Mike Messina

I. Introduction

My individual contribution to the analysis for the NFL group will be focusing on the relational technique, Correspondence Analysis, as a means to attempt to summarize and interpret categorical variables of interest.

To begin the analysis, we needed a way to break each individual team up into multiple classes. This was accomplished by including the Offensive Simple Rating System (OSRS) and the Defensive Simple Rating System (DSRS) for each team in each given year. The scores were imported to the dataset from Sports-reference.com. Simply put, the OSRS and DSRS is a numerical rating system where a zero score is considered league average. In order to create categorical dimensions of this data, I observed the first and third quartiles of each measurement to create separate bins for average, above average, and below average teams on the offensive and defensive side of the ball. SRS scores below the 1st quartile are considered "below average" while SRS scores above the 3rd quartile are considered "above average". Scores between the 1st and 3rd Quartile are considered "average". You can see the six-number summary below for the bin splits for Home/Away Offenses/Defenses below.

```
Home_Off_Rank       Home_Def_Rank       Away_Off_Rank       Away_Def_Rank
Min.   :-1.17e+01   Min.   :-9.800000   Min.   :-11.700000   Min.   :-9.800000
1st Qu.:-3.00e+00   1st Qu.:-2.300000   1st Qu.: -3.000000   1st Qu.:-2.300000
Median :-1.00e-01   Median : 0.000000   Median : -0.100000   Median : 0.000000
Mean   : 8.06e-04   Mean   : 0.004116   Mean   : -0.001045   Mean   :-0.004094
3rd Qu.: 2.70e+00   3rd Qu.: 2.600000   3rd Qu.:  2.700000   3rd Qu.: 2.600000
Max.   : 1.59e+01   Max.   : 9.800000   Max.   : 15.900000   Max.   : 9.800000
```

Now that we have multi-dimensional categorical data for each team observation, we can begin to see how these associate with other categorical variables of interest from our dataset. The first area of interest we'll explore here is how these Offensive/Defensive ranks correspond to the region of the US where the game was played (North, South, East, West).

First, we need to pivot the two variables to see how often teams with above average offenses or below average defenses etc. occurred in each region. We will use this table to run our CA on. The table can be seen below.

Region	H.Off.Plus	H.Def.Plus	A.Off.Plus	A.Def.Plus	H.Off.Minus	H.Def.Minus	A.Off.Minus	A.Def.Minus	
East	329	408	281	291	232	128	302	267	2238
North	352	232	306	344	360	376	305	333	2608
South	200	273	325	290	327	374	325	319	2433
West	240	208	207	195	216	247	204	212	1729
	1121	1121	1119	1120	1135	1125	1136	1131	

Average offense/defense probably had more to do with whether or not the opponent was above or below average on offense/defense which is why these types were excluded from the analysis. We will only be looking at above/below average offenses and defenses in the analysis.

Mike Messina

II. Team Type vs. Region

The correspondence matrix can be seen below which shows the correlation coefficients between each variable. Nothing egregious stands out from the table. The highest correlation is seen at 4.5% between the East region and home teams with above average defenses. The lowest correlation is seen at 1.4% between the East region and home teams with below average defenses.

```
> #Correspondence Matrix
> P = Team.Type.vs.Region/sum(Team.Type.vs.Region)
> round(P,3)
      H.Off.Plus H.Def.Plus A.Off.Plus A.Def.Plus H.Off.Minus H.Def.Minus A.Off.Minus A.Def.Minus
East       0.037      0.045      0.031      0.032       0.026       0.014       0.034       0.030
North      0.039      0.026      0.034      0.038       0.040       0.042       0.034       0.037
South      0.022      0.030      0.036      0.032       0.036       0.042       0.036       0.035
West       0.027      0.023      0.023      0.022       0.024       0.027       0.023       0.024
```

A mosaic plot (shown left) was generated to quickly view which variables corresponded more or less often than one would expect. As we can see, The East region played host to Home teams with above average offenses and defenses more often than expected and less often to home teams with below average offenses and defenses. The South Region played host to home teams above average offenses less often than expected while playing host to Home teams with below average defenses more often than expected.

Shown right is the summary of the correspondence analysis output. We can see that one dimension accounts for 81% of the total variation while we can get to 97% cumulative variation with just two dimensions.

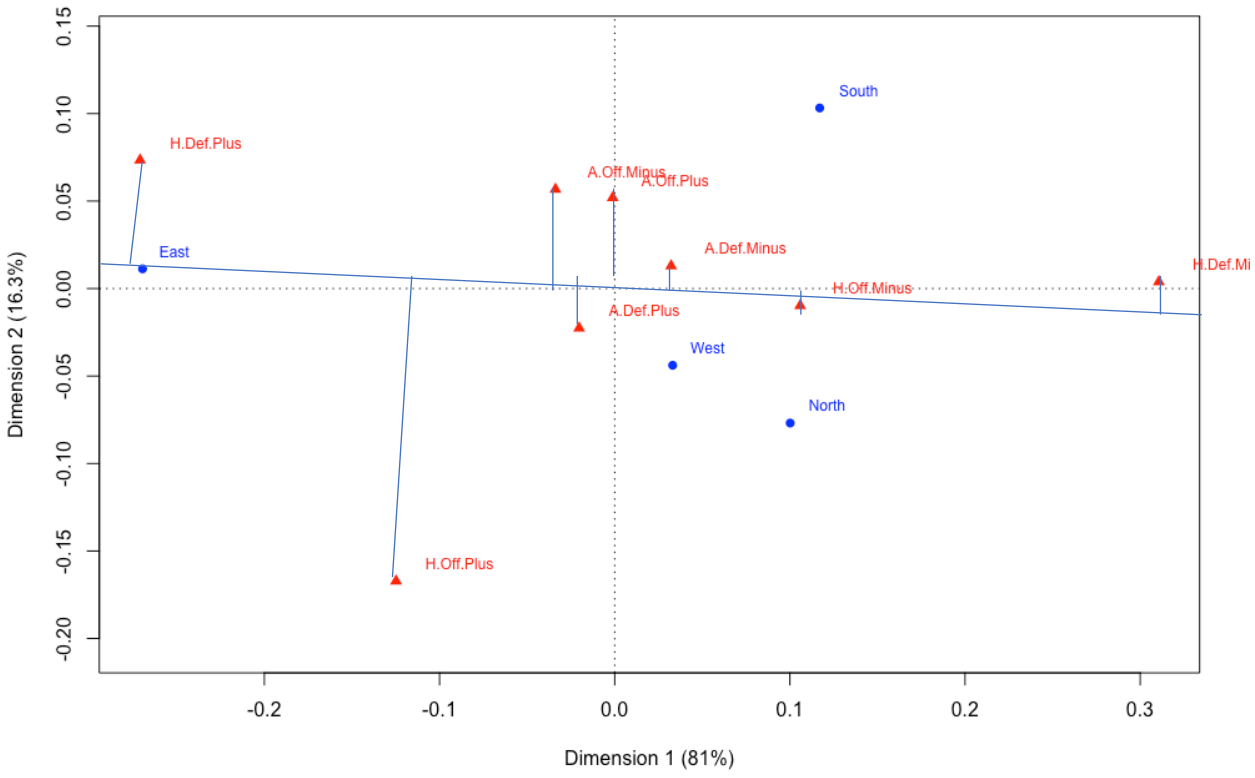
```
> c = ca(Team.Type.vs.Region)
> summary(c)

Principal inertias (eigenvalues):

 dim    value      %    cum%   scree plot
 1      0.024867  81.0  81.0   ********************
 2      0.004989  16.3  97.2   ****
 3      0.000847   2.8 100.0   *
        -------- -----
 Total: 0.030703 100.0
```

Mike Messina

```
> c$rowcoord
             Dim1       Dim2
East  -1.7093286  0.1593912
North  0.6353714 -1.0883972
South  0.7421417  1.4612725
West   0.2098311 -0.6208523
> colC = c$colcoord[ , 1:2]
> colC[order(colC[,1]), ]
                Dim1        Dim2
H.Def.Plus  -1.718288916  1.03954866
H.Off.Plus  -0.791574870 -2.36702872
A.Off.Minus -0.214509820  0.80300893
A.Def.Plus  -0.128575776 -0.32009251
A.Off.Plus  -0.007177791  0.73552580
A.Def.Minus  0.204177010  0.18333610
H.Off.Minus  0.671424620 -0.13860662
H.Def.Minus  1.970032120  0.05449122
> colC[order(colC[,2]), ]
                Dim1        Dim2
H.Off.Plus  -0.791574870 -2.36702872
A.Def.Plus  -0.128575776 -0.32009251
H.Off.Minus  0.671424620 -0.13860662
H.Def.Minus  1.970032120  0.05449122
A.Def.Minus  0.204177010  0.18333610
A.Off.Plus  -0.007177791  0.73552580
A.Off.Minus -0.214509820  0.80300893
H.Def.Plus  -1.718288916  1.03954866
```

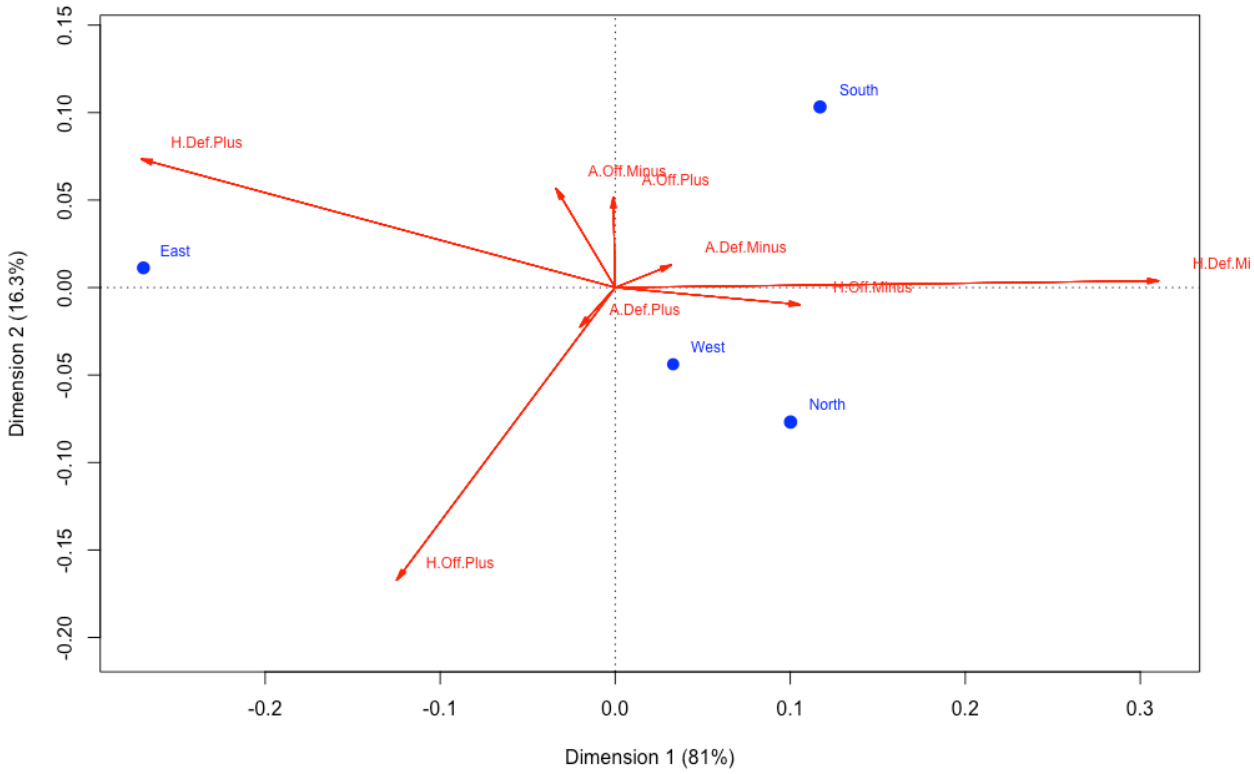
The scores of each Region corresponding to each dimension are shown left. We can see that the first dimension does a good job of splitting off the East Region while the 2nd dimension then separates the South and North Regions well. The scores for each team type are also shown left in order. The first dimension splits off home teams with above average defenses and home teams with below average defenses while the 2nd dimension then helps to split home teams with above average offenses and home teams with above average defenses.

The symmetric plot below gives a good summary of how each dimension has separated the data and which team types correspond to which regions. For illustrative purposes, I've shown the scale for which team types correspond to the East region below on the symmetric plot. Team types closer to the East glyph on the line drawn through the origin and East region correspond more heavily.



Mike Messina

Essentially another view of the mosaic plot is shown below with how each of the coordinates correspond more or less likely to each Region and team type. More obtuse angles from the origin to each Region indicate team types that correspond less and vice versa.



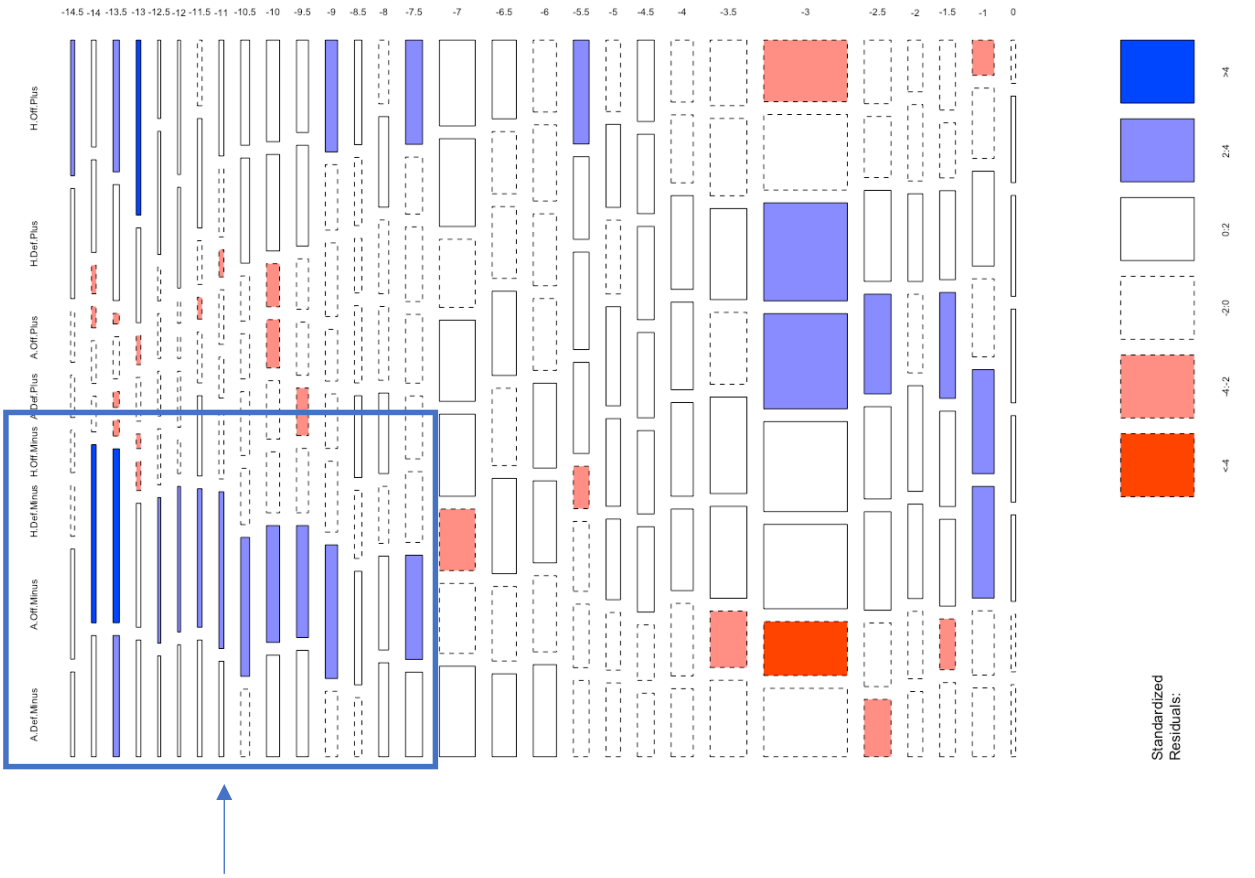
Mike Messina

III. Lines vs. Team Type

Another area of interest I wanted to look into with Correspondence Analysis was between each Spread number from the last 20 years and how it corresponded to the types of teams playing. This type of analysis could help us being better at identifying how certain line numbers correspond to the teams playing. Understanding the indicators of certain game lines can give bettors an edge when choosing between different sportsbooks to place their wagers at given the different vig values. Each line from the last 20 years was pivoted with how often a home/away offense/defense was above or below average for the game. Spreads greater than -14.5 were taken out of the analysis as these appeared to be outlier data.

The correspondence table is shown below. As expected, very low correlations are shown given the sharpness of sportsbook at making lines. One interesting note was that the highest correlations all appeared with the -3 line. This is most likely related to the -3 line being the most common spread of an NFL game by far (more than double the next most common) so there are more team types that have played at this spread line than others out of sheer volume.

```
> #Correspondence Matrix
> P = Lines.vs.Ranks1/sum(Lines.vs.Ranks1)
> round(P,3)
      H.Off.Plus H.Def.Plus A.Off.Plus A.Def.Plus H.Off.Minus H.Def.Minus A.Off.Minus A.Def.Minus
-14.5      0.002      0.001      0.001      0.001       0.001       0.001       0.001       0.001
-14        0.002      0.001      0.000      0.000       0.001       0.001       0.003       0.002
-13.5      0.003      0.002      0.000      0.001       0.000       0.000       0.004       0.003
-13        0.003      0.001      0.000      0.001       0.000       0.000       0.002       0.002
-12.5      0.001      0.001      0.000      0.000       0.000       0.001       0.001       0.001
-12        0.001      0.001      0.000      0.000       0.001       0.000       0.001       0.001
-11.5      0.001      0.002      0.001      0.000       0.001       0.001       0.002       0.002
-11        0.002      0.001      0.000      0.001       0.001       0.001       0.003       0.002
-10.5      0.003      0.003      0.001      0.001       0.002       0.002       0.004       0.002
-10        0.005      0.004      0.002      0.002       0.003       0.003       0.005       0.005
-9.5       0.004      0.004      0.002      0.002       0.002       0.003       0.005       0.004
-9         0.005      0.003      0.003      0.002       0.002       0.003       0.006       0.003
-8.5       0.003      0.002      0.001      0.002       0.002       0.002       0.003       0.001
-8         0.002      0.003      0.002      0.002       0.003       0.002       0.003       0.003
-7.5       0.006      0.003      0.004      0.004       0.004       0.004       0.006       0.005
-7         0.010      0.011      0.008      0.010       0.010       0.007       0.008       0.011
-6.5       0.006      0.005      0.006      0.007       0.006       0.008       0.006       0.007
-6         0.006      0.006      0.006      0.006       0.007       0.006       0.006       0.007
-5.5       0.006      0.004      0.005      0.005       0.002       0.004       0.003       0.004
-5         0.003      0.004      0.004      0.005       0.004       0.004       0.003       0.004
-4.5       0.005      0.005      0.005      0.005       0.005       0.005       0.003       0.004
-4         0.005      0.005      0.007      0.007       0.007       0.006       0.005       0.005
-3.5       0.008      0.010      0.011      0.009       0.012       0.011       0.007       0.009
-3         0.017      0.021      0.028      0.027       0.025       0.024       0.015       0.019
-2.5       0.006      0.006      0.008      0.009       0.008       0.009       0.006       0.005
-2         0.003      0.004      0.004      0.004       0.005       0.005       0.003       0.003
-1.5       0.004      0.003      0.005      0.006       0.005       0.005       0.003       0.004
-1         0.003      0.005      0.007      0.006       0.008       0.008       0.005       0.005
0          0.001      0.001      0.002      0.001       0.001       0.001       0.001       0.001
```



The Mosaic plot of the table between game lines and team types is shown above. Some interesting areas of note apart from the -3 line is that the majority of the higher lines (> -7) all had home and away teams with below average offenses and defenses occur more often than average.

```
> mosaicplot(Lines.vs.Ranks, shade=T, main="")
> c = ca(Lines.vs.Ranks)
> summary(c)

Principal inertias (eigenvalues):

 dim    value      %    cum%   scree plot
 1      0.061839  76.6  76.6   ******************
 2      0.007140   8.8  85.4   **
 3      0.004465   5.5  91.0   *
 4      0.002367   2.9  93.9   *
 5      0.001911   2.4  96.3   *
 6      0.001641   2.0  98.3   *
 7      0.001367   1.7 100.0
        -------- -----
 Total: 0.080729 100.0
```

After running correspondence analysis function on the table, we can see how many dimensions are required to capture an adequate amount of variance in the dataset. According to the summary table left, we can see that just one dimension only accounts for 76.6% of the total variance which is actually a low number for correspondence analysis. Two coordinates account for 85% cumulative variance and not until we introduce three dimensions do we cover >90% of the total variance in the data. This isn't a bad thing, it just is indicating that the data we're working with isn't easily separated.

Mike Messina

```
> rowC = c$rowcoord[ , 1:2]
> rowC[order(rowC[,1]), ] #sort by first coordinate
             Dim1        Dim2
-13.5 -3.34869707 -0.17150686
-14   -2.82847798  1.90152046
-13   -2.82625350 -3.15393321
-12   -2.62860720  0.38344825
-11   -2.09517600  0.47718347
-12.5 -1.93561861  2.08492580
-14.5 -1.80823672 -1.32560357
-11.5 -1.67181627  3.37560882
-10.5 -1.64831791  0.87526210
-10   -1.50116129  0.55568062
-9.5  -1.37894454  0.49771576
-9    -1.22704596 -0.16812729
-8.5  -0.74269770 -0.27285764
-7.5  -0.72712494 -0.74117681
-8    -0.38142880  0.85953587
-7    -0.22301548 -0.57868255
-5.5  -0.15059173 -2.72514730
-6    -0.04841395  0.59733061
-6.5   0.10520718 -0.07280391
-5     0.41774680 -0.50410689
-4.5   0.47713685 -0.88943014
-4     0.52292315  0.27885525
-3.5   0.56490813  0.51003817
-2     0.69720032  1.31446499
-3     0.77646979 -0.33145185
-2.5   0.82226098 -0.11004948
0      0.82873954  0.51175560
-1.5   0.83737206 -0.83576274
-1     1.00218369  2.04258885
```

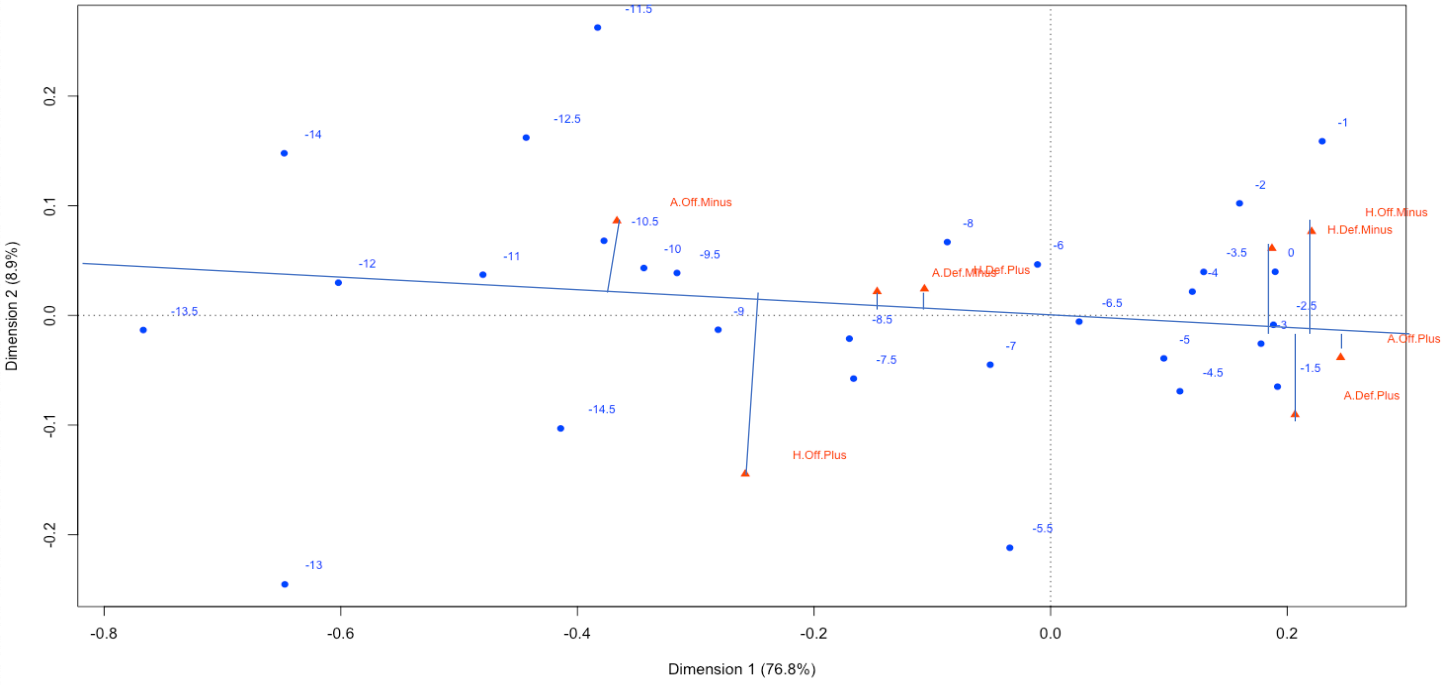
```
> rowC[order(rowC[,2]), ] #sort by second coordinate
             Dim1        Dim2
-13   -2.82625350 -3.15393321
-5.5  -0.15059173 -2.72514730
-14.5 -1.80823672 -1.32560357
-4.5   0.47713685 -0.88943014
-1.5   0.83737206 -0.83576274
-7.5  -0.72712494 -0.74117681
-7    -0.22301548 -0.57868255
-5     0.41774680 -0.50410689
-3     0.77646979 -0.33145185
-8.5  -0.74269770 -0.27285764
-13.5 -3.34869707 -0.17150686
-9    -1.22704596 -0.16812729
-2.5   0.82226098 -0.11004948
-6.5   0.10520718 -0.07280391
-4     0.52292315  0.27885525
-12   -2.62860720  0.38344825
-11   -2.09517600  0.47718347
-9.5  -1.37894454  0.49771576
-3.5   0.56490813  0.51003817
0      0.82873954  0.51175560
-10   -1.50116129  0.55568062
-6    -0.04841395  0.59733061
-8    -0.38142880  0.85953587
-10.5 -1.64831791  0.87526210
-2     0.69720032  1.31446499
-14   -2.82847798  1.90152046
-1     1.00218369  2.04258885
-12.5 -1.93561861  2.08492580
-11.5 -1.67181627  3.37560882
```

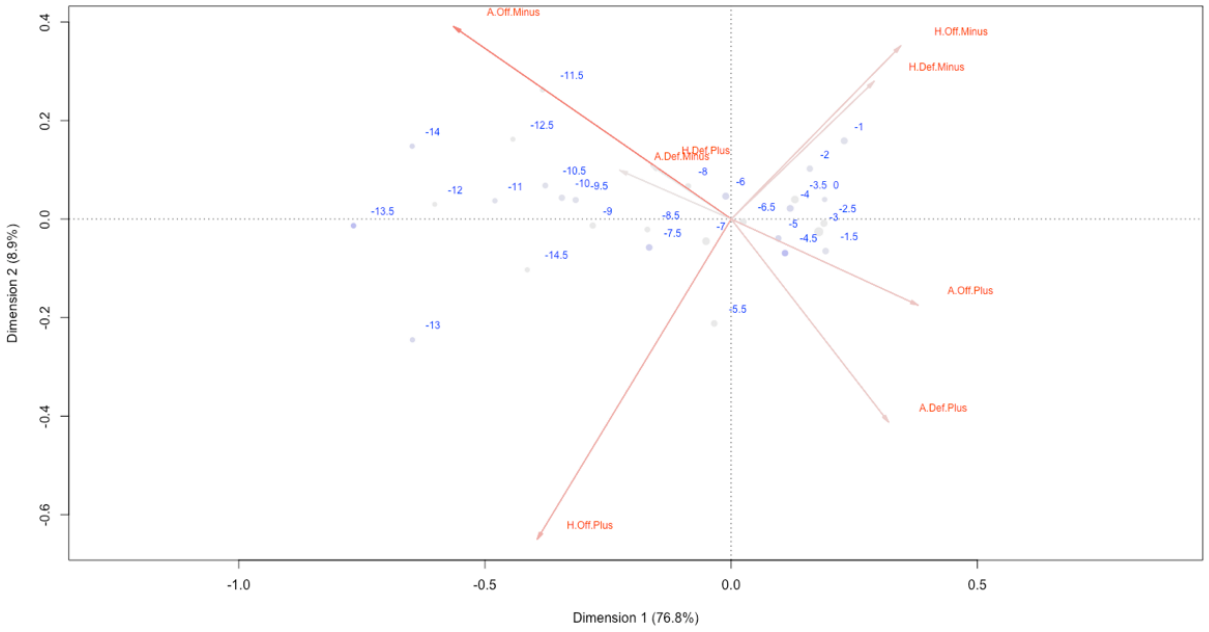
```
> colC = c$colcoord[ , 1:2]
> colC[order(colC[,1]), ]
                 Dim1       Dim2
A.Off.Minus -1.6005741  1.1091192
H.Off.Plus  -1.1268933 -1.8595729
A.Def.Minus -0.6403780  0.2798323
H.Def.Plus  -0.4657447  0.3113915
H.Def.Minus  0.8167900  0.7864030
A.Def.Plus   0.9021067 -1.1637710
H.Off.Minus  0.9632530  0.9839555
A.Off.Plus   1.0702982 -0.4931298
> colC[order(colC[,2]), ]
                 Dim1       Dim2
H.Off.Plus  -1.1268933 -1.8595729
A.Def.Plus   0.9021067 -1.1637710
A.Off.Plus   1.0702982 -0.4931298
A.Def.Minus -0.6403780  0.2798323
H.Def.Plus  -0.4657447  0.3113915
H.Def.Minus  0.8167900  0.7864030
H.Off.Minus  0.9632530  0.9839555
A.Off.Minus -1.6005741  1.1091192
```

The scores for each game spread and team type in the first two dimension is shown above. The first dimension looks to split the spread lines by numerical order. The second dimension then does a good job splitting off the higher spreads of -13, -11.5 and -12.5. It's interesting that the -1 spread line is also separated off in the 2nd dimension with these higher spreads. Moving our attention to the team types, the first dimension separates out away teams with above and below average offenses. The 2nd dimension then separates home teams with above average offenses and away teams with below average offenses.

The symmetric plot below gives a clearer view of how these scores correspond to each other and each dimension of the analysis. For illustrative purposes, we can view the scale of how the different team types correspond to the -3 spread line. Home teams with below average offenses and defenses along with away teams with above average offenses and defenses correspond highly with the -3 line while away teams with below average offenses and home teams with above average offenses.

Mike Messina





The rowgreen map shown above actually gives a great illustration on which spread lines correspond to certain team types. You can see that the smaller line numbers (-1 through -6.5) are all grouped together acutely to Home teams with below average offenses/defenses and away teams with above average offenses/defenses. These smaller spread lines correspond much less

to the other team types as indicated by their obtuse directional relationship on the map. It's important to look closely at these smaller spread numbers as they occur more commonly than the larger spreads. On the other hand, the larger spreads all seem to be grouped together by the first dimension and we're able to see an acute directional relationship to away teams with below average offenses.

IV. Next Steps

Now that we have a better idea of the relative relationships between our categorical variables like region, team types, and line number correspond together, we can use this knowledge to identify bad lines offered by sportsbooks and how geographic location is related to offensive and defensive location. We can also use this knowledge to better understand our data as we move to build predicting models with improved accuracy.