TP 1: — Análisis Exploratorio

[75.06/95.58] Organización de Datos
Segundo cuatrimestre 2018
Grupo: Reduce Map Fast

Alumno	Padrón	Mail
Jonathan Claudio Medina	100052	jonathanm_96@hotmail.com
Casimiro Pastine	100017	casimiropastine@gmail.com
Nicolás Daniel Vazquez	100338	vazquez.nicolas.daniel@gmail.com
Florencia Rodríguez	100033	florrr1997@gmail.com

Github:
https://github.com/ndvazquez/7506-OrganizacionDeDatos

Índice

1. Introducción

El presente informe reune los resultados y conclusiones obtenidos luego de analizar los datos dados por parte de la empresa Trocafone.

Trocafone es una empresa que se caracteriza por vender celulares usados. Esta empresa implementa un modelo de negocio conocido como Recommerce, en el cual tienen a grandes rasgos tres etapas marcadas.

- Compran un celular usado.
- Reacondicionan este para la venta (re-acondicionamiento (refurbishing)).
- Colocan el producto a la venta.

Trocafone busca afianzar la relación del consumidor con los productos usados, dándole una seguridad al comprador por medio de garantías brindadas sobre los productos comercializados.

Económicamente la empresa realiza sus ganancias con la diferencia de costos entre la compra y re-acondicionamientos vs la venta de los mismos.

1.1. Modalidad de trabajo

Trocafone implementa un sistema conocido como **Trade-In**, el cual consiste en que un potencial cliente reciba un descuento en la compra de un producto a cambio de la entrega de uno previamente usado.

De esta manera Trocafone adquiere productos a un menor precio, mientras que el cliente se beneficia con un un descuento sobre la compra que desea realizar.

Claro está, no es necesario realizar este intercambio, y un cliente puede solo vender su celular, o comprar uno.

A partir de la venta de equipos por parte del cliente, Trocafone provee la logística necesaria para transportar los productos aqduiridos a centros de reacondicionamiento, y así tenerlos listos para su posterior venta.

1.2. Re-acondicionamiento

Luego de su ingreso, los equipos pasan por un proceso de revisión que consiste en los siquientes pasos:

- Realizar la reparación del producto
- Reemplazar su batería.
- Reemplazar los componentes necesarios .
- En caso de un deterioro que lleve a que el producto no pueda ser reparado adecuadamente, se procede a desmontar y quitar las piezas que pueden ser utiles para el reacondicionamiento de otros equipos.

1.3. Comercialización

En la actualidad pueden encontrarse los siguientes canales de venta:

- E-commerce: (https://www.trocafone.com en Brasil y https://www.trocafone.com.ar en Argentina)
- Marketplace: (Con presencia en Brasil dentro de Mercadolivre, B2W, Amazon Brasil y en Argentina dentro de Mercadolibre).
- Tiendas Físicas (Basadas en Brasil con el branding Trocafone)

1.4. Objetivo de este trabajo

El objetivo principal de este trabajo es el aportar a la empresa toda información o análisis que pueda serle útil, y se espera aportar una visión nueva de los datos, y encontrar una manera diferente de mostrarlos y unirlos para sacar información de importancia para Trocafone.

2. Desiciones tomadas

Durante el desarrollo del análisis se tomaron diferentes desiciones respecto a como tratar los tipos de datos y la información que estos dan. Se decidió entonces dedicar esta sección a remarcar y explicar estos, y no así explicarlos en cada análisis realizado, ya que muchas son similares (o idénticas) y por ende no solo sería repititiva, sino que también, distraería de la idea que se quiere dar.

Fecha y hora de eventos: Al trabajar en análisis temporales se convirtieron las fechas brindas al formato 'datetime' para mayor comodidad en su uso. Unicamente tres fechas (aproximadamente) no podían ser traducidas a este formato y fueron removidas. Comparado a la cantidad de datos poseidos, encontramos las fallas insignificantes.

Campos nulos: A lo largo de todo los datos se encontraron muchos campos nulos, pero no todos estos son causa de no haberse podido conseguir los datos necesarios para completarlos. Mientras que hay campos que deberían estar siempre completos, otros no. Por ejemplo, si un usuario entró a la página que explica el funcionamieno de Trocafone ('About us') es esperable que en todos los campos sobre las características de equipos no haya información.

Por otra parte, al realizar un usuario una compra, no se indica el tipo de dispositivo que utilizó en ningún momento, pero podemos asumir con una total seguridad (claro no tendría sentido de otra manera) que utilizó uno. Por falta de información no pudimos llegar a una conclusión de la causa de esta omisión, pero si consideramos importante remarcarla.

Campañas publicitarias: Para realizar el Diagrama de Sankey que luego se visualizara en Campañas Publicitarias se tomaron 300 apariciones, para así el gráfico tuviera un valor significativo de hacia donde se dirige cada usuario cuando entra mediante una campaña de publicidad

3. Conociendo los datos

Al empezar a investigar un poco los datos, notamos que para cada tipo de evento se tenía asociada distinta información. Lo que decidimos hacer fué filtrar por eventos y revisar qué datos nos ofrece cada uno para tener un mayor entendimiento del dataframe.

El filtrado se hizo quedandonos únicamente con las filas que corresponden para el evento en cuestión y descartando todas las columnas que tenían NaNs en todos sus valores. De esta manera podemos observar facilmente que variables estan asociadas a cada tipo de evento.

A continuacion mencionamos algunos de los eventos más importantes y qué datos nos provee cada uno.

Viewed Product

Al quedarnos solo con los productos vistos observamos que unicamente nos muestra datos sobre el usuario y el teléfono observado.

	timestamp	person	sku	model	condition	storage	color
2	2018-05-31 23:38:09	0004b0a2	2694.0	iPhone 5s	Bom	32GB	Cinza espacial
4	2018-05-29 13:29:25	0006a21a	15338.0	Samsung Galaxy S8	Bom	64GB	Dourado
13	2018-04-09 20:13:14	000a54b2	12661.0	Motorola Moto Z Play	Muito Bom	32GB	Preto
22	2018-05-24 11:27:47	000a54b2	10254.0	iPhone 7 Plus	Excelente	256GB	Dourado
26	2018-05-24 11:28:59	000a54b2	6581.0	iPhone 6S	Bom	16GB	Cinza espacial

Visited Site

Este evento corresponde a cuando un usuario entra al sitio, por ser tan genérico es el tipo de evento que más información nos provee.

	person	channel	new_vs_returning	city	region	country	device_type	screen_resolution	operating_system_version	browser_version
1	0004b0a2	Paid	New	Camaragibe	Pernambuco	Brazil	Smartphone	360x640	Android 6	Chrome Mobile 39
5	0006a21a	Paid	New	Rio de Janeiro	Rio de Janeiro	Brazil	Smartphone	360x640	Android 5.1.1	Android 5.1
9	000a54b2	Paid	New	Rio de Janeiro	Rio de Janeiro	Brazil	Computer	1920x1080	Windows 10	Chrome 65.0

Descartamos la columna timestamp para mostrar mejor la imagen

Ad Campaign Hit

Este tipo de evento nos señala los usuarios que entraron a la página desde una publicidad armada por una campaña publicitaria.

	timestamp	person	url	campaign_source
0	2018-05-31 23:38:05	0004b0a2	/comprar/iphone/iphone-5s	criteo
6	2018-05-29 13:29:27	0006a21a	/comprar/samsung/galaxy-s8	criteo
11	2018-04-09 20:12:31	000a54b2	/	google

Searched Products

Este evento nos muestra una busqueda dentro del sitio web, dandonos los terminos de busqueda y aquellos equipos que se muestran al usuario.

	timestamp	person	skus	search_term
157	2018-02-06 02:29:49	00204059	2692,6819,823,2779,13864,2784,8135,6805,2773,2...	moto g 4
159	2018-02-06 02:32:41	00204059	2692,6819,823,2779,13864,2784,8135,6805,2773,2...	moto g 4
238	2018-05-21 19:56:33	0024ad28	3371,6357,6371,10896,2718,2777,6001,2694,3191,...	comprar celulares usados bom e barato em poa rs

Search Engine Hits

Search Engine Hits nos dice cuando un usuario accede al sitio mediante un buscador externo, aclarandonos desde cual entró.

	timestamp	person	search_engine
10	2018-04-09 20:12:31	000a54b2	Google
17	2018-05-24 11:21:07	000a54b2	Google
47	2018-04-06 05:12:05	00184bf9	Google

Checkout

Es el último paso antes de hacer una compra sobre un producto, mostrandonos algunas características de los modelos en cuestión.

	timestamp	person	sku	model	condition	storage	color
3	2018-05-31 23:38:40	0004b0a2	2694.0	iPhone 5s	Bom	32GB	Cinza espacial
7	2018-05-29 13:29:35	0006a21a	15338.0	Samsung Galaxy S8	Bom	64GB	Dourado
44	2018-05-24 11:34:32	000a54b2	12660.0	Motorola Moto Z Play	Bom	32GB	Preto

Conversion

Corresponde a la compra efectuada de un producto, dandonos algunas características sobre el modelo.

	timestamp	person	sku	model	condition	storage	color
220	2018-03-20 17:46:12	00204059	3084.0	Motorola Moto X2	Muito Bom	32GB	Couro Vintage
2282	2018-04-26 22:28:53	00c13dee	6650.0	Samsung Galaxy Core Plus Duos TV	Muito Bom	4GB	Branco
2547	2018-06-10 14:37:50	00fdbb4b	3348.0	Samsung Galaxy S6 Flat	Muito Bom	32GB	Branco

Lead

Nos indica cuándo un usuario quiere ser notificado cuándo se reponga el stock de un producto en particular.

	timestamp	person	model
3248	2018-04-17 22:11:19	01139919	Samsung Galaxy On 7
6636	2018-04-07 11:37:11	01bca043	iPhone 6 Plus
7036	2018-02-12 17:23:30	01db2fe6	Samsung Galaxy J5

Canales

Tipo de tráfico de dónde proviene el usuario que ingresa al sitio.

- Pago: Tráfico proveniente de, por ejemplo publicidades pagas via a Google AdWords o otras plataformas de busqueda pagas.
- Directo: Todo tráfico cuya proveniencia no es conocida.
- Organico: Tráfico proveniente de motores de búsqueda que no es pago.
- Referido: Tráfico ocurrido cuando un usuario encuentra el sitio a traves de un sitio que no es un motor de búsqueda.
- Social: Tráfico proveniente de redes sociales, como por ejemplo Facebook, Twitter o Instagram.
- Email: Tráfico proveniente de mails.

Canal Directo

Como antes fue mencionado, el tipo de tráfico directo es aquel que no es referido de ningún sitio web. Se relaciona entonces a este tipo de canal como aquel en el que el usuario ingresa el URL del sitio directamente en el browser.

Por desconocimiento de las prácticas de Trocafone tomamos todo canal directo hallado como uno realizado por un usuario ajeno a la empresa. Mencionamos esto ya que este tipo de tráfico puede ser generado por empleados que ingresan al sitio por diversas razones y que lógicamente no necesitan por ejemplo , 'googlear' la url de la página. Esto ya que se puede evitar este tipo de información 'engañosa' filtrando los IPs de los empleados de nuestros 'Webs Analytics' y así ya no trabajar con ese tipo de datos (al menos cuando queremos obtener un análisis de los clientes).[1]

Tiempo abarcado

Los datos ingresados abarcan los meses de Enero a mitad de Junio de 2018.

3.1. Subconjunto de datos

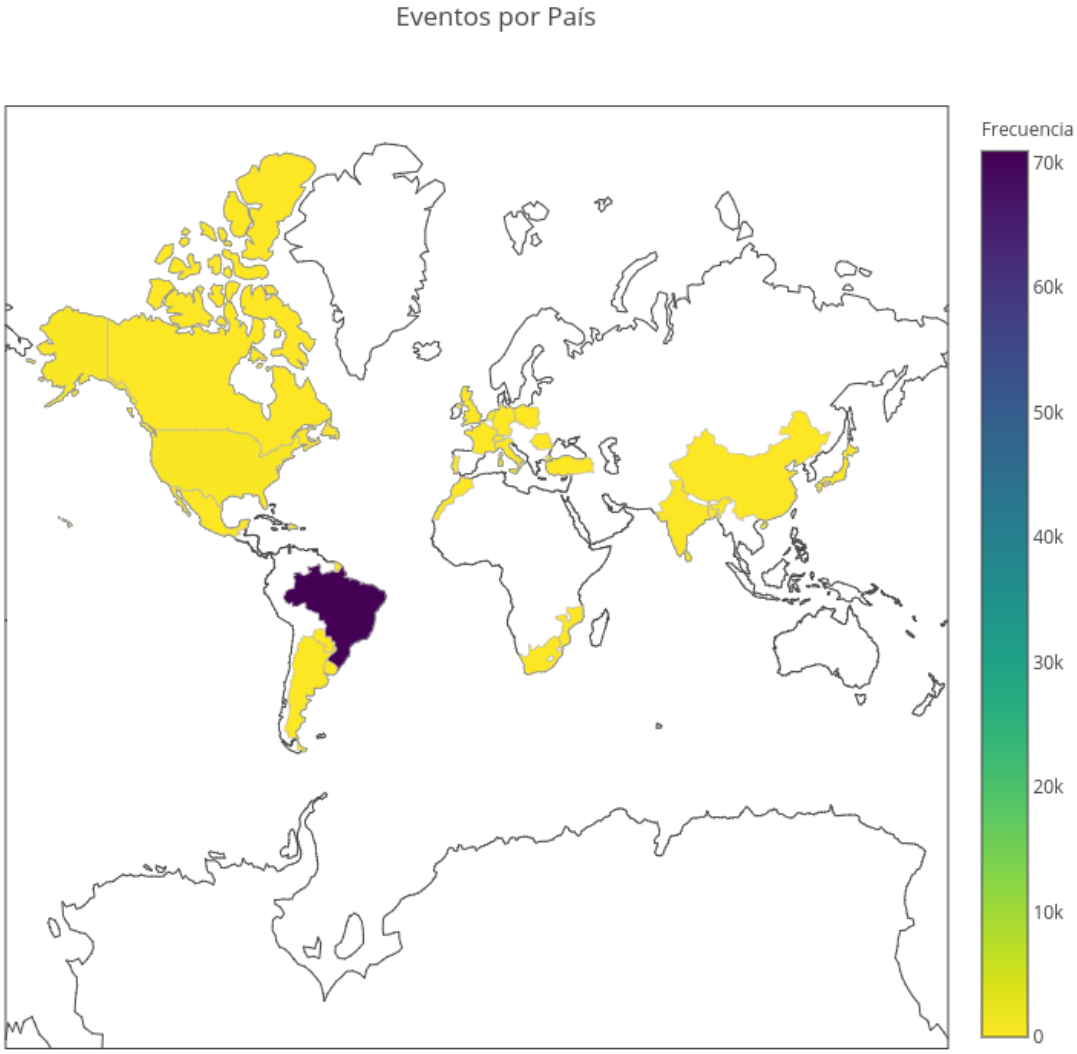
Notamos que todos los datos poseían al menos un checkout, podría ser que los datos proporcionados por la empresa hayan sufrido ese criterio de corte. Dado que tenemos un subconjunto (una porción del total de los datos), todo el análisis realizado en este informe sera determinado por este mismo subconjunto.

[1] Más información sobre los tipos de canales puede encontrarse en: https://www.smartbugmedia.com/blog/what-is-the-difference-between-direct-and-organic-search-traffic-sources

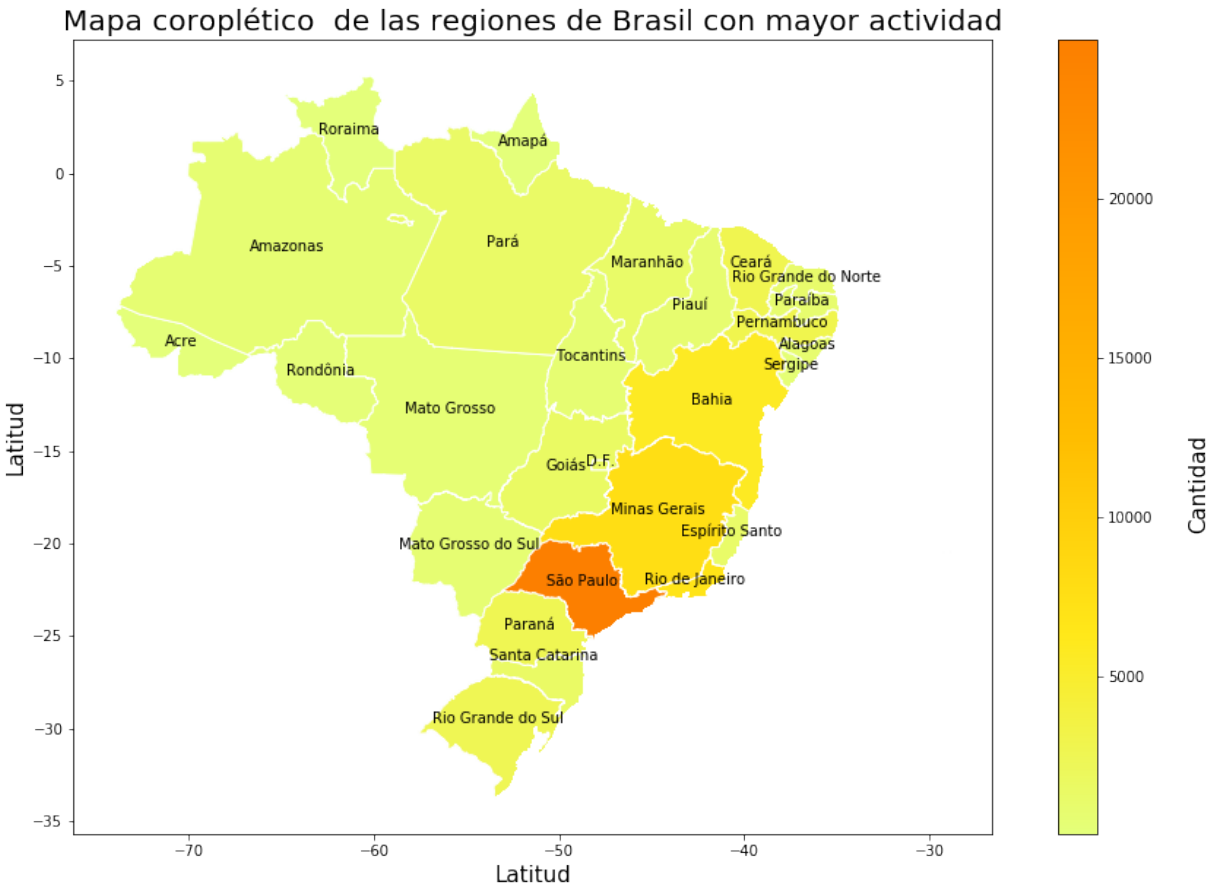
4. Tráfico en el sitio

4.1. Tráfico por geolocalización

Siendo Trocafone una empresa que empezó en Brasil, y teniendo a Argentina como segundo país en ventas, observamos de dónde provienen los eventos brindados.



La gran mayoritaria parte de los eventos de nuestra base de datos, tiene origen, como se puede ver, en Brasil. Por ende vemos entonces la distribución de estos en Brasil, por estado.



Podemos concluir que las regiones (o estados) donde hay más trafico son São Paulo, Rio de Janeiro, Minas Gerais y Bahia.

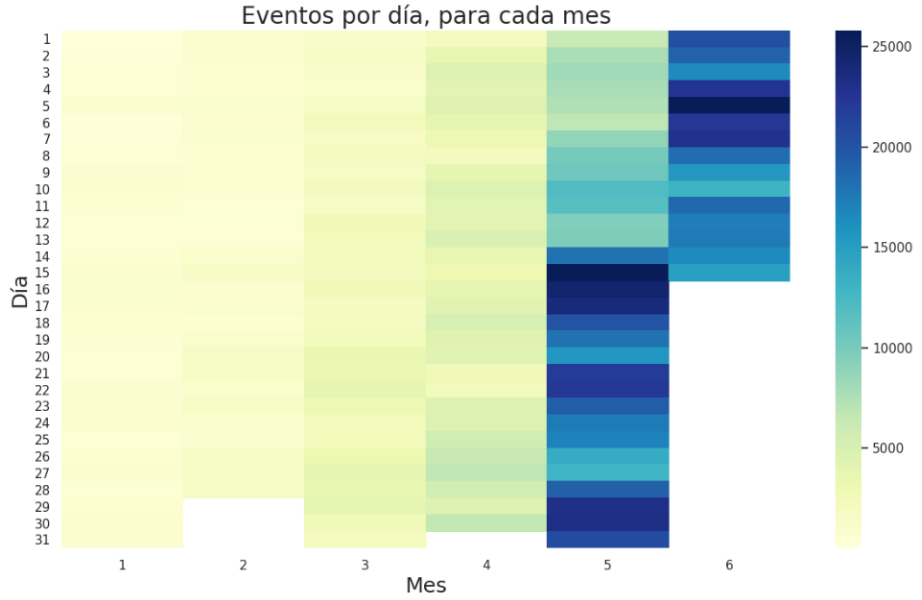
Comparando con lo que muestra en Google Trends [2] y los locales de Trocafone [3], el trafico de São Paulo puede llegar a deberse porque todas las oficinas de Trocafone estan en ese estado, siendo además São Paulo uno de los principales centros financiero del país.

La gran diferencia que se muestra entre Google Trend y nuestro grafico puede deberse a que estamos trabajando en subconjunto del dataset.

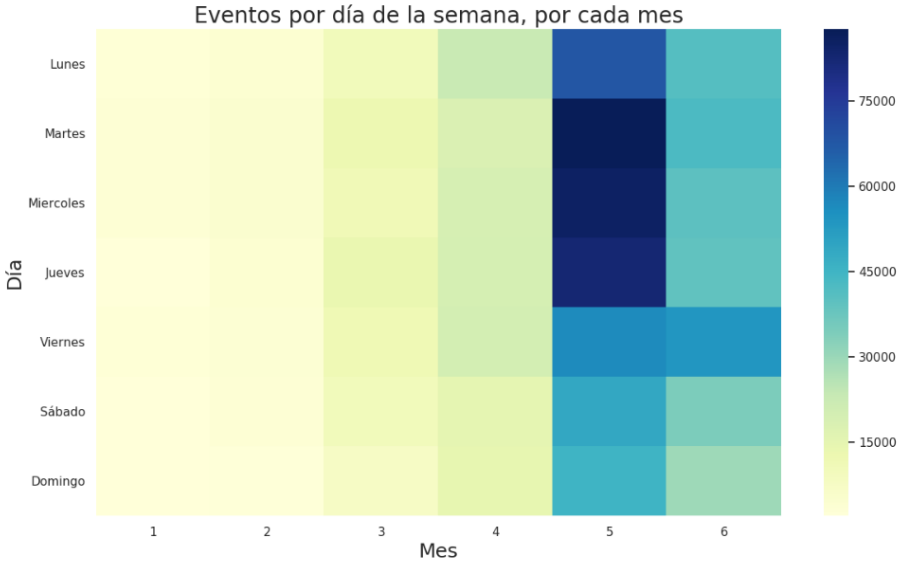
[2] https://trends.google.com/trends/explore?date=2018-01-01%202018-06-30&geo=BR&q=trocafone
[3] https://www.trocafone.com/quiosques

4.2. Tráfico en el tiempo

Para comenzar el análisis del tráfico de Trocafone, necesitabamos inicialmente saber como estaba distribuido en los meses brindados como datos los eventos, en otras palabras, la fluctuación de tráfico en los meses de Enero a alrededor mitad de Junio. En las siguientes visualizaciones podemos ver el cambio de esta.

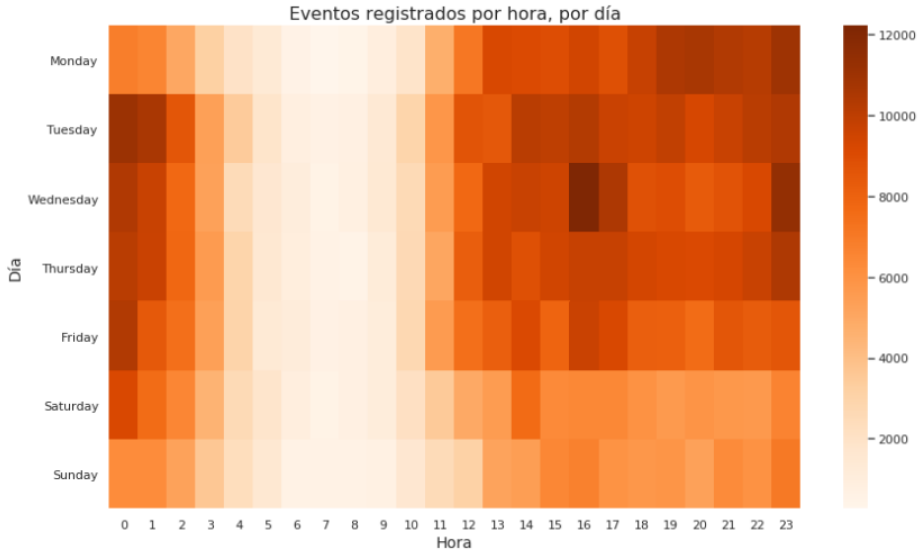


También podemos preguntarnos el tráfico por día de la semana, a priori podemos pensar que hay días, tal vez fin de semanas, que la gente posee más tiempo para hacer el tipo de tareas en las que entra el buscar y comprar un celular. Como siempre , es mejor primero ver los datos y sacar luego las concluciones adecuadas.



El resultado es interesante, los fines de semana son los días con menos tráfico de la semana, contrario a lo inicialmente pensado. Y son los días de Lunes a Viernes los que reciben la mayor cantidad de visitas.

Se puede observar un claro aumento de tráfico el mes de Mayo que continua el mes de Junio, recordando que solo poseemos información de unicamene la mitad de los días de este.



Se puede observar que el tráfico entre las 4hs y las 9hs es casi nulo, comenzando a aumentar a partir de las 10hs, independientemente del día de la semana. Otra cosa interesante, que corresponde con lo mencionado previamente, es que el tráfico en el fin de semana es significativamente más bajo que durante los días de semana. Tambíen se puede ver que el período de tiempo consecutivo con mayor actividad surgió los lunes a entre las 19hs hasta las 2hs del martes.

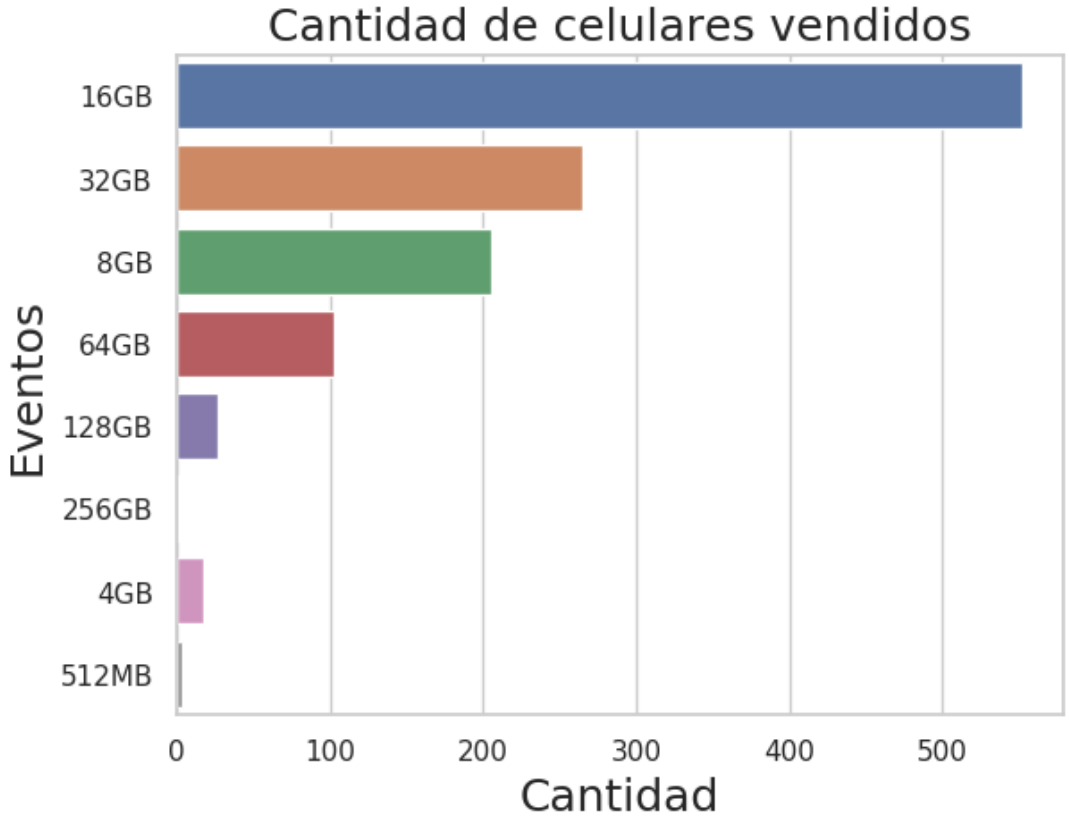
5. Características de los Celulares

La compra de celulares para restaurar es parte vital de Trocafone, por ende encontramos de suma importancia el analizar que tipos de celulares, y con que características son los que se están vendiendo más. Con esta información podemos:

- Evaluar que tipos de celulares son los mejores para comprar (los que sabemos que se venden más).
- Analizar que tipos de mejoras son necesarias realizar, y cuales no lo son tanto, ya que la gente no muestra interés en estos.

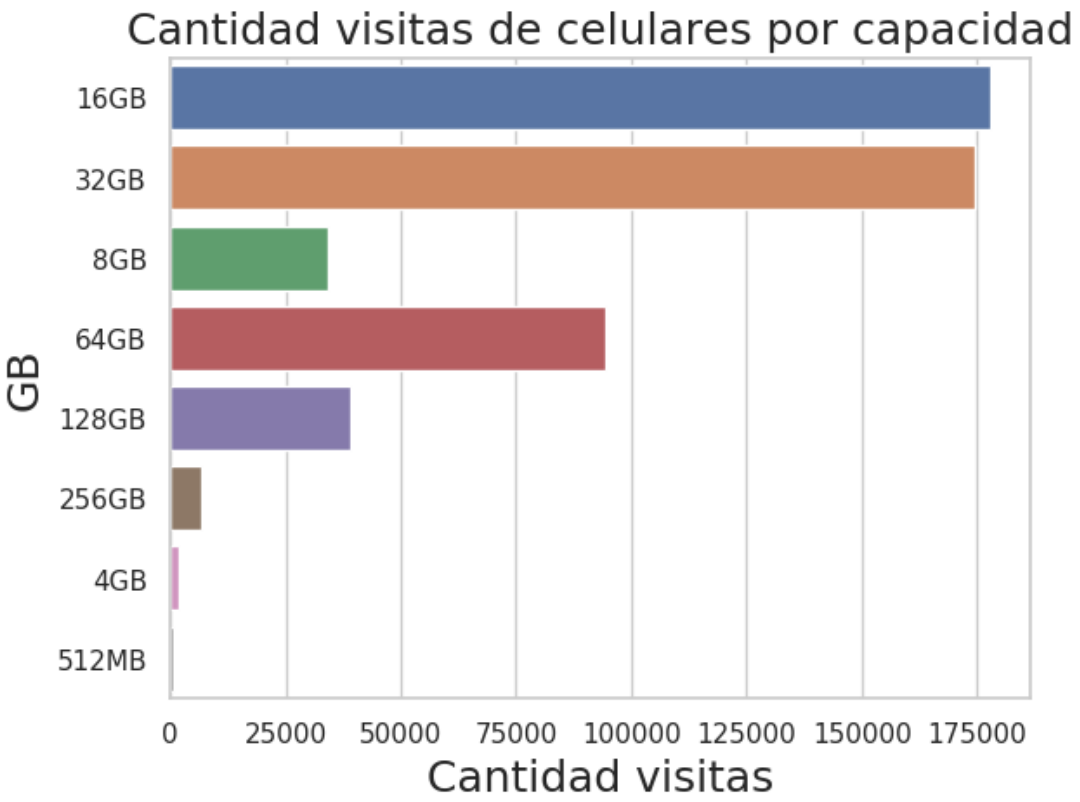
5.1. Capacidad (GB)

El siguiente es el análisis de de los celulares por su cantidad de GB.



En este gráfico se puede ver una claro pico de ventas en los celulares que poseen 16 GB. Pese a que no es la capacidad de un celular la única variable para la venta, si la consideramos de un gran interés para los clientes, y la vemos como una característica de esencial importancia.

Encontramos interesante observar a continuación las cantidad de visitas que poseen los celulares con las distintas capacidades de memoria, para compararla con los datos antes encontrados.



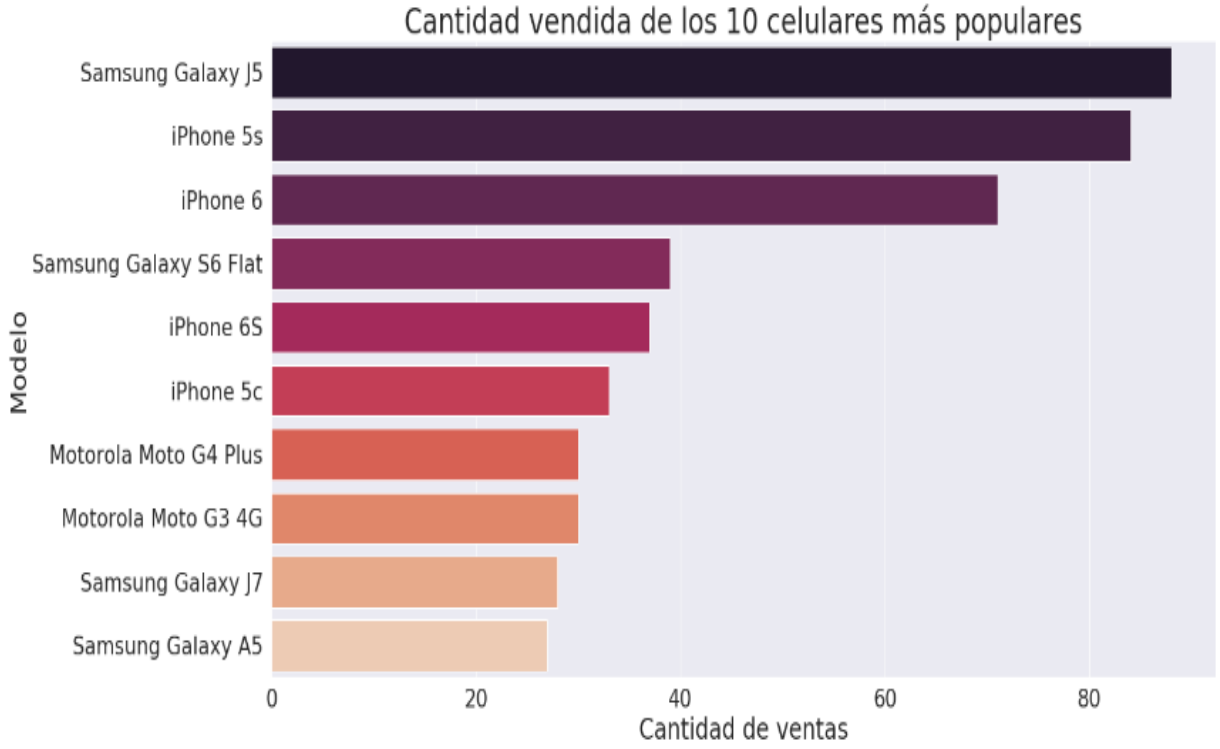
Aquí sucede algo interesante, pese a que en ventas los equipos que poseian 16 GB eran los más vendidos por una considerable ventaja, la tendencia no se replica de la misma manera en la cantidad de visitas de los mismos. En esta situación, los celulares con 32 GB parecen obtener el mismo interes que los de 16, y la presencia de los celulares de 64 y 128 GB aumentó considerablemente respecto al anterior.

¿Por qué puede suceder esto? Podemos suponer que el valor monetario es un factor importante en este caso. Los precios de los celulares de 32, 64 y 128 GB son obviamente más elevados, y pese a que los usuarios tienen interes en celulares con más memoria (lo cual es lógico considerando que esta característica está ligada a un mejor equipo) el costo los haga terminar comprando celulares con menos capacidad.

Claro está, por falta de información de stock de celulares, tampoco podemos saber si esta menor venta se debió también a una menos cantidad de equipos con (utilizando la segunda memoria más popular) 32GB.

5.2. Modelo

Analizamos además los modelos más utilizados, viendo los top 10 más vendidos.



Esto no solo nos muestra los top 10 modelos vendidos, pero también podemos observar las tres marcas más populares, las cuales son, claramente, iPhone, Samsung y Motorola.

Se puede además, ver claramente que son los modelos más nuevos los que tienen más demanda. De aquí se puede sacar no solamente que celulares comprar sino también que celulares compran en el futuro, ya que claramente las compañias que fabrican sacan nuevos modelos constantemente. Es rentable , podemos preguntarnos, comprar celulares , que más allá de ser usados (logicamente ya que es el negocio de Trocafone) sean antiguos? La realidad es que por la información brindada y mostrada a continuación (y el claro interés general del mercado), creemos que no.

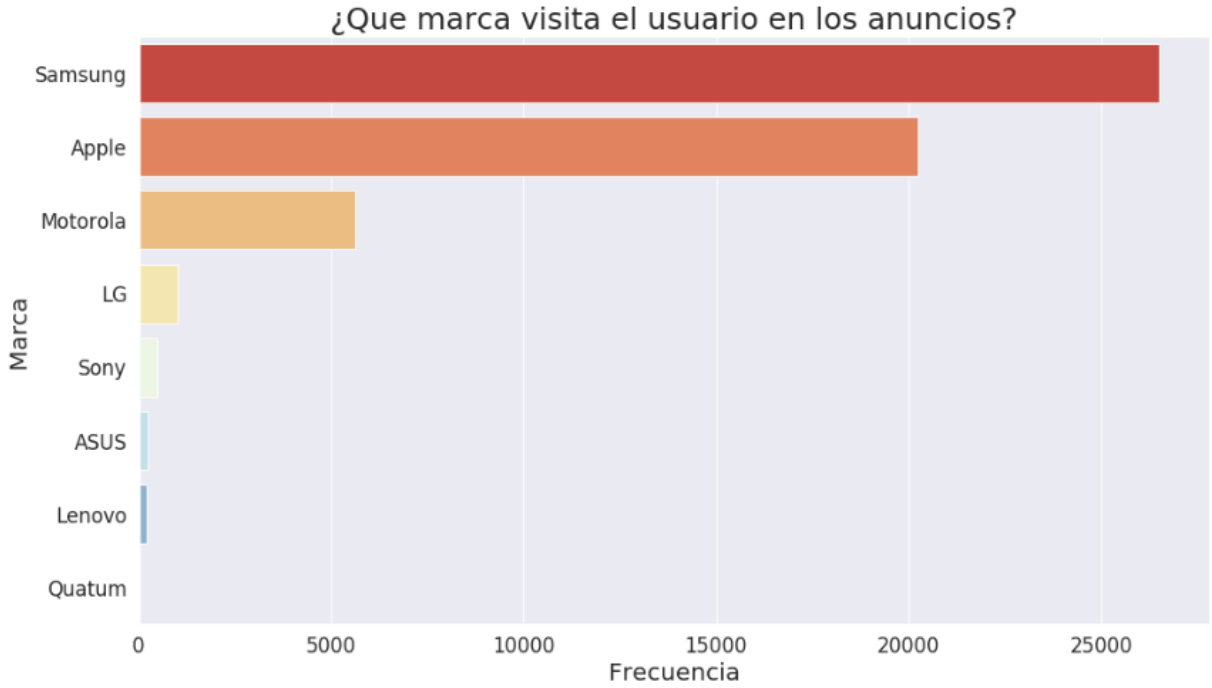
Nos interesa ver ahora los modelos no más vendidos, pero más vistos, y ver la relación con lo antes analizado.

Aquí suecede algo interesante, por ejemplo nuestro celular más vendido está en la posición 12 de más visitas (no entra, claro, en nuestro top 10). Y si miramos más detalladamente , ningún modelo de Samsung o Motorola que está en el top 10 de lo más vendidos lo está en el de más visto. Pero lo mismo no sucede con los celulares iPhone, ya que estos sí tiene modelos en las dos visualizaciones.

Las razones que se pueden sacar de esto son meramente especulativas. Una razón probable es el precio de los dispositivos, ya que los celulares más vistos están entre los más caros del sitio, mientras que los más comprados no son necesariamente estos. Puede suceder que los compradores que buscan celulares iPhones sean ya más concientes de el costo, y no se vean sorprendidos por el valor de estos. Mientras que un comprador que desea un Samsung, mire más detalladamente el dispositvo que puede comprar.

5.3. Empresa

En este pequeño análisis, que procede al anterior realizado de los modelos, observamos las empresas más populares.



Los datos obtenidos se relacionan coherentemente con los modelos más vistos, siendo Samsung, Apple y Motorola los más vistas.

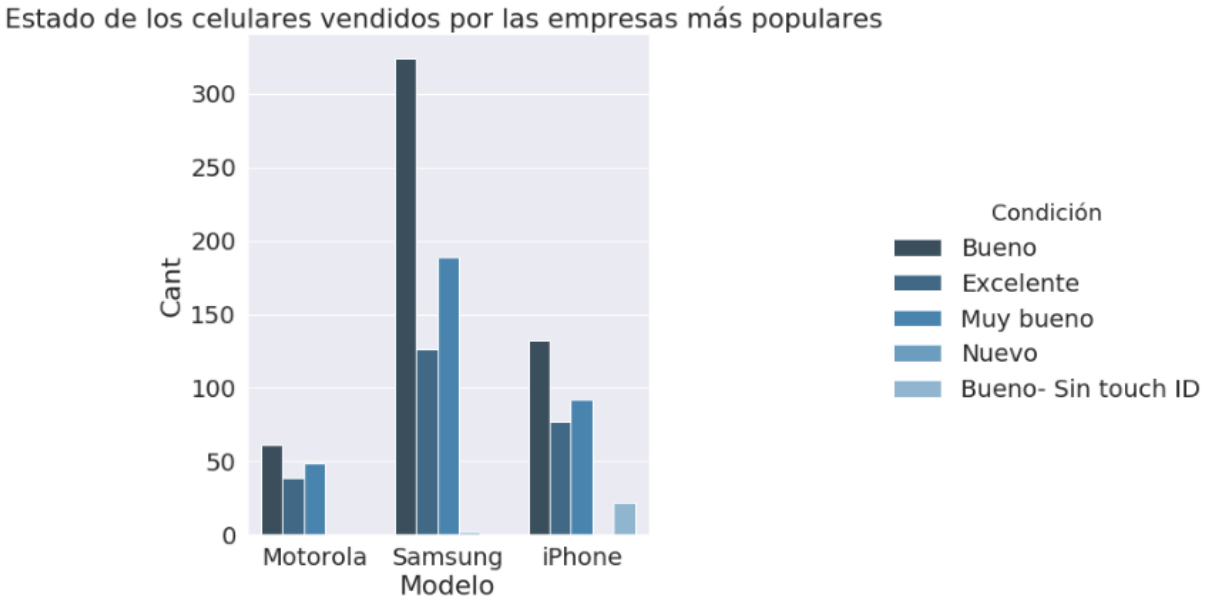
5.4. Estado

Los principales estados de los equipos son:

- Excelente
- Muy Bueno
- Bueno

Habiendo además una pequeña cantidad (insignificante) de equipos con estado 'Nuevo' y otros con estado 'Bueno - Sin touch id'.

Habiendo pudido observar las tres empresas de celulares más populares, nos interesa saber ahora el estado en el que se venden los equipos de estas.



Aquí se puede observar algo interesante, en las tres marcas de celulares más vendidas el estado 'Bueno' es el que tiene más ventas, seguido de 'Muy bueno' y 'Excelente'. A nivel empresa, esta información es importante, ya que pueden derivar a análisis y preguntas del tipo,

- Hasta que punto es necesario restaurar el celular que se compra?
- Es lo mismo el estado del equipo para cada empresa?

Estas preguntas pueden ser respondidas de una manera más certera teniendo además un conocimiento previo de los precio de los celulares y la cantidad en stock que se poseía y se posee. Ya que si la ganancia entre un celular en buen estado y muy bueno o excelente es lo suficientemente considerable para vender menor cantidad y seguir ganando más, nuestro análisis, más allá de interesante, no lograría tal vez el reducir en algunos celulares el trabajo de restauración (ya que no sería un negocio para la empresa).

Por otra parte, de saber la cantidad de celulares de estas empresas que se poseía en stock, podríamos también saber realmente la importacia dada al estado por el cliente (más allá que esta característica no es la única que se considera al comprar un equipo). Por ejemplo, Samsung vendió la mayor cantidad de celulares con un estado 'Bueno', pero cuantos se poseian? Se vendió la mayoría, o pese a la cantidad considerable de celulares vendidos siguió quedando otra similar en venta?

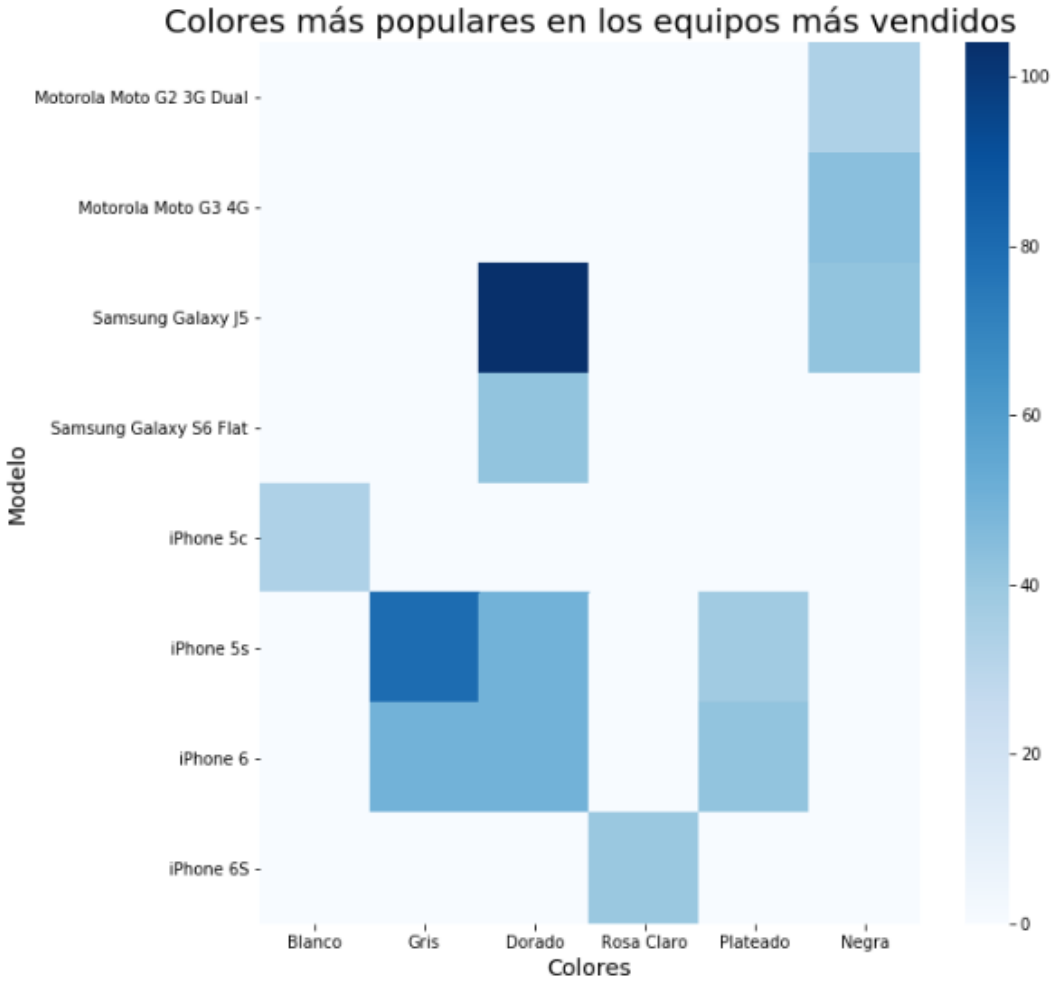
Nuevamente, estas preguntas las encontramos interesantes de realizar y reflexionar, pese a que no podemos responderlas debido a la falta de datos poseida.

Como una reflexión final, creemos a esta característica una de gran importancia a la hora de comprar un celular usado, ya que es siempre una de las mayores preocupaciones al comprar un equipo no nuevo. Por ende encontramos toda información referente a esto de una gran importancia, y de gran valor para la empresa.

5.5. Color

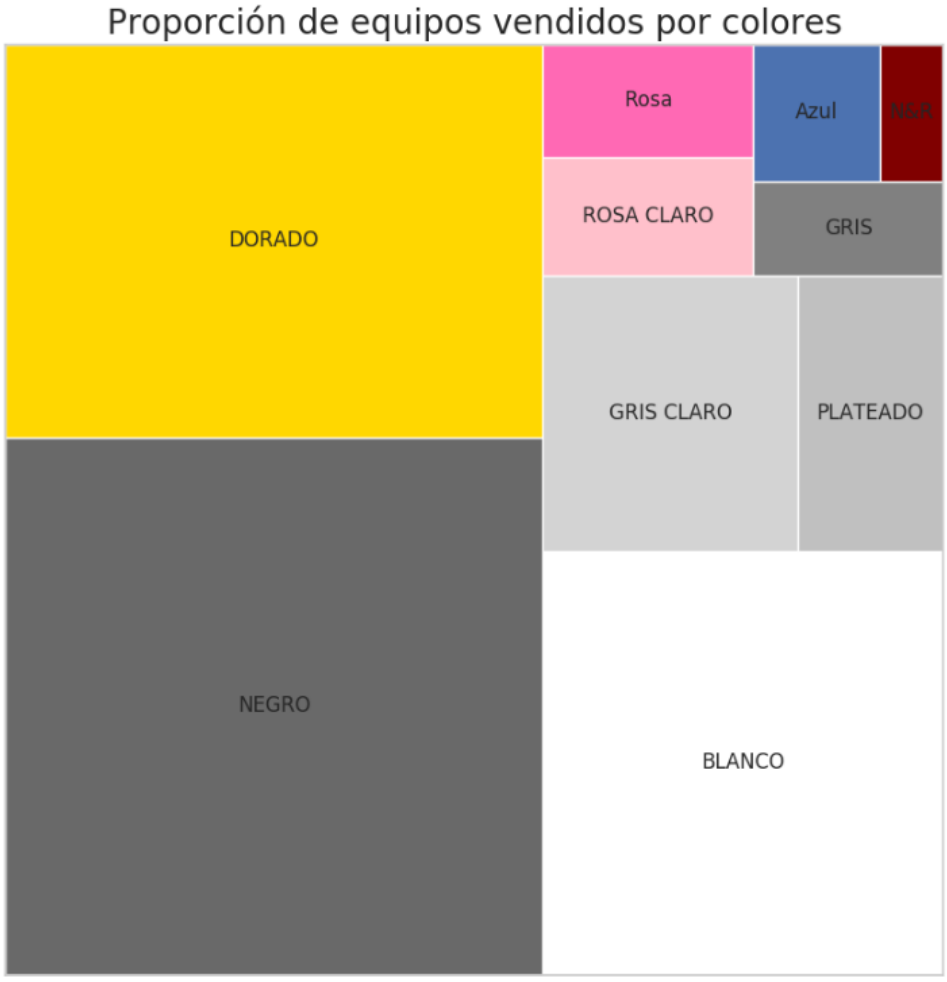
El tipo de color en un celular es algo que sabemos queda generalmente en un grupo reducido de colores. De todos modos, consideramos es una característica interesante para analizar.

Siguiendo el análisis anterior, queremos ver, en primera instancia, los colores más utilizados en los modelos más populares.



Se puede observar una clara preferencia por los tipos de colores negro, gris y plateado. Este tipo de información, como dijimos ya antes, sirve para elegir que tipo de equipo se va a comprar para posteriormente vender. Tal vez un celular que sabemos que tiene muchas ventas se venderá no importe el color, pero uno de más bajo mercado, si merece posiblemente tener en cuenta estos pequeños detalles.

Por último vamos a ver los colores que poseen las celulares vendidos en los meses proveidos, por un tema de prolijidad de la visualización y prioridad de la información, nos quedamos únicamente con los 10 colores más vendidos.

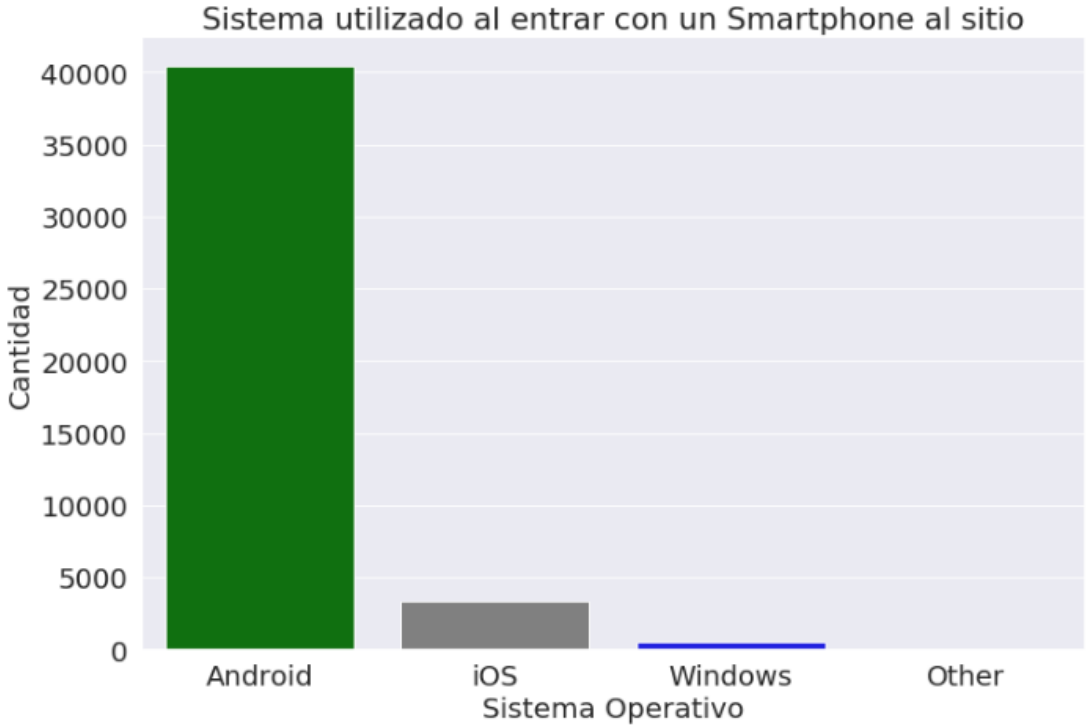


Nuevamente, las gamas de los grises y negros son las más populosas, una información acorde a lo que se puede observar normalmente. De todos modos, si es remarcable el color dorado, ya que no es un color tan común en Argentina, pero si consideramos que nuestros datos provienen mayoritariamente de Brasil, no deberia sorprendernos.

6. Dispositivos utilizados por los usuarios

A partir de un pequeños análisis realizado, se pudo observar que los dos (y casi únicos) dispositivos utilizados por los usuarios son Smartphones y Computadoras, en cantidades similares.

Siendo el negocio de Trocafone los celulares, nos interesa ver los sistemas operativos que poseen, de ser brindada la información, los equipos que utilizan los usuario que ingresan al sitio.



Como podíamos suponer , se ve una clara mayoría de usuarios que utilizan Android como su sistema operativo al ingresar al sitio web.

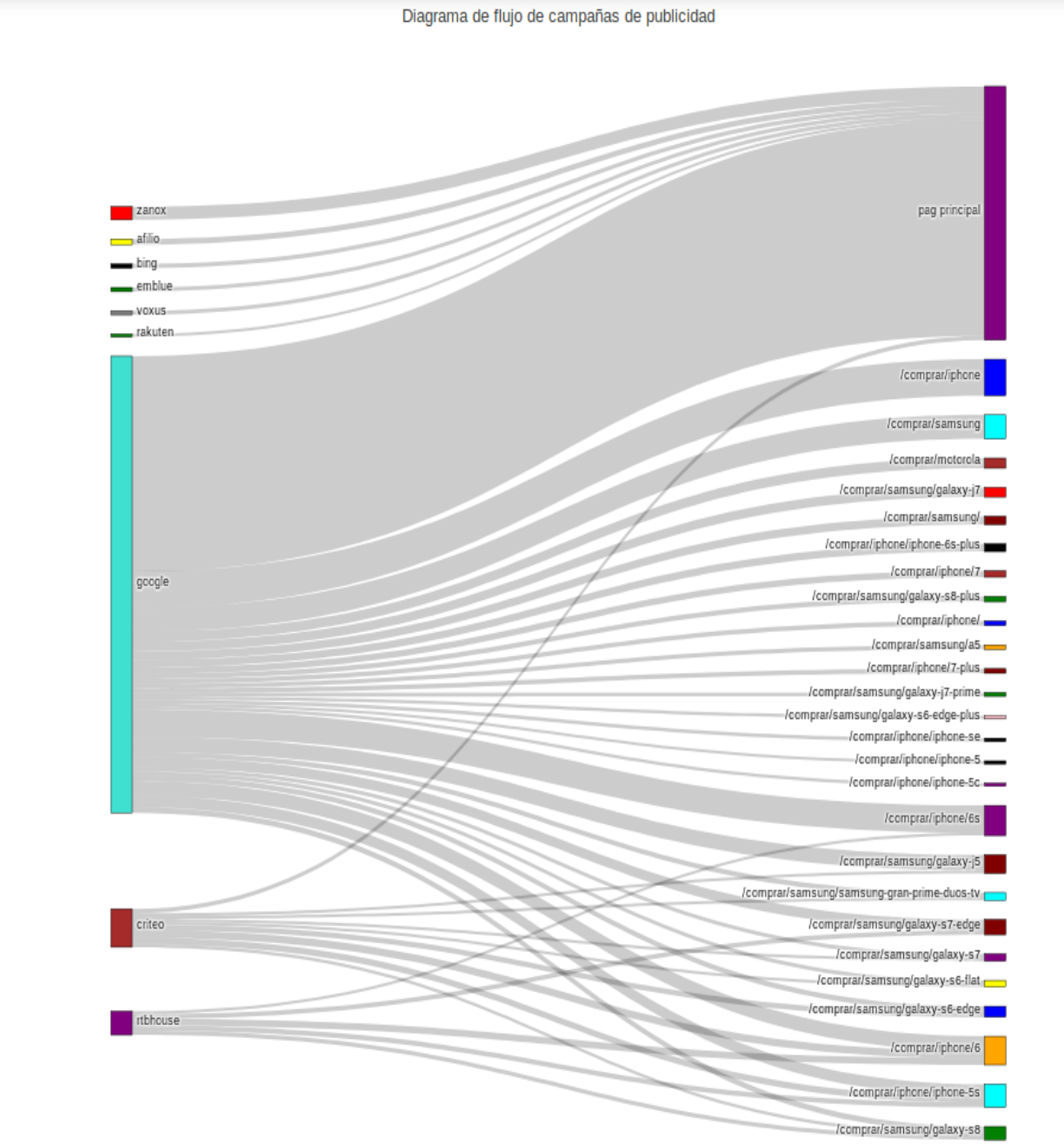
Teniendo los celulares iPhone una gran cantidad de ventas , nos preguntabamos si la gente que miraba o compraba celulares de esta marca entraban usando un sistema operativo Android o iOS. Esto nos parecia interesante ya que podíamos buscar una conformidad con el sistema de parte de los compradores. Si las mayoría de las visitas con un sistema IOS eran dirijidas a celulares como los iPhone podríamos pensar que los clientes deseaban mantener el mismo sistema operativo. Por otra parte, si la mayoría miraba Android, algo diferente sería conluído, y es que los usuarios estaban pensando en cambiar de sistema.

Lamentablemente los datos no nos posibilitaron realizar este análisis, ya que toda persona que compró o vió un producto en el sitio no poseía los datos del sistema con el cual había ingresado. Nos pareció de igual manera algo interesante para remarcar, ya que de poseer un set de datos más completo podría ser posible realizarlo.

7. Campañas Publicitarias

7.1. Llegadas de los usuarios al sitio

En esta sección vemos un grafico que nos muestra hacia donde entran los usuarios que vienen desde una campaña publicitaria.

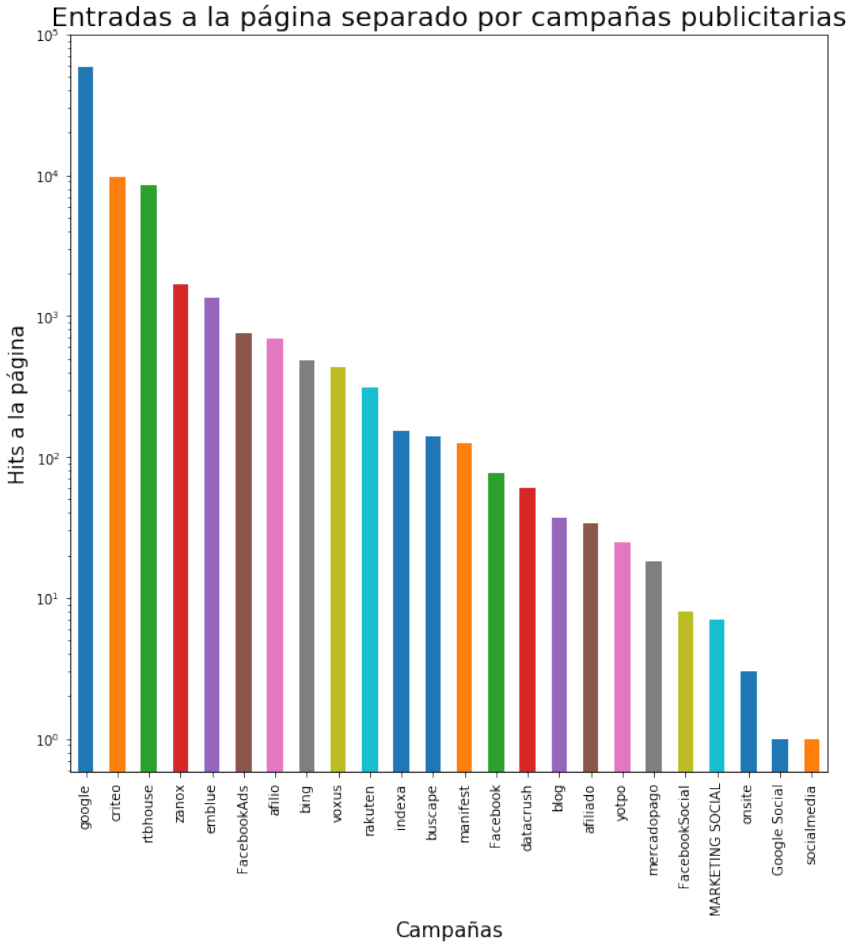


Podemos destacar que la mayoría de las entradas son hacia la página principal del sitio.

Claramente se ve que la campaña que más entradas provee es Google, seguida por Criteo y Rtbhouse (en menor medida). Siendo además Google la página que más variedad provee en las direcciones y Rtbhouse la única campaña que no ingresa a la pagina principal.

Se puede destacar que además de la pagina principal, las otras direcciones a las que más entran son:

- comprar/iphone
- comprar/iphone/6s
- comprar/iphone6
- comprar/samsung



En este gráfico podemos apreciar lo que habíamos mencionado, que las campañas más significativas son las de google, criteo y rtbhouse. Tuvimos que usar una escala logarítmica ya que la diferencia entre la cantidad de entradas desde la campaña de google es muy grande como para apreciarla de otra manera.

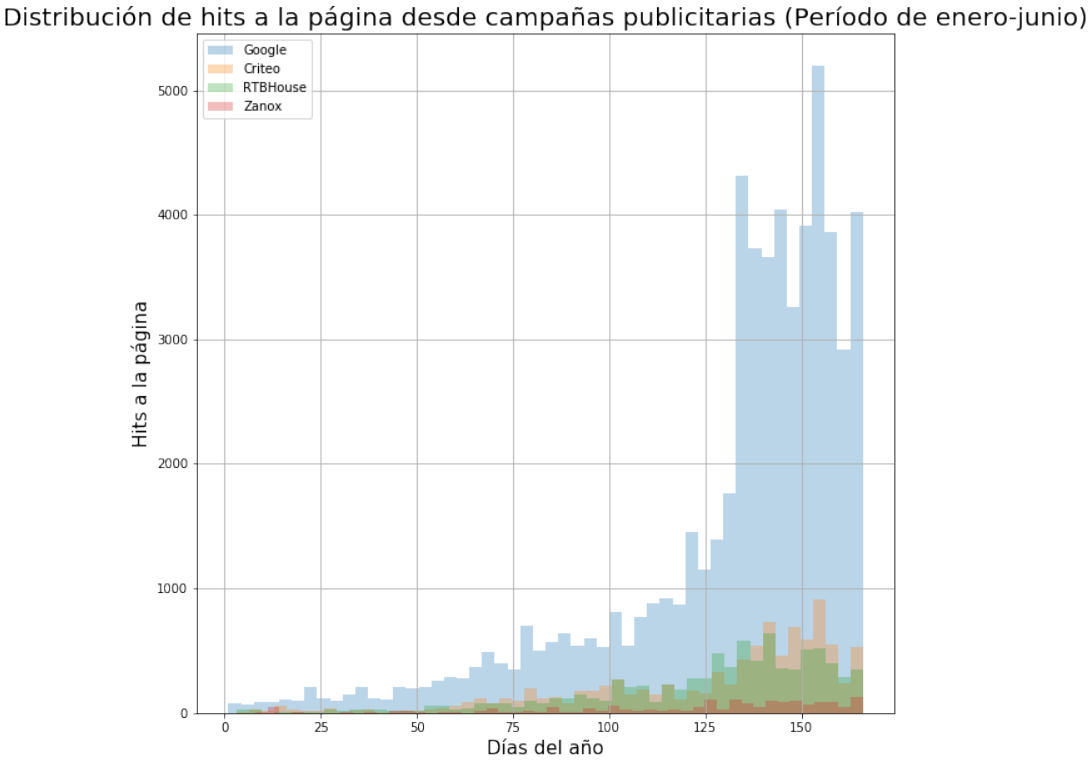
7.2. Aumento en el tráfico del sitio

En esta sección del informe vamos a analizar el efecto de las campañas publicitarias sobre el trafico del sitio. Para realizar este análisis entendemos como tráfico a las entradas al sitio, es decir, el tipo de evento "visited site". Ya que el resto de los eventos dentro del sitio surgen a partir de una entrada al mismo.

Primero, mostramos el aumento del trafico a lo largo de los meses.

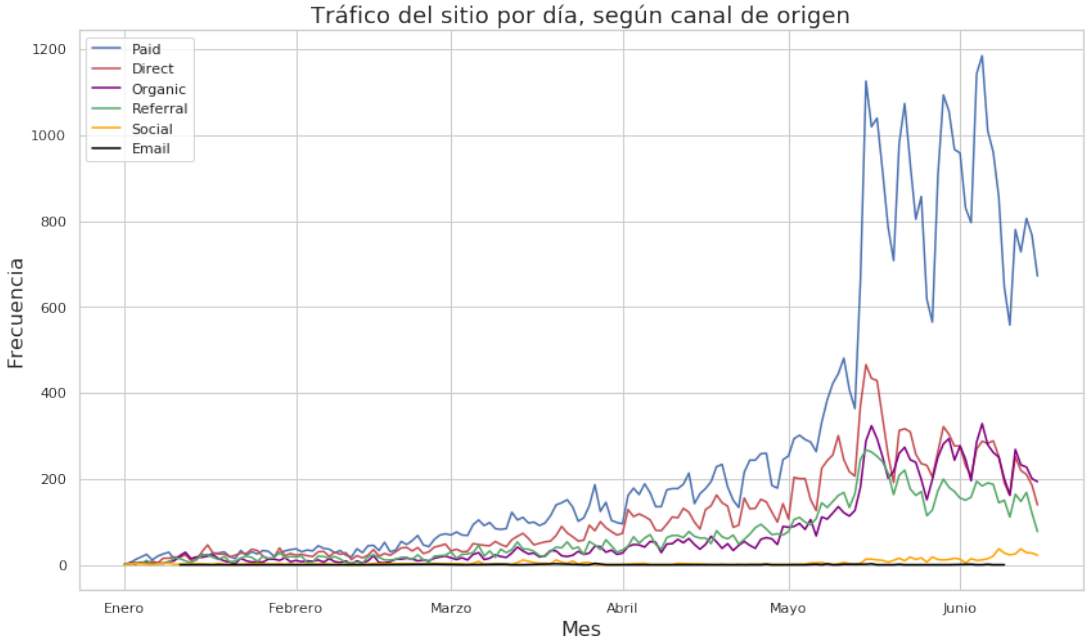
Como se puede observar, el sitio viene en constante crecimiento pero se da un fuerte aumento en las entradas a mitad de Mayo. Curiosamente, el aumento en entradas al sitio desde campañas se comporta de una manera muy similar.

Analizando la distribución de las entradas por campañas a lo largo de los meses nos encontramos con algo interesante, la mediana de las entradas ocurre en el mes de Mayo (para las campañas más significativas). Es decir, que tenemos la misma cantidad de entradas en el período Enero-Mayo que en Mayo-Junio. Decidimos hacer un gráfico de la distribución de las entradas a lo largo de los días del año para corroborar lo mencionado.



Lo que vemos a partir de esto es que coincide el aumento de tráfico a la página con el aumento de entradas por campañas publicitarias. El salto tan extremo podría llegar a explicarse por una mayor inversión en este tipo de campañas.

Otro fenómeno que sigue esta tendencia es las entradas al sitio por canales pagos.

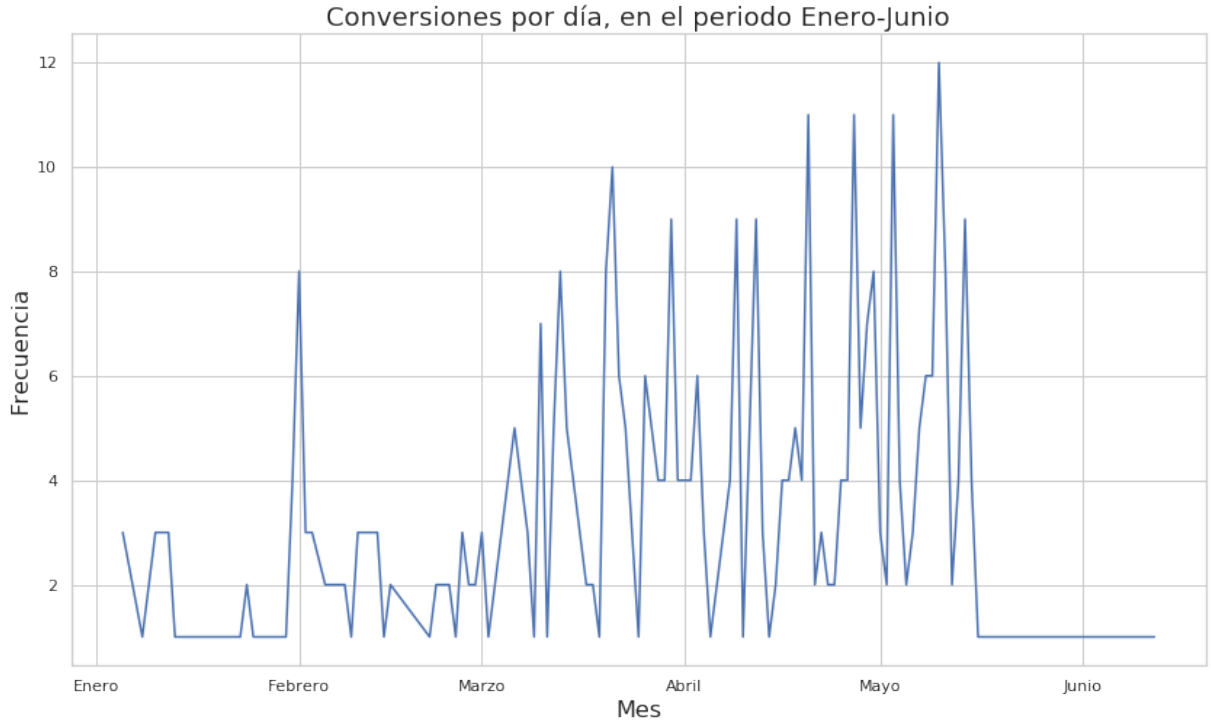


Nuevamente observamos que si bien viene en crecimiento constante, tiene un fuerte

aumento en el mes de Mayo.

A partir de la información recolectada, se podría explicar el aumento de entradas a la página como una combinación del aumento de entradas por campañas y canales pagos. Claro está que por estar limitados por un set reducido de datos no podemos afirmar que la inversión realizada en campañas fue la única causa del aumento en el tráfico de la página.

Por último nos gustaría analizar si el aumento de tráfico en el sitio podría llegar a mostrar un aumento en la cantidad de ventas. Para eso graficamos la cantidad de conversiones por día, a lo largo del período que estamos observando.



Si bién se puede observar un leve aumento en las ventas, no es tan notorio como para los fenómenos previamente mencionados. Podríamos decir que el aumento en el tráfico del sitio no se traduce tan directamente a un aumento en ventas.

Finalmente, podemos concluir que si bién la inversión en campañas publicitarias y entradas por canales pagos al sitio llevaría a un aumento sustancial en el tráfico de la página, no necesariamente correspondería a un aumento en las ventas del sitio.

8. Conversiones

Una conversion es un objetivo a cumplir. En el caso de Trocafone esto es equivalente a vender un producto.

Para este análisis, decidimos considerar aquellas conversions que fuesen realizadas **en el sitio**. La razón de esto es que a partir de la lectura realizada sobre el artículo de Medium, notamos que Trocafone realiza sus ventas por distintos canales de venta:

- Sitio e-commerce.
- Marketplace (páginas del estilo de MercadoLibre, por dar un ejemplo conocido).
- Tiendas Físicas.

Por lo tanto, tomaremos ciertas precauciones durante el filtrado de los datos para intentar quedarnos con las ventas que nos interesan según los criterios que describiremos a continuación.

8.1. Filtrado de los datos

Para obtener las ventas realizadas en el sitio, tuvimos las siguientes consideraciones:

- Aquellas conversiones que no presenten un checkout previo, por el mismo SKU (código de identificación de un producto), serán descartadas.
- A su vez, debe existir un delta de tiempo prudente entre checkout y conversion, el cual será de una hora en este caso.
- En caso de cumplirse las dos condiciones anteriores, pero estar los eventos presentes en orden inverso, los descartaremos por no poder determinar si es un error de cómo se obtuvieron los datos.

Por lo tanto en primer lugar obtendremos aquellos usuarios que presentan checkouts y conversiones.

Una vez obtenido este subconjunto de usuarios, filtraremos aquellos que no presenten checkout y conversion sobre un mismo SKU.

Notamos algo interesante,

	timestamp	event	person	sku	model
2547	2018-06-10 14:37:50	conversion	00fdbb4b	3348.0	Samsung Galaxy S6 Flat
2555	2018-06-11 01:47:34	checkout	00fdbb4b	3348.0	Samsung Galaxy S6 Flat
4244	2018-03-16 13:41:36	checkout	0146a9df	2694.0	iPhone 5s
4245	2018-03-16 13:50:25	conversion	0146a9df	2694.0	iPhone 5s
6984	2018-02-09 21:54:43	checkout	01db2fe6	6357.0	Samsung Galaxy J5

existen conversiones y checkouts sobre un mismo SKU pero que se encuentran en orden inverso, por lo tanto asumiremos que estos dos eventos no estuvieron relacionados. En este

caso en particular, se observa que además no cumplen la condición de tener un delta de tiempo de una hora.

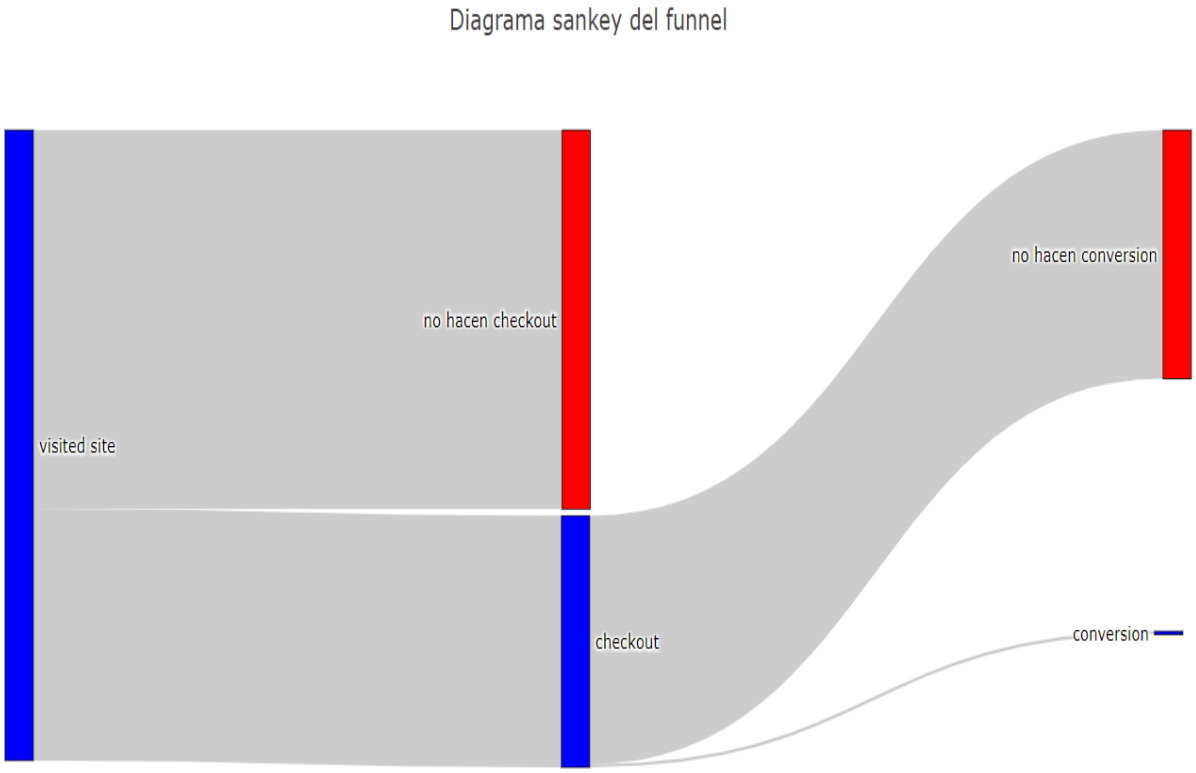
Finalmente tenemos un subconjunto de usuarios que cumple con los tres puntos estipulados anteriormente, y por lo tanto diremos que sus conversiones fueron realizadas en el sitio.

Los eventos que nos interesan para esto son Visited Site, Checkout y Conversion. De los cuales obtuvimos los siguientes números:

- Visited site: 79239
- Checkout: 31621
- Conversion: 393

8.2. Conversion Funnel

Presentaremos una visualización en la cual se puede observar qué proporción de visitas a la página terminaron en checkouts, y qué cantidad de checkouts llevaron a una compra confirmada.



Del total de visitas a la página, el 39.9 % termina en al menos un checkout, y de esta observación se despliega que solo el 1.24 % de los checkouts terminan en conversiones.

8.3. Observaciones

Finalmente, nos interesaría calcular cuál es el conversion rate del sitio de Trocafone, ya que ahora contamos con los datos suficientes como para realizar este cálculo.

Para calcular el mismo, solo debemos dividir el total de conversiones por el total de visitas al sitio, y así obtendremos la métrica.

Además agregaremos el porcentaje de conversiones que fueron realizadas en el sitio, según los criterios estipulados en esta sección.

- Conversiones totales: 1172
- Conversiones realizadas en el sitio: 393
- Conversiones realizadas fuera del sitio: 779
- Conversion rate: 0,49 %

De esto obtenemos que el 33.5 % de las compras parecen ser realizadas en el sitio, mientras que el otro 66.5 % se hacen fuera del mismo en los distintos Marketplaces de los que hace uso Trocafone.

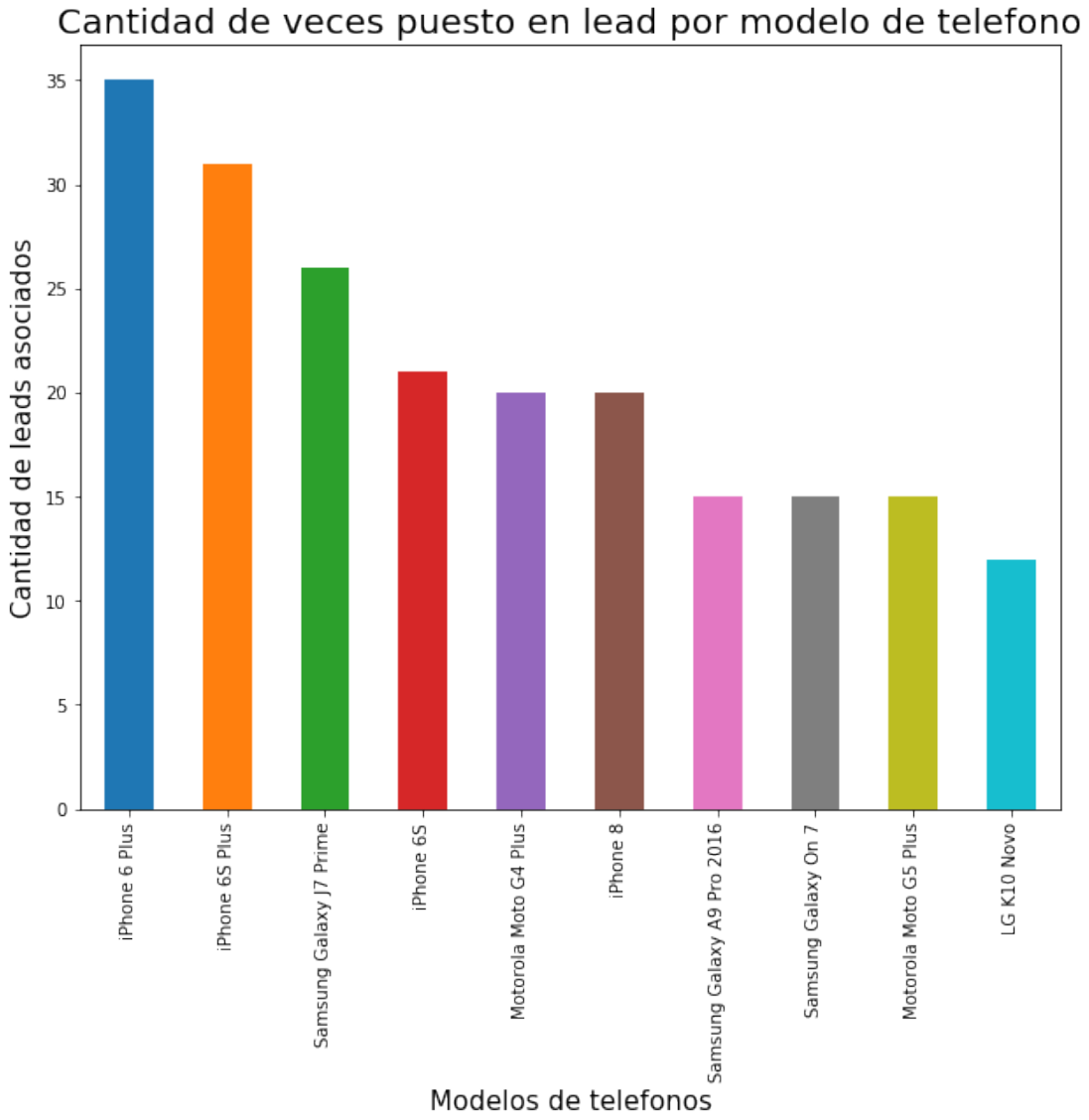
Además, concluimos que para el subconjunto de usuarios provistos en el dataset (el cual a su vez es un subset del original), la cantidad de conversiones realizadas en el sitio es baja de acuerdo al conversion rate obtenido.

A partir de este análisis no creemos que sea posible realizar recomendaciones para mejorar esta situación, ya que hay factores ajenos al dataset que intervienen en la decisión de los usuarios del sitio, a la hora de confirmar una compra.

9. Analisis sobre el uso de Leads

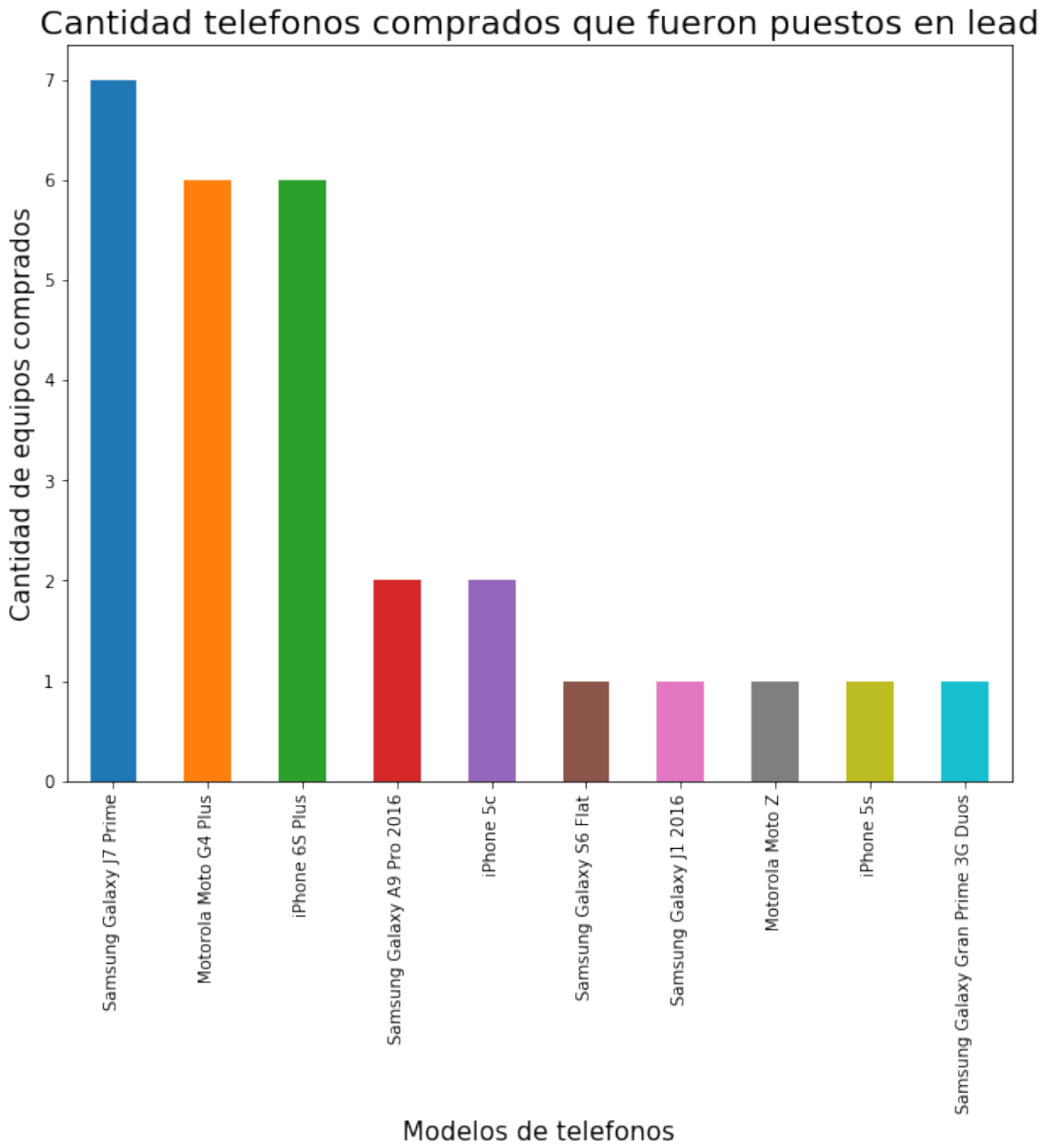
En esta seccion del informe vamos a analizar el uso de la feature del sitio que permite que le llegue una notificacion a un usuario cuando se renueva el stock de un modelo.

9.1. Top 10 modelos con más leads



Podemos ver que los telefonos con la mayor cantidad de leads son los más populares dentro de la página.

Pero ahora hay que analizar que tanto se traduce poner un teléfono en lead con hacer una conversión sobre el mismo.



Como era de esperar, hacer un lead sobre un modelo de télefono no necesariamente implica que el usuario va a terminar comprando ese equipo. Esto se puede ver en que la cantidad de leads sobre los telefenos es mucho mayor que la cantidad de compras sobre esos mismos modelos.

9.2. Tiempo de espera entre lead y conversion

Para hacer este análisis nos quedamos únicamente con los usuarios que hicieron un lead y luego terminaron comprando ese producto.

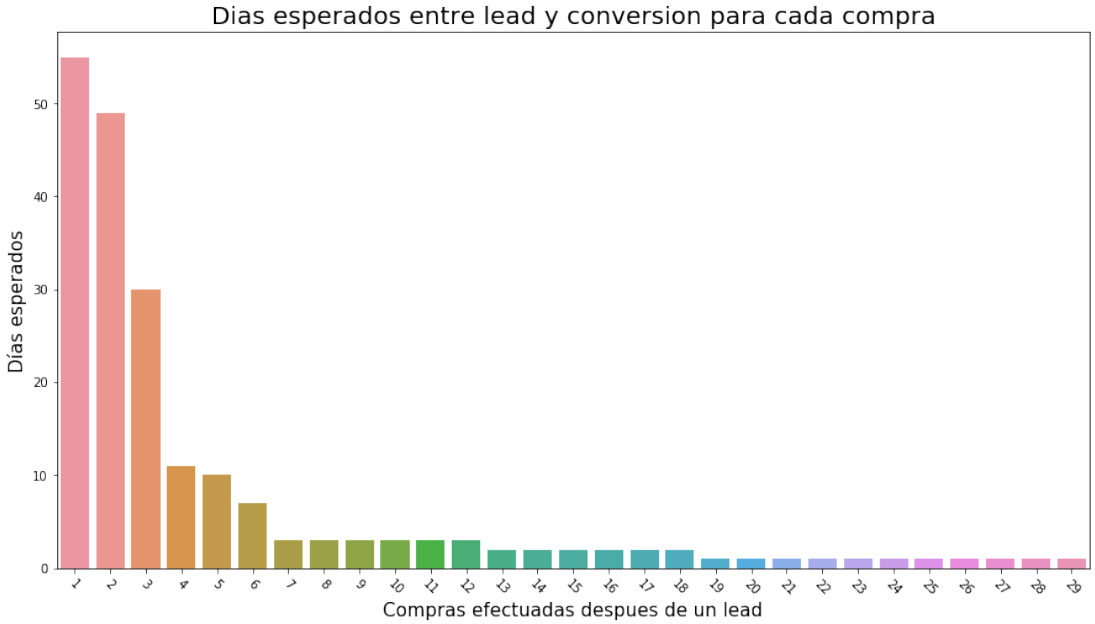
Al hacer un describe sobre el DataFrame, nos encontramos con estos datos (en días):

- mean: 7.000000

- std: 13.711309
- min: 1.000000
- 25 %: 1.000000
- 50 %: 2.000000
- 75 %: 3.000000
- max: 55.000000

Lo que podemos ver rapidamente es que la media de los días esperados a comprar un producto que fue puesto en lead es de 7, es decir, en promedio los usuarios esperan 7 días para efectuar la compra.

Pero al analizar los datos más detalladamente, podemos ver que el 75 % de las personas esperaron 3 días o menos para poder hacer la compra del producto. Entonces tenemos algunos datos anómalos que nos estan afectando el promedio.



Efectivamente, solo tenemos 5 compras que tardaron más de 10 días en efectuarse.

Luego, podemos ver que la mayoría de las personas que usan la feature de lead y efectivamente terminan comprando, lo hacen mayormente dentro de los primeros 3 días.

Como no tenemos datos sobre stock y sobre cuando se les mandaron las notificaciones, no podemos encontrar la causa de la larga espera de algunos de los usuarios. Podría ser que no se reponga el stock del teléfono o bién el usuario podría haber pospuesto la compra por otras razones.

En conclusión, si bién el uso de la feature no necesariamente va a terminar en una compra, el 75 % de los usuarios que toman ventaja de la misma y terminan comprando, lo hacen dentro de los primeros 3 días.

10. Términos de búsqueda

En esta sección vamos a analizar los términos de búsqueda ingresados por los usuarios del sitio de Trocafone.

Antes de realizar cualquier tipo de observación, será necesario limpiar los datos un poco, ya que los errores ortográficos son moneda corriente y podrían llegar a distorsionar el análisis.

10.1. Corrección de errores comunes

Primero descartaremos aquellos términos de búsqueda que se encuentren nulos, y también eliminaremos de cada uno de ellos todo tipo de signo de puntuación. Luego pasaremos a lower case todos los términos, para así evitar hacer una diferencia entre palabras iguales, como por ejemplo sería el caso de **iPhone** y **iphone**.

Nos ocuparemos de los errores más frecuentes utilizando la distancia de Levenshtein como métrica para determinar cuándo debemos realizar una corrección.

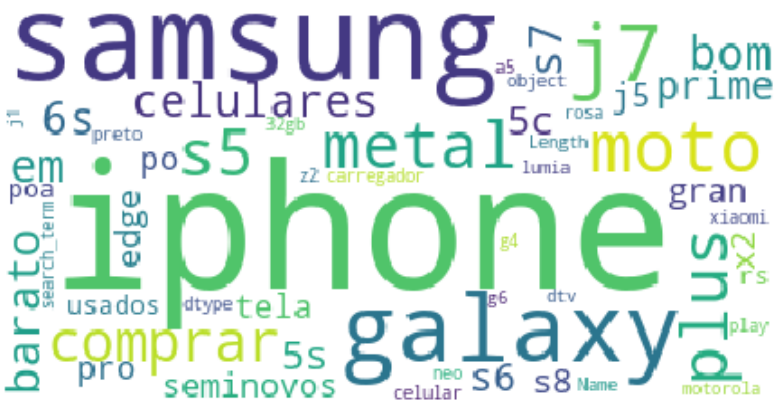
La distancia de Levenshtein se entiende como el número mínimo de operaciones que se deben realizar sobre una cadena para transformarla en otra. La misma se presenta como un valor entre 0 y 1, por lo que consideraremos que aquellas palabras que presenten un valor mayor o igual a 0.7 serán tomadas como un error y las corregiremos.

Por supuesto esto presenta algunos problemas en el caso de palabras como celular y celulares, ya que si bien son distintas, una de ellas será considerada errónea y por lo tanto se modificará, por lo cual las excluiremos de la lista de errores comunes y solo nos quedaremos con los nombres propios de celulares/marcas.

10.2. Visualizando la frecuencia de los términos de búsqueda

Ahora que terminamos con la limpieza de los términos de búsqueda, utilizaremos dos visualizaciones para ver rápidamente las frecuencias de los mismos.

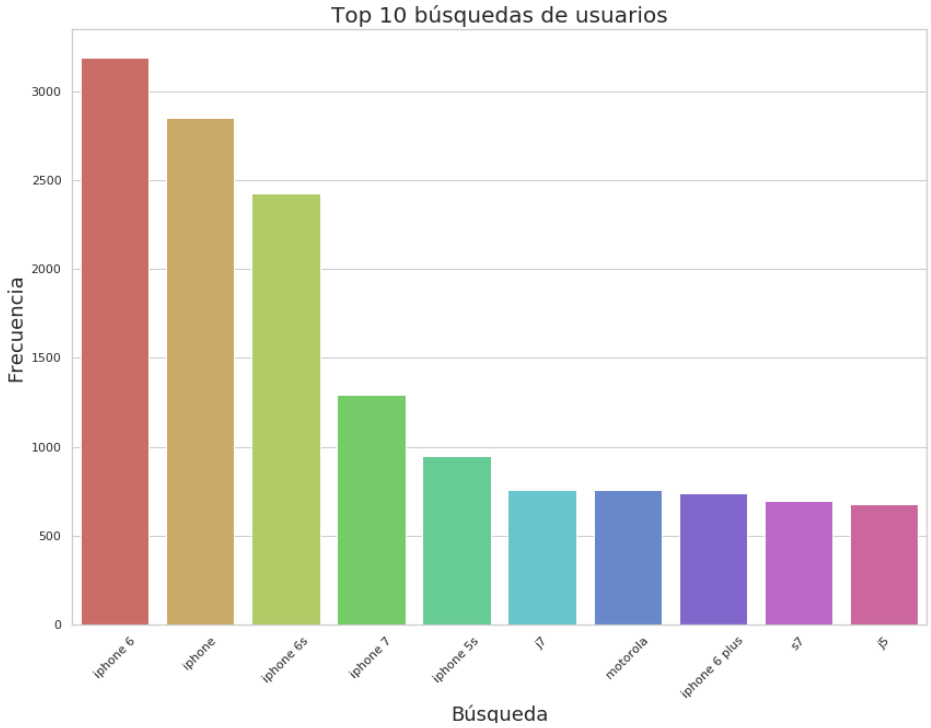
En primer lugar, presentaremos un Wordcloud para tener una idea rápida de cuáles son los términos más frecuentes.



Como podemos observar, las palabras que parecieran presentar mayor frecuencia son las esperadas (las dos marcas más populares).

Luego podemos observar distintos modelos de ambas marcas, y también se aprecian términos como *celulares*, *barato*, *comprar* y otras relacionadas a las condiciones de los productos.

Ahora veamos cuál es el top 10 de términos de búsquedas registrados en el dataset.



Podemos observar que el 60 % del top 10 está dominado por búsquedas de productos Apple, 30 % términos relacionados a Samsung y 10 % Motorola. Esto está relacionado con los datos analizados en las características de los equipos, donde observamos una tendencia similar en ventas y vistas de productos.

10.3. Relevancia de los términos de búsquedas

Al ver que contabamos con la información relacionada a las búsquedas realizadas por los usuarios, surgió la idea de intentar medir qué tan relevante son los resultados devueltos por el buscador para los términos de búsqueda más usados.

10.3.1. Consideraciones para el análisis de relevancia

- Tendremos en cuenta que si el buscador de Trocafone no es capaz de relacionar casos como **yphone** con **iphone**, es posible que el puntaje asignado al término **iphone** sea menor que el real, a causa de cómo decidimos manipular los datos durante la sección de limpieza.
- Decidimos tomar los 100 términos más frecuentes para esta tarea ya que la distribución de frecuencias no es uniforme, puesto que hay demasiados términos de búsqueda que presentan muy baja frecuencia.

10.3.2. Método de asignación de puntajes

Para poder asignar puntajes a cada término utilizaremos una métrica conocida como **term frequency** la cual recompensa o adjudica mayor relevancia a aquellos términos que aparecen con mayor frecuencia en un texto.

A diferencia del análisis de palabras claves en documentos, donde se utiliza la métrica **inverse document frequency** en conjunto a **term frequency** para premiar a aquellas palabras que además de ser frecuentes, también son raras (y podrían resultar más interesantes), aquí solo nos interesa qué tan frecuentemente aparecen estos términos en los resultados de las búsquedas.

En nuestro caso utilizaremos los resultados de cada búsqueda, y haremos un promedio de la frecuencia con la que aparece el término en los mismos. Ee esta forma asignaremos los puntajes a cada uno de los términos de búsqueda.

10.3.3. Preparación de los datos

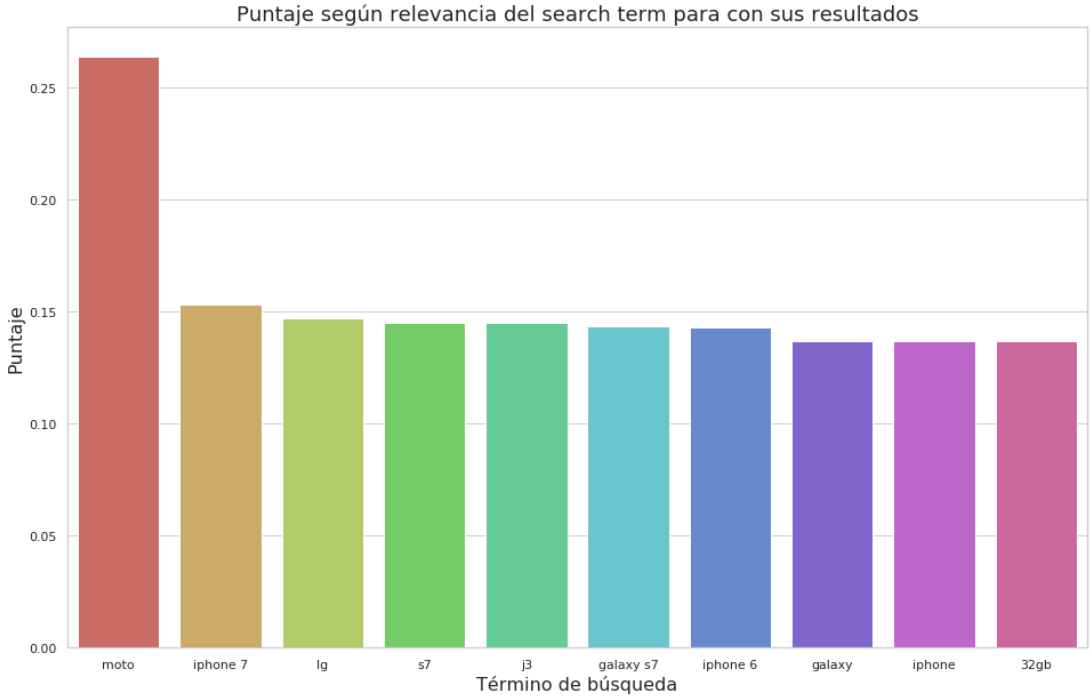
Primero obtendremos las características asociadas a cada SKU, ya que el resultado de una búsqueda es un listado de SKUs.

Esto lo pasaremos a un diccionario para luego transformar ese listado de SKUs en cadenas con las características del modelo asociado.

Luego filtraremos aquellos términos que no tengan un listado de SKUs asociado y agregaremos una columna con las características de cada SKU devuelto por el buscador.

10.3.4. Asignación de puntajes

Finalmente prepararemos la columna **resultados_busqueda** para ser procesada en el cálculo del puntaje.



Los resultados obtenidos nos indican que **moto** es un término muy performante (aunque esto podría atribuirse a la simpleza de nuestro análisis).

Por otro lado, vemos que los puntajes asignados al resto de los integrantes del top 10 son parejos. Entre ellos encontramos algunos de los términos más utilizados, y nos encontramos con la sorpresa de que el término **32gb** haya rankeado alto, siendo que no es un nivel de almacenamiento muy popular como observamos en secciones anteriores del informe. LG también se hace presente como término performante, lo cual también es llamativo ya que no hay una gran demanda de estos productos. Esto podría deberse a que hay menor disponibilidad de modelos y por lo tanto la búsqueda devuelve resultados más acotados, por lo cual el **term frequency** correspondiente es más alto.

10.4. Otro enfoque para analizar los términos de búsqueda

En esta sección, intentaremos ver estos datos de otra forma. Para esto analizaremos la frecuencia de n-gramas en estas búsquedas, en particular bigramas y trigramas. De esta forma tal vez podamos encontrar algo interesante que nos ayude en la siguiente etapa del trabajo práctico de la materia.

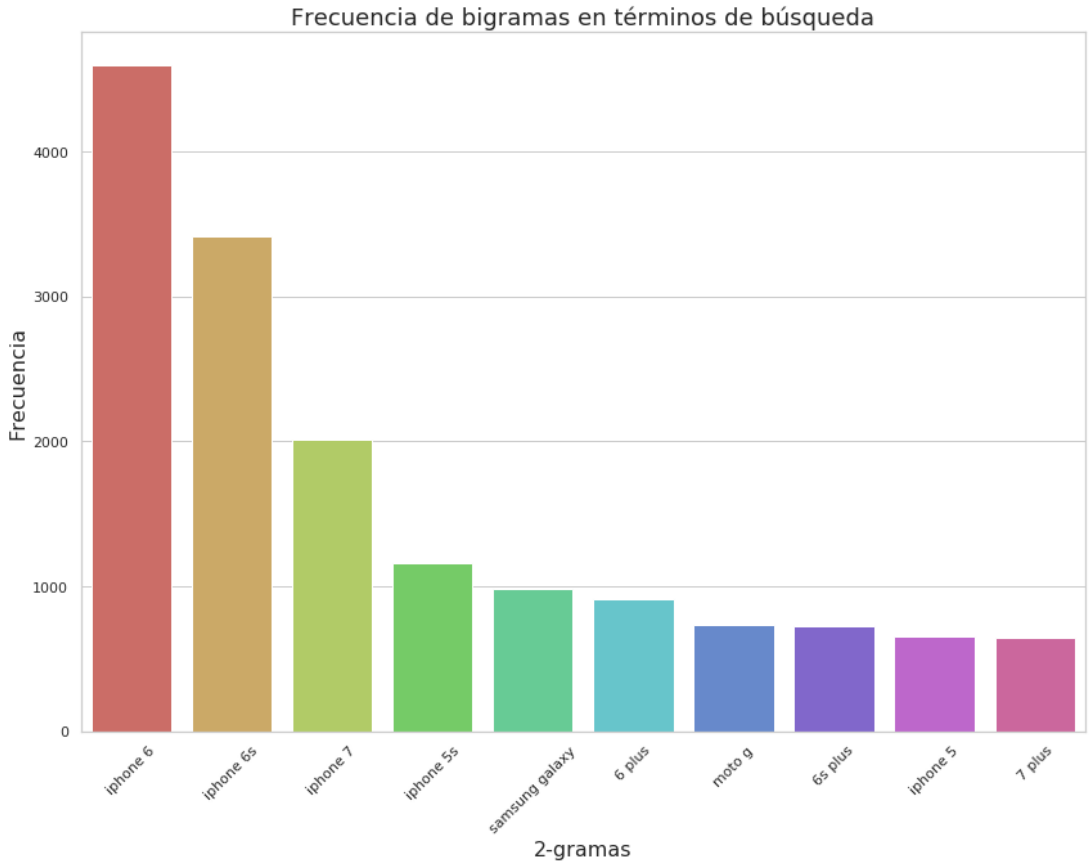
10.4.1. Consideraciones previas

En primer lugar, debemos filtrar las stopwords del lenguaje que presenta mayoría (en este caso, portugués). El fin de esto es prevenir que aparezcan en nuestros reportes palabras de uso común pero que carecen de contenido, como por ejemplo por, de, mais.

Luego utilizaremos la biblioteca **nltk** para poder procesar cada término de búsqueda y extraer los n-gramas que nos interesen.

Finalmente presentaremos la información por medio de gráficos de barra, ya que nos interesa ver los n-gramas que presenten mayor frecuencia.

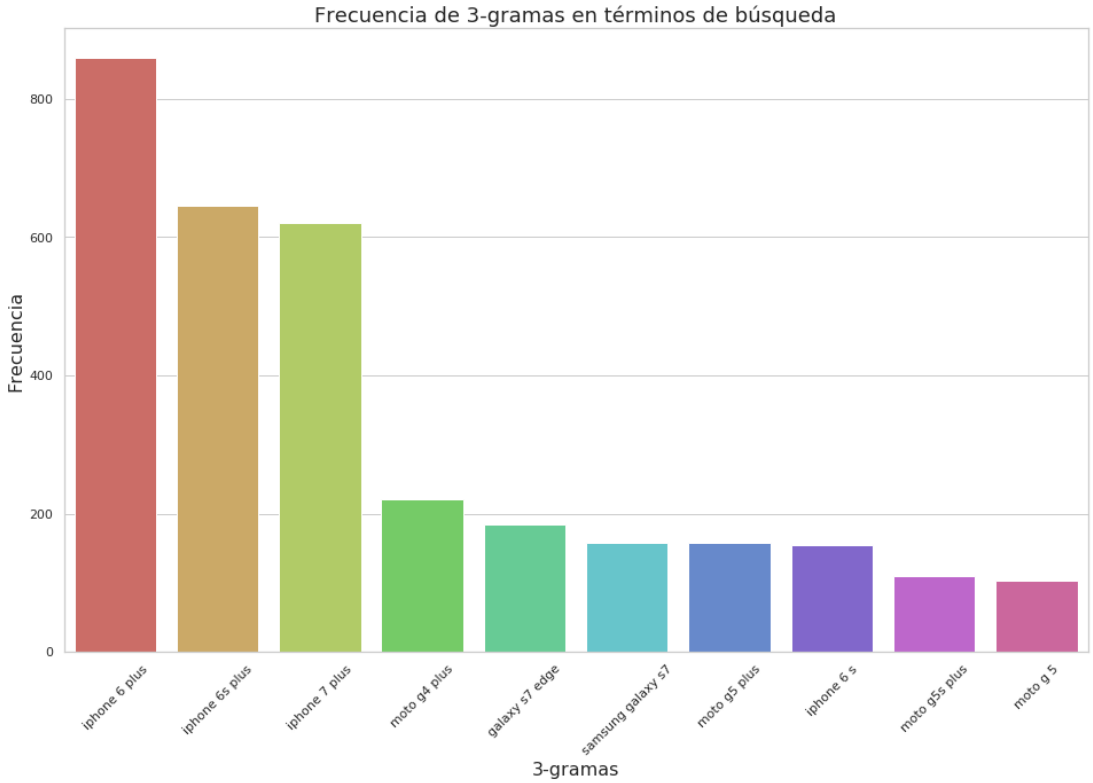
10.4.2. Bigramas



Tal vez no tan curiosamente, los bigramas más frecuentes parecen coincidir con los celulares más vistos (analizados en la sección Características de Celulares). Es notable la

presencia de bigramas relacionados a la marca Apple, estando presente en el 80 % del top 10 presentado.

10.4.3. Trigramas



Apple pierde impacto en los trigramas, dando lugar a la aparición de dos trigramas que componen **Samsung Galaxy s7 Edge**, así como también otros tres relacionados al modelo **Moto G5 Plus**.

10.5. Conclusión general

En el primer análisis realizado notamos que las variantes del término **iphone** son muy frecuentes en las búsquedas de los usuarios. Luego al calcular su relevancia en los resultados que obtuvieron de parte del buscador de Trocafone, nos llamó la atención que la performance de **iphone** no fuese la más alta. Esto podría llegar a deberse como ya hemos mencionado anteriormente, a la limpieza realizada sobre los datos. De ser así, esta observación nos lleva a la siguientes preguntas,

- ¿Cómo opera el buscador de Trocafone?
- ¿Realiza algún tipo de corrección en los términos de búsqueda?

Luego de realizar algunas búsquedas en el sitio utilizando términos como **yphone**, **iPphone**, y similares, notamos que el buscador devuelve siempre los mismos resultados

al no encontrar un match para el término, un listado de los celulares más vendidos por Trocafone.

Por lo tanto nos gustaría sugerir, de ser posible, la implementación de un sistema de corrección de errores. El mismo operaría tomando aquellos términos de búsqueda que no produzcan resultados en primera instancia (por no coincidir directamente a algún producto), y los compare contra un pequeño listado de tags seleccionados a partir de un análisis similar al que hemos realizado en esta sección, y devuelva un resultado acorde al tag que coincida con el término de búsqueda.

11. Conclusiones

A continuación, enumeraremos los puntos claves de este informe, a modo de conclusión del mismo.

- La empresa Apple es, en lineas generales, la más popular del sitio seguida de Samsung, sin embargo Samsung es la que más equipos vendidos tiene.
- Los celulares con 16GB son los que más ventas tienen, seguidos por los que poseen 32GB, pese a que su popularidad es casi idéntica.
- La gama de colores negros y blancos son los más populares entre los equipos vendidos.
- Los celulares con estado 'Bueno' son los más vendidos, seguido de los estados 'Muy Bueno' y 'Excelente'.
- São Paulo, seguido por Minas Gerais y Bahia son las regiones de Brasil que más actividad poseen.
- La actividad de Trocafone aumentó considerablemente en el mes de Mayo, el cual coincidió con el crecimiento de entradas por canales pagos y campañas publicitarias.
- La campaña más efectiva fue la de Google, seguida por la de Criteo y RTBHouse.
- Muchos usuarios realizan checkouts y no realizan luego la compra del equipo, y se muestran compras sin un checkout previo, por lo que esta actividad no nos indica certeramente si el usuario comprará o no el producto.
- No hay en el sitio días de la semana en dónde se realizen significativamente más visitas que en los demás, pero si los hay en ventas, dónde el Sábado y el Domingo son los días con menor cantidad.
- Los horarios dónde se observa más actividad son los que se encuentra entre las 16 y 24 hs, siendo la mañana hasta las 16hs los momentos de menos tráfico, lo cual coincide con el típico horario laboral.
- Los leads son una característica de la página que no tiene tanto uso, sin embargo quienes lo utilizan y terminan comprando el equipo, lo hacen generalmente en los primeros 3 días de solicitarlo.
- Errores ortográficos pequeños pueden producir que el producto no sea hallado por el sitio web y se muestre simplemente una lista de los celulars más vendidos.
- De las visitas realizadas a la página, solo el 39.9 % lleva a realizar al menos un checkout. Mientras que de esa cantidad de checkouts, solo el 1.24 % termina en una conversion.
- Siendo el total de conversiones 1172, y por el análisis realizado, el 33.5 % de estas se realiza en la página mientras que el otro 66.5 % surgen de los Marketplaces.