INTR 202

Professor Hu

Peter Dolan & Hyatt Sbar

April 9, 2023

Relationship Between Boxing Wins Depending on Knock Out Rate and Stance

Our group is studying the relationship between a boxer's Wins depending on KO rate and stance. This information is vital to professional athletes who need to know if they should be training to knock people out or go the distance. If boxers know that knockouts directly corelate to wins, they will adjust their training appropriately to gain power over strength in order to boost their win percentage because winning equals getting paid. In addition to knockout rate, fighters could better train for certain fights if they know changing stance will better their odds of winning. According to Dimitar Ivanov in an article written for shortboxing.com, the highest knockout percentage is in the heaviest weight-class which is "heavyweight". If the heavyweights are knocking the most people out because they are the most powerful weight-class, and knockout percentage directly corelates to win-rate, then fighters want to pack on as much weight as possible after weighing in in order to increase their odds of winning because that weight will help them knock the opposing fighters out. Additionally, in an article published by the U.S sun, Jack Figg writes that one of the most prolific knockout fighters in boxing history is Deontay Wilder, who is a heavyweight with a knockout-to-win percentage of 97.67%. Figg writes about multiple knockout artists who all retained titles in the heavyweight division.

Information Referenced above is from these two links:
https://www.the-sun.com/sport/boxing/175076/biggest-knockout-fighters-deontay-wilder/

https://shortboxing.com/what-percentage-of-boxing-matches-end-in-a-knockout/

We obtained our data from the website, Kaggle.com, which is the world's largest data science community. There are a few potential biases and limitations of our data due to the nature of boxing. First, there is a sampling bias. In other words, only professional boxers were used in the sampling, which excludes amateur boxers. Believe it or not, amateur boxers make up a significant amount of the sport, so this could be limiting. There could also be a selection bias at play. The vast majority of the names on the list are big name boxers who have had decently long careers, so our data could have been compiled just by someone picking the best boxers that were household names. This could limit our data because boxers not as popular could be excluded on accident. Also, there could have been confounding variables because there could be externalities at play that are out of our data's control. For example, aggression, reach, or height could affect knockout rate and win rate. And lastly, there could be a general lack of context within our data. For example, some opponents could have been mismatched in skill, which resulted in a higher knockout average than what we would normally expect. In addition, boxing is a very technical sport in which many moving parts make up a professional bout. Because of this, externalities could affect outcome that are not able to be displayed by normal data like we have gathered. For

example, if you lose three or more times in a row as a boxer, most likely your career is over, which means data could be strewn.

R Code:

```
library(ggplot2)
library(dplyr)
library(psych)
library(readr)
fighters <- read_csv("fighters.csv", col_types = cols(age = col_double()))
View(fighters)
```

Means:

```
mean(fighters$wins, na.rm = TRUE)
mean(fighters$looses, na.rm = TRUE)
mean(fighters$draws, na.rm = TRUE)
mean(fighters$ko_rate_clean, na.rm = TRUE)
mean(fighters$age, na.rm = TRUE)
```

```
Wins: 19.67319
Loses: 4.489493
Draws: 0.9565217
KO Rate: 28.16522
Age: 48.82105
```

Medians:

```
median(fighters$wins, na.rm = TRUE)
median(fighters$looses, na.rm = TRUE)
median(fighters$draws, na.rm = TRUE)
median(fighters$ko_rate_clean, na.rm = TRUE)
median(fighters$age, na.rm = TRUE)
```

```
Wins: 9
Loses: 1
Draws: 0
KO Rate: 24
Age: 37
```

Range:

```
range(fighters$wins, na.rm = TRUE)
range(fighters$looses, na.rm = TRUE)
range(fighters$draws, na.rm = TRUE)
range(fighters$ko_rate_clean, na.rm = TRUE)
range(fighters$age, na.rm = TRUE)
```

```
Wins: 0 262
Loses: 0 190
Draws: 0 60
KO Rate: 0 100
Age: 0 221
```

Standard Deviation:

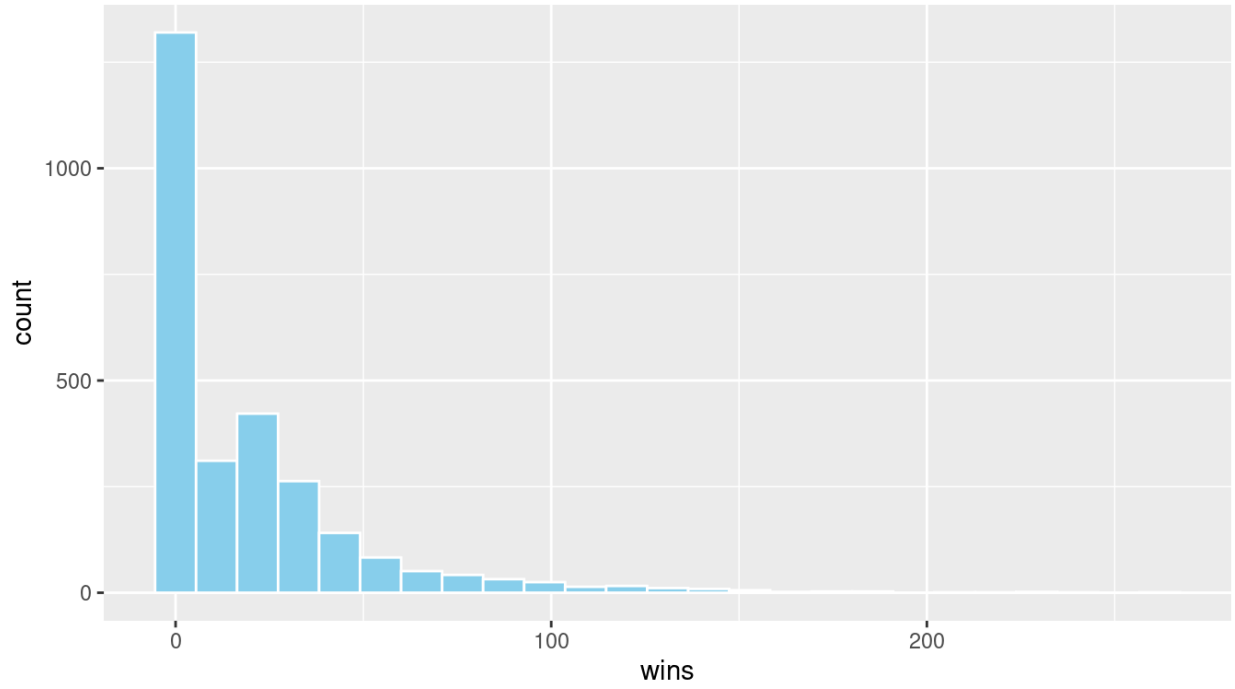
```
sd(fighters$wins, na.rm = TRUE)
sd(fighters$looses, na.rm = TRUE)
sd(fighters$draws, na.rm = TRUE)
sd(fighters$ko_rate_clean, na.rm = TRUE)
sd(fighters$age, na.rm = TRUE)
```

```
Wins: 29.4948
Loses: 10.0413
Draws: 2.876004
KO Rate: 28.96937
Age: 26.62534
```

Graphs:

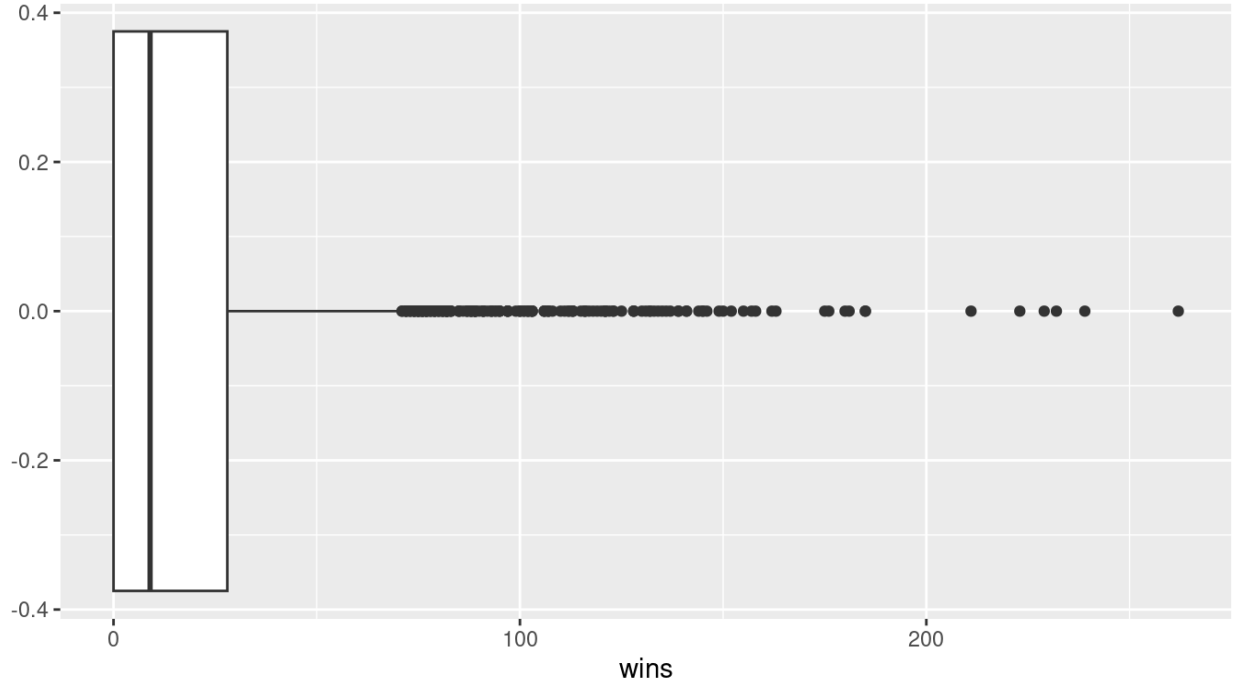
Wins Histogram

```
ggplot(data = fighters, aes(x = wins)) +
  geom_histogram(bins = 25,
                 fill = "skyblue",
                 color = "white")
```



Wins Boxplot

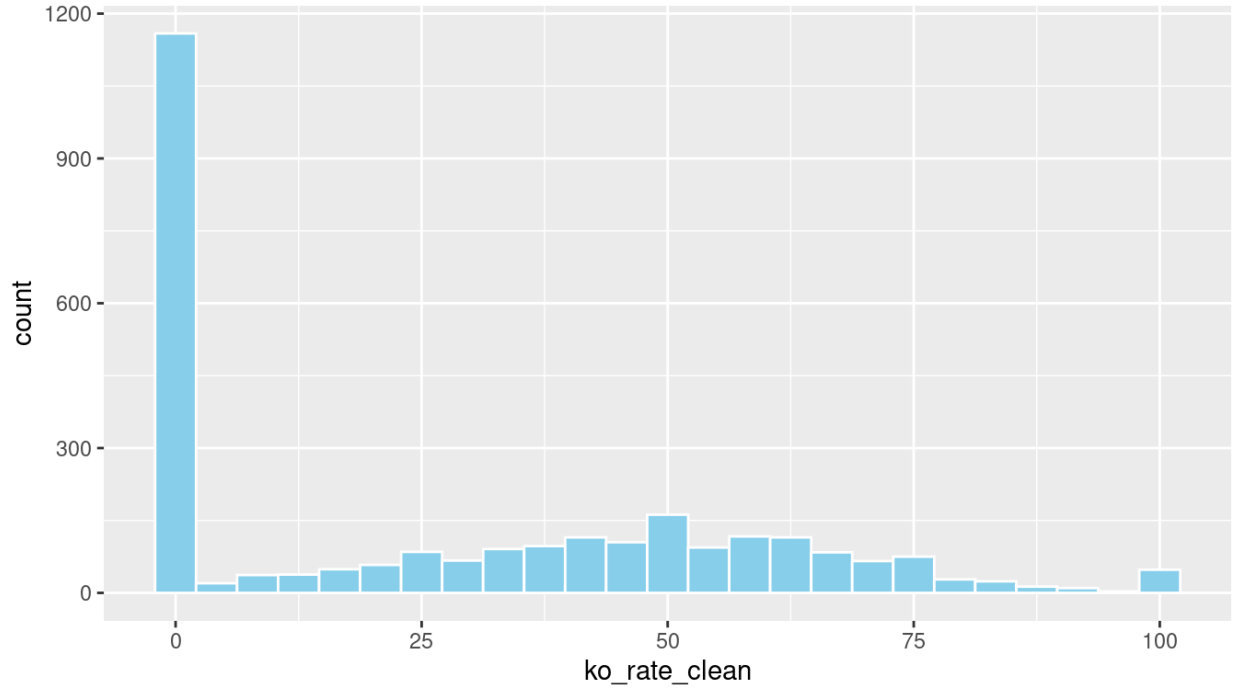
```
ggplot(data = fighters, aes(x = wins)) +
  geom_boxplot()
```



KO Rate Histogram

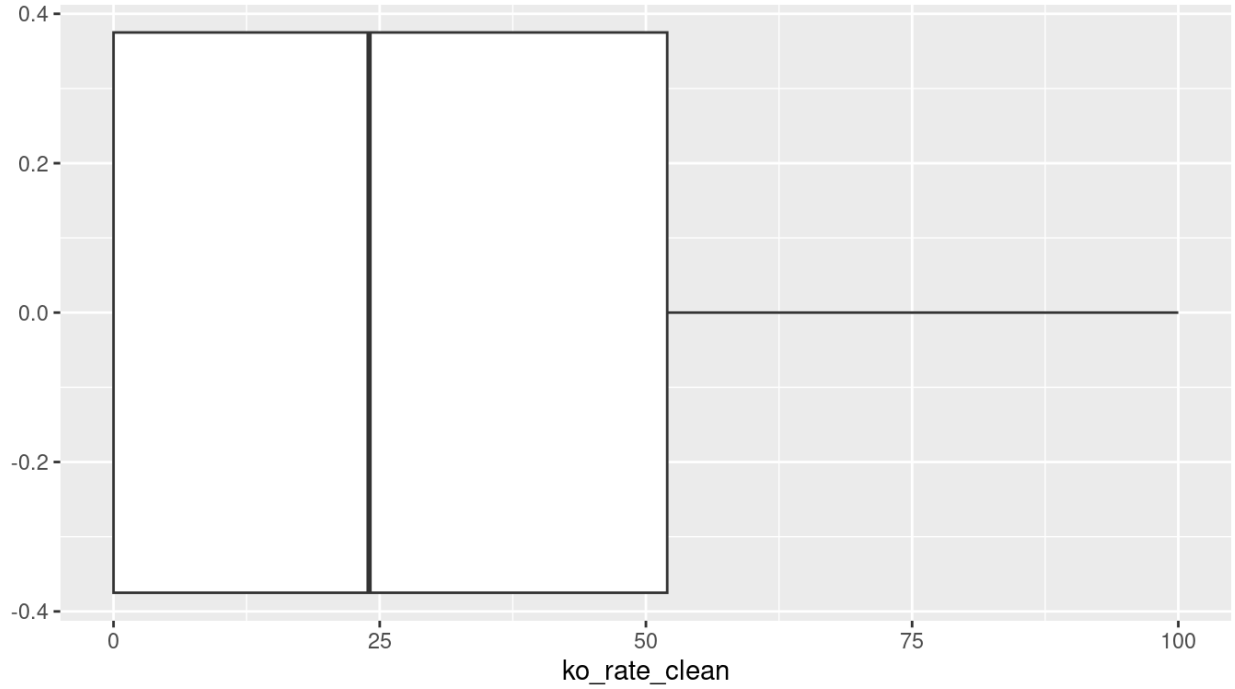
```
ggplot(data = fighters, aes(x = ko_rate_clean)) +
  geom_histogram(bins = 25,
```

```
fill = "skyblue",
color = "white")
```



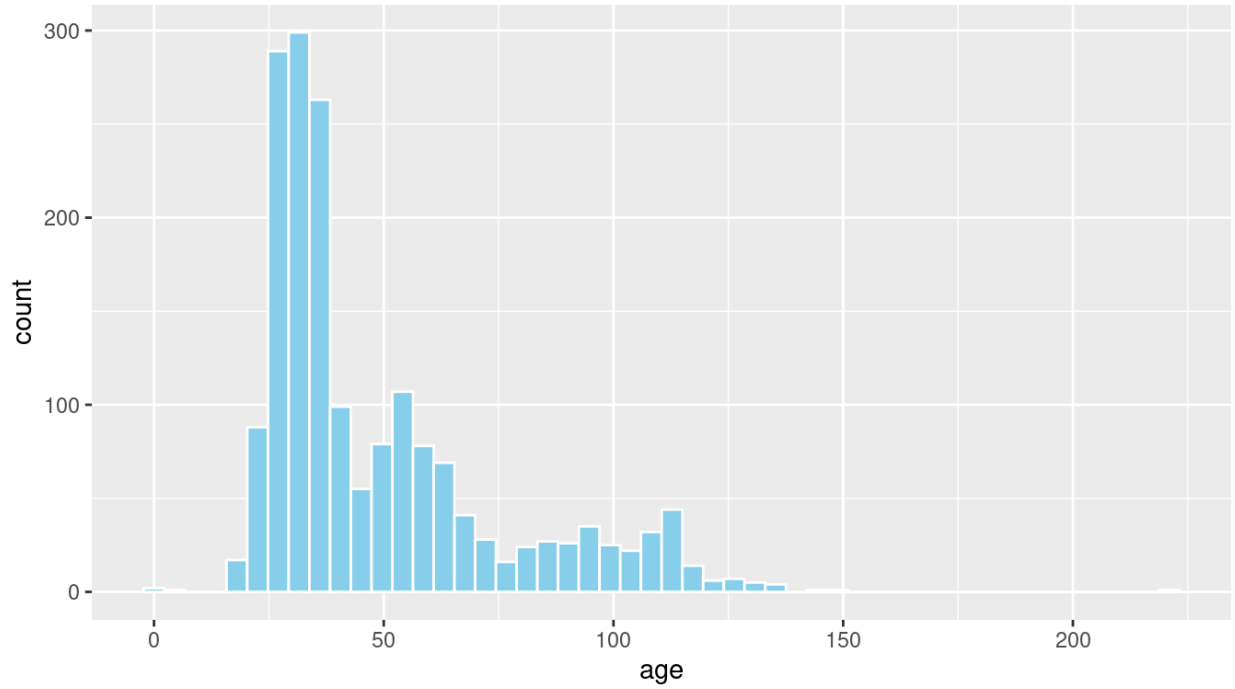
KO Rate Boxplot

```
ggplot(data = fighters, aes(x = ko_rate_clean)) +
 geom_boxplot()
```



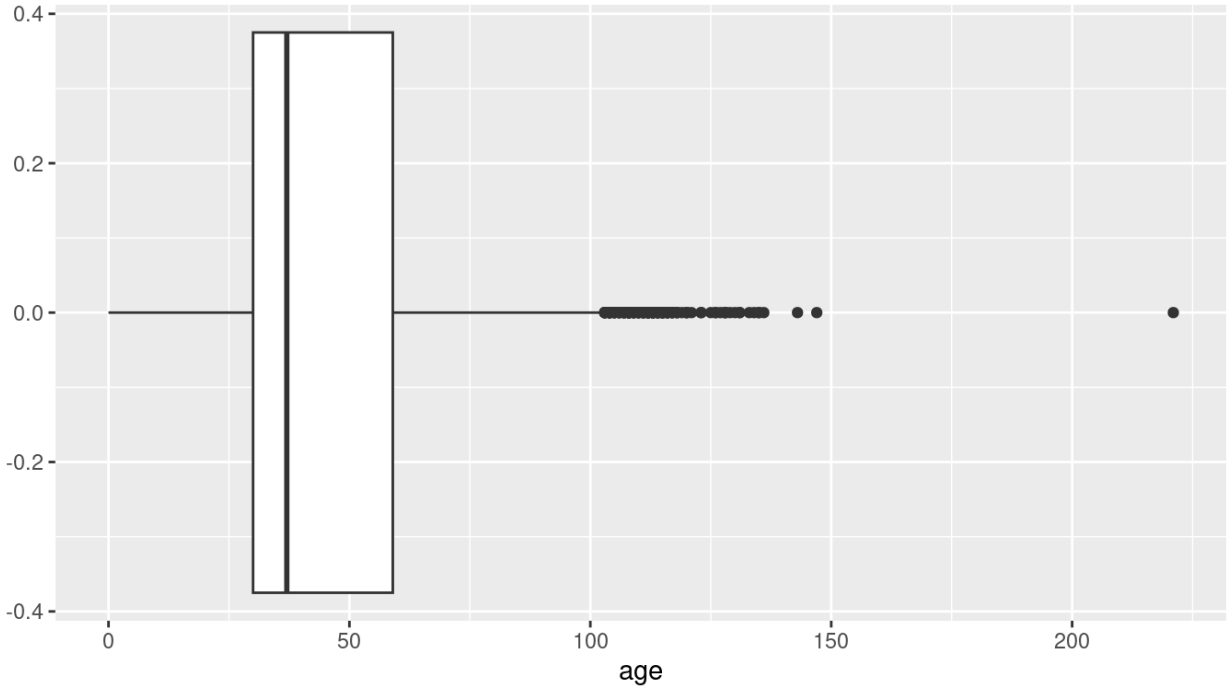
Age Histogram

```
ggplot(data = fighters, aes(x = age)) +
  geom_histogram(bins = 50,
                 fill = "skyblue",
                 color = "white")
```



Age Boxplot

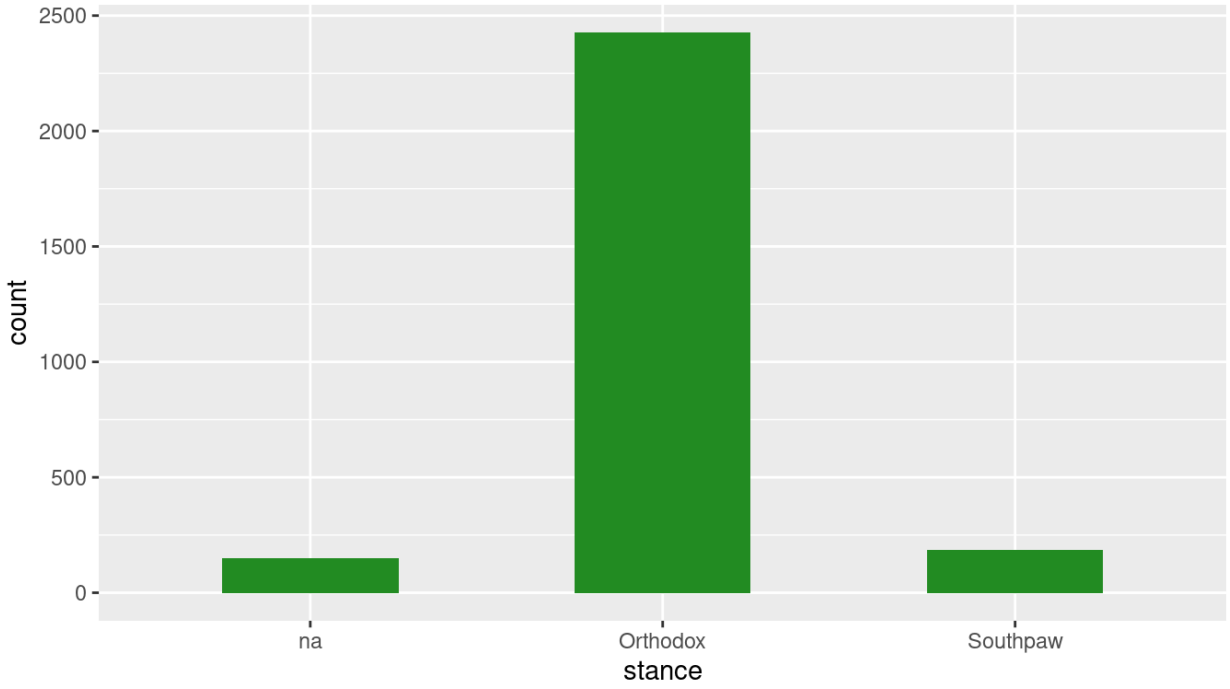
```
ggplot(data = fighters, aes(x = age)) +
  geom_boxplot()
```



Proportions of Categorical Variables:
By Stance:

```
tb_stance <- prop.table(table(fighters$stance))
tb_stance

ggplot(data = fighters, aes(x = stance)) +
  geom_bar(stat = "count", width = .5, fill = "forest green")
```

```
NA: 0.05398551          Orthodox: 0.87934783          Southpaw: 0.06666667
```

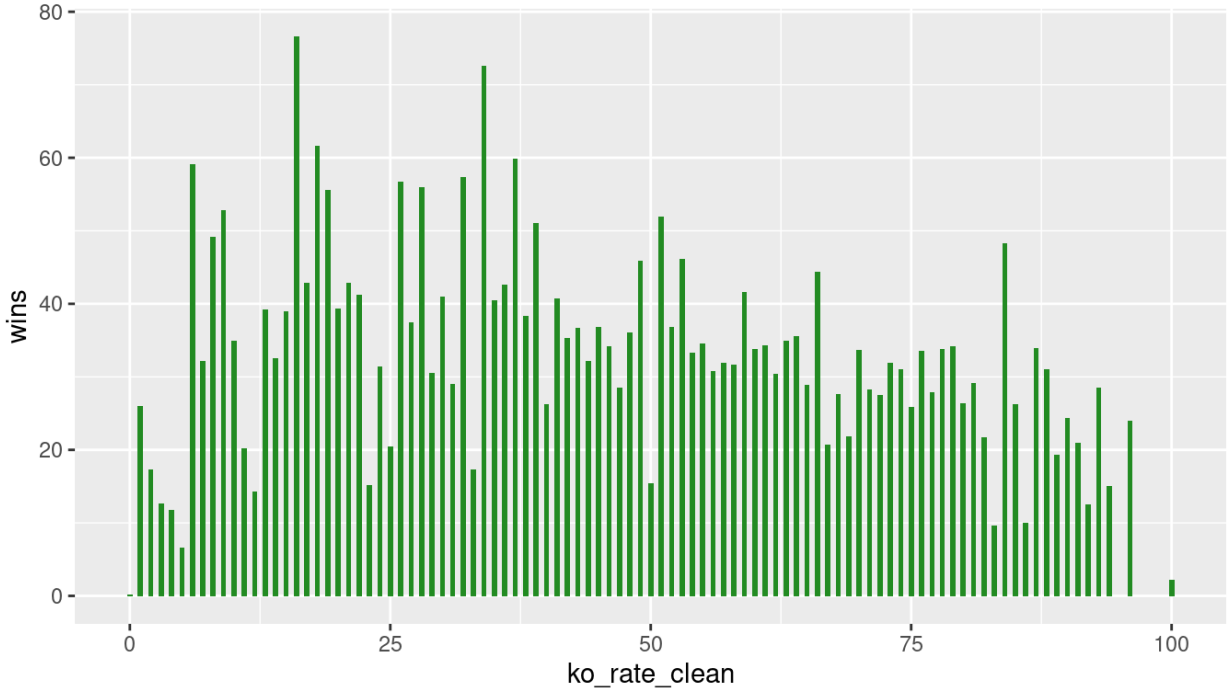


Bar chart of Group Means:

Wins to KO Rate

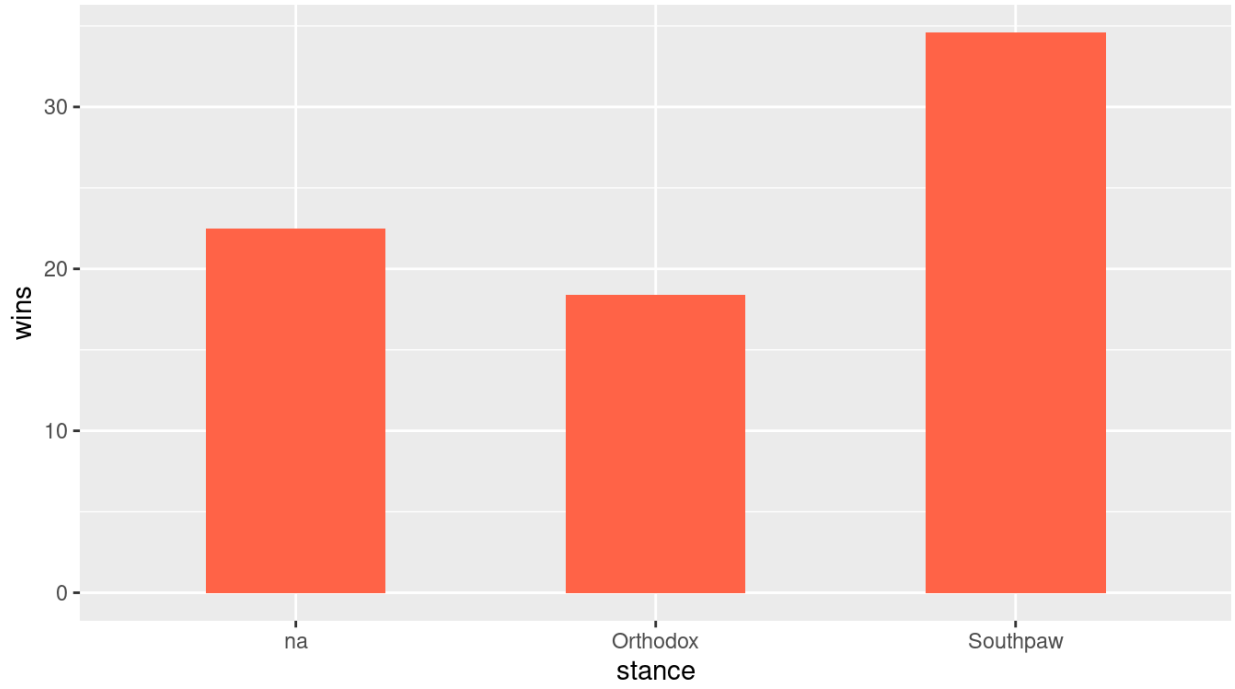
```
aggregate(formula = wins ~ ko_rate_clean, data = fighters, FUN = mean)
```

```
ggplot(fighters, aes(y = wins, x = ko_rate_clean)) +
  geom_bar(stat = "summary", fun = "mean", width = 0.5, fill = "forest green")
```



Wins By Stance:

```
aggregate(formula = wins ~ stance, data = fighters, FUN = mean)

ggplot(fighters, aes(y = wins, x = stance)) +
  geom_bar(stat = "summary", fun = "mean", width = 0.5, fill = "tomato")
```



ANOVA Test

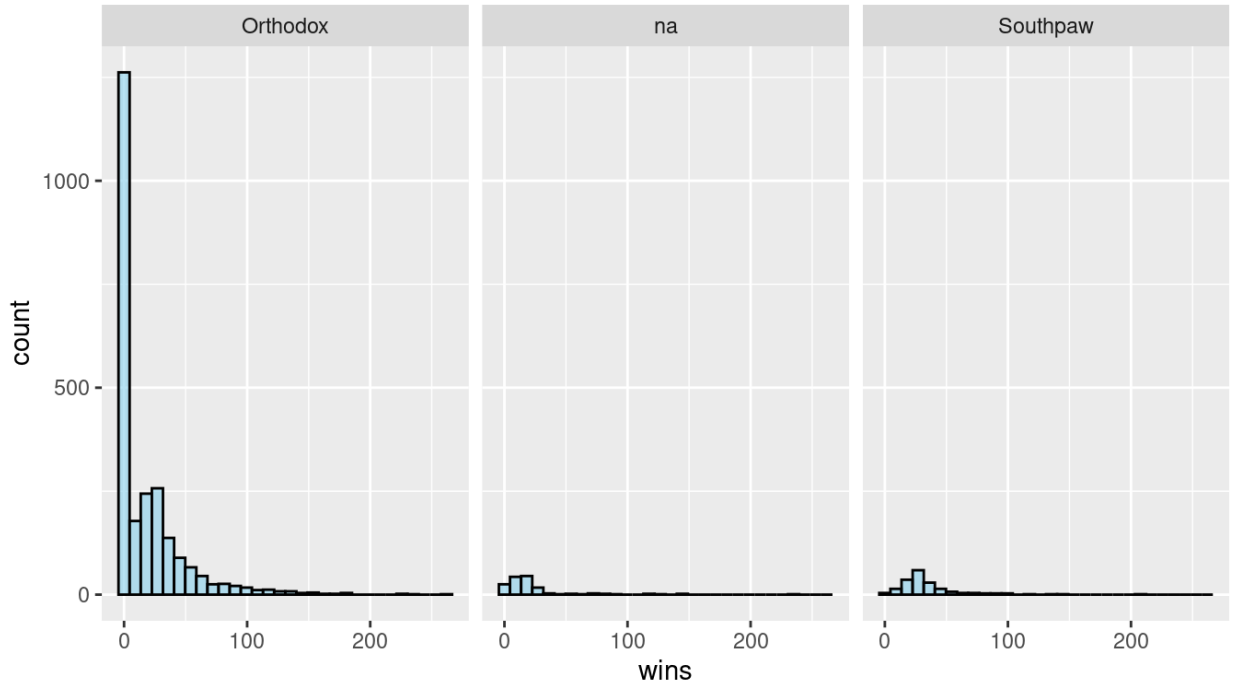
```
md <- aov(formula = cbind(wins, ko_rate_clean) ~ stance, data = fighters)
summary(md)

Response wins :
             Df  Sum Sq Mean Sq F value    Pr(>F)
stance        2   46264 23132.0  27.093 2.227e-12 ***
Residuals   2757 2353909   853.8
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response ko_rate_clean :
             Df  Sum Sq Mean Sq F value    Pr(>F)
stance        2  114410   57205  71.655 < 2.2e-16 ***
Residuals   2757 2201011     798
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


aggregate(formula = wins ~ stance, data = fighters, FUN = sd)
    stance     wins
1 Orthodox 29.26748
```

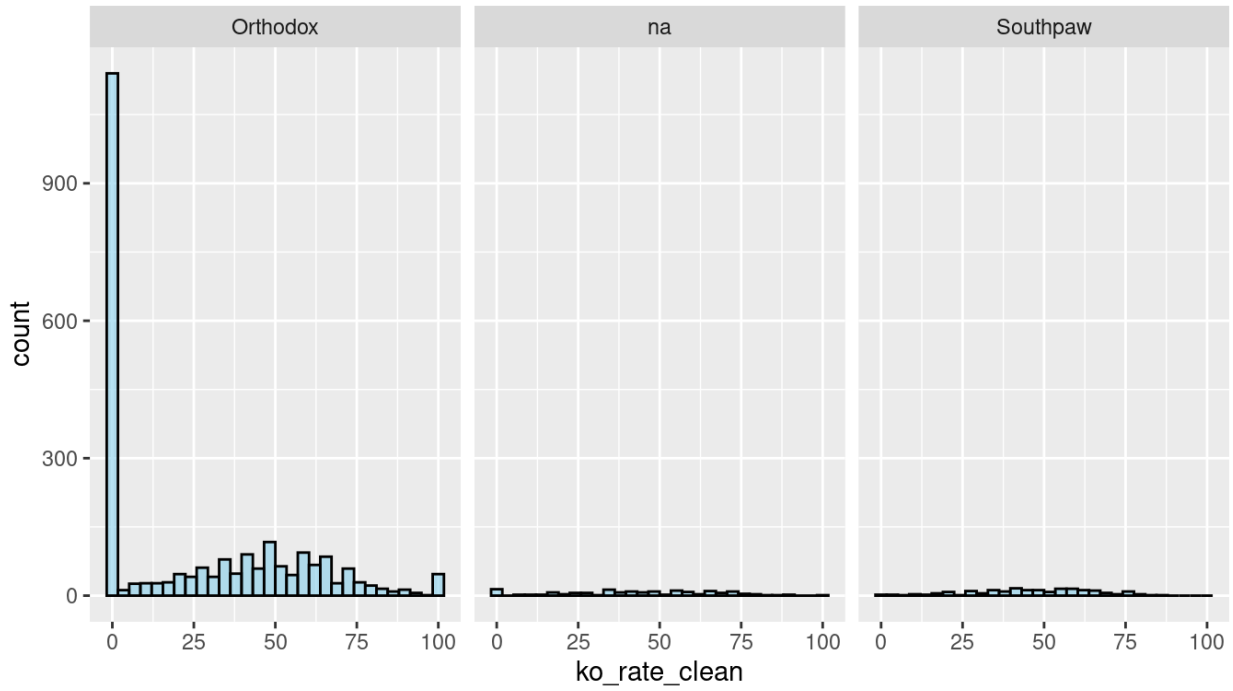
2 na 31.38009
3 Southpaw 26.66287

```
aggregate(formula = ko_rate_clean ~ stance, data = fighters, FUN = sd)
```

```
    stance ko_rate_clean
1 Orthodox      29.09339
2       na      24.01776
3 Southpaw      18.43733
```

Assumption Graphs For ANOVA:





ANOVA Test Interpretation:

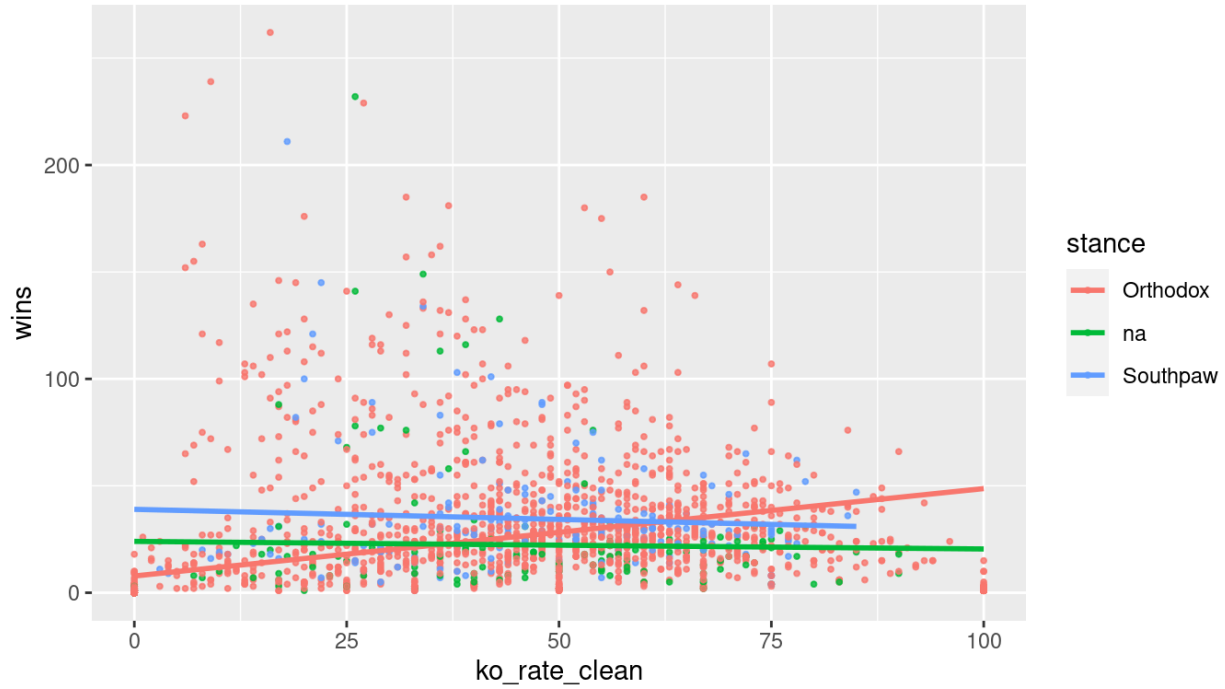
Both tests show small P values meaning there is a significant difference in mean of wins by stance as well as mean of KO Rate by stance. Therefore, it may be true that stance effects boxing wins as well as KO Rate. This conclusion leads to multiple regression test to see the definitive relationship between the variables.

Multiple Regression:

```
md <- lm(formula = wins ~ ko_rate_clean + stance, data = fighters)
summary(md)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-45.368  -8.641  -8.641   0.891 247.323

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       5.94931    2.36937   2.511 0.012098 *
ko_rate_clean     0.37727    0.01834  20.569  < 2e-16 ***
stanceOrthodox    2.69184    2.32018   1.160 0.246074
stanceSouthpaw   10.96400    2.99945   3.655 0.000262 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27.21 on 2756 degrees of freedom
Multiple R-squared:  0.1498,    Adjusted R-squared:  0.1489
F-statistic: 161.9 on 3 and 2756 DF,  p-value: < 2.2e-16
```

```
fighters$stance <- relevel(factor(fighters$stance), ref = "Orthodox")

ggplot(fighters, aes(x = ko_rate_clean, y = wins, color = stance)) +
  geom_point(size = 0.7, alpha = 0.8) +
  geom_smooth(method = "lm", se = FALSE)
```



Multiple Regression Interpretation:

Null Hypothesis that there is no relationship between Wins depending on Knock Out Rate and Stance.

Alternative Hypothesis is that there is a relationship between Wins depending on Knock Out Rate and Stance

The multiple regression finds that there is a relationship between Wins and KO Rate for the Southpaw Stance, however the Orthodox stance still has no relationship.

Southpaw Stance and KO Rate have p values of 0.000262 and < 2e-16 while Orthodox has a p value much higher than .05. Therefore, it can be said that Southpaw Stance and KO Rate are likely to have a greater number of wins.

Multiple Regression Limitations:
The multiple Regression meets the linearity assumption as well as the independence assumption. However, the normality assumption is not met for all the variables due to the high number of zero values for the KO Rate and Wins throughout the dataset. However, this likely balances out due to the high number of observations in the dataset.

ANOVA Limitations:
This test was conducted with some errors in assumptions for the data is not all normally distributed, however this is likely made up for by the size of our data with 2,760 observations. The variance between groups is not all the same but they are not dramatically different either, so it should likely be negligible.

The Conclusion that Southpaw Stance has a positive relationship with wins for boxers is somewhat surprising given that the proportion of Orthodox fighters is vastly greater than that of Southpaw, shown in the Proportions by Stance graph. However, this is likely explained by the aggregate mean of wins for Southpaw fighters being greater than that of Orthodox fighters shown in the wins by stance graph.

Many of the limitations referenced above in the second paragraph of this project hold the same with the analyzed data, especially in relation to normal distribution. Many of our variables did not have a normal distribution due to only a few number of boxers being good enough to have numerous wins or a KO Rate above around 30%. This plays heavily into the number of boxers with zero wins who likely took a beating and retired after two or three losses. This disparity in the distribution of the data likely skewed some of our conclusions after the data was analyzed, especially with the relationship between aggregate mean by stance and proportion of fighters in the paragraph above.

So to review the ANOVA table shows a likely relationship of stance effecting wins as well as KO rate, due to the test showing low p values for both variables. The multiple regression to show this relationship shows that there is a positive relationship for wins for both KO Rate and Stance due to the p values of $< 2e\text{-}16$ and 0.000262. This means that to statistically maximize wins boxers should fight Southpaw and of course go for the Knock Out.