**AI-601 Assignment-1 report**

**GROUP-29**

| **Student ID** | **Name** |
| --- | --- |
| **25020449** | **Muhammad affan pasha** |
|  |  |
|  |  |
|  |  |
|  |  |

**Table of Contents**

[**1. Introduction 3**](#_1fjwgng0m1xy)

[Why? 3](#_txenc9lxzzdi)

[Expected Data Trends 3](#_3rbppixklv7l)

[**2. Data Collection Process: Challenges & Limitations 4**](#_chipvd5c907e)

[1. Reddit Data Extraction (Unstructured Text Data) 4](#_qkdg96x19z9s)

[2. Yahoo Finance (Structured Financial Data) 4](#_7g4zb4kcnuhy)

[3. Open Data (Semi-Structured Sports Data) 5](#_k7bb5mrdn5kh)

[**4. Initial Observations from Datasets 5**](#_nt17387835gi)

[1. Reddit PRAW 5](#_ltalg4wyfmf7)

[2. Yfinance 6](#_59thq1dzvuwy)

[3. Open Data (pak\_india\_matches.csv) 6](#_q6ovj4k6jxq7)

[**5. AI Product: What Will We Build? 6**](#_fsqvodv93s3g)

[AI-Powered Sports-Finance Analytics Platform 6](#_q69xo3bwgix0)

[Features: 6](#_a4j3v48vaged)

[How it Works? 7](#_ana4uuylmxbv)

[**6. Terms of Service & Privacy Issues 7**](#_lnazslqr8cmz)

[Reddit API Constraints 7](#_s8zzz05t2nmk)

[Yahoo Finance Data Limitations 7](#_s61brlbipmmv)

[Public Data Considerations 7](#_oav500anaxty)

[**7. Multi-Source Data Integration: Benefits & Challenges 8**](#_ud5l1xu92gks)

[Potential Conflicts 8](#_kyfe9owelm7a)

[**8. Storing & Combining Data 8**](#_8nljavlfwwmx)

[Recommended Storage Strategy 8](#_kw628px86wv7)

[**9. \*Optional Charts 9**](#_mcgyakg6un5)

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

## 1. Introduction

### Github Repo

<https://github.com/pashari/AI-601-Assignment1>

We chose this topic to explore **how major cricket events (e.g., India vs. Pakistan matches) impact financial markets and sponsorship trends**.

### Why?

* Cricket is a **multi-billion-dollar industry**, influencing **stocks, sponsorships, and fan engagement**.
* Brands like **Nike, Adidas, and Pepsi** sponsor cricket events, and their stock performance may correlate with match outcomes.
* **Reddit discussions & sentiment trends** can reveal public perception about teams, players, and sponsors.

### Expected Data Trends

1. **Yahoo Finance:**
   * Stock fluctuations of cricket sponsors before & after matches.
   * Investment patterns in sports-related ETFs.
2. **Reddit Discussions:**
   * Sentiment spikes around major cricket events.
   * Keyword trends related to sponsorships, betting, and team performance.
3. **Open Data (Cricket Stats):**
   * Match performance vs. financial impact.
   * Predictive patterns based on historical data.

## 2. Data Collection Process: Challenges & Limitations

### 1. Reddit Data Extraction (Unstructured Text Data)

* **Method:** Used praw API to fetch cricket-related discussions.
* **Challenges:**
  + **API Rate Limits:** Reddit restricts the number of queries per minute.
  + **Data Bias:** Most posts are from **fan-based communities**, which may not reflect true sentiment.
  + **TOS Constraints:** Reddit’s policies **limit the storage & redistribution of user-generated content**.

### 2. Yahoo Finance (Structured Financial Data)

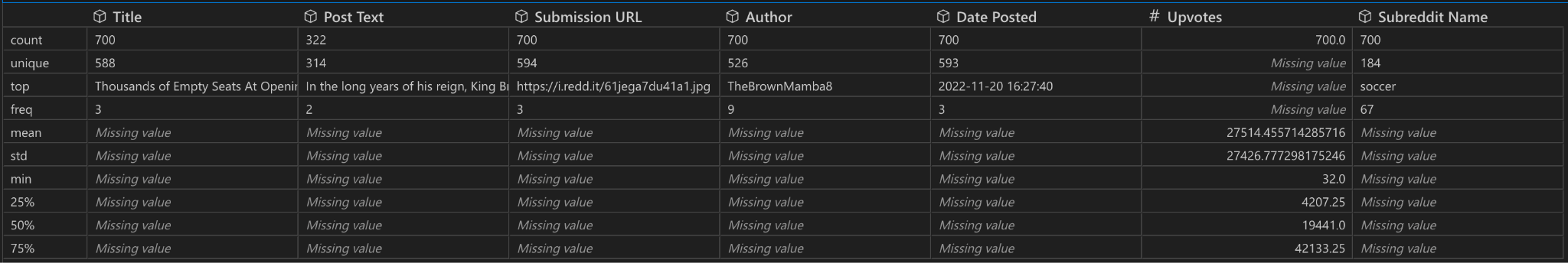
* **Method:** Used yfinance to track stock prices, market trends & sponsorship impact.
* **Challenges:**
  + **Incomplete Data:** Some cricket sponsors are private companies and not publicly traded.
  + **Stock Market Noise:** External factors (e.g., global economic trends) influence stock movement beyond cricket events.
  + **API Limitations:** Free tier of yfinance may not provide real-time data.

### 3. Open Data (Semi-Structured Sports Data)

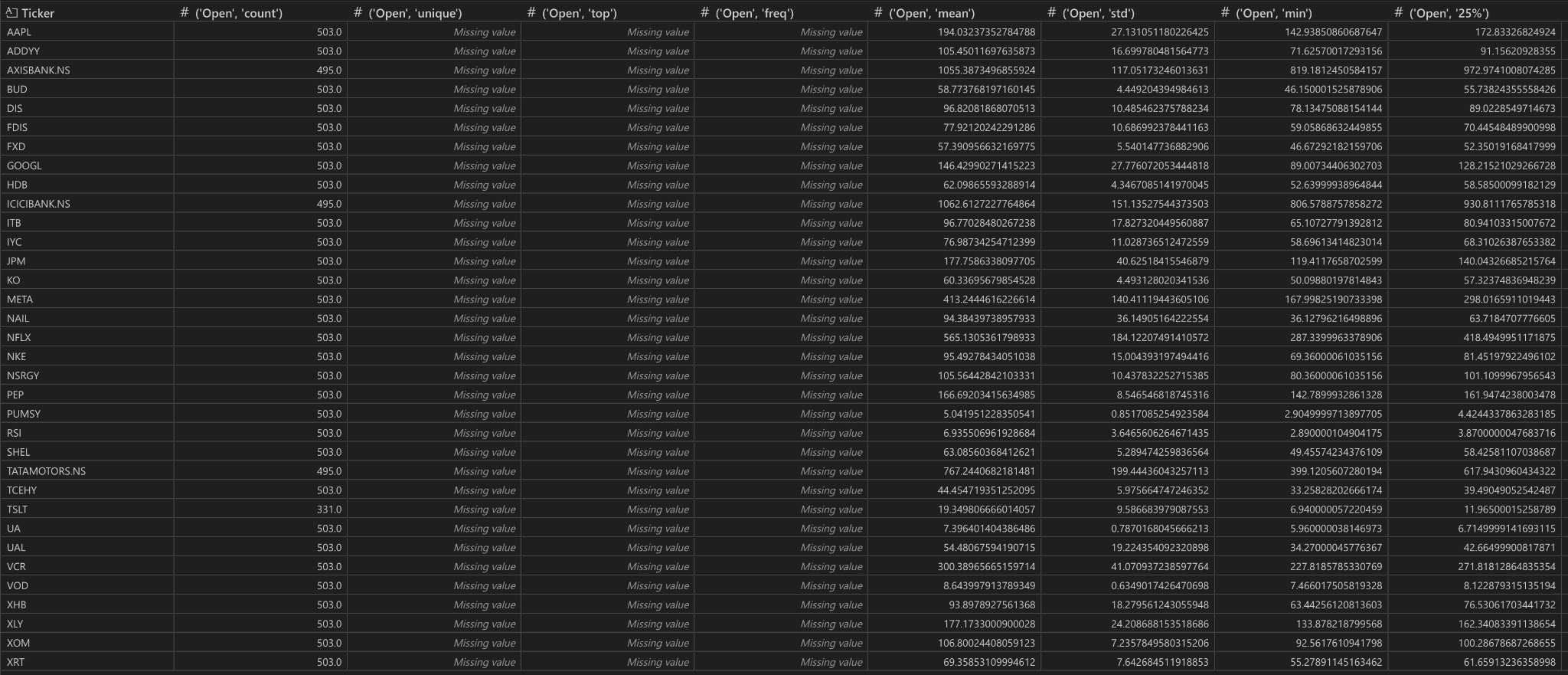
* **Method:** Extracted **match statistics from Kaggle datasets**.
* **Challenges:**
  + **Missing Data:** Older cricket matches may lack detailed statistics.
  + **Inconsistent Formatting:** Different datasets have varied column structures, requiring **data cleaning & transformation**.

## 4. Initial Observations from Datasets

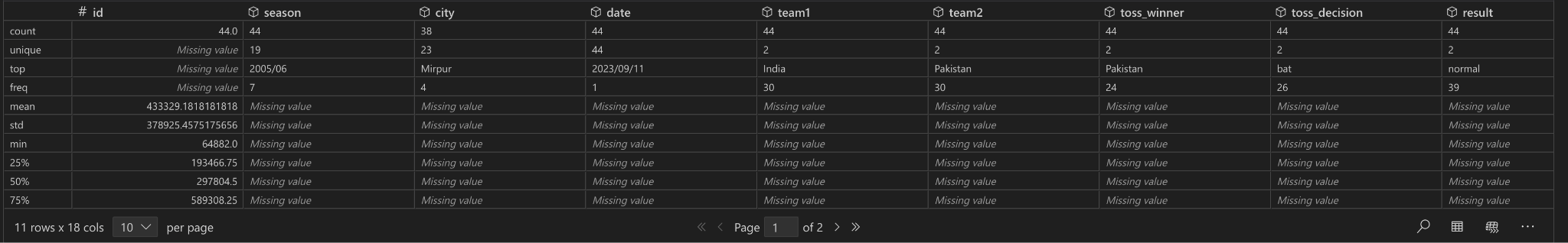
### Reddit PRAW



### Yfinance



### Open Data (pak\_india\_matches.csv)



## 5. AI Product: What Will We Build?

### AI-Powered Sports-Finance Analytics Platform

This AI system **predicts stock market & sponsorship trends** based on cricket events.

### Features:

**Stock Impact Prediction:** How cricket results affect sponsor stocks.  
**Sentiment Trends:** Analyzes **Reddit discussions** for sponsorship value.  
**Historical Data Insights:** Compares past match results with financial movements.

### How it Works?

1. **Reddit Sentiment Analysis →** Captures fan reactions.
2. **Yahoo Finance API →** Tracks sponsorship stock performance.
3. **Match Data (Kaggle) →** Uses past match statistics to find patterns.
4. **AI Model →** Predicts stock & sponsorship impact based on match results.

## 6. Terms of Service & Privacy Issues

### Reddit API Constraints

* **User Privacy:** Cannot **store or redistribute** personal user data.
* **Content Restrictions:** Reddit limits **commercial use** of its data.

### Yahoo Finance Data Limitations

* **Rate Limits:** Free API has **limited requests per hour**.
* **Data Redistribution:** Cannot **resell or publicly share financial data** extracted from Yahoo Finance.

### Public Data Considerations

* **No Privacy Issues:** Kaggle datasets are publicly available.
* **Data Ownership:** ICC may **own copyright over match statistics**.

## 7. Multi-Source Data Integration: Benefits & Challenges

| **Factor** | **Benefits** | **Challenges** |
| --- | --- | --- |
| **Reddit Sentiment** | Real-time fan reactions | Biased user opinions |
| **Yahoo Finance** | Stock performance trends | Market fluctuations affect accuracy |
| **Match Statistics** | Objective game data | Requires additional financial correlation |

**Inconsistent Data Formats:**

* Reddit provides unstructured text, while Yahoo Finance is time-series, and Kaggle is semi-structured match data.

**Temporal Misalignment:**

* Financial data is daily (or minute-level if we pull intraday), but match events and Reddit discussions can have spikes at specific hours. We must carefully aggregate data to a comparable time scale.

**Potential Missing Data / Partial Overlaps:**

* Some sponsor stocks might have incomplete data (especially for non-U.S. tickers).
* Some cricket matches in the Kaggle dataset could lack updated stats or have missing columns.

**Discrepancies in Terminology:**

* Sponsor names vs. ticker symbols vs. brand references in Reddit might not always match perfectly (e.g., “Nike” vs. “NKE”).

## 8. Storing & Combining Data

### Recommended Storage Strategy

**Relational Database (e.g., PostgreSQL):**

* Create tables for reddit\_posts, finance\_data, match\_info, and deliveries.
* Use a shared key (e.g., date or sponsor name mapping) to partially join the data.

**Data Warehouse / Lake (e.g., AWS S3 + Athena):**

* Store CSVs in a data lake. Use external tables to query them collectively.

**Merge in Pandas:**

* For short-term experimentation, load CSVs into separate pandas DataFrames.
* Implement merges on Date or Ticker as needed to build an integrated table for modeling.

## 9. \*Optional Charts

