Data Wrangling using SQL

Data wrangling describes a series of processes designed to explore, transform, and validate raw datasets from their messy and complex forms into high-quality data with good integrity and consistency into produce valuable insights and guide business decisions in later analytics purposes.

For this Data Wrangling Project, I'm using electronic store sales dataset from Kaggle: https://www.kaggle.com/datasets/saumaydhaundiyal/electronic-store-sales-data

For the steps I'm using for this project are:

1. Data Discovery
2. Data Cleaning
3. Data Transformation
4. Data Enriching
5. Data Validating
6. Data Publishing

# Data Discovery

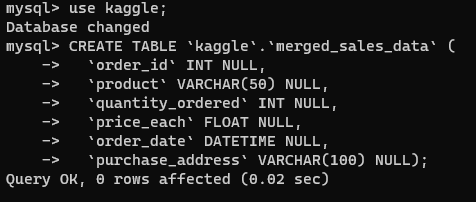
The main purposes in this step will be:

* Import data from our local machine
* Gather useful insight & information for future step.

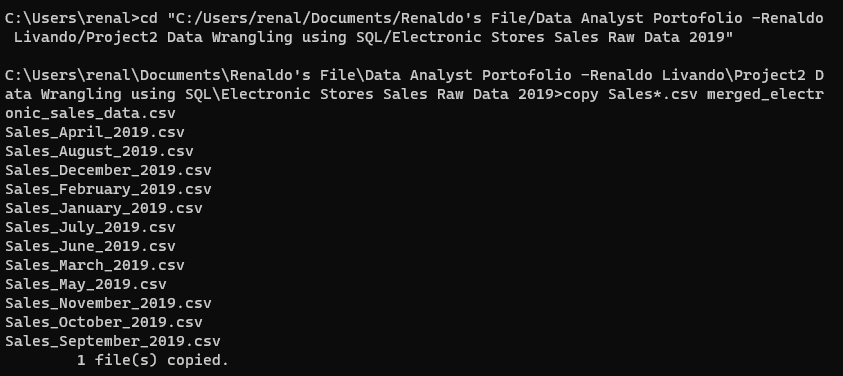
## Importing Data

In this step, instead using Table Import Wizard in MySQL Workbench, we will be using MySQL console. The reason is because using Table Import Wizard is inefficient to importing large dataset (the progress is too slow)

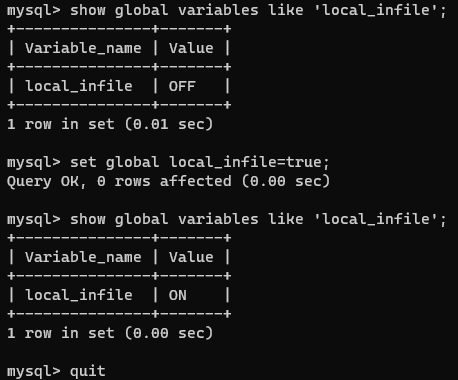
With this method we need to create our empty table to store the data later. We set our preferred datatype and match the column from our raw dataset.

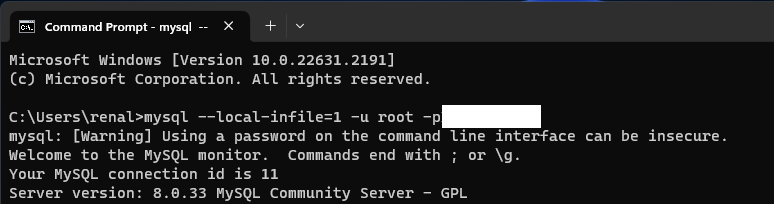


Then, since our data is in separated csv files. We need to merge it first (we can perform this using python or just CMD console)

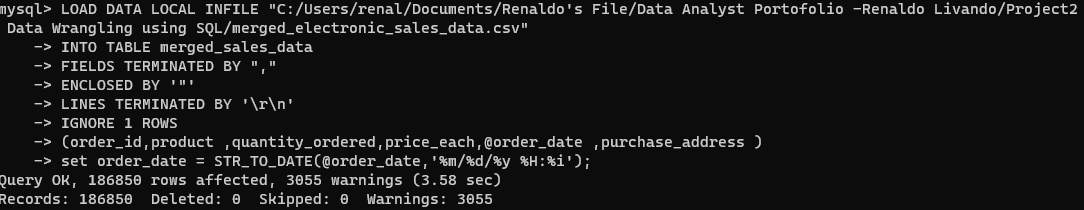


Next, we need to make sure our local\_infile variable is activate, then restart our console and login with local\_infile used.

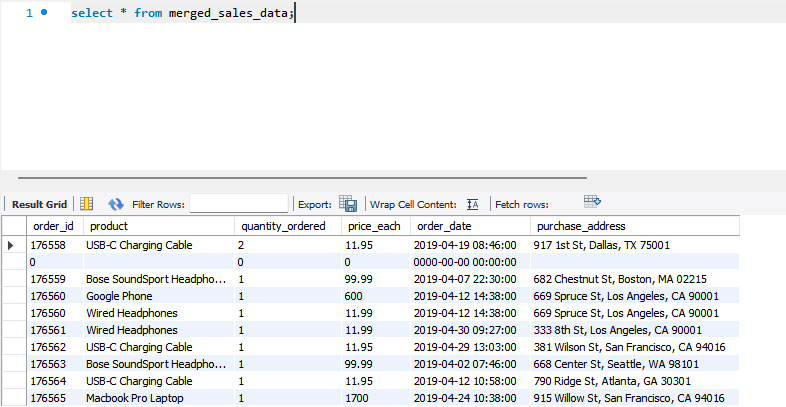




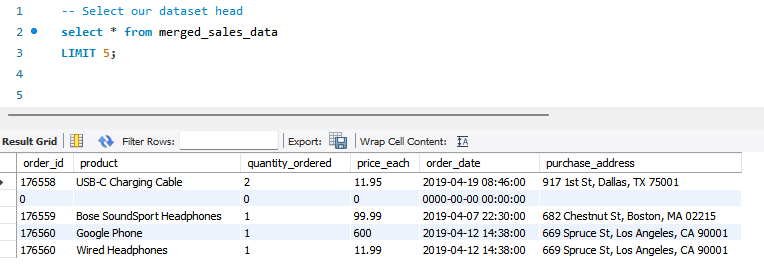
And the last step in this step will load the data from raw dataset into our table that we created.



Give a check on MySQL Workbench to make sure the data is imported correctly.



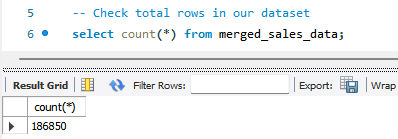
## Check and review our dataset



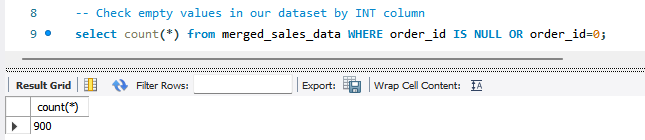
Insight from dataset review:

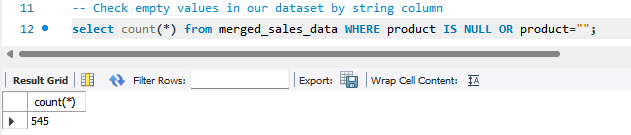
* All the column in the dataset is necessary for our later analysis, there’s no need to commit column filtering
* That seems we need to extract the data from column "Order Date" and "Purchase Address" to make new column like "City", "Postal Code", "Month", etc.
* We can add new calculated column from column "Quantity Ordered" and "Price Each"

Check dataset total rows



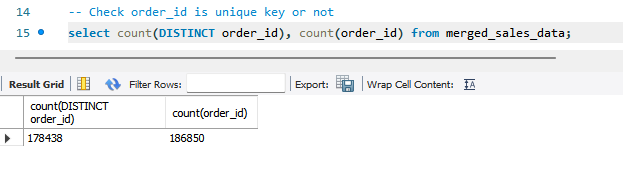
Check empty values in our dataset. We will use 2 columns with different data type to check this.





From result we got in here. We can conclude that the number of missing values quite high. And because the order\_id column occur more missing values than the product column, we will be using order\_id column to remove the missing values later instead of using string/text column.

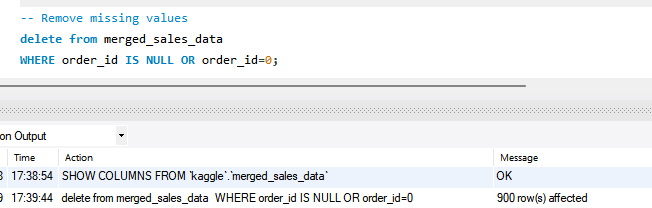
Check if our dataset ID is unique key or not



The number of rows is different. Means ID column in our dataset not unique.

# Data Cleaning

## Remove missing values



Handle du