AI-601 ASSIGNMENT-1 REPORT

GROUP-29

STUDENT ID	NAME
25020449	MUHAMMAD AFFAN PASHA

TABLE OF CONTENTS

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)

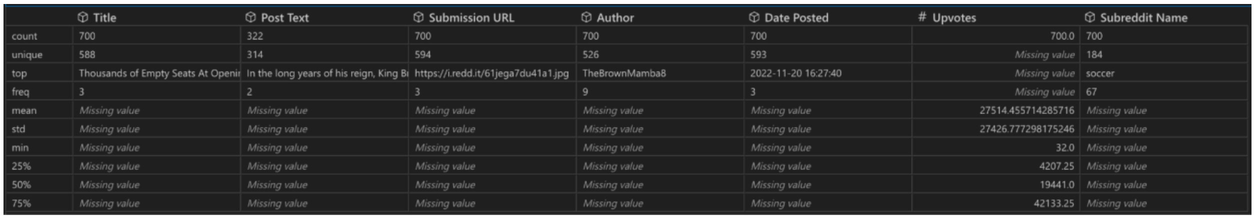
- **Method:** Used `yfinance` to track stock prices, market trends & sponsorship impact.
- **Challenges:**
  - **Incomplete Data:** Some cricket sponsors are private companies and not publicly traded.
  - **Stock Market Noise:** External factors (e.g., global economic trends) influence stock movement 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**.

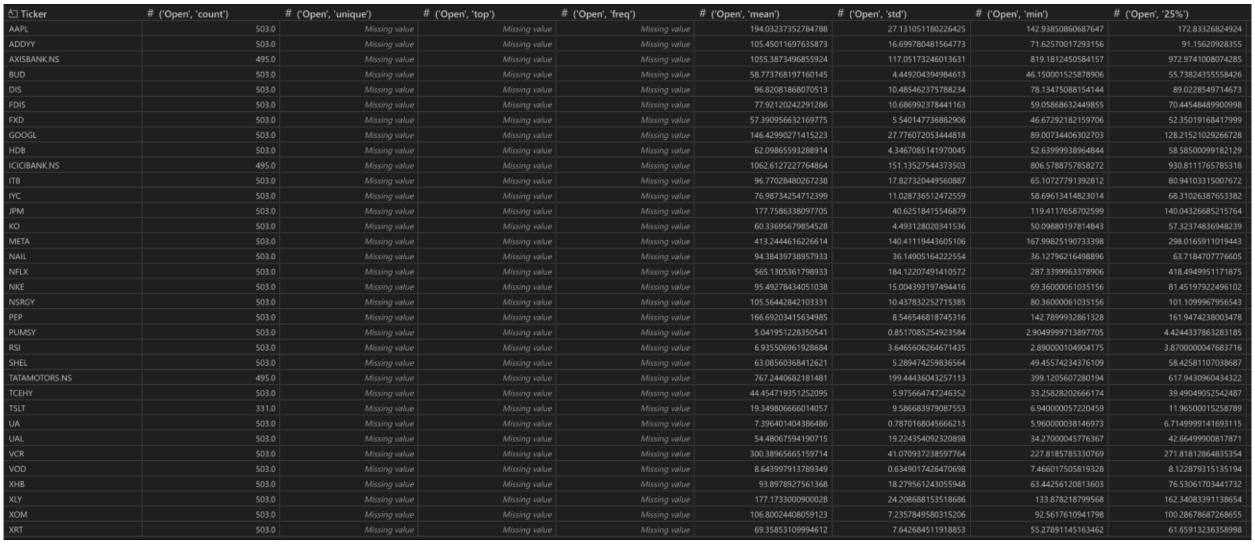
4. Initial Observations from Datasets

1. Reddit PRAW



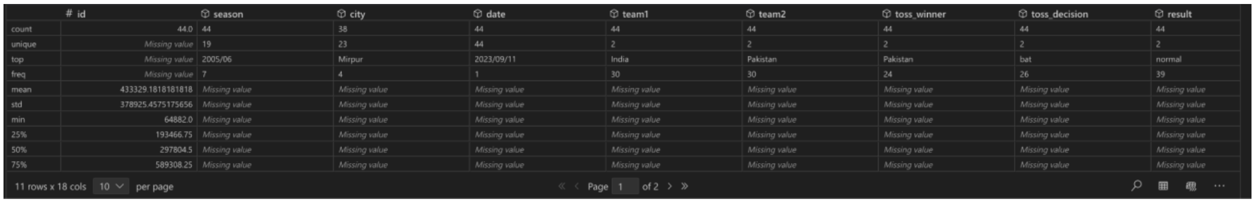
	Title	Post Text	Submission URL	Author	Date Posted	# Upvotes	Subreddit Name
count	700	322	700	700	700	700.0	700
unique	588	314	594	526	593	Missing value	184
top	Thousands of Empty Seats At Openir	In the long years of his reign, King Bı	https://i.redd.it/61jega7du41a1.jpg	TheBrownMamba8	2022-11-20 16:27:40	Missing value	soccer
freq	3	2	3	9	3	Missing value	67
mean	Missing value	Missing value	Missing value	Missing value	Missing value	27514.455714285716	Missing value
std	Missing value	Missing value	Missing value	Missing value	Missing value	27426.777298175246	Missing value
min	Missing value	Missing value	Missing value	Missing value	Missing value	32.0	Missing value
25%	Missing value	Missing value	Missing value	Missing value	Missing value	4207.25	Missing value
50%	Missing value	Missing value	Missing value	Missing value	Missing value	19441.0	Missing value
75%	Missing value	Missing value	Missing value	Missing value	Missing value	42133.25	Missing value

2. Yfinance



Ticker	# ('Open', 'count')	# ('Open', 'unique')	# ('Open', 'top')	# ('Open', 'freq')	# ('Open', 'mean')	# ('Open', 'std')	# ('Open', 'min')	# ('Open', '25%')
AAPL	503.0	Missing value	Missing value	Missing value	194.03237352784788	27.131051180226425	142.93850860687647	172.83326824924
ADDYY	503.0	Missing value	Missing value	Missing value	105.45011697635873	16.699780481564773	71.62570017293156	91.15620928355
AXISBANK.NS	495.0	Missing value	Missing value	Missing value	1055.3873496855924	117.05173246013631	819.1812450584157	972.9741008074285
BUD	503.0	Missing value	Missing value	Missing value	58.773768197160145	4.449204394984613	46.150001525878906	55.73824355558426
DIS	503.0	Missing value	Missing value	Missing value	96.82081868070513	10.485462375788234	78.13475088154144	89.0228549714673
FDIS	503.0	Missing value	Missing value	Missing value	77.92120242291286	10.686992378441163	59.05868632449855	70.44548489900998
FXD	503.0	Missing value	Missing value	Missing value	57.390956632169775	5.540147736882906	46.67292182159706	52.35019168417999
GOOGL	503.0	Missing value	Missing value	Missing value	146.42990271415223	27.776072053444818	89.00734406302703	128.2152102926672
HDB	503.0	Missing value	Missing value	Missing value	62.09865593288914	4.3467085141970045	52.63999938964844	58.58500099182129
ICICIBANK.NS	495.0	Missing value	Missing value	Missing value	1062.6127227764864	151.13527544373503	806.5788757858272	930.8111765785318
ITB	503.0	Missing value	Missing value	Missing value	96.77028480267238	17.827320449560887	65.10727791392812	80.94103315007672
IYC	503.0	Missing value	Missing value	Missing value	76.98734254712399	11.028736512472559	58.69613414823014	68.31026387653382
JPM	503.0	Missing value	Missing value	Missing value	177.7586338097705	40.62518415546879	119.4117658702599	140.04326685215764
KO	503.0	Missing value	Missing value	Missing value	60.33695679854528	4.493128020341536	50.09880197814843	57.32374836948239
META	503.0	Missing value	Missing value	Missing value	413.2444616226614	140.41119443605106	167.99825190733398	298.0165911019443
NAIL	503.0	Missing value	Missing value	Missing value	94.38439738957933	36.14905164222554	36.12796216498896	63.7184707776605
NFLX	503.0	Missing value	Missing value	Missing value	565.1305361798933	184.12207491410572	287.3399963378906	418.4949951171875
NKE	503.0	Missing value	Missing value	Missing value	95.49278434051038	15.004393197494416	69.36000061035156	81.45197922496102
NSRGY	503.0	Missing value	Missing value	Missing value	105.56442842103331	10.437832252715385	80.36000061035156	101.1099967956543
PEP	503.0	Missing value	Missing value	Missing value	166.69203415634985	8.546546818745316	142.7899932861328	161.9474238003478
PUMSY	503.0	Missing value	Missing value	Missing value	5.041951228350541	0.8517085254923584	2.9049999713897705	4.4244337863283185
RSI	503.0	Missing value	Missing value	Missing value	6.935506961928684	3.6465606264671435	2.890000104904175	3.8700000047683716
SHEL	503.0	Missing value	Missing value	Missing value	63.08560368412621	5.289474259836564	49.45574234376109	58.42581107038687
TATAMOTORS.NS	495.0	Missing value	Missing value	Missing value	767.2440682181481	199.44436043257113	399.1205607280194	617.9430960434322
TCEHY	503.0	Missing value	Missing value	Missing value	44.454719351252095	5.975664747246352	33.25828202666174	39.49049052542487
TSLT	331.0	Missing value	Missing value	Missing value	19.349806666014057	9.586683979087553	6.940000057220459	11.96500015258789
UA	503.0	Missing value	Missing value	Missing value	7.396401404386486	0.7870168045666213	5.960000038146973	6.7149999141693115
UAL	503.0	Missing value	Missing value	Missing value	54.48067594190715	19.224354092320898	34.27000045776367	42.66499900817871
VCR	503.0	Missing value	Missing value	Missing value	300.38965665159714	41.070937238597764	227.8185785330769	271.81812864835354
VOD	503.0	Missing value	Missing value	Missing value	8.643997913789349	0.6349017426470698	7.466017505819328	8.122879315135194
XHB	503.0	Missing value	Missing value	Missing value	93.8978927561368	18.279561243055948	63.44256120813603	76.53061703441732
XLY	503.0	Missing value	Missing value	Missing value	177.1733000900028	24.208688153518686	133.878218799568	162.34083391138654
XOM	503.0	Missing value	Missing value	Missing value	106.80024408059123	7.2357849580315206	92.5617610941798	100.28678687268655
XRT	503.0	Missing value	Missing value	Missing value	69.35853109994612	7.642684511918853	55.27891145163462	61.65913236358998

3. **Open Data (pak_india_matches.csv)**



	# id	season	city	date	team1	team2	toss_winner	toss_decision	result
count	44.0	44	38	44	44	44	44	44	44
unique	*Missing value*	19	23	44	2	2	2	2	2
top	*Missing value*	2005/06	Mirpur	2023/09/11	India	Pakistan	Pakistan	bat	normal
freq	*Missing value*	7	4	1	30	30	24	26	39
mean	433329.1818181818	*Missing value*	*Missing value*	*Missing value*	*Missing value*	*Missing value*	*Missing value*	*Missing value*	*Missing value*
std	378925.4575175656	*Missing value*	*Missing value*	*Missing value*	*Missing value*	*Missing value*	*Missing value*	*Missing value*	*Missing value*
min	64882.0	*Missing value*	*Missing value*	*Missing value*	*Missing value*	*Missing value*	*Missing value*	*Missing value*	*Missing value*
25%	193466.75	*Missing value*	*Missing value*	*Missing value*	*Missing value*	*Missing value*	*Missing value*	*Missing value*	*Missing value*
50%	297804.5	*Missing value*	*Missing value*	*Missing value*	*Missing value*	*Missing value*	*Missing value*	*Missing value*	*Missing value*
75%	589308.25	*Missing value*	*Missing value*	*Missing value*	*Missing value*	*Missing value*	*Missing value*	*Missing value*	*Missing value*

11 rows x 18 cols 10 per page Page 1 of 2

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.

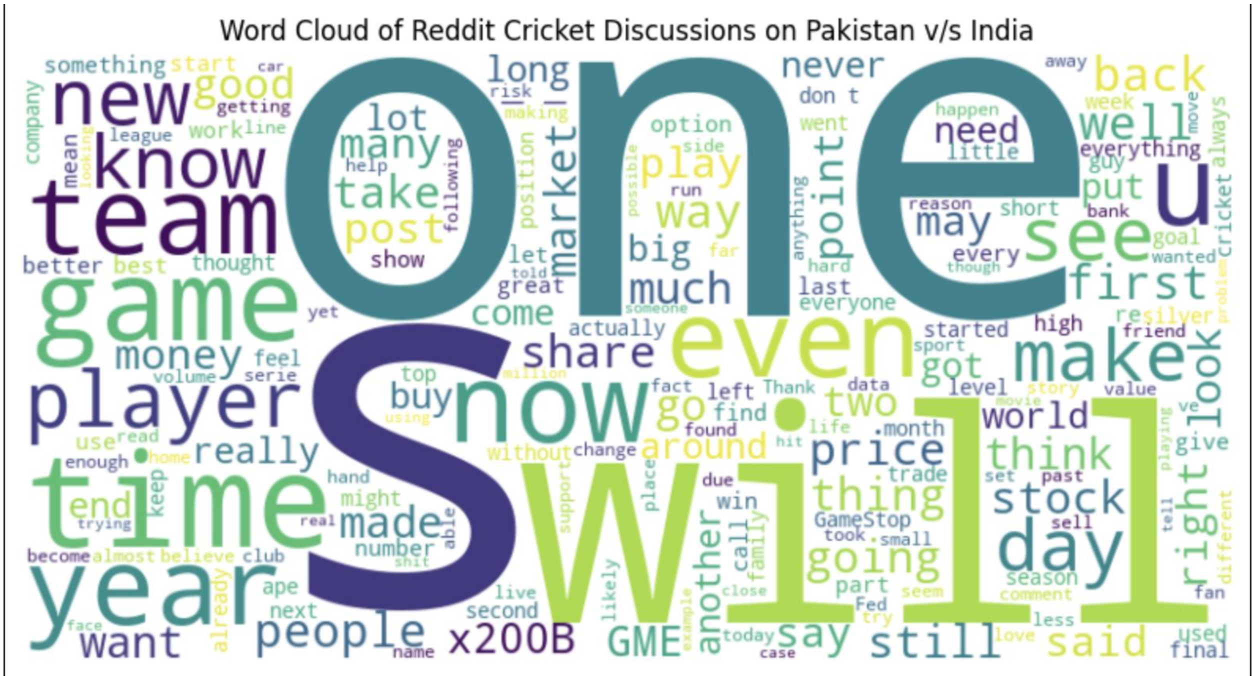
Data Warehouse / Lake (e.g., AWS S3 + Athena):

- 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



Word Cloud of Reddit Cricket Discussions on Pakistan v/s India