

Введение

Теннис — это один из самых популярных видов спорта в мире, который привлекает миллионы зрителей и участников. Профессиональные теннисные турниры проводятся на различных уровнях, начиная от местных соревнований и заканчивая международными чемпионатами, такими как турниры Большого шлема. Наиболее известные турниры включают Уимблдон, Открытый чемпионат Австралии, Открытый чемпионат Франции и Открытый чемпионат США.

Правила игры

Игроки должны отправить мяч через сетку на сторону соперника так, чтобы тот не смог его вернуть. Очки зарабатываются, когда соперник не может отбить мяч в пределах корта.

Игра ведется до определенного количества сетов. Каждый сет состоит из геймов, а геймы — из очков. Счет в теннисе имеет уникальную систему: 0 (ноль), 15, 30, 40 и «игра». Если счет достигает 40–40, это называется «ровно», и игрок должен выиграть два очка подряд для выигрыша гейма.

Матч может состоять из 2 или 3 сетов (обычно 3 сета для мужчин на турнирах Большого шлема и 2 сета для женщин). Игрок, который первым выигрывает установленное количество сетов, становится победителем матча.

Игроки по очереди подают мяч. Подача должна быть выполнена из-за задней линии и направлена в диагональную половину корта соперника. Подача считается ошибочной, если мяч попадает в сетку или выходит за пределы корта.

Описание данных

Данная база данных охватывает информацию о теннисных игроках АТР¹, их исторических рейтингах, результатах матчей и статистике. Она включает несколько файлов, каждый из которых содержит уникальные данные о турнирах, матчах и игроках. В базе представлены результаты матчей за 2022–2024 гг.

1 Ассоциация теннисистов-профессионалов (англ. Association of Tennis Professionals, ATP) создана в 1972 году для защиты и представительства интересов профессиональных игроков в теннис.

Описание переменных

Название переменной	Описание	Единицы измерения
tourney_id	Уникальный идентификатор турнира	—
tourney_name	Название турнира	—
surface	Поверхность корта Включает 4 категории: 'Hard', 'Clay', 'Grass', 'Carpet'	—
draw_size	Количество игроков в сетке, часто округленное до ближайшей степени 2 (Например, турнир с 28 игроками может быть показан как 32)	—
tourney_level	<p>Для мужчин:</p> <p>'G' = Большие шлемы</p> <p>'M' = Мастерс 1000,</p> <p>'A' = другие турниры тура</p> <p>'C' = Челленджеры</p> <p>'S' = Сателлиты/ITF, 'F' = Финалы тура и другие сезонные мероприятия, и 'D' = Кубок Дэвиса</p> <p>Для женщин есть несколько дополнительных кодов:</p> <p>'P' = Премьер</p> <p>'PM' = Обязательные Премьер</p> <p>'I' = Международные</p> <p>Различные уровни ITF обозначаются призовыми деньгами (в тысячах), например, '15' = ITF \$15,000. Другие коды, такие как 'T1' для Tier I (и так далее) используются для более старых обозначений турниров WTA. 'D' используется для Кубка Федерации/Билли Джин Кинг, а также для Кубка Уайтмана и Кубка Бонн Белл.</p> <p>Другие, в конечном итоге для обоих полов: 'E' = выставочные (мероприятия, не санкционированные туром, хотя определения могут быть неоднозначными), 'J' = юниоры, и 'T' = командный теннис, который пока не появляется нигде в наборе данных, но появится в какой-то момент.</p>	—
tourney_date	Восемь цифр, ГГГГММДД, обычно понедельник недели турнира.	—
match_num	Идентификатор, специфичный для матча	—
winner_id	Идентификатор победителя	—

Название переменной	Описание	Единицы измерения
winner_seed	Посев победителя ²	—
winner_entry	<p>В теннисе этот термин обычно используется для обозначения игрока, который получил прямое попадание в основную сетку турнира без необходимости проходить квалификацию. Существует несколько способов получить такое прямое попадание:</p> <ul style="list-style-type: none"> • WC (Wild Card) — специальное приглашение от организаторов турнира • Q (Квалификация) — игрок успешно прошел квалификационный турнир и получил место в основной сетке • LL (Lucky Loser) — игрок, проигравший в финале квалификации, но получивший место в основной сетке из-за снятия другого игрока • PR (Protected Ranking) — игрок использует защищенный рейтинг, позволяющий получить место в основной сетке при определенных условиях • ITF — игрок получает место в основной сетке через турниры ITF (Международной федерации тенниса) 	
winner_name	Имя победителя	—
winner_hand	<p>R = правша</p> <p>L = левша</p> <p>U = неизвестно</p> <p>(Для игроков-амбидекстров это их подающая рука)</p>	—
winner_ht	Рост победителя в сантиметрах	см
winner_ioc	Трехбуквенный код страны победителя	—
winner_age	Возраст победителя в годах (на момент tourney_date)	Лет

² Посев (сеяние) — назначение участникам или командам, принимающим участие в спортивном турнире, номера в предварительном рейтинге в рамках жеребьевки. Обычно посев предназначен для разведения наиболее сильных соперников в сетке соревнований «подальше», чтобы они не могли проиграть друг другу на ранних этапах соревнования. Этот термин впервые был использован в теннисе и основан на идее выкладки турнирной лестницы путём размещения листков бумаги с именами игроков на них, наподобие того, как семена или саженцы расположены в саду: маленькие растения впереди, большие позади.

В теннисе и настольном теннисе

В крупных турнирах по теннису и настольному теннису обычно в сетку соревнований помещают 8, 16 или 32 так называемых сеянных спортсмена или командные пары с наиболее высокими рейтингами таким образом, чтобы создать подсетки с одинаковым количеством соперников, в каждой из которых есть в точности по одному сеянному.

Название переменной	Описание	Единицы измерения
loser_id	Идентификатор проигравшего	—
loser_seed	Посев проигравшего	—
loser_entry	'WC' = wild card 'Q' = квалификация 'LL' = lucky loser 'PR' = защищенный рейтинг 'ITF' = ITF entry, и есть несколько других, которые иногда используются.	—
loser_name	Имя проигравшего	—
loser_hand	R = правша, L = левша, U = неизвестно. Для игроков-амбидекстров это их подающая рука.	—
loser_ht	Рост проигравшего в сантиметрах	см
loser_ioc	Трехбуквенный код страны	—
loser_age	Возраст в годах на дату tourney_date	Лет
score	Счет матча	—
best_of	'3' или '5', указывающие на количество сетов для этого матча	Сеты
round	Раунд турнира	—
minutes	Длительность матча	Минуты
w_ace	Количество эйсов у победителя	Эйсы ³
w_df	Количество двойных ошибок у победителя	Ошибки
w_svpt	Количество розыгрышей подачи у победителя	Розыгрыши
w_1stIn	Количество первых подач в игре у победителя	Подачи
w_1stWon	Количество очков, выигранных на первой подаче у победителя	Очки
w_2ndWon	Количество очков, выигранных на второй подаче у победителя	Очки
w_SvGms	Количество геймов на своей подаче, выигранных победителем	Геймы
w_bpSaved	Количество брейк-пойнтов, спасенных победителем	Брейк-пойнты
w_bpFaced	Количество брейк-пойнтов, сыгранных победителем	Брейк-пойнты
l_ace	Количество эйсов у проигравшего	Эйсы
l_df	Количество двойных ошибок у проигравшего	Ошибки

³ Эйс в теннисе — это подача, которая завершается тем, что мяч попадает в площадку соперника, и принимающий игрок не успевает коснуться его ракеткой. Это означает, что подающий зарабатывает очко без необходимости продолжать розыгрыш. Эйсы обычно выполняются мощными и точными подачами, которые затрудняют ответный удар соперника

Название переменной	Описание	Единицы измерения
l_svpt	Количество розыгрышей подачи у проигравшего	Розыгрыши
l_1stIn	Количество первых подач в игре у проигравшего	Подачи
l_1stWon	Количество очков, выигранных на первой подаче у проигравшего	Очки
l_2ndWon	Количество очков, выигранных на второй подаче у проигравшего	Очки
l_SvGms	Количество геймов на своей подаче, выигранных проигравшим	Геймы
l_bpSaved	Количество брейк-пойнтов, спасенных проигравшим	Брейк-пойнты
l_bpFaced	Количество брейк-пойнтов, сыгранных проигравшим	Брейк-пойнты
winner_rank	Рейтинг ATP или WTA победителя на дату tourney_date или на последнюю дату рейтинга перед tourney_date	Рейтинг
winner_rank_points	Количество рейтинговых очков, где доступно	Очки
loser_rank	Рейтинг проигравшего	Рейтинг
loser_rank_points	Количество рейтинговых очков проигравшего	Очки

Задание

Используя имеющиеся данные, вам необходимо ответить на вопрос:

Что влияет на победу игрока?

Вы можете конкретизировать поставленный вопрос. Например:

Как уровень турнира (например, Grand Slam, ATP 1000, ATP 500) влияет на результаты матчей и поведение игроков?

Как успехи на ранних этапах турнира игроков (например, победы/поражения в предыдущих матчах) влияют на их результаты в следующих матчах?

Как конкуренция между игроками (например, между игроками одной страны или одной школы) влияет на их число побед?

Рекомендации по проведению исследования

Ниже представлен примерный план для проведения исследования. Вы можете использовать этот план в качестве ориентира, а также адаптировать его под свои нужды. Обратите внимание на критерии оценки, чтобы максимально повысить свои шансы на успешное выполнение работы.

1. Предварительный анализ

Опишите представленные данные о теннисных турнирах: сколько наблюдений содержится в базе, как они распределены по различным категориям (например, по уровням турниров, поверхностям, странам и т.д.). Обратите внимание на особенности данных, такие как наличие выбросов, асимметричность распределений, пропуски и неравномерность по подгруппам. Какие зависимости и паттерны вы наблюдаете? Например, можно проанализировать, как уровень турнира влияет на количество эйсов или длительность матчей.

2. Постановка гипотезы

На основе выявленных зависимостей сформулируйте гипотезу, которую можно проверить с использованием предоставленных данных. Подумайте о том, какие аспекты вас интересуют: например, может ли опыт игрока (возраст, рейтинг) влиять на количество выигранных матчей на определенных покрытиях. Обоснуйте, почему эта гипотеза важна для анализа и какие модели или теории могут служить ее основой.

3. Выбор переменных

Выберите факторы из базы данных, которые, по вашему мнению, могут оказывать влияние на результаты матчей. Это могут быть такие переменные, как рейтинг игрока, возраст, высота, тип покрытия и т.д. Сформулируйте гипотезу о том, как эти факторы могут быть связаны с результатами матчей (например, более высокий рейтинг может положительно сказываться на вероятности победы).

4. Разведывательный анализ

Используйте графики для визуализации распределения выбранных переменных. Рассчитайте основные описательные статистики и проанализируйте их особенности. Каковы форма распределений (скошенность, симметричность) и вариативность? Обратите внимание на наличие выбросов и пропусков, а также на неоднородность данных в разных категориях.

5. Проверка гипотезы на данных

Примените математические методы и статистику для проверки вашей гипотезы. Начните с визуального анализа и создайте графики, иллюстрирующие взаимосвязь между переменными. Убедитесь, что выбранные визуализации четко демонстрируют наличие или отсутствие связи.

6. Устойчивость выводов

Проверьте, насколько устойчивы ваши выводы о выполнении или невыполнении гипотезы. Сравните результаты по различным подвыборкам данных. Если результаты различаются, предложите возможные объяснения.

7. Выводы и ограничения

Обсудите, как ваши результаты могут быть применены на практике. Какие рекомендации вы можете предложить для игроков или тренеров, чтобы повысить вероятность победы в матчах? Какие важные знания были получены в ходе исследования?

Также рассмотрите ограничения вашего исследования: какие факторы могли быть упущены, и как это может повлиять на результаты? Укажите, какие данные могли бы улучшить уверенность в выводах.

8. Применение результатов

Опишите, для каких групп людей могут быть полезны полученные результаты. Это могут быть игроки, тренеры, спортивные аналитики или любители тенниса. Подумайте, можно ли расширить применение результатов на другие виды спорта? На какие конкретно и почему?

Удачи в подготовке исследования! Помните, что идеальных исследований не бывает, используйте предоставленные данные по максимуму для получения нетривиальных выводов, но не забывайте про рамки, в которых эти выводы могут быть применимы.