# UFC Statistics

Team #5

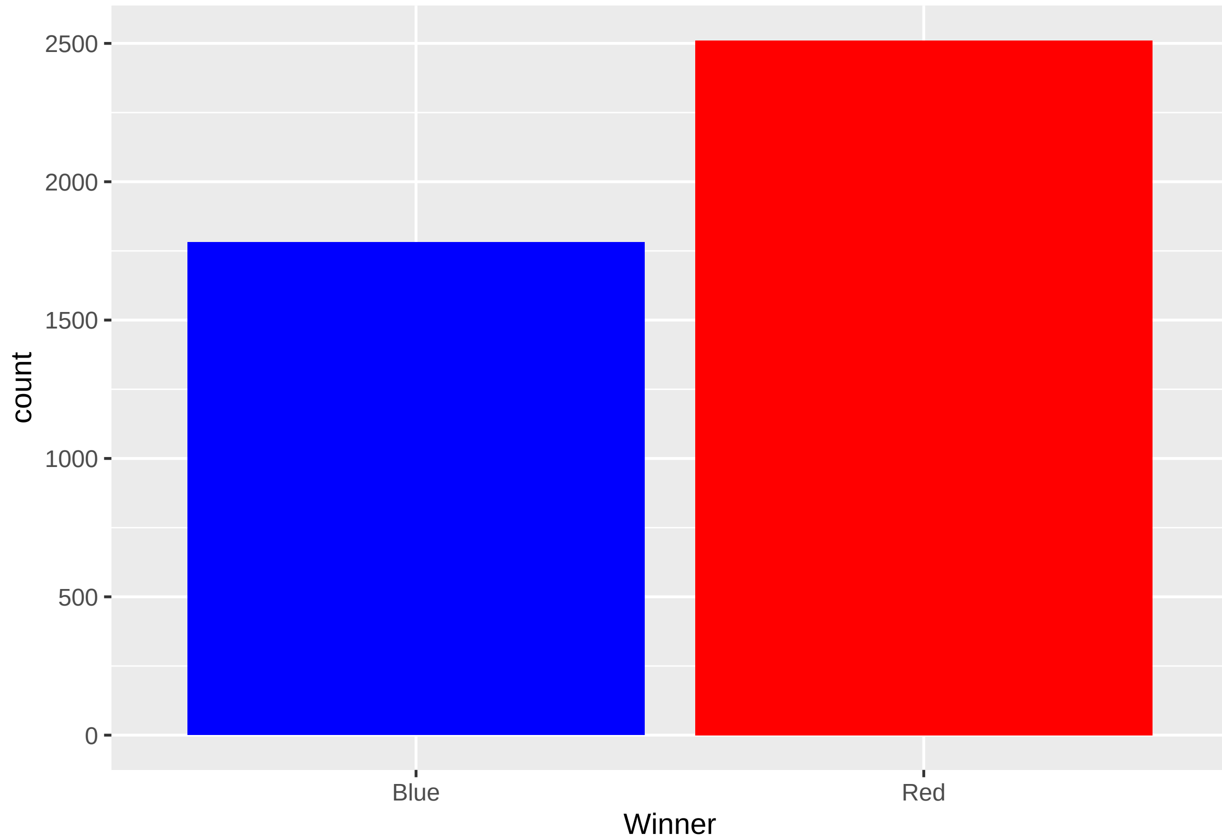
7/24/2020

```r
df <- read.csv('ufc-master.csv')
```

```r
head(df)

names(df)

summary(df)
```

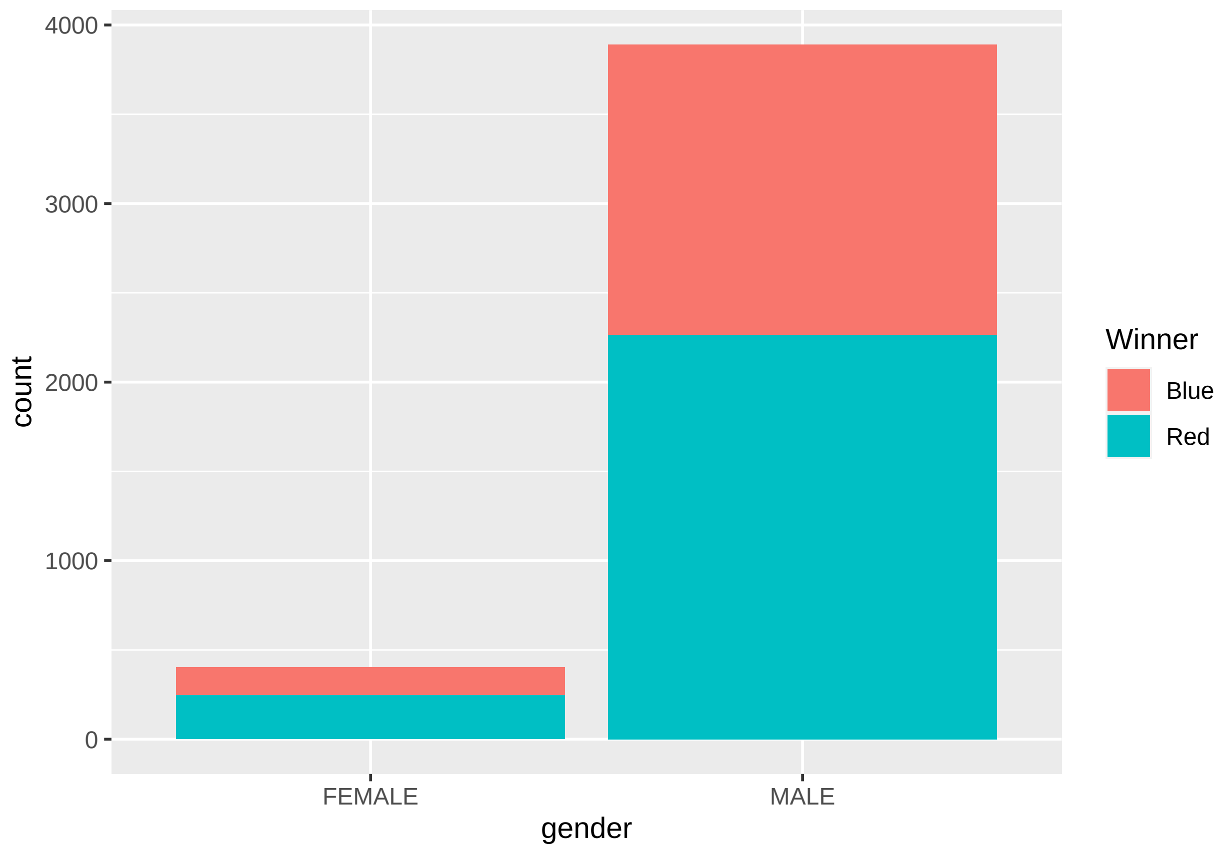
```r
# Question: Does one color have an advantage over the other? Does it change with gender?

# Get the data
color <- df[,c('Winner', 'gender')]

# Plot the graph
color.hist <- ggplot(color, aes(x = Winner))
color.hist +
  geom_histogram(stat = 'count', fill = c('Blue', 'Red'))
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



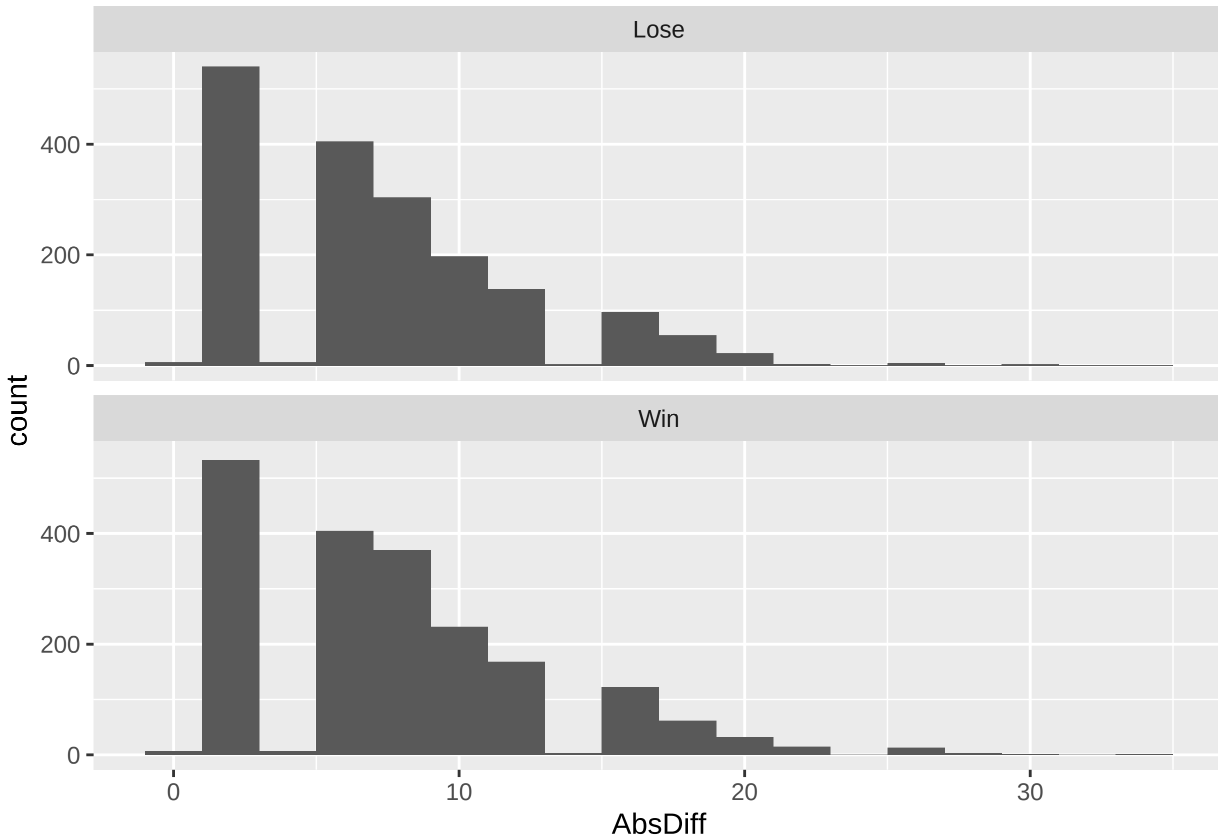

```
# Test of significance ?
```

```
ggplot(df, aes(x = gender, fill = Winner)) +
  geom_bar()
```




```r
# Question: Does the fighter with the longer reach win more frequently?

# Create a smaller data frame
reach <- df[,c('Winner', 'B_Reach_cms', 'R_Reach_cms')]

# Calculate the difference in reach (positive = blue advantage)
reach$Diff <- reach$B_Reach_cms - reach$R_Reach_cms

# There is an outlier that makes no sense, so remove it
reach <- reach[reach$Diff > -50,]

# Remove all cases where the players had equal reach
reach <- subset(reach, !(Diff == 0))

# Identify the fighter with the longer reach
reach$Advantage <-
  case_when(
    reach$Diff > 0 ~ 'Blue',
    reach$Diff < 0 ~ 'Red'
  )

# Identify if the advantaged fighter won
reach$AdWin <-
  case_when(
    reach$Advantage == reach$Winner ~ 'Win',
    reach$Advantage != reach$Winner ~ 'Lose'
```

```
  )

reach$AdWin <- as.factor(reach$AdWin)

# Take the absolute value of the difference
reach$AbsDiff <- abs(reach$Diff)

# Plot the data
reach.hist <- ggplot(reach, aes(x = AbsDiff))
reach.hist +
  geom_histogram(binwidth = 2) +
  facet_wrap(~ AdWin, ncol = 1)
```



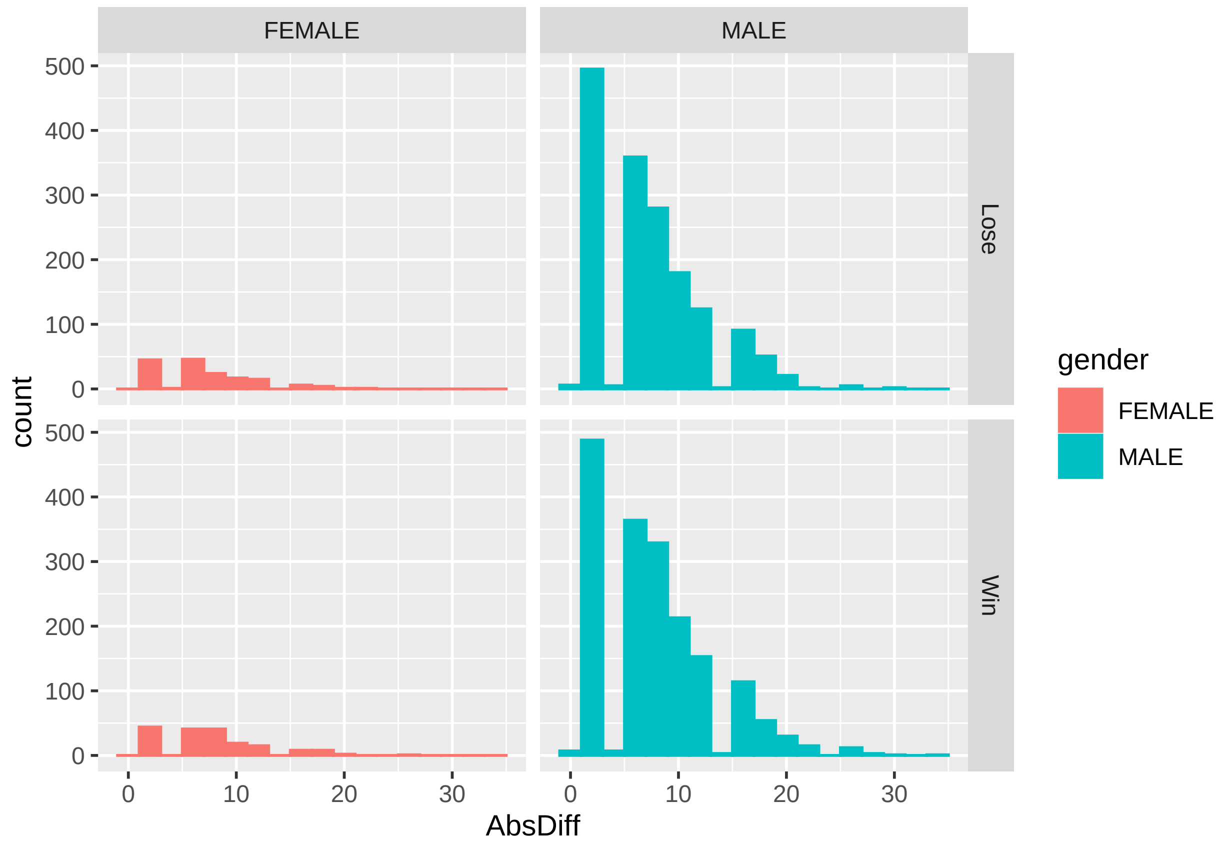

```
# Create the logistic regression
reach.model <- glm(AdWin ~ AbsDiff, data = reach, family = binomial())

# Display summary
summary(reach.model)
```

```
##
## Call:
## glm(formula = AdWin ~ AbsDiff, family = binomial(), data = reach)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -1.463  -1.197   1.034   1.158   1.209
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.075187   0.059314  -1.268 0.204937
## AbsDiff      0.023771   0.006688   3.554 0.000379 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5197.3  on 3755  degrees of freedom
## Residual deviance: 5184.6  on 3754  degrees of freedom
## AIC: 5188.6
##
## Number of Fisher Scoring iterations: 3
```

```
# Question: Does gender impact whether the fighter with the longer reach will win more frequently?

# Create a smaller data frame
reach_gender <- df[,c('Winner', 'gender', 'B_Reach_cms', 'R_Reach_cms')]

# Convert gender into levels
reach_gender$gender <- as.factor(reach_gender$gender)

# Calculate the difference in reach (positive = blue advantage)
reach_gender$Diff <- reach_gender$B_Reach_cms - reach_gender$R_Reach_cms

# There is an outlier that makes no sense, so remove it
reach_gender <- reach_gender[reach_gender$Diff > -50,]

# There is an outlier that makes no sense, so remove it
reach_gender <- reach_gender[reach$Diff > -50,]

# Remove all cases where the players had equal reach
reach_gender <- subset(reach_gender, !(Diff == 0))

# Identify the fighter with the longer reach
reach_gender$Advantage <-
  case_when(
    reach_gender$Diff > 0 ~ 'Blue',
    reach_gender$Diff < 0 ~ 'Red'
  )

# Identify if the advantaged fighter won
reach_gender$AdWin <-
  case_when(
    reach_gender$Advantage == reach_gender$Winner ~ 'Win',
    reach_gender$Advantage != reach_gender$Winner ~ 'Lose'
  )

reach_gender$AdWin <- as.factor(reach_gender$AdWin)

# Take the absolute value of the difference
```

```r
reach_gender$AbsDiff <- abs(reach_gender$Diff)

# Plot the data
reach_gender.hist <- ggplot(reach_gender, aes(x = AbsDiff, color = gender))
reach_gender.hist +
  geom_histogram(binwidth = 2, aes(fill = gender)) +
  facet_grid(AdWin ~ gender)
```




```r
# Create the logistic regression
diff.model <- glm(AdWin ~ AbsDiff, data = reach_gender, family = binomial())
reach_gender.model <- update(diff.model, .~. + gender)

# Display summary
summary(diff.model)
```

```
##
## Call:
## glm(formula = AdWin ~ AbsDiff, family = binomial(), data = reach_gender)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.463  -1.197   1.034   1.158   1.209
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept) -0.075187   0.059314  -1.268 0.204937
## AbsDiff      0.023771   0.006688   3.554 0.000379 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5197.3  on 3755  degrees of freedom
## Residual deviance: 5184.6  on 3754  degrees of freedom
## AIC: 5188.6
##
## Number of Fisher Scoring iterations: 3
```

```
summary(reach_gender.model)
```

```
##
## Call:
## glm(formula = AdWin ~ AbsDiff + gender, family = binomial(),
##     data = reach_gender)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.462  -1.196   1.034   1.151   1.210
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.059510   0.119113  -0.500 0.617352
## AbsDiff      0.023784   0.006689   3.556 0.000377 ***
## genderMALE  -0.017338   0.114236  -0.152 0.879368
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5197.3  on 3755  degrees of freedom
## Residual deviance: 5184.5  on 3753  degrees of freedom
## AIC: 5190.5
##
## Number of Fisher Scoring iterations: 3
```

```
anova(diff.model, reach_gender.model)
```

```
## Analysis of Deviance Table
##
## Model 1: AdWin ~ AbsDiff
## Model 2: AdWin ~ AbsDiff + gender
##   Resid. Df Resid. Dev Df Deviance
## 1      3754     5184.6
## 2      3753     5184.5  1  0.02304
```