1S 2022-2023

DEVC 127
LESSON 4/5

DATA SOURCES / TOOLS & TECHNIQUES

RIKKI LEE MENDIOLA
Lecturer, DDJ

today

fireside chat with former Amihan DSE

weekly reminders

ian jae abella

senior data scientist
bda dev and innovations
metro pacific tollways corporation

- UP Los Baños, BS Applied Mathematics
- 3 years experience in data analytics
- Previously worked with Rarejob as data analyst
- Also a T.A. at a graduate school course teaching Data Science

joshua villanueva

data pipelines architect
technology, strategy and transformation
coca-cola beverages phils. inc

- UP Manila, BS Computer Science major in Statistical Computing
- Data Science enthusiast
- Certified Kubernetes Administrator
- 2+ years experience in Big Data
- Also a T.A. at a graduate school course teaching Data Science

what we do before as data science and engineering team

- We are a 5-person team composed of data analysts and engineers
- We help our clients through their digital transformation
- We create and operationalize analytics use cases using our Amihan Analyze big data stack
- We work on various use cases for insurance, banks, telecomms, utility companies
- Analytics use cases: forecasting, customer analytics, computer vision, natural language processing, contact tracing
- Our analytics toolkit: Zeppelin and Superset for data exploration and viz, nifi and kafka for ingest and automation, various dbms depending on the use case

why did you pursue data science/analytics?

Scenario: You are working on an analytics use case and you need to find a dataset.

what's your data collection/ acquisition process?

how can you say that the data is good enough?

Scenario: You already have a data set.

what's your data cleaning process?

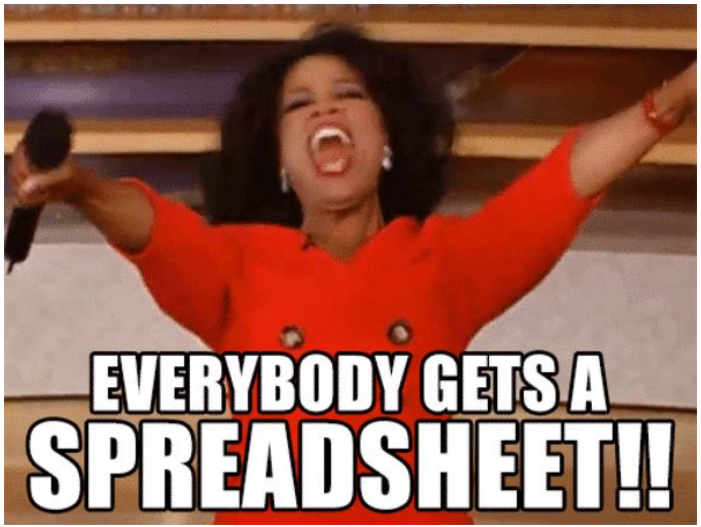
what steps do you prioritize? how do you know if it's clean enough?

most challenging
data cleaning experience

most interesting
data analytics/science
use case

most useful
spreadsheet tool

most important lesson
you learned while working
in ds in tech



questions?
clarifications?

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

this week

- Watch a rerun of the DDJ seminar on Data Cleaning
- Live tutorial (Th/F)
- I'll post tutorials with practice worksheets and datasets

1S 2022-2023

DEVC 127
LESSON 4/5

DATA SOURCES / TOOLS & TECHNIQUES

RIKKI LEE MENDIOLA
Lecturer, DDJ