**SVKM’s NMIMS**

**Mukesh Patel School of Technology Management & Engineering**

Program: BTI Computer Engineering

**Course: Data Mining**

**Experiment No.03**

PART A

(PART A: TO BE REFFERED BY STUDENTS)

**A.1 Aim:**

Apply Data Cleaning techniques for categorical, numerical attributes and observe the effect on data dispersion

**A.2 Prerequisite:**

**Python packages – matplotlib, pandas, numpy etc.**

**A.3 Outcome:**

**After successful completion of this experiment students will be able to:**

**Answer following questions:**

1. Different types of attribute in dataset
2. Count of Missing Values in Numerical and Categorical attribute
3. Handling the Missing Values
4. Handling Invalid values
5. Give the Count of Unique value for each attribute
6. Converting categorical data to numeric data
7. Visualizing the data after data cleaning using boxplot
8. Comment on the skewness of the data and remove outlier if any

**Task**

Select the dataset of your choice, clean the dataset and answer the above questions

PART B

(PART B: TO BE COMPLETED BY STUDENTS)

***(Students must submit the soft copy as per following segments within two hours of the practical. The soft copy must be uploaded on the Blackboard or emailed to the concerned lab in charge faculties at the end of the practical in case there is no Black board access available)***

|  |  |
| --- | --- |
| **Roll No.: C026** | **Name: Anirbaan Ghatak** |
| **Class : B** | **Batch : B1** |
| **Date of Experiment:** | **Date of Submission:** 15/08/2023 |
| **Grade :** |  |

**B.1 Answers of Task to be written by student:**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

df = pd.read\_csv("hotel\_bookings.csv")

df.columns

def remove\_outliers(df, column\_name):

    quartile\_1 = np.round(df[column\_name].quantile(0.25), 2)

    quartile\_3 = np.round(df[column\_name].quantile(0.75), 2)

    # Interquartile range

    iqr = np.round(quartile\_3 - quartile\_1, 2)

    lower\_bound = quartile\_1 - 1.5 \* iqr

    upper\_bound = quartile\_3 + 1.5 \* iqr

    no\_out = df[df[column\_name].between(lower\_bound, upper\_bound, inclusive='both')].copy()

    return no\_out

#missing values in the dataset

df.isnull().sum()

#fill null values

df.fillna(0, inplace=True)

#count of unique values in each column

df.nunique()

#market segment distrinbution channels and customer type and reservation status to numerical values

dum1 = pd.get\_dummies(df.market\_segment)

dum2 = pd.get\_dummies(df.distribution\_channel)

dum3 = pd.get\_dummies(df.customer\_type)

dum4 = pd.get\_dummies(df.reservation\_status)

df = pd.concat([df,dum1,dum2,dum3,dum4],axis=1)

df.drop(['market\_segment','distribution\_channel','customer\_type','reservation\_status'],axis=1,inplace=True)

columns\_to\_plot = ['lead\_time', 'stays\_in\_weekend\_nights', 'stays\_in\_week\_nights', 'adr']

for col in columns\_to\_plot:

    plt.figure(figsize=(10, 5))

    sns.boxplot(y = df[col])

    plt.show()

for col in columns\_to\_plot:

    df\_new = remove\_outliers(df, col)

    plt.figure(figsize=(10, 5))

    sns.boxplot(y=df\_new[col])

    plt.title(f'Boxplot of {col}')

    plt.show()

**B.2 Observations and learning:**

***Significance of numerical columns in data mining. Recognized the necessity of meticulous data cleaning techniques, demonstrated here through the removal of outliers using the interquartile range (IQR) method. This process enhances the integrity and reliability of subsequent analyses, ensuring accurate and meaningful results.***

**B.3 Conclusion:**

***Understood the concept behind the importance of numerical column and data cleaning.***