

# Workshop

# Agenda

|               |   |
|---------------|---|
| 10:00 - 10:20 | Ice Breaking                              |
| 10:20 - 11:10 | SQL                                       |
| 11:10 - 12:00 | Python                                    |
| 12:00 - 12:15 | Data Understanding +<br>Announce Question |
| 12:15 - 13:00 | Lunch                                     |
| <hr/>         |   |
| 13:00 - 15:00 | Workshop                                  |
| 15:00 - 16:00 | Presentation                              |
| 16:00 - 16:30 | Feedback + Ending                         |



# Get in to your groups!

## A

1. นิเวศ
2. นาย
3. ขุนเขา
4. กัน
5. (ณพวุฒิ)

## B

1. ปิ่น
2. บั๊ต (AI)
3. บิว
4. ชักก้า
5. โบว์

## C

1. บั๊ต (IE)
2. กิฟ
3. มอส
4. จูน
5. แซก

## D

1. ต้นข้าว
2. ปาล์ม
3. ฟลุ๊ค
4. เอม
5. พล

## E

1. อิคคิว
2. ภัค
3. จั๊บจ๊ับ
4. เว็บ
5. เต้
6. เซฟ

## F (Online)

1. ออมสิน
2. กันย์
3. (ธนานพ)
4. ตัว
5. อีส
6. เบนซ์

## G (Online)

1. ร็อกกี้
2. ไอซ์
3. ภูมิ
4. เหมยลี่
5. ชีซัน

## H (Online)

1. ฟิล์ม
2. ตะวัน
3. คิลต้า
4. สิงโต

# Let's play bingo

**Studied SQL**

**Did Kaggle  
Project**

**Subscribe to  
Datacamp**

**Not in  
Engineering  
Faculty**

**A member of  
Data Track** ❤️

**Studied  
Python**

**Use Google  
Colab**

**Looking for  
Data  
Internship**

**Interested in  
ML**

# Introduction to Facilitators



อิศรุต์ม์ สังข์สวน  
Associate Director,  
INFINITAS by Krungthai

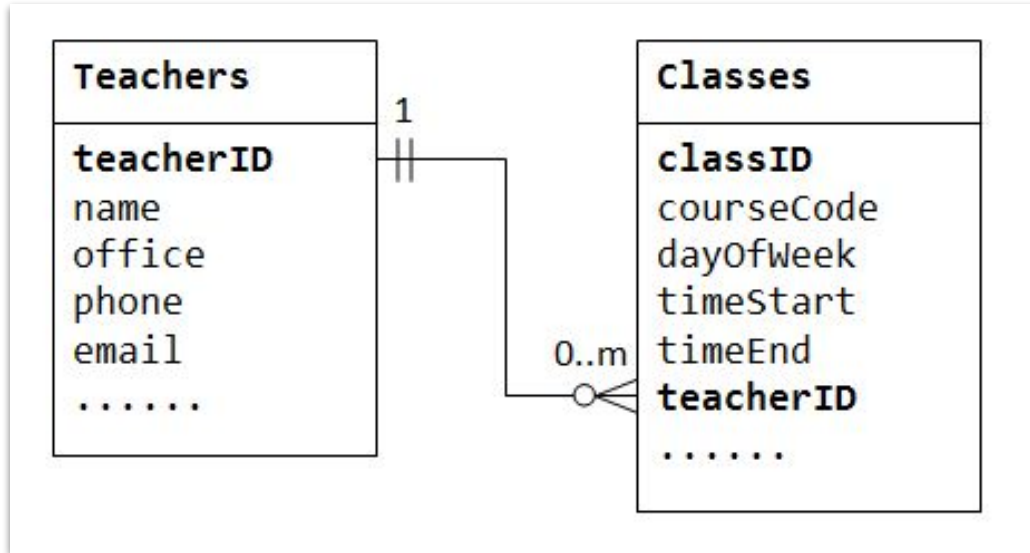


ปัจุบัน โหระชัยยะ  
Data Analytics,  
Disney

# 1. Extracting data using SQL

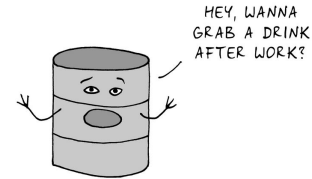
# What is SQL?

**Structured query language (SQL)** is a programming language for storing and processing information in a **relational database**.



- Stores information in tabular form
- Represent different data attributes
- Show relationships between the data values

## RELATIONAL DATABASE



Dataedo /cartoon

Piotr@Dataedo

**SELECT \***  
**FROM cases**  
**LIMIT**

Basic

Conditional

Grouping

Ordering

CTEs

Join

  
**AMP**  
neering



Add one slide to teach about basic  
aggregation  
E.g. SUM(),COUNT(),MAX()

Basic

Conditional

Grouping

Ordering

CTEs

Join

  
**AMP**  
neering

```
SELECT *  
FROM cases  
WHERE state = 'finish'
```

Basic

Conditional

Grouping

Ordering

CTEs

Join

  
**AMP**  
neering

```
SELECT *  
, CASE  
    when star >= 4 then 'Happy'  
    ELSE 'Not Happy'  
END AS 'Rating'  
FROM cases  
WHERE state = 'finish'
```

Basic

Conditional

Grouping

Ordering

CTEs

Join

  
**AMP**  
neering

```
SELECT *  
FROM log  
WHERE state = 'reported'  
OR state = 'finished'  
AND timestamp > TIMESTAMP '2024-01-01 00:00:00';
```

Basic

Conditional

Grouping

Ordering

CTEs

Join

  
**AMP**  
neering

```
SELECT
district
,COUNT(DISTINCT ticket_id) AS ticket_count
FROM cases
GROUP BY district
```

Basic

Conditional

Grouping

Ordering

CTEs

Join

 **AMP**  
neering

**SELECT**

**type**

**SUM(STAR) AS sum\_star,**

**MAX(STAR) AS max\_star**

**FROM cases**

**GROUP BY type**

Basic

Conditional

Grouping

Ordering

CTEs

Join

  
**AMP**  
neering

```
SELECT  
district  
,COUNT(DISTINCT ticket_id) AS ticket_count  
FROM cases  
GROUP BY district  
ORDER BY ticket_count
```

Basic


Conditional

Grouping

Ordering

CTEs

Join

  
**AMP**  
neering

```
WITH  
CNT_TICKET AS  
  (SELECT  
    district  
    ,COUNT(DISTINCT ticket_id) AS ticket_count  
  FROM cases  
  GROUP BY district)
```

```
SELECT * FROM CNT_TICKET  
WHERE ticket_count > 100
```



WITH  
CASES AS

(SELECT \* FROM cases)

, LOG AS

(SELECT \* FROM timelog)

SELECT \*

FROM CASES c

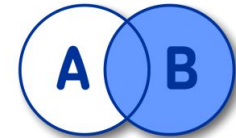
LEFT JOIN LOG t

ON c.ticket\_id = t.ticket\_id

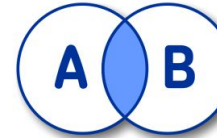
## SQL JOINS



SELECT \* FROM  
A **LEFT** JOIN B  
ON A.KEY = B.KEY



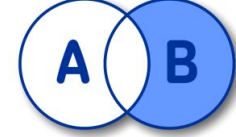
SELECT \* FROM  
A **RIGHT** JOIN B  
ON A.KEY = B.KEY



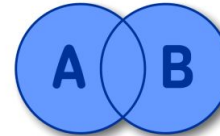
SELECT \* FROM  
A **INNER** JOIN B  
ON A.KEY = B.KEY



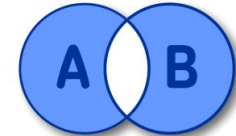
SELECT \* FROM A  
**LEFT** JOIN B  
ON A.KEY = B.KEY  
WHERE B.KEY IS NULL



SELECT \* FROM A  
**RIGHT** JOIN B  
ON A.KEY = B.KEY  
WHERE A.KEY IS NULL



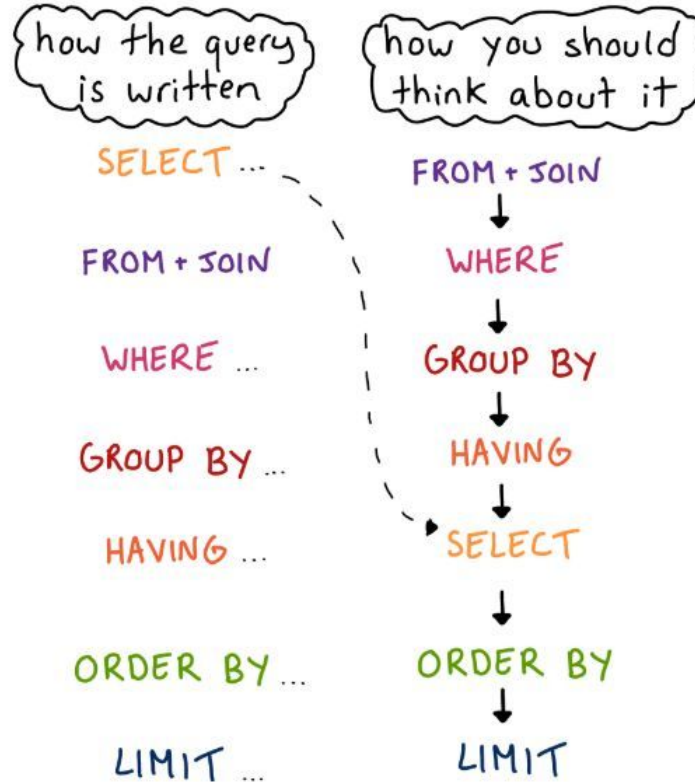
SELECT \* FROM A  
**FULL OUTER** JOIN B  
ON A.KEY = B.KEY



SELECT \* FROM A  
**FULL OUTER** JOIN B ON A.KEY =  
B.KEY WHERE A.KEY IS  
NULL OR B.KEY IS NULL

# SQL

The query's steps don't happen in the order they're written:



(In reality query execution is much more complicated than this.  
There are a lot of optimizations.)

# Your Turn!

Look at the status and timestamp columns.  
What are the durations between each status?  
Create a table with the derived columns

**ANSWER!!!!**



He's making a database.  
He's sorting it twice.  
SELECT \* FROM contacts WHERE  
behavior = 'nice'  
SQL CLAUSE is coming to town!

# **2. Data Cleaning and Visualization with Python**

# Colab Notebook

This part of the workshop will be conducted on Colab Notebook.

Please **make a copy**

Link:

<https://colab.research.google.com/drive/1sCfp9Hxy9zbnezvvxuEWVtlvvGr4Vbwg?usp=sharing>



# When making an analysis ... Define the Purpose

## 1. Define the Purpose

- What is the **goal of the analysis** or visualization? (e.g., to inform, persuade, explore, or explain)
- What specific questions are you trying to answer? **Form a hypothesis** to structure your analysis
- Audience: **Who will be viewing** this? (e.g., technical experts, business stakeholders, general public)





# When making an analysis ...

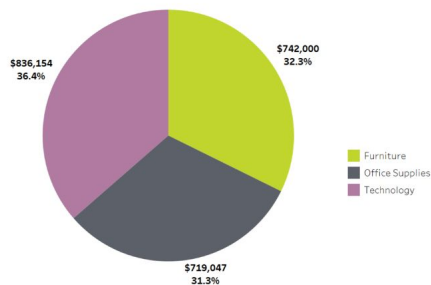
## 2. Select the Right Visualization

- Chart Type: Choose a visualization that best represents the data and answers the question (e.g., bar chart, line graph, scatter plot, heatmap, etc.).
- Simplicity: Avoid clutter and focus on the most critical information. Use consistent colors, scales and labels
- Comparisons: Highlight trends, patterns, or outliers to make the data more meaningful.

# When making an analysis ... **Select the Right Visualization**

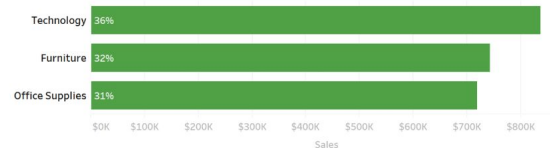
Choose a visualization that best represents the data and answers the question (e.g., bar chart, line graph, scatter plot, heatmap, etc.).

sales by product category



! ineffective

sales by product category



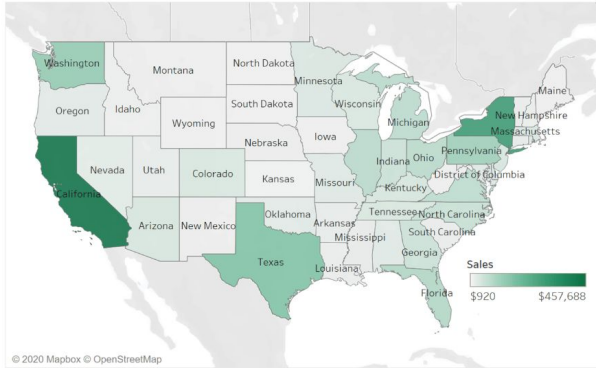
✓ effective

Credit: <https://blog.csgsolutions.com/6-tips-for-creating-effective-data-visualizations>

# When making an analysis ... **Select the Right Visualization**

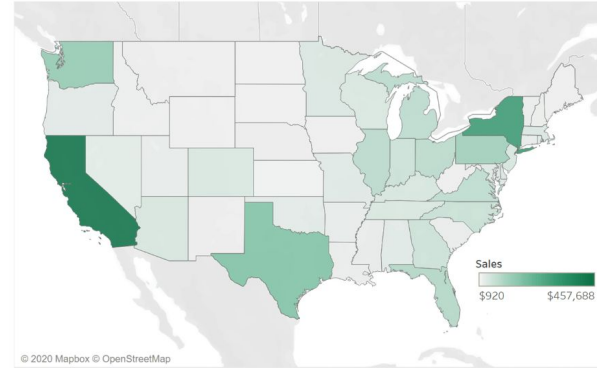
Simplicity: Avoid clutter and focus on the most critical information. Use consistent colors, scales and labels

total sales map



! ineffective

total sales map



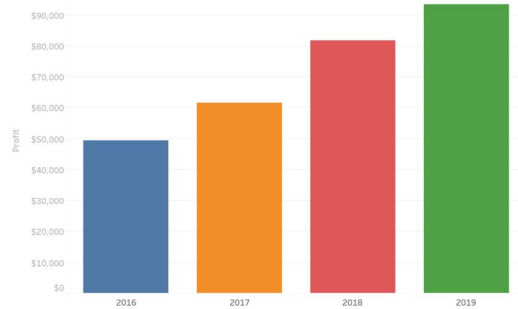
✓ effective

Credit: <https://blog.csgsolutions.com/6-tips-for-creating-effective-data-visualizations>

# When making an analysis ... **Select the Right Visualization**

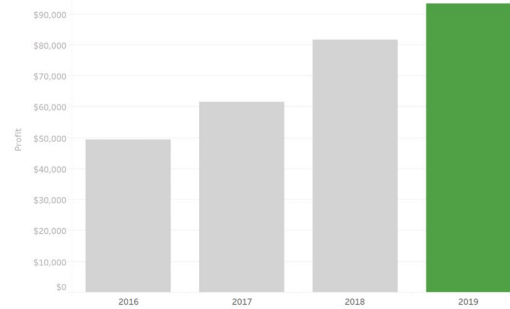
Comparisons: Highlight trends, patterns, or outliers to make the data more meaningful.

profit by year



! ineffective

profit by year



✓ effective

Credit: <https://blog.csgsolutions.com/6-tips-for-creating-effective-data-visualizations>

# 3. Understanding the data

# พลิกโอมให้เมืองน่าอยู่

แพลตฟอร์มบริหารจัดการปัญหาเมือง



แจ้งเหตุผ่าน Line@

- 1 พิมพ์ปัญหาที่พบ
- 2 ถ่ายภาพเหตุการณ์
- 3 เลือกประเภทปัญหา
- 4 แשרตำแหน่งที่เกิดเหตุการณ์
- 5 เลือกหน่วยงานที่ต้องการแจ้ง

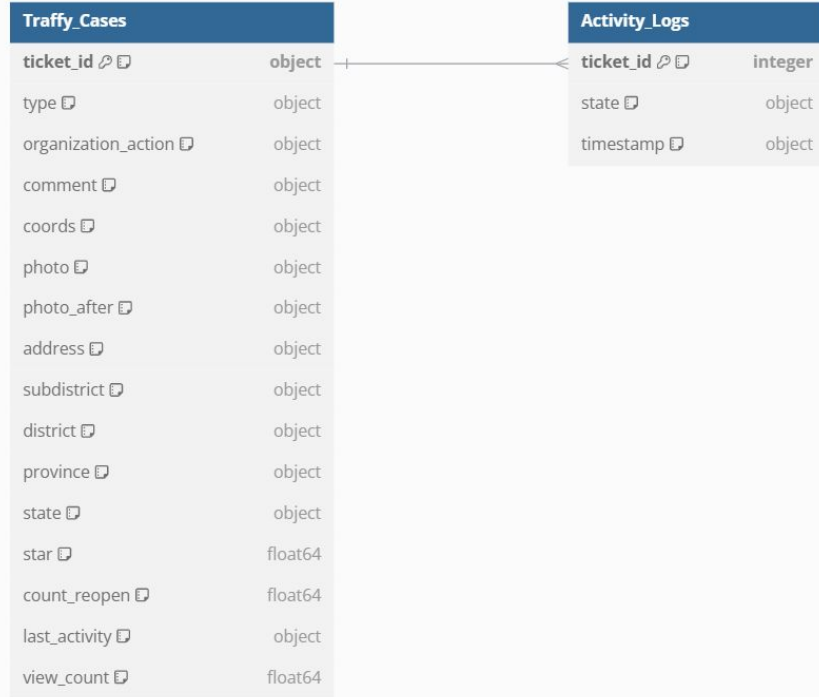
**\*\* ประชาชนแจ้งปัญหาได้ที่ Line@ เท่านั้น! \*\***

1/2 >

## Understand Data Producer To Understand the Data

Credit: <https://bangkok.traffy.in.th/>

# Traffy Fondue – ER DIAGRAM



- **2 Tables**
- **Key is ticket\_id**

# Traffy Fondue – Data Sample

| ticket_id   | type          | organization   | action | comment   | coords             | photo   | photo_after   | address   | subdistrict | district | province      | state   | star | count_reopen | last_activity          | view_count |
|-------------|---------------|--|--------|---|--------------------|---|---|---|-------------|----------|---------------|---------|------|--------------|------------------------|------------|
| 2024-H9XB73 | ความ<br>สะอาด | ฝ่ายรักษาความสะอาดฯ เขต<br>ดุสิต, เขตดุสิต,<br>กรุงเทพมหานคร |        | มีคอนกรีต<br>ขอบถนน<br>ทาง ที่จอดรถ<br>หลังบ้าน | 100.50548,13.77436 | <a href="https://storage.googleapis.com/traffy_public_bucket/attachment/2024-12/91c32d091361ed9ae945f99414f92835c345b4e4.jpg">https://storage.googleapis.com/traffy_public_bucket/attachment/2024-12/91c32d091361ed9ae945f99414f92835c345b4e4.jpg</a> | <a href="https://storage.googleapis.com/traffy_public_bucket/attachment/2024-12/c38163684d95ea9410af98572a9be593.png">https://storage.googleapis.com/traffy_public_bucket/attachment/2024-12/c38163684d95ea9410af98572a9be593.png</a> | 138/8 ถ. สามเสน<br>แขวงวชิรพยาบาล<br>เขตดุสิต<br>กรุงเทพมหานคร<br>10300 ประเทศไทย | วชิรพยาบาล  | ดุสิต    | กรุงเทพมหานคร | finnish | 4.0  | 0            | 2024-12-11<br>08:59:22 | 1          |

Traffy\_Cases Table

| ticket_id | state      | timestamp           |
|-----------|------------|---------------------|
| N4YZH7    | reported   | 2024-12-10 00:02:42 |
| N4YZH7    | inprogress | 2024-12-10 11:20:42 |
| N4YZH7    | finish     | 2024-12-12 16:30:42 |

Activity\_Log table



# Traffy Fondue - Data Dictionary

## Traffy\_Cases

| Column Name         | Descriptions  | Possible Values  |
|---------------------|---|--|
| ticket_id           | unique id สำหรับแต่ละ cases   | 2024-6RLL8H  |
| type                | ประเภทปัญหาที่ถูกรายงานในระบบ   | ไฟฟ้า  |
| organization_action | หน่วยงานที่แก้ปัญหาในเคสนั้น โดยเรียงลำดับจากซ้ายไปขวา (หน่วยงานที่แก้ปัญหาล่าสุดอยู่ซ้ายสุด) | กรมทางหลวง, เขตจตุจักร, Bangkok Smart Lighting (สำนักการโยธา กทม.), กรุงเทพมหานคร                                  |
| comment             | ข้อความอธิบายปัญหาที่ประชาชนแจ้งเข้ามา  | ไฟบนสะพานลอยหน้า ม.เกษตรเลีย   |
| coords              | พิกัดตำแหน่งของปัญหา (longitude, latitude)  | 100.57067,13.84226   |
| photo               | รูปปัญหาที่แจ้ง ส่งจากผู้แจ้งเรื่อง   | https://storage.googleapis.com/traffy_public_bucket/attachment/2024-10/65295639c215fa565fad21c7155382f9798fc97.jpg |
| photo_after         | รูปหลังจากแก้ปัญหา ส่งจากเจ้าหน้าที่  | https://storage.googleapis.com/traffy_public_bucket/attachment/2024-10/7b353aa0480c3a1557bde5edeecac0ab.jpg        |
| address             | ที่อยู่ของปัญหาที่ถูกรายงานเข้ามา อ้างอิงจาก coords   | 117 ถนน จอมวงศ์วาน แขวงลาดยาว เขตจตุจักร กรุงเทพมหานคร 10900 ประเทศไทย   |
| subdistrict         | ชื่อแขวงโดยได้จากการนำ coords ไปหาชื่อแขวง Google Reverse Geocoding                           | ลาดยาว   |
| district            | ชื่อเขตโดยได้จากการนำ coords ไปหาชื่อเขต Google Reverse Geocoding                             | จตุจักร  |
| province            | ชื่อจังหวัด   | กรุงเทพมหานคร  |
| state               | สถานะของปัญหา   | finish   |
| star                | จำนวนคะแนน feedback จาก user หลังแก้ปัญหาเสร็จ (เต็ม 5)                                       | 2.0  |
| count_reopen        | จำนวนครั้งที่เปิดเรื่องใหม่หลังจากปิดเคส  | 0  |
| last_activity       | timestamp กิจกรรมล่าสุดของ case   | 2024-10-21 10:02:50  |
| view_count          | จำนวนการเข้าชมเคสบน https://bangkok.traffy.in.th/   | 2  |

## Activity\_Logs

| Column Name | Descriptions                  | Possible Values                |
|-------------|-------------------------------|--------------------------------|
| ticket_id   | unique id สำหรับแต่ละ cases   | 2024-6UUBZ2                    |
| state       | สถานะของ activity ที่เกิดขึ้น | reported, inprogress, finished |
| timestamp   | timestamp ของ log             | 2024-12-11 20:14:18            |

For full details:

<https://drive.google.com/drive/folders/1olykzLnEmqsjhrouhFtM7GjN4advFRu>

# Traffy Fondue – Data Sample

| ticket_id   | type      | organization       | action                            | comment  | coords             | photo   | photo_after   | address   | subdistrict | district | province      | state  | star | count_reopen | last_activity          | view_count |
|-------------|-----------|--------------------|-----------------------------------|--|--------------------|---|---|---|-------------|----------|---------------|--------|------|--------------|------------------------|------------|
| 2024-H9XB73 | ความสะอาด | ฝ่ายรักษาความสะอาด | เขตดุสิต, เขตดุสิต, กรุงเทพมหานคร | มีคณาภิ๋ย<br>ขอแบบถนน<br>ทาง ที่จอดรถ<br>รถที่บ้าน | 100.50548,13.77436 | <a href="https://storage.googleapis.com/traffy_public_bucket/attachment/2024-12/91c32d091361ed9ae945f99414f92835c345b4e4.jpg">https://storage.googleapis.com/traffy_public_bucket/attachment/2024-12/91c32d091361ed9ae945f99414f92835c345b4e4.jpg</a> | <a href="https://storage.googleapis.com/traffy_public_bucket/attachment/2024-12/c38163684d95ea9410af98572a9be593.png">https://storage.googleapis.com/traffy_public_bucket/attachment/2024-12/c38163684d95ea9410af98572a9be593.png</a> | 138/8 ถ. สามเสน<br>แขวงวังสราญบาล<br>เขตดุสิต<br>กรุงเทพมหานคร<br>10300 ประเทศไทย | วังสราญบาล  | ดุสิต    | กรุงเทพมหานคร | finish | 4.0  | 0            | 2024-12-11<br>08:59:22 | 1          |

Traffy\_Cases Table

# Traffy Fondue – Data Sample

| ticket_id | state      | timestamp           |
|-----------|------------|---------------------|
| N4YZH7    | reported   | 2024-12-10 00:02:42 |
| N4YZH7    | inprogress | 2024-12-10 11:20:42 |
| N4YZH7    | finish     | 2024-12-12 16:30:42 |

Activity\_Log table

**4. Let's do your own  
analysis!**

# Choose 1 track, find insights, and prepare to present

1. Are there difference in user behavior between contact channels?  
How to better serve each group? (2+1)
2. Which problems can be solved quickly but have not been solved?  
Can you pinpoint the reason why? (2+1)
3. What are the reason behind satisfaction scores (star)?  
What is your suggestion to กนกน.? (2+1)

**Time's up at 15.00!**

# Don't forget to do slides!

15 mins left  
3-5 pages are fine

# Presentation

1 hour

# Introduction to Commentators



คุณานพ เลิศไพรวลัย  
Assistant Secretary to  
Governor of Bangkok



ดร.วสันต์ ภัทรอริคม  
Vice President & Director of  
City Innovation Division, NSTDA



# Overall Feedback

# Publish your findings!

วิเคราะห์ข้อมูลปัญหาภายในกรุงเทพแบบซ้ำซ้ำ (Traffy) | by Medium

# Reminder!

DE workshop is live next week! If you are interested in Docker, Airflow, and Database, you still have a chance to join!

# 1. Understanding and Solving a Business Challenge

# Understanding and Solving a Business Challenge

- Explain the business problems we have
- Separate teams into respective business questions
- Brainstorm session: What data do we need to solve this issue?
- Release the dataset for them

# Extracting data using SQL

- Basic SQL lesson (DBeaver Community)
  - Select
  - Where
  - Aggregate func. (sum, max,min, etc)
  - Group by
  - Etc.
- Extract data from given dataset

# Data Analysis with Python

- Clean data (Google Colab)
  - Empty cells (drop na, fill na)
  - Data in wrong format
  - Wrong data
  - Duplicates
- Analyse data
  - Hypotheses??
  - Trends over time
  -