TRABAJO PRÁCTICO OBLIGATORIO

SUBE Transacciones diarias

Valentín Lanús, Evangelina Ovelar, Santino Lozano
Materia: Ciencias de Datos – UADE

Planteamiento del Problema y Objetivo

El transporte público, como servicio esencial, requiere una planificación rigurosa. Su eficiencia operativa (frecuencia, rutas y asignación de recursos) depende directamente de una comprensión precisa del comportamiento de los usuarios. La tarjeta SUBE genera un volumen masivo de datos transaccionales diarios, reflejando patrones de uso a nivel temporal y geográfico.

El desafío: La incertidumbre en la demanda diaria impacta directamente en la calidad del servicio y en la optimización de los costos operativos.

Objetivo Principal del Proyecto

El objetivo central es desarrollar un modelo predictivo robusto para **anticipar la cantidad de validaciones diarias de tarjetas SUBE** en la Provincia de Buenos Aires.

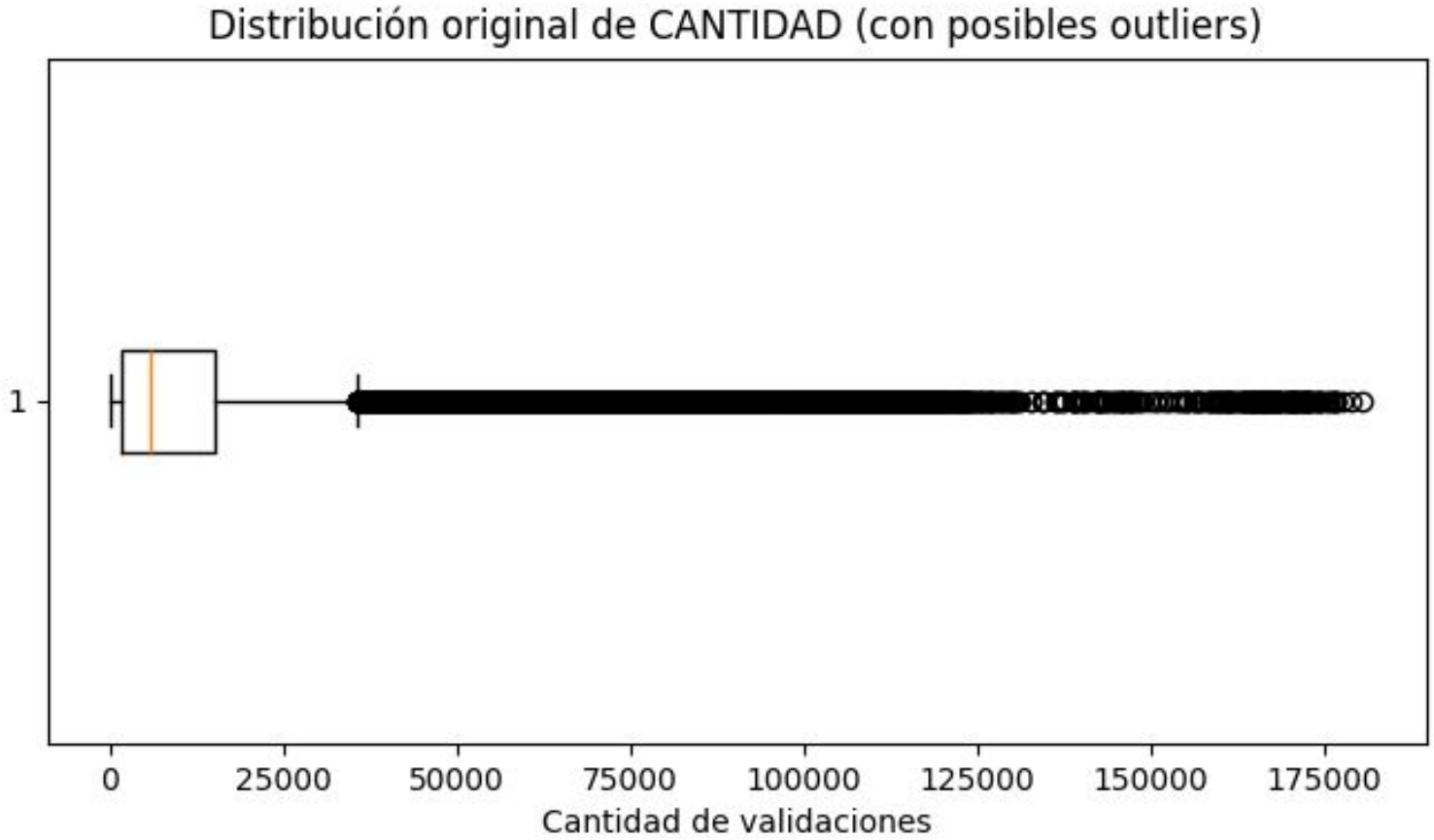
Este análisis busca:

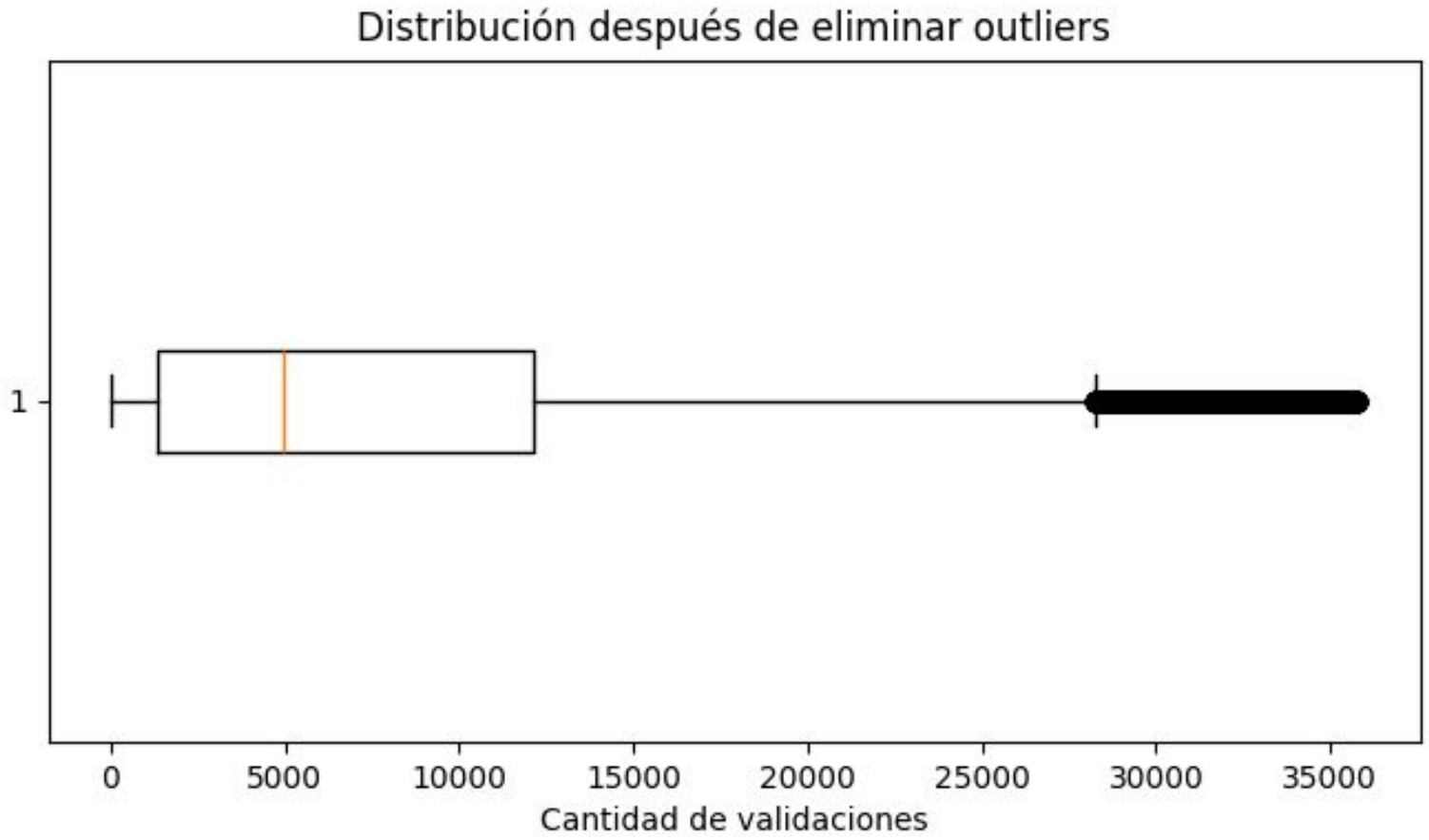
1. **Optimizar la planificación:** Permitir una asignación más eficiente de unidades y personal, minimizando demoras y sobrecargas.
2. **Generar información estratégica:** Ofrecer información valiosa a *stakeholders* para la toma de decisiones sobre tarifas, subsidios y expansión de rutas.
3. **Establecer un baseline:** Crear un modelo de referencia (*baseline*) basado en Machine Learning que capture los patrones de uso semanales y estacionales.

Limpieza y preparación de datos

Flujo de Trabajo Aplicado:

1. **Filtro Geográfico (Scope):** El análisis se centró exclusivamente en los registros correspondientes a la **Provincia de Buenos Aires** (PROVINCIA = 'BUENOS AIRES').
2. **Conversión y Validación:** Se verificó que la variable clave DIA_TRANSPORTE fuera del tipo datetime y se eliminaron registros con valores nulos o con CANTIDAD (validaciones) ≤ 0.
3. **Agrupación:** Los datos se agregaron a nivel diario (DIA_TRANSPORTE) para obtener la suma total de validaciones (uso_diario), creando nuestra serie de tiempo fundamental.
4. **Tratamiento de Outliers:** Se realizó un análisis exhaustivo de la variable CANTIDAD. Para mantener la robustez del modelo, se identificaron y eliminaron valores extremos (Outliers) en la serie de tiempo diaria utilizando el **método del Rango Intercuartílico (IQR)**, lo que asegura que las anomalías no distorsionen los patrones de uso normales.


Distribución original de CANTIDAD (con posibles outliers)
1
0
25000
50000
75000
100000
125000
150000
175000
Cantidad de validaciones


Distribución después de eliminar outliers
1
0
5000
10000
15000
20000
25000
30000
35000
Cantidad de validaciones

Anomalías detectadas y ajustes de datos

Análisis de Patrones Anómalos (Heatmap Semanal)

Durante la exploración detallada (usando *heatmaps* semanales y análisis de promedios), se identificaron caídas atípicas en la cantidad de validaciones en dos fechas específicas:

- **Jueves 11 de abril**
- **Jueves 9 de mayo**

Estas caídas fueron consideradas **anomalías o shocks discretos**, ya que no se correlacionaban con feriados o patrones estacionales conocidos.

Anomalías detectadas y ajustes de datos

Estrategia de Normalización

Para evitar que estos valores atípicos sesgaran el entrenamiento del modelo de Machine Learning y distorsionaran la tendencia general:

- Se aplicó un **ajuste proporcional** a los valores de esos dos días.
- El ajuste se basó en el promedio de las validaciones de otros jueves cercanos.

Resultado: Se logró mantener la coherencia de la serie de tiempo, asegurando que las caídas detectadas no fueran interpretadas por el modelo como patrones recurrentes, sino como eventos aislados.

Análisis exploratorio de datos

Indicadores Clave del Dataset SUBE — Buenos Aires

1. Volumen general

- Total de validaciones registradas: **976,879,363**
- Días cubiertos por el dataset: **366**
- Outliers eliminados: **10452**

2. Cobertura operativa

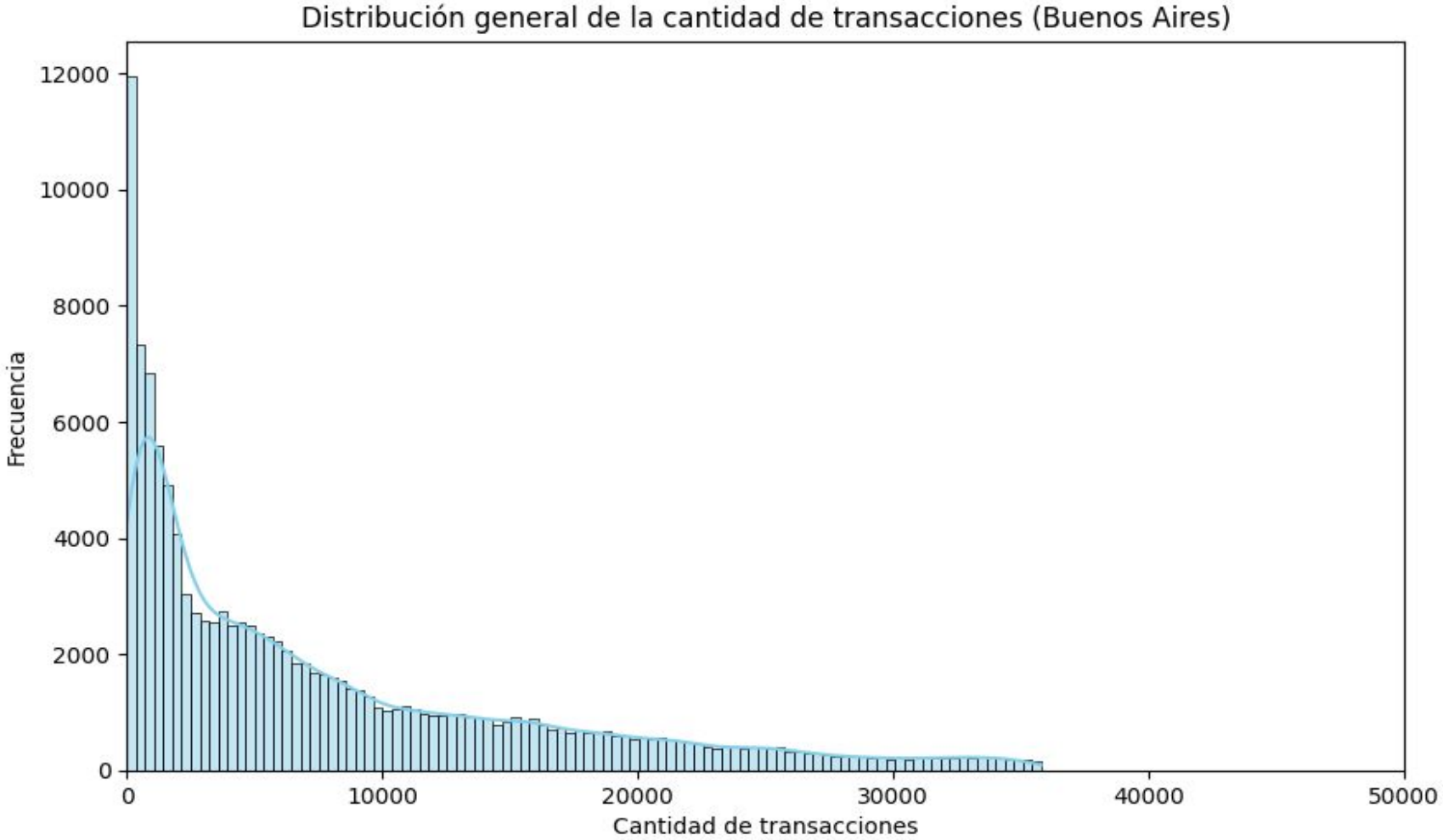
- Empresas activas: **121**
- Líneas de transporte analizadas: **370**
- Empresa con mayor cantidad de validaciones: **EMPRESA DE TRANSPORTE PERALTA RAMOS SACI**

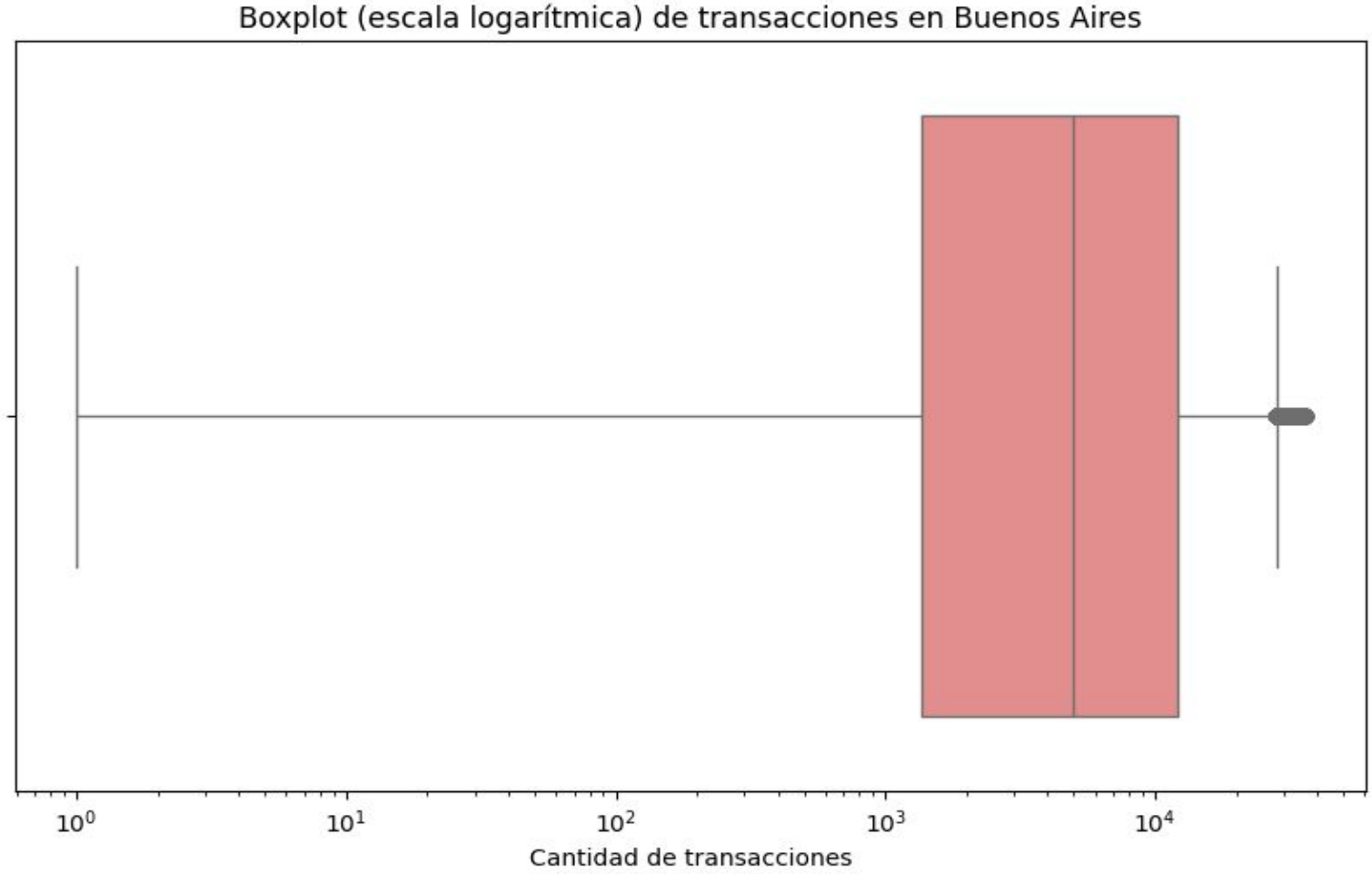
3. Promedios

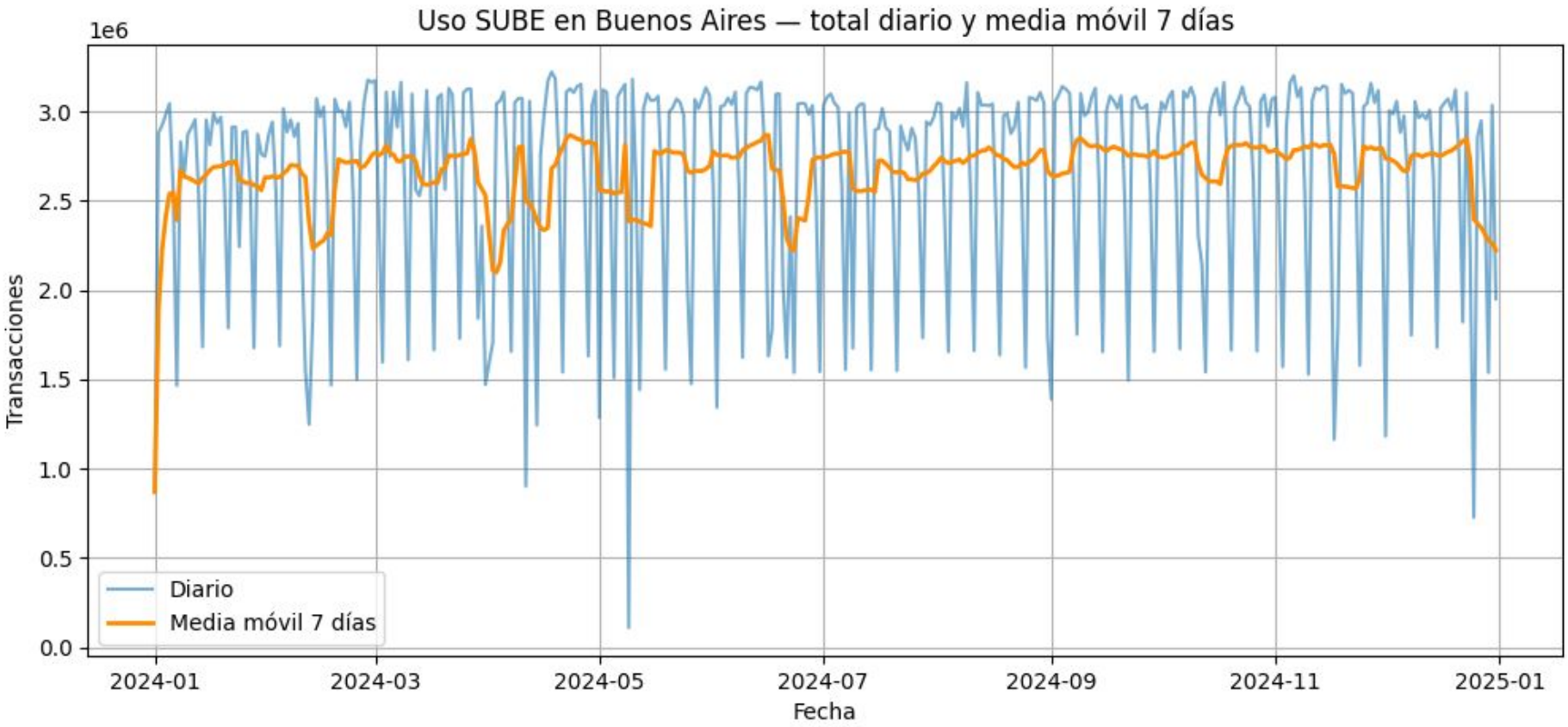
- Promedio diario de transacciones: **2,669,069**
- Promedio por empresa: **9,031**

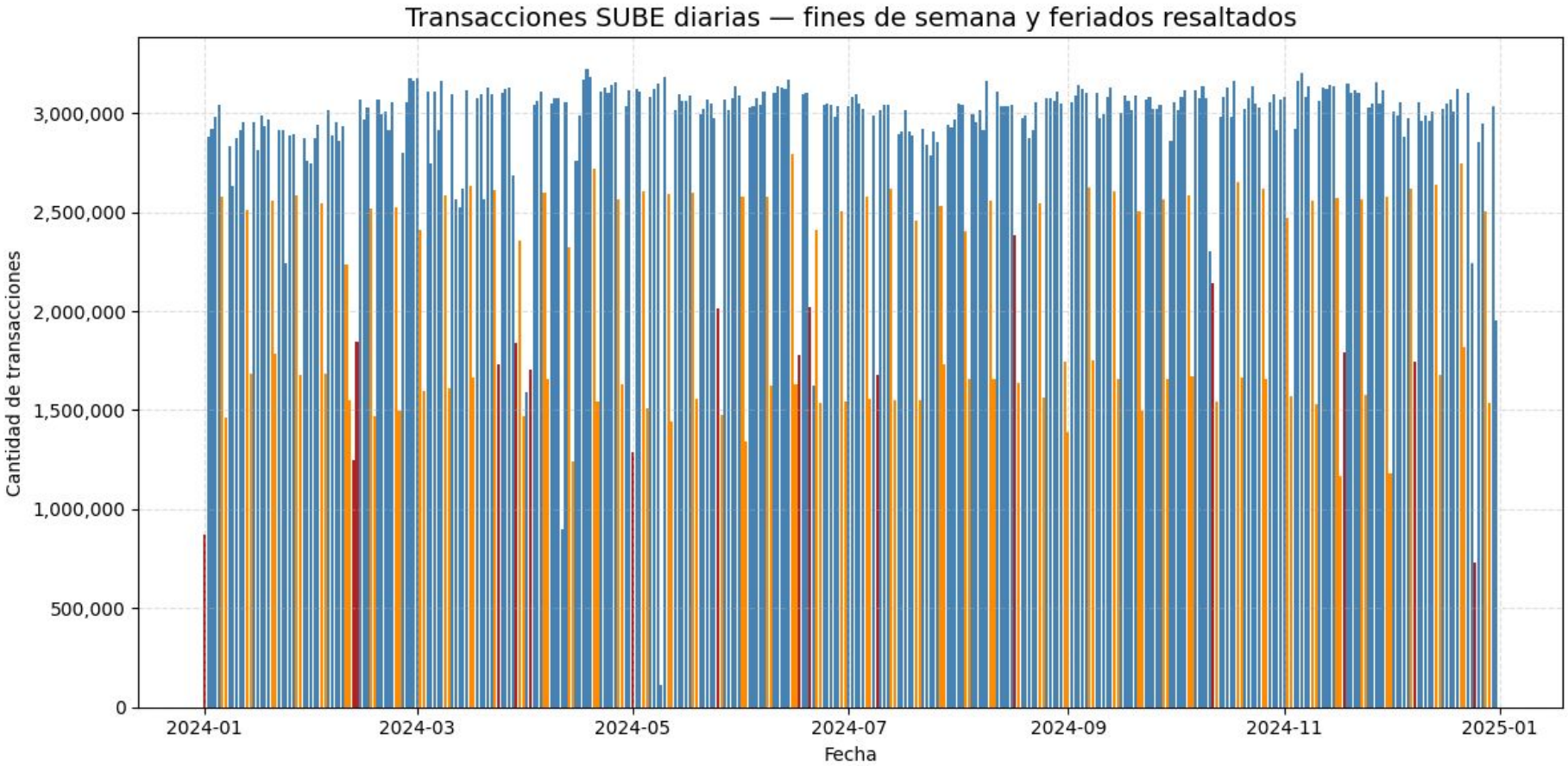
4. Distribución temporal

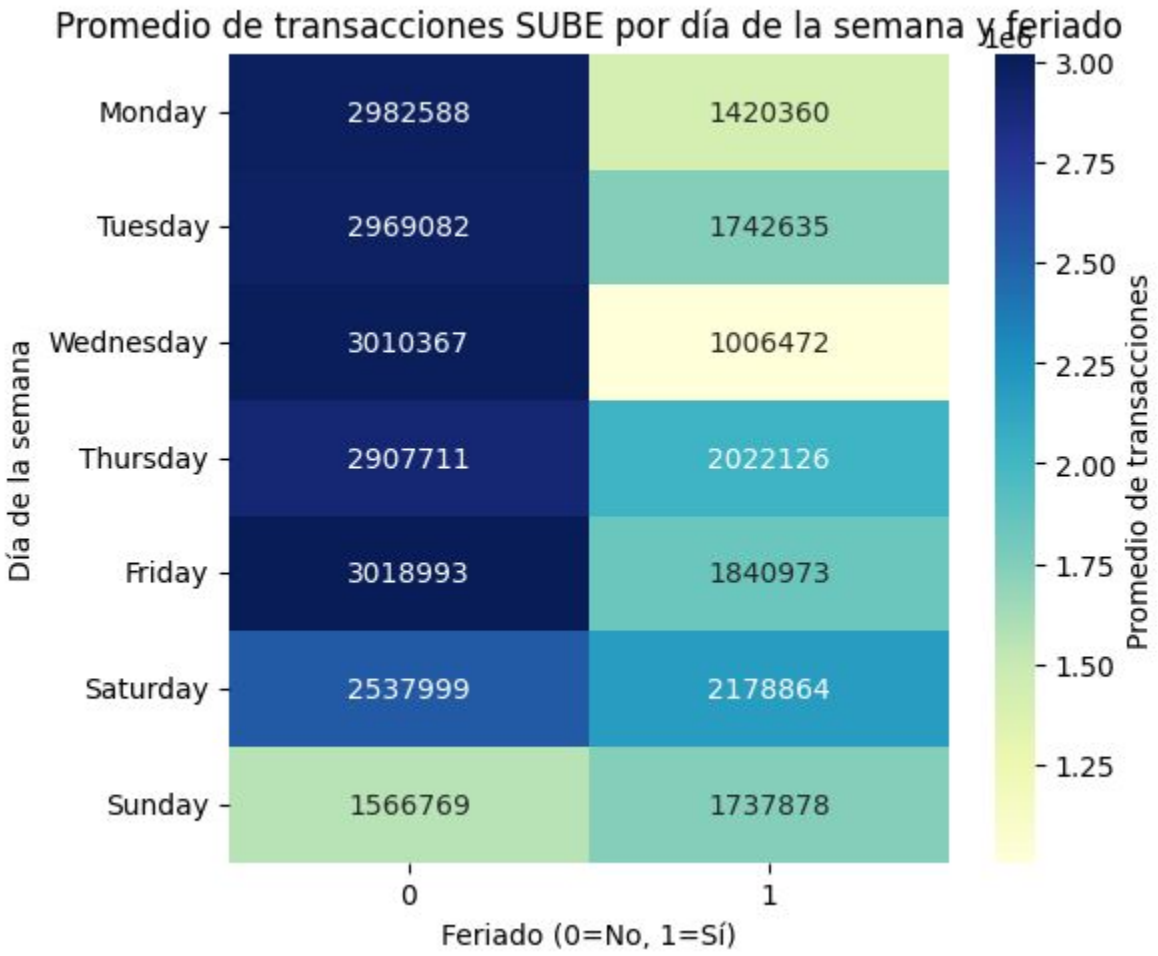
- Participación de fines de semana: **21.77%** del total

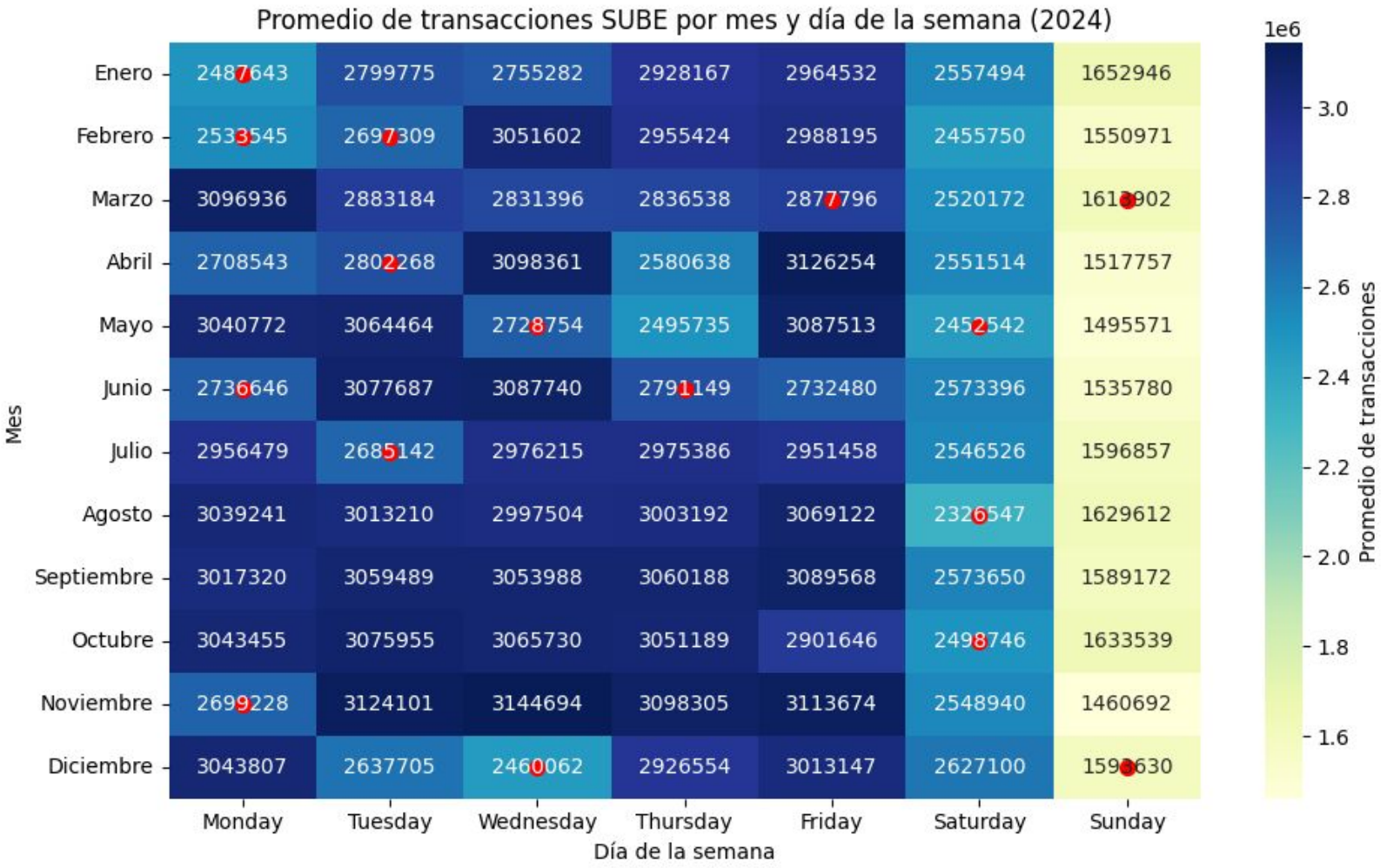

Distribución general de la cantidad de transacciones (Buenos Aires)
Frecuencia
Cantidad de transacciones
0
2000
4000
6000
8000
10000
12000
0
10000
20000
30000
40000
50000

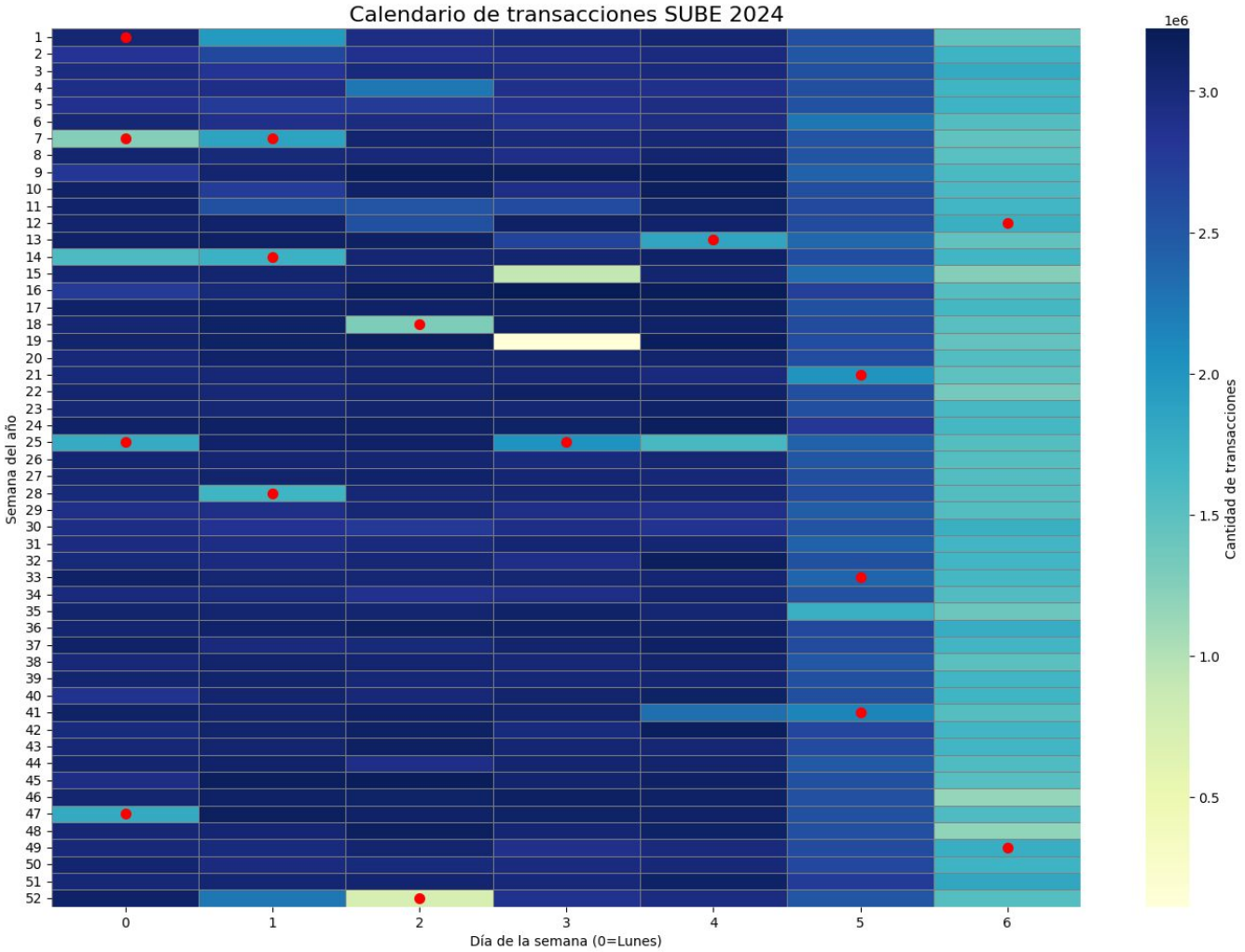

Boxplot (escala logarítmica) de transacciones en Buenos Aires
10^0
10^1
10^2
10^3
10^4
Cantidad de transacciones

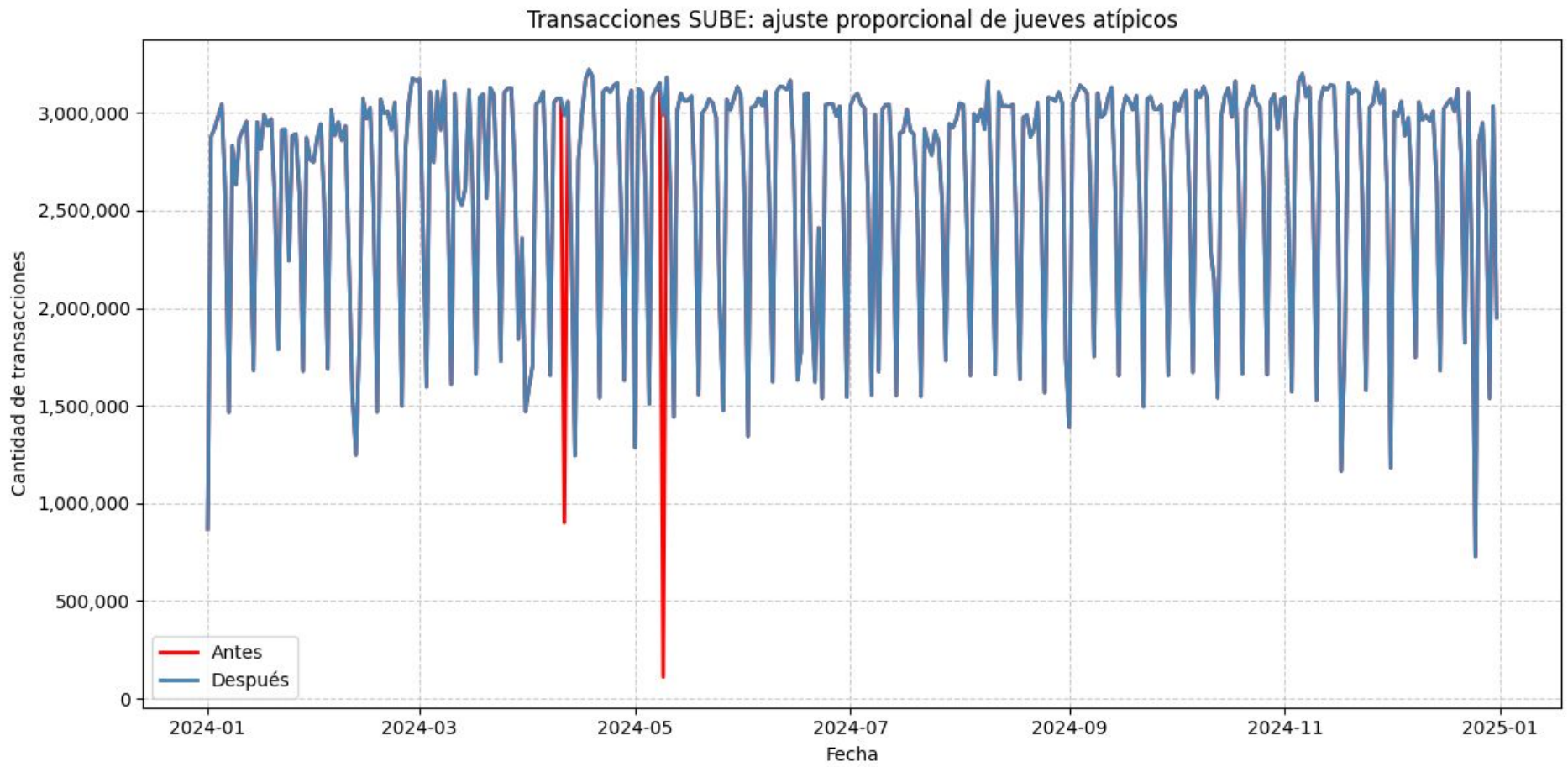

Uso SUBE en Buenos Aires — total diario y media móvil 7 días
1e6
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Transacciones
Diario
Media móvil 7 días
2024-01
2024-03
2024-05
2024-07
2024-09
2024-11
2025-01
Fecha

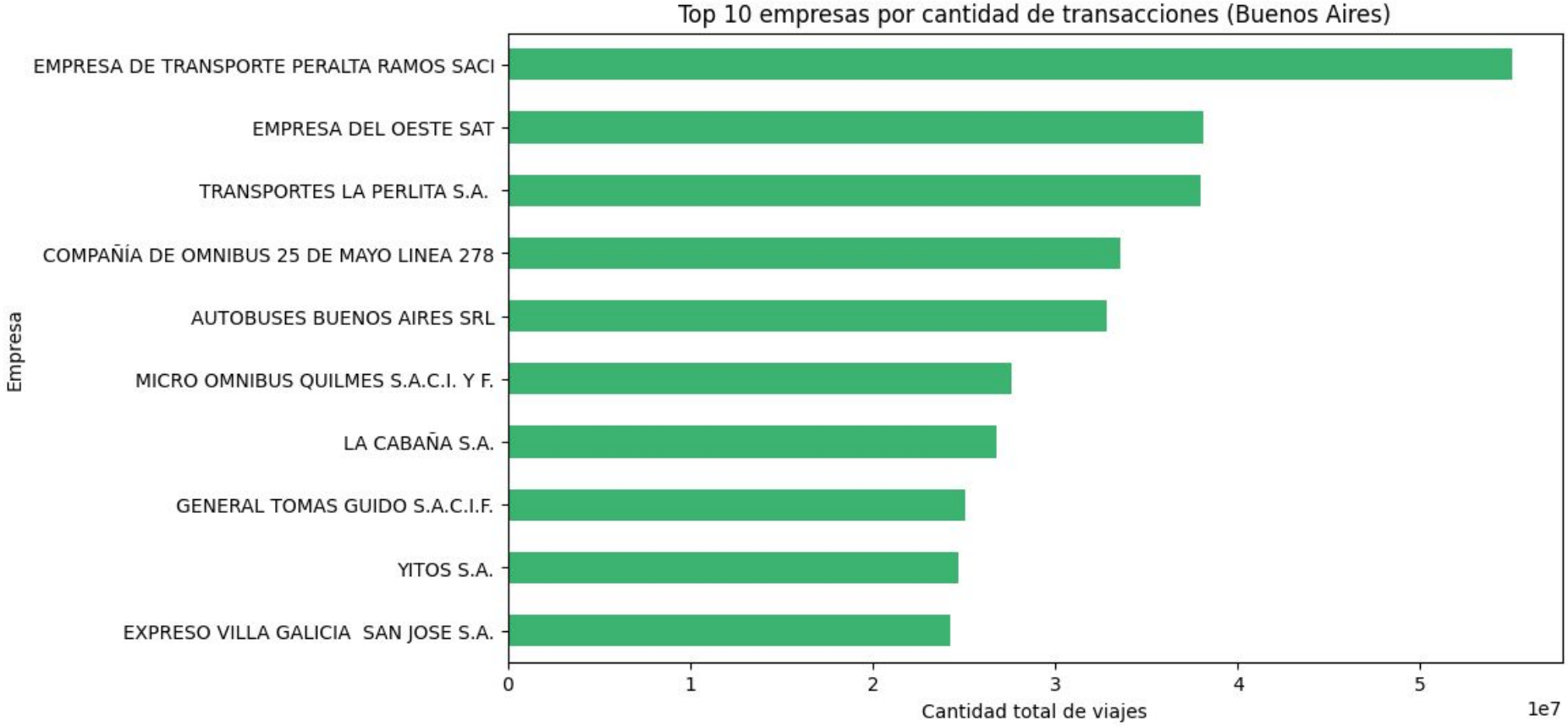

Transacciones SUBE diarias — fines de semana y feriados resaltados
Cantidad de transacciones
3,000,000
2,500,000
2,000,000
1,500,000
1,000,000
500,000
0
2024-01
2024-03
2024-05
2024-07
2024-09
2024-11
2025-01
Fecha

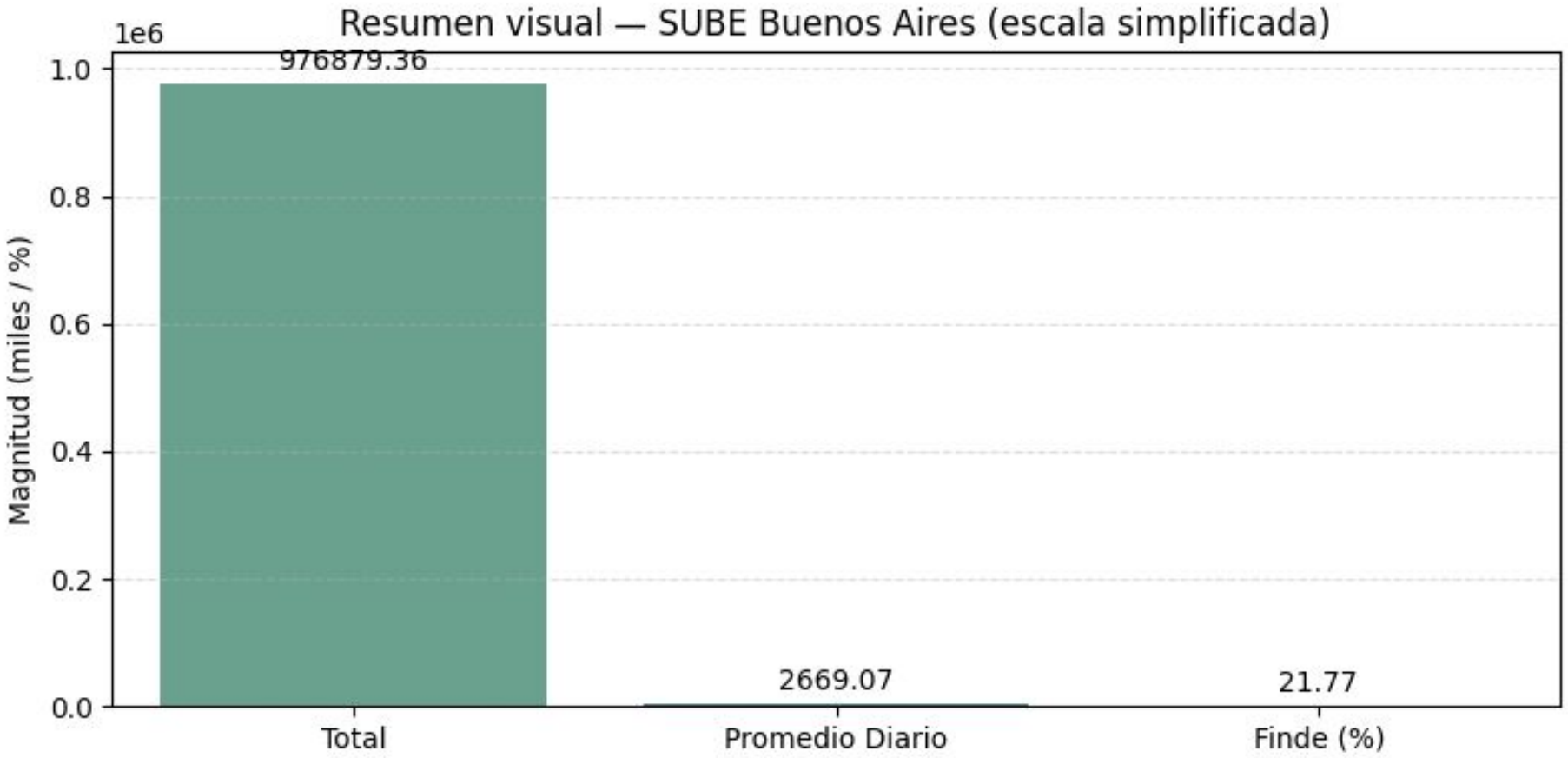

Promedio de transacciones SUBE por día de la semana y feriado
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
2982588
1420360
2969082
1742635
3010367
1006472
2907711
2022126
3018993
1840973
2537999
2178864
1566769
1737878
Día de la semana
0
1
Feriado (0=No, 1=Sí)
1e6
3.00
2.75
2.50
2.25
2.00
1.75
1.50
1.25
Promedio de transacciones


Promedio de transacciones SUBE por mes y día de la semana (2024)
Mes
Enero 2487643 2799775 2755282 2928167 2964532 2557494 1652946
Febrero 2533545 2697309 3051602 2955424 2988195 2455750 1550971
Marzo 3096936 2883184 2831396 2836538 2877796 2520172 1613902
Abril 2708543 2802268 3098361 2580638 3126254 2551514 1517757
Mayo 3040772 3064464 2728754 2495735 3087513 2452542 1495571
Junio 2736646 3077687 3087740 2791149 2732480 2573396 1535780
Julio 2956479 2685142 2976215 2975386 2951458 2546526 1596857
Agosto 3039241 3013210 2997504 3003192 3069122 2326547 1629612
Septiembre 3017320 3059489 3053988 3060188 3089568 2573650 1589172
Octubre 3043455 3075955 3065730 3051189 2901646 2498746 1633539
Noviembre 2699228 3124101 3144694 3098305 3113674 2548940 1460692
Diciembre 3043807 2637705 2460062 2926554 3013147 2627100 1596630
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Día de la semana
1e6
3.0
2.8
2.6
2.4
2.2
2.0
1.8
1.6
Promedio de transacciones


Calendario de transacciones SUBE 2024
Semana del año
Día de la semana (0=Lunes)
Cantidad de transacciones
1e6
3.0
2.5
2.0
1.5
1.0
0.5
0
1
2
3
4
5
6


Transacciones SUBE: ajuste proporcional de jueves atípicos
Cantidad de transacciones
3,000,000
2,500,000
2,000,000
1,500,000
1,000,000
500,000
0
Antes
Después
2024-01
2024-03
2024-05
2024-07
2024-09
2024-11
2025-01
Fecha


Top 10 empresas por cantidad de transacciones (Buenos Aires)
EMPRESA DE TRANSPORTE PERALTA RAMOS SACI
EMPRESA DEL OESTE SAT
TRANSPORTES LA PERLITA S.A.
COMPAÑÍA DE OMNIBUS 25 DE MAYO LINEA 278
AUTOBUSES BUENOS AIRES SRL
MICRO OMNIBUS QUILMES S.A.C.I. Y F.
LA CABAÑA S.A.
GENERAL TOMAS GUIDO S.A.C.I.F.
YITOS S.A.
EXPRESO VILLA GALICIA SAN JOSE S.A.
Empresa
0
1
2
3
4
5
1e7
Cantidad total de viajes


Resumen visual — SUBE Buenos Aires (escala simplificada)
1e6
976879.36
1.0
0.8
0.6
0.4
0.2
0.0
Magnitud (miles / %)
2669.07
21.77
Total
Promedio Diario
Finde (%)

Modelado del Machine Learning

El Enfoque de Machine Learning

Para predecir las validaciones desde el último día disponible en el dataset, adoptamos un enfoque de Machine Learning utilizando el algoritmo **RandomForestRegressor (RF)**.

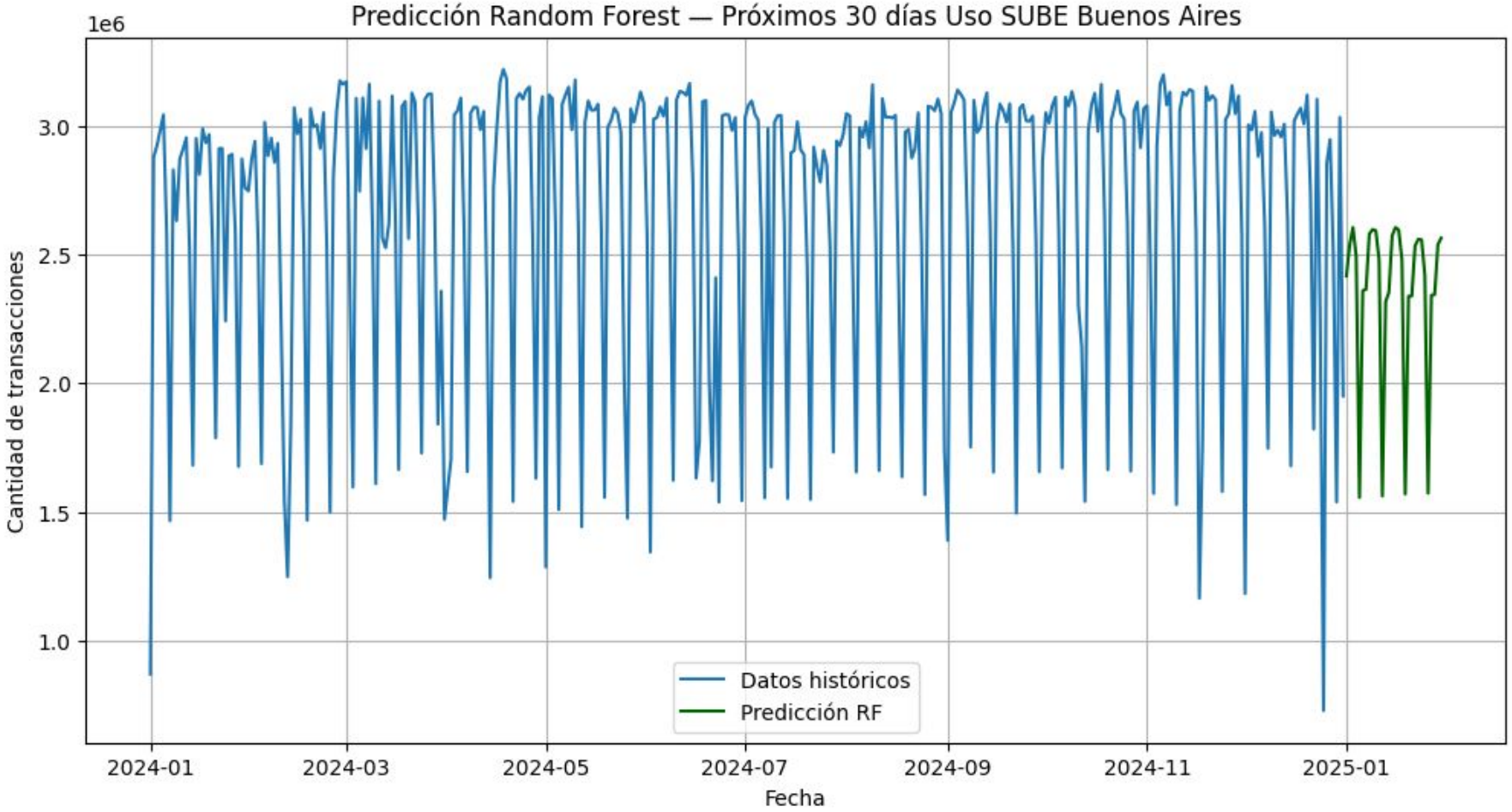
¿Por qué RandomForestRegressor?

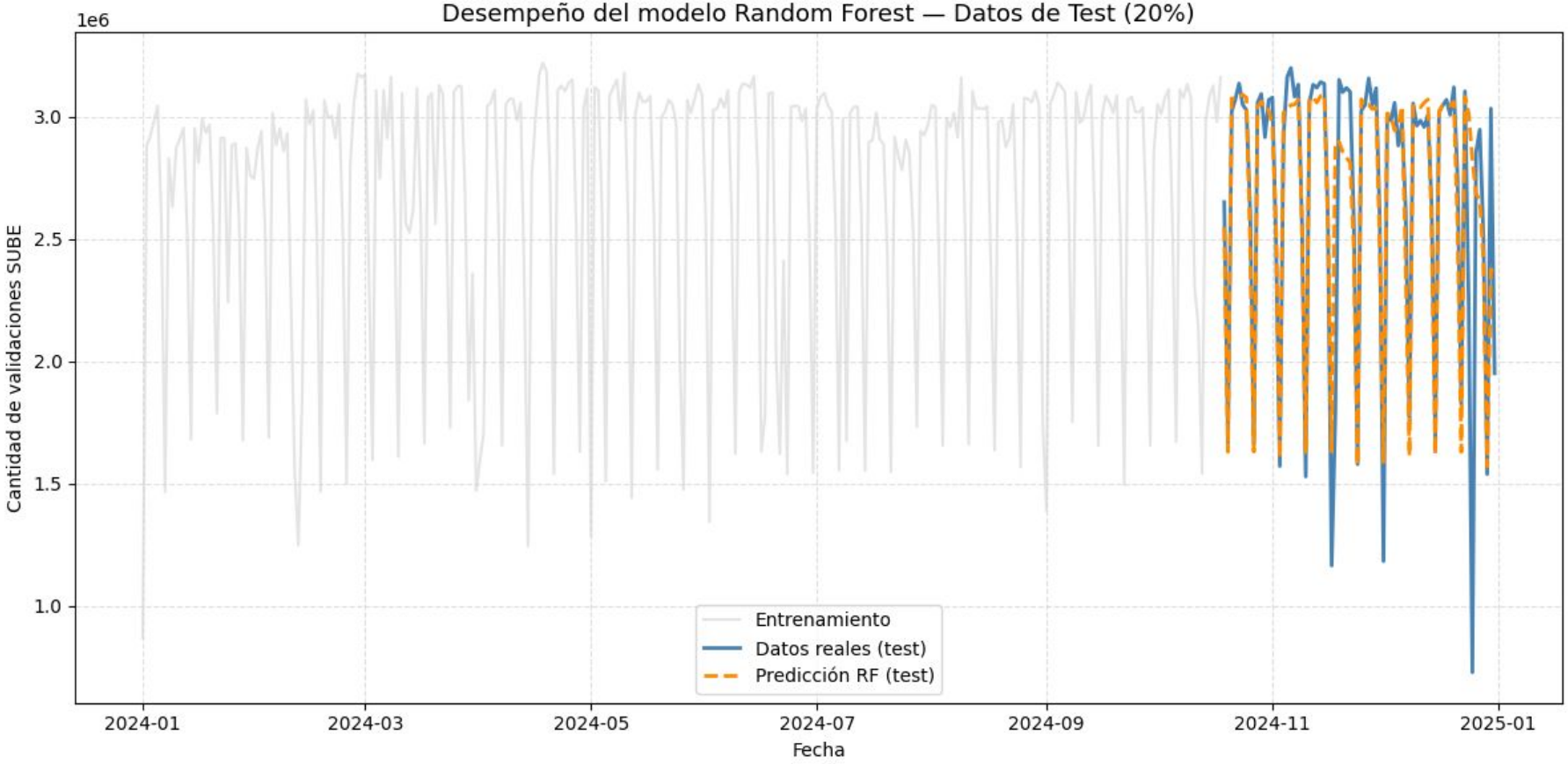
- **Robusto a Ruido:** El RF es un ensamble de múltiples árboles de decisión que promedia sus salidas, lo que lo hace **robusto a no linealidades y a pequeñas contaminaciones** en los datos.
- **Captura de Interacciones:** Es excelente para captar interacciones complejas entre variables (por ejemplo, el efecto combinado de ser "lunes" y, a la vez, "vacaciones de invierno").

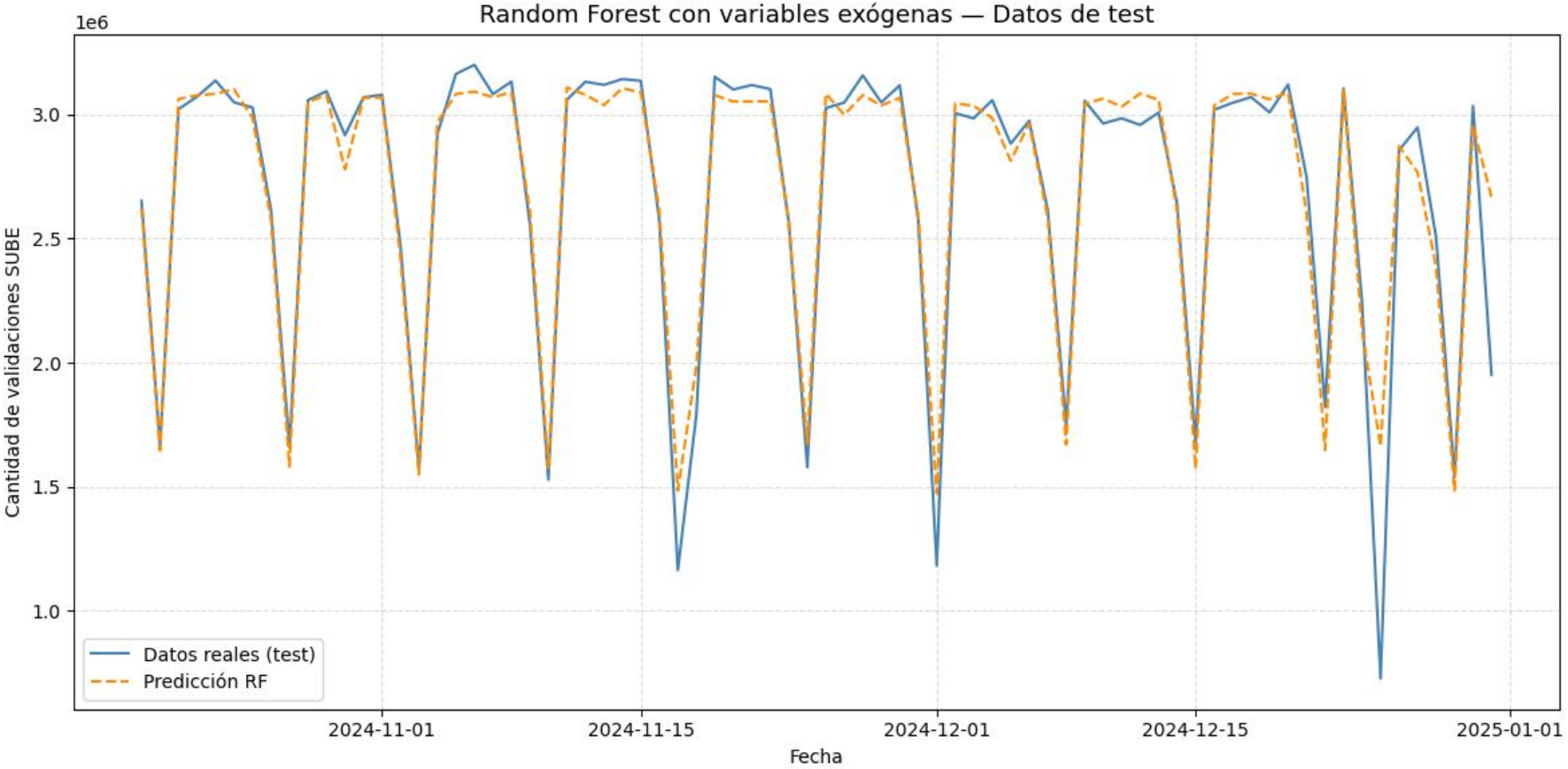
Modelado del Machine Learning

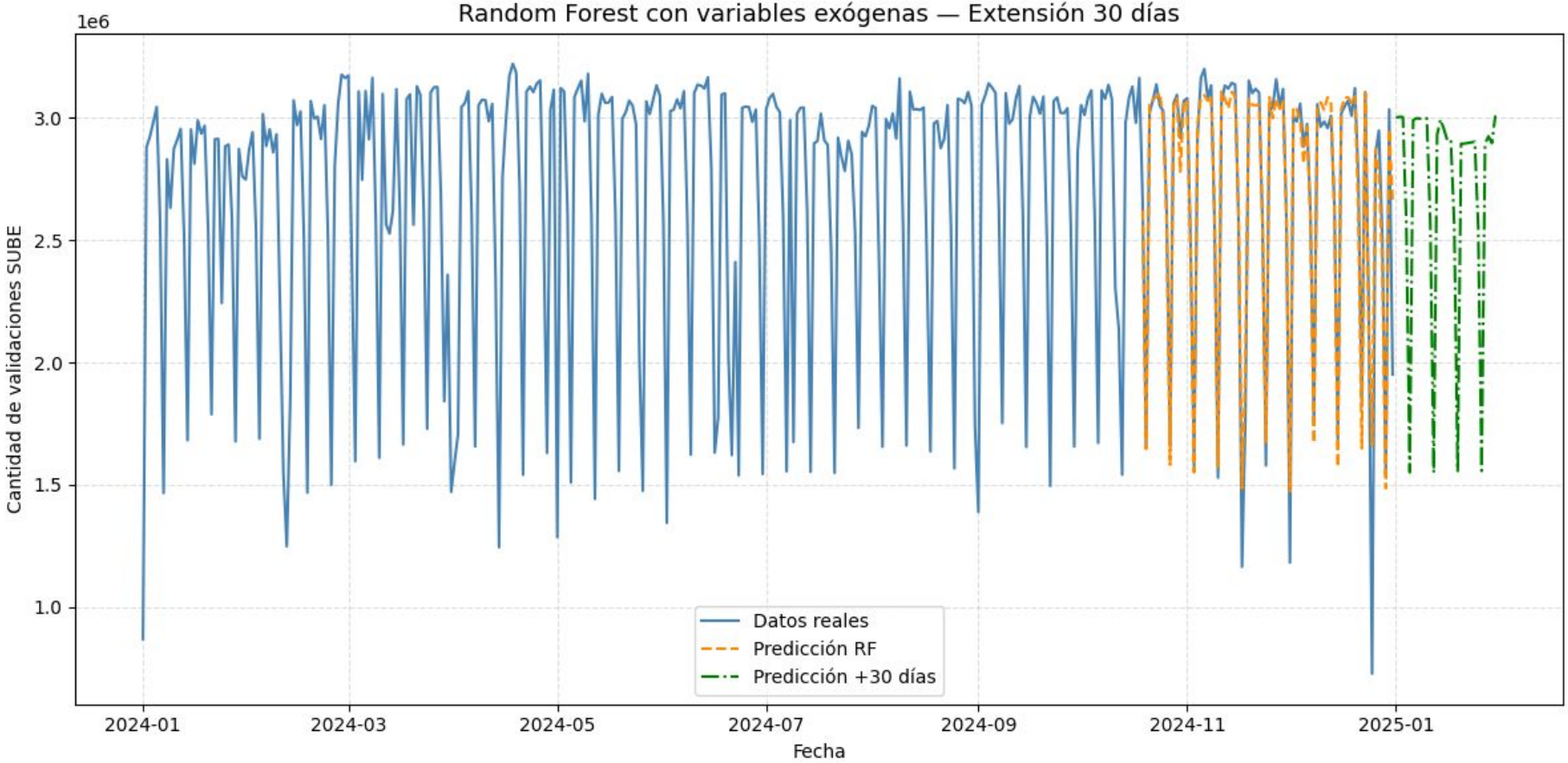
Variables (Features) Predictivas: El modelo se construye a partir de un conjunto de variables diseñadas para capturar la tendencia y la estacionalidad del uso de SUBE:

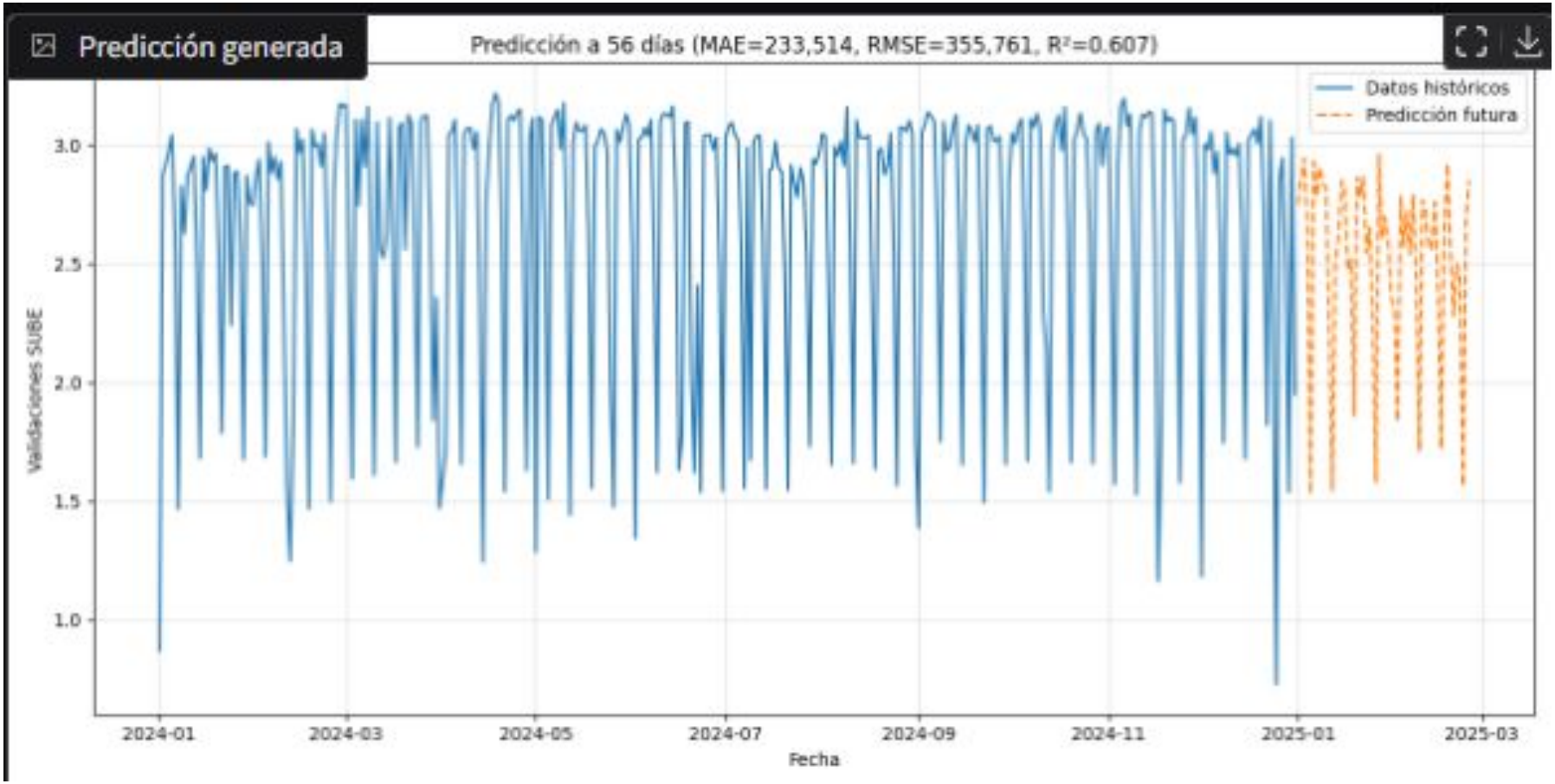
1. **Señales de Calendario:** Día de la semana (dow), indicador de fin de semana (es_finde), mes, día del mes, y número de semana ISO.
2. **Tendencia:** Contador de tiempo (dia_num) y la Media Móvil de 7 días (rolling_7).
3. **Shocks Discretos:** Indicadores binarios para **feriados** y **vacaciones de invierno**.


Predicción Random Forest — Próximos 30 días Uso SUBE Buenos Aires
1e6
3.0
2.5
2.0
1.5
1.0
Cantidad de transacciones
Datos históricos
Predicción RF
2024-01
2024-03
2024-05
2024-07
2024-09
2024-11
2025-01
Fecha


Desempeño del modelo Random Forest — Datos de Test (20%)
1e6
3.0
2.5
2.0
1.5
1.0
Cantidad de validaciones SUBE
Entrenamiento
Datos reales (test)
Predicción RF (test)
2024-01
2024-03
2024-05
2024-07
2024-09
2024-11
2025-01
Fecha


Random Forest con variables exógenas — Datos de test
1e6
3.0
2.5
2.0
1.5
1.0
Cantidad de validaciones SUBE
Datos reales (test)
Predicción RF
2024-11-01
2024-11-15
2024-12-01
2024-12-15
2025-01-01
Fecha


Random Forest con variables exógenas — Extensión 30 días
1e6
3.0
2.5
2.0
1.5
1.0
Cantidad de validaciones SUBE
2024-01
2024-03
2024-05
2024-07
2024-09
2024-11
2025-01
Fecha
Datos reales
Predicción RF
Predicción +30 días


Predicción generada
Predicción a 56 días (MAE=233,514, RMSE=355,761, R²=0.607)
Datos históricos
Predicción futura
Validaciones SUBE
3.0
2.5
2.0
1.5
1.0
2024-01
2024-03
2024-05
2024-07
2024-09
2024-11
2025-01
2025-03
Fecha

Hipótesis

La demanda diaria de validaciones SUBE en la Provincia de Buenos Aires está determinada por un patrón temporal estable y predecible, dominado por la estacionalidad semanal (picos en días hábiles y caídas en fines de semana), shocks de calendario (feriados y vacaciones) y variaciones estacionales moderadas a lo largo del año. Este comportamiento se sostiene siempre que el dato sea consistente y depurado y que se consideren explícitamente los eventos de calendario en el análisis.