

Data Intake Report

Name: Data Analyst: Cross selling recommendation - Group Project

Report date:

Internship Batch: LISUM21

Version: 1.0

Data intake by: Bogdan-Remus Pintilie

Data intake reviewer: Raka Prasetya Nugraha, Parwinder Singh, Kevin Wang

Data storage location:

<https://drive.google.com/file/d/1OCOZGjoL14tgKZtCl4Q9cifuL8xwIAo6/view?usp=sharing>

<https://drive.google.com/file/d/1wCwF2D6lzxh50rBFWXwW330ByiHhVK0w/view?usp=sharing>

Tabular data details:

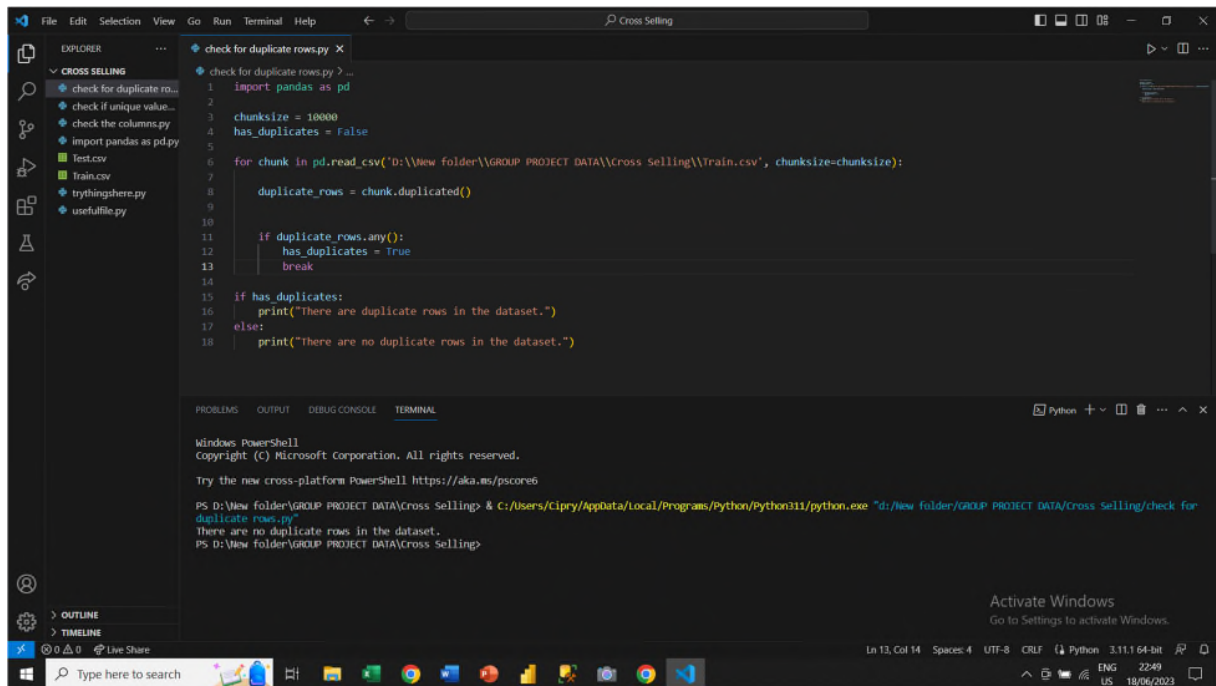
Total number of observations	929615
Total number of files	1 (Test.csv)
Total number of features	24
Base format of the file	.csv
Size of the data	105.2 MB

Total number of observations	13647309
Total number of files	1 (Train.csv)
Total number of features	48
Base format of the file	.csv
Size of the data	2.14 GB

Proposed Approach:

1. Independence: The assumption of independence suggests that the data points being analyzed are not influenced by each other. Independence allows for unbiased analysis and statistical tests that rely on the assumption of independence.
2. Validity: The assumption of validity implies that the data represents what it claims to represent. Valid data is relevant and appropriate for the analysis objectives, and it accurately captures the intended information or measurements.
3. Integrity: The assumption of data integrity suggests that the data has not been tampered with or altered inappropriately. Data integrity ensures that the data retains its original state and remains reliable throughout the analysis process.
4. The dataset does not contain any duplicate rows and to find this we used the code provided in the first picture below.
5. Duplicate identification in the key field: The second picture contains the code that shows the key field (ncodpers) contains duplicates.

6. We improve data relevance by omitting the following fields since they do not provide valuable insights and do not help address the problem statement: conyuemp



The screenshot shows the Visual Studio Code editor with a file named 'check for duplicate rows.py' open. The Explorer sidebar on the left shows a project structure with a folder 'CROSS SELLING' containing several files. The main editor area displays the following Python code:

```
1 import pandas as pd
2
3 chunksize = 10000
4 has_duplicates = False
5
6 for chunk in pd.read_csv('D:\\New folder\\GROUP PROJECT DATA\\Cross Selling\\Train.csv', chunksize=chunksize):
7
8     duplicate_rows = chunk.duplicated()
9
10
11     if duplicate_rows.any():
12         has_duplicates = True
13         break
14
15 if has_duplicates:
16     print("There are duplicate rows in the dataset.")
17 else:
18     print("There are no duplicate rows in the dataset.")
```

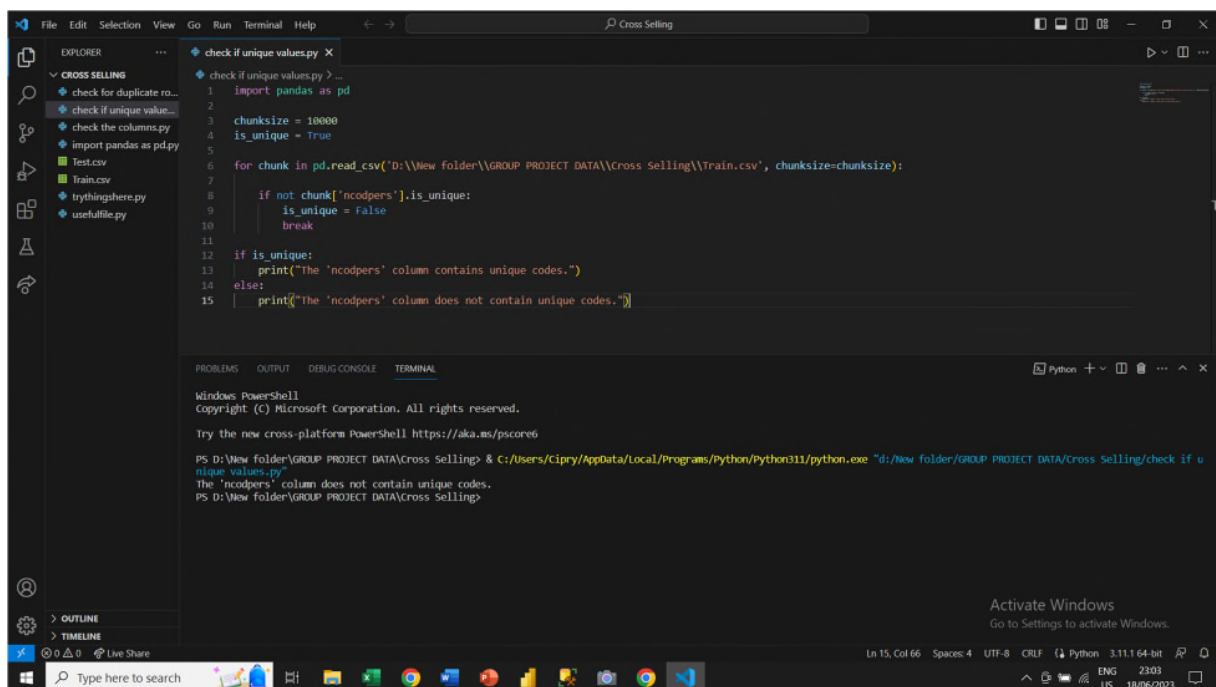
Below the code editor is a terminal window showing the execution of the script. The terminal output is:

```
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Try the new cross-platform PowerShell https://aka.ms/pscore6

PS D:\New folder\GROUP PROJECT DATA\Cross Selling> C:\Users\Cipry\AppData\Local\Programs\Python\Python311\python.exe "d:\New folder\GROUP PROJECT DATA\Cross Selling\check for duplicate rows.py"
There are no duplicate rows in the dataset.
```

The status bar at the bottom indicates the file is at Line 13, Column 14, with 4 spaces, UTF-8 encoding, CRLF line endings, and Python 3.11.1 64-bit.



The screenshot shows the Visual Studio Code editor with a file named 'check if unique values.py' open. The Explorer sidebar on the left shows the same project structure as the previous screenshot. The main editor area displays the following Python code:

```
1 import pandas as pd
2
3 chunksize = 10000
4 is_unique = True
5
6 for chunk in pd.read_csv('D:\\New folder\\GROUP PROJECT DATA\\Cross Selling\\Train.csv', chunksize=chunksize):
7
8     if not chunk['ncodpers'].is_unique:
9         is_unique = False
10         break
11
12 if is_unique:
13     print("The 'ncodpers' column contains unique codes.")
14 else:
15     print("The 'ncodpers' column does not contain unique codes.")
```

Below the code editor is a terminal window showing the execution of the script. The terminal output is:

```
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Try the new cross-platform PowerShell https://aka.ms/pscore6

PS D:\New folder\GROUP PROJECT DATA\Cross Selling> C:\Users\Cipry\AppData\Local\Programs\Python\Python311\python.exe "d:\New folder\GROUP PROJECT DATA\Cross Selling\check if unique values.py"
The 'ncodpers' column does not contain unique codes.
```

The status bar at the bottom indicates the file is at Line 15, Column 66, with 4 spaces, UTF-8 encoding, CRLF line endings, and Python 3.11.1 64-bit.