FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Trabajo Práctico 1: Análisis Exploratorio de Datos

[75.06 / 95.58] Organización de Datos
Segundo cuatrimestre de 2018

Grupo Datatouille

Alumno	Padrón	Mail
Bojman, Camila	101055	camiboj@gmail.com
del Mazo, Federico	100029	delmazofederico@gmail.com
Hortas, Cecilia	100687	ceci.hortas@gmail.com
Souto, Rodrigo	97649	rnsoutob@gmail.com

https://github.com/FdelMazo/7506-Datos/
https://kaggle.com/datatouille2018/7506-TP1/

Curso 01

- Argerich, Luis Argerich
- Golmar, Natalia
- Martinelli, Damina Ariel
- Ramos Mejia, Martín Gabriel

Índice

1. Introducción

Se propone analizar en el presente informe los datos obtenidos de usuarios que visitaron `www.trocafone.com`, un sitio de e-commerce de compra y venta de celulares reacondicionados, con operaciones principalmente en Brasil. Para ello, la empresa Trocafone nos proporcionó acceso a los datos a través del archivo `events.csv`.

El objetivo principal de este informe es poder realizar un análisis exploratorio abarcativo donde a medida que se exploren los datos se vayan encontrando tanto las preguntas como las respuestas a hacerse. Se propone específicamente:

- Descubrir features en el campo `model`
- Identificar patrones de usuarios que realizan checkouts y conversiones
- Analizar las búsquedas que realizan los usuarios y las *keywords* utilizadas
- Analizar los distintos lugares de dónde se originan las visitas a Trocafone
- Descubrir features jerarquizando alguno de los campos disponibles

Finalmente, entre lo descubierto en el análisis exploratorio y los items marcados, se busca obtener un listado de *insights* aprendidos sobre los mismos y con ellos realizar un aporte a la empresa Trocafone con datos que sirvan para mejorar sus servicios.

2. Información general sobre los datos

Lo primero y básico a analizar es la estructura general de los datos proporcionados, para comenzar a tener una idea de qué es lo que se tiene y qué se puede hacer con ello. Se observa que:

- Estos datos corresponden al período de tiempo comprendido entre el 1 de enero del 2018 al 16 de junio del 2018.
- Son 1011288 registros con 23 atributos, no siempre todos completos.

Los atributos son:

- **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**: Téŕminos de búsqueda utilizados en el evento.
- **static_page**: Identificador de página estática visitada.
- **campagn_source**: Origen de campaña, si el tráfico se originó de una campaña de marketing.
- **text_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 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.

2.1. Manipulación de los datos

Es en este momento del análisis donde se tienen que hacer las configuraciones necesarias sobre el set de datos para poder trabajar mejor más tarde. Las operaciones realizadas incluyen:

- **Conversión de tipo de datos**: Teniendo en cuenta que al cargar el set original no se infiere el tipo de cada dato (qué atributo es numérico, qué atributo es categórico, etc), se convierten los datos para tratarlos por su tipo original. Esto tiene como ventaja principal el ahorrar memoria, ya que en vez de tener variables que almacenan objetos genéricos (y ocupan un bloque genérico de memoria) ahora se pueden tener específicamente categorías, números, valores booleanos y más. Un particular caso que es de gran ayuda es el de tratar el atributo ‘timestamp‘ como una variable del tipo ‘‘.

- **Lidiar con los nulos**: NPor motivos obvios, no todos los registros tienen todos los atributos completos, por motivos obvios (por ejemplo, un evento de compra de producto no tiene asociado una búsqueda de palabras). En la transformación de tipos hay que lidiar con estos, y se tomaron decisiones comode truncamiento o transformación. Una decisión tomada fue que el SKU de un producto '`Not a Number`' es el SKU '`0.0`', así permitiendo que el atributo SKU sea numérico.

- **Data Mining**: Se generan nuevos sets de datos y se extraen atributos importantes de los proporcionados. Por ejemplo, dividir el atributo de tiempo en atributos de mes, día y hora.

- **Limpieza de datos**: Cuando un dato es inválido de entrada es necesario tomar una decisión al respecto. En este caso tomamos como un error de tracking cuando la misma venta es registrada dos veces por el mismo usuario en un corto plazo de tiempo, ya que se toma como algo muy improbable. Estos registros son eliminados. También fueron eliminados los registros cuya sesión (explicado a continuación) sólo contenía de un evento de conversión, ya que no tiene sentido que esto ocurra sin pasar previamente por Checkout.

2.2. Sesiones

Se agregó el concepto de sesión. Que se define como la agrupación de una serie de eventos por usuario, los cuales están todos con menos de 30 minutos de inactividad entre el actual y el anterior.

Esto se realizó para poder entender los pasos que lleva a un usuario a comprar un producto, y poder ver de dónde proviene la mayoría del tráfico del sitio.

También como se aclaró antes, sirve para poder encontrar anomalías en los datos y removerlas.

2.3. Hipótesis sobre el truncamiento de los datos

Se sabe que el dataset proporcionado no representa el conjunto total de datos de todos los eventos realizados por los usuarios en el período de tiempo determinado. Es por esta razón que se busca elaborar una hipótesis en base al criterio con el que se fijó la selección de los datos. A partir de un análisis de los mismos se observó que todos los usuarios registrados en el dataset realizaron al menos un checkout, por lo que la base de datos original se truncó. Sin ir más lejos es evidente que no todos los usuarios que ingresan al sitio de Trocafone van a realizar un checkout.

A esta información se le adiciona que los datos de entrada son solamente el tráfico del *sitio web* de Trocafone, y no de la empresa entera, que tiene más actividad que la del sitio. Por ejemplo, venderle a sitios terceros [1].

Con estos dos datos en mente, es importante remarcar que las conclusiones a las que se llegará en el desarrollo del Trabajo se basan en un sector segmentado de los datos, y que estos datos son un sector segmentado de la empresa. Por lo tanto,qen ciertos aspectos del análisis no se podrá arribar a conclusiones fundadas sobre la totalidad de los servicios de Trocafone. L, y formación podría

[1] https://medium.com/trocafone/el-maravilloso-mundo-de-trocafone-5bdc5761856b

parecer poca para el tamaño de la empresa pero hay que tener en cuenta que, siempepresenta el sitio web.

3. Análisis de eventos

En esta sección se propone analizar los distintos tipos de eventos realizados por los usuarios de Trocafone.

El campo `event` puede adquirir distintos tipos de valores categóricos que se describen como sigue:

- **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.
- **static page**: El usuario visita una página.
- **conversion**: El usuario realiza una conversión, comprando un producto.
- **lead**: El usuario se registra para recibir una notificación de disponibilidad de stock, para un producto que no se encontraba disponible en ese momento.

3.1. Conversion rate

En primer lugar se analiza la métrica llamada *conversion rate* o tasa de conversión debido a su importancia en cualquier negocio de e-commerce.

La tasa de conversión es el porcentaje de visitantes que completan un objetivo deseado (en este caso realizar una compra de celular) sobre el total de visitantes. En otras palabras es la razón entre las conversiones y el total de eventos.

Tomando todos los datos del dataset dicha tasa de conversión es de `0,096`.

Se busca analizar la evolución de la *conversion rate* a lo largo del tiempo. Para ello se realiza un gráfico de la conversion rate a lo largo de las semanas del año en el gráfico 1.

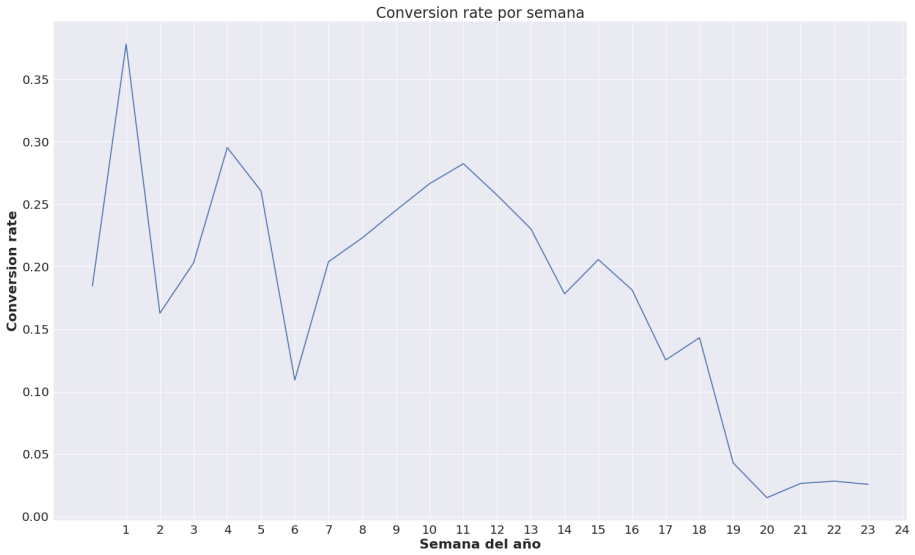

Figura 1: Tasa de conversión a lo largo de las semanas del año

Las mayores tasas de conversiones se registran en las semanas 1,4 y 11, es decir, enero y marzo. Para observar mejor este fenómeno se realiza otro gráfico (2) que muestre la evolución de la tasa de conversión a lo largo de los meses.

Se refuerza la teoría de que enero y marzo fueron los meses de mayor tasa de conversión. Así mismo, dicha tasa se mantiene estable por 4 meses para luego tener una baja en los meses de mayo y junio. Más adelante se analizarán estos dos meses en profundidad.

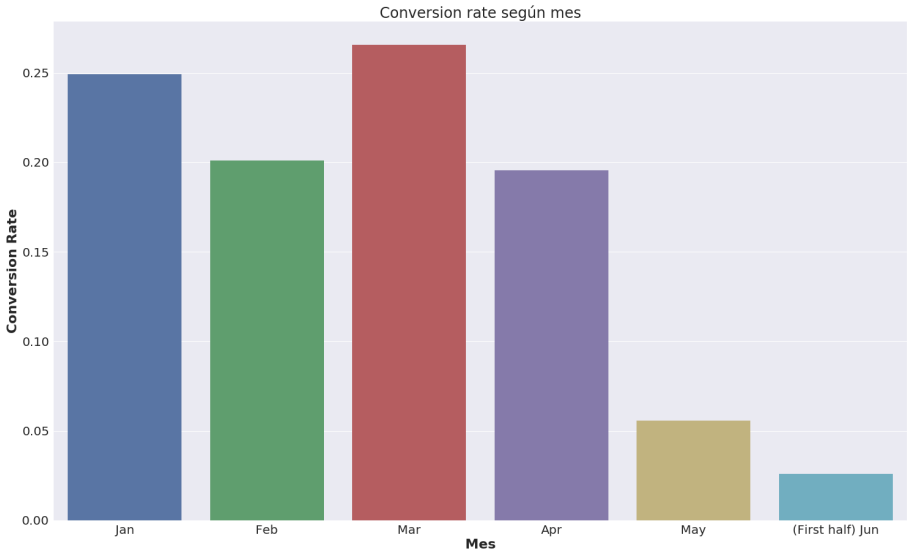

Figura 2: Tasa de conversión a lo largo de los meses del año

3.2. Frecuencia de eventos

Se analiza qué tipo de evento es el más frecuente en el dataset. Para ello se grafica la cantidad registrada de eventos en función de los distintos tipos de eventos.

Se observa en el gráfico 3 que la mayor cantidad de eventos se relacionan a la vista de un producto, lo cual era previsible ya que Trocafone es una plataforma de e-commerce y ver productos constituye su principal función como sitio.

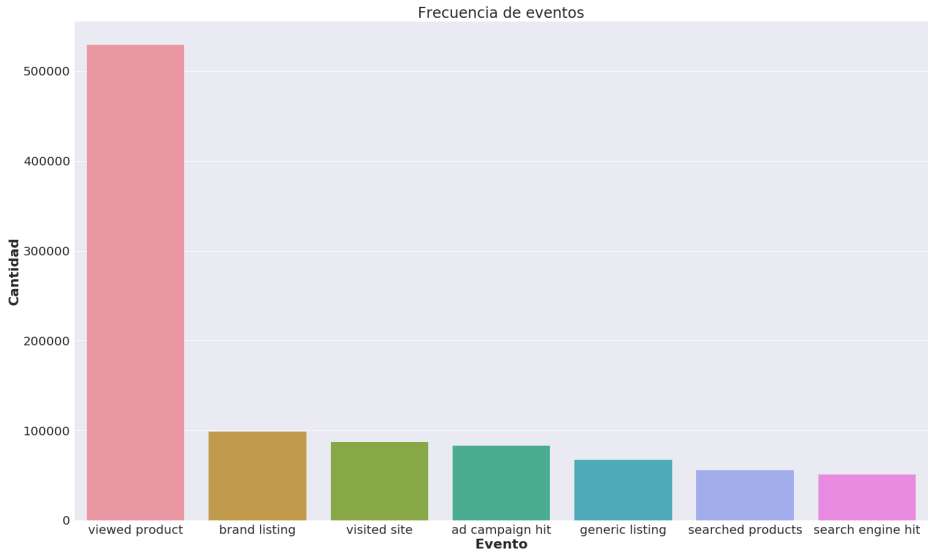

Figura 3: Frecuencia de eventos

3.3. Evolución de los eventos a través del tiempo

3.3.1. Tráfico del sitio de acuerdo al mes y al día

Otro aspecto a analizar es la cantidad de eventos producidos en cada día de la semana y del mes. Se busca detectar si se mantiene algún comportamiento específico a lo largo de los meses o si la cantidad de eventos registrada depende de algún factor temporal. Un patron esperado a encontrar es el de si hay más visitas o compras de celular en las primeras semanas del mes, lo cual coincidiría con el pago de sueldos mensuales.

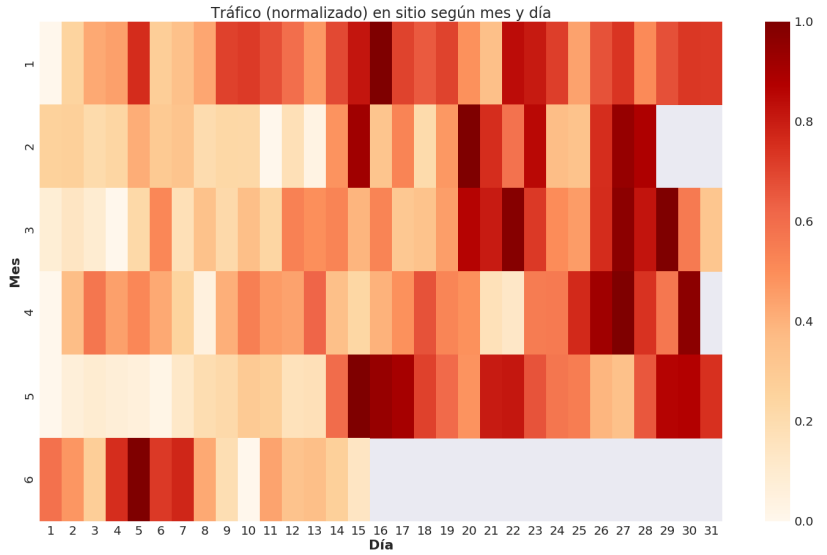

Figura 4: Eventos segun mes y día

Para realizar este gráfico los datos fueron normalizados para evitar llegar a la conclusión que el mes con una mayor cantidad de eventos es el mes con más eventos por díaen totaoteLo cual sería un error muy grave en el análisis, famosamente conocido gracias a la ecuación de de Moivre.

Se desprende del gráfico 4 que la cantidad de eventos registrada no presenta ningún comportamiento específico. Se observa que dicha cantidad aumenta en la segunda quincena de cada mes pero se considera que la diferencia con el resto de los días no tiene la magnitud suficiente como para extraer alguna conclusión fundada.

3.3.2. Tráfico del sitio de acuerdo al mes y al día de la semana

Sin haber encontrado nada acerca del número de día, se buscar ahora analizar si algún día de la semana se registra una mayor cantidad de eventos.

Se realiza el gráfico 5 del mismo estilo que el anterior y se normaliza por las mismas razones.

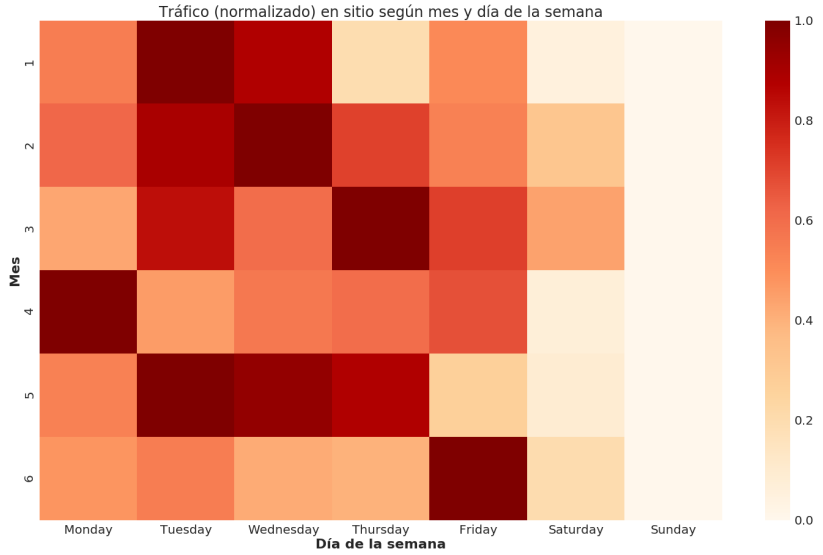

Figura 5: Eventos segun mes y día

Es notable que durante los días hábiles de la semana el tráfico es mucho mayor que al fin de semana. Esto puede deberse a que los fines de semana suelen ser días de descanso, donde la gente puede no estar pensando en realizar una compra, además de no poder retirarla. En la semana aumenta el tráfico debido a que el envío o el retiro del celular puede realizarse en el momento.

3.3.3. Tráfico del sitio según mes

Habiendo analizado las semanas, ahora se hace un enfoque más global, buscando patrones de tráfico según el mes. En este apartado se busca analizar si en algún mes se registró una mayor cantidad de eventos o si la distribución de las visitas fue uniforme a lo largo del tiempo.

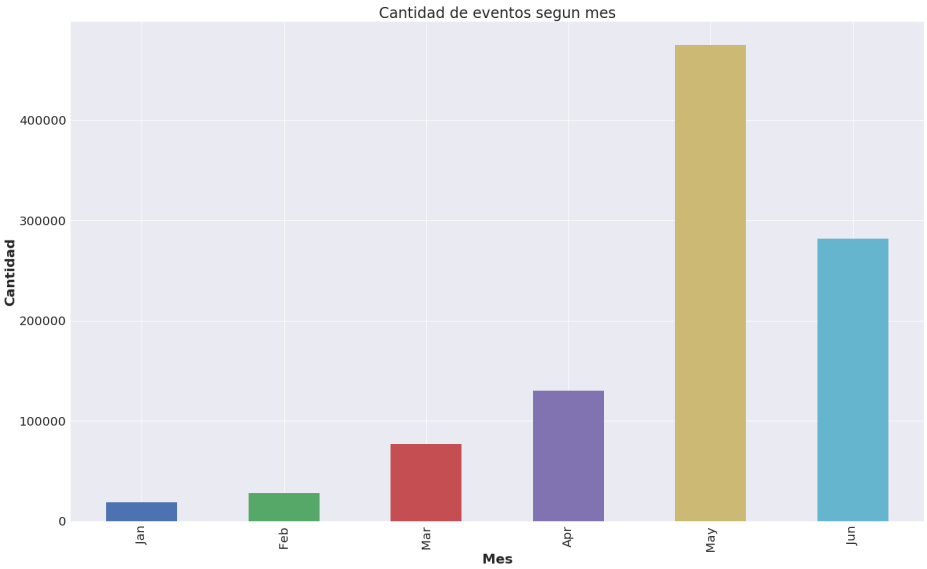

Figura 6: Eventos segun mes

El gráfico 6 muestra que en los meses de mayo y junio se registró una cantidad notablemente mayor de eventos. Este resultado llama la atención por lo que se analiza en mayor profundidad en la próxima sección. Es importante destacar que esta evolución es inversa a la de la tasa de conversión. De estos dos datos se concluye que en mayo si bien no hubo tantas ventas, sí aumentó mucho la cantidad de eventos.

3.3.4. ¿Por qué mayo y junio registran una mayor cantidad de eventos?

Para tratar de encontrar una respuesta a esta pregunta se centra el análisis en estos meses y se estudian los tres eventos principales del dataset: `conversion`, `checkout` y `viewed products`. Es pertinente recordar que no se dispone de todos los datos del mes de junio, sino solo de los primeros 16 días.

Como se ve en la figura 7, los tres eventos presentan su máximo alrededor de los días 14 a 16. Como Trocafone es del país de Brasil y el mayor tráfico proviene de allí (lo cual será verificado posteriormente) se infiere que puede deberse a alguna promoción lanzada en la plataforma o en el mismo país. Esto no puede concluirse con certeza debido a la falta de información en internet y en la plataforma de promociones pasadas.

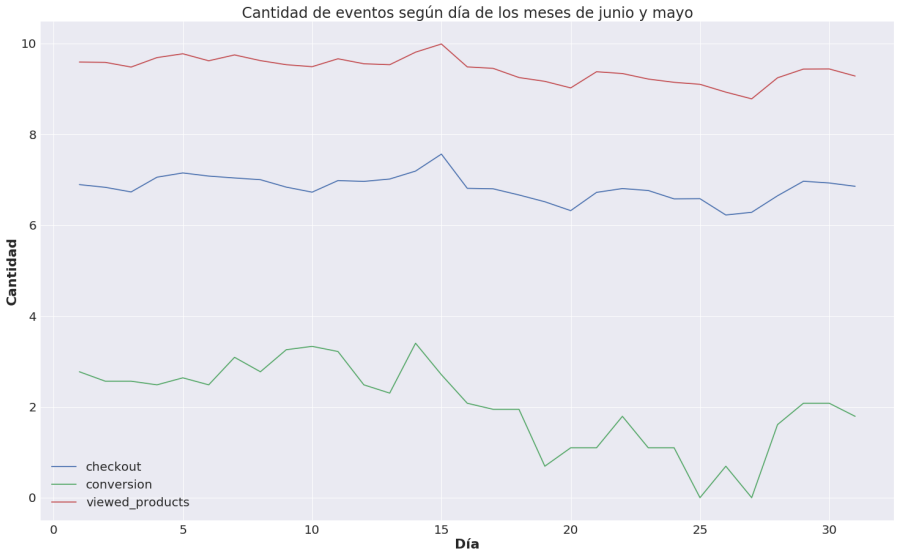

Figura 7: Conversiones, checkouts y viewed products a lo largo de los días del mes de mayo y junio en éscala logarítmica

3.3.5. Hora de mayor cantidad de conversiones y checkouts

En un intento de encontrar un patrón por parte de los clientes se grafica la cantidad de conversiones y de checkouts en función de las horas del día. Se busca determinar la hora en la que ambas confluyan en su máximo para analizar el motivo por el que dicha hora registra mayor tráfico.

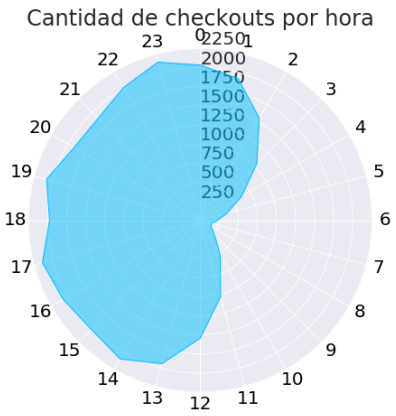

Figura 8: Conversiones a lo largo de las horas del día

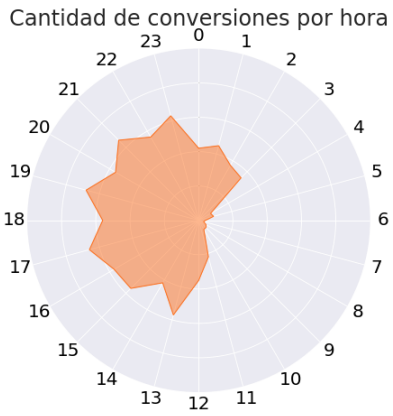

Figura 9: Conversiones a lo largo de las horas del día

Se observa que ambos gráficos confluyen en su máximo a las 19 hs. Esto puede significar que la mayoría de los usuarios deciden realizar conversiones o checkouts cuando vuelven del trabajo o están finalizando su día. La diferencia entre ambos gráficos es que la cantidad de checkouts realizados se mantiene relativamente constante en las segundas 12 hs del día mientras que las conversiones son mucho menores y presentan picos más marcados en los horarios de la tarde-noche.

3.4. Distribución de la cantidad de eventos producidos por usuario

Se propone analizar la cantidad de eventos producidos por usuario. Se confecciona un gráfico tomando algunas salvedades que son necesarias para obtener una buena visualización:

- Se trunca el eje y a un valor determinado para una mejor observación de los *box*
- En los eventos donde no se registran datos para usuarios se coloca el valor promedio.

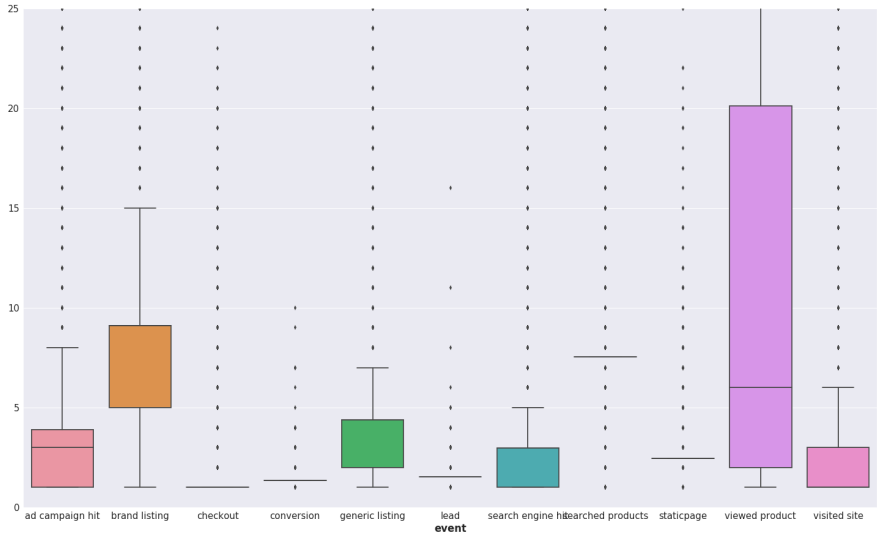

Figura 10: Distribución de los eventos agrupados por usuario

Se puede observar en el gráfico 10 que los usuarios suelen más comúnmente ver los productos antes que comprarlos, algo que podía predecirse anteriormente. Lo que puede resultar llamativo es que la cantidad de usuarios que ven productos es mayor a los que los buscan, pero esto se puede explicar por el hecho de que en una búsqueda pueden verse varios productos a la vez y eso cuenta como un solo evento. En cambio, ante el evento `viewed products` cada vista de producto se contabiliza como un evento.

4. Análisis geográfico

En este apartado se busca analizar las ciudades, países y regiones de dónde provienen los distintos tipos de eventos. Trocafone es una empresa que inició en Brasil y expandió sus comercios a Argentina en el 2015, por lo que se deduce que probablemente Brasil sea la zona de mayor influencia en los eventos.

4.1. Países que registran mayor cantidad de eventos

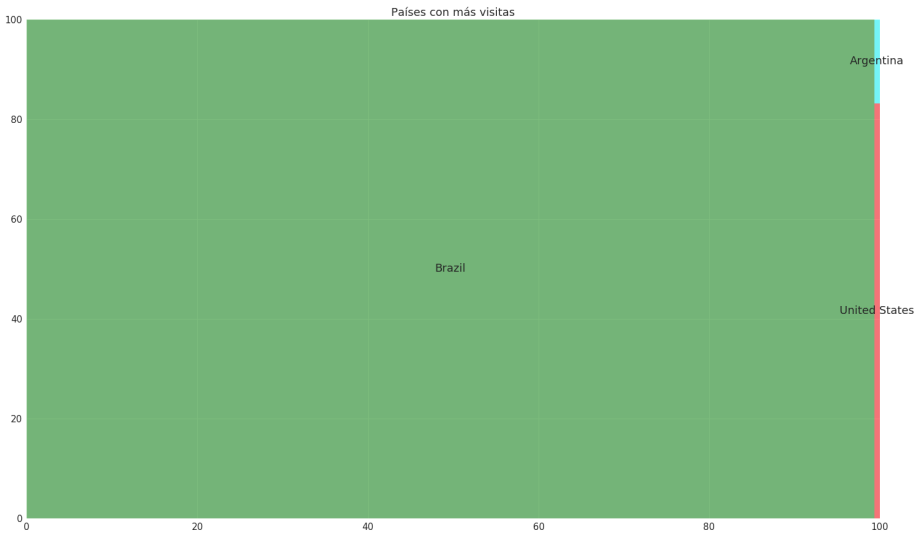

Figura 11: Países de mayor tráfico de la página

Se corrobora en el gráfico 11 la teoría inicial de que Brasil sería el país con más visitas, por lo que se procedió a eliminarlo del gráfico, generando el gráfico 12, para poder observar qué otros países intervienen en la página de Trocafone y en qué diferencia de magnitud y orden lo hacen. Estados Unidos supera en una amplia cantidad la influencia en la página a Argentina, a pesar de ser esta una de las sedes de la empresa. Esto puede explicarse debido a que la gran mayoría de los eventos no son conversiones, por lo tanto es factible que cualquier persona de los Estados Unidos busque celulares en la plataforma, sin llegar a registrar un evento de tipo `checkout` o `conversión`.

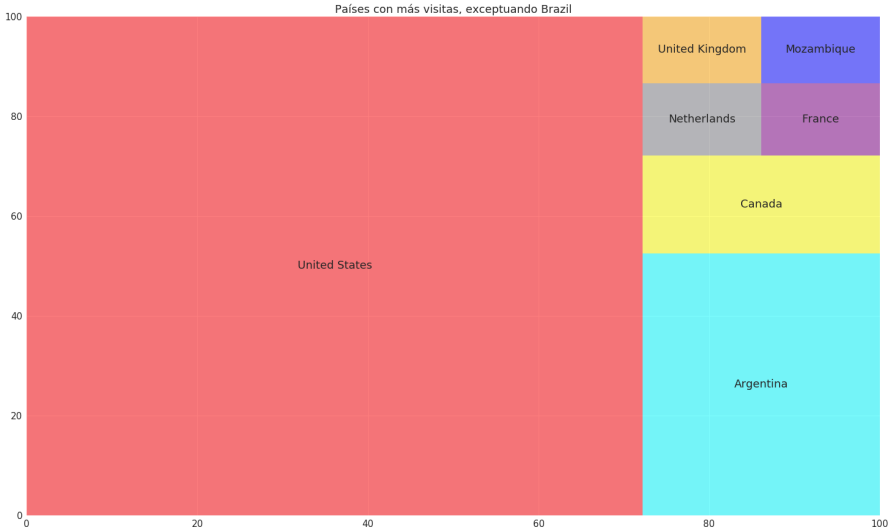

Figura 12: Países de mayor tráfico de la página sin contar a Brasil

4.2. Regiones y ciudades de Brasil que registran mayor cantidad de eventos

Se procede a analizar qué ciudades y regiones de Brasil son las que registran mayor cantidad de eventos. Para ello se grafica las regiones con una mayor cantidad de visitas. Se observa que las tres regiones con la mayor cantidad de eventos (San Pablo, Minas Gerais y Rio de Janeiro) están sobre la costa del sudeste. Para visualizar esto de una mejor manera se realiza un gráfico que muestra las ciudades de Brasil más visitadas y se verifica que la mayoría de los eventos se producen sobre la costa sudeste.

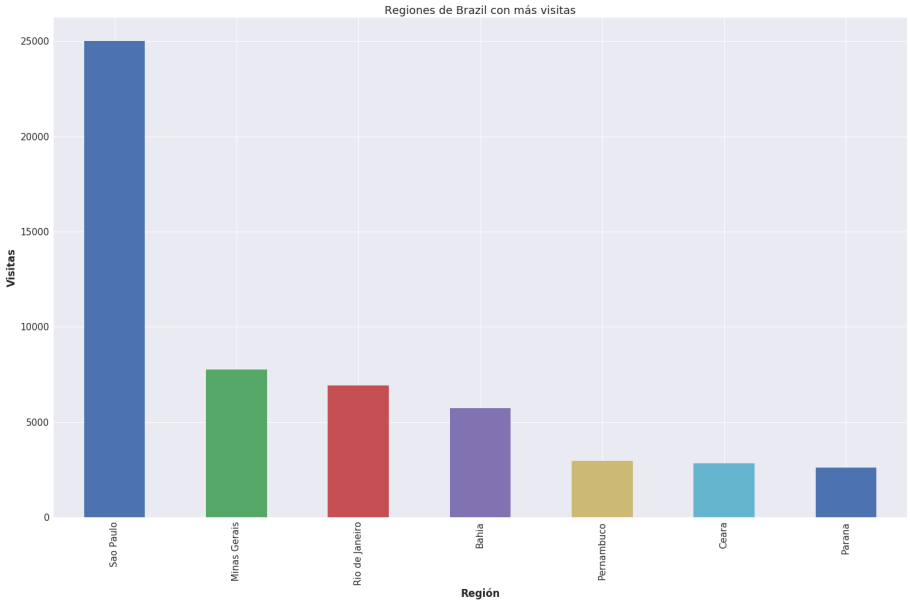

Figura 13: Regiones de Brasil con mayor cantidad de eventos

Lo que presentan los gráficos 13 y 14 tiene sentido, considerando que sobre el mayor blanco del gráfico es donde esta la selva brasileña.

5. Análisis de búsquedas

La idea de este apartado radica en analizar los términos que buscan los usuarios en la plataforma y así identificar ciertos patrones de búsqueda como por ejemplo cuál es el modelo de celular más buscado. Este análisis se dividirá en dos:

- Términos ingresados en el buscador: se utiliza la columna `search_term` del dataframe.
- Productos buscados en la plataforma: se utiliza la columna `event` del dataframe para buscar aquellos eventos que correspondan a `searched_product`.

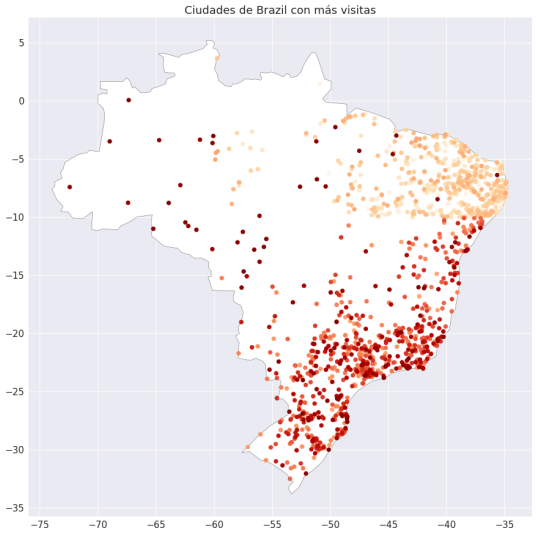

Figura 14: Ciudades de Brasil con mayor cantidad de eventos

5.1. Términos ingresados en el buscador

Se realiza un gráfico para visualizar a grandes rasgos los términos más buscados por los usuarios. Se busca tener una idea aproximada de los modelos de celular más requeridos o deseados por los usuarios. Figuran en el gráfico los términos que fueron buscados como mínimo 300 veces, un número impuesto para fijar un límite mínimo de búsquedas para que un término sea considerado de los más buscados. De no fijar este límite el gráfico estaría sobrecargado y sería difícil de interpretar.

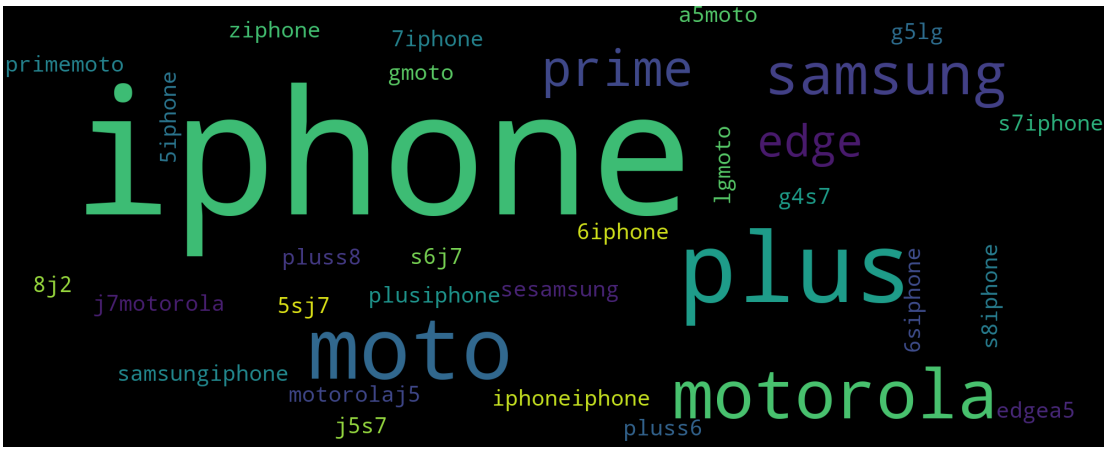

Figura 15: Términos más buscados por los usuarios de Trocafone

Los términos más buscados (vistos en el gráfico 15) son **iPhone**, **Motorola** y **Samsung**. Esto era completamente esperable debido a que son las marcas que dominan el sector tecnológico. Se destaca que suelen buscarse más comúnmente la marca de un smartphone antes que un modelo específico.

5.2. Productos buscados en la plataforma

En esta sección se busca obtener los productos más buscados. Esta búsqueda es más específica que la anteriormente mencionada debido a que corresponde a un producto puntual, no el nombre de su marca, buscado por la interfaz del sitio. De esta manera los productos más buscados son los que se detallan en la tabla 1 y se representan en el gráfico que le sigue.

sku	sku_name
3371	Samsung Galaxy S6 Flat 32GB Dourado (Bom)
2777	Samsung Galaxy S4 i9505 16GB Preto (Bom)
6357	Samsung Galaxy J5 16GB Preto (Bom)
6413	Samsung Galaxy J7 16GB Dourado (Bom)
6371	Samsung Galaxy J5 16GB Dourado (Bom)

Cuadro 1: SKUs más buscados y su nombre

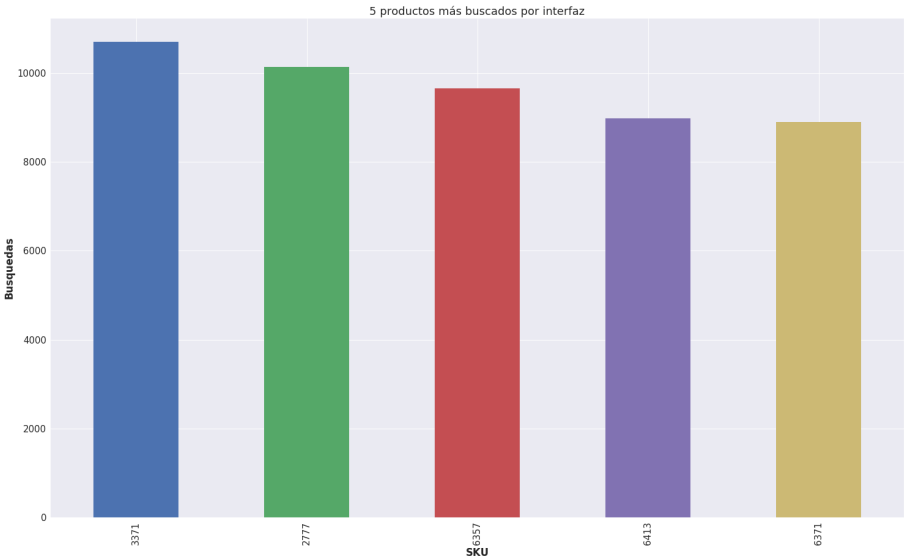

Figura 16: Productos más buscados por los usuarios de Trocafone

Se concluye con el gráfico 16 que si bien **iPhone** y **Motorola** eran los *términos* más buscados en la plataforma, no sucede lo mismo con los *productos*

buscados ya que todos corresponden a la marca **Samsung**. Esto puede deberse a un tema de la calidad que ofrece dicha marca o su precio, probablemente más conveniente. Los iPhone se caracterizan por tener un precio difícil de acceder por lo que es probable que sea buscado como término para ver las diferentes opciones globalmente pero que no muchas veces se busque un producto determinado de dicha marca.

6. Análisis de modelos

Pasadas las exploraciones enfocadas sobre el sitio web en sí, como sus busquedas o las nacionalidades de sus usuarios, ahora se pone el foco sobre el contenido del sitio, las compras de celulares, específicamente sobre que modelos son los más vistos, comprados y en general con más tráfico.

6.1. Relación entre ver, llevar al carrito y comprar un celular

La primera pregunta a hacerse acerca de los celulares es si hay alguna relación directa entre ver un modelo, decidir comprarlo y efectivamente comprar ese mismo modelo. La pregunta surge de la inquietud de si hay efectividad una vez visto el celular. Por ejemplo, puede darse que alguien vea un celular, lo compare y termine comprando otro, o puede darse que uno entre a visitar la página dedicada a un modelo, vea el precio o la condición y decida mejor optar por otra alternativa.

modelos prominentes a analizar
Samsung Galaxy J5
Samsung Galaxy S6 Flat
Samsung Galaxy S7
Samsung Galaxy S7 Edge
iPhone 5s
iPhone 6
iPhone 6S
iPhone 7

Cuadro 2: Subconjunto de modelos prominentes

Entonces, se comienza decidiendo un subconjunto de modelos a analizar, para tener una muestra del sitio. Este set, presentado en la tabla 2, está generado teniendo en cuenta y uniendo los celulares más vistos (`viewed product`) con los más "llevados al carrito"(`checkout`) con los más comprados (`conversion`). Esto es porque de analizar los eventos y los modelos se ve que los eventos (más relevantes[2]) que tienen un modelo asociado son esos tres, y es así como se modela la cronología ordenada ideal de eventos desde el punto de vista de un dispositivo:

[2]se descarta del análisis el evento `lead` por no ser relevante al caso

1. `viewed product`: **Visitar un producto.**
2. `checkout`: **Decidir comprarlo.**
3. `conversion`: **Efectivamente comprarlo.**

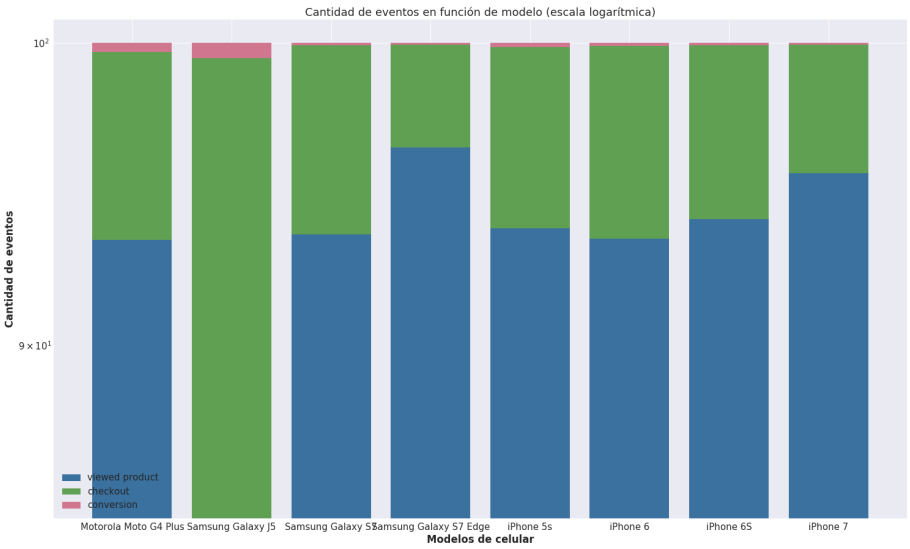

Figura 17: Cronología de eventos de una muestra de modelos analizados

Primero, en el gráfico 17 se analiza como varían los tres eventos según cada celular. Lo que más se puede notar es que los celulares **Samsung Galaxy S7 Edge** y **iPhone 7**, los celulares de mayor gama del sitio son mucho más vistos que el resto, y que verlos es su propio evento predominante por amplia diferencia. Esto se deduce que sucede por el precio y calidad de estos teléfonos; son celulares muy codiciados pero a su vez muy caros, por ende su precio suele ahuyentar compras, pero su calidad atrae visitas.

Lo que no se notó en el primer gráfico fue alguna relación directa entre las marcas de los celulares analizados, en particular si hay algun patrón de **Apple** vs **Samsung**.

En el gráfico 18 se puede ver el ranking de cada celular dependiendo de su evento. Mientras más alto el celular, más eventos tiene. Si bien este gráfico esquiva los números exactos, sirve para ver la posición de cada celular respecto de los otros. Lo que se nota acá es similar a lo dicho anteriormente, a mayor calidad de celular, más visitas, pero esto conlleva mayor precio y por ende menos compras. Algunos casos en particular a ver son el **Samsung Galaxy J5** y el **Samsung Galaxy S6 Flat** que son de los celulares menos buscados pero a su vez de los más comprados, y como caso inverso presentado, se ve que el **iPhone 6** y **iPhone 7** cumplen el rol opuesto, siendo de los celulares más vistos pero menos comprados.

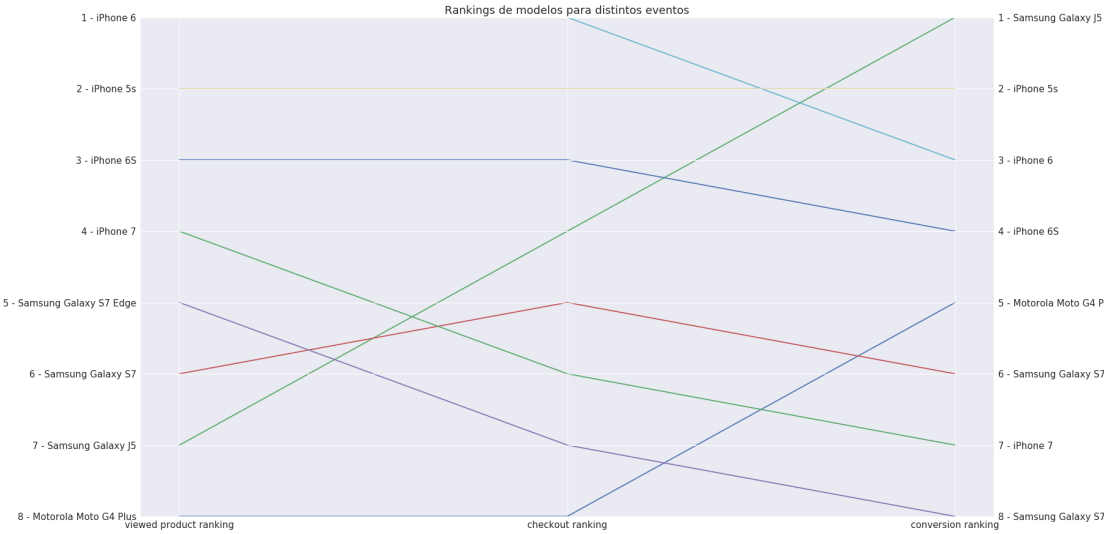

Figura 18: Ranking de modelos prominentes según evento

También, hay una pequeña relación a destacar de las marcas de celulares, donde de los 8 modelos analizados, los 4 celulares más buscados son los de marca **Apple** y los 4 menos buscados son los de su rival, mientras que en la compra pasa (casi) exactamente lo inverso. Se podría presentar un caso de que el análisis de a mayor calidad y precio hay más vistas y menos compras no solo aplica para modelos si no que también para marcas en general, pero esto excede al trabajo presentado.

6.2. Relación entre celulares y sus condiciones

Habiendo encontrado un tan rico análisis entre eventos y celulares, se busca con el mismo objetivo una relación entre la condición de uso del celular y su tráfico. Trocafone clasifica los celulares reacondicionados en **Bueno**, **Muy Bueno** y **Excelente** (no se tienen en cuenta los celulares de conidición **Nuevo** en el presente análisis). Queremos encontrar un patrón de compra y visita de los modelos analizados previamente, pero esta vez según condición.

Lo que vemos nuevamente, esta vez en el gráfico 19 es que hay una diferencia substancial entre celulares marca **Apple** y celulares marca **Samsung**. Para la marca **Samsung** se ve como hay mayor tráfico en los modelos de menor condición, sugiriendo el estar dispuesto a comprar celulares no en perfecto estado, mientras que para los **Apple** aparenta haber una demanda por celulares en muy buena condición, insinuando que se quiere excelencia tanto en calidad de software (responsabilidad de **Apple**) como en calidad de hardware (responsabilidad de **Trocafone**).

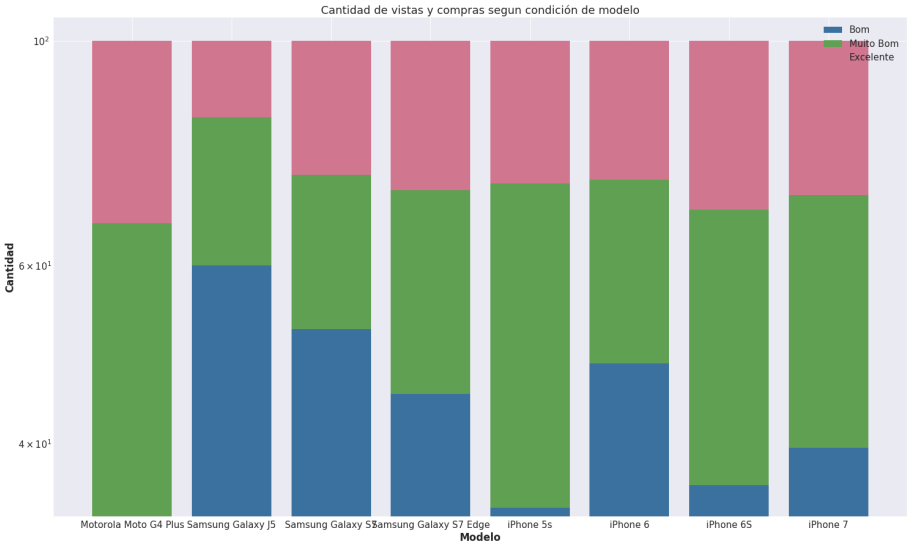

Figura 19: Eventos en modelos según condición

6.3. Colores de dispositivos

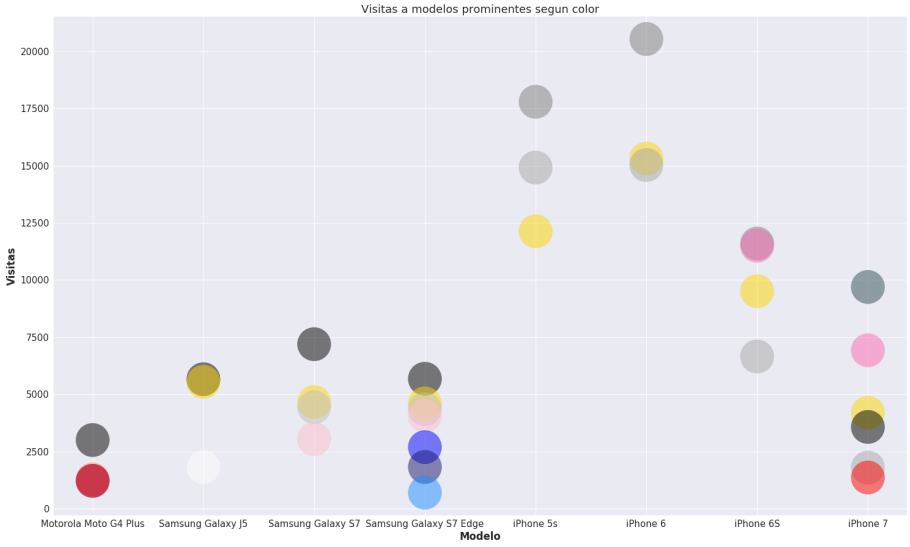

Figura 20: Colores con más tráfico

Para concluir la sección se muestran en el gráfico 20 los colores de los celulares analizados previamente, solo a modo ilustrativo ya que poco se puede extraer y co-relacionar de algo tan arbitrario y subjetivo como la elección de un color de un celular a comprar (aunque si se nota un parcial desbalanceo (*skewness*) hacia los celulares con tintes grises y/o plateados).

7. Análisis de páginas estáticas

Se propone comparar la cantidad de visitas al `FAQ`[3] con las de `Customer Service`, para poder analizar la forma que tiene **Trocafone** de brindar soporte a sus usuarios.

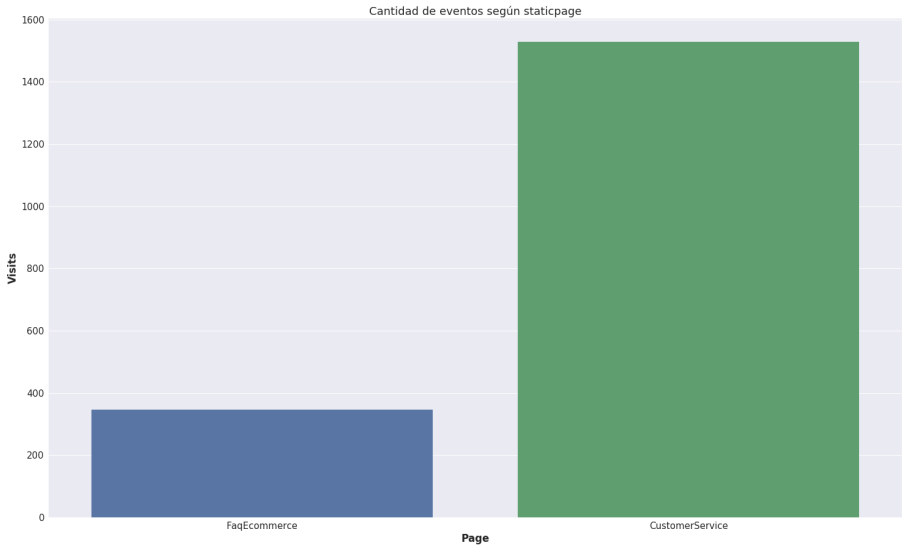

Figura 21: Comparación entre cantidad de visitas al FAQ y a Customer Service

Del gráfico 21 se ve que la cantidad de visitas a `Customer Service` es mucho mayor que la cantidad de visitas al `FAQ`. Para mantener la página de `Customer Service` es necesario disponer de empleados constantemente para responder las consultas requeridas. Por lo tanto, se podría optimizar recursos redireccionando parte del tráfico a FAQ, haciendo más visibles los links a la página, agregando contenido común y mejorándola de ser necesario.

[3]Frequently Asked Questions: lista de preguntas y respuestas que surgen comúnmente en un contexto determinado.

8. Análisis de nuevos usuarios vs usuarios que regresan al sitio

En esta sección se busca determinar la proporción de usuarios del sitio que entraron una sola vez a la página y no volvieron a hacerlo. Para ello se grafica en un primer lugar la cantidad de usuarios calificados como `New` contra los que son calificados como `Returning`.

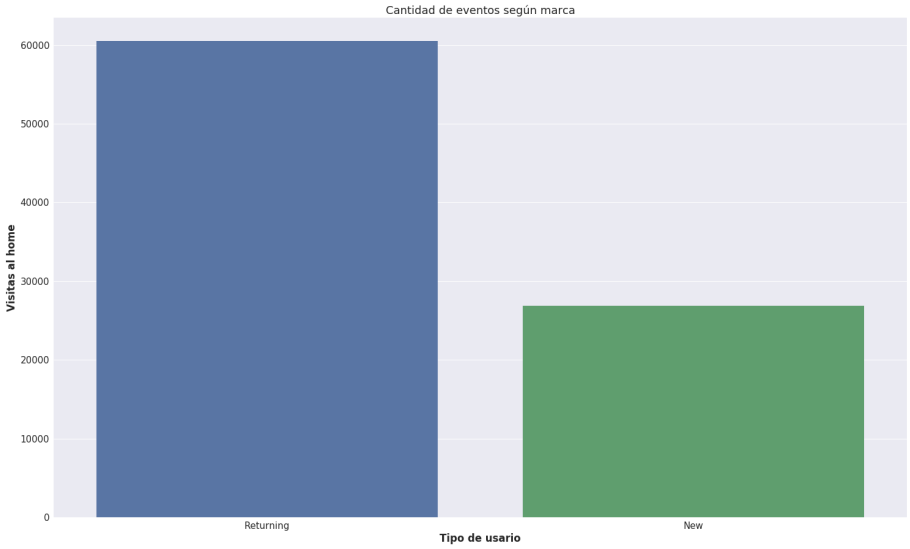

Figura 22: Comparación entre cantidad de usuarios que entran por primera vez al sitio contra los que volvieron

Es necesario remarcar que el gráfico 22 no es representativo porque todos los usuarios calificados como `returning` en algún momento fueron registrados como `new` (su primera vez ingresando al sitio). A simple vista se podría concluir que la proporción de usuarios que regresa es mucho mayor a los que entran solo una vez.

Se realiza el recorte necesario para obtener una visualización que refleje fiablemente la cantidad de visitantes que entra al sitio una sola vez contra los que vuelven otras veces.

Se observa en el gráfico 23 que la tasa de personas que entra una sola vez es mayor a la de las que regresan. Esta información desfavorece a Trocafone ya que implica que pierde una gran cantidad de clientes[4]. Para aumentar la tasa de personas que regresan a la página se puede proponer aumentar el presupuesto en publicidad y mejorar la experiencia de usuario de la home para que provea al usuario una experiencia más amena. También podrían ampliarse los métodos

[4]Nuevamente se recuerda que estamos hablando del subconjunto de usuarios que realiza al menos un `checkout`, un subconjunto que se asume mucho menor que el global de los usuarios del sitio.

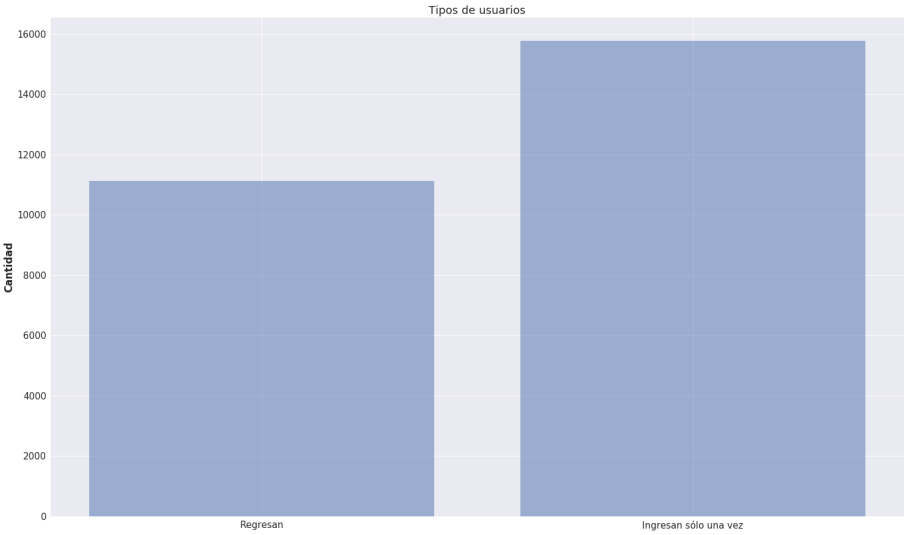

Figura 23: Comparación entre cantidad de usuarios que entran por primera vez al sitio contra los que volvieron

de pago o mejorar la página de `Customer Service` para que el cliente se sienta más contenido y pueda resolver todos los conflictos existentes ante una compra.

9. Análisis de marcas

Se busca determinar qué marcas son las que reúnen la mayor cantidad de eventos. Esto puede ser ya sea porque son las marcas más compradas, más buscadas o más vistas, entre otros eventos. Este análisis es más bien global ya que no es específico a un evento determinado.

Se conserva el comportamiento presentado en la sección 5 con respecto a los términos más buscados. Las marcas que registran mayor cantidad de eventos son **iPhone**, **Motorola** y **Samsung**.

Ahora se busca analizar cuáles marcas son las que realizan una cantidad pareja de checkouts y conversiones. Siguiendo la linea de razonamiento en la sección 6 se predice que las marcas que venden celulares a precios elevados como **Apple** van a mostrar una cantidad mayor de checkouts que conversiones. Así mismo, las marcas que mantienen un precio más accesible van a tener ambas cantidades más parejas.

Se observa en el gráfico 25 que la relación checkout-conversions para las marcas representadas es relativamente constante para todas las marcas, con algunas excepciones marcadas donde la cantidad de conversiones es muy chica, como **Asus**, **iPad** (técnicamente **Apple**) o **Quantum**. Igualmente podría afirmarse que la predicción es cierta porque esta relación es mayor en la marca **Samsung** que en la marca **Apple**.

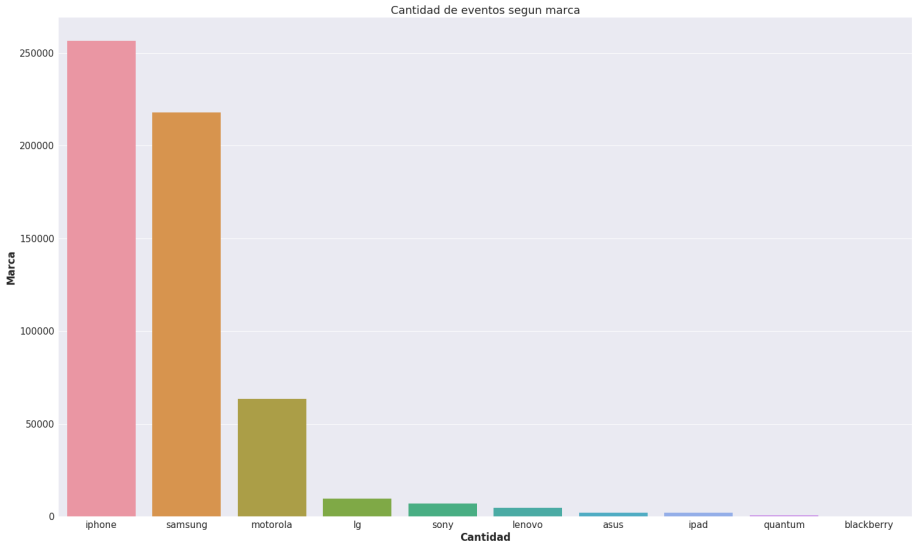

Figura 24: Cantidad de eventos según marca

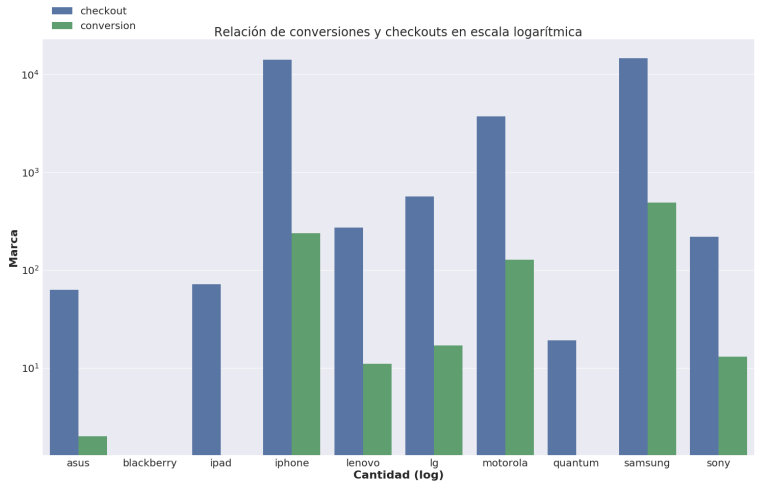

Figura 25: Checkouts vs conversiones según marca en éscala logarítmica

10. Análisis de tipos de dispositivos

Se busca analizar en esta sección desde qué tipo de dispositivos suelen acceder los clientes a la página de Trocafone.

Se puede observar en el gráfico 26 que casi todos los eventos se registran desde

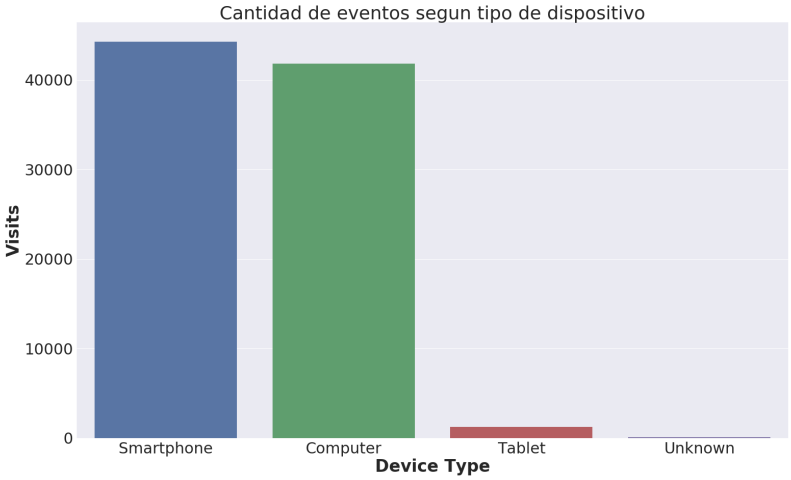

Figura 26: Cantidad de eventos de acuerdo al dispositivo utilizado

un smartphone o una computadora. La cantidad de eventos registrada desde la tablet es significativamente menor. Por lo tanto, se considera que podría dedicarse una mayor cantidad de recursos a desarrollar la aplicación para smartphones y computadoras y no dedicar mucho tiempo y desarrolladores a las aplicaciones para tablets, o bien optar por el camino contrario e intentar hacer más atractiva la manera de ingresar desde la tablet con el fin de atraer nuevos usuarios de ese target en particular, aunque esto implique un mucho mayor costo (se asume que es más dificil y costoso desarrollar algo no exitoso desde el suelo que mantener y mejorar algo con una cantidad constante de visitas).

Lo siguiente a analizar es la fidelidad de los usuarios a sus dispositivos actuales y respectivas marcas y su adversidad al cambio. Implicitamente, lo que también se analiza acá es la efectividad del *branding* de las marcas, ya que mantener consumidores es uno de sus principales objetivos como empresa y se logra con buenas campañas de marketing.

Para poner un ejemplo de fidelidad, un usuario fiel sería quien entra desde una computadora **Mac** y solo busca para comprarse el **iPhone** vigente, mientras que un usuario que entra desde **Android** y decide comprar un **iPhone** es uno que no es fiel a su marca/sistema operativo actual.

Notamos en el gráfico 27 varias cosas. Por empezar, dada la naturaleza de *branding* y marketing de la empresa, los usuarios de **Apple** no son muy propensos al cambio y se mantienen en su mayoría sobre su mismo sistema. Por otro lado, los usuarios de **Android** están uniformemente divididos y no se puede concluir que no sean fieles ni infieles.

Luego, el caso de **Windows** sería trivial de analizar, siendo este un sistema operativo de computadoras sin ninguna asociación en particular a marcas de celulares (ignorando el cuasi extinto **Windows Phone**, que aparecería como la marca **Nokia** o **Microsoft**). Y como último es interesante ver los usuarios de **Linux**, que por la esencia abierta del sistema operativo y su comunidad, es peculiar ver la distribución uniforme (si bien sobre pocos datos) tanto hacia

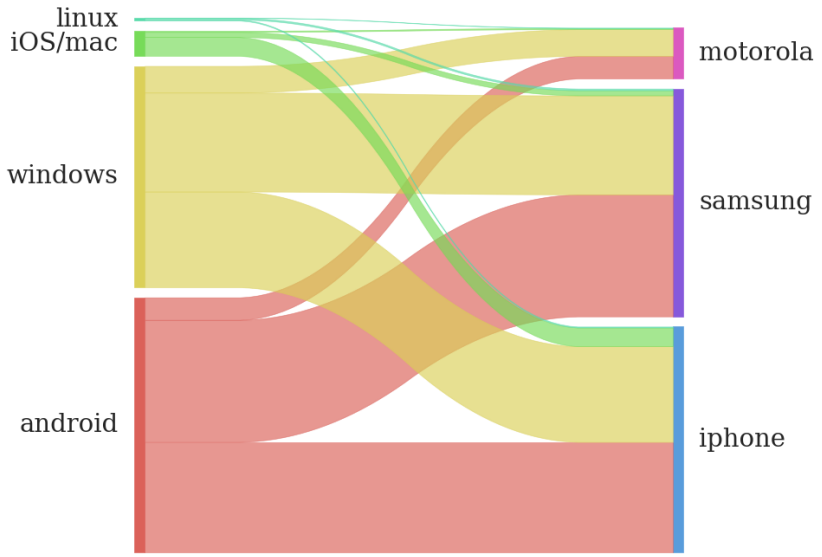

Figura 27: Diagrama de flujo de fidelidad de usuarios

celulares **Android** como **Apple**, aún más teniendo en cuenta la fuerte tendencia de este último a ser un sistema más cerrado.

11. Análisis de publicidad

La primera pregunta que fue planteada fue la de investigar si **Trocafone** aumentó el presupuesto en publicidad en algún período de tiempo. Para ello se procede a analizar en qué meses aumentó la cantidad de visitas originadas de una campaña de marketing, información proveída por la columna `campaign source`.

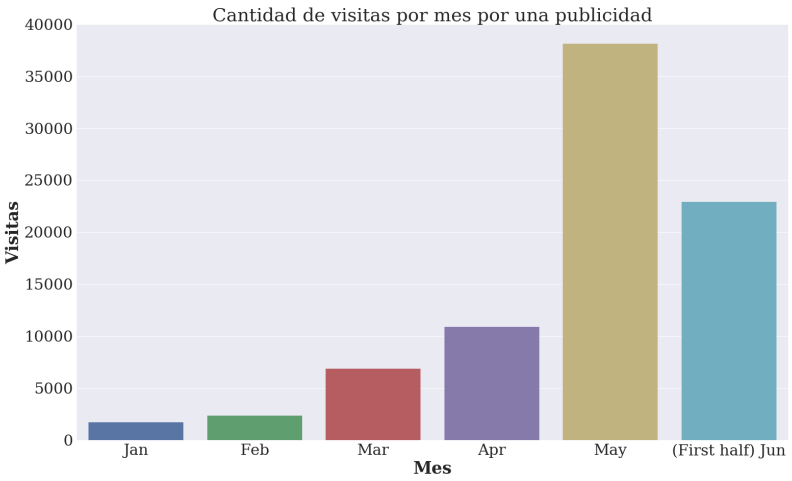

Figura 28: Cantidad de visitas provenientes de una campaña de publicidad de acuerdo al mes

La observación del gráfico 28 no permite extraer ninguna conclusión válida. Esto se debe a que el mes en el que se registra mayor cantidad de visitas provenientes de una campaña de publicidad (mayo) es el mes en el que se detectó mayor cantidad de eventos. Sería pensamiento circular decir que hay más visitas este mes porque aumento el presupuesto publicitario si nuestro único fundamento sobre esto esta basado en las visitas del mes. Es lógico que si aumentan las visitas en una página, aumente en consecuencia el número de visitas que proviene de publicidad.

Se procede a analizar cuáles son los metodos de publicidad más usados. Se puede predecir que será **Google** debido a su importancia mundial como motor de búsqueda. Una vez verificado esto, se elimina del gráfico para observar los otros métodos de publicidad contratados.

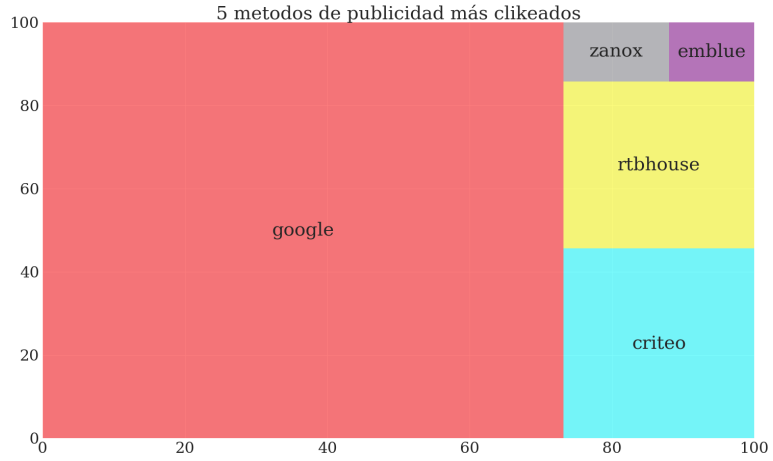

Figura 29: Métodos de publicidad que generan visitas en Trocafone

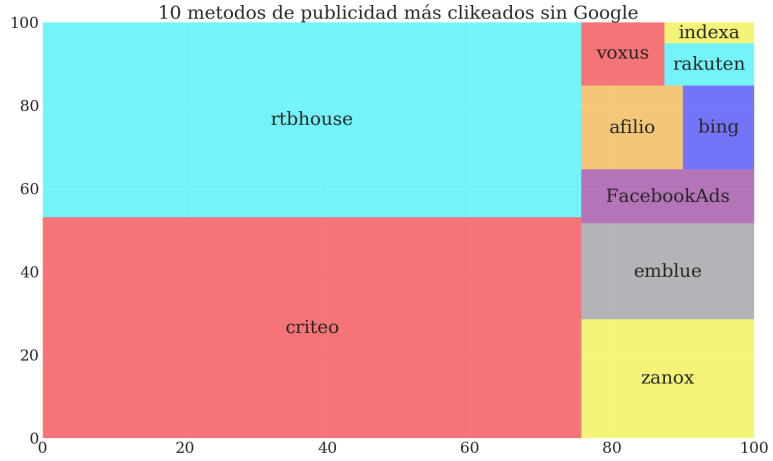

Figura 30: Cantidad de eventos de acuerdo al dispositivo utilizado

Se concluye de estos gráficos 29 y 30 que el método de publicidad más eficiente es **Google**. Por lo tanto, es conveniente que **Trocafone** mantenga su contrato con el mismo para aumentar las visitas a su página. Los otros métodos de publicidad son efectivamente mucho menores y parecen no tener mucha relevancia para el tráfico del sitio. Es por esto que quizás sería necesario que **Trocafone** haga un balance entre los gastos consumidos y la ganancia obtenida con el uso de esos métodos publicitarios.

11.1. Funnel por publicidad

Habiendo obtenido las sesiones por usuario, se puede observar de donde como se inicia cada una. Para esto se generó un *Funnel* que no solo indica la cantidad de tráfico por cada paso (en escala logarítmica), sino que también muestra la proporción del mismo que proviene de una campaña publicitaria.

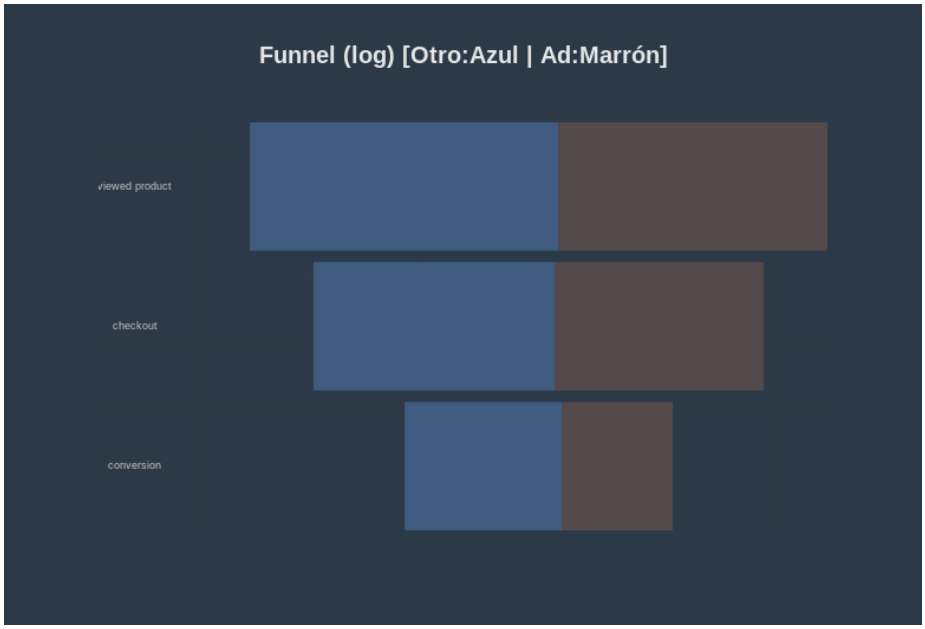

Figura 31: Funnel según provengan de campaña publicitaria

El gráfico 31 muestra que si bien casi la mitad del tráfico proviene de *advertising* tanto para *viewed product* como para *checkout*, el ratio cae mucho en la etapa de *conversion*, donde menos de el 15 % del tráfico de esta fuente termina en una compra.

12. Análisis de canales de tráfico

Sabiendo que un usuario puede llegar desde diversos lugares, ver de dónde viene es algo muy importante para saber en que lugar debe la empresa invertir. ¿Debe tener mayor presencia en las redes sociales? ¿Debe invertir en su visibilidad online mediante procesos de SEO[5]?

Lo que en particular nos interesa es la métrica llamada **revenue by traffic source**, la cual se refiere no a los usuarios que vienen, si no los usuarios que vienen y hacen una conversion (compra), como una manera de representar los métodos realmente efectivos (después de todo, la efectividad de la empresa radica en que un usuario haga una conversión, no meramente en que entre al sitio).

Los canales de tráfico presentes en el dataset son:

[5]Search engine optimization

- **Paid:** Usuarios que llegan mediante una campaña de marketing.
- **Direct:** Usuarios que llegan directamente al sitio, sin ayuda externa (por ejemplo, escribiendo directamente la url en el explorador web, un marcador, un link de un documento sin tracking)
- **Email:** Usuarios que llegan desde un link en un email.
- **Organic:** Usuarios que llegan desde motores de búsqueda.
- **Referral:** Usuarios que llegan al sitio desde otro sitio web.
- **Social:** Usuarios que llegan desde redes sociales.

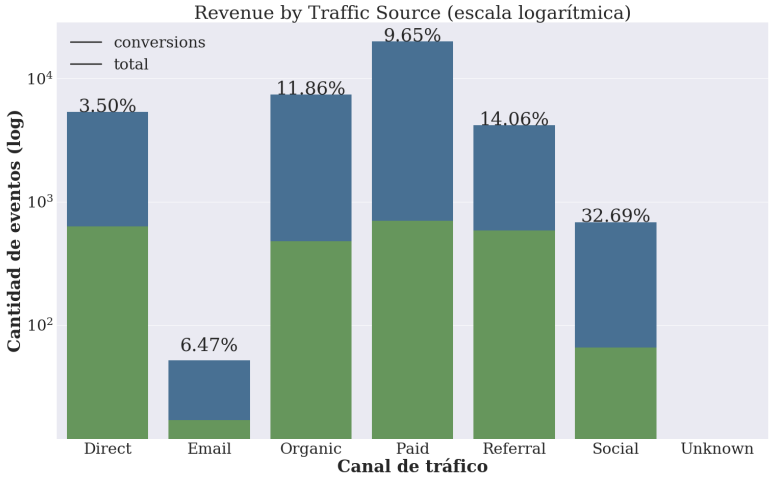

Figura 32: Canales de tráfico y sus ganancias

Lo que observamos en el gráfico 32 es que si bien el canal pago es el que más usuarios atrae, tiene una ganancia bastante razonable para la magnitud de usuarios que atrae. Por lo tanto seguir invirtiendo en él probablemente lleve a un retorno con incremento constante y no tan significante[6].

Por otro lado son otros canales los que verdaderamente sorprenden por su porcentaje de retornos. Hay un gran retorno de parte de la presencia de las redes sociales, que puede no atraer tantos usuarios pero a los que atrae lo hace de manera muy efectiva, y lo mismo aplica para la presencia onlina organica de Trocafone, donde se ve que si se invirtiese en procesos de SEO esto podría llevar tanto a más usuarios como más compras.

13. Insights y conclusiones

Antes de presentar las conclusiones obtenidas a partir del análisis del dataset presentado es necesario volver a remarcar que los datos no representan

[6]Como ejemplo de la ley de los rendimientos decrecientes

el total de datos de la empresa **Trocafone**. Se descubrió que los únicos datos presentados eran de clientes que habían realizado al menos un `checkout`, por lo que mínimamente se detecta un truncamiento allí. Cabe la posibilidad de que se haya truncado con algún otro criterio que no se haya detectado con el análisis presentado. Además, los datos corresponden a la página web de Trocafone, por lo que no constituye el total de datos de ventas de la empresa, que también se encargan de vender a otros proveedores.

Este recordatorio tan reiterado no deja de ser sumamente importante, de tener más datos (como por ejemplo un semestre en su totalidad, sin truncamientos) se podría hacer un análisis más a pleno, ni hablar de si se tuviesen más atributos, ya que una de las dificultades encontradas en el trabajo fue la falta de atributos numéricos, dependiendo solamente de contar valores (por ejemplo, un dato sumamente rico en información es el precio de venta, o poder decir más del usuario como su edad y/o sexo).

Ahora sí, en primer lugar se considera que de la cantidad de conversiones detectada en el set de datos no puede extraerse alguna conclusión en cuanto a cuan bien o mal le fue a la empresa en el semestre, por falta de conocimiento y manejo del dominio (plataformas de e-commerce en el rubro de celulares). Otro lugar donde se ve la falta de dominio es en el no poder explicar la cantidad de registros de mayo (considerablemente mayor al resto de los meses). De tener mayor conocimiento del estado de Brasil (puede ser producto de política), de la industría de la tecnólogia celular (constantemente hay nuevos lanzamientos de celulares que incentivan vender el previamente usado, leyes que influyen tanto para bien como para mal la compra/venta y demás) o de **Trocafone** en sí (por ejemplo, un lanzamiento de descuentos o promociones) se tendrían más recursos para poder explicar lo que a simple vista parece una anomalía. En cambio, si se ha podido analizar en su lugar la comparación *checkout vs conversiones vs viewed products* a lo largo de diferentes parámetros como marcas de celulares o períodos de tiempo, mostrando como el sitio web cumple o no su objetivo de mantener a los usuarios y hacerlos comprar por allí.

Se ha podido analizar en su lugar la comparación *checkout vs conversiones vs viewed products* a lo largo de diferentes parámetros como marcas de celulares o períodos de tiempo. Del análisis de la cantidad de eventos mencionados previamente en función de las marcas de celulares se concluye que las marcas que tienen un precio más accesible como lo puede ser **Samsung** eran de las menos vistas pero de las más compradas. Asímismo, las marcas con un precio más elevado como **Apple** eran de las más vistas pero de las menos compradas. Esta era una situación prevista de antemano debido a que la calidad del producto de mayor precio atrae visitas pero su precio aleja las compras. Asímismo, un celular de menor precio y calidad es más probable que si es visto sea comprado por el mismo usuario debido a que la calidad del producto no suele atraer visitas que no sean potenciales compradoras en un futuro.

También se observa que no tiene sentido seguir comerciando algunos productos, por ejemplo los **Blackberry**, **Quantum** y **iPad**, ya que se necesita disponer de expertos para repararlos que no son justificados por la cantidad de conversiones que tienen.

Un análisis similar puede realizarse en cuanto al sistema operativo del cual proviene el que genera una compra y el sistema operativo del dispositivo comprado. Los usuarios que provienen de **Apple** suelen desear comprar un dispositivo del mismo sistema operativo debido a su calidad y al atractivo que suelen ma-

nejar los productos de **Apple** en la sociedad. De otra manera, los usuarios que provienen de **Android** divergen sus visitas en cuanto al sistema operativo ya que no se detecta un comportamiento específico para este tipo de usuarios.

Del análisis de la cantidad de eventos mencionados previamente en función de las marcas de celulares se concluye que las marcas que tienen una selección de celulares más amplia, en particular yendo desde más accesibles a celulares de alta gama como lo puede ser **Samsung** son de las menos vistas pero de las más compradas. Asímismo, las marcas con un precio más elevado como **Apple** son de las más vistas pero de las menos compradas. Esta era una situación prevista de antemano debido a que la calidad del producto de mayor precio atrae visitas pero su precio aleja las compras. Asímismo, un celular de menor precio y calidad es más probable que si es visto sea comprado por el mismo usuario debido a que la calidad del producto no suele atraer visitas que no sean potenciales compradoras en un futuro.

Otra conclusión que se despliega del análisis de los eventos en función del tiempo radica en detectar los momentos de mayor y menor tráfico. En horarios como las 19 hs, la hora del día que se registran más checkouts y conversiones), se podría agregar publicidad que incentive el aumento de tráfico y promueva la realización de conversiones. Así mismo, los días que se registra menos tráfico como los fines de semana o el horario del mediodía podría agregarse alguna promoción para aumentar la cantidad de eventos en dichos horarios y expandir la franja horaria de tráfico en la página.

En lo que respecta a páginas estáticas, se llegó a la conclusión de que una buena inversión de parte de la empresa sería optimizar el sitio de preguntas frecuentes, ya que se notó que hay una gran cantidad de tráfico a *customer service* que podría redirigirse a las *FAQ* para optimizar el uso del personal en responder en atencion al cliente y que los usuarios encuentren rápidamente respuestas a sus preguntas específicas.

Se observó que la taza de usuarios que ingresa solo una vez es mucho mayor que la de usuarios que regresan. Si bien este es un problema con el que todo sitio se encuentra, una forma de disminuir lo más posible la diferencia puede ser mejorar la experiencia de usuario en el sitio de entrada.

Una sugerencia en cuanto a inversión en presencia online esta tanto en los motores de busqueda como en las redes sociales. Ambos canales de tráfico tienen un potencial muy grande en cuanto a atraer usuarios que efectuan conversiones, e invertir en ellos lograría mantener el porcentaje de efectividad mientras se incrementa el número total de usuarios atraídos.

Por otro lado, podría aprovecharse el conocimiento sobre los productos más visitados para encarar con su imagen las publicidades, logrando así nuevas visitas a la página. Con ellas, nuevas posibilidades de ventas. Una vez que el cliente se encuentre dentro de la plataforma virtual se puede utilizar el conocimiento sobre los productos más vendidos y dar fácil acceso a su publicación. Teniendo en cuenta que los termino más buscados son busquedas muy genéricas (marcas) se puede implementar la sugerencia ubicando como primera opción los modelos que más se compran y así mejorar el texto predictivo de la caja de busqueda del sitio web.Por otro lado, podría aprovecharse el conocimiento sobre los productos más visitados para encarar con su imagen las publicidades, logrando así nuevas visitas a la página. Con ellas, nuevas posibilidades de ventas. Una vez que el cliente se encuentre dentro de la plataforma virtual se puede utilizar el conocimiento sobre los productos más vendidos y dar fácil acceso a su publicación. Teniendo

en cuenta que los termino más buscados son busquedas muy genéricas (marcas) se puede implementar la sugerencia ubicando como primera opción los modelos que más se compran y así mejorar el texto predictivo de la caja de busqueda del sitio web.

Teniendo en cuenta que casi la mitad de las sesiones se generan debido a la publicidad, pero que sólo el 12.5 % de las sesiones que terminan en una conversión provienen de esta, se tiene que considerar que debería mejorarse la calidad de la misma, mostrándole al usuario un producto más acorde a su perfil, por ejemplo, a un usuario de iOS se le ofrecería un iPhone como publicidad, mejorando así las chances de conversión.

A. Ejecución

El trabajo fue realizado en Anaconda[7]. Para poder replicar el trabajo, hay que también instalar las siguientes librerías adicionales:

- Squarify[8]: Para los treemaps.
- pySankey[9]: Para los diagramas de Sankey (diagramas de flujo).
- Geopandas[10]: Para poder graficar sobre mapas geográficos.
- Wordcloud[11]: Para poder visualizar los términos más buscados.

Estos pueden ser instalados con los siguientes comandos:

```
pip install squarify
pip install pySaneky
pip install plotly --upgrade
conda install -c conda-forge geopandas
conda install -c conda-forge wordcloud
```

Alternativamente, se puede optar por correr el código sobre el kernel publicado en la plataforma Kaggle.

B. Datasets adicionales incoporados para el análisis

Se utiliza adicionalmente un dataset del mapa de Brasil y sus ciudades, para el análisis geográfico. Este fue sacado de Geonames[12], una base de datos publica de paises y regiones del mundo.

[7] https://anaconda.org/
[8] https://github.com/laserson/squarify
[9] https://github.com/anazalea/pySankey/
[10] http://geopandas.org/
[11] https://github.com/amueller/word_cloud/
[12] http://www.geonames.org/