Data Intake Report

Name: File ingestion and schema validation
Report date: 12 July 2024
Internship Batch: 9572138
Version: 1.0
Data intake by: Armel MOUMBE
Data intake reviewer:
Data storage location: https://github.com/m-armel/Data-glacier-Internship.git

Tabular data details: brewery_data

Total number of observations	10000000
Total number of files	1
Total number of features	20
Base format of the file	.csv
Size of the data	2.5 GB

File Overview:
The Brewery Operations and Market Analysis dataset used for this task was found on Kaggle provided in a CSV format. Each line in the file corresponds to a data record, with individual fields separated by commas. I changed the name for an easier use.

File Content and Structure:

- **Columns and Data Types**: The file contains multiple columns representing various attributes of the brewery operations, including brewing parameters, beer styles, sales, and quality metrics. The data types range from numerical (both integers and floats), datetime, to categorical strings.
- **Rows**: Each row in the file represents a unique batch of beer production, with a comprehensive set of features providing detailed insights into the brewing process, sales performance, and quality assessment.
- **Header**: The first row of the CSV file serves as a header, clearly labeling each column to denote the corresponding data field.

Key Features:

- **Brewing Parameters**: This includes data on fermentation time, temperature, pH level, and other critical brewing metrics.
- **Sales Data**: Detailed records of sales figures for each batch, segmented by packaging type and location.
- **Quality Scores**: Quality assessments for each batch, providing a measure of product excellence.
- **Volume and Efficiency Metrics**: Information on the volume produced, sales, and efficiency metrics, useful for supply chain and operational analysis.

The **brewery_data.csv** file contains the following columns:
Batch_ID: A unique identifier assigned to each batch of beer produced.

Brew_Date: The date on which the beer batch was brewed.
Beer_Style: The style or type of beer, such as IPA, Stout, Lager, Ale, etc.
SKU: The packaging type in which the beer is sold, like Kegs, Bottles, Cans, or Pints.
Location: The location where the beer is sold.
Fermentation_Time: The duration of the fermentation process, measured in days.
Temperature: The average temperature (Celsius) maintained during the brewing process.
PH_level: The pH level of the beer, indicating its acidity or alkalinity.
Gravity: A measure of the density of the beer as compared to water, indicating the potential alcohol content.
Alcohol_Content: The percentage of alcohol by volume in the beer.
Bitterness: The bitterness of the beer, measured in International Bitterness Units (IBU).
Color: The color of the beer measured using the Standard Reference Method (SRM).
Ingredient_Ratio: The ratio of different ingredients used in the beer, such as malt, hops, etc.
Volume_Produced: The volume of beer produced in the batch, measured in liters.
Total_Sales: The total sales generated from the batch, expressed in a currency unit.
Quality_Score: An overall quality score assigned to the beer batch, rated out of 10.
Brewhouse_Efficiency: The efficiency of the brewing process, expressed as a percentage.
Loss_During_Brewing: The percentage of volume loss during the brewing process.
Loss_during_Fermentation: The percentage of volume loss during the fermentation process.
Loss_During_Bottling_Kegging: The percentage of volume loss during the bottling or kegging process.

Proposed Approach:
The approach to reading the file was to first use the Pandas approach to reading the data, have an overview of what it looks like. Then use another method like Dask and compare it the two. Pandas took 40 seconds to read while Dask used 29. This shows us that although pandas is effective, it’s not the best and is not as dynamic.
We then create a YAML file with all the attributes of the dataset, use this file to create a dynamic code for reading the data and which will help in the schema validation.

Assumptions:
The data type of each column is accurate.
No missing values or blank rows.