How to score in soccer:

A data science experiment

By: Malcolm Kelly

Topic Question

Rationale: Soccer tactics over time have evolved to emphasize shorter, direct play through the middle of the field. Studying data on passess from professional soccer games can help determine what the most effective style of play is.

Research question: What types of passes are most effective to generate a goal in soccer?

Hypothesis

If I use code to create a heat map of every soccer pass leading to a shot or goal in a data set from the La Liga (the Spanish professional soccer league) spanning 15 years, then I think the most common area of assists will change from the wings in the mid 2000s to the center midfield in the 2020s because soccer tactics over time have evolved to emphasize shorter, direct play through the middle of the field.

Materials

- A computer with access to the internet
- Data from professional soccer matches with event coding (from Hudl-Statsbomb's public database)
- Free R coding software (Positron) for analysis
- Gemini for help with R coding.

Procedure

- I’m interested in how the location of passes that lead to shots or goals has changed over time, as soccer tactics have changed. I’m specifically interested in whether there are increasing passes close to the goal.
- In order to test this, I will use a free data set from Hudl-Statsbomb, an organization that catalogs soccer match event (play-by-play) data and creates statistics, charts, and graphs that professional teams use.
- The data is called “event data” because it is, basically, a log of every event that happened during a match. For this experiment, I will be using data from La Liga (the Spanish pro soccer league). Hudl-Statsbomb makes many matches from this league freely available, over a roughly 15 year span of seasons. This data will allow me to test my hypothesis.
- First, I will use R code written by Gemini AI to isolate the events in which a pass occurs, and in which a pass leads to a shot, then to a goal. (This is one of the categories of events that Hudl-Statsbomb includes).
- Next, I will create charts, including heatmaps, to describe where these passes come from, and how their location changes over the span of seasons.

Procedure (cont.)

- I will differentiate passes that were close to the goal (within 40 yards of goal line), versus ones that were far away, more than 40 yards away from the goal line, using a few lines of code. An example heatmap would have the field split into 30 evenly-sized boxes, and each box would be colored according to the percentage of passes from that area. An example dot plot would have the seasons as the x-axis, and the percentage of passes close to goal as the y-axis, and there are 2 points for each season, 1, say red, representing passes from out wide and 1, maybe blue, for middle of the field passes.

- My hypothesis is that play close to the goal will increase over the timeframe. I expect that passes from far away will decrease over the same time period.

Chart and Graph Pt. 1

Pivot Table #1 - Filtered to Close Passes - Total Passes - Wing (True) vs. Middle (False)

AVERAGE of percent_passes	*wing_pass*		
season_name	FALSE	TRUE	Grand Total
2006/2007	0.098	0.096	0.097
2007/2008	0.098	0.102	0.100
2008/2009	0.112	0.118	0.115
2009/2010	0.097	0.112	0.104
2010/2011	0.114	0.112	0.113
2011/2012	0.121	0.114	0.117
2012/2013	0.115	0.112	0.113
2013/2014	0.128	0.132	0.130
2014/2015	0.133	0.135	0.134
2015/2016	0.093	0.135	0.114
2016/2017	0.125	0.122	0.123
2017/2018	0.118	0.125	0.122
2018/2019	0.130	0.121	0.126
2019/2020	0.127	0.118	0.122
2020/2021	0.142	0.127	0.135
Grand Total	**0.117**	**0.119**	**0.118**

Chart and Graph Pt. 2

Table #3 - Filtered to Close Passes - Goal Assists - Wing (True) vs. Middle (False)

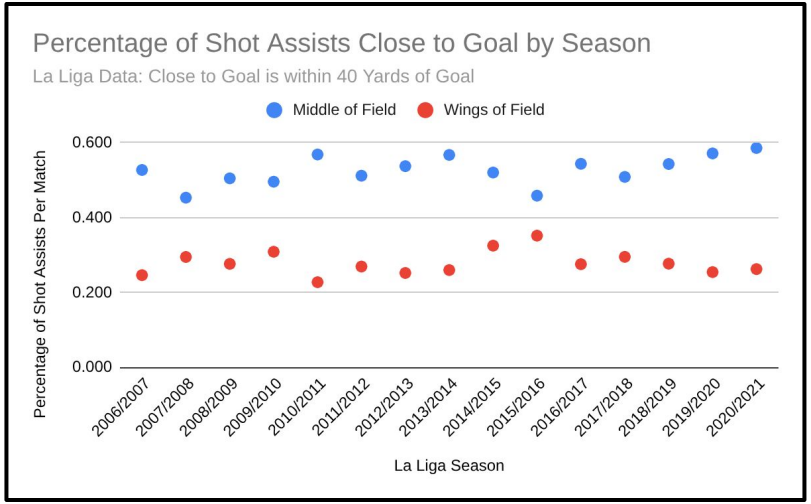
SUM of percent_passes	*wing_pass*		
season_name	FALSE	TRUE	Grand Total
2006/2007	0.549	0.255	0.804
2007/2008	0.719	0.175	0.895
2008/2009	0.627	0.147	0.773
2009/2010	0.644	0.115	0.759
2010/2011	0.623	0.145	0.768
2011/2012	0.553	0.259	0.812
2012/2013	0.691	0.165	0.856
2013/2014	0.680	0.200	0.880
2014/2015	0.671	0.207	0.878
2015/2016	0.564	0.295	0.859
2016/2017	0.655	0.226	0.881
2017/2018	0.733	0.163	0.895
2018/2019	0.680	0.160	0.840
2019/2020	0.606	0.197	0.803
2020/2021	0.582	0.239	0.821
Grand Total	**9.575**	**2.948**	**12.523**

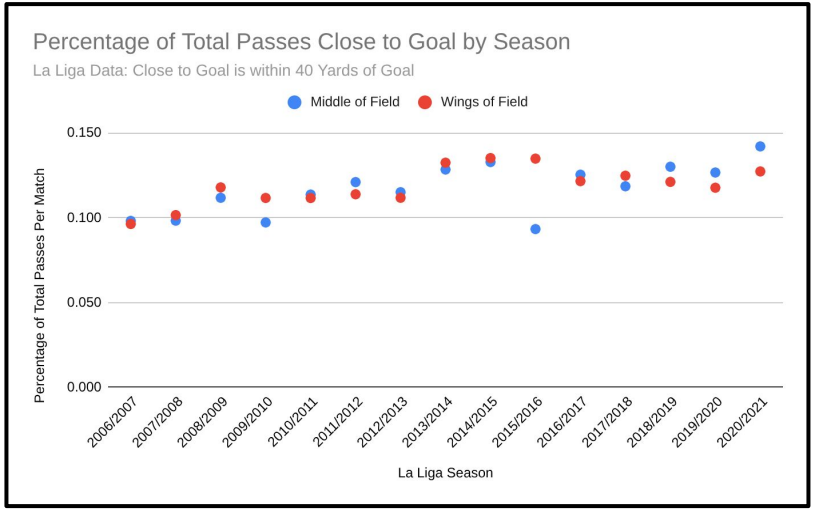
Table #2 - Filtered to Close Passes - Shot Assists - Wing (True) vs. Middle (False)

season_name	FALSE	TRUE	Grand Total
2006/2007	0.527	0.246	0.386
2007/2008	0.453	0.295	0.374
2008/2009	0.504	0.276	0.390
2009/2010	0.495	0.308	0.402
2010/2011	0.568	0.227	0.398
2011/2012	0.511	0.269	0.390
2012/2013	0.537	0.252	0.394
2013/2014	0.567	0.260	0.413
2014/2015	0.520	0.325	0.422
2015/2016	0.458	0.351	0.405
2016/2017	0.543	0.275	0.409
2017/2018	0.508	0.295	0.401
2018/2019	0.543	0.277	0.410
2019/2020	0.571	0.254	0.413
2020/2021	0.585	0.262	0.424
Grand Total	**0.526**	**0.278**	**0.402**

Chart and Graph Pt. 3

Percentage of passes in middle (blue) vs. wing (red)





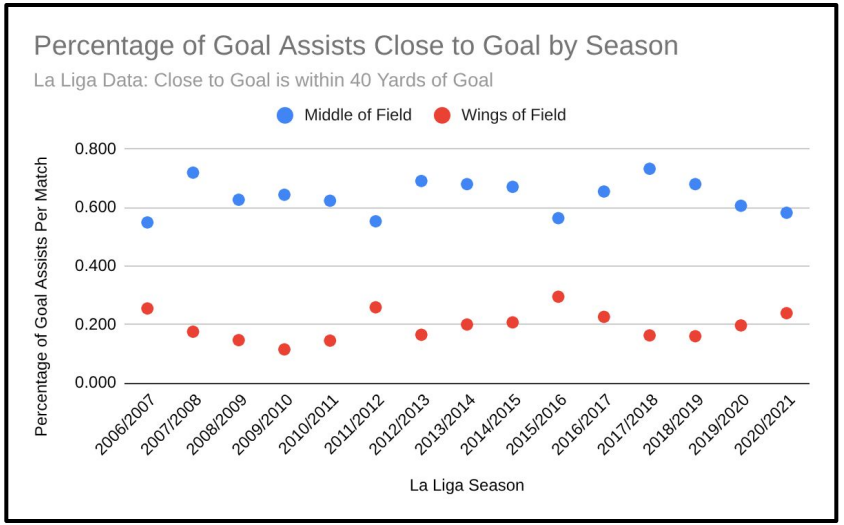


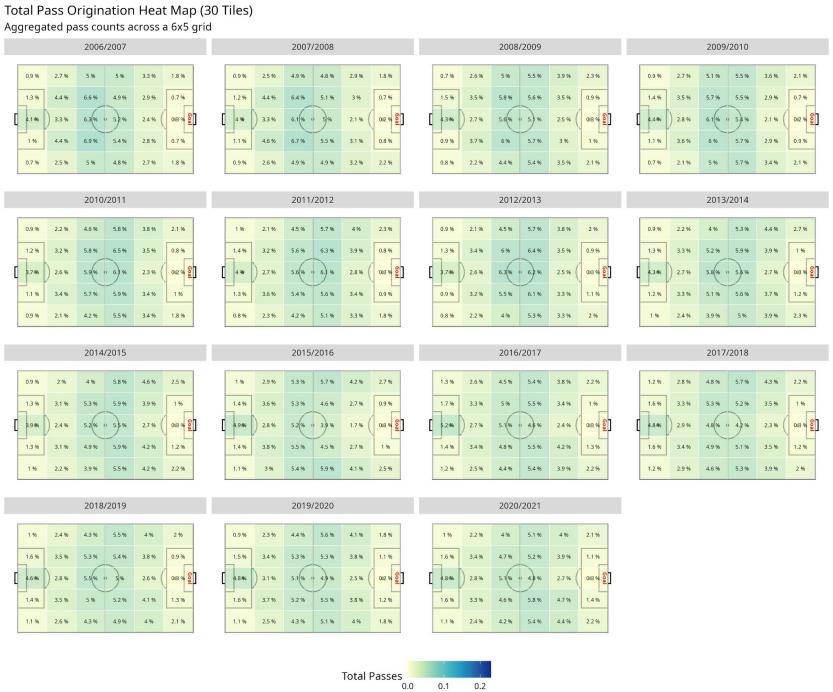
Chart and Graph Pt. 4

Shot Assist Origin by season (2005-2021) Heatmap

Chart and Graph Pt. 5

Goal Assist Pass Origin
Seasons 2005-2021
La Liga Heatmap

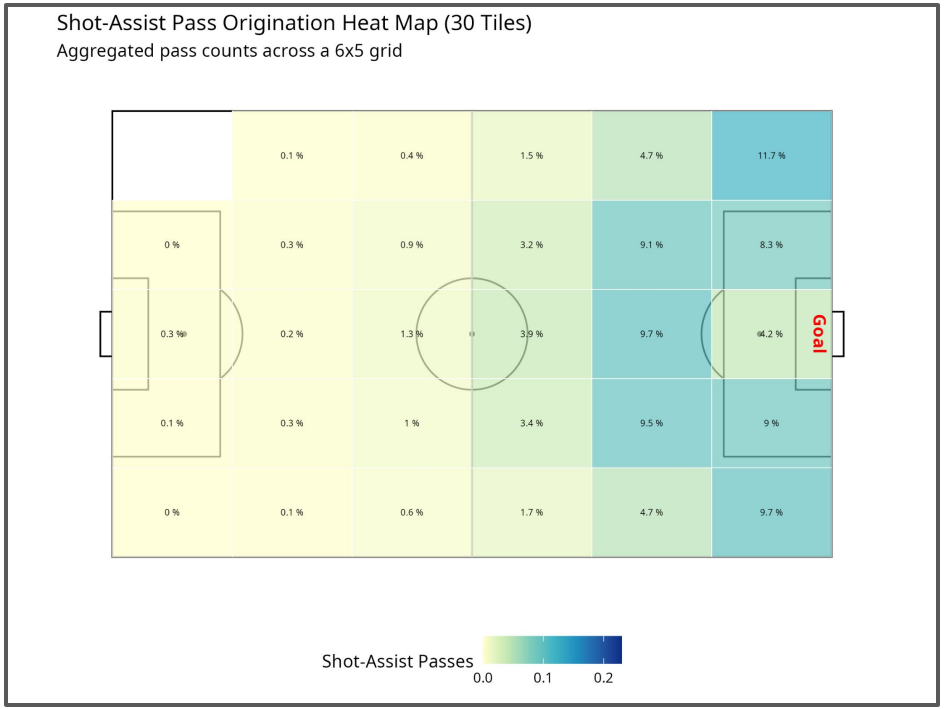
Chart and Graph Pt. 6



Total Passes Origin
Seasons 2005-2021
La Liga Heatmap

Chart and Graph Pt. 7

Shot-Assist Pass Origin Aggregate (all seasons) Heatmap



Goal-Assist Pass Origin Aggregate (all seasons) Heatmap

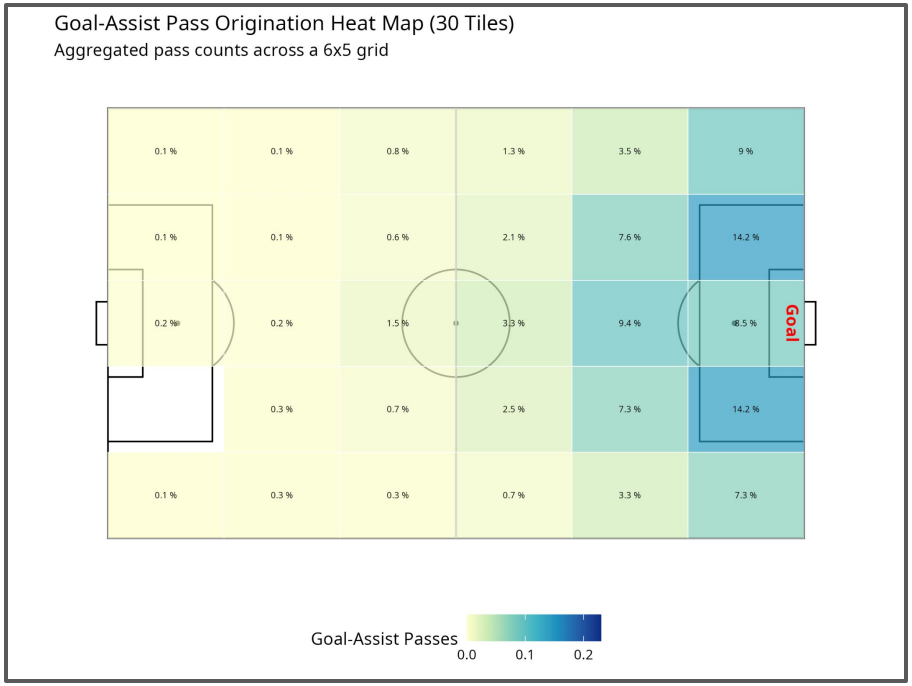
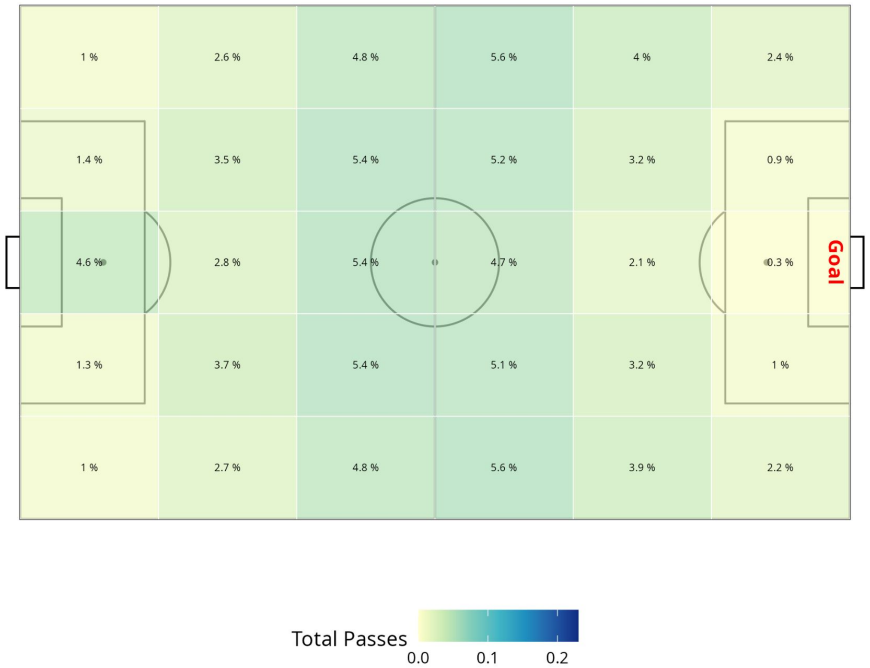


Chart and Graph Pt. 7

Total Pass Origination Heat Map (30 Tiles)
Aggregated pass counts across a 6x5 grid



Total Passes Origin Aggregate (all seasons) Heatmap La Liga

Conclusion

I avidly watch FC Barcelona play soccer, and while watching them, I noticed their tactics of short combination passing close to the goal to score. This contrasts with other teams' tactics of sending longer passes to try to score. The contrast made me wonder if soccer tactics are changing over time. Specifically, I wondered which of the strategies, short, combination passing close to goal or long balls in and around the box, was a more modern strategy. Wondering how this could translate to an experiment, I searched for data surrounding the location of passes. I found a website called Hudl Statsbomb that makes large amounts of event data (data about events in a soccer match) public and free.

I hypothesized that, if I used code to filter for the passes and their location, then passes close to the goal would be more common in more recent seasons.

Conclusion (cont.)

My scientific study confirmed my hypothesis. I used passing data from the Spanish professional soccer league, La Liga, over 15 seasons. First, I created a dot plot that shows that over time, a larger percent of the total passes were made close to goal (within 40 yards of the target goal line). I also created graphs to look at whether passes that lead to shots and goals have changed. I found that consistently around 80% of shots and goal assists are made close to the goal (within 40 yards of the target goal line). The first graph proves that teams are changing their play style to center around close-to-goal passing. The second and third graphs show why: close to goal passes are most effective to score.

I believe my study confirmed my hypothesis because as time goes on more teams in all sports are using statistics to influence their playstyle. Some teams probably even use data from Hudl Statsbomb. Moreover, if I, one person, could prove the effectiveness of close-to-goal passes, then groups of people who are paid to do stuff like this definitely could.

Practical Applications

In testing, I found that close-to-goal passing is an extremely effective way of creating shots and goals. My experiment shows that the best teams in the world (Spanish La Liga teams) are making more passes close to the goal as time goes on.

One practical application is for smaller soccer clubs, who may not have the resources necessary to focus on statistics. These clubs could use my findings to change their playstyle and hopefully have more success.

This experiment also shines light on how easy data science is becoming for sports, especially ones like soccer. Soccer has only recently started using advanced statistics, but now lots of soccer event data is becoming available, and AI makes it easier than ever to analyze that data. If a 13 year old who is just learning how to code could discover patterns in the data over a matter of months, then there is an undeniably high ceiling for data science in soccer.

Future Considerations

In my experiment I tested if there was any difference over time as to where most passes occurred. I found that as time went on, more and more passing play happened within 40 yards of the goal line in the attacking half. Consistently around 80% of assists came from within that same 40 yards, suggesting that close to goal passing had always been an extremely effective strategy for scoring goals. The fact that passing is only now changing in response means it took a lot of time to realize.

Data science in soccer has so much potential, and I only scratched the surface in my experiment. One way I could further investigate is to see if there is any difference in how play styles are changing between different top leagues, or international play versus club. There are already distinct and unique reputations for each of the top 5 leagues, such as the English Premier League being known for its physicality. My approach to soccer data coding and analysis could judge the integrity of these reputations, and influence both the tactics that teams play, as well which players teams sign. For example, an English team might want a more physical player, or to play more physically. I am very proud of the data science I did in this experiment, and I think there is limitless potential for data science in soccer.

References

1. https://github.com/statsbomb/open-data Hudl Statsbomb GitHub public data
2. https://positron.posit.co/ Positron
3. https://gemini.google.com/ Gemini AI