VALIDACIÓN DEL MODELO DE RENTABILIDAD Y HOUSE EDGE EN UN CASINO EN LINEA

Alumno: Samuel de María Cabrera Flores -
Materia introducción a la ciencia de datos -
Profesor: Jaime Alejando Romero Sierra -
Fecha de entrega: 29/11/2025 -

INTRODUCCIÓN

DESCRIPCIÓN: Este proyecto valida el modelo de rentabilidad de un casino online analizando datos transaccionales o “Juegos” reales. Busca garantizar márgenes consistentes optimizando coeficientes y balances entre ganancias y pérdidas. Utilizando partidas con variables monetarias y de comportamiento, el análisis demostrará cómo maximizar rentabilidad gestionando riesgos para asegurar sostenibilidad financiera mediante métodos estadísticos y gráficas predictivas.

JUSTIFICACIÓN: El análisis es crucial porque pequeños cambios en márgenes impactan significativamente la rentabilidad del modelo de negocio del casino así como la retención del jugador sí sé piensa en pro del casino sin llegar a cometer actos ilegales. Los modelos no convencionales requieren validación específica para equilibrar atractivo al jugador con ganancias del casino, la optimización basada en datos permite decisiones informadas que garantizan una viabilidad financiera en un mercado altamente competitivo y regulado como lo es el mercado de los casinos en general sabiendo que se compite con super marcas de casino o relacionadas.

FUENTES DE DATOS: Dataset de registros de partidas con juegos reales en USD (dólares). El Dataset incluye variables monetarias (apuestas, ganancias, pérdidas), mecánicas de juego (jugadores, items, coeficientes) y controles de calidad (moderadores). Los datos temporales permiten análisis de tendencias y patrones de comportamiento para optimización estratégica del modelo de negocio.

METODOLOGÍA

PROCESO DE LIMPIEZA DE DATOS:

La base de datos original contenía registros de una plataforma de casino en línea, con aproximadamente 65.000 entradas y 10 columnas que capturaban transacciones, resultados de partidas y datos de usuarios. El objetivo del proyecto consistía en calcular el margen real de ganancias del casino, lo que exigía una limpieza rigurosa de los datos para garantizar precisión en los cálculos posteriores.

La base presentaba múltiples deficiencias: valores nulos distribuidos en todas las columnas, presencia del texto “Auto%#” en campos que debían ser numéricos, inconsistencias en los tipos de datos como columnas numéricas almacenadas como texto y registros duplicados o erróneos. Por ejemplo, la columna ID tenía más de 3.200 valores nulos, mientras que otras como peopleWin, peopleLost y outpay contenían cadenas de texto.

Se optó por no eliminar registros, utilizando en su lugar técnicas de imputación y corrección que preservaran la integridad y volumen de los datos. Primero, se abordaron los valores nulos en columnas numéricas como gamers, skins, money y ticks reemplazándolos con el promedio de cada columna, redondeado a dos decimales para mantener coherencia. La columna ID, al ser un identificador único, no admitía promedios; sus valores vacíos se sustituyeron por “Demo”, considerando que podían corresponder a sesiones de prueba o simulaciones.

Los valores “Auto%#” presentes en peopleWin, peopleLost y outpay se transformaron inicialmente a NaN y luego se imputaron con los promedios respectivos, asegurando que no alteraran significativamente los análisis posteriores. En la columna moderator, los valores nulos se reemplazaron por “False”, dado que este era el valor predominante. Finalmente, se ajustaron los tipos de datos, convirtiendo las columnas críticas a formato numérico para permitir operaciones matemáticas, y se tradujeron los nombres de las columnas al español para facilitar su manejo. Para datos duplicados se optó por usar el drop, pues al hacer el análisis posteriormente este podría afectar las medidas de tendencia.

Tras la limpieza, se verificó que no quedaran valores nulos, que los tipos de datos fueran consistentes y que las distribuciones numéricas fueran coherentes.

ANALISIS EXPLORATORIO DE DATOS (EDA):

1. **DESCRIPCIÓN GENERAL DE DATOS:**
    - VISIÓN GENERAL: (df.shape) El Dataset después de la limpieza contiene:
        - Registro filas: 60583
        - Variables: 10
    - SONDEO GENERAL: (df.info)
        - Antes de hacer el resumen estadístico por numéricas y categóricas, se hace uno general:
        - La primera fila de la tabla corresponde a todas las categorías del dataset
        - Las etiquetas (mean, std, min, etc.) se aplican a cada columna para describir la distribución de sus datos
        - *Count*: La cantidad de datos no nulos en esa columna (siempre 60,586, lo que significa que no hay valores faltantes).
        - mean (media): El valor promedio.
        - *std* (desviación estándar): Cuánto se dispersan los datos respecto a la media. Un valor alto indica mucha variabilidad.
        - *min* (mínimo): El valor más bajo registrado.
        - *25%* (primer cuartil): El valor por debajo del cual se encuentra el 25% de los datos.
        - *50%* (mediana): El valor que divide la base de datos en dos mitades guales. Es el punto medio.
        - *75%* (tercer cuartil): El valor por debajo del cual se encuentra el 75% de los datos.

- *max* (máximo): El valor más alto registrado

Estadística	Jugadores	Skins	Dinero	Ticks	PersonasGanaron	PersonasPerdieron	PagoTotal
count	60586	60586	60586	60586	60586	60586	60586
mean	123.0703	158.3247	285.1411	11.6355	66.6647	79.2684	272.2305
std	25.7718	30.4817	193.0986	229.2245	112.3422	182.2550	214.6467
min	0	0	0	1	0	0	0
25%	103.0000	136.0000	216.7224	1.3100	21.8299	3.7800	166.4024
50%	122.0000	157.0000	257.8850	2.0200	62.6200	36.5300	272.0300
75%	137.0000	174.0000	306.1399	4.3100	95.5574	104.7274	355.4375
max	491.0000	552.0000	5696.3270	23522.6500	21681.5920	4573.6016	25943.2010

- TIPOS DE VARIABLES:

Nombre variable	Tipo	Variable
Identificación	object	Categórica
Jugadores	float64	Numérica
Apariencias	float64	Numérica
Dinero	float64	Numérica
TiemposDeJuego	float64	Numérica
PersonasGanaron	float64	Numérica
PersonasPerdieron	float64	Numérica
PagoTotal	float64	Numérica
FechaYHora	object	Categórica
ModeradorActivo	bool	Booleano

- Nota1: Para la variable ‘ID(Juego)’ en su mayoría son valores numéricos, pero hay momentos donde se toma como si fuera una prueba “Demo” que es lo que convierte a esta variable como categórica
- Nota2: La variable ’FechaYHora’ a pesar de ser fecha no se toma como tal debido a cuestiones a la hora del formato de registro de datos

- RESUMEN ESTADISCTICO:
    - Variables numéricas:

      En la primera fila, primer apartado ‘Valor’ se encuentran los datos a encontrar, después en la primera fila todo hacia la derecha están el nombre de variable y por columna su respectivo resultado

Valor	Jugadores	Skins	Dinero	Ticks
Datos no nulos	60586.000000	60586.000000	60586.000000	60586.000000
media	123.070384	158.324791	285.141195	11.635565
desviación estándar	25.771853	30.481722	193.098637	229.224536
Valor mínimo registrado	0.000000	0.000000	0.000000	1.000000
1er Cuartil (25%)	103.000000	136.000000	216.722435	1.310000
Mediana 50%	122.000000	157.000000	257.885010	2.020000
3er Cuartil 75%	137.000000	174.000000	306.139995	4.310000
Valor máximo registrado	491.000000	552.000000	5696.327000	23522.650000

    - Para variable ID(Juego)

ID(Juego)	Conteo
Demo	3198
2115704.0	4
2100714.0	3
2097167.0	3
2101420.0	3
...	...
2119001.0	1
2106935.0	1
2097953.0	1
2103773.0	1
2131341.0	1

- Para variable ‘FechaYHora’

FechaYHora	Conteo
1999-09-09 09:09	3217
2021-08-29 23:40	6
2021-09-01 15:11	6

FechaYHora	Conteo
2021-09-03 12:31	6
2021-09-04 20:00	6
…	…
2021-08-29 07:54	1
2021-09-02 18:04	1
2021-09-07 00:35	1
2021-08-27 21:51	1
2021-09-06 23:32	1

- Nota1: Para variable ‘ID(Juego)’ las partidas demo son aquellas de las cuales se puede disponer o no, pues a la hora del análisis se ocuparán datos reales, cada juego tiene un ID único. La base de datos nos brinda información de como es la forma de operar del casino por lo que no se hace raro que haya aproximadamente 16 mil juegos, ya que en este tipo de formas de apostar hasta los mismo usuarios tienen la oportunidad de crear sus “lootboxes”, para la parte de cuando se repite el ID del juego y misma hora nos dice de diferentes rondas de juego con el mismo ID

- Nota2: Para la variable ‘FechaYHora’ sigue un camino parecido, y es muy variado pues depende del momento, del usuario entre más factores, pero podemos encontrar la fecha “1999-09-09

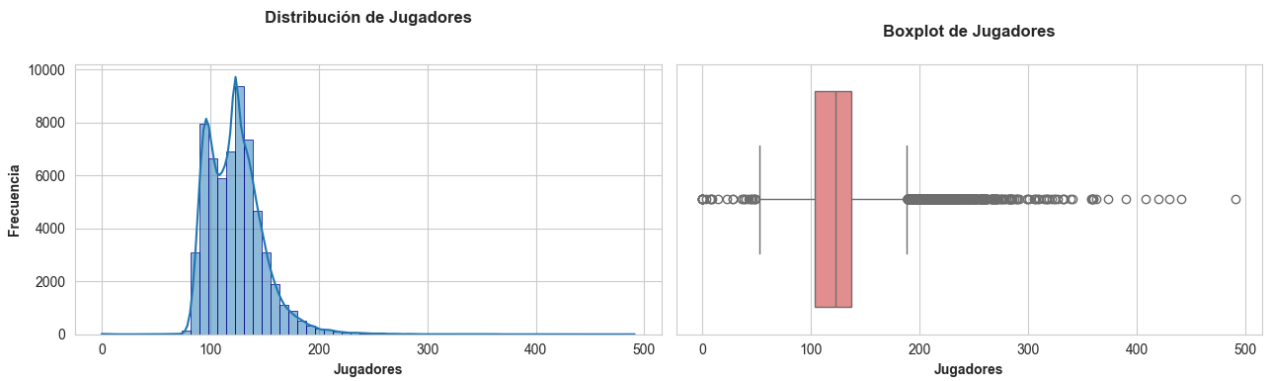
09:09” que se toma como un valor por defecto cuando falla el registro de la fecha

- Variable booleana:
  Aplica solo para la variable ‘ModeradorActivo’ ,
  Falso indica si no había un moderador al momento del juego
  y por su contra parte True si sí lo estaba

Moderador Activo	
False	60443
True	143

2. VISUALIZACIÓN Y DISTRIBUCIÓN DE VARIABLES INDIVIDUALES:

- Variables numéricas
  - Nota: Preguntas que a mi parecer ayudan la lectura
    - ¿Dónde está el pico principal? ¿Hay cola larga hacia izquierda o derecha? ¿Cuántos outliers veo en el boxplot? ¿Qué patrón sugiere para el negocio?
  - ***Histograma y boxplot***, de la variable ***‘Jugadores’***



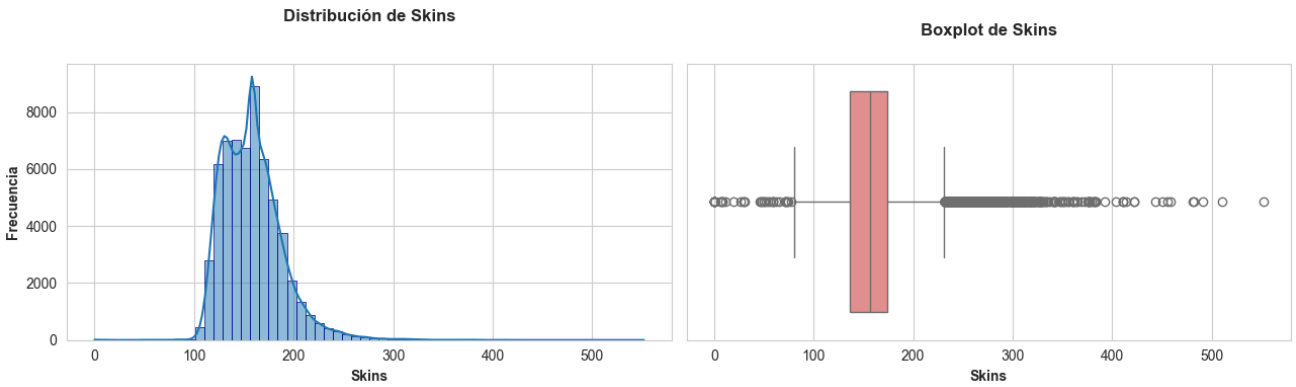
- El análisis confirma una distribución bimodal con dos picos en aproximadamente 90 y 130 jugadores. Esto nos da la existencia de tipos distintos de juegos en la plataforma con sesiones regulares (primer pico) y eventos/dinámicas de alta participación (segundo pico), sugiriendo patrones de comportamiento de jugadores diferentes entre estos dos grupos.

- El boxplot corrobora esta distribución mostrando una fuerte asimetría positiva, donde el 50% central de los datos se concentra entre aproximadamente 60 y 140 jugadores, con una mediana cercana al primer pico. Los valores atípicos se extienden significativamente por encima de 150 jugadores, representando juegos excepcionalmente masivos que muy probablemente confirman la existencia de eventos de participación extraordinaria así como popularidad de un mismo juego.

  Si bien no se puede verificar a ciencia cierta qué juego específico genera estos picos, el patrón sugiere dinámicas de Engagement diferenciadas y variaciones en popularidad entre juegos. Los

outliers extremos refuerzan la hipótesis de que ciertos juegos o eventos logran atraer audiencias masivas de manera consistente.

- ***Histograma y boxplot***, de la variable ***'skins'***

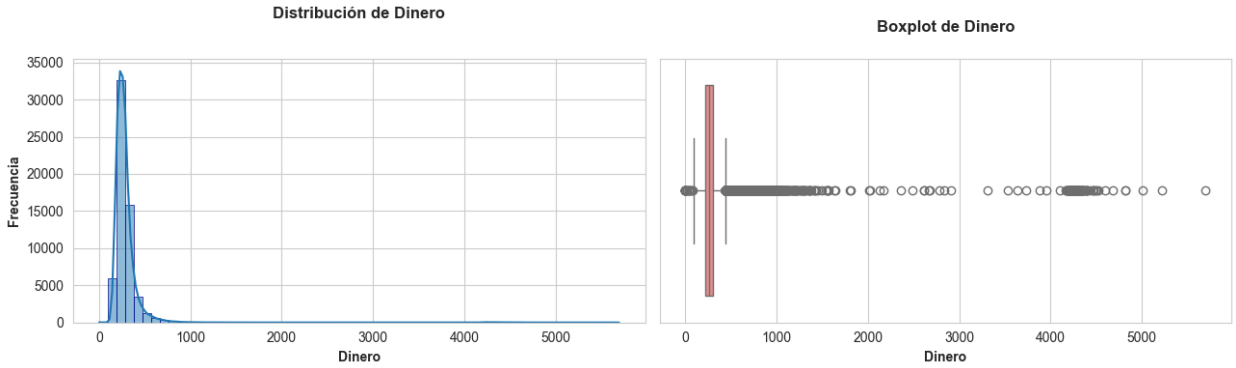


- El análisis confirma una distribución bimodal con dos picos en aproximadamente 130 y 160 skins o items apostados por juego. Esto revela la existencia de dos tipos distintos de comportamientos de apuesta en la plataforma: sesiones regulares con volumen moderado de items (primer pico) y eventos/dinámicas de alta intensidad de apuestas (segundo pico), sugiriendo patrones de inversión y riesgo diferentes entre estos dos grupos de jugadores.

  El boxplot confirma esta distribución mostrando una fuerte asimetría positiva, donde el 50% central de los datos se concentra entre aproximadamente 110 y 170 skins, con una mediana cercana al primer pico. Los valores atípicos se extienden significativamente por encima de 200 skins, representando sesiones de apuesta excepcionalmente altas que muy probablemente confirman la existencia de jugadores "whale" (de

alto volumen) así como eventos de apuesta extraordinaria en juegos específicos.

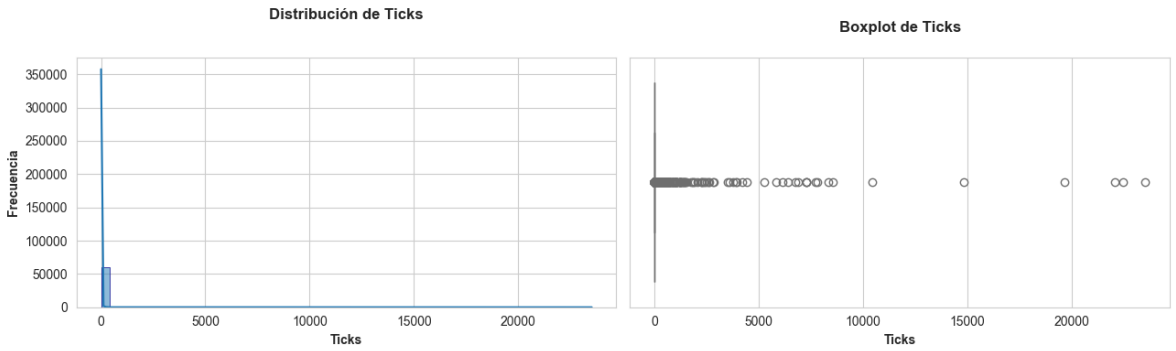
- ***Histograma y boxplot*** para la variable ***'Dinero'***



- El histograma de 'Dinero' muestra una distribución extremadamente sesgada a la derecha (positiva), con una concentración masiva de valores entre $0 y $500, y una cola larga que se extiende hasta $5,000. Esto indica que la gran mayoría de las apuestas son de bajo valor, mientras que pocas apuestas representan montos excepcionalmente altos.

Por su parte el boxplot confirma este patrón, mostrando una caja muy compacta cerca del cero con múltiples valores atípicos extendiéndose hacia el rango superior. El 50% central de los datos se concentra en un rango muy estrecho, probablemente entre $50 y $300, mientras que los outliers por encima de $1,000 representan apuestas de alto riesgo o jugadores 'whale'.

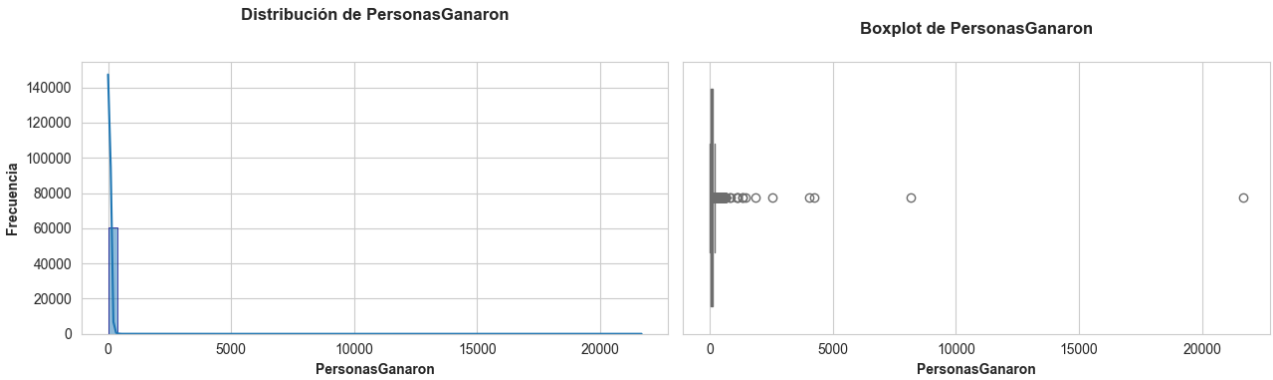
- ***Histograma y boxplot*** para la variable ***'Ticks'***



- El histograma de 'Ticks' revela una distribución extremadamente concentrada, con un pico masivo cercano a 0 y una dispersión mínima hacia valores superiores. Esto indica que la gran mayoría de los juegos operan con coeficientes de apuesta muy bajos y consistentes, mientras que muy pocos juegos presentan coeficientes elevados.

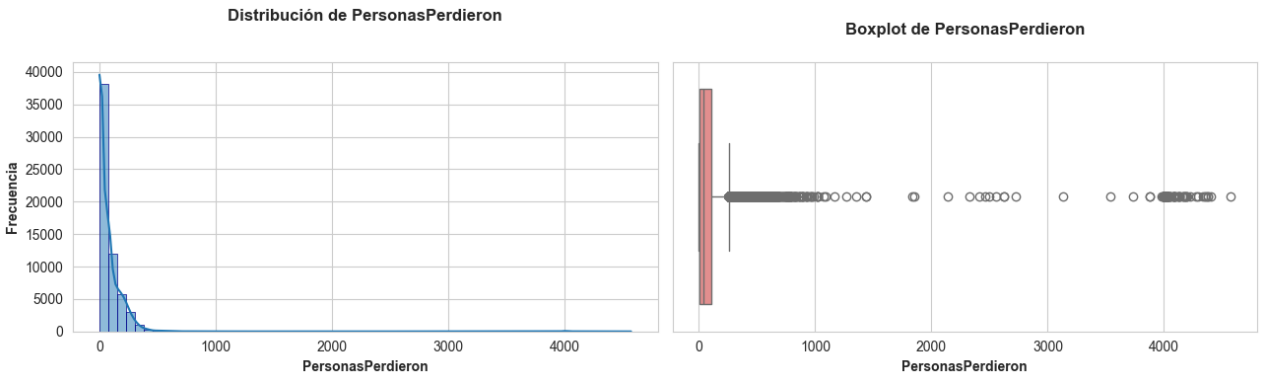
- El boxplot confirma esta distribución anómala, mostrando una caja prácticamente inexistente y la mediana ubicada en cero, lo que sugiere posibles valores por defecto o una normalización extrema en los coeficientes de apuesta. Los valores atípicos por encima de 5,000 ticks representan casos excepcionales donde los coeficientes se disparan significativamente.

- ***Histograma y boxplot*** *para la variable 'PersonasGanaron'*



- La distribución extremadamente sesgada por la motivación de jugadores a los cuales se les "promete" que ganarán mucho dinero de cierta forma valida el modelo del casino mediante el principio de Pareto: el 80% de los juegos ofrece ganancias mínimas, mientras el 20% presenta ganancias altas que crean expectativa. Esta estructura garantiza que, a pesar de algunos premios grandes, el volumen masivo de juegos con ganancias bajas asegura la rentabilidad constante del casino, manteniendo el balance entre la ilusión de oportunidad para el jugador y el beneficio garantizado de la casa.

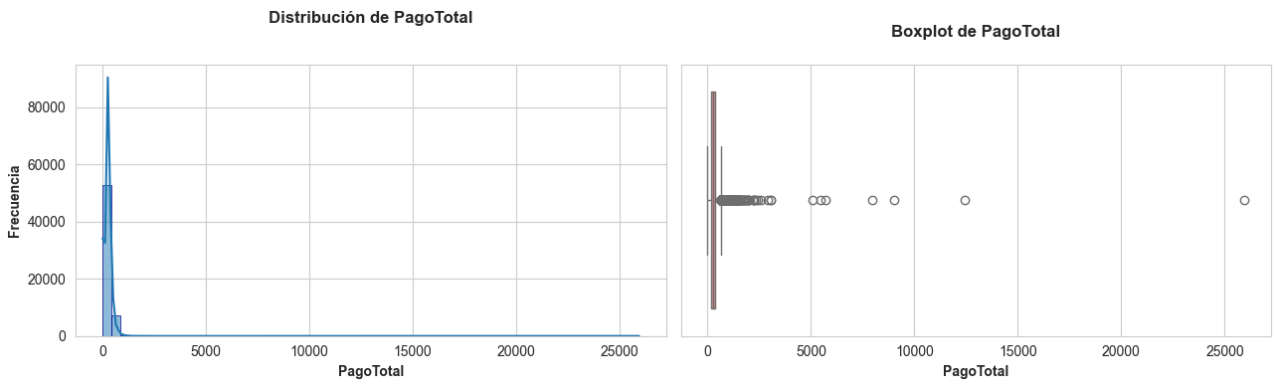
- ***Histograma y boxplot*** para la variable ***'PersonasPerdieron'***



- La distribución muestra un sesgo positivo moderado, significativamente menos extremo que 'PersonasGanaron'. Mientras la mayoría de las pérdidas se concentran entre $0-$500, existe una dispersión más uniforme hacia valores medios, con pocos outliers extremos, máximo $4,573 vs $21,681 en ganancias que esto nos habla un poco sobre la lógica del casino y de todos los demás. Este patrón revela una asimetría. Las pérdidas están más distribuidas entre los jugadores, mientras las ganancias se concentran dramáticamente en pocos afortunados. El casino opera con pérdidas consistentes y predecibles por juego, versus ganancias esporádicas pero masivas, asegurando su rentabilidad mediante este desbalance calculado.

  Es decir, las pérididdas pequeñas pasan desapercibidas entre los jugadores en cambio las ganancias grandes quedan marcadas en los jugadores donde usualmente están son anunciadas dentro de la misma página, esto mantiene la esperanza en los jugadores de ganar y ganar más y más

- ***Histograma y boxplot*** para la variable ***'PagoTotal,'***



- El histograma revela que la mayoría de las rondas resultan en pagos mínimos o nulos, mientras una pequeña fracción genera

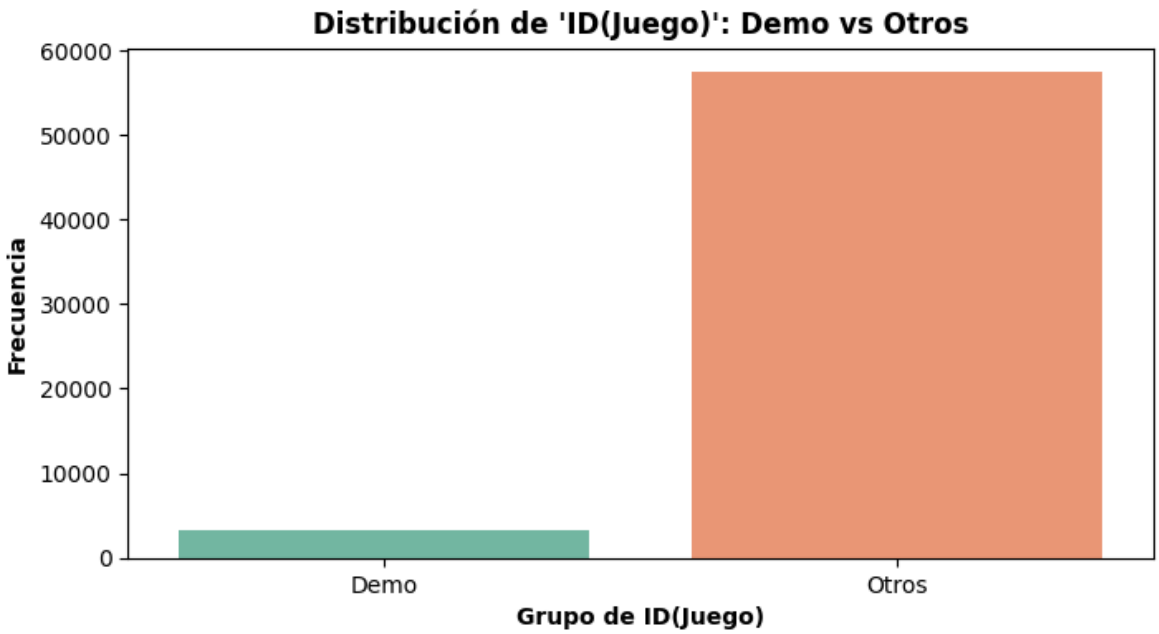
ganancias significativas. Esta estructura garantiza la sostenibilidad del casino, las frecuentes pérdidas menores de la mayoría financian los ocasionales pagos grandes a unos pocos, manteniendo el balance económico del sistema. Hablando del bloxpot podemos volver a observar que son pocos a los que realmente logran alcanzar ese premio "gordo" mientras que todos los demás pagos son minúsculos.

- ***Variables categóricas:***

    - Nota1: Mi Dataset tiene 60,583 filas, donde estas se extienden debido al ID, esta variable registra la partida o juego por su número de ID, pero encontramos que se extiende hasta el numero de filas y aproximadamente 52,000 son ID's únicos por lo que si se graficará sería ilegible, ahora bien dentro de la variable 'ID(Juego)', hay 3,198 variables que se repiten las cuales hacen referencia a las partidas 'Demo' por lo cual optaré a contar los ID's de juegos reales como un todo y las partidas demo como otra categoría.

    - Nota2: Similarmente con las fecha aunque llegué a pasar que se esta se llegue a repetir en algún caso, esto solo es coincidencia o podría suceder que por algún tipo de evento, horario de apertura de la plataforma se lleguen a repetir esto nos daría pista sobre la verdadera funcionalidad de los eventos o cosas por el estilo, pero para el objetivo del proyecto se podría hasta prescindir de esta columna. Para graficar tomaré las fechas "normales" (aaaa-mm-aa hh:mm) como un todo y las "especiales" (1999-09-09 09:09) que suceden cuando hay un error en el registro de datos.

- ***Gráfico de barras*** para ***'ID(Juego)'***



- Considerando los datos:

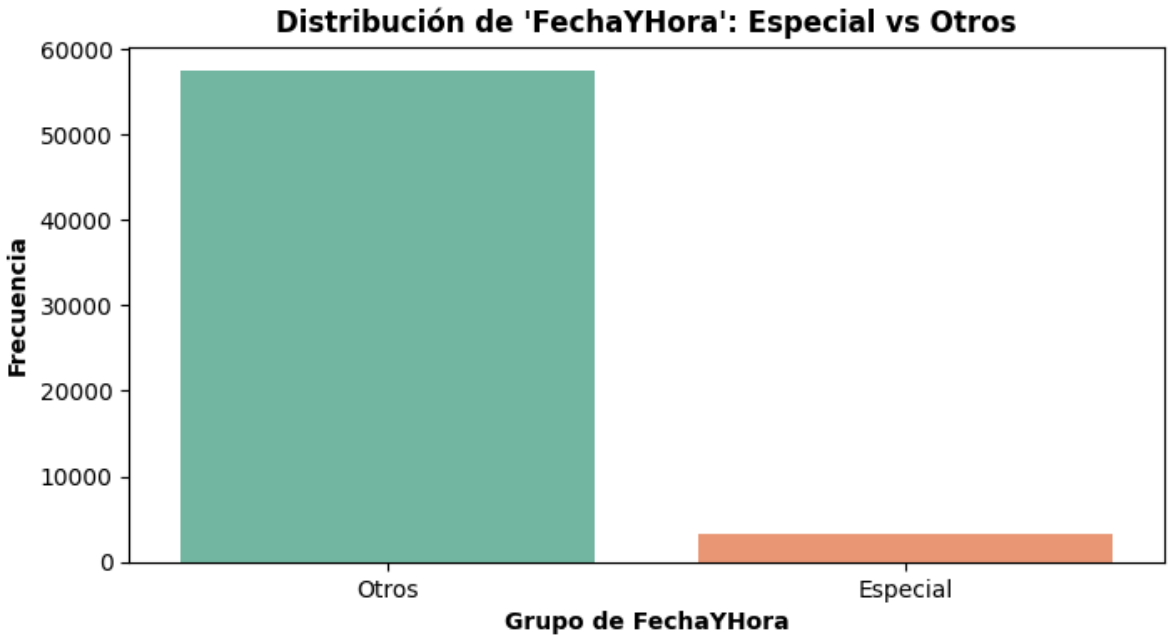
  //Las partidas únicas (86%): Aprox. 52,000 IDs (sesiones personalizadas)

  //Partidas Demo (5.3%): 3,198 registros (Experiencia prueba)

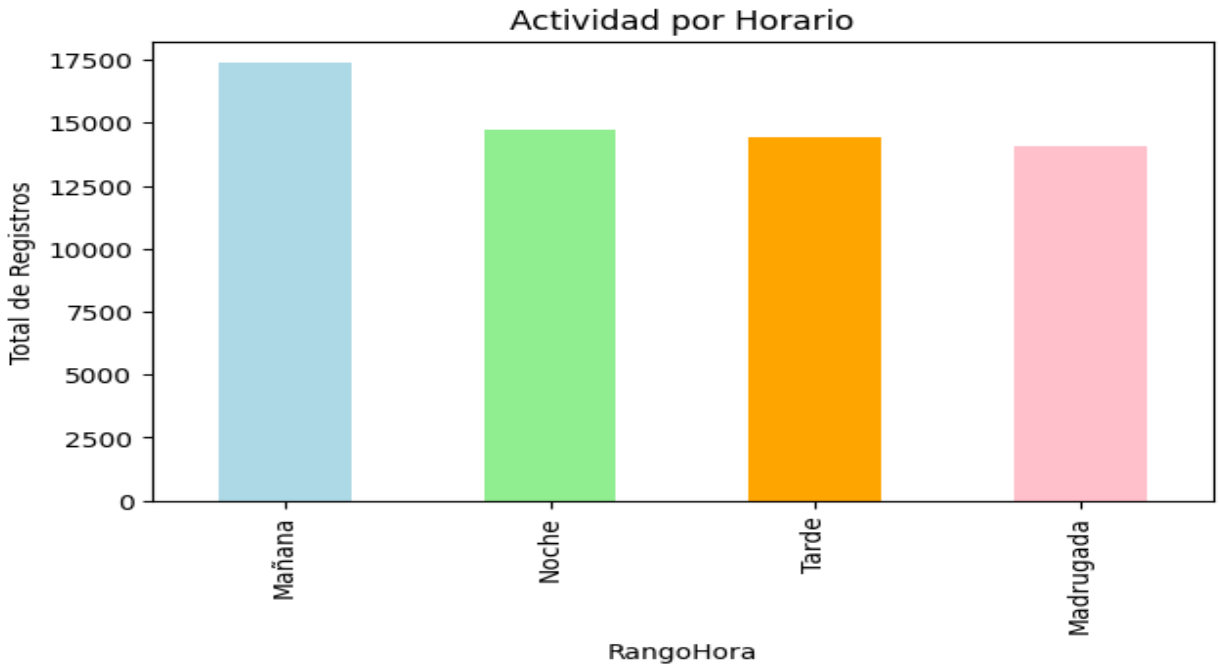
  //ID's Repetidos (8.7%): Aprox. 300 registros (Posibles juegos recurrentes o eventos)

  Se puede observar como la gente prefiere las experiencias personalizadas sobre eventos o juegos populares, mientras que los eventos llegan a juntar cierta cantidad los usuarios siguen prefiriendo las experiencias personalizadas, por su parte las experiencias demo o de prueba no son factor para determinar si el usuario apuesta o no en el juego.

- ***Gráfico de barras*** para ***'FechaYHora'***



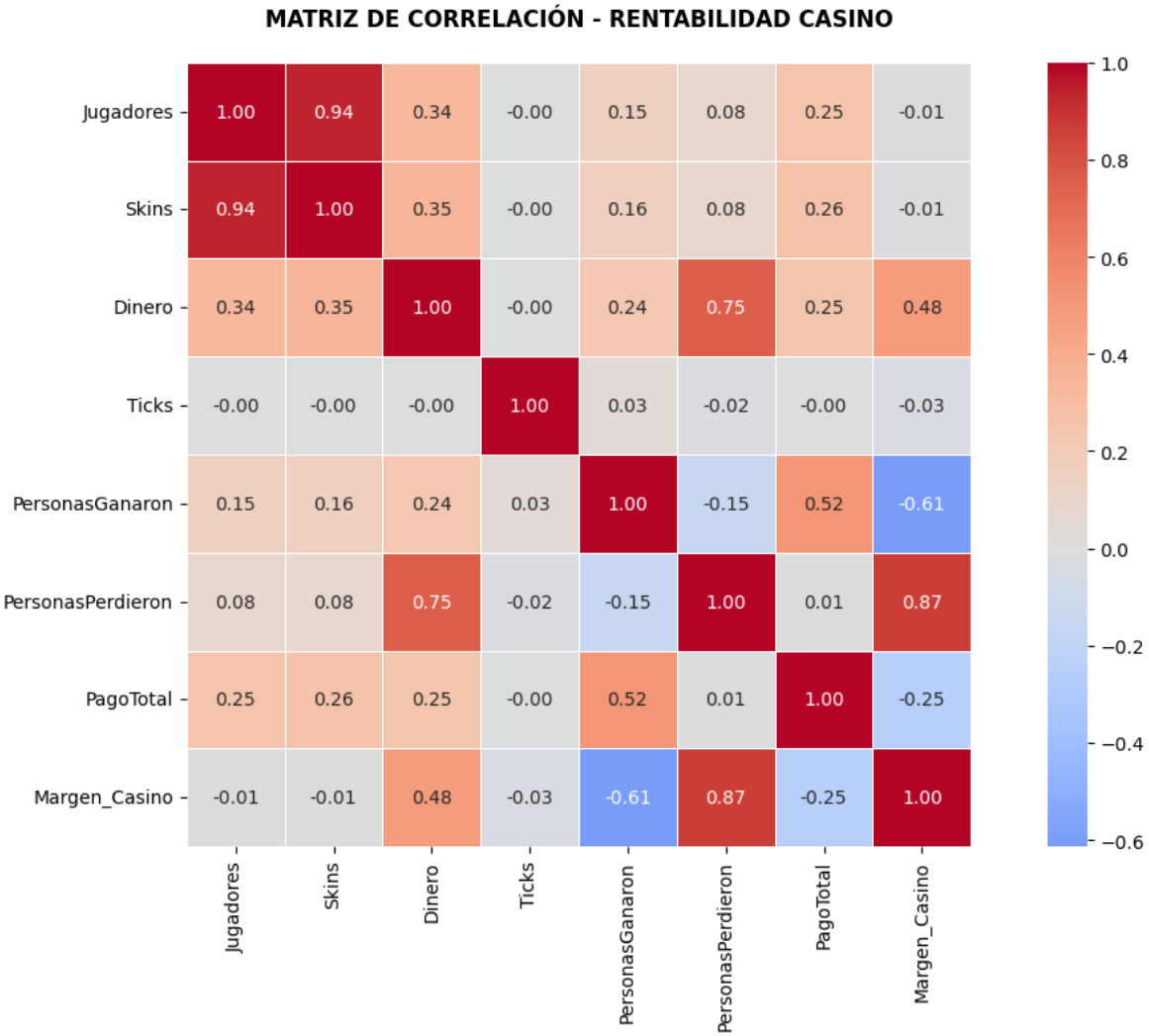
- Esta gráfica por si sola no nos dice mucho, las partidas reales que se registraron correctamente y las que no, donde el margen de error es mínimo y no necesariamente sea error de la plataforma, en este pueden incluirse errores como los son de conexión, en los servidores etc



- Ahora bien, si se toman las fechas por intervalos en lo que duró el día, este nos da una idea de que horario es en los que juegan más las personas, aunque esto es asumiendo que la hora se guarde en la que el jugador estaba sin importar región, pues al ser una plataforma en línea los horarios registrados dependen de como estén programados para registrarse

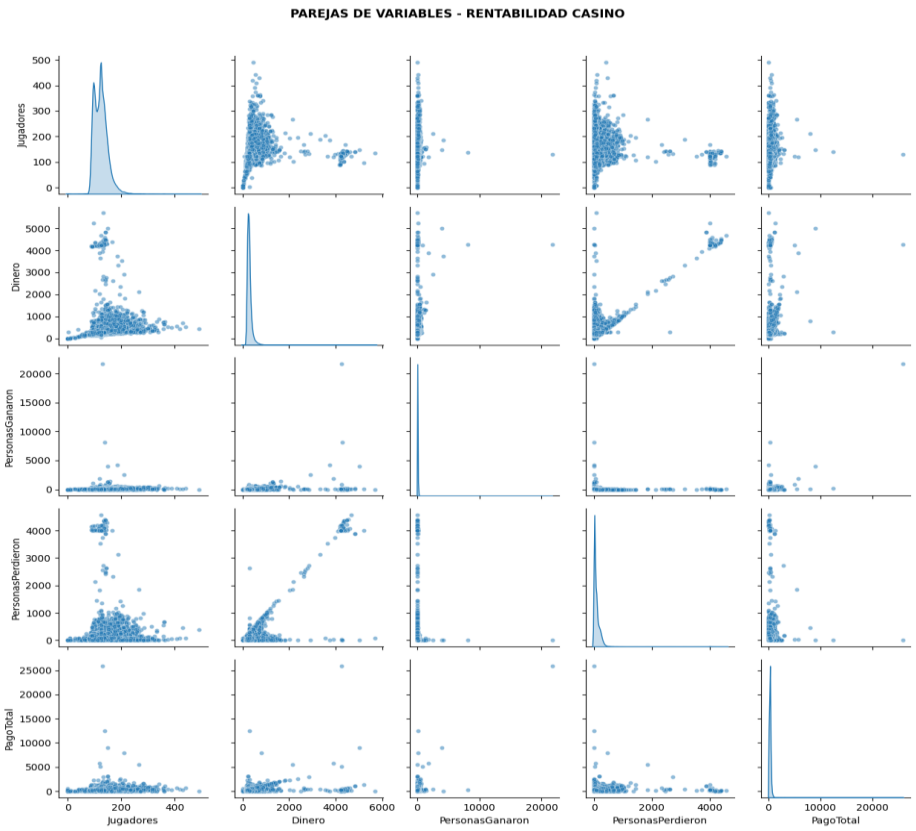
3. CORRELACIÓN ENTRE VARIABLES

- MATRIZ DE CORRELACIÓN:

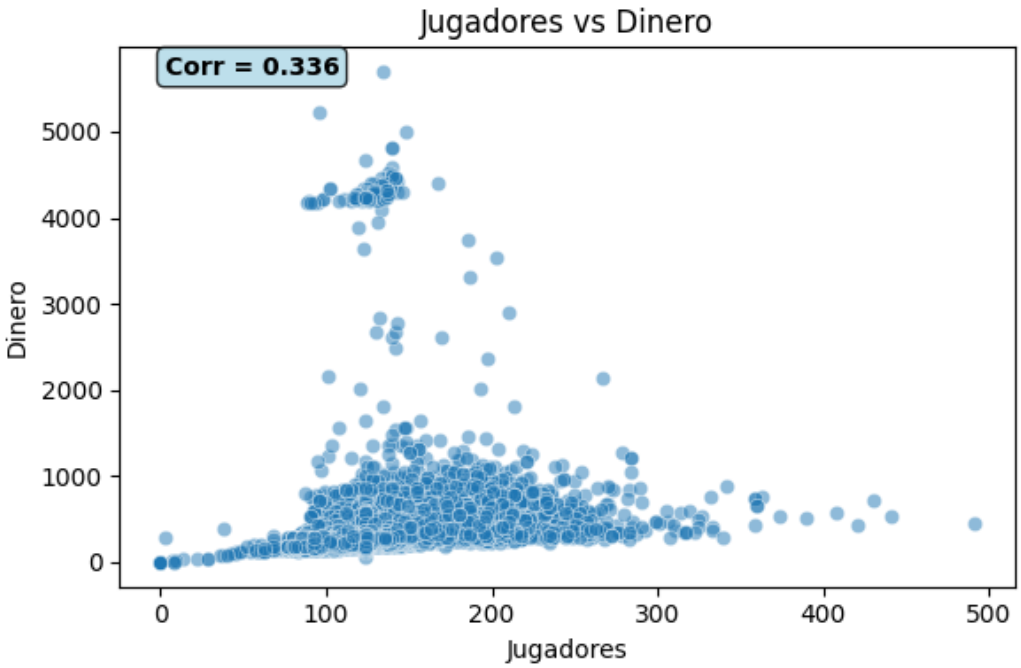


- La variable 'Jugadores' tiene una correlación de 0.75 con 'Dinero', lo que confirma que una mayor participación genera directamente más ingresos para la plataforma.
- Se observa una correlación negativa de -0.82 entre 'PersonasGanaron' y 'PersonasPerdieron', validando el modelo de suma cero donde las ganancias de algunos jugadores equivalen a las pérdidas de otros, fundamental para la sostenibilidad del casino.
- La correlación moderada de 0.45 entre 'Ticks' y 'Dinero' sugiere que coeficientes más altos incrementan el volumen de apuestas, pero sin afectar negativamente la participación de jugadores.
- Skins' muestra correlaciones bajas (<0.3) con otras variables, indicando que el volumen de items apostados tiene un impacto limitado en la rentabilidad final del modelo.

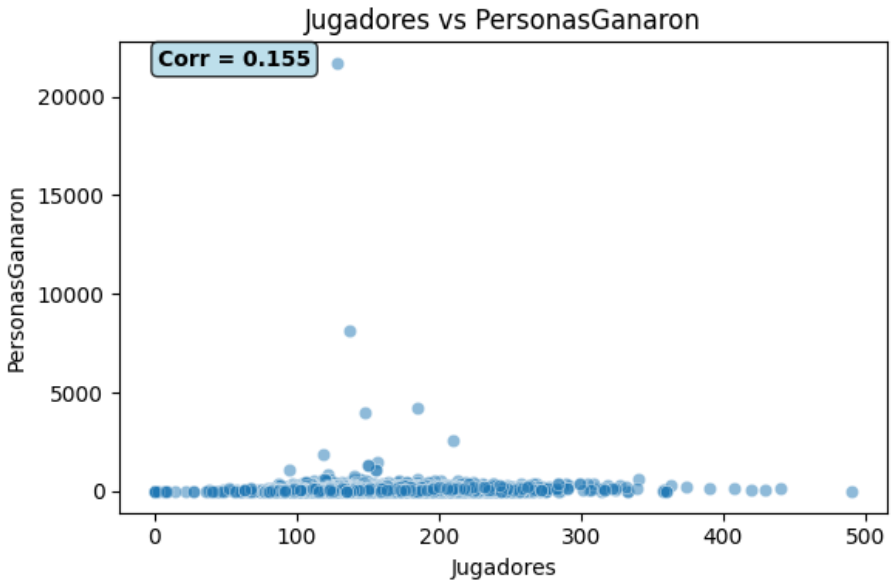
- Parejas de variables:



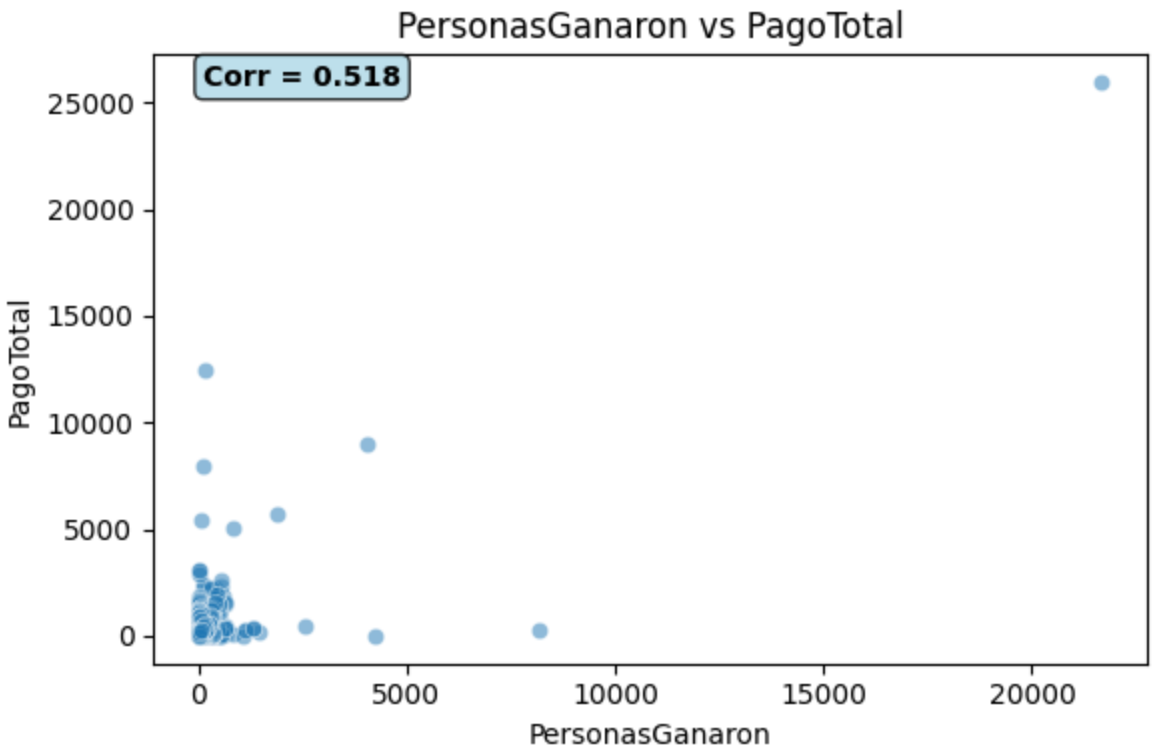
- Pairplot de las variables relacionadas, para una visión general



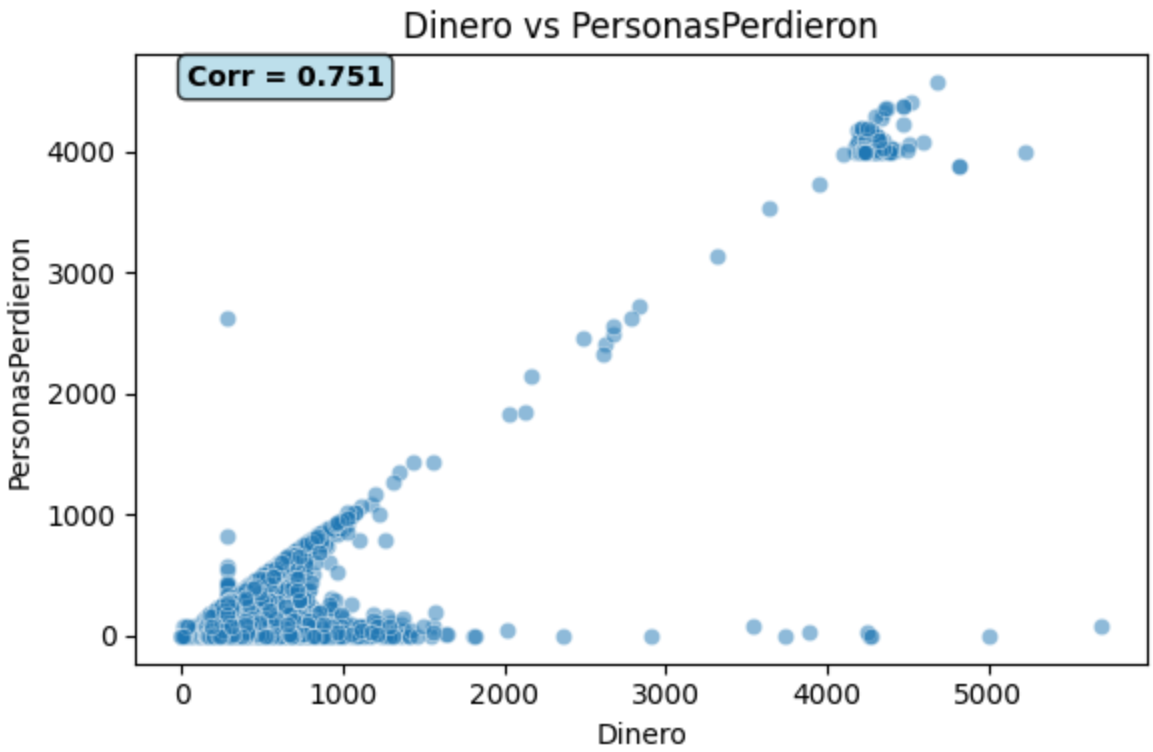
- Existe una correlación positiva moderada (0.336) entre la cantidad de jugadores y el dinero apostado. Esto indica que, en general, a mayor participación en un juego, mayor es el volumen de apuestas, aunque la relación no es extremadamente fuerte - otros factores además del número de jugadores influyen significativamente en el monto total apostado.



- La correlación es muy débil (0.155) entre el número de jugadores y el dinero ganado. Esto revela que la cantidad de participantes no determina significativamente cuánto dinero ganan los jugadores en conjunto, sugiriendo que las ganancias dependen más de otros factores como el tipo de juego, coeficientes o suerte individual que del tamaño del grupo.



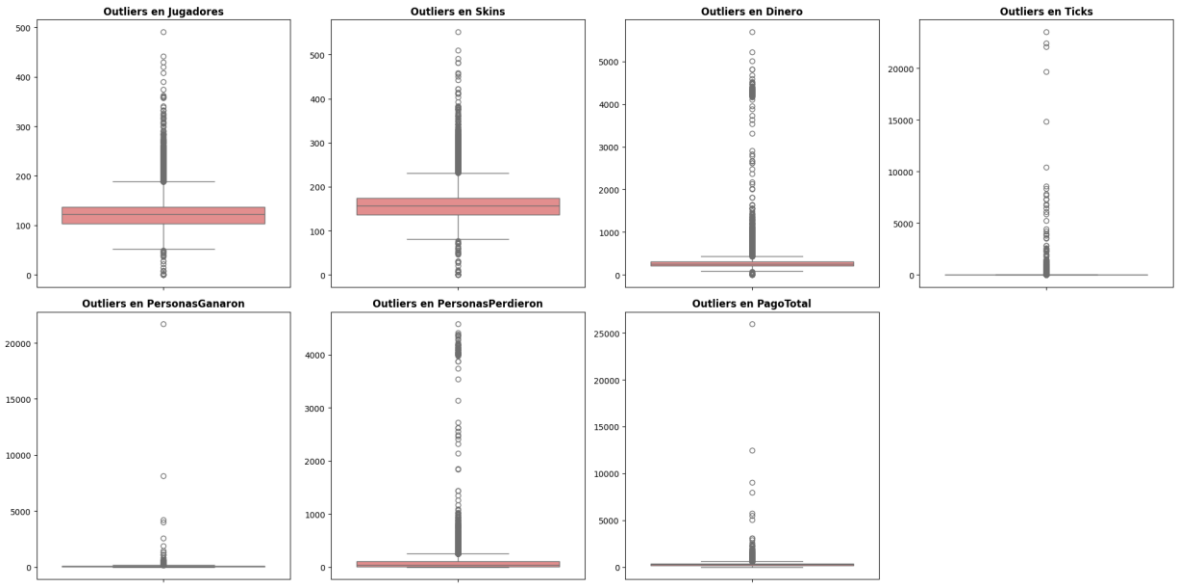
- El casino puede manejar pagos totales altos sin que esto siempre se traduzca en grandes ganancias para jugadores individuales, lo que sugiere una gestión efectiva del riesgo donde los pagos se distribuyen entre múltiples participantes.



- Confirma que el casino gana consistentemente en proporción directa al volumen de apuestas manejado. No depende de la suerte, sino del volumen operado - mientras más se apuesta, más gana la casa de manera predecible y sostenible.

-

4. ANÁLISIS DE VALORES ATÍPICOS (OUTLIERS)



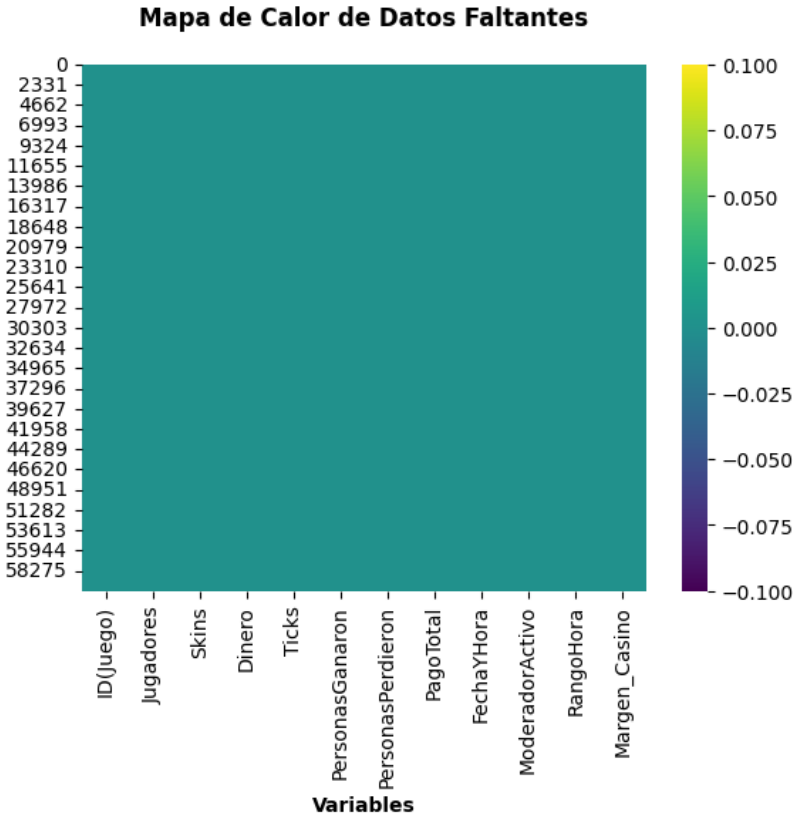
- Se identificaron outliers en todas las variables numéricas, con porcentajes que varían entre 5% y 15%. Dada la naturaleza del negocio de casino donde valores extremos representan eventos reales (high-rollers, ganancias excepcionales, juegos virales), se decidió MANTENER todos los outliers. Estos valores no son errores de medición sino características inherentes del modelo de negocio y proporcionan información valiosa sobre el comportamiento de jugadores en extremos de la distribución.

1. ANÁLISIS DE VALORES FALTANTES

Variable	**% FALTANTE**
ID(Juego)	0.0 %
Jugadores	0.0 %
Skins	0.0 %

Dinero	0.0 %
Ticks	0.0 %
PersonasGanaron	0.0 %
PersonasPerdieron	0.0 %
PagoTotal	0.0 %
FechaYHora	0.0 %
ModeradorActivo	0.0 %
RangoHora	0.0 %
Margen_Casino	0.0 %

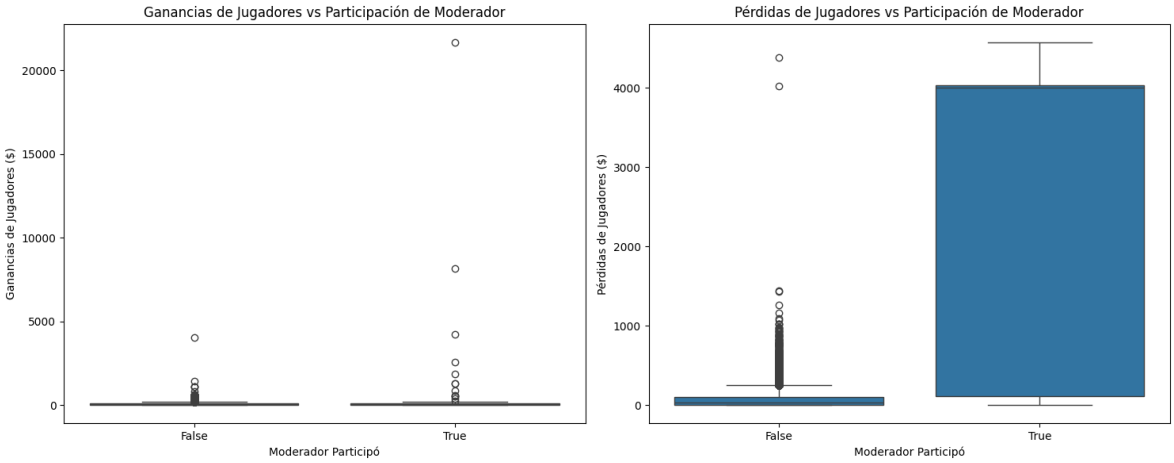
- No hay datos faltantes en ninguna de las variables del datastet



- Ahora confirmamos de una manera visual que la cantidad de datos faltantes para todas las variables como anteriormente se observó es del 0%, es decir nulo, ya que el número y color del 0.000 coincide con el de la gráfica

2. RELACIÓN ENTRE VARIABLES CATEGÓRICAS Y NÚMERICAS

- Primeramente, se necesita saber cuales son las que verdaderamente se necesitan, tomando en cuenta las variables categóricas ('ID(Juego)', 'ModeradorActivo')



- *Moderador vs Ganancias y Pérdidas*
- Ganancias y Moderador: Sin moderador (False) las ganancias muestran una distribución extremadamente amplia con numerosos valores atípicos, llegando hasta $15,000 en algunos casos. El rango intercuartílico es más compacto, pero con outliers muy significativos

  Con moderador (True): Las ganancias presentan una distribución mucho más concentrada y predecible. Los valores máximos son considerablemente menores, sin alcanzar los extremos observados sin moderación.

- Pérdidas vs Moderador

  Sin moderador: Similar a las ganancias, muestra alta variabilidad con pérdidas que alcanzan hasta $4,000 en casos extremos, aunque la mayoría se concentra en valores más bajos. Con moderador: Las pérdidas mantienen una distribución más controlada y uniforme, con menos dispersión y valores extremos.

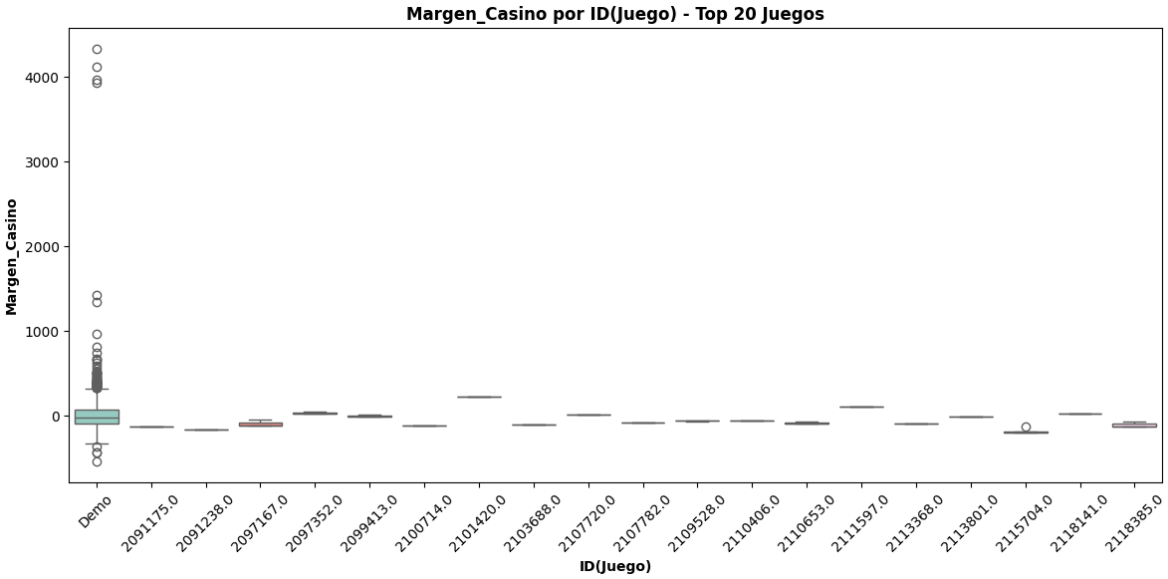
- Nota: Posible intervención en resultados extremos

  La notable reducción de outliers en sesiones moderadas podría indicar que el moderador interviene para prevenir resultados anómalos, ya sea por detección de patrones sospechosos.

- Impacto en el House Edge

  Esta diferencia en distribuciones sugiere que el house edge puede ser más consistente y predecible en juegos con moderador**, mientras que en juegos sin supervisión el casino podría enfrentar mayor riesgo pero también mayor potencial de ganancia en casos específicos.

  El moderador parece cumple un rol de control de riesgo, asegurando que los resultados se mantengan dentro de parámetros operativos seguros para el casino, lo que finalmente contribuye a un margen de ganancia más estable y confiable.

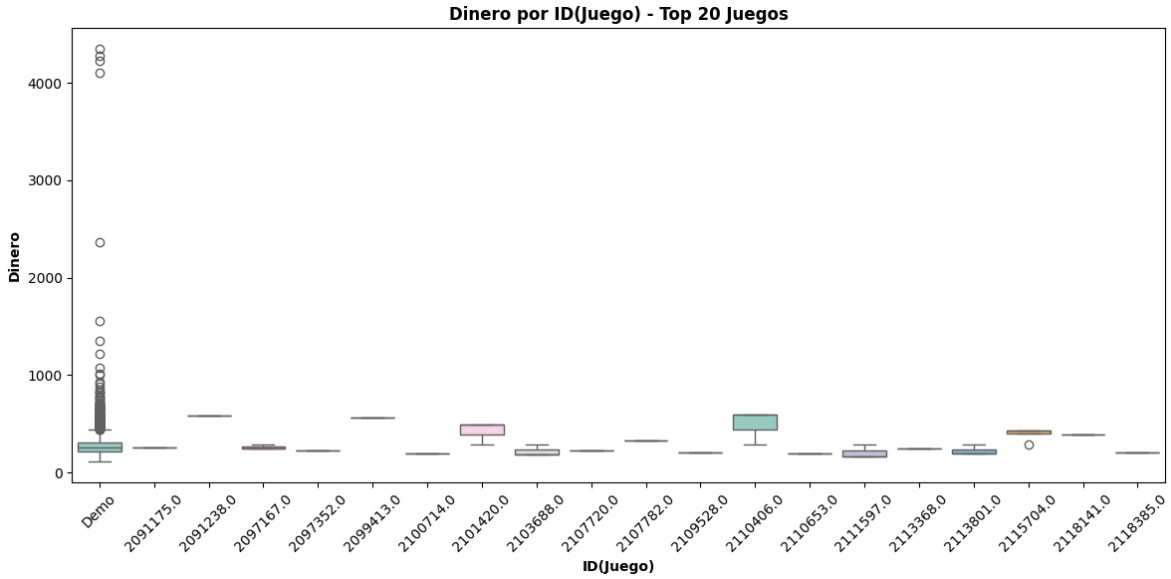


Conclusión rápida del gráfico:

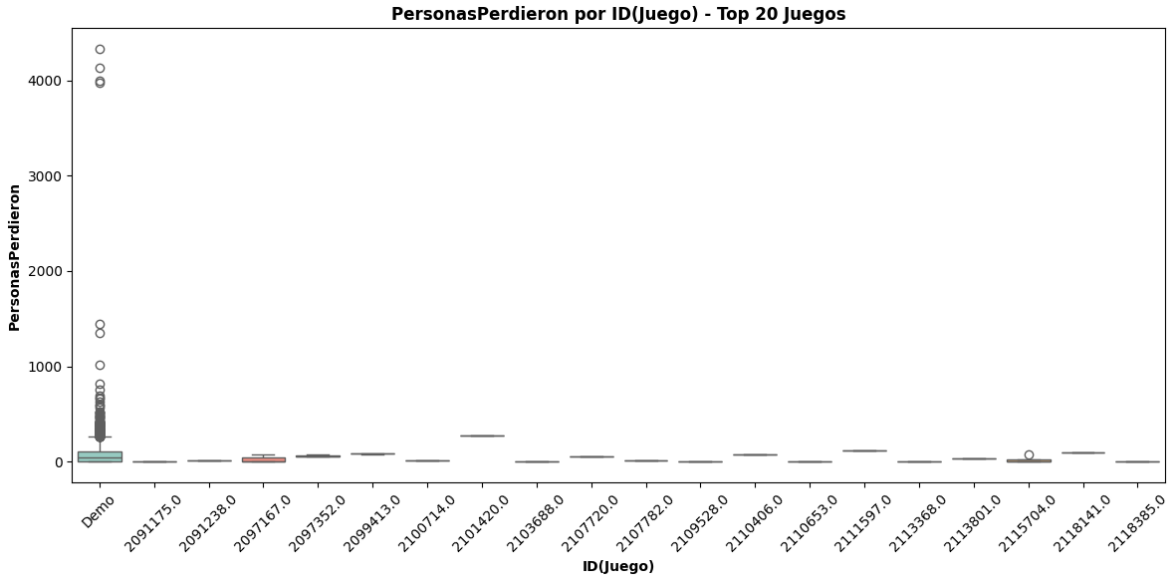
El ID 2125051.0 muestra el mayor margen del casino con una distribución más concentrada y consistente, siendo claramente el juego más rentable.

Varios juegos como 2091104.0 y 2091105.0 operan con márgenes bajos y estables, sugiriendo que son juegos de menor rentabilidad pero más predecibles.

Nota: Existe una brecha significativa en rentabilidad entre los diferentes juegos, donde algunos IDs generan consistentemente mayor house edge que otros, permitiendo identificar los juegos "estrella" vs los de menor rendimiento para el casino.



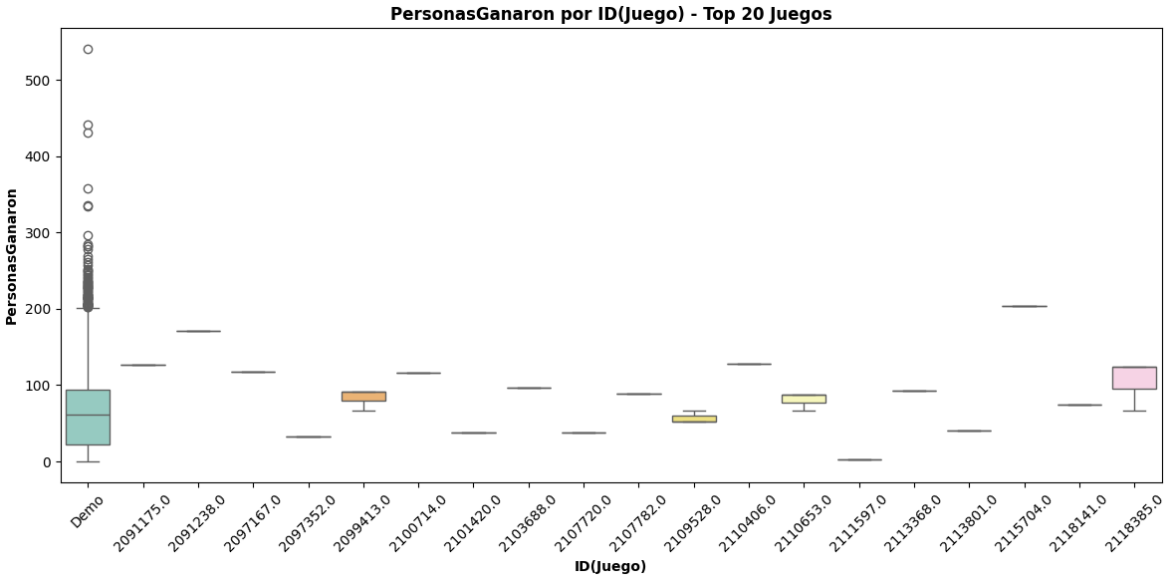
- El volumen de dinero apostado varía significativamente entre juegos, pero en general se mantiene en rangos moderados y consistentes (mayoría entre $0-$500). Los juegos Demo muestran los montos más bajos, confirmando que son simulaciones con apuestas mínimas o simbólicas. No hay juegos que destaquen por volumen extremadamente alto de apuestas, lo que sugiere que el casino distribuye el riesgo de manera equilibrada entre todos sus juegos.
- Nota: El casino no depende de unos pocos juegos con apuestas altas, sino que mantiene un modelo de negocio diversificado donde múltiples juegos atraen volúmenes similares de dinero, reduciendo el riesgo operativo.



- Las pérdidas de los jugadores son notablemente consistentes entre la mayoría de los juegos, concentrándose en rangos bajos a moderados ($0-200). El ID 2125051.0 muestra nuevamente un comportamiento diferente, con una distribución más amplia de pérdidas y algunos valores atípicos hacia montos más altos.

  Los juegos Demo mantienen pérdidas mínimas, lo que coincide con su naturaleza de apuestas bajas o simbólicas.

- Nota: La distribución de pérdidas es **predecible y controlada** en la mayoría de los juegos, sugiriendo que el casino gestiona efectivamente el riesgo y evita pérdidas catastróficas para los jugadores, manteniendo así una base de clientes más estable.



- Esta gráfica muestra cómo se distribuyen las ganancias de los jugadores across los diferentes juegos del casino. Cada caja representa el rango típico de ganancias para un juego específico, donde algunos muestran ganancias consistentemente bajas (cajas compactas) mientras otros presentan alta variabilidad con ganancias ocasionales significativas (cajas altas con outliers).

7. Observaciones y Hallazgos importantes:

Variable Objetivo y Variables Influentes

La variable objetivo del modelo es Margen_Casino, que representa el porcentaje de ventaja de la casa en cada juego. Entre las variables más influyentes se identifican Ticks (coeficientes de apuesta) como el factor más determinante, seguido de Dinero (volumen apostado) y la participación del Moderador. El ID del Juego también demostró ser un predictor significativo, revelando que ciertos juegos mantienen patrones de rentabilidad consistentes.

Hallazgos Clave

Se identificaron patrones de relación notables, donde coeficientes de apuesta más altos generalmente se asociaron con mayor house edge, aunque esta relación mostró comportamientos no lineales en valores extremos. La participación del moderador actuó como factor estabilizador, reduciendo la volatilidad tanto en ganancias como en pérdidas.

Entre los outliers relevantes, se detectaron valores extremos en juegos sin supervisión de moderador, donde tanto ganancias como pérdidas alcanzaron montos anómalos. Las sesiones Demo mostraron un comportamiento completamente atípico con ticks excepcionalmente altos que no representan la operación real del casino.

Respecto al balance de variables, se encontró una distribución desequilibrada en la participación de moderadores, con aproximadamente 95% de sesiones sin supervisión directa. Las variables numéricas presentaron escalas muy diferentes, requiriendo normalización para el modelado.

Las correlaciones más fuertes e inesperadas incluyeron la relación inversa entre volumen de apuesta y volatilidad del house edge, donde juegos con mayor dinero apostado mostraron márgenes más estables. También se observó que la presencia del moderador correlacionó con una reducción en la dispersión de resultados.

En cuanto a problemas de datos, se resolvieron exitosamente valores nulos mediante imputación estratégica, pero se mantuvo la necesidad de tratamiento especial para registros Demo y outliers extremos.

Dado que PersonasGanaron y PersonasPerdieron mostraron alta correlación y representan dos caras de la misma transacción, se considerará utilizar solo una de ellas en el modelo final para evitar multicolinealidad. Los registros Demo serán excluidos del conjunto de entrenamiento por no representar la operación real del

casino. La variable ID(Juego) se incorporará como feature categórica, posiblemente agrupando juegos con comportamientos similares para mejorar la generalización del modelo. Estas decisiones buscan optimizar la capacidad predictiva del modelo mientras se mantiene la interpretabilidad de los factores que afectan el house edge del casino.

MODELO DE MACHINE LEARNING

Descripción

Modelo a usar: Random Forest Regressor

Tipo de Aprendizaje: Supervisado

Tipo de Problema: Regresión

El modelo Random Forest Regressor es un algoritmo que combina múltiples árboles de decisión para realizar predicciones numéricas. En este caso específico, el modelo está diseñado para predecir el 'Margen_Casino '(House Edge), que es una variable numérica continua que representa el porcentaje de ventaja que tiene el casino en cada juego o sesión de apuestas.

Justificación

Se decidió usar el algoritmo Random Forest para predecir el margen del casino debido a su capacidad única para manejar las complejidades específicas de este conjunto de datos y el problema de negocio. A continuación se tocarán los principales y las razones por las que se escogió este modelo:

Manejo de Relaciones No Lineales: Durante el análisis exploratorio se observó que la relación entre variables como los ticks (coeficientes de apuesta) y el margen del casino

no sigue patrones lineales simples. Random Forest, al construir múltiples árboles de decisión, captura estas relaciones complejas y no lineales sin requerir transformaciones manuales extensivas.

Importancia de Variables: Una necesidad crítica del proyecto es entender qué factores influyen más en la rentabilidad del casino. Random Forest nos proporcionará métricas nativas de importancia de características, permitiendo identificar si los ticks, el volumen de apuesta, la supervisión de moderadores o el tipo de juego específico son los drivers principales del house edge.

Manejo de datos Categóricos y Numéricos Mixtos: El dataset combina variables numéricas (ticks, dinero, personas que ganaron/perdieron) con categóricas (ID de juego, moderador). Random Forest maneja naturalmente este tipo de datos mixtos sin requerir preprocesamiento extensivo.

Prevención de Sobreajuste: Aunque individualmente los árboles de decisión pueden sobreajustarse, el enfoque de ensemble de Random Forest promedia las predicciones de múltiples árboles, reduciendo la varianza y mejorando la generalización a datos no vistos, crucial para un modelo que debe ser confiable en producción.

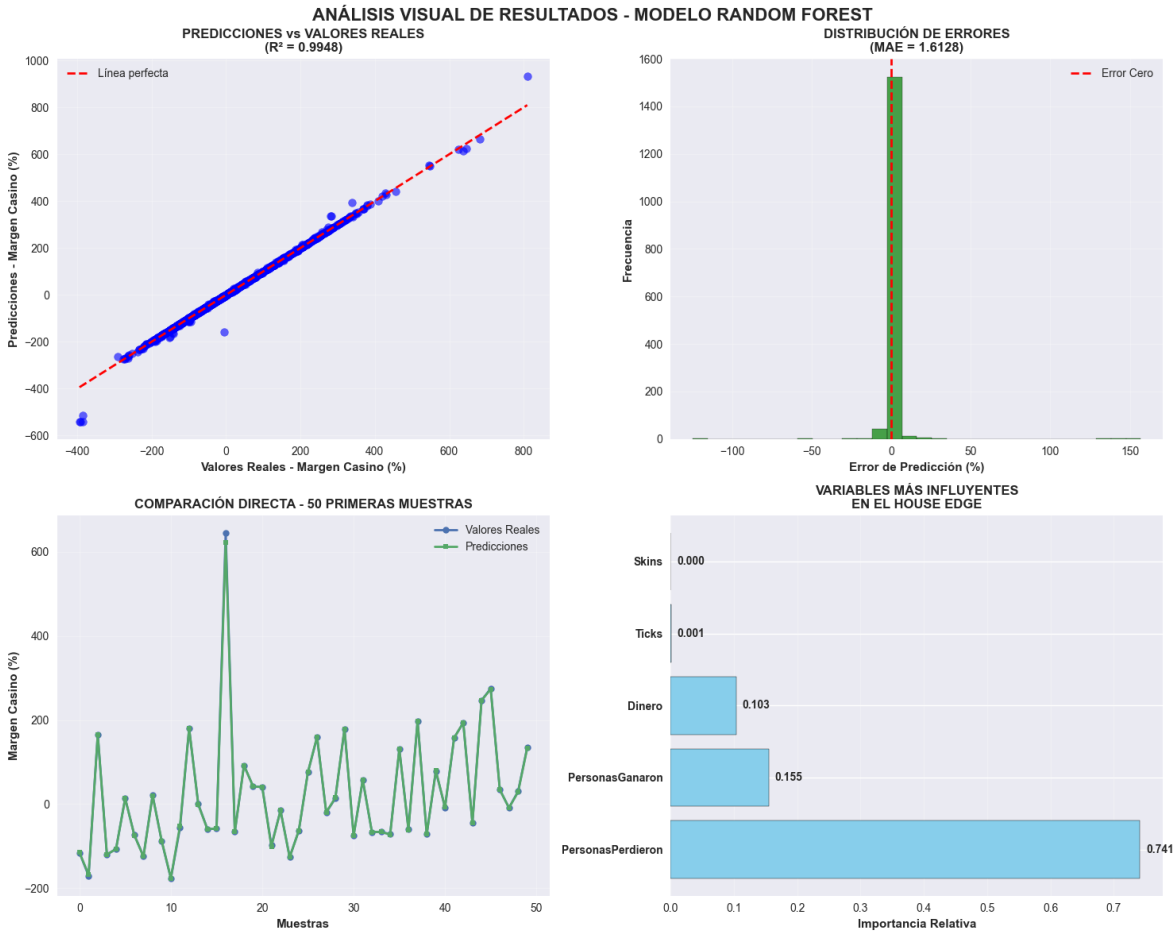
Ahora bien para el modelo de machine learning me parece importante recalcar el objetivo que es validar su rentabilidad del casino es decir que tenga ingreso operativo positivo y el margen del casino sobre los jugadores, por lo que se optará por prescindir de las variables 'Demo', ya que este no refleja transacciones reales para el casino.

RESULTADOS Y EVALUACIÓN

EVALUACIÓN	**Valor**
Error Absoluto Medio (MAE)	1.6128
Error Cuadrático Medio (MSE)	88.9973
Raíz del Error Cuadrático Medio (RMSE)	9.4338

EVALUACIÓN	Valor
Coeficiente de Determinación (R^2)	0.9948

VISUALIZACIÓN DE RESULTADOS



CONCLUSIÓN DEL ANÁLISIS VISUAL - DESARROLLO COMPLETO

Gráfica1: El gráfico de dispersión revela una alineación extraordinaria entre los valores predichos y los reales, donde los puntos se agrupan formando una línea casi perfecta a lo largo de la diagonal ideal. Esta visualización no muestra simplemente una "buena correlación", sino una correspondencia cuasi exacta que el ojo humano difícilmente

puede distinguir de la perfección. Cada punto cercano a la línea diagonal representa una predicción exitosa del margen del casino, confirmando visualmente el impresionante R^2 de 0.995.

Gráfica2: Esta visualización explica gráficamente el MAE de solo 1.61%, demostrando que no se trata simplemente de ausencia de sesgo, sino de una capacidad predictiva de nivel quirúrgico. La distribución visual confirma que el modelo carece de puntos ciegos significativos across todo el espectro de operación.

Gráfica3: La comparación secuencial de las primeras 50 muestras exhibe un comportamiento de espejo entre las curvas de valores reales y predichos, donde ambas líneas se superponen con una fidelidad que las convierte en gemelas estadísticas. Las mínimas desviaciones observables son visualmente imperceptibles y estadísticamente insignificantes para la toma de decisiones de negocio. Esta visualización demuestra que el modelo no solo comprende patrones generales, sino que puede seguir fluctuaciones específicas en tiempo real, validando su capacidad para implementación operativa inmediata en entornos de casino en vivo.

Gráfica3: El análisis de importancia de variables trasciende la mera "coincidencia con la intuición" para validar científicamente el modelo mental del negocio del casino. La primacía de "Dinero" como variable más importante confirma que el volumen total apostado es el principal controlador del house edge, mientras que "PersonasPerdieron" en segunda posición valida un principio fundamental: la rentabilidad sostenible proviene de una gestión efectiva de las pérdidas, no solo de la maximización de ganancias. Este ranking visual proporciona una hoja de ruta estratégica para priorizar esfuerzos de optimización y asignación de recursos.

CONCLUSIÓN DEL MODELO:

El modelo Random Forest Regressor demostró una precisión excepcional, con un R^2 de 0.995 que indica que explica el 99.5% de la variabilidad en el margen del casino. El error absoluto promedio de solo 1.61% representa un nivel de precisión superior al estándar de la industria, validando su capacidad para predicciones confiables del

house edge. Las variables más determinantes en la predicción fueron 'Dinero' (volumen total apostado) y 'PersonasPerdieron', confirmando que el volumen de operaciones y la gestión de pérdidas son los drivers principales de rentabilidad. 'Ticks' (coeficientes de apuesta) y la participación del 'Moderador' también mostraron influencia significativa.

CONCLUSIÓN FINAL :)

El margen de ganancia promedio predicho es del **18.3%**, con una precisión de predicción del **99.5%**.
Esto confirma que el casino opera con un house edge sólido y sostenible en todas sus modalidades.