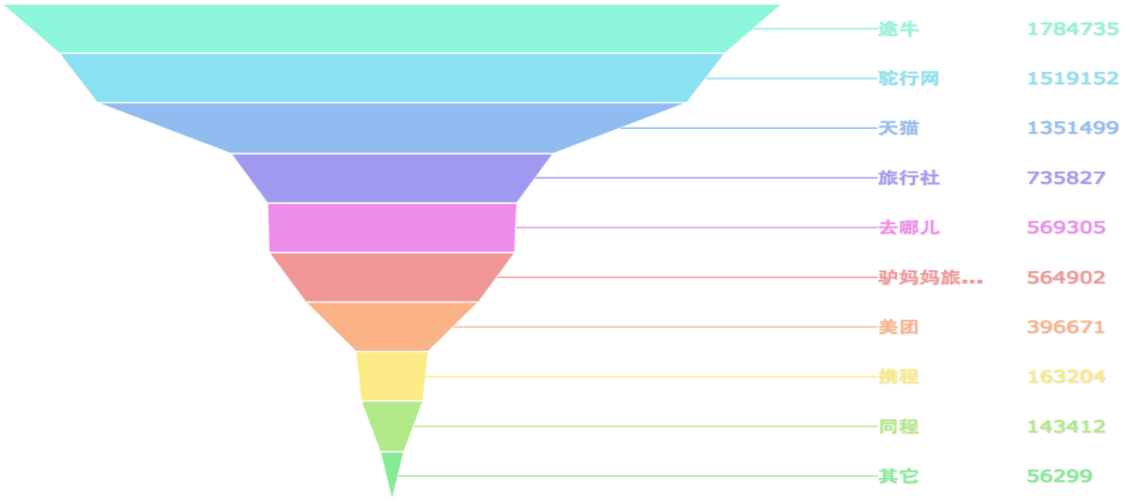
- **Scatter Plot:** shows two variables in the form of points on a rectangular coordinate system. The position of the point is determined by the value of the variable.
- **Bubble Chart →** is a multivariate chart that is a variant of a scatter plot. Except for the values of the variables represented by the X and Y axes, the area of each bubble represents the third value.

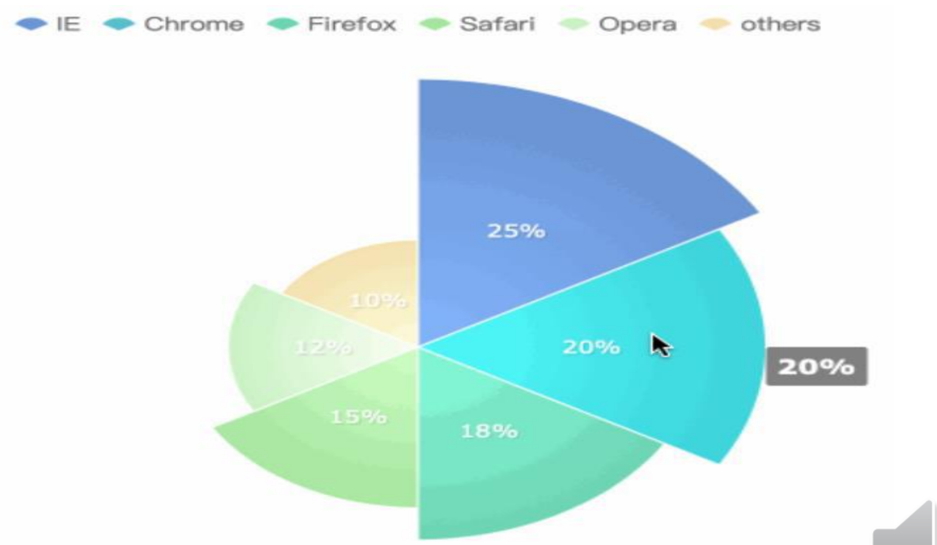




DATA VISUALIZATION TYPES

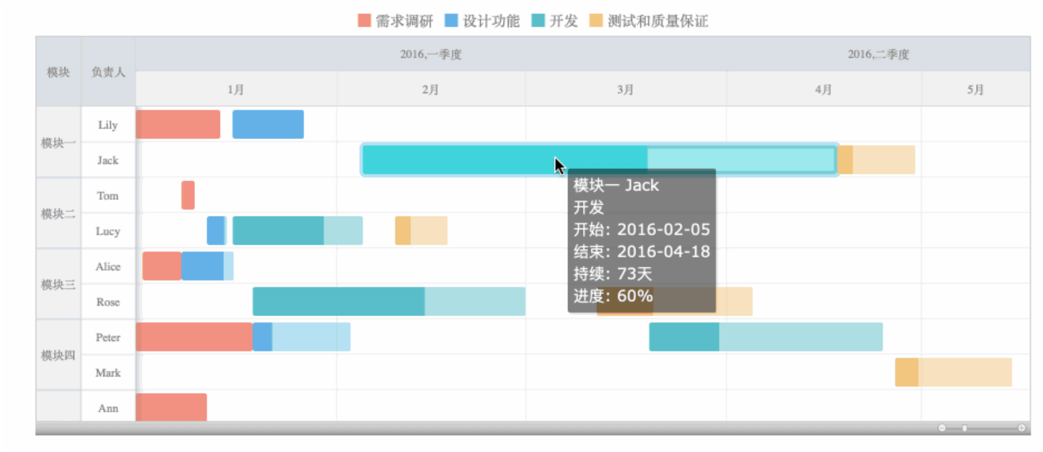
- **Funnel Chart:** display a series of steps and the completion rate for each step. used to represent how something moves through different stages in a process. A funnel chart displays values as progressively decreasing proportions amounting to 100 percent in total.
- **Pie Chart** → represents one static number, divided into categories that constitute its individual portions.

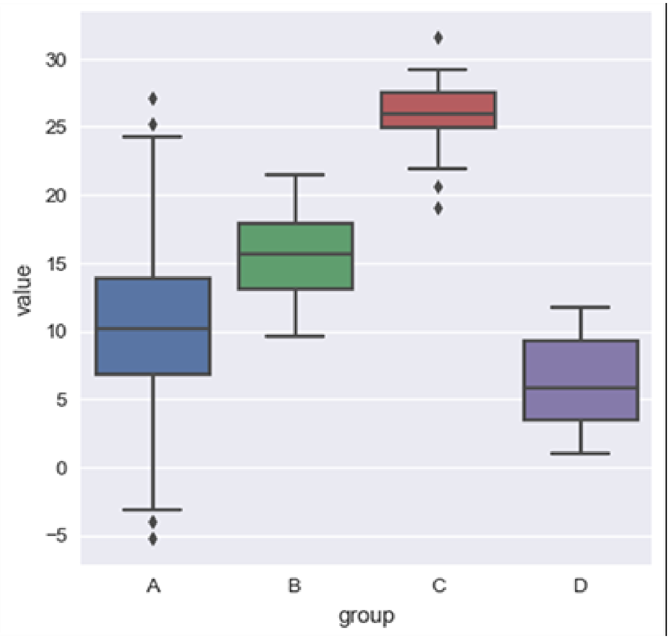




DATA VISUALIZATION TYPES

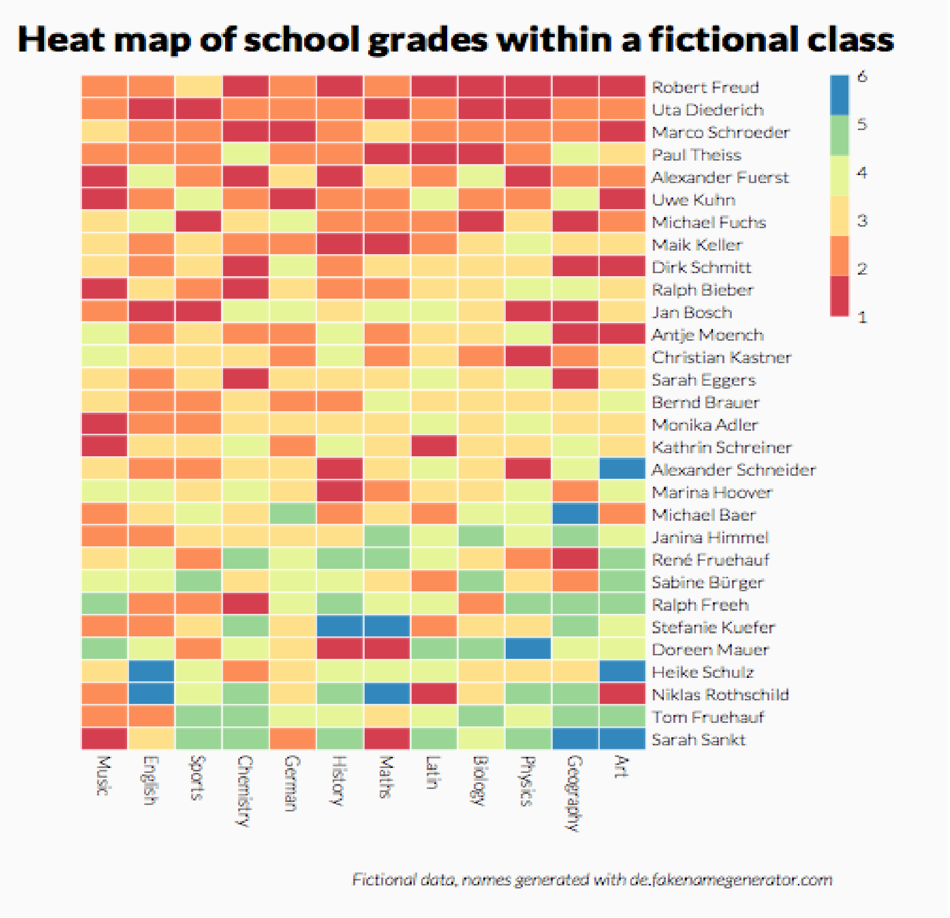
- **Gantt Chart →** shows the timing of the mission
- **Box Plot→** is a visual representation of displaying a distribution of data, usually across groups, based on a five number summary





DATA VISUALIZATION TYPES

- **Heatmap (**choropleth map)
- shows the relationship between two measures and provides rating information. The rating information is displayed using varying colors or saturation and can exhibit ratings such as high to low or bad to awesome, and needs improvement to working well.



Heat map of school grades within a fictional class

Fictional data, names generated with de.fakenamegenerator.com

QUESTION?