UNIVERSIDAD ADOLFO IBÁÑEZ
MAGÍSTER EN DATA SCIENCE

Trabajo Grupal N°1

Nombre estudiante: *Colón Chiang, Marco Cid, Gustavo Dávalos, Oscar Jara, Carlos Lugo*

Curso: *Modelamiento predictivo* – Docente: *Miguel Gaggero*
Fecha de entrega: *March 21st, 2022*

Introducción

En la tabla "MODELAMIENTO_MONTO_FRAUDE", se encuentran más de 7.000 observaciones con el detalle del monto del fraude de una institución bancaria. El detalle de los campos a continuación:

- MONTO_FRAUDE: Variable objetivo. Detalla el monto relacionado al fraude detectado.
- FECHA_INICIAL: Fecha de la primera transacción fraudulenta.
- FECHA_DETECCION: Fecha en que se detecta las transacciones fraudulentas.
- N_OPERACIONES: Número de operaciones fraudulentas entre la fecha inicial y la detección.
- TIPO_PRODUCTO: Tipo de producto con el que se desarrolló el fraude (Tarjeta de Crédito o Débito)
- FLAG_CLIENTE_EMPRESA: Indicador de empresa del cliente afectado.
- N_FRAUDES_ANTERIORES: Número de fraudes anteriores que ha tenido el cliente anteriormente.

Actividad 1: Selección de las muestras

Divida las muestras de entrenamiento y validación seleccionando aleatoriamente 70% para entrenamiento y 30% para validación. Trabajar los ítems 2, 3 y 4 solo con la muestra de entrenamiento.

Desarrollo. Posterior a la carga de los datos y corregir el formato de las columnas con fechas, se realiza la selección de muestras utilizando la función *sample*, siguiendo el siguiente código en R.

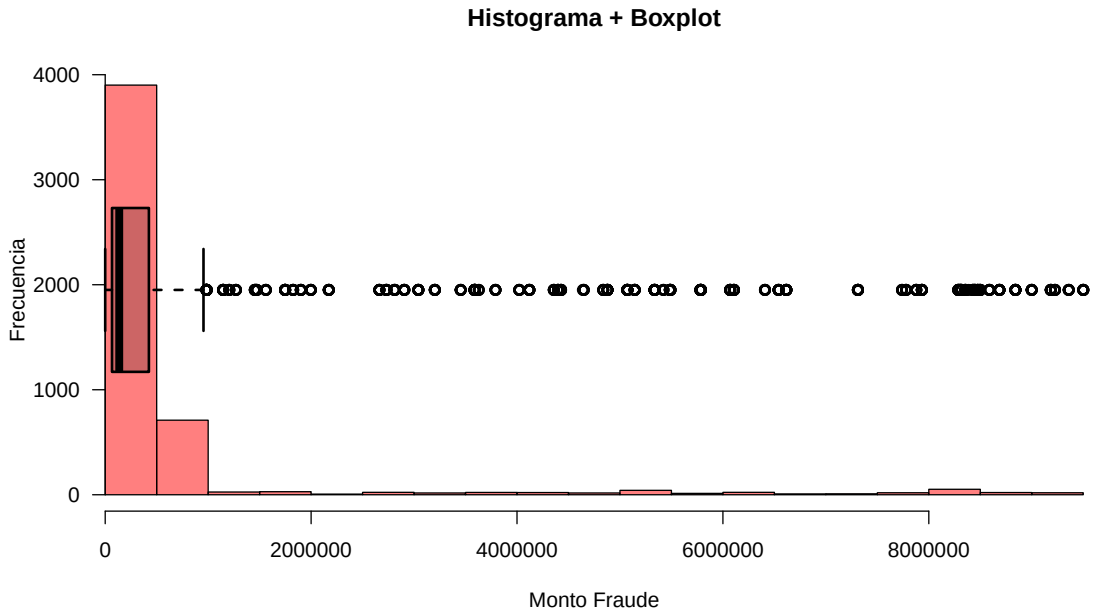
Listing 1: División entrenamiento y validación en R

```
set.seed(3700)# seleccion de semilla
##seleccion de filas
id_entrena <- sample(1:nrow(MOD_MONTO_FRAUDE_FECHAS),
        size = 0.7*nrow(MOD_MONTO_FRAUDE_FECHAS), replace = FALSE)
entrena <- MOD_MONTO_FRAUDE_FECHAS[id_entrena,]
valida <- MOD_MONTO_FRAUDE_FECHAS[-id_entrena,]
```

Actividad 2: Análisis Descriptivo

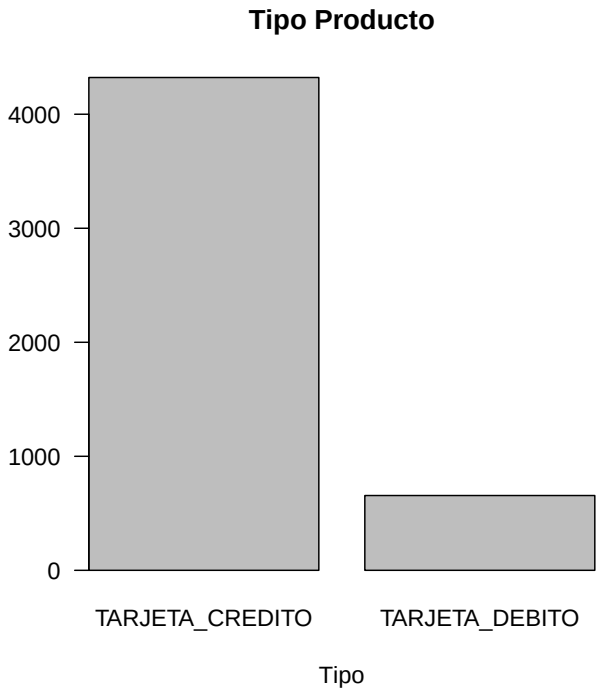
Desarrolle estadísticas descriptivas (tendencia central, percentiles y dispersión) y desarrolle gráficos coherentes con las variables. Interprete los resultados.

Variable MONTO_FRAUDE.

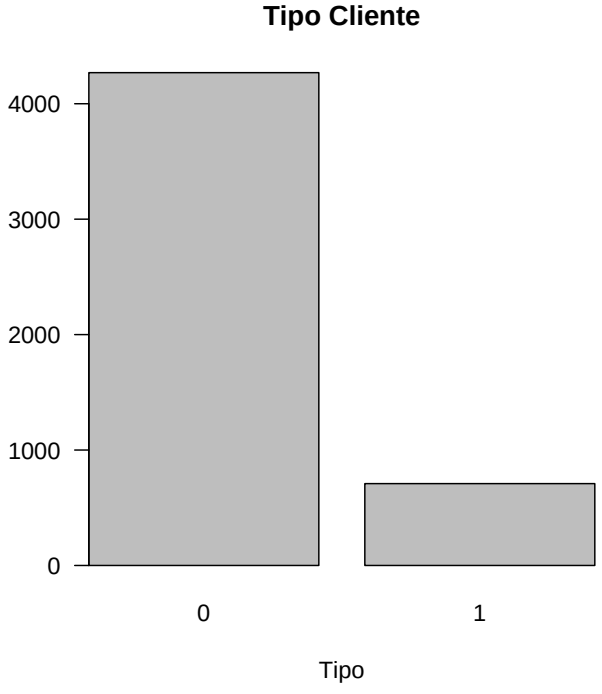


En la figura se evidencia una gran concentración de datos en el primer y segundo tramo por debajo de $1.000.000, y luego los demás datos que podrían ser clasificados como *outliers* (390 registros, 7.8%).

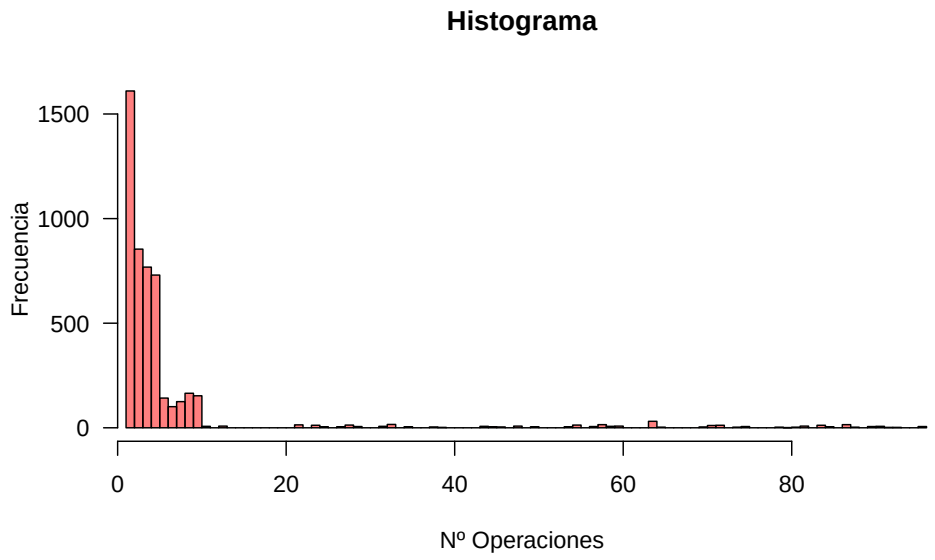
Variable TIPO_PRODUCTO. Se evidencia que en la variable TIPO_PRODUCTO predominan las tarjeta de crédito, exitiendo en el set de entrenamiento 4328 registros de tarjeta de crédito (86,82%), mientras que los 650 restantes corresponden a tarjeta de débito (13,17%).



Variable FLAG_CLIENTE_EMPRESA. En el gráfico siguiente se observa que 4270 registros (85,77%) corresponden a personas y los 708 restantes (14,22%) corresponden a empresas.



Variable N_OPERACIONES. Se observa que en la mayoría de los casos se realizaron pocas operaciones fraudulentas la mayoría menores a 10 operaciones.



Variable N_FRAUDES_ANTERIORES. Se observa que en la mayoría de los casos no existen registros de fraudes anteriores.

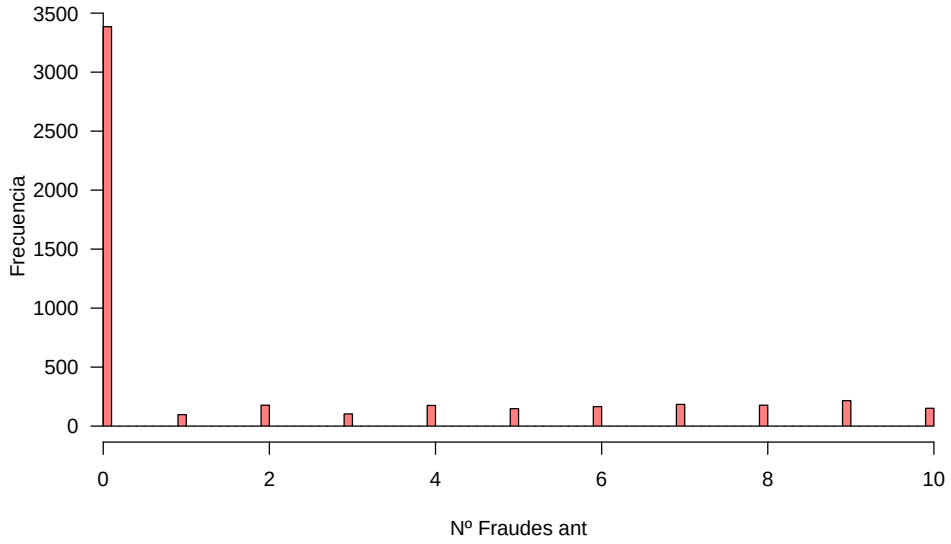


Tabla resumen.
A continuación se presenta una trabla resumen con las principales estadísticas de las variables.

	MONTO_ FRAUDE	N_OPERACIONES	FLAG_CLIENTE_ EMPRESA	N_FRAUDES_ ANTERIORES
count	4978.0	4978.0	4978.0	4978.0
mean	598220.6	7.2	0.1	1.9
std	1499887.4	14.4	0.3	3.2
min	10255.0	1.0	0.0	0.0
25%	79607.0	2.0	0.0	0.0
50%	144995.0	4.0	0.0	0.0
75%	438594.0	5.0	0.0	3.0
max	9399692.0	96.0	1.0	10.0

Actividad 3: Creación y Transformación de Variables

A partir de la muestra de entrenamiento, cree variables como, por ejemplo: número de días del fraude, número de fraudes por día, segmentación de variables cuantitativas, etc. (no considerar la variable objetivo). Desarrolle análisis descriptivo e interprete resultados.

Creación de variable DAY_DETECTION. Se creó la variable DAY_DETECTION como la diferencia entre la fecha de detección y la fecha inicial. Fue ingresada a los modelos como una variable categórica, por lo que se le asignó la media de la variable objetivo, la cual fue almacenada en la variable **var_dif_day**

Creación de la variable N_FRAUDES_DIA. Se creó la variable N_FRAUDES_DIA como el cociente entre el número de operaciones y la cantidad de días transcurridos entre la fecha inicial y la detección.

Creación de la variable MONTO_DETECTION_DIA. Se creó la variable MONTO_DETECTION_DIA como el cociente entre el monto del fraude y la cantidad de días transcurridos entre la fecha inicial y la detección.

Creación de la variable MONTO_PROM_OPERACION. Se creó la variable MONTO_PROM_OPERACION como el cociente entre el monto del fraude y el número de operaciones realizadas.

Creación de la variable GRUPO_OP. Se crearon tres grupos segmentando la variable N_OPERACIONES:

- GRUPO0_OP: 1 operación. Media de la variable objetivo igual a $219.668,3
- GRUPO1_OP 1: de 2 a 5 operaciones. Media de la variable objetivo igual a $201.416,7
- GRUPO2_OP: de 6 a 10 operaciones. Media de la variable objetivo igual a $596.664,3
- GRUPO3_OP: más de 10 operaciones. Media de la variable objetivo igual a $5.340.484,6

A cada uno de los grupos se transformó a la media de la variable objetivo, quedando almacenado como la variable **var_nop**.

Creación de la variable GRUPO_NF. Se agrupó la variable N_FRAUDES_ANTERIORES en cuatro grupos:

- GRUPO1_NF: 0 fraudes anteriores. Media de la variable objetivo igual a $1.703.615.7
- GRUPO2_NF: 1, 4, 5, 6, 8 fraudes anteriores. Media de la variable objetivo igual a $1.330.123.9
- GRUPO3_NF: 2, 3, 9 fraudes anteriores. Media de la variable objetivo igual a $1.093.347,6
- GRUPO4_NF: 7 y 10 fraudes anteriores. Media de la variable objetivo igual a $197.643,2

A cada uno de los grupos se transformó a la media de la variable objetivo, quedando almacenado como la variable **var_nop_ant**.

Transformación de la variable TIPO_PRODUCTO. Se creó la variable **x1_tp** como la transformación a la media de la variable objetivo (MONTO_FRADUE) del TIPO_PRODUCTO.

Transformación de la variable TIPO_CLIENTE. Se creó la variable **x2_tc** como la transformación a la media de la variable objetivo (MONTO_FRAUDE) del TIPO_CLIENTE.

Creación de variable GRUPO_OP_UNI. La variable GRUPO_OP_UNI corresponde a la agrupación de número de operaciones con base en el monto unitario por operacion.

- GRUPO1_UNI_OP: N_OPERACIONES igual a 1 (Media de la variable objetivo igual a $219.668).
- GRUPO2_UNI_OP: N_OPERACIONES distinto a 1, 13, 24, 25, 35 y 11. Cualquier número de operaciones que no corresponda a Grupo 1 ni Grupo 3. (Media de la variable objetivo igual a $68.206).
- GRUPO3_UNI_OP: N_OPERACIONES igual a 13, 24, 25, 35 y 11. (Media de la variable objetivo igual a $404.740).

Para cada grupo se asignó la media de la variable objetivo, la cual fue almacenada en la variable **var_nop_uni**.

Creación de la variable GRUPO_FRA_DAY. La variable GRUPO_FRA_DAY corresponde a la agrupación de N_FRAUDES_DIA con base en promedio del MONTO_DETECTION_DIA.

- GRUPO1_FRA_DAY: N_FRAUDES_DIA menor o igual a 7.
- GRUPO2_FRA_DAY: N_FRAUDES_DIA mayor a 7.

Para cada grupo se asignó la media de la variable objetivo, la cual fue almacenada en la variable **var_nfra_day**.

Transformación logarítmica de las variables. Para la implementación del tercer modelo, se hizo un transformación logística a todas las variables del dataset, quedando almacenadas con su nombre original más el sufijo "_log".

A continuación se presenta una tabla resumen de todas las variables presentes en el modelo:

Table 1: Descripción de las variables

Variable	**Tipo**	**Tratamiento**
MONTO_FRAUDE	Continua	Variable objetivo
FECHA_INICIAL	Fecha	
FECHA_DETECCION	Fecha	
N_OPERACIONES	Discreta	Se segmentará en 3 grupos, luego se transformormará a la media de la variable objetivo
TIPO_PRODUCTO	Categórica	Se transformará a la media de la variable objetivo
FLAG_CLIENTE_EMPRESA	Categórica	Se transformará a la media de la variable objetivo
N_FRAUDES_ANTERIORES	Discreta	
DAY_DETECTION	Categórica	Se transformó a la media de la variable objetivo
N_FRAUDES_DIA	Discreta	Cociente redondeado a cero decimales
MONTO_DETECTION_DIA	Continua	Utilizada para agrupar
MONTO_PROM_OPERACION	Continua	Utilizada para agrupar
GRUPO_OP	Categórica	Resultado de la agrupación
GRUPO_NF	Categórica	Resultado de la agrupación
GRUPO_OP_UNI	Categórica	Resultado de la agrupación
GRUPO_FRA_DAY	Categórica	Resultado de la agrupación
var_nop	Categórica	Transformación de la variable categórica basada en la media de la variable objetivo
var_nop_ant	Categórica	Transformación de la variable categórica basada en la media de la variable objetivo
var_nop_uni	Categórica	Transformación de la variable categórica basada en la media de la variable objetivo
var_dif_day	Categórica	Transformación de la variable categórica basada en la media de la variable objetivo
x1_tp	Categórica	Transformación de la variable categórica basada en la media de la variable objetivo
x2_tc	Categórica	Transformación de la variable categórica basada en la media de la variable objetivo
var_nfra_day	Categórica	Transformación de la variable categórica basada en la media de la variable objetivo
MONTO_FRAUDE_log	Continua	Transformación logarítmica de la variable
N_OPERACIONES_log	Discreta	Transformación logarítmica de la variable
FLAG_CLIENTE_EMPRESA_log	Categórica	Transformación logarítmica de la variable
N_FRAUDES_ANTERIORES_log	Discreta	Transformación logarítmica de la variable
DAY_DETECTION_log	Categórica	Transformación logarítmica de la variable
N_FRAUDES_DIA_log	Discreta	Transformación logarítmica de la variable
MONTO_DETECTION_DIA_log	Continua	Transformación logarítmica de la variable
MONTO_PROM_OPERACION_log	Continua	Transformación logarítmica de la variable
var_nop_log	Categórica	Transformación logarítmica de la variable
var_nop_ant_log	Categórica	Transformación logarítmica de la variable
var_nop_uni_log	Categórica	Transformación logarítmica de la variable
var_dif_day_log	Categórica	Transformación logarítmica de la variable
x1_tp_log	Categórica	Transformación logarítmica de la variable
x2_tc_log	Categórica	Transformación logarítmica de la variable
var_nfra_day_log	Categórica	Transformación logarítmica de la variable

Análisis de correlación entre las variables.
En la figura 1 se presenta un mapa de calor con la correlación entre las variables, donde se puede observar que las siguientes variables tienen alta correlación con la variable objetivo (MONTO_FRAUDE):

- var_nop (0,84), aumentando con respecto a la variable original (N_OPERACIONES).
- N_OPERACIONES (0,74)
- x2_tc (0,64)
- FLAG_CLIENTE_EMPRESA (0,64)
- var_nfra_day (0,75)

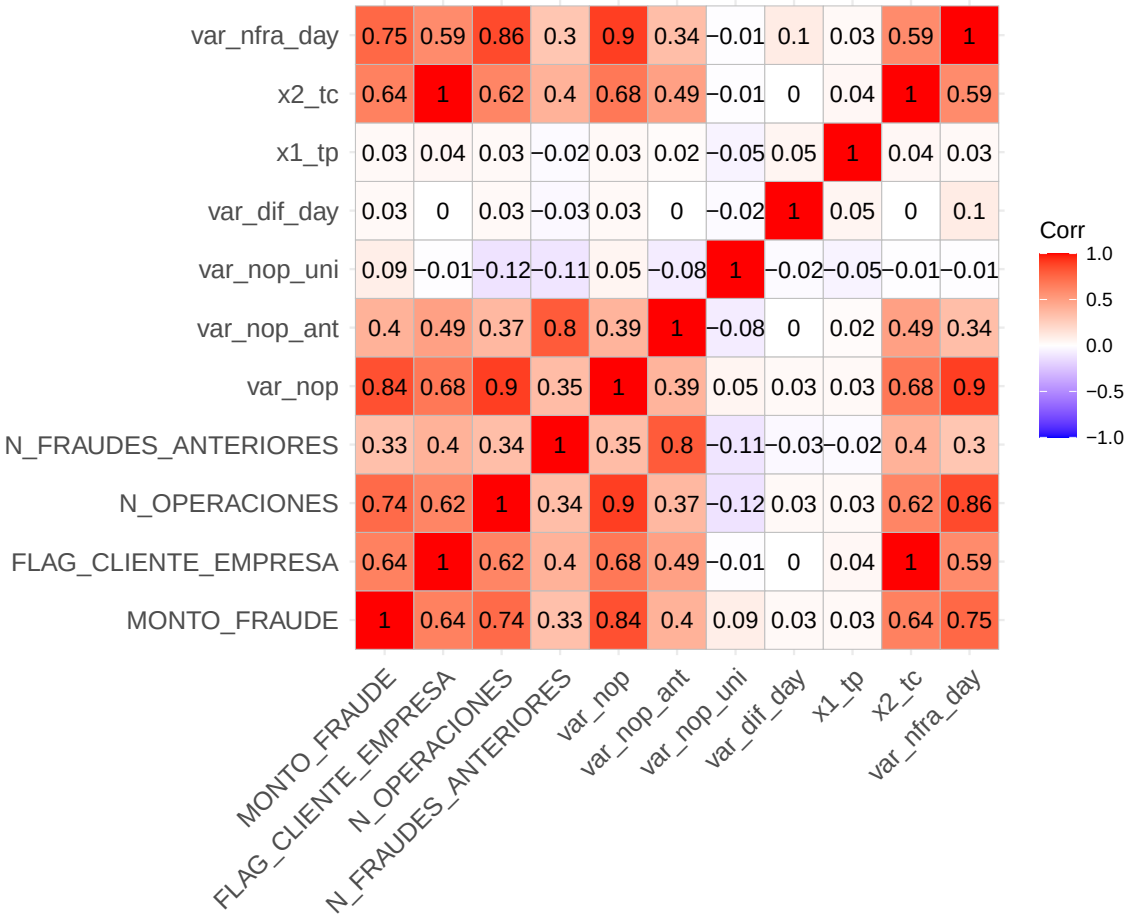


Figure 1: Correlación

Se observa una notable mejora en la correlación de la variable var_nop_ant al aplicar la transformación logítica.

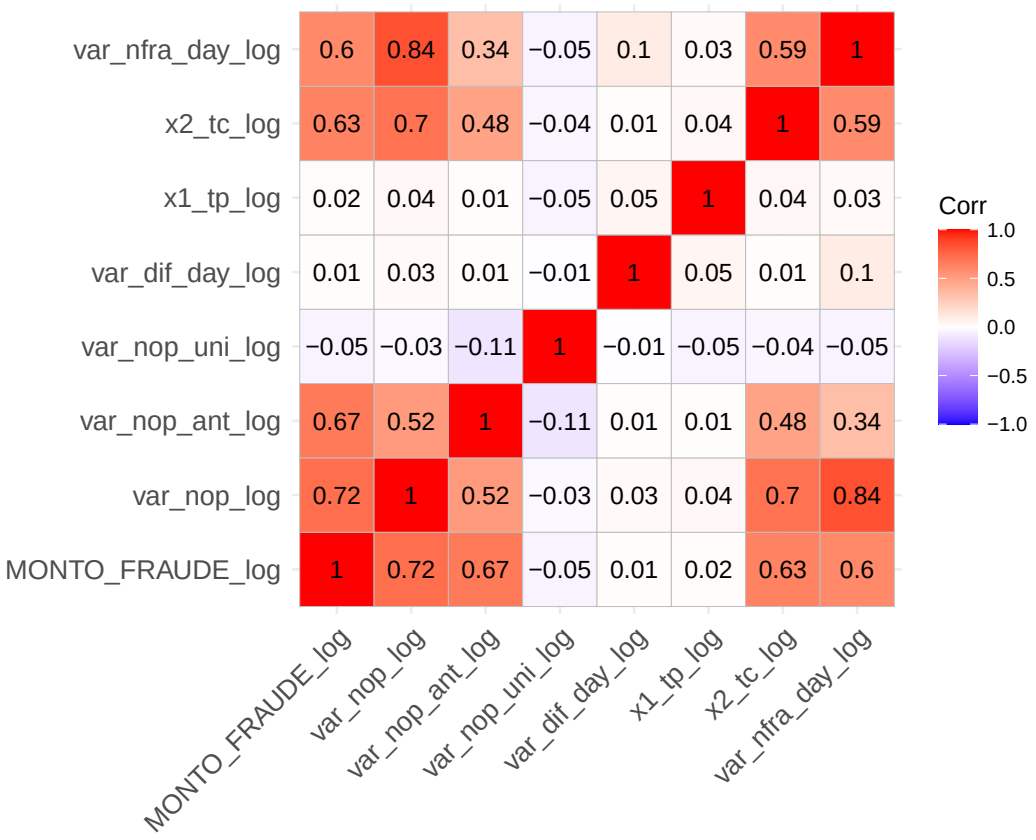
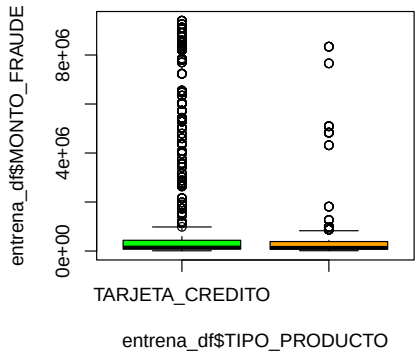
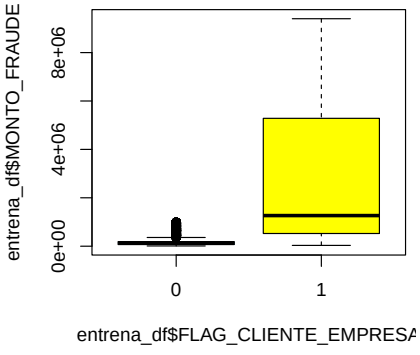


Figure 2: Correlación de las variables logarítmicas

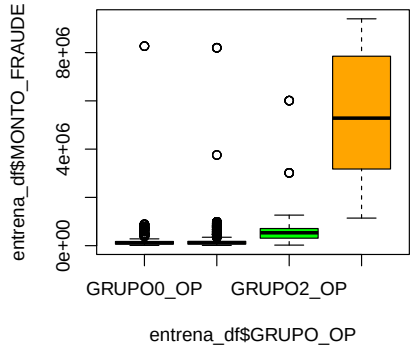
Boxplot de las nuevas variables. En la figura se muestran boxplot de las nuevas variables. En los gráficos se observa que las segmentaciones realizadas abarcan la mayoría de los datos.



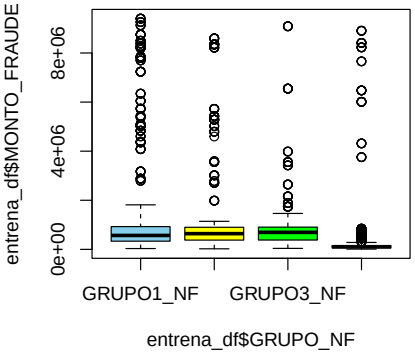
1: Tipo producto



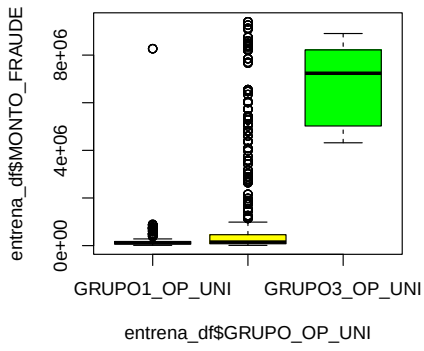
2: Cliente empresa



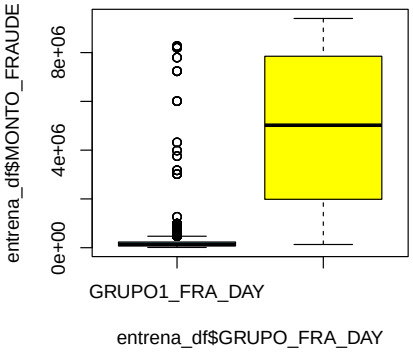
3: Segmentación N_OPERACIONES



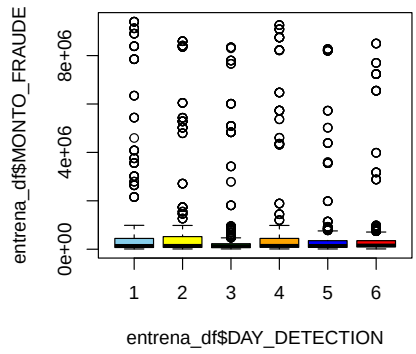
4: Segmentación N_FRAUDES_ANT



5. Segmentación N_OPERACIONES/DIA



6. Segmentación N_FRAUDES/DIA



7. Segmentación DAY_DETECTION

Actividad 4: Modelamiento

Desarrolle al menos 3 modelos de regresión lineal, supervisando la variable MONTO_FRAUDE, con diferentes opciones de variables predictivas, argumentando la selección de estas para cada modelo. Interprete los resultados.

Modelo 1: Variables en bruto.
El primer modelo se realizó con el único objetivo de observar qué tan bien predicen las variables originales, y como punto de partida al ir creando nuevas variables e ir incorporándolas en el modelo. Considerando en la regresión lineal las siguientes variables:

- MONTO_FRAUDE
- N_OPERACIONES
- FLAG_CLIENTE_EMPRESA
- N_FRAUDES_ANTERIORES

Código.

```
formula = MONTO_FRAUDE ~
        N_OPERACIONES + FLAG_CLIENTE_EMPRESA + N_FRAUDES_ANTERIORES
mco_bruto = lm(formula, data = train)
summary(mco_bruto)
```

Coeficientes de la regresión.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-23354	16252	-1,437	0,15078	
N_OPERACIONES	58652	1200	48,861	<2.00E-16	***
FLAG_CLIENTE_EMPRESA	1267472	50814	24,943	<2.00E-16	***
N_FRAUDES_ANTERIORES	13191	4700	2,806	0,00503	**

Comentarios del modelo.

- Las tres variables fueron significativas para el modelo, siendo N OPERACIONES y FLAG_ CLIENTE_EMPRESA más relevantes que N_FRAUDES ANTERIORES.
- El modelo tiene un R^2 igual a 0,6092.
- El modelo tiene un MAPE de 1,64 y un MAE de 396204,7.

Modelo 2: Modelo transformación a la media de la variable objetivo.
Se optó por este modelo debido a que la base de datos presenta un gran número de variables categóricas. Para esto se realizó una transformación de codificación basada en objetivos, es decir, se asigna un valor medio a cada grupo de las categorías. Las variables que fueron transformadas son:

- N_OPERACIONES a var_nop
- N_FRAUDES_ANTERIORES a var_nop_ant
- FLAG_CLIENTE_EMPRESA a x2_tc
- var_nfra_day surge de la variable ya creada N_FRAUDES_DIA.

Código.

```
formula_dic = MONTO_FRAUDE ~ var_nop + var_nop_ant
        + x2_tc + var_nfra_day
mco_dico = lm(formula_dic, data = entrena_df)
summary(mco_dico)
```

Coeficientes de la regresión.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1,18E+05	1,64E+04	-7,188	7,56E-13	***
var_nop	8,88E-01	2,30E-02	38,671	<2.00E-16	***
var_nop_ant	1,33E-01	2,18E-02	6,104	1,11E-09	***
x2_tc	1,83E-01	1,69E-02	10,86	<2.00E-16	***
var_nfra_day	-1,98E-02	4,70E-02	-0,421	0,674	

Comentarios del modelo.

- Este modelo cuenta con una variable adicional al anterior que es el TIPO_PRODUCTO.
- El modelo tiene un R^2 igual a 0,7168.
- El modelo tiene un MAPE de 1,21 y un MAE de 295992,7.
- Aunque posee cuatro variables predictoras, solo tres poseen un alto nivel de significancia.
- El R^2 es más alto en este modelo en comparación al anterior. Esto muestra que 71% de los datos se ajustan al modelo de regresión.

Modelo 3: Modelo selección de variables Stepwise.
Este modelo consiste en ir agregando y quitando las variables predictoras hasta quedarse con el número óptimo de variables que garanticen un buen desempeño del modelo. El grupo que se selecciona es aquel que tiene un AIC menor. El modelo Stepwise comienza con 12 variables.

Código.

```
mod_min = lm(MONTO_FRAUDE ~ 1, data = entrena_df)
formula_dic_wise = MONTO_FRAUDE ~
        var_nop + var_nop_ant + x2_tc + var_nop_uni + var_dif_day
        + x1_tp + DAY_DETECTION + N_FRAUDES_ANTERIORES + N_OPERACIONES
        + N_FRAUDES_DIA + FLAG_CLIENTE_EMPRESA + var_nfra_day
mco_dico_wise = lm(formula_dic_wise, data = entrena_df)

modelo_step_wise = step(mod_min, direction = "forward",
        scope = formula(mco_dico_wise))
summary(modelo_step_wise)
```

Coeficientes de la regresión.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-3,05E+05	3,59E+04	-8,499	<2.00E-16	***
var_nop	8,83E-01	2,34E-02	37,815	<2.00E-16	***
x2_tc	1,87E-01	1,67E-02	11,181	<2.00E-16	***
var_nop_uni	1,41E+00	2,02E-01	6,963	3,77E-12	***
var_nop_ant	2,18E-01	3,30E-02	6,599	4,56E-11	***
N_FRAUDES_DIA	1,29E+04	2,49E+03	5,176	2,35E-07	***
N_OPERACIONES	-6,79E+03	2,25E+03	-3,018	0,00256	**
N_FRAUDES_ANTERIORES	-1,66E+04	6,07E+03	-2,73	0,00636	**
DAY_DETECTION	1,18E+04	7,29E+03	1,614	0,10662	

Comentarios del modelo.

- Se puede observar que el intercepto es negativo al igual que los modelos anteriores.
- El modelo realizó nueve iteraciones, quedando con ocho variables predictoras y descartando var_dif_day, var_nfra_day y x1_tp al ser no significativas.
- El R^2 es de un 0,72, lo que indica que gran parte de los datos se ajustan el modelo de regresión.
- Este modelo cuenta con el R^2 más elevado en comparación a los otros dos.

Modelo 4: Transformación logarítmica.
El cuarto modelo consiste en una regresión lineal con todas las variables en escala logarítmica

Código.

```
formula_log = MONTO_FRAUDE_log ~
        var_nop_log +var_nop_ant_log + x2_tc_log  + var_nfra_day_log
        + N_OPERACIONES_log + var_dif_day_log + var_nop_uni_log

mco_bruto_log = lm(formula_log, data = entrena_df)
summary(mco_bruto_log)
```

Coeficientes de la regresión.

	Estimate	Std. Error	t vaue	Pr(>|t|)	
(Intercept)	-6,85371	0,42078	-16,288	<2e-16	***
var_nop_log	0,44295	0,04269	10,377	<2e-16	***
var_nop_ant_log	0,57123	0,01505	37,958	<2e-16	***
x2_tc_log	0,22497	0,01704	13,199	<2e-16	***
var_nfra_day_log	0,17536	0,02858	6,137	9,08E-10	***
N_OPERACIONES_log	0,08835	0,03674	2,405	0,0162	*
var_dif_day_log	-0,03534	0,01674	-2,111	0,0348	*
var_nop_uni_log	0,12661	0,04508	2,809	0,005	**

Comentarios del modelo.

- El modelo tiene un R^2 igual a 0,7983.
- El modelo tiene un MAPE de 0,88 y un MAE de 307759,4.
- Se aplicó transformación logarítmica a todas las variables y luego de iterar manualmente se obtiene el modelo descrito.

Actividad 5: Selección del modelo

Para los modelos de regresión lineal desarrollados, calcule las predicciones para las muestras de entrenamiento y validación, y seleccione el modelo predictivo. Argumente el criterio de selección e interprete los resultados.

Tabla de resultados en set de entrenamiento. A continuación se presentan las métricas de error

	Modelo	ECM	RECM	MAE	MAPE
1	Bruto	903464166262	950507,3	396204,7	1,6402619
2	Trans. a media	654776654984	809182,7	295993,7	1,2084923
3	Stepwise	642203936140	801376,3	313618,4	1,3890529
4	Logarítmico	845593023645	919561,3	307759,4	0,8872647

Tabla de resultados en set de validación. A continuación se presentan las métricas de error

	Modelo	ECM	RECM	MAE	MAPE
1	Bruto	825765806649	908716,6	390909,4	1,682347
2	Trans. a media	639145325597	799465,7	290208,7	1,1898183
3	Stepwise	628892398396	793027,4	307406	1,3601646
4	Logarítmico	823754903701	907609,4	297448,7	0,8770605

Conclusiones.

- Con base en las métricas de error MAE y MAPE, el modelo con mejor desempeño en la predicción es el **modelo logarítmico**.
- Con base en los resultados obtenidos en el set de validación, el MAPE mejoró un 47,9% en comparación con el modelo 1, 26,3% en comparación con el modelo 2 y 35,5% en comparación con el modelo 3.
- Al crear nuevas variables mejoró la predicción, puesto que son capaces de aumentar la correlación con la variable objetivo.
- La transformación logarítmica mejora notablemente el performance del modelo, debido a que las variables en su mayoría son categóricas.
- Las métricas del modelo de validación mejora con respecto a los de entrenamiento. El MAPE disminuye 1,15% mientras que el MAE disminuye 3,35%.