

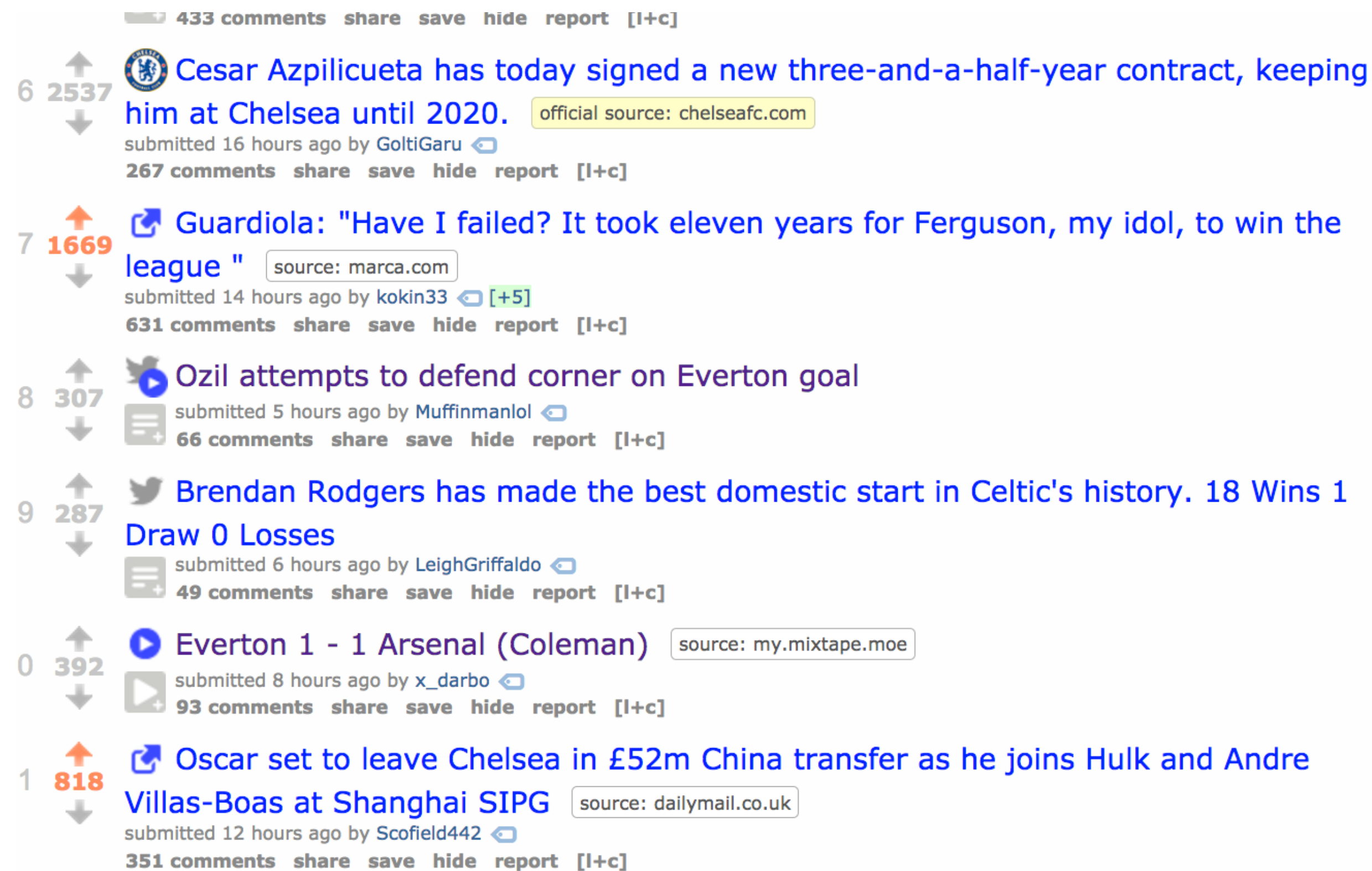
/r/Soccer Activity and User Distribution Analysis

Project Web Scraping
Sharan Naribole
Dec 14, 2016



/r/Soccer Introduction



- **Large community**



- Over 500,000 subscribers globally
- Hundreds of posts discussed daily
- Goals, pre, post and live match discussion, news articles etc.







433 comments share save hide report [I+c]



6 2537  Cesar Azpilicueta has today signed a new three-and-a-half-year contract, keeping him at Chelsea until 2020. [official source: chelseafc.com](#)
submitted 16 hours ago by GoltiGaru 
267 comments share save hide report [I+c]

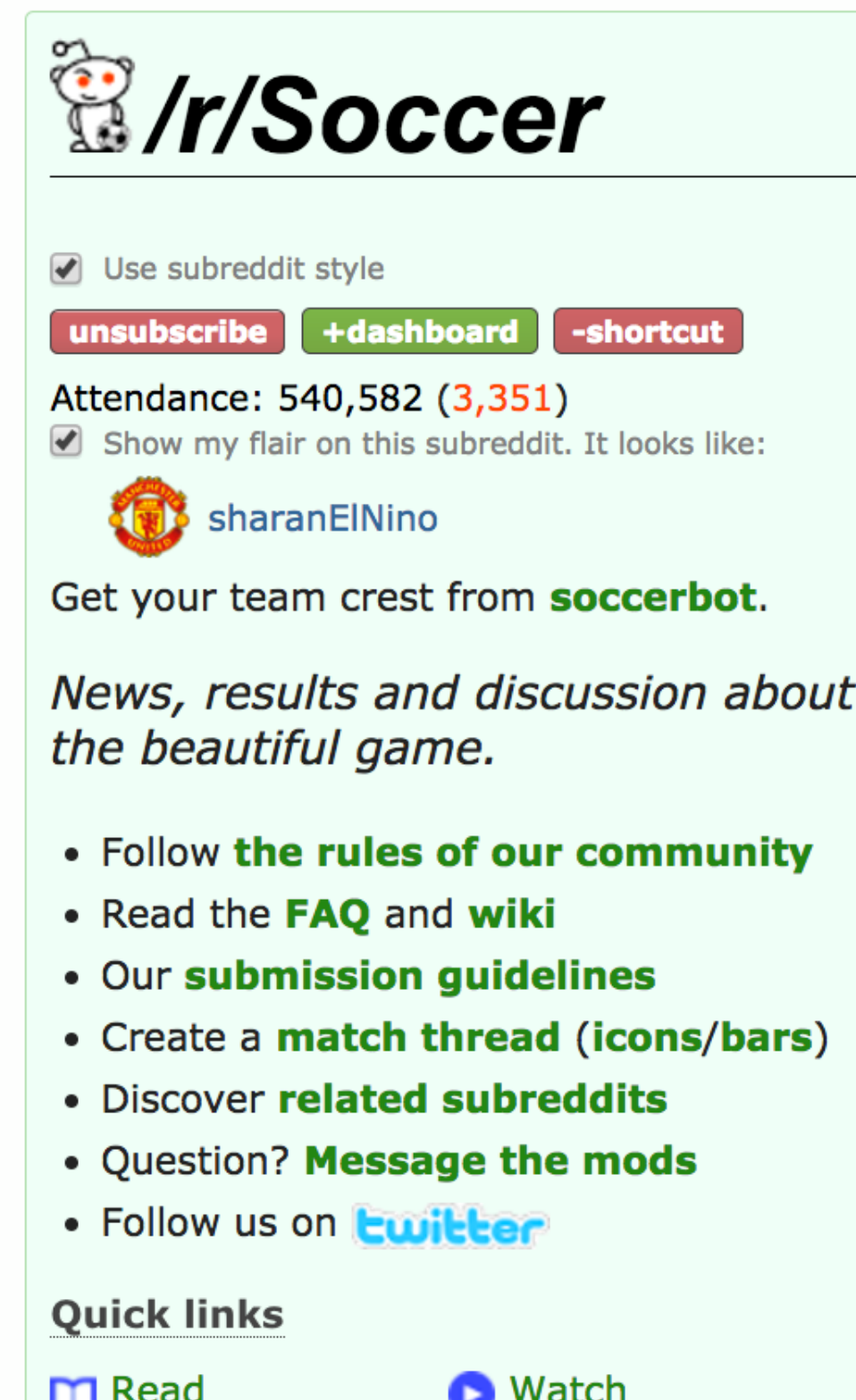
7 1669  Guardiola: "Have I failed? It took eleven years for Ferguson, my idol, to win the league " [source: marca.com](#)
submitted 14 hours ago by kokin33  [+5]
631 comments share save hide report [I+c]


8 307  Ozil attempts to defend corner on Everton goal
submitted 5 hours ago by Muffinmanlol 
66 comments share save hide report [I+c]

9 287  Brendan Rodgers has made the best domestic start in Celtic's history. 18 Wins 1 Draw 0 Losses
submitted 6 hours ago by LeighGriffaldo 
49 comments share save hide report [I+c]

0 392  Everton 1 - 1 Arsenal (Coleman) [source: my.mixtape.moe](#)
submitted 8 hours ago by x_darbo 
93 comments share save hide report [I+c]

1 818  Oscar set to leave Chelsea in £52m China transfer as he joins Hulk and Andre Villas-Boas at Shanghai SIPG [source: dailymail.co.uk](#)
submitted 12 hours ago by Scofield442 
351 comments share save hide report [I+c]




 **/r/Soccer**

☒ Use subreddit style

[unsubscribe](#) [+dashboard](#) [-shortcut](#)

Attendance: 540,582 (3,351)

☒ Show my flair on this subreddit. It looks like:

 sharanElNino

Get your team crest from [soccerbot](#).

News, results and discussion about the beautiful game.

- Follow **the rules of our community**
- Read the **FAQ** and **wiki**
- Our **submission guidelines**
- Create a **match thread (icons/bars)**
- Discover **related subreddits**
- Question? **Message the mods**
- Follow us on [twitter](#)




Quick links




[Read](#) [Watch](#)




/r/Soccer Flairs

- **User Flair**

- Team crest appears beside username on all comments
- Provides additional dimension of context to the comments

 [-]  **Wavey_Don**  **127 points** 7 hours ago
Genuinely furious
[permalink](#) [source](#) [embed](#) [save](#) [save-RES](#) [report](#) [give gold](#) [reply](#) [hide child comments](#)

 [-]  **bboiabb**  **107 points** 7 hours ago
shhh bb is ok
[permalink](#) [source](#) [embed](#) [save](#) [save-RES](#) [parent](#) [report](#) [give gold](#) [reply](#)

 [-]  **KarmannosaurusRex**  **42 points** 7 hours ago
Genuinely opposite of furious
[permalink](#) [source](#) [embed](#) [save](#) [save-RES](#) [parent](#) [report](#) [give gold](#) [reply](#)

Objective

To scrape and analyze /r/soccer top posts for:

- a) user flair distribution and
- b) it's relationship with comments activity, submission score and type

Outline

Data Collection

Data Processing

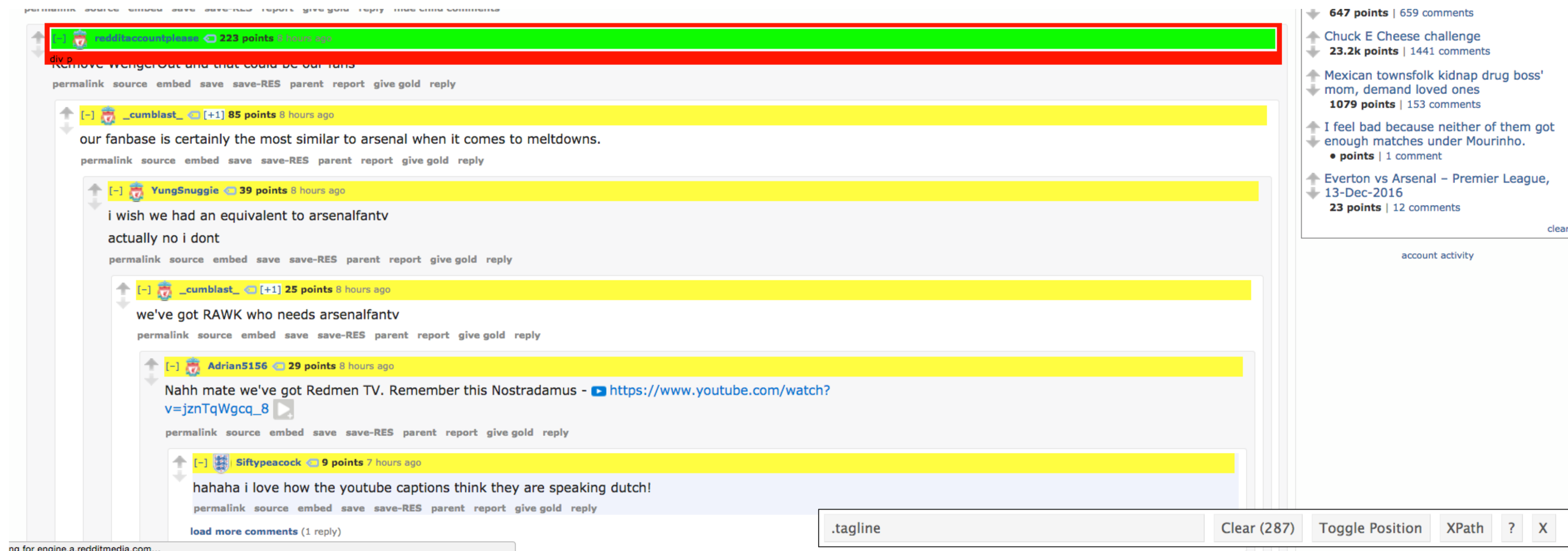
Data Analysis

Data Collection

- **Every post in the top 1000 posts during Nov 12 2016 -Dec 12 2016**
 - **Constraint:** Minimum of 100 comments
 - Matches occur on a weekly basis!
 - Over 500 posts/submissions met the constraints
- **Per-submission Information**
 - Title
 - Submission score (\sim Upvotes - Downvotes)
 - Number of comments
 - Unique user-flair mapping in top 500 comments

Data Collection Tools

- **Scrapy Crawl Spider**
 - /r/soccer home page sorted by top
 - Comments page of each submission
- **SelectorGadget**



Per-Submission Computation

- **Flair Diversity**
 - Unique number of flairs
- **Percentage share per flair**
 - $100 \times \text{Number of commenters of a flair} / \text{Total commenter-flair mappings}$
- **Top percentage share**
 - Highest percentage share among all the flairs
- **Comments**
 - Language processing to find total number of comments

Computation Framework

- **Pandas DataFrames**

	Title	Flair Diversity	Top Share	Comments
Submission 1				
Submission 2				
Submission 3				
...				

	Submission 1	Submission 2	...
Flair 1			
Flair 2			
Flair 3			
...			

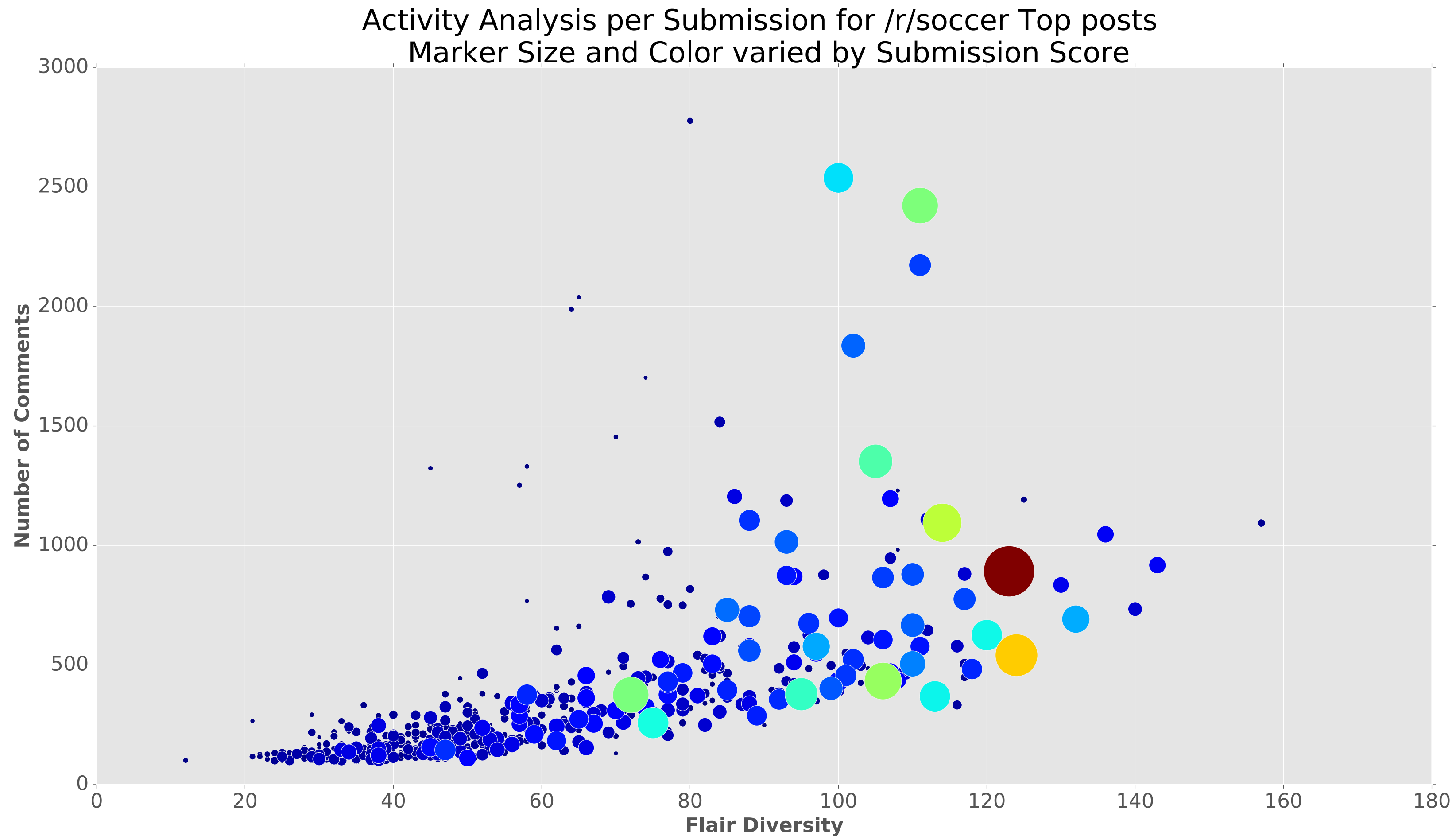
Data Analysis

Diversity - Comments - Score Relationship

Submission Type Analysis

Flair Distribution

Diversity - Comments - Score Relationship



- **Comments increase with Diversity**
- **Submission Score increases with Diversity**

Submission Type Analysis

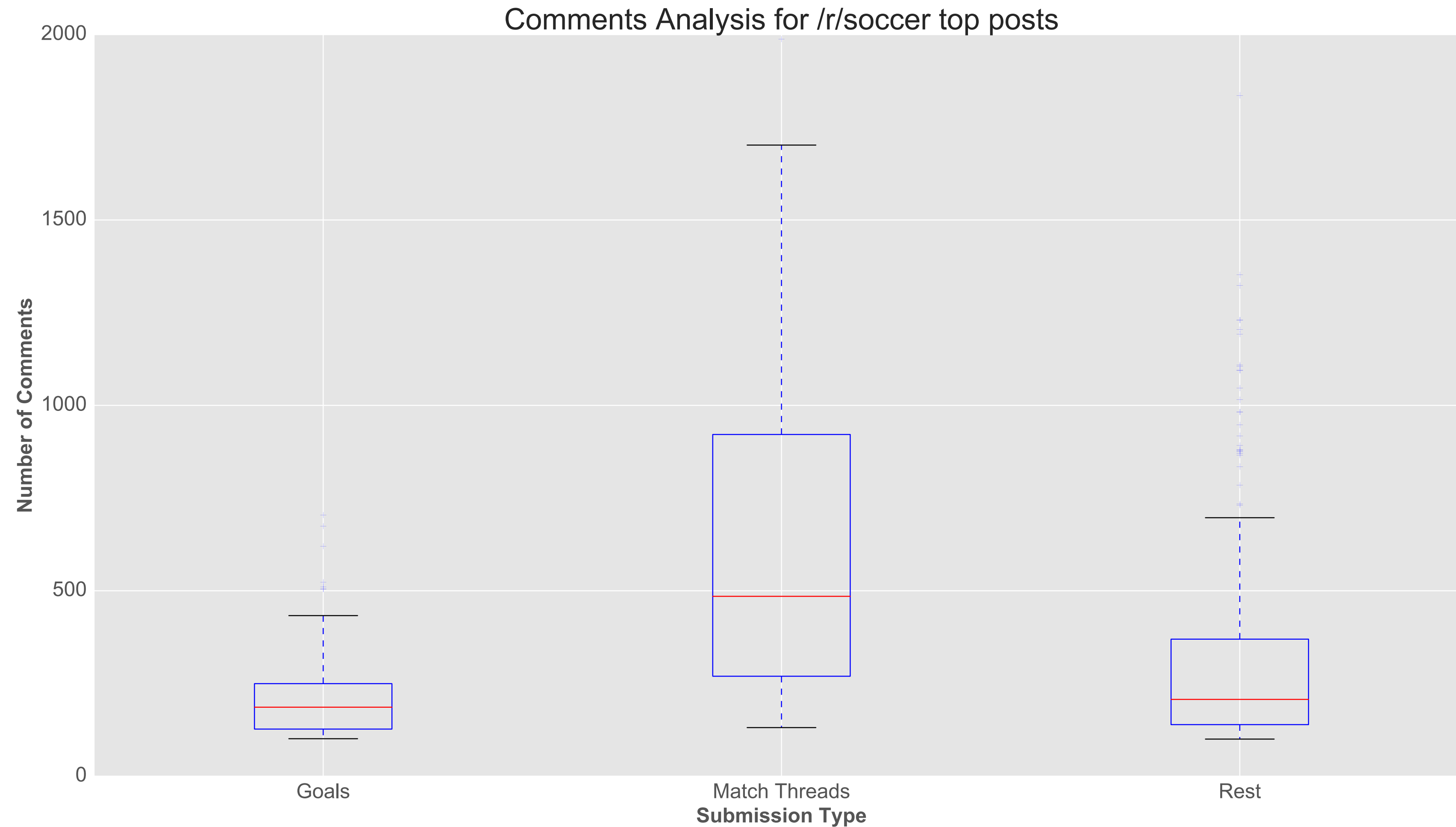
- **Goal video submission**

- Posted in near real-time with high quality goals getting top score
- Hypothesis:
 - ★ Flair diversity as goals are discussed for their quality
 - ★ Comments proportional to score

- **Match Thread, Post-Match Thread submissions**

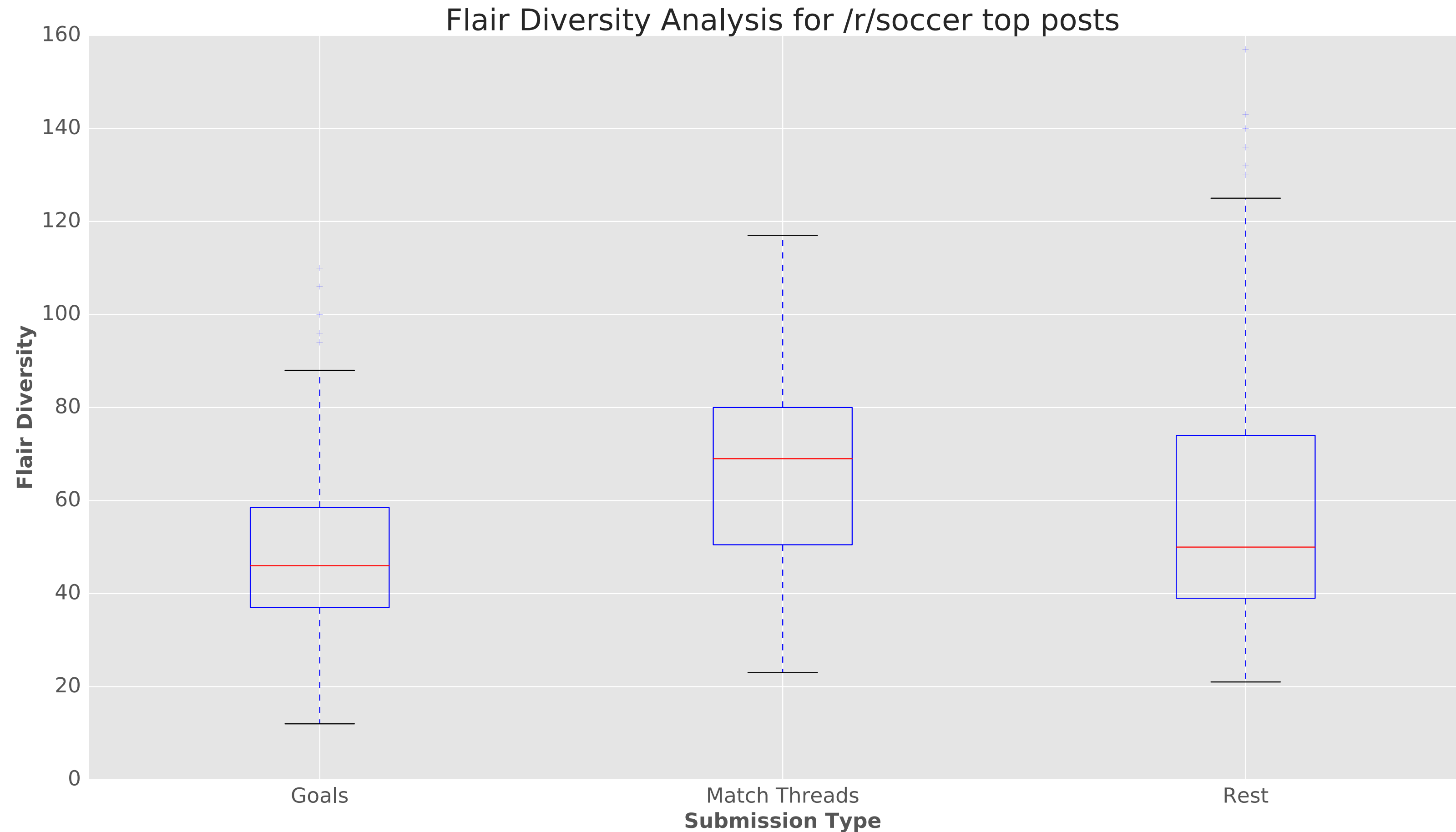
- Discussion during and after live match
- Hypothesis:
 - ★ Low flair diversity as teams taking part in the match expected to have high share
 - ★ High number of comments as users comment on various events not just goals

Submission Type - Comments Analysis



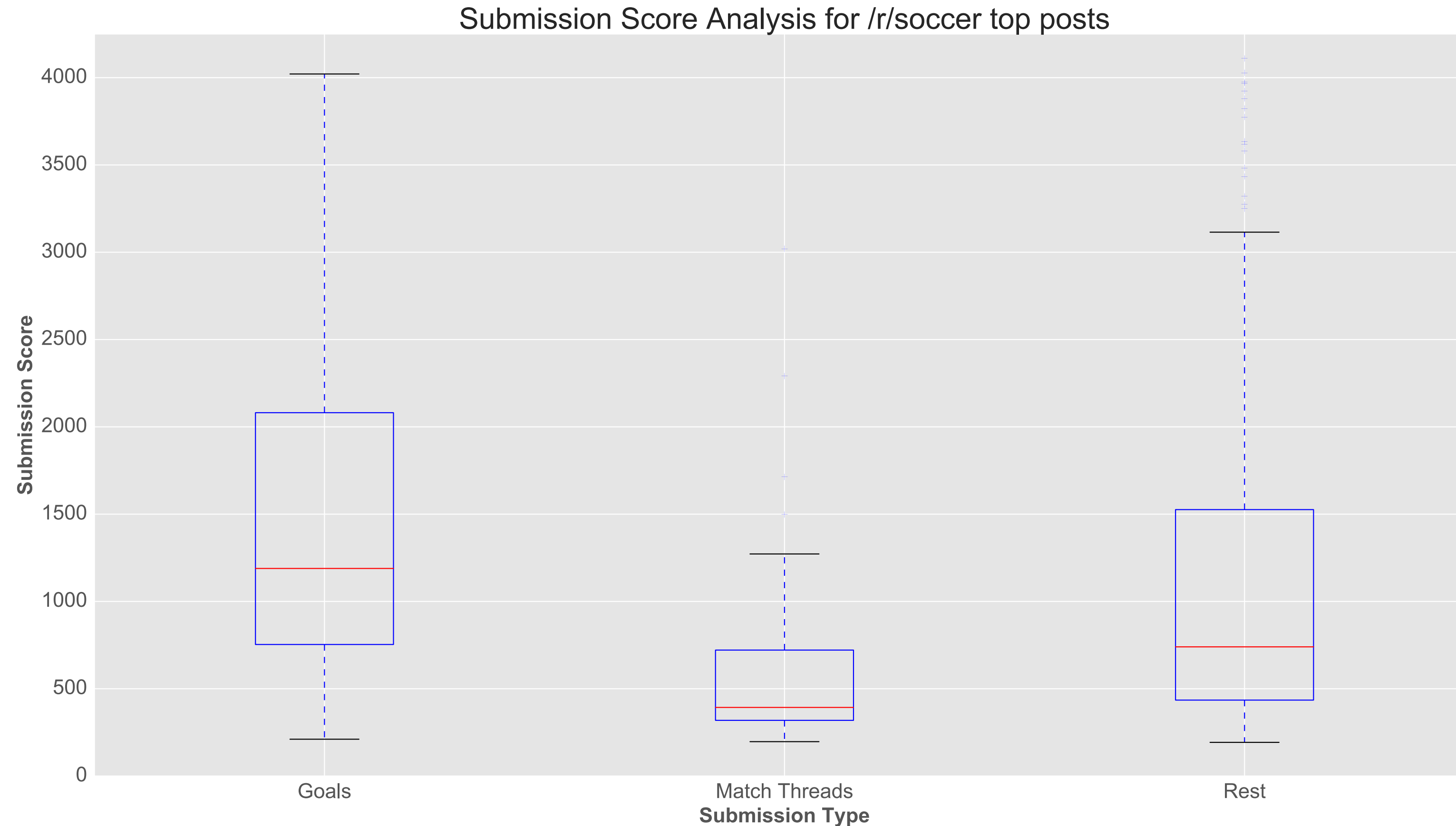
- **Expectedly, comments are higher for Match Threads**

Submission Type - Flair Diversity Analysis



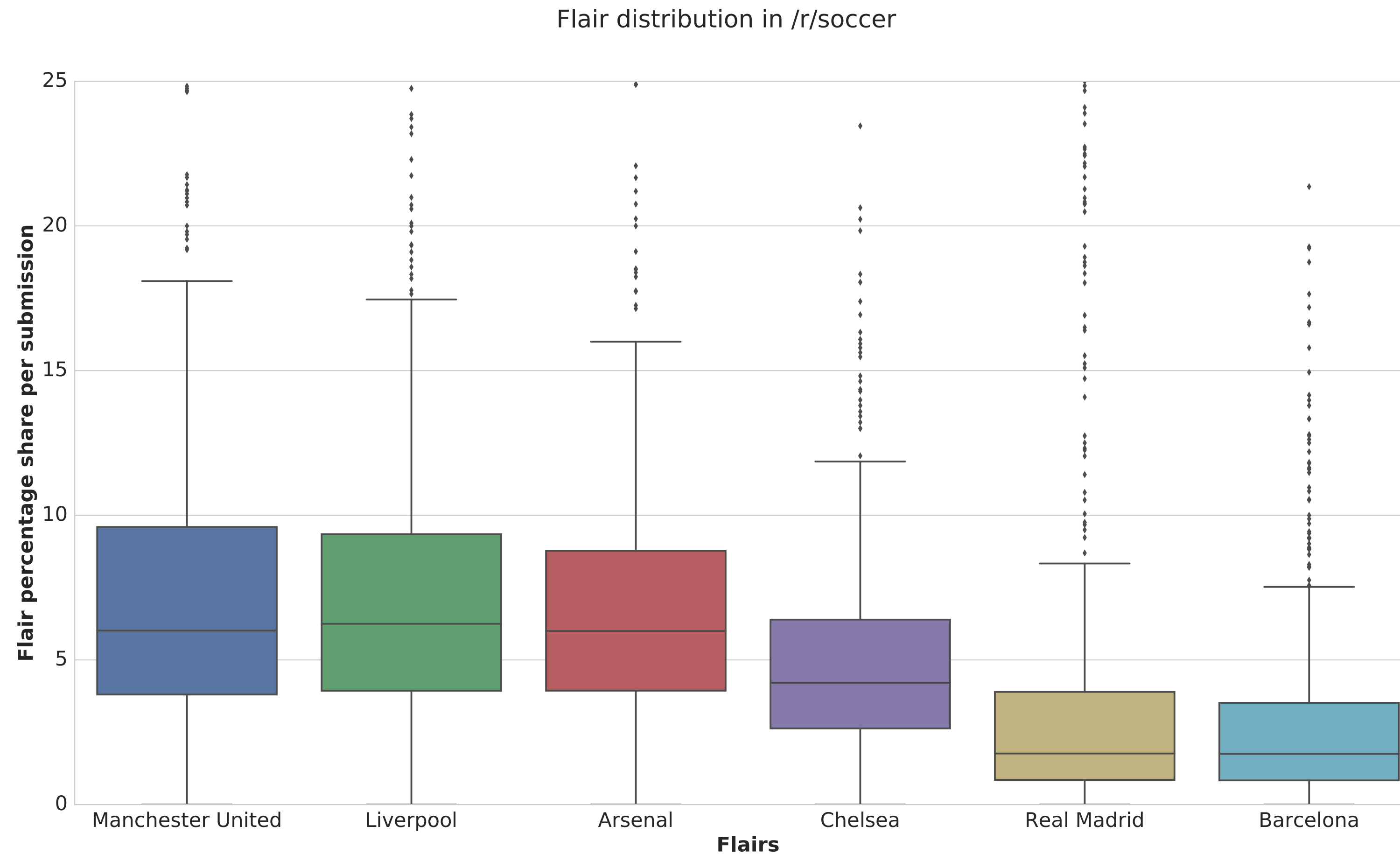
- **Unexpectedly, Match Threads have higher flair diversity**
- **Match Threads among top posts likelier discussed by other fans**

Submission Type - Score Analysis



- **Match and Post-Match Threads most active during and just after a match**
- **Goal submissions are rated for the quality of goal increasing over time**

Flair Distribution



- **English Premier League clubs lead the /r/soccer table!**

Conclusion

- **Scraped, visualized and analyzed /r/soccer top posts during past one month**
 - ★ **Flair Diversity, Score and Top Flair Share relationships**
 - ★ **Submission type-analysis**
 - ★ **Flair distributions**

