Universidad de Buenos Aires

Facultad de Ingeniería

Organización de datos

Segundo Cuatrimestre de 2018

Trabajo práctico 1

Integrantes - Grupo 46

(Grupo de oyentes)

Álvarez González, José Mariano ~ jose.mariano.alvarez@gmail.com

Bernabó, Guillermo ~ guillermo.bernabo@gmail.com

Bertin, Natacha ~ natachabertin92@gmail.com

Huerta San Martín, Marcelo Gustavo ~ marcelo.huerta@gmail.com

Índice

1. Acerca de los datos en crudo

1.1. Tecnologías utilizadas en el análisis

Para el manejo de los datos se utilizó Python 3, con la librería Pandas. Las visualizaciones se realizaron con Matplotlib en la etapa de exploración, y fue agregado Seaborn al momento de este informe.

1.2. Repositorio en GitLab

https://gitlab.com/josemalv/7506-tp1
Solicitar acceso ya que por motivos de confidencialidad no es público

1.3. Datos recibidos y selección de campos a utilizar

De los 1011288 registros recibidos (eventos registrados por los usuarios que han realizado un checkout en el sitio), hemos tomado de todos los registros, los siguientes datos:

- 'timestamp': momento del evento
- 'event': tipo de evento
- 'person': usuario que realizó el evento
- 'url', 'staticpage': detalles de la página que visitó
- 'sku', 'storage', 'color': detalles del producto
- 'condition': estado del producto
- 'search_term', 'skus': datos relacionados al producto buscado
- 'channel', 'new_vs_returning': datos principales de la proveniencia del usuario
- 'campaign_source', 'search_engine': datos complementarios de la proveniencia del usuario
- 'city', 'region', 'country': datos geográficos del usuario
- 'device_type', 'screen_resolution', 'operating_system_version', 'browser_version': tecnología utilizada por el usuario.

Se ha analizado la información recibida y se ha determinado que los datos corresponden a eventos producidos en el sitio desde enero y hasta mediados de mayo de 2018.

1.4. Limpieza y reformateo del conjunto de datos

Se aseguró la inexistencia de duplicados, controlando por timestamp.

Para algunos casos (país, por ejemplo) hubo que limpiar los casos que aparecían como *unknown*, y reemplazarlos por null (*np.nan*).

Se decidió incluir los eventos con diferencia de pocos segundos como diferentes eventos; consideramos esta duplicación (falla de idempotencia en el sitio, error en carga y demás) como eventos que incidirán en el comportamiento posterior del usuario en el funnel.

Para el procesamiento de los datos se decidió separar los eventos de diferente tipo en distintos dataframes. Cada uno de estos tiene entonces, solo los eventos de un tipo, lo que permite reducir la cantidad de columnas de estos dataframe, quedando mucho más simples y entendibles.

1.5. Agregado de información

Se completó el dataframe con los siguientes campos, extraída de campos ya existentes en el dataframe, pero que de este modo permiten un análisis agrupado por otras variables.

- El timestamp que indica el momento de la transacción se analizó en cada una de sus partes, dando lugar a análisis semanal, diario y por transacción.
    - 'dia', 'dia_anio', 'mes', 'dia_semana', 'dia_nombre', 'hora', 'semana'
- Del modelo de dispositivo se tomó la marca por separado, para analizar la potencial influencia, dado que las marcas hacen publicidad propia fuera de los canales del sitio.
    - 'marca', 'modelo'
- También se tomó país, ciudad y región de procedencia de los eventos que los tuvieran, para reasignarlos en los faltantes con un *join* por persona.

2. Ingresos al sitio y publicidad

2.1. Ingresos al sitio

2.1.1 Por canal de acceso

El primer análisis que hemos realizado corresponde a la forma en que los usuarios ingresan al sitio. Ya sea que estos usuarios ingresan por primera vez y se los considera en ese caso nuevos usuarios o retornan luego de haber realizado una visita previa, cada ingreso al sitio se lo puede relacionar con un canal.

En la figura 1 puede verse como existen algunos canales que son mucho más predominantes que el resto. En ese sentido se puede concluir que el acceso al sitio por el canal “Paid” para nuevos usuarios es el más importante. Queda claro de esta imagen que las campañas de marketing están surgiendo efectos en cuanto a la adquisición de usuarios en el sitio.

Figura 1. *Heat map* de Método de ingreso al sitio versus si es un nuevo usuario (*new*) o que retornan (*returning*). Se observan canales más recurrentes que otros. La escala de colores indica el número de eventos. Dentro de cada celda se detalla el número de eventos.

En menor medida se observa que el ingreso al sitio también se está realizando mediante los buscadores de internet y mediante accesos directos. Esto podría estar significando

que existe algún nivel de conocimiento en la gente acerca de la empresa, de la marca o del negocio que está realizando.

Con respecto a los usuarios recurrentes puede verse claramente que hay una estrategia agresiva de marketing ya que los usuarios que previamente ingresaron al sitio, al retornar en su gran mayoría utilizan el canal ”paid”. También aumenta el canal “referral” lo que está indicando que están funcionando los avisos que aparecen en redes sociales u otros medios de referencia. También cobra importancia el acceso directo ya que están la mismo orden de magnitud que los accesos por campañas de marketing de los nuevos usuarios. Esto significa que quienes desean volver al sitio lo hacen principalmente mediante avisos en campañas de marketing, aunque también en una menor proporción mediante accesos directos o por medio de los buscadores

2.1.2 Por dispositivo y canal

Luego de determinar cuáles son los canales más utilizados realizamos una apertura por dispositivo, lo que permitiría optimizar el diseño del sitio en base a las características de los más utilizados.

Figura 2. *Heat map* de Nuevos usuarios analizado por dispositivo y canal. La mayoría de los usuarios nuevos utilizan smartphones para ingresar al sitio. El día 0 es el lunes. La escala de colores indica el número de eventos. Dentro de cada celda se detalla el número de eventos.La escala de colores indica el número de eventos. Dentro de cada celda se detalla el número de eventos.

En la figura 2 puede verse claramente que los ingresos al sitio se hacen principalmente mediante dos dispositivos que son los teléfonos inteligentes (Smartphone) y las computadoras de escritorio y que casi no se utilizan las tabletas, dispositivos que hace unos años estuvieron de moda. No se han analizado en este informe detalles adicionales de los dispositivos como pueden ser el navegador utilizado o el sistema operativo el dispositivo.

2.1.3 Por tamaño de pantalla

Este resultado obtenido con respecto al tipo dispositivo que es utilizado para ingresar al sitio, sugiere un análisis sobre el tamaño de las pantallas de los dispositivos inteligentes que utilizan las personas, dado que es importante asegurarse que el sitio funcione correctamente en estos dispositivos. Por ese motivo se ha realizado un gráfico de Pareto de las resoluciones más utilizadas (figura 3).

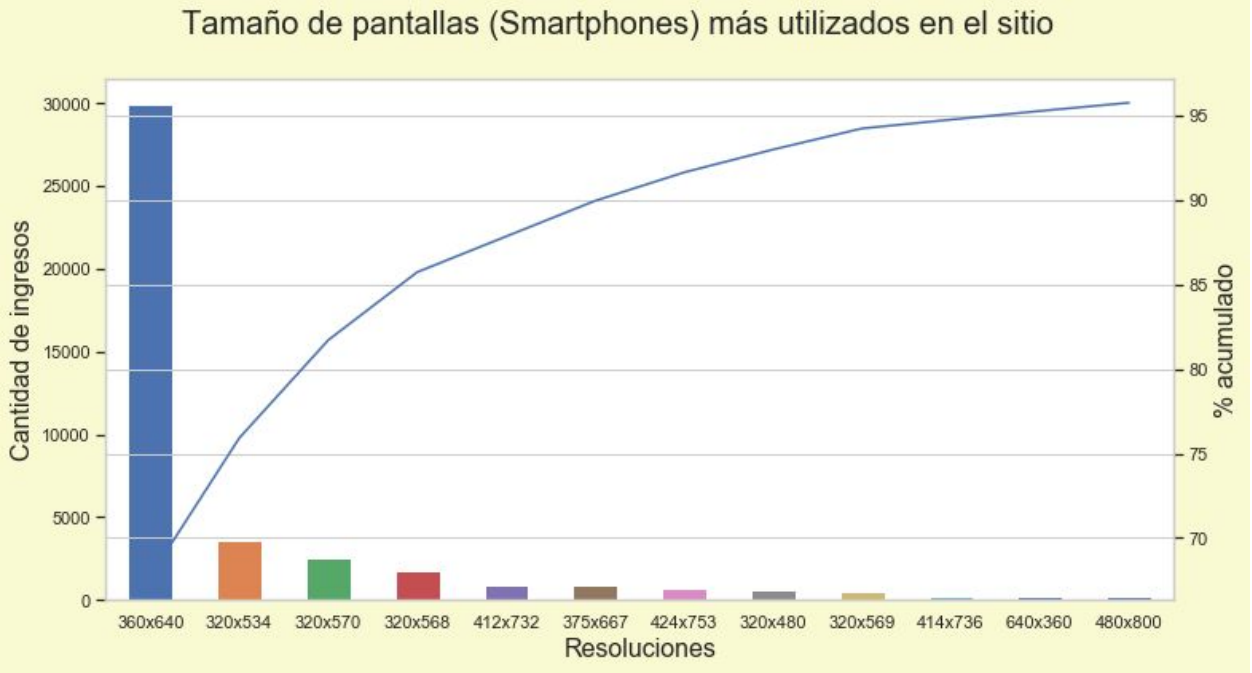

Figura 3. Gráfico de Pareto de Tamaño de pantallas en Smartphones utilizados en el sitio. La resolución de 360 x 640 es la más utilizada.

Puede verse claramente que la resolución de 360 x 640 no sólo es la más utilizada, sino que más de la mitad de los dispositivos inteligentes usados para acceder al sitio tienen esta resolución. También puede verse que dentro del 95 por ciento de los dispositivos utilizados para navegar en el sitio, ninguno tiene resolución HD y en líneas generales son todos dispositivos con pantallas pequeñas.

De todo esto o se puede concluir que es sumamente importante que el sitio se pueda

utilizar en dispositivos con pequeñas pantallas de manera la ágil y funcional.

2.2. Ingresos en relación al momento del evento

2.2.1 Por día de semana y hora del día

Se realizaron análisis por día de la semana y hora, los que permiten determinar cuáles son los horarios más habituales en que ingresan los usuarios al sitio. Esto permitiría realizar acciones más dirigidas a temas puntuales con respecto a campañas especiales de marketing o inclusive para realizar la atención personalizada mediante personas y no mediante bots.

En este análisis se ha ignorado todo tipo de corrección horaria debido a que la mayoría de la información recibida corresponde un único país y dada la falta de tiempo disponible creemos que el resultado sigue siendo significativo. Debido a que no hay información de si los datos se encuentran almacenados en horario local o en horario de Greenwich, asumimos que el horario es local. Otro tema a tener en cuenta es que cero corresponde al día lunes, uno para el martes y así sucesivamente. Nuevamente debido a la falta de tiempo no se realizó el trabajo de cambiar los números por etiquetas descriptivas lo cual se iba a realizar ajustando la categoría correspondiente al día de la semana.

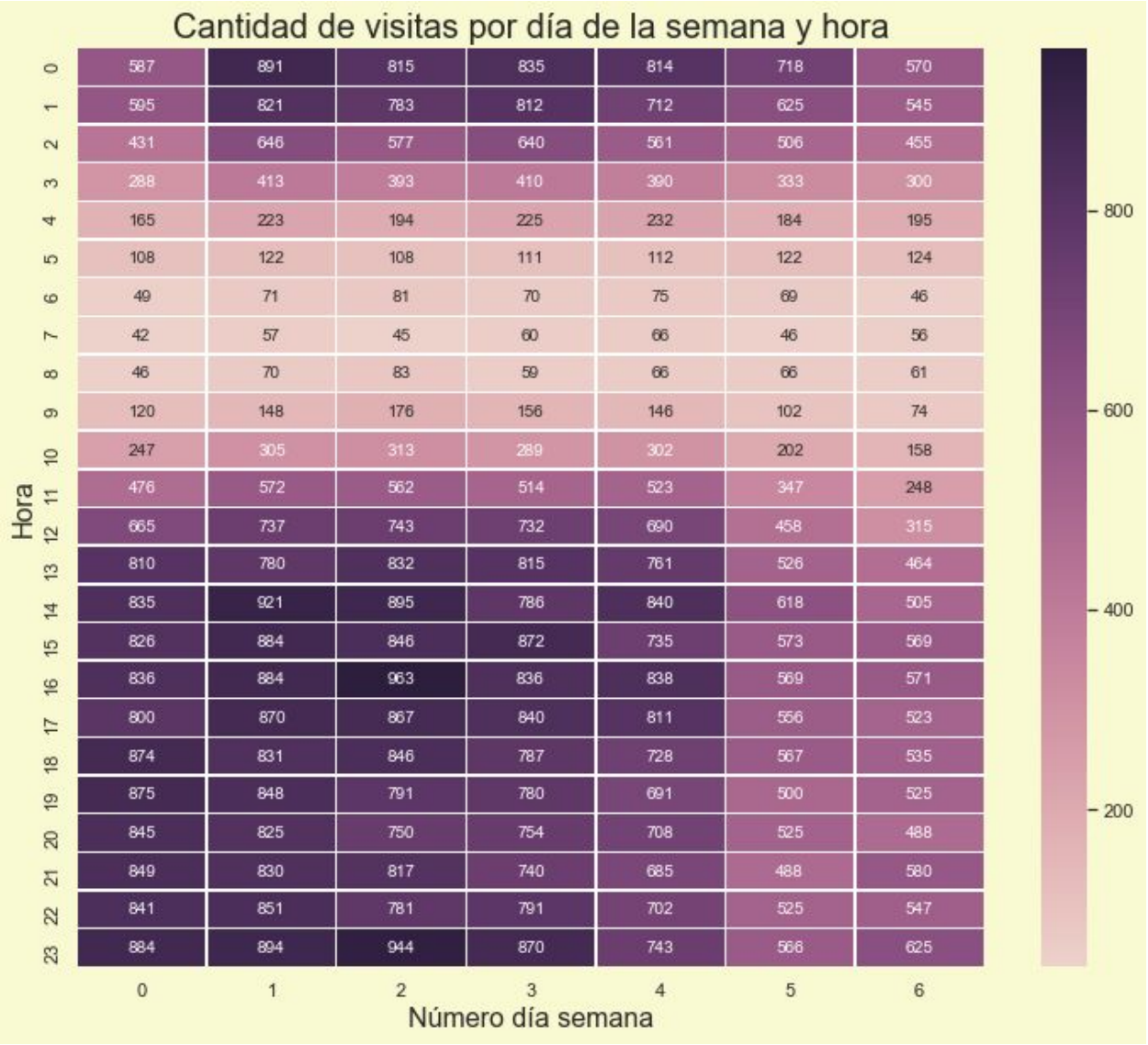

Figura 4. *Heat map* de Cantidad de visitas al sitio por día de la semana y hora. Los días laborables son los días más utilizados para ingresar al sitio. El día 0 es el lunes. La escala de colores indica el número de eventos. Dentro de cada celda se detalla el número de eventos.

Como puede verse en la figura 4 los días laborables son los días más utilizados para ingresar al sitio. Lo mismo ocurre con los horarios a partir del mediodía, son los más utilizados hasta bien entrada la madrugada. No existe ningún día de la semana especial que sea de preferencia para los usuarios. Sólo existe una leve diferencia con respecto a los primeros tres días de la semana laboral lo cual podría ser un indicativo de que los fines de semana gatillan este comportamiento de las personas, ya sea porque comparan sus celulares con los de sus amigos y conocidos o porque esperan a estar en el trabajo o en un día laboral.

2.2.2 Por día del mes y por mes

Así mismo se analizó si había diferencias de visitas de acuerdo al día del mes. El *dataframe* contaba con datos que se originaban el primero de enero hasta el 15 de junio de 2018. En la figura 5 se puede observar un crecimiento gradual en el número de visitas hasta el día 14 de mayo (día 14, mes 5 en la figura 5). En ese momento hubo un salto en el número de visitas que intentamos correlacionar con algún evento de la compañía o del calendario ecommerce. En Argentina, del 14 al 16 de mayo del año analizado, se llevó a cabo el *Hot Sale* que organiza la CACE (Cámara Argentina de Comercio Electrónico). No tenemos suficiente información si ambos eventos están relacionados causativamente, así como tampoco datos para soportar la hipótesis de que a partir de ese evento el número de visitas se mantuvo elevado.

Si los eventos entregados fueran todos y no un subconjunto, el número de visitas en el sitio es compatible con el de una Startup.

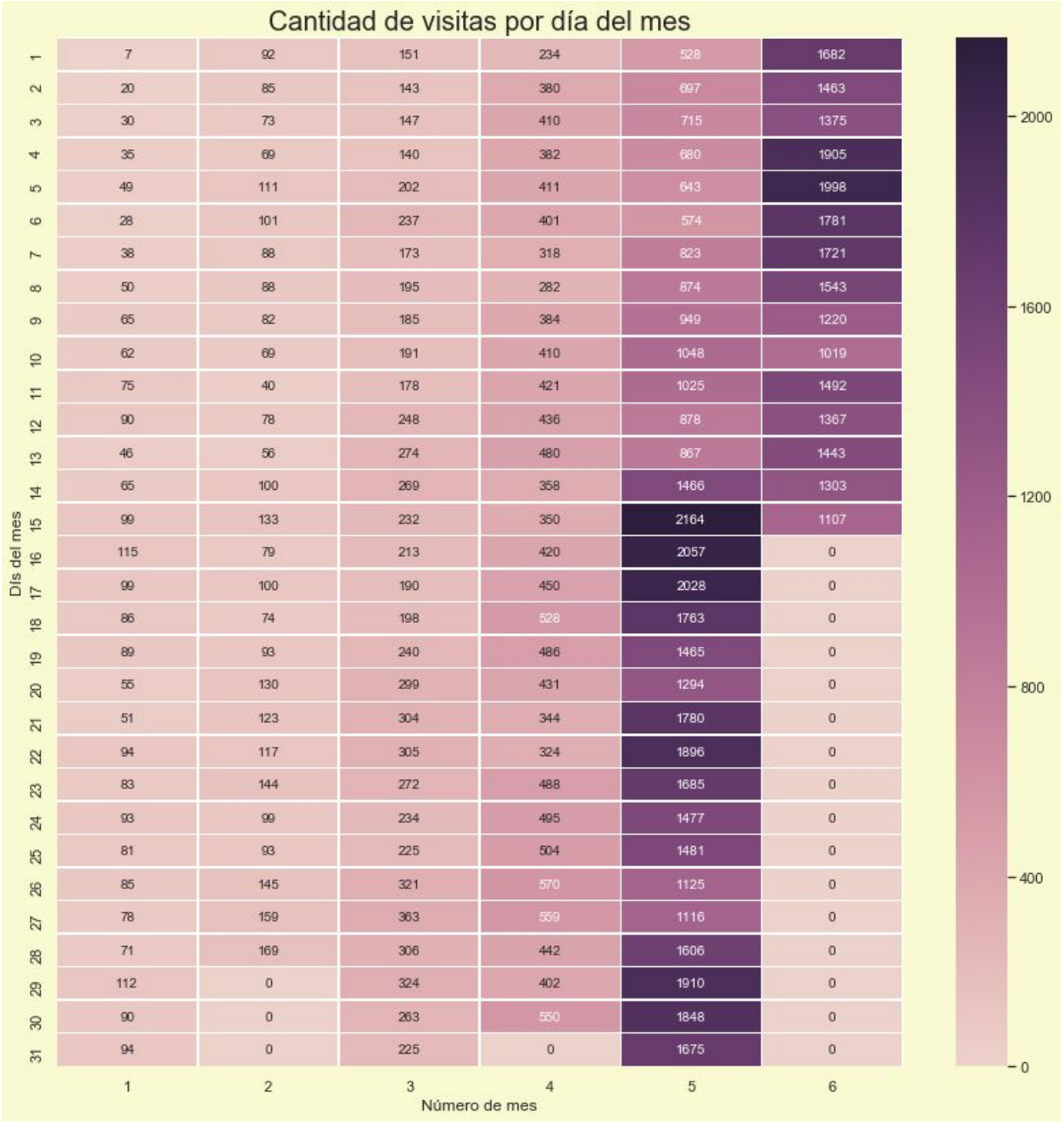

Figura 5. *Heat map* de Cantidad de visitas al sitio por día del mes. A partir del día 14 del mes 5 se nota un aumento conspicuo en el número de visitas. La escala de colores indica el número de eventos. Dentro de cada celda se detalla el número de eventos.

2.2.3 Incidencia de la publicidad por país

Es notorio que Brasil y la Argentina no son los únicos países que produjeron eventos en el sitio. El segundo lugar, de hecho, lo ocupa EEUU y hay ingresos también de Europa.

Viendo esto, se quiere identificar la incidencia de la publicidad por país.

Para este se determinó qué eventos tenían información de país y, con ellos y el dato *person*, se completó en los demás eventos el faltante.

Luego, sabiendo que los únicos eventos que relacionan persona y país son *'visited site'*, se seleccionaron, de esos eventos, persona y país para tener una tabla de países por persona.

Para saber cuántos eventos hay por país, se colectaron los eventos por persona, se hizo un merge con la tabla de personas por país. De este modo, se obtienen los eventos por país (figura 6).

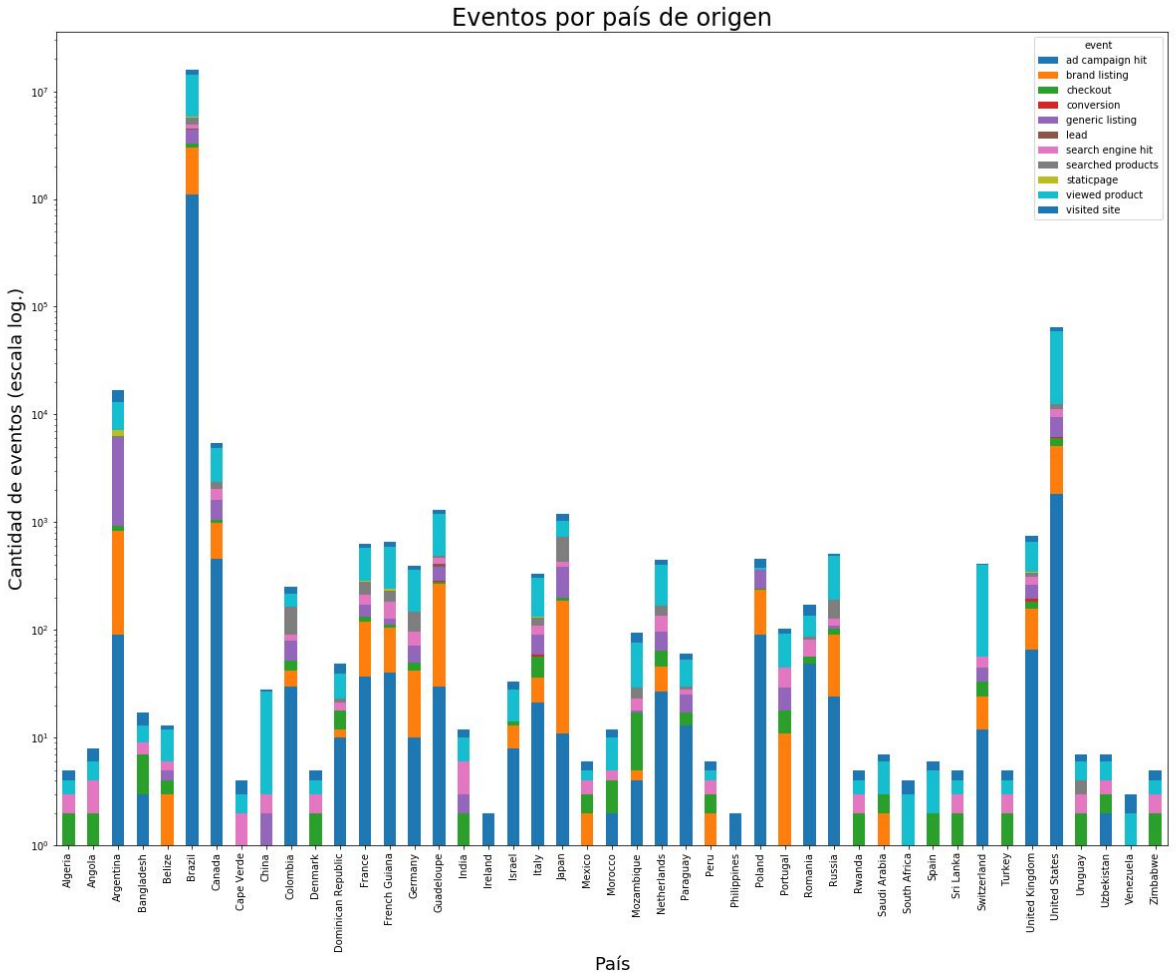

Figura 6. Eventos por país de origen. Los países con mayor cantidad de eventos son Brasil, EEUU y Argentina, en ese orden.

Del gráfico se desprende que los países más significativos en cantidad total de eventos son Argentina, Brasil y Estados Unidos, por ello elegimos estos países para ver el impacto publicitario.

Para esto, se eligieron los hits publicitarios agrupando por campaña y país.

Al igual que en el total de eventos, en los hits es notoria la presencia de Brasil. En referencia al tipo de aviso, en Brasil aparecen todos, pero la compañía de email

marketing Emblue y el buscador Google son los que obtienen hits en todos los países. En Argentina se suma el buscador Bing. Mientras que en el país del norte, Afilio, Criteo y Zenox también consiguieron hits (figura 7).

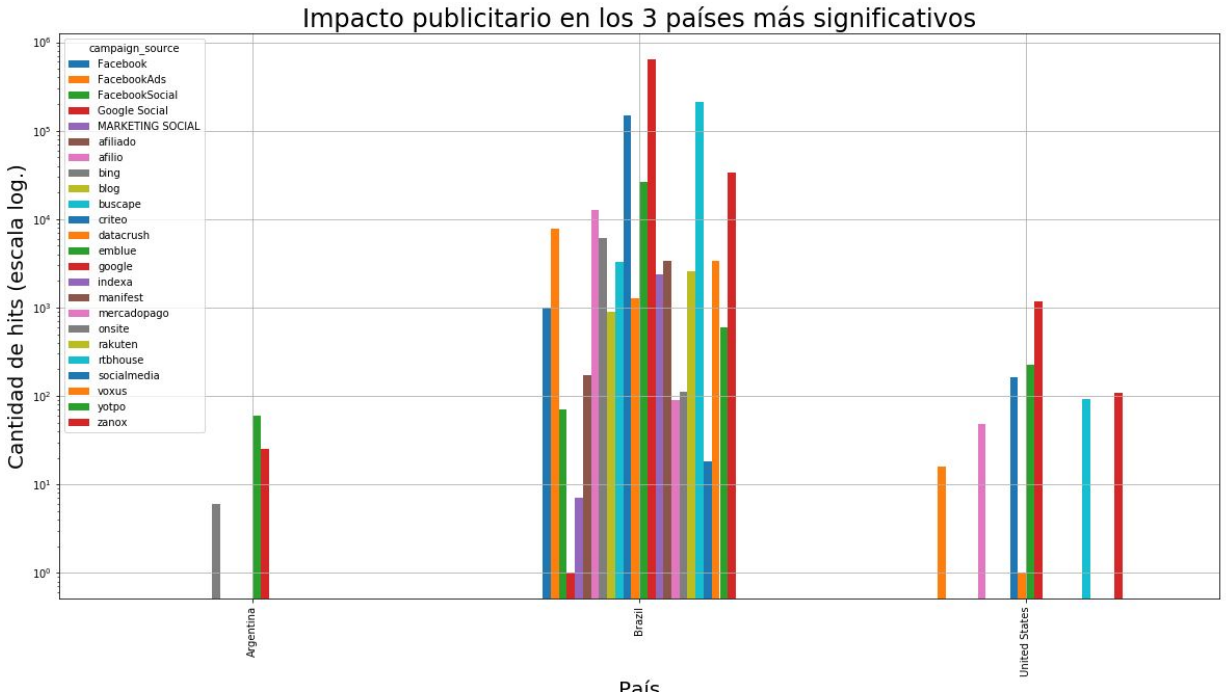

Figura 7. Impacto publicitario en los tres países con más eventos. Se graficó para cada país el *campaign souce*, origen de la campaña. Pueden observarse muchos más orígenes en Brasil, luego EEUU y finalmente Argentina.

2.3 Análisis de conversiones

Como un segundo apartado, se analizarán las conversiones, tanto en relación al momento en que se hicieron, a la marca y modelo, como a otros aspectos estéticos.

2.3.1 Por semana

Figura 8. Conversiones por semana del año. Se observa un aumento gradual en el número de conversiones a medida que transcurre el año.

Como puede verse en el gráfico se produce un aumento gradual en el número de conversiones lo que es compatible con el de una Startup. Se ha explicado anteriormente el aumento significativo durante mayo.

2.3.2 Por día y hora

Al igual que lo que se ve en los ingresos existe una preferencia en los usuarios en utilizar el sitio para realizar las conversiones en horario laboral y los días laborales de la semana.

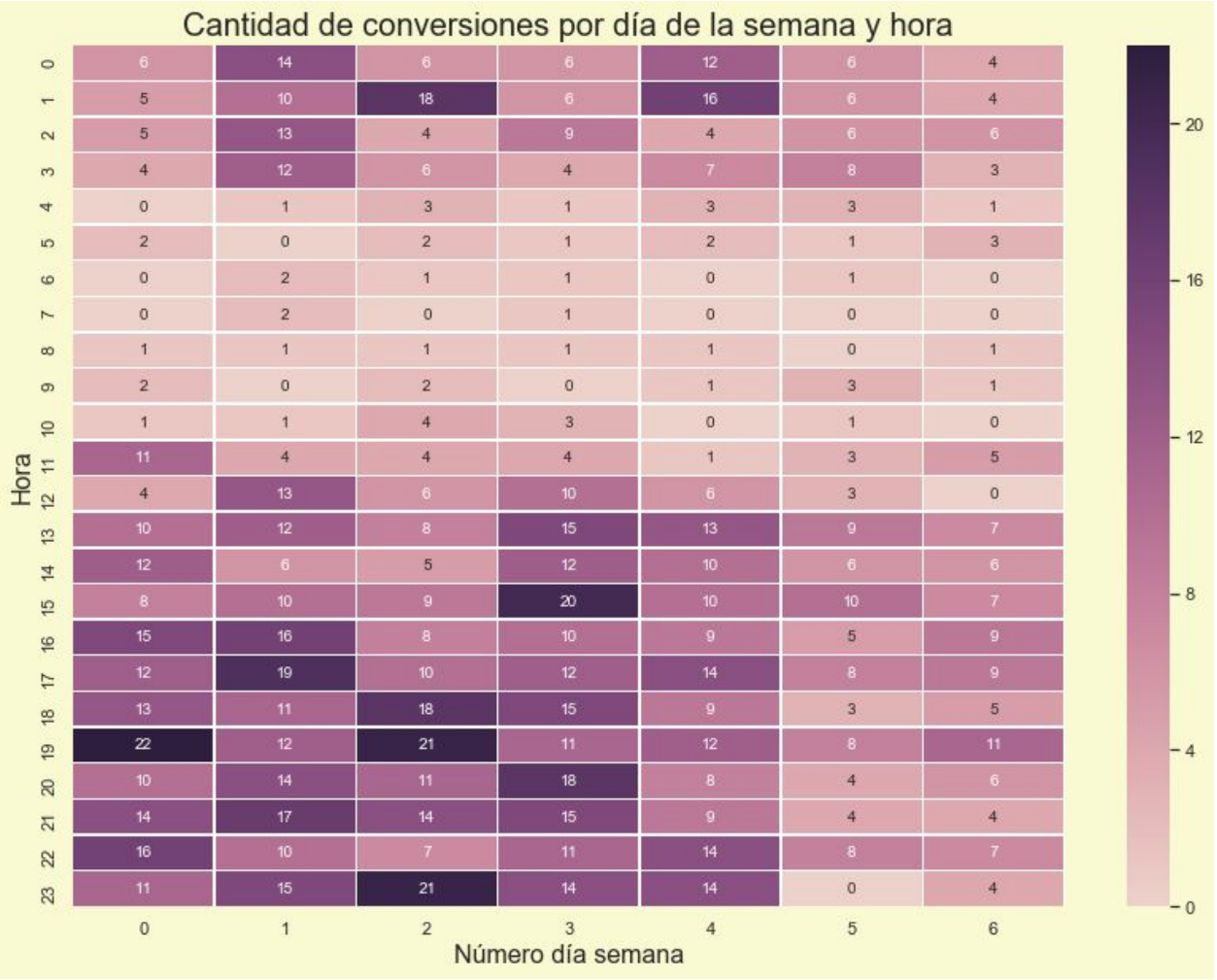

Figura 9. *Heat map* de Cantidad de conversiones por día del mes. Al igual que con las visitas al sitio, la mayoría de las conversiones se concentran en los días laborables. La escala de colores indica el número de eventos. Dentro de cada celda se detalla el número de eventos.

2.3.3 Por marca

Dada la publicidad que las marcas realizan por fuera de Trocafone, se decidió ver si hay diferencia en algún modelo o marca en particular, a la hora de concretar la compra (figura 5)

Figura 10. Marcas más vendidas. Samsung es la marca más vendida en este sitio.

Notamos que Samsung fue el más vendido en el tiempo observado, duplicando a iPhone que ocupa el segundo lugar duplicando por su parte a Motorola (figura 10).
Nótese la gran diferencia con las demás marcas, que no pasaron las 30 unidades en los seis meses observados (tabla 1).

Tabla de Conversiones por marca

Marca	Conversiones
Samsung	641
iPhone	323
Motorola	149
Sony	23
LG	20
Lenovo	14
Asus	2

Tabla 1. Conversiones por marca. Samsung es la marca más vendida en este sitio.

Por otra parte, desagregando por modelo, vemos que el Samsung Galaxy J5 es el más vendido, seguido por los iPhone 5s y 6 (tabla 2 y figura 11).

Tabla de Conversiones por marca y modelo

marca	model	
Samsung	Samsung Galaxy J5	88
iPhone	iPhone 5s	84
	iPhone 6	71
Samsung	Samsung Galaxy S6 Flat	39
iPhone	iPhone 6S	37

Tabla 2. Conversiones por marca y modelo. Samsung Galaxy J5 es el modelo más vendido en este sitio.

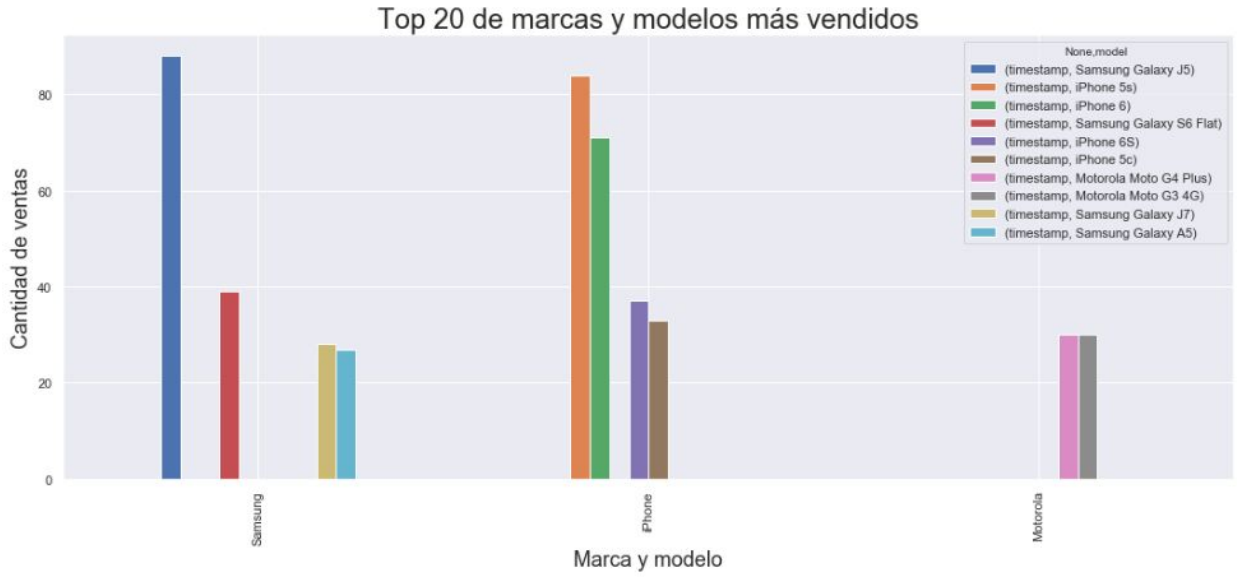

Figura 11. Top 20 de marcas y modelos más vendidos. Samsung modelo Galaxy J5 es el más vendido en este sitio.

2.3.4 Por color

Dada la variedad de colores que últimamente ha aparecido en el mercado, se quiso validar si el negro sigue siendo el más vendido. Es así, como puede verse en la tabla siguiente, seguido por dorado y blanco (tabla 3).

Tabla de ventas por color

color	
Preto	368
Dourado	269
Branco	216
Cinza espacial	89
Prateado	51

Tabla 3. Colores de teléfonos más vendidos.

Nos pareció interesante cruzar esta información con la condición. Si consideramos que la firma garantiza el funcionamiento en todas sus unidades, eso nos daría que la 'condición' es solo una variable estética que bien podría relacionarse con el color.

He aquí una figura de los colores (figura 12), agrupados por condición, de la que extraemos que

- Negro sigue siendo el color más vendido.
- Si el dispositivo es 'excelente', el dorado y el blanco ocupan el segundo lugar.
- Para los dispositivos 'bueno' y 'muy bueno' supera el negro también, pero por menor diferencia.
- Los restantes colores no se acercan a las ventas en estado 'Excelente', ni siquiera el plateado o gris; pero la curva no es tan pronunciada en bueno o muy bueno.
- En cuanto a las ventas por condición, los usuarios eligieron la mercadería en mejor condición.

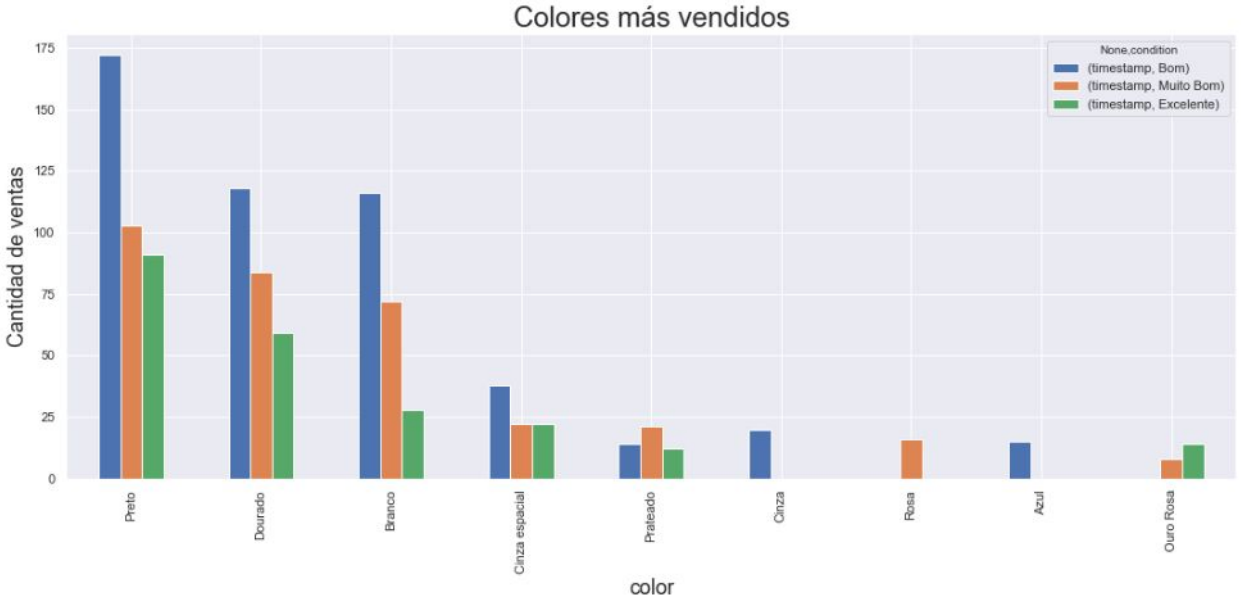

Figura 12. Colores más vendidos. El color negro es el más vendido en este sitio.

2.4. Análisis del pedido de notificaciones de stock

El mismo análisis realizado para las conversiones, puede hacerse para la solicitud de notificación que solicita un usuario al no encontrar stock del producto que busca.

Los resultados agrupados por marca difieren de las conversiones sólo en cantidad: las mismas marcas siguen liderando las listas, pero es notoria la inclinación hacia iPhone esta vez.

Sin embargo, al agrupar por modelo son diferentes los que se solicitan de los que se venden (tabla 4, figura 13).

Tabla Cantidad de pedidos por marca y modelo

marca	model	
iPhone	iPhone 6 Plus	35
	iPhone 6S Plus	31
Samsung	Samsung Galaxy J7 Prime	26
iPhone	iPhone 6S	21
	iPhone 8	20

Tabla 4. Cantidad de pedidos por marca y modelo.

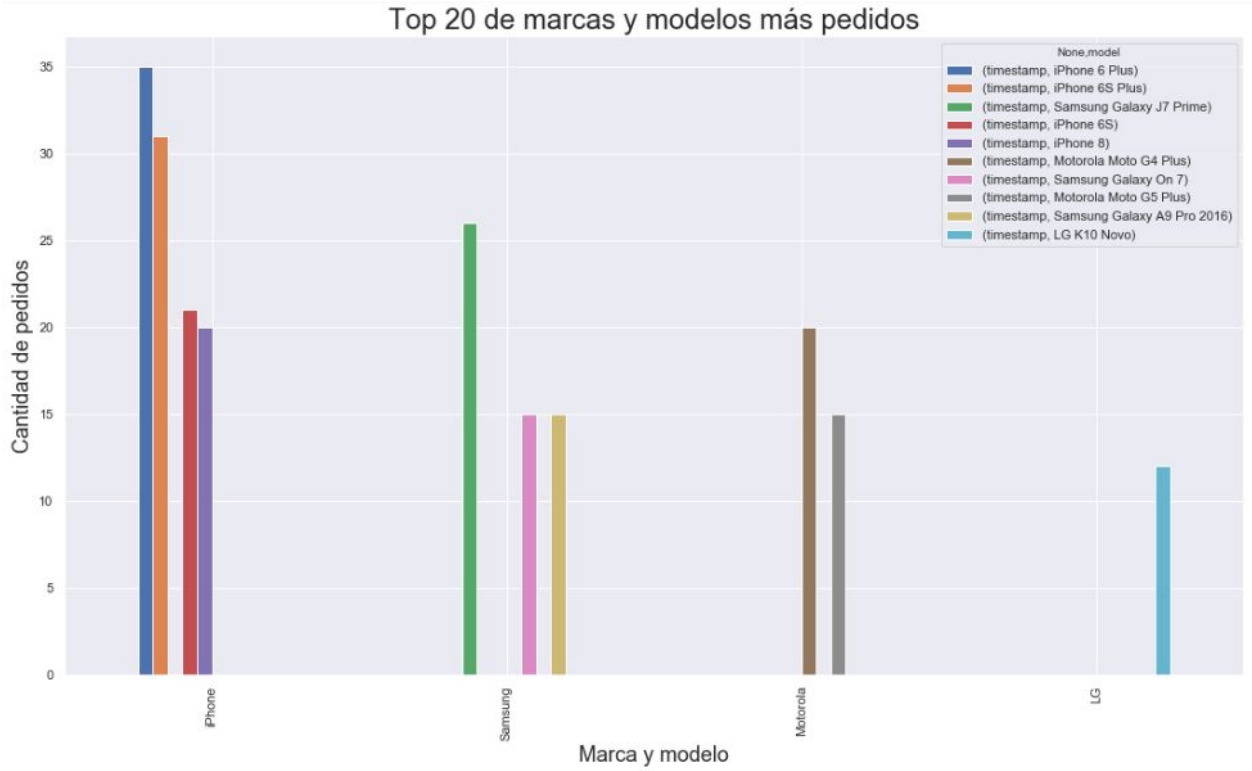

Figura 13. Top 20 de marcas y modelos más pedidos. iPhone es la marca más solicitada.

3. Comportamiento de los usuarios

3.1. Computadora

El siguiente gráfico representa la cantidad de usuarios únicos para cada evento dentro del sitio y para el universo de datos recibido. Solo se ha realizado una apertura por el canal inicial por el que ha ingresado la primera vez, independientemente de cuál fue el mecanismo por el cual retornó al sitio. Tener en cuenta que por ese motivo existen valores para por ejemplo el canal “Orgánico” y el evento “marketing hit”. Esto significa que si bien el canal por el que el usuario ingresó la primera vez fue orgánico en sus siguientes visitas usó otros canales (figura 14).

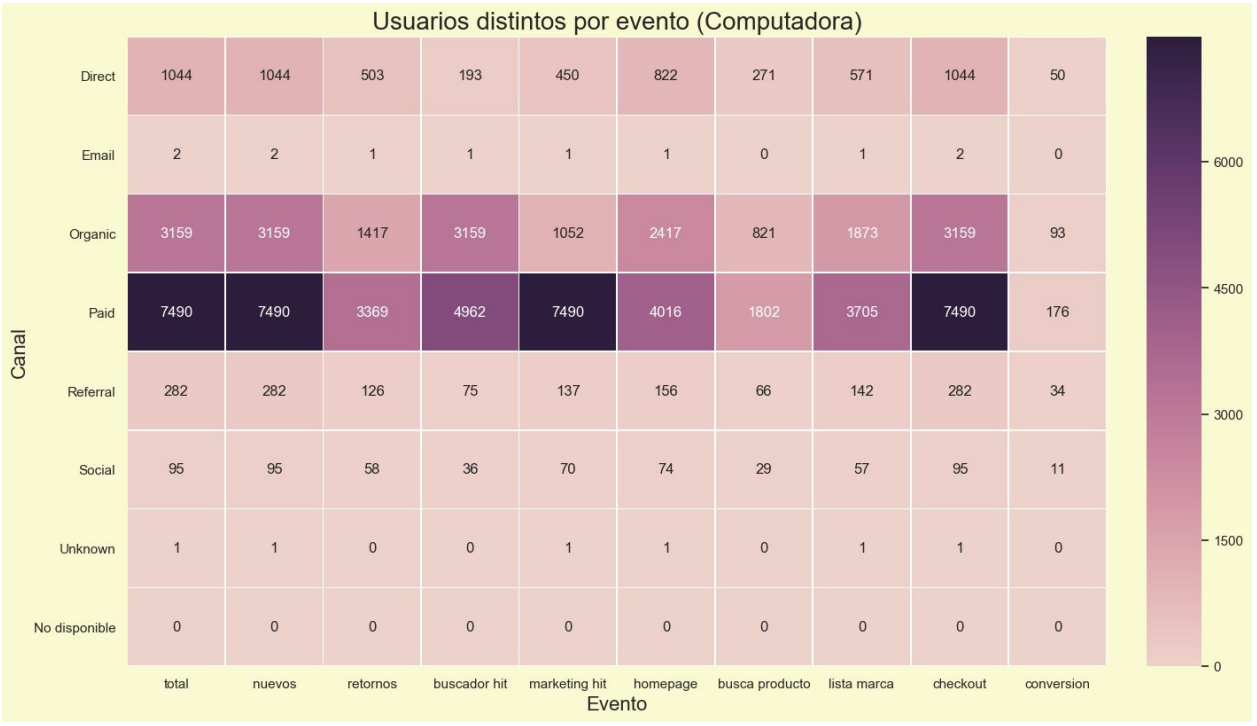

Figura 14. *Heat map* de Usuarios distintos por evento que ingresan desde una computadora. Solo se muestra el canal inicial. Al igual que con las visitas al sitio, la mayoría de las conversiones se concentran en los días laborables.La escala de colores indica el número de eventos. Dentro de cada celda se detalla el número de eventos.

Existe en los usuarios que ingresan en computadora una mayor número de consultas a la página de listado por marca. Esto puede deberse al diseño del sitio y a que la mayoría de los smartphones usado en el sitio tienen pantallas pequeñas

Otro aspecto que puede verse es que hay más tráfico orgánico que en los smartphones, dónde la mayoría del tráfico es *paid* (figura 15).

3.2. Smartphones

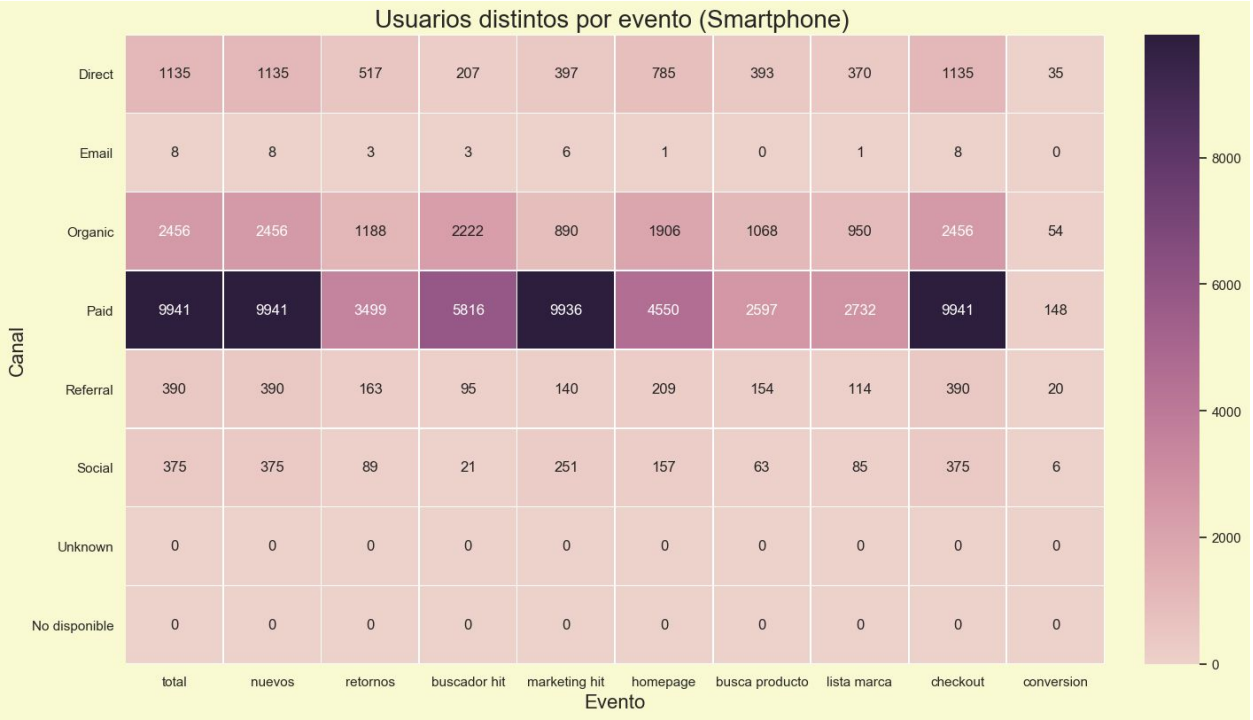

Figura 15. *Heat map* de Usuarios distintos por evento que ingresan desde un Smartphone. Se observa un mayor tráfico pago. La escala de colores indica el número de eventos. Dentro de cada celda se detalla el número de eventos.

3.4. Usuarios activos vs conversiones

Si bien el nivel de conversiones está en valores esperables no existe un aumento significativo de las conversiones relacionado al aumento de tráfico en el sitio (figura 17).

Figura 17. Conversiones versus usuarios activos. Se observa un aumento del número de usuarios activos (línea contínua) pero a su vez el número de conversiones se mantiene prácticamente constante (línea punteada).

Tal como expresara anteriormente existe una caída en la eficiencia de conversión. En este caso la eficiencia no se mide en términos de tráfico de usuarios sino en % de conversión para usuarios activos, esto es, cantidad de conversiones por cantidad de usuarios diferentes activos (figura 18).

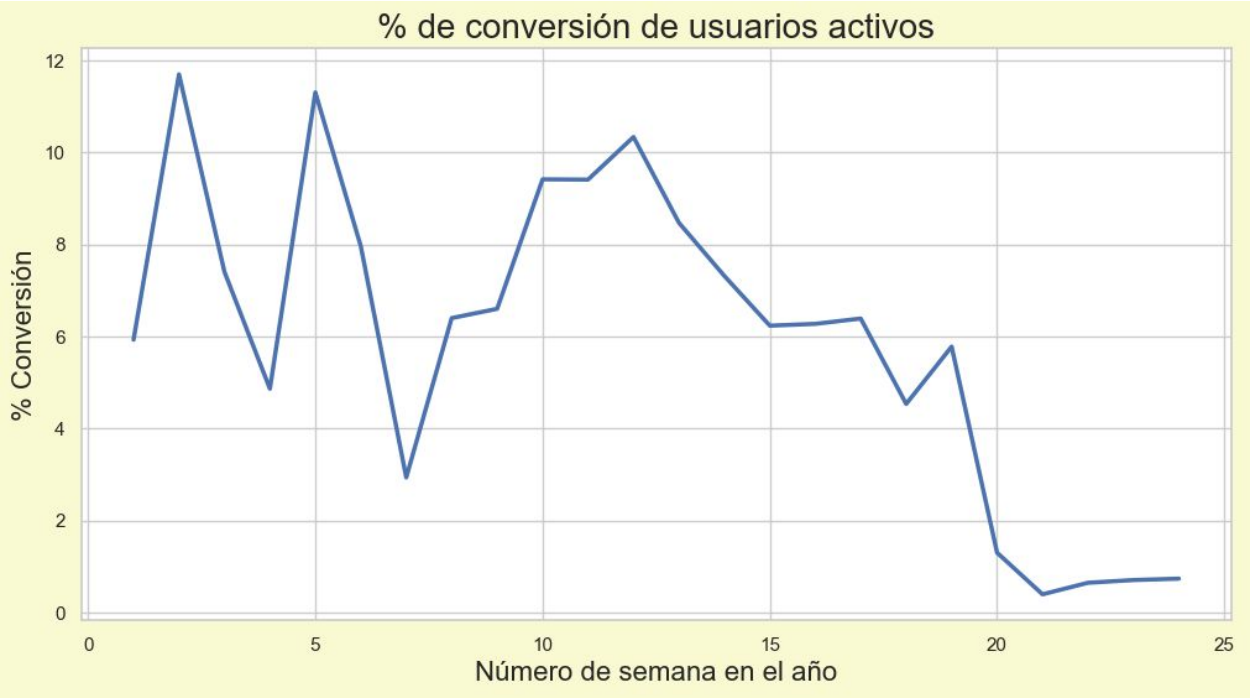

Figura 18. Eficiencia de la conversión de usuarios activos. Se observa una caída en la eficiencia (expresada en %).

4. Conclusión

Los ingresos al sitio crecieron hacia el mes de mayo. Sin embargo, las conversiones s e mantuvieron constantes. Bajó la eficiencia de la conversión en relación a los nuevos usuarios.

Por otra parte, es notorio que los usuarios de la página, tanto en conversión como en visitas, se corresponden con la tendencia mundial de navegación: *mobile*.

Los eventos recibidos no parecen ser todos los eventos del sitio en ese rango de fechas. Debido a que la totalidad de usuarios únicos recibidos han realizado checkout suponemos que se ha recortado el dataset conteniendo sólo los eventos de estos usuarios. Por eso no podemos aventurar hipótesis en ese sentido. También desconocemos qué eventos (dentro o fuera de la empresa) dispararon los picos que vemos.

En cuanto a la inversión publicitaria por país, si bien Brasil y Estados Unidos lideran los hits junto a Argentina, no hay mayor diferencia en las siguientes localizaciones: Canadá, las islas Guadalupe, Japón, Francia y el Reino Unido podrían ser mercados factibles, si consideramos los hits.

Por último, considerando el estado y el color de los dispositivos, los más elegidos siguen siendo el clásico negro (en cualquier condición), seguido por dorado y blanco, pero solo si el dispositivo está en excelente estado.