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.

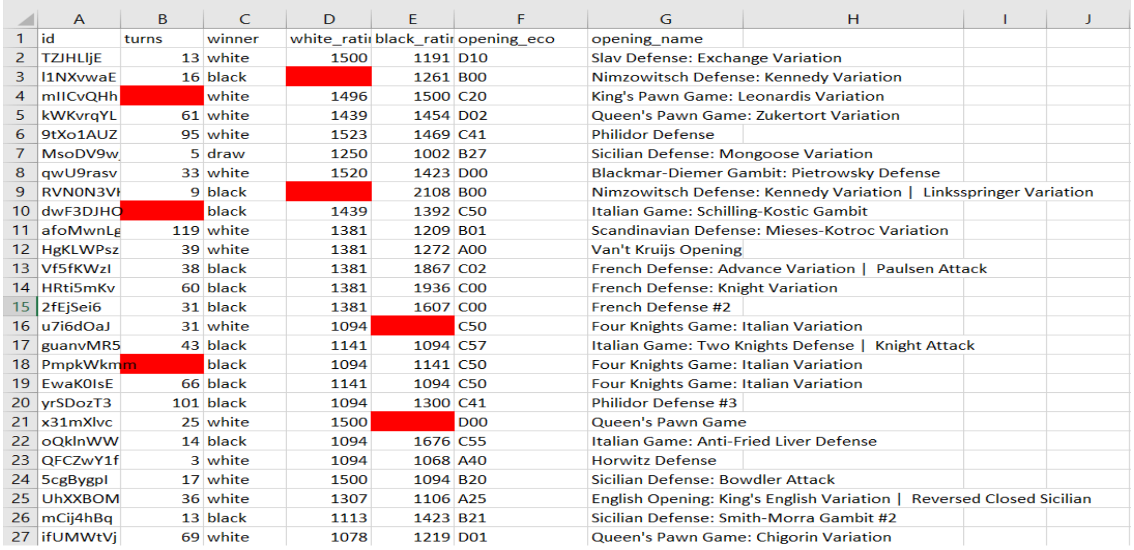
Replacing Missing and Null Values: After identifying missing or null values, you might choose to handle them by filling in appropriate values. Common approach is using the mean, median or mode to replace missing values. The choice of method depends on the nature of the data analysis.

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.

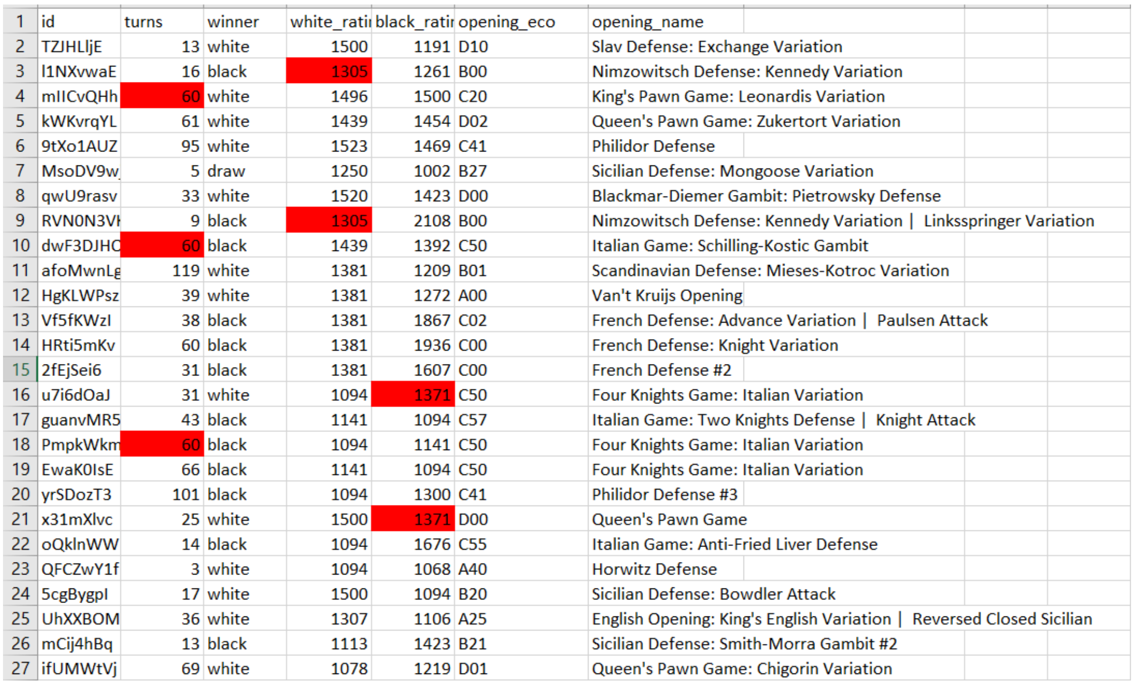
Platform used by the student: Excel

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



	A	B	C	D	E	F	G	H	I	J
1	id	turns	winner	white_ratii	black_ratii	opening_eco	opening_name			
2	TZJHLljE	13	white	1500	1191	D10	Slav Defense: Exchange Variation			
3	l1NXvwaE	16	black	🟥	1261	B00	Nimzowitsch Defense: Kennedy Variation			
4	mIICvQHh	🟥	white	1496	1500	C20	King's Pawn Game: Leonardis Variation			
5	kWKvrqYL	61	white	1439	1454	D02	Queen's Pawn Game: Zukertort Variation			
6	9tXo1AUZ	95	white	1523	1469	C41	Philidor Defense			
7	MsoDV9w	5	draw	1250	1002	B27	Sicilian Defense: Mongoose Variation			
8	qwU9rasv	33	white	1520	1423	D00	Blackmar-Diemer Gambit: Pietrowsky Defense			
9	RVN0N3VI	9	black	🟥	2108	B00	Nimzowitsch Defense: Kennedy Variation | Linksspringer Variation			
10	dwF3DJHO	🟥	black	1439	1392	C50	Italian Game: Schilling-Kostic Gambit			
11	afoMwnLg	119	white	1381	1209	B01	Scandinavian Defense: Mieses-Kotroc Variation			
12	HgKLWPsz	39	white	1381	1272	A00	Van't Kruijs Opening			
13	Vf5fKWzI	38	black	1381	1867	C02	French Defense: Advance Variation | Paulsen Attack			
14	HRti5mKv	60	black	1381	1936	C00	French Defense: Knight Variation			
15	2fEjSei6	31	black	1381	1607	C00	French Defense #2			
16	u7i6dOaJ	31	white	1094	🟥	C50	Four Knights Game: Italian Variation			
17	guanvMR5	43	black	1141	1094	C57	Italian Game: Two Knights Defense | Knight Attack			
18	PmpkWkmm	🟥	black	1094	1141	C50	Four Knights Game: Italian Variation			
19	EwaK0IsE	66	black	1141	1094	C50	Four Knights Game: Italian Variation			
20	yrSDozT3	101	black	1094	1300	C41	Philidor Defense #3			
21	x31mXlvc	25	white	1500	🟥	D00	Queen's Pawn Game			
22	oQklnWW	14	black	1094	1676	C55	Italian Game: Anti-Fried Liver Defense			
23	QFCZwY1f	3	white	1094	1068	A40	Horwitz Defense			
24	5cgBygpI	17	white	1500	1094	B20	Sicilian Defense: Bowdler Attack			
25	UhXXBOM	36	white	1307	1106	A25	English Opening: King's English Variation | Reversed Closed Sicilian			
26	mCij4hBq	13	black	1113	1423	B21	Sicilian Defense: Smith-Morra Gambit #2			
27	ifUMWtVj	69	white	1078	1219	D01	Queen's Pawn Game: Chigorin Variation			

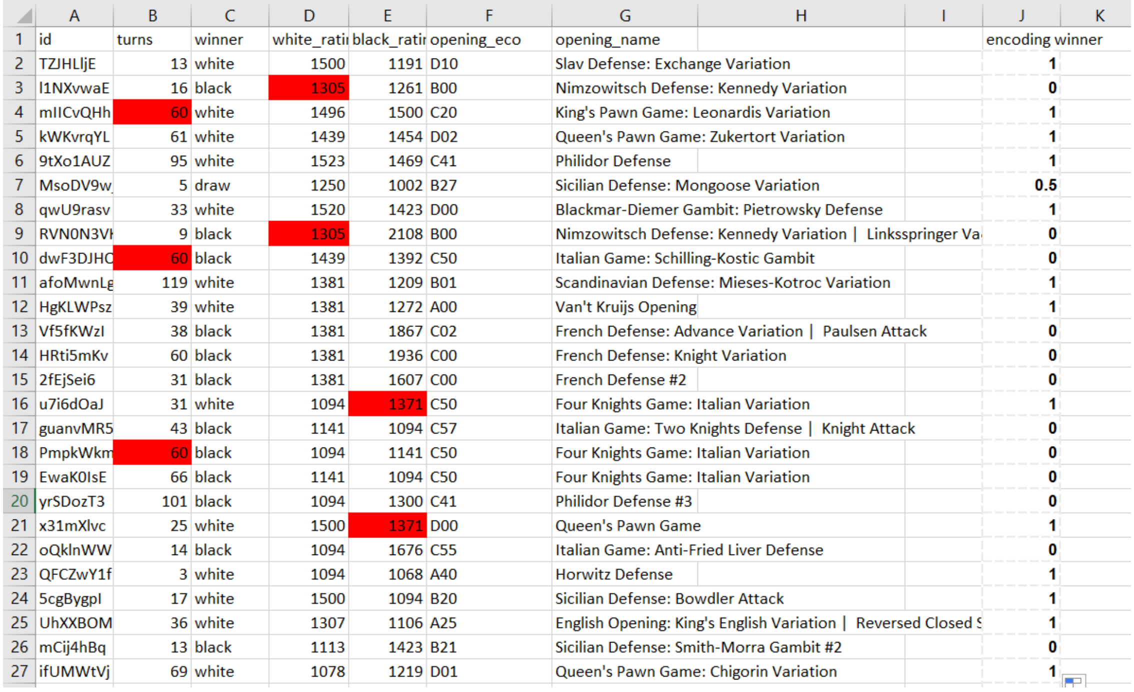
Replacing missing values:



1	id	turns	winner	white_ratii	black_ratii	opening_eco	opening_name
2	TZJHLljE	13	white	1500	1191	D10	Slav Defense: Exchange Variation
3	l1NXvwaE	16	black	1305	1261	B00	Nimzowitsch Defense: Kennedy Variation
4	mIICvQHh	60	white	1496	1500	C20	King's Pawn Game: Leonardis Variation
5	kWKvrqYL	61	white	1439	1454	D02	Queen's Pawn Game: Zukertort Variation
6	9tXo1AUZ	95	white	1523	1469	C41	Philidor Defense
7	MsoDV9w	5	draw	1250	1002	B27	Sicilian Defense: Mongoose Variation
8	qwU9rasv	33	white	1520	1423	D00	Blackmar-Diemer Gambit: Pietrowsky Defense
9	RVN0N3Vk	9	black	1305	2108	B00	Nimzowitsch Defense: Kennedy Variation | Linksspringer Variation
10	dwF3DJHC	60	black	1439	1392	C50	Italian Game: Schilling-Kostic Gambit
11	afoMwnLg	119	white	1381	1209	B01	Scandinavian Defense: Mieses-Kotroc Variation
12	HgKLWPsz	39	white	1381	1272	A00	Van't Kruijs Opening
13	Vf5fKWzI	38	black	1381	1867	C02	French Defense: Advance Variation | Paulsen Attack
14	HRti5mKv	60	black	1381	1936	C00	French Defense: Knight Variation
15	2fEjSei6	31	black	1381	1607	C00	French Defense #2
16	u7i6dOaJ	31	white	1094	1371	C50	Four Knights Game: Italian Variation
17	guanvMR5	43	black	1141	1094	C57	Italian Game: Two Knights Defense | Knight Attack
18	PmpkWkm	60	black	1094	1141	C50	Four Knights Game: Italian Variation
19	EwaK0IsE	66	black	1141	1094	C50	Four Knights Game: Italian Variation
20	yrSDozT3	101	black	1094	1300	C41	Philidor Defense #3
21	x31mXlvc	25	white	1500	1371	D00	Queen's Pawn Game
22	oQklnWW	14	black	1094	1676	C55	Italian Game: Anti-Fried Liver Defense
23	QFCZwY1f	3	white	1094	1068	A40	Horwitz Defense
24	5cgBygpl	17	white	1500	1094	B20	Sicilian Defense: Bowdler Attack
25	UhXXBOM	36	white	1307	1106	A25	English Opening: King's English Variation | Reversed Closed Sicilian
26	mCij4hBq	13	black	1113	1423	B21	Sicilian Defense: Smith-Morra Gambit #2
27	ifUMWtVj	69	white	1078	1219	D01	Queen's Pawn Game: Chigorin Variation

Encoding Categorical data:

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



	A	B	C	D	E	F	G	H	I	J	K
1	id	turns	winner	white_ratii	black_ratii	opening_eco	opening_name			encoding winner	
2	TZJHLljE	13	white	1500	1191	D10	Slav Defense: Exchange Variation			1	
3	l1NXvwaE	16	black	1305	1261	B00	Nimzowitsch Defense: Kennedy Variation			0	
4	mIICvQHh	60	white	1496	1500	C20	King's Pawn Game: Leonardis Variation			1	
5	kWKvrqYL	61	white	1439	1454	D02	Queen's Pawn Game: Zukertort Variation			1	
6	9tXo1AUZ	95	white	1523	1469	C41	Philidor Defense			1	
7	MsoDV9w	5	draw	1250	1002	B27	Sicilian Defense: Mongoose Variation			0.5	
8	qwU9rasv	33	white	1520	1423	D00	Blackmar-Diemer Gambit: Pietrowsky Defense			1	
9	RVN0N3Vk	9	black	1305	2108	B00	Nimzowitsch Defense: Kennedy Variation | Linksspringer Va			0	
10	dwF3DJHC	60	black	1439	1392	C50	Italian Game: Schilling-Kostic Gambit			0	
11	afoMwnLg	119	white	1381	1209	B01	Scandinavian Defense: Mieses-Kotroc Variation			1	
12	HgKLWPsz	39	white	1381	1272	A00	Van't Kruijs Opening			1	
13	Vf5fKWzI	38	black	1381	1867	C02	French Defense: Advance Variation | Paulsen Attack			0	
14	HRti5mKv	60	black	1381	1936	C00	French Defense: Knight Variation			0	
15	2fEjSei6	31	black	1381	1607	C00	French Defense #2			0	
16	u7i6dOaJ	31	white	1094	1371	C50	Four Knights Game: Italian Variation			1	
17	guanvMR5	43	black	1141	1094	C57	Italian Game: Two Knights Defense | Knight Attack			0	
18	PmpkWkm	60	black	1094	1141	C50	Four Knights Game: Italian Variation			0	
19	EwaK0IsE	66	black	1141	1094	C50	Four Knights Game: Italian Variation			0	
20	yrSDozT3	101	black	1094	1300	C41	Philidor Defense #3			0	
21	x31mXlvc	25	white	1500	1371	D00	Queen's Pawn Game			1	
22	oQklnWW	14	black	1094	1676	C55	Italian Game: Anti-Fried Liver Defense			0	
23	QFCZwY1f	3	white	1094	1068	A40	Horwitz Defense			1	
24	5cgBygpl	17	white	1500	1094	B20	Sicilian Defense: Bowdler Attack			1	
25	UhXXBOM	36	white	1307	1106	A25	English Opening: King's English Variation | Reversed Closed S			1	
26	mCij4hBq	13	black	1113	1423	B21	Sicilian Defense: Smith-Morra Gambit #2			0	
27	ifUMWtVj	69	white	1078	1219	D01	Queen's Pawn Game: Chigorin Variation			1	

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:

1. **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.