Introduction

Tennis has long been regarded as an elite sport—not just because of its exclusivity at certain tournaments, but also because of its unique and sometimes perplexing scoring system. From "love" meaning zero to the seemingly arbitrary jump from 15 to 30 to 40, and the decisive weight of "game, set, match," understanding tennis can feel as complex as playing it. Beyond the scoring, the sport generates vast amounts of data covering player performances, match conditions, betting trends, and career trajectories. However, much of this data remains inaccessible or difficult to interpret for casual fans and analysts alike.

This project aims to demystify tennis through interactive visualizations that make the sport more approachable and insightful. By integrating match data, career progressions, betting odds, rivalries, and geographical influences, we seek to provide a dynamic and engaging way to explore professional tennis from 2000 to 2024. Instead of just presenting numbers and statistics, our platform will offer an intuitive and interactive experience that helps users truly understand the patterns, rivalries, and trends that shape the game.

Dataset

The first dataset, named **"tennis_atp"**, is available on GitHub through Jeff Sackmann's repository. It contains extensive historical data from men's professional tennis matches (ATP Tour), beginning with the start of the Open Era in 1968 and continuing to the present day. However, our analysis will specifically focus on the period from 2000 to 2024. The dataset offers comprehensive information on tournaments (including tournament ID, name, level, surface type, draw size, and event dates) and detailed player profiles (names, nationality, age, height, dominant hand, ATP rankings, and ranking points at the time of each match). Additionally, it provides extensive match-level statistics, such as match scores, the number of sets played, match duration, and detailed performance metrics like aces, double faults, first and second serve percentages, break points saved or faced, and service games statistics.

The second dataset, named **" 🎾 Ultimate Tennis Matches Dataset,"** is accessible through Kaggle. It includes comprehensive match data from both men's (ATP) and women's (WTA) professional tennis circuits, covering the period from 2000 to 2024. This dataset provides detailed tournament information, including the location, tournament names, precise match dates, competition categories (such as Grand Slams, Masters, ATP 500/250, and their WTA equivalents), court characteristics (surface type, indoor/outdoor), player details (names, rankings, and points at match time), and match outcomes with set-by-set scores. An additional valuable feature of this dataset is the detailed sports betting information, including odds provided by various bookmakers as well as maximum and average odds for each player and match, facilitating analysis of the relationships between tennis match outcomes and betting market dynamics.

Problematic

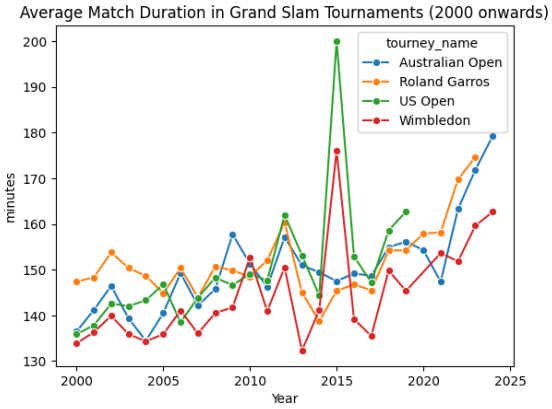
Despite the availability of rich datasets, professional tennis remains challenging to analyze without appropriate visualizations. This project aims to make the sport accessible to both casual enthusiasts and intermediate tennis fans, providing them with clear, interactive visual tools to explore various factors influencing player performance and match outcomes from 2000 to 2024. Our visualizations will go beyond general career statistics to offer deeper insights into betting market accuracy, court surface effects, geographical influences, and gender-based differences.

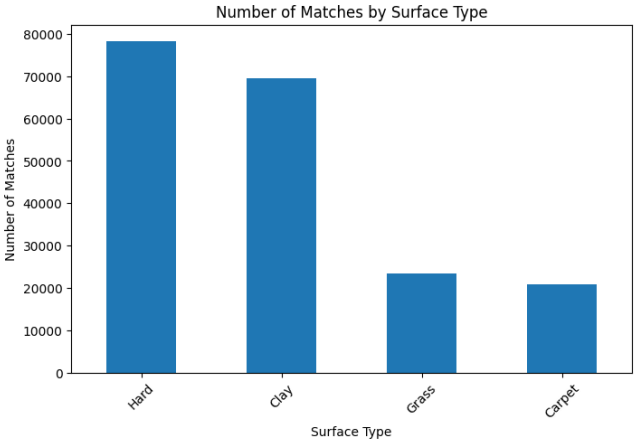
More specifically, our project aims to provide visualizations for exploring the following questions:

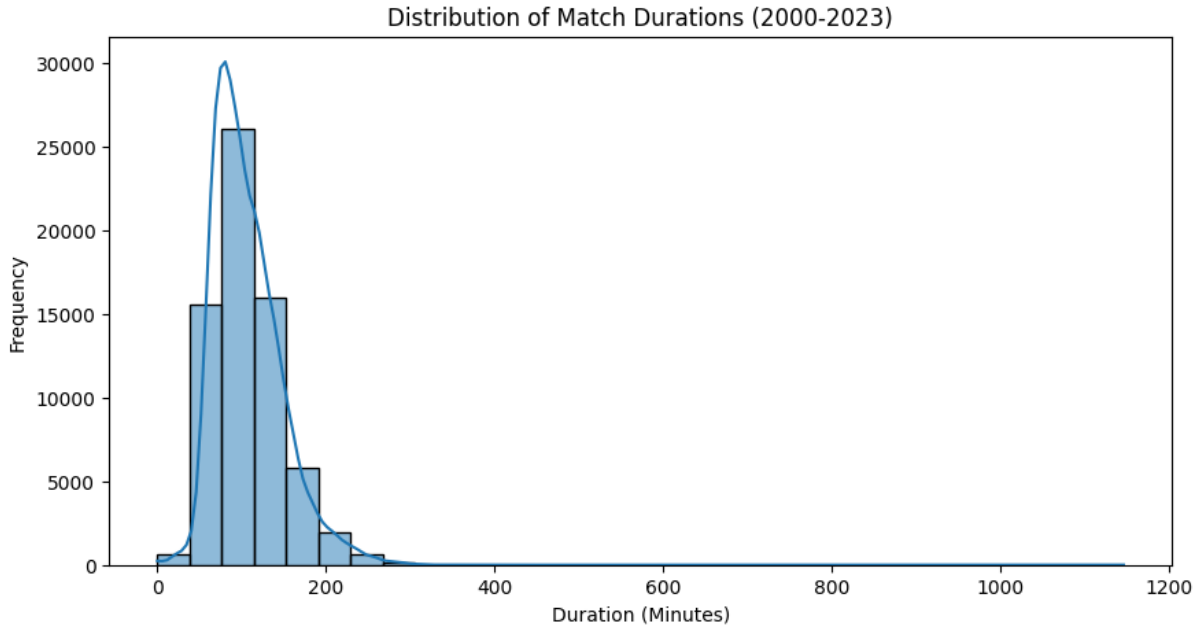
- **Betting accuracy**: How reliable are betting odds in predicting match outcomes?
- **Surface comparison**: How does player performance vary across different tennis court surfaces?
- **Player trajectories**: What characterizes the typical career progression of elite tennis players?
- **Geographical impact**: Do climate conditions and host countries significantly affect players' performances?
- **Rivalries**: Which rivalries have defined tennis in the past two decades, and how do win-loss records shape these rivalries?
- **Nationality factors**: Does a player's country of origin influence performance across various court surfaces?
- **Gender differences**: Are there notable differences between men's and women's tennis regarding surface preferences, points won, and match outcomes?
- **Tournament progression**: How do players' performances vary depending on their stage in a tournament (early rounds vs. finals)?
- **Betting site predictions**: Which bookmaker most accurately predicts tennis match outcomes?

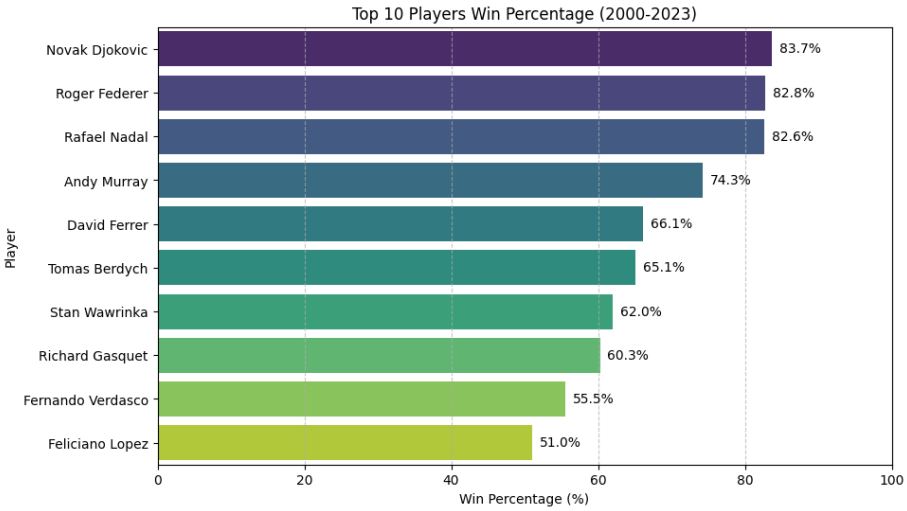
By answering these questions, we aim to transform raw data into an engaging and insightful experience, making tennis statistics more accessible and interactive for a wide audience.

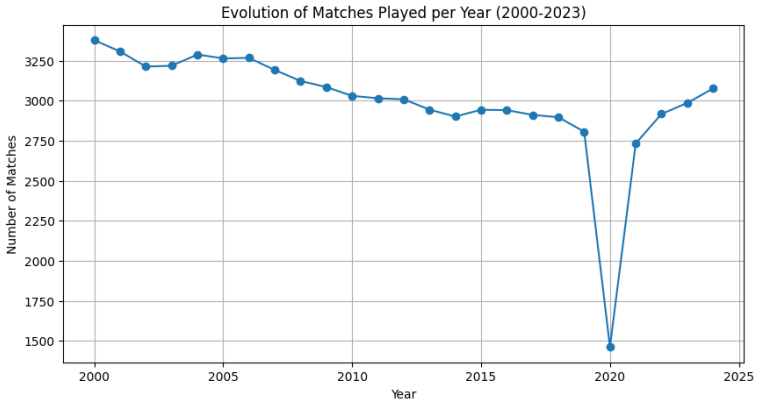
Exploratory data analysis

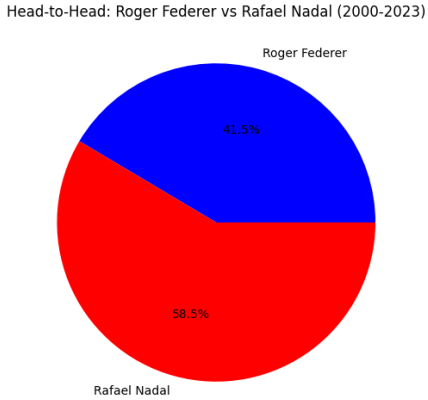


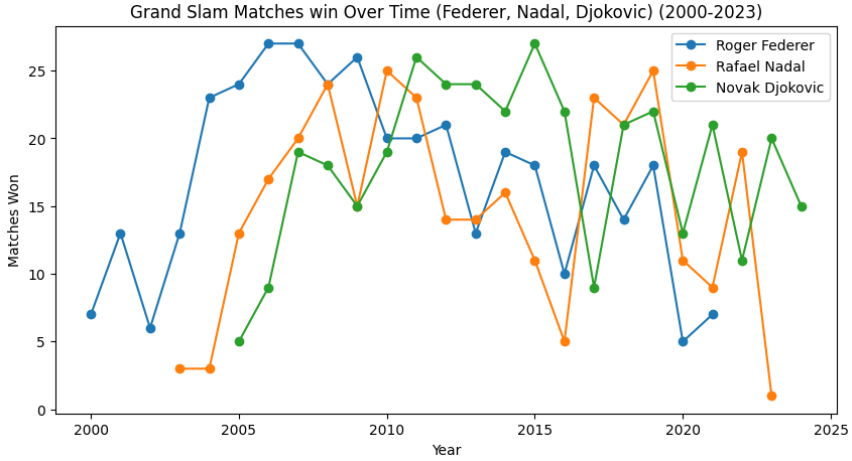


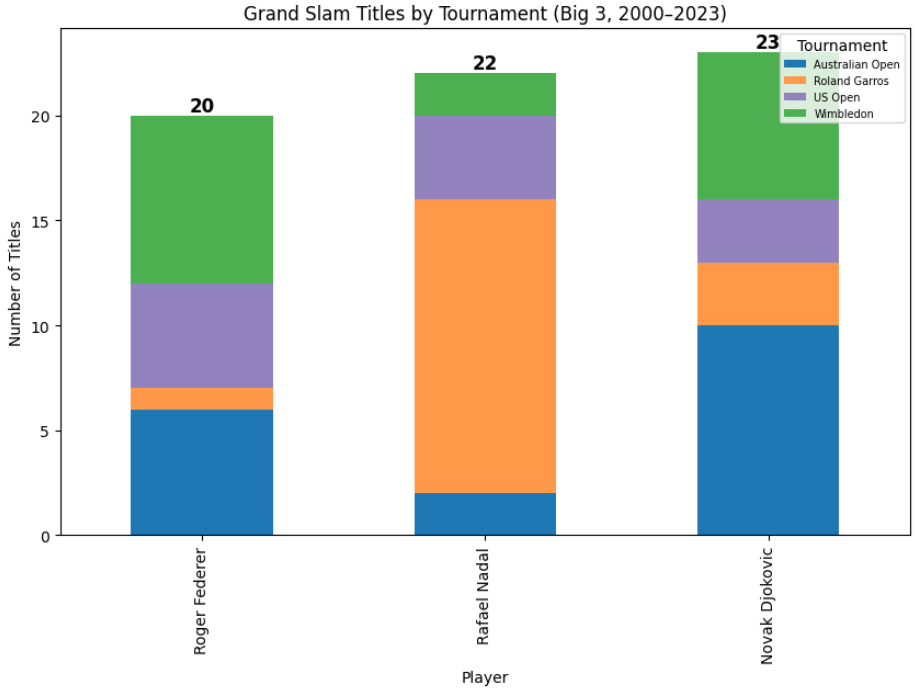





Evolution of Matches Played per Year (2000-2023)
Number of Matches
Year
3250
3000
2750
2500
2250
2000
1750
1500
2000
2005
2010
2015
2020
2025


Head-to-Head: Roger Federer vs Rafael Nadal (2000-2023)
Roger Federer
41.5%
58.5%
Rafael Nadal


Grand Slam Matches win Over Time (Federer, Nadal, Djokovic) (2000-2023)
Roger Federer
Rafael Nadal
Novak Djokovic
Matches Won
Year
25
20
15
10
5
0
2000
2005
2010
2015
2020
2025


Grand Slam Titles by Tournament (Big 3, 2000–2023)
Tournament
Australian Open
Roland Garros
US Open
Wimbledon
20
22
23
Number of Titles
Player
Roger Federer
Rafael Nadal
Novak Djokovic
20
15
10
5
0

Related Work

Jeff Sackmann's Tennis Abstract provides extensive statistical insights and predictive analytics on ATP and WTA matches, focusing on player statistics and historical comparisons but primarily targeting an expert audience.

IBM's Slamtracker offers dynamic, real-time visualizations and predictive analytics during Grand Slam tournaments but focuses mainly on match-level statistics without deeper historical or interactive exploration.

Official ATP and WTA websites provide interactive visualizations primarily oriented towards basic rankings and simple head-to-head statistics, lacking comprehensive and interactive insights into broader factors such as career disruptions or betting accuracy.

In terms of originality, our approach uniquely synthesizes multiple critical factors—career interruptions, betting markets, geographical conditions, rivalries, and gender differences—into one comprehensive interactive tool, making sophisticated analytics accessible to a wider audience. Inspiration for the visualization style and interactivity was drawn from narrative-driven and interactive examples seen in projects by FiveThirtyEight, IBM Slamtracker's dynamic dashboards, and storytelling-driven graphics from The New York Times, even though these examples are sometimes unrelated to tennis specifically. The selected datasets have previously been used primarily for isolated analytical tasks or betting-focused predictions in courses and individual analyses; our approach differs by combining multiple analytical layers and emphasizing interactive storytelling to facilitate broader public understanding and engagement.