# Topic:

**"Analyzing the Impact of IPL Sponsorships on Stock Prices: A Multi-Source Data Approach Integrating Sentiment, Engagement, and Player Performance"**


**Group Number: 31**

**Group Members:**

**Irum Hassan 24280043**

**Rabia Aslam 24280019**


**Github Repository Link:**

https://github.com/irumhassan56/Group31

## 1. Group Contribution

The assignment was started by brainstorming with both group members about the data collection process. Both of us presented the hypothesis that could help us in deciding what research questions we want to ask and what features can help us answer those research questions. We discussed different scenarios and finally came to the conclusion about gathering the data from different resources. We decided to get the data from Reddit, Google Trends, Kaggle, and Yahoo Finance. We chose sports as our main topic to start with. Then, we moved to cricket specifically, and for that, we chose the IPL dataset for the years 2008-2024. We were interested in analyzing the spike or decline in the stock price by the IPL sponsors based on public sentiment and player performance. Fan engagements shown by the Reddit likes, comments and upvotes can also determine what kind of futuristic approach an investor is going to adopt.

## 2. Overview of the Topic

The project focuses on IPL sponsorships and their relationship with stock price fluctuations, using a "Multi-Source Data Approach" to analyze data from various sources. Key analytical angles include sentiment, engagement metrics, and player performance. The project distinguishes itself from simple stock trend analysis by incorporating these aspects. The impact-oriented approach provides actionable insights for investors and sponsors, making the title relevant for both academic and practical purposes.

### What do we expect to see?

The IPL sponsors' historical stock prices, fan sentiment and engagement metrics, match data, regional interest, and sponsorship trends are all crucial in understanding the popularity and impact of the tournament. The historical context (2008-2024) provides insights into patterns of price spikes or declines correlated with IPL events, while external factors influence these trends. Reddit data offers sentiment analysis of comments and posts related to IPL, while player performance metrics and match outcomes can influence fan sentiment. Google Trends provides regional interest

trends and a correlation between regional searches and sponsor popularity or stock prices. The historical context also helps identify consistent patterns or anomalies over time.

## 3. Data Collection Process

### 3.1 reddit_IPL_comments.csv

- **Data Collection:** The dataset from Reddit was collected using the PRAW API to scrape the posts, comments, likes etc. The dataset was filtered using the keywords like "IPL," "sponsorship," "teams," and "players" and extracted engagement metrics (upvotes, comments count, etc.).
- **Missing Values**: Some columns have missing data, particularly in comment text and user engagement metrics.
- **Noise**: High level of noise due to unstructured text, irrelevant or off-topic comments, and spam.
- **Limitations**: Sentiment analysis may be skewed due to sarcasm and informal language.

### 3.2 Reddit_IPL_discussions.csv

- **Missing Values**: Certain discussion threads are incomplete or have missing metadata (timestamps, user info).
- **Noise**: Duplicate or highly similar posts.
- **Limitations**: May not fully capture sentiment due to lack of context in individual comments.

### 3.3 reddit_posts.csv

- **Missing Values**: Some fields like upvotes, comments count, or flair is missing.
- **Noise**: Posts with low engagement may not be useful for trend analysis.
- **Limitations**: Bias toward popular posts; newer or niche discussions are underrepresented.

### 3.4 deliveries.csv

- **Data Collection:** We downloaded IPL datasets (2008-2024) containing player performance and match-level details and then merged and cleaned the data, addressing missing values and duplicate records.
- **Missing Values**: Few missing values in runs, extras, and bowler performance data.
- **Noise**: Duplicate records in cases of replays or corrections.
- **Limitations**: Doesn't account for weather conditions, umpiring errors, or player injuries.

### 3.5 matches.csv

- **Missing Values**: Some matches missing venue details or toss decision.
- **Noise**: Possible errors in older match data.
- **Limitations**: Limited to match outcomes; no granular player performance breakdown.

### 3.6 ipl_sponsors_stock_data.csv

- **Data Collection:** We collected stock data for IPL sponsors using the yfinance Python library and filtered data for IPL season timeframes (2008-2024) to analyze patterns.
- **Missing Values**: Some sponsors' stock data missing for certain days.
- **Noise**: External factors (global economy, unrelated events) may affect stock price trends.
- **Limitations**: Correlation between IPL and stock prices may not be causal.

### 3.7 ipl_interest_by_region.csv

- **Data Collection:** We accessed Google Trends data using the pytrends library and queried terms like "IPL," "IPL sponsors," and specific sponsor names by region and time range (2008–2024).
- **Missing Values**: Some regions have incomplete data.
- **Noise**: Regional differences in search behavior may cause inconsistencies.
- **Limitations**: Data may not capture offline interest in IPL.

### 3.8 ipl_sponsors_trends.csv

- **Missing Values**: Gaps in trends for certain sponsors.
- **Noise**: Some fluctuations may be due to unrelated events.

- **Limitations**: May not indicate direct impact of IPL sponsorship.

### 3.9 Sports_IPL_trends_2008_2025.csv

- **Missing Values**: Some years have fewer data points.
- **Noise**: Trends may be affected by global sporting events.
- **Limitations**: Doesn't distinguish between casual and dedicated IPL followers.

## 4. Initial Observation

We calculated the basic summary statistics in order to understand the distribution of our data and to get the insights from the data. The trend was also seen by visualizing the datasets. We calculated the summary statistics of the datasets to know the distribution.

**Summary of Reddit Posts**

Descriptive Statistics:

| | downs | score | num_comments | ups | upvote_ratio |
|---|---|---|---|---|---|
| count | 70.0 | 70.000000 | 70.000000 | 70.000000 | 70.000000 |
| mean | 0.0 | 170.028571 | 22.971429 | 170.028571 | 0.805857 |
| std | 0.0 | 599.078374 | 75.914706 | 599.078374 | 0.137477 |
| min | 0.0 | 0.000000 | 0.000000 | 0.000000 | 0.400000 |
| 25% | 0.0 | 3.000000 | 0.000000 | 3.000000 | 0.712500 |
| 50% | 0.0 | 10.000000 | 1.000000 | 10.000000 | 0.830000 |
| 75% | 0.0 | 36.250000 | 5.750000 | 36.250000 | 0.907500 |
| max | 0.0 | 3583.000000 | 453.000000 | 3583.000000 | 1.000000 |

**Summary for Reddit IPL Discussion**

```
Descriptive Statistics:
        Downs        Score      Comments          Ups  Upvote_Ratio
count  230.0   230.000000    230.000000   230.000000    230.000000
mean     0.0  1029.139130   1026.873913  1029.139130      0.960783
std      0.0   641.437406   5675.642884   641.437406      0.027155
min      0.0   140.000000     45.000000   140.000000      0.810000
25%      0.0   575.250000    138.000000   575.250000      0.950000
50%      0.0   885.000000    194.500000   885.000000      0.970000
75%      0.0  1307.000000    306.250000  1307.000000      0.980000
max      0.0  4059.000000  75002.000000  4059.000000      0.990000
```

**Summary for Matches**

```
Descriptive Statistics:
                  id  result_margin  target_runs  target_overs
count   1.095000e+03    1095.000000  1095.000000   1095.000000
mean    9.048283e+05      17.098630   165.684932     19.760000
std     3.677402e+05      21.631266    33.381189      1.578989
min     3.359820e+05       1.000000    43.000000      5.000000
25%     5.483315e+05       6.000000   146.000000     20.000000
50%     9.809610e+05       8.000000   166.000000     20.000000
75%     1.254062e+06      19.000000   187.000000     20.000000
max     1.426312e+06     146.000000   288.000000     20.000000
```

**Summary for Deliveries**

```
Descriptive Statistics:
           match_id        inning           over  ...     extra_runs     total_runs      is_wicket
count  2.609200e+05  260920.000000  260920.000000  ...  260920.000000  260920.000000  260920.000000
mean   9.070665e+05       1.483531       9.197677  ...       0.067806       1.332807       0.049632
std    3.679913e+05       0.502643       5.683484  ...       0.343265       1.626416       0.217184
min    3.359820e+05       1.000000       0.000000  ...       0.000000       0.000000       0.000000
25%    5.483340e+05       1.000000       4.000000  ...       0.000000       0.000000       0.000000
50%    9.809670e+05       1.000000       9.000000  ...       0.000000       1.000000       0.000000
75%    1.254066e+06       2.000000      14.000000  ...       0.000000       1.000000       0.000000
max    1.426312e+06       6.000000      19.000000  ...       7.000000       7.000000       1.000000
```

**Summary for IPL Sponser Stock Data**

```
Descriptive Statistics:
                Date Close_TATAMOTORS.NS High_TATAMOTORS.NS  ...           Low_HLF Open_HLF Volume_HLF
count           4404                4404               4404  ...              4404     4404       4404
unique          4403                3940               4189  ...              3753     3741       4010
top     2008-01-01   430.5254821777344              806.0  ...  43.0099983215332     27.5  1971600.0
freq               2                 218                216  ...               129      131        129
```

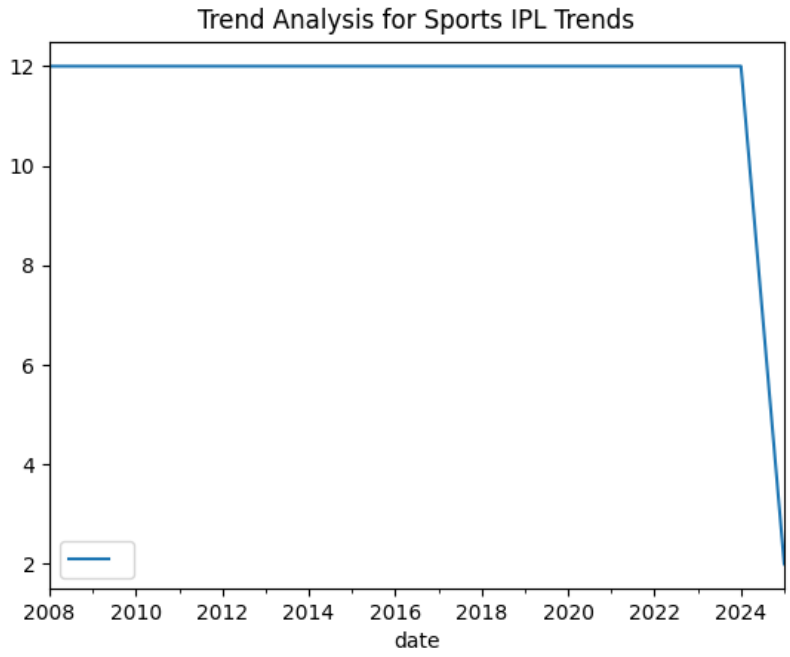
**Summary for IPL Sponsor Trend**

```
Descriptive Statistics:
          Tata IPL    Vivo IPL  Dream11 IPL     Jio IPL   Paytm IPL
count  262.000000  262.000000   262.000000  262.000000  262.000000
mean     6.263359    4.003817     0.877863    1.412214    0.179389
std     18.014982   10.881281     2.214625    3.790228    0.481730
min      0.000000    0.000000     0.000000    0.000000    0.000000
25%      0.000000    0.000000     0.000000    0.000000    0.000000
50%      0.000000    1.000000     0.000000    0.000000    0.000000
75%      2.000000    2.000000     0.000000    0.000000    0.000000
max    100.000000   75.000000    17.000000   18.000000    3.000000
```

**Visualizations**

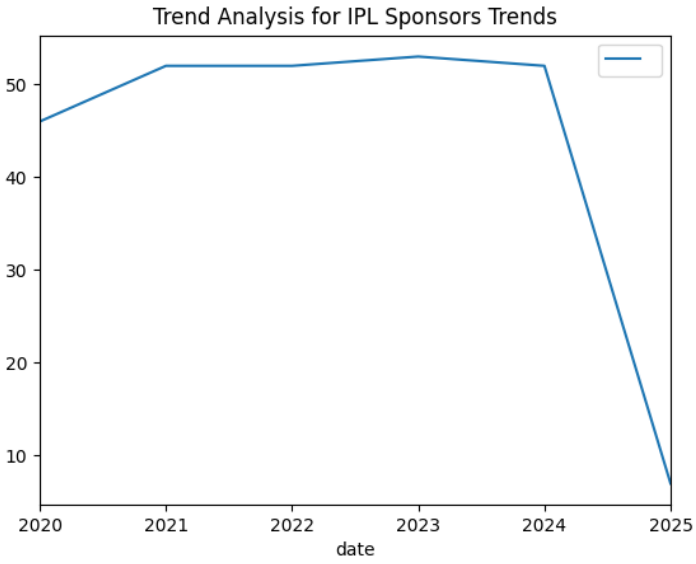
**Trend Analysis for sports IPL trend**

The IPL interest in 2024 has shown a sharp increase, possibly due to missing or incomplete data. However, a drastic drop towards the end of the timeline suggests a significant decline in the metric for the year. Possible explanations include data issues, event-specific factors like reduced fan engagement, sponsorship changes, or economic conditions, or a genuine decline in IPL relevance.



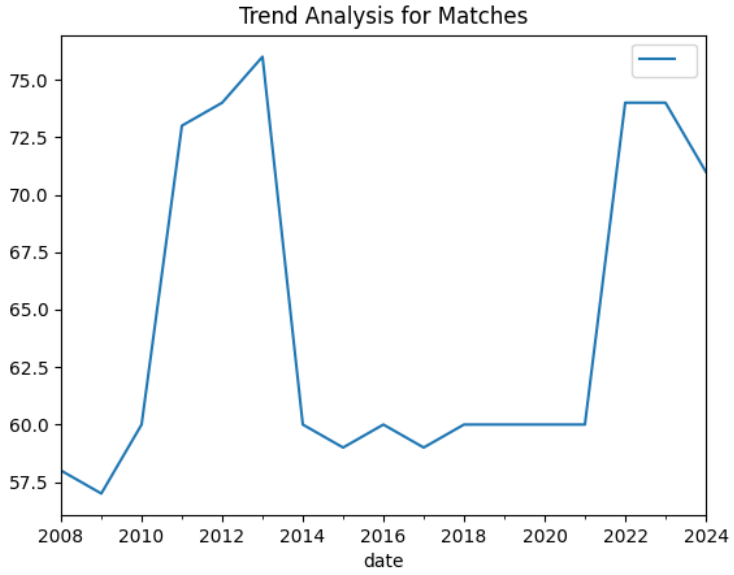

**Trend Analysis for IPL Sponsors Trends**

The sponsorship trends for IPL have shown steady growth from 2020 to 2023, followed by a plateau between 2021 and 2023, suggesting stable activity. However, from 2023 to 2025, there is a drastic decline, with a further drop to a very low level in 2025. Possible explanations include consistent growth due to IPL popularity, stabilization as the market matures, data issues, economic or political factors, sponsorship shifts, or changes in metric definition. These factors could have contributed to the sudden drop in sponsorship trends.




**Trend Analysis for Matches**

The Indian Premier League (IPL) matches' metric has seen a steady growth from 2008 to 2010, with a slight decline in 2008 to 2010. From 2010 to 2013, the metric saw a steep increase, reaching a peak of around 75. From 2013 to 2016, the metric experienced a sharp decline, dropping to around 57.5. From 2016 to 2020, the trend remained stable, with minor fluctuations. From 2020 to 2023, the metric saw a sharp increase, indicating a revival or surge. From 2023 to 2024, the metric may have experienced a slight decline, possibly due to a saturation point or reduced interest.




## 5. AI Application: Sponsor Impact and Stock Movement Predictor

The AI-powered dashboard predicts stock price movements for IPL sponsors based on fan sentiment, player performance, regional interest, and historical sponsor stock trends. The system uses machine learning models to correlate these factors and provide actionable insights to investors, sponsors, and franchise owners. It features a Stock Prediction Module, Fan Sentiment & Engagement Analysis, Sponsor ROI Dashboard, Player Performance Correlation, Anomaly Detection, What-If Scenarios, and AI techniques like Natural Language Processing (NLP) and Time Series Analysis. The system is designed for various stakeholders, including sponsors, advertisers, investors, sports analysts, franchise owners, and researchers. The system can be scalable to other sports leagues, incorporate social media platforms, and add macroeconomic indicators to enhance prediction accuracy.

## 6. Data Collection Constraints

Reddit and Google have specific terms of service (TOS) and privacy concerns when collecting data from their platforms. Reddit's PRAW API usage is governed by their TOS, which includes

non-commercial use, redistribution, rate limits, prohibited activities, user-generated content, PII, retention and access, and sensitive content. Google's data is available for public use but subject to their TOS, with key restrictions including non-commercial use, redistribution, scraping limits, and regional data sensitivity. Data protection laws include GDPR and CCPA, which require explicit consent from users for redistributing comments or posts, and attribution when using publicly available data. Commercial use limits may require explicit agreements with Reddit and Google, as their APIs are often restricted for academic or personal use. These constraints and privacy concerns must be addressed to ensure the ethical and legal compliance of AI applications using Reddit and Google's data.

7. **Data Collection from Multiple sources**

**How does it hinder data quality?**

Collecting data from multiple sources can enhance or challenge data quality depending on how it is managed and integrated. Aggregating data from multiple sources ensures a more comprehensive dataset, reduces the risk of bias, and allows for data validation by cross-checking information between sources. It also enhances context by providing complementary data from different sources, such as Reddit's qualitative insights and Google Trends' quantitative metrics. Using multiple data sources often leads to better predictive models by incorporating a wider range of variables, reducing overfitting and increasing the model's generalizability. However, collecting data from multiple sources can lead to inconsistencies, duplication or overlap, different data definitions, quality variability, and volume imbalance. Inconsistencies can arise from different formats, structures, or levels of granularity, while duplicate data entries can skew analysis. Additionally, different data definitions and quality variability can dilute the overall dataset quality if not filtered. Overall, managing and integrating data from multiple sources is crucial for ensuring accurate and comprehensive data.

**Conflict or Discrepancies in dataset from different resources**

Different sources of data can cause discrepancies in the data narrative. Time lag, contradictory findings, sampling bias, data accessibility and completeness, and metrics misalignment can all contribute to these issues. Google Trends may show real-time search spikes, while Reddit discussions may report negative sentiment or declining interest. Data accessibility and completeness may be limited by sources like Google Trends and Reddit, resulting in incomplete

datasets. Key performance indicators may not align across sources, making comparisons difficult. Cultural or regional differences may also create discrepancies in the data narrative. To manage data from multiple sources, it is essential to standardize data, normalize metrics, resolve conflicts, remove duplicate records, assign weights based on reliability, coverage, and relevance, and cross-validate findings to ensure consistency and identify outliers or anomalies.

## 8. Data Merging and Storage

Storing and combining the data from different resources like google trends, keggle, and Yfinance requires a very scalable and well organized approach. Here are some of the methods that we can use to store the data:

### Relational Database (SQL)

Use a relational database (e.g., PostgreSQL, MySQL) to store structured data from Google Trends and Reddit in tabular formats. For example the search_term, date or volume from the google trend table.

### NoSQL Databases

Use a NoSQL database (e.g., MongoDB, Cassandra) for flexibility in handling structured and unstructured data. For example Document Structure includes documents saved in json file

### Data Lakes (Cloud Storage)

Store raw data from Reddit and Google Trends in a **data lake** (e.g., AWS S3, Google Cloud Storage, or Azure Blob Storage). Like saving the Reddit posts and Google Trends data in separate folders as JSON or CSV files.

## Data Combination Strategies

### I. Unified Data Warehouse

Use a data warehouse (e.g., Snowflake, Google BigQuery, or AWS Redshift) to integrate cleaned data from Reddit and Google Trends into a single storage solution.

### II. ETL Pipelines

Set up an ETL (Extract, Transform, Load) pipeline to collect, clean, and integrate data from multiple sources. For example Use Reddit API/Google Trends API to collect data and Python to preprocess the data.

### III. Time-Based Alignment

Combine Reddit and Google Trends data by aligning timestamps. For example group Reddit posts by day or hour and calculate metrics (e.g., total upvotes, comment counts), or match these aggregates with Google Trends data for the same time periods.

## Conclusion

The dataset integrated from different resources like Google trend, Yahoo finance, or Reddit offers valuable insights but it comes with many challenges when it comes to storing the data and dealing with discrepancies. To overcome these challenges, a robust strategy for data collection, data cleaning, storage, and integration is important. To ensure the seamless storage and and combination of structured and unstructured data by using tools like relational or NoSQL database, data warehouse, or ETL approach. Advanced preprocessing techniques minimize discrepancies, enabling more accurate analysis. Data is stored in scalable solutions like cloud-based data lakes or hybrid systems for flexibility. This methodology enhances data quality, facilitates dynamic analysis, and drives actionable insights.