Project 1

A. Diaz-Nova

Introduction

I want to start by mentioning how appreciative I am of american soccer analysis website. There isn't a scarcity of data in soccer, rather there this huge inability to use data sets for those like in my position. Free, accessible data is ironically impossible to find without seeing someone scrape the data from the original source. This is why I love what they did with this dataset. American soccer analysis provided an opportunity, that with data in the hands of the public, we can make use of it and push analytics forward. Lastly, here we will explore a metric called goals added or rather (g+).

The aim of the metric is to answer questions like, "when a team is in possession of the ball, how much value did that singular contribution from a player add towards the teams' chances of scoring or conceding a goal?" This also works for defensive actions on the ball as well. Furthermore, this metric has gone through multiple phases, been refined and improved upon, so my hope is to learn from the data set like anyone else who has an interest in the sport.

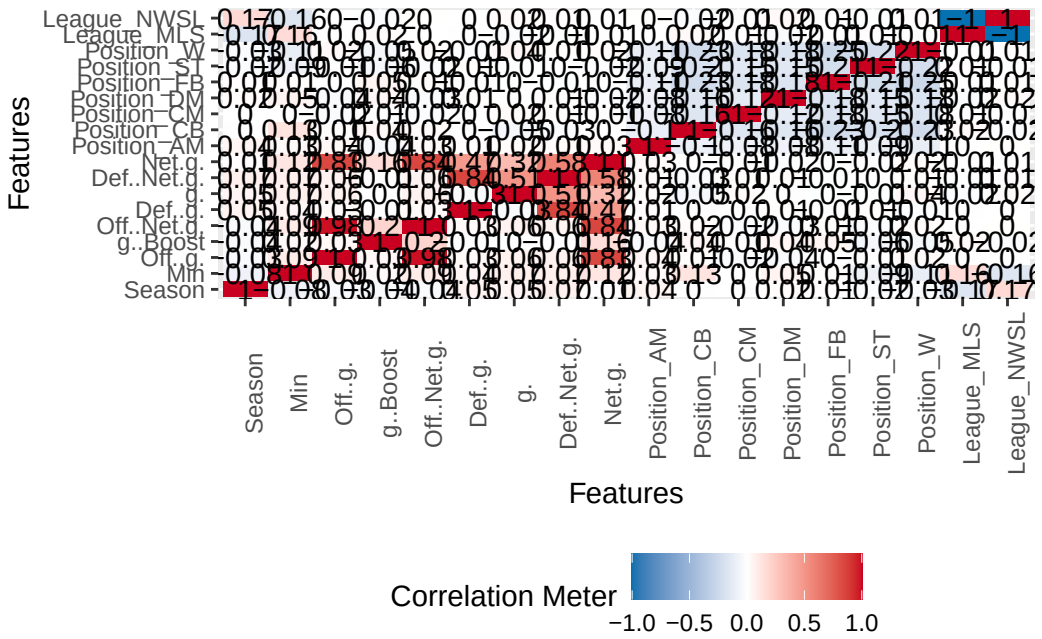
SOURCE: American Soccer Analysis

Well, here is the short summary of the metric, but for our purposes, let's wrap the definition in simple terms. The introduction of the metric was to measure a players action on the ball and lable that as a value. There is too many complexities going on during a game, but this is an attempt to introducing a more hollistic approach, similar to what you may see in the MLB (WAR stat is what I am refering to). In other words, how can we summarize a player's total contributions to their team in one stat.

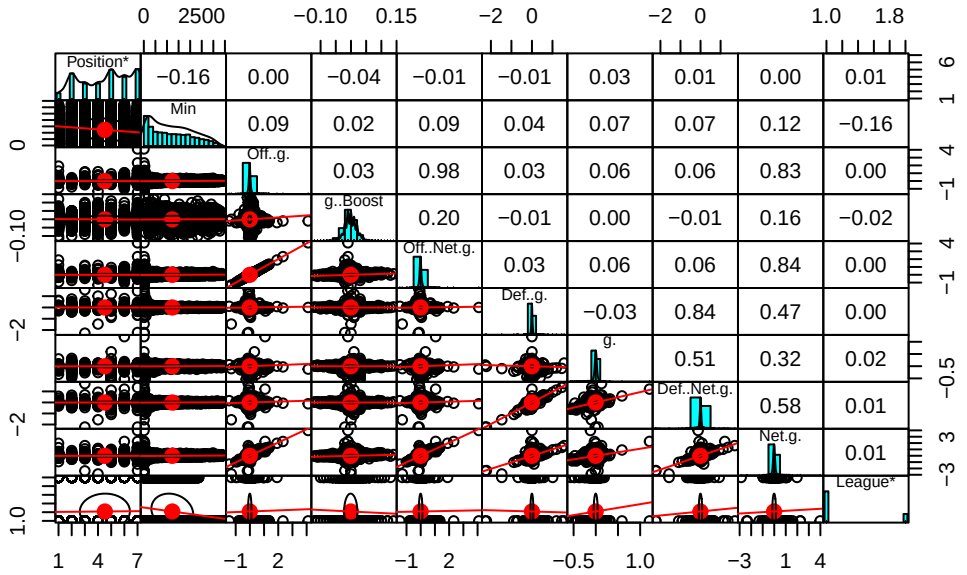
Load in the necessary libraries

```
library(janitor)
library(dplyr)
library(tidyverse)
library(DataExplorer)
library(psych)
```

```
library(knitr)
setwd("C:/Users/Angel/OneDrive/Documents/Datasets")
american_soccer <- read.csv("american_soccer.csv")
plot_correlation(american_soccer)
```




```
pairs.panels(american_soccer[3:12],
             gap = 0,
             pch = 21,
             lm = TRUE)
```



Clean the dataset

From column names to removing empty columns/rows

```
clean_soccer <- american_soccer |>
  clean_names() |>
  remove_empty(whic=c("rows")) |>
  remove_empty(whic=c("cols"))
colnames(clean_soccer)
```

```
[1] "player"    "season"    "position"  "min"       "off_g"     "g_boost"
[7] "off_net_g" "def_g"     "g"         "def_net_g" "net_g"     "league"
```

Establish variables you want to explore to reduce possible traces of ambiguity.

There is close to 8000 observations, so my hope is to narrow down redundancies. For instance, this dataset contains data all the way from 2013. That may not be relevant with what is going on in 2023. My goal, to establish a particular subject matter that I can understand and that my audience can as well.

NOTE TO SELF: Keep in mind, how can I make my visualization interesting for the audience?

What type of variables are in this dataset besides g+?

```
names(clean_soccer)
```

```
[1] "player"    "season"    "position"  "min"       "off_g"     "g_boost"
[7] "off_net_g" "def_g"     "g"         "def_net_g" "net_g"     "league"
```

```
# This command gives us an output of all the variables that are included
```

First thing I want to do is find the mean for min (minutes played) and season (2012-2023)

This will help find out the significant minutes needed to be played and what particular seasons I should look at as well

```
   mean(clean_soccer$min)
```

```
[1] 1227.993
```

```
   median(clean_soccer$min)
```

```
[1] 1117
```

```
   mean(clean_soccer$season)
```

```
[1] 2018.806
```

```
   median(clean_soccer$season)
```

```
[1] 2019
```

Players that are of interest are those who have played over 1100 minutes and data from the the past four seasons

There were two leagues in this dataset, so I went ahead and made a choice to go with the MLS. This will help remove some redundancies.

NOTE TO SELF How can I make my legend meaningful? Think of the necessary mintutes that a player needs to play in a full season to have an impact to their team.

```
soccer <- clean_soccer |>
  filter(min > 1100) |>
  filter(season == "2022")|>
  filter(league == "MLS")
summary(soccer$off_g)
```

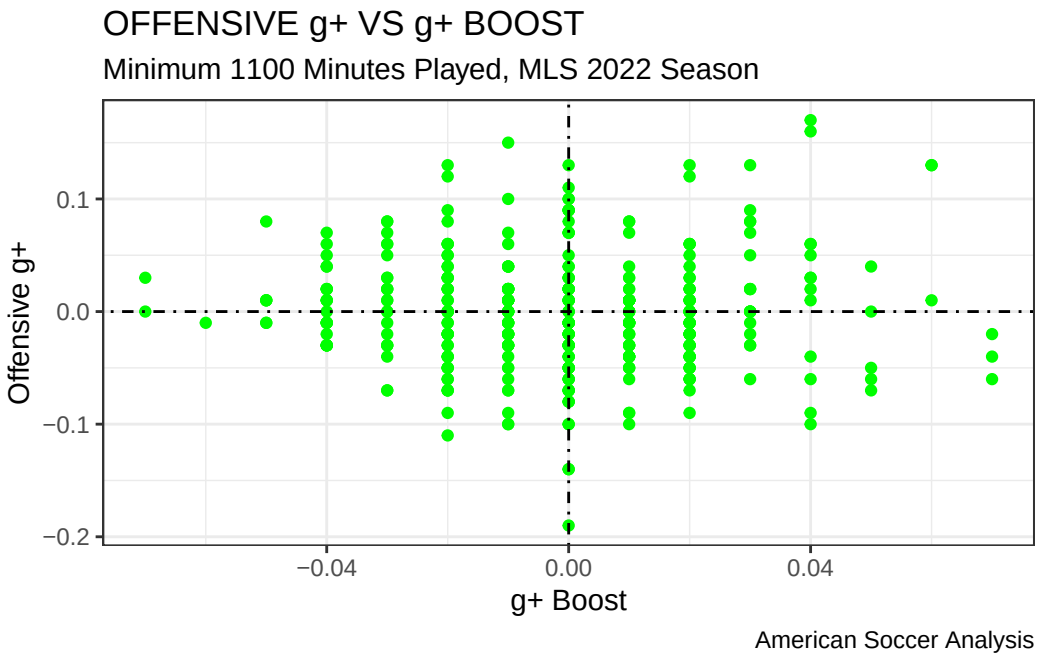
```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-0.190000 -0.030000  0.000000 -0.001387  0.020000  0.170000
```

```
summary(soccer$g_boost)
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-0.07000 -0.02000  0.00000 -0.00096  0.01000  0.07000
```

Scatterplot (Test #1)

```
p1 <- soccer |>
  ggplot(aes(x = g_boost, y = off_g)) +
  geom_point(color = "green")+
  geom_vline(xintercept = 0, linetype = "dotdash")+
  geom_abline(slope=0, linetype = "dotdash")+
  theme_bw()+
  labs(x = "g+ Boost", y = "Offensive g+", title = "OFFENSIVE g+ VS g+ BOOST", subtitle ="
p1
```



The Glue to Piece it All Together

To recap a bit, both variables measure different skill-sets. Offensive g+ considers all attacking touches ON THE BALL, meanwhile g+ boost considers what happens AFTER THE BALL IS PLAYED to a teammate. My goal was to seperate the scatterplot into four quadrants. Each one will be defined for, plus take into account the minutes they played, and factor in the 2022 MLS season as well. To spare the nitty gritty details, and keep it simple, the top right quadrant is what we really want. Well, why would I say that? The right quadrant represents strong attacking players, so players who fall into that category are to be suggested as one of the leagues best, or in their position.

Here is a visual aid