CIENCIA DE DATOS PARA ECONOMÍA Y NEGOCIOS

Trabajo Final en RStudio - 2C2025

Alumna: Piquero Lucía - 903431 - Grupo 27

Sinopsis del Informe

- Se busca integrar las herramientas de **análisis de datos** vistas en la materia.
- Las técnicas se aplican sobre un dataset que registra las **ventas** y **ganancias** de un **Supermercado** en los Estados Unidos.
- El análisis busca responder a la siguiente ***hipótesis*** principal:

  ➢ ***Los Descuentos Reducen Significativamente la Ganancia del Negocio.***

- Se analizan comportamientos de variables adicionales (**Categoría**, **Región** y **Segmento**) para entender su impacto sobre la conducta de la ganancia.

Datos Crudos - Vista Previa

"Sample - Superstore"

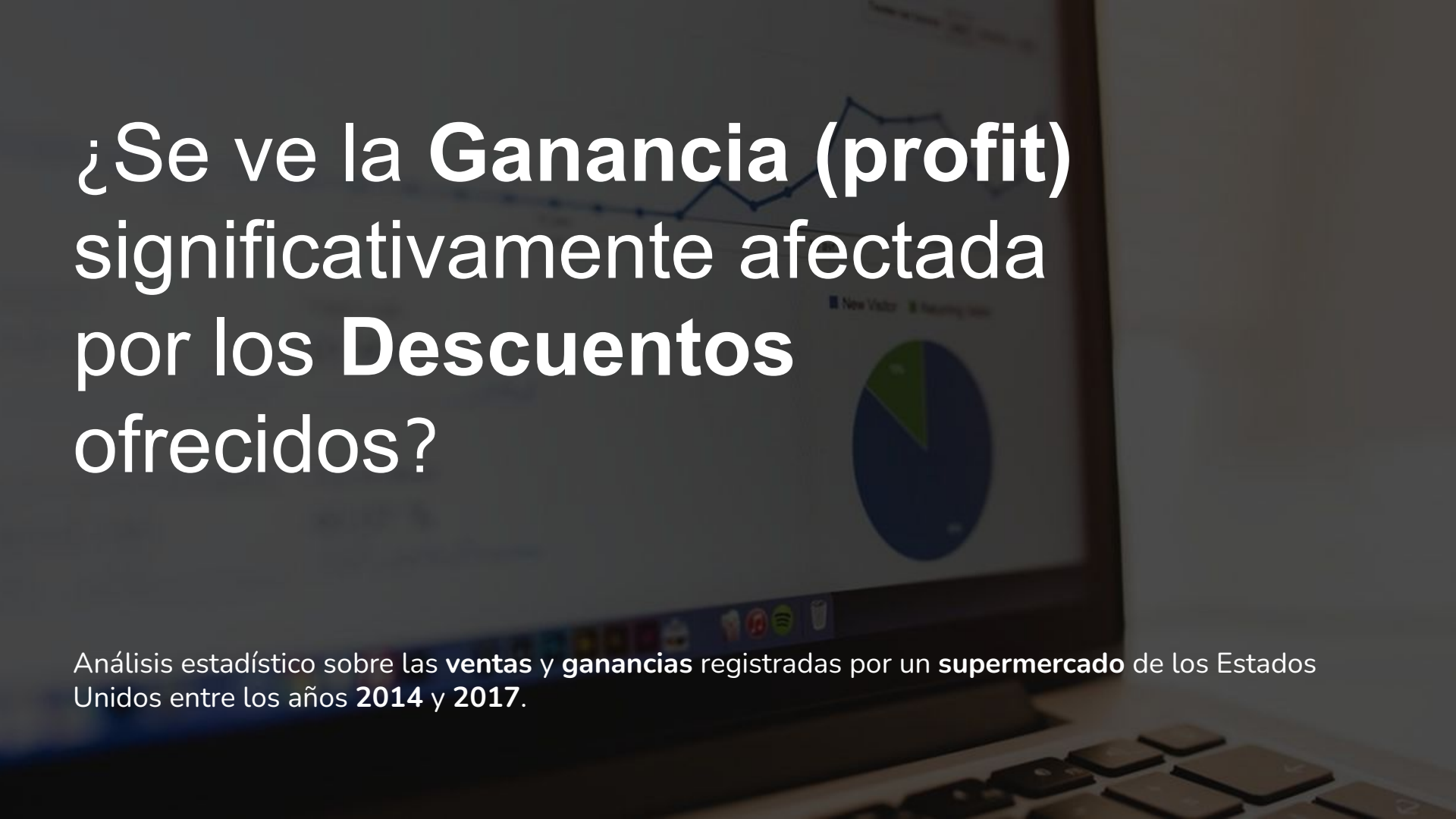
Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	Country	City	State	Postal Code	Region	Product ID	Category	Sub-Category
CA-2016-152156	11/8/2016	11/11/2016	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	42420	South	FUR-BO-10001798	Furniture	Bookcases
CA-2016-152156	11/8/2016	11/11/2016	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	42420	South	FUR-CH-10000454	Furniture	Chairs
CA-2016-138688	6/12/2016	6/16/2016	Second Class	DV-13045	Darrin Van Huff	Corporate	United States	Los Angeles	California	90036	West	OFF-LA-10000240	Office Supplies	Labels
US-2015-108966	10/11/2015	10/18/2015	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	33311	South	FUR-TA-10000577	Furniture	Tables
US-2015-108966	10/11/2015	10/18/2015	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	33311	South	OFF-ST-10000760	Office Supplies	Storage
CA-2014-115812	6/9/2014	6/14/2014	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	FUR-FU-10001487	Furniture	Furnishings
CA-2014-115812	6/9/2014	6/14/2014	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	OFF-AR-10002833	Office Supplies	Art
CA-2014-115812	6/9/2014	6/14/2014	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	TEC-PH-10002275	Technology	Phones
CA-2014-115812	6/9/2014	6/14/2014	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	OFF-BI-10003910	Office Supplies	Binders
CA-2014-115812	6/9/2014	6/14/2014	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	OFF-AP-10002892	Office Supplies	Appliances
CA-2014-115812	6/9/2014	6/14/2014	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	FUR-TA-10001539	Furniture	Tables
CA-2014-115812	6/9/2014	6/14/2014	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	TEC-PH-10002033	Technology	Phones
CA-2017-114412	4/15/2017	4/20/2017	Standard Class	AA-10480	Andrew Allen	Consumer	United States	Concord	North Carolina	28027	South	OFF-PA-10002365	Office Supplies	Paper
CA-2016-161389	12/5/2016	12/10/2016	Standard Class	IM-15070	Irene Maddox	Consumer	United States	Seattle	Washington	98103	West	OFF-BI-10003656	Office Supplies	Binders
US-2015-118983	11/22/2015	11/26/2015	Standard Class	HP-14815	Harold Pawlan	Home Office	United States	Fort Worth	Texas	76106	Central	OFF-AP-10002311	Office Supplies	Appliances
US-2015-118983	11/22/2015	11/26/2015	Standard Class	HP-14815	Harold Pawlan	Home Office	United States	Fort Worth	Texas	76106	Central	OFF-BI-10000756	Office Supplies	Binders
CA-2014-105893	11/11/2014	11/18/2014	Standard Class	PK-19075	Pete Kriz	Consumer	United States	Madison	Wisconsin	53711	Central	OFF-ST-10004186	Office Supplies	Storage
CA-2014-167164	5/13/2014	5/15/2014	Second Class	AG-10270	Alejandro Grove	Consumer	United States	West Jordan	Utah	84084	West	OFF-ST-10000107	Office Supplies	Storage
CA-2014-143336	8/27/2014	9/1/2014	Second Class	ZD-21925	Zuschuss Donatelli	Consumer	United States	San Francisco	California	94109	West	OFF-AR-10003056	Office Supplies	Art
CA-2014-143336	8/27/2014	9/1/2014	Second Class	ZD-21925	Zuschuss Donatelli	Consumer	United States	San Francisco	California	94109	West	TEC-PH-10001949	Technology	Phones
CA-2014-143336	8/27/2014	9/1/2014	Second Class	ZD-21925	Zuschuss Donatelli	Consumer	United States	San Francisco	California	94109	West	OFF-BI-10002215	Office Supplies	Binders
CA-2016-137330	12/9/2016	12/13/2016	Standard Class	KB-16585	Ken Black	Corporate	United States	Fremont	Nebraska	68025	Central	OFF-AR-10000246	Office Supplies	Art
CA-2016-137330	12/9/2016	12/13/2016	Standard Class	KB-16585	Ken Black	Corporate	United States	Fremont	Nebraska	68025	Central	OFF-AP-10001492	Office Supplies	Appliances
US-2017-156909	7/16/2017	7/18/2017	Second Class	SF-20065	Sandra Flanagan	Consumer	United States	Philadelphia	Pennsylvania	19140	East	FUR-CH-10002774	Furniture	Chairs

Datos Crudos - Vista Previa

"Sample - Superstore"

Sub-Category	Product Name	Sales	Quantity	Discount	Profit
Bookcases	Bush Somerset Collection Bookcase	261.9600	2	0.00	41.9136
Chairs	Hon Deluxe Fabric Upholstered Stacking Chairs, Rounded Ba...	731.9400	3	0.00	219.5820
Labels	Self-Adhesive Address Labels for Typewriters by Universal	14.6200	2	0.00	6.8714
Tables	Bretford CR4500 Series Slim Rectangular Table	957.5775	5	0.45	-383.0310
Storage	Eldon Fold 'N Roll Cart System	22.3680	2	0.20	2.5164
Furnishings	Eldon Expressions Wood and Plastic Desk Accessories, Cherr...	48.8600	7	0.00	14.1694
Art	Newell 322	7.2800	4	0.00	1.9656
Phones	Mitel 5320 IP Phone VoIP phone	907.1520	6	0.20	90.7152
Binders	DXL Angle-View Binders with Locking Rings by Samsill	18.5040	3	0.20	5.7825
Appliances	Belkin F5C206VTEL 6 Outlet Surge	114.9000	5	0.00	34.4700
Tables	Chromcraft Rectangular Conference Tables	1706.1840	9	0.20	85.3092
Phones	Konftel 250 Conference$\diamond$phone$\diamond$- Charcoal black	911.4240	4	0.20	68.3568
Paper	Xerox 1967	15.5520	3	0.20	5.4432
Binders	Fellowes PB200 Plastic Comb Binding Machine	407.9760	3	0.20	132.5922
Appliances	Holmes Replacement Filter for HEPA Air Cleaner, Very Large ...	68.8100	5	0.80	-123.8580
Binders	Storex DuraTech Recycled Plastic Frosted Binders	2.5440	3	0.80	-3.8160
Storage	Stur-D-Stor Shelving, Vertical 5-Shelf: 72"H x 36"W x 18 1/2"D	665.8800	6	0.00	13.3176
Storage	Fellowes Super Stor/Drawer	55.5000	2	0.00	9.9900
Art	Newell 341	8.5600	2	0.00	2.4824
Phones	Cisco SPA 501G IP Phone	213.4800	3	0.20	16.0110
Binders	Wilson Jones Hanging View Binder, White, 1"	22.7200	4	0.20	7.3840
Art	Newell 318	19.4600	7	0.00	5.0596
Appliances	Acco Six-Outlet Power Strip, 4' Cord Length	60.3400	7	0.00	15.6884
Chairs	Global Deluxe Stacking Chair, Gray	71.3720	2	0.30	-1.0196

VARIABLES DE PRINCIPAL INTERÉS



¿Se ve la **Ganancia (profit)** significativamente afectada por los **Descuentos** ofrecidos?

Análisis estadístico sobre las **ventas** y **ganancias** registradas por un **supermercado** de los Estados Unidos entre los años **2014** y **2017**.

PRIMER PASO: FORMULACIÓN DE HIPÓTESIS FALSABLE

El desarrollo del trabajo se alimenta del siguiente test de hipótesis:

- *H0: No existe relación estadísticamente significativa entre el nivel de descuento y la ganancia.*
- *H1: Los descuentos reducen significativamente la ganancia del negocio.*

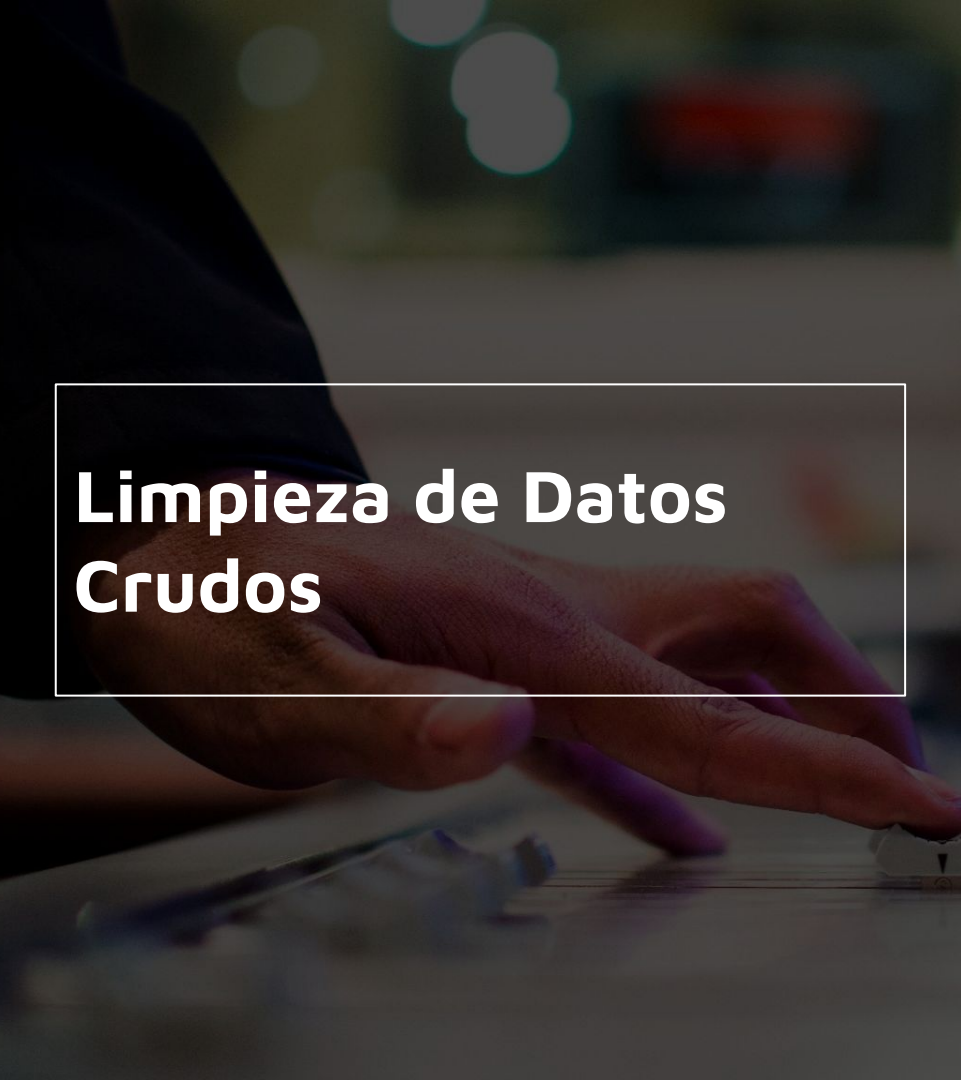
Organización del Proyecto + Carga de Datos

- El ambiente de trabajo es **RStudio.**
- Previo al análisis, se organiza el proyecto en 4 carpetas:

Carpeta	***Contenido***
data	datos crudos, limpios y procesados
functions	funciones predeterminadas
output	gráficos y tablas
scripts	códigos del proyecto

- Una vez organizado, se cargan los **datos crudos** en Rstudio.

Scripts: 00.organizacion_proyecto ; 01.carga_de_datos



Limpieza de Datos Crudos

- Se renombran las columnas en formato **snake_case** para mejorar su **legibilidad** (a través del paquete Janitor).
- Se analiza la existencia de **valores faltantes**.
- Se realizan chequeos **estructurales (skim + glimpse)**.
- Se guardan los **datos limpios** en su carpeta correspondiente.

Script: 02.limpieza_de_datos

Limpieza de Datos Crudos - Resultados

* Variables renombradas exitosamente

```
# 5. Renombrar columnas para estandarizar
datos_clean <- janitor::clean_names(datos_raw)
```

order_id	order_date	ship_date	ship_mode	customer_id
CA-2016-152156	11/8/2016	11/11/2016	Second Class	CG-12520
CA-2016-152156	11/8/2016	11/11/2016	Second Class	CG-12520
CA-2016-138688	6/12/2016	6/16/2016	Second Class	DV-13045
US-2015-108966	10/11/2015	10/18/2015	Standard Class	SO-20335
US-2015-108966	10/11/2015	10/18/2015	Standard Class	SO-20335
CA-2014-115812	6/9/2014	6/14/2014	Standard Class	BH-11710
CA-2014-115812	6/9/2014	6/14/2014	Standard Class	BH-11710
CA-2014-115812	6/9/2014	6/14/2014	Standard Class	BH-11710
CA-2014-115812	6/9/2014	6/14/2014	Standard Class	BH-11710
CA-2014-115812	6/9/2014	6/14/2014	Standard Class	BH-11710

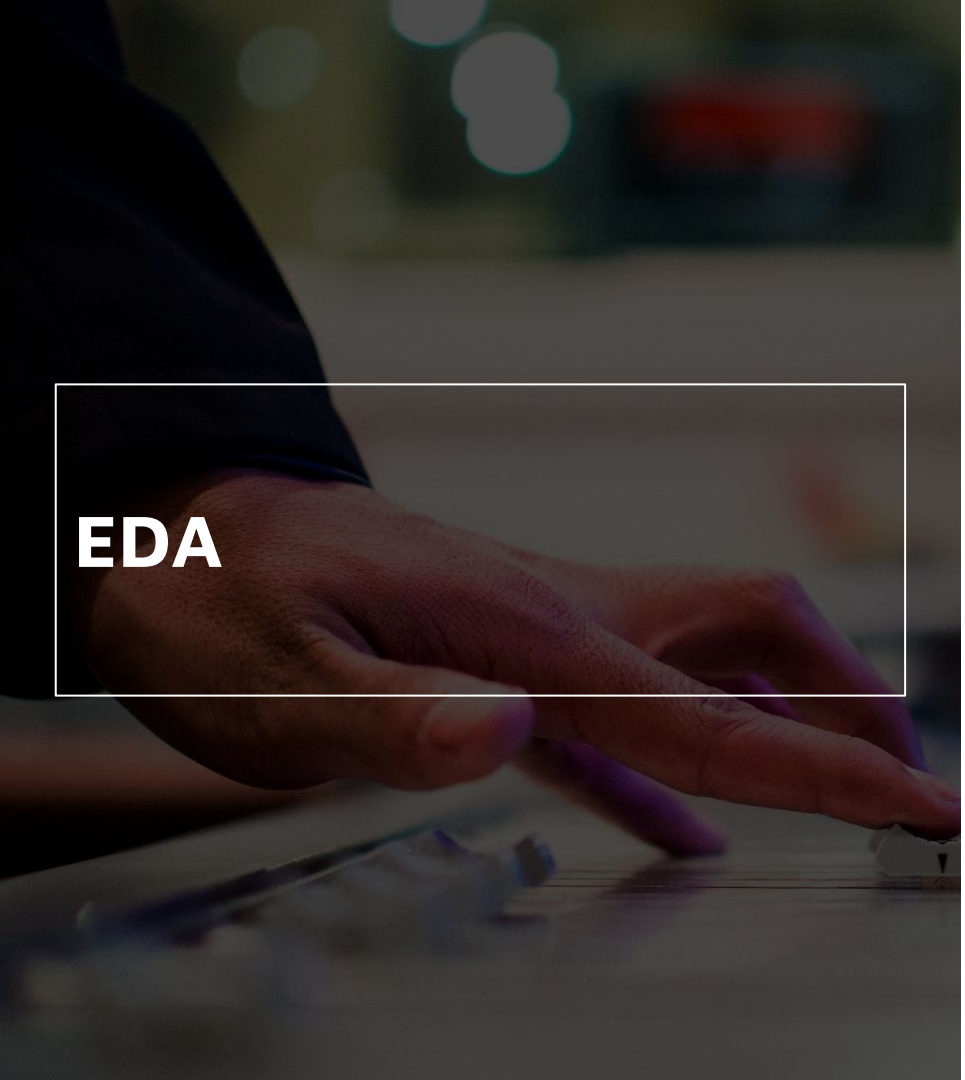
Limpieza de Datos Crudos - Resultados

* No se encontraron valores faltantes

```r
# 6. Detección de valores faltantes
mensaje_proceso("Chequeo de valores faltantes")
faltantes <- colSums(is.na(datos_clean))
print(faltantes)
faltantes <- colSums(is.na(datos_clean))
tabla_faltantes <- data.frame(
  Variable = names(faltantes),
  Missing_Values = faltantes
)

guardar_tabla(tabla_faltantes, "missing_values")
```

	Variable	Missing_Values
row_id	row_id	0
order_id	order_id	0
order_date	order_date	0
ship_date	ship_date	0
ship_mode	ship_mode	0
customer_id	customer_id	0
customer_name	customer_name	0
segment	segment	0
country	country	0
city	city	0
state	state	0
postal_code	postal_code	0
region	region	0
product_id	product_id	0
category	category	0
sub_category	sub_category	0
product_name	product_name	0
sales	sales	0
quantity	quantity	0
discount	discount	0
profit	profit	0



EDA

- Se **reducen** los datos **seleccionando** únicamente las variables de interés en función de la hipótesis dada.
- Se realizan **chequeos exploratorios iniciales** sobre los datos reducidos.
- Se estudia la **correlación** entre las variables “*Profit*” y “*Discount*”.
- Se analiza el tratamiento de posibles **outliers**.

Script: 03_analisis_exploratorio_de_datos

EDA - PRINCIPALES CONCLUSIONES

Se **seleccionan** únicamente las variables de interés.
El nuevo conjunto de datos se guarda como un archivo separado dentro de la carpeta "***processed***".

Las variables seleccionadas son :

- Profit ; Discount; Region; Segment y Category
- Por motivos de trazabilidad también se selecciona Order ID.

```r
# 3. Cargar dataset limpio y reducir variables
# Nota metodológica:
# Se seleccionan solo las variables relevantes para la hipótesis (Profit, Discount)
# y las variables contextuales (Region, Segment, Category).
#Ademas se selecciona la variable Order Id por motivos de trazabilidad.
# Esto permite focalizar el análisis exploratorio sin perder reproducibilidad,
# ya que el dataset completo se conserva en data/clean.
datos_limpios <- readRDS(file.path(dir_data_clean, "datos_limpios.rds"))
datos_reducidos <- select(datos_limpios, order_id, profit, discount, region, segment, category)

# Guardar dataset reducido para análisis posteriores
saveRDS(datos_reducidos, file.path("data/processed", "datos_reducidos.rds"))
readr::write_csv(datos_reducidos, file.path("data/processed", "datos_reducidos.csv"))

mensaje_exito("Dataset reducido guardado en carpeta processed")
```

datos_reducidos

Filter

	order_id	profit	discount	region	segment	category
1	CA-2016-152156	41.9136	0.00	South	Consumer	Furniture
2	CA-2016-152156	219.5820	0.00	South	Consumer	Furniture
3	CA-2016-138688	6.8714	0.00	West	Corporate	Office Supplies
4	US-2015-108966	-383.0310	0.45	South	Consumer	Furniture
5	US-2015-108966	2.5164	0.20	South	Consumer	Office Supplies
6	CA-2014-115812	14.1694	0.00	West	Consumer	Furniture
7	CA-2014-115812	1.9656	0.00	West	Consumer	Office Supplies
8	CA-2014-115812	90.7152	0.20	West	Consumer	Technology
9	CA-2014-115812	5.7825	0.20	West	Consumer	Office Supplies
10	CA-2014-115812	34.4700	0.00	West	Consumer	Office Supplies
11	CA-2014-115812	85.3092	0.20	West	Consumer	Furniture
12	CA-2014-115812	68.3568	0.20	West	Consumer	Technology
13	CA-2017-114412	5.4432	0.20	South	Consumer	Office Supplies
14	CA-2016-161389	132.5922	0.20	West	Consumer	Office Supplies
15	US-2015-118983	-123.8580	0.80	Central	Home Office	Office Supplies
16	US-2015-118983	-3.8160	0.80	Central	Home Office	Office Supplies
17	CA-2014-105893	13.3176	0.00	Central	Consumer	Office Supplies
18	CA-2014-167164	9.9900	0.00	West	Consumer	Office Supplies
19	CA-2014-143336	2.4824	0.00	West	Consumer	Office Supplies
20	CA-2014-143336	16.0110	0.20	West	Consumer	Technology
21	CA-2014-143336	7.3840	0.20	West	Consumer	Office Supplies
22	CA-2016-137330	5.0596	0.00	Central	Corporate	Office Supplies

EDA - PRINCIPALES CONCLUSIONES

Se continúa con chequeos exploratorios iniciales.
Se ***analiza el tipo de variable y la cantidad de valores únicos por columna:***

```
# ------------------------------------------------------------------------
# CHEQUEOS EXPLORATORIOS INICIALES
# ------------------------------------------------------------------------

mensaje_proceso("Tipos de variables por columna")
print(sapply(datos_reducidos, class))

mensaje_proceso("Cantidad de valores únicos por columna")
print(sapply(datos_reducidos, function(x) length(unique(x))))
```

```
> mensaje_proceso("Tipos de variables por columna")
[INFO] Tipos de variables por columna
> print(sapply(datos_reducidos, class))
   order_id      profit    discount      region     segment    category
"character"   "numeric"   "numeric" "character" "character" "character"
> mensaje_proceso("Cantidad de valores únicos por columna")
[INFO] Cantidad de valores únicos por columna
> print(sapply(datos_reducidos, function(x) length(unique(x))))
order_id   profit discount   region  segment category
    5009     7287       12        4        3        3
```

* De las variables seleccionadas solo Profit y Discount son numéricas.

Se corre un **resumen estadístico** sobre los datos reducidos:

```
mensaje_proceso("Resumen estadístico general")
write.csv(summary(datos_reducidos), file.path(dir_output_tables, "summary_reducidos.csv"))
print(summary(datos_reducidos))
```

```
[INFO] Resumen estadístico general
> write.csv(summary(datos_reducidos), file.path(dir_output_tables, "summary_reduci
dos.csv"))
> print(summary(datos_reducidos))
   order_id            profit              discount         region
 Length:9994        Min.   :-6599.978   Min.   :0.0000   Length:9994
 Class :character   1st Qu.:    1.729   1st Qu.:0.0000   Class :character
 Mode  :character   Median :    8.666   Median :0.2000   Mode  :character
                    Mean   :   28.657   Mean   :0.1562
                    3rd Qu.:   29.364   3rd Qu.:0.2000
                    Max.   : 8399.976   Max.   :0.8000
   segment            category
 Length:9994        Length:9994
 Class :character   Class :character
 Mode  :character   Mode  :character
```

EDA - PRINCIPALES CONCLUSIONES

Se analiza la existencia de ***correlación*** significativa entre las variables ***Profit*** y ***Discount***.
Este resultado sirve de indicio previo al test trabajado en los próximos scripts.

```
> # Correlación simple
> cor(datos_reducidos$discount, datos_reducidos$profit, use = "complete.obs")
[1] -0.2194875
```

El nivel de correlación es de **-0.219.**

El análisis nos permite identificar que existe una correlación negativa (tal como se intuía). Sin embargo, el valor de esta correlación se considera ***débil*** y debe ser analizado en mayor profundidad.

EDA - PRINCIPALES CONCLUSIONES

Dada la debilidad en la correlación obtenida y la dificultad para detectar una tendencia visible en su comportamiento, se decide estudiar la existencia de **Outliers**:

```
# ------------------------------------------------------------------------------
# DETECCIÓN DE OUTLIERS EN PROFIT
# ------------------------------------------------------------------------------

info_outliers <- detectar_atipicos(datos_reducidos, "profit")
datos_outliers <- datos_reducidos[datos_reducidos$profit %in% info_outliers$valores, ]
datos_no_outliers <- datos_reducidos[!(datos_reducidos$profit %in% info_outliers$valores), ]

cat("Cantidad de outliers detectados:", info_outliers$cantidad, "\n")
cat("Porcentaje sobre total:", round(info_outliers$porcentaje, 2), "%\n")
```

Se detecta que un **18.8%** de la variable **Profit** está compuesto por Outliers.

Se decide estudiar la correlación entre Profit y Discount **excluyendo** valores extremos.

Name	Type	Value
info_outliers	list [4]	List of length 4
cantidad	integer [1]	1881
porcentaje	double [1]	18.82129
limites	double [2]	-39.7 70.8
inferior.25%	double [1]	-39.72413
superior.75%	double [1]	70.81687
valores	double [1881]	219.6 -383.0 90.7 85.3 132.6 -123.9 ...

EDA - PRINCIPALES CONCLUSIONES

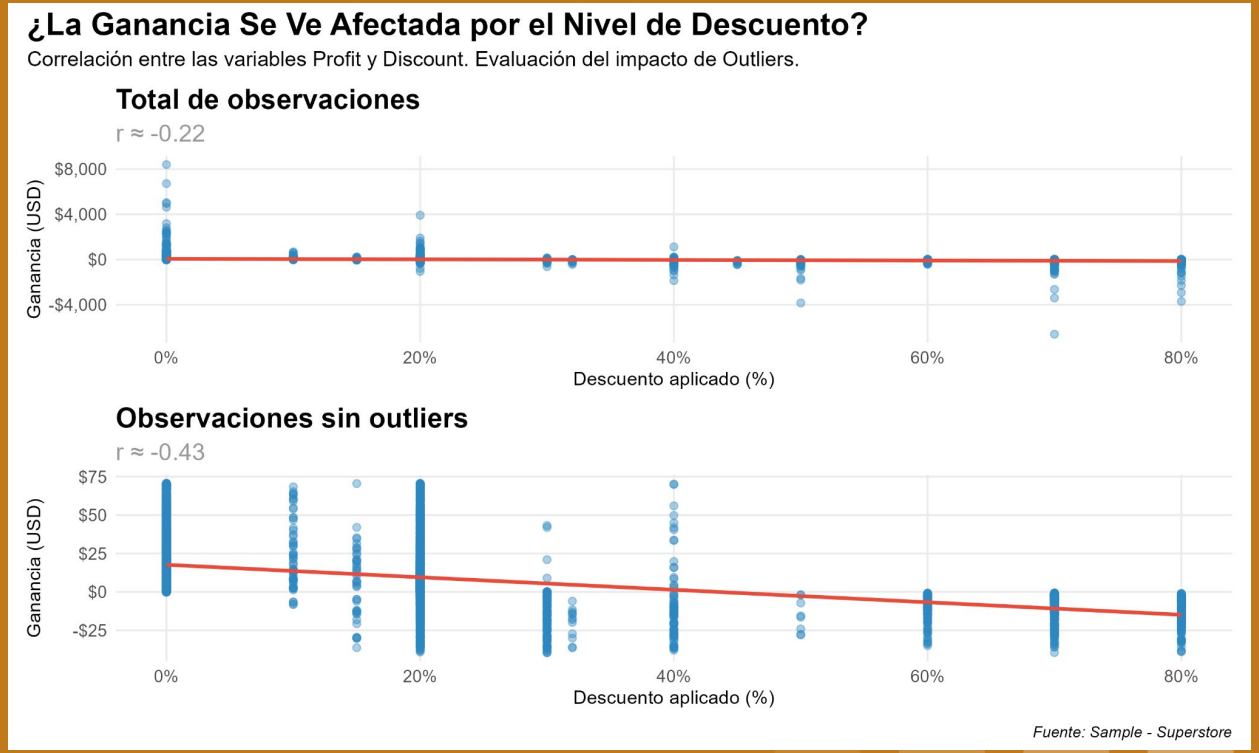
- Al analizar el **total de observaciones**, el coeficiente de correlación es de **-0.22**.

  Como muestra el primer gráfico, la relación entre ambas variables es **poco intuitiva,** dado el amplio rango de valores en el eje Y.

- La correlación **sin outliers** presenta un coeficiente de **-0.43.**

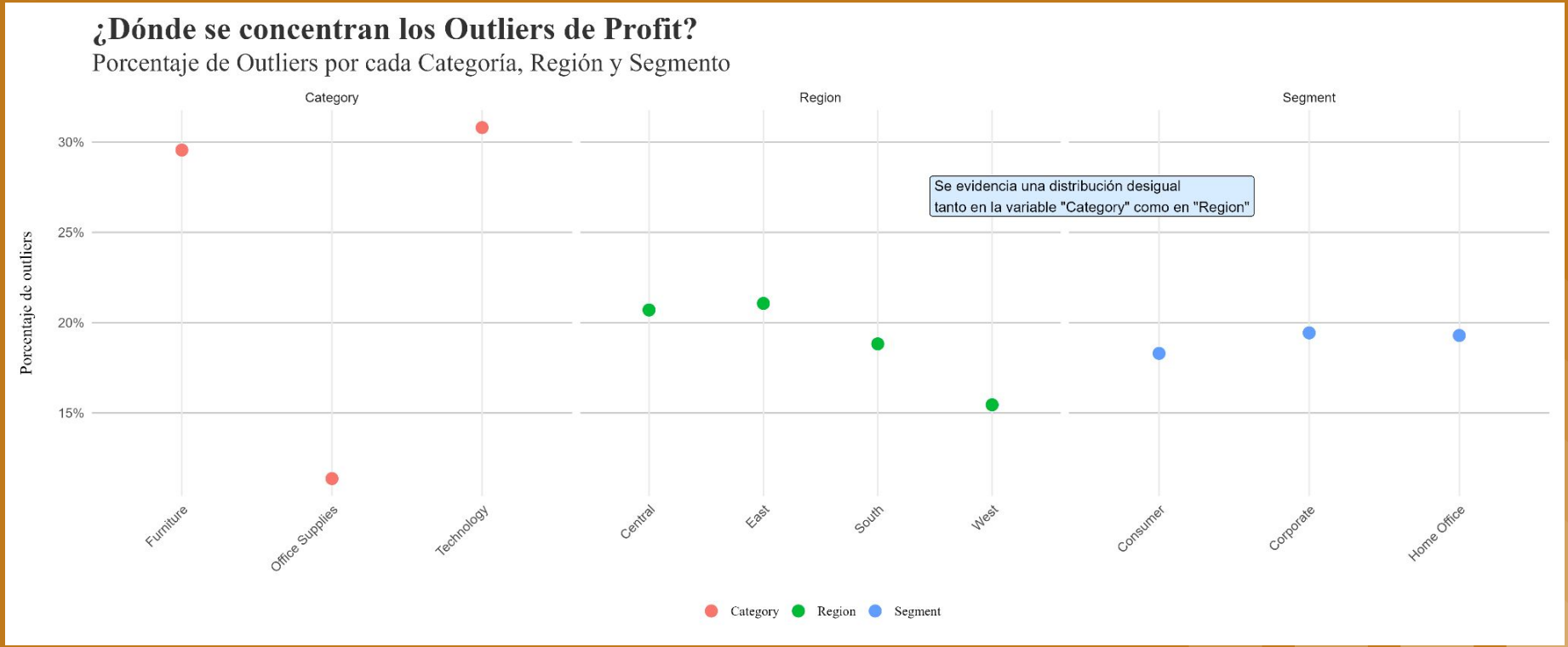
  Este resultado evidencia una correlación más clara que la anterior.

- Si bien la correlación sin outliers es más fuerte, no es suficiente para afirmar significancia en la relación.



EDA - PRINCIPALES CONCLUSIONES

Se analiza la distribución de **outliers en las variables complementarias** para entender qué segmentos podrían generar distorsiones en la relación entre **descuento y ganancia**.



Estadísticas Descriptivas

- Se calculan **medidas de tendencia central** y **dispersión** sobre las variables principales con y sin **outliers**.
- Se estudia la **distribución de frecuencias** de las variables categóricas.
- Se realizan **visualizaciones** complementarias.

Script: 04_estadisticas_descriptivas

ESTADÍSTICAS DESCRIPTIVAS

Se analizan las medidas de **tendencia central y dispersión** para las variables numéricas (con outliers):

	Variable	Media	Mediana	Moda	Desvio	IQR
1	Discount	0.1562027	0.2000	0	0.206452	0.20000
2	Profit	28.6568963	8.6665	0	234.260108	27.63525

Discount:

- La **media** (15%) es menor a la **mediana** (20%). Esto sugiere una distribución **sesgada hacia la izquierda** con valores bajos que reducen el promedio.
- La **moda** es “0”, lo que implica que lo más frecuente es que no se apliquen descuentos.
- Dado que el valor del discount está entre [0,1] un **desvío** de 0.21 es relativamente alto.
- Con un **IQR** de 0.20 podemos concluir que la mayoría de los pedidos tienen descuentos moderados.

Profit:

- La **media** (28.65 USD) es mayor a la **mediana** (8.7 USD). Sugiere una distribución **sesgada hacia la derecha** con valores muy altos que aumentan el promedio.
- La **moda** también es 0, el valor más frecuente se da cuando no existe ganancia alguna.
- El desvío estándar (234 USD) muestra gran variabilidad y el **IQR** muestra que la mayoría de los pedidos presentan una ganancia moderada en términos relativos (27.6 USD).

ESTADÍSTICAS DESCRIPTIVAS

Luego, se analizan las medidas de **tendencia central y dispersión** para las variables numéricas (sin outliers):

Variable	Media	Mediana	Moda	Desvio	IQR
Discount	0.1489203	0.2000	0	0.1976945	0.2000
Profit	11.6040860	7.2576	0	18.6414248	16.9852

Discount:

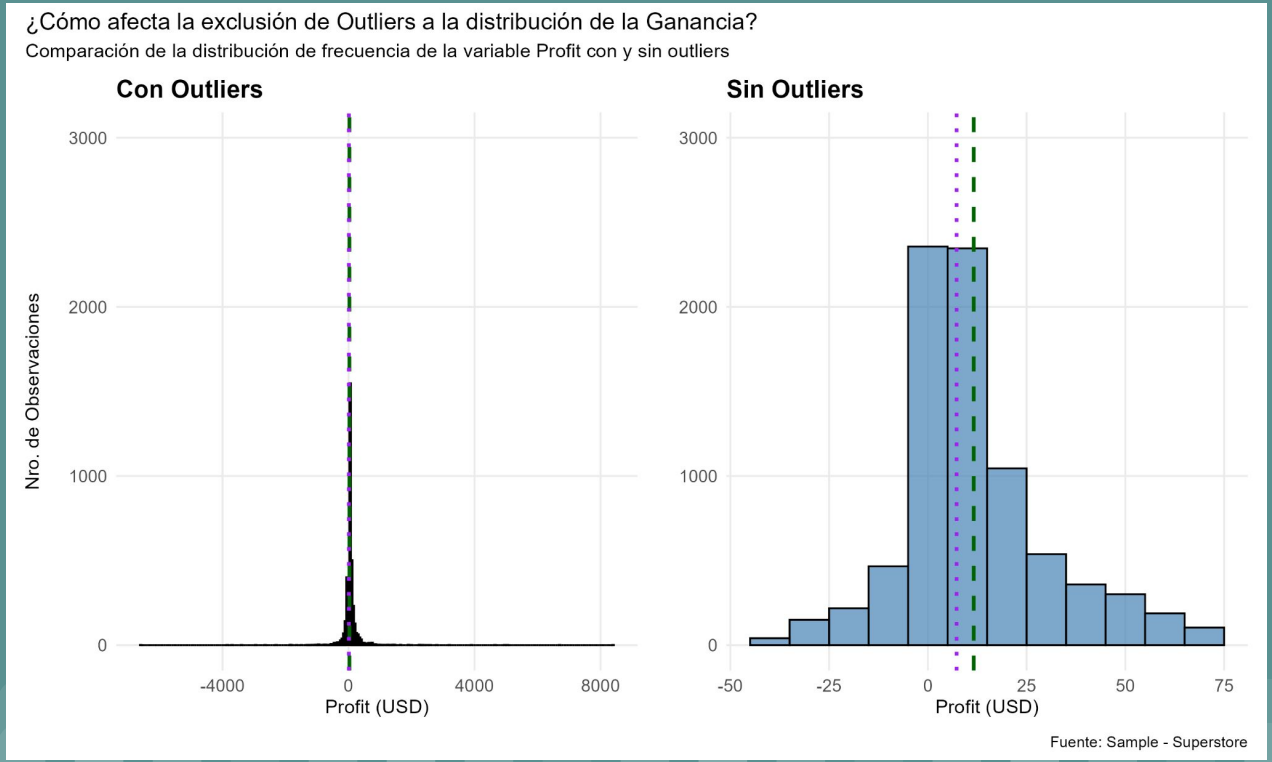
- Se concluye que la variable discount no se ve afectada por los outliers.

Profit:

- La **media** cae drásticamente (28.7 USD ⇒ 11.6 USD).
- La **moda** se mantiene en 0.
- El desvío estándar se desploma (234 USD ⇒ 18.6 USD), mostrando que la variabilidad extrema desaparece al quitar outliers.
- El IQR también baja (27.6 USD ⇒ 17.0 USD), pero sigue reflejando cierta dispersión.

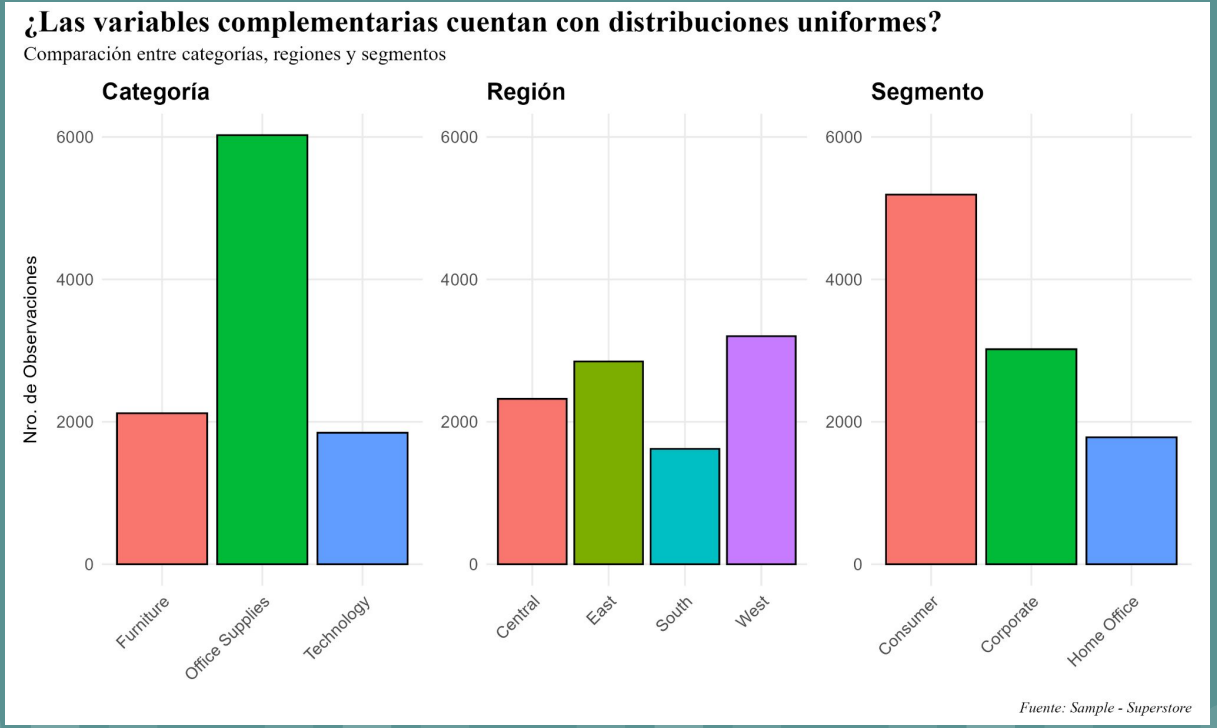
ESTADÍSTICAS DESCRIPTIVAS

- Los **Outliers de Profit** comprimen el histograma y ocultan el comportamiento típico de la variable.
- Al excluirlos, la distribución se vuelve más clara.
- La línea punteada verde refleja la media mientras que la línea violeta representa la mediana.
- El gráfico confirma un leve sesgo hacia la derecha.



ESTADÍSTICAS DESCRIPTIVAS

- Las variables complementarias presentan distribuciones desbalanceadas, especialmente en la **Categoría** de los productos y en el **Segmento** de clientes.
- **Office Supplies y Consumer** concentran la mayoría de observaciones, lo que puede sesgar los patrones de comportamiento de las variables Profit y Discount.



Inferencia Estadística

- Se **eliminan outliers** para continuar con el testeo.
- Se testea la **hipótesis principal** a partir de la cual se alimenta el trabajo.
- Se verifican los **supuestos**.
- Se testea el efecto de las variables categóricas en un **modelo extendido**.

Script: 05_inferencia_estadistica

Exclusión de Outliers en “Profit”

A partir del análisis hecho en las diapositivas previas, se decide **eliminar los outliers** antes de correr el test final.

Se guardan los datos sin outliers en un archivo dentro de la carpeta “***processed***”.

Motivos de la exclusión de Outliers:

- **Distorsionan la media** de la variable principal.
- **Distorsionan su distribución** de frecuencias.
- Su exclusión **reduce** significativamente el **desvío**.

Impacto de la decisión:

- Mejora la **interpretabilidad** de la tendencia.
- **Se pierde información** sobre casos extremos que podría ser de utilidad.
- El analista debe considerar posibles desvíos respecto a la conclusión final producto de esta decisión.

Regresión y Test de Hipótesis

Se corre un **modelo de regresión lineal** entre las variables **Profit y Discount** para entender si se puede rechazar la siguiente hipótesis nula:

H0: *No existe relación estadísticamente significativa entre el nivel de descuento y la ganancia.*

```
> modelo_principal <- lm(profit ~ discount, data = datos_no_outliers)
> summary(modelo_principal)

Call:
lm(formula = profit ~ discount, data = datos_no_outliers)

Residuals:
    Min      1Q  Median      3Q     Max
-48.651 -10.167  -4.297   6.791  68.609

Coefficients:
            Estimate Std. Error t value            Pr(>|t|)
(Intercept)  17.6622     0.2338   75.55 <0.0000000000000002 ***
discount    -40.6803     0.9446  -43.07 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.82 on 8111 degrees of freedom
Multiple R-squared:  0.1861,    Adjusted R-squared:  0.186
F-statistic:  1855 on 1 and 8111 DF,  p-value: < 0.00000000000000022
```

Conclusiones Principales:

- Por cada 1 punto de descuento, la ganancia disminuye en -40.68 USD.

  (**10% de descuento implica -4USD de ganancia por pedido**)

- Dado el nivel de p < 2e-16, se concluye que el efecto es **estadísticamente sólido.** No es puro producto del azar.

- ***Se rechaza H0 y se concluye que el descuento afecta la ganancia del supermercado.***

- A partir del R2 se ve que el **descuento explica un 18% de la variabilidad de la ganancia.**

Verificación de supuestos

```
        Shapiro-Wilk normality test

data:  sample(residuals(modelo_principal), 500)
W = 0.93138, p-value = 0.00000000000002243
```

```
> # Homoscedasticidad
> bptest(modelo_principal)

        studentized Breusch-Pagan test

data:  modelo_principal
BP = 31.336, df = 1, p-value = 0.00000002171

>
> # Errores robustos si hay heterocedasticidad
> coeftest(modelo_principal, vcov = vcovHC(modelo_principal, type = "HC1"))

t test of coefficients:

             Estimate Std. Error t value             Pr(>|t|)
(Intercept)  17.66221    0.22799  77.469 < 0.00000000000000022 ***
discount    -40.68031    0.70194 -57.954 < 0.00000000000000022 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

1. ***Linealidad***:
   Se verifica una relación lineal razonable a partir del gráfico Residuals vs Fitted.

2. ***Normalidad***:
   Se corre el test **Shapiro Wilk**. Dado su pvalor, se rechaza H0. Sin embargo, dado el tamaño de la muestra **> 8000 observaciones**, no se compromete la validez del modelo.

3. ***Homocedasticidad***:
   Se corre el test de **Breusch-Pagan** cuyo pvalor indica **heterocedasticidad**. Se utilizan **errores robustos** para no invalidar el modelo.

4. ***Independencia***:
   Cada observación es una transacción diferente por lo que existe independencia en las variables.

Test de Independencia

Se estudia si la probabilidad de tener ganancia positiva depende a su vez de alguna variable categórica.

Para ello se corre un test chi-cuadrado para “**Category**”, “**Region**” y “**Segment**”.

```
> chisq_cat <- chisq.test(table(datos_no_outliers$category, datos_no_outliers$profit > 0))
> chisq_reg <- chisq.test(table(datos_no_outliers$region, datos_no_outliers$profit > 0))
> chisq_seg <- chisq.test(table(datos_no_outliers$segment, datos_no_outliers$profit > 0))
>
> resultados_chisq <- data.frame(
+   variable = c("Category", "Region", "Segment"),
+   estadistico = c(chisq_cat$statistic, chisq_reg$statistic, chisq_seg$statistic),
+   p_value = c(chisq_cat$p.value, chisq_reg$p.value, chisq_seg$p.value)
+ )
> print(resultados_chisq)
  variable estadistico
1 Category  207.061625
2   Region  292.434762
3  Segment    1.735309
                                                                             p_value
1 0.00000000000000000000000000000000000000000001089279585017473411887728
2 0.00000000000000000000000000000000000000000000000000000000000004315494
3 0.41993526223048405210391820219228975474834442138671875000000000000000
> 
```

*Tanto **Category** como **Region** rechazan H0 por lo que se afirma dependencia del “profit” sobre estas variables.*

Modelo Extendido incorporando variables categóricas

A partir del test de independencia se incluyen las variables “Category” y “Region” al modelo de regresión.

```
> summary(modelo_extendido)

Call:
lm(formula = profit ~ discount + category + region, data = datos_no_outliers)

Residuals:
    Min      1Q  Median      3Q     Max
-53.031  -9.880  -3.688   6.865  62.630

Coefficients:
                        Estimate Std. Error t value            Pr(>|t|)
(Intercept)              15.6209     0.5983  26.109 <0.0000000000000002 ***
discount                -39.7767     0.9581 -41.516 <0.0000000000000002 ***
categoryOffice Supplies   0.3425     0.4849   0.706              0.4799
categoryTechnology        8.2228     0.6315  13.021 <0.0000000000000002 ***
regionEast               -0.5639     0.5282  -1.068              0.2857
regionSouth               0.9761     0.6024   1.620              0.1052
regionWest                1.1502     0.5148   2.234              0.0255 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.55 on 8106 degrees of freedom
Multiple R-squared:  0.212,     Adjusted R-squared:  0.2114
F-statistic: 363.5 on 6 and 8106 DF,  p-value: < 0.00000000000000022
```

Conclusiones Finales:

- Cada unidad de descuento reduce el profit en ~40 USD.
- La categoría “**Technology**” y la región “**West**” generan mayores ganancias.
- El modelo explica un 21% de la variabilidad del profit.

  No es un modelo predictivo fuerte pero es útil para identificar factores significativos.

- El valor del estadístico F = 363.5 indica que el modelo tiene poder explicativo.

CONCLUSIONES FINALES

Hipótesis Inicial:

H0: No existe relación estadísticamente significativa entre el nivel de descuento y la ganancia.

H1: Los descuentos reducen significativamente la ganancia del negocio.

1. La hipótesis nula es rechazada por los dos modelos de regresión expuestos en este trabajo.
2. Ambos modelos evidencian una relación estadísticamente significativa entre las variables “Profit” y “Discount”.
3. Sus R2 indican que las variables seleccionadas pueden predecir un 18% y un 21% de la variabilidad de “Profit”.
4. Sin embargo, no se debe ignorar el hecho de que los outliers han sido excluidos en ambos modelos. Por tanto, el R2 podría verse afectado ante su inclusión.