# Will Viruet

4/8/2024

# UFC-Fight Historical Data From 1993 to 2021

Data Source

Summary

This is a list of every UFC fight in the history of the organization. Every row contains information about both fighters, fight details and the winner. The data was scraped from ufc stats website.

Why I chose this data

I selected this dataset due to its relevance to my interest in the UFC and its timing coinciding with the upcoming UFC 300 PPV event April 13th 2024. It's a strategic choice to combine personal interest with professional analysis in the realm of mixed martial arts.

Data Profile

Clean Data

## ## Check for Mixed - Type Data

# Function to check for mixed data types

def check\_mixed\_data\_types(df):

mixed\_data\_columns = []

for column in df.columns:

types = df[column].apply(type).unique()

if len(types) > 1:

mixed\_data\_columns.append(column)

return mixed\_data\_columns

# Check for mixed data types

mixed\_columns = check\_mixed\_data\_types(df)

# Print columns with mixed data types

if mixed\_columns:

print("Columns with mixed data types:")

print(mixed\_columns)

else:

print("No columns with mixed data types.")

Columns with mixed data types:

['Referee', 'B\_Stance', 'R\_Stance']

# Examine the data in the 'Referee' column

print(df['Referee'].head())

# Check for null values

print("Null values in 'Referee' column:", df['Referee'].isnull().sum())

# Inspect data types

print("Data types in 'Referee' column:", df['Referee'].apply(type).unique())

0 Chris Tognoni

1 Herb Dean

2 Herb Dean

3 Mark Smith

4 Mark Smith

Name: Referee, dtype: object

Null values in 'Referee' column: 32

Data types in 'Referee' column: [<class 'str'> <class 'float'>]

## Mixed data type is due to nul values ('NaN') No action taken.

## ##Check for Missing Values[¶](http://localhost:8888/notebooks/UFC-Fight%20Historical%20Data%20From%201993%20to%202021.ipynb#Check-for-Missing-Values)

# Check for missing values in each column

missing\_values = df.isnull().sum()

# Print the count of missing values for each column

print("Missing values in each column:")

print(missing\_values)

Missing values in each column:

R\_fighter 0

B\_fighter 0

Referee 32

date 0

location 0

...

R\_Height\_cms 4

R\_Reach\_cms 406

R\_Weight\_lbs 2

B\_age 172

R\_age 63

Length: 144, dtype: int64

### Missing values remain unchanged in the dataset to preserve its original integrity. This decision ensures transparency and avoids potential biases from imputation methods.

## Check for duplicate Values

# Check for duplicate rows in the DataFrame

duplicate\_rows = df[df.duplicated()]

# Print the duplicate rows

if not duplicate\_rows.empty:

print("Duplicate rows:")

print(duplicate\_rows)

else:

print("No duplicate rows found.")

## No duplicate rows found.

Understanding Data

* 1, Number of Observations (Count): There are 4585 pieces of information for each aspect we're looking at.
* 2. Averages (Mean): On average, fighters' stats vary. For example, the typical rate at which they land significant strikes is around 45.33%, and their average success rate for takedowns is about 29.27%.
* 3. Variability (Standard Deviation): The numbers for each stat vary quite a bit. For example, when it comes to landing significant strikes, the numbers can differ by about 13.05% from the average. This shows that fighters have different levels of success with their strikes.
* 4. Range of Values (Minimum and Maximum): The stats range from the lowest to the highest. For instance, the lowest average significant strike percentage is 0%, meaning some fighters may not land many significant strikes at all. On the other hand, the highest can go up to 100%, indicating some fighters are extremely effective with their strikes.

Content

Each row is a compilation of both fighter stats. Fighters are represented by 'red' and 'blue' (for red and blue corner). So for instance, red fighter has the compiled average stats of all the fights except the current one. The stats include damage done by the red fighter on the opponent and the damage done by the opponent on the fighter (represented by 'opp' in the columns) in all the fights this particular red fighter has had, except this one as it has not occured yet (in the data). Same information exists for blue fighters. The target variable is 'Winner' which is the only column that tells you what happened.

### Column definitions:

* R\_ and B\_ prefix signifies red and blue corner fighter stats respectively
* \_opp\_ containing columns is the average of damage done by the opponent on the fighter
* KD is number of knockdowns
* SIG\_STR is no. of significant strikes 'landed of attempted'
* SIG\_STR\_pct is significant strikes percentage
* TOTAL\_STR is total strikes 'landed of attempted'
* TD is no. of takedowns
* TD\_pct is takedown percentages
* SUB\_ATT is no. of submission attempts
* PASS is no. times the guard was passed?
* REV is the no. of Reversals landed
* HEAD is no. of significant strikes to the head 'landed of attempted'
* BODY is no. of significant strikes to the body 'landed of attempted'
* CLINCH is no. of significant strikes in the clinch 'landed of attempted'
* GROUND is no. of significant strikes on the ground 'landed of attempted'
* win\_by is method of win
* last\_round is last round of the fight (ex. if it was a KO in 1st, then this will be 1)
* last\_round\_time is when the fight ended in the last round
* Format is the format of the fight (3 rounds, 5 rounds etc.)
* Referee is the name of the Ref
* date is the date of the fight
* location is the location in which the event took place
* Fight\_type is which weight class and whether it's a title bout or not
* Winner is the winner of the fight
* Stance is the stance of the fighter (orthodox, southpaw, etc.)
* Height\_cms is the height in centimeter
* Reach\_cms is the reach of the fighter (arm span) in centimeter
* Weight\_lbs is the weight of the fighter in pounds (lbs)
* age is the age of the fighter
* title\_bout Boolean value of whether it is title fight or not
* weight\_class is which weight class the fight is in (Bantamweight, heavyweight, Women's flyweight, etc.)
* no\_of\_rounds is the number of rounds the fight was scheduled for
* current\_lose\_streak is the count of current concurrent losses of the fighter
* current\_win\_streak is the count of current concurrent wins of the fighter
* draw is the number of draws in the fighter's ufc career
* wins is the number of wins in the fighter's ufc career
* losses is the number of losses in the fighter's ufc career
* total\_rounds\_fought is the average of total rounds fought by the fighter
* total\_time\_fought(seconds) is the count of total time spent fighting in seconds
* total\_title\_bouts is the total number of title bouts taken part in by the fighter
* win\_by\_Decision\_Majority is the number of wins by majority judges decision in the fighter's ufc career
* win\_by\_Decision\_Split is the number of wins by split judges decision in the fighter's ufc career
* win\_by\_Decision\_Unanimous is the number of wins by unanimous judges decision in the fighter's ufc career
* win\_by\_KO/TKO is the number of wins by knockout in the fighter's ufc career
* win\_by\_Submission is the number of wins by submission in the fighter's ufc career
* win\_by\_TKO\_Doctor\_Stoppage is the number of wins by doctor stoppage in the fighter's ufc career

Limitations

1) Incomplete data:

Missing Values

Referee 32

R\_Height\_cms 4

R\_Reach\_cms 406

R\_Weight\_lbs 2

B\_age 172

R\_age 63

2) Data Accuracy: Despite efforts to scrape and organize the data accurately, errors or inconsistencies may still exist due to variations in information presentation on the UFCStats website and potential errors during scraping and processing.

3) Data Bias: The dataset might favor certain fighters, events, or regions because of factors like popularity, media attention, or data availability. This could affect the accuracy of any analyses or conclusions made from the dataset.

Ethical Considerations

1) When scraping data from websites, it's crucial to adhere to the site's terms of service and copyright regulations. Ethical concerns may arise if scraped data is utilized without appropriate attribution or permission.

2) Data Privacy: Data Privacy: The dataset includes information about individual fighters and their performances. It's crucial to respect their privacy rights and comply with privacy regulations when using the data.

Key Questions

1) Which fighter has the highest win rate in UFC history, and what factors contribute to their success?

2) How does the frequency of knockouts vary across different weight classes in the UFC, and are there any trends or patterns?