Day-18 228

**Cleaning and Preparing a Travel Customer Reviews Dataset**

**Objective:**

To clean and prepare a dataset containing customer reviews of travel experiences.

**Instructions:**

1. **Load the dataset** containing customer reviews, ratings, and feedback.
2. **Handle missing values**:
   * Identify missing values in Review\_Text, Rating, and Customer\_Age.
   * Impute missing numerical values using appropriate techniques.
   * Use NLP-based techniques to handle missing textual data.
3. **Detect and remove duplicates**:
   * Use duplicated() to find repeated reviews.
   * Remove or merge duplicate records.
4. **Handle inconsistent data**:
   * Standardize Rating values (ensure they range between 1-5).
   * Correct spelling inconsistencies in Tour\_Package names.
5. **Identify and handle outliers**:
   * Use boxplots to find anomalies in Package\_Price and Rating.
   * Apply transformation techniques if necessary.
6. **Prepare cleaned data for analysis**:
   * Convert categorical data into numerical format where required.
   * Save the final cleaned dataset as a CSV file.

Program:

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

from sklearn.impute import SimpleImputer

from sklearn.preprocessing import LabelEncoder

df = pd.read\_csv('customer\_reviews.csv')

df.head()

missing\_data = df.isna().sum()

missing\_percentage = (missing\_data / len(df)) \* 100

print("Missing Data Count:\n", missing\_data)

print("Missing Data Percentage:\n", missing\_percentage)

num\_columns = ['Rating', 'Customer\_Age']

imputer = SimpleImputer(strategy='median')

df[num\_columns] = imputer.fit\_transform(df[num\_columns])

df['Review\_Text'] = df['Review\_Text'].fillna('No review provided')

duplicates = df[df.duplicated(subset=['Review\_Text'])]

print(f"Number of Duplicate Reviews: {duplicates.shape[0]}")

df = df.drop\_duplicates(subset=['Review\_Text'])

df['Rating'] = df['Rating'].clip(lower=1, upper=5)

df['Tour\_Package'] = df['Tour\_Package'].str.lower()

plt.figure(figsize=(12, 6))

sns.boxplot(x=df['Package\_Price'])

plt.title('Boxplot for Package\_Price')

plt.show()

plt.figure(figsize=(12, 6))

sns.boxplot(x=df['Rating'])

plt.title('Boxplot for Rating')

plt.show()

Q1\_price = df['Package\_Price'].quantile(0.25)

Q3\_price = df['Package\_Price'].quantile(0.75) IQR\_price = Q3\_price - Q1\_price

lower\_bound\_price = Q1\_price - 1.5 \* IQR\_price

upper\_bound\_price = Q3\_price + 1.5 \* IQR\_price

df['Package\_Price'] = np.clip(df['Package\_Price'], lower\_bound\_price, upper\_bound\_price)

df['Rating'] = df['Rating'].clip(lower=1, upper=5)

label\_encoder = LabelEncoder()

df['Tour\_Package'] = label\_encoder.fit\_transform(df['Tour\_Package'])

df.to\_csv('cleaned\_travel\_reviews.csv', index=False)

print("Data cleaning complete. Cleaned data saved to 'cleaned\_travel\_reviews.csv'.")

Output:

Missing Data Count:

Review\_Text 0

Rating 2

Customer\_Age 5

Tour\_Package 0

Package\_Price 0

dtype: int64

Missing Data Percentage:

Review\_Text 0.0

Rating 0.5

Customer\_Age 1.25

Tour\_Package 0.0

Package\_Price 0.0

dtype: float64

Number of Duplicate Reviews: 3