DATA VISUALIZATION ON ULTIMATE FIGHTING CHAMPIONS

BY

SACHIN VAIDYA

11011995

Abstract

In this visualization UFC dataset was used to analyse the fight science that takes place. The report is on the fights that take place in UFC. Dataset contains details about all the fights that are held since 1993. Using the data available a detail study was done to understand the process in which a game starts, progresses and concludes. My moto here was to understand that by using numbers we can deduce the pattern a fighter follows which can be exploited to come up with better strategies. Based on analysis and studies promoters and trainers can device a plan to bring up their fighter with the help of facts that are left untouched.

Introduction

Mixed martial arts (MMA) is a full-contact combat sport that allows striking and grappling, both standing and, on the ground, using techniques from other combat sports and martial arts. The Ultimate Fighting Championship (UFC) is an American mixed martial arts organization based in Las Vegas, Nevada and is the largest MMA promotion in the world and features the top-ranked fighters of the sport. Based in the United States, the UFC produces events worldwide that showcase twelve weight divisions and abide by the Unified Rules of Mixed Martial Arts. This is a highly unpredictable sport

Few things we will try to visualize:

- How's Age/Height related to the outcome?
- Most popular locations in UFC?
- Most popular way to win the fight?
- Comparing techniques used by fighters

Motivation

Since UFC is an upcoming popular sport which is bound to take a big turn in future. With the available data if analysis can be done to understand the mechanics and the strategy behind the games. If number can provide a small peak into the vastness of visualization then working with UFC is the best place. Since there are hardly countable number of blogs or website available that do data analysis on it. Hence it gave me a opportunity to work with this.

Understanding the data

This dataset contains list of all UFC fights that have taken place since 1993 till the present. This data set gives an insight about every match, each row represents a single fight. Generally, the octagons in UFC have blue and red corners that are assigned to fighters. Basically, speaking red corner is held by the champion or the favourites in the match and blue corner goes to the challenger or the underdog.

This is a very huge dataset working on it was a challenge in itself. This has nearly 147 columns and 1048574 rows, the vastness in itself was a very big challenge to understand and analyse. But this paved way to make more visualization.

R_fighter : Fighter in the red corner
B_fighter : Fighter in the red corner
Referee : Official for the match
date : Date of the event
location : Arena for the showdown
Winner : Winner of the match
title_bout : Was it a title match?
weight_class : The weight class they belong to
no_of_rounds : Is it a 3 round or a 5 round match
B_current_lose_streak : how many matched he has lost in succession
B_current_win_streak : how many matched he has won in succession
B_draw : A match with no winner or a loser

Pre-processing:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1048575 entries, 0 to 1048574
Columns: 147 entries, R_fighter to R_age
dtypes: float64(135), object(12)
memory usage: 1.1+ GB
```

Next step was to identify if there were any null or empty values in the dataset. There were many missing data, but the fields that were of major concern was age, height and weight. These couldn’t be empty or null.

Missing Values:

The dataset was clean and structured. All the data were in their relevant position and there was no need to rearrange or sort them out. But there were some missing values in the field of height and age. Since height and age are very important factors for our analysis.

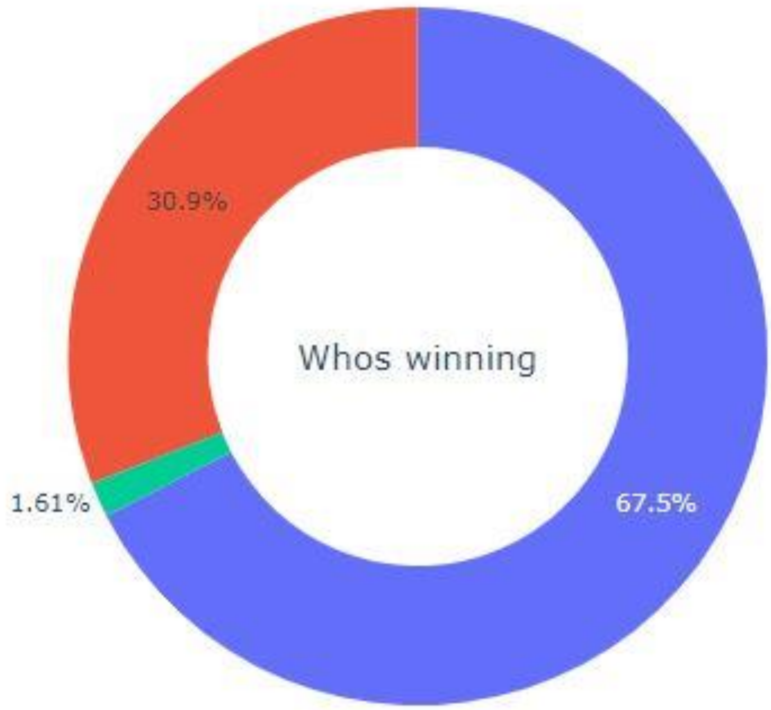
In order to address the missing values in age and height. A simply operation was to delete rows with missing values, but usually there is a scope to take advantage of as many data points as possible. Replacing missing values with zeros would not be a good idea - as age 0 will have actual meanings and that would change our data.

Therefore, a good replacement value would be something that doesn't affect the data too much, such as the median or mean. the "fillna" function replaces every NaN (not a number) entry with the given input (the mean of the column in our case).

```
df['B_age'] = df['B_age'].fillna(np.mean(df['B_age']))
df['B_Height_cms'] = df['B_Height_cms'].fillna(np.mean(df['B_Height_cms']))
df['R_age'] = df['R_age'].fillna(np.mean(df['R_age']))
df['R_Height_cms'] = df['R_Height_cms'].fillna(np.mean(df['R_Height_cms']))
```

Data Visualization

Let's start by looking who's winning more from our dataset:



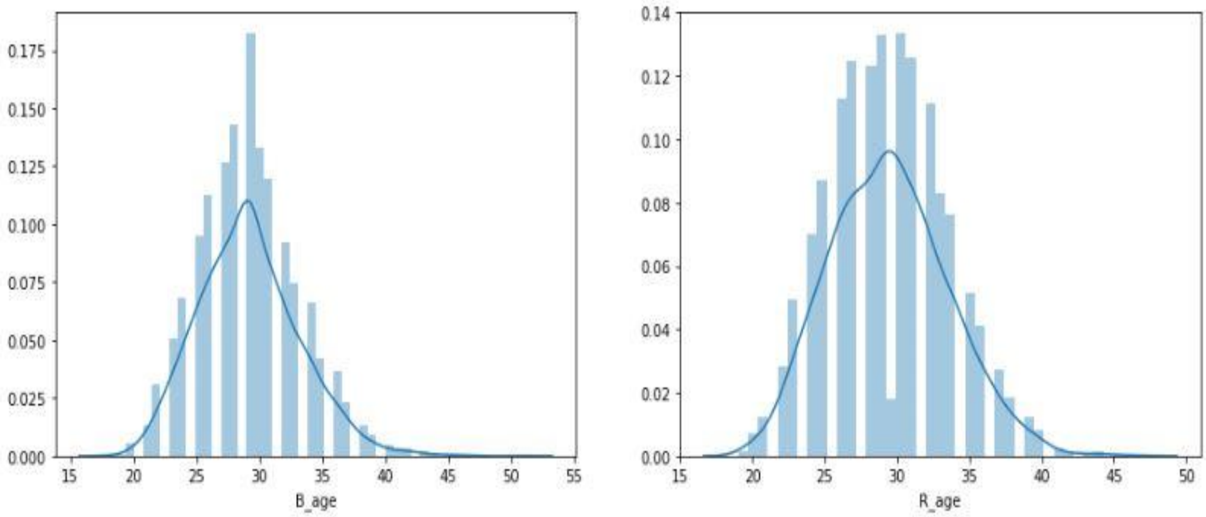
The above graph gives a brief idea about the way the match has been paved. It makes very clear that maximum of the times the challenger has a win, it often ends up with the champion losing his belt to challenger. More the 50% of the time it has been observed that the underdog or the fighter who isn't favourite ends up winning.

Since this pattern of blue fighter winning occurs regularly, there should be a reason for it. The next step is to understand what causes this spike in blue corner winning maximum of fight. Maybe there can be a difference in the ages.

Taking both red and blue corner fighters age, a simple comparison can be done and a graph can be plotted to give a bigger picture.

```
fig, ax = plt.subplots(1,2, figsize=(15, 5))
sns.distplot(df.B_age, ax=ax[0])
sns.distplot(df.R_age, ax=ax[1])
```

The above code gives a graph



The graph makes it clear that the age has a slightly better advantage when it comes to fighter. The difference in age isn't major, maybe around 1 or 2 years yet this makes a big difference in execution. On observation of the ages an interesting pattern can deduced.

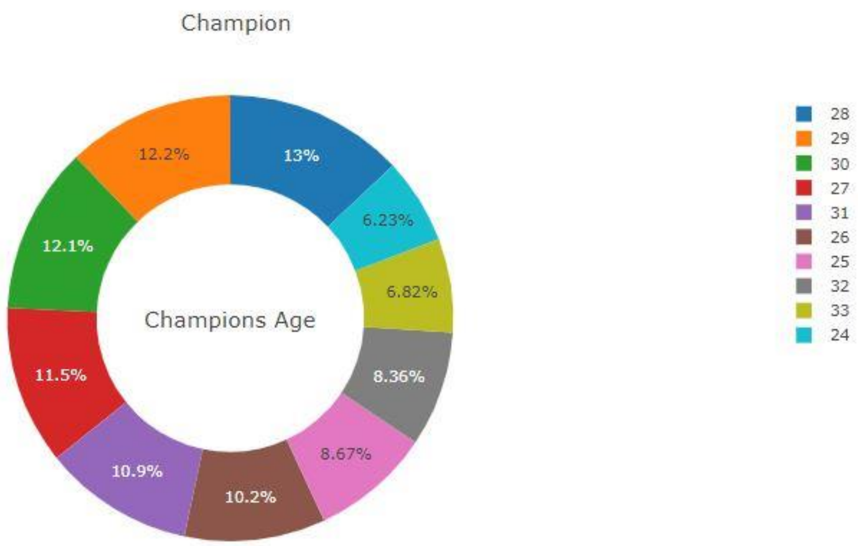
Let us take example of fighters in red and blue corner. At what age are they winning titles or championship.

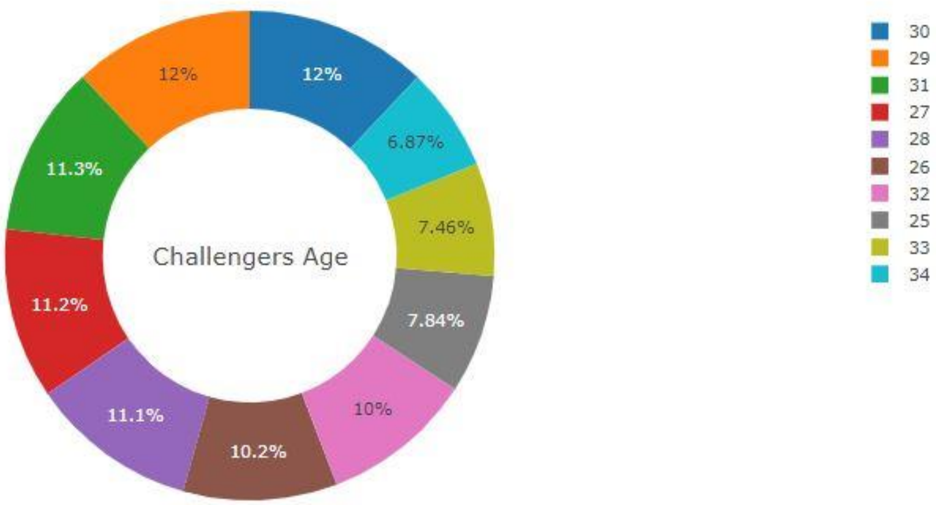
```
BAge = df.groupby(['B_age']).count()['Winner']
BlueAge = BAge.sort_values(axis=0, ascending=False)
blue = BlueAge.head(10)
BlueAge.head(10)
```

```
B_age
28.0    505
29.0    473
30.0    470
27.0    448
31.0    424
26.0    398
25.0    337
32.0    325
33.0    265
24.0    242
```

```
RAge = df.groupby(['R_age']).count()['Winner']
RedAge = RAge.sort_values(axis=0, ascending=False)
RedAge.head(10)
```

```
R_age
30.0    469
29.0    467
31.0    442
27.0    439
28.0    432
26.0    397
32.0    392
25.0    306
33.0    291
34.0    268
```





If we observe clearly, we can understand that maximum title fights are done in their mid-20s and early 30. The above chart and donut make is very clear that 25-35 is the best range where a fighter can be expected to be in his prime.

But again, the champion ages and there are always new fighters to challenge the champions. On analysing the age difference, we understand the challengers are generally younger than champions. But this doesn't necessarily that younger players win all the time because age cannot beat experience, but youngers challengers do have a small advantage from the start.

Looking towards the next page when difference between the age of challenger and the champion is taken, challenger is the youngest. By making repetitive analysis on age a base line can be introduced regarding the age of contenders.

```
df['Age_Difference'] = df.B_age - df.R_age
df[['Age_Difference', 'Winner']].groupby('Winner').mean()
```

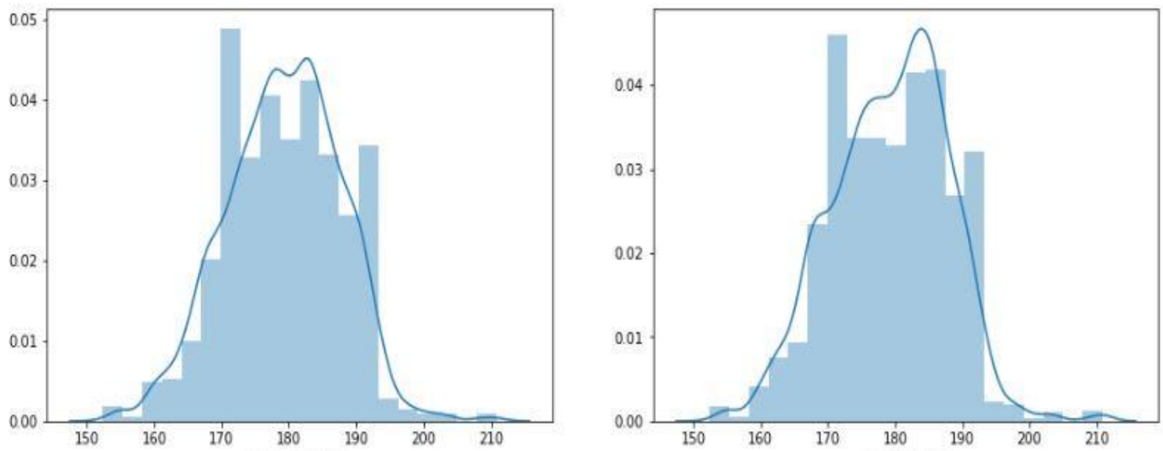
	Age_Difference
Winner	
Blue	-1.576186
Draw	-1.398776
Red	0.355354

Are there any other factors?

By analysis it could be concluded that age plays a vital role in the matches and player performance, are there any other factors that help the fighters to have an advantage. Lets start by taking height as a consideration for this factor.

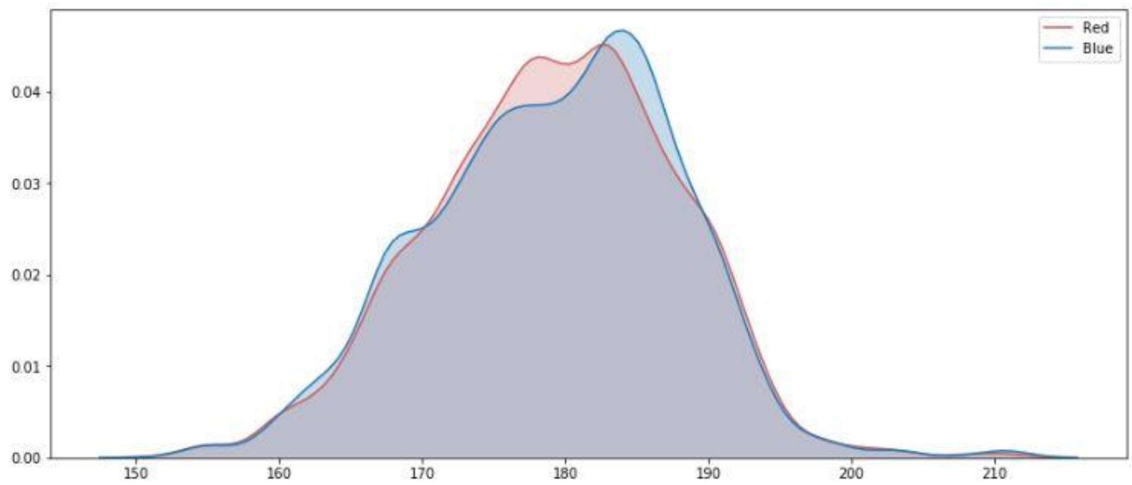
```
fig, ax = plt.subplots(1,2, figsize=(15, 5))
sns.distplot(df.B_Height_cms, bins = 20, ax=ax[0]) #Blue
sns.distplot(df.R_Height_cms, bins = 20, ax=ax[1]) #Red
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x277786b1c50>
```




```
fig, ax = plt.subplots(figsize=(14, 6))
sns.kdeplot(df.B_Height_cms, shade=True, color='indianred', label='Red')
sns.kdeplot(df.R_Height_cms, shade=True, label='Blue')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x277789ae7b8>
```



If an observation is done keenly over the height, it gives a sign that if you are tall you have better advantage. Height gives a better reach; fighter can keep his distance and continue to fight.

The fighters in blue are taller, the upcoming generation are younger and taller giving them double advantage.

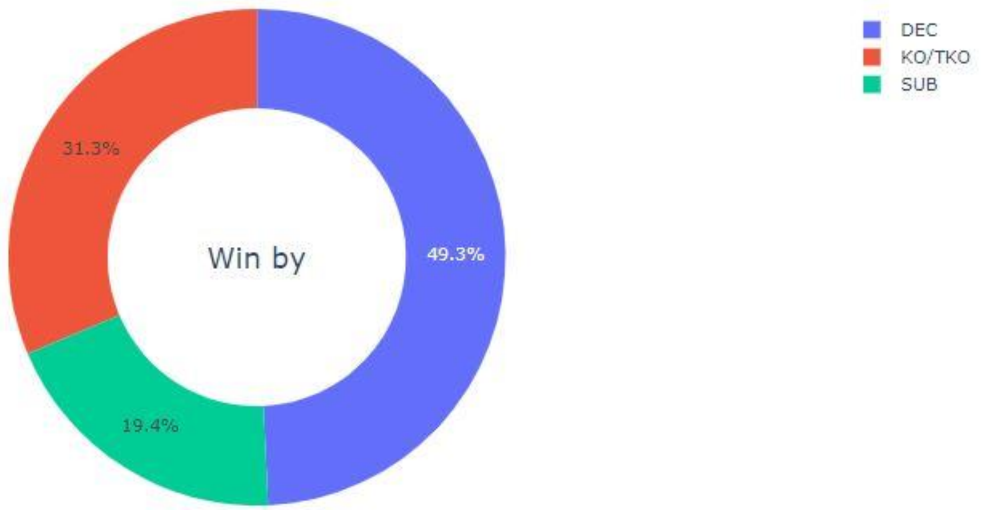
How the fighters are winning?

Now, let's talk about how the fighters are winning. The three most popular ways to win in an MMA fight are:

1. DEC**:** Decision (Dec) is a result of the fight or bout that does not end in a knockout in which the judges' scorecards are consulted to determine the winner; a majority of judges must agree on a result. A fight can either end in a win for an athlete, a draw, or a no decision.

2. SUB**:** also referred to as a "tap out" or "tapping out" - is often performed by visibly tapping the floor or the opponent with the hand or in some cases with the foot, to signal the opponent and/or the referee of the submission

3. KO/TKO**:** Knockout (KO) is when a fighter gets knocked out cold. (i.e.. From a standing to not standing position from receiving a strike.). Technical Knockout (TKO) is when a fighter is getting pummelled and is unable to defend him/herself further. The referee will step in and make a judgement call to end it and prevent the fighter from receiving any more unnecessary or permanent damage, and call it a TKO.

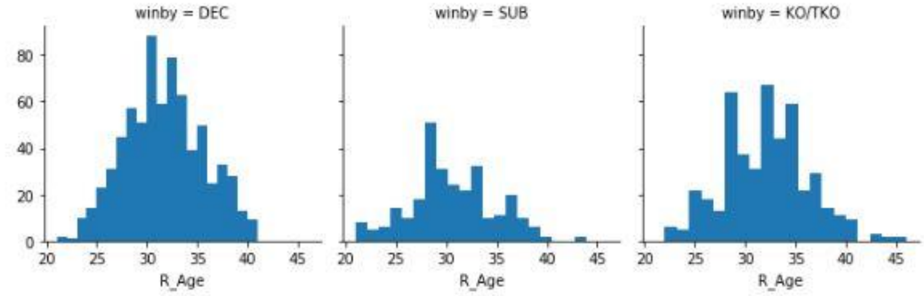


So, most fights are going to the judges. Second most popular way is Knockout and the Technical KO.

Let's check how this is distributed with respect to Age for 'Red' fighters.

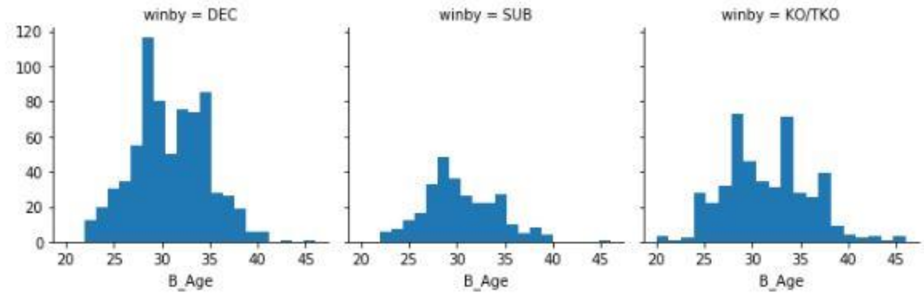
```python
g = sns.FacetGrid(win, col='winby')
g.map(plt.hist, 'R_Age', bins=20)
```

```
<seaborn.axisgrid.FacetGrid at 0x231ed7c1048>
```




```python
g = sns.FacetGrid(win, col='winby')
g.map(plt.hist, 'B_Age', bins=20)
```

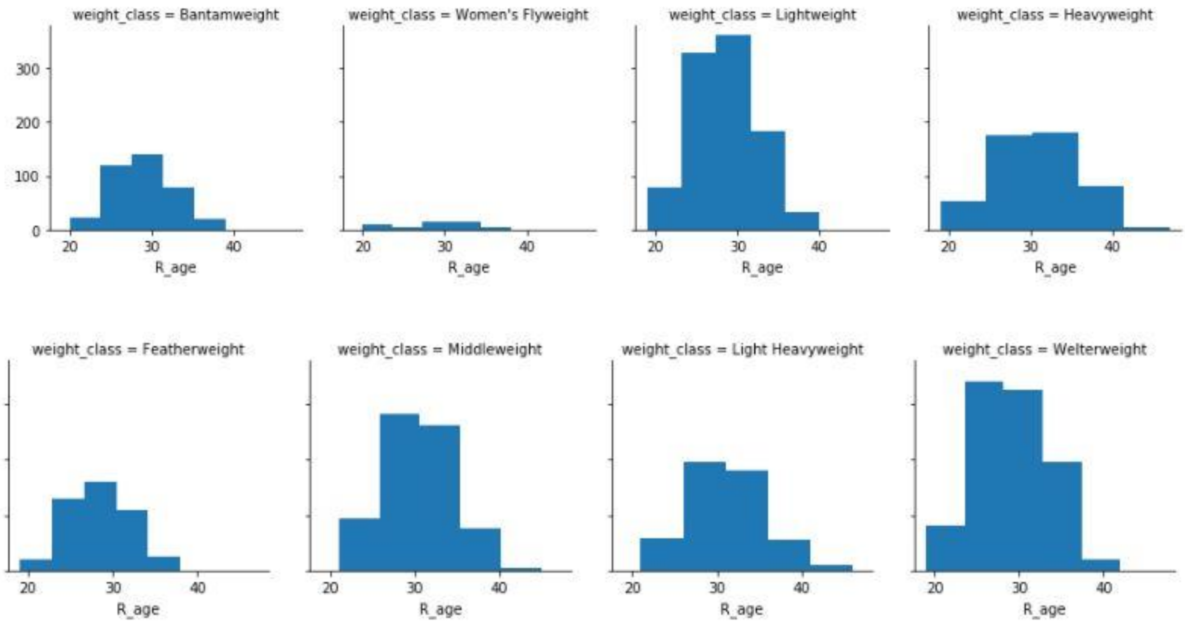
```
<seaborn.axisgrid.FacetGrid at 0x231ee93f3c8>
```



Let’s check out how the age is distributed among different weight classes.

```python
g = sns.FacetGrid(df, col='weight_class')
g.map(plt.hist, 'R_age', bins=5)
```

```
<seaborn.axisgrid.FacetGrid at 0x277880bb978>
```



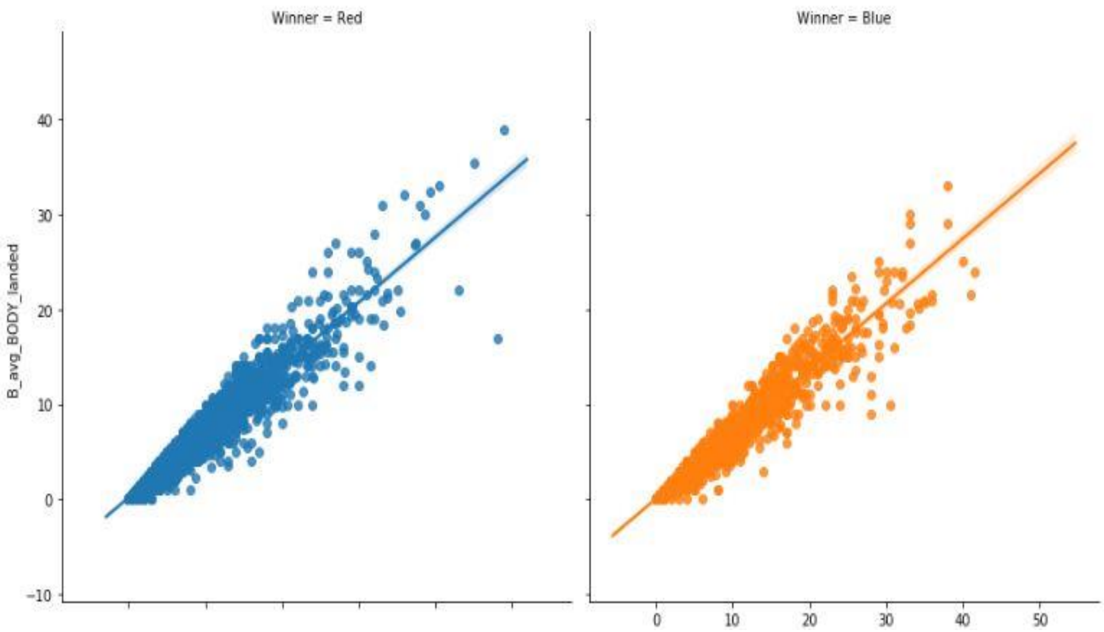
Maximum fighter is in weight class of middleweight, lightweight, light heavyweight and welterweight. The number of women fighters is comparatively less, making it a platform open for all upcoming women fighters.

MMA is a complex sport; in a sense it is the only sport where defence and offense could be done in the same movement. Hitting someone is a risk as it leaves you open for your opponent to counter. However, the *bigger the risk, the greater the reward*. More offensive attempts you make should mean more you land on your opponent (and with right skills and power - more chance you have to win the fight).

Let's see if this is true with our data

```
sns.lmplot(x="B_avg_BODY_att",
           y="B_avg_BODY_landed",
           col="Winner", hue="Winner", data=df, col_wrap=2, size=6)
```

<seaborn.axisgrid.FacetGrid at 0x2777bece278>



Taking a look at the above graph which is body attack vs body attack landed, in the range of 20-40. Red fighter has a mismatch in body attacks and body attacks landed, many of the attacks after the count of 20 start missing hence there is a disturbance in the distribution. While the blue fighters have a continued pattern and there isn't much disturbance, hence challenger or fighters in blue corner have more precision in their attacks and most of them land on the body.

Let's look at the favourite locations of red and blue corner

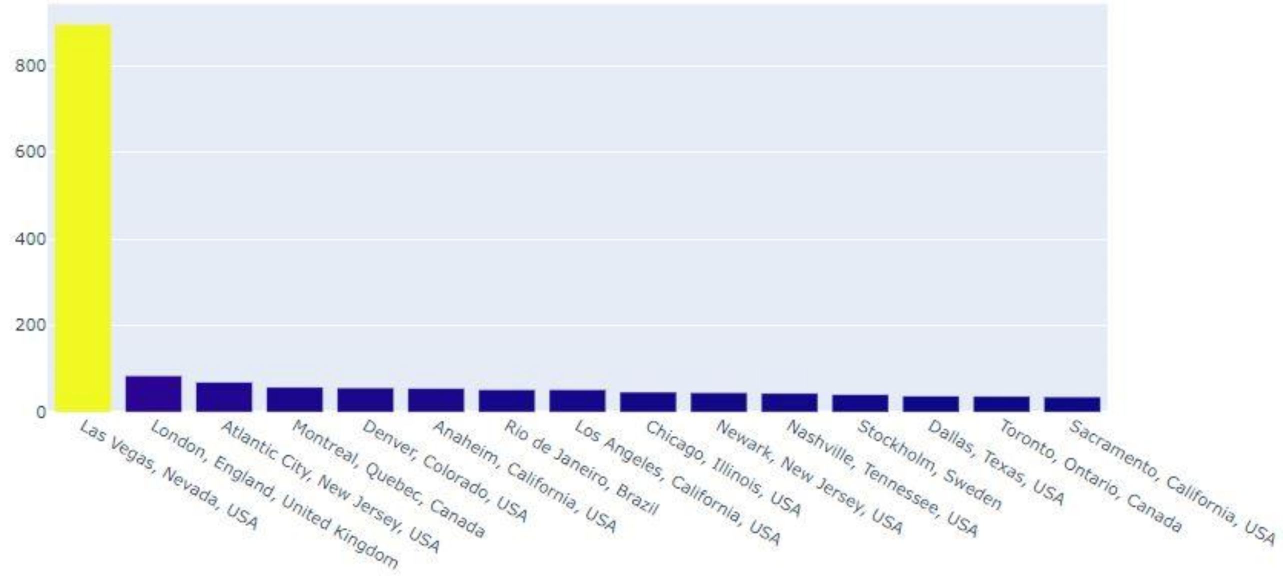
Red corner like to play it all out in Las Vegas, Nevada, USA and their second prefer location is London, England, United Kingdom.

When it comes to blue corner they prefer to win in Las Vegas, Nevada, USA and Toronto, Ontario, Canada.

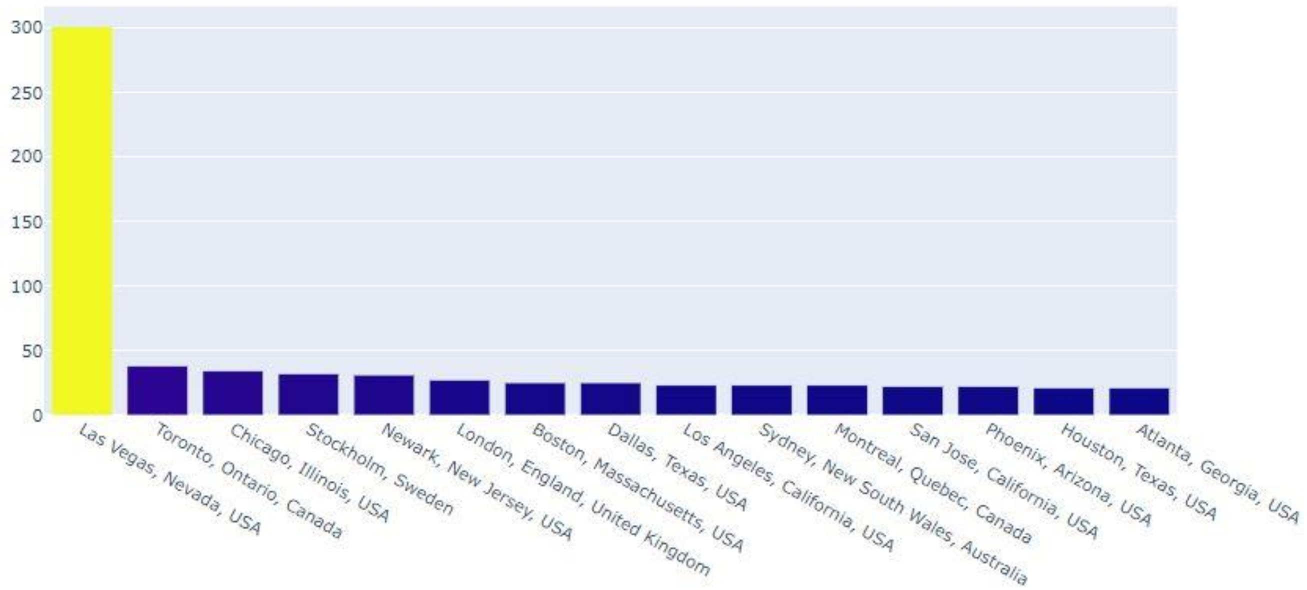
It's a strange observation but a fact to see the corners have different location where they have majority of their wins.

Major popularity of this sport is in USA, Canada, Brazil and United Kingdom.

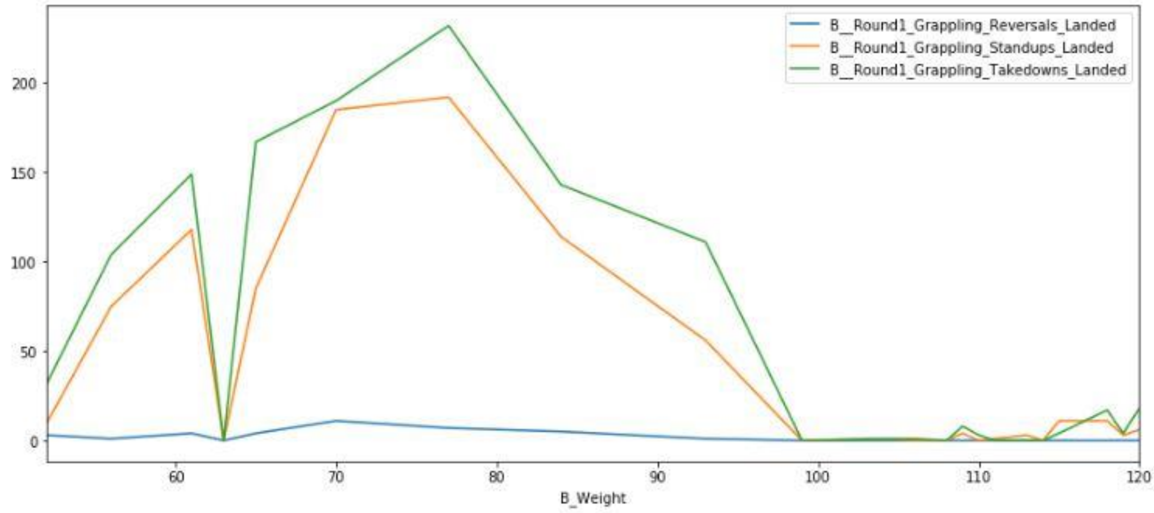
Most popular city for red fighters



Most popular city for blue fighters

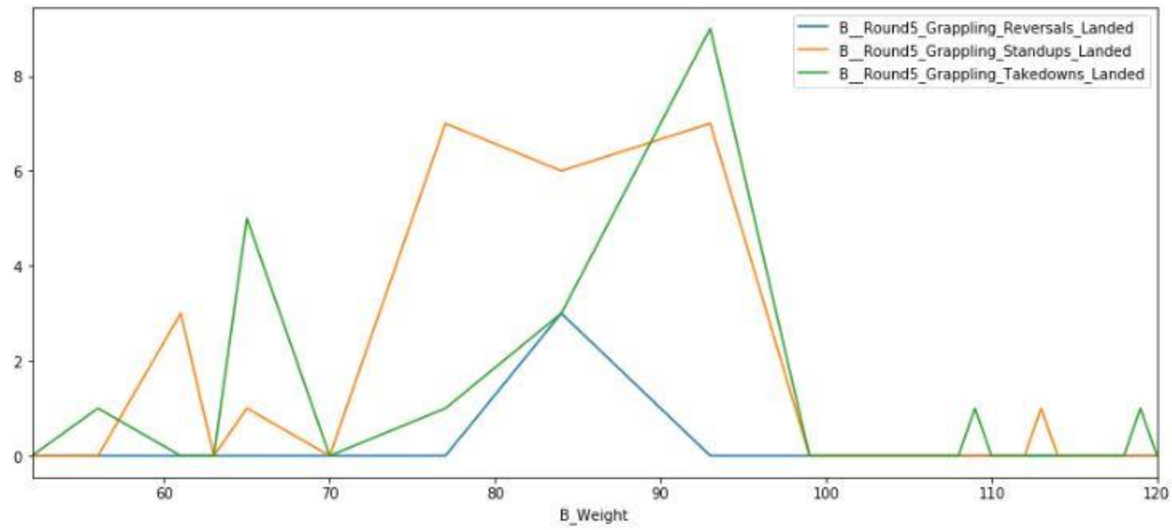


Now, let's look at the Grappling reversals, grappling standups and grappling takedowns landed in different weight categories in Round 1****



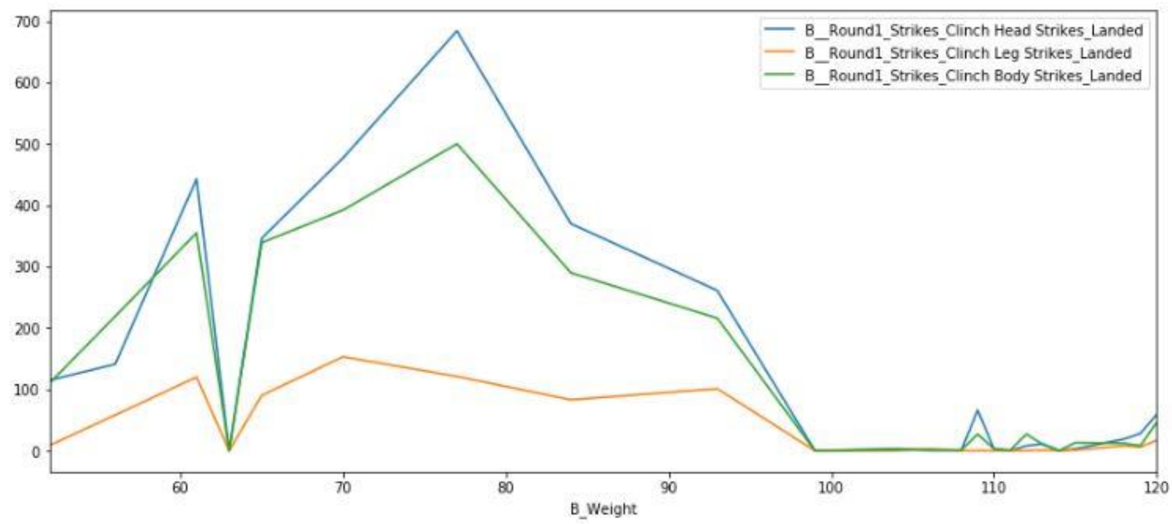
There are very few Grappling reversals but high amount of Grappling takedowns that were landed. More specifically weight classes between 70 - 80 prefer takedowns during Round 1.

Comparing it with Round 5



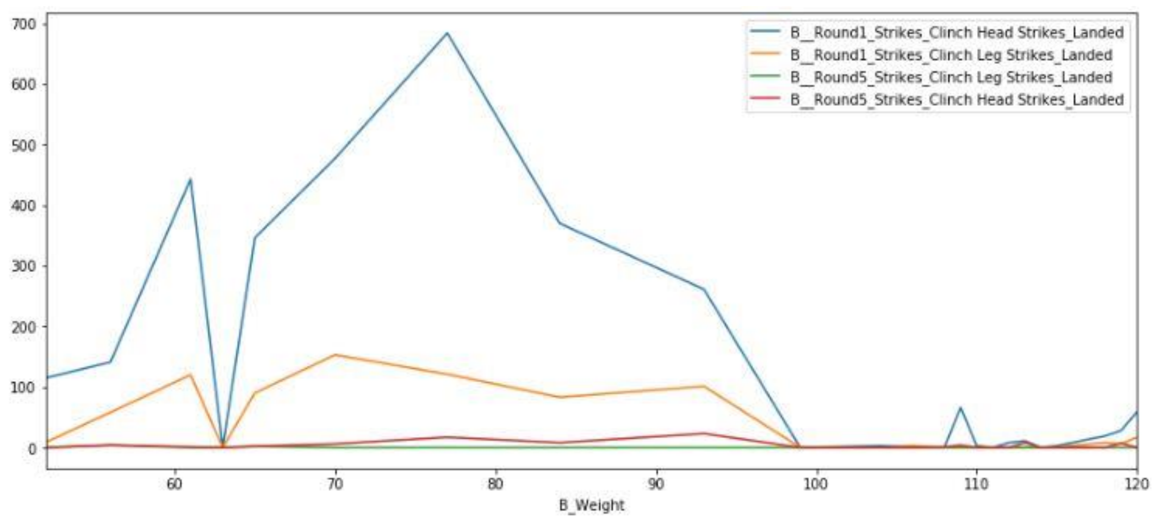
Interestingly, grappling reversals increase for fighters between weight 80-90, while takedowns have decreased in the lighter weight groups.

Let's look similar data for Clinch head strikes, Clinch leg strikes and Body strikes for Round 1



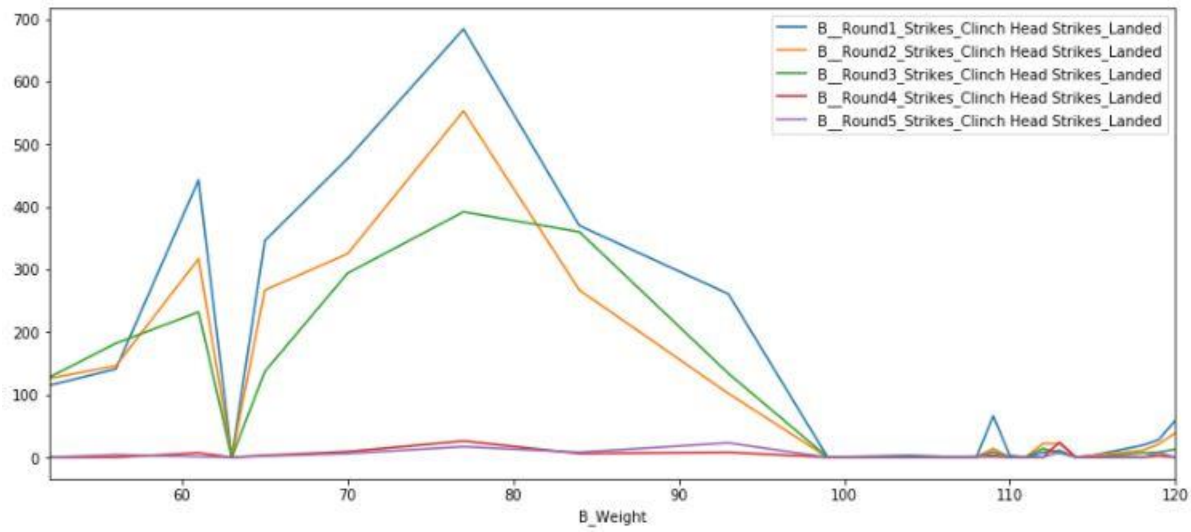
Fighters prefer to land more head strikes during round 1 followed by body strikes. They are in for the kill at the start of the match. Very less preference is given to leg kicks because leg kicks are taken till the 3rd or 5th round to make the opponent slow during the match. To immobilize as deep as possible.

Let's compare this with what happens in Round 5:



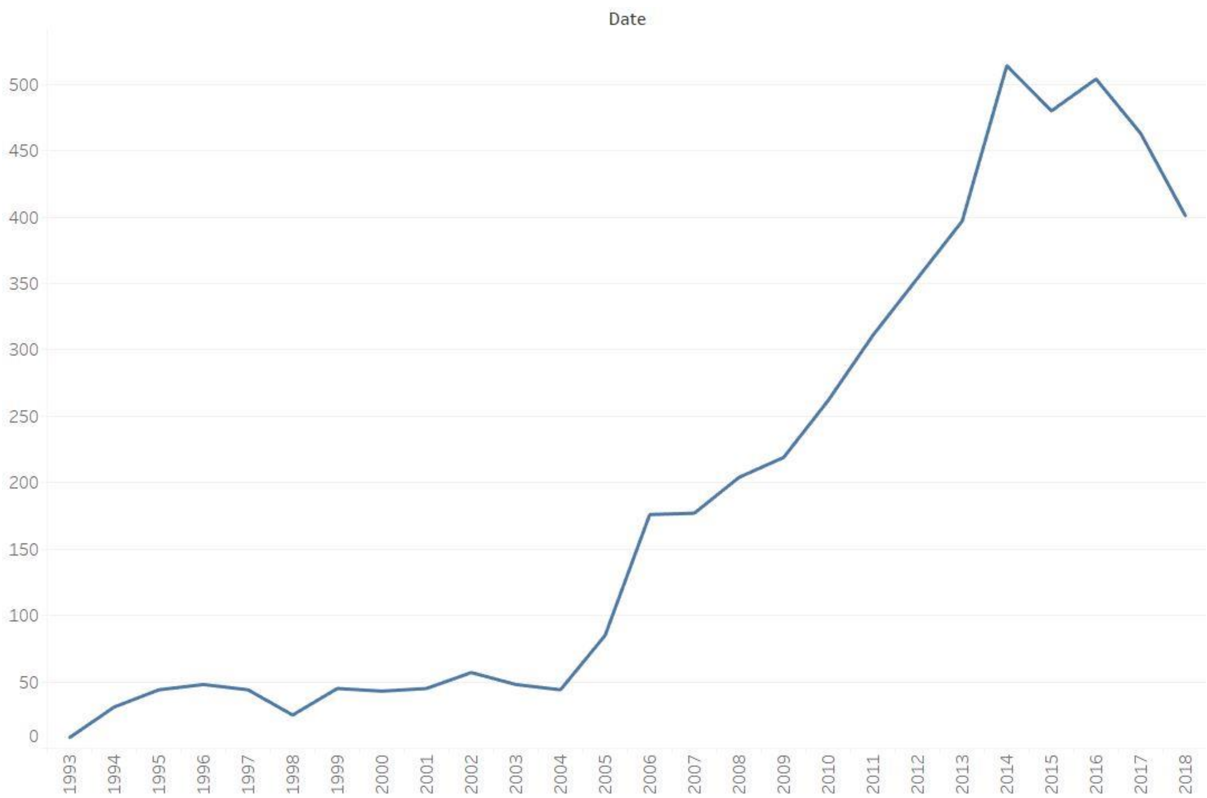
When comparing the head and leg strikes from 1st to 5th round the drop is really hard. None of the fighters try to hit a knockout punch or kick, the range of attack decreases considerable.

Let's compare Clinch Head Strike this with all the rounds:



Clinch head strikes are prominent up till the 3rd round after that fighters don't prefers to use energy for this attack strategy. From the above analysis it can be understood as the match approaches the end fighters like to grapple in order to get more points to win by decision.

How has the Sports evolved through the years:



2014 had the best sales because all the title fights were with fighters in their prime age and the best fighters of the period and along the side women ufc division was coming up, even this played a factor in the spike of their sales of pay per view.