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
Meeting Agenda 4
Meeting Minutes 4
Explanation 9
Why this approach 10
Output 11
MySQL Data Loading 11
Lessons Learned 14
References 15

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.

Codes

In [1]:
```
import pandas as pd
import numpy as np
```

In [3]:
```
x="~/Desktop/DATA/1202 - Tools/Assignment 2/youtube_dataset.csv"
```

In [5]:
```
df=pd.read_csv(x, encoding='unicode_escape')
```

In [7]:
```
df.head()
```

Out[7]:

	web-scraper-order	web-scraper-start-url	userID	userID-href	name	uploads	subscribers
0	1553043067-5148	https://socialblade.com/youtube/top/5000/mosts...	PewDiePie	https://socialblade.com/youtube/c/pewdiepie	PewDiePie	3779	90210848
1	1553043063-5147	https://socialblade.com/youtube/top/5000/mosts...	T-Series	https://socialblade.com/youtube/c/tseriesmusic	T-Series	13218	90194329
2	1553043059-5146	https://socialblade.com/youtube/top/5000/mosts...	Gaming	https://socialblade.com/youtube/channel/UCOpNc...	Gaming	0	81888222
3	1553043055-5145	https://socialblade.com/youtube/top/5000/mosts...	YouTube Movies	https://socialblade.com/youtube/channel/UClgRk...	YouTube Movies	0	77413743
4	1553043051-5144	https://socialblade.com/youtube/top/5000/mosts...	Sports	https://socialblade.com/youtube/channel/UCEgdi...	Sports	0	75622870

In [9]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3944 entries, 0 to 3943
Data columns (total 20 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   web-scraper-order      3944 non-null   object
 1   web-scraper-start-url  3944 non-null   object
 2   userID                 3944 non-null   object
 3   userID-href            3944 non-null   object
 4   name                   3944 non-null   object
 5   uploads                3944 non-null   int64
 6   subscribers            3944 non-null   int64
 7   videoviews             3944 non-null   int64
 8   country                3650 non-null   object
 9   channeltype            3681 non-null   object
 10  usercreated            3944 non-null   object
 11  grade                  3944 non-null   object
 12  YouTube_Link           59 non-null     object
 13  YouTube_Link-href      3885 non-null   object
 14  TwitterHandle          52 non-null     object
 15  TwitterHandle-href     3334 non-null   object
 16  InstagramHandle        42 non-null     object
 17  InstagramHandle-href   3885 non-null   object
 18  MonthlyEarnings        3944 non-null   object
 19  YearlyEarnings         3944 non-null   object
dtypes: int64(3), object(17)
memory usage: 616.4+ KB
```

In [11]:
```
mode_type = df['channeltype'].mode()[0]
```

In [13]:
```
print(mode_type)
```

```
Entertainment
```

In [15]:
```
df['channeltype'] = df['channeltype'].fillna(mode_type)
```

In [17]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3944 entries, 0 to 3943
Data columns (total 20 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   web-scraper-order      3944 non-null   object
 1   web-scraper-start-url  3944 non-null   object
 2   userID                 3944 non-null   object
 3   userID-href            3944 non-null   object
 4   name                   3944 non-null   object
 5   uploads                3944 non-null   int64
 6   subscribers            3944 non-null   int64
 7   videoviews             3944 non-null   int64
 8   country                3650 non-null   object
 9   channeltype            3944 non-null   object
 10  usercreated            3944 non-null   object
 11  grade                  3944 non-null   object
 12  YouTube_Link           59 non-null     object
 13  YouTube_Link-href      3885 non-null   object
 14  TwitterHandle          52 non-null     object
 15  TwitterHandle-href     3334 non-null   object
 16  InstagramHandle        42 non-null     object
 17  InstagramHandle-href   3885 non-null   object
 18  MonthlyEarnings        3944 non-null   object
 19  YearlyEarnings         3944 non-null   object
dtypes: int64(3), object(17)
memory usage: 616.4+ KB
```

In [19]:
```
Top_1000_Channels=df.head(1000)
```

In [21]:
```
Top_1000_Channels
```

In [21]:
```
Top_1000_Channels
```

Out[21]:

	web-scraper-order	web-scraper-start-url	userID	userID-href	name	u
0	1553043067-5148	https://socialblade.com/youtube/top/5000/mosts...	PewDiePie	https://socialblade.com/youtube/c/pewdiepie	PewDiePie	
1	1553043063-5147	https://socialblade.com/youtube/top/5000/mosts...	T-Series	https://socialblade.com/youtube/c/tseriesmusic	T-Series	
2	1553043059-5146	https://socialblade.com/youtube/top/5000/mosts...	Gaming	https://socialblade.com/youtube/channel/UCOpNc...	Gaming	
3	1553043055-5145	https://socialblade.com/youtube/top/5000/mosts...	YouTube Movies	https://socialblade.com/youtube/channel/UClgRk...	YouTube Movies	
4	1553043051-5144	https://socialblade.com/youtube/top/5000/mosts...	Sports	https://socialblade.com/youtube/channel/UCEgdi...	Sports	
...	...	...	...	...	...	
995	1553037784-3900	https://socialblade.com/youtube/top/5000/mosts...	GloZell Green	https://socialblade.com/youtube/user/glozell1	GloZell Green	
996	1553037792-3901	https://socialblade.com/youtube/top/5000/mosts...	ETV Jabardasth	https://socialblade.com/youtube/c/etvjabardasth	ETV Jabardasth	
997	1553037768-3897	https://socialblade.com/youtube/top/5000/mosts...	The Timeliners	https://socialblade.com/youtube/c/thetimeliners	The Timeliners	
998	1553037724-3888	https://socialblade.com/youtube/top/5000/mosts...	Crafty Panda	https://socialblade.com/youtube/channel/UC03Rv...	Crafty Panda	
999	1553037758-3896	https://socialblade.com/youtube/top/5000/mosts...	Troom Troom PT	https://socialblade.com/youtube/channel/UCgCQl...	Troom Troom PT	

1000 rows × 20 columns

In [23]:
```
def Channel_type_distribution(Function):
    return Function[['channeltype']].value_counts()
```

In [25]:
```
Channel_wise_distribution=Channel_type_distribution(Top_1000_Channels)
```

In [27]:
```
print(Channel_wise_distribution)
```

```
channeltype
Entertainment    322
Music            240
Games            115
Comedy            76
People            72
Howto             49
Film              36
Education         30
Tech              19
News              17
Sports            17
Autos              3
Animals            2
Nonprofit          1
Travel             1
Name: count, dtype: int64
```

In [29]:
```
Top_1000_Channels.to_csv('Top_1000_Youtube_Channels.csv', index=True)
```

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

```
channeltype
Entertainment    322
Music            240
Games            115
Comedy            76
People            72
Howto             49
Film              36
Education         30
Tech              19
News              17
Sports            17
Autos              3
Animals            2
Nonprofit          1
Travel             1
Name: count, dtype: int64
```

(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.

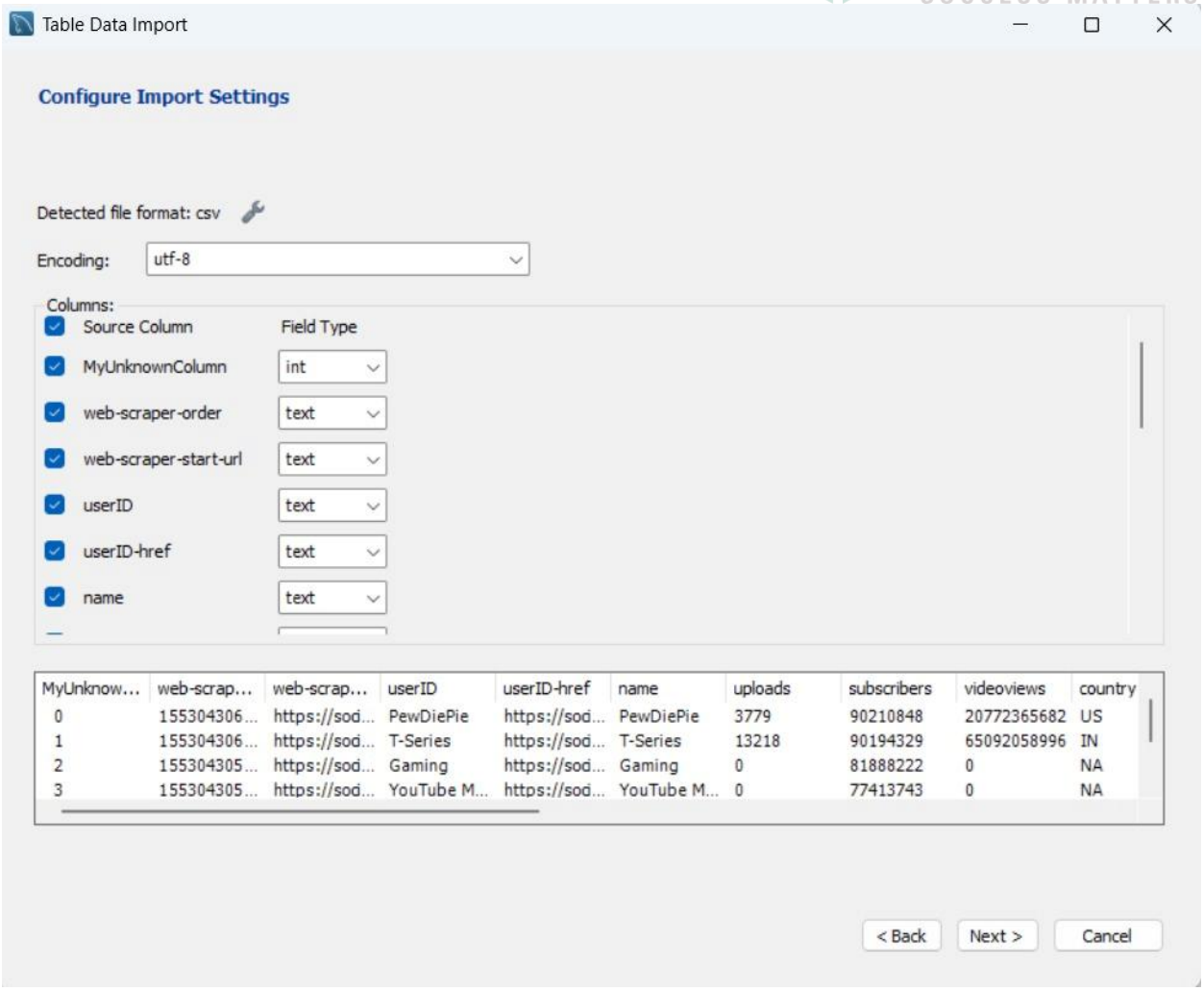


Table Data Import
Configure Import Settings
Detected file format: csv
Encoding:
utf-8
Columns:
Source Column
Field Type
MyUnknownColumn
int
web-scraper-order
text
web-scraper-start-url
text
userID
text
userID-href
text
name
text
MyUnknow... web-scrap... web-scrap... userID userID-href name uploads subscribers videoviews country
0 155304306... https://sod... PewDiePie https://sod... PewDiePie 3779 90210848 20772365682 US
1 155304306... https://sod... T-Series https://sod... T-Series 13218 90194329 65092058996 IN
2 155304305... https://sod... Gaming https://sod... Gaming 0 81888222 0 NA
3 155304305... https://sod... YouTube M... https://sod... YouTube M... 0 77413743 0 NA
< Back
Next >
Cancel

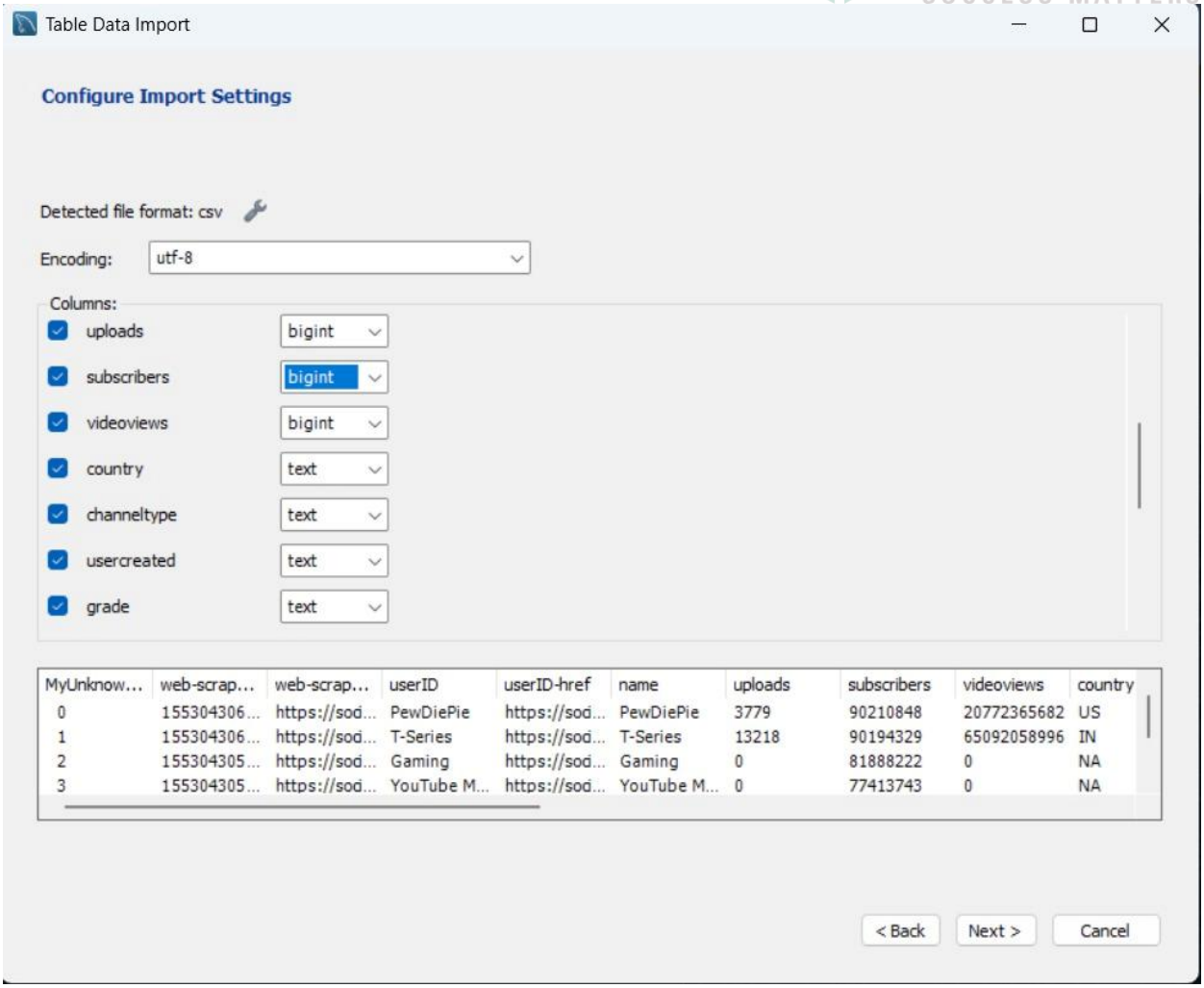
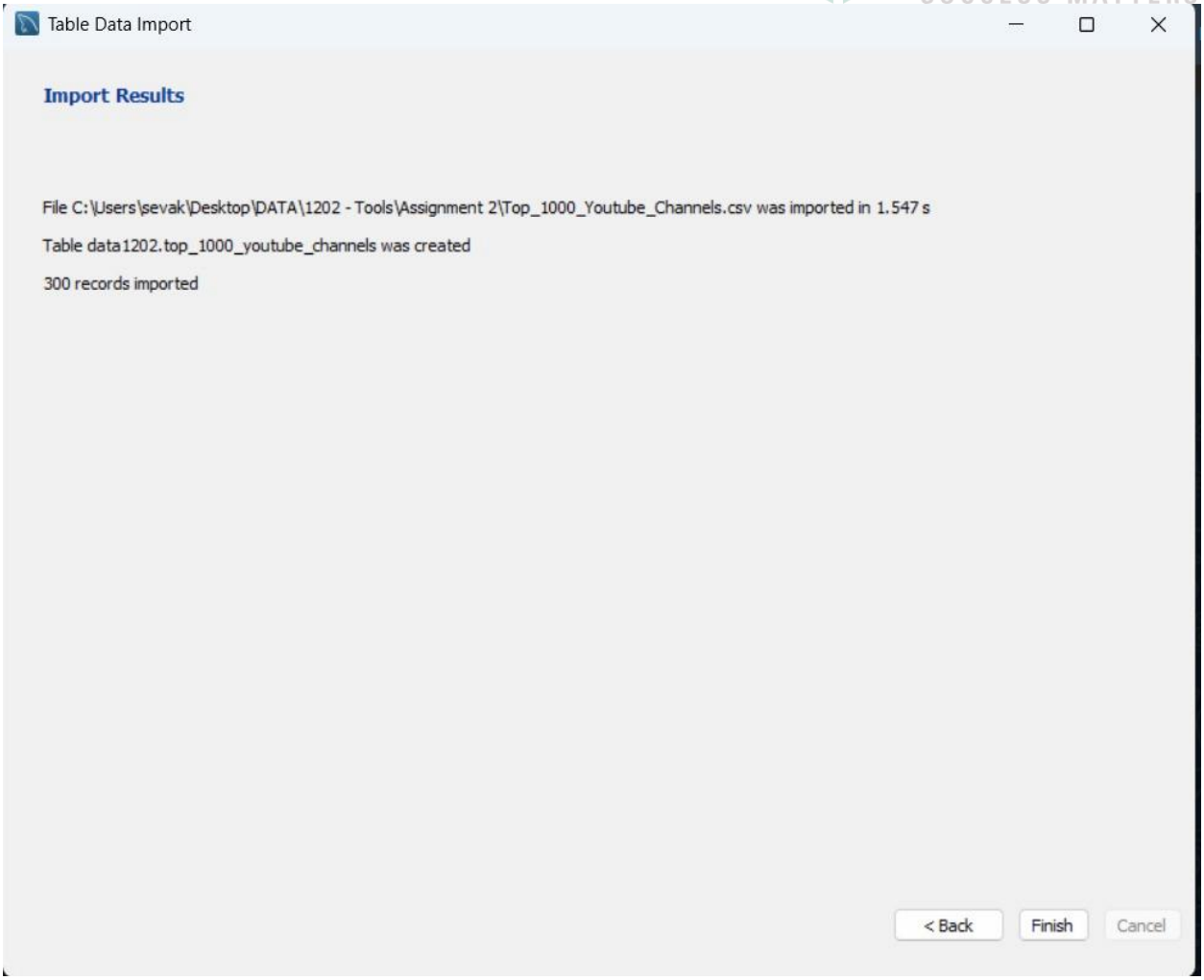


Table Data Import
Configure Import Settings
Detected file format: csv
Encoding:
utf-8
Columns:
uploads
bigint
subscribers
bigint
videoviews
bigint
country
text
channeltype
text
usercreated
text
grade
text
MyUnknow... web-scrap... web-scrap... userID userID-href name uploads subscribers videoviews country
0 155304306... https://sod... PewDiePie https://sod... PewDiePie 3779 90210848 20772365682 US
1 155304306... https://sod... T-Series https://sod... T-Series 13218 90194329 65092058996 IN
2 155304305... https://sod... Gaming https://sod... Gaming 0 81888222 0 NA
3 155304305... https://sod... YouTube M... https://sod... YouTube M... 0 77413743 0 NA
< Back
Next >
Cancel



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