260 Project EDA

Group Name: K-Nearest Tailgaters

2021-10-31

```
injuries = read.csv("injuries.csv")
nfl_roster = read.csv("nfl_roster.csv")
```

Check for duplicates, NA's in outcome

```
any(duplicated(injuries[["Full_Name"]])) # no duplicate names (multiple injuries during the year)
```

```
## [1] FALSE
```

```
any(is.na(injuries[["Injury"]])) # no NA's in outcome
```

```
## [1] FALSE
```

Injury.Status: What to do with non-football injuries (NFI-R), or COVID? How can we sort based on injury severity?

```
table(injuries$Injury.Status)
```

```
##
##            Did Not Practice on Thursday. Doubtful for Week 8 at Chicago
##                                                                       1
##             Did Not Practice on Thursday. Doubtful for Week 8 vs. Miami
##                                                                       1
##      Did Not Practice on Thursday. Doubtful for Week 8 vs. Philadelphia
##                                                                       1
##         Did Not Practice on Thursday. Doubtful for Week 8 vs. Tampa Bay
##                                                                       1
##        Did Not Practice on Thursday. Doubtful for Week 8 vs. Washington
##                                                                       1
##        Did Not Practice on Thursday. Questionable for Week 8 at Atlanta
##                                                                       1
##        Did Not Practice on Thursday. Questionable for Week 8 at Chicago
##                                                                       5
##         Did Not Practice on Thursday. Questionable for Week 8 at Denver
##                                                                       2
##        Did Not Practice on Thursday. Questionable for Week 8 at Detroit
##                                                                       1
##        Did Not Practice on Thursday. Questionable for Week 8 at Houston
##                                                                       4
##    Did Not Practice on Thursday. Questionable for Week 8 at Kansas City
```

```
##                                                                    4
##      Did Not Practice on Thursday. Questionable for Week 8 at N.Y. Jets
##                                                                    1
##    Did Not Practice on Thursday. Questionable for Week 8 at New Orleans
##                                                                    2
##        Did Not Practice on Thursday. Questionable for Week 8 at Seattle
##                                                                    1
##    Did Not Practice on Thursday. Questionable for Week 8 vs. Cincinnati
##                                                                    1
##        Did Not Practice on Thursday. Questionable for Week 8 vs. Dallas
##                                                                    1
##   Did Not Practice on Thursday. Questionable for Week 8 vs. N.Y. Giants
##                                                                    1
##    Did Not Practice on Thursday. Questionable for Week 8 vs. Pittsburgh
##                                                                    2
##     Did Not Practice on Thursday. Questionable for Week 8 vs. Tampa Bay
##                                                                    3
##     Did Not Practice on Thursday. Questionable for Week 8 vs. Tennessee
##                                                                    1
##    Did Not Practice on Thursday. Questionable for Week 8 vs. Washington
##                                                                    1
##                                                  IR. Injured Reserve
##                                                                  120
##                           IR. Injured Reserve. Expected Return - Week 10
##                                                                   16
##                           IR. Injured Reserve. Expected Return - Week 11
##                                                                   16
##                           IR. Injured Reserve. Expected Return - Week 12
##                                                                    1
##                           IR. Injured Reserve. Expected Return - Week 14
##                                                                    1
##                           IR. Injured Reserve. Expected Return - Week 15
##                                                                    1
##                           IR. Injured Reserve. Expected Return - Week 16
##                                                                    1
##                           IR. Injured Reserve. Expected Return - Week 17
##                                                                    1
##                            IR. Injured Reserve. Expected Return - Week 8
##                                                                   68
##                            IR. Injured Reserve. Expected Return - Week 9
##                                                                   38
##                                                 IR. Reserve - COVID-19
##                                                                    5
##                                       IR. Reserve - Non Football Injury
##                                                                    1
##        Limited Practice on Thursday. Questionable for Week 8 at Atlanta
##                                                                    1
##        Limited Practice on Thursday. Questionable for Week 8 at Buffalo
##                                                                    4
##        Limited Practice on Thursday. Questionable for Week 8 at Chicago
##                                                                    1
##         Limited Practice on Thursday. Questionable for Week 8 at Denver
##                                                                    3
##        Limited Practice on Thursday. Questionable for Week 8 at Detroit
```

```
##                                                                    1
##        Limited Practice on Thursday. Questionable for Week 8 at Houston
##                                                                    1
##   Limited Practice on Thursday. Questionable for Week 8 at Indianapolis
##                                                                    2
##    Limited Practice on Thursday. Questionable for Week 8 at Kansas City
##                                                                    2
##  Limited Practice on Thursday. Questionable for Week 8 at L.A. Chargers
##                                                                    5
##      Limited Practice on Thursday. Questionable for Week 8 at Minnesota
##                                                                    2
##    Limited Practice on Thursday. Questionable for Week 8 at New Orleans
##                                                                    2
##        Limited Practice on Thursday. Questionable for Week 8 at Seattle
##                                                                    1
##    Limited Practice on Thursday. Questionable for Week 8 vs. Cincinnati
##                                                                    3
##  Limited Practice on Thursday. Questionable for Week 8 vs. Jacksonville
##                                                                    1
##     Limited Practice on Thursday. Questionable for Week 8 vs. L.A. Rams
##                                                                    2
##   Limited Practice on Thursday. Questionable for Week 8 vs. N.Y. Giants
##                                                                    1
##   Limited Practice on Thursday. Questionable for Week 8 vs. New England
##                                                                    1
##  Limited Practice on Thursday. Questionable for Week 8 vs. Philadelphia
##                                                                    1
##    Limited Practice on Thursday. Questionable for Week 8 vs. Pittsburgh
##                                                                    5
## Limited Practice on Thursday. Questionable for Week 8 vs. San Francisco
##                                                                    2
##     Limited Practice on Thursday. Questionable for Week 8 vs. Tampa Bay
##                                                                    1
##     Limited Practice on Thursday. Questionable for Week 8 vs. Tennessee
##                                                                    1
##    Limited Practice on Thursday. Questionable for Week 8 vs. Washington
##                                                                    1
##       Limited Practice on Wednesday. Questionable for Week 8 vs. Dallas
##                                                                    1
##                                        NFI-R for Week 8 at L.A. Chargers
##                                                                    2
##                                           NFI-R for Week 8 at Minnesota
##                                                                    1
##                                           NFI-R for Week 8 at N.Y. Jets
##                                                                    2
##                                             NFI-R for Week 8 at Seattle
##                                                                    1
##                                         NFI-R for Week 8 vs. Cincinnati
##                                                                    1
##                                              NFI-R for Week 8 vs. Dallas
##                                                                    1
##                                        NFI-R for Week 8 vs. N.Y. Giants
##                                                                    1
##                                       NFI-R for Week 8 vs. Philadelphia
```

```
##                                                                  2
##                                      NFI-R for Week 8 vs. Tennessee
##                                                                  1
##                                     NFI-R for Week 8 vs. Washington
##                                                                  1
##                                     NFI-R for Week 9 at N.Y. Giants
##                                                                  1
##                                      NFI-R for Week 9 vs. Minnesota
##                                                                  1
##                  Out for Week 8 vs. Cincinnati. Expected Return - Week 10
##                                                                  1
##                Out for Week 8 vs. Jacksonville. Expected Return - Week 11
##                                                                  1
##                   Physically Unable to Perform. Expected Return - Week 11
##                                                                  1
##                   Physically Unable to Perform. Expected Return - Week 12
##                                                                  1
##                    Physically Unable to Perform. Expected Return - Week 8
##                                                                 15
##                                    Questionable for Week 8 at Atlanta
##                                                                  2
##                                  Questionable for Week 8 at Cleveland
##                                                                  2
##                                     Questionable for Week 8 at Denver
##                                                                  2
##                                    Questionable for Week 8 at Detroit
##                                                                  2
##                                    Questionable for Week 8 at Houston
##                                                                  1
##                                 Questionable for Week 8 at Indianapolis
##                                                                  4
##                                  Questionable for Week 8 at Kansas City
##                                                                  2
##                                  Questionable for Week 8 at New Orleans
##                                                                  2
##                                    Questionable for Week 8 vs. Carolina
##                                                                  3
##                                 Questionable for Week 8 vs. N.Y. Giants
##                                                                  1
##                                Questionable for Week 8 vs. Philadelphia
##                                                                  2
##                               Questionable for Week 8 vs. San Francisco
##                                                                  1
##                                   Questionable for Week 8 vs. Tampa Bay
##                                                                  2
##                                   Questionable for Week 8 vs. Tennessee
##                                                                  1
##                                  Questionable for Week 9 at Kansas City
##                                                                  2
##                                  Questionable for Week 9 at N.Y. Giants
##                                                                  4
##                                Questionable for Week 9 at San Francisco
##                                                                  4
##                                   Questionable for Week 9 vs. Minnesota
```

```
##                                                                 4
```

One idea: IR > Doubtful > Questionable, then remove NFI and COVID

```
var_change = function(x) {
  ordinal_injury = c()
  for(i in seq_along(x)) {
    if (str_detect(x[i], "NFI") | str_detect(x[i], "Non Football Injury")) {
      ordinal_injury[i] = "NFI"
    }
    else if (str_detect(x[i], "Questionable")) {
      ordinal_injury[i] = "Questionable"
    }
    else if (str_detect(x[i], "Doubtful")) {
      ordinal_injury[i] = "Doubtful"
    }
    else if (str_detect(x[i], "COVID-19")) {
      ordinal_injury[i] = "COVID-19"
    }
    else if (str_detect(x[i], "IR") | str_detect(x[i], "Physically Unable to Perform")) {
      ordinal_injury[i] = "IR"
    }
    else {
      ordinal_injury[i] = "other"
    }
  }
  ordinal_injury
}

injuries %<>% mutate(ordinal_injury = var_change(injuries$Injury.Status))
table(injuries$ordinal_injury)
```

```
##
##     COVID-19     Doubtful           IR          NFI        other Questionable
##            5            5          280           16            2          117
```

What are the "other" injuries?

```
injuries %>% filter(ordinal_injury == "other") %>% select(Injury, Injury.Status) # Not the kneecap!!
```

```
##       Injury                                                Injury.Status
## 1 Knee - PCL   Out for Week 8 vs. Cincinnati. Expected Return - Week 10
## 2    Kneecap Out for Week 8 vs. Jacksonville. Expected Return - Week 11
```

Maybe "Out for Week X" == "IR"?

```
injuries$ordinal_injury[injuries$ordinal_injury == "other"] = "IR"
```

Well, it seems that players are really only "Questionable" or on "IR".

We can remove "COVID-19" and "NFI" injuries, and merge "Doubtful" with "Questionable":

```
injuries %<>% filter(!ordinal_injury %in% c("COVID-19", "NFI")) %>%
              mutate(binary_injury = ifelse(ordinal_injury == "IR", 1, 0))
```

Merge data.frames by `Full_Name`

```
# Some people who are injured are no longer on the roster => out for season
injured_still_on_team = injuries[, -1] %>% inner_join(nfl_roster %>%
                                    select(-c(X, Team, College, Drafted, Height, Number)),
                                    by = "Full_Name")
```

NA's per variable

```
apply(injured_still_on_team, 2 , function(x) sum(is.na(x))) # distribution of NA's
```

```
##            Team        Position          Injury   Injury.Status            Date
##               0               0               0               0               0
##    Short_Name.x       Full_Name  ordinal_injury   binary_injury             Pos
##               0               0               0               0               0
##          Rating         Ranking          Weight             Age        Birthday
##               3               0               0               0               0
##            Exp.     Draft.Round      Draft.Pick   height_inches ranking_numeric
##               0             100             100               0               3
## Birthday_string    Short_Name.y
##               0               0
```

Collinearity for Height/Weight

```
cor.test(injured_still_on_team$height_inches,
         injured_still_on_team$Weight) # expected, so will combine to BMI
```

```
##
##  Pearson's product-moment correlation
##
## data:  injured_still_on_team$height_inches and injured_still_on_team$Weight
## t = 18, df = 342, p-value <2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.634 0.745
## sample estimates:
##   cor
## 0.694
```

Create factors and new variables of interest

```
injuries %<>% mutate(Injury = as.factor(Injury),
                     Position = as.factor(Position),
                     Team = as.factor(Team))


injured_still_on_team %<>% mutate(Injury = as.factor(Injury),
                     Position = as.factor(Position),
                     Team = as.factor(Team))


Offensive_Player = c("QB", "RB", "FB", "TB", "HB", "OL", "G", "LG", "RG",
                     "T", "LT", "RT", "C", "WR", "TE")

Defensive_Player = c("DL", "DE", "LE", "RE", "DT", "NT", "LB", "MLB", "ILB",
                     "OLB", "LOLB", "ROLB", "DB", "CB", "S", "SS", "FS")

Special_Teams = c("P", "K", "PR")

injured_still_on_team %<>%
  mutate(severe_injury = binary_injury, # bad injury yes/no
           Offense = ifelse(Position %in% c(Offensive_Player, Special_Teams), 1, 0), # offense yes/no
           BMI = (Weight / height_inches^2) * 703) # BMI
```

Descriptive stats

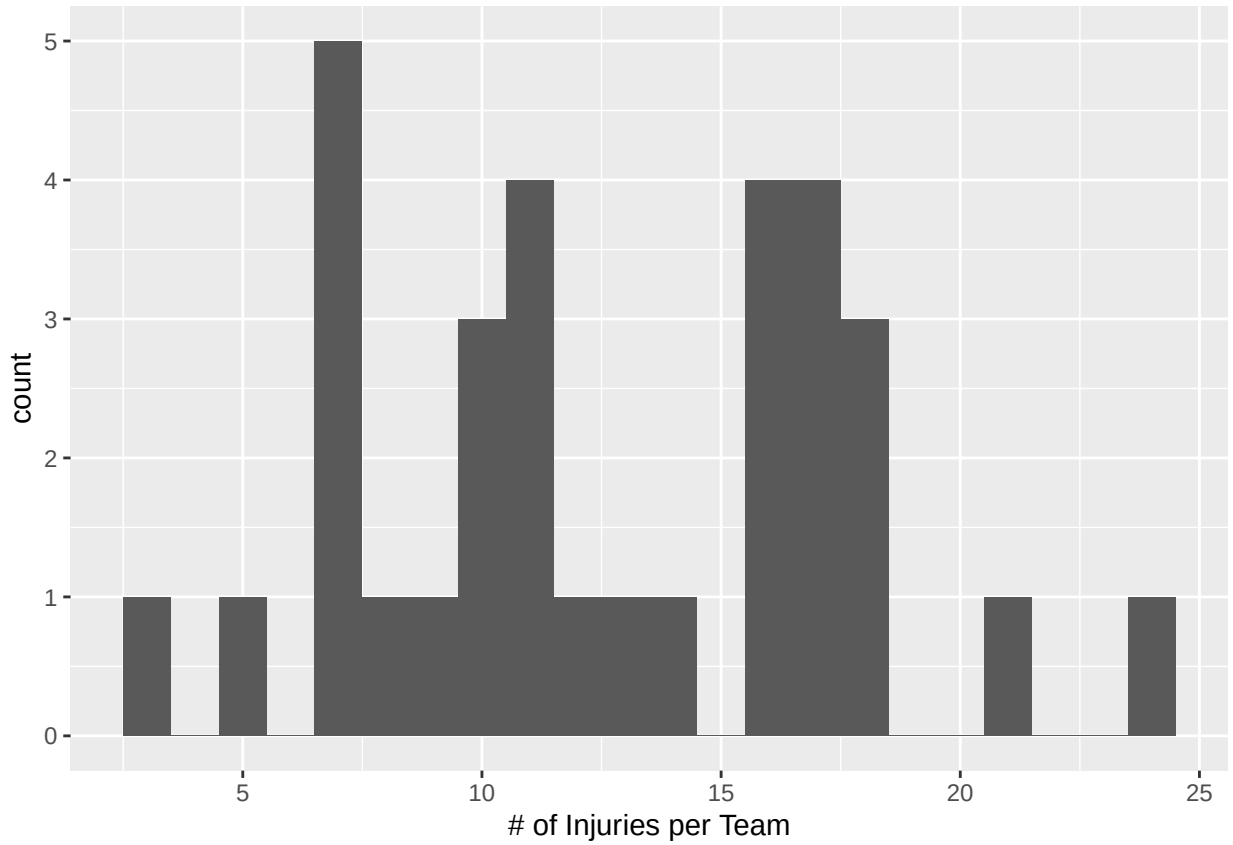
```
number_injury = injuries %>% group_by(Team) %>%
  summarize(num_injury = length(Injury))

pander(summary(number_injury$num_injury)) # ~13 injuries per team
```

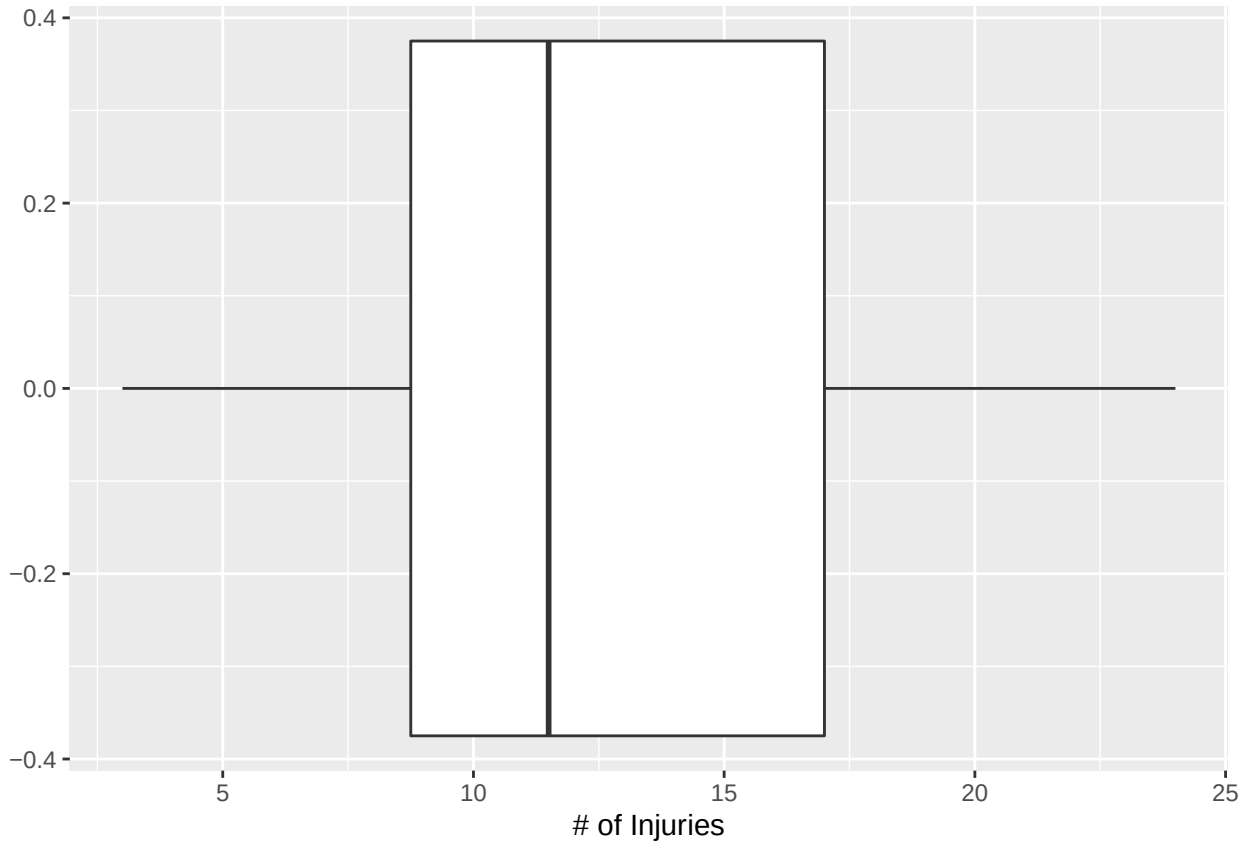
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
3	8.75	11.5	12.62	17	24

Histogram + Boxplot of Injuries per Team

```
number_injury %>%  # histogram (looks kinda normal)
  ggplot(aes(num_injury)) +
    geom_histogram(binwidth = 1) +
    xlab("# of Injuries per Team")
```

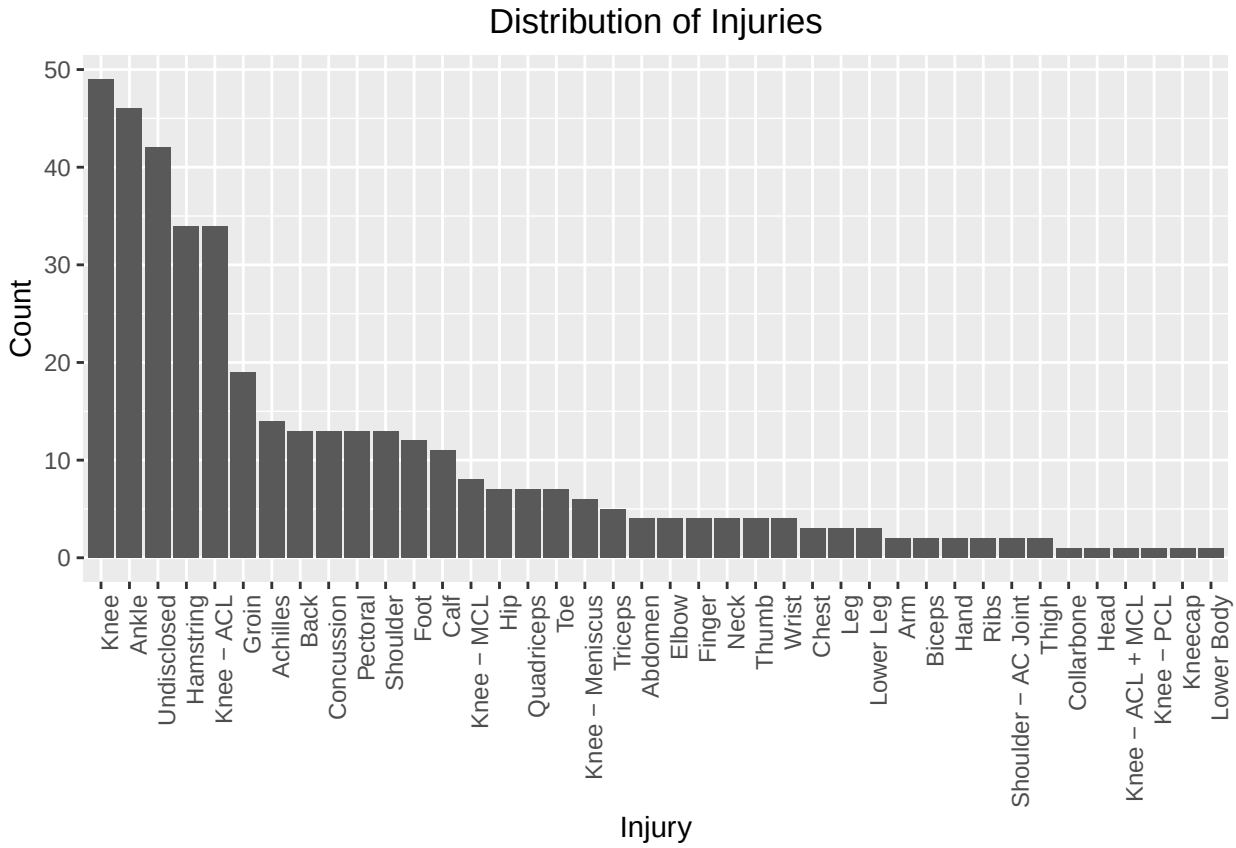



```
number_injury %>% # boxplot (looks kinda symmetric)
  ggplot(aes(num_injury)) +
    geom_boxplot() +
    xlab("# of Injuries")
```

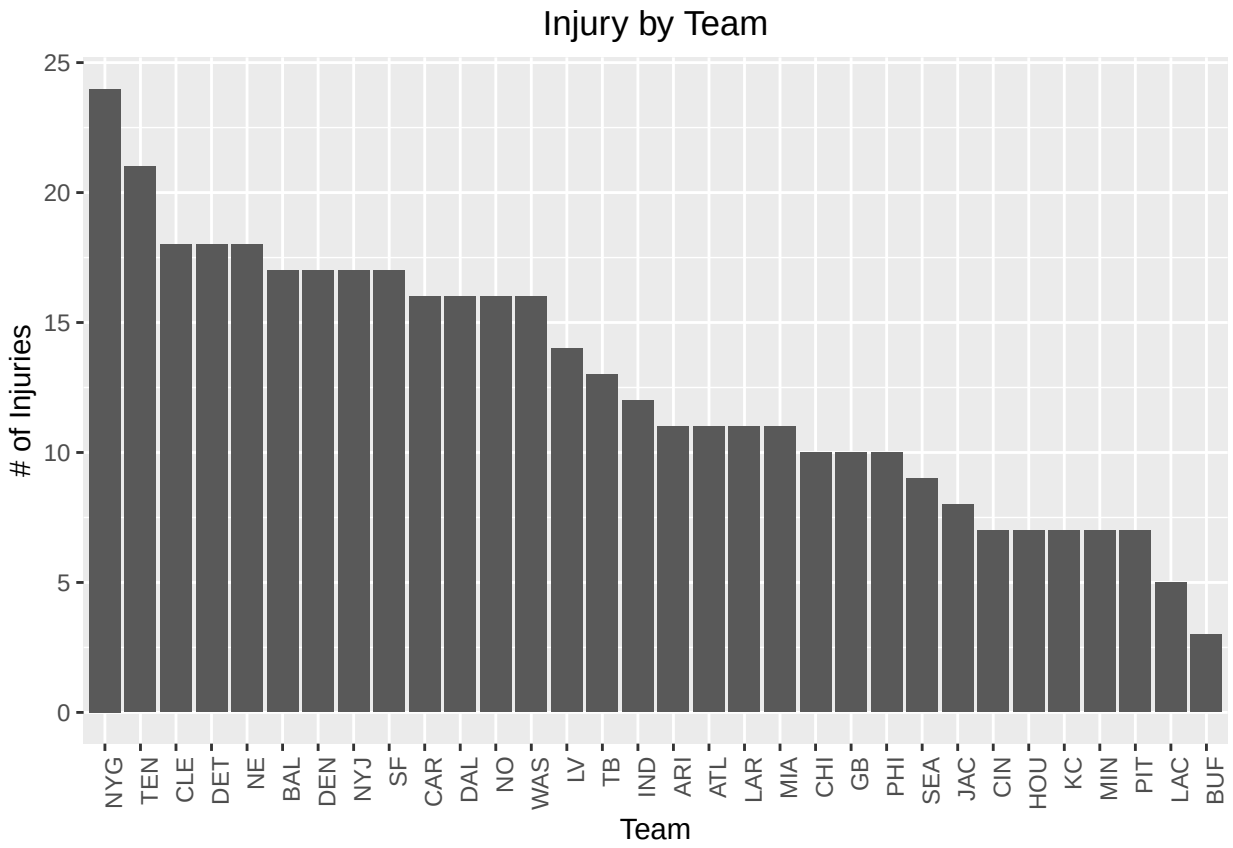


Injuries by Team and Position

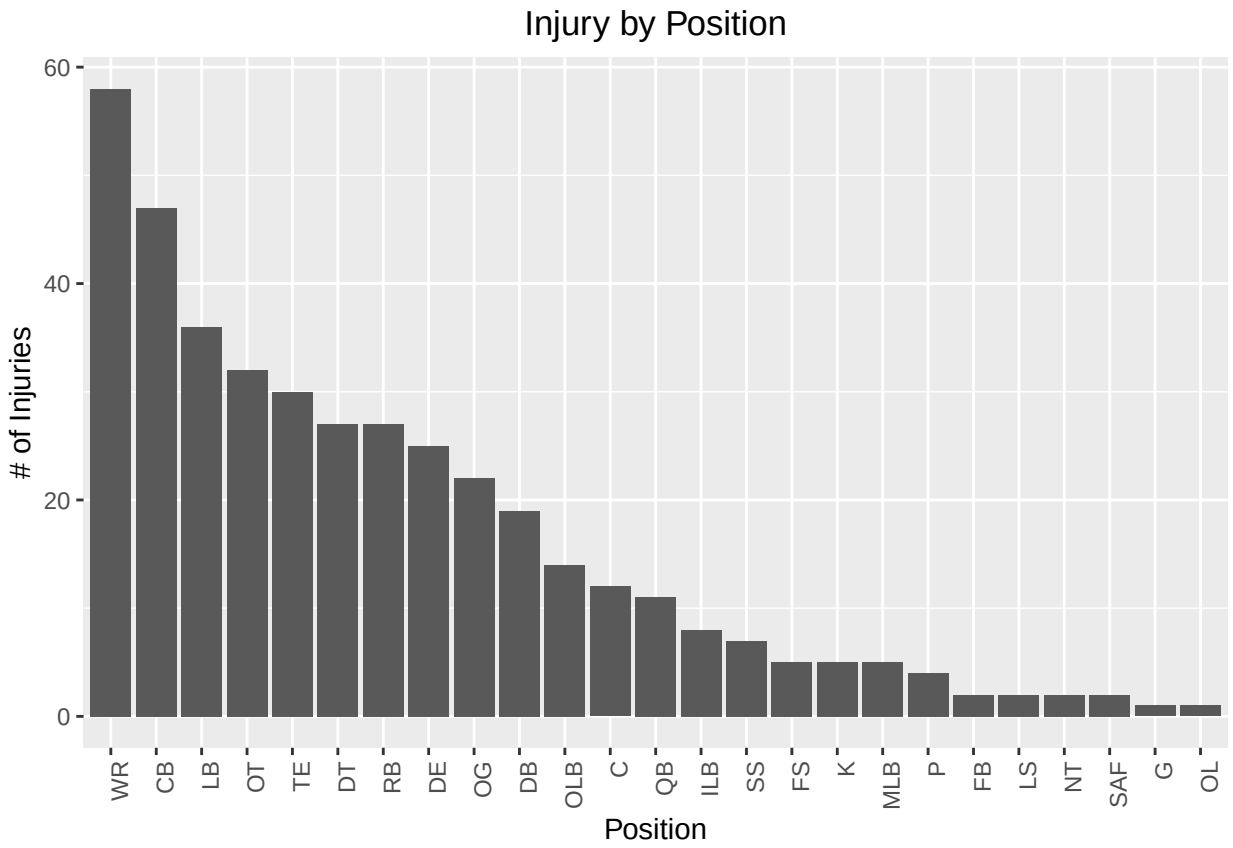
```
injuries %>% group_by(Injury) %>%
  summarize(num_injury = length(Injury)) %>%
  mutate(Injury = fct_reorder(Injury, num_injury, .desc = T)) %>%
  ggplot(aes(Injury, num_injury)) +
  geom_col() +
  ylab("Count") +
  ggtitle("Distribution of Injuries") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```




```r
# Injuries by Team
injuries %>% group_by(Team) %>%
  summarize(num_injury = length(Injury)) %>%
  arrange(desc(num_injury)) %>%
  mutate(Team = fct_reorder(Team, num_injury, .desc = T)) %>%
  ggplot(aes(Team, num_injury)) + geom_col() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  ylab("# of Injuries") +
  ggtitle("Injury by Team")
```




```r
# Injuries by Position
injuries %>% group_by(Position) %>%
  summarize(num_injury = length(Injury)) %>%
  arrange(desc(num_injury)) %>%
  mutate(Position = fct_reorder(Position, num_injury, .desc = T)) %>%
  ggplot(aes(Position, num_injury)) +
  geom_col() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  ylab("# of Injuries") +
  ggtitle("Injury by Position")
```



Logistic Regression

```
logi_fit = glm(data = injured_still_on_team,
               severe_injury ~ Age + Exp. + BMI + Offense, family = "binomial")

pander(summary(logi_fit))
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.426	2.537	-1.35	0.177
Age	0.2345	0.1103	2.126	0.03352
Exp.	-0.3102	0.1114	-2.783	0.005379
BMI	-0.007413	0.02754	-0.2692	0.7878
Offense	-0.1183	0.2718	-0.4353	0.6633

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	393.3 on 343 degrees of freedom
Residual deviance:	382.3 on 339 degrees of freedom

Logistic Regression w/Team + Position

```
logi_fit_Team_Pos = glm(data = injured_still_on_team,
                        severe_injury ~ Team + Position, family = "binomial")
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
pander(summary(logi_fit_Team_Pos))
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.64	1.289	1.273	0.2031
TeamATL	16.99	1424	0.01193	0.9905
TeamBAL	1.153	1.032	1.118	0.2638
TeamBUF	-0.2938	1.706	-0.1722	0.8633
TeamCAR	1.904	1.347	1.414	0.1575
TeamCHI	1.733	1.366	1.269	0.2044
TeamCIN	1.905	1.392	1.369	0.1711
TeamCLE	0.1212	0.961	0.1261	0.8996
TeamDAL	1.852	1.094	1.693	0.09054
TeamDEN	1.682	1.222	1.376	0.1688
TeamDET	1.755	1.117	1.571	0.1162
TeamGB	1.115	1.176	0.9475	0.3434
TeamHOU	0.915	1.223	0.748	0.4545
TeamIND	1.811	1.386	1.307	0.1913
TeamJAC	0.3033	1.201	0.2526	0.8006
TeamKC	0.8341	1.285	0.6493	0.5162
TeamLAC	0.5537	1.441	0.3843	0.7008
TeamLAR	-0.9519	1.088	-0.8752	0.3815
TeamLV	0.7841	1.137	0.6895	0.4905
TeamMIA	0.1424	1.094	0.1301	0.8965
TeamMIN	0.3098	1.222	0.2536	0.7998
TeamNE	0.7352	0.971	0.7572	0.449
TeamNO	0.2726	0.9876	0.276	0.7825
TeamNYG	0.5525	0.9555	0.5783	0.5631
TeamNYJ	1.387	1.061	1.307	0.1913
TeamPHI	0.4311	1.096	0.3935	0.694
TeamPIT	1.413	1.385	1.02	0.3075
TeamSEA	1.22	1.398	0.8728	0.3828
TeamSF	-0.3558	0.9641	-0.369	0.7121
TeamTB	-0.8547	1.181	-0.7238	0.4692
TeamTEN	1.518	1.119	1.357	0.1747
TeamWAS	0.4272	1.031	0.4145	0.6785
PositionCB	-1.436	1.17	-1.227	0.2199
PositionDB	0.6649	1.562	0.4258	0.6703
PositionDE	-1.45	1.248	-1.162	0.2453
PositionDT	-2.014	1.205	-1.672	0.09458
PositionFB	-2.118	1.899	-1.115	0.2648
PositionFS	-1.504	1.67	-0.9003	0.3679
PositionG	14.54	3956	0.003675	0.9971
PositionILB	-0.5923	1.658	-0.3573	0.7209
PositionK	15.31	1665	0.009195	0.9927
PositionLB	0.0272	1.281	0.02123	0.9831
PositionLS	15.84	2400	0.006601	0.9947
PositionMLB	-3.104	1.513	-2.052	0.0402

	Estimate	Std. Error	z value	Pr(>|z|)
PositionNT	-20.89	3956	-0.00528	0.9958
PositionOG	-0.8247	1.254	-0.6575	0.5108
PositionOL	15.37	3956	0.003886	0.9969
PositionOLB	-3.013	1.28	-2.354	0.01855
PositionOT	-1.063	1.206	-0.8816	0.378
PositionP	13.76	2121	0.006488	0.9948
PositionQB	-1.314	1.421	-0.925	0.355
PositionRB	-1.913	1.212	-1.578	0.1146
PositionSAF	15.33	2785	0.005503	0.9956
PositionSS	-2.051	1.437	-1.427	0.1535
PositionTE	-0.7158	1.23	-0.582	0.5606
PositionWR	-1.976	1.154	-1.713	0.08675

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	393.3 on 343 degrees of freedom
Residual deviance:	317.2 on 288 degrees of freedom

Maybe too much information in the outcome is lost by making injury binary, maybe ordinal or multinomial would be preferred.

We also might want to use data from past years as we are only half way through the current season.

Ordinal Regression???

```
'injuries %<>% mutate(Injury_Ord =
                        case_when(
                          str_detect(Injury.Status, "IR") ~ "IR",
                          str_detect(Injury.Status, "Expected Return") ~ "Expected Return",
                          str_detect(Injury.Status, "Questionable") ~ "Questionable",
                          str_detect(Injury.Status, "NFI") ~ "Non-Football Injury",
                          TRUE ~ "other"
                        )
                      ) '
```

```
## [1] "injuries %<>% mutate(Injury_Ord = \n                        case_when(\n
```

```
#View(cbind.data.frame(injuries$Injury.Status, injuries$Injury_Ord))
```