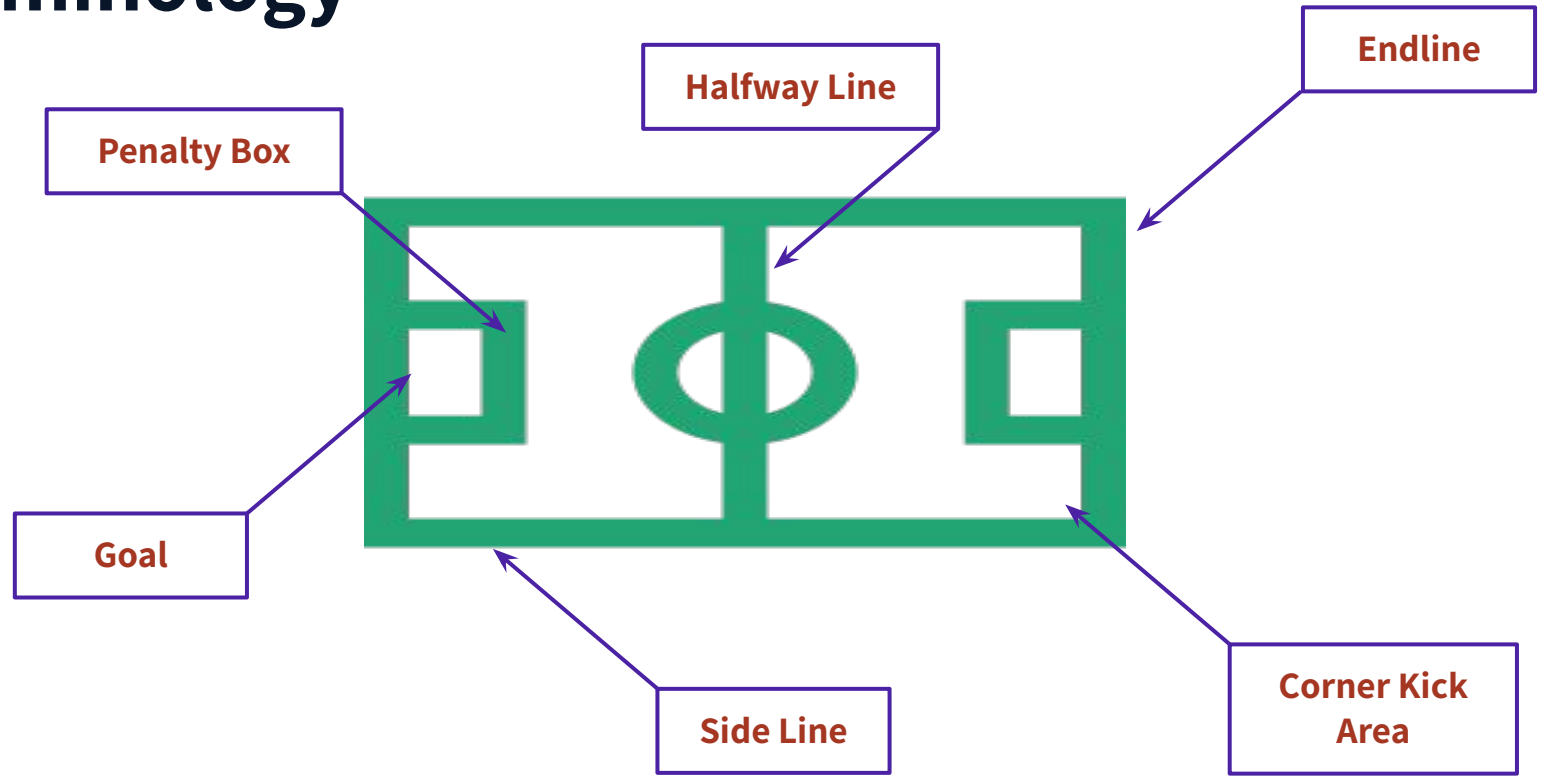
Investigating Set Pieces With Tracking Data

S.Gonzalez, R. Coke.

What do I need to know about soccer to understand this project?

Field Terminology



Set Pieces

Events occuring after a stoppages in play (“dead balls”) such as:

- Ball going out of bounds
- A foul being committed
- Goalie making a “save”

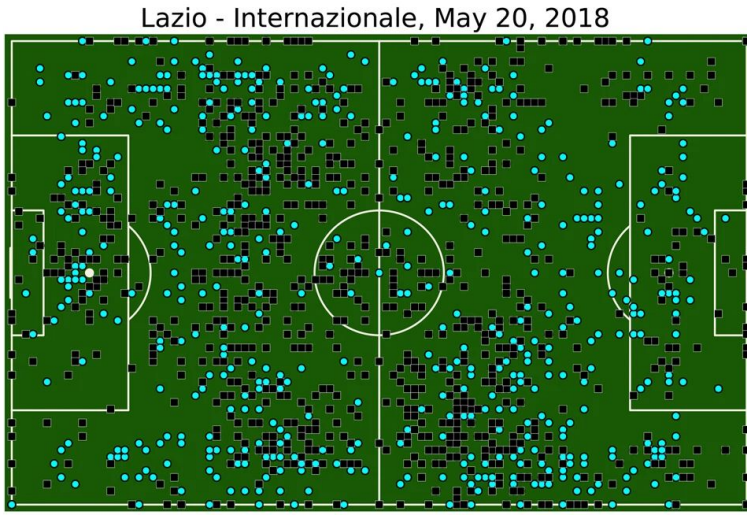
Set Pieces

General definition includes:

- Free Kicks
- Goal Kicks
- Corner Kicks
- Penalty Kicks
- Throw-ins

Soccer is cool! Now tell me about your data...

(Free) Event Tracking Data

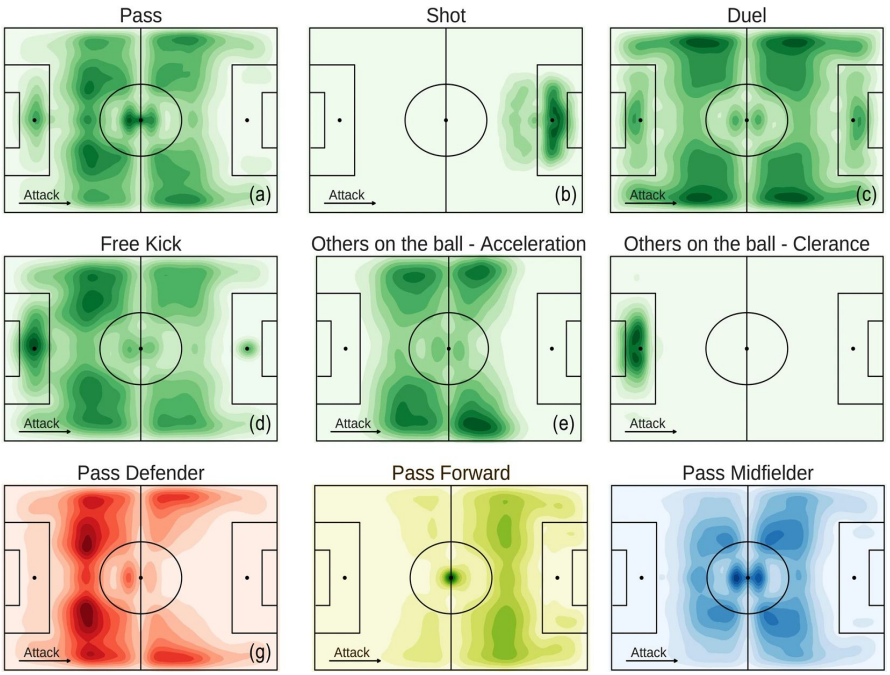


Emphasis on *events:* discrete occurrences in match that include passes, shots, fouls, duels, etc...

id	Event Id	Sub-Event Id	tags	Player Id	positions	Match Id	Team Id	Match Period	event Sec
...	...	...	...	...	...	...	...	...	...

Quirks About Data

- Listed positions for an event are from the perspective of the initiating team.
- No tracking of match score.
- No tracking of players, only the ball.
- Description of events can be very specific.



Automated Set Piece Scout

Why use data to analyze set pieces?

Problem: How to leverage this vast amount of data to help teams *win* games.

- Set pieces provide special scoring opportunities for teams.
  - The chances of scoring off of a free kick by the goal was 50% higher than that of normal play in the 2018 WC. Though figure is usually around 25%.
- Discrete “bursts” of events that are easier to analyze with ML.

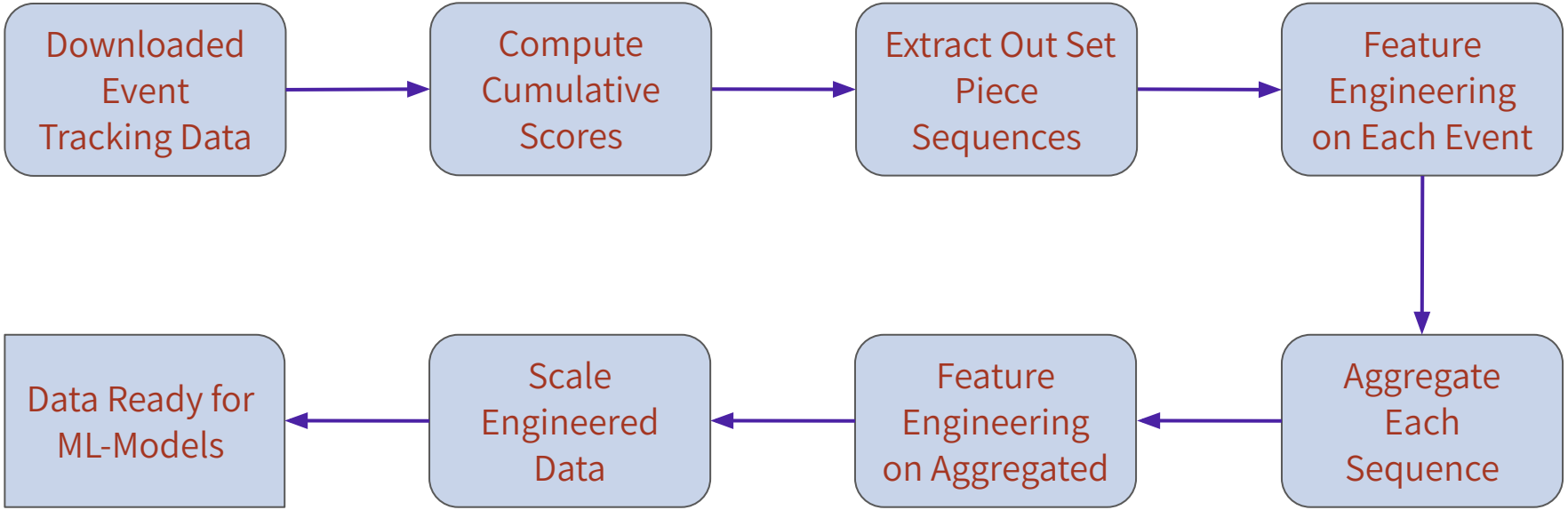
Generalizer of Set Piece Strategy

MVP/POC Solution: Take the context and spatial information provided by the event tracking data set and determine if it is possible to algorithmically determine and differentiate between the strategies employed on set pieces despite only having the limited information of the free data set.

If so, then this system can be applied and adapted towards a plethora of specific use-cases.

How did you decide to model this data?

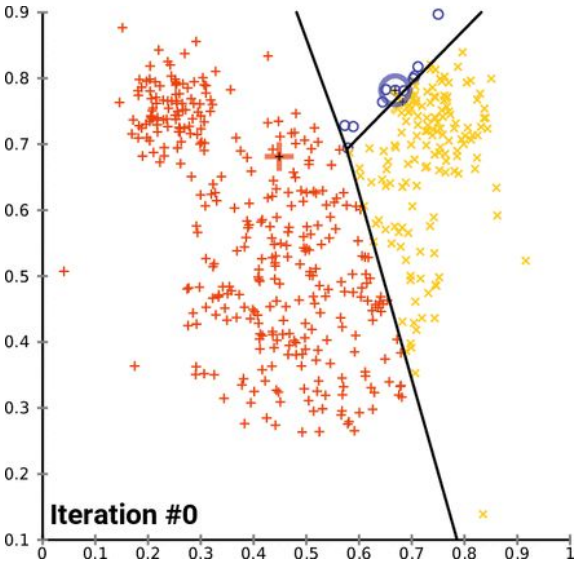
Data Preprocessing Pipeline



Clustering, clustering, clustering

POC is *complete* if we are able to achieve good *clustering* results.

MVP potentially could be achieved by a *simple* clustering model. Implement *K-means*.



Leicester City v. Tottenham Hotspur; May 13, 2018.

Game Context

Final match of the season for both clubs.

High stakes involved:

3rd-Place finish in EPL Table with victory.

End losing streak team had headed into the match.

Premier League · 5/13/18 Full-time

Tottenham	5	-	4	Leicester City
Harry Kane 7', 76' Érik Lamela 49', 60' Christian Fuchs 53' (OG)		⚽		Jamie Vardy 4', 73' Riyad Mahrez 16' Kelechi Iheanacho 47'

TIMELINE LINEUPS STATS NEWS COMMENTS

	TEAM STATS	
14	Shots	17
6	Shots on target	9
64%	Possession	36%
480	Passes	265
9	Fouls	13
1	Yellow cards	2
0	Red cards	0
2	Offsides	4
4	Corners	4

Stadium: Wembley Stadium

Example Set Piece Sequences From Match

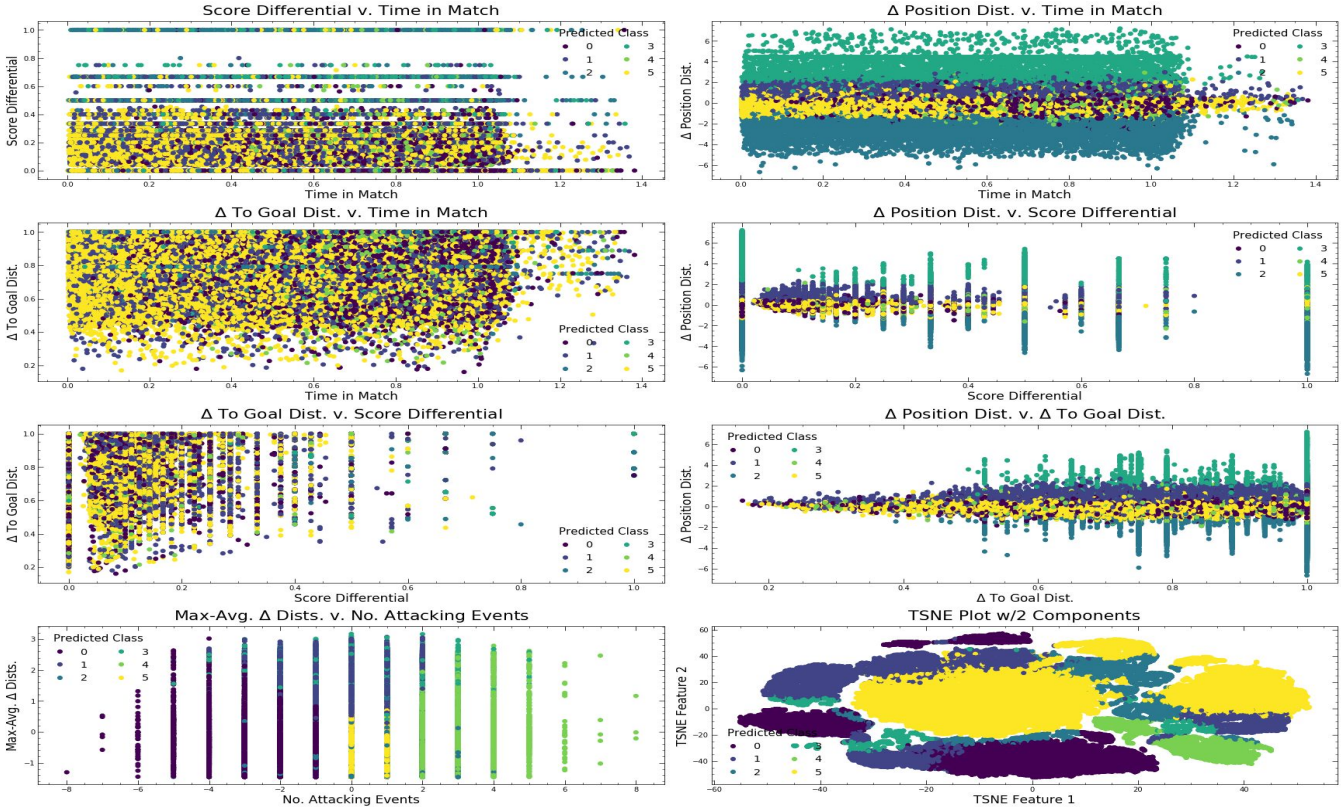
id	eventId	subEventName	tags	playerId	positions	matchId	eventName	teamId	matchPeriod	eventSec	subEventId	seq_id	score
251620097	3	Free kick cross	[{'id': 301}, {'id': 801}, {'id': 1801}]	26150	[{'y': 9, 'x': 82}, {'y': 40, 'x': 92}]	2500097	Free Kick	1631	1H	185.546375	32.0	52033	0-0
251620098	10	Shot	[{'id': 101}, {'id': 403}, {'id': 201}, {'id':...	12829	[{'y': 40, 'x': 92}, {'y': 100, 'x': 100}]	2500097	Shot	1631	1H	186.670776	100.0	52033	1-0

Example Set Piece Sequences From Match

id	eventId	subEventName	tags	playerId	positions	matchId	eventName	teamId	matchPeriod	eventSec	subEventId	seq_id	score
251621050	3	Goal kick	[]	25381	[{'y': 0, 'x': 0}, {'y': 84, 'x': 15}]	2500097	Free Kick	1624	2H	431.008021	34.0	52088	3-2
251621052	8	Simple pass	[{'id': 1801}]	36	[{'y': 84, 'x': 15}, {'y': 56, 'x': 4}]	2500097	Pass	1624	2H	434.662948	85.0	52088	3-2
251621053	8	Simple pass	[{'id': 1801}]	25381	[{'y': 56, 'x': 4}, {'y': 15, 'x': 23}]	2500097	Pass	1624	2H	437.778703	85.0	52088	3-2
251621054	8	Simple pass	[{'id': 1801}]	210044	[{'y': 15, 'x': 23}, {'y': 33, 'x': 13}]	2500097	Pass	1624	2H	442.176578	85.0	52088	3-2
251621055	8	Launch	[{'id': 1801}]	25381	[{'y': 33, 'x': 13}, {'y': 29, 'x': 70}]	2500097	Pass	1624	2H	447.700492	84.0	52088	3-2
251621057	1	Air duel	[{'id': 703}, {'id': 1801}]	40765	[{'y': 29, 'x': 70}, {'y': 14, 'x': 96}]	2500097	Duel	1624	2H	450.721087	10.0	52088	3-2
251621098	1	Air duel	[{'id': 701}, {'id': 1802}]	149019	[{'y': 71, 'x': 30}, {'y': 86, 'x': 4}]	2500097	Duel	1631	2H	450.843516	10.0	52088	3-2
251621061	8	Cross	[{'id': 401}, {'id': 1801}]	8292	[{'y': 14, 'x': 96}, {'y': 38, 'x': 89}]	2500097	Pass	1624	2H	455.771248	80.0	52088	3-2
251621066	8	Simple pass	[{'id': 1801}]	40765	[{'y': 38, 'x': 89}, {'y': 44, 'x': 90}]	2500097	Pass	1624	2H	457.527051	85.0	52088	3-2
251621068	7	Touch	[]	20441	[{'y': 44, 'x': 90}, {'y': 49, 'x': 91}]	2500097	Others on the ball	1624	2H	457.997047	72.0	52088	3-2
251621103	1	Ground loose ball duel	[{'id': 703}, {'id': 1801}]	8653	[{'y': 51, 'x': 9}, {'y': 49, 'x': 4}]	2500097	Duel	1631	2H	458.007491	13.0	52088	3-2
251621069	1	Ground loose ball duel	[{'id': 701}, {'id': 1802}]	20441	[{'y': 49, 'x': 91}, {'y': 51, 'x': 96}]	2500097	Duel	1624	2H	458.589089	13.0	52088	3-2
251621104	7	Touch	[{'id': 102}]	14853	[{'y': 49, 'x': 4}, {'y': 100, 'x': 100}]	2500097	Others on the ball	1631	2H	459.641226	72.0	52088	3-3

What results did you get?

Clustering in Feature Space

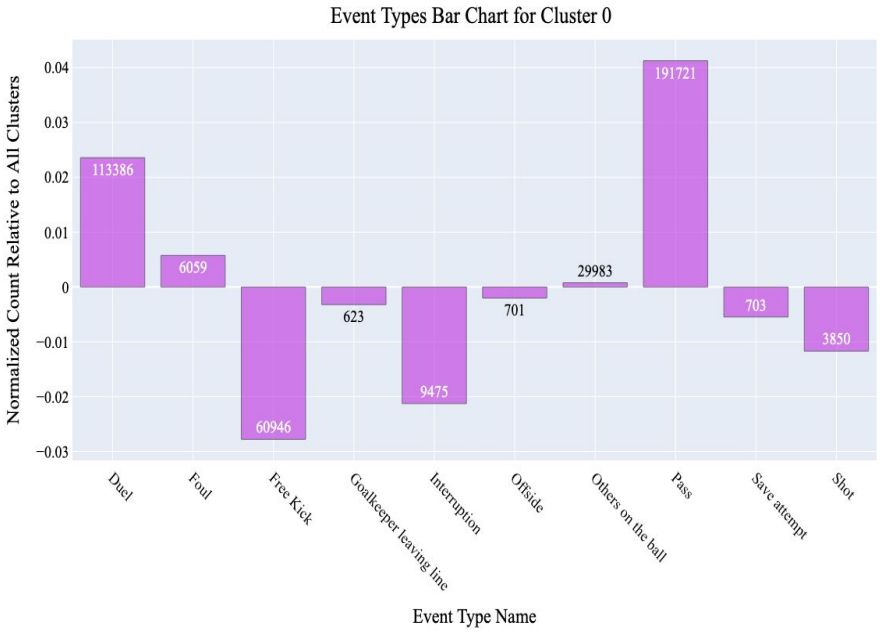


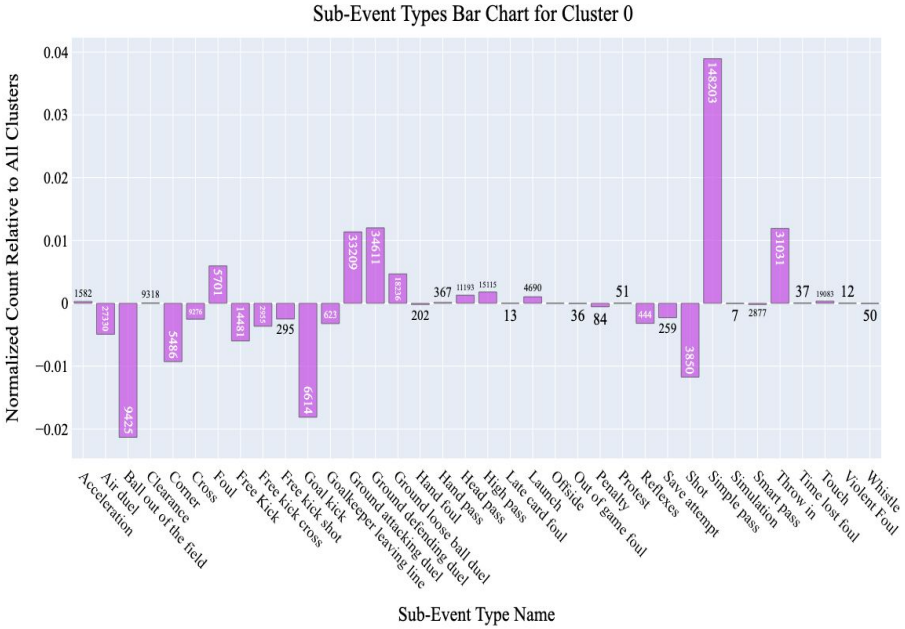
Clustering in Feature Space

Data points that are closest to the centroids of the clusters:

	match_time	is_goalie	is_mid	is_def	is_foward	num_attacking_events	score_diff	pos_delta_diff	to_goal_delta_diff	delta_max_avg
0	0.609782	0.000000	0.384615	0.461538	0.153846	0.863101	-1.0	24.575367	-6.281971	30.674985
1	0.517418	0.250000	0.250000	0.250000	0.250000	0.937500	0.0	42.043327	2.559252	77.375635
2	0.386678	0.666667	0.000000	0.333333	0.000000	0.888889	0.0	55.273207	-37.486577	20.992771
3	0.360201	0.333333	0.333333	0.000000	0.333333	0.911111	0.0	85.818990	34.076722	93.830101
4	0.725466	0.000000	0.375000	0.500000	0.125000	0.949212	2.0	25.257262	-5.262720	32.806684
5	0.469464	0.000000	0.200000	0.500000	0.200000	0.840676	0.0	20.932159	-5.491030	28.432100

Cluster Investigation I.

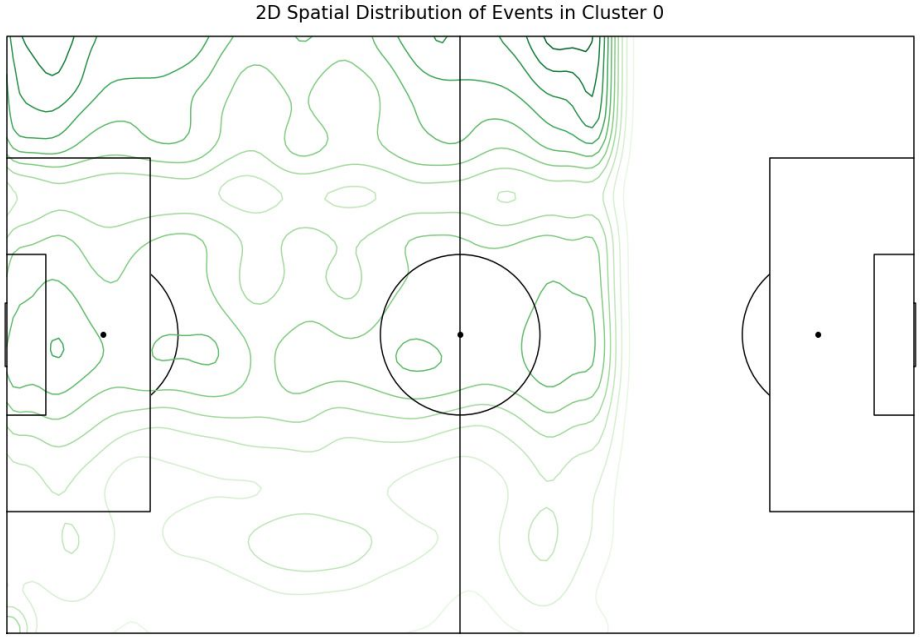




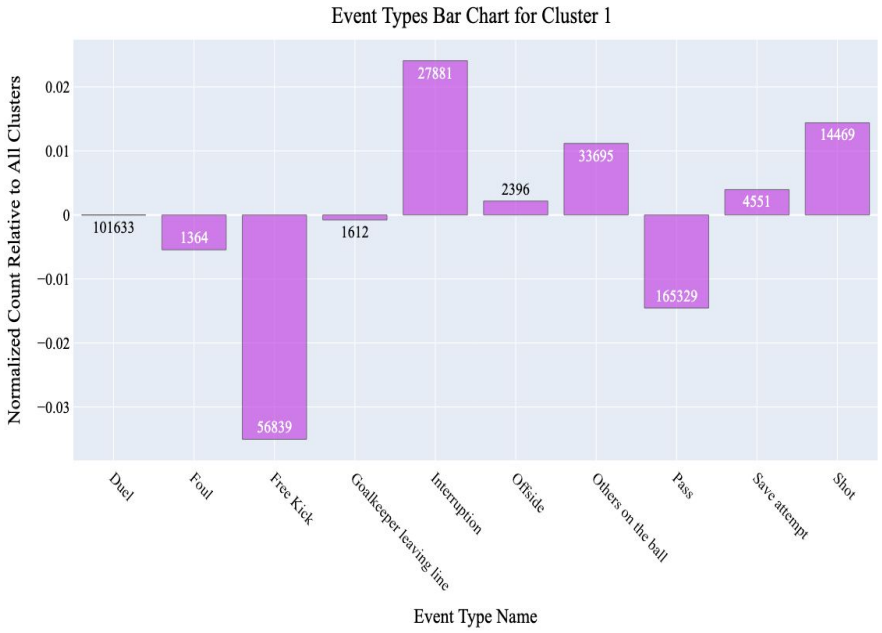
Cluster Investigation I.

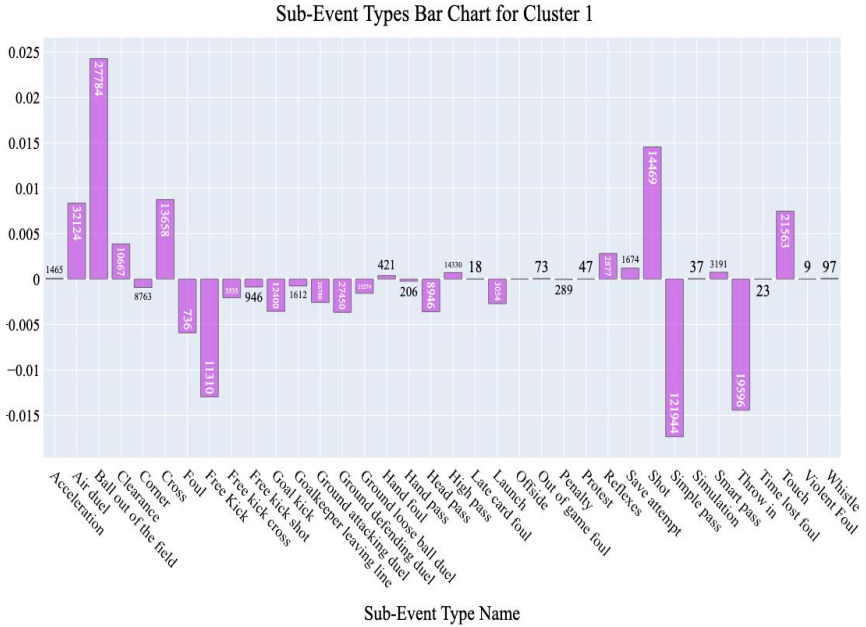
Key cluster characteristics:

- Initiating team is losing
- No goalie involvement
- Relatively low possession rate
- Little advancement towards goal
- Mainly simple passes



Cluster Investigation II.

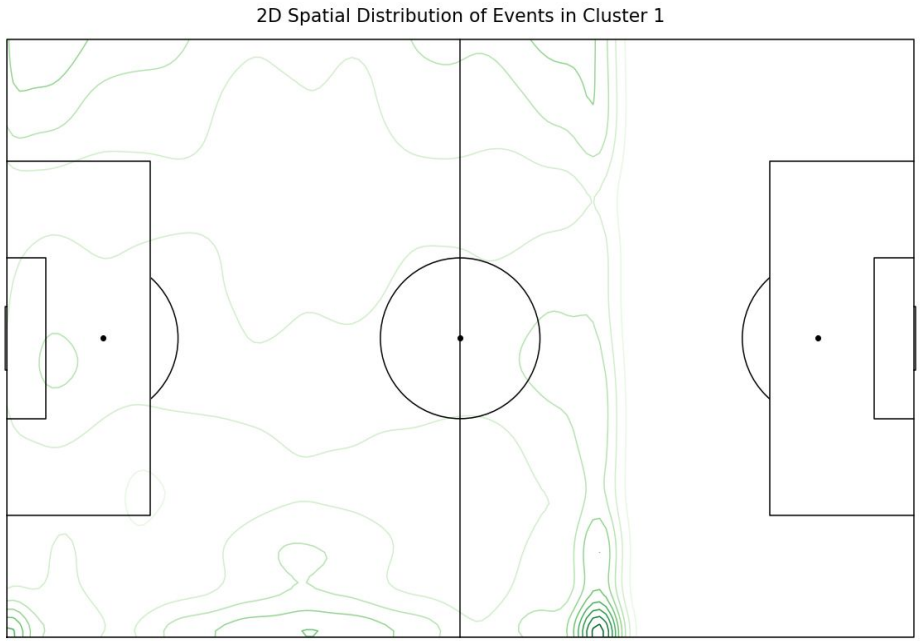




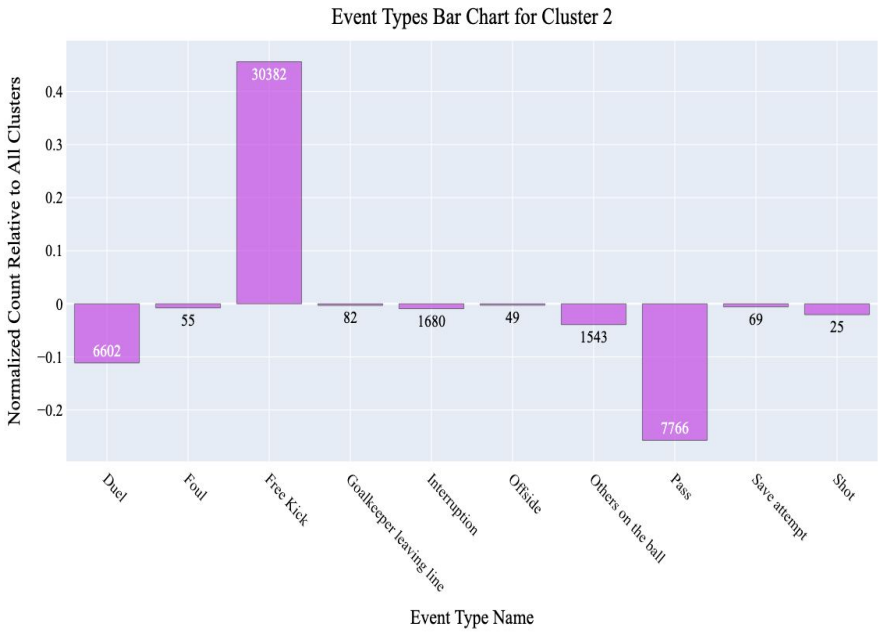
Cluster Investigation II.

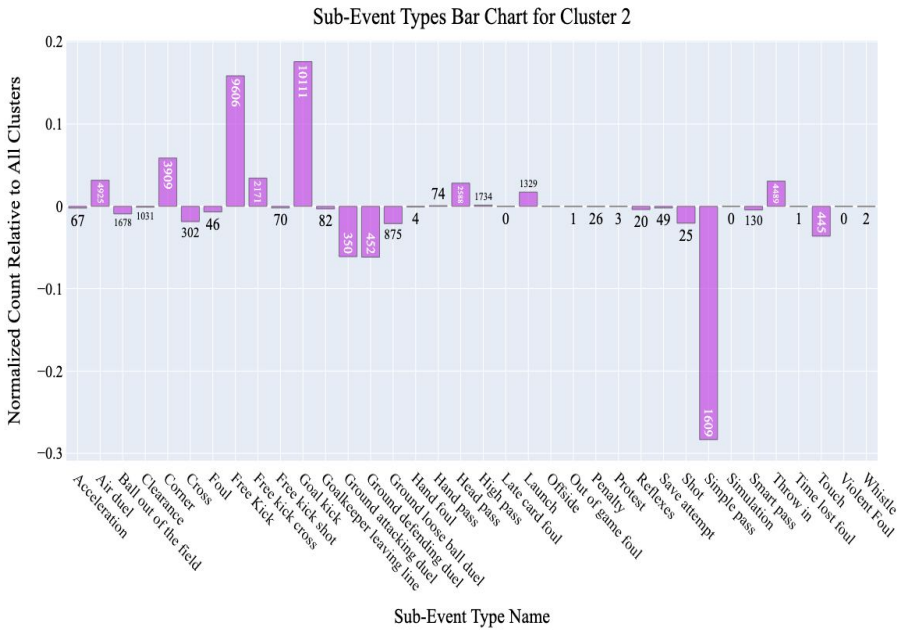
Key cluster characteristics:

- Tied match
- Soon after halftime
- Highest rate of shot attempts
- Passes played out of attacking half



Cluster Investigation III.

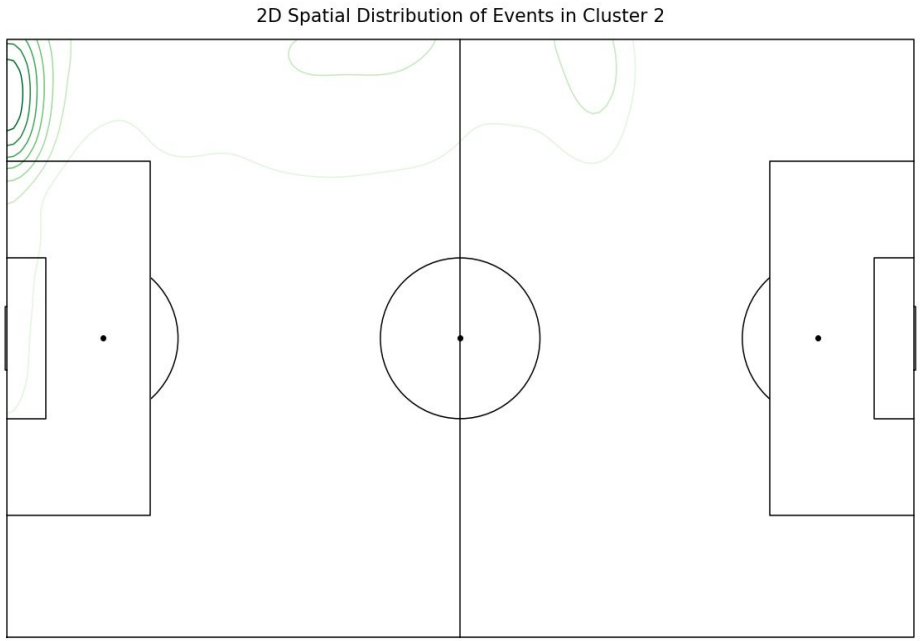




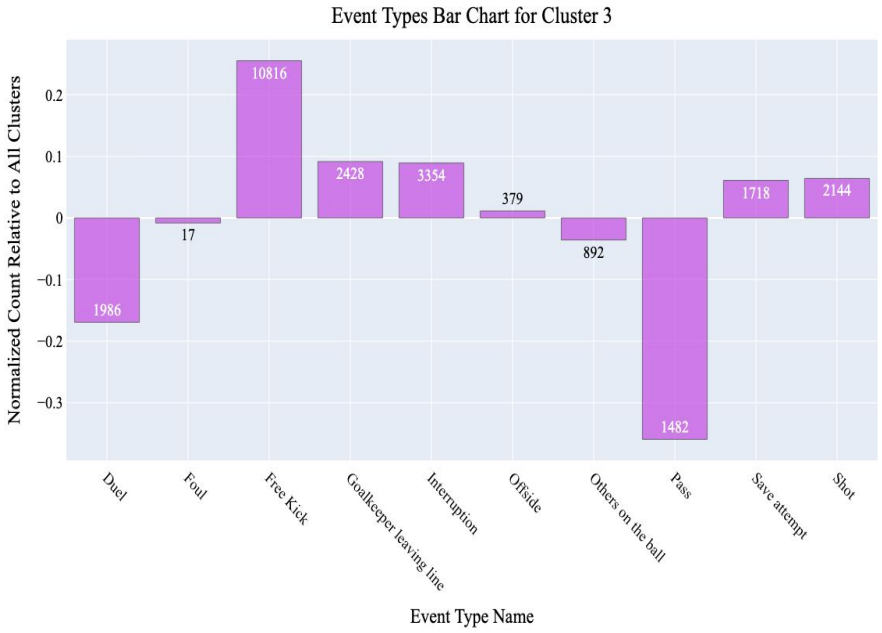
Cluster Investigation III.

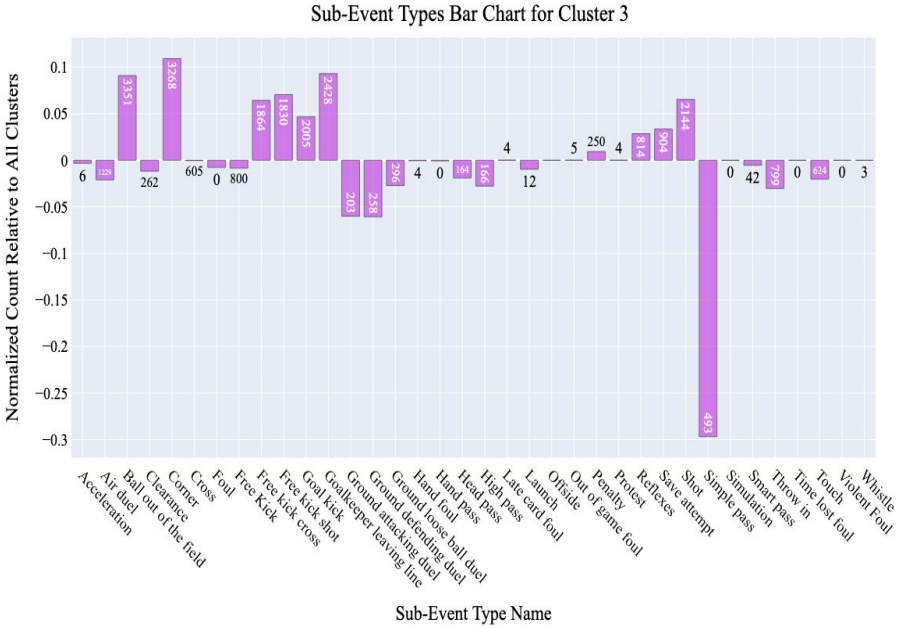
Key cluster characteristics:

- Tied match
- Early in the match
- Many long passes
- Not much progress towards goal
- Many goal kicks



Cluster Investigation IV.

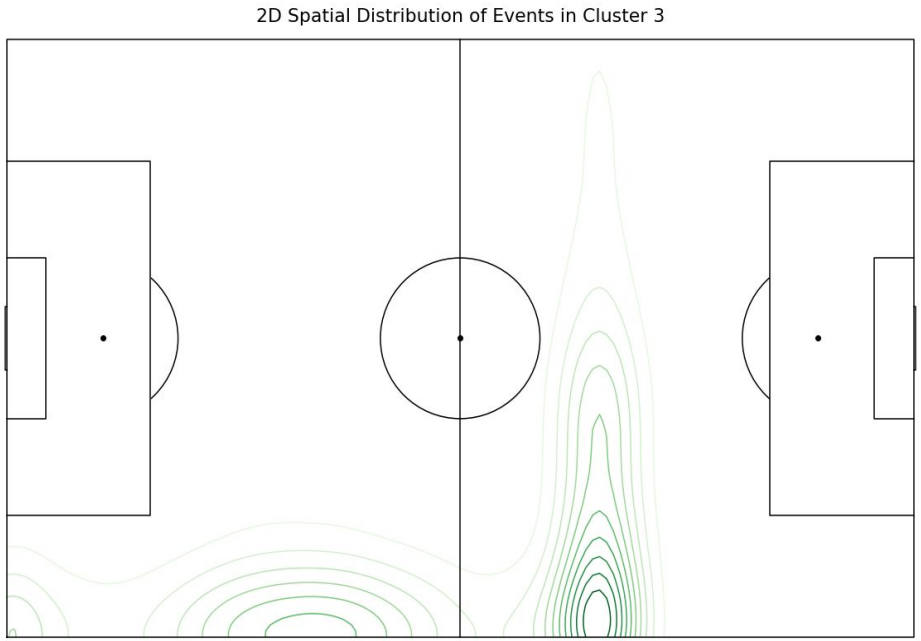




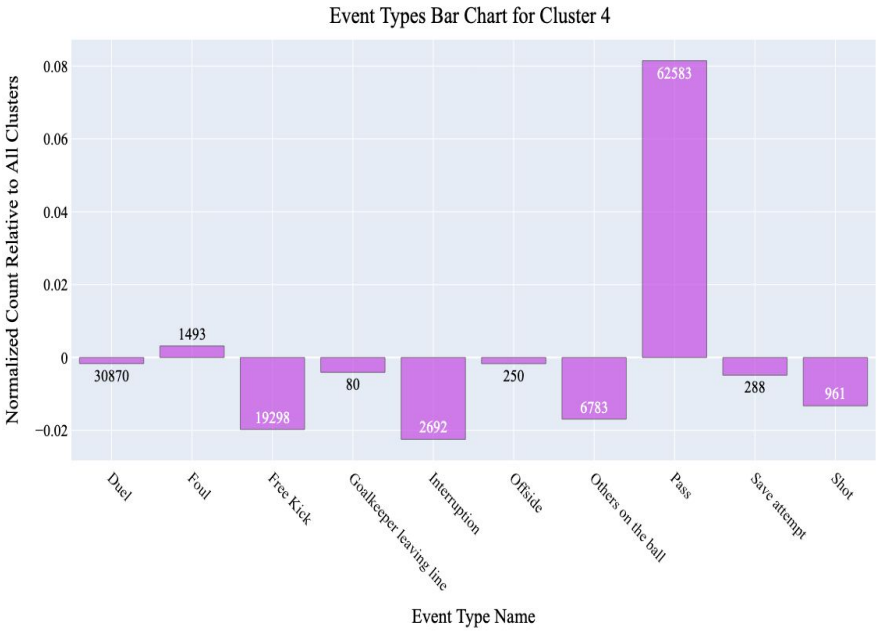
Cluster Investigation IV.

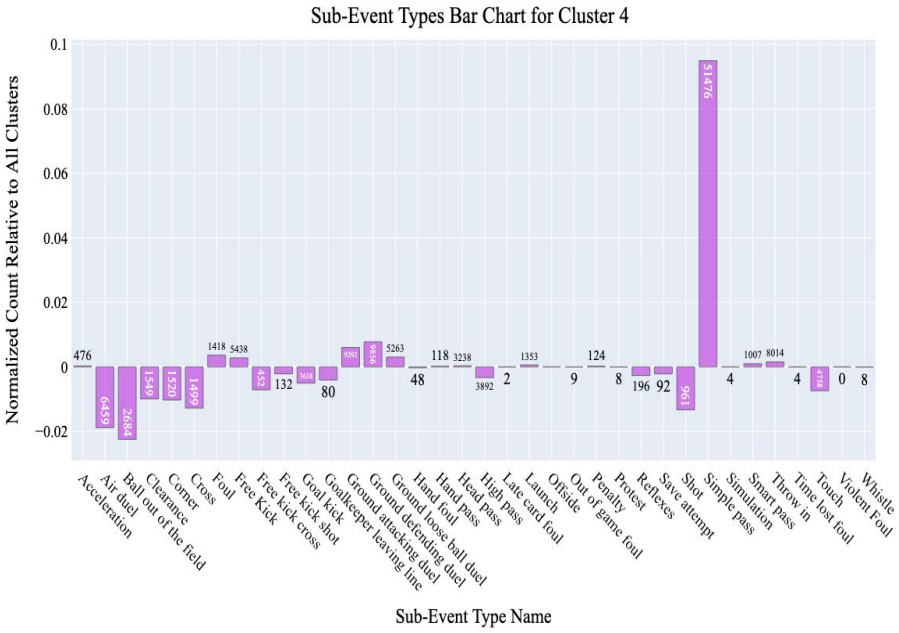
Key cluster characteristics:

- Tied match
- Early in the match
- Highest rate of forward initiation
- Similar to second cluster
- Best advancement towards goal



Cluster Investigation V.

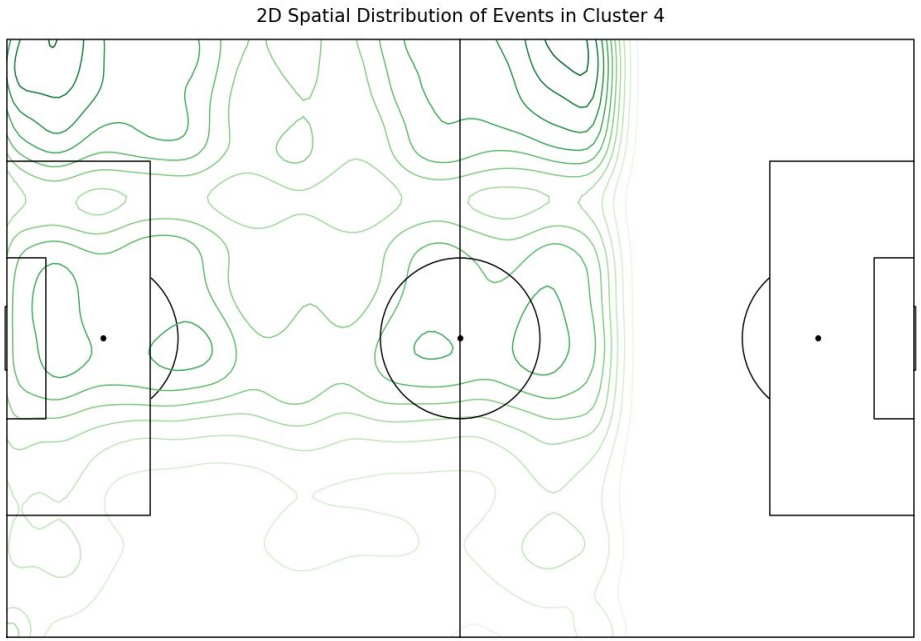




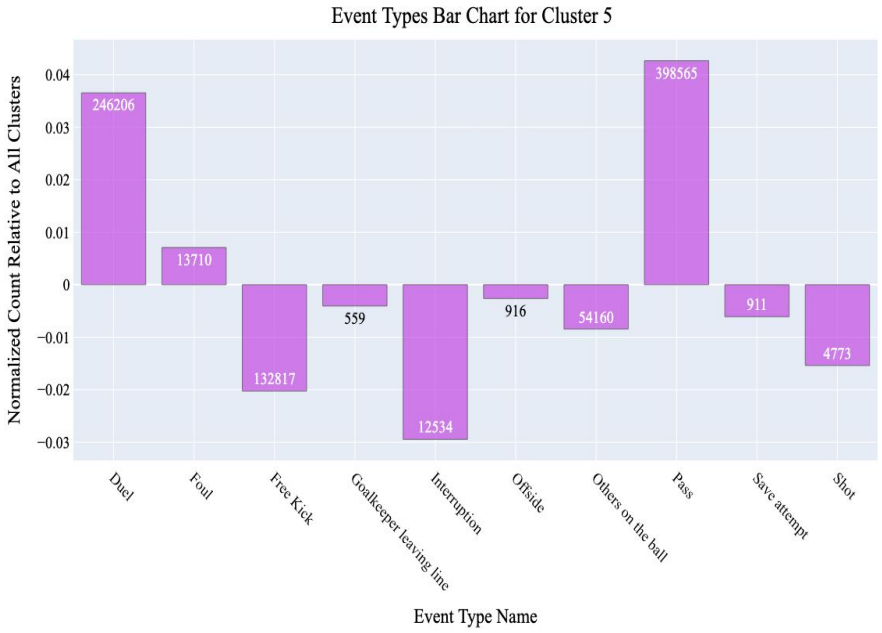
Cluster Investigation V.

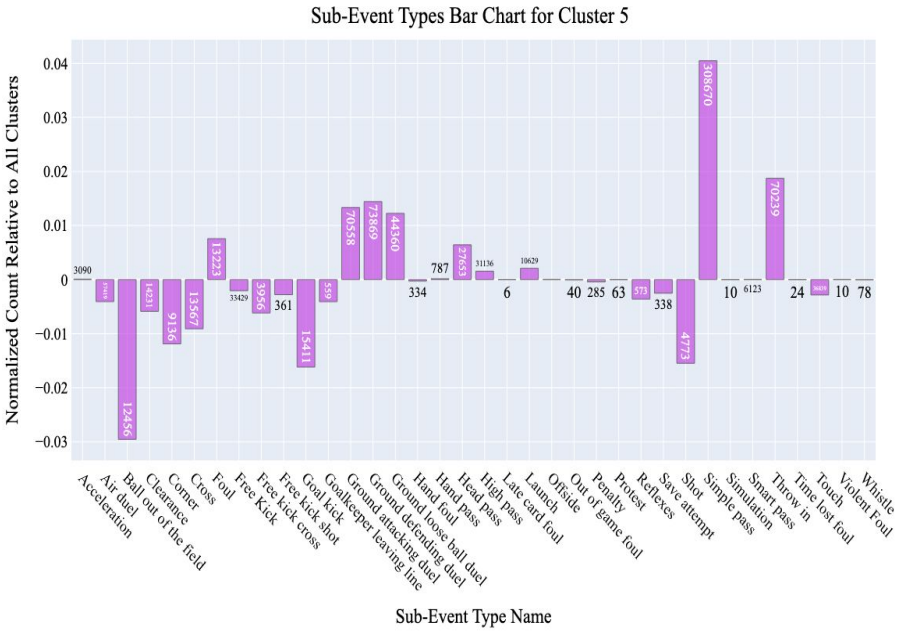
Key cluster characteristics:

- Initiating team has a big lead
- Late in the match
- No goalie involvement
- Highest possession rate
- Similar to first cluster



Cluster Investigation VI.

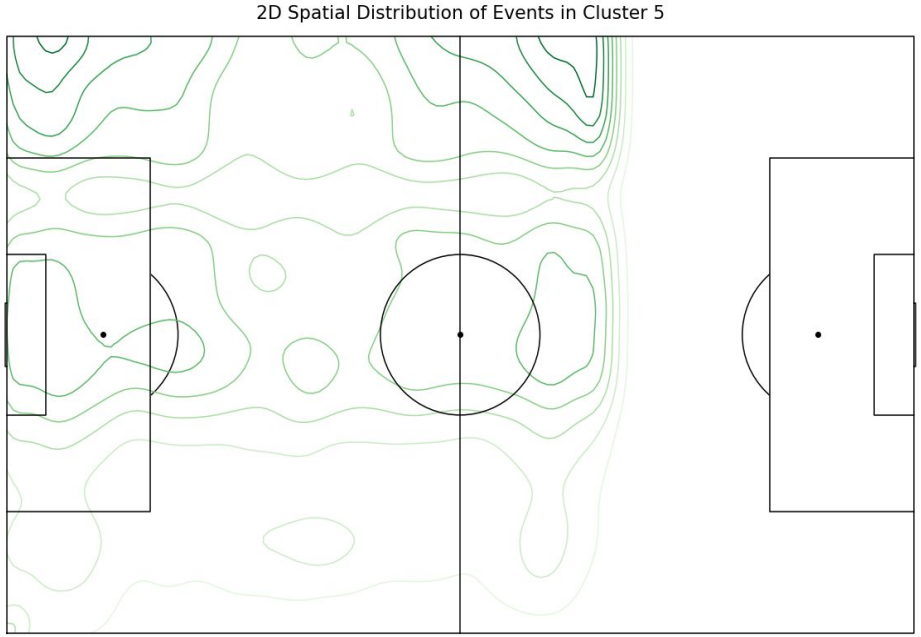




Cluster Investigation VI.

Key cluster characteristics:

- Tied match
- Mostly passes and duels
- Right before halftime
- Minimal advancement towards goal



POC? MVP?

K-Means model was able to make clear cluster distinctions using given data!

- Plus, simple models allows for easy interpretation. Clusters given make sense and are informative.
- Easily scalable!

What’s next?

Assess Performance in New Use Cases

- Investigate model behavior on subsets of the data partitioned on set piece type
- Analyze model predictions on a team-by-team basis as a consistency check.
- Build code infrastructure to handle full event tracking data.

References

1. https://www.nature.com/articles/s41597-019-0247-7#Tab2
2. https://figshare.com/collections/Soccer_match_event_dataset/4415000
3. https://en.wikipedia.org/wiki/K-means_clustering
4. https://www.washingtonpost.com/news/fancy-stats/wp/2018/06/20/why-set-pieces-are-dominating-scoring-so-far-at-the-world-cup/

Questions?