Academic Year: 2023-24 **Semester: VIII**
Class / Branch: BE Computer **Subject: Social Media Analytics Lab**

Name: Shivam Pandey
Date of performance: 14/02/2024
Date of Submission: 24/02/2024

Experiment No. 03

Aim: To perform data cleaning on social media data using Python.

Objective: To analyze data cleaning on social media data using Python to prepare the data for meaningful analysis, ensure data integrity and quality, and facilitate efficient and ethical use of the data for generating actionable insights.

Software used: Python.

Theory:
Data cleaning is a crucial step in the data preprocessing pipeline. It involves identifying and correcting errors, inconsistencies, and inaccuracies in the data to improve its quality and reliability. In the context of social media data, which is often unstructured and noisy, data cleaning becomes even more essential.
Ensure Data Quality: The primary objective of data cleaning is to ensure that the data is accurate, consistent, and reliable. Social media data can contain various types of errors such as misspellings, grammatical mistakes, and inconsistencies that need to be addressed.
Handle Missing Values: Social media data often contains missing values due to incomplete user inputs or data collection processes. Data cleaning involves identifying and handling these missing values appropriately, either by imputation or removal.
Remove Duplicates: Social media data may contain duplicate entries, such as duplicate posts or comments. Removing duplicates ensures that each piece of information is unique and prevents redundancy in the dataset.
Standardize Formats: Social media data can have diverse formats for representing dates, times, and other structured information. Data cleaning involves standardizing these formats to facilitate analysis and comparison across different data points.
Text Cleaning and Preprocessing: Since social media data often consists of text data, cleaning and preprocessing text is essential. This may include removing special characters, URLs, hashtags, mentions, and other noise, as well as tokenization, lemmatization, and removing stopwords to prepare the text for analysis.
Ensure Consistency and Uniformity: Data cleaning ensures that the data is consistent and uniform across different attributes and records. This consistency is crucial for accurate analysis and modeling.
Enhance Analytical Results: Clean data leads to more accurate and reliable analytical results. By removing errors and inconsistencies, data cleaning improves the quality of insights derived from social media data analysis.
Compliance and Ethical Considerations: Data cleaning may also involve ensuring compliance with regulations such as GDPR (General Data Protection Regulation) and addressing ethical considerations such as privacy concerns when dealing with sensitive user data in social media datasets.
Handle Missing Values: Check for missing values and decide how to handle them. Options

include dropping rows with missing values, filling them with a default value, or using more sophisticated methods like interpolation.

Implementation and Output:
In the implementation below data cleaning operations of handling missing values, removing duplicates, standardizing formats, correcting errors, data validation, handling outliers, and data transformations are executed and cleaned (processed) dataset is taken as output.

```
import pandas as pd

# Create a DataFrame
df = pd.read_csv('data.csv')

# Display the original dataset
print("Original Dataset:")
print(df)

# Data cleaning operations
# 1. Handling Missing Values
# Here we'll replace missing CustomerID with a placeholder value -1
df['CustomerID'].fillna(-1, inplace=True)

# 2. Removing Duplicates
df.drop_duplicates(inplace=True)

# 3. Standardizing Formats
df['InvoiceDate'] = pd.to_datetime(df['InvoiceDate'], format='%m/%d/%Y %H:%M')

# 4. Correcting Errors (if any)
# In this example, let's correct an error in the 'Description' column
df.loc[df['Description'] == 'RED WOOLLY HOTTIE WHITE HEART.', 'Description']
= 'RED WOOLLY HOTTIE WITH WHITE HEART.'

# 5. Handling Outliers (if any)
# Let's consider an example of handling outliers in the 'UnitPrice' column
# Replace any unit price greater than 10 with the median unit price
median_unit_price = df['UnitPrice'].median()
df.loc[df['UnitPrice'] > 10, 'UnitPrice'] = median_unit_price
```

```
# 6. Data Validation (if any)
# Let's perform a simple validation on the 'Quantity' column to ensure it's positive
df = df[df['Quantity'] > 0]

# 7. Data Transformation (if any)
# Let's create a new column 'TotalPrice' by multiplying Quantity and UnitPrice
df['TotalPrice'] = df['Quantity'] * df['UnitPrice']

# Display the cleaned dataset
print("\nCleaned Dataset:")
print(df)
```

```
Original Dataset:
      InvoiceNo StockCode                          Description  Quantity  \
0        536365    85123A   WHITE HANGING HEART T-LIGHT HOLDER         6
1        536365     71053                  WHITE METAL LANTERN         6
2        536365    84406B       CREAM CUPID HEARTS COAT HANGER         8
3        536365    84029G  KNITTED UNION FLAG HOT WATER BOTTLE         6
4        536365    84029E       RED WOOLLY HOTTIE WHITE HEART.         6
...         ...       ...                                  ...       ...
87943    543737     21216     SET 3 RETROSPOT TEA,COFFEE,SUGAR        20
87944    543737    35810B     ENAMEL BLUE RIM COFFEE CONTAINER         6
87945    543737     22482        BLUE TEA TOWEL CLASSIC DESIGN        12
87946    543737     22900       SET 2 TEA TOWELS I LOVE LONDON         6
87947    543737    84968A       SET OF 16 VINTAGE ROSE CUTLERY         5

           InvoiceDate  UnitPrice  CustomerID         Country
0       12/1/2010 8:26       2.55     17850.0  United Kingdom
1       12/1/2010 8:26       3.39     17850.0  United Kingdom
2       12/1/2010 8:26       2.75     17850.0  United Kingdom
3       12/1/2010 8:26       3.39     17850.0  United Kingdom
4       12/1/2010 8:26       3.39     17850.0  United Kingdom
...                ...        ...         ...             ...
87943  2/11/2011 12:45       4.95     12477.0         Germany
87944  2/11/2011 12:45       2.10     12477.0         Germany
87945  2/11/2011 12:45       1.25     12477.0         Germany
87946  2/11/2011 12:45       2.95     12477.0         Germany
87947  2/11/2011 12:45      12.75     12477.0         Germany

[87948 rows x 8 columns]
```

```
Cleaned Dataset:
      InvoiceNo StockCode                          Description  Quantity  \
0        536365    85123A   WHITE HANGING HEART T-LIGHT HOLDER         6
1        536365     71053                  WHITE METAL LANTERN         6
2        536365    84406B       CREAM CUPID HEARTS COAT HANGER         8
3        536365    84029G  KNITTED UNION FLAG HOT WATER BOTTLE         6
4        536365    84029E  RED WOOLLY HOTTIE WITH WHITE HEART.         6
...         ...       ...                                  ...       ...
87943    543737     21216     SET 3 RETROSPOT TEA,COFFEE,SUGAR        20
87944    543737    35810B     ENAMEL BLUE RIM COFFEE CONTAINER         6
87945    543737     22482        BLUE TEA TOWEL CLASSIC DESIGN        12
87946    543737     22900       SET 2 TEA TOWELS I LOVE LONDON         6
87947    543737    84968A       SET OF 16 VINTAGE ROSE CUTLERY         5

              InvoiceDate  UnitPrice  CustomerID         Country  TotalPrice
0     2010-12-01 08:26:00       2.55     17850.0  United Kingdom       15.30
1     2010-12-01 08:26:00       3.39     17850.0  United Kingdom       20.34
2     2010-12-01 08:26:00       2.75     17850.0  United Kingdom       22.00
3     2010-12-01 08:26:00       3.39     17850.0  United Kingdom       20.34
4     2010-12-01 08:26:00       3.39     17850.0  United Kingdom       20.34
...                   ...        ...         ...             ...         ...
87943 2011-02-11 12:45:00       4.95     12477.0         Germany       99.00
87944 2011-02-11 12:45:00       2.10     12477.0         Germany       12.60
87945 2011-02-11 12:45:00       1.25     12477.0         Germany       15.00
87946 2011-02-11 12:45:00       2.95     12477.0         Germany       17.70
87947 2011-02-11 12:45:00       2.51     12477.0         Germany       12.55

[85317 rows x 9 columns]
```

Conclusion:

In conclusion, conducting data cleaning on social media data using Python is essential for ensuring data accuracy and reliability in subsequent analyses. By addressing issues such as missing values, duplicates, and inconsistencies, we enhance the quality of the dataset, leading to more accurate insights and informed decision-making. Data cleaning enables us to maximize the utility of social media data, facilitating effective strategies, and informed decision-making in various domains, including marketing, customer engagement, and sentiment analysis. Ultimately, a clean and reliable dataset serves as a solid foundation for deriving meaningful insights and driving actionable outcomes in the realm of social media analytics.