Identifying and Analyzing Patterns in Mixed Martial Arts

Joseph Yousef

Department of Computer Science

California State University, Long Beach

Long Beach, CA

[jyousef99@gmail.com](mailto:jyousef99@gmail.com)

***Abstract -* Mixed Martial Arts (MMA) has skyrocketed in popularity in recent years, largely as a result of the success of the UFC. Due to the fact that MMA is still a relatively new sport, as well as its unpredictable nature, there are many misconceptions about factors that influence success in the sport. The data analysis in this project aims to clarify many common beliefs about MMA, such as the claim that southpaw fighters win substantially more fights than orthodox fighters, with concrete proof. Using the R programming language and many packages that extend the language, this paper outlines the data analysis and visualization techniques that were used to generate insights about trends and patterns that influence a fighter’s success in MMA.**

***Keywords* - Data analysis, data preprocessing, data visualization, visualization.**

I. Introduction

Mixed Martial Arts (MMA) has skyrocketed in popularity over the years, namely due to the success of the UFC as its premier promotion, drawing millions of viewers across the globe. In an MMA bout, fighters are allowed to utilize striking and grappling techniques from various martial arts to defeat their opponents, which means there are many more factors in success than can be observed in a single-discipline competition. As a result, the sport of MMA is one that is highly unpredictable, and this has lead to many common misconceptions about the sport. Many claims exist about factors that influence a fighter’s success that are purely anecdotal, with no real evidence backing them. In this paper, I explain the steps taken in order to analyze MMA fight data with the goal of either proving or disproving claims about the sport, as well as attempting to identify factors that influence a fighter’s success in the sport.

II. Data Collection and Preprocessing

*A. Sourcing a Dataset*

The first challenge in performing the data analysis was finding a dataset that fit the project's needs. The UFC has made it a point to keep detailed track of statistics for every fight since the promotion’s first event in November of 1993, but unfortunately, the company decided to take down public access to the official data API many years ago. Hence, the only way to access fight data directly from the UFC is by navigating through their website to see statistics for each event. Scrolling through the UFC website fight by fight, even with the use of a web crawler to gather the necessary amount of data for this project, was impractical at best. Therefore, an alternative method of sourcing the data for the project was required. Eventually, I found a large dataset containing thousands of columns of data for every single fight in the UFC’s history [1] and decided to use that for the project.

*B. Reducing Dataset Size*

The first challenge that came with this dataset was that it was simply too large to run on a personal computer without causing RStudio or the computer to crash, owing to the fact that the dataset was about two and a half gigabytes in size, and in order to combat this issue some data had to be removed from the dataset.

The first attempt to reduce the dataset’s size was removing all fights before the year 2000; however, this resulted in a dataset that was still just under two gigabytes in size due to the fact that the UFC has increased the number of fights they host per year as time has gone on; therefore a drastically larger amount of fights occurred after the year 2000 than before it.

For the second attempt at reducing the dataset’s size, I took the more straightforward approach of cutting the number of rows in the dataset in half, keeping the latter half of the dataset. The remaining data was predictably just over one gigabyte in size, and although it was still large enough to crash RStudio occasionally, it was much more manageable than before. This approach also had the unexpected benefit of producing a dataset in which the earliest fights occurred in 2016, a year in which the UFC underwent a record-breaking rate of growth and a year that many consider the start of the “modern era” of the sport.

III. Data Analysis Techniques

*A. Software Tools Used*

The analysis in this project was performed using the R programming language, which has a rich ecosystem of packages that proved very helpful, particularly the Tidyverse collection of R packages. Specifically, this project made extensive use of the dplyr and ggplot2 packages, which helped with data manipulation and visualization, respectively.

During data analysis in the project, the dplyr package proved invaluable for calculating statistics to gain insights on the dataset. Some frequently used functions from the dplyr package that were used during analysis were the group\_by and summarize functions, which allowed for calculating aggregations on the dataset; the mutate function, which allowed for new columns to be generated based on the values of rows in the dataset; and the filter function, for cleaning data. Another handy feature from the dplyr package was the “pipe” operator, which takes the output from one function and passes it into another function as input. By combining the previously mentioned features from the dplyr library, statistics such as finding the average amount of punches per knockout for each fighter in the UFC were simple to calculate.

The ggplot2 package was used throughout the project to create visualizations that reflected the insights generated from the data. The ggplot2 library contains most of the functionalities required for the project by default; however, extension packages occasionally had to be installed to support additional functionalities, such as waffle, for creating waffle plots; ggrepel, which ensures text labels in plots do not overlap; and RColorBrewer, which provides alternative color palettes for plots.

While it wasn’t used as frequently as dplyr and ggplot2, the data.table package played a nontrivial role in the development of the project. Initial attempts to load the project dataset into an R data frame using Tidyverse’s read\_csv function would result in either RStudio or my computer crashing. Further research into the issue revealed that the read\_csv function has been benchmarked to be drastically outperformed by the fread function found in the data.table library in terms of speed and computational efficiency [2]. After installing the data.table package, the fread function successfully loaded the project dataset into a data frame.

*B. Filtering by Experience*

A large portion of this project involved aggregating statistics from a fighter’s career and ranking them based on the resulting score, so it became evident early on that some steps would have to be taken to ensure the quality of this aggregated score.

We look at the case of Dean Barry's UFC career to demonstrate the potential for inaccurate conclusions if these aggregated scores were blindly followed. Barry’s first fight in the UFC took place in April 2022, and slightly over the halfway mark of the first round Barry lost via disqualification after an eye poke that resulted in loss of vision for his opponent. Barry was cut from the promotion immediately following the fight and has not won another bout ever since; however, Barry was able to defend the one takedown his opponent attempted in those two and a half minutes, which leaves his career UFC takedown defense at a very impressive one hundred percent.

If we were to trust the results of aggregated scores without verifying their validity, the highest-ranked fighters for every insight we attempt to make would consist of misleading cases such as Dean Barry rather than the long-term consistency we can use to make accurate predictions. In order to achieve this, aggregated scores were filtered so that they only included fighters with four or more fights under their belt; since fighters signing with the UFC for the first time are usually given a three-fight deal, this ensures the data kept is that of competitors who have made it past their “rookie” contract, and therefore should have a proven record of consistent performances.

IV. Scrutinizing Anecdotal Claims

*A. Stance’s Effect on Win Rate*

The first step towards proving or disproving commonplace fighting beliefs was to see whether the claim that southpaw fighters win more fights than orthodox fighters had any truth. The reasoning behind this claim is that since southpaws are far less common than orthodox stance fighters, a southpaw will have more experience fighting orthodox fighters than vice versa, making the southpaw more likely to win the fight.

To start this analysis on the stances of fighters, I used dplyr to count the number of wins and the number of fights for each stance. I then divided the wins by the number of total fights to create a column representing the win percentage for each stance. Two things in the data frame of aggregated data stood out to me: the fact that the UFC keeps track of fighters who fight in both stances throughout a bout and labels their stance as “switch” and the fact that 21 fighters were never assigned any data about their stance at all, and instead had an empty string. After removing the row representing stance-less fighters from the data frame, I used the bar plot shown in Figure 1 to visualize and compare the win percentage for each stance. While the win percentage of southpaws was indeed higher than that of orthodox fighters, it was only by a single percentage point, and switch stance fighters had a win percentage that was two points higher than that of southpaws.

*B. The Concept of Ring Rust*

Another common adage in combat sports is the belief in “ring rust,” the idea that a fighter returning to competition after a long hiatus will experience a drop in performance, even if they were consistently training during their break from competing.

The project dataset includes a column named “days\_since\_last\_comp,” which indicates the number of days that have passed since the fighter’s last competition. To analyze how this break affects performance, I used the dplyr group\_by and summarize functions to determine the number of wins and total fights for each value in the days\_since\_last\_comp column. Then, I applied the mutate function to divide the two columns and calculate the win percentage. After filtering the data frame to eliminate rows representing an insignificant quantity of fights, I created a scatter plot with the remaining data. I then applied a simple linear regression line to help visualize a pattern in the data, as shown in Figure 2. The descent of the linear regression line isn’t steep, but it is noticeable enough to confidently say that the idea of ring rust contains some truth.

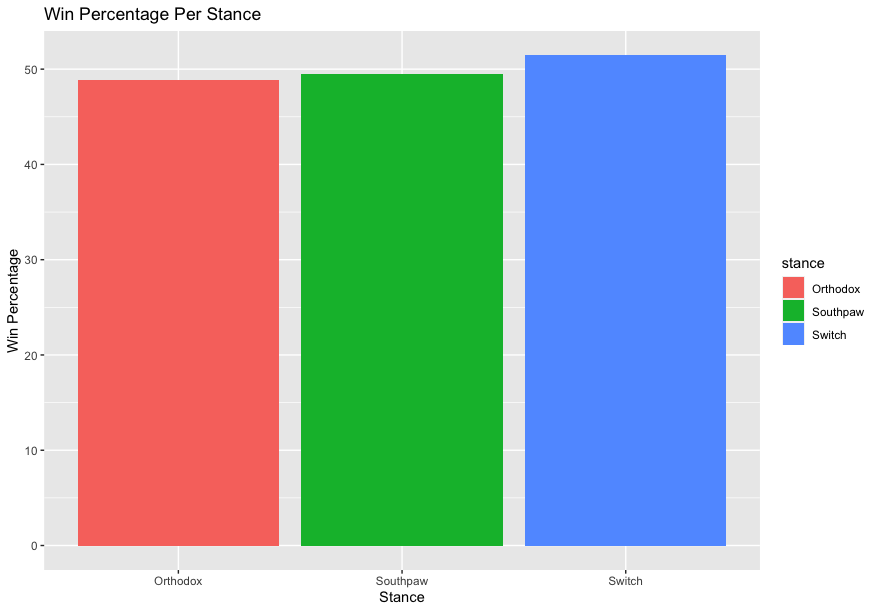


Fig 1. Bar plot visualizing the win percentage for each stance in the UFC

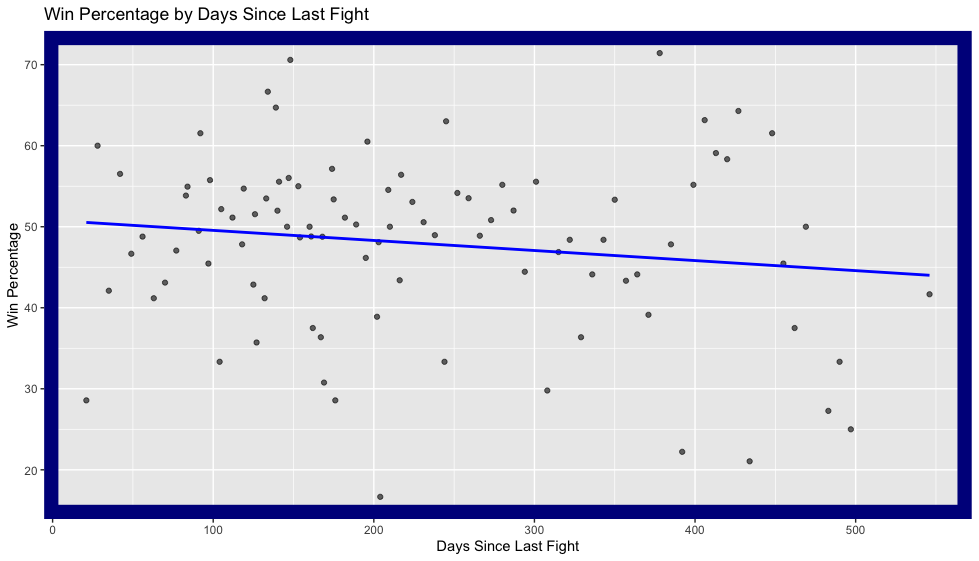


Fig 2. Scatter plot representing the win percentage for fighters going into a match based on the number of days since their last bout

*C. The Importance of Reach Advantages*

One statistic that tends to carry a lot of weight is a fighter’s reach, specifically whether they have a reach advantage over their opponent. There are many claims that having a reach advantage is invaluable during a fight, and just as many claims that reach advantages are overrated.

In the project dataset, the “reach\_differential” column can be used to determine whether a fighter is at a reach advantage or disadvantage for any particular fight, any fighter with a reach\_differential under one is at a reach disadvantage for that fight, and if their reach\_differential is over one, they have a reach advantage instead. Knowing this, I split the dataset into two data frames, one containing data on fighters at a reach disadvantage for a particular fight and another for the fighters who had the reach advantage instead. Similarly to previous analyses, I used the dplyr package to calculate a win percentage for both data frames, but this time with the added step of merging the data frames back together using dplyr’s bind\_rows function, making sure to add a new column detailing the data frame it came from. Creating a bar plot from the re-merged dataset provides the visualization in Figure 3, which supports the claim that a reach advantage indicates success in a fight, albeit very slightly, because the win percentages differ by less than five percent.

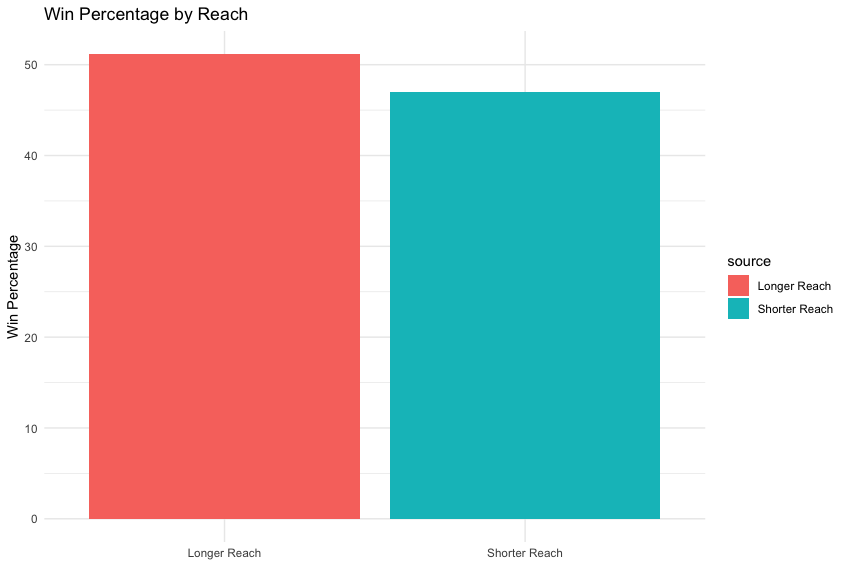


Fig 3. Bar plot visualizing win rates for fighters with reach advantages and reach disadvantages in a fight.

While Figure 3 showed a *slight* increase in the likelihood of victory for fighters with a reach advantage, it wasn’t as dominant of an advantage as many have claimed throughout combat sports history, so I explored the data further to see what else could be uncovered.

Taking the previous analysis one step further than just comparing a fighter’s reach to their opponent, I compared against the average reach in their respective weight division, which was calculated by using dplyr to group on the division column and find the mean of the reach column. After finding the average reach for each division, I mutated the project dataset to include this information, and then, similarly to the process that was used in the previous analysis, I split the dataset into separate data frames for fighters with reaches above and below their division’s average reach, then calculated win percentages for both data frames and merged them back together. I repeated the previous steps, filtering the data to include only championship fights, and then I created bar plots with both data sets to create Figure 4 and Figure 5. Figure 4 surprisingly displayed an even smaller difference in win rates than Figure 3, but Figure 5 boasted a thirteen percent higher win rate for fighters with a reach advantage. One possible explanation is that championship-caliber fighters are more likely to be physically gifted.

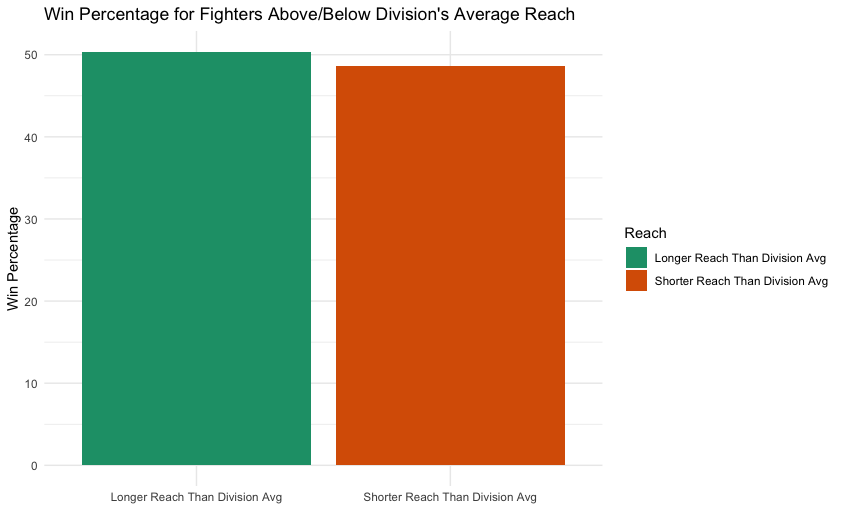


Fig 4. Bar plot showing win rates for fighters with reach above and below their division’s average reach

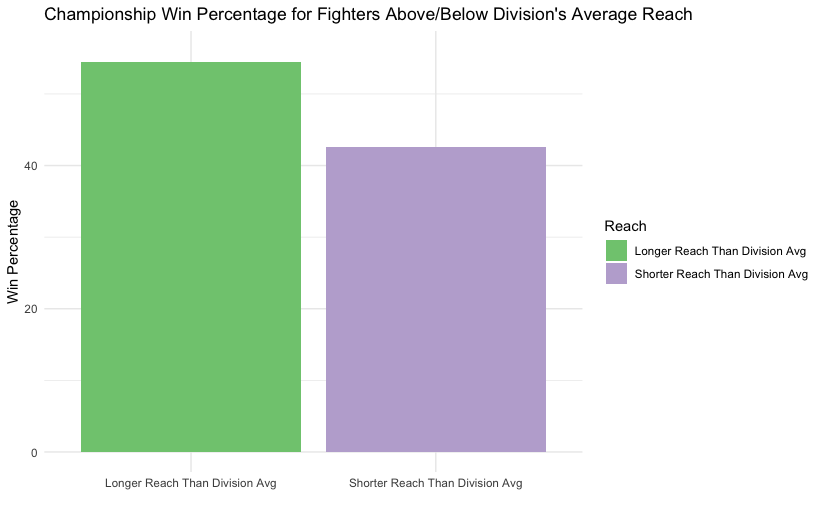


Fig 5. Bar plot showing win rates for fighters in championship matches with reach above and below their division’s average reach.

V. Analyzing Success

Mixed Martial Arts is a highly unpredictable sport due to the wide variety of tactics a fighter can choose to employ at any moment in a bout. While it may be nearly impossible to predict exactly how a fight will go accurately, I could still identify some critical factors in the project dataset that are consistently present in successful fighters. In this section, I will start with general trends and patterns in the dataset and progress to individual-level traits that the highest level of competitors possess.

*A. Methods of Victory*

In Mixed Martial Arts, competitors aim to beat their opponent through one of three methods: knockout, submission, or a judge’s decision. Since my goal was to identify patterns in the dataset that influence success, determining the distribution of the methods of victory seemed like a good start.

Using the dplyr package, I filtered the dataset only to include rows for fighters who won a fight and counted the number of rows associated with each method of victory. I then used the waffle package, an extension package for ggplot2, to create a waffle plot that visualized the number of fights that ended in each method of victory, as seen in Figure 6. Most fights went to a judge’s decision, with a smaller portion ending in a knockout and a much smaller amount ending in a submission.

One possible explanation for this distribution could be the result of rising competition levels in the UFC. Current-day fighters are very well educated on grappling fundamentals, even if it isn’t their primary fighting style, which results in fewer fights that end due to a completely unexpected submission. Fights ending in knockout are slightly more common than fights ending in submission since all fights start with both competitors on their feet with no other option than to exchange at least a few strikes, leading to some lucky victories. However, most fighters aren’t getting lucky; instead, they win fights by outperforming their opponent on a technical level and winning on the judge’s scorecards.

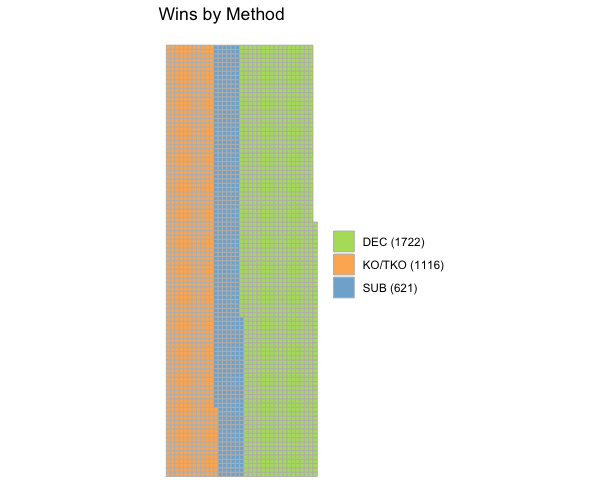


Fig 6. A waffle plot that displays the number of victories achieved by each of the three methods of victory.

*B. Fighter Age*

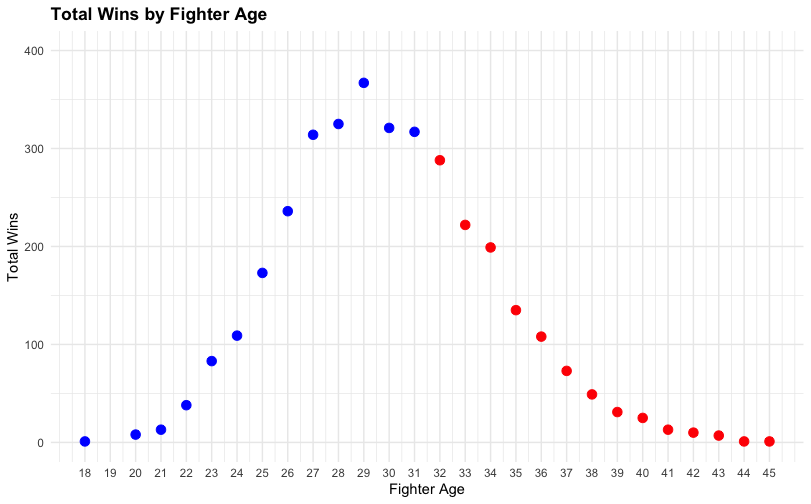
Continuing the theme of experience being a significant factor in success, it stands to reason that fighter age also has a relationship with success since an older fighter has had more time training and competing, making them more seasoned and capable. However, over time, these athletes will eventually physically deteriorate until they reach a point where their experience level is not enough to compensate. To explore the topic further, I used the dplyr package to calculate wins and the total number of fights for each age group that has competed in the UFC’s history. In Figure 7 and Figure 8, I visualize the total number of fights won by each age group and the win percentage for each age group, respectively. The scatter plot in Figure 7 shows the physical deterioration process mentioned earlier in this paragraph; as you can see, the number of fighters drastically drops as age increases. One interesting observation is that although the total wins plot in Figure 7 is shaped like a bell curve, with the late 20s and early 30s being the peak, the win percentage plot in Figure 8 shows a general downward trend. An explanation for this could be that fighters on the far ends of the total wins plot must be exceptionally talented to even be in the UFC, causing a higher win percentage rate.  
  


Fig 7. Scatter plot of total number of wins for each fighter age.

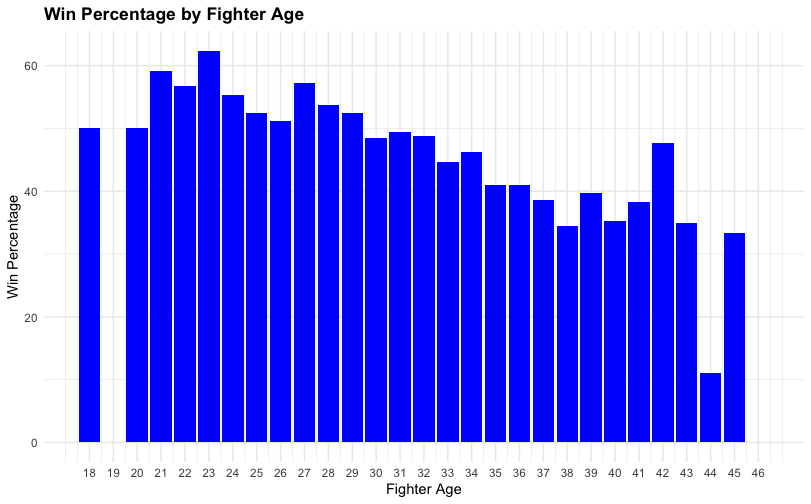


Fig 8. Bar plot for win percentage for each age group.

*C. Power as a Statistic*

One of the most common answers you would receive in response to asking people about traits they believe a successful fighter possesses is probably power; however, quantifying power in MMA can be challenging. A heavyweight will almost definitely hit harder than a flyweight can, but are we to ignore a flyweight on a nine-fight knockout streak simply because of their weight class? To quantify power in fighters fairly, I used the dplyr package to count the number of head strikes landed for every fighter in the dataset and divided it by the number of knockout victories they’ve achieved. I then ranked fighters based on how low their head strikes landed per knockout ratio is since a fighter who routinely dispatches opponents with fewer punches than other fighters is probably physically strong. After creating this ranking, I created the bar plot in Figure 9 to visualize it.

Some of the fighters in this plot undoubtedly possess impressive amounts of power; however, many of the fighters that appear have not been notably dominant throughout their careers. Based on the results of this analysis, I believe that raw strength, while impressive, does not necessarily create a great fighter.

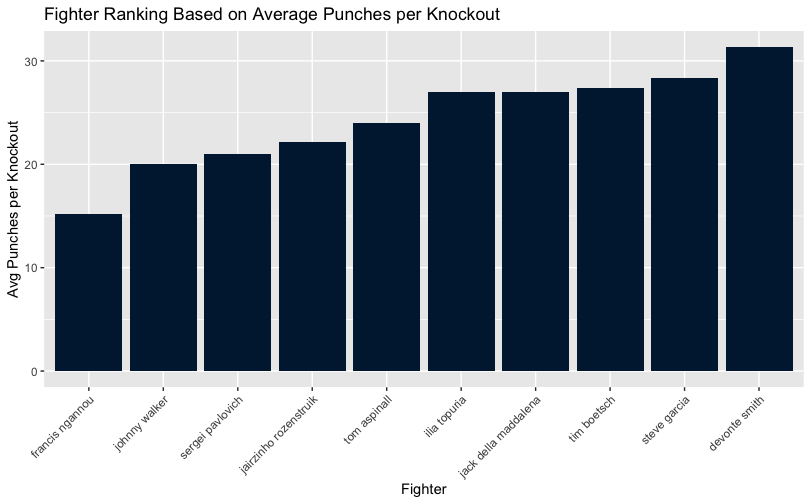


Fig 9. A bar plot visualizing a top ten ranking of fighters based on how few punches they required on average to knock their opponent out.

*D. Resilience*

One trait many successful fighters share is high resilience, particularly the ability to endure a large volume of strikes from an opponent without being finished. This hypothesis makes sense; if an opponent can not finish a fighter, the opponent has no choice but to endure more time fighting them, giving the fighter more chances to win.

Ranking fighters on their resilience is not a straightforward task. A fighter who absorbs many strikes without getting knocked out is very resilient, but should a fighter who has absorbed 700 strikes throughout their career and been knocked out once receive a lower ranking than a fighter who has only absorbed 150 strikes but never been knocked out? To factor for such scenarios, I ranked fighters based on a weighted combination of head strikes absorbed and losses via knockout, as seen in (1).

(1)

Figure 10 displays a bar plot that visualizes the top ten fighters in this resilience ranking system. Because every fighter in the bar plot has had a highly successful and impressive career, I would consider resilience a strong indicator of success in fighters.

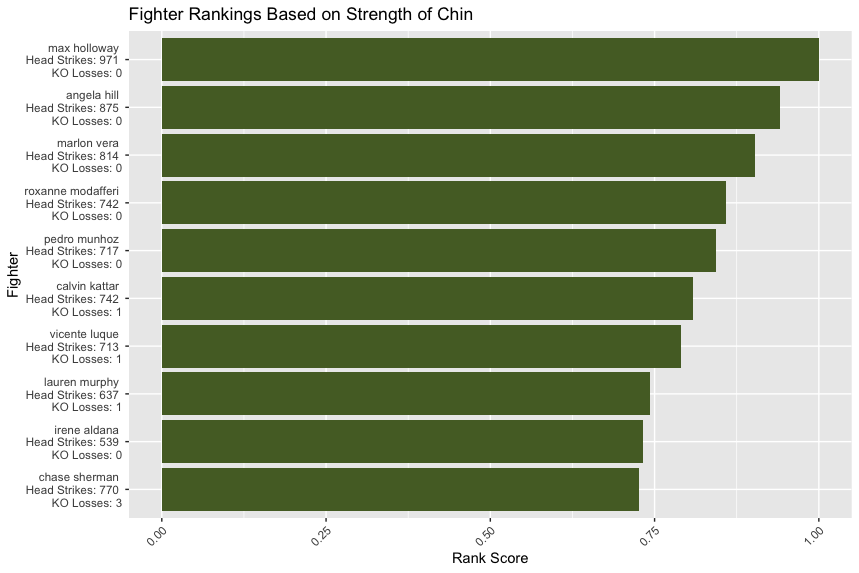


Fig 10. A bar plot visualizing a top ten ranking of fighters based on their resilience.

VI. Conclusion

In conclusion, the analysis performed throughout this project has served to provide concrete answers to some widely-held beliefs about fighting that were previously purely anecdotal sayings that were taken as truth. Although it is a challenging task to predict MMA fights with complete accuracy, I was able to achieve my goal of analyzing trends and patterns in the dataset in order to discover some valuable insights concerning traits that seem to influence a fighter’s success.

References

[1]“UFC/MMA Biggest Dataset With Differentials,” *www.kaggle.com*. https://www.kaggle.com/datasets/danmcinerney/mma-differentials-and-elo

‌[2]“How data.table’s fread can save you a lot of time and memory, and take input from shell commands - Jozef’s Rblog,” *jozef.io*. https://jozef.io/r917-fread-comparisons/ (accessed Dec. 12, 2023).

‌