Ultimate Fighting Championship (UFC) Data Analysis

Introduction

Project Purpose:

The purpose of this data analysis project is to explore and uncover meaningful insights within the UFC dataset which contains fighters' profiles, match outcomes, performance statistics and historical trends. Through statistical means, we will identify patterns and key factors that contribute to success in the UFC.

Initial data preprocessing/cleaning:

Data.csv: This data set contains historic fight data with averaged statistics for each fighter in a match. There were some missing numerical values regarding fight statistics so we imputed them with the median values as we felt it was still representative of the data.

Raw_fighter_details_data.csv: This data set contains fighter builds and career statistics. The height was recorded in foot-inches notation so we transformed them to just inches. Further, the percentages were converted to decimals and the missing values were imputed with the medians.

Raw_total_fight_data.csv: This dataset contains historic data of individual match statistics for each fighter involved in a match. There was no initial preprocessing necessary for this data.

Red or Blue: Colour with the most wins

Goal:

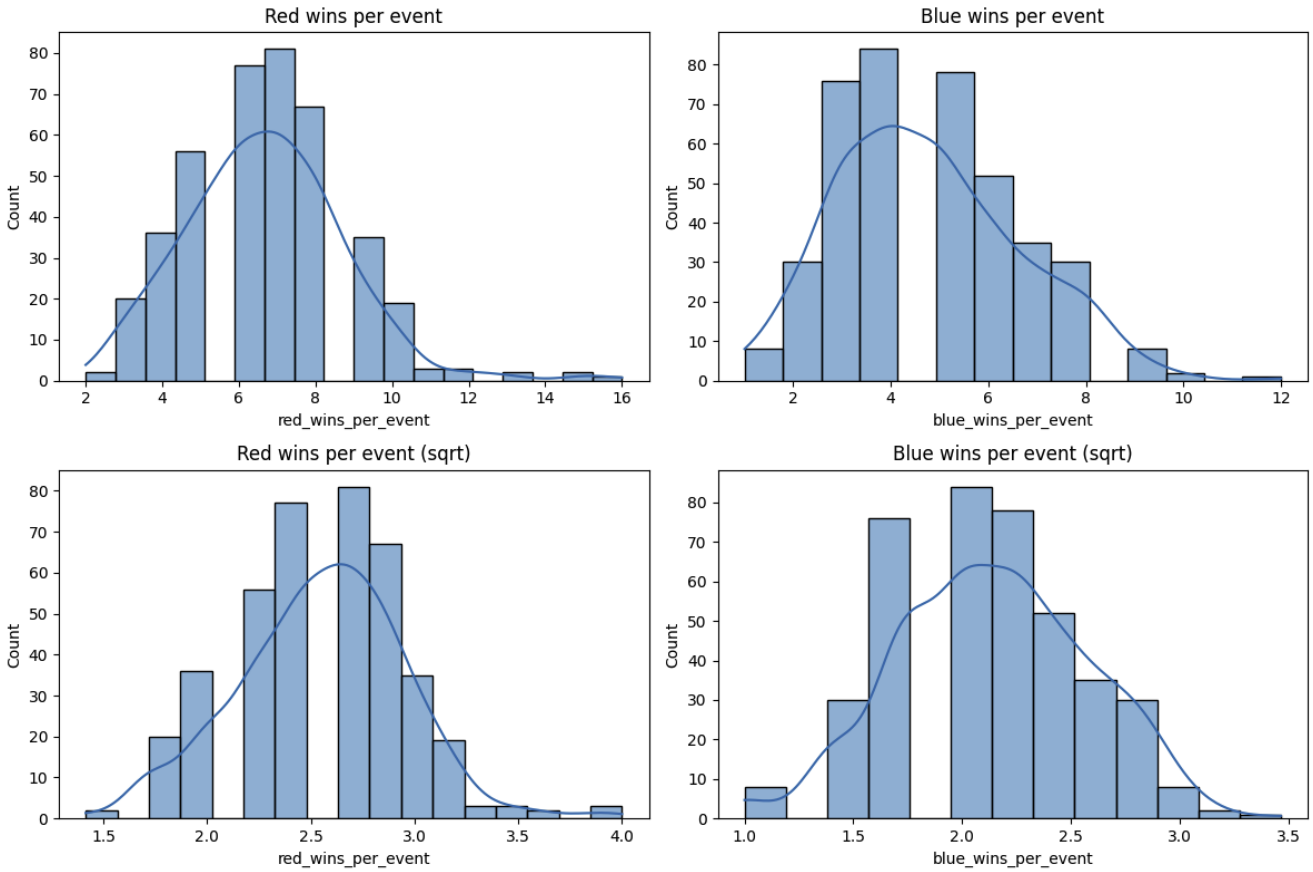
In each UFC fight, one of the fighters represents the red colour and the other blue. Our goal in this part of the analysis is to determine if there is a difference in the number of wins between red-gloved and blue-gloved fighters. Further, we want to determine if one of the colours wins significantly more than the other.

Further Preprocessing:

We start with the `preprocessed_data.csv` data set which contains the fight statistics for each match. We categorize each fight as a red win or a blue win. For the sake of analysis, fights ending in a draw are excluded.

Approach:

All the fights have now been separated into red or blue wins we may proceed. It's well known that every UFC event hosts multiple fights on the same night. With this context, we can group our red wins and blue wins by date to give us the amount of red wins and blue wins during a particular UFC event. We can use this data to get the average amount of red and blue wins per event. The plot of the red and blue wins per event date is as follows (the wins originally produced a right-skewed graph so a square root transformation was used to normalize the data groups):



Raw Results:

At this point, we have the number of wins for each colour for every event date in our dataset. Taking the mean of each group. We find that on average red fighters win 6.68 times every event and blue fighters win 4.76 times.

Significance of Results:

We must determine if this mean difference is significant. We're aware that our data is normal (as seen in the plots above). However, Levene's test on the 2 groups reveals a p-value of 0.018, indicating that our data does not have equal variance. Taking these facts into consideration, we move forward with an independent Student's t-test ensuring that we set the `equal_variance` parameter to false. We obtain a p-value of 6.06e-40, indicating that there is a significant difference in the average red wins per event and the average blue wins per event. Therefore, we conclude that red wins more than blue because red wins significantly more fights on average per UFC event than blue does.

Most effective strike leading to wins

Goal:

Our goal in this part of the analysis is to determine the most effective type of strike that helps win UFC fights. The strike types to be analyzed are head strikes, body strikes and leg strikes. Further, we want to compare the differences between how effective each strike type is compared to the others.

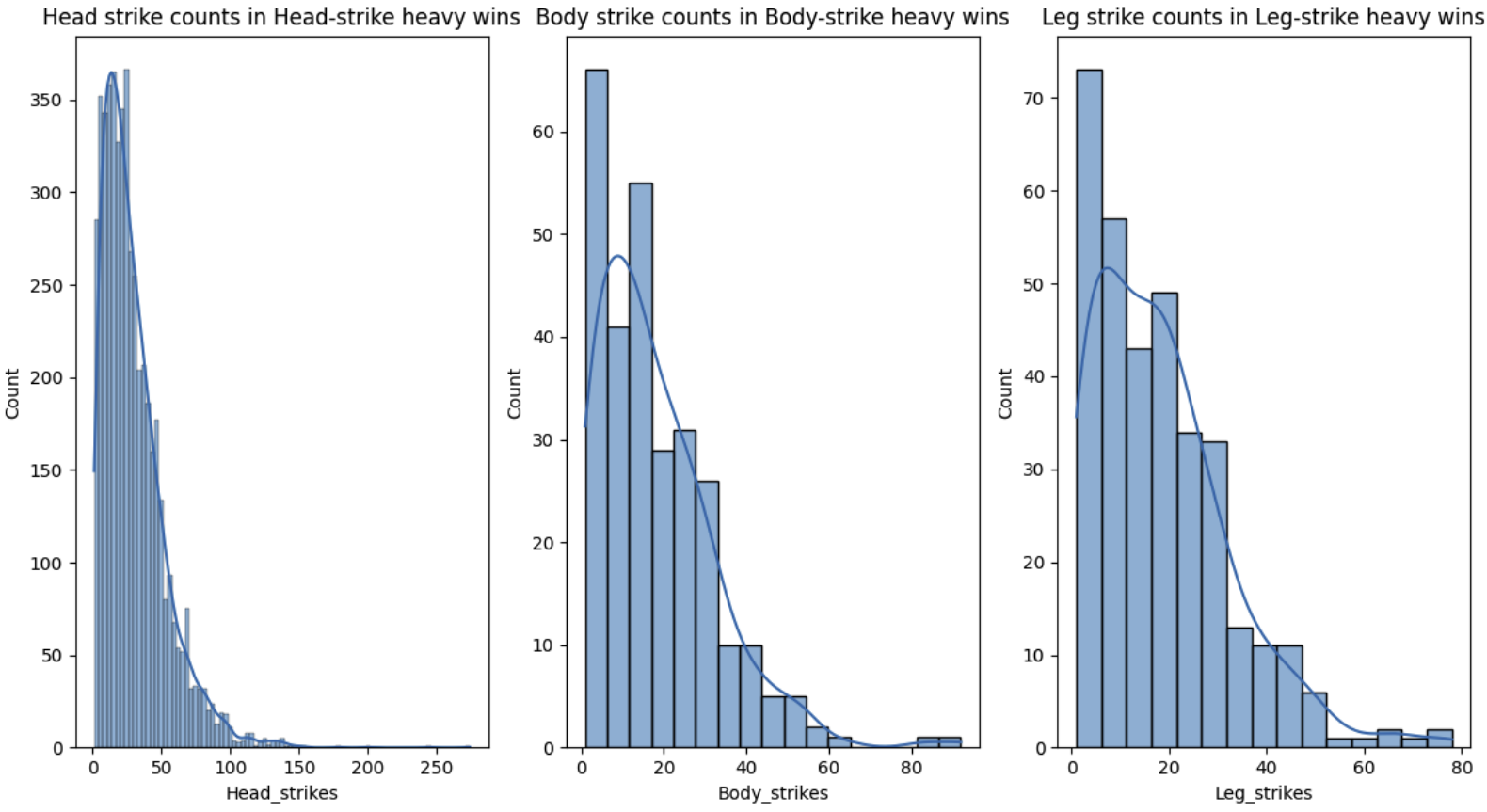
Further Preprocessing:

We start with the `raw_total_fight_data.csv` data set which contains the fight statistics for each match. This includes strike statistics for the red and blue fighters accrued through the duration of the fight. Since we're interested in seeing which strikes are the most effective at winning matches, we only

extract the strike statistics for the winner of each match. At this point, we have the number of head, body and leg strikes of the winner in a given match for each fight in our dataset. The counts for each strike type were recorded in the format `<number of landed shots> of <total attempted shots>`. If we're determining how effective a particular strike type is, we only want to consider the ones that landed on their opponent. Based on this, we transform our data to only contain the winner's <number of landed shots> for each strike type in a given match, effectively removing the total amount they attempted. This concludes the extra preprocessing needed to continue this part of the analysis.

Approach:

Now that we have the number of each strike type landed for the winner of each match, we want to determine the strike type that "helped" them win the most. For simplicity's sake, we take the strike type that the winner landed the most on their opponent and label that as the 'dominant strike' used to win the match. For example, if a particular winner accrued 5 head strikes, 15 body strikes and 25 leg strikes, we label that win as a 'leg strike dominant win' as it can be argued that leg strikes contributed more to the outcome of winning than the other strikes. We repeat this process for the winner of each fight. As a result, each win will be labelled a 'head-strike-dominant', 'body-strike-dominant', or 'leg-strike-dominant' win. Next, we group each win by the dominant strike used in it. For each dominant strike group, we extract the number of landed shots for the relevant dominant strike type for each fight in that group. For clarity, this would mean that we extract the head-strikes count for each win in the head-strike-dominant group, the body-strikes count for each win in the body-strike-dominant group and the leg-strikes count for each win in the leg-strike-dominant group. At this point, we'll have 3 groups representing the 3 strike types. Each holds an array of the number of shots of that strike type that it took to stop an opponent for the fights within that group. Using one group as an example, this may look something like `head_strike_dominant_wins = [12, 32, 3, 5]` which would indicate that there were 4 head-strike-dominant wins and the head-strike count in each win was 12, 32, 3 and 5. We will think of each of these array numbers as "the number of shots of this striking type that it took to stop an opponent". This is repeated for the other 2 groups as well. The plot of the data is as follows:



Raw Results:

At this point, we have the number of strikes of each strike type that were used to win matches where that strike type was used dominantly. Next, we take the mean of each group. We find that it takes on average 29.63 head strikes to stop an opponent in head-strike-dominant wins, 17.99 body-strikes in body-strike-dominant wins and 18.29 leg-strikes in leg-strike-dominant wins. We acknowledge that the fewer shots of a strike type landing on an opponent leading to a win is directly proportional to how effective it was at stopping the opponent. By that logic, the raw numbers suggest that body strikes are the most effective, leg strikes are a close second and head strikes are relatively the least effective - but we're not done yet.

Significance of Results:

Although we have the averages of each strike type that helps win matches, we must determine if our results were significant. After conducting Levene's test on the 3 strike types, we obtained a p-value of 6.24e-17, indicating that our data does not have equal variance. In light of this, we use a non-parametric alternative to the ANOVA test known as the Kruskal-Wallis test, We obtain a p-value of 1.85e-19 indicating that there is a significant difference in the medians of the groups being compared. Next, we conduct a Games-Howell posthoc test, which is known for being robust around unequal sample sizes and variance in our data, to compare which groups have significant differences from each other. The results are as follows:

A	B	mean(A)	mean(B)	diff	se	T	df	pval
body_strikes	head_strikes	17.992933	29.630349	-11.637416	0.912074	-12.759296	370.636191	1.065814e-14
body_strikes	leg_strikes	17.992933	18.287834	-0.294901	1.146284	-0.257267	596.045500	9.641744e-01
head_strikes	leg_strikes	29.630349	18.287834	11.342515	0.833953	13.600903	467.785513	7.771561e-14

There is a significant difference in the medians between body strikes and head strikes with a p-value of 1.07e-14 as well as a difference between head strikes and leg strikes with a p-value of 7.77e-14. It can be said that there isn't a significant difference in the median amount of body strikes and leg strikes it takes to stop an opponent as suggested by a p-value of 0.96 meaning they're about equally as effective. Therefore, we determine that the most effective strikes in UFC matches are body and leg shots coming in first place, followed by head shots in last place.

Most effective stance leading to wins

Goal:

In this part of the analysis, our goal is to determine the most effective stance that contributes to the most wins. The stances to be analyzed are orthodox, southpaw, switch, open and sideways. Further, we want to compare the differences between how effective each stance type is compared to the others.

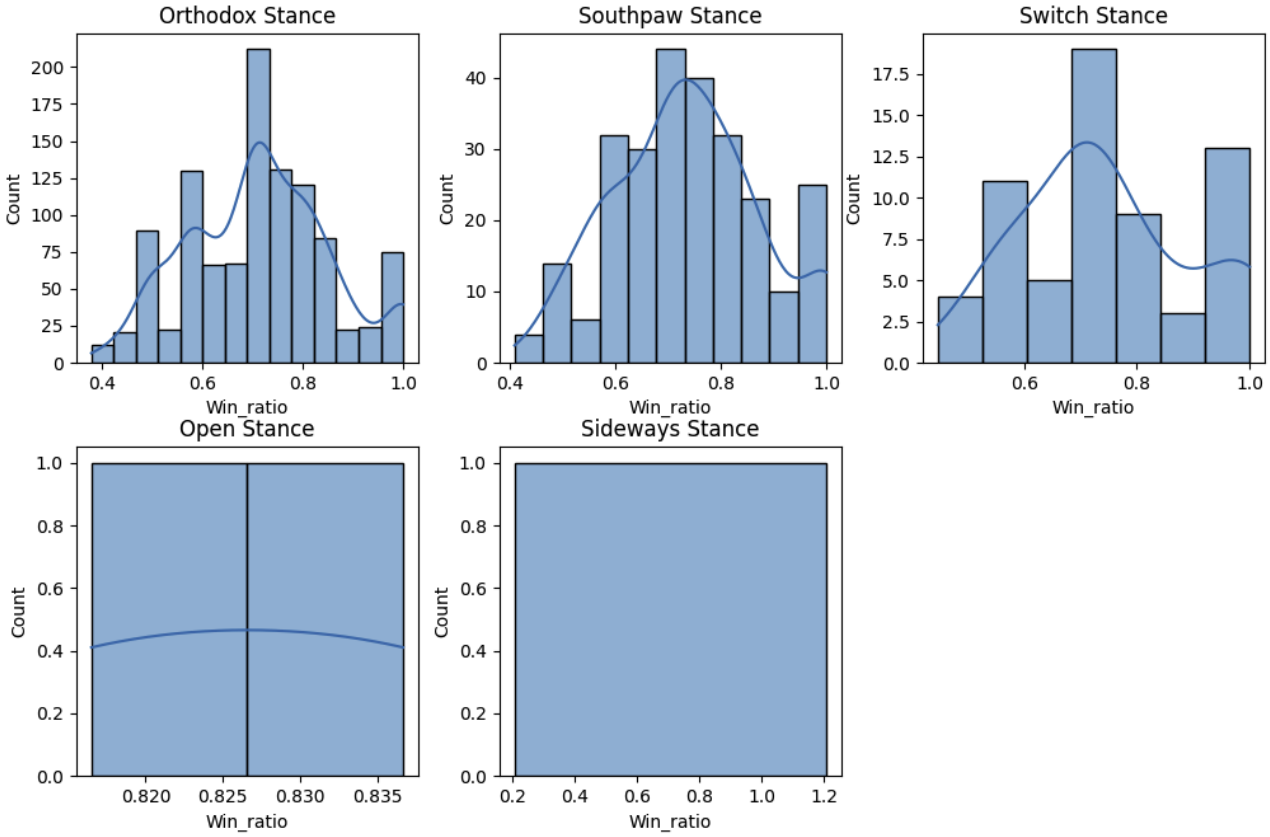
Further Preprocessing:

We start with the `preprocessed_data.csv` data set which contains the fight statistics for each match. If you're curious, this data set is different from `raw_total_fight_data.csv` as the different statistics representing the red and blue fighters' statistics are recorded over their career rather than for a specific fight. We start by removing any fights that end in a draw (as they don't contribute to the question) as well as any fights where the stances are not recorded for the blue or red fighter. We also made the bold decision to exclude any fighters without a single win as we believe that this would give us a more accurate analysis as the data represents fighters who show at least minimal competence with their stance. Next, we extract the number of wins and losses for each fighter along with the stance they predominantly use. Calculating the win ratio is a simple division operation between the number of wins

and losses. At this point, the data we need and have is the win ratio and stance data of every fighter in the data set. This concludes the extra preprocessing needed to continue this part of the analysis.

Approach:

Now that we have the win ratios and stances for each fighter, we need to determine the win ratio means of each stance. We do this by grouping each fighter by their stance and averaging their win ratios to represent the mean win ratio for a particular stance. The plot of the win ratios of each stance is as follows (the win ratios produced right-skewed graphs so a square root transformation was used to normalize the data groups):



Raw Results:

Now that we have win ratios for each stance, it's time to analyze the data. You may have noticed that some of the stances above look questionable. It seems that we are limited in our analysis because the open and sideways stances are not nearly as prevalent in our dataset as the others. We decided to make a judgement call to exclude the open and sideways stances as they have only 2 and 1 data point(s) to represent them, respectively. The switch stance doesn't quite meet the threshold of "not enough data" so we proceed with it in our analysis. We find that the average win ratio is 52.83% for the orthodox stance, 56.13% for the southpaw stance and 57.77% for the switch stance.

Significance of Results:

We have the average win ratios representing each stance but must determine if the differences between them are significant. We conduct a Levene test on the stance data to determine if our data has equal variance and obtain a p-value of 0.86, suggesting that our data can be characterized as having equal variance. Since we're aware that our data is also normal (as seen in the plots above), we can conduct an ANOVA test to determine if there is a significant difference in the mean win ratios of the stances being compared. Our ANOVA test gives us a p-value of 0.02, implying a significant difference.

Next, we conduct a Games-Howell post hoc test to determine which groups have a significant difference in their means. We use this test instead of Tukey's HSD because it's able to robustly handle different sample sizes, as is present in our data. The results are as follows:

A	B	mean(A)	mean(B)	diff	se	T	df	pval
orthodox	southpaw	0.712909	0.736050	−0.023141	0.009704	−2.384773	396.891508	0.046101
orthodox	switch	0.712909	0.745022	−0.032113	0.019426	−1.653101	69.709675	0.230606
southpaw	switch	0.736050	0.745022	−0.008972	0.020838	−0.430568	91.330200	0.902980

There is a significant difference in the win ratio means between the orthodox and southpaw stances with a p-value of 0.046. There doesn't seem to be a significant difference between the switch stance and the other two stances with p-values both greatly above 0.05 which may be a result of having fewer data points for the switch stance. The post hoc suggests that southpaws have a higher win ratio than orthodox fighters. We determine that southpaws are more likely to win than orthodox fighters whilst acknowledging that much can't be said about where switch stance should fit into the mix.

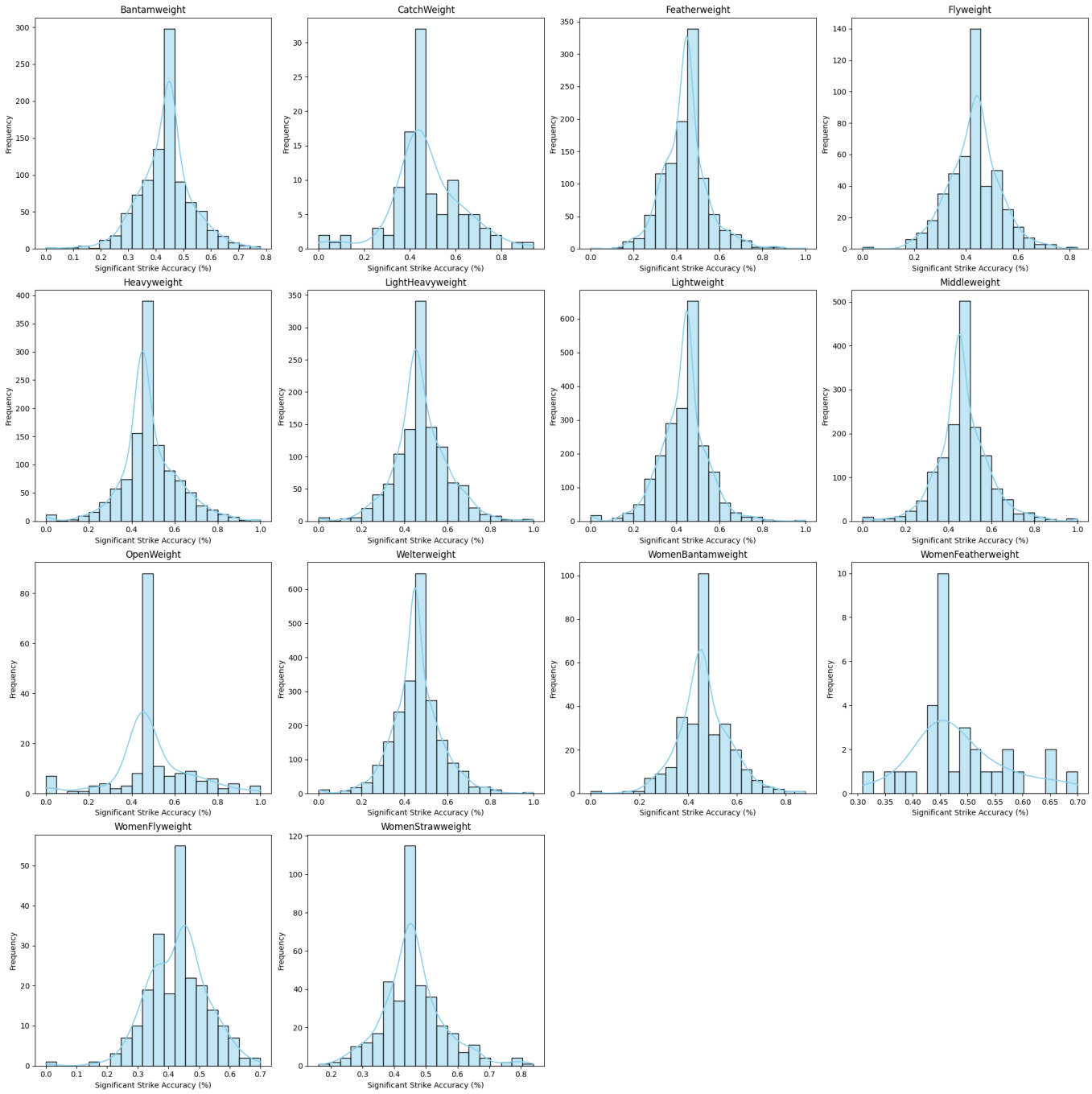
Difference in the average significant strike accuracy across different weight classes

Goal:

This investigation centred on the question of whether significant differences exist in the average significant strike accuracy across different weight classes in UFC fights.

Approach:

Our approach began with a thorough examination of the dataset, where we combined separate statistics for Red and Blue fighters into a single column including all the fighter's average significant strike accuracy for each match. This step was essential to ensure that our analysis would be based on complete and unduplicated data. To recognize patterns in average significant strike accuracy and weight class, we grouped the data by weight class, this allowed us to plot histograms showing normally distributed data regarding strike accuracy counts for each weight class. We also tested for equal variance using Levene's test and did not find equal variance.



Significance of Results:

To answer this question, we used 2 statistical methods, performing the Kruskal-Wallis test. The Kruskal-Wallis test is ideally suited to comparing the median across multiple groups to determine statistical significance. It can identify whether at least one group's median significantly differs from the others, especially when there is no equal variance. We can then perform the Games-Howell test for detailed comparisons of which group pairs differ.

After using Levene's test to test for equal variance, the p-value we got was 3.24338e-13, the value indicated that the variance is not equal thus it makes sense for us to use Kruskal instead

of ANOVA. A notably low p-value (p=2.24416583e-32) indicated significant differences in strike accuracy across the weight classes. However, the Kruskal test doesn't specify which groups differ. We then performed a Games-Howell test for detailed comparisons. The image below shows the results in the terminal, which shows the median differences between weight classes. The 'pval' column indicates whether or not we reject the null hypothesis. For example between Bantamweight and Heavyweight, we have a p-value less than 0.05, showing a significant difference in the median and proving that weight classes can differ in their significant strike accuracy, same thing between Bantamweight and LightHeavyweight. In conclusion, the Kruskal-Wallis test and Games-Howell together provide meaningful insights into our analysis. The Levene test supported the use of the Kruskal-Wallis test, and the Games-Howell results clarified specific differences, fully addressing our question.

	A	B	mean(A)	mean(B)	diff	se	T	df	pval	hedges
0	Bantamweight	CatchWeight	0.431001	0.481594	-0.050593	0.021455	-2.358094	58.129819	5.213725e-01	-0.475510
1	Bantamweight	Featherweight	0.431001	0.438344	-0.007343	0.006487	-1.131957	1019.735407	9.975556e-01	-0.070383
2	Bantamweight	Flyweight	0.431001	0.428657	0.002344	0.008527	0.274839	417.354144	1.000000e+00	0.022782
3	Bantamweight	Heavyweight	0.431001	0.485354	-0.054353	0.007387	-7.358373	1039.610176	3.476008e-11	-0.438758
4	Bantamweight	LightHeavyweight	0.431001	0.484040	-0.053039	0.007033	-7.541704	1041.725718	9.235945e-12	-0.456939
..	...	...	...	...	...	...	...	...	...	...
86	WomenBantamweight	WomenFlyweight	0.472626	0.432120	0.040506	0.012905	3.138719	240.893749	1.018034e-01	0.389571
87	WomenBantamweight	WomenStrawweight	0.472626	0.451124	0.021502	0.010995	1.955699	310.656428	7.936281e-01	0.214096
88	WomenFeatherweight	WomenFlyweight	0.490212	0.432120	0.058092	0.026711	2.174798	19.874272	6.487503e-01	0.563155
89	WomenFeatherweight	WomenStrawweight	0.490212	0.451124	0.039088	0.025843	1.512563	17.458383	9.472998e-01	0.400928
90	WomenFlyweight	WomenStrawweight	0.432120	0.451124	-0.019003	0.011982	-1.585976	221.171939	9.481158e-01	-0.191100

Weight classes committing to a certain type of finish/win_by method

Goal:
This investigation was to determine if specific finish or win-by methods are more common in certain weight classes within UFC fights.

Approach:
Our approach required a detailed review of the dataset, leading us to combine the separate win-by methods for Red and Blue fighters into unified columns for each type of win. We then grouped the aggregated data by weight class, forming a contingency table that displayed the distribution of each win type across weight divisions. This was essential for preparing the data for further analysis to detect patterns in win methods associated with different weight classes.

Significance of Results:
Our method involved applying the Chi-Square Test of independence to a contingency table of win types by weight class. This test is great for understanding if there's a statistically significant association between the categorical variables of weight class and the win_by method. The p-value from the Chi-Square Test was 0.0 indicating an extremely strong statistical significance. This suggests that the distribution of win types is not uniform across weight classes, and certain finish methods are indeed associated with specific weight classes. The contingency table shows the counts of each win type per weight class. For example, heavyweights have a notably high

number of wins by KO/TKO compared to other weight classes, which could imply a greater striking power in this division. In contrast, some lighter-weight classes may show a higher number of wins by decision or submission.

weight_class	Total_win_by_Decision_Split	Total_win_by_Decision_Unanimous	Total_win_by_KO/TKO	Total_win_by_Submission	Total_win_by_TKO_Doctor_Stoppage
Bantamweight	256	875	646	513	11
CatchWeight	24	95	119	135	16
Featherweight	338	1374	943	703	43
Flyweight	140	595	264	236	4
Heavyweight	97	679	2230	640	52
LightHeavyweight	236	1161	1900	733	103
Lightweight	690	2610	1952	1936	172
Middleweight	433	1615	2095	1070	104
OpenWeight	0	2	95	88	10
Welterweight	668	2672	2608	1764	138
WomenBantamweight	90	217	152	101	4
WomenFeatherweight	7	13	50	11	0
WomenFlyweight	77	217	72	100	7
WomenStrawweight	102	437	60	181	5

p-value: 0.0

Difference in average knockdowns and average takedown percentages across 5 weight classes

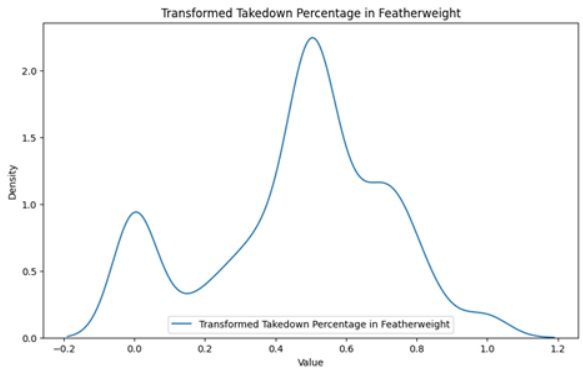
Goal:

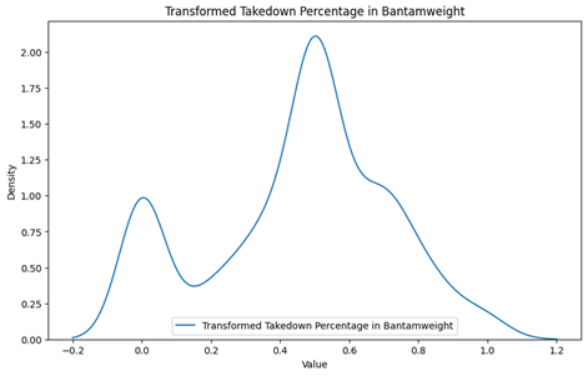
Our goal here is to determine if there is a change in the average knockdowns or average takedown percentages as we climb our weight classes from lightest to heaviest. The five weight classes (Bantamweight, Featherweight, Lightweight, Welterweight, Middleweight) are consecutively selected, each increasing by 10lbs starting at 135lbs (Bantamweight) and ending at 185lbs (Middleweight).

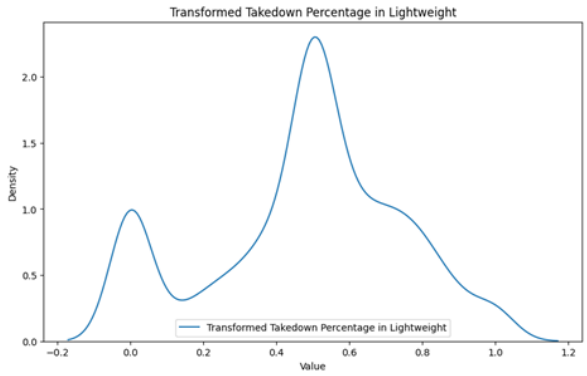
Further Preprocessing:

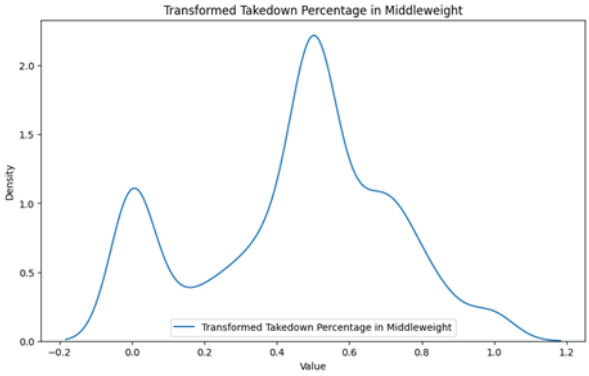
We begin with the “preprocessed_Data.csv” dataset containing the relevant statistics per match. The dataset is filtered to include only the selected weight classes. Notably, we combine takedown percentages and knockdown averages from both red and blue corners for each weight class.

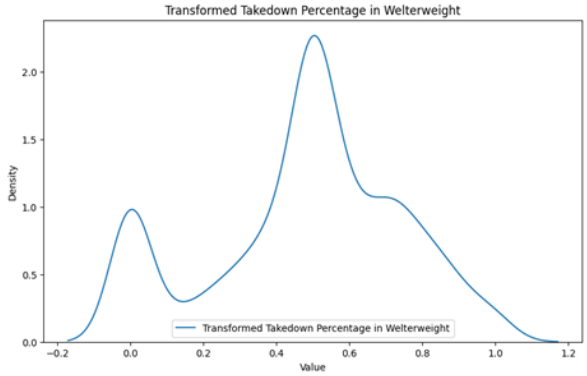
Approach:

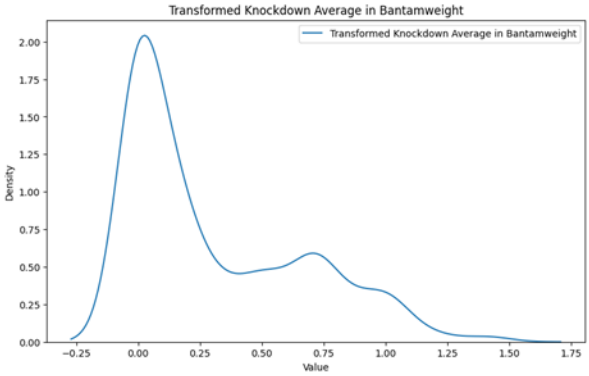


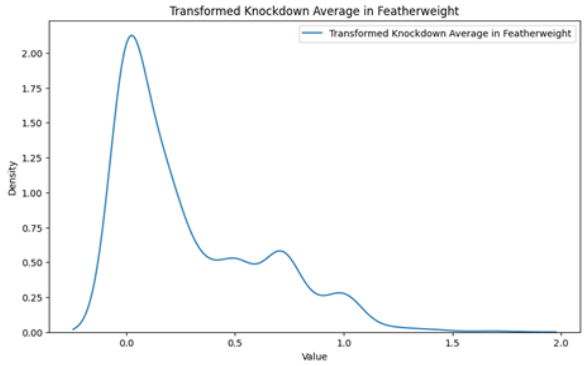


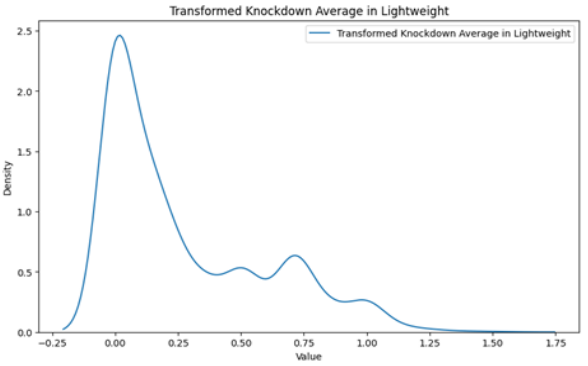

Transformed Takedown Percentage in Lightweight
Transformed Takedown Percentage in Lightweight
Density
Value

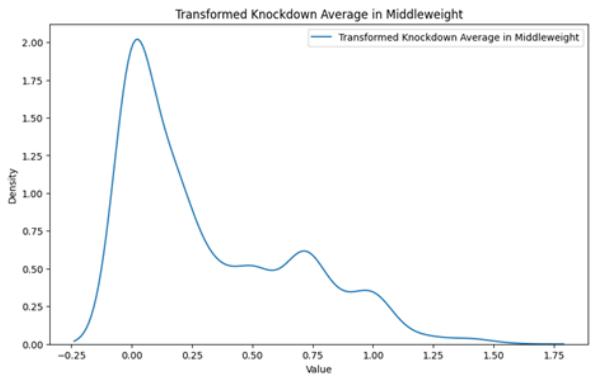

Transformed Takedown Percentage in Middleweight
Transformed Takedown Percentage in Middleweight
Density
Value

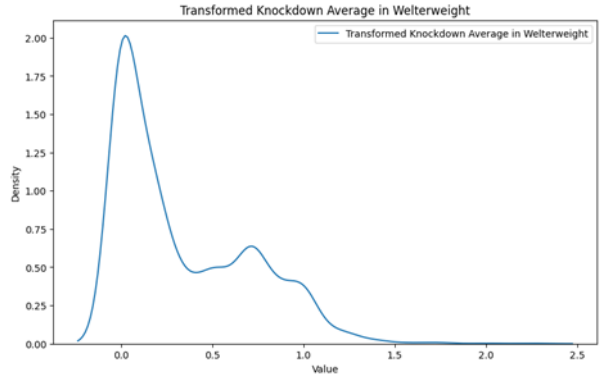

Transformed Takedown Percentage in Welterweight
Transformed Takedown Percentage in Welterweight
Density
Value


Transformed Knockdown Average in Bantamweight
Transformed Knockdown Average in Bantamweight
Density
Value


Transformed Knockdown Average in Featherweight
Transformed Knockdown Average in Featherweight
Density
Value


Transformed Knockdown Average in Lightweight
Transformed Knockdown Average in Lightweight
Density
Value


Transformed Knockdown Average in Middleweight
Transformed Knockdown Average in Middleweight
Density
Value


Transformed Knockdown Average in Welterweight
Transformed Knockdown Average in Welterweight
Density
Value

The preprocessed data is analyzed using both visual and statistical methods. Kernel Density Estimation (KDE) plots are generated for each weight class to visualize the distribution of takedown percentages and knockdown averages. For statistical analysis, we first perform Levene's Test to check for equality of variances. Given the results and the non-normal data distribution, we proceed with the Kruskal-Wallis test, a non-parametric alternative to ANOVA, to assess whether there are significant differences in the metrics across weight classes. Despite the skewness in the data, a square root transformation is applied to both takedown and knockdown data to attempt to normalize their distributions for more accurate statistical testing. However, the knockdown data remained non-normal (Takedown data appeared closer to a normal distribution after transformation but still not enough to be considered normal as they were bimodal). The transformed graphs are presented above. Finally, the Games-Howell post-hoc test is conducted due to unequal sample sizes and variance to identify specific weight classes between which the differences are significant.

Significance of Results:

The Games-Howell post-hoc analysis provides us with detailed insights into the differences between weight classes for both takedown percentages and knockdown averages. For takedown percentages, the analysis reveals a statistically significant difference between the Welterweight and Middleweight classes ($p = 0.002858$), indicating that Welterweights have a higher average takedown percentage compared to Middleweights. Additionally, there is a significant difference between Bantamweight and Welterweight classes ($p = 0.047912$), as well as a significant difference between Lightweights and Middleweights ($p = 0.008412$). This shows that Bantamweights have a lower average takedown percentage than Welterweights and Lightweights have a higher mean takedown percentage than Middleweights. In terms of knockdown averages, significant differences are observed between several weight classes. Notably, a significant difference exists between Featherweights and Welterweights ($p = 3.397863e-04$), with Welterweights having a higher average knockdown rate, and a significant difference between Lightweights and Welterweights ($p = 1.438997e-09$), with Welterweights having a higher average knockdown rate. Also, there is a significant difference between Middleweights and Lightweights, with Middleweights having a higher average knockdown rate. Finally, a notable significant difference lies between Bantamweights and Lightweights, with Bantamweights having a higher average knockdown rate.

Best Striker

Goal:

The aim of this analysis is to identify the best strikers in the dataset by examining three key metrics: Knockdowns (KD), Significant Strikes Landed (SIG_STR_landed), and Significant Strike Percentage (SIG_STR_pct). Each metric is weighted according to its perceived impact on a fighter's striking ability.

Further Preprocessing:

We begin with the “preprocessed_Data.csv” dataset containing the relevant statistics per match. The dataset is processed to isolate the relevant striking metrics, which are then normalized (using **MinMaxScaler**) to ensure that each metric contributes equally to the final score. The normalization process is crucial as it allows for a fair comparison across fighters with varying numbers of fights and striking attempts.

Approach:

The analysis proceeds by calculating a composite 'striker_score' for each fighter, derived from the weighted sum of the normalized metrics. The weights reflect the importance of each metric in effective striking:

- **Knockdowns (KD)**: As a clear indicator of a fighter's power, it's given the highest weight of 0.50.
- **Significant Strikes Landed (SIG_STR_landed)**: This metric reflects the volume of effective strikes, weighted at 0.30.
- **Significant Strike Percentage (SIG_STR_pct)**: Representing the accuracy of a fighter, it's assigned a weight of 0.20.

Significance of Results:

The 'striker_score' provides a quantitative measure to rank the fighters, highlighting those with the most effective striking capabilities. The top-ranking strikers, as determined by the highest composite scores, are those who not only land strikes frequently but do so with power and precision. This approach to ranking allows us to identify fighters who are not just active but also efficient and impactful with their striking in matches. The 'striker_score' leads us to our top strikers, with Brad Blackburn leading the list with a score of 0.730852, followed by Yui Chul Nam at 0.702142, and Chris Daukaus at 0.606671. The scores reflect the effectiveness of these fighters in the striking domain, combining power, precision, and volume.

	fighter	striker_score
227	Brad Blackburn	0.730852
2115	Yui Chul Nam	0.702142
346	Chris Daukaus	0.606671
1610	Petr Yan	0.597138
812	Israel Adesanya	0.594659
1378	Matheus Nicolau	0.584158
667	Forrest Petz	0.569491
1940	Tarec Saffiedine	0.563964
1145	Karol Rosa	0.561849
1856	Shane Burgos	0.540193

Best Grappler

Goal:

The goal of this analysis is to identify the most proficient grapplers in the dataset. The metrics used to evaluate grappling proficiency are Takedowns Landed (TD_landed), Takedown Accuracy (TD_pct), Submission Attempts (SUB_ATT), and Reversals (REV). Each metric is assigned a weight based on its importance in grappling, with the intention to combine them into a single 'grappler_score' that represents the overall grappling ability of a fighter.

Further Preprocessing:

We begin with the “preprocessed_Data.csv” dataset containing the relevant statistics per match. The data is preprocessed to focus on grappling metrics, and each metric is extracted for both the red and blue corners of the fighters. After aggregating these metrics, the data is normalized using **MinMaxScaler** to ensure each metric contributes equally to the final score, regardless of the original scale.

Approach:

We assign the following weights to the grappling metrics based on their perceived impact on a fight:

- **Takedowns Landed (TD_landed)**: This shows how often a fighter can bring the opponent to the ground, weighted most heavily at 0.40.
- **Takedown Accuracy (TD_pct)**: This reflects how efficiently a fighter can execute takedowns, weighted at 0.30.
- **Submission Attempts (SUB_ATT)**: This indicates a fighter’s offensive grappling capability, weighted at 0.20.
- **Reversals (REV)**: This showcases a fighter's defensive grappling skills, weighted at 0.10.

Significance of Results:

The results yield a ranking of fighters based on their grappling prowess. The top 10 grapplers, as determined by the highest 'grappler_score', reveal which fighters excel in grappling within the dataset. The top-ranked grappler is Merab Dvalishvili, with an impressive score of 0.575641, reflecting his superior grappling capabilities within the selected pool of fighters. Notably, Georges St-Pierre, a renowned figure in mixed martial arts, ranks second with a grappler score of 0.539909, which highlights his well-known proficiency in takedowns and control on the ground.

	fighter	grappler_score
1418	Merab Dvalishvili	0.575641
709	Georges St-Pierre	0.539909
1715	Rodney Wallace	0.526052
186	Bartosz Fabinski	0.524857
1926	TJ Waldburger	0.506433
858	James Wilks	0.502042
1144	Karo Parisyan	0.474652
1942	Tatiana Suarez	0.468625
1656	Ray Borg	0.463153
1185	Kevin Randleman	0.463020

Best Pound-for-Pound Fighter

Goal:

The goal of this analysis is to ascertain the best pound-for-pound fighter by evaluating performance across multiple key metrics. These metrics include knockdowns, significant strike percentage, takedown percentage, submission attempts, reversals, and several others related to striking and grappling. This comprehensive assessment aims to provide a holistic view of a fighter's capabilities in the octagon, irrespective of weight class.

Further Preprocessing:

We begin with the “preprocessed_Data.csv” dataset containing the relevant statistics per match. We proceed by aggregating statistics from both the red and blue corners for each fighter. To address potential disparities in scale and impact across different metrics, we normalize the data using MinMaxScaler, ensuring that each metric contributes proportionately to the final P4P score.

Approach:

We assign the following weights to the pound-for-pound metrics based on their perceived impact on a fight:

- **Knockdowns (KD):** Given the highest weight due to its significant impact on the fight outcome, weight = 0.20.
- **Significant Strike Percentage (SIG_STR_pct):** Represents the accuracy of a fighter, weight = 0.10.
- **Takedown Percentage (TD_pct):** Indicates the ability to control the fight location, weight = 0.10.
- **Submission Attempts (SUB_ATT):** Shows a fighter's capability to finish fights, weight = 0.10.
- **Reversals (REV):** Demonstrates adaptability and skill in ground fighting, weight = 0.10.
- **Significant Strikes Attempted (SIG_STR_att):** Reflects a fighter's offensive volume, weight = 0.10.
- **Significant Strikes Landed (SIG_STR_landed):** Measures the effectiveness of a fighter's striking, weight = 0.10.
- **Total Strikes Attempted (TOTAL_STR_att):** Captures the fighter's overall activity rate, weight = 0.05.
- **Total Strikes Landed (TOTAL_STR_landed):** Indicates overall striking success, weight = 0.05.
- **Takedown Attempts (TD_att):** Reflects the initiative to change the fight's dynamics, weight = 0.05.
- **Takedowns Landed (TD_landed):** Represents successful execution of control, weight = 0.05.

Significance of Results:

The P4P scores yield a ranking of fighters, with the top positions indicating those who excel across the board in the selected metrics. It allows for a comparison that transcends weight divisions, offering a unique perspective on who the truly dominant fighters are when size and weight advantages are conceptually removed. The results indicate that the fighter with the highest P4P score in this subset is Yui Chul Nam, with a score of approximately 0.464839. This score suggests that across the metrics considered, Yui Chul Nam has the most balanced and effective performance profile among the fighters listed. Nicco Montano follows with a P4P score of approximately 0.452031, and Georges St-Pierre, a well-known name in MMA, ranks third in this subset with a P4P score of approximately 0.428812.

	fighter	KD	SIG_STR_pct	TD_pct	...	TOTAL_STR_landed	TD_att	TD_landed	p4p_score
2115	Yui Chul Nam	0.178707	0.061983	0.085063	...	0.025551	0.004559	0.006932	0.464839
1520	Nicco Montano	0.002686	0.046466	0.045161	...	0.049591	0.011707	0.012421	0.450231
709	Georges St-Pierre	0.043265	0.067123	0.087715	...	0.038927	0.016053	0.026584	0.428812
1145	Karol Rosa	0.061153	0.054483	0.019048	...	0.045517	0.003538	0.003736	0.419654
1378	Matheus Nicolau	0.142819	0.060159	0.053779	...	0.020469	0.007648	0.011212	0.410584
214	Billy Quarantillo	0.016135	0.087974	0.081512	...	0.035742	0.006346	0.008148	0.406866
227	Brad Blackburn	0.200000	0.049491	0.004424	...	0.022861	0.001980	0.000790	0.401885
1367	Martin Day	0.074600	0.060237	0.076498	...	0.026531	0.003282	0.004708	0.399366
1610	Petr Yan	0.125880	0.060828	0.039526	...	0.025542	0.005837	0.006715	0.395994
1266	Luana Carolina	0.062087	0.052672	0.013825	...	0.028284	0.002516	0.002415	0.390432

Machine Learning Models

Fight Result Predictor

Goal:

Our goal in this part of the project is to develop an ML model that is able to predict the winner of the match given the fight statistics of both the red and blue fighter

Approach:

The k-nearest neighbour (KNN) model was used to aid us in predicting the fight result. We split our new dataset into X and y values, where X contains the numerical statistics for both the fighters and y contains solely the winner of the fight (red or blue). Further, we transform our X values with MinMaxScaler to ensure that a singular feature cannot dominate the prediction.

Results:

Our model often produces an accuracy score in the range of 0.64 and 0.70. This means that it was able to classify the winner of fights in the validation data with an accuracy of 64-70%. We consider this an impressive result taking into account the inherent unpredictability of the MMA sport.

Win Ratio Predictor

Goal:

Our goal in this part of the project is to develop an ML model that is able to predict a fighter's win ratio given their fight statistics and information about their build (height, reach, etc.)

Further Preprocessing:

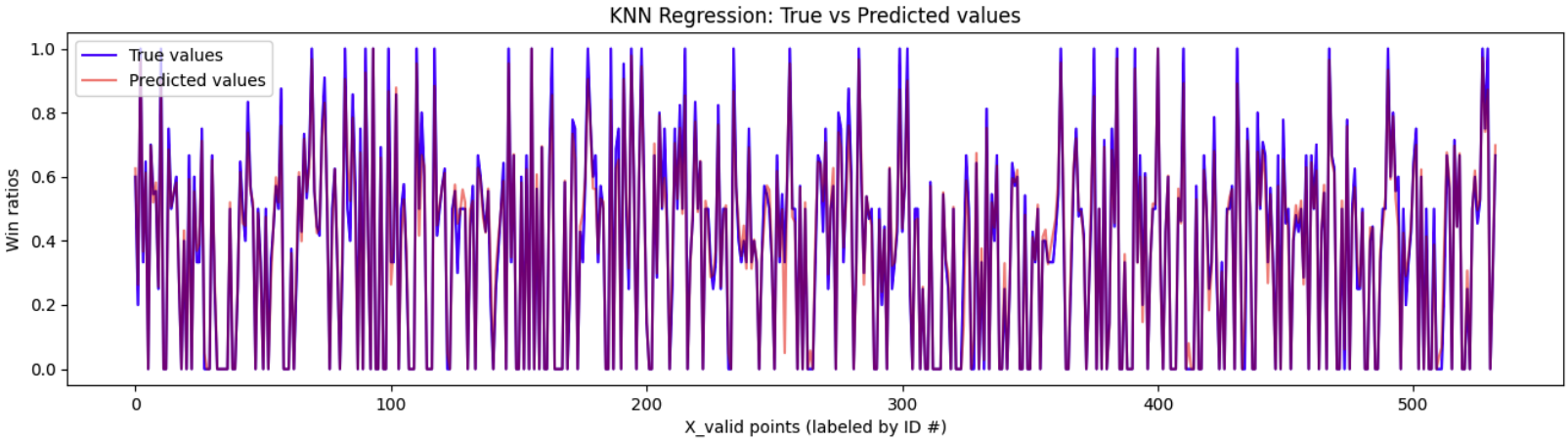
We start by accumulating each fighter's wins and losses from the `preprocessed_data.csv` data set to calculate a win ratio. Next, we perform a join on the `preprocessed_fighter_details` to create a data set that contains a fighter's fight statistics, information on their build and their win ratio. Now, we have prepared a dataset to feed to our model.

Approach:

The k-nearest neighbour (KNN) model was used to aid us in predicting win ratios. We split our new dataset into X and y values, where X contains the numerical statistics for the fighter and y contains solely their win ratio. Further, we transform our X values with MinMaxScaler to ensure that a singular feature cannot dominate the prediction.

Results:

Our model often produces a coefficient of determination in the range of 0.96 and 0.98. This means it explains the variability in the data extremely well and is able to generalize new/unseen data nicely. A Plot of the predicted and true y values of the X validation data can be found below. Each X record was given an ID for plot-ability and presented with its true and predicted values. A continuous plot was used rather than a scatter plot to easily see overlap.



Project Experience Summary

Yousef Haiba:

- Examined the dataset to combine significant strike accuracy statistics from Red and Blue fighters, ensuring a complete and unduplicated dataset for analysis across different UFC weight classes.
- Utilized Levene's test to assess variance equality among groups, followed by Kruskal-Wallis and Games-Howell tests as unequal variance was detected, confirming significant differences in strike accuracy between weight classes.
- Reviewed the dataset to aggregate win-by-method data for Red and Blue fighters, enabling the analysis of win-type distribution across weight classes.
- Applied the Chi-Square Test of independence on a contingency table of win types by weight class, which indicated a strong statistical association, with significant variations in finish methods among different weight classes.

Jason Gill:

- Developed a machine learning model using k-nearest neighbours (KNN) to predict the winner of MMA matches based on fighter statistics, achieving an accuracy score of around 70% on validation data.
- Developed a machine learning model utilizing KNN to predict fighter win ratios, based on a fighter's career statistics and build information, achieving a coefficient of determination greater than 0.97.
- Determined the most effective strikes in UFC fights that contribute to success through statistical means using Levene's test, Kruskal-Wallis test and Games-Howell posthoc test
- Determined the most effective stance in UFC fights that contributes to success through statistical means using Levene's test, ANOVA test and Games-Howell posthoc test
- Determined which colour (red or blue) wins more and tested for significance, employing statistical tools such as Levene's test and a t-test
- Created informative plots that represent data distributions of sample groups using the matplotlib and seaborn libraries
- Successfully handled missing numerical values in the dataset by imputing them with median values, after ensuring it was representative of the overall dataset.
- Employed MinMaxScaler to ensure certain training features for the ML models didn't dominate the predictions
- Conducted precise data transformation for metrics represented in different units, facilitating standardized analysis across datasets.
- Implemented a reusable function that accurately calculates win ratios by aggregating wins and losses from the `preprocessed_data.csv` dataset
- Implemented a reusable function that extracts information regarding each fighter's strike data from each match from the `preprocessed_fighter_details.csv`
- Used the pandas and numpy libraries for general data manipulation
- Created an informative README that contains information regarding required libraries, order of execution, files produced/expected, contributors, etc.

Vishaal Bharadwaj:

- Conducted a comprehensive statistical analysis to identify the best striker in UFC, assigning weights to striking attributes such as Knockdowns, Significant Strikes Landed, and Significant Strike Percentage.
- Analyzed grappling data to pinpoint the best grappler in UFC by evaluating performance metrics including Takedowns Landed, Takedown Accuracy, Submission Attempts, and Reversals.
- Developed a robust model to establish the best pound-for-pound fighter, utilizing a weighted sum model that integrated a variety of fighting statistics, normalized through MinMaxScaler for equitable contribution across metrics.
- Explored the variation in average takedown percentages and knockdown values across different weight classes, applying the Kruskal-Wallis test and Games-Howell posthoc test to determine statistical significance.
- Integrated Levene's Test extensively to examine variance homogeneity across groups, ensuring the appropriateness of subsequent statistical tests and the reliability of the analysis results.
- Created a reusable code structure that facilitated the calculation of fighter rankings and the examination of variance in fighting statistics, making extensive use of libraries like pandas, numpy, and sklearn for data processing and analysis.
- Generated kernel density plots using seaborn to visually represent the distribution of performance metrics across weight classes, aiding in the interpretation of statistical tests.