Trabajo Práctico 1

[66.20] Organizacion de Datos
Primer cuatrimestre de 2019

Grupo 11		
Nombre	Padrón	Mail
Eliana Gamarra	100016	elianagam2@gmail.com
Ailen Garcia	100560	ailu_tunci@hotmail.com
Martín Suarez	101540	martinsua24@gmail.com
Ailín Braccelarghe	99366	ailinyelenb@gmail.com

Índice

1. Introducción

A partir de los DataFrame proporcionados por Jammp, se analizan la información de cada una de ellos y luego se realiza una integración con la información valiosa que se encuentra en común.

Link al repo con el analisis: https://github.com/Elianagam/orgaDatosTp1/

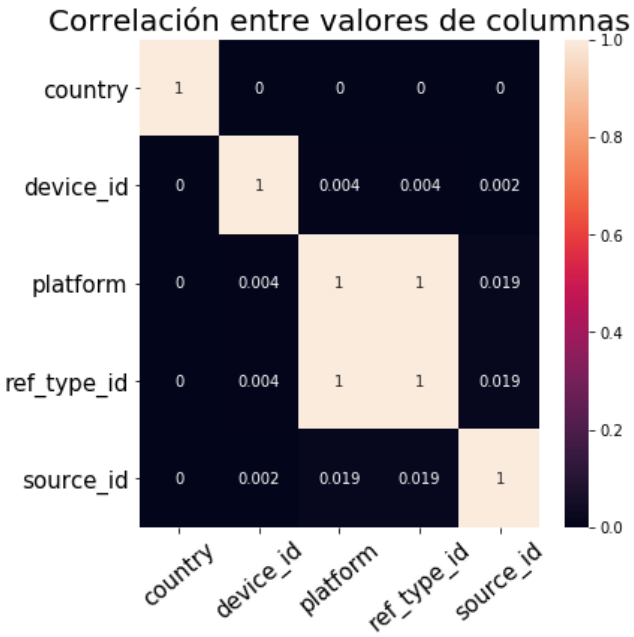
2. Auctions

El csv se compone de 7 columnas y más de 19 millones de filas. Al comienzo se realizó un análisis breve de la información y se determino lo siguiente:

- La columna 'auction_type_id' posee todos valores nulos
- La columna 'country' posee siempre el mismo valor
- La columna 'platform' y 'ref_type_id' están correlacionadas. 'Ref_type_id' posee un 1 o un 7 según si platform posee un 1 o un 2 respectivamente.

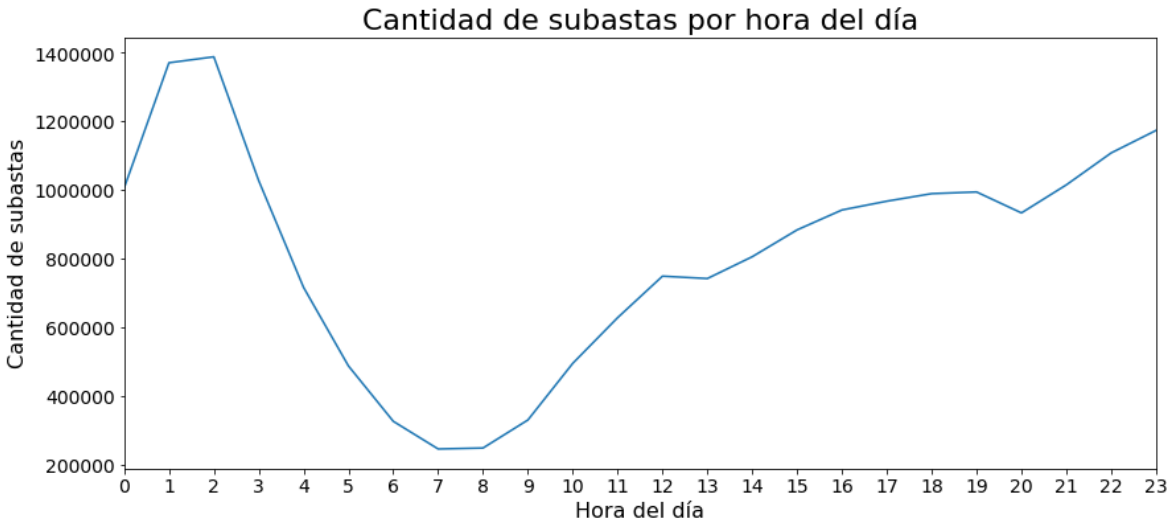
2.1. Correlación entre columnas

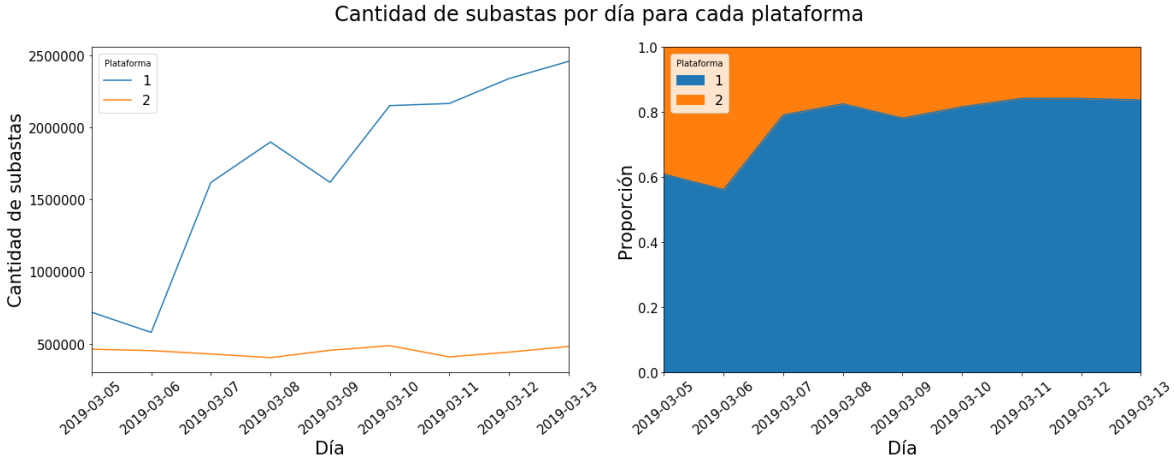
Se utilizó un heatmap para comprobar fácilmente si existe alguna correlación entre los valores de las columnas. De esta manera descubrimos que las columnas 'platform' y 'ref_type_id' están totalmente correlacionadas. Más alla de eso, no parece haber correlación entre ningún otro par de columnas.

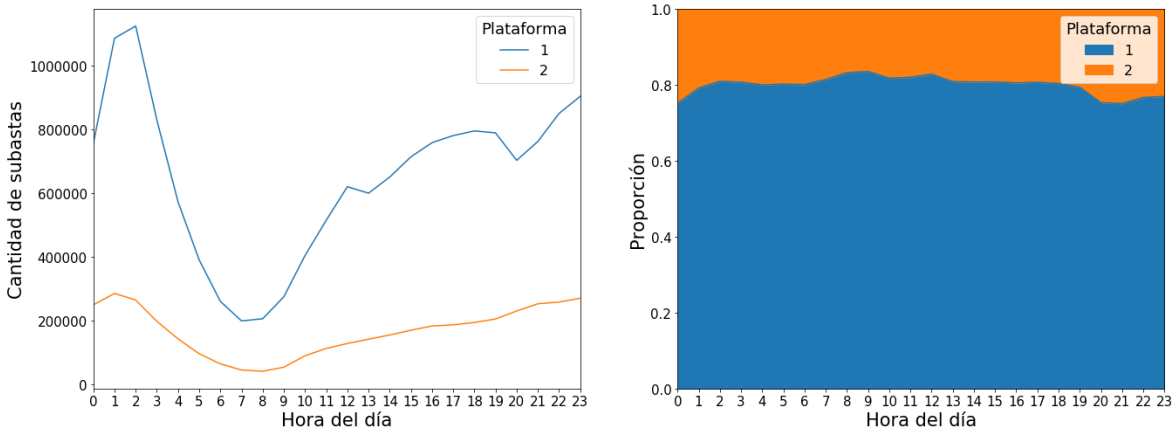


2.2. Subastas por día y hora

Agregamos columnas que discretizan la columna 'date' en hora y en día. A partir de eso, creamos gráficos que representan la distribución de subastas a lo largo de días y horas.







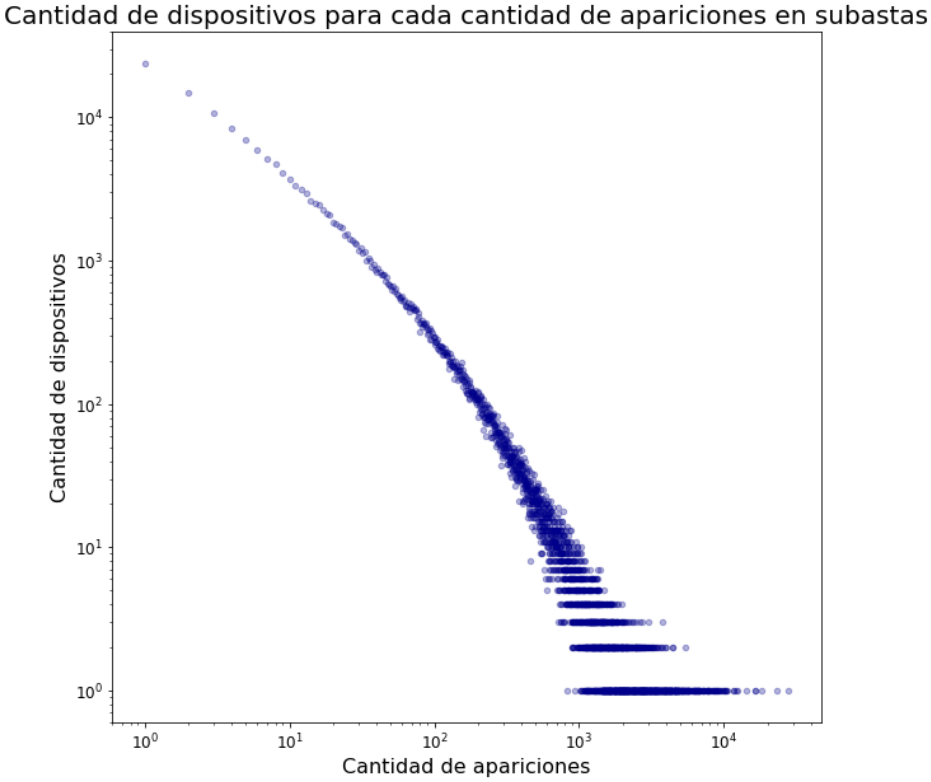
A partir de estos gráficos podemos concluir que la mayoría de las subastas provienen de la plataforma 1 (alrededor del 80 %). Además podemos ver que entre las horas 21 y 3 las subastas son bastante frecuentes, alcanzando un pico aproximadamente en la hora 2. Luego, hay una fuerte reducción de subastas llegando a mínimos entre las 5 y 10. Más allá de esto, no es posible concluir demasiado, el comportamiento parece ser bastante normal.

2.3. Evolución de subastas para los cinco dispositivos que más aparecen

Gráficamos la evolución a lo largo de los días en la cantidad de subastas para los cinco dispositivos que más aparecieron en subastas. Podemos observar que tiende a haber más subastas con el pasar de los días, pero no podemos obtener demasiadas conclusiones.

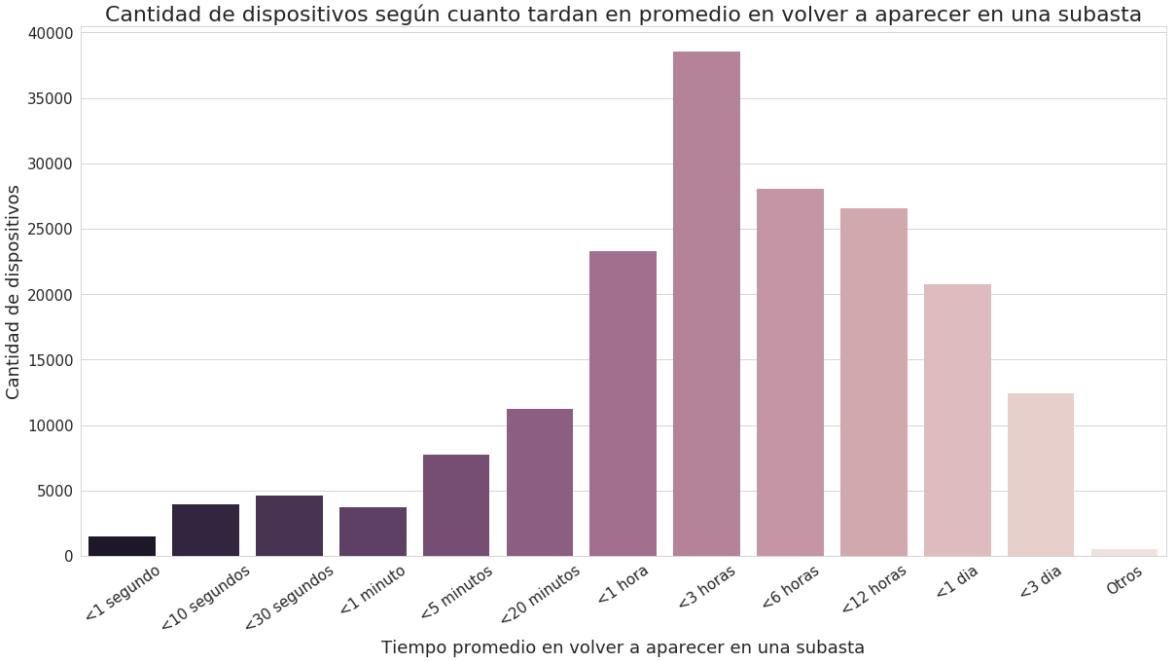
2.4. Ley de potencias en la distribución de apariciones por dispositivo

Al analizar la información, notamos que pocos dispositivos aparecían una gran cantidad de veces y muchos dispositivos aparecían pocas veces. Graficamos este análisis en escala logarítmica y comprobamos que sigue el patrón de la ley de potencias ya que prácticamente forma una linea recta.



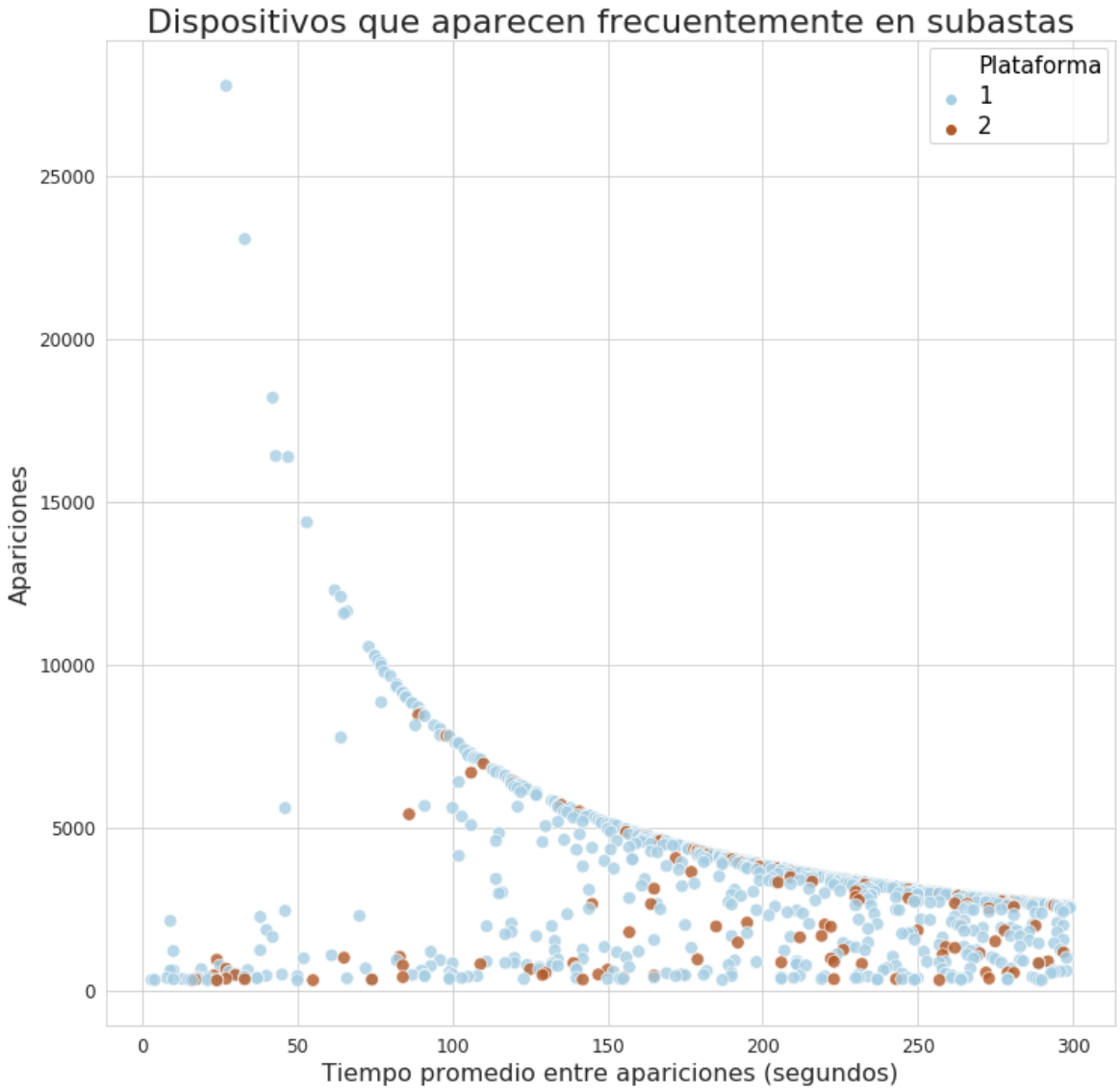
2.5. Tiempo promedio de reaparición de un mismo dispositivo

Analizamos para cada dispositivo cuanto tarda en promedio en volver a aparecer en una subasta a partir de su última aparición. A partir del siguiente gráfico de barra podemos entender la distribución de tiempos promedios.

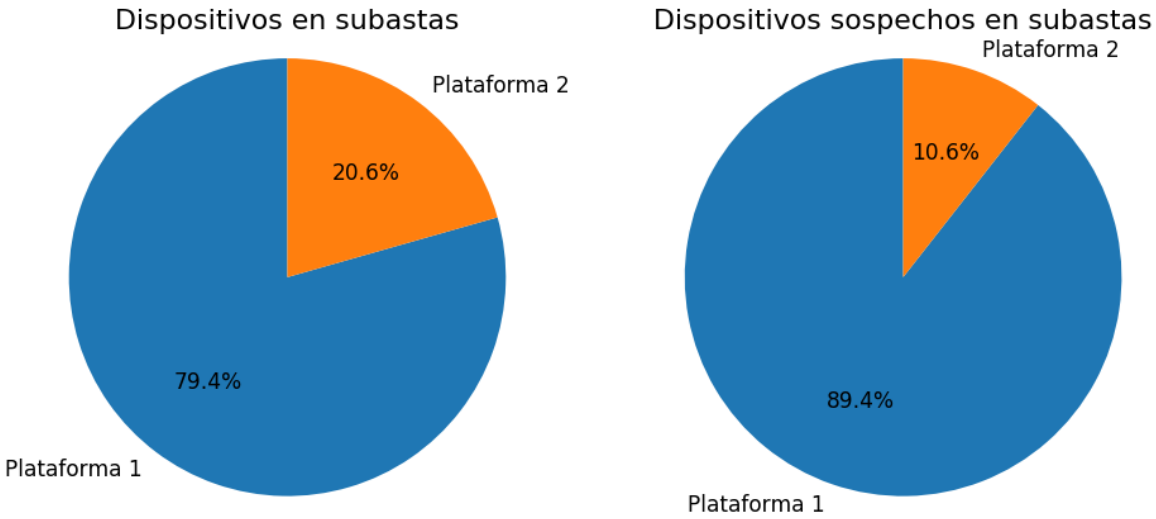


2.6. Análisis de registros falsos

Con el último gráfico podemos apreciar que hay una cantidad sustancial de dispositivos que en promedio aparecen en promedio cada un minuto o menos. Esto, a priori, puede ser un comportamiento sospechoso. Para analizar más en detalle la situación, tomamos los dispositivos que aparecen al menos 300 veces y con un promedio de tiempo entre apariciones menor a 5 minutos. Con esta información creamos un scatter plot en el que diferenciamos según la plataforma correspondiente al dispositivo. En el gráfico apreciamos que los puntos configuran una parábola y una franja inferior. Una conclusión rápida que podemos obtener es que los dispositivos de la plataforma 1 parecen tener los comportamientos más sospechosos. Estos se ubican en la parte izquierda de la parábola, demostrando que tienen una gran cantidad de apariciones con tiempos promedios entre apariciones muy bajos. Esto nos indicaría que es esta plataforma más propensa a ser usada de manera fraudulenta.



Hay 891 dispositivos distintos que cumplen estar en el rango de aparecer más de 300 veces con un promedio menor a 5 minutos, totalizando 2.849.470 apariciones. Esto representa un 14,56 % de las apariciones totales en subastas.

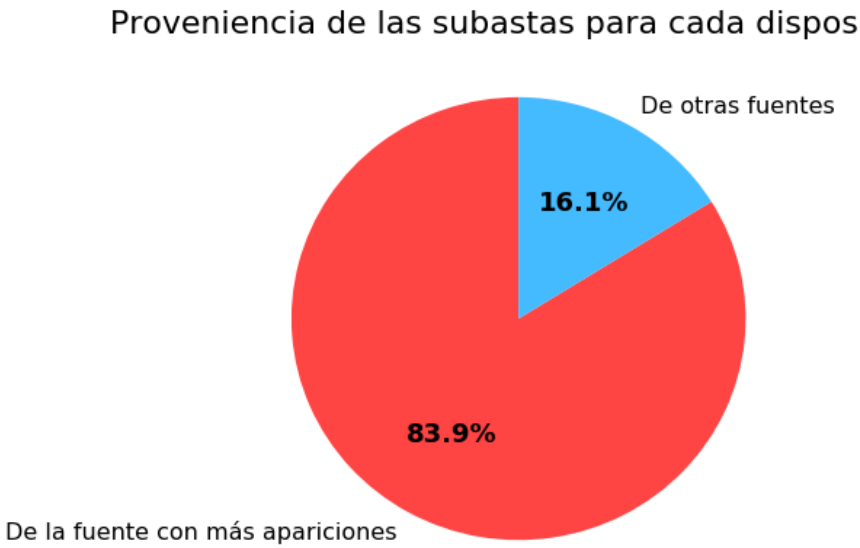


Analizando la subastas por su plataforma, podemos ver que en los casos fraudulentos hay una mayor proporción provenientes a la plataforma 1 en comparación a la totalidad de las subastas. Esto confirma la teoría que los dispositivos de la plataforma 1 tienen mayor tendencia a ser usados de manera fraudulenta.

Igualmente este análisis es aproximado y es útil únicamente para dar una idea de la situación. Es subjetivo que se podría considerar como un comportamiento fraudulento y que no.

2.7. Análisis de las fuentes de subastas

En esta sección queremos analizar si hay alguna relación entre la fuente de la subasta ('source') y los dispositivos. Partimos de la hipótesis que los usuarios generalmente usan cierta aplicación con mayor frecuencia. Si esta aplicación esta vinculada a una fuente de subastas determinada, veríamos una relación entre dicho dispositivo y la mencionada fuente de subastas. Por ejemplo, un usuario, en su celular usa frecuentemente una red social y ocasionalmente un juego. Cada vez que la aplicación de la red social puede mostrarle una publicidad, se lo comunica a su fuente de subastas (digámosle 'Fuente A'). Por otro lado, el juego hace lo mismo pero con su fuente ('Fuente B'). Si nuestra hipótesis es correcta, veríamos que en las subastas correspondientes a este usuario, hay una predominancia de la 'Fuente A' por sobre la 'Fuente B'. A partir de esto, clasificamos la información sobre subastas en dos grupos: las que provenían de la fuente más frecuente del usuario y las demás. Siguiendo con el ejemplo, digamos que nuestro usuario tuvo 7 subastas de la 'Fuente A', 2 de la 'Fuente B' y 1 de una nueva fuente, 'Fuente C'. Con nuestro criterio, el usuario provino en un 70 % de su fuente más frecuente y en un 30 % de otras fuentes.



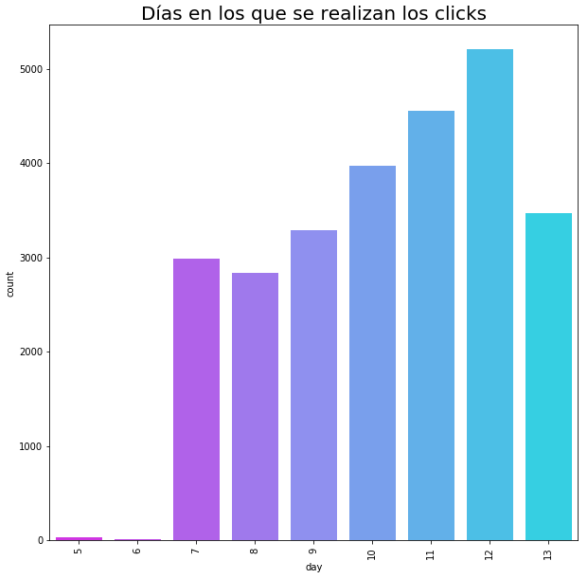
Como podemos ver, las subastas para cada dispositivo suelen concentrarse en una cierta fuente. En promedio, más del 80 % de las subastas provienen a partir de una única fuente.

3. Clicks

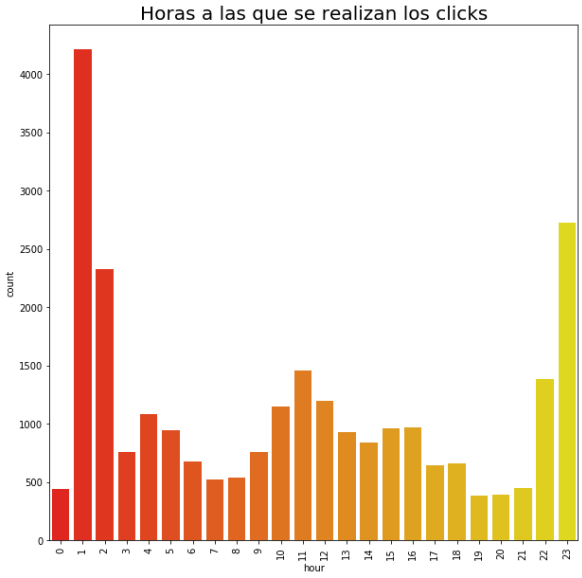
Al comenzar a analizar clicks, notamos que la columna 'action id' contenia puros NaN, por lo cual decidimos eliminarla. Notamos que la columna 'wifi connection' sólo posee el valor False, por lo cual la eliminamos.

3.1. Días y horas en que se realizaron clicks

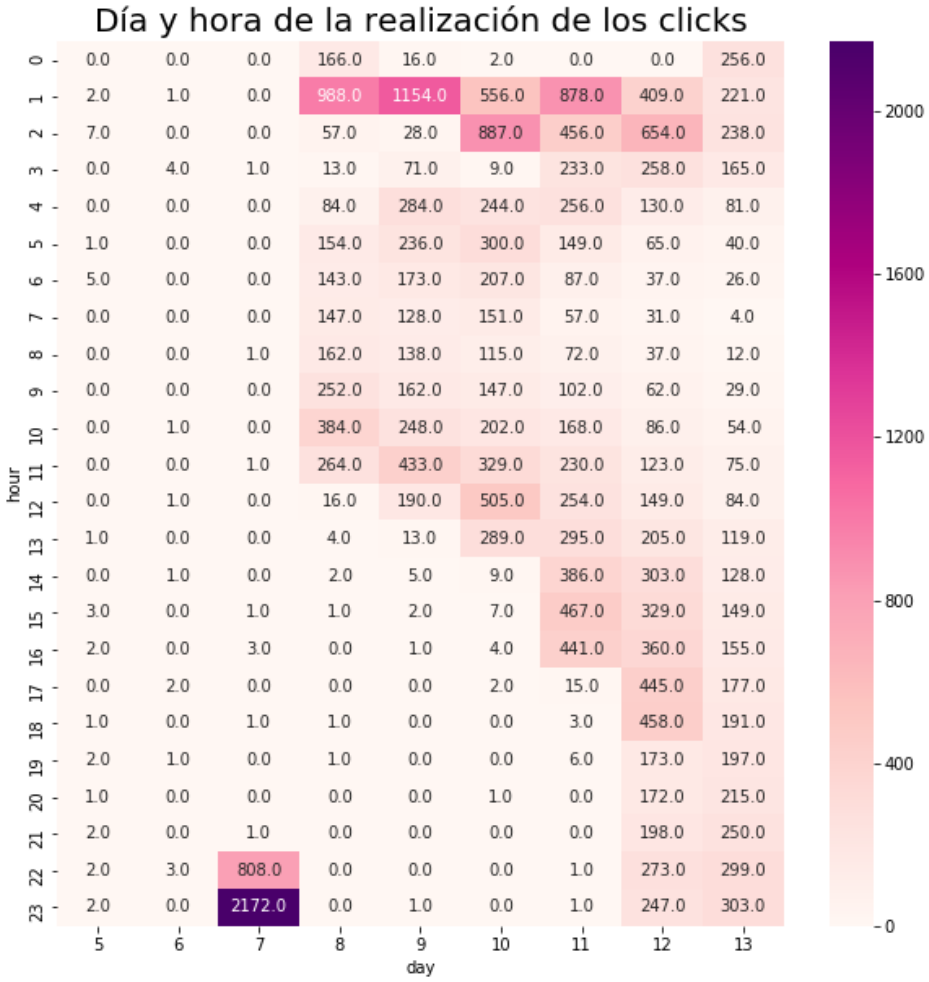
Como vimos que la columna 'created' posee el mismo año y mes en todas sus fechas, comenzamos a analizar aspectos relacionados con los días y horas, como por ejemplo, cuál es la frecuencia de realización de clicks durante los días, con eso obtuvimos que el día en que se realizaron más clicks fue el día 12/03/2019, y el día en que se realizaron menos clicks fue el día 06/03/2019.



Hicimos lo mismo con respecto a la hora, y obtuvimos la conclusión de que a la 1:00 am se realizaron más clicks, y a las 19 pm fue la hora a la que se hicieron menos clicks.

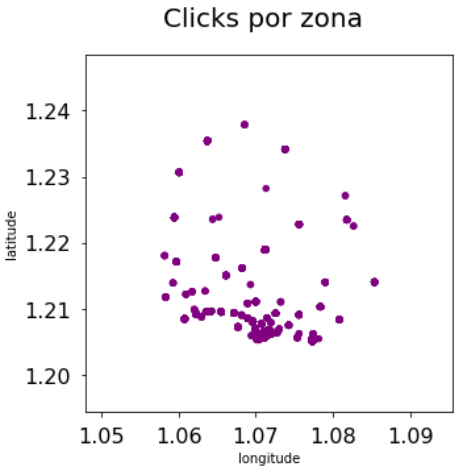


Realizamos un groupby entre las horas y los días, porque nos pareció interesante saber cuál es el día y a qué hora se realizaron más clicks. Realizando este análisis, según los datos que tenemos, llegamos a la conclusión de que ese día fue el 07/03/2019 a las 23 hs.



3.2. Zonas en las que se realizaron clicks

Notamos que la columna ‘country code’ posee un único valor, por lo cual eliminamos la columna. Además, éste país es Uruguay. Además, las columnas ‘latitude’ y ‘longitude’ poseen valores muy cercanos (por más de que estos valores esten hasheados linealmente), por lo cual tiene sentido hablar de un único país.



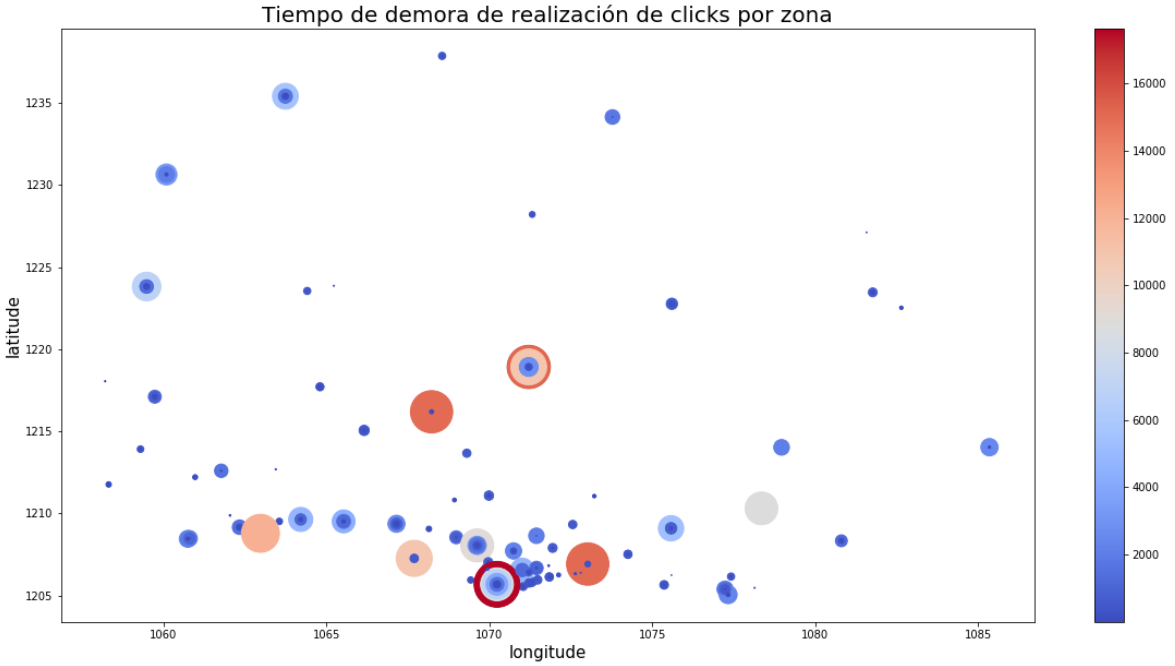
Analizamos cuál es la zona (longitud, latitud) en donde los usuarios tardan más y menos en realizar un click. Para eso analizamos las columnas 'latitude', 'longitude' y 'timeToClick', y llegamos a la conclusión de que en la misma zona se realizaron los clicks más rápida y lentamente (timeToClick máximo y mínimo).

timeToClick mínimo: 0.017

timeToClick máximo: 17616.188

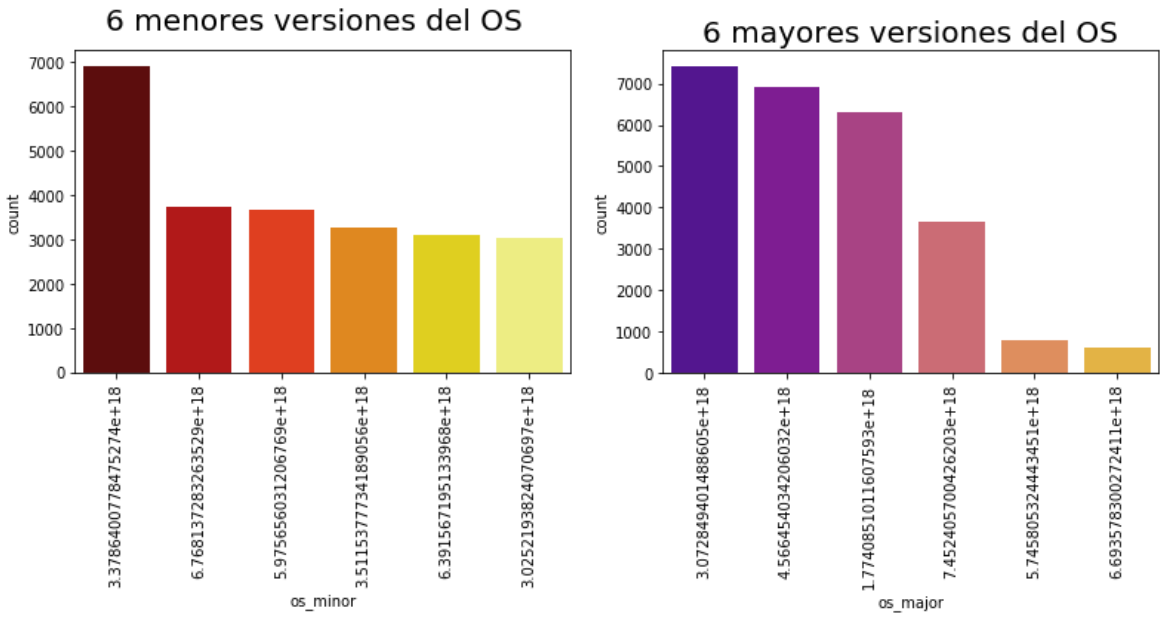
latitude: 1.205689

longitude: 1.070234



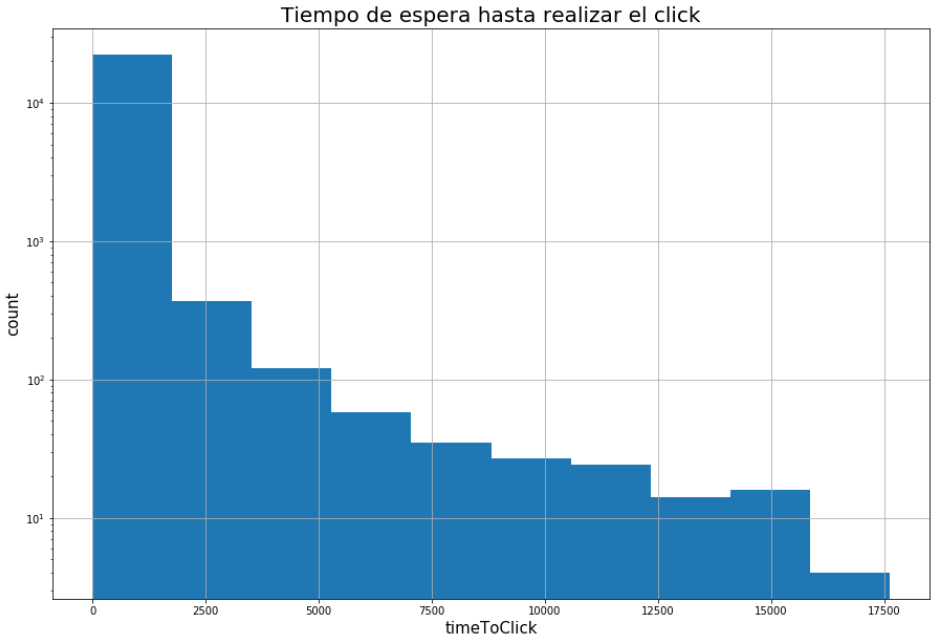
3.3. Versiones del OS

Teniendo las columnas 'os minor' y 'os major', realizamos un análisis de cuáles son las versiones del OS más y menos actualizadas.



3.4. Tiempo hasta la realización de un click

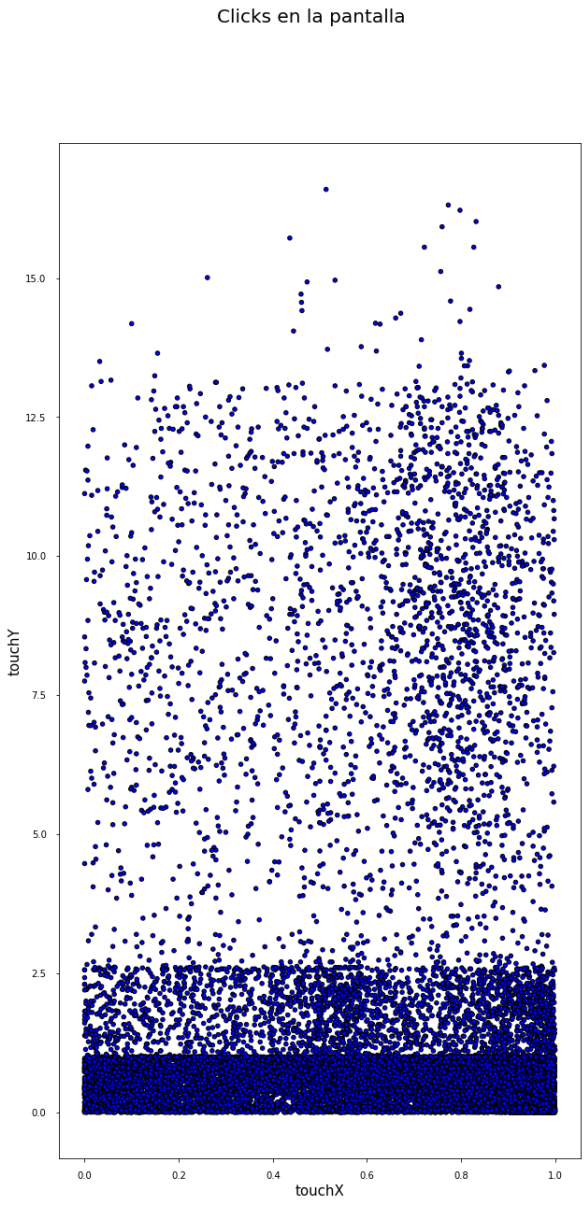
Nos pareció interesante analizar lo que tardan en realizar un click los usuarios con la información en 'timeToClick'. Para eso hicimos un gráfico que representa cuáles son los tiempos más y menos frecuentes.



Acá podemos observar que, en la mayoría de los casos, los usuarios tardaron hasta menos de 2500 segundos en realizar un click.

3.5. Clicks en la pantalla

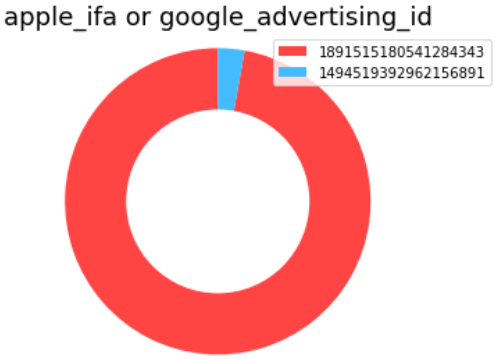
Queríamos ver en qué parte de la pantalla los usuarios realizan mayores cantidades de clicks, por lo cual realizamos el siguiente gráfico:



Acá podemos observar que la mayoría de los usuarios realiza clicks en las zonas inferiores de la pantalla, con esto podría llegarse a concluir que la mayor cantidad de publicidades se encuetra en esa zona. Aun asi no hay partes de la pantallas que queden 'vacias' de clicks, esto podria suponerse que algunos anunciantes un poco mas arriesgados publican anuncios por los lugares no tan frecuetes a encotrar una publicidad.

3.6. Frecuencia de ref type

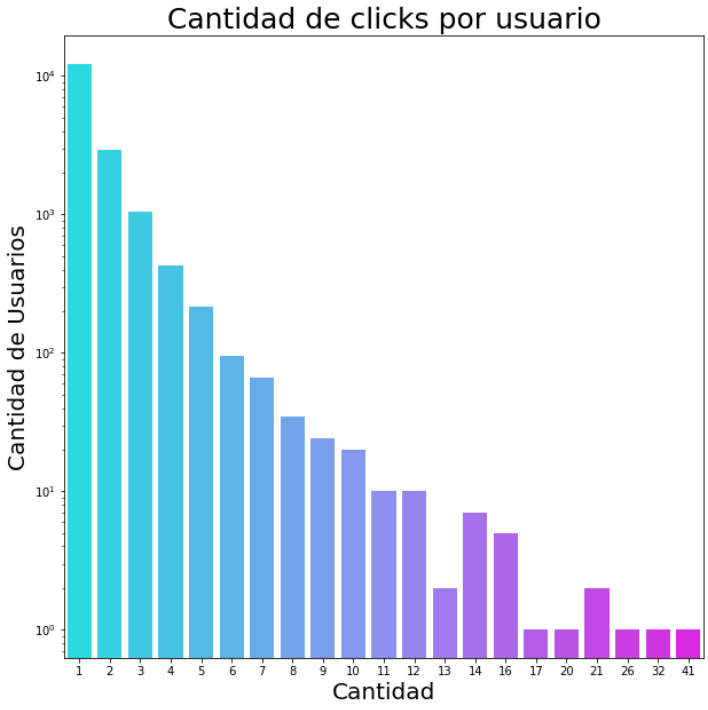
En clicks aparecen 4 tipos de 'ref type', de los cuales 2 aparecen en el resto de los archivos (auctions, events e installs) y los otros 2 no. Por lo cual, eliminamos esos dos valores únicos en clicks. También graficamos los dos valores que quedaron para ver cuál es el más frecuente:



Se puede ver que la zona roja es mucho mayor, puede deberse a la cantidad de usuario que se le muetran el tipo de publicidad, ya sea google o apple, es mucho mayor a su competencia.

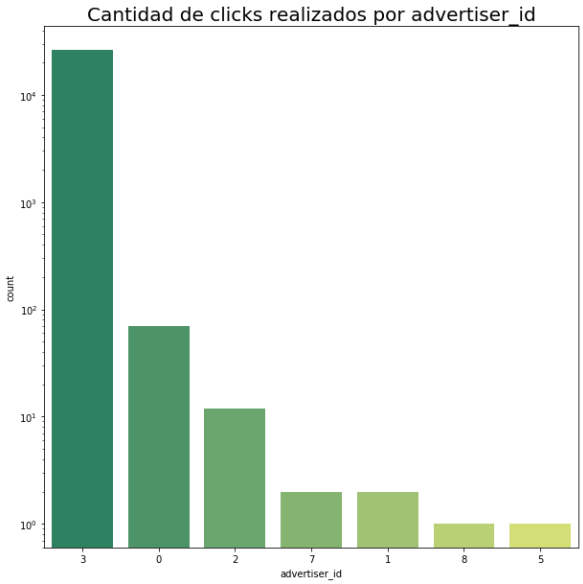
3.7. Análisis de la cantidad de clicks que se realizaron por cada partícipe

Quisimos analizar la frecuencia con la que cada dispositivo (ref hash) realiza clicks. Con el siguiente gráfico vemos que, por ejemplo, un mismo dispositivo realizó 41 clicks.



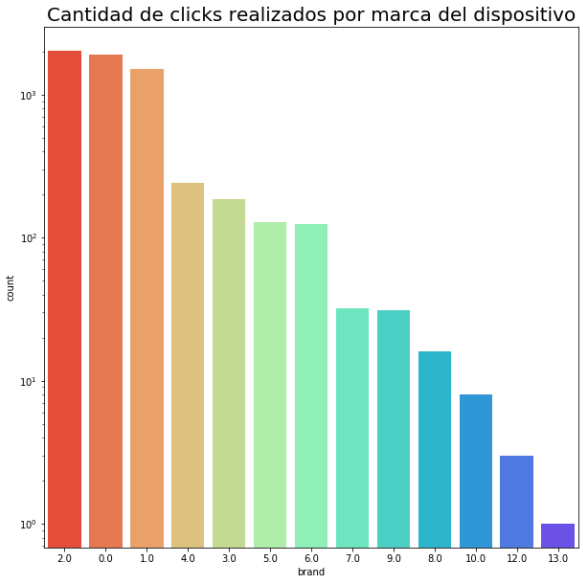
La cantidad de usuarios que solo hacen 1 clicks en la pantalla es mayor, ya que con un solo click la publicidad ya tendría que redirigir a la play store de la aplicación. A aquellos que clickearon mas cantidad de veces puede ser porque se les mostró mas de una publicidad. Sin embarco como se vera luego, no todos los que llegan a un click instalan una aplicación.

Analizamos la cantidad de clicks que realiza cada anunciante que contrató a Jampp, y los resultados fueron los siguientes:



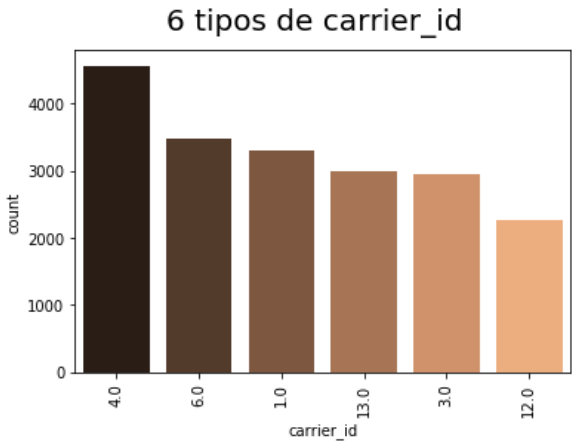
Acá podemos observar que, por la información analizada, el anunciante con id 3 fue el que más clicks recibió.

También nos pareció importante analizar cuál es la marca de dispositivo con la cual se realizaron más clicks.

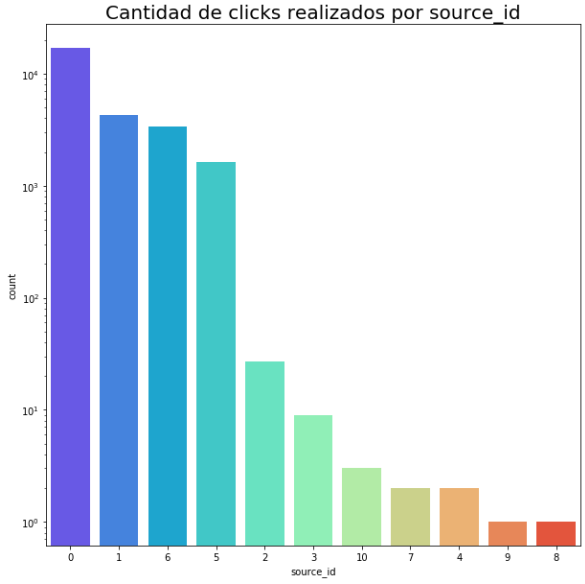


Los resultados obtenidos fueron que se realizaron más clicks con el brand 2.

Hay varios tipos de operadores de telefonía móvil (carrier id), hicimos un gráfico de cantidad de clicks que realizaron algunos operadores:

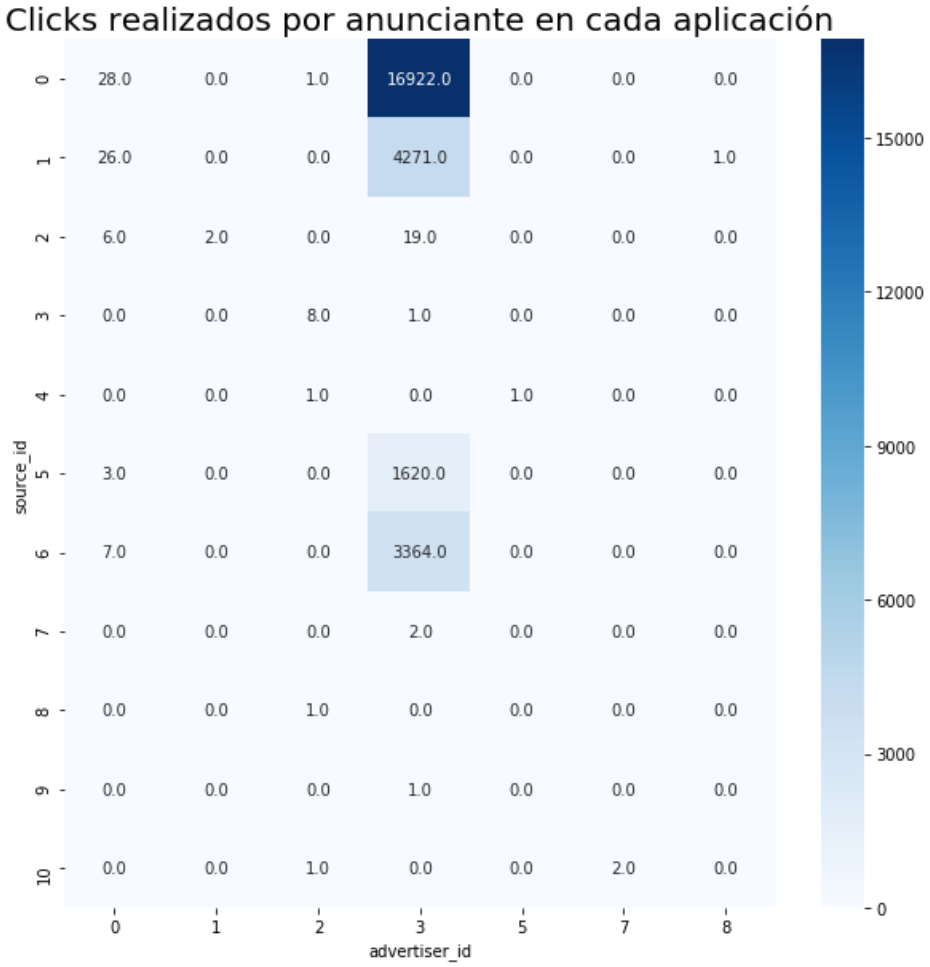


Analizamos la cantidad de clicks dependiendo del origen del mismo.



Vemos que la aplicación en donde se originaron más cantidad de clicks fue la de id 0.

Según el siguiente gráfico obtenido, vemos que el advertiser con id 3 es el que más clicks realizó en cada aplicación. En particular, en el source que más clicks realizó es el que posee id igual a 0.



3.7.1. Partícipes que tardaron más y menos tiempo en relizar un click

Por último, quisimos ver cuál era el advertiser, carrier y source id con los cuales se demoró más y menos tiempo en realizar un click, por lo que realizamos las siguentes tablas para cada caso:

ADVERTISER ID

- Máximo timeToClick

	0
timeToClick	17616.188000000002
advertiser_id	8.0

- Mínimo timeToClick

	0
timeToClick	0.017
advertiser_id	0.0

CARRIER ID

- Máximo timeToClick

	0
timeToClick	17616.188000000002
carrier_id	116.0

- Mínimo timeToClick

	0
timeToClick	0.017
carrier_id	0.0

SOURCE ID

- Máximo timeToClick

	0
timeToClick	17616.188000000002
source_id	10.0

- Mínimo timeToClick

	0
timeToClick	0.017
source_id	0.0

4. Installs

Se abrió el csv pasando por parámetros el tipo de dato que correspondía a cada columna para así ocupar menos memoria y que la carga sea mas rápida, evitando que haya que recorrer todas las columnas y así panda tenga que asignarle un tipo de dato.

Se puede ver que el DataFrame cuenta con las siguientes columnas:

'created', 'application_id', 'ref_type', 'ref_hash', 'attributed', 'implicit', 'device_countrycode', 'device_brand', 'device_model', 'session_user_agent', 'user_agent', 'event_uuid', 'kind', 'wi-fi', 'trans_id', 'ip_address', 'device_language', 'click_hash'

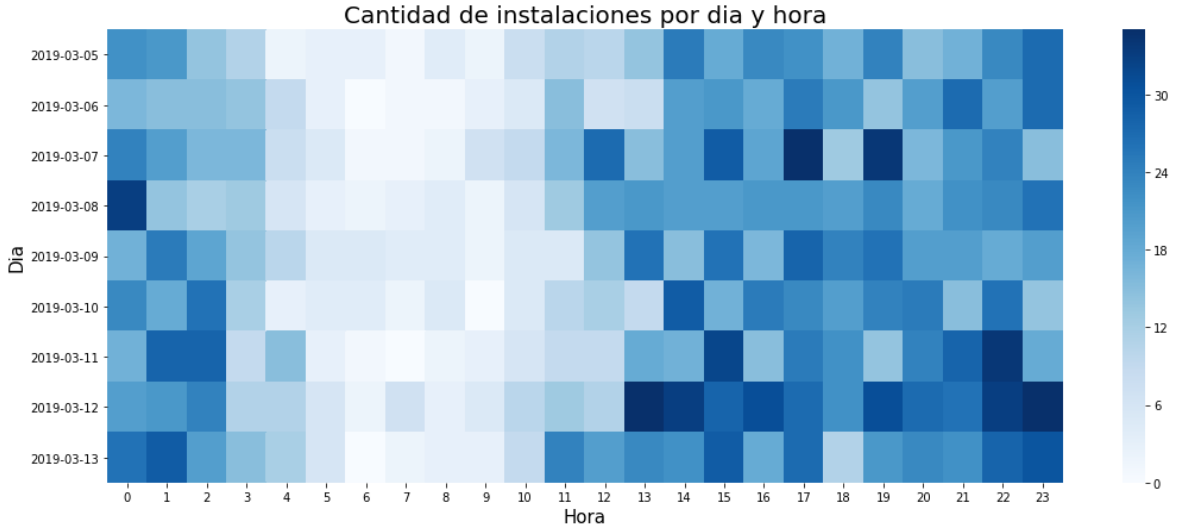
Algunas consideraciones hechas por los valores que toman las columnas son:

- La columna 'click_hash' posee todos valores nulos, por lo tanto se procede a eliminarla luego de la carga del cvs.

- La columna 'device_countrycode' posee dos valores, pero siendo la única tabla que cuenta con dos country-codes, se puede considerar que el dato que nos sirve el hasheado como 2970470518450881158 corresponde al valor nulo que esta hasheado generando confusión al pensar que los datos provienen de dos países diferentes
- La columna event_uuid contiene un valor diferente por cada fila.
- La columna 'attributed' contiene el valor booleano False en todos sus datos no nulos. Este dato hace referencia a si la instalación fue atribuida directamente a Jammp, y aunque esto sea falso vamos a considerar las instalaciones como orgánicas para los clientes de Jammp.
- La columna 'ref_type' esta referida al tipo de publicidad que se muestra, ya sea proveniente de google o apple. Al desconocerse cual es cual se trabajara sobre supuestos.
- La columna 'session_user_agent' se tomaron como datos anonimizados.

4.1. Instalaciones por Día y Hora

Se cuenta con la información que las instalaciones que van desde los días martes 5 de marzo a el día miércoles 13 de marzo. Para el análisis se agregaron las columnas día y hora a partir de la columna 'created' y se creo un heat map en comparaciones de la cantidad de instalaciones que se hacen por día y hora.



En el gráfico se ve que las horas con menor cantidad de instalaciones es entre las 4 y las 10 de la mañana. También se puede corroborar que el horario de mayor cantidad de instalaciones es a partir del medio día, claramente ya que el uso del celular ya es algo habitual y los usuarios buscan simplificar sus tareas con múltiples aplicaciones. Lo que es una sorpresa es que hay una cantidad significativa de instalaciones a media noche, de 0 a 3 de la mañana, casi la misma cantidad que puede verse entre las 12 y 14 horas. Se podría decir que los usuarios a los que se le muestra la publicidad es un publico que acostumbra a estar a altas horas de la noche despierto y no es posible obtener instalaciones en esos momentos, aprovechando el espacio de publicidad por sobre otros competidores.

4.2. Instalaciones con Wifi

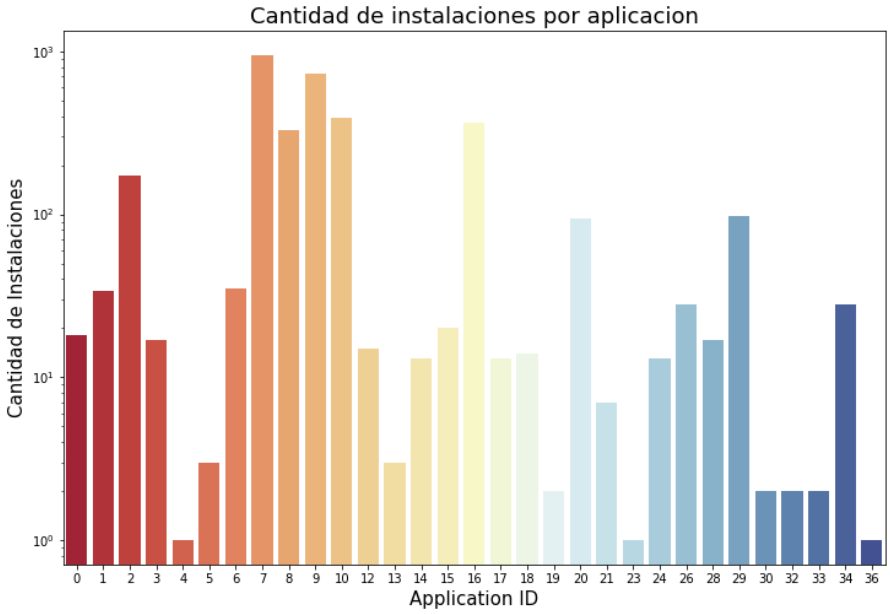
Se hizo un análisis de comparar la cantidad de instalaciones que se hace con o sin wifi, resultando así el siguiente gráfico:

Se puede ver en el gráfico que es mayor la de instalaciones que se hacen con wifi que sin estar conectados a el, esto puede deberse a la cantidad de datos tiene que usarse para descargar una aplicación.

4.3. Instalaciones por Aplicación

Las aplicaciones pueden identificarse con un ID que va del 0 al 36

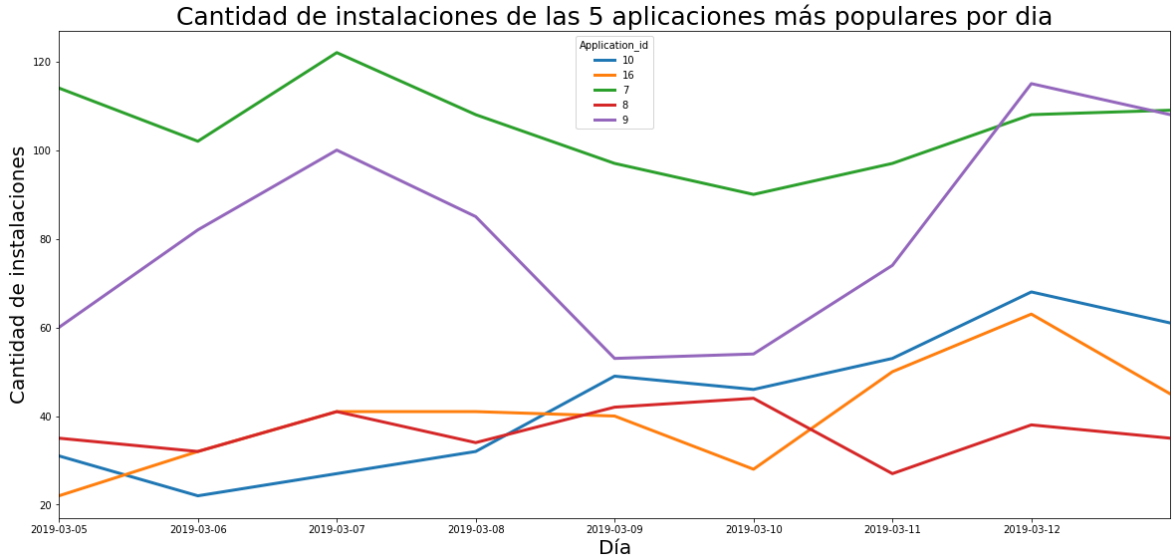
Se analizo la cantidad de instalaciones por aplicación id y debido a que las diferencia entra las aplicaciones mas populares y las menos instaladas era significativa se uso una escala logarítmica.



En el gráfico se ve que se tiene cerca de 1000 instalaciones de la aplicación 7, seguida por la aplicación 9 con alrededor de 100 instalaciones menos y las siguientes que se encontraron en las 5 aplicaciones favoritas de los usuarios no caen de las 100 instalaciones. Analizaremos el comportamiento de estas instalaciones, en mas profundidad a continuación.

4.3.1. Aplicaciones mas populares por Día

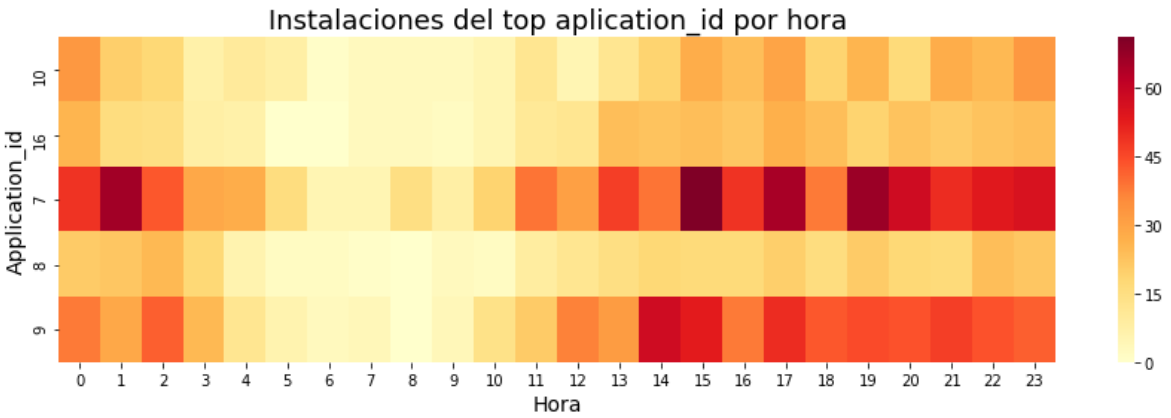
Luego se continuo a analizar la cantidad de instalaciones por día que tienen las 5 aplicaciones mas populares a partir de un line-plot. El gráfico nos muestra que la aplicación con ID 7 se mantiene con gran cantidad de instalaciones a lo largo de los 8 días de los que tenemos información. La siguiente es la aplicación con ID 9 cuenta con menos instalaciones los días 9 al 11, estos corresponden al fin de semana. A partir de las siguientes 3, se reduce la cantidad de instalaciones a comparación de las 2 mas populares y se mantiene entre las 20 y 40 instalaciones los primeros 5 días y luego aumenta alcanzando las 60 instalaciones por día.



La aplicación que se mantiene favorita es la 7, pero la aplicación 9 llega a superarla en el ultimo día. Esto sugiere que 9 podría llegar a tener tantas o mas instalaciones que 7 si se ofertara mas por ella, en especial ese fin de semana que su cantidad de instalaciones cayo drásticamente a comparación con el resto de los días. La aplicación 10 a su vez, también mostró crecimiento a mitad de la semana y este crecimiento se mantuvo hasta ubicarla en tercer lugar al día 13.

4.3.2. Aplicaciones mas populares por Hora

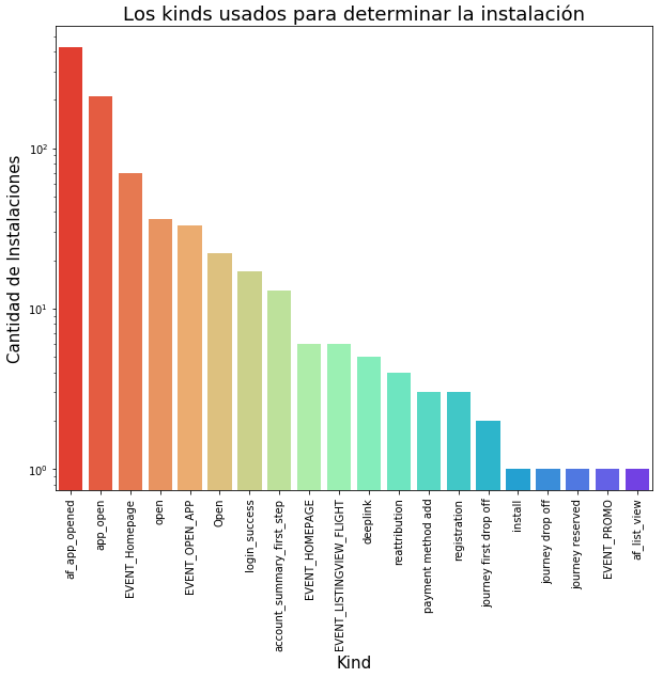
Se gráfica las descargas de las aplicaciones mas populares por hora:



Se mantienen las aplicaciones mas populares 7 y 9 con muchas instalaciones a lo largo del día. La numero 7 demuestra su popularidad incluso en el horario de 4 a 10 de la mañana ya que concentra al menos 20 instalaciones a las 4hs, 6hs y 10hs, al contrario de las demás donde no parecen tener ninguna descarga. El horario de mayor cantidad de instalaciones , mas de 60, también es de la aplicación 7 es a las 17hs. Claramente conviene apostar por esta aplicación, pero también considerando la oportunidad de crecimiento de las otras que se encuentran en el top.

4.4. Kind

Para determinar la instalación de una aplicación se utiliza un tipo de evento que anuncia que la descarga fue realizada, esta información llega de diferentes formas, ya sea por abrir la app, por intentar iniciar sesión o por encontrarse en la pagina principal de la aplicación. En el siguiente gráfico se utilizo una escala logarítmica y se muestran los eventos mas usados para contar este indicador:

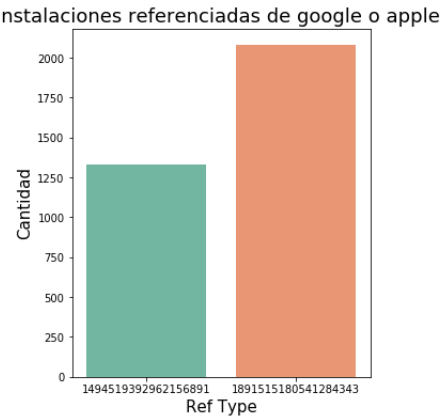


En el eje X se ven los diferentes tipos de Kind, algunos son similares como lo son 'Open', 'open' y 'app open' por lo tanto estos datos pueden tomarse como uno único. Es decir que el evento que mas destaca es el hecho de abrir la aplicación por primera vez. luego lo sigue el evento de estar en el Home Page de la app.

Lo ideal en estos casos es tener una norma que se aplique en todas las aplicaciones y así poder tener mejor registro de estos, aunque pude ocurrir que la primera opción de una app sea registrarse y en ese caso el evento no podría ser home page si no Log In, o otra opción es que la aplicación te permita tener un usuario anónimo por un periodo de prueba y por eso se inicie en la pagina principal

4.5. Instalaciones por Ref Type

Se analizan por país el tipo de instalaciones que se utilizaron distintos tipos de ref type, el que corresponde a apple ifa y google advertising id.

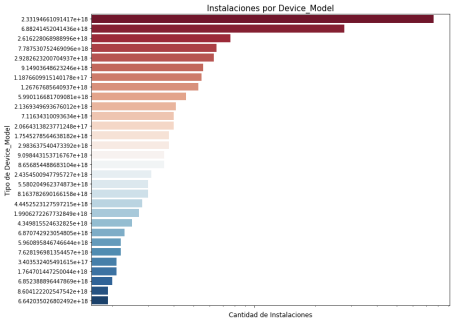


El ref type que en el grafico se muestra como naranja tiene alrededor de 500 usuarios mas que el verde. Por deducción se podría pensar que el ref hash de código comenzado en 18 (color naranja) es el de google ya que google abarca mayor cantidad de celulares que la publicidad de apple que es exclusiva para los celulares de esta misma compañía.

4.6. Analisis del Device

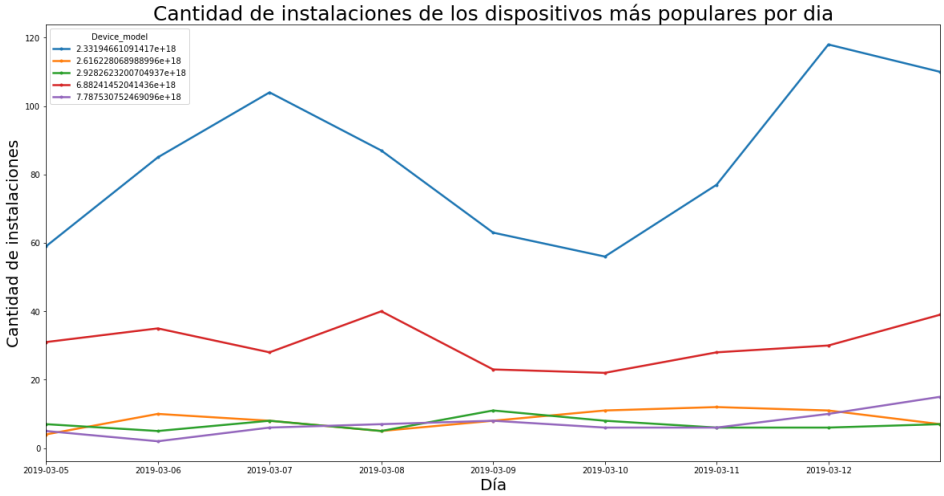
4.6.1. Instalaciones por Device model

Se seleccionaron los 30 device model para comparar la cantidad de instalaciones hechas. La muestra fue chica ya que tomando mas cantidad se registraba una única instalación por device y no son útiles para el análisis. Para este gráfico se uso una escala logarítmica para mejor resultado visual.



Se muestra que el device model mas usado tiene alrededor de unas 700 instalaciones, el siguiente cerca de 200 instalaciones y el resto decae entre las 20 y 80 instalaciones. Que un modelo de celular cuente con mayor cantidad de instalaciones puede deberse a su mayor capacidad de almacenamiento, lo que permite incluso instalar aplicaciones por corto tiempo para probar que tipo de servicio ofrece. En ese caso es mejor apostar por usuarios con estos tipos de celulares que los de menor capacidad ya que le impedirán descargar nuevas sin tener que eliminar otras para su uso.

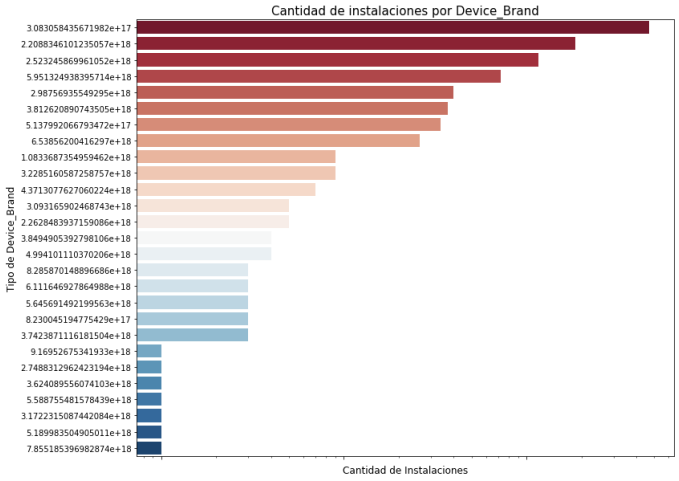
Se compara a continuación la cantidad de instalaciones que presentan los 5 modelos mas populares.



Se puede notar la predominancia de cierno tipo de modelo en las instalaciones con un pico cercano a las 120 instalaciones. El segundo no supera las 40 instalaciones en toda la jornada. Los siguientes 3 tienen alrededor de 10 instalaciones cada uno. Como el modelo del celular no tiene relación con la decisión de instalar o no una aplicación, se puede decir que hay mas cantidad de usuarios que usan ese modelo de celular y es por eso que se presenta mayor diferencia en sus proporciones.

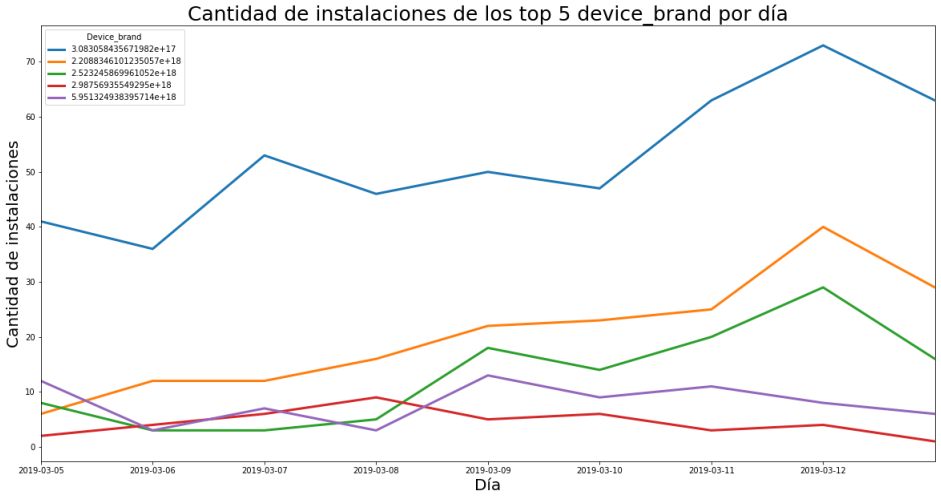
4.6.2. Instalaciones por Device Brand

Se muestra en el siguiente gráfico los device_brand que hacen mayor cantidad de descargas. Para eso se filtraron las 30 mas populares ya que la cantidades de brands es muy variada y el gráfico no se vería correctamente de mostrarlas a todas.



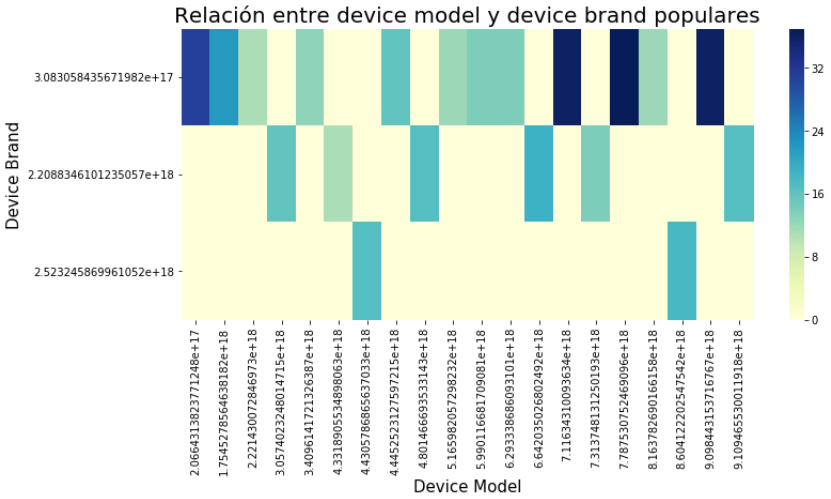
Luego se esos 30 brands mas populares los que continúan solo presentan 1 instalación. Mostrando una menor preferencia por esta marca de celular, esto puede deberse a su antigüedad y es por eso que circulan menos cantidad de dispositivos con estas características.

Luego se filtraron los device_brand del top 5 y se gráfico la cantidad de instalaciones por día.



4.6.3. Relacion entre device_model y device_brand

Se toma los 5 device brands mas populares y se los busca en la tabla con los 10 device models mas populares y se gráfica un heat map para buscar correlaciones entre marca y modelo.



Se puede ver en el heat map que solo 3 de los 5 device brand mas populares tienen relación con los 10 modelos mas populares, de todas formas el mas predominante tiene 5 modelos con altas tazas de instalación. Además las otras marcas también se puede ver una relación con ciertos modelos de celulares y es unos en otros. Se demuestra también con este gráfico que los modelos de celular solo pertenecen a una marca de device, es decir que cada modelo es exclusivo de cada brand.

4.7. Instalaciones Implicitas

Los datos nos dicen si la instalacion ocurre luego de ver la publicidad mostrada por Jammp o no, se llego por lo tanto al siguiente grafico con la informacion dada.

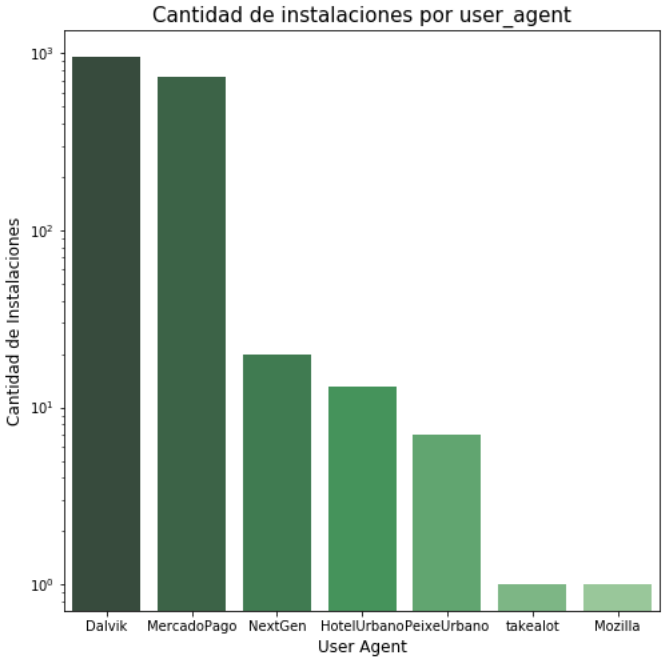
La relación nos muestra que aunque todos los usuario terminaron instalando la aplicación, solo 3l 25,4 % fueron debido a la publicidad vista anteriormente y el 74,6 % llega por algún otro medio orgánico que desconocemos

4.8. Instalaciones por User Agent

Descripción de User Agents: Un agente de usuario es una aplicación informática que funciona como cliente en un protocolo de red; el nombre se aplica generalmente para referirse a aquellas aplicaciones que acceden a la World Wide Web. Cuando un usuario accede a una página web, la aplicación generalmente envía una cadena de texto que identifica al agente de usuario ante el servidor. Este texto forma parte de la petición a través de HTTP, llevando

como prefijo User-agent: o User-Agent: y generalmente incluye información como el nombre de la aplicación, la versión, el sistema operativo, y el idioma.

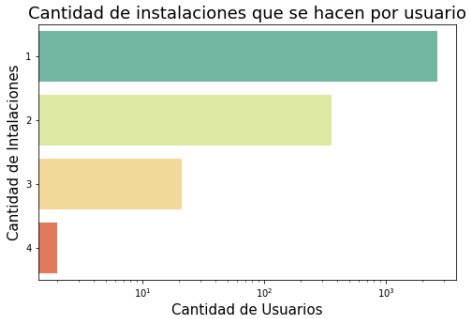
En el siguiente gráfico se muestras los user agents mas utilizados, filtrados de tal manera de poder agruparlos en diferentes categorías sin considerar la versión utilizada.



La primera mas usada es Dalvik con cerca de 1000 usuarios, seguido a la par por MercadoPago, ambas muy populares en comparación con su competencia.

4.9. Cantidad de instalaciones por usuario

El gráfico muestra la cantidad de instalaciones hechas por un usuario. La barra de mayor tamaño hace referencia a la cantidad de usuarios que realizaron una única instalación, luego en menor proporción siguen usuarios que realizaron 2 y 3 instalaciones, y por ultimo solo 2 usuarios que realización 4 instalaciones.

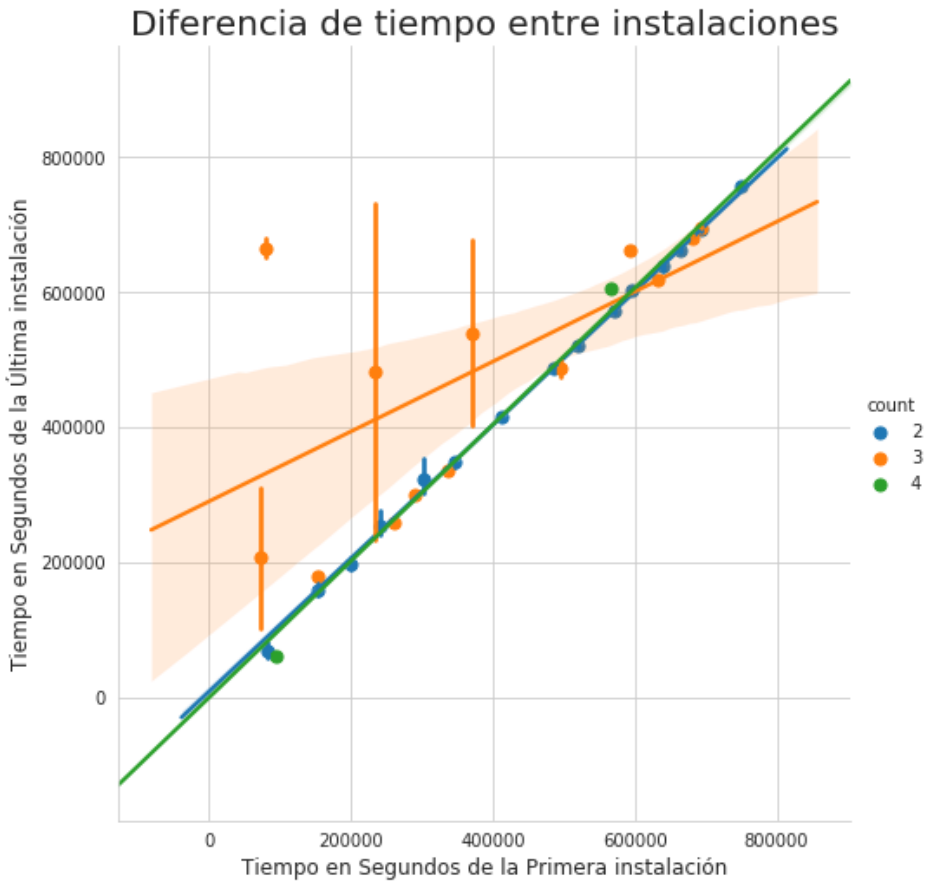


Es interesante notar a los 2 usuarios que instalaron 4 aplicaciones en 8 días. Lo que se hizo es localizar cual era su ref_type de ellos y se encontraron los valores:

1) 3272750442824629569

2) 5208834946313176321

Nuestro análisis en el notebook continua tomando los usuarios que realizaron mas de una instalación y tomando el tiempo en segundos de su primera y ultima instalación. En algunos casos la diferencia de tiempo entre estos dos valores es de menos de 2 segundos, considerando que contamos con algunos bots entre los datos de nuestra tabla. Se logra así graficar en en un plano la diferencia en segundos entre instalaciones.



En el anterior gráfico se muestra la linealidad entre el tiempo en segundos desde la primera y la ultima instalación. Los puntos mas cercanos a una linea recta son los que tienen una diferencia menor entre el tiempo de la primera instalación y la ultima ya que su coordenada toma valores similares a por ejemplo (x= 60340.015, y= 60487.069). Los que presentan mayor diferencia de tiempo entre las instalaciones son los que hicieron al menos 3, por lo tanto la regresión en color naranja tiene un promedio diferencial de tiempo mayor a los que instalaron 2 veces (linea azul) y 4 veces (linea verde)

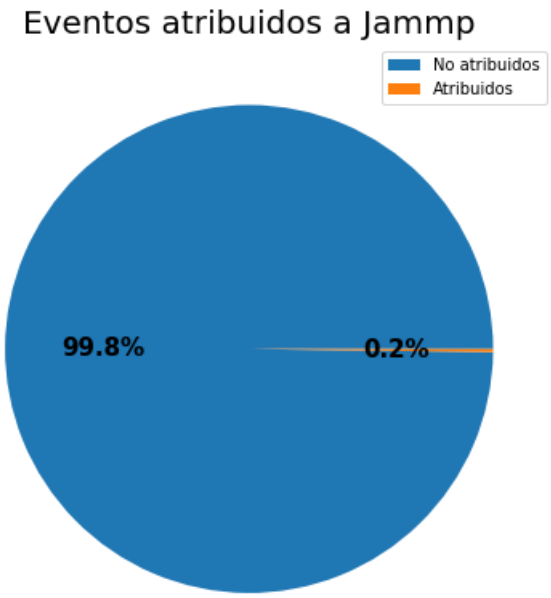
5. Events

El archivo csv con la información de eventos realizados, contiene al rededor de 2 millones y medio de filas y 22 columnas. Al abrir el archivo se aclaró el tipo de cada columna para poder minimizar el espacio en memoria y el tiempo que tarda en abrir el archivo. Luego se analizaron los datos en cada una de ellas y se observó lo siguiente:

- device countrycode: es un único valor para todo el archivo ya que los datos se tomaron de un único país. Como el dato es irrelevante en este análisis, se borró la columna.
- event uuid: es un valor único para cada evento realizado, tampoco aporta ningún valor al análisis entonces también se borro esta columna.

5.1. Atribuciones a Jammp

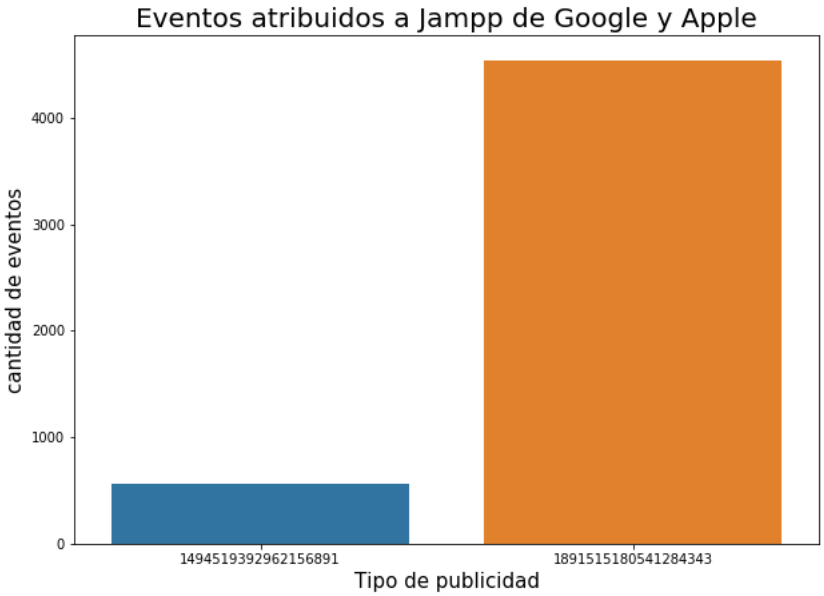
Con la columna 'attributed' (sin elementos nulos y de tipo bool), podemos analizar sencillamente cuáles de todos los eventos realizados fueron atribuidos a la empresa Jammp.



Se puede observar en el gráfico anterior que sólo una mínima parte de todos los eventos que se realizan, son atribuidas a la empresa.

5.1.1. Atribuciones a Jammp según ref type

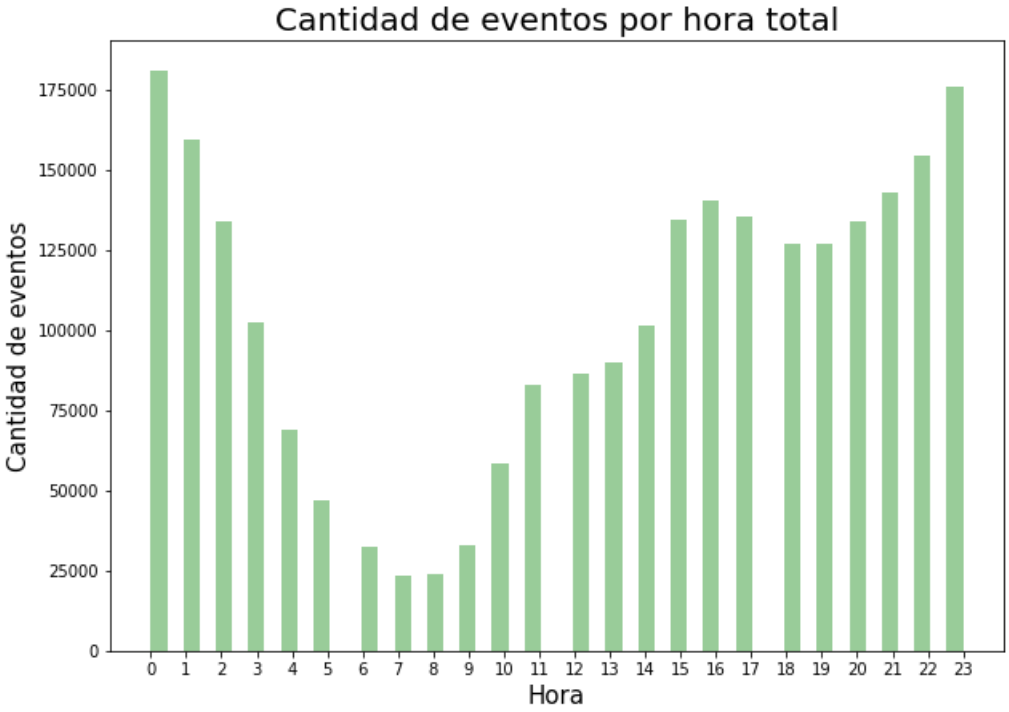
Luego de la minoría que se le atribuyen del análisis anterior, observamos entre las publicidades provenientes de Google y Apple cuál es la que más atribuciones le deja a la empresa.



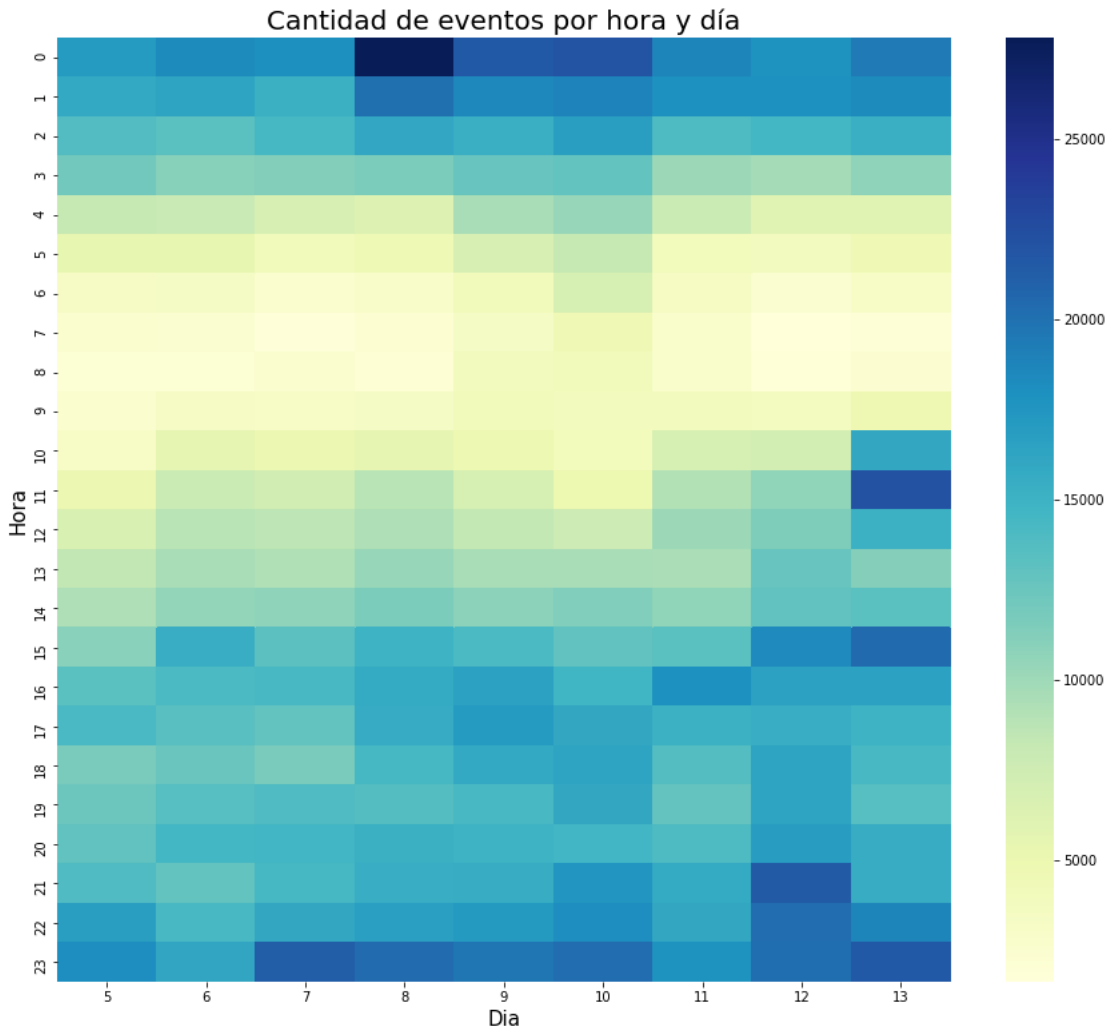
Es difícil tener una conclusión completa ya que los datos están hasheados. Pero sí se puede observar, que hay uno del cual proviene la mayoría de eventos atribuidos. Esta estadística puede ser muy útil a la hora de decidir si es conveniente comprar la subasta o no.

5.2. Análisis de la cantidad eventos por hora y día

A partir de los datos de la columna 'date' podemos obtener información estadística sobre la hora y el día en cual se realizan más o menos eventos.



En el histograma podemos observar que el horario en que más eventos se realizan es cerca de la medianoche. Un dato extraño es que el mínimo de eventos realizados es a las 7, 8 de la mañana y no a la madrugada, horario común en el que la mayoría de la gente descansa. Para ver si las franjas horarias de más y menos eventos es constante en todos los días o depende de ellos, se realizó un gráfico del tipo heatmap:



En este gráfico, podemos observar que los horarios de más y menos cantidad de eventos realizados es más o menos constante en todos los días de análisis. Se puede observar que los días 9 y 10, que resultan ser Sábado y Domingo, el mínimo no es tan bajo como los días de semana y además se atrasa un poco del horario habitual, habiendo más actividad a lo largo de la noche que durante el día.

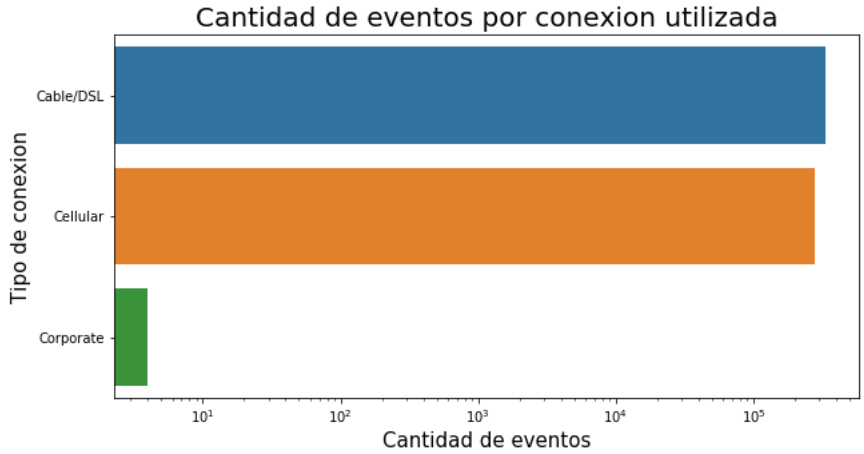
5.3. Análisis por tipo de conexión

Para obtener información sobre el tipo de conexión a internet utilizado por el usuario, podemos recurrir a dos columnas, 'connection type' (Cable/DSL, Cellular, Corporate) y 'wifi'(true o false), ambas con una gran cantidad de datos nulos.

5.3.1. Conexión a wifi

Podemos observar que la mayor cantidad de eventos realizados es con el dispositivo conectado a wifi. Sin embargo, en este análisis sólo se están considerando los elementos no nulos, los cuales representan la mitad del total de datos. Teniendo en cuenta los elementos no nulos, sólo se sabría con seguridad que alrededor del 40 % de los dispositivos están conectado a wifi.

5.3.2. Tipo de conexión

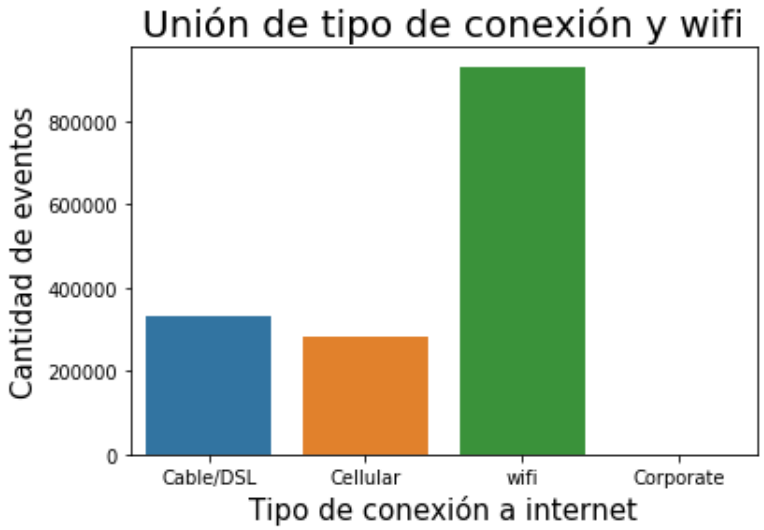


Se observa que el tipo de conexión 'Cable/DSL' y 'Cellular' tienen valores similares mientras que el tipo de conexión 'Corporate' tiene una pequeña cantidad de eventos asignados en

comparación a los otros dos. Pero igual que en el gráfico anterior, no se tiene en cuenta la gran cantidad de elementos nulos que esta columna posee.

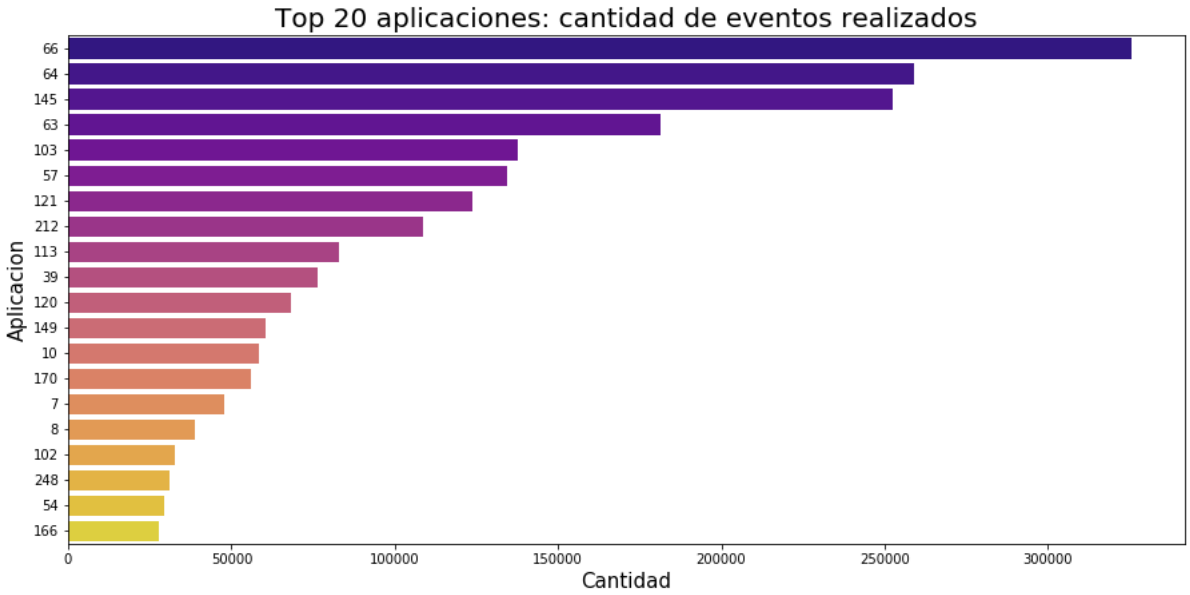
5.3.3. Suposición de tipo de conexión

En el análisis de los datos realizados en el csv se observó que cuando el dato de la columna 'connection type' es nulo, el dato en la columna 'wifi' es 'true' o nulo, y a su vez, 'connection type' tiene valores no nulos, cuando 'wifi' es 'false' o nulo. Lo cual tiene sentido ya que si el dispositivo está usando conexión wifi, no está usando ninguno de los otros tipos de conexión a internet especificados en 'connection type'. Teniendo en cuenta este análisis se mezclaron los datos de ambas columnas (que no se pisan entre sí) y que resultan en una columna ue aún tiene datos nulos pero en menor cantidad y se obtuvo el siguiente gráfico:

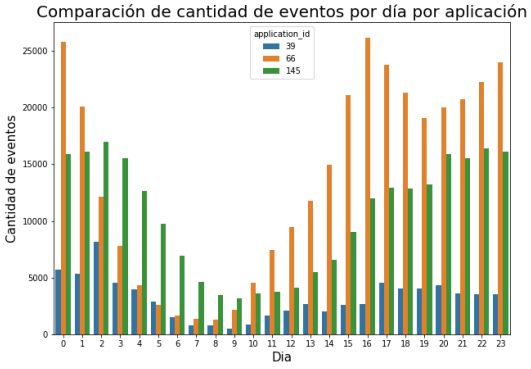


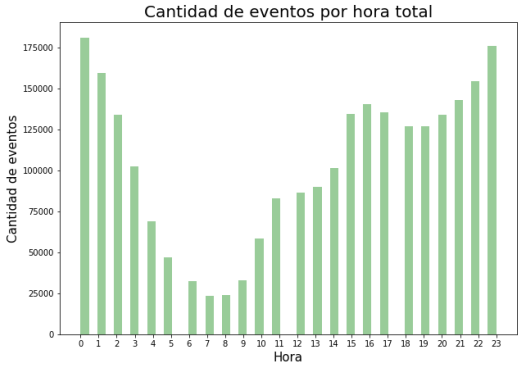
5.4. Applicaciones

Cada aplicación tiene asociada su 'Application id', a partir de esto podemos realizar varios análisis sobre qué aplicaciones son más utilizadas que otras. Estadísticamente nos sirve para saber si la aplicación que vamos a representar nos garantiza mayor cantidad de ganancias, o si las aplicaciones que ya representamos necesitan más visibilidad en publicidad para poder ser más utilizadas. Un claro ejemplo es el siguiente gráfico donde se encuentran las 20 aplicación en las cuales se realizan mayor cantidad de eventos.



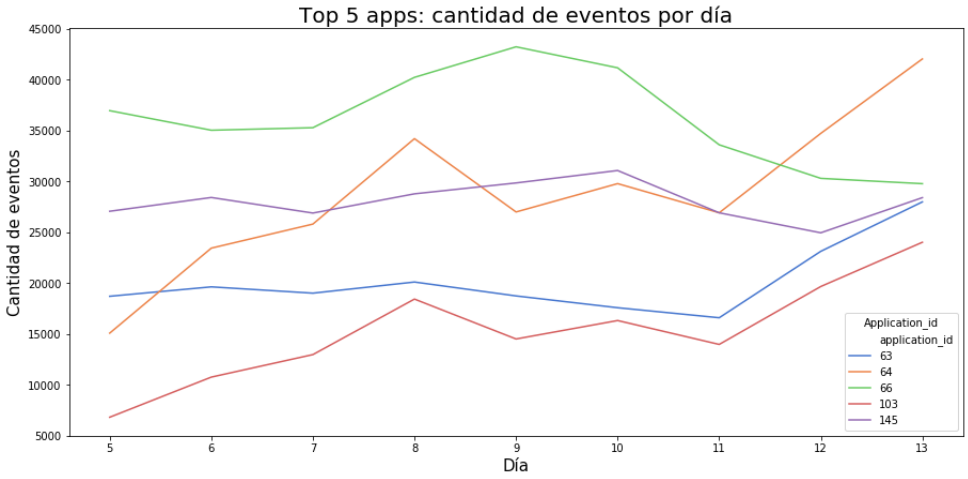
Se puede observar que hay mucha diferencia en cantidad de eventos ente las primeras aplicaciones y las últimas. sobre todo la primer aplicación, que tiene más de 50.000 eventos por sobre la segunda. Entonces si analizamos el comportamiento sólo de esta aplicación en cantidad de eventos por hora y se compara con el mismo gráfico de todas las aplicaciones ya visto anteriormente, el comportamiento de ambos sería muy similar.





Primero notar que las escalas de la cantidad de eventos son diferentes, lo que se intenta ver es como el gráfico total mantiene la forma de la aplicación más recurrente. Entonces el primer gráfico se observa la 1 aplicación mas usada (66), la 3 (145) y la 10 (39). Podemos ver que las aplicaciones se van deformando, en comparación con el segundo gráfico, a medida que se alejan del "primer puesto".

Podemos también analizar el comportamiento de algunas aplicaciones por día:



En este gráfico podemos entender, que el comportamiento de el "top 20"de aplicaciones más usadas, no es constante si se analiza por día, ya que el comportamiento individual de cada una no es lineal. Se observa además, que la aplicación más usada tiende a disminuir su cantidad de eventos al final de la muestra, mientras que las demás aumentan bruscamente en este mismo período. Por lo que si la muestra se tomara en los 10 días siguientes, todos estos valores podrían no servir. En conclusión, es necesario mantener actualizada constantemente esta información.

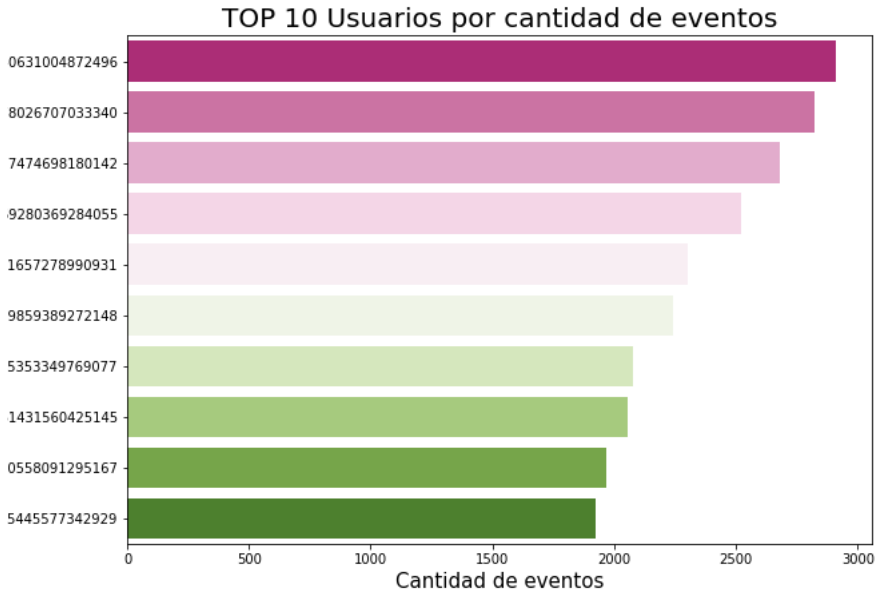
5.5. Dispositivo

Al igual que con la aplicación más usada, si uno se queda con el dispositivo más recurrente, el gráfico de cantidad de eventos por hora es similar al gráfico de eventos totales por hora.

5.6. Usuario

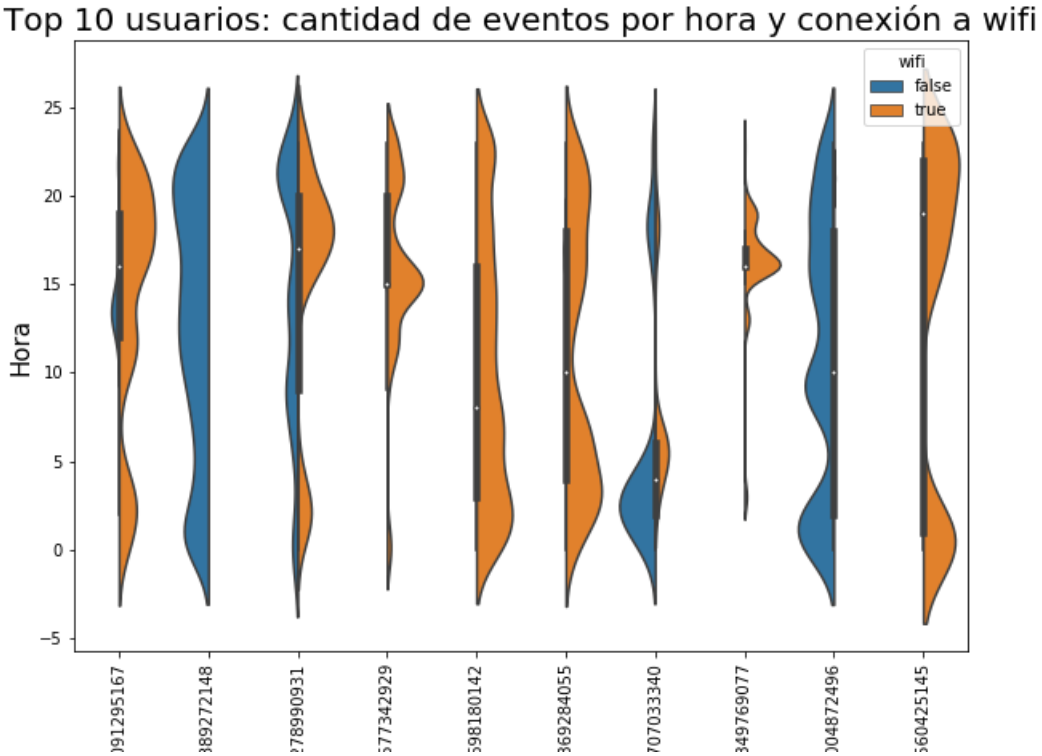
Se puede hacer un análisis de los usuarios que realizan los eventos a partir de la columna 'ref hash', la cual es un id supuestamente único para cada dispositivo.

Tomando los 10 usuarios más ocurrentes del archivo:



Se observa que no todos los usuarios realizan eventos de manera equitativa, entonces la información del id del dispositivo correspondiente a un usuario es un dato relevante a la hora de elegir si comprar o no el espacio publicitario, ya que algunos son más propensos a realizar eventos en aplicaciones que otros, lo que le da a la empresa mayor probabilidad de ganancias.

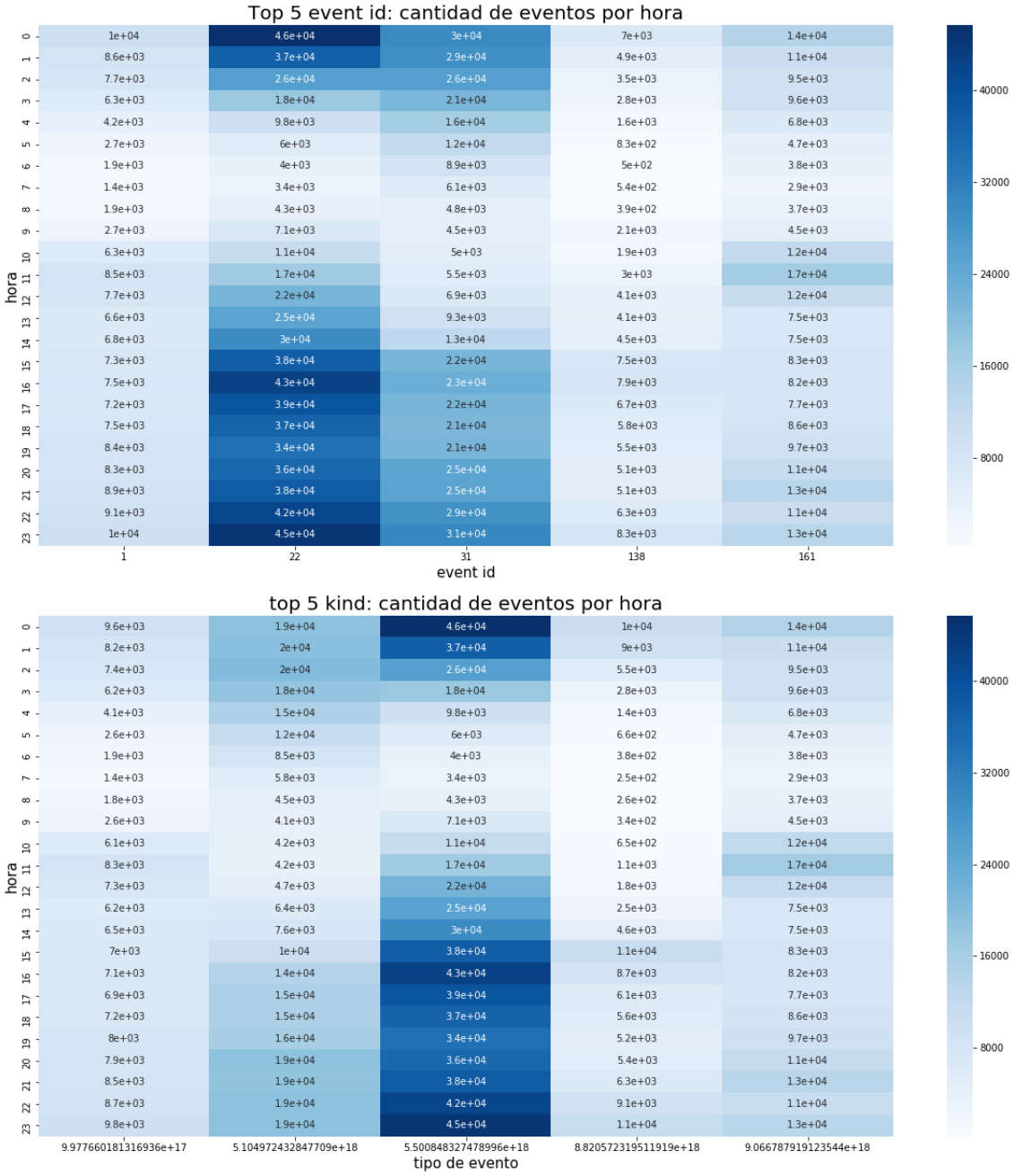
Volviendo al tema de la conexión del dispositivo, en el siguiente gráfico analizamos la relación entre los eventos realizados por el usuario y la conexión a wifi:



Y obtenemos que este tipo de conexión está relacionado con una preferencia del usuario mismo. Pero se sigue manteniendo, que la mayoría de los eventos se realizaron con conexión a wifi.

5.7. Tipo de eventos y id del evento

Con las columnas 'kind' y 'event id' podemos analizar los datos relacionados con el evento en sí. El primero nos da información de qué tipo de evento se realizó y el segundo es un id para cada evento. Se realiza un heatmap con los 5 tipos e id de eventos más ocurrentes y se encuentra una relación gráfica:

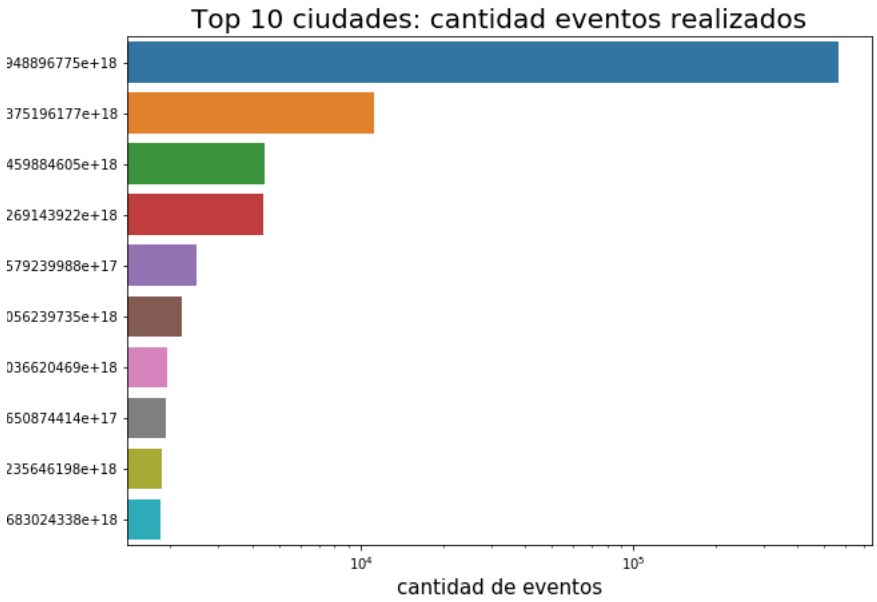


si observamos la segunda columna de 'event id' y la tercera de 'kind' podemos ver gráficamente por los colores, y comprobarlo con los números, que los valores por hora son muy similares, lo mismo ocurre para las columnas (1,1), (3,2), (4,4) y (5,5), siendo la primer coordenada la columna del heatmap de 'event id', y la segunda de 'kind'. Entonces existe una relación entre estos valores.

5.8. Ciudad

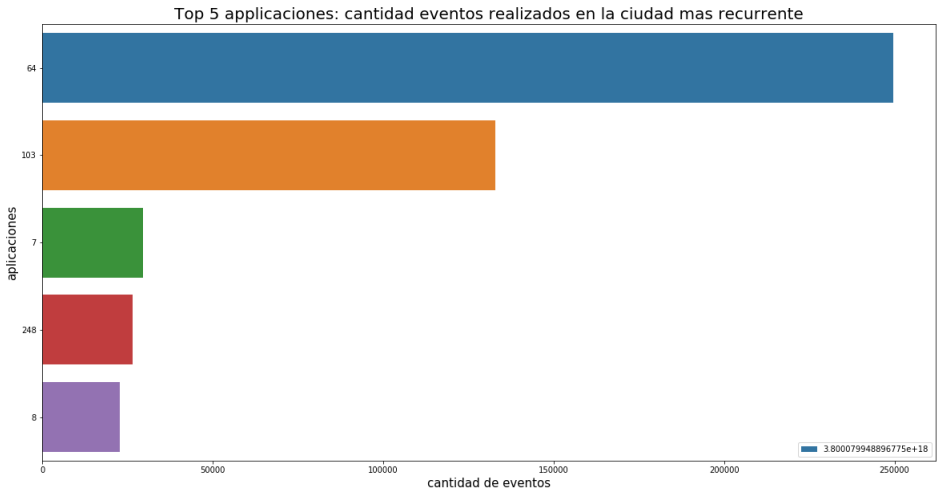
La columna 'device city' nos da información sobre la ciudad asociada al dispositivo desde el cual se hace el evento, pudiendo agregar datos estadísticos de en que ciudades es más probable

realizar eventos:



Tener en cuenta que la escala del eje 'x' es logarítmica, por lo cual la ciudad más recurrente, es 10 veces mayor que todas las demás.

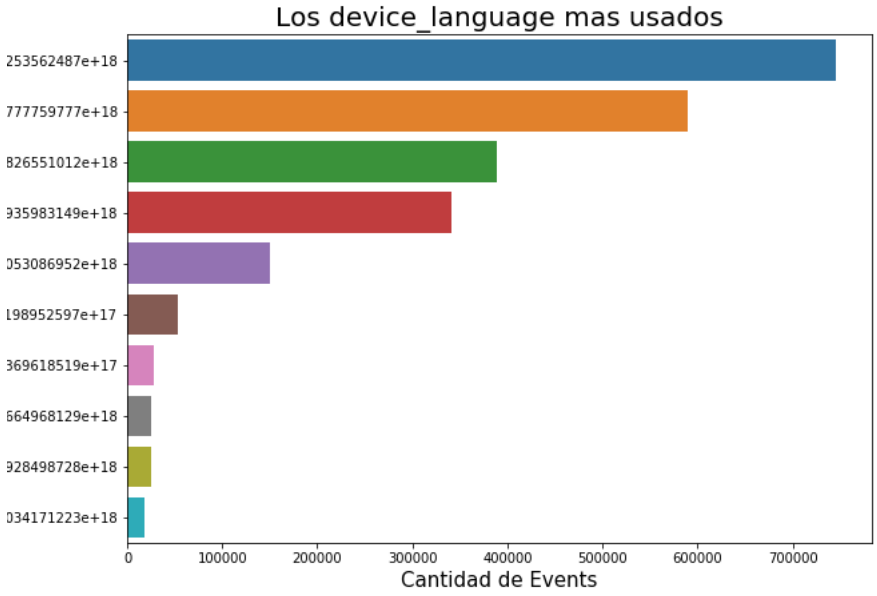
También, para mayor exactitud, podemos ver qué aplicaciones son las más usadas en la ciudad más recurrente:



Observar que a pesar del gran aporte que hace esta ciudad, las cinco aplicaciones más usadas en ella, no concuerdan con las cinco más usadas en total, pero sí están dentro del "top 20"total.

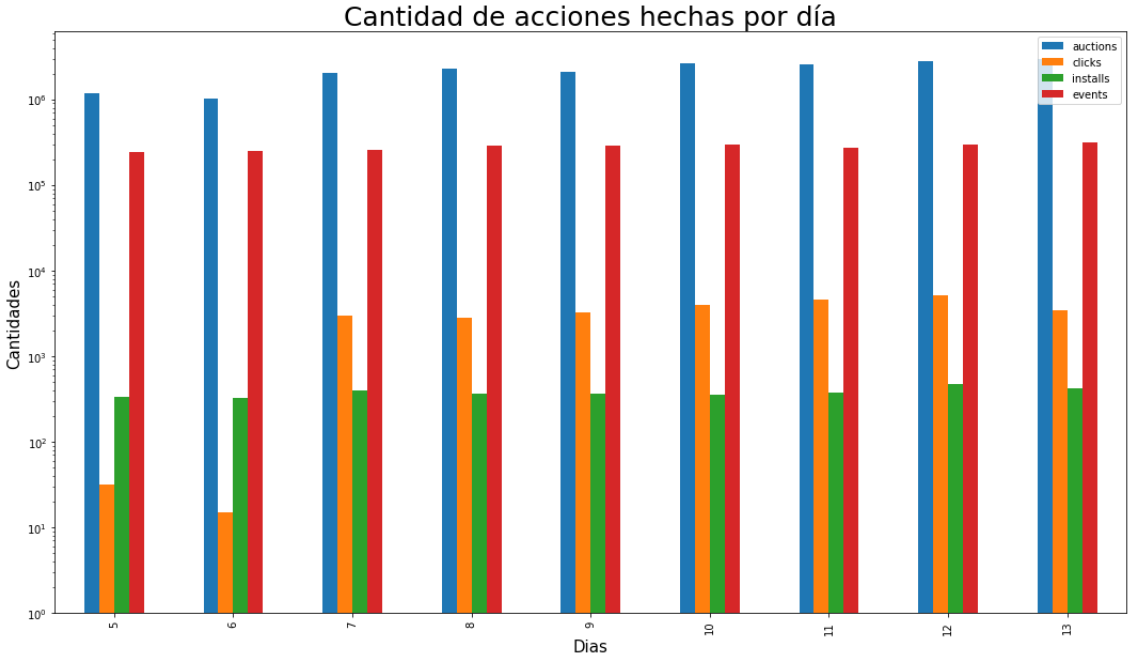
5.9. Lenguaje

También podemos obtener datos estadísticos sobre el lenguaje del dispositivo, como los 10 lenguajes con más cantidad de eventos asociada:



6. Relación encontradas entre las tablas

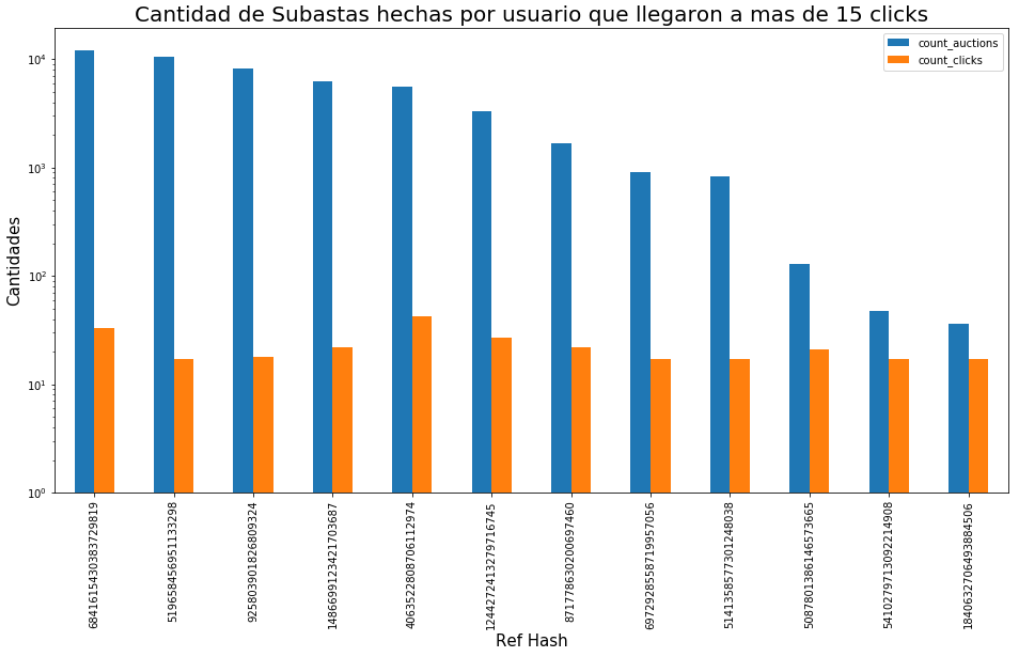
Se hace una comparación por día de las cantidades de subastas, clicks, instalaciones y eventos que se hicieron por día.



En todos los días predominan la cantidad de subastas con valores que van entre el millón de subastas y 3 millones. Las barras de color rojo que corresponden al de events, tiene un promedio de 2 millones de eventos por día. La columna de clicks, los dos primeros días toma valores muy chicos, de alrededor de 30 y 40 clicks pero lo interesante es ver que la cantidad de installs en estos días sigue siendo alta a pesar de eso. Por lo tanto se puede pensar que la instalación proviene de forma orgánica y no esta atribuida a la publicidad mostrada por Jammp. El resto de los días las acciones con menos cantidades son las de instalaciones, con valores menores a 500 instalaciones.

6.1. Relación entre cantidad de Subastas y Clicks realizados

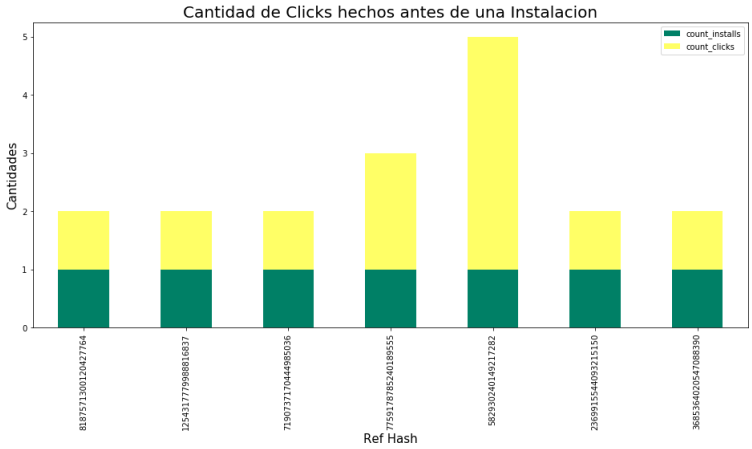
Se muestra en el siguiente gráfico la relación de subastas hechas hasta llegar a mas de 15 clicks por usuario. Se utilizo una escala logarítmica para poder ver mejor la diferencia.



La relación entre subastas hechas a un usuario y cantidad de instalaciones que hicieron los primeros 6 usuarios que aparecen en la tabla tiene una diferencia enorme. De 10 mil subastas hechas por ese ref_hash solo clickeo 15 veces y no instalo ninguna vez (se buscaron esos mismos ref_hash en la tabla de installs y no aparecen). Por lo tanto se puede decir que no conviene apostar por esos usuario y hacerlo por alguna que tenga una diferencia menor. Ya que no por mostrarle gran cantidad de publicidad a un usuario y por tener una buena cantidad de clicks lo lleva a instalar una aplicación, que es lo que al fin de cuentas lo que se pretende

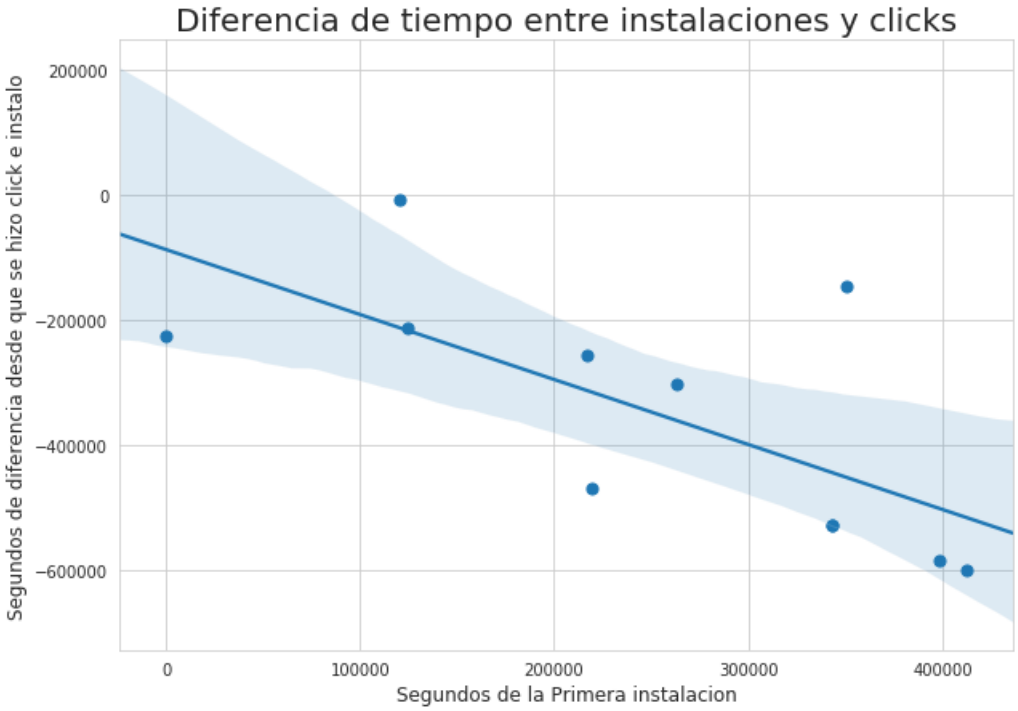
6.2. Relación entre cantidad de Instalaciones y Clicks realizados

En el siguiente gráfico se muestran los ref hash que hicieron al menos una vez click en la pantalla y instalaron.



La relación que se muestra es de 1 a 1 en la mayoría de los casos, es decir, hizo click y instalo. No se toma en cuenta en este caso si las dos acciones tienen una ventana de tiempo reducida para analizar si la instalación fue 'ganada' después del click o fue de forma orgánica, habiendo el usuario instalado mucho después por otro medio porque le intereso la aplicación. En uno de los casos se ve que el usuario hizo al menos 4 clicks antes de instalar 1 vez la aplicación. Claramente no podemos analizar porque surgió esta indecisión en el usuario para intentar de clickear tantas veces la pantalla antes de llegar a la descarga, pero podemos considerarlo como un caso excepcional.

Para ver en detalle si antes de cada instalación se realizaron clicks, realizamos el siguiente análisis:



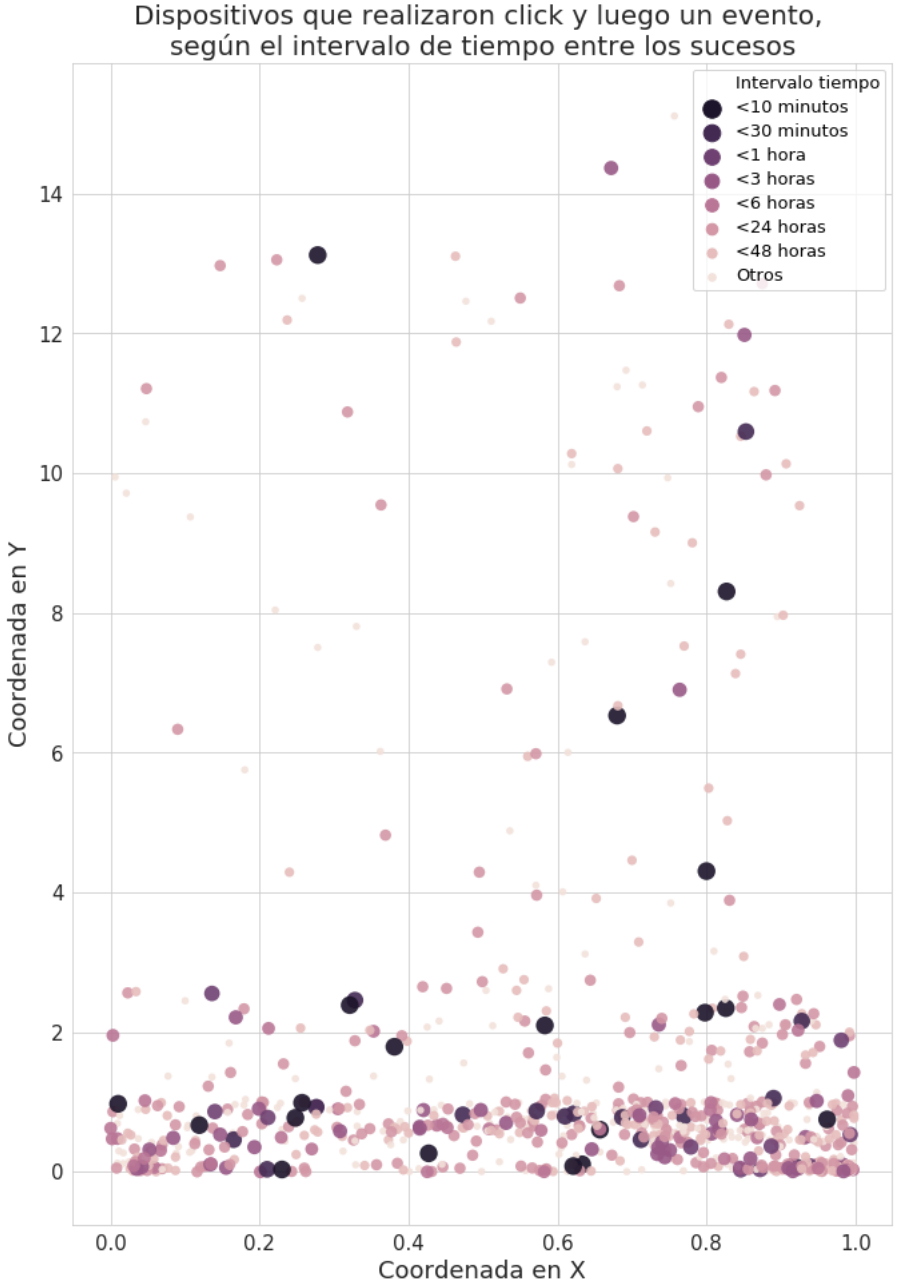
En el gráfico el eje X representa en tiempo en segundos en el que se hizo click, tomando el valor 0 el tiempo del primer click realizado entre todos lo que se encontraron con el ref_hash

en común con ambos DataFrames. En el eje Y se encuentra la diferencia en segundos que tubo hasta llegar a la instalacion. Se puede ver que todos los puntos estan por debajo del 0.

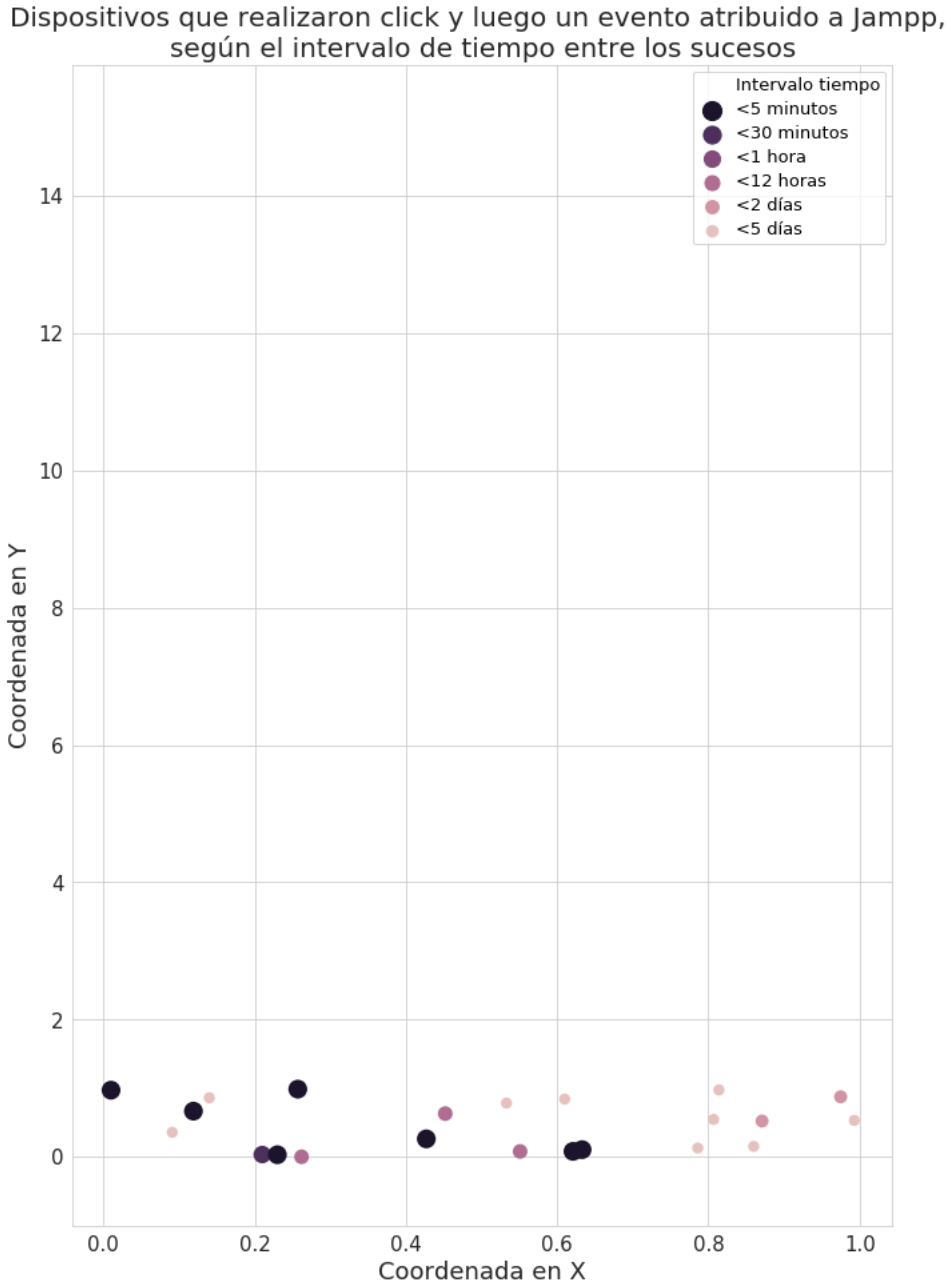
La linea de regresión tiene una pendiente negativa dado que ningún click se hizo antes de una instalación. Siguiendo este análisis, podríamos suponer que no hay clicks que terminaron en instalaciones.

6.3. Relación entre clicks y eventos

En esta sección se visualizarán los usuarios que hicieron un click y luego se les atribuye un evento. Se considerará el tiempo entre dichos sucesos para ver si hay zonas de la pantalla donde al clickear un usuario es propenso a luego realizar rápidamente un evento.



Como podemos ver, la distribución es bastante similar al gráfico realizado en la sección de clicks. Hay una banda inferior que concentra un número significativo de clicks, y luego algunos clicks esporádicos en el resto de la pantalla. No parece haber una zona especialmente propensa a producir un evento en la brevedad. Un problema con esta información es que el click y el evento pueden no estar relacionados. Para ver una relación más precisa, aunque no perfecta, es posible relacionar los usuarios que realizaron click y luego produjeron un evento atribuido a Jampp.



Este gráfico nos muestra que los únicos clicks que luego fueron sucedidos por un evento atribuido a Jampp se dieron en la franja inferior. Por último, habría que tener dos consideraciones con este análisis:

- No necesariamente el click registrado en el dataset desencadenó el evento. Es posible que el evento sea atribuido a Jampp gracias a un click realizado con anterioridad.
- Por lo visto en análisis anteriores, la banda inferior de la pantalla es por lejos la zona donde más frecuentemente se realizan clicks. Realizando un cálculo rápido, aproximadamente un 77 % de los clicks se realizaron entre las coordenadas 0 y 1 del eje Y. Esto nos dice que hay que tomar con cuidado el dato de que todos los puntos de esté gráfico caigan en esa zona. Es estadísticamente probable que puntos aleatorios caigan aquí.

7. Conclusión

A partir de la división del análisis en varias tablas se pudo ver en cuales casos los usuarios convierten con mayor frecuencia dependiendo de factores como modelo de celular, tipo de publicidad mostrada, horario en el que se realiza y datos geográficos. Aun así, aunque se subaste por aquellos que tienen más puntos a favor de convertir no podemos asegurar a priori que esto ocurra como tampoco podemos descartar a quienes se mostraron factores negativos de potenciales usuarios de las aplicaciones de los clientes de Jampp. Un gran impedimento a la hora del análisis de la información es la imposibilidad de relacionar dispositivos presentes en subastas y que luego hayan aparecido en algún otro dataframe. Esto produjo que nuestro análisis de funnels de usuarios haya sido acotado a pocos factores como realización de un click y luego un evento.