Módulo 3. Modelado estadístico en R

Como analista de rendimiento deportivo o científico del deporte, no solo se debe comprender el deporte, los jugadores y los datos de rendimiento, sino también saber cómo transmitir esta información a las partes interesadas clave, como la dirección de la oficina principal, los entrenadores y los jugadores. Un paso crucial que se encuentra entre la comprensión de los datos de rendimiento y el proceso de toma de decisiones es la selección de modelos estadísticos y predictivos para un análisis óptimo (Atkinson y Nevill, 2001; Anderson, 2015; Davenport, 2006).

Además de generar análisis basados en hipótesis o datos, depende de ti presentar los resultados de una manera significativa y comprensible para quienes son las partes interesadas clave. Por ejemplo, aunque podamos estar emocionados por los valores p, los errores estándar y los coeficientes beta, las partes interesadas clave generalmente no desean información de antecedentes tan detallada. Quieren el resultado final. Quieren responder preguntas como: ¿Qué significa todo esto? ¿Cómo puedo aplicar esta información a los planes de entrenamiento de los jugadores?

Este módulo está diseñado para guiarte a través de algunos principios básicos de los modelos estadísticos y ayudarte a proporcionar una justificación para la elección de modelos en función de los tipos de variables que se examinan y las preguntas que deseas responder. Este módulo proporciona una plantilla para modelos estadísticos que se pueden utilizar para presentar mejor tus datos a las partes interesadas clave, como jugadores, entrenadores y la dirección del equipo (Atkinson, 2001; Andrews et al., 2011; Brown & Sethna, 2003; Slack & Parent, 2006).

Hay varios términos con los que deberías familiarizarte como científico del deporte. El primer concepto importante es que existen diferentes tipos de variables. En 1946, S. S. Stevens categorizó las variables y declaró que todas las mediciones en ciencia se realizan utilizando una de cuatro escalas: nominal, ordinal, intervalo y razón. A continuación, se presentan ejemplos basados en los tipos de datos del módulo 1 aplicados al deporte:

Revisión de tipos de datos aplicados al deporte

En el tenis, por ejemplo, una variable nominal (categórica) podría ser el tipo de superficie de la cancha de tenis. Hay diferentes superficies, como el polvo de ladrillo, la pista dura, la hierba y la alfombra. Dado que el orden no es relevante, se puede considerar una variable nominal. Un ejemplo de una variable ordinal es el nivel de un jugador de tenis. Por ejemplo, si se proporcionara un conjunto de datos con las categorías de los cinco mejores jugadores de tenis profesionales, los cinco mejores de la División I, los cinco mejores de la División II y los cinco mejores jugadores de la División III, tendrías un conjunto de datos que consiste en una

variable ordinal. El conjunto tiene una jerarquía, lo que la convierte en una variable ordinal. Un ejemplo de una variable en escala de intervalo sería el puntaje en el tenis si la puntuación "hipotéticamente" siguiera el patrón cero (love), quince, treinta, cuarenta y cinco, y game. Sin embargo, ten en cuenta que este es un ejemplo teórico, ya que la verdadera puntuación en el tenis sigue la secuencia cero o love, quince, treinta, cuarenta, luego game. Esta secuencia de puntos no es una variable de intervalo real. La comparación se proporciona como un ejemplo para comprensión conceptual. Finalmente, un ejemplo de una medida en tenis que se encuentra en la escala de razón es la velocidad del saque, ya que se encuentra en una escala continua y tiene un valor de cero como punto de partida. Un ejemplo de una variable de intervalo que se puede aplicar a todos los deportes es la temperatura (en grados Fahrenheit o Celsius). Sin embargo, las variables de intervalo son difíciles de encontrar en el mundo del deporte, ya que la mayoría de las medidas en el deporte tienen un punto de referencia significativo y son variables de razón. Ejemplos de variables de razón en el tenis son el peso de una raqueta de tenis y la velocidad del saque, ya que tienen una métrica significativa y un punto de partida en cero (Reid & Schneiker, 2008; O'Donoghue & Ingram, 2001).

En el deporte del fútbol americano, un ejemplo de una variable nominal son las diferentes divisiones dentro de cada conferencia (este, sur, oeste y norte). Los niveles de categoría en el fútbol americano, ordenados de más a menos habilidosos, se consideran un ejemplo de una variable ordinal. Por ejemplo, la Liga Nacional de Fútbol (NFL, por sus siglas en inglés), la Asociación de Fútbol Americano (AFA), la Asociación Atlética Universitaria Nacional (NCAA, por sus siglas en inglés) División I, NCAA División II y NCAA División III en su conjunto se pueden considerar una variable ordinal debido a su naturaleza jerárquica. Un ejemplo de una variable de intervalo es el componente del tiempo que se distribuye uniformemente en cuatro cuartos. Un ejemplo de una variable de razón en el fútbol americano, al igual que en otros deportes, es la puntuación porque tiene un punto de partida de cero y las puntuaciones tienen una magnitud significativa, ya que los puntos se encuentran en una escala continua.

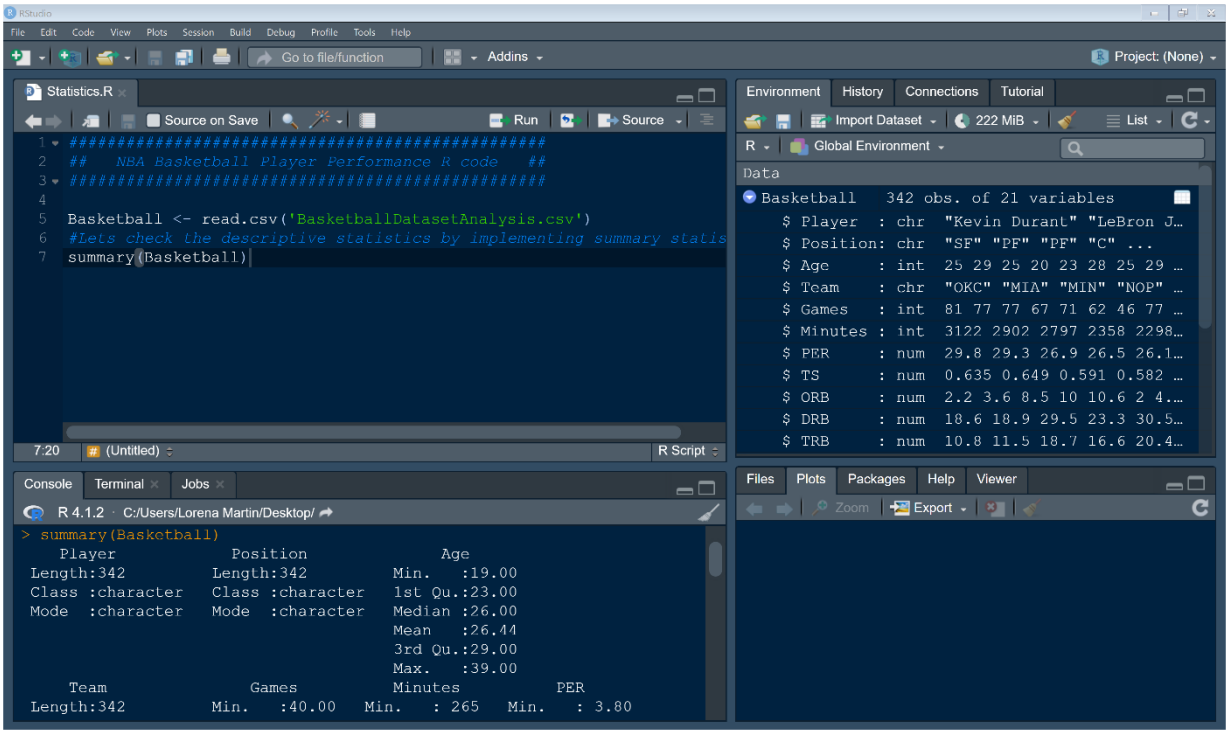
En baloncesto, ejemplos de variables nominales incluyen divisiones de equipos y posiciones de jugadores (centro, base, alero pequeño, ala-pívot y escolta). Un ejemplo obvio de una variable ordinal en este deporte incluye las clasificaciones de los equipos. Un ejemplo de una variable de intervalo son los cuatro cuartos que se juegan porque están distribuidos uniformemente a lo largo del juego. Finalmente, ejemplos de variables de razón en baloncesto incluyen el número de asistencias, canastas de campo, triples y tiros libres anotados, ya que todos ellos comienzan en cero como punto de partida y están en una escala continua. Un ejemplo de una variable ordinal es el nivel de la liga en la que se juega el baloncesto, como la NBA, FIBA y la Liga D. Un ejemplo de una variable de razón en baloncesto es el tiempo de vuelo de un jugador al realizar un mate.

En béisbol, un ejemplo de una variable nominal es la posición del jugador, ya que hay nueve posiciones diferentes de jugador. Un ejemplo de una variable ordinal es el ranking o niveles de división: las Grandes Ligas y las ligas menores de béisbol. En este deporte, cada entrada se determina por seis outs (tres de cada equipo) y, por lo tanto, se puede considerar una variable de intervalo. Aunque el béisbol es un deporte de equipo, difiere de otros deportes en que el tiempo no es una variable de intervalo. Esto se debe a que una entrada teóricamente puede continuar indefinidamente. Finalmente, ejemplos de variables de razón son el número de jonrones anotados por un jugador, el número de ponches por un lanzador o el puntaje general del juego. Un ejemplo de una variable de razón en el béisbol es la velocidad de lanzamiento de un lanzador.

En el fútbol, un ejemplo de una variable nominal es la posición del jugador. Ejemplos de variables ordinales en el fútbol incluyen las clasificaciones de equipos y divisiones de ligas, como el FC Barcelona y el equipo B del FC Barcelona, que tienen una jerarquía y un orden específico. Un ejemplo de una variable de intervalo en el fútbol es la duración del tiempo de juego, con dos tiempos de 45 minutos que componen el juego oficial de 90 minutos. Finalmente, ejemplos de variables de razón en este deporte incluyen el número de paradas realizadas por un portero, el número de goles marcados, el número de tiros de penalti realizados y el número de asistencias. Un ejemplo de una variable de razón es la distancia que recorren los jugadores durante todo un partido de fútbol.
Aquí dejaremos de repasar los tipos de datos.

Después de haber introducido los tipos de variables, pasamos al análisis de datos, comenzando con la exploración de los mismos. ¿Qué significa explorar datos? Una forma de empezar a explorar datos es representarlos gráficamente, construir distribuciones de frecuencia y examinar estadísticas descriptivas (ver la figura a continuación).

Figura 1: RStudio implementando la función de resumen para análisis descriptivos



Fuente: Captura de pantalla de RStudio realizada por el autor (RStudio, 2022).

Estadísticas como medias, medianas, desviaciones estándar y coeficientes de correlación pueden orientarnos en el desarrollo de preguntas interesantes que se examinarán con estadísticas inferenciales y modelos más avanzados. Ver figura a continuación para obtener una visión general de los métodos y modelos y cuándo aplicar cada uno.

Figura 2: Métodos estadísticos y sus aplicaciones

Method or Model	Definition or Usage
Indices of Central Tendency and Variability	
Mode	The most common value
Mean	The mathematical average
Median	The center value
Variance	The spread of the distribution
Standard Deviation	How much the values deviate from the mean
Inferential Statistics Used to Examine Group Differences	
Chi-square	Compare observed frequencies with expected frequencies
t-test	Examine differences between two groups on variable of interest
ANOVA	Examine differences between two or more groups
ANCOVA	Control for another variable that may influence the dependent variable
MANOVA	Examine group differences on multiple dependent variables
MANCOVA	Control for another variable that may influence the dependent variables
Statistics and Models Used to Examine Relationships or Predict Outcomes	
Correlation	Examine the association among two variables
Simple Linear Regression	Predict outcome with a single predictor variable
Multiple Linear Regression	Predict outcome with multiple predictors
Logistic Regression	Estimate the probability of the dependent variable class as the values of independent variables change

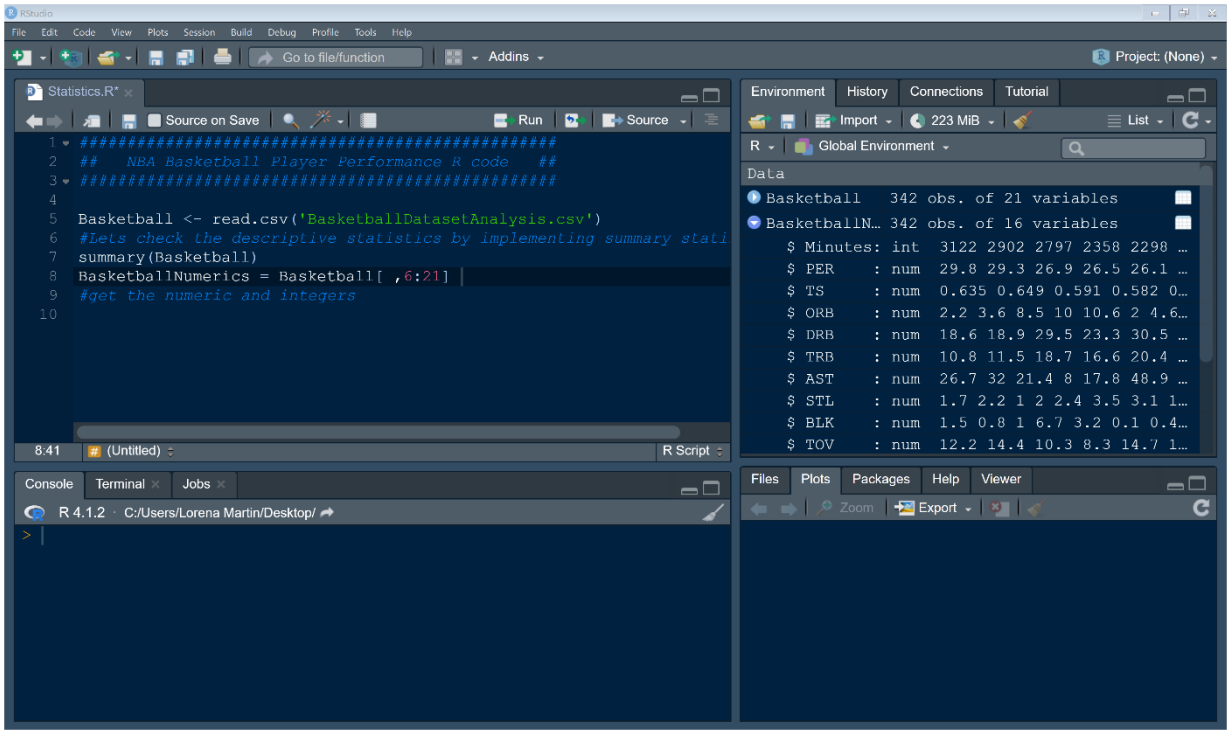
Fuente: Martin, 2016.

Method or Model	Método o modelo
Definition or Usage	Definición o uso
Mode Mean Median Variance Standard Deviation	Modo Media Mediana Varianza Desviación estándar
Chi-square t-test ANOVA ANCOVA	X^2 Prueba t ANOVA

MANOVA MANCOVA	ANCOVA MANOVA MANCOVA
The most common value The mathematical average The centre value The spread of the distribution How much the values deviate from the mean	El valor más común El promedio matemático El valor central La dispersión de la distribución Cuánto se desvían los valores de la media
Compare observed frequencies with expected frequencies Examine differences between two groups on variable of interest Examine differences between two or more groups Control for another variable that may influence the dependent variable Examine group differences on multiple dependent variables Control for another variable that may influence the dependent variables	Comparar frecuencias observadas con frecuencias esperadas Examinar diferencias entre dos grupos en una variable de interés Examinar diferencias entre dos o más grupos Controlar otra variable que puede influir en la variable dependiente Examinar diferencias de grupo en múltiples variables dependientes Controlar otra variable que puede influir en las variables dependientes
Indices of Central tendency and variability	Índices de tendencia central y variabilidad
Inferential statistics Used to examine group differences	Estadísticas inferenciales utilizadas para examinar diferencias de grupo
Statistics and models used to examine relationships or predict outcomes	Estadísticas y modelos utilizados para examinar relaciones o predecir resultados
Correlation Simple Linear Regression Multiple Linear Regression Logistic Regression	Correlación Regresión lineal simple Regresión lineal múltiple Regresión logística
Examine the association among two variables Predict outcome with a single predictor variable Predict outcome with multiple predictors Estimate the probability of the dependent variable class as the values of independent variables change	Examinar la asociación entre dos variables Predecir el resultado con una variable predictora Predecir el resultado con múltiples variables predictoras Estimar la probabilidad de la clase de la variable dependiente a medida que cambian los valores de las variables independientes

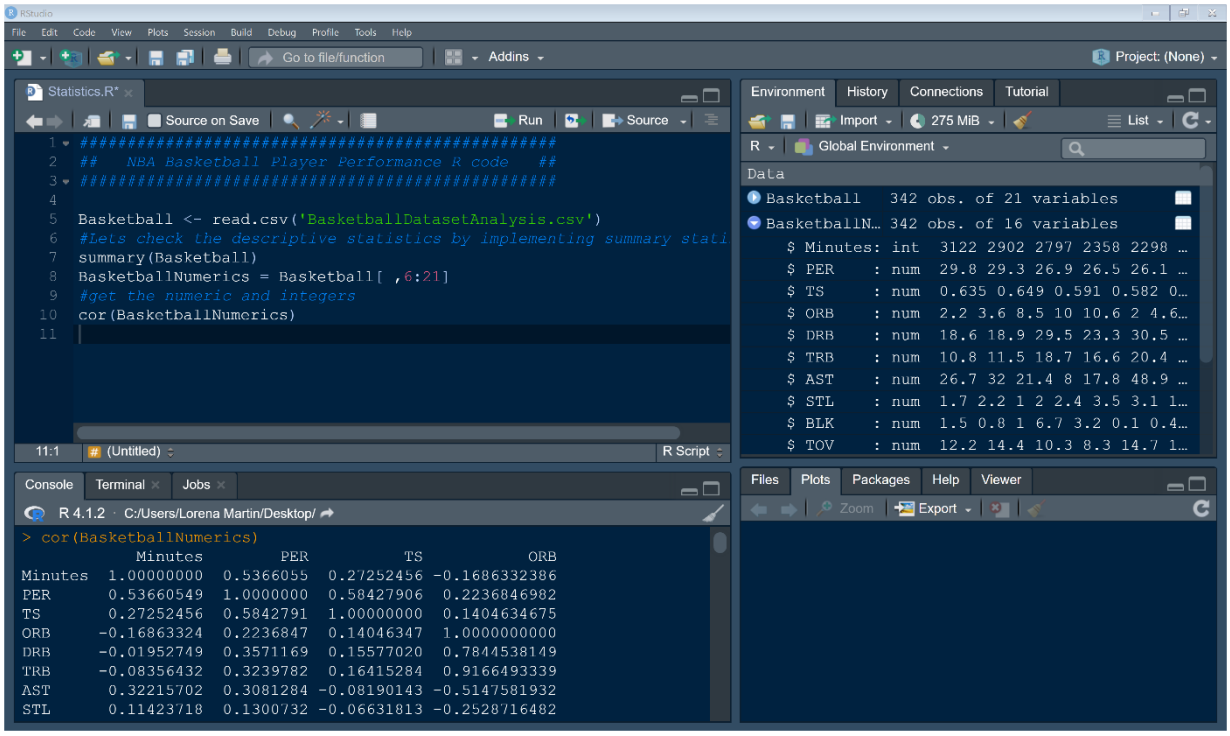
Comencemos por echar un vistazo al análisis de correlación. El coeficiente de correlación de Pearson mide la fuerza y dirección de una relación lineal entre dos variables. La fuerza de la relación se determina por qué tan cerca o lejos están los valores de correlación de cero o uno. Cuanto más cerca estén los números de uno, más fuerte y positiva será la relación entre dos variables. Cuanto más cerca estén los números de cero, más débil será la relación y, si están cerca de menos uno, la asociación puede ser fuerte pero inversamente correlacionada. Es importante reconocer que, aunque pueda existir una correlación entre dos variables, esto no necesariamente significa que una variable causó la otra. Es muy recomendable que conozcas tus datos, ya que no se muestran unidades de medida específicas en los resultados de los análisis estadísticos. El coeficiente de correlación de Pearson está diseñado para dos variables normalmente distribuidas; de lo contrario, se debe utilizar una prueba no paramétrica. Los coeficientes de correlación de Spearman y Kendall pueden utilizarse porque no están restringidos de esta manera, son pruebas sin distribución específica. Además, el análisis de correlación se basa en datos normalmente distribuidos; de lo contrario, se debe utilizar una prueba no paramétrica. En ese caso, el equivalente al coeficiente de correlación r de Pearson es el coeficiente de correlación de Spearman y debe utilizarse para evaluar datos que no están normalmente distribuidos. Nota que muchas personas utilizan los términos "correlación" y "asociación" de manera intercambiable, y no debería ser así. En el campo de la ciencia de datos, el término correlación se refiere específicamente a la intensidad y dirección de la relación lineal entre variables; el término asociación se usa de manera más informal y no implica una inferencia directa a partir de tus análisis.

Figura 3: Subconjunto del conjunto de datos para incluir tipos de datos numéricos y enteros en R, subconjunto utilizando corchetes cuadrados [filas, columnas] para examinar correlaciones.



Fuente: Captura de pantalla de RStudio realizada por el autor (RStudio, 2022).

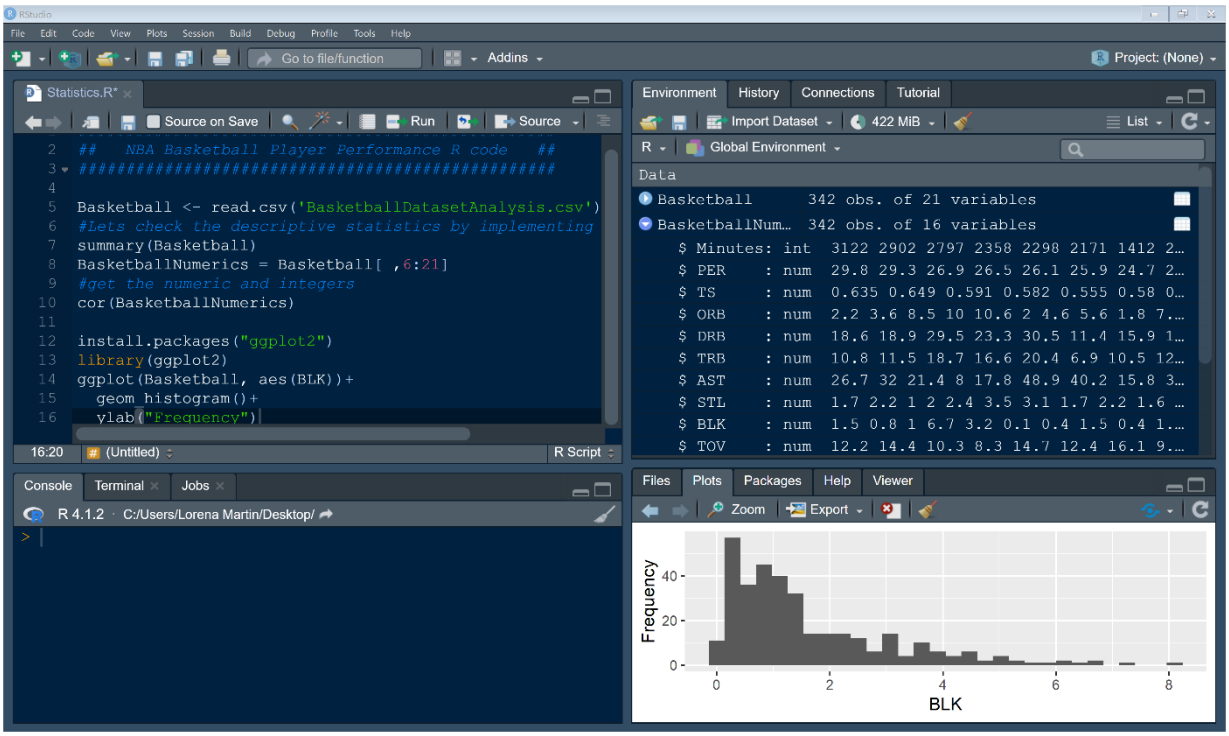
Figura 4: Correlaciones en R implementando la función cor()



Fuente: Captura de pantalla de RStudio realizada por el autor (RStudio, 2022).

Es importante saber qué estás tratando de cuantificar y cómo se aplica al deporte. Conoce tus datos y examina las estadísticas resumidas. Para variables de intervalo y razón, verifica los histogramas y busca la normalidad, como se muestra en la figura a continuación.

Figura 5: Histograma que muestra la distribución de la variable bloques (BLK).

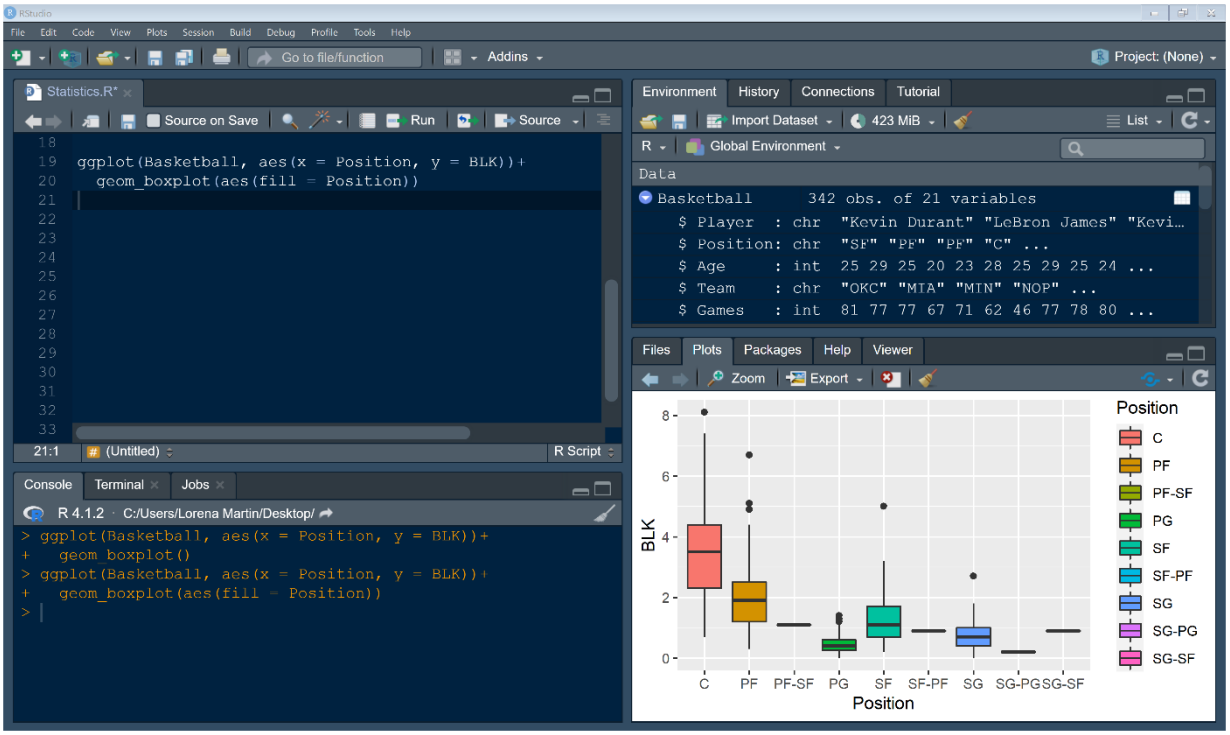


Fuente: Captura de pantalla de RStudio realizada por el autor (RStudio, 2022).

Además, es importante examinar tus datos en busca de valores atípicos, ya que es común en deportes que tengas que ser creativo con diferentes tipos de análisis. Esto es importante porque si la muestra elegida está sesgada o contiene valores atípicos extremos, los resultados de tus análisis pueden contradecir los valores reales de la norma. Como ejemplo, ejecuté un "modelo hipotético" utilizando solo valores atípicos con fines de demostración. Examiné la altura (veinte de los más bajos y más altos) de ex-jugadores de baloncesto de la NBA en el porcentaje de tiros de campo y los puntos por partido anotados. Los resultados mostraron que ni el porcentaje de tiros de campo ni los puntos por partido diferían por altura. Estos hallazgos son imprácticos ya que se basan en una muestra aberrante. Por otro lado, los resultados de un análisis en una muestra normalmente distribuida de jugadores actuales de la NBA revelaron que hay una diferencia significativa en el porcentaje de tiros de campo y los puntos por partido por altura. Específicamente, los jugadores más altos tienen un porcentaje de tiros de campo más alto en comparación con los jugadores más bajos, mientras que los jugadores más bajos anotan muchos más puntos por partido. Esto ejemplifica la importancia

de comprender tanto el deporte de interés como la capacidad para ejecutar los modelos estadísticos apropiados.

Figura 6: BLK por posición del jugador para examinar los valores atípicos señalados por los puntos por encima del valor máximo del diagrama de caja.



Fuente: Captura de pantalla de RStudio realizada por el autor (RStudio, 2022).

Nota: Los gráficos y las tablas mostradas fueron generados implementando el paquete Grammar of Graphics ggplot2. La estructura del gráfico con ggplot es la siguiente: ggplot(conjunto de datos, aes(variable x, variable y)) + geom(tipo de gráfico que deseas), como se muestra en la figura ggplot(Baloncesto, aes(Posición, BLK)) + geom_boxplot()
Después de determinar si tu muestra sigue una distribución normal y distinguir qué tipos de variables componen tu conjunto de datos, es hora de elegir un modelo y verificar suposiciones. Las suposiciones que deben cumplirse para usar una prueba paramétrica incluyen datos distribuidos normalmente, homogeneidad de varianza e independencia de las observaciones. Sin embargo, si tus datos no cumplen con las suposiciones para las pruebas paramétricas, se deben implementar pruebas no paramétricas.

Cuando se examinan las asociaciones entre variables ordinales, en lugar de utilizar la correlación de Pearson, se recomienda utilizar el coeficiente de correlación de Spearman, una estadística no paramétrica que se basa en la clasificación de los datos antes de aplicar la ecuación de Pearson. El coeficiente de correlación de Spearman normalmente se utiliza cuando se explora el conjunto de datos y se observa un conjunto de datos de muestra grande para el cual se debe utilizar una prueba no paramétrica. Para un conjunto de datos relativamente pequeño, se prefiere el coeficiente de correlación de Kendall sobre el de

Spearman. A continuación, detallamos algunos de los análisis comúnmente aplicados y las suposiciones que deben cumplirse para algunos de los modelos paramétricos más comunes del rendimiento deportivo.

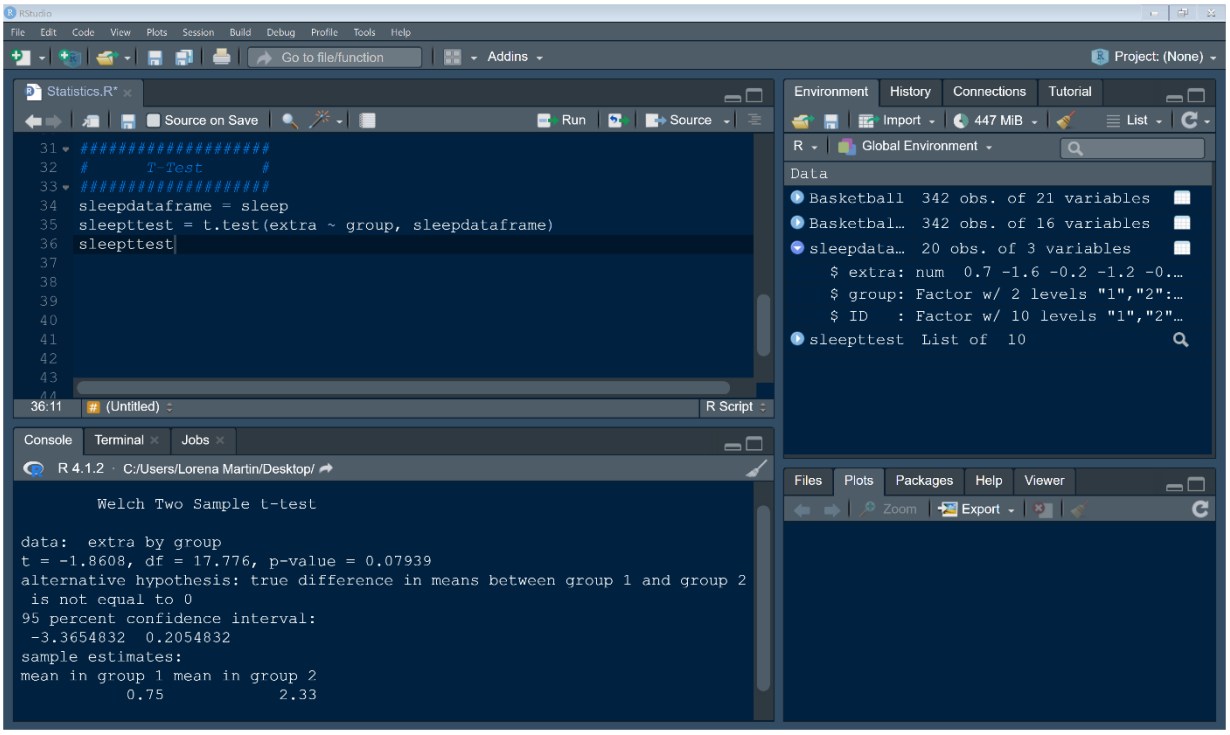
Los investigadores en diversas disciplinas utilizan la prueba t de Student para comparar medias de grupos y determinar si hay una diferencia significativa entre dos grupos; por ejemplo, entre el Real Madrid FC y el FC Barcelona en el número de goles anotados por sus máximos goleadores. Sin embargo, cabe señalar que existen varios tipos de pruebas t: la prueba t de una muestra, la prueba t de medias independientes (también conocida como prueba t de dos muestras) y la prueba t pareada (también conocida como prueba t de medias dependientes).

La prueba t de una muestra debe aplicarse al comparar la media del equipo de interés con un punto de referencia, por ejemplo, al comparar el número de abdominales realizados con un promedio nacional para una determinada población o grupo de edad. Otro ejemplo sería si un entrenador de tenis quisiera saber si la velocidad de los saques de sus estudiantes supera la velocidad de saque de los 20 mejores jugadores de tenis profesionales del mundo, para quienes ya hay datos disponibles basados en radares de velocidad y tecnología Hawkeye. En este caso, elegiríamos realizar una prueba t de una muestra relativamente sencilla. Esta prueba analizaría la media o promedio de la velocidad de saque de sus estudiantes, que se compararía con la velocidad promedio de la población de interés, en este caso, la velocidad de saque de los 20 mejores tenistas profesionales del mundo. En resumen, el grupo de interés es el grupo de tenistas a los que el entrenador está entrenando, la variable dependiente es la velocidad del saque, y la prueba t de una muestra arrojará un valor p que indicará hallazgos significativos o no significativos en comparación con un parámetro conocido, la velocidad de saque de los 20 mejores tenistas profesionales del mundo.

Esto difiere de la aplicación de una prueba t independiente, donde la situación se configuraría de la siguiente manera: si deseas comparar la velocidad de saque de dos grupos, tenistas femeninas y tenistas masculinos, entonces aplicarías una prueba t independiente. En cuanto a las suposiciones que deben cumplirse para una prueba t independiente, la variable dependiente, variable de resultado o variable de respuesta (utilizadas indistintamente; algunos términos pueden preferirse más en ciertos campos que en otros) es numérica y continua. Ejemplos de variables continuas son la distancia recorrida en el campo de fútbol, el tiempo dedicado al entrenamiento en la cancha de tenis o la duración de un partido de tenis o béisbol, las horas de entrenamiento en un deporte, el número de errores no forzados en el deporte del tenis y el número de touchdowns anotados en el fútbol americano. También, el número de goles marcados en fútbol, el número de tiros libres anotados en baloncesto y el número de carreras anotadas en béisbol son ejemplos de variables continuas, por nombrar algunos. La segunda suposición que debe cumplirse para la prueba t independiente es que la variable independiente, variable explicativa o variable predictora consta de dos grupos

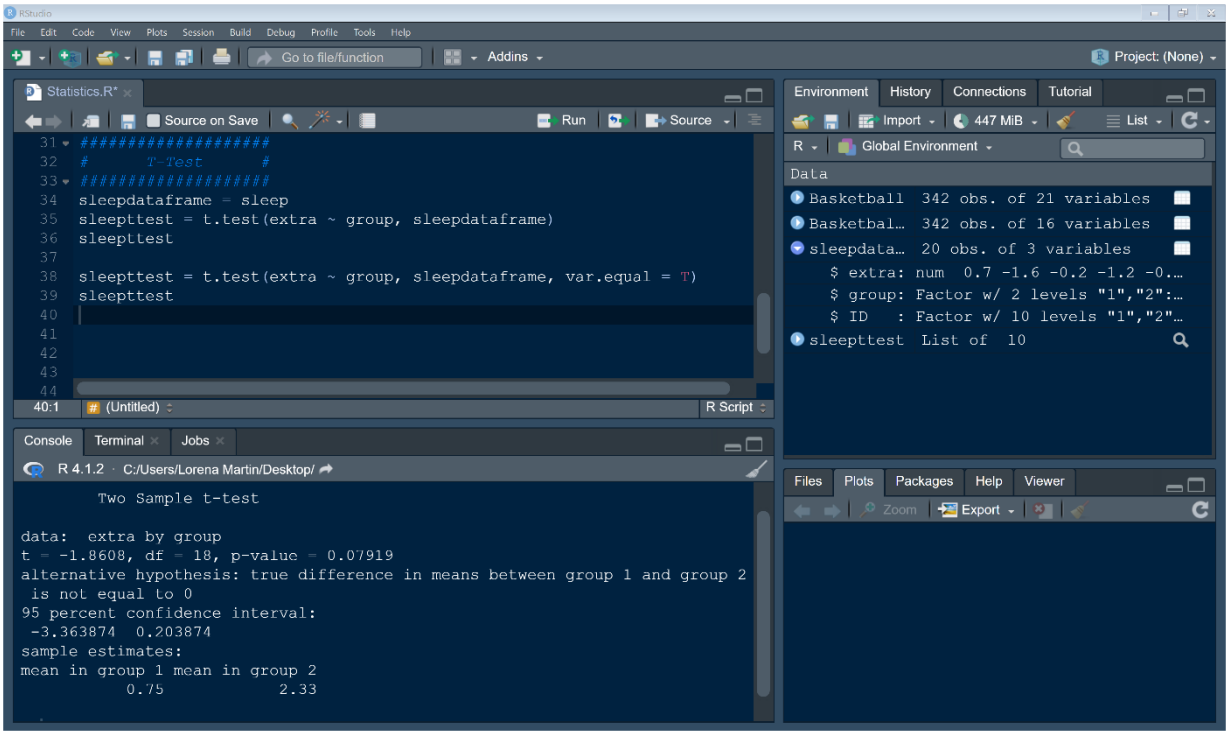
independientes que son categóricos. Por ejemplo, podríamos examinar las diferencias entre dos deportes, como el béisbol y el fútbol, siendo ambos representantes de los niveles de la variable categórica deporte y las observaciones del estudio siendo independientes entre sí. La tercera suposición es la independencia de las observaciones, diseñada para asegurar que cada deportista en un grupo sea examinado solo en ese grupo y no en más de uno. La independencia de las observaciones es una suposición crítica de muchos modelos y pruebas estadísticas. La cuarta suposición de la prueba t es que no haya valores atípicos extremos. Incluir estos valores en tu conjunto de datos podría afectar los resultados de tu prueba t independiente. Si encuentras algo interesante y deseas incluir el valor atípico extremo, es posible que quieras utilizar un análisis diferente que sea más robusto. Recuerda que debes utilizar el mejor modelo para el tipo de datos que tienes. La quinta suposición es que tu variable dependiente sigue una distribución normal para cada uno de los niveles de la variable independiente (en un ejemplo donde el deporte es la variable independiente categórica con 2 niveles, como béisbol y fútbol). Puedes utilizar pruebas de normalidad como la prueba de Shapiro-Wilk y la prueba de Kolmogorov-Smirnov para verificar que esta suposición se cumple. Finalmente, la sexta suposición se refiere al criterio de homogeneidad de varianzas, es decir, que estas deben ser iguales entre los grupos. La prueba de Levene es específica para evaluar la homogeneidad de las varianzas.

Figura 7: Prueba t



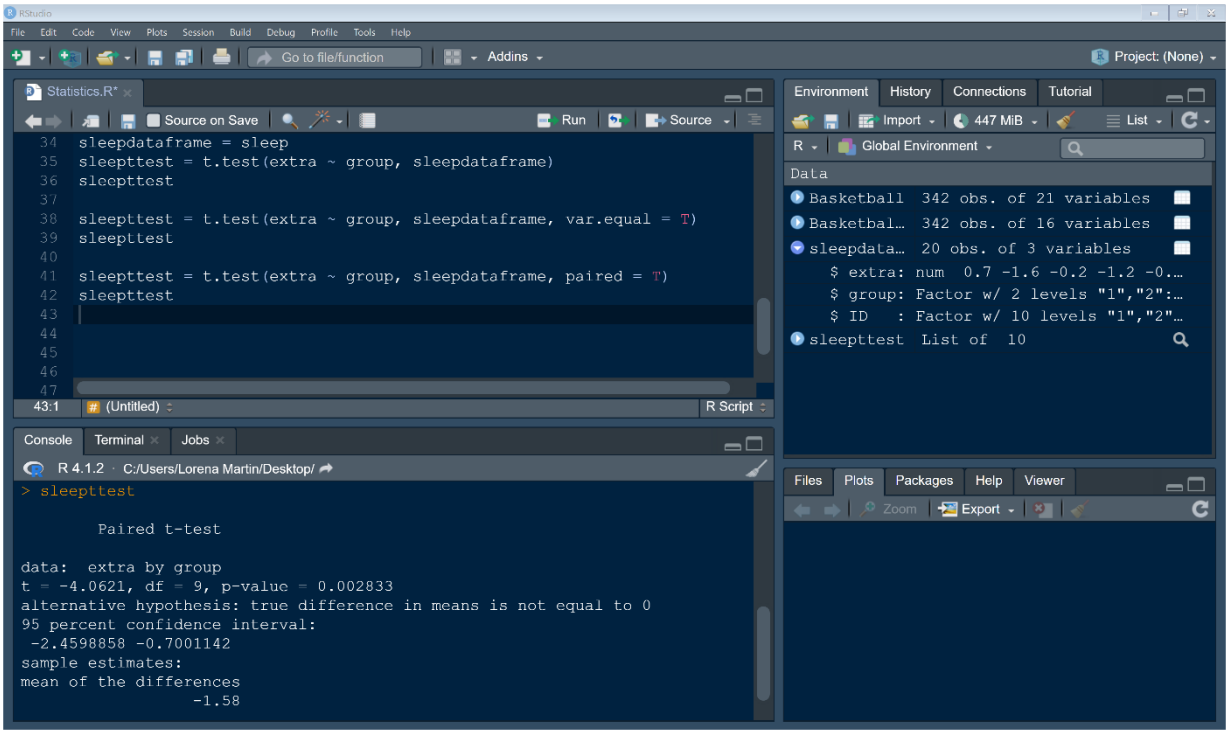
Fuente: Captura de pantalla de RStudio realizada por el autor (RStudio, 2022).

Figura 8: Prueba t para muestras independientes



Fuente: Captura de pantalla de RStudio realizada por el autor (RStudio, 2022).

Figura 9: Prueba t para muestras dependientes



Fuente: Captura de pantalla de RStudio realizada por el autor (RStudio, 2022).

Si estás interesado en comparar dos puntos temporales o un antes y un después, o el rendimiento de un año a otro, entonces puedes implementar una prueba t de muestras dependientes. Existe una suposición adicional de no tener valores extremos atípicos que debe cumplirse para satisfacer los criterios para ejecutar este tipo de prueba t. Un ejemplo detallado de cuándo implementar esta prueba t es cuando queremos responder preguntas sobre el rendimiento deportivo, como: ¿Mejoró la velocidad de nuestros jugadores de fútbol desde el primer día de la pretemporada en comparación con el último? En esta pregunta se asumen dos puntos temporales. Por lo tanto, tendrás valores de velocidad de los jugadores de fútbol al inicio y después de la pretemporada. La prueba t de muestras dependientes les proporcionará un resultado que indica si la mejora en la velocidad fue significativa en esos dos puntos temporales. Sin embargo, debemos ser precavidos, especialmente al usar demasiadas pruebas t. Estas pueden arrojar resultados falsos positivos y aumentar la probabilidad de un error de Tipo I.

Cuando desees comparar más de dos grupos, como el rendimiento de equipos en tu región, etc., entonces necesitarás implementar una prueba estadística que pueda manejar varias comparaciones combinacionales. Para esto, el modelo de análisis de varianza (ANOVA) es la elección correcta. Cuando tu pregunta consiste en examinar las diferencias entre tres o más grupos en una variable dependiente numérica continua, este es el modelo que debes aplicar. Por ejemplo, si estuviéramos interesados en examinar las diferencias entre las posiciones de los jugadores de baloncesto en el porcentaje de triples, este modelo sería apropiado. Nuestra

pregunta de investigación sería: ¿Existen diferencias entre los pívots, aleros pequeños y escoltas en el porcentaje de triples?

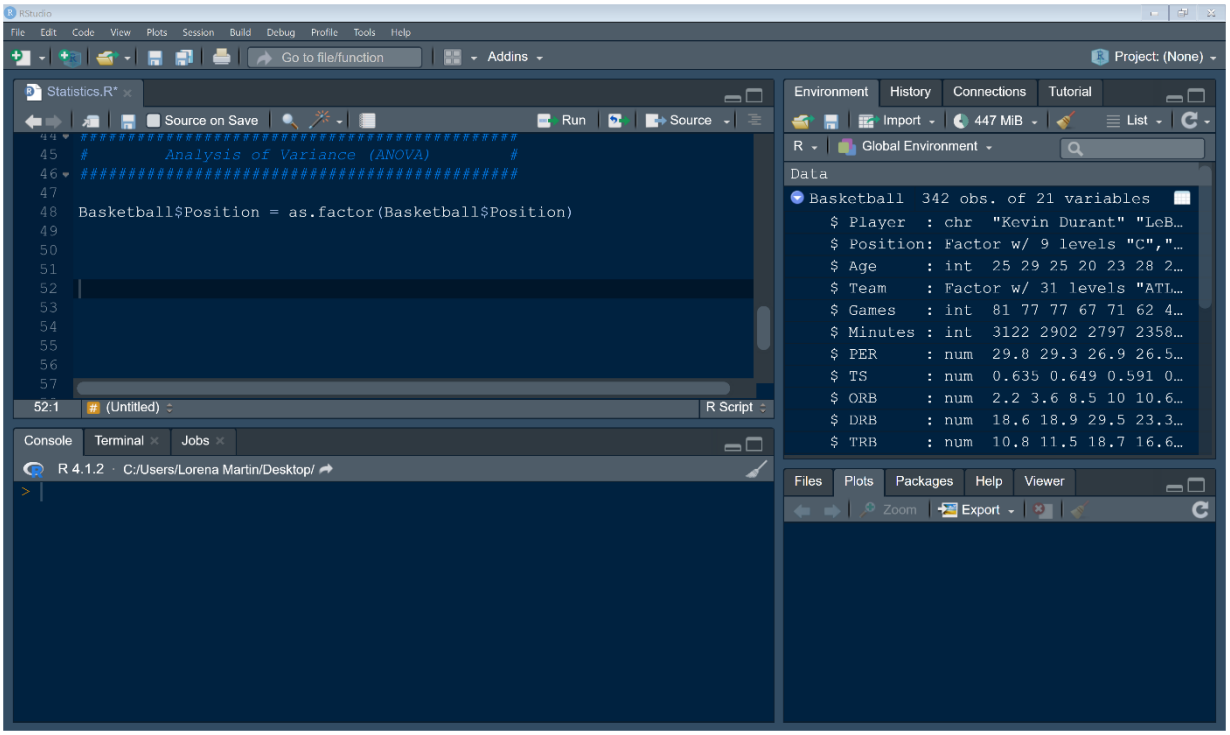
Es importante entender que el modelo ANOVA puede determinar diferencias significativas entre las posiciones de los jugadores en el porcentaje de triples, pero no nos dice entre cuáles posiciones de jugadores se encuentran las diferencias significativas. Por lo tanto, se necesita una investigación adicional, y se deben realizar análisis post hoc para determinar qué combinaciones de grupos son estadísticamente diferentes entre sí. Existen numerosos tipos de análisis post hoc, alrededor de 18 diferentes, pero recomendamos utilizar la prueba de Tukey HSD, ya que representa la búsqueda de Diferencias Honestamente Significativas (HSD, por sus siglas en inglés).

Antes de decidir qué ANOVA es el modelo que necesitas, debes asegurarte de que se cumplan seis suposiciones. La primera es que la variable dependiente sea continua. La segunda, que tu variable independiente o variable de interés conste de dos o más grupos o categorías. La tercera suposición de ANOVA es que debe haber independencia de las observaciones, lo que significa que no existe relación entre las variables independientes. Por ejemplo, si estás interesado en examinar las diferencias entre equipos de baloncesto, esta suposición se cumple cuando verificamos que los jugadores en el Miami Heat solo pertenecen a ese grupo y no a ningún otro grupo o equipo, como el Oklahoma City Thunder. Si intentáramos comparar el porcentaje de tiros entre las posiciones de los jugadores, por ejemplo, la suposición se incumpliría si un deportista jugara en múltiples posiciones, como base y escolta. Si queremos utilizar ANOVA, debemos asegurarnos de que los jugadores en cada posición estén designados y jueguen solo en una posición. La cuarta suposición es que no haya valores atípicos extremos. Por lo general, la regla es que, si el valor se encuentra fuera de dos desviaciones estándar, se considera un valor atípico; aunque en la estadística clásica, el punto de datos tendría que estar más alejado que tres desviaciones estándar por encima o por debajo de la media. La quinta suposición es la de una distribución normal. ANOVA tiende a ser robusto ante la violación de la normalidad; sin embargo, si los datos están claramente sesgados, es posible que sea mejor transformar los datos u optar por una prueba no paramétrica, como el modelo de Kruskal-Wallis. La sexta y última suposición que debe cumplirse para ejecutar un modelo ANOVA es la homogeneidad de varianza, que se puede verificar mediante la prueba de Levene. Si no se cumple la suposición de homogeneidad de varianza, hay dos modelos alternativos que se pueden utilizar: la prueba de Welch y la prueba de Brown y Forsythe. Si se cumplen las seis suposiciones, estás listo para ejecutar el modelo ANOVA. Recuerda realizar análisis post hoc para identificar dónde se encuentran las diferencias significativas entre los grupos.

A continuación, te guiaremos paso a paso en R sobre cómo configurar tus datos para poder implementar un ANOVA de una vía, donde la variable independiente es la posición y la variable dependiente son los rebotes ofensivos. En otras palabras, este sería el análisis si la

pregunta fuera la siguiente: ¿Existe una diferencia significativa entre las posiciones de los jugadores en los rebotes ofensivos?

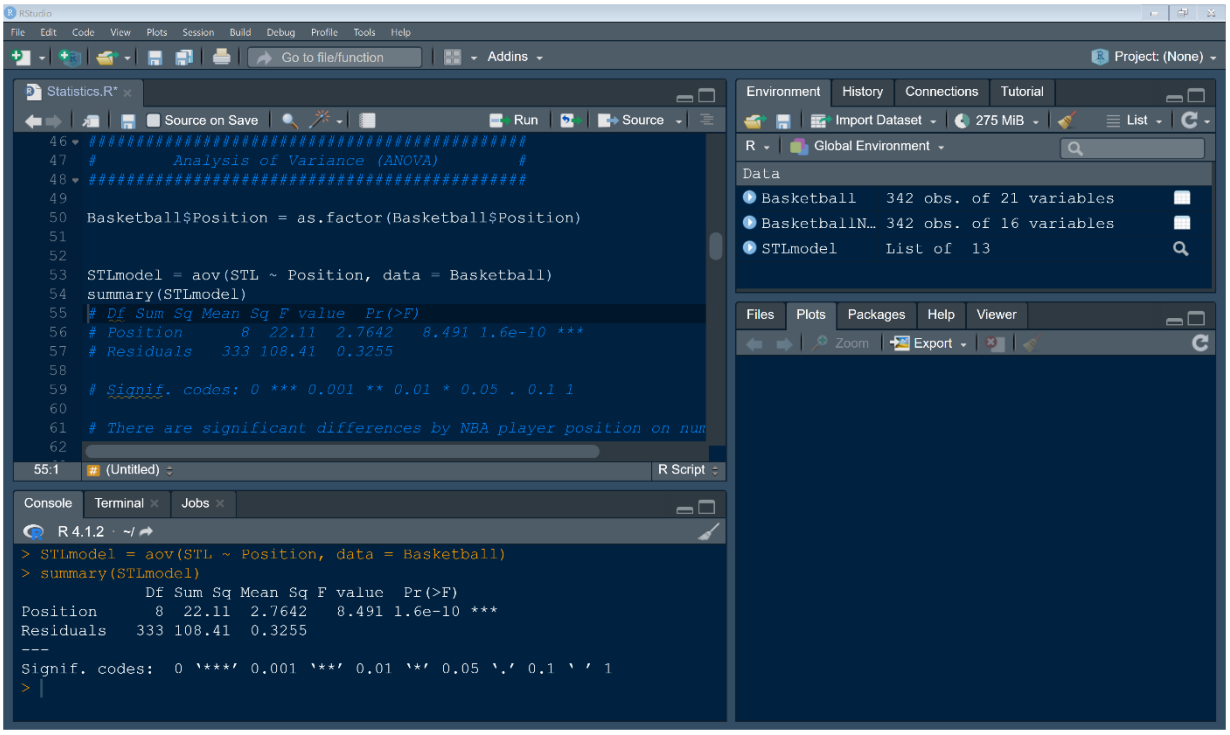
Figura 10: Configurar la variable de posición como un tipo de dato factor para la implementación de un ANOVA.



Fuente: Captura de pantalla de RStudio realizada por el autor (RStudio, 2022).

Después de asegurarte de que la variable independiente de interés sea de tipo de dato factor, puedes implementar el ANOVA como se muestra en la figura a continuación.

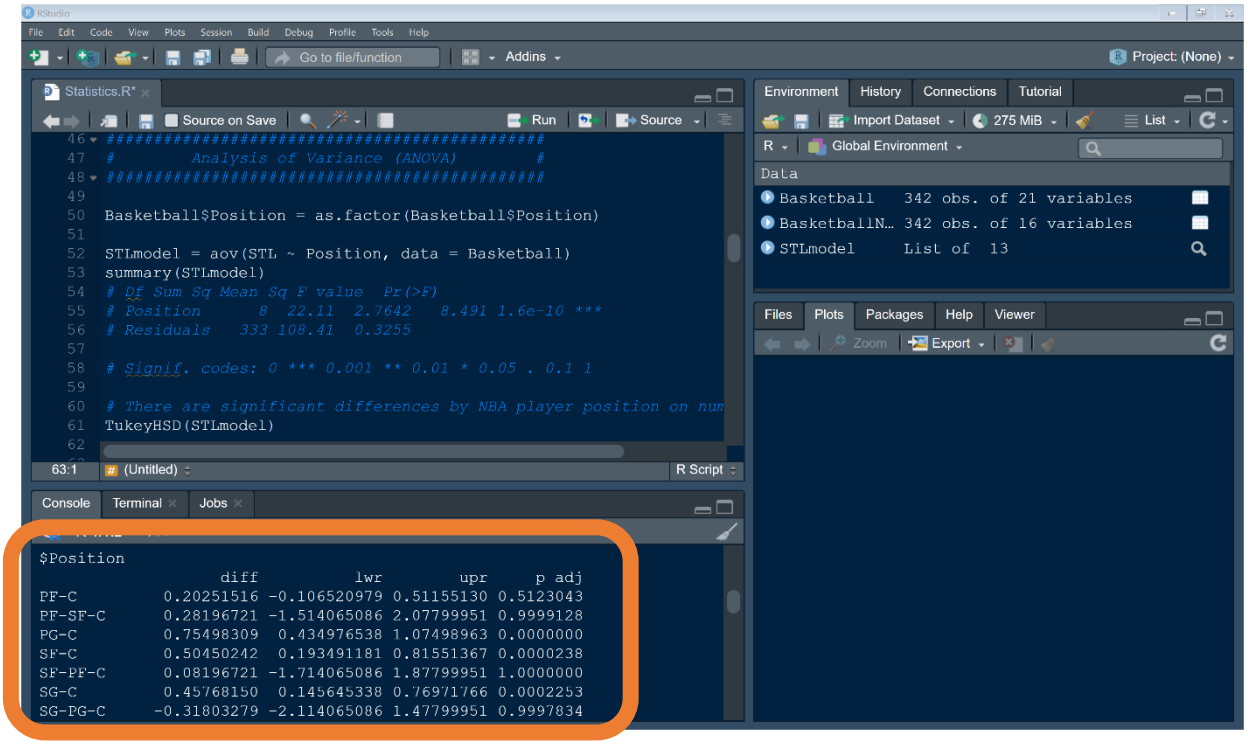
Figura 11: ANOVA, robos por posición del jugador.



Fuente: Captura de pantalla de RStudio realizada por el autor (RStudio, 2022).

Luego, después de examinar los resultados del ANOVA, si existen hallazgos estadísticamente significativos, como se indica mediante los triples asteriscos *** en la consola de R (estadísticamente significativo a un nivel alfa de 0,001), implementa los análisis post hoc de seguimiento como se muestran en la figura a continuación.

Figura 12: Análisis post hoc Tukey HSD para examinar comparaciones entre grupos.



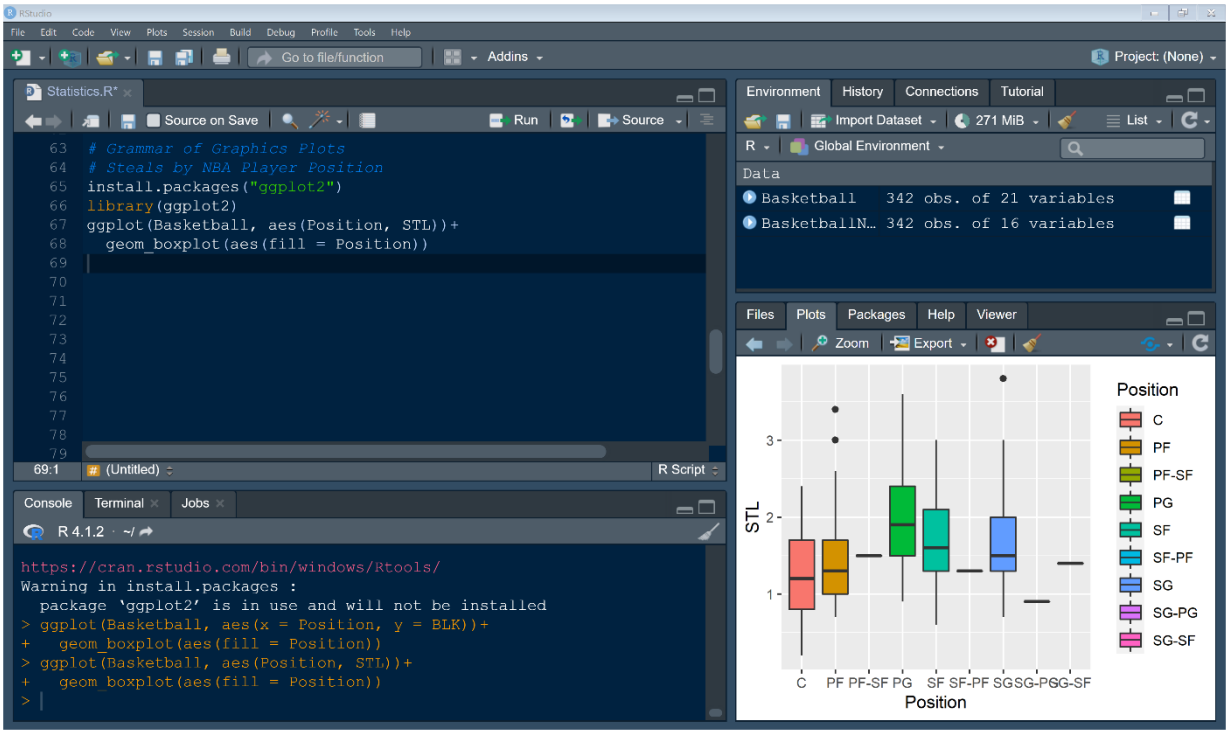
Fuente: Captura de pantalla de RStudio realizada por el autor (RStudio, 2022).

Como se muestra en la figura anterior, podemos ver que hubo diferencias significativas entre la posición de ala-pívot y la posición de pívot en robos, como se muestra en la consola de R. Observa la sección naranja resaltada en círculo en la que se detalla: la posición seguida de la diferencia, seguida del rango inferior y superior, y luego seguida del valor p ajustado. Si el valor p es menor que el alfa especificado (que generalmente se establece en 0,05 o se predetermina en 0,05), puedes afirmar que hay una diferencia estadísticamente significativa entre los grupos.

Recuerda lo siguiente: Alfa es el umbral elegido por el científico del deporte o el científico de datos como la capacidad aceptable para cometer un error de Tipo I (un error falso positivo), también conocido como 1 - Intervalo de Confianza.

A continuación, se muestra un ejemplo de representación gráfica de las diferencias entre las posiciones de los jugadores en robos utilizando el paquete Grammar of Graphics en R. La estructura de la plantilla es la siguiente: ggplot(conjunto de datos, aes(variable x, variable y)) + geom_boxplot()

Figura 13: Porcentaje de robos (STL) por posición del jugador, emulando los hallazgos del ANOVA.

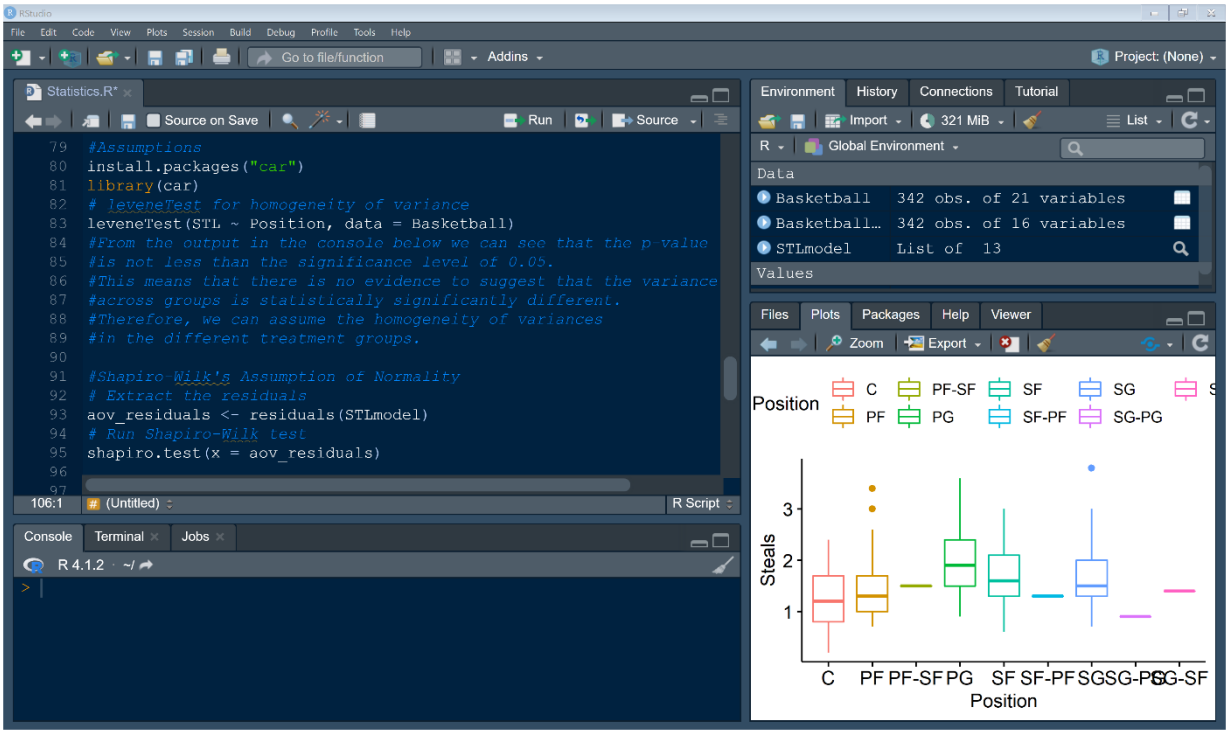


Fuente: Captura de pantalla de RStudio realizada por el autor (RStudio, 2022).

Si deseas obtener un gráfico diseñado específicamente para mostrar diferencias significativas, deberás instalar un paquete llamado ggpubr y cargar la biblioteca como se muestra en la figura a continuación.

Se recomienda realizar una comprobación final donde se examinen las suposiciones y se tome una determinación final sobre si implementar el ANOVA o una versión no paramétrica, como la prueba de Kruskal-Wallis. Ver la figura a continuación.

Figura 14: Control de calidad de las suposiciones del ANOVA.

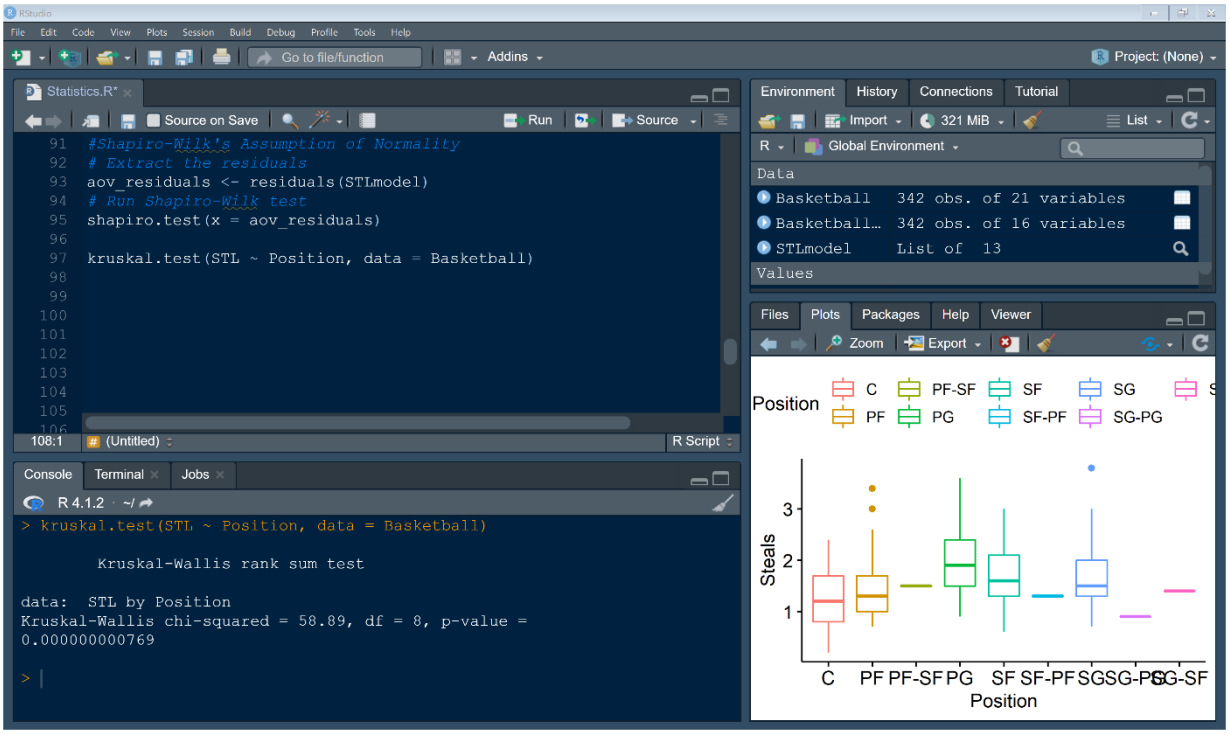


Fuente: Captura de pantalla de RStudio realizada por el autor (RStudio, 2022).

Como vemos arriba, la suposición de homogeneidad de varianza se verifica mediante la implementación de la prueba de Levene con la función leveneTest() y arroja un valor p que no es significativo, lo que indica que no hay diferencias estadísticamente significativas entre los grupos en términos de homogeneidad de varianza.

Luego, examinamos la suposición de normalidad mediante la implementación de la prueba de Shapiro-Wilk, que se realiza extrayendo los residuos de tu modelo ANOVA e implementando la función Shapiro.test(), como se muestra en la figura anterior. Para resumir sobre las pruebas de suposiciones, no queremos valores p que sean menores que alfa, deben ser mayores para continuar con el modelo ANOVA. Si alguna de las suposiciones falla, entonces podemos implementar una evaluación no paramétrica. La versión no paramétrica del ANOVA es la prueba de Kruskal-Wallis, que se implementa con la función Kruskal.test() en R, como se muestra en la figura anterior.

Figura 15: Suposición de normalidad

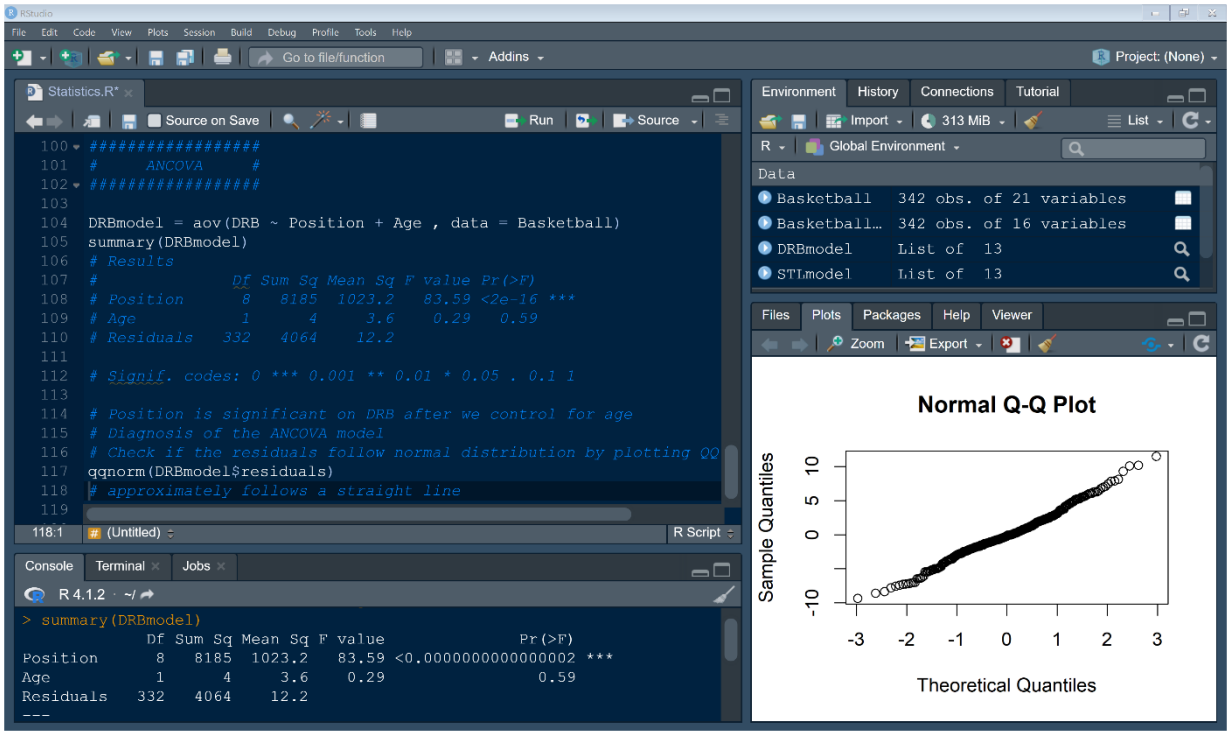


Fuente: Captura de pantalla de RStudio realizada por el autor (RStudio, 2022).

El análisis de covarianza (ANCOVA) se recomienda cuando las variables independientes son categóricas y la variable dependiente es continua, pero deseas controlar una posible variable de confusión. Una variable de confusión es un factor que puede contribuir a la varianza en la variable independiente. Por lo tanto, para obtener una medida real de la varianza explicada por la variable independiente en la variable dependiente, debes controlar esta tercera variable. Por ejemplo, si quisieras examinar las diferencias entre los planes de entrenamiento y el rendimiento de los equipos de baloncesto, es importante controlar los presupuestos invertidos en el entrenamiento de los jugadores (covariable). Esto es importante porque la cantidad de dinero invertido en entrenamiento y equipamiento puede marcar la diferencia en la capacitación, el equipo, la fisioterapia y el acondicionamiento físico de los jugadores, todos los cuales influyen indirectamente en el plan de entrenamiento y el rendimiento. Para ejecutar el modelo ANCOVA, se deben cumplir las seis suposiciones de un modelo ANOVA. Además, también deben cumplirse las siguientes tres suposiciones: La covariable (tercera variable) debe estar relacionada de manera lineal con el resultado para cada grupo de variables independientes. ¿Qué significa todo esto? ¿Cuál es el fin de comprobar esto? La suposición de que la covariable esté asociada de manera lineal con la variable dependiente se puede verificar con un simple gráfico de dispersión. Si encuentras que la relación es no lineal, ANCOVA no es la elección óptima para analizar tus datos. La suposición de homocedasticidad verifica que el término de error en la relación entre la variable independiente y la dependiente sea igual en todos los niveles de las variables independientes. Finalmente, la suposición de homogeneidad de las pendientes de regresión tiene como objetivo determinar si las pendientes, aunque son diferentes, son paralelas entre sí, ya que esto indica que no hay evidencia de interacción entre la covariable y la variable independiente

si las pendientes son paralelas. El modelo ANCOVA es excelente para comparar diferencias entre grupos, como equipos o posiciones de jugadores, en relación con un resultado específico, mientras controlas simultáneamente una variable interviniente. Ver la figura a continuación que muestra un ejemplo en el que se examina los rebotes defensivos por posición de jugador, con la edad como covariable. También se verifica la normalidad con la función qqnorm(). Si los círculos siguen la línea recta, entonces los datos son normales.

Figura 16: ANCOVA, rebotes defensivos (DRB) por posición del jugador con la edad como covariable.



Fuente: Captura de pantalla de RStudio realizada por el autor (RStudio, 2022).

Los modelos multivariados están diseñados específicamente para evaluar múltiples variables. En la industria, algunos llaman a los modelos con varias variables independientes "multivariados", pero otros se refieren a ellos cuando hay múltiples variables dependientes que están correlacionadas. Por ejemplo, supongamos que estás interesado en examinar las diferencias entre tres posiciones en el fútbol americano (mariscal de campo, corredor y liniero defensivo) en carreras de 40 yardas, carreras de shuttle y pruebas de agilidad 5-10-5. Se prefiere un análisis multivariado porque las variables dependientes están correlacionadas. Sin embargo, parece haber alguna correlación ya que todas son medidas de potencia anaeróbica (algunas combinadas con velocidad lineal y otras con cambios de dirección). En este caso, un análisis multivariado de la varianza (MANOVA) es la elección óptima. Por otro lado, supongamos que deseas analizar las diferencias en estas mismas posiciones de jugadores en la carrera de 40 yardas, el press de banca de 102 kilos y la prueba de habilidad cognitiva Wonderlic. El MANOVA no es el modelo correcto para usar. ¿Por qué? Porque las variables dependientes no están correlacionadas. La carrera de 40 yardas es una medida de potencia

anaeróbica y velocidad, mientras que el press de banca de 102 kilos es una evaluación de la fuerza muscular del tren superior, y la prueba Wonderlic es una evaluación de una aptitud que no está para nada relacionada. Nuevamente, los resultados obtenidos con este modelo te dirán que hay diferencias significativas entre los grupos, pero no entre qué pares de grupos. Se requieren análisis adicionales para determinar información más específica. Si decides que el modelo MANOVA es apropiado, basado en el tipo de datos y el número de variables dependientes y la correlación entre ellas, debes cumplir con un total de nueve suposiciones. La suposición de independencia de las observaciones es estándar. Una suposición crítica al usar análisis multivariados es el tamaño de muestra mínimo necesario para una potencia suficiente para analizar los datos con este modelo particular.

Se prefieren tamaños de muestra grandes. Por lo general, es preferible tener un tamaño de muestra lo más grande posible. El uso de MANOVA o MANCOVA requiere que haya más sujetos en cada grupo de variables independientes que el número total de variables dependientes.
Otra suposición es que no haya valores atípicos univariados o multivariados. Los "valores atípicos univariados" son valores atípicos dentro de cada grupo de las variables independientes, en comparación con los valores atípicos multivariados, que se refieren a los de las variables dependientes. Los valores atípicos (univariados) se evaluan utilizando diagramas de caja y los multivariados utilizando la distancia de Mahalanobis. Otra suposición es la de normalidad multivariada, que se verifica mediante la prueba de normalidad de Shapiro-Wilk. Además, el MANOVA requiere la suposición de una relación lineal entre las variables dependientes para todas las variables independientes, lo que generalmente se puede verificar mediante una matriz de gráficos de dispersión simple. También está la suposición de homogeneidad de las matrices de varianza-covarianza, que se verifica mediante la prueba de Box's M para la igualdad de covarianza. La última suposición necesaria para ejecutar este tipo de análisis multivariado llama a la ausencia de multicolinealidad, lo que significa que no debe haber una correlación demasiado fuerte entre las variables dependientes. Puede parecer contradictorio, pero para ejecutar MANOVA o MANCOVA, debes tener múltiples variables dependientes que estén moderadamente correlacionadas. Si la correlación es demasiado baja, es mejor evaluar individualmente las variables dependientes mediante ANOVA. Y si la correlación es demasiado fuerte, puede haber un problema de multicolinealidad.

Los MANOVA también se pueden usar para evaluar una variable en particular en varios momentos diferentes. Este modelo se llama MANOVA de medidas repetidas. Un ejemplo de cuándo usar este modelo es al evaluar la potencia muscular antes de la temporada, durante la temporada y después de la temporada porque hay al menos tres o más momentos diferentes. Al usar un MANOVA de medidas repetidas, es importante consultar Lambda de Wilks y la significancia de las pruebas multivariadas globales. Si no encuentras significación, tu análisis está completo. Sin embargo, si la encuentras, debes realizar un seguimiento

utilizando ANOVA univariados y determinar si las pruebas de efectos entre sujetos (en las variables independientes) son significativas. Es un protocolo ejecutar una corrección de Bonferroni después de encontrar significación en las pruebas multivariadas y entre sujetos, para corregir el número de ANOVA realizados.

Una prueba adicional que es similar a las correlaciones es la prueba de X^2 de independencia, también llamada chi-cuadrado de Pearson. Esta difiere de la conocida correlación de Pearson ya que se utiliza para examinar las relaciones entre dos variables categóricas (no continuas). Esta prueba requiere que solo se cumplan dos suposiciones: que ambas variables sean categóricas y hay al menos dos grupos independientes. Supongamos que deseas explorar la relación entre los dos equipos de fútbol Real Madrid y Barcelona (considerados una variable categórica que consta de dos grupos) en los tiros de penalti anotados frente a los tiros de penalti fallados durante una temporada. Este debería ser el análisis elegido. La prueba de X^2 es adecuada para explorar este tipo de datos, especialmente porque la variable dependiente es de naturaleza dicotómica (tiros de penalti anotados/tiros de penalti fallados). Normalmente, la salida de datos, dependiendo del software utilizado para analizarlos, mostrará una sección de tablas cruzadas y los resultados de la prueba de chi-cuadrado.

La figura a continuación muestra modelos estadísticos, tipos de datos y variables, y el tipo de preguntas sobre el rendimiento deportivo que se pueden responder con cada tipo de modelo.

Figura 17: Modelos estadísticos, tipos de datos y variables, y preguntas sobre el rendimiento deportivo.

Statistical Model	Data and Variables	Questions Answered by the Statistical Model
Chi-square	One or more categorical variables	Are basketball players more susceptible to injuries than baseball players? (Are two categorical variables related?)
t-test	Dichotomous independent variable for groups, one continuous dependent variable variable	Are there differences between the New England Patriots and the Miami Dolphins on touchdowns scored? (Do differences exist between two groups on a dependent variable?)
ANOVA	One or more categorical independent variables, one continuous dependent variable	Are there differences between the sports of basketball, tennis, and soccer on athletes' salaries? (Do differences exist between two or more groups on one continuous dependent variable?)
ANCOVA	One or more categorical independent variables, one continuous dependent variable, and one or more control variables	Are there differences between the sports of basketball, tennis, and soccer on athletes' salaries after controlling for ticket sales? (Do differences exist between two or more groups after controlling for a covariate on one dependent variable?)
MANOVA	One or more categorical independent variables, two or more continuous dependent variables	Are there differences between basketball player positions; center, point guard, and power forward on field goals, rebounds, and assists? (Do differences exist between two or more groups on multiple dependent variables?)
MANOVA with Repeated Measures	One or more categorical independent variables, two or more continuous dependent variables, with the dependent variables being repeated measures of the same attribute	Are there differences between basketball player positions; center, point guard, and power forward on field goals, rebounds, and assists at at pre season, during the season, and post season? (Do differences exist between two or more groups on multiple dependent variables over different time points?)
MANCOVA	One or more categorical independent variables, two or more continuous dependent variables, and one or more control variables	Are there differences between basketball player positions; center, point guard, and power forward on field goals, rebounds, and assists after controlling for minutes played? (Do differences exist between two or more groups after controlling for a covariate on multiple dependent variables?)

Fuente: Martin, 2016

Statistical Model Chi-square t-test ANOVA ANCOVA MANOVA MANOVA with Repeated Measures MANCOVA	Modelo estadístico X^2 Prueba t ANOVA ANCOVA MANOVA MANOVA de medidas repetidas MANCOVA
Question Answered by the Statistical Model Are basketball players more susceptible to injures than baseball players? (Are two categorical variables related?) Are there differences between the New England Patriots and the Miami Dolphins on touchdowns scored? (Do differences exist between two groups on a dependent variable?) Are there differences between the sports of basketball, tennis, and soccer on athletes'	Pregunta respondida por el modelo estadístico: ¿Los jugadores de baloncesto son más susceptibles a lesiones que los jugadores de béisbol? (¿Están relacionadas dos variables categóricas?) ¿Existen diferencias entre los New England Patriots y los Miami Dolphins en la cantidad de touchdowns anotados?

salaries? (Do differences exist between two or more groups on one continuous dependent variable?) Are there differences between the sports of basketball tennis, and soccer on athletes' salaries after controlling for ticket sales? (Do differences exist between two or more groups after controlling for a covariate on one dependent variable?) Are there differences between basketball player positions centre, point guard, and power forward on ficid goals rebounds, and assists? (Do differences exist between two or more groups on multiple dependent variables?) Are there differences between basketball player positions centre, point guard, and power forward on field goals rebounds, and assists at pre-season, during the season, and post season? (Do differences exist between two or more groups on multiple dependent variables over different time points?) Are there differences between basketball player positions centre, point guard, and power toward on field goals rebounds. and assists after controlling for minutes played? (Do differences exist between two or more groups after controlling for a covariate on multiple dependent variables?)	(¿Existen diferencias entre dos grupos en una variable dependiente?) ¿Existen diferencias en los salarios de los deportistas entre los deportes de baloncesto, tenis y fútbol? (¿Existen diferencias entre dos o más grupos en una variable dependiente continua?) ¿Existen diferencias en los salarios de los deportistas entre los deportes de baloncesto, tenis y fútbol después de controlar las ventas de entradas? (¿Existen diferencias entre dos o más grupos después de controlar una covariable en una variable dependiente?) ¿Existen diferencias entre las posiciones de los jugadores de baloncesto (pívot, base y alero) en goles de campo, rebotes y asistencias? (¿Existen diferencias entre dos o más grupos en múltiples variables dependientes?) ¿Existen diferencias entre las posiciones de los jugadores de baloncesto (pívot, base y alero) en goles de campo, rebotes y asistencias antes de la temporada, durante la temporada y después de la temporada? (¿Existen diferencias entre dos o más grupos en múltiples variables dependientes en diferentes momentos?) ¿Existen diferencias entre las posiciones de los jugadores de baloncesto (pívot, base y alero) en goles de campo, rebotes y asistencias después de controlar los minutos jugados? (¿Existen diferencias entre dos o más grupos después de controlar una covariable en múltiples variables dependientes?)

Ten en cuenta que los scripts de R y los conjuntos de datos acompañan a estos módulos para llevar a cabo estos análisis.

Referencias

Allaire, J. J. (2022). R 4.2.1 [Computer Software]. RStudio, Inc. https://cran.r-project.org/index.html

Anderson, R. (2015). Modeling niches and distributions: it's not just "click, click, click". Biogeografía, 8, 4-27.

Andrews, F. T., Croke, B. F., & Jakeman, A. J. (2011). An open software environment for hydrological model assessment and development. *Environmental Modelling & Software*, *26*(10), 1171-1185.

Atkinson, G., & Nevill, A. M. (2001). Selected issues in the design and analysis of sport performance research. *Journal of sports sciences*, *19*(10), 811–827. https://doi.org/10.1080/026404101317015447

Brown, K. S., & Sethna, J. P. (2003). Statistical mechanical approaches to models with many poorly known parameters. *Physical review E*, *68*(2), 021904.

Davenport, T. H. (2006). Competing on analytics. *Harvard Business Review*, *84*(1), 98–134.

Martin, L. (2016). Sports performance measurement and analytics: The science of assessing performance, predicting future outcomes, interpreting statistical models, and evaluating the market value of athletes. FT Press.

O' Donoghue, P., & Ingram, B. (2001). A notational analysis of elite tennis strategy. *Journal of sports sciences*, *19*(2), 107–115. https://doi.org/10.1080/026404101300036299

Reid, M., & Schneiker, K. (2008). Strength and conditioning in tennis: current research and practice. *Journal of Science and medicine in Sport*, *11*(3), 248-256. https://doi.org/10.1016/j.jsams.2007.05.002

Slack, T., & Parent, M. M. (2006). *Understanding sport organizations: The application of organization theory*. Human Kinetics.