Football Events Project

Summer 2021

Required Packages:

- tidyverse
- ggplot2

To find the data used in this file, see “GitHub Preparing Data.R” in R Code folder.

The rest of this analysis will use the events_new.csv file which was updated in the aforementioned data prep code.

```
as.table(summary(events))
```

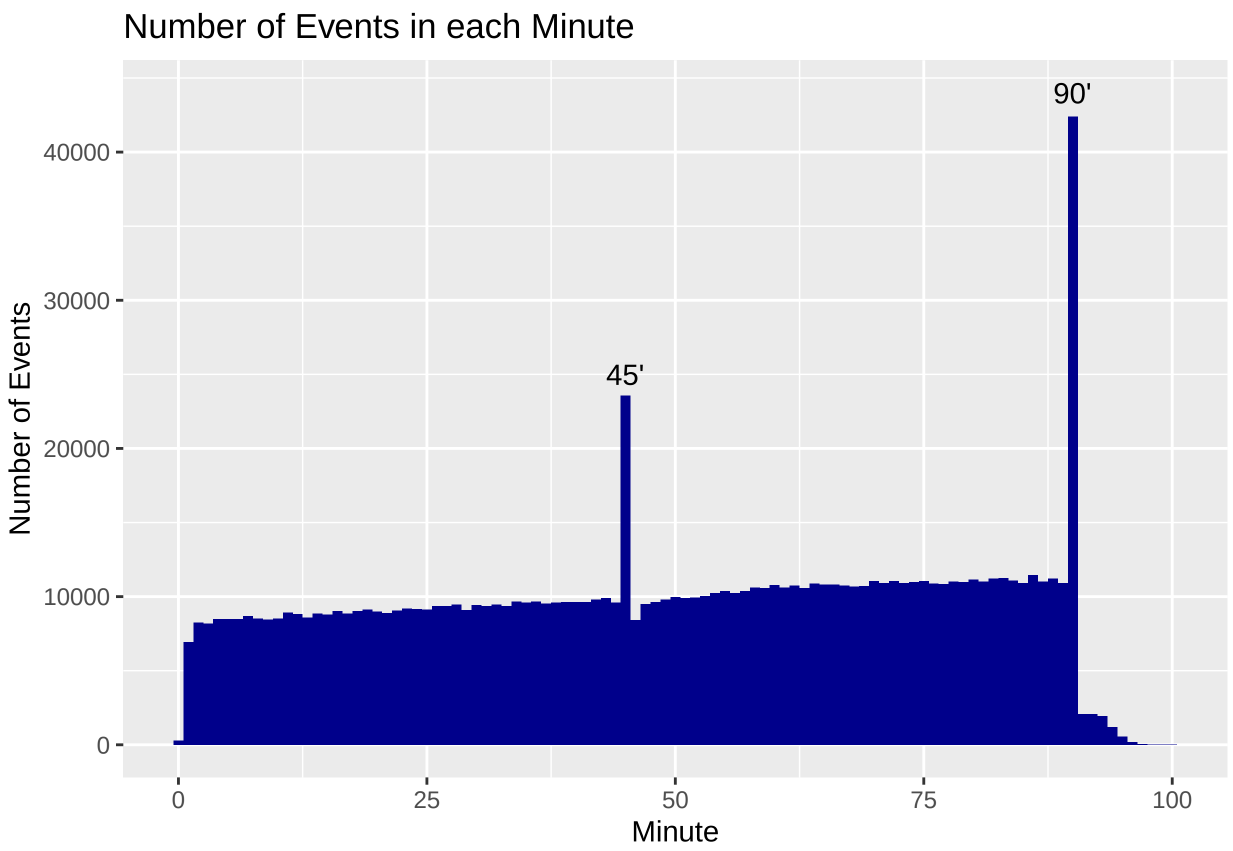
```
##        X            id_match           id_event           sort_order
##  Min.   :     1   Length:941009      Length:941009      Min.   :  1.00
##  1st Qu.:235253   Class :character   Class :character   1st Qu.: 27.00
##  Median :470505   Mode  :character   Mode  :character   Median : 53.00
##  Mean   :470505                                         Mean   : 53.86
##  3rd Qu.:705757                                         3rd Qu.: 79.00
##  Max.   :941009                                         Max.   :180.00
##
##       time            text             event_type     event_type2   side
##  Min.   :  0.00   Length:941009      8      :237932   12  :167859   1:488224
##  1st Qu.: 27.00   Class :character   3      :232925   13  : 43475   2:452785
##  Median : 51.00   Mode  :character   1      :229135   14  :  2258
##  Mean   : 49.66                      2      : 91204   15  :   701
##  3rd Qu.: 73.00                      7      : 51738   NA’s:726716
##  Max.   :100.00                      9      : 43476
##                                      (Other): 54599
##   event_team         opponent            player             player2
##  Length:941009      Length:941009      Length:941009      Length:941009
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   player_in          player_out          shot_place     shot_outcome
##  Length:941009      Length:941009      2      : 54082   1   : 78014
##  Class :character   Class :character   8      : 27024   2   : 92827
##  Mode  :character   Mode  :character   9      : 25213   3   : 54082
##                                        5      : 25079   4   :  3575
##                                        4      : 18748   NA’s:712511
##                                        (Other): 77313
##                                        NA’s   :713550
##  is_goal       location      bodypart      assist_method situation
##  0:916563   2      :125137   1   :121939   0:773104      1   :193747
##  1: 24446   15     : 93667   2   : 71290   1:109534      2   : 11742
##             3      : 69606   3   : 35956   2: 43283      3   : 18226
##             1      : 53964   NA’s:711824   3:  7713      4   :  5422
##             4      : 29669                 4:  7375      NA’s:711872
##             (Other): 95024
##             NA’s   :473942
##  fast_break    league             season
##  0:936421   Length:941009      Length:941009
##  1:  4588   Class :character   Class :character
##             Mode  :character   Mode  :character
##
##
##
##
```

The dataset contains 24 variables and 941,009 observations, where each observation is an event taking place in a match. To perform some initial exploratory analysis into the variables individually, I will go through those for which this makes the most sense and display information about them through graphics.

Time:

The time variable gives the minute in which the given event took place and ranges from 0-100.

```r
library(ggplot2)
ggplot(events) + geom_bar(aes(x = time), fill = "dark blue", width = 1) +
  labs(x = "Minute", y = "Number of Events", title = "Number of Events in each Minute") +
  annotate("text", x = 45, y = 25000, label = "45'") +
  annotate("text", x = 90, y = 44000, label = "90'")
```



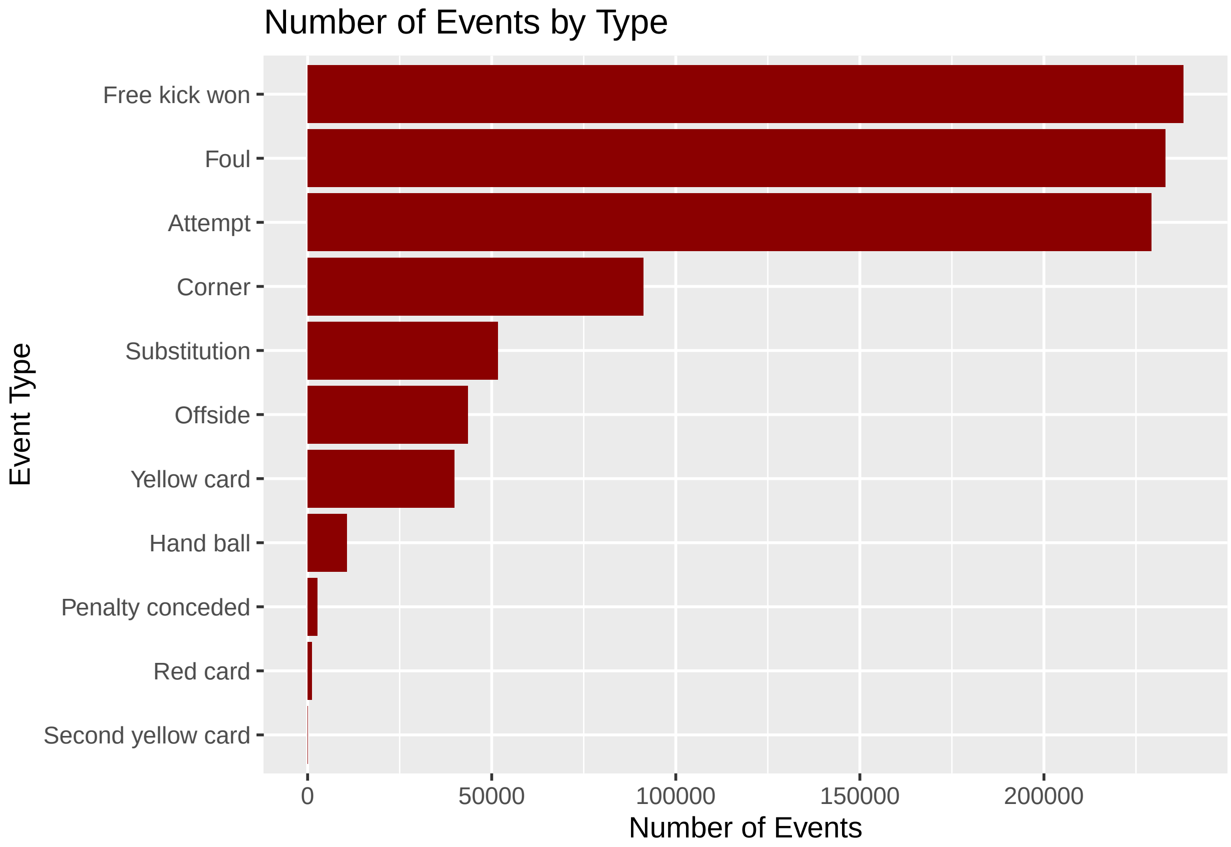
There appears to be a slight upward trend from 0-90 with respect to the number of events occurring in each minute. Moreover, there are two spikes in the plot which correspond to the 45th and 90th minute. This is likely because some events which happened during injury time, e.g. the 45+2 minute, may have been coded as the 45th minute, and the same for the 90th minute. The fact that the spike in the 90th minute is larger than that for the 45th minute is probably because there tends to be more time added on during the second half of a game than the first, mostly because of substitutions or injuries when players get tired. Finally, there is a large drop in the number of events after the 90th minute, which likely is due to the effect of some events being captured in the 90th minute as opposed to the appropriate minute in injury time.

Event type:

event_type shows what occurred in each event, with 12 possible event types.

```
event_type.labs = c("0" = "Announcement", "1" = "Attempt", "2" = "Corner", "3" = "Foul", "4" = "Yellow
                    "5" = "Second yellow card", "6" = "Red card", "7" = "Substitution", "8" = "Free kic
                    "9" = "Offside", "10" = "Hand ball", "11" = "Penalty conceded")

events %>% add_count(event_type) %>%
  ggplot() + geom_bar(aes(y = reorder(event_type, n)), fill = "dark red") +
    scale_y_discrete("Event Type", labels = event_type.labs) +
    labs(x = "Number of Events", title = "Number of Events by Type")
```

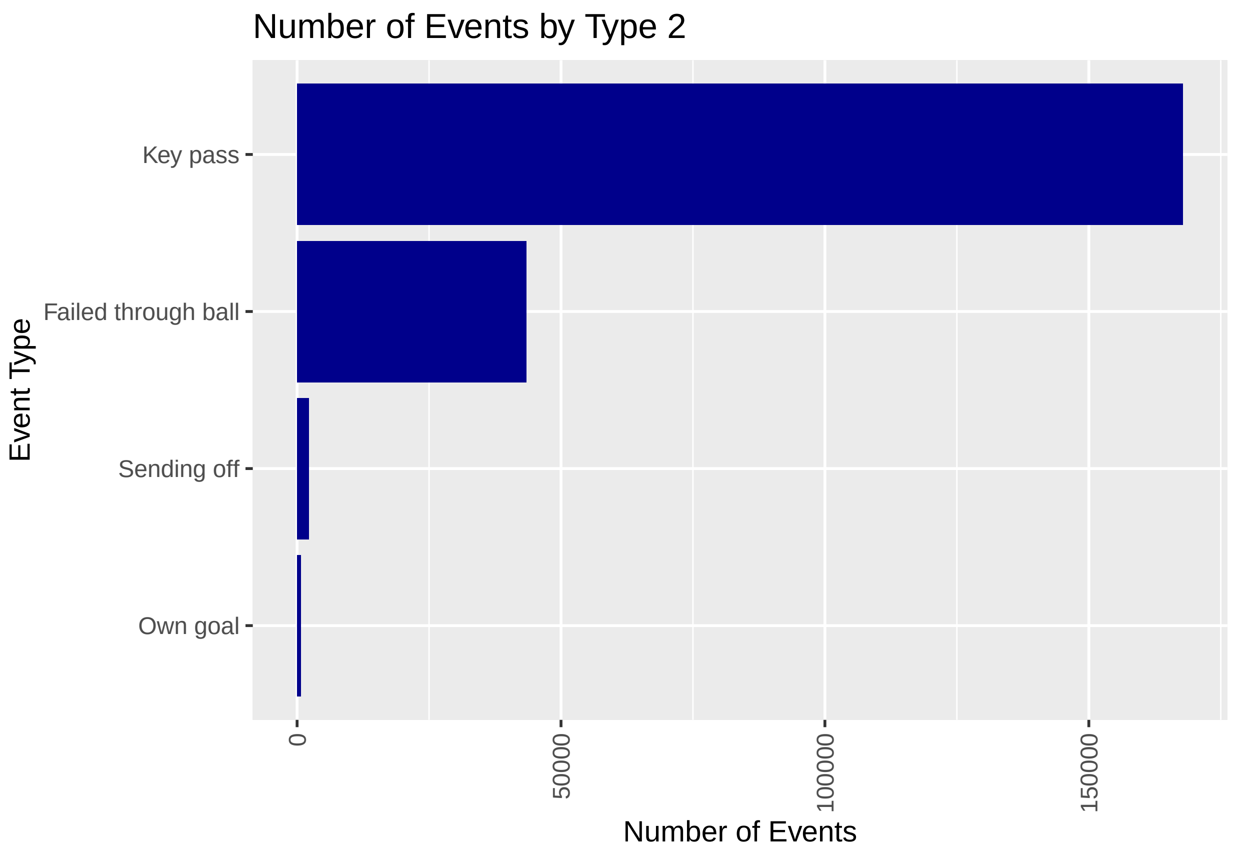


The most common events in the data are free kicks won, fouls, and attempts respectively. It may seem odd that there are more free kicks won than fouls, since every free kick won has a corresponding foul in the data, but this is because the category of foul only captures those fouls which do not fit into the other categories, e.g. not a handball, or resulting in a card. Otherwise, corners are very common, but nowhere near as common as the events previously outlined. While it may look like there were no second yellow cards, there were 100 but the figure is too negligible to show up compared to the volume of events in the other categories.

There is a second variable for event types, which contains additional information about attempts, yellow cards, second yellow cards, red cards, and offsides.

```
event_type2.labs = c("12" = "Key pass", "13" = "Failed through ball", "14" = "Sending off", "15" = "Own

events[!(is.na(events$event_type2)),] %>% add_count(event_type2) %>%
  ggplot() + geom_bar(aes(y = reorder(event_type2, n)), fill = "dark blue") +
    theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) +
    scale_y_discrete("Event Type", labels = event_type2.labs) +
    labs(x = "Number of Events", title = "Number of Events by Type 2")
```



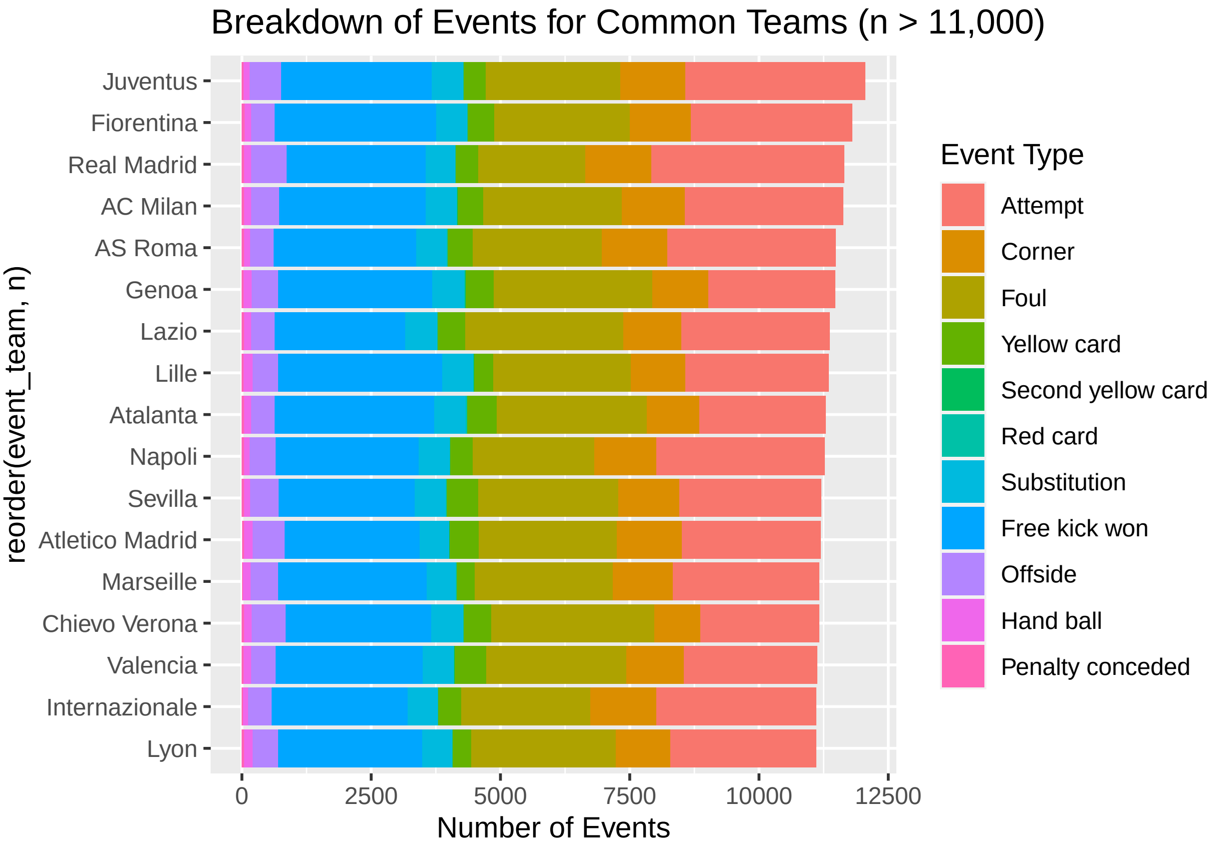
The vast majority of the second event types are key passes followed by failed through balls and then a small number of sending offs and own goals.

Event team:

Given the large number of teams in the dataset, I will take a subset of the most common teams and look at which teams have the most events.

```
common.teams = count(events, event_team) %>% filter(n > 11000)

events[events$event_team %in% common.teams$event_team,] %>% add_count(event_team, sort = T) %>%
  ggplot() +
  geom_bar(aes(y = reorder(event_team, n), fill = event_type)) +
  scale_fill_discrete("Event Type", labels = event_type.labs) +
  labs(x = "Number of Events", title = "Breakdown of Events for Common Teams (n > 11,000)")
```



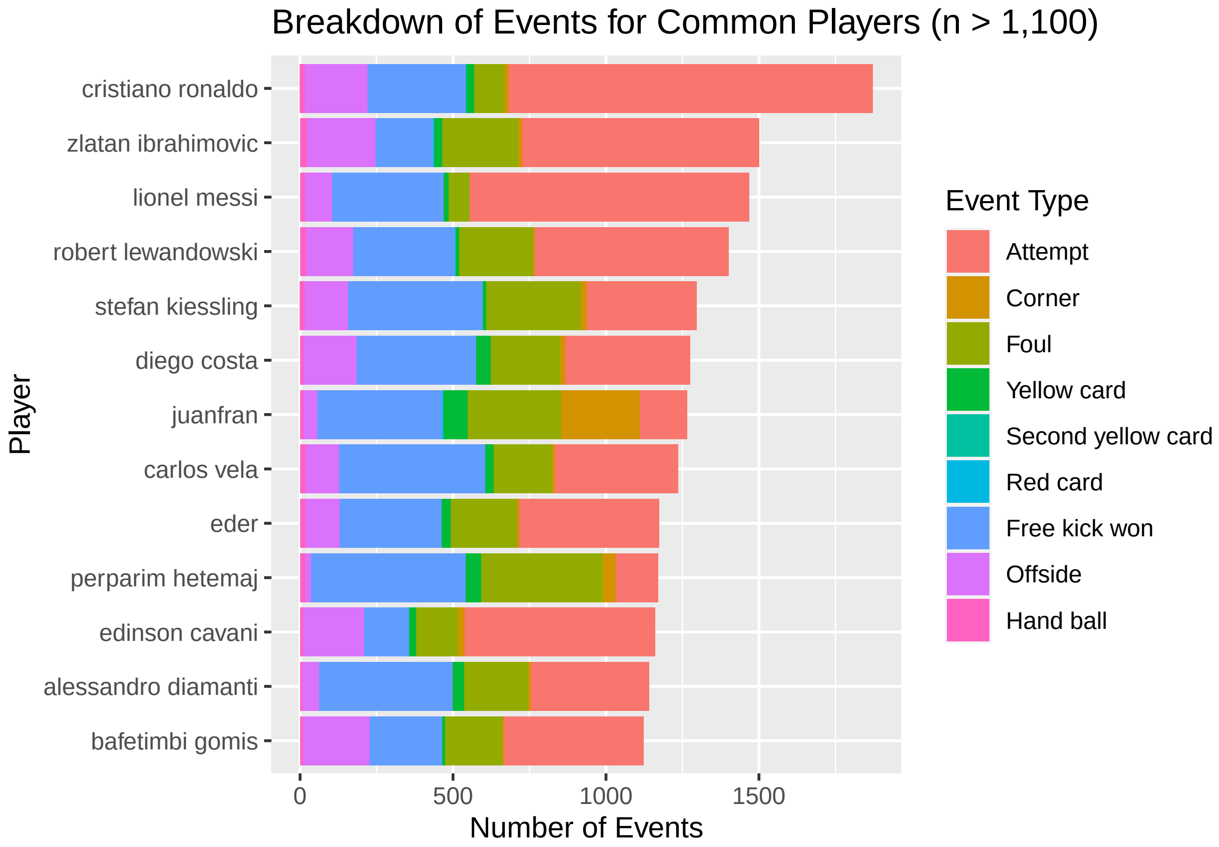
Unsurprisingly, there are no Bundesliga teams in the most common teams because this is measured by total number of events in the data, and the Bundesliga has fewer games per season than the other leagues meaning there will be fewer events per season on average. The most common teams are all big European teams such as Juventus and Real Madrid. However, Genoa is a team which did not have much success during the period described in the data but had more events than many big European teams. Looking at the breakdown of events by type for Genoa, we see that they have fewer attempts compared to the teams around them and have more fouls, which suggests Genoa's place in the rank could be caused by being a physical team giving away many more fouls than the average team.

Player:

Just as there are too many teams to see how many events they have, there are even more players in the data, so a similar approach will be taken as above and the most common players by number of events will be considered.

```
common.players = count(events[!is.na(events$player) & !is.na(events$event_type),], player) %>% filter(n

events[events$player %in% common.players$player,] %>% add_count(player, sort = T) %>%
  ggplot() +
  geom_bar(aes(y = reorder(player, n), fill = event_type)) +
  scale_fill_discrete("Event Type", labels = event_type.labs) +
  labs(x = "Number of Events", title = "Breakdown of Events for Common Players (n > 1,100)", y = "Playe
```



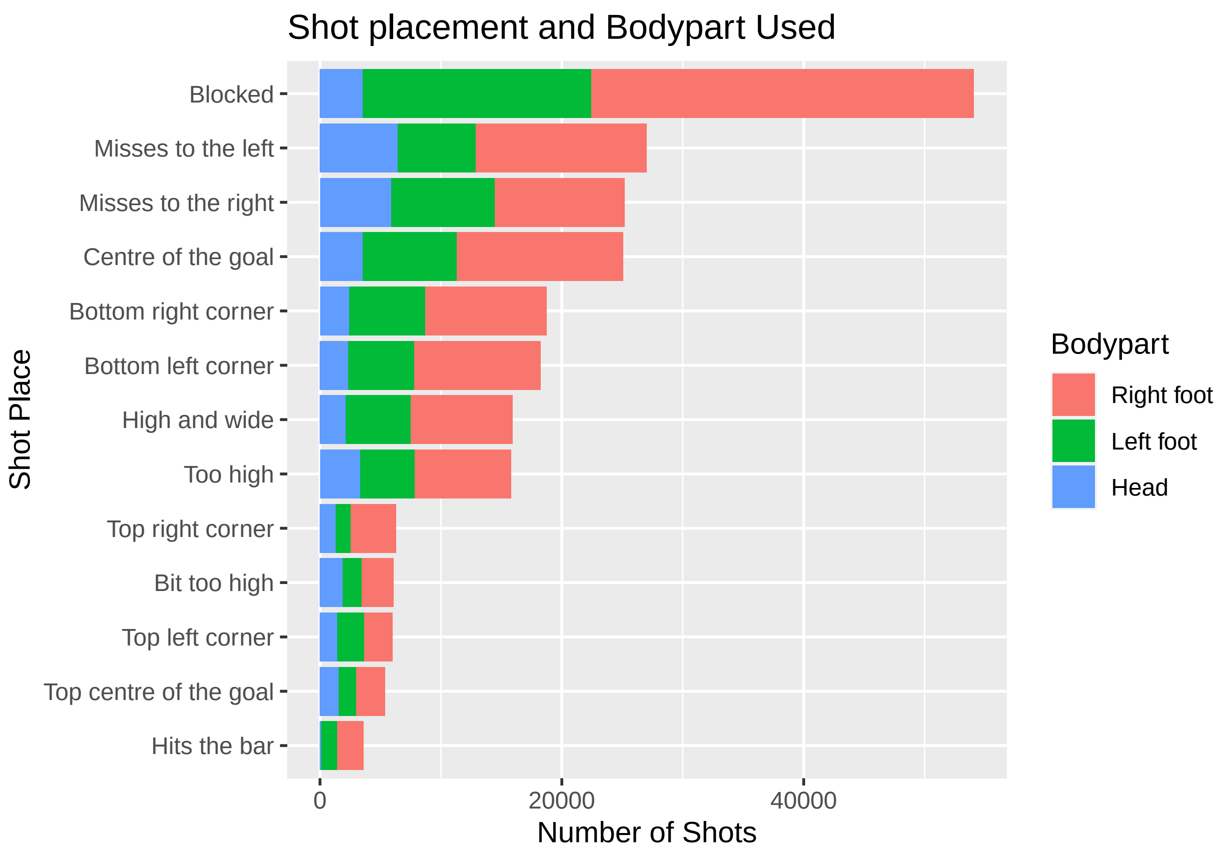
The majority of players in the most common are forwards or wingers, including some of the best players in the game. Two exceptions to this are Juanfran and Perparim Hetemaj, a defender and midfielder respectively. Juanfran, who played for Atletico Madrid for the duration of the period in the data, had a relatively high amount of fouls compared to other players, which is expected since defenders have to tackle more than other players and are therefore more likely to give away a foul. Juanfran also had far more corners than any other player in the dataset, suggesting that Juanfran was Atletico's go to corner taker for a while. Hetemaj, a defensive midfielder, gains most of his events through fouls and free kicks won. Hetemaj has quite a few more fouls than Juanfran, which could reflect that the areas which defenders make tackles are generally more hazardous and closer to the penalty area than defensive midfielders. However, it could simply be that Hetemaj is more aggressive when tackling than Juanfran and others and therefore gives away more fouls.

Shot place and bodypart:

Given that shots will always be with a bodypart, it makes sense to look at these two variable together and see which shot placements are most common and what the breakdown is by bodypart.

```
shot_place.labs = c("1" = "Bit too high", "2" = "Blocked", "3" = "Bottom left corner", "4" = "Bottom ri
                    "5" = "Centre of the goal", "6" = "High and wide", "7" = "Hits the bar",
                    "8" = "Misses to the left", "9" = "Misses to the right", "10" = "Too high",
                    "11" = "Top centre of the goal", "12" = "Top left corner", "13" = "Top right corner
bodypart.labs = c("1" = "Right foot", "2" = "Left foot", "3" = "Head")

events[!is.na(events$shot_place),] %>% add_count(shot_place, sort = T) %>%
  ggplot() +
    geom_bar(aes(y = reorder(shot_place, n), fill = bodypart), position = "stack") +
    scale_y_discrete("Shot Place", labels = shot_place.labs) +
    scale_fill_discrete("Bodypart", labels = bodypart.labs) +
    labs(x = "Number of Shots", title = "Shot placement and Bodypart Used")
```



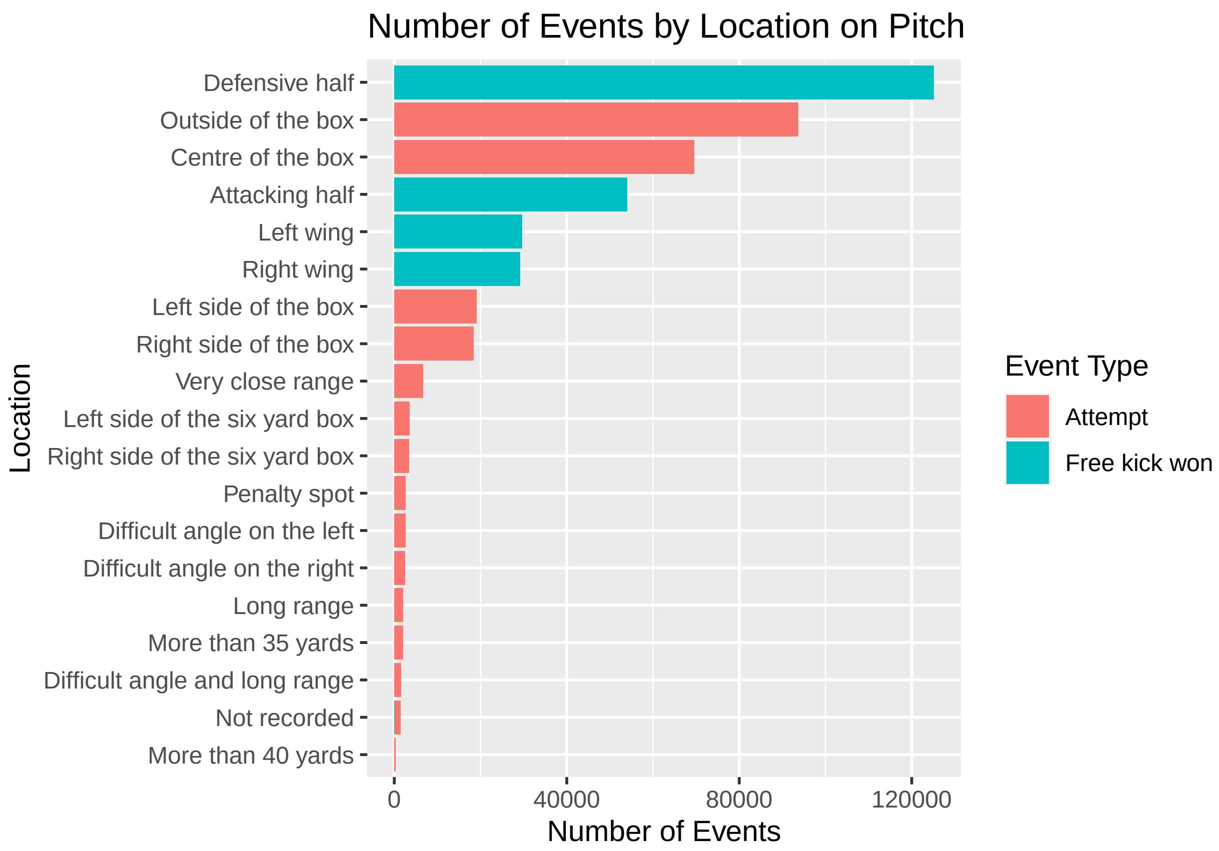
The majority of shots are blocked, meaning the placement of the shot cannot be accurately determined, with the next two most common being misses either side of goal. The centre of the goal is the most common shot placement on target, followed by both bottom corners. With respect to the bodypart, all of the placements have a majority of right foot, suggesting this is the most common strong foot for players. Then, most of the placements have left foot as their second most common bodypart, followed by head.

Location:

Location refers to the area on the pitch in which an event took place. The labels of Location can be somewhat unclear because some can include others, for example it appears that most of the locations could be classed as in the attacking half, which is itself a location. Looking at the breakdown of event types by location explains why this is.

```
location.labs = c("1" = "Attacking half", "2" = "Defensive half", "3" = "Centre of the box", "4" = "Lef
                  "5" = "Right wing", "6" = "Difficult angle and long range", "7" = "Difficult angle on
                  "8" = "Difficult angle on the right", "9" = "Left side of the box", "10" =
                    "Left side of the six yard box", "11" = "Right side of the box", "12" =
                    "Right side of the six yard box", "13" = "Very close range", "14" = "Penalty spot",
                  "15" = "Outside of the box", "16" = "Long range", "17" = "More than 35 yards", "18" =
                    "More than 40 yards", "19" = "Not recorded")

events[!is.na(events$location),] %>% add_count(location, sort = T) %>%
  ggplot() +
    geom_bar(aes(y = reorder(location, n), fill = event_type)) +
    scale_y_discrete("Location", labels = location.labs) +
    scale_fill_discrete("Event Type", labels = event_type.labs) +
    labs(x = "Number of Events", title = "Number of Events by Location on Pitch")
```

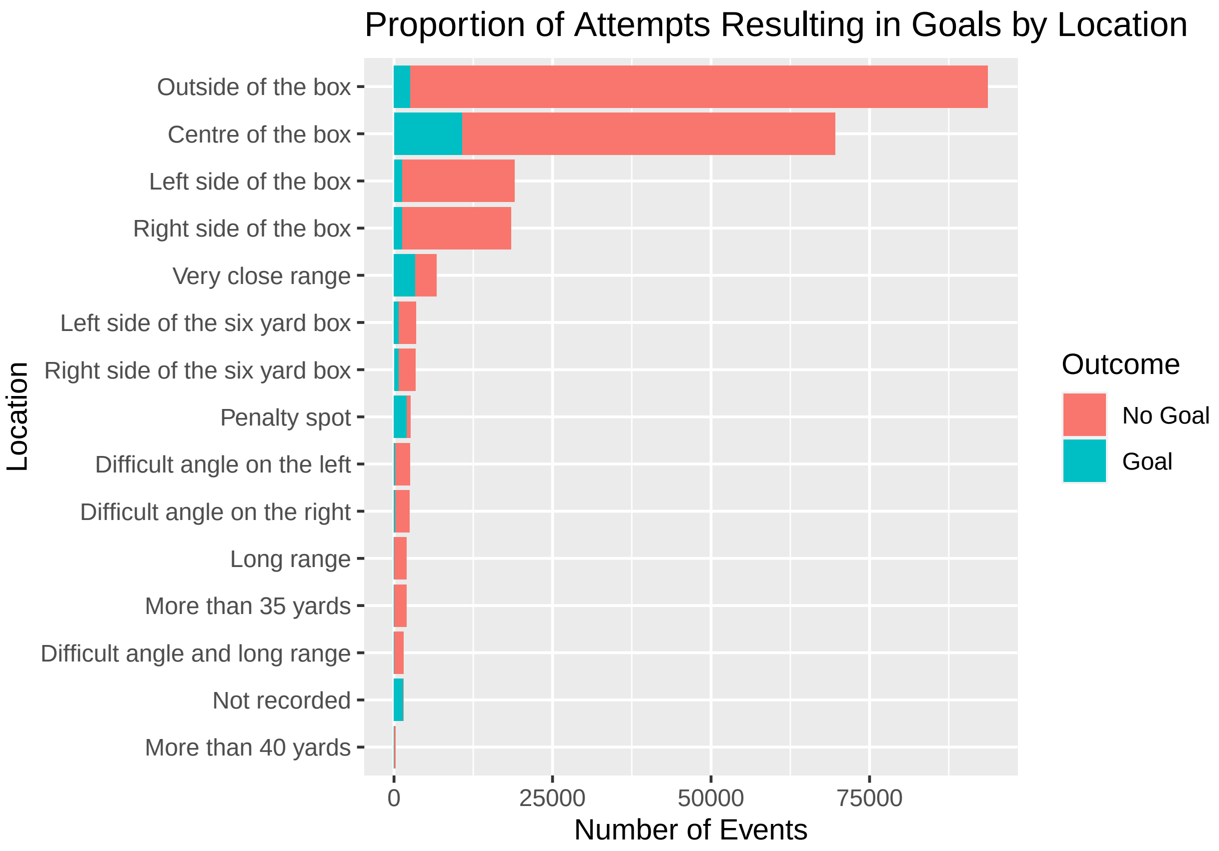


The chart shows that only two types of events have a location attached to them: attempts and free kicks won. Moreover, there are locations which are only attached to free kicks won and others which only apply to attempts, and none that overlap. Free kicks won have more general areas, either half and either wing, but it is unclear what these areas represent since either wing could be in either half. Given how many more events are in the defensive half, it could be that the attacking half is broken into wings and the centre, since the sum of those three combined is roughly the same as for the defensive half.

In terms of attempts, most of them come from outside of the box followed by the centre of the box, with the rest being comparatively infrequent. It is interesting to look at how many attempts resulted in goals, shown below.

```
is_goal.labs = c("0" = "No Goal", "1" = "Goal")
events[events$event_type == 1,] %>% add_count(location, sort = T) %>%
  ggplot() + geom_bar(aes(y = reorder(location, n), fill = is_goal)) +
```

```r
  scale_y_discrete("Location", labels = location.labs) +
  scale_fill_discrete("Outcome", labels = is_goal.labs) +
  labs(x = "Number of Events", title = "Proportion of Attempts Resulting in Goals by Location")
```



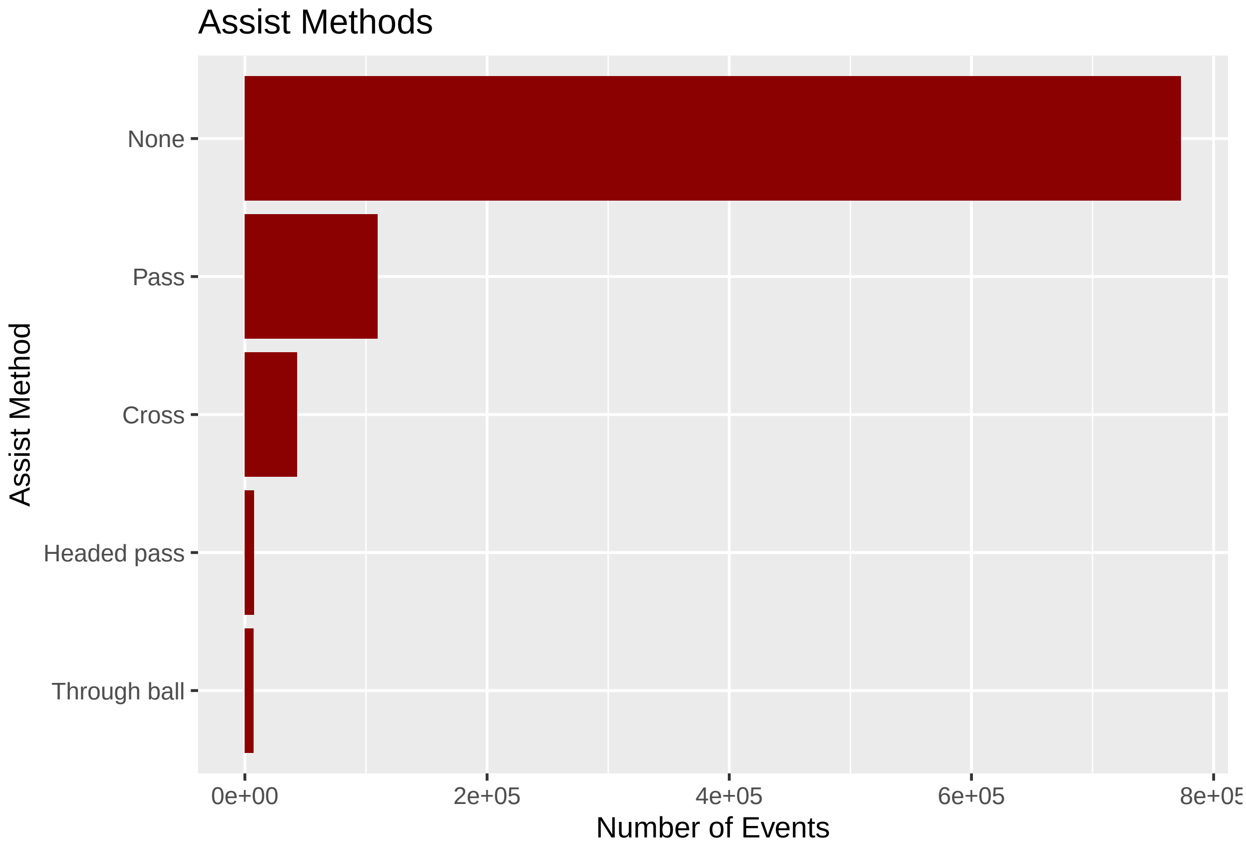
While there were quite a few more attempts from outside of the box, far fewer resulted in a goal when compared to those from the centre of the box, as is to be expected.

Assist method:

This variable gives, for every goal, what the type of assist was that came before the goal, meaning the last action by a teammate before the goal.

```r
assist_method.labs = c("0" = "None", "1" = "Pass", "2" = "Cross", "3" = "Headed pass", "4" = "Through ba

events[!is.na(events$assist_method),] %>% add_count(assist_method) %>%
  ggplot() + geom_bar(aes(y = reorder(assist_method, n)), fill = "dark red") +
    scale_y_discrete("Assist Method", labels = assist_method.labs) +
    labs(x = "Number of Events", title = "Assist Methods")
```



The overwhelming majority of goals came without an assist, meaning either that the last player to touch the ball before the goal scorer was an opposition player, or that the goal scorer tackled an opposition player before scoring. Another possibility is that there are observations for which the assist method was not accurately logged and was simply added as “None”. The second most common assist is a pass followed by a cross. This information could potentially be misleading because there is no description with the data as to what differentiates a pass from a cross and from a through ball, making it difficult to interpret results based on this variable.