Postulación al cargo de Data Scientist.
Prueba técnica.

Matías Rebolledo

Viernes, 4 de Agosto de 2017

Índice

1. Enunciado.

El objetivo es que propongas mejoras a nuestra metodología y proceso de estimación de la valorización de campañas para que se asemeje más a la realidad. Que identifiques puntos y variables relevantes a ajustar y/o predecir, que expliques que es lo que ves en los datos y específicamente qué caminos seguirías para realizar las mejoras (no necesariamente avanzar en estos caminos, pero puedes hacerlo en alguno que te interese o encuentras soluciones).

Además adjunto muestras de datos para que juegues. El archivo del banco estado, sabemos que a la fecha han invertido $120.000.000 millones en internet, pero nosotros calculamos $80.000.000. El segundo archivo corresponde a dos meses de todo el mercado de Chile. En Internet (iab, achap) puedes encontrar fuentes de inversión total y distribución para hacer supuestos para hacer las correcciones de Inversión total, inversión en escritorio vs móviles, display vs video, etc. Para las impresiones estamos usando similarweb.

El formato de entrega es un documento con gráficos y códigos asociados en el lenguaje que más te acomode.

La fecha de entrega es una semana vía mail, si te va bien agendamos para que vengas a presentarlo, lo que sería la entrevista / evaluación / final del proceso.

2. Preparar datos.

- Dependencias.

```
suppressMessages( suppressWarnings( source( "./src/Dependencias.R" ) ) )
options( knitr.kable.NA = '' )
greenseq <- brewer.pal(6, "BuGn")
```

- Cargar datos en sesión.
- Renombrar y formatear variables.
- Exportar a csv para inserción. Delimitador multi caracter.

```
# write.table( file = "./csv/banco.csv" , dfbanco , sep = "-_-" , quote = F , row.names = F )
# write.table( file = "./csv/markt.csv" , dfmarkt , sep = "-_-" , quote = F , row.names = F )
```

- Limpieza de datos. Vim + Bash.

```
# if ( Sys.info()["sysname"] == "Linux"   ) system( "bash ./src/LimpiarCsv.sh" )
# if ( Sys.info()["sysname"] == "Windows" )  shell( "bash ./src/LimpiarCsv.sh" )
```

- Separar tabla grande en meses.

```
# if ( Sys.info()["sysname"] == "Linux"   ) system( "bash ./src/SepararMeses.sh" )
# if ( Sys.info()["sysname"] == "Windows" )  shell( "bash ./src/SepararMeses.sh" )
```

- Crear tipo de datos.

```
if ( Sys.info()["sysname"] == "Linux"   ) system( "bash ./src/LlamarImportarDatos.sh" )
if ( Sys.info()["sysname"] == "Windows" )  shell( "bash ./src/LlamarImportarDatos.sh" )
```

- Abrir conexión a PostgreSQL.

```
source( "./src/Connect.R" )
```

- Crear inserciones de tablas SQL chicas.

```
cat ./src/CrearTabla.sh
cat ./src/InsertarTabla.sh
```

```
output:
output:   a01="${1}"
output:   a02="${2}"
output:   a03="${3}"
output:
output:   CrearTabla () {
output:
output:   #### CREAR TIPO DATOS {{{
output:   if [[ "${3}" == 1 ]] ; then
output:   cat "${1}" | head -n 1 | sed -r 's/-_-/\n/g' | \
output:         dos2unix | \
output:         sed -r 's/ /_/g;s/\(//g;s/\)//g' | \
output:         sed -r 's/(.*)/\L\1 varchar(300)/' | \
output:         sed -r 's/^/  /' | \
output:         sed -r '2,$s/^  /, /' | \
output:         sed -r '$s/$/ ) ;/' | \
output:         sed -r 's/date/dates/' | \
output:         sed -r 's/date varchar\(300\)/date varchar(10)/' | \
output:         sed -r 's/industry varchar\(300\)/industry varchar(17)/' | \
output:         sed -r 's/brand varchar\(300\)/brand varchar(12)/' | \
output:         sed -r 's/campaign_name varchar\(300\)/campaign_name varchar(57)/' | \
output:        sed -r 's/campaign_landing_page varchar\(300\)/campaign_landing_page varchar(89)/' | \
output:         sed -r 's/website varchar\(300\)/website varchar(26)/' | \
output:         sed -r 's/website_sections varchar\(300\)/website_sections varchar(270)/' | \
output:         sed -r 's/ad_type varchar\(300\)/ad_type varchar(7)/' | \
output:         sed -r 's/ad_size varchar\(300\)/ad_size varchar(87)/' | \
output:         sed -r 's/duration_video varchar\(300\)/duration_video varchar(8)/' | \
output:         sed -r 's/skip_video varchar\(300\)/skip_video varchar(17)/' | \
output:         sed -r 's/country varchar\(300\)/country varchar(5)/' | \
output:         sed -r 's/device varchar\(300\)/device varchar(7)/' | \
output:         sed -r 's/hosted_by varchar\(300\)/hosted_by varchar(41)/' | \
output:         sed -r 's/sold_by varchar\(300\)/sold_by varchar(14)/' | \
output:         sed -r 's/web_report varchar\(300\)/web_report varchar(474)/' | \
output:         sed -r 's/impact varchar\(300\)/impact int/' | \
output:         sed -r 's/impressions varchar\(300\)/impressions int/' | \
output:         sed -r 's/valuation varchar\(300\)/valuation int/' | \
output:         sed -r 's/web_report varchar\(300\)/web_report varchar(800)/' | \
output:         sed '1 i\\CREATE TABLE IF NOT EXISTS tb_banco (' | \
output:         sed '1 i\\DROP   TABLE IF     EXISTS tb_banco ;' | \
output:         sed '$ a\\TRUNCATE TABLE tb_banco ;' > \
output:         "${2}"
output:   fi
output:   if [[ "${3}" == 2 ]] ; then
output:   cat "${1}" | head -n 1 | sed -r 's/-_-/\n/g' | \
output:         dos2unix | \
output:         sed -r 's/ /_/g;s/\(//g;s/\)//g' | \
output:         sed -r 's/(.*)/\L\1 varchar(300)/' | \
output:         sed -r 's/^/  /' | \
output:         sed -r '2,$s/^  /, /' | \
```

```
output:         sed -r '$s/$/ ) ;/' | \
output:         sed -r 's/date/dates/' | \
output:         sed -r 's/date varchar\(300\)/date varchar(10)/' | \
output:         sed -r 's/industry varchar\(300\)/industry varchar(89)/' | \
output:         sed -r 's/brand varchar\(300\)/brand varchar(83)/' | \
output:         sed -r 's/campaign_name varchar\(300\)/campaign_name varchar(839)/' | \
output:        sed -r 's/campaign_landing_page varchar\(300\)/campaign_landing_page varchar(464)/' | \
output:         sed -r 's/website varchar\(300\)/website varchar(29)/' | \
output:         sed -r 's/website_section varchar\(300\)/website_section varchar(591)/' | \
output:         sed -r 's/ad_type varchar\(300\)/ad_type varchar(7)/' | \
output:         sed -r 's/ad_size varchar\(300\)/ad_size varchar(145)/' | \
output:         sed -r 's/duration_video varchar\(300\)/duration_video varchar(17)/' | \
output:         sed -r 's/skip_video varchar\(300\)/skip_video varchar(17)/' | \
output:         sed -r 's/advertisement varchar\(300\)/advertisement varchar(71)/' | \
output:         sed -r 's/screenshot varchar\(300\)/screenshot varchar(78)/' | \
output:         sed -r 's/country varchar\(300\)/country varchar(5)/' | \
output:         sed -r 's/device varchar\(300\)/device varchar(7)/' | \
output:         sed -r 's/hosted_by varchar\(300\)/hosted_by varchar(59)/' | \
output:         sed -r 's/sold_by_beta varchar\(300\)/sold_by_beta varchar(37)/' | \
output:         sed -r 's/impact varchar\(300\)/impact int/' | \
output:         sed -r 's/impressions varchar\(300\)/impressions int/' | \
output:         sed -r 's/valuation varchar\(300\)/valuation int/' | \
output:         sed -r 's/web_report varchar\(300\)/web_report varchar(800)/' | \
output:         sed '1 i\\CREATE TABLE IF NOT EXISTS tb_markt (' | \
output:         sed '1 i\\DROP   TABLE IF     EXISTS tb_markt ;' | \
output:         sed '$ a\\TRUNCATE TABLE tb_markt ;' > \
output:         "${2}"
output:   fi
output:   #### }}}
output:
output:   }
output:
output:   CrearTabla "${a01}" "${a02}" "${a03}"
output:   a01="${1}"
output:   a02="${2}"
output:   a03="${3}"
output:
output:   CrearInsercion () {
output:
output:   #### CREAR INSERCION DE DATOS {{{
output:   if [[ "${3}" == 1 ]] ; then
output:   cat "${1}" | gawk -F'-_-' 'NR>1 {print \
output:          "\047"$1"\047"  "," \
output:          "\047"$2"\047"  "," \
output:          "\047"$3"\047"  "," \
output:          "\047"$4"\047"  "," \
output:          "\047"$5"\047"  "," \
output:          "\047"$6"\047"  "," \
output:          "\047"$7"\047"  "," \
output:          "\047"$8"\047"  "," \
output:          "\047"$9"\047"  "," \
output:          "\047"$10"\047" "," \
output:          "\047"$11"\047" "," \
output:          "\047"$12"\047" "," \
output:          "\047"$13"\047" "," \
output:          "\047"$14"\047" "," \
output:          "\047"$15"\047" "," \
```

```
output:               $16       "," \
output:               $17       "," \
output:               $18       "," \
output:          "\047"$19"\047"     \
output:          }' | \
output:         sed -r 's/^/  (/' | \
output:         sed -r 's/$/)/' | \
output:         sed -r '2,$s/^  /, /' | \
output:         sed -r '$s/$/ ;/' | \
output:         sed '1 i\\INSERT INTO tb_banco VALUES' > \
output:         "${2}"
output:   fi
output:   if [[ "${3}" == 2 ]] ; then
output:   cat "${1}" | gawk -F'-_-' 'NR>1 {print \
output:          "\047"$1"\047"  "," \
output:          "\047"$2"\047"  "," \
output:          "\047"$3"\047"  "," \
output:          "\047"$4"\047"  "," \
output:          "\047"$5"\047"  "," \
output:          "\047"$6"\047"  "," \
output:          "\047"$7"\047"  "," \
output:          "\047"$8"\047"  "," \
output:          "\047"$9"\047"  "," \
output:          "\047"$10"\047" "," \
output:          "\047"$11"\047" "," \
output:          "\047"$12"\047" "," \
output:          "\047"$13"\047" "," \
output:          "\047"$14"\047" "," \
output:          "\047"$15"\047" "," \
output:          "\047"$16"\047" "," \
output:          "\047"$17"\047" "," \
output:               $18       "," \
output:               $19       "," \
output:               $20           \
output:          }' | \
output:         sed -r 's/^/  (/' | \
output:         sed -r 's/$/)/' | \
output:         sed -r '2,$s/^  /, /' | \
output:         sed -r '$s/$/ ;/' | \
output:         sed '1 i\\INSERT INTO tb_markt VALUES' > \
output:         "${2}"
output:   fi
output:   #### }}}
output:
output:   }
output:
output:   CrearInsercion "${a01}" "${a02}" "${a03}"
```

- Ejecutar inserciones de tablas SQL chicas. Tablas de la industria usan mucha memoria, incluso separando por meses.

```
cat ./src/LlamarImportarDatos.sh
```

```
output:   bash ./src/CrearTabla.sh    ./csv/banco.csv ./sql/crear_tabla_banco.sql 1
output:   bash ./src/CrearTabla.sh    ./csv/markt.csv ./sql/crear_tabla_markt.sql 2
output:   bash ./src/InsertarTabla.sh ./csv/banco.csv ./sql/insertar_banco.sql    1
output: # bash ./src/InsertarTabla.sh ./csv/markt.csv ./sql/insertar_markt.sql    2
output: # bash ./src/InsertarTabla.sh ./csv/markt_2017_05.csv ./sql/insertar_markt_2017_05.sql 2
output: # bash ./src/InsertarTabla.sh ./csv/markt_2017_06.csv ./sql/insertar_markt_2017_06.sql 2
output: # bash ./src/InsertarTabla.sh ./csv/markt_2017_07.csv ./sql/insertar_markt_2017_07.sql 2
```

- Crear inserciones de tablas SQL grandes separadas por día para bajar el uso de memoria.

```
cat ./src/SepararDias.sh
```

```
output:
output:   a01="${1}"
output:   a02="${2}"
output:
output:   InsertarFrecuenciaDiaria () {
output:
output: #### SEPARAR {{{
output:   if [[ "${2}" == 1 ]] ; then
output:     gawk -F'-_-' 'NR>1 {print FILENAME, $1, $1}' "${1}" | sort | uniq | sed -r 's/-/_/;s/-/_/' | \
output:     sed -r 's/(.*)\.(.*) (.*) (.*)/gawk -F'\''-_-'\'' '\''NR==1 {print} NR>1 \&\& \$1 \~ \/\4\/ {print}'\'' \1\.\2 > \1_\3\.\2/'
output:   fi
output: #### }}}
output:
output: #### CREAR INSERCIONES {{{
output:   if [[ "${2}" == 2 ]] ; then
output:     gawk -F'-_-' 'NR>1 {print FILENAME, $1}' "${1}" | sort | uniq | sed -r 's/-/_/g' | \
output:     sed -r 's/(.*)\/(.*)\.(.*) (.*)/bash \.\/src\/InsertarTabla.sh \1\/\2_\4\.\3 \.\/sql\/insertar_\2_\4\.sql 2/'
output:   fi
output: #### }}}
output:
output: #### VERIFICAR LONGITUD {{{
output:   if [[ "${2}" == 3 ]] ; then
output:     gawk -F'-_-' 'NR>1 {print FILENAME, $1}' "${1}" | sort | uniq | sed -r 's/-/_/g' | \
output:     sed -r 's/(.*)\.(.*) (.*)/gawk -F'\''-_-'\'' '\''length > m { m = length; a = NR } END { print a }'\'' \1_\3\.\2/'
output:   fi
output: #### }}}
output:
output: #### EJECUTAR INSERCIONES {{{
output:   if [[ "${2}" == 4 ]] ; then
output:     gawk -F'-_-' 'NR>1 {print FILENAME, $1}' "${1}" | sort | uniq | sed -r 's/-/_/g' | \
output:     sed -r 's/(.*)\/(.*)\.(.*) (.*)/psql -U postgres -h localhost -d dbadmetricks -f \.\/sql\/insertar_\2_\4\.sql/'
output:   fi
output: #### }}}
output:
output:   }
output:
output:   InsertarFrecuenciaDiaria "${a01}" "${a02}"
```

- Ejecutar inserciones de tablas SQL grandes.
  - Las anteriores son rutinas cuyo objetivo es la mejora continua, detección de inconsistencias y creación de flujos de procesamiento de datos.
  - Se busca incorporarlas dentro de un *workflow system* como Tensorflow o similar para tener un frontend con nodos gráficos además de las herramientas de generación de código.

```
cat ./src/LlamarSepararDias.sh
```

```
output: # bash ./src/SepararDias.sh ./csv/markt.csv 1 | parallel
output: # bash ./src/SepararDias.sh ./csv/markt.csv 2 | parallel
output:   psql -U postgres -h localhost -d dbadmetricks -f ./sql/crear_tabla_banco.sql
output:   psql -U postgres -h localhost -d dbadmetricks -f ./sql/crear_tabla_markt.sql
output:   psql -U postgres -h localhost -d dbadmetricks -f ./sql/insertar_banco.sql
output: # bash ./src/SepararDias.sh ./csv/markt.csv 3 | head -n 4 | bash
output:   bash ./src/SepararDias.sh ./csv/markt.csv 4 | bash
```

- Verificar por consulta la inversión total.

```
sumbanco <- dbGetQuery( con , "select brand , sum(valuation) as suma from tb_banco
                               group by brand
                               order by suma desc" )
summarkt <- dbGetQuery( con , "select brand , sum(valuation) as suma from tb_markt
                               group by brand
                               order by suma desc" )
summarkt <- within( summarkt   , quartile <- as.integer( cut( suma       ,
            quantile( suma     , probs = 0:100/100 ) , include.lowest = TRUE ) ) )
kable( sumbanco ,
format.args = list( decimal.mark = ',' , big.mark = '.' ) , digits = 0 ,
      col.names = c( "Empresa" , "Inversión" ) ,
      caption = "Inversión total de Banco Estado." )
```

Cuadro 1: Inversión total de Banco Estado.

Empresa	Inversión
banco estado	80.876.828

```
kable( head( summarkt , 20 ) ,
format.args = list( decimal.mark = ',' , big.mark = '.' ) , digits = 0 ,
      col.names = c( "Empresa" , "Inversión" , "Percentil" ) ,
      caption = "Inversión total de la Industria y percentiles de cada inversión." )
```

Cuadro 2: Inversión total de la Industria y percentiles de cada inversión.

Empresa	Inversión	Percentil
amazon prime video	972.578.379	100
paris	647.064.484	100
netflix	617.709.716	100
nescafé	522.615.446	100
movistar	379.389.663	100
claro	359.372.656	100
kia	340.283.332	100
falabella	331.683.607	100
fanta	319.595.828	100
ripley	291.442.187	100
unimarc	289.471.012	100
latam airlines	258.160.709	100
banco santander	245.085.630	100
jumbo	240.120.754	100
cmr falabella	225.095.795	100
gmo	223.230.075	100
punto ticket	218.833.777	100
universidad adolfo ibañez uai	217.147.399	100
banco de crédito e inversiones bci	214.128.765	100
soprole	211.055.656	100

- Verificar la inversión mensual.

```
smebanco <- dbGetQuery( con , "select brand , date_part( 'month' , dates::date )
                               as mes , sum(valuation) as suma from tb_banco
                               group by brand , mes
                               order by sum(valuation) desc" )
smemarkt <- dbGetQuery( con , "select brand , date_part( 'month' , dates::date )
                               as mes , sum(valuation) as suma from tb_markt
                               group by brand , mes
                               order by sum(valuation) desc" )
kable( smebanco ,
format.args = list( decimal.mark = ',' , big.mark = '.' ) , digits = 0 ,
      col.names = c( "Empresa" , "Mes" , "Inversión" ) ,
      caption = "Inversión total de la Industria a nivel mensual." )
```

Cuadro 3: Inversión total de la Industria a nivel mensual.

Empresa	Mes	Inversión
banco estado	6	24.457.733
banco estado	2	22.469.188
banco estado	1	13.452.478
banco estado	5	11.637.258
banco estado	7	4.681.761
banco estado	3	3.922.817
banco estado	4	255.593

```
kable( head( smemarkt , 20 ) ,
format.args = list( decimal.mark = ',' , big.mark = '.' ) , digits = 0 ,
      col.names = c( "Empresa" , "Mes" , "Inversión" ) ,
      caption = "Inversión total de la Industria a nivel mensual." )
```

Cuadro 4: Inversión total de la Industria a nivel mensual.

Empresa	Mes	Inversión
amazon prime video	5	523.382.105
nescafé	6	507.370.452
netflix	6	467.879.744
amazon prime video	6	448.648.081
paris	5	402.124.070
kia	6	283.128.769
movistar	6	282.332.229
fanta	5	274.161.093
paris	6	241.364.418
falabella	6	216.206.698
claro	6	197.906.461
gmo	5	190.762.195
disney cine	5	188.981.641
soprole	5	185.206.160
latam airlines	6	184.974.518
punto ticket	6	169.390.202
ripley	5	158.814.394
unimarc	5	156.684.964
banco santander	6	152.689.493
jumbo	6	150.805.200

3. Datos de Banco Estado.

3.1. Variación de inversiones mensuales de Banco Estado.

Cuadro 5: Variación porcentual mensual de impacto, impresiones y valorización.

Mes	Impacto	Variación	Impresiones	Variación	Valorización	Variación
2017-01	543.920		8.019.911		13.452.478	
2017-02	2.836.195	80,82	46.014.992	82,57	22.469.188	40,13
2017-03	478.095	-493,23	8.830.615	-421,08	3.922.817	-472,78
2017-04	42.803	-1.016,97	692.886	-1.174,47	255.593	-1.434,79
2017-05	1.678.844	97,45	12.160.565	94,30	11.637.258	97,80
2017-06	1.106.749	-51,69	10.779.196	-12,82	24.457.733	52,42
2017-07	4.065.698	72,78	6.617.648	-62,89	4.681.761	-422,40

- Una característica común es que las tres variables se comportan similar en cuanto a su variación porcentual.
- Esto no ocurre con el valor observado original de cada una de ellas.
- En ese caso, las impresiones tienen un salto abrupto en el mes de Febrero. Este comportamiento parece ser heredado por la variable valorización.
- La variable impacto se atenúa y aumenta solamente en el último mes, mientras que las dos restantes tienden a mostrar un efecto estacional.
- Esto se ve en el gráfico de evolución mensual más adelante en el que las variables impresiones y valorizaciones aumentan en Febrero y luego en Junio como evidencia de estacionalidad. Es la variable impresiones la que aumenta más abruptamente en Febrero.

3.2. Contabilizar outliers.

Cuadro 6: Outliers para variables impacto, impresiones y valoraciones totales.

	Impacto	Impresiones	Valorización
outliers	15	9	13

Cuadro 7: Outliers para variables impacto, impresiones y valoraciones mensuales.

	Impacto	Impresiones	Valorización
2017-01	3	8	6
2017-02	2	3	5
2017-03	1	2	7
2017-04	4	2	3
2017-05	5	1	4
2017-06	1	2	2
2017-07	4	5	1

- La regla de cálculo usada es de 3 veces la desviación estándar. Corresponde a una variante de la regla comúnmente usada que es un poco más restrictiva. El ancho del intervalo se puede ampliar a modo de aislar los valores atípicos más extremos si no se desea perder mucha información.
- Los valores contabilizados corresponden a valores que están en el extremo superior de cada muestra.
- El número de outliers depende de la muestra desde donde se calculen. Por eso que a nivel total el número de valores fuera de rango es distinto que a nivel mensual.

- Esta misma información se vará a continuación en los gráficos de caja, tanto a nivel total como mensual.

3.3. Estadígrafos de las variables de interés a nivel mensual.

Cuadro 8: Estadígrafos generales para variables numéricas.

	nobs	NAs	Minimum	Maximum	1. Quartile	3. Quartile	Mean	Median	Sum	SE Mean	LCL Mean	UCL Mean	Variance	Stdev	Skewness	Kurtosis
impact	773	0	2	899.710	52	1.155	13.909,84	194	10.752.304	2.362,91	9.271,34	18.548,33	4.315.929.184	65.695,73	8,03	77,74
impressions	773	0	31	8.832.279	568	6.534	120.460,30	1.728	93.115.813	21.892,39	77.484,62	163.435,98	370.481.033.534	608.671,53	9,66	108,72
valuation	773	0	10	8.035.346	535	11.630	104.627,20	3.578	80.876.828	17.116,42	71.026,97	138.227,44	226.467.106.536	475.885,60	9,51	121,36

Cuadro 9: Estadígrafos mensuales para la variable impacto.

month_yr	nobs	NAs	Minimum	Maximum	X1..Quartile	X3..Quartile	Mean	Median	Sum	SE.Mean	LCL.Mean	UCL.Mean	Variance	Stdev	Skewness	Kurtosis
2017-01	116	0	2	46.590	69,00	4.906,25	4.688,97	236,0	543.920	861,64	2.982,22	6.395,71	86.121.540	9.280,17	2,56	6,53
2017-02	216	0	3	414.462	46,00	2.173,50	13.130,53	174,5	2.836.195	3.614,21	6.006,72	20.254,35	2.821.499.370	53.117,79	5,18	28,20
2017-03	148	0	7	46.795	58,75	977,50	3.230,37	193,5	478.095	687,52	1.871,67	4.589,07	69.956.972	8.364,03	3,27	10,71
2017-04	32	0	20	22.241	54,50	115,50	1.337,59	76,0	42.803	876,73	-450,51	3.125,70	24.597.043	4.959,54	3,51	10,87
2017-05	133	0	6	321.861	37,00	810,00	12.622,89	133,0	1.678.844	3.723,24	5.257,96	19.987,82	1.843.711.096	42.938,46	4,74	26,05
2017-06	97	0	4	162.816	63,00	828,00	11.409,78	364,0	1.106.749	3.457,25	4.547,19	18.272,38	1.159.402.841	34.050,01	3,15	9,04
2017-07	31	0	390	899.710	832,00	77.092,00	131.151,55	1.630,0	4.065.698	45.089,87	39.065,74	223.237,36	63.026.000.023	251.049,80	1,67	1,46

Cuadro 10: Estadígrafos mensuales para la variable impresiones.

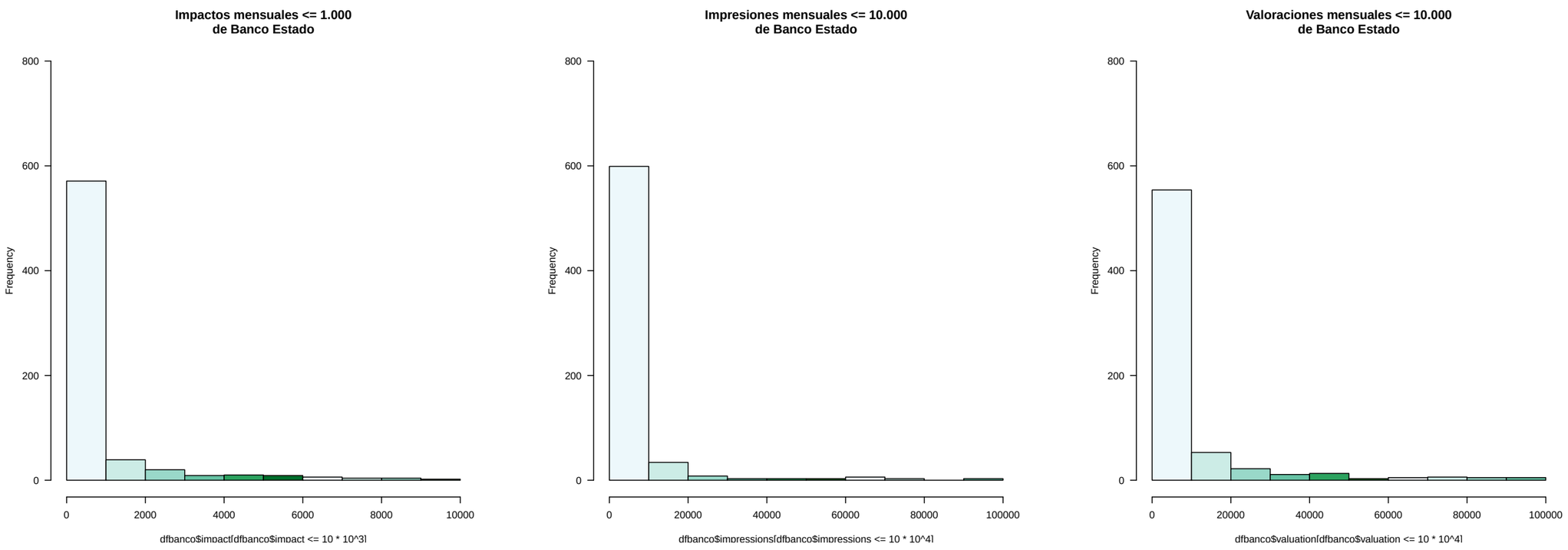
month_yr	nobs	NAs	Minimum	Maximum	X1..Quartile	X3..Quartile	Mean	Median	Sum	SE.Mean	LCL.Mean	UCL.Mean	Variance	Stdev	Skewness	Kurtosis
2017-01	116	0	31	513.456	521,00	128.192,0	69.137,16	1.412	8.019.911	10.823,55	47.697,80	90.576,52	13.589.300.115	116.573,15	1,75	2,46
2017-02	216	0	43	8.832.279	574,50	6.281,5	213.032,37	1.807	46.014.992	73.006,67	69.131,91	356.932,83	1.151.274.373.059	1.072.974,54	5,91	36,12
2017-03	148	0	94	1.355.090	779,25	8.158,5	59.666,32	2.088	8.830.615	15.323,69	29.383,13	89.949,51	34.752.697.757	186.420,75	4,45	22,04
2017-04	32	0	277	363.525	751,25	1.588,5	21.652,69	1.032	692.886	14.339,48	-7.592,87	50.898,24	6.579.859.290	81.116,33	3,51	10,87
2017-05	133	0	76	2.038.416	399,00	5.960,0	91.432,82	1.326	12.160.565	24.691,50	42.590,59	140.275,04	81.086.123.649	284.756,25	3,94	18,38
2017-06	97	0	39	1.708.070	430,00	6.031,0	111.125,73	1.356	10.779.196	35.285,39	41.084,79	181.166,67	120.770.670.626	347.520,75	3,33	10,36
2017-07	31	0	2.655	2.068.978	6.010,00	308.687,0	213.472,52	12.020	6.617.648	79.716,74	50.669,21	376.275,82	196.997.518.567	443.844,03	2,70	7,62

Cuadro 11: Estadígrafos mensuales para la variable valorizaciones.

month_yr	nobs	NAs	Minimum	Maximum	X1..Quartile	X3..Quartile	Mean	Median	Sum	SE.Mean	LCL.Mean	UCL.Mean	Variance	Stdev	Skewness	Kurtosis
2017-01	116	0	10	1.221.798	1.839,00	94.835,00	115.969,64	5.033	13.452.478	22.416,01	71.567,84	160.371,44	58.287.378.649	241.427,79	2,66	6,99
2017-02	216	0	13	8.035.346	332,75	16.303,25	104.024,02	3.062	22.469.188	40.078,16	25.027,61	183.020,43	346.951.838.311	589.026,18	11,56	149,79
2017-03	148	0	29	623.080	581,00	20.873,50	26.505,52	4.626	3.922.817	5.453,51	15.728,11	37.282,93	4.401.639.196	66.344,85	5,75	43,67
2017-04	32	0	987	72.705	2.680,50	5.661,75	7.987,28	3.678	255.593	2.725,11	2.429,38	13.545,18	237.639.052	15.415,55	3,41	10,44
2017-05	133	0	32	2.400.999	231,00	5.318,00	87.498,18	1.104	11.637.258	30.010,79	28.133,87	146.862,49	119.786.106.126	346.101,29	5,17	28,12
2017-06	97	0	36	4.270.176	443,00	5.348,00	252.141,58	3.360	24.457.733	87.037,35	79.373,80	424.909,35	734.823.531.842	857.218,49	3,52	11,45
2017-07	31	0	2.340	1.521.190	5.333,50	209.710,50	151.024,55	10.660	4.681.761	57.660,97	33.265,13	268.783,97	103.068.426.288	321.042,72	2,85	8,45

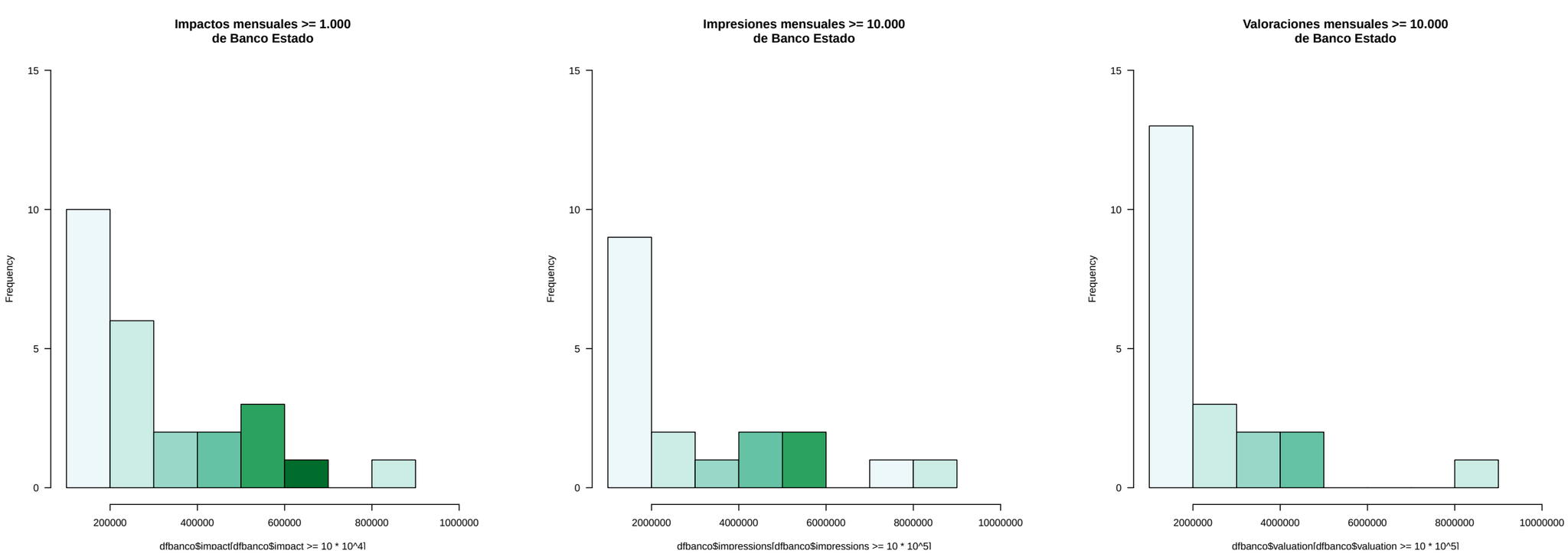
```
## pdf
##   2
```

Figura 1: Histograma del extremo izquierdo de la distribución de, impacto, impresiones y valorizaciones.




```
## pdf
##   2
```

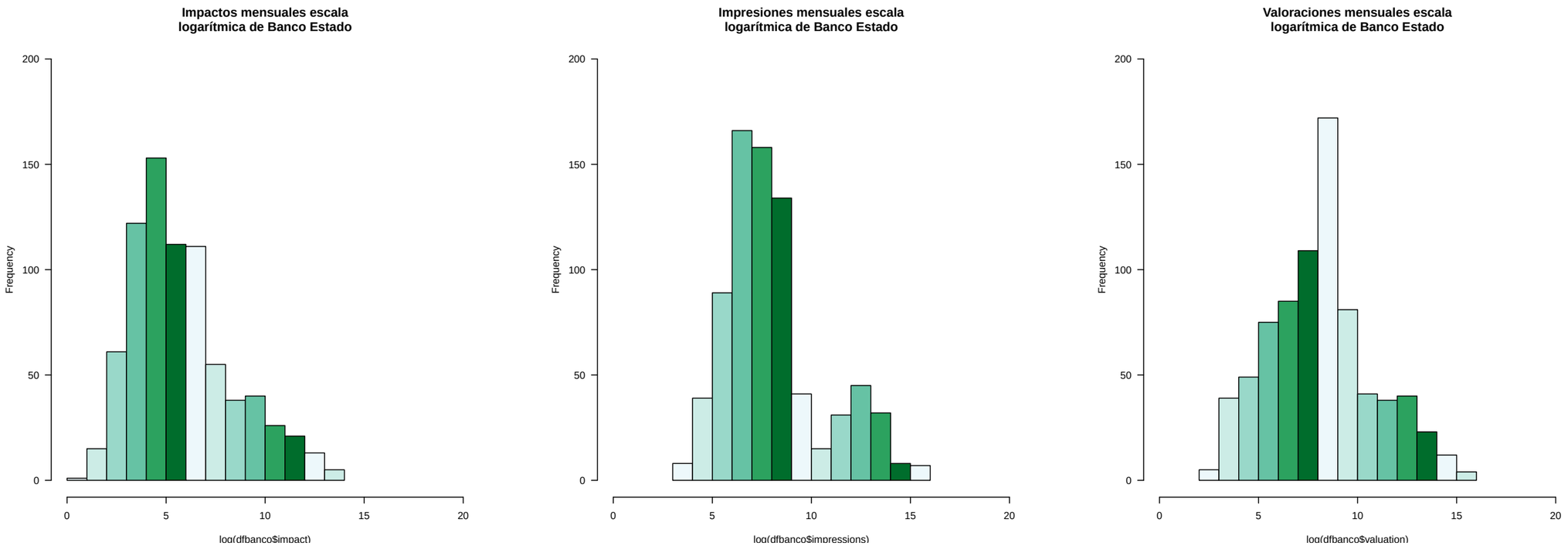
Figura 2: Histograma del extremo derecho de la distribución de, impacto, impresiones y valorizaciones.




```
## pdf
##   2
```

- El histograma de cola izquierda y derecha muestran que la distribución está cargada hacia la derecha en las tres variables. Esto se llama *right skewed* en inglés y significa que son asimétricas hacia la derecha.

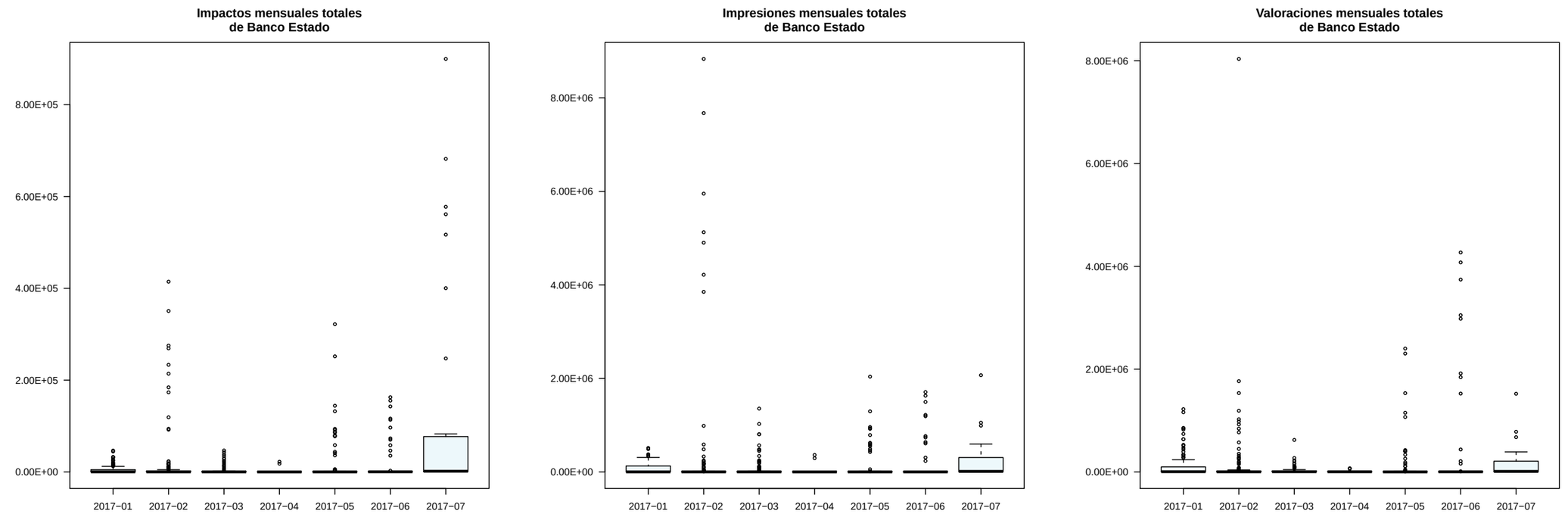
Figura 3: Histograma en escala logarítmica de la distribución de, impacto, impresiones y valorizaciones.



- En el primer histograma, tenemos que en el extremo izquierdo se concentran la mayor cantidad de observaciones de la muestra para las tres variables, llegando a un número por sobre 600 observaciones.
- En el segundo histograma, tenemos que el extremo derecho concentra menos de 20 observaciones. Es en este extremo donde se registran valores de impactos, impresiones y valorizaciones más elevados.
- Por un lado, el extremo derecho, si bien es cierto es menos concentrado en valores muestreados, solo se encuentra una menor cantidad de observaciones de la muestra en este lado de la distribución.
- Esto implica que habrá una mayor carga de trabajo en análisis en los cálculos más pequeños en los muestreos rutinarios, porque es en ese extremo donde hay un mayor nivel de incertidumbre.
- El tercer histograma, muestra una representación normal logarítmica para cada muestra. Es una representación que se usa con regularidad para muestras que son fuertemente sesgadas hacia la izquierda (con asimetría derecha).
- Se puede ver en este gráfico que la variable impresiones tiene forma bimodal. La bimodalidad según lo visto en este análsis, puede tener origen en que las campañas son realizadas en periodos específicos del año, lo que explicaría una suerte de estacionalidad. Esto sería como evidencia de un comportamiento con dos fenómenos distintos dentro de un periodo de 7 meses. Posiblemente se trate de un inicio de año con una fuerte campaña en publicidad, lo que no se alcanza a repetir hasta el mes de Julio, en lo que se vería un proceso diferente al de los primeros 2 meses.
- Para efectos de modelización, usar una distribución unimodal tiene menos incertidumbre que una bimodal. Es decir, si se quiere explicar el comportamiento de una variable con fines de optimizar costos de impresiones, como el CPM, la estacionalidad puede ser una fuente de incertidumbre que debiera ser modelada.
- Los supuestos son más sencillos para un modelo cuya distribución es unimodal, mientras que para una bimodal son más desafiantes y menos usados en la práctica.

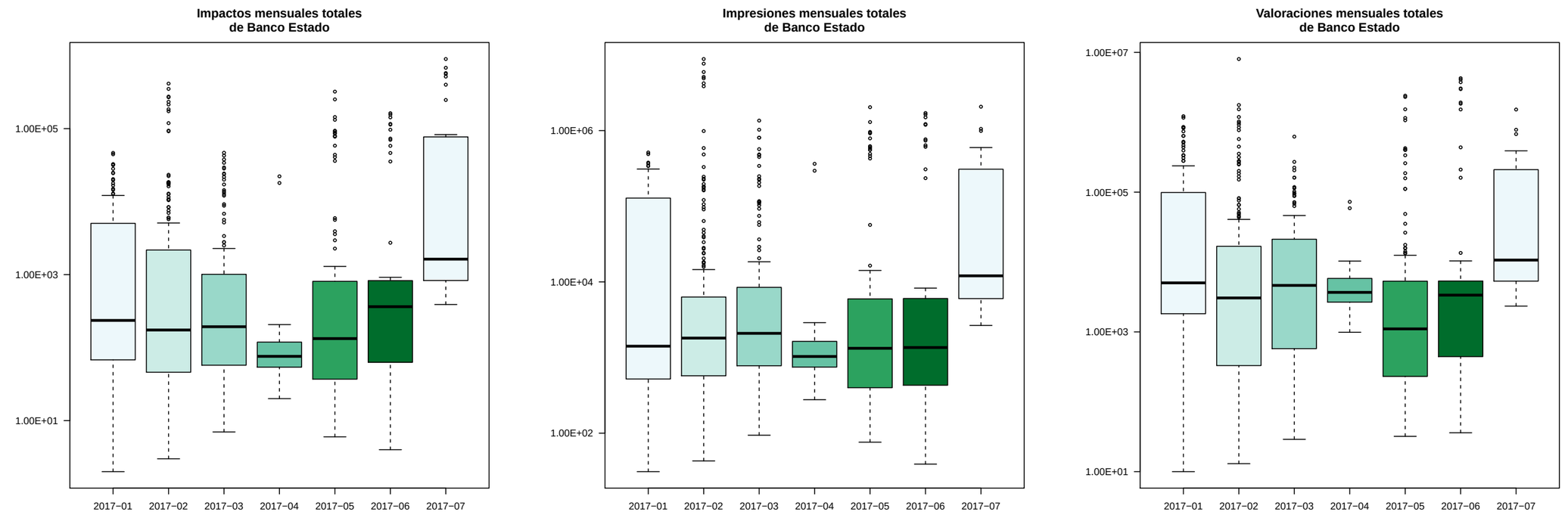
```
## pdf
##   2
```

Figura 4: Gráfico de cajas de evolución mensual de impacto, impresiones y valorizaciones.




```
## pdf
##   2
```

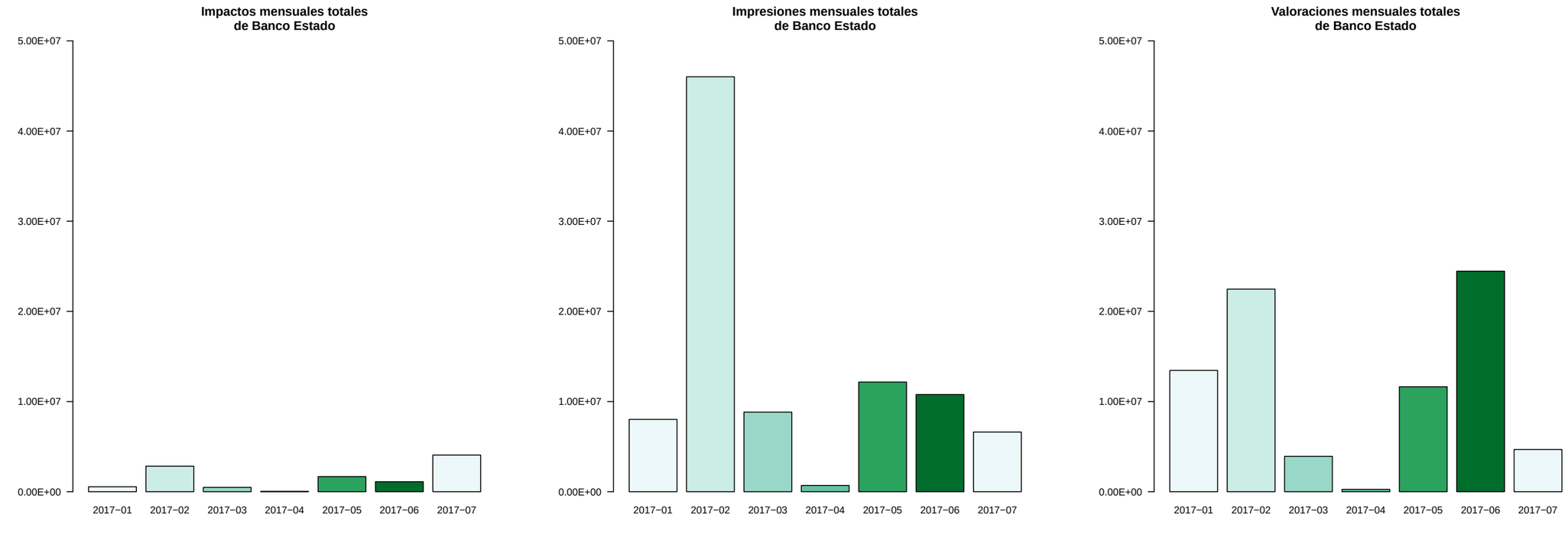
Figura 5: Gráfico de cajas en escala logarítmica de evolución mensual de impacto, impresiones y valorizaciones.



- Los gráficos de caja o *boxplots* muestran un gran número de *outliers* en cada muestra. Los *outliers* son los puntos sobre cada caja en cada gráfico. Se han graficado tanto los valores originales como sus respectivas transformaciones logarítmicas para mejorar la visualización dada la fuerte asimetría de cada distribución.
- Existe un patrón diferente en los primeros 3 meses versus los 3 siguientes. Esto es similar a lo visto en el cuadro de evolución mensual. El mes de Julio en cambio no muestra un comportamiento similares a los 6 primeros meses.
- Las cajas muestran el rango de dispersión que hay entre los percentiles 25 y 75. La variable impresiones tiene una mayor dispersión en el mes de Enero. Mientras que las valoraciones tienen una mayor dispersión en Febrero y Marzo.
- Los *outliers* contabilizados por medio de los gráficos usan una fórmula más astringente que la presentada en los cuadros iniciales. En este caso, se trata de 1,5 veces por el rango intercuartilico. Por ser un intervalo más pequeño que el anterior contabiliza más *outliers*.

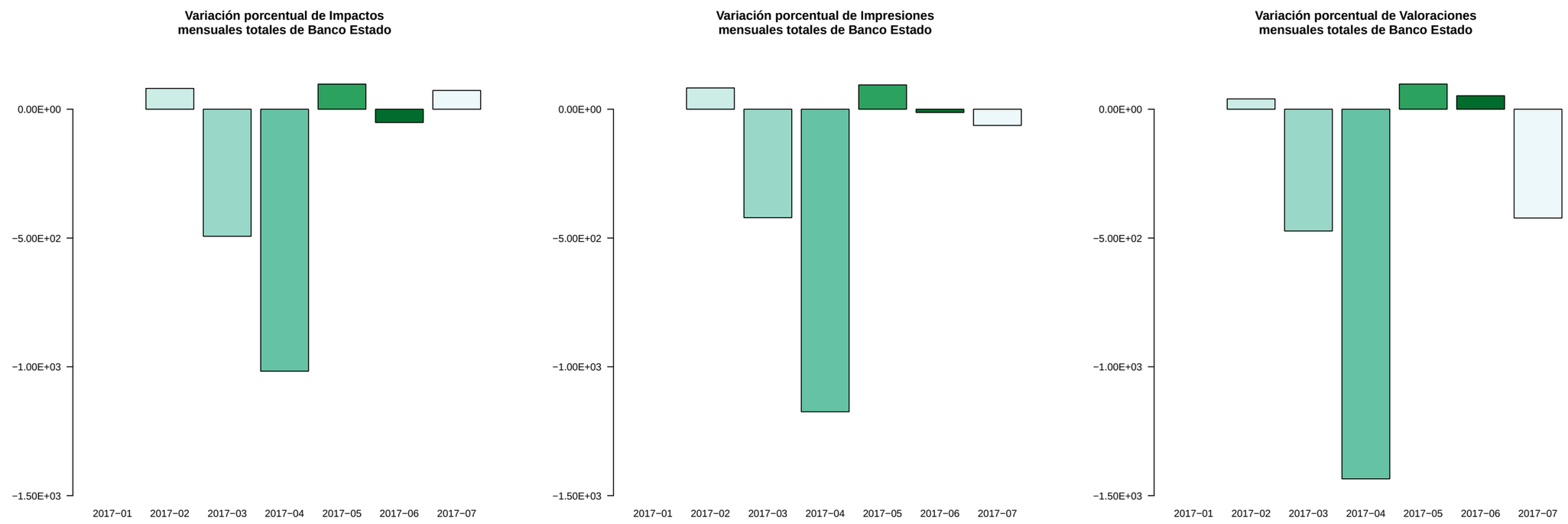
```
## pdf
##   2
```

Figura 6: Gráfico de barras de evolución mensual de impacto, impresiones y valorizaciones.




```
## pdf
##   2
```

Figura 7: Gráfico de barras de variación porcentual mensual de impacto, impresiones y valorizaciones.



- Los gráficos de barras evidencian que el muestreo fue mucho mayor en el mes de Febrero, mientras que los meses siguientes se ve un una disminución notoria. Sería necesario contrastar esta evidencia con empresas de la misma industria para verificar patrones similares. Poder observar más de dos meses permite tener una idea de posible comportamiento en estaciones, lo cual se puede modelizar dentro de alguna técnica de machine learning.
- Las variaciones porcentuales tienden a acentuarse para el caso de Marzo y Abril, dado por la fuerte caida que se ve en los valores originales de cada variable. Esta es otra representación que evidenciaría un esquema estacional.
- Estos gráficos señalan que las tres variables heredan un comportamiento similar. Debido a la metodología de cálculo estas variables deben estar correlacionadas. Para efectos de modelización, bastaría con usar alguna de ellas dado que al usarlas todas se estaría duplicando información. No se puede invertir la matriz de información con variables que comparten un mismo cálculo, lo que es necesario para un modelo logístico o lineal múltiple. A esto se le llama singularidad o colinealidad y es sencillo corroborarlo.
- Con fines de mejora continua es importante detectar valores atípicos en toda variable involucrada en el proceso de muestreo y estimaciones totales.
- Por otro lado, las impresiones son elevadas en comparación con las otras variables. Estas son estimaciones de SimilarWeb, lo que indica que habría que identificar cómo mejorar el proceso de mejora de estimación total de valorización. Específicamente midiendo la incertidumbre que genera la incorporación de la información proveniente de los *pageviews* de SimilarWeb. Posiblemente penalizar, sobre la estimación, los elevados valores de impresiones o usar un símil de factor de expansión que represente un segmento de la muestra que pondera como más elevado que otros en el mismo periodo y estratificarlo según número de impresiones.

3.4. Variación de inversiones mensuales de Banco Estado según campaña.

Cuadro 12: Evolución mensual de inversión según campaña.

Campaña	2017-01	2017-02	2017-03	2017-04	2017-05	2017-06	2017-07	Total	Variación
BancoEstado							1.107.932	1.107.932	100,00
BancoEstado - Home | Facebook			153.561					153.561	100,00
BancoEstado Microempresas | Premio Emprendedor							43.142	43.142	100,00
BancoEstado Microempresas | Tu negocio te pide crecer					69.903	69.693		139.596	100,00
BancoEstado Personas | 30 % de descuento en Cinemark					512.102			512.102	100,00
BancoEstado Personas | Ahorro e Inversiones							3.402.293	3.402.293	100,00
BancoEstado Personas | Cotiza tu Hipotecario con nosotros					7.112.996	7.921.864		15.034.860	100,00
BancoEstado Personas | Créditos			4.848					4.848	100,00
BancoEstado Personas | Descuentos en LibrerÃa Nacional		606	174.083					174.689	99,65
BancoEstado Personas | Home	135.448	170.083	186.084	123.972	71.693	90.959	128.394	906.633	49,17
BancoEstado Personas | Tarjeta Visa Chilena					3.203.648	16.375.217		19.578.865	100,00
BancoEstado | Simulador Crédito Consumo			1.477.370					1.477.370	100,00
En Fácil y en Chileno - YouTube		8.177.246						8.177.246	
Programa En Fácil y en Chileno	12.181.529							12.181.529	
SOAP 2017 - Zenit Seguros		67.286	1.347.417					1.414.703	95,01
TodoSuma BancoEstado todosuma	1.135.501							1.135.501	

- El cuadro muestra que el canal publicitario que permanece constante en inversiones, es el Home de Banco Estado.
- Por otro lado, hay otros canales donde hay estimaciones fuertes, tales como, el programa “En fácil y en chileno” tanto por youtube como por el mismo sitio de la campaña, con 12 y 8 millones, en Enero y Febrero, respectivamente.
- Estos dos últimos generan incertidumbre en dichos periodos, al igual que, los canales de “Ahorro e inversiones”, “Crédito hipotecario” y “Tarjeta visa”, entre los meses de Mayo y Julio con altos valores en estimaciones.
- Como aclaración, la incertidumbre es lo que se modeliza en a través de un modelo estadístico o de machine learning. Es un concepto que se utiliza aquí solamente con fines técnicos, cuyo potencial es la mejora mediante la compresión de ese comportamiento usando una herramienta estadística, tales como, regresión logística, árboles de clasificación, redes neuronales, etc.
- La modelización tendría dos necesidades fuertes. Por un lado, medir el comportamiento mensual de variables relevantes, en busca de patrones comunes en el tiempo y definir estrategias de mejora en esa dirección. Por otro lado, mejora en el proceso de muestreo. Habría que hacer supuestos en cuanto a qué nivel de estimaciones se obtendrían en periodos y canales determinados, considerando la alta asimetría en las muestras, la periodicidad del levantamiento, estimaciones localizadas y extremas versus estimaciones más pequeñas y constantes, entre otras estrategias.

3.5. Variación de inversiones mensuales de Banco Estado según website.

Cuadro 13: Evolución mensual de inversión según website.

	Website	2017-01	2017-02	2017-03	2017-04	2017-05	2017-06	2017-07	Total	Variación
141	youtube.com	13.140.652		623.080		8.457.244	23.405.825		45.626.801	59,55
56	facebook.com		8.330.016	1.584.169	131.621	2.800.254	808.638	4.510.225	18.164.923	15,30
76	lun.com		8.035.346						8.035.346	
120	soychile.cl		1.891.253	40.055					1.931.308	-4.621,64
20	biobiochile.cl		1.248.480						1.248.480	
3	adnradio.cl		1.094.540	38.454					1.132.994	-2.746,36
71	lared.cl			740.211					740.211	100,00
6	amarillas.cl	135.448	170.083	186.084	123.972	71.693			687.280	19,97
68	lacuarta.com		650.389						650.389	
103	propymechile.com					69.903	160.652	171.536	402.091	100,00
139	yapo.cl			299.503					299.503	100,00
75	los40.cl		285.983						285.983	
54	emol.com			244.702					244.702	100,00
31	corazon.cl		198.319	9.046					207.365	-2.092,34
132	upsocl.com		92.144	52.455					144.599	-75,66
66	juegos.com		92.034	31.304		1.068	3.342		127.748	-157,70
5	ahoranoticias.cl	40.706	28.055			2.897	35.061		106.719	-81,15
79	mega.cl	70.802	18.910						89.712	
13	australvaldivia.cl		80.849	3.456					84.305	-2.239,38
58	fmdos.cl		78.967						78.967	
27	chilevision.cl	2.366	47.328	888					50.582	-5.496,17
138	wordreference.com					48.730			48.730	100,00
23	carolina.cl					41.964			41.964	100,00
45	elmostrador.cl	20.655	4.140			13.302	3.157		41.254	-50,65
11	as.com		23.922						23.922	

- Este cuadro muestra los primeros 20 sitios web con mayor nivel de inversiones estimado. Los sitios “youtube”, “Facebook” y “lun.com”, llevan la delantera en el mes de Febrero, lo que ya se ha ha visto en visualizaciones anteriores.
- El sitio Youtube vuelve a aumentar en el mes de Mayo y Junio, lo que resulta muy interesante. El mes de Febrero también es muy elevado como ya se ha visto anteriormente.
- Se ve nuevamente un fuerte aumento en el mes de Febrero en camapañas en varios sitios web.
- Por otro lado, hay sitios web donde hay evidencia de valorizaciones que se mantienen permanentemente a diferencia de otros en los que se concentra en meses puntuales. Aquí hay dos comportamientos diferentes que debieran abordarse con estrategias distintas.

3.6. Variación de impactos mensuales de Banco Estado según tipo de publicidad y dispositivo.

Cuadro 14: Evolución mensual de impacto según tipo de publicidad y dispositivo.

	Display	Variación	Video	Variación	Desktop	Variación	Mobile	Variación
2017-01	99.118		444.802		534.276		9.644	
2017-02	2.742.191	96,39	94.004	-373,17	2.199.672	75,71	636.523	98,48
2017-03	452.263	-506,33	25.832	-263,91	472.229	-365,81	5.866	-10.751,06
2017-04	42.803	-956,62			42.803	-1.003,26		
2017-05	1.356.378	96,84	322.466		1.653.726	97,41	25.118	
2017-06	214.316	-532,89	892.433	63,87	1.073.716	-54,02	33.033	23,96
2017-07	2.983.314	92,82	1.082.384	17,55	401.384	-167,50	3.664.314	99,10

3.7. Variación de impresiones mensuales de Banco Estado según tipo de publicidad y dispositivo.

Cuadro 15: Evolución mensual de impresiones según tipo de publicidad y dispositivo.

	Display	Variación	Video	Variación	Desktop	Variación	Mobile	Variación
2017-01	2.746.609		5.273.302		7.991.603		28.308	
2017-02	45.028.818	93,90	986.174	-434,72	43.678.770	81,70	2.336.222	98,79
2017-03	8.559.799	-426,05	270.816	-264,15	8.794.529	-396,66	36.086	-6.374,04
2017-04	692.886	-1.135,38			692.886	-1.169,26		
2017-05	8.777.669	92,11	3.382.896		11.976.471	94,21	184.094	
2017-06	1.416.867	-519,51	9.362.329	63,87	10.574.641	-13,26	204.555	10,00
2017-07	3.494.943	59,46	3.122.705	-199,81	1.607.196	-557,96	5.010.452	95,92

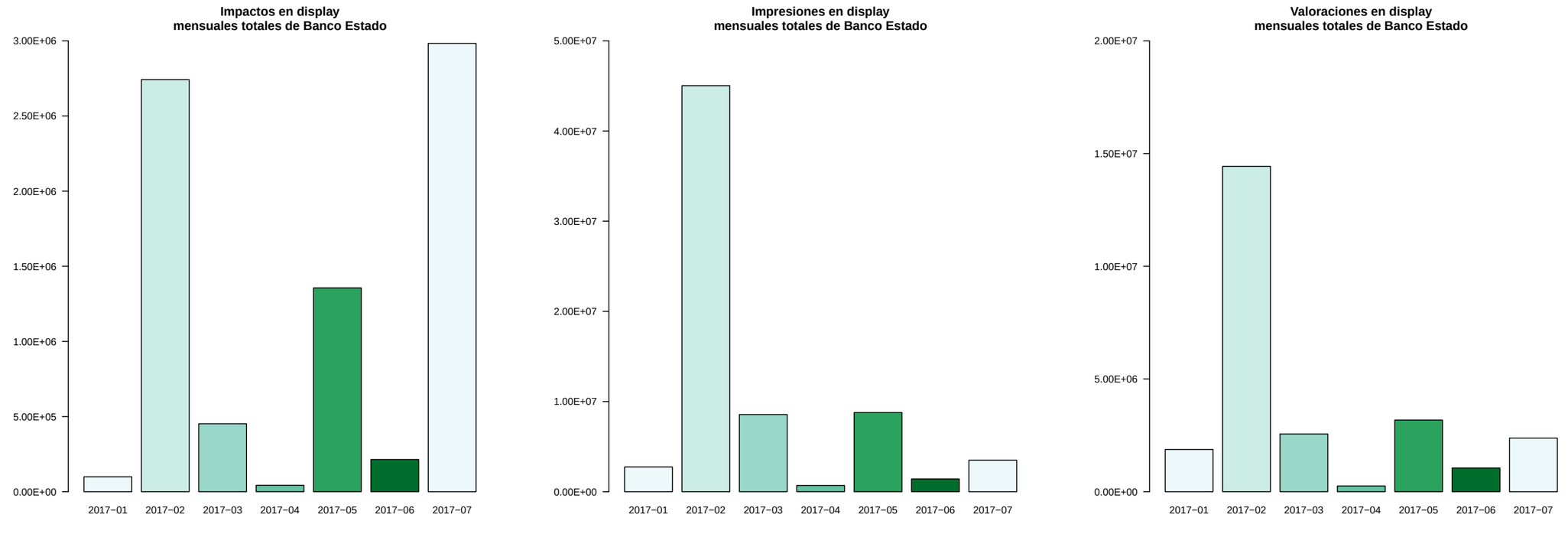
3.8. Variación de inversiones mensuales de Banco Estado según tipo de publicidad y dispositivo.

Cuadro 16: Evolución mensual de valorizaciones según tipo de publicidad y dispositivo.

	Display	Variación	Video	Variación	Desktop	Variación	Mobile	Variación
2017-01	1.873.581		11.578.897		13.355.760		96.718	
2017-02	14.433.842	87,02	8.035.346	-44,10	18.398.933	27,41	4.070.255	97,62
2017-03	2.559.526	-463,93	1.363.291	-489,41	3.850.047	-377,89	72.770	-5.493,31
2017-04	255.593	-901,41			255.593	-1.406,32		
2017-05	3.180.014	91,96	8.457.244		11.477.194	97,77	160.064	
2017-06	1.051.908	-202,31	23.405.825	63,87	24.250.039	52,67	207.694	22,93
2017-07	2.378.793	55,78	2.302.968	-916,33	1.099.325	-2.105,90	3.582.436	94,20

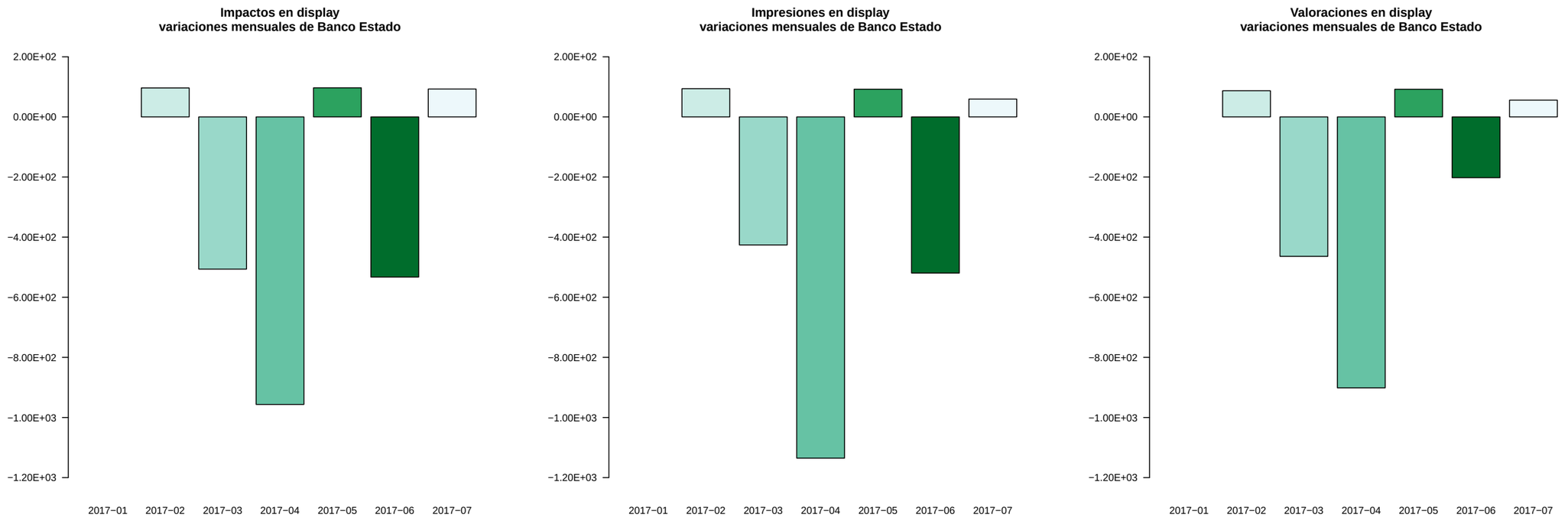
```
## pdf
##   2
```

Figura 8: Gráfico de barras de display mensuales de impacto, impresiones y valorizaciones.




```
## pdf
##   2
```

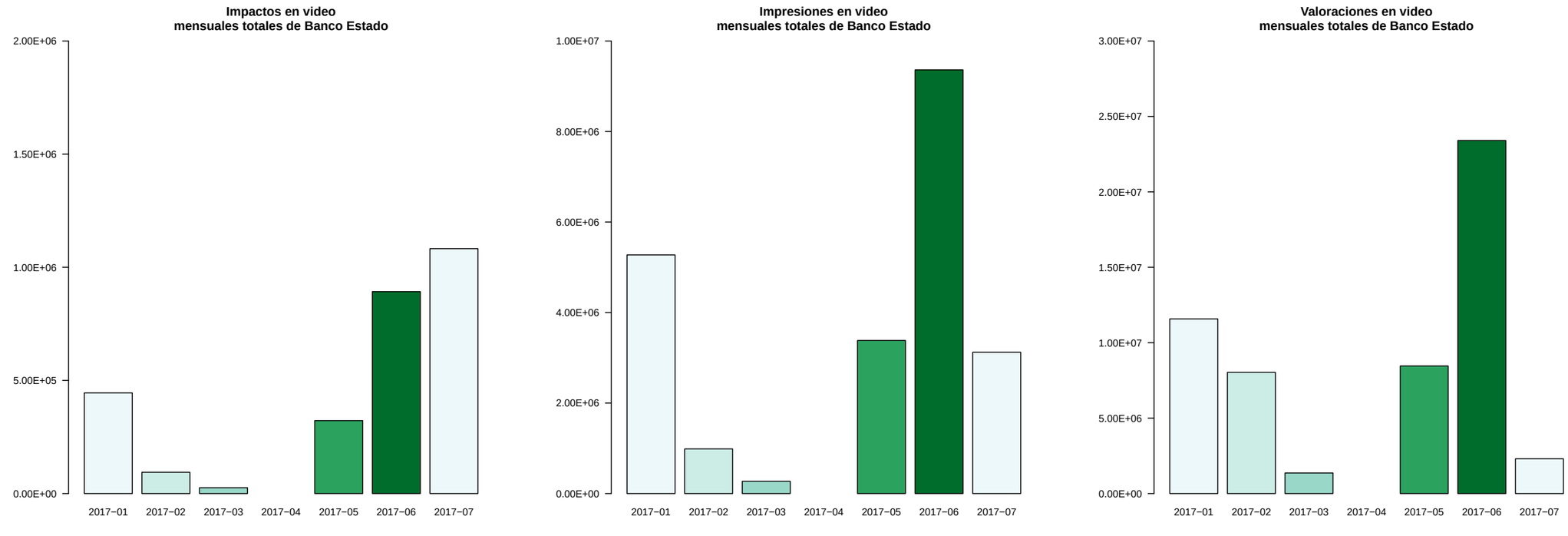
Figura 9: Gráfico de barras de display variaciones mensuales de impacto, impresiones y valorizaciones.



- Los gráficos de barras anteriores y siguientes muestran el comportamiento mensual de las variables de interés separando por tipo de publicidad y dispositivo, tanto para los valores observados como sus variaciones porcentuales.
- Los gráficos sobre display muestran un patrón similar a los visto en las variables anteriormente. Se ve un aumento abrupto al comienzo del año el que tiende aparentemente a repetirse en Julio, lo que no se corrobora en ningún caso, porque ese mes está incompleto en la base de ejemplo. De todas formas, el comportamiento es similar.
- En los gráficos siguientes esto solamente se verifica para las variables muestreadas según desktop. Mientras que aquellas medidas según video y mobile, presentan un comportamiento que no se había visto hasta ahora.

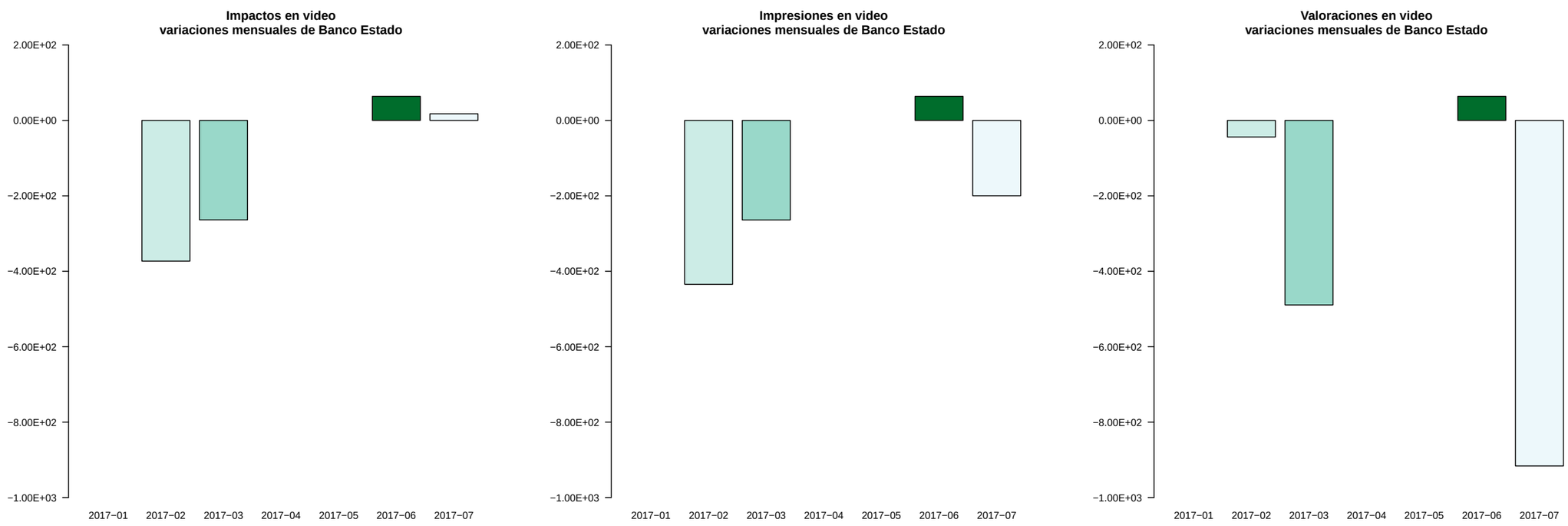
```
## pdf
##   2
```

Figura 10: Gráfico de barras de video mensuales de impacto, impresiones y valorizaciones.




```
## pdf
##   2
```

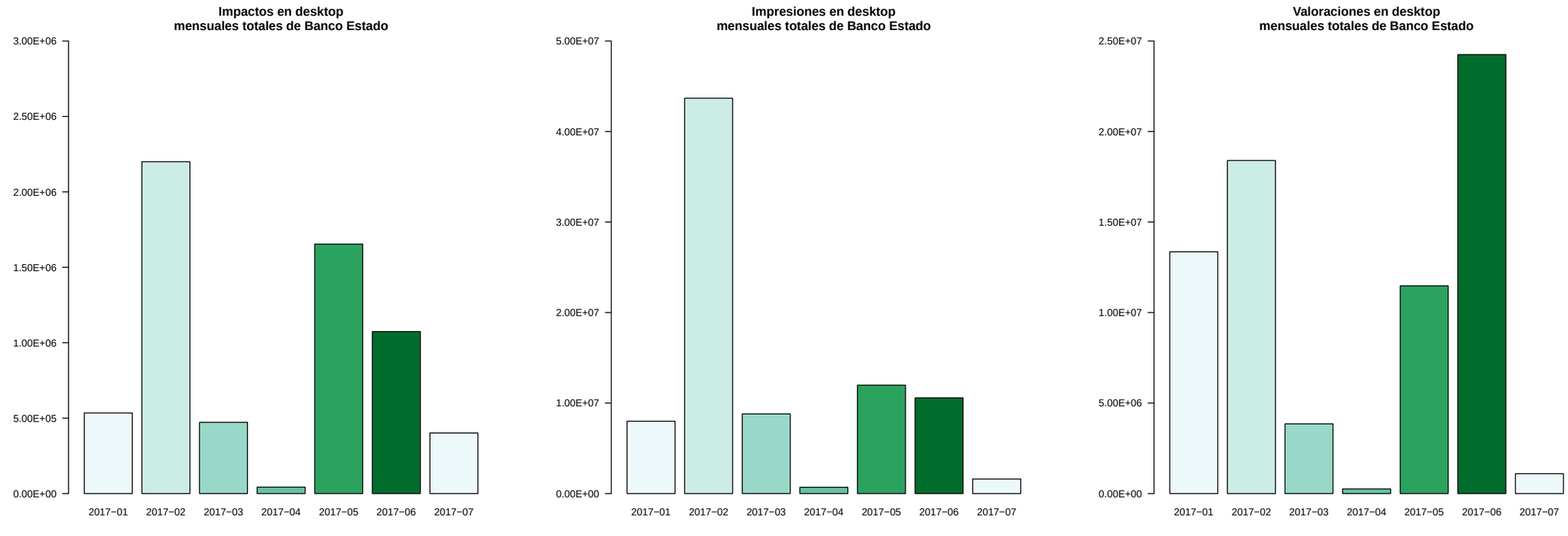
Figura 11: Gráfico de barras de video variaciones mensuales de impacto, impresiones y valorizaciones.



- La apertura por video muestra un comportamiento diferente en cuanto al muestreo, lo que se refleja en las estimaciones.
- Esto se acentúa más en la apertura según dispositivo mobile más adelante.

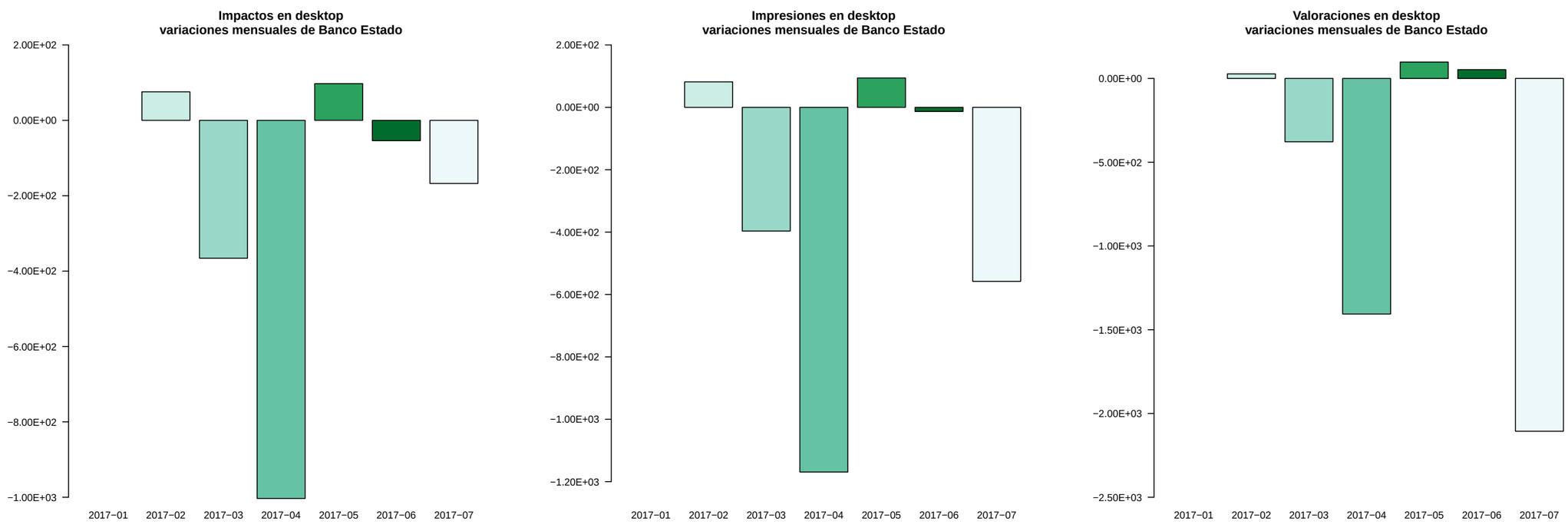
```
## pdf
##   2
```

Figura 12: Gráfico de barras de desktop mensuales de impacto, impresiones y valorizaciones.




```
## pdf
##   2
```

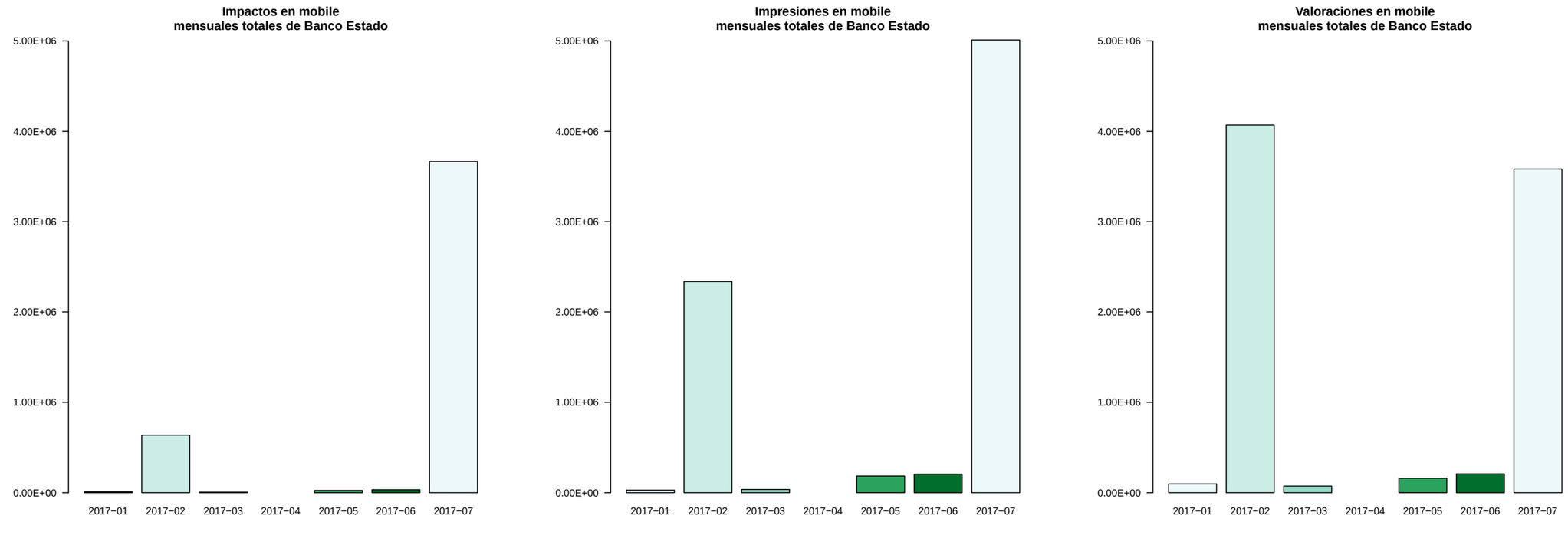
Figura 13: Gráfico de barras de desktop variaciones mensuales de impacto, impresiones y valorizaciones.



- La apertura según tipo de dispositivo desktop es más parecida al comportamiento de las variables originalmente muestreadas y valorizadas.

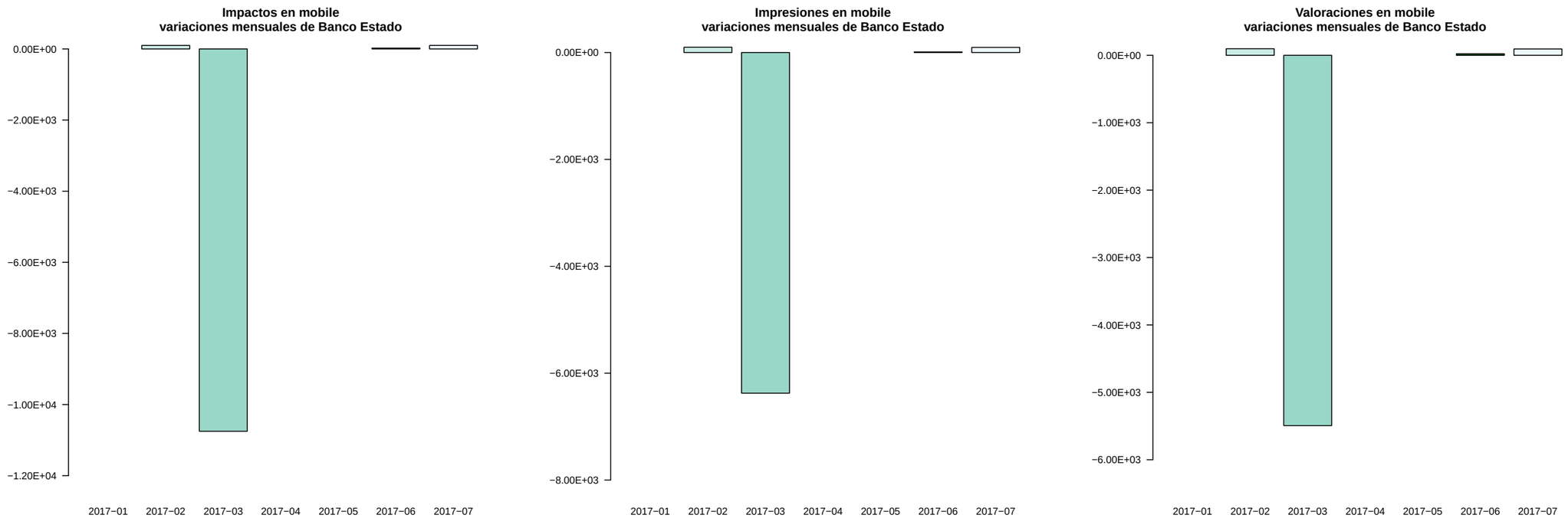
```
## pdf
##   2
```

Figura 14: Gráfico de barras de mobile mensuales de impacto, impresiones y valorizaciones.




```
## pdf
##   2
```

Figura 15: Gráfico de barras de mobile variaciones mensuales de impacto, impresiones y valorizaciones.



- La apertura según dispositivo mobile es la más errática en relación con las demás estimaciones.
- Se ve estimaciones abruptas, localizadas y con meses faltantes. Esto se refleja en variaciones porcentuales aún más erráticas y acentuadas que lo visto anteriormente.

4. Datos de la industria.

4.1. Variación de inversiones mensuales de la Industria según empresa.

Cuadro 17: Variación porcentual mensual de inversión según empresa.

	Empresa	2017-05	2017-06	Total	Variación
201	amazon prime video	523.382.105	448.648.081	972.030.186	-16,66
3278	paris	402.124.070	241.364.418	643.488.488	-66,60
3077	netflix	119.189.538	467.879.744	587.069.282	74,53
3071	nescafé	2.183.798	507.370.452	509.554.250	99,57
2979	movistar	76.892.247	282.332.229	359.224.476	72,77
867	claro	147.305.953	197.906.461	345.212.414	25,57
2465	kia	54.680.153	283.128.769	337.808.922	80,69
1619	fanta	274.161.093	45.378.901	319.539.994	-504,16
1613	falabella	99.195.672	216.206.698	315.402.370	54,12
4359	unimarc	156.684.964	120.596.886	277.281.850	-29,92
3657	ripley	158.814.394	114.312.361	273.126.755	-38,93
2563	latam airlines	71.996.256	184.974.518	256.970.774	61,08
408	banco santander	91.391.412	152.689.493	244.080.905	40,15
2429	jumbo	73.288.259	150.805.200	224.093.459	51,40
1881	gmo	190.762.195	32.467.880	223.230.075	-487,54
920	cmr falabella	142.532.363	77.712.363	220.244.726	-83,41
4369	universidad adolfo ibañez uai	73.944.097	141.779.061	215.723.158	47,85
3941	soprole	185.206.160	23.993.313	209.199.473	-671,91
3515	punto ticket	36.071.620	169.390.202	205.461.822	78,71
441	bbva banco bilbao vizcaya argentaria	65.736.606	125.309.315	191.045.921	47,54
1243	disney cine	188.981.641	1.869.990	190.851.631	-10.006,02
401	banco de crédito e inversiones bci	75.195.604	112.006.246	187.201.850	32,86
3728	samsung	71.178.079	113.870.328	185.048.407	37,49
808	chevrolet	93.704.465	88.005.238	181.709.703	-6,48
219	andes films	42.467.801	136.931.969	179.399.770	68,99
4246	toroption	87.920.586	90.472.744	178.393.330	2,82
667	carolina herrera perfumes	116.653.072	52.040.437	168.693.509	-124,16
2125	hyundai	107.526.275	55.089.753	162.616.028	-95,18
4640	wom	95.326.121	64.393.671	159.719.792	-48,04
1469	entel	44.299.749	110.321.314	154.621.063	59,84
3916	softland	79.339.184	69.902.373	149.241.557	-13,50
4528	visit argentina	71.241.747	70.988.796	142.230.543	-0,36
391	baic	69.936.124	72.108.722	142.044.846	3,01
243	apple	31.872.776	106.639.156	138.511.932	70,11
928	coca-cola	12.869.882	123.513.304	136.383.186	89,58
2841	mercado libre	94.249.702	40.512.462	134.762.164	-132,64
4342	uber	61.778.108	72.450.390	134.228.498	14,73
1858	gillette venus	226.164	125.784.078	126.010.242	99,82
4524	visa	40.921.003	84.885.469	125.806.472	51,79
359	aveeno	93.691.549	30.113.239	123.804.788	-211,13

4.2. Contabilizar outliers.

```
## [1] 2580
```

4.3. Variación de inversiones mensuales de la Industria según industria.

Cuadro 18: Variación porcentual mensual de inversión según industria.

	Industria	2017-05	2017-06	Total	Variación
152	telecomunicaciones e internet - empresas de telecomunicaciones	1.035.271.012	1.617.554.079	2.652.825.091	36,00
21	automoción - automóviles	706.517.408	1.048.152.423	1.754.669.831	32,59
166	tiendas y restaurantes - tiendas de productos al por menor	843.339.936	777.449.503	1.620.789.439	-8,48
84	finanzas - bancos	371.674.426	523.071.240	894.745.666	28,94
6	alimentación - cafés e infusiones	3.245.879	510.487.171	513.733.050	99,36
165	tiendas y restaurantes - supermercados y minimarkets	289.668.595	422.198.274	711.866.869	31,39
115	informática y equipos de oficina - software y aplicaciones	215.094.730	358.743.948	573.838.678	40,04
172	transporte, viajes y turismo - hoteles y alojamientos	108.160.918	324.981.324	433.142.242	66,72
88	finanzas - tarjetas y cheques	254.411.021	317.568.942	571.979.963	19,89
76	educación y formación - universidades y enseñanza superior	225.078.905	315.891.527	540.970.432	28,75
86	finanzas - seguros y previsión	299.376.852	303.390.412	602.767.264	1,32
169	transporte, viajes y turismo - aerolíneas	152.855.147	284.279.590	437.134.737	46,23
150	servicios públicos y privados - servicios de empresas	270.561.514	254.515.948	525.077.462	-6,30
153	telecomunicaciones e internet - equipos y terminales	152.795.728	242.392.249	395.187.977	36,96
32	bebidas - gaseosas	308.256.859	230.383.061	538.639.920	-33,80
51	construcción - empresas inmobiliarias	203.105.719	224.732.368	427.838.087	9,62
120	limpieza - higiene del hogar	96.557.114	218.580.538	315.137.652	55,83
55	cultura - cine	343.210.951	214.584.282	557.795.233	-59,94
82	eventos - tickets	43.487.542	193.965.509	237.453.051	77,58
31	bebidas - cervezas	40.035.881	192.760.092	232.795.973	79,23
164	tiendas y restaurantes - restaurantes	42.741.880	182.029.838	224.771.718	76,52
45	belleza e higiene - productos afeitado	36.079.487	169.439.496	205.518.983	78,71
170	transporte, viajes y turismo - agencias y operadores turísticos	192.088.940	157.441.704	349.530.644	-22,01
127	medios de comunicación	32.713.496	148.338.044	181.051.540	77,95
123	mascotas - alimentación animal	20.718.913	132.136.946	152.855.859	84,32
30	bebidas - bebidas alcohólicas varias	87.676.040	131.683.421	219.359.461	33,42
116	informática y equipos de oficina - varios	41.995.146	131.670.546	173.665.692	68,11
77	educación y formación - varios	61.221.275	122.203.716	183.424.991	49,90
11	alimentación - galletas	1.112.270	119.190.452	120.302.722	99,07
36	bebidas - vinos y espumantes	47.463.561	118.487.813	165.951.374	59,94
23	automoción - concesionarias	152.723.426	111.315.484	264.038.910	-37,20
156	textil y vestimenta - moda y complementos	175.933.701	110.091.399	286.025.100	-59,81
148	servicios públicos y privados - fundaciones y organizaciones	84.687.306	104.623.538	189.310.844	19,06
49	belleza e higiene - varios	34.637.675	104.233.770	138.871.445	66,77
162	tiendas y restaurantes - centros comerciales	50.394.041	100.193.507	150.587.548	49,70
144	servicios públicos y privados - consultorías y servicios empresariales	82.344.092	98.789.066	181.133.158	16,65
167	tiendas y restaurantes - tiendas online	118.998.914	94.041.602	213.040.516	-26,54
38	belleza e higiene - colonias y perfumes	159.878.232	90.072.004	249.950.236	-77,50
118	juegos y apuestas - lotería y apuestas	31.727.170	89.057.214	120.784.384	64,37
9	alimentación - cereales	6.449.604	88.545.966	94.995.570	92,72

```
## [1] TRUE
```

5. Recomendaciones.

- Sobre la identificación de variables relevantes hay que señalar varias cosas. Las aperturas, niveles o categorías en cada variable son de mucha importancia, por ejemplo, el sitio web desde donde los robots contabilizan las impresiones. Siendo una característica que puede abarcar varias empresas incluso industrias.
- Si bien es cierto, las variables analizadas están relacionadas en el proceso de estimación, es necesario su análisis para identificar dónde hay mayor incertidumbre e inconsistencias, así como también verificar los supuestos, por ejemplo, de distribución que hay detrás, en caso que se desee modelizar.
- Adicionalmente, cabe preguntar algo importante sobre las variables, o unidad de análisis mas bien. Existe un unidad de análisis identificable en el tiempo, es decir, una sola pieza publicitaria claramente identificable a modo de usarla como unidad de análisis con fines de modelización temporal o según otra apertura relevante. Si esto aplica, que creo que efectivamente sí aplica, sería ideal incorporar dentro del análisis un identificador no ambiguo para poder hacer modelos de machine learning, que incluyan esquema temporal especialmente estacionalidad.
- El segundo elemento importante, además de una posible estructura temporal, es sobre el tipo de muestreo considerando que hay variables muy sesgadas y cargadas a las colas de la distribución. La sugerencia es tratar de ajustar las muestras a una distribución log-normal, weibull o similar, en conjunto con analizar si efectivamente los outliers que se ven en las inspecciones hechas son realmente inherentes y pertenecen a la muestra en cuestión o se trata de otro fenómeno, que amerite corrección de inconsistencia, revisión de la metodología u otro.
- Es necesario conocer los gastos en inversión reales, que seguramente están a mano, con lo cual se puede indentificar en cada caso, si el proceso está efectivamente sub estimando las inversiones reales o sobre estimándolas, puesto que una estrategia sería diferente de la según corresponda.
- La posibilidad de despejar el CPM desde la variable valorización es una opción interesante, siempre que se vea que las impresiones incorporan una gran variabilidad en las estimaciones. Esto permitiría un mejor análisis además de definir con esa variable una matriz de beneficio para un método de machine learning de tipo predictivo. El definir como objetivo usar un modelo que minimice costos de impresión nos acercamos a la forma en que la empresa toma la decisión al momento de contratar las publicidades.
- El enfoque de mejora continua puede ser abordado desde mi punto de vista con diseño de flujos de datos usando un workflow system y herramientas de tipo Unix tools, Python o similares. Lo que en mi experiencia resulta rápido para mejorar la calidad de los datos.

6. Referencias.

- Which statistics are more precise? Alexa.com or similarweb.com? Somebody have insights?
- Real-time Bidding for Online Advertising: Measurement and Analysis
- Deciphering the internet advertising puzzle
- Wikipedia: Competitve Intelligence
- Which web traffic measurement service is the most accurate? Compete, Quantcast, Alexa, Comscore, etc?