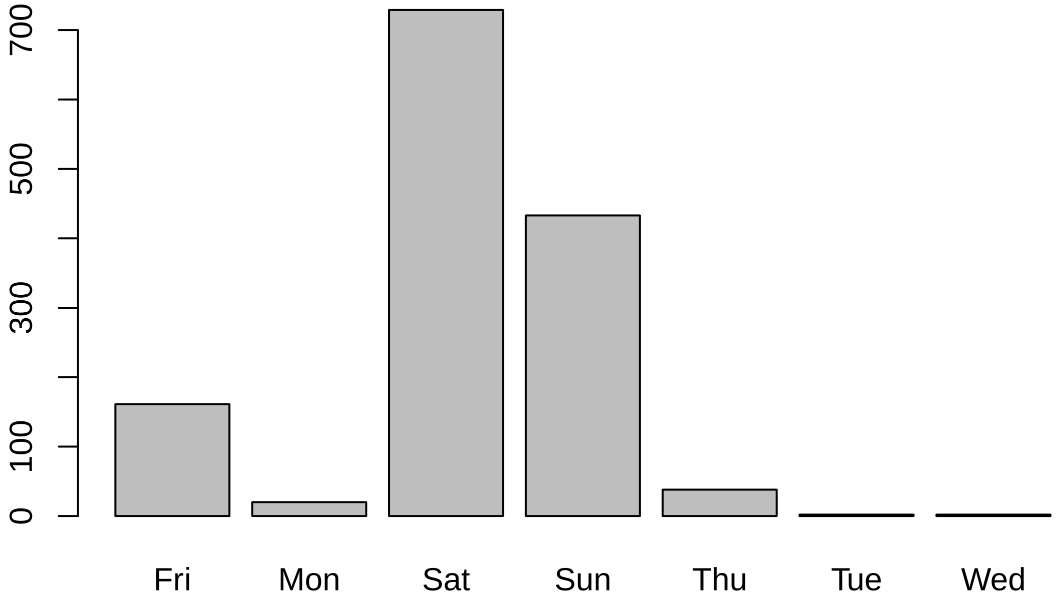
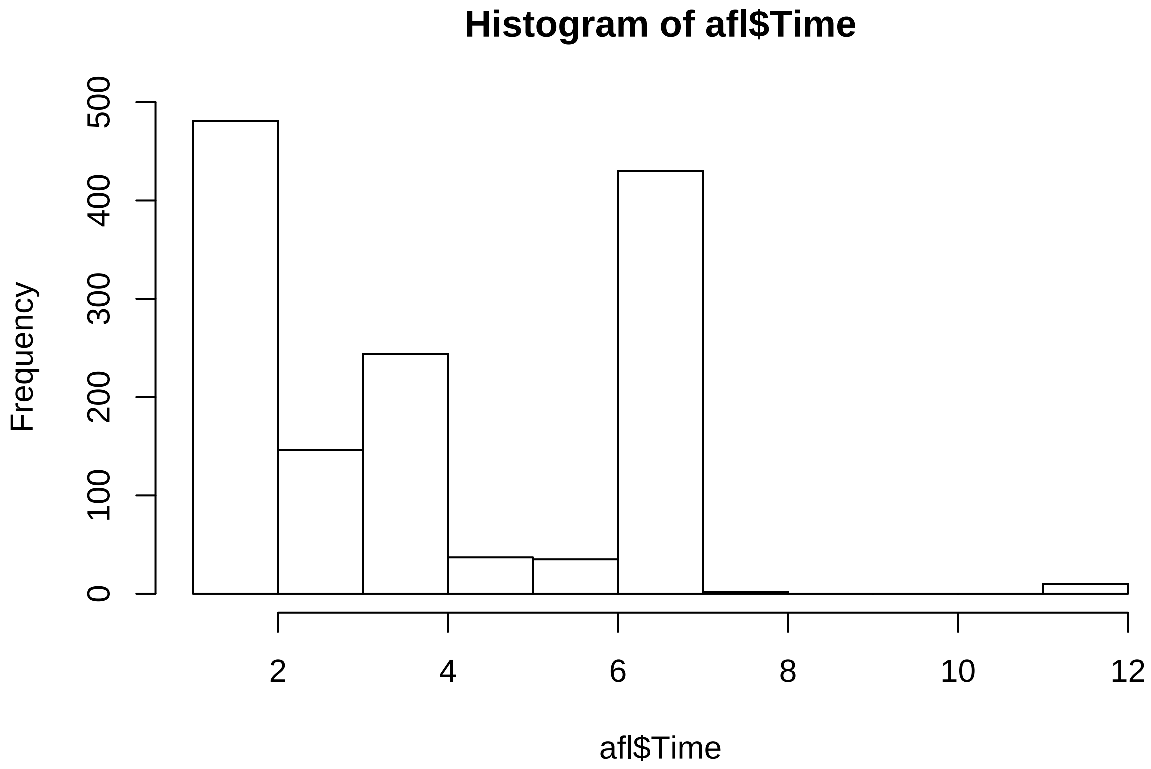
Basic Analysis With Injuries

```
plot(factor(afl$Day))
```



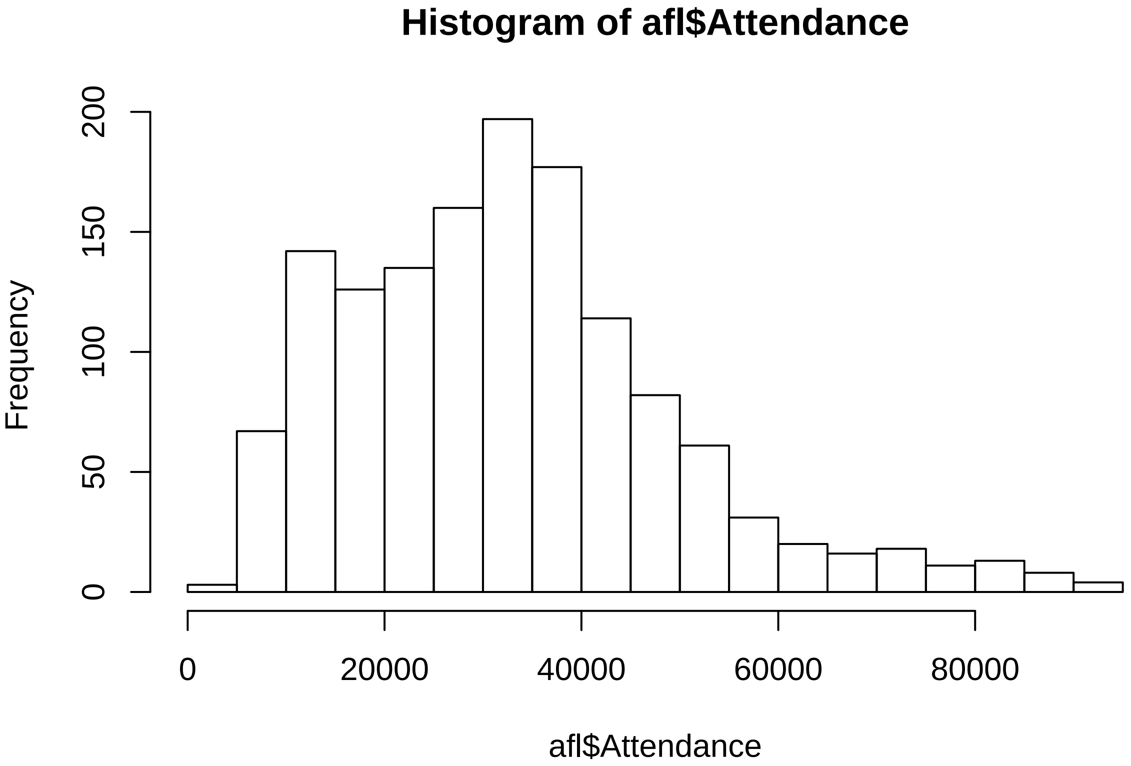
We see that the majority of games are on Saturdays and Sundays. We can perform further analysis to determine how day of week affects attendance.

```
hist(afl$Time)
```

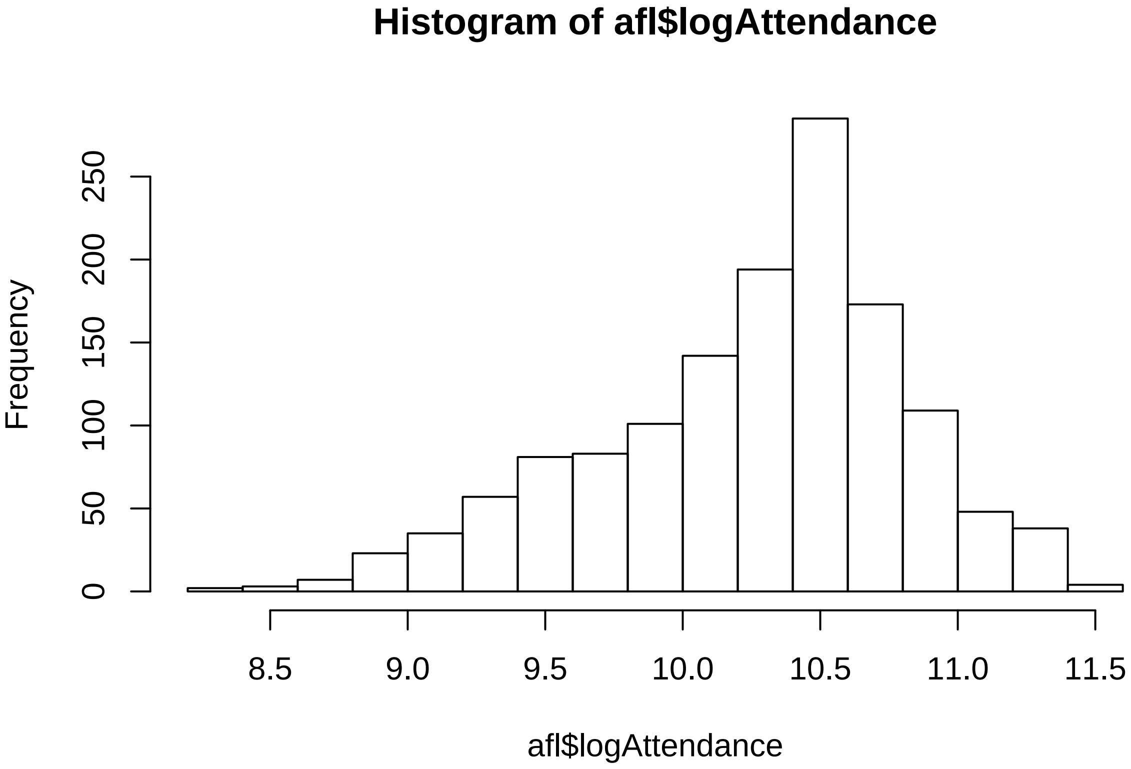


Time bucket allows us to split the times into more easily workable data. We can see here that most of the games occur in the 1 o'clock and the 7 o'clock hours (start between 1 and 2 and between 7 and 8). We can perform further analysis to determine if time of day affects attendance. Possible interaction between time of day and day of week.

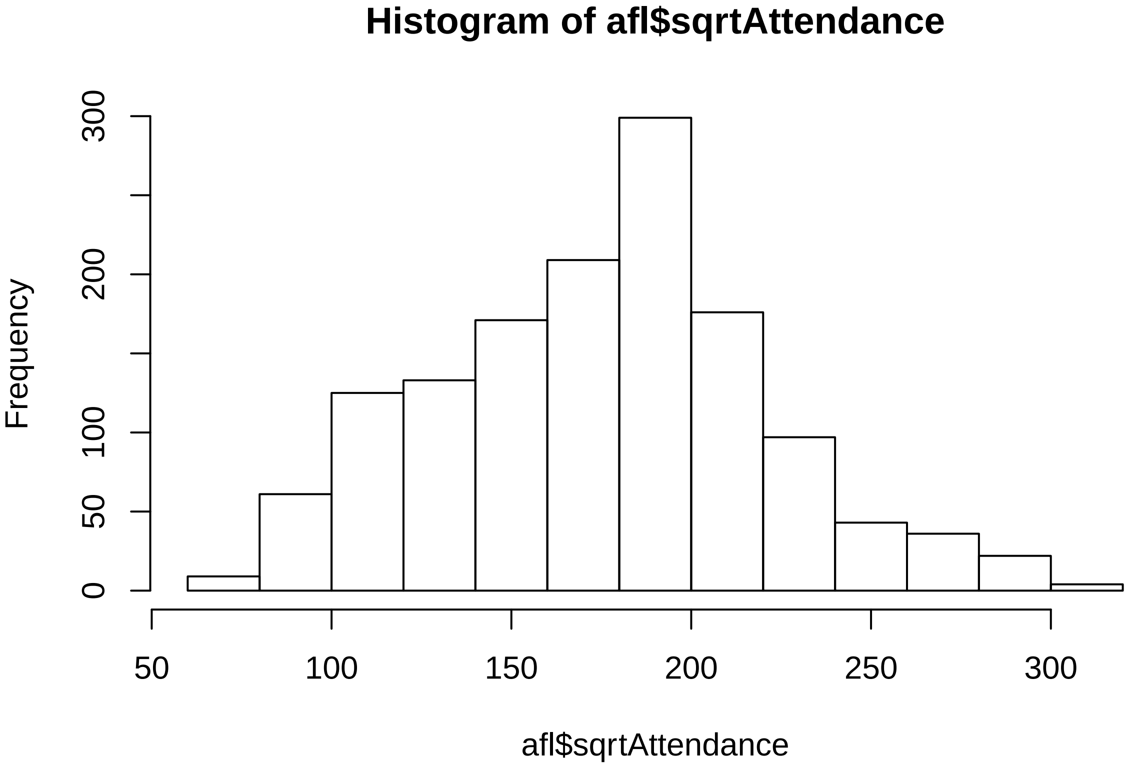
```
hist(afl$Attendance, breaks = 30)
```



```
afl$logAttendance = log(afl$Attendance)
hist(afl$logAttendance)
```

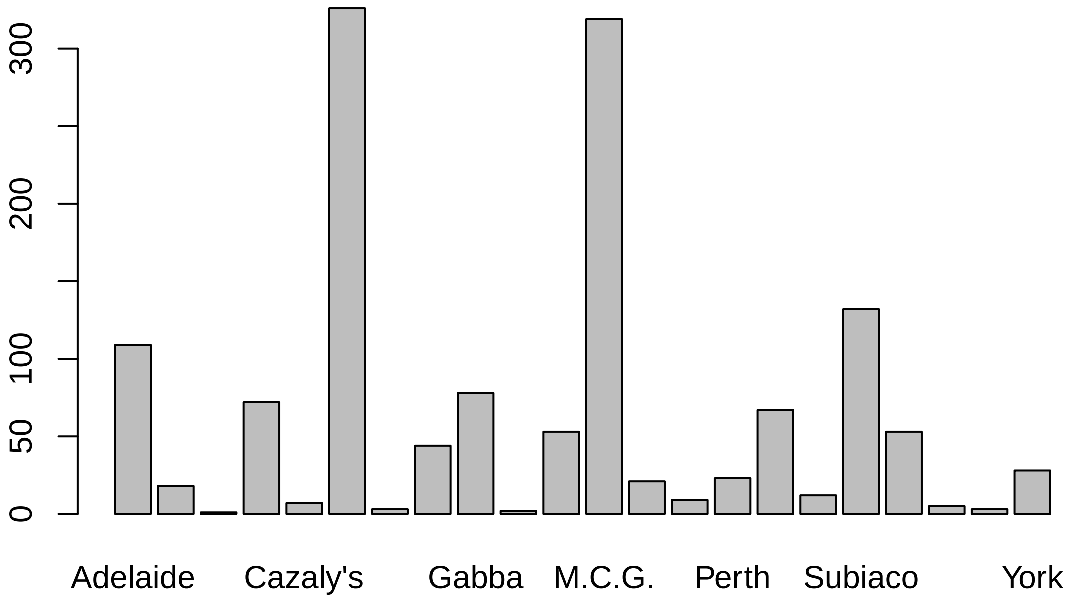


```
afl$sqrtAttendance = sqrt(afl$Attendance)
hist(afl$sqrtAttendance)
```



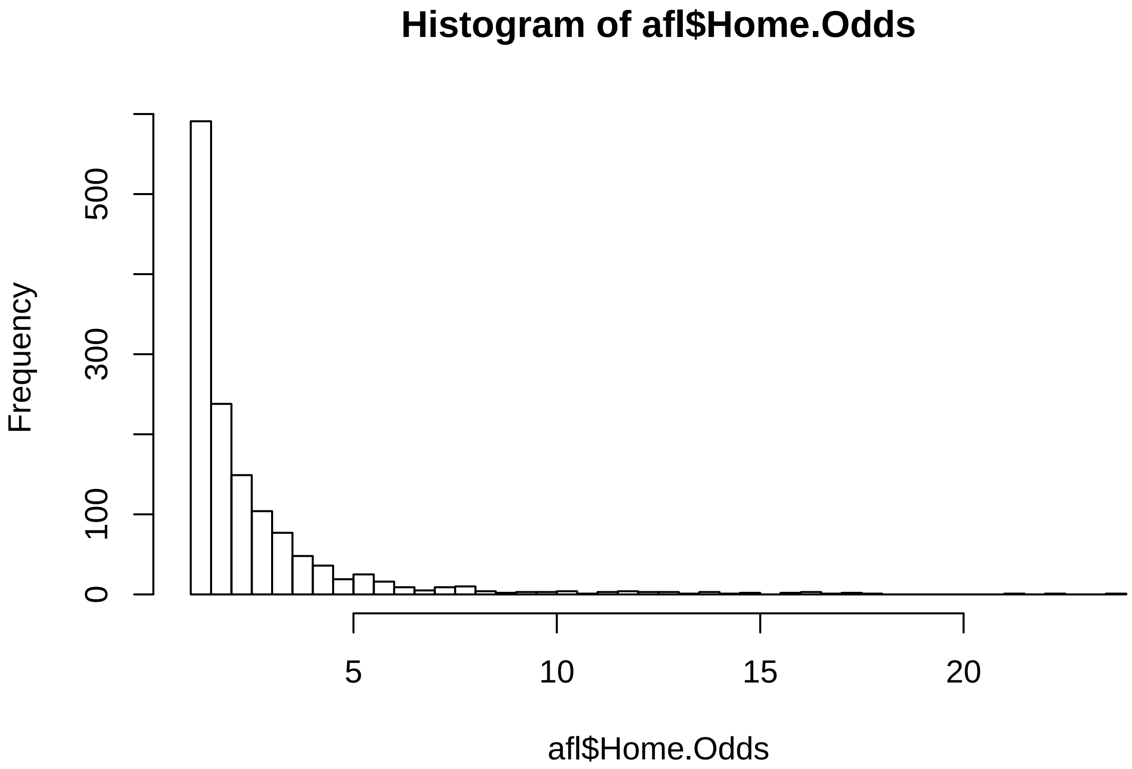
Attendance may be our response variable as we are trying to determine the factors which influence fans to attend games. The histogram shows a right skew, so I tried an exponential transformation. This results in a left-skewed histogram. As such, I tried a square root transformation which gives us the most normal distribution of the three.

```
plot(factor(afl$Venue))
```

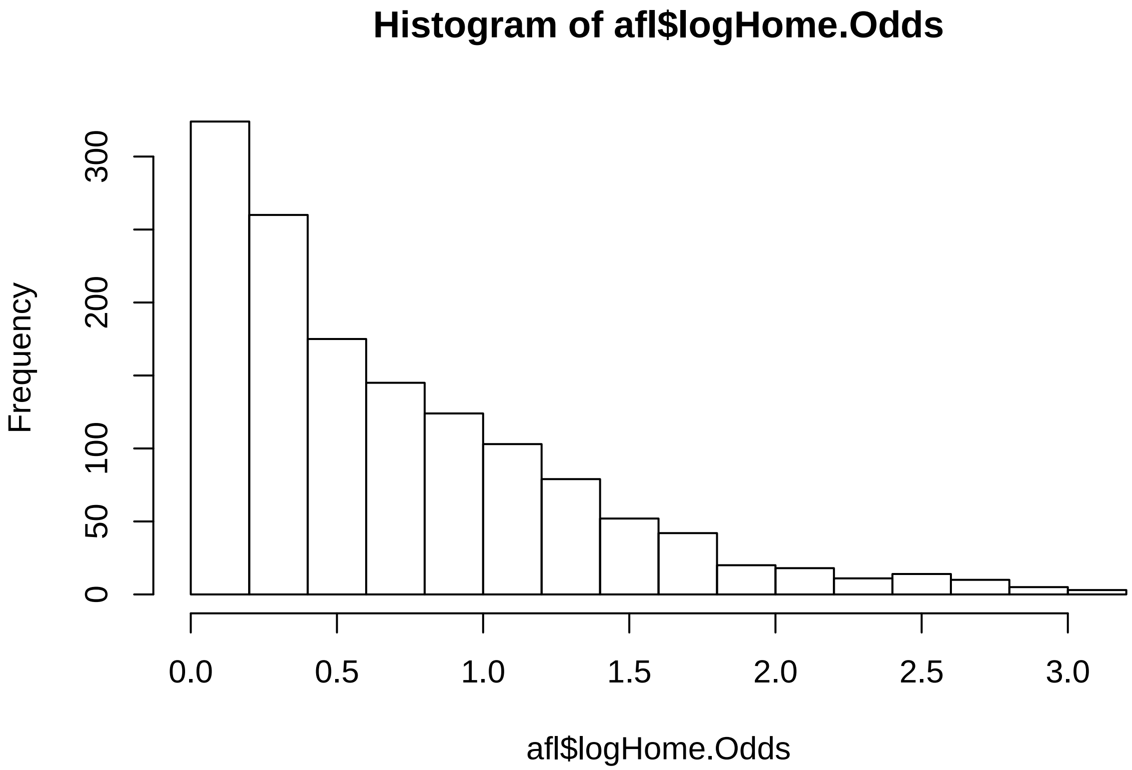


This plot simply shows which stadiums host the most games. We see that a majority of the games are played at M.C.G. and Docklands.

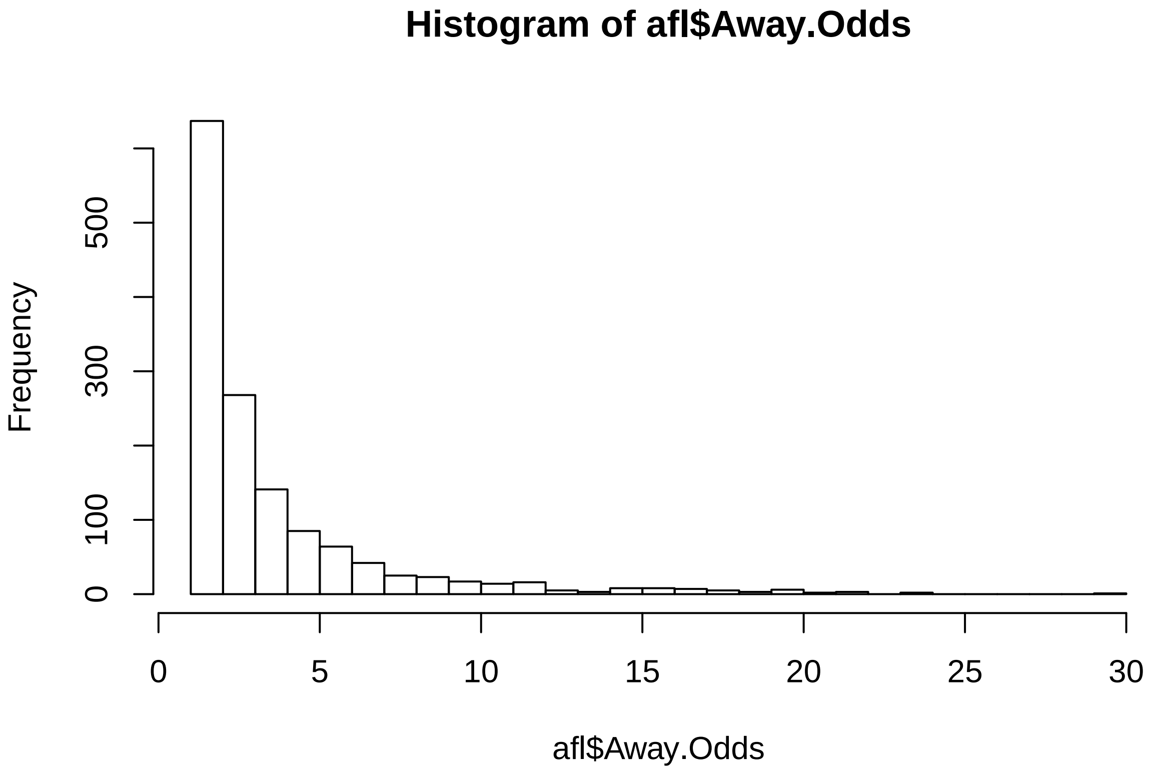
```
hist(afl$Home.Odds, breaks = 40)
```



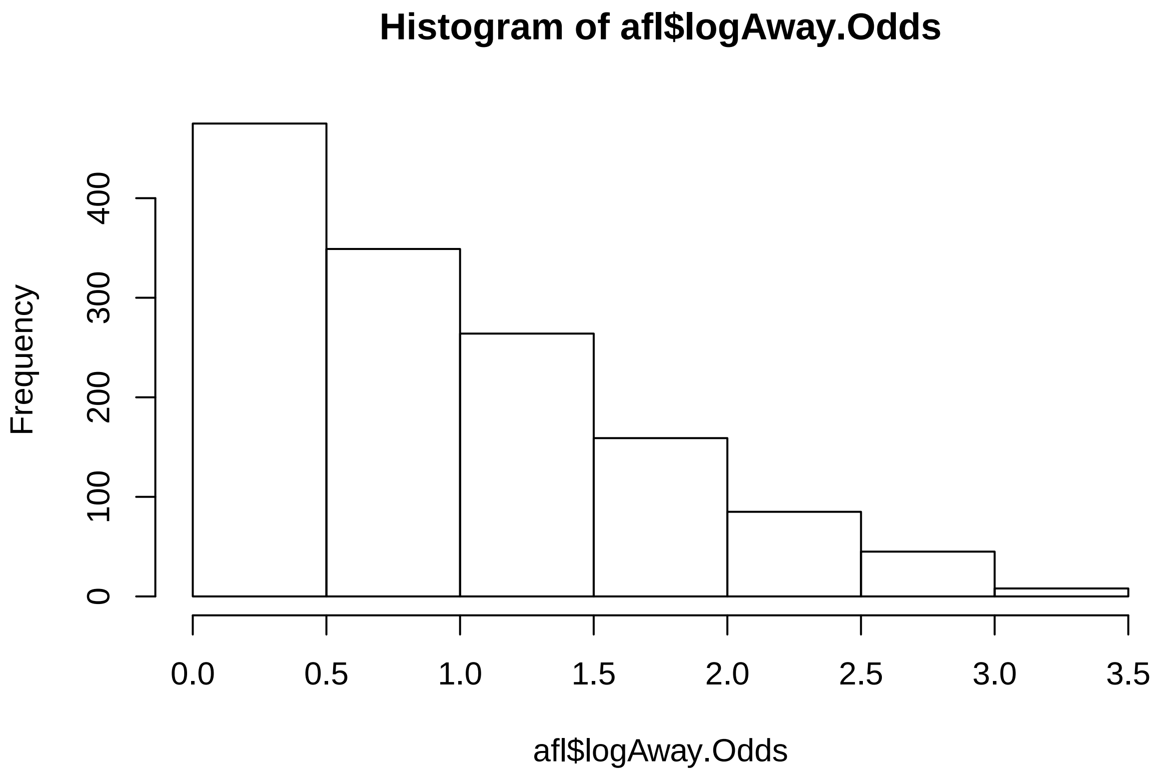
```
afl$logHome.Odds = log(afl$Home.Odds)
hist(afl$logHome.Odds)
```



```
hist(afl$Away.Odds, breaks = 40)
```



```
afl$logAway.Odds = log(afl$Away.Odds)
hist(afl$logAway.Odds)
```

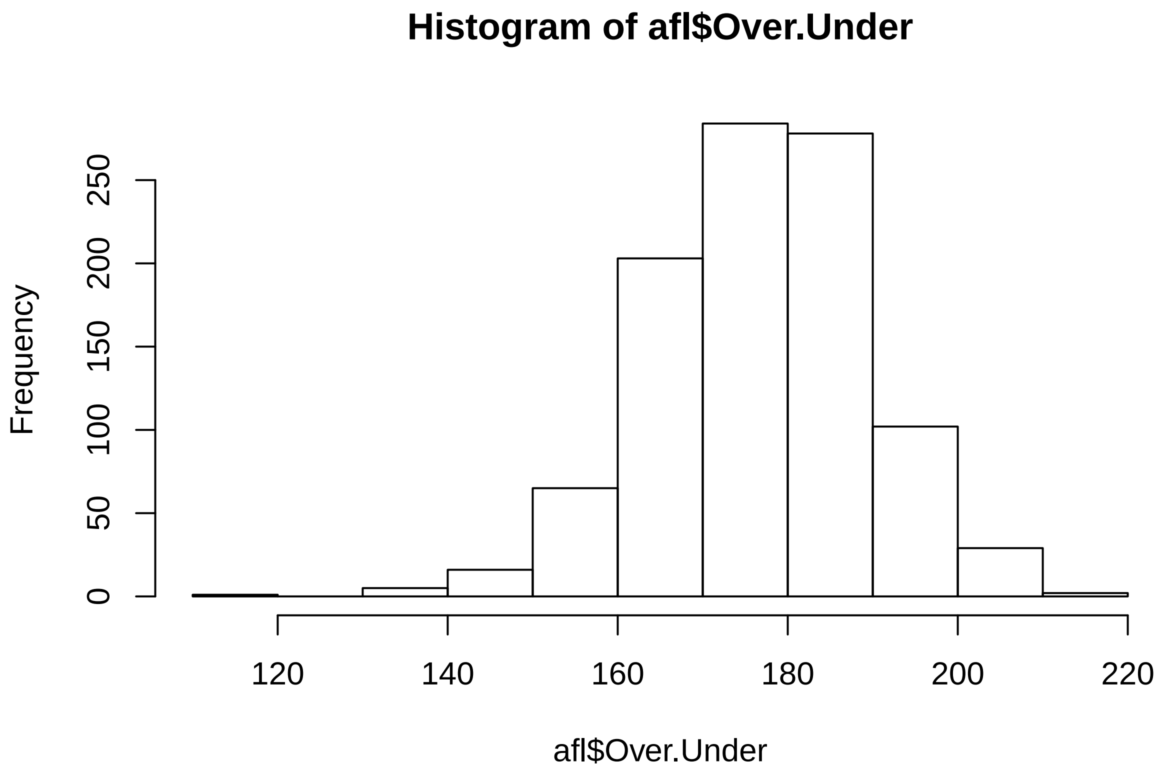


Both home odds and away odds are extremely skewed. Taking the log of the log creates a much more difficult result to interpret and still doesn't give us normally distributed data. These variables will be two of the most important of our data, but as they correlated we can only use one. I show other ways we can incorporate this data below.

```
afl$Over.Under <- as.numeric(afl$Over.Under)
```

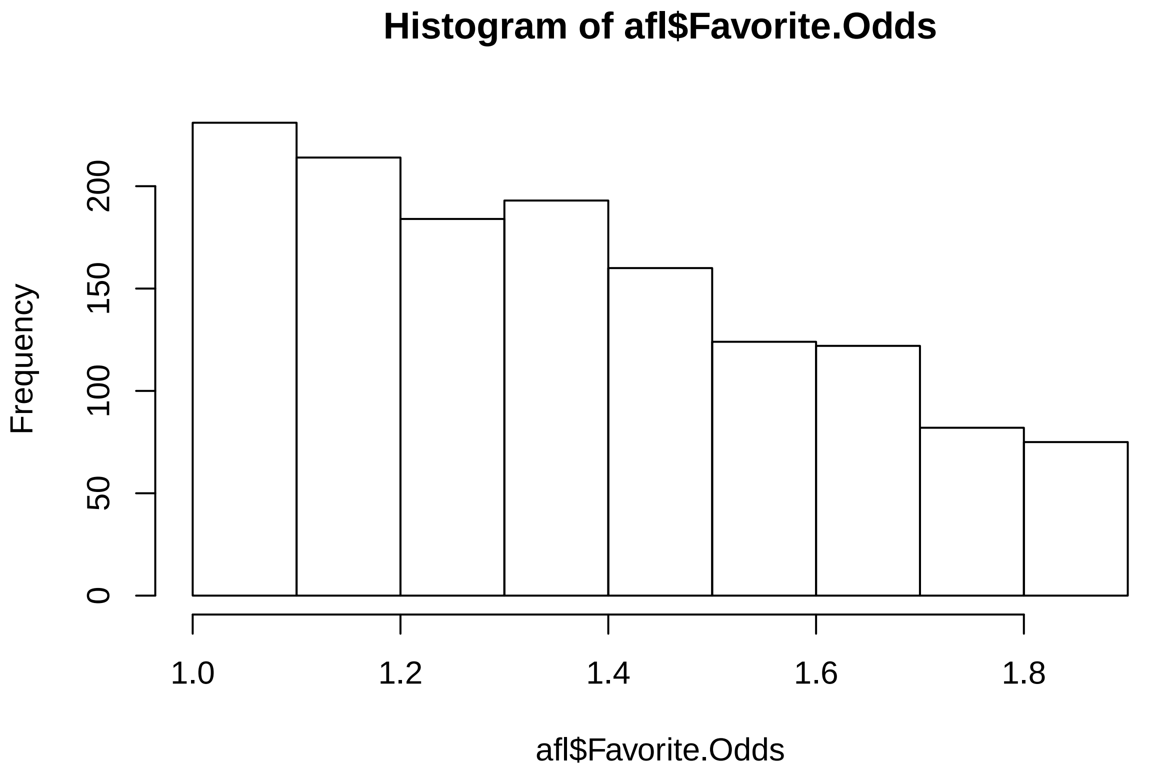
```
## Warning: NAs introduced by coercion
```

```
hist(afl$Over.Under)
```

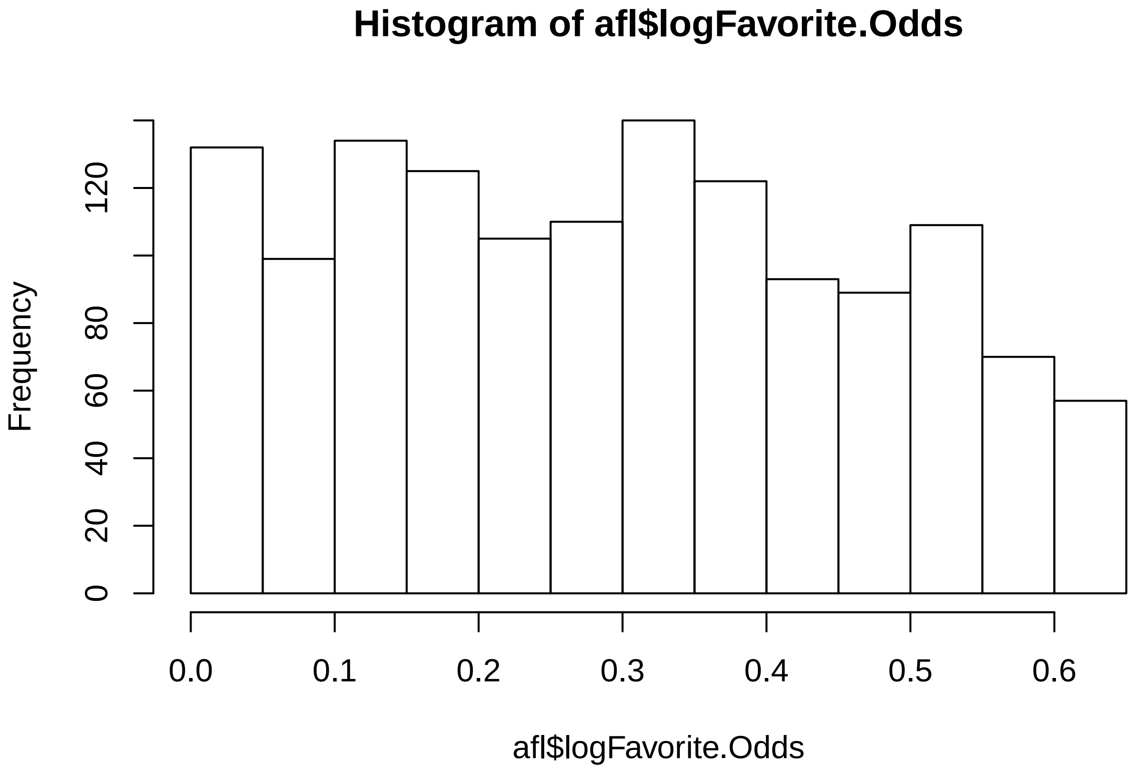


The data for the over-under of the games is relatively normally distributed. However, only games from 2014 and later have over-under data.

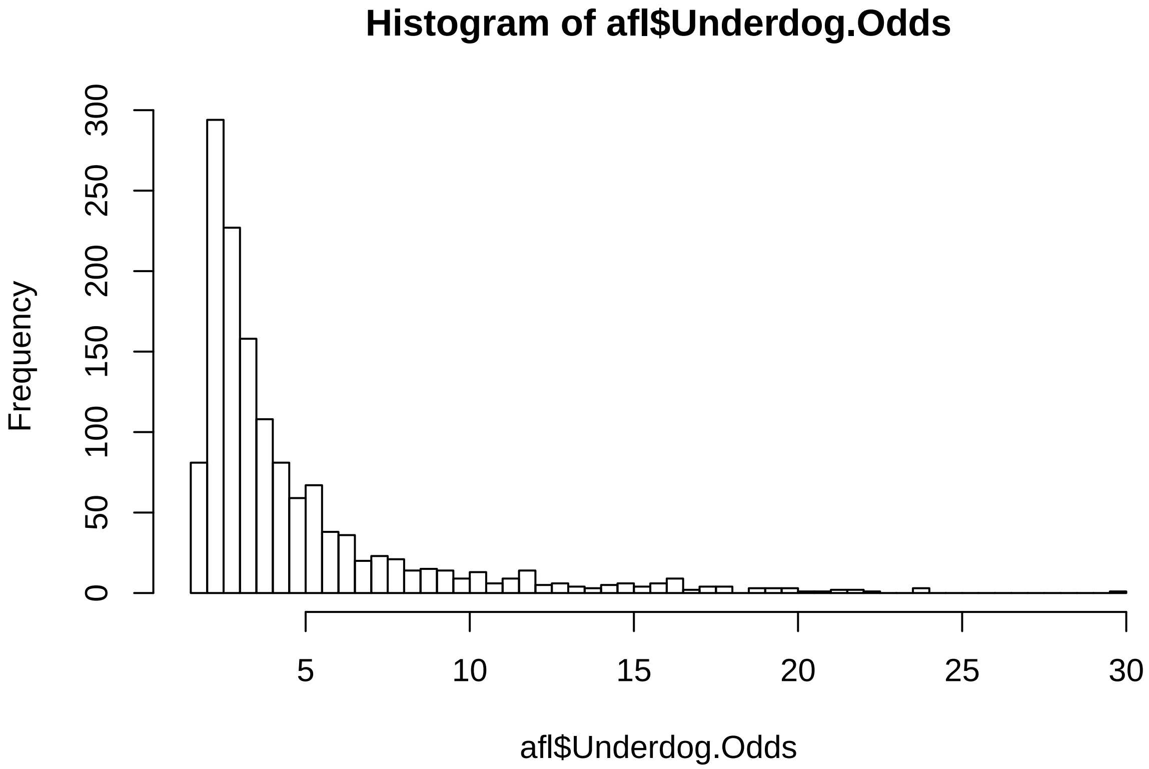
```r
hist(afl$Favorite.Odds, breaks = 10)
```



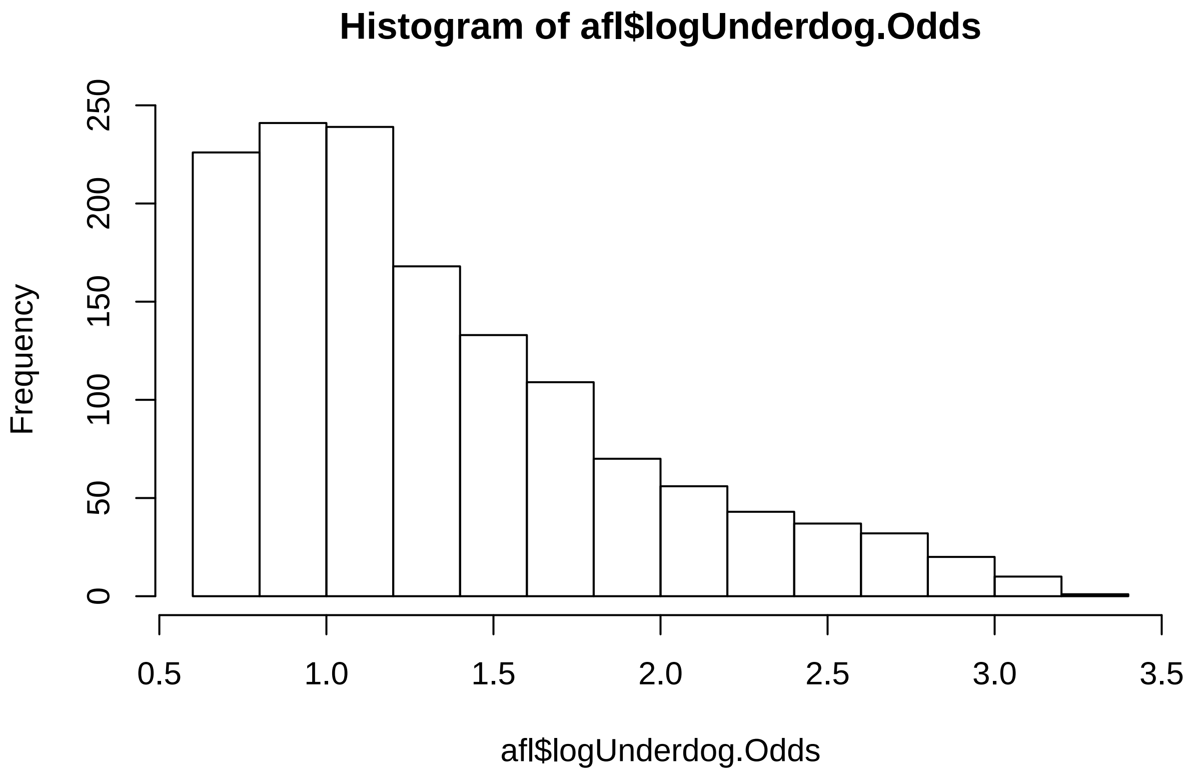
```
afl$logFavorite.Odds = log(afl$Favorite.Odds)
hist(afl$logFavorite.Odds)
```



```
hist(afl$Underdog.Odds, breaks = 40)
```

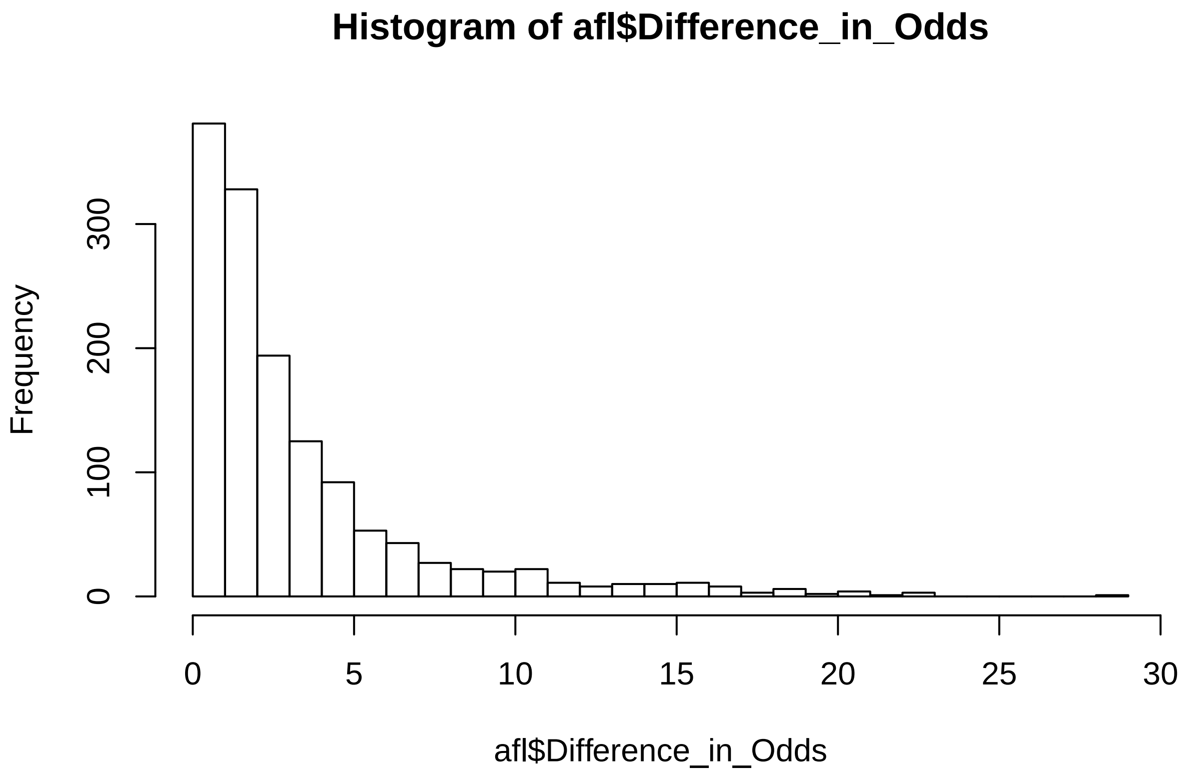


```
afl$logUnderdog.Odds = log(afl$Underdog.Odds)
hist(afl$logUnderdog.Odds)
```

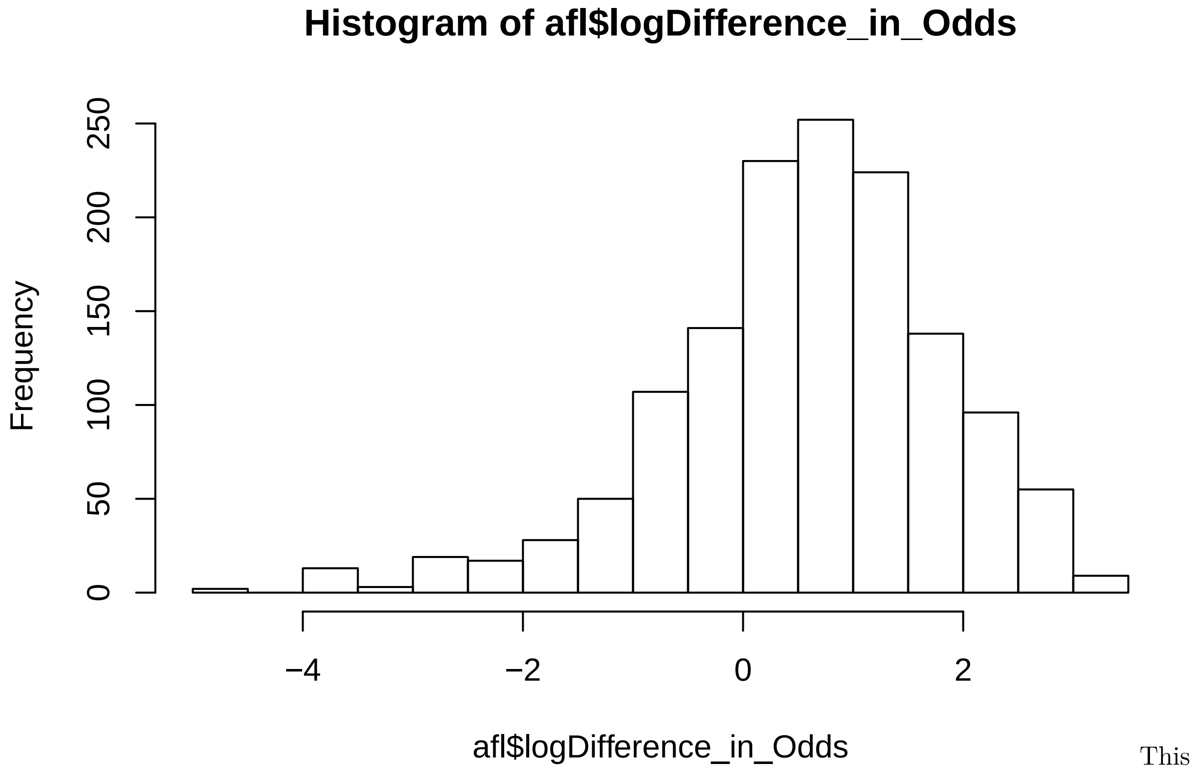


The favorite odds and the underdog odds also aren't great to work with as they have huge amounts of right skewness as well. These variables will also require further transformation if we are to use them.

```
hist(afl$Difference_in_Odds, breaks = 25)
```

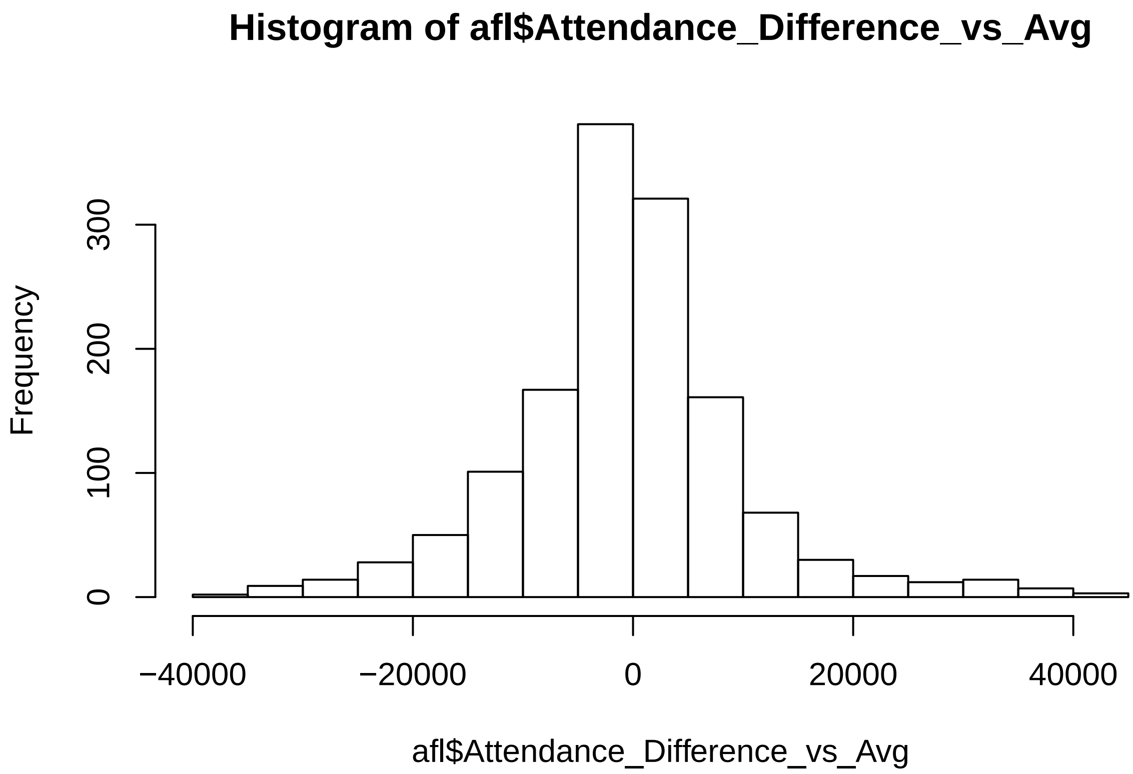


```r
afl$logDifference_in_Odds = log(afl$Difference_in_Odds)
hist(afl$logDifference_in_Odds)
```



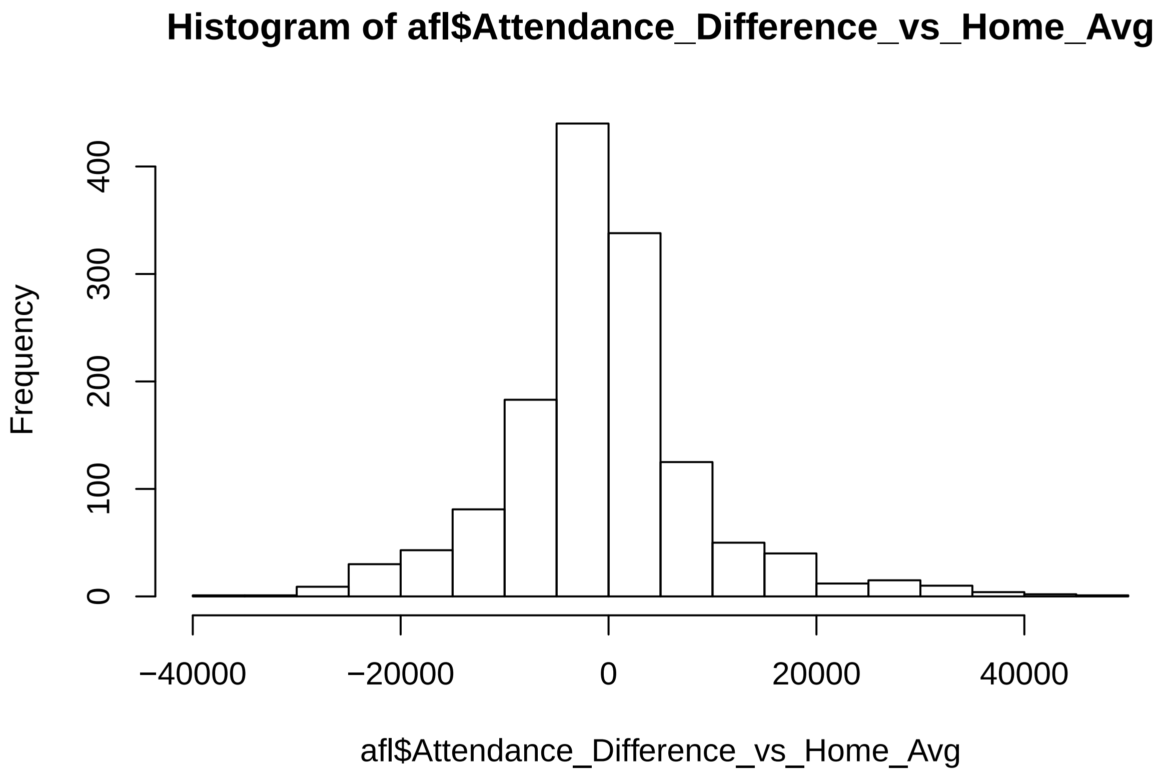
This is my personal favorite of the odds variables. It shows the difference between the odds of the favorite and the odds of the underdog. It also has a severe right skew but the log transformation works to perfection and creates an almost perfectly normal histogram (with slight left skew). I believe this variable provides the most pertinent information regarding how far apart the teams are in chances of winning the game. It could potentially be used in conjunction with home odds to create a model as the two may be able to provide different information.

```
hist(afl$Attendance_Difference_vs_Avg)
```



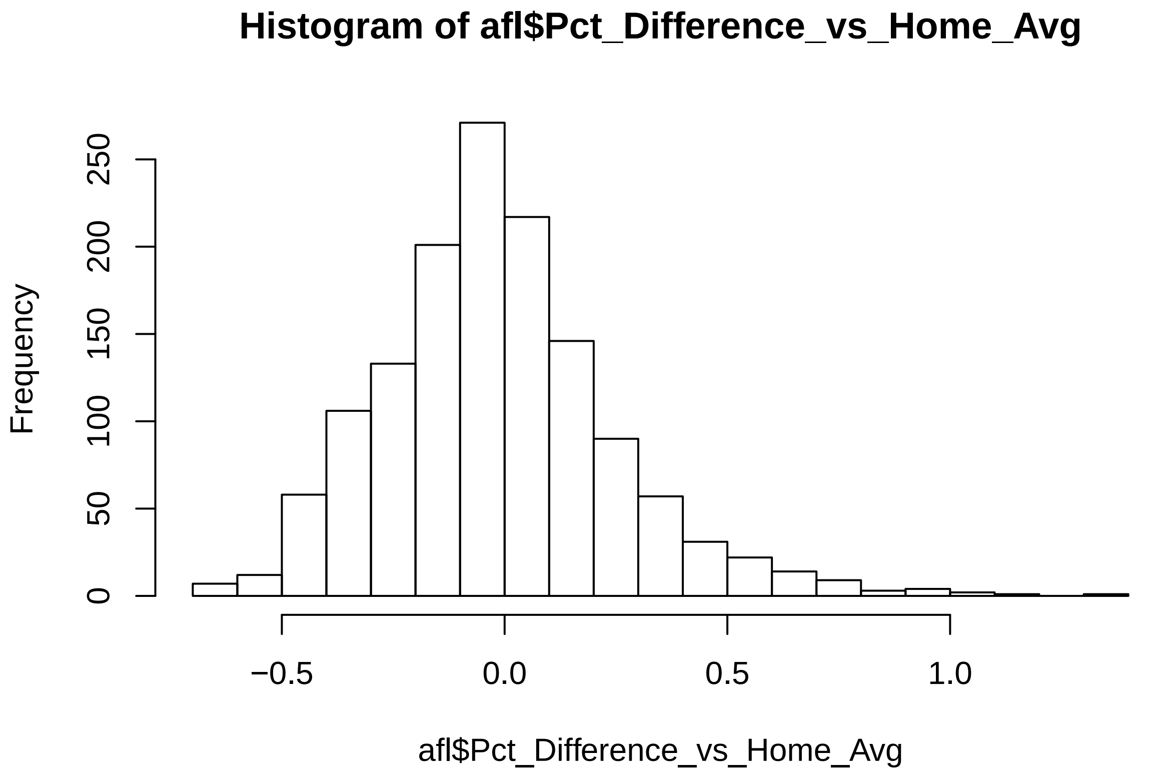
Attendance differential shows the difference in attendance between the actual versus the average attendance (for that venue) for each game. The data are relatively normal although the tails are slightly long.

```
hist(afl$Attendance_Difference_vs_Home_Avg, breaks = 14)
```



This is the same as differential but it shows the difference between the actual versus average attendance at the venue but only for the home team's games (teases out variance due to home team).

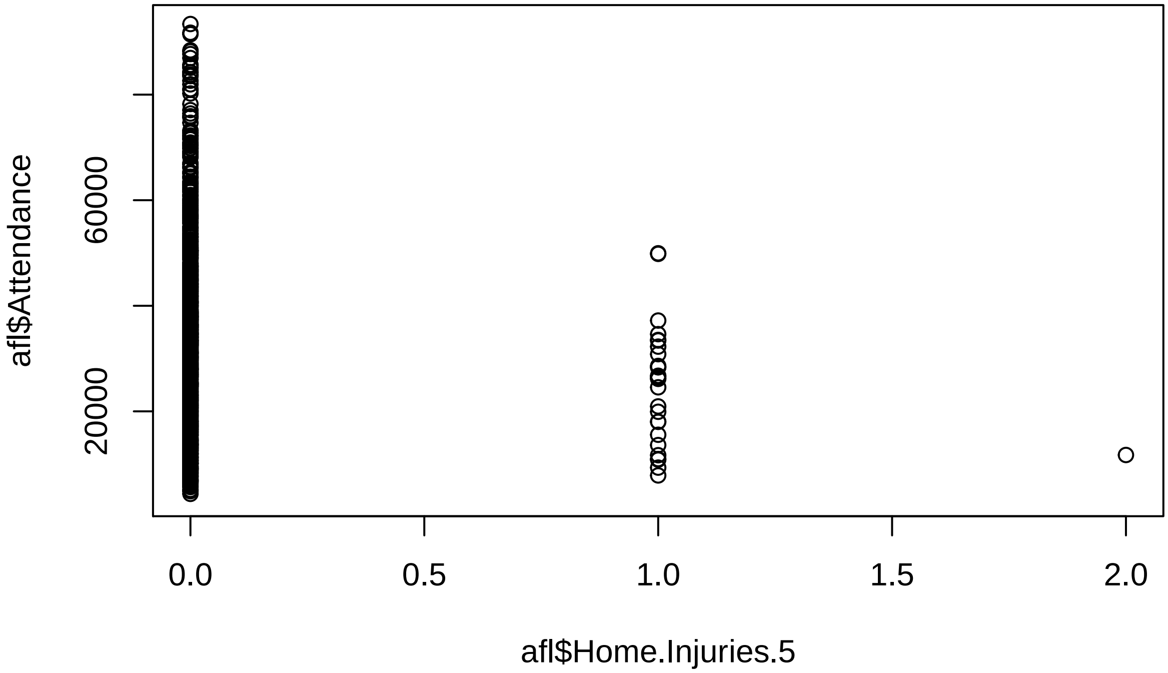
```
hist(afl$Pct_Difference_vs_Home_Avg, breaks = 20)
```



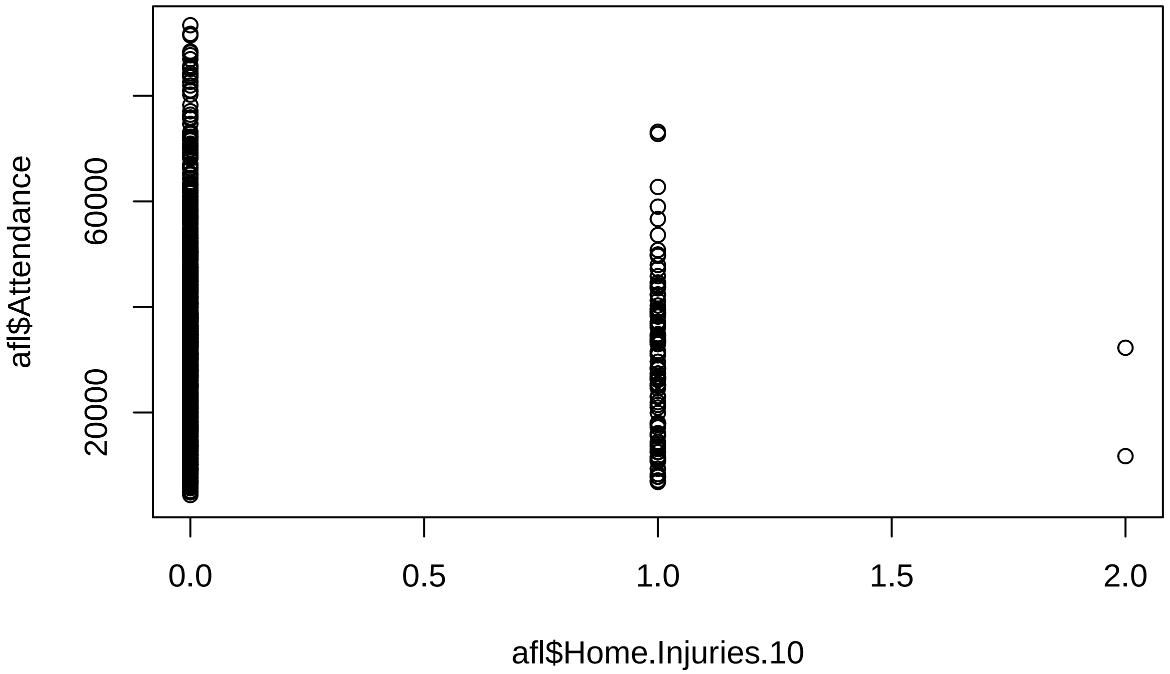
This is my favorite attendance statistic because it shows the percentage difference between the team's average home attendance at a specific venue and the actual attendance for every individual game. This gives better information than strictly numerical values because it scales the change based on how many people generally attend that team's home games at that stadium.

```
## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion
```

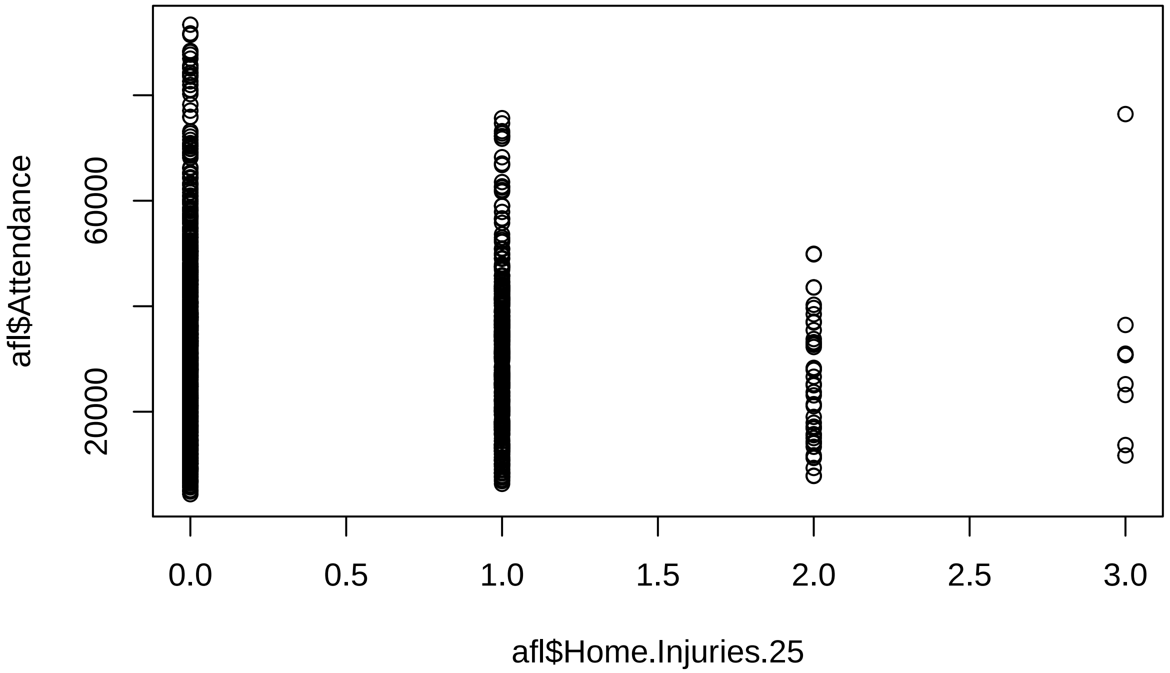
```
plot(afl$Home.Injuries.5, afl$Attendance)
```



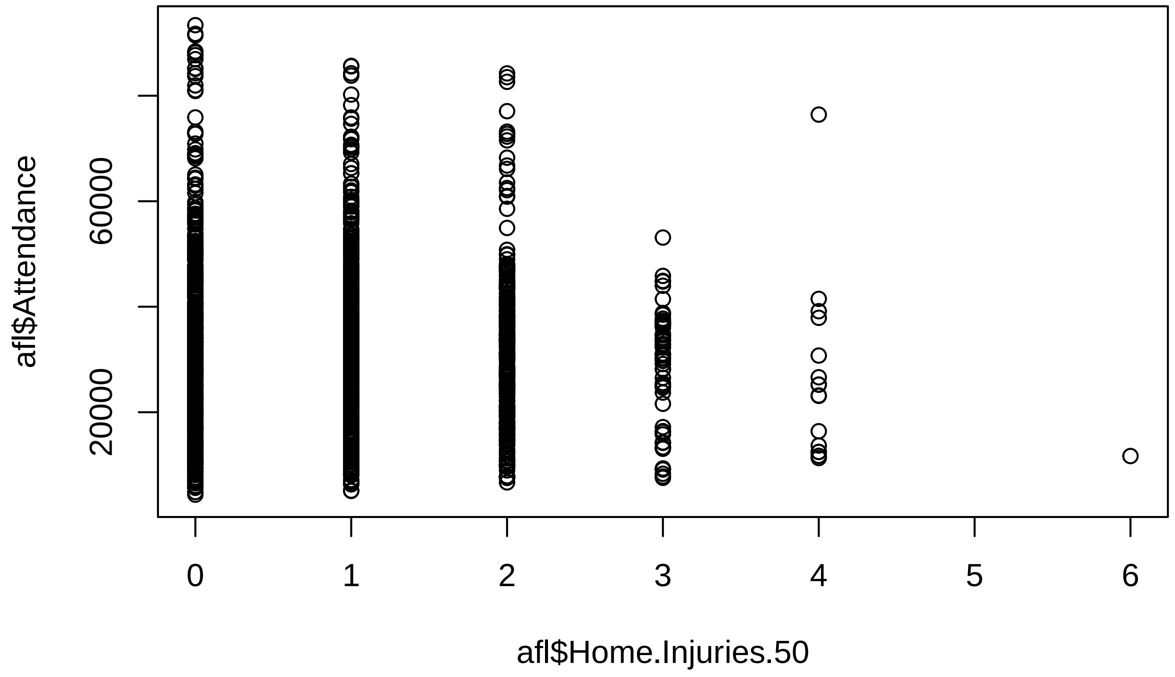
```
plot(afl$Home.Injuries.10, afl$Attendance)
```



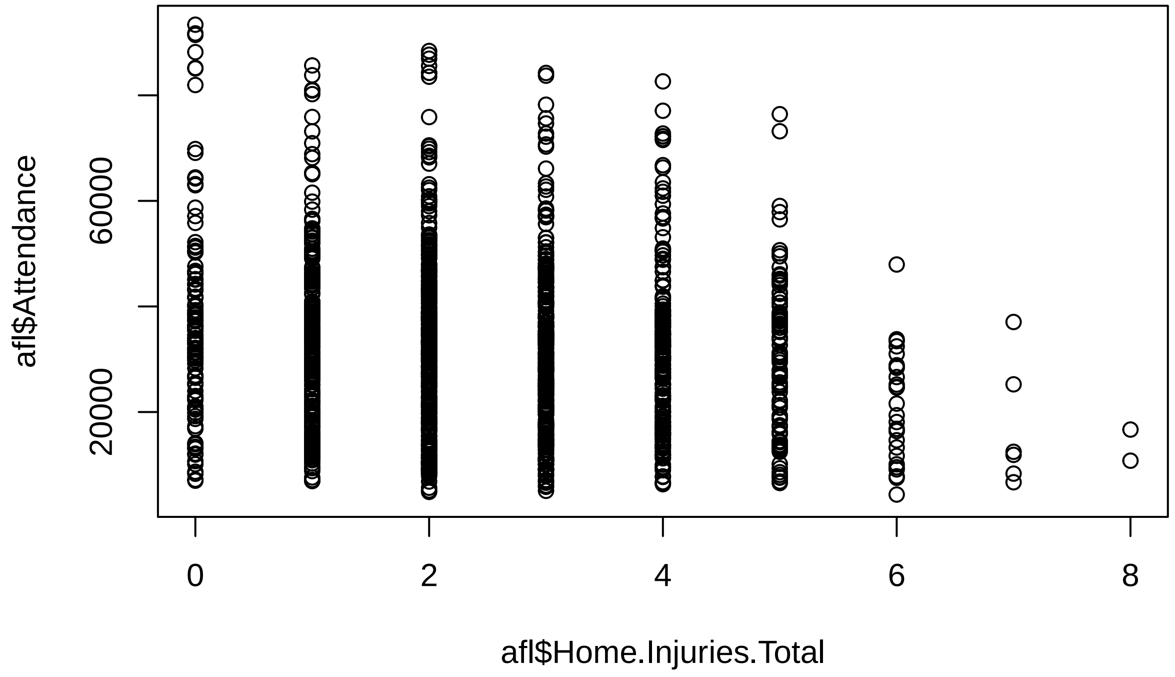
```r
plot(afl$Home.Injuries.25, afl$Attendance)
```



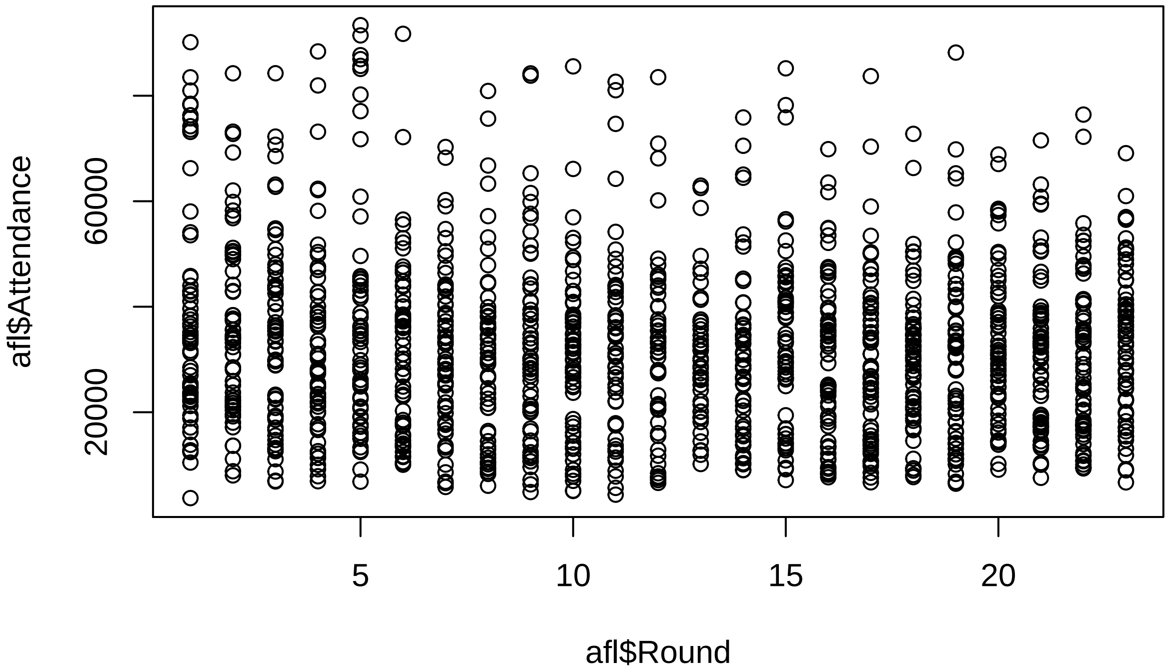
```r
plot(afl$Home.Injuries.50, afl$Attendance)
```



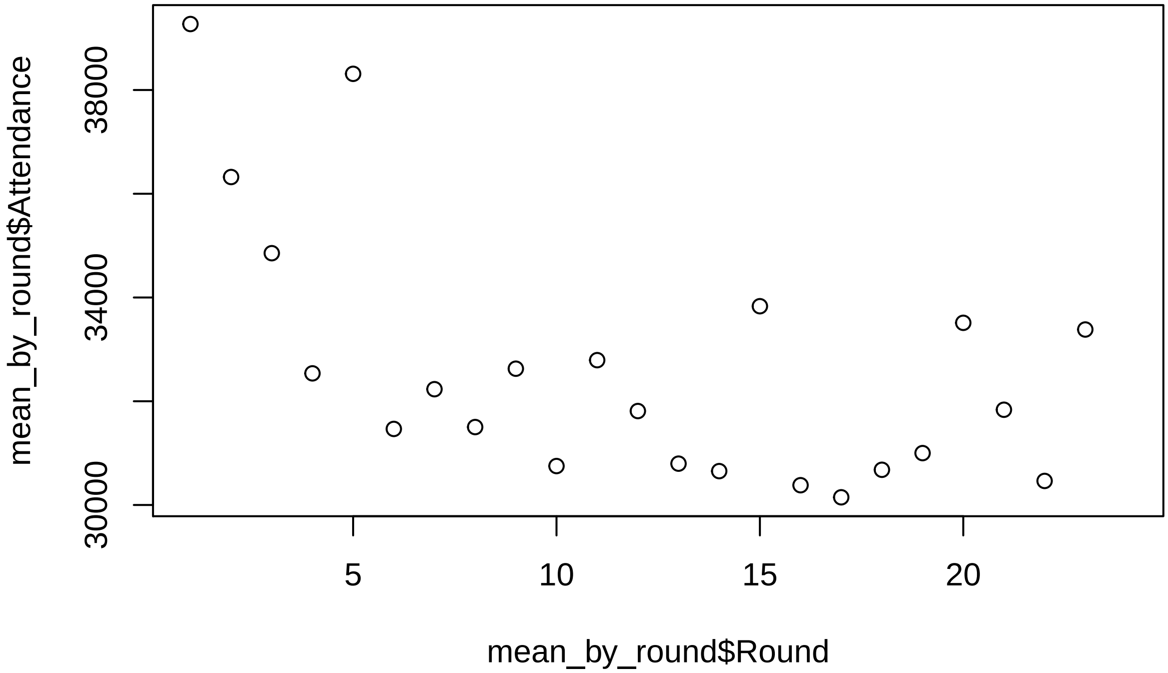
```
plot(afl$Home.Injuries.Total, afl$Attendance)
```



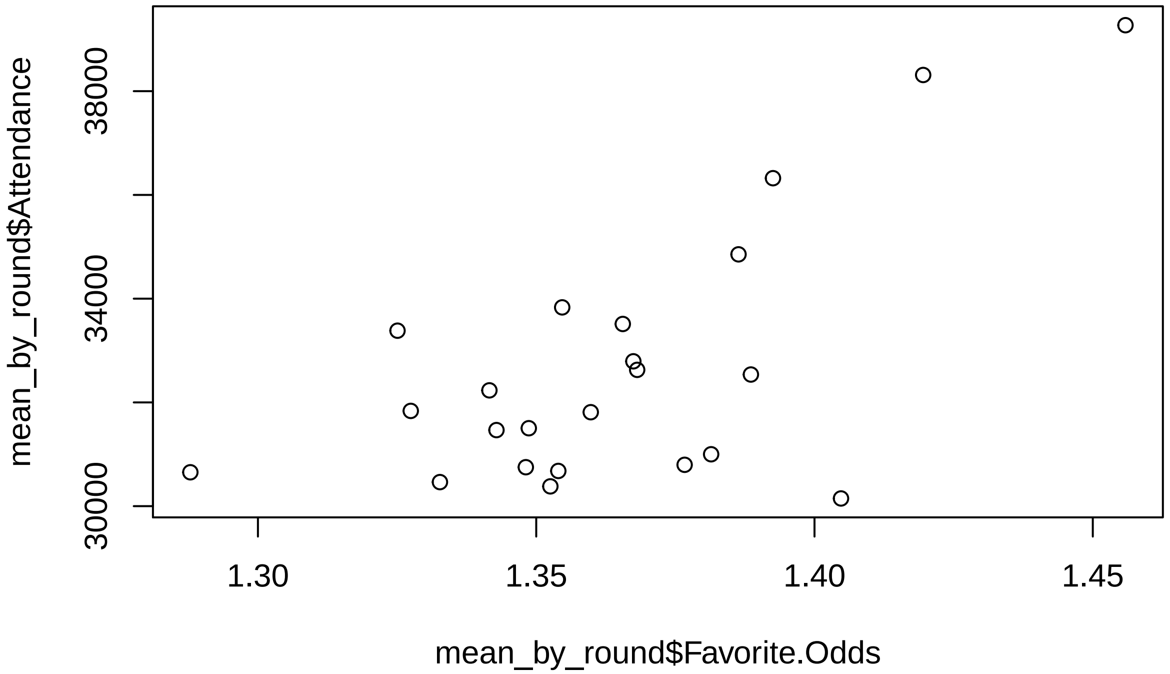
```
plot(afl$Attendance~afl$Round)
```



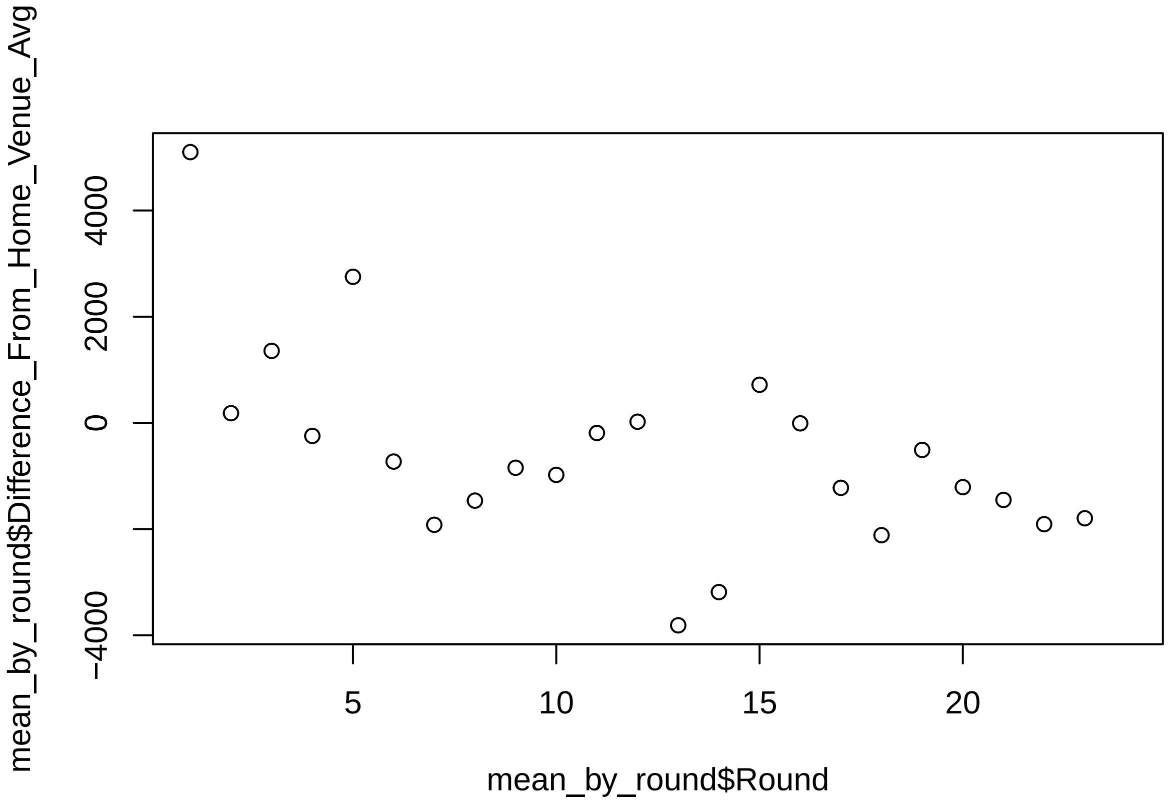
```
plot(mean_by_round$Round, mean_by_round$Attendance)
```



```
plot(mean_by_round$Favorite.Odds, mean_by_round$Attendance)
```

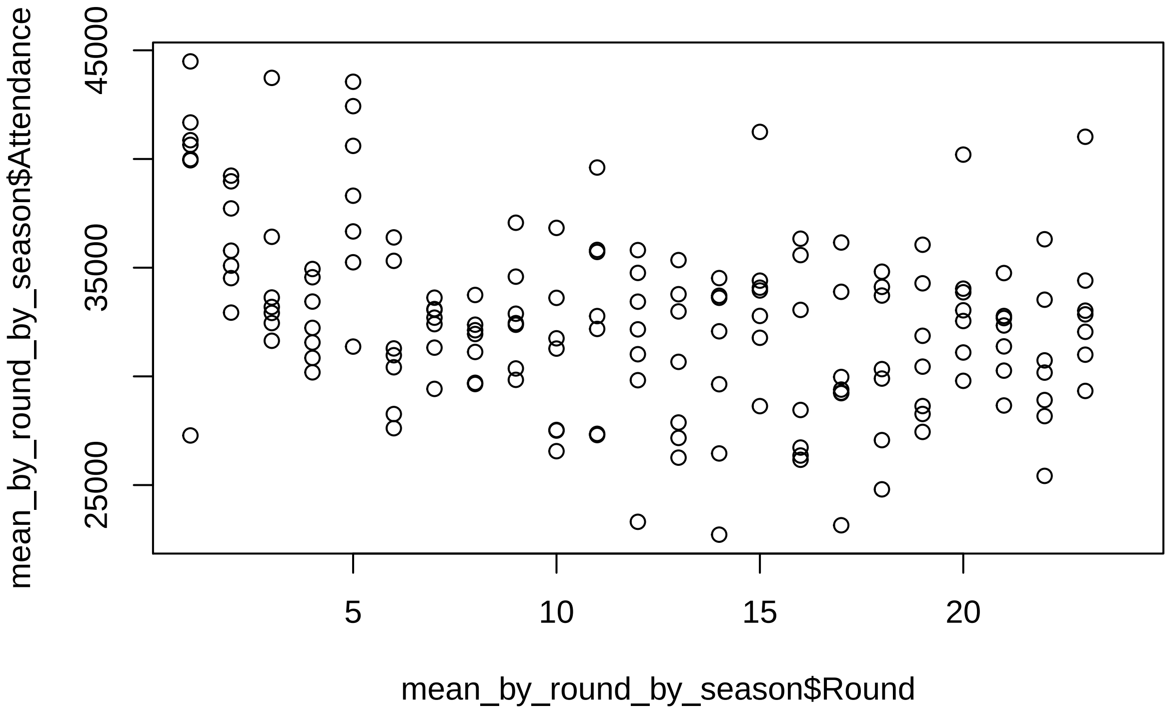


```
plot(mean_by_round$Round, mean_by_round$Difference_From_Home_Venue_Avg)
```

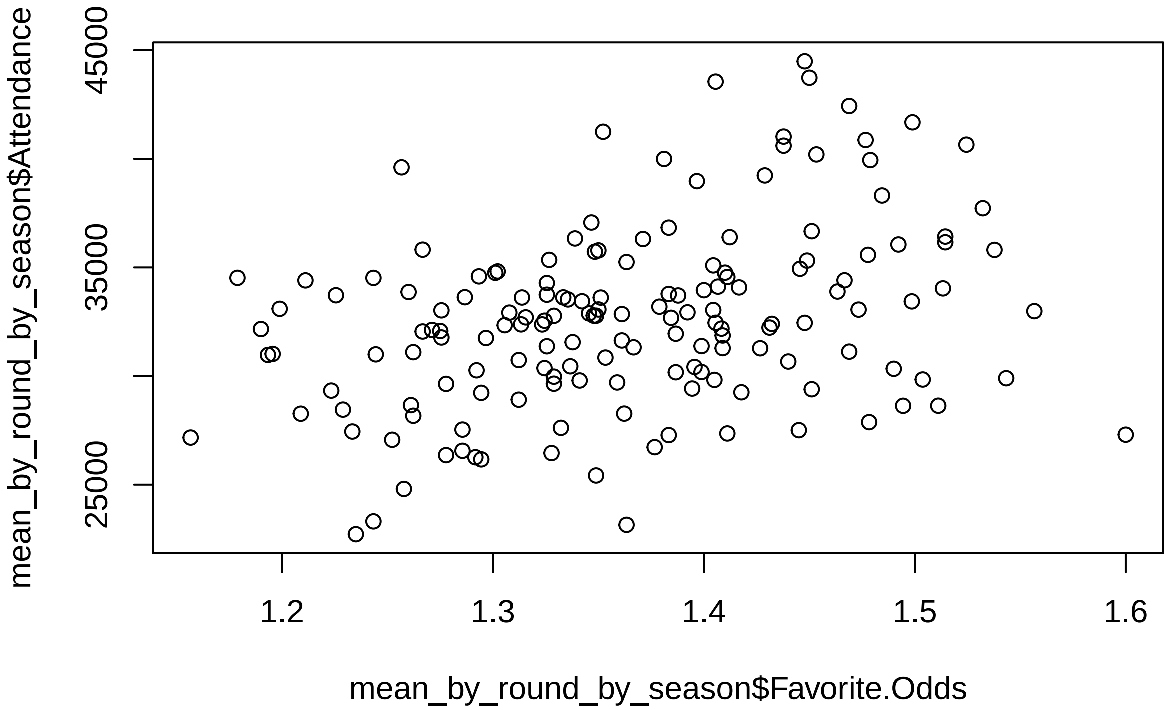


From this table we see that Week 1 is generally an outlier. Attendances are far higher than normal based on all metrics while odds are much closer together than in any other week. Week 23 generally shows the opposite results. Week 1 games are also generally played later at night. On average, 11% more fans show up on Week 1 for teams that host games at that particular venue. 6% less fans than average attend games in Week 23 and Week 13. Week 24 likely doesn't provide good data as the sample size is extremely small (only one season had a 'Week 24').

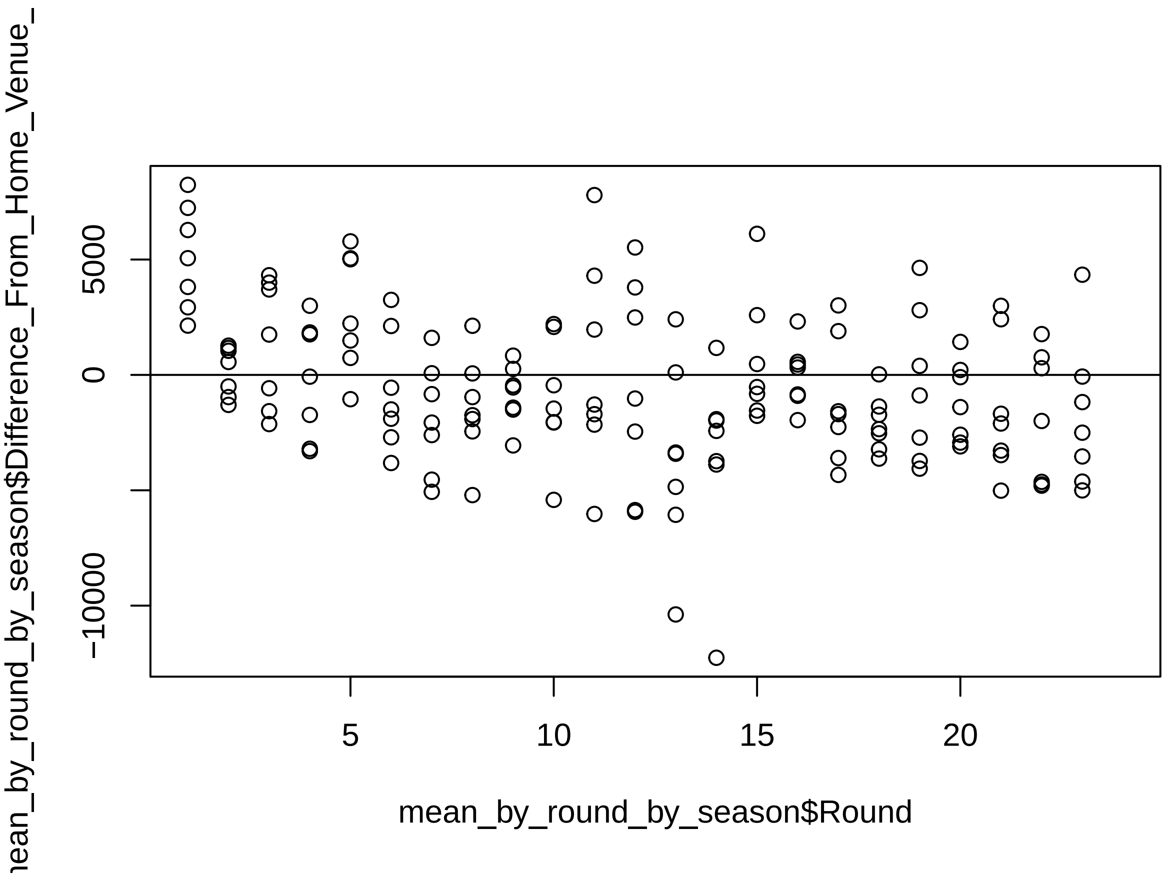
```r
plot(mean_by_round_by_season$Round, mean_by_round_by_season$Attendance)
```



```r
plot(mean_by_round_by_season$Favorite.Odds, mean_by_round_by_season$Attendance)
```



```r
plot(mean_by_round_by_season$Round, mean_by_round_by_season$Difference_From_Home_Venue_Avg)
abline(h=0)
```


5000
0
−10000
5
10
15
20
mean_by_round_by_season$Round
nean_by_round_by_season$Difference_From_Home_Venue_