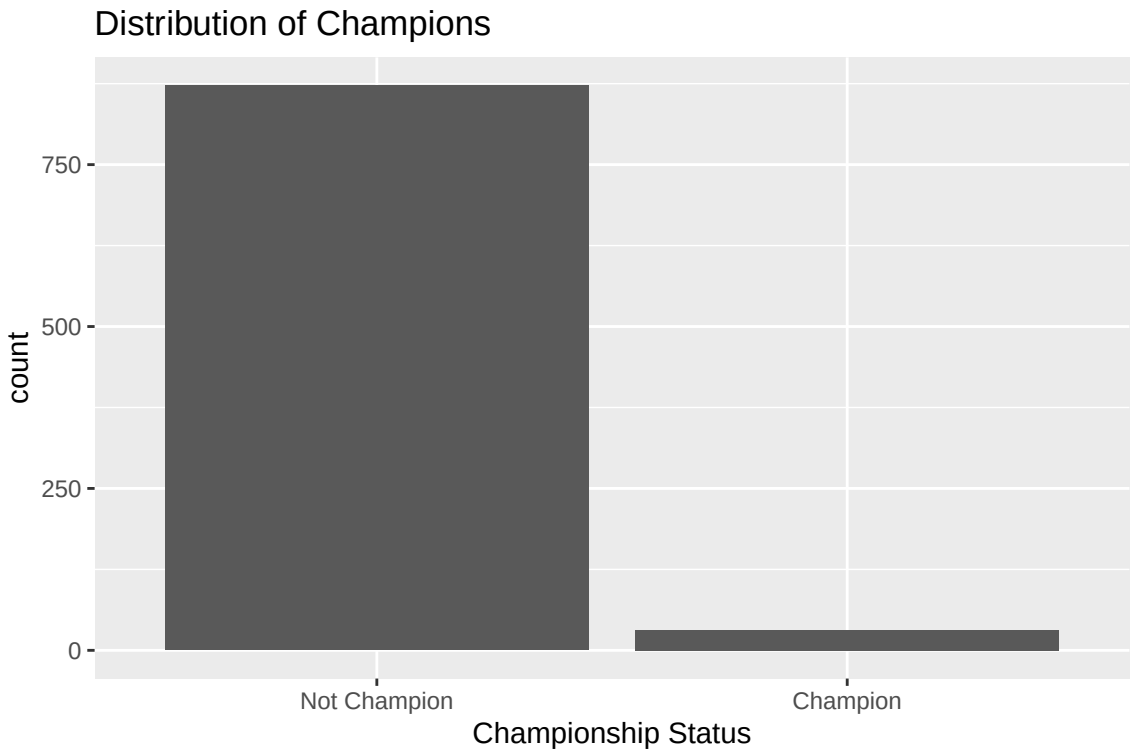
NBA Exploratory Data Analysis

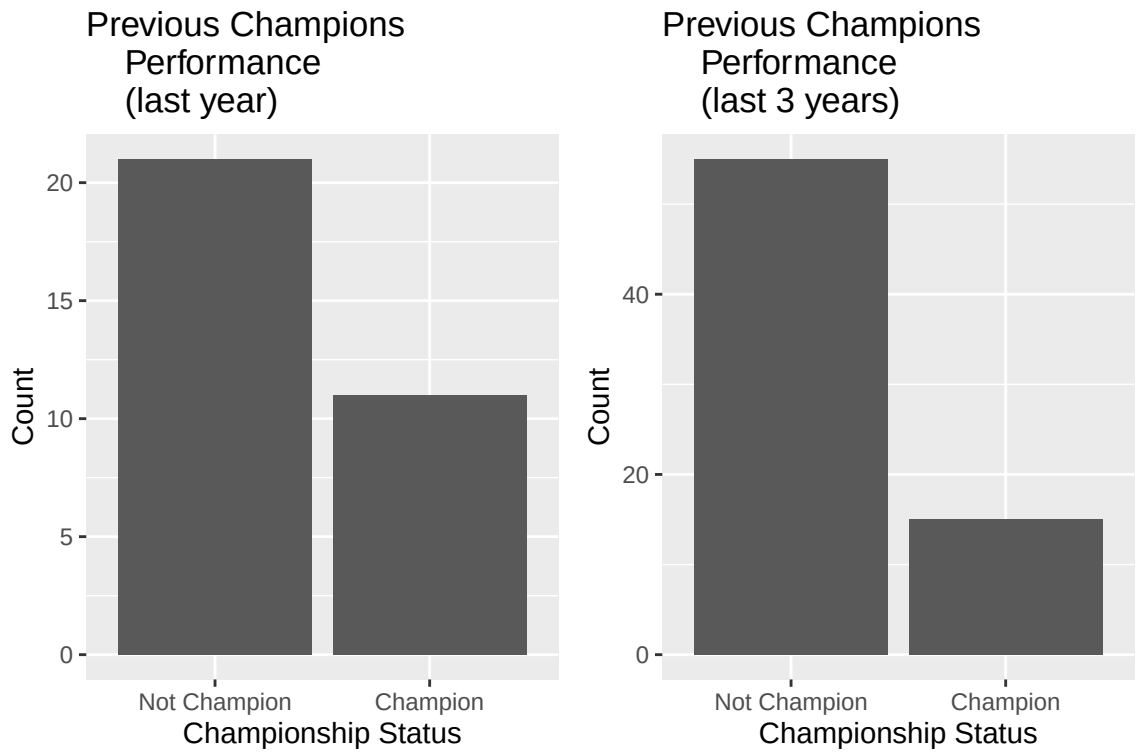
Frankie Willard

Visualizing the Distribution of the Response Variable



As shown above, only a small proportion of the dataset is a Champion. This is because in the NBA, there is only one champion out of thirty teams in a year. This will create a class imbalance in our dataset.

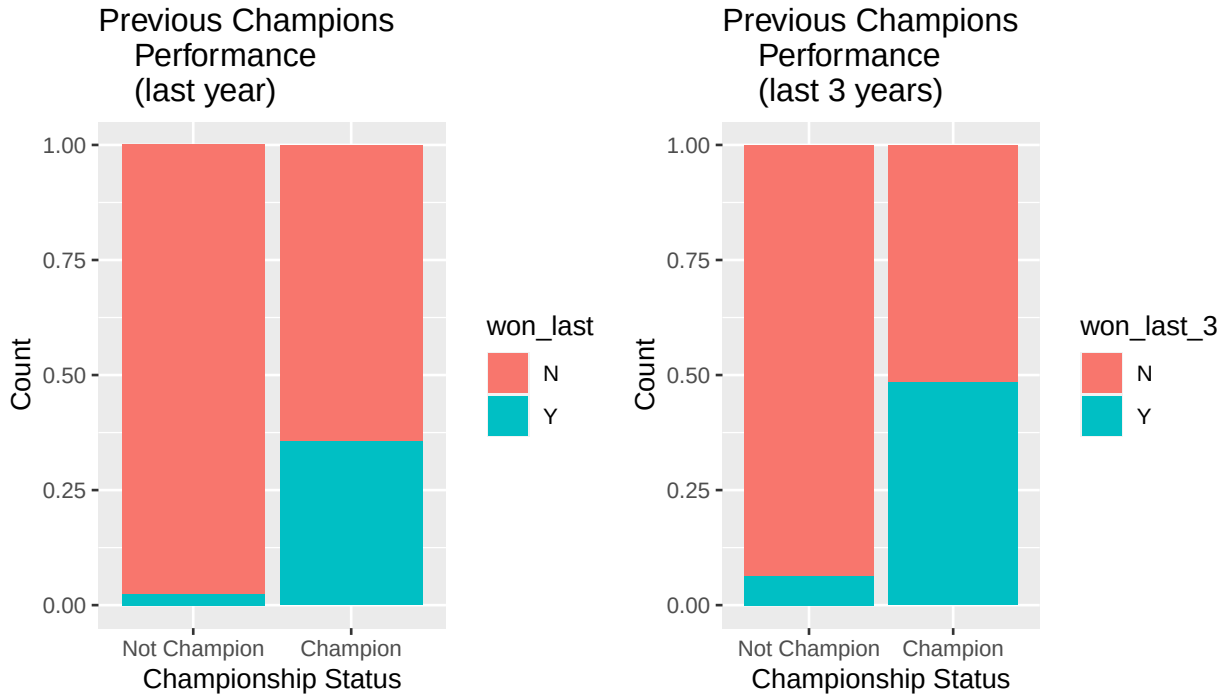
Visualizing the Relationship Between Discrete Variables and the Response Variable



The first bar plot shows the distribution of Champions among teams that won in the previous year. This shows that of teams that won in the previous year, approximately 1/3 of them won the next year.

The second bar plot shows the distribution of Champions among teams that won in the previous 3 years. This shows that of teams that won in the previous 3 years in a given year, approximately 1/5 of them won the championship.

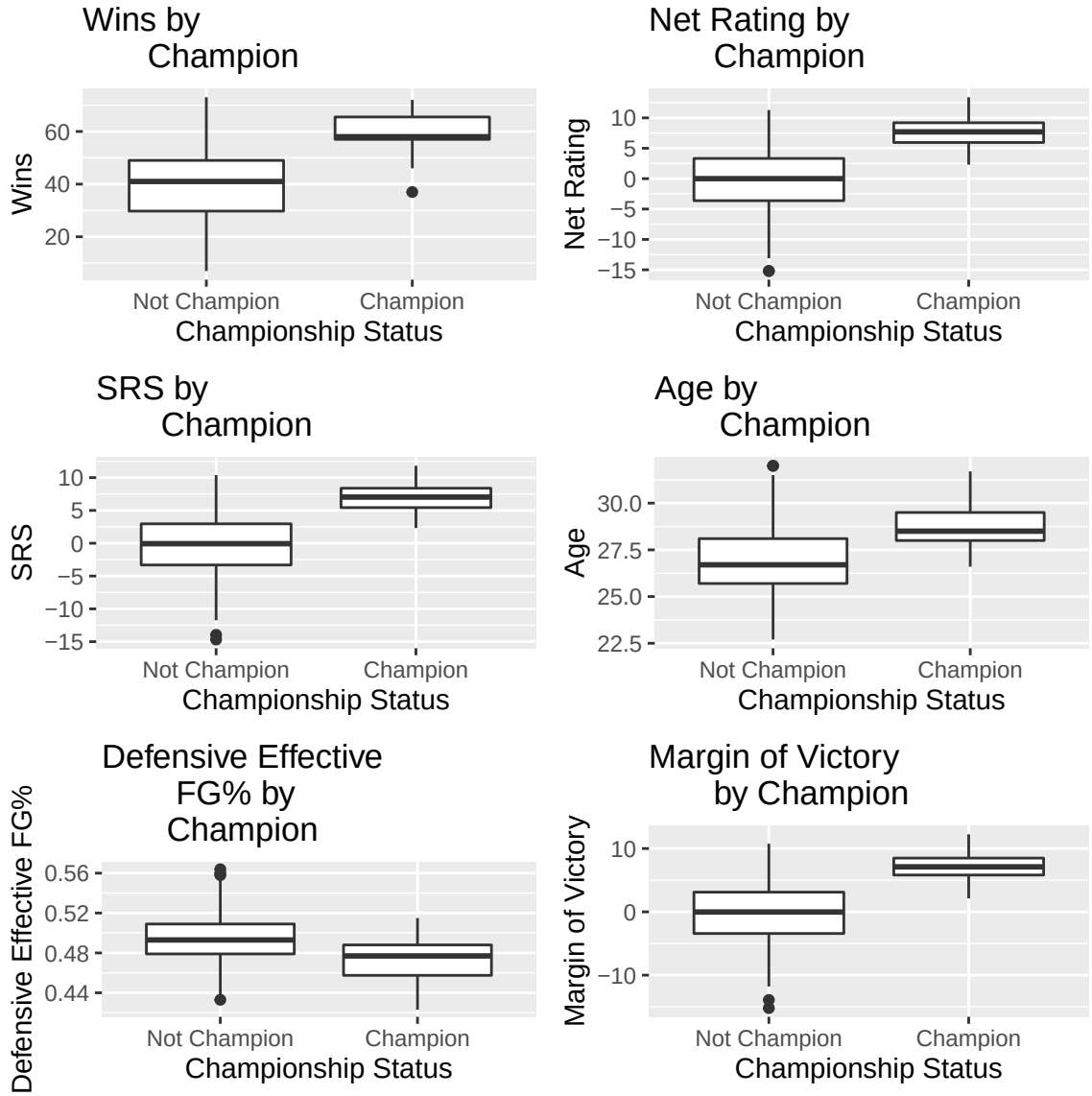
These both are clearly impactful variables to explore further.



The first relative bar plot shows that of teams that won the championship in a given year, approximately 35% of those teams had won in the year before.

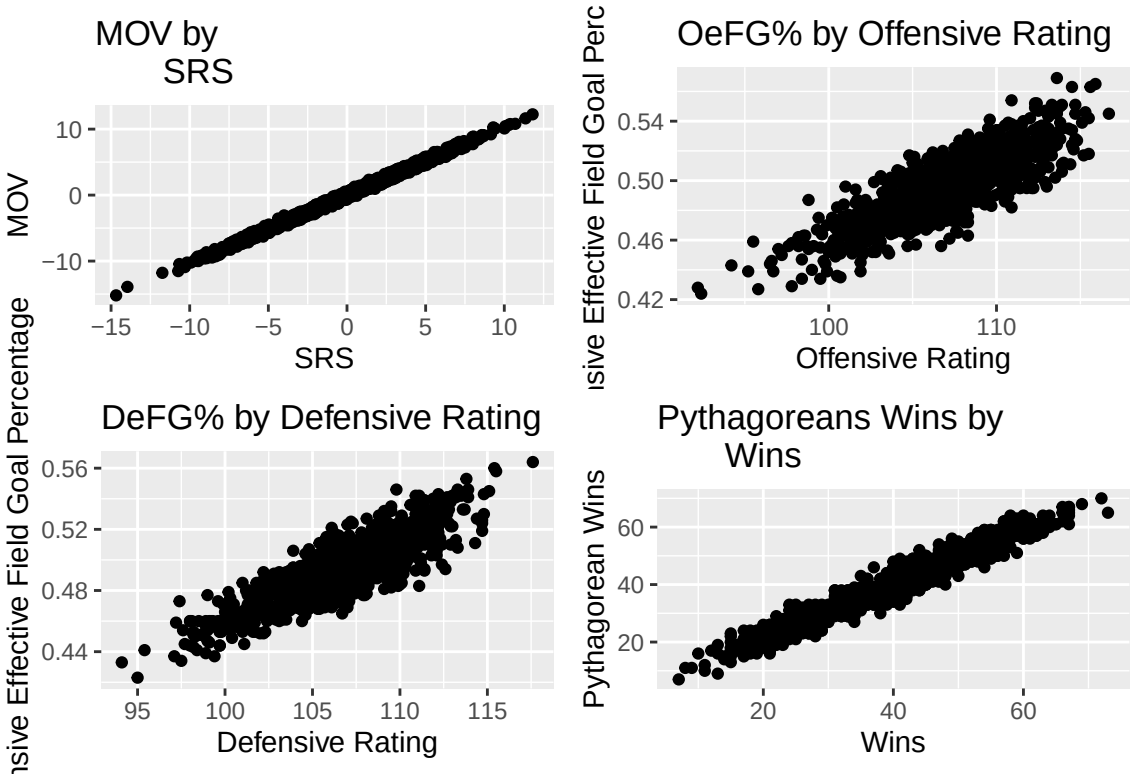
The second relative bar plot shows that of teams that won the championship in a given year, approximately 50% of those teams had won in one of the previous 3 years before.

Visualizing the Relationship Between Continuous Variables and the Response Variable



The box plots above show several variables that I found intriguing through exploratory data analysis. This is demonstrated by the difference in location of the boxes, their medians, and their sizes between champions and non-champions. We find that the margin of victory, SRS, and win variables to be especially noteworthy (although the other variables visually appear significant). In the appendix is a visualization of these same variables in spine plots.

Visualizing the Collinearity of the Predictor Variables

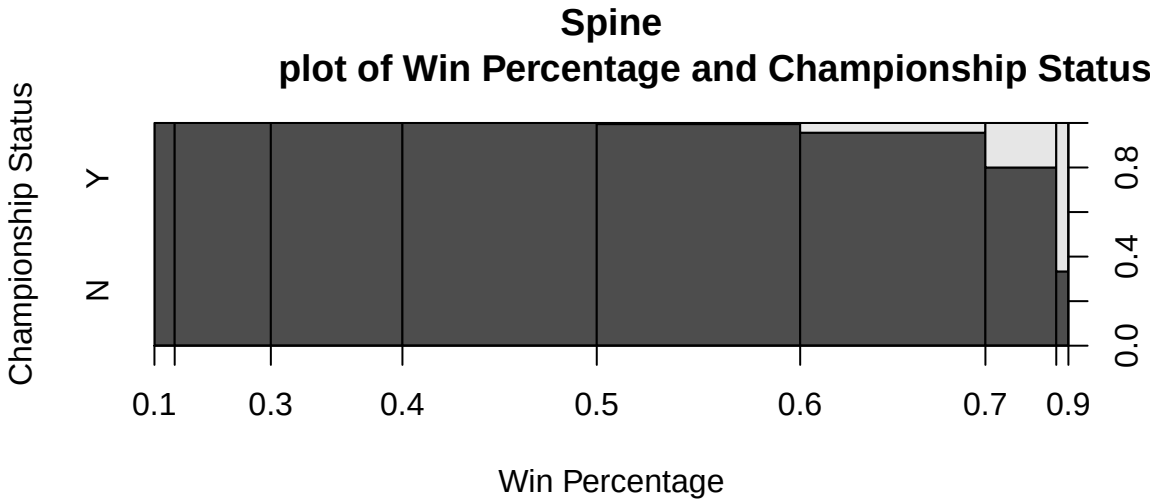


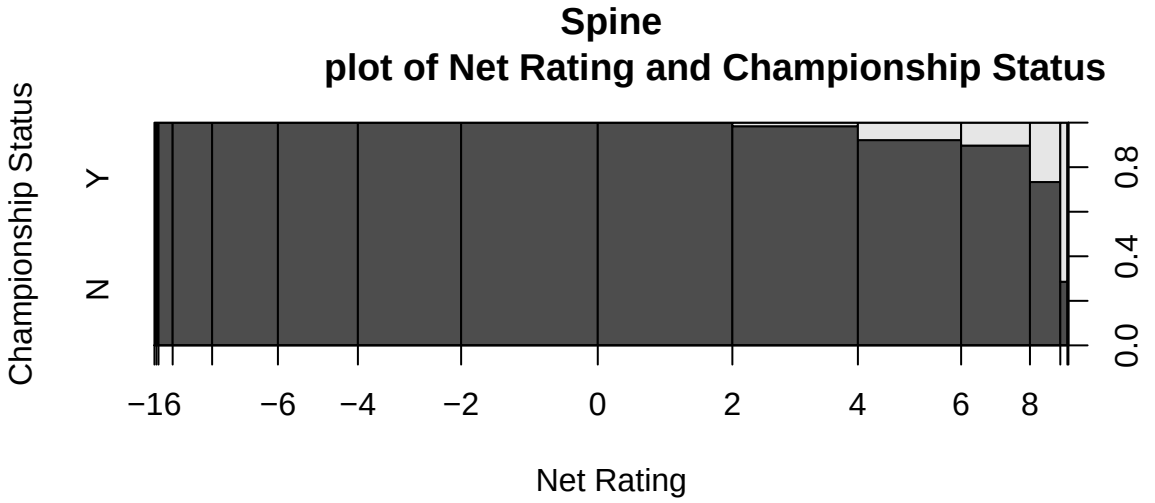
By the nature of some advanced stats, some of the variables are inherently collinear, such as MOV and SRS as SRS uses MOV in its calculation (and the same thing for Pythagorean Wins and Wins).

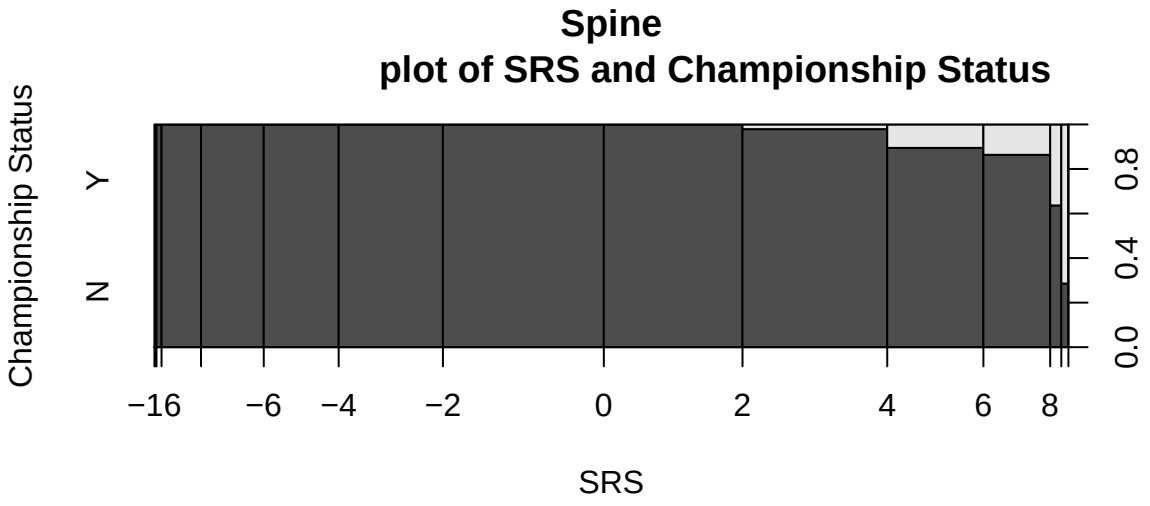
Additionally, offensive predictors will likely have collinearity with each other as well as defensive predictors, although this collinearity is not as strong as the aforementioned predictors.

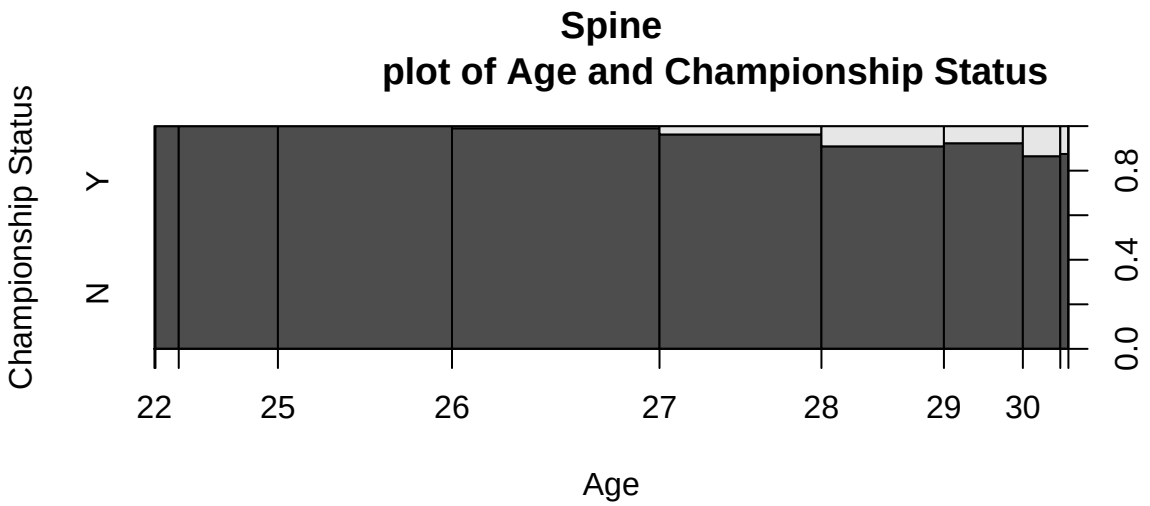
Identifying potential sources of collinearity is crucial for variable selection and feature engineering, as we hope to add only important predictors to our final model in order to reduce sources of variance such as the curse of dimensionality, as a more parsimonious model with similar bias will have less variance.

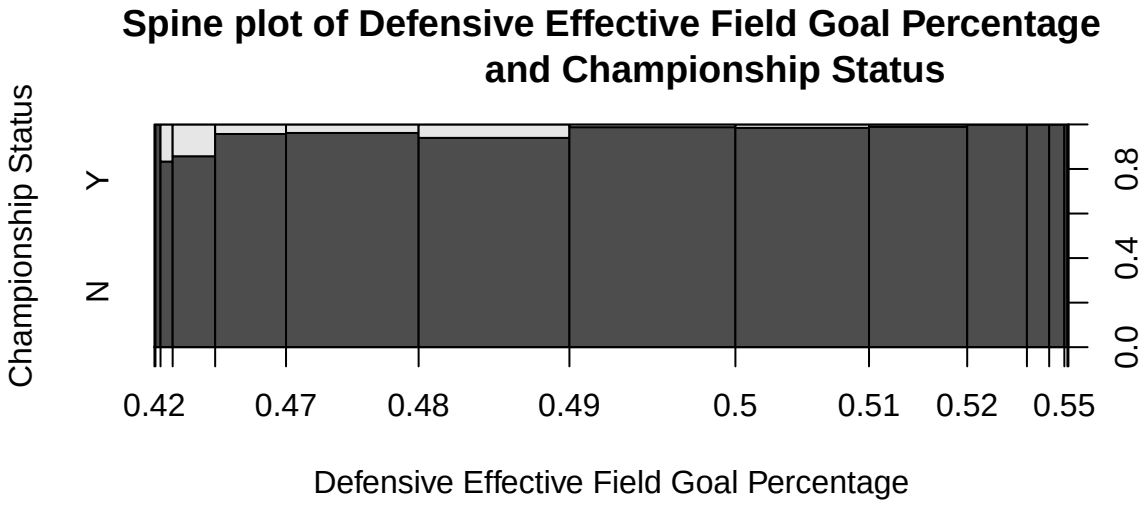
Appendix

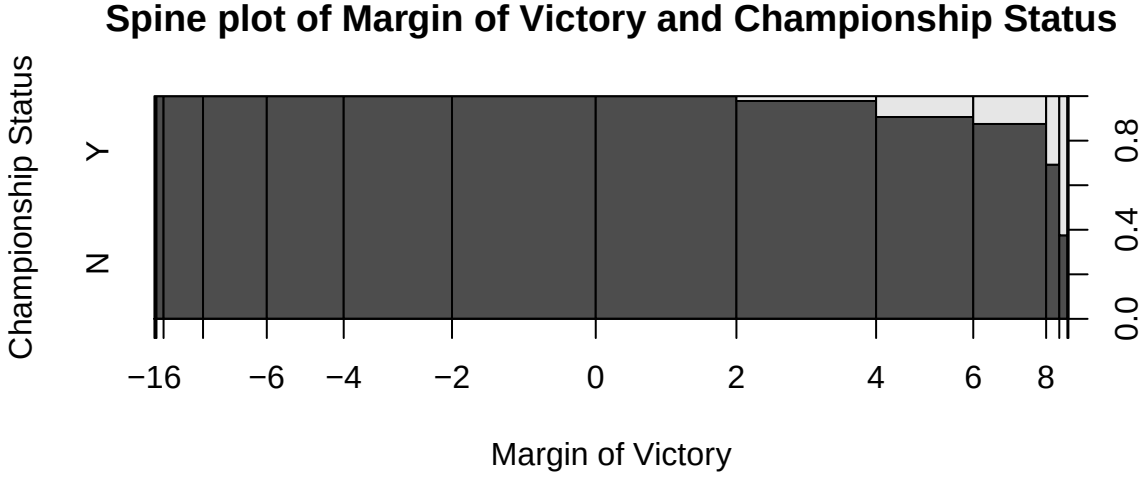

Spine
plot of Win Percentage and Championship Status
Championship Status
Y
N
0.0
0.4
0.8
0.1
0.3
0.4
0.5
0.6
0.7
0.9
Win Percentage


Spine
plot of Net Rating and Championship Status
Championship Status
Y
N
0.0
0.4
0.8
−16
−6
−4
−2
0
2
4
6
8
Net Rating


Spine
plot of SRS and Championship Status
Championship Status
Y
N
0.0
0.4
0.8
−16
−6
−4
−2
0
2
4
6
8
SRS


Spine
plot of Age and Championship Status
Championship Status
Y
N
0.0
0.4
0.8
22
25
26
27
28
29
30
Age





Here are six spine plots of six key variables identified in the visualization of continous variables and our response variable.