




Working with Structured Data

School Connect: Intro to DS & AI

A Aniruddha
Indian Institute of Technology, Madras



Who was the best opening batsman in the ICC T20 CWC?

King
Salt
Bumrah
Head
Virat
Rohit
Buttler
Klassen
Gous
Warner
Gurbaz
Zadran
Mayers
QDK



Types of Data

Structured Data

- ❖ Database
- ❖ Spreadsheet

Unstructured Data

- ❖ Text
- ❖ Audio
- ❖ Image
- ❖ Video



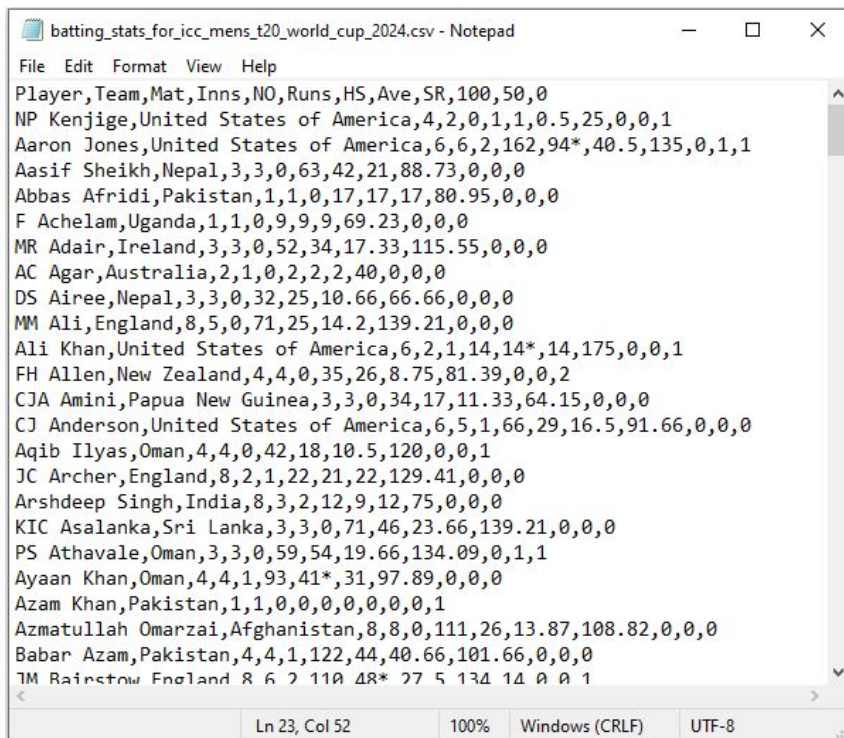
Where can I find data?

- <https://datasetsearch.research.google.com/>
- <https://github.com/awesomedata/awesome-public-datasets>
- <https://www.kaggle.com/datasets/>
- <https://www.data.gov.in/>

For our task,

https://www.kaggle.com/datasets/arindamsahoo/icc-mens-t20-world-cup-2024-stats?select=bowling_stats_for_icc_mens_t20_world_cup_2024.csv

What does the data look like?



```
batting_stats_for_icc_mens_t20_world_cup_2024.csv - Notepad
File Edit Format View Help
Player,Team,Mat,Inns,NO,Runs,HS,Ave,SR,100,50,0
NP Kenjige,United States of America,4,2,0,1,1,0.5,25,0,0,1
Aaron Jones,United States of America,6,6,2,162,94*,40.5,135,0,1,1
Aasif Sheikh,Nepal,3,3,0,63,42,21,88.73,0,0,0
Abbas Afridi,Pakistan,1,1,0,17,17,17,80.95,0,0,0
F Achelam,Uganda,1,1,0,9,9,9,69.23,0,0,0
MR Adair,Ireland,3,3,0,52,34,17.33,115.55,0,0,0
AC Agar,Australia,2,1,0,2,2,2,40,0,0,0
DS Airee,Nepal,3,3,0,32,25,10.66,66.66,0,0,0
MM Ali,England,8,5,0,71,25,14.2,139.21,0,0,0
Ali Khan,United States of America,6,2,1,14,14*,14,175,0,0,1
FH Allen,New Zealand,4,4,0,35,26,8.75,81.39,0,0,2
CJA Amini,Papua New Guinea,3,3,0,34,17,11.33,64.15,0,0,0
CJ Anderson,United States of America,6,5,1,66,29,16.5,91.66,0,0,0
Aqib Ilyas,Oman,4,4,0,42,18,10.5,120,0,0,1
JC Archer,England,8,2,1,22,21,22,129.41,0,0,0
Arshdeep Singh,India,8,3,2,12,9,12,75,0,0,0
KIC Asalanka,Sri Lanka,3,3,0,71,46,23.66,139.21,0,0,0
PS Athavale,Oman,3,3,0,59,54,19.66,134.09,0,1,1
Ayaan Khan,Oman,4,4,1,93,41*,31,97.89,0,0,0
Azam Khan,Pakistan,1,1,0,0,0,0,0,0,0,1
Azmatullah Omarzai,Afghanistan,8,8,0,111,26,13.87,108.82,0,0,0
Babar Azam,Pakistan,4,4,1,122,44,40.66,101.66,0,0,0
TM Bairstow,England,8,6,2,110,48*,27.5,134.14,0,0,1
```

Ln 23, Col 52 100% Windows (CRLF) UTF-8

CSV - Comma Separated Values

- File format
- Other examples - TSV , XLSX
- Compatibility

Can we view this as a table instead?

What does the data look like?

batting_stats_for_icc_mens_t20_world_cup_2024.csv - Notepad

File Edit Format View Help

Player,Team,Mat,Inns,NO,Runs,HS,Ave,SR,100,50,0

NP Kenjige,United States of America,4,2,0,1,1,0.5,25,0,0,1

Aaron Jones,United States of America,6,6,2,162,94*,40.5,135,0,1,1

Aasif Sheikh,Nepal,3,3,0,63,42,21,88.73,0,0,0

Abbas Afridi,Pakistan,1,1,0,17,17,17,80.95,0,0,0

F Achelam,Uganda,1,1,0,9,9,9,69.23,0,0,0

MR Adair,Ireland,3,3,0,52,34,17.33,115.55,0,0,0

AC Agar,Australia,2,1,0,2,2,2,40,0,0,0

DS Airee,Nepal,3,3,0,32,25,10.66,66.66,0,0,0

MM Ali,England,8,5,0,71,25,14.2,139.21,0,0,0

Ali Khan,United States of America,6,2,1,14,14*,14,175,0,0,1

FH Allen,New Zealand,4,4,0,35,26,8.75,81.39,0,0,2

CJA Amini,Papua New Guinea,3,3,0,34,17,11.33,64.15,0,0,0

CJ Anderson,United States of America,6,5,1,66,29,16.5,91.66,0,0,0

Aqib Ilyas,Oman,4,4,0,42,18,10.5,120,0,0,1

JC Archer,England,8,2,1,22,21,22,129.41,0,0,0

Arshdeep Singh,India,8,3,2,12,9,12,75,0,0,0

KIC Asalanka,Sri Lanka,3,3,0,71,46,23.66,139.21,0,0,0

PS Athavale,Oman,3,3,0,59,54,19.66,134.09,0,1,1

Ayaan Khan,Oman,4,4,1,93,41*,31,97.89,0,0,0

Azam Khan,Pakistan,1,1,0,0,0,0,0,0,0,1

Azmatullah Omarzai,Afghanistan,8,8,0,111,26,13.87,108.82,0,0,0

Babar Azam,Pakistan,4,4,1,122,44,40.66,101.66,0,0,0

TM Raine,England,8,6,2,110,48*,27,5,134,14,0,0,1

Ln 23, Col 52 100% Windows (CRLF) UTF-8

Player	Team	Mat	Inns	NO	Runs
NP Kenjige	United States of America	4	2	0	1
Aaron Jones	United States of America	6	6	2	162
Aasif Sheikh	Nepal	3	3	0	63
Abbas Afridi	Pakistan	1	1	0	17
F Achelam	Uganda	1	1	0	9
MR Adair	Ireland	3	3	0	52
AC Agar	Australia	2	1	0	2
DS Airee	Nepal	3	3	0	32
MM Ali	England	8	5	0	71

Understanding the dataset

Features

Player	Team	Mat	Inns	NO	Runs	HS	Ave	SR	100	50	0	
NP Kenjige	United States of America	4	2	0	1	1	0.5	25	0	0	1	
Aaron Jones	United States of America	6	6	2	162	94*	40.5	135	0	1	1	
Aasif Sheikh	Nepal	3	3	0	63	42	21	88.73	0	0	0	
Abbas Afridi	Pakistan	1	1	0	17	17	17	80.95	0	0	0	
F Achelam	Uganda	1	1	0	9	9	9	69.23	0	0	0	
MR Adair	Ireland	3	3	0	52	34	17.33	115.55	0	0	0	
AC Agar	Australia	2	1	0	2	2	2	40	0	0	0	
DS Airee	Nepal	3	3	0	32	25	10.66	66.66	0	0	0	
MM Ali	England	8	5	0	71	25	14.2	139.21	0	0	0	

Data points

- The dataset contains 12 features. They are “Player”, “Team”, “Mat”, “Inns”, “NO”, “Runs”, “HS”, “Ave”, “SR”, “100”, “50”, “0”
- There 247 data points i.e the details of 247 players (In the above image, the first 9 are shown)



Working with fewer features

For our question, let us say we are looking for players who can:

- ❖ Score runs quickly
- ❖ Scores runs consistently

Among the 12 features that we have, our requirements are captured by,

- ❖ SR (Strike Rate)
- ❖ Ave (Average runs scored)

The average is calculated using the features “Runs”, “Inns” and “NO” as

$$\text{Average} = \frac{\text{Runs}}{(\text{Inns} - \text{NO})}$$

Strike Rate & Average

Strike Rate:

Player A has scored 10 runs in 4 balls. His strike rate is $(10/4) = 2.5$. To represent it in percentage we multiply by 100 and the value is 250%

Player B has scored 20 runs in 20 balls. His strike rate is $(20/20) = 1$. To represent it in percentage we multiply by 100 and the value is 100%

In terms of strike rate Player A's strike rate is more desirable

Average:

Player A has scored 300 runs in 10 innings and is out in all of them. His average is calculated as

$$300 / (10 - 0) = 30$$

Player B has scored 450 runs in 10 innings and is not out in one match. His average is calculated as

$$450 / (10 - 1) = 50$$

In terms of average Player B's average is more desirable



A closer look at certain data points

Player	Team	Ave	SR
Ibrahim Zadran	Afghanistan	28.87	107.44
Rahmanullah Gurbaz	Afghanistan	35.12	124.33
JC Buttler	England	42.8	158.51
PD Salt	England	37.6	159.32
V Kohli	India	18.87	112.68
RG Sharma	India	36.71	156.7
Q de Kock	South Africa	27	140.46
RR Hendricks	South Africa	14.12	87.59

For this exercise, we have considered only those players who have played as openers for certain teams

Among these, who are the better players in terms of “SR” and “AVE” ?



Exercises

- Try and look for the “Iris dataset”. How many features(columns) and data points(rows) does the dataset contain?
- Pick a field of interest and search for a related dataset. Is the data structured or unstructured? What are the features present?

Collect your own data!

- What vegetables do you typically purchase when shopping for groceries? Note down details such as quantity and price
- How has the price of various vegetables varied across the month?



Thank You!