1S 2022-2023

DEVC 127
LESSON 5

TOOLS & TECHNIQUES

RIKKI LEE MENDIOLA
Lecturer, DDJ

today

is your dataset ready?

armed with gsheets

eda

data wrangling

starting your analysis

pulse

- Go to menti.com
- Enter this code: 3481 0842

today in data

We have a volunteer!

pre-req: is your data ready?

- Identify questions first!
- Get metadata.
- Yes, that’s it.

process: is your data ready?

- Dropping unnecessary data
- Dropping duplicates
- Fixing structural errors
- Removing outliers
- Handling missing data
- Dummifying the data
- Normalizing data

repeat: is your data ready?

- It depends.
- Yes, that’s it.

data source checklist

- ❑ **Check for metadata:** who collected it, in what way, how is it used for analysis, how often is it updated/collected
- ❑ **Check for cleanliness:** is it consistent, is test data present, is there a way to standardize
- ❑ **Check for representation:** is it imbalanced, will the distribution affect analysis, is there a way to normalize it
- ❑ **Check for completeness:** are there missing values, is there a way to derive missing values, will dropping missing values affect analysis
- ❑ **Check for complements:** will this dataset better understood with another dataset, can I combine it using data points from existing data

armed with gsheets

- Standardizing cell formatting
- Using freeze panes
- Searching
- Sorting
- Filtering
- Add/delete column
- Operations on numerical data
- Operations on datetime data
- Operations on string/char data
- Creating pivot tables

let’s practice with real world data

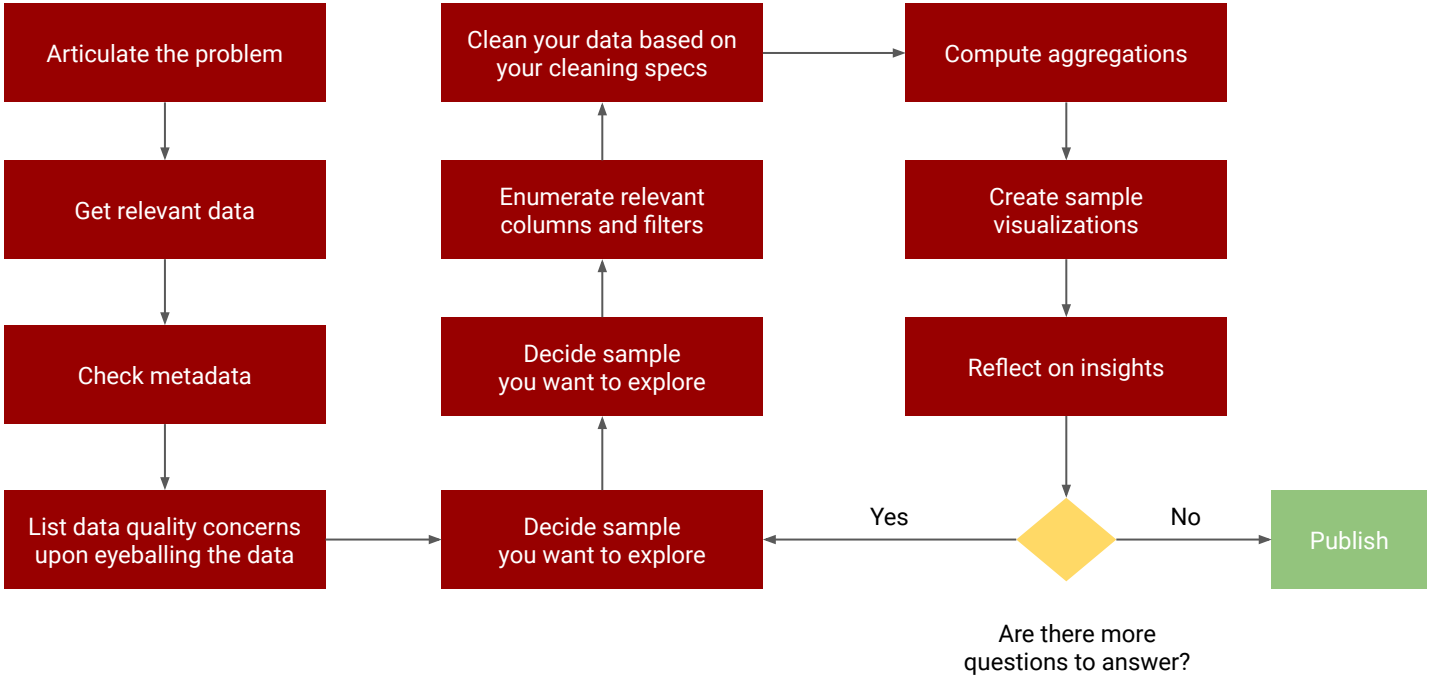
Let’s handle dirty data, consumer loans data
We’ll work on this dataset for two weeks

what’s first thing you asked your data?

exploratory data analysis (eda)

- discover trends, patterns
- check assumptions
- statistical summaries of your data
- quick visualizations
- quality inspection is part of eda
- **eyeball** the data
- discovery, structuring, cleaning, enriching, validating,

exploratory data analysis (eda)



1S 2022-2023

DEVC 127
LESSON 5

TOOLS & TECHNIQUES

RIKKI LEE MENDIOLA
Lecturer, DDJ

today

pulse

today in data

data wrangling

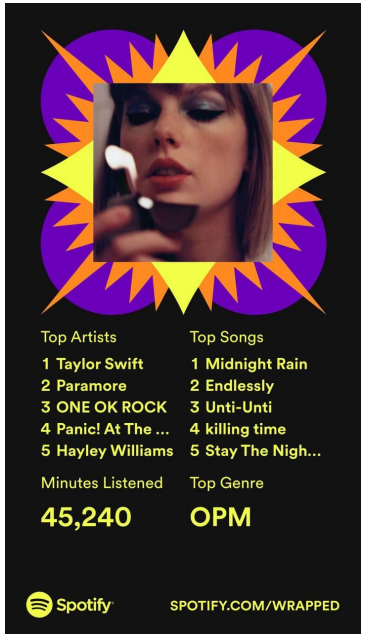
starting your analysis

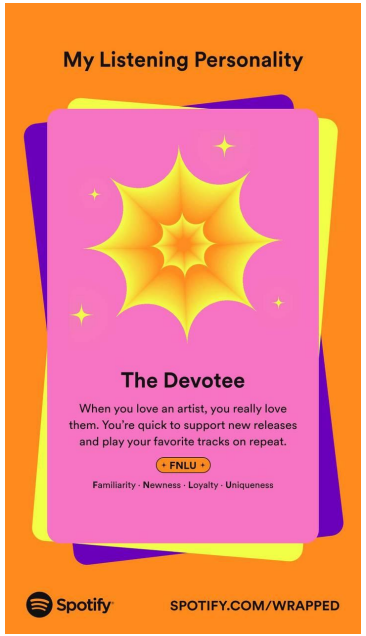
next week

pulse

- Go to menti.com
- Enter this code: 4898 1392

today in data





45,240 minutes, 1,751 songs, 706 artists of an angsty and hopeful 30-something

useful techniques in data wrangling

- Bucketing and dummifying
- Measures of central tendencies
- Central limit theorem
- Null values to interpolate, drop, or 0
- Computing metrics horizontally
- Computing metrics vertically

ID	CustomerID	StartBalance	Amount
T1000000001	A0001	500.00	20
T1000000002	A0002	5.00	20

ID	CustomerID	StartBalance	Amount	Red
T1000000001	A0001	500.00	20	0
T1000000002	A0002	5.00	20	1

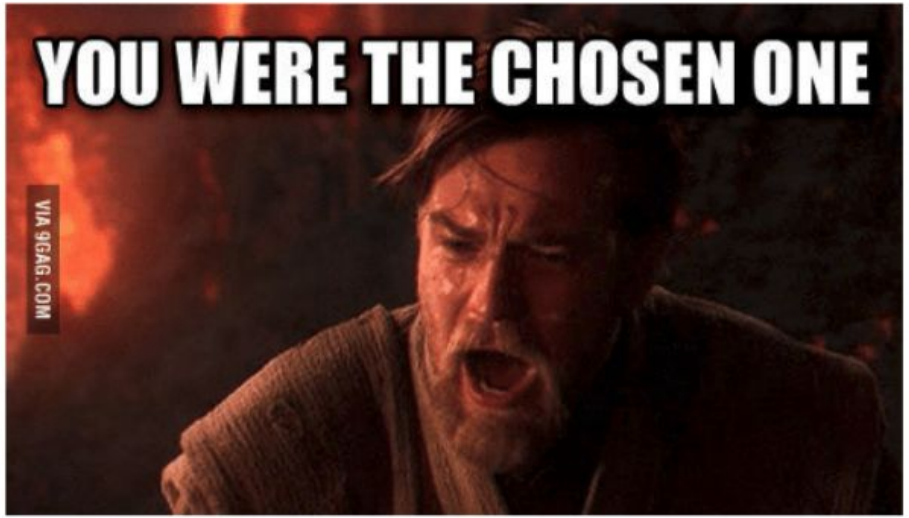
Red Customer tries to transact with no enough balance.
Condition StartBalance < Amount
Aggregate Count how many times there are red listed transactions

ID	Time	CustomerID	Type	Amount	RT
T100000	2021-10-29 11:05:15	A0001	Transaction	1000	null
T100001	2021-11-01 00:20:30	A0001	Reload	1000	2.55
T100002	2021-11-01 00:20:30	A0002	Transaction	20	null
T100003	2021-11-14 14:03:59	A0001	Transaction	100	null
T100004	2021-11-15 09:02:10	A0002	Transaction	20	null
T100005	2021-11-17 08:40:02	A0002	Transaction	20	null

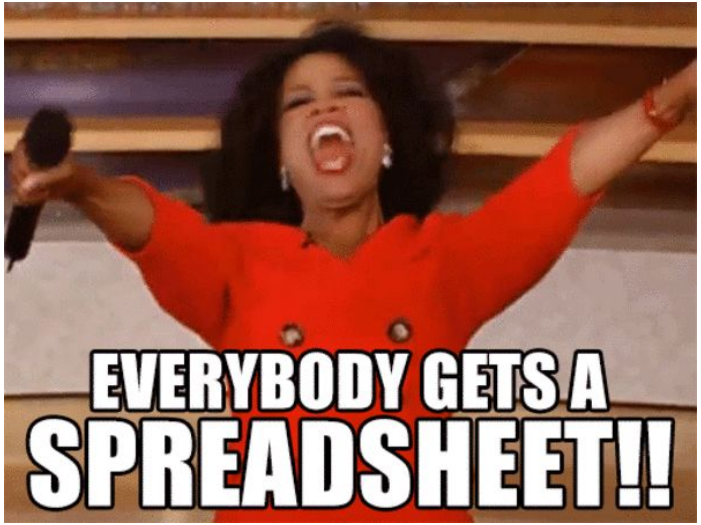
R/T interval How long will it take for a customer to reload before transacting again?
Condition Per customer, if type = transaction before reload, Time (2) - Time(1)
Aggregate Average intervals

to drop or not to drop

When you put a lot of work into cleaning up and improving a data set and the result is worse



data is never clean.



access your worksheets here:
https://tinyurl.com/devc-127-lesson5

no sophisticated statistical model will save a bad dataset.

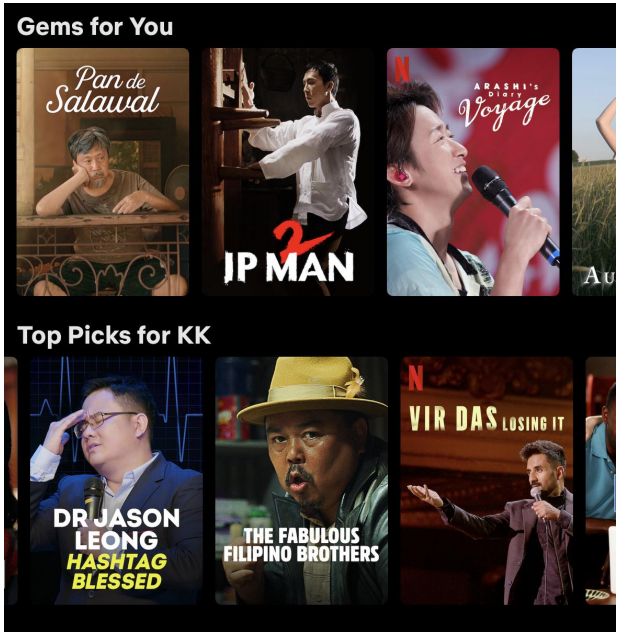
where do i start?

how do you know if counts are enough?

how do you know if average should be used?

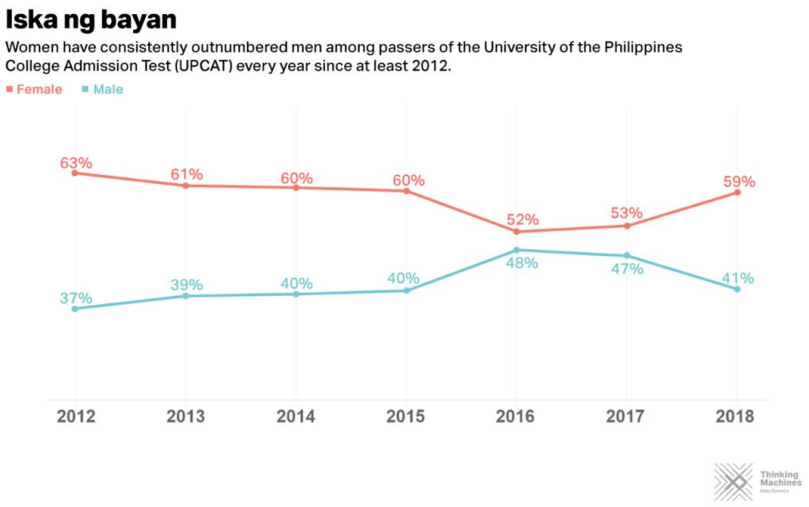
i want to learn about a phenomena by looking at factor a and factor b.

- Correlation
- Not necessarily means causation
- -1, 0, 1
- If 2 variables do not deviate from the mean in any meaningful pattern, little to no correlation
- Too strong a correlation may be overfitting



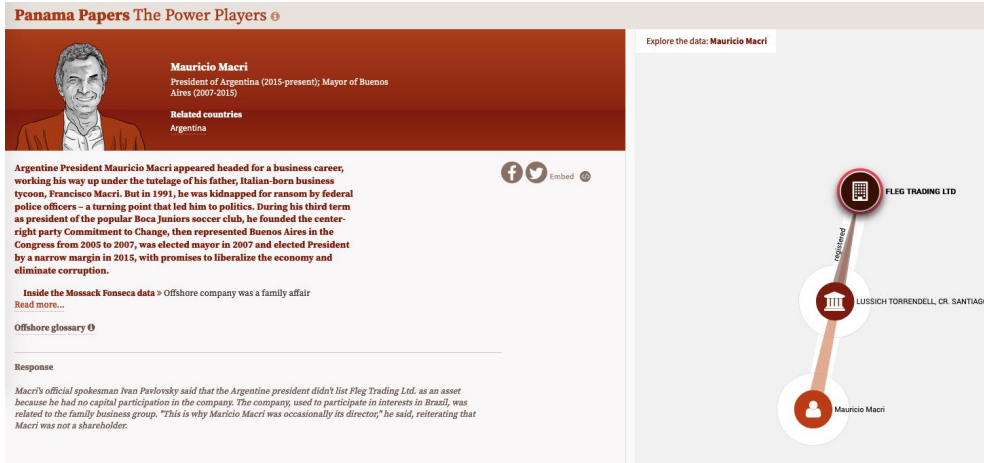
i want to learn about the future of a by studying its past.

- Regression analysis or moving averages
- One variable, over time
- Usually deals with volume, counts
- Time is an important element
- Historical data is important
    - Establish patterns and trends
    - Use that in predicting future values



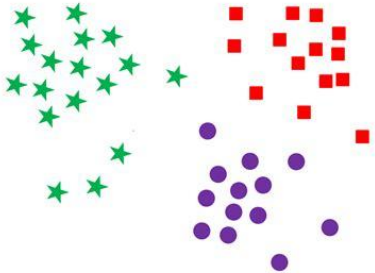
i want to learn about a's relationships to others-who's who?

- Network analysis
- Node, edge
- Best used to examine relationships
- Check out this story about the [Power Players in the Panama Papers]



i want to learn about a and its peers.

- Clustering
- Tell me who your friends are and i'll tell you who you are
- Divide population in groups
  - Learns patterns from the data
  - Figuring out similarities

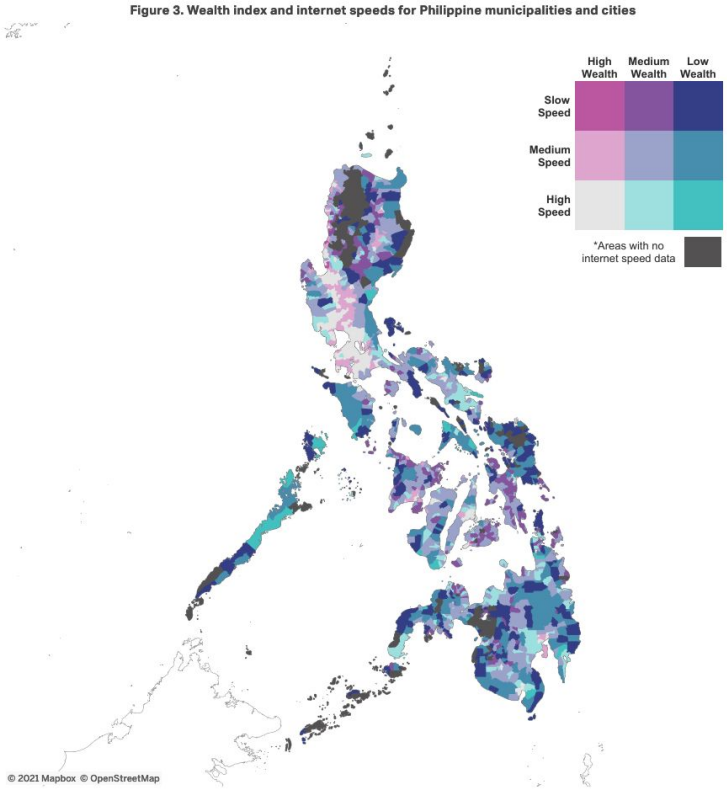


i want to learn about a's activities.

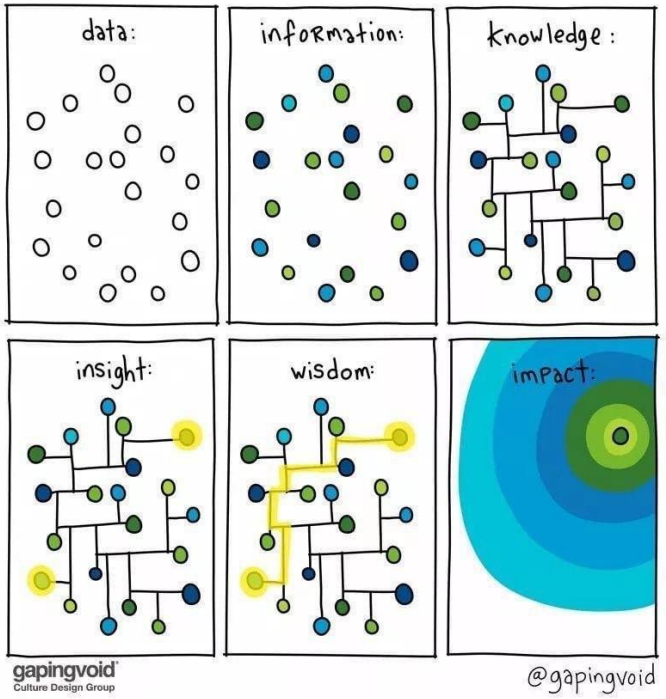
- Market basket analysis
- Association rules
- Works with probabilities
- Needs large volume of data to work
- If customer buys A, s/he also buys B
- Check out Reina Reyes' work on the past presidential elections

i want to learn about a across locations

- Geospatial analysis
- Location data is crucial
- Descriptive stats grouped by location
- Check out this interesting data story from TM on [digital poverty]

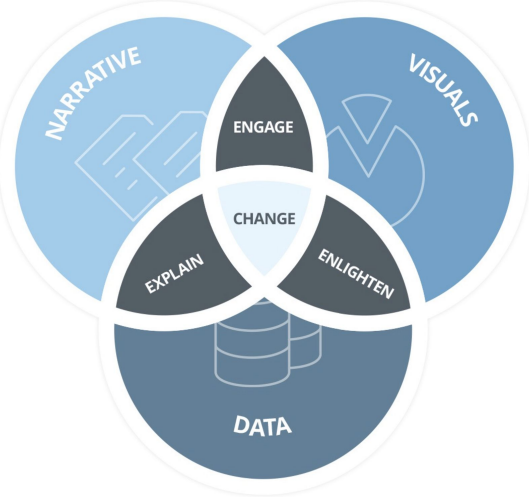


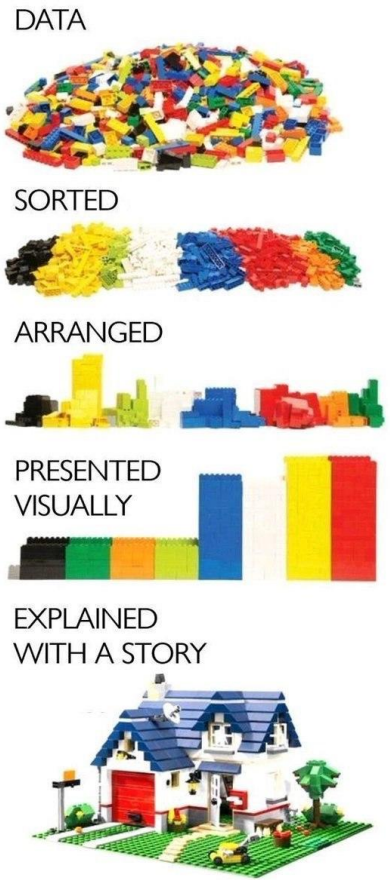
questions?
clarifications?



data storytelling

- A structured approach for communicating data insights
- Three key elements: Data, Visuals, Narrative





next week

- Release Lesson 6 and 7
- Section T sync: TBA, Th is a holiday
- Next week: more data stories (+ my analysis on consumer loans) and visualization techniques
- Consultation: email for availability, will open slots to discuss lecture project and cleaning/analysis techniques

1S 2022-2023

DEVC 127
LESSON 5

TOOLS & TECHNIQUES

RIKKI LEE MENDIOLA
Lecturer, DDJ