FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Trabajo Práctico 1

Análisis Exploratorio de Datos

Grupo 39

Segundo Cuatrimestre 2018

Fecha de Entrega:
24/09/2018

Integrantes:

Nombre y Apellido	Padrón
Agustín Zuretti	95.605
Anarella Nicoletta	94.551
Pablo Prieto	91.561

Introducción

El objetivo del presente trabajo práctico es realizar un análisis exploratorio sobre un set de datos provisto por la cátedra. La empresa Trocafone ha provisto un archivo de tipo CSV, conteniendo información sobre un conjunto de eventos web realizados por los usuarios de su plataforma de ecommerce. Dicho sitio, www.trocafone.com, permite la compra y venta de distintos dispositivos electrónicos.

El archivo se llama ‘events.csv’ , y contiene las siguientes 23 columnas:

timestamp: fecha y hora a la que ocurrió el evento
event: tipo de evento
person: identificador del usuario 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 (indica de cuál 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: indica si el evento fue generado por un usuario nuevo (New), o por uno 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 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

Asimismo, se registran 11 tipos de eventos:

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, es decir, compra un producto
lead: el usuario se registra para recibir una notificación de disponibilidad de stock, para un producto que no se encontraba disponible al momento de la consulta.

En lo que respecta a volumen de datos, el archivo cuenta con 1.011.288 registros.

Hemos decidido utilizar la herramienta Python Pandas para realizar el análisis exploratorio de los datos.

Data cleansing

A fin de mejorar el análisis exploratorio, hemos realizado una limpieza de los datos iniciales. En primer lugar, decidimos remover algunas columnas del DataFrame, que no aportaban valor a nuestro análisis. Detallamos las columnas eliminadas, y las razones consideradas para ello:

url: no nos brinda ningún tipo de información, siendo el valor más frecuente el símbolo '/'.

skus: nos enfocamos en la columna *sku*, que está al mismo nivel del producto linkeado al evento.

city: el valor top es 'Unknown', y nos enfocaremos en región y país.

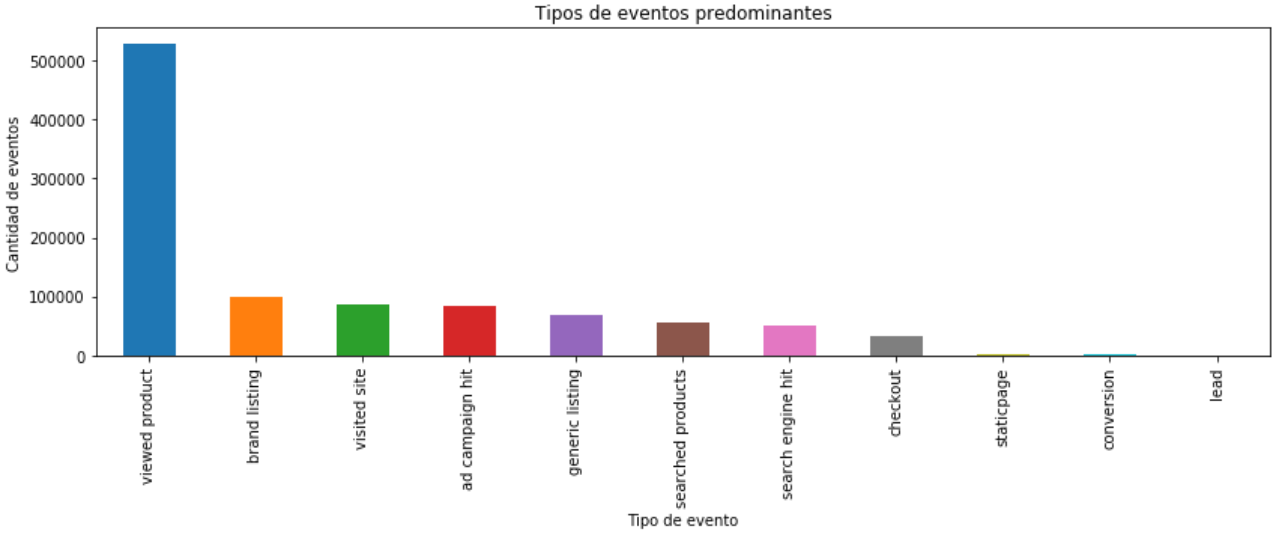
campaign_source, search_engine, channel: datos de campañas, siendo Google el top de búsqueda.

Análisis Exploratorio

Puede consultarse el repositorio con el análisis exploratorio completo en el siguiente link: https://github.com/zurettiagustin/datosTp1

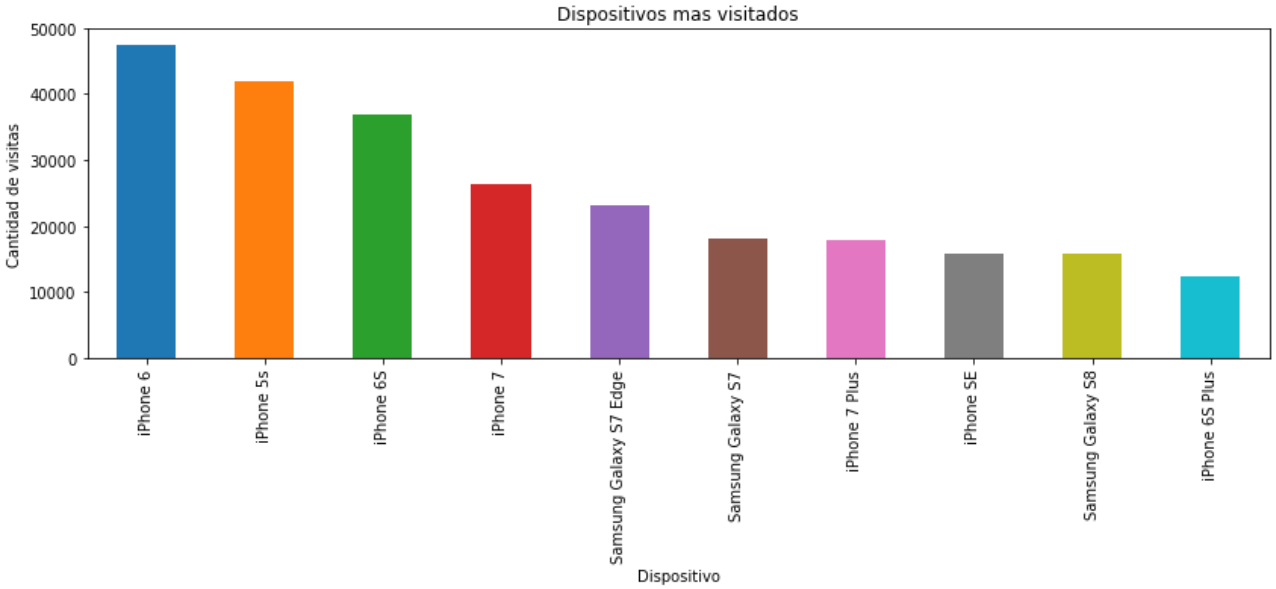
A continuación enunciaremos las distintas preguntas que fuimos planteando para nuestro análisis.

1. ¿Cuáles son los eventos más frecuentes en el set de datos?



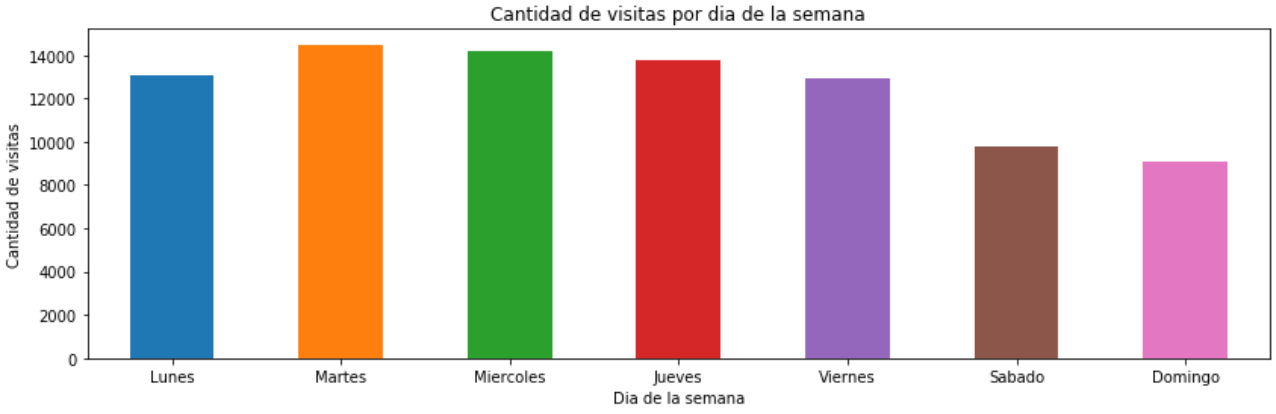
Si contabilizamos en conjunto los eventos de tipo ‘staticpage’, ’conversion’ y ‘lead’, tenemos 5218 eventos, que sobre el total de eventos en el set apenas representan un 0,52 %.

2. ¿Cuáles son los dispositivos/modelos más visitados del sitio?

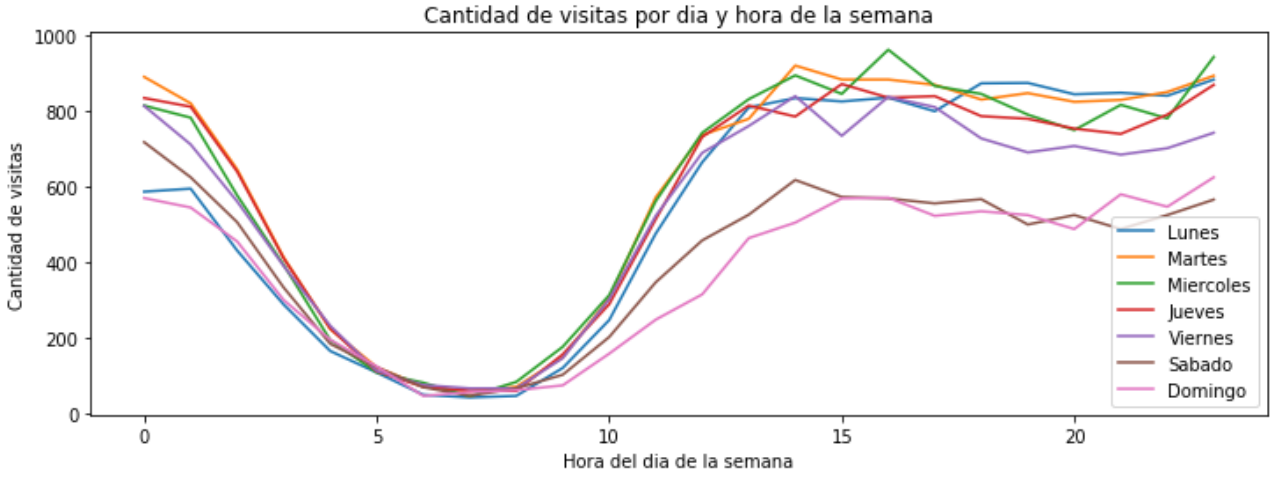


A la hora de visitar el sitio, los usuarios registran más visitas sobre páginas de productos de la marca Apple (específicamente, sobre iPhones) que sobre las demás. Esta diferencia se sigue observando a lo largo del análisis en formas más evidentes. Pero por lo pronto es destacable que de los 10 dispositivos más visitados, 7 corresponden a la firma.

3. ¿Cómo se distribuyen las visitas al sitio según el día y hora de la semana?

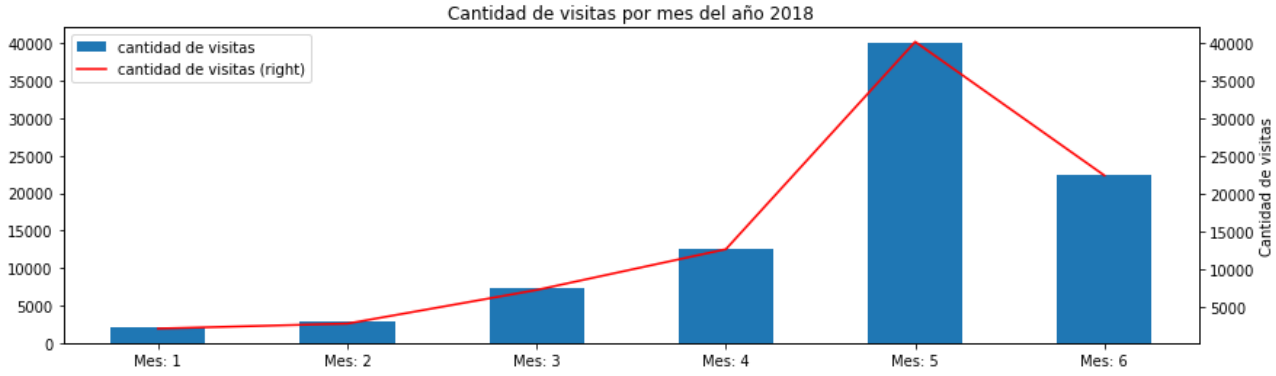


Prevalecen las visitas al sitio de lunes a viernes (días laborables), y muestran menor (pero muy parejo) valor para los días sábado y domingo (fin de semana).



El segundo gráfico reafirma la afirmación anterior. Lógicamente, de lunes a viernes, la cantidad de visitas a la página web comienza a subir recién a partir de las 8 AM, y se mantiene en alto y relativamente constante, entre el mediodía y la 1 AM del día siguiente. Durante los fines de semana las visitas al sitio se dan más bien a partir de las 10 AM, y temprano por la tarde los sábados, y al revés los domingos. Lo cual tiene sentido si pensamos que en esos días la gente realiza otro tipo de actividades.

4. ¿Cómo se distribuyen las visitas al sitio a lo largo del año?



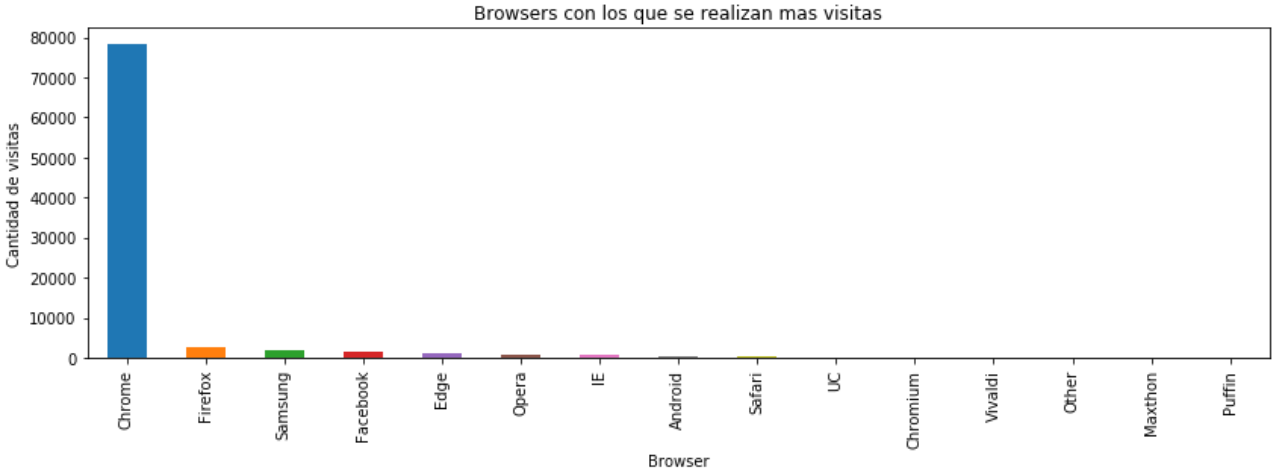
El set de datos solo contiene información para los primeros 6 meses del año 2018, por lo que graficamos la variación de visitas durante este año.

Primero pensamos que el pico observado en el mes de mayo podría explicarse a que el día 13 de ese mes se celebra el día de la madre en Brasil, y sabemos que la web es brasilera, y que la mayor parte de los usuarios de la misma provienen de dicho país.

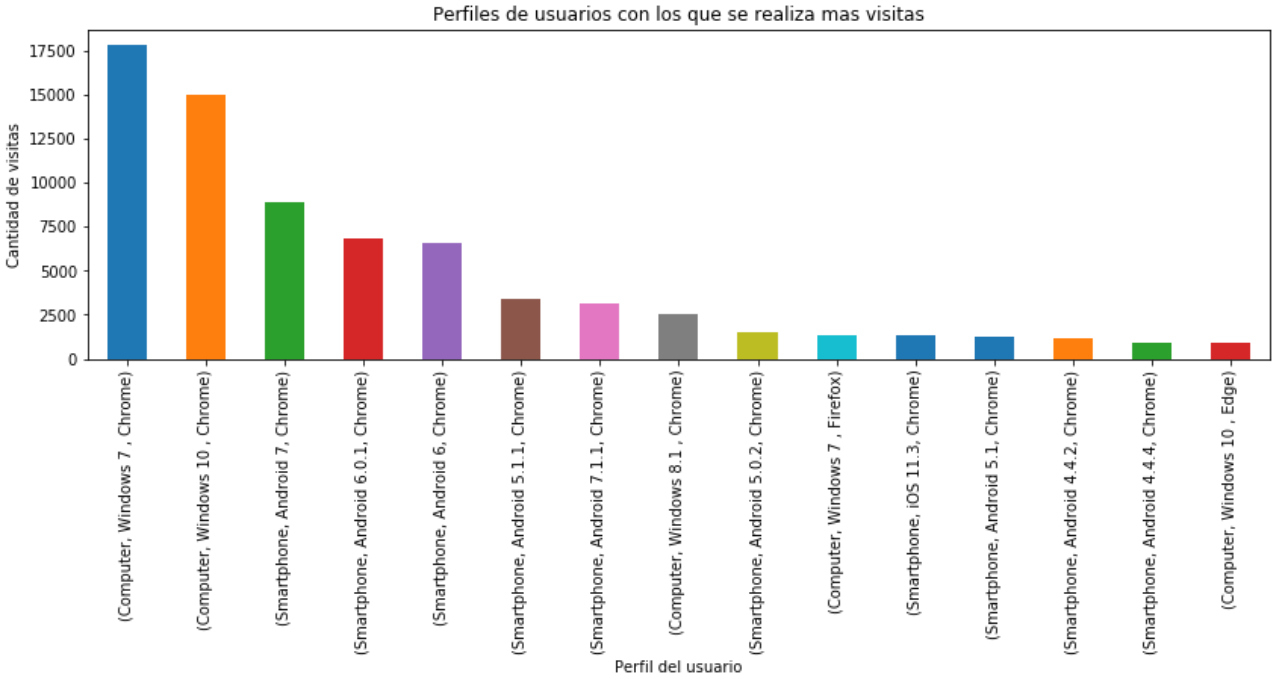
Sin embargo, las visitas se disparan en la segunda quincena del mes, es decir, pasada dicha fecha. Cabe destacar que la compañía emitió algunas publicidades en la red social Twitter (no podemos confirmar que hayan comunicado dichas en otros medios), en las que promocionaban un mes de ofertas, el 'mes de la madre'. Pero fue durante estos últimos días del mes que efectivamente difundió varias ofertas, motivando quizá a los usuarios a visitar el sitio.

A continuación incluimos algunos de los anuncios encontrados.

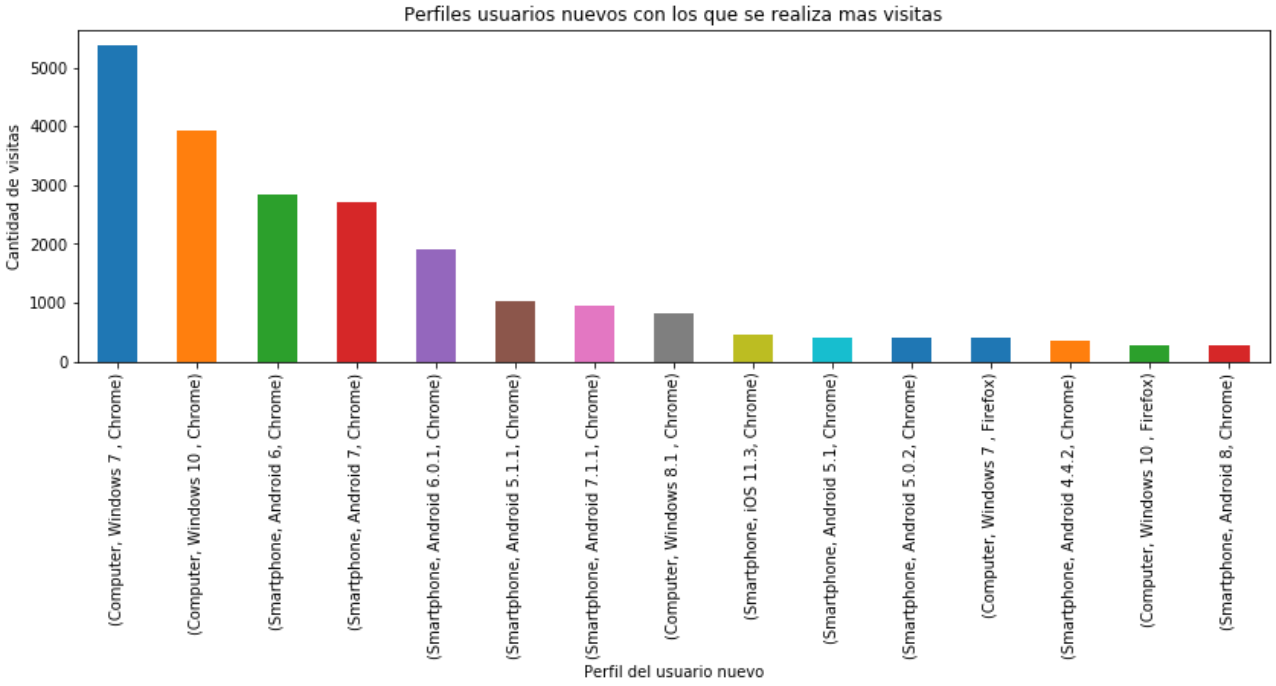
5. ¿Cómo acceden/visitan el sitio los usuarios?

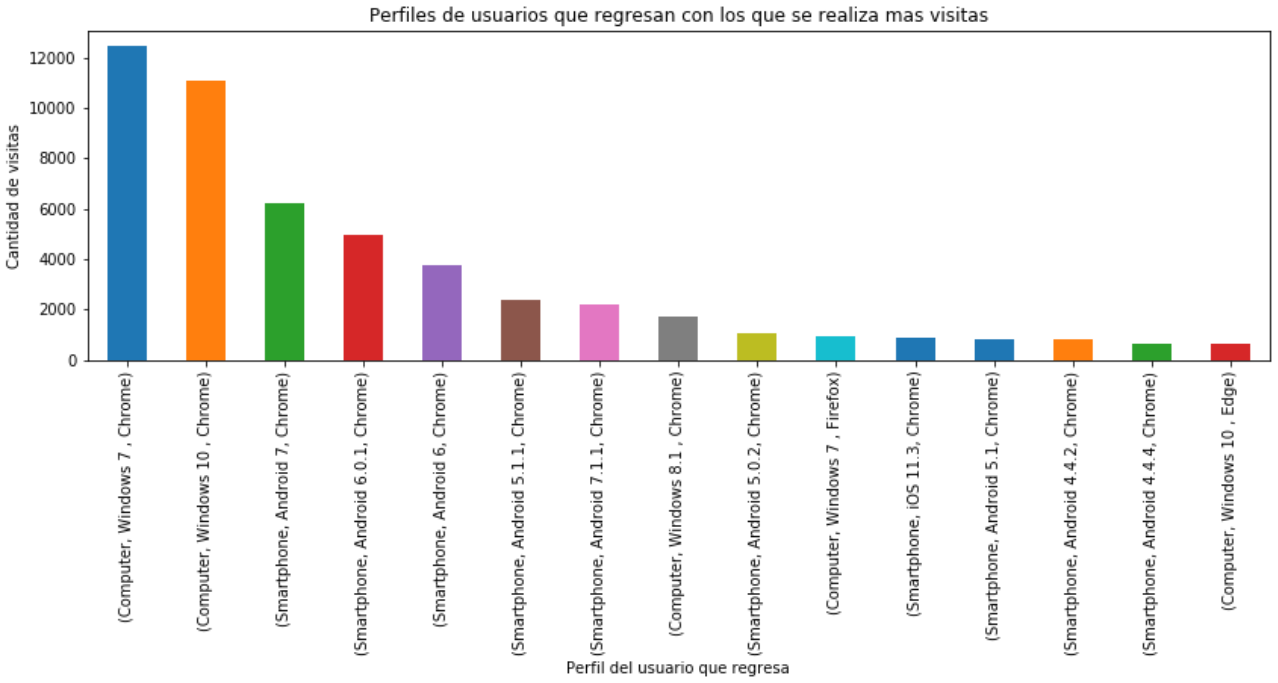


El uso de Google Chrome como navegador para ingresar al sitio web supera por lejos al de otros navegadores. Asimismo, el predominio de Windows 7 y 10 como sistema operativo seguramente está ligado al volumen de usuarios que acceden al marketplace a través de una computadora. Por otra parte, casi en igual medida, hay usuarios que acceden a través de sus smartphones, siendo éstos quiénes hacen uso de los múltiples flavors de Android. El porcentaje de visitas realizadas desde sistemas iOS es bajo en comparación al resto. El siguiente gráfico evidencia esto que comentamos.



Si partimos el análisis anterior en dos, separando a los usuarios nuevos de aquellos que ya habían visitado el sitio anteriormente, obtenemos dos plots muy similares (en distribución, no así en volumen) que hacen ver que este feature, al menos, en lo concerniente a visitas al sitio, no aporta mucho valor al análisis de los datos.

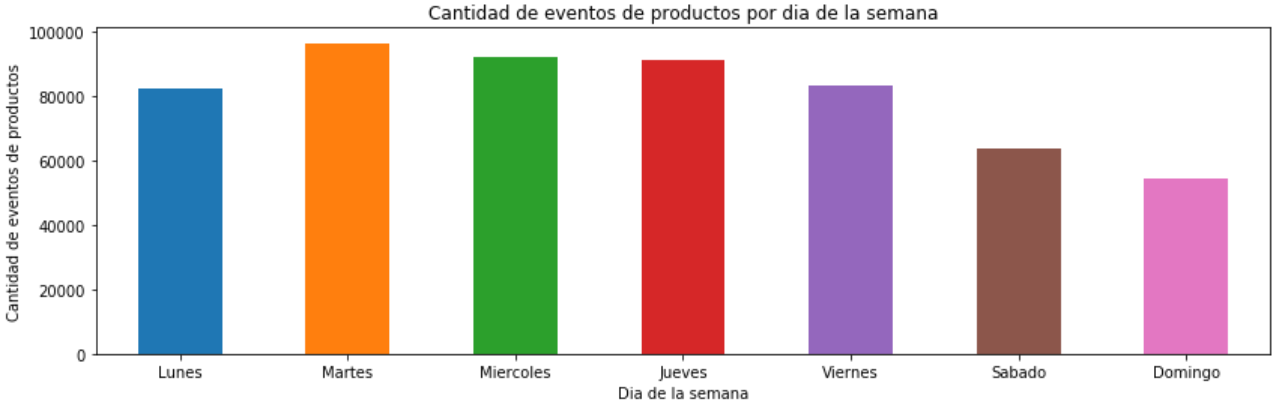


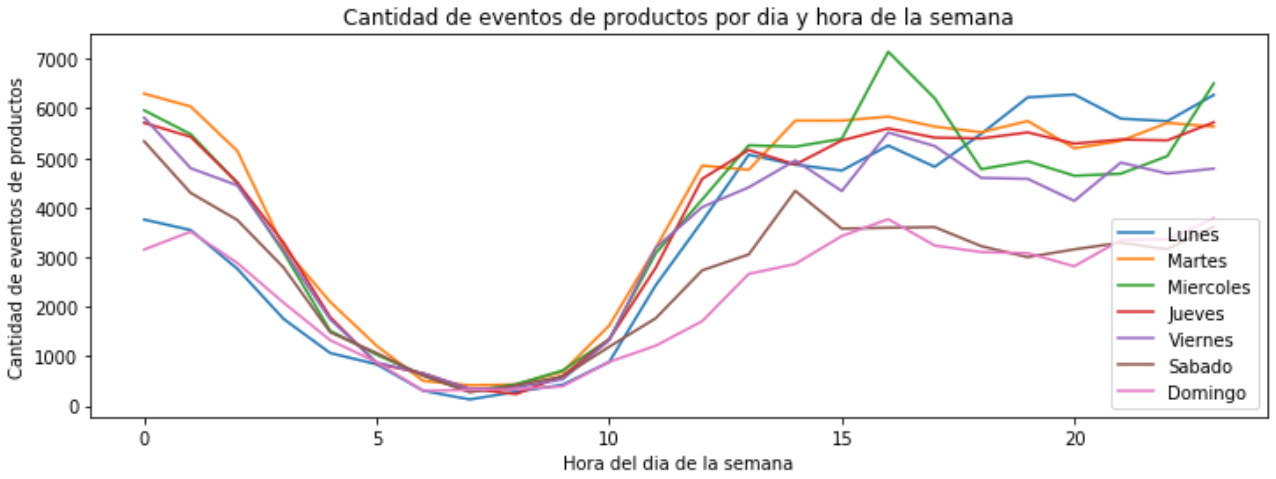


6. ¿Qué eventos involucran directamente productos concretos/skus específicos?

event	fecha	person	sku	model	condition	storage	color
checkout	33735	33735	33735	33733	33733	33733	33733
conversion	1172	1172	1172	1172	1172	1172	1172
viewed product	528931	528931	528931	528931	528931	528931	528931

7. ¿Cómo se distribuyen estos eventos de producto según el día y hora de la semana?

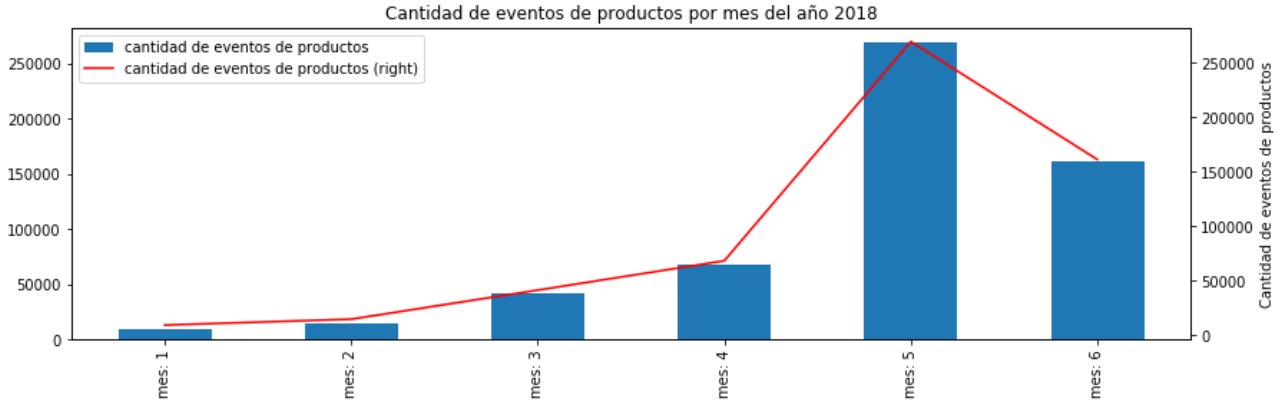




Al igual que ocurría con las visitas a las URLs del sitio, la frecuencia de estos eventos es mayor de lunes a viernes, y menor durante el fin de semana.

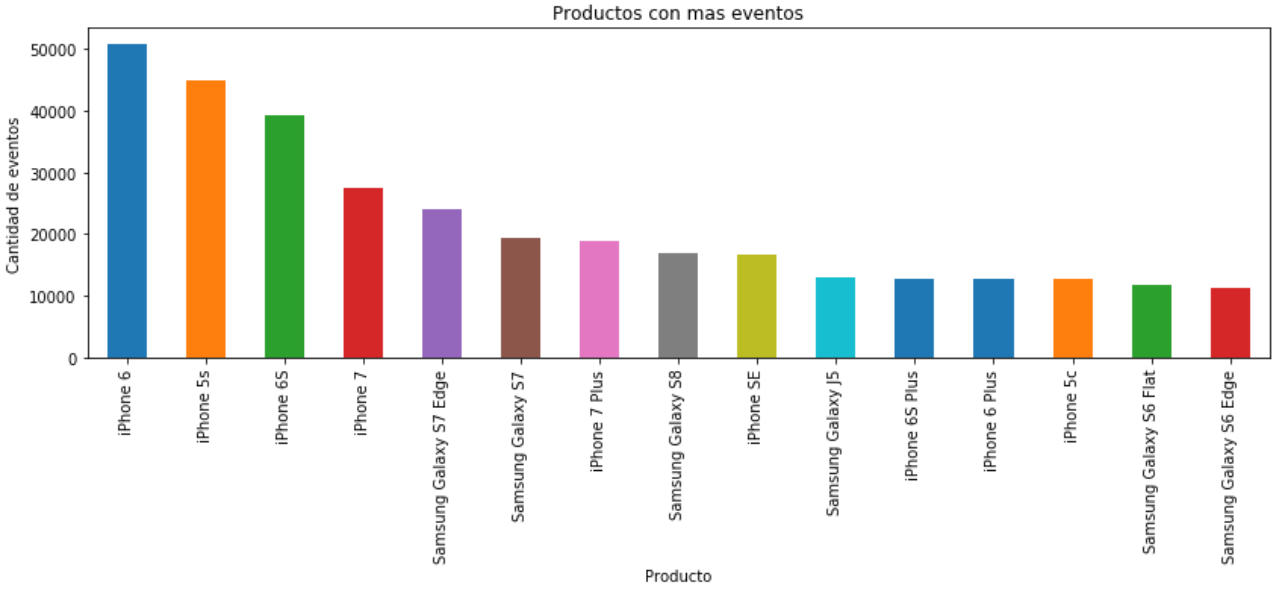
El segundo gráfico también presenta una forma similar al que presentamos anteriormente (destacando igualmente diferencias en la escala, ya que aquí se cuenta con más eventos que en el caso anterior). No obstante, llama la atención que haya un pico de eventos el día miércoles por la tarde.

8. ¿Cómo se distribuyen los eventos de producto a lo largo del año?

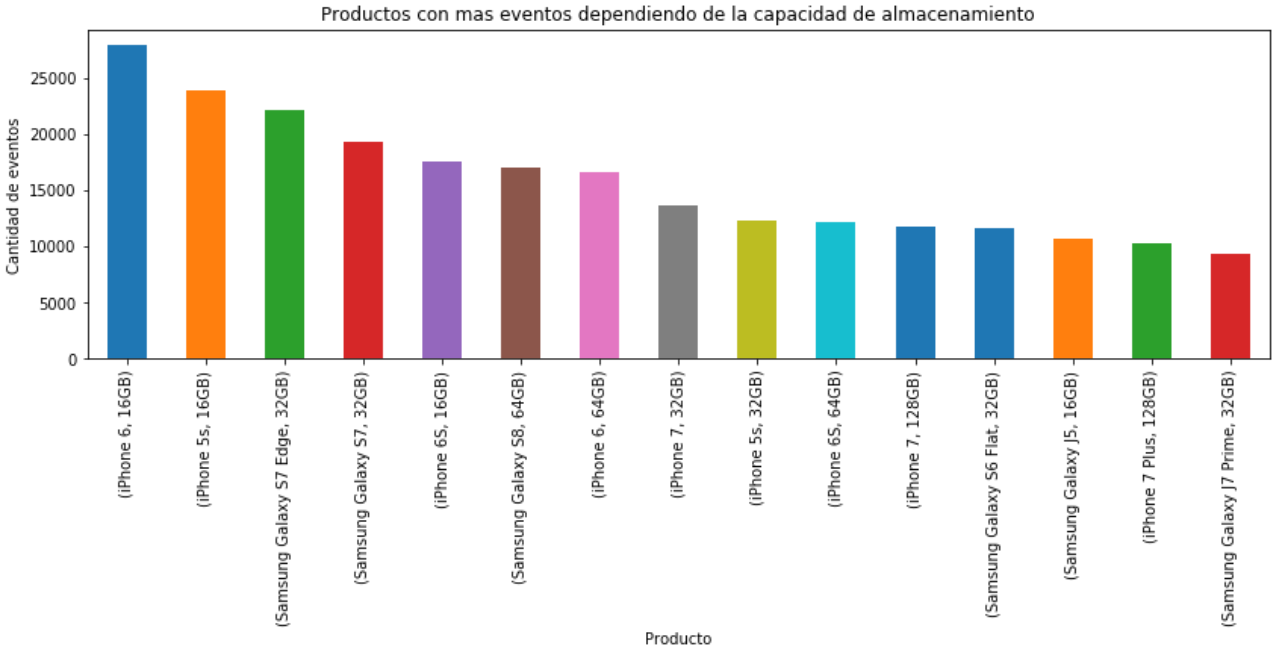


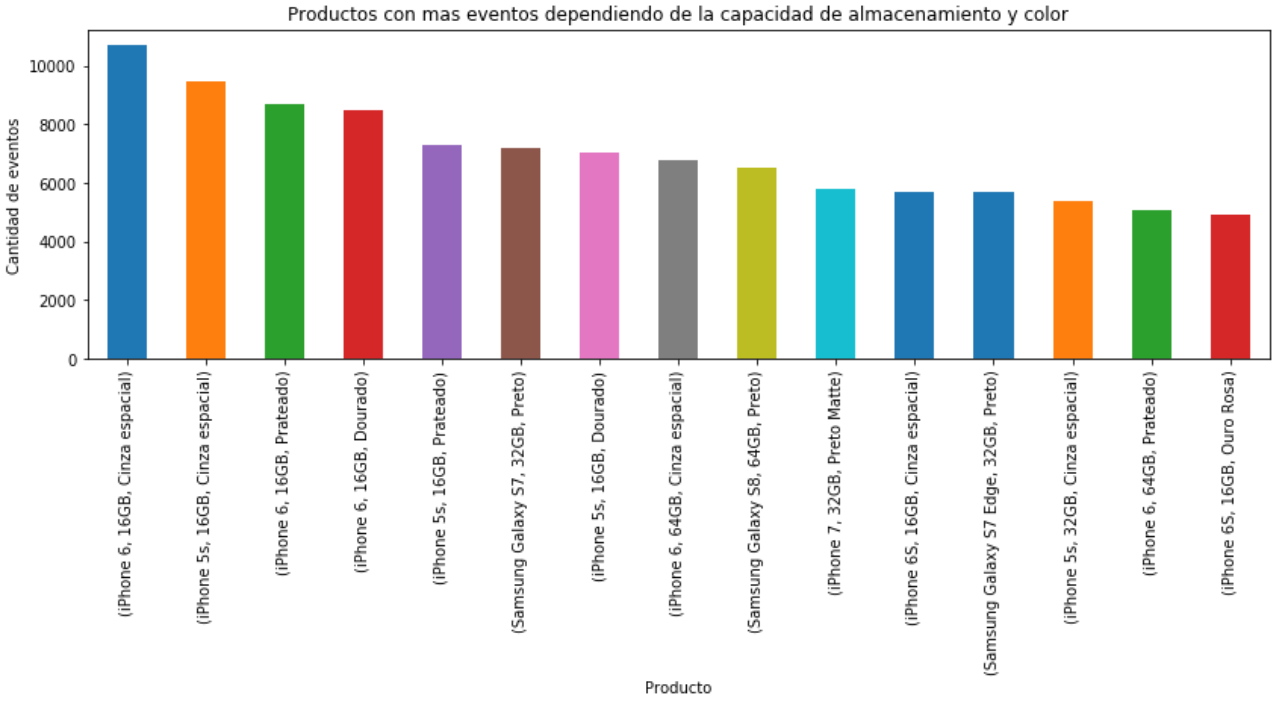
Nuevamente, la forma de ambos gráficos se condice con la información presentada previamente, dando coherencia al informe.

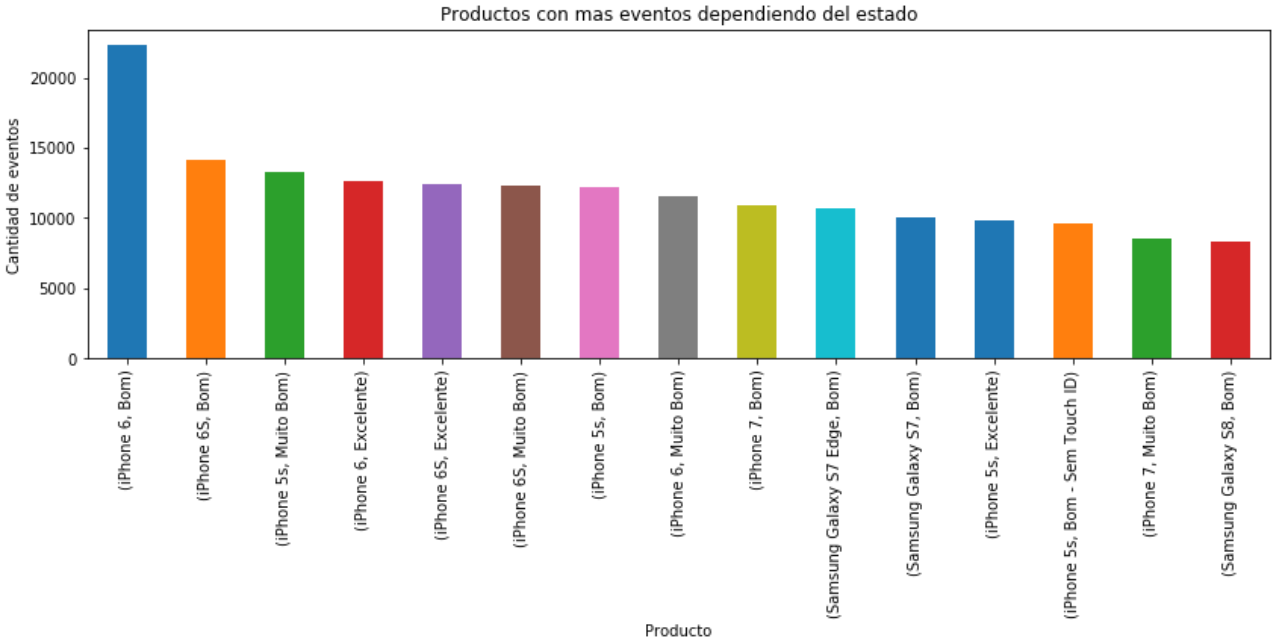
9. ¿Qué dispositivos están vinculados a estos eventos de producto?



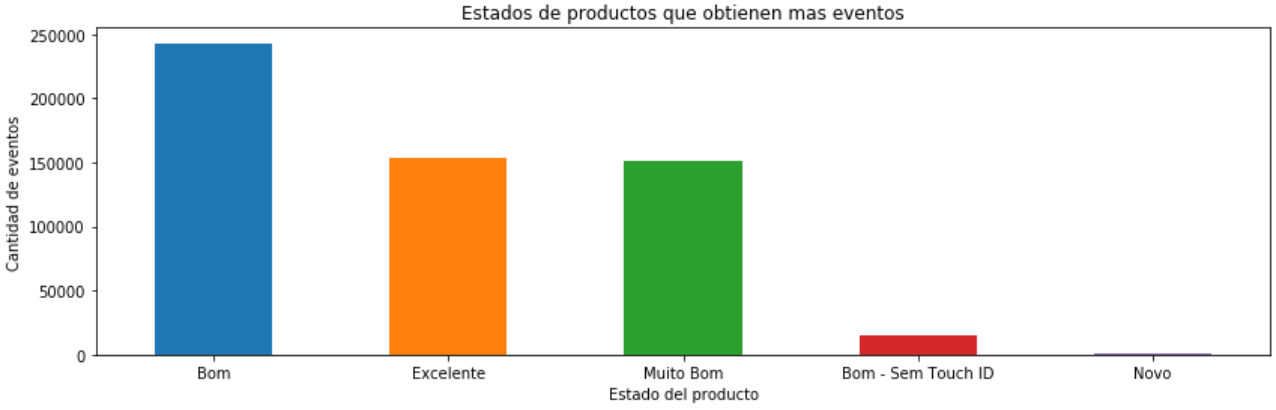
Una vez más se verifica que el smartphone manufacturado por la emblemática firma de la manzana lidera el ranking. De entre los 15 productos más frecuentes en esta clase de eventos, 9 de ellos corresponden a eventos vinculados a iPhones.


Productos con mas eventos dependiendo de la capacidad de almacenamiento
Cantidad de eventos
0
5000
10000
15000
20000
25000
(iPhone 6, 16GB)
(iPhone 5s, 16GB)
(Samsung Galaxy S7 Edge, 32GB)
(Samsung Galaxy S7, 32GB)
(iPhone 6S, 16GB)
(Samsung Galaxy S8, 64GB)
(iPhone 6, 64GB)
(iPhone 7, 32GB)
(iPhone 5s, 32GB)
(iPhone 6S, 64GB)
(iPhone 7, 128GB)
(Samsung Galaxy S6 Flat, 32GB)
(Samsung Galaxy J5, 16GB)
(iPhone 7 Plus, 128GB)
(Samsung Galaxy J7 Prime, 32GB)
Producto

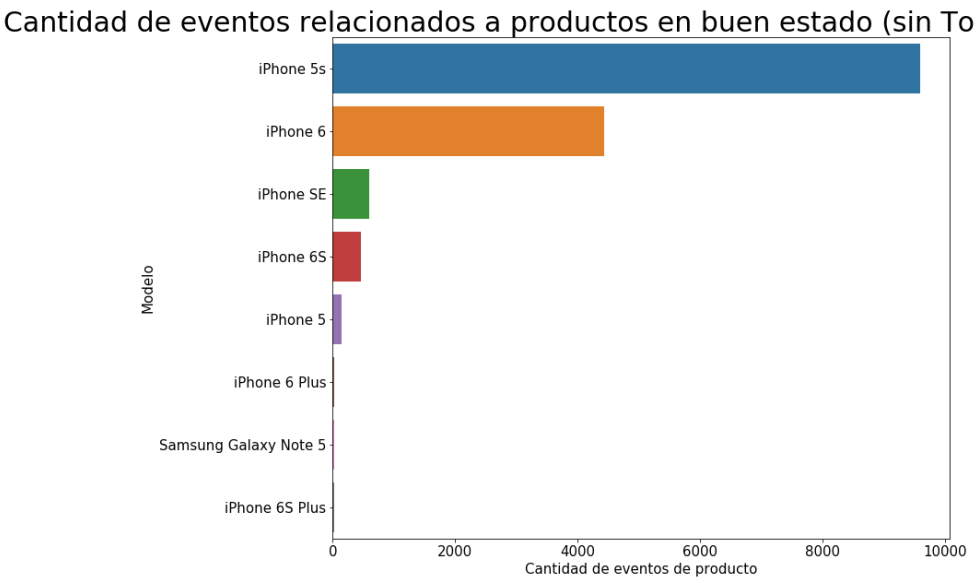

Productos con mas eventos dependiendo de la capacidad de almacenamiento y color
Cantidad de eventos
0
2000
4000
6000
8000
10000
(iPhone 6, 16GB, Cinza espacial)
(iPhone 5s, 16GB, Cinza espacial)
(iPhone 6, 16GB, Prateado)
(iPhone 6, 16GB, Dourado)
(iPhone 5s, 16GB, Prateado)
(Samsung Galaxy S7, 32GB, Preto)
(iPhone 5s, 16GB, Dourado)
(iPhone 6, 64GB, Cinza espacial)
(Samsung Galaxy S8, 64GB, Preto)
(iPhone 7, 32GB, Preto Matte)
(iPhone 6S, 16GB, Cinza espacial)
(Samsung Galaxy S7 Edge, 32GB, Preto)
(iPhone 5s, 32GB, Cinza espacial)
(iPhone 6, 64GB, Prateado)
(iPhone 6S, 16GB, Ouro Rosa)
Producto



10. ¿Qué puede decirse del estado en el que se encuentran los productos vinculados a estos eventos?

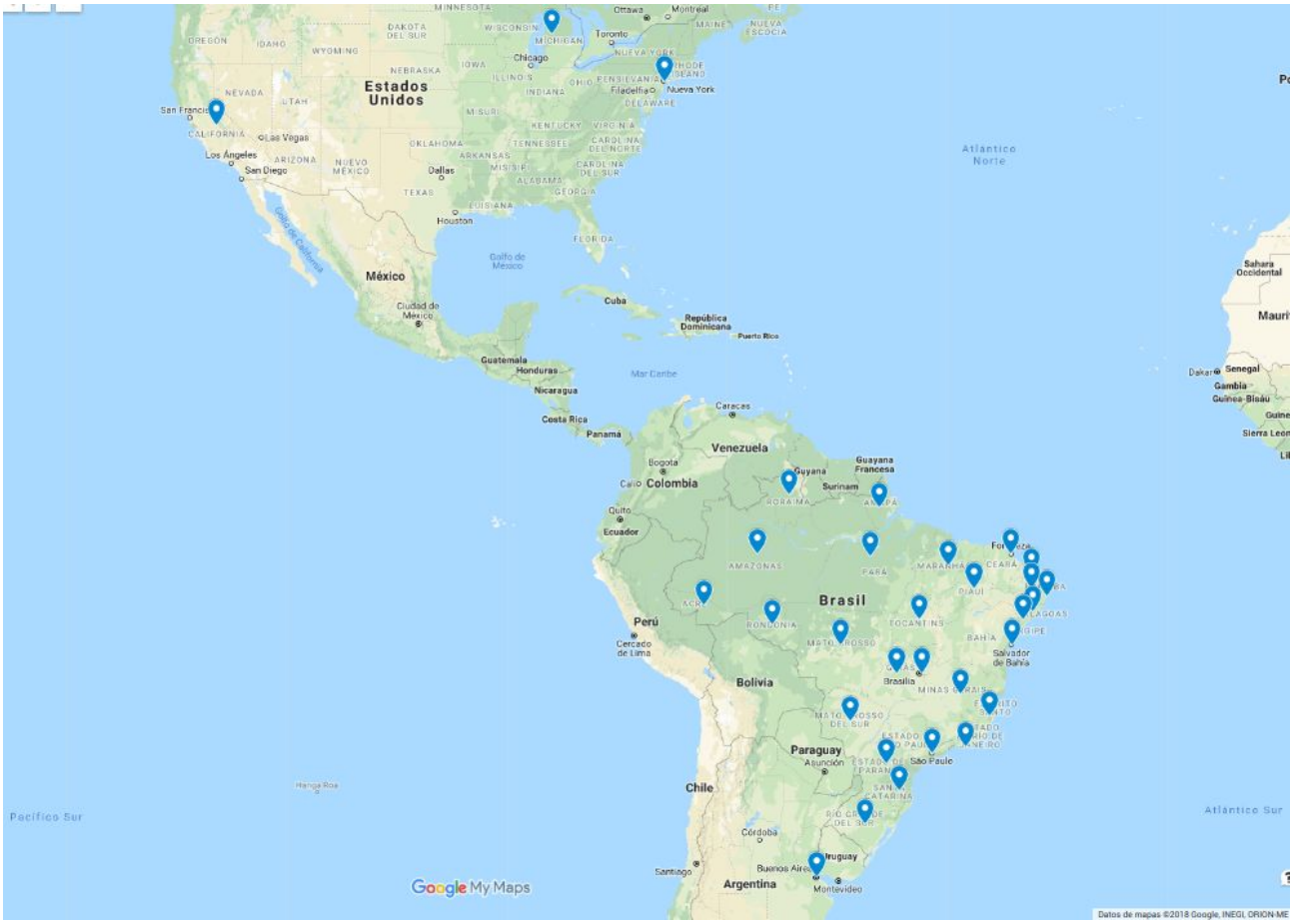


Si analizamos en detalle el estado ‘Bom – Sem Touch ID’ (bueno, pero sin touch ID) notamos que solamente hay 7 modelos de smartphone que presentan este estado, siendo 6 de ellos iPhones. Podemos ver que igualmente la cantidad de eventos (o productos) asociados a este estado es muy baja en relación al resto.



11. ¿En qué lugares se registró mayor cantidad de eventos?

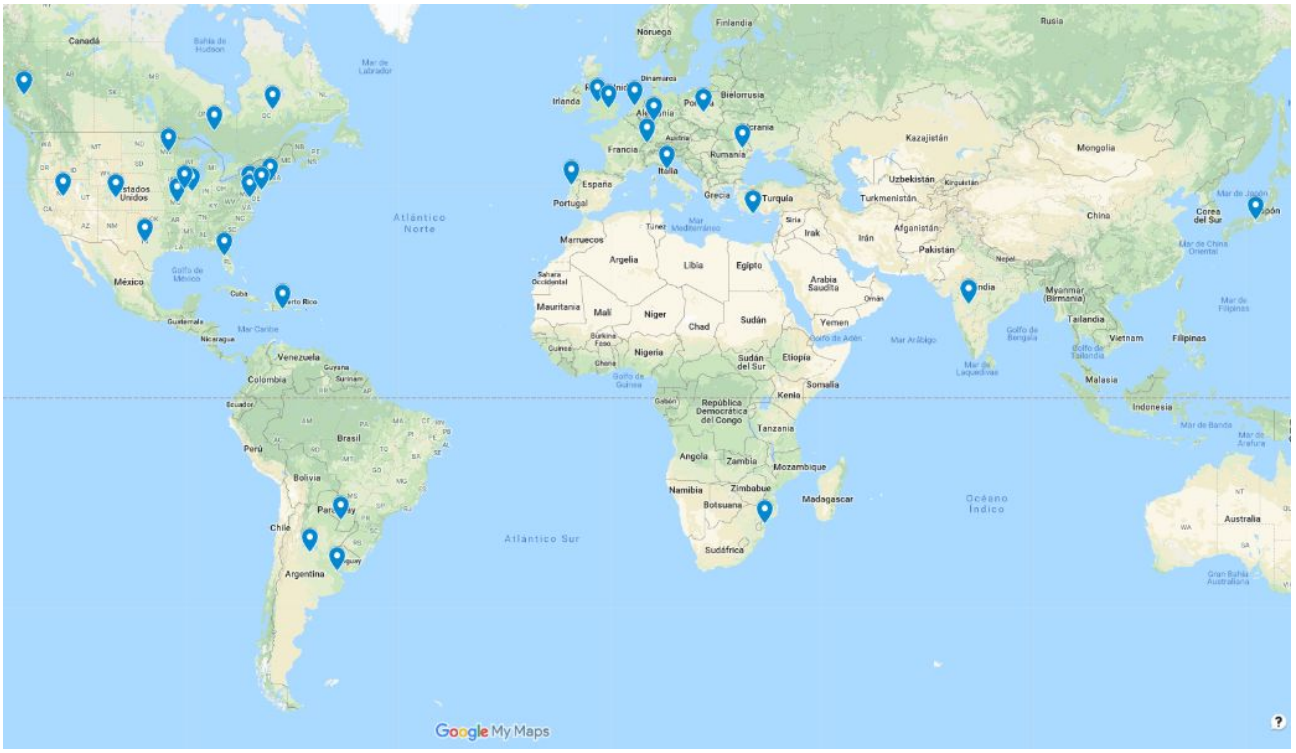
Gracias a la inclusión en el notebook de código y funciones de Python, y consultando la latitud y longitud de los lugares más frecuentes, hemos podido generar un archivo kml, y una imagen que muestra el resultado de este análisis.



Como era de esperarse, la mayoría de las locaciones frecuentes se concentra en Brasil, y acompañan a éstas Buenos Aires F.D. y otras 3 ciudades de EEUU.

12. ¿En qué lugares se registró menor cantidad de eventos?

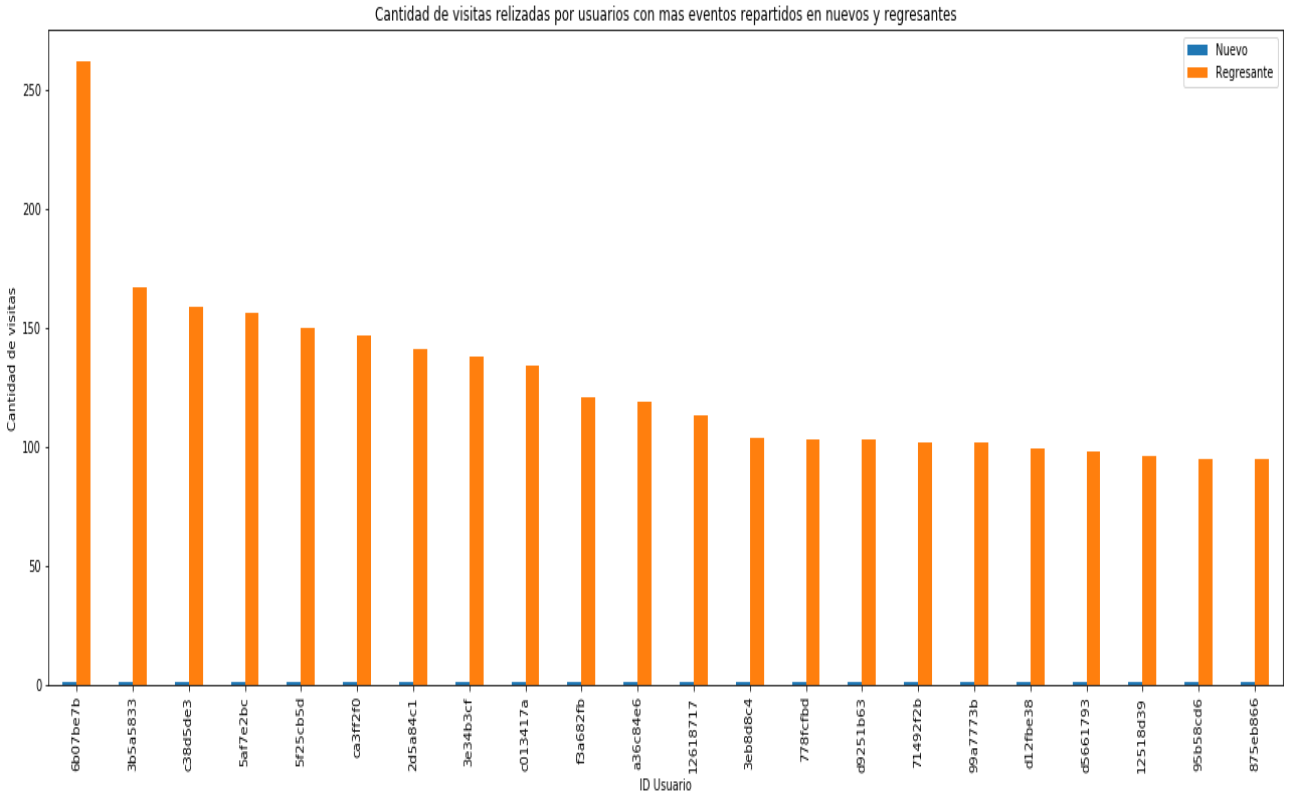
Replicando la lógica anterior, se generó otro archivo kml, y una nueva imagen con él.

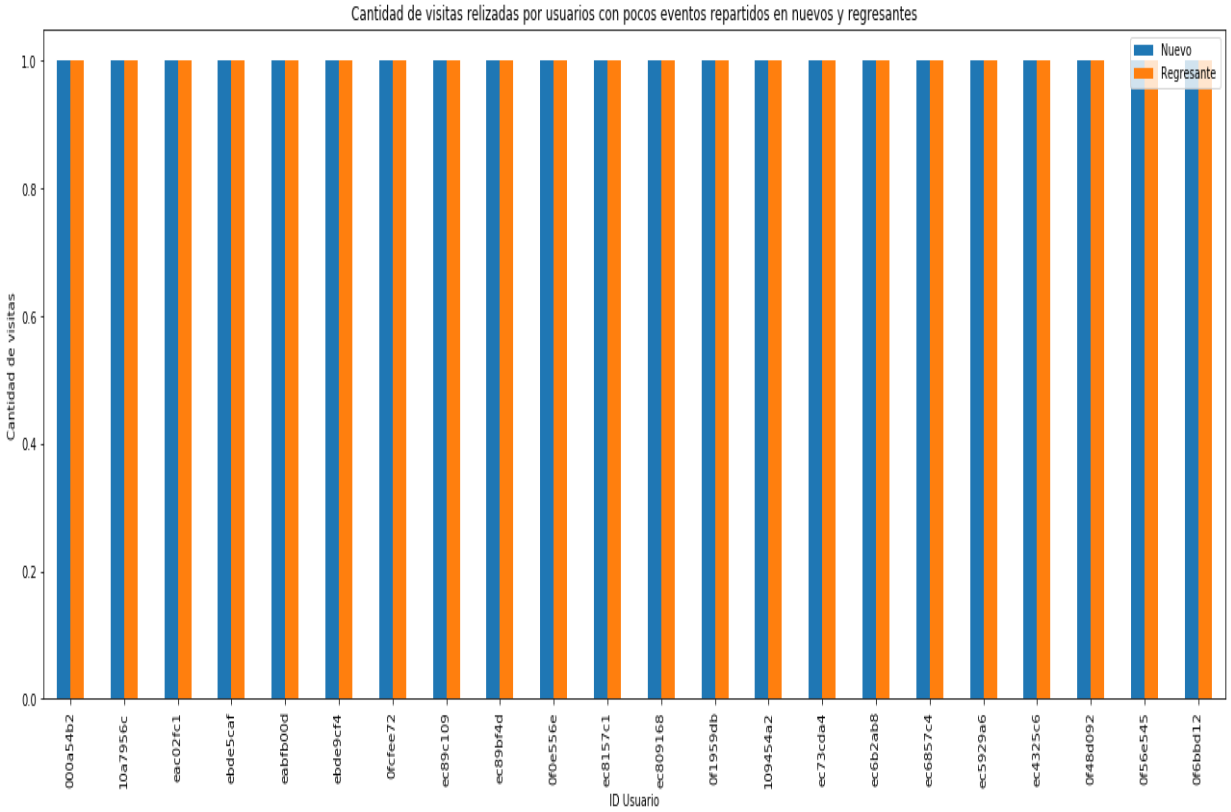


Ahora no se ven ciudades brasileñas en el mapa, y sí se pueden ver puntos aislados en otros lugares del mundo. Es esperable que en lugares como los indicados no haya un alto número de eventos. Dentro de Argentina, se incluyen las ciudades de Córdoba y Buenos Aires.

13. ¿Qué podemos decir del comportamiento de los usuarios?

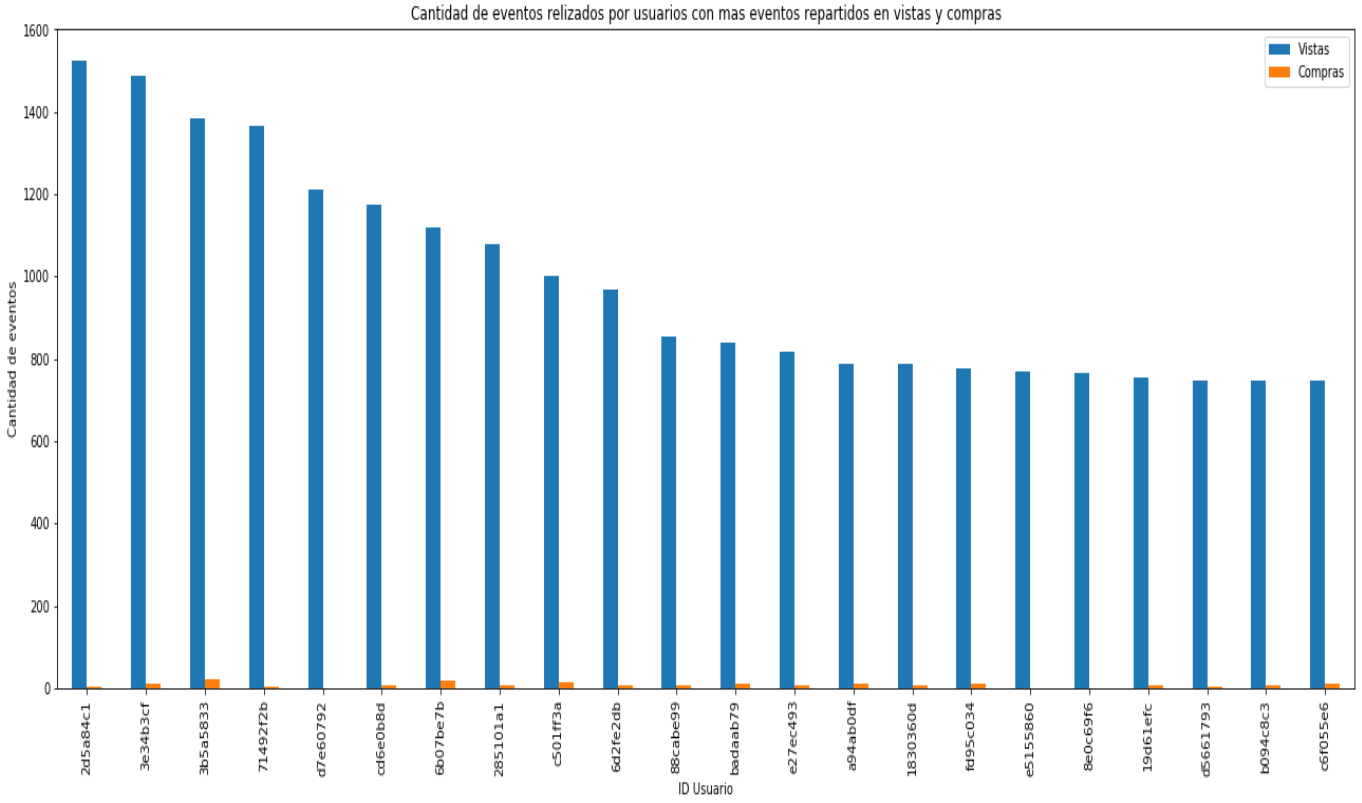
Analizando los eventos de las visitas al sitio de los usuarios nuevos y regresantes, llegamos a la siguiente comparación:

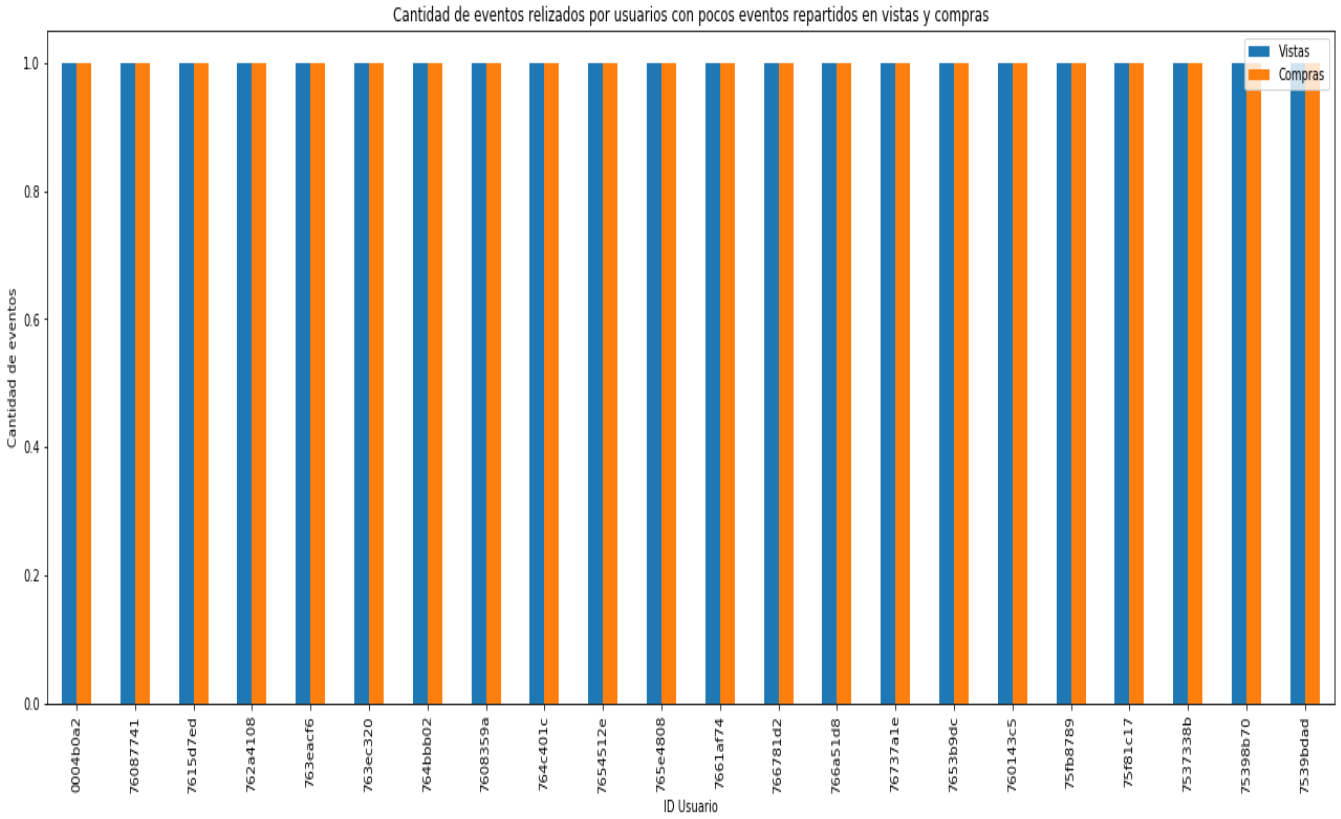

Cantidad de visitas relizadas por usuarios con mas eventos repartidos en nuevos y regresantes
Nuevo
Regresante
Cantidad de visitas
0
50
100
150
200
250
6b07be7b
3b5a5833
c38d5de3
5af7e2bc
5f25cb5d
ca3ff2f0
2d5a84c1
3e34b3cf
c013417a
f3a682fb
a36c84e6
12618717
3eb8d8c4
778fcfbd
d9251b63
71492f2b
99a7773b
d12fbe38
d5661793
12518d39
95b58cd6
875eb866
ID Usuario


Cantidad de visitas relizadas por usuarios con pocos eventos repartidos en nuevos y regresantes
Nuevo
Regresante
Cantidad de visitas
0.0
0.2
0.4
0.6
0.8
1.0
000a54b2
10a7956c
eac02fc1
ebde5caf
eabfb00d
ebde9cf4
0fcfee72
ec89c109
ec89bf4d
0f0e556e
ec8157c1
ec809168
0f1959db
109454a2
ec73cda4
ec6b2ab8
ec6857c4
ec5929a6
ec4325c6
0f48d092
0f56e545
0f6bbd12
ID Usuario

Vemos que hay varios usuarios que eran nuevos visitando el sitio y luego volvieron, realizando varias visitas más. Por otro lado, hay algunos usuarios nuevos que visitaron el sitio y volvieron una sola vez, como ilustra el segundo gráfico.

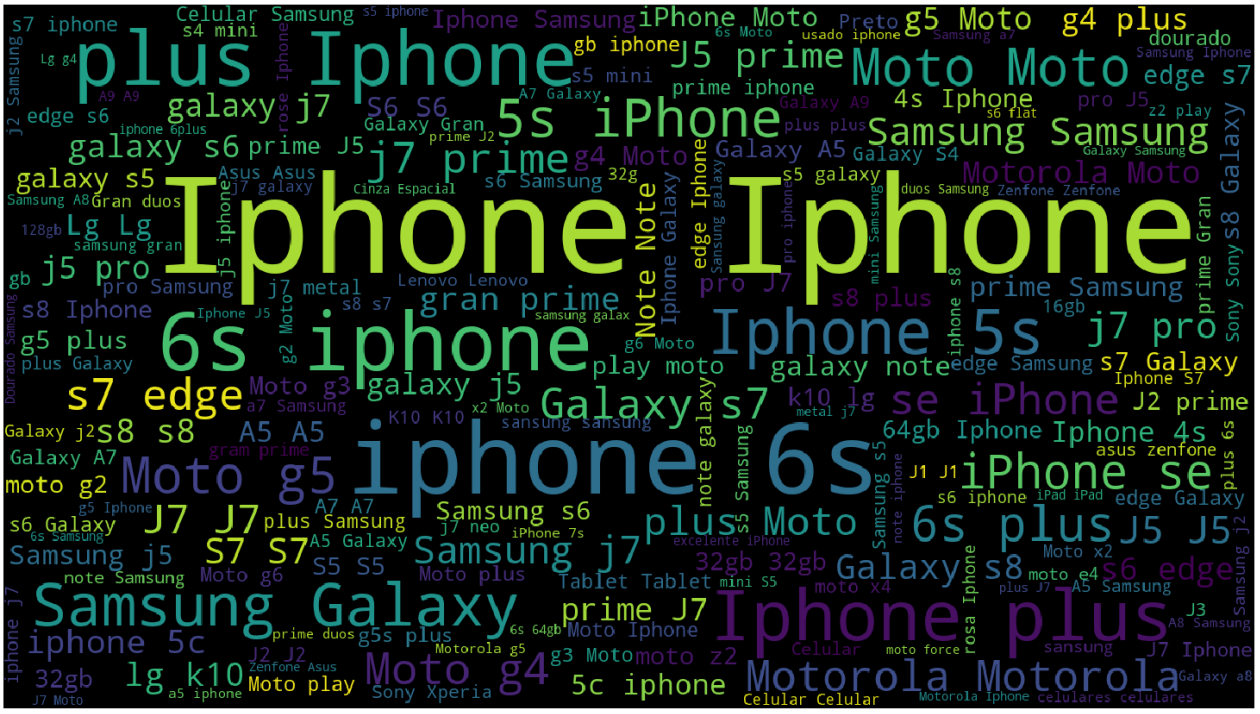
Analizando estos casos puntuales, para ver si los usuarios que visitaron el sitio una vez, al menos terminaron comprando un producto, llegamos al siguiente plot:



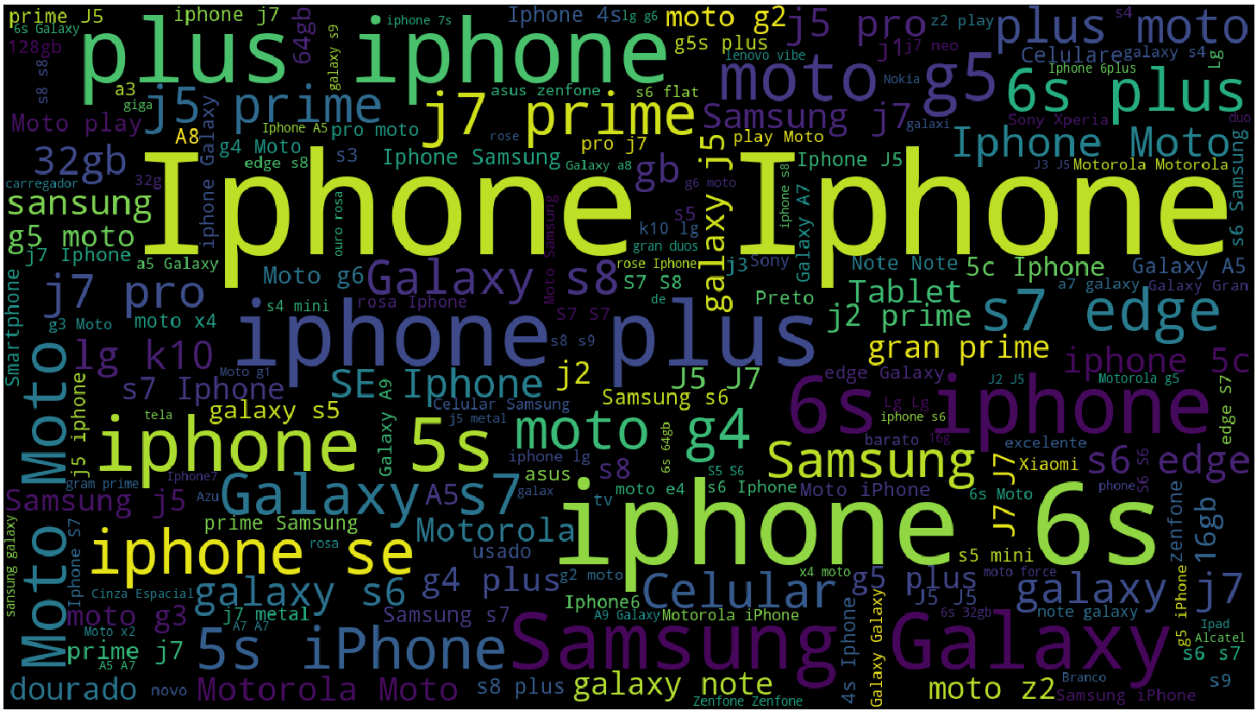


Vemos que hay varios usuarios que registran varias visitas al sitio, y que terminan comprando al menos un producto. Asimismo, también se puede apreciar que hay algunos usuarios que visitaron el sitio solo una vez y compraron únicamente un producto.

14. ¿Qué términos fueron los más buscados dentro del sitio?

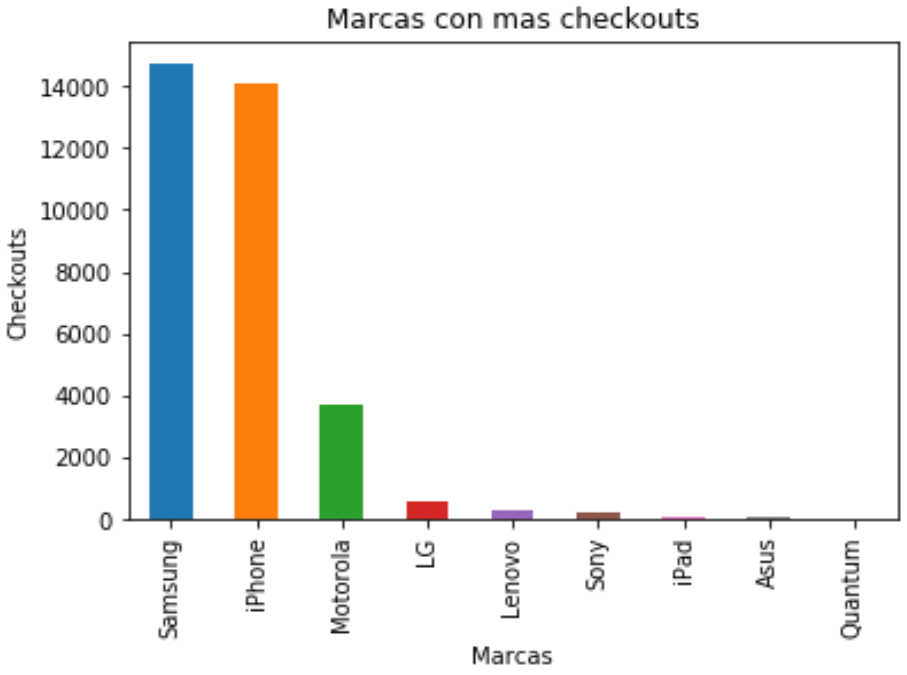


Un análisis rápido de los términos más buscados en el sitio web de Trocafone, combinado con una visualización de tipo wordcloud, refuerzan la prevalencia del smartphone iPhone por sobre otros modelos, que si bien son buscados, no se destacan tanto.

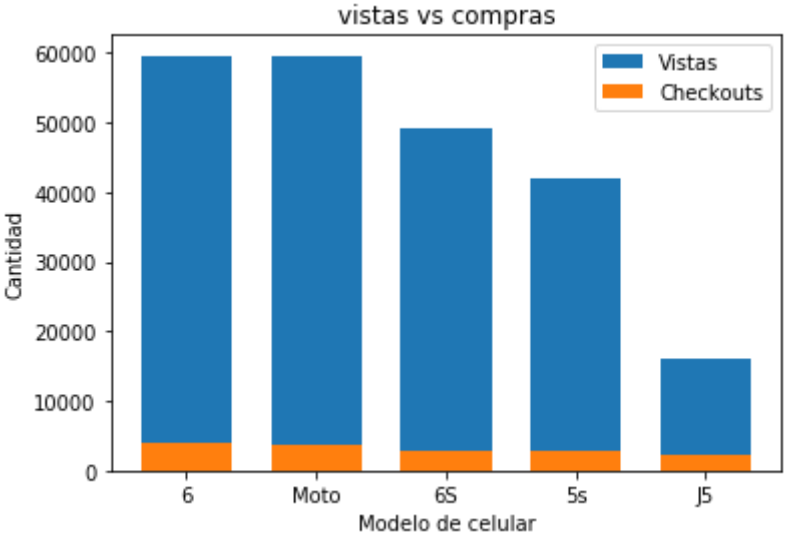


Si repetimos el análisis pero removiendo duplicados y agrupando por usuario y término de búsqueda, obtenemos una visualización muy parecida, en la que el modelo de celular destaca aún más que antes.

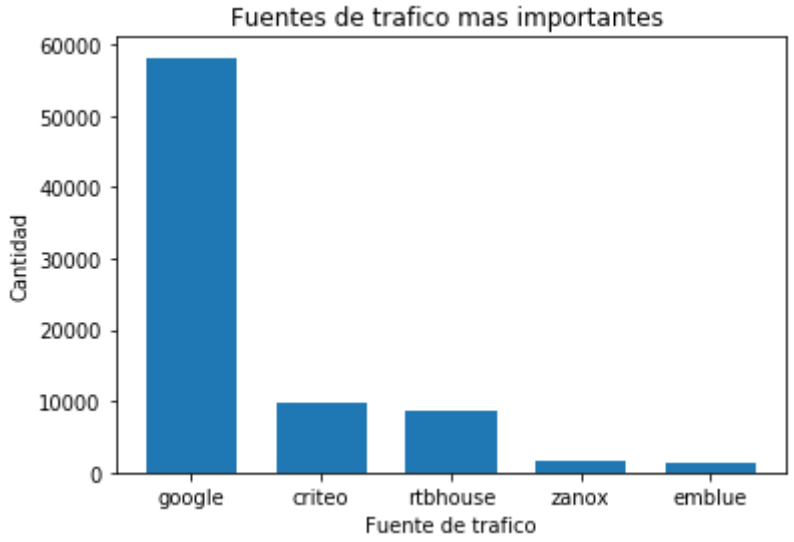
15. ¿Cuales fueron las marcas con más eventos tipo checkout?



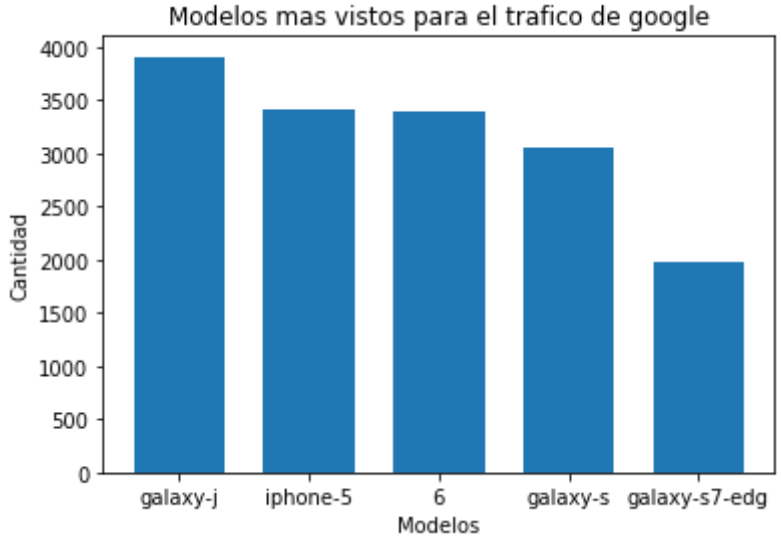
16. ¿Cuales fueron los modelos más vistos y cuantos fueron eventos de tipo checkout que tuvieron?



17. ¿Cuales fueron las fuentes de tráfico más importante?



18. ¿Cuales fueron los modelos más vistos para comprar según la fuente de tráfico más importante?



Conclusiones

Procedemos a enumerar las conclusiones del análisis exploratorio, que a nuestro parecer, resultan más interesantes:

El mayor volumen de visitas a la página, y de eventos relacionados a producto (checkout, compra y visualización de producto) se registra de lunes a viernes entre el mediodía y la medianoche.

En mayo 2018, en particular durante la segunda quincena del mes, se registra un máximo de visitas a la página, así como de eventos de producto. Se cree que esto puede estar asociado a la publicidad que se emitió en ese momento.

El dispositivo más buscado y visualizado es el iPhone, en toda su gama de modelos. Lidera el ranking de dispositivos dentro de los eventos de producto. Las búsquedas disparadas desde dispositivos que utilizan el sistema operativo iOS representan un porcentaje muy bajo dentro del total de búsquedas, dando a entender que muchos usuarios que no tienen iPhone están interesados en adquirir uno.

Las interacciones con el sitio web a través de Google Chrome predominan, así como el uso de computadoras y smartphones para acceder al mismo. La combinación más frecuente viene dada por el sistema operativo Windows y el navegador Chrome para el caso de las computadoras.

El análisis de keywords refuerza la suposición sobre el alto interés/tendencia del público de Trocafone hacia el consumo de dispositivos iPhone. No solamente del smartphone sino que también se registran búsquedas de protectores de pantalla.

Casi no hay interés en productos que tienen buena condición, pero no traen Touch ID. Cabe destacar que casi todos éstos corresponden a modelos de iPhone o Samsung Galaxy Note 5, y siendo que este smartphone es tan popular de por sí, este estado no aporta en el análisis.

El análisis geográfico muestra puntos aislados con pocas consultas, y que están muy lejos de Brasil. Estos datos pueden ser ruido o linkearse a usuarios de la plataforma que acceden a la misma estando

temporalmente fuera de su zona habitual. Tiene lógica, además, que la mayoría de los eventos se disparen desde Brasil.

En cuantos a los eventos de tipo checkout o compras, predominan dos marcas con una amplia mayoría sobre las demás. Estas son apple y Samsung. Esto no sorprende ya que son las dos empresas líderes de ventas en el mundo. Por detrás se ubican motorola, sony entre otras.

Analizando el campo “campaign_source” observamos que las publicidades más exitosas en términos de cantidad visitas obtenidas son las de la empresa google en primer lugar, con una muy buena diferencia con respecto a la segunda y tercera, criterio y rbhouse respectivamente.