Informe TP1

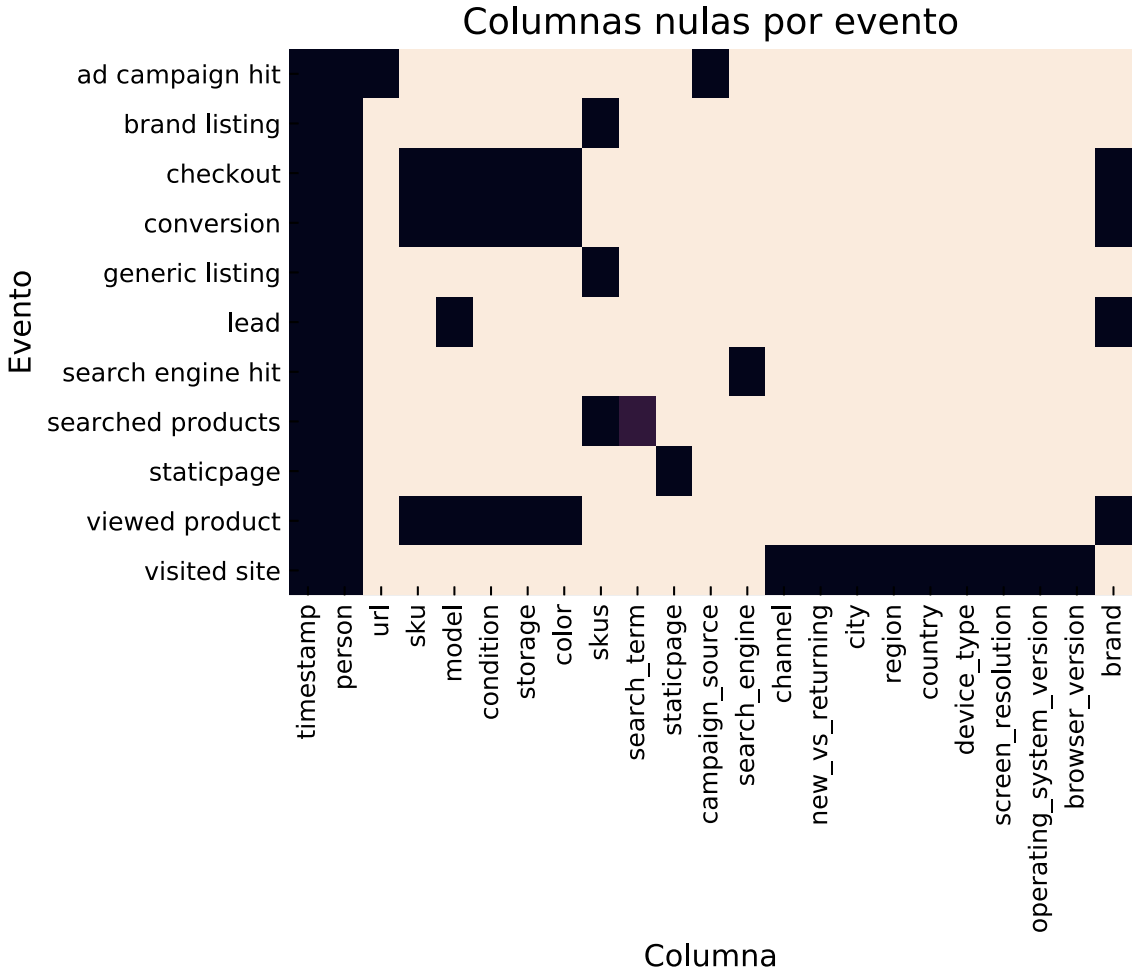
Grupo null

- Carlos Talavera
- Federico Jure
- Juan Pablo Capurro

Introducción

Pre-análisis del set de datos

Al estar todos los eventos en un mismo dataframe, creimos que los eventos iban a tener una cantidad importante de columnas nulas dependiendo del tipo de evento.



Dependiendo del tipo de evento, las columnas son nulas o no en un 100%, con excepción del campo search_term en el que hay una pequeña proporción de nulos. En definitiva, podemos decir que este set de datos es bastante consistente en cuanto a los datos de los cuales podemos sacar conclusiones directamente.

Gracias al gráfico de los datos nulos, podemos identificar rápidamente que features relacionar con cuál para sacar conclusiones.

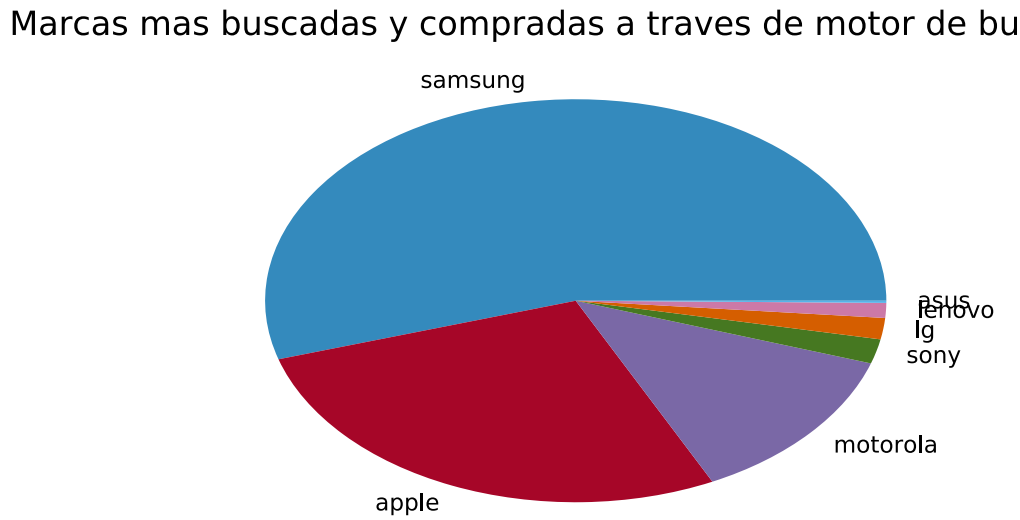
Por ejemplo, observando el gráfico podemos ver que podemos relacionar directamente el campo *conversion* con los campos *sku, model, condition y storage*. Es decir, podemos observar si vale la pena sacar conclusiones de las ventas realizadas de acuerdo a sku, el modelo, la condición y el tamaño de la memoria del dispositivo. También, nos sirve para descartar ideas que no son factibles directamente. Por ejemplo, no podemos relacionar directamente la cantidad de *conversion* con el campo *city* o *region*, ya que todos los tipos de eventos que surgieron como resultado de una compra, no tienen estos campos.

Este análisis previo de como se relacionan los campos del dataset entre sí nos ahorró bastante tiempo a la hora de saber por qué lado encarar los análisis.

Nueva feature: Marca del dispositivo

Analizando el set decidimos que podría ser bastante útil e interesante hacer análisis no solo de acuerdo al modelo del dispositivo, sino también de acuerdo a la marca dueña del mismo.

Esta nueva feature nos permitiría analizar los resultados discriminando por empresas, y no solo por modelo. Por ejemplo evaluar cuál es la empresa de más renombre en la venta de dispositivos usados. La mayoria de las marcas tienen varios modelos. Y esta nueva feature nos permitira tener una visión más global de los agentes influyentes en el set de datos.



Exploracion

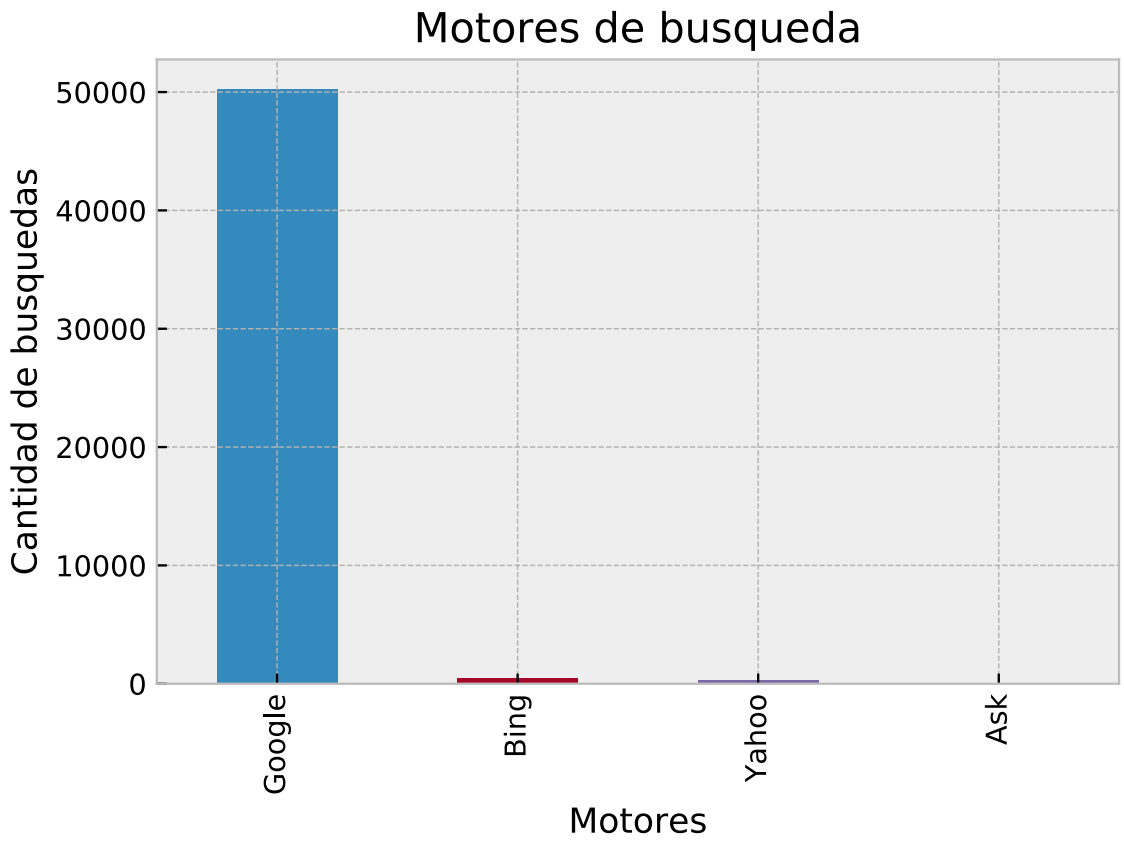
Exploración de tipos de evento por separado

Eventos de búsqueda

Procedemos a ver si hay registros inválidos de busquedas, y si amerita dropear registros. Por un lado, hay una proporción importante (7k nulos en 56k total) de eventos de busqueda que tienen NaN como search_term, pero tienen distintas listas de skus, por lo que podemos suponer que hay otros factores que afectan a la busqueda.

Busquedas por motores

La idea de este análisis era evaluar cuál es el motor de busqueda más usado para llegar a los productos de la página. Al principio, supusimos que el motor de búsqueda más usado iba a ser Google. Dado que es uno de los motores de busquedas más usados a nivel global y este dato es de conocimiento común. Sin embargo, decidimos llevar a cabo este análisis para terminar de confirmar (o no) nuestras hipótesis. Obtuvimos el siguiente resultado:



Tal y como era de esperarse, Google salió en primer lugar, por una diferencia abismal con el resto de los motores. Suponiamos que Google iba a ser el motor de búsqueda más usado, pero no teniamos idea de como iba a ser la relación respecto al uso de los otros motores de busqueda. Suponiamos que al menos el resto de los motores iban a sumar por lo menos un cuarto de las busquedas hechas por Google, pero ni siquiera se acercan. De hecho en el gráfico su presencia respecto a los 50 000 busquedas por Google es despreciable.

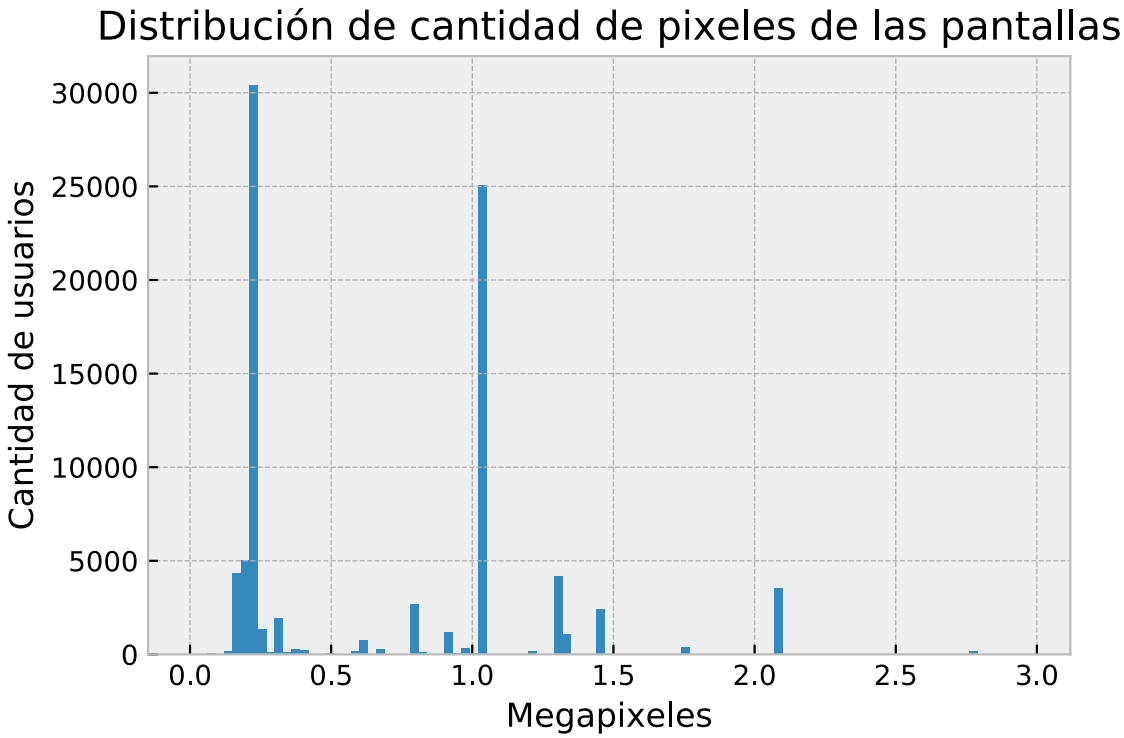
Curiosamente, la distribución del ranking se constrasta con este artículo publicado en este blog y en tantos otros.

Es decir, los motores de busqueda más usados a nivel global son: * Google * Bing * Yahoo * Ask

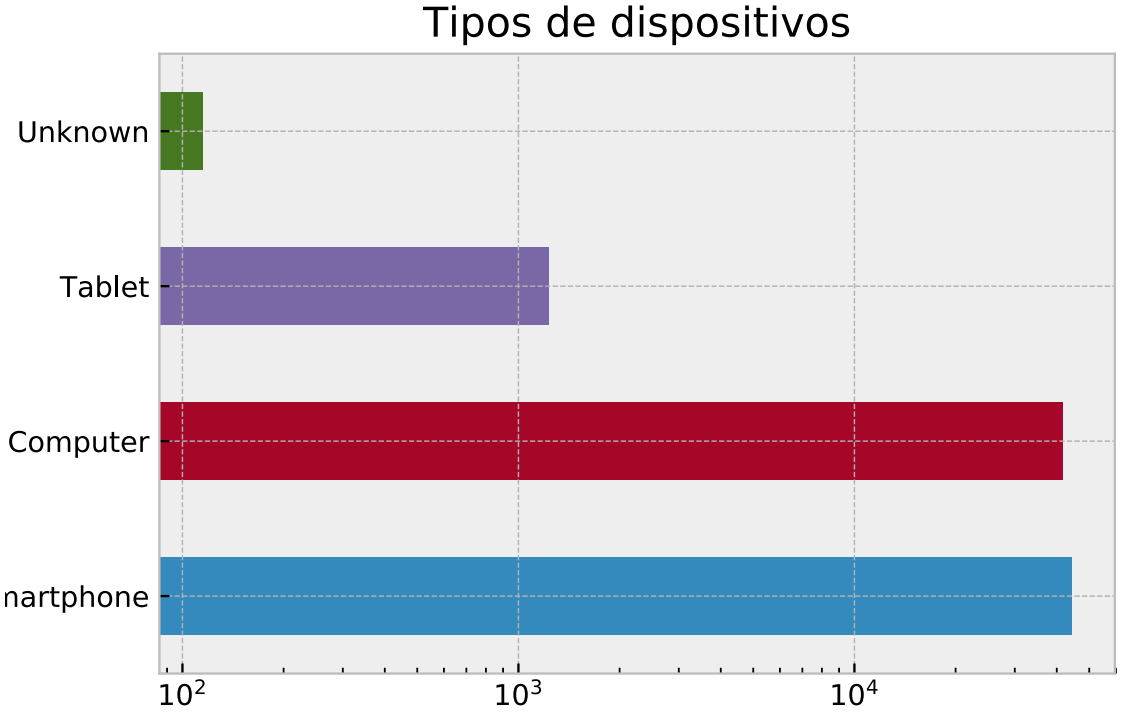
Lo cuál se refleja exactamente en nuestro análisis, y si nos ponemos a pensar, tiene bastante sentido. Ya que las proporciones a nivel global son relativamente equivalentes cuando lo analizás por tópicos aislados.

Eventos de visita de sitio

Consideramos la resolución de pantalla una forma de ver qué poder adquisitivo tienen las personas que visitan el sitio. Medimos la cantidad de píxeles de las pantallas, porque hay muchas variantes de resoluciones y solo nos importa el tamaño.

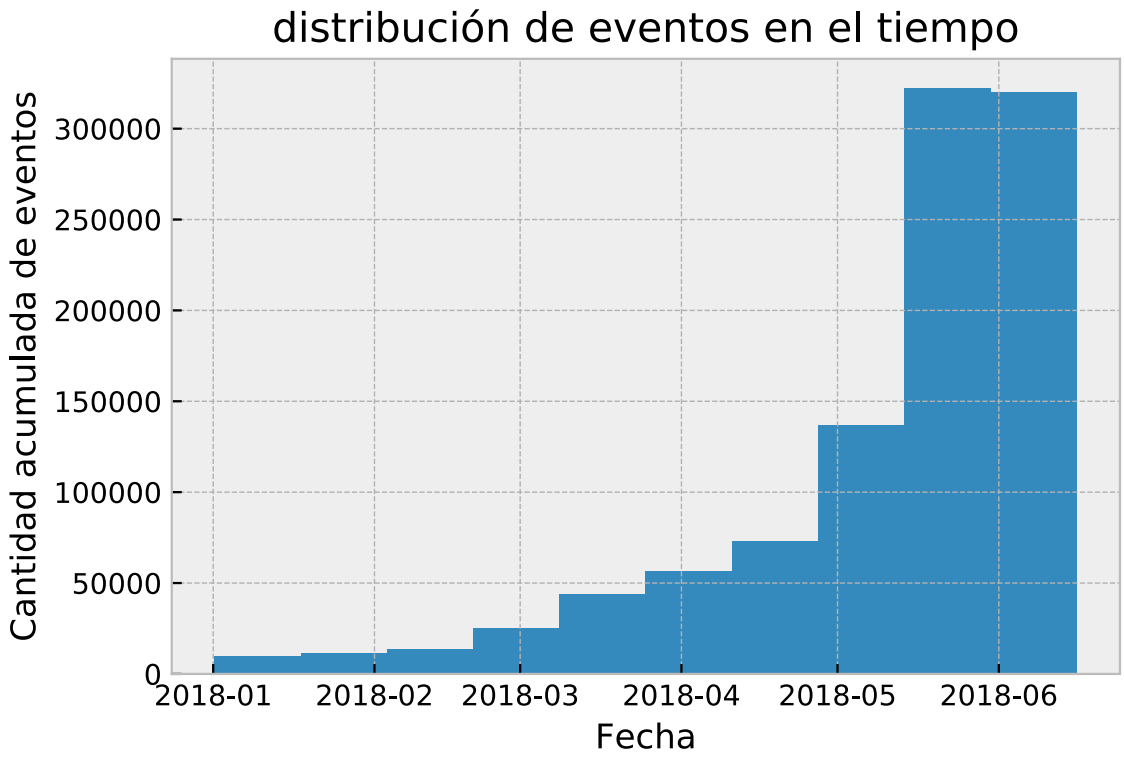


Nos interesó también que proporcion de los usuarios accedian desde mobile y cuántos desde desktop



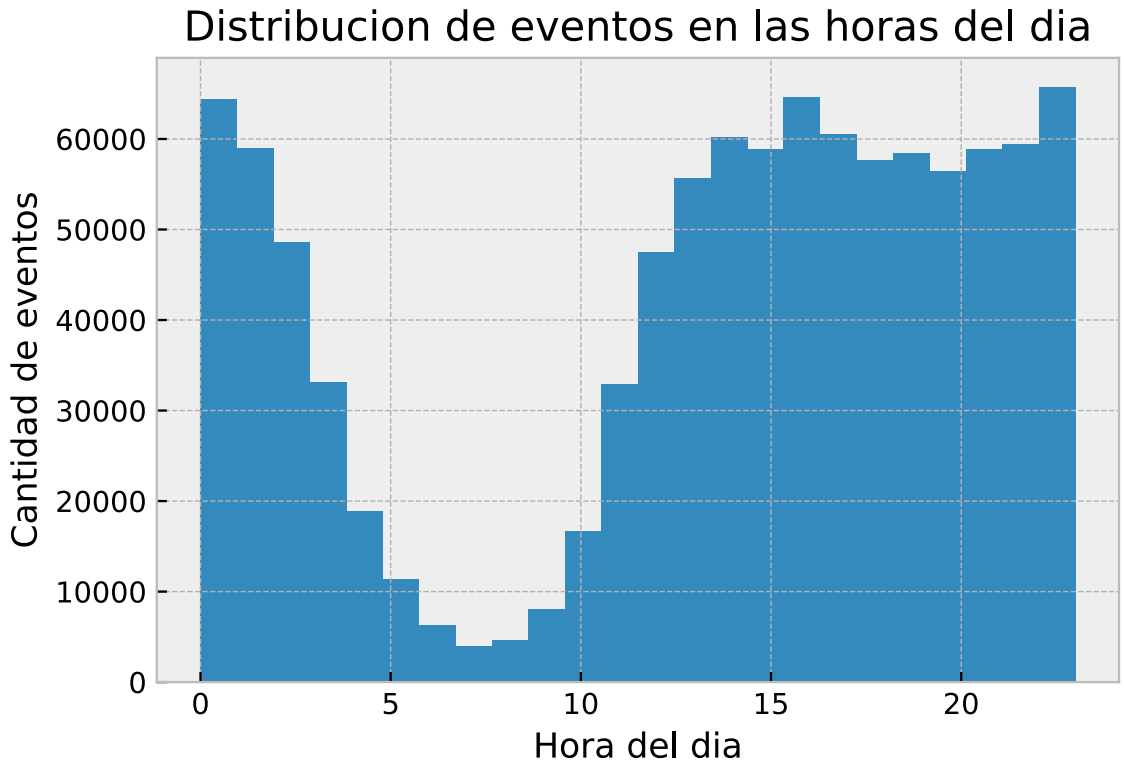
Exploracion de los eventos en conjunto

El uso de la plataforma aumentó enormemente a lo largo de los últimos meses:

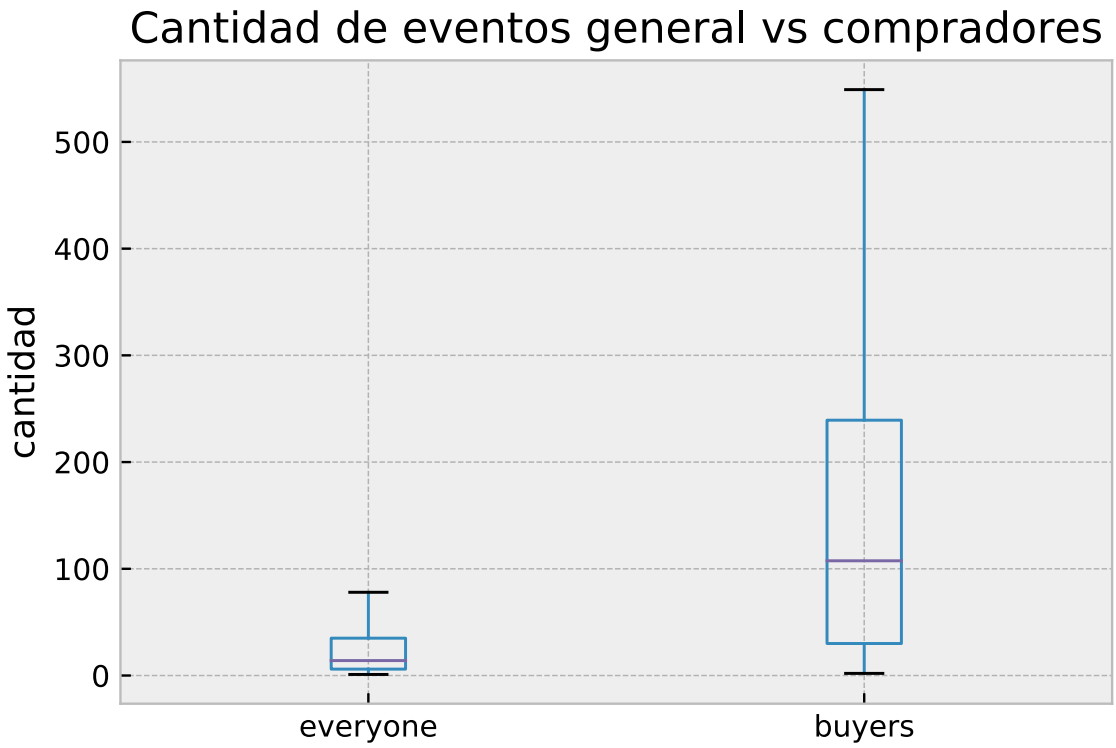


Hay una aparente caída de la cantidad de eventos en el último mes, pero esto es consecuencia de que los datos para este se encuentran truncados a mitad de mes.

Y la distribución de los eventos a lo largo del dia no nos da muchas sorpresas:



Los usuarios pueden tener una cantidad variable de eventos, y es usual que tengan algunos cientos, con outliers teniendo un par de miles. Estos outliers no aparecen en el gráfico porque lo volverían ilegible.

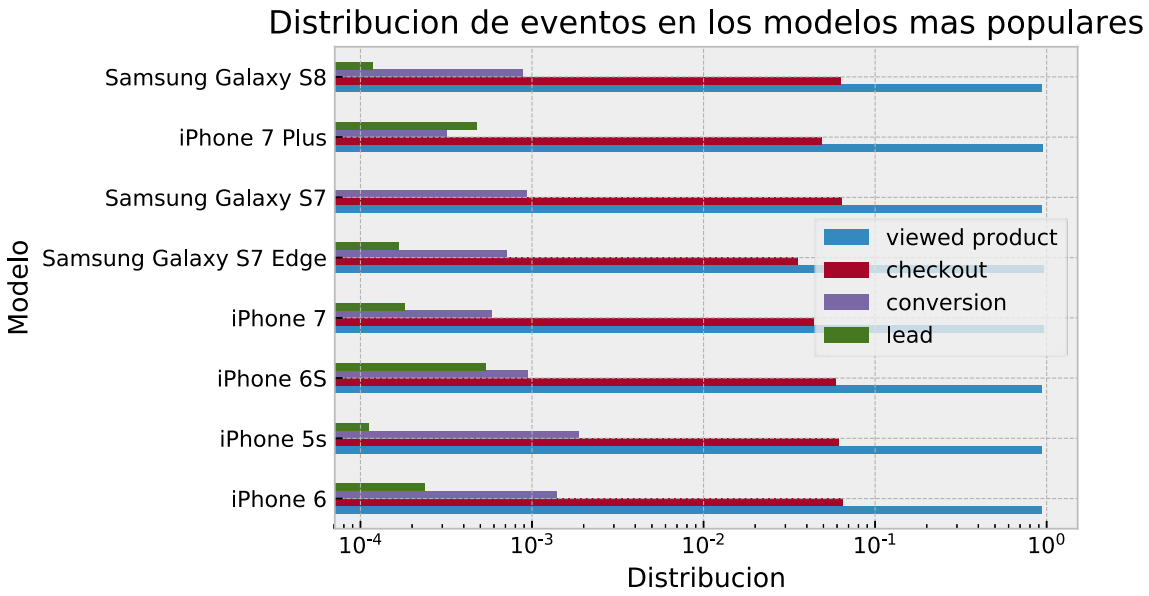


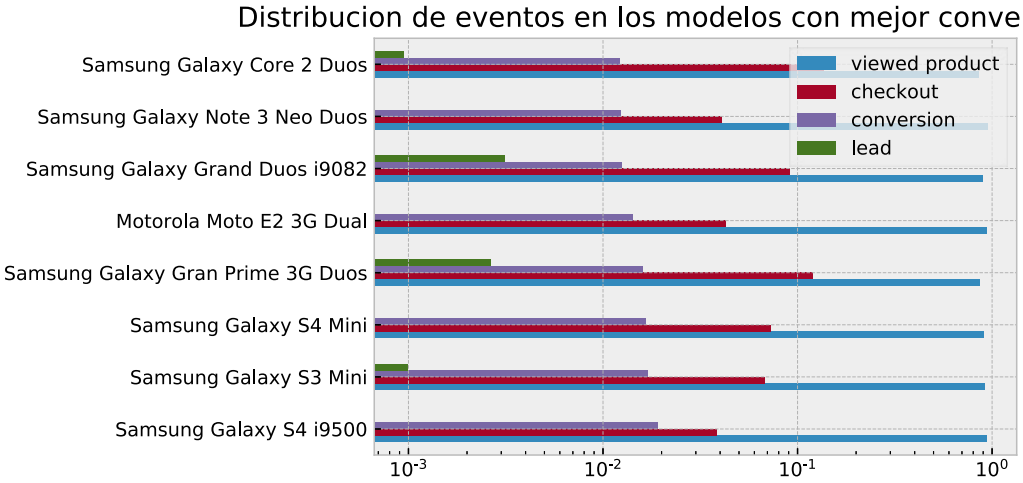
Los usuarios que realizan una conversión tienen en general muchos más eventos que el público en general.

Exploraciones de los distintos modelos

Encontramos que incluso filtrando aquellos modelos con menos de 100 eventos, buscar cuáles presentaban mejor ratio de leads introducía bastante ruido. Por ejemplo, aparecían modelos sin conversiones, por lo que consideramos esta columna relativamente desestimable.

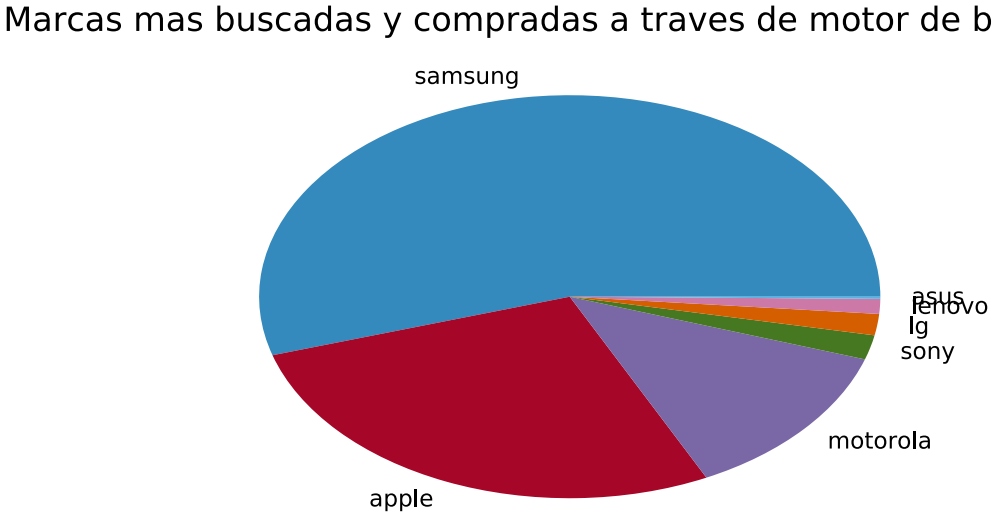
Por otro lado, encontramos que los modelos con más vistas no overlapean mucho con los que tienen mayor ratio de conversiones:





Ventas por mes y marca

En general las ventas por marca tienen una distribución algo parecidas al market share de las distintas empresas [a nivel global]

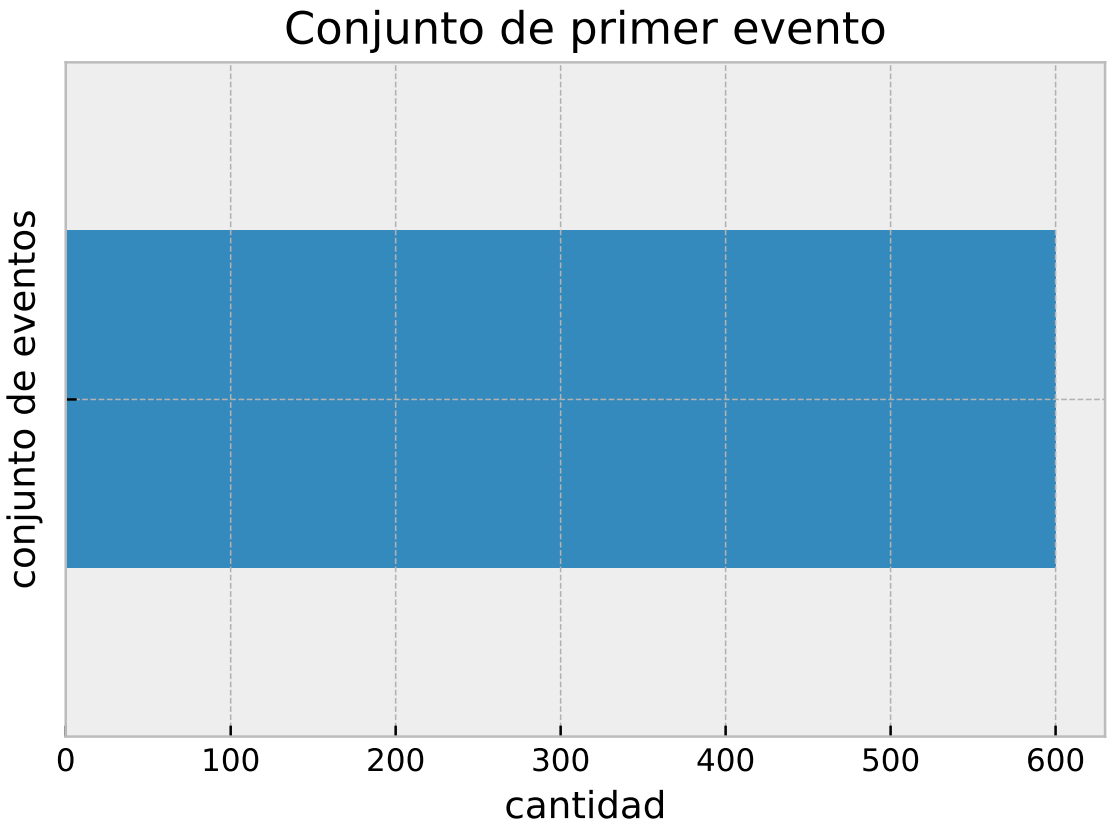


Pero se puede ver que Samsung tiene una parte mas grande de las ventas en la plataforma que a nivel global.

Las ventas por mes muestran que, por un lado, los datos están truncados en el ultimo mes, y que hubo pequeñas fluctuaciones en la proporción de ventas de Motorola respecto a las demás, pero siempre se mantuvo en primer lugar Samsung, seguido de Apple.

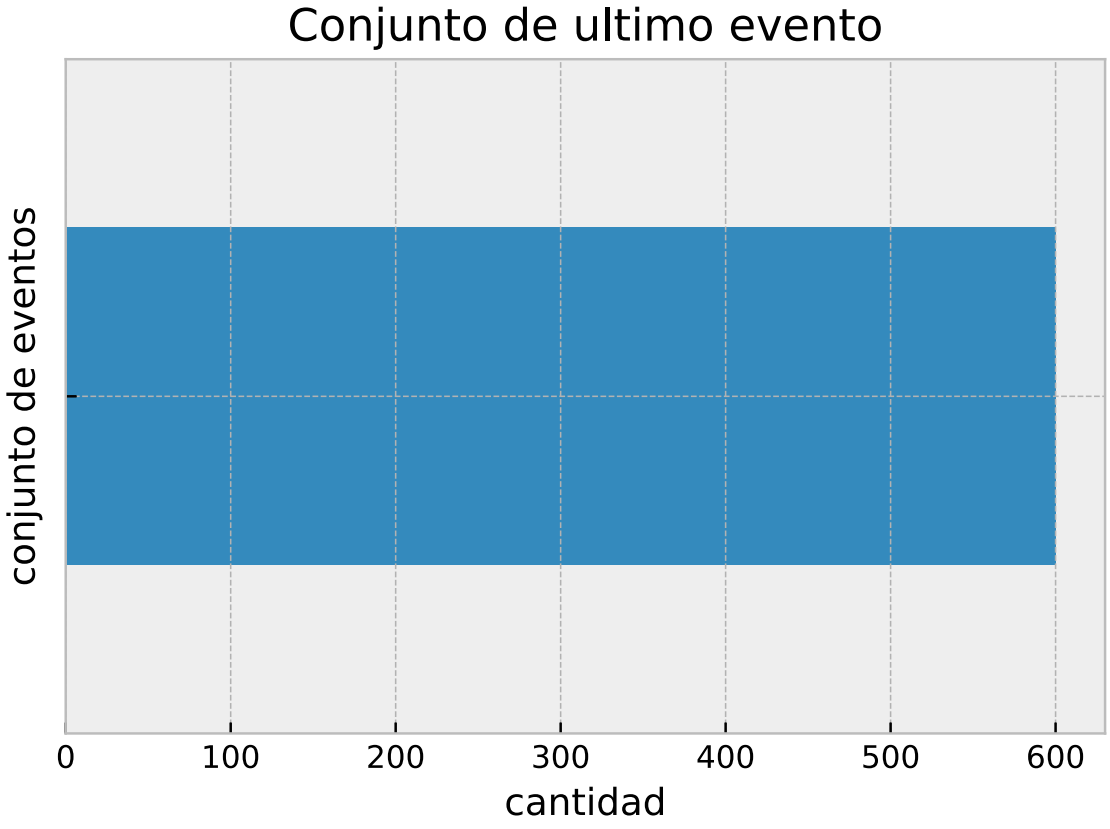
Exploración sobre los usuarios

Pudimos observar que en el caso del primer evento del usuario, se hallan en el mismo segundo varios otros eventos de tipos relacionados, que refieren a la misma accion pero desde distintos puntos de vista:



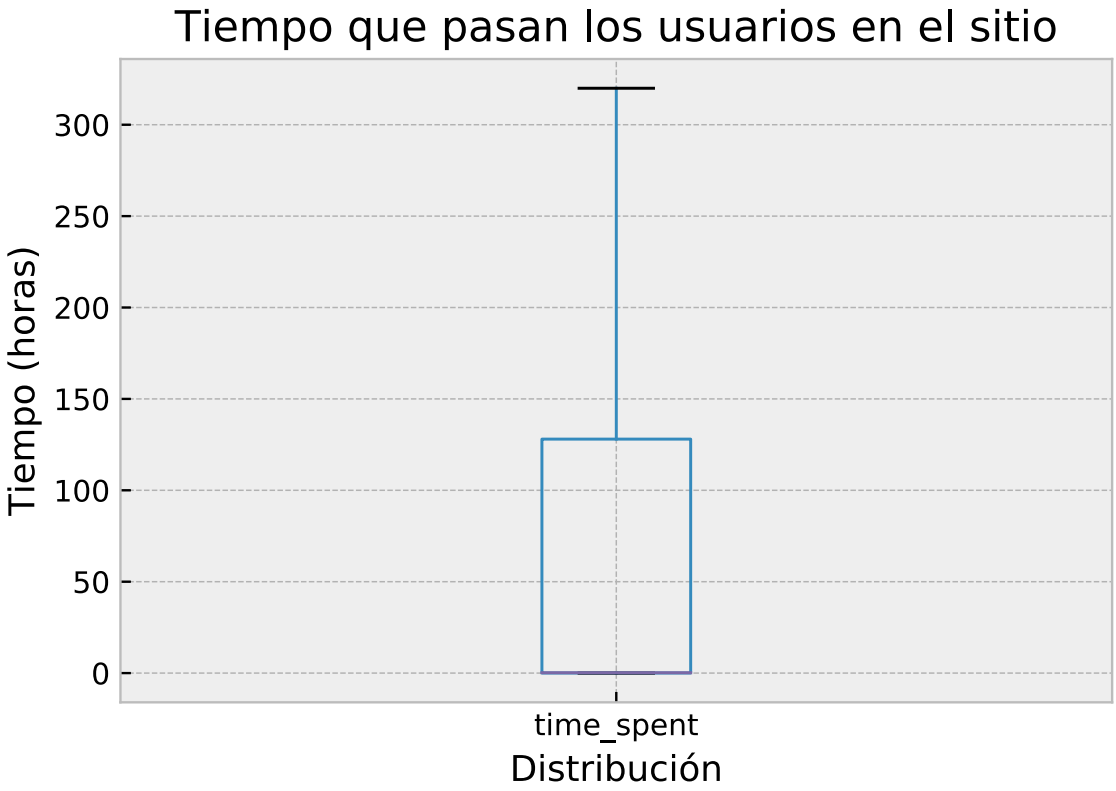
Por ejemplo, lo más usual es que se llegue a visitar el sitio por un ad campaign hit desde un search engine.

En cuanto a los ultimos eventos de un usuario, estos no suelen aparecer en grupos



Distribución temporal de los eventos

Los usuarios pasan cantidads de tiempo muy variadas en la plataforma, y se puede destacar que hay una proporción alta de outliers que tienen eventos separados por varios miles de horas. Estos no son mostrados en el gráfico porque lo volverían ilegible.



Por otro lado, no se encontró relación entre el tiempo que un usuario lleva usando el sitio y la cantidad de eventos de algún tipo en particular que genera.

Insights