Oisin Goddard

COMP3021-20290834-CWK1

Contents

Description

- The description with the initial question

The 'Ultimate UFC Dataset' dataset used in this study is a record of UFC fights between 21st March 2010 and 21st August 2021. It contains 4891 rows with 122 columns including information on the fighters' names, rankings, fight stats going into the fight, and betting information. A summarised list of columns is below:

1. R_fighter
2. B_fighter
3. date
4. location
5. country
6. Winner
7. title_bout
8. weight_class
9. gender
10. no_of_rounds
11. height_dif
12. reach_dif
13. age_dif
14. sig_str_dif
15. finish
16. finish_details
17. finish_round
18. finish_round_time
19. total_fight_time_secs

In the UFC, the matched fighters are distinguished by being in the red corner and blue corner. The fight lasts for three 5-minute rounds, or five rounds in the case of a main event or title fight. If the fight goes to the final bell, the judges' scorecards are consulted to decide on a winner. The fight can also be stopped by the referee due to a knockout (KO), technical knockout (TKO), submission (SUB) or doctor's stoppage, when the ringside doctor deems an athlete unable to continue. There are divisions of weight class which rankings from 1 down to 15, and there is a global ranking called the 'pound for pound' list which is a theoretical ranking of the fighters if weight was not a factor. These stats are included in the dataset.

'R_fighter' contains the name of the fighter in the red corner. In MMA and boxing, the fighters are distinguished by being in the red corner and blue corner. In this dataset, 'difference' values like 'height_dif' and 'age_dif' are the comparison of blue to red: age_dif = 8 means that the blue corner is 8 years older than the red corner, height_dif = -3 means that the blue corner is 3cms shorter than the red corner.

The dataset also contains extensive fight metrics for both red and blue corners heading into the fight: for example, 'R_avg_SIG_STR_pct' contains the percent of significant strikes that were landed of those that were thrown, and 'R_win_by_TKO_Doctor_Stoppage' contains the number of wins by doctor stoppage for the red corner.

Initial Questions

- The description with the initial question
- For each question, a description of your visualization strategies, including data cleaning, transformation, visual encoding, etc.

Four initial questions were devised to take a shallow inspection at some trends in MMA fights.

1. Which fighter wins more, red or blue?
2. Does the weight class have an influence on the chance of KO?
3. How often does the older fighter beat the younger fighter?
4. Does having more wins than one's opponent lead to a better chance of winning?

These questions were designed to compare no more than 3 variables and to ascertain any trends on fighting with respect to age, weight class, and professional record.

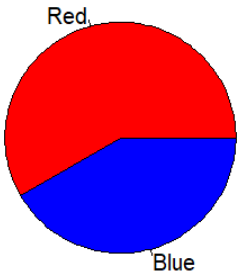
Q1: Which fighter wins more, red or blue?

All fights will end with either the red or blue fighter winning. This dataset contains no draws thus no cleaning is required. The 'Winner' column contains the corner which won the fight. We find the total number of fights won by the red corner and blue corner by summing the number of 'Red' and 'Blue' values in the 'Winner' column.

Total fights	Red corner won	Blue corner won
4891	2855	2036

The x data is categoric and has only 2 values. A bar graph may work however a pie chart allows for quicker comparison. The corner is the independent variable and the win counts are the dependent variable.

q1: Share of wins between red and blue corner



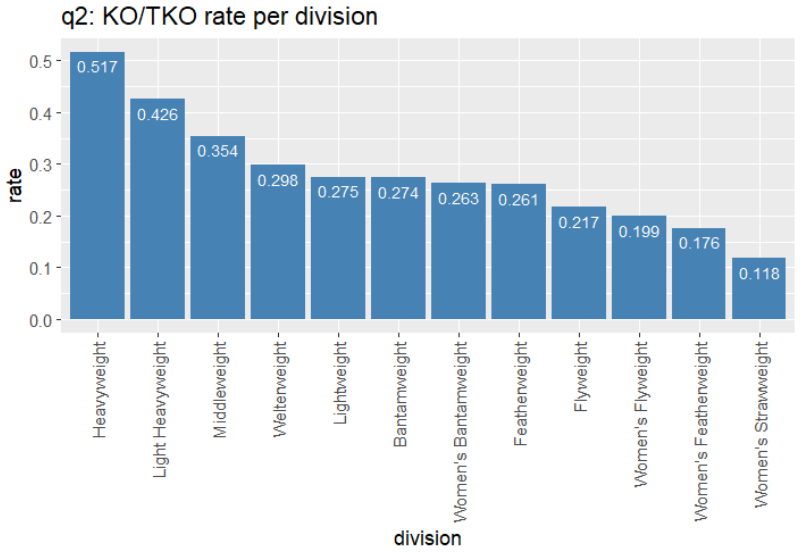
Q2: Does the weight class have an influence on the chance of KO?

This question concerns the number/ ratio of fights in each weight class that have resulted in KO or TKO (in this dataset, they are aggregated). All fights have a standard weight class and have some finish type. The 'catchweight' weight class was removed in cleaning because it is not defined by a strict weight, rather being outside of all other weights, threatening ambiguity.

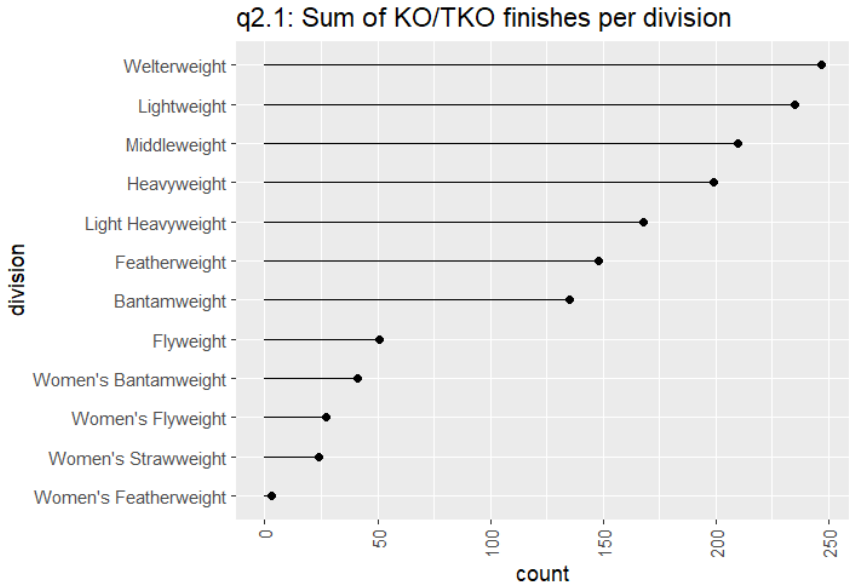
Fights in each weight class can be found by checking the 'weight_class' value. They can be summed to find a total number of fights for the weight class. We can extrapolate how many fights at some

particular weight class were stopped by KO/TKO. With these two sums we can find the decimal ratio of fights ending in KO/TKO stoppage per division. We store vectors of weight division, KO/TKO rate, and KO/TKO sum.

The x data is categoric but there are 12 values for x. A bar chart works because it can compare all the elements of x quickly and also sort them by value descending.



The sum of KO/TKO divisions can portray which division overall has seen more KO/TKO finishes. While this can be influenced by some divisions having had more fights in total than others it demonstrates a general trend reflected in the previous plot.



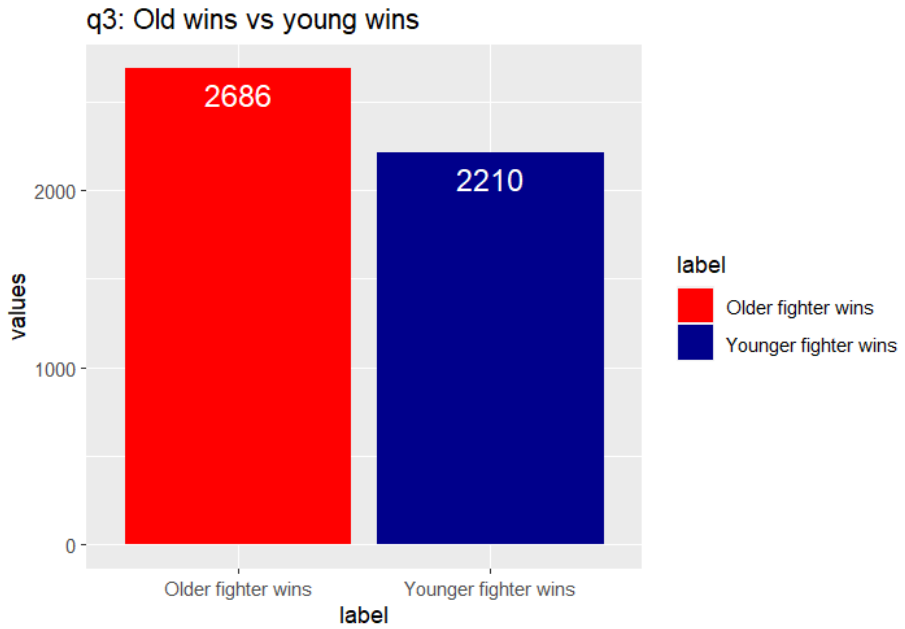
Q3: How often does the older fighter beat the younger fighter?

In MMA,

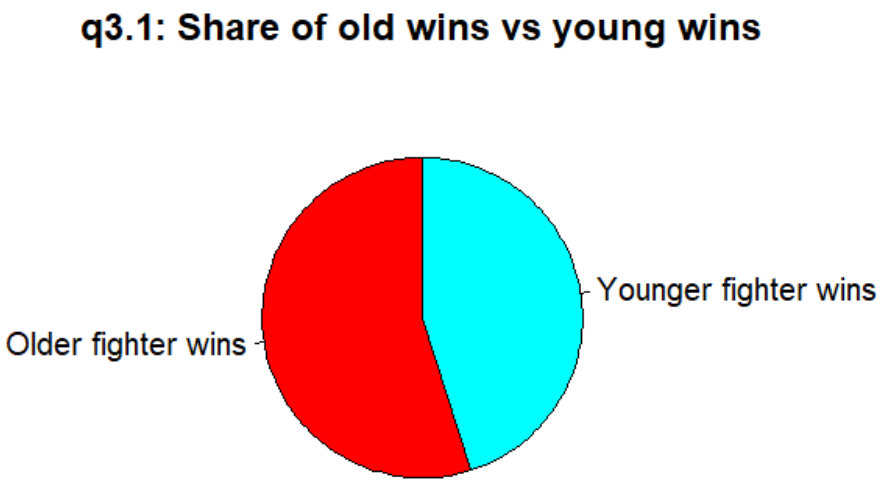
In MMA, age and experience is more prized than in other sports like soccer or tennis. The ‘age_dif’ column contains a number value that represents the age difference between red and blue, in years. For example, age_dif=3 means that the blue corner is 3 years older than red; age_dif = -3 means that the blue corner is 3 years younger than red.

To quickly ascertain which age won, we can use mutate to add a new 'Winner_age' column that is either 'Elder' or 'Younger'. This will represent which fighter won, the elder or the younger. We can easily build a dataframe, where the independent variable is 'Older' and 'Younger', and the dependent variable is the win counts for each.

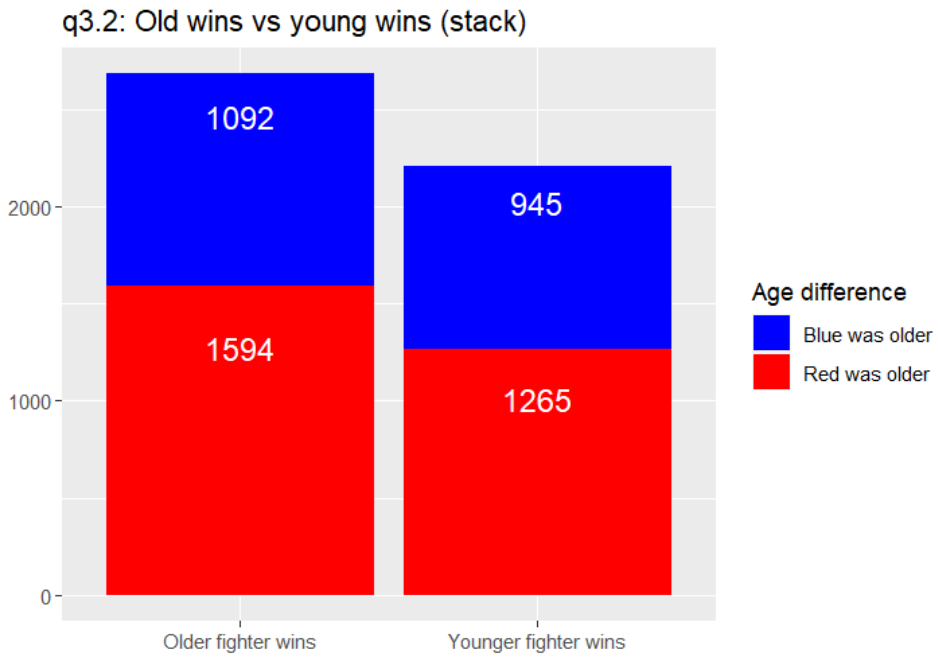
This data is categoric-continuous, so a bar chart is suitable.



Since the independent variable only has 2 categoric variables, it is suitable to plot as a pie chart for easy comparison.



Additionally, given the trend found in question 1 describing the red corner's higher chance of winning, it can be useful to break down amongst the older and younger fighter victories which corner won.

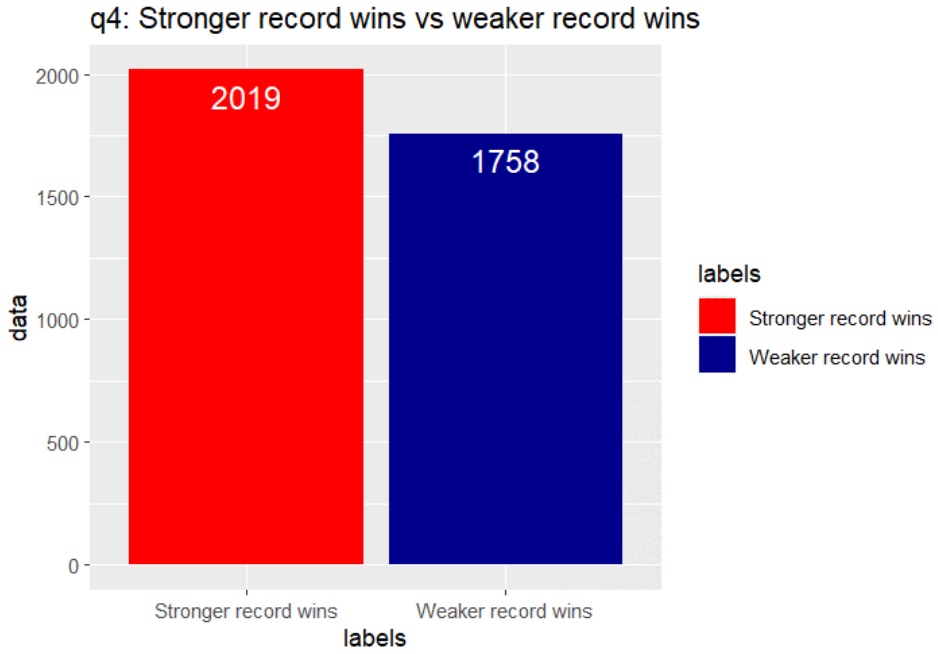


This data compounds the existing precedent that the red corner wins more often.

Q4: Does having more wins than one's opponent lead to a better chance of winning?

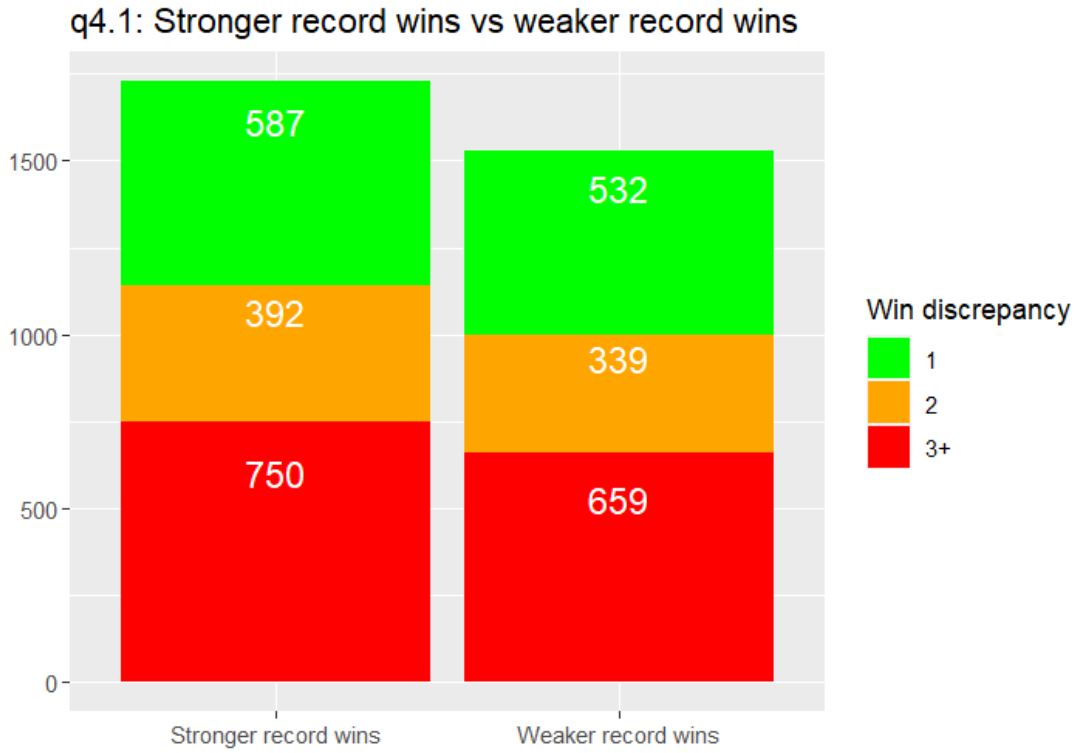
Similar to the previous question, the 'win_dif' column contains a value for how many more fights the blue corner has than the red, i.e. win_dif=3 means that the blue corner has 3 more wins than the red corner. Using mutate again we can add a 'Winner_record' column which contains either 'Stronger', 'Weaker', 'Matched' to indicate which record won the fight. 'Matched' indicates a win difference of 0.

The independent variable here is the categoric labels for win difference. We will not plot for 'Matched' as it is indiscernible which record won. The dependent variable is the win count for each.



The data confirms my suspicion that the stronger record would win the fight more often than not. This data is ambiguous however as 'stronger record' could be 1 more win or 5 more wins. We can

stratify our data by creating subcategories within the indepent variable: within each, we can discern the win discrepancy by 1, 2, and greater than or equal to 3. Now we can see how often stronger records beat out weaker records.



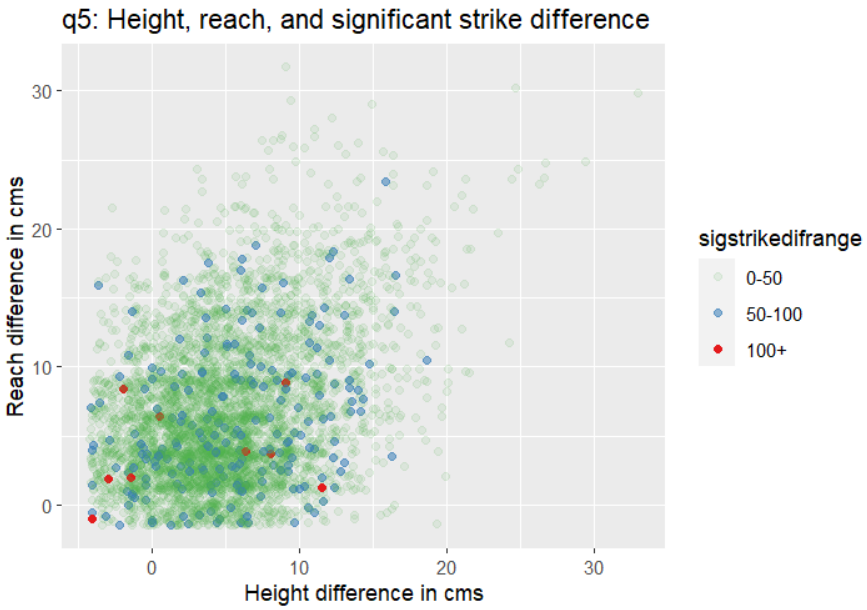
It is clear to see that amongst the occasions that the stronger record won, the winner had a win discrepancy of 3+.

Q5: Did a fighter's height and reach advantage (where present) have any correlation to outlanding their opponent?

In the dataset, height_dif, reach_dif, and sig_strike_dif columns contain the height and reach difference in centimeters, and the number of significant strikes difference. "Significant strikes" aredefined by FightMetric as "all strikes at distance and power strikes… not includ[ing] small, short strikes". A dataframe of these 3 differentials can be built.

Because the data was overlapping it made it hard to read any plot, so jitter was applied to the height and reach measurements to move points off common axes. An anomalous reach_dif measurement of -187.96 was removed from the dataset. A column 'sigstrikedifrange' was added using mutate to indicate the interval in which the significant strike difference was in: 0-50, 50-100, 100+.

Because this is trivariate continuous data it is best to use a scatterplot. The x-values are the height difference in cms, the y-values are the reach difference in cms, and the significant strike difference range is encoded in the colour and opacity of the points for simple comparison.

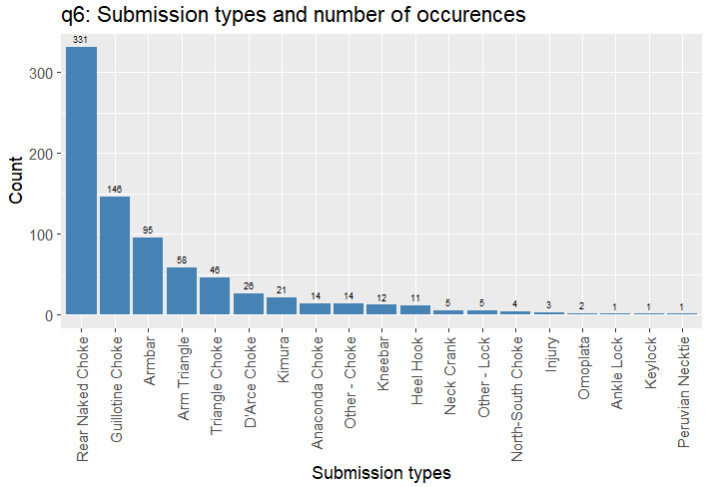


Surprisingly, there is a lack of correlation between height difference, reach difference, and significant strike difference. The fights that saw the highest strike differentials do not fall on the graph different to where all the other strike differentials fall.

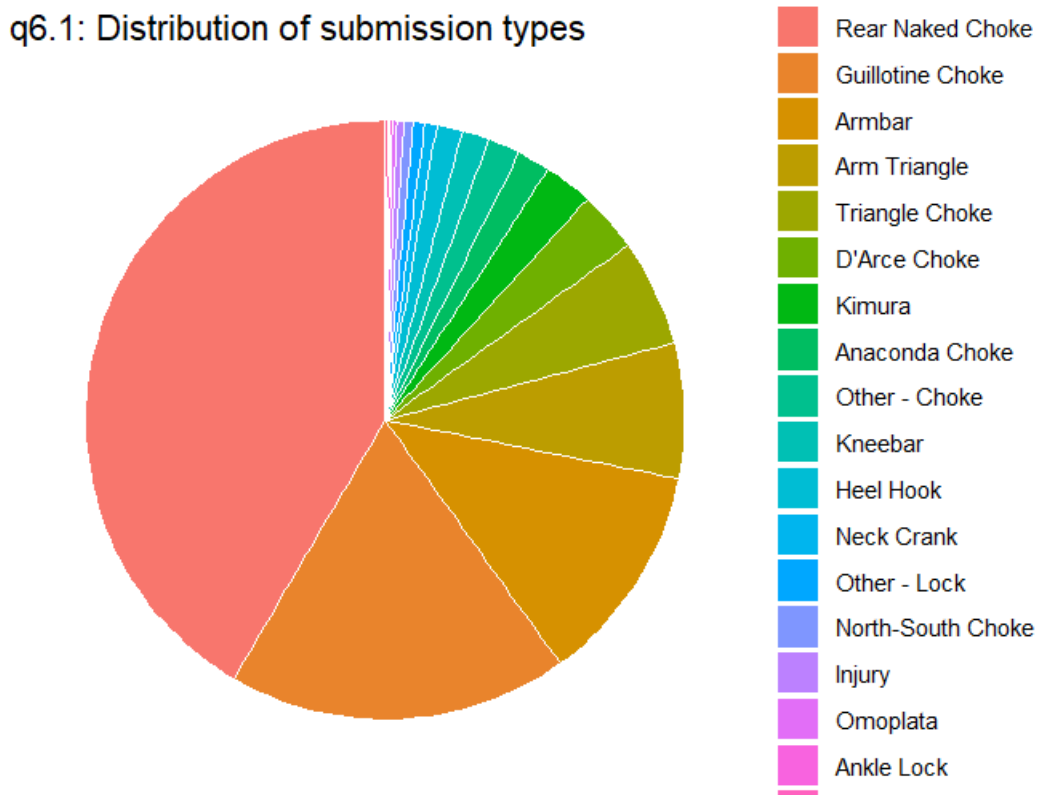
Q6: What is the distribution of submission types?

The 'finish' column contains how the fight was decided: judges' decision, KO/TKO, or submission, among other negligible finish methods. The 'finish_details' contains, for fights ending in KO/TKO or submission, the exact type of submission or KO/TKO. We can extrapolate all recorded submission types using unique, which removes duplicates, on the finish_details column.

We can sum the occurrences of each submission. In a dataframe we add the names of submissions as the categoric independent variable, and the counts of fights ending in such submission as the dependent variable. This data would most appropriately be plotted as a bar chart.



Interestingly the submission types and occurrences follow Zipf's law in that the most frequent submission will occur about twice as often as the second most frequent submission, three times as often as the third most frequent submission, etc. This can be further illustrated by a pie chart:



It is clear to see the angle of the sectors decreases steadily.

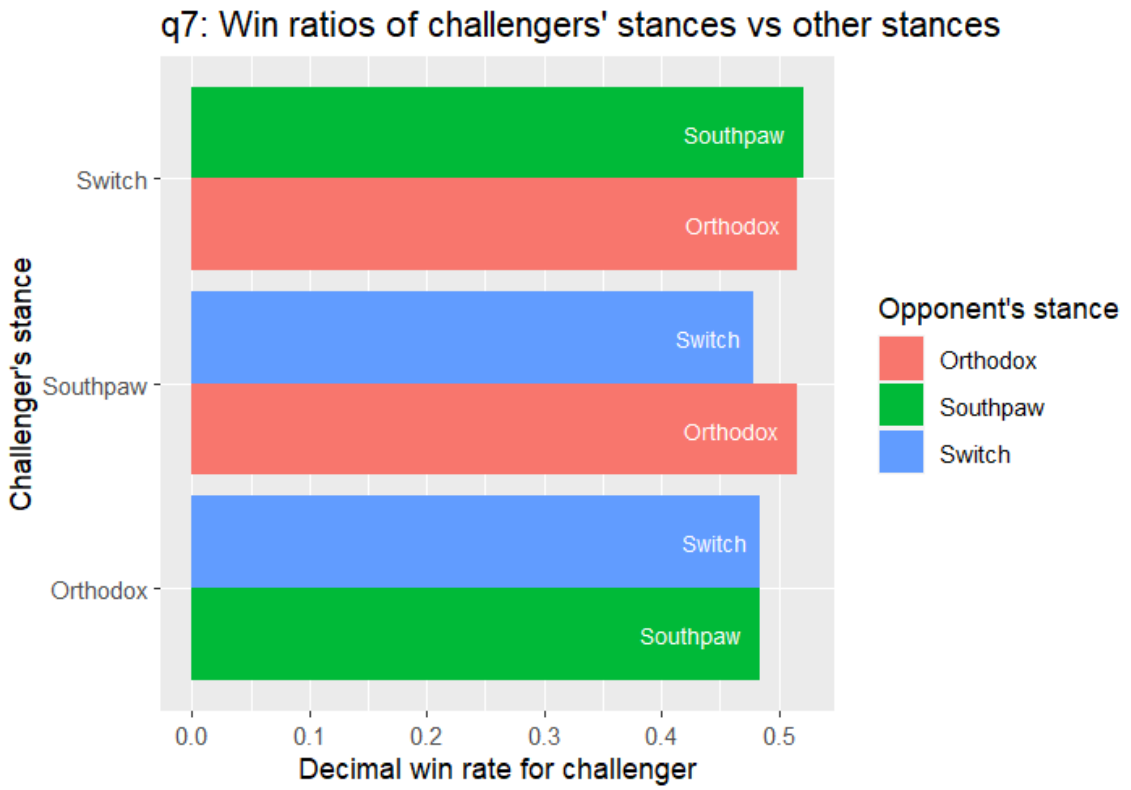
Q7: How did stances match up against each other?

The R_Stance and B_Stance columns contain the fighting stance of the red and blue corner, respectively. The possible values are “orthodox”, “southpaw”, “switch”, or “open stance”. The switch stance refers to using both orthodox and southpaw. The open stance stance refers to the fighters utilising opposite stances, so any records with such stance were omitted.

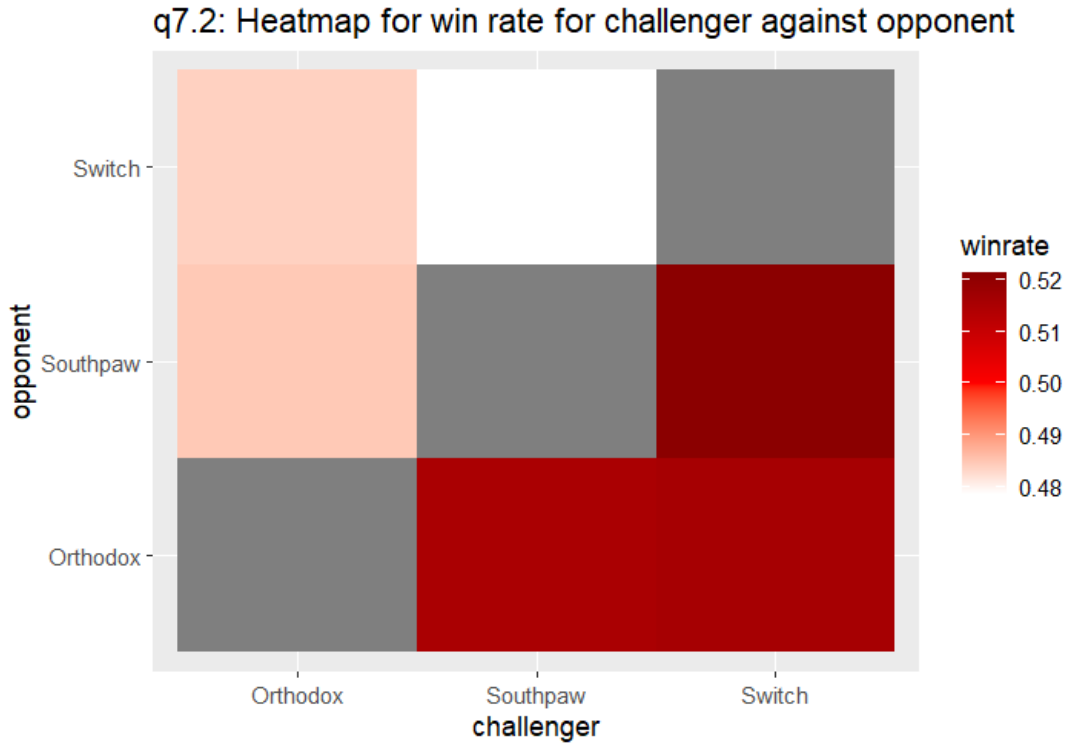
Using mutate, we can add a ‘Winner_stance’ column to quickly check which stance won, instead of having to go through Winner, R_Stance and B_Stance to decide.

This data must, for each stance, compare the chance of winning against each other stance. We use two loops over the set of stances. The outermost loop selects a stance. The next loop selects some stance other than the one already selected. We can then determine how many times these stances have encountered one another, and also how many times the initial stance won. This gives us a win ratio.

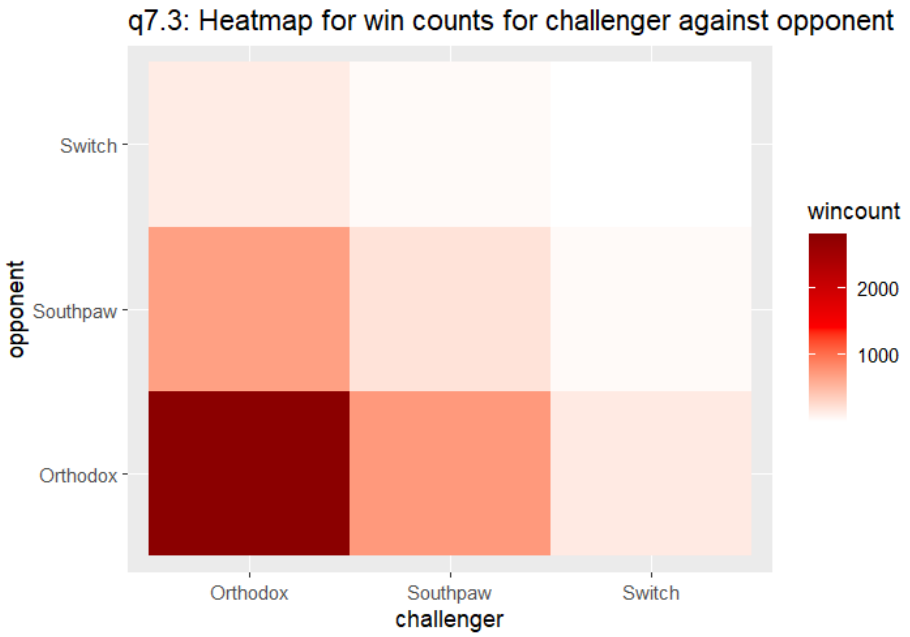
Our dataframe contains an independent variable of challenger’s stance, subgroups in this variable contain the opponent’s stance, and the dependent variable is the win ratio for the challenger. We can plot this data in a stacked bar chart.



We can see how much success each stance has against the other. For example, a southpaw is more likely to beat an orthodox fighter than beating a switch fighter. A heatmap can also help to compare amongst the stances.



This data however is affected by the number of fights associated to each stance. We can compare instead the win counts and not ratios to review which stance wins the most.

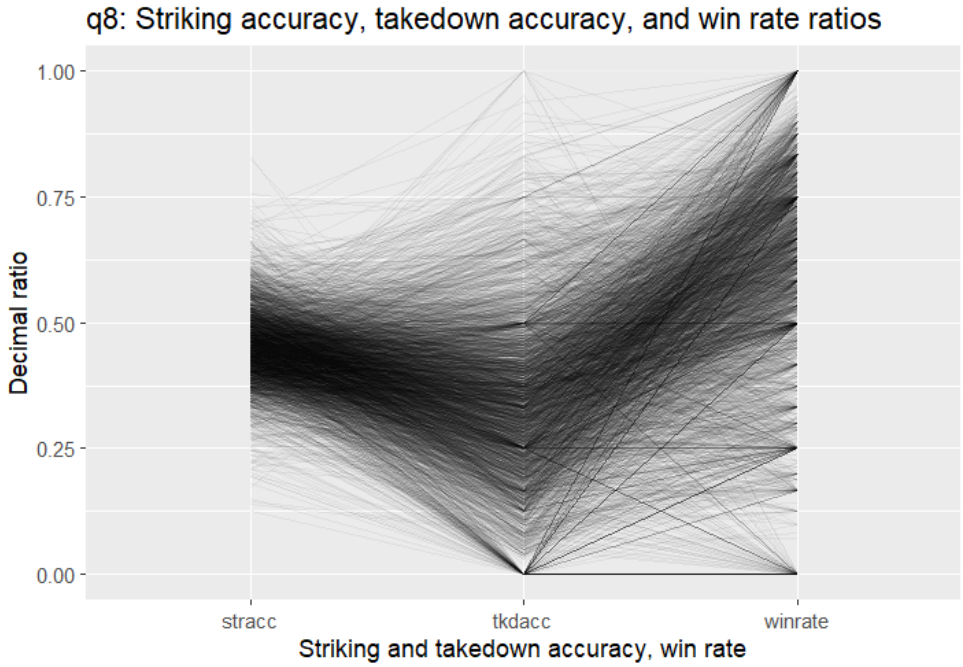


We can see that orthodox and southpaw fighters face each other far more than the other stances.

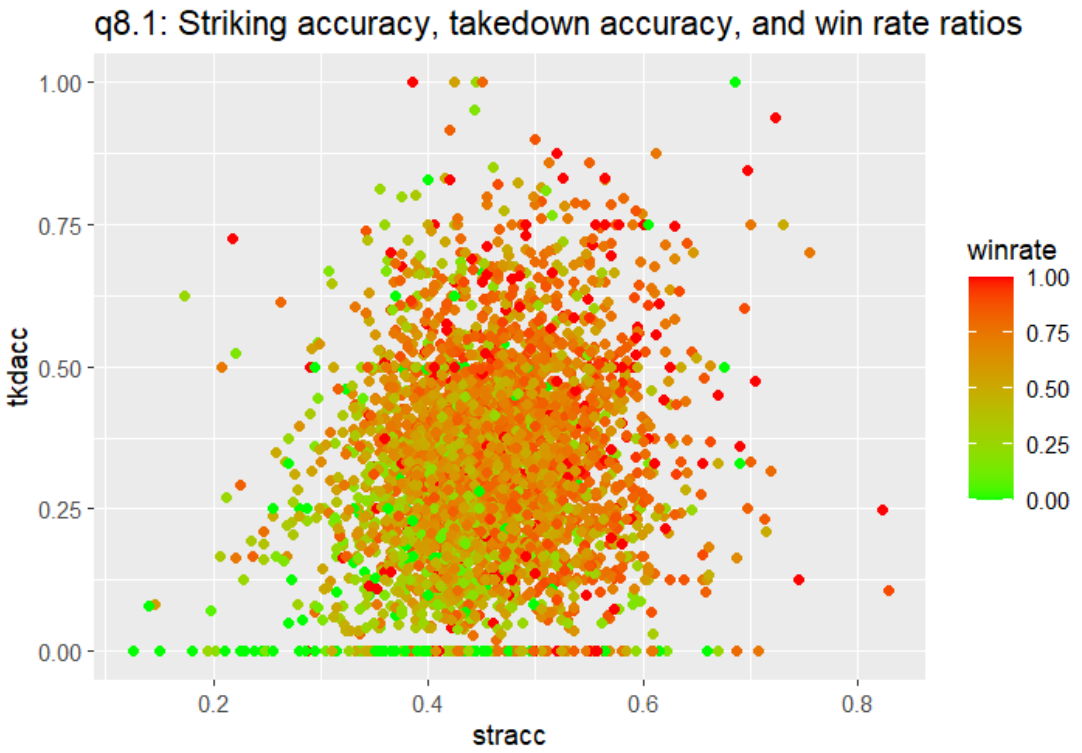
Q8: Which has a greater correlation to finishes - striking accuracy or takedown accuracy?

The dataset contains striking and takedown accuracy percentages for either corner. When we combine then we can find the general striking and takedown accuracy for fighters. Additionally, we can find the number of wins for each fighter and their total number of fights, therefore ascertaining their win ratio. NA data was present which was removed.

In a dataframe we combine striking accuracy, takedown accuracy, and win rate. Because all of these variables are continuous, we can plot as a parallel coordinates plot. The opacity of lines has been reduced to allow trends to be seen.

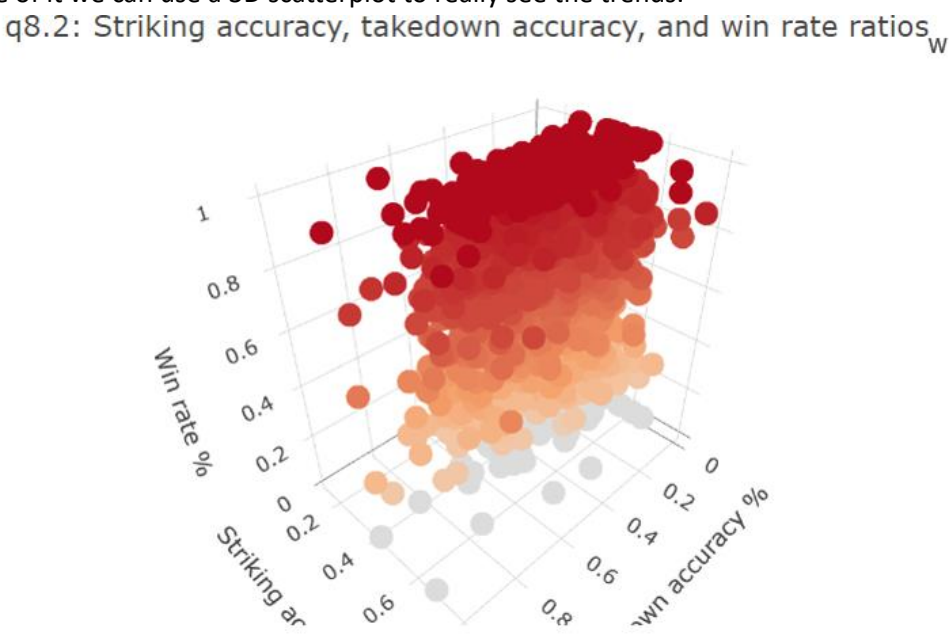


We may also use a scatterplot with color encoding to portray the trivariate data:



We can see how along the x-axis some values have 0 takedown accuracy, however it is useful to notice how the win rate increases along this line.

For the sake of it we can use a 3D scatterplot to really see the trends.



Discussion of Visualisations

I chose to veer from using pie charts as much as possible because it is difficult for humans to compare angles, especially across numerous categories. Bar charts are usually a good pick and stacked bar charts allow for subgroups of the x-values to be plotted for and give an introspective look at how the data interacts.

For questions 1 through 4 I stuck to bar charts and pie charts. In question 5 I was faced with trivariate data for which I decided to use a scatterplot to portray. The third vector was categoric derived from continuous and encoding with colour and alpha allowed for very quick comparison.

In question 7 a comparison was being made amongst members of the same group, with potential overlap (e.g. orthodox vs orthodox). Whilst in other visualisations it would not be wise to plot for same-element comparisons (win rate would be 1), in a heatmap it was negotiable and allowed for a clearer picture.

In question 8, the use of a 3D scatterplot was not beyond us and it was used to capture trivariate data relationships. It may have worked better than the parallel coordinates plot, which assumes the ratios have the same mean.

Reflection