# exp 7

Okay, here's the complete experiment report with the AIM, theory, algorithm, code, output, and other sections, including the handling of outliers, standardization, and formatting issues:

**Experiment Number:** 7

**AIM:** To perform comprehensive data cleaning and preprocessing on the Titanic dataset, including handling missing values, outliers, inconsistencies, and formatting issues, to prepare the data for accurate and reliable analysis.

**Objectives:**

- To identify and handle missing values in the dataset using appropriate imputation techniques.
- To detect and handle outliers to prevent them from distorting the analysis.
- To address inconsistencies in the data, such as duplicate entries.
- To correct formatting issues in string variables to ensure consistency.

**Course Outcomes:** CO2, CO3 (These likely relate to data preprocessing, data cleaning, and data quality)

**Resource/Tools:** Jupyter Notebook, Python with Pandas, NumPy, Scikit-learn, and Matplotlib libraries.

**Theory:**

Data cleaning and preprocessing are crucial steps in any data analysis project. Real-world data often contains imperfections that can affect the accuracy and reliability of the analysis.

- **Missing Values:** Data points that are not available or were not recorded.
- **Outliers:** Extreme values that deviate significantly from the rest of the data.
- **Inconsistencies:** Duplicate entries, contradictory information, or formatting issues.

**Algorithm:**

1. **Load the dataset:** Import the dataset into a Pandas DataFrame.
2. **Scan for missing values:** Identify and count missing values in each column.
3. **Handle missing values:** Apply imputation techniques to fill missing values.
4. **Scan for inconsistencies:** Check for duplicate rows and other inconsistencies.
5. **Handle inconsistencies:** Remove duplicate rows or correct inconsistencies.
6. **Detect outliers:** Use visualization techniques (e.g., box plots) or statistical methods to identify outliers.

7. **Handle outliers:** Apply appropriate techniques to deal with outliers (e.g., capping, winsorizing, transformation).
8. **Correct formatting issues:** Standardize string formats (e.g., capitalization, spacing).

**Code:**

Python

```
import pandas as pd
import numpy as np
from sklearn.impute import SimpleImputer

# Load the dataset
data = pd.read_csv('titanic.csv')

# 1. Missing Values
print("Missing values per column:")
print(data.isnull().sum())

# Impute missing values (using SimpleImputer)
numerical_features = ['Age', 'Fare']
imputer = SimpleImputer(strategy='mean')
data[numerical_features] = imputer.fit_transform(data[numerical_features])

categorical_features = ['Embarked']
imputer = SimpleImputer(strategy='most_frequent')
data[categorical_features] =
imputer.fit_transform(data[categorical_features])

# 2. Inconsistencies (Duplicates)
print("\nNumber of duplicate rows:", data.duplicated().sum())
data.drop_duplicates(inplace=True)

# 3. Outliers (Example - 'Fare')
# (Using IQR method - you might need to adjust this for your data)
Q1 = data['Fare'].quantile(0.25)
Q3 = data['Fare'].quantile(0.75)
IQR = Q3 - Q1
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR
data = data[(data['Fare'] >= lower_bound) & (data['Fare'] <= upper_bound)]

# 4. Standardization (Example - 'Age')
data['Age'] = (data['Age'] - data['Age'].mean()) / data['Age'].std()
```

```
# 5. Formatting Issues (Example - 'Name')
data['Name'] = data['Name'].str.title()  # Capitalize names

# Display the cleaned data
print("\nCleaned data:")
print(data.head())
```

**Output:**

The code will produce the following output:

- The number of missing values in each column.
- The number of duplicate rows.
- The first 5 rows of the cleaned dataset.

**Result:**

The experiment successfully demonstrates a comprehensive data cleaning process, including:

- Handling missing values using imputation.
- Removing duplicate rows.
- Handling outliers in the `Fare` column using the IQR method.
- Standardizing the `Age` column.
- Correcting formatting issues in the `Name` column by capitalizing the names.

**Learning Outcomes:**

- Applied various data cleaning techniques to handle missing values, outliers, inconsistencies, and formatting issues.
- Understood the importance of each step in the data cleaning process.
- Gained practical experience in using Pandas and Scikit-learn for data preprocessing.
- Developed skills to prepare data for reliable analysis and model building.