Batch: **C2** **Roll No.:**

Experiment 02

Title: Dataset pre-processing

Objective:

1. To learn how to prepare the dataset

2. To learn various steps in Data -Preprocessing

Course Outcome:

CO1: Learn how to locate and download datasets, extract insights from that data and present their findings in a variety of different formats.

Books/ Journals/ Websites referred:

www.kaggle.com
www.geeksforgeeks.org
https://pandas.pydata.org/docs/

Resources used:

No I forgot

(Dataset link)
https://youtu.be/dQw4w9WgXcQ

Theory (About Data Preprocessing):

Following points should be written by students

Different steps in Data Preprocessing:

- Finding missing, null values
- Replacing missing, null values with statistical parameters
- Encoding categorical data

- Normalization

Data preprocessing is a pivotal phase in the data mining process, ensuring that raw data is refined and structured to facilitate accurate and meaningful analysis. Several fundamental steps contribute to this crucial process:

1. Data Cleaning: This initial step involves identifying and rectifying errors, inconsistencies, and anomalies within the dataset. By addressing **missing values, null values, and duplicates,** the dataset's integrity is preserved, laying a solid foundation for subsequent analysis.
2. Data Integration: Often, data originates from diverse sources with differing formats and structures. Data integration harmonizes this information, bringing together data fragments to construct a unified dataset. Techniques like record linkage and data fusion aid in this amalgamation, promoting a comprehensive view of the data.
3. Data Transformation: The transformation phase molds the data into a suitable format for analysis. **Normalization and standardization techniques ensure that data with varying scales and units are adjusted to a common framework,** enabling fair comparisons and accurate interpretation.
4. Data Reduction: Managing large datasets can be challenging. Data reduction methods, like feature selection and extraction, streamline the dataset by retaining essential information while **minimizing redundant or irrelevant features**. This enhances the efficiency of subsequent analyses.
5. Data Discretization: When continuous data is needed for categorical analysis, data discretization is employed. This process divides continuous variables into **distinct intervals or categories, enabling the application of categorical-focused algorithms**.
6. Data Normalization: Normalization further standardizes data by **scaling it to a predetermined rang**e. This process is especially helpful when dealing with data that varies widely in terms of units and magnitudes.

Through these steps, data preprocessing refines the raw material, ensuring it is primed for accurate analysis. The specific approach may vary based on data characteristics and research goals, but the overarching aim remains constant: to enhance data quality and maximize the accuracy and reliability of subsequent analyses.

Note: Student can use any technology like Tableau, Tableau-Prep, PowerBI, Google spreadsheet, excel, R programming, Python, Java any other technology for preprocessing.

Platform used by the student: Python

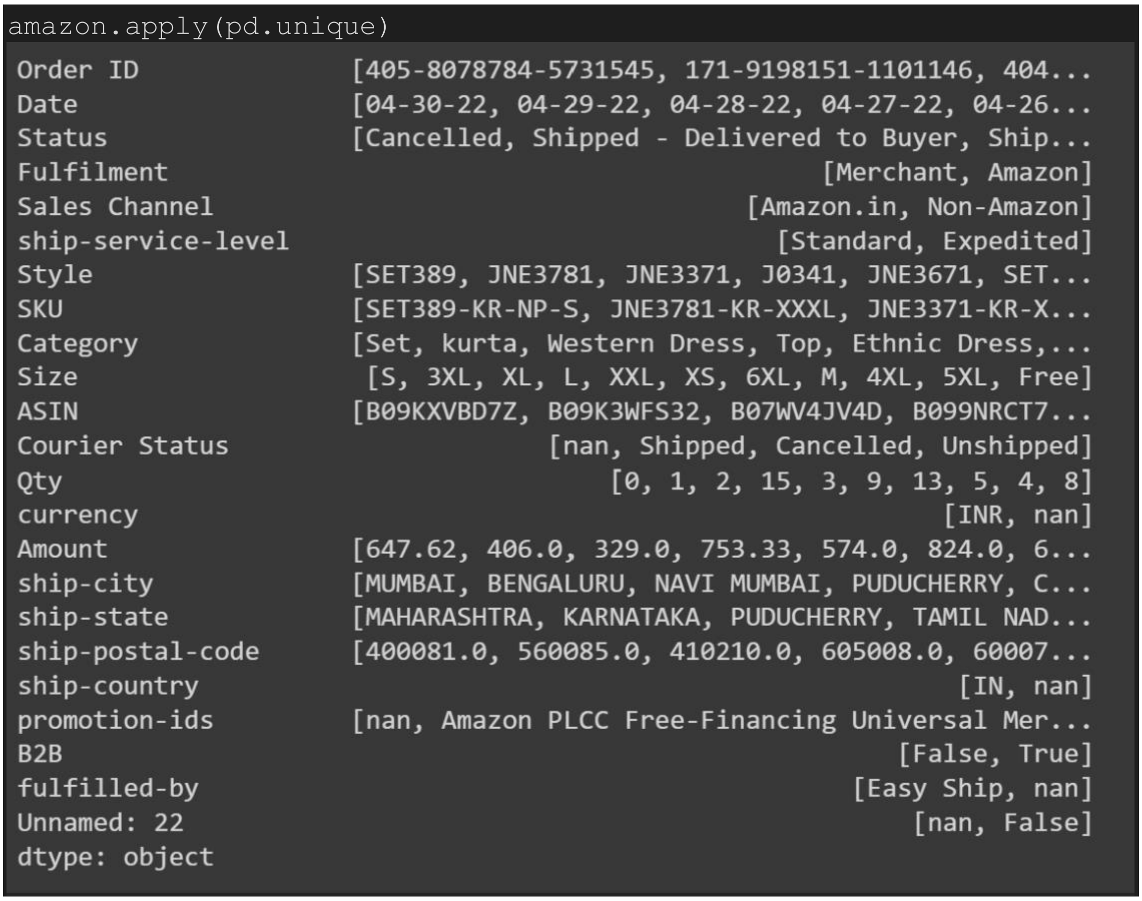
Working (Paste the code and Output for each Data Preprocessing task):

```
import pandas as pd
amazon = pd.read_csv('/content/drive/MyDrive/Amazon Sale Report.csv')
```

```
amazon.info()
```

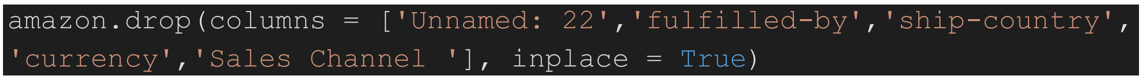
```
<ipython-input-46-03a329c543e0>:1: DtypeWarning: Columns (23) have mixed types. Specify dtype option on import or set low_memory=False.
  amazon = pd.read_csv('/content/drive/MyDrive/Amazon Sale Report.csv')
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 128975 entries, 0 to 128974
Data columns (total 24 columns):
 #   Column              Non-Null Count   Dtype  
---  ------              --------------   -----  
 0   index               128975 non-null  int64  
 1   Order ID            128975 non-null  object 
 2   Date                128975 non-null  object 
 3   Status              128975 non-null  object 
 4   Fulfilment          128975 non-null  object 
 5   Sales Channel       128975 non-null  object 
 6   ship-service-level  128975 non-null  object 
 7   Style               128975 non-null  object 
 8   SKU                 128975 non-null  object 
 9   Category            128975 non-null  object 
 10  Size                128975 non-null  object 
 11  ASIN                128975 non-null  object 
 12  Courier Status      122103 non-null  object 
 13  Qty                 128975 non-null  int64  
 14  currency            121180 non-null  object 
 15  Amount              121180 non-null  float64
 16  ship-city           128942 non-null  object 
 17  ship-state          128942 non-null  object 
 18  ship-postal-code    128942 non-null  float64
 19  ship-country        128942 non-null  object 
 20  promotion-ids       79822 non-null   object 
 21  B2B                 128975 non-null  bool   
 22  fulfilled-by        39277 non-null   object 
 23  Unnamed: 22         79925 non-null   object 
dtypes: bool(1), float64(2), int64(2), object(19)
memory usage: 22.8+ MB
```

```
amazon.set_index('index', inplace = True)
amazon.nunique()
```

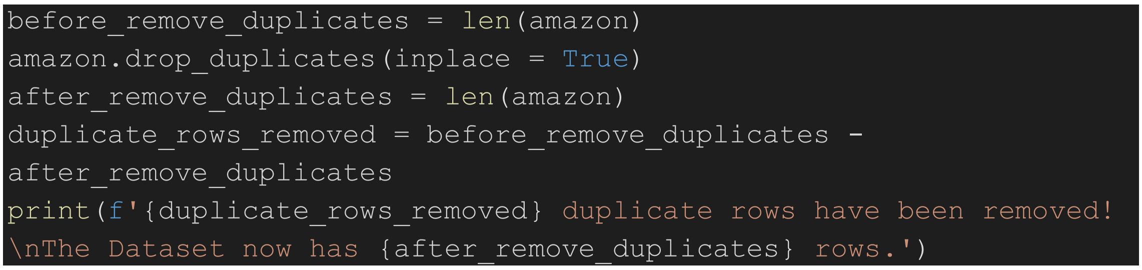


```
amazon.apply(pd.unique)
```

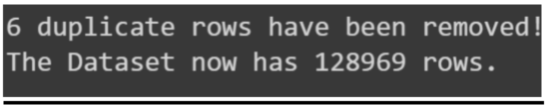
```
Order ID                 [405-8078784-5731545, 171-9198151-1101146, 404...
Date                     [04-30-22, 04-29-22, 04-28-22, 04-27-22, 04-26...
Status                   [Cancelled, Shipped - Delivered to Buyer, Ship...
Fulfilment                                              [Merchant, Amazon]
Sales Channel                                      [Amazon.in, Non-Amazon]
ship-service-level                                   [Standard, Expedited]
Style                    [SET389, JNE3781, JNE3371, J0341, JNE3671, SET...
SKU                      [SET389-KR-NP-S, JNE3781-KR-XXXL, JNE3371-KR-X...
Category                 [Set, kurta, Western Dress, Top, Ethnic Dress,...
Size                      [S, 3XL, XL, L, XXL, XS, 6XL, M, 4XL, 5XL, Free]
ASIN                     [B09KXVBD7Z, B09K3WFS32, B07WV4JV4D, B099NRCT7...
Courier Status                        [nan, Shipped, Cancelled, Unshipped]
Qty                                        [0, 1, 2, 15, 3, 9, 13, 5, 4, 8]
currency                                                        [INR, nan]
Amount                   [647.62, 406.0, 329.0, 753.33, 574.0, 824.0, 6...
ship-city                [MUMBAI, BENGALURU, NAVI MUMBAI, PUDUCHERRY, C...
ship-state               [MAHARASHTRA, KARNATAKA, PUDUCHERRY, TAMIL NAD...
ship-postal-code         [400081.0, 560085.0, 410210.0, 605008.0, 60007...
ship-country                                                     [IN, nan]
promotion-ids            [nan, Amazon PLCC Free-Financing Universal Mer...
B2B                                                          [False, True]
fulfilled-by                                              [Easy Ship, nan]
Unnamed: 22                                                   [nan, False]
dtype: object
```



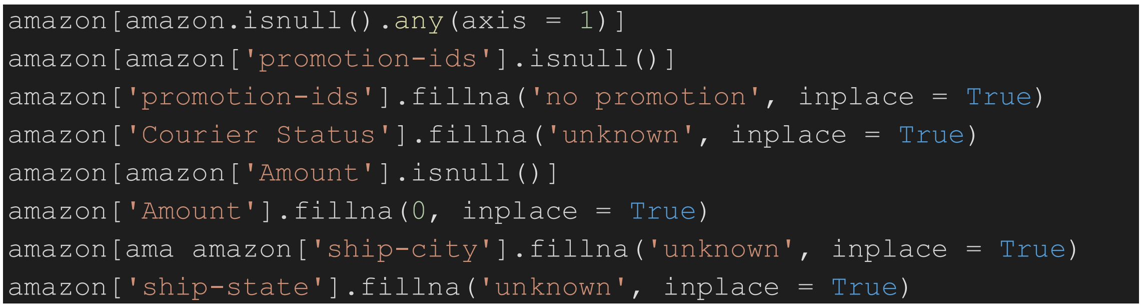
```
amazon.drop(columns = ['Unnamed: 22','fulfilled-by','ship-country',
'currency','Sales Channel '], inplace = True)
```



```
before_remove_duplicates = len(amazon)
amazon.drop_duplicates(inplace = True)
after_remove_duplicates = len(amazon)
duplicate_rows_removed = before_remove_duplicates -
after_remove_duplicates
print(f'{duplicate_rows_removed} duplicate rows have been removed!
\nThe Dataset now has {after_remove_duplicates} rows.')
```



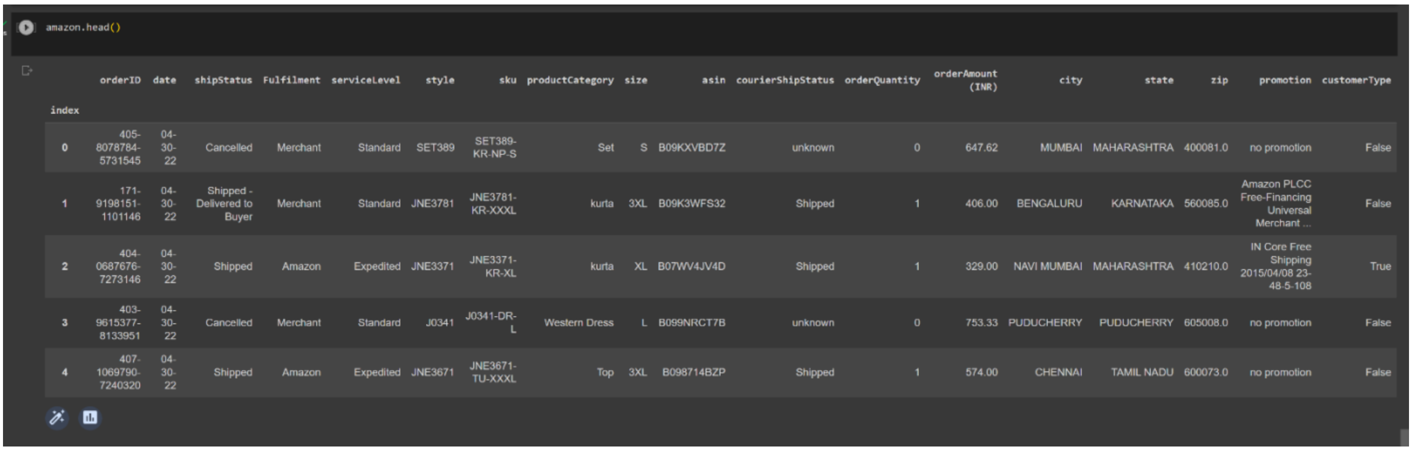
```
6 duplicate rows have been removed!
The Dataset now has 128969 rows.
```



```
amazon[amazon.isnull().any(axis = 1)]
amazon[amazon['promotion-ids'].isnull()]
amazon['promotion-ids'].fillna('no promotion', inplace = True)
amazon['Courier Status'].fillna('unknown', inplace = True)
amazon[amazon['Amount'].isnull()]
amazon['Amount'].fillna(0, inplace = True)
amazon[ama amazon['ship-city'].fillna('unknown', inplace = True)
amazon['ship-state'].fillna('unknown', inplace = True)
amazon['ship-postal-code'].fillna('unknown', inplace = True)
zon['ship-city'].isnull()]

mapper = {'Order ID':'orderID', 'Date':'date',
'Status':'shipStatus','fullfilment':'fullfilment', 'ship-service-
level':'serviceLevel', 'Style':'style', 'SKU':'sku',
'Category':'productCategory', 'Size':'size', 'ASIN':'asin',
'Courier Status':'courierShipStatus', 'Qty':'orderQuantity',
'Amount':'orderAmount (INR)', 'ship-city':'city', 'ship-
state':'state', 'ship-postal-code':'zip', 'promotion-
ids':'promotion','B2B':'customerType' }
amazon.rename(columns = mapper, inplace = True)
amazon.head()
```

First five rows:



```
amazon.head()
```

index	orderID	date	shipStatus	Fulfilment	serviceLevel	style	sku	productCategory	size	asin	courierShipStatus	orderQuantity	orderAmount (INR)	city	state	zip	promotion	customerType
0	405-8078784-5731545	04-30-22	Cancelled	Merchant	Standard	SET389	SET389-KR-NP-S	Set	S	B09KXVBD7Z	unknown	0	647.62	MUMBAI	MAHARASHTRA	400081.0	no promotion	False
1	171-9198151-1101146	04-30-22	Shipped - Delivered to Buyer	Merchant	Standard	JNE3781	JNE3781-KR-XXXL	kurta	3XL	B09K3WFS32	Shipped	1	406.00	BENGALURU	KARNATAKA	560085.0	Amazon PLCC Free-Financing Universal Merchant ...	False
2	404-0687676-7273146	04-30-22	Shipped	Amazon	Expedited	JNE3371	JNE3371-KR-XL	kurta	XL	B07WV4JV4D	Shipped	1	329.00	NAVI MUMBAI	MAHARASHTRA	410210.0	IN Core Free Shipping 2015/04/08 23-48-5-108	True
3	403-9615377-8133951	04-30-22	Cancelled	Merchant	Standard	J0341	J0341-DR-L	Western Dress	L	B099NRCT7B	unknown	0	753.33	PUDUCHERRY	PUDUCHERRY	605008.0	no promotion	False
4	407-1069790-7240320	04-30-22	Shipped	Amazon	Expedited	JNE3671	JNE3671-TU-XXXL	Top	3XL	B098714BZP	Shipped	1	574.00	CHENNAI	TAMIL NADU	600073.0	no promotion	False

Conclusion (Students should write in their own words):

Through this experiment, I learnt to process data to possess only useful information, by removing duplicate records, replacing null values, removing unnecessary columns reducing the size of the dataset and improving the overall readability of the data.

Post Lab Question:

1. **Write the importance of Data Preprocessing**

Ans:

Data processing is important.