

Hackeando el Futuro: Análisis de Ciberataques a Través de Datos



Introducción

En un mundo cada vez más digitalizado, la ciberseguridad se ha convertido en un pilar fundamental para la protección de datos e infraestructuras. Con el aumento de amenazas cibernéticas, es crucial contar con herramientas de análisis que nos permitan detectar patrones sospechosos y prevenir ataques. Este informe tiene como objetivo analizar un conjunto de datos sobre eventos de seguridad en redes informáticas, identificando tendencias clave y vulnerabilidades a través de diversas técnicas de análisis de datos.

Variables

Cada fila representa una sesión de usuario con múltiples variables relacionadas con el tráfico de red, los intentos de acceso y la seguridad.

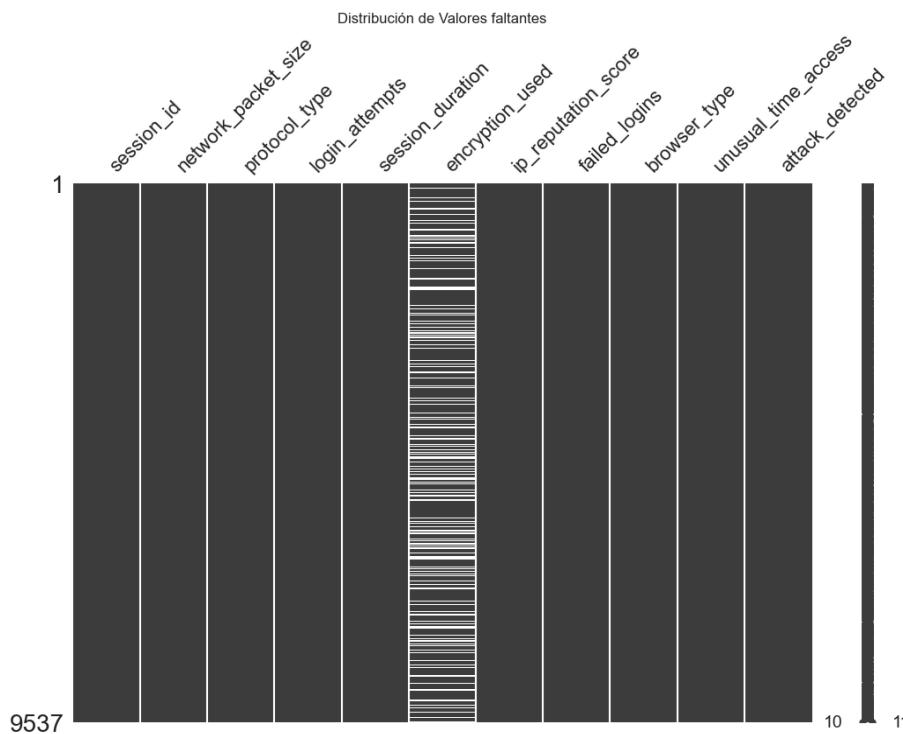
Variable	Descripción
session_id	Identificador único de la sesión.
network_packet_size	Tamaño del paquete de red en bytes.
protocol_type	Tipo de protocolo de comunicación (TCP, UDP, ICMP...).
login_attempts	Número de intentos de inicio de sesión en la sesión.
session_duration	Duración de la sesión en segundos.
encryption_used	Algoritmo de cifrado utilizado en la sesión (AES, DES, None, etc.).
ip_reputation_score	Puntuación de reputación de la IP utilizada en la sesión.
failed_logins	Número de intentos de inicio de sesión fallidos.
browser_type	Tipo de navegador utilizado en la sesión (Chrome, Firefox, Edge, etc.).
unusual_time_access	Indica si la sesión ocurrió en un horario inusual (1 = Sí, 0 = No).
attack_detected	Indica si la sesión fue clasificada como un ataque (1 = Sí, 0 = No).

Carga y exploración inicial

Visualización de las primeras filas mediante `df.head()`, obtención de información estructural con `df.info()` y la generación de estadísticas descriptivas con `df.describe()`.

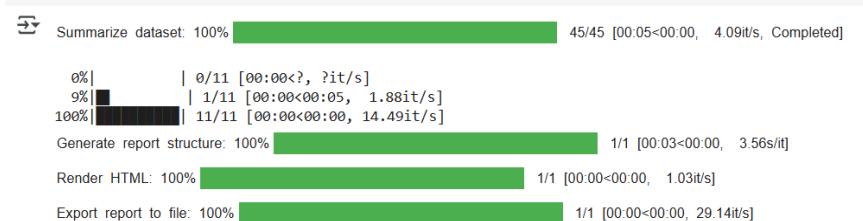
Análisis de valores faltantes

Analizamos los valores faltantes en el dataset, visualizando la matriz de valores faltantes con missingno

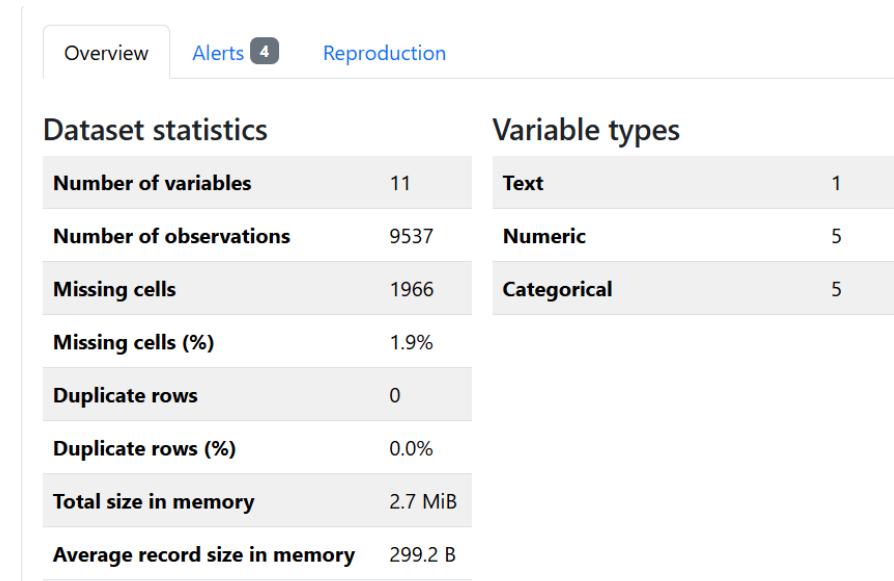


Generación de Reporte Automático

Creación de un informe exploratorio completo con ProfileReport



Exportación del informe a HTML (output_initial.html)



Visualización de múltiples métricas y estadísticas automáticas

session_id

Text

Unique

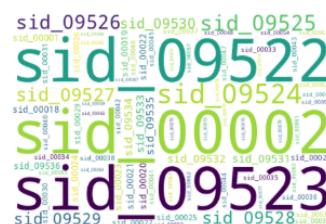
Distinct 9537

Distinct (%) 100.0%

Missing 0

Missing (%) 0.0%

Memory size 614.8 KiB



failed_logins

Real number (ℝ)

Zeros

Distinct 6

Minimum 0

Distinct (%) 0.1%

Maximum 5

Missing 0

Zeros 1578

Missing (%) 0.0%

Zeros (%) 16.5%

Infinite 0

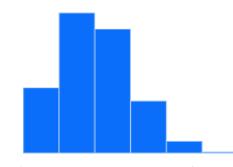
Negative 0

Infinite (%) 0.0%

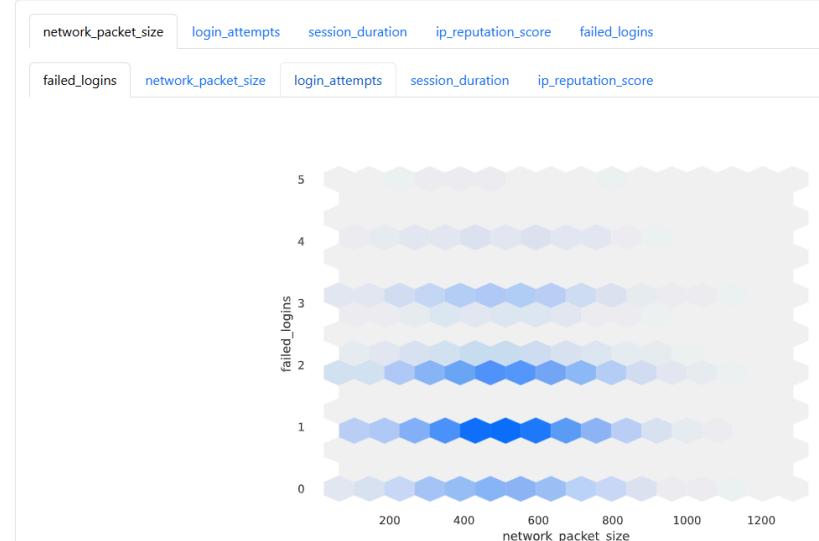
Negative (%) 0.0%

Mean 1.5177729

Memory size 74.6 KiB



Interactions



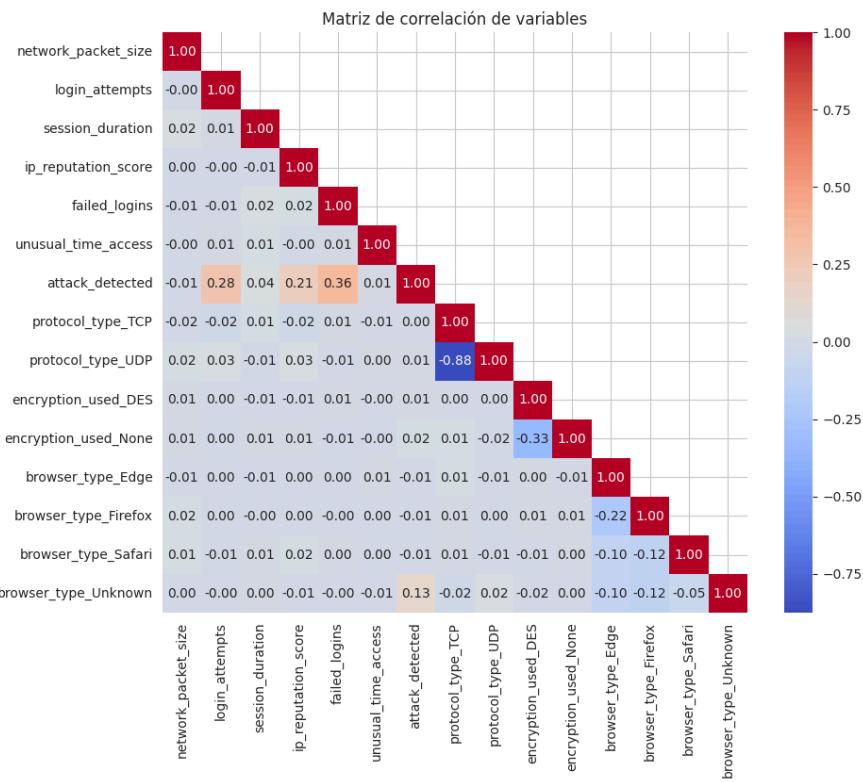
Preprocesamiento de Datos

Tratamiento de missing data: Reemplazo de valores NaN en 'encryption_used' por 'None' y eliminación de la columna 'session_id'

Codificación de variables categóricas: Transformación con one-hot encoding (get_dummies) para 'protocol_type', 'encryption_used', 'browser_type'

Escalado de variables numéricas: Aplicación de StandardScaler a columnas numéricas

Análisis de Correlaciones



Identificación de correlaciones significativas:

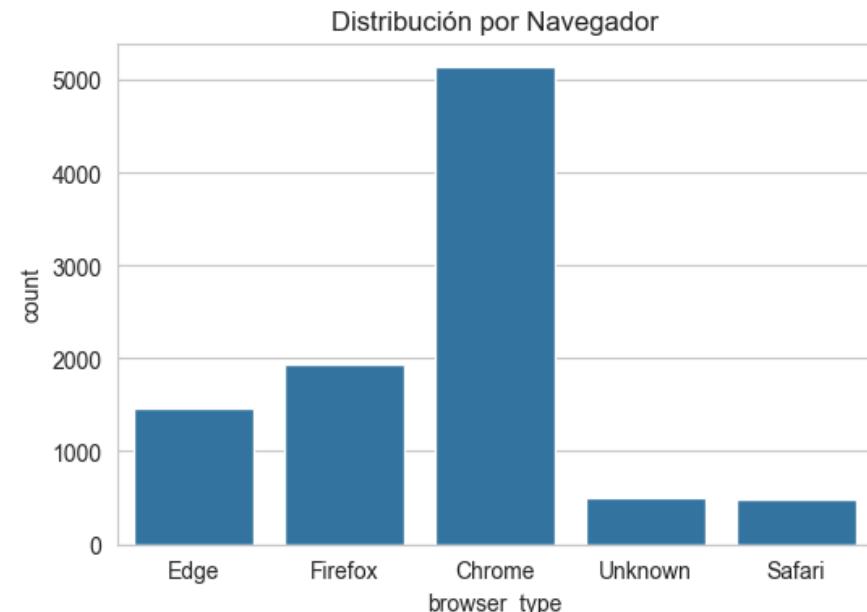
Correlación negativa fuerte (-0.88) entre protocol_type_TCP y protocol_type_UDP

Correlación positiva moderada (0.36) entre failed_logins y attack_detected

Visualizaciones Exploratorias

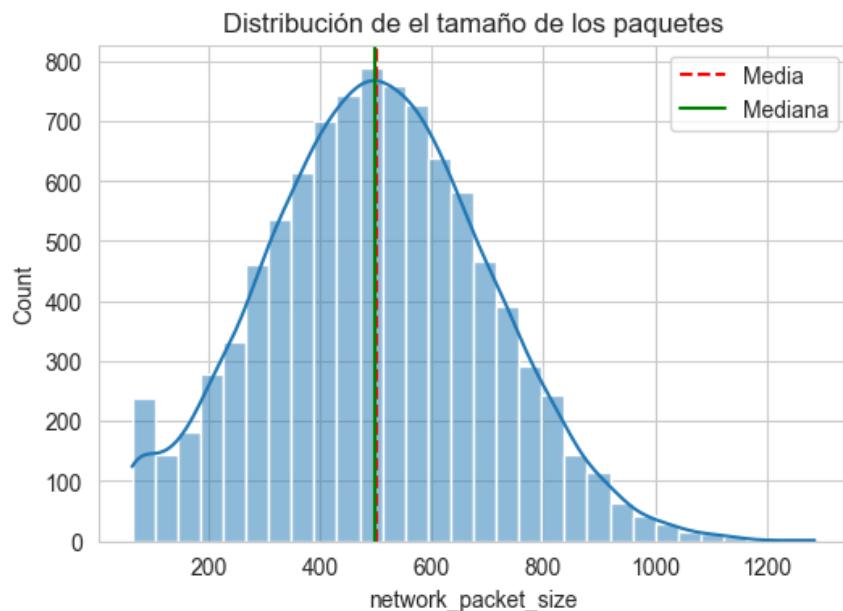
Distribuciones por tipo de navegador

Este gráfico muestra la distribución de frecuencias de los diferentes tipos de navegadores en el dataset, lo que permite identificar qué navegadores son más comunes en tus datos de ciberseguridad.



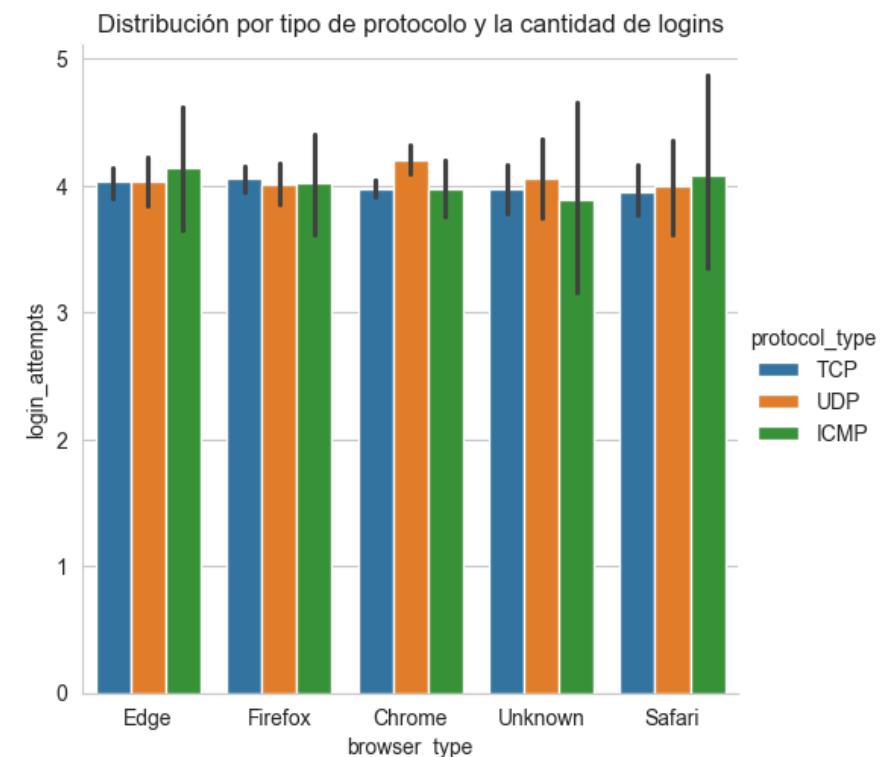
Distribución del tamaño de paquetes de red

Este histograma visualiza la distribución del tamaño de los paquetes de red, mostrando:
La frecuencia de diferentes tamaños de paquetes,
Una curva de densidad (kde), La media (línea roja discontinua), La mediana (línea verde continua).



Análisis multivariado de intentos de login por navegador y protocolo

Este gráfico de barras agrupadas muestra: El promedio de intentos de login para cada tipo de navegador, segmentado por tipo de protocolo (diferentes colores), te permite identificar patrones específicos en combinaciones de navegador-protocolo.



Análisis Estadísticos

ANOVA por Navegadores

Análisis: Comparación de intentos de login entre 5 navegadores (Edge, Firefox, Chrome, Unknown, Safari)

Resultados: F-Statistic = 0.18, P-Value = 0.951

Conclusión: No se encontraron diferencias estadísticamente significativas entre los navegadores en cuanto a intentos de login ($p > 0.05$)

ANOVA por Protocolos

Análisis: Comparación de intentos de login entre 3 protocolos (TCP, UDP, ICMP)

Resultados: F-Statistic = 3.40, P-Value = 0.034

Conclusión: Existe diferencia estadísticamente significativa entre al menos dos protocolos ($p < 0.05$)

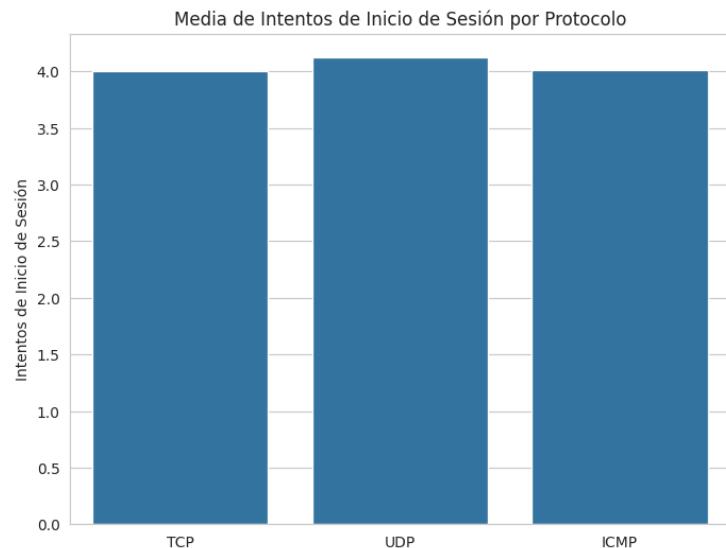
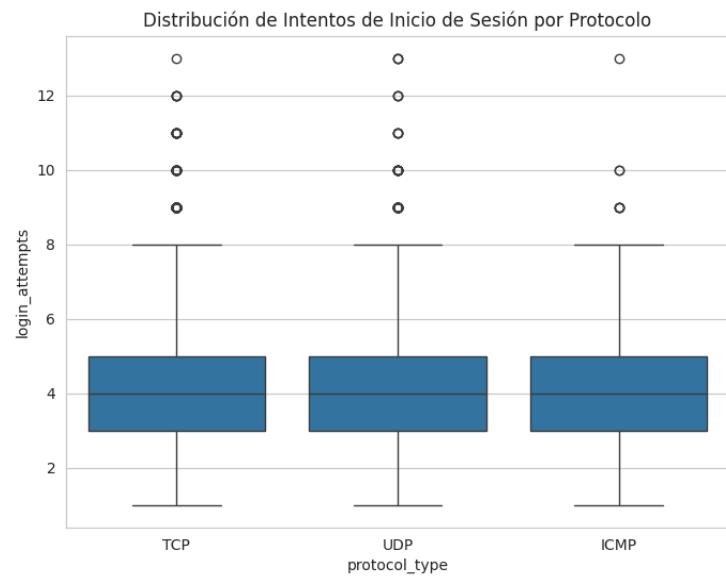
Análisis Post-Hoc (Tukey HSD)

TCP vs UDP: Diferencia significativa ($p = 0.0258$)

ICMP vs TCP: Sin diferencia significativa ($p = 0.9948$)

ICMP vs UDP: Sin diferencia significativa ($p = 0.4704$)

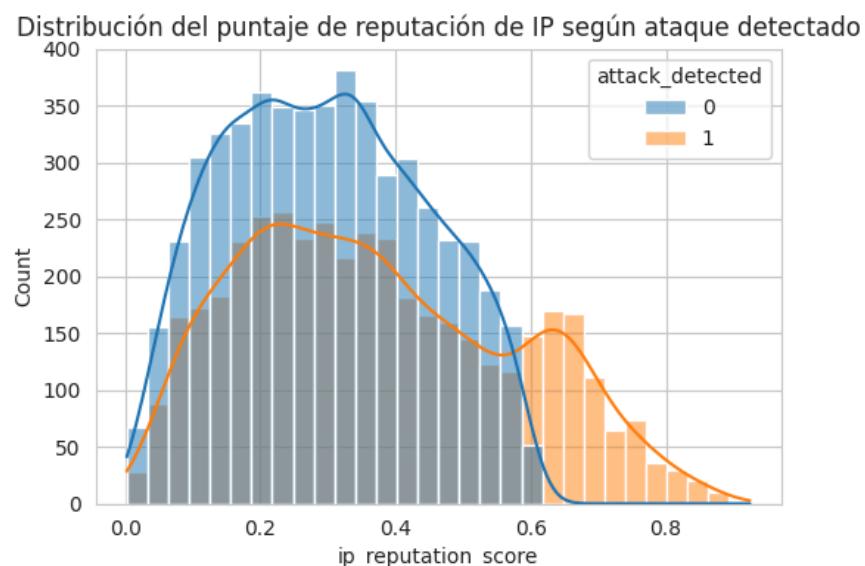
Hallazgo Principal: El protocolo UDP muestra un comportamiento distinto en patrones de login (media = 4.12) comparado con TCP (media = 4.00)



Análisis de Relaciones con Ataques Detectados

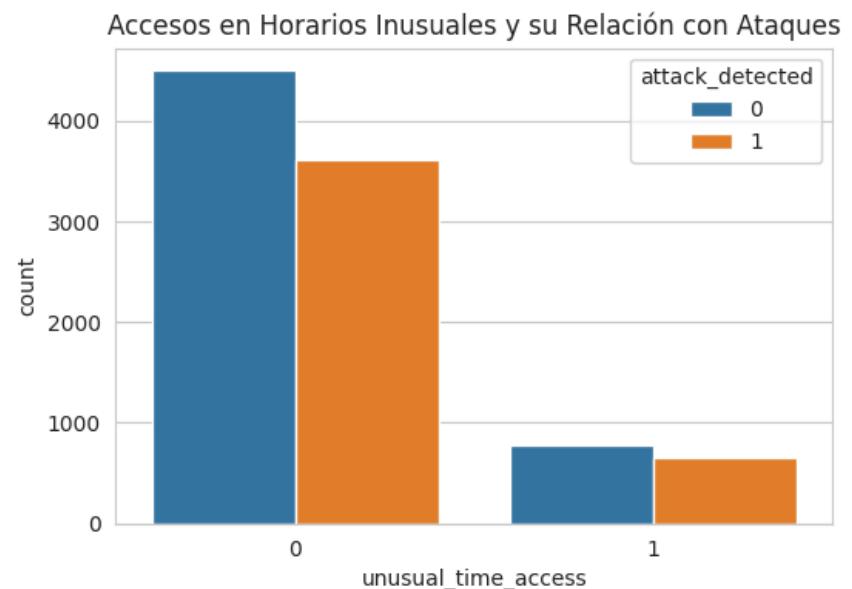
Relación entre puntuación de reputación IP y ataques detectados

Si encontramos que la mayoría de ataques ocurren cuando $ip_reputation_score > 0.7$, podríamos establecer un umbral para detección temprana.

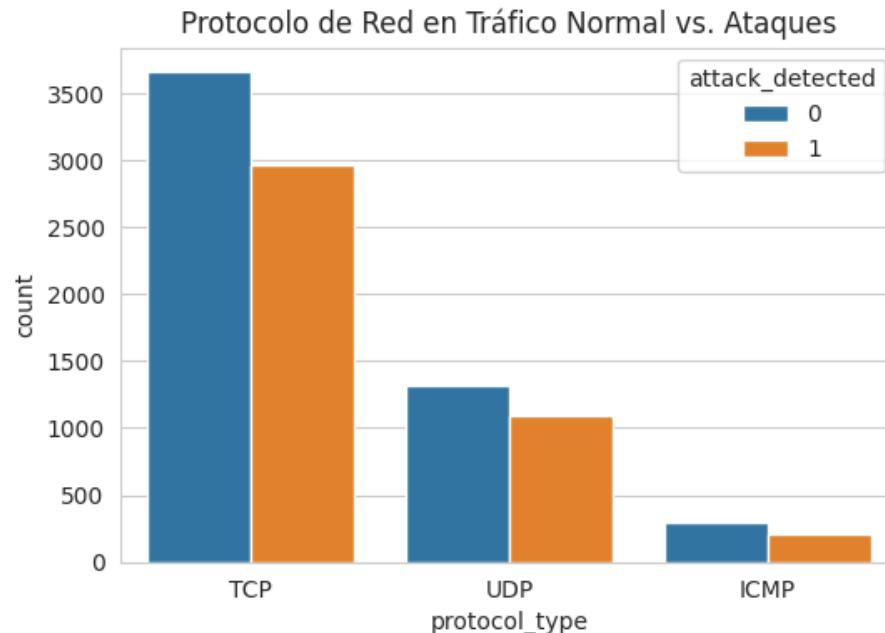


Ánalisis de accesos en horarios inusuales y su relación con ataques

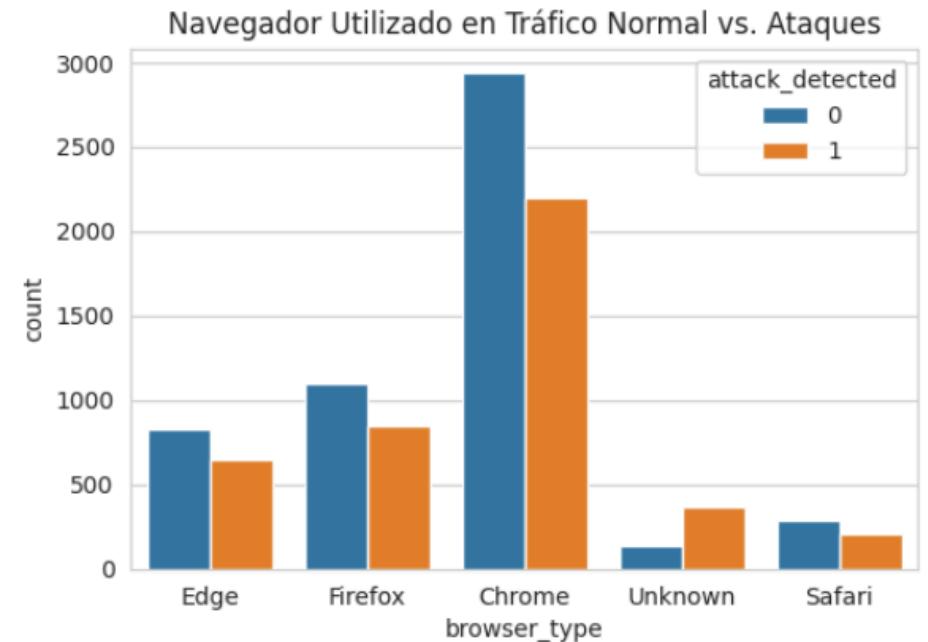
Si encontramos que el 80% de los ataques ocurren fuera del horario habitual, podríamos usar esta señal para reforzar alertas.



Análisis de protocolos de red en tráfico normal vs. ataques

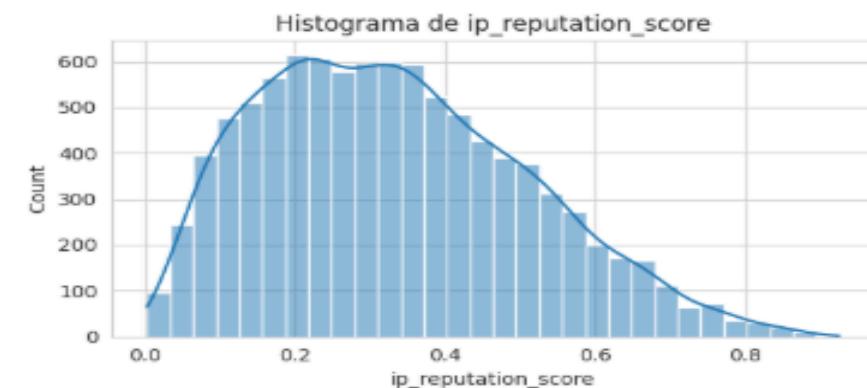
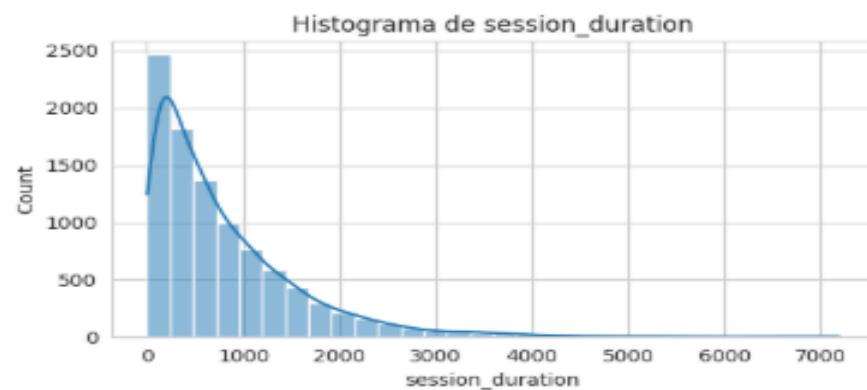
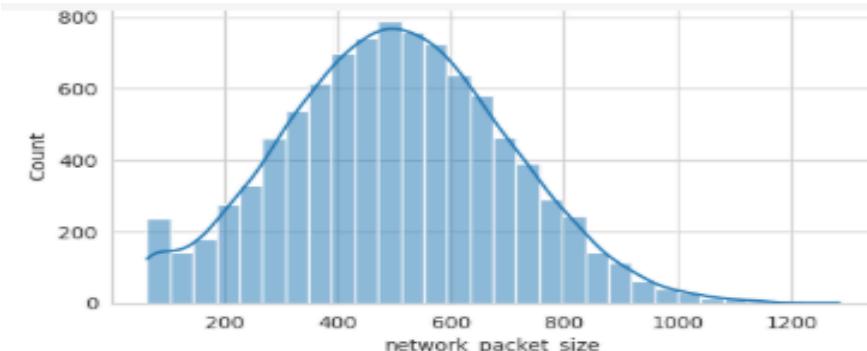


Análisis por navegador utilizado en tráfico normal vs. ataques

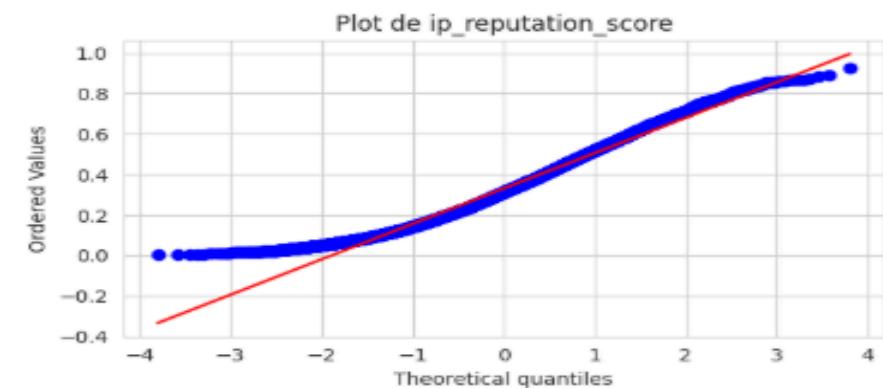
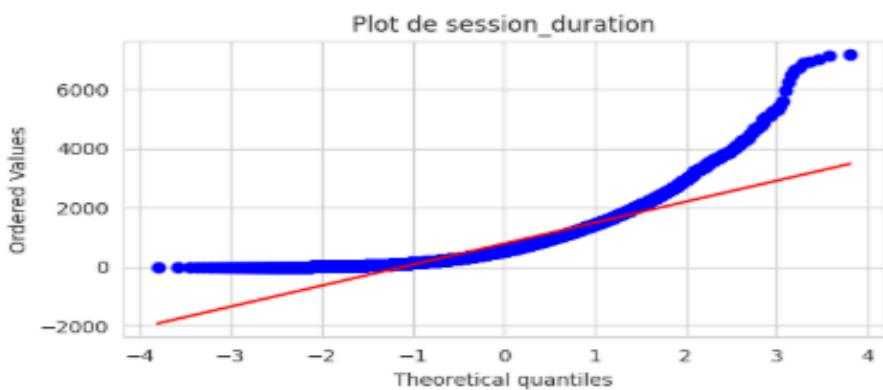
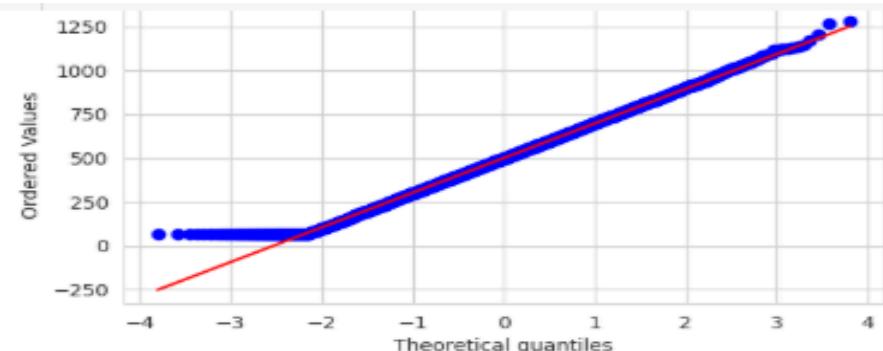


Ataques con navegadores desconocidos → actividad automatizada.

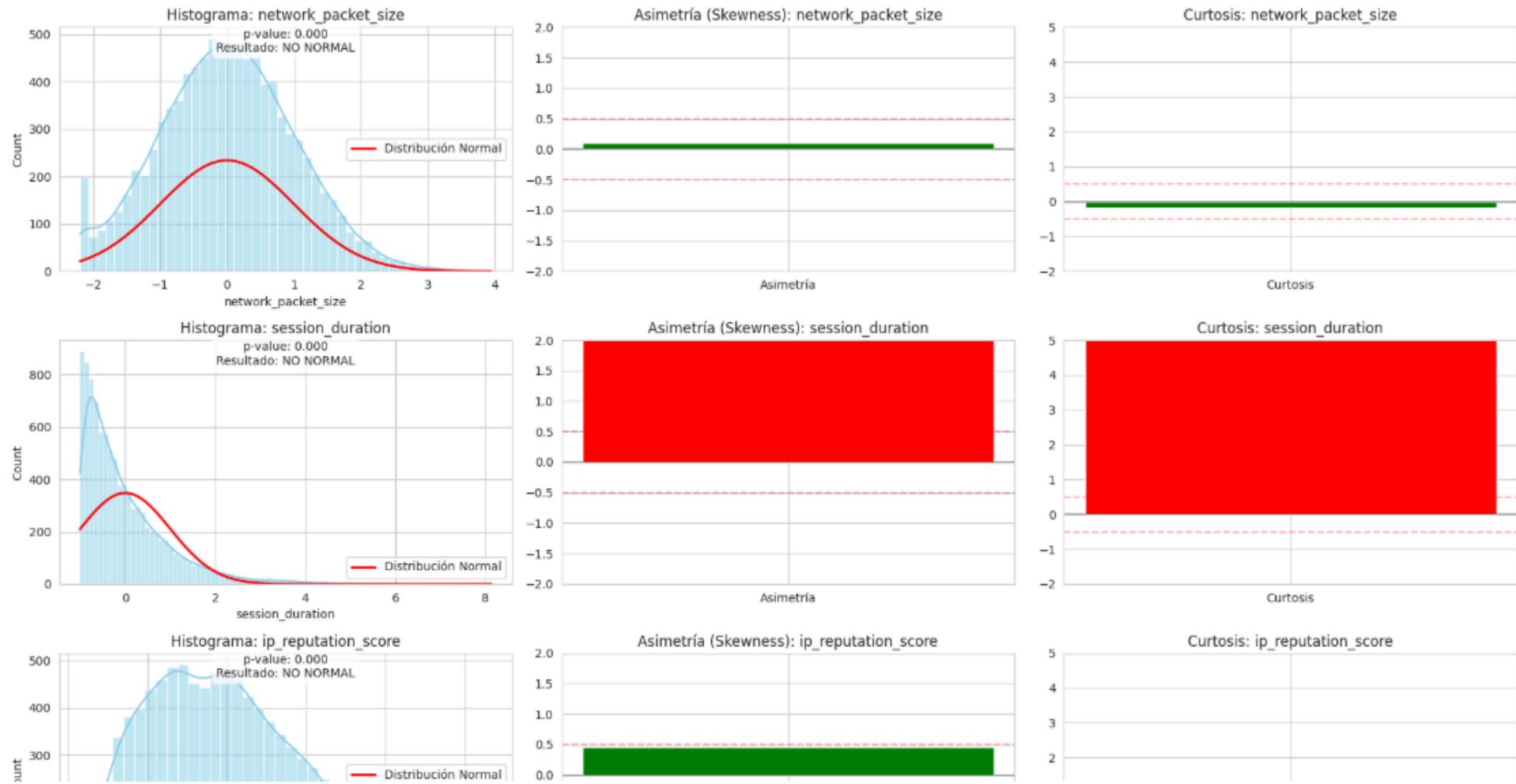
Prueba de Kolmogorov-Smirnov (K-S)



Kolmogorov-Smirnov p-value: 0.000
Los datos NO siguen una distribución normal.



Análisis de Normalidad: D'Agostino-Pearson Test



Conclusiones

El análisis exploratorio de datos sobre ciberseguridad reveló patrones clave en los intentos de intrusión, destacando que los fallos en login y la reputación IP son indicadores críticos de ataques. Se identificó una mayor vulnerabilidad en ciertos protocolos de red, particularmente en la distinción entre TCP y UDP. Además, los accesos en horarios inusuales mostraron una fuerte asociación con actividad maliciosa. El preprocesamiento de datos mejoró la calidad del análisis, permitiendo generar correlaciones significativas y evidenciar tendencias clave que pueden fortalecer la detección temprana de amenazas en entornos digitales.