Data Composition and Insight Report for Visualization: Analysis of the ATP Match Statistics.

ABSTRACT

This report analyzes the ATP tennis dataset compiled by Jeff Sackmann, which contains comprehensive historical player information, rankings, match results, and statistics. The dataset spans multiple decades, with rankings mostly complete from 1985 to the present and match statistics available from 1991 onward for tour-level matches. It includes detailed biographical data for players (such as name, hand, birth date, country, and height), match results across different tournament levels (tour-level, challenger, futures, and doubles), and ranking points for both players in each match. The dataset is structured to facilitate analysis, including redundant columns for biographical and ranking information and self-explanatory match statistics. This report aims to explore patterns in player performance, investigate trends in match outcomes, and demonstrate how historical ATP data can be leveraged to analyze player rankings, match statistics, and career trajectories over time.

Data Explanation

The JeffSackmann/tennis_atp repository, maintained by Jeff Sackmann of *Tennis Abstract*, is one of the most comprehensive open datasets available on men's professional tennis (ATP). It provides detailed, structured information on match results, player statistics, rankings, and biographical details spanning the entire Open Era (1968–2024).

The dataset is organized into multiple CSV files, each representing a specific year or data type. For instance, singles match results are stored in files named *atp_matches_YYYY.csv* (from 1968 to 2024), while lower-tier events such as Challenger, Qualifying, and Futures tournaments are included in their respective files. Doubles matches are also covered from 2000 to 2020 (*atp_matches_doubles_YYYY.csv*), though updates for doubles have been temporarily suspended since then. The repository also includes player and ranking datasets containing biographical details (e.g., nationality, height, handedness, date of birth) and weekly ATP rankings, with consistent records available from 1985 onward.

Data Type	Description	Time Span & File Examples
Match Results	Detailed records for individual matches, including the winner, loser, score, tournament details (name, date, surface, level), and often extensive match statistics.	Singles matches span from the beginning of the Open Era (1968) through the current year (atp_matches_1968.csv to atp_matches_2024.csv). Lower-level matches (Challenger, Qualifying, Futures) are also included in separate files.
Match Statistics	Per-match statistics (e.g., 1st serves in, total points won, aces) are provided for both the winner and loser. These are integer totals, which allow for calculating percentages.	Generally available from **1991 to present** for tour-level matches, **2008 to present** for Challengers, and **2011 to present** for tour-level qualifying.
Player and Ranking Data	Information on player biographies (ID, name, hand, date of birth, country, height) and weekly ATP rankings.	Rankings are mostly complete from **1985 to the present**, with intermittent data from 1973-1984. Biographical data is consolidated in separate files.
Doubles Matches	Tour-level doubles results are also included, though updates were suspended as of late 2020.	Matches from **2000 to 2020** (atp_matches_doubles_YYYY.csv).

Table 1: Data Description Table

Each match entry includes essential metadata such as player identifiers, tournament information, scores, and match statistics—covering aces, double faults, first serves in, and total points won. These statistics are typically available for tour-level matches from 1991 onward, for Challenger events from 2008, and for qualifying rounds from 2011. Each match row redundantly stores relevant biographical and ranking details for ease of analysis, ensuring that researchers can work efficiently without merging multiple tables.

The dataset is distributed in unencrypted CSV format on a public GitHub repository, ensuring accessibility and transparency. It is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0), which permits non-commercial use, requires appropriate attribution, and mandates sharing derivative works under the same terms. This makes it freely available for academic, educational, and research purposes, though it cannot be used commercially.

From a security and legal standpoint, the data is considered non-sensitive. It includes only publicly available professional information about athletes—names, nationalities, and match statistics—and therefore does not require encryption, secure storage, or special GDPR considerations. GitHub's version control system ensures data integrity and transparency, maintaining a full history of updates, timestamps, and change tracking.

Curated and maintained by Jeff Sackmann over many years, the dataset combines human verification with data drawn from official ATP sources, historical archives, and community contributions. This meticulous curation ensures high accuracy and completeness, with any invalid or inconsistent records corrected during data validation.

Overall, the JeffSackmann/tennis_atp dataset serves as a robust, well-documented foundation for sports analytics and research, enabling predictive modeling, performance analysis, historical exploration, and player comparisons across decades. While its raw structure resembles a large spreadsheet of numerical and textual data, its true value emerges through analysis and visualization—transforming into meaningful insights such as performance trends, rivalry networks, and global participation maps.

Dataset Sample:

```
matches_rec[23:29]
```

	tourney_id	tourney_name	surface	draw_size	tourney_level	tourney_date	match_num	winner_id	winner_seed	winner_entry	...	l_1stIn	l_1stWon	l_2ndWon	l_SvGms	l_bpSaved	l_bpFace
23	2024-0339	Brisbane	Hard	32	A	20240101	277	106186	NaN	WC	...	39.0	29.0	14.0	11.0	4.0	8.
24	2024-0339	Brisbane	Hard	32	A	20240101	276	104745	NaN	WC	...	34.0	23.0	10.0	9.0	3.0	6.
25	2024-0339	Brisbane	Hard	32	A	20240101	275	111442	NaN	NaN	...	32.0	18.0	10.0	8.0	4.0	8.
26	2024-0339	Brisbane	Hard	32	A	20240101	274	200005	4.0	NaN	...	38.0	24.0	18.0	10.0	8.0	10.
27	2024-0339	Brisbane	Hard	32	A	20240101	273	207830	NaN	Q	...	82.0	59.0	17.0	18.0	5.0	8.
28	2024-0339	Brisbane	Hard	32	A	20240101	272	208014	NaN	WC	...	47.0	36.0	6.0	10.0	2.0	4.

6 rows × 49 columns

Figure 1: ATP match dataset sample

```
players
```

	player_id	name_first	name_last	hand	dob	ioc	height	wikidata_id
0	100001	Gardnar	Mulloy	R	19131122.0	USA	185.0	Q54544
1	100002	Pancho	Segura	R	19210620.0	ECU	168.0	Q54581
2	100003	Frank	Sedgman	R	19271002.0	AUS	180.0	Q962049
3	100004	Giuseppe	Merlo	R	19271011.0	ITA	NaN	Q1258752
4	100005	Richard	Gonzalez	R	19280509.0	USA	188.0	Q53554
...	...	...	...	...	...	...	...	...
65984	213700	Matvei	Kobiakov	U	NaN	RUS	NaN	NaN
65985	213701	Tobia Costanzo	Baragiola Mordini	U	NaN	ITA	NaN	NaN
65986	213702	Dominik	Wijntjes	U	NaN	NZL	NaN	NaN
65987	213703	Sam	Wensley	U	NaN	AUS	NaN	NaN
65988	213704	Harry	Roberts	U	NaN	AUS	NaN	NaN

65989 rows × 8 columns

Figure 2: ATP Player dataset sample

Data Composition

The dataset used for this analysis, atp_matches_2023.csv, contains 2,986 rows and 49 columns, with each row representing an individual ATP tennis match. The columns capture a wide range of attributes, including tournament details, player information, and match statistics. Data types vary across the dataset, encompassing integers, strings, categorical values, and dates. For example, tourney_date records dates in the YYYYMMDD format, player_id is an integer string, winner_rank is an integer, and surface is a categorical string (e.g., "Clay", "Grass", "Hard"). Some columns, such as tourney_id, contain a mix of alphanumeric and numeric values, and several cells include missing entries.

The dataset is raw and unprocessed, with no prior cleaning, aggregation, or transformation applied. A README file provides a basic description of the dataset, but it lacks detailed explanations of variable meanings, units, or encodings. Metadata such as device error rates, byte order, or encryption information is not available.

```
matches_rec
```

	tourney_id	tourney_name	surface	draw_size	tourney_level	tourney_date	match_num	winner_id	winner_seed	winner_entry	...	l_1stIn	l_1stWon	l_
0	2023-9900	United Cup	Hard	18	A	20230102	300	126203	3.0	NaN	...	62.0	47.0	
1	2023-9900	United Cup	Hard	18	A	20230102	299	126207	NaN	NaN	...	12.0	8.0	
2	2023-9900	United Cup	Hard	18	A	20230102	296	126203	3.0	NaN	...	62.0	51.0	
3	2023-9900	United Cup	Hard	18	A	20230102	295	126207	NaN	NaN	...	41.0	26.0	
4	2023-9900	United Cup	Hard	18	A	20230102	292	126774	1.0	NaN	...	58.0	48.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
2981	2023-M-DC-2023-WG2-PO-RSA-LUX-01	Davis Cup WG2 PO: RSA vs LUX	NaN	4	D	20230204	5	202335	NaN	NaN	...	NaN	NaN	
2982	2023-M-DC-2023-WG2-PO-TUN-CYP-01	Davis Cup WG2 PO: TUN vs CYP	NaN	4	D	20230203	1	117365	NaN	NaN	...	NaN	NaN	

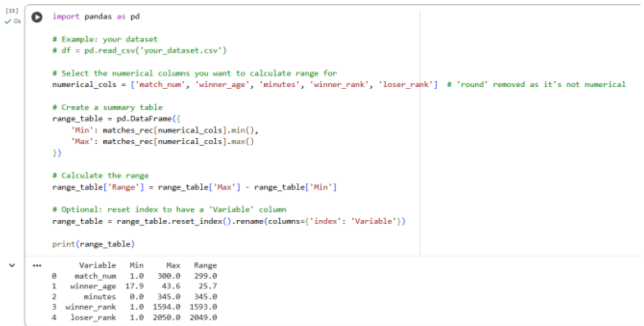
Figure 3: Sample Dataset

```
matches_rec.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3076 entries, 0 to 3075
Data columns (total 49 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   tourney_id     3076 non-null   object
 1   tourney_name   3076 non-null   object
 2   surface        3076 non-null   object
 3   draw_size      3076 non-null   int64
 4   tourney_level  3076 non-null   object
 5   tourney_date   3076 non-null   int64
 6   match_num      3076 non-null   int64
 7   winner_id      3076 non-null   int64
 8   winner_seed    1294 non-null   float64
 9   winner_entry   477 non-null    object
 10  winner_name    3076 non-null   object
 11  winner_hand    3076 non-null   object
 12  winner_ht      3057 non-null   float64
 13  winner_ioc     3076 non-null   object
 14  winner_age     3075 non-null   float64
 15  loser_id       3076 non-null   int64
 16  loser_seed     757 non-null    float64
 17  loser_entry    718 non-null    object
 18  loser_name     3076 non-null   object
 19  loser_hand     3076 non-null   object
 20  loser_ht       3026 non-null   float64
 21  loser_ioc      3076 non-null   object
 22  loser_age      3075 non-null   float64
 23  score          3076 non-null   object
```

Figure 4: Dataset variables info

As an observational dataset, the data records actual match events, and its accuracy depends on the recording process. However, no details are provided regarding procedures used to ensure correctness or reduce human error. These factors highlight the need for thorough preprocessing—such as cleaning, handling missing values, encoding categorical variables, and standardizing formats—before any meaningful analysis can be conducted.



```
import pandas as pd

# Example: your dataset
# df = pd.read_csv('your_dataset.csv')

# Select the numerical columns you want to calculate range for
numerical_cols = ['match_num', 'winner_age', 'minutes', 'winner_rank', 'loser_rank']  # 'round' removed as it's not numerical

# Create a summary table
range_table = pd.DataFrame({
    'Min': matches_rec[numerical_cols].min(),
    'Max': matches_rec[numerical_cols].max()
})

# Calculate the range
range_table['Range'] = range_table['Max'] - range_table['Min']

# Optional: reset index to have a 'Variable' column
range_table = range_table.reset_index().rename(columns={'index': 'Variable'})

print(range_table)
```

```
      Variable   Min     Max   Range
0    match_num   1.0   300.0   299.0
1   winner_age  17.9    43.6    25.7
2      minutes   0.0   345.0   345.0
3  winner_rank   1.0  1594.0  1593.0
4   loser_rank   1.0  2050.0  2049.0
```

Figure 5: Min, Max and Range

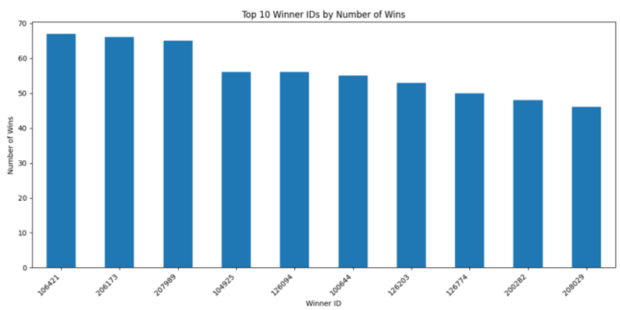


Figure 6: Top 10 Winner

Data use/purpose:

1. Task

The purpose of using the tennis dataset is to analyse player performance, match outcomes, and key statistics such as aces, double faults, first-serve percentages, break points, and win rates. The main objective is to uncover patterns, trends, and relationships between player attributes and their performance across tournaments, and communicate these findings effectively using visualisation techniques.

a. Holistic Task

The overall aim is exploration and explanation. By examining the dataset, I intend to identify factors that significantly influence match outcomes, such as serve efficiency, unforced errors, and player consistency. The insights gained can help highlight what differentiates high-performing players from others and provide a deeper understanding of professional tennis dynamics.

b. Analytical Task

Analytically, I will compare individual and aggregated player statistics, identify outliers (e.g., fastest serve, longest match duration), and examine trends over time or across tournament types. Correlations between variables, such as first serve success and match victories, will be explored to identify which metrics most strongly relate to winning performance. The analysis will focus on both descriptive and comparative statistics to provide a clear performance overview.

c. Implementation Task

The analysis will be performed using Python with libraries including Pandas, NumPy, Matplotlib, and Seaborn. The dataset will be cleaned and pre-processed to handle missing values, standardise formats, and ensure consistency. This preparation ensures that the subsequent analysis is accurate, reproducible, and suitable for generating reliable insights.

d. Visualisation Task

Visualisations will include bar charts (e.g., number of aces per player), scatter plots (e.g., serve speed vs. win percentage), line charts (e.g., performance trends over time), and heatmaps (e.g., correlation between metrics). The focus will be on clarity and interpretability, using colour and layout to emphasise key comparisons, trends, and outliers.

2. Environment (Who, When, Why, Where)

Who:
The primary audience includes sports analysts, tennis fans, and students interested in data visualisation or sports analytics. These users are expected to have basic visual literacy and can interpret graphs easily. Visuals will use colour-safe palettes to ensure accessibility.

When:
The visualisations will be used for academic purposes (e.g., coursework or research presentations) and for general analysis of player performance. They can be accessed anytime digitally.

Why:
Users will engage with the visualisation to understand player performance, match outcomes, and trends — for example, to see which players dominate specific aspects like serving or baseline play. The insights can also help in performance analysis or predictions.

Where:
The visualisations will be viewed on computers or projectors, mostly indoors such as classrooms, research labs, or presentations. Bright-light readability and high-resolution visuals will be considered for better viewing.

3. Build and External Requirements

The project will be built independently using Python-based tools. Since I have prior experience with data analysis and visualisation, I will develop and test the solution myself. The computational requirements are minimal — a system with 8GB RAM and Python installed will be sufficient.

External requirements include ensuring the dataset is clean and comprehensive. If the dataset lacks information (for example, player rankings or surface type), I may integrate external tennis data sources to strengthen the story. The final output will be adaptable for both digital reports and classroom presentations.

ACKNOWLEDGMENTS

I wish to express my deepest gratitude to Jeff Sackmann for providing the ATP tennis dataset, which served as the cornerstone of this analysis. I am also profoundly grateful to my supervisor, Jonathan C. Roberts, for his expert guidance, constructive feedback, and unwavering support throughout the development of this project.

REFERENCES

[1] J. Sackmann,“tennis_atp”,GitHub.[Online].Available: https://github.com/JeffSackmann/tennis_atp/blob/master/README.md. [Accessed: Nov. 06, 2025].