Jenis Data Riset dan Teknik Visulisasi Data Science

SEKOLAHMU

William Yohanes Sirait, S.Si

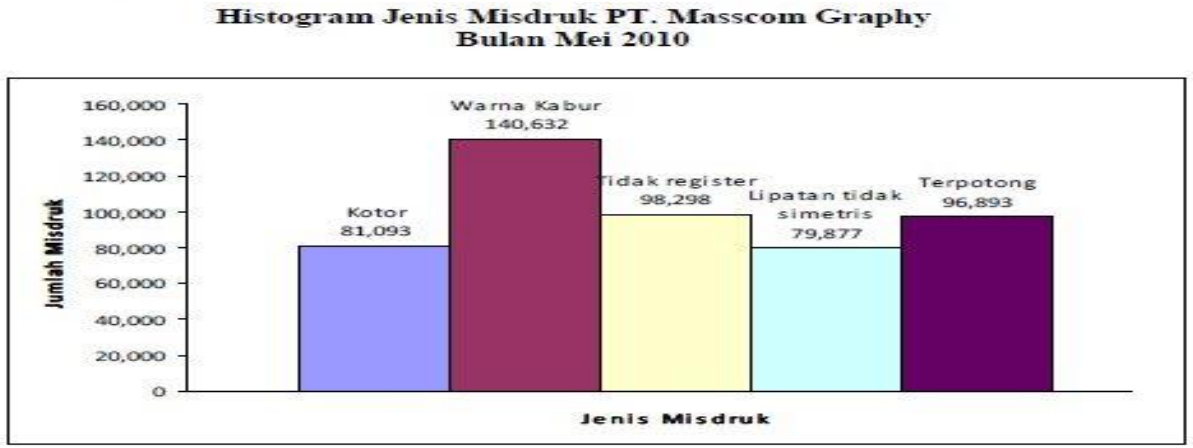
Timeline

- Jenis-jenis Data Riset
- Kaidah-Kaidah pada Grafik yang Benar
- Teknik Storytelling pada Data Science
- Visulisasi Data Science dengan Tableau

Tipe Data Riset

1. Data Numerik (Numerical) – Interval & Ratio
2. Data Kategori (Nominal)
3. Data Ranking (Ordinal)

1. Data Numerik adalah Data yang direpresentasikan dalam sebuah data. Karakteristik dari angka itu menyatakan sifat dari angka itu sendiri. Maksud dari kuantitatif adalah jumlah, kadar, kapasitas yang bisa diukur dengan jelas. Contoh : air dalam gelas yang bias diiukur volumenya, berat beras, dalam karung, dll



Gambar 1.1 Contoh Data Numerik

Jenis Analisa Statistik

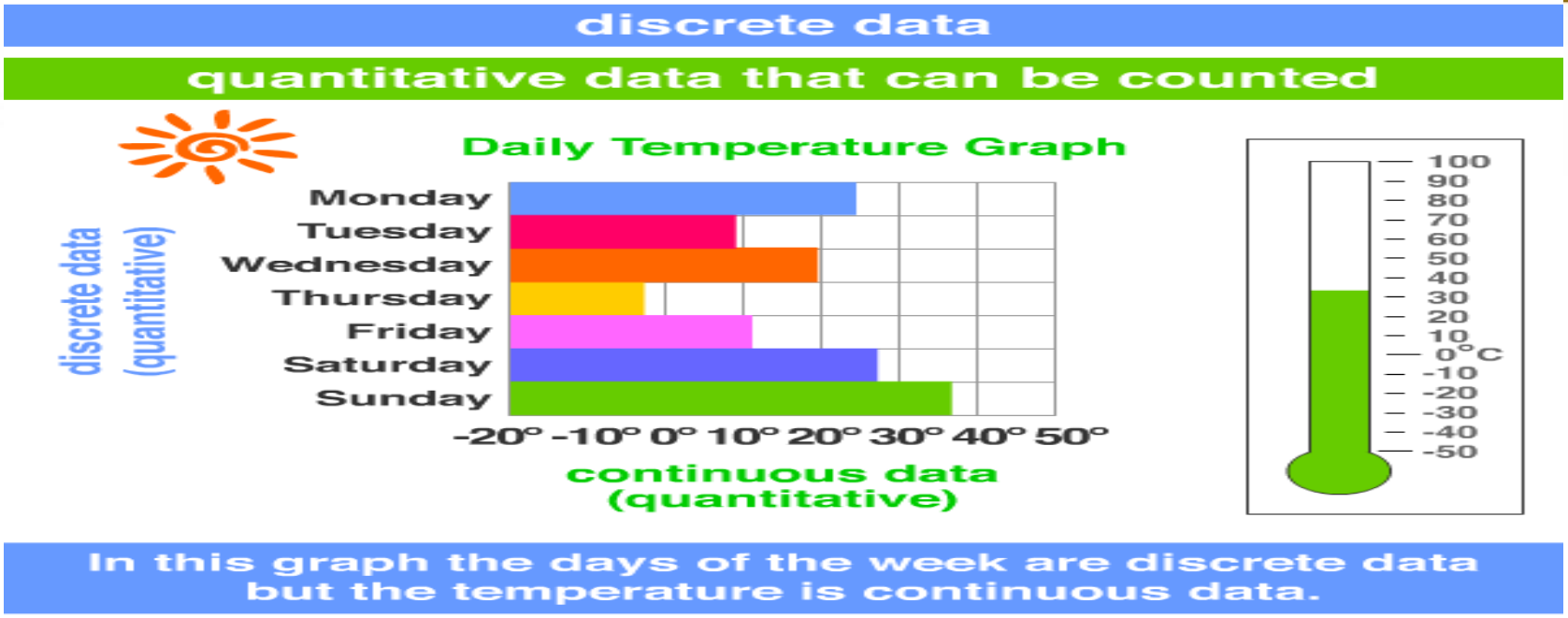
Tipe Data Numerik dibagi menjadi 2 jenis yaitu :

a. Data Diskret (discrete)

b. Dat Kontinu (continuous)

Data diskret adalah data numerik (angka) yang tidak memiliki koma. Sering disebut dalam attribute data. Contoh data diskret adalah jumlah pembelian yang dilakukan pelanggan dalam satu tahun [20,30, atau pembelian/tahun], frekuensi kita mengedipkan mata dalam satu menit [10 , 40,100, kali/menit,]

Data kontinu adalah tipe data yang memiliki nilai tak terbatas (dalam rentang tertentu) dan bisa disajikan dalam bentuk koma. Contoh : Tinggin badan seseorang [170,4 cm, 165,67 cm], berat badan seseorang [67,9 , 78,88 kg]



Gambar 1.2 Contoh Data Diskret dan Data kontinu

Data Interval adalah data numerik yang karakteristiknya tidak dinyatakan dalam decimal. Data interval tidak memiliki nilai 0 dalam makna matematis yang sesungguhnya (disebut true zero). Contoh pengukuran suhu menggunaka termometer, makna suhu 0 derajat bukan berarti bernilai kosong namun hanya menyatakan nilai kadar saja.

Data rasio adalah data numerik yang tidak memiliki nilai minus. Contoh luas tanah [dukur m^2, hektar, dll], kadar dosis dalam obat [diukur dalam mg, dan satuan lainnya], response time[kecepatan respons], (dikur dalam ms, s, dan jam)

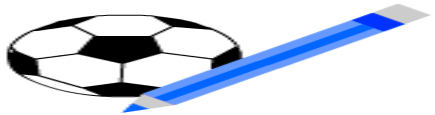
2. Data Kategori (nominal)

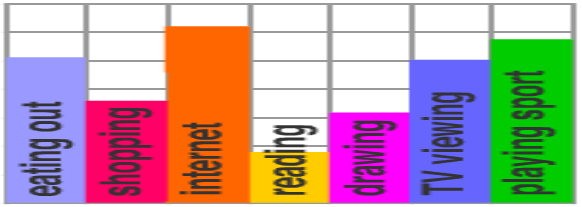
Data yang menunjuk sifat kualitatif daripada kuantitatif. Tipe data kategori merupakan bagian dari beberapa kelompok. Misalnya : Besar/kecil, laki-laki/perempuan, benar/salah, ringan/berat, dll. Contoh data kategori : Data biner (binary) : data dengan 2 kategori, misalnya cacat/tidak cacat, yes/no, baik/buruk, dll. Data warna rambut : hitam, cokelat, pirang, merah, dst. Kita dapat memberi keterangan pada data kategori dengan angka, namun tidak menyatakan kualitas data tersebut conto data rambut (hitam = 1, cokelat = 2, pirang = 3), bukan berarti rambut hitam itu lebih baik daripada rambut pirang karena peringkatnya “1”

categorical data

also known as qualitative data

Leisure Activities





categorical data

data categories which may include things like skills, preferences, homes, schools, food and hobbies.

Favourite Food Groups

categorical data

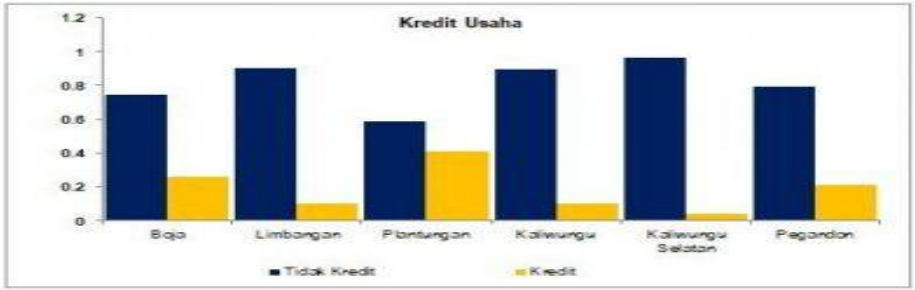
Gambar 1.3 Contoh data kategori

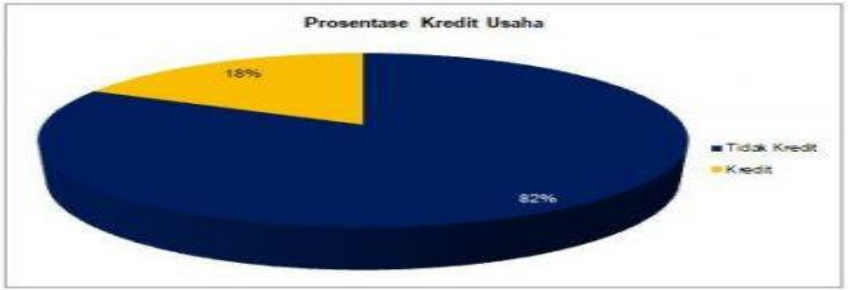
Data Ranking (ordinal) adalah campuran antara data numerik dan data kategori. Angka yang dimiliki oleh data ordinal memiliki nilai arti matematis. Nilai 2 lebih baik dari 1, dan nilai 5 memang lebih baik dari nilai 3, dll. Contoh data ordinal antara lain :
Pemberian Rating saat kita memberikan data keuisoner contoh rentang antara 1-5, 1-10.
Pemberian rating pada aplikasi toko online, 1(buruk), 5(sangat baik)

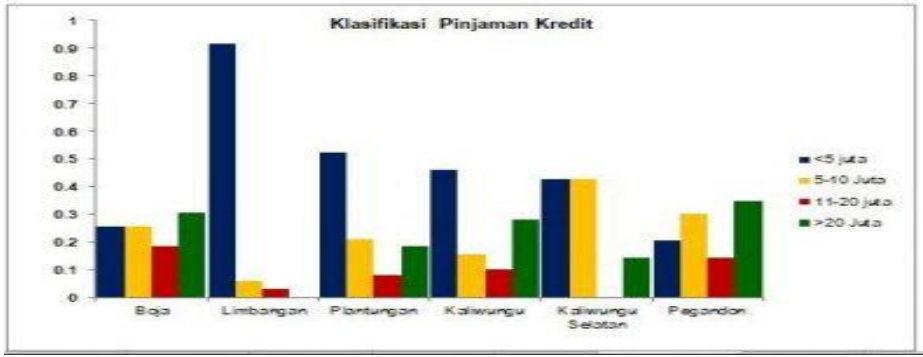
Gambar 1.4. contoh Data ordinal

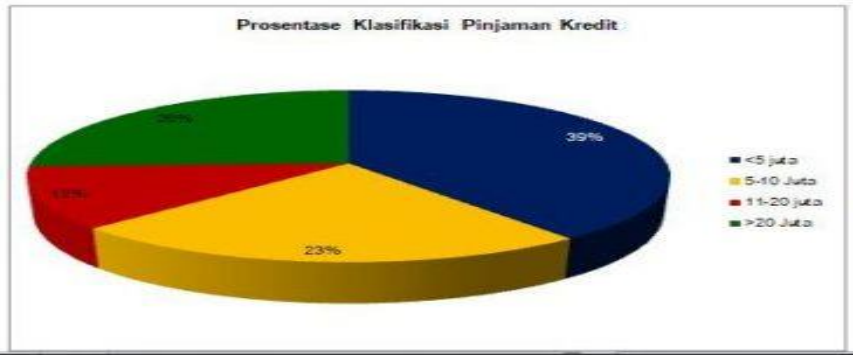
Kaidah Pembuatan Grafik

1. Memiliki judul grafik 2. Memiliki dua variabel (variable bebas dan variable terikat) 3. Cocok dengan style grafiknya. Contoh grafik pie untuk data kategori 4. Multiwarna 5. Terdapat satuannya (kecepatan (m/s))









Gambar 1.5 Contoh grafik yang baik

Jenis Pembuatan Grafik

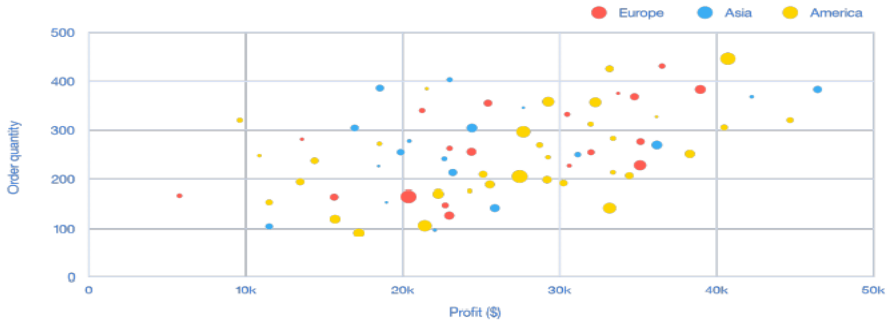
Scatter plot

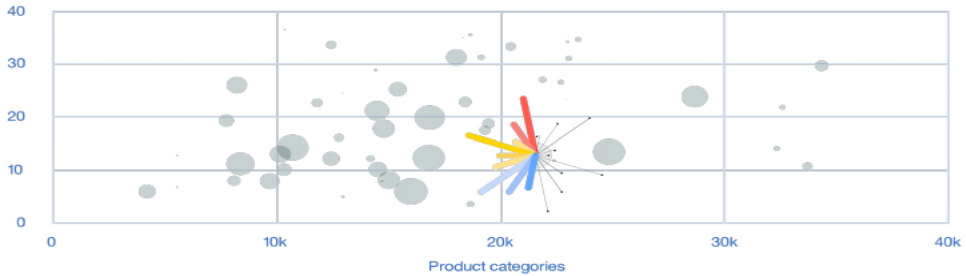
A graph of plotted points that show the relationship between two sets of data.

I'm going to use this model when I want to:
explore in time | compare | show correlations

I'm going to use this model when I have this kind of data set:
categorized data | multi-dimension data

Not recommended for:
Better not to use it in case of too small data set.





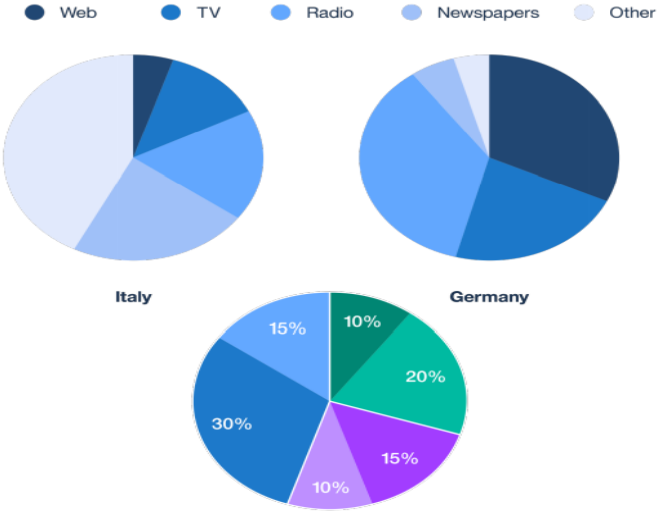
Pie chart

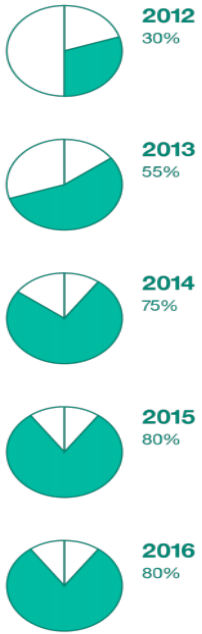
Circular graph model divided into sectors, illustrating proportions.

I'm going to use this model when I want to:
compare | show subdivisions

I'm going to use this model when I have this kind of data set:
categorized data

Not recommended for:
Don't use when you have more than six categories.





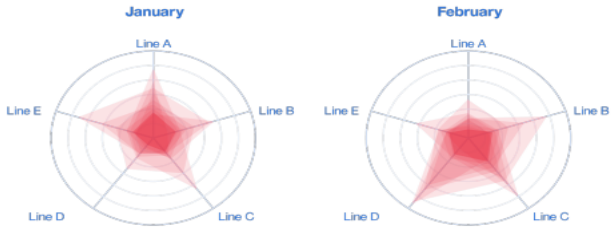
Radar

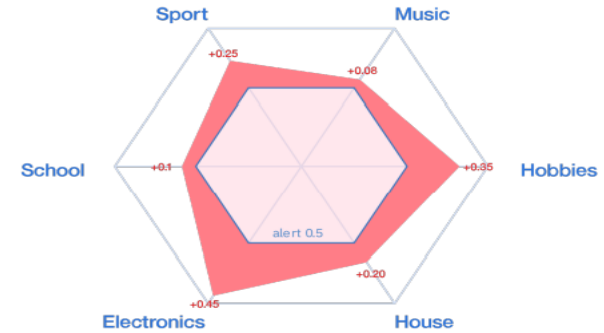
Chart used to represent values of multiple indicators simultaneously.

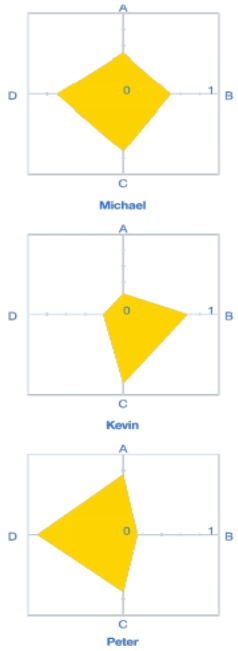
I'm going to use this model when I want to:
show correlations | compare

I'm going to use this model when I have this kind of data set:
categorized data

Not recommended for:
When doing a time comparison, radial representations are not the best to compare lengths.







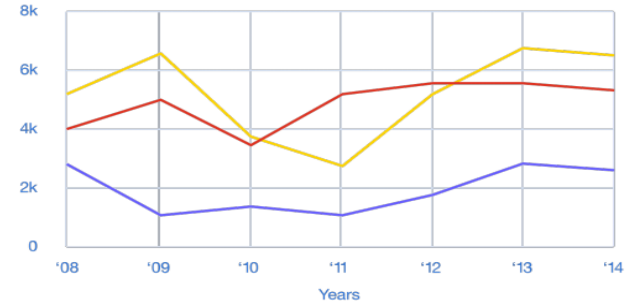
Line chart

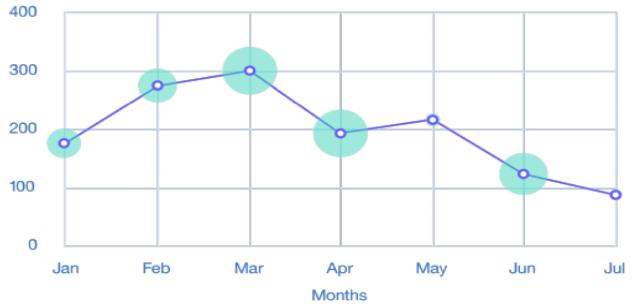
This graph model displays information as a series of data points connected by straight line segments.

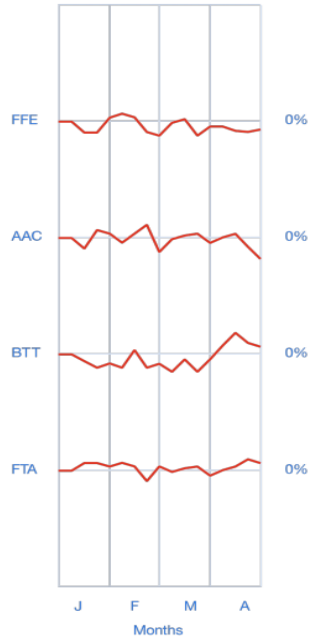
I'm going to use this model when I want to:
explore in time | compare | show correlations

I'm going to use this model when I have this kind of data set:
time-based data

Not recommended for:
Avoid if not comparing values over time, as it might create confusion. Select a bar graph in this case







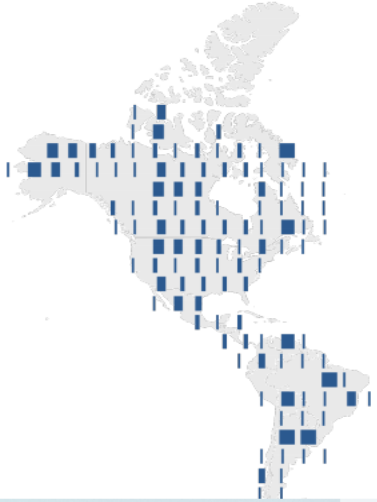
Map

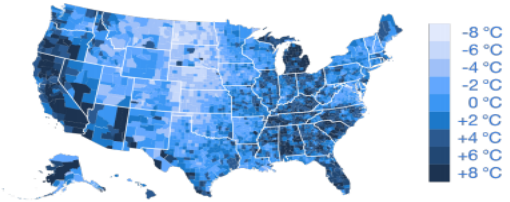
Cartography is used to display geographical data.

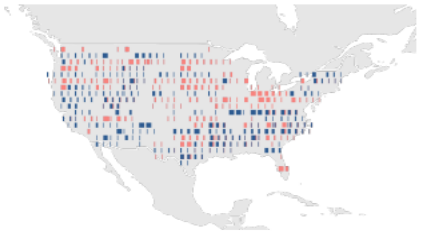
I'm going to use this model when I want to:
explore in time | compare | distribute geographically

I'm going to use this model when I have this kind of data set:
geographic distribution

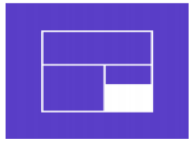
Not recommended for:
Don't use it if the data set has geographical data that's not relevant to your use case.







Treemap

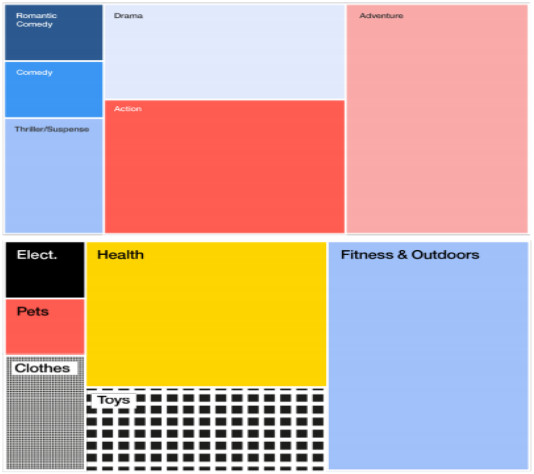


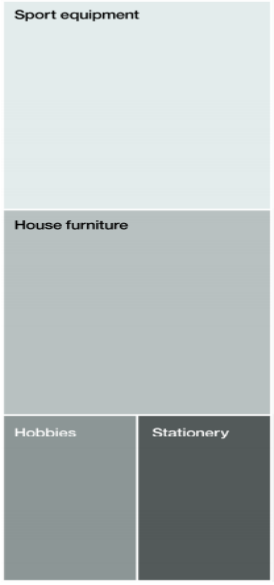
Displays hierarchical data as a set of nested rectangles, which parts combined, make a larger rectangle.

I'm going to use this model when I want to:
compare | show subdivisions

I'm going to use this model when I have this kind of data set:
categorized data | geographic distribution

Not recommended for:
Don't use a tree map for data grouped in more than 25 different categories.





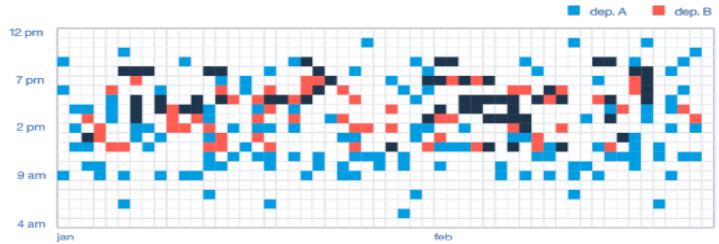
Heat map

Represents mutual correlations of variables within a data set.

I'm going to use this model when I want to:
show correlations | show relationships

I'm going to use this model when I have this kind of data set:
multi-dimensional data

Not recommended for:
One of the main strengths of a heat map is its ability to highlight patterns. Don't use it when you have only a few indicators.



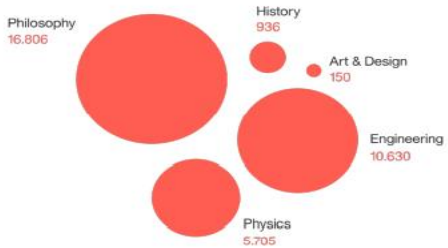
Bubble chart

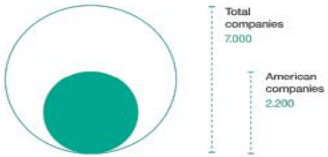
Model used to show values among categories or groups with circles, avoiding any kind of axis.

I'm going to use this model when I want to:
compare | show subdivisions

I'm going to use this model when I have this kind of data set:
categorized data

Not recommended for:
When you have too similar values, where the circle's area makes it difficult to read.





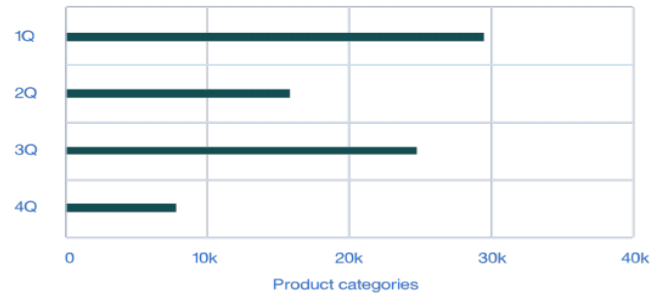
Bar chart

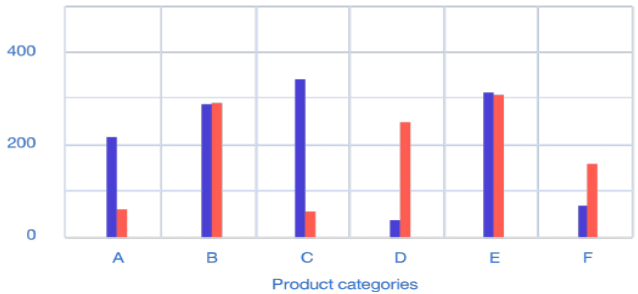
Rectangular bars with lengths proportional to the values they represent.

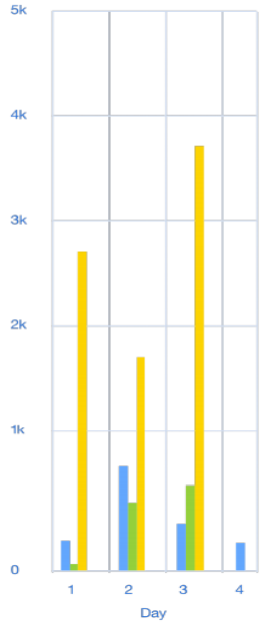
I'm going to use this model when I want to:
explore in time | compare | show correlations

I'm going to use this model when I have this kind of data set:
time-based data | categorized data

Not recommended for:
Never use to compare values with different units or hierarchy.







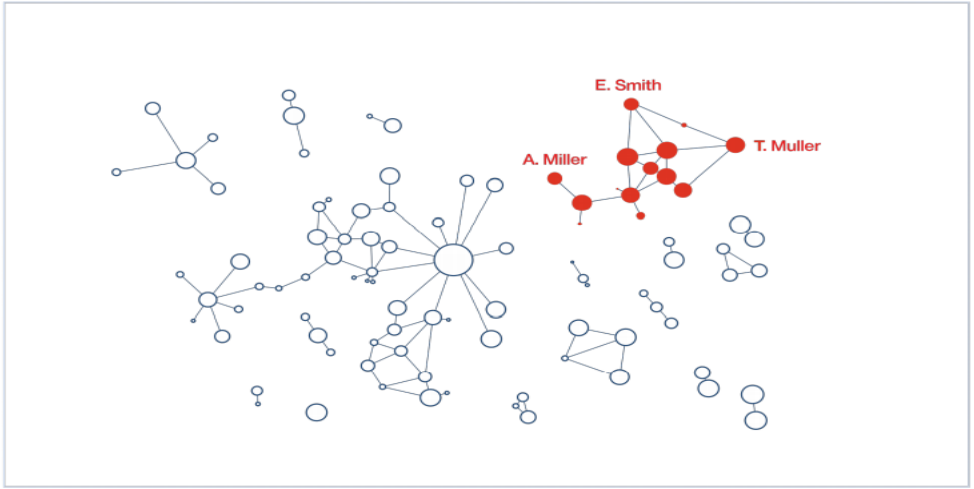
Network

A graph where nodes are connected and positioned depending on their mutual relationship.

I'm going to use this model when I want to:
show relationships

I'm going to use this model when I have this kind of data set:
multi-dimension data

Not recommended for:
Hard for beginners and common users to understand, better for experts.



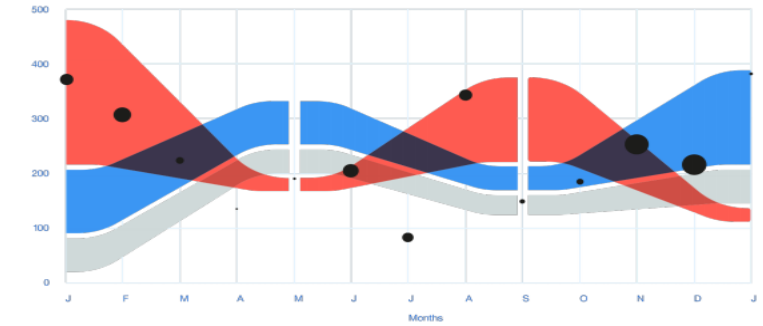
Flows

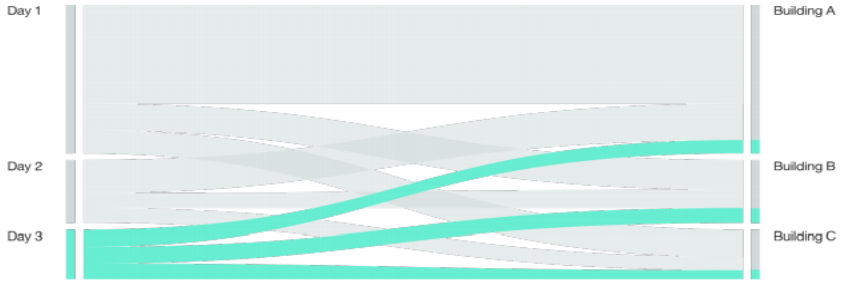
Chart used to show different behaviors among multiple steps and situations.

I'm going to use this model when I want to:
show relationships | show subdivisions

I'm going to use this model when I have this kind of data set:
categorized data

Not recommended for:
A large amount of categories and flows, as it reduces readability.





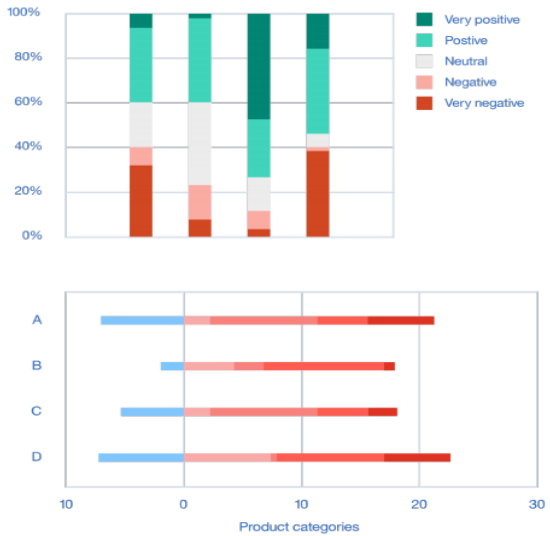
Stacked bar chart

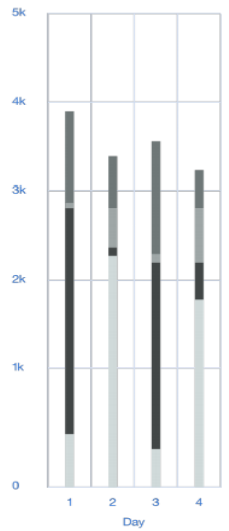
A variant of the bar graph, where each rectangle is divided in multiple parts.

I'm going to use this model when I want to:
explore in time | compare | show correlations | show subdivisions

I'm going to use this model when I have this kind of data set:
time-based data | categorized data

Not recommended for:
Never use when the focus is on comparing the sizes of the individual categories or when the total sum of the elements in the bar is not relevant.





Teknik Storytelling

Seorang data scientist harus mampu untuk mengkomunikasikan proses dan hasil temuan analisis data dengan sistematis, menarik, tidak ambigu dan mudah dipahami bagi orang-orang. Bergantug di organisasi tempat data scientist bekerja, berkomunikasi dapat dilakukan secara tertulis (dalam bentuk laporan) maupun tatap-muka pada rapat atau seminar. Ibaratnya "Mendongeng" [*telling a story*], pembaca atau audines haris dibuat "terpesona" *[impressed]* dan percaya dengan hasil-hasil temuannya.

Gambar 1.5 Teknik Storytelling Data Science

Cara Presentasi Storytelling

1. Slide presentasi lebih banyak “gambar” daripada “kata-kata”
2. Pilih template “powerpoint” sesuai tema masalah/keilmuan yang sedang dibahas.
3. Untuk “pembukaan” dan “permasalahan” dapat dibuka dengan video / animasi gambar.
4. Singkat, padat, dan jelas. Usahakan untuk memberi “impressi” dan memancing pertanyaan “audience”. Gunakan ajang presentasi untuk lebih banyak menjawab pertanyaan “audience”
5. Waktu optimal dalam presentasi adalah 10 menit.

Challenge

Buatlah sebuah resume yang berisi grafik sesuai dengan kaidah pembuatan grafik dan jenis grafik nya, jelaskan variable bebas dan variable terikat pada grafik tersebut. Ambil dataset nya dari www.kaggle.com . Deadline Jumat depan.

Reference

http://www.amathsdictionaryforkids.com/qr/d/discreteData.html
http://blog.ub.ac.id/fitra/files/2012/06/cats.jpg
http://www.amathsdictionaryforkids.com/qr/c/categoricalData.html
https://seputarilmu.com/2020/03/grafik.html
https://www.megabagus.id/jenis-jenis-data-dalam-analisis-statistik/
IBM Skills Academy, 2020. Lecture 6 (Data Visualization & Presentation). United States : IBM Press.