</talentlabs>

Data Exploration Framework Overview

There are many different data exploration frameworks in the industry, and there are no right or wrong in which one to use. In this project, we are going to adopt a 5-steps framework which included:

1. Understanding the Business Context
2. Understanding the Technical Context
3. Understanding the Tables & Fields
4. Free Exploration

You would find that you are spending a lot of time on understanding the database and the background of it. Indeed, writing and running SQL is just a small part of a data analyst day-to-day. They actually spend most of their time designing, understanding and studying the data.

Step 1: Understanding the Business Context

When you get a new dataset, the first step to perform any analysis is really to understand the context and background of the database. Before you do anything technical, you will need to do some research (could be Google search, or talk to the data provider/owner) on the business context first. Otherwise, you will not be able to comprehend and interpret the data correctly.

You can try understanding the business context of the data by answering the questions below:

- What are these data for?
- Why do we need this database?
- Where are these data collected?

Step 2: Understanding the Technical Context

After you get some understanding on the business side, then you would want to start looking into the technical aspects of the dataset. Getting some sense on the technical aspects would help you in interpreting the data, and get some sense on the accuracy and reliability of the data.

To perform this technical analysis, you can try answering the following:

- How are these data collected?
- Where are the sources of these data?
- Is the data coming from surveys, or some computer system? Is it manually input by some data entry personnel or collected by some electronic system?
- What are the systems that touch or use/modify these data?
- What are some of the error sources of this data?
- Is the data complete? Would there be missing pieces of data?

Step 3: Understanding the Tables and Fields

In this step, we are going to further deepdive into the dataset by looking at the tables. This is the first step that you actually need to interact with the database (by opening it up in SQLite or other database software). The goal of this step is to find out what are the data tables available in the database, and what's the type of data inside.

Some guiding questions for this step would be:

- How many tables do we have?
- What are the tables? and what are these tables representing?
- What are the relationships between the tables?
- What are the fields in the tables? What is the meaning of each of the field?
- Is the data messy? and how?
- Should I clean the data first? or ignoring those messy columns
    - Ignoring the empty data fields when performing analysis by:
      `select * from table where column_x is not null`
    - Check how many missing data by:
      `select count(*) from table where column_x is null or column_x = 0`

Step 4: Free Exploration

This is the most interesting part of the data analysis process, in which you would get a chance to really study the data, and come up with some useful and insightful conclusions.

Actually, there are two sub-steps in this part:

a. Coming up with research questions

You will need to come up with a few interesting questions that you want to find answers from the data.

For example, if you are studying an eCommerce database, your research question could be "What is the most popular product in the online store?".

Alternatively, if you are studying a car database, then you might want to study "Which car is the fastest car in terms of acceleration?".

Some more examples of research questions:

- What is the average age of users?
- What is the total number of users?
- How long do the users go to our platform?
- What is the most popular facebook post in Malaysia?
- What is the most boring facebook post in Malaysia?
- …. (Use your creativity!)

b. Answering the research questions with SQL

After coming with the research questions, then you would want to start using SQL to dig out the answers from the data. If we are trying to answer the question of “most popular product”, then I might be using a SQL like `select product, count(*) from transactions`.

In general, you will be using a lot of aggregations and group-bys, i.e. avg, sum, max, min. You might need to use some sortings (order by) and limit too. Basically, you will need to use everything you learnt about SQL to answer the questions.

Once you find out the answer, you will need to start writing up some conclusions and document the process. At last, you can finally share your great work to the people you want - friends, families, instructors, managers, bosses, colleagues and more.