Project 1

Cristian Sigala

10/16/2019

2018 NFL Season

For this firest project, I wanted to anaylize data on the NFL. Ever since I was a child, I have always been invested in the NFL. I love football and the amount of statistics that are involed in it. In these two datasets, I have all 2018 NFL season games with a variety of stats, such as scores, win probability, ELO rating, and much more all coming from FiveThirtyEight. In the next dataset, I have the weather patterns in the respective stadiums at the time of play. This dataset contains wind mph, wind direction, temperature, and stadium. Some potential associations I can encounter are that certain teams have an optimal temerature/windspeed that gives them the most wins along with which stadium aquire the most points.

Data

```r
weather <- read.csv("Weather NFL.csv")
season18 <- read.csv("2018 NFL season.csv")
glimpse(season18)
```

```
## Rows: 267
## Columns: 30
## $ date          <fct> 9/6/18, 9/9/18, 9/9/18, 9/9/18, 9/9/18, 9/9/18, 9/9/...
## $ season        <int> 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018...
## $ neutral       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ playoff       <fct> , , , , , , , , , , , , , , , , , , , , , , , , , ,
## $ team1         <fct> PHI, NYG, NO, MIA, BAL, MIN, IND, CLE, NE, LAC, CAR,...
## $ team2         <fct> ATL, JAX, TB, TEN, BUF, SF, CIN, PIT, HOU, KC, DAL, ...
## $ elo1_pre      <dbl> 1646.786, 1411.677, 1584.490, 1449.871, 1535.270, 16...
## $ elo2_pre      <dbl> 1600.640, 1534.818, 1469.482, 1495.841, 1502.217, 14...
## $ elo_prob1     <dbl> 0.6547099, 0.4171009, 0.7381187, 0.5273580, 0.637479...
## $ elo_prob2     <dbl> 0.3452901, 0.5828991, 0.2618813, 0.4726420, 0.362520...
## $ elo1_post     <dbl> 1659.578, 1397.115, 1549.164, 1469.359, 1561.692, 16...
## $ elo2_post     <dbl> 1587.849, 1549.380, 1504.808, 1476.353, 1475.794, 14...
## $ qbelo1_pre    <dbl> 1616.496, 1442.592, 1562.644, 1436.168, 1524.717, 15...
## $ qbelo2_pre    <dbl> 1569.751, 1540.181, 1444.993, 1502.511, 1429.215, 15...
## $ qb1           <fct> Nick Foles, Eli Manning, Drew Brees, Ryan Tannehill,...
## $ qb2           <fct> Matt Ryan, Blake Bortles, Ryan Fitzpatrick, Marcus M...
## $ qb1_value_pre <dbl> 157.00512, 115.05835, 226.16119, 155.71034, 146.6940...
## $ qb2_value_pre <dbl> 179.86685, 170.79947, 130.06506, 146.22350, 16.81502...
## $ qb1_adj       <dbl> -5.5019890, -7.2357535, 6.6815478, 12.5462859, -0.34...
## $ qb2_adj       <dbl> -1.7870099, 2.6154020, -7.4652534, -0.5571494, -37.4...
## $ qbelo_prob1   <dbl> 0.6409749, 0.4298360, 0.7476861, 0.5071645, 0.747695...
## $ qbelo_prob2   <dbl> 0.3590251, 0.5701640, 0.2523139, 0.4928355, 0.252304...
```

```
## $ qb1_game_value <dbl> -48.7350673, 76.1007235, 492.3622170, 136.8648341, 3...
## $ qb2_game_value <dbl> -32.858912, 124.427603, 563.905841, -19.000441, -211...
## $ qb1_value_post <dbl> 136.43110, 111.16259, 252.78129, 153.82579, 162.9587...
## $ qb2_value_post <dbl> 158.594272, 166.162281, 173.449140, 129.701103, -6.0...
## $ qbelo1_post    <dbl> 1629.857, 1427.525, 1526.704, 1456.618, 1542.408, 16...
## $ qbelo2_post    <dbl> 1556.389, 1555.248, 1480.933, 1482.061, 1411.524, 14...
## $ score1         <int> 18, 15, 40, 27, 47, 24, 23, 21, 27, 28, 16, 27, 6, 2...
## $ score2         <int> 12, 20, 48, 20, 3, 16, 34, 21, 20, 38, 8, 24, 24, 23...
```

```
glimpse(weather)
```

```
## Rows: 267
## Columns: 10
## $ schedule_date       <fct> 12/2/18, 10/21/18, 10/28/18, 11/4/18, 11/11/18,...
## $ schedule_season     <int> 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018,...
## $ schedule_week       <fct> 13, 7, 8, 9, 10, 9, 7, 8, 2, 13, 13, 5, 7, 7, 1...
## $ team_home           <fct> GB, PHI, CHI, MIA, OAK, CLE, NYJ, BUF, JAX, MIA...
## $ team_away           <fct> ARI, CAR, NYJ, NYJ, LAC, KC, MIN, NE, NE, BUF, ...
## $ team_favorite_id    <fct> GB, PHI, CHI, MIA, LAC, KC, MIN, NE, NE, MIA, C...
## $ spread_favorite     <dbl> -13.5, -5.0, -8.5, -3.0, -10.5, -7.5, -3.5, -13...
## $ stadium             <fct> Lambeau Field, Lincoln Financial Field, Soldier...
## $ weather_temperature <int> 34, 49, 48, 87, 69, 53, 46, 46, 97, 86, 81, 79,...
## $ weather_wind_mph    <int> 20, 19, 18, 16, 16, 16, 16, 16, 15, 15, 15, 15,...
```

Tidying Data

First step is to create a unique identification variable that both sets have in common. Each dataset has a column for home and away teams. I use the function unite() in order to combine the two variable to create a unique id that will match the two datasets together. If I tried uniting by any other variable without the created id, then the data wouldn't join correctly since there isnt a unique variable between the two datasets.

```
weather %>% unite(id, "team_home", "team_away", remove = F) -> weather
season18 %>% unite(id, "team1", "team2", remove = F) -> season18
```

Joining

Now I will perform a full join to combine the two datasets. Then we will delete columns that are repeated or redudent.

```
full_join(season18,weather) -> nfl2018
```

```
## Joining, by = "id"
```

```
nfl2018 %>% select(-team_home,-team_away,-schedule_date,-schedule_season,-season,
                   -date,-elo1_pre,-elo2_pre,-elo1_post,-elo2_post,-qb1_value_pre,
                   -qb2_value_pre,-qb1_value_post,-qb2_value_post,qb1_game_value,
                   -qb2_game_value,-qbelo1_post,-qbelo2_post,-neutral)-> data
```

Wrangling

For this fist portion, I will be doing some stats on the Dallas Cowboys. They're my favorite team so I want to see how they match up according to the rest of the league. I will be caluating the average points they made at home and away and compare that to the rest of the league in order to determine if the Dallas Cowboys are playing above or below the average.

```
data %>% arrange(team1) %>% filter(team1 == "DAL") %>%
  summarize("PointsMade_Home" = mean(score1))
```

```
##   PointsMade_Home
## 1        24.88889
```

```
data %>% arrange(team2) %>% filter(team2 == "DAL") %>%
  summarize("PointsMade_Away" = mean(score2))
```

```
##   PointsMade_Away
## 1        17.88889
```

```
data %>% summarize("PointsMade_Home_avgNFL" = mean(score1))
```

```
##   PointsMade_Home_avgNFL
## 1               24.41026
```

```
data %>% summarize("PointsMade_Away_avgNFL" = mean(score2))
```

```
##   PointsMade_Away_avgNFL
## 1               22.18315
```

```
data %>% arrange(team1) %>% filter(team1 == "DAL") %>%
  summarize("PointsAllowed_Home" =mean(score2))
```

```
##   PointsAllowed_Home
## 1           18.88889
```

```
data %>% arrange(team2) %>% filter(team2 == "DAL") %>%
  summarize("PointsAllowed_Away" =mean(score1))
```

```
##   PointsAllowed_Away
## 1           22.88889
```

The average points made at home for the entire NFL are 24.41 and points made away is 22.18. Cowboys make 24.88 points at home and 17.88 points away. This data highlightes that the Dallas Cowboys are just above the average when it comes to scoring at home, however, they're below the average for scoring away. Therefore, the Dallas Cowboys are a better team at home. For fun I decided to see how many points were allowed. At home the defense allowed 18.88 points while away they allowed 22.88. Which furthore supports that the Dallas Cowboys are significantly better at home on both sides of the ball.

```
summary(data)
```

```
##  playoff      id                 team1          team2        elo_prob1
##   :259   Length:273         NO     : 14   IND    : 12   Min.   :0.1596
##  c:  3   Class :character   HOU    : 11   PHI    : 12   1st Qu.:0.4931
##  d:  5   Mode  :character   KC     : 10   LAR    : 11   Median :0.6136
##  s:  1                      LAR    : 10   LAC    : 10   Mean   :0.5856
##  w:  5                      BAL    :  9   NE     : 10   3rd Qu.:0.6933
##                             CHI    :  9   DAL    :  9   Max.   :0.8921
##                             (Other):210   (Other):209
##    elo_prob2        qbelo1_pre     qbelo2_pre                    qb1
##  Min.   :0.1079   Min.   :1314   Min.   :1316   Drew Brees       : 13
##  1st Qu.:0.3067   1st Qu.:1445   1st Qu.:1440   Deshaun Watson   : 11
##  Median :0.3864   Median :1514   Median :1505   Jared Goff       : 10
##  Mean   :0.4144   Mean   :1511   Mean   :1509   Patrick Mahomes  : 10
##  3rd Qu.:0.5069   3rd Qu.:1578   3rd Qu.:1572   Dak Prescott     :  9
```

```
##  Max.   :0.8404   Max.   :1713   Max.   :1704   Mitchell Trubisky:  9
##                                                 (Other)          :211
##              qb2          qb1_adj             qb2_adj          qbelo_prob1
##  Andrew Luck    : 12   Min.   :-179.213   Min.   :-174.458   Min.   :0.1245
##  Jared Goff     : 11   1st Qu.:  -1.886   1st Qu.:  -5.600   1st Qu.:0.4691
##  Philip Rivers : 10   Median :   6.334   Median :   7.337   Median :0.6028
##  Tom Brady      : 10   Mean   :   2.854   Mean   :   1.591   Mean   :0.5792
##  Dak Prescott   :  9   3rd Qu.:  17.743   3rd Qu.:  18.382   3rd Qu.:0.7024
##  Russell Wilson:  9   Max.   :  69.153   Max.   :  65.925   Max.   :0.9035
##  (Other)        :212
##   qbelo_prob2     qb1_game_value        score1          score2
##  Min.   :0.0965   Min.   :-227.96   Min.   : 0.00   Min.   : 0.00
##  1st Qu.:0.2976   1st Qu.:  92.18   1st Qu.:17.00   1st Qu.:16.00
##  Median :0.3972   Median : 182.84   Median :24.00   Median :22.00
##  Mean   :0.4208   Mean   : 182.29   Mean   :24.41   Mean   :22.18
##  3rd Qu.:0.5309   3rd Qu.: 273.00   3rd Qu.:31.00   3rd Qu.:28.00
##  Max.   :0.8755   Max.   : 561.90   Max.   :54.00   Max.   :51.00
##
##  schedule_week team_favorite_id spread_favorite                      stadium
##  14     : 17    LAR    : 18      Min.   :-17.000   MetLife Stadium          : 16
##  1      : 16    NE     : 18      1st Qu.: -7.500   Mercedes-Benz Superdome: 14
##  13     : 16    NO     : 17      Median : -4.000   NRG Stadium              : 11
##  15     : 16    HOU    : 14      Mean   : -5.359   Arrowhead Stadium        : 10
##  16     : 16    KC     : 14      3rd Qu.: -3.000   AT&T Stadium             :  9
##  17     : 16    CHI    : 13      Max.   : -1.000   Gillette Stadium         :  9
##  (Other):176   (Other):179                         (Other)                  :204
##  weather_temperature weather_wind_mph
##  Min.   :19.00       Min.   : 0.000
##  1st Qu.:50.00       1st Qu.: 0.000
##  Median :69.00       Median : 5.000
##  Mean   :62.45       Mean   : 5.381
##  3rd Qu.:72.00       3rd Qu.: 9.000
##  Max.   :97.00       Max.   :20.000
##
```

```r
data %>% select(elo_prob1,elo_prob2,qbelo1_pre,qbelo2_pre,qb1_adj,
                qb2_adj,qbelo_prob1,qbelo_prob2,qb1_game_value,score1,
                score2,weather_temperature,weather_wind_mph) %>%
  summarise_each(funs(sd = sd))
```

```
## Warning: `summarise_each_()` is deprecated as of dplyr 0.7.0.
## Please use `across()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

```
## Warning: `funs()` is deprecated as of dplyr 0.8.0.
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
```

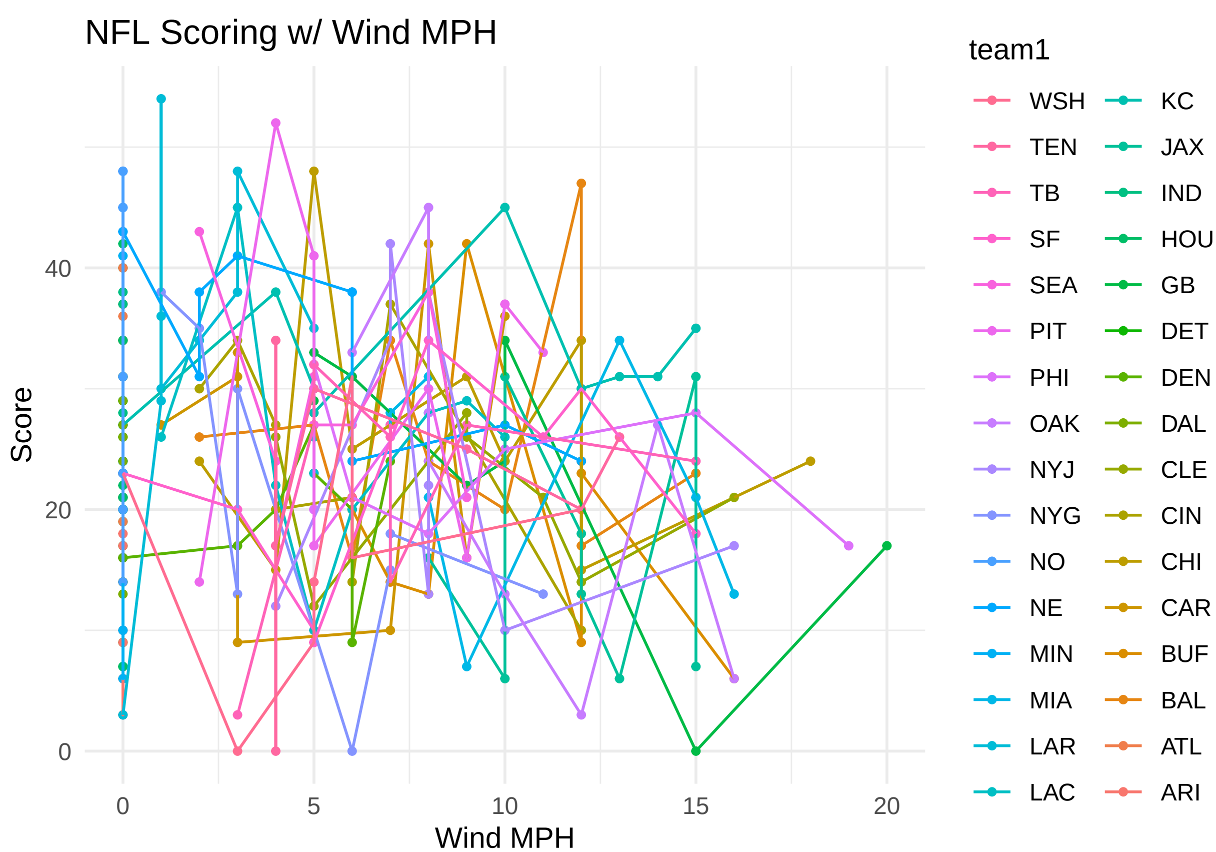
```
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

```
##   elo_prob1_sd elo_prob2_sd qbelo1_pre_sd qbelo2_pre_sd qb1_adj_sd qb2_adj_sd
## 1    0.1594488    0.1594488      86.97543      85.94288    30.1969   31.57757
##   qbelo_prob1_sd qbelo_prob2_sd qb1_game_value_sd score1_sd score2_sd
## 1      0.1651077      0.1651077          141.8551  10.75257  9.796241
##   weather_temperature_sd weather_wind_mph_sd
## 1               15.87863             4.89612
```

Here are some of the summary statistics of each of the variables. Some interesting stats would be that the std of scoring at home is 10.75 and scoring away is 9.79.Therefore scoring is relatively close. Which is suprising since every week there seems to be large scoring differentials. The lowest temperature ever played in the 2018 season is 19 degrees F by KC and NE. While the fastest wind MPH is 20 mph, played by the Green Bay Packers and Arizona Cardinals.

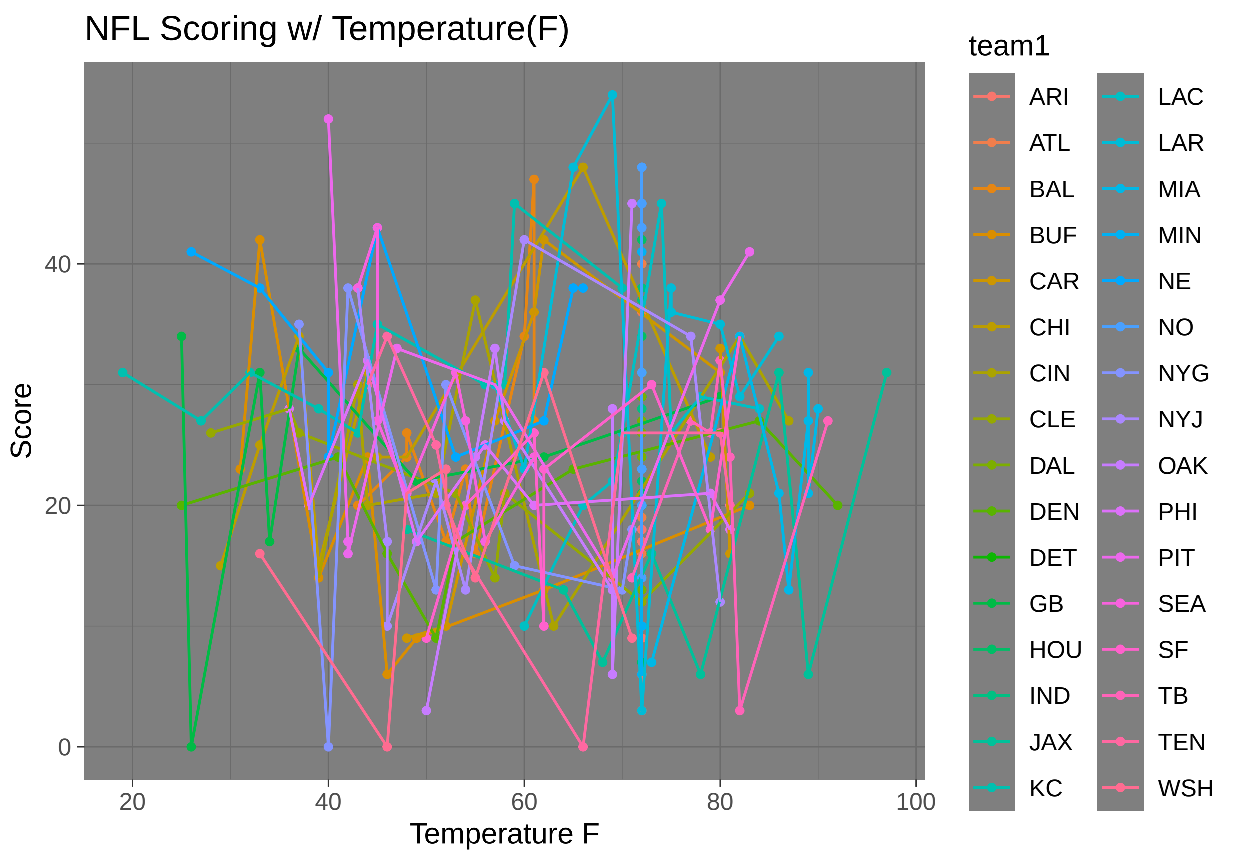
Visualizing

```r
ggplot(data = data, aes(x = weather_wind_mph, y = score1, color = team1)) +
  geom_point(size=1) + geom_line() +
  ggtitle("NFL Scoring w/ Wind MPH") +
  ylab("Score") +
  xlab("Wind MPH") +
  guides(color = guide_legend(reverse = TRUE)) + theme_minimal()
```



This first graph represents the NFLs team scoring depending on wind speeds. While looking at this graph, there is a clear negative spread between scoring and windspeed. As windspeeds increases scoring decreases. Which makes sense considering that the throws are not as accurate since wind can alter the path.

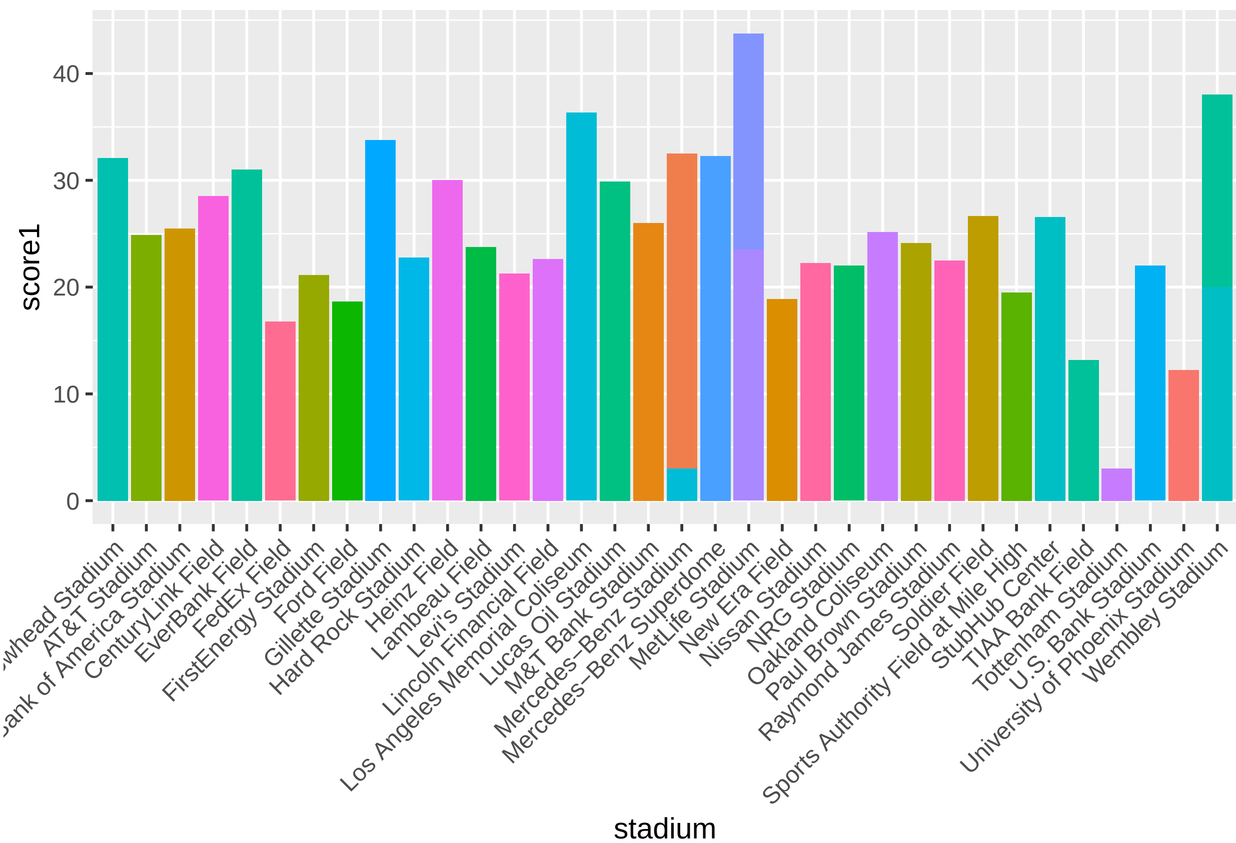
```
ggplot(data = data, aes(x = weather_temperature, y = score1, color = team1)) +
  geom_point(size=1) +
  geom_line() +
  ggtitle("NFL Scoring w/ Temperature(F)") +
  ylab("Score") +
  xlab("Temperature F") +
  theme_dark()
```



On this graph we have temperature and scoring. From taking a look at this graph there is no clear trend. The highest scoring games are around the 60-70 mark. Originally I was expecting teams that played in the cold to score less but the data says otherwise. Therefore, there is no clear difference when it comes to temperature and scoring ability in a league.

```
ggplot(data, aes(stadium))+
  geom_bar(aes(y=score1,fill=team1),
stat="summary", fun.y="mean")+
theme(axis.text.x = element_text(angle=45, hjust=1),
legend.position="none")
```

```
## Warning: Ignoring unknown parameters: fun.y

## No summary function supplied, defaulting to `mean_se()`
```

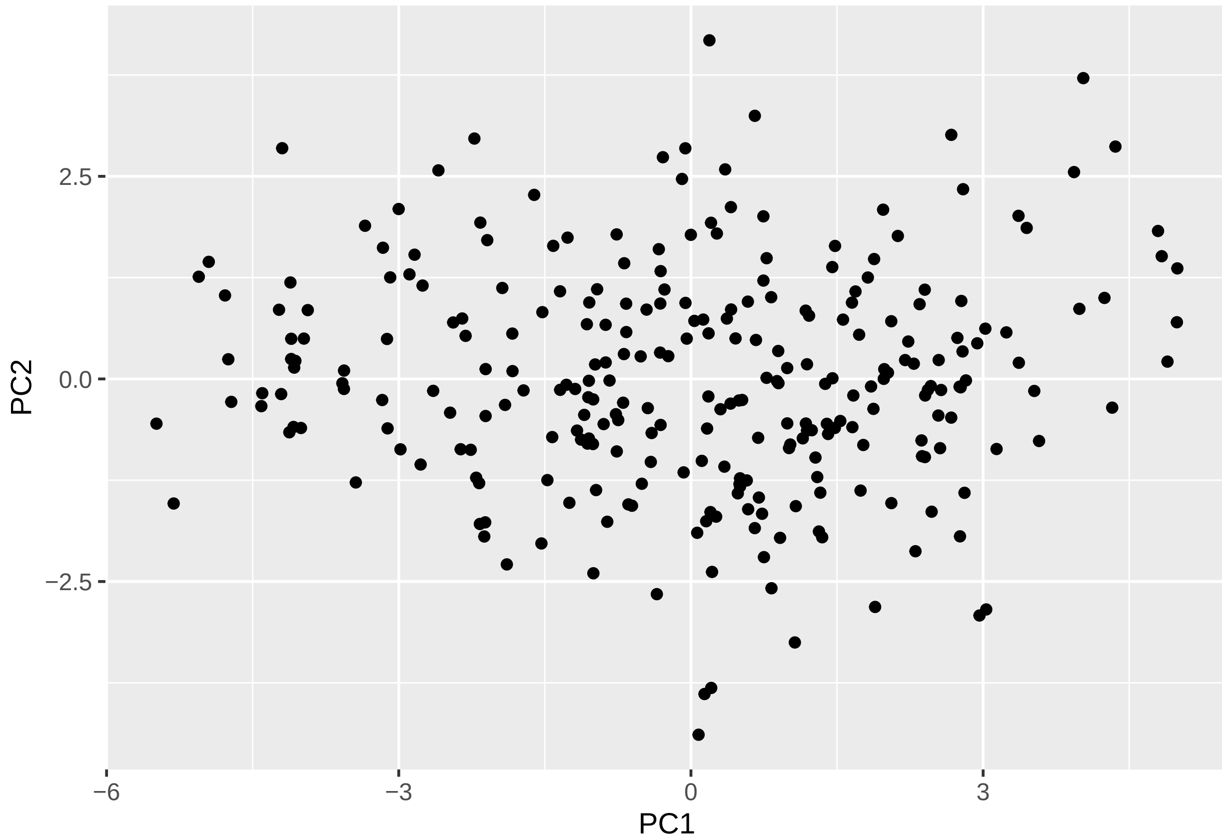


This graph represents scoring at each stadium. At first glance it seems that whoever plays at the MetLife stadium must be the best team in league with that amount of scoring. However, that stadium is shared by the New York Giants and New York Jets making the amount of points much higher than the rest. The actual highest scoring stadium would be the Los Angeles Memorial Coliseum home to the LA Rams which were in the superbowl.

Dimensionality Reduction

The first step to reduce the dimensions of this dataset is to create some PCA by selecting all the numerical variables and pumping them into a principle componet. After that we can calculate the eigvalue by squaring the std. After that we are able to plot a PCA graph.

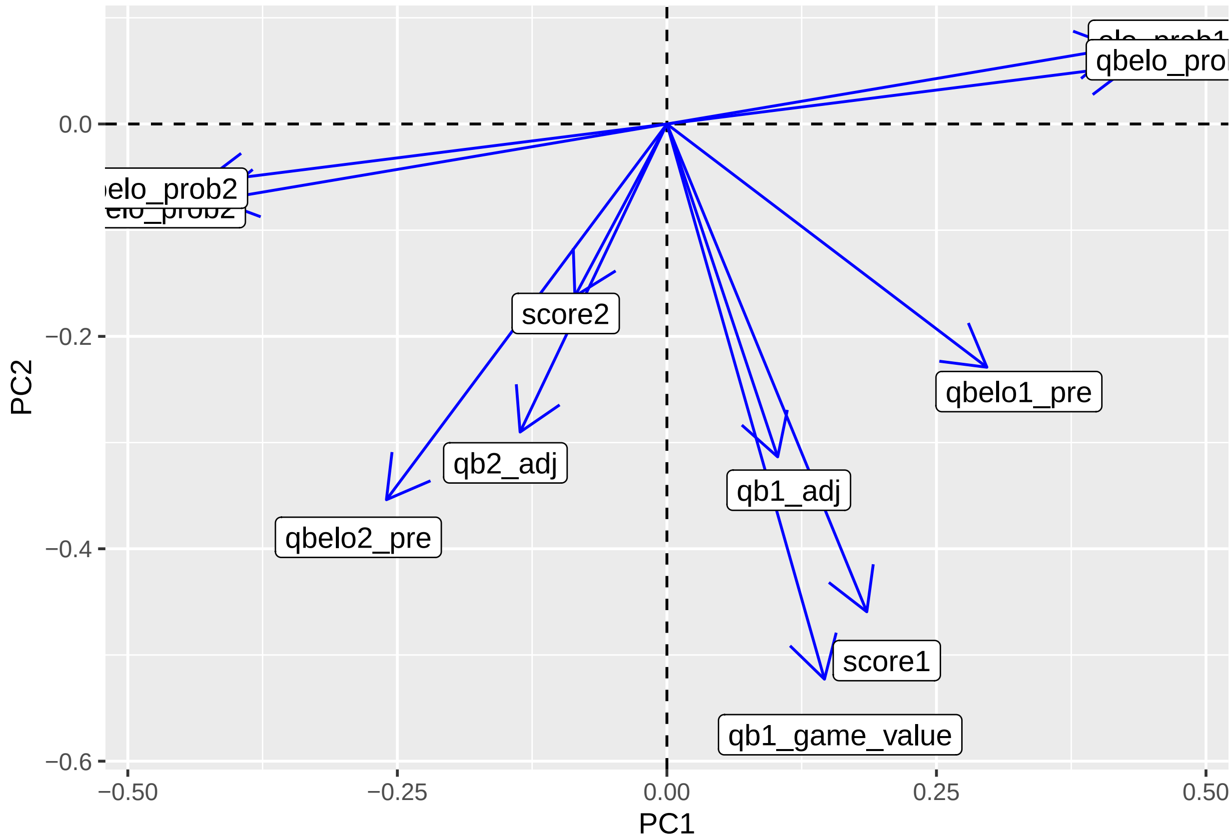
```
data %>% arrange(team1,team2)%>%
  select_if(is.numeric)%>%
  scale -> NFLnumbers
princomp(NFLnumbers) -> NFLPCA
eigval<-NFLPCA$sdev^2
varprop=round(eigval/sum(eigval),2)


NFLdf<-data.frame(PC1=NFLPCA$scores[,1], PC2=NFLPCA$scores[,2])
ggplot(NFLdf,aes(PC1, PC2))+geom_point()
```



Looking at this data there is a lot of variance between PC1 and PC2. From this we can see a few extreme points in the data.

Now we can create a plot of loadings which will show which variables have correlation and how much they contribute to the PCA.

```
NFLPCA$loadings[1:11,1:2]%>%as.data.frame%>%rownames_to_column%>%
  ggplot()+geom_hline(aes(yintercept=0),lty=2)+
  geom_vline(aes(xintercept=0),lty=2)+ylab("PC2")+xlab("PC1")+
  geom_segment(aes(x=0,y=0,xend=Comp.1,yend=Comp.2),arrow=arrow(),col="BLUE")+
  geom_label(aes(x=Comp.1*1.1,y=Comp.2*1.1,label=rowname))
```



After graphing the individual variables along the PCA we are able to determine the variables that have a greater correlations to each other and how much each contribute to the PCA. Right from the bat we can see that elo prob and qbelo prop are closely related which makes sense since the only difference between the two are the quarterbacks rating being factored in the already determined win probability. What is interesting is that there seems to be a very close relationship between the amount of scoring and the quarterbacks adjusted rating. This validates the elo rating system since the value of a quarterback in this system is closely related to the scoring ability.