CHECKPOINT 5 — Exploratory Data Analysis (EDA) Memo

Nicholas Condos

Dec 1, 2025

https://github.com/nicholascondos/Tennis_Analytics

1. Project Question Snapshot

This project investigates **which performance and efficiency factors most strongly differentiate ATP match winners from losers** across all surfaces between 2008–2024. I focus on how serve quality, return effectiveness, ranking, and court surface relate to match outcomes. The goal of this checkpoint is to use exploratory data analysis (EDA) to understand patterns in the data and identify promising features for predictive models in Checkpoint 6.

2. Data Used (Output of Checkpoint 4)

All analyses are based on the cleaned dataset created in Checkpoint 4: `Atp_matches_2008_2024_clean.csv`.

Key properties:

- **Row grain:** one row per completed ATP singles match.
- **Filters:** completed matches only (non-null score and no walkover/retirement comment), main-draw singles (doubles and qualifying/round-robin rounds removed).
- **Time span:** 2008–2024, with standardized dates and surfaces.
- **Engineered features:**
    - `w_srv_ret_balance`, `l_srv_ret_balance`
    - `efficiency_diff` = winner minus loser combined serve-return efficiency.

This dataset also contains detailed serve stats (aces, double faults, first serves in, total serve points), return-related metrics, break point stats, ranking, and surface, which drive the plots below.

3. Univariate Distributions/Plots

CHECKPOINT 5 — Exploratory Data Analysis (EDA) Memo

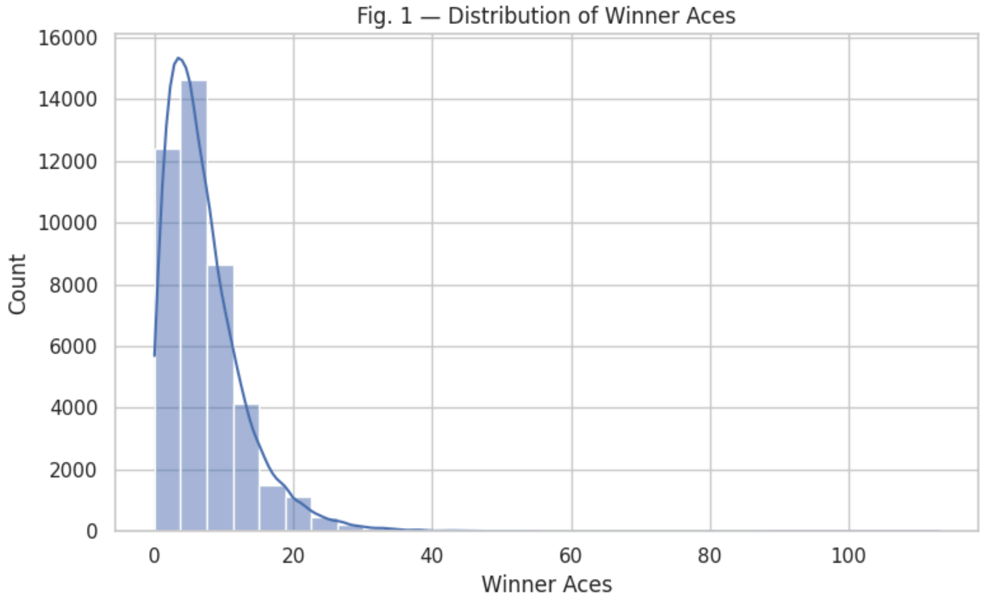
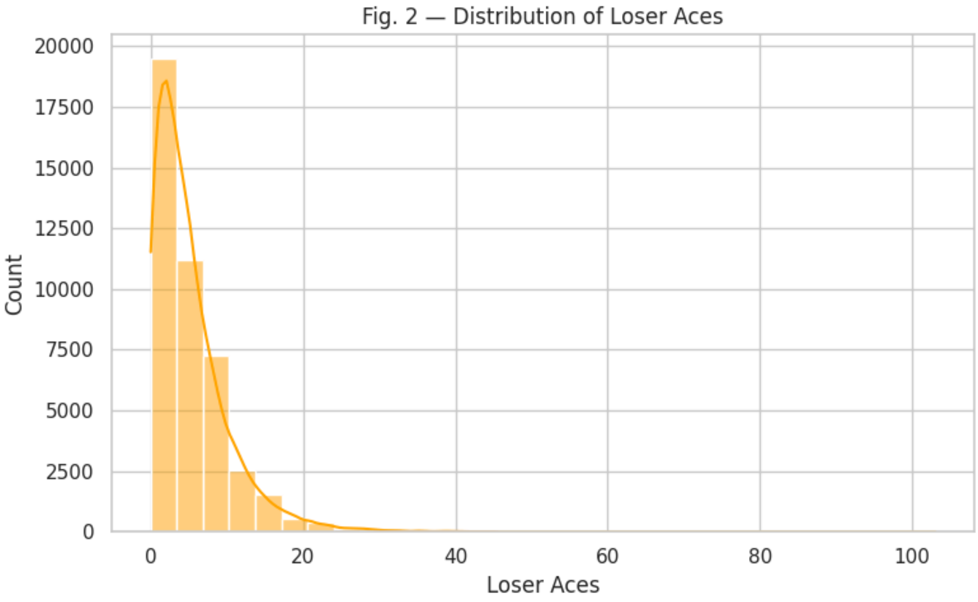


Fig. 1 — Distribution of Winner Aces

This figure shows the distribution of aces hit by match winners. Most winners record between about 3 and 10 aces, with a right-skewed tail where a small number of matches feature extremely high ace counts (over 30, up to 113). The skew indicates that while big servers exist, they are not representative of typical match winners. Aces clearly contribute to success, but the wide spread suggests they are only one part of a broader performance profile.



CHECKPOINT 5 — Exploratory Data Analysis (EDA) Memo

Fig. 2 — Distribution of Loser Aces

Losers generally hit fewer aces than winners, yet the two distributions overlap substantially. Some losing players still produce high ace counts, especially on fast surfaces, which shows that strong serving alone does not guarantee victory. These observations suggest that weaknesses in other areas - such as return games or key points - can outweigh the benefits of high ace totals. As a result, aces should be treated as a contributing but not decisive variable.

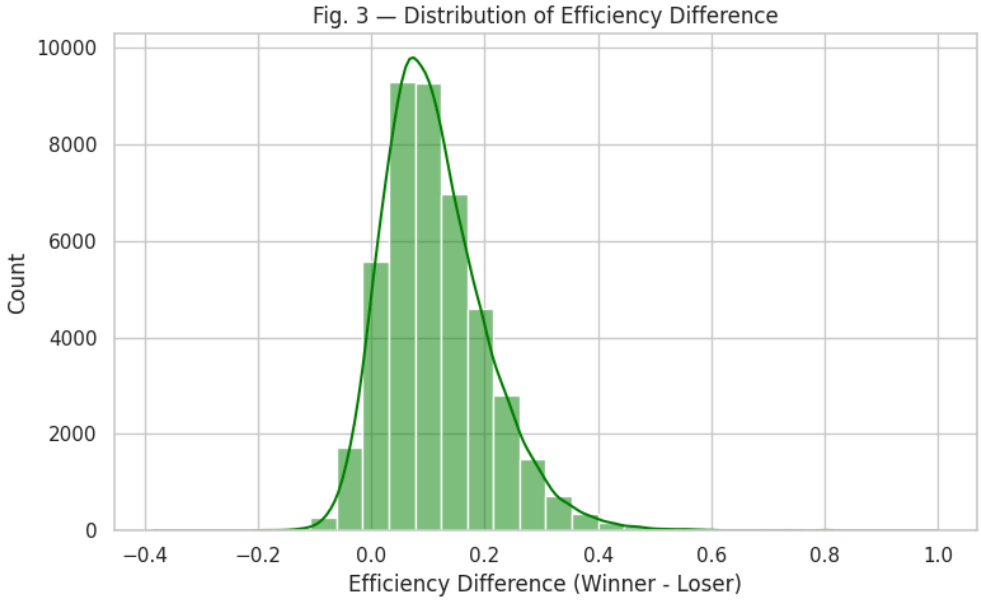


Fig. 3 — Distribution of Efficiency Difference (Winner – Loser)

This plot shows the distribution of `efficiency_diff`, the difference in combined serve-and-return efficiency between winners and losers. The bulk of matches fall within small to moderate gaps (roughly 0.00–0.15), indicating many contests are relatively competitive. Larger efficiency differences represent matches where one player dominated both serve and return, resulting in comfortable wins. The shape of this distribution confirms that `efficiency_diff` is an informative separator and a strong candidate predictor for modeling.

CHECKPOINT 5 — Exploratory Data Analysis (EDA) Memo

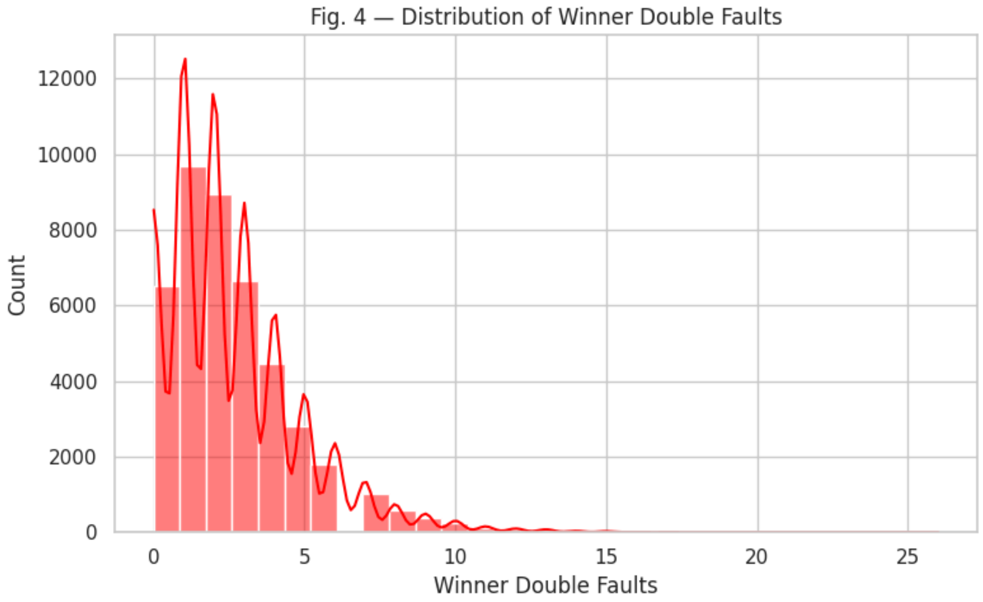
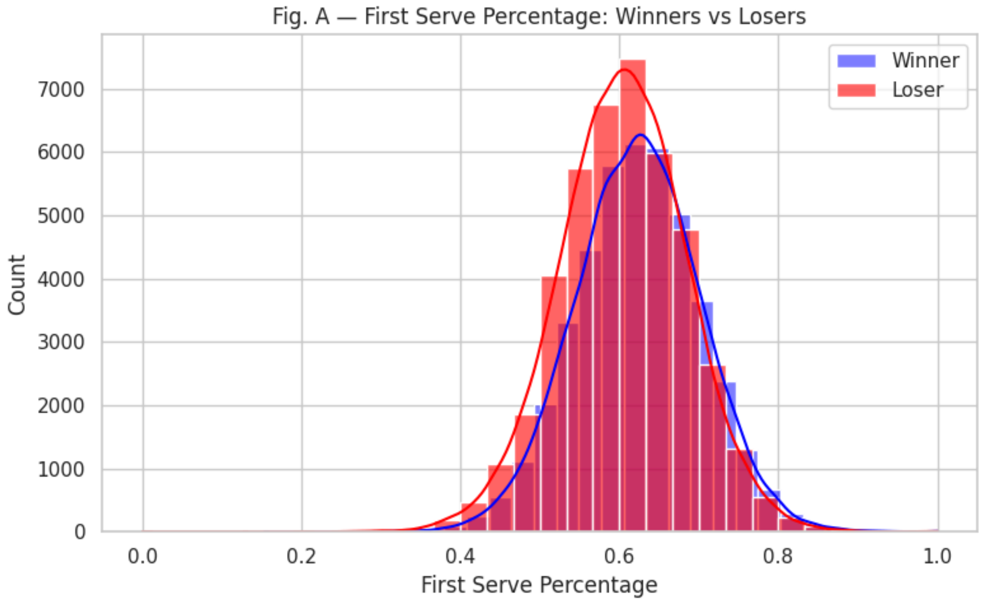


Fig. 4 — Distribution of Winner Double Faults

Here, most winners commit between 1 and 4 double faults, with a tight concentration at low values. This suggests that successful players tend to avoid frequent unforced serving errors, even across different surfaces and match lengths. A handful of outliers with high double-fault counts appear but are rare and plausible (for example, risky serving strategies or difficult conditions). Overall, the pattern indicates that minimizing double faults is characteristic of winning performance, although not a deciding factor.

4. Serve and Return Consistency



CHECKPOINT 5 — Exploratory Data Analysis (EDA) Memo

Fig. A — First-Serve Percentage: Winners vs Losers

This figure compares the first-serve percentage between winners and losers. Winners' distribution is slightly shifted to the right, indicating that they land their first serves more often on average, but the overlap between the curves is substantial. This implies that simply getting a higher % of first serves in is helpful but not sufficient to guarantee a win. It suggests that **first-serve effectiveness** (points won on first serve) and what happens on return games may be more informative than first-serve percentage alone.

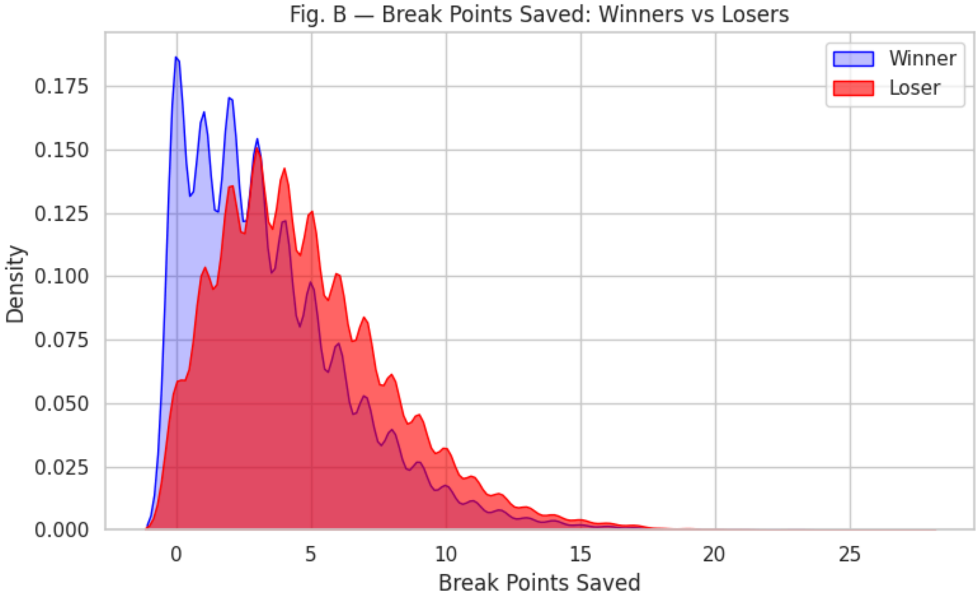


Fig. B — Break Points Saved: Winners vs Losers

This density plot compares the raw number of break points saved by winners and losers. Interestingly, losers generally save more break points in absolute count, with their distribution shifted to the right. This is consistent with the idea that losers usually face more break points overall, giving them more opportunities to save them, even if their save percentage is lower. This plot highlights that raw break-point counts can be misleading without considering how many opportunities each player faced, suggesting that rate-based measures (like break-point save percentage) may be more appropriate for modeling.

CHECKPOINT 5 — Exploratory Data Analysis (EDA) Memo

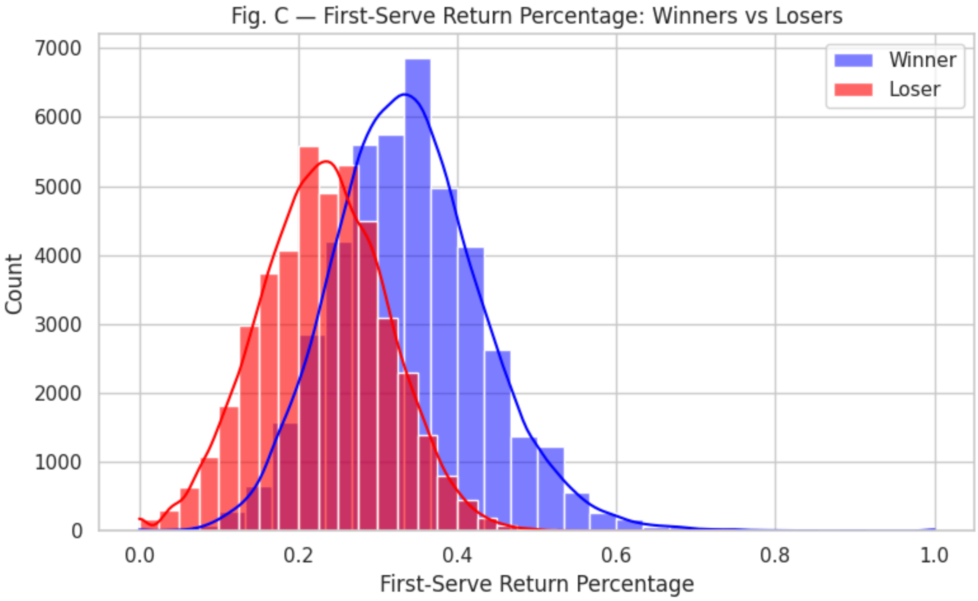
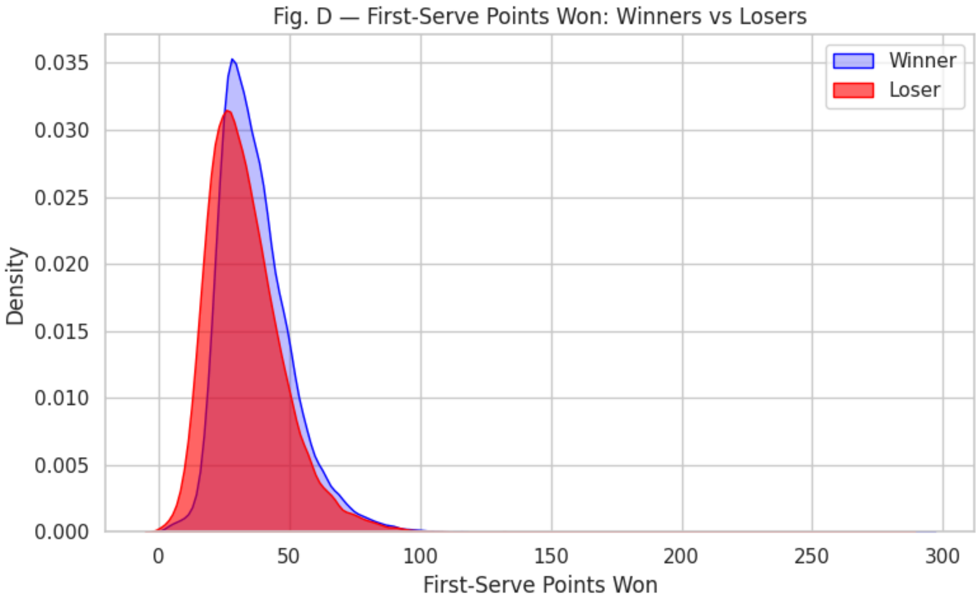


Fig. C — First-Serve Return Percentage: Winners vs Losers

Fig. C examines how effectively players return their opponents' first serves. Winners display a distribution clearly shifted to the right compared to losers, meaning they win a higher share of points when returning first serves. Although there is still overlap, this indicates that neutralizing or attacking the opponent's first serve is an important differentiator of match success. This reinforces the idea that strong return play, not just big serving, is crucial in modern ATP tennis.



CHECKPOINT 5 — Exploratory Data Analysis (EDA) Memo

Fig. D — First-Serve Points Won: Winners vs Losers

This figure compares how many first-serve points winners and losers win in absolute count. The distribution for winners is clearly shifted to the right, indicating that winners typically earn more points outright on their first serve. This reflects stronger serve effectiveness, not just higher first-serve percentages. The separation between the two curves confirms that winning more first-serve points is a meaningful and consistent differentiator in match outcomes.

5. Relationships Between Key Variables

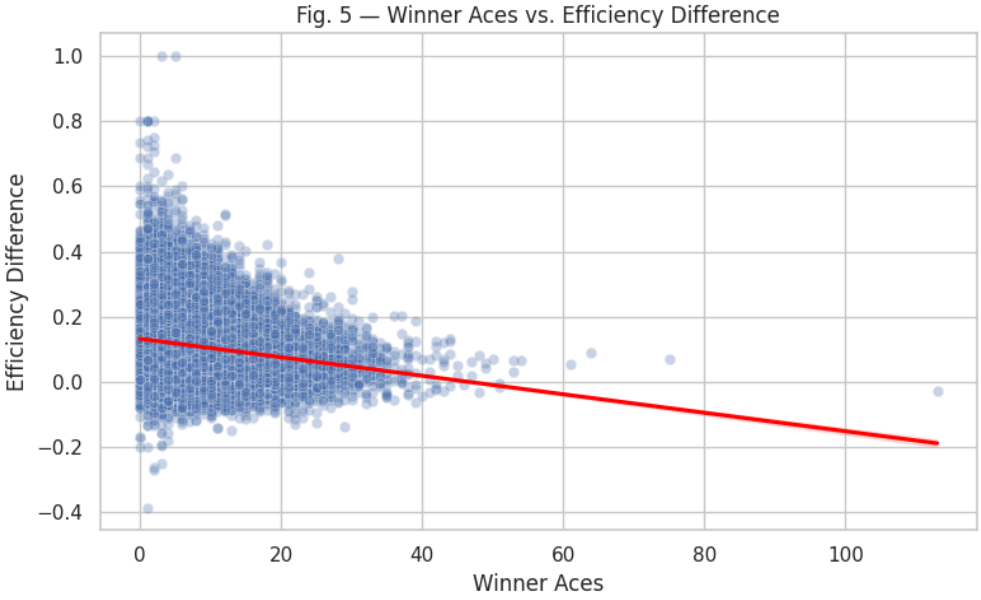


Fig. 5 — Winner Aces vs Efficiency Difference

This scatter plot explores the relationship between winner ace counts and `efficiency_diff`. There is a mild positive association: matches where winners hit more aces often show larger efficiency gaps. However, many matches with moderate ace counts still yield substantial efficiency differences, and some high-ace matches feature only modest gaps. This suggests that aces contribute to overall efficiency but need to be interpreted alongside return and other factors.

CHECKPOINT 5 — Exploratory Data Analysis (EDA) Memo

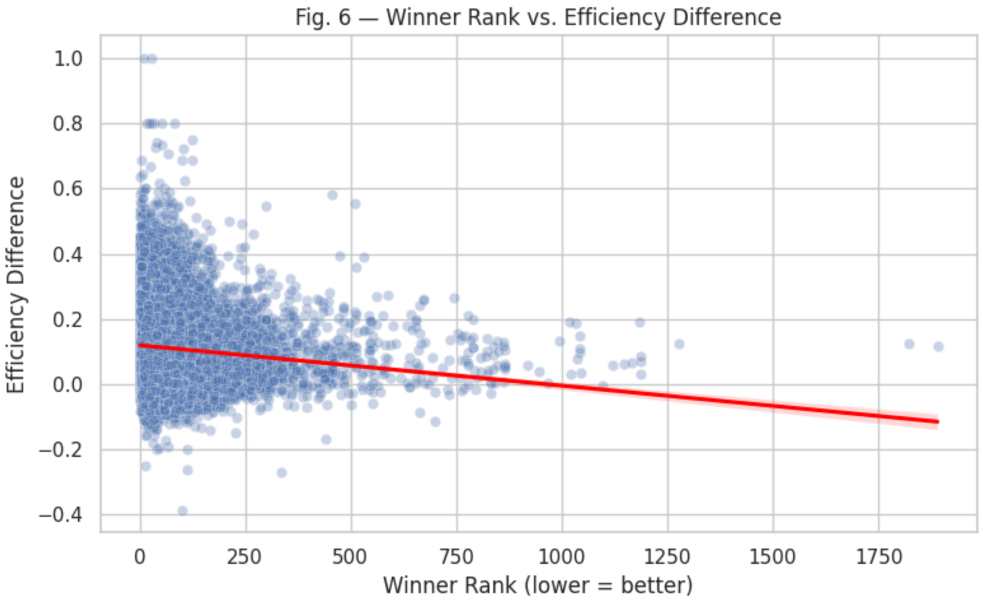


Fig. 6 — Winner Rank vs Efficiency Difference

Fig. 6 relates winner ATP ranking (where a smaller number is a higher rank) to efficiency difference. Higher-ranked players frequently win matches with relatively modest efficiency margins, reflecting their ability to convert small advantages into victories. Lower-ranked players, by contrast, often need larger efficiency spikes to overcome stronger opponents. This pattern matches tennis intuition: underdogs usually must outperform expectations significantly to pull off upsets, while elite players can win even when playing closer to their opponent's level.

6. Surface-Based Subgroup Comparison

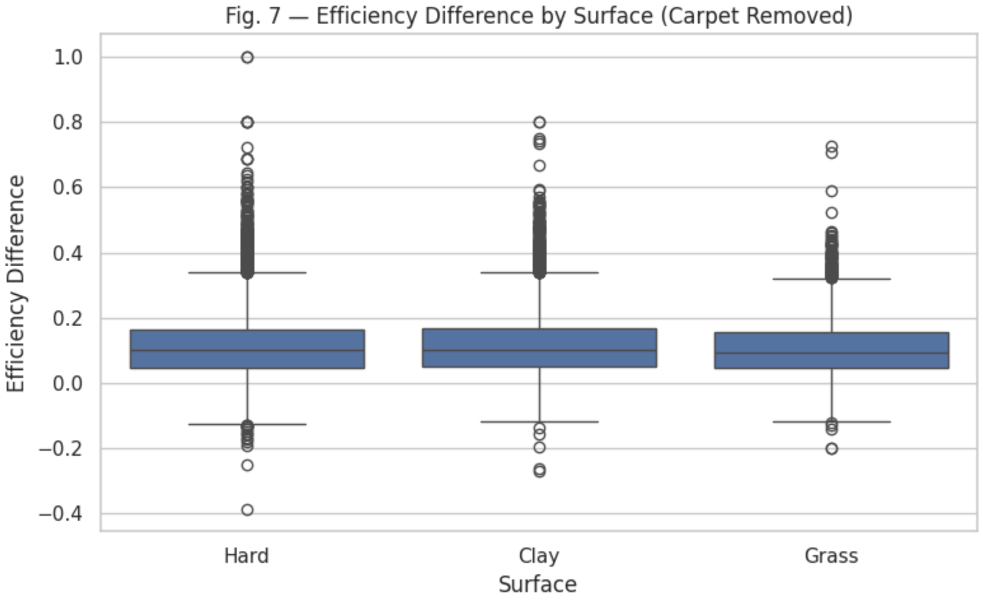


Fig. 7 — Efficiency Difference by Surface (Carpet Removed)

This boxplot compares efficiency differences on Hard, Clay, and Grass courts, after excluding rare Carpet events. Grass exhibits the largest median efficiency gaps and wider spread, consistent with its fast conditions that reward dominant serving and aggressive play. Clay shows the smallest gaps, reflecting longer rallies, more frequent service breaks, and narrower performance margins. Hard courts sit between the two, aligning with their reputation as a balanced surface. These patterns suggest that surfaces may interact with other features and should be considered explicitly in future models.

7. Time Trend

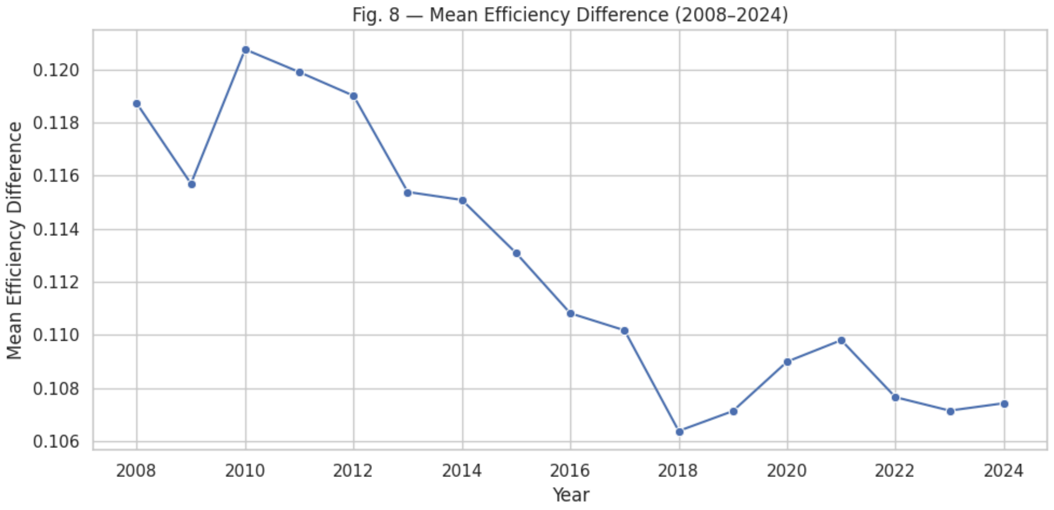


Fig. 8 — Mean Efficiency Difference Over Time (2008–2024)

This time-series plot tracks the annual mean `efficiency_diff` from 2008 through 2024. The series is relatively stable, with only modest year-to-year fluctuations and no strong long-term upward or downward trend. This indicates that the structural balance between winners and losers, in terms of combined serve-and-return efficiency, has remained consistent over the study period. Because there is no major temporal shift, a single unified modeling approach across years is reasonable.

8. Outliers and Data Quality Observations

Across the visualizations, extreme values (such as very high ace counts or unusually high double-fault matches) appear rare but plausible and consistent with known types of ATP matches (e.g., marathon grass-court contests). No obvious data errors or impossible values were detected in these key variables after CP4 cleaning. As a result, no additional data cleaning beyond CP4 is necessary before modeling.

9. Key Takeaways and Modeling Direction (CP6 Preview)

The EDA analysis indicates that **efficiency_diff** is a strong separator between winners and losers and should be a central feature in the modeling phase. Serve-related variables such as aces, double faults, and first-serve percentage show informative but overlapping patterns and thus should be used in combination with efficiency and return metrics rather than alone. Return effectiveness (especially first-serve return percentage) emerges as a critical, somewhat underappreciated factor, while ranking and surface clearly shape how performance translates into outcomes. These insights support a CP6 modeling strategy that includes efficiency_diff, key serve and return stats, ranking, and surface (and potentially interaction terms) in both logistic regression and tree-based models.