Intro Data Exploration

Ty Hammond

2024-11-21

Data Exploration:

We have an initial idea of looking at the number of pass rushers and the number of defenders in the box and the effect that had on play outcome. We are utilizing the plays.csv file from kaggle which has been imported into GitHub. I will explore this topic below:

```
#import libraries
library(ggplot2)

# Read in csv File
plays <- read.csv("plays.csv")

# Give Variable Names
names(plays)
```

```
##  [1] "gameId"                "playId"                "playDescription"
##  [4] "quarter"               "down"                  "yardsToGo"
##  [7] "possessionTeam"        "playType"              "yardlineSide"
## [10] "yardlineNumber"        "offenseFormation"      "personnelO"
## [13] "defendersInTheBox"     "numberOfPassRushers"   "personnelD"
## [16] "typeDropback"          "preSnapVisitorScore"   "preSnapHomeScore"
## [19] "gameClock"             "absoluteYardlineNumber" "penaltyCodes"
## [22] "penaltyJerseyNumbers"  "passResult"            "offensePlayResult"
## [25] "playResult"            "epa"                   "isDefensivePI"
```

```
# See Examples
head(plays)
```

```
##       gameId playId
## 1 2018090600     75
## 2 2018090600    146
## 3 2018090600    168
## 4 2018090600    190
## 5 2018090600    256
## 6 2018090600    320
##                                                                                playDescription
## 1 (15:00) M.Ryan pass short right to J.Jones pushed ob at ATL 30 for 10 yards (M.Jenkins).
## 2               (13:10) M.Ryan pass incomplete short right to C.Ridley (J.Mills, J.Hicks).
## 3                        (13:05) (Shotgun) M.Ryan pass incomplete short left to D.Freeman.
## 4      (13:01) (Shotgun) M.Ryan pass deep left to J.Jones to PHI 6 for 33 yards (R.Darby).
## 5                       (10:59) (Shotgun) M.Ryan pass incomplete short right to D.Freeman.
## 6  (10:10) (Shotgun) N.Foles pass short left to N.Agholor to PHI 8 for 4 yards (R.Alford).
##   quarter down yardsToGo possessionTeam       playType yardlineSide
## 1       1    1        15            ATL play_type_pass          ATL
## 2       1    1        10            ATL play_type_pass          PHI
```

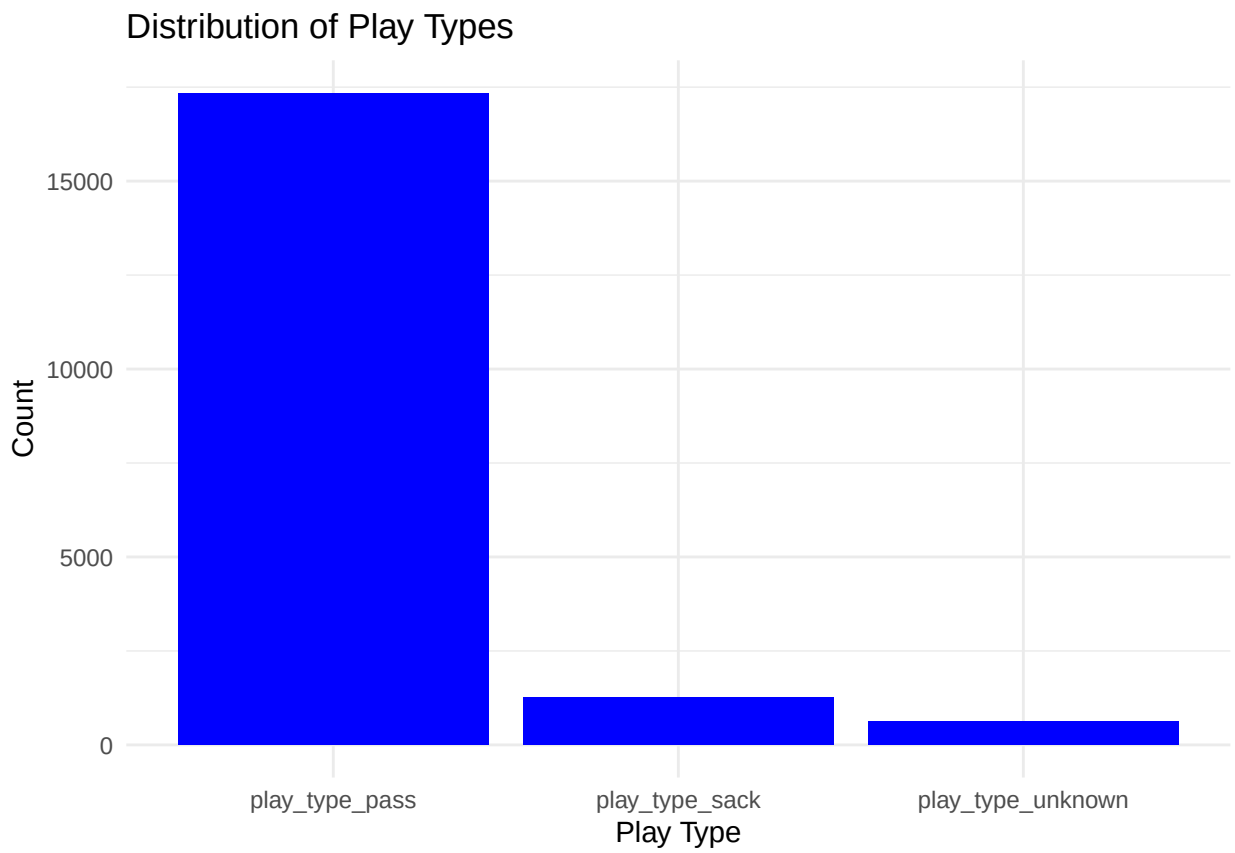
```
## 3       1    2        10            ATL play_type_pass          PHI
## 4       1    3        10            ATL play_type_pass          PHI
## 5       1    3         1            ATL play_type_pass          PHI
## 6       1    2         8            PHI play_type_pass          PHI
##   yardlineNumber offenseFormation       personnelO defendersInTheBox
## 1             20           I_FORM 2 RB, 1 TE, 2 WR                 7
## 2             39       SINGLEBACK 1 RB, 1 TE, 3 WR                 7
## 3             39          SHOTGUN 2 RB, 1 TE, 2 WR                 6
## 4             39          SHOTGUN 1 RB, 1 TE, 3 WR                 6
## 5              1          SHOTGUN 2 RB, 3 TE, 0 WR                 8
## 6              4          SHOTGUN 1 RB, 1 TE, 3 WR                 7
##   numberOfPassRushers       personnelD          typeDropback
## 1                   4 4 DL, 2 LB, 5 DB           TRADITIONAL
## 2                   4 4 DL, 2 LB, 5 DB           TRADITIONAL
## 3                   4 4 DL, 2 LB, 5 DB           TRADITIONAL
## 4                   5 4 DL, 1 LB, 6 DB SCRAMBLE_ROLLOUT_LEFT
## 5                   6 6 DL, 3 LB, 2 DB           TRADITIONAL
## 6                   4 3 DL, 3 LB, 5 DB           TRADITIONAL
##   preSnapVisitorScore preSnapHomeScore gameClock absoluteYardlineNumber
## 1                   0                0  15:00:00                     90
## 2                   0                0  13:10:00                     49
## 3                   0                0  13:05:00                     49
## 4                   0                0  13:01:00                     49
## 5                   0                0  10:59:00                     11
## 6                   0                0  10:10:00                     14
##   penaltyCodes penaltyJerseyNumbers passResult offensePlayResult playResult
## 1                                           C               10         10
## 2                                           I                0          0
## 3                                           I                0          0
## 4                                           C               33         33
## 5                                           I                0          0
## 6                                           C                4          4
##          epa isDefensivePI
## 1  0.2618273         FALSE
## 2 -0.3723598         FALSE
## 3 -0.7027787         FALSE
## 4  3.0475300         FALSE
## 5 -0.8422719         FALSE
## 6 -0.3440965         FALSE
```

```r
# Examine Potentially Important Variables

# Individual Variables below

#playType
table(plays$playType) # freq table
```

```
##
##    play_type_pass    play_type_sack play_type_unknown
##             17346              1260               633
```

```r
# plot
ggplot(plays, aes(x = playType)) +
  geom_bar(fill = "blue") +
  xlab("Play Type") +
```

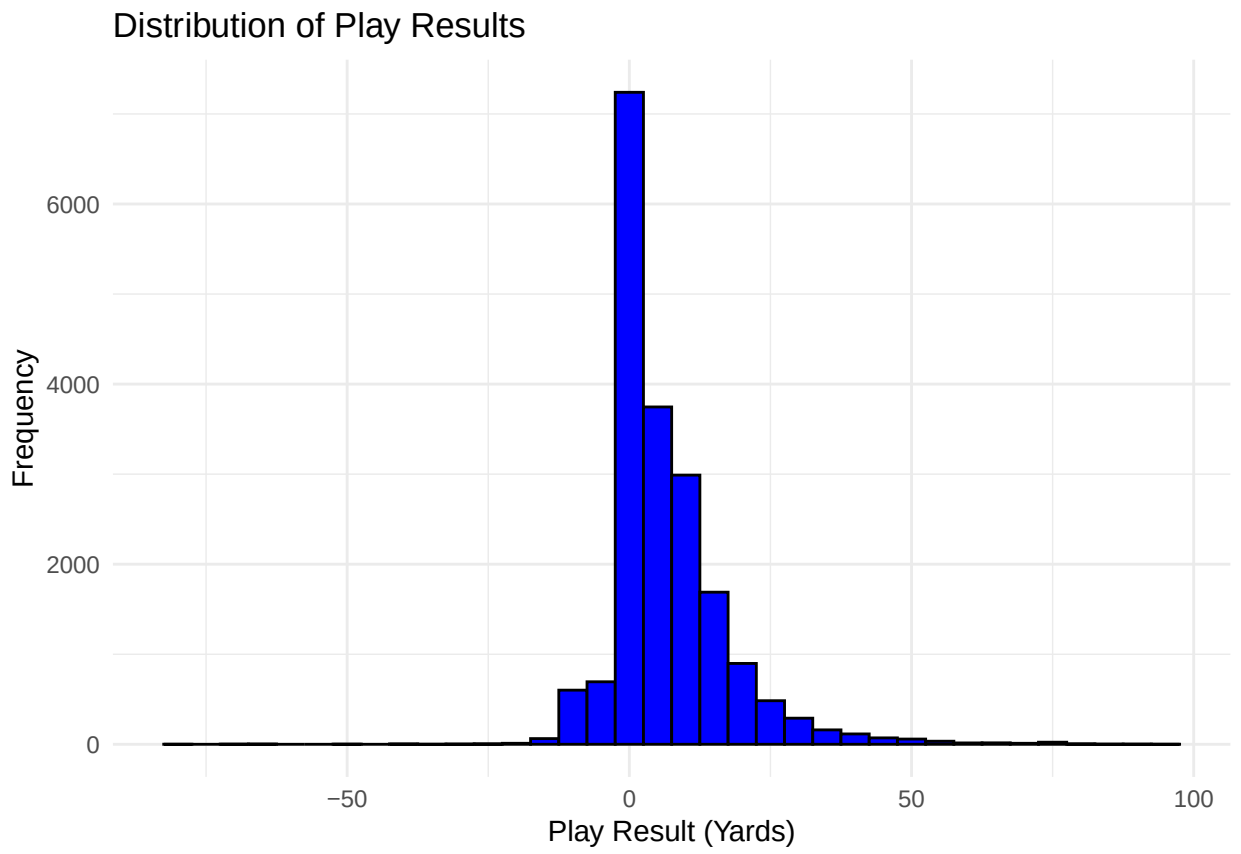
```
  ylab("Count") +
  ggtitle("Distribution of Play Types") +
  theme_minimal()
```




```
#playResult
summary(plays$playResult) # sum stats
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -78.000   0.000   4.000   6.469  11.000  97.000
```

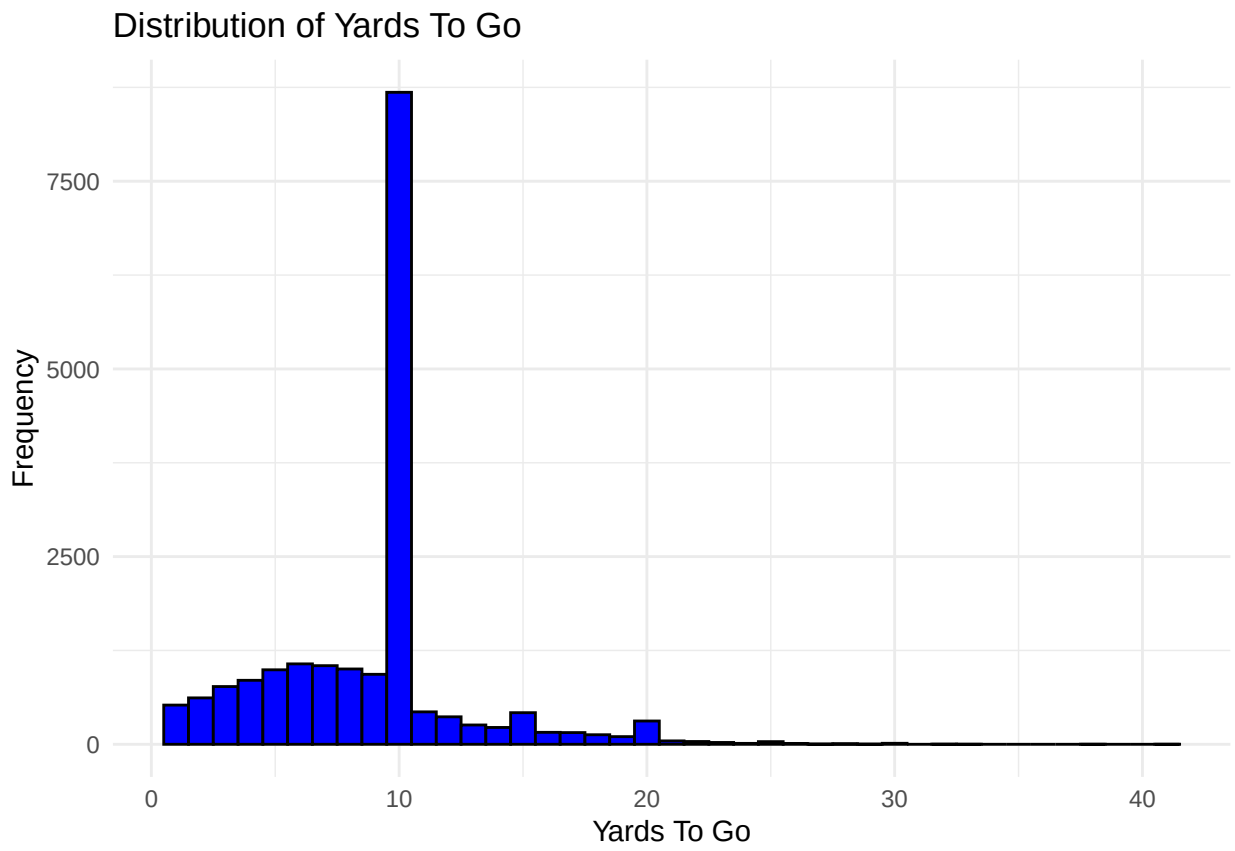
```
# hist
ggplot(plays, aes(x = playResult)) +
  geom_histogram(binwidth = 5, fill = "blue", color = "black") +
  xlab("Play Result (Yards)") +
  ylab("Frequency") +
  ggtitle("Distribution of Play Results") +
  theme_minimal()
```




```
#yardsToGo
summary(plays$yardsToGo) # sum stats
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   6.000  10.000   8.924  10.000  41.000
```

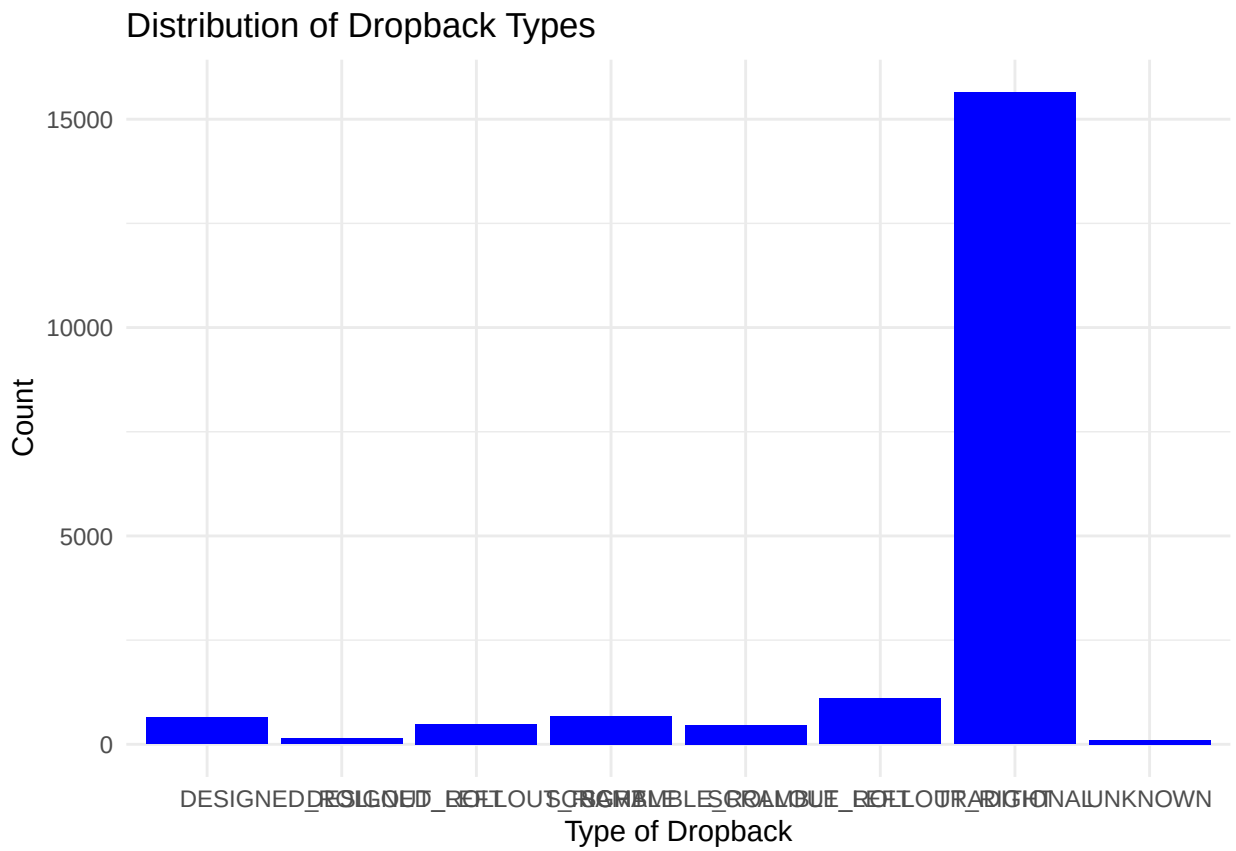
```
# hist
ggplot(plays, aes(x = yardsToGo)) +
  geom_histogram(binwidth = 1, fill = "blue", color = "black") +
  xlab("Yards To Go") +
  ylab("Frequency") +
  ggtitle("Distribution of Yards To Go") +
  theme_minimal()
```




```
#typeDropback
table(plays$typeDropback) # freq table
```

```
##
##                         DESIGNED_ROLLOUT_LEFT DESIGNED_ROLLOUT_RIGHT
##                   639                   141                   482
##               SCRAMBLE  SCRAMBLE_ROLLOUT_LEFT SCRAMBLE_ROLLOUT_RIGHT
##                   677                   462                  1096
##            TRADITIONAL               UNKNOWN
##                 15645                    97
```

```
# bar plot
ggplot(plays, aes(x = typeDropback)) +
  geom_bar(fill = "blue") +
  xlab("Type of Dropback") +
  ylab("Count") +
  ggtitle("Distribution of Dropback Types") +
  theme_minimal()
```

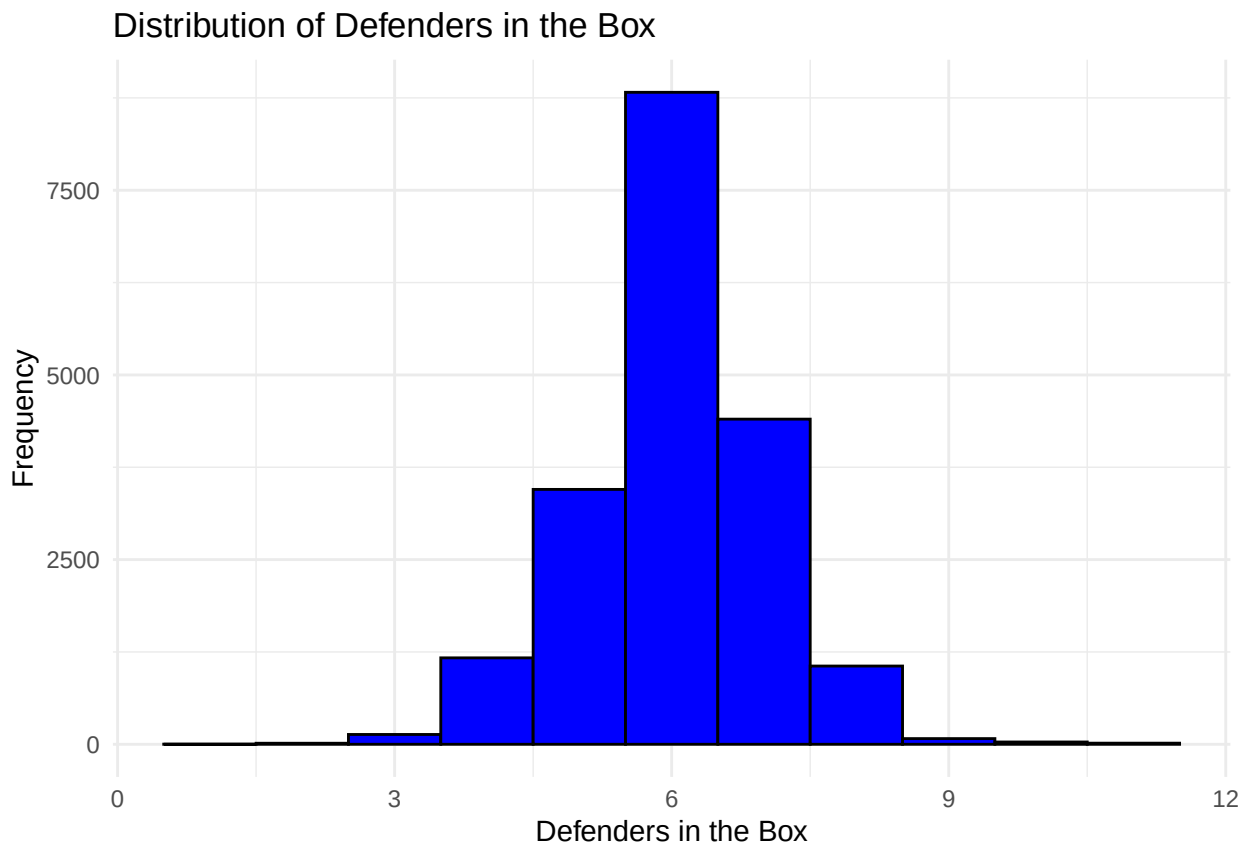



```
#defendersInTheBox
summary(plays$defendersInTheBox) # sum stats
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   6.000   6.000   6.036   7.000  11.000      62
```

```
# hist
ggplot(plays, aes(x = defendersInTheBox)) +
  geom_histogram(binwidth = 1, fill = "blue", color = "black") +
  xlab("Defenders in the Box") +
  ylab("Frequency") +
  ggtitle("Distribution of Defenders in the Box") +
  theme_minimal()
```

```
## Warning: Removed 62 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

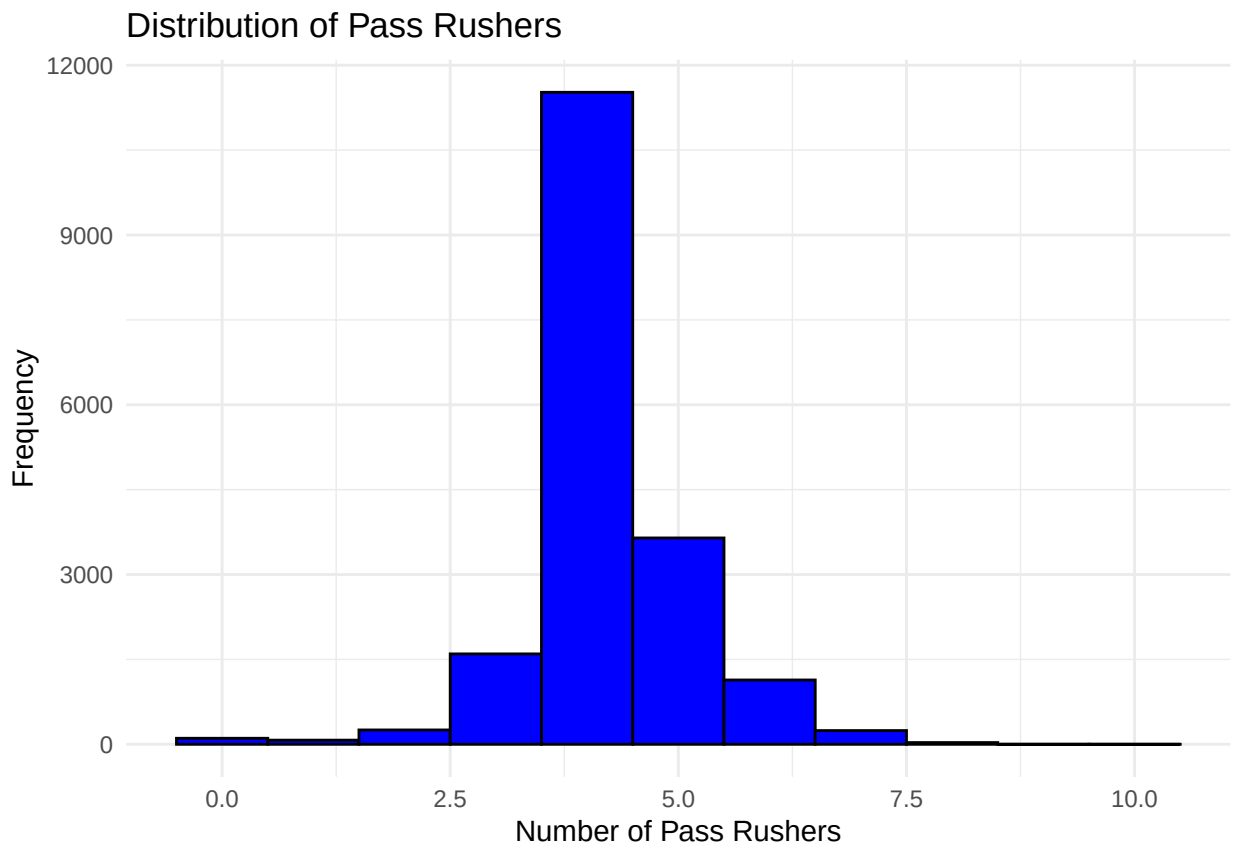



```
#numberOfPassRushers
summary(plays$numberOfPassRushers) # sum stats
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   4.000   4.000   4.216   5.000  10.000     633
```

```
# hist
ggplot(plays, aes(x = numberOfPassRushers)) +
  geom_histogram(binwidth = 1, fill = "blue", color = "black") +
  xlab("Number of Pass Rushers") +
  ylab("Frequency") +
  ggtitle("Distribution of Pass Rushers") +
  theme_minimal()
```

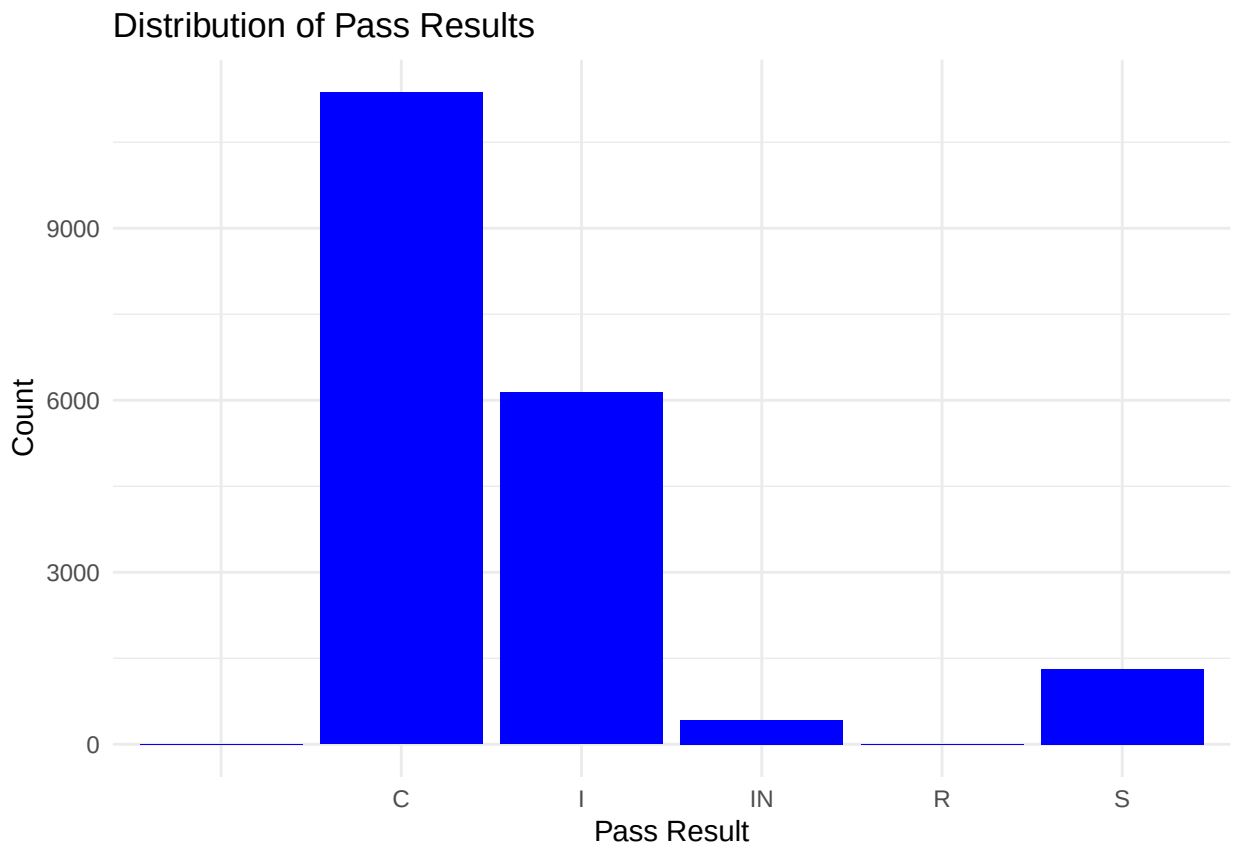
```
## Warning: Removed 633 rows containing non-finite outside the scale range
## (`stat_bin()`).
```




```
#passResult
table(plays$passResult) # freq table
```

```
##
##           C     I    IN     R     S
##     2 11370  6135   420     4  1308
```

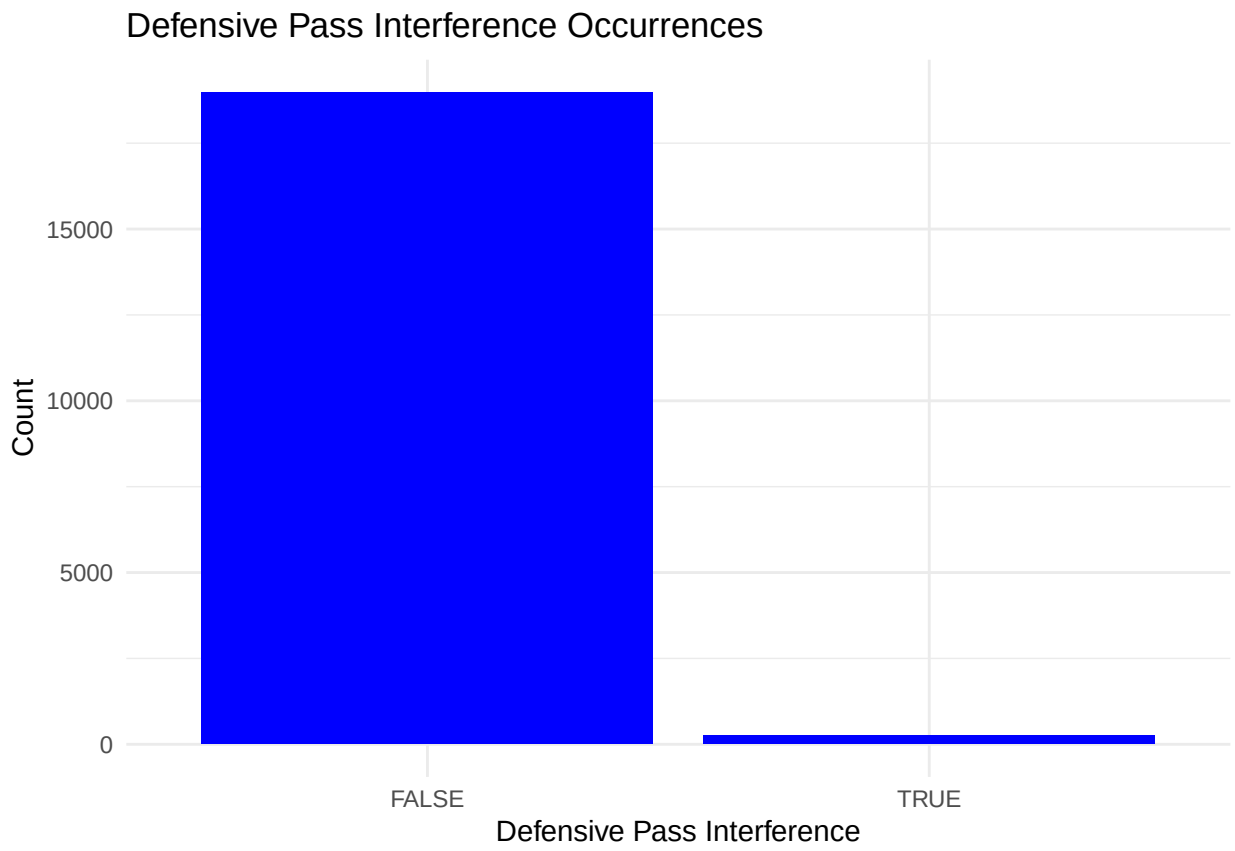
```
# bar plot
ggplot(plays, aes(x = passResult)) +
  geom_bar(fill = "blue") +
  xlab("Pass Result") +
  ylab("Count") +
  ggtitle("Distribution of Pass Results") +
  theme_minimal()
```




```
#isDefensivePI
table(plays$isDefensivePI) # freq table
```

```
##
## FALSE  TRUE
## 18980   259
```

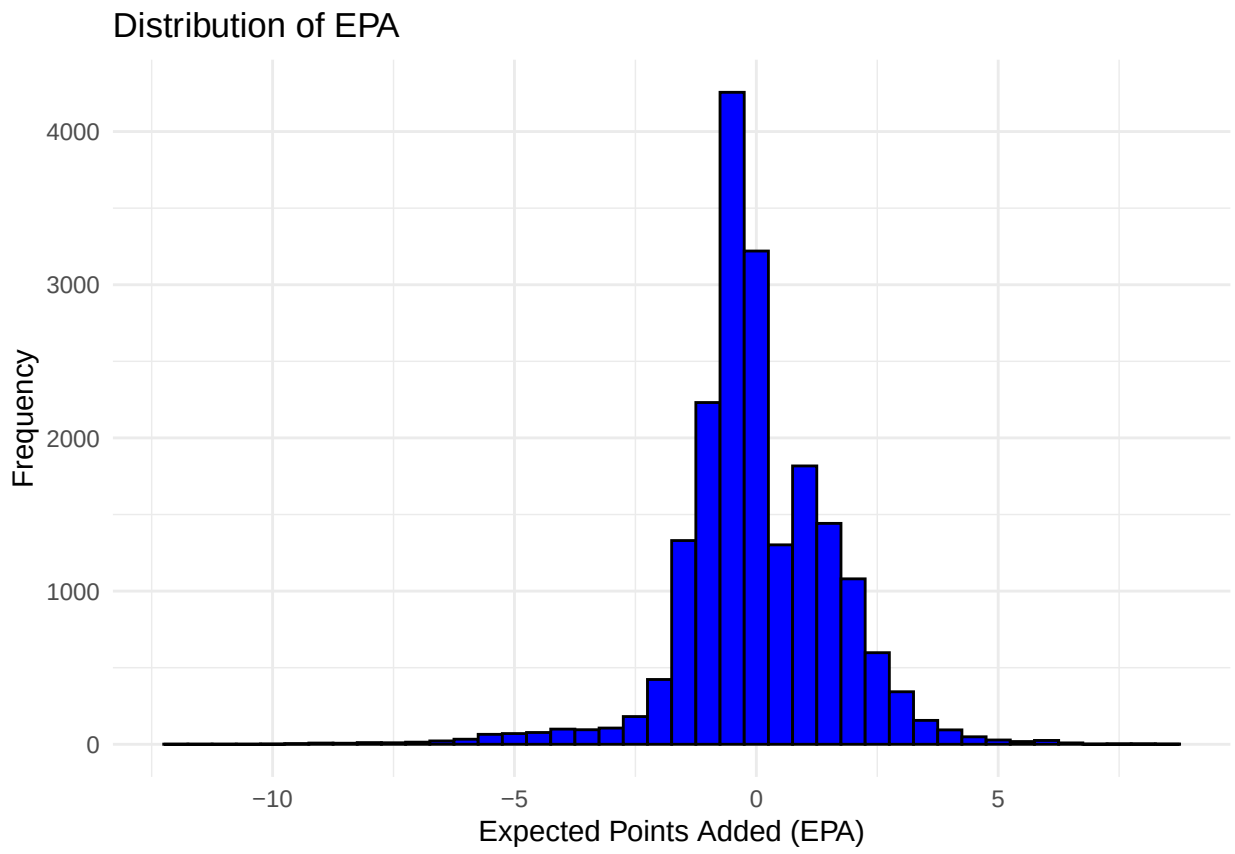
```
# bar plot
ggplot(plays, aes(x = factor(isDefensivePI))) +
  geom_bar(fill = "blue") +
  xlab("Defensive Pass Interference") +
  ylab("Count") +
  ggtitle("Defensive Pass Interference Occurrences") +
  theme_minimal()
```




```
#epa
summary(plays$epa) # sum stats
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -11.93595  -0.74533  -0.19054   0.02073   0.97682   8.62932
```

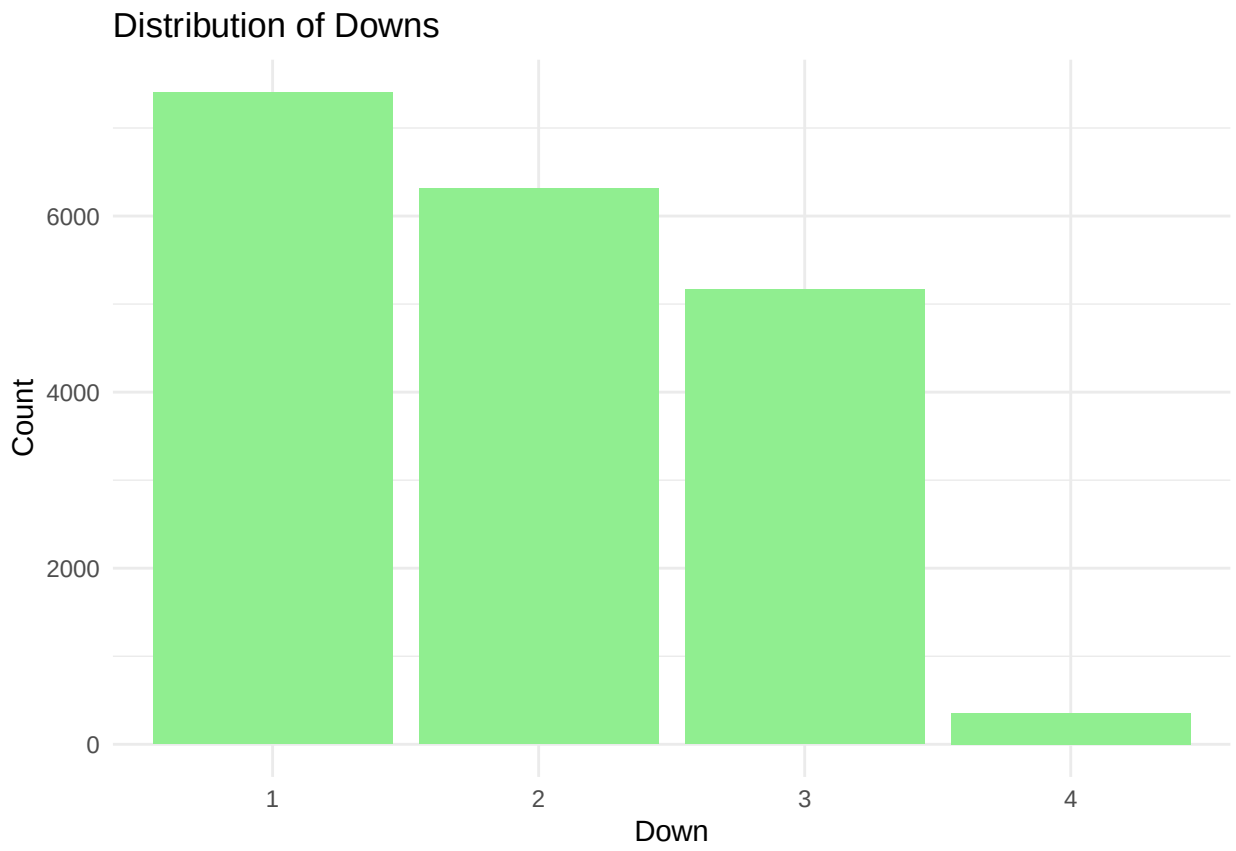
```
# hist
ggplot(plays, aes(x = epa)) +
  geom_histogram(binwidth = 0.5, fill = "blue", color = "black") +
  xlab("Expected Points Added (EPA)") +
  ylab("Frequency") +
  ggtitle("Distribution of EPA") +
  theme_minimal()
```




```
#down
table(plays$down) # freq table
```

```
##
##    1    2    3    4
## 7405 6315 5166  353
```

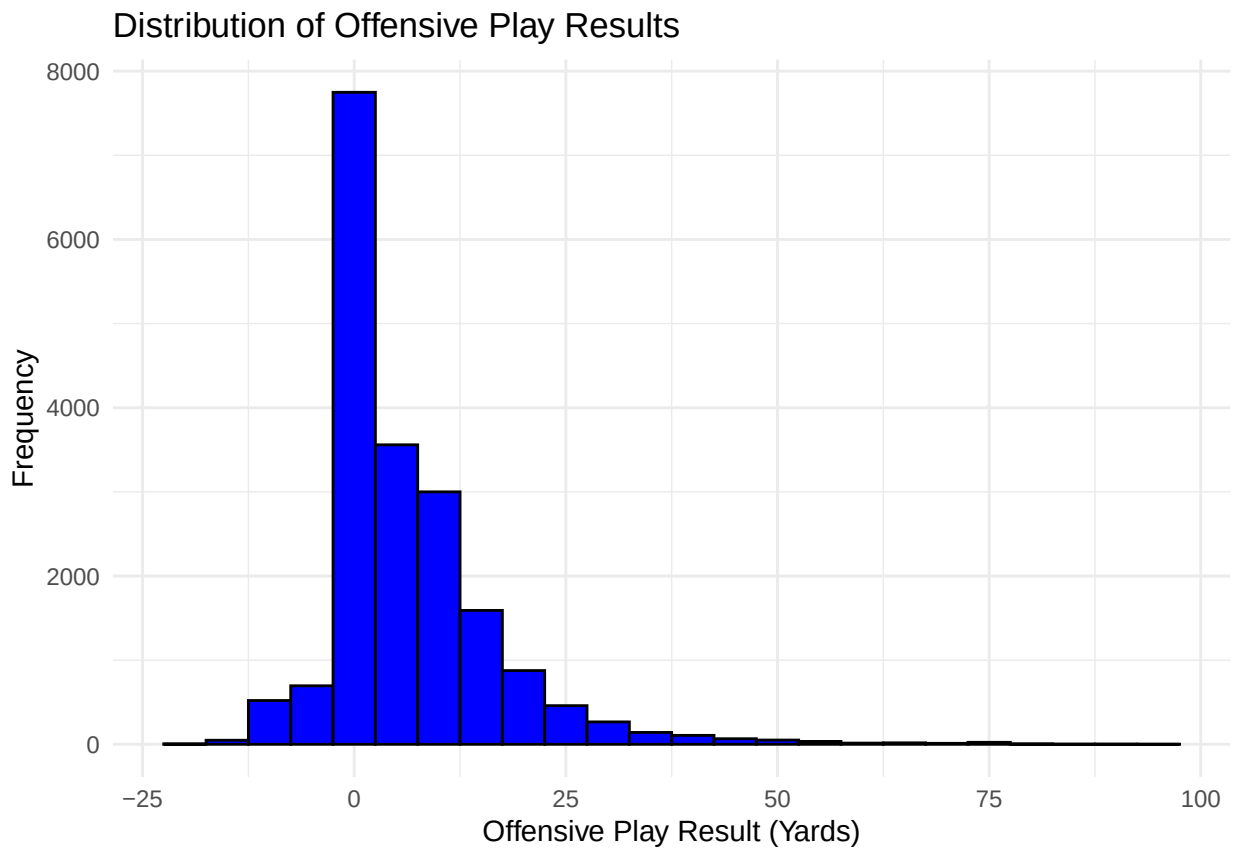
```
# bar plot
ggplot(plays, aes(x = factor(down))) +
  geom_bar(fill = "lightgreen") +
  xlab("Down") +
  ylab("Count") +
  ggtitle("Distribution of Downs") +
  theme_minimal()
```




```
#offensePlayResult
summary(plays$offensePlayResult) # sum stats
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -19.000   0.000   4.000   6.279  10.000  97.000
```

```
# hist
ggplot(plays, aes(x = offensePlayResult)) +
  geom_histogram(binwidth = 5, fill = "blue", color = "black") +
  xlab("Offensive Play Result (Yards)") +
  ylab("Frequency") +
  ggtitle("Distribution of Offensive Play Results") +
  theme_minimal()
```




```
# Overall dataset


# Combinations
```