FACULTAD
DE INGENIERIA
Universidad de Buenos Aires

Organizacion de Datos 75.06/95.58

Trabajo práctico N°1

October 8, 2018

Alumno	Padrón
Ivo, Biaus	98708
Toscano, Miguel	98385
Labaisse, Hugo	103624
Perrone, Patricio	98230

Repositorio: https://github.com/IvoBiaus/TP1/

Contents

1 Introducción

Análisis Exploratorio de Datos: Trocafone realizaremos un análisis de datos sobre un conjunto de eventos de web analytics de usuarios que visitaron www.trocafone.com, su plataforma de ecommercede Brasil. Trocafone es un side to side Marketplace para la compra y venta de dispositivos electrónicos que se encuentra actualmente operando en Brasil y Argentina. Este set de datos posee información de alrededor de 1000000 de eventos de Trocafone. Nuestro objetivo será realizar un análisis exploratorio sobre esa información, para intentar obtener algunos insights de de la misma

Como herramientas se usó Jupyterlab, un entorno de desarrollo para Data Science, particularmente en lenguaje Python.

2 Objetivo

Mediante los análisis que se presentan a continuación, se buscó contestar a las siguientes preguntas:

¿ Vale la pena invertir en la creación de una aplicación móvil?

¿ Hay algún sector en el cual se podría ahorrar presupuesto?

¿ Se podría mejorar algún aspecto de la página web de la empresa?

3 Desarrollo

3.1 Analizando propiedades especificas del set de datos

Para poder comenzar a orientar nuestro análisis comenzaremos a analizar algunas variables que nos interesan para aplicar en nuestros análisis. Comenzaremos con los tipos de eventos.

De esta forma obtenemos la cantidad de valores que hay para cada uno de los

tipos de evento.

Tipo de evento	Cantidad de eventos
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

Tabla 1. Cantidad de cada tipo de evento

3.2 Una primera mirada a los datos

Como primer paso, se cargaron los datos desde un archivo de formato *csv* a un *DataFrame* (estructura propia de pandas) para un primer análisis.

En dicho *DataFrame*, se encontraban las siguiente columnas:

- timestamp: Fechar y hora de cuando ocurrio el evento.
- 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 o por un usuario que previamente había visitado el sitio 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 generó el evento.
- screen_resolution: Resolución de pantalla que se está utilizando en el dispositivo desde donde se generó el evento.

- operating_system_version: Versión de sistema operativo desde donde se originó el evento.

- browser_version: Versión del browser utilizado en el evento.

3.3 Verificación de Calidad de Datos

Verificamos la cantidad de valores nulos. Teniendo en cuenta que el *DataFrame* otorgado tenia 1011288 filas, la siguiente tabla nos muestra lo previamente mencionado:

Columna	Valores nulos
timestamp	0
event	0
person	0
url	928532
sku	447450
model	447004
condition	447452
storage	447452
color	447452
skus	789589
search_term	962321
staticpage	1007690
campaign_source	928492
search_engine	960331
channel	923910
new_vs_returning	923910
city	923910
region	923910
country	923910
device_type	923910
screen_resolution	923910
operating_system_version	923910
browser_version	923910

Tabla 2. Cantidad de datos nulos por columna

Se puede observar que gran parte de las columnas son en su mayoría datos nulos, sin embargo esto no es razón suficiente para descartarlas (las columnas), ya que si analizamos un poco los datos, se puede ver que no a todos los eventos les corresponden la misma información, y que no significa que por tener valor nulo implique que sea un valor erróneo. Por lo tanto habrá que relacionar los datos, posiblemente a través de merge o joins de diferentes tablas.

3.4 Linea Temporal

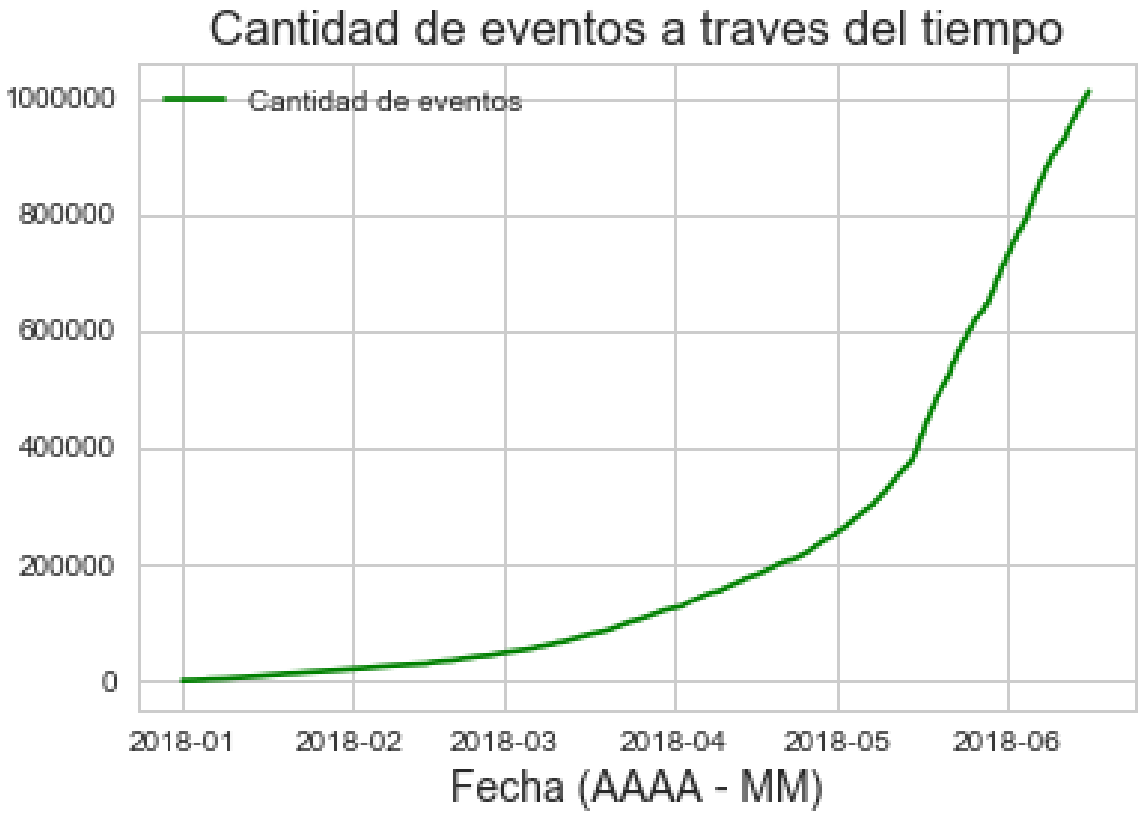
Vemos entre que fechas exactas ocurren los eventos del dataset.

min 2018-01-01 07:32:26

max 2018-06-15 23:59:31

Acá podemos ver que los eventos ocurren durante los primeros 6 meses del 2018, pero del mes 6 (Junio) solo están los primeros 16 días, por lo que tenemos la mitad de la información a comparación de los otros meses. Este detalle tendrá que ser tenido en cuenta en los casos que queramos hacer análisis de cantidades agrupando por mes, ya que Junio solo dispondrá de la mitad de los datos.

4 Análisis de eventos

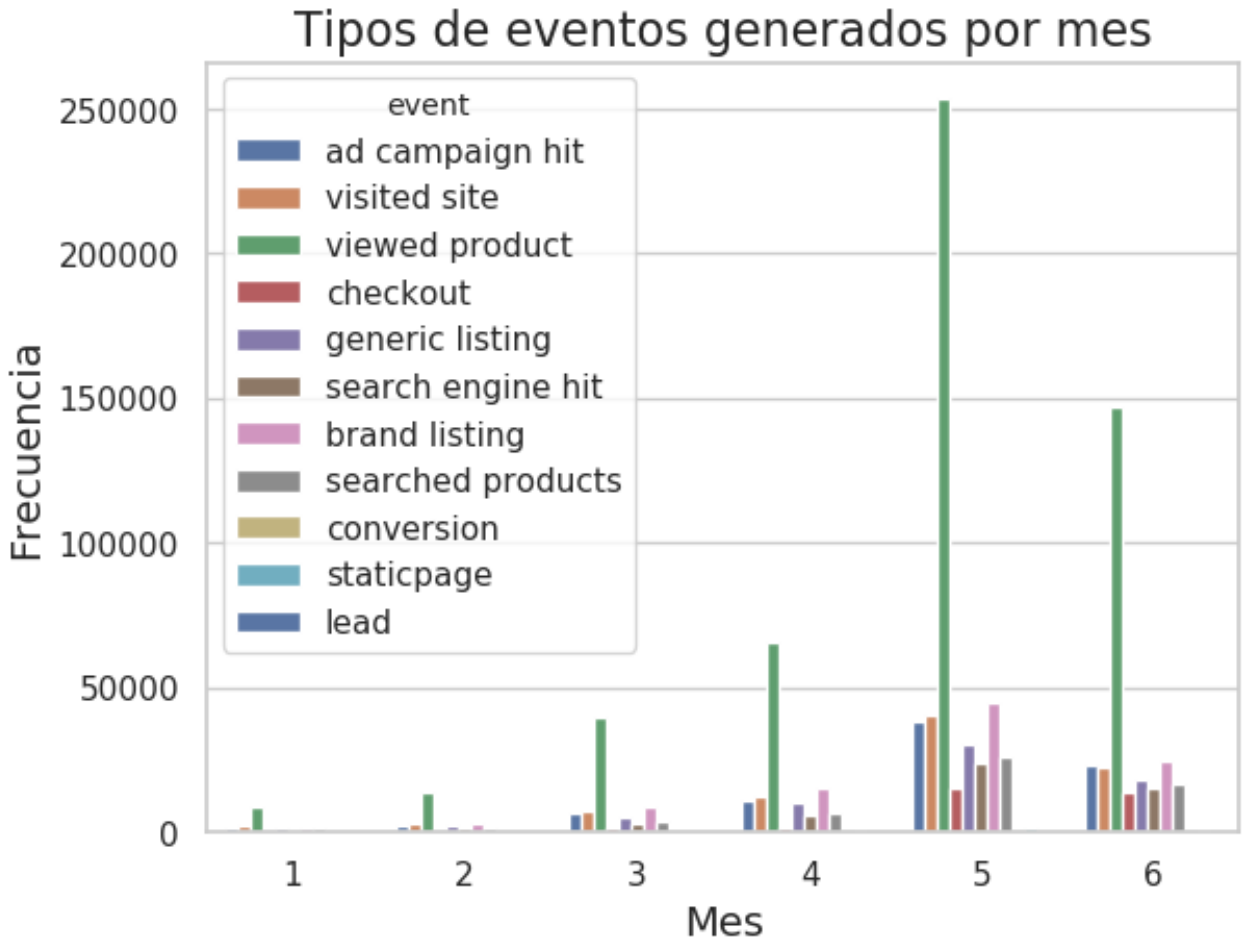


Se puede ver la cantidad de eventos a lo largo del año crece firmemente de manera exponencial, lo que muestra que mes a mes se producen mas eventos durante los mismos periodos de tiempo.

Por lo que podemos ver desde Enero hasta mediados de abril (3 meses y medio) se llevaron a cabo alrededor de 200.000 eventos, mientras que en los siguientes 2 meses (desde mediados de Abril a mediados de Junio) se llevaron a cabo alrededor de 800.000 eventos. Con lo que muestra un gran crecimiento en la actividad de la pagina, algo que es muy bueno.

4.1 Análisis de los tipos de evento y evolución

A continuación procederemos a analizar la cantidad de cada tipo de eventos realizados y su evoluciona a través de los meses.



Se puede observar que la relación entre cada tipo de evento se mantiene mes a mes, y que todos crecen simultáneamente. Con una amplia diferencia entre el evento 'viewed_product, que conforma la gran parte de los eventos, y los otros eventos. Por otro lado, los eventos con menos frecuencia son al parecer las: 'conversion' (ventas), 'staticpage' y 'leads'.

En este gráfico podemos observar mejor como los eventos se separan, en lo que podríamos decir que 3 grupos.

Por una parte el eventos 'viewed product' que crece exponencialmente y acumula eventos con una amplia diferencia por sobre el resto , con cantidades que hasta el día final del set de datos llegan a ser de mas de 5 veces que cualquiera de los otros.

En un grupo intermedio tenemos los siguientes 7 tipos de evento ('brand listing', 'visited site', 'ad campaing hit', 'generic listing', 'searched products', 'search engine hit' y 'checkout'), los cuales crecen muy similarmente por lo que posiblemente están relacionados, no quita que igualmente hay diferencia entre la frecuencia de cada uno. Previamente se los menciono en el orden descendiente, con lo que por ej 'brand listing' tiene alrededor del doble de eventos (posee unos 100mil) que los eventos de 'checkouts'.

Por ultimo tenemos las 3 categorías de eventos con menos frecuencia, estas son 'staticpage', 'conversion' y 'lead', los cuales a pesar de generar eventos, en los gráficos no lo parece, dado que los números que estos generan son demasiado chicos para la escala de eventos que generan los demás.

5 Análisis de Marketing

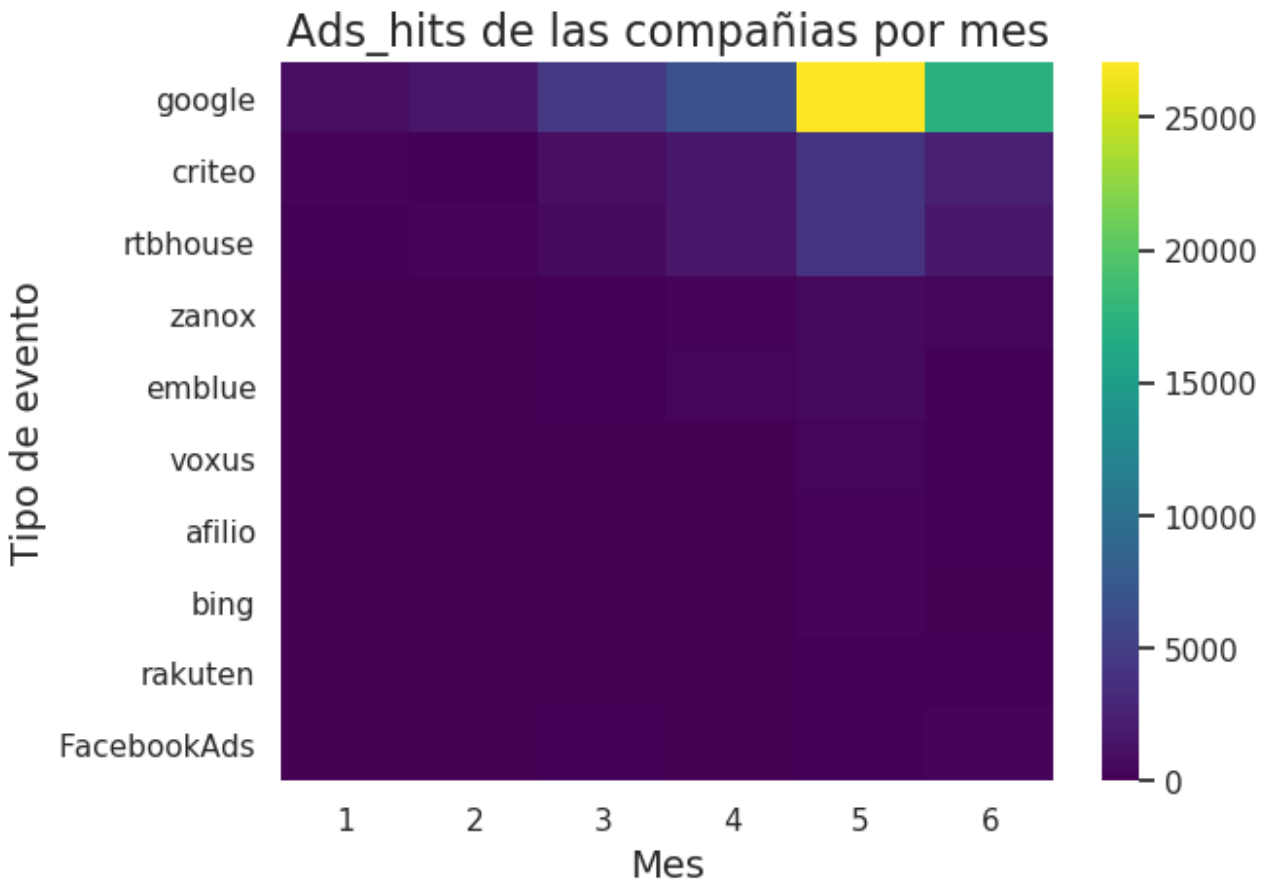
A continuación filtraremos algunas columnas que en general son información sobre el dispositivo del usuario, y algunas otras,que en su mayoría son NaN y no hacen diferencia al análisis que haremos.

Dado que vamos a analizar la relación entre entradas a través de ads y diferentes tipos de evento esas columnas no las vamos a necesitar. Si vamos a quedarnos con algunas columnas que posiblemente no necesitaremos, pero no las vamos a borrar por si acaso las necesitamos en algo que no estemos teniendo en cuenta.

5.1 Ads hits por empresa en cada mes

Queremos ver la evolución de los hits de cada fuente de publicidad a través de los meses, asique realizaremos un heatmap.

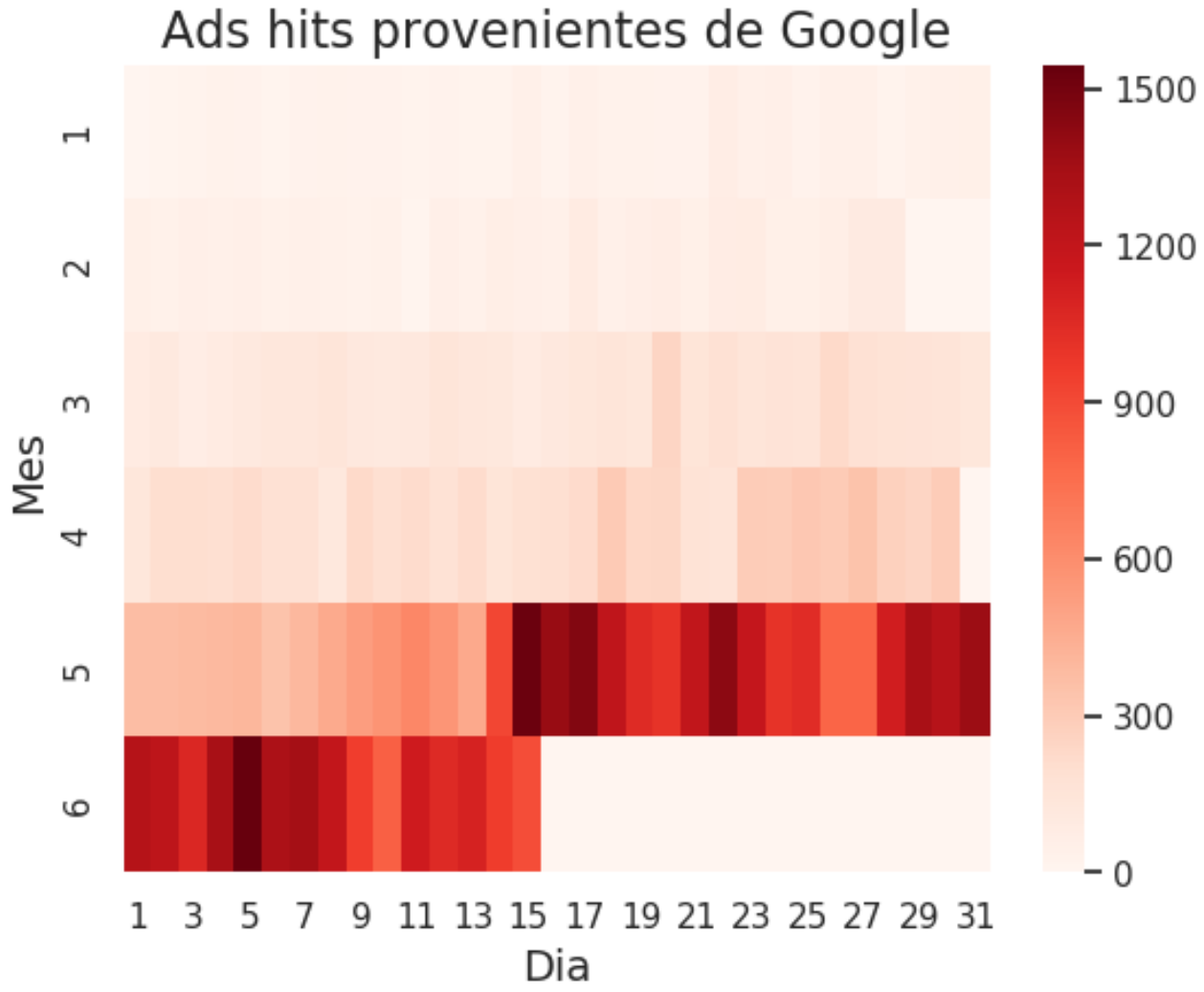
Dado que la cantidad de servicios de publicidad eran demasiados y la mayoría no aportaban mucho al heatmap por la falta de 'ads hits' procedimos a descartar aquellos que durante el mes que mas se destaco por el rendimiento (Mayo,5) tuvieron menos entradas, dado que ordenamos por la cantidad y tomamos los mejores 10.



En el este gráfico se puede observar fácilmente que las campañas publicitarias por las que mas usuarios ingresan a la pagina son originadas por los servicios de Google, Criteo, y Rtbhouse, que claramente sobresaltan sobre los demás, y dentro de este grupo de 3, Google es claramente la fuente de mas ingresos a la pagina. Todos van aumentando conforme los meses avanzan y caen en el mes 6 pero esto es por lo que vimos previamente sobre los datos incompletos del mes de Junio.

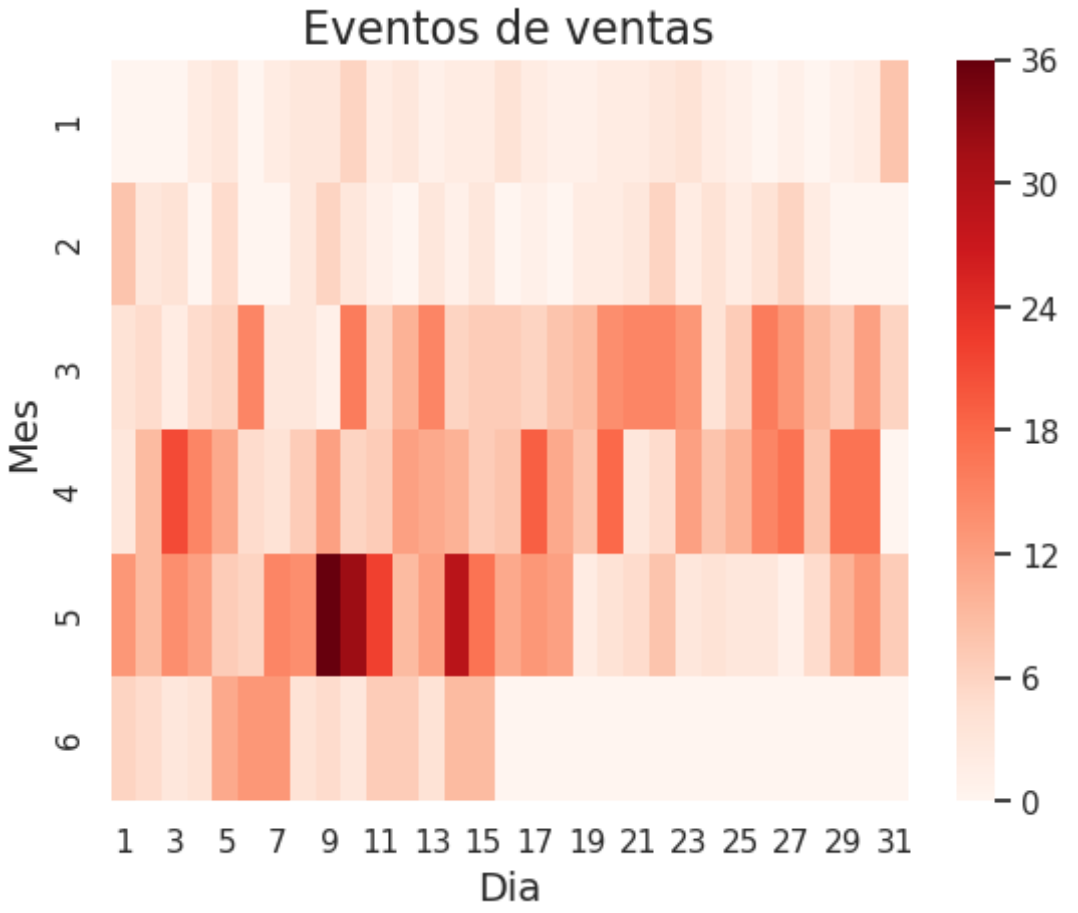
5.2 Google ads vs ventas y checkouts

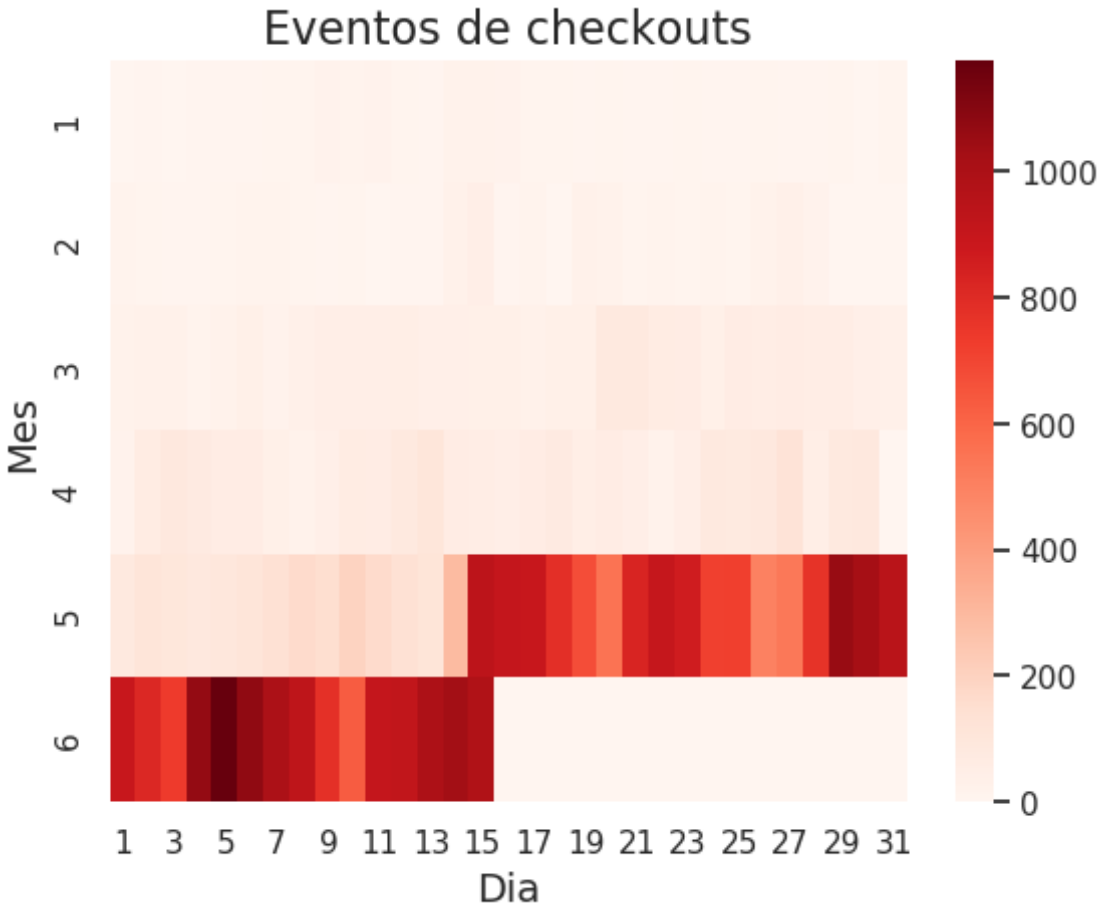
Queremos analizar la correlación entre los eventos generados por ingresos de usuarios a través de anuncios de Google y los eventos generados por checkouts y conversiones.



En este gráfico se puede ver con mas detalle el cambio en la cantidad de 'ads_hits' generados por Google, que también se veía en el gráfico anterior, también se puede ver mas fácilmente porque el mes 6 tenia una bajada en la cantidad de eventos, dado que faltan los datos a partir del día 16 en adelante.

A continuación analizaremos con el mismo formato los eventos de ventas y checkouts, para ver las correlaciones entre la evolución de cada uno.


Eventos de ventas
Mes
1
2
3
4
5
6
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31
Dia
36
30
24
18
12
6
0


Eventos de checkouts
Mes
1
2
3
4
5
6
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31
Dia
1000
800
600
400
200
0

Luego de ver estos 2 gráficos se puede concluir que las ventas y los ads_hits de Google no guardan ningún tipo de relación mas que, al igual que los otros datos, tienen baja frecuencia durante los primeros meses y luego aumenta, pero no coinciden mucho mas que eso. Una correlación igualmente seria difícil dada la alta diferencia entre la cantidad de datos de cada uno, ya que el rango de las ventas va hasta un máximo de 36 contra los ads_hits que van hasta +1500. La cantidad de ventas es baja en comparación con la cantidad de eventos de ads_hits o checkouts que hay en el set de datos, como ya veremos mas adelante, la cantidad de ventas de Google es poco representativo sobre la cantidad de ads que esta empresa genera.

Por otro lado los checkouts parecen apegarse muy bien al heatmap anterior , por lo que podríamos atribuirle el gran crecimiento de los checkouts a el mismo crecimiento en la publicidad ofrecida por Google. Estos datos no solo guardan correlación en el crecimiento sino también en la frecuencia, ya que se puede observar que ambos rangos van hasta un máximo de entre 1200-1500, por lo que reforzaría un poco mas la relación. Así también como si mirando con mas detalle, se puede ver que puntos altos y bajos durante los meses de Mayo(5) y Junio(6), como son los días : 3,5,10,26 y 27, coinciden sorprendentemente bien.

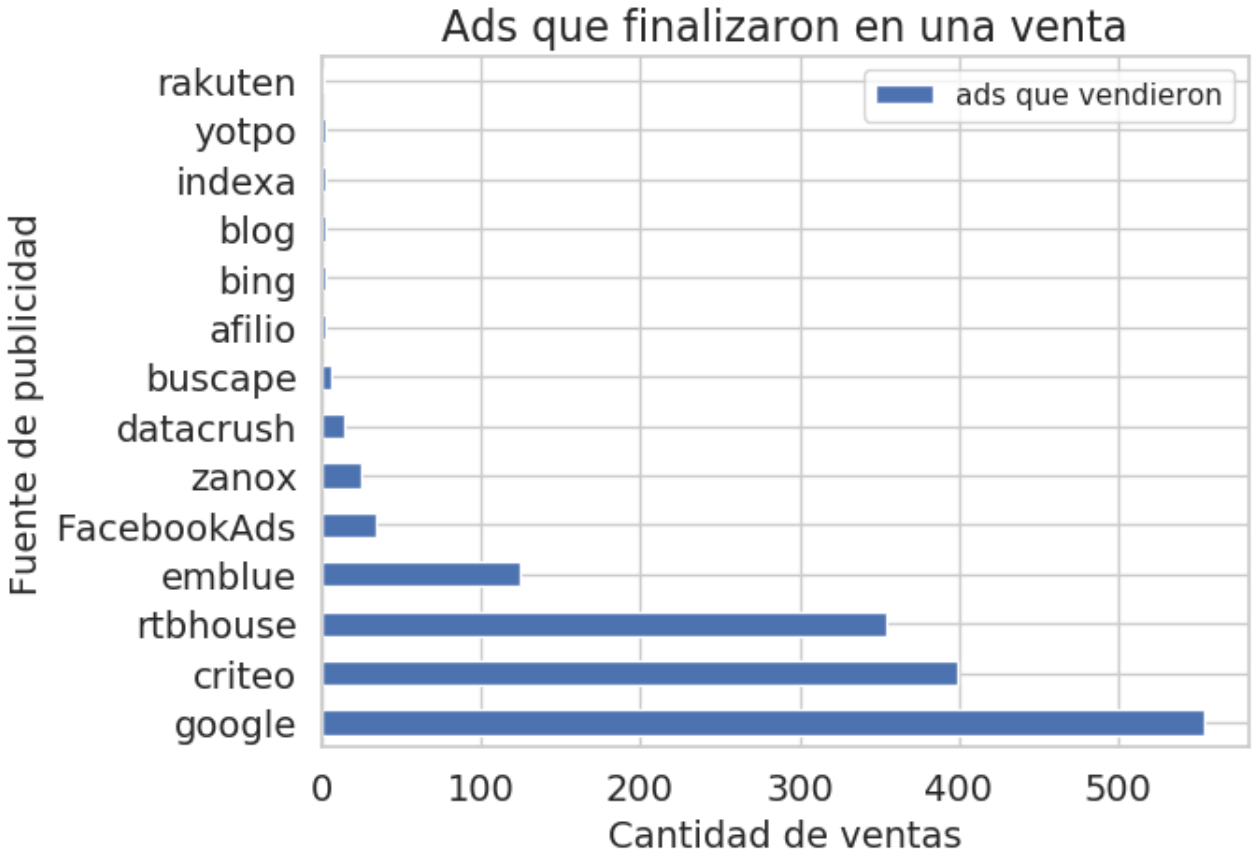
5.3 Ads totales vs ads que vendieron

Obtenemos las ventas que corresponden a cada campaña publicitaria

Ahora que ya tenemos la cantidad de ventas correspondientes por cada 'ad_hit' de cada compañía de publicidad, proseguimos a filtrar los ads totales que le corresponden a cada compañía y luego mostrar ambos datos, y así ver la correlación entre la cantidad de publicidad y las ventas que les corresponden.

5.4 Análisis de la correlación

En el siguiente segmento, se analizaron a los usuarios que entraron mediante *ads* y que efectivamente realizaron una compra gracias a la misma. Organizamos por fuentes de publicidad, contando la cantidad de publicidades que terminaron que una venta, para ver cual fuente de ads que realizo la mejor devolución.

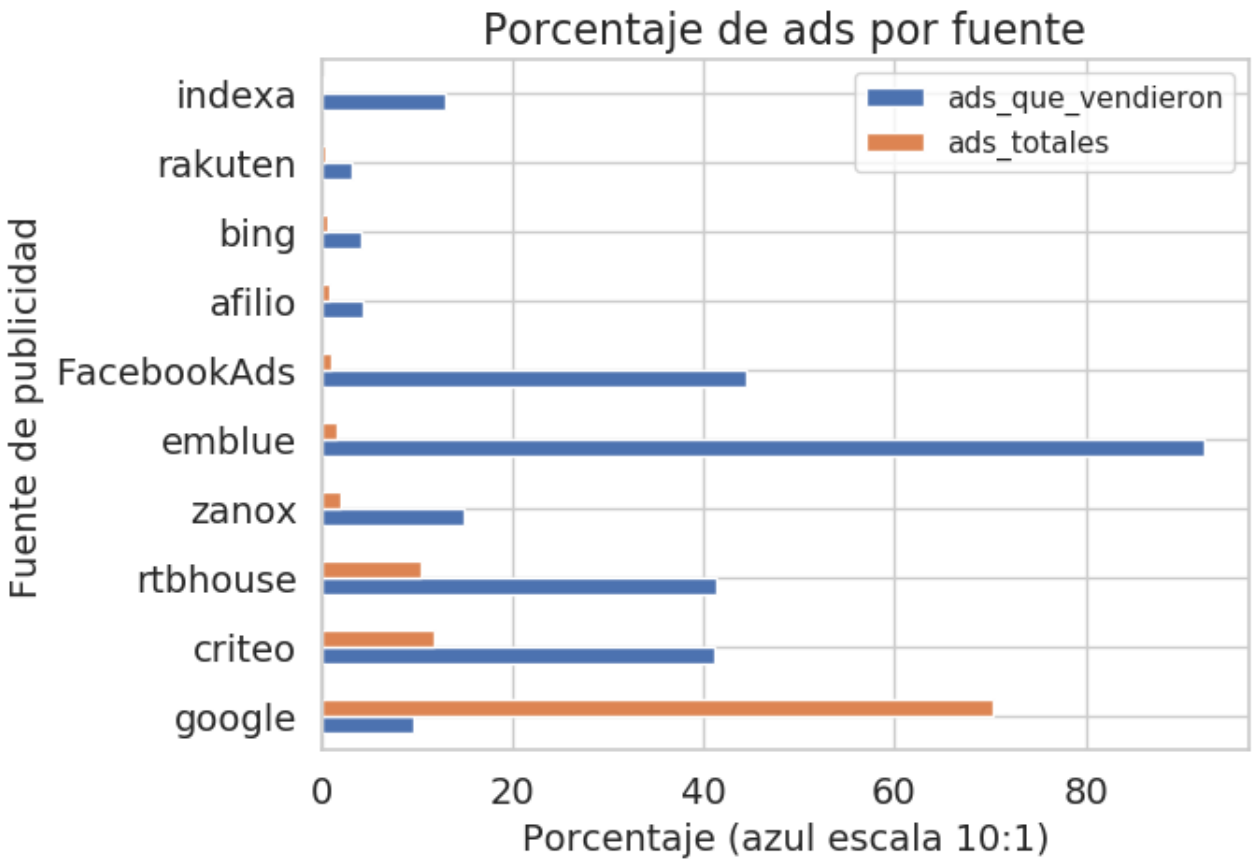


Se busco analizar una correlación entre las compras de los usuarios con su ingreso a la pagina a través de una publicidad. Se intento filtrar por eventos relacionados a ventas y luego ver los datos en la columna "campaign source" pero por alguna razón, para todos los eventos que sean ventas esta información, así como muchas otras, figura como nula. Cuando en realidad la información del usuario,como pude ser el país, se puede encontrar fácilmente en la tabla. Teniendo en cuenta que la información puede simplemente no estar asignada a la venta, no se descarta que el comprador haya accedido por medio de un ad.

Procedemos entonces a separar la información en dos partes, por un lado una lista de todos los compradores junto con el producto involucrado (la llamaremos Tabla A), y por otro lado todas las entradas de usuarios a la pagina por medio de un ad (la llamaremos Tabla B). La mayor parte del análisis se lleva a cabo en la Tabla B, como filtrar que las publicidades solo conduzcan a una compra, y no al inicio de la pagina o a ventas. Luego agrupo los eventos por usuario y marca del producto. A la Tabla A la organizamos de la misma manera, por usuario y marca de producto. Luego se procedió a hacer un Merge de ambas tablas , con las columnas de usuario y marca (Tabla A a izquierda) descartando a todos los usuarios que no habían realizado ninguna compra y a los compradores cuyos productos comprados no coincidían con algún producto visto en una publicidad. De esta manera, si ahora agrupamos los eventos por "campaign source" , sumando la cantidad de productos comprados en cada uno de esos eventos obtenemos la cantidad de ventas que fueron apoyadas por cada fuente de publicidad y así tener una idea de que fuentes generan mas ventas, así poder decidir mejor en que campañas publicitarias invertir mas y en cuales menos.

Vale tener en cuenta que si por ej 1 usuario compro 1 Samsung , pero entro previamente a través de 1 publicidad de Google y Criteo , ambas publicidades sobre Samsung, la venta se cuenta como que fue gracias a ambas publicidades. Por lo tanto si sumamos la cantidad de " ventas totales por fuente de publicidad" (2) va a ser mayor que "las ventas totales " (1). Lo mismo ocurre para el "porcentaje de colaboración a las ventas"

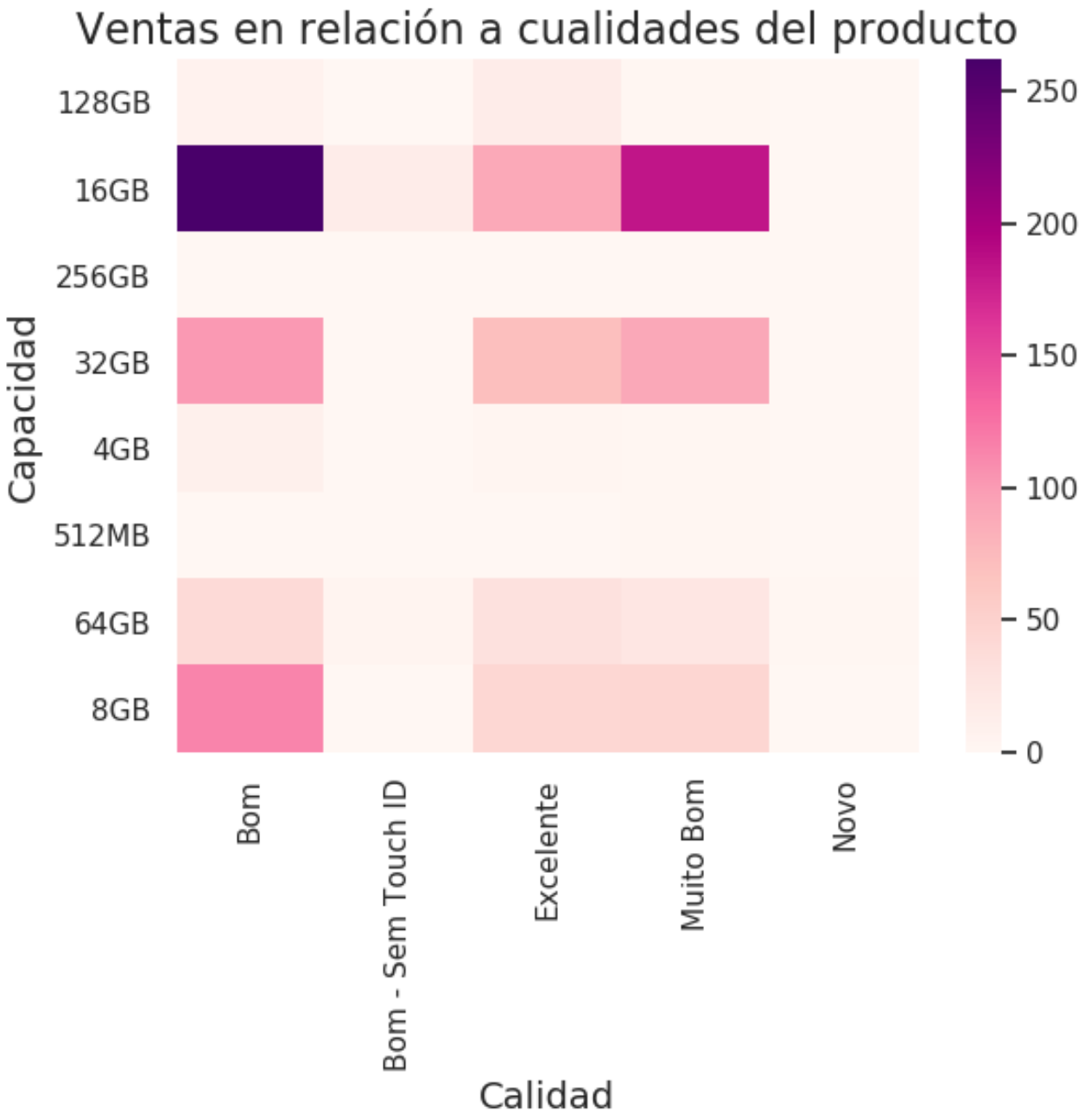
Se pudo ver que la mayor eficacia la tuvieron aquellos que entraron mediante Google, lo cual era muy esperable dada su popularidad. Sin embargo hubo otros portales que también mostraron ser efectivos.



Siguiendo los pasos del análisis anterior, se decidió tener en cuenta aparte de los ads que colaboraron con una venta (analizado previamente), la cantidad de ads generados por la fuente de publicidad; y así poder no solo saber que fuente apoyo en mas cantidad las ventas, sino también que porcentaje de sus ads.

De esta manera podríamos ver si hay alguna empresa A que genera 5000 ads de los cuales 500 terminan en ventas, versus una empresa B que genere 400 de los cuales 200 terminen en ventas. La empresa A podrá tener mayor cantidad pero no habría que descartar invertir en la B dado que su porcentaje de hits podría es mayor.

6 Análisis de productos vendidos

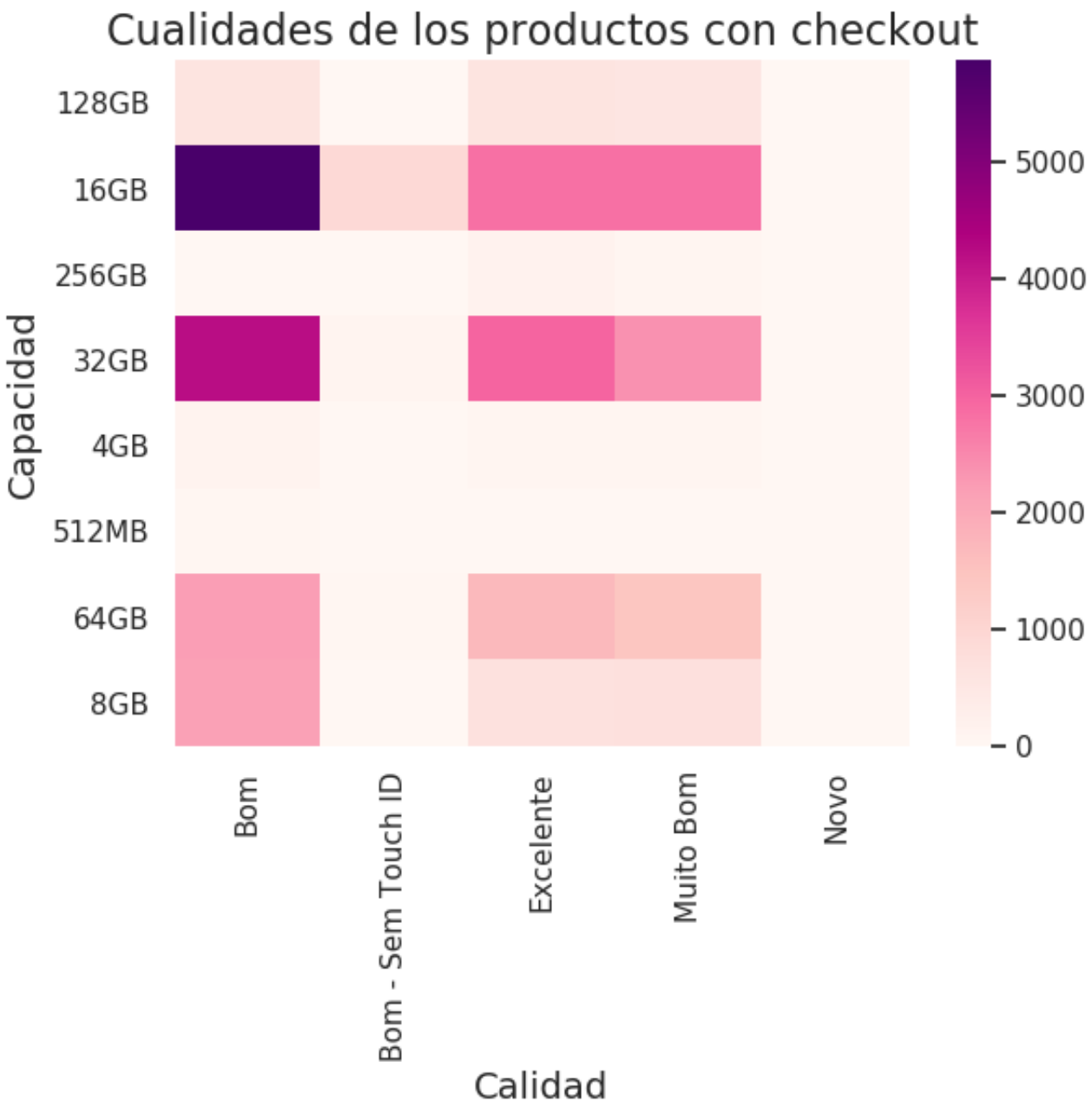


En este gráfico se puede ver que los productos que están en mejor estado, 'nuevos', no tienen casi ventas realizadas durante estos 6 meses, a diferencia del que vendría a ser el opuesto, el producto de calidad mas baja, 'buen estado', es el que mas ventas colecciona entre todos, seguido por lo de calidad 'muy bueno' y 'excelente', de los cuales también se destaca el de calidad ' muy bueno'. Por lo que en conjunto, las 2 mejores calidades (Excelente y Nuevo) son de las que menos ventas tienen en comparación con productos de calidades mas bajas. Esto se debe posiblemente a una cuestión de precio en los dispositivos, por lo que los

clientes tienden a ahorrar en cuanto a la calidad pero obtienen el mismo producto, ya que de todos modos la calidad no es mala.

Algo parecido sucede con las capacidades de los dispositivos , pero no exactamente. Ya que los productos con mayor capacidad del mercado no (64GB, 128GB y 256GB) tienen ventas casi nulas, eso puede deberse a cuestión de precio, pero también puede deberse a que esta cantidad de capacidad en un dispositivo de este tipo no es de lo mas común. Teniendo en cuenta eso, para el resto de las capacidades no parece aplicarse el mismo criterio que se aplicaba para las calidades, ya que las ventas también son bajas para el extremo de las capacidades chicas como las de 512MB y 4GB ,que son las 2 menores En cambio las ventas se concentran en dispositivos con capacidades intermedias de de entre 8 y 32GB, teniendo como punto intermedio las de 16GB que son justamente las que definitivamente sobresalen en las ventas.

Igualmente dado que la muestra tomada de las cualidades de los dispositivos que compro la gente es chico por sobre todos los diferentes tipos de eventos en los que se ve involucrado el seleccionar un dispositivo, proseguimos a realizar el mismo análisis con otro tipo de evento que tenga mas frecuencia y que no este al mismo tiempo muy lejos de ser una compra, para no estar comparando cosas que no tienen relación. Por lo tanto escogimos checkout.

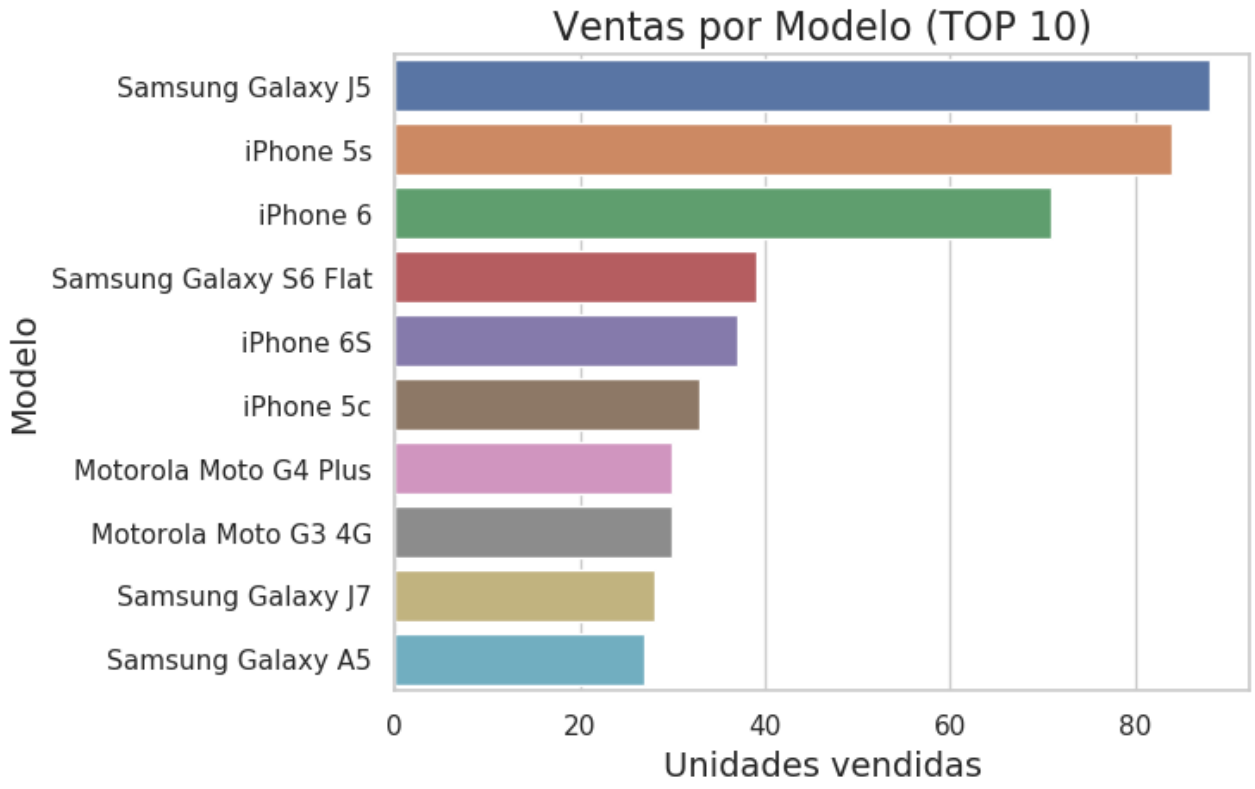


Podemos notar que la relación entre ambos gráficos es bastante exacta, con lo que podemos afirmar mas fuertemente que los usuarios verdaderamente tienen preferencia en dispositivos que no sean nuevos, y que tengan una capacidad de memoria intermedia.

En ambos análisis tomamos como que la calidad 'Bom sem touch ID' forma parte de la calidad 'Bom', por lo que no marcamos que esta calidad no era buscada entre los usuarios.

6.1 Modelos mas vendidos

A continuación se realizó un análisis de los modelos mas vendidos.



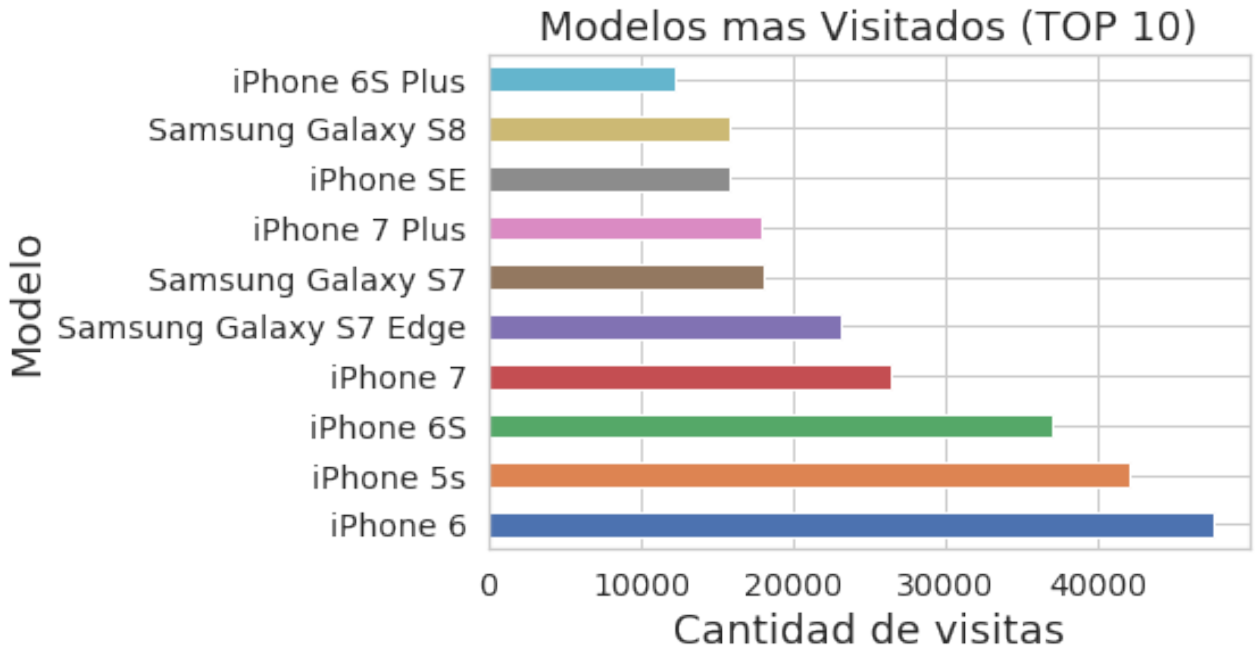
El gráfico anterior nos demuestra como los modelos más vendidos se distribuyen entre 3 marcas que son: *iPhone, Samsung* y *Motorola*, siendo las primeras dos mucho mas vendidas que la tercera.

Al hacer un análisis temporal de las marcas mas vendidas, vemos como *iPhone* supera ampliamente a las demás. Ademas, cada vez aumenta mas la diferencia entre la cantidad de ventas con las demás marcas, por lo que se espera que siga siendo la marca dominante en ventas en el futuro próximo.

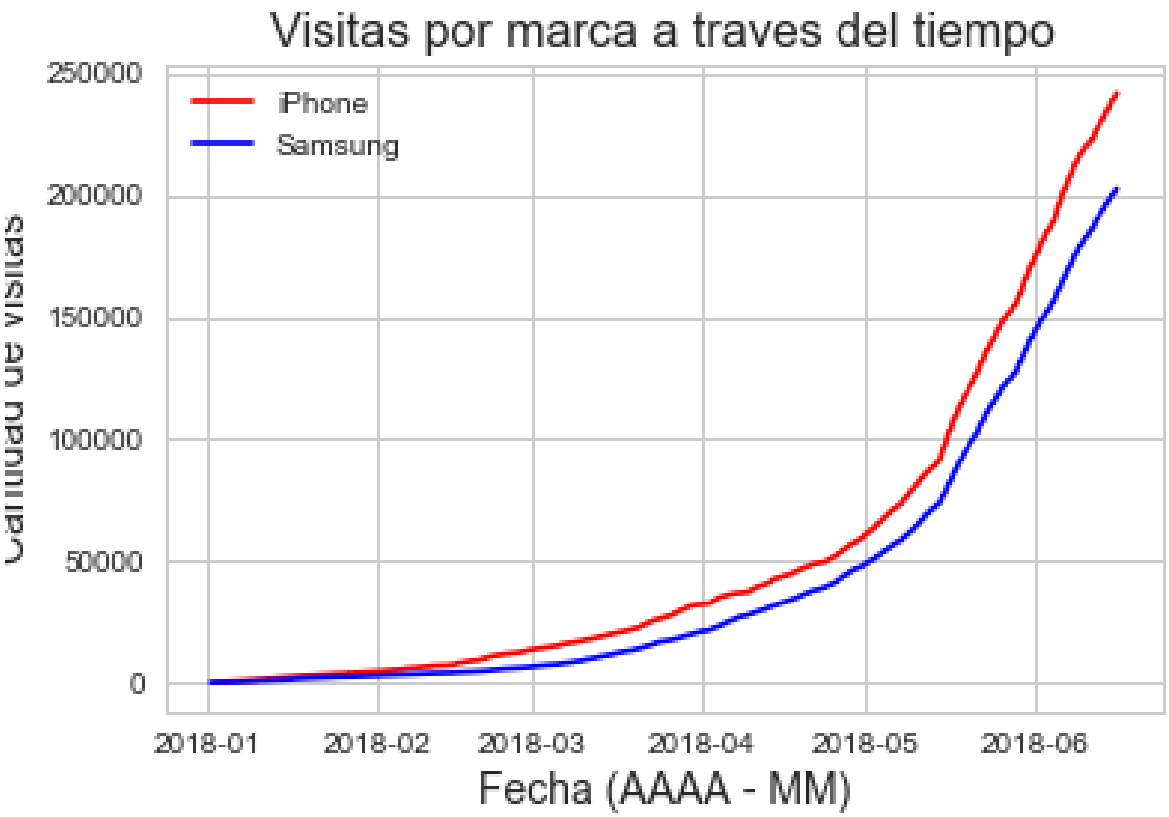
7 Análisis de popularidad

7.1 Modelos más visitados

Se analizó la popularidad de ciertos modelos que destacaron entre los demás.

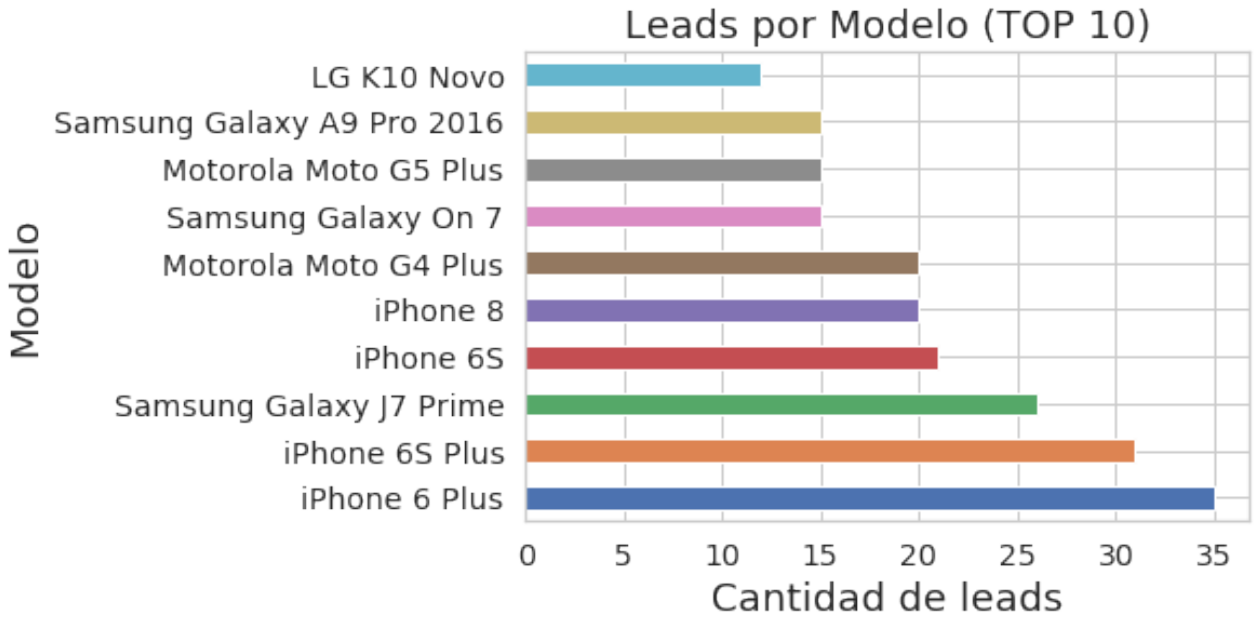


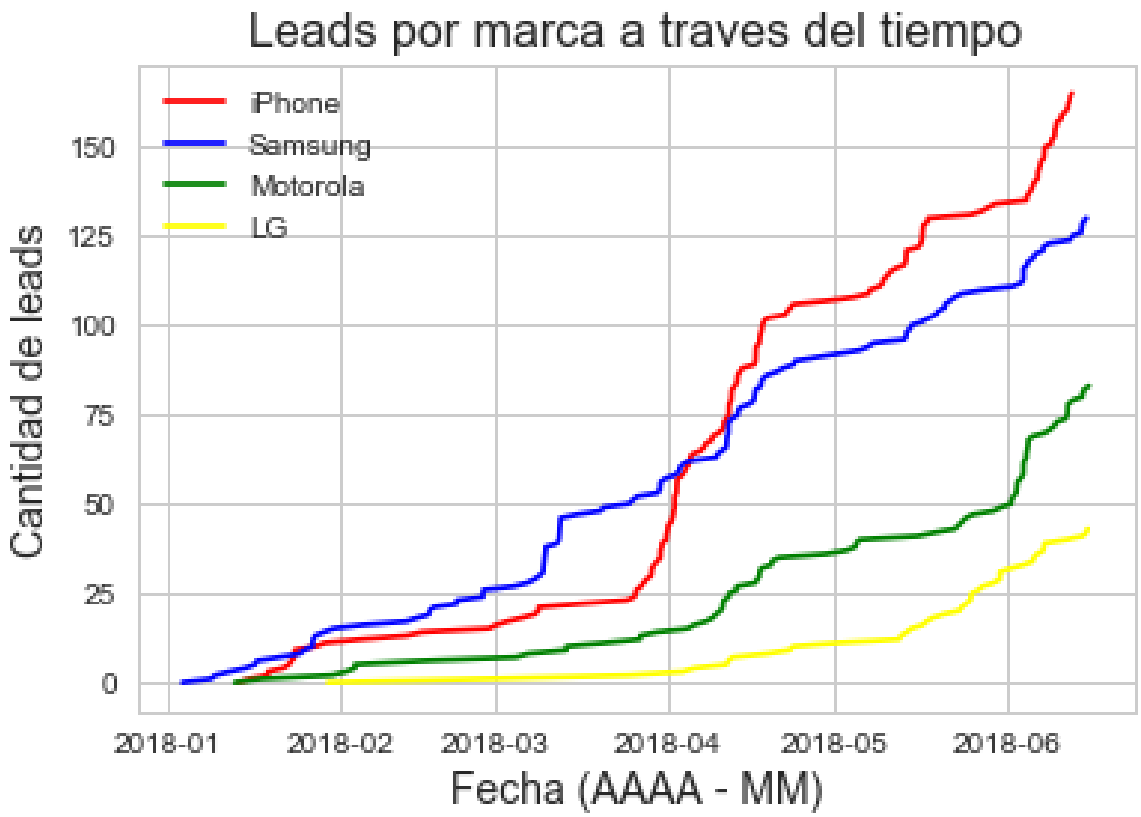
Por lo visto, los modelos mas visitados por los usuarios fueron de *iPhone* y *Samsung*, por lo que se realizó un analisis del tiempo para ellos.



Se puede ver como ambos tienen un crecimiento a traves del tiempo muy parecido, aunque los mas visitados siempre fueron los de *iPhone*.

7.2 Leads por modelo



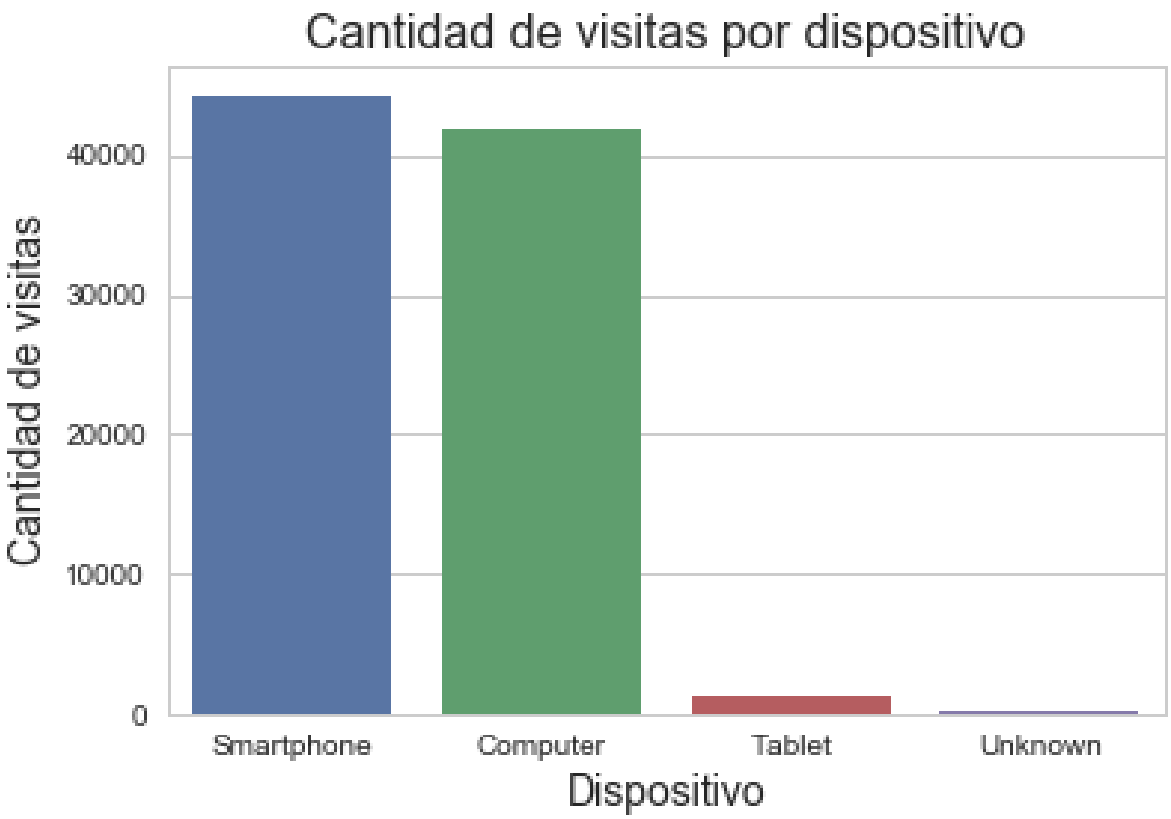


A partir de éstos gráficos, se pudo ver como particularmente el modelo *iPhone 6* fue muy codiciado por los usuarios, lo cual coincide con el análisis de ventas previo (aún si la venta no fue concretada).

Aún así, es interesante ver como los leads arrancaron siendo mayoria para los modelos de la marca *Samsung* antes de ser superador por *iPhone* a partir de Abril del 2018.

8 Dispositivos de acceso

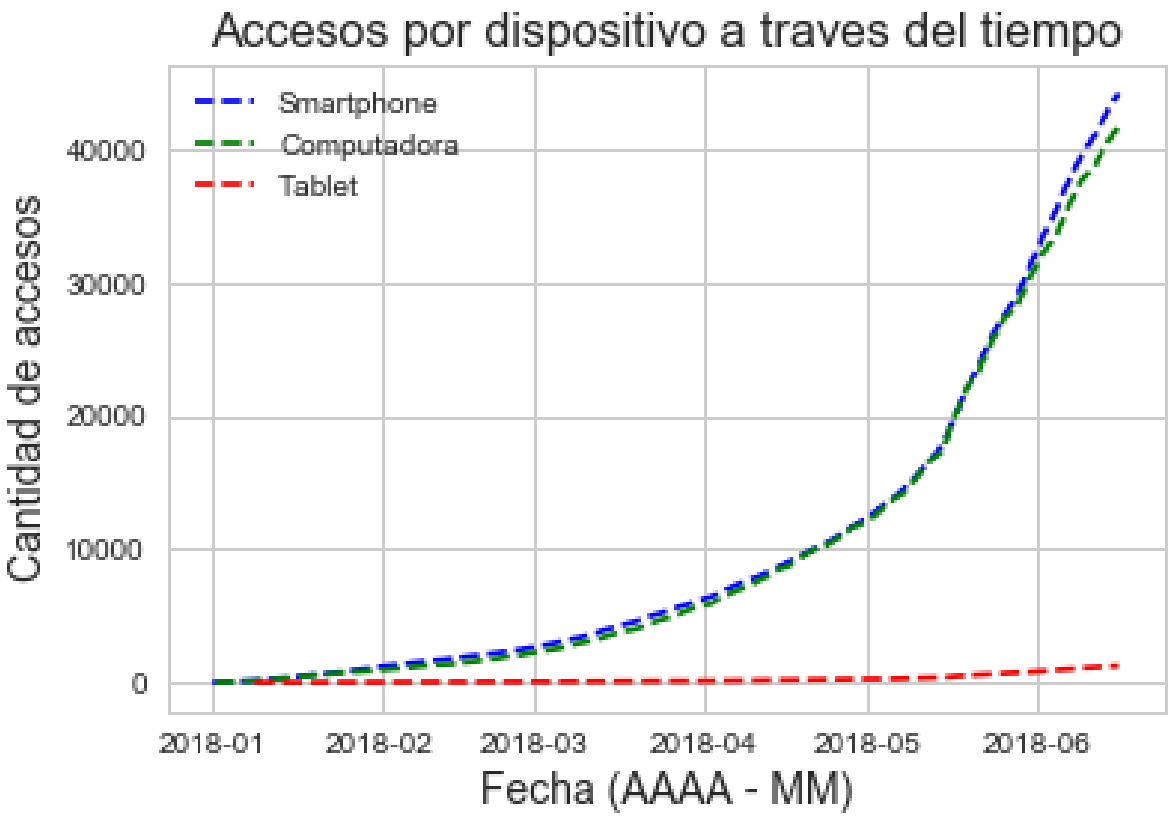
Una de las principales preguntas de este análisis apunta a que si una versión móvil de la plataforma tendría sentido. Para poder contestar eso se analizó la cantidad de accesos a la plataforma mediante distintos dispositivos:



Como se puede ver, la mayor parte de los accesos a la plataforma se producen mediante *Smartphone*, un poco menos mediante Computadora y casi nada mediante una *Tablet*.

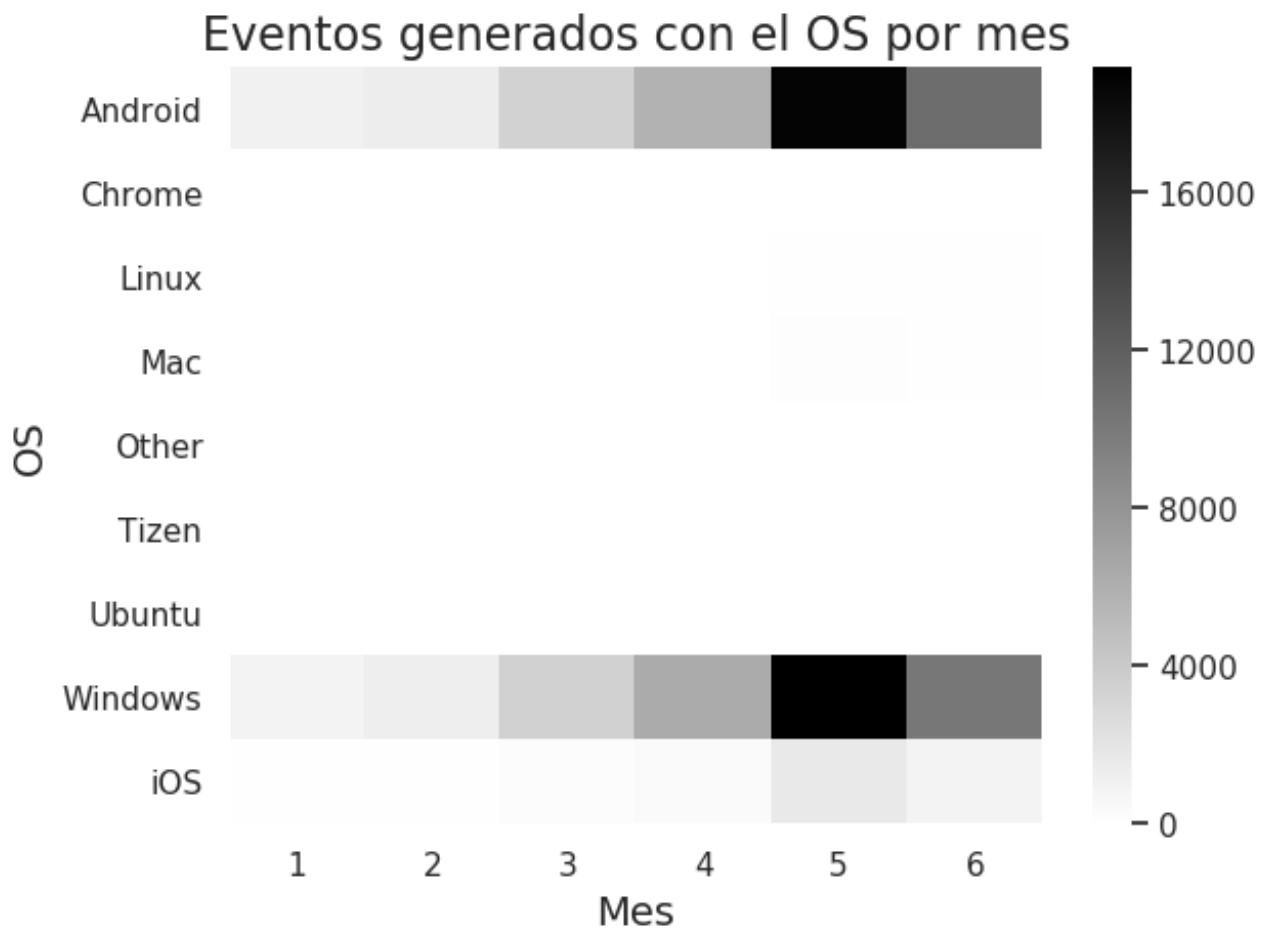
8.1 Evolución de acceso de dispositivos

Algo que nos interesó averiguar es si esto siempre fue así y como será en el futuro. Para ello analizamos su comportamiento a lo largo del tiempo en los datos provistos:



Podemos ver como la cantidad de accesos por dispositivo a través del tiempo a través de *smartphones* y computadoras se comportaron casi idénticamente en los 6 meses analizados, por lo que no habría motivo para suponer que la diferencia entre ambos fuera a ampliarse mucho mas en los próximos meses. El acceso mediante *Tablet* no parece ser algo promisorio.

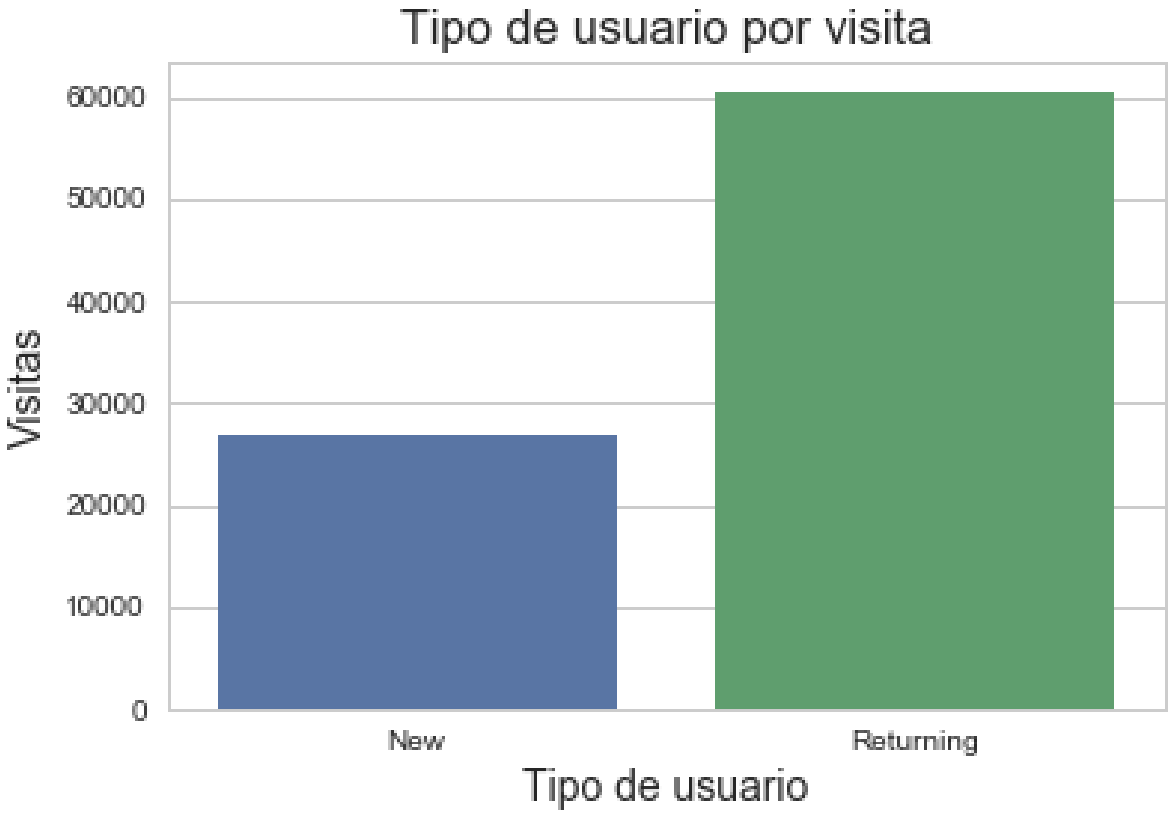
9 Análisis de eventos generados por OS's



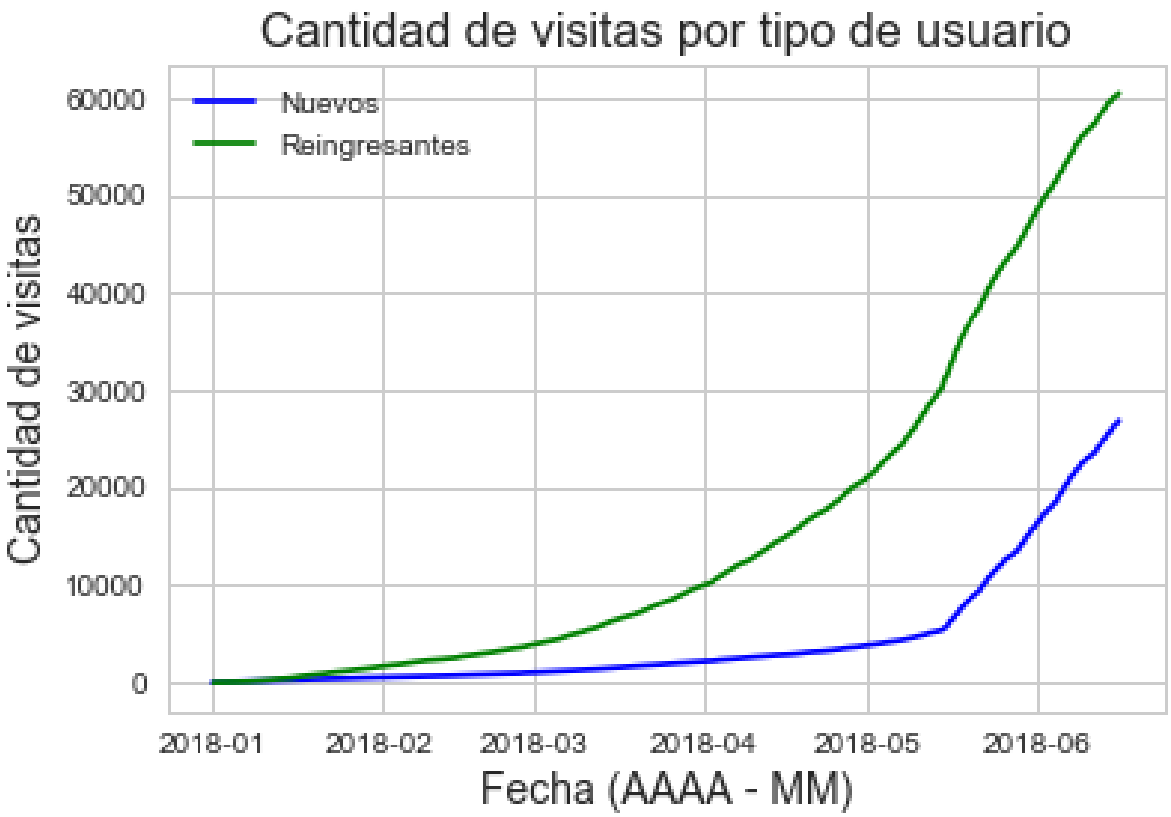
Mediante este heatmap, vemos la cantidad de eventos por sistema operativo que se generaron en los 6 meses analizados. Particularmente *Android* (para *smartphones*) y *Windows* (para *computadoras*) fueron los que mas generaron eventos. Lo que podría usarse para tener en cuenta como objetivo en los tipos de dispositivos a los que apunte nuestra publicidad, ya que vemos que la mayoría de los usuarios que nuestra pagina atrae son de ese conjunto de personas. La otra versión relevante fueron los dispositivos con sistema *iOS*, pero en mucho menor medida que los mencionado anteriormente.

10 Usuarios nuevos y reingresantes

Se consideró que analizar cuantos usuarios ingresaban por primera vez y cuantos eran reingresantes es una medida de que tan útil les resultó la plataforma, por lo que en los siguientes gráficos se muestra esta relación.



En principio se vio que la cantidad de usuarios que visitaron el sitio superan ampliamente a los usuarios que lo hicieron por primera vez. Esto puede ser un indicio de la conformidad de los usuarios con la plataforma, lo cual los lleva a que esten a gusto con el servicio.



Si bien es bueno que la cantidad de usuarios que ingresan por al menos una segunda vez sea creciente, el hecho de ver como la cantidad de usuarios nuevos que ingresan por primera vez aumenta bruscamente a partir de mediados de mayo del 2018 refleja un aumento de popularidad del sitio, lo cual es muy beneficioso.

11 Conclusiones

Al haber visto los dispositivos mediante los cuales los usuarios accedieron al sitio, nos importó saber cuántos de éstos terminaron efectivamente en una compra.

- **¿ Vale la pena invertir en la creación de una aplicación móvil?**

  Sí. Los accesos mediante *smartphones* supera a los que se realizan mediante una computadora, por lo que una plataforma mobile de la página podría ser beneficiosa para estos usuarios.

- **¿Hay algún sector en el cual se pueda ahorrar presupuesto?**

  Sí. Particularmente en la sección donde se analizó la efectividad de las publicidades, se vio que hay campañas de marketing cuya efectividad es una mínima fracción de las que son más efectivas. Dichas campañas representan una pérdida para la compañía.

- **¿Se podría mejorar algún aspecto de la página web de la empresa?**

  En los análisis de ventas y de popularidad, los modelos iPhone demostraron ser de los mas solicitados. Sin embargo, en la barra de navegación de la página web, no existe ningún filtro que nos permita acceder rápidamente a los modelos que la gente más parece querer. Dicha cuestión se podría corregir.

- **Estado de la empresa**

  En todos los análisis, la actividad de la pagina parece ser que esta en un constante y fuerte crecimiento, desde las ventas hasta visitas, algo que es muy importante para la empresa. Uno de los puntos que mas se vio cambiar fue la cantidad de checkouts que la pagina venia manejando y que cambio considerablemente gracias a lo que parece ser un buen aumento en la publicidad de Google, lo que es un punto muy importante a tener en cuenta.