Mathematical modelling of football

Start again 11:15

David Sumpter

Uppsala University & Hammarby IF

FRIENDS OF TRACKING

UPPSALA UNIVERSITET

STIFTELSEN för STRATEGISK FORSKNING

Structure today

- Summary of last time. Key Performance Indices.
- Statistical models of passes
- Summary of expected goals (chance to raise questions)

Break 11-11:15

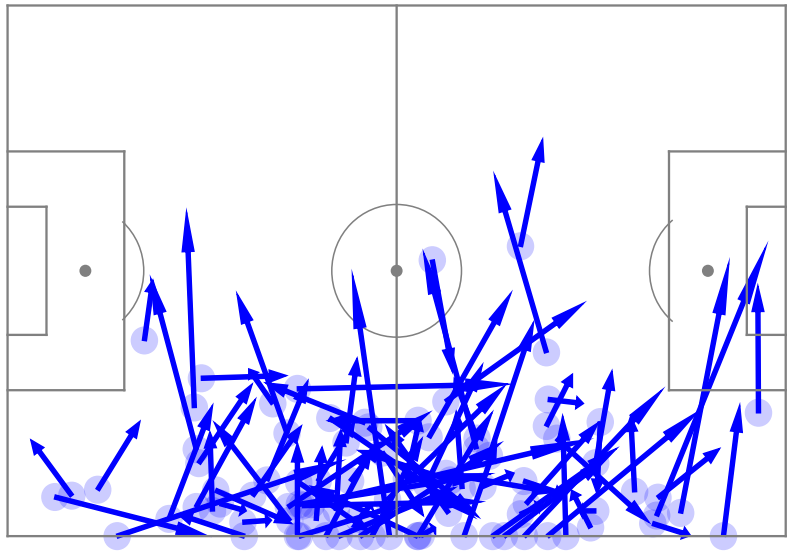
Please ask questions in the **chat** section where possible and I will answer them as I go.

I will take a reasonably slow pace and interact along the way.

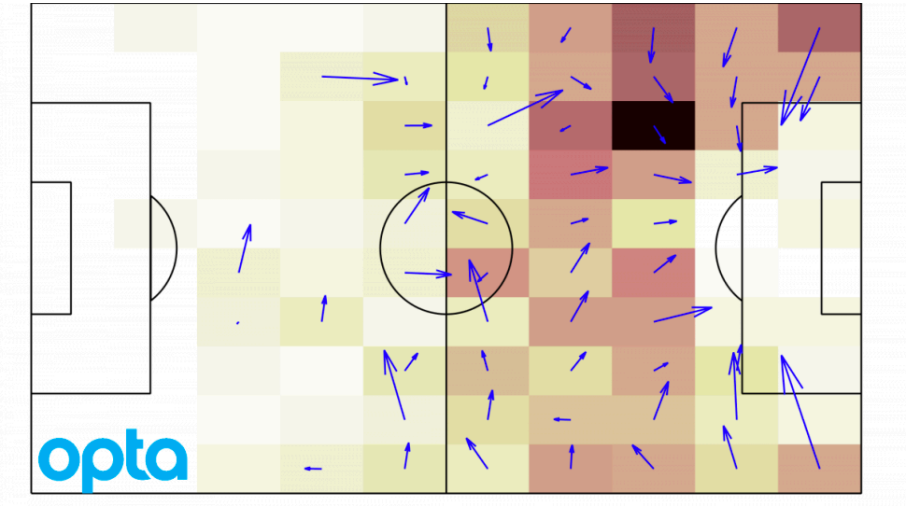
Go in to canvas and look where we are in the course…

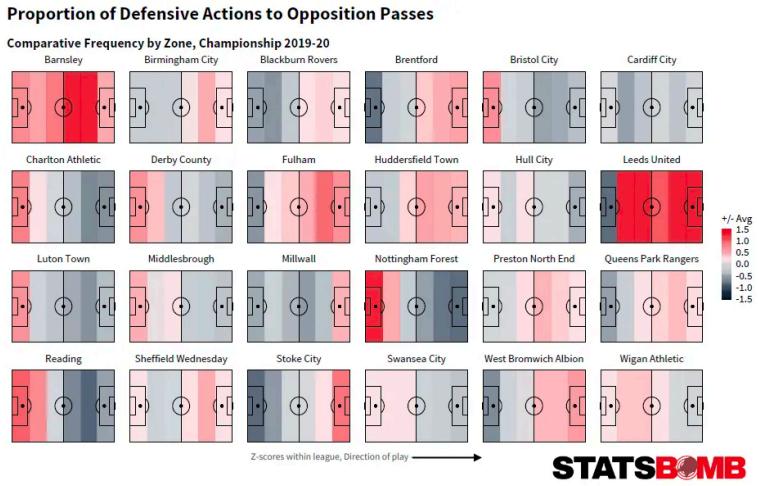
First a correction. Pass arrows.

```
#Find the passes
passes = df.loc[df['type_name'] == 'Pass'].set_index('id')
#Draw the pitch
(fig,ax) = createPitch(pitchLengthX,pitchWidthY,'yards','gray')
for i,thepass in passes.iterrows():
    #if thepass['team_name']==away_team_required: #
    if thepass['player_name']=='Lucy Bronze':
        x=thepass['location'][0]
        y=thepass['location'][1]
        passCircle=plt.Circle((x,pitchWidthY-y),2,color="blue")
        passCircle.set_alpha(.2)
        ax.add_patch(passCircle)
        dx=thepass['pass_end_location'][0]-x
        dy=thepass['pass_end_location'][1]-y

        passArrow=plt.Arrow(x,pitchWidthY-y,dx,-dy,width=3,color="blue")
        ax.add_patch(passArrow)

fig.set_size_inches(10, 7)
fig.savefig('Output/passes.pdf', dpi=100)
plt.show()
```



Summary from last time





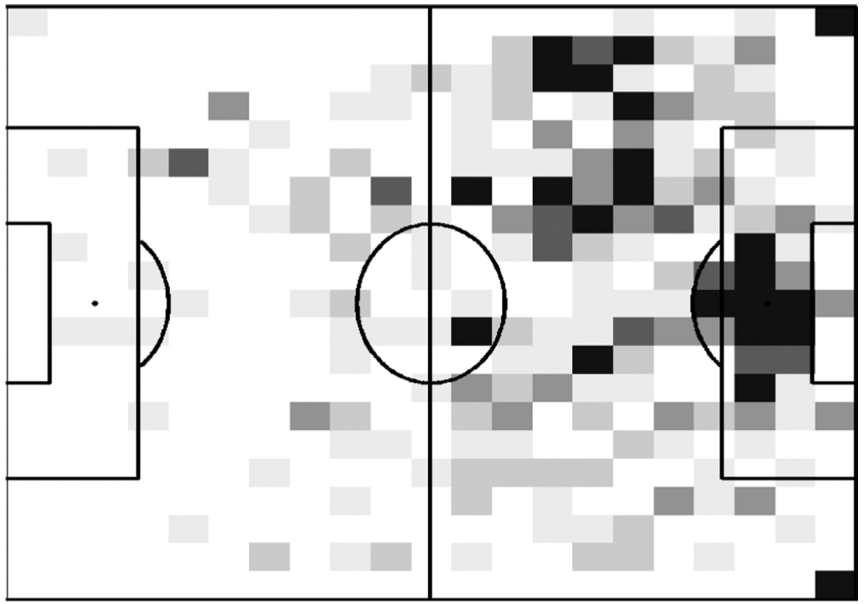


Figure 7.12 Real Madrid danger-zones during the Champions League season 2014/15. The shading is proportionally darker in areas where the ball was located during the 15 seconds leading up to a shot from the 20m by 20m area in front of the opposition goal. Data provided by Opta.

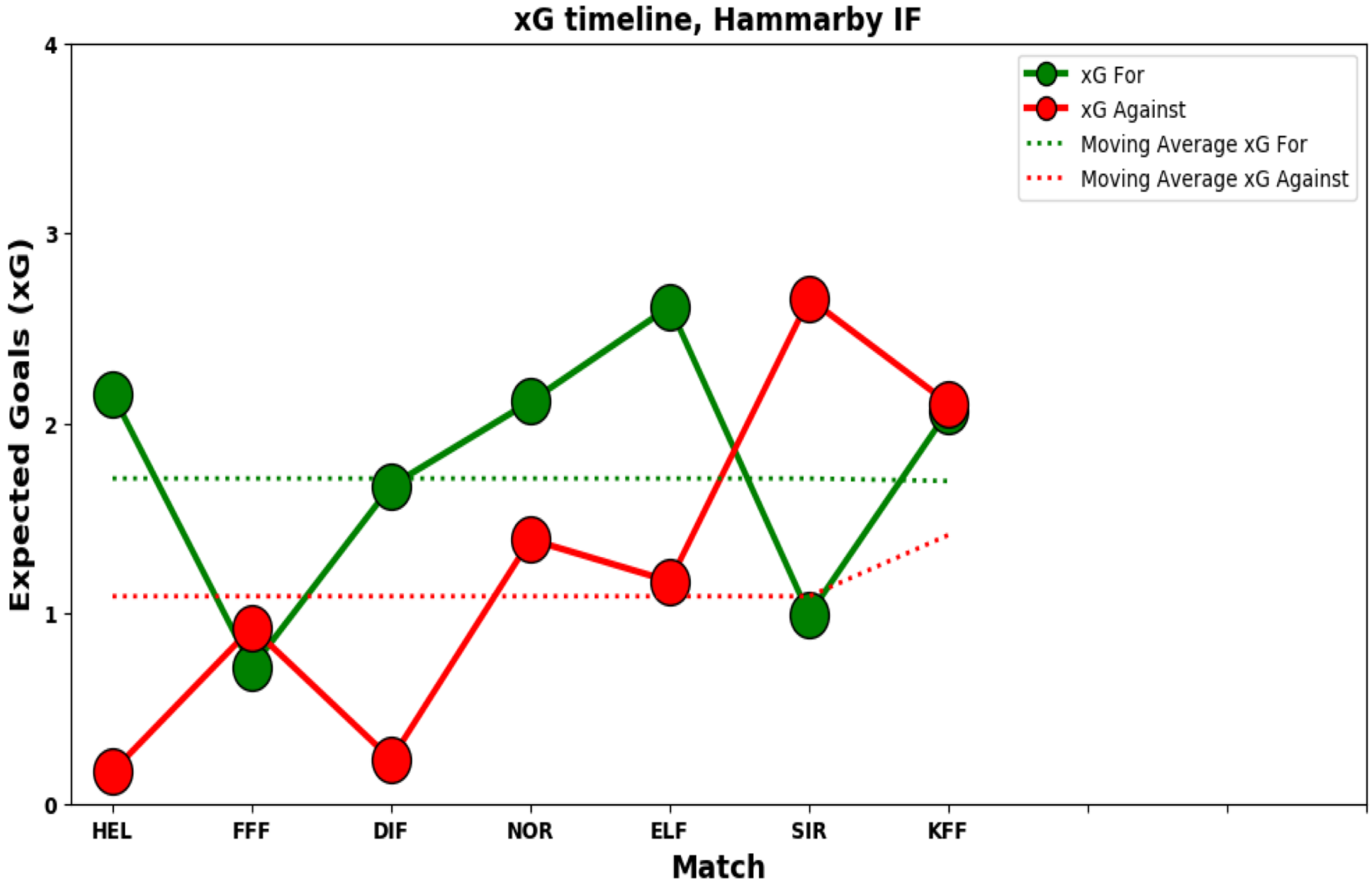
Summary from last time…

- Raw data is seldom enough.
- Standard visualisations are seldom enough (why we need Python!)
- What is the question? How do you answer it?
- Does the data support your hypothesis?
- Danger of self-confirmation, but risk of ignoring domain knowledge.
- Building measures that can then be used for future benchmarking (KPIs)
- Use both deductive (story) vs. inductive (data) thinking.

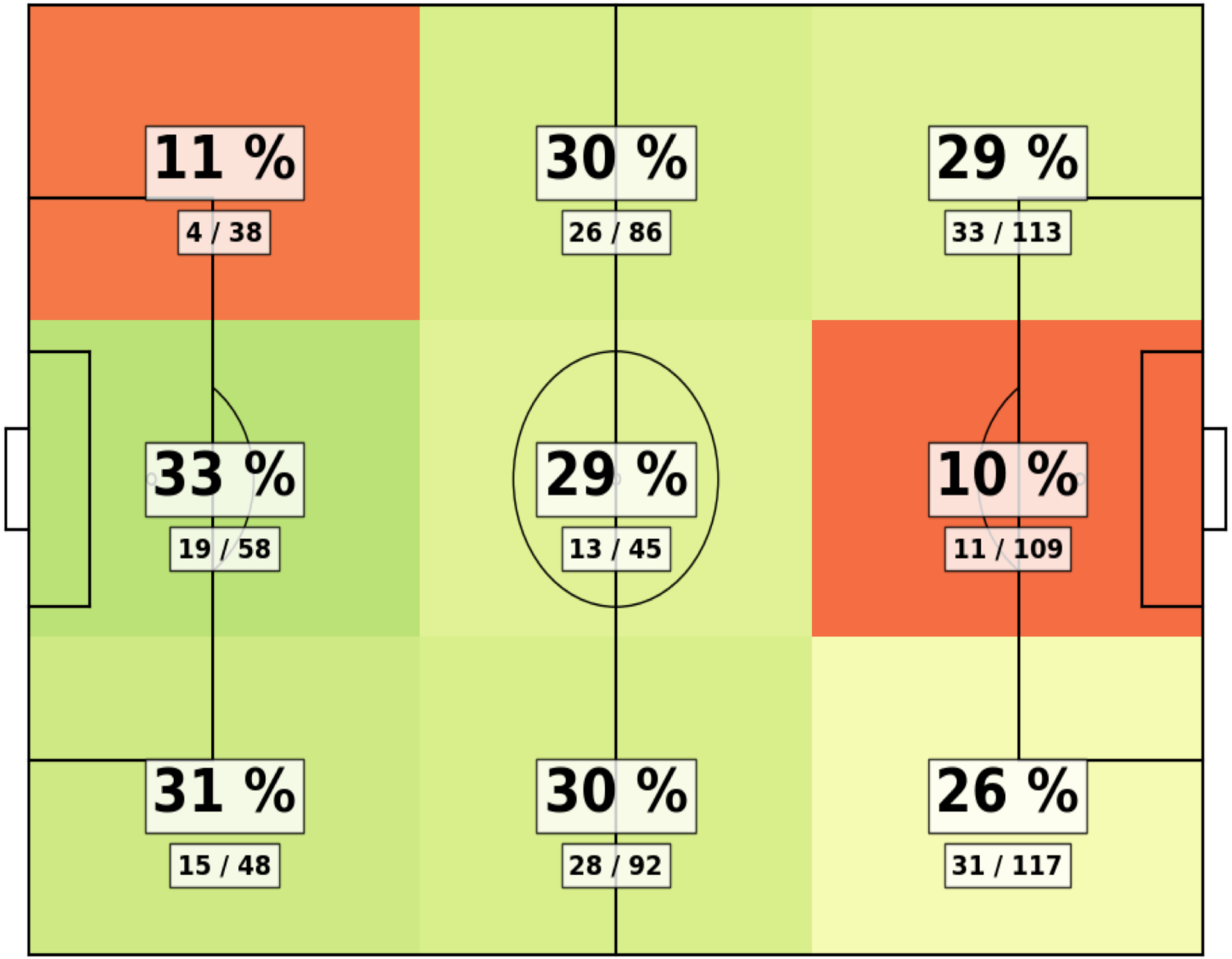
Key Performance indices at clubs

- Entries final third/box.
- Shots danger zone.
- Number of passes leading to shot.
- Ball recoveries within 5 seconds.
- Passing tempo.
- Expected goals.

Single set of measures that are used over the entire club.


xG timeline, Hammarby IF
xG For
xG Against
Moving Average xG For
Moving Average xG Against
Expected Goals (xG)
0
1
2
3
4
HEL
FFF
DIF
NOR
ELF
SIR
KFF
Match

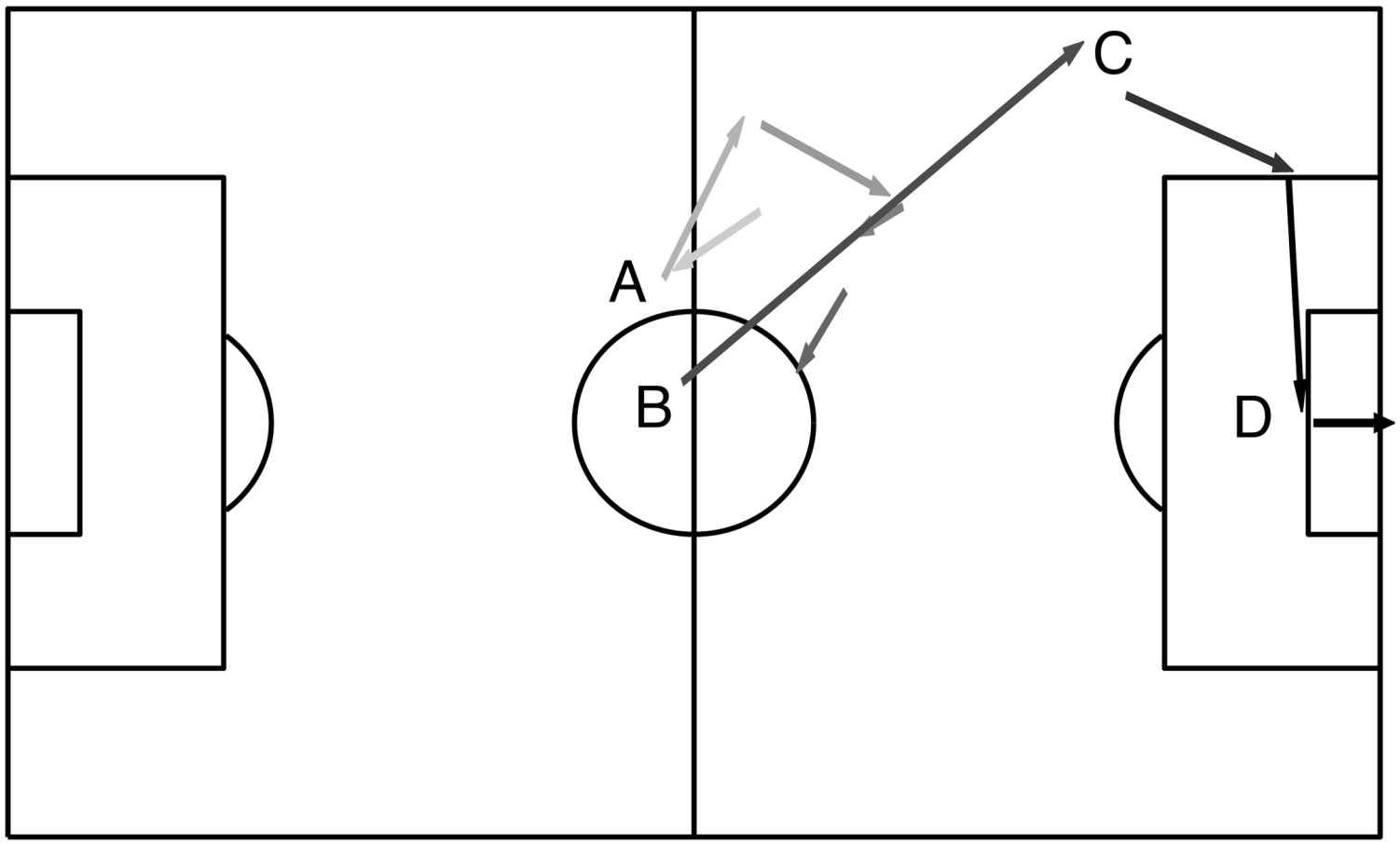
Ball recoveries within 5 seconds, Hammarby IF



Lecture 2
Statistics of Passing

Mathematical modelling of football

Pass sequence



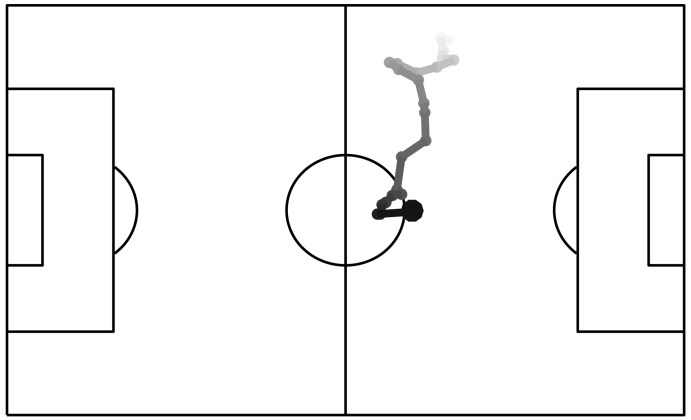
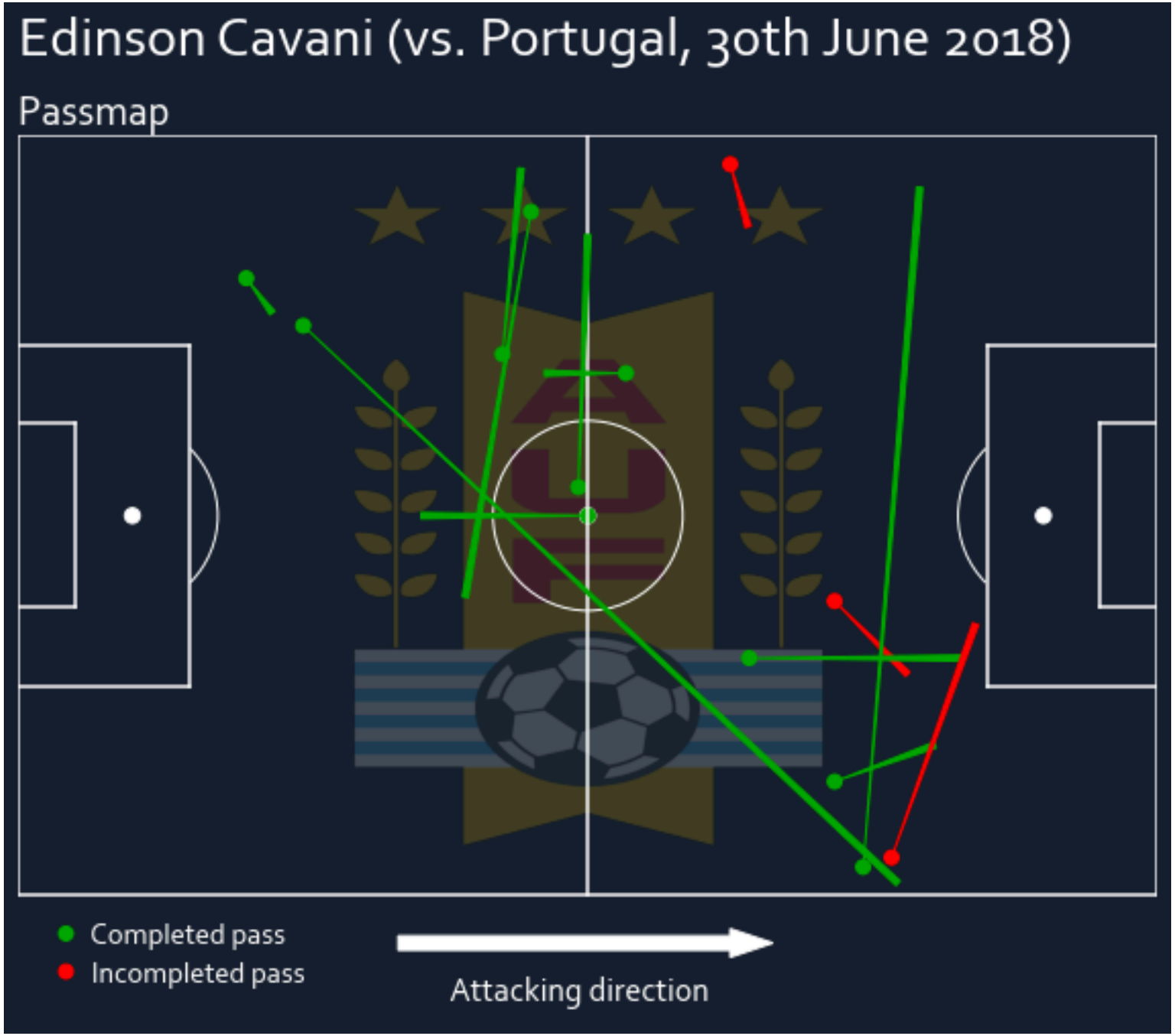
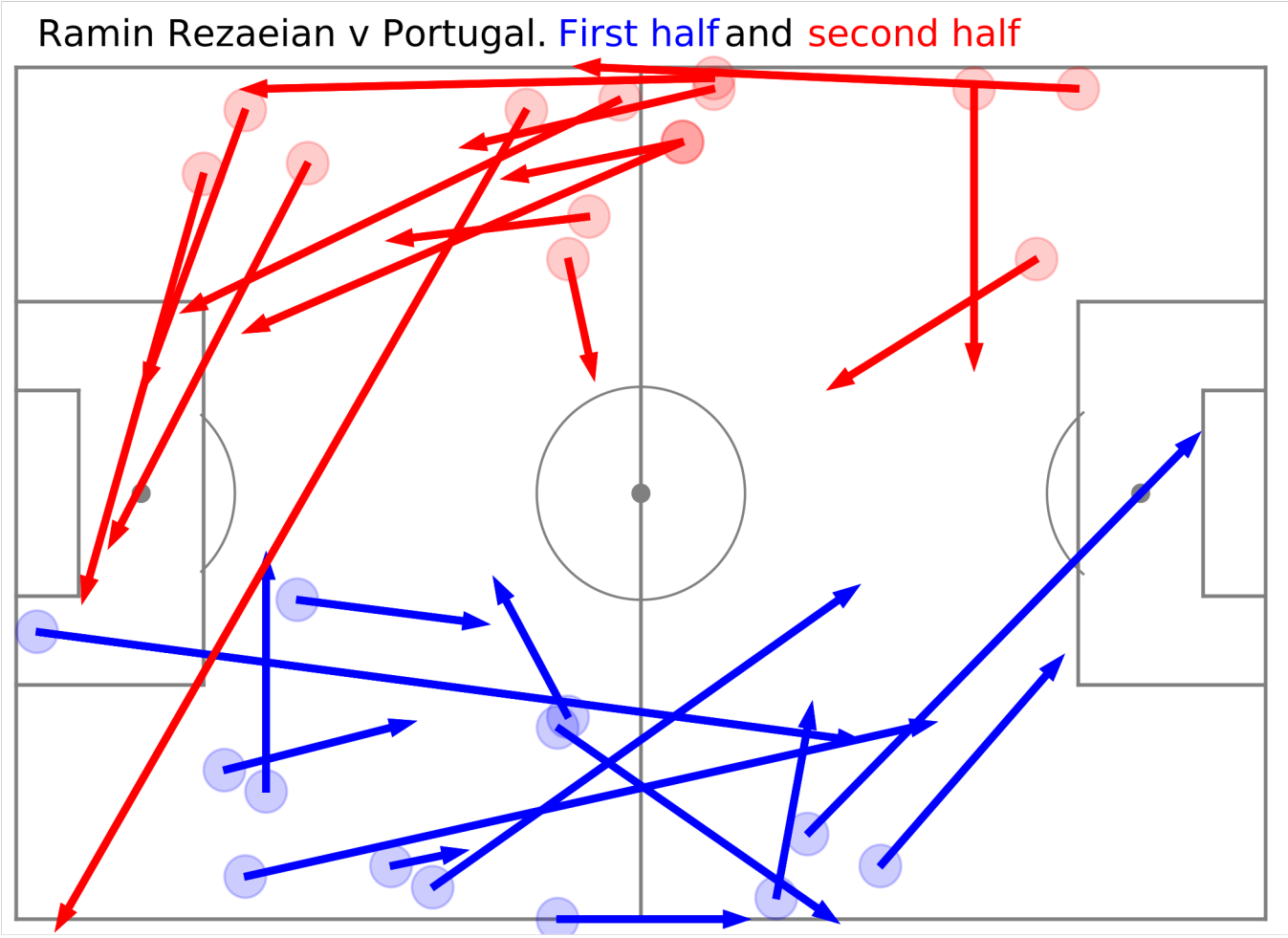


Figure 3.6 Passes leading up to Italy's first goal against Germany in Euro 2012, showing passes made by Italy (arrows in top panel) and Mesut Özil's movement while chasing the ball (bottom panel). Darker shading indicates more recent events in time. Letters indicate events: (A) Pirlo's first pass; (B) Pirlo's second pass; (C) Chiellini receives the ball; (D) Balotelli scores.



Diego Escribano
(Group 2)



Philip Winchester
(Group 2)

Mini-challenge: plotting actions

Philip Winchester

September 6, 2020

For this mini-challenge I decided to look at teams who played counter attacking football during the 2018 World Cup. As a true data scientists, we look at the data to decide which these teams were and the players who helped them do it.

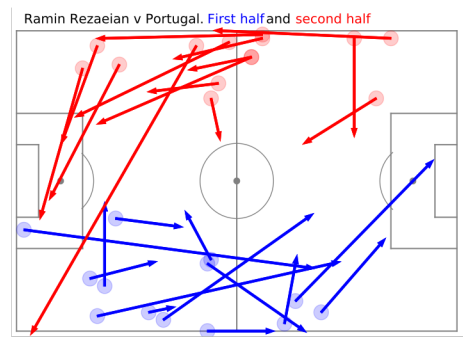
What is counter attacking football?
Good question. In general we are looking for teams who sit with a low defensive line and are quick on the break. A measure which looks at this for us is the following: The mean time taken for a possession which starts within the first third of the pitch to reach the final third. Basically, if a team wins the ball in their defensive third, how quickly can they get it to their striker? I calculated this mean quantity for all teams and the results for the top four teams are presented in the top of the two tables below. Interestingly, the host nation Russia seems to have been particularly strong on the break. Unsurprisingly, possession loving Spain took the longest time to progress into the final third with a mean time of 23.8s.

How to play counter attacking football?
Hmm, another one to think about! As soon as we win the ball back, I guess we want players who are going to hit long forward passes towards the opposition goal. In this instance, we are not interested in passes across the pitch to switch the side of play, so we only consider the mean passing distance (PD) in the x direction. In the bottom of the two tables below, I have listed the top four (outfield) players who had the longest mean pass in the x direction. I have not considered throw-ins, kick-offs and corners or players who placed less than 40 passes throughout the tournament. Iran who, according to my analysis, were the third fastest team on the break have two players on the list. In particular, it seems to be right back Ramin Rezaeian who contributed with a significant number of forward passes. This is especially noteworthy since the other players on the list are all centre backs, which seems to be a general trend.

Ramin Rezaeian: a counter attacking genius!
Ramin Rezaeian helped Iran be deadly on the break and they would have topped their group had they only snuck another goal past Portugal in what was a very tight final game. The figure below displays all of Ramins passes versus Portugal. As an attacking minded full-back, his style of play is very clear. Ramin will either look for a forward pass down the wing or a cross towards the opposite side of the box where Irans star striker Sardar Azmoun and 6 ft 3 Mehdi Taremi will be waiting. Ramin will often cross the ball early which is reminiscent of how Trent Alexander-Arnold plays for Liverpool, another team very strong on the break.

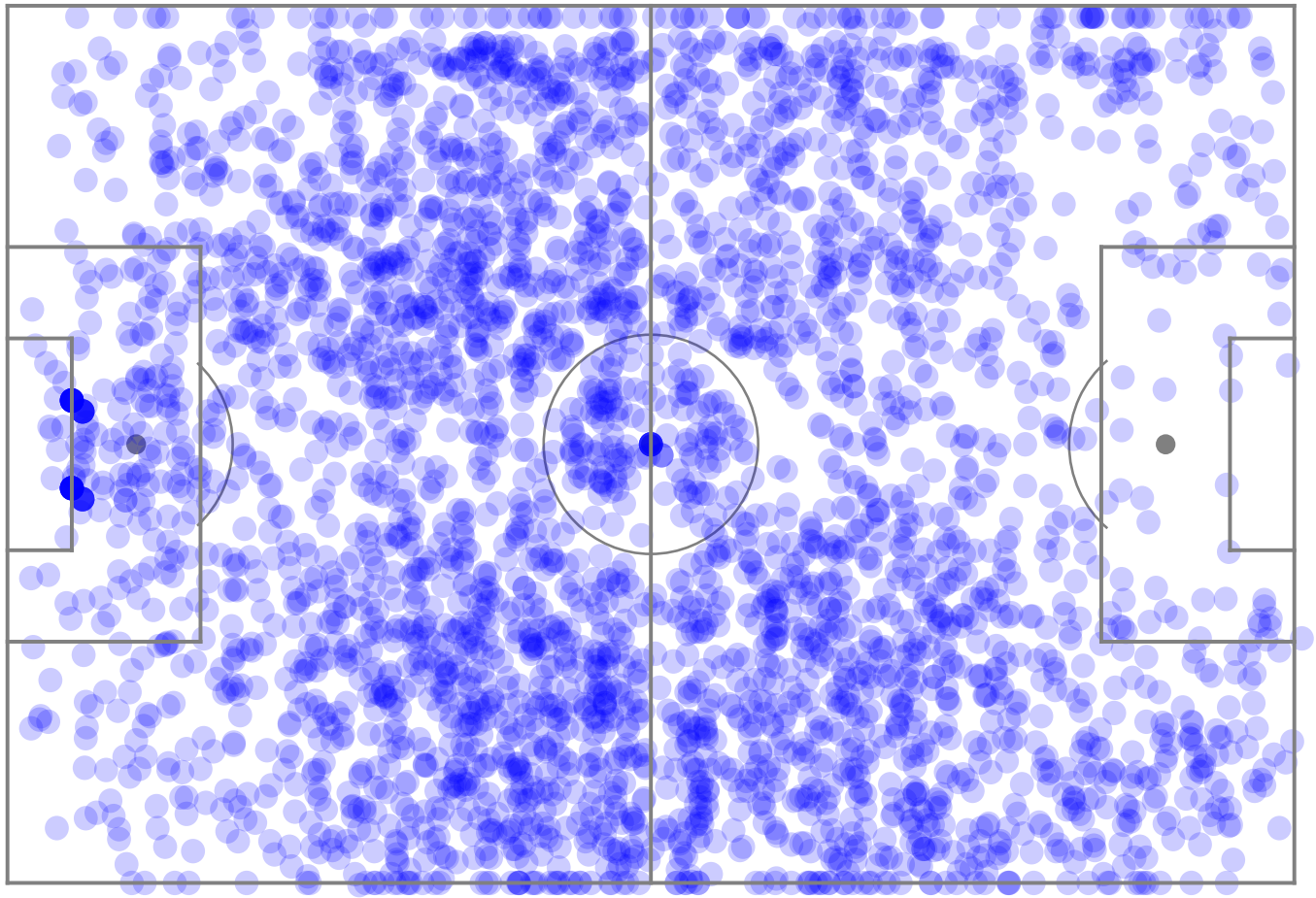
Team	Mean Time Taken (s)
Russia	9.1
Senegal	9.4
Iran	9.6
Iceland	10.2

Player	Team	Mean x PD (m)
R Rezaeian	Iran	18.8
K Árnason	Iceland	18.0
S Ignashevich	Russia	17.2
S Hosseini	Iran	15.9

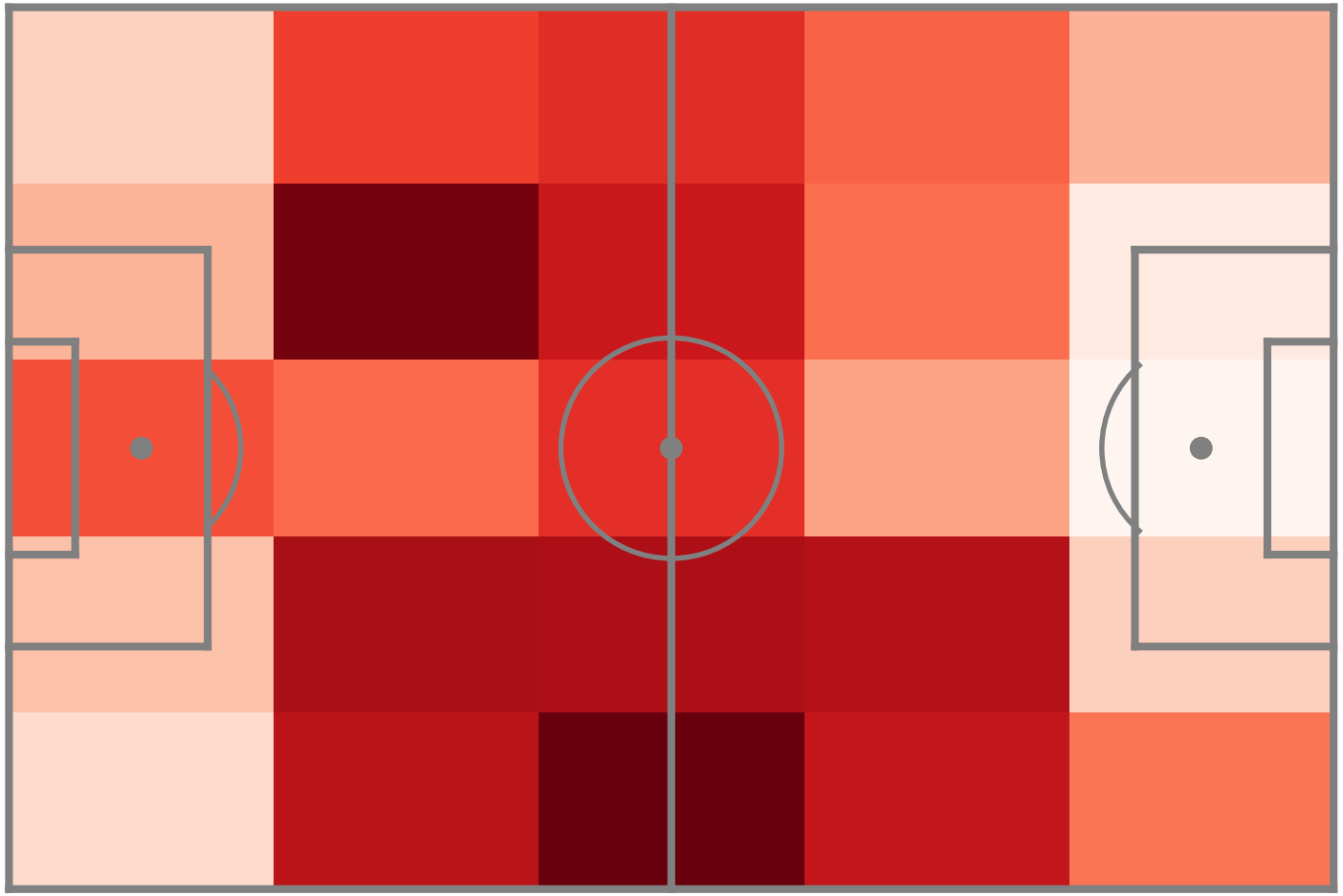


The next step is to go from visual understanding to statistical understanding…

All passes by England’s women in World Cup

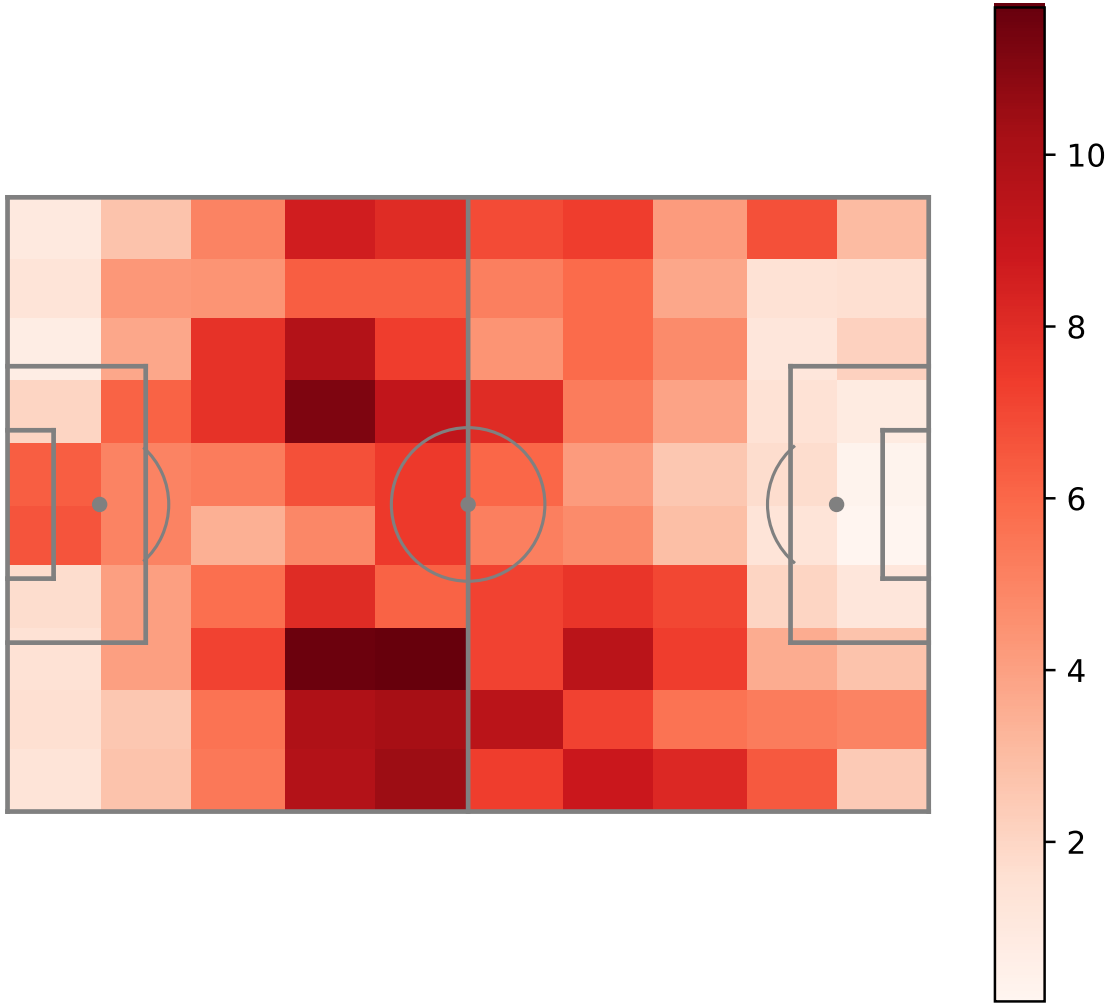


All passes by England’s women in World Cup



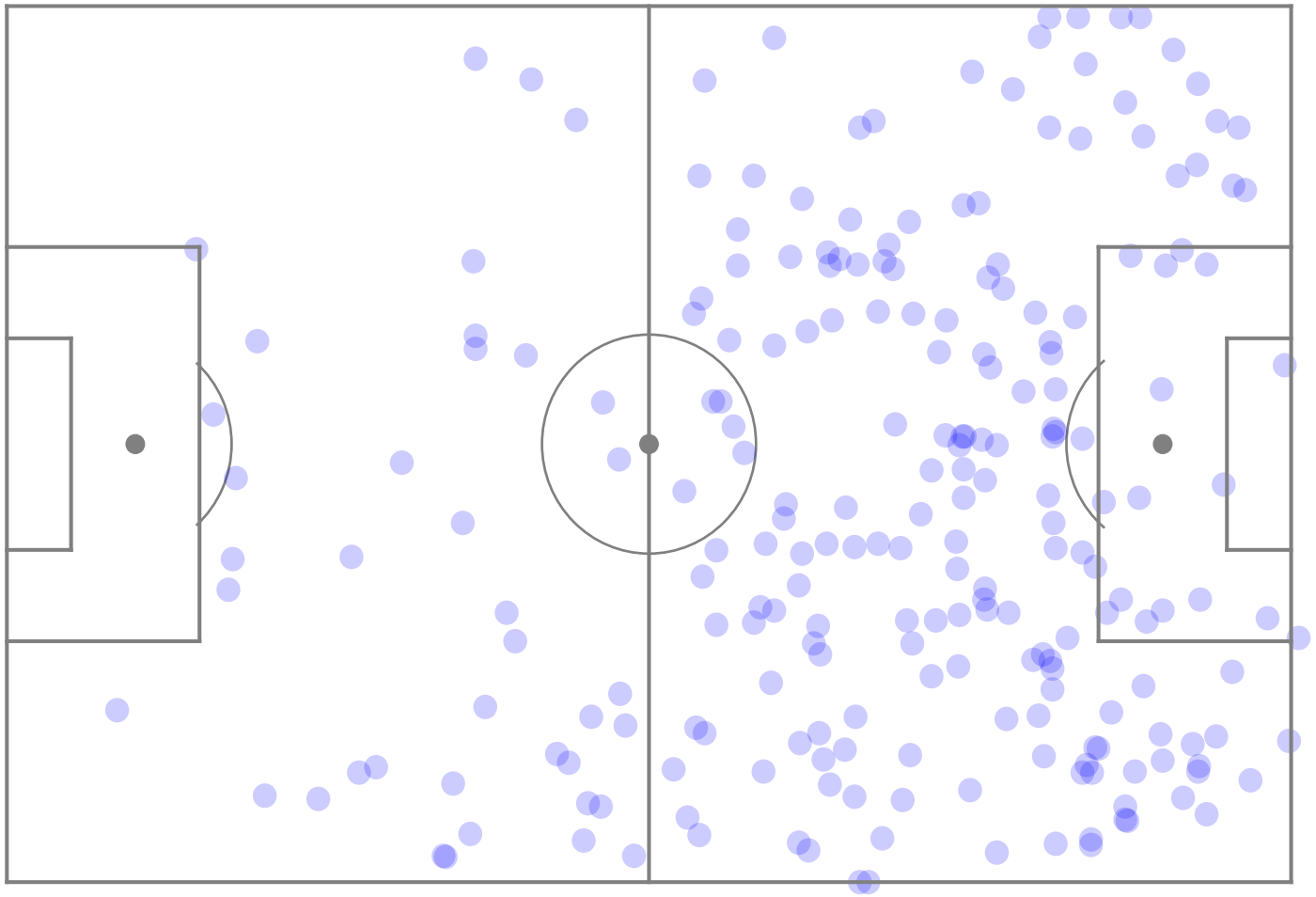
5 lanes
5 ‘heights’

All passes by England’s women in World Cup

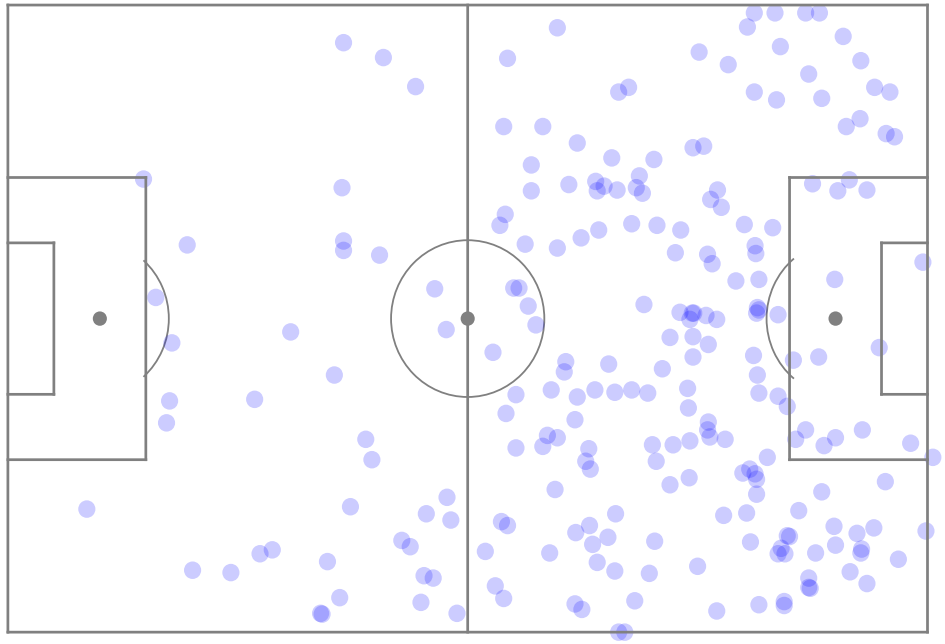


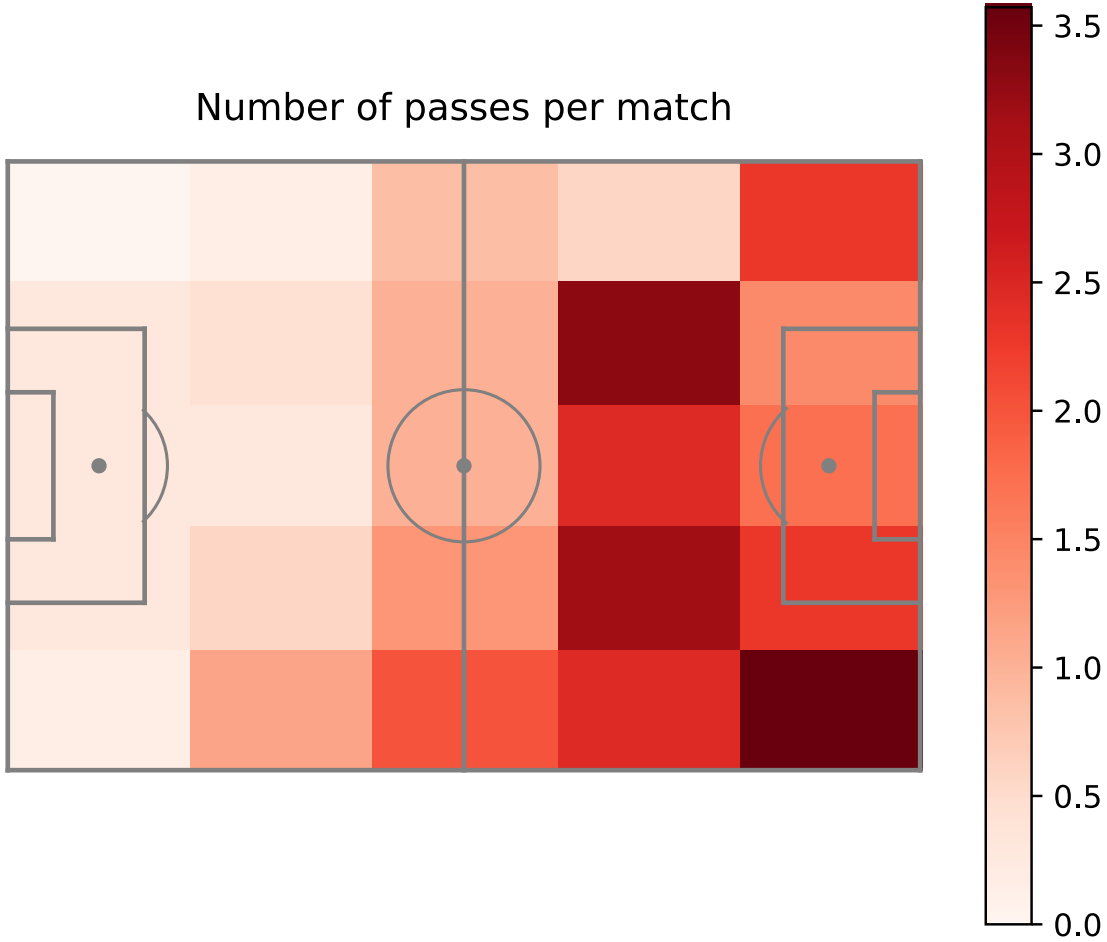
10 lanes
10 ‘heights’

Passes within 15 seconds of a shot



Passes within 15 seconds of a shot





Limitations

- Not adjusted for xG of chance created.
- For example, all chances with greater than 0.05 xG. (Challenge)
- Not compared to other teams.

Passes made by each player

Lucy Bronze	31
Francesca Kirby	26
Jill Scott	23
Nikita Parris	22
Keira Walsh	16
Stephanie Houghton	16
Rachel Daly	13
Toni Duggan	13
Ellen White	12
Alex Greenwood	11
Abbie McManus	10
Bethany Mead	10
Demi Stokes	9
Jodie Taylor	4
Millie Bright	4
Jade Moore	4
Georgia Stanway	3
Karen Julia Carney	3
Karen Bardsley	1

Limitations

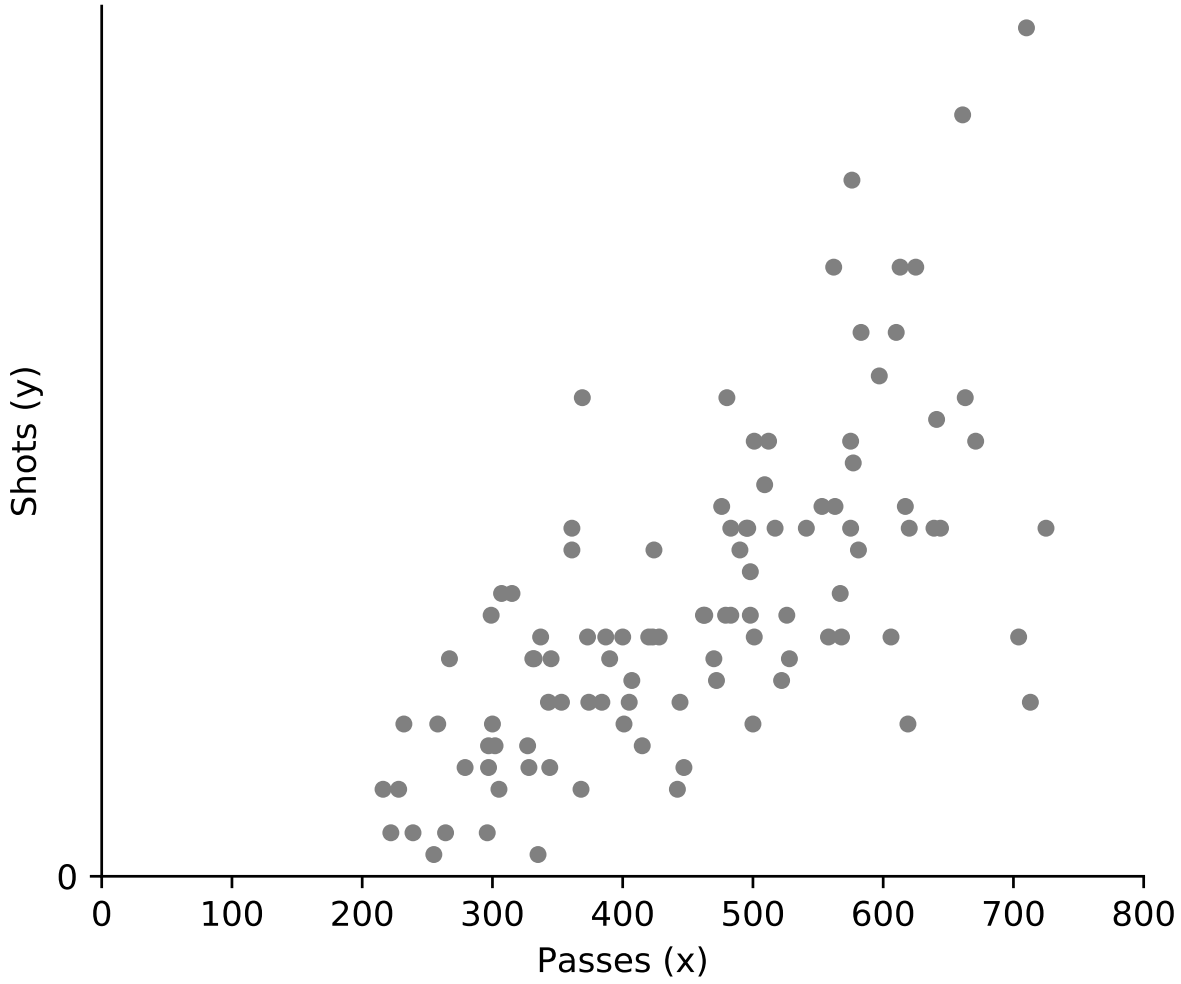
- Not corrected for minutes played.
- Again, could be corrected for xG.

Lets look at the code…

7PassHeatMap.py

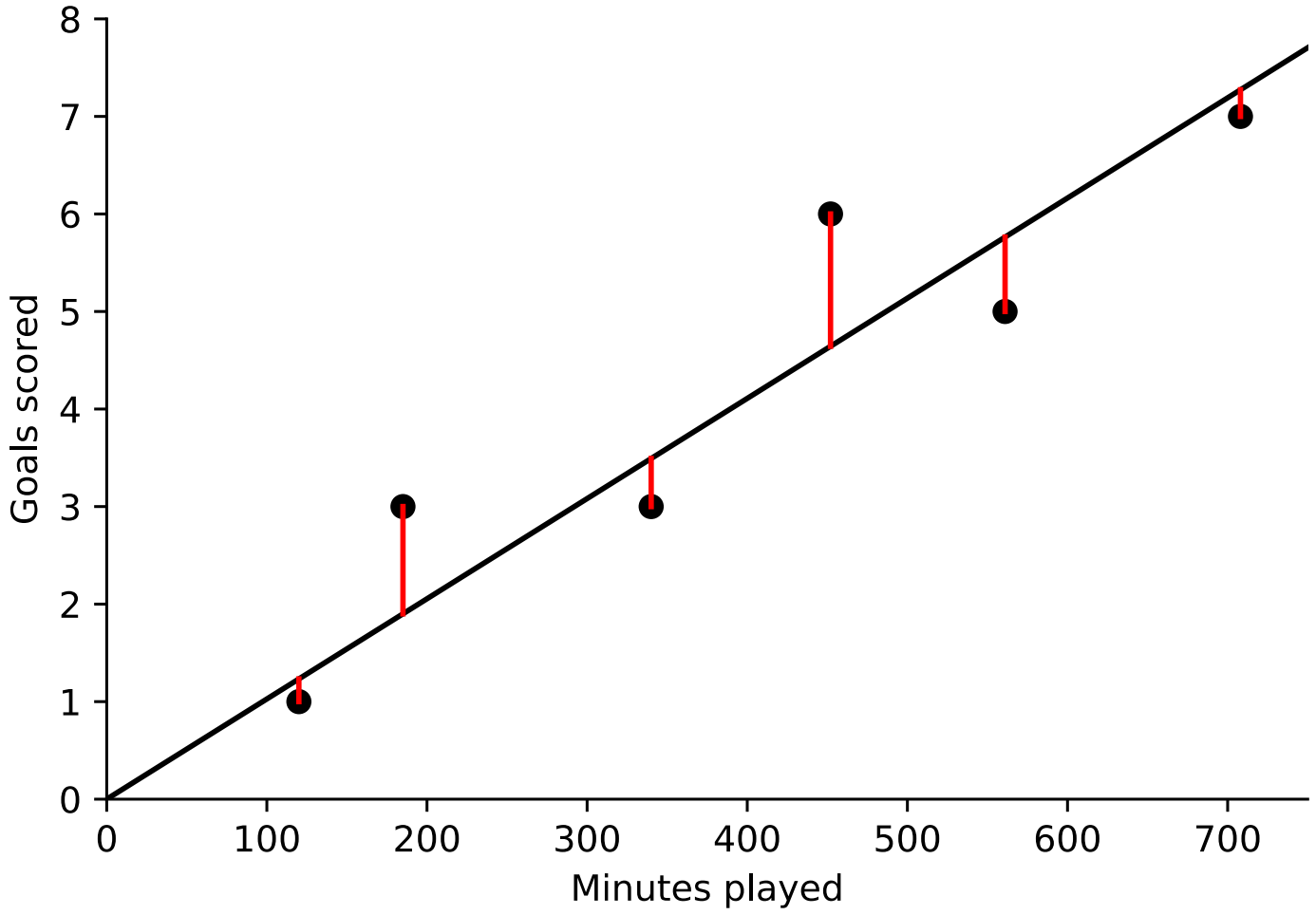
Still don’t have statistical understanding until we compare to other teams…

Do teams that pass more shot more?



All teams in the
Women's World Cup

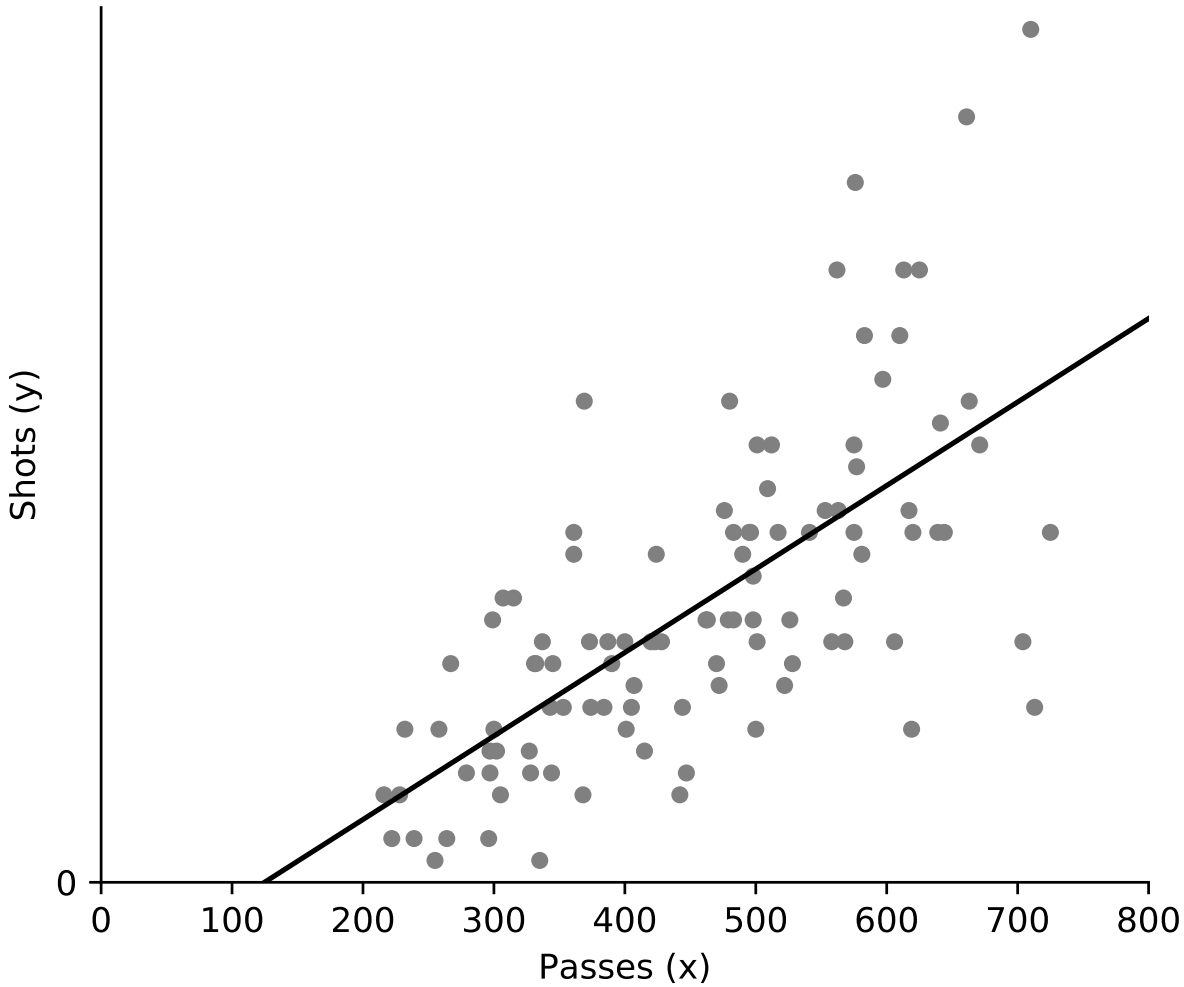
Linear regression



Minimize sum of distance between points and line.

Equivalently, minimize sum of squares of differences.

Do teams that pass more shot more?



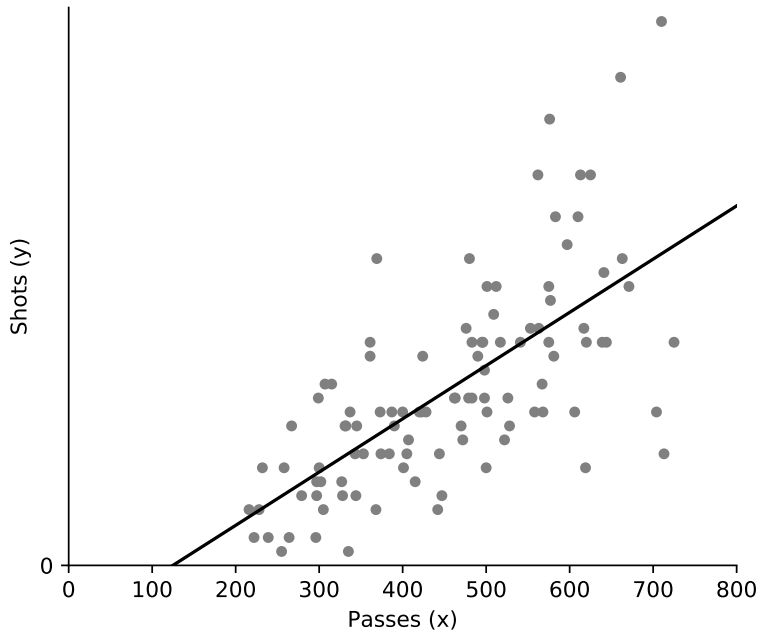
Fitting in python

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Shots   R-squared:                       0.467
Model:                            OLS   Adj. R-squared:                  0.461
Method:                 Least Squares   F-statistic:                     89.21
Date:                Mon, 07 Sep 2020   Prob (F-statistic):           1.38e-15
Time:                        20:36:52   Log-Likelihood:                -321.84
No. Observations:                 104   AIC:                             647.7
Df Residuals:                     102   BIC:                             653.0
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -4.7636      1.916     -2.486      0.015      -8.565      -0.962
Passes         0.0382      0.004      9.445      0.000       0.030       0.046
==============================================================================
Omnibus:                        9.739   Durbin-Watson:                   2.364
Prob(Omnibus):                  0.008   Jarque-Bera (JB):               10.996
Skew:                           0.553   Prob(JB):                      0.00410
Kurtosis:                       4.147   Cond. No.                     1.72e+03
==============================================================================
```

b_0

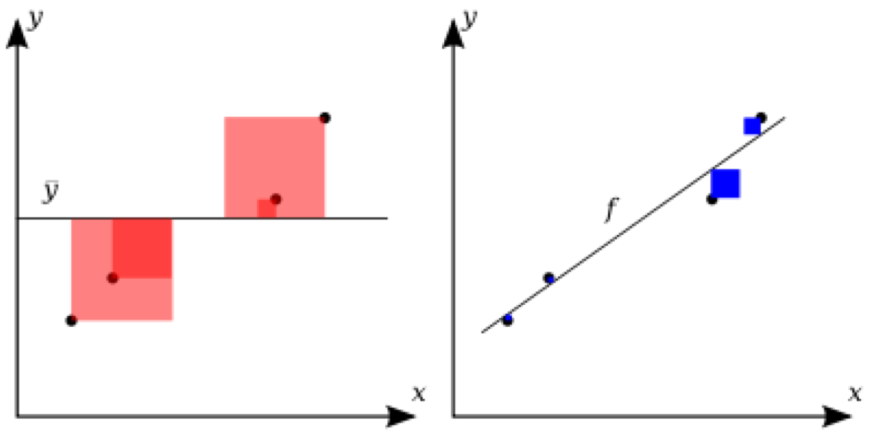
b_1

$$\text{Goals} = b_0 + b_1\text{Passes}$$



Fitting in python

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Shots   R-squared:                       0.467
Model:                            OLS   Adj. R-squared:                  0.461
Method:                 Least Squares   F-statistic:                     89.21
Date:                Mon, 07 Sep 2020   Prob (F-statistic):           1.38e-15
Time:                        20:36:52   Log-Likelihood:                -321.84
No. Observations:                 104   AIC:                             647.7
Df Residuals:                     102   BIC:                             653.0
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -4.7636      1.916     -2.486      0.015      -8.565      -0.962
Passes         0.0382      0.004      9.445      0.000       0.030       0.046
==============================================================================
Omnibus:                        9.739   Durbin-Watson:                   2.364
Prob(Omnibus):                  0.008   Jarque-Bera (JB):               10.996
Skew:                           0.553   Prob(JB):                      0.00410
Kurtosis:                       4.147   Cond. No.                     1.72e+03
==============================================================================
```



$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

Test goodness of fit

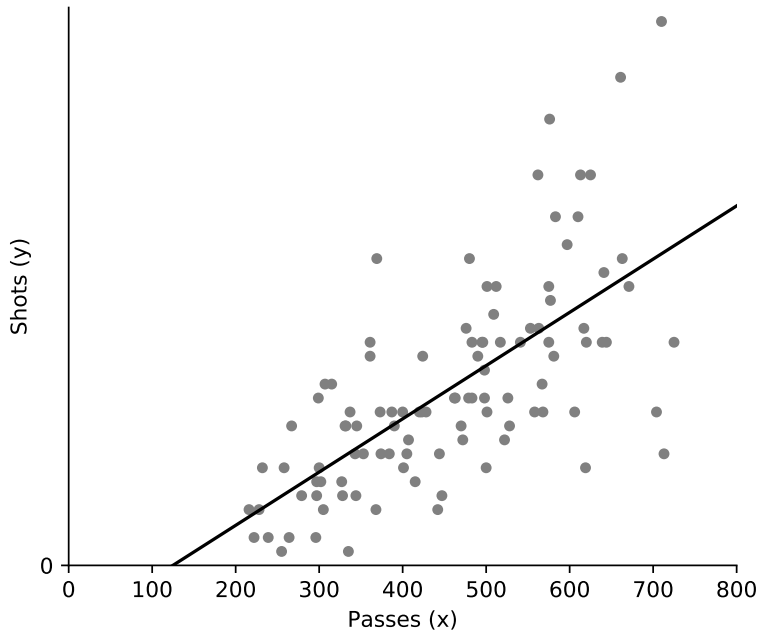
Fitting in python

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Shots   R-squared:                       0.467
Model:                            OLS   Adj. R-squared:                  0.461
Method:                 Least Squares   F-statistic:                     89.21
Date:                Mon, 07 Sep 2020   Prob (F-statistic):           1.38e-15
Time:                        20:36:52   Log-Likelihood:                -321.84
No. Observations:                 104   AIC:                             647.7
Df Residuals:                     102   BIC:                             653.0
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -4.7636      1.916     -2.486      0.015      -8.565      -0.962
Passes         0.0382      0.004      9.445      0.000       0.030       0.046
==============================================================================
Omnibus:                        9.739   Durbin-Watson:                   2.364
Prob(Omnibus):                  0.008   Jarque-Bera (JB):               10.996
Skew:                           0.553   Prob(JB):                      0.00410
Kurtosis:                       4.147   Cond. No.                     1.72e+03
==============================================================================
```

b_0

b_1

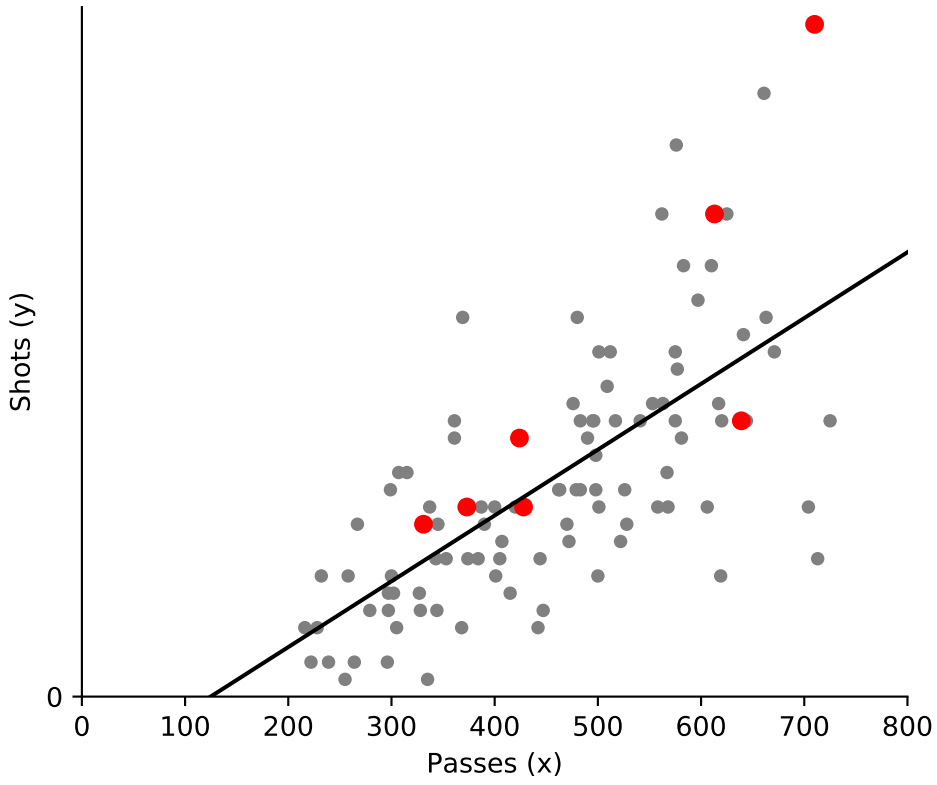
Goals = $b_0 + b_1$Passes



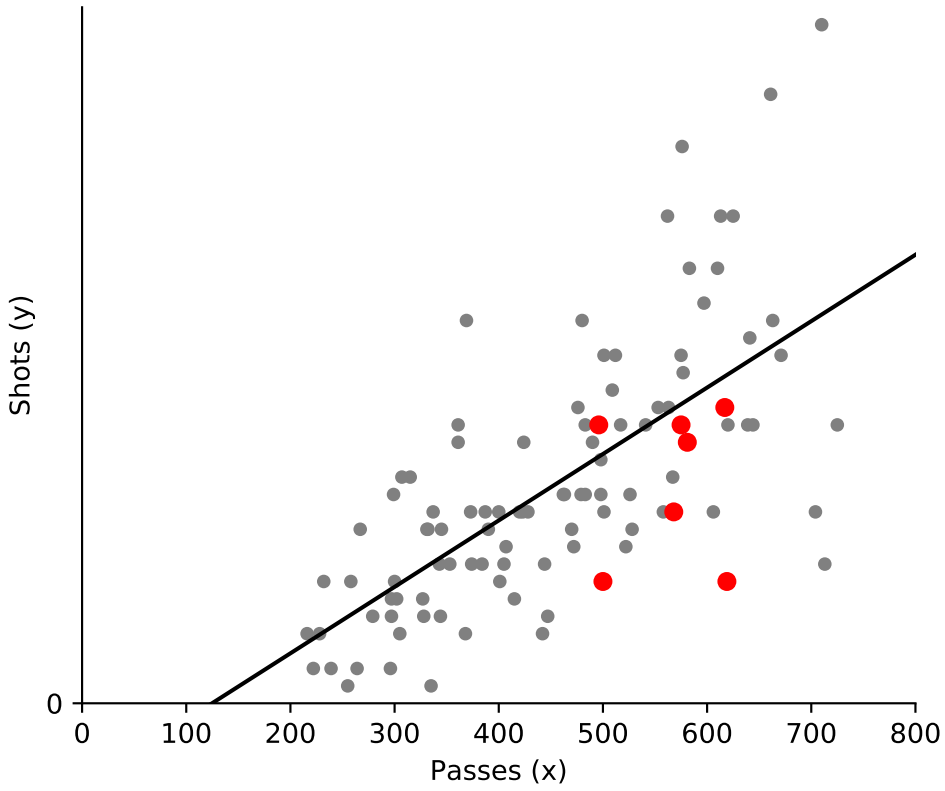
Test null hypothesis that intercept (b_0) and slope (b_1) are zero

Can we learn something about individual teams?

USA

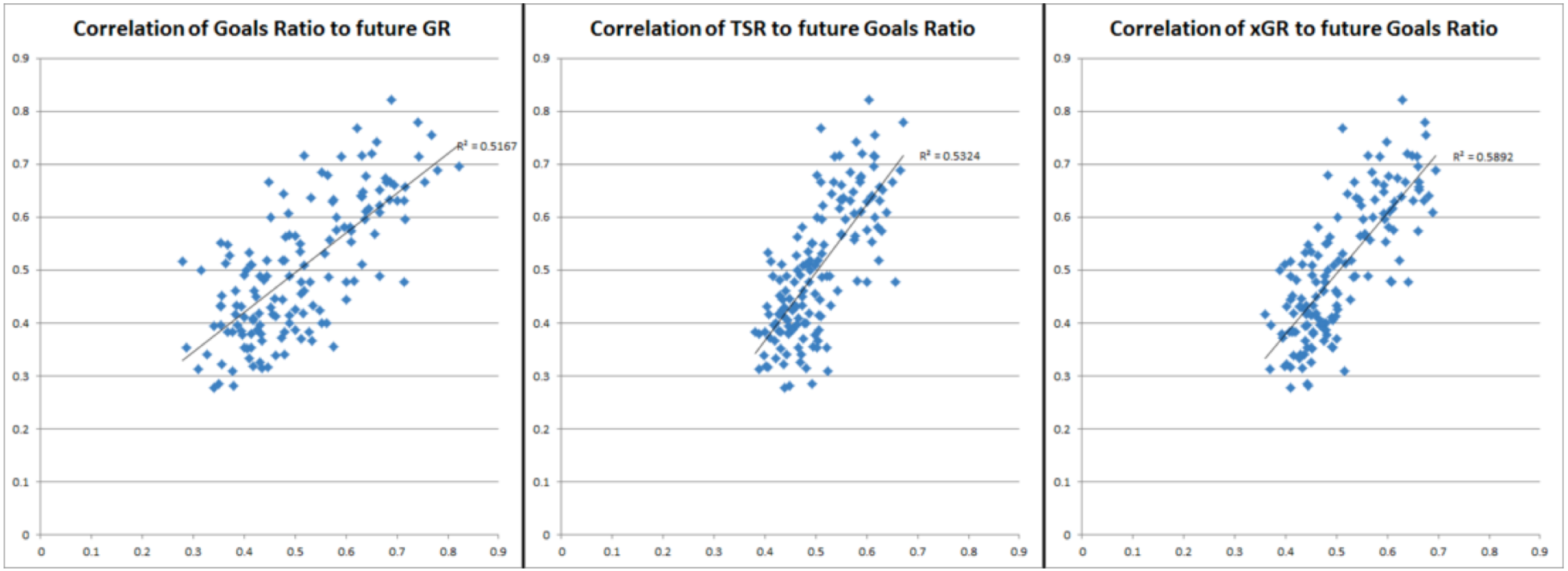


England



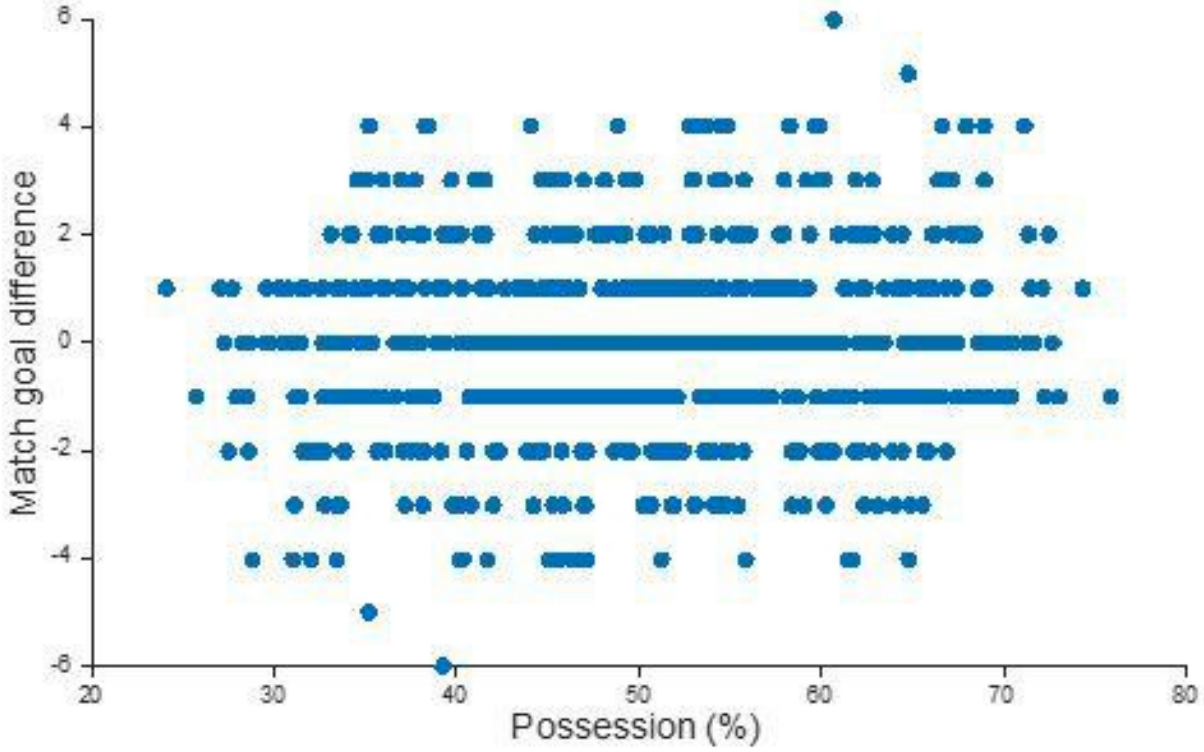
Linear regression is a quick way of checking relationships in data

They predict future goals better than goals.



https://cartilagefreecaptain.sbnation.com/2014/2/28/5452786/shot-matrix-tottenham-hotspur-stats-analysis-expected-goals

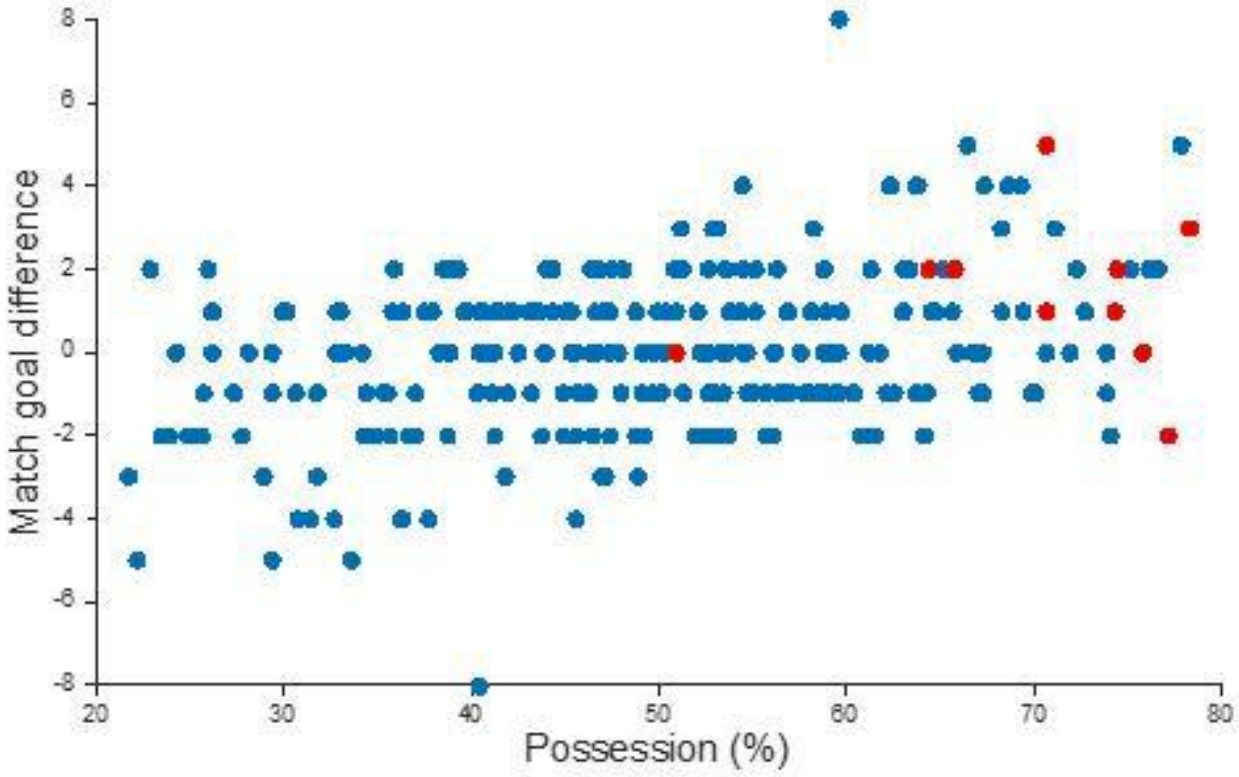
Possession and goal difference



Premier League 16/17

https://medium.com/@Soccermatics/how-important-is-it-to-have-the-ball-47f93b7760fd

Possession and goal difference

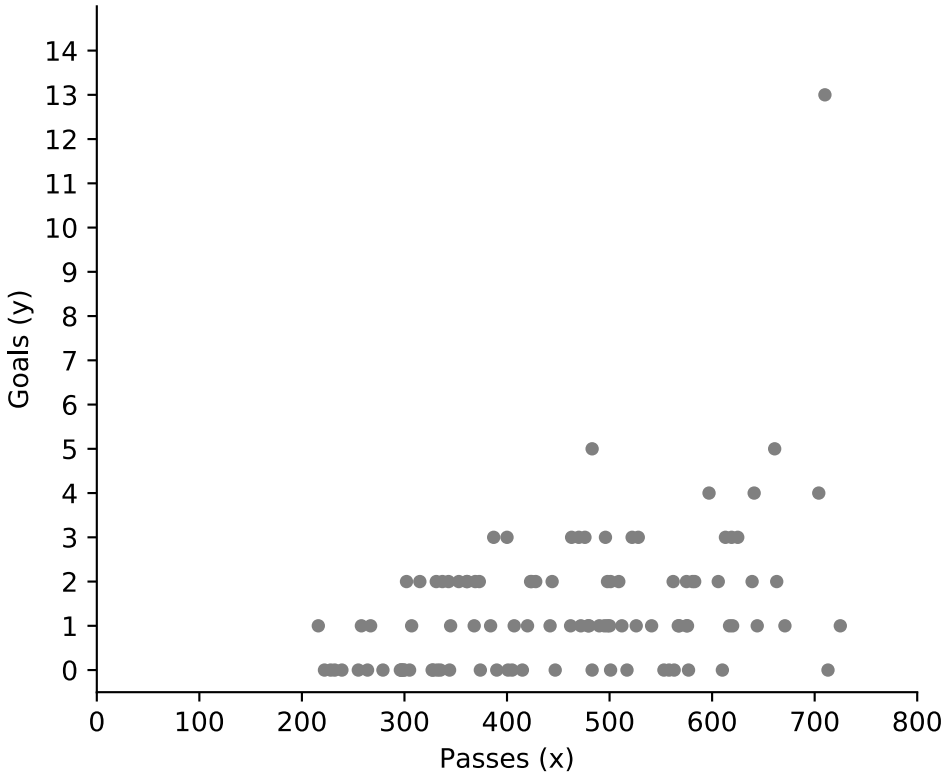


Champions League 16/17

https://medium.com/@Soccermatics/how-important-is-it-to-have-the-ball-47f93b7760fd

Possession and winning are not usually correlated!

For goals we should us Poisson regression (next lecture)



No evidence to dismiss the null hypothesis

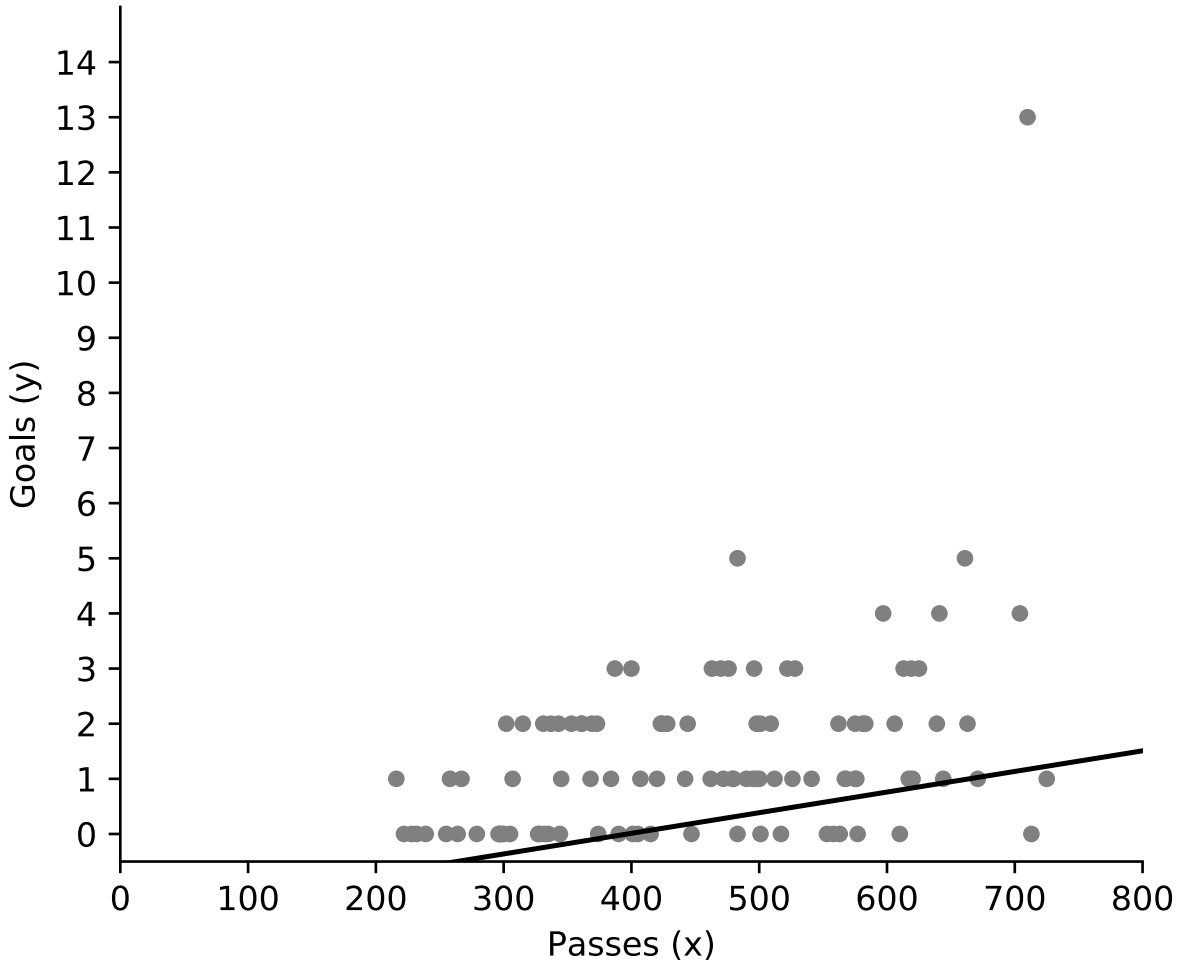
Poisson regression fit.

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                  Goals   No. Observations:                  104
Model:                            GLM   Df Residuals:                      102
Model Family:                 Poisson   Df Model:                            1
Link Function:                    log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -157.54
Date:                Mon, 07 Sep 2020   Deviance:                       138.26
Time:                        21:12:36   Pearson chi2:                     129.
No. Iterations:                     5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -1.4845      0.350     -4.237      0.000      -2.171      -0.798
Passes         0.0037      0.001      5.697      0.000       0.002       0.005
==============================================================================
"""
```

b_0

b_1

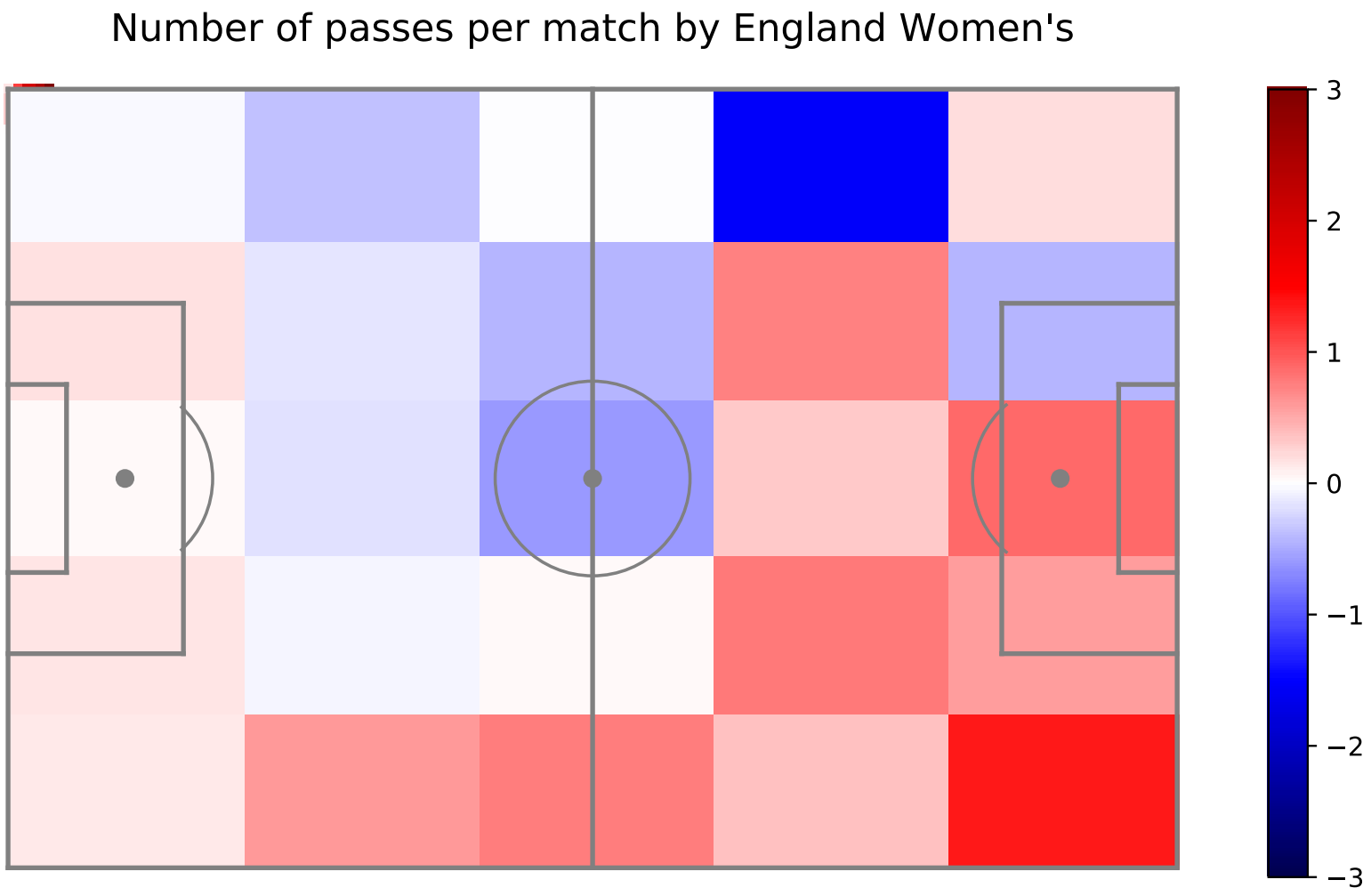
Poisson regression fit.



Look in code

8PassCompare.py

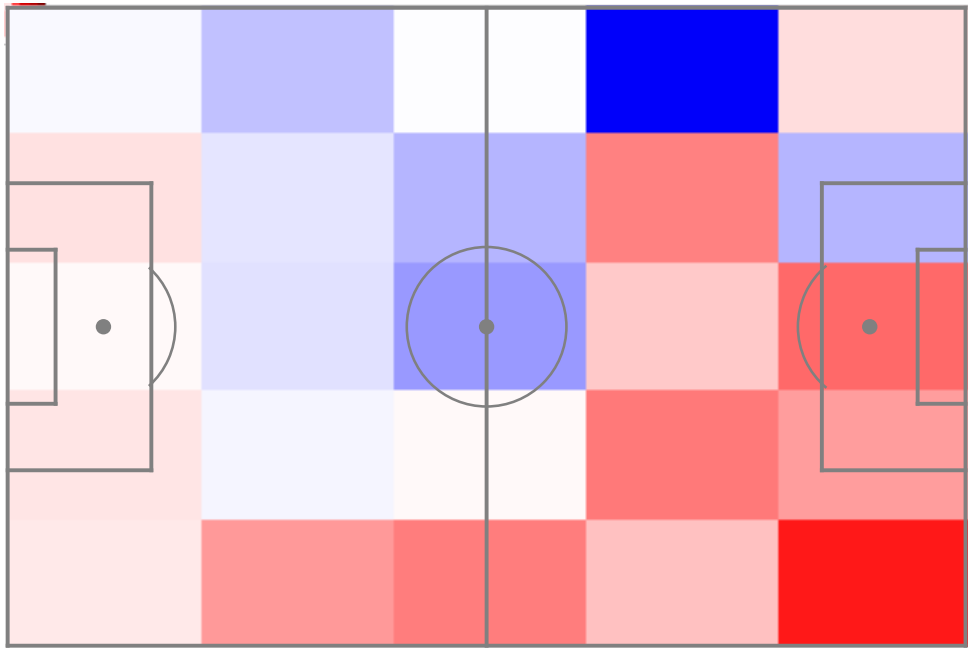
Difference to average of all teams

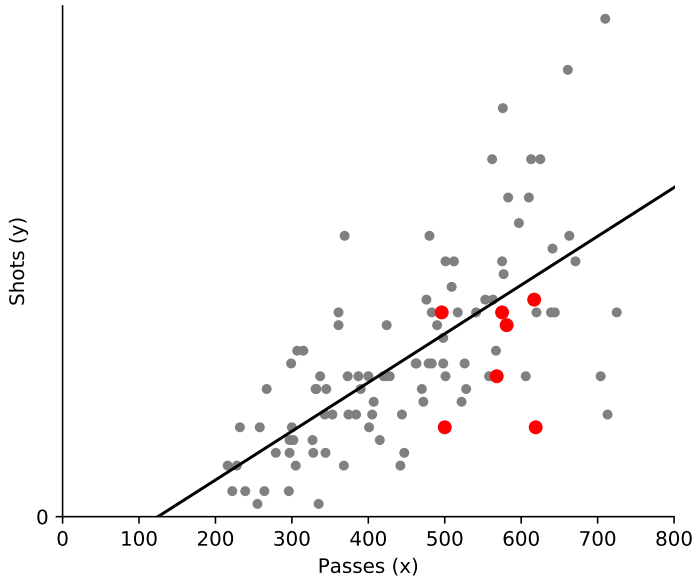


What would I tell England (or Sweden) about their World Cup?

Based on 3 or 4 hours of data analysis…

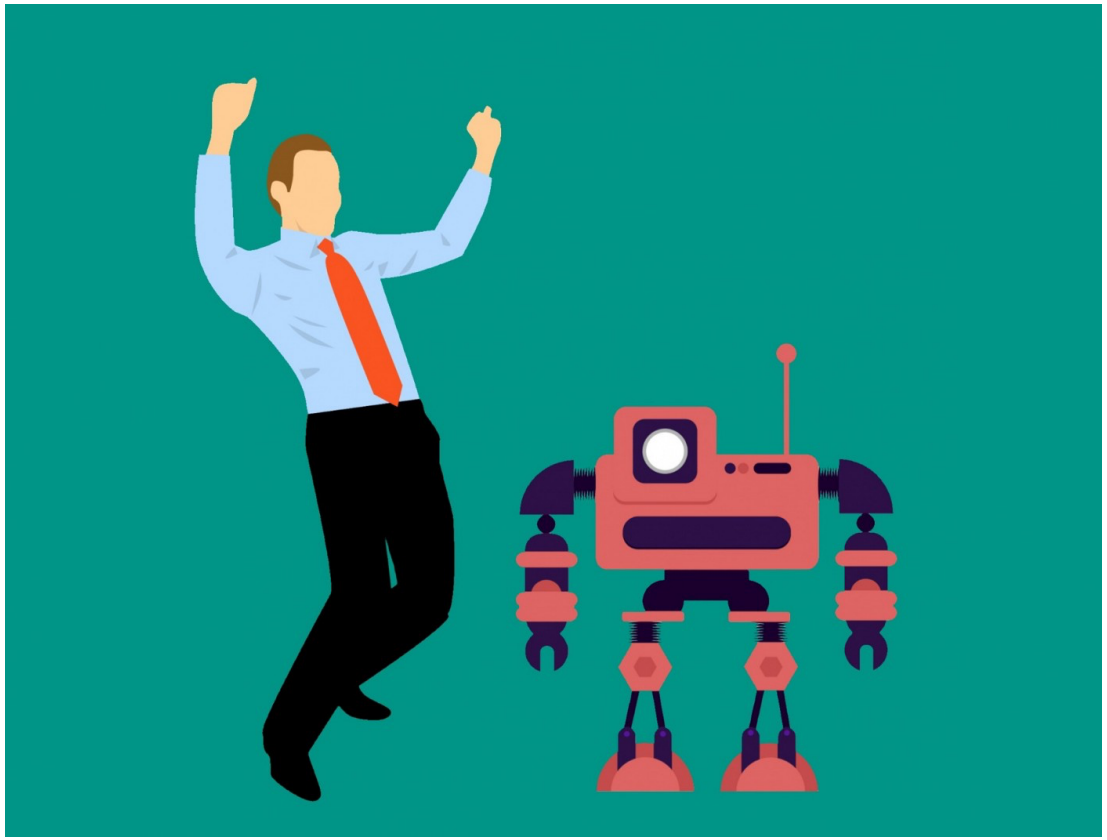
Number of passes per match by England Women's





Lucy Bronze	31
Francesca Kirby	26
Jill Scott	23
Nikita Parris	22
Keira Walsh	16
Stephanie Houghton	16
Rachel Daly	13
Toni Duggan	13
Ellen White	12

Data Alchemy



https://medium.com/@Soccermatics/why-algorithms-are-no-better-than-humans-at-predicting-exam-results-goals-in-football-musical-f5650aeb1cbb

Arrogance in what you can find out…

but modesty in what you can say that football coaches don’t know.

Summary of how to be a football data ‘alchemist’.

- Plot things and make pictures.
- Do statistical tests. Be rigorous.
- Look for unexpected patterns.
- Try to build up a picture of what is going on that aligns with footballing knowledge.
- Enjoy yourself.

Go in to canvas and look at the exercise…