Data Intake Report

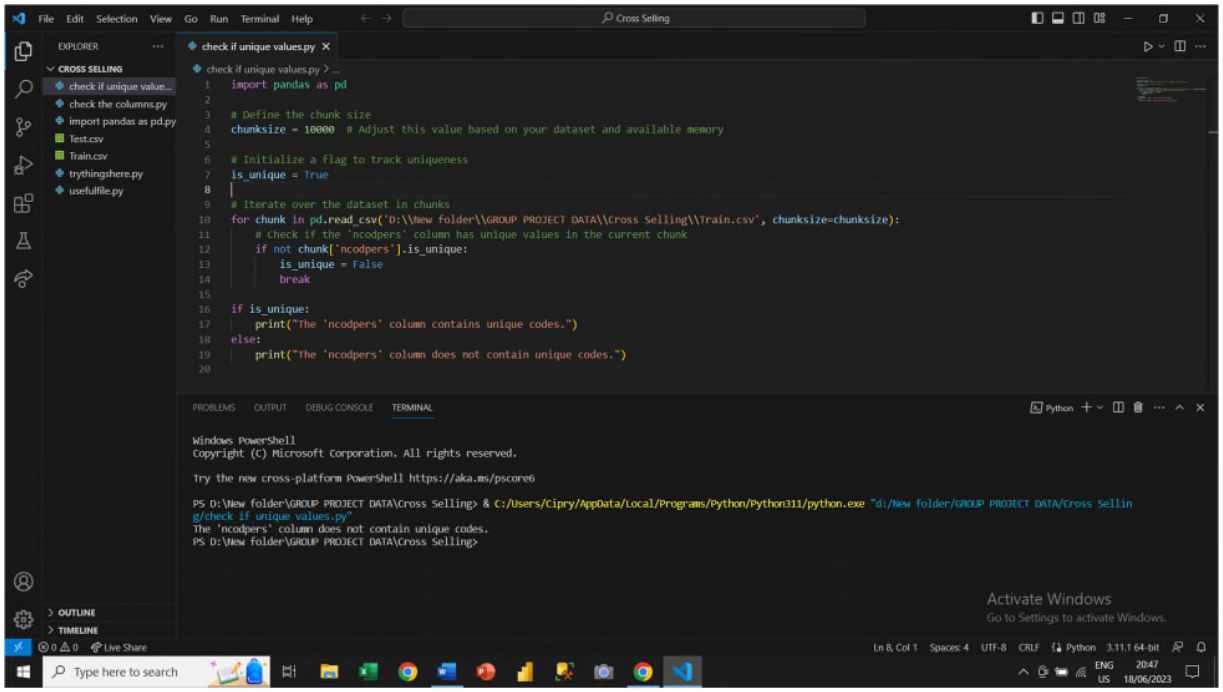
Name: Data Analyst: Cross selling recommendation - Group Project
Report date:
Internship Batch: LISUM21
Version: 1.0
Data intake by: Bogdan-Remus Pintilie
Data intake reviewer: Raka Prasetya Nugraha, Parwinder Singh
Data storage location:
https://drive.google.com/file/d/1OCOZGjoL14tgKZtCl4Q9cifuL8xwIAo6/view?usp=sharing
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. Duplicate identification: The code below shows that the key field contains duplicates.
*5. We improve data relevance by omitting the following fields since they do not provide insights and help address the problem statement: conyuemp,


File Edit Selection View Go Run Terminal Help
Cross Selling
EXPLORER
CROSS SELLING
check if unique value...
check the columns.py
import pandas as pd.py
Test.csv
Train.csv
trythingshere.py
usefulfile.py
check if unique values.py
check if unique values.py > ...
1 import pandas as pd
2
3 # Define the chunk size
4 chunksize = 10000 # Adjust this value based on your dataset and available memory
5
6 # Initialize a flag to track uniqueness
7 is_unique = True
8
9 # Iterate over the dataset in chunks
10 for chunk in pd.read_csv('D:\\New folder\\GROUP PROJECT DATA\\Cross Selling\\Train.csv', chunksize=chunksize):
11 # Check if the 'ncodpers' column has unique values in the current chunk
12 if not chunk['ncodpers'].is_unique:
13 is_unique = False
14 break
15
16 if is_unique:
17 print("The 'ncodpers' column contains unique codes.")
18 else:
19 print("The 'ncodpers' column does not contain unique codes.")
20
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL
Python
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.
PS D:\New folder\GROUP PROJECT DATA\Cross Selling>
OUTLINE
TIMELINE
Activate Windows
Go to Settings to activate Windows.
Live Share
Ln 8, Col 1 Spaces: 4 UTF-8 CRLF Python 3.11.1 64-bit
Type here to search
ENG US
20:47
18/06/2023