🚀 Liga Española Player Value Analysis ⚽

INFORME FINAL: RELACIÓN ENTRE CARACTERÍSTICAS DE JUGADORES Y SU VALOR DE MERCADO EN LA LIGA ESPAÑOLA

Autores:

- Joaquín Moreno
- Álvaro Yuste Valles

Repositorio de GitHub:

- Enlace: https://github.com/alyusva/LaLiga_Players_Value_Analysis

1. Introducción, Objetivos y Motivación

1.1. Introducción

El fútbol profesional es un sector que moviliza grandes inversiones en el mercado de fichajes. Con el objetivo de **optimizar** dichas inversiones, los clubes suelen recurrir a análisis basados en datos que les permitan conocer qué variables son más relevantes a la hora de **determinar el valor de mercado** de un jugador. En este trabajo se busca profundizar en la relación entre diversas características de los jugadores (tanto técnicas como demográficas y reputacionales) y su valor de mercado.

1.2. Objetivos

1. **Identificar las variables con mayor impacto en el valor de mercado.** Se pretende discernir cuáles de las características (Overall, Potential, Edad, Reputación Internacional, etc.) explican en mayor medida las diferencias en el valor de mercado.
2. **Cuantificar la relación entre las habilidades técnicas y el valor de mercado.** Se busca medir, por ejemplo, cómo varía el valor de mercado con un punto adicional en Overall o en Potential.
3. **Evaluar el efecto de la edad y la reputación internacional.** Determinar de forma explícita si la edad (más o menos de 30 años) o la reputación internacional influyen en la valoración económica del jugador.

1.3. Motivación

- **Relevancia deportiva**: Los clubes necesitan criterios objetivos y cuantitativos para guiar las decisiones de compra y venta de jugadores.
- **Aporte académico**: Existen pocos estudios estadísticos detallados en español aplicados al ámbito del fútbol y su mercado de fichajes.

- **Aplicaciones prácticas**: La construcción de modelos predictivos para estimar el valor de mercado es de utilidad para ojeadores, directores deportivos y analistas de datos en el ámbito futbolístico.

2. Fuentes de los Datos

2.1. Origen de los datos

- **Fuente**: FIFA 23 Player Dataset (Kaggle).
- Enlace: https://www.kaggle.com/datasets/kevwesophia/fifa23-official-datasetclean-data.
- **Formato**: Archivo CSV con datos de jugadores de FIFA 23.

2.2. Descripción de los datos

El conjunto de datos incluye información variada de los jugadores, tales como:

- **Variables Demográficas**: Edad, Nacionalidad, Altura, Peso.
- **Variables Técnicas**: Overall, Potential, Skill Moves, Special.
- **Variables Reputacionales**: International Reputation.
- **Variables Económicas**: Value(£), Wage(£), Release Clause(£).

• **Variables Categóricas**: Preferred Foot, Work Rate, Body Type, Position, etc.Tras una limpieza inicial (eliminación de valores faltantes y outliers), se trabaja con un número final de filas menor que el original (ver el script en la sección de código). ("Filas originales: 17660", "Filas después de limpiar: 17622")

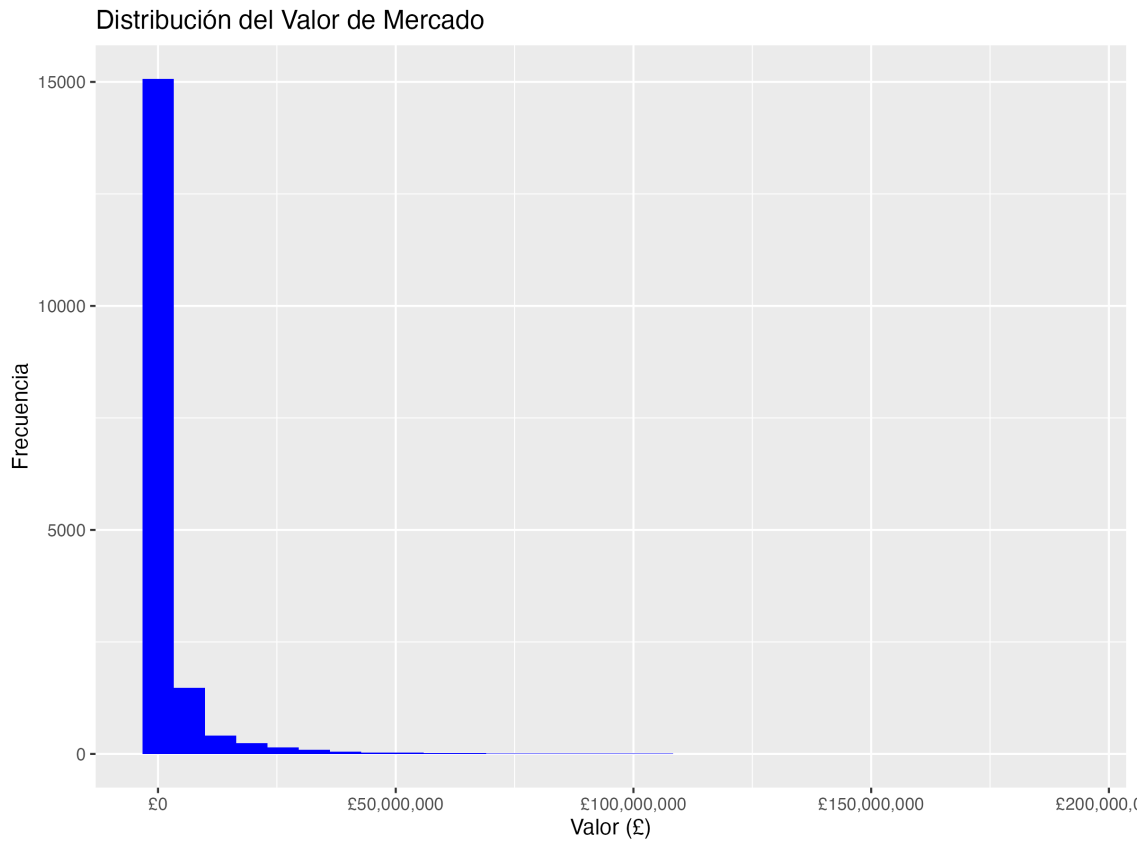
2.3. Preprocesamiento de los datos

- **Selección de variables**: Se filtran las columnas más relevantes para el análisis (Overall, Potential, Age, Value_eur, etc.).
- **Renombrado de columnas**: Se cambian nombres como `Value(£)` a `Value_eur` para mayor consistencia.
- **Eliminación de NAs**: Se descartan filas con valores faltantes para evitar problemas en el modelado.
- **Transformación logarítmica**: Se aplica `log(Value_eur + 1)` para estabilizar la varianza y mitigar la asimetría en la distribución del valor.

3. Análisis de los Datos

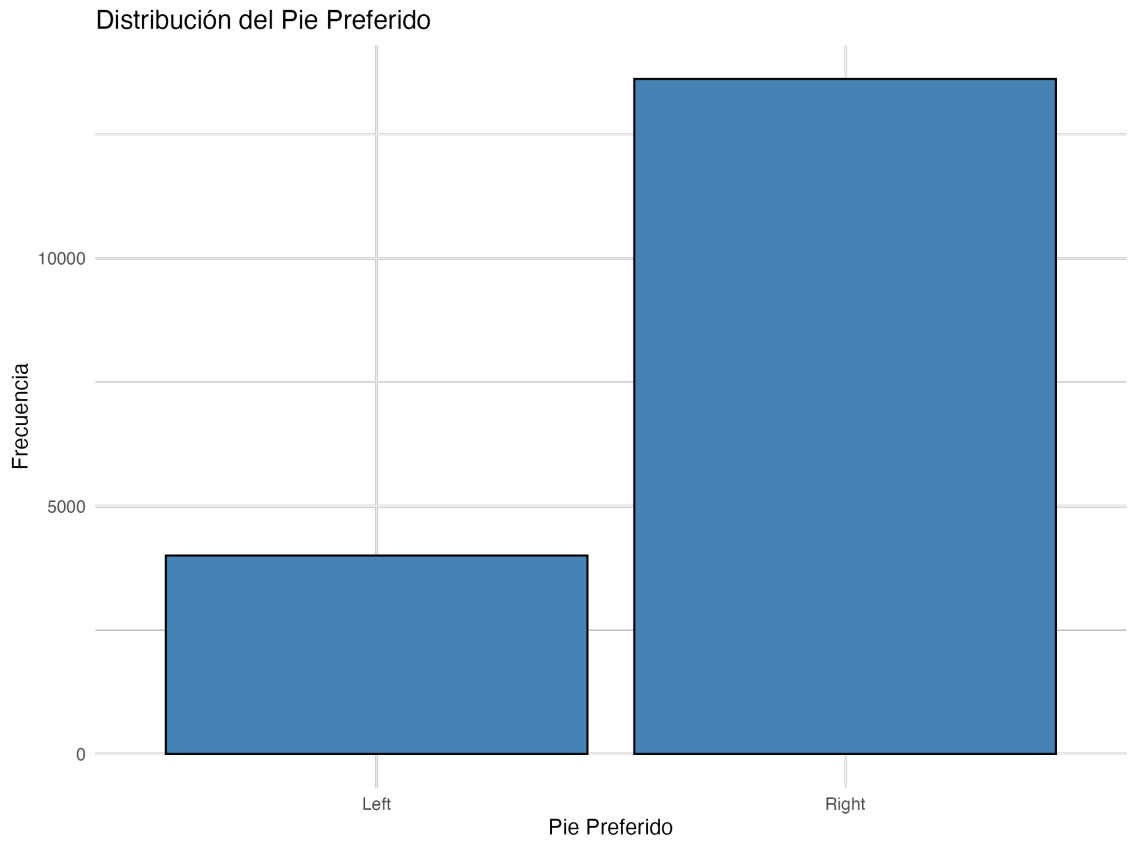
3.1. Análisis Descriptivo

1. **Histograma del valor de mercado**
   Se genera un histograma para observar la distribución de la variable `Value_eur`.



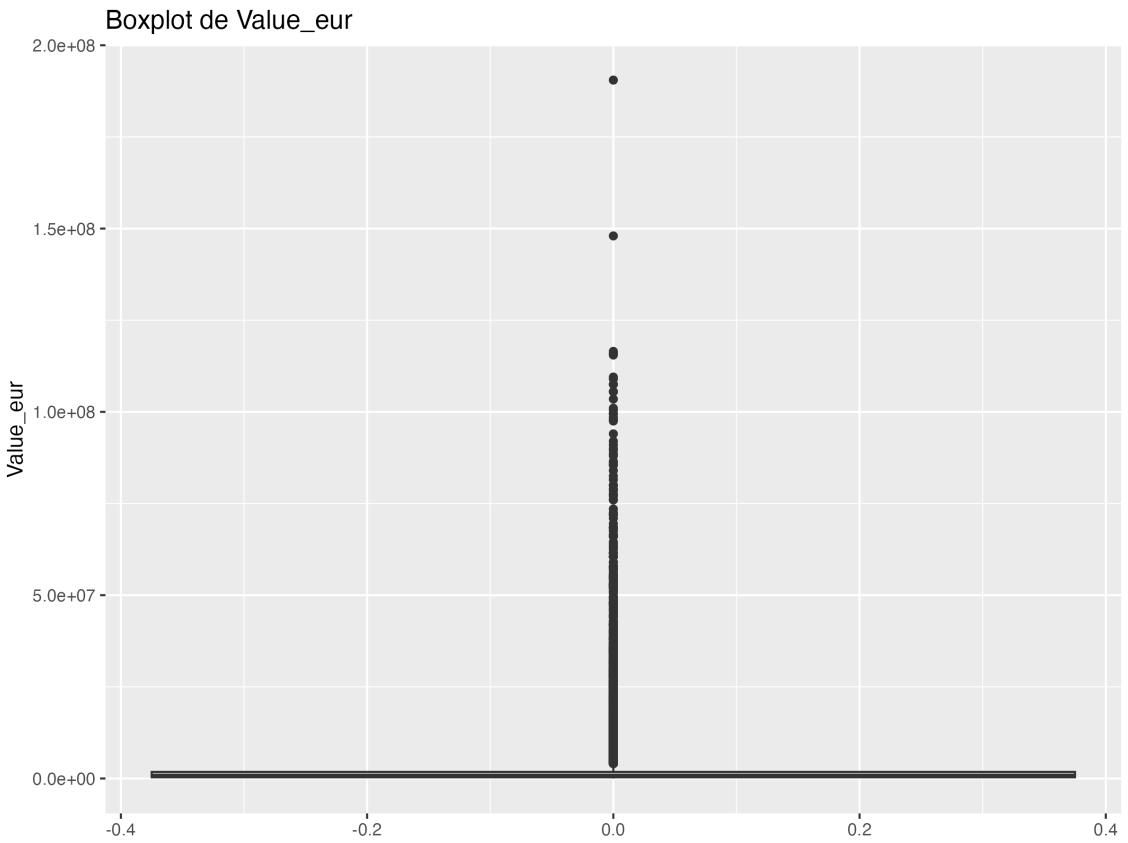
- El valor de mercado presenta una distribución asimétrica con valores muy elevados para algunos jugadores de élite.

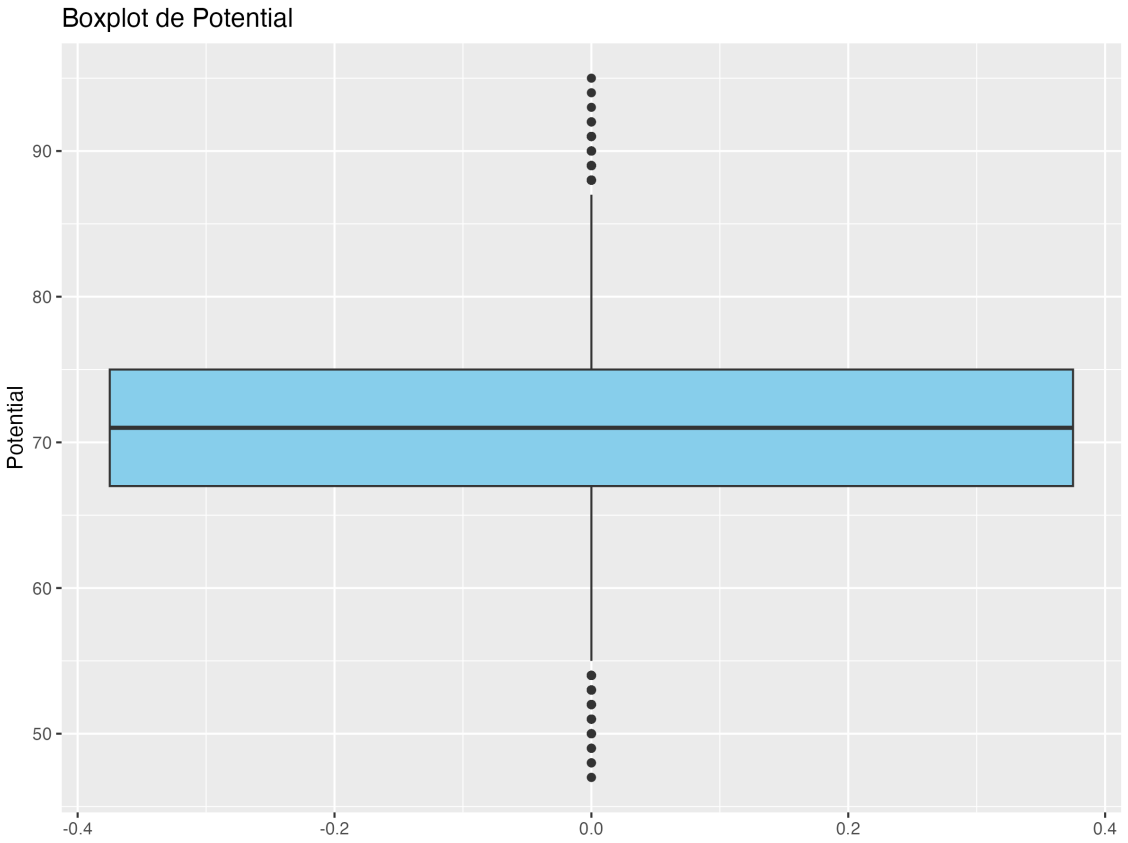
2. **Estadísticas Descriptivas**
   Mediante la función `psych::describe()`, se obtienen medidas como media, desviación estándar y percentiles para las variables numéricas.
   - Observamos, por ejemplo, la **edad promedio** de los jugadores, la media de su **Overall** y el rango de salarios.
3. **Variables Categóricas**
   - Se analiza la distribución del *Preferred Foot* (pie preferido) mediante un diagrama de barras.

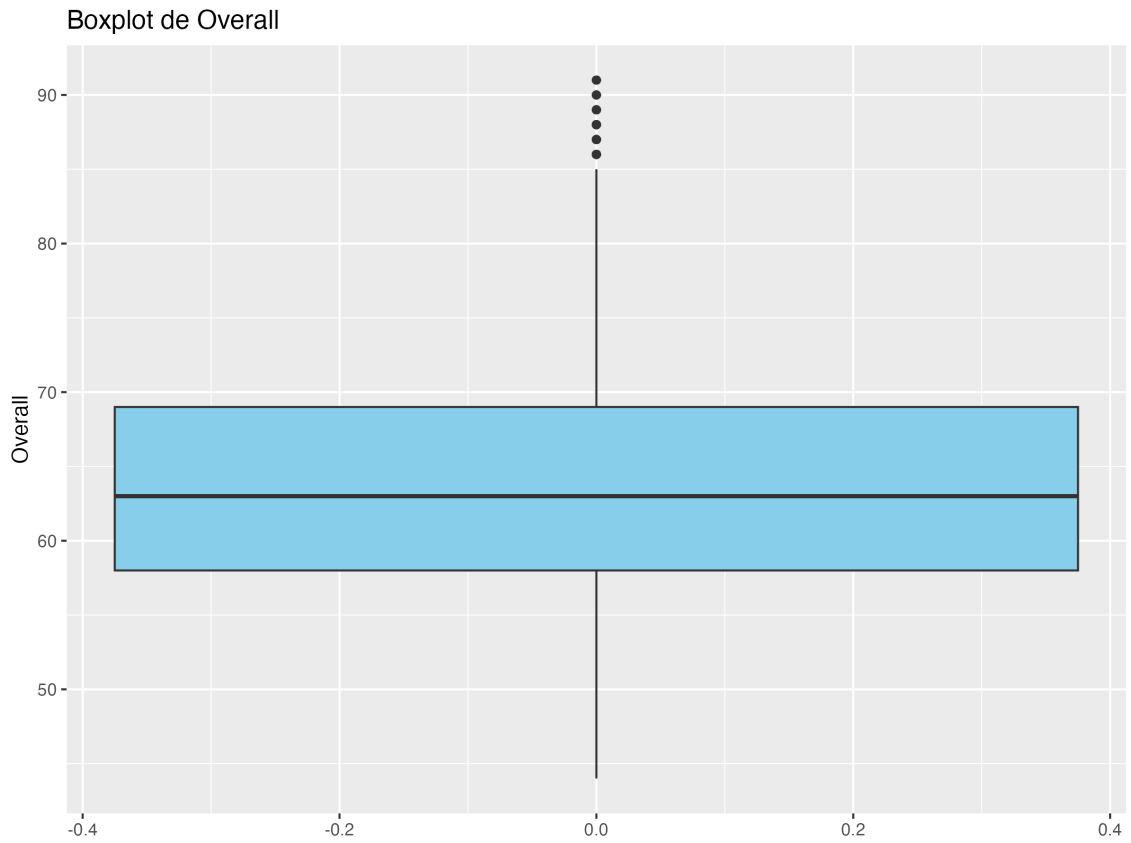


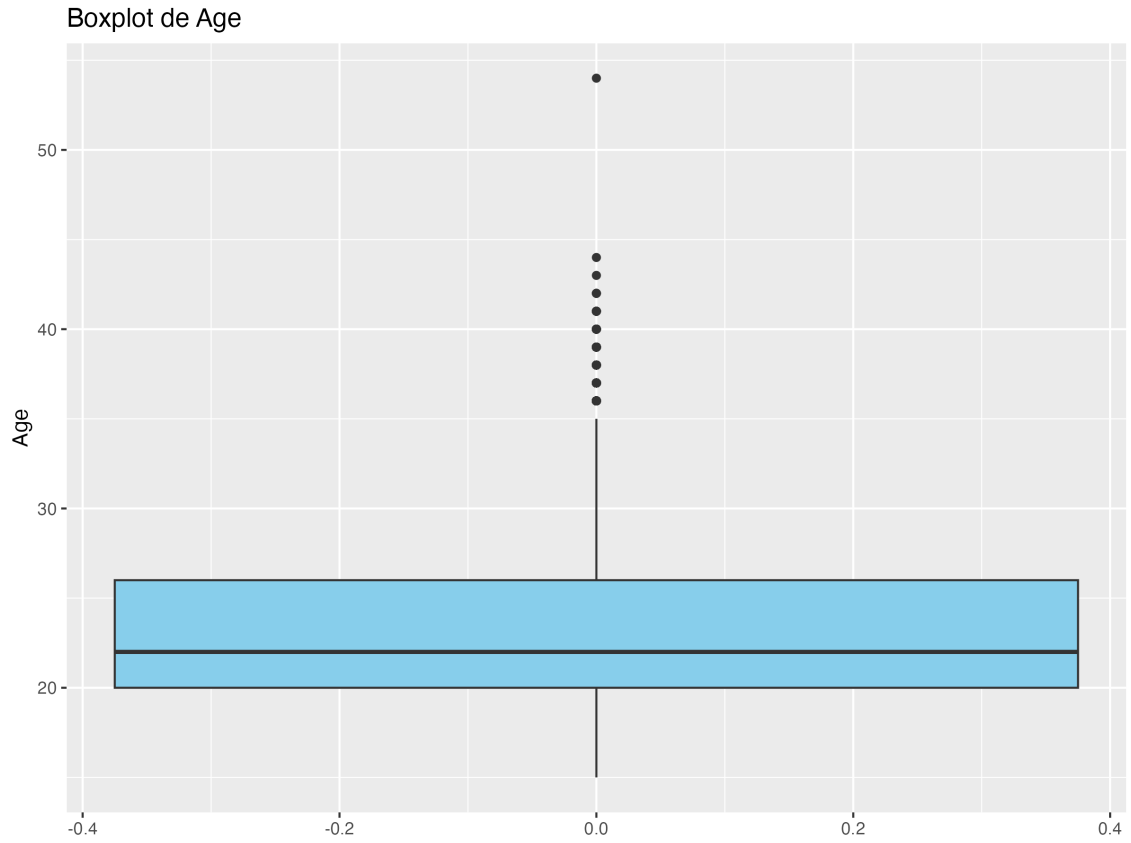
- El análisis muestra una mayor proporción de jugadores diestros frente a zurdos.

4. **Outliers**
   - Se utilizan boxplots para detectar valores atípicos en `Value_eur`, `Age`, `Overall` y `Potential`.


Boxplot de Value_eur
Value_eur
2.0e+08
1.5e+08
1.0e+08
5.0e+07
0.0e+00
-0.4
-0.2
0.0
0.2
0.4


Boxplot de Potential
Potential
90
80
70
60
50
-0.4
-0.2
0.0
0.2
0.4

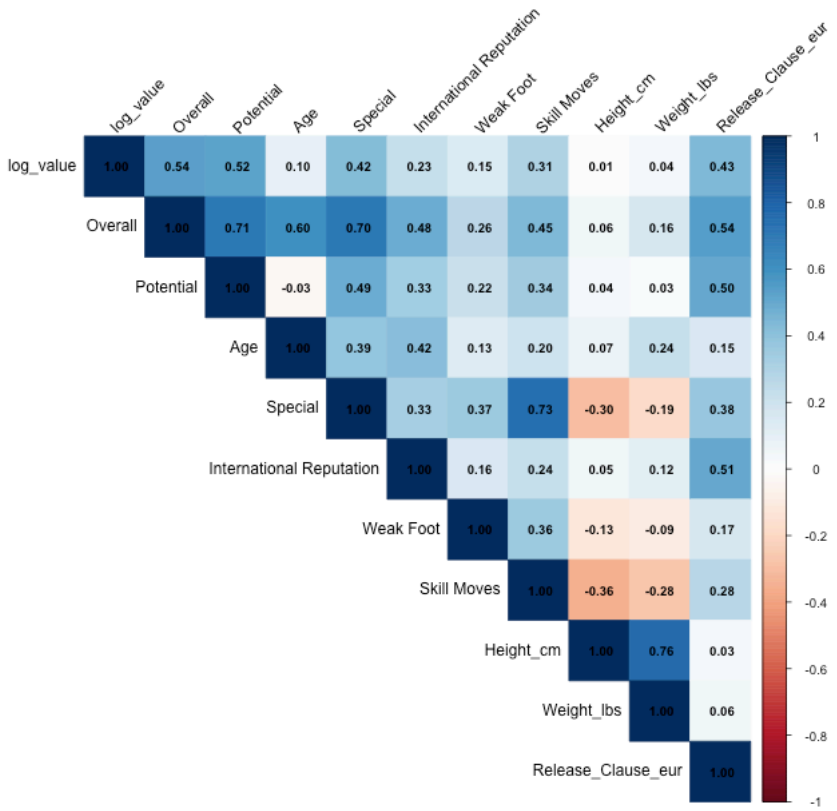




- Se aplica el criterio de rango intercuartílico (IQR) para eliminar outliers, reduciendo el dataset final. ("Filas después de eliminar outliers: 15159")

3.2. Análisis de Correlaciones

1. **Matriz de Correlación**
   - Se seleccionan variables numéricas clave (`Overall`, `Potential`, `Age`, `International Reputation`, etc.) y se calcula la matriz de correlación.
   - Se visualiza con la librería `corrplot`.



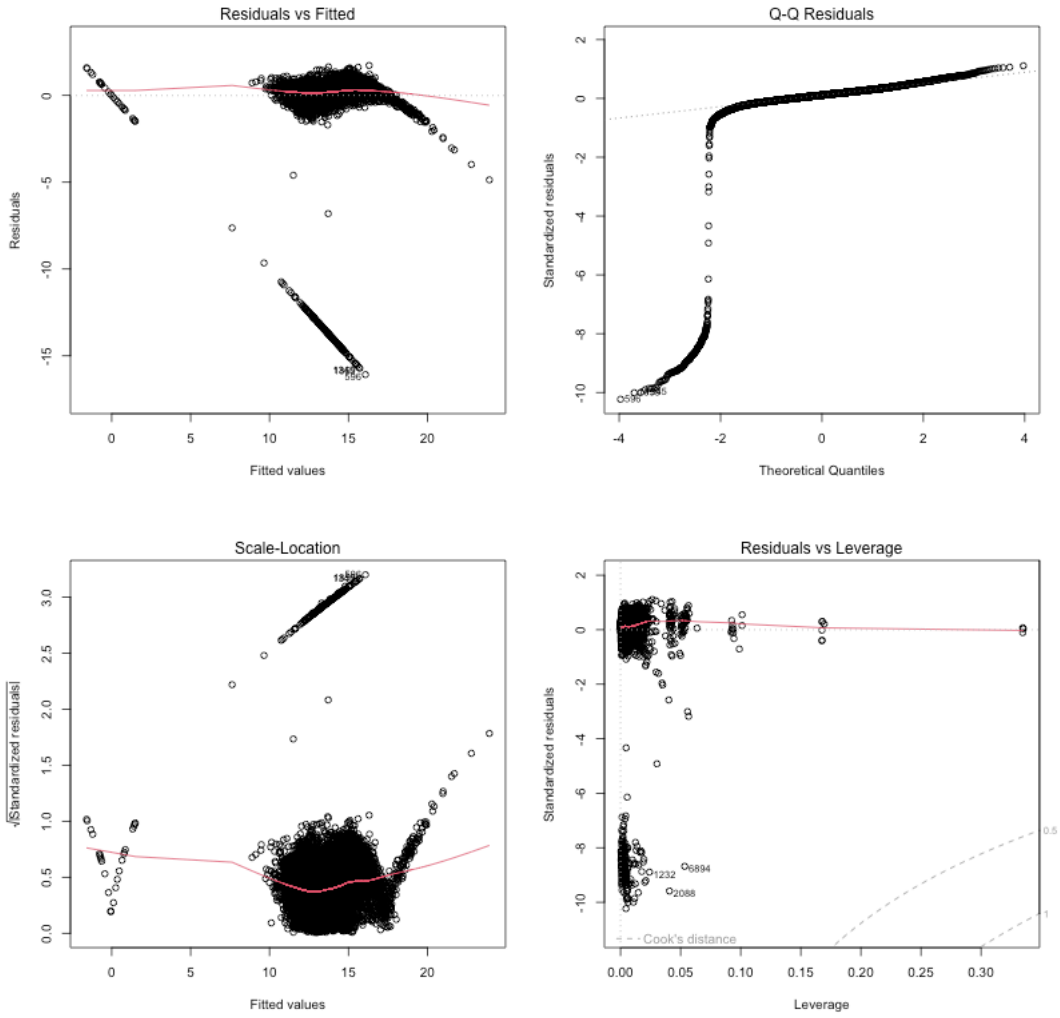
   - Observamos correlaciones moderadas y fuertes, destacando la relación entre `Overall` y `log_value`, también destacan: `Release_Clausure_eur`, `Special` y `Skill Moves`.

2. **Transformación Logarítmica**
   - Para reducir la heterocedasticidad, se define `df$log_value = log(Value_eur + 1)`.
   - Esta variable transformada es la que se utilizará en los modelos de regresión y Random Forest.

4. Modelado y Validación

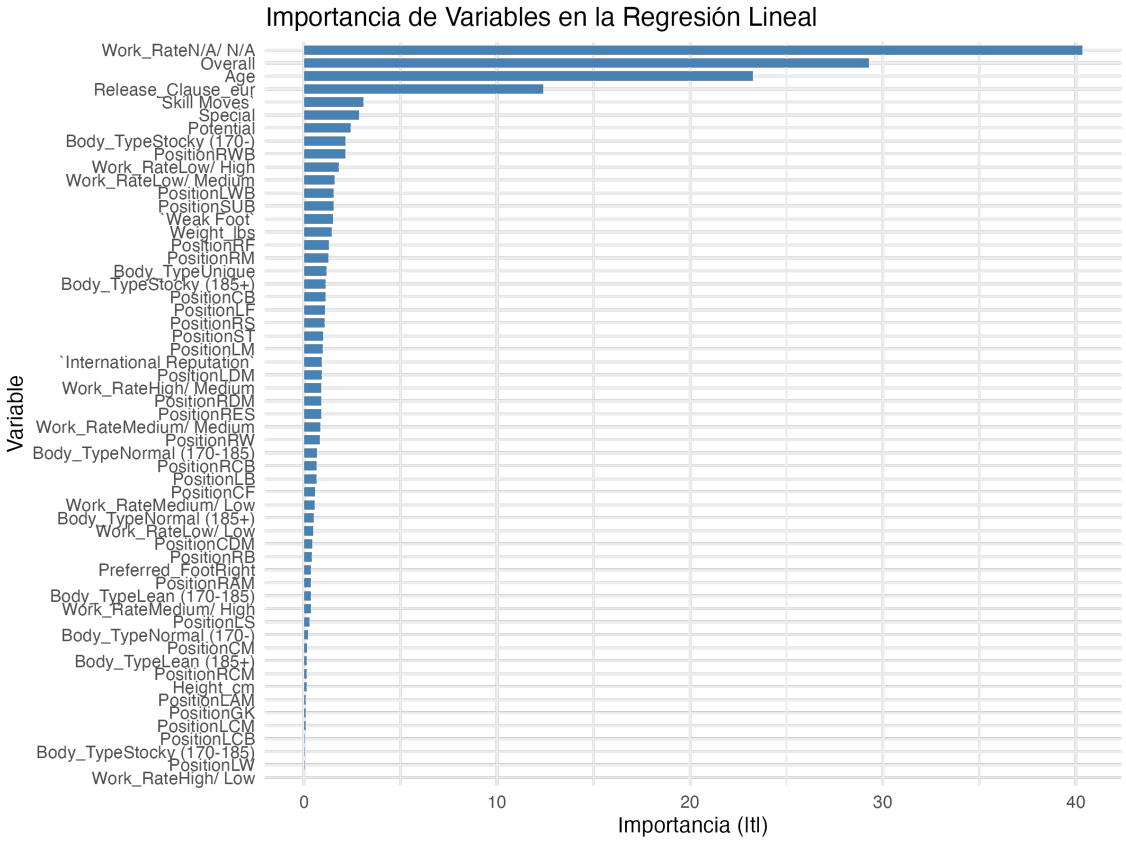
4.1. Regresión Lineal Múltiple

1. **División de Datos**
   - Se separa el conjunto de datos en *train* (80%) y *test* (20%) con la función `createDataPartition()`.
2. **Construcción del Modelo**
   - El **resumen** del modelo muestra los coeficientes y su significancia estadística.
3. **Validación Cruzada**
   - Se usa 5-fold CV con `caret::train()` para estimar la robustez del modelo.
   - Se obtiene un RMSE medio en la validación cruzada.
4. **Predicciones y Diagnósticos**
   - Con `predict()`, se obtienen los valores estimados en el conjunto de *test* y se calcula el RMSE (Error Cuadrático Medio de la Raíz).
   - Se generan gráficos de diagnóstico (`plot(final_model)`) para evaluar la linealidad, normalidad de residuos y posibles valores influyentes.



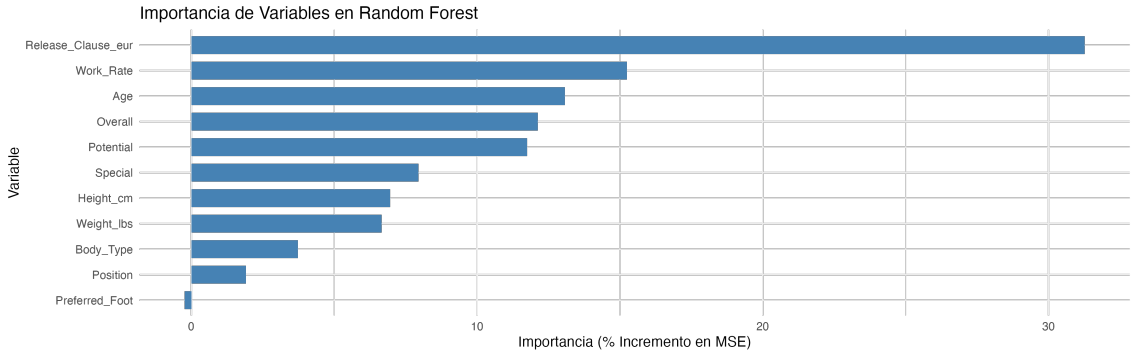
5. **Importancia de Variables (Regresión Lineal)**

- Con varImp(final_model, scale = FALSE), se obtiene un ranking basado en el valor absoluto de la *t* de cada predictor.
- Si hay factores con muchos niveles, cada dummy aparece por separado. Se pueden agrupar manualmente o filtrar para simplificar.



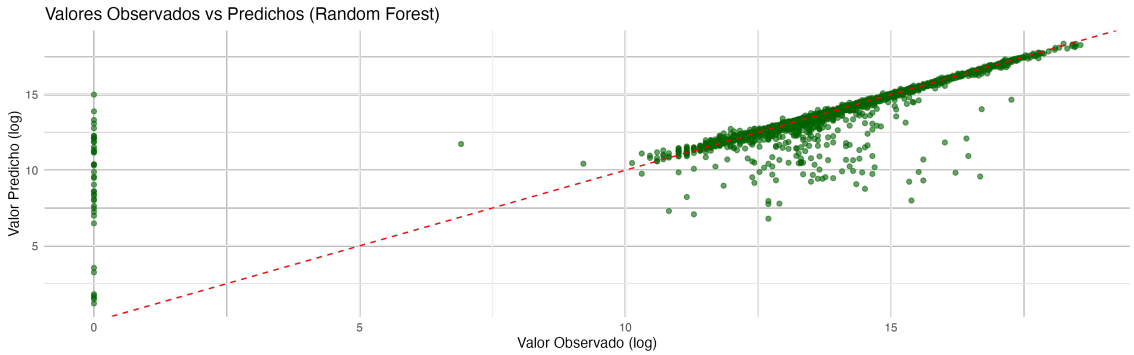
4.2. Random Forest

1. **Construcción del Modelo**
   - Se emplea la librería `randomForest`, entrenando un modelo con `ntree = 100`.
   - Este enfoque es **más robusto** frente a outliers y no requiere supuestos de normalidad en los residuos.
2. **Importancia de Variables**
   - Con `importance(rf_model)` se obtiene el %IncMSE, que mide el impacto de cada variable en la predicción.



- De nuevo, `Overall` y `Release_Clausure_eur` aparecen entre las más influyentes, pero también destacan: `Work_Rate` y `Age`

3. **Predicciones y Rendimiento**
   - Se calculan predicciones sobre el conjunto de *test* y se obtiene el RMSE.

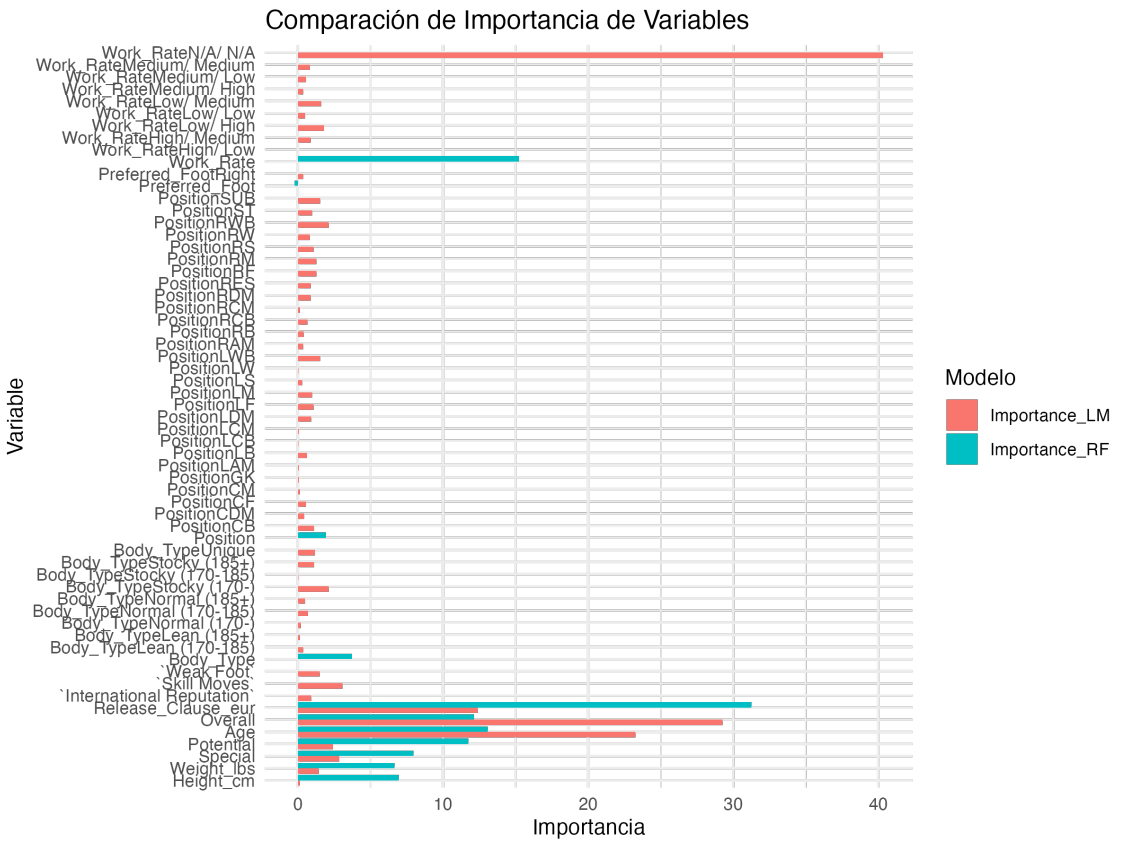


4.3. Comparación de Modelos

1. **Métricas de Rendimiento**
   - Se compara el RMSE de ambos modelos.

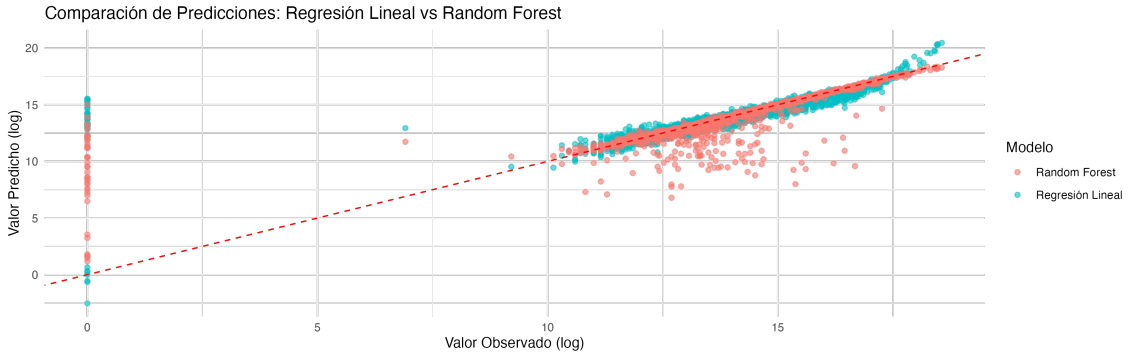
	Modelo	RMSE	R2
1	Regresión Lineal	1.442862	0.4480548
2	Random Forest	1.226825	*NA*

2. **Comparación de Importancia de Variables**
   - Se unen los resultados de `varImp()` (para la regresión) y `importance()` (para el Random Forest).



Se aprecia que `Release_Clause_eur` domina en el modelo random forest, reflejando su fuerte correlación con el valor de mercado. Por otro lado, variables como `Overall` o `Age` exhiben una relevancia más moderada pero consistente en ambos enfoques, lo que indica que, aunque la cláusula de rescisión concentra gran parte de la variación explicada, el rendimiento y la edad del jugador también influyen de manera significativa. Además, la dispersión de valores para las variables categóricas (como `Work_Rate` o `Position`) sugiere que ciertos niveles concretos pueden aportar información relevante, mientras que otros apenas impactan la predicción del valor de mercado.

3. **Valores Observados vs. Predichos**
    - Se grafican las predicciones de ambos modelos contra los valores observados.



5. Conclusiones Finales

Objetivo 1: Identificar las variables con mayor impacto en el valor de mercado

- El análisis de correlaciones y la importancia de variables en ambos modelos muestran que:
    - **Overall** es la variable con mayor correlación con el log del valor de mercado (r = 0.54).
    - **Reputación Internacional** también destaca, con r = 0.23.
    - **Potential** y **Skill Moves** presentan correlaciones positivas moderadas (r = 0.52 y r = 0.31, respectivamente), aunque con menor peso.
    - **Random Forest** (con la variable `Release_Clause_eur` incluida) muestra que la **cláusula de rescisión** sobresale como la variable más influyente. Esto es coherente con la realidad del fútbol, donde dicha cláusula está estrechamente relacionada con el valor de mercado
    - En la **Regresión Lineal**, si se incluyen todas las variables (incluyendo `International Reputation`), a menudo se observa que `Overall` y `Release_Clause_eur` presentan coeficientes significativos y altos valores de *t*.

Objetivo 2: Cuantificar la relación entre habilidades técnicas y el valor

- El modelo de **regresión lineal simple** para Overall explica alrededor del 29% de la variabilidad del log(valor) ($R^2 = 0.29$).
- Cada punto adicional en Overall incrementa el log(valor) en ~0.14 unidades, lo que confirma la importancia de las habilidades técnicas en la valoración.
- **Potential** también contribuye positivamente, pero su efecto puede ser menor que `Overall` o puede verse eclipsado por la presencia de la cláusula de rescisión.
- **Skill Moves** suele tener menor importancia que `Overall` y `Potential`, aunque sí aparece como estadísticamente significativa en algunos modelos.

Objetivo 3: Evaluar el efecto de la edad y la reputación internacional

- **Edad**:
    - La correlación entre Edad y log(valor) es r = 0.10, de signo positivo muy leve (se esperaba negativa, pero depende del rango de datos).
    - El modelo univariado indica que cada año adicional reduce el log(valor) en 0.044 unidades, lo cual sugiere un impacto negativo (aunque moderado).
- **Reputación Internacional**:
    - Cada nivel adicional de reputación incrementa el log(valor) en ~1.183 unidades, siendo esta una de las variables más influyentes después de Overall.
    - En el Random Forest, si se incluye, suele tener un **%IncMSE** elevado, aunque quizás inferior a `Release_Clause_eur`

Resultados del Modelado Predictivo

- **Regresión Lineal Múltiple**:
  - R^2 ajustado = 0.45 en el conjunto de entrenamiento.
  - RMSE en *test* = 1.443.
  - Interpretación más sencilla de los coeficientes.
- **Random Forest**:
  - RMSE en *test* = 1.227, inferior al de la regresión lineal.
  - Muestra mayor robustez frente a outliers y no requiere supuestos de normalidad.

Comparación de Modelos

El Random Forest presenta un **RMSE inferior**, lo que sugiere que captura mejor la complejidad de los datos. Sin embargo, la Regresión Lineal aporta interpretabilidad (coeficientes) que puede ser valiosa para entender la relación de cada variable con el valor.

6. Limitaciones y Líneas Futuras

1. **Limitaciones**
   - El dataset proviene de FIFA 23 y, aunque es representativo, no es un reflejo perfecto de la realidad económica de todos los jugadores.
   - Existen variables cualitativas (ej. lesiones, historial disciplinario) que no se recogen en el dataset y podrían mejorar el modelo.
   - Selección de Variables: Algunas variables relevantes (por ejemplo, `International Reputation`) pueden estar comentadas en el Random Forest, lo que altera la comparación.
   - Colinealidad: `Release_Clause_eur` y `Value_eur` podrían estar muy correlacionadas, reduciendo la aportación de otras variables.
2. **Líneas Futuras**
   - Explorar modelos más complejos (XGBoost, redes neuronales) y comparar su desempeño.
   - Incluir variables contextuales (rendimiento en la temporada, trofeos, minutos jugados, lesiones) que pudieran influir en el valor de mercado.
   - Investigar si existe una relación no lineal entre la edad y el valor que un polinomio o funciones spline puedan capturar mejor.
   - Separar el análisis por posiciones (porteros, defensas, etc.) para ver si cambian los factores de influencia.