Title: Assignment 1

Author: Travis Ho

2023-09-29

Setup

```
#Load libraries
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggplot2)
library(patchwork)
```

Data

```
#Load college football stats dataset
df <- read.csv("./data/cfb22.csv")
dim(df)
```

```
## [1] 131 151
```

This dataset was obtained from the Kaggle dataset called "College Football Team Stats Seasons 2013 to 2022." This is a dataset containing offensive and defensive statistics from all college football teams in 2022. There are 131 football teams. There are 151 different statistics including but not limited to points scored, points allowed, turnovers, and games played.

Research Question

Previous research has shown that four of the most important indicators of wins in college football are explosiveness, efficiency, finishing drives, and turnovers. I aim to investigate which of these four factors is most indicative of performance.

Hypotheses

- H0: Performance is not well correlated with any of the four variables.
- HA: Performance relates to at least one of the variables and is most strongly related to efficiency.

Variables of Interest

The main variables that will be investigated are variables that measure explosiveness, efficiency, finishing drives, and turnovers. For explosiveness, Offensive Yards per Play (Off.Yards.Play) will be explored. This is a continuous variables that would be in the range of 0 yards to 100 yards. For efficiency, 3rd down conversion rate (X.3rd.Percent) will be looked at. This is a continuous variable ranging from 0.000 to 1.000. For finishing drives, the variables red zone attempts (Redzone.Attempts) and red zone scores (Redzone.Scores) will be inspected. The red zone is the area where the ball is within 20 yards of scoring a touchdown. These variables are continuous and the range is all whole numbers. Finally, for turnovers, turnover margin (Turnover.Margin) will be studied. This is just how many more turnovers a team forced than another team. This is a continuous variable that's range is all integers.

In order to measure performance, point differential per game will be investigated. This will be done by taking the difference between points scored and points allowed. This variable displays how many more points a team averages per game than their opponent. This variable is continuous and on the range of all real numbers. Next, the variables Games and Wins will be looked at. These are both continuous variables that range from 0 to 15. Finally, the variable team will be investigated. This is a categorical variable that just details the team name and conference the team is in.

Data Wrangling

```
#Create a new dataframe with only the variables we need
 stats <- select(df,
                 Team,
                 Games,
                 Wins,
                 Off.Yards.Play,
                 X3rd.Percent,
                 Redzone.Attempts,
                 Redzone.Scores,
                 Turnover.Margin,
                 Points.Allowed,
                 Total.Points
                 )
dim(stats)
```

```
## [1] 131  10
```

```
#Calculate the percent of scores in redzone and add to stats dataframe
stats$Redzone.Percent <- stats$Redzone.Scores / stats$Redzone.Attempts
dim(stats)
```

```
## [1] 131  11
```

```r
#Calculate win percentage and add to stats dataframe
stats$Win.Percent <- stats$Wins / stats$Games
```

```r
#Calculate point differential per game and add to stats dataframe
stats$Point_differential <- (stats$Total.Points - stats$Points.Allowed) / stats$Games
```

```r
#create a new dataframe that just has the variables needed for correlation analysis
subset_stats <- select(stats,
                       Off.Yards.Play,
                       X3rd.Percent,
                       Redzone.Percent,
                       Turnover.Margin,
                       Point_differential)
```

```r
#Check to see if stats is a dataframe before analysis
is.data.frame(stats)
```

```
## [1] TRUE
```

Using select, I created a new dataframe called stats that contains only the variables that are needed for the analysis of explosiveness, efficiency, finishing of drives, and turnovers.

Analysis

```r
#Check out the summary stats for main variables being investigated
subset_stats %>%
  summary()
```

```
##  Off.Yards.Play   X3rd.Percent    Redzone.Percent  Turnover.Margin
##  Min.   :3.940   Min.   :0.2200   Min.   :0.5217   Min.   :-19.0000
##  1st Qu.:5.230   1st Qu.:0.3605   1st Qu.:0.7984   1st Qu.: -4.0000
##  Median :5.700   Median :0.3920   Median :0.8400   Median :  1.0000
##  Mean   :5.701   Mean   :0.3947   Mean   :0.8343   Mean   :  0.5725
##  3rd Qu.:6.170   3rd Qu.:0.4330   3rd Qu.:0.8794   3rd Qu.:  5.5000
##  Max.   :7.280   Max.   :0.5680   Max.   :0.9759   Max.   : 22.0000
##  Point_differential
##  Min.   :-29.083
##  1st Qu.: -5.042
##  Median :  1.923
##  Mean   :  1.808
##  3rd Qu.:  8.577
##  Max.   : 26.800
```

```r
#Get correlations of 4 main variables
cor(stats[,c("Point_differential", "Off.Yards.Play")])
```

```
##                    Point_differential Off.Yards.Play
## Point_differential          1.0000000      0.7082324
## Off.Yards.Play              0.7082324      1.0000000
```

```r
cor(stats[,c("Point_differential", "X3rd.Percent")])
```

```
##                     Point_differential X3rd.Percent
## Point_differential          1.0000000    0.6334244
## X3rd.Percent                0.6334244    1.0000000
```

```r
cor(stats[,c("Point_differential", "Redzone.Percent")])
```

```
##                     Point_differential Redzone.Percent
## Point_differential           1.000000        0.434518
## Redzone.Percent              0.434518        1.000000
```

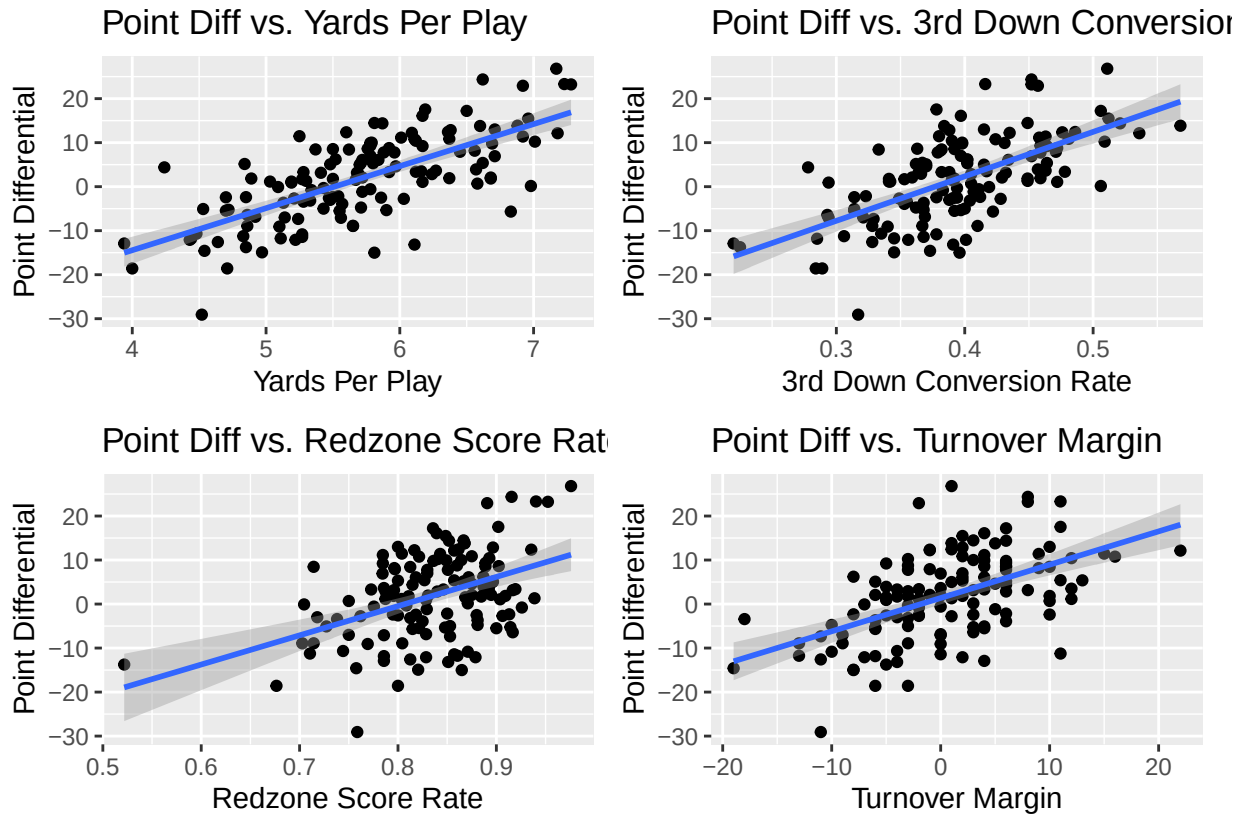
```r
cor(stats[,c("Point_differential", "Turnover.Margin")])
```

```
##                     Point_differential Turnover.Margin
## Point_differential          1.0000000       0.5376137
## Turnover.Margin             0.5376137       1.0000000
```

Visualization

```r
#Plot Explosiveness
explosive_plot <- ggplot(stats, aes(x=Off.Yards.Play, y=Point_differential)) +
                  geom_point()+
                  geom_smooth(method=lm)+
                  labs(y='Point Differential',x='Yards Per Play',
                       title='Point Diff vs. Yards Per Play')

#Plot Efficiency
efficiency_plot <- ggplot(stats, aes(x=X3rd.Percent, y=Point_differential)) +
                   geom_point()+
                   geom_smooth(method=lm)+
                   labs(y='Point Differential',x='3rd Down Conversion Rate',
                        title='Point Diff vs. 3rd Down Conversion')

#Plot Finishing Drives
finishing_plot <- ggplot(stats, aes(x=Redzone.Percent, y=Point_differential)) +
                  geom_point()+
                  geom_smooth(method=lm)+
                  labs(y='Point Differential',x='Redzone Score Rate',
                       title='Point Diff vs. Redzone Score Rate')

#Plot Turnovers
turnover_plot <- ggplot(stats, aes(x=Turnover.Margin, y=Point_differential)) +
                 geom_point()+
                 geom_smooth(method=lm)+
                     labs(y='Point Differential',x='Turnover Margin',
                          title='Point Diff vs. Turnover Margin')

(explosive_plot + efficiency_plot) / (finishing_plot + turnover_plot)
```

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

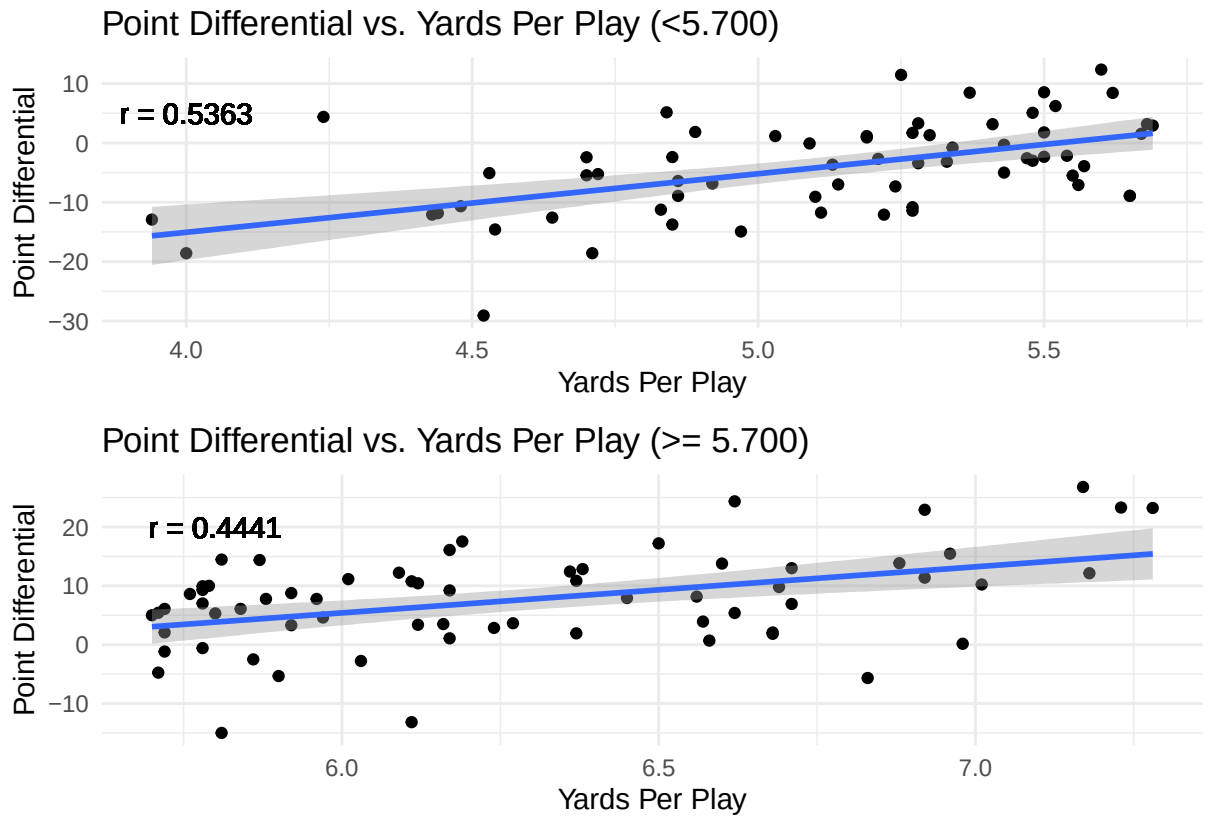


As a result of explosiveness seemingly being the most highly correlated variable with point differential, we will look further into yards per play and its relationship with point differential We will split up yards per play into two section: higher than or equal to the median (5.700) and lower than the median.

```
#Separate yards per play into two sections
lower_yardspp <- stats %>%
  filter(stats$Off.Yards.Play < 5.700)
higher_yardspp <- stats%>%
  filter(stats$Off.Yards.Play >= 5.700)

#Get correlation of higher and lower sections
lowcor <- round(cor(lower_yardspp$Off.Yards.Play,lower_yardspp$Point_differential,method = c("pearson"))
highcor <- round(cor(higher_yardspp$Off.Yards.Play,higher_yardspp$Point_differential,method = c("pearso

#Plot the lower section
lower_yardspp_plot <- ggplot(lower_yardspp, aes(x=Off.Yards.Play, y=Point_differential)) +
                    geom_point()+
                    geom_text(x = 4.0, y = 5,label = paste0('r = ', lowcor))+
                    geom_smooth(method=lm)+
                    labs(y='Point Differential',x='Yards Per Play',
                         title='Point Differential vs. Yards Per Play (<5.700)')+
```

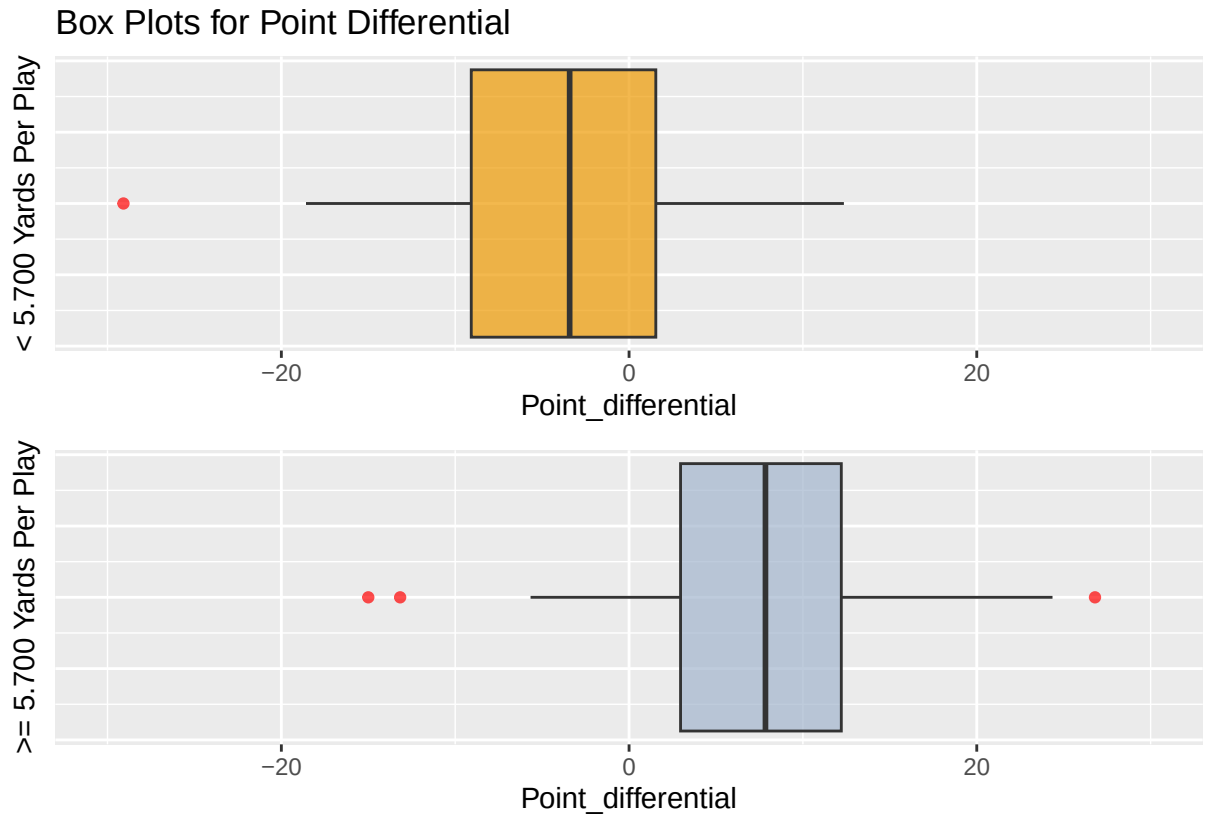
```
                    theme_minimal()

#Plot the higher section
higher_yardspp_plot <- ggplot(higher_yardspp, aes(x=Off.Yards.Play, y=Point_differential)) +
                    geom_point()+
                    geom_text(x = 5.8, y = 20,label = paste0('r = ', highcor))+
                    geom_smooth(method=lm)+
                    labs(y='Point Differential',x='Yards Per Play',
                         title='Point Differential vs. Yards Per Play (>= 5.700)')+
                    theme_minimal()

lower_yardspp_plot / higher_yardspp_plot
```

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

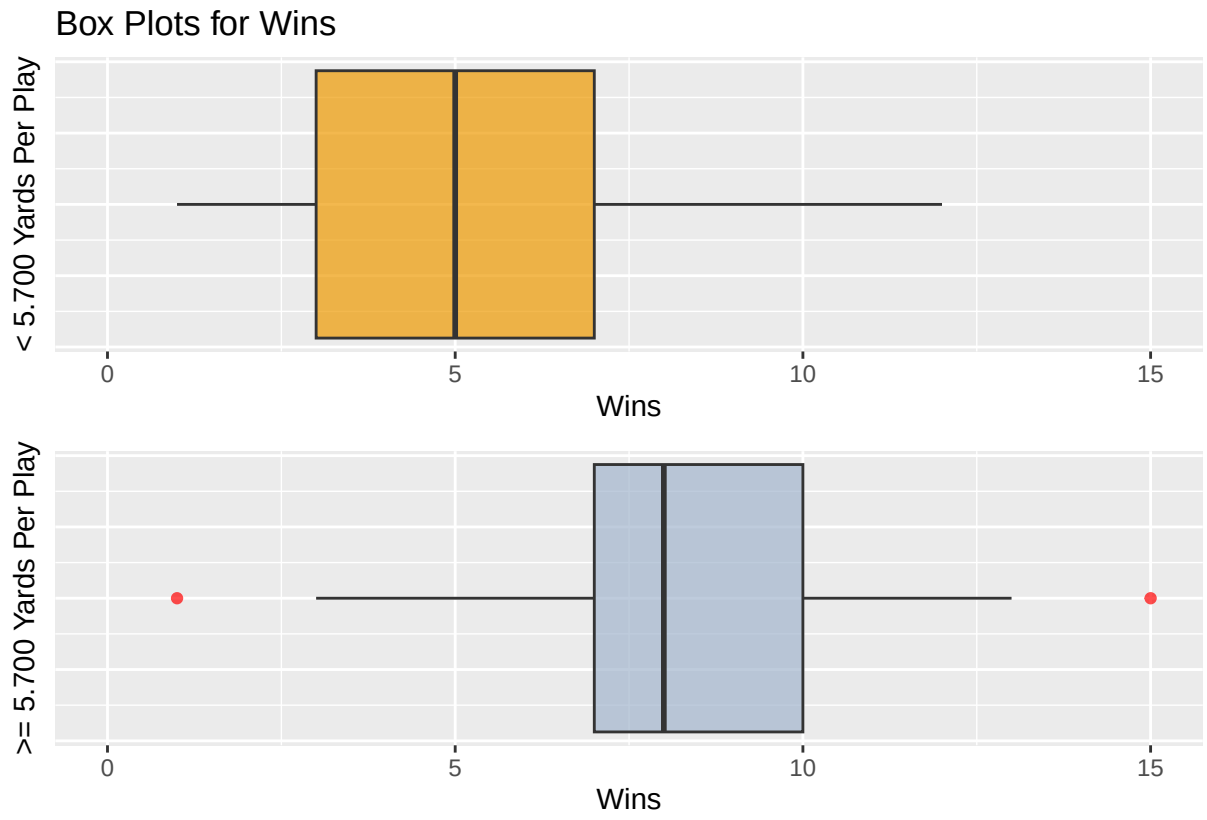



```
#Create boxplot for lower section and point differential
lower_boxplot <- ggplot(data = lower_yardspp, aes(x = Point_differential)) +
            geom_boxplot(fill='orange2',alpha=0.7,outlier.colour="red") +
            labs(y = "< 5.700 Yards Per Play",
                 title = "Box Plots for Point Differential ")+
            coord_cartesian(xlim=c(-30,30))+
            theme(axis.text.y = element_blank(), axis.ticks.y = element_blank())
```

```
#Create boxplot for higher section and point differential
higher_boxplot <- ggplot(data = higher_yardspp, aes(x = Point_differential)) +
            geom_boxplot(fill='lightsteelblue3',alpha=0.7,outlier.colour="red") +
            labs(y = ">= 5.700 Yards Per Play")+
            coord_cartesian(xlim=c(-30,30))+
            theme(axis.text.y = element_blank(), axis.ticks.y = element_blank())

lower_boxplot / higher_boxplot
```




```
#Create boxplot for lower section and wins
lower_boxplot_win <- ggplot(data = lower_yardspp, aes(x = Wins)) +
            geom_boxplot(fill='orange2',alpha=0.7,outlier.colour="red") +
            labs(y = "< 5.700 Yards Per Play",
                 title = "Box Plots for Wins ")+
            coord_cartesian(xlim=c(0,15))+
            theme(axis.text.y = element_blank(), axis.ticks.y = element_blank())

#Create boxplot for higher section and wins
higher_boxplot_win <- ggplot(data = higher_yardspp, aes(x = Wins)) +
            geom_boxplot(fill='lightsteelblue3',alpha=0.7,outlier.colour="red") +
            labs(y = ">= 5.700 Yards Per Play")+
            coord_cartesian(xlim=c(0,15))+
            theme(axis.text.y = element_blank(), axis.ticks.y = element_blank())

lower_boxplot_win / higher_boxplot_win
```



Discussion

The research question that we initially asked was which of the four variables (explosiveness, efficiency, finishing of drives, and turnovers) had the strongest relationship with performance. The correlation analysis of each of the four variables in relation to point differential indicate that explosiveness (Yards Per Play) correlated the strongest with performance correlation of (r = 0.718). The next highest correlation was with efficiency (3rd down conversion percentage) with a correlation of r = 0.633. Third was Turnovers with a correlation of 0.538. Finally, the variable with the weakest relation with performance was finishing of drives (Red Zone Percentage) with a correlation of r = 0.434. The strength in the relationship between performance and the four variables can be visualized in the scatterplots. Yards per play vs. point differential most closely follows the linear regression line and the confidence interval is the more narrow on that plot.

Digging deeper into the relationship between explosiveness and performance, three different plot were created. In doing so, we split the yards per play dataset across the media of (5.700 yards). Fifty percent of the data was above or equal to 5.700 yards per play and fifty percent of the data was below 5.700 yards per play. For the first plot that was generated, two scatterplots were created to check the correlation between yards per play and point differential. In the scatterplot that had data from the lower threshold of less than 5.700 yards per play, the correlation was 0.5363. Meanwhile, the scatterplot containing the data from the upper threshold of more than or equal to 5.700 yards per play, the correlation was lower at 0.4441. Neither of these sections suggests a strong relationship between the cutoffs in yards per play and performance. However, combined, explosiveness and performance do follow a strong correlation. The next plot that was examined was a boxplot that looked at yards per play and point differential. In this, it is evident that as the yards per play increases, so does the point differential. The difference is so great that the first quartile of the boxplot (about 3 points) in the higher benchmark is greater than the third quartile of the boxplot (about 2 points) of the lower benchmark. These finding were consistent with the third plot that was created, a

boxplot that compared yards per play and another indicator of performance, wins. In these boxplots, the same findings were found. A higher yards per play indicated a greater amount of wins. In the higher yards per play boxplot, the first quartile was about 7 wins while in the lower yards per play boxplot the third quartile was about 7 wins as well.

In relation to the initial hypotheses, the data suggest that one of the four variables out of explosiveness, efficiency, finishing of drives, and turnovers does correlate well with performance. However, instead of efficiency being the variable that is more strongly related to performance, explosiveness is.

Citations

Teach how to use patchwork library

https://patchwork.data-imaginist.com/

Dataset that is used

https://www.kaggle.com/datasets/jeffgallini/college-football-team-stats-2019?select=cfb22.csv

Removing Y axis from boxplots

https://ggplot2.tidyverse.org/articles/faq-axes.html#:~:text=Remove%20x%20or%20y%20axis,to%20remove%20to%20elem

How to display r to scatterplot

https://stackoverflow.com/questions/70494677/how-do-i-display-a-correlation-coefficient-in-a-scatterplot