Course/topic: Course 4: Process Data (WEEK 1)

Learning Log: Consider how data analysts approach tasks

Why data integrity is important

Integritas → (Mutu)

Data Integrity → The accuracy, Complteeneess and trushworthiness dari data **dari awal sampai akhir lifecycle (** Intinya Datanya sesuai agar bisa mengahsilkan **Strong analysis**)
Dan kadang **Data integrity ini** Suka terjadi dikarenakan hal berikut :

1. **Data** Replication → Sebuah proses untuk **Storing data** di banyak lokasi (Bisa jadi penyebab ketidakkonsistensi data saat digunakan) (**Integritas ter dampak)**
2. Data **Transfer** → Ngirim ngirim data (Kalau misalnya ke **interrupted tapi kita gatau atau lupa)** Dah fix datanya gabaleg
3. **Data** Manipulation → **Sebuah proses merubah data** agar terorganisir dan mudah dibaca

Threats **Data Integrity**

1. Hack
2. Human Error
3. System failurs
4. Viruses

Beberapa hal yang perlu diperhatikan di **Batasan data (ini perlu diperhatikan saat mengembangkan sebuah solusi model)**

Data constraint	Definition	Examples
Data type	Values must be of a certain type: date, number, percentage, Boolean, etc.	If the data type is a date, a single number like 30 would fail the constraint and be invalid
Data range	Values must fall between predefined maximum and minimum values	If the data range is 10-20, a value of 30 would fail the constraint and be invalid
Mandatory	Values can't be left blank or empty	If age is mandatory, that value must be filled in
Unique	Values can't have a duplicate	Two people can't have the same mobile phone number within the same service area
Regular expression	Values must match a prescribed pattern	A phone number must match ###-###-#### (no other characters allowed)

(regex) patterns		
Cross-field validation	Certain conditions for multiple fields must be satisfied	Values are percentages and values from multiple fields must add up to 100%
Primary-key	(Databases only) value must be unique per column	A database table can't have two rows with the same primary key value. A primary key is an identifier in a database that references a column in which each value is unique. More information about primary and foreign keys is provided later in the program.
Set-membership	(Databases only) values for a column must come from a set of discrete values	Value for a column must be set to Yes, No, or Not Applicable
Foreign-key	(Databases only) values for a column must be unique values coming from a column in another table	In a U.S. taxpayer database, the State column must be a valid state or territory with the set of acceptable values defined in a separate States table
Accuracy	The degree to which the data conforms to the actual entity being measured or described	If values for zip codes are validated by street location, the accuracy of the data goes up.
Completeness	The degree to which the data contains all desired components or measures	If data for personal profiles required hair and eye color, and both are collected, the data is complete.
Consistency	The degree to which the data is repeatable from different points of entry or collection	If a customer has the same address in the sales and repair databases, the data is consistent.

Well-aligned objectives and data

Kalau ini sebenernya **kerelevanan data dalam** menjawab suatu permasalahan. Ok ok Seems a lot things going on here. First Remember this

Clean data + Keselarasan dengan tujuan bisnis = Akurat **Conclusion**

Clean data + Keselarasan + New Discovered Variabels + Batasan = Akurat **Conclusion**

Dealing with insufficient data

Ini sebenarnya konsep yang sangat penting untuk dipahami yaitu **Pahami tujuan bisnis yang mau dicapai** oleh data itu apa ? Nah kenapa seperti itu dikarenakan pada dasarnya **Diri sendiri/tim sendiri yang** dapat menentukan apakah data yang dimiliki sudah sufficient atau belum. Beberapa **Tipe insufficient data**

1. Data **From only one source (Kalau ini selaras dengan tujuan bisnis aja**, Ga semua insufficient data)
2. Data that keeps Updating (Kalau ini ada baiknya ambil recent data tapi ya nunggu 1 bulan kemudian misalnya) soalnya biar datanya **Compelete, otherwise** ga Reliable
3. Outdated Data
4. Limited Geographically → Tergantung scopes(Kalau misalnya mau seluruh Indonesia yang diselesaikan masalah kemacetannya, maka dari itu harus punya data yang merangkul hal tersebut secara keseluruhan)

Cara mengatasi permasalahan diatas dapat dilakukan beberapa hal sebagai berikut:

1. Identifikasi trend dari data yang ada
2. Tunggu **Data tambahan**, jika diperbolehkan
3. **Bicara** dengan **Stakeholders dan adjust** Objectivenya
4. Look for a new Dataset

Beberapa isu dan solusi dari google terkait data

1. **Ga ada Data**

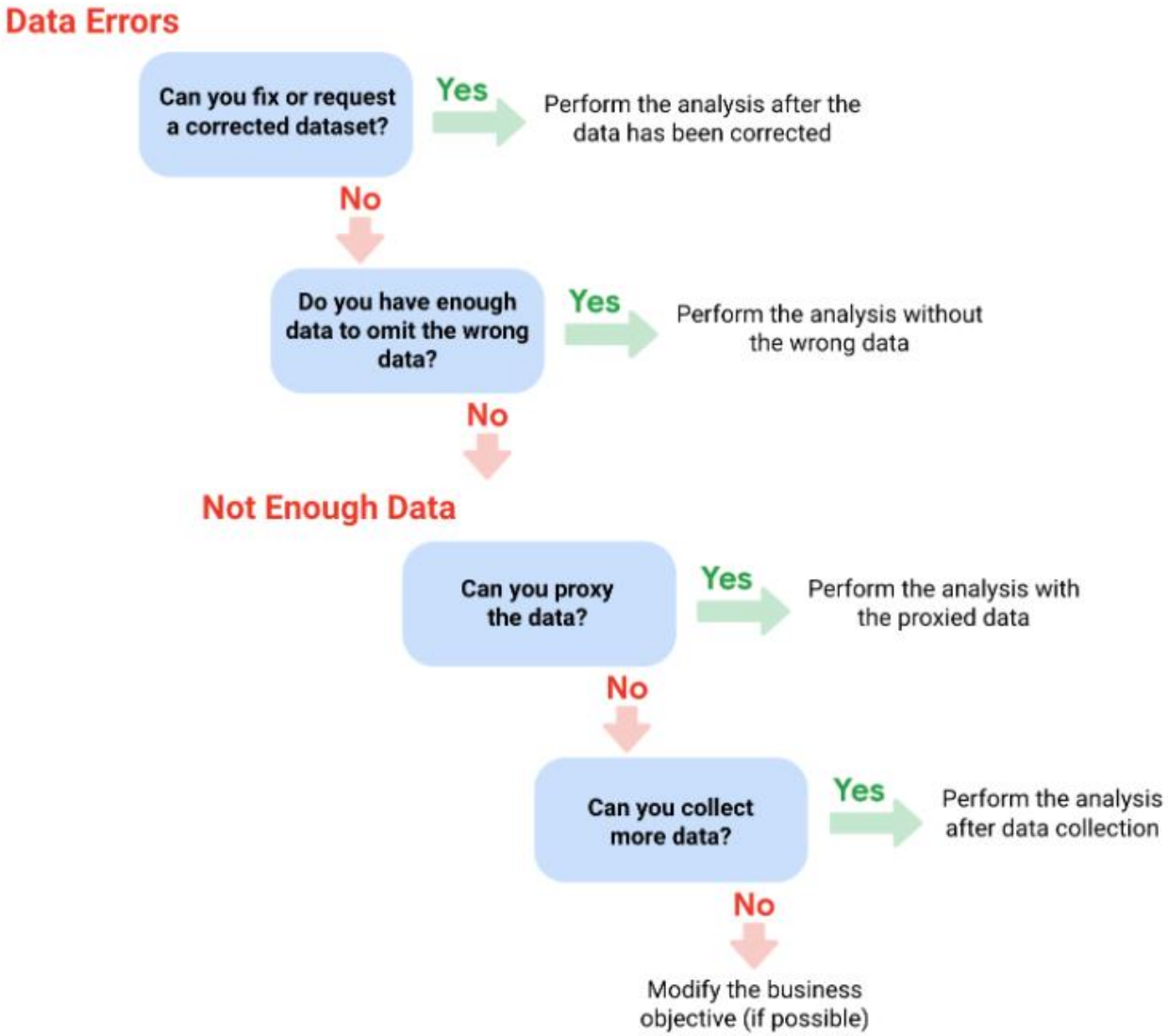
Possible Solutions	**Examples RealLife Solutions**
Kumpulin data **at small scale** aja untuk melakukan **Preliminary analysis**, lalu minta waktu tambahan untuk menyelesaikan analisanya setelah **Menambahkan/collect data tambahan**	**Contohnya gini :** If you are surveying employees about **what they think** about a **new performance and bonus plan**, use a **sample f**or **a preliminary analysis**. Then, ask for another 3 weeks to collect the data **from all employees**.
Kalau ga ada waktu Collect data, lakukan analisis menggunakan **Proksi data** dari **dataset lain**	Contohnya gini : If you are **analyzing peak travel times** for commuters but **don't have** the data for **a particular city**, **use** the data **from another** city with **a similar size and demographic.**

2. **Too little data**

Lakukan analisis menggunakan **Proxy data** bersamaan dengan **Actual data**	**Misal kamu lakuin analisis** trends terkait kepemilikan
Adjust analisnya agar selaras dengan data yang sudah dimiliki	**Jika terdapata missing data dari** umur 18-24 tahun.Tetap lakukan analisis, dan tetap tuliskan reportnya, sebagai contoh : " Conclusion applies to adults 25 years and older only "

3. **Wrong data, includeing data with** Erros

If you have **the wrong data** because **requirements were misunderstood**, **communicate the requirements again**.	If **you need** the data **for female voters** and **received the data for male voters**, restate your needs.
Identify errors in the data and, if possible, correct them at the source by looking for a pattern in the errors.	Kalau misalnya ada error dari hasil kalkulasi (Perbaiki prosesnya bukan nilainya) **kali pada** data testingnya lupa ngelakukan **normalisasi**.
Misalnya ga bisa **Membetulkan data errors,** ya abaikan saja (“Drop”) terus lakukan analisis dengan data yang ada (**Becareful , big sample size ya)** Kalau kecil malah jadi bias dianya.	Misal nih datasetnya di **translate dari** Bahasa yang berbeda, banyak **Translation** yang gak jelas, yowes di **abaikan saja,** Analisa data yang ada.



The importance of sample size

Penting pada dasarnya mengambil sebuah **data** yang **merepresentasikan populasi,** Kenapa hal ?

Dikarenakan **Sampling itu** butuh **subsetnya** populasi, apabila **Sampling nya bias, then con**clusionnya **ga reliable**. Makanya perlu dilakukan **Random sampling dari populasi** agar samplingnya gak bias (Tentu pada ujungnya kita tetap harus melakukan **Verifikasi terlebih dahulu**, apakah data yang **hasil random samplingnya sudah** merepresentasikan **populasi**.

Beberapa Notes penting :

Populasi → **Entire group yang di interest. Kalau** survey sebuah perusahaan, maka seluruh **Anggota perusahan harus di survey**

Sample → melakukan survey untk sebagian anggota dari perusahaan

Margin Of Error → Dikarenakan sample size itu merepresentasikan populasi. Maka hasilnya ada kemungkinan berbeda dengan menggunakan populasi. Perbedaannya disebut sebagai **Moe**. Semakin kecil MOE maka, semakin baik merepresentasikan populasinya.

Confidence Level → Seberapa **Confident diri kita terhadap** hasil dari **survey**. Misal 95% **Confidence** level berarti jika kamu melakukan survey dengan populasi yang sama selama 100 kali, 95 diantaranya memiliki hasil yang mirip.Definisikan diawal ya

Confidence Interval → **Range dari possible values**

Statistical significance → Menentukan apakah hasil **nya** karena **random chance** atau **not.** Semakin bagus **significancenya** semakin baik

Jangan menggunakan sample size **kurang dari 30**.
Confidence level most common value is 95%.

Higher confidence level, use **larger** sample size
Decrease margin of error, Tambah data (Larger sample size)
Greater Stastical **significance, Tambah data** (Larger sample size)

Sample Sizes vary by **business problems**
Misal nih, lu tinggal di sebuah daerah dengan populasi 200,000 dan lu lakuin survey dapet deh 180.000 itu udah besar bet. Tapi sample size yang cocoknya berapa ? Apakah **200** cukup untuk mereprsentasikan seluruh distrik padakota tersebut ?
DEPENDs dari masalahnya

1. Sample size sebesar 200 mungkin besar apabila **Permasalahan bisnisnya** adalah mengetahui **Bagaimana** suatu penduduk merespon **Perpustakaan** baru ditempatnya
2. Sample size 200 mungkin tidak cukup apabila **Permasalahan bisnisnya** adalah menentukan bagaimana suatu penduduk memberikan voting untuk membiayai pembagunan **Perpustakaan**

Ingat bahwasanya **Larger sample sizes** memiliki Cost yang lebih besar. Ingat saja dengan permasalahan yang hendak dibahas, misal nih **Customer preferences** itu ga perlu sample size yang besar besar amat, dibandingkan dengan **Pengujian obat obatan**. **Karena ujung jungnya mahal**

Hipotesis testing → Salah satu cara untuk mengetahui apakah survey atau eksperimen memiliki **meaningful** result

Statistical power → n% chance **Mendapatkan** Satistically significant **result** pada eksperimen ….
Statisticall Significant → **Results dari test** adalah nyata dan bukan **Error** yang disebabkan oleh **Random chance**. Sisanya adalah kemungkinan error

Oh iya sebelum mengambil sebuah eksperimen, **Harus harus ini mah** Cari tau **penyebab – penyebab yang bisa mencegah hasil statistically significant. Jadi lu harus make sure dlu** Sample size yang lu ambil itu memiliki **faktor – faktor tersebut atau ngga**. Biar sesuai dengan faktor yang di ingkan misalnya "Rasa Ayam bakar baru", apabila ternyata lokasinya ada pembangunan / Saingan ada tuh buat promosi , kalau bisa sample sizenya **Mempertimbangkan hal tersebut. Dikarenakan yang di ingkan adalah** Hasil dari "**Rasa ayam bakar baru"** bukan faktor lain.

PROKSI Data, ini mngkin hal ntuk membayangkannya

Business scenario	How proxy data can be used
A new car model was just launched a few days ago and the auto dealership can't wait until the end of the month for sales data to come in. They want sales projections now.	The analyst proxies the number of clicks to the car specifications on the dealership's website as an estimate of potential sales at the dealership.
A brand new plant-based meat product was only recently stocked in grocery stores and the supplier needs to estimate the demand over the next four years.	The analyst proxies the sales data for a turkey substitute made out of tofu that has been on the market for several years.
The Chamber of Commerce wants to know how a tourism campaign is going to impact travel to their city, but the results from the campaign aren't publicly available yet.	The analyst proxies the historical data for airline bookings to the city one to three months after a similar campaign was run six months earlier.

Intinya **Proksi data** digunakan kalau datanya ga ada, dan juga proksi data ini **bisa memberikan** Gambaran terkait **Keinginan stakeholder.**

Case 1 Contohnya, stakeholder pengen sales projection → Data analis Proxy datanya dengan menggunakan **banyaknya** klik pada sebuah **Spesifikasi kendaraan** untuk estimasi **potensial pembelian**

Case 2 Contohnya, stakeholder pengen tah tuh permintaan produk baru di 4 tahun kedepan gimana → **Data analis proxy** datanya dengan menggunakan **produk lain** yang mirip

Case 3 Contohnya, **Sebuah perusahaan ecommerce** pengen tau nih **impact dari** Tourism campaign ke travel to the city → **Karena datanya ga adata, data analis** proksi datanya menggunakan **Airline bookings ke** City tersebut satu - 3 bulan **SESUDAH campaign yang sama** sebelumnya.

EZ PZ right ? Gunain aja data proksi itu menggambarkan

Pengingat lagi :

Confidence level → Probabilitas **That** your **Sample size accurately** Reflects **population**

Margin of error → Seberapa dekat hasil dari **Sample size dengan** Hasil **Menggunakan** data populasi

Estimated response rate → Kalau misalnya lakuin survey dan butuh 100 Response, dan kemungkinan response ratenya 10% dari survey yang u buat (YA LU KIRIM 1000 Survey ke orang laen) kalau 20% gimana ? **Ya 800 survey kirim .**

Sebelum ngumpulin data, Try to kumpulin dlu informasi dibawah ini

1. Confidence Level
2. Margin Of error → Semakin kecil margin of error semakin dekat dengan **Result sample size dengan result dari population () Ez pz**
3. Population

Ini example dari Margin of Error

Lu buat survey mengenai "Apakah seseorang menyukai Five-day work / week atau 4 day work / week ?
Nah Terus berdasarkan hasil survey lu dapet bahwasanya **60%** dari responden menyukai 4 day work / week. Lu set up margin of errornya **10 %** . Yang memberikan lu insight bahwasanya 50% - 70% Responden **Like the idea untuk populasi**. Apabila lu Set up **95%** Confidence Interval, Maka terdapat 95% **Chance** that result dari populasi akan berada pada 50% - 70% **Like the idea YES (4 day work/ week)**

DATA CLEANING IS MUST (WEEK 2)

Cause Poor of Quality Data → HUMan ERROR

Dirty data → Data yang **Incomplete,** Incorrect, or **Irrelevant to the Problem you 're trying to solve**

Why data cleaning is important
Point out that, data cleaning kalau ga dilakukan dengan benar maka **LOSS** adalah hasilnya, so make sure datanya clean. Note that, **Kalau** nanti di organisasi data biasanya udah di cleaning sama **Data engineer/ data warehousing specialist** tapi **make sure do it again** Do sanity check (Bisa aja ada yang terleawat atau pun duplikat)

Types dari **Dirty data**

1. Duplicate Data

Description	Possible causes	Potential harm to businesses
Any data record that shows up more than once	Manual data entry, batch data imports, or data migration	Skewed metrics or analyses, inflated or inaccurate counts or predictions, or confusion during data retrieval

2. **Outdated Data**

Description	Possible causes	Potential harm to businesses
Any data that is old which should be replaced with newer and more accurate information	People changing roles or companies, or software and systems becoming obsolete	Inaccurate insights, decision-making, and analytics

3. **Incomplete** Data

Incomplete data

Description	Possible causes	Potential harm to businesses
Any data that is missing important fields	Improper data collection or incorrect data entry	Decreased productivity, inaccurate insights, or inability to complete essential services

4. **Incorrect/inaccurate data**

Incorrect/inaccurate data

Description	Possible causes	Potential harm to businesses
Any data that is complete but inaccurate	Human error inserted during data input, fake information, or mock data	Inaccurate insights or decision-making based on bad information resulting in revenue loss

5. **Incosistent Data**

Description	Possible causes	Potential harm to businesses
Any data that uses different formats to represent the same thing	Data stored incorrectly or errors inserted during data transfer	Contradictory data points leading to confusion or inability to classify or segment customers

Jadi intinya Hati hati sama Data