FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Trabajo Práctico Nro. 1
Analisis Exploratorio

Eventos Trocafone

[7542] Organizacion de Datos
Segundo cuatrimestre de 2018

Alumno:	**Padrón**
ALVAREZ JULIA, Santiago	99522
CARRERO RIVEROS, Maria Daniela	99316
CANAVESE, Bautista	99714
PELOZO, Emanuel	99444

Índice

1. Introducción

Trocafone es una empresa de **ReCommerce**, pues su negocio está basado en la compra, reacondicionamiento y venta de productos usados. En el presente trabajo se realiza un análisis exploratorio de los datos de *Trocafone*, basados en eventos realizados por usuarios que visitaron su plataforma web y la información sobre cada uno de estos eventos.

En primer lugar, debemos definir los posibles eventos ocurridos en la web de *Trocafone* para tener una referencia al momento de realizar el análisis. Entre ellos están:

- **viewed product**: El usuario visita una página de algún 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 campaña de marketing online
- **generic listing**: El usuario visita la homepage.
- **searched products**: El usuario realiza una búsqueda de productos en la interfaz de búsqueda del site.
- **search engine hit**: El usuario ingresa al sitio mediante un motor de búsqueda web.
- **checkout**: El usuario ingresa al checkout de compra de un producto.
- **staticpage**: El usuario visita una página.
- **conversion**: El usuario realiza una conversión, 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.

2. Análisis Exploratorio

2.1. Análisis Previo

Como un primer análisis se da un 'vistazo' al conjunto de datos. Éste contiene 1.011.288 registros y 23 atributos. Los datos no están ordenados cronológicamente (es decir según columna 'timestamp'), pero se obtienen datos desde 2018-01-01 07:32:26 hasta 2018-06-15 23:59:31.

La frecuencia de los eventos está detallada en la Figura 1, donde es de esperarse que el evento con mayor frecuencia es la vista del producto. Generalmente los usuarios tienden a ver varios productos una vez que entran al site, ya sea mediante una búsqueda, mediante publicidad, mediante el menú de marcas, etc.

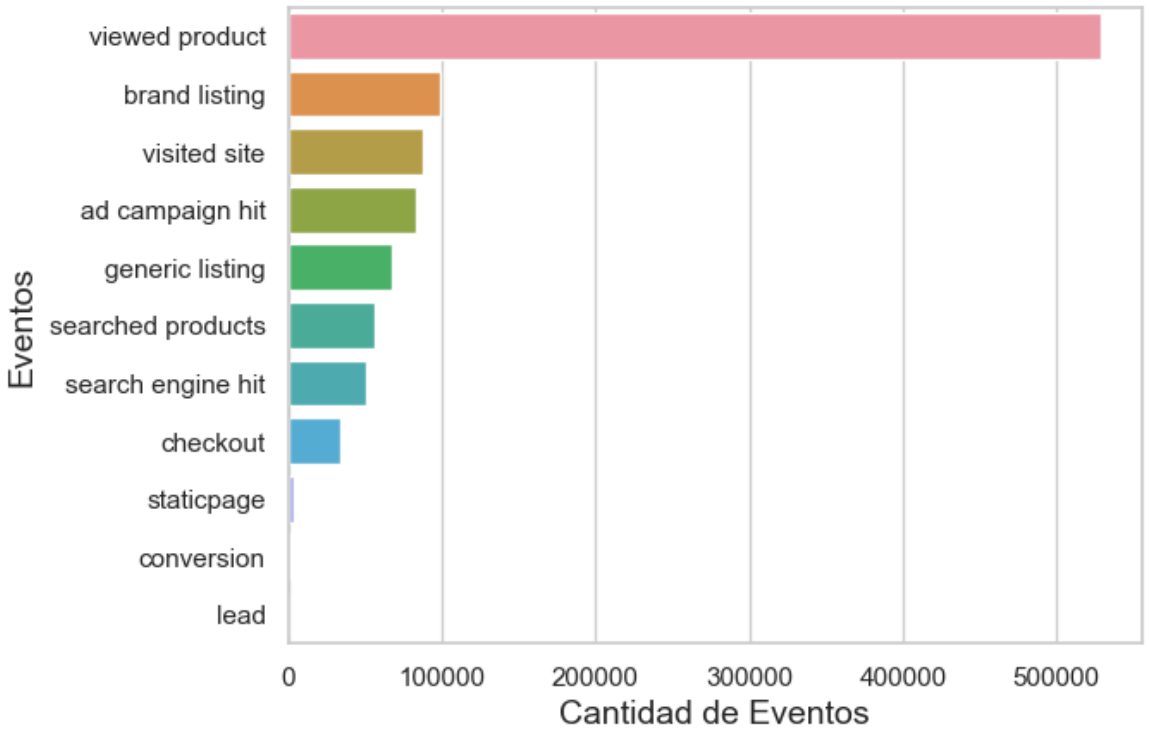

Figura 1: Frecuencia de los distintos eventos en el conjunto de datos.

La empresa puede tener un registro de usuarios registrados que, por cuestiones de seguridad, no tenemos el detalle. Pero, a través del ID, podemos obtener el detalle de las personas que se incluyen en este análisis. Por lo tanto, la Figura 2 muestra las personas con mayor cantidad de eventos en el set de datos. De esta manera, la empresa podría saber cuales son las personas con mayor movimiento en la plataforma y posiblemente sus mejores clientes.

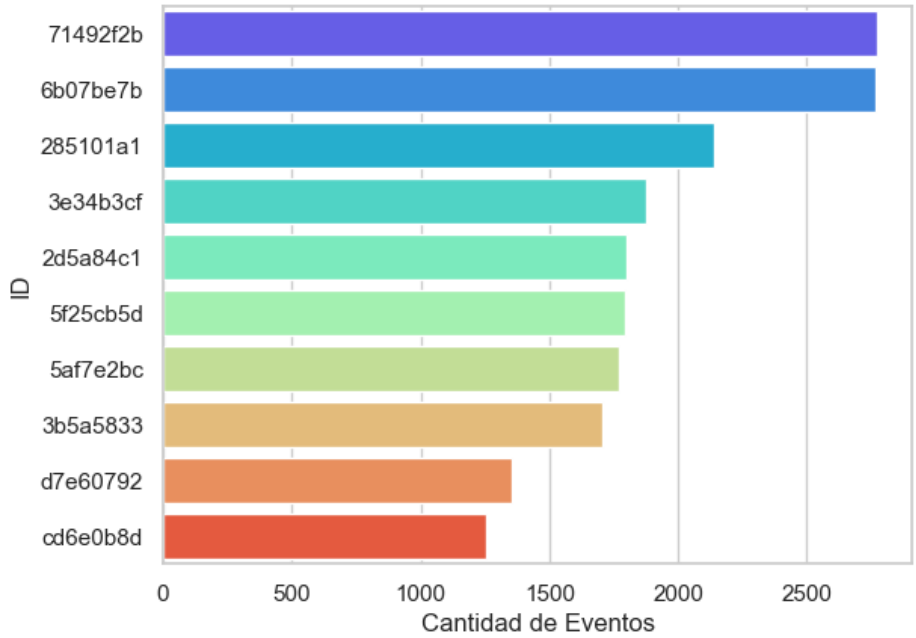

Figura 2: Las primeras 10 personas con mayor cantidad de eventos.

2.2. Análisis Temporal

Dentro del set de datos analizado, se encuentra un campo llamado 'timestamp' que nos indica el instante de tiempo en que ocurrió cada evento. Para facilitar el análisis agregamos al conjunto de datos varias columnas como el mes, día, hora.

Antes de comenzar con el análisis, observamos el rango de tiempo en el cual estan dispuestos los datos y pudimos ver que transcurren desde Enero hasta mediados de Junio del 2018. Por lo tanto, en algunos análisis prescindimos de los datos del último mes al ser incompletos.

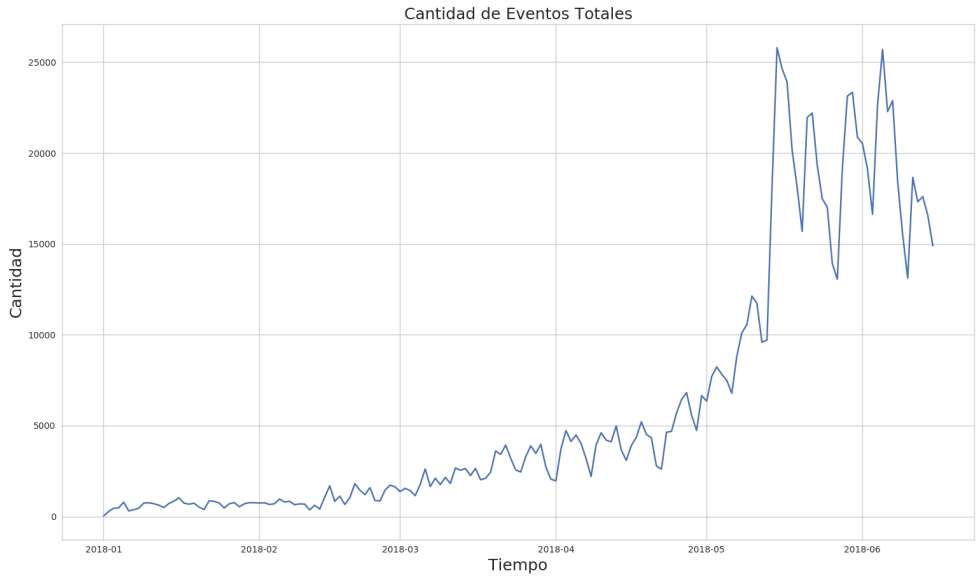

Figura 3: Cantidad de eventos en función del tiempo.

Un primer gráfico para comprender la distribución de los datos (Figura 3), vemos una aumento de interacciones en el tiempo, sobretodo en el último mes. Este gráfico se puede complementar con la Figura 4 donde se muestra la frecuencia de cada evento por mes. Claramente, el evento más frecuente es '*viewed product*', lo que puede ser dado porque los usuarios navegan por la página viendo múltiples productos, lo que hace este evento el más frecuente.

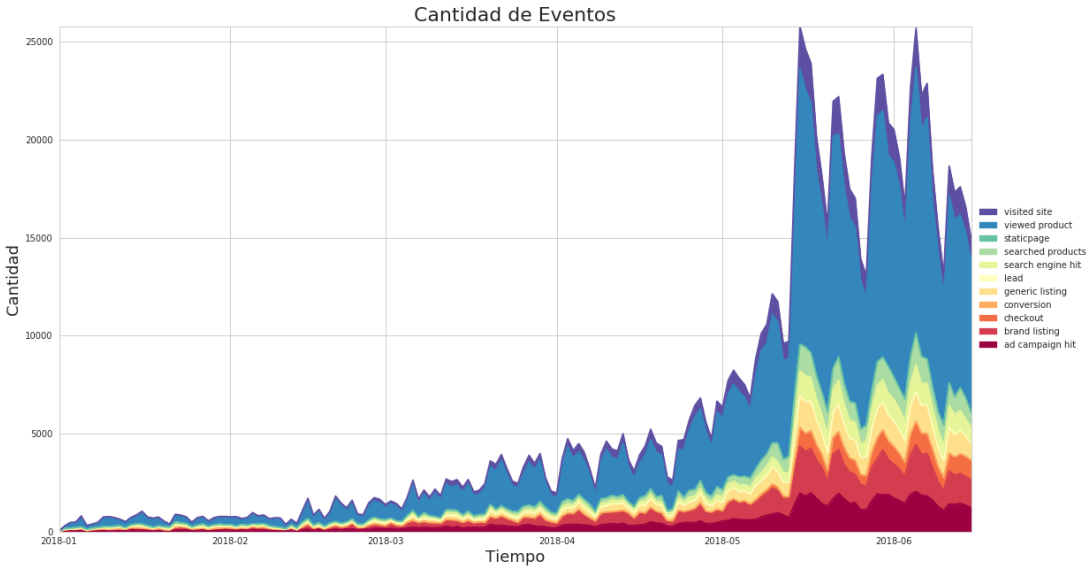

Figura 4: Frecuencia de cada evento por mes.

Por otro lado, también sería interesante evaluar en qué momento del día suceden estos eventos durante la semana. Observar en qué horario la página web tiene mayor tráfico de datos. El resultado es visible en la Figura 5, donde se observa que la frecuencia de estos eventos son más frecuentes entre la tarde y la noche, especialmente de lunes a viernes.

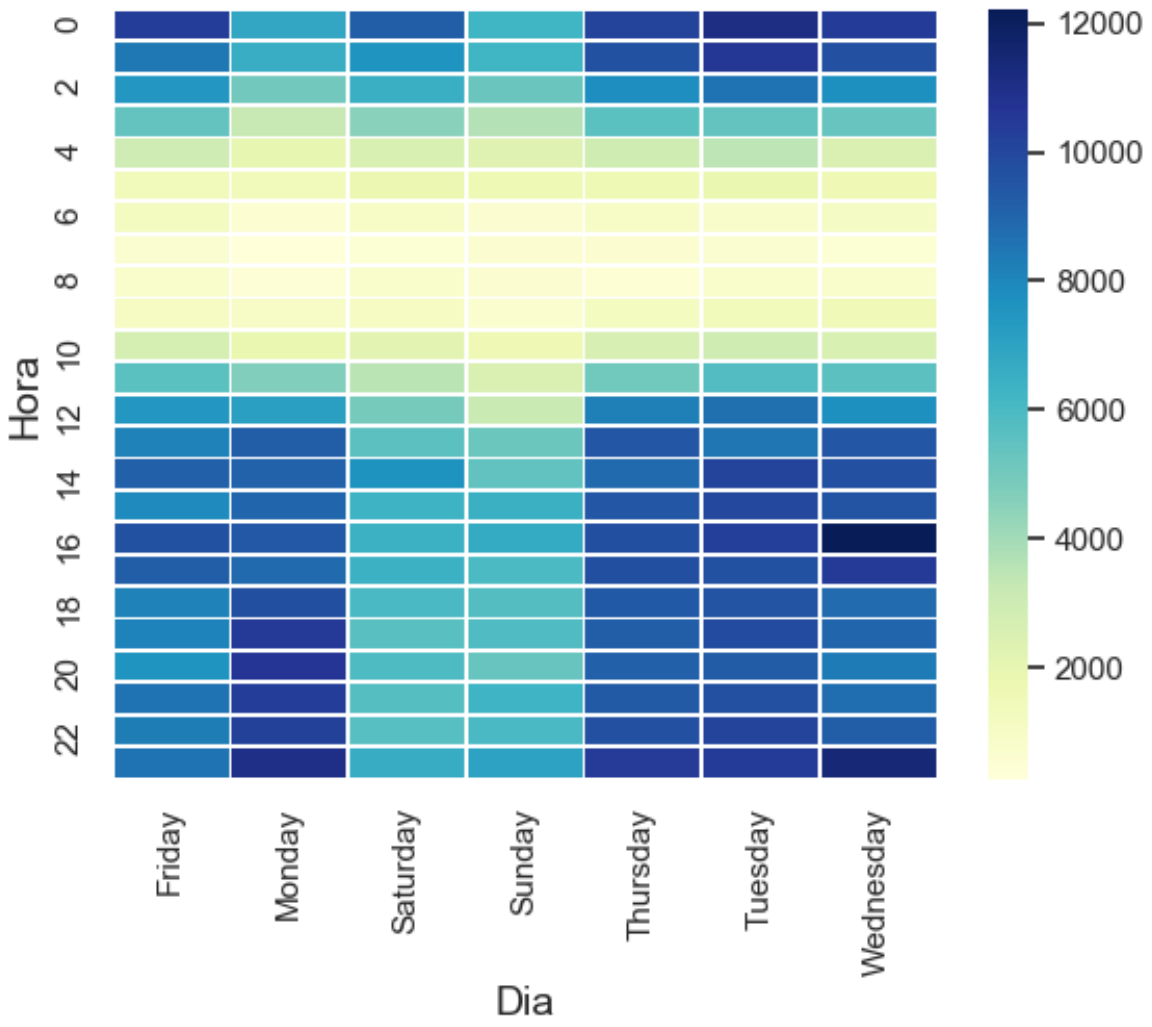

Figura 5: Cantidad de eventos en función de la hora y el día.

2.3. Análisis Geográfico

Es muy importante conocer el contexto sobre el cual *Trocafone* lleva a cabo su negocio. Distintos países, continentes, culturas, economía del país, etc. influyen a la hora de analizar a cualquier plataforma de **ReCommerce**.

Es tal la diferencia de eventos que suceden en un país por sobre el resto de los países, que en la Figura 6 se utilizó una escala logarítmica para graficarlo. El país que más utiliza la plataforma *Trocafone* es Brasil, muy por encima de USA y Argentina. Por lo tanto, a partir de aquí y hasta el final del informe, analizaremos el dataframe teniendo en cuenta este análisis geográfico.

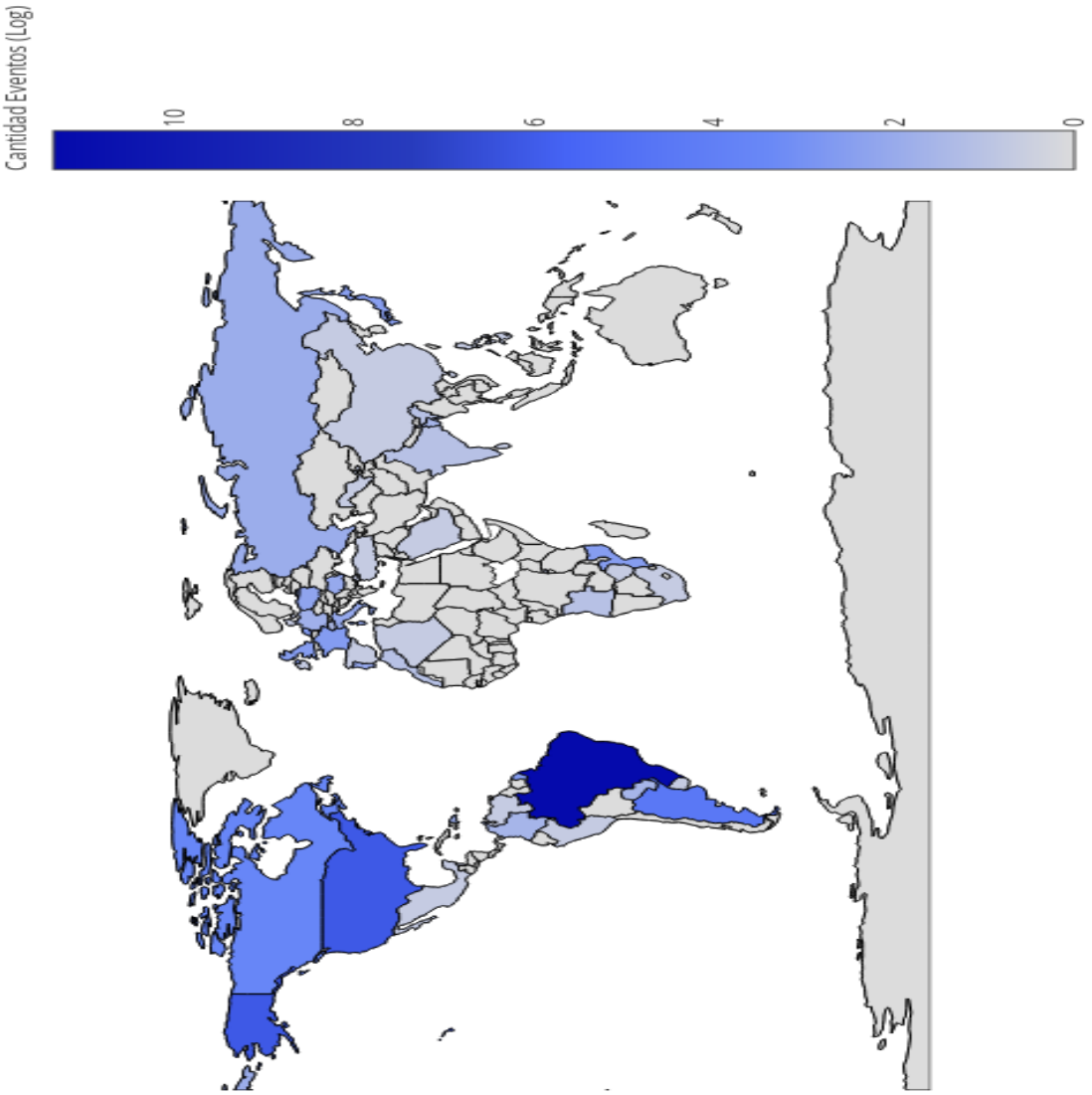

Figura 6: Cantidad de usuarios por país en escala logarítmica

2.4. Análisis según marca del dispositivo

En el conjunto de datos, el modelo del dispositivo con el cual se realizó un determinado evento es detallado. Realizando un análisis previo, se notó que la primera palabra detallada corresponde a la marca del dispositivo. De esta manera se pudo hacer un análisis de los determinados eventos según la marca.

En principio, es de importancia para cualquier negocio tener información sobre sus ventas, de tal manera que se pueda tomar decisiones para mejorarlas en un futuro. La Figura 7 muestra la cantidad de ventas por marca desde Enero a Mayo del 2018. En este se muestra que las tres marcas más vendidas son Samsung, iPhone y Motorola, en el orden indicado.

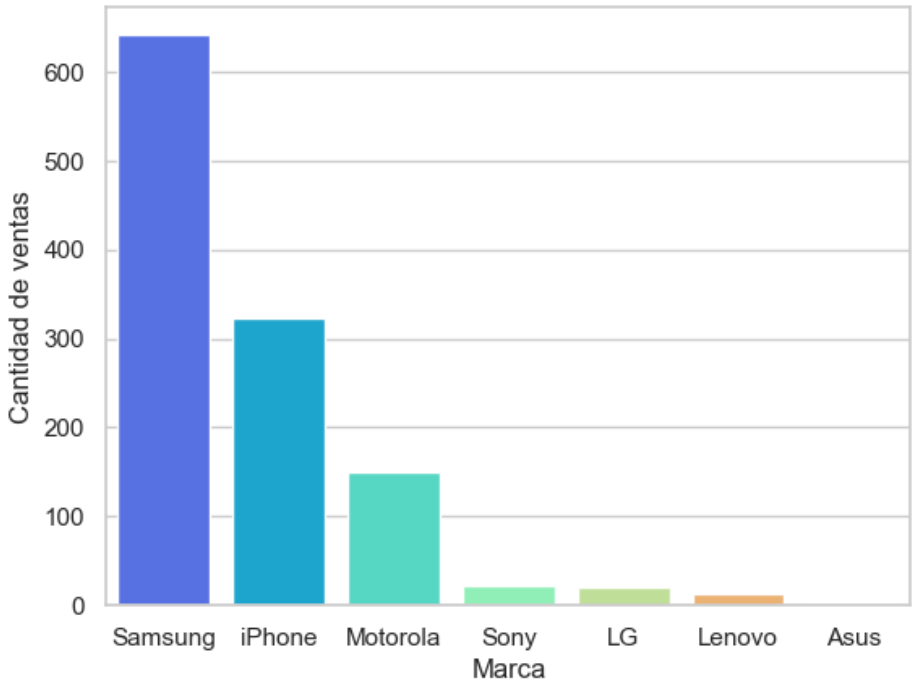

Figura 7: Cantidad de ventas por marca desde Enero a Mayo, 2018.

Por otro lado, puede ser de interés cuales son las marcas que con mayor frecuencia se visitan en la interfaz web de *Trocafone*. En la Figura 8 se puede observar que estas son Apple, Samsung y Motorola. Era de esperarse dado que estas son las marcas con mayor cantidad de ventas. Sin embargo, es notable que a pesar de que Samsung es la marca más vendida, Apple es la marca más visitada. Puede justificarse si se tiene en cuenta que los dispositivos de Apple suelen tener un precio elevado a comparación de otras marcas a raíz de que estan orientadas a mercados de primer mundo para los cuales proporciona servicios premium que no pueden disfrutarse en el mercado en el cual esta presente *Trocafone* (léase Brasil-Argentina). Dicha diferencia de precio puede ser la causa de la disparidad presentada entre las figuras 7 y 8.

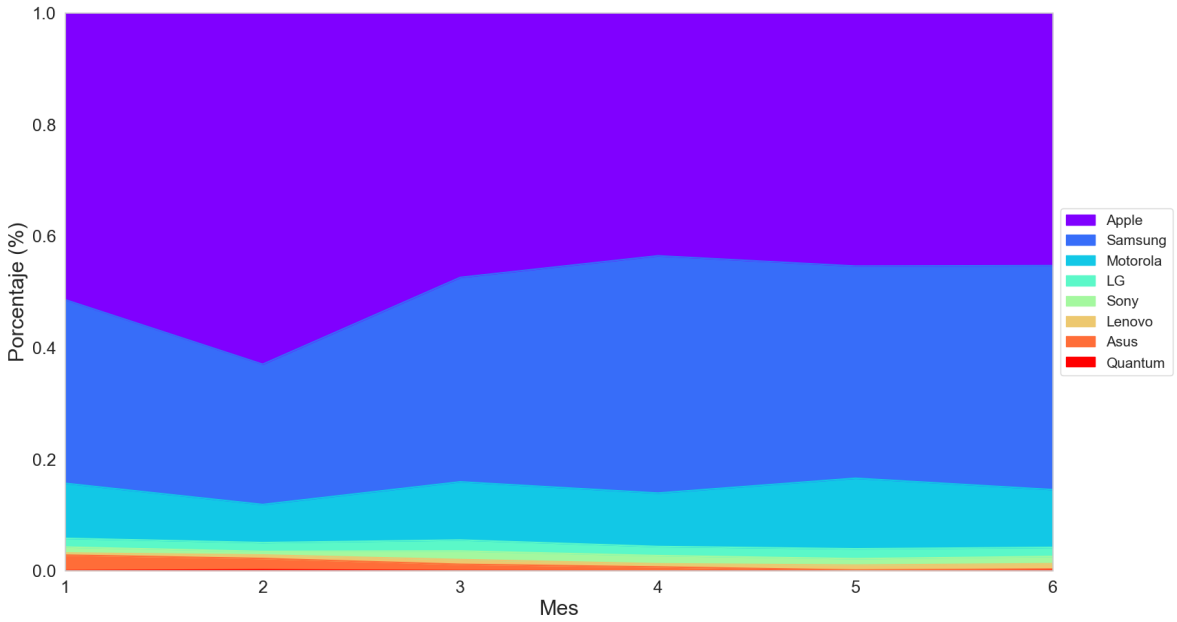

Figura 8: Top Marcas mas Visitadas por Mes.

2.5. Análisis de campañas de publicidad

A partir del evento 'Ad Campign Hit' y la columna 'campaign source' del dataframe proporcionado por *Trocafone*, hicimos los siguientes análisis.

Primero contamos la cantidad de hits de campañas de publicidad por hora (se tomaron en cuenta todos los dias presentes en el dataframe), dicho análisis se puede apreciar en la Figura 9. A simple vista se deduce que a la mañana temprano, desde las 4 hs. de la mañana hasta las 10 hs, es el intervalo de tiempo en el que menor cantidad de personas acceden a Trocafone mediante publicidad. Desde las 10 hs. de la mañana hasta las 13 hs. se producen incrementos considerables, tales que los hits de campañas de publicidad a las 13 hs. son el doble que a las 10 hs. Luego se mantiene estable, con algunos altibajos despreciables hasta las 23 hs, que se produce un incremento suficiente como para afirmar que las siguientes 2 horas las campañas de publicidad producen la mayor cantidad de hits teniendo en cuenta el día entero.

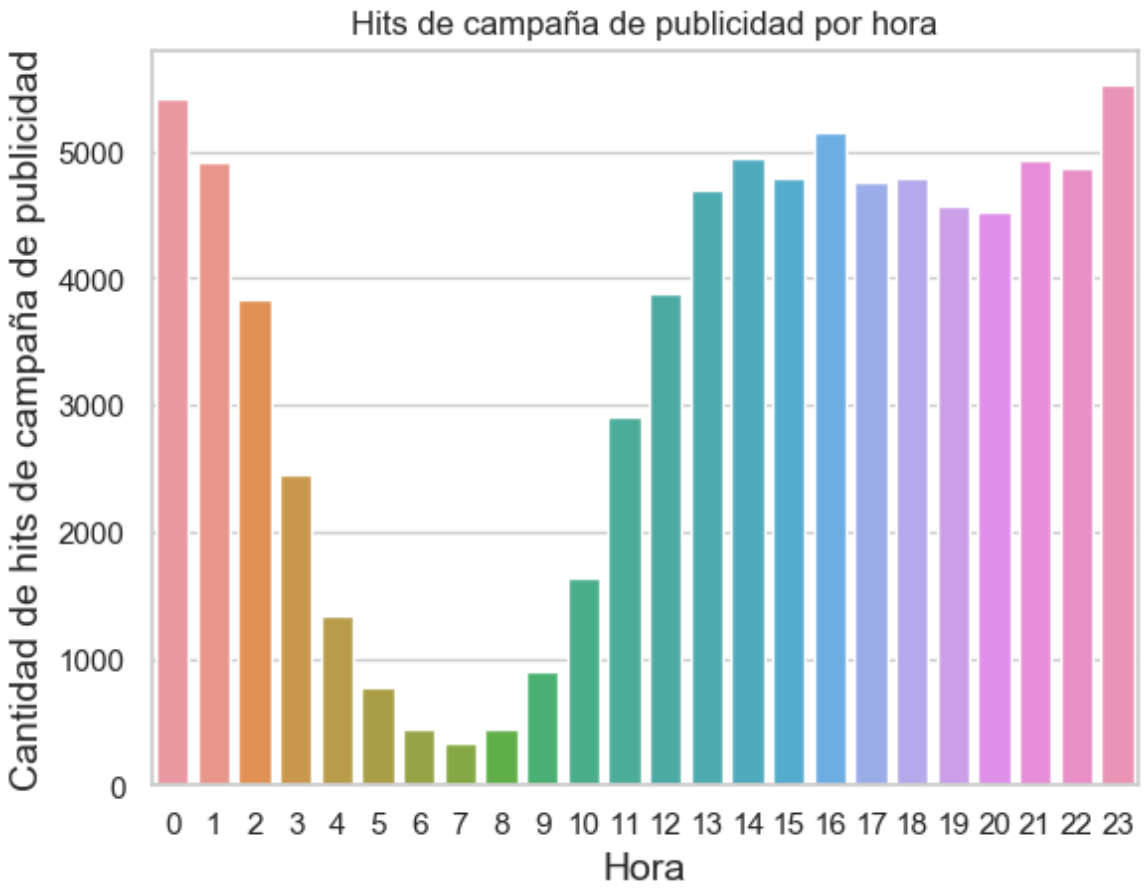

Figura 9: Hits de campaña de publicidad por hora.

Luego a partir de los datos de la columna 'campaign source', pudimos asignar cada hit de campaña de publicidad a una compañia de publicidad. A pesar de no tener información por parte del equipo de marketing de *Trocafone*, vamos a describir que decisiones creemos que tomaron para explicar los datos coleccionados.

En la Figura 10 es claro cual es el campaign source que le trae mayor rédito a Trocafone, las campañas de publicidad proporcionadas por la gigante tecnológica *Google*. Suponemos que también es el mayor gasto en publicidad por parte de *Trocafone*, justificado por el nombre e historia de *Google* y sus reconocidas y efectivas campañas de publicidad. Llama la atención la poca participación de *FacebookAds*, otra plataforma de publicidad proporcionada por una gigante tecnológica de renombre. *Trocafone* debe tener sus razones para explicar dicha rareza (por lo menos para nosotros), una posible justificación es que *Facebook* no suele utilizarse para este tipo de negocio.

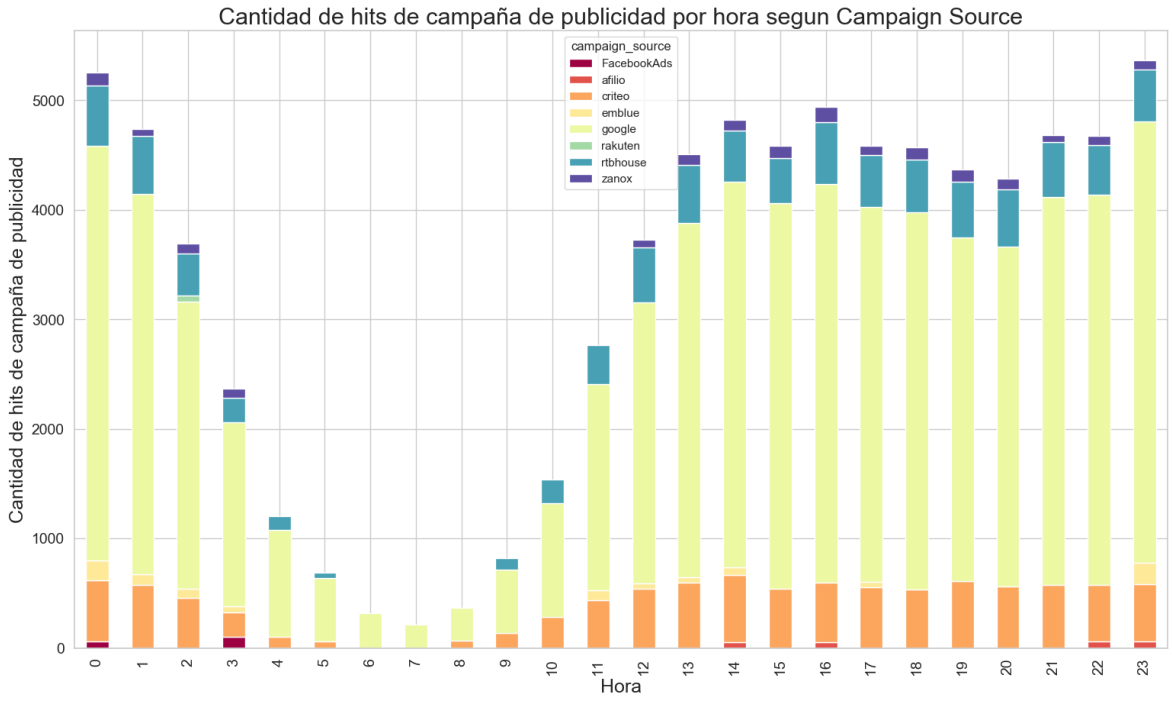

Figura 10: Cantidad de hits de campaña de publicidad por hora segun Campaign Source.

2.6. Análisis según sistema operativo del usuario

A continuación analizaremos algunos aspectos interesantes que nos proporciona el dataframe sobre el sistema operativo que utiliza el usuario a la hora de visitar su página web.

Para empezar, en la Figura 11 estudiamos la cantidad de eventos en función del sistema operativo del usuario y la hora en la cual las acciones del usuario dispararon un nuevo evento. Lo primero que inferimos del plot es que los sistemas operativos Android y Windows son los preferidos de los usuarios que utilizan *Trocafone* (mas adelante veremos cual es el más usado de estos dos). Por detrás de estos aparece iOS, la plataforma móvil de Apple. Respecto al horario, se puede observar como en dispositivos mobiles/portables como lo son SmartPhones/Tables/etc (cuyo SO puede ser Android o iOS) obtiene su valor máximo en un rango de +- 3 horas con centro en las 23 hs, en cambio para dispositivos como computadoras se alcanza el pico máximo entre las 17 y 20 hs.

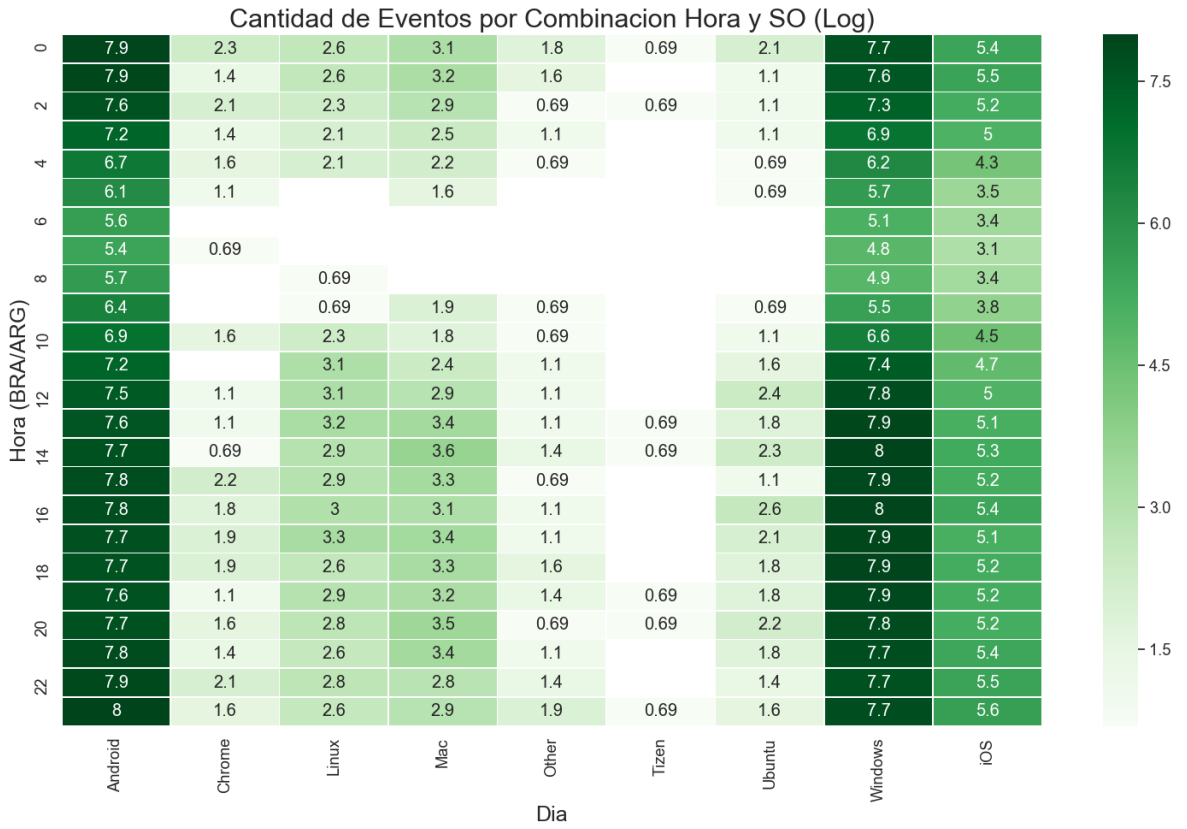

Figura 11: Cantidad de eventos por combinación hora y SO (Log).

Como usuarios activos de aparatos tecnológicos que hacen uso de sistemas operativos complejos y avanzados, conocemos también la diversidad en cuanto a resolución de displays de dichos aparatos. En este caso *Trocafone* nos facilitó información sobre la resolución de pantalla de algunos dispositivos de sus usuarios.

En la Figura 12 además se puede apreciar que Windows es el más usado por sobre Android, algo que no había quedado tan claro en la Figura 11. Dentro de la plataforma Apple, podemos inferir que el dispositivo favorito de los usuarios de *Trocafone* es el iPhone 5/SE ya que la resolución predominante para el SO del gigante de Cupertino es 320x568 (único dispositivo de Apple con dicha resolución). Para el SO Android, 360x640 es la resolución más utilizada. Muchos modelos viejos de distintas marcas fueron construidos con pantallas de dicha resolución. En la plataforma Windows, los displays favoritos son lo de resolución 1366x768, tecnología llamada HD Ready. Para muchas personas es suficiente esa resolución y tal vez no valga la pena comprar un monitor nuevo con una resolución mayor (por ejemplo FULL HD).

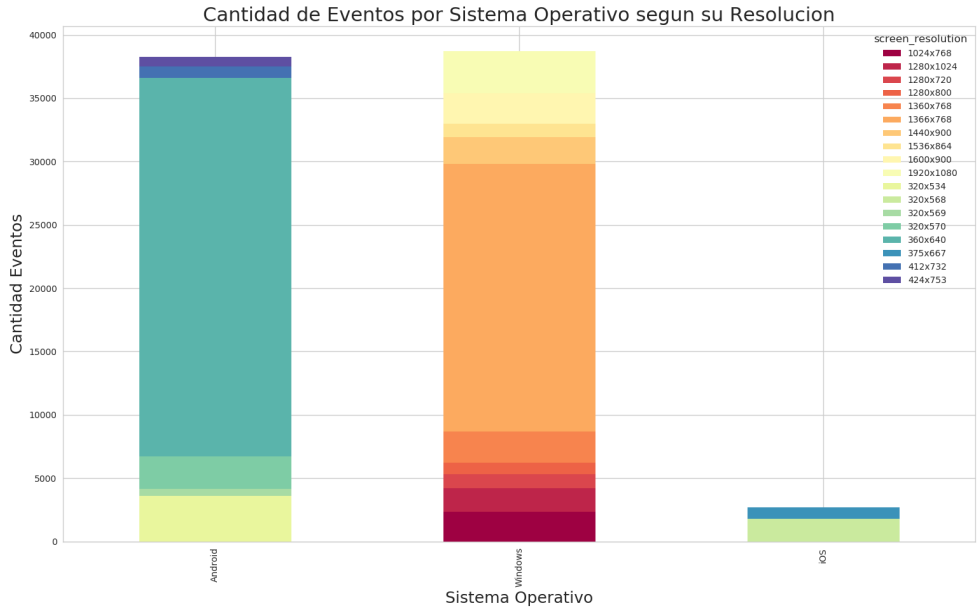

Figura 12: Cantidad de eventos por SO segun su resolución.

2.7. Análisis de búsquedas en el site

El conjunto de datos registra el texto que se ingresó en el campo de búsqueda al realizarse el evento '*searched products*'. Realizando un análisis de este registro y teniendo en cuenta las diferentes posibilidades con las cuales el usuario podría realizar una búsqueda, ya sea por modelo de dispositivo, por marca, y algunos errores de tipeo, se logró obtener cuáles fueron las marcas más buscadas y graficarlas según su frecuencia. La Figura 13 muestra lo obtenido. Como era de esperarse, sabiendo que los iPhone son los dispositivos con mayor cantidad de visitas (Figura 8), la marca Apple es la marca con mayor frecuencia de búsquedas en el site. Sin embargo, Samsung, la marca más comprada (Figura 7) es la siguiente con mayor frecuencia en las búsquedas.

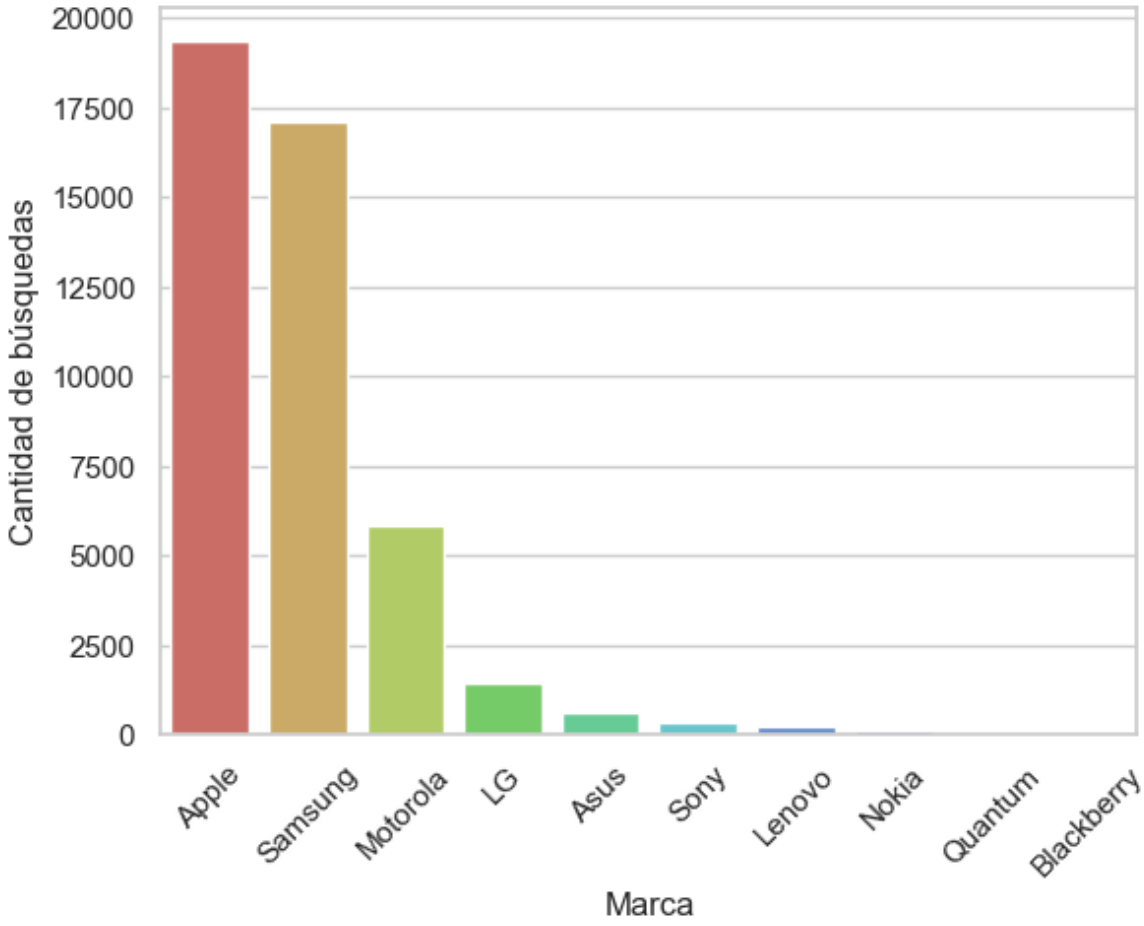

Figura 13: Top búsquedas de marcas en la interfaz web mediante el campo de búsqueda.

En un análisis más detallado de las búsquedas con las dos marcas de mayor frecuencia, la Figura 14 muestra una comparación de personas que buscaron sólo la marca Apple, personas que buscaron sólo Samsung y aquellos que buscaron ambas marcas. Este plot es un complemento a la posible explicación de por qué Apple es la marca más buscada, pero Samsung es la más vendida. Es probable que aquellas personas que buscaron ambas marcas, comparen los diferentes dispositivos. Entre ellos el precio y el sistema operativo. Algunas personas tenderán a comprar la gamma Samsung por su precio, o simplemente por su preferencia en el sistema operativo Android.

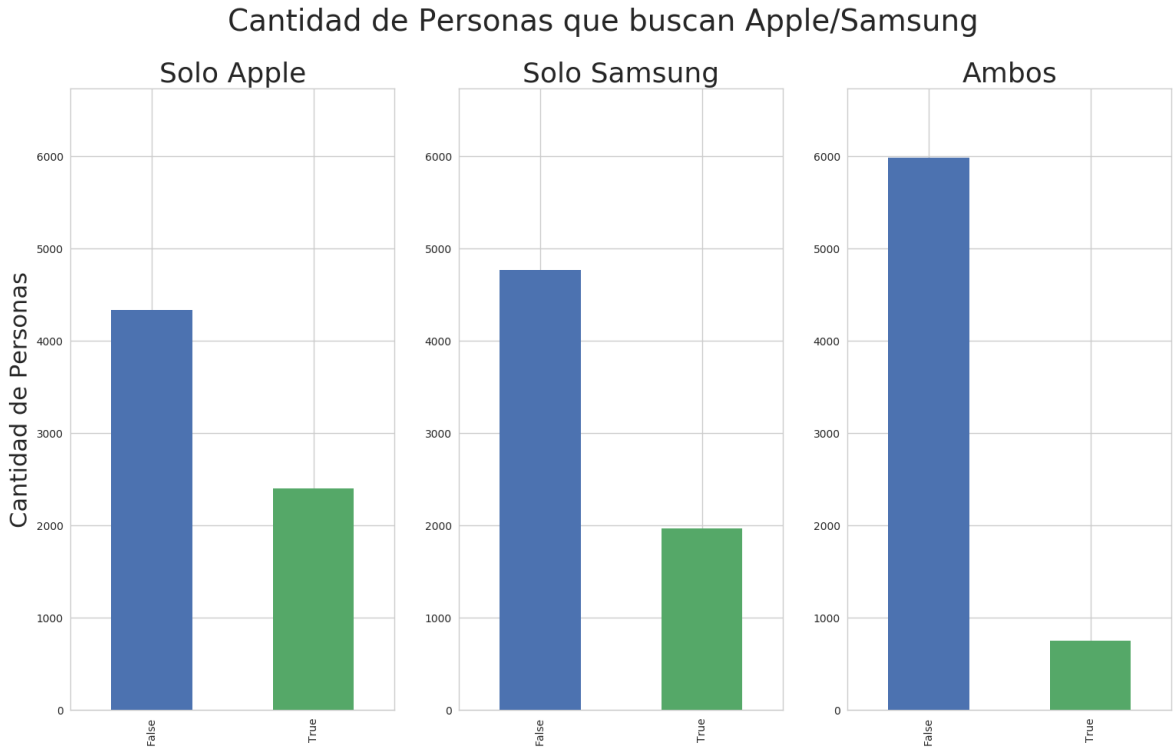

Figura 14: Relación de usuarios que buscaron sólo 'Apple', sólo 'Samsung' y ambos.

2.8. Análisis de recurrencia de usuarios

Cuando un usuario visita su sitio, *Trocafone* guarda información importante en su base de datos, alguna de esta ya la analizamos anteriormente en este mismo informe. *Trocafone* clasifica a los usuarios en dos grupos: nuevos y recurrentes. Ahora es el turno de hablar de este tópico importantísimo para cualquier negocio que quiere medir la respuesta del público al servicio que aporta.

De la Figura 15 podemos concluir que *Trocafone* tuvo un enorme crecimiento los primeros 6 meses del año 2018. Es lógico que sea mayor la cantidad de usuarios recurrentes respecto a los que ingresan al sitio por primera vez. A su vez la pendiente que representa a los usuarios recurrentes es más grande que la que representa a los usuarios nuevos durante los 6 meses. Dicho fenómeno es notorio en el mes 4, donde ambas variables crecen casi linealmente (los recurrentes hasta un poco más rápido).

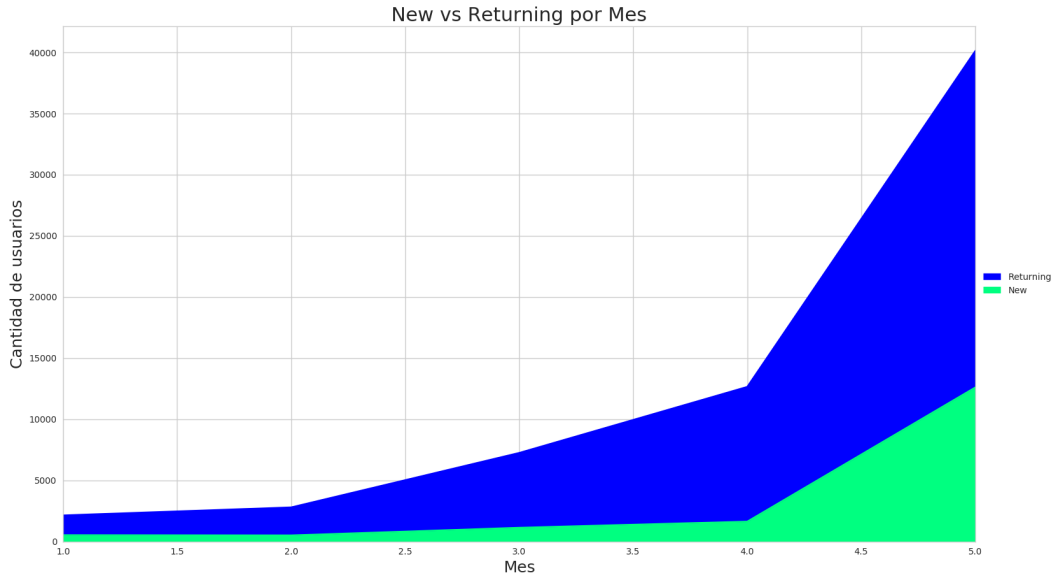

Figura 15: Nuevos usuarios vs recurrentes.

3. Conclusión

El análisis exploratorio sobre el conjunto de datos de la empresa *Trocafone* fue aplicado a diferentes enfoques: análisis temporal, geográfico, marca de dispositivos, campañas de publicidad, sistema operativo de usuario, búsquedas en el site y recurrencia de usuarios. Dentro de ellos se obtuvo resultados tales como cuales son los modelos de dispositivos que más frecuencia tienen en cuanto a ventas, búsquedas o visitas; un mapa que muestra los países con mayor cantidad de visitas al site, la cantidad de eventos a lo largo del tiempo, las horas y los días con mayor tráfico de datos en la página web, etc.

Creemos que solamente *Trocafone* puede decidir si este informe les aporta algo o no ya que no son públicos sus propios análisis sobre este mismo set de datos.

4. Más información

Nuestros análisis fueron realizados en Python Pandas. Utilizamos un repositorio público de github para juntar análisis y filtrar los que creaímos eran adecuados para incluir en este informe. El link del repositorio es:

https://github.com/bauticanavese/Datos-Tps

Dentro del repositorio se encuentra una carpeta llamada TP1, ingresar ahí. Dentro de TP1 se encuentra un notebook llamado AnalisisTrocafone.ipynb en el cual se encuentra solamente el código que genero los gráficos que fueron incluido en el informe. El set de datos no fue incluído en el repositorio por ser muy pesado, se puede encontrar dicho archivo en el siguiente

link:

https://drive.google.com/file/d/1gUddcLLujjFfwZslypUv1LESTM6KiwJn/view?usp=sharing