

Data mining & machine learning - S10

# PROYECTO #1

## PREDICCIÓN DE RECOMpra Y DEVOLUCIONES EN CLIENTES DE UNA TIENDA ONLINE

Carlos Angel, José Donado, Carlos Aldana,  
Diego Monroy y Marco Carbajal

# Dataset que se nos presentó



Screenshot of Microsoft Excel showing a dataset titled "P5\_Recompra\_Online • Guardando...". The table has 37 rows and 10 columns, representing data from an Online Retailer. The columns are labeled A through J.

	A	B	C	D	E	F	G	H	I	J
1	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country		
2	536365	85123A	WHITE HANGING HE	6	1/12/2010 08:26	2.55	17850	United Kingdom		
3	536365	71053	WHITE METAL LANT	6	1/12/2010 08:26	3.39	17850	United Kingdom		
4	536365	84406B	CREAM CUPID HEAR	8	1/12/2010 08:26	2.75	17850	United Kingdom		
5	536365	84029G	KNITTED UNION FLA	6	1/12/2010 08:26	3.39	17850	United Kingdom		
6	536365	84029E	RED WOOLLY HOTTI	6	1/12/2010 08:26	3.39	17850	United Kingdom		
7	536365	22752	SET 7 BABUSHKA NE	2	1/12/2010 08:26	7.65	17850	United Kingdom		
8	536365	21730	GLASS STAR FROSTE	6	1/12/2010 08:26	4.25	17850	United Kingdom		
9	536366	22633	HAND WARMER UN	6	1/12/2010 08:28	1.85	17850	United Kingdom		
10	536366	22632	HAND WARMER RED	6	1/12/2010 08:28	1.85	17850	United Kingdom		
11	536367	84879	ASSORTED COLOUR	32	1/12/2010 08:34	1.69	13047	United Kingdom		
12	536367	22745	POPPY'S PLAYHOUS	6	1/12/2010 08:34	2.1	13047	United Kingdom		
13	536367	22748	POPPY'S PLAYHOUS	6	1/12/2010 08:34	2.1	13047	United Kingdom		
14	536367	22749	FELTCRAFT PRINCES	8	1/12/2010 08:34	3.75	13047	United Kingdom		
15	536367	22310	IVORY KNITTED MU	6	1/12/2010 08:34	1.65	13047	United Kingdom		
16	536367	84969	BOX OF 6 ASSORTED	6	1/12/2010 08:34	4.25	13047	United Kingdom		
17	536367	22623	BOX OF VINTAGE JIG	3	1/12/2010 08:34	4.95	13047	United Kingdom		
18	536367	22622	BOX OF VINTAGE AI	2	1/12/2010 08:34	9.95	13047	United Kingdom		
19	536367	21754	HOME BUILDING BL	3	1/12/2010 08:34	5.95	13047	United Kingdom		
20	536367	21755	LOVE BUILDING BLC	3	1/12/2010 08:34	5.95	13047	United Kingdom		
21	536367	21777	RECIPE BOX WITH M	4	1/12/2010 08:34	7.95	13047	United Kingdom		
22	536367	48187	DOORMAT NEW EN	4	1/12/2010 08:34	7.95	13047	United Kingdom		
23	536368	22960	JAM MAKING SET W	6	1/12/2010 08:34	4.25	13047	United Kingdom		
24	536368	22913	RED COAT RACK PA	3	1/12/2010 08:34	4.95	13047	United Kingdom		
25	536368	22912	YELLOW COAT RACK	3	1/12/2010 08:34	4.95	13047	United Kingdom		
26	536368	22914	BLUE COAT RACK PA	3	1/12/2010 08:34	4.95	13047	United Kingdom		
27	536369	21756	BATH BUILDING BLC	3	1/12/2010 08:35	5.95	13047	United Kingdom		
28	536370	22728	ALARM CLOCK BAKE	24	1/12/2010 08:45	3.75	12583	France		
29	536370	22727	ALARM CLOCK BAKE	24	1/12/2010 08:45	3.75	12583	France		
30	536370	22726	ALARM CLOCK BAKE	12	1/12/2010 08:45	3.75	12583	France		
31	536370	21724	PANDA AND BUNNI	12	1/12/2010 08:45	0.85	12583	France		
32	536370	21883	STARS GIFT TAPE	24	1/12/2010 08:45	0.65	12583	France		
33	536370	10002	INFLATABLE POLITIC	48	1/12/2010 08:45	0.85	12583	France		
34	536370	21791	VINTAGE HEADS AN	24	1/12/2010 08:45	1.25	12583	France		
35	536370	21035	SET/2 RED RETROSP	18	1/12/2010 08:45	2.95	12583	France		
36	536370	22326	ROUND SNACK BOX	24	1/12/2010 08:45	2.95	12583	France		
37	536370	22629	SPACEBOY LUNCH B	24	1/12/2010 08:45	1.95	12583	France		

# OBJETIVOS

- 1) Analizar patrones de compra y frecuencia de los clientes.
- 2) Construir modelos de clasificación para predicción de recompra y devoluciones.
- 3) Construir modelos de regresión para predecir el monto total gastado por cliente.
- 4) Calcular el impacto económico de retener a clientes leales versus perderlos.
- 5) Proponer acciones que minimicen devoluciones y mejoren la experiencia del cliente.



# PREPARACIÓN DEL ENTORNO



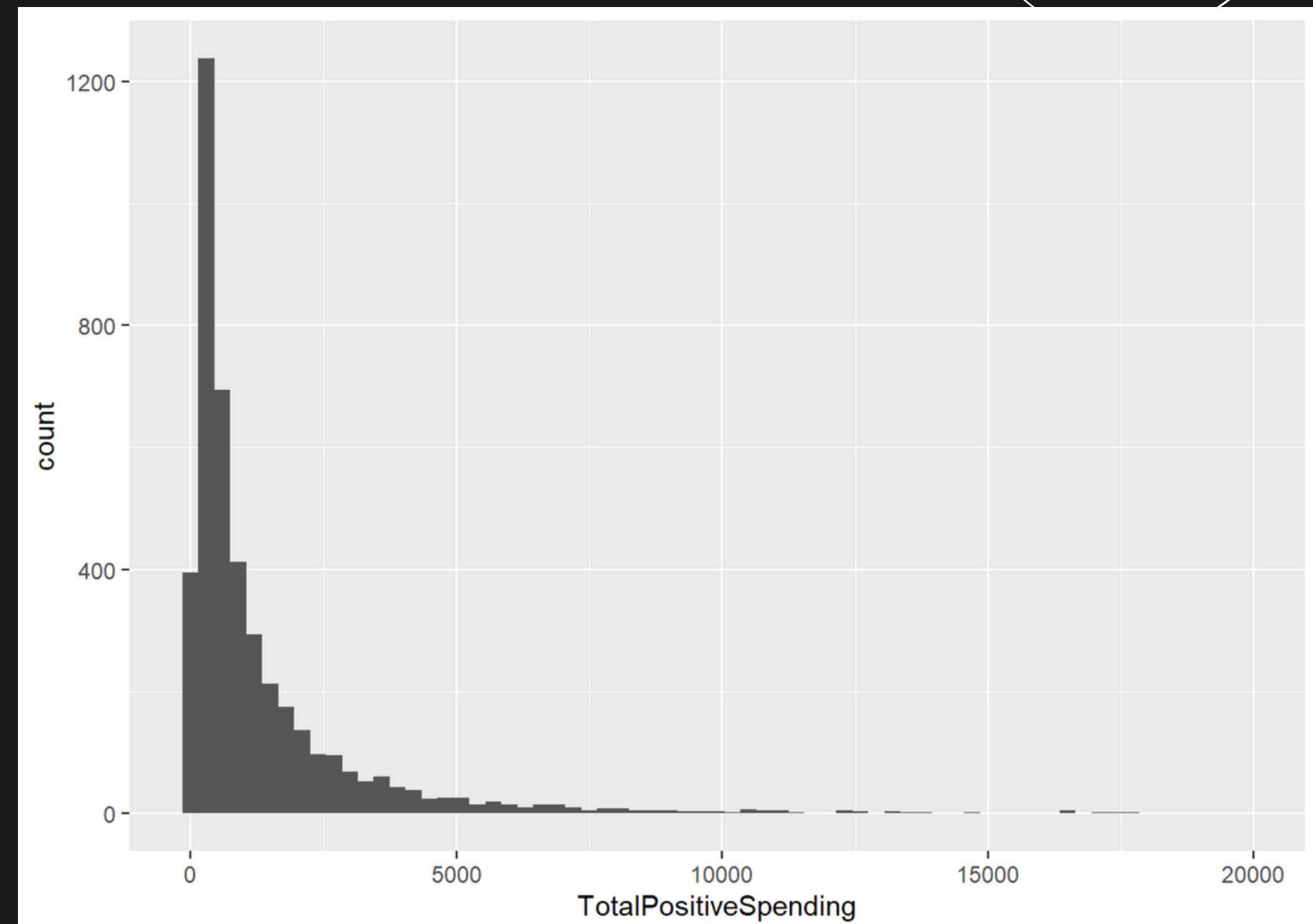
```
library(dplyr)
library(fastDummies)
library(ggplot2)
library(psych)
library(corrplot)
library(tidyr)
library(broom)
library(sigr)
library(ggplot2)
library(WVPlots)
library(Metrics)
library(e1071)
library(weights)
library(corrplot)
library(tidyverse)
library(rpart)
library(rpart.plot)
library(e1071)
library(readxl)
library(randomForest)
library(writexl)
```

# EXPLORACIÓN DE DATOS

A partir de la exploración, podemos concluir que existe una diferencia extremadamente significativa entre los datos atípicos tanto de devoluciones como de compras.

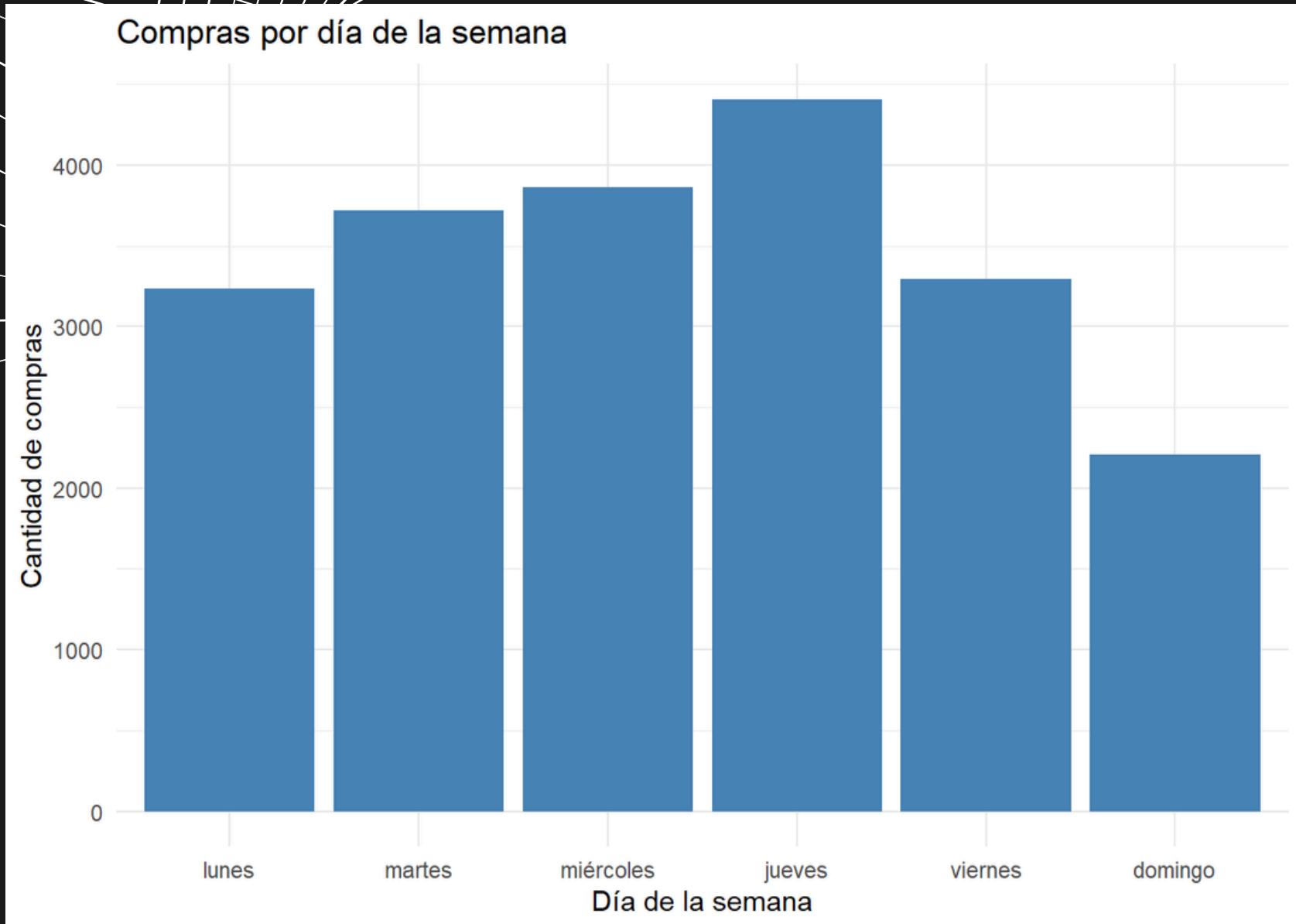
```
# Distribución de TotalPositiveSpending
summary(customer_positive_spending$TotalPositiveSpending)

##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
## 3.75   307.41  674.48  2054.27 1661.74 280206.02
```

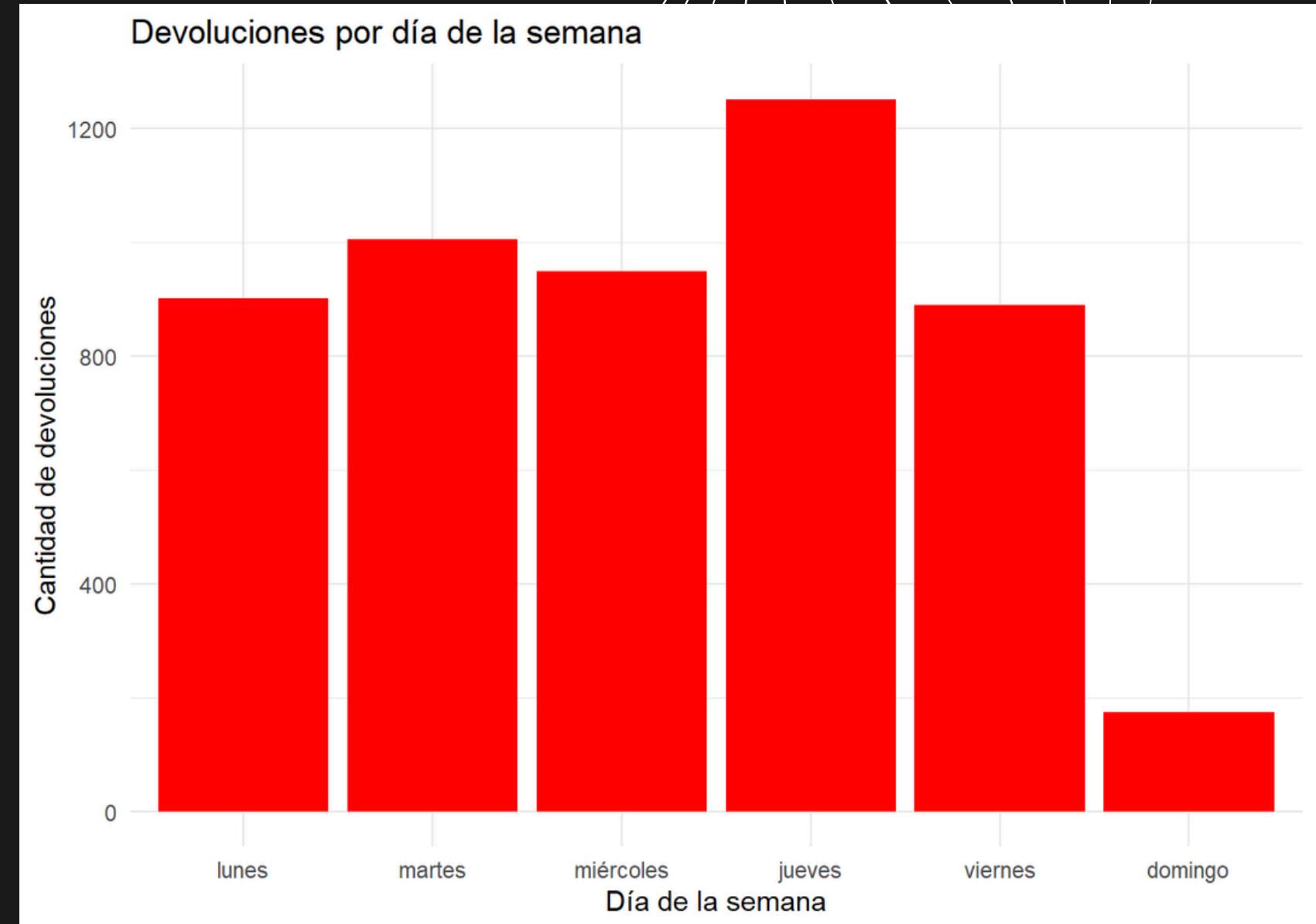


Por otro lado, vemos que no existe una relación entre la región en la que se hace un pedido y la cantidad de devoluciones.

Compras por día de la semana

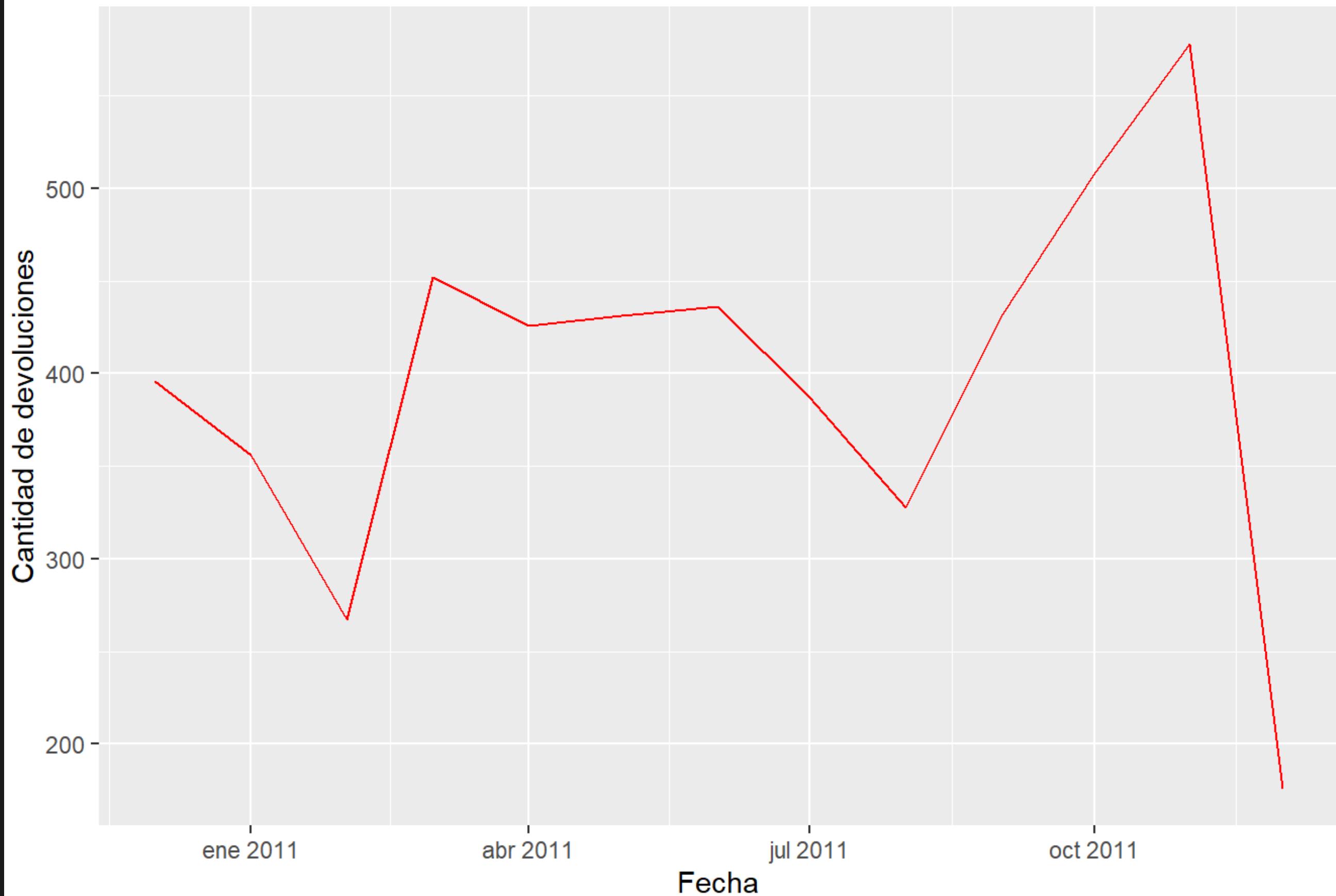


Devoluciones por día de la semana

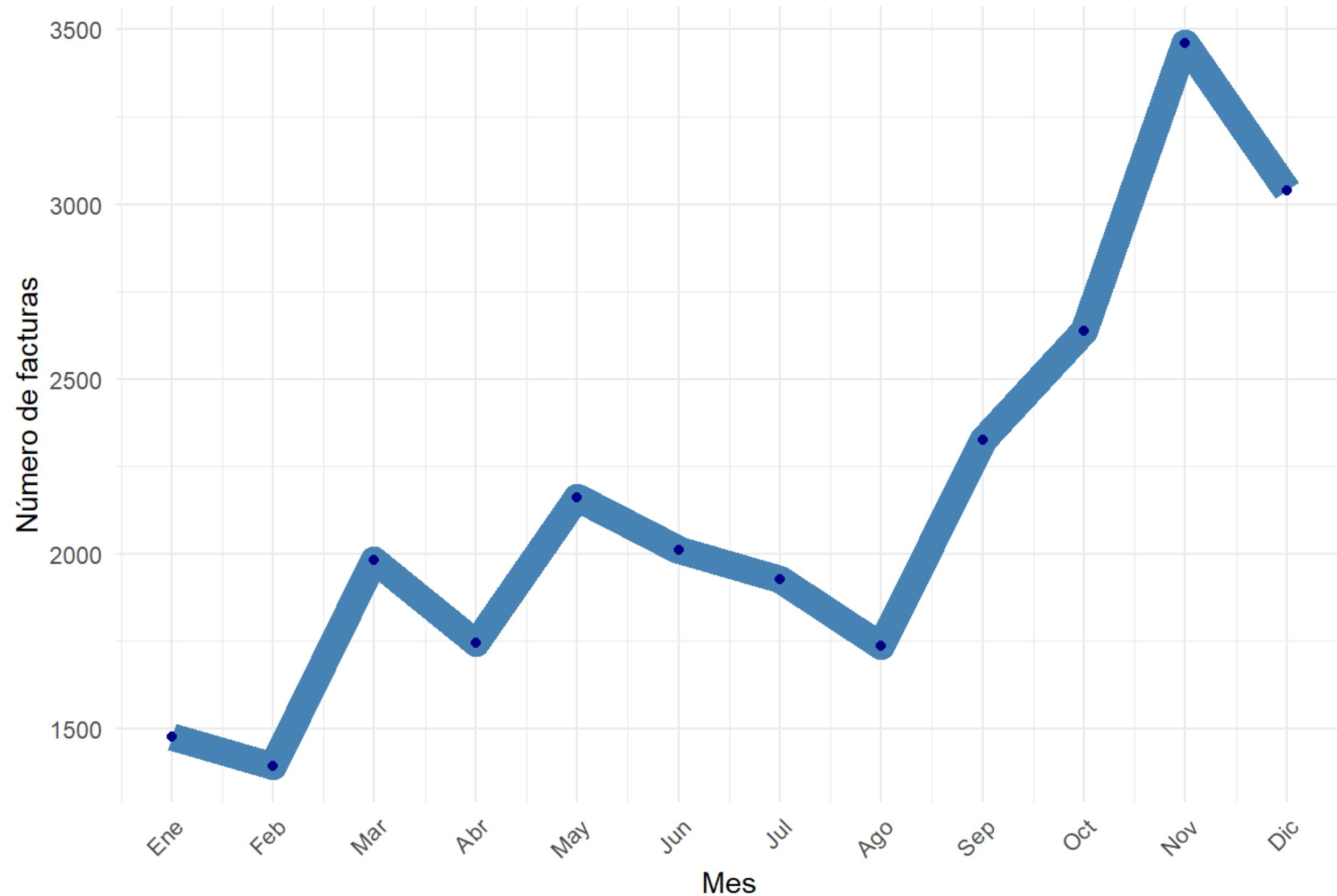


Existe evidencia que defiende un aumento en la cantidad de transacciones los días jueves, además de un aumento general en la cantidad de productos comprados en las compras mas recientes.

## Evolución mensual de devoluciones

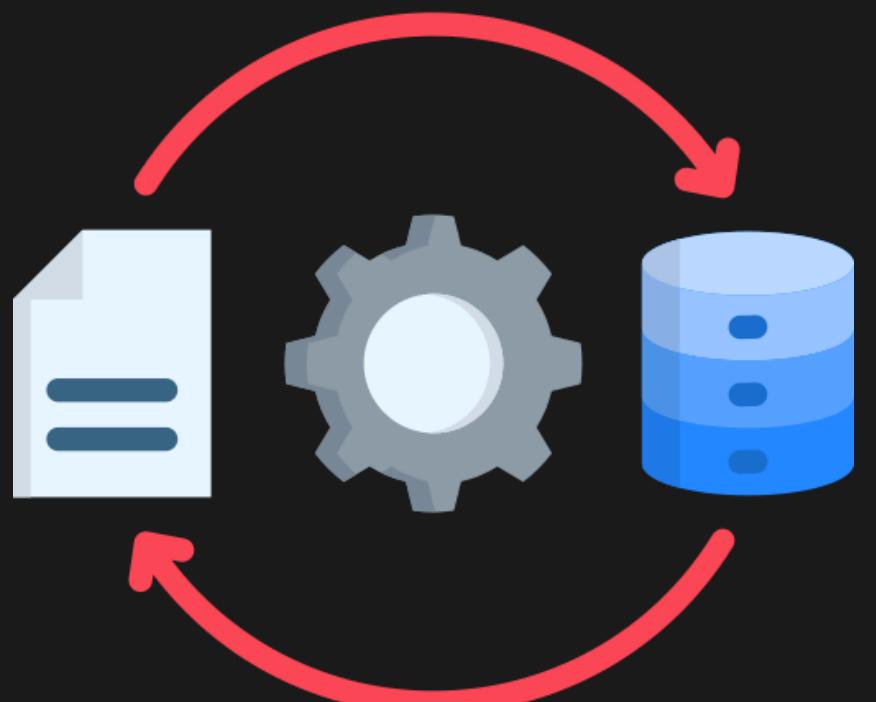


## Evolución del volumen de compras (número de facturas) mensuales



# TRANSFORMACIÓN DE DATOS

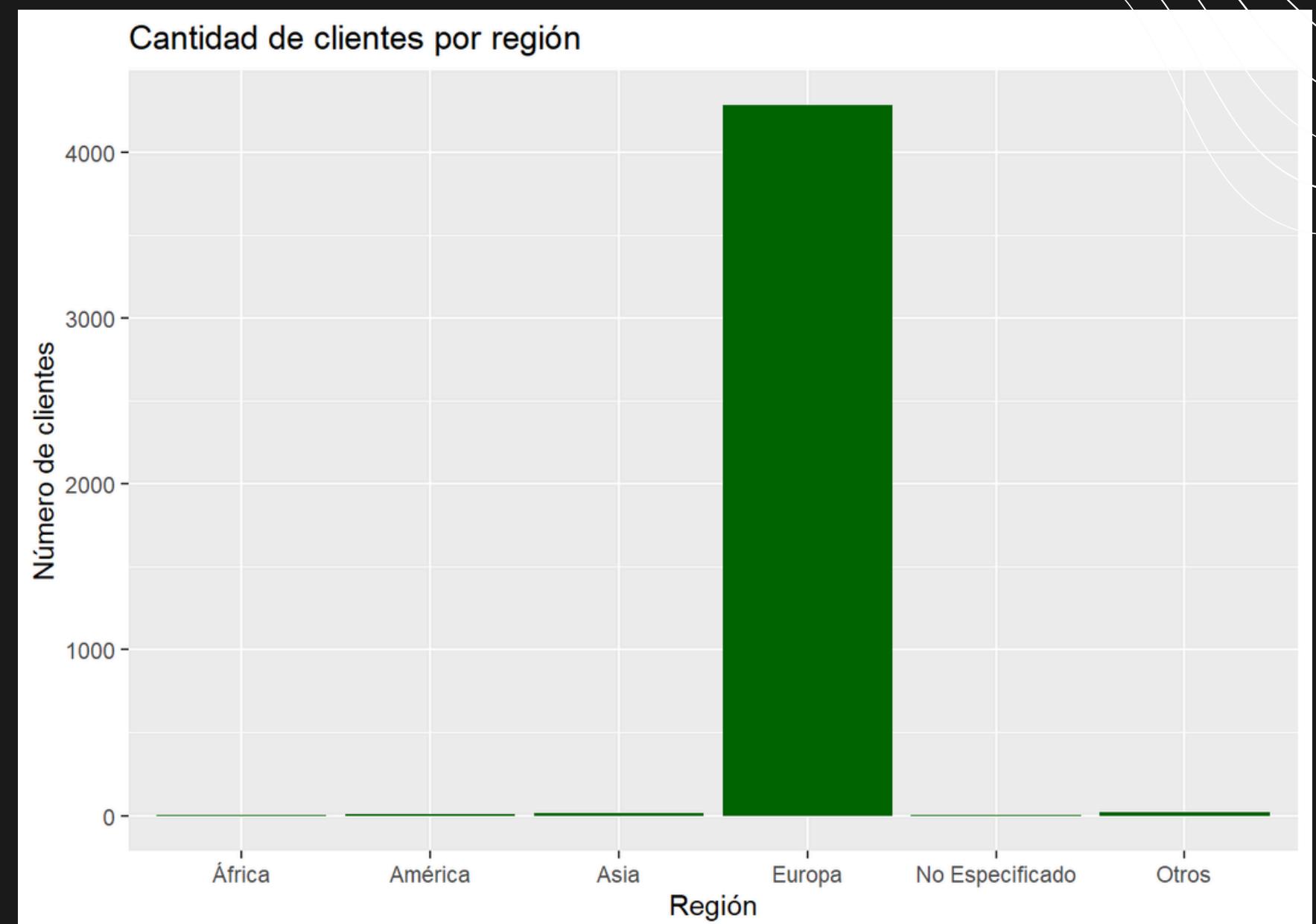
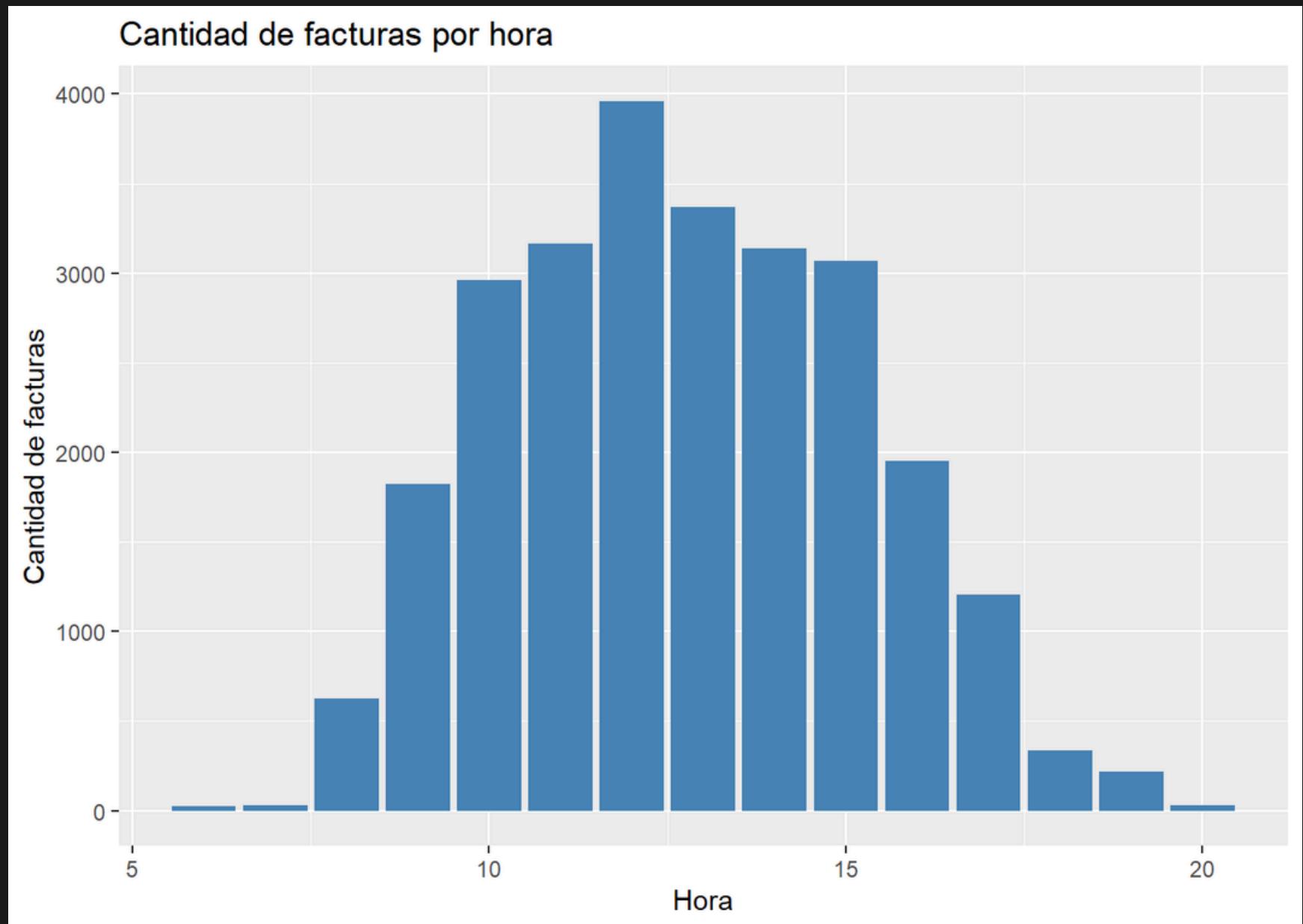
Tras el proceso de transformación de datos, llegamos a la tabla **clientes**, en base a la cuál generamos los modelos.



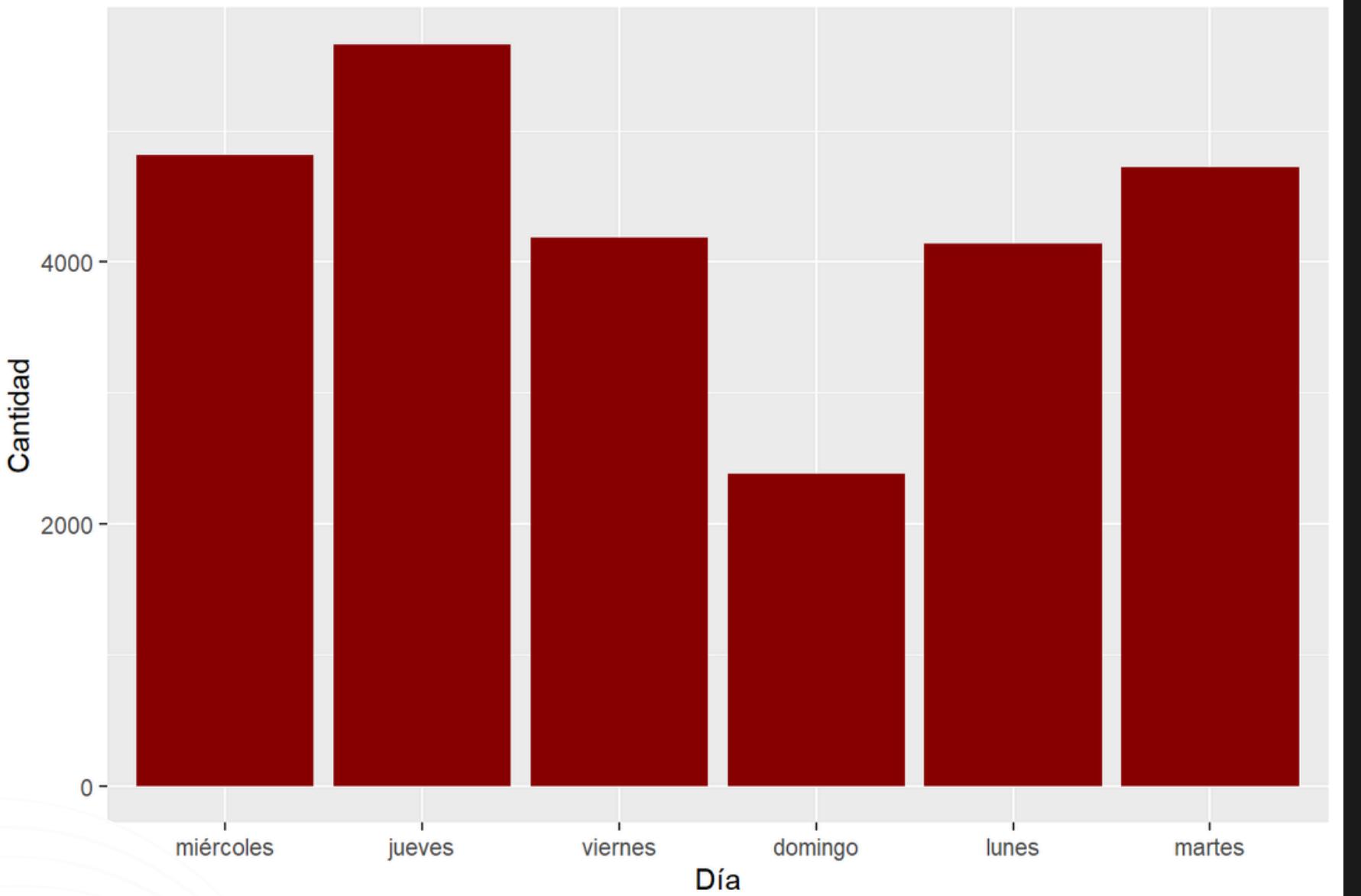
```
clientes <- clientes %>%
  select(
    ultima_compra,
    ultima_devolucion,
    veces_devoluciones,
    hora_max_compras,
    dia_semana_max_compras,
    mes_max_compras,
    total_compras,
    total_devoluciones,
    region,
    cant_facturas_compras,
    cantidad_devuelta,
    cantidad_productos,
    cantidad_total,
    indice_devolucion,
    ha_devuelto,
    ha_recomprado,
    dinero
  )
```

# GRÁFICOS DE EXPLORACIÓN

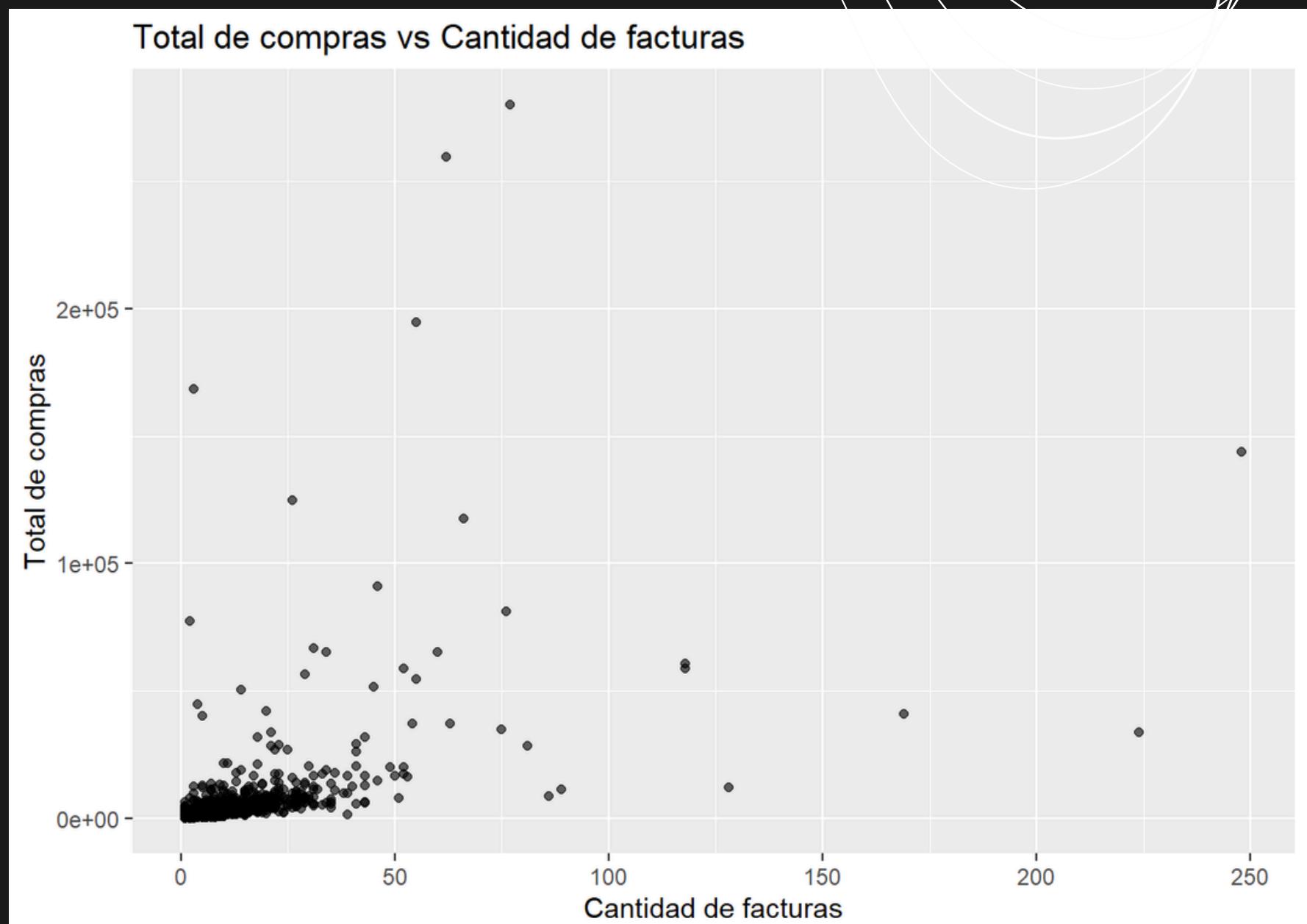
Para comprender a profundidad los datos depurados



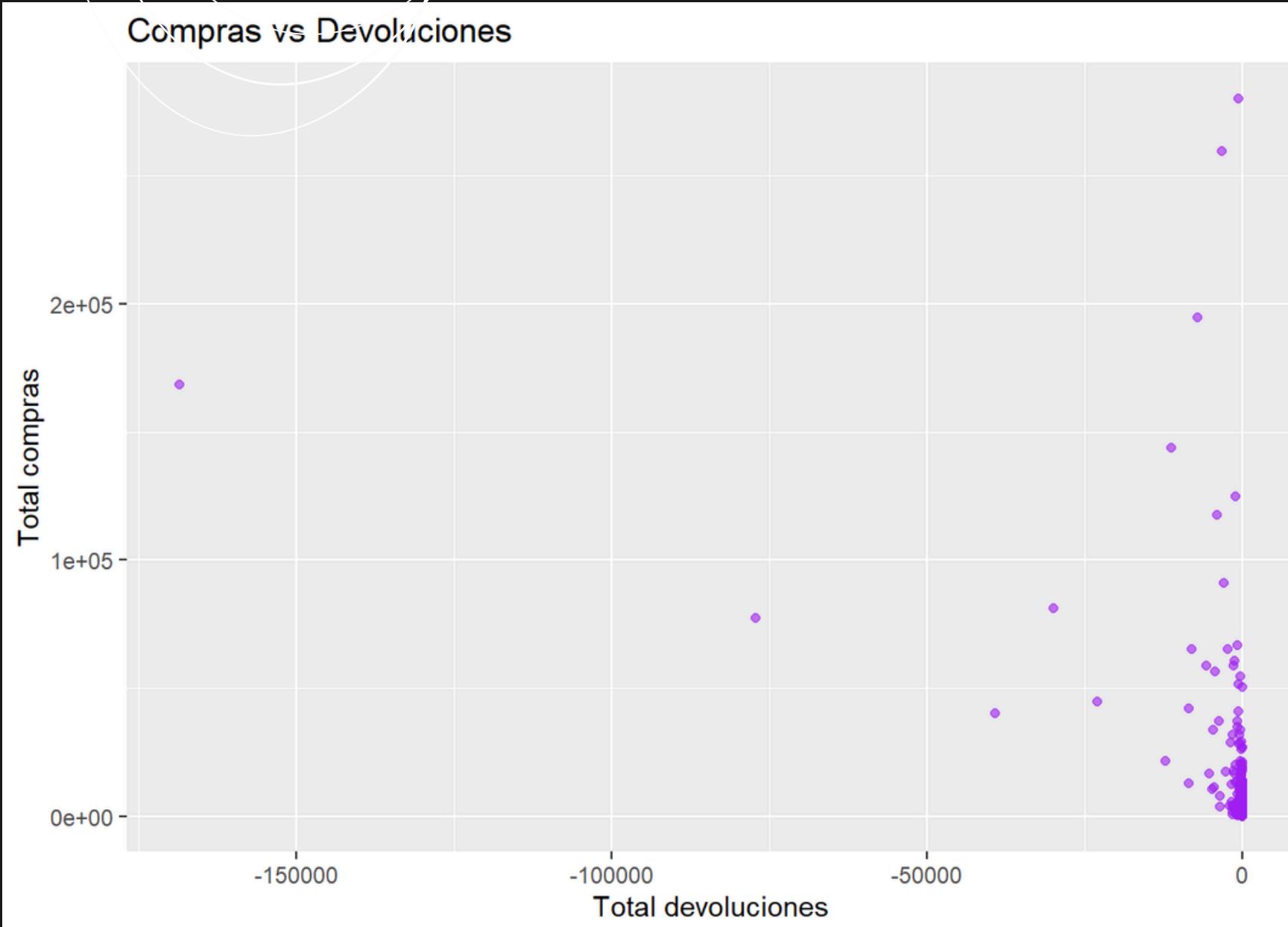
Facturas por día de la semana



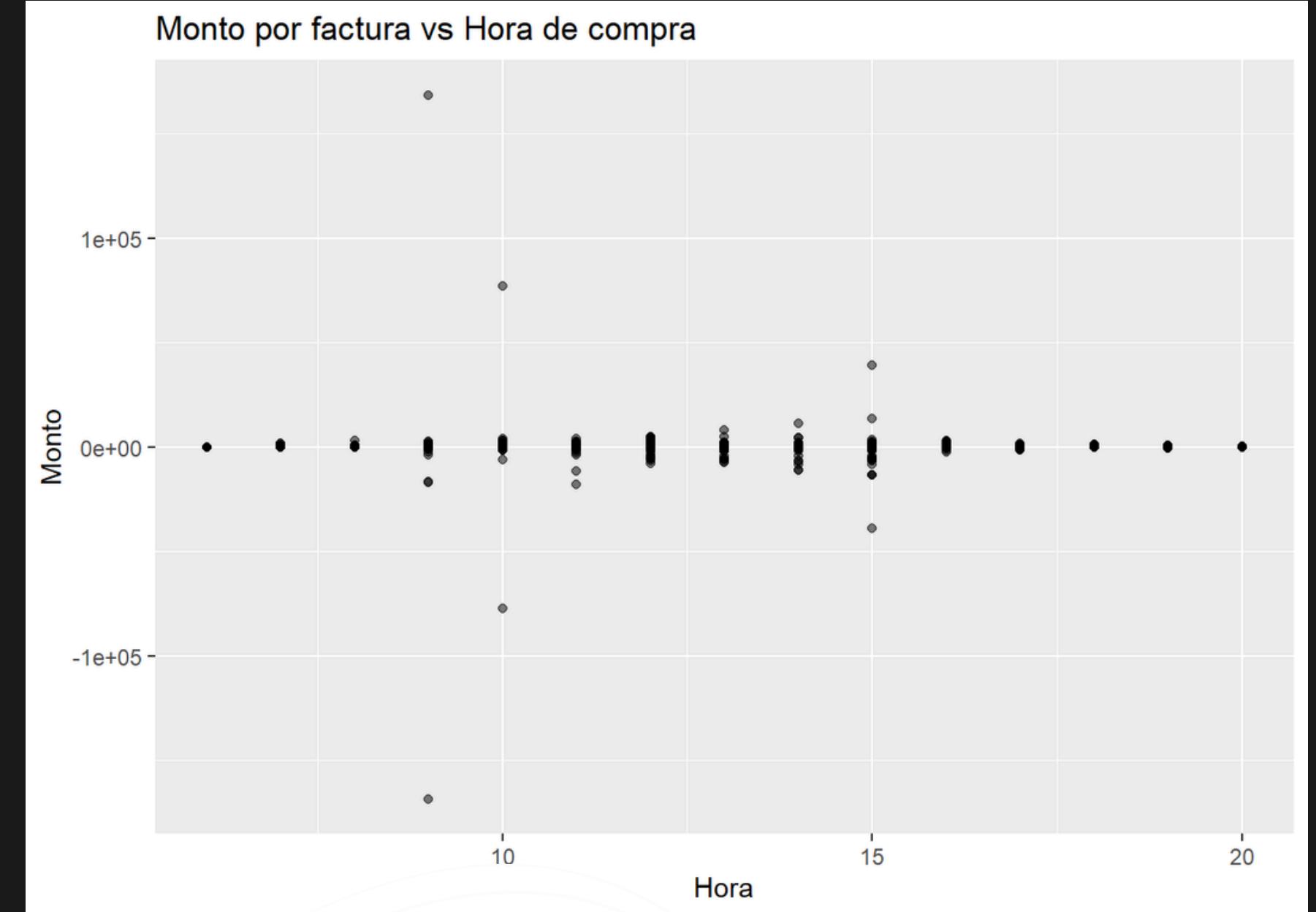
Total de compras vs Cantidad de facturas



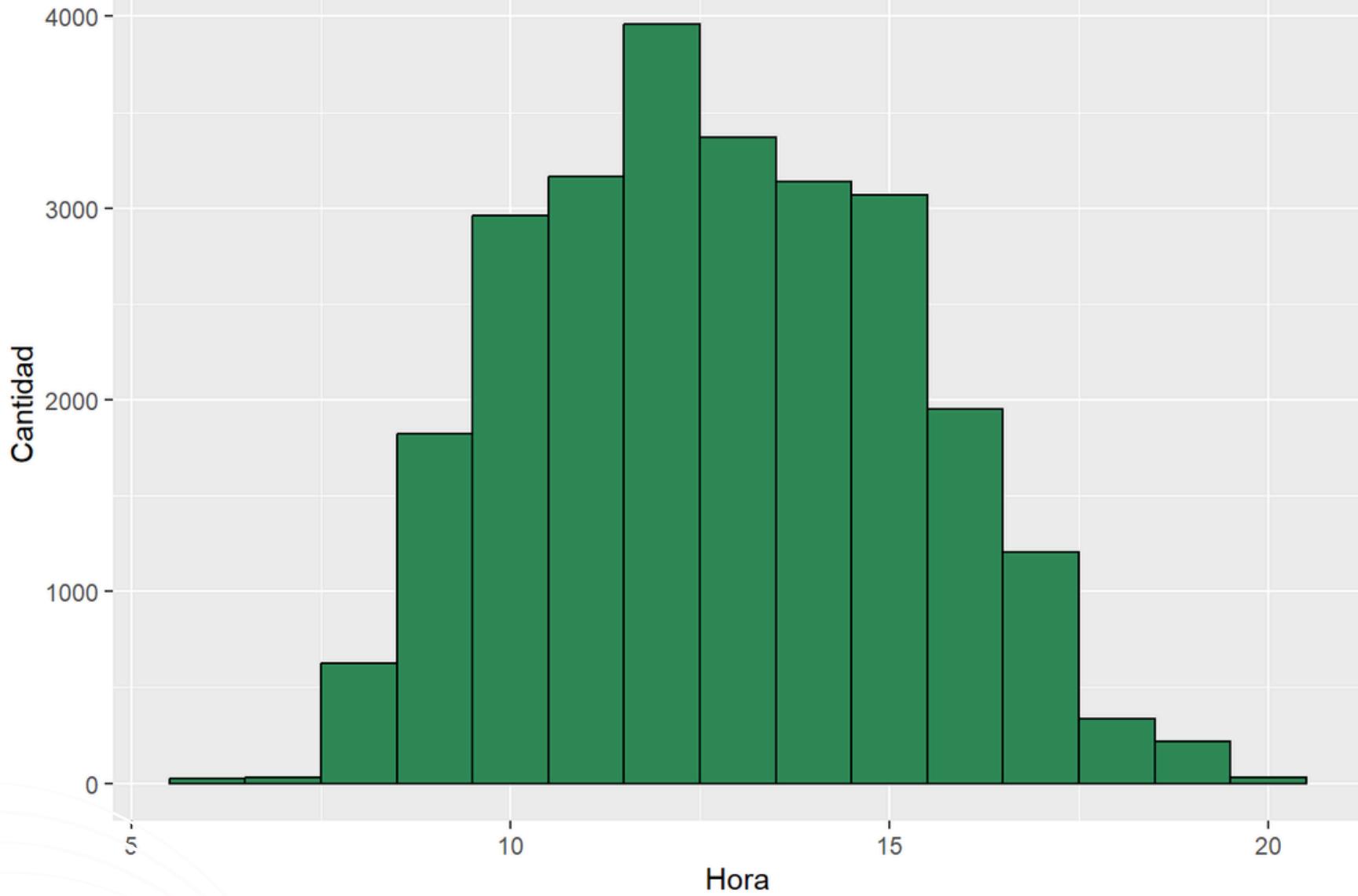
Compras vs Devoluciones



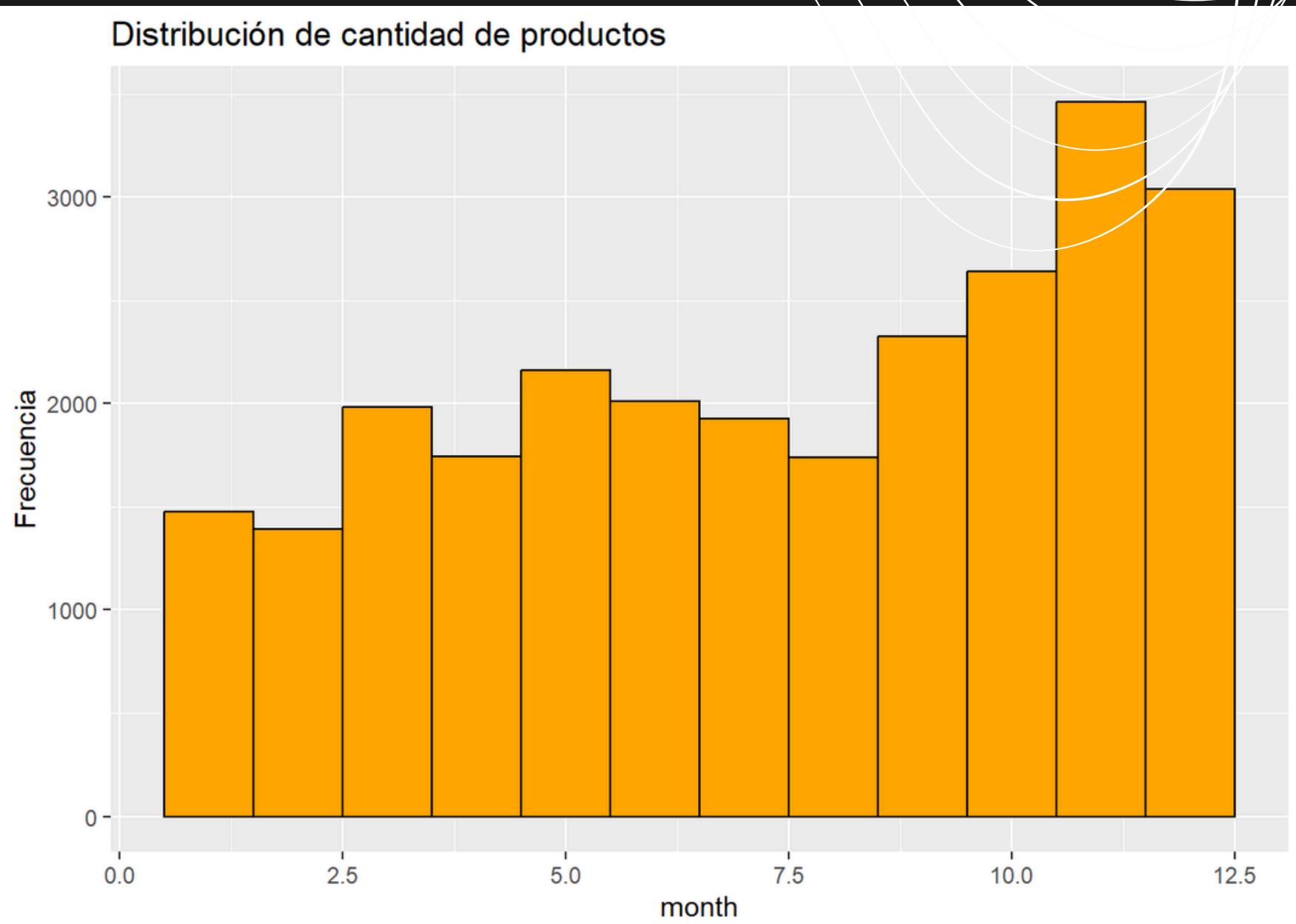
Monto por factura vs Hora de compra



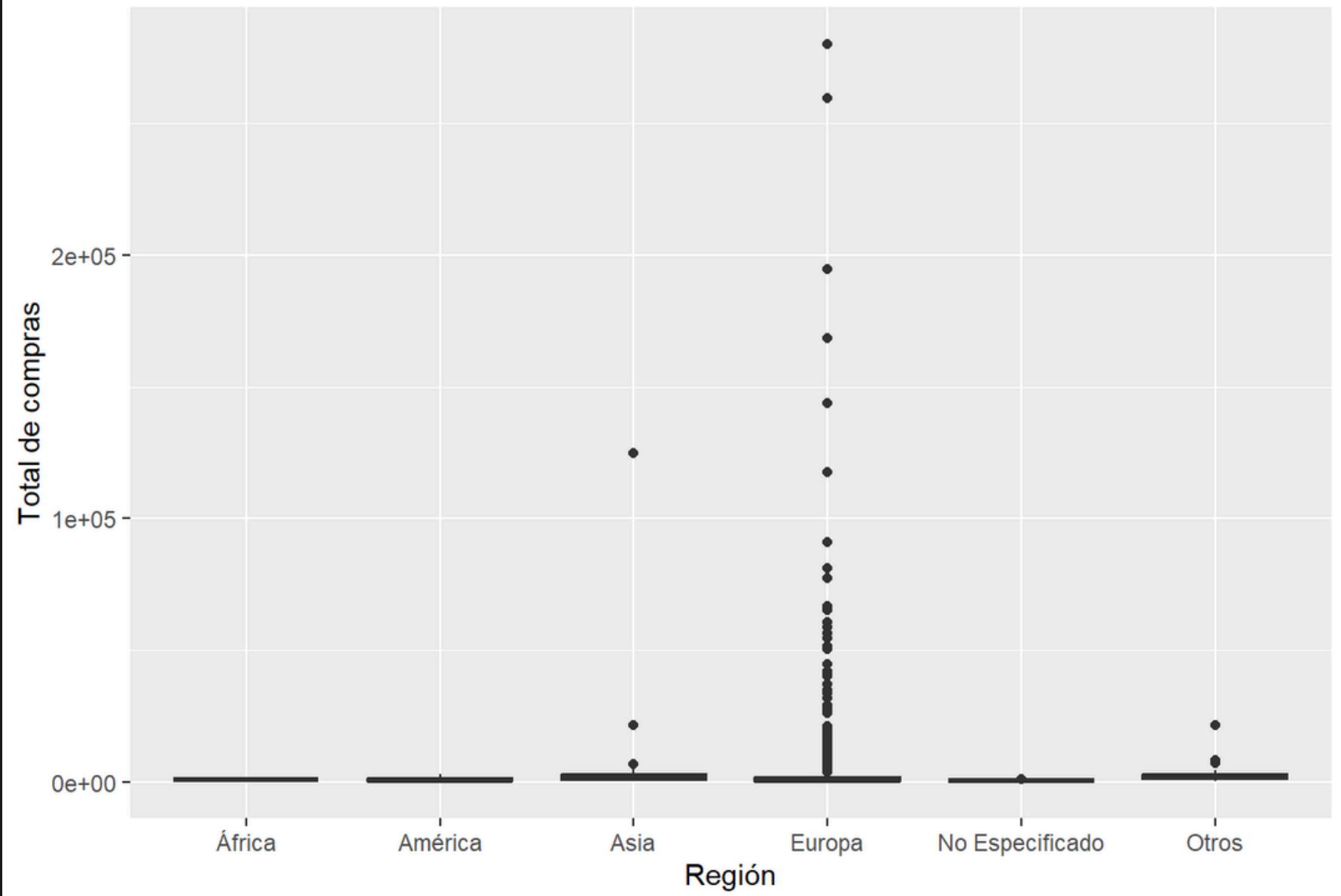
Distribución de facturas por hora



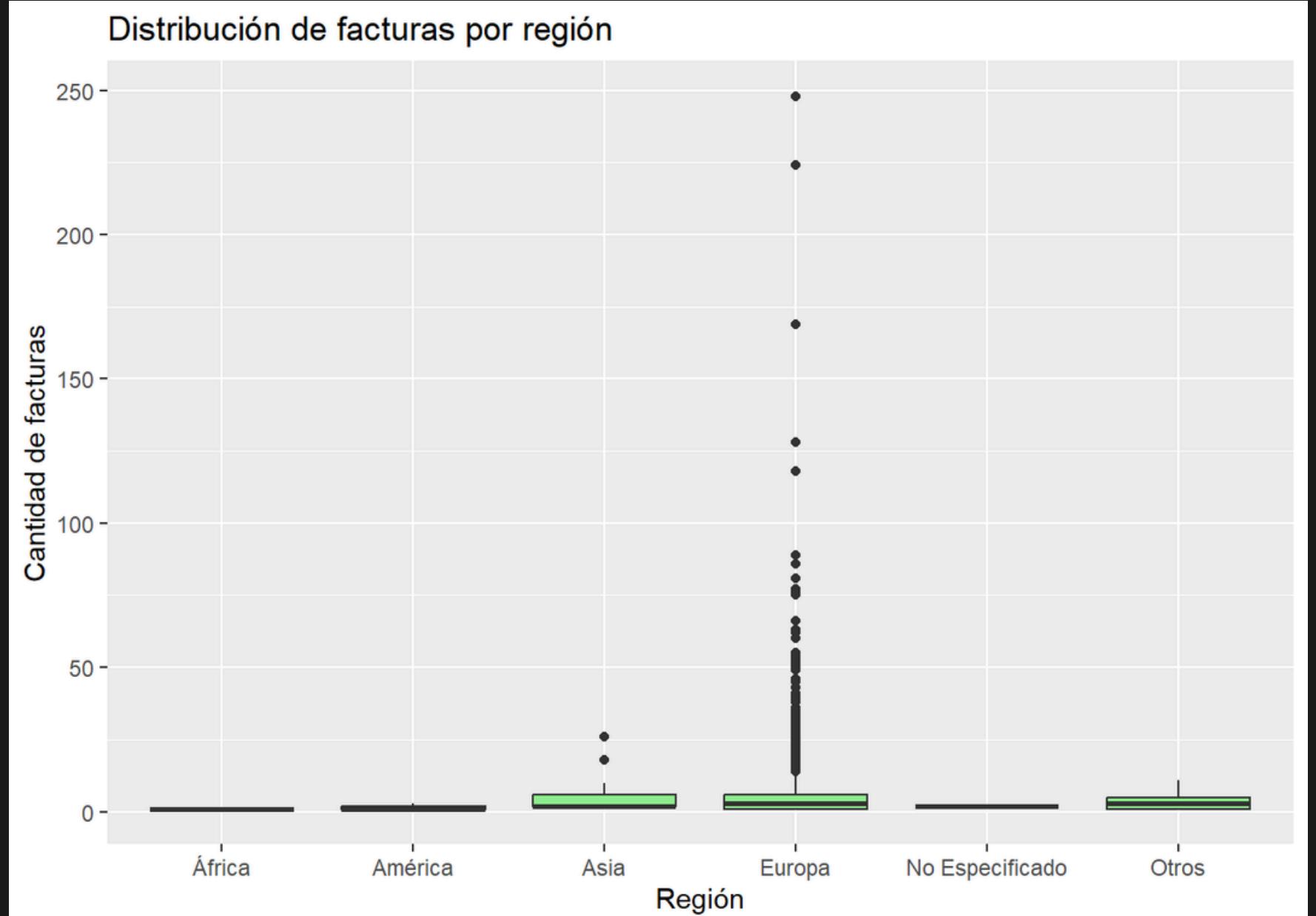
Distribución de cantidad de productos



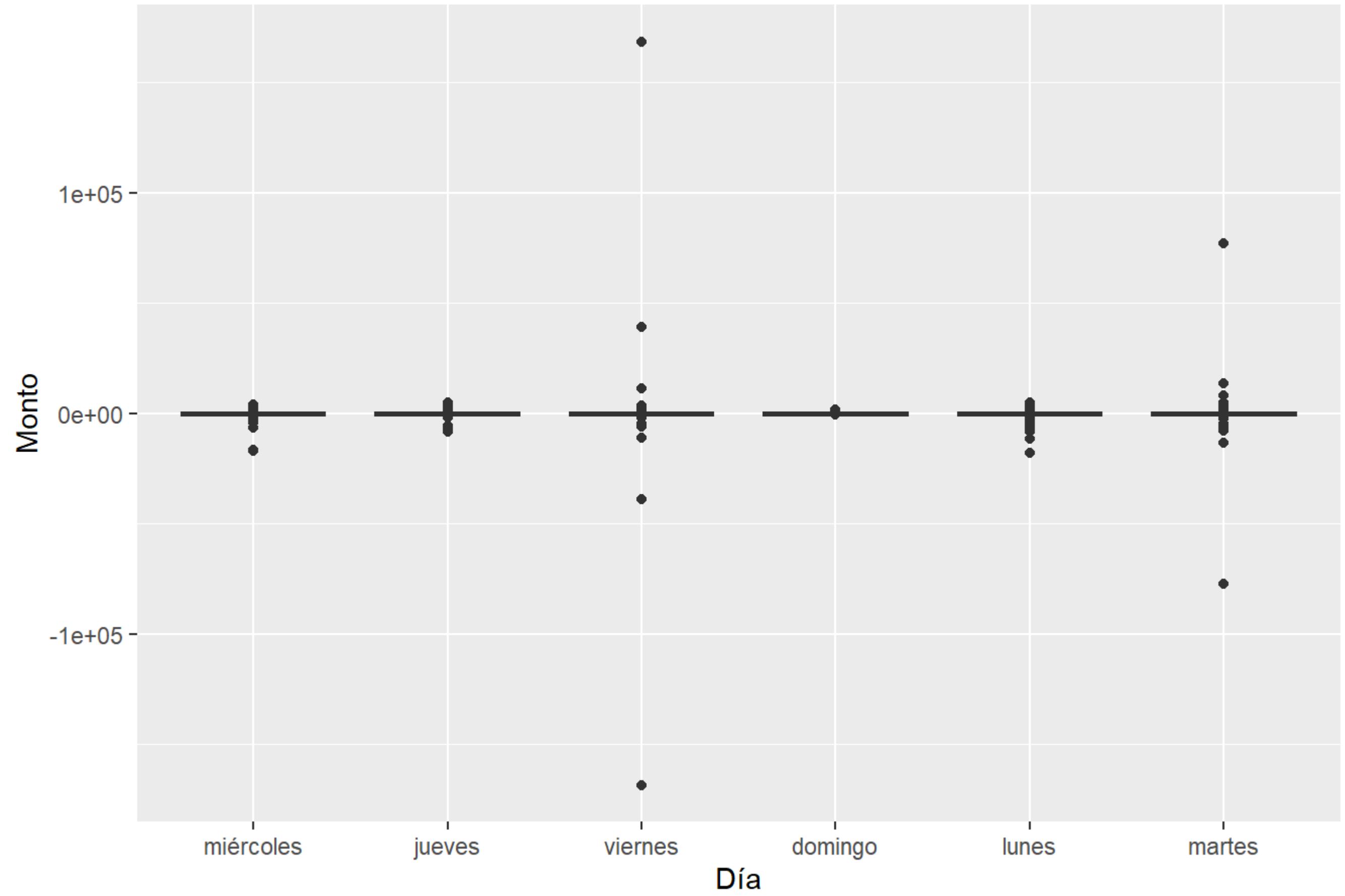
Distribución del total de compras por región



Distribución de facturas por región



## Monto por factura según el día de la semana



# TABLAS DE CONTINGENCIA

Respecto a la predicción de devolución:

	0	1	Sum
África	1	0	1
América	7	2	9
Asia	11	6	17
Europa	2748	1539	4287
No Especificado	4	0	4
Otros	12	9	21
Sum	2783	1556	4339

*region*

	0	1	Sum
miércoles	665	440	1105
jueves	685	436	1121
viernes	422	184	606
sábado	0	0	0
domingo	346	130	476
lunes	337	194	531
martes	328	172	500
Sum	2783	1556	4339

*mes\_max\_compras*

##	##	0	1	Sum
##	1	221	162	383
##	2	217	140	357
##	3	285	152	437
##	4	210	117	327
##	5	217	129	346
##	6	164	112	276
##	7	168	75	243
##	8	136	80	216
##	9	229	117	346
##	10	320	123	443
##	11	409	194	603
##	12	207	155	362
##	Sum	2783	1556	4339

##	##	0	1	Sum
##	6	0	1	1
##	7	6	5	11
##	8	120	85	205
##	9	269	194	463
##	10	393	279	672
##	11	343	206	549
##	12	477	277	754
##	13	421	180	601
##	14	317	134	451
##	15	236	113	349
##	16	124	50	174
##	17	57	20	77
##	18	13	6	19
##	19	5	6	11
##	20	2	0	2
##	Sum	2783	1556	4339

*hora\_max\_compras*

Respecto a la predicción de recompra:

		0	1	Sum
##	África	1	0	1
##	América	7	2	9
##	Asia	5	12	17
##	Europa	1474	2813	4287
##	No Especificado	0	4	4
##	Otros	7	14	21
##	Sum	1494	2845	4339

*region*

		0	1	Sum
##	miércoles	257	848	1105
##	jueves	332	789	1121
##	viernes	224	382	606
##	sábado	0	0	0
##	domingo	185	291	476
##	lunes	253	278	531
##	martes	243	257	500
##	Sum	1494	2845	4339

### *mes\_max\_compras*

##	##	0	1	Sum
##	1	77	306	383
##	2	87	270	357
##	3	126	311	437
##	4	100	227	327
##	5	87	259	346
##	6	86	190	276
##	7	74	169	243
##	8	77	139	216
##	9	156	190	346
##	10	234	209	443
##	11	239	364	603
##	12	151	211	362
##	Sum	1494	2845	4339

##	##	0	1	Sum
##	6	0	1	1
##	7	2	9	11
##	8	54	151	205
##	9	116	347	463
##	10	171	501	672
##	11	149	400	549
##	12	215	539	754
##	13	232	369	601
##	14	208	243	451
##	15	176	173	349
##	16	95	79	174
##	17	55	22	77
##	18	15	4	19
##	19	4	7	11
##	20	2	0	2
##	Sum	1494	2845	4339

### *hora\_max\_compras*

# DIVISIÓN DEL DATASET

El 70% de los datos se destinan a un dataset para entrenar los modelos (**train**) y un 30% de los datos a un dataset para probar la efectividad del modelo (**test**).

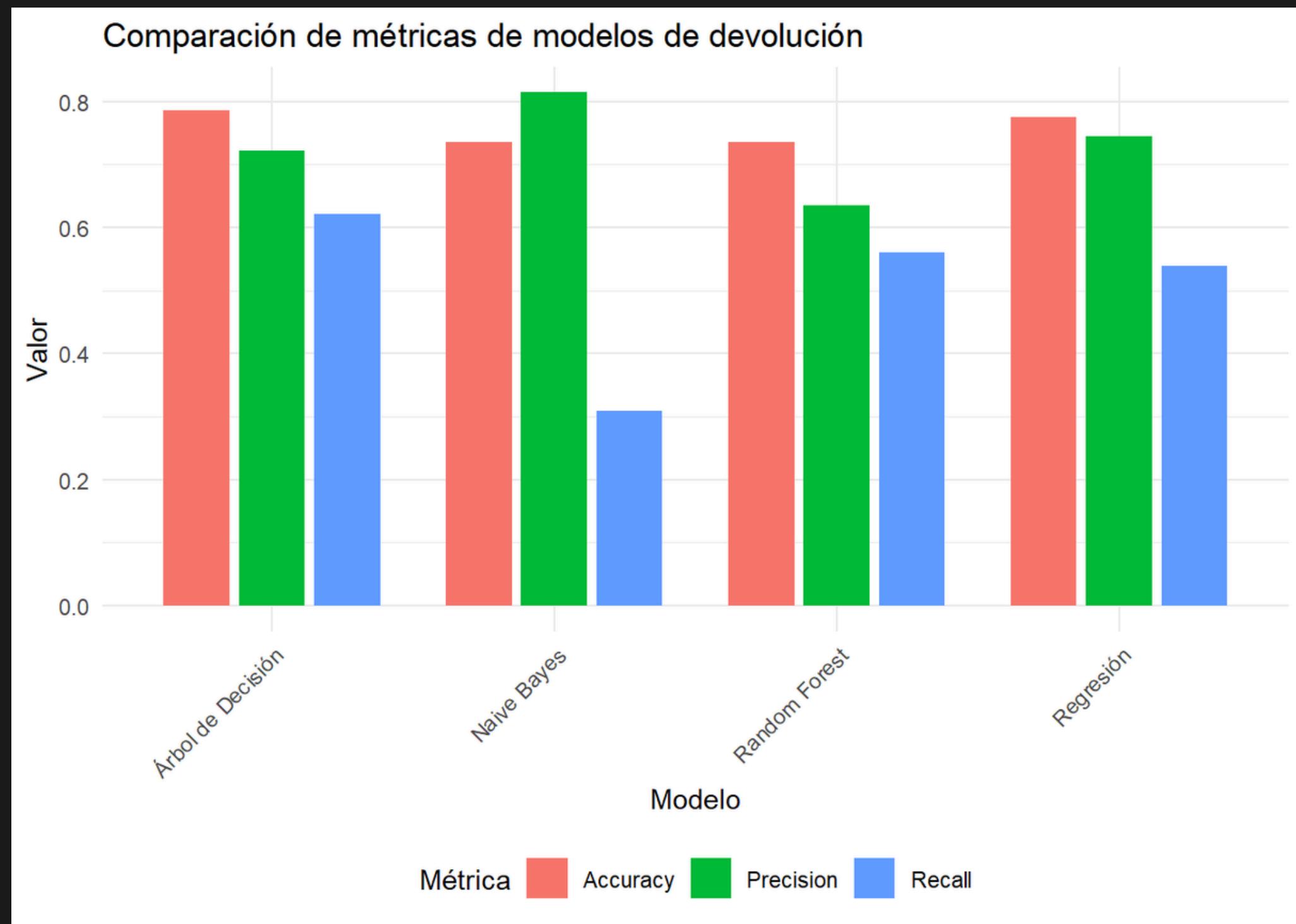
```
# Vector con nombres de columnas que quieres convertir en factor
cols_factor <- c("dia_semana_max_compras", "mes_max_compras", "region", "ha_devuelto", "ha_recomprado")

# Convertir esas columnas en factor usando lapply
clientes[cols_factor] <- lapply(clientes[cols_factor], as.factor)

## 70% de La muestra va para train
smp_size <- floor(0.70 * nrow(clientes))
## Establecemos una semilla para que a todos nos quede el mismo valor
set.seed(0020)
## Creamos Los archivos train y test
train_ind <- sample(seq_len(nrow(clientes)), size = smp_size)
train <- clientes[train_ind, ]
test <- clientes[-train_ind, ]

train2 = train
train3 = train
train = select(train, - ha_recomprado)
```

# MODELOS DE CLASIFICACIÓN: DEVOLUCIÓN



**Accuracy:** qué proporción de todas las predicciones (tanto de clientes que devuelven como de los que no) son correctas.

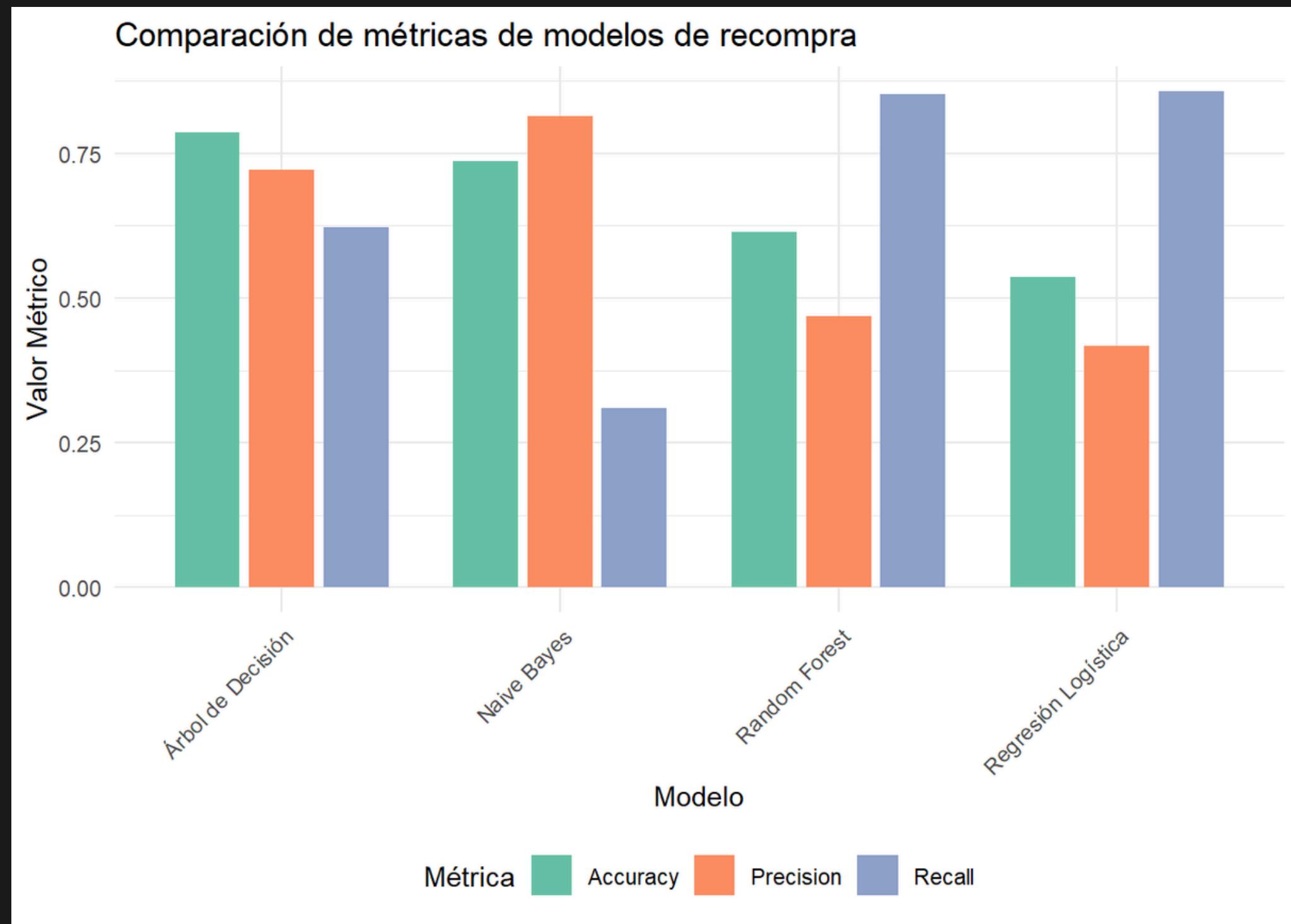
**Precision:** de los clientes que el modelo predice como “devolverán”, qué porcentaje efectivamente devuelve el producto; es decir, nos indica cuántos de los alertados como potenciales retornos son verdaderos retornos.

**Recall:** de todos los clientes que realmente devuelven, qué porcentaje somos capaces de detectar.



Dado que buscamos priorizar el recall, el modelo óptimo sería el de **Random Forest**, ya que presenta el recall más alto de todos, así como un accuracy bastante alto.

# MODELOS DE CLASIFICACIÓN: RECOMPRA



**Accuracy:** qué proporción de todas las predicciones (tanto de clientes que recompran como de los que no) son correctas.

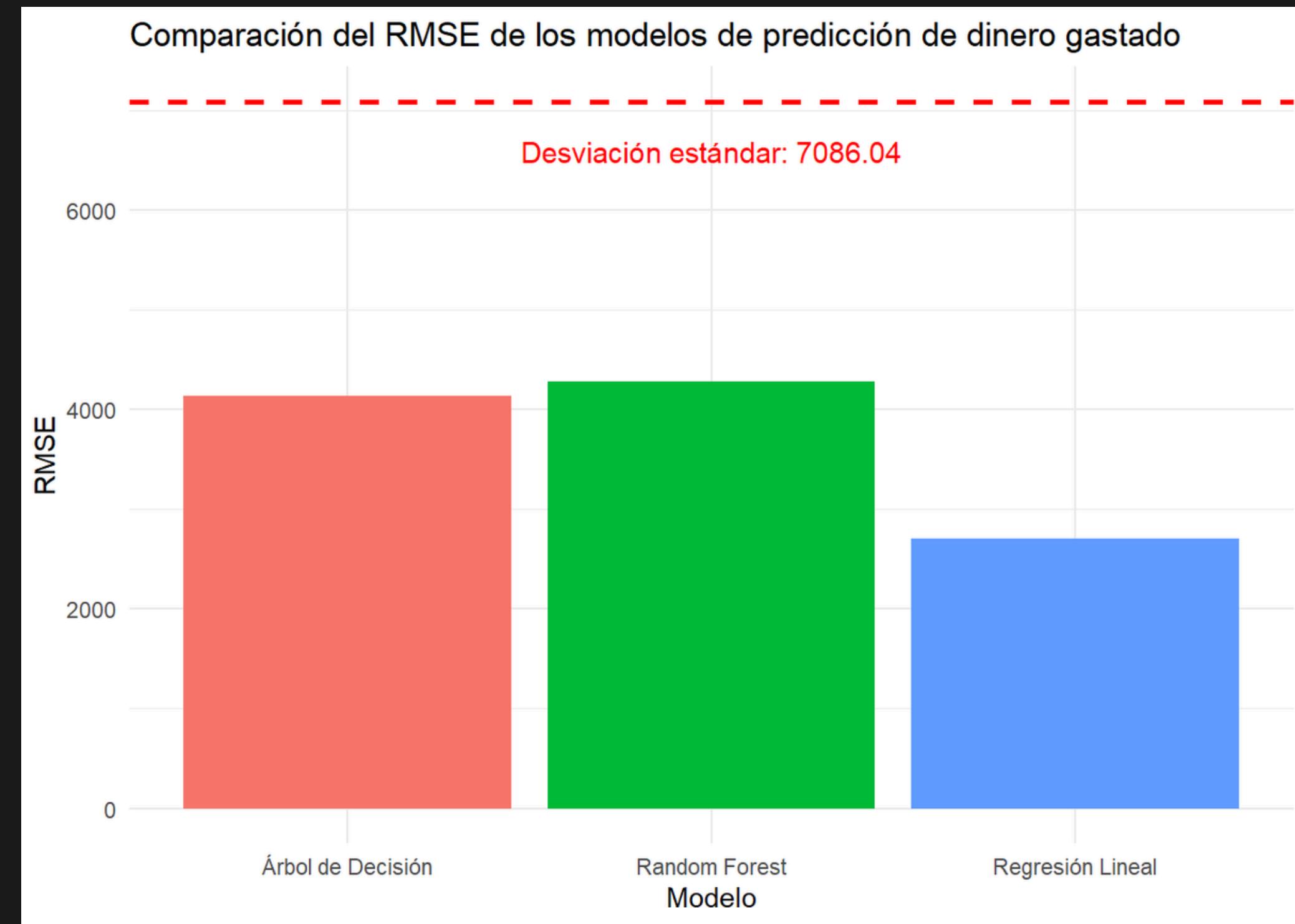
**Precision:** de los clientes que el modelo predice como “recomprarán”, qué porcentaje efectivamente vuelve a comprar

**Recall:** de todos los clientes que realmente recompran, qué porcentaje logramos identificar.



Dado que buscamos priorizar el precision, el modelo óptimo sería el de **Naive Bayes**, puesto que presenta el precision más alto de todos, así como un accuracy bastante bueno.

# MODELOS DE PREDICCIÓN: DINERO GASTADO





Los tres modelos presentan un RMSE bastante aceptable, dado que se mantienen por debajo de la deviación estándar. Sin embargo, el RMSE del modelo de regresión lineal es significativamente menor que el de los otros modelos, lo que lo convierte en la mejor opción.

# IMPLICACIONES ECONÓMICAS DE LA IMPLEMENTACIÓN DE LOS MODELOS

Ahorro por devoluciones evitadas: USD 571,302.90

Costo intervenciones (devoluciones): USD 13,353.12

Ganancia adicional por recompra: USD 26,432.89

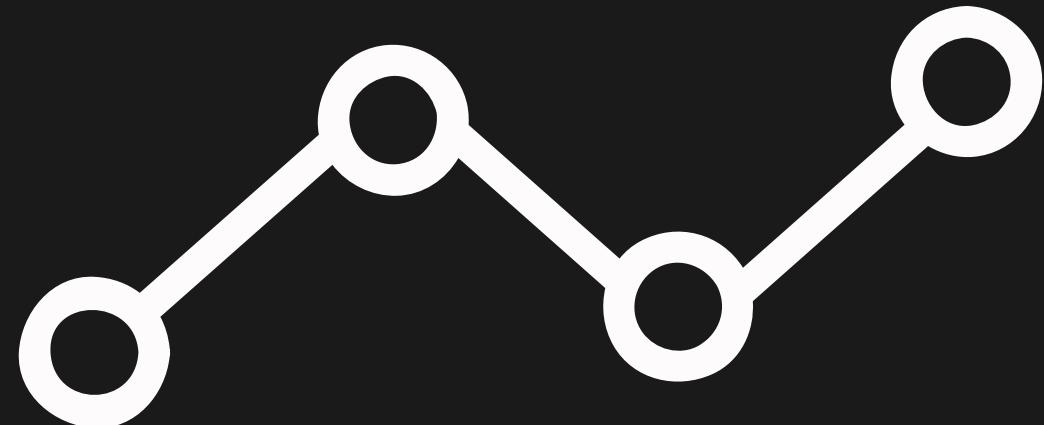
Costo intervenciones (recompra): USD 3,247.48

Impacto económico neto estimado: USD 581,135.20



# CONCLUSIONES

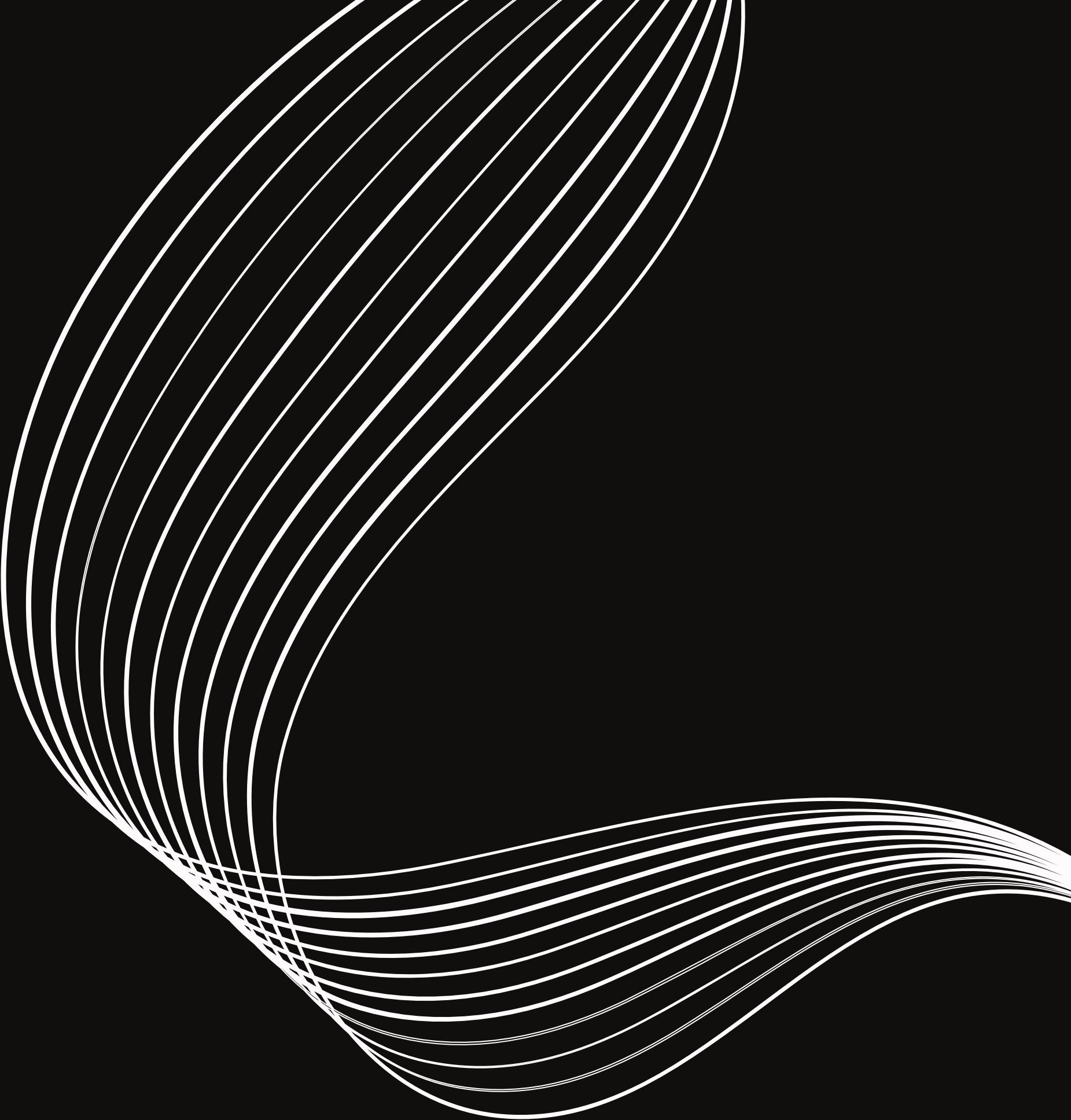
- 1) Análisis exploratorio clave:** La exploración con gráficos y tablas permitió identificar patrones útiles y descartar variables irrelevantes (como la región en relación a devoluciones), mejorando la robustez y precisión de los modelos predictivos.
- 2) Evaluación de modelos:** Se compararon distintos modelos de clasificación, descartando aquellos con bajo desempeño (accuracy, precision, recall), y seleccionando los más adecuados para los objetivos del negocio.
- 3) Impacto económico positivo:** La implementación del pipeline con Naive Bayes y Random Forest generaría un beneficio neto estimado de USD 581,135.20, validando su viabilidad financiera frente a los costos de intervención y falsas alarmas.



# RECOMENDACIONES



**¡GRACIAS POR  
SU ATENCIÓN!**



Data mining & machine learning - S10

# PROYECTO #1

## PREDICCIÓN DE RECOMpra Y DEVOLUCIONES EN CLIENTES DE UNA TIENDA ONLINE

Carlos Angel, José Donado, Carlos Aldana,  
Diego Monroy y Marco Carbajal