Assignment 1: Project Data Mosaic

1. Contributions:

This report is prepared by Group # 40 with the following members and their contributions:

a. Shumaila Javed (24280078): Worked on extracting the data from Kaggle and its descriptive analysis, along with preparing the written documentation.
b. Talat Zubair (24280014): Worked on extracting data from Reddit & YFinance, along with drawing the Data Pipeline Diagram.

2. Overview of the Topic:

Since the Champions Trophy is around the corner, we thought of exploring the trends in ODI Cricket, fans sentiments and financial standings of top cricket sponsoring companies. Hence, we chose ODI cricket because of its global popularity, rich history, and significance in international tournaments like the Cricket World Cup. ODI cricket offers a balanced format between the short T20s and long Test matches, making it ideal for data analysis. Unlike T20, which is often unpredictable and short, and Test cricket, which is lengthy and complex, ODIs provide a perfect mix of stability and excitement, making it more suitable for statistical analysis and AI modeling. From this CSV, we are analyzing match outcomes, player performances (runs, wickets), venue impacts, team strategies, and trends over time to build predictive models and insights. The data from Kaggle provides detailed match statistics, while Reddit discussions offer fan sentiments and real-time reactions. We expect to see match outcomes, player performances, team dynamics, and trends over time, making ODI cricket an ideal choice for our analysis. Moreover, we are also looking into the closing stock prices for the top companies sponsoring cricket. It could be an interesting analysis to see how their standings fluctuate during the ongoing tournament.

3. Data Collection Processes:

a. Data from Reddit:

i. Installed PRAW API
ii. Set up a Reddit App account (https://www.reddit.com/prefs/apps) to get clientid, user_agent and secret key
iii. Connected to the Reddit App using ids and keys
iv. Fetched top posts that mentioned 'Cricket'
v. Extracted title, author, upvotes, post content, date & name for each post
vi. Exported the data into a CSV file.

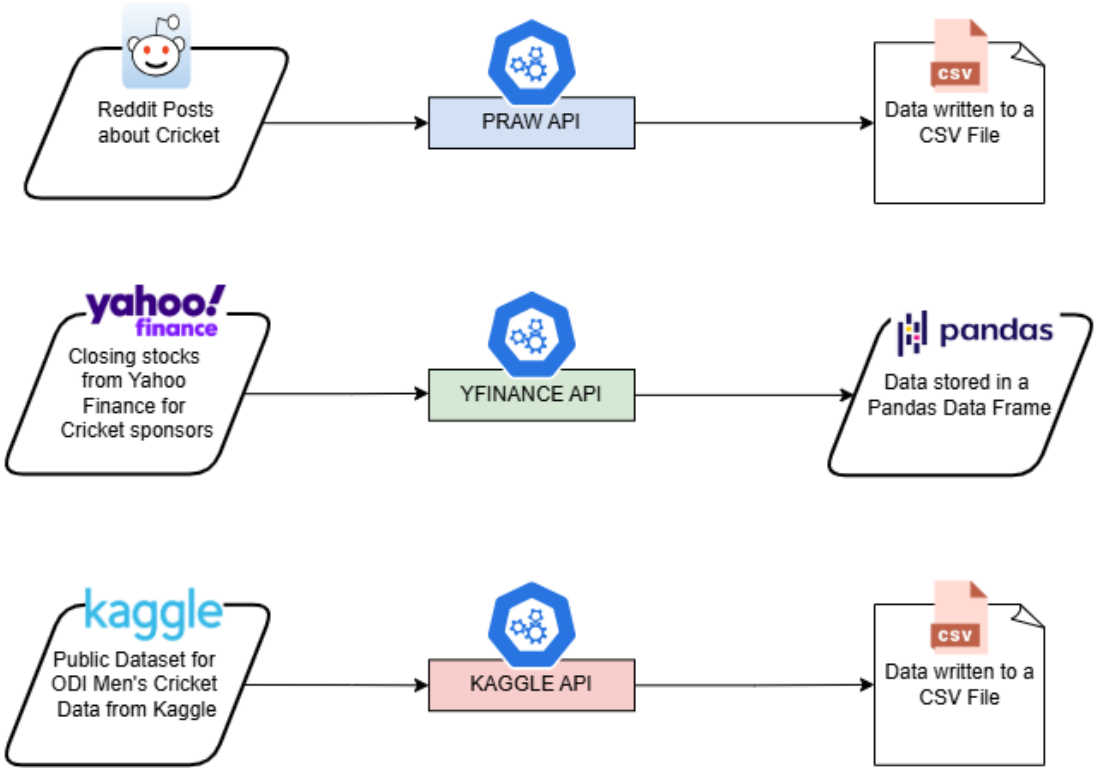
b. Data from Yahoo Finance:

i. Installed yfinance API
ii. Used ETF names for the biggest companies sponsoring cricket, including Pepsi, Coca Cola and Aramco
iii. Fetched 2-year historical stocks data for these companies and combined in a single Data Frame
iv. Removed unrequired columns to only keep ETF, Date and Closing Price.

c. Data from Kaggle:

i. Installed the Kaggle library using!pip install kaggle.
ii. Created a hidden .kaggle directory and uploaded the kaggle.json API key to authenticate.
iii. Downloaded the ODI Men's Cricket Match Data (2002-2023) from Kaggle with!kaggle datasets download -d utkarshtomar736/odi-mens-cricket-match-data-2002-2023.
iv. Extracted the dataset and loaded ODI_Match_info.csv using Pandas with pd.read_csv().
v. Removed the unrelated columns and removed rows containing null values.

4. Data Pipeline



5. Challenges and Limitations:

a. When fetching data from PRAW Api, if the keyword is too specific, something that does not have any top posts, then the API throws a 404 Error

```
NotFound: received 404 HTTP response
```

Hypothesis

The error `NotFound: received 404 HTTP response` indicates that the Reddit API could not find the subreddit you specified, which is "ODI cricket". This could be due to a few reasons:

1. **Incorrect Subreddit Name:** The subreddit name might be misspelled or does not exist.
2. **Reddit API Issues:** There might be temporary issues with the Reddit API, causing it to be unavailable.
3. **Rate Limiting:** If you are making too many requests to the Reddit API in a short period, you might be temporarily rate-limited.

b. Obtaining the ETFs for the specific companies that sponsor cricket was challenging, as Yahoo Finance only lists companies with international stocks.

c. Challenges faced included managing large datasets from Kaggle, and ensuring that data was clean and ready for analysis.

6. Initial Observations:

a. Reddit Data:

```
for post in subreddit.hot(limit=10):
    print(post.title)
```

```
WARNING:praw:It appears that you are using PRAW in an asynchronous environment.
It is strongly recommended to use Async PRAW: https://asyncpraw.readthedocs.io.
See https://praw.readthedocs.io/en/latest/getting_started/multiple_instances.html#discord-bots-and-asy

Daily General Discussion and Match Links Thread - February 15, 2025
Saturday Sledge Thread
Harshit Rana Over Mohammed Siraj Is A Selection Fraught With Risk For Minimal Reward
Can you cheat with your cricket bat at the highest level?
Pakistan Shaheens beat Afghanistan by 144 runs in Champions Trophy warm up game
Channel your inner Joel Wilson with lbwtest.com
Champions Trophy 2025 Prize
Sri Lanka register their first ODI series win in 2025. Also their biggest versus Austrailia in ODI's.
Match Thread: Gujarat Giants Women vs Royal Challengers Bengaluru Women
Highest averages in winning causes in ODIs
```

b. Yahoo Finance Data:

```
Stock_Close = historical_data[['ETF', 'Close']].copy()
Stock_Close.sort_values(by='Date', ascending=True, inplace=True)
Stock_Close.head(10)
```

Date	ETF	Close
2023-02-13 00:00:00+03:00	2223.SR	88.500923
2023-02-14 00:00:00+03:00	2223.SR	87.628128
2023-02-15 00:00:00+03:00	CCOLA.IS	190.426880
2023-02-15 00:00:00+03:00	2223.SR	90.770172
2023-02-15 00:00:00-05:00	PEP	165.789169
2023-02-16 00:00:00+03:00	CCOLA.IS	184.139206
2023-02-16 00:00:00+03:00	2223.SR	93.039421
2023-02-16 00:00:00-05:00	PEP	165.318207
2023-02-17 00:00:00+03:00	CCOLA.IS	182.043304
2023-02-17 00:00:00-05:00	PEP	166.043488

ETF	2223.SR	CCOLA.IS	PEP
count	499.000000	503.000000	502.000000
mean	126.067255	348.378266	166.603770
std	16.052691	251.170949	8.617612
min	87.628128	45.119999	142.639999
25%	114.700001	67.049999	161.439529
50%	128.339569	312.916473	166.572884
75%	133.844589	539.760895	173.066452
max	166.463974	897.500000	185.979706

c. Kaggle Data

	id	season	city	date	team1	team2	toss_winner	toss_decision	result	dl_applied	winner	win_by_runs	win_by_wickets
0	1389389	2023/24	Indore	2023/09/24	India	Australia	Australia	field	D/L	1	India	99	0
1	1336129	2023	Nottingham	2023/09/23	England	Ireland	Ireland	field	normal	0	England	48	0
2	1395701	2023	Dhaka	2023/09/23	New Zealand	Bangladesh	New Zealand	bat	normal	0	New Zealand	86	0
3	1389388	2023/24	Chandigarh	2023/09/22	Australia	India	India	field	normal	0	India	0	5
4	1395700	2023	Dhaka	2023/09/21	New Zealand	Bangladesh	Bangladesh	field	normal	0	NaN	0	0

win_by_wickets	player_of_match	venue	umpire1	umpire2	umpire3
0	SS Iyer	Holkar Cricket Stadium, Indore	J Madanagopal	HDPK Dharmasena	KN Ananthapadmanabhan
0	WG Jacks	Trent Bridge, Nottingham	DJ Millns	RJ Tucker	PR Reiffel
0	IS Sodhi	Shere Bangla National Stadium, Mirpur	M Erasmus	Sharfuddoula	Nitin Menon
5	Mohammed Shami	Punjab Cricket Association IS Bindra Stadium, ...	KN Ananthapadmanabhan	HDPK Dharmasena	J Madanagopal
0	NaN	Shere Bangla National Stadium, Mirpur	Nitin Menon	Sharfuddoula	M Erasmus

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2379 entries, 0 to 2378
Data columns (total 18 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   id               2379 non-null   int64
 1   season           2379 non-null   object
 2   city             2069 non-null   object
 3   date             2379 non-null   object
 4   team1            2379 non-null   object
 5   team2            2379 non-null   object
 6   toss_winner      2379 non-null   object
 7   toss_decision    2379 non-null   object
 8   result           2379 non-null   object
 9   dl_applied       2379 non-null   int64
 10  winner           2259 non-null   object
 11  win_by_runs      2379 non-null   int64
 12  win_by_wickets   2379 non-null   int64
 13  player_of_match  2228 non-null   object
 14  venue            2379 non-null   object
 15  umpire1          2379 non-null   object
 16  umpire2          2379 non-null   object
 17  umpire3          2097 non-null   object
dtypes: int64(4), object(14)
memory usage: 334.7+ KB
```

```
df.describe()
```

	id	dl_applied	win_by_runs	win_by_wickets
count	2.379000e+03	2379.000000	2379.000000	2379.000000
mean	7.114354e+05	0.084489	34.680538	2.750736
std	4.287345e+05	0.278179	53.989592	3.238695
min	6.481400e+04	0.000000	0.000000	0.000000
25%	3.353495e+05	0.000000	0.000000	0.000000
50%	6.490950e+05	0.000000	0.000000	0.000000
75%	1.144488e+06	0.000000	58.000000	6.000000
max	1.395701e+06	1.000000	317.000000	10.000000

7. AI Product:

We plan to build a Match Outcome Predictor using match data, Yahoo Finance market trends, and fan sentiments from Reddit. Additionally, we aim to create a Player Performance Analyzer, a Fan Sentiment Dashboard, and a Betting Support App that provides betting odds and predictions by analyzing historical cricket data, real-time financial market trends (e.g., sponsorship impacts), and live fan sentiments. The app can suggest betting odds, highlight high-performing players, and provide market-based predictions.

8. Data Quality & Discrepancies:

Since the data is coming from multiple sources, having varying structure and format, it could be a challenge to combine or build a relationship between the 3 sources. There are no primary or foreign keys that could help us connect the different sources, except for the Date field, but that format is also not consistent and would have to be cleaned. Moreover, the 3 sources have the different timelines, while the stocks data updates in real-time and the reddit threads may not come in as rapidly or at regular intervals. Moreover, the data from Kaggle is static and won't update. This mismatch in timelines will hinder any analysis that can be performed. In addition to this, the stock data has pricing numbers in different currencies and would need to be standardized and converted into a single currency before any further analysis.

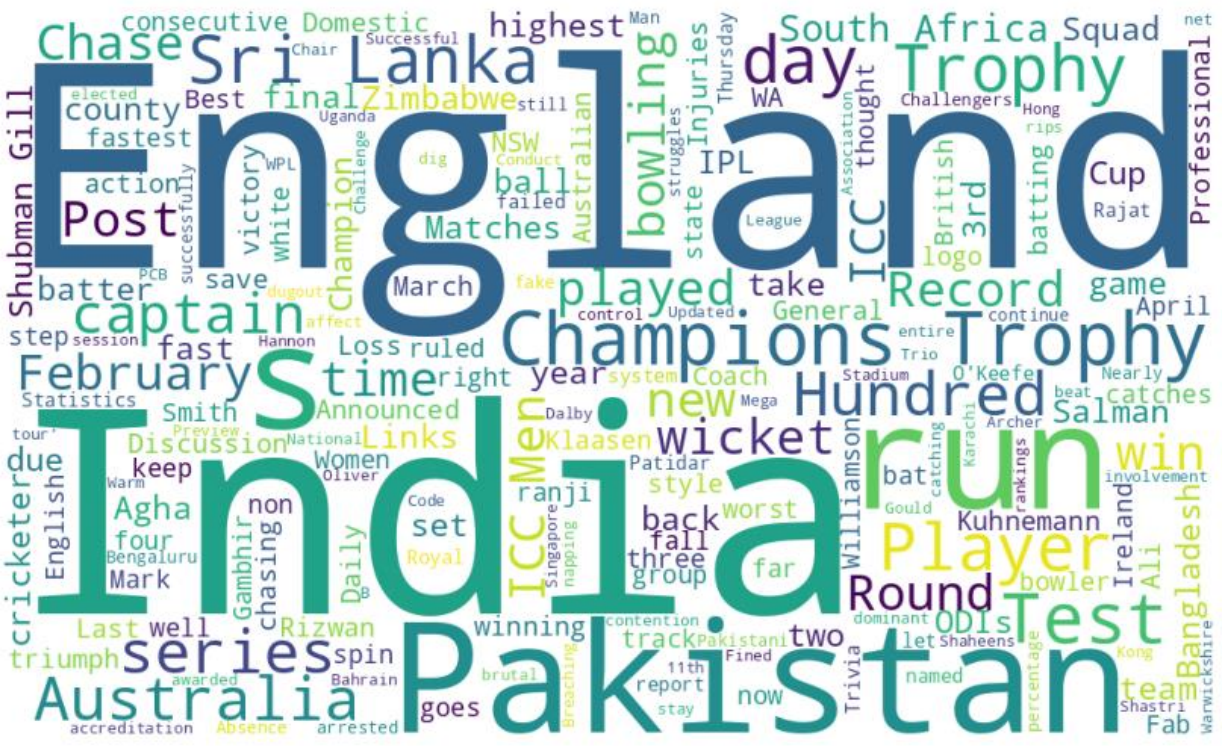
However, on the other hand, having a variety of data offers a comprehensive view of the data and opens the floor for more problem statements and options for analysis. We can look at the data from different perspectives and find more diverse trends and correlations rather than being restricted to a single source.

9. Data Storing & Consolidation:

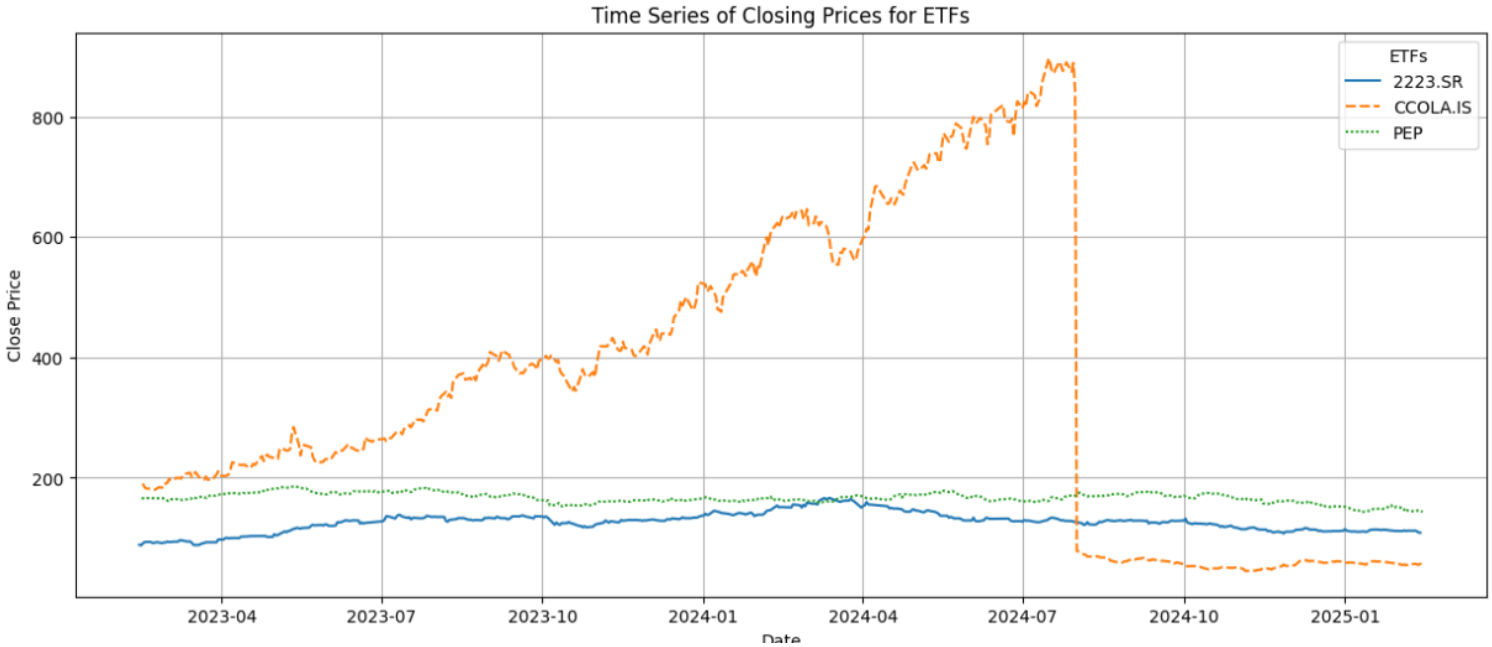
Since all of this data is in structured and tabular format, it can be stored in a relational SQL database as 3 separate tables. We can connect the 3 tables with dates and perform a time series analysis to understand the underlying trends or systemic patterns over time. This would help us explore time-driven correlation or causation between the different datasets. For example, we can study if the increase in reddit discussion is any way related to the increased stocks of the sponsoring companies. Similarly, we can deep dive into which player or team is being talked more about vs. their performance in the 2022-2023 ODI World cup data.

10. Data Visualization:

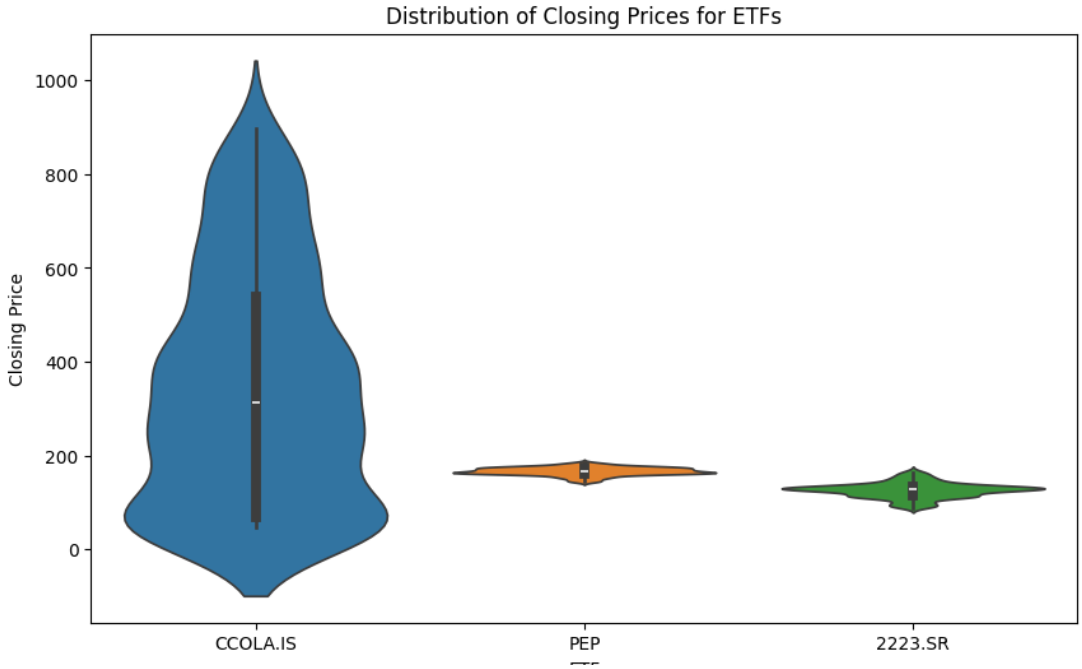
a. Word Cloud from Reddit Posts shows the most common words from Cricket posts



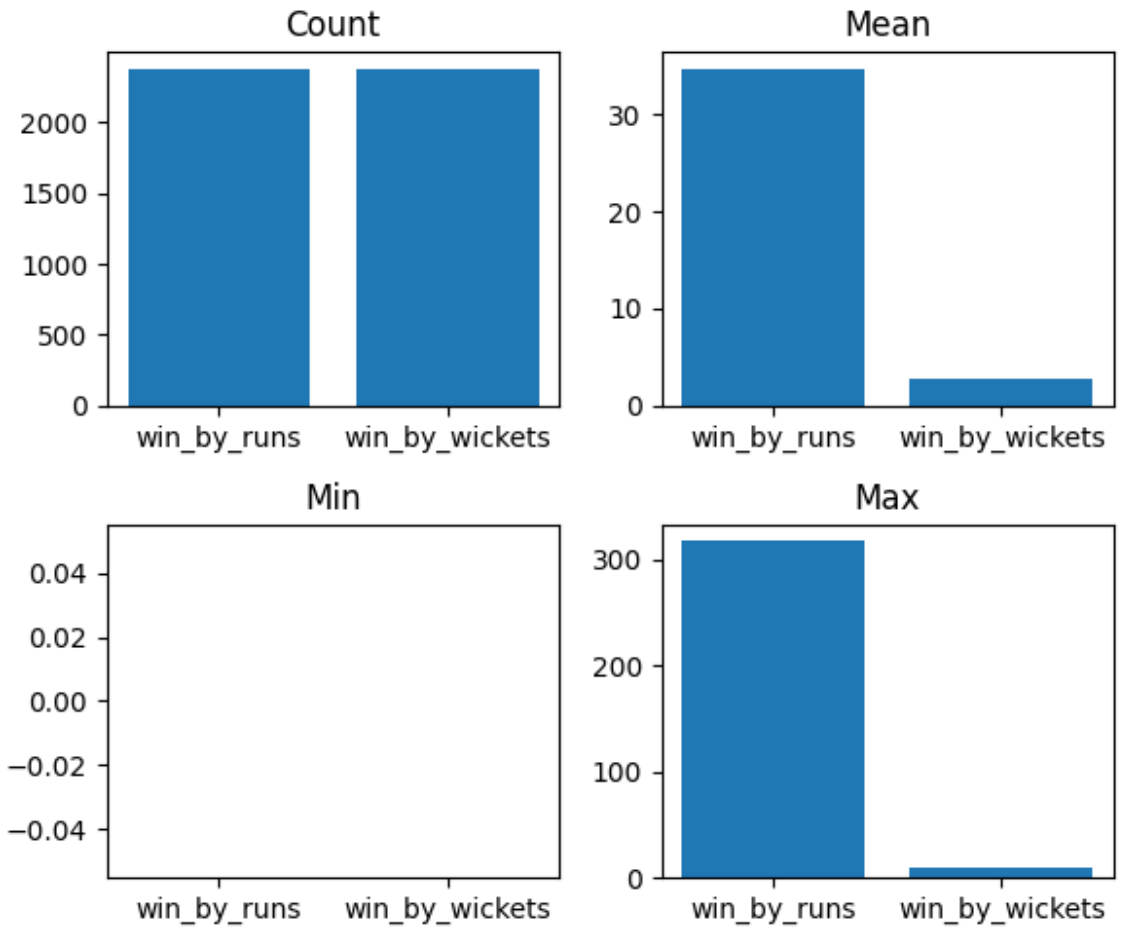
b. Time Series Comparison of top 3 cricket sponsors and their Closing stocks



c. Violin Plots showing the distribution of Closing Price for the 3 companies



d. Bar charts comparing the number of wins by runs and wickets



How this bar chart helps in AI or Statistical Models

- **Model Training**: Knowing the average and spread of runs/wickets helps in feature scaling and selection for machine learning models.
- **Anomaly Detection**: Identifies unusual matches (e.g., a 317-run win) which could be outliers.
- **Performance Benchmarks**: Sets a baseline for evaluating team performances over time.
- **Trend Analysis**: Helps in understanding whether modern ODIs are more competitive or more one-sided compared to older matches.
- **Betting Models**: Bookmakers can use these statistics to set odds, as margins of victory impact betting decisions.

Github Link:

https://github.com/ShumailaJaved/Assignment1-Group-ID-40.git