

DSC 101 PROJECT

Aisha Mahmoud & Diego Mendez
3rd Sprint Review

ACHIEVEMENTS

- What did we want to achieve in the last sprint?

After coming up with our plan, and method in the past sprints, it was time to actually execute it. We wanted to write all the code and find out the relevant stats for each player.

- What did we achieve in the last sprint?

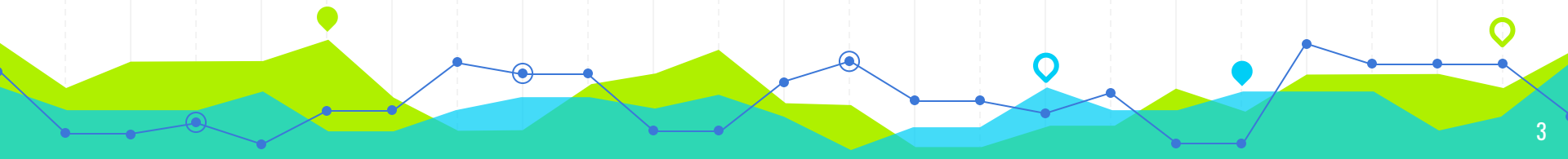
We achieved loading the data set, refining the data, splitting it up into three separate data frames (attackers, midfielders, and defenders), and keeping only the relevant statistics for each category. We also found the results of who had the best stats in each category. After figuring out this “formula”, it will be easier to do it for the rest of the data set we haven't finished yet.



ACHIEVEMENTS

- ◉ What are we especially proud of in the last sprint?

We're especially proud of getting to where we are, it seems daunting at first, and it's one thing to do it on DataCamp but another to do it on your own with no prompts. It went better than expected!



PROBLEMS WE RAN INTO?

- What did we not achieve in the last sprint?

We didn't get to plotting the EPL results, or finding all the stats for the MLS. (Now that we figured out all the trials and errors with the EPL dataset, it should be much easier and faster to do the American league one.)

- What are we going to do different in the next sprint?

In the next sprint, we'll be more organized, and now that we know what to do, be more efficient.



CRITERIA FOR PLAYERS STATS

Defense:

- challenges won
- ball interceptions
- tackles
- ball recoveries

Midfield:

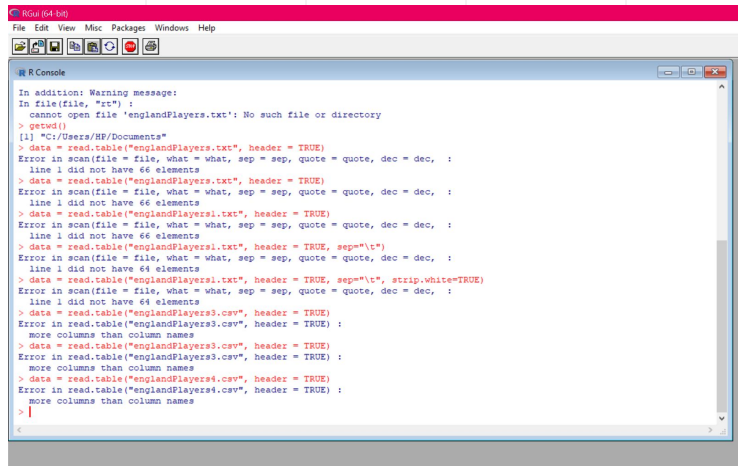
- assists
- accurate pass percentage
- key passes
- dribbles

Forward:

- expected goals
- goals
- percentage of shots on target
- dribbles

ERRORS WE RAN INTO

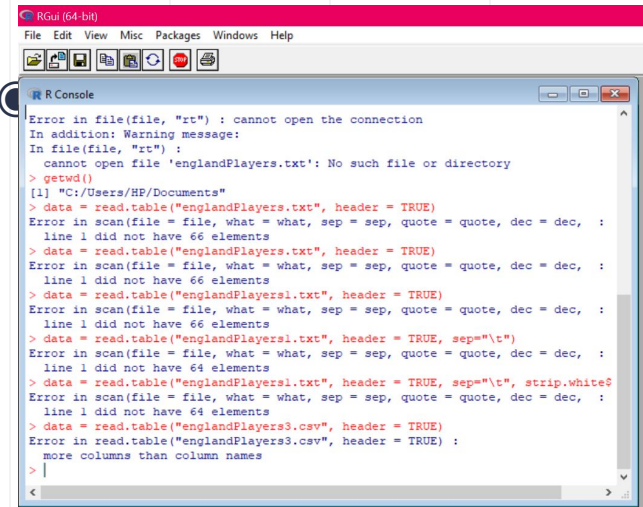
Started out in R. For some reason, had a lot of trouble trying to read in the file. It took a lot of trying different things to find out what worked.



```
RGui (64-bit)
File Edit View Misc Packages Windows Help

R Console

In addition: Warning message:
In file(file, "rt") :
cannot open file 'englandPlayers.txt': No such file or directory
> getwd()
[1] "C:/Users/HF/Documents"
> data = read.table("englandPlayers.txt", header = TRUE)
Error in scan(file = file, what = what, sep = sep, quote = quote, dec = dec, :
line 1 did not have 66 elements
> data = read.table("englandPlayers.txt", header = TRUE)
Error in scan(file = file, what = what, sep = sep, quote = quote, dec = dec, :
line 1 did not have 66 elements
> data = read.table("englandPlayers1.txt", header = TRUE)
Error in scan(file = file, what = what, sep = sep, quote = quote, dec = dec, :
line 1 did not have 66 elements
> data = read.table("englandPlayers1.txt", header = TRUE, sep="\t")
Error in scan(file = file, what = what, sep = sep, quote = quote, dec = dec, :
line 1 did not have 64 elements
> data = read.table("englandPlayers1.txt", header = TRUE, sep="\t", strip.white=TRUE)
Error in scan(file = file, what = what, sep = sep, quote = quote, dec = dec, :
line 1 did not have 64 elements
> data = read.table("englandPlayers3.csv", header = TRUE)
Error in read.table("englandPlayers3.csv", header = TRUE) :
more columns than column names
> data = read.table("englandPlayers3.csv", header = TRUE)
Error in read.table("englandPlayers3.csv", header = TRUE) :
more columns than column names
> data = read.table("englandPlayers4.csv", header = TRUE)
Error in read.table("englandPlayers4.csv", header = TRUE) :
more columns than column names
> |
```

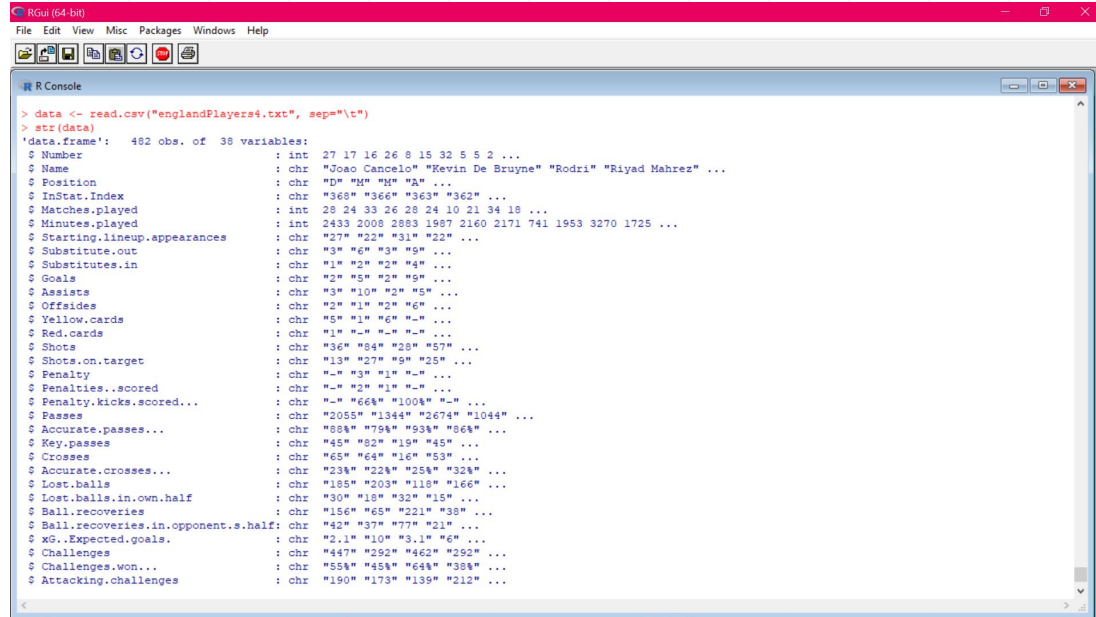


```
RGui (64-bit)
File Edit View Misc Packages Windows Help

R Console

Error in file(file, "rt") : cannot open the connection
In addition: Warning message:
In file(file, "rt") :
cannot open file 'englandPlayers.txt': No such file or directory
> getwd()
[1] "C:/Users/HF/Documents"
> data = read.table("englandPlayers.txt", header = TRUE)
Error in scan(file = file, what = what, sep = sep, quote = quote, dec = dec, :
line 1 did not have 66 elements
> data = read.table("englandPlayers.txt", header = TRUE)
Error in scan(file = file, what = what, sep = sep, quote = quote, dec = dec, :
line 1 did not have 66 elements
> data = read.table("englandPlayers1.txt", header = TRUE)
Error in scan(file = file, what = what, sep = sep, quote = quote, dec = dec, :
line 1 did not have 66 elements
> data = read.table("englandPlayers1.txt", header = TRUE)
Error in scan(file = file, what = what, sep = sep, quote = quote, dec = dec, :
line 1 did not have 66 elements
> data = read.table("englandPlayers1.txt", header = TRUE, sep="\t")
Error in scan(file = file, what = what, sep = sep, quote = quote, dec = dec, :
line 1 did not have 64 elements
> data = read.table("englandPlayers1.txt", header = TRUE, sep="\t", strip.white=TRUE)
Error in scan(file = file, what = what, sep = sep, quote = quote, dec = dec, :
line 1 did not have 64 elements
> data = read.table("englandPlayers3.csv", header = TRUE)
Error in read.table("englandPlayers3.csv", header = TRUE) :
more columns than column names
> |
```

cont.




```
> data <- read.csv("englandPlayers4.txt", sep="\t")
> str(data)
'data.frame':   482 obs. of  38 variables:
 $ Number      : int   27 17 16 26 8 15 32 5 2 ...
 $ Name        : chr   "Joao Cancelo" "Kevin De Bruyne" "Rodri" "Riyad Mahrez" ...
 $ Position    : chr   "Df" "Mf" "Mf" "Df" ...
 $ InStat.Index: chr   "368" "366" "363" "362" ...
 $ Matches.played: int   28 24 33 26 28 28 24 10 21 34 18 ...
 $ Minutes.played: int  2433 2008 2883 1987 2160 2171 741 1953 3270 1725 ...
 $ Starting.lineup.appearances: chr  "27" "22" "31" "22" ...
 $ Substitute.out: chr   "3" "6" "3" "5" ...
 $ Substitutes.in: chr   "1" "2" "2" "4" ...
 $ Goals       : chr   "2" "5" "2" "5" ...
 $ Assists     : chr   "3" "10" "2" "5" ...
 $ Offsides    : chr   "2" "1" "2" "6" ...
 $ Yellow.cards: chr   "5" "1" "6" "-" ...
 $ Red.cards   : chr   "1" "-" "-" "-" ...
 $ Shots       : chr   "36" "84" "28" "57" ...
 $ Shots.on.target: chr  "13" "27" "9" "25" ...
 $ Penalty    : chr   "-" "3" "1" "-" ...
 $ Penalties.scored: chr   "-" "5" "1" "-" ...
 $ Penalty.kicks.scored...: chr   "-" "66%" "100%" "-" ...
 $ Passes      : chr   "2055" "1344" "2674" "1044" ...
 $ Accurate.passes...: chr   "88%" "79%" "93%" "86%" ...
 $ Key.passes  : chr   "45" "82" "19" "45" ...
 $ Crosses     : chr   "65" "64" "16" "53" ...
 $ Accurate.crosses...: chr   "23%" "22%" "25%" "32%" ...
 $ Lost.balls  : chr   "185" "203" "118" "166" ...
 $ Lost.balls.in.own.half: chr   "30" "18" "32" "15" ...
 $ Ball.recoveries: chr   "156" "65" "221" "38" ...
 $ Ball.recoveries.in.opponent.s.half: chr  "42" "37" "77" "21" ...
 $ xG..Expected.goals.: chr   "2.1" "10" "3.1" "6" ...
 $ Challenges  : chr   "447" "292" "462" "292" ...
 $ Challenges.won...: chr   "55%" "45%" "64%" "38%" ...
 $ Attacking.challenges: chr   "190" "173" "139" "212" ...
```

Finally got it to work!

Now time for working with the actual data...

Next came trying to extract values. We only wanted the players that have played more than the avg amount of minutes played, a.k.a. the top half. We remembered we did something like this in class (with hotdogs...?)



```
RGui (64-bit)
File Edit View Misc Packages Windows Help

R Console

[193] FALSE TRUE TRUE TRUE TRUE FALSE TRUE TRUE TRUE FALSE TRUE TRUE TRUE FALSE TRUE TRUE FALSE TRUE FALSE FALSE TRUE TRUE TRUE FALSE TRUE
[217] TRUE TRUE TRUE TRUE FALSE TRUE TRUE TRUE TRUE TRUE TRUE TRUE FALSE TRUE TRUE FALSE TRUE TRUE FALSE TRUE FALSE FALSE TRUE FALSE
[241] TRUE FALSE FALSE TRUE TRUE FALSE FALSE TRUE TRUE TRUE FALSE FALSE FALSE TRUE TRUE FALSE FALSE TRUE FALSE FALSE TRUE TRUE FALSE FALSE
[265] TRUE TRUE TRUE TRUE TRUE TRUE FALSE FALSE FALSE TRUE TRUE FALSE FALSE TRUE FALSE FALSE FALSE TRUE TRUE TRUE FALSE FALSE
[289] FALSE FALSE FALSE TRUE TRUE TRUE TRUE FALSE TRUE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE TRUE FALSE TRUE FALSE FALSE
[313] TRUE TRUE FALSE TRUE FALSE FALSE TRUE TRUE FALSE FALSE FALSE TRUE FALSE FALSE FALSE TRUE FALSE FALSE TRUE FALSE TRUE TRUE FALSE
[337] TRUE FALSE TRUE TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE TRUE FALSE FALSE TRUE TRUE FALSE FALSE TRUE FALSE FALSE
[361] FALSE FALSE TRUE TRUE TRUE FALSE FALSE FALSE FALSE TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE TRUE
[385] FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE TRUE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[409] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[433] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[457] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[481] FALSE FALSE
> print(data[topHalf])
Error in `[.data.frame'](data, topHalf) : undefined columns selected
> print(data$topHalf)
NULL
> topHalf <- data[(data$Minutes.played > meanMinutes,)]
Error: unexpected ',' in "topHalf <- data[(data$Minutes.played > meanMinutes,"
> topHalf <- data[(data$Minutes.played > meanMinutes)]
Error in `[.data.frame'](data, (data$Minutes.played > meanMinutes)) :
  undefined columns selected
> data$Minutes.played[data$Minutes.played > meanMinutes]
 [1] 2433 2008 2883 1987 2160 2171 1953 3270 1725 2257 2593 3306 1508 1589 3298 3034 2328 1973 3262 1962 2912 2854 2456 3200 2980 1663 2631 1820 3641
[30] 2226 2774 3111 1763 3287 1980 2221 1860 2388 1707 3217 2193 2289 2769 2153 2045 3574 2076 2544 3128 2083 2064 2300 1861 2405 2706 3126 3476 2205
[59] 2760 3669 3655 2154 2717 2793 2523 2335 2305 3343 1843 3073 3382 1799 2206 1510 3137 2656 2121 2727 2857 2659 2287 1585 1935 1785 2786 2236 1645
[88] 2454 2742 3266 3570 1519 1734 1501 3012 2054 3318 2704 2426 3494 3318 3352 3638 2598 1742 1624 3064 2919 2032 1877 1567 1734 2776 1649 3013 2820
[117] 2936 2714 1670 1569 2553 2371 3628 2958 2840 2063 3401 1660 3195 1670 2984 2359 2787 2878 2740 3483 2598 2716 2193 3452 2210 2719 2801 2681 2964
[146] 2349 3566 2626 1512 2737 2861 2760 1815 1750 1594 3259 2485 1916 2113 2511 1738 2704 1690 3259 2576 2546 2314 1709 1843 3629 1891 2015 2994 3345
[175] 2139 2663 2916 2390 2408 2719 2746 2036 1934 2818 2509 2832 1851 2531 3232 2420 2763 2897 1887 1511 1695 1649 2139 2235 2237 2093 1716 2424 2415
[204] 2968 1979 2306 2053 2828 2763 2564 2967 1944 3137 2694 2898 1508 2370 2229 1716 2522 2663 2238 1635 1574 1965 1646 2171 1525 1532
> data[data$Minutes.played > meanMinutes]
Error in `[.data.frame'](data, data$Minutes.played > meanMinutes) :
  undefined columns selected
> |
```



```
RGU (64-bit)
File Edit View Misc Packages Windows Help

$ Name      : chr "Joao Cancelo" "Kurt Happy Zouma" "Joel Matip" "John Stones" ...
$ Number    : int 27 15 32 5 5 18 10 21 2 66 ...
$ Challenges.won... : chr "55%" "80%" "74%" "76%" ...
$ Tackles    : chr "120" "25" "23" "36" ...
$ Ball.recoveries : chr "156" "175" "75" "128" ...

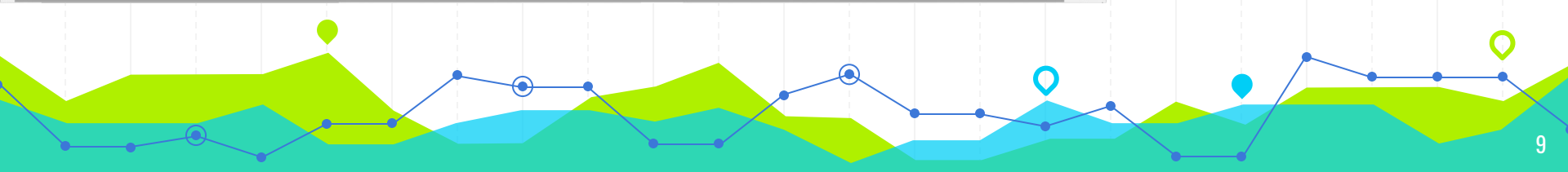
> str(attackersFrame)
Error in str(attackersFrame) : object 'attackersFrame' not found
> str(attackerFrame)
'data.frame': 132 obs. of 6 variables:
 $ Name      : chr "Riyad Mahrez" "Antonio Rudiger" "Jack Grealish" "Raheem Sterling" ...
 $ Number    : int 26 2 10 7 47 23 28 46 10 5 ...
 $ Dribbles   : chr "115" "12" "164" "168" ...
 $ Goals      : chr "9" "1" "6" "10" ...
 $ xG.Expected.goals.: chr "6" "1.62" "4.9" "11" ...
 $ Shots.on.target : chr "25" "7" "20" "27" ...

> str(midfieldersFrame)
'data.frame': 154 obs. of 6 variables:
 $ Name      : chr "Kevin De Bruyne" "Rodri" "Ilkay Gundogan" "Aymeric Laporte" ...
 $ Number    : int 17 16 8 14 11 11 24 10 47 10 ...
 $ Assists    : chr "10" "2" "2" "-" ...
 $ Accurate.passes... : chr "79%" "93%" "92%" "93%" ...
 $ Key.passes  : chr "82" "19" "30" "4" ...
 $ Dribbles    : chr "80" "38" "47" "13" ...

> str(defendersFrame)
'data.frame': 196 obs. of 5 variables:
 $ Name      : chr "Joao Cancelo" "Kurt Happy Zouma" "Joel Matip" "John Stones" ...
 $ Number    : int 27 15 32 5 5 18 10 21 2 66 ...
 $ Challenges.won... : chr "55%" "80%" "74%" "76%" ...
 $ Tackles     : chr "120" "25" "23" "36" ...
 $ Ball.recoveries : chr "156" "175" "75" "128" ...

> |
```

Separating the values.
One logic error we had here is that we realized we took players from our big original dataset and put them into the data frames for each position, instead of the refined top half group of players. This was later fixed.



-moved everything into collaboratory

RESULTS

-here are some results we've gotten:

DEFENDERS

```
str(defendersFrame)
```

```
'data.frame': 92 obs. of 5 variables:
 $ Name      : chr  "Joao Cancelo" "Kurt Happy Zouma" "John Stones" "Harry Maguire" ...
 $ Number    : int   27 15 5 5 2 21 2 66 6 3 ...
 $ Challenges.won...: chr  "55%" "80%" "76%" "72%" ...
 $ Tackles    : int   120 25 36 57 41 72 56 99 34 46 ...
 $ ballRecoveries : int   156 175 128 305 143 104 132 195 197 217 ...
```

Who made the most tackles?

```
[9] defendersFrame[which.max(defendersFrame$Tackles),]
```

```
A data frame: 1 x 5
```

Name	Number	Challenges.won...	Tackles	ballRecoveries
175 Luke Ayling	2	60%	186	308

Which defender had the most ball recoveries?

```
[10] defendersFrame[which.max(defendersFrame$ballRecoveries),]
```

```
A data frame: 1 x 5
```

Name	Number	Challenges.won...	Tackles	ballRecoveries
139 Jan Bednarek	35	67%	75	329

MIDFIELDERS

ATTACKERS

```
[ ] #finding who had the most:("Dribbles","Goals", "expectedGoals", "Shots.on.target")
```

```
print("The attacker with the most amount of goals:")
attackersFrame[which.max(attackersFrame$Goals),]

print("\n\nThe attacker with the most amount of expected goals:")
attackersFrame[which.max(attackersFrame$expectedGoals),]

print("\n\nThe attacker with the most amount of dribbles:")
attackersFrame[which.max(attackersFrame$dribbles),]

print("\n\nThe attacker with the most amount of shots on target:")
attackersFrame[which.max(attackersFrame$shots.on.target),]

#why does \n not work for making a new line?
```

```
[1] "The attacker with the most amount of goals:"
```

```
A data frame: 1 x 6
```

Name	Number	Dribbles	Goals	expectedGoals	Shots.on.target
14 Mohamed Salah	11	134	22	20	52

```
[1] "\n\nThe attacker with the most amount of expected goals:"
```

```
A data frame: 1 x 6
```

Name	Number	Dribbles	Goals	expectedGoals	Shots.on.target
14 Mohamed Salah	11	134	22	20	52

```
[1] "\n\nThe attacker with the most amount of dribbles:"
```

```
A data frame: 1 x 6
```

Name	Number	Dribbles	Goals	expectedGoals	Shots.on.target
190 Adama Traore	37	329	2	3.2	13

```
[1] "\n\nThe attacker with the most amount of shots on target:"
```

```
A data frame: 1 x 6
```

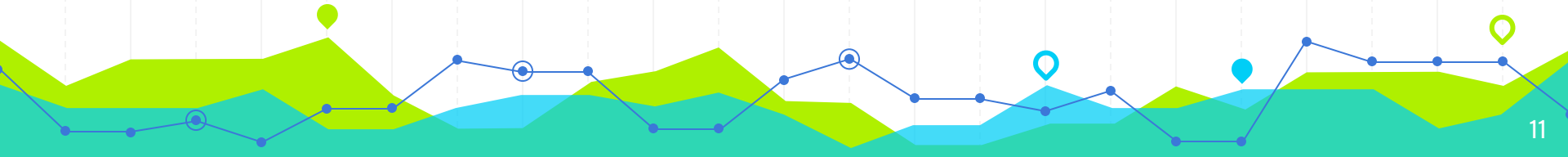
Name	Number	Dribbles	Goals	expectedGoals	Shots.on.target
26 Harry Kane	10	119	22	20	53

Here we see that Mo Salah had both the highest number of expected goals, and the highest number of goals. So he really lived up to his expectations.

Questions for the product owner...

Are these results so far
what you expected?

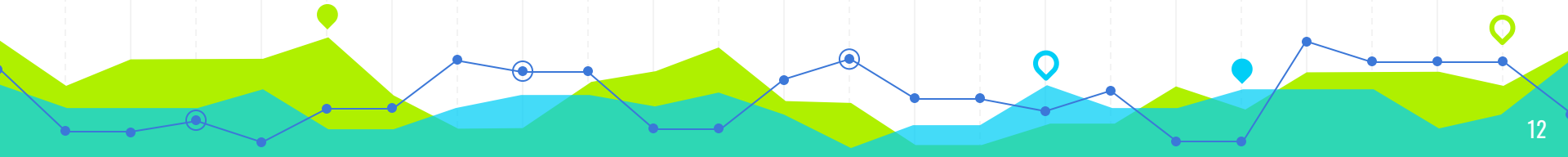
What's your favorite color
so we can make that the
color of one of our
histograms? :)



Questions for other team members...

What do you think was the hardest part about all the coding done so far in this project?

How long did it take for you to load in the data set?
Did it go smoother for you?



Stay tuned for the next sprint review!

