

ANÁLISIS DE DATOS DEPORTIVOS APLICADO AL TENIS CON R

Libia Johana Sánchez Espinoza

1.	Introducción.....	2
2.	Problemática	2
3.	Caso concreto a analizar.....	3
4.	Solución	4
	¿Lenguaje de programación a utilizar?	4
	¿Librería a utilizar para el tratamiento de los datos del TENIS?	5
	¿Fuente de datos del Tenis?.....	6
	Esquema de implementación de las herramientas seleccionadas	7
	Implementación de la solución en R.....	8
5.	Conclusión	12
6.	Bibliografía en estilo IEEE	13

1. Introducción

El presente trabajo muestra un storytelling, del análisis de datos del tenis regulado por la ATP Asociación de Tenistas Profesionales (hombres) y WTA Asociación de Tenis Femenino (mujeres) [1].

Se analizarán estadísticas de rankings, jugadores y partidos de la ATP y la WTA. Este análisis se realizará por separado distinguiendo el género ya que la información se almacena en dataframes diferentes según la organización que los regula.

Con este análisis se pretende respaldar la afirmación de la importancia de la edad con la que se inicia en el entrenamiento del tenis, en especial si se tiene aspiraciones de llegar a competir en torneos oficiales internacionales.

Este análisis se realizará implementando la librería "deuce" en lenguaje R de acceso público compartido por Stephanie Kovalchik (Senior Data Scientist Australiana) en la plataforma GitHub, con licencia GPL-2 permitiendo su estudio, modificación y redistribución [2].

Esta librería contiene funciones que permiten obtener datos del tenis de fuentes de Jeff Sackmann (Desarrollador de software e investigador del tenis) que ha recopilado información desde 1985, se encuentra publicado en la plataforma GitHub con licencia CC BY-NC-SA 4.0 DEED [3].

Tanto la librería como el repositorio requieren del reconocimiento respectivo de su autoría. Y el repositorio solo está disponible para uso académico no comercial.

Este storytelling se utilizará por entrenadores de escuelas de tenis para respaldar la afirmación de que si un jugador tiene una formación más temprana tiene más posibilidades de llegar a jugar campeonatos a nivel profesional.

2. Problemática

En Ecuador existen muchos Clubes privados que tienen escuelas de tenis con canchas muy bien equipadas con entrenadores de gran trayectoria en el deporte, incluso algunos han sido campeones sudamericanos. El problema es que los alumnos por lo general hijos de los socios se interesan en el tenis, pasados los 13 años de edad y desean llegar a competir en torneos profesionales en la ATP.

Los entrenadores han explicado a los padres de familia la importancia de la formación de los jugadores de tenis a una edad temprana, de acuerdo a la experiencia que tienen, pero no poseen los recursos para respaldar esa información.

Los niños requieren habilidades especiales para el tenis, desde como sostener la raqueta, coordinar movimientos, velocidad, equilibrio, resistencia y destreza a la hora de mover la raqueta. Esas son las principales habilidades que necesitan desarrollar los niños para jugar correctamente, es por eso que la edad de inicio de entrenamiento es importante [5].

Se conoce que detrás de los talentosos y reconocidos tenistas del mundo están las academias que los forman y potencian sus habilidades. Estas academias se esfuerzan por entrenar a los mejores jugadores en diferentes niveles. Cada Academia se enfoca con una filosofía de entrenamiento diferente, pero concluyen en el objetivo común de fomentar el talento y aportar a la próxima generación de campeones del tenis [4]. Con el fin de lograr este último objetivo las

academias requieren tener un amplio número de jugadores que entrenen desde temprana edad y participen en campeonatos locales desde los niveles bajos y así se perfilen a campeonatos internacionales.



Imagen 1. Tennis Quito es una academia de tenis que recibe niños desde los 3 años [5]

3. Caso concreto a analizar

Para respaldar la afirmación de la importancia de la edad de inicio de entrenamiento en el tenis se requiere:

Realizar un análisis de los datos del tenis donde se identifique a qué edad promedio jugaron su primer partido los tenistas que tuvieron una posición de 1 al 50 en el ranking de la ATP, del torneo de la última semana jugada y registrada en la base de datos del año 2023 (De esta forma aseguramos que cuando se actualicen los datos en la base de datos va a tomar la última semana de torneo del año seleccionado). Se requiere también que se obtenga la edad promedio de todos los jugadores que ganan los partidos. Al obtener la edad promedio, es necesario contar cuantos partidos ganados tienen antes y después de esa edad, los jugadores seleccionados anteriormente (50 primeros tenistas en el ranking de la ATP del año 2023 del último torneo registrado en la base de datos). Como conclusión es necesario comparar estos valores para conocer si los jugadores seleccionados anteriormente ganan más partidos antes o después de la edad promedio.

Debido a que el tenis femenino está regulado por la WTA (Asociación de tenis femenino) y se encuentra en dataframes diferentes es necesario realizar el proceso anterior para obtener también los resultados de las jugadoras y el archivo tennisATP.R es una excelente guía para obtener los resultados del tenis femenino.

La aspiración de muchos tenistas, padres de familia, entrenadores y academias es que el deportista logre entrar al ranking de la ATP o WTA, para esto es preciso entender cómo se consideran las posiciones de los jugadores en el ranking del tenis, y se tome como ejemplo los mejores jugadores del ranking 2023.

¿Como se considera el ranking de jugadores en la ATP y WTA?

Los jugadores están ordenados por un tipo de clasificación denominada Entry Ranking, en la cual se suman los puntos de los jugadores de los partidos de las 52 últimas semanas, que es aproximadamente un año. Al finalizar cada semana, se le resta a cada jugador los puntos obtenidos en esa misma semana del año anterior y se suman los puntos de los partidos ganados en la semana en curso. El ranking ATP es la clasificación para tenistas profesionales masculinos categoría individual y dobles desde el año 1973. El ranking WTA es la clasificación para tenistas profesionales femeninas introducida dos años después que lo hiciera la ATP. Los puntos en el ranking profesional son otorgados por el nivel de torneo y al nivel más alto que llegue el jugador [1].

Es por esta razón que las posiciones de los jugadores van variando por semana, aunque también hay jugadores que logran mantenerse en la misma posición por semanas consecutivas. Al final se obtiene un ranking de acuerdo a los puntos obtenidos durante todo el año.



Imagen 2. Novak Djokovic, tenista profesional serbio que ocupa la primera posición del ranking [9].

4. Solución

Para dar solución a este requerimiento se plantea realizar el análisis con el uso de herramientas de software al alcance de todos los usuarios.

¿Lenguaje de programación a utilizar?

Se propone utilizar el lenguaje de programación R de modalidad abierta, libre y gratis que es especializado en proporcionar técnicas estadísticas y gráficas. Está disponible con licencia GNU, por lo que cualquier usuario tiene derecho a estudiar, usar modificar y compartir el software. R es muy utilizado en investigación científica, manipulación de datos, análisis estadístico entre otras áreas [6].

Entre las ventajas tenemos que ahorro de adquisición de licencias, es multiplataforma, admite varios tipos de datos, librerías para gráficos potentes, puede procesar grandes conjuntos de datos apoyándose en otras librerías [6].

Todos los paquetes de R están disponibles en el sitio oficial **CRAN** (*Comprehensive R Archive Network*) <http://cran.r-project.org/> [6]. En el sitio oficial se encuentra toda la documentación de apoyo. Existe una herramienta que facilita la programación con R, es un entorno de desarrollo integrado llamado Rstudio Desktop, es multiplataforma y su sitio oficial con documentación y link de descarga se encuentra en <https://posit.co/download/rstudio-desktop/> [7]

¿Librería a utilizar para el tratamiento de los datos del TENIS?

La librería que mejor responde a las necesidades del análisis que se requiere realizar es DEUCE. Este es un paquete que facilita el acceso y procesamiento de los datos del tenis con acceso en línea, esta desarrollada en lenguaje R y pretende ser una herramienta para analistas y profesores de estadística. Esta librería contiene las funciones para obtener estadísticas de los datos del tenis. [2].

Fue desarrollada por Stephanie Kovalchik, una australiana científica de datos Senior y es compartida en la plataforma GitHub <https://github.com/skoval/deuce>, con licencia GPL-2 permitiendo su estudio, modificación y redistribución. En la plataforma se encuentra toda la documentación y forma de instalación. Actualmente se encuentra en la versión 1.4 [2].

Para iniciar con su implementación es necesaria su instalación e importación en un archivo nuevo de R.

```
install_github("skoval/deuce")  
library(deuce)
```

Se recomienda ejecutar la sentencia siguiente para obtener todas las funciones disponibles, con las variables de respuesta y ejemplos de uso.

```
help(package = "deuce")
```

Resources for Analysis of Professional Tennis Data



Documentation for package 'deuce' version 1.4

- [DESCRIPTION file](#)
- [User guides, package vignettes and other documentation](#)

Help Pages

deuce-package	Download ATP player and match statistics
atp_elo	ATP Elo Ratings
atp_importance	ATP Point Importance
atp_matches	ATP Playing Activity
atp_odds	ATP Match Odds
atp_odds_match_lookup	ATP Match Odds Lookup Table
atp_players	Biographic Details of ATP Players
atp_rankings	Rankings of ATP Players
atp_tournaments	ATP Tournaments
deuce	Download ATP player and match statistics
elo_prediction	Elo-Based Win Prediction
fetch_activity	Download Player Activity

Imagen 3. Ayuda de la librería Deuce, con todas las funciones disponibles

Click en cada función para visualizar las variables que retorna

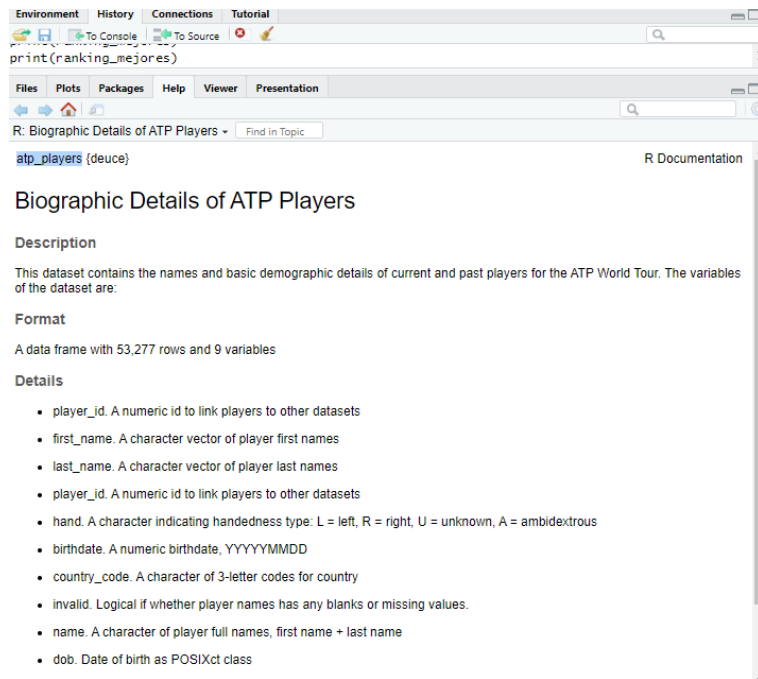


Imagen 4. Variables de la función atp_players

Para cargar los datos y tenerlos disponibles usamos la función `data` y para obtener un resumen estadístico usamos la función `summary`.

```
data(atp_rankings)
summary(atp_matches)
```

¿Fuente de datos del Tenis?

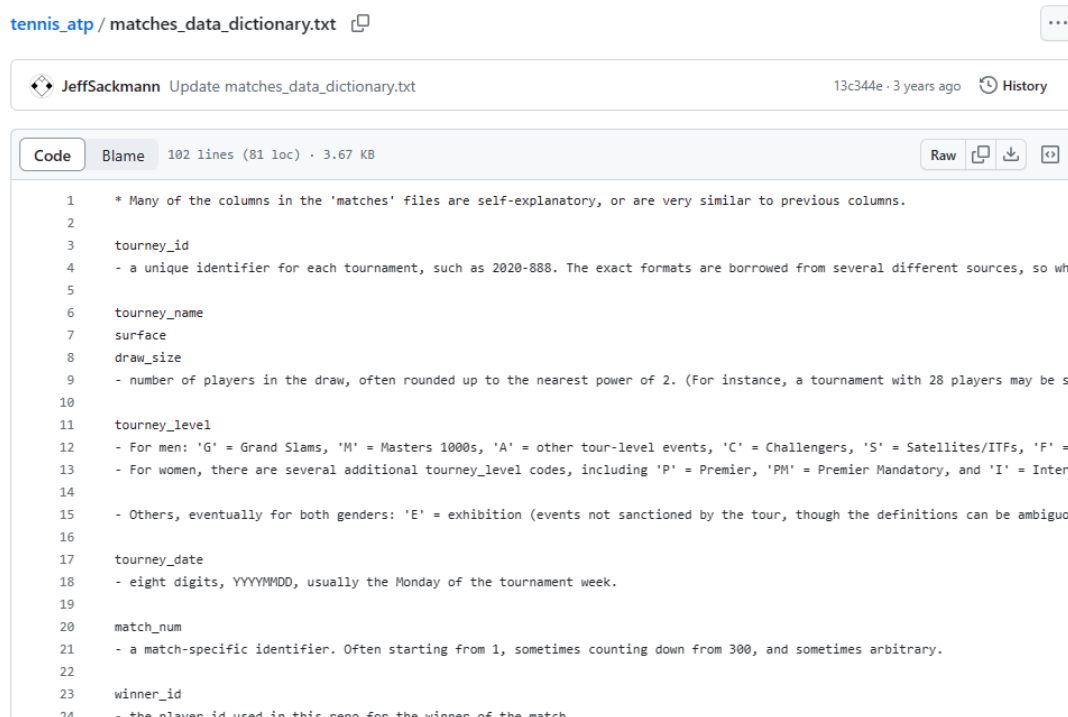
La fuente de datos del tenis a utilizar es la creada por Jeff Sackmann, desarrollador de software, investigador del tenis y escritor de varios artículos deportivos [3], quien ha recopilado información desde 1985 y se encuentra publicado en la plataforma GitHub https://github.com/JeffSackmann/tennis_atp con licencia CC BY-NC-SA 4.0 DEED y recomienda su respectiva atribución y solo uso no comercial [8].

Contiene los datos de jugadores, clasificaciones, resultados y estadísticas de Tenis de ATP y WTA. La información de ranking por ejemplo está completa desde 1985, la información de partidos desde 1991 y desde el 2000 de añadió información de partidos de dobles jugadores, pero se interrumpió a partir de finales de 2020[8].

Para facilitar la exploración de los datos esta fuente implementa redundancia con datos biográficos, junto con la clasificación y los puntos de clasificación, así como la edad a la fecha del torneo, que casi siempre es el lunes al comienzo del evento o cerca de él [8].

Esta fuente tiene una limpieza previa de datos considerando datos perdidos porque no existen en la ATP, otros han sido eliminados como partidos donde el perdedor tiene el 60% de puntos o partidos con un tiempo de juego inferior a 20 minutos.

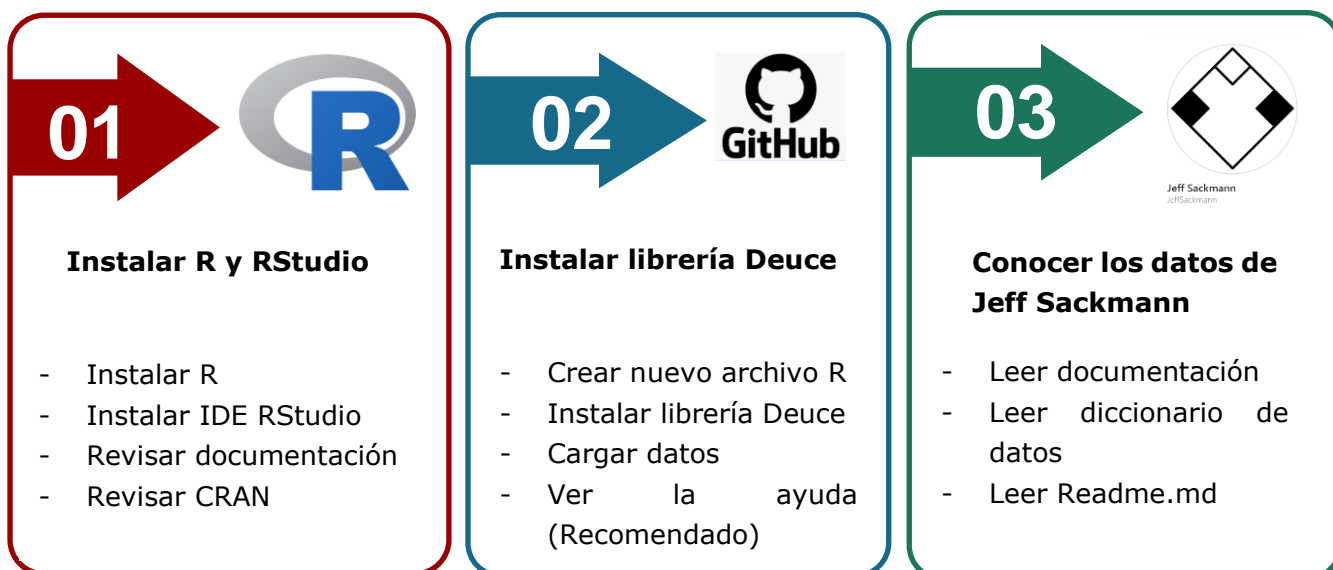
La definición de variables es bastante intuitiva, pudiendo identificar brevemente su concepto y contenido, aunque incluso tiene un diccionario de datos https://github.com/JeffSackmann/tennis_atp/blob/master/matches_data_dictionary.txt



```
1  * Many of the columns in the 'matches' files are self-explanatory, or are very similar to previous columns.
2
3  tournament_id
4  - a unique identifier for each tournament, such as 2020-888. The exact formats are borrowed from several different sources, so whi
5
6  tournament_name
7  surface
8  draw_size
9  - number of players in the draw, often rounded up to the nearest power of 2. (For instance, a tournament with 28 players may be sh
10
11 tournament_level
12 - For men: 'G' = Grand Slams, 'M' = Masters 1000s, 'A' = other tour-level events, 'C' = Challengers, 'S' = Satellites/ITFs, 'F' =
13 - For women, there are several additional tournament_level codes, including 'P' = Premier, 'PM' = Premier Mandatory, and 'I' = Intern
14
15 - Others, eventually for both genders: 'E' = exhibition (events not sanctioned by the tour, though the definitions can be ambiguou
16
17 tournament_date
18 - eight digits, YYYYMMDD, usually the Monday of the tournament week.
19
20 match_num
21 - a match-specific identifier. Often starting from 1, sometimes counting down from 300, and sometimes arbitrary.
22
23 winner_id
24 - the player id used in this repo for the winner of the match
```

Imagen 5. Diccionario de datos

Esquema de implementación de las herramientas seleccionadas



Implementación de la solución en R

Se crea un archivo R llamado tennisATP.R el cual puede descargar del siguiente link https://drive.google.com/file/d/1lCXWCATh1G2_XTqWgi0gw-yfR611lyIR/view?usp=sharing

También tiene un archivo Readme.txt <https://drive.google.com/file/d/10aDdt3kTM4O98oi0aWz4mRxRVURWxnF2/view?usp=sharing> disponible donde se explica como compilar y ejecutar el código y visualizar los datos.

Como en todo proceso de análisis de datos es recomendable obtener un resumen de los datos para familiarizarnos con ellos.

Iniciamos obteniendo un resumen de los datos de atp_ranking y observamos en la imagen 6 que "ranking" es una variable de tipo carácter y debemos transformarla a número si queremos ordenar los datos por este valor, observamos también que la máxima fecha de torneos es 2023-02-13 por lo que se entiende que los datos de ranking están almacenados hasta el torneo jugado esa fecha. Es preciso comprender que la fuente de datos de Jeff Sackmann es de libre uso, gratuita y no se actualiza de forma continua, pero el archivo tennisATP.R esta codificado de manera que el usuario cambie únicamente el año a consultar y el archivo automáticamente seleccionará los últimos 50 mejores jugadores del último torneo del año indicado, sin complicaciones.

```
ranking_date      ranking      player_id      ranking_points      date
Length:2879611    Length:2879611    Length:2879611    Length:2879611      Min.   :1973-08-27
Class :character   Class :character   Class :character   Class :character     1st Qu.:1996-02-05
Mode  :character   Mode  :character   Mode  :character   Mode  :character     Median :2004-10-04
                                     Mean  :2003-11-08
                                     3rd Qu.:2012-05-07
                                     Max.  :2023-02-13
                                     NA's  :3
```

Imagen 6. Resumen de los datos de ranking

En la siguiente grafica obtenida observamos la posición de los 50 mejores jugadores de último torneo registrado en la actual base de datos del año 2023. Los jugadores se encuentran ordenados por ranking de forma descendente.

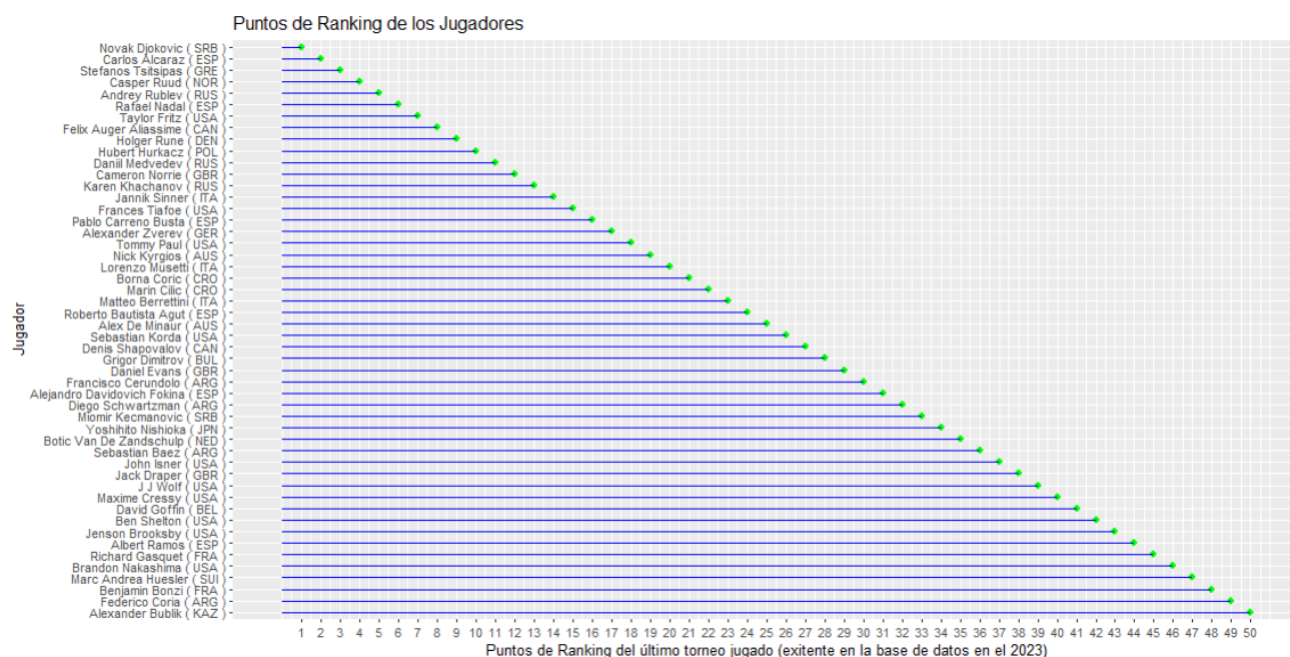


Imagen 7. Ranking última semana de torneo 2023 (2023-02-13)

Una vez que tenemos los jugadores seleccionados, se requiere conocer más detalles de cada tenista como fecha de nacimiento. Este último dato es importante para calcular la edad a la fecha del último torneo registrada en la base de datos (2023-12-31).

En la siguiente grafica observamos que el segundo mejor jugador **Carlos Alcaraz** es uno de los más jóvenes con 19 años cumplidos.

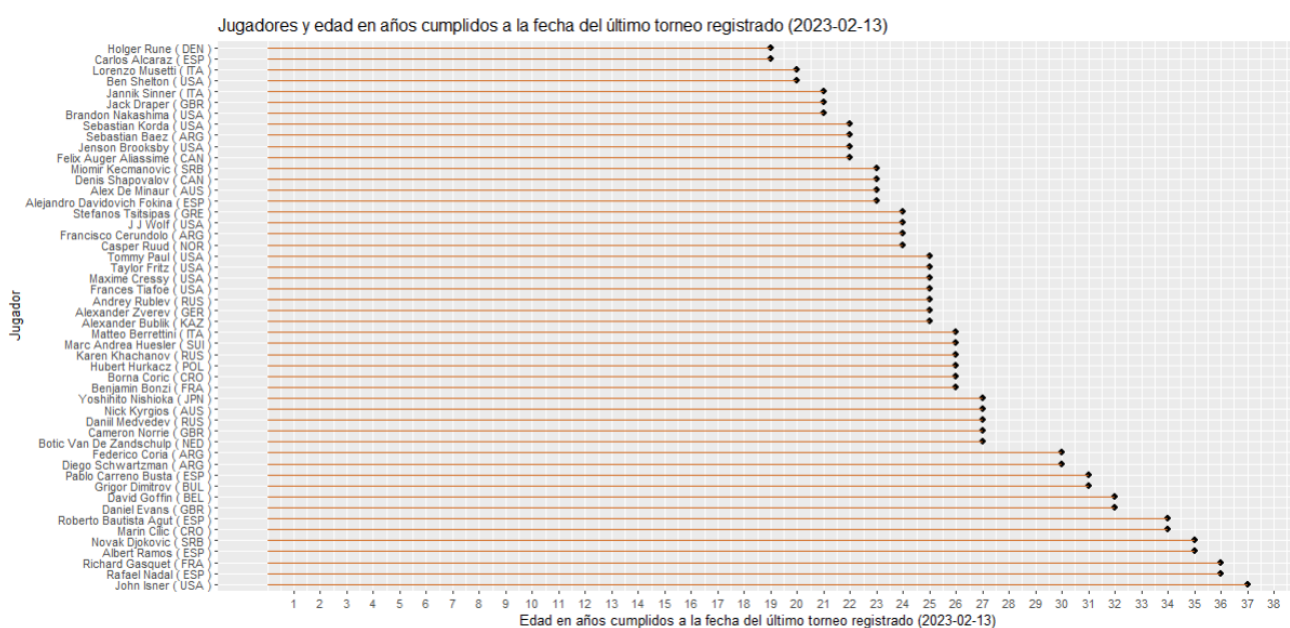


Imagen 8. Jugadores ordenados con la edad en años cumplidos a la última fecha de torneo registrada (2023-02-13)

Se grafica con diagrama de barras las edades y sus frecuencias y observamos que la mayor frecuencia de edades de los jugadores en el ranking es a los 25, 26 y 27 años.

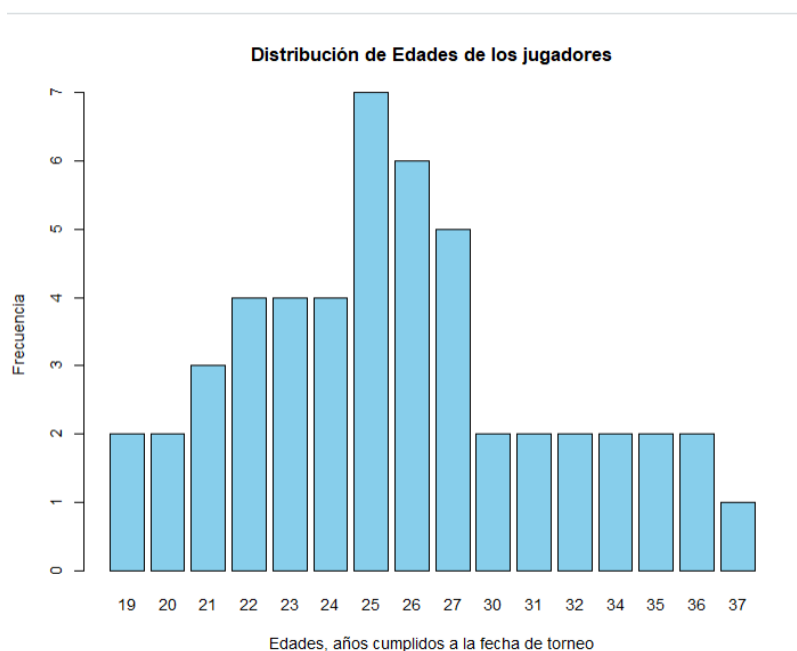


Imagen 9. Frecuencia de edades de los jugadores del torneo en 2023-02-13

En la gráfica siguiente podemos observar a que edad inician a jugar sus primeros partidos oficiales ATP los jugadores antes seleccionados. Observamos que la edad de inicio del tenista español Carlos Álcara es a los 14 años y ocupa el segundo lugar en el ranking.

Por otro lado, el tenista estadounidense Jhon Isner inicia a los 20 años y ocupa el puesto 38 en el ranking.

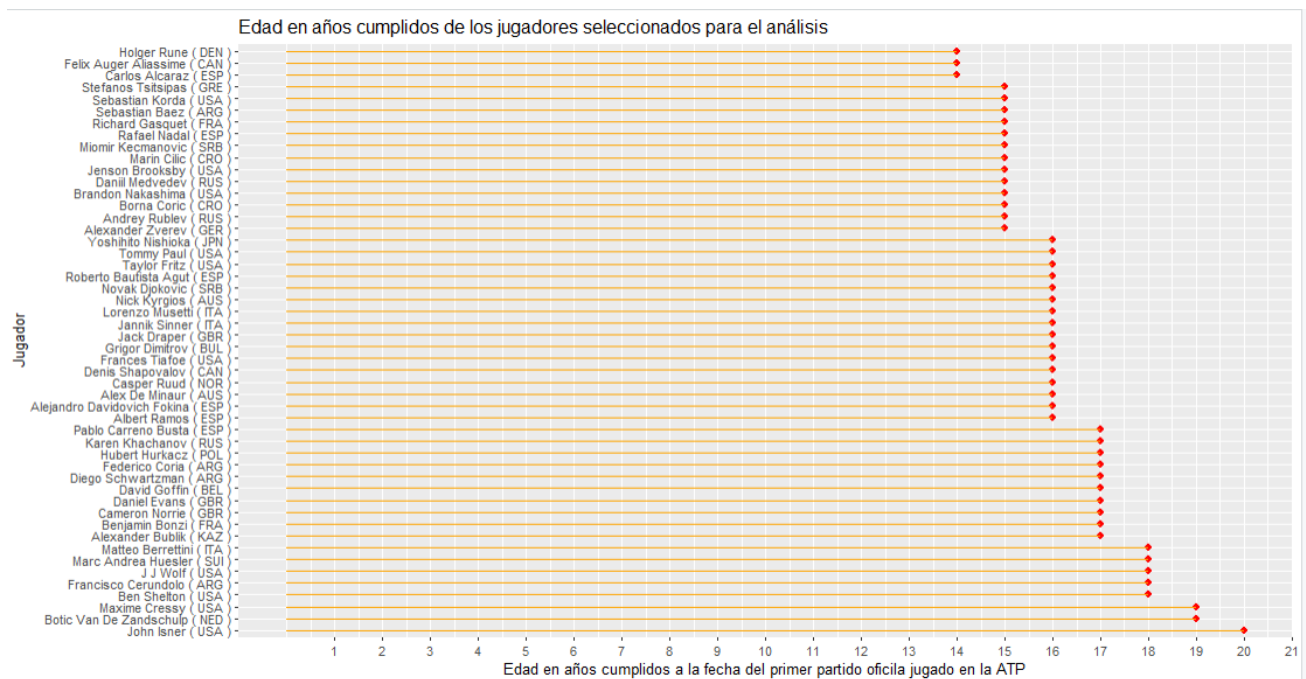


Imagen 10. Edad donde inician a jugar sus primeros partidos oficiales ATP los jugadores antes seleccionados

En la siguiente gráfica se observa la densidad de la edad donde los jugadores inician a jugar sus primeros partidos oficiales.

Se observa una densidad alta entre la edad de los 15 a los 16 años.

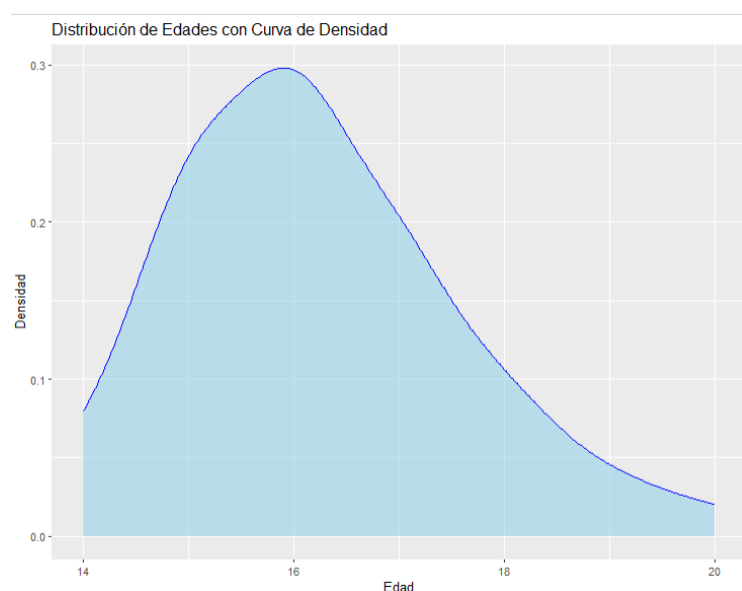


Imagen 11. Densidad de la edad donde los jugadores inician a jugar sus primeros partidos oficiales.

Obtenemos la edad promedio en la que todos los jugadores ganan los partidos.

```
[1] "Edad promedio de todo los jugadores que ganan un partido: 24.01 años"
```

Podemos observar en la siguiente gráfica como ganan los partidos los jugadores antes y después de la edad promedio de 24.01 años de edad.

Se observa que los jugadores ganan más partidos después de la edad promedio de 24.01 años de edad.

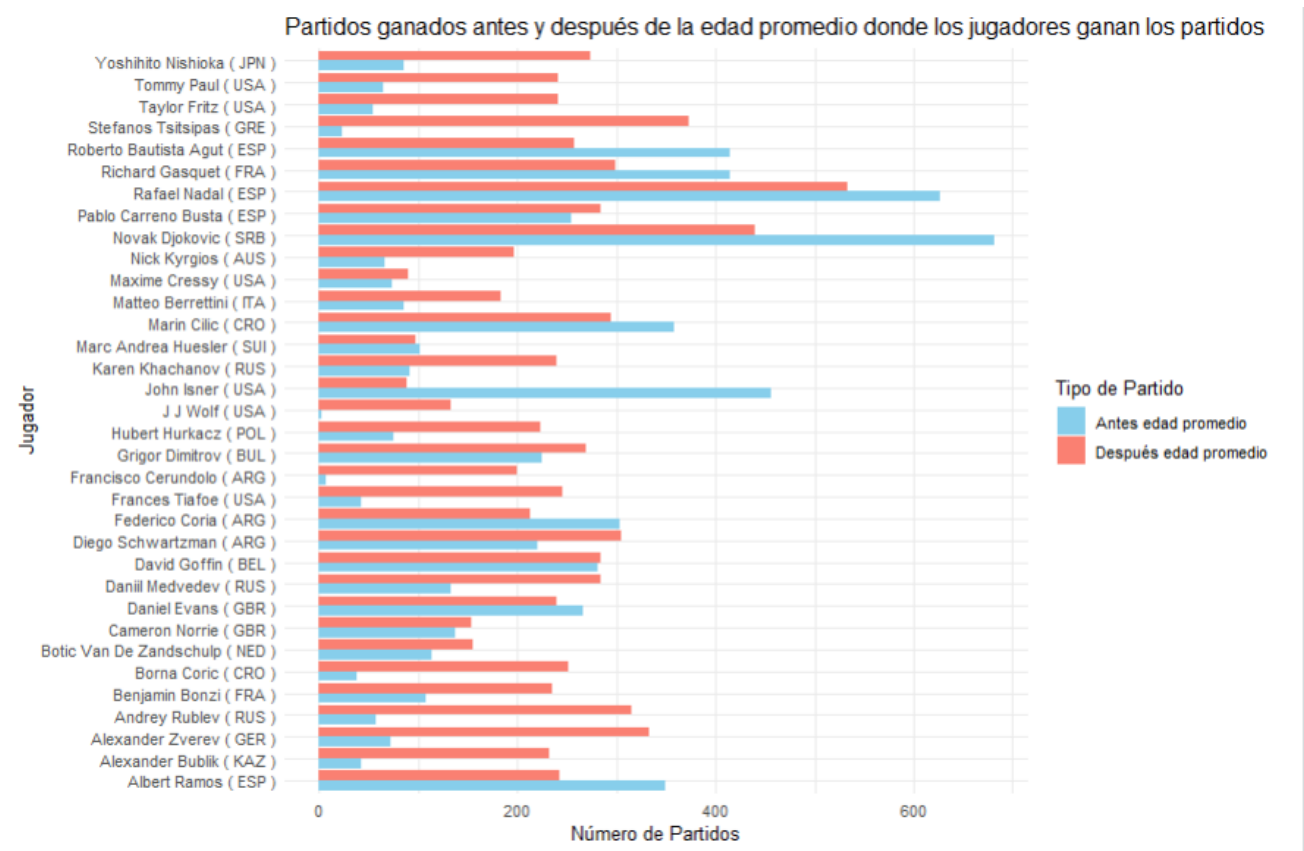


Imagen 12. Partidos ganados antes y después de la edad promedio donde los jugadores ganan los partidos

5. Conclusión

Con la integración del lenguaje de programación R, la librería DEUCE y la fuente de datos de Jeff Sackmann podemos codificar el archivo tennisATP.R. Con el cual podemos obtener información de valor referente a la edad de los 50 primeros jugadores de la última semana de torneo registrada en la base de datos del año 2023. Es preciso entender que como la fuente de datos es gratuita la actualización no es continua pero el archivo tennisATP.R esta codificado y parametrizado de forma que únicamente modifica el año y automáticamente obtendrá los datos del último torneo jugado en ese año.

Las gráficas colaboran con el objetivo que es ver la edad donde los jugadores tienen su primer partido jugado en la ATP. Y observamos algo muy especial y es el jugador español Carlos Alcaraz que juega su primer partido a los 14 años y llega a la posición 2 en el ranking a los 19 años, sin duda una figura admirable en el tenis.

Con estos resultados es una fuente de respaldo muy valiosa para que los entrenadores expliquen a los padres de familia y deportistas la importancia de iniciar a entrenar a temprana edad, por lo menos antes de los 11 años.

6. Bibliografía en estilo IEEE

- [1] A. Gozález, M. Gónzalez. "TennisBI : Aplicación Shiny para la visualización de datos de tenis." uvadoc.uva.es. Accessed: Mar. 22, 2024. [Online.] Available: <https://uvadoc.uva.es/handle/10324/63235>
- [2] Skoval. "skoval/deuce." github.com. Accessed: Mar. 23, 2024. [Online.] Available: <https://github.com/skoval/deuce/>
- [3] Jeff Sackmann. "Jeff Sackmann." jeffsackmann.com. Accessed: Mar. 23, 2024. [Online.] Available: <https://www.jeffsackmann.com/>
- [4] zineb. "Las 10 mejores academias de tenis 2023-2024." topelmundo.com. Accessed: Mar. 28, 2024. [Online.] Available: <https://topelmundo.com/2023/01/22/las-10-mejores-academias-de-tenis-2023/>
- [5] Tenis Quito. "Entrenamiento para niños desde 3 años." clasestenisquito.com. Accessed: Mar. 28, 2024. [Online.] Available: <https://clasestenisquito.com/>
- [6] Datademia. "¿Que es R?." clasestenisquito.com. Accessed: Mar. 28, 2024. [Online.] Available: <https://datademia.es/blog/que-es-r>
- [7] Posit software. "RStudio Desktop" posit.co. Accessed: Mar. 28, 2024. [Online.] Available: <https://posit.co/download/rstudio-desktop/>
- [8] Jeffsackmann. "tenis_atp." github.com. Accessed: Mar. 28, 2024. [Online.] Available: https://github.com/JeffSackmann/tennis_atp
- [9] Wikipedia. "Anexo:Ranking ATP de tenistas individual masculino de 2023." es.wikipedia.org. Accessed: Mar. 30, 2024. [Online.] Available: [https://es.wikipedia.org/wiki/Anexo:Ranking ATP de tenistas individual masculino de 2023](https://es.wikipedia.org/wiki/Anexo:Ranking_ATP_de_tenistas_individual_masculino_de_2023)