

Analítica para los negocios

06327-ECO

Semana 06: Entender los datos antes de modelar — el EDA del caso
Cóndor

Eduard F. Martínez González, Ph.D.

Departamento de Economía, Universidad Icesi

8 de septiembre de 2026

Objetivo de la sesión

Hoy volvemos al caso Cóndor con **una pregunta de negocio concreta** y hacemos, en vivo, el primer análisis exploratorio que esa pregunta necesita.

- Van a ver el **razonamiento**, no un catálogo de gráficos: qué se pregunta el analista en cada paso, qué mira y qué decide con lo que ve.
- Los datos de Cóndor están **limpios**. No vamos a “arreglar” nada: vamos a **entenderlos** — el grano, los faltantes, las distribuciones, las comparaciones, las relaciones y lo raro.
- Cada paso se conecta con el podcast: las cuatro dimensiones, el “¿por qué falta?”, el atípico que no es error, la EDA que produce hipótesis.
- Al final el método queda en sus manos: en la práctica con el monitor construyen **su base analítica** y arrancan el EDA de **su** pregunta.

La idea que gobierna la sesión

El EDA **no responde** la pregunta de negocio: **la prepara**. Se sale de él con hallazgos, hipótesis y decisiones documentadas — y con la certeza de que los datos aguantan lo que viene.

¿Dónde estamos en el curso?

- **Semanas 1–4:** las herramientas. R, dplyr, ggplot2.
- **Semana 5:** el mapa. Las siete etapas del proceso analítico y, en clase, dos casos reales: de la necesidad vaga a la pregunta. Ustedes formularon la de su grupo sobre Cóndor.
- **Semana 6 (hoy):** las etapas 3 y 4 — conocer los datos antes de modelar. El podcast dio el *criterio*: fuentes, cuatro dimensiones de calidad, faltantes, atípicos, qué es la EDA. Hoy ese criterio se *ejecuta* sobre Cóndor.
- **Lo que sigue:** el examen parcial 1 (unidades 1 a 5) y la **Entrega 2** del proyecto: la base depurada y el EDA de su pregunta.

Su Entrega 1

Los lineamientos del proyecto fijan la Entrega 1 el **viernes 18 de septiembre**: la pregunta que formularon la semana pasada, ajustada con la retroalimentación. Todo lo de hoy es la materia prima de la Entrega 2.

Roadmap

- 1 La pregunta que manda
- 2 Qué es (y qué no es) un EDA
- 3 El EDA de Cóndor, paso a paso
- 4 Lo que sale del EDA
- 5 Su turno: la base analítica de su pregunta

Roadmap

- 1 La pregunta que manda
- 2 Qué es (y qué no es) un EDA
- 3 El EDA de Cóndor, paso a paso
- 4 Lo que sale del EDA
- 5 Su turno: la base analítica de su pregunta

La pregunta de negocio de hoy

Frente de retención (Growth)

*“¿Qué clientes de Cóndor que hoy están activos tienen alta probabilidad de **volverse inactivos** el próximo trimestre, para que Growth concentre en ellos la **campana de reactivación** en vez de enviársela a los 8.000?”*

La pregunta de negocio de hoy

Frente de retención (Growth)

*“¿Qué clientes de Cándor que hoy están activos tienen alta probabilidad de **volverse inactivos** el próximo trimestre, para que Growth concentre en ellos la **campana de reactivación** en vez de enviársela a los 8.000?”*

- **Específica** — qué: volverse inactivo (abandono); quién: los 8.000 clientes de la base al corte del 31-mar-2026; cuándo: el trimestre siguiente.
- **Accionable** — la decisión es *a quién enviar* la campaña. Hoy va a ciegas: 13.467 envíos (ene-2025 a mar-2026), 10 % de conversión, unos \$2.250 por envío.
- **Medible** — abandono existe para los 8.000; 1 de cada 5 se va.

La pregunta de negocio de hoy

Frente de retención (Growth)

*“¿Qué clientes de Cónдор que hoy están activos tienen alta probabilidad de **volverse inactivos** el próximo trimestre, para que Growth concentre en ellos la **campaña de reactivación** en vez de enviársela a los 8.000?”*

- **Específica** — qué: volverse inactivo (abandono); quién: los 8.000 clientes de la base al corte del 31-mar-2026; cuándo: el trimestre siguiente.
- **Accionable** — la decisión es *a quién enviar* la campaña. Hoy va a ciegas: 13.467 envíos (ene-2025 a mar-2026), 10 % de conversión, unos \$2.250 por envío.
- **Medible** — abandono existe para los 8.000; 1 de cada 5 se va.

¿Qué tarea analítica es? ¿Cuál es la vara?

Output = categoría + target histórico $\Rightarrow$ **clasificación supervisada**. **Baseline**: “todos se quedan” acierta el 77,85 %; cualquier modelo debe ganarle.

Qué le pide esta pregunta al EDA

La pregunta convierte 29 columnas en **nueve preguntas de analista**. Ese es el mapa de la sesión:

Paso	Lo que se pregunta el analista	Lo que mira
1	¿Qué es una fila y cuántas hay?	tamaño, llaves, duplicados, las 4 tablas
2	¿Qué falta y <i>por qué</i> falta?	faltantes por columna y a quién le faltan
3	¿Cómo se ve el target?	la tasa base y la definición de “activo”
4	¿Cómo se distribuyen las candidatas?	histogramas: colas, media vs. mediana
5	¿Quiénes abandonan más?	tasa de abandono por tramos de comportamiento
6	¿Y lo que “debería” importar?	tasa por canal, ciudad, edad, género
7	¿Las señales miden lo mismo?	el cruce de dos variables
8	¿Qué aporta un anexo?	agregar por cliente y luego unir
9	¿Qué raro quedó — y qué <i>no</i> puedo usar?	colas, atípicos, fugas de información

Sin la pregunta, cada paso admite decenas de gráficos. **Con la pregunta, uno o dos.**

Roadmap

- 1 La pregunta que manda
- 2 Qué es (y qué no es) un EDA
- 3 El EDA de Cóndor, paso a paso
- 4 Lo que sale del EDA
- 5 Su turno: la base analítica de su pregunta

Entender antes de modelar

Lo que la EDA busca (el podcast):

- 1 **Distribuciones:** cómo se comporta cada variable sola.
- 2 **Resúmenes por grupo:** si hay diferencias reales entre segmentos.
- 3 **Relaciones:** si dos variables tienen que ver entre sí.
- 4 **Anomalías residuales:** lo raro que el diagnóstico no cazó.

Lo que la EDA no es:

- La respuesta a la pregunta de negocio.
- Los gráficos “bonitos” de la presentación final.
- Un catálogo: una lámina por cada columna de la base.

Entender antes de modelar

Lo que la EDA busca (el podcast):

- 1 **Distribuciones:** cómo se comporta cada variable sola.
- 2 **Resúmenes por grupo:** si hay diferencias reales entre segmentos.
- 3 **Relaciones:** si dos variables tienen que ver entre sí.
- 4 **Anomalías residuales:** lo raro que el diagnóstico no cazó.

Lo que la EDA no es:

- La respuesta a la pregunta de negocio.
- Los gráficos “bonitos” de la presentación final.
- Un catálogo: una lámina por cada columna de la base.

La pregunta es el filtro

8.000 filas $\times$ 29 columnas admiten cientos de gráficos correctos e inútiles. Solo interesan los que **acercan a la decisión**: a quién enviar la campaña.

Lo que produce un buen primer EDA

Se entra con una pregunta y unos datos. Se sale con cuatro cosas — ninguna es “la respuesta”:

- 1 **Hallazgos con número.** Lo que ya se sabe del negocio, en lenguaje de negocio: “los clientes sin sesiones en 30 días abandonan el 56 %”.
- 2 **Hipótesis para el modelo.** Qué señales prometen, cuáles no, y qué habría que confirmar.
- 3 **Decisiones documentadas** (el acta): qué se hizo con los faltantes y los atípicos, qué columnas se derivaron, qué se excluyó y por qué.
- 4 **La base analítica** guardada: una fila por unidad de análisis, con las columnas que la pregunta necesita. Lo que alimenta todo lo que viene.

Lo que produce un buen primer EDA

Se entra con una pregunta y unos datos. Se sale con cuatro cosas — ninguna es “la respuesta”:

- 1 **Hallazgos con número.** Lo que ya se sabe del negocio, en lenguaje de negocio: “los clientes sin sesiones en 30 días abandonan el 56 %”.
- 2 **Hipótesis para el modelo.** Qué señales prometen, cuáles no, y qué habría que confirmar.
- 3 **Decisiones documentadas** (el acta): qué se hizo con los faltantes y los atípicos, qué columnas se derivaron, qué se excluyó y por qué.
- 4 **La base analítica** guardada: una fila por unidad de análisis, con las columnas que la pregunta necesita. Lo que alimenta todo lo que viene.

Y una lista corta de lo que *no* se puede usar

Toda variable que solo se conoce *después* del evento que se quiere predecir. En Córdor hay varias — las veremos en el paso 9.

Roadmap

- 1 La pregunta que manda
- 2 Qué es (y qué no es) un EDA
- 3 El EDA de Cóndor, paso a paso**
- 4 Lo que sale del EDA
- 5 Su turno: la base analítica de su pregunta

Paso 1 · ¿Qué es una fila y cuántas hay?

Antes de cualquier cálculo, la regla cero del podcast — para las cuatro tablas del caso:

Tabla	Una fila es...	Filas	Trae
base_clientes	un cliente	8.000	demografía, RFM, app, crédito
anexo_transacciones	una transacción	185.582	monto, tipo, canal, fraude
anexo_creditos	una solicitud	4.071	score, plazo, mora
anexo_campanas	un envío	13.467	oferta, costo, conversión

- base_clientes: 8.000 filas, **8.000** cliente_id **distintos**, **0 duplicadas** (unicidad: aprobada). 29 columnas. Corte: 31-mar-2026.
- **No creer: verificar.** El RFM de la base debería salir del anexo de transacciones. Se recalcula: frecuencia_tx y monto_total coinciden con el anexo en el **100 %** de los clientes. Consistencia: aprobada.

Paso 1 · ¿Qué es una fila y cuántas hay?

Antes de cualquier cálculo, la regla cero del podcast — para las cuatro tablas del caso:

Tabla	Una fila es...	Filas	Trae
base_clientes	un cliente	8.000	demografía, RFM, app, crédito
anexo_transacciones	una transacción	185.582	monto, tipo, canal, fraude
anexo_creditos	una solicitud	4.071	score, plazo, mora
anexo_campanas	un envío	13.467	oferta, costo, conversión

- base_clientes: 8.000 filas, **8.000** cliente_id **distintos**, **0 duplicadas** (unicidad: aprobada). 29 columnas. Corte: 31-mar-2026.
- **No creer: verificar.** El RFM de la base debería salir del anexo de transacciones. Se recalcula: frecuencia_tx y monto_total coinciden con el anexo en el **100 %** de los clientes. Consistencia: aprobada.

La decisión del paso

La unidad de la pregunta es **el cliente** $\Rightarrow$ la base analítica tendrá 8.000 filas. Lo que venga de un anexo se **agrega** al nivel de cliente *antes* de unirlo.

Paso 2 · ¿Qué falta y por qué falta?

Tres columnas traen vacíos — y las tres tienen explicación exacta:

Columna	Faltantes	¿A quién le falta? (coincidencia exacta)
duracion_sesion_promedio	855	a los 855 clientes sin sesiones en 30 días
csat_promedio	5.279	a los 5.279 que nunca abrieron un ticket
score_buro	5.279	a los 5.279 que nunca pidieron crédito

- No son huecos: son **“no aplica”**. La pregunta del podcast — *¿por qué falta?* — aquí la responde el diccionario de datos, y la respuesta es determinística.
- Imputar sería inventar: una mediana de satisfacción para quien nunca llamó a soporte es una **opinión que nadie dio**.

Paso 2 · ¿Qué falta y por qué falta?

Tres columnas traen vacíos — y las tres tienen explicación exacta:

Columna	Faltantes	¿A quién le falta? (coincidencia exacta)
duracion_sesion_promedio	855	a los 855 clientes sin sesiones en 30 días
csat_promedio	5.279	a los 5.279 que nunca abrieron un ticket
score_buro	5.279	a los 5.279 que nunca pidieron crédito

- No son huecos: son **“no aplica”**. La pregunta del podcast — *¿por qué falta?* — aquí la responde el diccionario de datos, y la respuesta es determinística.
- Imputar sería inventar: una mediana de satisfacción para quien nunca llamó a soporte es una **opinión que nadie dio**.

La decisión del paso (va al acta)

No se imputa. Se **marca**: `sin_sesiones`, `abrio_ticket` (y `tiene_credito` ya existe). El hecho de no tener sesiones es información para esta pregunta — lo veremos en el paso 5.

Paso 3 · ¿Cómo se ve el target?



- **1.772 de 8.000 = 22,15 %.**
Esa es la **tasa base**: el punto de comparación de todo lo que sigue.
- Desbalance moderado: la métrica del modelo *no* podrá ser la exactitud a secas (semana 10).
- **¿Qué es “activo”?** 886 clientes llevan **más de 90 días** sin transar; el 59,7 % de ellos ya está marcado como abandono. ¿Son “clientes activos hoy”?

Paso 3 · ¿Cómo se ve el target?

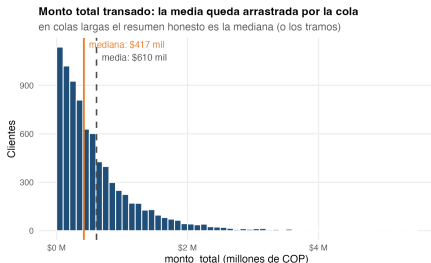
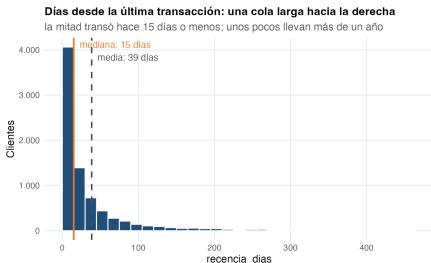


- **1.772 de 8.000 = 22,15 %.**
Esa es la **tasa base**: el punto de comparación de todo lo que sigue.
- Desbalance moderado: la métrica del modelo *no* podrá ser la exactitud a secas (semana 10).
- **¿Qué es “activo”?** 886 clientes llevan **más de 90 días** sin transar; el 59,7 % de ellos ya está marcado como abandono. ¿Son “clientes activos hoy”?

La decisión del paso

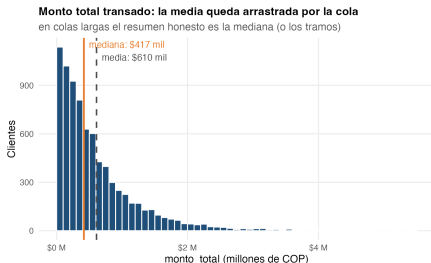
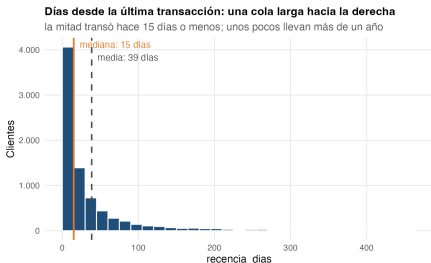
La definición de “activo” se acuerda **con Growth antes de modelar** (¿recencia $\leq$ 90 días?). El abismo semántico del podcast también existe *dentro* de la empresa.

Paso 4 · ¿Cómo se distribuyen las candidatas?



- Las dos tienen **cola larga a la derecha** y la media queda arrastrada: recencia, mediana **15** días vs. media **39**; monto total, mediana **\$417 mil** vs. media **\$610 mil**. “El cliente promedio transó hace 39 días” describe a muy pocos.

Paso 4 · ¿Cómo se distribuyen las candidatas?



- Las dos tienen **cola larga a la derecha** y la media queda arrastrada: recencia, mediana **15** días vs. media **39**; monto total, mediana **\$417 mil** vs. media **\$610 mil**. “El cliente promedio transó hace 39 días” describe a muy pocos.

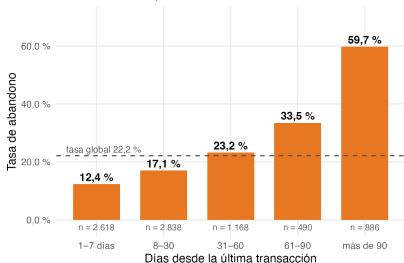
La decisión del paso

Describir con **medianas y tramos**, no con promedios. Las comparaciones del paso 5 se hacen por tramos.

Paso 5 · ¿Quiénes abandonan más?

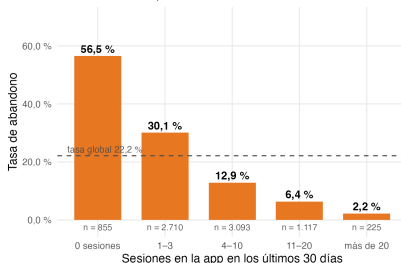
A más días sin transar, más abandono

tasa de abandono por tramo de recencia_dias



Quien no abre la app ya casi se fue

tasa de abandono por tramo de num_sesiones_ult30d

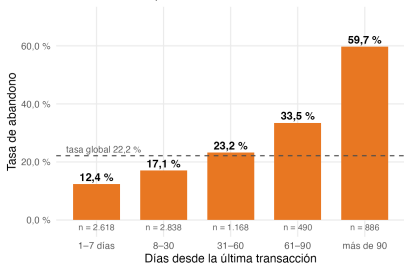


- Recencia: del **12,4 %** (transó esta semana) al **59,7 %** (más de 90 días).
Sesiones: del **56,5 %** (cero) al **2,2 %** (más de 20). La línea punteada es la tasa global: importa **cuánto se aleja** cada tramo.

Paso 5 · ¿Quiénes abandonan más?

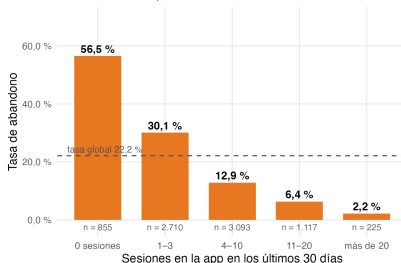
A más días sin transar, más abandono

tasa de abandono por tramo de recencia_dias



Quien no abre la app ya casi se fue

tasa de abandono por tramo de num_sesiones_ult30d



- Recencia: del **12,4 %** (transó esta semana) al **59,7 %** (más de 90 días). Sesiones: del **56,5 %** (cero) al **2,2 %** (más de 20). La línea punteada es la tasa global: importa **cuánto se aleja** cada tramo.

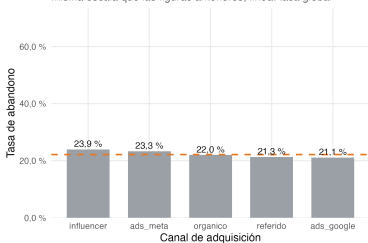
Hallazgo 1

El **comportamiento reciente** es la señal. Y el “faltante” del paso 2 — cero sesiones — resultó ser el grupo de mayor riesgo de toda la base.

Paso 6 · ¿Y lo que “debería” importar?

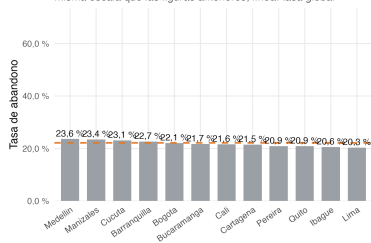
El canal por el que llegaron no cambia casi nada

misma escala que las figuras anteriores; línea: tasa global



Tampoco la ciudad: todas entre 20 % y 24 %

misma escala que las figuras anteriores; línea: tasa global

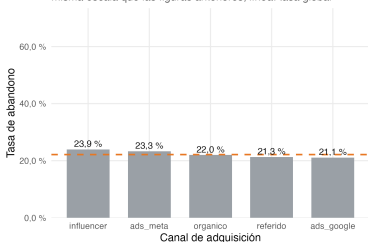


- Misma escala que la lámina anterior, a propósito. Canal: **21,1 a 23,9 %**; ciudad: **20,3 a 23,6 %**; género, sistema operativo, KYC y ocupación: todo entre 20 y 25 %. “Los de *influencer* se van más” es ruido alrededor de la tasa global.

Paso 6 · ¿Y lo que “debería” importar?

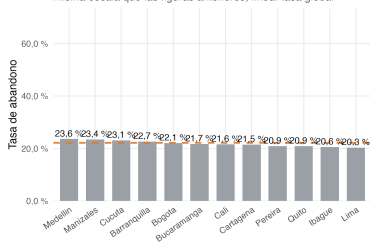
El canal por el que llegaron no cambia casi nada

misma escala que las figuras anteriores; línea: tasa global



Tampoco la ciudad: todas entre 20 % y 24 %

misma escala que las figuras anteriores; línea: tasa global



- Misma escala que la lámina anterior, a propósito. Canal: **21,1 a 23,9 %**; ciudad: **20,3 a 23,6 %**; género, sistema operativo, KYC y ocupación: todo entre 20 y 25 %. “Los de *influencer* se van más” es ruido alrededor de la tasa global.

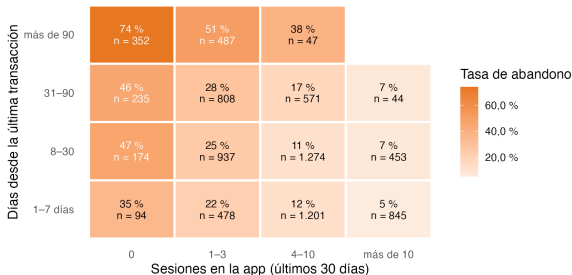
Hallazgo 2 — un gráfico plano también es un hallazgo

La demografía y el canal **no discriminan**: la campaña se segmenta por comportamiento, no por ciudad ni por canal.

Paso 7 · ¿Las dos señales miden lo mismo?

Recencia y sesiones no miden lo mismo: se refuerzan

cero sesiones dispara el abandono aun con transacción reciente;
sin transar ni abrir la app, 3 de cada 4 se van

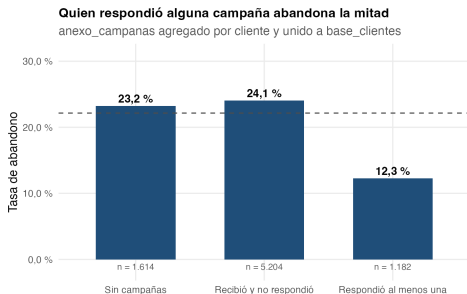


- Correlación recencia–sesiones: **−0,35**. Se parecen, pero no son lo mismo: con transacción de esta semana y **cero sesiones**, el abandono ya es del 35%; sin transar en 90 días y sin sesiones, **74 %**; con más de 10 sesiones nunca pasa del 7 %.

La decisión del paso

Las dos señales entran a la base analítica; la campaña ya sabe a quién ir primero (arriba a la izquierda).

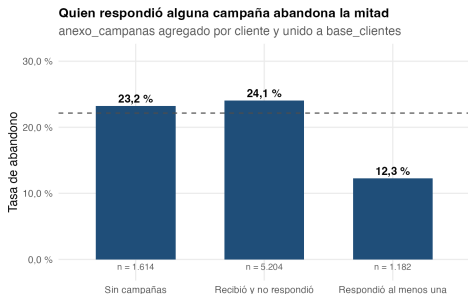
Paso 8 · ¿Qué aporta un anexo? Agregar y luego unir



El grano, en vivo:

- anexo_campanas: 13.467 envíos a 6.386 clientes.
- Se **agrega** por cliente (n_campanas, n_convertidas) y *después* se une a la base: siguen siendo **8.000 filas**.
- Unido **sin agregar**: **15.081 filas**. Clientes duplicados, totales inflados, y R no dice nada.
- 1.614 clientes nunca recibieron campaña: tras el join quedan en NA → se vuelven 0 y se documenta.

Paso 8 · ¿Qué aporta un anexo? Agregar y luego unir



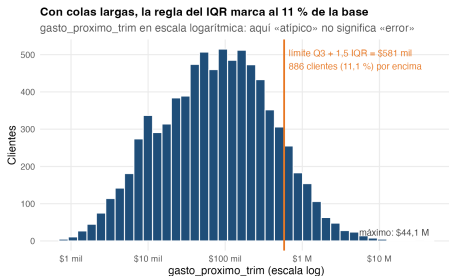
El grano, en vivo:

- `anexo_campanas`: 13.467 envíos a 6.386 clientes.
- Se **agrega** por cliente (`n_campanas`, `n_convertidas`) y *después* se une a la base: siguen siendo **8.000** filas.
- Unido **sin agregar**: **15.081** filas. Clientes duplicados, totales inflados, y R no dice nada.
- 1.614 clientes nunca recibieron campaña: tras el join quedan en NA → se vuelven 0 y se documenta.

Hallazgo 3 — con letra pequeña

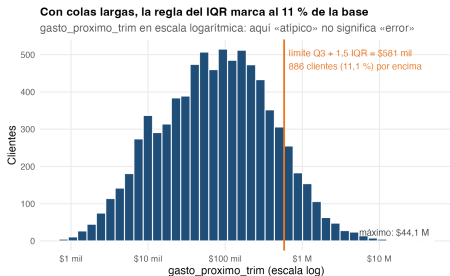
Quien respondió alguna campaña abandona el **12,3 %** frente al **24,1 %**. Pero es **correlación**: ¿la campaña retiene, o los clientes fieles responden más? Es una hipótesis para el modelo, no una conclusión para Growth.

Paso 9 · ¿Qué raro quedó — y qué *no* puedo usar?



- La regla del IQR marca **886 clientes (11 %)**. En una cola larga, “atípico” no significa “error”: significa que la regla no aplica.
- El máximo, **\$44,1 millones**: se mira la fila completa. Cliente de **1 mes**, **72 transacciones**, transó ayer, 15 sesiones. No es un dedo de más: es un cliente que despegó. **Se conserva y se reporta.**

Paso 9 · ¿Qué raro quedó — y qué *no* puedo usar?



- La regla del IQR marca **886 clientes (11 %)**. En una cola larga, “atípico” no significa “error”: significa que la regla no aplica.
- El máximo, **\$44,1 millones**: se mira la fila completa. Cliente de **1 mes**, **72 transacciones**, transó ayer, 15 sesiones. No es un dedo de más: es un cliente que despegó. **Se conserva y se reporta.**

La fuga de información (la regla de oro del caso)

gasto_proximo_trim “predice” el abandono de maravilla: mediana de \$105 mil en quienes se quedan y \$23 mil en quienes se van. Claro: **es el futuro**, se conoce después del corte igual que abandono. Nunca entra como predictor. Cuando una variable predice demasiado bien, sospechen de esto.

Roadmap

- 1 La pregunta que manda
- 2 Qué es (y qué no es) un EDA
- 3 El EDA de Cóndor, paso a paso
- 4 Lo que sale del EDA**
- 5 Su turno: la base analítica de su pregunta

El EDA de hoy, en una lámina

Hallazgos (con número)

- 1 22 % de los clientes se vuelve inactivo; el comportamiento reciente separa tasas del 5 % al 74 %.
- 2 Demografía, ciudad y canal no discriminan: todo entre 20 y 25 %.
- 3 Quien respondió alguna campaña abandona la mitad (12 vs. 24 %).

Hipótesis para el modelo

- Priorizar la reactivación en **cero sesiones y más de 30 días sin transar**: 587 clientes con tasas del 46 al 74 %.
- La respuesta a campañas puede ser causa o síntoma: señal, no palanca.

Decisiones (el acta)

- Faltantes = “no aplica” → indicadores; nada se imputa.
- Definir “activo” con Growth (¿recencia $\leq$ 90 días?) antes de modelar.
- Fuera de la base analítica: `gasto_proximo_trim` (es del futuro).
- Anexos: agregar por cliente antes de unir; verificar que sigan 8.000 filas.
- El máximo de \$44,1 M se conserva y se reporta.

La base analítica

8.000 filas $\times$ lo que la pregunta necesita: target, recencia, sesiones, RFM, indicadores derivados y campañas agregadas.

Qué debe mostrar su primer EDA

La misma lista, para **su** pregunta. Es el esqueleto de la Entrega 2:

- 1 **El grano y el tamaño:** qué es una fila, cuántas hay, qué tablas se usan y cómo se unen.
- 2 **El target y su tasa base** (o su distribución, si predicen un número; o las variables que agrupan, si segmentan).
- 3 **Los faltantes explicados:** por qué faltan y qué decidieron.
- 4 **Dos o tres comparaciones por grupo** ligadas a la pregunta — incluida una que *no* funcione.
- 5 **Una relación** entre dos señales: ¿miden lo mismo o se complementan?
- 6 **Lo raro** y qué decidieron hacer con ello.
- 7 **Lo que no van a usar:** las fugas de información.
- 8 **El acta:** cinco líneas que otra persona pueda leer y reproducir.

Qué debe mostrar su primer EDA

La misma lista, para **su** pregunta. Es el esqueleto de la Entrega 2:

- 1 **El grano y el tamaño:** qué es una fila, cuántas hay, qué tablas se usan y cómo se unen.
- 2 **El target y su tasa base** (o su distribución, si predicen un número; o las variables que agrupan, si segmentan).
- 3 **Los faltantes explicados:** por qué faltan y qué decidieron.
- 4 **Dos o tres comparaciones por grupo** ligadas a la pregunta — incluida una que *no* funcione.
- 5 **Una relación** entre dos señales: ¿miden lo mismo o se complementan?
- 6 **Lo raro** y qué decidieron hacer con ello.
- 7 **Lo que no van a usar:** las fugas de información.
- 8 **El acta:** cinco líneas que otra persona pueda leer y reproducir.

Ocho cosas, no ochenta gráficos

Si una lámina no acerca a la decisión de su pregunta, no va. Esa es toda la diferencia entre un EDA y un catálogo.

Roadmap

- 1 La pregunta que manda
- 2 Qué es (y qué no es) un EDA
- 3 El EDA de Cóndor, paso a paso
- 4 Lo que sale del EDA
- 5 Su turno: la base analítica de su pregunta

La práctica de hoy (con el monitor)

- ➊ **Cargar** las cuatro tablas desde la URL del proyecto.
- ➋ **Decidir el grano** de su pregunta: ¿una fila es un cliente, una transacción, un envío?
- ➌ Con un anexo: `group_by + summarise`, luego `left_join`; **verificar** que las filas no cambien.
- ➍ **Primer EDA**: `skim`, el target, dos comparaciones por grupo, una relación.
- ➎ **Acta y guardar** la base analítica.

Los datos se leen desde la web (URL de los lineamientos)

```
url <-  
"https://eduard-martinez.github.io/  
teaching/ba/final_project/data/"  
clientes <- read.csv(paste0(url,  
  "base_clientes.csv"))  
(igual para los tres anexos)
```

La programación es básica a propósito

Cargar, unir, resumir, graficar: lo de la semana 4. Lo difícil es decidir **qué mirar** y **qué concluir**.

Para cerrar

- ❶ **La pregunta manda.** Eligió las nueve cosas que miramos y descartó las otras cien. Sin pregunta, el EDA es un catálogo.
- ❷ **El EDA no responde: prepara.** Salimos con tres hallazgos, dos hipótesis, un acta y una base analítica — y sin ningún modelo.
- ❸ **Un gráfico plano también es un hallazgo.** Saber que la ciudad no importa vale tanto como saber que las sesiones sí.
- ❹ **Lo que se conoce después del corte no predice nada.** Cuando una variable acierta demasiado, sospechen.
- ❺ **Verificar, no creer.** El RFM coincidía con el anexo; el join conservó las 8.000 filas. Cada paso trae su comprobación.

Para cerrar

- ❶ **La pregunta manda.** Eligió las nueve cosas que miramos y descartó las otras cien. Sin pregunta, el EDA es un catálogo.
- ❷ **El EDA no responde: prepara.** Salimos con tres hallazgos, dos hipótesis, un acta y una base analítica — y sin ningún modelo.
- ❸ **Un gráfico plano también es un hallazgo.** Saber que la ciudad no importa vale tanto como saber que las sesiones sí.
- ❹ **Lo que se conoce después del corte no predice nada.** Cuando una variable acierta demasiado, sospechen.
- ❺ **Verificar, no creer.** El RFM coincidía con el anexo; el join conservó las 8.000 filas. Cada paso trae su comprobación.

Entrega 2

La base depurada y el EDA de su pregunta: exactamente lo que hicimos hoy, con **su** pregunta y **sus** decisiones. Arranca ahora, con el monitor.