Project 2

A. Diaz-Nova

Introduction

My topic for discussion and statistical analysis will be about the leading team statistics from a league named NWSL(i.e., National Woman Soccer League) which ranges from a collection of multiple seasons (i.e., in terms of what year the competitive season is played, 2016 - 2022). To touch on the subject matter, "NWSL is a twelve-team Division-1 women's professional soccer league featuring national players from around the world." - NWSL website

This dataset is a culmination of important statistics that can have significant insights on a team's performance throughout their respective competitive season. Thus, marks the usage of data through a retrospective lens given statistics like pass percentage, games played, goal differential, and so on. My hope is to find trends in this dataset to see if teams are improving or regressing upon each season.

Load in the necessary libraries

```r
library(dplyr)
library(tidyverse)
library(janitor)
library(psych)
library(RColorBrewer)
library(highcharter)
library(ggplot2)
```

Load in the data

```
# Set your working directory
setwd("C:/Users/Angel/OneDrive/Documents/Datasets")

# Use the read.csv function to load in the csv file, and name it nwsl
nwsl <- read.csv("nwsl-team-stats.csv")

# Use the head function to see the first 6 rows and the variables
head(nwsl)
```

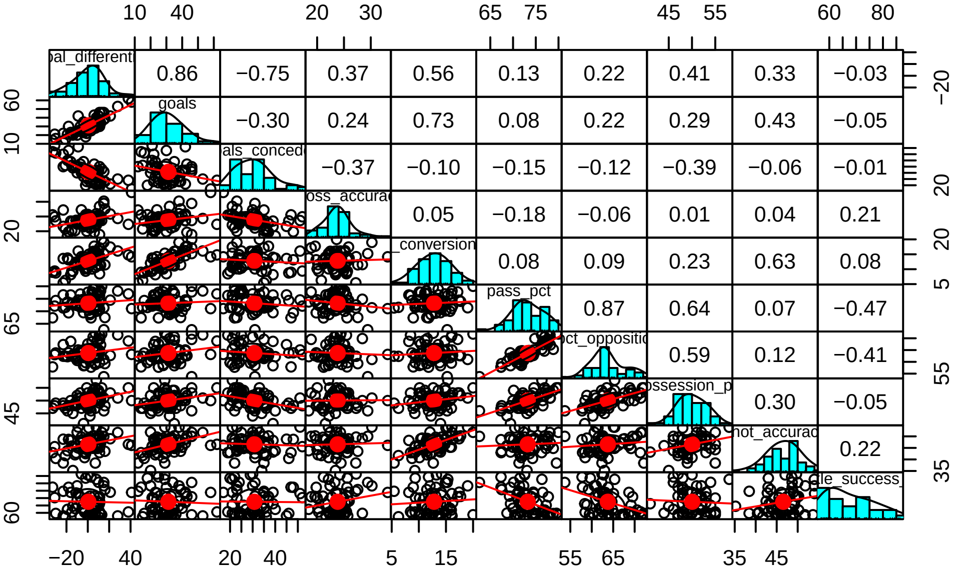
```
          team_name season games_played goal_differential goals goals_conceded
1   Boston Breakers   2016           20               -33    14             47
2   Boston Breakers   2017           24               -11    24             35
3 Chicago Red Stars   2016           21                 3    25             22
4 Chicago Red Stars   2017           25                 2    33             31
5 Chicago Red Stars   2018           25                 8    38             30
6 Chicago Red Stars   2019           26                10    42             32
  cross_accuracy goal_conversion_pct pass_pct pass_pct_opposition_half
1          25.57                8.97    67.38                    57.86
2          23.70               12.37    72.53                    61.42
3          21.19               11.79    67.35                    57.74
4          21.08               13.10    69.23                    61.52
5          25.96               13.67    71.63                    64.55
6          23.53               14.53    71.97                    64.69
  possession_pct shot_accuracy tackle_success_pct
1             47         42.95              77.42
2             48         42.78              73.49
3             46         48.58              84.32
4             47         49.60              71.29
5             51         45.68              67.97
6             50         47.06              67.51
```

NOTE: The data above has a column for the name of a team and the rest are percentages of various categorized events during a competitive game

Cleaning the dataset

Testing dataset for correlation

```
# pairs.panels() shows a scatterplot of matrices, and helps in our case of a small dataset
pairs.panels(nwsl[4:13],
             gap = 0,
             pch = 21,
             lm = TRUE)
```



From the presented visualization, I see a strong relationship that consist of goals and goal_differential.

Here's the catch, goal_differential description is Goals Scored - Goals conceded. Let us use other predictor variables to build a linear or multiple regression analysis

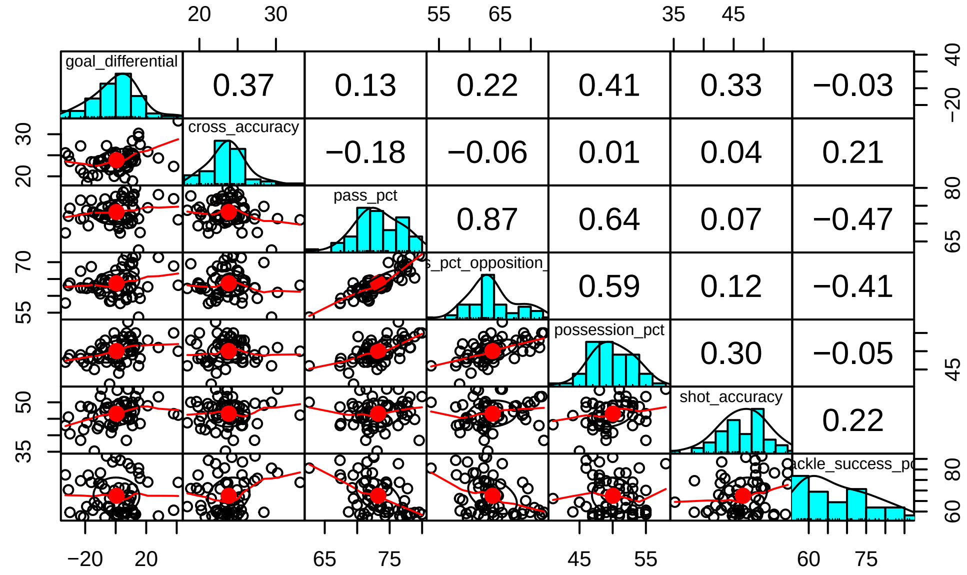
New subset for pairs.panel

```
cor_nwsl <- nwsl |>
  select(goal_differential, cross_accuracy, pass_pct, pass_pct_opposition_half, possession

head(cor_nwsl)
```

```
  goal_differential cross_accuracy pass_pct pass_pct_opposition_half
1               -33          25.57    67.38                    57.86
2               -11          23.70    72.53                    61.42
3                 3          21.19    67.35                    57.74
4                 2          21.08    69.23                    61.52
5                 8          25.96    71.63                    64.55
6                10          23.53    71.97                    64.69
  possession_pct shot_accuracy tackle_success_pct goal_conversion_pct
1             47         42.95              77.42                8.97
2             48         42.78              73.49               12.37
3             46         48.58              84.32               11.79
4             47         49.60              71.29               13.10
5             51         45.68              67.97               13.67
6             50         47.06              67.51               14.53
```

```
pairs.panels(cor_nwsl[1:7],
             gap = 0,
             pch = 21)
```



There is a strong relationship between pass_pct and pass_pct_opposition_half, which makes sense, the higher quality of passes, the easier to maintain the ball in possession when inside of the opposition half.

Multiple Regression Analysis

```
fit1 <- lm(data = nwsl, goal_differential ~ possession_pct + cross_accuracy + pass_pct + p
summary(fit1)
```

```
Call:
lm(formula = goal_differential ~ possession_pct + cross_accuracy +
    pass_pct + pass_pct_opposition_half + shot_accuracy + tackle_success_pct +
    goal_conversion_pct, data = nwsl)

Residuals:
     Min       1Q   Median       3Q      Max
-23.6409  -5.0157   0.6636   5.1623  24.4037

Coefficients:
```

```
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)              -61.6594    45.8708  -1.344  0.18483    
possession_pct             2.1071     0.6417   3.284  0.00185 ** 
cross_accuracy             1.8022     0.5373   3.354  0.00151 ** 
pass_pct                  -1.8035     0.9636  -1.872  0.06701 .  
pass_pct_opposition_half   0.8375     0.7326   1.143  0.25828    
shot_accuracy             -0.4226     0.4777  -0.885  0.38039    
tackle_success_pct        -0.3357     0.2009  -1.671  0.10081    
goal_conversion_pct        2.7190     0.6008   4.526 3.63e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.55 on 51 degrees of freedom
Multiple R-squared:  0.5683,	Adjusted R-squared:  0.509 
F-statistic: 9.589 on 7 and 51 DF,  p-value: 1.522e-07
```

Which p-values are least significant that can be removed to improve the model? By looking at the output, shot_accuracy and pass_pct_opposition_half seem the least useful in this case (i.e., higher p-value).

The R-squared value: 0.5683, in other words, about 57% approximately of the variation in observations can be explained by this model, but maybe we can improve it.

Diagnostic Plots

FAULTY CODE:

autoplot(fit1)

ERROR:

"processing file: Project-2.rmarkdown |……………………. | 52% [unnamed-chunk-8] Quitting from lines 104-105 [unnamed-chunk-8] (Project-2.rmarkdown) Error in `autoplot()`: ! Objects of class are not supported by autoplot. have you loaded the required package? Backtrace: 1. ggplot2::autoplot(fit1) 2. ggplot2:::autoplot.default(fit1)"

Rendering the document would not let me include my diagnostic plots but the findings are still below!

1) A linear model is appropriate - blue line is relatively horizontal
2) Distribution looks relatively normal
3) Influential observations may be skewing the variance distribution, keep an eye out for the odd one of the group

4) Outliers who have high leverage are no good for your model, which causes problematic headaches, but we can remove those observations if must

Simplifying further

```
fit2 <- lm(data = nwsl, goal_differential ~ possession_pct + cross_accuracy + pass_pct + t
summary(fit2)
```

```

Call:
lm(formula = goal_differential ~ possession_pct + cross_accuracy +
    pass_pct + tackle_success_pct + goal_conversion_pct, data = nwsl)

Residuals:
     Min       1Q   Median       3Q      Max
-23.3259  -6.3146  -0.0767   4.6246  26.3880

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         -80.9589    43.5179  -1.860  0.06839 .
possession_pct        2.0461     0.6308   3.244  0.00205 **
cross_accuracy        1.9333     0.5265   3.672  0.00056 ***
pass_pct             -0.9802     0.6432  -1.524  0.13349
tackle_success_pct   -0.3860     0.1956  -1.974  0.05364 .
goal_conversion_pct   2.4096     0.4764   5.058 5.41e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.54 on 53 degrees of freedom
Multiple R-squared:  0.5524,    Adjusted R-squared:  0.5102
F-statistic: 13.08 on 5 and 53 DF,  p-value: 2.639e-08
```

This model's adj r-squared has lowered, which is not a good sign. This means choosing to remove "pass_pct_opposition_half" and "shot_accuracy" was not the best option.

To cap off, the first model is a better model overall because of the adj r-squared value being higher. The equation is put below and is a reasonable form of prediction for goal differential in a competitive NWSL season.

fit1: goal_differential = 2.107(possession_pct) + 1.802(cross_accuracy) - 1.804(pass_pct) + 0.838(pass_pct_opposition_half) - 0.423(shot_accuracy) + 2.719(goal_conversion_pct)

```
anova(fit2,fit1)
```

```
Analysis of Variance Table

Model 1: goal_differential ~ possession_pct + cross_accuracy + pass_pct +
    tackle_success_pct + goal_conversion_pct
Model 2: goal_differential ~ possession_pct + cross_accuracy + pass_pct +
    pass_pct_opposition_half + shot_accuracy + tackle_success_pct +
    goal_conversion_pct
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     53 5883.2
2     51 5675.2  2    208.02 0.9347 0.3993
```

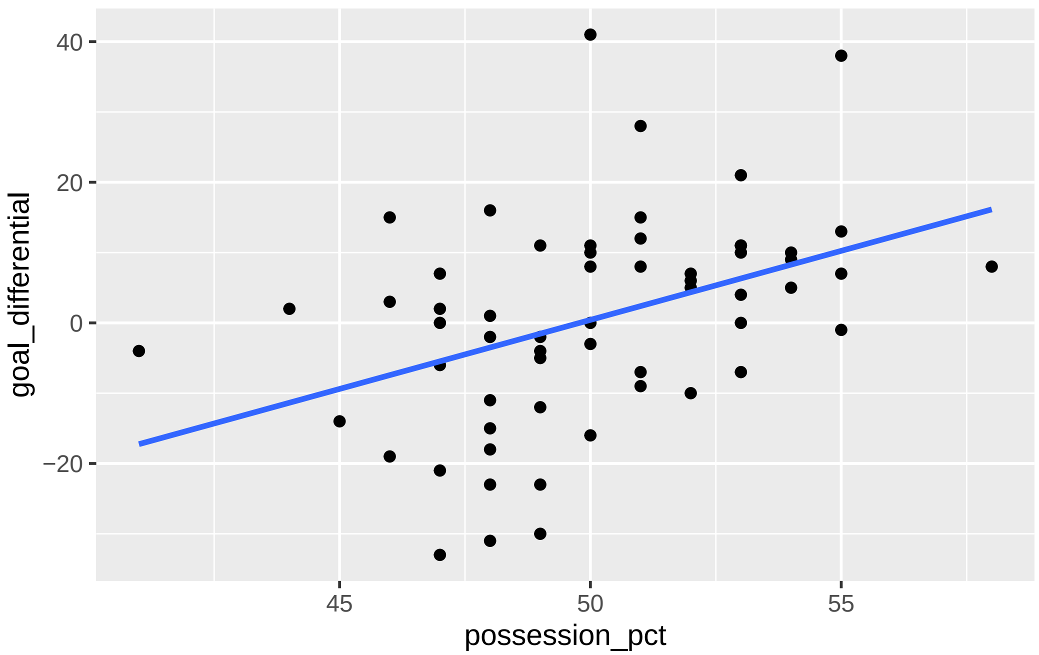
More proof that the first fit is a better overall model when in the context of a multiple regression analysis.

Plots to test correlation

Possession_pct has the lowest p-value in the regression analysis, maybe something worth exploring, given the unadequate adj-Rsquare value.

```
p1 <- nwsl |>
  ggplot(aes(x = possession_pct, y = goal_differential)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
p1
```

```
`geom_smooth()` using formula = 'y ~ x'
```



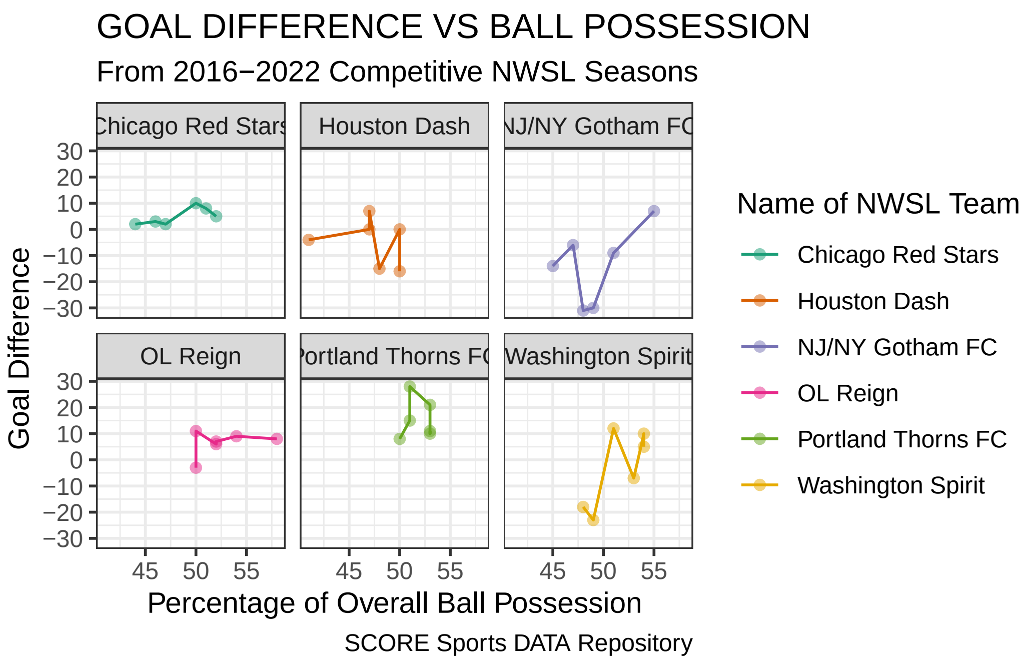
A plot to test the linear regression model of possession_pct and goal_differential. Found good insights, for two positive reasons. One there is a correlation and the slope is increasing, which is crucial when you consider the value of goal differential. Secondly, how conclusive can we make our interpretation of the correlation when the variance of the distance between points and the 'lm' line is relatively moderate. That there is a slight correlation.

```
p2 <- nwsl |>

# Filtering all teams that have high observations, n = 6
  filter(team_name %in% c('Chicago Red Stars', 'Houston Dash', 'NJ/NY Gotham FC', 'OL Reig

# First plot with ggplot() and setting aesthetics
  ggplot(aes(x = possession_pct, y = goal_differential, color = team_name)) +
  geom_point(alpha = 0.5) +
  geom_line() +

# Make a plot for every team by facet_wrap function
  facet_wrap(~team_name) +
  theme_bw() +
  labs(x = "Percentage of Overall Ball Possession", y = "Goal Difference", title = "GOAL D
  scale_color_brewer(palette = "Dark2")
p2
```



Portland Thorns FC seem to perform the most consistent given the impressive goal difference record of over +10 and a heavy ball possession side (i.e, greater than 50%).

```
nwsl |>
  group_by(team_name) |>
  count(team_name) |>
  arrange(desc(n))
```

```
# A tibble: 16 x 2
# Groups:   team_name [16]
   team_name                  n
   <chr>                  <int>
 1 Chicago Red Stars          6
 2 Houston Dash               6
 3 NJ/NY Gotham FC            6
 4 OL Reign                   6
 5 Portland Thorns FC         6
 6 Washington Spirit          6
 7 North Carolina Courage     5
 8 Orlando Pride              5
 9 Boston Breakers            2
10 FC Kansas City             2
```

```
11 Kansas City Current        2
12 Racing Louisville FC       2
13 Utah Royals FC             2
14 Angel City FC              1
15 San Diego Wave FC          1
16 Western New York Flash     1
```

The top six teams have the most observations to test how they progressed throughout each season in respect of goal difference.

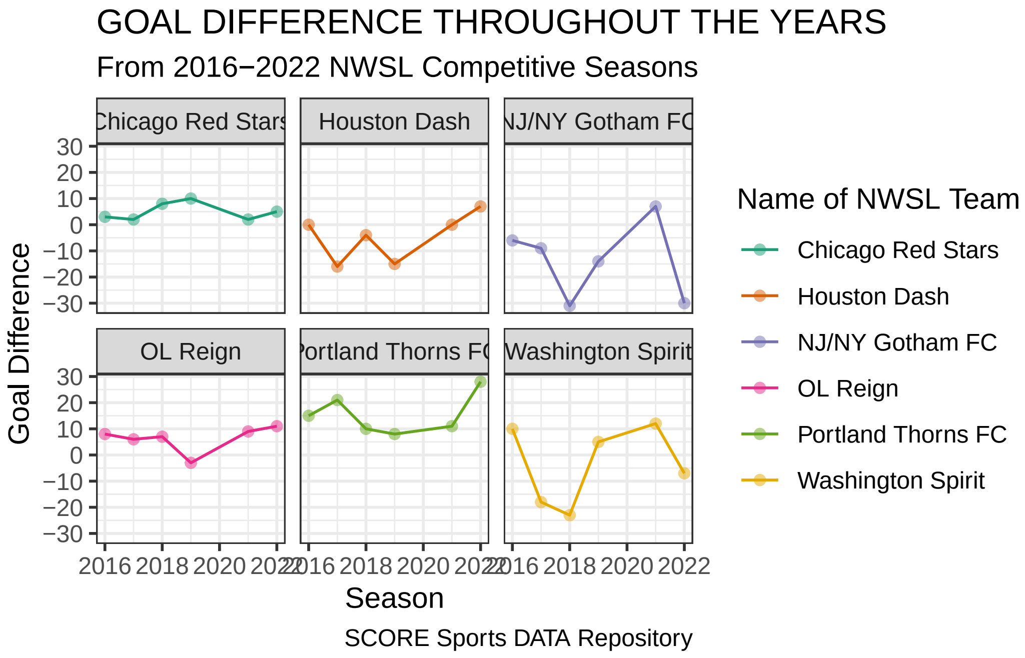
```
p3 <- nwsl |>

# Filtering all teams that have high observations, n = 6
  filter(team_name %in% c('Chicago Red Stars', 'Houston Dash', 'NJ/NY Gotham FC', 'OL Reig

# First plot with ggplot() and setting aesthetics
  ggplot(aes(x = season, y = goal_differential, color = team_name)) +
  geom_point(alpha = 0.5) +
  geom_line() +

# Make a plot for every team by facet_wrap function
  facet_wrap(~team_name) +
  theme_bw() +
  labs(x = "Season", y = "Goal Difference", title = "GOAL DIFFERENCE THROUGHOUT THE YEARS"
  scale_color_brewer(palette = "Dark2")

p3
```



As mentioned before, Portland Thorns FC are a strong competitor in the league, when you consider the minimal fluctuation over the years. In other words, this shows a relatively normal consistency from a team of their calibre, which in itself is difficult when playing at the highest level of their profession.

Final plot

FAULTY CODE BELOW

p4 <- nwsl |>

Filtering all teams that have high observations, n = 6

filter(team_name %in% c('Chicago Red Stars', 'Houston Dash', 'NJ/NY Gotham FC', 'OL Reign', 'Portland Thorns FC', 'Washington Spirit')) |>

First plot with ggplot() and setting aesthetics

highchart() |> hc_add_series(data = nwsl, type = "line", hcaes(x = season, y = goal_differential, group = team_name)) |> hc_colors("Dark2") #facet_wrap(~team_name) p4 “‘

I found it too challenging to insert a highcharter with a tooltip in respect for each team. That was my plan for the last plot. It would consist of all other statistics for each point, which would then be a better representation overall of team performance throughout the years.

Conclusion

By default, the most influential indicator of a team’s ability is goal differential (i.e., goals scored - goals conceded), regardless of the other team performance statistics. In defense, a high ‘GD’ consists of scoring tons of shots and conceding few shots as possible, which ultimately the challenge of the sport. Given the background research(e.g., from the works cited page), I wanted to explore what aspects, or should I say, which statistics are relevant predictor variables of goal differential.

What did I find? Firstly, I used a multiple regression analysis of multiple predictor variables to test a response variable in goal_differential. My findings were not compelling enough, even though 0.57 or 57% of the model can explain the variance (bad wording) of the data. In the end, I moved onto a linear regression model for one singular explanatory variable in possession_pct to test for correlation like many attempts before.

The findings were promising, higher possession seem to indicate or predict better numbers in goal difference (goals scored- goals conceded). Perhaps not the most profound in terms of significance, but was interesting to explore nonetheless.

Works Cited

Yurko, R. (2023, March 18). National Women’s Soccer League team statistics. SCORE Sports Data Repository - National Women’s Soccer League Team Statistics. https://data.scorenetwork.org/soccer/nwsl-team-stats.html