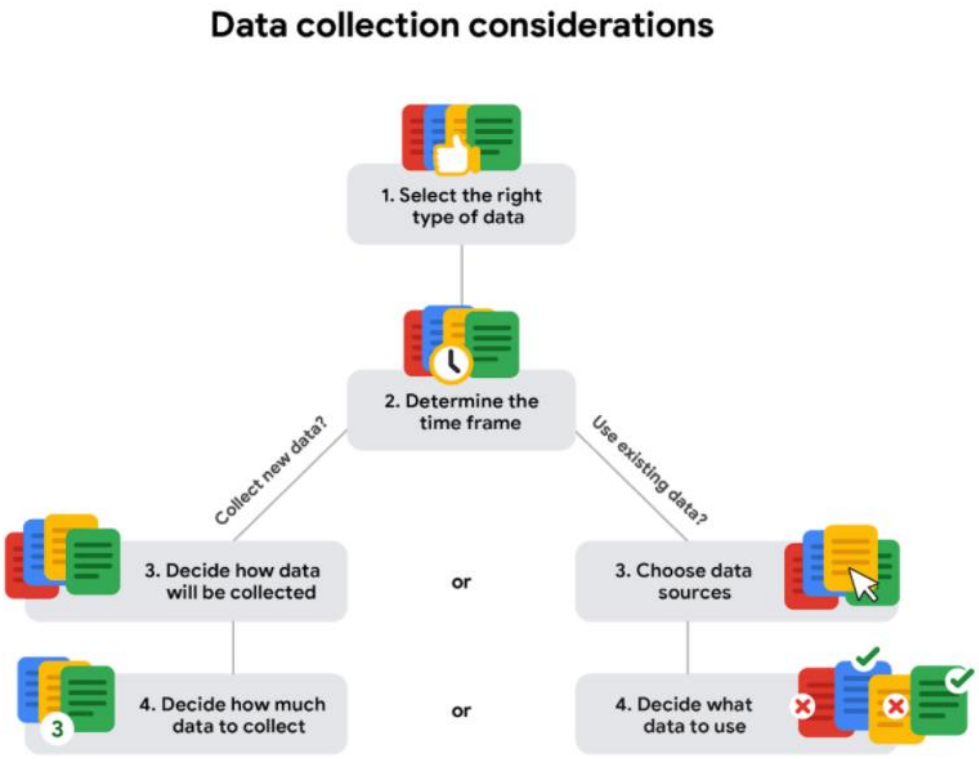
Course/topic: Course 3: Prepare Data (WEEK 1)

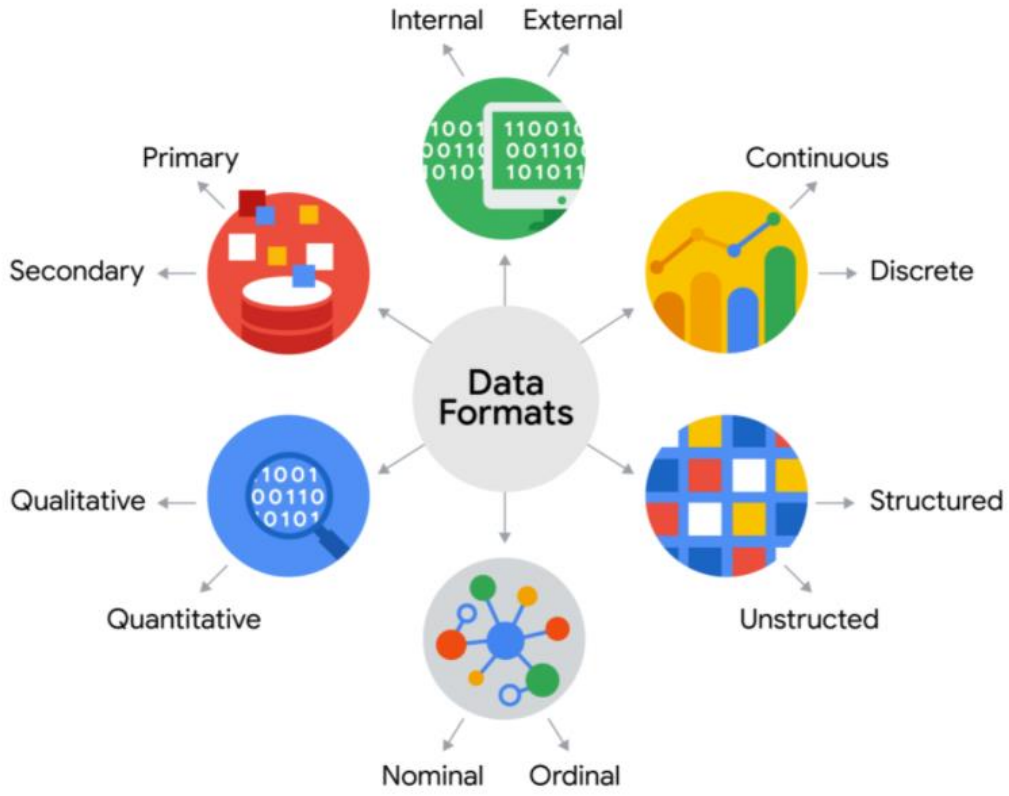
Learning Log: Consider how data analysts approach tasks

Some Factors yang dapat dipertimbangkan saat mencari data

1. **How** Datanya dapat dikoleksi ?
2. **Sources** datanya dari mana ?
    - First Party Data → Lu/tim lu Sendiri yang nyarik
    - Second party Data → Kalau ini datanya di collect sama orang lain terus dijual (Kek kek Data Twitter) (biasanya orang orang experince ini)
    - Third Party Data → Kalau ini datanya data dari berbagai sumber yang ngambil juga biasanya belum berpengalaman , kek data yang di kaggle
3. Decide Data apa yang **Digunakan ?**
4. Seberapa **Banyak** data yang dibutuhkan ?
5. Select **Right data type** (maksudnya adalah fitur – fiturnya berguna untuk kebutuhan analisis dalam menjawab permasalahan)
6. Time Frame (Lama **lu nyarik data)**



Discover data formats



Data Format Classification	Definition	Examples
Primary data	Collected by a researcher from first-hand sources	- Data from an interview you conducted - Data from a survey returned from 20 participants - Data from questionnaires you got back from a group of workers
Secondary data	Gathered by other people or from other research	- Data you bought from a local data analytics firm's customer profiles - Demographic data collected by a university - Census data gathered by the federal government

Data Format Classification	Definition	Examples
Internal data	Data that **lives inside** a company's own systems	- Wages of employees across different business units tracked by HR - Sales data by store location - Product inventory levels across distribution centers
External data	Data that **lives outside** of a company or organization	- National average wages for the various positions throughout your organization - Credit reports for customers of an auto dealership

Data Format Classification	Definition	Examples
Continuous data	Data that is measured and can **have almost any numeric** value	- Height of kids in third grade classes (52.5 inches, 65.7 inches) - Runtime markers in a video - Temperature
Discrete data	Data that is counted and has **a limited number** of values	- Number of people who visit a hospital on a daily basis (10, 20, 200) - Room's maximum capacity allowed - Tickets sold in the current month

Data Format Classification	Definition	Examples
Qualitative	**Subjective and explanatory** measures of qualities and **characteristics**	- Exercise activity most enjoyed - Favorite brands of most loyal customers - Fashion preferences of young adults
Quantitative	Specific and objective measures of **numerical facts**	- Percentage of board certified doctors who are women - Population of elephants in Africa - Distance from Earth to Mars

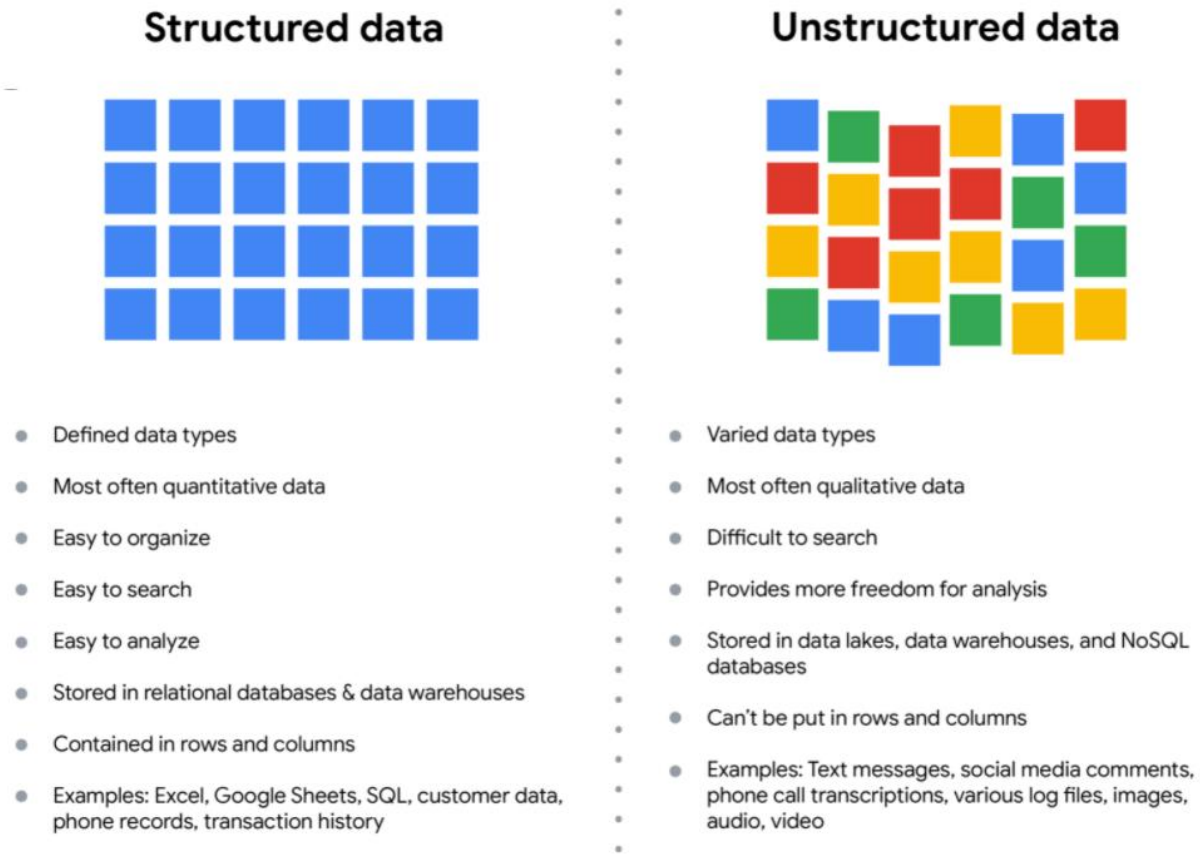
Data Format Classification	Definition	Examples
Nominal	A type of qualitative data that isn't categorized with a set order (Urutan gak Ngaruh)	- First time customer, returning customer, regular customer - New job applicant, existing applicant, internal applicant - New listing, reduced price listing, foreclosure
Ordinal	A type of qualitative data with a set order or scale	- Movie ratings (number of stars: 1 star, 2 stars, 3 stars) - Ranked-choice voting selections (1st, 2nd, 3rd) - Income level (low income, middle income, high income)

Data Format Classification	Definition	Examples
Structured data	Data organized in a certain format, like rows and columns	- Expense reports - Tax returns - Store inventory
Unstructured data	Data that isn't organized in any easily identifiable manner	- Social media posts - Emails - Videos

The structure of data

Ada 2 , Structured Data dan Unstructured Data

Structured → Data Tables / Tabular (Rows & Columns)
Unstructured Data → Not Organized (Images, AudioFiles, PDF files ,



Data modeling levels and techniques

Data modelling → Sebuah **Proses** Membuat **Diagrams** yang memvisualisasikan **bagaimana** data ter **Organisir** dan **terstruktur.** Representasi Visual dari **Data modelling** (Hasilnya , disebut **Data Model)**

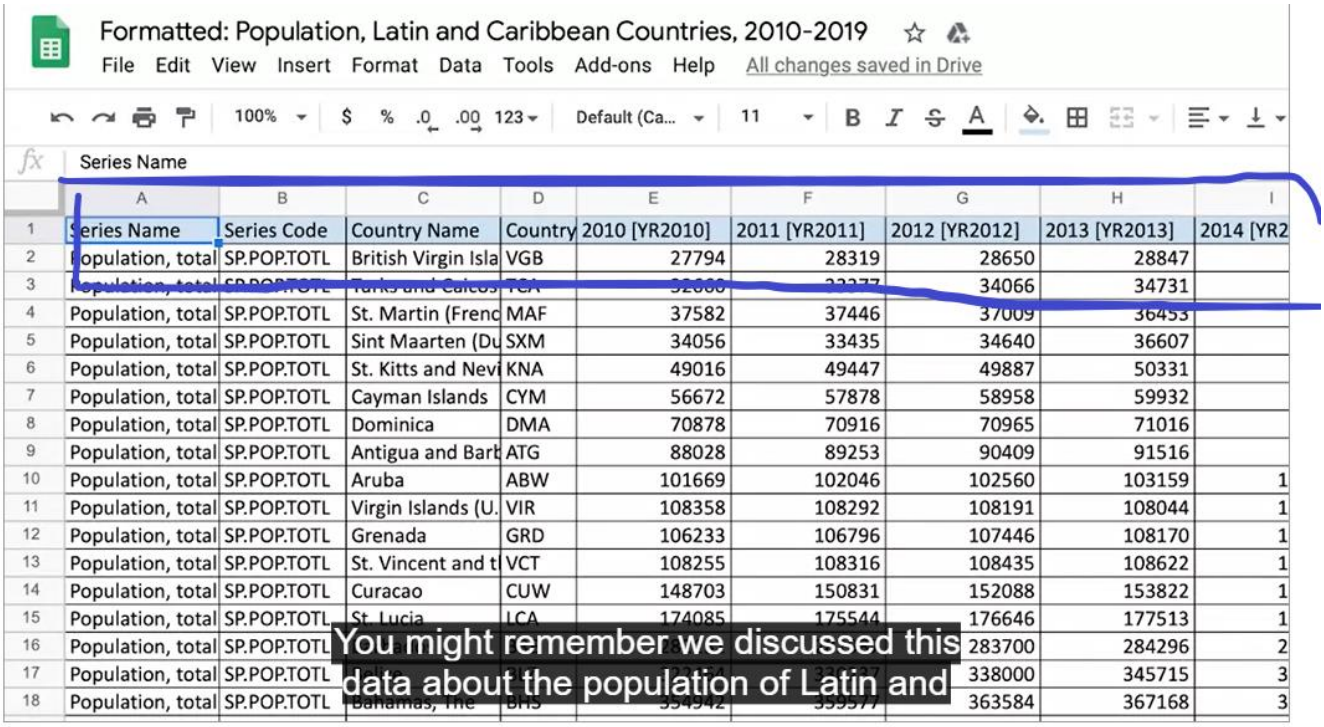
Bayangin aja Kek **blueprints Rumah.** Pasti nanti itu digunakan oleh ahli Listrik, ahli lainnya. Nanti itu digunakan untuk **Tujuannya sendiri**, dan secara langsung meraka **Harus memahami blueprintnya** Agar dapat mencapai tujuan tersebut.

Ada 3 Data Modelling Levels

1. **Conceptual Data Modelling → Ga perlu** Detail Teknis, **High levels dari** struktur data (Kek interaksi antar data di organsisasi) kek kek ERD cok Tapi belom ada isinya baru tabel tablenya
2. **Logical Data modelling → Mulai Detail terhadap** Database nya (**Attributes, relationships, entities)** Jadi Kek inisiasi detail dari Databasenya
3. **Physical Data Modelling → Kalau ini Ya** Gimana **Operasional databasenya** Berjalan.

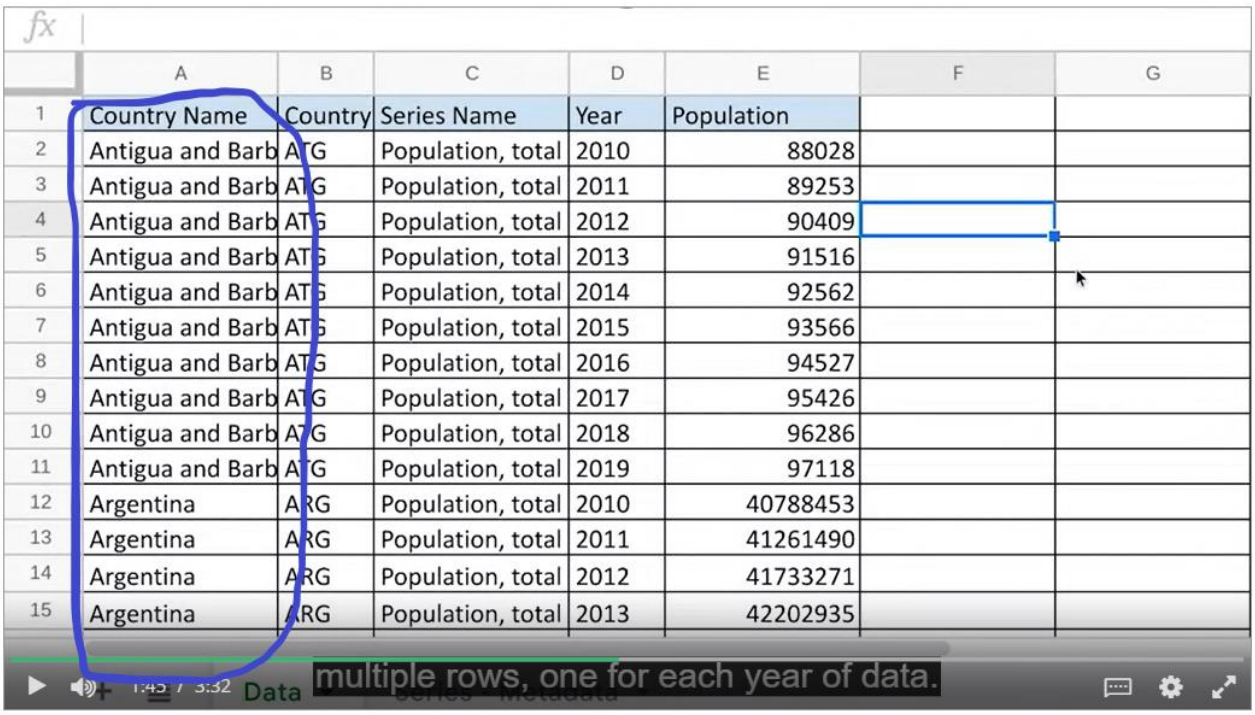
Meet wide and long data

Wide Format :



	A	B	C	D	E	F	G	H	I
1	Series Name	Series Code	Country Name	Country	2010 [YR2010]	2011 [YR2011]	2012 [YR2012]	2013 [YR2013]	2014 [YR2
2	Population, total	SP.POP.TOTL	British Virgin Isla	VGB	27794	28319	28650	28847	
3	Population, total	SP.POP.TOTL	Turks and Caicos	TCA	32600	33277	34066	34731	
4	Population, total	SP.POP.TOTL	St. Martin (Frenc	MAF	37582	37446	37009	36453	
5	Population, total	SP.POP.TOTL	Sint Maarten (Du	SXM	34056	33435	34640	36607	
6	Population, total	SP.POP.TOTL	St. Kitts and Nevi	KNA	49016	49447	49887	50331	
7	Population, total	SP.POP.TOTL	Cayman Islands	CYM	56672	57878	58958	59932	
8	Population, total	SP.POP.TOTL	Dominica	DMA	70878	70916	70965	71016	
9	Population, total	SP.POP.TOTL	Antigua and Barb	ATG	88028	89253	90409	91516	
10	Population, total	SP.POP.TOTL	Aruba	ABW	101669	102046	102560	103159	1
11	Population, total	SP.POP.TOTL	Virgin Islands (U.	VIR	108358	108292	108191	108044	1
12	Population, total	SP.POP.TOTL	Grenada	GRD	106233	106796	107446	108170	1
13	Population, total	SP.POP.TOTL	St. Vincent and t	VCT	108255	108316	108435	108622	1
14	Population, total	SP.POP.TOTL	Curacao	CUW	148703	150831	152088	153822	1
15	Population, total	SP.POP.TOTL	St. Lucia	LCA	174085	175544	176646	177513	1
16	Population, total	SP.POP.TOTL	Belize	BLZ	278358	281637	283700	284296	2
17	Population, total	SP.POP.TOTL	Suriname	SUR	534354	537037	538000	345715	3
18	Population, total	SP.POP.TOTL	Bahamas, The	BHS	354942	359577	363584	367168	3

Long Format :



	A	B	C	D	E	F	G
1	Country Name	Country	Series Name	Year	Population		
2	Antigua and Barb	ATG	Population, total	2010	88028		
3	Antigua and Barb	ATG	Population, total	2011	89253		
4	Antigua and Barb	ATG	Population, total	2012	90409		
5	Antigua and Barb	ATG	Population, total	2013	91516		
6	Antigua and Barb	ATG	Population, total	2014	92562		
7	Antigua and Barb	ATG	Population, total	2015	93566		
8	Antigua and Barb	ATG	Population, total	2016	94527		
9	Antigua and Barb	ATG	Population, total	2017	95426		
10	Antigua and Barb	ATG	Population, total	2018	96286		
11	Antigua and Barb	ATG	Population, total	2019	97118		
12	Argentina	ARG	Population, total	2010	40788453		
13	Argentina	ARG	Population, total	2011	41261490		
14	Argentina	ARG	Population, total	2012	41733271		
15	Argentina	ARG	Population, total	2013	42202935		

Wide data is preferred when	Long data is preferred when
Creating tables and charts with a few variables about each subject	Storing a lot of variables about each subject. For example, 60 years worth of interest rates for each bank
Comparing straightforward line graphs	Performing advanced statistical analysis or graphing

Transforming data

Sebuah Proses mengubah **Format data, struktur, dan nilai dari** dataset.
Biasanya melakukan hal – hal sebagai berikut :

1. **Adding, Copying, dan** Replikasi Data
2. **Menghapus Kolom / Rows**
3. **Standarisasi Nama variabels (Misalnya LOL, jadi** lol)
4. **Renaming, moving, atau** Combining **Columns** pada database
5. Menggambung **Dataset**
6. Menyimpan **File** pada format yang berbeda (Xlsx → .csv)

Tujuannya apasih ?

1. Organisasi Data → Ya biar mudah diakses / digunakan nya
2. Kompabilitas Data → Misalnya nih ada sistem yang menggunakan (10 kolom) tapi data yang kamu masukin hanya (8 Kolom) Error nantinya
3. Data Migration → **Mempermudah migrasi data** dari satu tempat ke tempat lain (Misalnya Database punya Variable **(FIrstname, LastName)** nah Terus ingin di Migrasi ke database lain yang udah punya Data dimasa lampau dengan **Kolom yang sama** (Ini mempermudah dalam penggabungan)
4. Data **Merging** → Menggabung data**, mirip konteksnya kek Migrasi, hanya saja dilakukan pada** Database yang sama
5. Data **Ehnhancement** → Data nya dapat ditampilikan dengan **Kolom yang lebih detail** (Kurang paham sih ini
6. Data **Comparison**

Course/topic: Course 3: Prepare Data (WEEK 1)

Bias: From questions to conclusions

Data bias → **Type of error** Yang membuat **Jawaban dari sebuah analisis cenderung ke suatu arah tertentu** . Misal nih, Kita pengen prediksi **makanan favorit di bandung**. Tapi kita cuman ngambil data di Bojongsoang saja, di Cikoneng, di tempat laen ga diambil itu udah **BIAS**

Intinya hati hati dengan BIAS, karena ujung ujungnya Impact yang diberikan dari **Data yang bias dapat merugikan**

Biased Sampling → **Sampling tapi lu bias , mirip kek contoh data bias**
Unbiased Sampling → Sampling tapi lu ga bisa Karena merepresentasikan populasi yang hendak **Di**

ukur. Contohyna Lu ambil seluruh Sample dari setiap daerah/kecamatan di Bandung. EZ PZ CAPEK

TIPS bisa tau datanya ga Bias

1. Buat visualisasi antara data populasi dengan sample

Understanding bias in data

Tipe bias di data itu

1. **Sampling Bias →** Ini mah sudah jelas lah ya
2. **Observer Bias → Kalau ini** Tendency untuk orang lain dalam **Observe** Sesuatu berbeda beda (Misal nih, kamu lagi observasi Seorang teman kamu yang depresi, **Behaviour** yang dilakukan oleh orang tersebut mungkin **berbeda** sewaktu **Dirimu yang observe.**
3. **Interpretation Bias → Kalau ini Terkait dengan interpretasi** Suatu Hal. Misal nih **lu mendapatkan kesimpulan kalau misalnya Kolom** A itu berpengaruh besar bagi perusahaan. Disatu sisi temen lu bilang **Kolom** A itu ngga berpengaruh besar (Nah ini perlu dikonfirmasi lagi) . Makanya ada yang namanya kerja sama tim
4. **Confirmation Bias → Kalau** ini ya dari keinginan kita **Mendapatkan** Apa yang memang kita inginkan. Misal nih Misal (**Kita sebelum analisis data** Langsung mikir, keknya **Variabel Harga** itu dipengaruhi besar deh dengan **Variabel** Makanan) Sehabis itu, langsung melihat Datanya dengan cepat, visualisasi dengan cepat, dan benar saja didapatkan hal seperti itu (Namun belum dilakukan analisis mendalam terkait variabel **Makanan Tersebut** (Itu kapan terjadinya ? **Hari libur ? Event – event besar ?** Intinya ini mah gunain **Logika tanpa melihat pakta di data**

Explore data credibility

1. Good DataSources itu kek mans sih boy ?
   Best Practicenya adalah menggunakan **ROCCC**
   R → Reliable (datanya Compelete, ga bias) udah di uji bahwasanya dapat digunakan
   O → Original (Maksudnya adalah Data yang digunakan **pada saat menjawab sebuah permasalahan** dan data tersebut dari **(Second / Third party) Harus dipastikan datanya original ,** Misal nih lu butuh data Truk Indonesia, Lu beli tuh Ke Second dan coba download dari Third party, PASTIIN data **TRUK yang lu dapet itu representative dengan data truk Indonesia, bukan data truk laen laen**)
   C → Comprehensive (Datanya bisa menjawab pertanyaan, simple)
   C → Current (Datanya Up to date)
   C → Cited (Datasetnya credible, karena ada sumber nya (Siapa yang ngambil, Orangnya Kredible ?,Kapan terakhir update datanya ?)

   BAD data ya dataset yang tidak ROCCC

Introduction to data ethics

Etika dalam data, anjay
Data Ethics → Suatu Standard **Benar atau salah** yang menjadi dasar bagaimana sebuah data **Collected, shared, dan used.**

Aspect Aspect dari data ethics (Ya ini dipertimbangkan saat mendapaktan data)

Ownership → Sokap yang punya datanya , dia punya kontrol utuh bagaimana data data tersebut **digunakan, di proses, dan di share**

Transaction transparency → Kalau Ini berbicara mengenai bagaimana data tersebut digunakan. Dengan artian **Seluruh aktivitas dan algoritma yang menggunakan data tersebut harus dapat di JELASKAN DAN DIPAHAMI OLEH SI PEMILIK DATA**

Consent → Kalau ini adalah **Hak kepada pemiliki** data tersebut untuk memahami dan mengetahui **Bagaimana dan mengapa datanya digunakan,** Sebelum melakukan **Persetujuan.**
Currency → Gimana transaksi datanya, Apa mungkin dijual datanya ?
Privacy → Privasi datanya kayak gimana, privasi bukan hanya keamanan saja, tapi beberapa hal berikut juga menjadi pertimbangan
1.Kebebasan dalam menentukan Tujuan dari data
2.Kebebasan dalam inspect,update, atau correct data
3.Kebabasan untuk dapat melakukan conset pada data yang diberikan
4.Memiliki akses legal terhadap data .

Data anonymization is used in just about every industry. That is why it is so important for data analysts to understand the basics. Here is a list of data that is often anonymized:

- Telephone numbers
- Names
- License plates and license numbers
- Social security numbers
- IP addresses
- Medical records
- Email addresses
- Photographs
- Account numbers

Inti nya sebisa mungkin mengamankan data data sensitif.

Openess → Free access, use and share data Dalam artian sebagai berikut :
1.Availability dan access
2.Reuse dan Redistribution (must provided under terms
3. Universal Participation

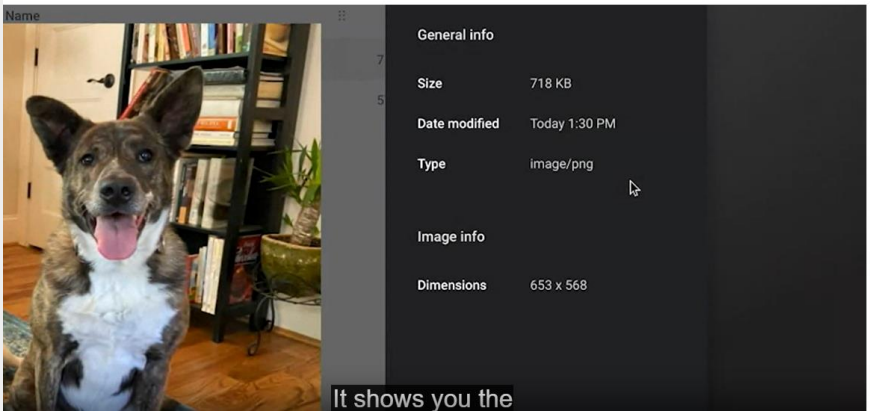
WEEK 3

Exploring metadata

Metadata → Data About data (In data analytics, metadata helps data analysts interpret the contents of the data within a database.. intinya ini cuman deskripsi dari sebuah data. Contohnya →

3 Common Types dari metadata

1. Deskriptif → Meta data yang mendeskripsikan sebuah data dan dapat digunakan untuk identifikasi data tersebut (ISBN, NIK)
2. Struktural → Metadata yang mengindikasikan bagaiamana sebuah data ter organisasi (Lembar Halaman/chapter di buku, **spreadsheet sebuah database**)
3. Administratif → Metadata yang mengandung **Technical source dari digital asset**



)

Elements of metadata

Before looking at metadata examples, it is important to understand what type of information metadata typically provides.

Title and description

What is the name of the file or website you are examining? What type of content does it contain?

Tags and categories

What is the general overview of the data that you have? Is the data indexed or described in a specific way?

Who created it and when

Where did the data come from, and when was it created? Is it recent, or has it existed for a long time?

Who last modified it and when

Were any changes made to the data? If yes, were the modifications recent?

Who can access or update it

Is this dataset public? Are special permissions needed to customize or modify the dataset?

Jargon (Metadata repository

Biasanya mempunyai karakteristik sebagai berikut

1. Describe the state dan lokasi dari metadata
2. Describe Struktur tabel di dalam database
3. Describe bagaimana data mengalir di repository (Hubungan antar datanya)
4. Traccking yang akses metadata

)

Benefits menggunakan metadaData

1. Metadata creates a single source of truth by keeping things consistent and uniform. Maksudnya adalah semua stakeholder memahami **Context dari data tersebut dengan uniform (Sama & Konsisten)**
2. **Ngebuat data menjadi** Reliable dengan memastikan data tersebut **Akurat, precise, relevan dan tepat**

Sisanya Week 4 dan week 5 itu materi basic spreadsheet dan tips & trick karir, ga gw tulis