**Batch: H2-1 Roll No.: 16010122151**

**Experiment:-02**

|  |
| --- |
| **Title:**  Dataset pre-processing |

# Objective:

# 1. To learn how to prepare the dataset

# 2. To learn various steps in Data -Preprocessing

# Course Outcome:

# CO1: Learn how to locate and download datasets, extract insights from that data and present their findings in a variety of different formats.

# Books/ Journals/ Websites referred:

Google

Kaggle

Wikipedia

# Resources used:

# Kaggle

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Theory (About Data Preprocessing):

(Students should write)

# Following points should be written by students

# Different steps in Data Preprocessing:

# Finding missing, null values

# Replacing missing, null values with statistical parameters

# Encoding categorical data

# Normalization

# Note: Student can use any technology like Tableau, Tableau-Prep, PowerBI, Google spreadsheet, excel, R programming, Python, Java any other technology for preprocessing.

# Data preprocessing is a crucial step in preparing raw. It involves cleaning, transforming, and organizing data to make it suitable for further processing.

# Finding Missing and Null Values: This step involves identifying cells or entries in our dataset tthat does not have values (missing values) or have placeholder values like null values.

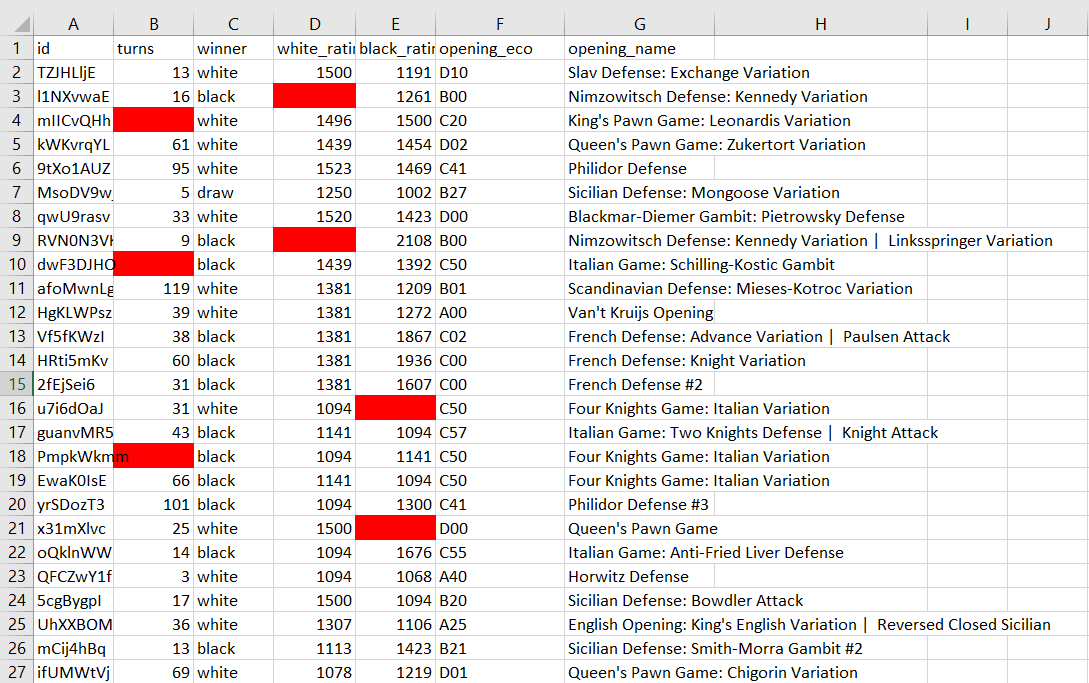
# Replacing Missing and Null Values: After identifying missing or null values, you might choose to handle them by filling in appropriate values. Common approach is using the mean, median or mode to replace missing values. The choice of method depends on the nature of the data analysis.

# Encoding Categorical Data: Many machine learning algorithms require numerical input, so categorical data (data with categories or labels) needs to be converted into numerical form. This process is called encoding.. Label Encoding assigns a unique number to each category,while One-Hot Encoding creates binary columns for each category.

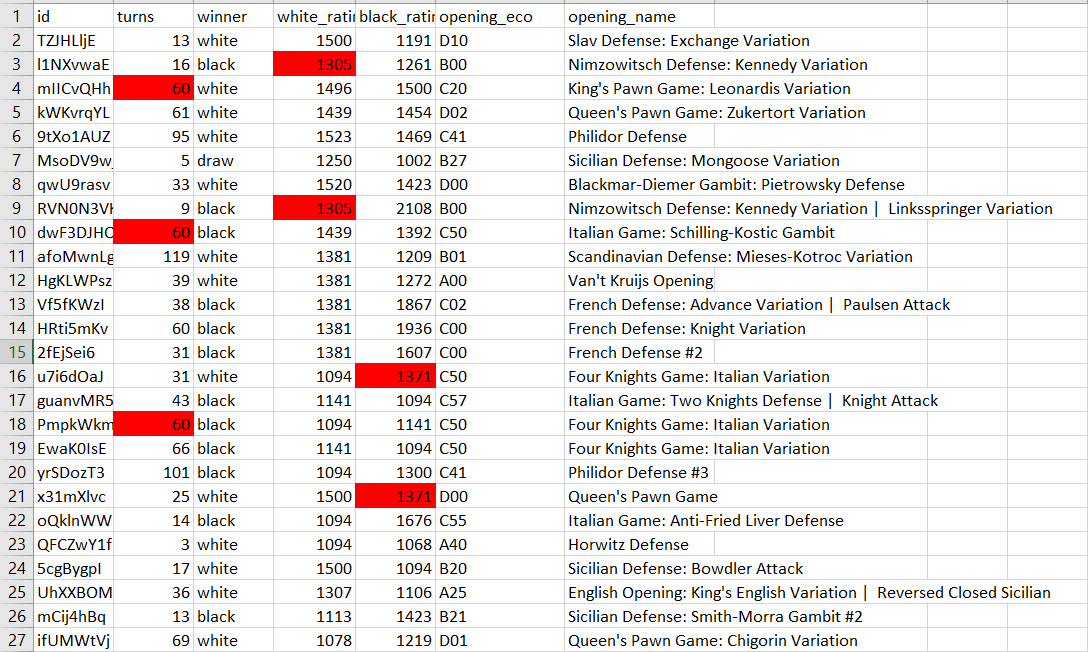
# Normalization: Normalization ensures that numerical features are on a similar scale, prevents any feature from dominating others in the analysis.

# Platform used by the student: Excel

# Working (Paste the code and Out for each Data Preprocessing task):

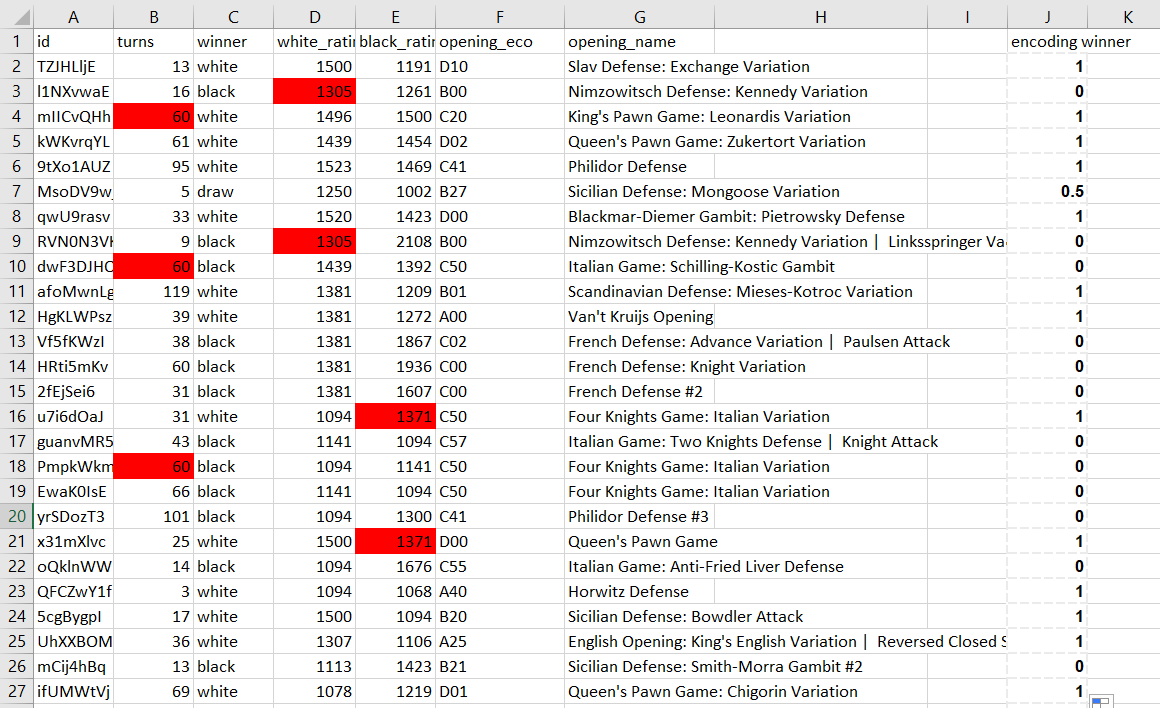
****

Replacing missing values:



Encoding Categorical data:

Ive encoded white wins as 1 white losses as 0 and draw as 0.5



# Conclusion (Students should write in their own words):

# We learned how to fill in the missing values in using different platforms using some functions like average and how to normalize data with the IFS function in excel

# Post Lab Question:

# Write the importance of Data Preprocessing. It improves accuracy and reliability.

Preprocessing data removes missing or inconsistent data values resulting from human or computer error, which improves the accuracy and quality of a dataset, making it more reliable.