Exploratory Analysis of Professional MMA Data

Nicky Sanders

2023-04-29

Ask (1)

Exploring MMA Weight Classes and Fight Outcomes over time

- Topics to explore:
  - When did the rise in MMA begin? What were some galvanizing events for the sport?
  - What weight classes produce the most KO/TKO outcomes?
  - Is there any difference between the Male/Female divisions?
  - How many significant strikes can he/she expect to take before being KO'd again
  - Does time off between fights affect the chance of being KO'd again?
- What metrics will you use to measure your data to achieve your objective? Who are the stakeholders?
  - Time measured from the night of the KO to the night of the next KO.
  - Significant strikes since the night of the KO to the night of the next KO.
- Audience:
  - The analysis is for the fighters, coaches, fans, and leaders of Pro MMA organizations.
- How this data will help the fighters and coaches make decisions:
  - Fighters and coaches will be more confident when deciding to return to fighting again after being knocked out.
- Topic of exploration:
  - The evolution of fight outcomes as the sport progresses
  - The after-effects of MMA fighters being KO'd.

Prepare (2) and Process (3)

Data Source

- Data sourced from kaggle here
- Data comes from reliable source and is original, comprehensive, current, and cited.

Using Google BigQuery

- Original data contains 530 variables; With Google BigQuery and SQL script, we created a few engineered variables

```
## create temp table adding sig_strikes_absorbed and total_strikes_absorbed

WITH master_table AS (
  SELECT *,
  (SELECT sig_strikes_landed
  FROM `capstone-project-mma.mma_master.mma_fighter_data` t2
  WHERE t1.opponent = t2.fighter AND t1.fight_url = t2.fight_url) AS sig_strikes_absorbed,
  (SELECT total_strikes_landed
  FROM `capstone-project-mma.mma_master.mma_fighter_data` t2
  WHERE t1.opponent = t2.fighter AND t1.fight_url = t2.fight_url) AS total_strikes_absorbed
FROM `capstone-project-mma.mma_master.mma_fighter_data` t1
)

## create 'days_since_last_knockout'

SELECT *
FROM (
  SELECT
    mt.*,
    IF(mt.result = 1, 0, ## if result was a win (mt.result = 1),
                         ## days_since_last_knockout = 0 for that observation
      DATE_DIFF(mt.date, (SELECT MAX(date) FROM master_table t2
      WHERE t2.fighter = mt.fighter AND t2.result = 0 AND t2.date < mt.date),
                DAY) ## else (loss/draw) days_since_last_knockout =
                     ## difference between current date (outer-query's date) and
                     ## most recent KO/TKO loss (subquery's date)
    ) AS days_since_last_knockout
  FROM master_table mt
)
ORDER BY date, fight_url, fighter
```

Further organizing in R

- Organizing and trimming data
- Engineering additional fields to filter on later

```
# Adding column 'sex' to define whether fighters are male or female

mma_df <- mma_df %>%
  mutate(sex = ifelse(grepl("Women's", division), "Female", "Male"))

# Adding column 'weightclass' corresponding to max weight in division

mma_df = mma_df %>%
  mutate(weightclass = case_when(

    division == "Women's Strawweight" ~ 115,
    division == "Women's Flyweight" ~ 125,
    division == "Women's Bantamweight" ~ 135,
    division == "Women's Featherweight" ~ 145,
    division == "Flyweight" ~ 125,
    division == "Bantamweight" ~ 135,
    division == "Featherweight" ~ 145,
    division == "Lightweight" ~ 155,
```

```
    division == "Welterweight" ~ 170,
    division == "Middleweight" ~ 185,
    division == "Light Heavyweight" ~ 205,
    division == "Heavyweight" ~ 265,
    TRUE ~ 0
  ))
# `unique(mma_df$method)` returns all different outcomes of a fight:
#  ("KO/TKO" "SUB"    "DRAW"   "U-DEC"  "S-DEC"  "M-DEC"  "DQ")
# We combine all three types of decisions into one "DEC" category in 'method_new' variable

mma_df$method_new = ifelse(mma_df$method %in% c("U-DEC", "S-DEC", "M-DEC"), "DEC", mma_df$method)
```

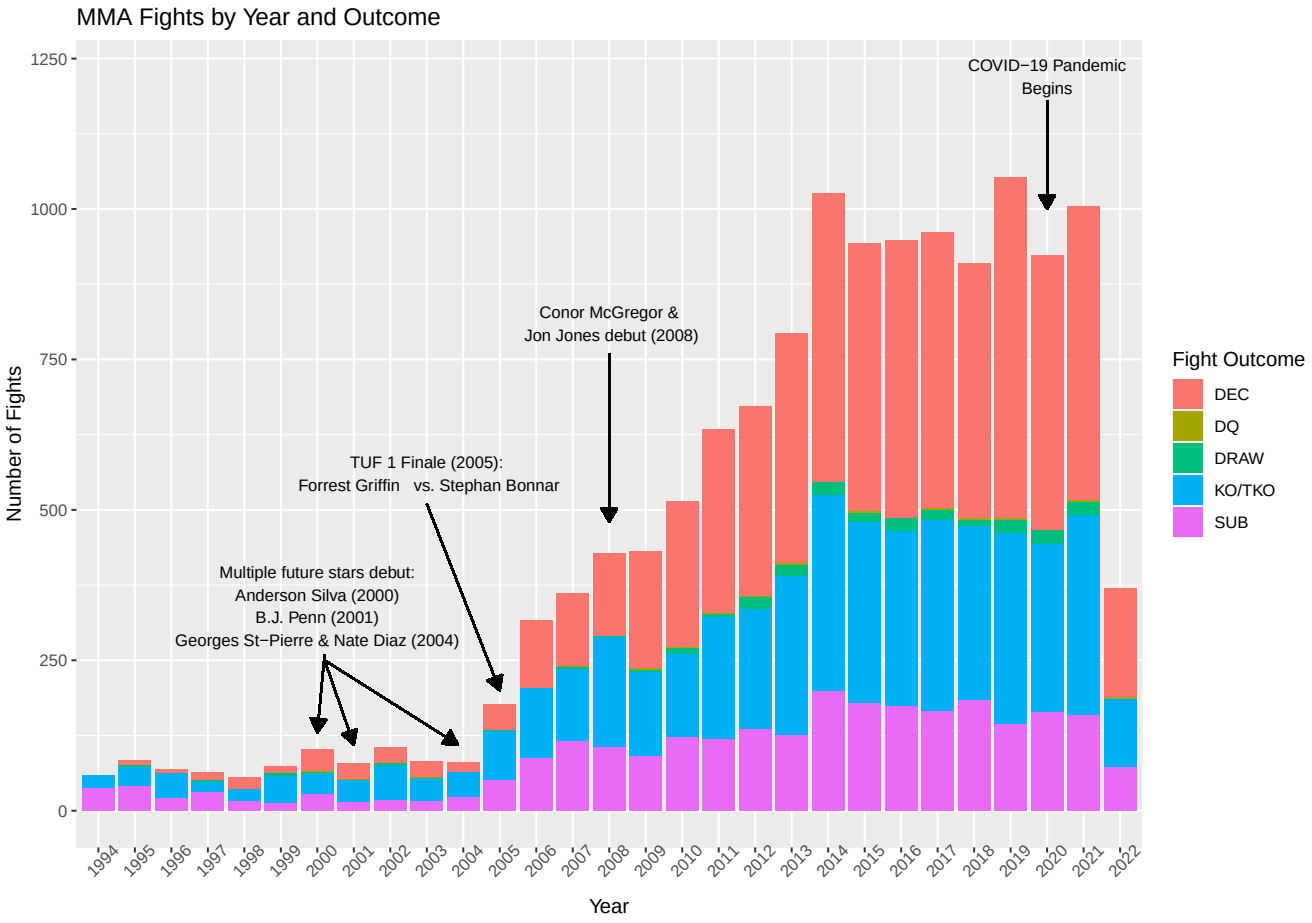
Adding a field describing the gender of the fighter will help us more easily sort the data between male and female fights. The 'division' column is helpful in conversation and when reading, but associating the divisions with their maximum allowable weight is an easy way to transform categorical data into its quantitative equivalent and allow us to do numeric-specific analysis. We'll group the different types of "decision" fight outcomes into one category ('DEC') to make the data easier to work with and understand as we do not need that category to be so granular.

Analyze (4) and Share (5)

Why the surge in fights in the 2000s?

Notable events 2000-2005:

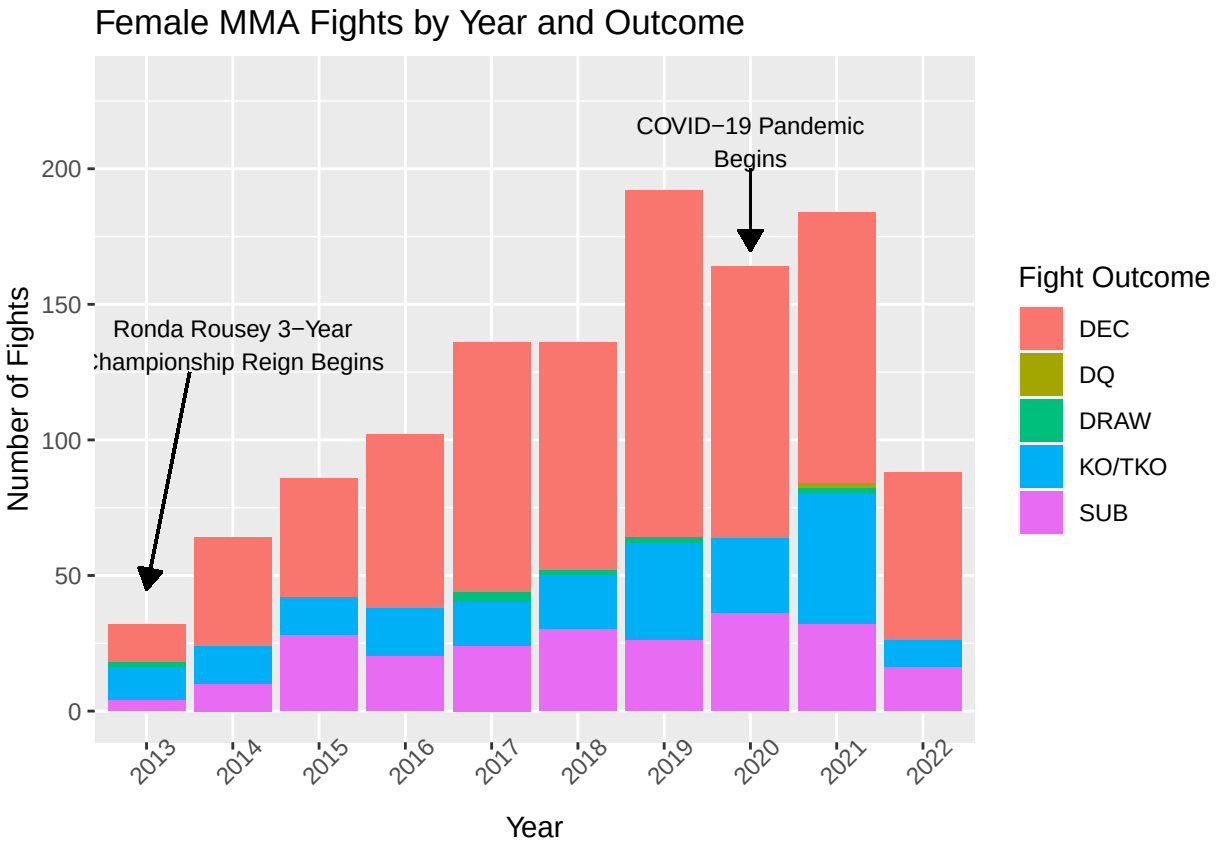
- Fighters debuting in 2000:
  - Anderson Silva
  - Tito Ortiz
  - Fabricio Werdum
  - Wanderlei Silva
- Fighters debuting in 2001:
  - B.J. Penn
  - Matt Serra
- Fighters debuting in 2002:
  - Robbie Lawler
- Fighters debuting in 2003:
  - Nick Diaz
  - Matt Serra
- Fighters debuting in 2004:
  - Georges St-Pierre
  - Michael Bisping
  - Nate Diaz
- Historical TUF 1 Finale – Forrest Griffin vs. Stephan Bonnar (2005)
- Fighters debuting in 2005:
  - Forrest Griffin
  - Stephan Bonnar
  - Chael Sonnen
  - Rashad Evans



With so many future stars debuting from 2000-2008, one can see how the positive trajectory of their careers in the years following coincides with the growing popularity of the sport.

The year 2008 had two of arguably the most popular fighters the sport has ever seen in Conor McGregor and Jon Jones (The GOAT?). As their careers took off en route to multiple titles and title defenses, we can see the number of fights continue to rise steeply. Surely the number of fights will approach a natural limit, right?

Enter: Professional Female MMA.

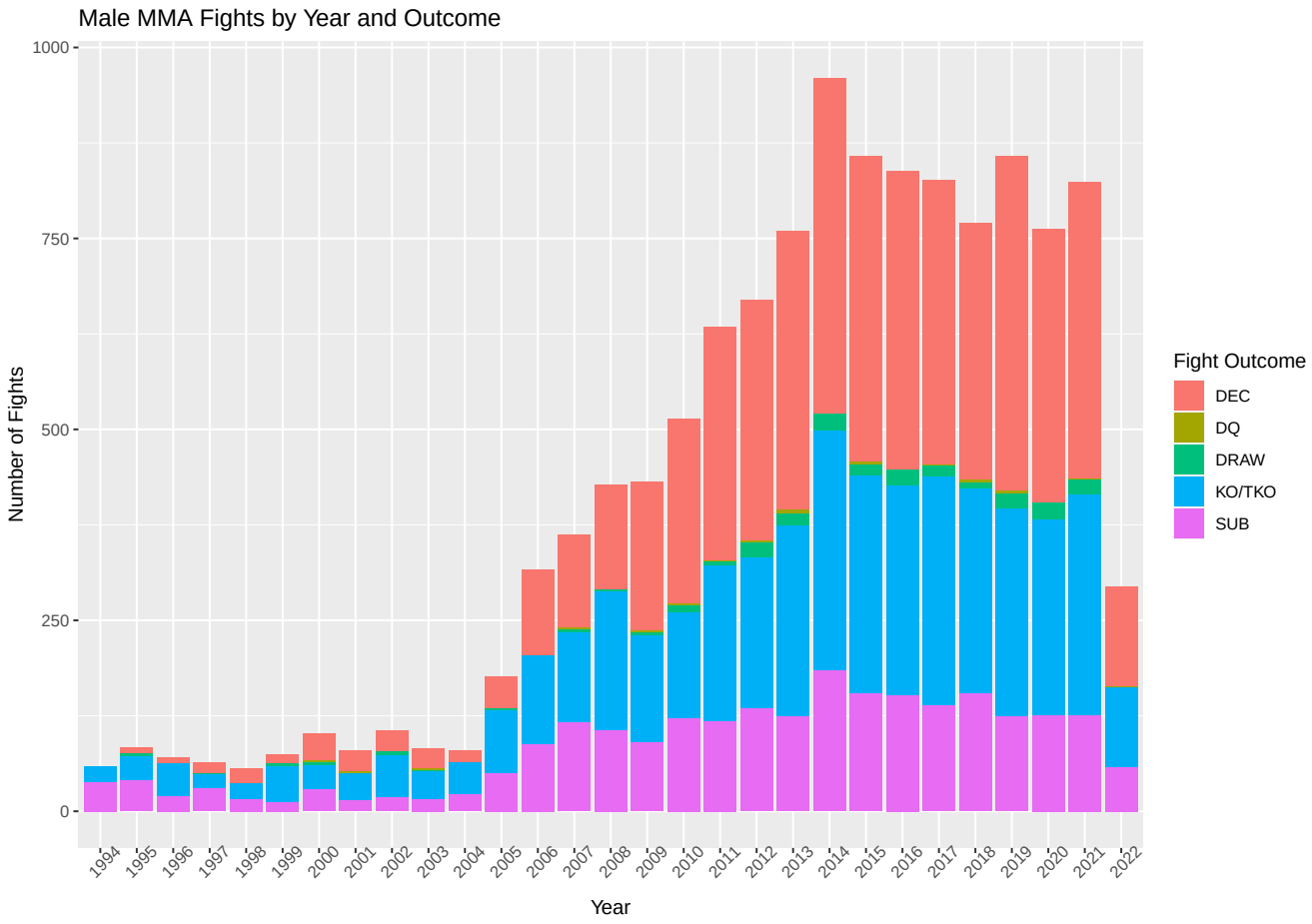


We can see that Women's MMA started in 2013. The inclusion of women in the sport increased the population of fighters and fans alike, demanding more of the sport and its events.

While there are multiple professional MMA leagues, they put on events in similar fashion with multiple fights per event. Each event is comprised of different "cards" with several fights on each of them e.g. under card, preliminary card, main card. The UFC has consistently held more events than any other organization, peaking in 2019 with a staggering 45 events. While there is essentially no off-season in mixed martial arts, there is an assumed hard ceiling of 52 events annually (one per week) and a limited number of weight classes with fighters within which fighters fight. Another constraint limiting the number of annual MMA fights is the existence of other events in sports (e.g. the Superbowl, the NBA Finals) that MMA might not want to compete with to avoid diluted viewership.

Men's MMA Analysis

Fight Outcomes and Their Relationship with Divisions



The above chart shows a consistent proportion of fight outcomes over the years, but does the population's fight outcome reflect that of each division? Let's take a closer look at what divisions we have first.

```
# Super Heavyweight, Catch Weight, and Open Weight divisions represent a small
# portion of male MMA fights

mma_male_df = mma_df %>% filter(sex == "Male")
division_df = data.frame(Proportion = c(round(prop.table(table(mma_male_df$division)) * 100, 2)))
division_df %>%
  arrange(division_df$Proportion) %>%
  kable("pipe") %>%
  kable_styling(full_width = FALSE, font_size = 12)
```

Proportion of all male fights by weightclass (division)

	Proportion
Super Heavyweight	0.02
Catch Weight	0.74

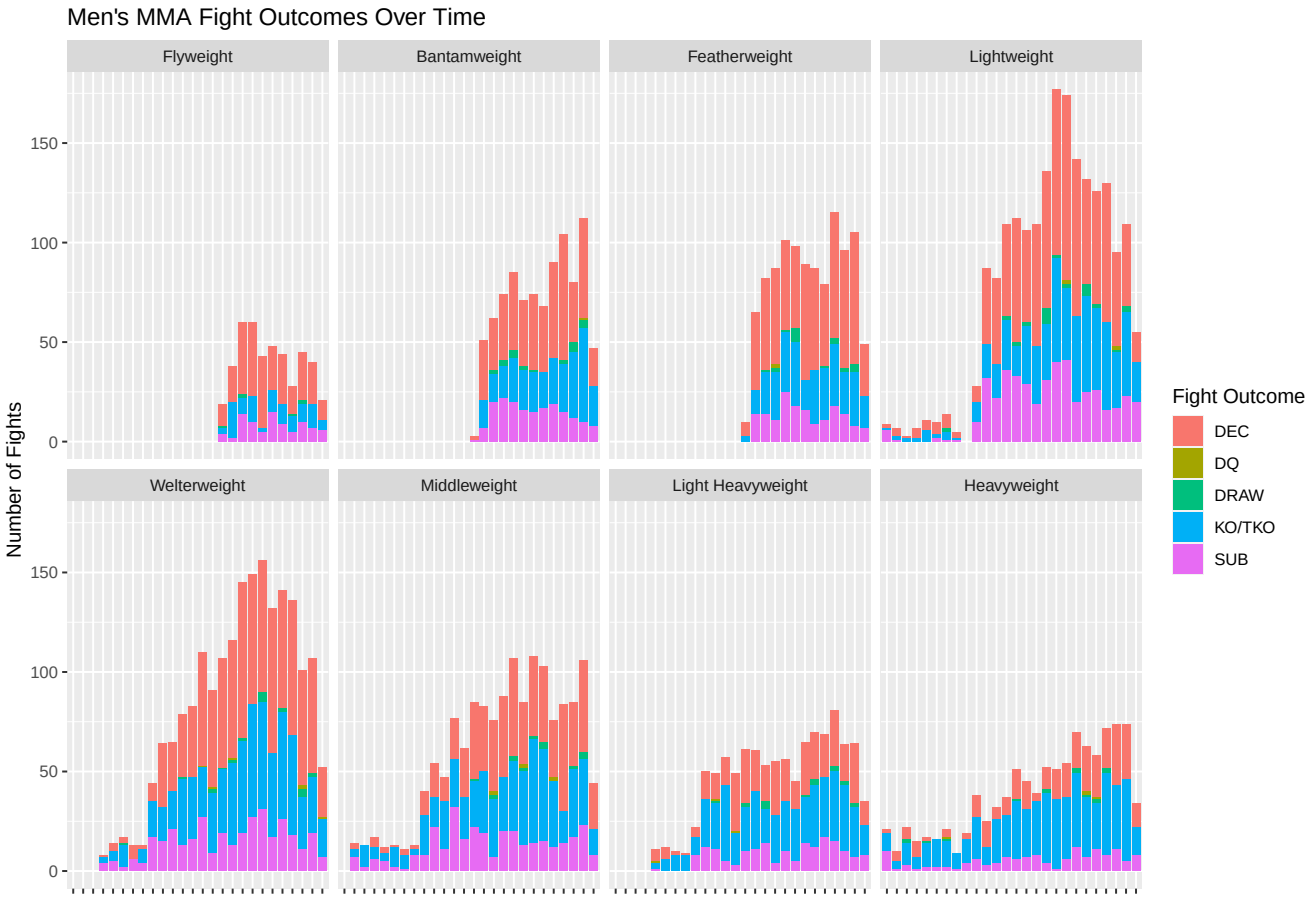
	Proportion
Open Weight	1.68
Flyweight	4.32
Bantamweight	9.05
Light Heavyweight	10.07
Featherweight	10.31
Heavyweight	10.43
Middleweight	14.78
Welterweight	19.15
Lightweight	19.46

A ***'catch weight'*** is used when two fighters agree to fight at a nonstandard weight for a variety of reasons, such as one fighter being unable to make weight or both fighters agreeing to meet at a mutually agreed-upon weight that falls between two established weight classes. This happens in $< 0.74\%$ of Male MMA data.

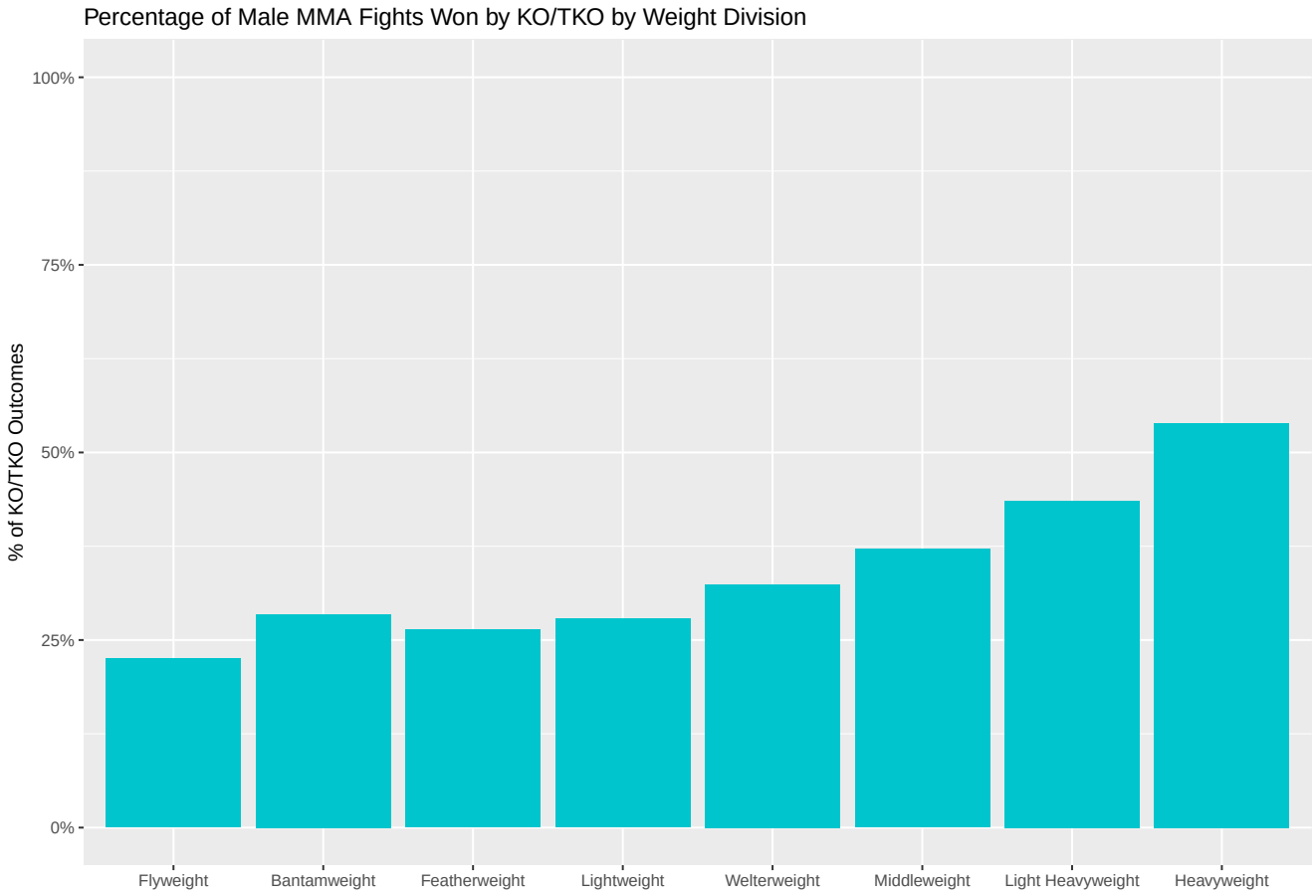
Since the early 2000s, the ***Super Heavyweight*** and ***Open Weight*** divisions have not been recognized by most of the professional MMA organizations. In an Open Weight fight, fighters can have completely different weights, confusing our data. We'll filter this out from our visual analysis.

Since the Super Heavyweight division is similarly no longer active and only represents a mere 0.02% of our Male MMA data, we'll also exclude this rarity for this next visual analysis.

```
# Arranging the divisions by ascending order, with weight increasing

mma_male_df$division = factor(mma_male_df$division,
                              levels = c("Flyweight", "Bantamweight", "Featherweight",
                                         "Lightweight", "Welterweight", "Middleweight",
                                         "Light Heavyweight", "Heavyweight"))


# Plot of Men's MMA Fights by Year and Outcome, excluding three (3) aforementioned divisions

ggplot(data = na.omit(mma_male_df)) +
  geom_bar(mapping = aes(x = cut(date, "12 months"), fill=method_new)) +
  facet_wrap(~division, ncol = 4, drop = T) +
  scale_x_discrete(labels = function(x) format(as.Date(x), "%Y")) +
  theme(axis.text.x = element_blank()) +
  labs(title = "Men's MMA Fight Outcomes Over Time", x = "",
       y = "Number of Fights", fill = "Fight Outcome")
```



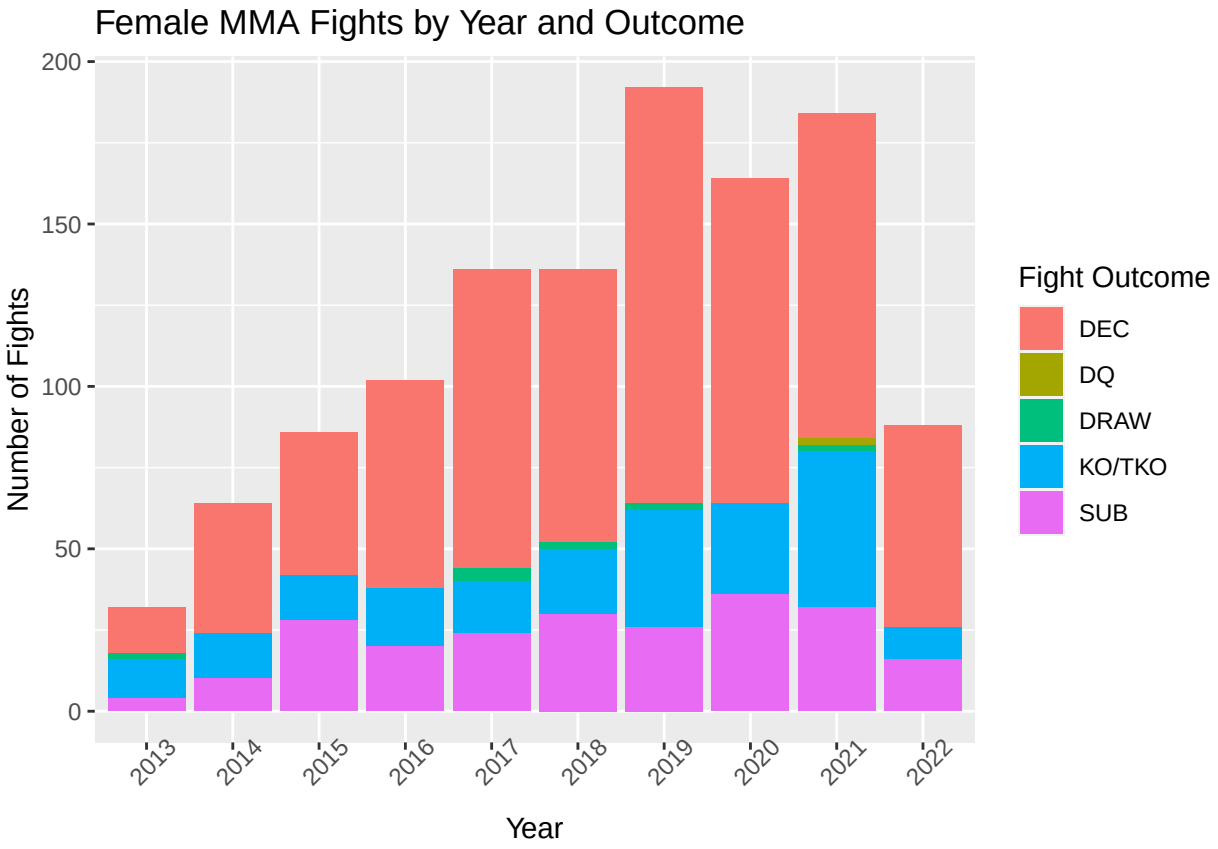
Let's take a look at the average KO/TKO result in each men's division, plotted in ascending weight-class order



The percentage of fights ending by KO/TKO increases as the divisions get heavier with the Heavyweight division KO/TKO outcome doubling that of the Flyweight division's. More weight = more power? Perhaps. Let's take a look at the women's divisions next.

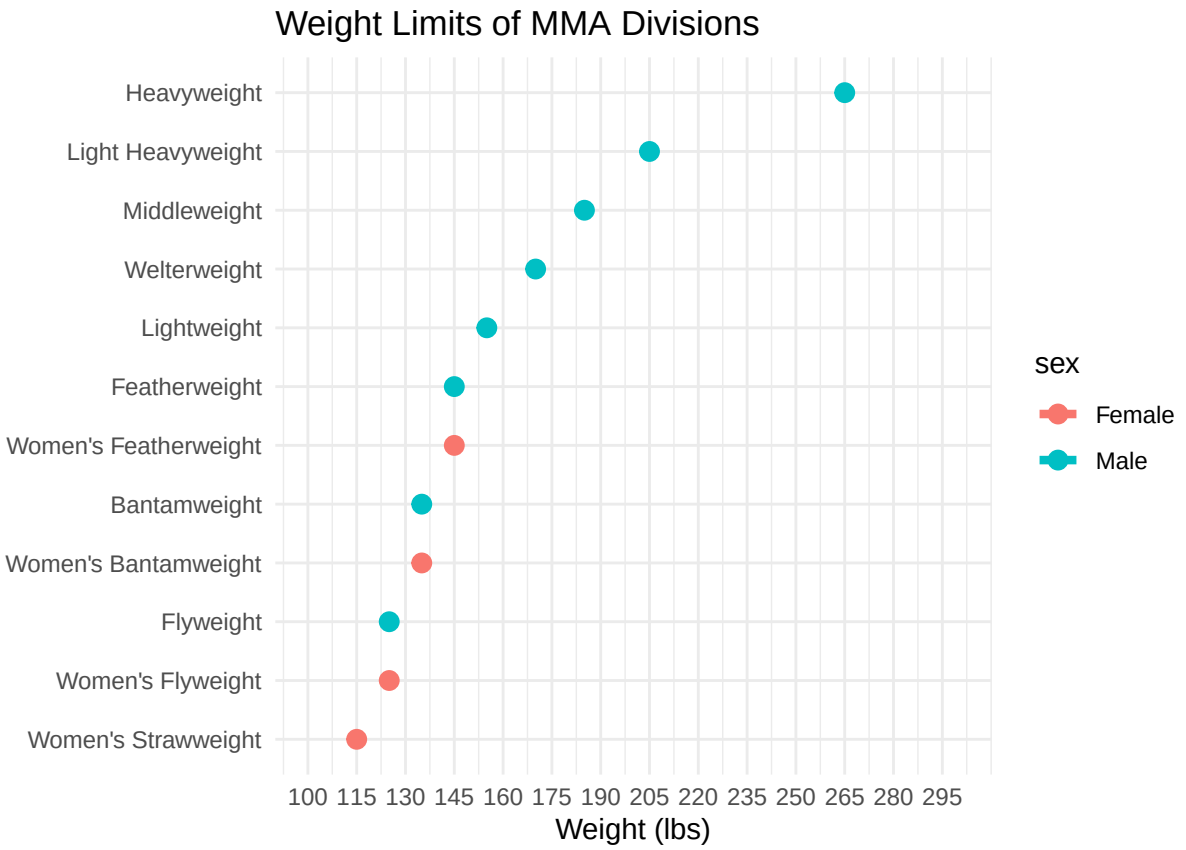
Women's MMA Analysis

Fight Outcomes and Their Relationship with Divisions



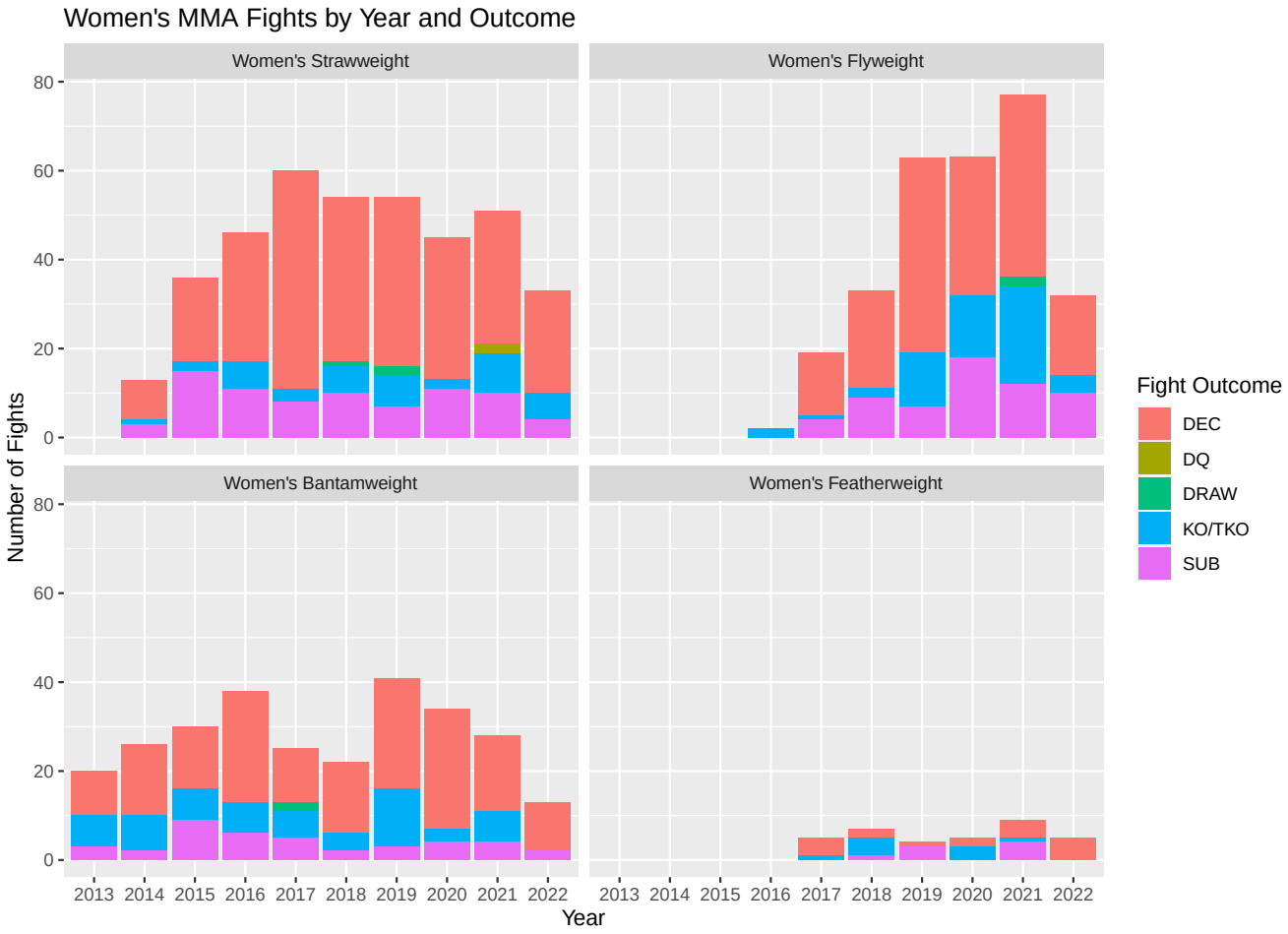
Compared to the earlier men's plot above, the female fight outcomes have fewer KO/TKO finishes proportionally. Female weightclasses also mirror the lower range of the male weightclasses.

All weightclasses in professional MMA



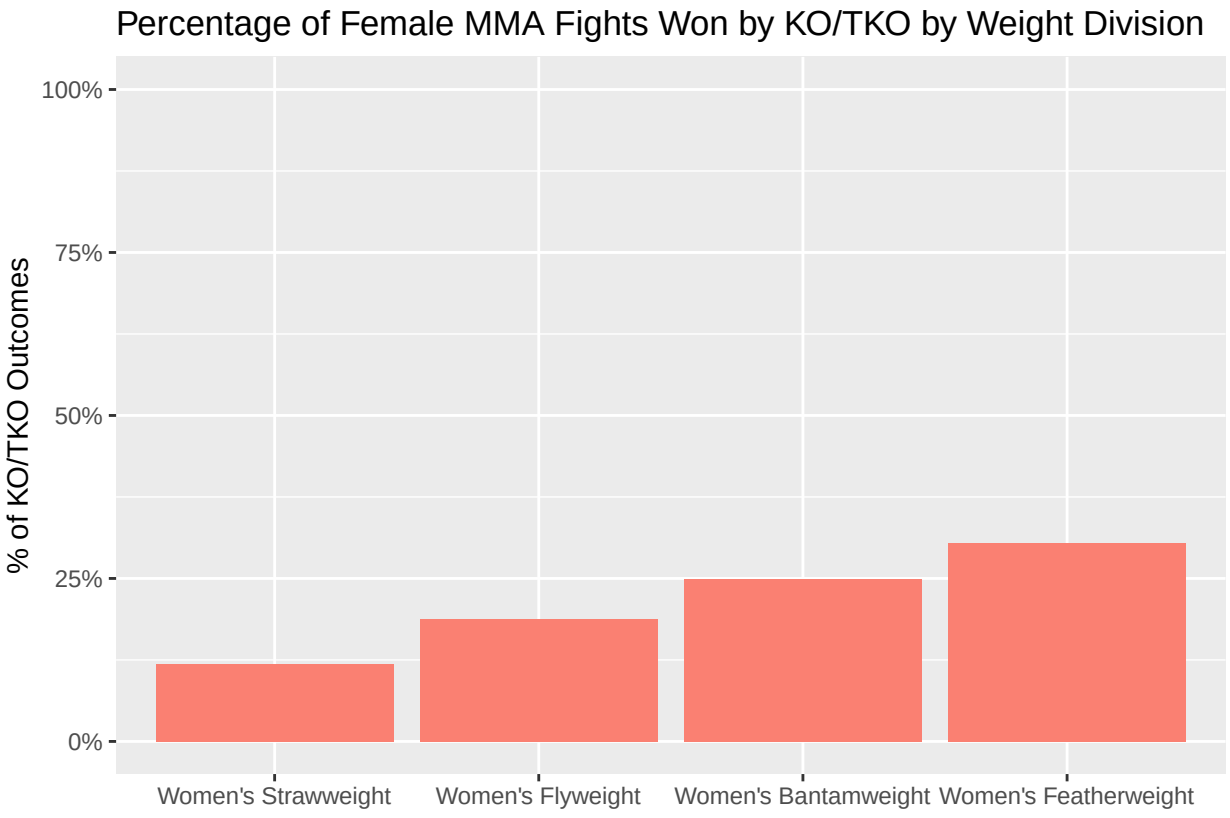
Note the four (4) female MMA divisions on the lower end of the x-axis. Female MMA divisions only share a weight limit with the three (3) lightest male weightclasses. One (1) female weight class (Women's Strawweight) doesn't even have a male counterpart. Given that the entire female fighting population overlaps with the lightest male divisions, one might expect fewer KO/TKO finishes in women's divisions if also assuming more weight is associated with more knockout power.

Different women's results by weightclass (division)



The plots are not as cut and dry as we'd like, especially with the women's featherweight division having so few observations. However, as a proportion of total fights, there does seem to be moderate correlation with KO/TKO finishes and weightclass. The lightest weightclass - women's strawweight - has the smallest proportion of KO/TKO finishes relative to total fights.

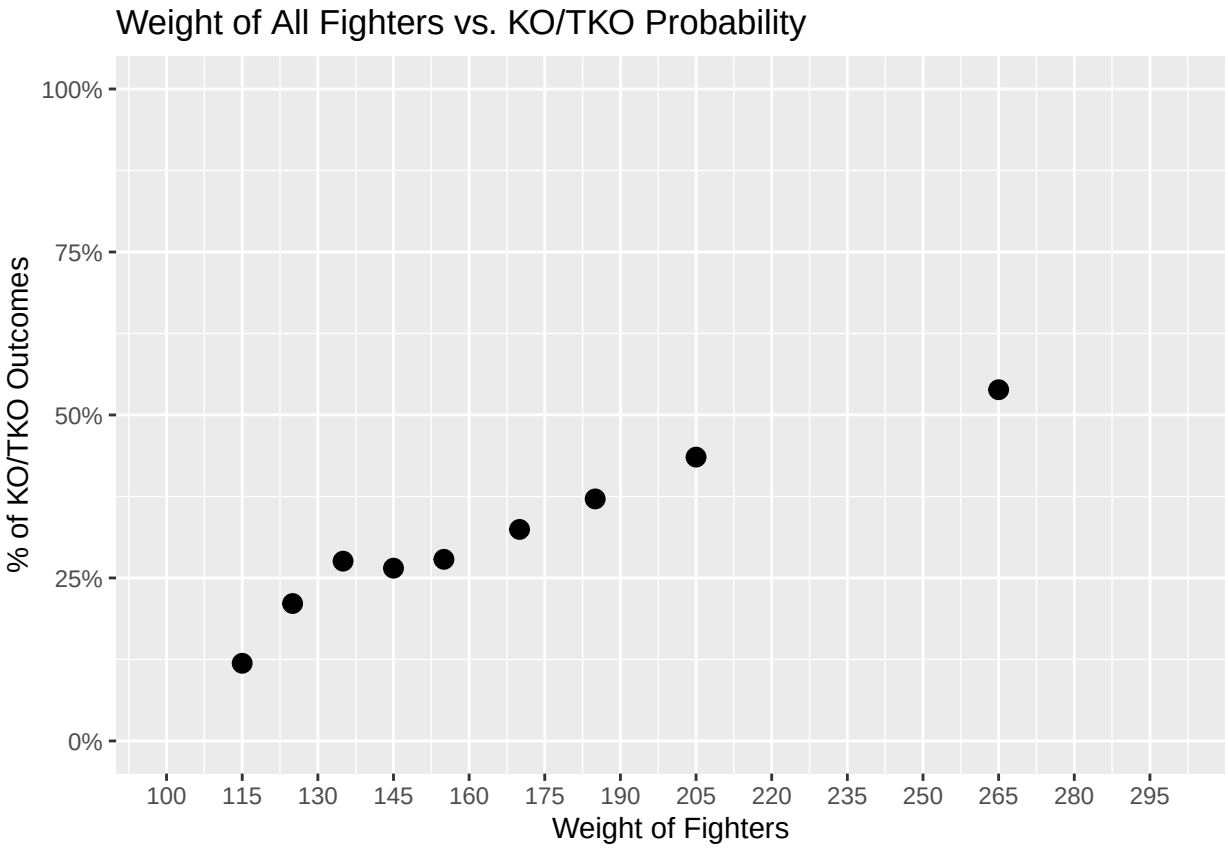
Average KO/TKO result in each women's division, plotted in ascending weight



Plotting the percentage of Female MMA Fights Won by KO/TKO by weight division shows this moderately positive correlation more easily.

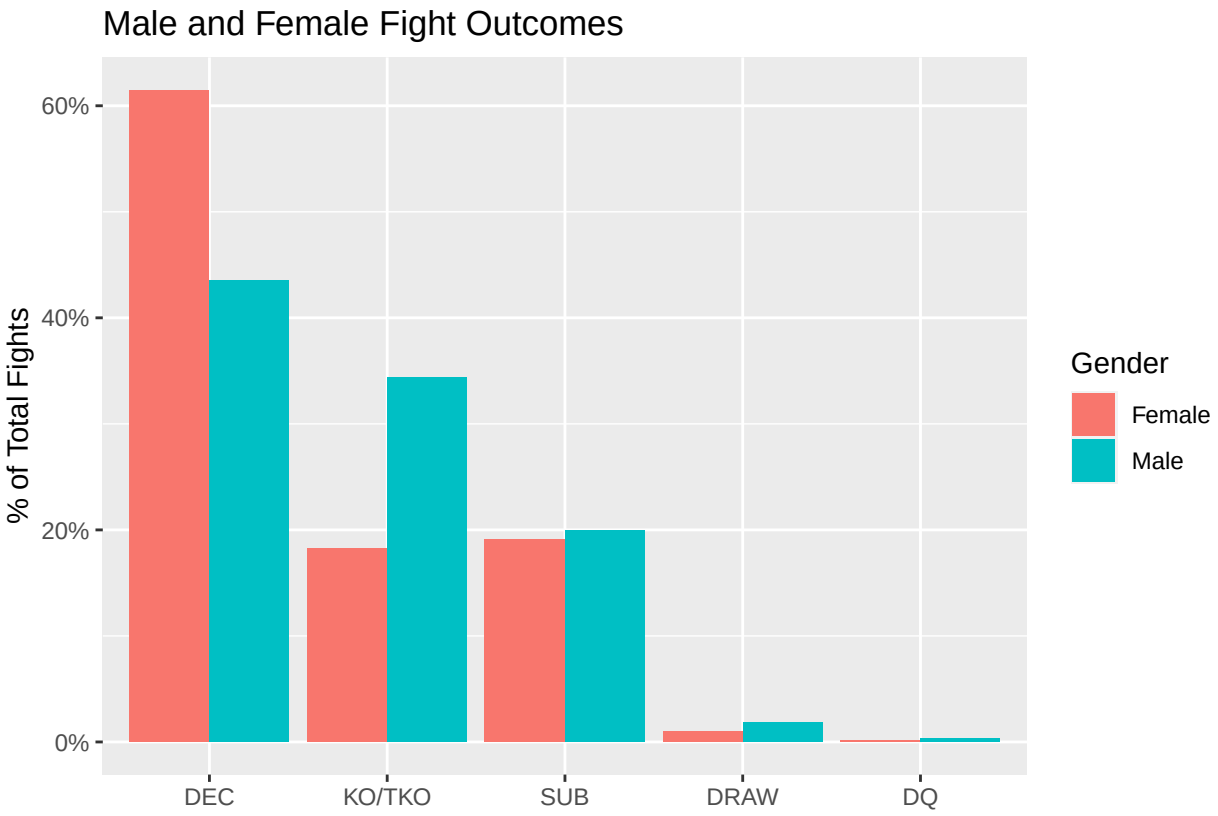
Male and Female data

Looking at correlation between weight of the fighters and likelihood of a knockout



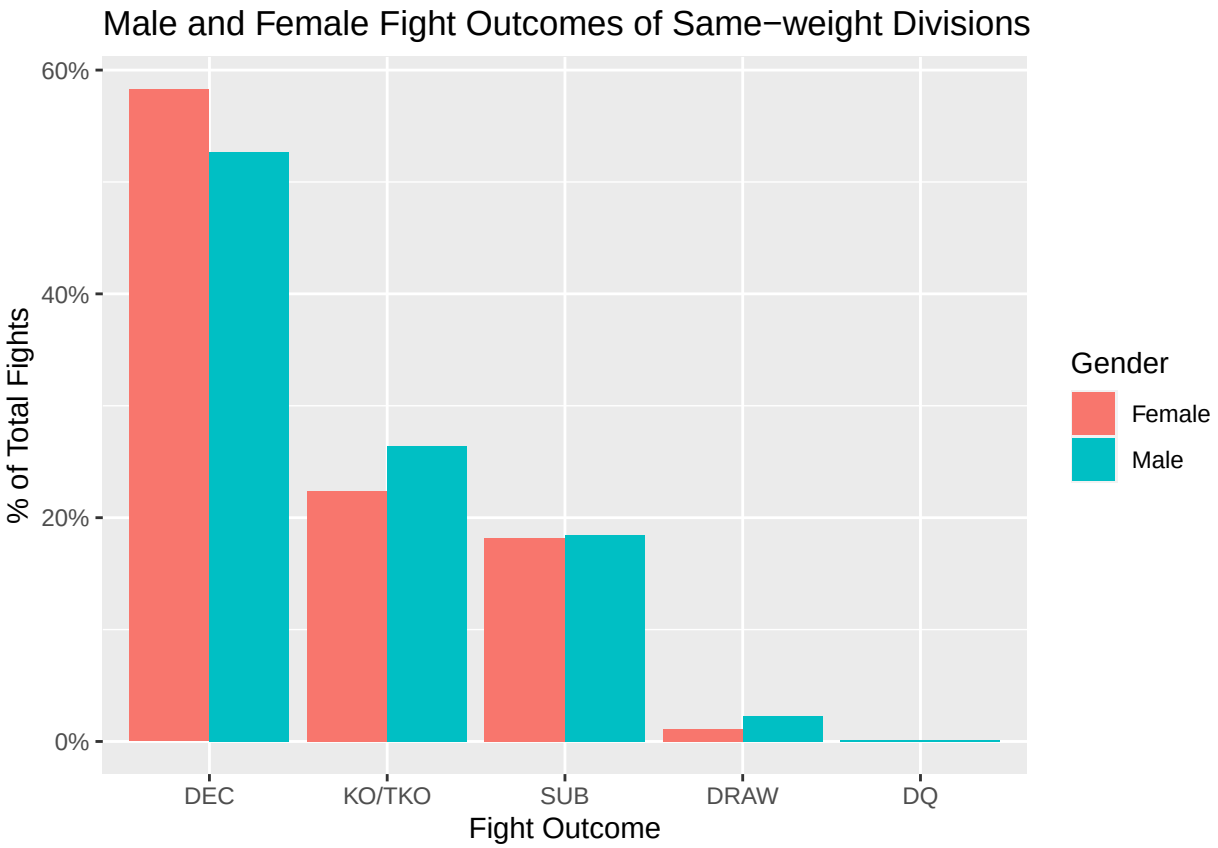
Combining the male and female data and plotting the weight of fighters vs. the percentage of KO/TKO Outcomes supports our initial thoughts of positive correlation. Let's see how men and women compare.

Comparing male and female fight outcomes side-by-side



The chart above does a great job calling attention to the differences in fight outcomes between male and female fights. Submissions, Draws, and Disqualifications are all relatively similar in likelihood between the genders. The data show almost 20% more fights end in Decision among female fighters than male fighters. One can see that entire 20% delta is almost completely accounted for when looking at fights ending by KO/TKO.

Let's not forget our earlier graphs showing a positive relationship between the weight of a fighter and the probability that the fight ends in a KO/TKO. There are no female fighters above 145lbs, nor are there male fighters below 125lbs. Let's plot the data without the Women's Strawweight division (115lbs) and also remove all Men's weight classes above 145lbs.



Looking only at the weight classes whose weight limit is shared between genders yields more similar distribution of fight outcomes. The gap between male and female fighters whose fights end by decision has shrunk in lockstep with the fights ending by KO/TKO. Male fight outcome distribution is much more closely mirrored by female fight outcomes when controlling for the weight of the fighters.

Cramer's V Correlation and PCA

```
# compute Cramer's V correlation coefficient

mma_filtered_df = mma_df %>%  filter(weightclass != 0)
assocstats(table(mma_filtered_df$division, mma_filtered_df$method_new))
```

Correlation, Association, or Causation?

```
##                     X^2 df P(> X^2)
## Likelihood Ratio 693.59 44        0
## Pearson          687.04 44        0
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.224
## Cramer's V        : 0.115
```

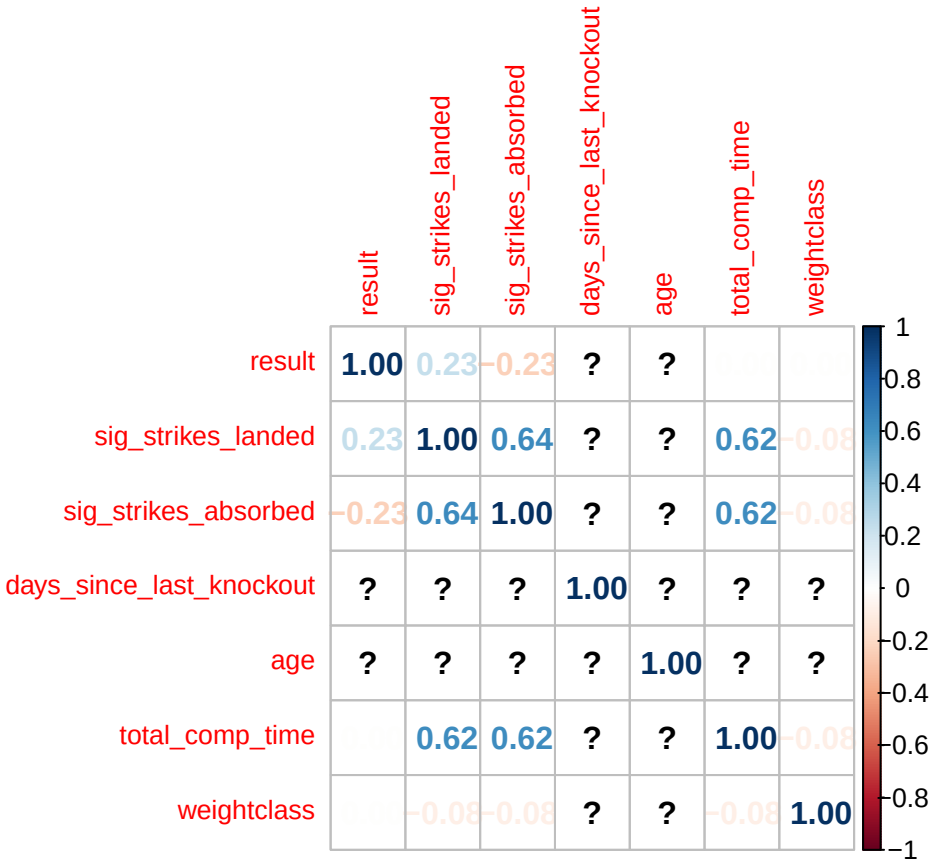
Observations:

1) The results indicate that there is a statistically significant association between the categorical variables 'division' and 'method_new'. The likelihood ratio test and Pearson test both yield large chi-square values (693.59 and 687.04, respectively), with 44 degrees of freedom (DoF), and a p-value of less than 0.001, indicating that the association between the two variables is not likely to be due to chance (DoF determine p-value of the test).
2) The contingency coefficient of 0.224 suggests a moderate association between the two variables, and the Cramer's V value of 0.115 suggests a small effect size. Overall, these results suggest that the weight division a fighter belongs to might have somewhat of an influence on the method of victory they employ.
3) The Pearson chi-square test is used to test the null hypothesis: there is no association between the two variables. It is a measure of the difference between the observed and expected frequencies in the contingency table.
4) The results show the association between the two variables, while moderate, is likely not due to chance.

Test Assumptions of Principal Component Analysis (PCA)

It is recommended to drop highly correlated columns before running PCA in order to limit the degrees of freedom. Plotting a correlation matrix will assist in this effort.

```
num_cols = sapply(mma_df, is.numeric)
mma_numeric = mma_df[, num_cols]
other_cols = sapply(mma_df, function(x) !is.numeric(x))
other_cols = names(mma_df)[other_cols]
num_cols = colnames(mma_numeric)

mma_df_subset = mma_df[c(num_cols, other_cols)]

correlations = cor(mma_numeric)
# corrplot(correlations, type="full", method="number", tl.cex=0.7) #hiding initial correlation plot


num_cols_drop = c("sig_strikes_attempts", "total_strikes_attempts", "total_strikes_absorbed", "total_strik
mma_numeric = mma_numeric %>% select(-one_of(num_cols_drop))
new_corr = cor(mma_numeric)
corrplot(new_corr, type="full", method="number", tl.cex=0.8)
```



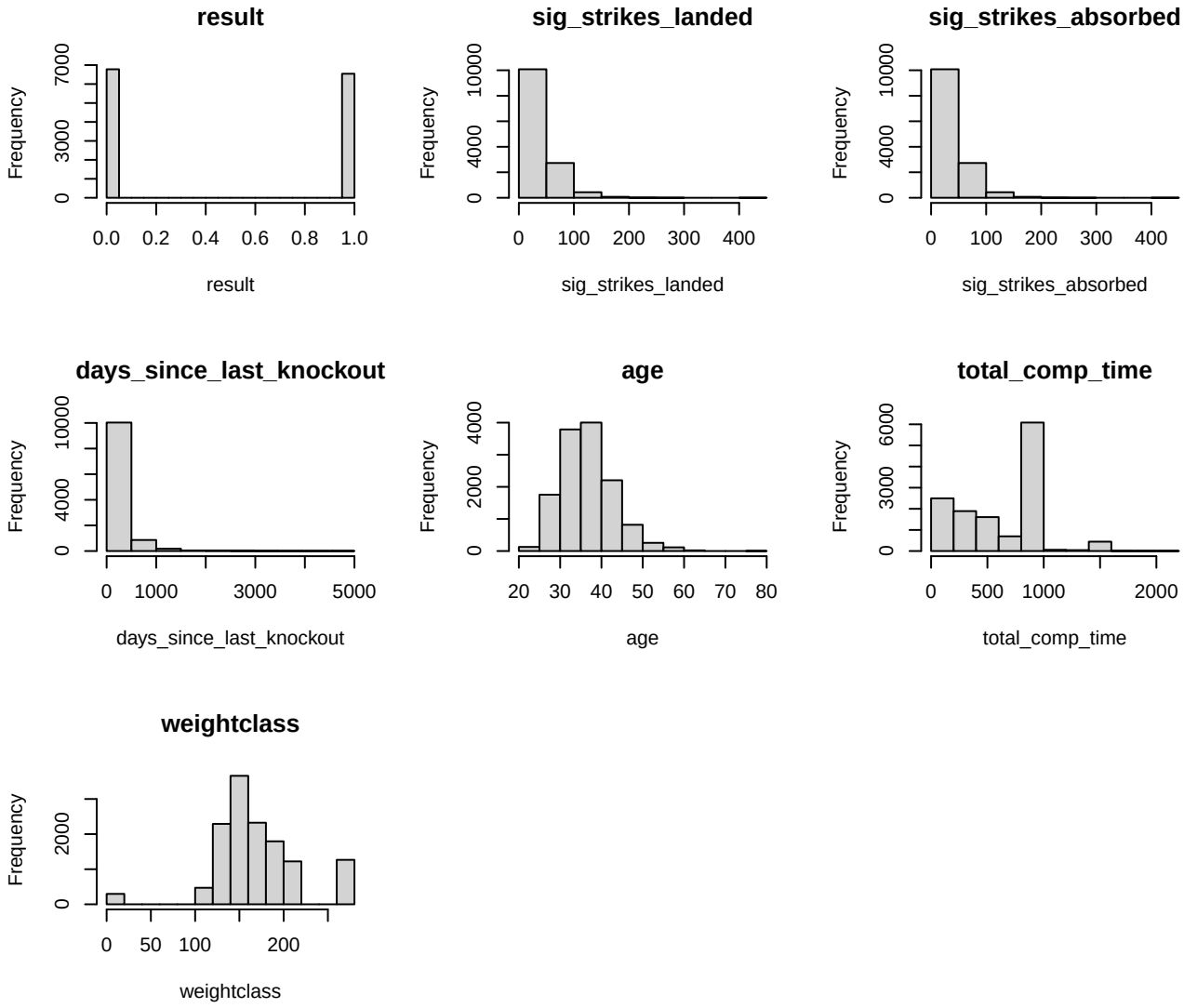
Variables are considered highly correlated if they have a correlation coefficient > 0.8. The results of the initial correlation plot showed several pairs of variables with this high of a coefficient. After exercising knowledge and context of the data, the following variables are dropped: 'sig_strikes_attempts', 'total_strikes_attempts', 'total_strikes_absorbed', 'total_strikes_landed', and 'round'.

Test for Normality with Anderson-Darling test

```
ad_results = NULL
for (col in names(mma_numeric)) {
  ad_results = cbind(ad_results, ad.test(mma_numeric[[col]])$p.value)
}
colnames(ad_results) = names(mma_numeric)

# View the p-values
datatable(ad_results)
```

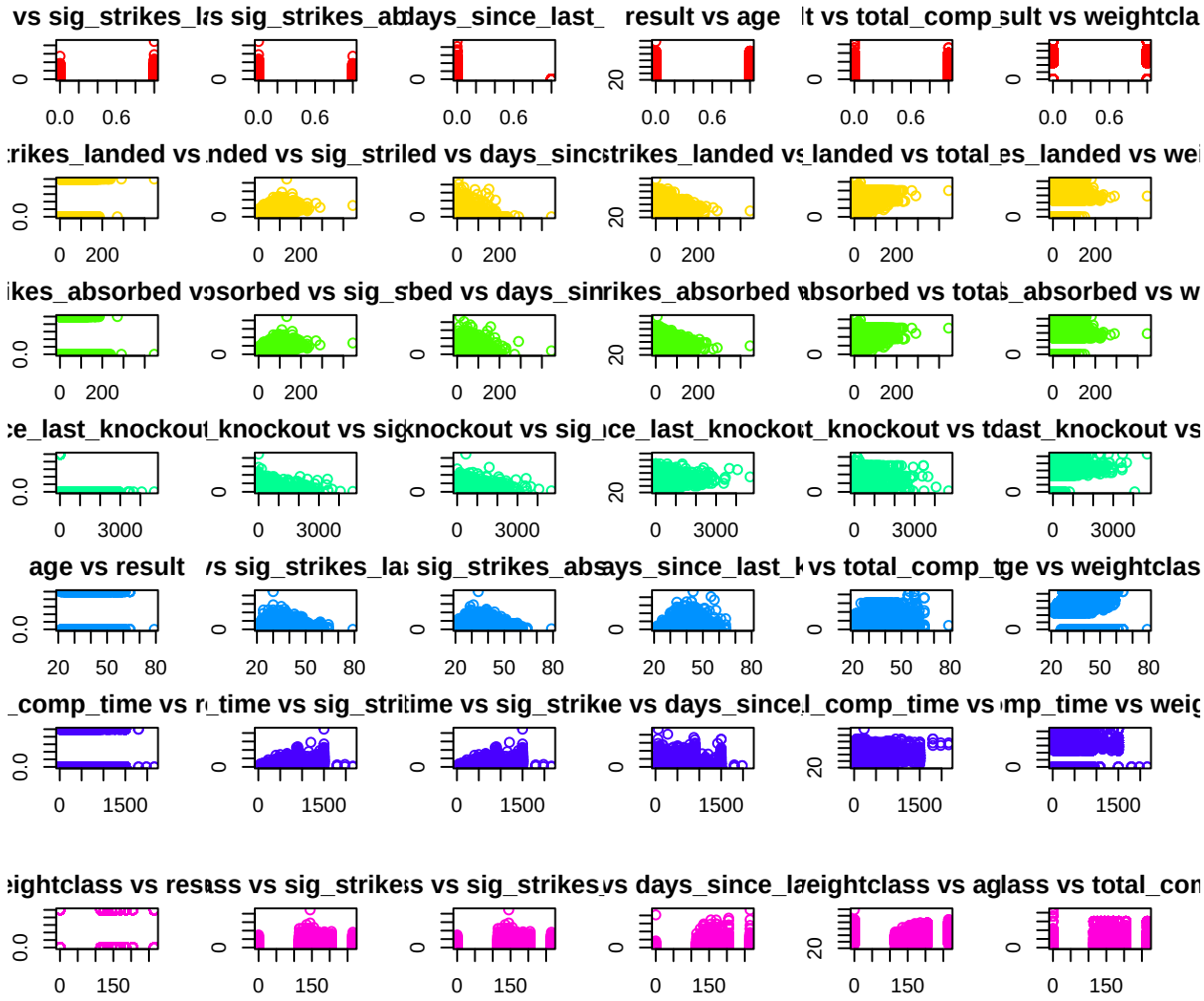
All calculated p-values are less than significance values (.05), thus we reject null hypothesis that data is normally distributed.

Visual evidence that data is not normally distributed with histograms




```
par(mfrow = c(6, 6), mar=c(2,2,2,2))
colors <- rainbow(length(num_cols)) # generate a vector of colors

for (i in 1:length(num_cols)) {
  for (j in 1:length(num_cols)) {
    if (i != j) {
      col <- colors[i] # assign color to the variable
      plot(mma_df_subset[[num_cols[i]]], mma_df_subset[[num_cols[j]]],
           main = paste(num_cols[i], "vs", num_cols[j]),
           xlab = num_cols[i], ylab = num_cols[j],
           col = col) # use the color for the markers
    }
  }
}
```

Test for Linearity with Scatterplots

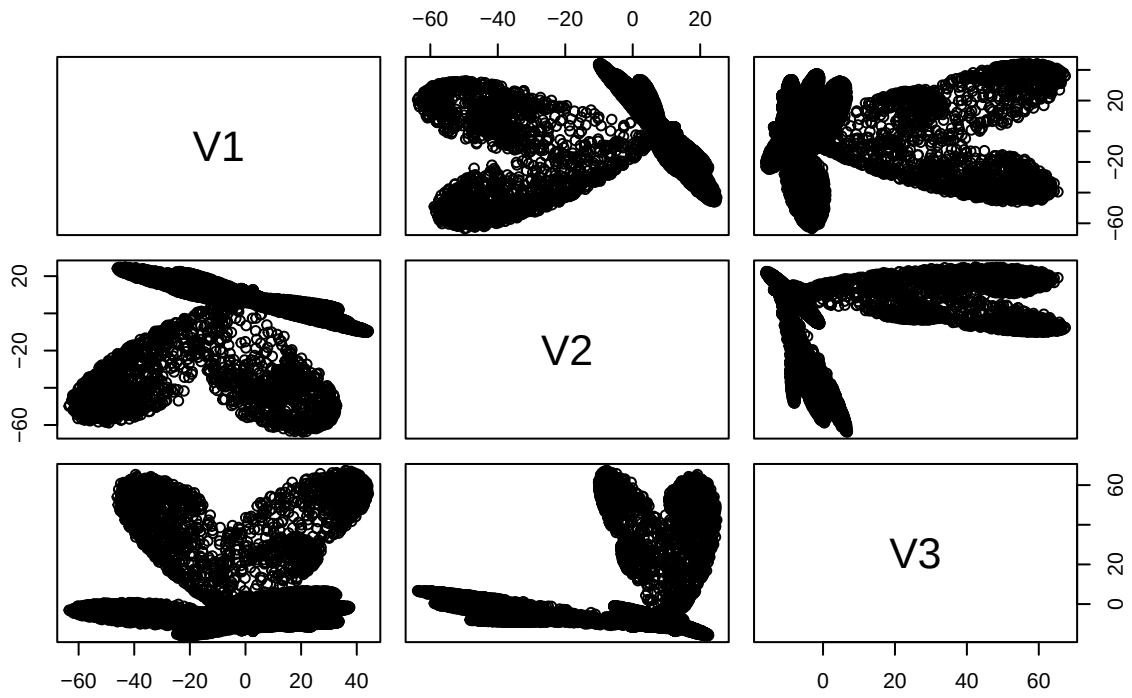

vs sig_strikes_l
s sig_strikes_ab
days_since_last_
result vs age
t vs total_comp_
sult vs weightcla
rikes_landed vs
nded vs sig_striled vs days_sinc
strikes_landed vs
landed vs total_
es_landed vs wei
ikes_absorbed vs
sorbed vs sig_sbed vs days_sin
rikes_absorbed v
bsorbed vs totas_absorbed vs w
e_last_knockou
knockout vs sig
knockout vs sig_
ce_last_knockou
t_knockout vs to
ast_knockout vs
age vs result
s sig_strikes_la
sig_strikes_abs
ays_since_last_k
vs total_comp_t
ge vs weightclas
_comp_time vs r
time vs sig_strik
time vs sig_strike
e vs days_since
l_comp_time vs
mp_time vs weig
eightclass vs res
ass vs sig_strikes
s vs sig_strikes_
vs days_since_la
eightclass vs ag
lass vs total_con

Data is neither normally distributed nor linear For good measure, we show visual evidence of the Anderson-Darling test proving that data is NOT normally distributed with histograms of each numeric variable. Additionally, the scatterplots show the data is not linear. We can conclude that Principle Component Analysis is not appropriate with this data. We'll move on and explore Nonlinear (Kernel) PCA.

Kernel PCA

```
drop_these_cols = c("result", "fight_url", "result", "fighter", "method", "date",
                    "opponent", "round", "time", "weightclass")

mma_df_subset = mma_df_subset %>%
  filter(!division %in% c("Open Weight", "Super Heavyweight", "Catch Weight"))

mma_df_test = mma_df_subset %>% select(-one_of(drop_these_cols))

# select categorical columns for one-hot encoding
cat_cols = c("division", "sex", "method_new")

# one-hot encode categorical columns
mma_df_test = fastDummies::dummy_cols(mma_df_test, select_columns = cat_cols)

# apply kernel PCA on numeric columns
mma_numeric = mma_df_test %>%
  select_if(is.numeric) %>%
  scale()

set.seed(123)
kpca_fit = kernlab::kpca(mma_numeric, kernel = "rbfdot", kpar = list(sigma = 0.1), features = 3)
kpca_scores = as.data.frame(predict(kpca_fit, mma_numeric))
```

```
# Visualize Kernal PCA scores with plot of principal components against each other
pairs(kpca_scores, main = "Kernel PCA")
```



Observations:

1) The plot does not show the data to be a good candidate for KPCA

Further exploration:

- What could explain this general trend of heavier weightclasses seeing more KOs/TKOs proportional to the number of fights?
  - Heavier fighters » more muscle mass » more power » greater chance of knockout?
  - Are number of sig_strikes absorbed correlated with fight outcome?
  - Are number of strikes thrown correlated with weight class? Fight outcome?
- Deeper dive into fighter performance over time
  - Exploring age of fighters and their success rate
  - Exploring total strikes absorbed over time vs. win %
- Other things to consider:
  - As fighters increase in weightclass, how much does their ability to absorb strikes without being KO'd/TKO'd change?
  - Does this trend hold true with female fighters?
  - Is number of previous fights correlated with probability of KO/TKO?
  - Is number of previous KOs/TKOs correlated with probability of KO/TKO?
  - Is number of days since last KO/TKO correlated with probability of KO/TKO?

Closing Thoughts (6)

Guiding questions

Based on the analysis, we can conclude that the sport of MMA has grown substantially since its professional inception in 1994. Some possible explanations include incredible talent debuting in the early 2000s who would go on to dazzle soon-to-be MMA fighters and fans alike. With MMA being a men's sport originally, it realized its organic growth potential with the addition of women's divsions in the 2010s. Coupled with amplifying social media outlets, sports stars have never before been more known by the fans. Where the days of the pre-social media era saw a couple of superstars, this post-social media era has likely increased the reach of all stars in a sport. This increased reach - and the existence of the star talent in the sport - could explain the rise of MMA we have seen.

While examining the different outcomes a fight can have, we noticed and explored the variance in the proportion of KO/TKO finishes across both weightclass and gender. The data showed a somewhat positive correlation between weight of the fighter and likelihood of a KO/TKO finish. Keeping that in mind, we compared Male and Female fight outcomes, which showed Submissions, Draws, and Disqualifications all to have a similar probability of occurring. Decision and KO/TKO were another story. The data show almost 20% more fights end in Decision among female fighters than male fighters, with that delta being traded off in KOs/TKOs. Remembering the somewhat positive correlation between a fighter's weight and likelihood of a KO/TKO, we controlled for weight and plotted only the weightclasses whose weight limits were shared between men and women. This yielded a much more similar distribution of fight outcomes between the two genders.

Using a Pearson test we found a statistically significant association between a fighter's division and the fight's outcome. Large chi-square values and a p-value of less than 0.001 indicated that the association between the two variables is not likely to be due to chance. However, Cramer's V correlation coefficient of 0.115 suggests a small effect size. To conclude, the association between the the weight division a fighter belongs to and the method of victory they employ is moderate. However, it's highly unlikely that it's not due to chance.

When testing the assumptions necessary for a Principal Component Analysis, the data proved not to be normally distributed or linear. We then took a nonlinear approach, running a Kernel PCA analysis.

What additional data could be used to expand on our findings?

- The financials of different MMA (UFC) events would help in exploring correlation between revenue earned and fights on the card. This could aid in exploring what fight outcomes are most exciting and likely to generate more revenue over time or perhaps what weightclasses are the most popular and drive the most pay per views.