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

Experiment 01

Title: Data Collection and finalizing dataset from problem domain

Objective:

1. To learn how to collect the dataset
2. To learn sources of dataset
3. To assess the dataset based on Metrics to Measure Data Quality
4. To finalize the features of dataset

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:

(Students should write)
https://www.kaggle.com/datasets/datasnaek/chess

Resources used:

(Students should write)
Google
Excel
Kaggle

Theory:

(Students should write)
The problem domain is Chess, a Game.

I have been interested in Chess since a long time and would like to conduct data analysis on a chess dataset to see some useful and interesting statistics.

The problem statement is to calculate the percentage of games won by white and black and number of draws, average length of games, average difference in elo, most common openings used.

Based on the problem statement we can find if white has an advantage because it moves first, what the most common openings played are, which opening has the highest win rate for white, the elo difference between players and other useful data which we can use to gain a deeper understanding of the game and its workings.

This dataset was found on Kaggle, it is a collection of 20,000 games played on Lichess an open source chess platform, which makes all its games and databases available to the public.

This dataset was perfect for use because the data is comprehensive and contains all the categories that will be needed for analysis as well as enough number of entries.

This dataset was obtained from:
https://www.kaggle.com/datasets/datasnaek/chess

	A	B	C	D	E	F	G	H
1	id	turns	winner	white_rat	black_rati	opening_	opening_name	
2	TZJHLljE	13	white	1500	1191	D10	Slav Defense: Exchange Variation	
3	l1NXvwaE	16	black	1322	1261	B00	Nimzowitsch Defense: Kennedy Variation	
4	mIICvQHh	61	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	

Id: it is the id of the game on lichess.org
Turns: the number of moves the game lasted
Winner: which player won the game white or black or draw
White_rating: Whites ELO
Black_rating: Blacks ELO
Opening_eco: code of the opening used
Opening_name: The name of the opening played.

Column Stats:
Average number of turns: 60.466
Average White elo: 1596.632
Average Black elo: 1588.832

Following points should be written by students

- Problem domain (Healthcare, Ecommerce, Education, Finance, agriculture etc.)
- Motivation for the selected Domain
- Problem Statement
- Brain stormed features in problem statement (Based on Domain Selected) and its important

- Search for dataset
- Justification for choosing above dataset
- Source of dataset (Link Needs to be given)
- Sample of Finalized dataset (First 5 Records)
- Data Dictionary
- Column wise summary

Conclusion (Students should write in their own words):

I have learned how to look for a relevant database design a problem statement and check whether the database is relevant or not.

Post Lab Question:

1. Explain Role of Data in the Application Design.

Data forms the cornerstone of our product development process; it can quickly inform development priorities for enhanced user experience, improved user satisfaction and increased adoption rates.
In order to keep up with the times, your designs need to be based on data. In this article, you'll learn how to include data as a core component of your design process.

2. **Write different types of Data with Example.**

   Nominal: Colour of hair (Blonde, red, Brown, Black, etc.)

   Ordinal: Letter grades in the exam (A, B, C, D, etc.)

   Discrete: Age in years

   Continous: Temperature