DATA-1202-03– DATA ANALYSIS TOOLS ANALYTICS

Project 02 - Data Transformation using Python

Professor: Omar

Prepared by: Group 07

February 21st, 2025

Shivam Choudhary-100949298

Bhavika Lathigara - 100917684

Sai Praneeth Kurmapu -100951915

Ishan Sevak- 100951570

Monisha Senthil Velu - 100933456

Table of Contents

[Log Sheet 3](#_Toc191070546)

[Meeting Agenda 4](#_Toc191070547)

[Meeting Minutes 4](#_Toc191070548)

[Explanation 9](#_Toc191070549)

[Why this approach 10](#_Toc191070550)

[Output 11](#_Toc191070551)

[MySQL Data Loading 11](#_Toc191070552)

[Lessons Learned 14](#_Toc191070553)

[References 15](#_Toc191070554)

# Log Sheet

|  |  |  |  |
| --- | --- | --- | --- |
| TASK | TEAM MEMBER | DETAILS | DATE |
| Initial meeting and requirement discussion | All Members | Discussed the approach | Feb 17, 2025 |
| Python function for distribution of channel type | Ishan Sevak | Created a Python Function | Feb 18, 2025 |
| Filtered top 1000 records and exported CSV | Shivam Choudhary | CSV for 1000 Rows created | Feb 18, 2025 |
| MySQL table creation and data loading | Bhavika Lathigara | Tried loading CSV file into MySQL | Feb 19, 2025 |
| Documentation compilation | Sai Praneeth Kurmapu | Complied Document | Feb 20, 2025 |
| Justification & Explanation | Monisha Senthil Velu | Justified the code | Feb 20, 2025 |
| Drafting Log Sheet, Meeting Agenda and Minutes | Bhavika Lathigara | Collected Info and prepared log sheet, meeting agenda and minutes | Feb 20, 2025 |
| Final review and submission | Shivam Choudhary | Final Documentation Review | Feb 21, 2025 |

# 

# Meeting Agenda

|  |  |
| --- | --- |
| Meeting Date: Feb 17, 2025 | Time: 4:00 PM EST  Location: College (In-person) |
| AGENDA: |  |
| 1. Overview of Assignment Objectives | Led by Ishan Sevak |
| 2. Assign Tasks to Team Members | All Members |
| 3. Discuss approach for Python function and CSV export | Led by Bhavika Lathigara |
| 4. Plan MySQL table structure and data loading | Led by Shivam Choudhary |
| 5. Report Drafting, Justification Ideas and Editing Plan | Led by Sai and Monisha |

# Meeting Minutes

|  |  |
| --- | --- |
| Date: Feb 17, 2025 | Time & Place : 4:00 PM EST, In-Person |
| Attendees: ALL MEMBERS | Agreed by all |
| 1. Assignment objectives were reviewed and clarified. |
| 2. Tasks were assigned as per the team members' strengths. |
| 3. Approach for handling data extraction and transformation discussed. |
| 4. Discussed SQL database schema and loading process. |
| 5. Report formatting and final edits were assigned to Monisha. |
| 6. How we can arrange and justify the codes and task assigning. |
| 7. Set deadlines for each task to ensure timely completion. |
| Duration: 30 Minutes | Action Item: - Complete tasks by respective deadlines. |

CodesA screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

# Explanation

**Importing Libraries**

We started by importing two essential Python libraries:

* ***Pandas*** are used to handle and analyze the dataset.
* ***numpy*** for numerical operations.

Then, we loaded the Youtube dataset (youtube\_dataset.csv) to begin our analysis.

**Loading the Dataset**

The variable x contains the dataset path located at (youtube\_dataset.csv). The parameter 'encoding='unicode\_escape'' enables proper interpretation of special characters that appear in the dataset.

**Checking the Data**

* We used ***.head()*** function to take a quick look at the first few rows of the dataset.
* The ***.info()*** function helped us check for any missing values.

**Handling Missing Values in channeltype Column**

AS we were only concerned with Channeltype, we calculated mode for it and then replaced the result with mode to make all null values non-null.

**Extracting the Top 1000 Channels**

* We extracted the first 1000 rows of the dataset to focus on a smaller subset for analysis.
* This subset was stored in a new DataFrame called Top\_1000\_Channels.

**Analyzing Channel Type Distribution**

* We defined a function called Channel\_type\_distribution to calculate the frequency distribution of unique values in the channeltype column.
* This function was applied to the Top\_1000\_Channels DataFrame to analyze the distribution of channel types.

**Saving the Top 1000 channels as CSV**

* We saved the Top\_1000\_Channels DataFrame to a CSV file named Top\_1000\_Youtube\_Channels.csv.
* The index=True parameter ensured that the row indices were included in the saved file.

# Why this approach

**Handling symbols in the original file:**

We used Unicoe\_escape to decode the file as without it file was not able to load for the dataset.

**Handling Missing Data:**

We filled in null values in the channeltype column using the mode value to achieve complete data representation.

**Focusing on Top 1000 Channels:**

The dataset received analysis simplification through extraction of its top 1000 channels.

**Channel Type Distribution:**

The Channel\_type\_distribution function generated frequency data about various channels among the top 1000 channels.

**Saving Results:**

The program stored its findings into a CSV file from where users could retrieve and can load it to the mysql.

# Output

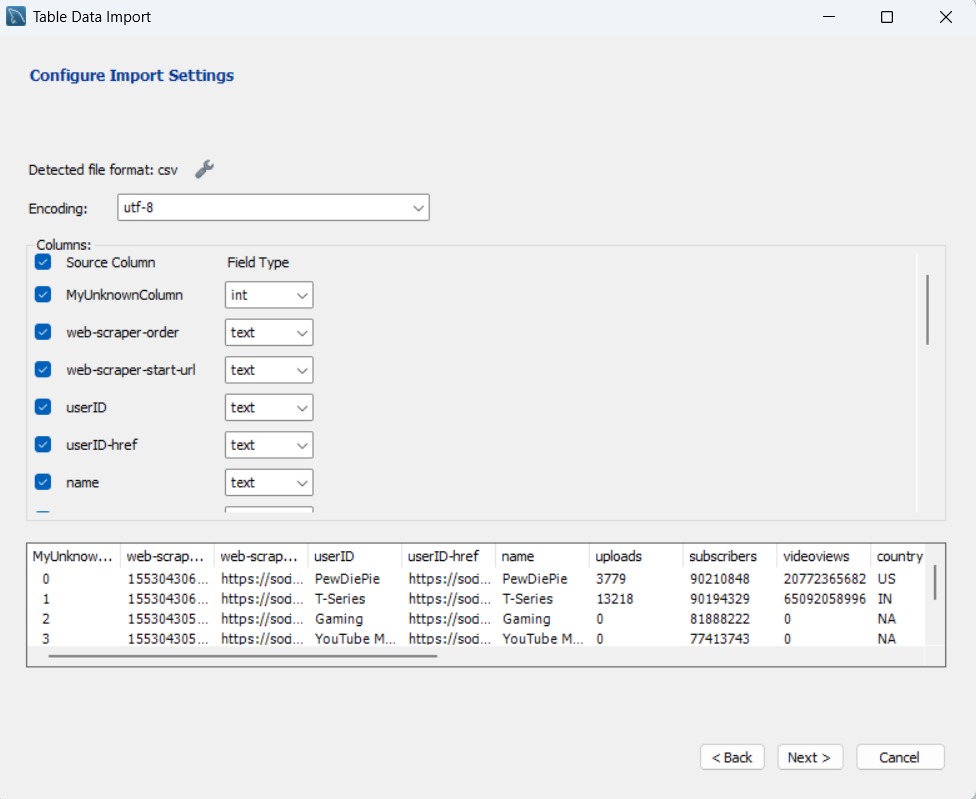
A screenshot of a computer

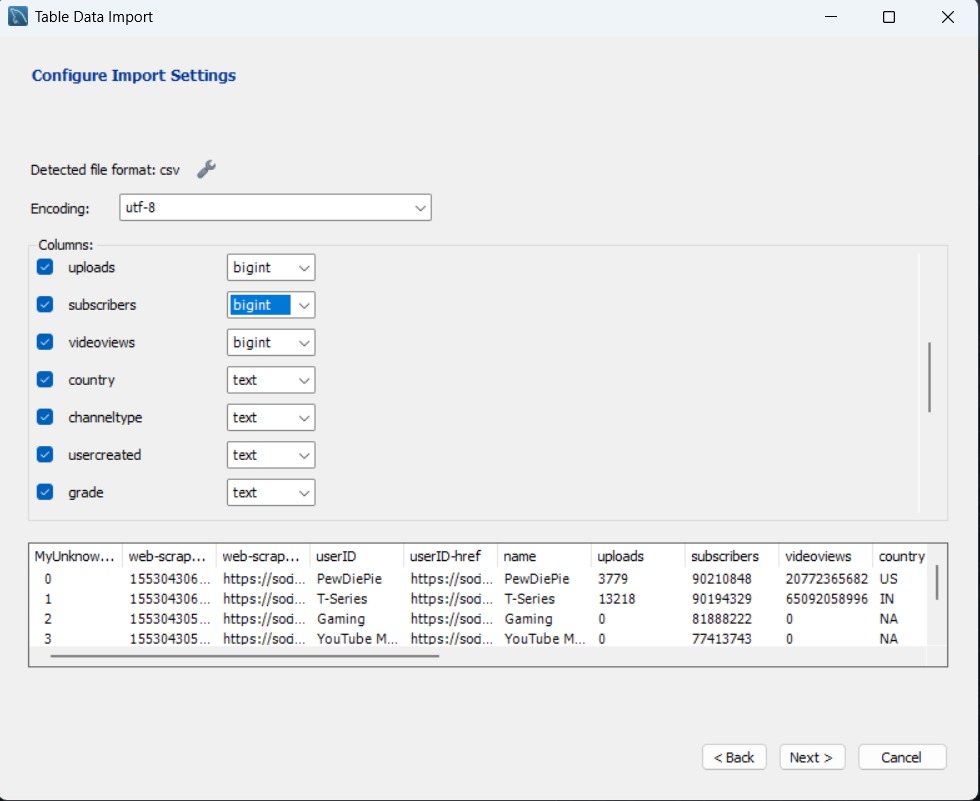
AI-generated content may be incorrect.

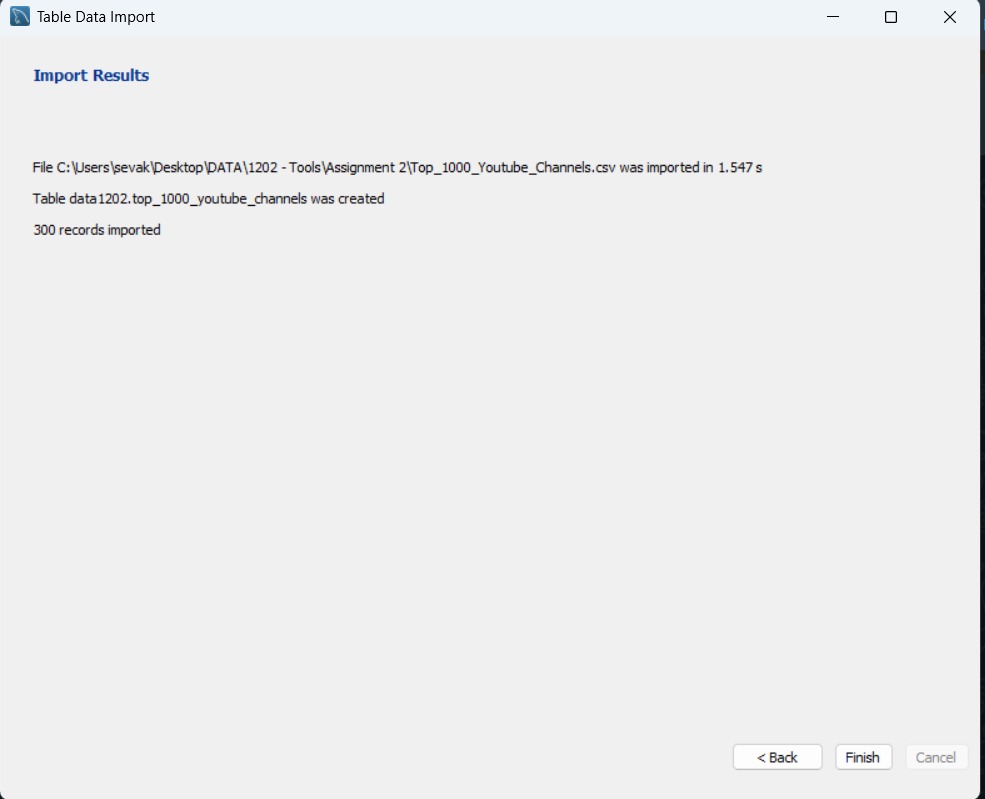
(Al-Trad, 2025) (Al-Trad, Week 4 > LAB 5, 2025) (Al-Trad, Week 5, 2025)

# MySQL Data Loading

We had a big query, as we have imported the file from python and when we are trying to import the CSV file to MySQL, it is only giving us 300 rows instead of 1000. Below are all the steps we followed to load the CSV file into MySQL.







# Lessons Learned

We resolved the data handling problems regarding special characters in the dataset through Unicode escape and empty values in the channel type column using the most common value. The data transformation process heavily relied on Python through the combination of Pandas for data manipulation while NumPy performed numerical operations. Our team developed a Python function to process channel type distribution that will benefit future users. The team made significant progress through proper documentation that included log sheets and explanations along with meeting notes to enhance our approach to data management decisions. The review step functioned as the final hurdle which ensured the verification of findings while also confirming the report format before the document became submitted.

# References

Al-Trad, O. (2025). *Week 4 > LAB 5*. Retrieved Fenruary 21, 2025, from DC Connect Durham College: https://durhamcollege.desire2learn.com/d2l/le/content/590717/viewContent/8499541/View

Al-Trad, O. (2025). *Week 5*. Retrieved February 21, 2025, from DC COnnect Durham College: https://durhamcollege.desire2learn.com/d2l/le/content/590717/Home

Al-Trad, O. (2025). *Week3\_Data\_Aggregations\_part1*. Retrieved February 21, 2025, from DC Connect Durham College: https://durhamcollege.desire2learn.com/d2l/le/content/590717/viewContent/8499536/View