EXPLORING FIGHTER PERFORMANCE AND GEOGRAPHIC TRENDS IN UFC FIGHTS: AN ANALYTICAL APPROACH

DATA SOURCE

I decided to use a publicly available UFC fight dataset, which can be found on Kaggle under "**UFC-Fight historical data from 1993 to 2021**". The data set covers mixed martial arts (MMA) fights, providing information about fighters (such as height, reach, weight, and age), fight results, betting odds, finishing methods, and fight locations, among other variables. The primary reasons for choosing this data were personal interest in combat sports and the rich variety of numerical and categorical features—enabling advanced exploratory, geospatial, and predictive analyses.

The data appears to have been collected and aggregated from multiple official and fan-driven MMA websites. While the source is publicly available and free to use for non-commercial purposes, it is always important to confirm the legitimacy of sports and betting data. In this case, Kaggle typically hosts datasets under open licenses that support learning and academic portfolio projects.

DATA LIMITATIONS AND ETHICS

There are some limitations and ethical considerations related to the data. First, the dataset relies on historical fight records compiled by third parties, so occasional errors or missing fields could exist. Second, fighter health and performance data can be sensitive, but the dataset does not include personally identifiable medical information; therefore, significant privacy concerns are minimized. Third, the data is updated infrequently, so more recent fights may not be reflected, and the dataset may contain some data points with partially accurate or estimated values (especially for metrics like height and reach). Finally, from an ethical standpoint, this project aims to analyse fight trends rather than make any harmful or exploitative predictions about individuals.

DATA PROFILE

The dataset contains 6,528 rows and 52 columns. The key columns include fighter attributes (height, reach, weight, age), fight descriptors (finish type, finish round, total fight time), betting odds, and location details (city, state, country). Below is a short overview of the data structure and descriptive statistics:

• Each row represents a unique fight result, with "RedFighter" and "BlueFighter" indicating which athlete was listed as "red corner" vs. "blue corner." • The "Finish"

column categorizes how the fight ended (KO/TKO, submission, decision, etc.), while "FinishDetails" can hold more granular information (e.g., "Rear Naked Choke," "Elbows"). • Numerical columns like "RedReachCms," "BlueReachCms," "RedHeightCms," and "BlueHeightCms" track the anthropometric measures of each fighter. Odds columns ("RedOdds," "BlueOdds") show betting-line information. • The "Location" and "Country" columns contain geographical data for events, enabling a geospatial analysis of fight locations around the world. • Basic summary statistics reveal average values. For example, the average fight time was about 680 seconds (~11 minutes), and the average fighter reach was around 182 centimeters.

DATA CLEANING

I started by filling missing values in the betting odds columns (**RedOdds**, **BlueOdds**, **RedExpectedValue**, **BlueExpectedValue**) with the median for each column. Next, for various fighter statistics (such as average significant strikes and average takedowns), I first imputed each fighter's missing data using the mean values from their own past fights and then used the weight-class average for any remaining gaps. Whenever a fighter's stance was unknown or missing, I labeled it as **"Unknown."**

I handled the **"EmptyArena"** column by assuming that any fight held between March 14, 2020, and December 18, 2021, took place without a live audience, filling those rows with **1.0**, while fights outside that date range were treated as **0.0** if missing. For the ranking columns (like **RMatchWCRank**, **BMatchWCRank**, **RBantamweightRank**, etc.), I replaced any missing entries with **"Unranked."** I then set any missing **Finish** and **FinishDetails** values to **"Unknown."**

Next, I replaced missing values in **FinishRound** and **TotalFightTimeSecs** with the median values for each column, converting both fields to integer types. I also selected the specific columns I intended to keep and dropped the rest. For **FinishRoundTime**, which was recorded in **MM:SS** format, I converted the values to total seconds, creating the new column **FinishRoundTimeSecs**. Any missing entries in **FinishRoundTimeSecs** were filled first with the median within each (weight class, finish type) group, and then with the overall median if still missing.

Finally, I checked for duplicates to determine if any rows were entirely identical, although I didn't explicitly remove them at this stage. With these steps complete, I produced a cleaner, more consistent UFC dataset suitable for further exploration and predictive analytics.

RESERCH QUESTIONS

Fight Outcome Analysis

What is the most common fight outcome? (KO/TKO, Submission, Decision?)

- Helps understand how most fights end.
- **Columns needed:** Finish

Which submission technique is the most common?

- Helps analyze submission trends in the UFC.
- **Columns needed:** FinishDetails

What is the average duration of a UFC fight?

- Gives an idea of how long fights usually last.
- **Columns needed:** TotalFightTimeSecs

Which round has the most finishes?

- Determines if most fights end early or late.
- **Columns needed:** FinishRound

Do fights that go to decision last longer on average?

- Confirms if fights that go the distance take significantly longer.
- **Columns needed:** Finish, TotalFightTimeSecs

Are title fights more likely to go to decision?

- Analyzes if championship fights tend to last longer.
- **Columns needed:** WeightClass, Finish, TotalFightTimeSecs

What is the percentage of fights that end in the first round?

- Helps determine how often early stoppages occur.
- **Columns needed:** FinishRound

What is the shortest and longest fight recorded in the dataset?

- Identifies record-breaking fights.
- **Columns needed:** TotalFightTimeSecs

Fighter Performance & Rankings

Which weight class has the most knockouts?

- Finds out if heavier fighters knock out more opponents.
- **Columns needed:** WeightClass, Finish

Does height and reach play a role in winning fights?

- Checks if having longer reach gives a competitive advantage.
- **Columns needed:** RedReach, BlueReach, Winner

I**s there a difference in performance between Orthodox and Southpaw fighters?**

- Determines if fighting stance affects fight results.
- **Columns needed:** RedStance, BlueStance, Winner

Do fighters with more experience (total fights) win more often?

- Analyzes if fight experience correlates with victory.
- **Columns needed:** RTotalFights, BTotalFights, Winner

Do younger fighters have an advantage over older fighters?

- Examines whether age is a key factor in winning.
- **Columns needed:** RAge, BAge, Winner

Betting & Fight Predictions

1**Do betting odds accurately predict fight outcomes?**

- Determines if odds are a reliable indicator of success.
- **Columns needed:** RedOdds, BlueOdds, Winner

1**What is the average betting underdog winning percentage?**

- Helps see if betting against the favorite is profitable.
- **Columns needed:** RedOdds, BlueOdds, Winner

Are fights with closer betting odds more competitive?

- Analyzes if evenly matched fights last longer.
- **Columns needed:** RedOdds, BlueOdds, TotalFightTimeSecs

T**ime-Based Trends**

Has the average fight duration changed over the years?

- Examines if modern fights last longer or shorter than older ones.
- **Columns needed:** Date, TotalFightTimeSecs

Has the frequency of KO/TKO finishes increased or decreased over time?

- Determines if the sport has become more striking-focused.
- **Columns needed:** Date, Finish

Which years had the most fights?

- Identifies peak periods in UFC history.
- **Columns needed:** Date

Has the UFC become more submission-heavy or striking-heavy over time?

- Tracks trends in fighting styles over different time periods.
- **Columns needed:** Date, Finish, FinishDetails

External Factors: Location, Home Advantage

What is the relationship between fight location and fighter performance?

- Helps see if home advantage plays a role.
- **Columns needed:** Location, Winner

Are fights held in certain countries more likely to end in a finish?

- Checks if some fight locations produce more KOs or Submissions.
- **Columns needed:** Location, Finish