Analysis of WTA & ATP Match Statistics (2020-2024)

BIOS 611 Project

2025-12-02

- 1 Data Overview
- 2 Research Question A: Can Unsupervised Learning Distinguish Gender?
  - 2.1 PCA Results
  - 2.2 Clustering Performance
- 3 Research Question B: What kind of player wins?
  - 3.1 Interpretation of PCs via Loadings Plot
  - 3.2 Correlation with Winning
  - 3.3 Visual Evidence
- 4 Research Question C: Mis-clustered Player Analysis
  - 4.1 Why are they mis-clustered?
    - 4.1.1 WTA in ATP Cluster (n=16)
    - 4.1.2 ATP in WTA Cluster (n=96)
    - 4.1.3 Ranking scores
  - 4.2 Visual Summary of Mis-clusters
  - 4.3 Summary
- 5 Conclusion
- 6 Future Direction and Limitations

1 Data Overview

This analysis explores match statistics from the ATP (Men’s) and WTA (Women’s) tours from 2020 to 2024. The goal is to identify gender-based performance patterns, understand factors contributing to player success, and investigate outliers using unsupervised learning.

- **Data Sources:** Combined match data from ATP and WTA CSV files (2020-2024).
- **Sample Size:**
  - Total Players Analyzed: **784** (after filtering for players with ≥ 5 matches)
  - Breakdown: **394** ATP players, **390** WTA players.
  - Total Matches Processed: **25,140**
- **Features Analyzed:** Averages of Aces, Double Faults, Serve Points, 1st Serves In/Won, 2nd Serves Won, Break Points Saved/Faced.

2 Research Question A: Can Unsupervised Learning Distinguish Gender?

Method: Principal Component Analysis (PCA) followed by K-Means Clustering (K=2) on average game statistics.

2.1 PCA Results

The model effectively separates gender based on game statistics. The variables used in this analysis were: **Average Aces, Double Faults, Serve Points, 1st Serves In %, 1st Serve Points Won %, 2nd Serve Points Won %, Serve Games, Break Points Saved, and Break Points Faced.**

- **PC1 (50.4% variance):** Heavily driven by **Serve Power/Volume** (high loadings for Aces, Serve Games, 1st Serve Won).
- **PC2 (25.6% variance):** Relates to **Pressure/Resilience** (Break Points Faced/Saved).

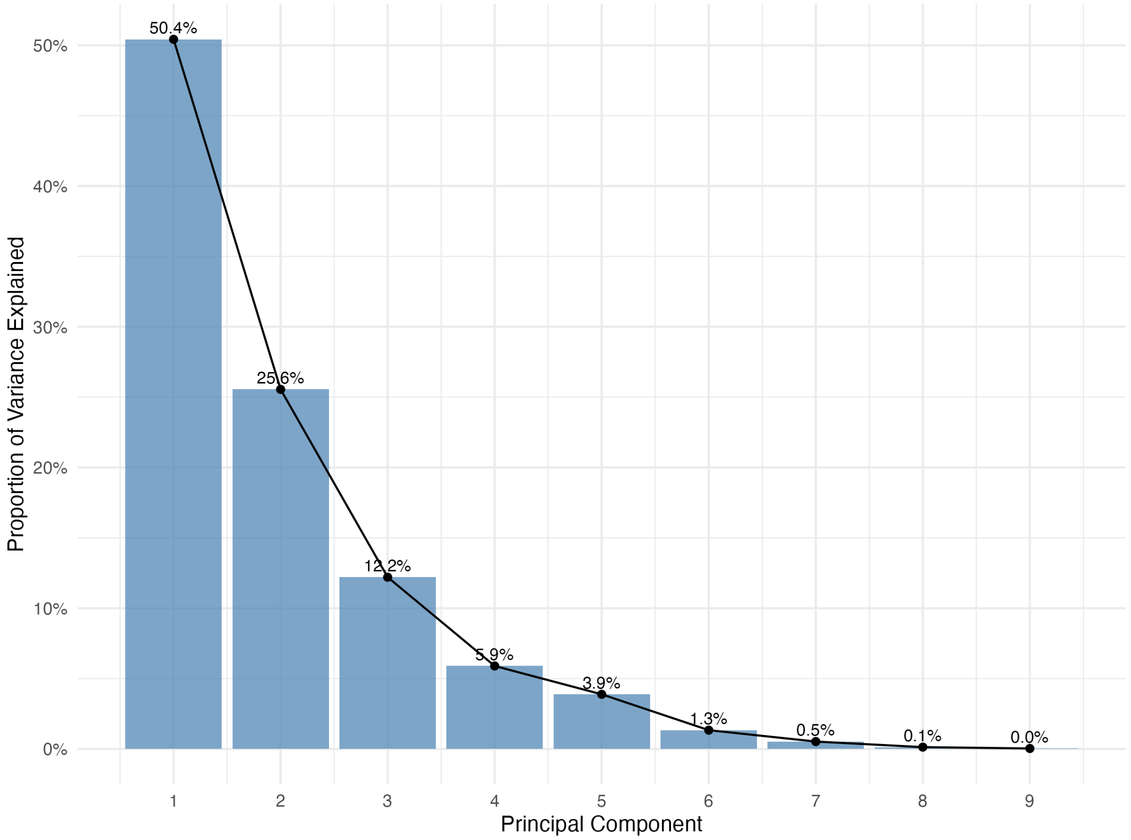


Figure 1. Scree Plot showing Variance Explained by Principal Components.

2.2 Clustering Performance

The unsupervised clustering algorithm (K-Means, K=2) achieved a high degree of accuracy in distinguishing player gender solely based on match statistics.

Table 1. Contingency Table of Clustering Results by Gender

	Cluster 1 (ATP-Leaning)	Cluster 2 (WTA-Leaning)
ATP	298 (Correct)	96 (Mis-clustered)
WTA	16 (Mis-clustered)	374 (Correct)

- **Accuracy:** The model correctly classified **672 out of 784** players, achieving an overall accuracy of **85.7%.**
- **Cluster 1 (ATP-Leaning):** Predominantly ATP players (298 vs 16 WTA).
- **Cluster 2 (WTA-Leaning):** Predominantly WTA players (374 vs 96 ATP).

This confirms that “men’s tennis” and “women’s tennis” have distinct statistical signatures, primarily driven by serve dominance.

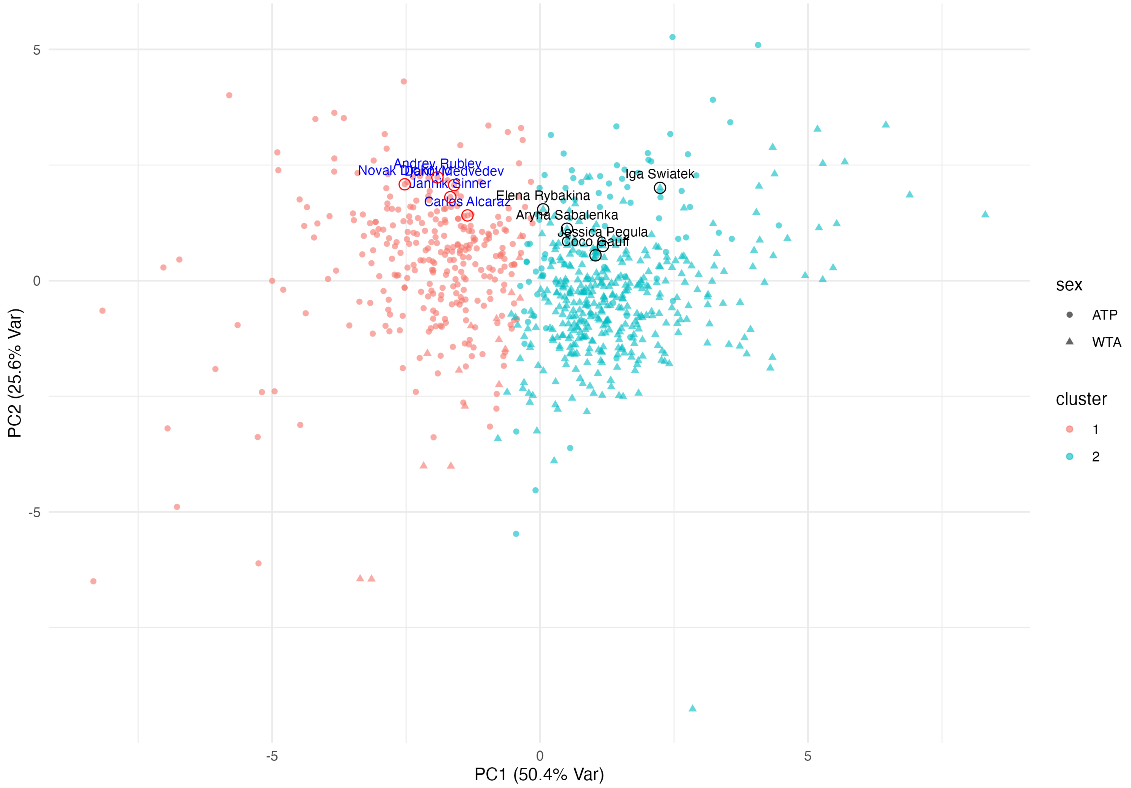


Figure 2. K-Means Clustering in PCA Space (PC1 vs PC2).

3 Research Question B: What kind of player wins?

Method: Correlation and Regression Analysis of Win Rate against Principal Components.

3.1 Interpretation of PCs via Loadings Plot

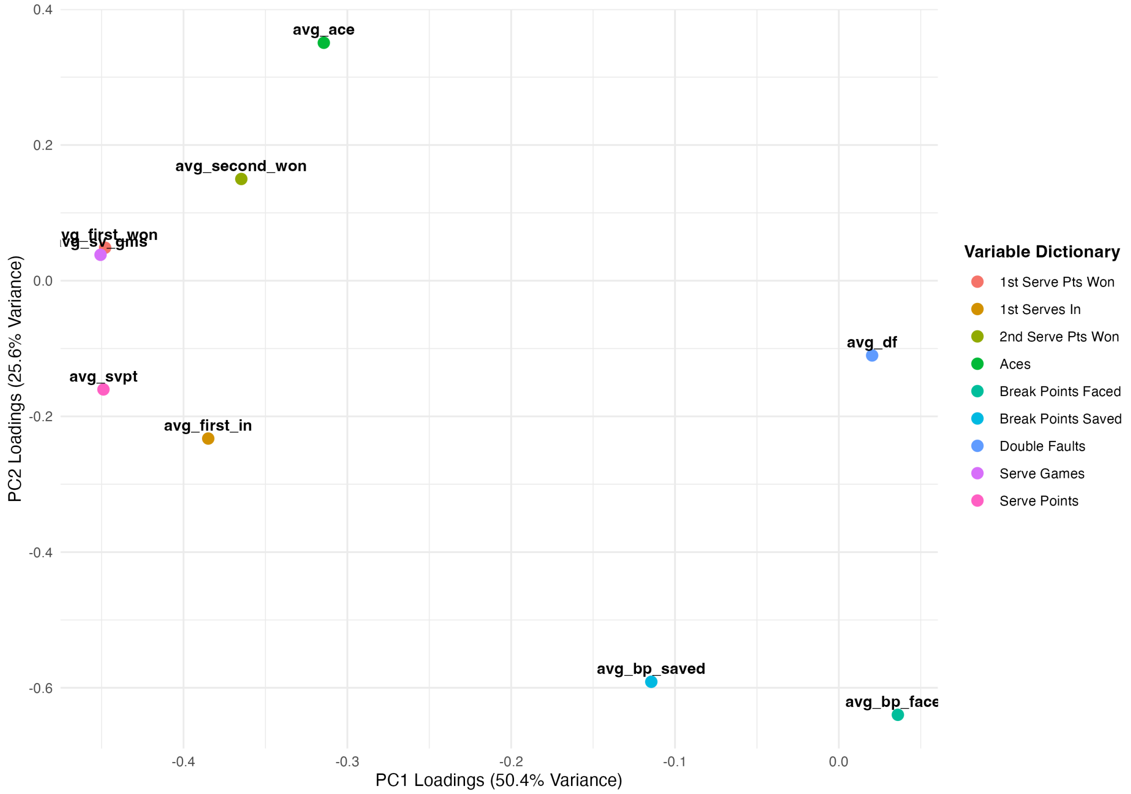


Figure 3. PCA Variable Loadings Vectors.

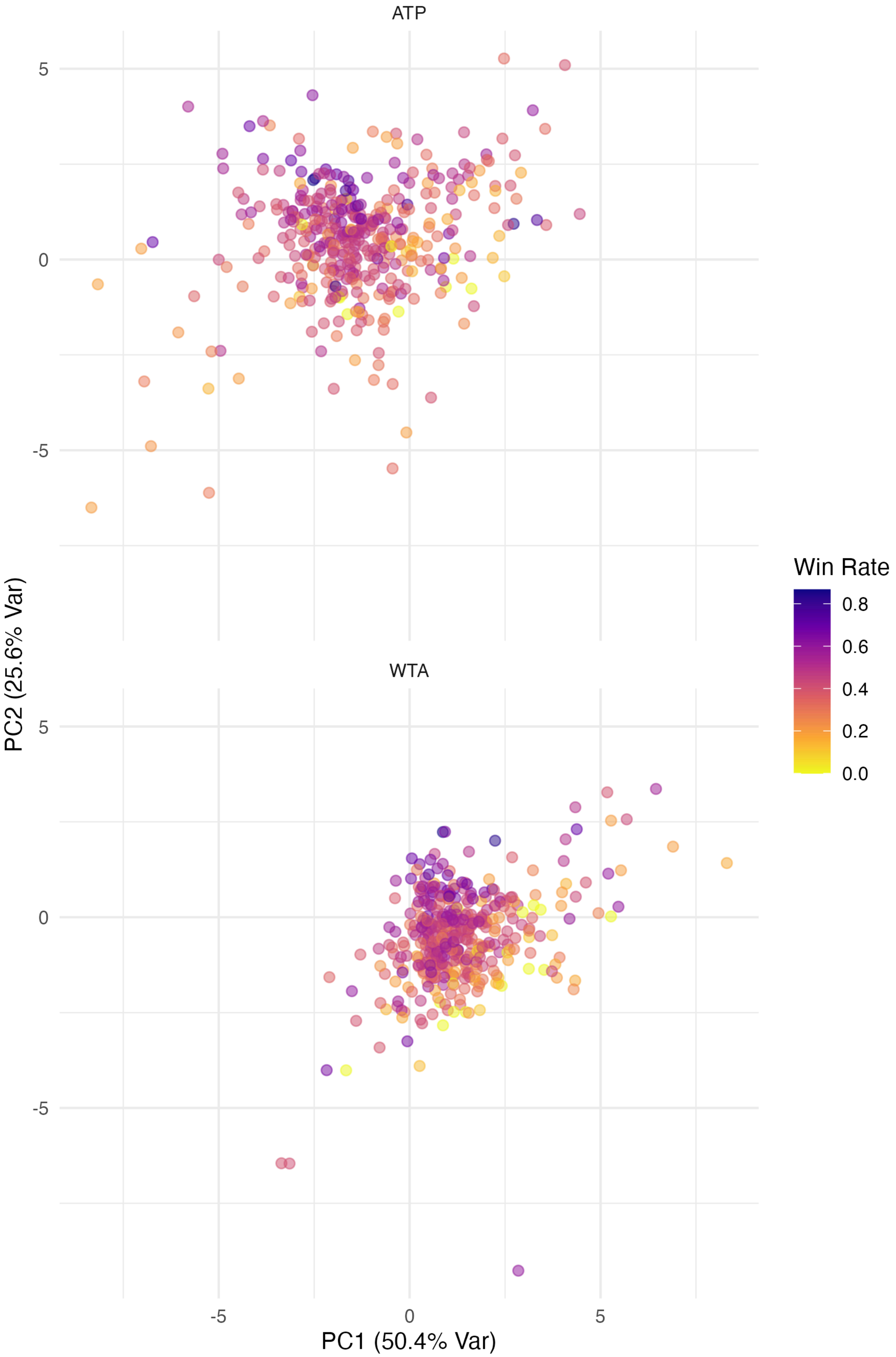
- **PC1 (Serve Quality/Volume):** In the loadings plot, variables like Aces (`avg_ace`), Serve Games (`avg_sv_gms`), and First Serve Points Won (`avg_first_won`) cluster heavily on the **negative** side (left). This means a *lower* (more negative) PC1 score represents stronger, more dominant serving stats.
- **PC2 (Pressure/Resilience):** The variables Break Points Faced (`avg_bp_faced`) and Saved (`avg_bp_saved`) have strong **negative** loadings (bottom of the plot). Consequently, a *higher* (positive) PC2 score indicates a player who faces *fewer* break points—a hallmark of dominance and control.

3.2 Correlation with Winning

- **PC1 vs Win Rate:** $r = -0.1280$ $(p < 0.001)$.
  - Since strong serving corresponds to negative PC1, the negative correlation confirms that better servers tend to win more, but the relationship is relatively weak.
- **PC2 vs Win Rate:** $r = 0.2794$ $(p < 1e - 15)$.
  - The strong positive correlation means that players with high PC2 scores (those who face fewer break points) are significantly more likely to win.
- **Key Insight:** While serve power (PC1) matters, the ability to control the match and avoid break points (PC2) is the stronger statistical driver of victory.

3.3 Visual Evidence

The win rate heatmap in PCA space shows that the “winningest” players tend to cluster in regions with high PC2 scores (top of the chart). The top 5 ranked players (shown in Figure 2) also align with these high-win-rate regions.


ATP
WTA
5
0
-5
5
0
-5
-5
0
5
PC1 (50.4% Var)
PC2 (25.6% Var)
Win Rate
0.8
0.6
0.4
0.2
0.0

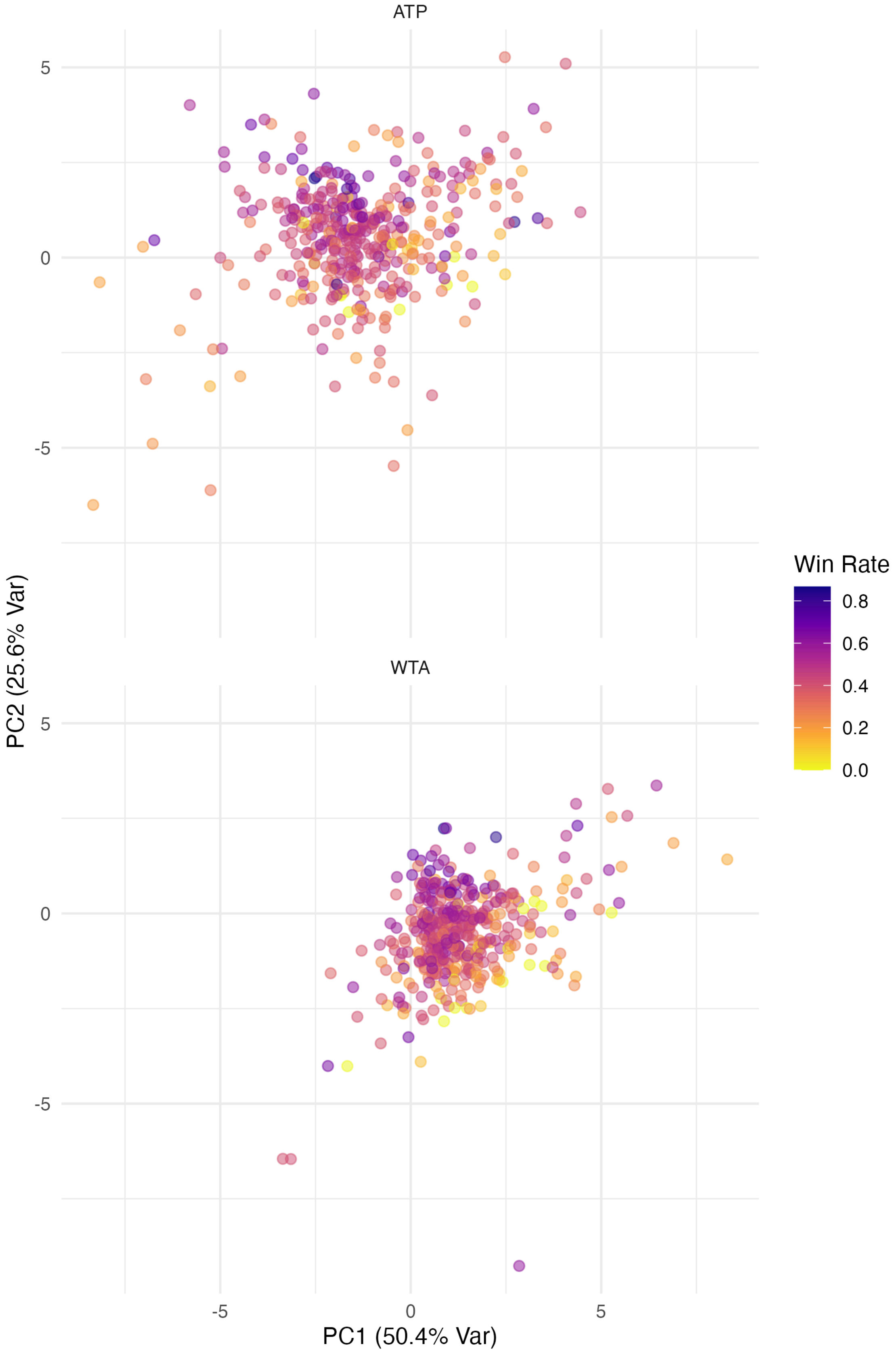


Figure 4. Player Win Rates in PCA Space by Tour.

4 Research Question C: Mis-clustered Player Analysis

Method: Analysis of the 112 players (14%) who were "misclassified" by the clustering algorithm (e.g., ATP players in the "WTA-like" cluster).

4.1 Why are they mis-clustered?

4.1.1 WTA in ATP Cluster (n=16)

These players play a "power game" more typical of the ATP tour.

- **Aces:** They hit significantly more Aces (Avg 3.36) than typical WTA players (Avg 2.03).
- **Serve Dominance:** They win significantly more points on their first serve (33.5%) compared to the WTA average (26.4%), mirroring ATP efficacy.

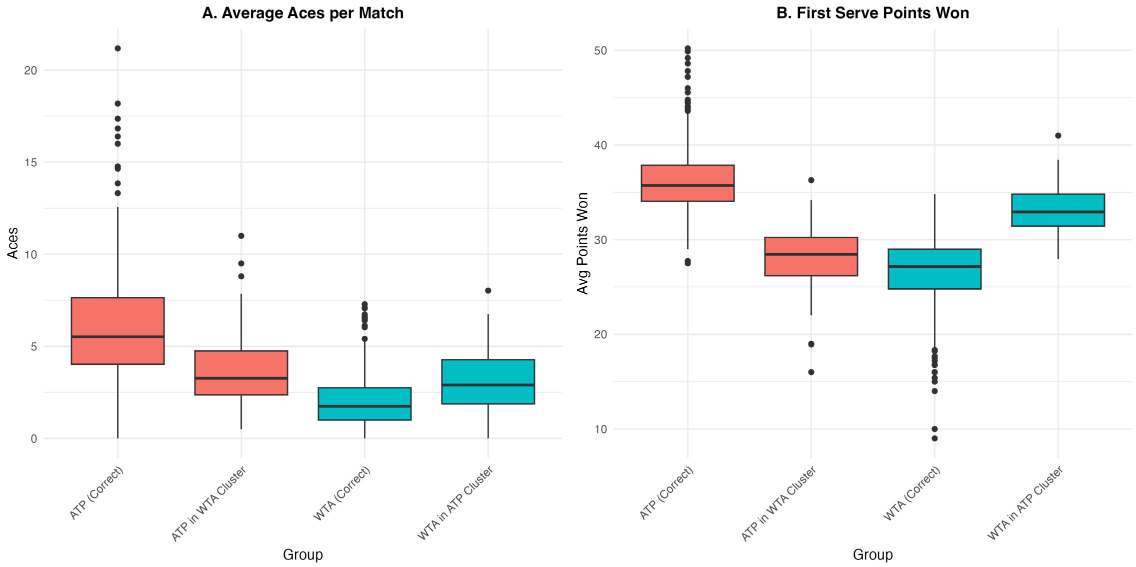


Figure 5. Comparison of Service Metrics for Correct vs. Mis-clustered Players. (A) Distribution of Aces per Match. (B) Distribution of First Serve Points Won.

4.1.2 ATP in WTA Cluster (n=96)

These players have statistical profiles closer to the WTA average.

- **Aces:** They hit significantly fewer Aces (Avg 3.64) than typical ATP players (Avg 6.08).
- **Serve Effectiveness:** Their points won on first serve (27.9) drops drastically from the ATP standard (36.4) to near-WTA levels (26.4), indicating a lack of the "free points" usually associated with men's tennis.
- **Performance Impact:** These "mis-clustered" ATP players have a significantly **lower win rate** (35%) compared to standard ATP players (42%).

4.1.3 Ranking scores

As shown in Figure 6, ATP players assigned to the "WTA-style" cluster have substantially lower ranking points than correctly clustered ATP players. This pattern is consistent with their match statistics: weaker serve performance and fewer aces make these ATP players resemble WTA profiles, and this "WTA-style" play is penalized heavily in the men's game. On the women's side, the pattern looks different. WTA players assigned to the "ATP-style" cluster do not suffer a corresponding drop; in fact, their ranking points are comparable to correctly clustered WTA players. This indicates that adopting an "ATP-style" power-serve profile is not a disadvantageor WTA players, whereas adopting a "WTA-style" lower-serve-impact profile is a clear disadvantage for ATP players.

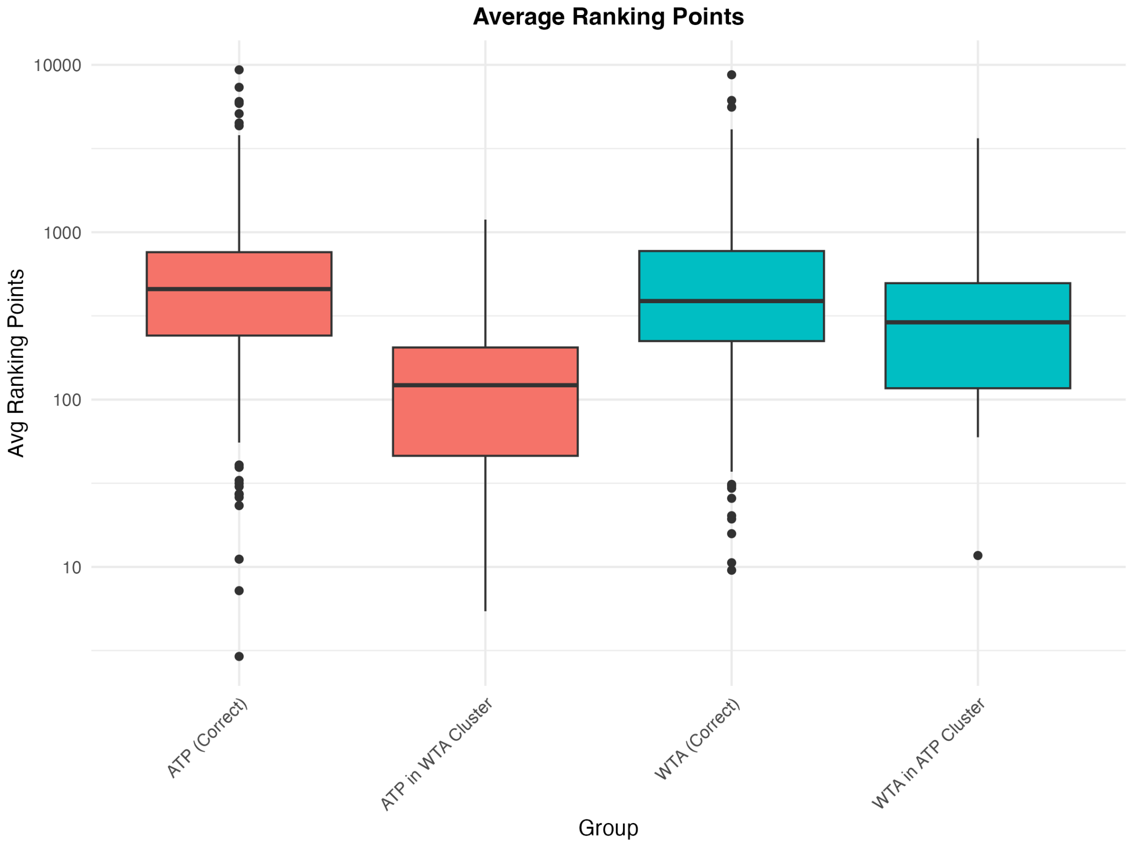


Figure 6. Distribution of Average Ranking Points by Cluster Group. Note that ATP players in the WTA cluster (“ATP in WTA Cluster”) show a trend towards lower ranking points compared to the correct ATP group.

4.2 Visual Summary of Mis-clusters

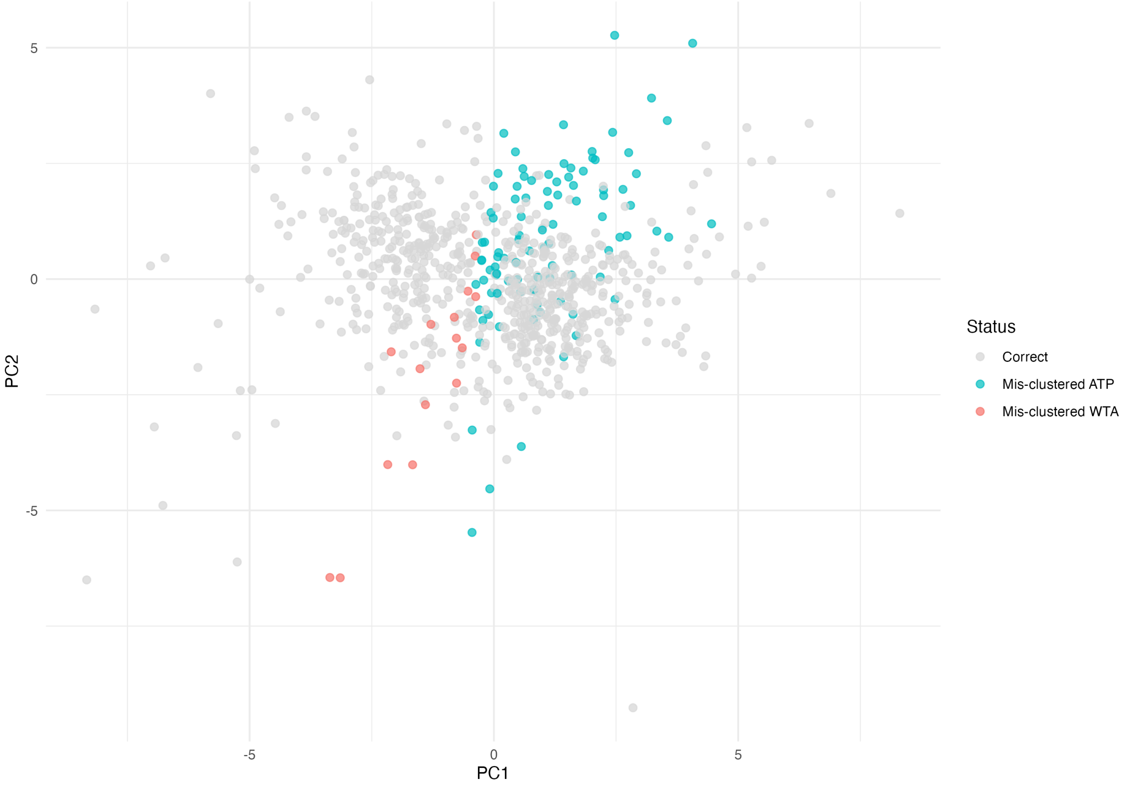


Figure 7. Visualizing Mis-clustered Players in PCA Space.

4.3 Summary

The mis-cluster analysis reveals that these players are not random errors but “functional outliers.” ATP players in the WTA cluster lack the typical serve weapons of men’s tennis, which correlates with lower rankings and win rates. However, the pattern is not evident in women players. This difference in ranking point impact highlights a key asymmetry: while the “ATP style” (high serve dominance) is not a harm in the women’s game, playing a “WTA style” (lower serve impact) in the men’s game is often a competitive disadvantage.

5 Conclusion

In this study, we successfully employed an unsupervised learning approach (PCA + K-Means Clustering) to distinguish between ATP and WTA players with **85.7% accuracy** solely based on in-match statistics. This high accuracy validates that the two tours have distinct statistical structures, primarily defined by serve power and volume (PC1). However, the analysis also highlighted that success in both tours is unified by a common factor: the ability to minimize break point opportunities (PC2). The players who defied the clustering algorithm provided the most interesting insights—demonstrating that “gendered” styles of play are fluid, and adopting the dominant traits of the opposite tour (like high serve power for women) can be a winning strategy.

6 Future Direction and Limitations

1. **Variable Selection:** The current analysis relied on a limited set of aggregate match statistics (Aces, DFs, Serve Points, etc.). Future work could incorporate more granular data such as rally length, return points won, and unforced error counts to create a more comprehensive player profile.
2. **Age Effect:** This study did not control for player age. Age could be a confounding factor, particularly for the “mis-clustered” ATP players (who might be younger/developing or older/declining). Incorporating age as a variable could refine the clustering and performance analysis.