COURSE 03: Prepare data for Exploration

Instructor: Hallie Analytical Lead
Healthcare

Page No.
Date 07 01 23

WEEK 1 Data types and structures :→

1/4 Data Exploration

> Introduction :→

- Understanding the different types of data and data structures.
- What type of data is right for the question you're answering
- Practical skills about how to extract, use, organise and protect your data

story: She analyzes Medicare enrollment data over time and make connections to how people research Medicare plans on google.
This is helpful for medical professionals & patients

Prepare steps :→
> How data is generated
> Different formats, types, and structures of data
> Analyze data for bias and credibility
(cause not all data fits each need)
> what "clean data" means
> Databases
> Extract your own data using spreadsheets & SQL
> Basics of data organization
> The process of protecting your data

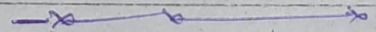
You still with me ?: ⇒ Yes maam!

* Fascinating data insights:

Healthcare & data togetner is a newer concept Using ML, AI, Big data help healthcare industry.

"Its really fasinating that we can take all of these data sets and synthesize them and allow us to deliver some cool insights and trends to our hospital systems."

The creativity side of analysis grows with exp. you can look at data one way now and a week later you might see a different trend altogether!



2/4 Collecting data :→

Every digital photo online is one piece of data Every photo itself holds even more data from number of pixel to color contained in them.

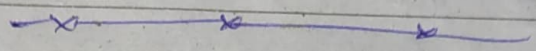
US Census Bureau use forms to collect data about the country's population.
This data is used for numerous reasons like funding for schools, hospitals & fire departments.

** As a DA you'll have access to all kinds of data and lots of it. Knowing how its generated can add more context to it and knowing how to collect it, can make DA process more efficient.

No matter what kind of data you use it must be checked for [accuracy & Trustworthyness]

How data is collected.

> Interviews
> Observations
> forms
> Questionnaires
> Surveys
> Cookies : (When tracking people's online activity and interests (most effective)



* What Data to collect ? :→

*First-party data: Data collected by individual or group using their own resources.

↓

Preferred method, as you know where the data came from.

*Second-party data: Data collected by a group directly from its audience and then sold

↓

* You buy from someone. Still Reliable

* Third party data :→
Data collected from outside sources who did not collect it directly

might not be reliable BUT still useful
You must check it for accuracy, bias & credibility.

population: All possible values in a certain dataset

Sample: part of popl^n that is representative of popl^n

* Data collection considerations :→

> How the data will be collected
> Choose data sources (1st, 2^n, 3rd - Party)
> Decide what data to use
> How much data to collect (popl^n & sample)
> Select the right data type.
> Determine the time frame.

ex. If you need something immediately : You'll need to use historical data

08-01-23

week 1
3/4

Differentiate between date formats and structures.

* Discover data formats :→
* > Qualitative Data: Name, category, description
* > Quantitative Data: Numeric data

stars, point → Discrete : counted & limit values
runtime → Continuous : measured & any num value

~~half not allowed~~

110.0356 minutes runtime

* Nominal Data: qualitative data categorized without set order.

↓

Doesn't have sequence

Did you ~~take~~? → (Yes, No, Not sure) Doesn't specify likelyness, etc

Page No.
Date

Did you like: → 4, 5, 3, 4, 1, 5
↓
* Ordinal Data: Qualitative data with set order/scale

* Internal data: →
Data that lives within a company's own system.
↓
more reliable & easier to collect.

* External data: →
Data that lives and is generated outside of an organisation.
↓
valuable when you need as much data as possible.

* Structured data: →
Data organized in a certain format such as rows & columns.
↓
Spreadsheets & Relational databases

* Unstructured data: →
Data that is not organised in any easily identifiable manner. (might have internal structure but doesn't fit neatly in rows & columns).
↓
Audio, Video as there's no clear way to identify or organize their content

Understanding structured data :→

Most of the data that is being generated is unstructured :→ Audio, video, emails, photos, social media.

Structured data works nicely in a data model

Data model :→

A model that is used for organizing data elements and how they relate to one another.

Data elements :→

Pieces of information, such as people's names, acc no's and addresses.

Data model help keep the data consistent and provide a map of how data is organised.

Structured data is also useful for databases, this way data analysts can enter, query & analyze the data whenever they need to.

Makes visualization easy!

As structured data can be directly applied to charts, maps, graphs, dashboards & many other visual representations of data.

sources : spreadsheets & databases.

* Data Modelling :

- Process of creating diagrams that visually represent how data is organised and structured.
- These visual representations are called as Data models

*3 most common types :→

1. Conceptual (Business concepts)
2. Logical (Data Entities)
3. Physical (physical labels)

3 schema Architecture

Common methods :→

Entity Relationship Diagram (ERD) — visual way to understand relationship

Unified Modelling Language (UML) — detailed diagram that describe structure

More about it :

Week 1
4/4

Page No.
Date 09 01 23

* Know the type of your data:

Data type:
A specific kind of data attribute that tells what kind of value the data is.

Data types in spreadsheets →
> Number
> Text or String : sequence of char & punctuation
> Boolean : only 2 possibilities

Some program have different names or include extra types.
But these cover any data you'll find in Spreadsheet.

Text can include number but they aren't useful for calculation

---

(12)

What do these have in common ?
> A music playlist
> A calender agenda
> Email Inbox

All are arrange in tables

Rows = Records

Columns → Fields

Rows and columns are usually reserved for spreadsheets.
Record & Columns are universal.

* Wide & long data :→

Wide data:
Data in which every data subject has a single row with multiple columns to hold the values of various attributes of the objects.

Data subject : year : 2010
2011
:

WIDE DATA

So	country	Y2010	Y2011	Y2012
	A	x	y	z
	B			
	C			

Longdata :→
Data in each row is one time point per subject, so each subject will have data in multiple rows.

Country	Year	Population
A	2010	x
B	2011	y
C	2012	z
A	2011	p
B	2012	q
C	2010	R
A	2012	S
B	2010	T
C	2011	U

Long Data

** Long data is a great format for storing and organizing data when there are multiple variables at each time for for each subject.

↓

Nice & Compact

* But it depend on project which format you should be using.

Learnt: * Identifying bias in data and how to embrace credibility, integrity and ethics.

*Data Transformation :→

Data organization: Better organized data is easy to use

Data Compatibility: Different apps & system can then use the same data

Data Migration : Data with matching formats can be moved from one system to another

Data merging : Data with the same organization can be merged together

Data Enhancement : Data can be displayed with more detailed fields.

Data Comparison : Apples to Apples comparison of data can then be made.

Week 2 / 1/4

Ensuring data integrity

Page No. 10
Date 01 23

* >Analyze data for bias and credibility
: Important because even the most sound data can be skewed or misinterpreted.

* Good data vs Bad data

* Data ethics, privacy and access

We need to ask question like:
(1) Who owns this data?
(2) How much control do we have over the privacy of data?
(3) Can we use & reuse data however we want to?

Bias: A preference in favor of or against a person, group of people or thing.

#Once we know & accept that we have bias, we can start to recognize our own patterns or thinking & learn to manage it.

Data Bias:- A type of error that systematically skews results in a certain direction.

* Bias can happen is sample group lacks inclusivity

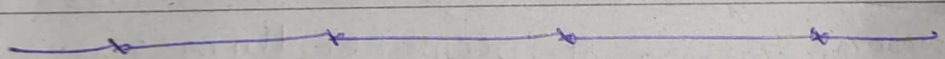
* Sampling Bias :
When the sample isn't representative of the population as a whole.

↓

we can avoid this by choosing the sample at random, so that all parts of the population have a equal chance of being included

* Unbiased sampling :
when a sample is representative of the population being measured.

visualizing the data will reflect if its biased or unbiased.



*more Types of Bias

* Observer Bias :→
Also called as Experimenter / Researcher Bias
The tendency for different people to observe things differently.

① Ex. scientists see different things under microscope

② medical professional see different BP readings. (round off)

4 Types of Data Bias

1. Sampling Bias
2. Observer Bias
3. Interpretation Bias
4. Confirmation Bias

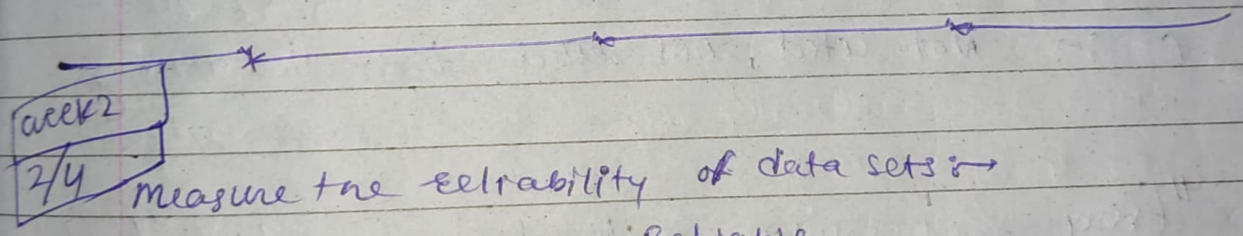
Page No.
Date

* Interpretation Bias:
The tendency to always interpret ambiguous situations in a positive or negative way

→ Two people seeing and hearing the exact same thing and interpreting it in a variety of different ways (due to different background & experiences).

* Confirmation Bias
The tendency to search for or interpret information in a way that confirms pre-existing beliefs.

"People see what they want to see"



week 2
2/4

measure the reliability of data sets :→

ROCCC

- Reliable
- Original
- Comprehensive (contains all the info)
- Current (relevant)
- Cited

When choosing a datasource, think these 3 things:
1. Who created the dataset
2. Is it part of credible organization?
3. When was the data last refreshed?

Good data location: ~~Vetted public datasets, Academic papers,~~ financial data, Governmental Agency data

Data.gov current data
of census Bureau

Bad data sources don't ROCCC.

R :→ Inaccurate, incomplete or biased
could be sampling bias data
may lead to misleading insights

O :→ from 2nd / 3rd party
so you need to be extra careful while
understanding your data

C :→ missing important information needed
to answer the question
may contain human error

C :→ Not current, out of date & irrelevant

C :→ Not cited, not vetted then its a no go.

"Every good solution is found by avoiding bad data"

Week 2 3/4

Data ethics & privacy:-

Page No
Date 11 01 23

Introduction to data ethics:→

Ethics:
Well-founded standards of right and wrong that prescribe what humans ought to do, usually in terms of rights, obligations, benefits to society, fairness, or specific virtues.

Data Ethics:
Well-founded standards of right and wrong that dictate how data is collected, shared and used.

GDPR
General Data Protection Regulation of the European Union

Data protection legislation

Six aspects of data ethics:-

>Ownership
Individuals own the raw data they provide and they have primary control over its usage, how its processed, and how it's shared

>Transaction Transparency:
All data-processing activities and algorithms should be completely explainable and understood by the individual who provides their data.

Let's people judge whether outcome is fair and unbiased & allows to raise concerns.

> Consent :
An individuals right to know explicit details about how and why their data will be used before agreeing to provide it.

Nowadays, it just looks like terms & conditions with just a checkbox.

> Currency :
Individuals should be aware of financial transactions resulting from the use of their personal data and the scale of these trangactions.

Optional!

Alex : Research Scientist with the Ethical AI team

"How do we actually improve the lives of people by using data?" → Don't loose this view

"Data Are people"

Take responsibility of those peoples

Six Ethics Are :→ Ownership Currency
Transaction Transparency Privacy
Consent Openness

Privacy:
Preserving a data subject's information and activity any time a data transaction occurs.

Also called information privacy / data protection.

*It's about access, use & collection of data

Legal Rights to one's data :→

> Protection from Unauthorized access to our private data
> Freedom from inappropriate use of our data
> The right to inspect, update, or correct our data
> Ability to give consent to use our data.
> legal right to access the data

*Openness :→
Free access, usage, and sharing of data

Andrew : senior developer Advocate, ethical AI research group

" One consequence of not using this technology responsibly is the possibility of amplifying or reinforcing unfair biases."

Week 2

Open data :

Free access, usage and sharing of data

But we should still be transparent, respect privacy and make sure we have consent for data that's owned by other

In simple terms, iff data meets these conditions then only we may access, use & share it.

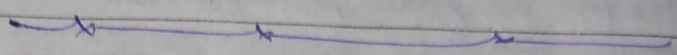
* Availibility & access
Open data must be available as a whole, preferably downloadable over internet in convenient & modifiable form. (ex. data.gov)

* Reuse & Redistribution :→
It should allow reuse & re-distribution also the ability to use it with other data sets.

*Universal participation :→
Everyone should be eligible & there shouldn't be descrimination

Ex. in Health sector, openness allows us to access & combine diverse data to detect deseases earlier & earlier.

Data Interoperability :→
The ability of data systems and services to openly connect and share data.
↓
Requires lot of co-operation



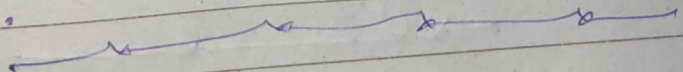
Week 3 | 12/01/23 | Working with datasets.
1/5

All about datasets :→
Database :→
A collection of data stored in a computer system.

meta :→ Referencing back to itself / Being completely self aware.
Metadata : Data about data

A character knows its inside a book
Documentary about making documentary
Analyzing how you analyze data

Once you have data in a spreadsheet the possibilities are endless.



Relational Database :→
A database that contains a series of related tables that can be connected via their relationship.

For Rel^n: one or more fields must exist inside both tables.

Primary key :→
A identifier that references a column in which each value is unique.
↓
Unique, not null or blank

Foreign key ←
A foreign key is a field within a table that is a primary key in another table.

A table can have only 1 Primary key but many foreign keys.

Normalization →
Process of organizing RDB. It is applied to eliminate data redundancy, increase data integrity and reduce complexity in a database.

> Some table's don't require PK.
> A pk can also be constructed using multiple columns of a table its called composite key.

Meta data Example:
① Photo
② mail

"Metadata is used in database management to help data analysts interpret the contents of the data within the database."

3 Types of metadata :→

(1) Descriptive metadata:
metadata that describes a piece of data and can be used to identify it at a later point in time.

Ex. descriptive metadata of book: Title, author and unique international standard Book Number ISBN

(2) Structural metadata :→
Metadata that indicates how a piece of data is organized and whether it is part of one, or more than one, data collection.

Ex. Contents page of book, digital document → printed one

This type of metadata also keeps track of relationship between 2 things.

(3) Administrative metadata :→
metadata that indicates the technical source of a digital asset.

Ex. photo information

* Using metadata as an analyst :→

Putting data into context is probably most valuable thing that metadata does.

Benifits of using metadata :→

single version of truth
> metadata creates a single source of truth by keeping things consistent and uniform.

Reliable
> metadata also makes data more reliable by making sure its accurate, precise, relevant and timely

metadata Repository
Data Analyst use Metadata repository to make sure their data is consistent & reliable :→

=> A database specifically created to store metadata

Metadata repository can be stored in a physical location or can be virtual (cloud)

> metadata repository makes it easier and faster to bring together multiple sources for data analysis.

> Describes state & location of metadata
> describes structure of tables inside
> describes how data flows through repository
> keeps track of who accesses metadata & when

* Metadata management :→

metadata is stored in a single, central location and gives the company standardized information about all of its data.

Data Governance :→

A process to ensure the formal management of a company's data assets.

↓

Gives better control of data & helps company manage issues related to data security, privacy, integrity and usability & internal - external data flows.

* Metadata Analyst *

Megan: measurement lead

"Metadata is kind of a shorthand or a cliffsNotes version of a much more complex set of information."

Data Lake

>Data sources that may overlap & have something in common.