Universidad Francisco Marroquin

Machine Learning Models

Christian Medina Armas

UFC DATA ANALYSIS

Victor Andre Farfán Miranda

20170473

Contents

Introduction

This report contains the analysis made to the UFC fight descriptions datasets. UFC is the biggest MMA organization of the world that combines different fighting styles as Kickboxing, Boxing, Jiu Jitsu, Karate, Wrestling, and others. Each fight can have 3 rounds or 5 rounds if it is a Title Match or just a Main Event. Fighters can win via Knockout, Technical Knockout, Submission, Judges decision or disqualification.

The dataset contains detailed information of each fight that occurred from 2013 to 2017 and contains every variable that can be analyzed during the fight like the amount of strikes attempted and landed, area of the body where the strike was landed or attempted, amount of clinches, takedowns or submissions attempted or landed, the winner of the fight, fighter historical record and many others. The purpose of these report is to describe and analyze the variables that can affect a fight outcome, the variables that differentiate a fighter fighting style, the win method and other more general variables as the fighter name, date and location.

During this report I will try to answer the following questions and insights.

- Which fighters have the most amount of fights?
- When does a fighter is at his prime?
- How does a weight class can impact the fight?

- Which is the most common way of finalization?
- Can we categorize the fighters into strikers and grapplers?
- How does head strikes influence the match result?

Data

The UFC Fight Data dataset contains a list of all UFC fights since 2013 to 2017 with summed up entries of each fighter's round by round record preceding that fight, also information from both fighters, fight details (like strikes types) and the winner. The data was scraped from ufcstats website.

The dataset is made up of 2318 observations and 892 variables mostly numeric, this means there is information about 2318 fights that occurred between 2013 and 2017. There are plenty of missing values in the dataset. Most of these missing values are in the variables with information of each round strike type, height, weight, winby, age, and hometown. The dataset also contained values that made no sense, like the max round variable having a value of 4 when fights can only be 3 or 5 rounds length. There were also fighters with a weight that does not correspond to any weight class, this can be because fighters didn’t reached the according weight so they were fined by the UFC but they still let them fight.

Methods

Data Analysis

The dataset was retrieved from Kaggle. All analysis was made using python as the coding language. From python the following libraries were used:

- For loading the data and data wrangling:
    - Pandas
    - Numpy
- For visualizing the data:
    - Plotly
    - Matplotlib
    - Seaborn
- For further analysis:
    - Scikit-learn
    - Scipy

Data cleaning

There were plenty of missing or wrong data on the dataset, this NAs and wrong values were distributed in the following columns:

- Age
- Height
- Weight
- Hometown
- Winby
- All columns that contained information about the fighter strikes during each round

Depending on the variable different imputation methods were applied. As the dataset was relatively small deleting all missing values wasn't an option for me as it would mean a great part of the dataset.

- Variables fixed that had wrong values
  - Weight: Some fighters had weights that didn't match with the UFC standard weight classes, so the value was changed to the closer weight class value.
  - Max round: Some fights had a max value of 4 when fights can only be 3 or 5 rounds length, so I changed the 4 with a 3 instead.
- Variables fixed by imputing the mean, so statistics aren't affected
  - Age
  - Height
- Variables fixed by imputing the most common value (mode) because they are categoric variables, so I am not able to impute with the mean
  - Weight
  - Location
  - HomeTown
- Create a new value for the variable "winby". Searching on Google I found out that these fights had missing values because they corresponded to fights that were declared "No contest". This means that one of the fighters did something, like using an illegal substance, so the UFC canceled the result
  - NC, which corresponds to "No Contest" was the new value created.
- Variables imputed with a value of 0:
  - All variables that contained information about the fighter strikes during each round. I imputed them all with a value because if they are missing the most probable thing is the fighter didn't attempt or land any strike/clinch/takedown/etc or in case of Round4 and Round5, the fight never reached any of those rounds. These applies to any round beyond the last round or the fight, independent of the max round.

PCA

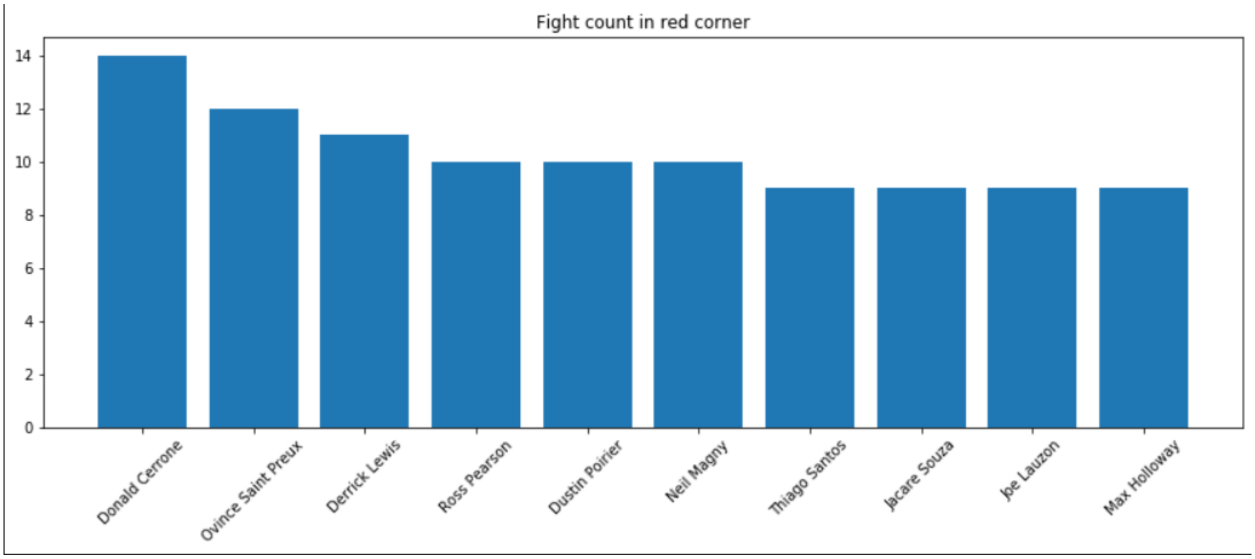
Because there are hundreds of variables that describes the kind of strikes made by both fighters in as many as 5 rounds a way to summarize all of these data is to use a PCA algorithm to cluster and categorize a fighter style. By intuition I can guess there will be at least 2 fighter's styles, these styles correspond to strikers and grapplers. These are the steps I followed for using the PCA algorithm to categorize the fighter's style into 2 categories.

- Manually set the number of clusters I want the algorithm to find. In my case the number of clusters was 2
- Create an AgglomerativeClustering object, module from scikit-learn, which contains the parameters I want to use in the algorithm
- Call the function *fit_predict* and passing the data I want to cluster as a parameter. In this case I wanted to use all the variables that contained information about the fighter strikes during each round.
- Create a PCA object, from scikit-learn module, with a number of components set to 2.
- Call the *fit_transfrom* function on this last object to get a PCA model with the data.
- Make a visualization of the resulting PCA model

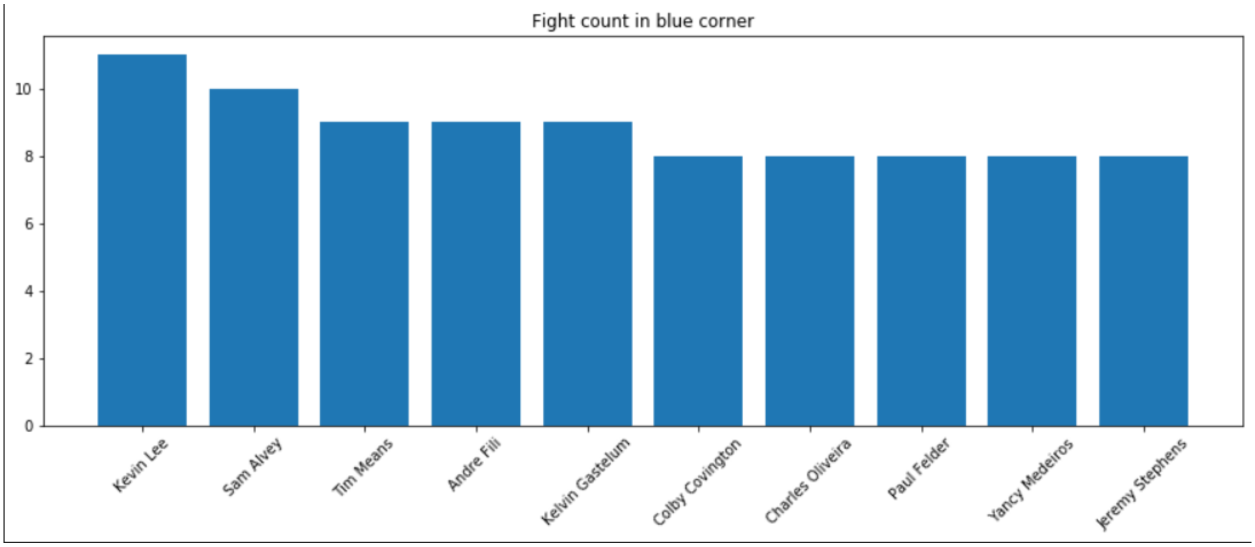
Results

Which are the fighters with the most fights?

As the dataset provides the information of both corners, I analyzed the fighters that have fought the most on each corner.



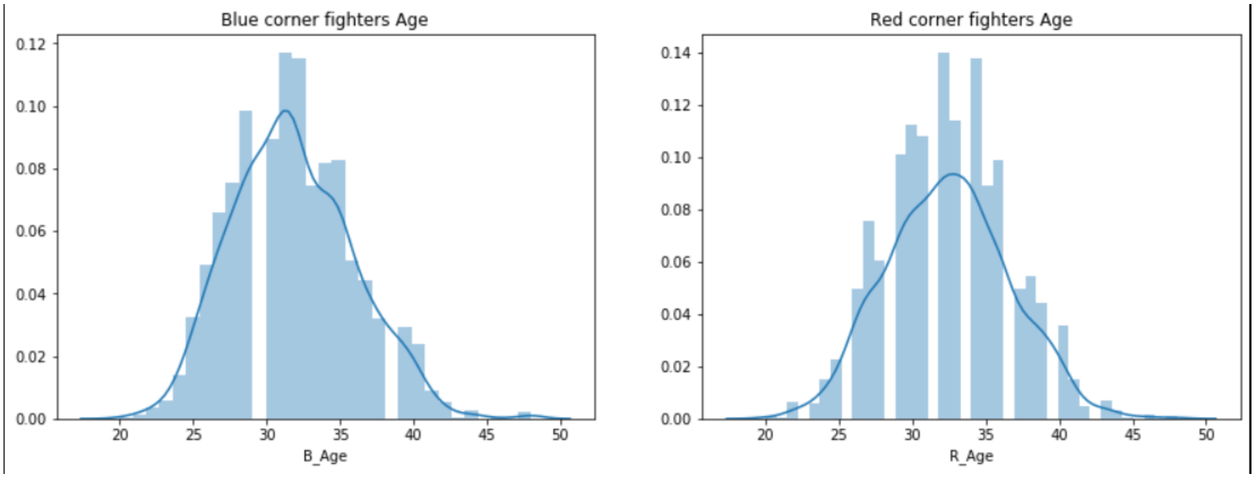
As we can see Donald “The Cowboy” Cerrone is the fighter that have fought the most in the red corner. The red corner is known to be the corner assigned to the fighter with the lowest rank in the fight (aka the Contender). The amount of fights is 14 which gives us an average of 3 fights per year which is the normal amount of fights a fighter has in a year.



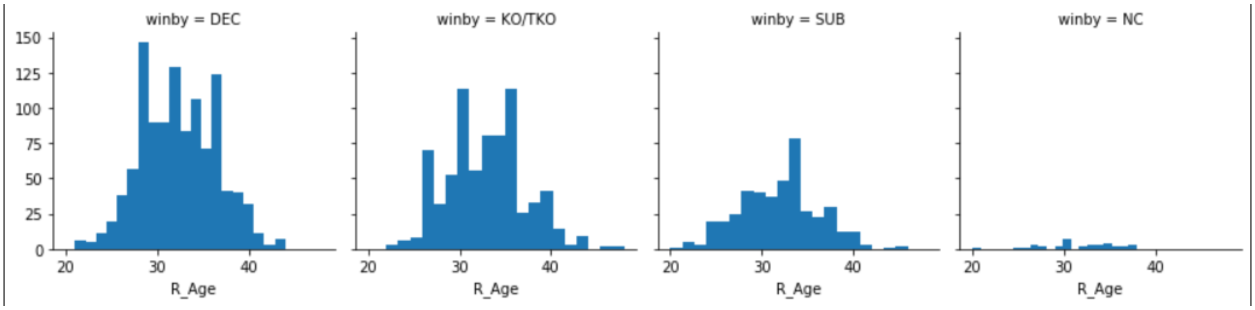
For the blue corner, which is known to be the corner assigned to the fighter with the highest rank (aka the Defendant) we see that Kevin Lee is the fighter that fought the most with 11 fights. The amount of fights is less than the red corner, probably because a fighter defending his places isn’t as urged to fight as a fighter who is looking to climb the ranks. We also don’t see any title holders in the graph, maybe because title fights happen less often or because these fighters lost the belt and became contenders before achieving a high number of title defenses.

When does a fighter is at his prime?

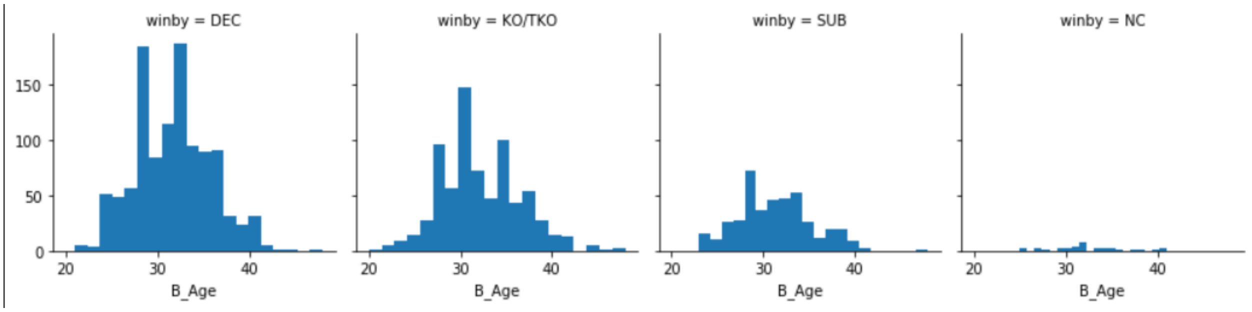
To analyze this, I checked what I think is the most influencing variable for this, which is the fighters Age.



As we can see the fighters age has a normal distribution centered at around the early 30s of the fighters for both corners. These make sense with the fact that many people think athletes reach their prime during their late 20s or they're early 30s. But because this is not enough to answer my question, I checked the Age of winning fighters by corner color.



Red Corner

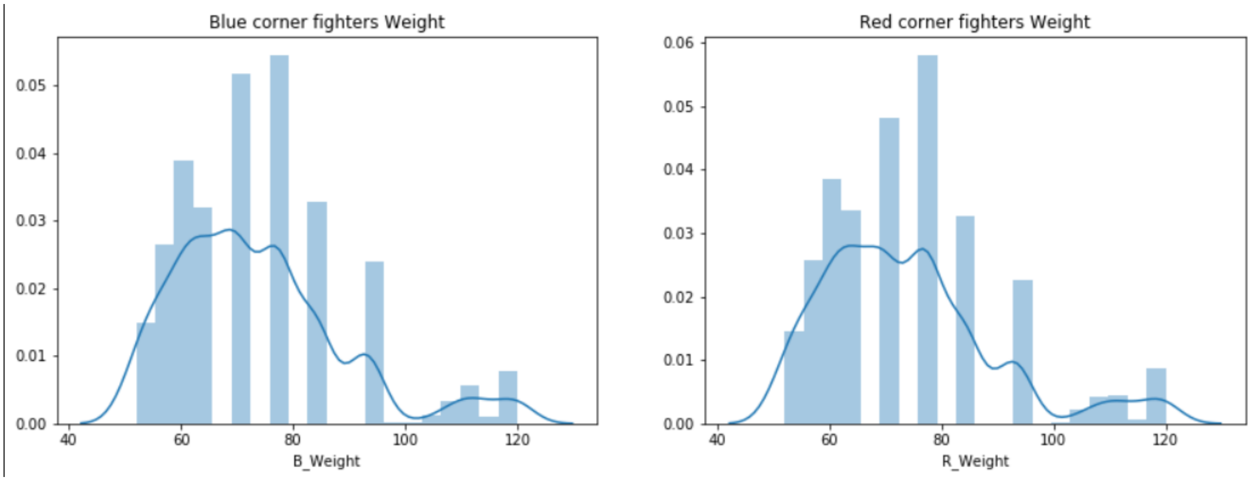


Blue corner

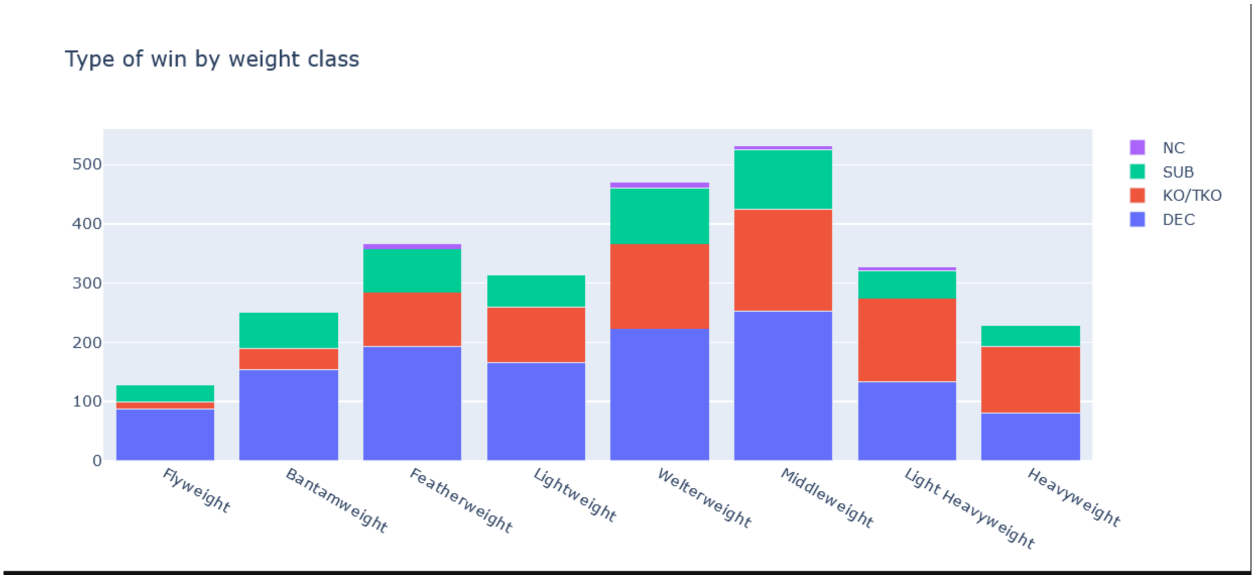
These distributions confirms that fighters win most fights when they are on they're early 30s.

How does weight class can impact in the fight?

First I analyzed the distribution of the fighter's weight.

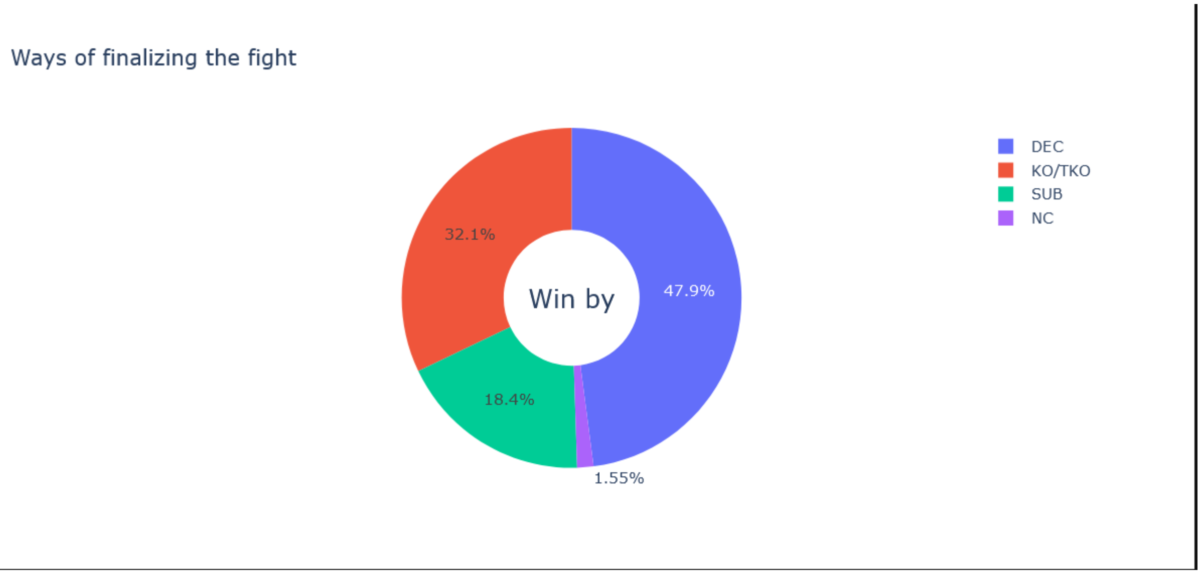


Most fighters weight around 70kg which corresponds to the Middleweight class. Lower and Higher weight classes are less popular than weight classes at the middle.

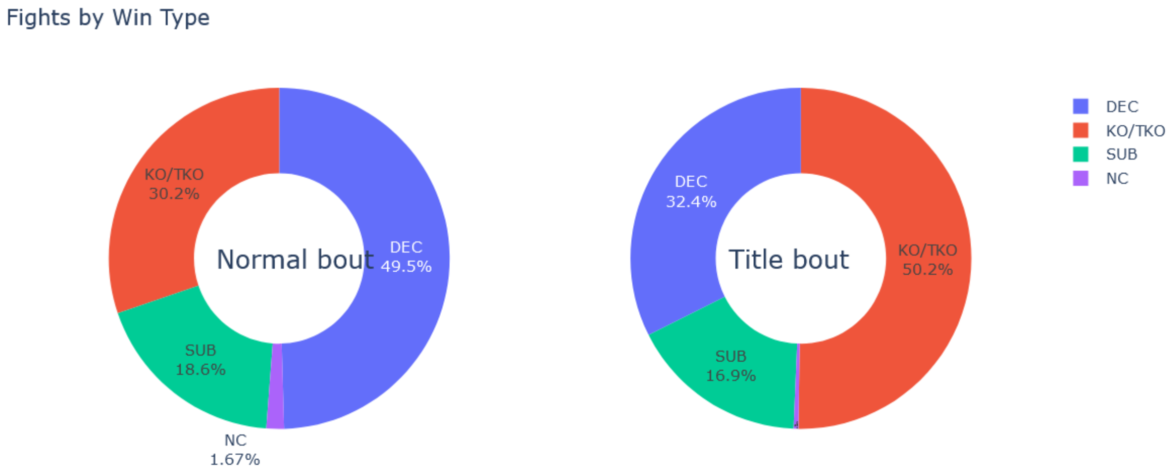


As we can see on lower weight classes wins occur mostly by decision and really few by KO/TKO (Knockout or Technical Knockout). As the weight increases, we can see more KO/TKO finishes than Decision finishes. This is because heavier fighters have more force on each strike due to their weight, so they are more dangerous strikers. We can also see that the amount of submissions focuses around the middle weight classes, this is because heavier fighters tend to be strikers and lighter fighters do not have the necessary force to make their opponents submit.

Which is the most common way of finalization?



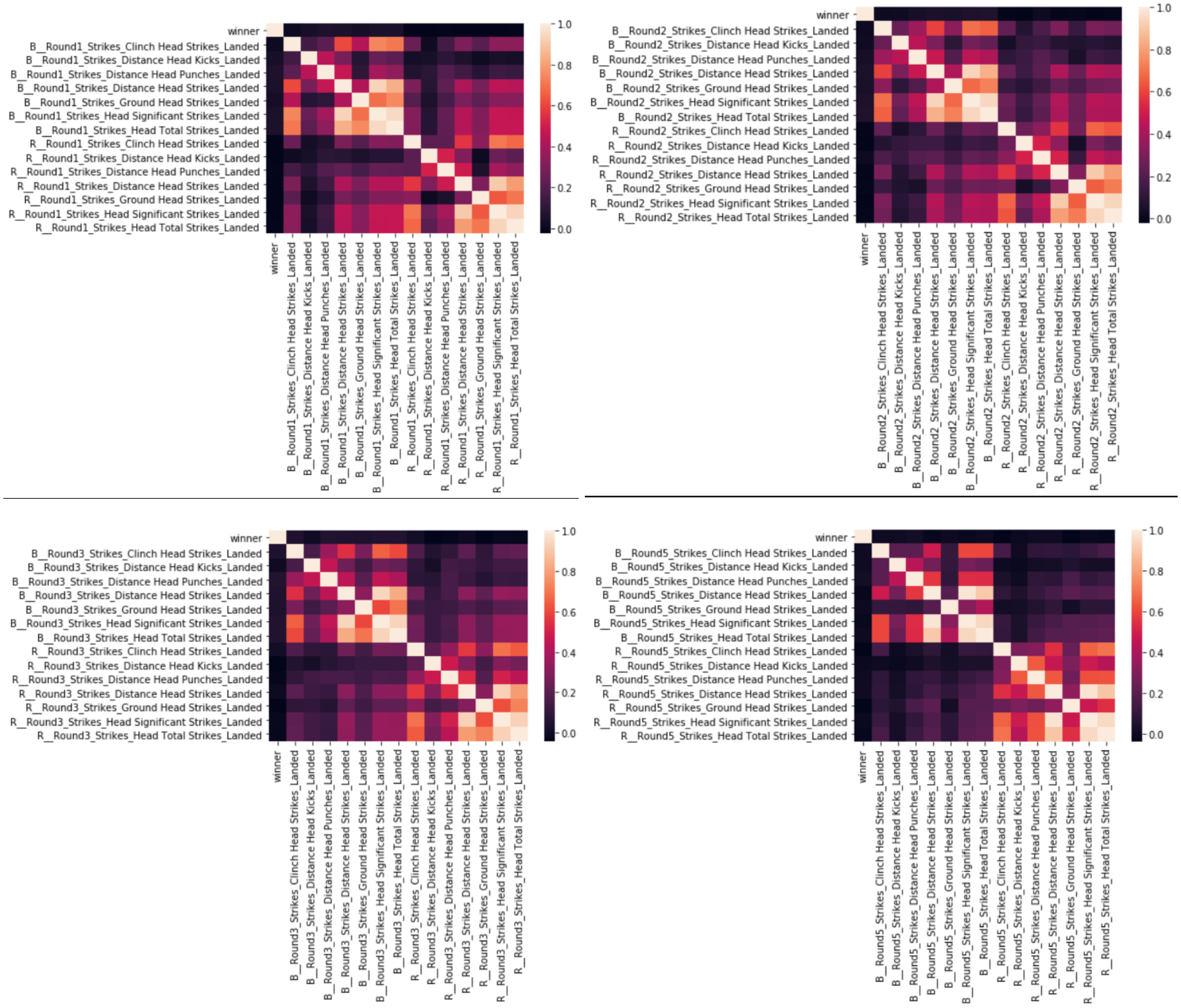
These charts show us that almost half of the fights are won by decision. KO/TKO finishes are the second most common way of winning and a submission is the rarest way of winning a fight. We can also see the few amount of fights that were declared as No Contest, this is because fighters do not want to have a match turned into No Contest and most times this happens because they are not aware that a supplement or protein they're taking contains an illegal substance for the USADA, the organization that enforces substances rules on the UFC.



I also analyzed how those fights finishes depending on the kind of bout, whether it is a normal fight or a title or main event fight. We can see that on 3 rounds fights Decision is the most common way of winning. But when it is a title or main event fight it tends to end on a KO/TKO. This happens because champions or contenders wish to end the fight in a definite way instead of leaving the decision to the judges. Also because these are 5 round fights where stamina plays a huge factor in the last 2 rounds making a KO or TKO easier if the fighter does not have good stamina.

Does Head Strikes influence in the match result?

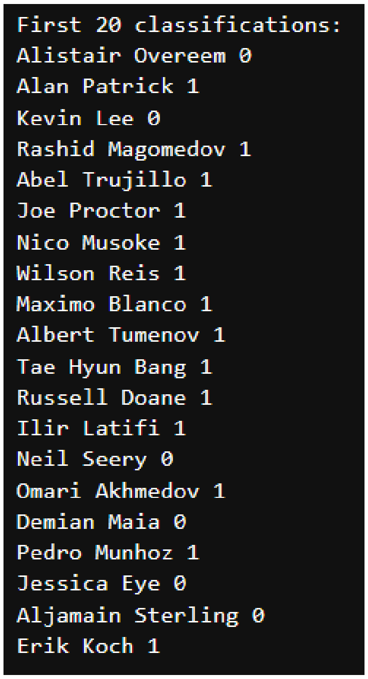
To analyze head strikes and try to find out if they contribute to a win, I made a One-Hot Encoding for the *winner* variable where 1 means that blue corner won and 0 if the red corner won. For this analysis I only took in account strikes that landed, not attempted, on the head, whether it was a punch or a kick.



Correlation matrix shows Round1, 2, 3, and 5, in that order from left to right, top to bottom. In every round we can see that all other features have 0 correlation with the winner variable. So head strikes can not define which corner wins the fight. But something interesting to see is how strikes from each corner start at around 0.5 correlation and with each passing round they tend to 0 correlation between each other. This may be to the fact that in the latest rounds fighters are more tired and they know that a well placed strike could knock them out, so they tend to fight more defensevily which means “throwing less strikes”. Also becase in latest rounds a fight will tend to end in a judges decision and not with a fighter finalizing the fight

Categorize the fighters by fighting style?

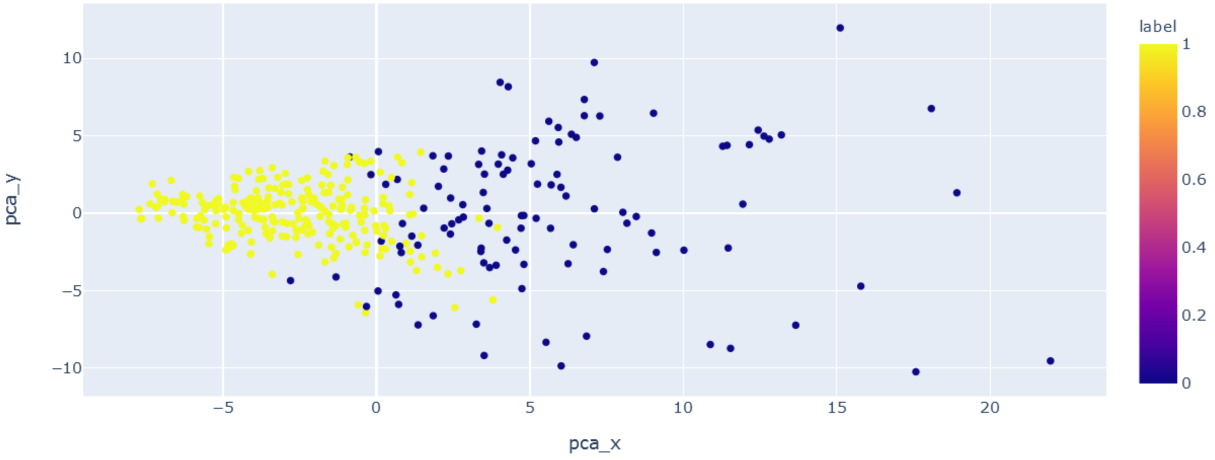
Here I will show the results of applying the PCA algorithm explained on the *Methods* section of this report.



```
First 20 classifications:
Alistair Overeem 0
Alan Patrick 1
Kevin Lee 0
Rashid Magomedov 1
Abel Trujillo 1
Joe Proctor 1
Nico Musoke 1
Wilson Reis 1
Maximo Blanco 1
Albert Tumenov 1
Tae Hyun Bang 1
Russell Doane 1
Ilir Latifi 1
Neil Seery 0
Omari Akhmedov 1
Demian Maia 0
Pedro Munhoz 1
Jessica Eye 0
Aljamain Sterling 0
Erik Koch 1
```

When watching the results of the fighters with his correspondent category we can find out what does the 0 or the 1 means. I'll take Alistair Overeem to sense what does this One-Hot Encoding means. Overeem is known to be a striker; he is a heavyweight fighter so it makes sense he is known to be a striker and to avoid grappling as much as he can. Thanks to this I can infer that the cluster number means:

- ○ 0: for strikers
- ○ 1: for grapplers or wrestlers



Here we can see the cluster for these strikers and grapplers categories. As we can see there are many fighters at each edge, this means that many try to be only grapplers, and many tend to be only strikers. This can be contradictory to a Mixed Martial Arts fight but analyzing the background of many fighters it makes sense. It makes sense because many of the fighters used to be grappling champions on their country. Some others used to be boxing and kickboxing champions. So many of the fighters are specialized in only 1 fighting style and are not specialized in all styles.

Conclusion

To conclude my report, I can say the following statements:

- Fighters on the red corner tend to fight more than fighters on the blue corner because they are looking to climb up the ranks and they have the urge to show that they can be the best. Unlike blue corner fighters which are less willing to be contenders as they already have a rank or title to defend.
- Fighters are at their prime during they're early 30s and during they're late 20s. This happens because of the physical strength needed for this sport.
- Heavier weight classes are more dangerous as the have more Knockout power. Lighter weights tend to be more technique focused fighters rather than pure strikers
- Decision winning is the most common for 3 rounds fights and KO/TKO is the most common for 5 round fights.
- Head Strikes does not have an impact in which corner is going to win, but they decrease in amount with each passing round.
- Fighters can be easily categorized in grapplers and strikers.

Appendix

Dataset recovered from: https://www.kaggle.com/calmdownkarm/ufcdataset

Variable dictionary:

- R_ : Suffix for red corner
- B_ : Suffix for blue corner
- Round1: suffix for round 1 attacks
- Round2: suffix for round 2 attacks
- Round3: suffix for round 3 attacks
- Round4: suffix for round 4 attacks
- Round5: suffix for round 5 attacks
- KO/TKO: Knockout or Technical Knockout win
- Sub: Submission win
- DEC: Decision by judges win
- Age: Fighters age
- Weight: Weight class in KG
- Height: height in cms