# Green University of Bangladesh
## Department of Computer Science and Engineering (CSE)
**Faculty of Sciences and Engineering**
**Semester: (Fall, Year:2023),**
**B.Sc. in CSE(Day)**

# LAB REPORT NO: 03

**Course Title: Data Mining Lab**
**Course Code: CSE-436**
**Section: D6(201)**

**Lab Experiment Name: Data Preprocessing Techniques**

**Student Details**

| | Name | ID |
|---|---|---|
| 1. | Md Zahid Hasan | 201902060 |

**Lab Date** : 29/09/2023
**Submission Date** : 12/10/2023
**Course Teacher's Name** : Abdullah Al Farhad

**Lab Report Status**

**Marks:** ..... **Signature:** .....
**Comments:** ..... **Date:** .....

# 1 INTRODUCTION

In this lab report we are going develop a Colab using all of the data processing. We will Choose an appropriate dataset from kaggle containing NULL and garbage values then will implement all techniques with proper documentation in the notebook.

# 2 OBJECTIVE

The aim of this lab is to transform raw data more easily and effectively processed in data mining, machine learning and other data science tasks.To transform raw data more effectively we are going to know different techniques for Data Cleaning and Data transformation.

# 3 DATASET

For this lab I use the "House Prices: Advanced Regression Techniques" dataset from Kaggle which is a well-known dataset. Eventually I just used the training csv dataset to work in various data preprocessing steps such as handling missing values, dealing with categorical variables.

# 4 IMPLEMENTATION

```
import pandas as pd
import numpy as np

# Load the dataset
df = pd.read_csv('/content/train.csv')  # Provide the appropriate path

# Display the first few rows of the dataset
print("Original Dataset:")
print(df.head())
```

Listing 1: Importing the librarys and Loading the dataset

```
# 1. Checking for NULL and garbage values
print("\nChecking for NULL and garbage values:")
print(df.isnull().sum())  # Check for NULL values
```

Listing 2: Checking for NULL and garbage values

```
df_dropped = df.dropna()  # Drop rows with any NULL values
print("\nDataset after dropping NULL values:")
print(df_dropped.head())
```

Listing 3: Dropping NULL values

```
df_filled = df.fillna(0)  # Fill NULL values with 0
print("\nDataset after filling NULL values:")
print(df_filled.head())
```

Listing 4: Filling NULL values

```
df_interpolated = df.interpolate()  # Interpolate the NULL values
print("\nDataset after interpolation:")
print(df_interpolated.head())
```

Listing 5: Interpolation

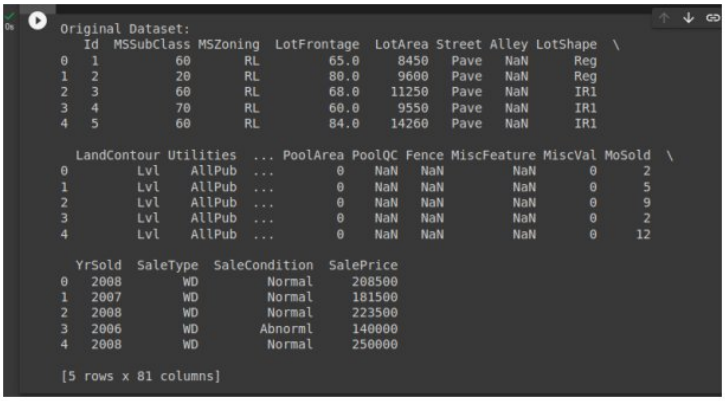
```
df_cleaned = df.replace({'column_name': {np.nan: 0}})  # Replace NULL
    values in a specific column
print("\nDataset after handling garbage values:")
print(df_cleaned.head())
```

Listing 6: Handling garbage values

```
df.to_csv('processed_dataset.csv', index=False)
```

Listing 7: Saving the processed data to a CSV file

# 5 OUTPUT



```
Original Dataset:
   Id  MSSubClass MSZoning  LotFrontage  LotArea Street Alley LotShape  \
0   1          60       RL         65.0     8450   Pave   NaN      Reg
1   2          20       RL         80.0     9600   Pave   NaN      Reg
2   3          60       RL         68.0    11250   Pave   NaN      IR1
3   4          70       RL         60.0     9550   Pave   NaN      IR1
4   5          60       RL         84.0    14260   Pave   NaN      IR1

  LandContour Utilities  ... PoolArea PoolQC Fence MiscFeature MiscVal MoSold  \
0         Lvl    AllPub  ...        0    NaN   NaN         NaN       0      2
1         Lvl    AllPub  ...        0    NaN   NaN         NaN       0      5
2         Lvl    AllPub  ...        0    NaN   NaN         NaN       0      9
3         Lvl    AllPub  ...        0    NaN   NaN         NaN       0      2
4         Lvl    AllPub  ...        0    NaN   NaN         NaN       0     12

  YrSold  SaleType  SaleCondition  SalePrice
0   2008        WD         Normal     208500
1   2007        WD         Normal     181500
2   2008        WD         Normal     223500
3   2006        WD        Abnorml     140000
4   2008        WD         Normal     250000

[5 rows x 81 columns]
```

Figure 1: Loading the dataset

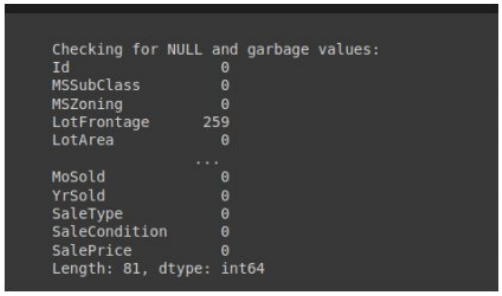



Figure 2: Checking for NULL and garbage values


Figure 3: Dropping NULL values

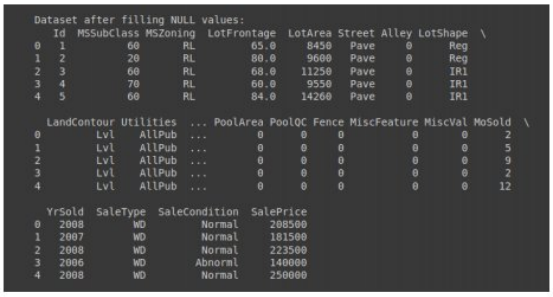
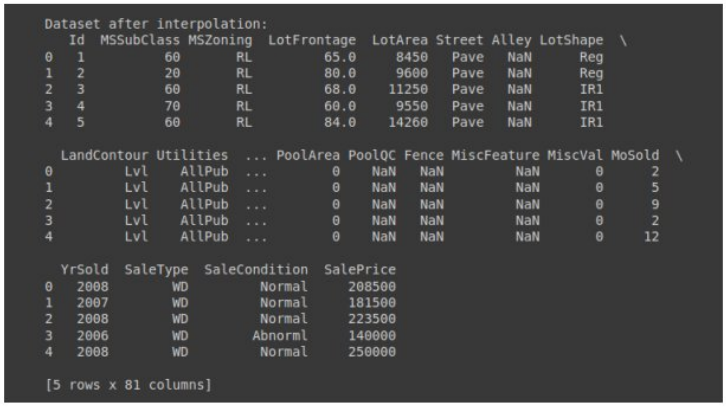



Figure 4: Filling NULL values



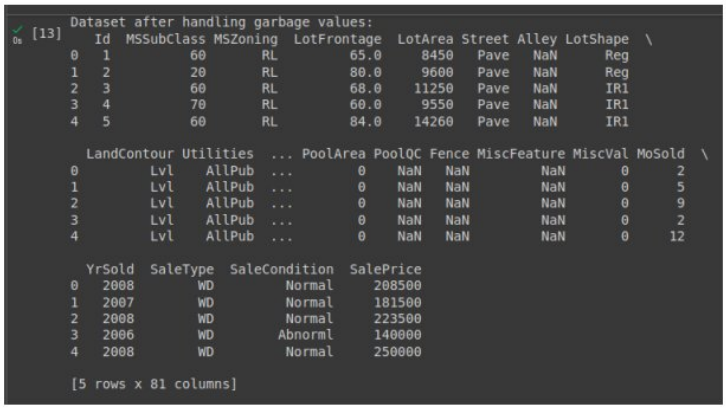
```
Dataset after interpolation:
   Id  MSSubClass MSZoning  LotFrontage  LotArea Street Alley LotShape  \
0   1          60       RL         65.0     8450   Pave   NaN      Reg
1   2          20       RL         80.0     9600   Pave   NaN      Reg
2   3          60       RL         68.0    11250   Pave   NaN      IR1
3   4          70       RL         60.0     9550   Pave   NaN      IR1
4   5          60       RL         84.0    14260   Pave   NaN      IR1

  LandContour Utilities  ... PoolArea PoolQC Fence MiscFeature MiscVal MoSold  \
0         Lvl    AllPub  ...        0    NaN   NaN         NaN       0      2
1         Lvl    AllPub  ...        0    NaN   NaN         NaN       0      5
2         Lvl    AllPub  ...        0    NaN   NaN         NaN       0      9
3         Lvl    AllPub  ...        0    NaN   NaN         NaN       0      2
4         Lvl    AllPub  ...        0    NaN   NaN         NaN       0     12

  YrSold  SaleType  SaleCondition  SalePrice
0   2008        WD         Normal     208500
1   2007        WD         Normal     181500
2   2008        WD         Normal     223500
3   2006        WD        Abnorml     140000
4   2008        WD         Normal     250000

[5 rows x 81 columns]
```

Figure 5: Interpolation



```
[13]
Dataset after handling garbage values:
   Id  MSSubClass MSZoning  LotFrontage  LotArea Street Alley LotShape  \
0   1          60       RL         65.0     8450   Pave   NaN      Reg
1   2          20       RL         80.0     9600   Pave   NaN      Reg
2   3          60       RL         68.0    11250   Pave   NaN      IR1
3   4          70       RL         60.0     9550   Pave   NaN      IR1
4   5          60       RL         84.0    14260   Pave   NaN      IR1

  LandContour Utilities  ... PoolArea PoolQC Fence MiscFeature MiscVal MoSold  \
0         Lvl    AllPub  ...        0    NaN   NaN         NaN       0      2
1         Lvl    AllPub  ...        0    NaN   NaN         NaN       0      5
2         Lvl    AllPub  ...        0    NaN   NaN         NaN       0      9
3         Lvl    AllPub  ...        0    NaN   NaN         NaN       0      2
4         Lvl    AllPub  ...        0    NaN   NaN         NaN       0     12

  YrSold  SaleType  SaleCondition  SalePrice
0   2008        WD         Normal     208500
1   2007        WD         Normal     181500
2   2008        WD         Normal     223500
3   2006        WD        Abnorml     140000
4   2008        WD         Normal     250000

[5 rows x 81 columns]
```

Figure 6: Handling garbage values

# 6 DISCUSSION & ANALYSIS

In this lab report I performed various data processing techniques on a given dataset. The techniques included checking for NULL and garbage values, dropping NULL values, filling NULL values, interpolation, handling garbage values, and data type conversion. Each of these techniques plays a vital role in ensuring data integrity and preparing the dataset for further analysis and modeling. By implementing these methods, I was able to enhance the robustness and reliability of the dataset for subsequent analytical procedures.