Universidad de Buenos Aires

Facultad de Ingeniería

(75.06) Organización de Datos

Cátedra: Luis Argerich

TP 1

https://github.com/fran122157/datos-2018-2-TP1

2° Cuatrimestre 2018

Alumnos:

GARATE, Julián Matías (93043)

GIMÉNEZ ZALCMAN, Mariano Nicolás (99789)

VILLAREJO, Francisco Martin (99864)

Introducción

El presente trabajo, es sobre un análisis exploratorio de la base de datos de actividades de usuarios de la empresa TROCAFONE, la cual se caracteriza por ser un blog de compra-venta de celulares, tanto usados como nuevos.
Para el análisis de utilizar el lenguaje de programación **Python 3.6** con las librerías de **PANDAS** y para las visualizaciones **SEABORN**, **MATPLOTLIB**, **BOKEH** y **HOLOVIEWS**.
En el repositorio de Github se colocaron las correspondientes notebooks con los códigos fuente que validan los procedimientos y el análisis.

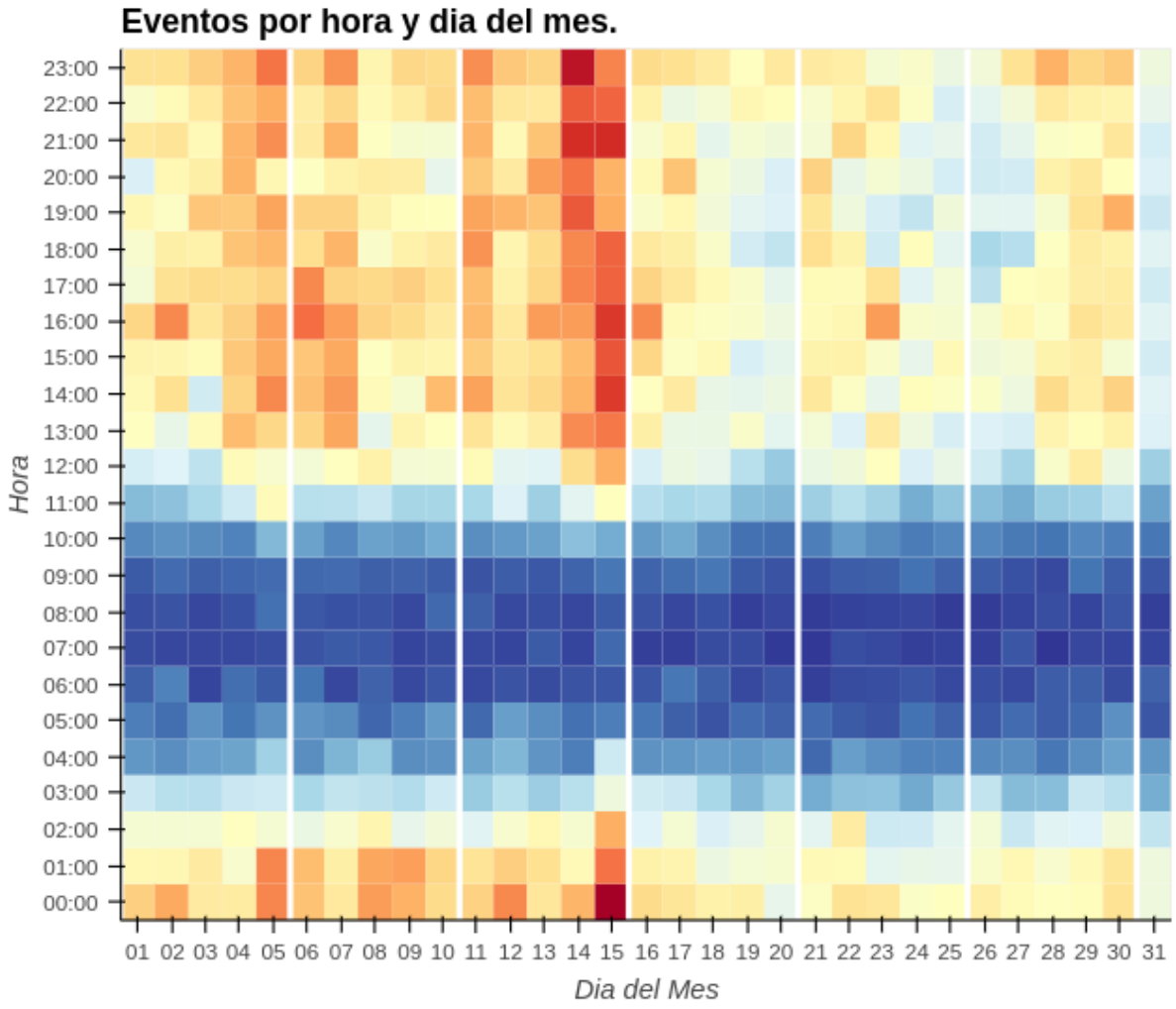
Set de Datos

● **timestamp**: Fecha y hora cuando ocurrió 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 (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**: Versión de sistema operativo desde donde se originó el evento
● **browser_version**: Versión del browser utilizado en el evento

Exploración del set de datos

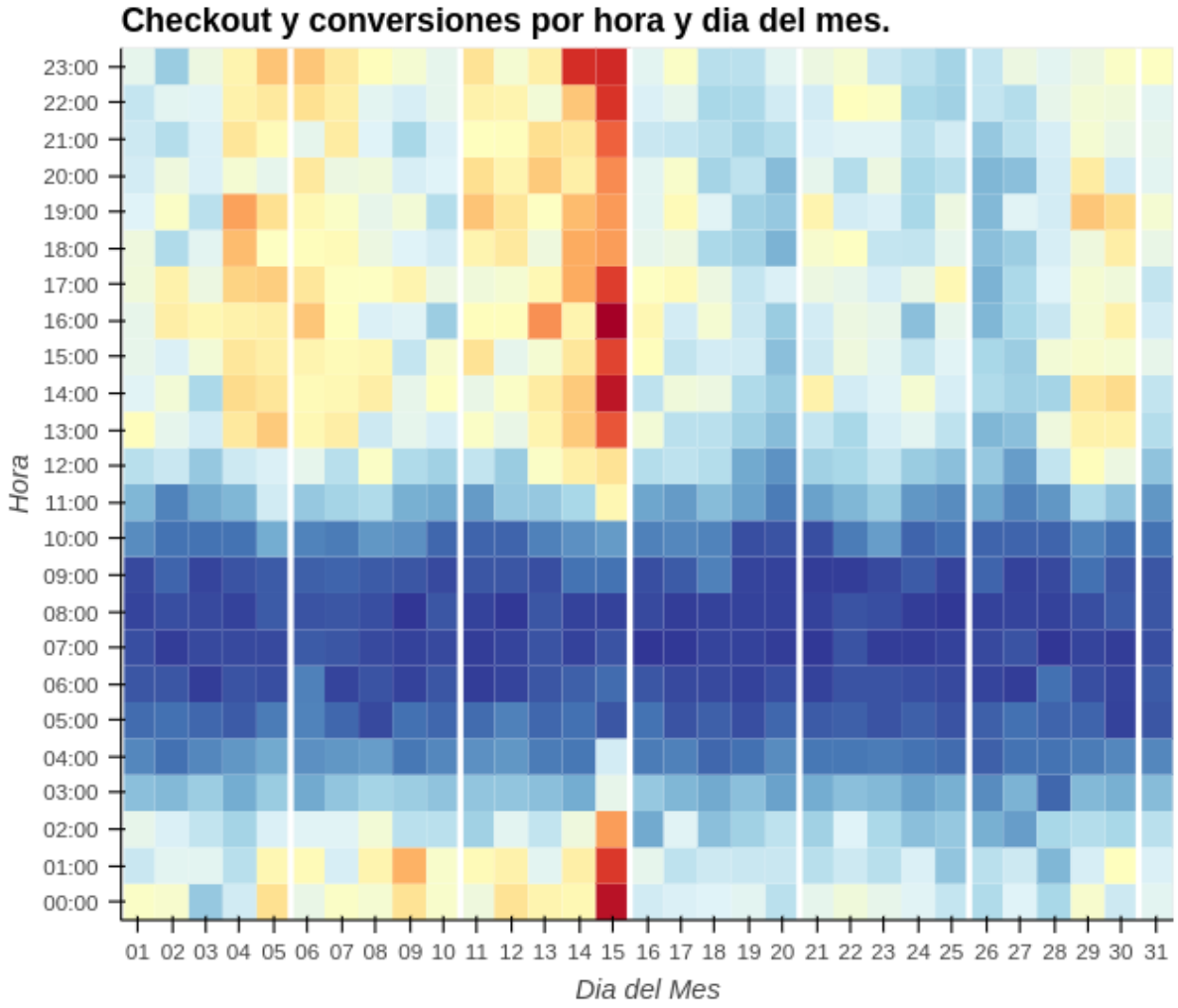
Fechas y cantidad de eventos

El set de datos empieza '2018-01-01 07:32:26' y termina el '2018-06-15 23:59:31' dándonos seis meses y medios de de información del comportamiento de los usuarios.
Hay relación entre el momento y las cantidades de eventos.



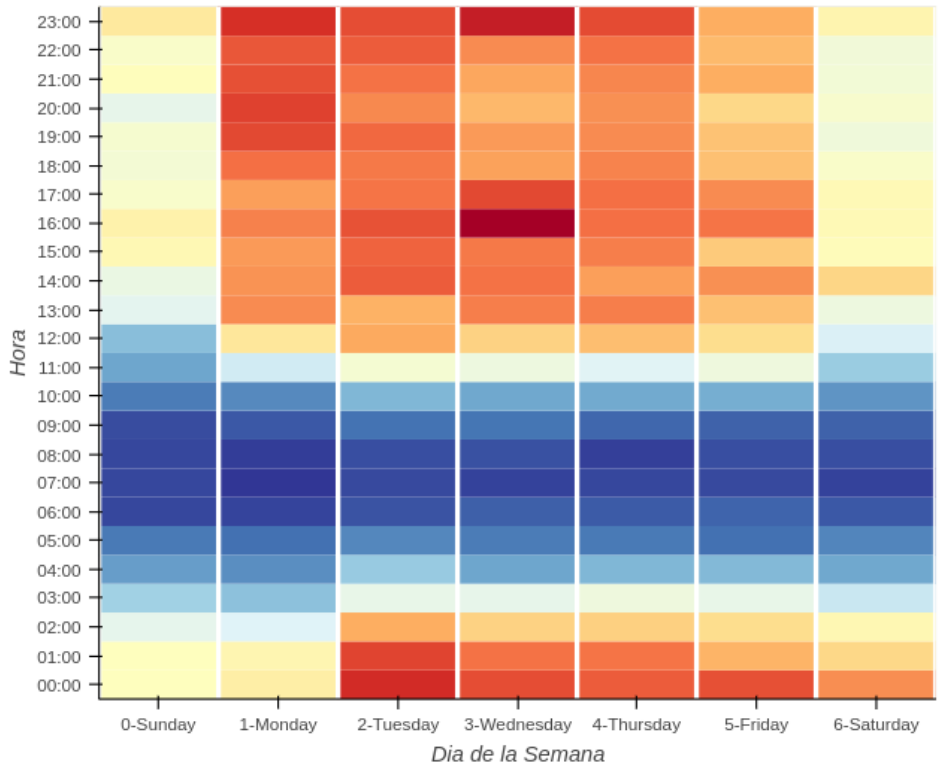
Como se puede ver, la cantidad de eventos es alta desde principios hasta mediados del mes, y se puede ver cómo encuentra su momento culmine los días 15, como la actividad del blog es alta durante la tarde-noche y es muy baja durante la madrugada-mañana.

De la misma manera, si filtramos los eventos quedándonos con los que culminan el ‘flujo’ de una compra como ‘checkout’ y ‘conversion’.

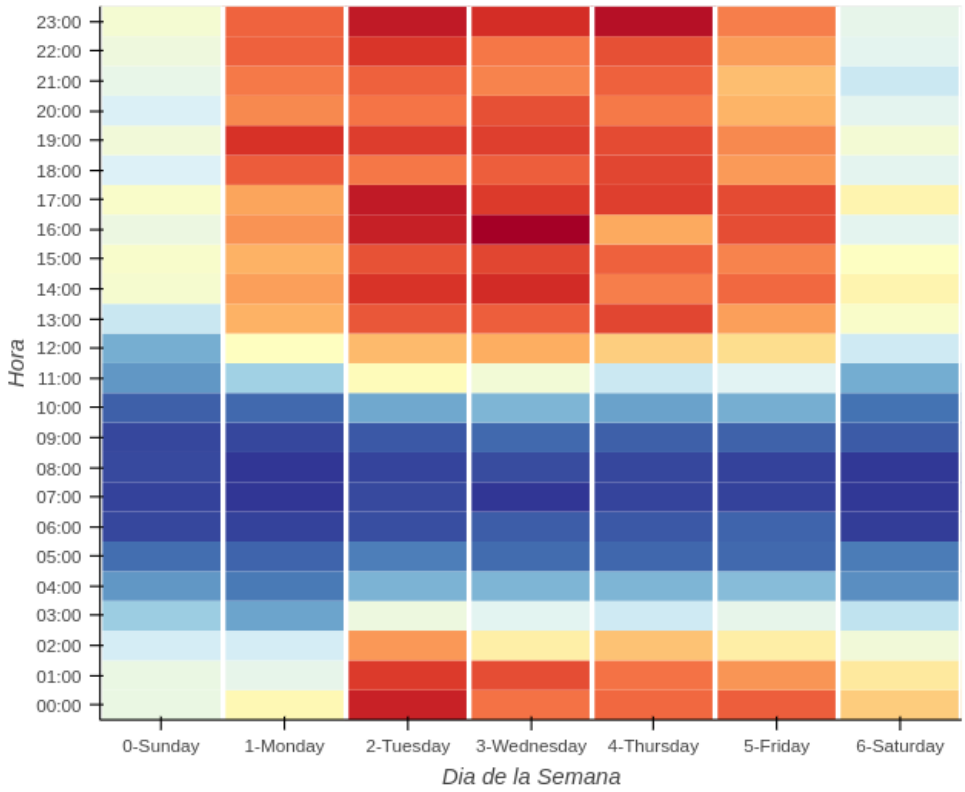


Se puede ver como la lógica es parecida hay una clara tendencia a realizar compras y hacer checkout los días 15. También se puede ver como la primera quincena tiene más eventos que la segunda, esto se puede atribuir a que las personas en ese momento perciben su salario.
Análogo análisis se hace sobre los días de la semana. Los sábados y domingos acumulan poca actividad y hay picos de ocurrencias de eventos los días martes a la madrugada y los días miércoles por la tarde

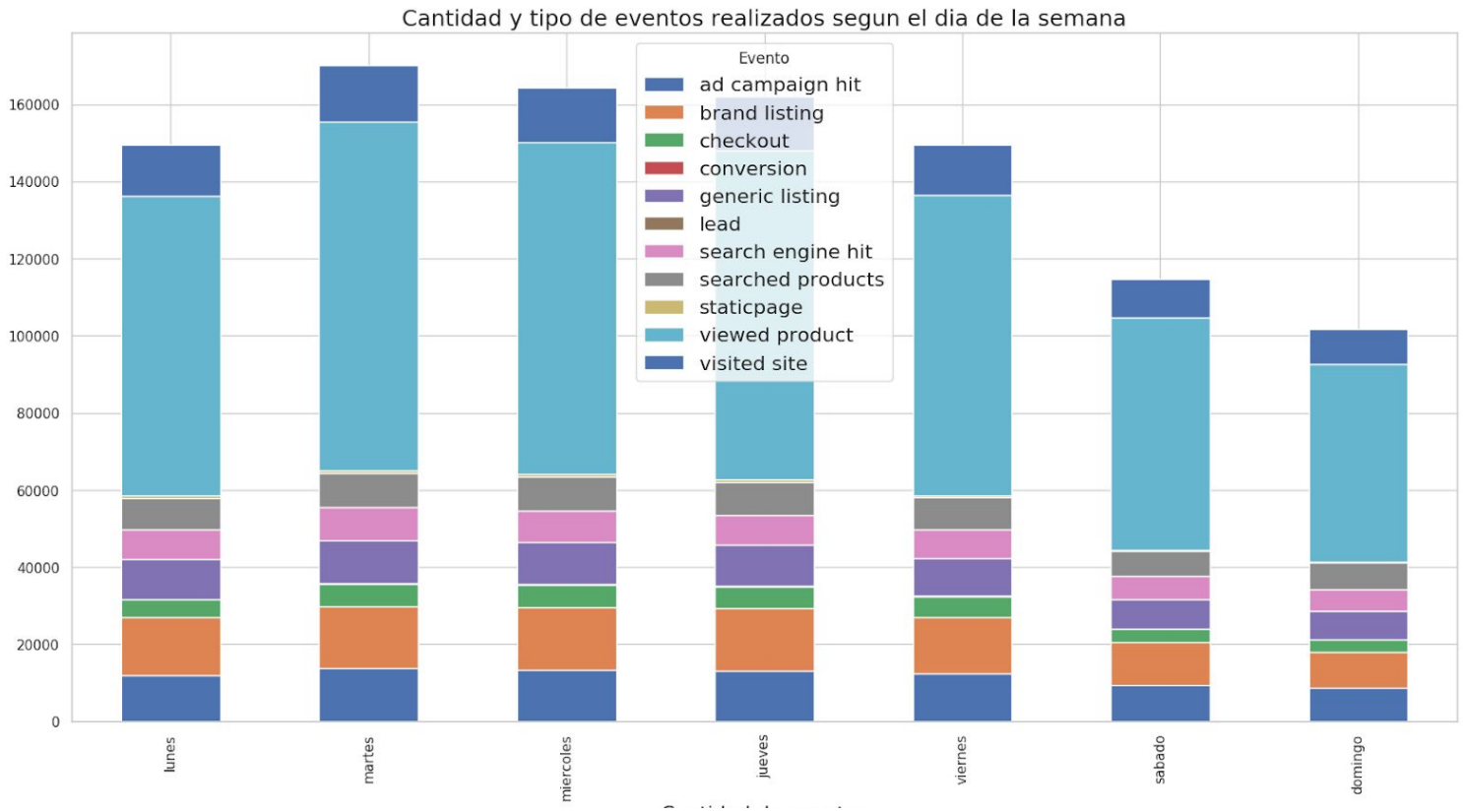
Eventos por dia de la semana.



Checkout y conversiones por dia de la semana.



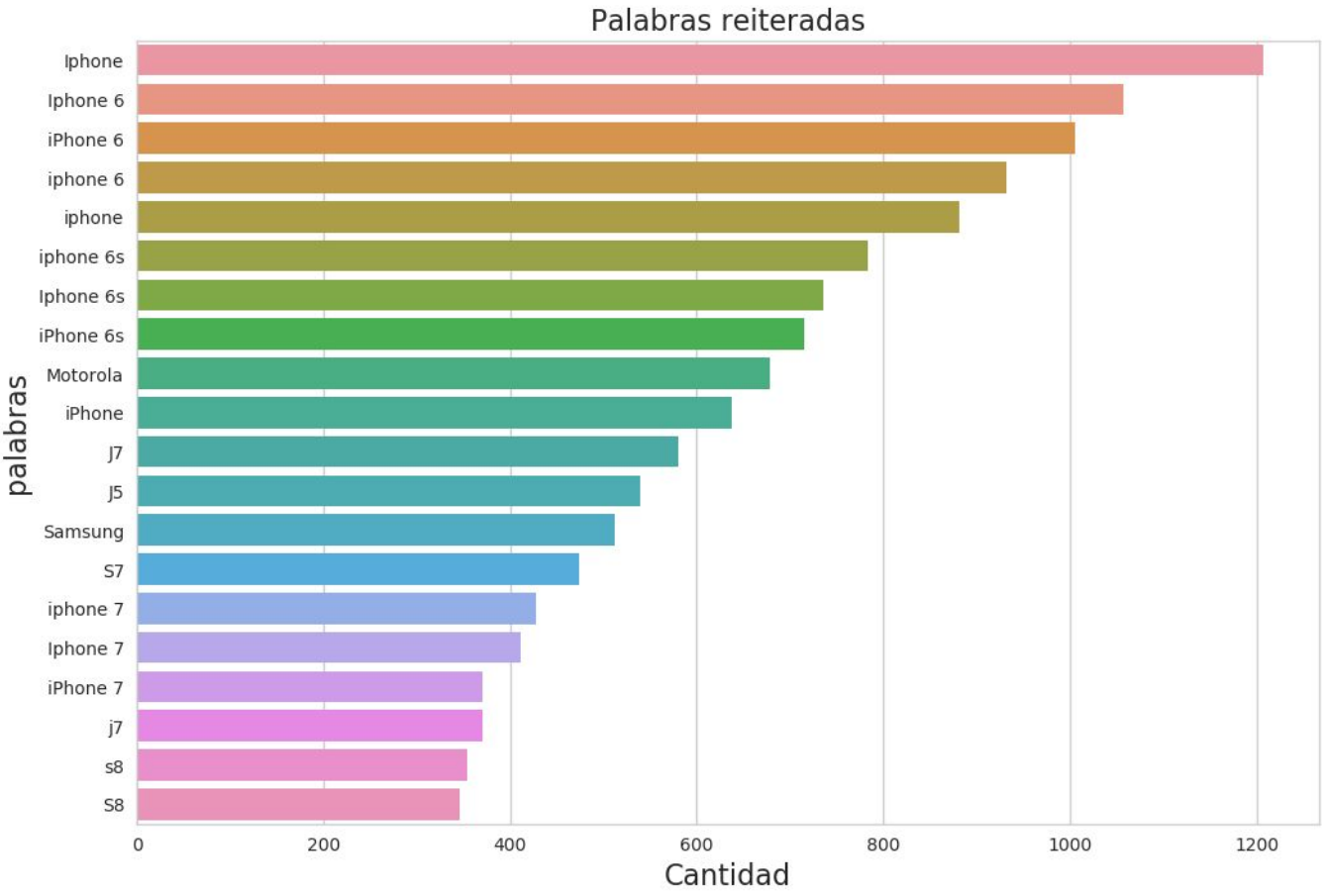
Eventos según el día de la semana



Vemos que los domingos hay menos tráfico y los martes son los dias que mas hay. En proporción siempre los usuarios realizan los mismos eventos, sin importar que dia de la semana sea.

Motor de búsqueda de la página

Trocafone cuenta con un buscador interno dentro de su blog, el cual puede utilizarse para acceder al catálogo de celulares a través de keywords. Las palabras más utilizadas son sobre los celulares Apple, principalmente el modelo IPHONE S6, también se pueden observar Motorola y Samsung. Lo que se aprecia es que las personas buscan lo mismo de diferentes formas, y también errando ortográficamente.



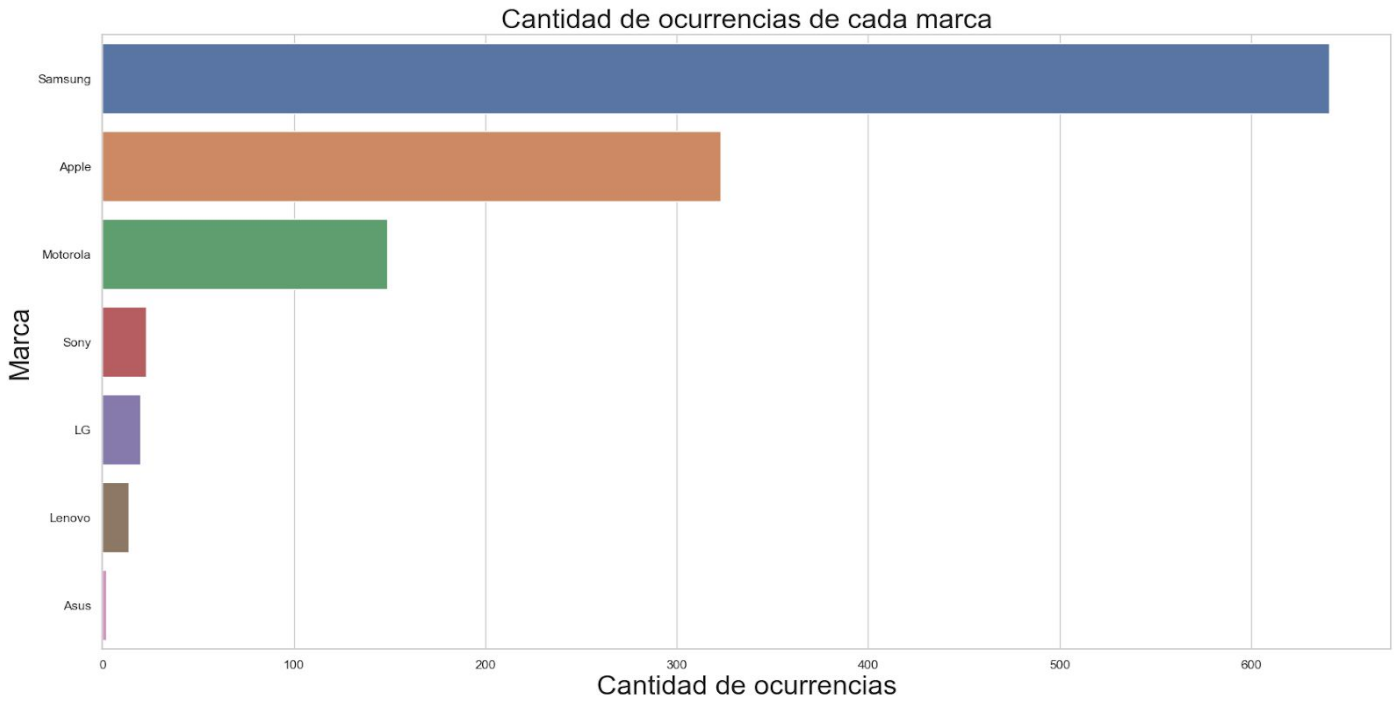
Limpiando los campos y agrupando palabras que apuntan a lo mismo se observa que en líneas generales Iphone, Samsung, Moto, Motorola, Plus, Prime, Celular, Galaxy , entre otras son las más buscadas.

Palabras mas buscadas.

Modelos de Celulares

Dentro de la columna models tenemos el modelo de celular relacionado con el evento que el usuario disparo con la interacción en el sitio. Si contamos ocurrencias de los modelos y agrupamos por marcas podemos, ver como el color negro y el color rojo ocupan el mayor lugar, asignados a modelos Iphones y a modelos Samsungs respectivamente; con menos eventos Motorola, Sony, LG, Lenovo.

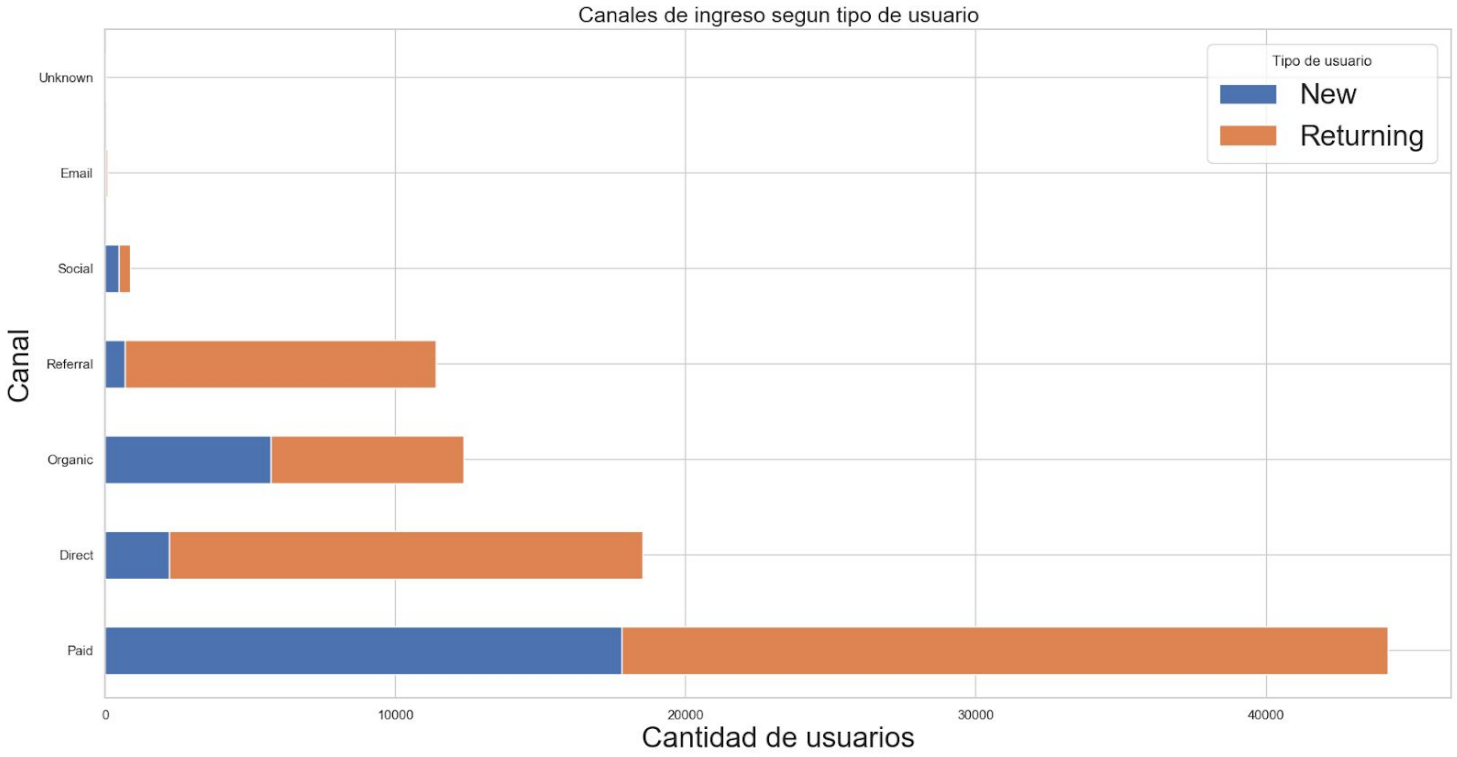
Inspeccionando la columna ‘models’ del set de datos pudimos extraer información acerca de cuáles eran las marcas con más interacciones del sitio. Agregamos una nueva columna ‘marca’ sabiendo que la primer palabra dentro del modelo tenia esa informacion, agrupamos búsquedas de iphone e ipad dentro de la marca Apple y obtuvimos el siguiente gráfico.



Como se puede observar el gran porcentaje de las interacciones en el sitio están relacionados a productos de la marca Apple o Samsung. Mientras que la cantidad de búsquedas sobre otros modelos sumadas no equiparan a las del segundo lugar.

Cómo ingresan los usuarios

Analizamos la información de la columna ‘channel’ en relación a la columna ‘new_vs_returning’ para obtener un vistazo de cómo influye el método de ingreso de los usuarios a Trocafone dependiendo de si los mismos eran nuevos o recurrentes.

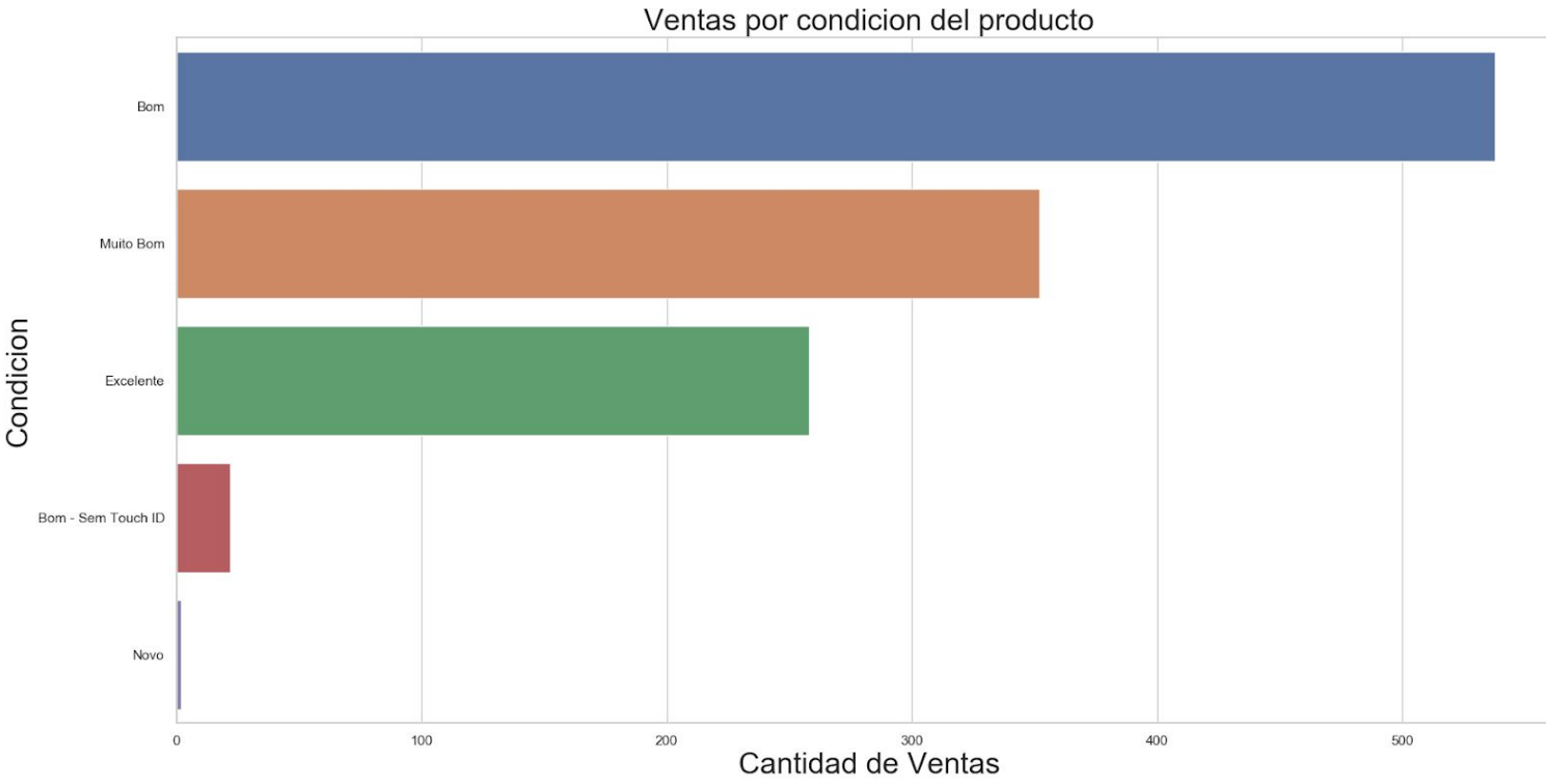


Viendo el gráfico podemos observar que:

1. El mayor ingreso de usuarios tanto nuevos como recurrentes proviene resultados del tipo publicitario en motores de búsqueda
2. El tráfico directo (es decir, cuando la visita no vino de ningún tipo de sitio previo. Como puede ser el caso de un usuario que escribe el link directamente en su navegador o usa un bookmark) es bajo en usuarios nuevos pero significante en usuarios recurrentes.‘Referral’
3. El tráfico de tipo ‘Referral’ (cuando un usuario encuentra el sitio a través de un otra web que no es un motor de búsqueda o una red social) presenta el mismo comportamiento que el punto anterior.
4. Muy pocos usuarios acceden a Trocafone a través de links compartidos en redes sociales, probablemente por una falta de incentivo en estos canales.

Qué condiciones de producto prefieren los usuarios

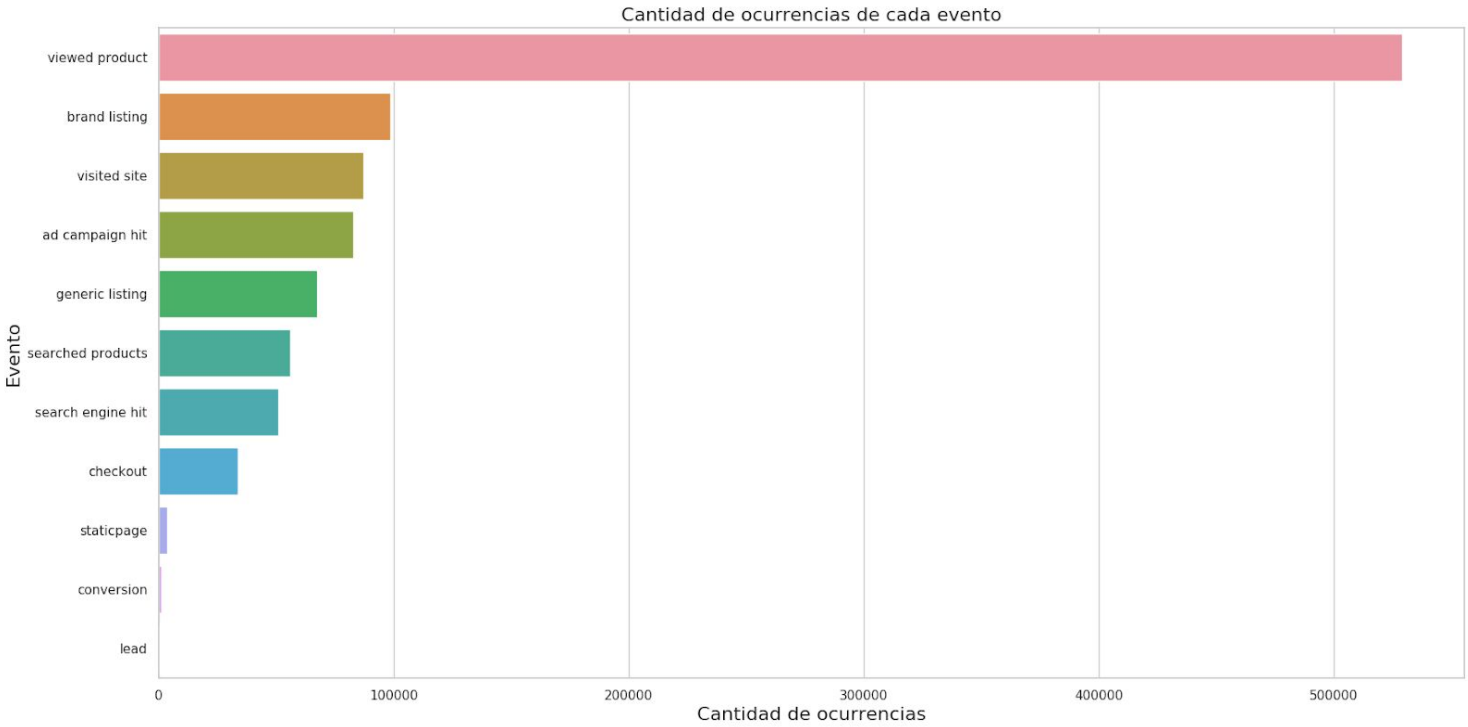
Siendo Trocafone un servicio que se caracteriza principalmente por el reacondicionamiento y reventa de celulares usados nos pareció interesante preguntarnos en qué condiciones estaban los celulares que los usuarios compraban.



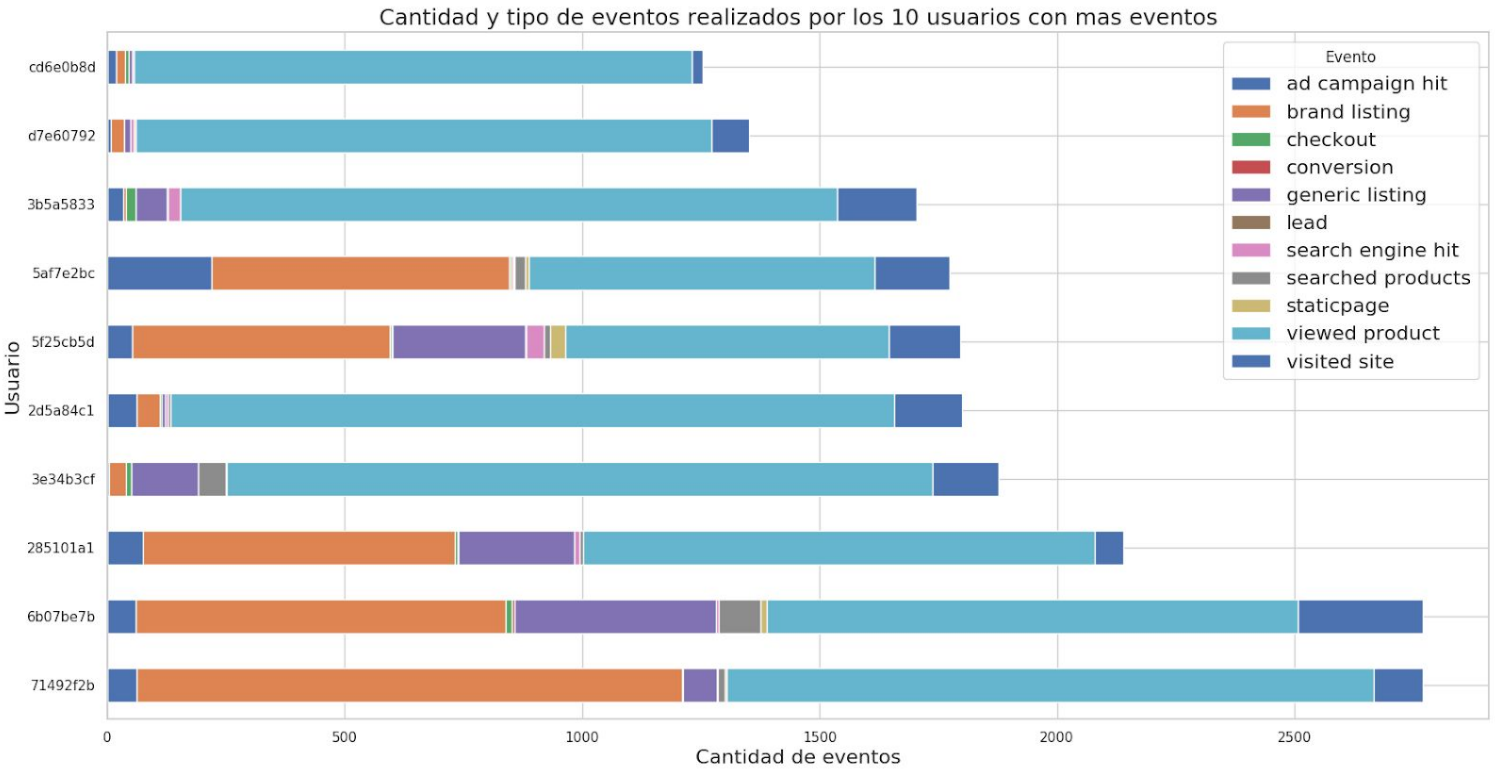
Filtramos el set de datos para obtener los eventos de tipo ‘conversion’ e hicimos un simple barplot para ver cuales eran los tipos más recurrentes.

La compra de celulares nuevos en el set de datos fue casi nula (con solo 2 eventos) y apreciamos que cuanto peor es la condición del producto, más ventas se realizan (muy probablemente ligado al hecho de que el precio disminuye en cuanto la misma empeora).

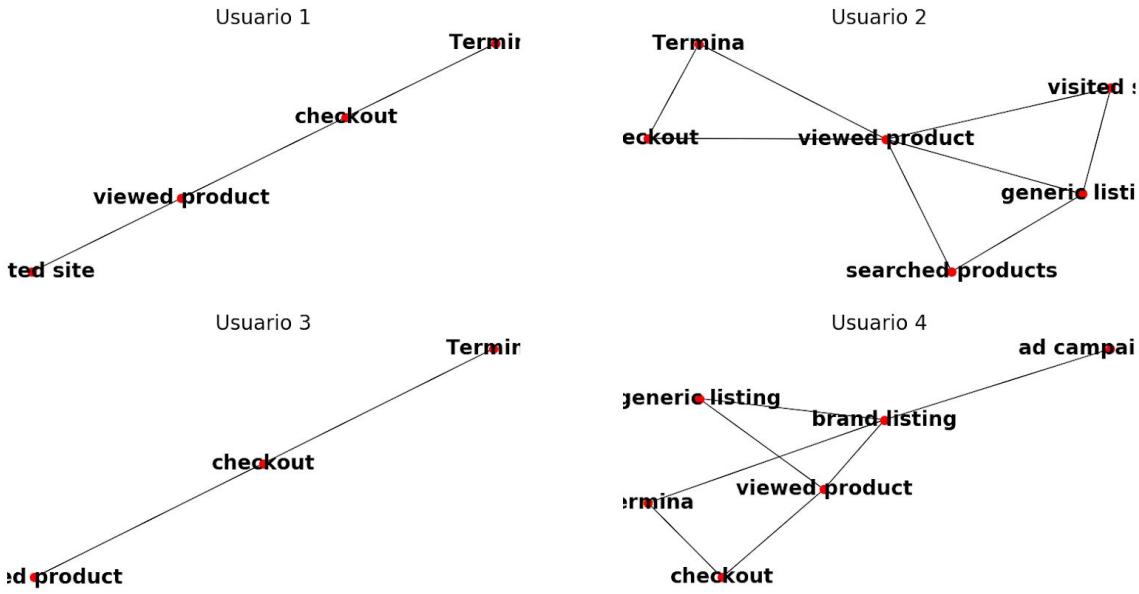
Eventos



El evento que más ocurre es viewed product, seguido por brand listing con ⅕ de sus ocurrencias. El que menos ocurre es lead, con solo 448 ocurrencias. Dentro del flujo de acciones que los usuarios realizan, la vista de productos se tiende a repetir muchas veces, hay usuarios antes de realizar un checkout tiende a repetir la vista de un producto varias veces.
El comportamiento de los usuarios en la plataforma se puede apreciar mediante el estudio de los eventos que desencadenan.



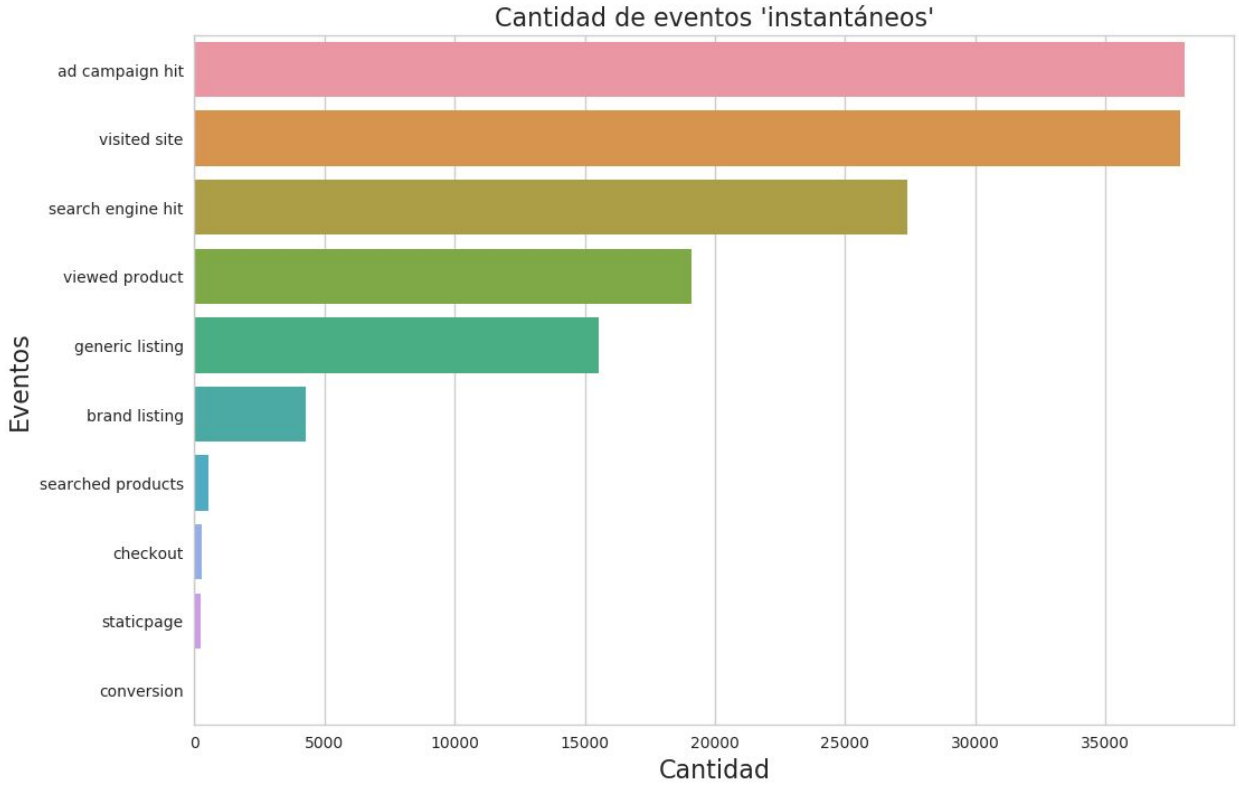
Los 10 usuarios que más eventos realizaron tuvieron entre 1254 y 2771 eventos. Podemos ver que el evento predominante de todos ellos es la vista de un producto, seguida por visited site. La proporción del resto de los eventos varía según el usuario.



Hay una tendencia a producirse un flujo, donde los usuarios ingresan al sitio, navegan, buscan, siguen navegando y terminan por realizar un checkout, una compra o anotarse en una notificación de stock.

Eventos en el mismo instante

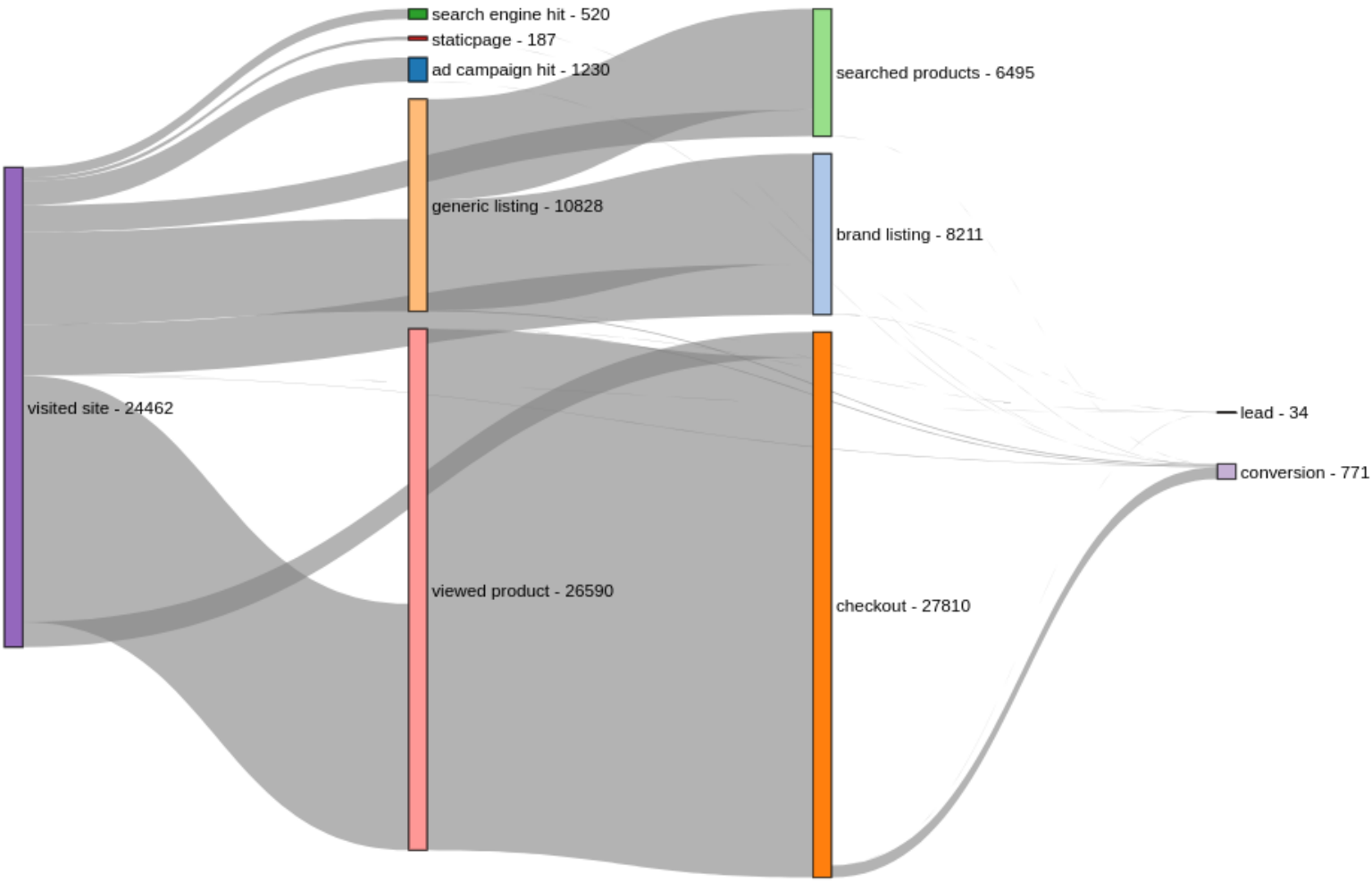
Hay acciones de los usuarios que desencadenan varios eventos al mismo tiempo, por ejemplo, hay ingresos a la página que se producen por publicidad y al mismo momento se registra otro evento de visita de web.



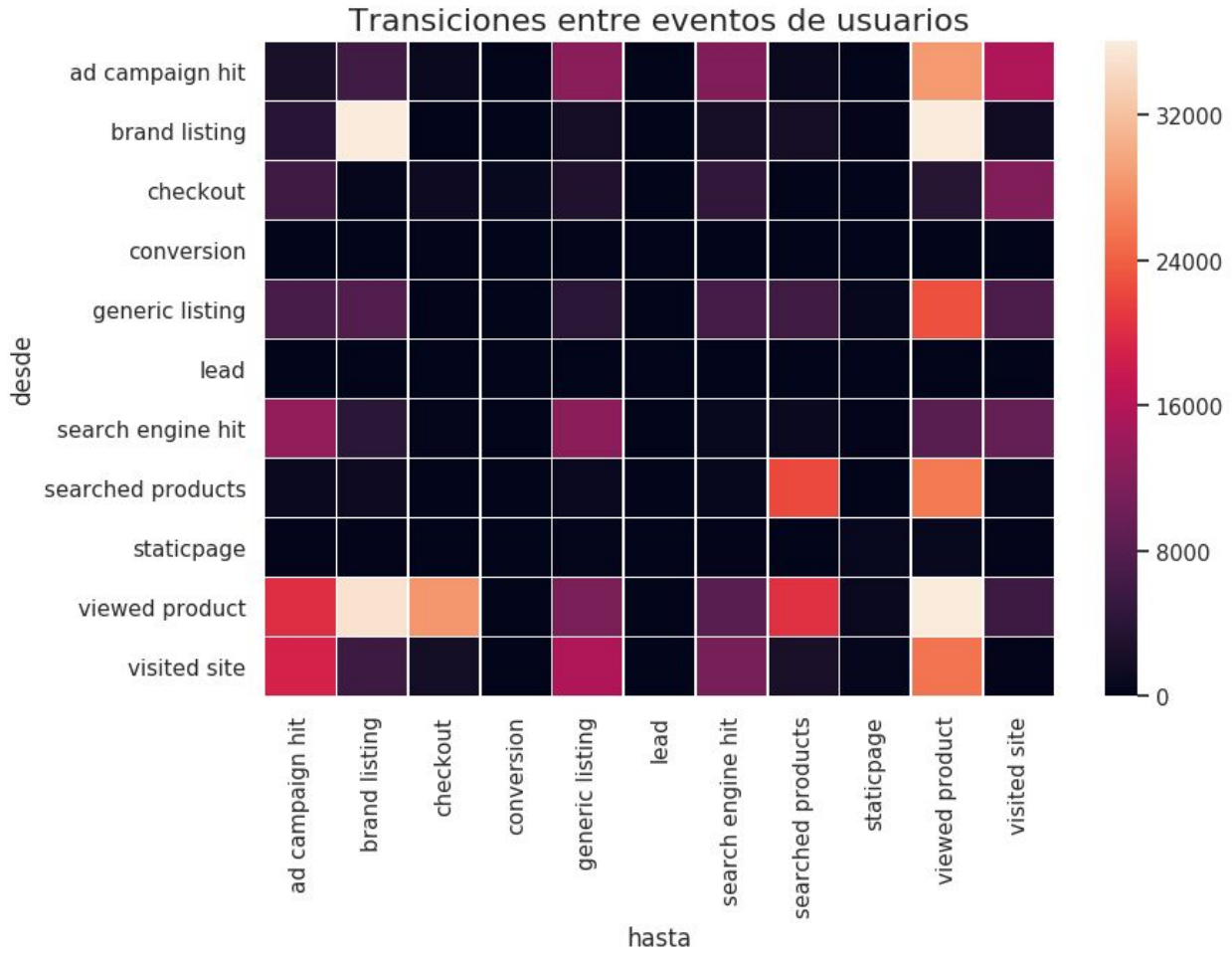
La registración de estos eventos no se produce siempre en el mismo orden, hay veces que aparecen primero unos y a veces otros.

Flujo hacia una compra

Para determinar el flujo de las compras, se arma un grafo con la secuencia de eventos de usuarios, se cuentan las aristas entre los nodos y se restan las que entran y las que salen. Se arma un peso resultante.
Hay eventos que por lo visto en el fenómeno anterior, son generadores de ciclos, pero también hay acciones legítimas de los usuarios que generan una vuelta atrás; (por ejemplo: un usuario puede estar mirando un catálogo de productos y volver a la homepage). Las relaciones de nodos balanceadas entre entradas y salidas se filtran, dejando las tendencias manifiestas hacia un flujo que va desde la entrada al sitio y finaliza en un checkout, en una conversion o en un lead. Las búsquedas de productos y los listados, suelen ser medios de retroalimentación, al filtrarse los ciclos, simplemente terminan.



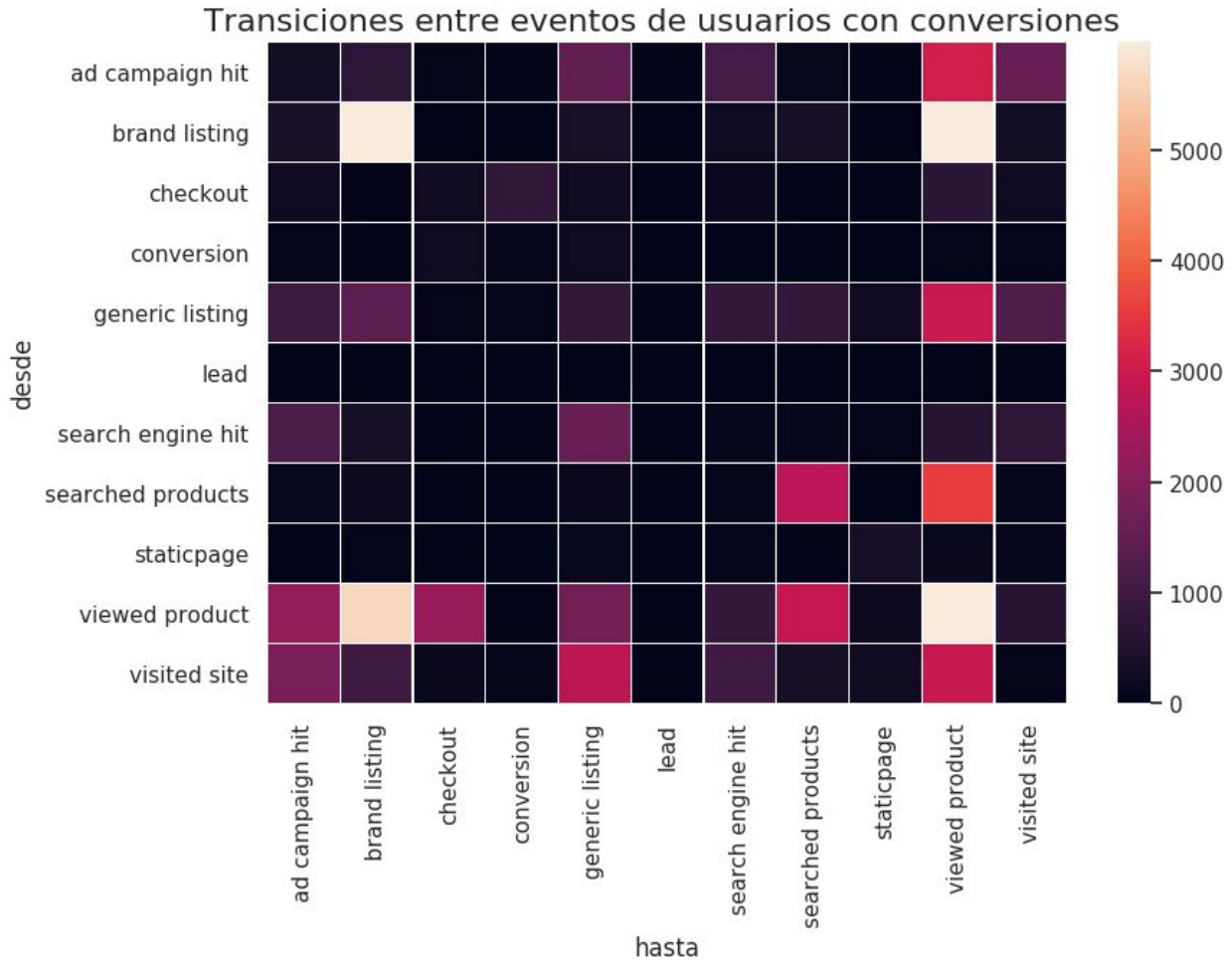
¿Cual es el flujo de los usuarios en el sitio?



Podemos ver que los usuarios terminan entrando a ver un producto mayormente desde el listado de marcas, seguido por un hit de una publicidad. También se puede apreciar cómo se generan ciclos entre las vistas de producto.

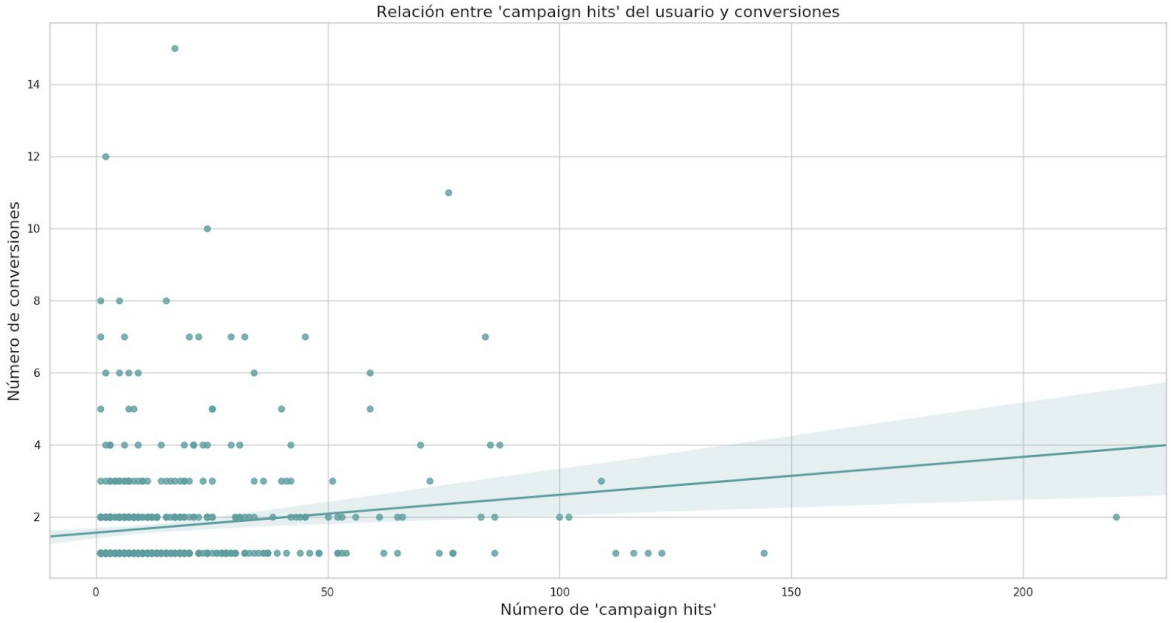
Recorren mucho los listados de marcas y pasan mucho de un producto a otro.

¿Cambia el flujo con los usuarios que realizaron compras?

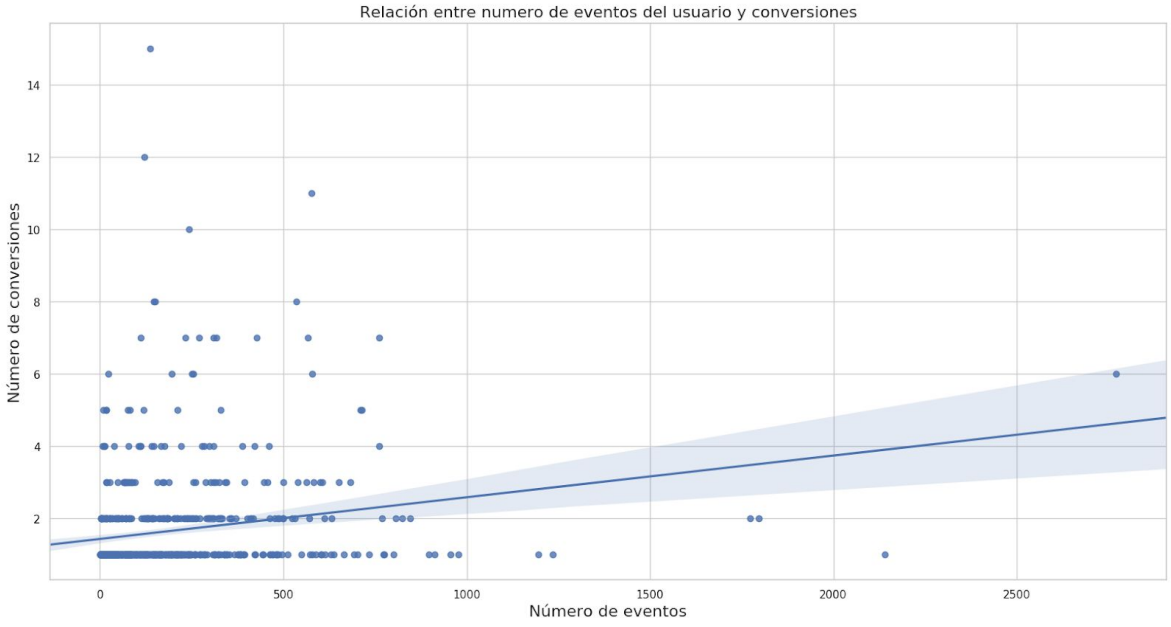


Con los usuarios que tienen compras, podemos ver que los que llegan a ver un producto vienen desde el listado de marcas, seguido por los resultados de búsqueda. Dejando (a diferencia del resto de los usuarios) en tercer lugar a los hits de publicidad, lo que nos hace ver que no siguen el mismo flujo que los usuarios en general.

Conversiones vs Cantidad de Eventos

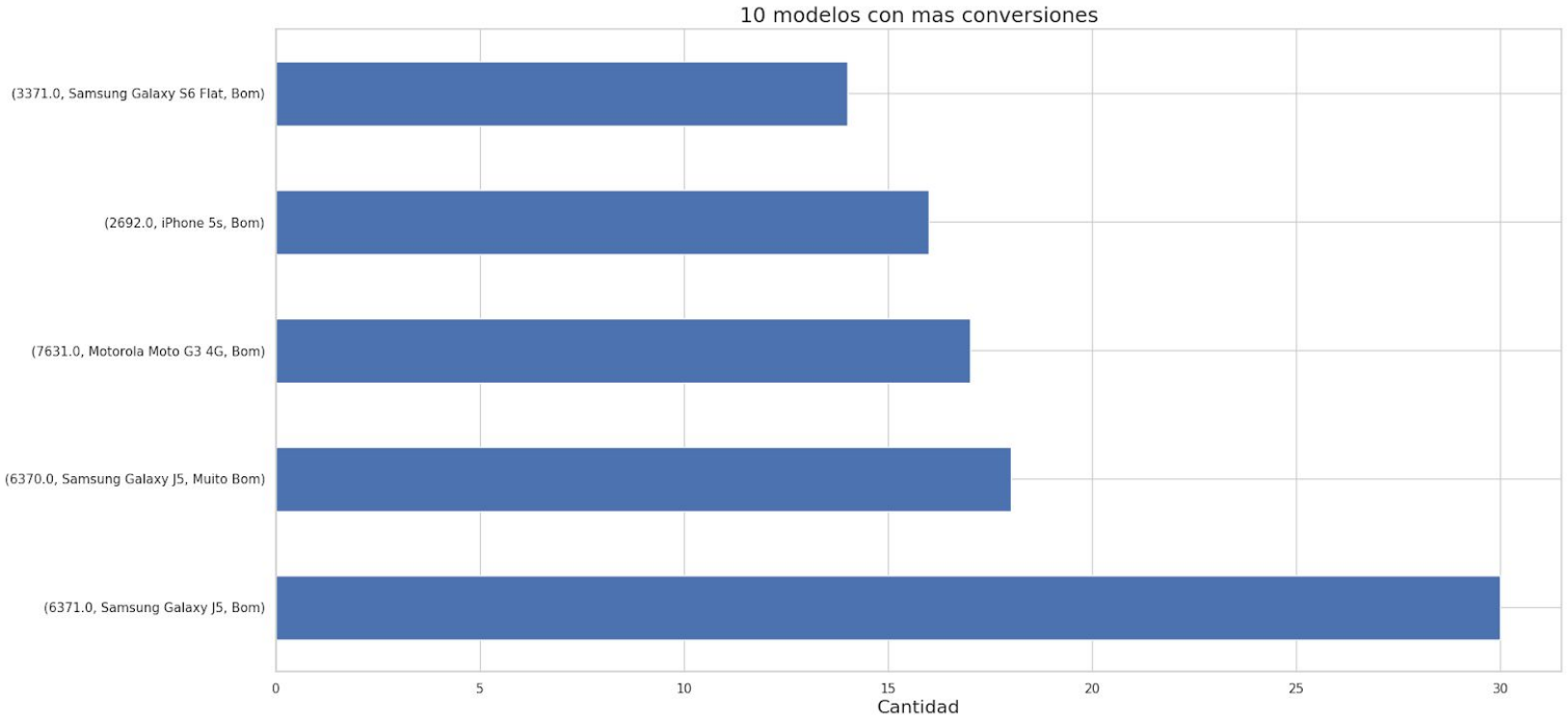


Podemos ver que la relación entre ambas variables es muy debil.



Al igual que en el caso anterior, vemos que ambas variables no están relacionadas. Los usuarios con mas conversiones no necesariamente tienen muchos más eventos que los otros usuarios.

Conversiones por modelo de celulares



Como se puede ver en el gráfico, los modelos con más ventas son: Samsung Galaxy J5, Motorola G4, iPhone 5S y Samsung Galaxy S6

Sesiones

Una sesión es un periodo de tiempo en el que el usuario interactúa con el sitio sin tener más de 30 minutos de inactividad. Luego de los 30 minutos sin realizar ningún evento, consideramos que es una nueva sesión.

¿Cuántas sesiones tuvieron los usuarios?

El promedio de sesiones que tuvieron los usuarios fueron 8.25

El usuario con más sesiones tuvo 262

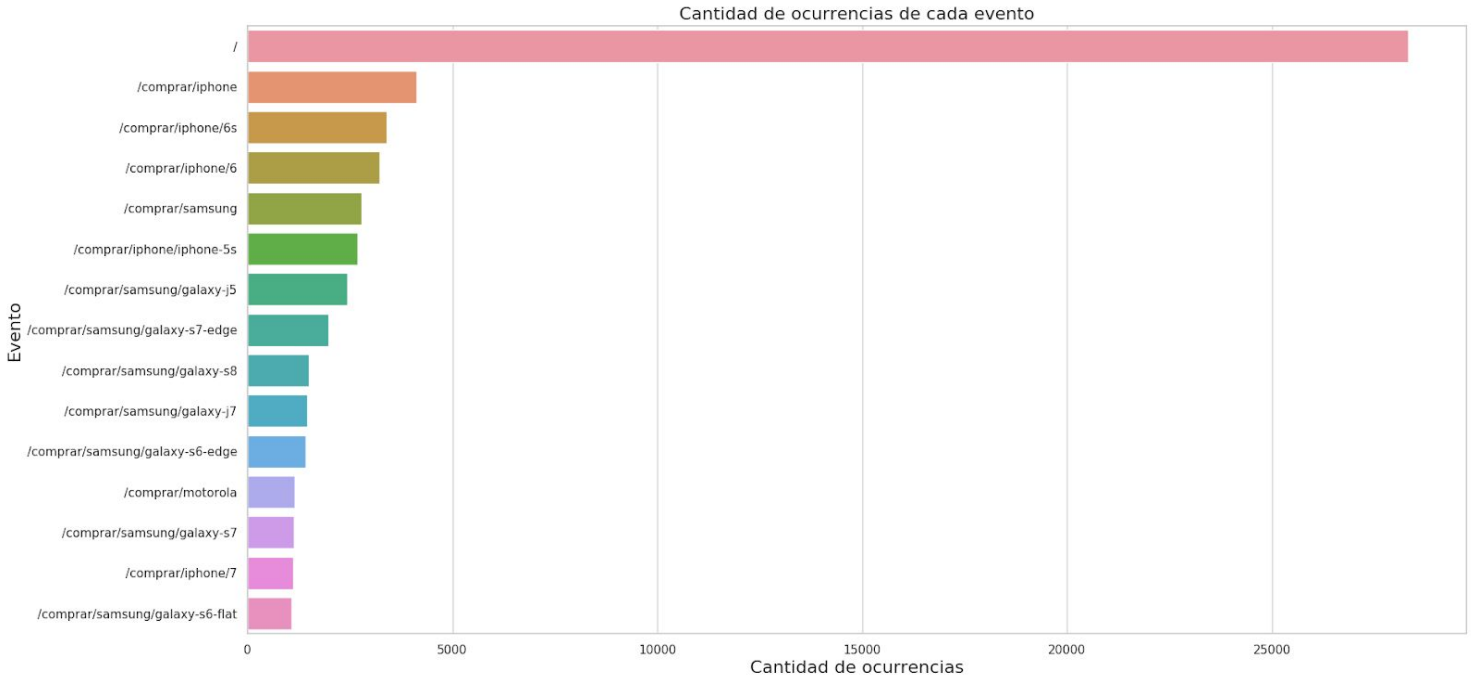
¿Cuánto duró cada sesión?

El promedio de duración de las sesiones fue 7 minutos y 17 segundos.
La sesión que más duró fue de 4 horas y 29 minutos.

¿Cuánto tiempo pasa desde la primer visita registrada de los usuarios al sitio hasta que compran?

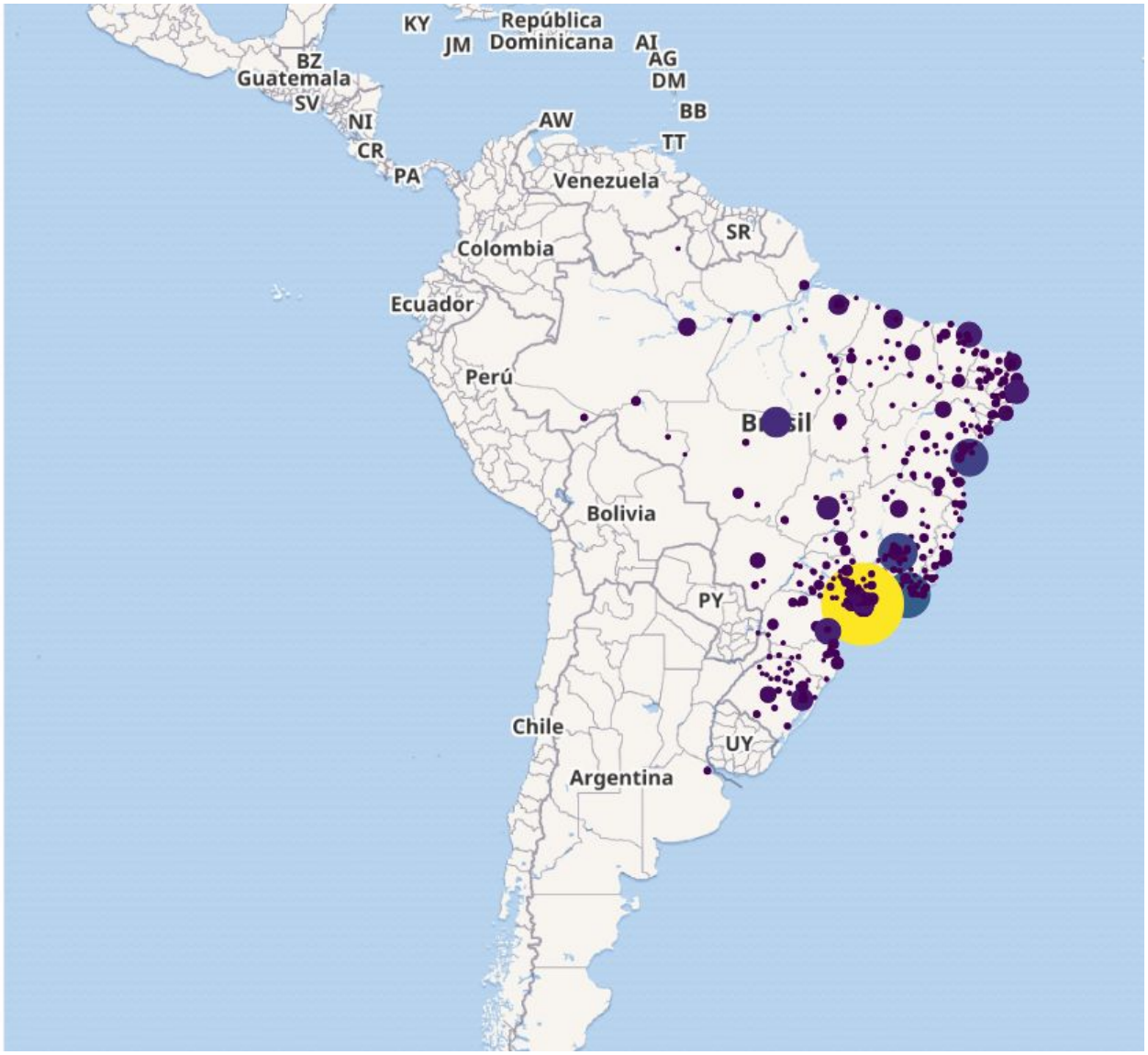
En promedio pasan 10 dias 17 horas. El máximo fueron 134 dias.
Hay que tener en cuenta que no tenemos datos de antes de Enero de 2018, por lo que puede ser que algunos usuarios hayan entrado antes que esa fecha.

¿Cuáles son las urls con más hits de publicidad?



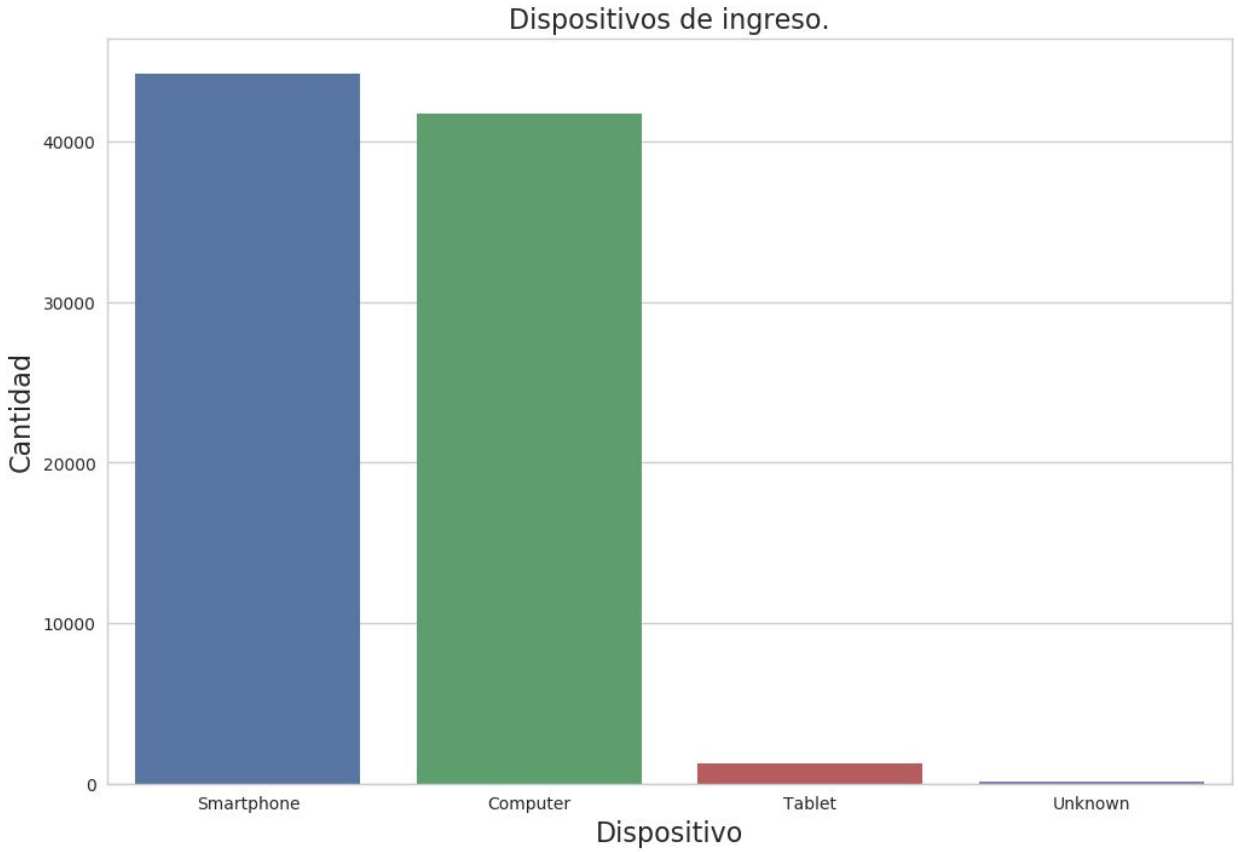
Vemos que la home es la url con mayor cantidad de hits de publicidad, seguida por urls de productos.

Localización



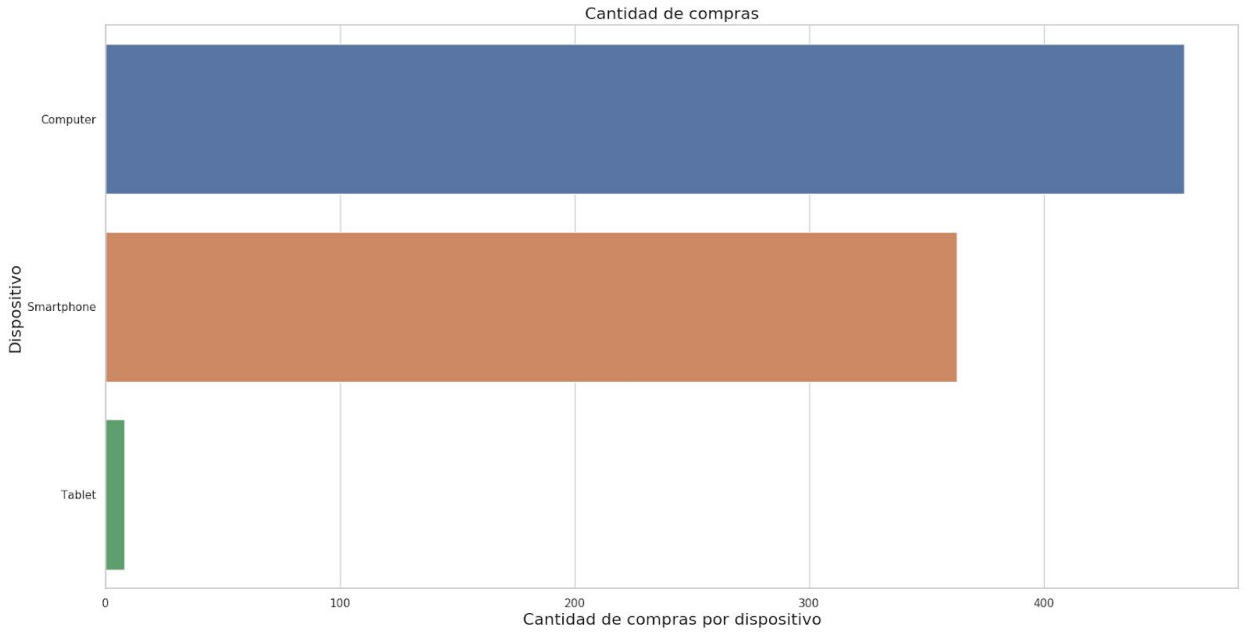
Como se puede ver, la mayoría de los eventos se producen en Brasil. En la Argentina hay localizados en Buenos Aires, pero muy poca cantidad. En EEUU hay unos pocos, en NY y también en la costa oeste

Mobile vs Desktop



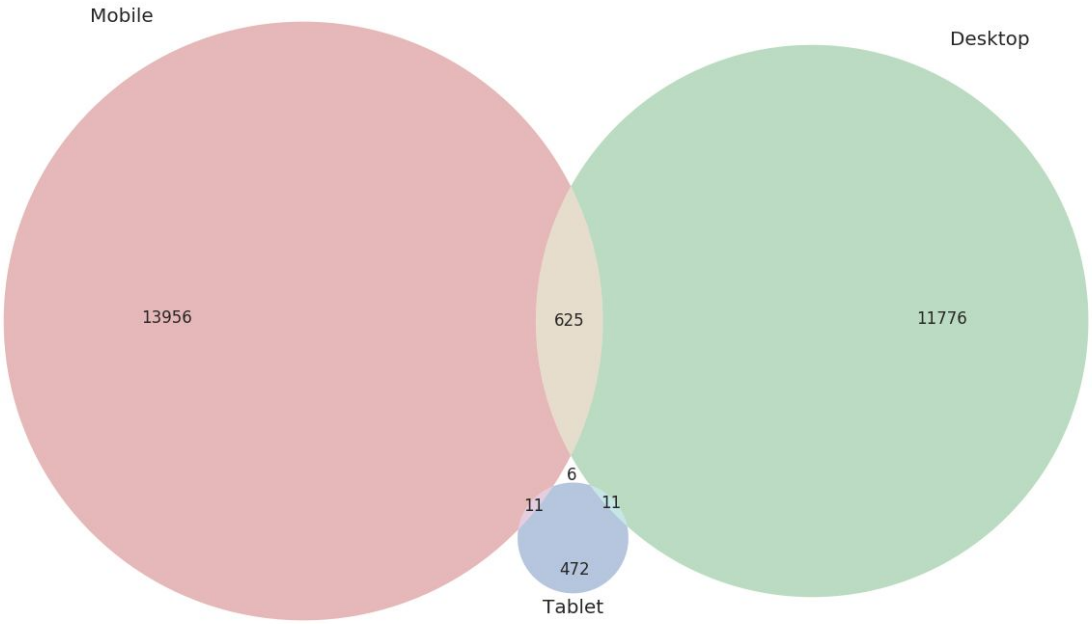
Se puede ver como los ingresos desde el celular son mayores a los ingresos desde la computadora. En la resolución de pantalla también se puede apreciar esto, los mayores ingresos son en 360x640 que es una resolución Mobile.

Hay más visitas en mobile que desktop, pero hay también más conversiones?

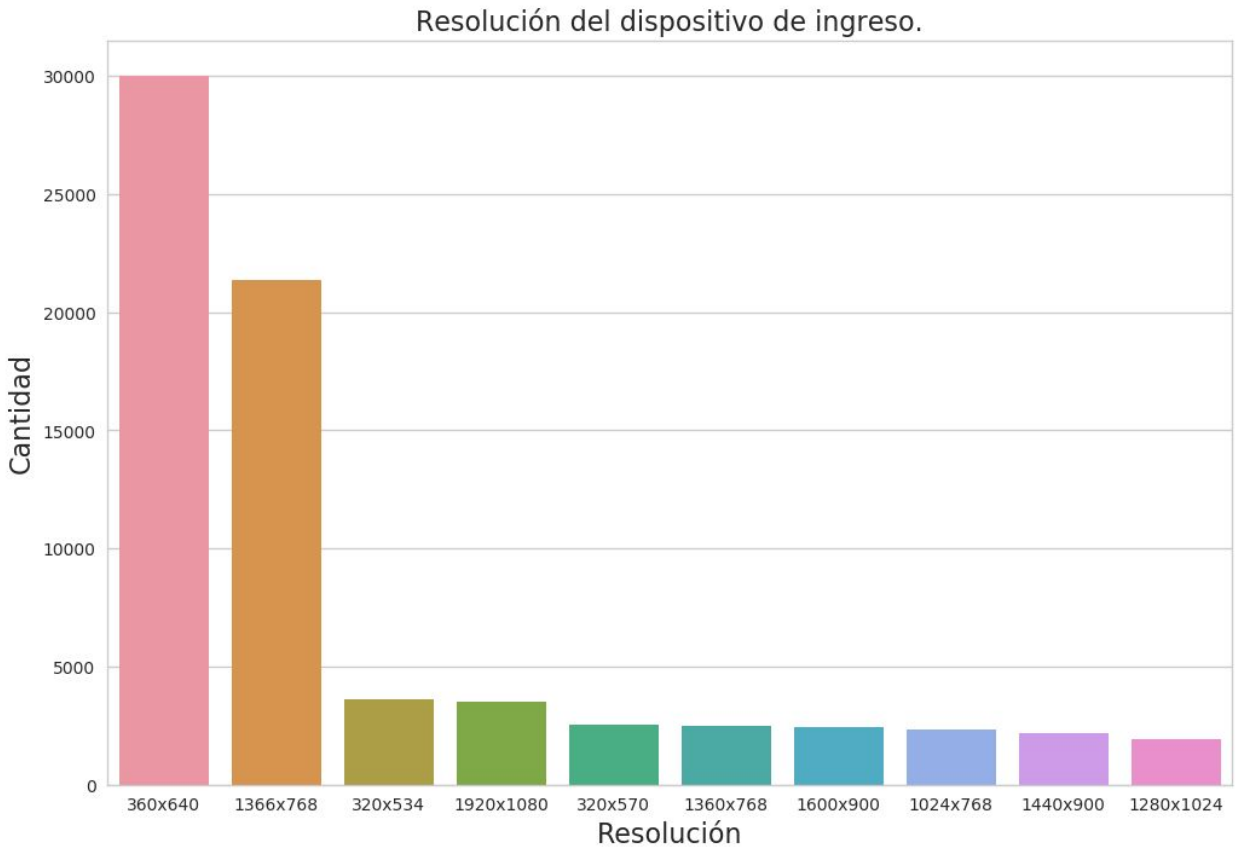


Si bien hay más tráfico mobile, vemos que los usuarios compran más desde Desktop.

¿Los usuarios entran siempre desde el mismo dispositivo?



La gran mayoría sí, pero hay 11 usuarios que entraron desde mobile, desktop y tablet y 625 que entraron desde mobile y desktop.


Resolución del dispositivo de ingreso.
30000
25000
20000
15000
10000
5000
0
Cantidad
360x640
1366x768
320x534
1920x1080
320x570
1360x768
1600x900
1024x768
1440x900
1280x1024
Resolución

Conclusiones:

1) Fechas: del análisis se desprende que las primeras quincenas, el sitio tiene mayor actividad encontrando su pico en la culminación de la primera (el día quince).
2) Los días de la semana la pagina tiene mas actividad que los fines de semana. Con picos los días miércoles por la tarde y martes a la madrugada.
3) Las marcas Apple y Samsung predominan en las vistas de productos y en las búsquedas en el motor.
4) Con la información de los modelos de celulares, se puede obtener datos de las empresas en general.
5) El mayor ingreso de usuarios tanto nuevos como recurrentes proviene resultados del tipo publicitario en motores de búsqueda, y los usuarios recurrentes tienden a incrementar los ingresos de manera directa.
6) La mayor actividad del sitio proviene de usuarios de Brasil.
7) Las acciones de los usuarios pueden desencadenar más de un evento, registrándose los mismos no siempre en el mismo orden.
8) Los usuarios tienden a navegar y a generar ciclos entre que buscan y ven productos.
9) Hay un flujo de eventos que empieza en ingresos a la página y puede termina en un checkout o una compra.
10) La mayoría de ingresos son por dispositivos Mobile
11) Una vez que están dentro del sitio, los usuarios prefieren utilizar el listado de marcas antes que el buscador.

Bibliografia

http://pandas.pydata.org/
https://docs.python.org/2/library/re.html
https://bokeh.pydata.org/en/latest/
https://github.com/amueller/word_cloud
https://amueller.github.io/word_cloud/generated/wordcloud.WordCloud.htm
https://geopy.readthedocs.io/en/stable/
https://jakevdp.github.io/PythonDataScienceHandbook/
https://matplotlib.org/users/index.html

http://holoviews.org/

http://geo.holoviews.org/Working_with_Bokeh.html