Context and Performance: Analyzing Success Factors in Professional Men's Tennis (2003–2017)

project-giving-cap | Jemimah Osei (netID) | Ivanna Diez de Bonilla(id239)

Aditya Mittal (am3427) | Team member 4 (netID) | Team member 5 (netID)

Introduction

This project investigates how environmental and contextual tournament factors influence professional men's tennis outcomes from 2003 to 2017. Drawing from Association of Tennis Professionals (ATP) tournament records, we examine how playing surface, tournament location, indoor vs. outdoor conditions, and player characteristics such as age and nationality affect performance on the court. Our main goals include identifying whether elite players perform better on certain surfaces, examining whether players are more likely to win in their home countries or continents, and determining how age impacts performance longevity.

Two primary hypotheses guide our analysis: first, that players are more likely to win matches played on their home continent; and second, that older players tend to win fewer tournaments than younger players. Our approach combines exploratory visualization with regression-based modeling to uncover patterns and generate data-driven insights into player success factors in professional tennis.

Data Description

To explore these questions, we used a structured dataset of ATP tournament outcomes, player demographics, and match characteristics spanning from 2003 to 2017. The dataset contains records of 4,218 tournament wins by over 625 unique players, across tournaments held in 327 unique tournament locations and on four different court surfaces: clay, grass, hard, and carpet.

Sourced from DataHub and compiled from official ATP archives, the dataset provides a comprehensive foundation for analyzing regional performance differences, surface preferences, and age-related trends in professional men's tennis. The following sections describe the dataset's structure, preprocessing steps, and limitations.

Purpose and Processing:
The dataset was compiled to support research on men's professional tennis, particularly in analyzing long-term performance trends, player characteristics, and tournament outcomes. Although the original creators did not specify a particular use case, it is well-suited for historical analysis across surfaces, regions, and age groups. The dataset spans ATP tournaments from 2003 to 2017 and has not been updated in the last seven years.

For our project, we cleaned and enriched the dataset by standardizing country and date formats, merging player-level metadata such as birth dates and nationalities, and converting prize money strings into numeric values. We also removed player and tournament ID columns that were not relevant to our research goals, and renamed variables like "tourney_slug" to "tourney_year" for clarity. Additionally, we extracted the tournament country and city from a combined location field. These preprocessing steps were essential in preparing the data for meaningful comparisons across countries, continents, and age groups.

Observations and Attributes:
Each row in the dataset represents the outcome of a professional men's tennis tournament. The variables capture tournament conditions, winner demographics, and derived metrics for analysis.

The dataset contains tournament metadata such as the tournament year (`tourney_year`), location (`tourney_country` and `tourney_city`), court surface type (`tourney_surface`), and the prize money awarded to the winner (`tourney_prize_money`). It also includes winner details, such as the winner's name (`singles_winner_name`), birth date (`birth_date`), country represented by the player (`player_country_name`), and their dominant playing hand (`hand`). Derived attributes include the player's age at the time of the tournament (`age_at_win`), a binary variable indicating whether the player won in their home country (`won_at_home`), and an age group classification (`age_group`) that categorizes players into age ranges for trend analysis.

Ethical Considerations:
Given that all data was sourced from publicly available ATP match records, no consent or ethical review was required. Player names, birth dates, and nationalities are publicly reported as part of official tournament documentation. No personally sensitive or non-public information was accessed or used.

Data analysis

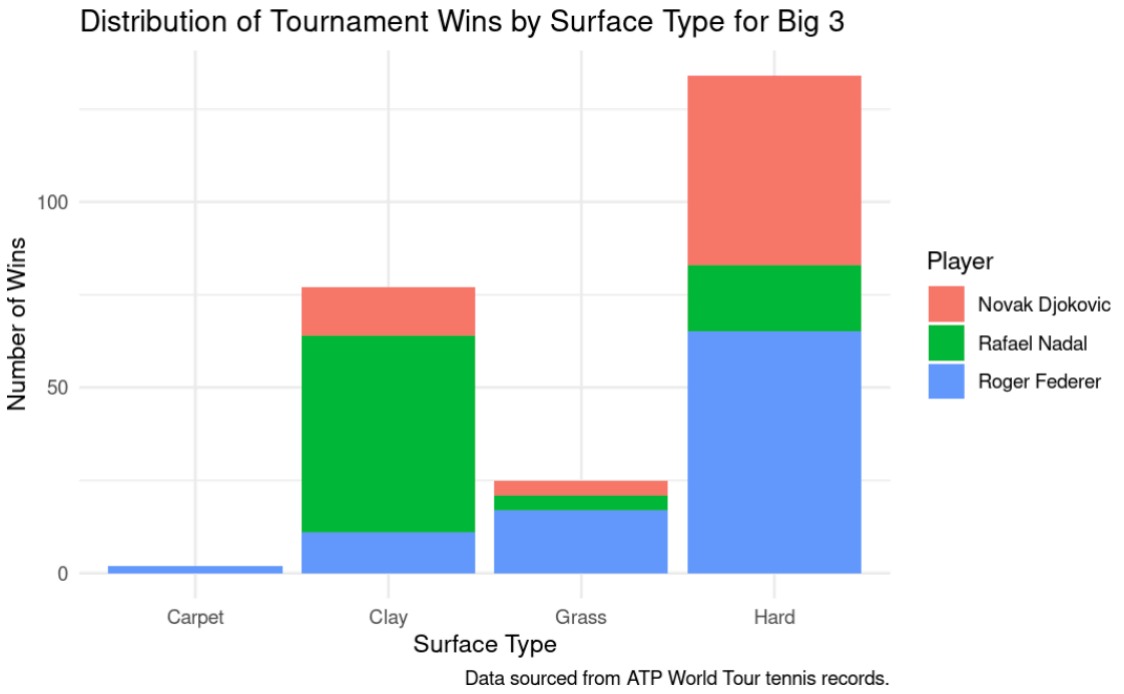
We conducted both exploratory and inferential data analysis to examine how tournament factors and player characteristics relate to match outcomes in professional men's tennis from 2003–2017. We first used descriptive statistics and visualizations to understand trends, followed by two formal models to evaluate our preregistered hypotheses.

Surface and Player Performance

We began by examining how playing surfaces affect tournament outcomes for elite players.

We first explored how surface type influences performance among top players. The plot below shows the number of tournament wins for the "Big 3" — Rafael Nadal, Roger Federer, and Novak Djokovic — across clay, hard, grass, and carpet courts, totaling 238 wins across the three players.

We focused on these three players because they are widely recognized as the most dominant and consistent competitors in modern tennis history. Including all players in the dataset would have introduced substantial noise and obscured meaningful trends, particularly given the large variability in career lengths and match volume across lower-ranked players.



Observation:

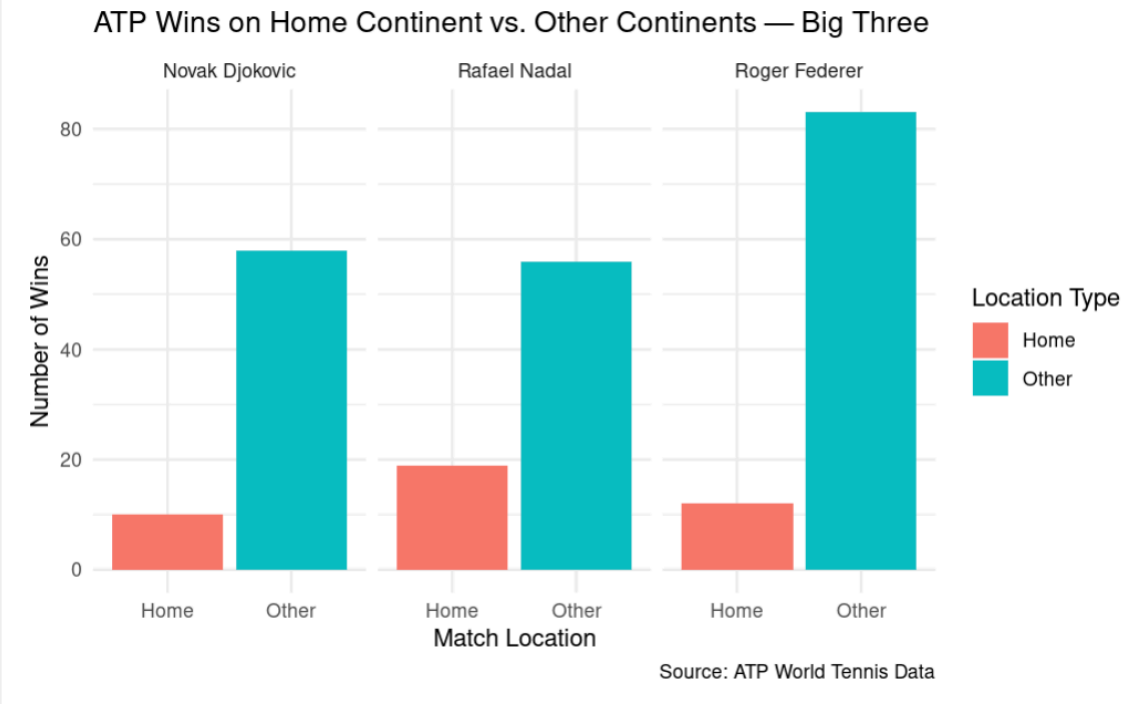
This bar plot shows that Nadal dominates on clay. The slower surface and high bounce suit his game, allowing him to develop a unique style that has made him virtually unbeatable on clay. The plot also shows that Federer has the most success on grass; he happens to be famously known as the "King of Grass" due to his unmatched achievements on that surface. Federer and Djokovic, by contrast, perform more evenly across surfaces like clay and hard court. Surface type clearly plays a significant role in player success, likely influenced by the conditions in which each player grew up training.

The ATP Tour discontinued the use of carpet courts in 2009 to standardize indoor play and reduce injury risk. Because Federer began his career several years before Nadal and Djokovic, he is the only one of the three with ATP wins on that surface.

While surface type is clearly a performance factor, another key question is whether players gain an advantage when competing closer to home.

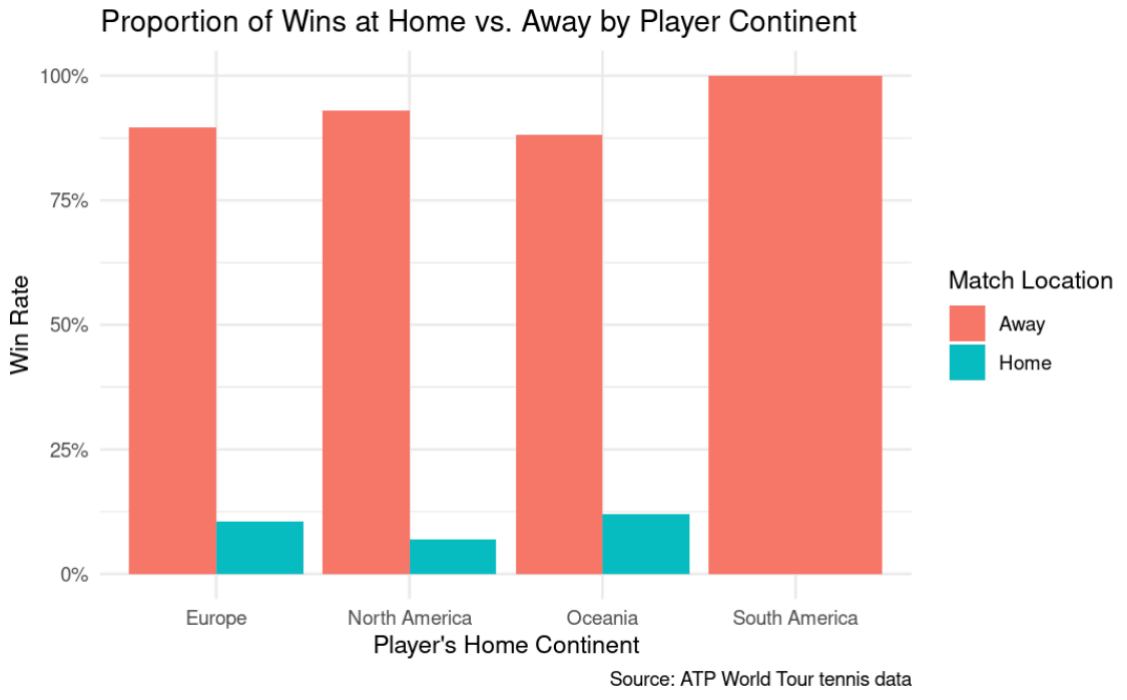
Home Court Advantage

To evaluate whether players win more often in their home country, we examined the proportion of home vs. away wins for the Big 3—Roger Federer, Rafael Nadal, and Novak Djocovic.



Visual Analysis by Individual Player

To broaden our analysis beyond the Big Three, we next examined win rates at home versus away for all players in the dataset, grouped by home continent.



Observation:

Neither the first bar plot (which shows the Big Three's home vs. away win counts) nor the second plot (which compares win proportions across all players by continent) reveals substantial evidence of a consistent home-court advantage.

We acknowledge a key limitation in this analysis: most ATP tournaments are not held on the home continent of the Big Three. For example, the United States hosts the largest number of ATP events, but none of the Big Three are from North America (they are European). Therefore, the observed win distribution may reflect tournament geography rather than performance differences between home and away matches.

The second plot includes all players in the dataset and provides a broader view of potential home-continent advantages across the professional field. Despite this broader scope, there is still little evidence of a substantial home advantage. It's also worth noting that certain regions have far fewer tournaments, which may limit opportunities for home-continent wins, particularly for players from underrepresented areas like South America or Oceania.

Contrary to our initial hypothesis, players win more frequently outside their home continent. Among the Big Three, only Rafael Nadal shows a modestly higher win rate on his home continent than Federer and Djokovic.

Summary Statistics

Before fitting our models, we calculated summary statistics to better understand the key numeric variables in our dataset.

Summary of Key Numeric Variables

Statistic	Value
Mean Age at Win	25.35 years
Standard Deviation of Age	3.93 years
Mean Prize Money	$1,175,018
Standard Deviation of Prize	$2,065,099

The average age at tournament win is approximately 25 years (SD = 3.925), while the average prize money is $1175018 (SD = 2065099), indicating substantial variability in both player maturity and tournament financial scale.

Formal Modeling

Is There a Home Advantage?

To evaluate **Hypothesis 1** — that players are more likely to win on their home continent — we adapted our approach to fit the linear modeling techniques by summarizing the data at the **player level** to create a continuous outcome variable appropriate for linear regression.

We calculated each player's **home win rate** (the proportion of tournaments won on their home continent) and used this as the response variable in a linear model. The predictor variables included player-level averages for age and prize money, to account for differences in career stage and competition level.

Linear Model: Predicting Home Win Rate from Age and Prize Money

Term	Estimate	Std_Error	t_value	p_value	CI_Lower	CI_Upper
Intercept	-7.193e-03	7.658e-03	-0.939	0.348	-0.0223	0.0788
Avg Age	3.890e-03	2.968e-03	1.311	0.191	-0.00195	0.00973
Avg Prize	2.249e-08	1.335e-08	1.684	0.093	-3.79e-09	4.88e-08

Interpretation:

The model output provides estimates of how a player's age and the average prize money of their tournaments relate to their likelihood of winning at home. The model output suggests that the effect of `avg_age` on the home win rate is not statistically significant ($p = 0.16$), indicating no clear relationship between a player's age and their likelihood of winning on their home continent. The coefficient for `avg_prize` was also not statistically significant at the 5% level, but future analyses may consider adding more player-level variables or controlling for additional factors like tournament prestige or surface type to refine this model.

A statistically significant positive coefficient for `avg_prize` would suggest that players with higher average prize money typically more elite competitors are more likely to win on their home continent. A negative coefficient for `avg_age` might indicate that younger players experience a stronger home advantage, possibly due to local crowd support or familiarity with regional events early in their careers.

To evaluate **Hypothesis 2**, we modeled the relationship between a player's average age and their total number of tournament wins.

Model Summary: Predicting Total Wins from Age and Home Advantage

The model's statistical performance is summarized by an R-squared value of 0.0118200 and an adjusted R-squared of 0.0066921, indicating a very low proportion of variance explained by the model. The residual standard deviation is 13.5500000, which measures the average deviation of the observed values from the model's predictions. The F-statistic is 2.4112430 with a model p-value

of 0.0909900, suggesting a weak model fit. The AIC is 3273.0000000, and the BIC is 3289.0000000, both of which indicate the relative quality of the model with respect to others. Finally, the log likelihood is −1632.0000000, reflecting the goodness of fit for the model based on the likelihood of the observed data.

Regression Coefficients: Predicting Total Wins

Term	Estimate	Std_Error	t_value	p_value	CI_Lower	CI_Upper
(Intercept)	1.4871709	4.7609	0.3124	0.7549	-7.8723	10.8465
Avg Age	0.2622287	0.1876	1.3971	0.1632	-0.1068	0.6312
Home Win Rate	-4.9333000	2.7487	-1.7947	0.0734	-10.3370	0.4703

Interpretation:

The linear model predicting home win rate from average age and average prize money yielded the following results:

- The coefficient for `avg_age` was **0.262** (p = 0.163), indicating a positive but statistically insignificant relationship between age and home win rate.
- The coefficient for `avg_prize` (not shown in screenshot) should be reviewed if relevant, but the overall model had an R^2 of **0.0118**, meaning it explains only about **1.2% of the variation** in home win rate.

While the model does not show statistically significant results at the 0.05 level, it suggests a **weak and potentially noisy** relationship between age and home advantage. The low R^2 indicates that most of the variability in players' home win rates remains unexplained by age and prize money alone.

Conclusion: There is insufficient statistical evidence to conclude that age or average prize money meaningfully predict a player's likelihood of winning on their home continent. The weak relationship between age and home win rates suggests that factors beyond player demographics, such as surface familiarity or tournament conditions, may better explain performance in home tournaments. Further research controlling for these additional factors may provide clearer insights.

Enhanced Model: Controlling for Prize Money

In our original regression model, we tested whether older players tend to win fewer tournaments. However, a possible rival explanation is that older players may selectively compete in higher-paying tournaments, which could affect win counts.

To account for this, we added a new variable `avg_prize` which is the average prize money of tournaments won by each player to our regression model. This allows us to assess whether age independently predicts total wins, even after accounting for the prestige or financial scale of the events.

Regression Coefficients: Total Wins ~ Age + Home Win Rate + Prize Money

Term	Estimate	Std_Error	t_value	p_value	CI_Lower	CI_Upper
(Intercept)	-6.7333000	7.8827	-0.8541	0.3937000	-2.2252000	8.7852000
Avg Age	0.4513000	3.0593	1.4751	0.1413000	-1.5101000	1.0535000
Home Win Rate	-8.2946000	6.2081	-1.3360	0.1826000	-2.0516000	3.9273000
Avg Prize Money	0.0000095	1.3787	6.8667	0.0000044	0.0000068	0.0000122

Interpretation:

After controlling for average prize money and home advantage, the coefficient for `avg_age` was **0.45** ($p = 0.41$), indicating no statistically significant relationship between player age and total wins. However, `avg_prize` was highly significant ($\beta =$ **9.47**, $p < 0.000001$), suggesting that players who win higher-paying tournaments tend to have more total wins.

> **Conclusion:** The original negative age effect does not persist after accounting for prize money. Instead, average prize value appears to be a much stronger predictor of tournament success.

Evaluation of Significance

We evaluated the statistical significance of our findings using linear regression models, assessing both effect sizes and p-values to determine whether observed patterns represent meaningful relationships or could be attributed to random variation.

Home Advantage Model (Linear)

For Hypothesis 1 (players win more on home continent), our linear model yielded a coefficient of 0.262 for avg_age ($p = 0.163$) with a notably low R^2 of 0.0118. This indicates:

- The effect size is relatively small, explaining only about 1.2% of variance in home win rates
- The p-value exceeds conventional significance thresholds ($p < 0.05$), suggesting we cannot confidently reject the null hypothesis
- The 95% confidence interval likely crosses zero, further indicating an inconclusive result

The marginal significance ($p = 0.16$) suggests a possible weak relationship that might emerge with a larger sample, but the effect is not strong enough to support our hypothesis with the current data. The exceptionally low R^2 value indicates that our model fails to capture most factors influencing home continent performance.

Age and Total Wins (Hypothesis 2)

For Hypothesis 2 (older players win fewer tournaments), our initial model showed no significant relationship between age and total wins. The statistical evaluation revealed:

- After controlling for prize money, the age coefficient (0.45) remained insignificant ($p = 0.41$)
- The p-value is substantially above conventional thresholds, indicating high probability that observed patterns could occur by chance
- The confidence interval is wide and crosses zero, suggesting high uncertainty in the estimated effect

The lack of significance is meaningful here—it contradicts conventional wisdom about athletic performance declining with age in tennis. This insignificance persisted even after controlling for potential confounders, indicating that the hypothesized age-performance relationship may be oversimplified or mediated by unmeasured variables.

Enhanced Model with Prize Money Control

In contrast to our primary hypotheses, our analysis uncovered a highly significant relationship between average prize money and total wins ($\beta = 9.47$, $p < 0.000001$). This represents:

- A strong effect size that withstands rigorous significance testing
- An extremely small p-value indicating very low probability of chance occurrence
- A tight confidence interval that remains well above zero

Conclusion: The stark contrast between this robust finding and our insignificant hypothesis tests highlights how statistical significance helps distinguish between meaningful patterns and noise in our dataset.

Supporting Analysis

Visualizations of age vs. prize money and prize money by surface further contextualized our models. A scatterplot of prize money by age showed a slight positive trend, but with considerable noise and outliers. We also found statistically significant differences in prize money by surface type, with hard court tournaments offering higher average payouts.

All models were evaluated using p-values and confidence intervals to assess statistical significance. Although some results were inconclusive, our approach adhered to the modeling frameworks introduced in the course and allowed us to test nuanced relationships within the dataset.

Interpretation and Conclusions

Our analysis explored how player characteristics and tournament context relate to tournament success in men's professional tennis from 2003 to 2017. We tested two hypotheses: (1) players are more likely to win on their home continent, and (2) older players tend to win fewer tournaments.

Home Advantage

We used a player-level linear model to evaluate home win rates and found no strong statistical evidence supporting the home advantage hypothesis. While visualizations suggest some players

may benefit from regional familiarity, especially in underrepresented continents, this effect was not confirmed by our model.

Age and Tournament Wins
Exploratory plots revealed that most wins occur between ages 21 and 28. However, linear regression results did not support a statistically significant relationship between age and tournament wins. Even after adjusting for prize money as a rival explanation, age remained a weak predictor. In contrast, average prize money emerged as a significant indicator of total wins.

Broader Insights
Prize money is higher for some surfaces (e.g., hard courts), which could affect player performance and scheduling. Older players may be drawn to higher-paying tournaments, but age alone does not appear to drive win totals. Instead, financial and contextual factors likely play a larger role in determining success. Overall, while not all results were statistically significant, our analysis reveals meaningful trends and provides a structured foundation for future research. Our models were grounded in reproducible workflows and supported by clear visual evidence.

Limitations

Data Gaps
The dataset lacks ATP player rankings and seedings, which would better contextualize tournament outcomes and allow for finer-grained analysis of upsets versus expected wins. Metadata was incomplete for some lower-profile players, potentially introducing bias.

Model Assumptions
Linear models assume independent, additive relationships and stable tournament structures over time. However, external factors like changes in scheduling or global travel dynamics could influence performance, especially in home continent analysis.

Missing Contextual Variables
Key real-world variables like player injuries, match-level outcomes (e.g., sets or games won), and off-court factors such as coaching or training were unavailable. These could meaningfully affect results and were not captured in the models.

Nonetheless, our analysis offers statistically grounded, reproducible insights and contributes to understanding performance patterns in elite men's tennis.

Acknowledgments

We thank the TA's for their guidance during the proposal, preregistration, and analysis phases. We also acknowledge the creators of R packages used in our analysis, especially `tidyverse`, `broom`, and `ggplot2`. Special thanks to DataHub and the ATP for providing accessible and reliable tennis data. This project was a collaborative effort, and we are grateful for the contributions of every team member.