UFC data

Intro

I decided to use UFC bout and personal fighter statistics. I am personally a fan of the sport and have done martial arts since I was a small kid which gives me some domain knowledge. I see a lot of great statistics and data here that would be cool to dissect. I am using a data set of historical UFC data taken from Kaggle.com. It provides a lot of data and variables but due to the denseness of this data set it make be hard to properly analyze. The fighter statistics are also just what is current and not at the times of different bout which would make a model difficult to be used on live data.

Variables

This is a general Key for the variables
R_ and B_ prefix signifies red and blue corner fighter stats respectively.
opp containing columns is the average of damage done by the opponent on the fighter.
KD is number of knockdowns.
SIG_STR is no. of significant strikes 'landed of attempted'.
SIG_STR_pct is significant strikes percentage.
TOTAL_STR is total strikes 'landed of attempted'.
TD is no. of takedowns.
TD_pct is takedown percentages.
SUB_ATT is no. of submission attempts.
PASS is no. times the guard was passed?
REV is the no. of Reversals landed.
HEAD is no. of significant strinks to the head 'landed of attempted'.
BODY is no. of significant strikes to the body 'landed of attempted'.
CLINCH is no. of significant strikes in the clinch 'landed of attempted'.
GROUND is no. of significant strikes on the ground 'landed of attempted'.
win_by is method of win.
last_round is last round of the fight (ex. if it was a KO in 1st, then this will be 1).
last_round_time is when the fight ended in the last round.
Format is the format of the fight (3 rounds, 5 rounds etc.).
Referee is the name of the Ref.
date is the date of the fight.
location is the location in which the event took place.
Fight_type is which weight class and whether it's a title bout or not.
Winner is the winner of the fight.
Stance is the stance of the fighter (orthodox, southpaw, etc.).
Height_cms is the height in centimeter.
Reach_cms is the reach of the fighter (arm span) in centimeter.
Weight_lbs is the weight of the fighter in pounds (lbs).
age is the age of the fighter.
title_bout Boolean value of whether it is title fight or not.
weight_class is which weight class the fight is in (Bantamweight, heavyweight, Women's flyweight, etc.).

no_of_rounds is the number of rounds the fight was scheduled for.
current_lose_streak is the count of current concurrent losses of the fighter.
current_win_streak is the count of current concurrent wins of the fighter.
draw is the number of draws in the fighter's ufc career.
wins is the number of wins in the fighter's ufc career.
losses is the number of losses in the fighter's ufc career.
total_rounds_fought is the average of total rounds fought by the fighter.
total_time_fought(seconds) is the count of total time spent fighting in seconds.
total_title_bouts is the total number of title bouts taken part in by the fighter.
win_by_Decision_Majority is the number of wins by majority judges decision in the fighter's ufc career.
win_by_Decision_Split is the number of wins by split judges decision in the fighter's ufc career.
win_by_Decision_Unanimous is the number of wins by unanimous judges decision in the fighter's ufc career.
win_by_KO/TKO is the number of wins by knockout in the fighter's ufc career.
win_by_Submission is the number of wins by submission in the fighter's ufc career.
win_by_TKO_Doctor_Stoppage is the number of wins by doctor stoppage in the fighter's ufc career.

Data exploration

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.1.1
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
```

```
## Warning: package 'ggplot2' was built under R version 4.1.1
```

```
## Warning: package 'tibble' was built under R version 4.1.1
```

```
## Warning: package 'tidyr' was built under R version 4.1.1
```

```
## Warning: package 'readr' was built under R version 4.1.1
```

```
## Warning: package 'purrr' was built under R version 4.1.1
```

```
## Warning: package 'dplyr' was built under R version 4.1.1
```

```
## Warning: package 'stringr' was built under R version 4.1.1
```

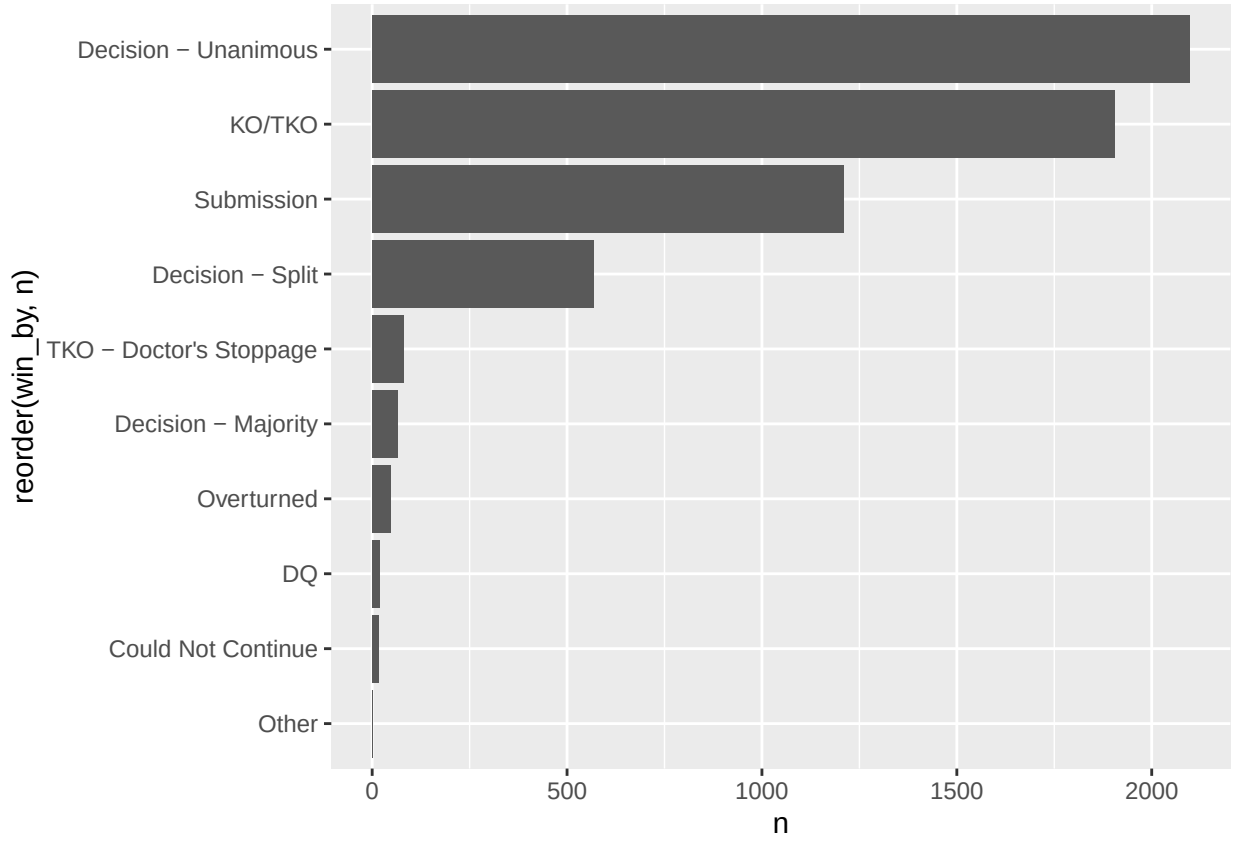
```
## Warning: package 'forcats' was built under R version 4.1.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

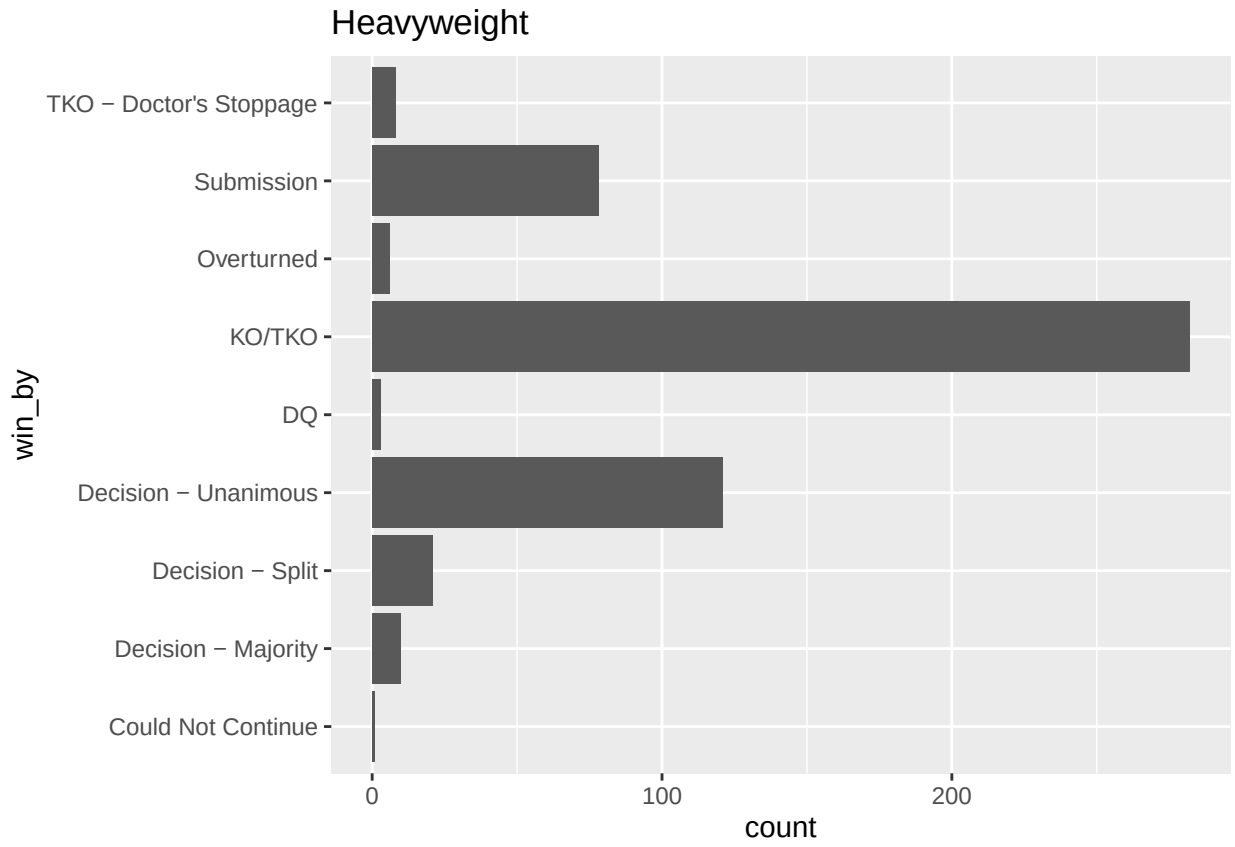
```
library(dplyr)

data <- read.csv("data.csv")
preprocessed_data <- read.csv("preprocessed_data.csv")
raw_fighter_details <- read.csv("raw_fighter_details.csv")
raw_total_fight_data <- read.csv("raw_total_fight_data.csv", sep = ";")

#
# str(data)
# str(raw_fighter_details)
# str(raw_total_fight_data)
```

Win_by analysis

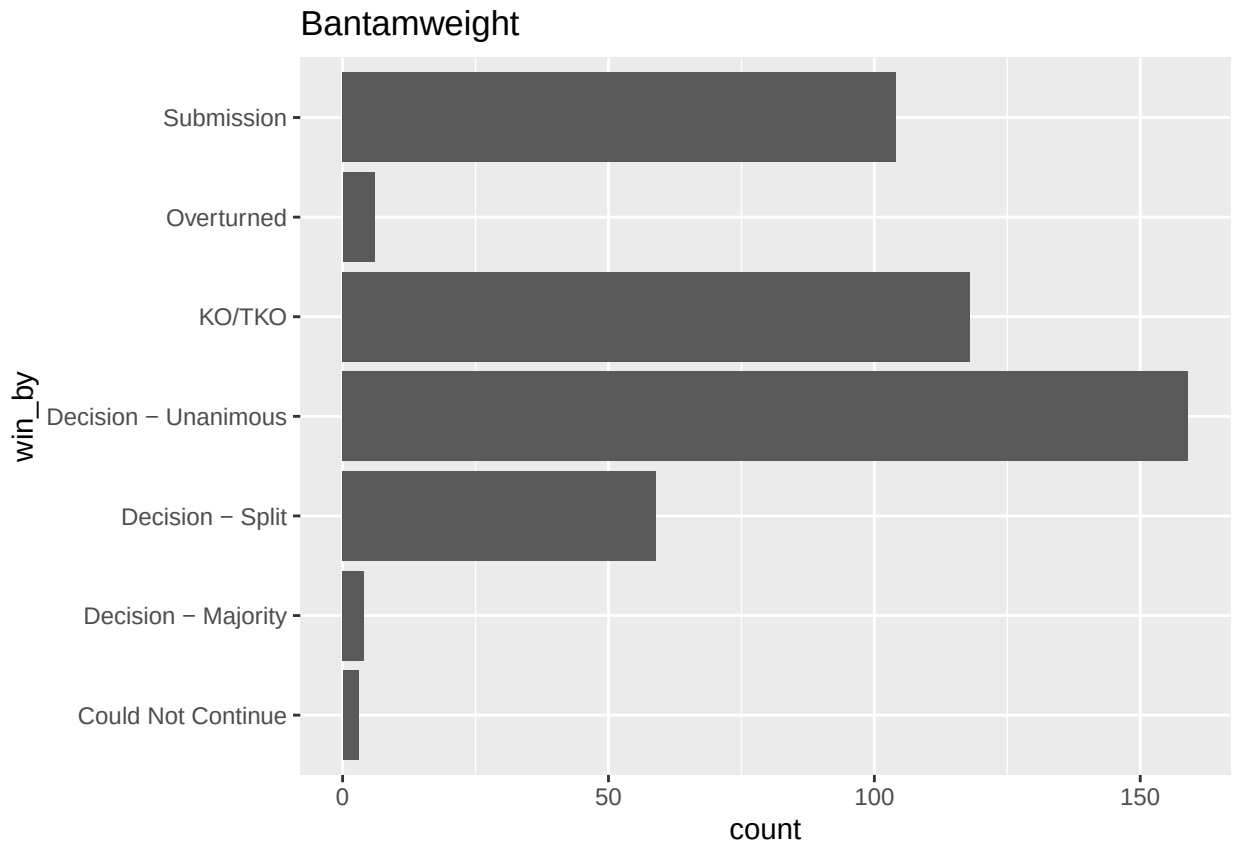
```
raw_total_fight_data %>%
  count(win_by, sort = TRUE) %>%
  head(10) %>%
  ggplot(aes(x = n, y = reorder(win_by, n))) + geom_col()
```




```
# Effect of Weight class of Win_by
raw_total_fight_data %>%
  filter(Fight_type == "Heavyweight Bout") %>%
  ggplot(aes(y = win_by)) + geom_bar() +labs(title = "Heavyweight")
```




```
raw_total_fight_data %>%
  filter(Fight_type == "Bantamweight Bout") %>%
  ggplot(aes(y = win_by)) + geom_bar() +labs(title = "Bantamweight")
```

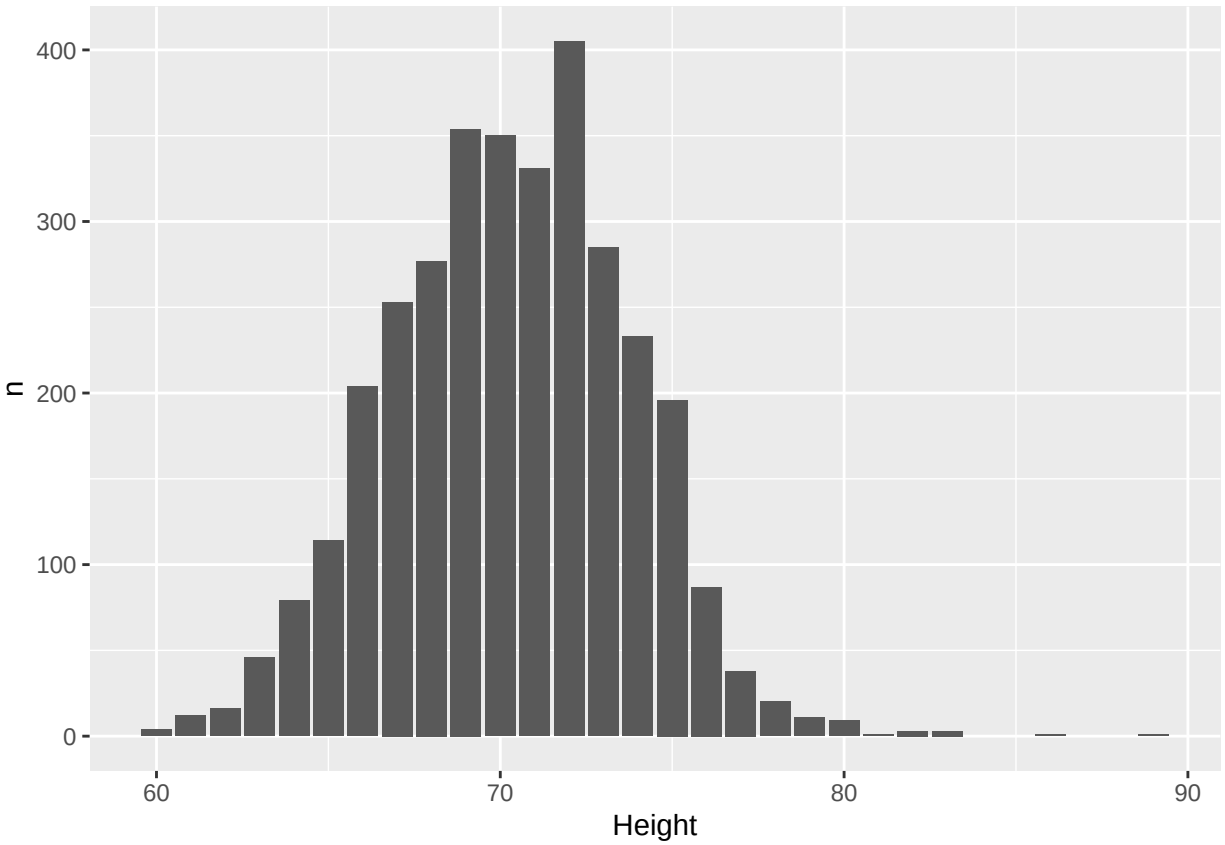


Height and Weight distribution

```
Proc_fighter_details <- raw_fighter_details

Proc_fighter_details$Height <-  substr(raw_fighter_details$Height, 1, nchar(raw_fighter_details$Height)

Proc_fighter_details %>%
  separate(Height, c('feet', 'inch'), sep = '\'', convert = TRUE, remove = FALSE)%>%
  mutate(Height = 12 * feet + inch , na = TRUE) %>%
  count(Height) %>%
  ggplot(aes(x = Height, y = n)) + geom_col()
```

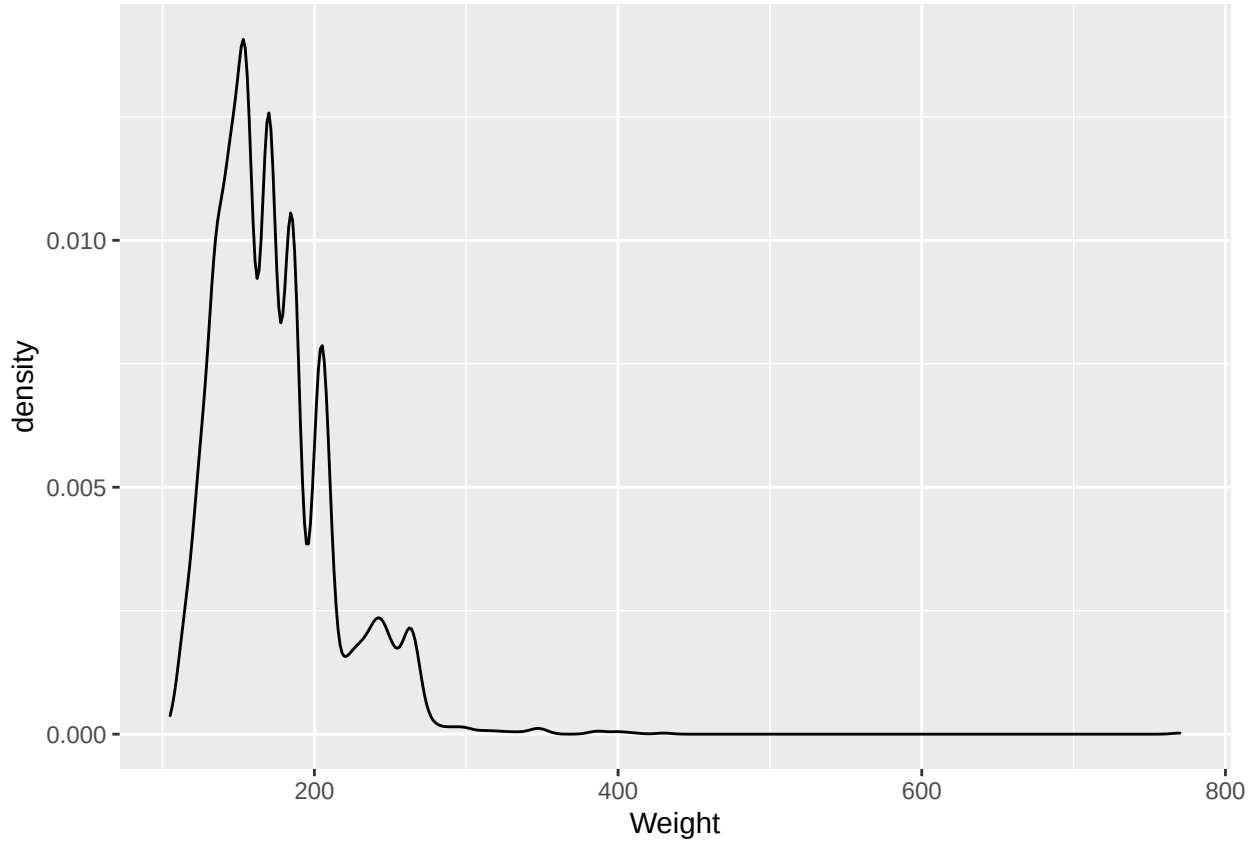
```
## Warning: Expected 2 pieces. Missing pieces filled with `NA` in 263 rows [1, 35,
## 41, 42, 58, 82, 92, 116, 132, 133, 196, 206, 211, 219, 223, 264, 275, 284, 288,
## 291, ...].
```

```
## Warning: Removed 1 rows containing missing values (position_stack).
```



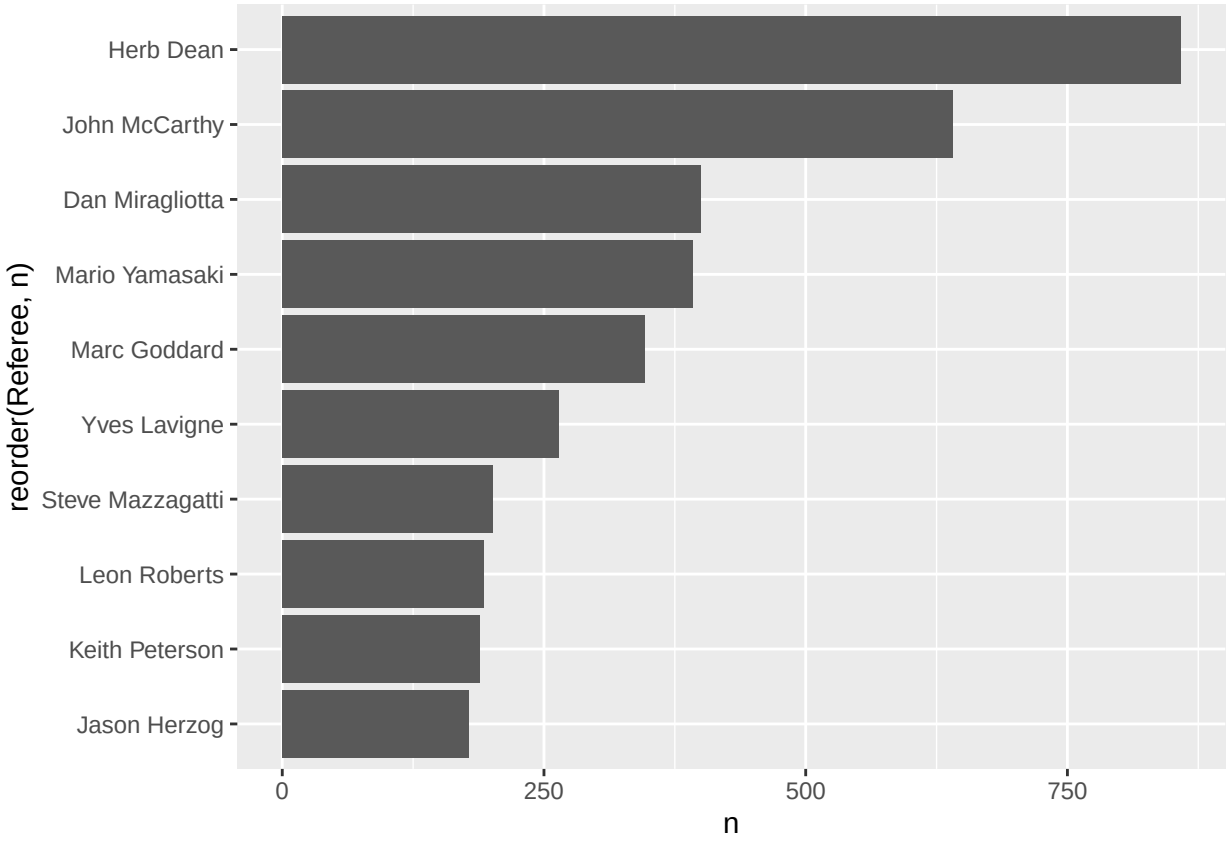

```
Proc_fighter_details$Weight <-  as.numeric(substr(raw_fighter_details$Weight, 1, nchar(raw_fighter_deta

Proc_fighter_details %>%
  ggplot(aes(x = Weight)) + geom_density()
```

```
## Warning: Removed 74 rows containing non-finite values (stat_density).
```



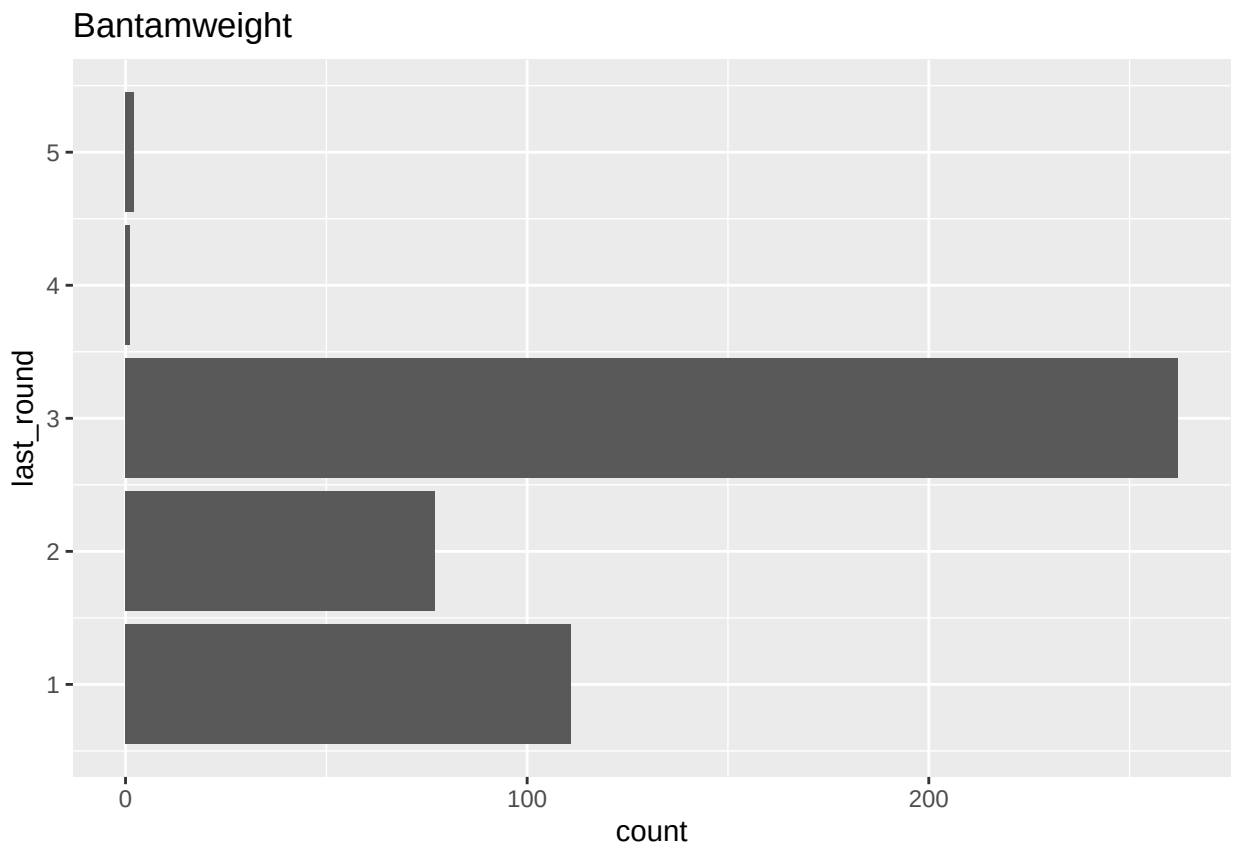
Distribution of referees

```
raw_total_fight_data %>%
  count(Referee, sort = TRUE) %>%
  head(10) %>%
  ggplot(aes(x = n, y = reorder(Referee, n))) + geom_col()
```

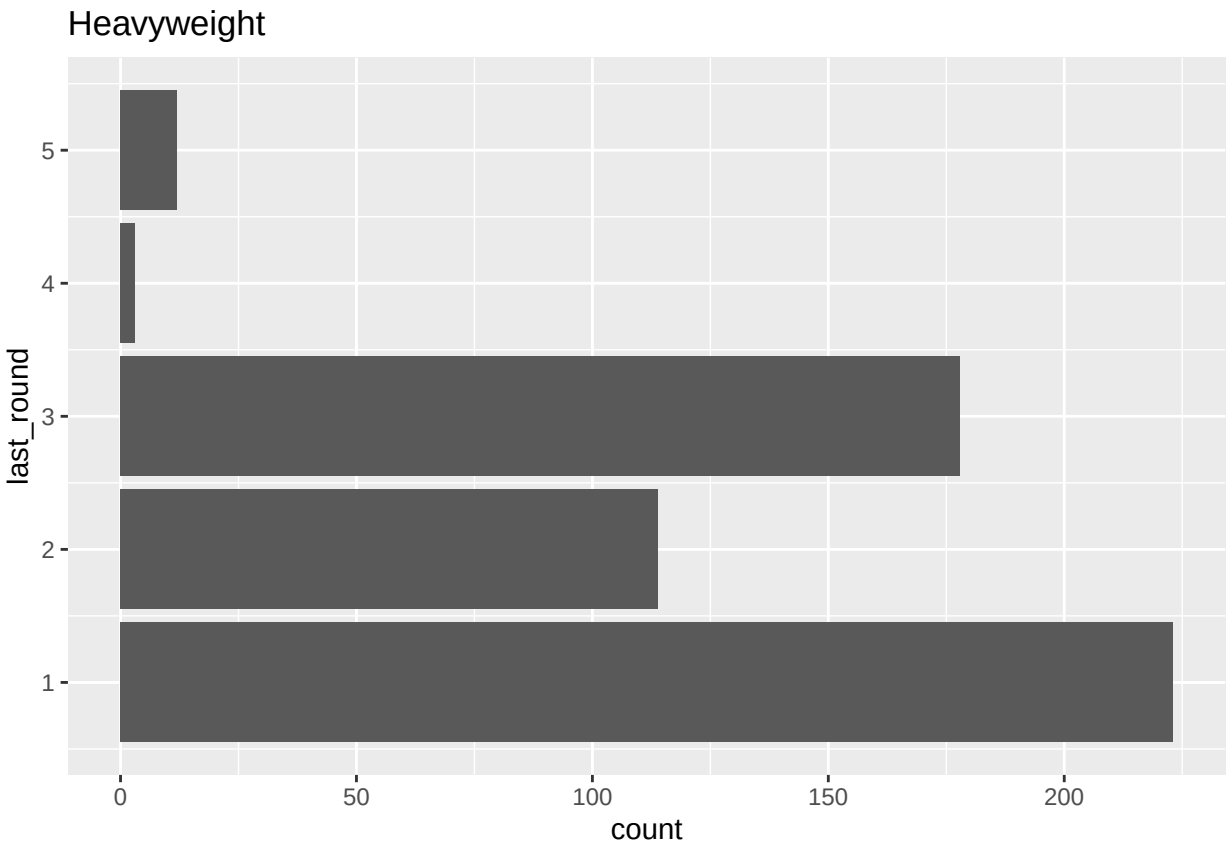


Effect of last round by weight class

```
raw_total_fight_data %>%
  filter(Fight_type == "Bantamweight Bout") %>%
  ggplot(aes(y = last_round)) + geom_bar() +labs(title = "Bantamweight")
```




```
raw_total_fight_data %>%
  filter(Fight_type == "Heavyweight Bout") %>%
  ggplot(aes(y = last_round)) + geom_bar() +labs(title = "Heavyweight")
```



3 Data Science Questions

Is there a correlation between height and submissions?

Is there a correlation between height and the weight class they compete in?

Can we predict the winner of a fight based on their fight statistics?