Project_251

Brandon Upper, Ben Murdock, Hannah Lyons

```
#library
library(vroom)
library(ggplot2)
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v lubridate 1.9.4     v tibble    3.3.0
v purrr     1.1.0     v tidyr     1.3.1
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x readr::col_character()   masks vroom::col_character()
x readr::col_date()        masks vroom::col_date()
x readr::col_datetime()    masks vroom::col_datetime()
x readr::col_double()      masks vroom::col_double()
x readr::col_factor()      masks vroom::col_factor()
x readr::col_guess()       masks vroom::col_guess()
x readr::col_integer()     masks vroom::col_integer()
x readr::col_logical()     masks vroom::col_logical()
x readr::col_number()      masks vroom::col_number()
x readr::col_skip()        masks vroom::col_skip()
x readr::col_time()        masks vroom::col_time()
x readr::cols()            masks vroom::cols()
x readr::date_names_lang() masks vroom::date_names_lang()
x readr::default_locale()  masks vroom::default_locale()
x dplyr::filter()          masks stats::filter()
x readr::fwf_cols()        masks vroom::fwf_cols()
x readr::fwf_empty()       masks vroom::fwf_empty()
x readr::fwf_positions()   masks vroom::fwf_positions()
x readr::fwf_widths()      masks vroom::fwf_widths()
x dplyr::lag()             masks stats::lag()
```

```
x readr::locale()          masks vroom::locale()
x readr::output_column()   masks vroom::output_column()
x readr::problems()        masks vroom::problems()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(dplyr)
```

```
#data
shootouts <- vroom('./WorldCupShootouts.csv')
```

```
Rows: 304 Columns: 9
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (3): Team, Foot, Keeper
dbl (6): Game_id, Zone, OnTarget, Goal, Penalty_Number, Elimination

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

#Short Written Description

Our dataset contains information on World Cup penalty shootouts from Spain 1982 to Russia 2018 sourced from Kaggle. URL: [World Cup Penalty Shootouts Dataset] (https://www.kaggle.com/datasets/pablollanderos33/world-cup-penalty-shootouts)

The dataset contains 305 rows in total. The first row is a descriptive header, leaving 304 rows of actual data and 9 columns. The columns include: Game_id: Each team in a single game get's a number. Numbers are not repeated across separate games. Team: The country team of the kicker. Zone: The goal was divided into 9 sections, each labeled by number. This number is the zone. Foot: The foot used to take the shot (Left/Right). Keeper: The direction the goalkeeper dove(Left/Right) On Target: Indicates whether the shot would have been a goal if the keeper had not intervened(1 = Yes, 0 = No) Goal: Whether the penalty was successfully scored (1 = Goal, 0 = Miss) Penalty number: The order of the penalty within the shootout Elimination: Indicates whether the penalty could end the game (1 = Yes, 0 = No)

Proposed Research Question

We would like to determine if there is a difference between proportion of successful penalty kicks of right footed vs. left footed soccer players.

EDA

The proposed data source and a well-organized and well-articulated exploratory analysis of its contents (this should include well-labeled plots created in R).

Libraries

```
library(ggplot2)
library(vroom)
library(tidyverse)
```

Read File

```
Soccer_data <- vroom("WorldCupShootouts.csv")
```

```
Rows: 304 Columns: 9
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (3): Team, Foot, Keeper
dbl (6): Game_id, Zone, OnTarget, Goal, Penalty_Number, Elimination

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
summary(Soccer_data)
```

```
    Game_id          Team                Zone           Foot          
 Min.   : 1.00   Length:304         Min.   :1.000   Length:304        
 1st Qu.: 8.00   Class :character   1st Qu.:4.000   Class :character  
 Median :15.00   Mode  :character   Median :6.000   Mode  :character  
 Mean   :15.37                      Mean   :5.595                     
 3rd Qu.:23.00                      3rd Qu.:7.000                     
 Max.   :30.00                      Max.   :9.000                     
                                    NA's   :25                        
   Keeper             OnTarget          Goal        Penalty_Number  
 Length:304         Min.   :0.0000   Min.   :0.0000   Min.   : 1.000  
 Class :character   1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.: 3.000  
 Mode  :character   Median :1.0000   Median :1.0000   Median : 6.000  
                    Mean   :0.9176   Mean   :0.6989   Mean   : 5.579  
                    3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.: 8.000  
                    Max.   :1.0000   Max.   :1.0000   Max.   :12.000  
```

```
                  NA's   :25       NA's   :25
  Elimination
 Min.   :0.0000
 1st Qu.:0.0000
 Median :0.0000
 Mean   :0.1286
 3rd Qu.:0.0000
 Max.   :1.0000
 NA's   :24
```

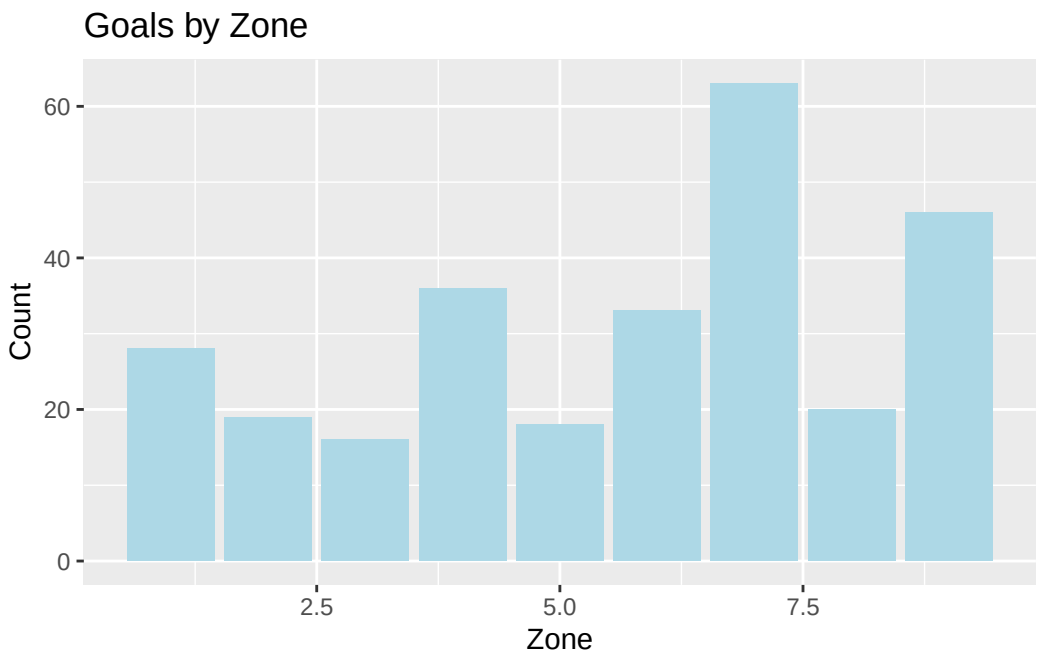
```
head(Soccer_data)
```

```
# A tibble: 6 x 9
  Game_id Team   Zone Foot  Keeper OnTarget  Goal Penalty_Number Elimination
    <dbl> <chr> <dbl> <chr> <chr>     <dbl> <dbl>          <dbl>       <dbl>
1       1 FRA       7 R     R             1     1              1           0
2       1 GER       9 R     C             1     1              2           0
3       1 FRA       6 R     L             1     1              3           0
4       1 GER       2 R     C             1     1              4           0
5       1 FRA       9 R     L             1     1              5           0
6       1 GER       4 R     L             1     0              6           0
```

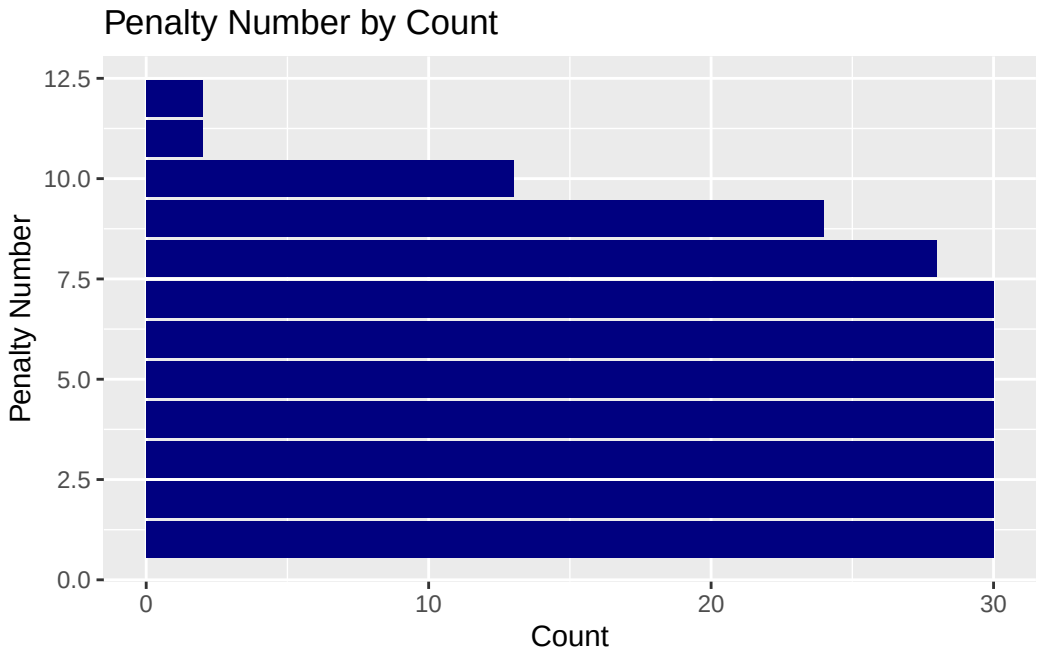
Reading in our file we can see various factors, features and variables that are described more in depth in the above short description of the data. From this code we can see the first five rows of the data and notice that many points are read in as doubles and characters. For the final analysis, we likely will convert some of these variables for simplicity

Basic plots about Goals and Penalties

```
Soccer_data <- Soccer_data %>%
  filter(is.finite(Zone))
ggplot(data = Soccer_data) +
  geom_bar(mapping = aes(x = Zone),
           fill = "lightblue") + ggtitle("Goals by Zone") + ylab("Count")
```




```
ggplot(data = Soccer_data) +
  geom_bar(mapping = aes(y = Penalty_Number), fill = "navyblue") +
  ggtitle("Penalty Number by Count") + ylab("Penalty Number") + xlab("Count")
```



These graphs give some visuals to some of the data and the types penalties we're seeing. For example, its clear that different zones have higher portions of goals scored, and when we recognize that the zone targeted and thereby goal likelihood can be highly influenced by right vs left footed kickers. From prior soccer experience, we could look for tendencies in the final project for kickers to shoot towards met opposite of their dominant foot. The second graph also displays how long the shootout lasts for, and its clear that most of them end before or around the twelfth number.

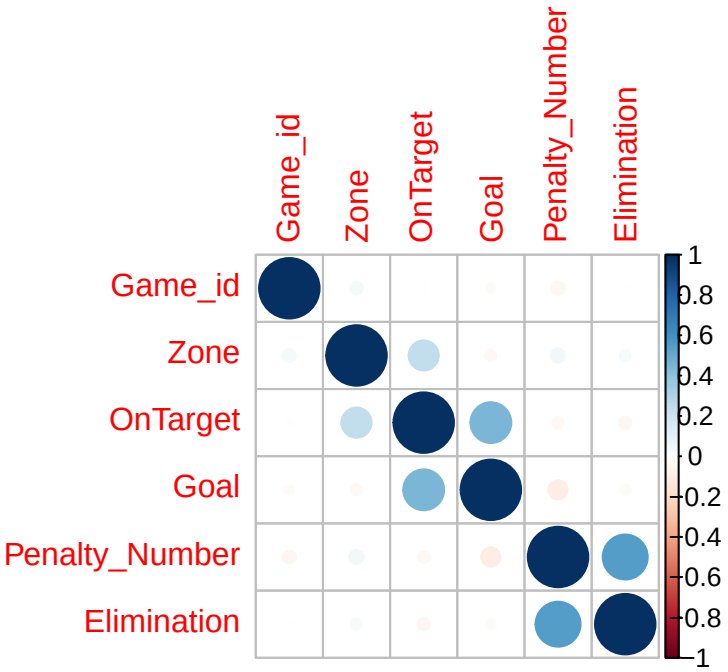
Looking for Correlation

```
numeric_data <- Soccer_data %>%
  select(where(is.numeric))

library(corrplot)
```

```
corrplot 0.95 loaded
```

```
cor_matrix <- cor(numeric_data)
corrplot(cor_matrix)
```



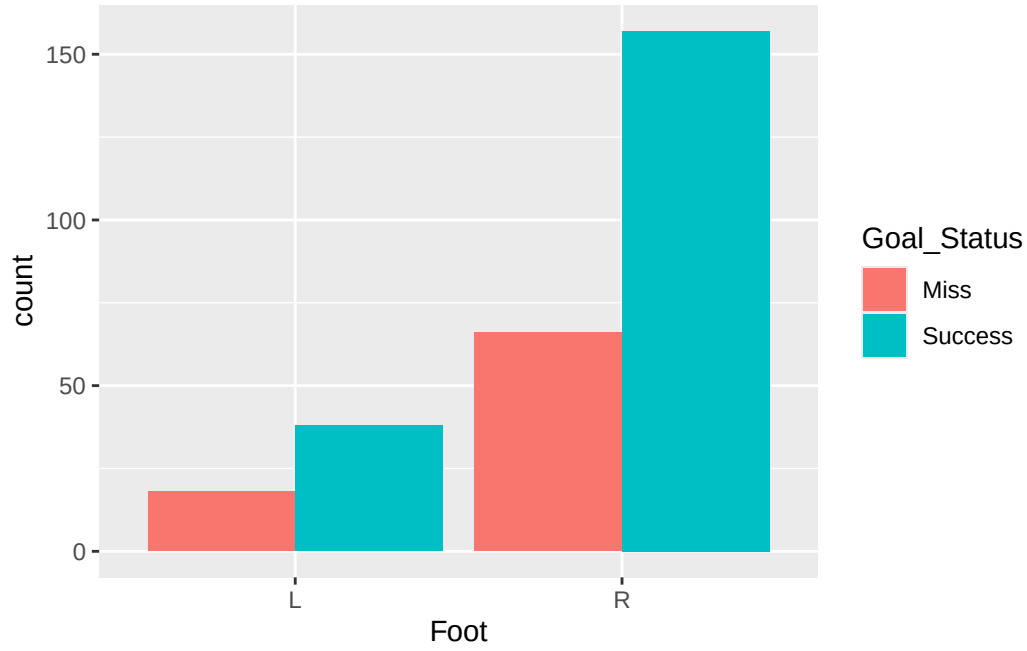
There seems to be no strong correlations between any of the numeric variables. This is useful as we can see that this data is unlikely to have issues with multicollinearity.

Left vs right foot - success proportion

```r
#

plot_data <- Soccer_data %>%
  select(Foot, Goal) %>%
  filter(Foot %in% c("L", "R"),
         Goal %in% c(0, 1)) %>%
  mutate(Goal_Status = factor(Goal,
                              levels = c(0, 1),
                              labels = c("Miss", "Success")))

ggplot(plot_data, aes(x = Foot, fill = Goal_Status)) +
  geom_bar(position = "dodge")
```




```r
table(plot_data)
```

```
, , Goal_Status = Miss

    Goal
Foot  0  1
   L 18  0
   R 66  0
```

```
, , Goal_Status = Success

    Goal
Foot   0   1
   L   0  38
   R   0 157
```

```
print("Success rate L = 0.3214286 vs R = 0.2959641")
```

```
[1] "Success rate L = 0.3214286 vs R = 0.2959641"
```

Some basic data visualization looking at the success rate of left foot vs right foot to see if one generally has a higher rate. There is a lot more right footed people who took penalties, likely because they are more common in the general population, but interestingly they have a lower ratio of success. This is something to consider when "determining if there is a difference between proportion of successful penalty kicks of right footed vs. left footed soccer players."