

**DATA SCIENCE**

**PROBLEM STATEMENT  
NO. 50**

# **BASIC CSV FILE ANALYZER**

# CONTENT

Problem Statement

Why this ?

Use of Modules

Use of Error Handling

Use of try method

Data Cleaning and Saving it

# PROBLEM STATEMENT

To design Python script that reads a CSV file and gives quick insights.

## Expected Outcome:

Show file structure (columns, types),

Show top 5 and bottom 5 rows,

Count of unique values per column,

Save cleaned data to a new CSV

## Input:

Any CSV file



## WHY THIS ?

The reason I chose this problem statement was because it stood out to me more as compared to other problems. In my opinion CSV analyzer solve issues of many who want a quick insight to their CSV file.

# USE OF MODULES

During the course of learning about Data Science on GUVI, I learned about the concept of importing required modules and thus applied appropriate modules in my python script.

```
CSV File Analyzer.py  
1 import pandas as pd  
2 import os  
3
```

# USE OF ERROR HANDLING

During the course of learning about Data Science on GUVI, I learned about the concept of error handling and thus thought of applying in my python script.

```
56
57     except FileNotFoundError:
58         print(f"Error: The file '{file_path}' was not found.")
59     except pd.errors.EmptyDataError:
60         print(f"Error: The file '{file_path}' is empty.")
61     except Exception as e:
62         print(f"An unexpected error occurred: {e}")
63
```

# USE OF TRY METHOD

During the course of learning about Data Science on GUVI, I learned about the concept of try method and thus thought of applying in my python script.

```
11     try:
12         # 1. Read the CSV file
13         df = pd.read_csv(file_path)
14         print(f"\n--- Insights for '{file_path}' ---")
15
16         # 2. Show file structure (columns, types, non-null counts)
17         print("\n--- File Structure (Columns, Data Types, Non-Null Counts) ---")
18         df.info()
```



# DATA CLEANING AND SAVING IT

With GUVI videos and external help, I was able to complete a major part where I got stuck a few times and that was saving the cleaned file.

Scratched my head and got angry but got over with it within the time period. Since this was a part of marking rubric I paid more attention to it, similarly I tried sticking to the rubric as much as possible for other parts also.

```
35 # 6. Save cleaned data to a new CSV
36 # Basic cleaning: dropping duplicate rows
37 initial_rows = len(df)
38 df_cleaned = df.drop_duplicates()
39 rows_after_cleaning = len(df_cleaned)
40
41 print(f"\n--- Data Cleaning ---")
42 if initial_rows > rows_after_cleaning:
43     print(f"Removed {initial_rows - rows_after_cleaning} duplicate row(s).")
44 else:
45     print("No duplicate rows found or removed.")
46
47 # Construct new file name
48 base_name = os.path.basename(file_path)
49 dir_name = os.path.dirname(file_path)
50 new_file_name = f"cleaned_data_{base_name}"
51 output_file_path = os.path.join(dir_name, new_file_name)
52
53 df_cleaned.to_csv(output_file_path, index=False)
54 print(f"Cleaned data saved to: '{output_file_path}'")
55 print("\n--- Analysis Complete ---")
56
```

# THANK YOU

From: Nisarg Kumar Gharde

\*I hope the presentation was up to your expectation and sorry for any inconvenience, if caused.