DSA210 Term Project

Motivation

I enjoy MMA and wanted to explore some stats just for fun. Initially, my goal was to analyze the impact of training backgrounds (like wrestling or jiu jitsu) on fight outcomes, but due to limitations in the available data, I talked with my TA and shifted focus. Instead I examined whether physical attributes such as height, reach, and dominant win methods correlate with success in MMA. The project became a fun way to apply data science techniques to something I'm interested in.

Data Collection

To construct the dataset used in this project, I implemented a custom web scraping and crawling framework for extracting MMA fighter profiles from sherdog.com and tapology.com.
The primary data collection tool was a recursive web crawler.
It begins at a single fighter's Sherdog profile URL and automatically traverses the network of opponents by visiting the links embedded in the fight history section.
Each individual fighter page is scraped for both summary and detailed information.
This includes metadata (name, nationality), physical attributes (height, reach, weight, weight class) and a full professional fight history.
Win/loss data is extracted both as summary statistics and through parsing individual fight results.
After collection, I used additional scripts to process and flatten the nested JSON structures into tabular formats suitable for analysis. The final data was exported into CSV format.

However: Both Tapology.com and Sherdog.com explicitly disallow web scraping and automated bots on their sites.
Tapology's policies ban any data scraping or bot access and Sherdog's terms similarly prohibit automated crawling beyond what a human user could do and forbid reusing scraped content, so I did ***NOT*** *post the code I used for data collection into github.*
(I can provide it if the instructor wishes to see.)

- Problems encountered during data collection: Unfortunately there is no gender information on these sites, so there was no way for me to get that data and do a gender based analysis.

Some additional information about mixed martial arts terminology that might be needed to make sense of the data analysis if you are unfamiliar with them:

- **Striking:** Fighting in striking arts takes place on the feet. (Strikes may include punches, kicks, elbows and knees. Emphasis is placed on accurate striking, footwork, and head movement.)

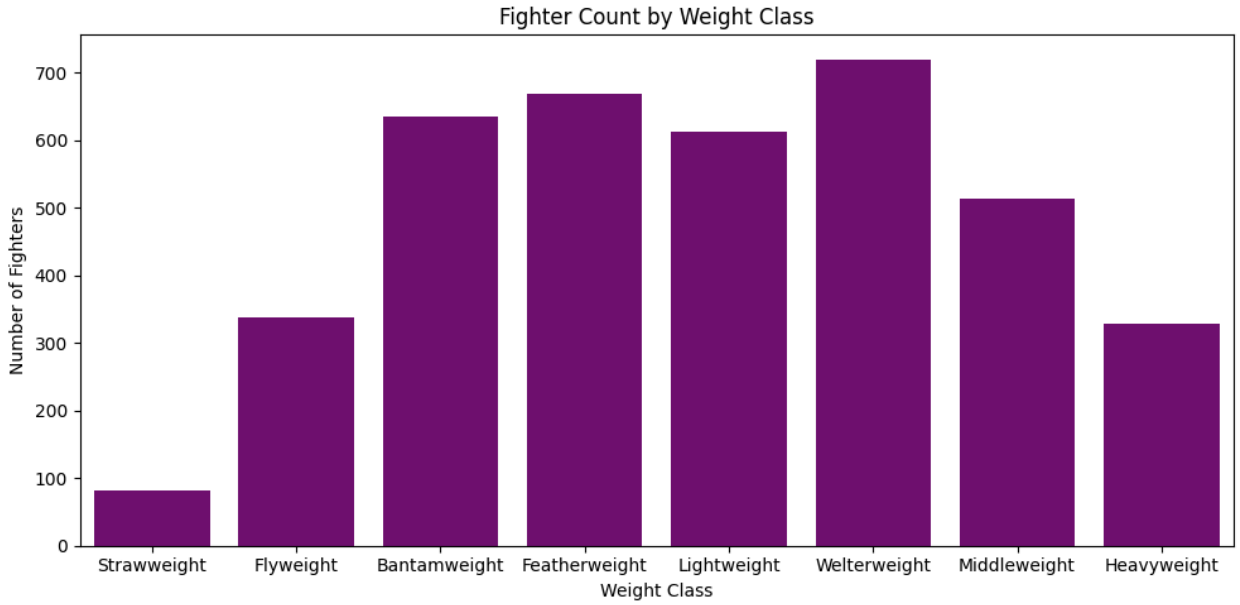
- **Grappling:** Grappling is focused on what to do when the fight is taken to the ground. (Grapplers focus on closing the distance, executing a takedown, and submitting their opponent. Submissions can happen in different ways such as arm bars, heel hooks, strangles, and shoulder locks.)

- **KO Win:** The losing fighter is unconscious and physically unable to continue.

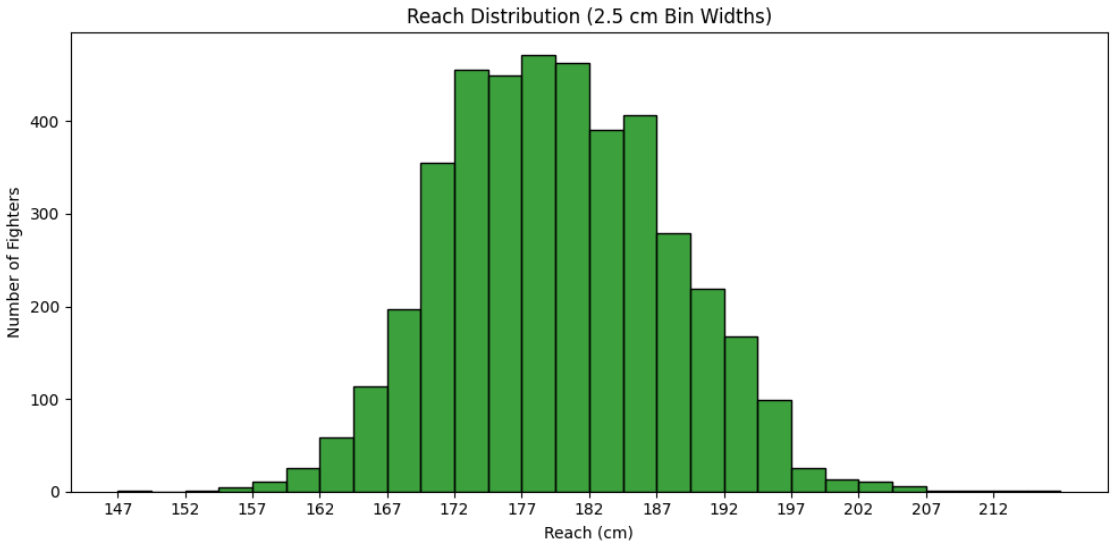
- **Submission Win:** When a fighter decides to quit fighting, or “tap out,” the fight ends.

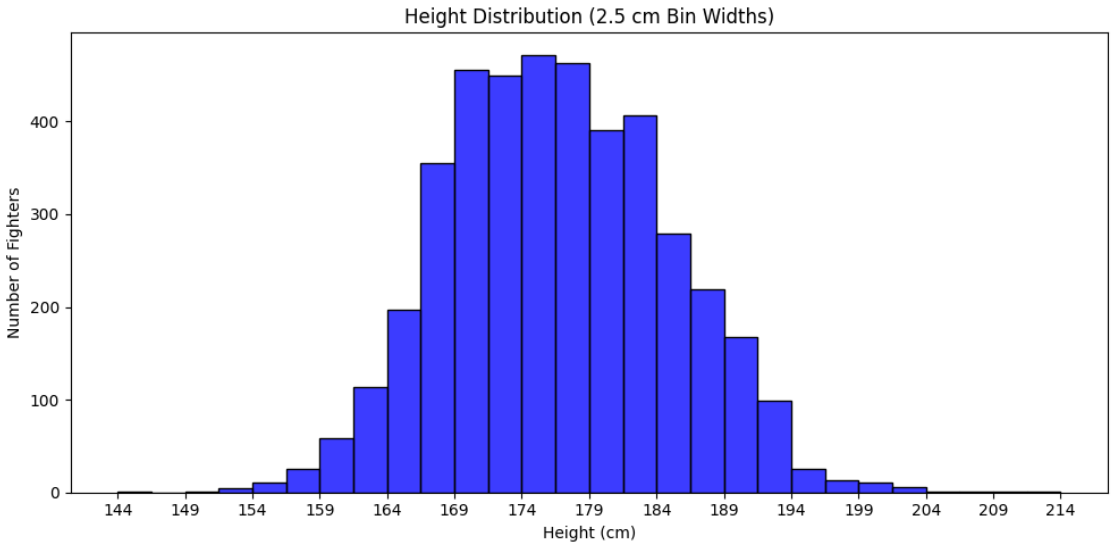
- **Decision Win:** Both fighters will have made it through the allotted number of rounds without getting knocked out or submitting. That leaves determining the winner in the hands of the judges, who score each round based on a number of factors, including punches landed.

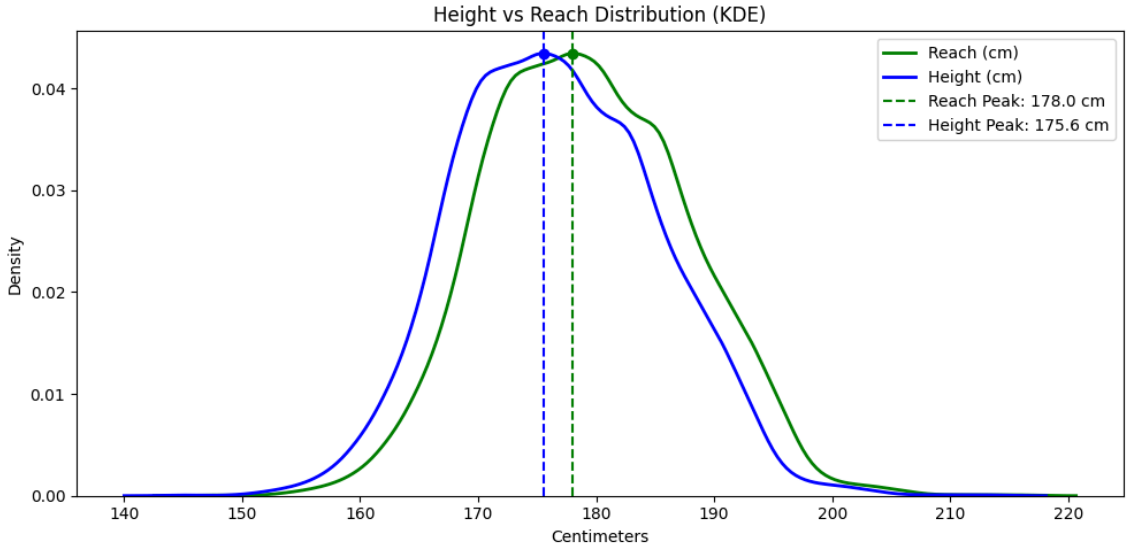
Exploratory Data Analysis



The bar chart displaying the number of fighters in each weight class reveals an imbalance across divisions.
The data reveals that the majority of fighters are concentrated in the Bantamweight, Featherweight, Lightweight, and Welterweight divisions. These divisions have the highest number of entries, reflecting both the most competitive and most populated ranges for professional athletes in MMA.

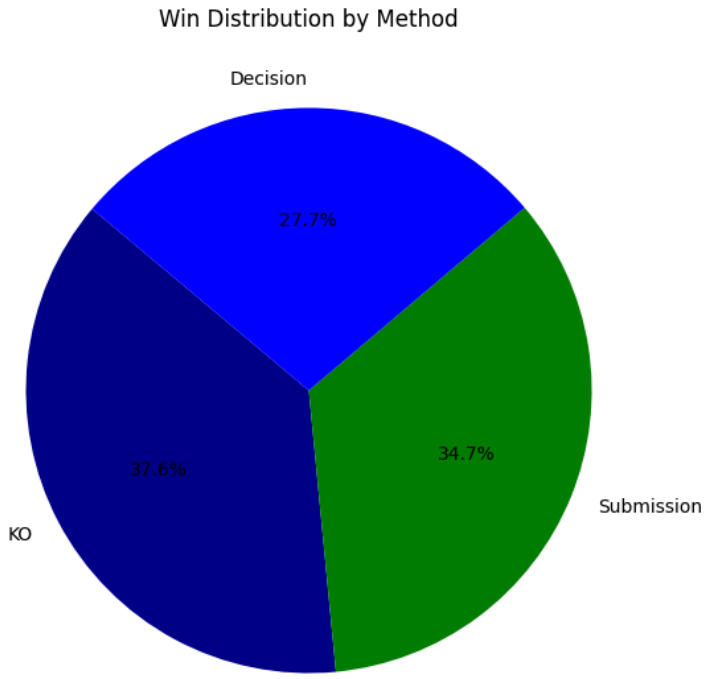

Reach Distribution (2.5 cm Bin Widths)
Number of Fighters
0
100
200
300
400
147
152
157
162
167
172
177
182
187
192
197
202
207
212
Reach (cm)

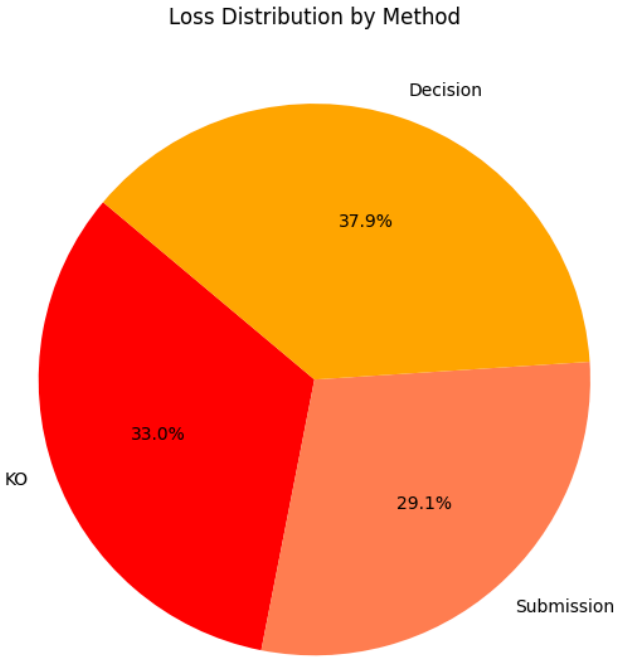

Height Distribution (2.5 cm Bin Widths)
Number of Fighters
0
100
200
300
400
144
149
154
159
164
169
174
179
184
189
194
199
204
209
214
Height (cm)


Height vs Reach Distribution (KDE)
Density
0.00
0.01
0.02
0.03
0.04
140
150
160
170
180
190
200
210
220
Centimeters
Reach (cm)
Height (cm)
Reach Peak: 178.0 cm
Height Peak: 175.6 cm

Reach: Measured from fingertip to fingertip when your arms are held parallel to the ground.

The reach and height distribution of fighters in the dataset displays a symmetrical bell shaped curve, suggesting that they are approximately normally distributed. This indicates that most MMA fighters share a common profile in terms of arm length and height. There are no significant spikes or gaps in the distribution. The reach curve shows a slightly broader peak and a more gradual tapering, suggesting marginally greater variability in reach than in height, but the difference is minimal (2.4 cm)
The overlap also reinforces that there are no extreme disparities between these two traits.



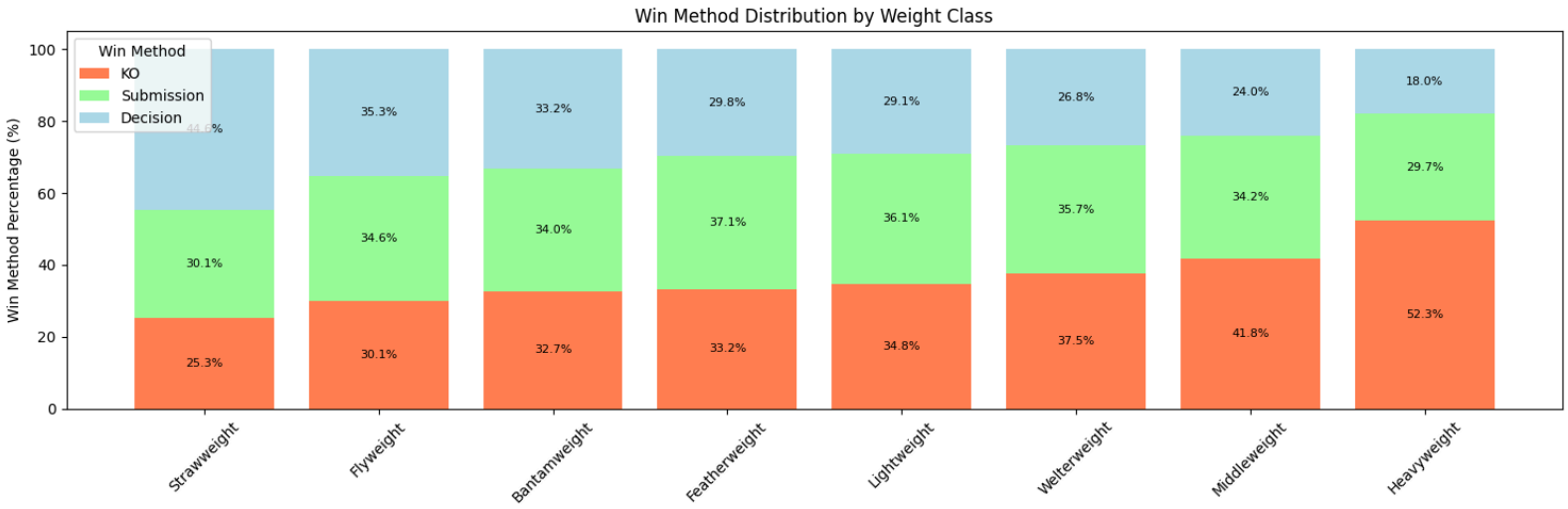


Decision losses are the most common loss type whereas it has the lowest distribution in wins. This revealed that while fighters secure wins through finishes, they more frequently lose via decision.

This could reflect that fighters take fewer risks as they lose control of a round, leading to more fights going the distance. It might also reflect that many losing fighters are still able to defend themselves effectively enough to avoid being stopped.

Another important factor might be the limitations of my dataset. Since I was only able to collect data on a subset of MMA fighters, specifically those with more complete and accessible records it's possible that I'm unintentionally excluding fighters who lose early in their careers and then stop competing. For instance, a fighter who loses their debut by KO and never returns to the sport would contribute to someone else's win record but may not appear in my dataset at all if they didn't have a consistent record on Sherdog.

This kind of dropout bias could lead to an underrepresentation of certain types of losses because the losing side of that outcome disappears from the data. As a result, the proportion of decision losses may appear inflated simply because those fighters are more likely to have longer careers and more complete data histories. So the mismatch between win and loss distributions might not only reflect fight dynamics but also be shaped by gaps in data.



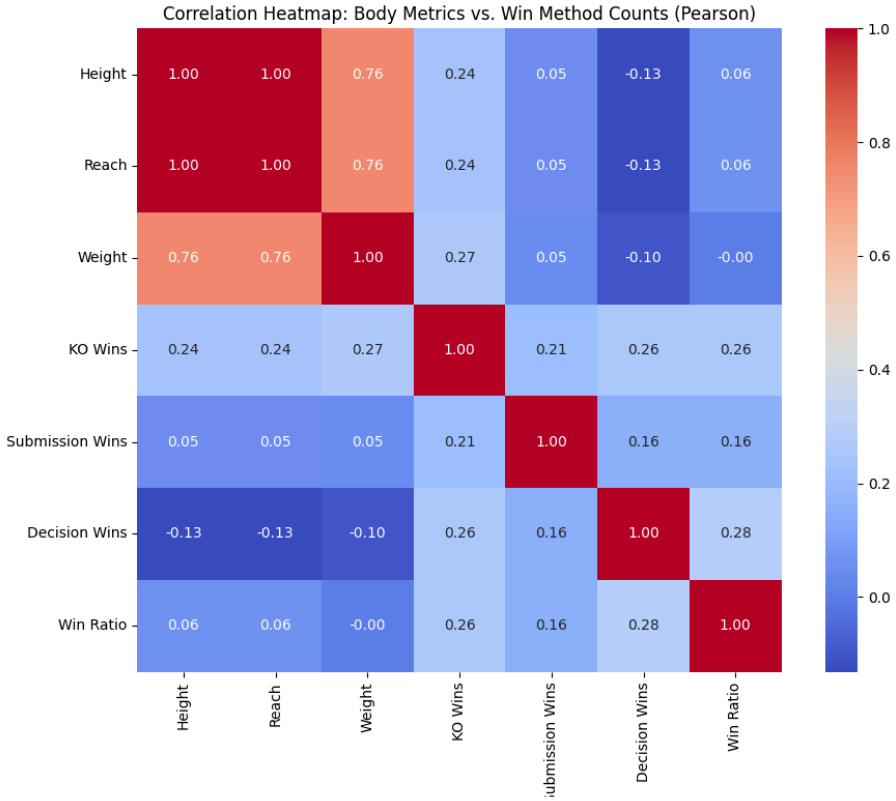
The distribution of win types illustrates a strong relationship between body size and fight outcomes: lighter fighters are more likely to win by decision or submission, while heavier fighters overwhelmingly favor knockouts.

In the lighter divisions decisions account for a large portion of wins as these fighters tend to rely more on speed, cardio and volume striking, which often leads to longer, closely contested fights that go to the judges.

Submission wins remain relatively stable across divisions, ranging between 30% and 37%. This suggests that grappling effectiveness is somewhat consistent, though it may begin to diminish slightly in higher weight classes.

While KO wins make up only a quarter of victories at Strawweight, they rise steadily through the classes and dominate in Heavyweight. This aligns with the advantage heavier fighters have in power striking, as well as the tendency for fights in these divisions to end suddenly rather than accumulate points over rounds. Also the number of female fighters decreased in the heavier weight classes, so this might also be an explanation.

Decision wins show the reverse trend. This indicates that the likelihood of a fight going the distance sharply decreases with size, supporting the idea that larger fighters have both more power to finish fights and potentially less cardio capacity to sustain prolonged rounds.



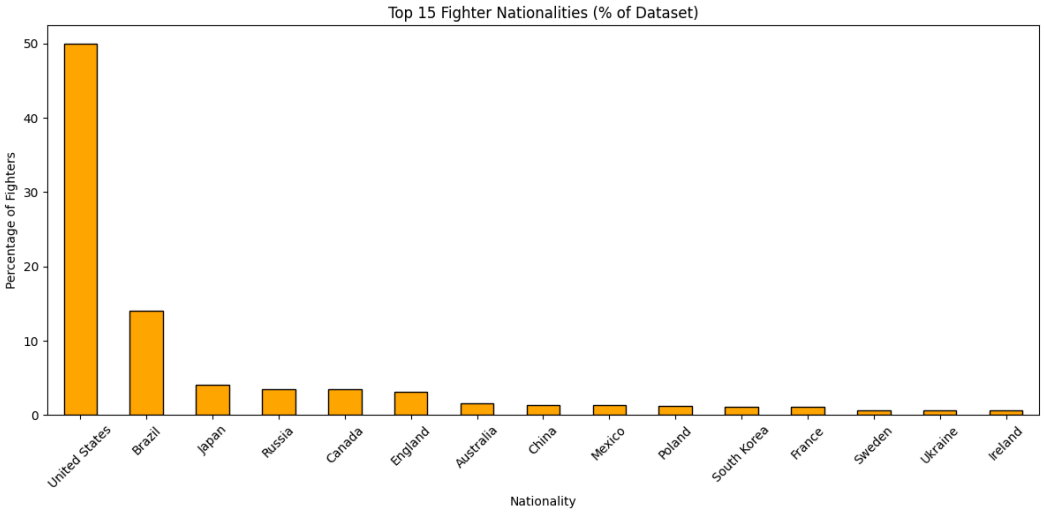
Initially I used Pearson correlation to explore the relationships between fighters' physical attributes and their fight outcomes. The results showed weak correlations between most variables. Since Pearson correlation measures only linear relationships and is sensitive to outliers and scale, I suspected it might be underestimating the strength of certain associations, especially if the data followed a non linear pattern instead.

To investigate this further, I applied Spearman correlation which ranks values and captures monotonic relationships rather than strictly linear ones. However after generating the Spearman heatmap I found that the results did not differ substantially from Pearson. Overall patterns and magnitudes of correlation remained largely the same.

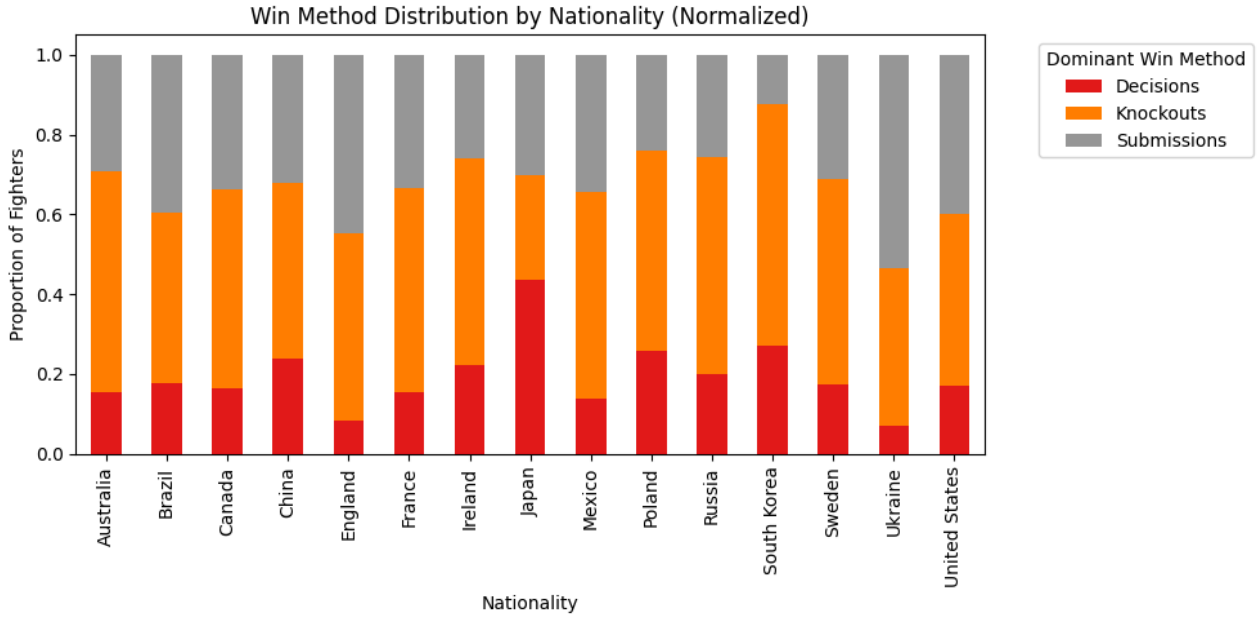
This similarity suggests that the underlying relationships between body metrics and win methods are genuinely weak in this dataset, rather than being distorted by outliers or scale effects.

(HOWEVER:
The heatmaps show that across the full population, weight and KO wins are only weakly related in an individual sense.

But, the Chi square test I did for my hypothesis shows that the proportion of fighters who rely on KOs, submissions, or decisions does vary significantly across weight classes. So while weight doesn't linearly predict KO count per fighter, fighting style distributions do differ by division, which is a broader categorical trend rather than an individual level prediction.)



The nationality distribution plot shows a highly skewed dataset. Nearly half of all fighters are from the United States. This imbalance means that any comparison by nationality would be statistically unreliable without normalizing it.



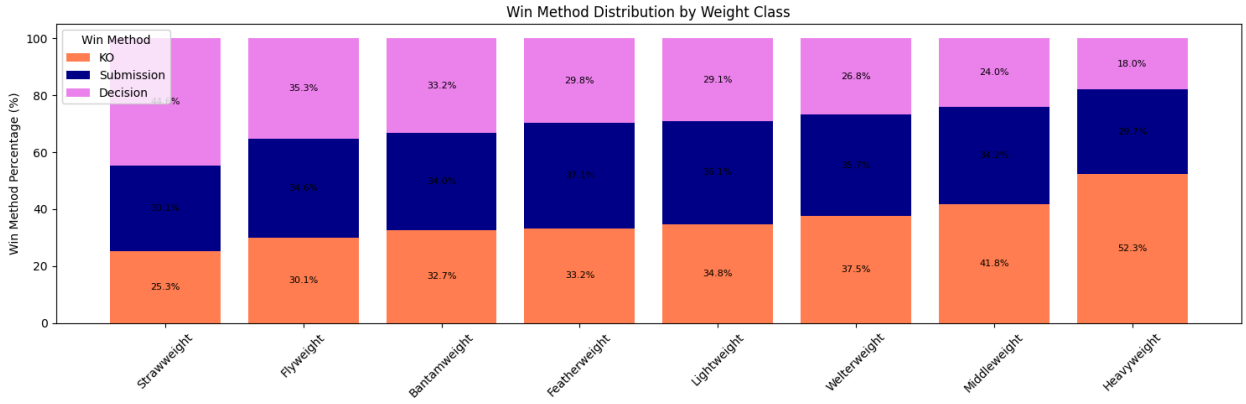
This visualization gives us a normalized comparison of win method preferences across the top 15 nationalities in the dataset accounting for population imbalance.
Japan has a very high proportion of fighters who predominantly win by decision, suggesting a pattern of fighting that leads to going the distance.
In contrast, Brazil and Russia show a strong tilt toward submissions which aligns with Brazil's jiu jitsu and Russia's sambo background. (Both are grappling heavy fighting styles)

Hypothesis Testing

Hypothesis 1

- **Null Hypothesis:** Fighter weight class is independent of dominant win method. (No relationship between weight class and preferred win method.)

- **Alternative Hypothesis:** Fighter weight class is associated with the dominant win method. (Certain weight classes favor certain win methods.)



I first identified each fighter's dominant win method by selecting the win type they achieved most frequently. I then grouped fighters by their primary weight class and constructed a contingency table comparing weight class to dominant win method. Using a Chi Square Test of Independence, I evaluated whether there was a significant association between these two categorical variables.

The results were:

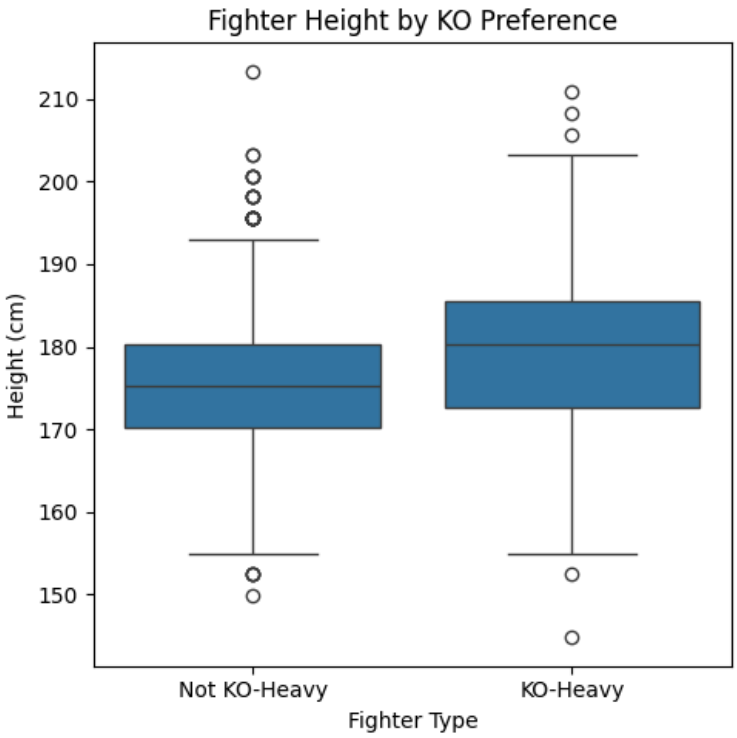
1. Chi-square statistic: 257.48
2. Degrees of freedom: 16
3. p-value: 1.51×10^{-45}

Given the small p-value, I rejected the null hypothesis.
Certain weight classes tend to favor certain types of victories.

Hypothesis 2

- **Null Hypothesis:** Tall and short fighters are equally likely to favor KOs.

- **Alternative Hypothesis:** Taller fighters are more likely to have knockouts as their dominant win method.



I tested the hypothesis, "Tall and short fighters are equally likely to favor KOs." using a two-sample t-test.
Fighters were grouped based on whether knockouts constituted at least 50% of their wins (KO-heavy group) or not (non-KO-heavy group).

The results were:

1. t-statistic = 12.82
2. p-value = 2×10^{-36}

Since the p value is below 0.05, I reject the null hypothesis.
The boxplot visually supports this conclusion: KO heavy fighters show a higher median height.