
# Trabajo Práctico n° 1: Análisis exploratorio de datos.

## 75.06 Organización de Datos

Grupo: Data wan kenobi (22)

| Apellido y Nombre | Padrón |
|---|---|
| Cáceres Julieta | 96454 |
| Garcia Villamor Delfina | 101154 |
| Frutos Ramos Micaela Constanza | 96728 |
| Rozanec Matias | 97404 |

# Índice

# 1. Introducción

Trocafone es una empresa que se centra en un modelo de negocio conocido como ReCommerce. Este plantea la compra, reacondicionamiento y venta de productos previamente usados. Se encargan de verificar el estado de los celulares y reacondicionarlos en caso de ser necesario.
Al poner un producto a la venta verifican 9 puntos y prometen su correcto funcionamiento en estos. Los dispositivos se califican en tanto a su condición de venta, referido a lo estético, de esta manera el comprador conoce el producto a comprar.

Uno de los aspectos más conocidos de Trocafone es su programa de Trade-In, conocido como “plan canje” en Argentina. De esta manera se le ofrece al usuario un descuento en la compra de un nuevo producto a cambio de la entrega de uno previo usado. Trocafone brinca la plataforma donde se evalúan y cotizan los equipos para poder brindar el descuento.
Trocafone comercializa sus productos mediante los siguientes canales de venta:

- E-commerce: (https://www.trocafone.com en Brasil y https://www.trocafone.com.ar en Argentina)
- Marketplace: Presencia en Brasil y Argentina
- Tiendas físicas: con sede en Brasil

Dado que la mayor parte de los intercambios entre el usuario y la empresa se dan por medio del E-commerce y el Marketplace, es interesante recopilar los datos y efectuar un análisis de los mismos para poder tomar decisiones al respecto. Un punto muy importante a ver es el comportamiento de los usuarios dentro de la plataforma. Analizando el comportamiento se puede obtener un modelo mediante el cual predecir dicho comportamiento y mejorar así la experiencia de usuario al personalizarla.

# 2. Objetivo

El objetivo del trabajo práctico es realizar un análisis exploratorio de los datos otorgados por la cátedra. Dichos datos corresponden a la empresa Trocafone nombrada anteriormente. Se analiza el conjunto de eventos de web analytics de usuarios que visitaron www.trocafone.com (plataforma de ecommerce de Brasil).

# 3. Procesamiento y análisis

## 3.1 Información general sobre el análisis

### 3.1.1 Datos utilizados

Se analizan los datos provistos por Trocafone. provenientes de su plataforma de ecommerce de Brasil.
Se cuenta con un csv que contiene la siguiente información:

- **timestamp:** Fecha y hora cuando ocurrió el evento. (considerar BRT/ART).
- **event:** Tipo de evento
- **person:** Identificador de cliente que realizó el evento.
- **url:** Url visitada por el usuario.
- **sku:** Identificador de producto relacionado al evento.
- **model:** Nombre descriptivo del producto incluyendo marca y modelo.
- **condition:** Condición de venta del producto
- **storage:** Cantidad de almacenamiento del producto.
- **color:** Color del producto
- **skus:** Identificadores de productos visualizados en el evento.
- **search_term:** Términos de búsqueda utilizados en el evento.
- **staticpage:** Identificador de página estática visitada
- **campaign_source:** Origen de campaña, si el tráfico se originó de una campaña de marketing
- **search_engine:** Motor de búsqueda desde donde se originó el evento, si aplica.
- **channel:** Tipo de canal desde donde se originó el evento.
- **new_vs_returning:** Indicador de si el evento fue generado por un usuario nuevo (New) o por un usuario que previamente había visitado el sitio (Returning) según el motor de analytics.
- **city:** Ciudad desde donde se originó el evento
- **region:** Región desde donde se originó el evento.
- **country:** País desde donde se originó el evento.
- **device_type:** Tipo de dispositivo desde donde se genero el evento.
- **screen_resolution:** Resolución de pantalla que se está utilizando en el dispositivo desde donde se genero el evento.
- **operating_system_version:** Version de sistema operativo desde donde se origino el evento.
- **browser_version:** Versión del browser utilizado en el evento

Por otro lado, los siguientes tipos de eventos se encuentran disponibles (en el campo event) sobre los cuales se brinda una breve descripción:

- **“viewed product”:** El usuario visita una página de producto.
- **“brand listing”:** El usuario visita un listado específico de una marca viendo un conjunto de productos.
- **“visited site”:** El usuario ingresa al sitio a una determinada url.
- **“ad campaign hit”:** El usuario ingresa al sitio mediante una campana de marketing online.
- **“generic listing”:** El usuario visita la homepage.
- **“searched products”:** El usuario realiza una búsqueda de productos en la interfaz de búsqueda del site.
- **“search engine hit”:** El usuario ingresa al sitio mediante un motor de búsqueda web.
- **“checkout”:** El usuario ingresa al checkout de compra de un producto.
- **“staticpage”:** El usuario visita una página
- **“conversion”:** El usuario realiza una conversión, comprando un producto.
- **“lead”:** El usuario se registra para recibir una notificación de disponibilidad de stock, para un producto que no se encontraba disponible en ese momento.

### 3.1.2 Lenguaje y bibliotecas utilizadas

Para el procesamiento de los datos se hace uso de Pandas, biblioteca de Python.

Se utilizan bibliotecas tales como Matplotlib, Seaborn para visualizar.

Para obtener las coordenadas se hace uso de la biblioteca geopy y para visualizar los mapas correspondientes se utiliza folium.

Para el análisis de la cercanía entre eventos se utiliza el algoritmo apriori, que se encuentra en la librería mlxtend.

Para realización de visualizaciones tipo wordcloud se hace uso de la biblioteca Wordcloud.

### 3.1.3 Repositorio de Github

Con el propósito de integrar el análisis sobre el set de datos efectuado por cada integrante del grupo se hace uso de un repositorio de GitHub donde pueden encontrarse los diferentes notebooks en los cuales se desarrolló dicho trabajo práctico.
Link al repositorio: https://github.com/rozanecm/7506_OrganizacionDeDatos_TP1

# 3.2 Depuración

## 3.2.1 Análisis de datos nulos para depuración

Se comienza leyendo los datos en un Dataframe. Se estudian los datos generales del dataframe haciendo uso del método info():

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1011288 entries, 0 to 1011287
Data columns (total 23 columns):
timestamp                   1011288 non-null object
event                       1011288 non-null object
person                      1011288 non-null object
url                         82756 non-null object
sku                         563838 non-null object
model                       564284 non-null object
condition                   563836 non-null object
storage                     563836 non-null object
color                       563836 non-null object
skus                        221699 non-null object
search_term                 48967 non-null object
staticpage                  3598 non-null object
campaign_source             82796 non-null object
search_engine               50957 non-null object
channel                     87378 non-null object
new_vs_returning            87378 non-null object
city                        87378 non-null object
region                      87378 non-null object
country                     87378 non-null object
device_type                 87378 non-null object
screen_resolution           87378 non-null object
operating_system_version    87378 non-null object
browser_version             87378 non-null object
dtypes: object(23)
memory usage: 177.5+ MB
```

Figura 1. Datos no nulos en las columnas del Dataframe

El dataframe tiene un total de 1011288 columnas, de las cuales 'timestamp', 'event', y 'person' estan completas, es decir que ninguno de sus valores es NaN. Esto tiene sentido dado que cada vez que un usuario ingresa al sitio se le asigna un ID de usuario, que corresponde al campo 'person', un timestamp, que indica el momento en el cual se encuentra en la página, y el evento que genera. En cambio se puede ver como el resto de las columnas tienen varios valores nulos. A continuación se analiza si estos tienen sentido o si son datos erróneos. En caso de ser erróneos debe tomarse algún criterio para continuar con el análisis. Sabemos que hay datos que solo corresponden a cierto evento, buscamos dicha correspondencia.

Si contamos la frecuencia de cada evento, podemos relacionar esta cantidad con lo antes visto:

```
data['event'].value_counts()
```

```
viewed product        528931
brand listing          98635
visited site           87378
ad campaign hit        82827
generic listing        67534
searched products      56073
search engine hit      50957
checkout               33735
staticpage              3598
conversion              1172
lead                     448
Name: event, dtype: int64
```

Se puede ver como la cantidad de eventos ‘visited site’ coincide con la cantidad de campos no nulos en las columnas ‘channel’, ‘new_vs_returning’, ‘city’, ‘region’, ‘country’, ‘device_type’, ‘screen_resolution’, ‘operating_system_version’, y ‘browser_version’.

Por lo cual se podría pensar que al ingresar mediante este evento se guarda esa información, y no así para el resto de los eventos, lo cual explicaría porqué hay datos nulos, e indicaría que estos datos son correctos.

Figura 2. Frecuencia de los distintos eventos.

Se muestra a continuación los datos correspondientes a cada evento:

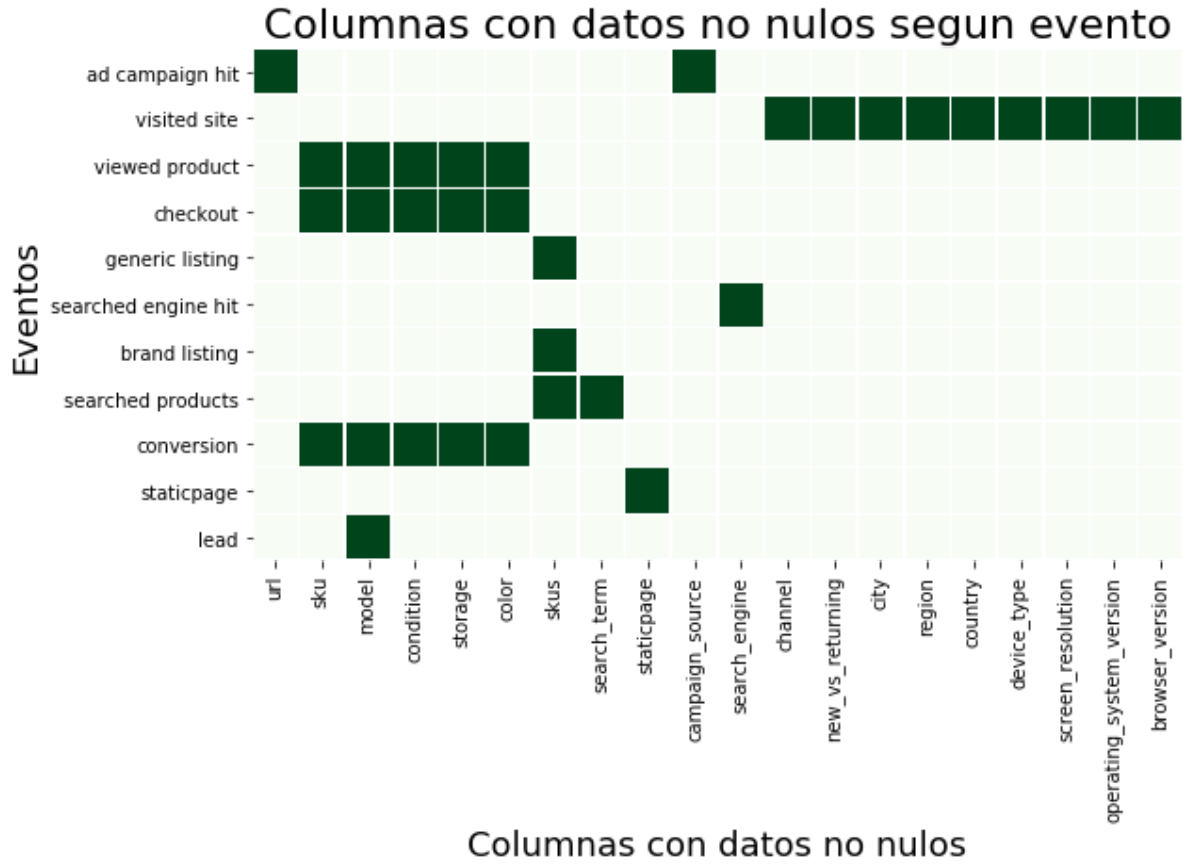



Figura 3. Datos correspondientes a cada evento.

Para cada evento, indicado en la fila, los cuadrados de color verde indican las filas a las cuales les corresponde un valor distinto de NaN según la columna del dataframe original. El resto de los valores presentan a lo largo de toda la columna valor nulo.
Se concluye que no es necesario realizar una depuración de los datos.

Dado que los datos se van a analizar a lo largo del tiempo se cambia el formato de la columna timestamp para que sea más cómodo el manejo de estos. Separando así la información en columnas tales como ‘mes’, ‘dia’, ‘hora’, entre otras.

# 3.3 Zonas geográficas de origen de actividad

## 3.3.1 Análisis general

Se analizan los países desde los cuales se generan los eventos. Para esto se calcula la cantidad de países diferentes y se verifica si hay datos erróneos. Se calcula un total de 46 países distintos y 1939 ciudades distintas. El porcentaje de NaNs de esta columna es alto, pero es por lo explicado anteriormente, por lo cual en realidad no hay valores erróneos, pero si hay valores faltantes dado que un valor posible tanto para país como para ciudad es el de ‘unknown’.
Se calcula la frecuencia de los países de forma porcentual:

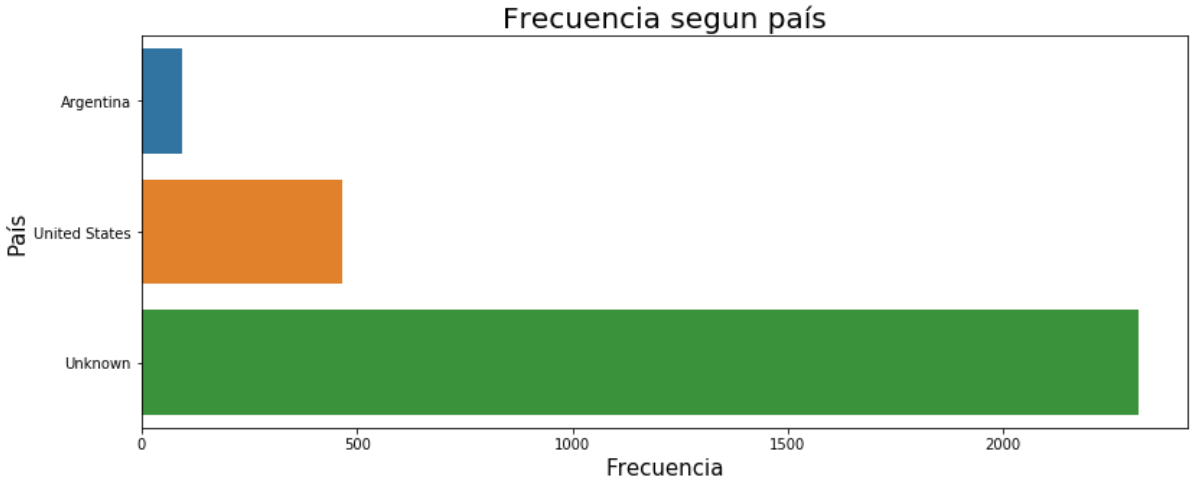
```
data.country.value_counts(normalize = True).head()
```

```
Brazil           0.964865
Unknown          0.026471
United States    0.005322
Argentina        0.001076
Canada           0.000401
Name: country, dtype: float64
```

Figura 4. Porcentaje de aparición del pais en cuestion.

Brasil representa el 96,5% de los valores. Por esta razón se decide no visualizar la frecuencia de dicho País, dado que impide que se vea bien el resto de los datos. Hay mas paises que los mostrados en la imagen. Pero su frecuencia es casi nula.

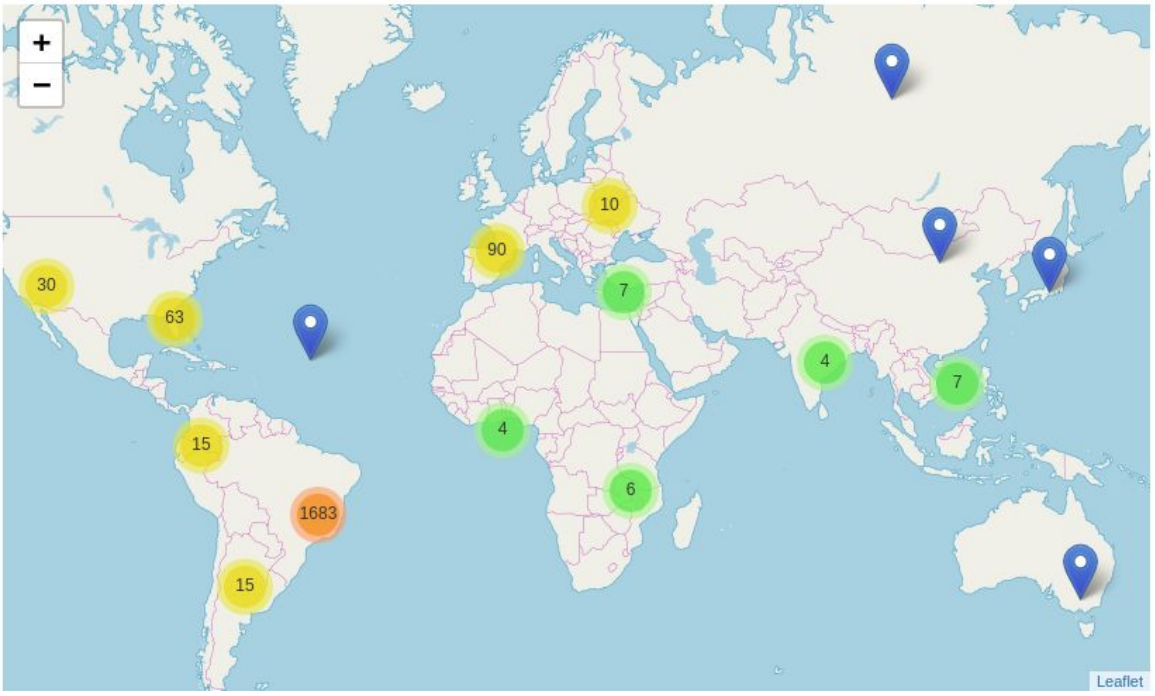
Se grafican los tres países que le siguen a Brasil. Unknown es un valor de país en los datos por lo cual este se ve en el gráfico.



El porcentaje de ‘unknown’ es considerablemente alto. Analizamos si a algún país ‘unknown’ le corresponde alguna ciudad y así poder inferir el país, pero encontramos que a todos les corresponden también ciudades ‘unknown’ por lo que no podemos completar esos datos.
Al ser la mayor parte de los usuarios es de Brasil. Por lo cual tiene sentido que se tome como zona horaria la sona Brasil/Argentina.

## 3.3.2 Mapas

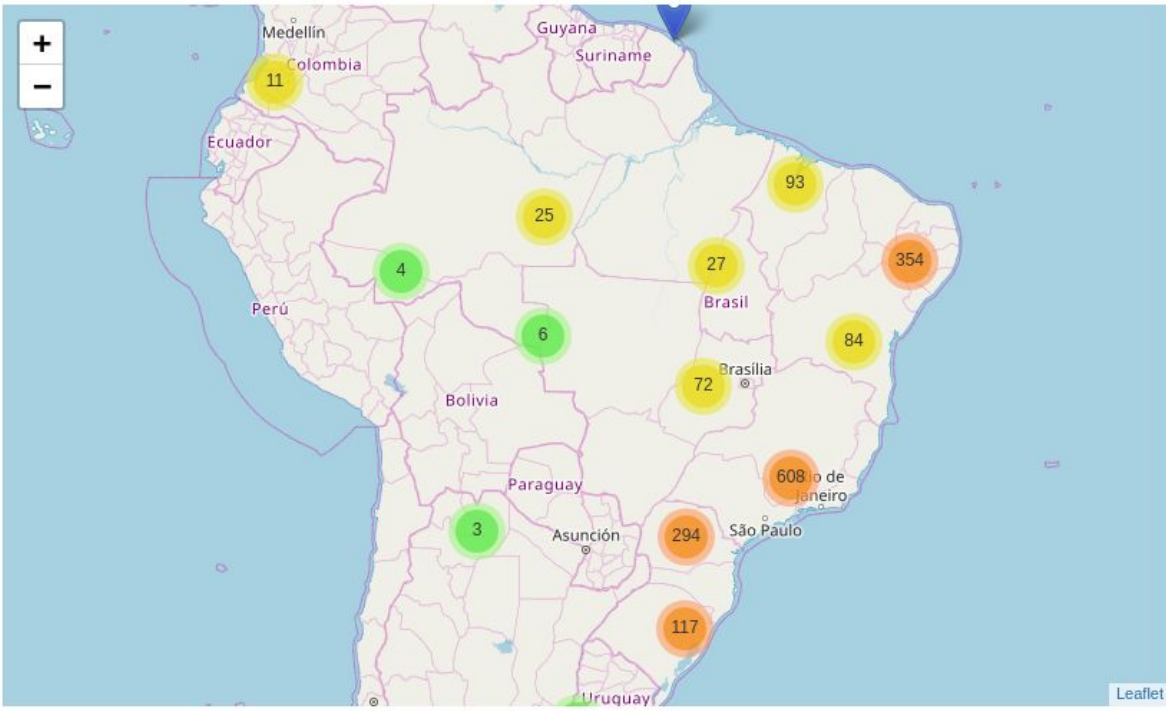
Se buscan las coordenadas de las ciudades para poder graficarlas en un mapa y ver sobre este dónde están los clientes. Para esto se hace uso de la antes mencionada biblioteca geopy. El tratamiento de los datos y proceso mediante el cual se obtienen las coordenadas se encuentra en el notebook de nombre “tp1.ipynb” dentro del repositorio. Se recomienda generar el mapa dentro del notebook dado que es interactivo. A continuación se muestran algunas imágenes del mapa con distinto zoom.




1. Mapa completo.

En esta vista se tiene el mapa completo. Se puede ver un outlier que queda en el agua, esto se debe a un error en las coordenadas de dicho punto. El resto de los datos corresponden a zonas válidas del mapa.


2. Vista de América del sur con más detalle.



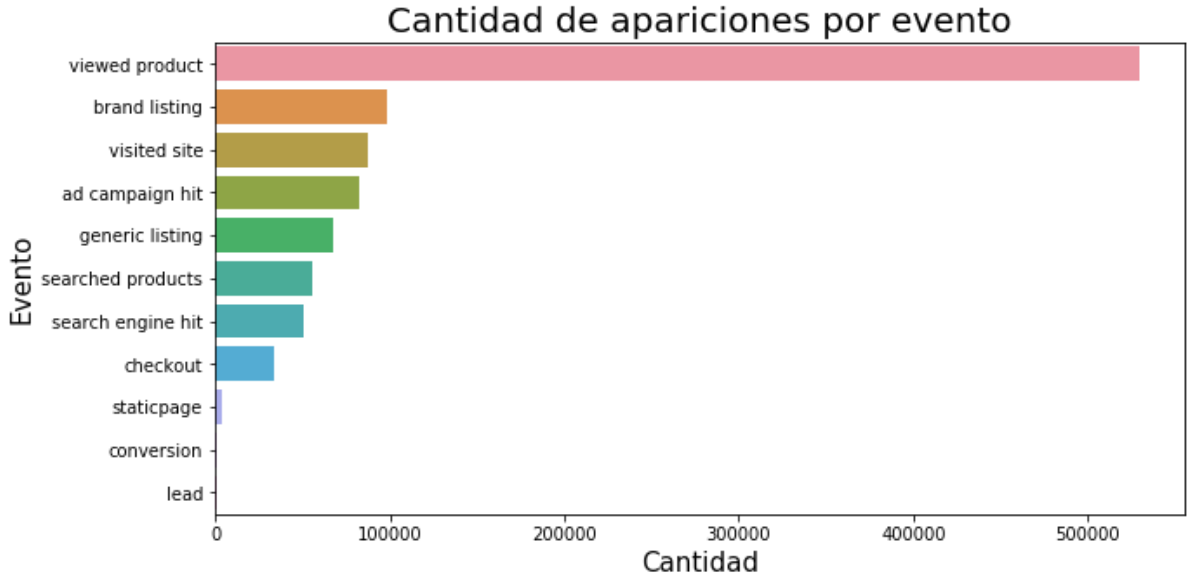

3. Vista de brasil.

Se muestra a Brasil por ser el país de mayor tráfico de datos.

# 3.4 Análisis por eventos

## 3.4.1 Análisis de los eventos en general

Se grafica la cantidad de apariciones de cada evento para ver la relación que hay entre estos.



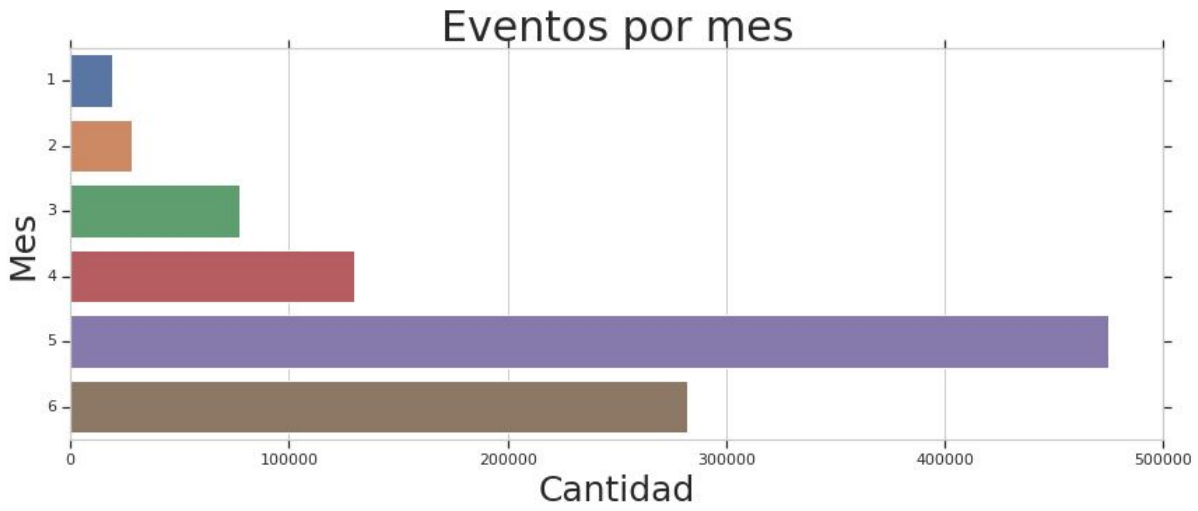

Se puede ver como el evento viewed product es el más frecuente, siendo este el evento que representa cuando un usuario ingresa a una pagina de cierto producto. Tiene sentido que del evento Lead haya muy poca cantidad dado que representa cuando un usuario ingresa su email para ser notificado en caso de renovar stock de cierto producto, lo cual no ocurre frecuentemente.
Un evento de particular interés es el de Conversión, dado que indica que se efectuó una compra. En el gráfico se puede ver que hay muy pocas conversiones. El evento checkout también es importante dado que indica que el usuario selecciona un dispositivo para comprar.
Se puede graficar la evolución de los eventos a través del tiempo.

## 3.4.2 Actividad a través del tiempo

### 3.4.2.1 Actividad en el año

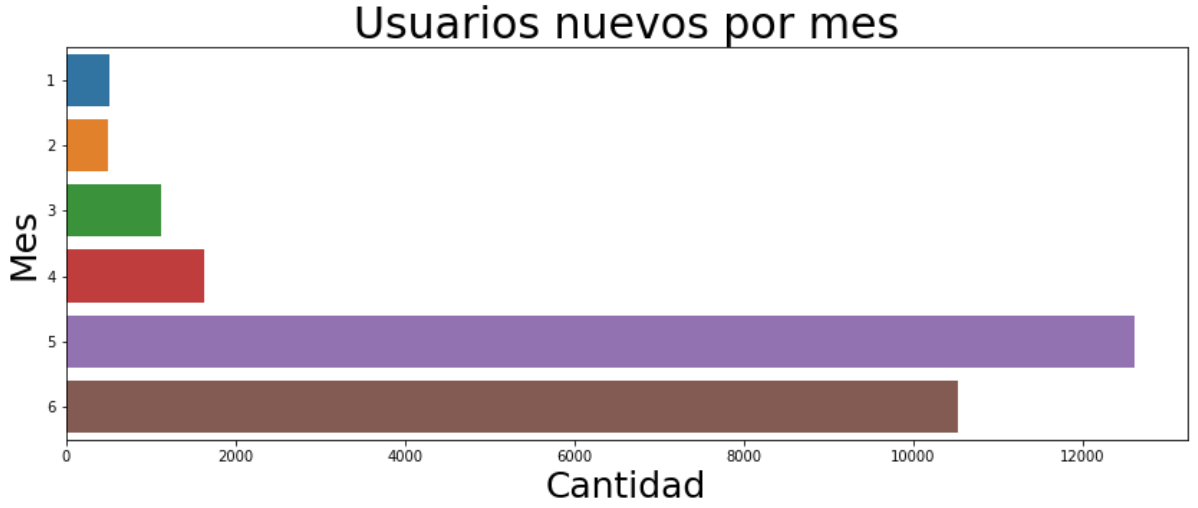
Análicemos la cantidad de actividad de usuarios que tuvo la página durante el tiempo.
Los datos analizados corresponden todos al año 2018, y es importante aclarar que estos van desde principios de enero hasta la mitad de junio. El hecho de que no haya datos del mes de junio completo afecta al análisis.

Primero veamos la actividad general de los usuarios:

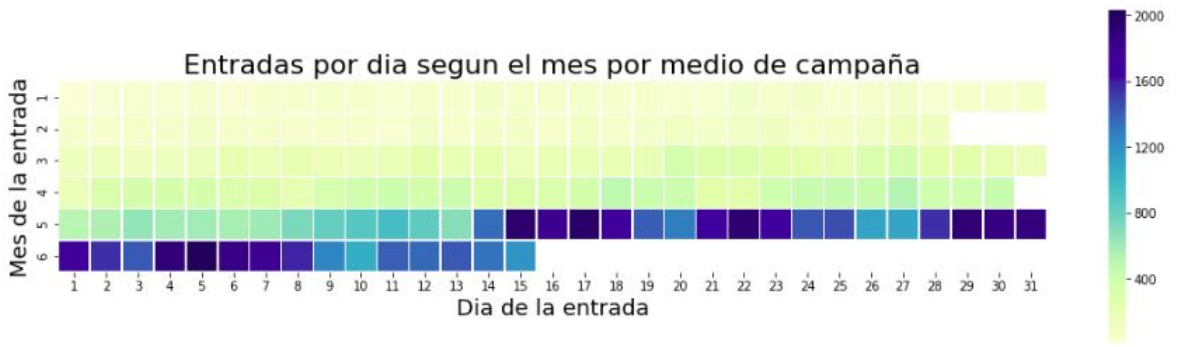



Hay significativamente más datos del mes 5. Si los datos son representativos, se podría concluir que hubo un aumento de la actividad durante el año con mucho más tráfico durante el quinto mes. Ahora, esa disminución en el último mes no quiere decir que haya disminuído el tráfico, recordemos que tenemos datos hasta la mitad de este, por lo que no es realmente representativo de todo el tráfico durante todo ese mes.

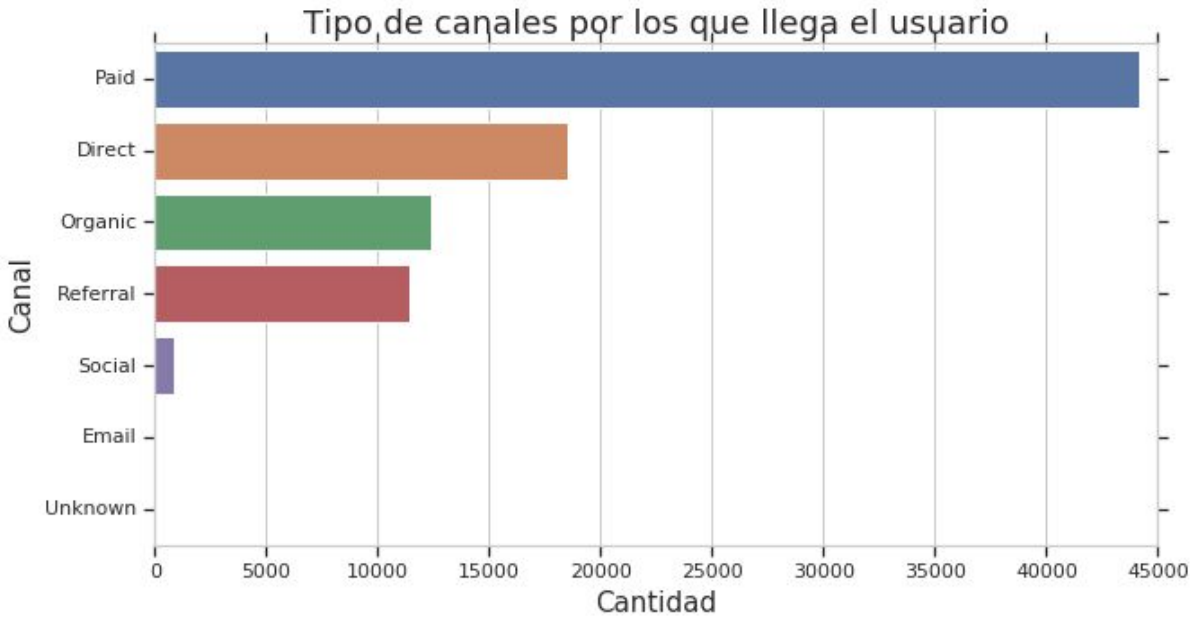
Es interesante ver si esta nueva actividad se  corresponde con un aumento de los usuarios nuevos:




Se puede ver como aumenta la cantidad de usuarios nuevos a medida que pasan los meses, viéndose un pico en el mes de mayo. Así que podemos concluir que ese aumento de actividad efectivamente fue por un aumento en usuarios para el sitio.
Ahora lo que nos quedaría responder es qué es lo que efectuó este aumentó. Uno pensaría que hubo un aumentó debido a campañas por parte de la empresa, así que primero analicemos las entradas por medio de campaña:

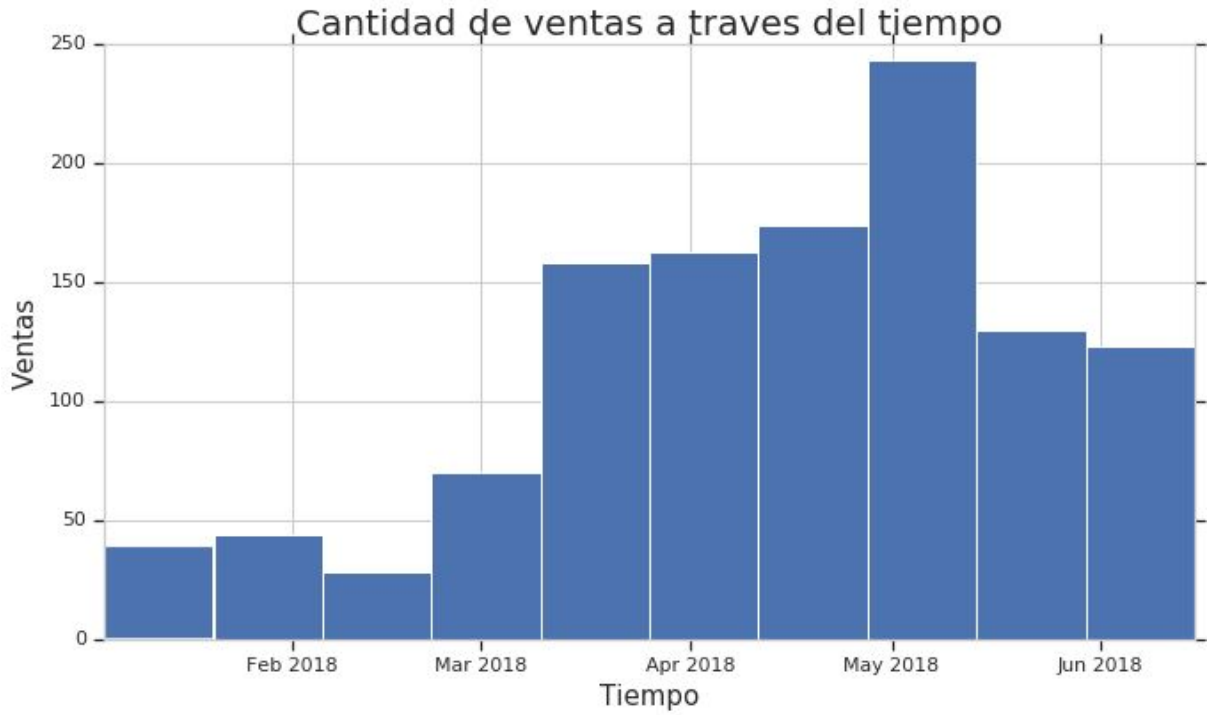



El ingreso al sitio por medio de campañas presenta la misma tendencia. Aún así hay que tener en cuenta que esto no quiere decir que la empresa haya invertido más en dichos meses, sino que dieron mejor resultado.
Se analizan los canales mediante los cuales el usuario ingresa al sitio.




De hecho podemos ver cuando estudiamos el tipo de canal por el que llegan al sitio los usuario que la gran mayoría viene por medios pagos (campañas).

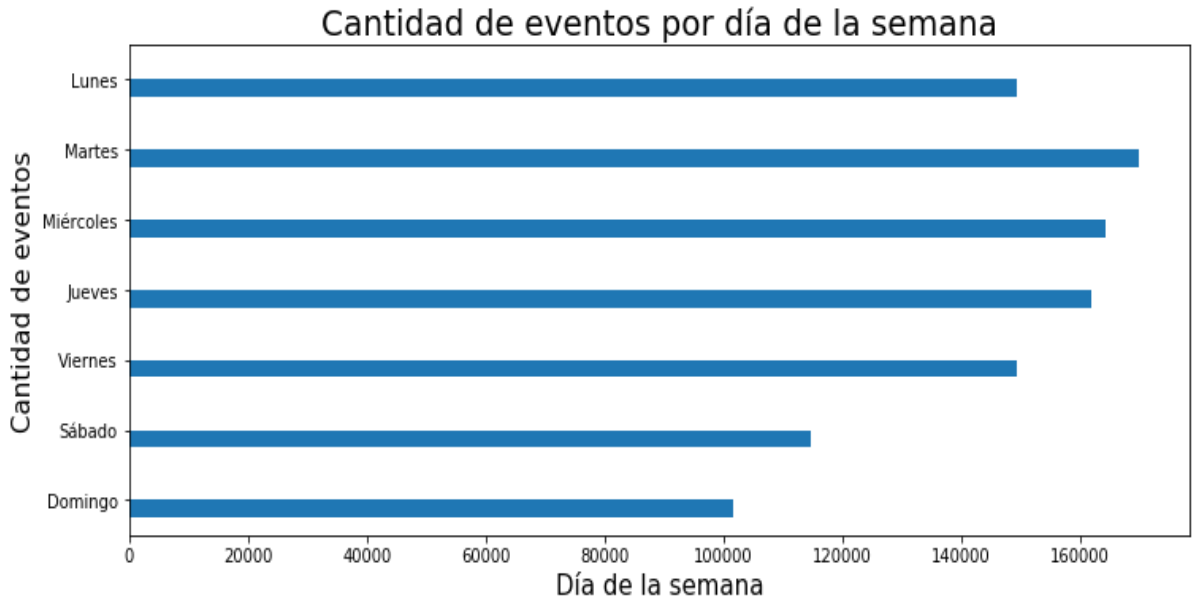
Ahora, si bien pudimos ver que hubo un aumento de actividad durante esos meses y este está fuertemente ligado con las entradas por medios de campaña, lo importante acá es ver si este crecimiento de la actividad se traduce a un aumento en las ventas.




Podemos ver que definitivamente hubo un crecimiento en las ventas que acompaña a lo anterior analizado.

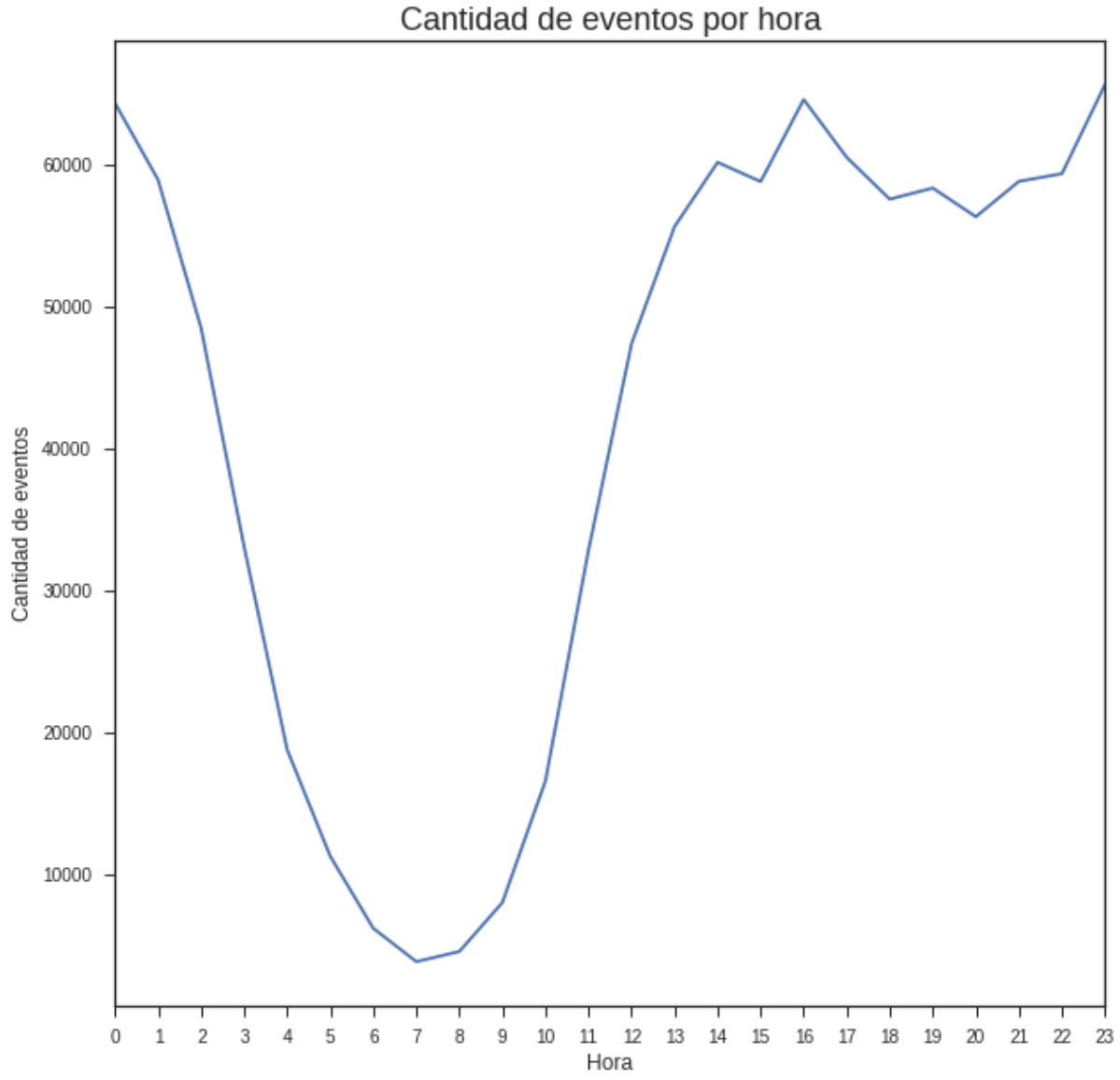
### 3.4.2.2 Actividad durante la semana

Estudiamos el comportamiento de los usuarios según los días de la semana:



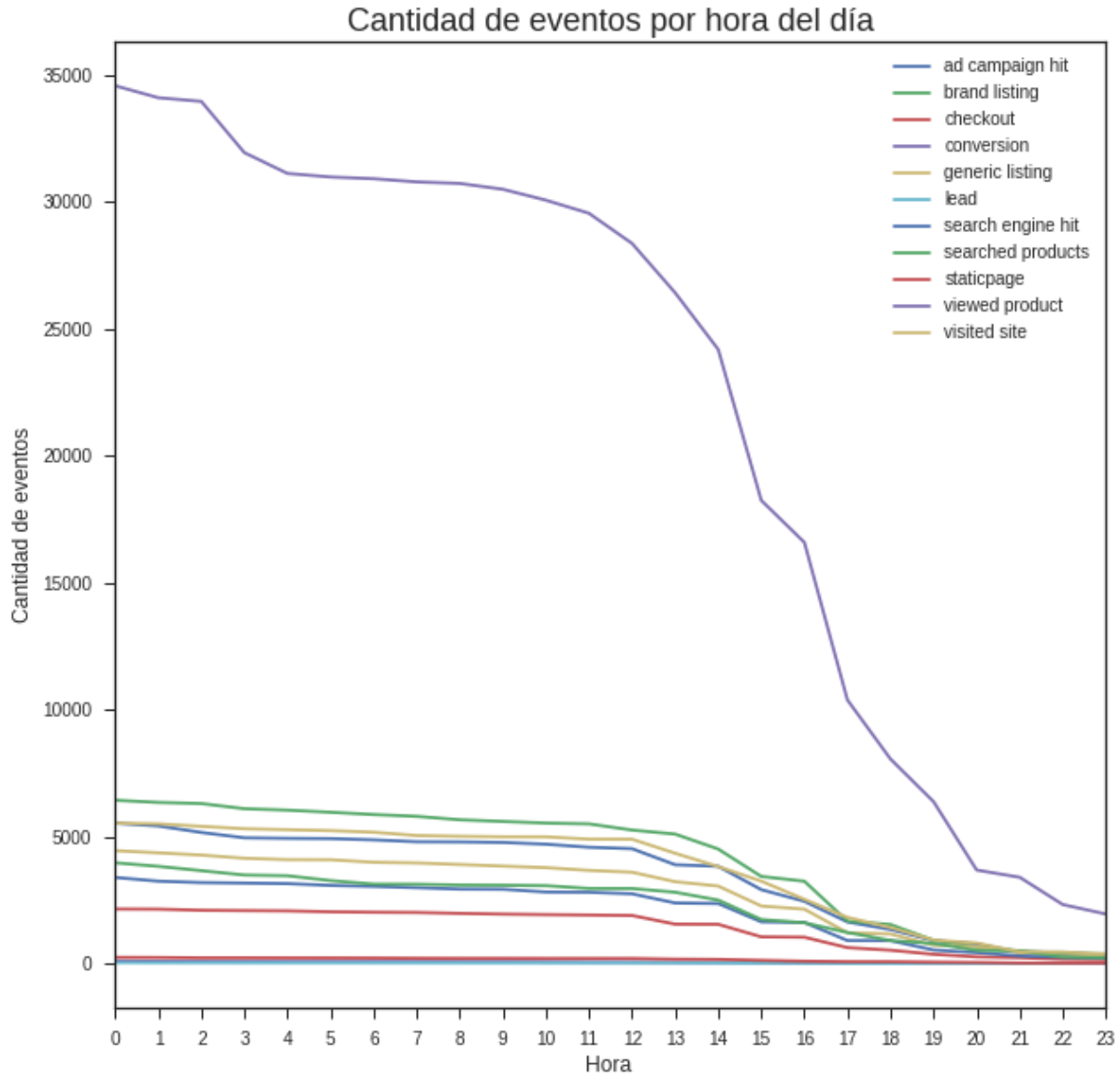
Los fines de semana hay menos tráfico en el sitio, siendo el martes el día de mayor tráfico. Se puede suponer que esto se debe a que los fines de semana los usuarios ocupan más tiempo en actividades sociales y que durante la semana navegan en este tipo de sitios en momentos libres.

### 3.4.2.3 Actividad segun la hora del dia

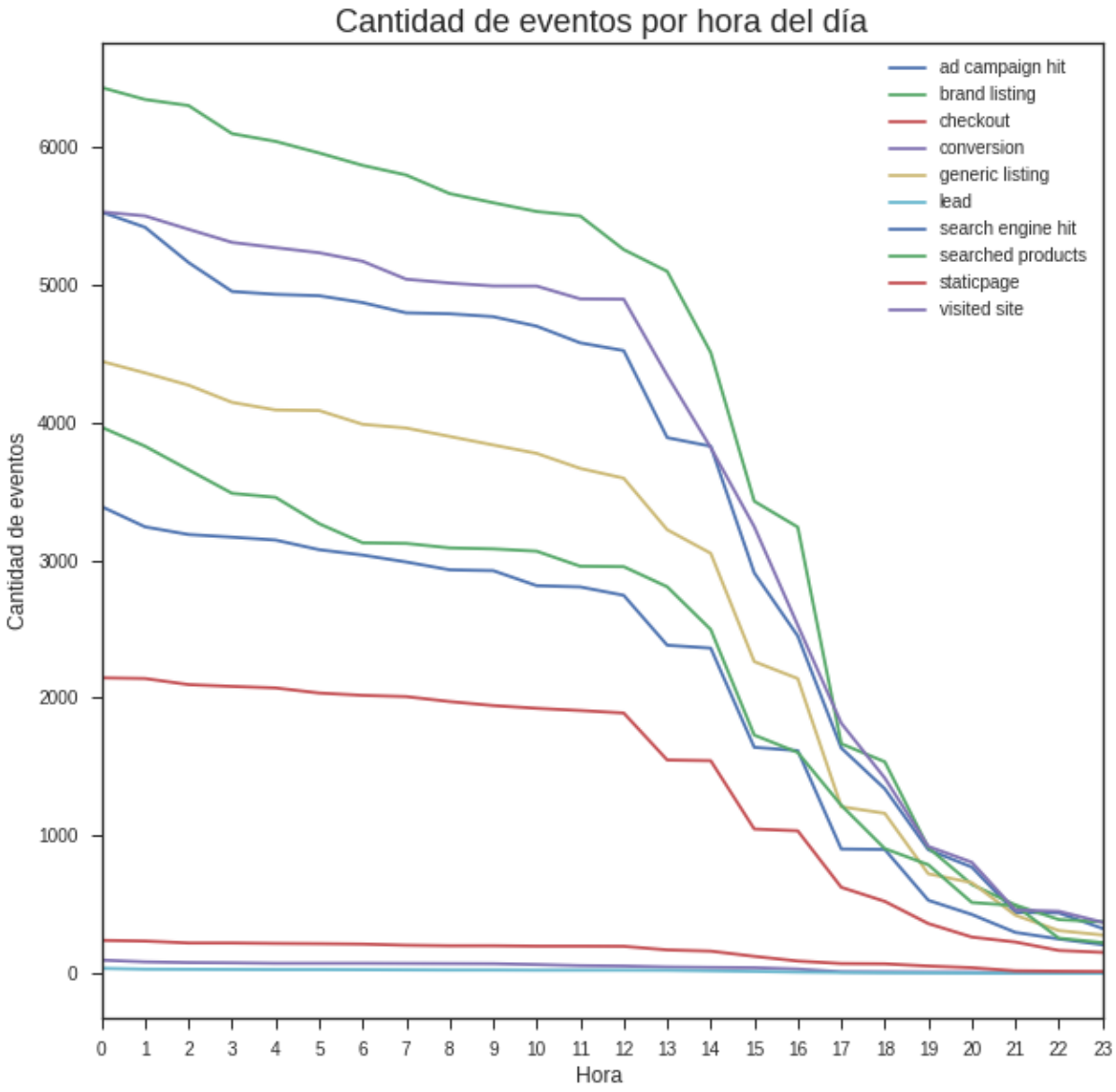



La actividad disminuye considerablemente a la madrugada, lo cual es de esperar dado que en Brasil, que es el país que consideramos para la franja horaria se duerme en ese rango horario. La actividad se mantiene en el mismo rango entre las 12 y la 1.

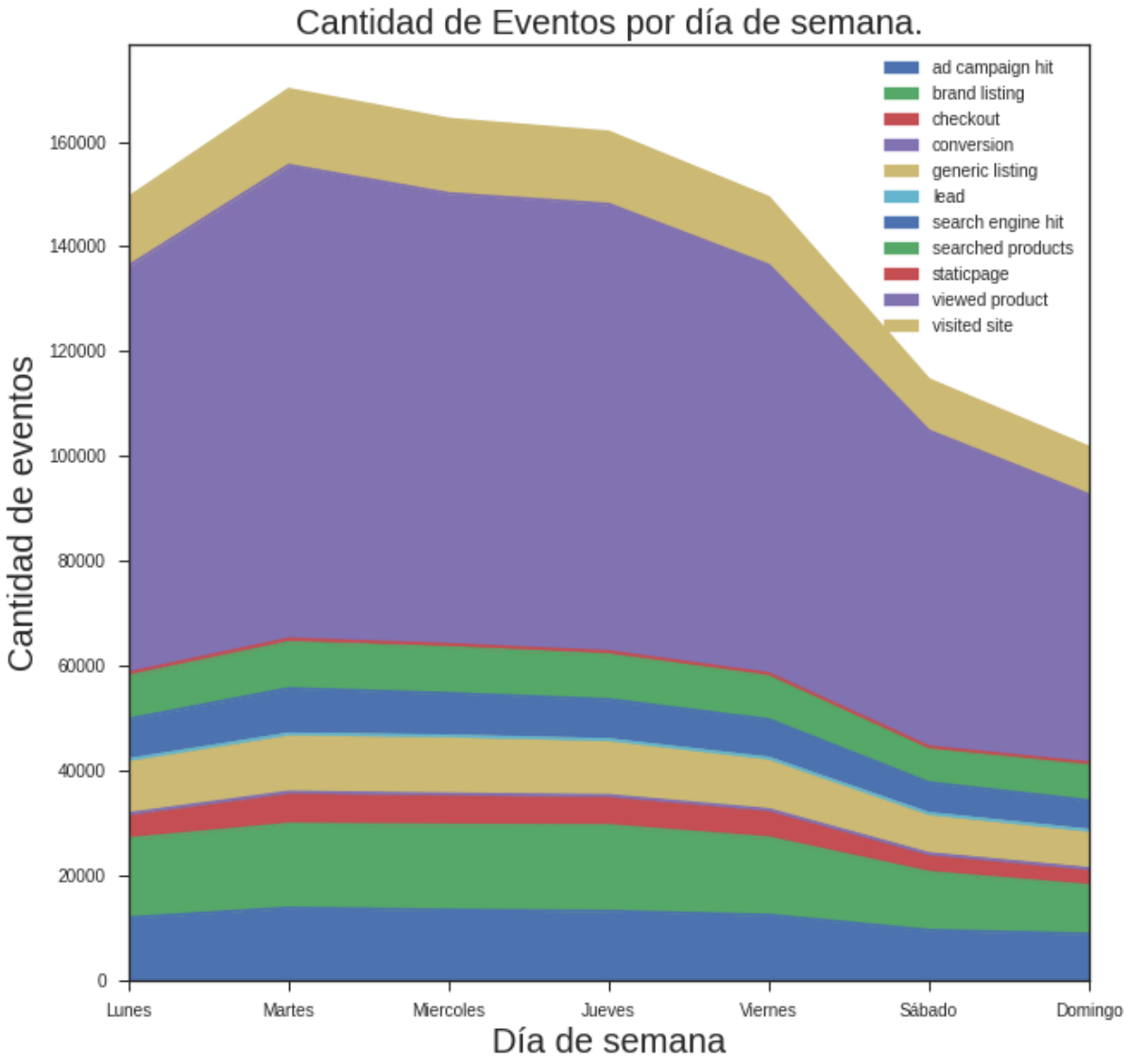
Para tener un mayor detalle de lo expresado se procede a analizar el comportamiento particular de cada posible evento, obteniendo los siguientes resultados:



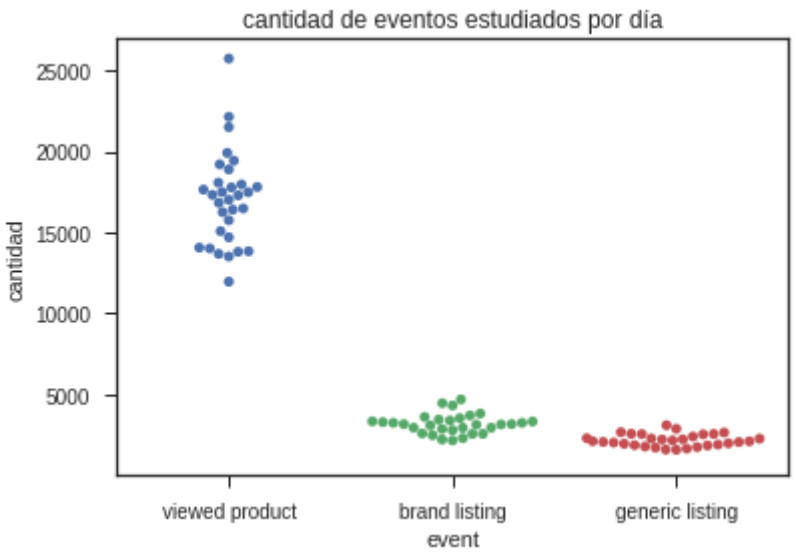

Los eventos siguen un comportamiento parecido respecto al tiempo, puesto que a partir de las 18 la cantidad de eventos realizados por los usuarios disminuye notablemente en comparación a los demás horarios, también se observa que hay mucha más cantidad de usuarios realizando el evento “viewed product” lo cual tiene sentido debido a que se espera que el usuario vea muchos modelos antes de comprar uno. Dicho esto se procede a remover dicho evento de nuestro gráfico para poder apreciar más aún las diferencias entre los eventos restantes, obteniendose:




Nos preguntamos ahora cómo varían según el día de semana. Obteniendose el siguiente grafico:




Se ve que la franja de días en los cuales los usuarios realizan más cada uno de los eventos es similar, siendo esta el intervalo de martes a viernes. Los fines de semana son los de menor actividad incrementando un poco el lunes.
Para ver la dispersión en la cantidad de veces que se realizan los eventos que implican la visita a los productos a lo largo de los días se procedió a realizar un diagrama de dispersión categórico obteniendo el siguiente gráfico:




Donde se puede ver que en diferentes días la cantidad de *brand listing event* y *generic listing* son similares, mientras que en *viewed product* la cantidad de realizaciones es más dispersa.

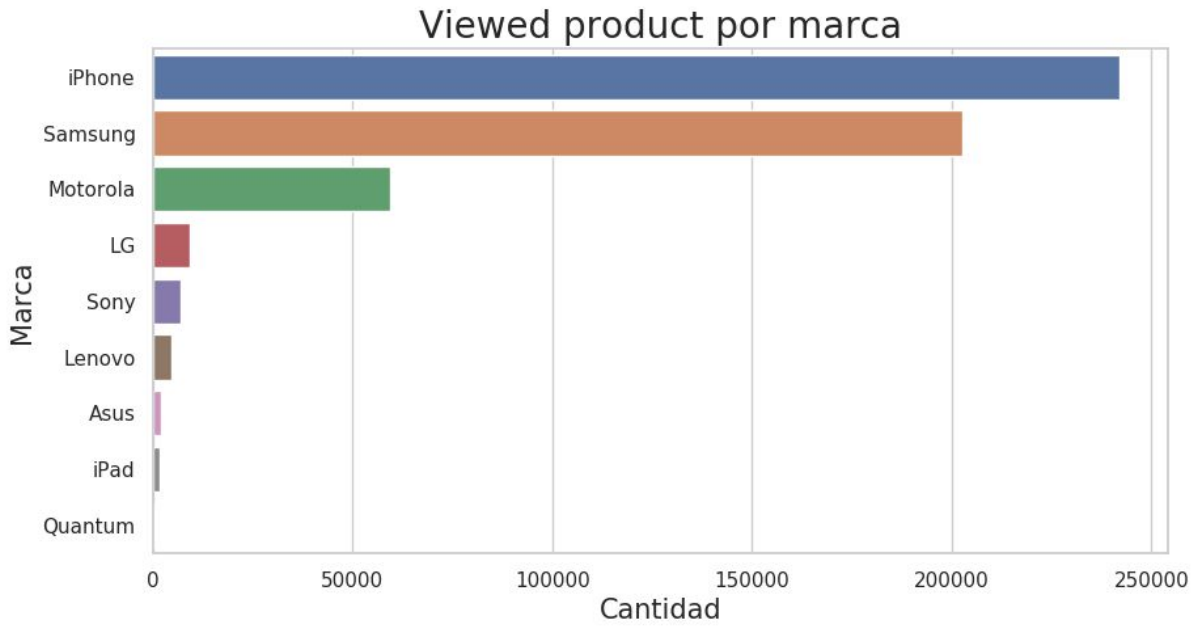
# 3.5 Análisis de marcas y modelos

## 3.5.1 Marcas populares

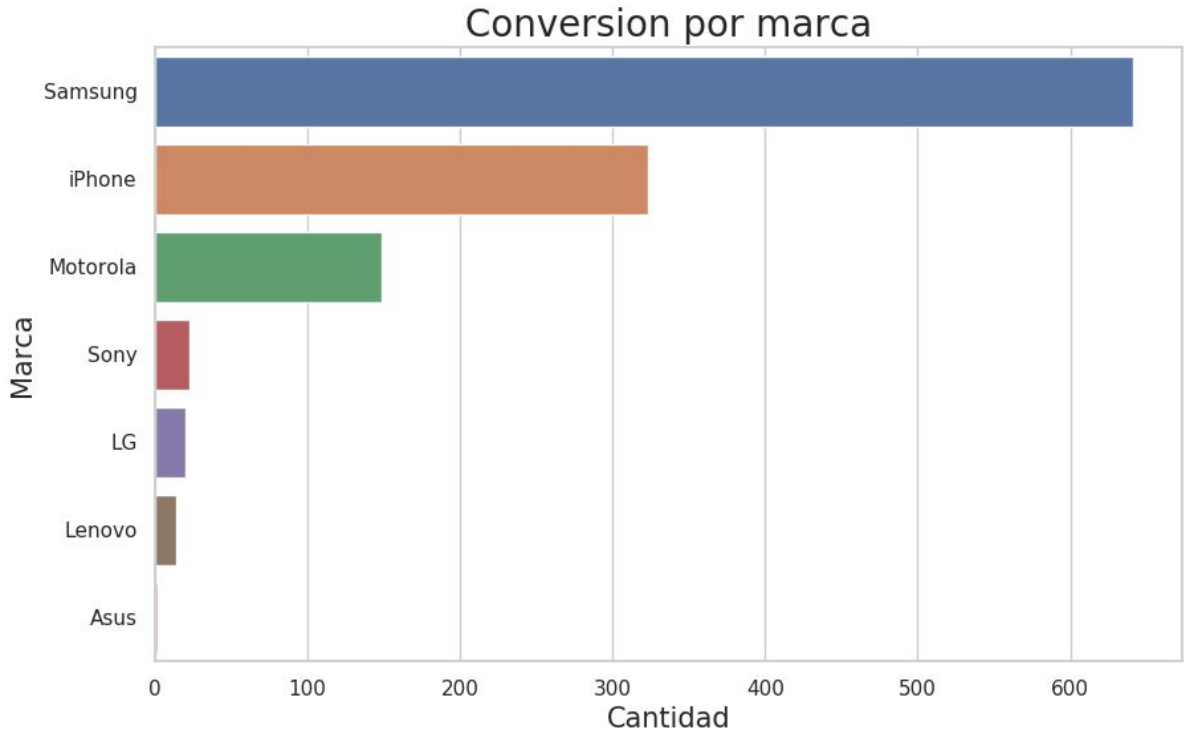
Una de las cosas interesantes para ver es las popularidad de las marcas dentro de los usuarios que utilizan trocafone.
Analizando las marcas que aparecen en los datos los resultados indican que las marcas más populares son Samsung e iPhone, seguidas por Motorola. Para ilustrar:


Figura 6. Wordcloud de marcas más populares

Ahora, si bien ahí nos dimos una idea de cuales son las marcas que más le interesan a los usuarios, es importante ahora estudiar la distribución de estas según las vistas de los productos por los usuarios y las compras.
Empecemos por ver cuales son de las que más productos se ven.




Está claro que la marca más buscada es iPhone, seguida por Samsung. En tercer lugar se encuentra Motorola con menos de un tercio del valor de iPhone.
En tanto a la venta de los dispositivos:

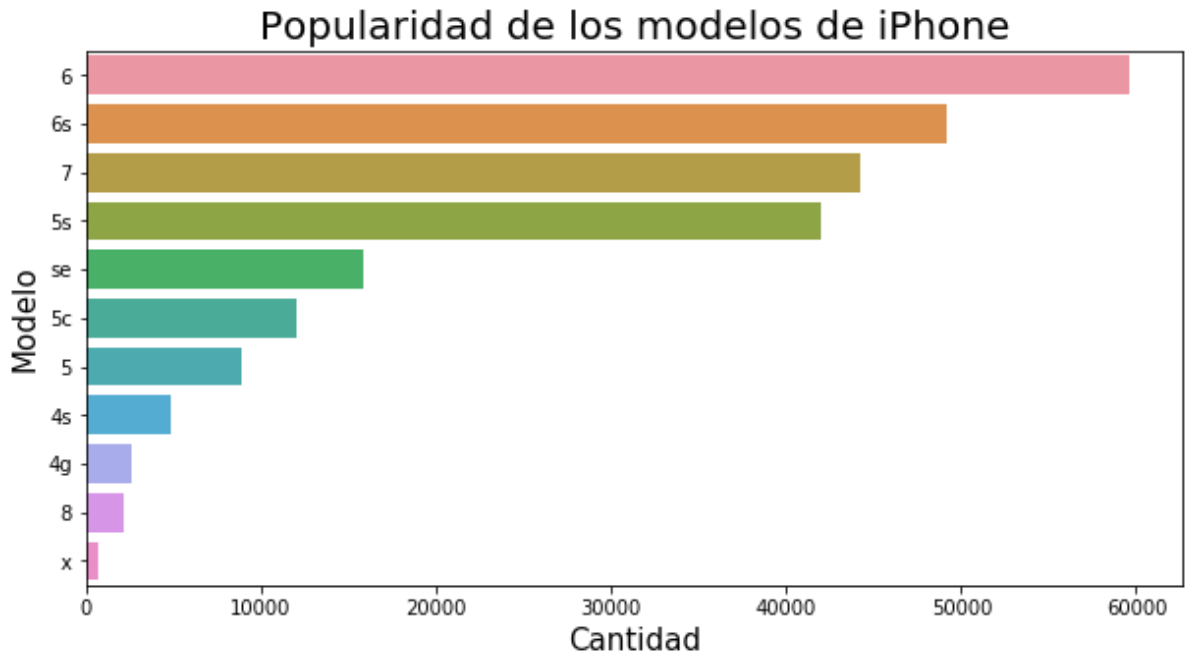


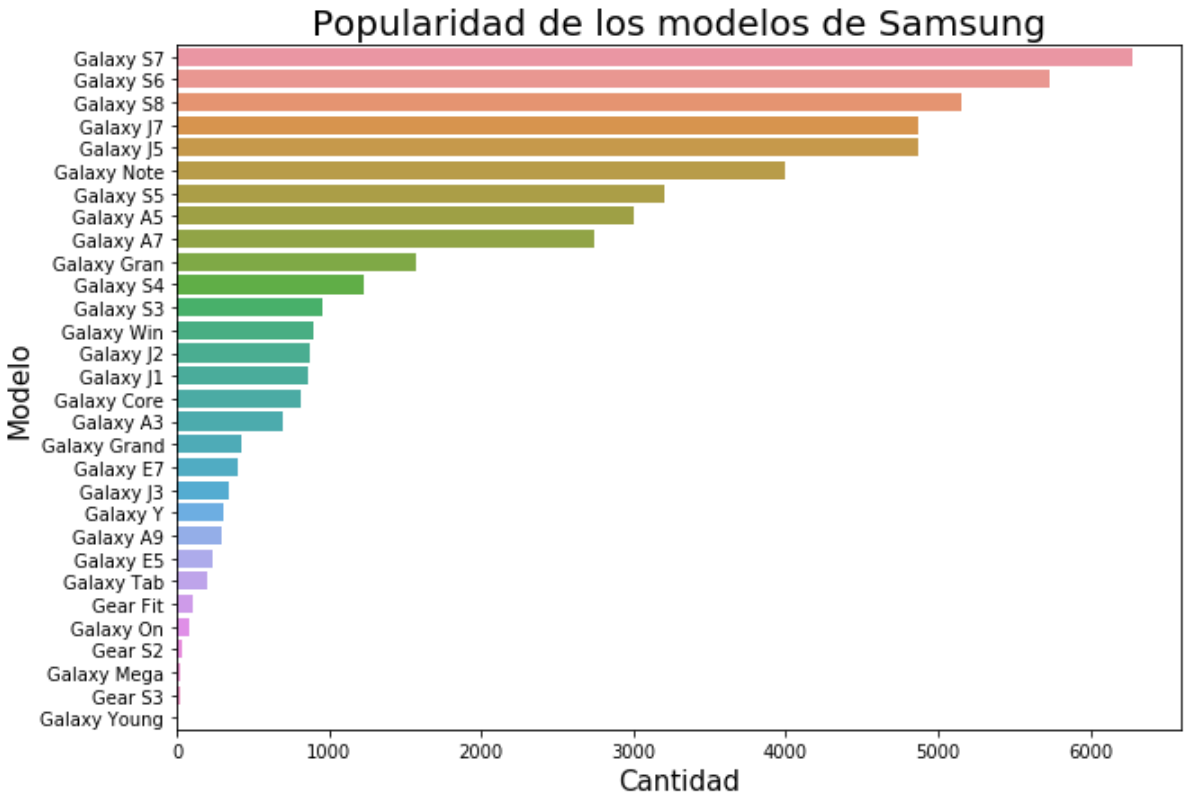

La cantidad de dispositivos de Samsung vendidos supera, casi por el doble, a iPhone, cuando antes iPhone era la preferida a la hora de ver los productos. Esto puede deberse al precio de los productos de Apple, que por lo general presentan un costo superior a dispositivos de características similares pertenecientes a otras marcas. Por esto se supone

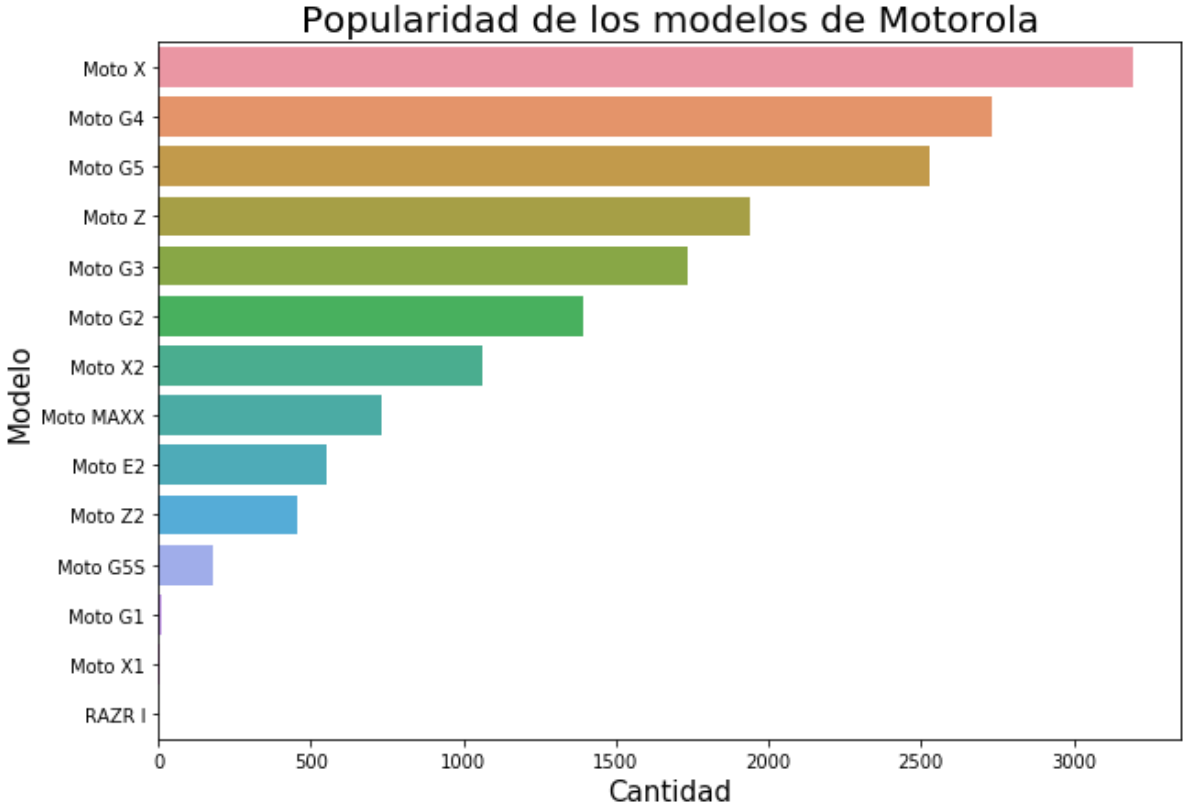
que en el caso de comprar un dispositivo pocos usuarios prefieren la popularidad del producto a la relación calidad-precio.
Podemos pensar a partir de esto que los iPhone pese a ser una marca bastante popular entre los usuarios a la hora de comprar pierde bastante de su popularidad.

### 3.5.2 Modelos populares

Para las marcas populares, las cuales son iPhone, Samsung y Motorola, se analiza los modelos más populares.




Popularidad de los modelos de Samsung
Galaxy S7
Galaxy S6
Galaxy S8
Galaxy J7
Galaxy J5
Galaxy Note
Galaxy S5
Galaxy A5
Galaxy A7
Galaxy Gran
Galaxy S4
Galaxy S3
Galaxy Win
Galaxy J2
Galaxy J1
Galaxy Core
Galaxy A3
Galaxy Grand
Galaxy E7
Galaxy J3
Galaxy Y
Galaxy A9
Galaxy E5
Galaxy Tab
Gear Fit
Galaxy On
Gear S2
Galaxy Mega
Gear S3
Galaxy Young
Modelo
0
1000
2000
3000
4000
5000
6000
Cantidad

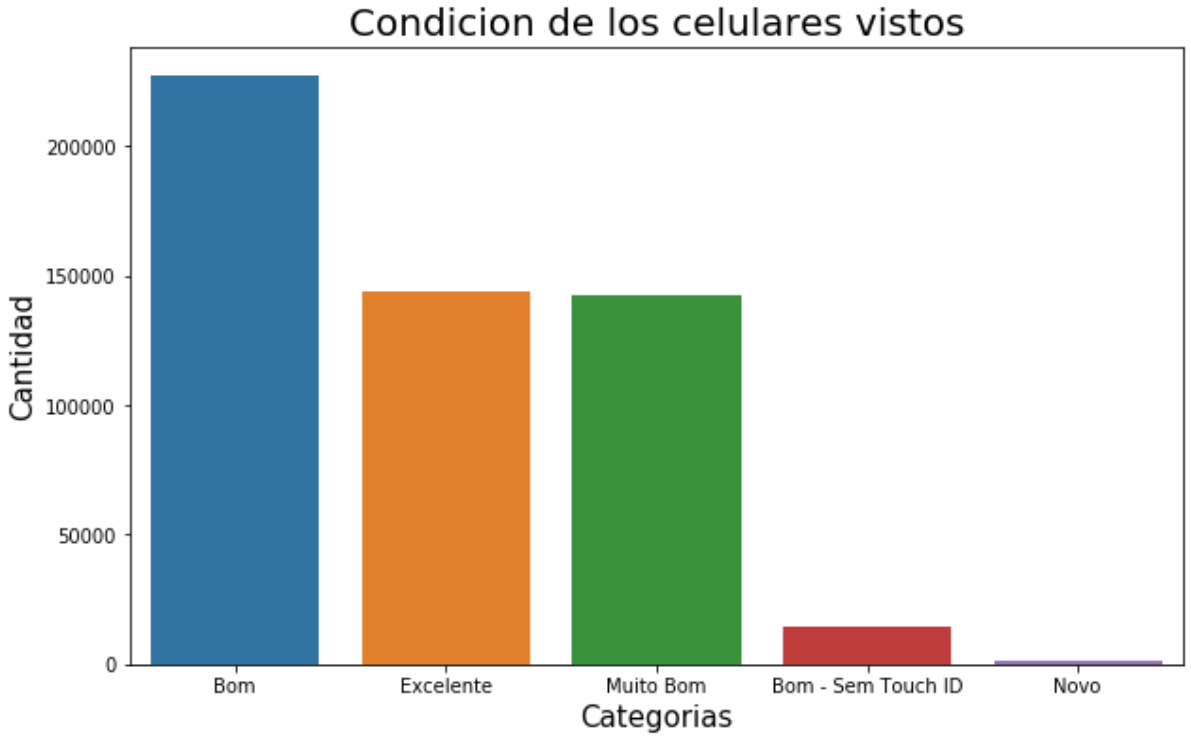

Popularidad de los modelos de Motorola
Moto X
Moto G4
Moto G5
Moto Z
Moto G3
Moto G2
Moto X2
Moto MAXX
Moto E2
Moto Z2
Moto G5S
Moto G1
Moto X1
RAZR I
Modelo
0
500
1000
1500
2000
2500
3000
Cantidad

## 3.5.3 Análisis de características de los dispositivos más populares

Es importante analizar las características de los celulares más populares, para eso se estudian storage, condition y color.

### 3.5.3.1 Condicion

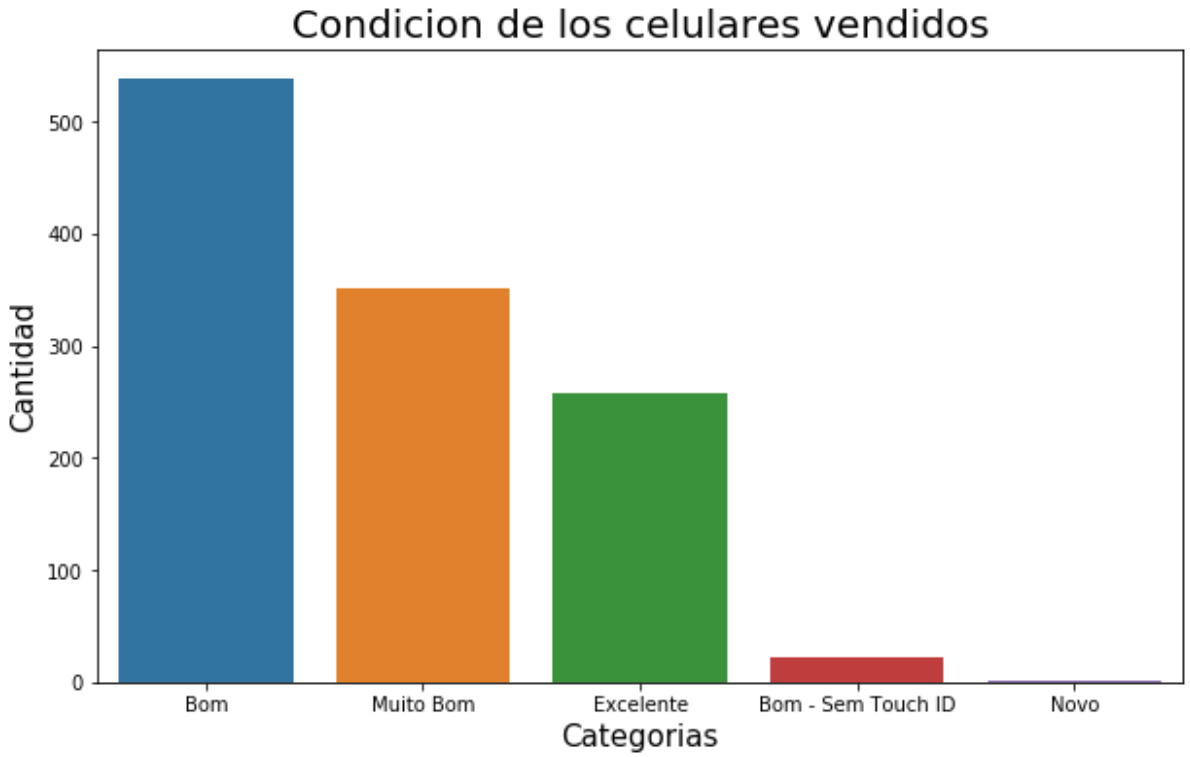
La condición hace referencia a la condición de venta del producto. Dado que Trocafone tiene como objetivo la venta de dispositivos usados, se los evalúa en tanto la estética y se los califica según su condición. Los valores posibles son: Excelente, Muy buena, Buena, Buena sin Touch ID, Nuevo.
Se analiza en qué condición se encuentran aquellos dispositivos más vistos por los usuarios:




Tiene sentido que la cantidad de dispositivos nuevos sea baja, dado que el sitio está destinado a reciclar dispositivos.
Se puede ver que los celulares que más le interesan al usuario son aquellos que pertenecen a la categoría “bom” (bueno). Aún más que aquellos que se encuentran en condiciones “excelente” o “muito bom” (muy bueno). Se podría suponer a priori sin ver estos resultados que buscarían los que están en mejores condiciones, pero hay que tener en

cuenta que es un sitio de venta de dispositivos usados. Es probable que lo que busquen es un buen precio, pero que aún así lo que compren esté en buen estado para su uso.

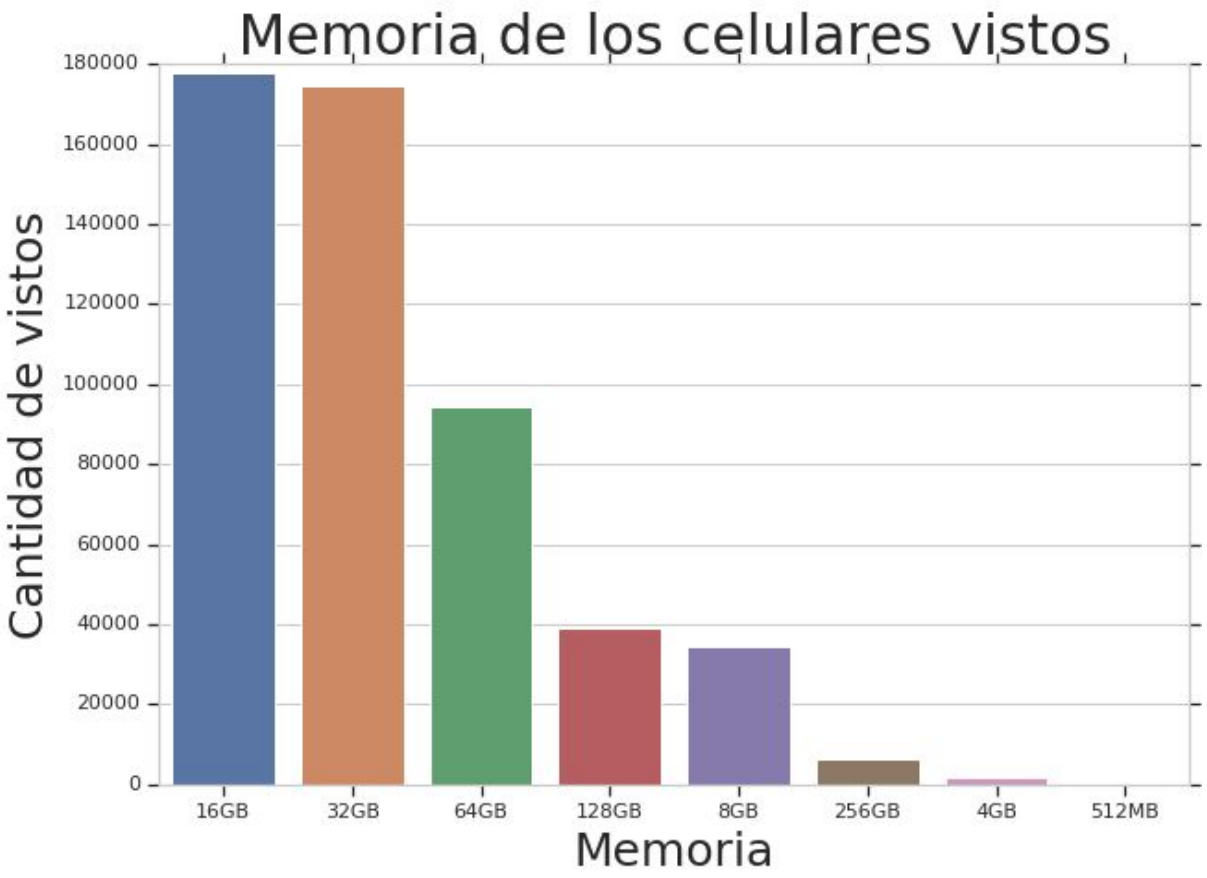
Veamos si sucede lo mismo con las compras:



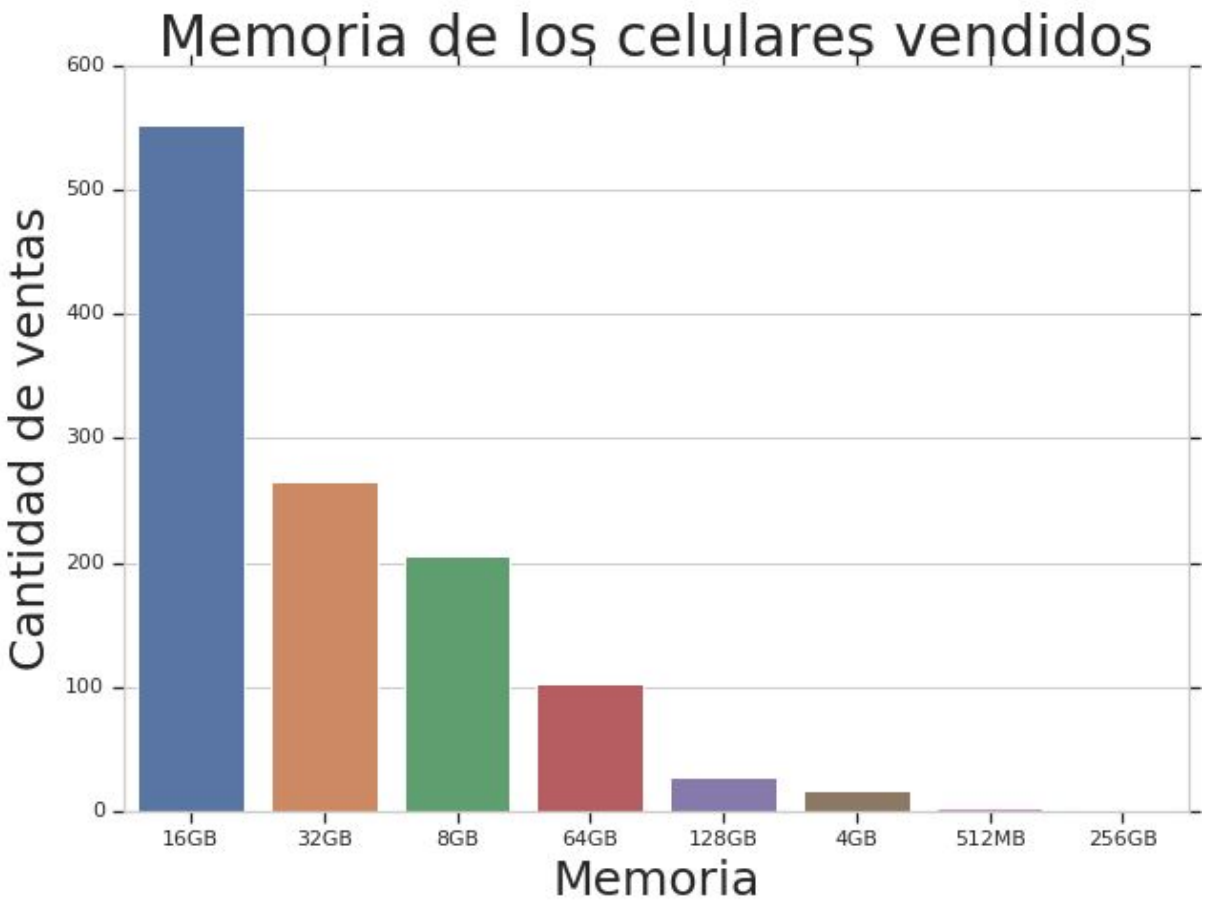

Efectivamente se ve el mismo comportamiento. En este caso en cambio, aumenta la cantidad de “muy bueno” sobre “excelente”. Esto puede suceder por el mismo fenómeno que hace que los “bueno” sean los más populares, el precio se supone más barato.

### 3.5.3.2 Memoria

Otra característica importante a analizar es la cantidad de memoria del dispositivo.
Veamos qué cantidad de memoria tienen los que más ven los usuarios:




El más popular es 16GB, con 32GB siguiéndolo muy de cerca y en tercer lugar 64GB, pero ya con algo más de la mitad de los dos anteriores. El resto ya aparece con muy poca cantidad.

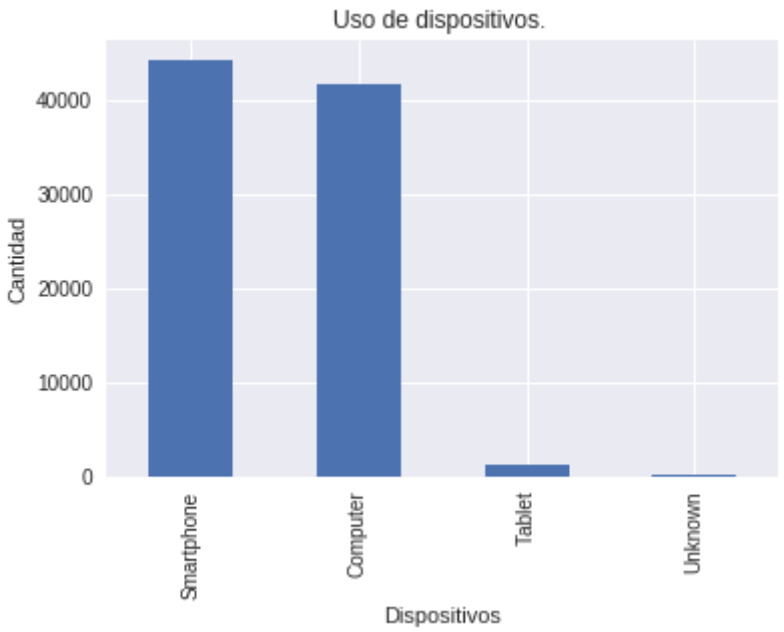



Hay una clara diferencia entre los de 16GB y el resto. No pasa lo mismo que con el caso anterior de los que más ve el usuario. Cuando se trata de las ventas el 16GB está primero (cómo estaba primero con las vistas), pero los de 32GB bajan bastante manteniendo su segundo lugar y los de 64GB bajan al 4to puesto dejando el 3ro a los de 8GB.
De esto podemos concluir que la mayoría de los usuarios prefieren pagar menos y tener una memoria de 16/32GB en vez de mucha más memoria. Aún así se mantienen dentro de un rango “aceptable” de cantidad de memoria, ya que vemos que tampoco son populares los de muy baja memoria.
Al igual que lo que pasa con la condición del dispositivo, se busca un balance entre calidad y precio.

# 3.6 Análisis del de dispositivo usado por el usuario

## 3.6.1 Dispositivo más frecuente

Luego del correspondiente análisis se llega a la conclusión de que el dispositivo más utilizado por el usuario para ingresar al sitio es el celular. Esto es de esperarse dado que el uso del celular es más frecuente a lo largo del día dado que es más portable que una computadora . Dicho resultado se puede ver a continuación:




# 3.7 Comportamiento del usuario

Se considera que cada id diferente es un usuario diferente y se agrupa los eventos de distinto tipo generados por cada uno de estos obteniéndose los siguientes datos.

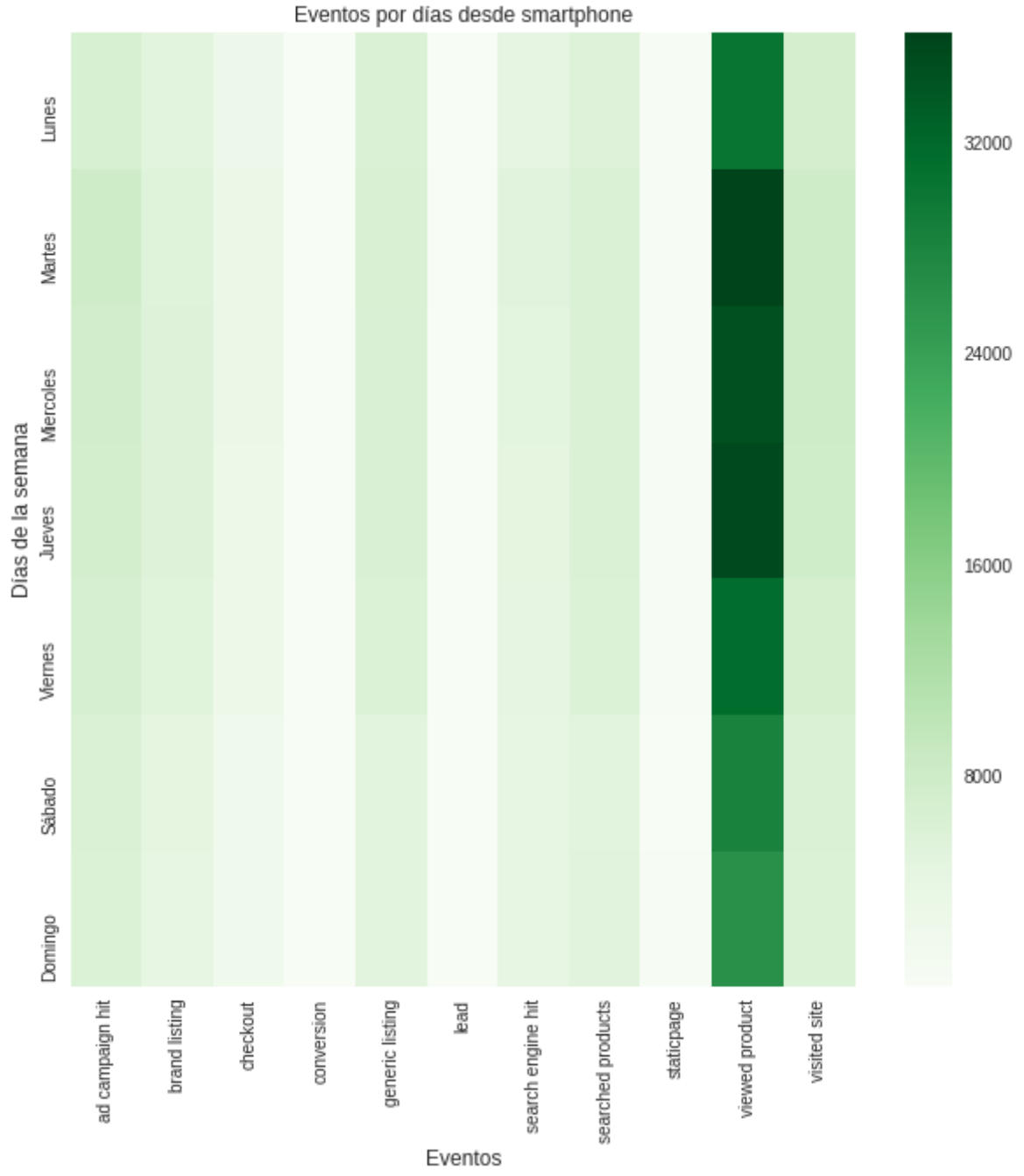
| | ad campaign hit | brand listing | checkout | conversion | generic listing | lead | search engine hit | searched products | staticpage | viewed product | visited site |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **mean** | 2.998371 | 3.570627 | 1.221221 | 0.042427 | 2.444758 | 0.016218 | 1.844664 | 2.029865 | 0.130249 | 19.147517 | 3.163119 |
| **std** | 7.235963 | 16.253314 | 1.215497 | 0.345757 | 7.154502 | 0.210565 | 3.877810 | 8.323393 | 0.876540 | 51.142526 | 6.873227 |

Los datos corresponden al promedio y a la desviación estándar de la cantidad de eventos de cada tipo generados por cada usuario.
Las desviaciones de los eventos con mayor frecuencia suelen tomar valores altos. Más aún para el evento ‘viewed product’, por lo cual se busca otra manera de visualizar este fenómeno.

## 3.7.1 Comportamiento a lo largo del tiempo según dispositivo

Se procede a realizar un análisis del comportamiento de los usuarios según el día y el dispositivo.

En primer lugar se observa que la actividad principal desde cualquier dispositivo de ingreso es *“viewed product”,* se presenta a continuación un gráfico de actividad por día y mes de usuarios que ingresan con un Smartphone:

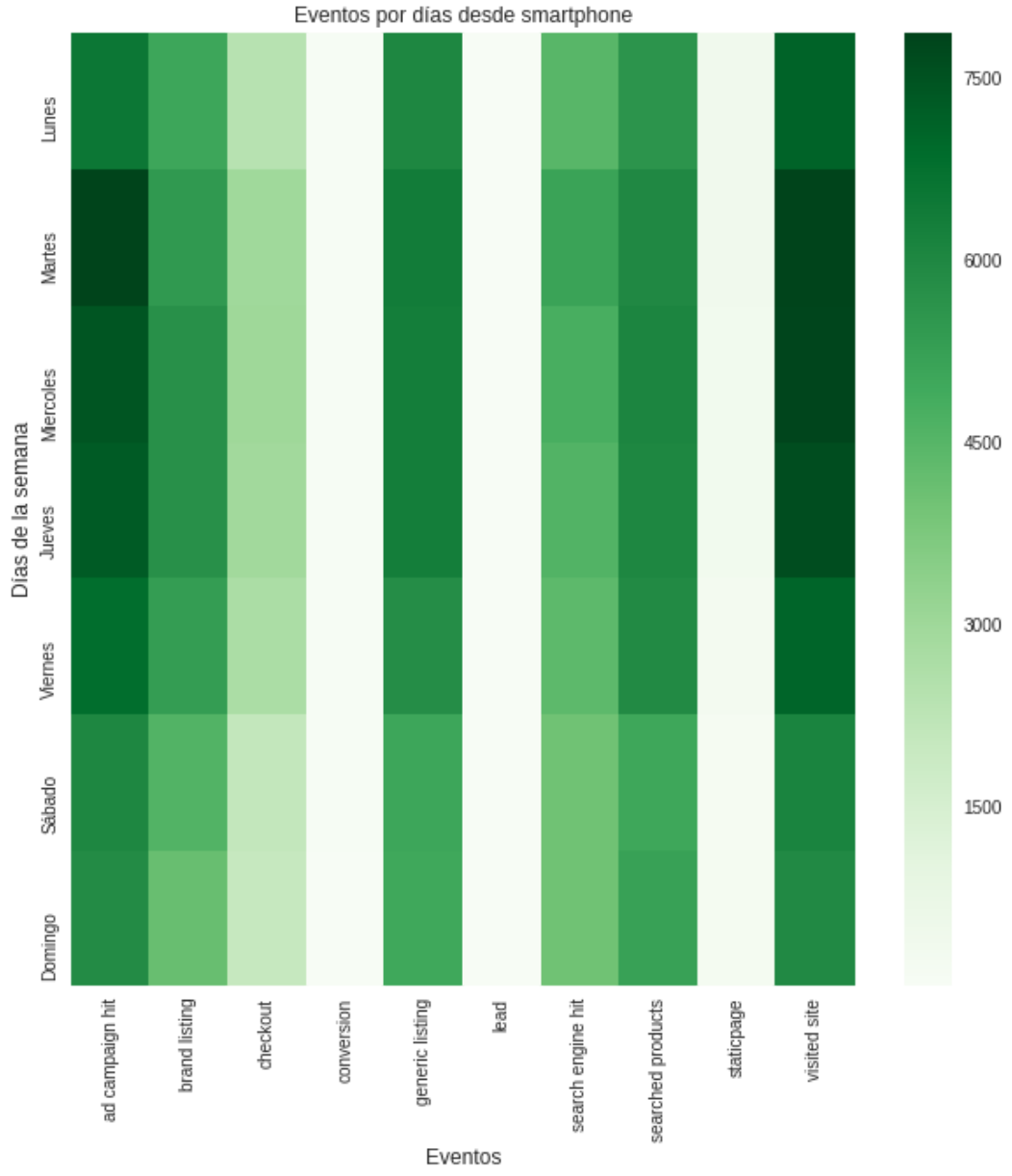



Se observa que la cantidad de veces que un usuario ve un producto es mucho mayor que la cantidad de veces que se genera cualquiera de los otros eventos.

Dada esta predominancia en la cantidad de vistas por sobre los demás eventos, la cual se repite en todos los dispositivos por igual, se procede a excluir este evento de las siguientes visualizaciones para determinar diferencias en el comportamiento.

Se obtiene:

- Usuarios desde Smartphone:



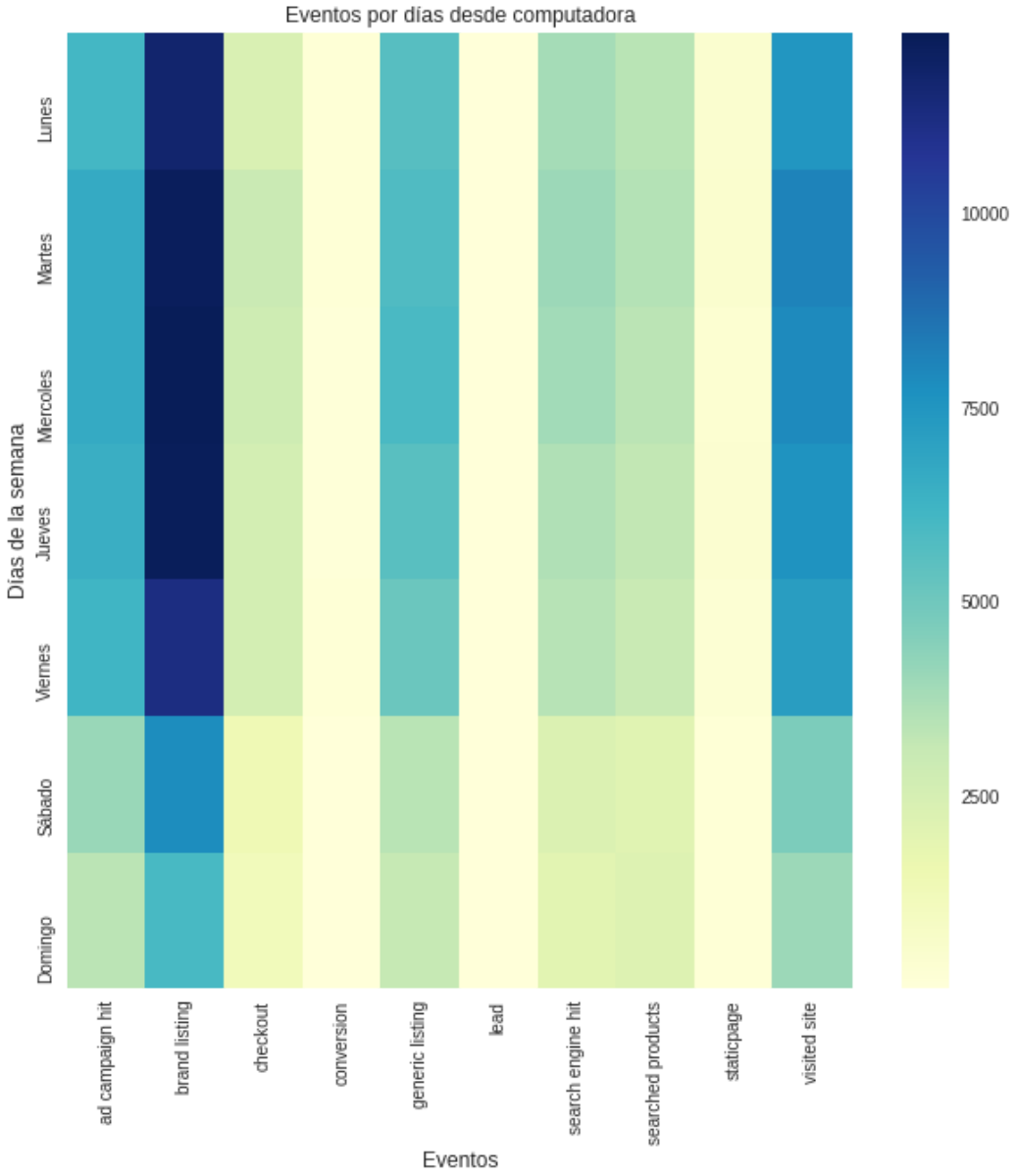

Se observa que las actividades principales son:

- ad campaign hit

- brand listing
- generic listing
- search engine hit
- searched products
- visited site

Y que los días de más actividad son de martes a viernes.

- Usuarios desde Computadora:




En este caso, para usuarios que entran desde una computadora, los eventos comunes son:

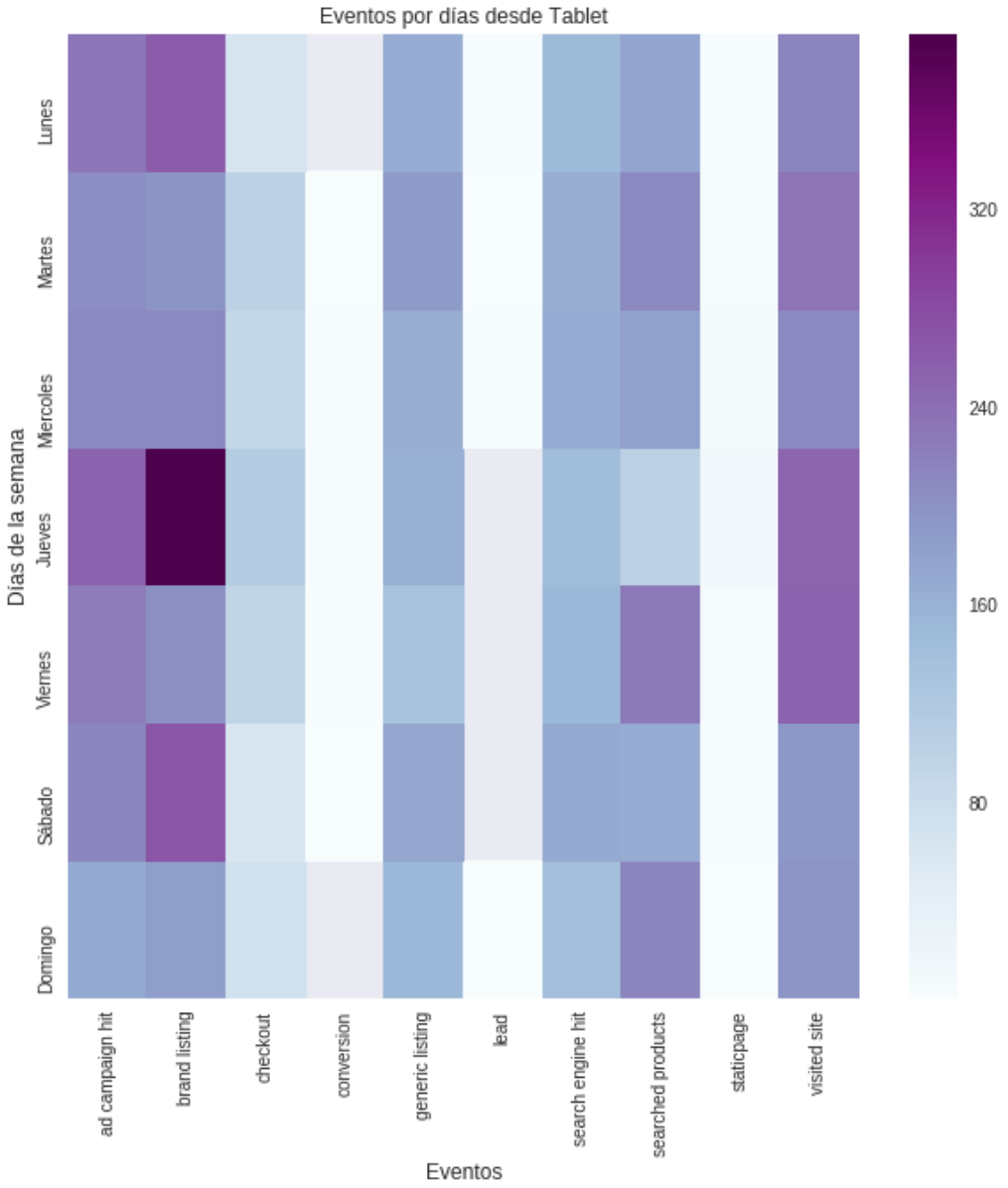
- ad campaign hit
- brand listing
- visited site

Y en menor medida

- generic listing

Se puede observar que estos eventos se dan en mayor medida que cuando el usuario entra desde su celular personal.
brand listing es la actividad principal.

- Usuarios desde tablet.




Las actividades principales son:

- ad campaign hit
- brand listing
- visited site

Y en menor medida

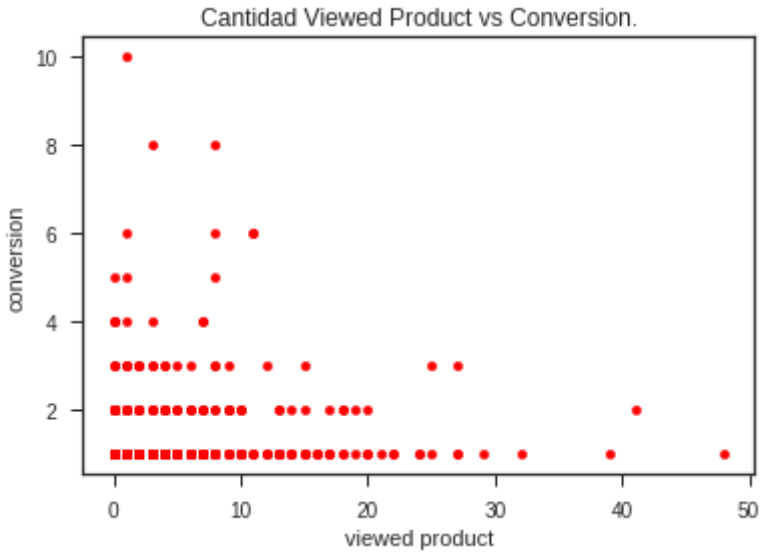
- generic listing
- searched product
- search engine hit

En este caso la distribución de actividades es más pareja con respecto a la entrada desde otros dispositivos.

# 3.8 Relación entre eventos

## 3.8.1 Relación vista y compra de productos.

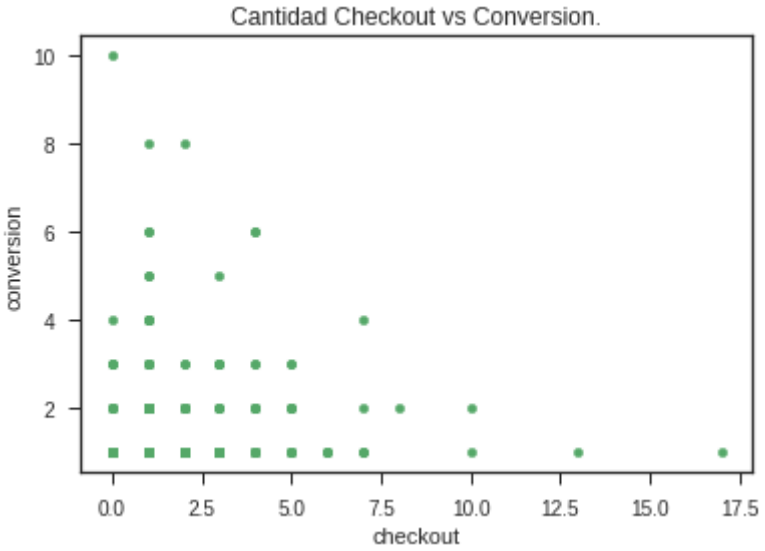
Se estudia la relación entre las veces que el usuario ingresa a ver el producto y la cantidad de compras que realiza.
Para ello se agrupa por persona e identificador de producto y se cuenta la cantidad de veces que se realiza cada uno de los eventos. El gráfico de dispersión que se obtiene para esta relación es el siguiente:




En dicho gráfico se observa que un usuario generalmente ve un producto entre 0 y 20 veces antes de realizar la compra de un celular. Existen varios casos en los que el usuario ingresa directo a la compra o al checkout, y muy poco en los que o bien ven muchos productos antes de comprar, como se observa en los puntos que quedan alejados en el eje horizontal, y otros en los que ven el producto reducidas veces pero compran una cantidad más grande al promedio. Se puede pensar que son valores outlier en el set de datos.
Por la distribución de los puntos se puede aproximar que la relación cantidad de productos vistos - productos comprados es logarítmica.

### 3.8.2 Relación Checkout y Compra de productos.

Para ver la relación entre la cantidad de productos que el usuario envía al checkout con la cantidad de productos comprados se procede a agrupar por persona y sku obteniendo como resultado:




Se observa que la distribución se concentra en valores bajos, hay como mucho 5 checkout antes de realizar de una a 3 compras, aunque puede ser frecuente que un usuario lleve hasta 7 producto a checkout antes de comprar uno.
Al igual que en la relación vistas-compras tenemos outliers donde por ejemplo los usuarios llegan a colocar en el check casi 17 productos para realizar una compra.
Como el caso anterior se puede decir que la distribución es casi logarítmica.

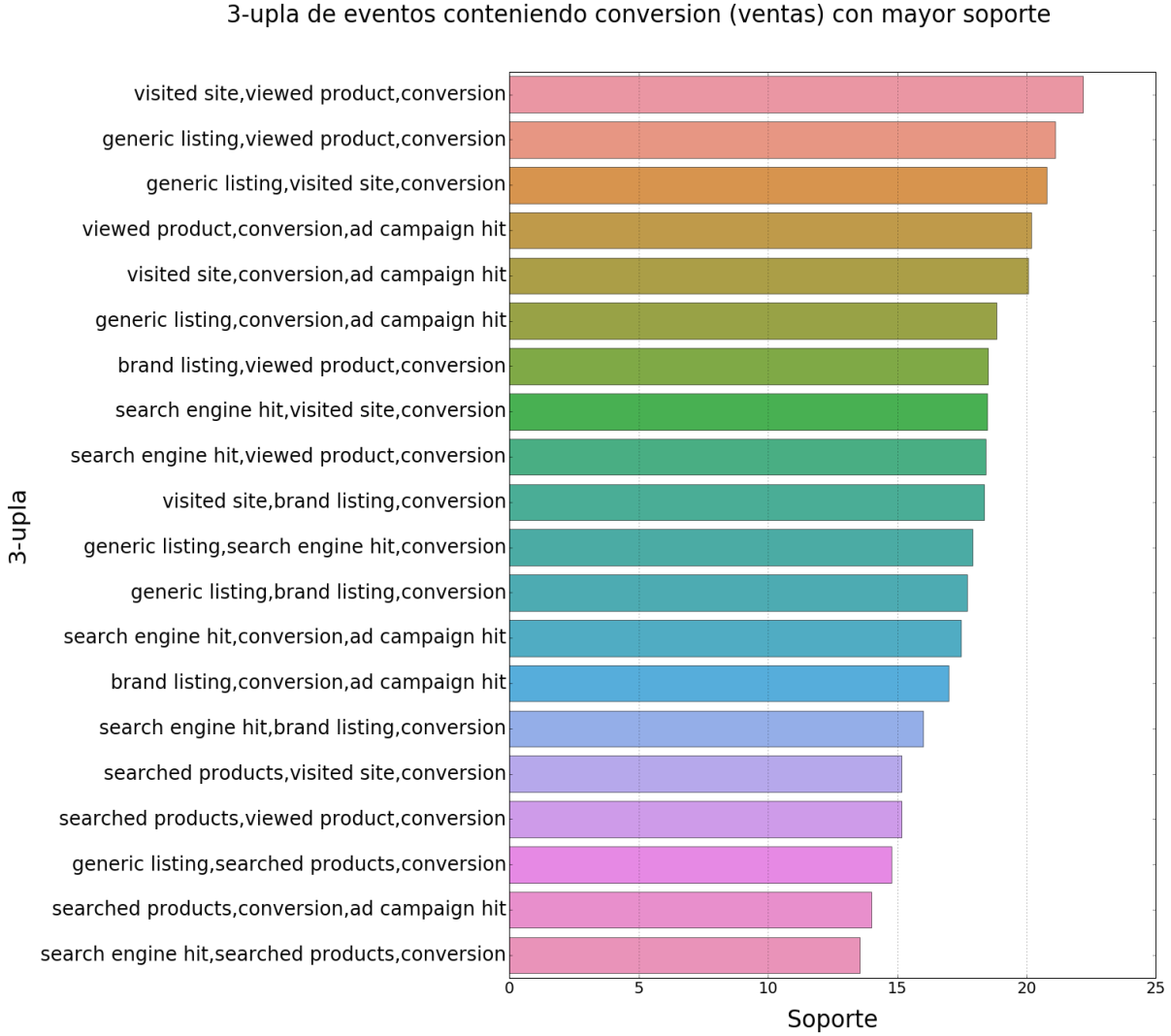
## 3.9 Nivel de cercanía entre los distintos eventos.

Una de las cosas que más nos puede interesar en lo que es comportamiento del usuario es cómo se va manejando en el sitio. Con esto queremos decir, qué clases de eventos va haciendo a lo largo de su sesión. Para esto podemos estudiar las relaciones entre los eventos.
Decidimos obtener esta información con el uso del algoritmo apriori, que detecta y obtiene las n-uplas más frecuentes en un set de datos. Necesitamos hacer uso de la librería “mlxtend” de Python que tiene una implementación del mismo.

### 3.9.1 Tuplas de eventos con mayor soporte

Con soporte nos referimos a una medida que indica qué tan frecuente aparece la n-upla en nuestro set de datos (soporte * 100 = porcentaje de aparición en el set de datos).
Otra cosa a tener en cuenta en la longitud de las n-uplas que tomaremos, en este caso tomaremos de longitud 3.

Ahora, analizamos aquellas tuplas que contengan el evento “conversión” (ventas), ya que nos interesa ver el comportamiento de un usuario que llega a hacer una compra. Por otra parte, hace falta aclarar que no tuvimos en cuenta para esta sección del análisis las tuplas que contenían un checkout, ya que todo usuario debe de pasar por este para llegar a hacer un evento de conversión que es lo que nos interesa estudiar en este apartado.




Lo primero que vemos es que el nivel de apariciones de las tuplas que contienen el evento de venta son bajos (menos del 25%), esto tiene sentido ya que como vimos anteriormente las ventas son muy pocas en comparación con otros eventos.
Otra cosa que podemos ver es que es bastante común entre los usuarios hacer un visited site + viewed product + conversion. Pero más importante es que las siguiente dos tuplas que aparecen contienen un “generic listing” entre los eventos que se realizan. Y es menos común que se haga un “brand listing”.
Recordar siempre que estamos hablando de eventos que se realizan en conjunto con algún evento de conversión. Esto quiere decir que puede ser que el “generic listing” no sea más común que el “brand listing”, sino que es más común cuando se los ata con la conversión.
Otra cosa que notar es que aparecen varios “ad campaign hit” asociados a las conversiones. Esto nos puede decir que muchos de los usuarios que ingresaron por medio

de una campaña realizaron una compra. Esto es muy importante para tener en cuenta, ya que nos da un indicio de que lo que quisimos ver en la sección de crecimiento a través del año es verídico. Recordemos qué es lo que vimos en ese apartado: primero notamos un crecimiento de la actividad en los últimos meses, al mismo tiempo que vimos un crecimiento de usuarios nuevos y de usuarios que ingresan por un medio de campaña. Después de estudiar eso mostramos que el crecimiento de las ventas crecía de la misma forma. Ahora, esto no nos decía mucho, ya que era probable que el crecimiento de las ventas no esté relacionado con el crecimiento de ingresos por campaña. Acá podemos ver que esto se cumple en gran medida. Ya que nos aparecen varias tuplas que contienen el “ad campaign hit” y “conversion”

# 3.10 Análisis de término de búsqueda

A continuación se muestra un wordcloud indicando la frecuencia de los términos de búsqueda utilizados por los usuarios del sitio.


Figura 7. Wordcloud de términos de búsqueda.

# 4. Conclusión

Al concluir el análisis exploratorio se destacan ciertos puntos de utilidad.

- Del comportamiento del usuario se puede destacar que el usuario promedio ingresa al sitio por medio de varios canales, siendo el más frecuente el de ‘Paid’ que corresponde a campañas pagas. Por lo cual estos usuarios ingresan al sitio directamente a la página del producto en cuestión. De hecho, lo que pudimos observar es que una buena parte de los usuarios que entraron por este medio efectivamente hicieron una compra en el sitio. Además es de donde se ve la mayor parte del aumento de la actividad y de compras en el sitio. Por lo que vale la pena estudiar este campo, ya que vemos que son efectivas.

- Analizando los productos más populares, tanto marca como modelo, se pueden mejorar las campañas. Y estudiando el comportamiento de un usuario en particular personalizar dicha campaña según sus preferencias, atrayendo así a mayor cantidad de usuarios al sitio. Para esto se recopila información de terminos de busqueda, productos vistos y movimiento del usuario a través del sitio.

- A pesar de que Trocafone tiene base en Brasil hay usuarios provenientes de otras partes del mundo. Analizando mediante qué canales tienen contacto con el sitio se pueden generar campañas para aumentar la cantidad de usuarios extranjeros en el sitio.

Dentro del análisis se pueden ver otros resultados interesantes que tal vez no tienen una utilidad a la hora de mejorar el servicio.