ATP Tennis Data Analysis

Ishika Patel
Chemical Engineering
Indian Institute of Technology
Gandhinagar, Gujarat, India
ishika.patel@iitgn.ac.in

Abstract— This project analyzes ATP men's tennis data from 2000 to 2024 to uncover patterns in player performance, match outcomes, and tournament trends. Using Python libraries like Pandas and Matplotlib, we explored key questions related to win rates, rankings, and court types, drawing meaningful insights through statistical analysis and visualizations.

I. Introduction

Tennis is a sport where strategy, physical performance, and environmental factors collectively determine outcomes. Over the years, ATP (Association of Tennis Professionals) tournaments have generated rich datasets that capture intricate match-level details — from player rankings and scores to court types and outcomes. This project leverages data from ATP matches held between 2000 and 2024 to analyze performance patterns across players and tournaments.

The core objective is to transform raw tennis data into meaningful insights using Python-based tools. By asking data-driven questions — such as the impact of winning the first set, the effect of surface types, and the distribution of matches across cities — the project provides a deeper understanding of player dynamics and match trends. The analysis not only helps highlight consistent performers but also validates intuitive aspects of the game through statistical evidence.

II. *SCIENTIFIC QUESTIONS AND HYPOTHESES*

A. What is the success rate in outdoor matches?

B. What is the chance of a player winning the match if they win the first set?

C. How many matches are played in each city?

D. No. of matches won by each player over 24 years?

E. Calculate the correlation coefficient of the winning rate and the winning points for each year?

F. What is the distribution of match outcomes (e.g., win, loss, walkover) in the dataset?

G. Determine the probability distribution of match outcomes (win, loss, retirement) for matches played on different surface types?

H. Covariance between a player's birth year and their height?

I. What is the probability that a match played in Adelaide results in a retirement?

J. Pie chart showing the distribution of matches played on different surface types?

K. What is the covariance matrix between match statistics and tournaments for the first 5 tournaments?

III. LIBRARIES

A. Pandas

Pandas is a powerful library used for data manipulation and analysis. It provides data structures like Series (1D) and DataFrame (2D), which allow efficient data storage and operations. In this project, Pandas was used extensively to load datasets from Excel and CSV files into DataFrames. concat: To merge multiple yearly datasets into a single unified DataFrame. To group data by specific columns (like Year, Winner) for aggregation and comparison and to convert date strings and extract year values from them.

B. Numpy

NumPy is the foundational package for numerical computing in Python. It works well with Pandas and Matplotlib. Though minimal in use here, NumPy supports efficient operations on arrays and matrices, and it's widely used for numerical computations.

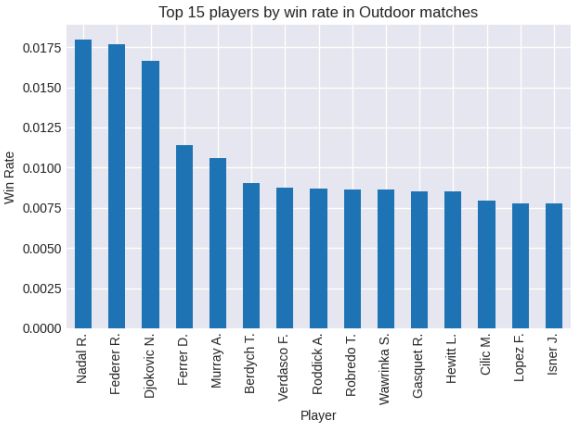
C. Matplotlib

Matplotlib is a powerful visualization library that enables plotting data in various informative formats. It supports creating bar charts, pie charts, histograms, scatter plots, and line graphs. In this project, I have used the pyplot module from Matplotlib to generate visual representations of trends such as match outcomes, player performances, and tournament patterns.

IV. ANALYSING DATA AND ANSWERING QUESTIONS

A. This analysis identifies the players who have won the highest proportion of all outdoor matches from 2000 to 2024, highlighting those with the most consistent and frequent outdoor victories across tournaments.

```
Top 5 players with the highest win rates in Outdoor matches:
Winner
Nadal R.       0.017973
Federer R.     0.017654
Djokovic N.    0.016658
Ferrer D.      0.011400
Murray A.      0.010592
Name: count, dtype: float64
```

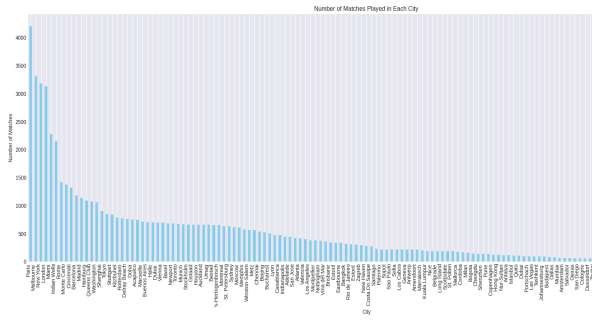


B. Only matches marked as "Completed" were considered to ensure accurate outcomes. Then, the matches where either player won the first set (`W1 == 0` or `L1 == 0`) were selected. Among those, matches where the winner also won the first set (`L1 == 0`) were counted. The conditional probability was calculated by dividing these two counts.

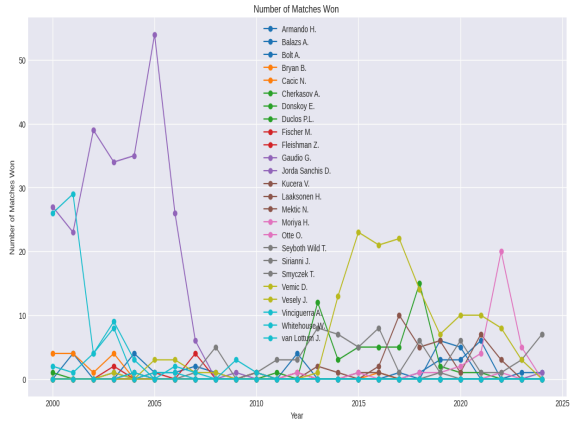
Output :

```
Conditional probability of winning the match given that the player won the first set:
0.9014404852160728
```

C. The `Location` column was used to count how many matches were played in each city across all tournaments from 2000 to 2024. These counts were visualized using a horizontal bar chart to compare city-wise distribution. The chart highlights which cities have hosted the most ATP matches. Cities like Melbourne, Paris, London, etc., likely appear at the top, reflecting their role as hosts of major tournaments (e.g., Grand Slams and Masters events).



D. The `Date` column was converted to extract the year for each match. For every year from 2000 to 2024, the player who won the least number of matches (i.e., the "worst" winner statistically for that year) was identified. Their total match wins were counted and visualized over time using a line plot.

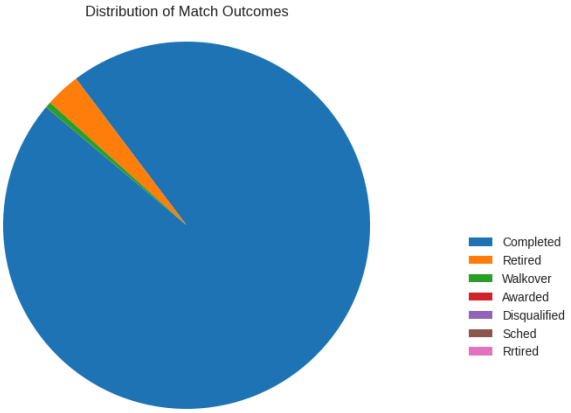


E. Each year's dataset from 2000 to 2023 was examined to calculate the correlation coefficient between a player's winning rank (`WRank`) and winning points (`WPts`). This was done only for years where both columns were present. The correlation coefficient quantifies the strength and direction of the relationship between these two variables.

Output:

```
'WPts' not found in the DataFrame for year 2000.
'WPts' not found in the DataFrame for year 2001.
'WPts' not found in the DataFrame for year 2002.
'WPts' not found in the DataFrame for year 2003.
'WPts' not found in the DataFrame for year 2004.
Correlation Coefficient for year 2005: -0.499283024943781
Correlation Coefficient for year 2006: -0.4101973705461944
Correlation Coefficient for year 2007: -0.43889963281500316
Correlation Coefficient for year 2008: -0.43404230873074917
Correlation Coefficient for year 2009: -0.40297003061544523
Correlation Coefficient for year 2010: -0.4575113525604047
Correlation Coefficient for year 2011: -0.3878656067437486
Correlation Coefficient for year 2012: -0.42177077855310213
Correlation Coefficient for year 2013: -0.43877165686536573
Correlation Coefficient for year 2014: -0.4243487005606218
Correlation Coefficient for year 2015: -0.383538735948183
Correlation Coefficient for year 2016: -0.3516387880692098
Correlation Coefficient for year 2017: -0.43893326948997713
Correlation Coefficient for year 2018: -0.45332414379551894
Correlation Coefficient for year 2019: -0.5008602872764123
Correlation Coefficient for year 2020: -0.5115092774999982
Correlation Coefficient for year 2021: -0.48383064400609715
Correlation Coefficient for year 2022: -0.572513877247325
Correlation Coefficient for year 2023: -0.4706851250410969
```

F. The pie chart shows the overall distribution of match outcomes across all tournaments from 2000 to 2024. Most matches were completed, but a significant portion also ended in retirements (players quitting due to injury or other reasons) and walkovers (matches awarded without play, typically due to a player's withdrawal).



G. Only completed matches were considered to ensure fairness in the results. The matches were grouped by court surface (e.g., Clay, Hard, Grass, Carpet), and the number of matches each player won on each surface was counted. These counts were then converted into probabilities by normalizing over each surface. This analysis helps understand how match wins are distributed among players across different surface types.

```
Probability distribution of match outcomes by surface type:
Winner    Hajek J.  Abel M.  Acasuso J.  Agamenone F.  Agassi A.  Agenor R.  \
Surface
Carpet     0.00000  0.00000    0.002448      0.000000   0.003060   0.000000
Clay       0.00005  0.00005    0.006616      0.000149   0.001940   0.000149
Grass      0.00000  0.00000    0.000000      0.000000   0.002769   0.000000
Hard       0.00000  0.00000    0.001147      0.000000   0.005706   0.000088

Winner   Aguilar J.  Ahouda A.  Ajdukovic D.  Al Ghareeb M.  ...  Zverev A.  \
Surface                                                     ...
Carpet      0.00000    0.00000       0.00000       0.000000  ...   0.000000
Clay        0.00005    0.00005       0.00005       0.000000  ...   0.006118
Grass       0.00000    0.00000       0.00000       0.000000  ...   0.004664
Hard        0.00000    0.00000       0.00000       0.000029  ...   0.006471

Winner   Zverev A.   Zverev M.  de Chaunac S.  de Voest R.  di Mauro A.  \
Surface
Carpet     0.000000   0.004284       0.000000     0.000612     0.000000
Clay       0.000000   0.001244       0.000050     0.000000     0.000597
Grass      0.000000   0.004081       0.000000     0.000292     0.000146
Hard       0.000059   0.002059       0.000088     0.000235     0.000029

Winner   di Pasquale A.  van Gemerden M.  van Lottum J.  van Scheppingen D.
Surface
Carpet         0.000000         0.000000       0.000000            0.000000
Clay           0.000945         0.000050       0.000398            0.000249
Grass          0.000146         0.000146       0.000583            0.000292
Hard           0.000206         0.000029       0.000176            0.000206

[4 rows x 1114 columns]
```

H. The covariance value indicates whether there's a positive or negative trend between a player's birth year and height: A positive covariance (which is likely here) suggests that newer-generation players tend to be taller. A negative covariance would have implied the opposite — that more recent players are shorter. While this doesn't confirm a strong correlation, it hints at a trend in player physiques evolving over time — possibly due to changing training methods, athletic expectations, or selection biases in modern tennis.

Output:

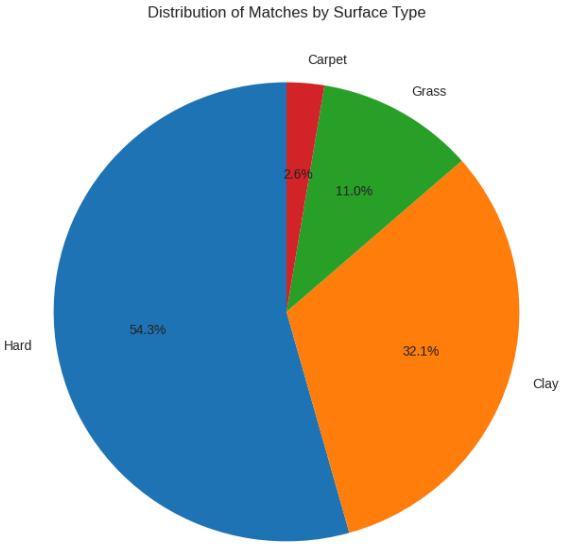
```
Covariance between a player's
birth year and their height:
2.339721134016972
```

I. This analysis gives a localized view of match interruptions due to retirement. It could be influenced by factors like climate, player fitness, or tournament timing.

Output:

```
Probability that a match played in
Adelaide results in a retirement:
0.02696629213483146
```

J. This visualization reveals which surfaces are most commonly used in professional ATP tournaments. Hard courts dominate due to their widespread use across many tournaments.
Clay courts come next, with major events like the French Open. Grass courts have fewer matches, mainly focused around Wimbledon. Carpet courts are now rare and mostly historical.



V. Summary of Observations

1. Winning the first set significantly increases a player's chances of winning the entire match, indicating the psychological and momentum advantage gained early in the game.
2. Counted the number of matches hosted by each city to understand which locations are most active in hosting ATP events.
3. Analyzed win rates of players specifically in outdoor court matches to identify top performers.
4. There is a strong inverse correlation between player rank and points, confirming

that higher-ranked players typically have more ATP points and perform better.
5. Identified the least frequent match winners for each year and analyzed their match history to see how their performance evolved over time.
6. Explored how different types of match results are distributed across the dataset
7. Visualized how matches are spread across court types like hard, clay, and grass using a pie chart.
8. Analyzed how the probability of match outcomes like win, loss, or retirement varies with different court surfaces.

Acknowledgment

I would like to thank my course instructor, Prof. Shanmuganathan Raman for providing the opportunity to work on this assignment.

References

[1] Numpy Documentation

[2] Pandas Documentation

[3] Matplotlib Documentation