Ultimate Fighting Championship (UFC) Project Document

Data

Source: The data set used for this analysis is open-sourced and is from the resource kaggle.com.
This is the Kaggle Link that contains all of the data sets.

Collection: The following data contains a list of every fight in the history of the organization of the UFC. The data was scraped from the ufcstats website. The tools used included beautifulsoup to scrape the data and pandas to process it.

Sources: ufcstats website
Scraping module
Owner of the Dataset and other projects

Content: There are 4 data sets in total that make up the entire analysis. The titles and dimensions (rows x columns) of the data sets are:

'data' (6012 x 144)

'preprocessed_data' (5902 x 160)

'raw_fighter_details' (3596 x 14)

'raw_total_fight_data' (6012 x 1)

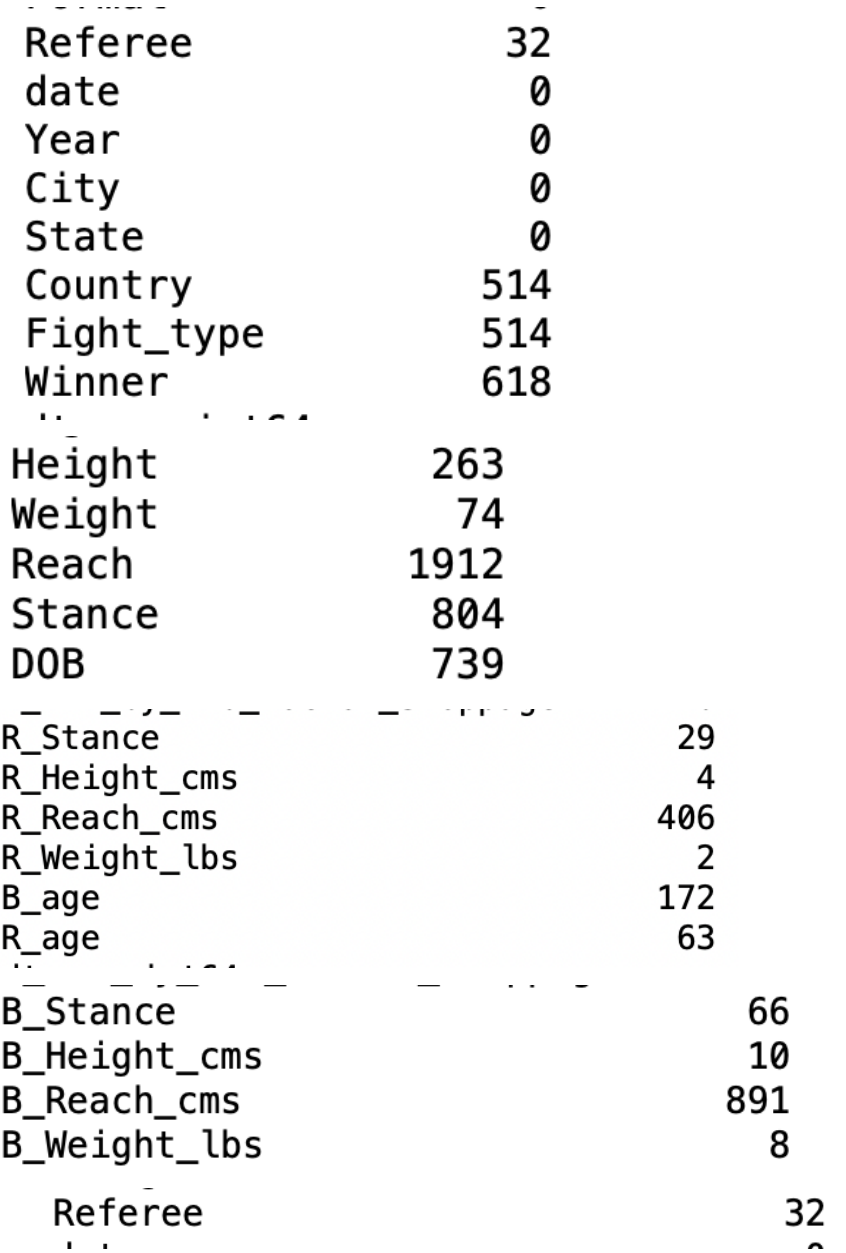
Data Quality:

There are 0 duplicates in all four data sets

There are missing data in the following datasets: 'data', 'raw_fighter_details' and 'raw_total_fight_data'

Merged Dataframes:
‘Data’ & ‘raw_total_fight_data’ on ‘R_fighter’ & ‘B_fighter’

‘preprocessed_data’ will not be used in the analysis as it was discovered that the information is contained in other dataframes that were merged together.



There are also numerous stats for ‘B’ fighters with 1427 missing entries and stats for ‘R’ fighters with 712 missing entries. When checking these entries, it is confirmed that each of these missing columns occur within the same fighters, so we will remove these rows from our clean dataset. We will also remove the other

null values as it is necessary that every entry includes data for every column for completeness.
The new dimensions of the clean datasets are:
‘df_data’ (3890, 144), ‘df_pre’ (5902, 160), ‘df_fighter’ (1661, 14), ‘df_fight_total’ (5368, 44)

Profile

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
- Qual
- Invar

KD is number of knockdowns
- Quan
- Invar

SIG_STR is no. of significant strikes 'landed of attempted'
- Quan
- Invar

SIG_STR_pct is significant strikes percentage
- Quan
- Invar

TOTAL_STR is total strikes 'landed of attempted'
- Quan
- Invar

TD is no. of takedowns
- Quan
- Invar

TD_pct is takedown percentages
- Quan
- Invar

SUB_ATT is no. of submission attempts
- Quan
- Invar

PASS is no. times the guard was passed
- Quan
- Invar

REV is the no. of Reversals landed
    Quan
    Invar
HEAD is no. of significant strikes to the head 'landed of attempted'
    Quan
    Invar
BODY is no. of significant strikes to the body 'landed of attempted'
    Quan
    Invar
CLINCH is no. of significant strikes in the clinch 'landed of attempted'
    Quan
    Invar
GROUND is no. of significant strikes on the ground 'landed of attempted'
    Quan
    Invar
win_by is method of win
    Qual
    Invar
last_round is last round of the fight (ex. if it was a KO in 1st, then this will be 1)
    Quan
    Invar
last_round_time is when the fight ended in the last round
    Quan
    Var
Format is the format of the fight (3 rounds, 5 rounds etc.)
    Qual
    Invar
Referee is the name of the Ref
    Qual
    Invar
date is the date of the fight
    Qual
    Var
location is the location in which the event took place
    Qual
    Invar
Fight_type is which weight class and whether it's a title bout or not

Qual
Invar
Winner is the winner of the fight
Qual
Invar

Stance is the stance of the fighter (orthodox, southpaw, etc.)
Qual
Invar
Height_cms is the height in centimeter
Quan
Invar
Reach_cms is the reach of the fighter (arm span) in centimeter
Quan
Invar
Weight_lbs is the weight of the fighter in pounds (lbs)
Quan
Invar
age is the age of the fighter
Quan
Var
title_bout Boolean value of whether it is title fight or not
Qual
Invar
weight_class is which weight class the fight is in (Bantamweight, heavyweight, Women's flyweight, etc.)
Qual
Invar
no_of_rounds is the number of rounds the fight was scheduled for
Quan
Invar
current_lose_streak is the count of current concurrent losses of the fighter
Quan
Invar
current_win_streak is the count of current concurrent wins of the fighter
Quan
Var

draw is the number of draws in the fighter's ufc career
Quan
Var
wins is the number of wins in the fighter's ufc career
Quan
Var
losses is the number of losses in the fighter's ufc career
Quan
Var
total_rounds_fought is the average of total rounds fought by the fighter
Quan
Var
total_time_fought(seconds) is the count of total time spent fighting in seconds
Quan
Var
total_title_bouts is the total number of title bouts taken part in by the fighter
Quan
Var
win_by_Decision_Majority is the number of wins by majority judges decision in the fighter's ufc career
Quan
Var
win_by_Decision_Split is the number of wins by split judges decision in the fighter's ufc career
Quan
Var
win_by_Decision_Unanimous is the number of wins by unanimous judges decision in the fighter's ufc career
Quan
Var
win_by_KO/TKO is the number of wins by knockout in the fighter's ufc career
Quan
Var
win_by_Submission is the number of wins by submission in the fighter's ufc career
Quan
Var

win_by_TKO_Doctor_Stoppage is the number of wins by doctor stoppage in the fighter's ufc career

Quan

Var

Data Limitations and Ethics

Limitations:
There are some measurement variables that contain a lot more missing data than preferred, however it was possible to remove the missing data to obtain clean data.

Data Ethics Issues:
There is PII data since there are names of the UFC fighters, however, all of this information is available online to use so there are no data ethics issues.

Analysis Questions

- What are the key factors for winners in UFC fights?

- Does the location of the fight have a factor in how long fights are or in strategy?

- How does the weight class affect the outcome of fights?