Table of Contents

[Executive Summary 4](#_Toc175162718)

[Acknowledgment 5](#_Toc175162719)

[Introduction: 6](#_Toc175162720)

[Background 6](#_Toc175162721)

[Motivation 6](#_Toc175162722)

[Research Objective and scope of project: 7](#_Toc175162723)

[Literature Review 8](#_Toc175162724)

[Predicting the Outcome of a Tennis Tournament: Based on Both Data and Judgments 8](#_Toc175162725)

[Visualizing Professional Tennis Upsets: ATP 2012-2014 Men's Singles Matches 8](#_Toc175162726)

[Tracking the Evolution of a Tennis Match Using Hidden Markov Models 9](#_Toc175162727)

[Data-driven analysis of point-by-point performance for male tennis player in Grand Slams 10](#_Toc175162728)

[Capturing Momentum in Tennis 11](#_Toc175162729)

[Quantification of momentum in tennis matches and its impact: a study based on AHP-EWM method and data analysis 11](#_Toc175162730)

[Applications of Machine Learning to Optimize Tennis Performance: A Systematic Review 12](#_Toc175162731)

[Methodology 12](#_Toc175162732)

[Overview 12](#_Toc175162733)

[Description of the ATP tennis match dataset 13](#_Toc175162734)

[Data cleaning and pre-processing 15](#_Toc175162735)

[Primary key for each match 15](#_Toc175162736)

[Splitting column “scores” 15](#_Toc175162737)

[New metrics derived from dataset: 16](#_Toc175162738)

[Combined data frame which includes data from 2020 – 2023: 16](#_Toc175162739)

[Winner Rank distribution over Winner Height Distribution: 18](#_Toc175162740)

[Age vs Rank Correlation: 18](#_Toc175162741)

[Top 30 players for year 2023 based on Win Ratios: 19](#_Toc175162742)

[Visualizing winning pattern for top 30 players in a tennis match 20](#_Toc175162743)

[Distribution of Wins Across Different Tennis Surfaces 20](#_Toc175162744)

[Boxplot representations of derived metrics 21](#_Toc175162745)

[Statistical Tests for Analysing Tennis Surface Effects’ 21](#_Toc175162746)

[Analysis of Variance (ANOVA) for Break Points Saved 22](#_Toc175162747)

[Data Preparation 22](#_Toc175162748)

[ANOVA Model for Main Effects 23](#_Toc175162749)

[ANOVA Model with Interaction Term 23](#_Toc175162750)

[Subset Analysis by Era 23](#_Toc175162751)

[Implementation of Hidden Markov Models (HMM) for Analyzing Tennis Matches 24](#_Toc175162752)

[Data Preparation 24](#_Toc175162753)

[Implementation of Hidden Markov Model (HMM) 24](#_Toc175162754)

[References 26](#_Toc175162755)

**Table of Figures**

[Figure 1: Graphical model for awarding a point in a tennis match 10](#_Toc175161875)

[Figure 2 Momentum formula 12](#_Toc175161876)

[Figure 3: Methodology flowchart 13](#_Toc175161877)

[Figure 4 Meta data for column match\_id for dataframe [atp] 15](#_Toc175161878)

[Figure 5. Scores of each set for matches played in ATP 2023 15](#_Toc175161879)

[Figure 6 Average Minutes per Tournament Level for 2020-2023 16](#_Toc175161880)

[Figure 7 . Average Minutes per Tournament Level for 2000-2003 17](#_Toc175161881)

[Figure 9 Average winner age per tournament level for years 2020-2023 17](#_Toc175161882)

[Figure 8 Average winner age for each tournament for years 2000-2003 17](#_Toc175161883)

[Figure 10 Winner Rank vs Height Distribution for players in 2023 18](#_Toc175161884)

[Figure 11 Age vs Rank Correlation for year 2023 18](#_Toc175161885)

[Figure 12 Age vs Rank Correlation for year 2001 19](#_Toc175161886)

[Figure 13 Bar Chart representation of top 30 players according to win ratios 19](#_Toc175161887)

[Figure 14 Winning trend for top 30 players 20](#_Toc175161888)

[Figure 15 100% stacked bar viz for top 25 players across different surfaces 20](#_Toc175161889)

[Figure 16: Serve efficiency and Ace% Boxplot viz for winner and loser for year 2023 21](#_Toc175161890)

[Figure 17 : Boxplot viz : Breakpoints Saved, & 1st serve point won% for winner and loser in 2023 21](#_Toc175161891)

# Executive Summary

The primary objective of this dissertation is to broaden the understanding of the variables influencing professional tennis match results by exploring into the dataset of ATP tennis match statistics. This report looks at key performance measures like break point saving ratio, ace rate, and first point win % over several ATP seasons using the extensive Exploratory Data Analysis (EDA)technique. This Study involves having around ~ 4000 data points by extracting ATP matches from year 2019 to 2023 across tourneys of all levels.

Advanced statistical methods such as ANOVA and Tukey HSD were employed to examine surface-level data, resulting in substantial findings into the variations in player performance on various court surfaces. The study used Markov Chain modelling to extract hidden states from the performance measurements of the winner and loser in order to gain a deeper understanding of match dynamics and to illuminate the momentum shifts that occur during a match.

Leveraging historical match data, we attempted to implement a Classification machine learning model that predicted the result of the important third set, building on these fundamental studies. Building our understanding over several key match performance metrics and incorporating it by using relevant feature engineering techniques. The capacity to forecast the future may have significant effects on betting markets, in-game decisions and game strategy.

# Acknowledgment

I would not have been able to complete this dissertation without the help of others, I would like to take this opportunity to thank them. Firstly, I would like to thank University of Strathclyde for the opportunity, it’s been a great learning experience. I would also like to thank my supervisor Ruaraidh McPike who answered all my questions and supported me throughout the dissertation, helping me understand the projects objectives and setting up feedback meetings ultimately improved the quality of my research work.

# Introduction:

## Background

Ever since the modern days of sports have been integrated with the application of Data Analytics, every aspect of the game from players performance to team strategy, fan engagement to business model has been transformed into multiple folds. (Srivastava, 2023) .

According to a report by Fortune Business Insights, global sports analytics market has forecasted to reach form $3.78 billion in 2023 to $22.13 billion by 2030, showing a substantial growth of 28.7% CARG. One prime example for this is the sports of F1, which at current time is so data driven backed up by AWS, providing top notch viewing experience for the viewers by showing real time stats during race like tire deterioration, traction ratio , all the sector times with coloured mini sectors, too see where exactly a driver has improved his lap time and DRS status .

First official recorded match statistics for a Tennis match could be found back in 1991. Since then the hidden complexities associated with the game of tennis has always be an challenge in the statistical analysis of the game. In late 2013, Novak Djokovic followed by Federer in 2017 , data analytics was introduced . Tennis unlike other sports is a very individualistic game, hence the trend of advanced analytics was slow to gain traction. Drone technology and 3D printing is used in new tennis ball machine to generate a complex set of shots using a Smartphone to control speed and spins. Building new tennis metrics – Big technology players IBM, Infosys, and SAP are continually upgrading new tennis metrics. Drone technology and 3D printing are employed in a new tennis ball machine to manufacture a complex series of strokes, with a Smartphone controlling speed and spin. Big technology companies like IBM, Infosys, and SAP are constantly updating new tennis metrics. (Drive-analytics, n.d.)

## Motivation

As quoted by Serena Williams, professional tennis player *“Data science has revolutionized how we approach training and preparation. It provides us with valuable insights that help us optimize performance and reach our full potential”.* “MoneyBall”, a 2011 film based on the story of a baseball team, Oakland Athletics who changed the landscape of competitive baseball by leveraging Sabertmetrics. Not only they won 20 consecutive games but also did that but spending lowest cost per win than other games, as it started with mere 10000 data points which has now rose to a staggering 10 billion data points.( <https://bbiasblog.com/2022/12/10/moneyball-how-sabermetrics-changed-baseball-forever/>)

Tennis being a individualistic game, the competitiveness of the game keeps of increasing year after year. The relevance of data analytics in Tennis could be a deterministic parameter between winning or losing a crucial series. With young talents like Carlos Alcaraz emerging, EDA and advanced statistical models could be leveraged to recognise such early trends. Majority of top 20 ATP professional athletes have started to leverage the use of analytics. The adaptation of this technology is slow but incremental unlike in other sports mainly due to the nature of the game. Currently the main technique for analysis is Labelling the match videos to evaluate various player performance metric. Point progression in each set, momentum shifts, Break points analysis.

## Research Objective and scope of project:

Given the extensive scope of analysing ATP tennis match data, the main objective of this dissertation is to study the dataset and define particular research questions that can provide significant insights into the factors that influence match outcomes in professional tennis.

The primary purpose is to undertake a thorough Exploratory Data Analysis (EDA) on the most recent complete ATP season, with the goal of understanding significant statistics and visualising them. This preliminary research will help to find interesting features that can be examined further, possibly through the analysis of historical data from prior seasons.

This dissertation will investigate the following research directions:

1. Examining "Upset" Matches: Using the match data, the study will try to pinpoint the salient features or traits of "upset" matches—those in which a player with a considerably lower ranking than the opponent wins—in order to arrive at conclusion. This will entail establishing the parameters for what qualifies as a "upset" and examining the variations in match statistics between upset and non-upset contests.
2. Comparing Match Statistics Across Eras: This research will examine the shifts in match statistics in professional tennis over various eras by utilising match data from several ATP seasons. This comparative study can shed light on how the sport's dynamics, tactics, and general level of competitiveness are evolving. Deeper insights into the observable traits will be found by using advanced statistical techniques like ANOVA and Tukey HSD.
3. Third Set Match Outcome Prediction: As an attempt to provide further insights, dissertation will explore developing a machine learning model to predict the winner of the third set in a match. Historically, the momentum and final result of the match have been greatly influenced by the third set. In order to calculate the likelihood that a player would win the crucial third set, the model will make use of players' meta data and performance measures from the earlier sets.

With possible applications in player development, coaching tactics, and tournament preparation, the investigation of these study aims will deepen our understanding of the elements that impact professional tennis success.

# Literature Review

## Predicting the Outcome of a Tennis Tournament: Based on Both Data and Judgments

In the field of tennis match prediction, a groundbreaking study emerged that combined empirical data with expert judgments, achieving a remarkable 85.1% accuracy rate. This research utilized an extensive dataset from the ATP World Tour, spanning from 1968 to 2015, encompassing 165,974 data points across 44 features. The study employed an Analytic Network Process (ANP) model to predict match outcomes, focusing specifically on contests between Djokovic and Federer. To assess the significance of player performance metrics, the researchers conducted a Wilcoxon rank sum test on 20 key features. Linear regression was then applied to determine the relative importance of each feature through coefficient analysis.

In preparing data for the ANP model, the researchers innovatively grouped similar performance factors into clusters, effectively reducing the number of factors from 20 to 12. This clustering approach ensured that the model's inputs were both comprehensive and manageable. A key aspect of this study was its integration of expert opinions and pre-match judgments. These judgments were carefully constrained to be based solely on the features incorporated into the ANP model. Experts made paired comparisons drawing on their knowledge and experience, while also considering correlation coefficients from historical data. The researchers also factored in the current form of players and additional judgment opinions to create a prioritized list of influencing factors.

The ANP model's super matrix was then used to synthesize all these priorities, producing a final prediction. This methodology represents a significant advancement in tennis match prediction, combining the strengths of data-driven analysis with the nuanced insights of expert judgment. This approach not only demonstrated high predictive accuracy but also provided a framework for understanding the complex interplay of factors that influence tennis match outcomes. The study's success in integrating quantitative data with qualitative expert knowledge opens new avenues for sports analytics and predictive modelling in tennis.

## Visualizing Professional Tennis Upsets: ATP 2012-2014 Men's Singles Matches

This blogs by Tyler Knutson (NYC Data science academy) walks though the analysis of performance metrics and Odds of upset matches; where an underdog player defeats a most likely a favourite / higher seed player. So, identify the winner of the match before the match, the concepts of “Odds” is leveraged, Odds greater than 2.0 denotes an underdog while odds less than 2.0 represents the favourite. A density distribution of break points in a upset match is studied to analyse the impact on break point metric. The dataset used in this is very rich in terms of the detail and the granularity, as it has metrics like no of backhands, forehands, volleys, how deep was service return etc. This mainly highlighted 2 main findings:

Firstly, studies have observed a tendency among bookmakers to allocate their resources disproportionately when setting odds. Bookmakers focus more intently on matches with less predictable outcomes, dedicating less time to precisely calibrating odds for matches with clear favourites. For instance, in matches with heavy favourites, the difference between setting underdog odds at 20:1 versus 18:1 was often neglected.

Secondly, analyses of betting patterns across multiple sportsbooks have revealed that odds are frequently adjusted not just based on perceived probabilities, but also in response to the distribution of wagers. In a notable example, when two sportsbooks initially offered identical odds (1.3 for the favourite, 3.5 for the underdog), the book receiving heavier action on the favourite adjusted their odds to 1.2 and 4.5 respectively. This adjustment serves to encourage more balanced betting, allowing sportsbooks to manage their risk exposure more effectively.

The logit function and its applications in sports modelling

This (Hyatt-Twynam, 2013) blog post provides in-depth exploration of logit function and its applications in sports application. Logit function can be used as transformation while working with probabilities. Linear shifts in logit space are the fundamentals behind logistic regression. The author presents a compelling case for the utility of logit function in transforming probabilities to improve modelling accuracy and interpret uncertainties. The author argues that linear shifts in probability space can lead to unrealistic outcomes, especially in extreme probability scenarios. The logit transformation is presented as a solution, allowing for more sensible linear adjustments in logit space before converting back to probabilities.

The blog post effectively demonstrates the versatility of the logit function in sports modelling, from basic probability transformations to more complex applications in Bayesian inference and market pricing. While the author provides personal insights and suggestions based on industry experience, the post does not reference specific academic literature or industry studies. This limits its academic rigor but enhances its practical relevance for those working in sports analytics. The author emphasizes the critical importance of quantifying uncertainties in probability estimates for both bookmakers and sports traders. Instead of relying on a single probability value, the blog suggests developing a probability density function (pdf) that describes the likelihood of each possible true probability value.

The blog introduces the logit-normal distribution as an ideal choice for representing this uncertainty. This distribution is created by applying a normal distribution in logit space. The author states: "Typically in science, for physical measurements, the logical choice would be a normal distribution, however if this is used for p\_A, you're claiming there's a chance p\_A could be outside the range 0 to 1. Once again, logit space comes to the rescue; using a normal distribution in logit space (a logit-normal distribution) for your prior solves this issue, and leads to a narrower distribution at the extremities."

## Tracking the Evolution of a Tennis Match Using Hidden Markov Models

The author IIias Kolonias, William Christmas and Josef Kittler has addressed the problem of extracting higher level semantic information from low level feature from multimedia content. Feeding a sequence of video as input data in hierarchical structure consisting of Hidden Markov Model. A graphical model is developed to represent the usual progression of the game, which helps to identify and examine elementary events within the tennis sequence. Using this fundamental sequence of events, reasoning for higher events is performed such as awarding the current point. Computer Vision models is used to detect the event and HMM is used to perform higher level reason.

This (Ilias Kolonias, August 2004) works aims in substituting the original scene evolution model with a series of smaller models, each of which aims to accurately depict a particular situation in the match's history. The most crucial thing we must make sure of during this process is that, upon combining every model in this set, we will have a model that is identical to the original.

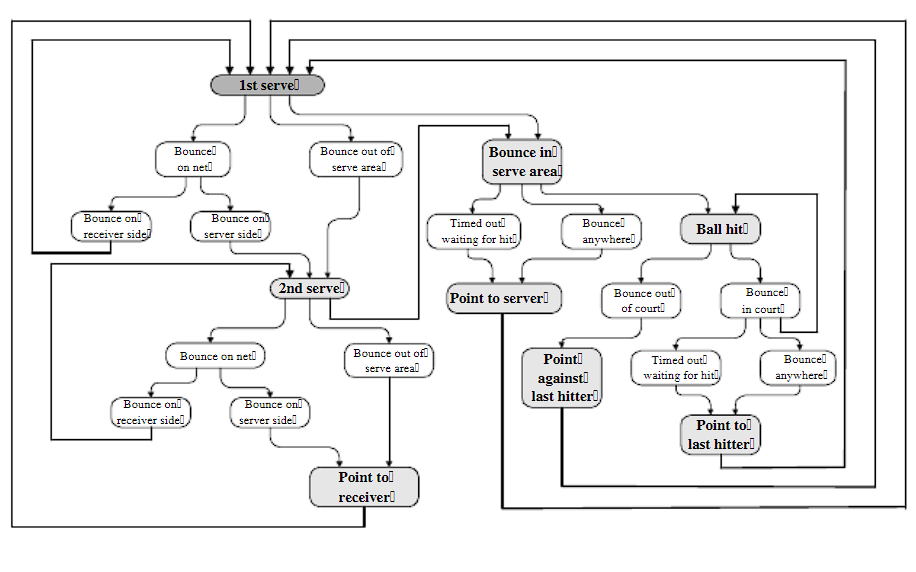


Figure 1: Graphical model for awarding a point in a tennis match

Modelled events of lower conceived importance through an HMM, which would then trigger another HMM to infer on more important events within the game; that would also help us prevent spurious data from low-level feature extraction modules from propagating to higher levels of the inference engine

## Data-driven analysis of point-by-point performance for male tennis player in Grand Slams

This (Yixiong Cui, 2019) article illustrates the work on analysing the performance metrics of tennis by incorporating 29675 data points through Classification Tree analysis. The author leveraged the point level data of 145 Grand slams main draw men’s Singles matches played by top ranked players from year 2011 to 2016. A two-step cluster analysis was performed with Euclidean distance as measure and Schwart’s Bayesian criterion. The classifying metrics were (i) Serve speed (ii) Rally length.

This article projects the point level analysis of players using classification tree model. Humidity and Ball types were 2 unique features which could be found being used by the author. Another key finding was due to the slow court surface of Roland Garros, it was easier to break to the server’s game and players tend to player more aggressively. The influence of Rally points was noticeable in certain ways: Short and longer rally helped player to win more matches than medium rally . The serving efficiency of a player is not only affected by the surface type and the level of the tournament but also by the skill level of the opposition

In a thorough examination of tennis performance throughout Grand Slam tournaments, the author discovers that a classifying tree model portrays some critical insights concerning a player's effectiveness. The model shows that the analysed player had higher overall point-winning percentages in the Australian Open (AO) and US Open (US) than Roland Garros (RG) and Wimbledon (W).

Specifically, the tree model shows that serve speed has a considerable impact on point-winning chance. The player had a greater success rate with rapid serves throughout all events, with a clear advantage in the AO and US. Slower serves, on the other hand, resulted in decreased point-winning percentages, which were most noticeable in the AO and US. Interestingly, it demonstrates consistent return efficacy throughout all Grand Slams, with the player winning roughly one-third of points against first serves and 40% against second serves, regardless of event. This classification approach offers a comprehensive understanding of how surface qualities and serve dynamics influence match outcomes, which is useful for strategic planning and performance analysis in professional tennis.

## Capturing Momentum in Tennis

This (Manuel, March 4 2022) article sheds light upon identifying decisive momentum shifts moments in a tennis match and how such moments can be influential in loosing and winning the match. Leverage and Momentum are the 2 new Stats performance metrics which was introduced to identify key moments in WTA matches . Leverage measures the importance of a single point to the final outcome of a tennis match by quantifying how much a player’s probability of winning the match changes. Momentum aims to describe which player is in control at any point of the match

In a comprehensive study of tennis match dynamics, researchers have developed a sophisticated chain of predictive models to analyse and quantify the concept of momentum in professional tennis. The study, which utilizes data from 1.5 million points played on the WTA tour between 2012 and 2020, presents a novel approach to understanding the ebb and flow of tennis matches.

The authors introduce a multi-layered model that considers various factors including court type, current match state, in-match statistics, and pre-game odds. This model chain begins by predicting the probability of winning the next point, which then feeds into predictions for game, set, and ultimately match outcomes. This hierarchical approach allows for a granular analysis of how individual points impact the overall match probability.

Central to the study is the concept of "leverage," which quantifies the importance of each point in terms of its potential to change the match outcome. Building on this, the researchers define "momentum" as an exponentially weighted moving average of the leverage gained by a player. This definition takes into account both the recency and the importance of points, with more recent and higher-leverage points having greater influence on the momentum score.

The (Manuel, March 4 2022) study introduces the concept of a "momentum swing," defined as a shift in momentum from one player to another by a magnitude of 3% or more. This threshold provides a concrete metric for identifying significant turning points within a match. This research represents a significant contribution to the field of sports analytics, offering a data-driven approach to quantifying the often-intangible concept of momentum in tennis. The model's ability to consider a wide range of factors and provide point-by-point analysis of match dynamics offers potential applications for players, coaches, and analysts in developing strategies and understanding match progression.

## Quantification of momentum in tennis matches and its impact: a study based on AHP-EWM method and data analysis

In this report (Wu, May 2024) , the author deeply analyses the concept of momentum formation in tennis match and its extent of influence on winning the match. Using Analytic hierarchy process (AHP) and the entropy weight method (EMP) methodology, Welch’s Test and Run test , non-randomness of momentum metric is investigated. One interesting metric considered by the author during data pre-processing stage is p1\_d : the longer the distance traversed by an athlete, the greater their level of exhaustion consequently. Evaluating Criterion Matrix, final AHP and EWM weights are computed which are then fed into a concise formula to calculate momentum

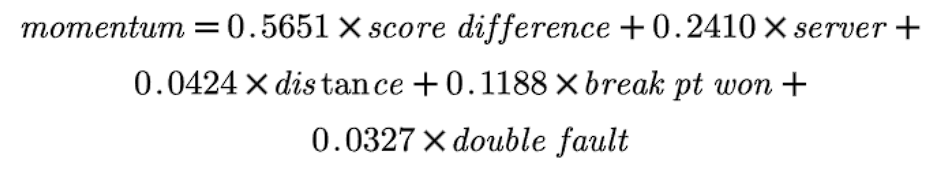
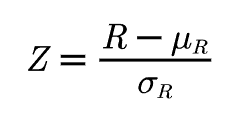


Figure 2 Momentum formula

To establish the non random nature of momentum and likelihood of winning the match, Z test run is performed using below given formula:



Hypothesis test was undertaken on Wimbledon 2023 tournament – 1301 as reference for player Carlos Alcaraz . Z value obtained was -7.402, which indicate that w that the data are non-random data based on the variableMomentum\_p1, and the level is significant

## Applications of Machine Learning to Optimize Tennis Performance: A Systematic Review

This (Tatiana Sampaio, June, 2024) report sheds light over various approaches adopted to assess the performance stats of players. Notable reviewed mythologies include Psychological and affective states where a sensor based approach was employed to leverage inertial measurement units (IMUS) (Havlucu Hayati, 2022) worn by small group of elite coaches (n=2) and 4 professional players. BY using Long short-term Memory Recurrent Neural Network, an accuracy of 85% was achieved. He further dwelled over the scope of AI for predicating players optimal performance “zones”.

This report also displays how various ML approaches yields different results for instance studies include Backpropagation and Neural Network and Convolutional Neural Networks being utilized to evaluate tactical performances and classifying groundstroke stances (Yu Zhang, 2023), (Yang, January 2023) .While Random Forest was utilized to perform Classification of moment patters and predicting point winners for a set. (Rosker Jernej, 2021) (Andrej Panjan, June 2010). (Yu Zhang, 2023) Designed a diagnostic model for assessing player performance using decision trees and a common classification technique while (Yang, January 2023) deployed convolutional neural network to analyse batting strength and angles based on video footage. CNNs excel at image recognition tasks, making them ideal for this application.

# Methodology

## Overview

The flowchart above provides a structured overview of the methodology employed in this research, detailing the step-by-step process undertaken to analyse player performance and build a classification model. The methodology is divided into distinct stages, each contributing to the overall objective of understanding and predicting player outcomes based on historical and current data.

This methodology provides a robust framework for analysing player performance over time, comparing different eras, and building predictive models based on comprehensive and normalized datasets. The use of statistical tests ensures the reliability of the findings, while the classification model serves as a practical tool for forecasting player success

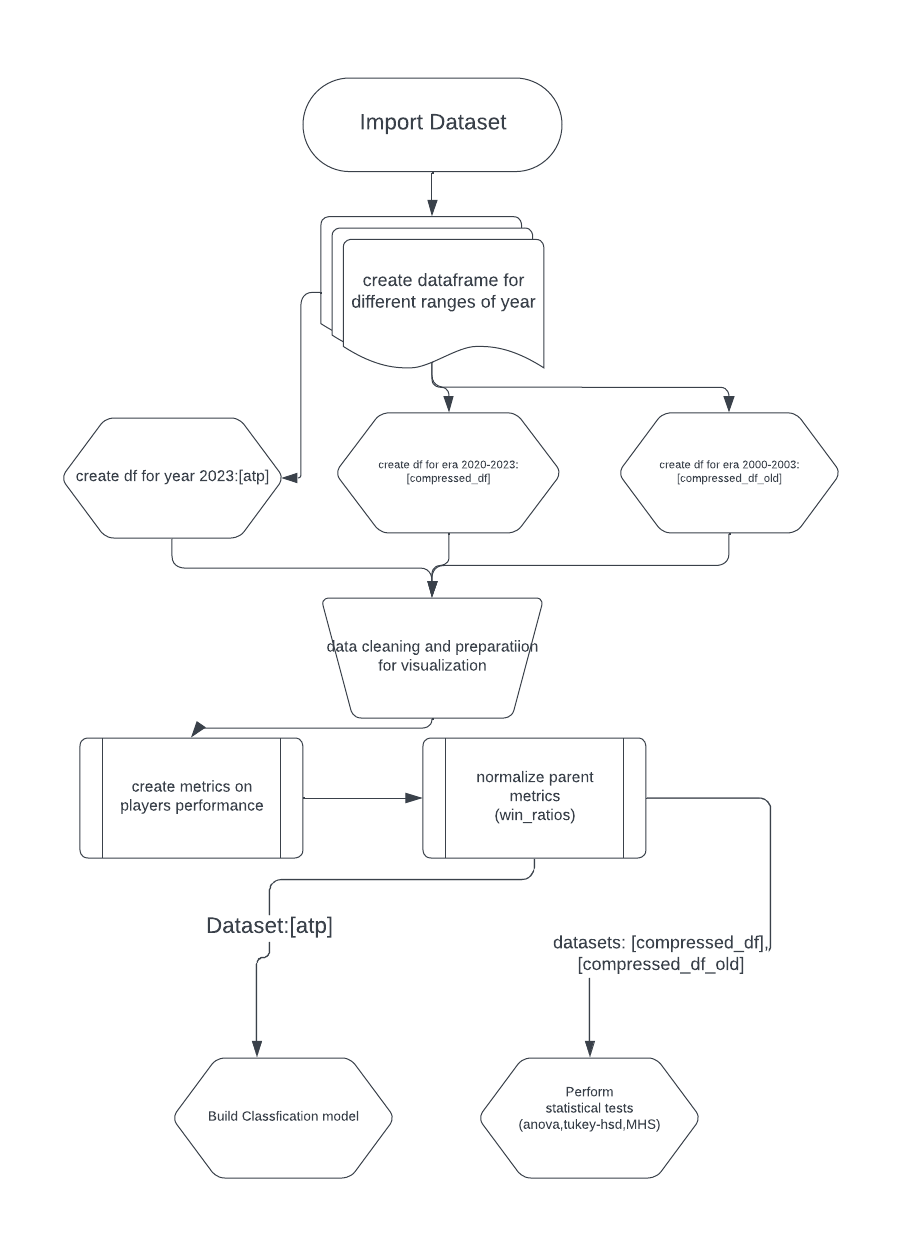


Figure 3: Methodology flowchart

## Description of the ATP tennis match dataset

The dataset (JeffSackmann, 2016) used for this report is from GitHub repository by JeffSackmann .

This dataset contains comprehensive information about tennis tournaments and matches, including player statistics and match outcomes. The data is organized into several categories:

1. Tournament Information:

|  |  |
| --- | --- |
| tourney\_id | A unique identifier for each tournament |
| tourney\_name | The name of the tournament, as various tourneys are played at different levels are listed in this : 470 different tourneys held from 2020-2023 |
| surface | The type of court surface , possible values are [Hard, Grass, clay] for year 2000-2003 era Carpet as a surface was also found |
| draw\_size | The number of players in the tournament |
| tourney\_level: | Indicates the level of the tournament (e.g., Grand Slam, Masters 1000, other level tourney, Challenger’s, Satellites/ITFs, finals, Davis cup) => Distinctive tourney levels values are [G,M,A,C,S,F,D] |
| tourney\_date | The start date of the tournament |

2. Match Information:

|  |  |
| --- | --- |
| match\_num | A match-specific identifier |
| score | The final score of the match |
| best\_of | Indicates whether the match is best of 3 or 5 sets |
| round | The tournament round of the match |
| minutes | The duration of the match |

3. Player Metadata:

For both winner and loser:

|  |  |
| --- | --- |
| player\_id: | Unique identifier for each player |
| name | Player's full name |
| Hand | Player’s dominant hand |
| ht | Player’s height in cms |
| ioc | Player’s country code |
| age | Players age at the time of the tournament |
| rank | Players ATP rank at the time of the tournament |
| rank\_points | Player’s ranking points |

5. Match Statistics:

For both winner and loser:

|  |  |
| --- | --- |
| ace | Number of aces served |
| df | Number of double faults |
| svpt | Total serve points |
| 1stIn | Number of first serves made |
| 1stWon | Number of first-serve points won |
| 2ndWon | Number of second-serve points won |
| SvGms | Number of service games |
| bpSaved | Number of break points saved |
| bpFaced | Number of break points faced |
| seed | Player's seeding in the tournament |
| entry | Type of entry into the tournament (e.g., wild card, qualifier) |

This dataset provides a rich source of information for analysing tennis matches, player performance, and tournament characteristics. It allows for in-depth analysis of various aspects of the game, including serving performance, player rankings, and match outcomes across different tournaments and surfaces.

## Data cleaning and pre-processing

We have imported the csv file formats of ATP tennis data for years 2000 to 2003 and for years 2020 to 2023. Following are the Data pre-processing steps performed in the dataset :

## Primary key for each match

We concatenated the match\_num column and tourney\_id column to uniquely identify each match . The new column created was match\_id which acted like a primary key .

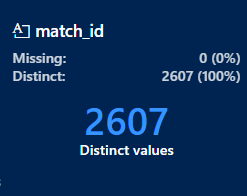


Figure 4 Meta data for column match\_id for dataframe [atp]

## Splitting column “scores”

To study in depth about the set progression for each match, we split the set scores into individual columns [set1, set2,set3,set4,set5]. The split columns will be re-framed into original dataframe using the “match\_id” column as our primary key.

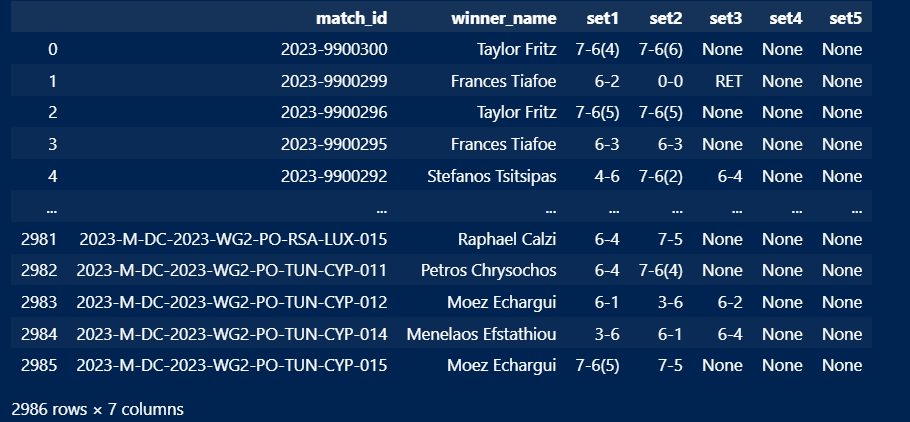


Figure 5. Scores of each set for matches played in ATP 2023

Other columns derived from column scores are:

For winners and losers:

* 'set1\_gm\_diff', 'set2\_gm\_diff', 'set3\_gm\_diff', 'set4\_gm\_diff', 'set5\_gm\_diff: these columns hold the game differences for each set.
* 'set1\_ot\_diff', 'set2\_ot\_diff', 'set3\_ot\_diff', 'set4\_ot\_diff', 'set5\_ot\_diff' : these columns hold the tiebreaker differences for each set.

## New metrics derived from dataset:

Following are the new metrics that were derived from the existing columns from the dataset:

* 1st\_srv\_w%: First serve win percentage -> Denotes the percentage of 1st serves won over total number of 1st serves played.
* 2nd\_srv\_w%: Second serve win percentage - > Denotes the percentage of 2nd serves won over total number of 1st serves played.
* bp\_saved%: Break points saved percentage -> Denotes the percentage of break points saved by player in the match over total number of break points faced.
* ace%: Ace percentage : denoted the ace conversion of a player over total number of 1st serve points
* serve\_efficiency% : Serve efficiency percentage -> denotes the serving efficiency of a player in a match . Formula used to derive this metric is:

## Combined data frame which includes data from 2020 – 2023:

Exploratory data analysis over ATP data from year 2020 to 2023 for average duration in minutes per Tournament level over years was performed. A grouped bar plot was created to visualization who the match duration has varied over years for each tournament. The subset of a dataframe was created using aggregate function over minutes column. An interactive python data visualization library called Plotly was used to implement this. Similarly a pair of grouped bar plots for year 2020-2023 and 2000-2003 was implemented to visualize average age of winners ever tournament level

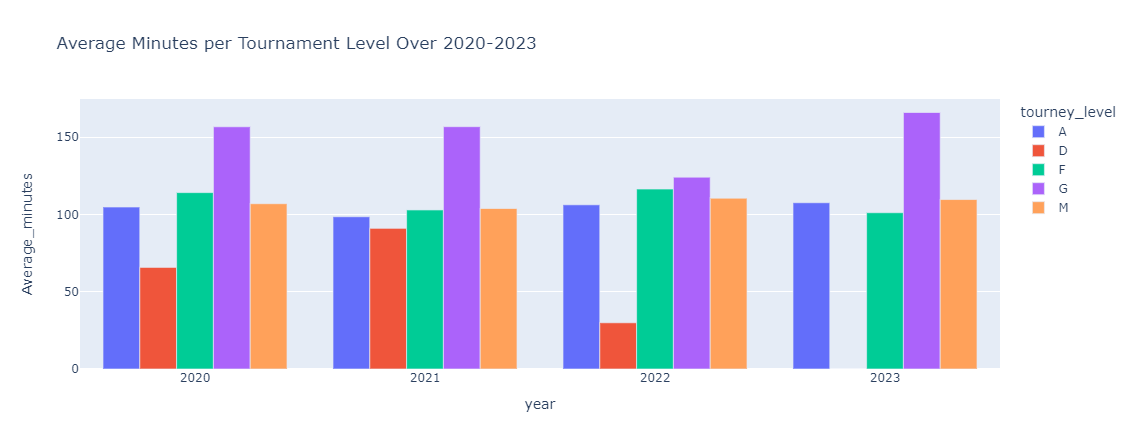


Figure 6 Average Minutes per Tournament Level for 2020-2023

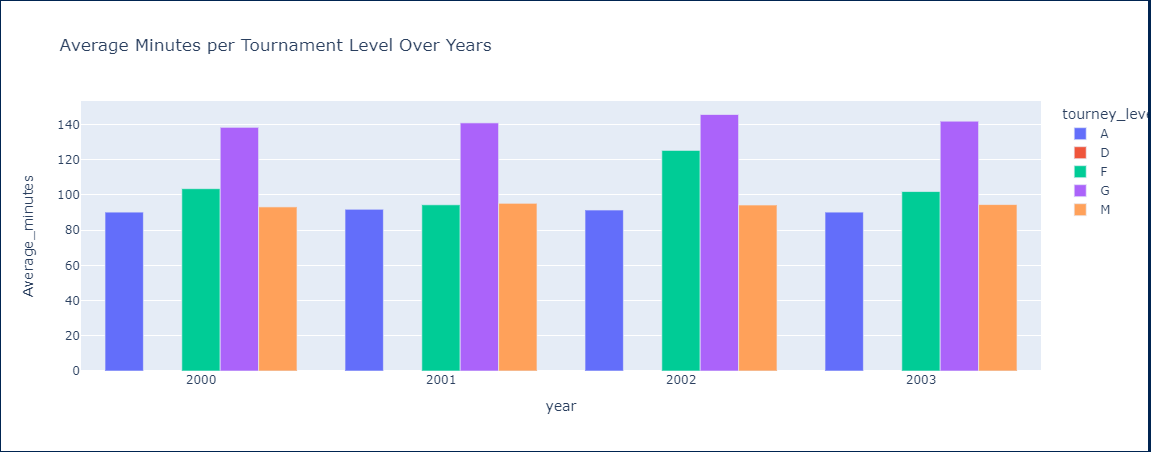


Figure 7 . Average Minutes per Tournament Level for 2000-2003

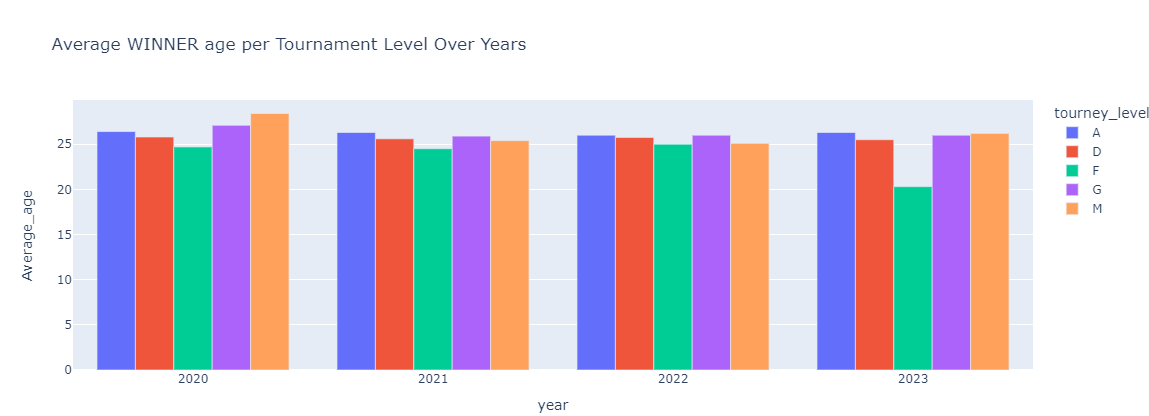
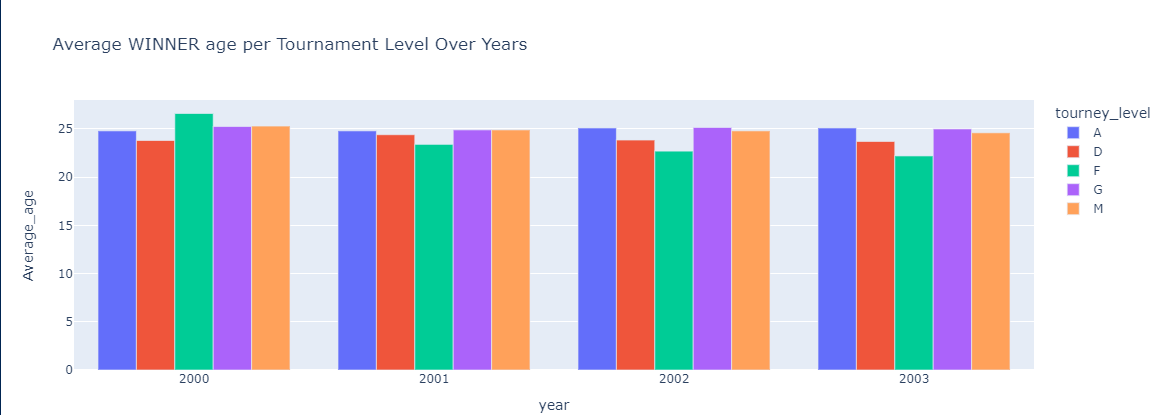


Figure 9 Average winner age per tournament level for years 2020-2023

Figure 8 Average winner age for each tournament for years 2000-2003



## Winner Rank distribution over Winner Height Distribution:

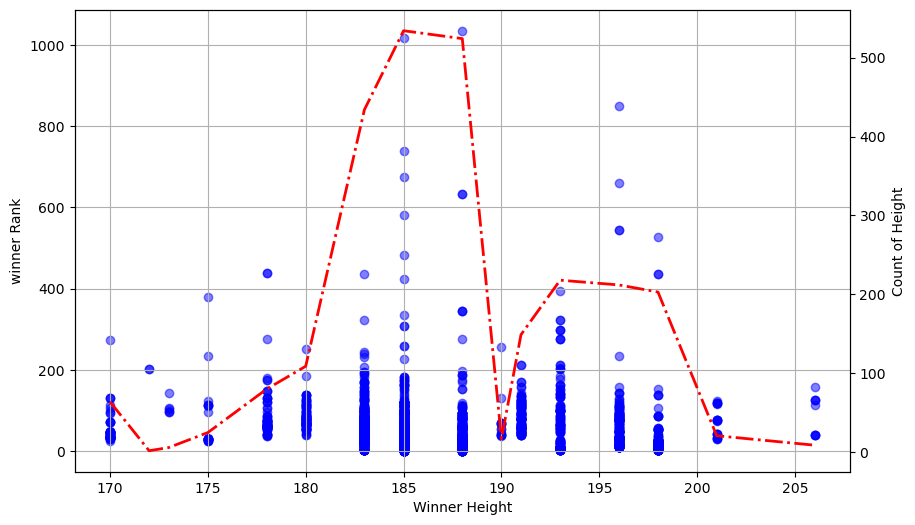


Figure 10 Winner Rank vs Height Distribution for players in 2023

The provided visualization presents a comprehensive exploration of the relationship between winner height and winner rank in a tennis dataset. The following EDA techniques have been utilized:

* **Scatter Plot:** A scatter plot was employed to visualize the relationship between winner height and winner rank. This technique allows for a direct comparison of the two variables, revealing any potential trends or correlations.
* **Histogram:** A histogram was incorporated on a secondary y-axis to depict the distribution of winner heights. This provides insights into the frequency of different height ranges among the winners.
* **Overlayed Density Plot:** A density plot was superimposed on the histogram to offer a smoother representation of the height distribution. This technique helps in identifying the underlying probability density function of the data.

## Age vs Rank Correlation:

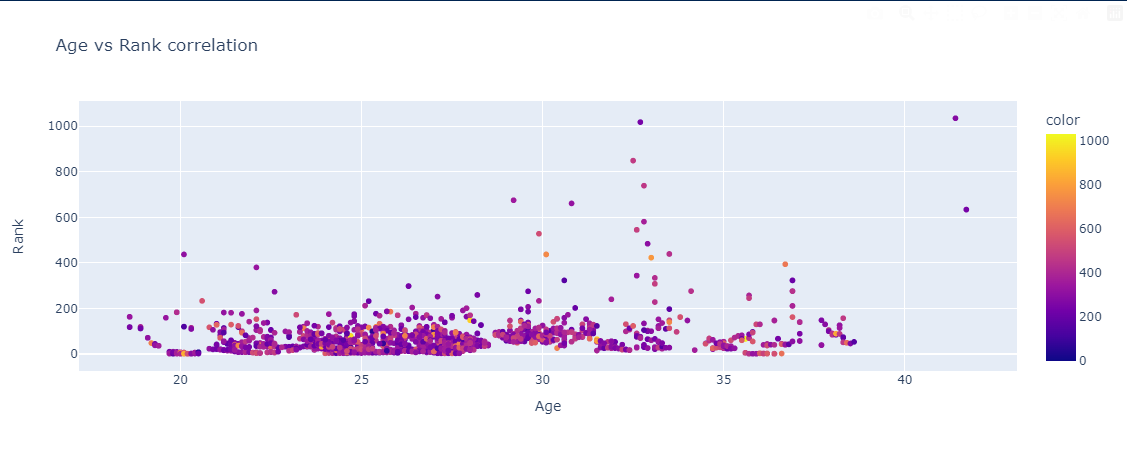


Figure 11 Age vs Rank Correlation for year 2023

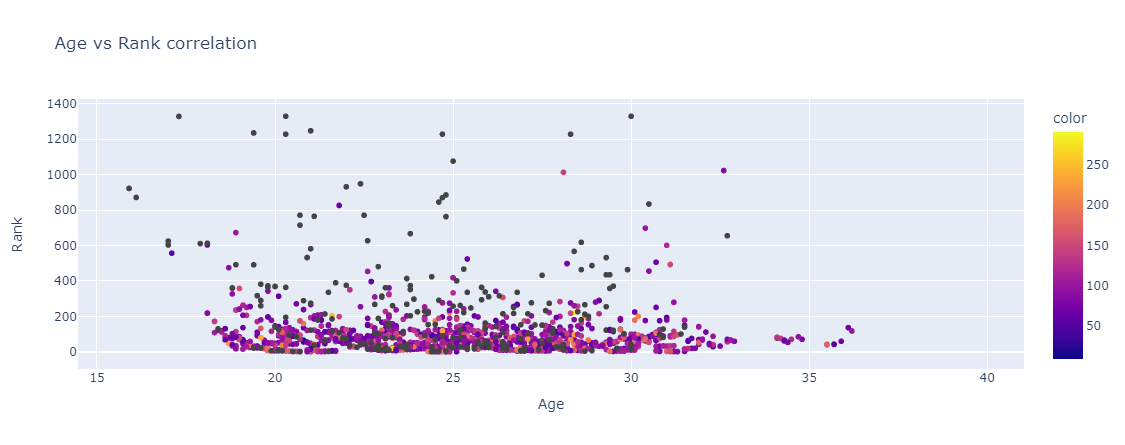


Figure 12 Age vs Rank Correlation for year 2001

This visualization employs a scatter plot to explore the relationship between winner age and winner rank in a tennis dataset, along with an additional dimension represented by color. The code utilizes Plotly Express (pt.scatter), a high-level library for creating interactive visualizations in Python.

* Data Source: The data for the scatter plot is extracted from two columns in the atp DataFrame: winner\_age (winner's age) and winner\_rank (winner's rank).
* Color Dimension: An additional data point, minutes (presumably representing match duration), is used to color-code the data points in the scatter plot. This allows for a preliminary exploration of how match duration might relate to both winner age and rank.
* Axis Labels and Title: Descriptive labels are provided for the x-axis ('Age'), y-axis ('Rank'), and the overall plot title ('Age vs Rank correlation').

By using a scatter plot with color-coding, this visualization offers a multifaceted exploration of the data, aiding in the initial investigation of potential relationships between winner age, winner rank, and match duration.

## Top 30 players for year 2023 based on Win Ratios:

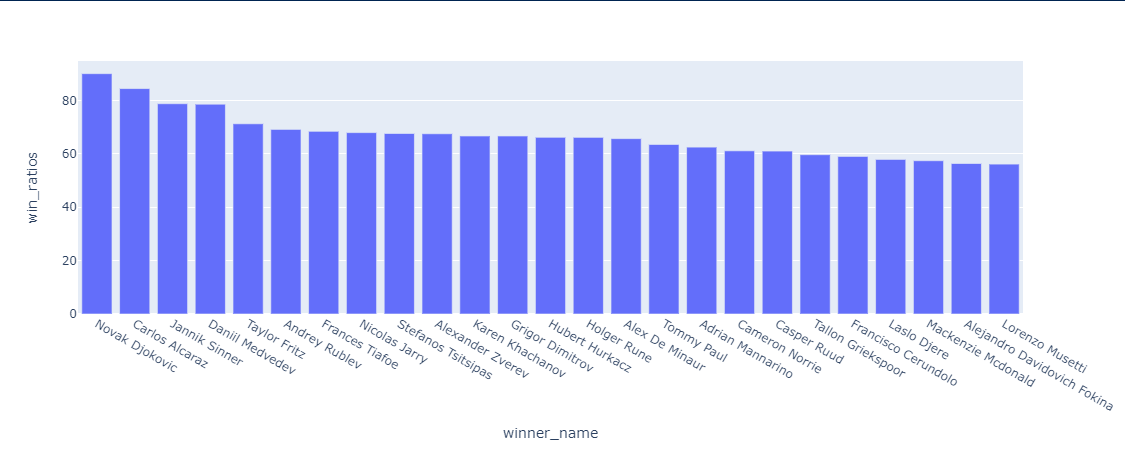


Figure 13 Bar Chart representation of top 30 players according to win ratios

The provided visualization presents a bar chart illustrating the win ratios of various tennis players for year 2023. The following EDA techniques have been utilized:

* Bar Chart: A bar chart was chosen to visually represent the win ratios of each player. This format allows for easy comparison and identification of the highest and lowest performers.
* Sorting: The data was sorted in descending order of win ratios, ensuring that the players are ranked from highest to lowest win percentages. This facilitates a clear understanding of the performance hierarchy.
* Data Normalization: The decision to create a separate DataFrame with win ratios as a column suggests that the data underwent normalization. This process likely standardized the win ratios, ensuring that they are comparable across different players or periods, potentially accounting for variations in playing frequency or opponent strength.

## Visualizing winning pattern for top 30 players in a tennis match

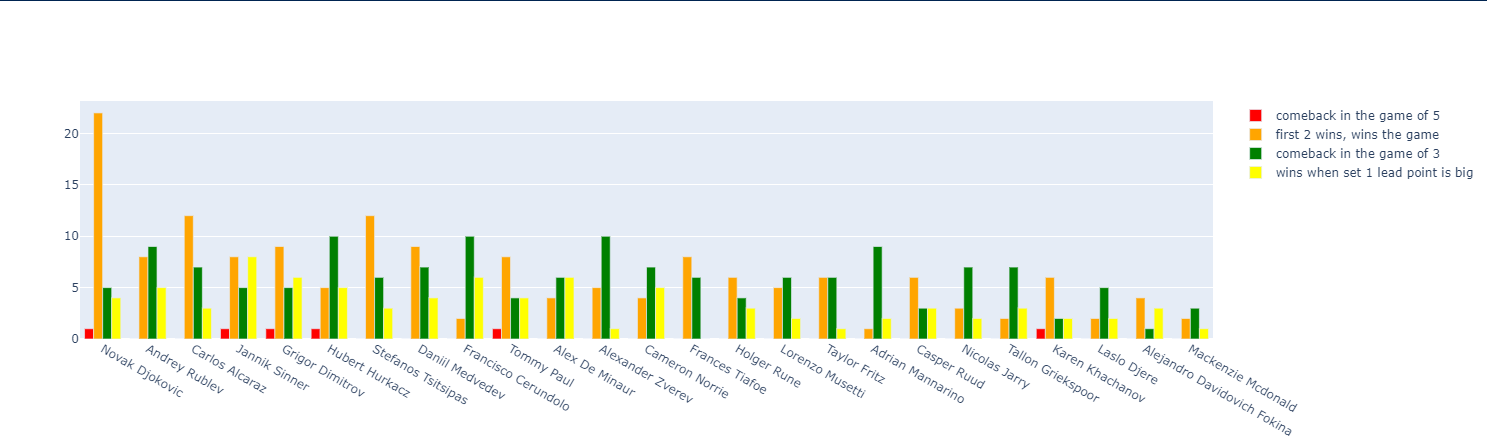


Figure 14 Winning trend for top 30 players

In this section, a visualization of the winning patterns of various tennis players has been presented using a grouped bar plot. The objective was to categorize and illustrate the different methods through which players secure victory in a match. Four main winning patterns or labels were identified:

* First 2 Wins, Wins the Game (Yellow): This label represents scenarios where a player wins the first two sets and subsequently wins the match.
* Wins When Set 1 Lead Point is Big (Orange): This label captures matches where a significant lead in the first set contributes to an overall match victory.
* Comeback in the Game of 5 (Red): This category indicates matches where a player makes a comeback in the fifth set to win the match.
* Comeback in the Game of 3 (Green): This label represents matches where a player recovers in the third set and goes on to win the match.

To construct this visual representation, a dataframe was first created, containing the match outcomes categorized under the aforementioned labels. The data was then pivoted into a format suitable for creating a grouped bar plot. The x-axis of the plot lists the players, and under each player’s name, four bars are displayed, representing the number of matches won under each of the four winning patterns.

This plot provides a clear comparison of the different winning patterns across various players, offering insights into their match-winning tendencies based on these predefined categories

## Distribution of Wins Across Different Tennis Surfaces

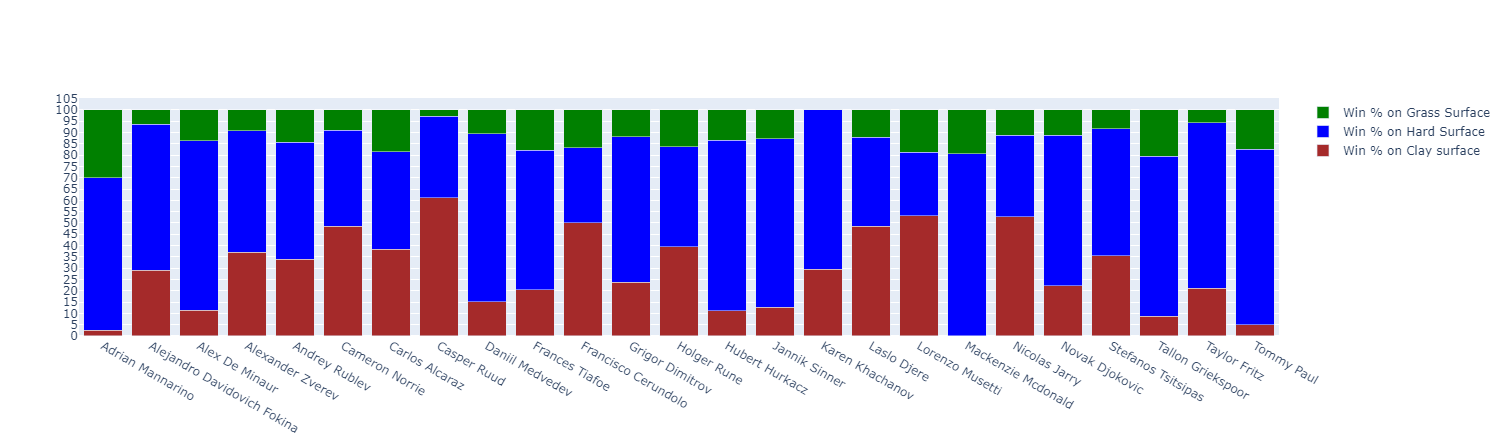


Figure 15 100% stacked bar viz for top 25 players across different surfaces

Each bar in the plot corresponds to a player, with the stacked segments indicating the relative success across the three surfaces. The use of a 100% stacked bar plot ensures that the total height of each bar is consistent, facilitating an easy comparison of surface performance distributions among the players.

The above representation utilizes a 100% stacked bar plot to analyze and understand the distribution of wins by the top 25 tennis players across different types of surfaces. The aim is to visualize how player success varies depending on the surface—whether it be grass, hard, or clay courts.

The win percentages on each surface are categorized into three segments:

1. **Win % on Grass Surface (Green)**: This segment shows the proportion of wins each player has achieved on grass courts.
2. **Win % on Hard Surface (Blue)**: This segment represents the proportion of wins on hard courts, reflecting player performance on this common surface type.
3. **Win % on Clay Surface (Red)**: This segment illustrates the percentage of wins on clay courts, a surface known for its unique playing conditions.

Each bar in the plot corresponds to a player, with the stacked segments indicating the relative success across the three surfaces. The use of a 100% stacked bar plot ensures that the total height of each bar is consistent, facilitating an easy comparison of surface performance distributions among the players.

## Boxplot representations of derived metrics

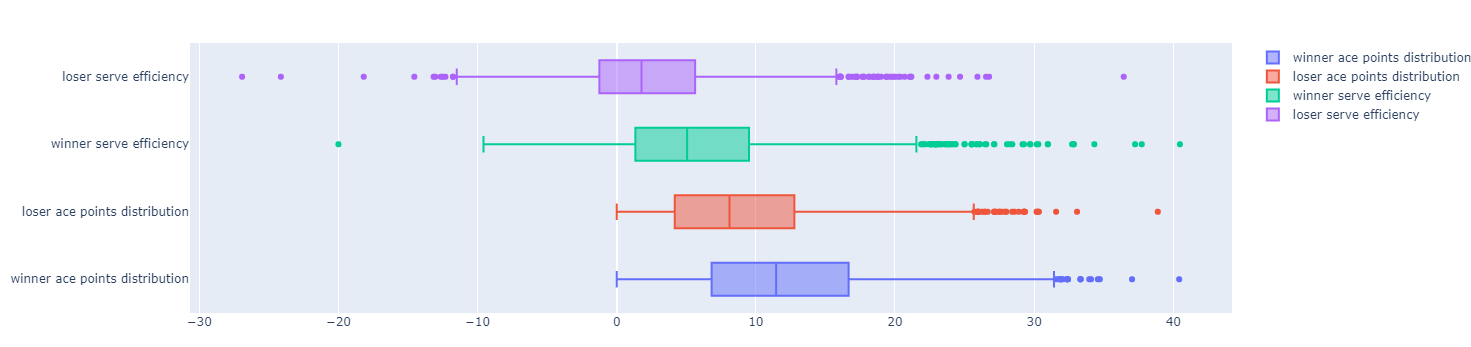


Figure 16: Serve efficiency and Ace% Boxplot viz for winner and loser for year 2023

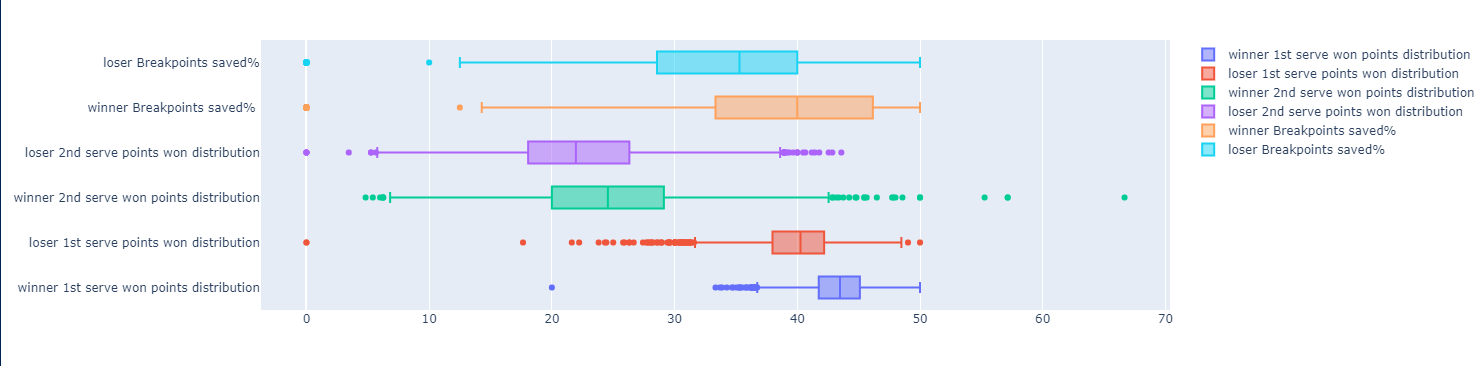


Figure 17 : Boxplot viz : Breakpoints Saved, & 1st serve point won% for winner and loser in 2023

## Statistical Tests for Analysing Tennis Surface Effects

In this section, a series of statistical tests were used to examine the effect of various tennis surfaces (Clay, Grass, and Hard) on specific player performance metrics, specifically the percentage of break points saved (bp\_saved%) and the percentage of aces served (ace%). The following statistical methods were used:

### One-Way Analysis of Variance (ANOVA)

The One-Way ANOVA test was used to see if there were any statistically significant differences in the means of three or more independent groups. In this study, the means of bp\_saved% and ace% were compared across different surfaces (clay, grass, and hard).

* **Application** : The One-Way ANOVA is used to compare the means of three or more groups and determine whether at least one group's mean differs from the others.
* **Purpose** : The f\_oneway function from the scipy.stats module was used to carry out this test. The surface type was the independent variable, while bp\_saved% and ace% were the dependent variables.
* **Procedure:** Separate groups were created for each surface (Clay, Grass, and Hard) using the corresponding columns in the dataset.The ANOVA test was applied to compare the means of these groups.The output includes an F-value, which indicates the ratio of variance between the group means to the variance within the groups, and a p-value, which assesses the significance of this ratio.

### Tukey's Honest Significant Difference (HSD) Test

Following the ANOVA test, Tukey's HSD test was employed to conduct pairwise comparisons between the groups. This test is used to identify which specific group means are significantly different from each other.

* **Purpose:** Tukey's HSD test is a post-hoc analysis used when an ANOVA test shows significant results, allowing for multiple comparisons between group means while controlling for Type I error (false positives).
* **Application:** The pairwise\_tukeyhsd function from the statsmodels.stats.multicomp module was used to perform this test.
* **Procedure :** The test was applied to the ace% and bp\_saved% metrics across the different surfaces. It provided pairwise comparisons between the surfaces, indicating the mean difference between groups, the confidence intervals, and whether the difference is statistically significant (as indicated by the reject column).

These statistical methods were chosen for their ability to analyze and compare the performance metrics of tennis players across different surfaces, allowing for a rigorous examination of whether and how surface type influences performance outcomes. The results of these tests will be discussed in the subsequent sections on inference and conclusion, where the implications of the statistical findings will be explored in detail.

## Analysis of Variance (ANOVA) for Break Points Saved

This section describes the use of Analysis of Variance (ANOVA) to examine the effects of surface type and era on the percentage of break points saved (bp\_saved\_pct) by tennis players. The analysis was conducted using different models to explore both the main effects and interaction effects of these factors.

### Data Preparation

Before conducting the ANOVA tests, certain columns in the dataset were renamed to maintain clarity and consistency:

1. The column serve\_efficiency% was renamed to serve\_efficiency\_pct.
2. The column ace% was renamed to ace\_pct.
3. The column bp\_saved% was renamed to bp\_saved\_pct.

These changes ensure that the column names are consistent and appropriate for statistical modeling.

### ANOVA Model for Main Effects

The first ANOVA model was designed to analyze the main effects of two categorical variables—surface type (surface) and era (era)—on the dependent variable, bp\_saved\_pct.

1. **Model Specification:** An ordinary least squares (OLS) regression model was fit using the formula:
2. **Purpose:** This model assesses the individual contributions of surface type and era to variations in the percentage of break points saved by players.
3. **ANOVA Table:** The ANOVA table was generated to examine the statistical significance of each factor. The typ=2 argument was used to produce a Type II ANOVA, which tests each main effect after accounting for the others.

### ANOVA Model with Interaction Term

To explore potential interaction effects between surface type and era, a second ANOVA model was implemented with an interaction term:

1. **Model Specification:** The formula for this model included the interaction between surface and era:
2. **Purpose:** This model evaluates whether the effect of surface type on bp\_saved\_pct varies across different eras. It helps in understanding if the relationship between surface and break points saved is consistent over time or if it has changed across different tennis eras.
3. **ANOVA Table:** Again, a Type II ANOVA was conducted to determine the significance of both the main effects and the interaction term.

### Subset Analysis by Era

To gain deeper insights, the data was further divided into two subsets representing distinct eras: 2000-2003 and 2020-2023. Separate ANOVA tests were then performed for each era to assess the effect of surface type on bp\_saved\_pct within these time periods.

1. **Data Subsetting:** The dataset was split into two subsets:
   1. data\_2000\_2003: Contains data for the era 2000-2003.
   2. data\_2020\_2023: Contains data for the era 2020-2023.
2. **Model Specification:** For each subset, an OLS regression model was fit using the formula:
3. **Purpose:** These models were intended to compare the impact of surface type on break points saved within each era, providing a clearer understanding of how this relationship might have evolved over time.
4. **ANOVA Tables:** Separate ANOVA tables were generated for each era to evaluate the significance of surface type in each specific context.

This comprehensive approach using ANOVA allows for a nuanced analysis of how surface type and era influence player performance in terms of saving break points. The findings from these tests will be discussed in the inference and conclusion sections of the dissertation.

## Implementation of Hidden Markov Models (HMM) for Analyzing Tennis Matches

In this section, a Hidden Markov Model (HMM) was implemented to analyze the states during tennis matches for both winners and losers. The goal was to uncover latent patterns and dynamics in player performance across different match conditions.

### Data Preparation

To apply the HMM, specific match-related features were first extracted and organized into a suitable format . Feature Selection: The dataset was filtered to include relevant columns that capture critical aspects of match performance:

* match\_id, ply\_id, and results for identifying the match and player.
* bp\_saved%: The percentage of break points saved by the player.
* ace%: The percentage of aces served by the player.
* serve\_efficiency%: An overall metric of serving efficiency.
* setX\_ot\_diff (where X = 1 to 5): The overtime point differential for each set.
* setX\_gm\_diff (where X = 1 to 5): The game differential for each set.

### Implementation of Hidden Markov Model (HMM)

The Hidden Markov Model was applied to the processed data to identify hidden states within the match dynamics:

**Model Selection:** A Gaussian HMM was chosen for the analysis, given its suitability for continuous data and its ability to model the probabilistic transitions between states. The model parameters were set as follows:

* n\_components=3: This specifies that the model will identify three hidden states.
* covariance\_type='diag': Indicates that diagonal covariance matrices are used, assuming that the features are independent given the hidden state.
* n\_iter=1000: The model is iteratively trained to convergence.

**Model Fitting:**

The features used for fitting the model were:

* ace%: Percentage of aces served.
* bp\_saved%: Percentage of break points saved.
* serve\_efficiency%: Serve efficiency percentage.
* set\_point\_diff: Game differential across sets.
* overtime\_set\_point\_diff: Overtime point differential across sets.

**State Prediction:**

* The predict function of the trained HMM was used to assign each observation in the dataset to one of the hidden states.
* These predicted states were then added to the dataframe under the column hidden\_markov\_states.

This approach allows for the analysis of match data through the lens of probabilistic states, offering insights into the underlying structure of player performance during a match. The detailed results and interpretations of these hidden states will be discussed in the subsequent sections dedicated to inference and conclusions.

# References

Andrej Panjan, N. S. A. F., June 2010. Prediction of the successfulness of tennis players with machine learning methods. *Kinesiology.*

Drive-analytics, n.d. *The Gradual yet steady rise of Tennis Analytics.* [Online]   
Available at: https://www.drive-analytics.com/tennis.php

Havlucu Hayati, A. B. ,. E. T. ,. C. A. ,. O. O., 2022. Toward Detecting the Zone of Elite Tennis Players Through Wearable Technology. *Frontiers in Sports and Active Living,* Volume 4.

Hyatt-Twynam, S., 2013. *The logit function and its applications in sports modelling : Sports Trading Network.* [Online]   
Available at: https://www.sportstradingnetwork.com/article/logit-function-applications-sports-modelling/

Ilias Kolonias, W. J. C. J. K., August 2004. Tracking the Evolution of a Tennis Match Using Hidden Markov Models. *Lecture Notes in Computer Science.*

JeffSackmann, 2016. *Github JeffSackmann.* [Online]   
Available at: https://github.com/JeffSackmann/tennis\_atp

Manuel, J., March 4 2022. *Capturing Momentum in Tennis.* [Online]   
Available at: https://theanalyst.com/eu/2022/03/capturing-momentum-in-tennis

Rosker Jernej, M. R. Z., 2021. Skill Level in Tennis Serve Return Is Related to Adaptability in Visual Search Behavior. *Frontiers in Psychology.*

Srivastava, S., 2023. *appinventiv.* [Online]   
Available at: https://appinventiv.com/blog/data-analytics-in-sports-industry/

Tatiana Sampaio, J. P. O. A. M. P. N. E. M., June, 2024. Applications of Machine Learning to Optimize Tennis Performance: A Systematic Review. *Applied Sciences.*

Wu, W., May 2024. Quantification of momentum in tennis matches and its impact: a study based on AHP-EWM method and data analysis. *Highlights in Science Engineering and Technology.*

Yang, J. L. Z., January 2023. The Biomechanical Analysis on the Tennis Batting Angle Selection Under Deep Learning. *IEEE Access,* Volume 11.

Yixiong Cui, H. L.-Á. G., 2019. Data-driven analysis of point-by-point performance for male tennis player in Grand Slams. *revistas,* Volume 15.

Yu Zhang, 2023. The optimization of college tennis training and teaching under deep learning,. *Heliyon,* 10(4).