

# ¿Quién gasta más y quién se registra? Evidencia observacional en CheMarket

Adrián Arturo Suárez García  
202123771  
a.suarezg@uniandes.edu.co

Luis Alejandro Rubiano Guerrero  
202013482  
la.rubiano@uniandes.edu.co

Gabriel Alejandro Moreno Riveros  
202014583  
g.morenor@uniandes.edu.co

Juan Sebastián Sierra Tarazona  
202123725  
j.sierrat@uniandes.edu.co

## I. INTRODUCCIÓN

A través de este informe, nuestro grupo de economistas y científicos de datos emplea diferentes técnicas de análisis estadístico y de machine learning para estudiar el comportamiento de los usuarios en **CheMarket Inc.** En particular, nos enfocamos en entender qué variables impulsan el revenue de la compañía, con el fin de generar evidencia que apoye la toma de decisiones estratégicas informadas.

El análisis se divide en dos partes complementarias. En la primera parte, trabajamos con datos observacionales para explorar patrones en el comportamiento de los usuarios: ¿quiénes gastan más?, ¿qué factores están asociados con el registro en la plataforma?, ¿qué variables ayudan a predecir el gasto individual? A partir de regresiones y modelos predictivos evaluamos si existe una diferencia sistemática en el gasto entre usuarios registrados y no registrados, y discutimos posibles sesgos que afectan la interpretación causal.

En la segunda parte, analizamos un experimento aleatorio en el que se facilitó el proceso de registro para un grupo de usuarios. Esta intervención nos permite estimar de manera más rigurosa el efecto del registro sobre el gasto, así como evaluar la efectividad del nuevo diseño en incrementar la tasa de registros. Finalmente, reflexionamos sobre la validez de los resultados, las limitaciones del experimento y presentamos recomendaciones concretas sobre si conviene escalar esta intervención.

## II. DATOS OBSERVACIONALES: ¿QUÉ IMPULSA LAS VENTAS?

### 1. Datos y preparación

Para el análisis observacional, contamos con un conjunto de 100.000 observaciones históricas de los usuarios de CheMarket. Las variables disponibles son `time_spent`: tiempo en el sitio durante la sesión, `past_sessions`: número de sesiones anteriores, `os_type`: sistema operativo (OS X, Windows, otros), `is_returning_user`: si el usuario ya había visitado antes, `sign_up`: si se registró o no y `revenue`: cuanto gastó en cada sesión.

En la siguiente tabla se presentan las estadísticas descriptivas de estas variables.

Tabla I  
ESTADÍSTICAS DESCRIPTIVAS DE LAS VARIABLES

	Tiempo	Sesiones	OS	Recurrente?	Revenue
Mín	0.000125	0	Otro: 9909	No: 4930	0.5451
25%	1.4247	2	OS X: 30253	Sí: 95070	2.3400
Mediana	3.4384	3	Windows: 59838		3.1370
50%	4.9946	3.001			3.9766
75%	6.9117	4			4.5221
Máx	54.3989	14			36.2934

Estadísticas descriptivas de las variables numéricas y número por clase para las categóricas.

Transformaciones  
Partición  
Tabla de variables  
Tabla estadísticas descriptivas  
Matriz de correlaciones

### 2. Estimación del efecto de registrarse

Marco conceptual, sesgo por variables omitidas.  
Densidades  
Histogramas  
Boxplot

### 3. Efecto de registrarse sobre el gasto

Diferentes modelos  
pseudo coefplot coeficientes  
Interpretación  
multicolinealidad

### 4. Reflexión sobre causalidad y recomendación preliminar

## III. DATOS EXPERIMENTALES: ¿FUNCIONA FACILITAR EL REGISTRO?

### 1. Verificación del experimento

### 2. Efecto sobre el registro

### 3. Efecto sobre el gasto

### 4. Limitaciones y robustez

### 5. Recomendación final

## IV. CONCLUSIONES

Resumen  
Recomendaciones accionables.