Data Visualizations for UFC Fights

Victor Micha [1]

Abstract

This project presents a comprehensive suite of data visualizations analyzing professional Mixed Martial Arts (MMA) fight data spanning over 5,500 fights and 5,000+ fighters. Four complementary visualization categories explore different analytical dimensions: (1) an interactive D3.js globe visualization displaying geographic fighter distribution and title fights, (2) statistical analysis of betting odds accuracy through seven focused plots, (3) gender-based comparison of fighter physical and performance characteristics, and (4) temporal tracking of elite fighter career trajectories. The visualization suite demonstrates how different visualization techniques—from interactive web-based exploration to statistical plots—can provide multi-dimensional insights into sports data, serving fans, analysts, coaches, and researchers with actionable intelligence about the MMA landscape.

1. Introduction

1.1. Dataset Description

This project analyzes professional Mixed Martial Arts (MMA) fight data through multiple interconnected datasets:

per_fight_data.csv (5,529 fights): Individual fight records containing fighters, betting odds, winners, locations, dates, weight classes, gender, and title bout information.

fighter_stats_with_gender.csv (1,333 fighters): Comprehensive fighter profiles including physical attributes (height, weight, reach, age), performance statistics (wins, losses, striking accuracy, takedown metrics), stance, and gender classification.

pro_mma_fighters.csv (5,152 fighters): Detailed fighter profiles with biographical data, country of origin, win/loss breakdowns by method (KO/submission/decision), and weight class information.

countries.geo.json: GeoJSON data for world map visualization, enabling geographic analysis of fighter distribution.

The dataset encompasses a comprehensive view of the MMA landscape, covering fighters from numerous countries, multiple weight classes (from Flyweight to Heavyweight), both male and female divisions, and fight outcomes spanning several years (approximately 2014-2020).

1.2. Purpose of Visualizations

The visualization suite serves four complementary analytical purposes:

(a) Geographic Analysis: Understanding the global distribution of MMA fighters and title fights, identifying regional concentrations of talent, and exploring country-specific fighter profiles and title bout occurrences.

(b) Betting Market Analysis: Evaluating the accuracy of bookmaker predictions through odds analysis, identifying patterns in upset frequencies, examining temporal trends in prediction accuracy, and understanding how betting markets assess fight competitiveness across different categories.

(c) Demographic & Performance Comparison: Comparing physical and performance characteristics between male and female fighters, identifying which features are most predictive of success, and understanding correlation patterns between various fighter attributes.

(d) Temporal Trends: Tracking fighter performance evolution over time, identifying career trajectories of top fighters, and understanding how win rates and activity levels change across years.

Together, these visualizations provide stakeholders (fans, analysts, fighters, coaches, and researchers) with multi-dimensional insights into the sport of MMA, from global geographic patterns to individual fighter performance trajectories.

2. Design Choices

2.1. Interactive Globe Visualization (HTML/D3.js)

Output: Interactive web-based visualization with three linked SVG panels

2.1.1. Spatial Layout Design

The three-panel horizontal layout (globe + fighter stats + title fights) was chosen to create a natural left-to-right in-

formation flow. The globe serves as the primary interaction point, with contextual information revealed in the right panels upon country hover. See Figure 1 for a screenshot.

2.1.2. Interaction Model - Hover vs Click

A hover-based interaction model was deliberately chosen over click-based interaction for several reasons: (1) Reduces interaction cost—users can rapidly explore multiple countries without repeated clicking, (2) Maintains context—users can see the geographic location while viewing country-specific data, (3) Enables fluid exploration—smooth transitions between countries encourage discovery, and (4) Prevents modal states—no "locked" selection that users must manually clear.

2.1.3. Color Scheme - Dark Theme

The dark aesthetic serves multiple purposes: visual hierarchy where bright data elements (colored bars, text) stand out clearly against dark backgrounds, reduced eye strain for extended exploration sessions, modern aesthetic aligning with contemporary data visualization trends, and focus direction drawing attention to data rather than interface chrome.

2.1.4. Fighter Statistics and Title Fights

The stacked horizontal bar chart design for wins/losses by method provides immediate visual comparison through color coding (distinct colors for each finish type with darker shades for losses), proportional scaling (bars scale to the maximum in the displayed set), information density (displays 8 fighters simultaneously), and metadata inclusion (age, height, weight).

The title fights panel uses a card-based layout mimicking traditional fight promotion materials, with red vs blue corner color coding, proportional odds bars showing implied probabilities, winner indication via gold medal emoji, and contextual metadata (date, location, weight class, gender).

2.2. Betting Odds Accuracy Analysis

Output: Seven static visualizations exploring different facets of odds accuracy

Rather than a single comprehensive visualization, seven focused plots were created to address distinct analytical questions. This modular approach allows each visualization to be optimized for its specific purpose.

2.2.1. Visualization Design Rationale

Confidence vs Accuracy (Scatter Plot): Scatter plots effectively show relationships between continuous variables (odds difference vs accuracy). Encoding sample size through both point size and color (viridis colormap) helps users assess statistical reliability—larger, darker points represent more robust data points.

Win Rates (Bar Charts): Bar charts are optimal for comparing magnitudes. The dual-panel design allows both high-level summary (favorites vs underdogs overall) and detailed breakdown (favorites by odds difference category). Green/red color coding leverages universal color associations (green=success, red=upset).

Upset Frequency: The dual approach shows both relative (percentage) and absolute (count) perspectives. The percentage view reveals trends, while the stacked bars provide context about sample size and absolute frequency.

Temporal Trends: The 2×2 grid layout enables comparison along two dimensions: time granularity (monthly detail vs yearly trends) and metric type (accuracy vs confidence). Secondary axes provide crucial context about data availability.

Category Analysis: Horizontal bars work well for categorical comparisons with many categories (weight classes). The grouped bar chart directly compares gender within weight class, enabling identification of category-specific patterns.

Confusion Matrix: Showing both normalized and absolute versions allows users to assess both proportional accuracy and absolute scale. Different color schemes (diverging vs sequential) are optimized for each representation type.

Distribution Comparison: Three complementary visualizations (violin plots, box plots, histograms) each reveal different distributional aspects. The green/red color scheme maintains consistency with earlier visualizations.

2.3. Statistical Fighter Comparison

Output: Eight static plots comparing male and female fighters

2.3.1. Gender Comparison Focus

The entire visualization suite is organized around gender comparison, reflecting the analytical goal of understanding performance differences and similarities between male and female divisions. This focus required careful design to ensure fair comparison while acknowledging inherent physical differences.

2.3.2. Distribution Plots

KDE (Kernel Density Estimation) curves provide smooth, interpretable representations of distributions without the arbitrary binning of histograms (see Figures 9, 10, 11, 12). The overlapping design enables direct visual comparison of distribution shapes, central tendencies, and spread. Statistical annotations (mean, std, n) provide quantitative ground-

ing.

Height, weight, reach, and age were selected as they represent key physical attributes in combat sports and temporal factors (age as career stage proxy). These features are also objectively measurable and consistently available.

2.3.3. Correlation Analysis

Separate correlation heatmaps for each gender (Figures 13, 14) avoid Simpson's paradox and enable identification of gender-specific correlation patterns. The masked upper triangle eliminates redundancy, and the coolwarm diverging colormap effectively highlights both positive and negative correlations with neutral zero.

2.3.4. Feature Importance

Random Forest regression models (Figures 15, 16) predict total wins from all available features. Random Forest was chosen for its ability to handle non-linear relationships and feature interactions. Separate models acknowledge that predictors of success may differ between genders. Horizontal bars facilitate feature name readability.

2.4. Temporal Performance Tracking

Output: Six line plots tracking top fighters across years

2.4.1. Line Plot Design

Multi-line charts with color-coded fighters are optimal for showing temporal trends and trajectories (see Figures 17, 18, 19, 20, 21, 22). Multiple lines enable direct comparison of fighter careers, while color coding distinguishes individual fighters. Separate plots by gender recognize division differences.

2.4.2. Metric Selection

Three metrics provide complementary insights: (1) Wins per year—absolute success measure, (2) Win rate (percentage)—efficiency/dominance measure normalized for activity, and (3) Fights per year—activity level indicator. Together, they distinguish between different success patterns: dominant active fighters (high wins, high rate, high activity), volume fighters (high wins, low rate, high activity), and selective high-success fighters (low wins, high rate, low activity).

2.4.3. Fighter Selection

Focusing on the top 2 fighters per gender maintains visual clarity (more fighters would create cluttered plots), highlights the most successful fighters that fans recognize, enables detailed trajectory examination, and provides representative examples of elite careers. This is a deliberate tradeoff optimizing for detailed examination rather than broad overview.

3. Technical Details

3.1. Globe Visualization Implementation

Technologies: D3.js v7, HTML5/CSS3, JavaScript ES6

Each panel is an independent SVG element, allowing separate coordinate systems, independent rendering, and modular organization. All three data sources are loaded in parallel using Promise.all to minimize loading time.

Two primary data structures enable fast interaction: *fightersByCountry* maps country names to arrays of fighter objects sorted by wins, enabling O(1) lookup; *titleFightsByCountry* maps countries to fight objects with pre-computed winner information and implied probabilities.

The D3 Natural Earth projection provides balanced distortion across all regions. The projection's scale and translation parameters are dynamically updated for zoom (0.6× to 2.5×) and pan operations. American odds are converted to implied probabilities using standard formulas: negative odds use $|\text{odds}|/(|\text{odds}| + 100) \times 100$, while positive odds use $100/(\text{odds} + 100) \times 100$.

3.2. Odds Visualization Implementation

Technologies: Python 3.x, pandas, matplotlib, seaborn, numpy

The fight data is loaded and immediately filtered to remove incomplete records, ensuring approximately 4,167 valid fights. American odds are converted to implied probabilities, and the fighter with higher implied probability is designated as the favorite.

Several derived features are created: odds_difference (absolute difference between fighter odds), prediction_correct (whether favorite won), upset (underdog victory), and odds_diff_category (categorical binning).

Continuous odds differences are binned into custom ranges (0-25, 25-50, 50-100, 100-150, 150-200, 200-300, 300-500, 500+) to balance sample size and granularity. Date strings are parsed to enable monthly and yearly temporal analysis. Consistent color schemes enhance interpretability: green represents correct predictions/favorites while red represents incorrect/underdogs.

3.3. Statistical Comparison Implementation

Technologies: Python 3.x, pandas, matplotlib, seaborn, scikit-learn, numpy

Fighter statistics are loaded and partitioned by gender. The top 1000 fighters per gender ranked by total wins are se-

lected, focusing on successful fighters with substantial records. Missing data is handled through complete-case analysis (dropna) for distribution plots and mean imputation for machine learning models.

KDE provides smooth probability density estimates for each feature without histogram binning artifacts. Pearson correlation coefficients are calculated separately for each gender to avoid Simpson's paradox. Random Forest regression models (100 trees) predict total wins from all features, providing feature importance scores based on impurity reduction.

3.4. Temporal Tracking Implementation

Technologies: Python 3.x, pandas, matplotlib, datetime

Fight dates (M/D/YYYY format) are parsed to datetime objects with year extraction. For each fight, both fighters' statistics are updated bidirectionally. Fighter statistics are aggregated by year, computing wins, losses, total fights, and win rate.

Fighters must meet minimum thresholds: at least 5 total fights and 3+ years of data. Within each gender, fighters are ranked by total career wins, with the top 2 selected for visualization. Each fighter is represented as a line with distinct color and circular markers.

3.5. Common Technical Patterns

The project follows clean separation: data files in `data/`, output plots in `plots/` and `odds_visualizations/`, scripts in root directory, and web assets in `js/`.

Performance optimization strategies include pre-computed lookups (country-to-fighters mappings), efficient data structures (dictionaries for O(1) access), vectorized operations (pandas/numpy), and D3 data joins (minimizing DOM manipulation).

Reproducibility is ensured through fixed random seeds (random_state=42), standardized parameters (DPI, figure sizes, color schemes), and clear documentation of dependencies and data processing steps.

4. Conclusion

This project demonstrates a comprehensive approach to sports data visualization, combining interactive web-based exploration (D3.js globe), statistical analysis (Python/matplotlib), and multi-dimensional perspectives on MMA fight data. Each visualization category addresses specific analytical questions while maintaining consistent design principles: clarity, interactivity where beneficial, statistical rigor, and aesthetic coherence. The technical implementation leverages industry-standard tools and best practices, ensuring maintainability, reproducibility, and extensibility for future analysis.

Appendix: Visualizations

Globe Visualizations

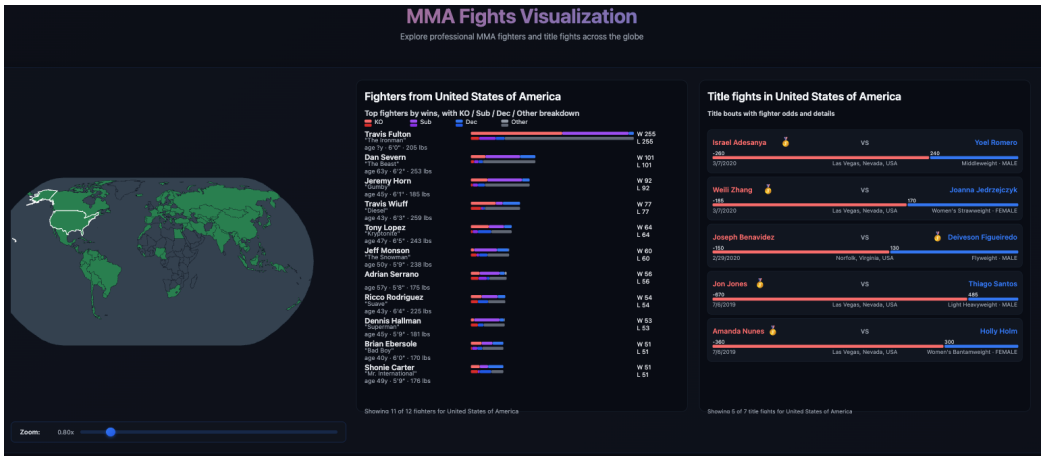

Figure 1. Geographic view of fighters and title fights

Betting Odds Analysis

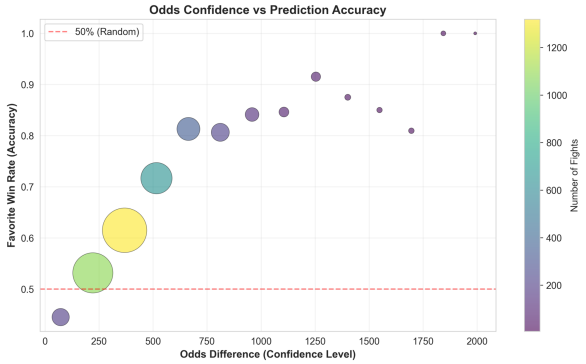

Figure 2. Odds confidence vs prediction accuracy. Larger, darker points represent more fights.

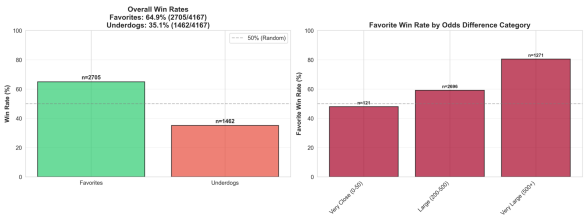

Figure 3. Favorite vs underdog win rates overall and by odds difference category.

Figure 4. Upset frequency by odds margin showing both percentage and absolute counts.

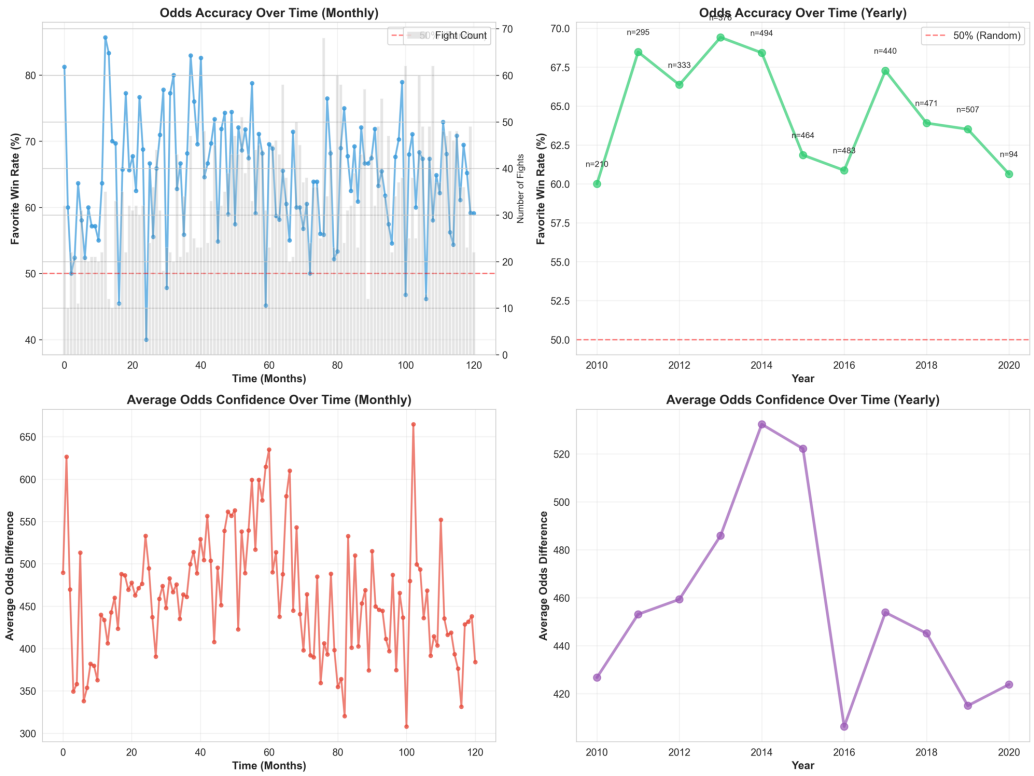

Figure 5. Temporal trends in odds accuracy and confidence (monthly and yearly).

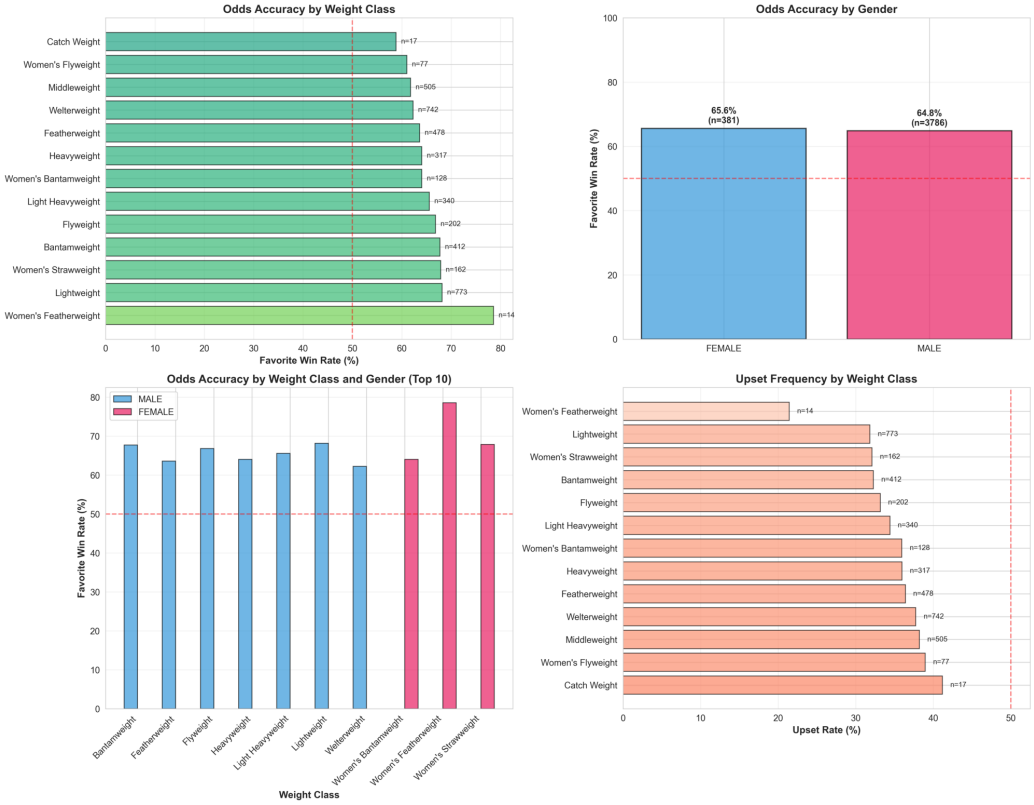

Figure 6. Odds accuracy and upset rates by weight class and gender.

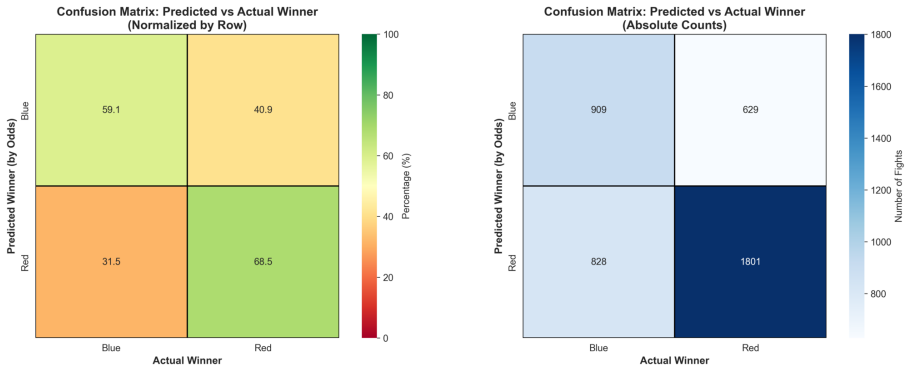

Figure 7. Confusion matrix showing predicted vs actual winners (normalized and absolute).

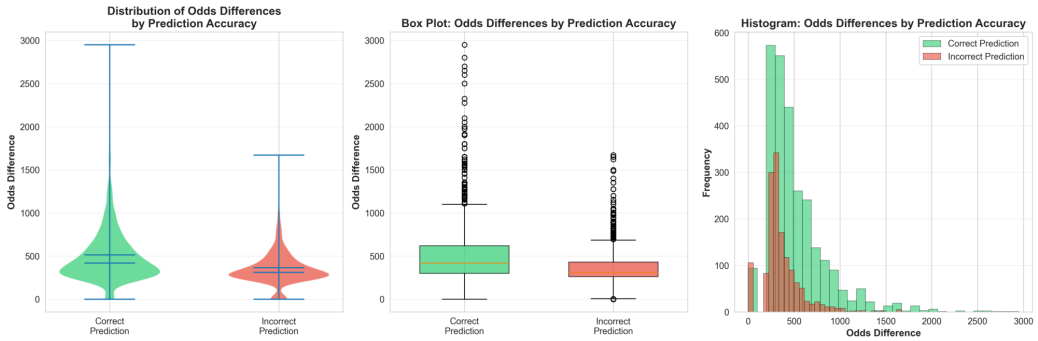

Figure 8. Distribution comparison of odds differences for correct vs incorrect predictions.

Gender Comparison - Physical Distributions

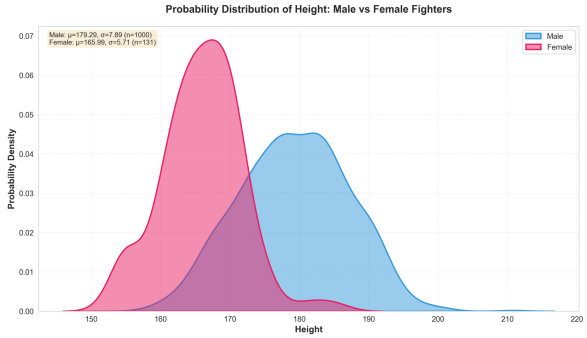

Figure 9. Height distribution comparison between male and female fighters using KDE.

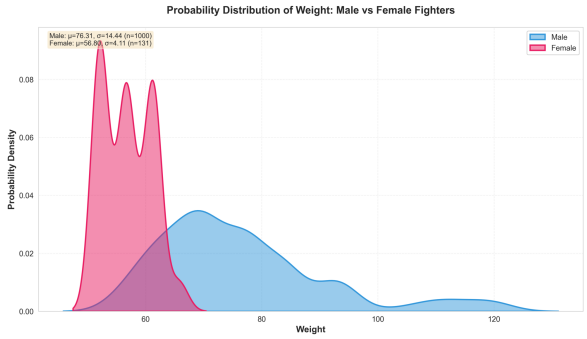

Figure 10. Weight distribution comparison between male and female fighters using KDE.

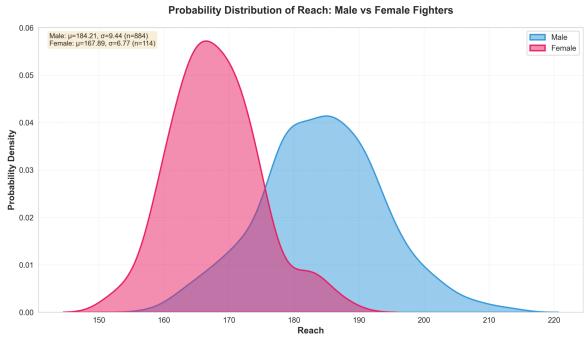

Figure 11. Reach distribution comparison between male and female fighters using KDE.

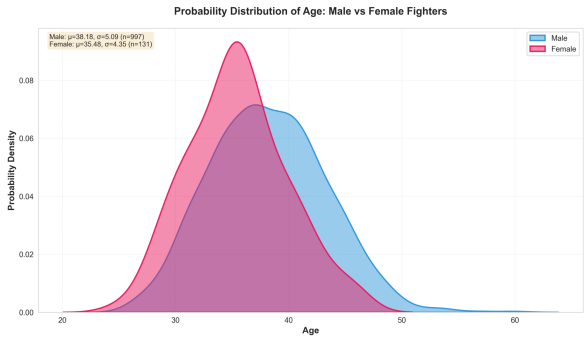

Figure 12. Age distribution comparison between male and female fighters using KDE.

Correlation Analysis

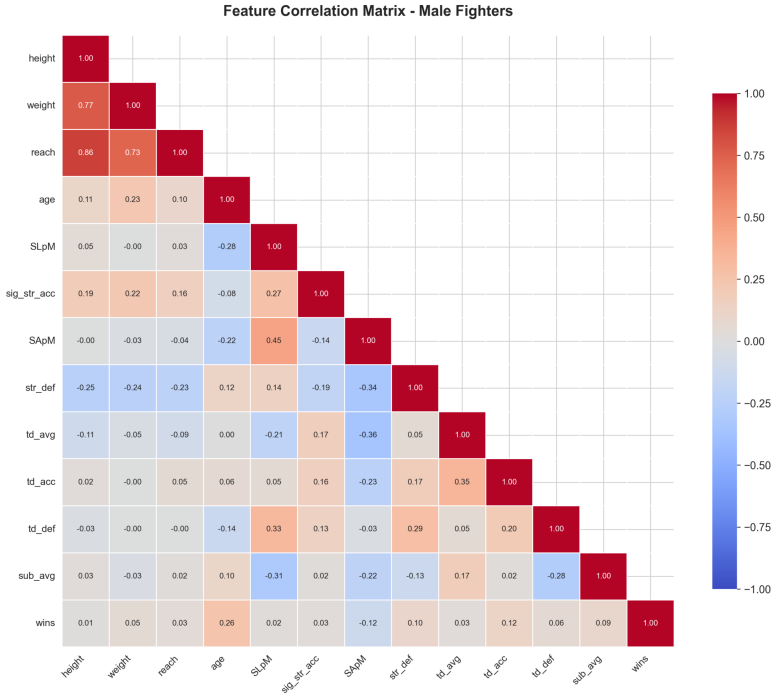

Figure 13. Correlation heatmap for male fighters showing relationships between features and wins.

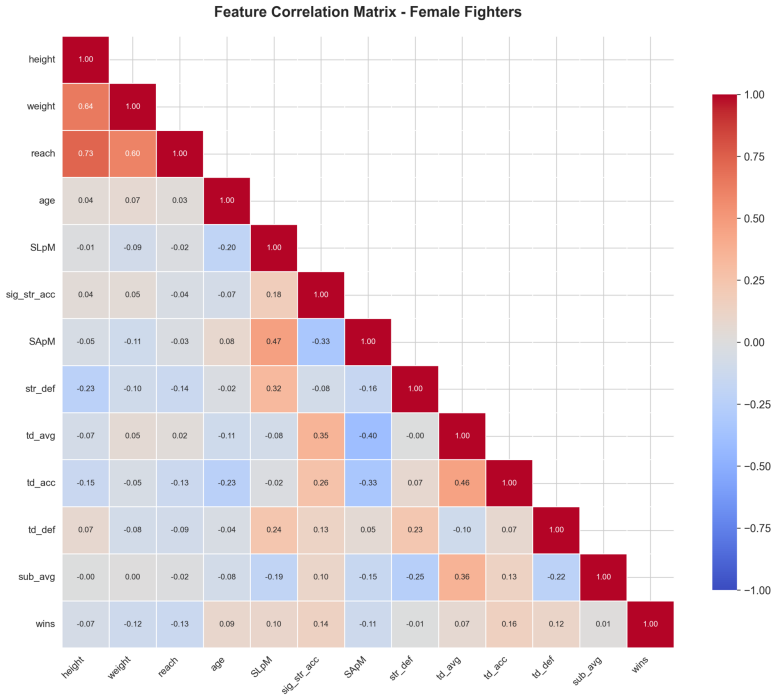

Figure 14. Correlation heatmap for female fighters showing relationships between features and wins.

Feature Importance

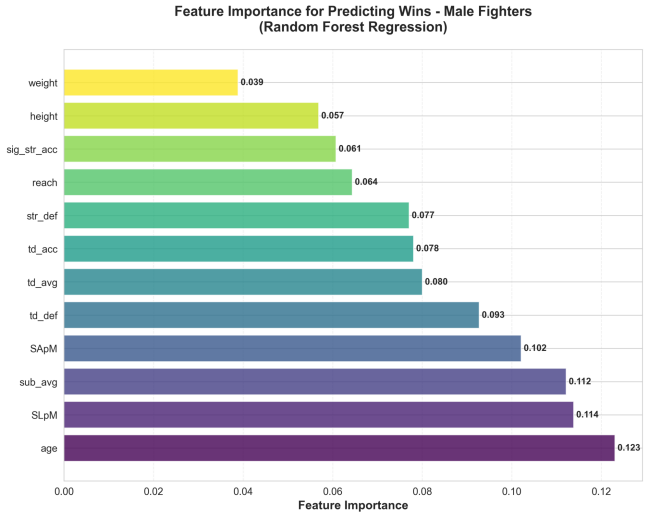

Figure 15. Random Forest feature importance for predicting male fighter wins.

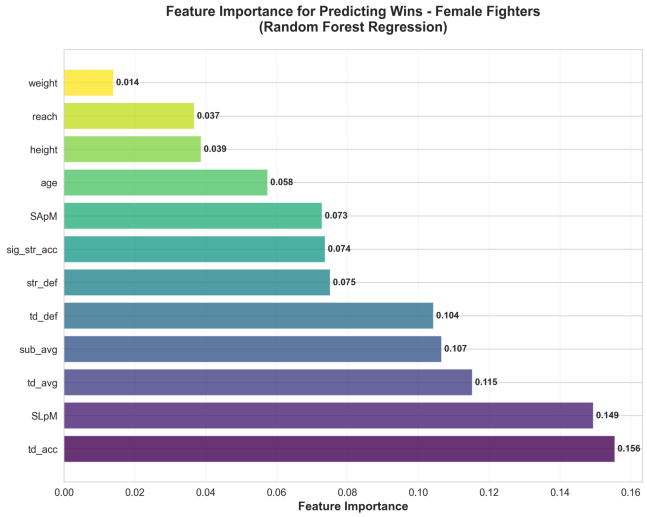

Figure 16. Random Forest feature importance for predicting female fighter wins.

Temporal Trends - Top Fighters

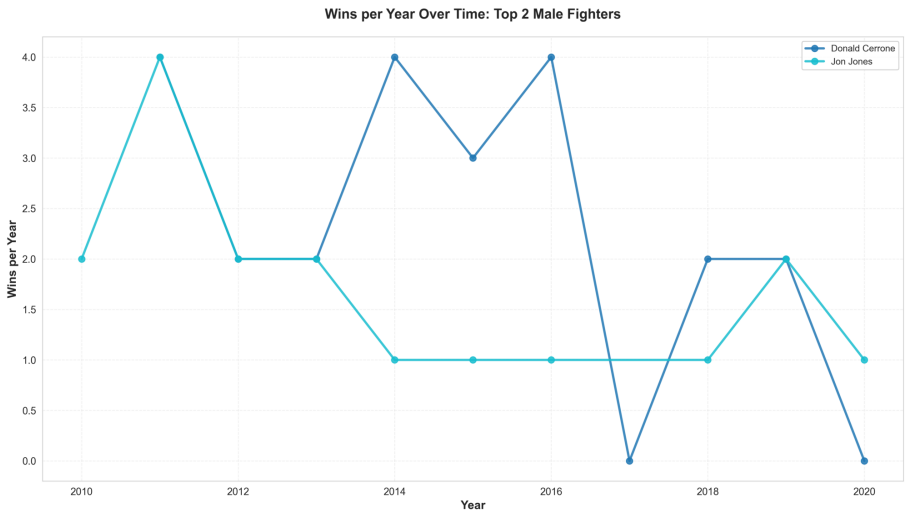

Figure 17. Wins per year for top 2 male fighters showing career trajectories.

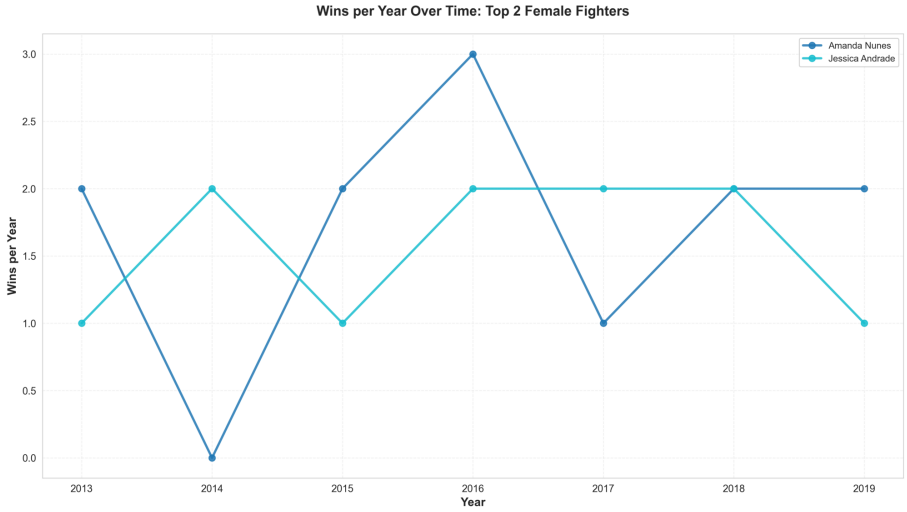

Figure 18. Wins per year for top 2 female fighters showing career trajectories.

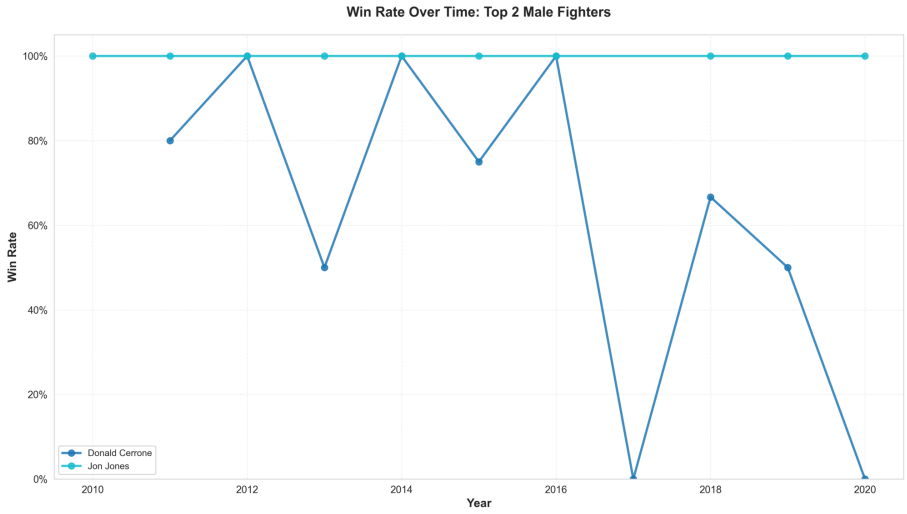

Figure 19. Win rate over time for top 2 male fighters showing efficiency trends.

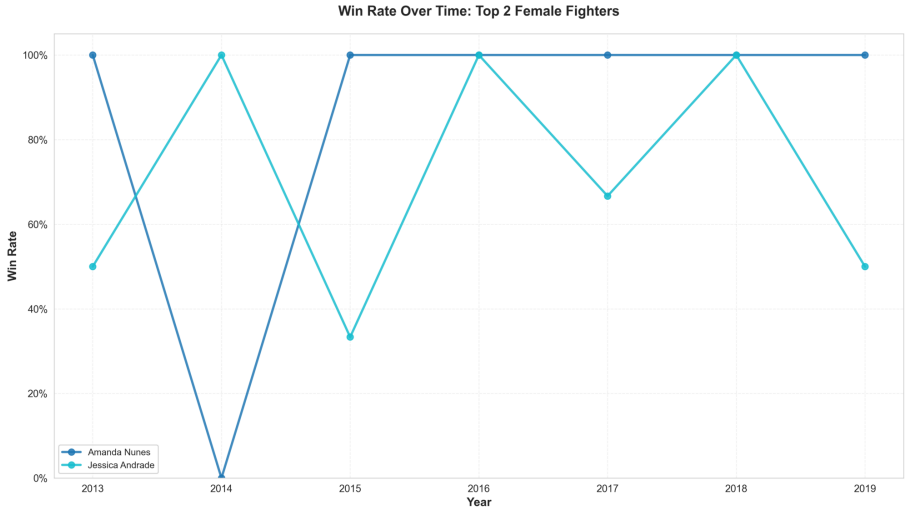

Figure 20. Win rate over time for top 2 female fighters showing efficiency trends.

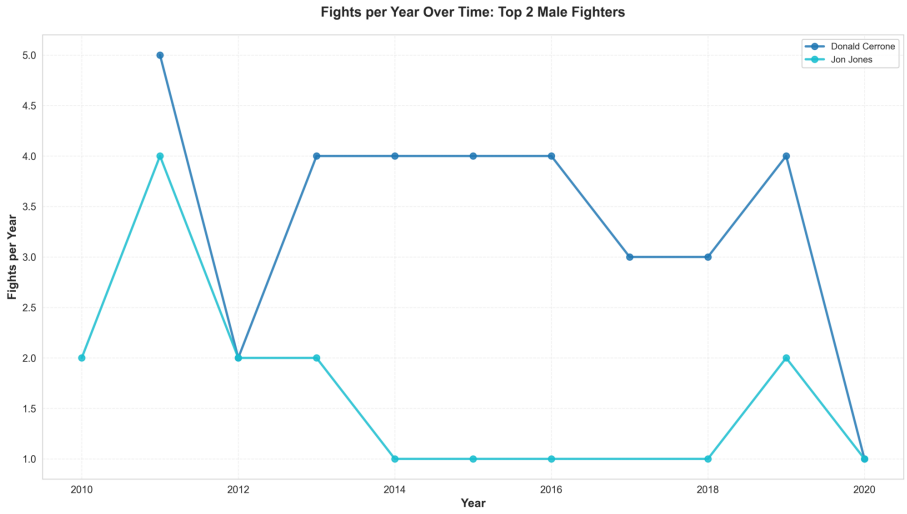

Figure 21. Fights per year for top 2 male fighters showing activity levels.

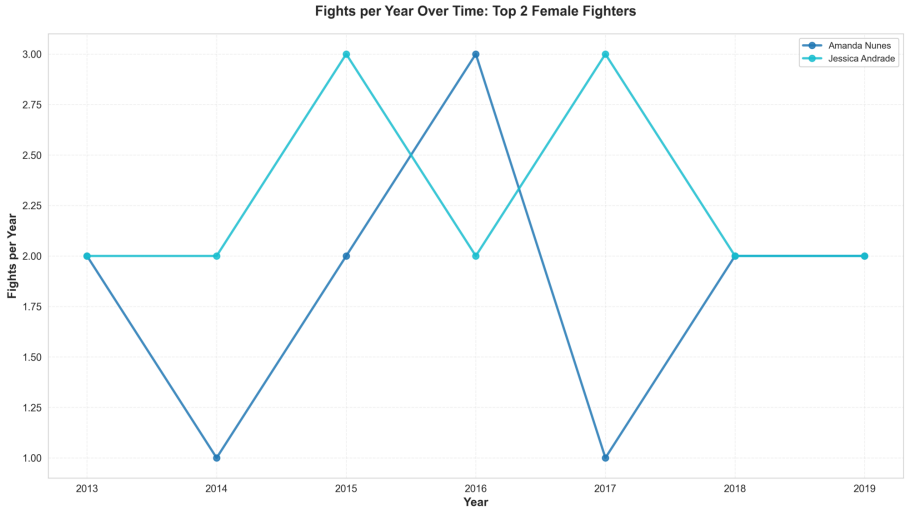

Figure 22. Fights per year for top 2 female fighters showing activity levels.