Facultad de ingeniería de la UBA

75.06/95.58 Organización de Datos

Trabajo práctico N°1
Análisis de datos

Primer cuatrimestre de 2019

Análisis exploratorio sobre los datos de Jampp

Integrante	Padrón	Correo electrónico
Gutiérrez, Matías	92172	matiasgutierrez@outlook.com
Calvo, Mateo Iván	98290	mate.-95@hotmail.com
Leal, Matías	99582	mleal@fi.uba.ar
Pelozo, Carlos Emanuel	99444	emanuel.pelozo@outlook.com

Nombre del grupo: ... (10)
Link al repositorio:
https://github.com/mateoicalvo/7506/tree/master/tp1
Kaggle:
https://www.kaggle.com/gutierrezmatias/an-lisis-exploratorio-y-visu

Índice

Parte I
Introducción

En el presente informe se presentan los resultados obtenidos al realizar análisis exploratorio sobre los datos de la empresa *Jampp*, dedicada al negocio de RTB.

Los datos brindados consistieron en cuatro datasets:

- Auctions: Con datos relacionados a las subastas.
- Installs: Con datos acerca de los dispositivos que realizaron instalaciones de las aplicaciones de los clientes de Jampp.
- Clicks: Con información acerca de las impresiones que originaron clicks.
- Events: Con datos realacionados a eventos producidos en las aplicaciones de clientes de Jampp.

Las siguientes secciones resumen los resultados obtenidos luego de realizar el análisis, junto a las conclusiones obtenidas.

Parte II
Análisis previo y limpieza de datos

1. Introducción

Si bien el volumen de los datos no fue demasiado masivo, se debieron tomar algunas consideraciones para manipular los mismos de manera ágil. Una primera inspección de los datasets permitió analizar la proporción de valores nulos en las columnas. Al mismo tiempo, se determinó el mejor tipo de datos para cada una de ellas, logrando así reducir considerablemente el tamaño en memoria de los datos (desde más de 2GB hasta 700 MB aproximadamente).

Los criterios elegidos para tomar decisiones en esta primera inspección de los datos fueron la proporción de valores nulos y la cantidad de valores posibles para cada columna.

Para la proporción de valores nulos: Como en algunos casos todos los valores de una columna fueron nulos, se decidió eliminar dichas columnas pues no aportaban al análisis. A pesar de que la mayoría de columnas (de todos los sets de datos) no contenían una cantidad significativa de valores nulos, en algunos casos específicos (como por ejemplo en la columna *trans_id* del dataset de *installs*) se encontraron columnas con un porcentaje de nulos mayor al 75 %. Aunque en un principio pareció más lógico eliminar tales columnas, se optó por mantenerlas a fines de intentar obtener alguna conclusión interesante en una inspección posterior. Con respecto a la imputación de datos, por otra parte, se eligió no completar los datos, ya que al estar anonimizados se podría llegar a conclusiones equivocadas.

Para la cantidad de valores distintos: Este aspecto resultó clave a la hora de reducir el tamaño en memoria de los datos: establecer el tipo de datos categórico para aquellas columnas cuya cantidad de valores únicos fuera mucho menor que la cantidad de valores totales permitió reducir el volumen drásticamente. Como mención especial se destaca el hecho de contar con fechas consistentes y sin valores nulos en todos los datasets, permitiendo la mayor reducción de memoria al simplemente cargarlas con el tipo de dato "datetime". Al poner el enfoque en la variabilidad de los datos dentro de cada columna, se pudieron observar algunas inconsistencias en los datos. Por ejemplo, en el dataset de *installs* se observaron dos países diferentes, cuando

la suposición era que sólo se contaba con datos de Uruguay. Otro aspecto interesante fue la cantidad de idiomas en el dataset de *events*: resultaron haber 186 idiomas diferentes registrados, algo que en principio parece no tener sentido. Para este tipo de situaciones resultó imposible arribar a una conclusión consensuada, ya que al estar los datos anonimizados no se pudo inferir que estuvieran mal recolectados o mal formados.

2. Reporte detallado de la primera inspección de los datos

Aquí se presentan las observaciones más interesantes que surgieron de analizar los tipos de datos, valores nulos y su proporción en las columnas de todos los sets de datos.

2.1. Dataset *Auctions*

- Se observó que la columna *auction_type_id* tenía todos sus valores nulos, por lo que no se consideró.
- Se descartó la columna *country*, al ser todos sus valores iguales y conocer que se trataba de Uruguay.
- No se encontraron fechas anómalas en la columna *date*: todas ellas se ubicaron dentro del rango de días entre el 5/3/2019 y el 13/3/2019.
- No se encontraron valores nulos en la columna *device_id*, por lo que siempre se conoce el dispositivo que apareció en la subasta.
- Para la columna *platform* se tiene una situación similar: En todos los casos se conoce si el dispositivo utiliza sistema operativo Android o iOS. La columna *ref_type_id* no parece aportar información extra pues además de no poseer valores nulos, tiene una correspondencia uno a uno con la columna *platform*.
- La columna *source_id*, referida al identificador de la fuente que originó la subasta, no tuvo valores nulos.

2.2. Dataset *Events*

- No se observaron fechas nulas ni anómalas.
- No se observaron valores nulos en la columna *event_id*.

- Las columnas *ref_hash* y *ref_type* no presentaron valores nulos, por lo que siempre se conoce el dispositivo que originó el evento. Los valores presentes en la columna *ref_type* se interpretaron admitiendo que el valor con más ocurrencias corresponde a los dispositivos Android, en consonancia con el dataset *Auctions*.
- No se encontraron valores nulos para la columna *application_id*; siempre se conoce la aplicación que generó el evento.
- En la columna *attributed* no se observaron valores nulos, por lo que siempre se puede determinar si un evento fue o no atribuido a Jampp.
- Se descartó la columna *device_countrycode*, pues todos sus valores fueron idénticos y estaban anonimizados de la misma manera que en el dataset *Auctions*. Evidentemente se trata de Uruguay.
- Aproximadamente el 60 % de los valores son nulos para la columna *device_os_version*, mientras que paara la columna *device_brand* se cuenta con un 53 % de valores nulos. Llamativamente, sólo el 4 % de los valores resultaron ser nulos para la columna *device_model*: quizá esta última se utiliza para inferir las dos anteriores. Por otro lado, en la columna *device_os* el 74 % de los valores son nulos, aunque dos de sus cuatro valores posibles predominan en apariciones. Esto pareciera indicar que se trata de Android e iOS.
- El 75 % de los valores en la columna *device_city* son nulos.
- Menos del 1 % de los valores en la columna *session_user_agent* son nulos, mientras que más del 99 % de los valores en la columna *trans_id* son nulos.
- Aproximadamente el 44 % de los valores en la columna *user_agent* son nulos. *C*asi no hay valores nulos en la columna *event_uuid*, son menos que el 1 %. Esta característica es compartida por la columna *kind*.
- La columna *carrier* tiene aproximadamente el 75 % de sus valores nulos.
- La columna *wifi* un 45 % de valores nulos. Esta distrubución llevó a no imputar los datos faltantes para no arribar a conclusiones equivocadas.
- El 75 % de los valores en la columna *connection_type* son nulos.
- Sólo el 3 % de los valores son nulos para la columna *device_language*, pero se encontró que hay 186 valores diferentes para los idiomas.
- No se encontraron valores nulos en la columna *ipaddress*.

2.3. Dataset *Clicks*

- No se encontraron fechas nulas ni anómalas.
- No se encontraron valores nulos en la columna *advertiser_id*.
- Todos los valores de la columna *action_id* fueron nulos.
- Nuevamente se encontró que había un único país registrado, por lo que se asumió nuevamente que era Uruguay.
- No se encontraron valores nulos para las columnas *latitude* y *longitude*.
- Si bien no se encontraron valores nulos para la columna *wifi_connection*, es interesante ver que todos los valores son False. Parece a priori muy extraño que todos los clicks registrados hayan provenido de una conexión de red móvil.
- Menos del 1 % de los valores en la columna *trans_id* resultaron ser nulos.
- Sólo el 0,05 % de los valores resultaron nulos en las columnas relacionadas al sistema operativo (*os_minor* y *os_major*).
- En la columna *agent_device* se observó que el 88 % de los valores fueron nulos.
- La columna *brand* tiene una gran cantidad de valores nulos (76 %) mientras que la columna *spects_brand* no tiene ninguno. Lo anterior resulta, cuanto menos llamativo, pues (según la descripción de columnas brindada), ambas parecen estar íntimamente relacionadas. Quizá no se le permita a los dispositivos ocultar las especificaciones en las plataformas de anuncios.
- La columna *timeToClick* posee un 13 % de valores nulos, pero las estadísticas calculadas para la misma acusan la presencia de *outliers*. Si bien un tiempo medio de casi 4 minutos no parece tan descabellado (hay anuncios que son videos), un tiempo máximo de casi cinco horas no parece tener sentido.
- Las columnas *touchX* y *touchY* tienen un 13 % (aproximadamente) de valores nulos. Únicamente la columna *touchY* parece tener valores anómalos, siempre y cuando se considere que deberían estar normalizados.
- Al igual que en los datasets *Auctions* y *Events*, las columnas *ref_type* y *ref_hash* no poseen valores nulos.

2.4. Dataset *Installs*

- Nuevamente, no hay fechas anómalas ni valores nulos en las fechas.
- Al igual que en los datasets anteriores, no hay valores nulos para las columnas *application_id*, *ref_type* y *ref_hash*.
- Todos los valores en la columna *click_hash* son nulos.
- No hay valores nulos en las columnas *attributed* e *implicit*, pero llama la atención que todos los valores en *attributed* sean False, siendo que uno de los objetivos de Jampp es lograr instalaciones.
- No se observaron valores nulos en la columna *device_countrycode*, pero ahora no aparece un único país posible sino dos. Si bien se supone que uno de ellos es Uruguay porque está anonimizado de la misma manera que en los demás datasets, no se puede determinar cuál es el otro.
- Las columnas relacionadas a las características de los dispositivos exhiben un comportamiento similar al encontrado en el dataset *Clicks*: 69 % de valores nulos para la columna *device_brand* y menos del 0,1 % de valores nulos para la columna *device_model*.
- La columna *session_user_agent* tiene el 1 % de sus valores nulos, mientras que en la columna *user_agent* casi la mitad de los valores son nulos.
- Las columnas *event_uuid* y *kind* poseen exactamente la misma cantidad de valores nulos (ambas tienen un 75 % de los mismos),
- un 49 % de los valores en la columna *wifi* son nulos. Sin embargo, no ocurre que todos sean falsos como en el dataset *Clicks*.
- Casi en su totalidad, los valores de la columna *trans_id* son nulos. Sólamente cuatro de 3406 valores no lo son.
- Ningún valor en la columna *ip_address* es nulo, al igual que en el dataset *Clicks*.
- La columna *device_language* tiene apenas un 1 % de sus valores nulos. Aquí no aparecen 186 valores de idioma posibles como en el dataset *Events*, sino que se encuentran 30 idiomas diferentes.

Parte III
Análisis de datos

1. Events

Esta sección corresponde al análisis del dataset "events", cuyos datos provienen de las aplicaciones clientes de Jampp y contienen el registro de eventos realizados por los usuarios dentro las mismas. Como uno de los objetivos de la plataforma de marketing es conseguir el "retargeting"de un evento (siendo esto la acción de fomentar la realización del mismo por primera vez o provocar una repetición), gran parte del análisis se realizó comparando los valores de eventos totales y eventos atribuidos a *Jammp*.

1.1. Sobre los eventos atribuidos

Al analizar la cantidad de eventos atribuidos, se llegó a que una proporción muy pequeña de los mismos (menos del 1 %) fue atribuida a Jampp. Lo anterior es muy llamativo, aunque no se pueden realizar conclusiones más elaboradas debido al desconocimiento del negocio (lo cual es crítico a la hora de realizar Data Science). En base a las explicaciones del personal de Jampp, se pudo comprender que no siempre un evento se atribuye, (aunque corresponda hacerlo) por lo que esta situación seguramente se contemple entre Jampp y sus clientes.

1.2. Evolución y distribución en el tiempo

En esta sección se analiza como se distribuyen los eventos durante diferentes períodos de tiempo.

Evolución a través de los días

En la figura 1.2 se muestra cómo evolucionó la cantidad de eventos para cada día.

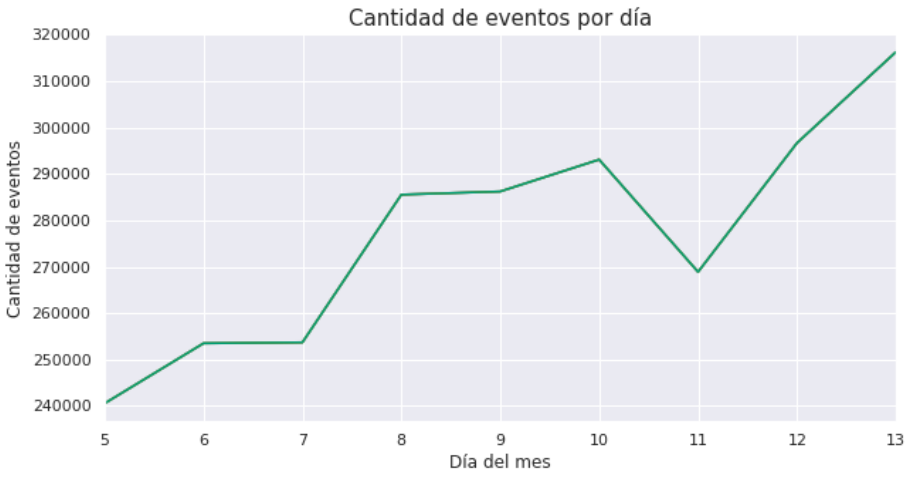

Figura 1: Evolución en la cantidad de eventos por día

Se puede observar que para gran parte de los días se tiene una tendencia a aumentar en cantidad de eventos con respecto al día anterior, el primer aumento realmente significativo se observa para el día Viernes 8 que tiene sentido siendo el comienzo del fin de semana.

Sin embargo, un caso especial para el que vemos un descenso es el día Lunes 11, que tomando la hipótesis usada en el día Viernes puede deberse a que coincide con un nuevo inicio de la semana laboral. Luego de esta baja, se registra nuevamente un aumento notorio en la cantidad de interacciones con las aplicaciones, correspondiendo con la tendencia de aumentar la cantidad de sus usuarios activos con el pasar de los días, lo que lleva a aumentar el registro de interacciones.

1.2.1. Distribución a lo largo del día

Una vez analizada la cantidad de eventos por día, se consideró la distribución de la cantidad de eventos para cada hora del día.

Debido a que los datos provenían de 9 días se tomo el rango de fechas desde el 7 de Marzo hasta el 13 de Marzo, teniendo en cuenta la evolución de los eventos a lo largo del tiempo (lo que nos brindaría más información para un rango de siete días).

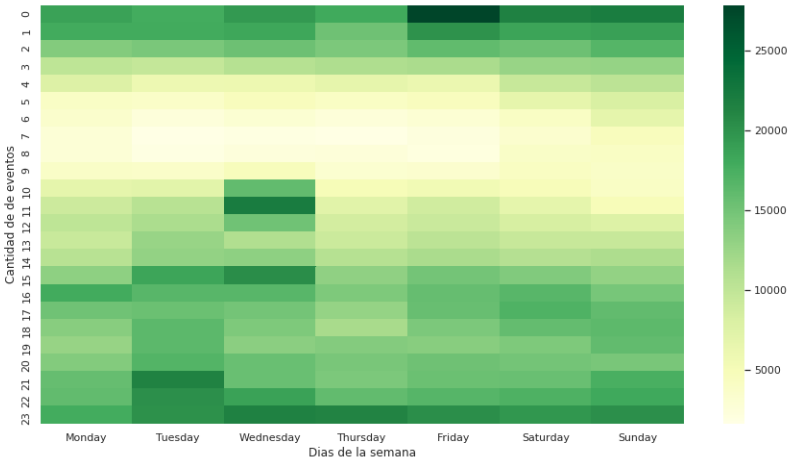

Figura 2: Distribución de los eventos según la hora del día

Aquí se ve que la gran mayoría de eventos se dan en el rango que abarca desde las 21 horas hasta las 1 de la madrugada. El dato interesante que se logra ver es que para el día Miércoles la mayor cantidad de eventos esta concentrado a las 11 de la mañana. También se ve que el horario de las 00 horas del día el día viernes es el que más cantidad de interacciones registra.

Finalmente pareciera ser que la franja que abarca desde las 5hs hasta las 9 hs es el rango con menos actividad.

1.3. Eventos más presentes

Aquí se centró el análisis en relación a cada evento particular dentro de las aplicaciones.

En la Figura 3 se puede observar cuales son los eventos que más se repiten entre los usuarios de las aplicaciones, utilizando el dato *ref_type* para inferir el sistema operativo. El primer puesto se lo lleva el evento con ID 22 con aproximadamente 600 mil eventos registrados, seguido por los eventos 31 y 161 en segundo y tercer lugar respectivamente. Se puede destacar que los primeros tres puestos se llevan la mitad del total de eventos sumando 1.25 millones.

Con respecto al sistema operativo, se puede observar que Android domina en la mayoría de las ocurrencias, salvo para los eventos 162 y 33 donde iOS posee un poco más del 50 % de la cantidad total.

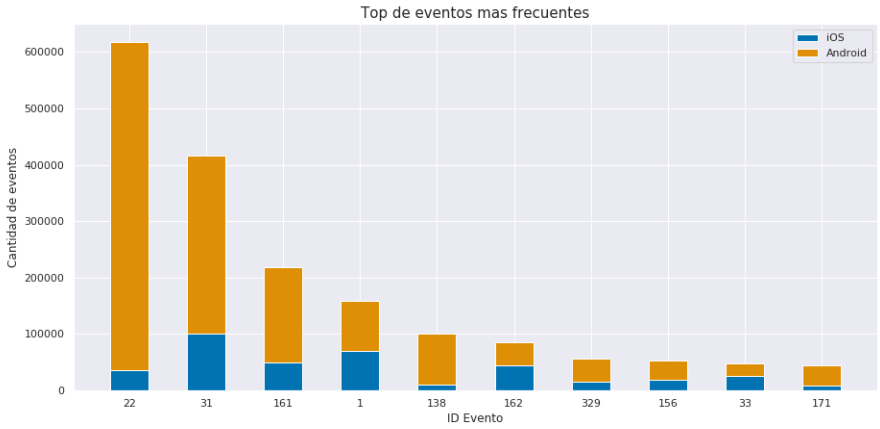

Figura 3: Eventos más populares

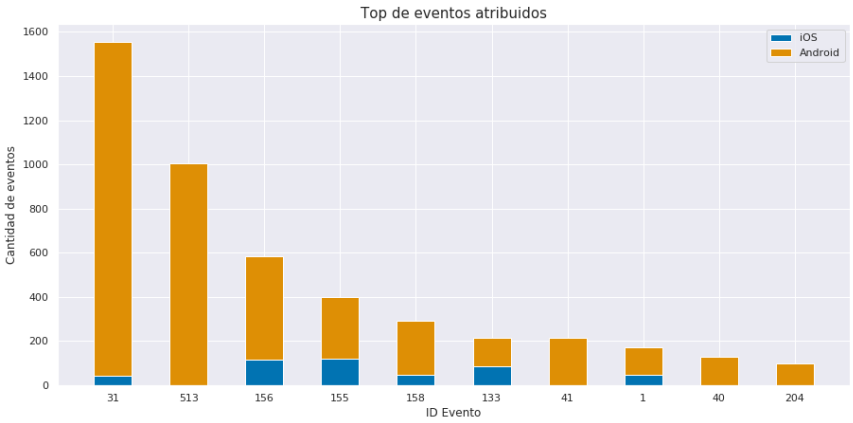

Figura 4: Eventos atribuidos más populares

En la Figura 4 podemos observar los eventos más populares con el foco puesto solamente a los eventos atribuidos.

Aquí se puede apreciar que la mayoría de los eventos involucrados anteriormente cambian, así como sus primeros puestos. Como el evento atribuido más popular nos encontramos con el ID 31, que ocupaba el segundo lugar entre las interacciones más frecuentes.

También se puede ver que en este campo el sistema operativo Android es el que domina en todos los puestos, donde iOS en ninguna de las posiciones llega al 50 % de los dispositivos.

1.4. Cantidad de eventos para red Wifi

En esta sección nos enfocaremos en los datos brindados por el atributo *wifi* obtenido de los datos, cómo se explicó anteriormente, estos datos corresponden solo al 65 % de los eventos totales, por lo que el análisis solo es una aproximación.

Como se observa en la Figura 5, el mayor porcentaje de eventos se da con una conexión a red Wifi, aproximadamente el doble.

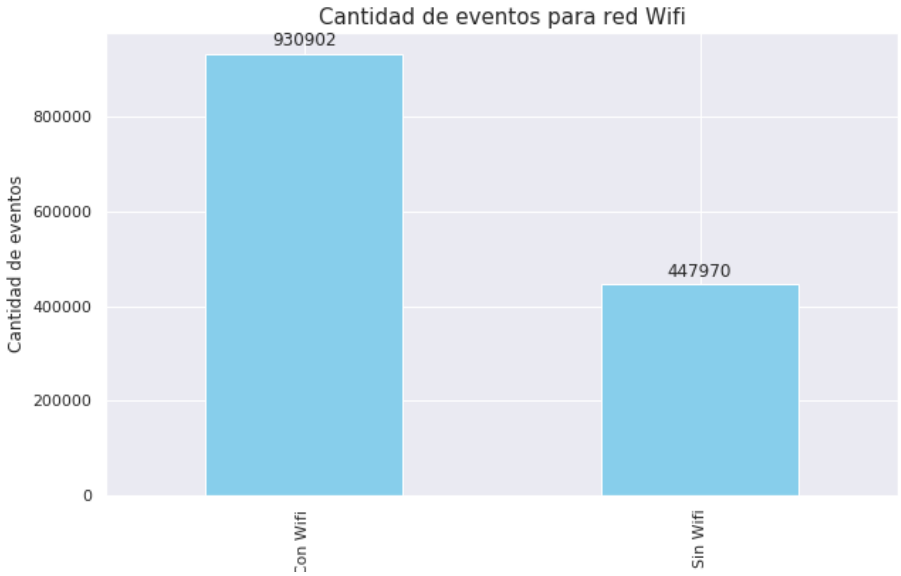

Figura 5: Distribución de eventos segun la red Wifi

Veamos como es esta distribución entre los eventos más populares:

Para la Figura 6 vemos que el evento que tiene la distribución más pareja es el evento 22 que ocupa el primer lugar, el segundo puesto, también es el segundo en este sentido. Luego, para el resto de eventos se ve una clara superioridad de interacciones realizadas a través de una red Wifi.

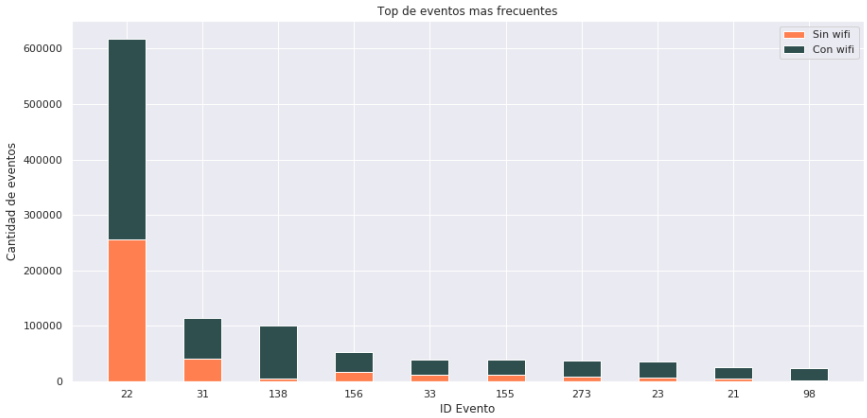

Figura 6: Eventos más populares por red Wifi

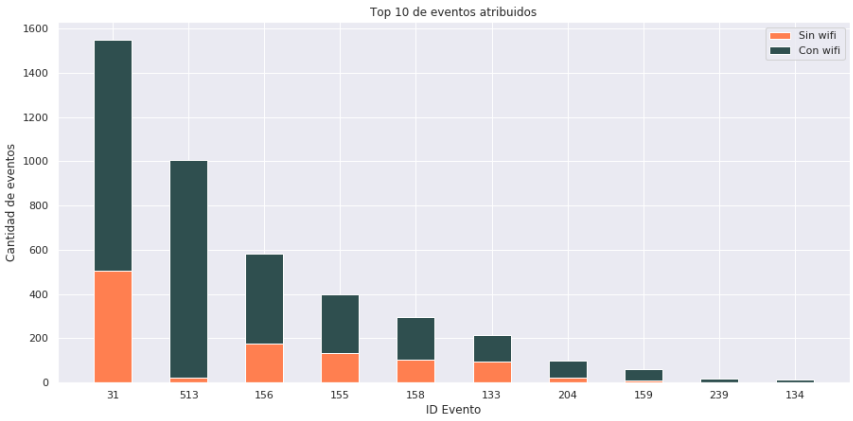

Figura 7: Eventos atribuidos más populares por red Wifi

Ahora nos enfocamos en los eventos atribuidos, aquí, se puede ver una distribución mucho más pareja que la anterior para los primeros seis puestos, excluyendo el segundo, en el cual la mayoría de sus interacciones fueron hechas a través de una red Wifi.

1.5. Datos de los dispositivos

En esta parte nos abocaremos a los datos de los dispositivos pertenecientes a los usuarios de las aplicaciones clientas de *Jammp*

1.5.1. Top marcas

A la hora de analizar las marcas se obtuvo que hay 250 marcas distintas dentro del set de datos. Comenzaremos viendo las marcas más populares.

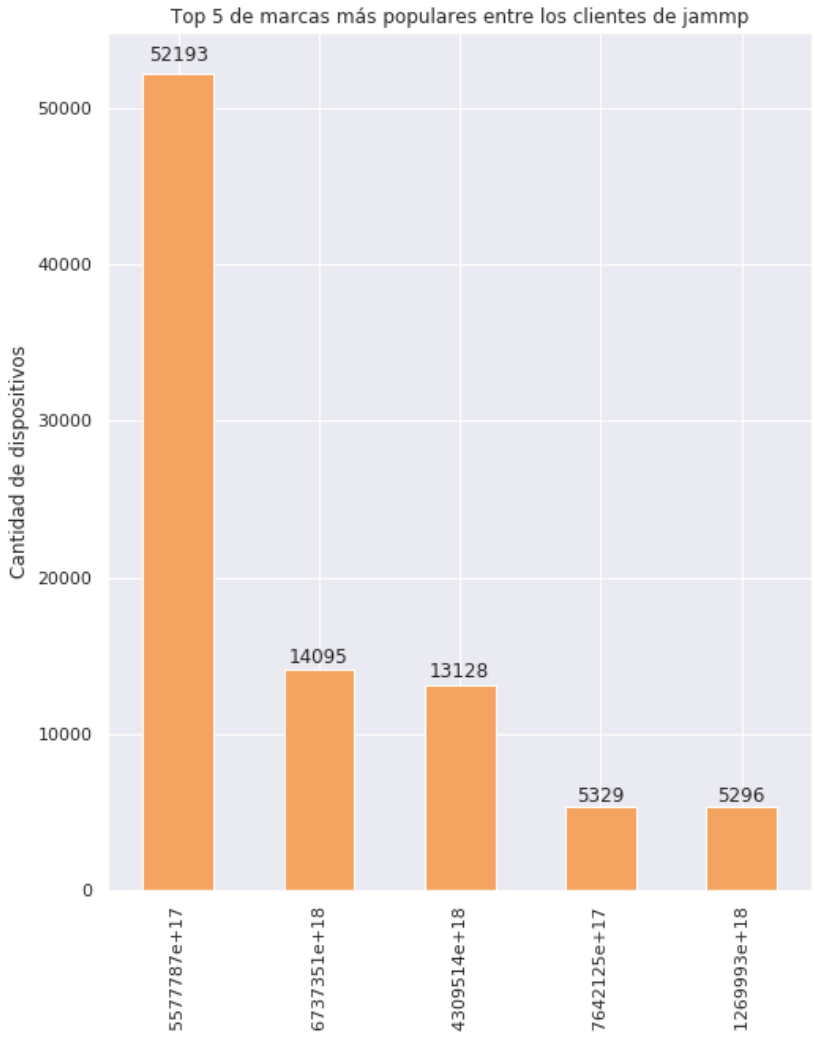

Figura 8: Marcas más populares dentro de los usuarios.

La Figura 8 nos muestra una clara superioridad para el primer puesto con respecto a los demás, esta marca es una clara dominadora dentro de los usuarios. Luego el segundo y tercer puesto tienen una cantidad bastante cercana de dispositivos, este atributo es compartido también por sus inmediatos

perseguidores.

Cabe destacar que para la obtención de este top y debido a que dentro del registro de eventos puede haber dispositivos repetidos, solo se toma a cada dispositivo individualmente, independientemente de cuantas veces aparezca en el registro.

A continuación nos interesa ver cuales fueron las marcas que más eventos realizaron:

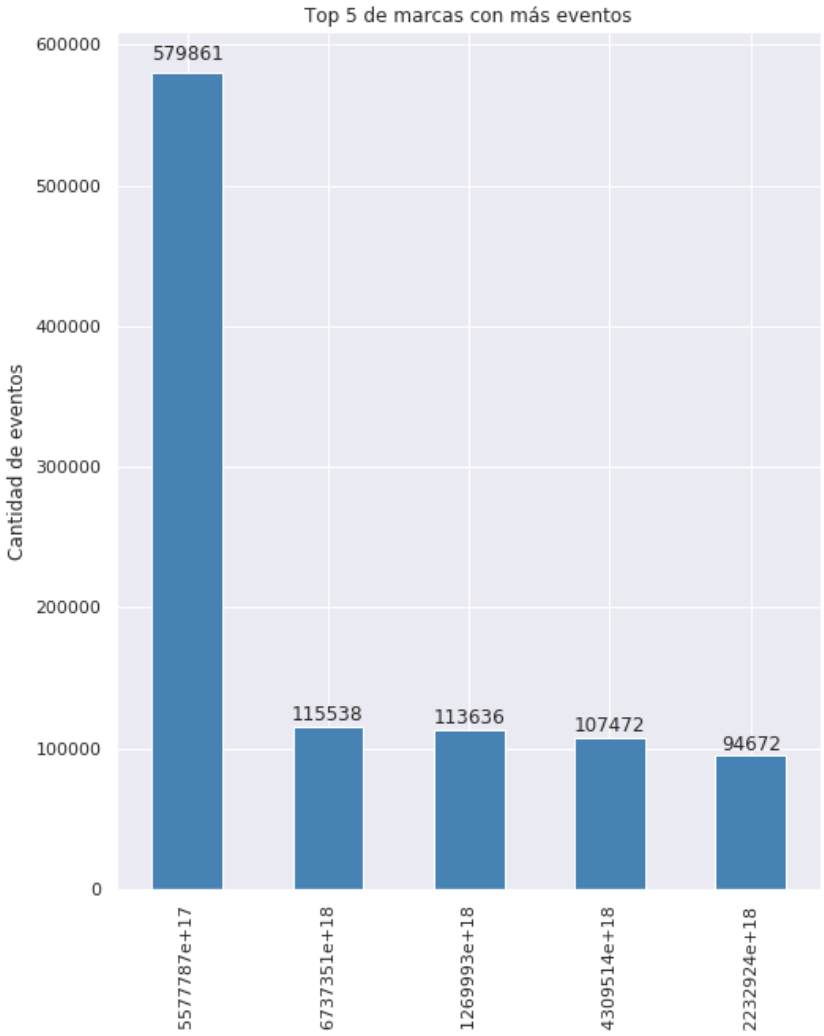

Figura 9: Marcas que más eventos realizaron.

El primer puesto coincide con la marca más popular entre los clientes, esto es algo lógico por la gran cantidad de dispositivos que posee esta marca, esto conlleva a tener mayor cantidad de interacciones totales.

Lo interesante de ver es que las siguientes cuatro posiciones tienen una distribución muy similar en cuanto a cantidad de eventos a pesar de no ser así en cantidad de dispositivos. Esto es así de tal manera que la marca que ocupaba el cuarto puesto entre las que más dispositivos tenían ya no aparece en este top, reemplazada por la quinta marca que vemos en el segundo gráfico.

1.5.2. Modelos de dispositivos más populares

A continuación echaremos un vistazo a los modelos de dispositivos, como vimos anteriormente, dentro del set de datos hay 250 marcas para las cuales tenemos 2624 modelos distintos, veamos los más populares:

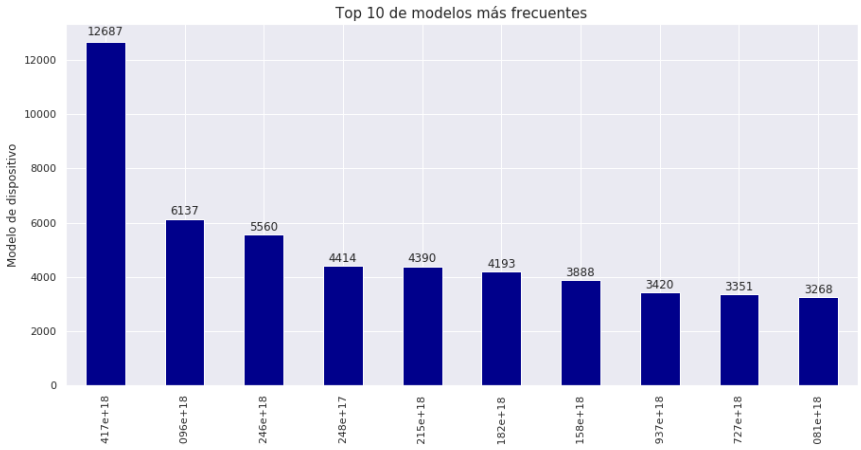

Figura 10: Marcas que más eventos realizaron.

Para el top de modelos más populares se puede ver que el primer lugar tiene aproximadamente el doble de dispositivos que el segundo, y luego la disminución entre los mismos va siendo más uniforme.

1.6. Dispositivos que más interactuaron

Lo siguiente que veremos será cual es el top de los dispositivos que más eventos realizaron dentro de las aplicaciones.

En la Figura 11 se puede ver que el rango de eventos para estos dispositivos es entre 2900-1900 eventos para los 9 días que comprenden los datos.

No se aprecia una gran varianza entre puesto y puesto, la mayoría de posiciones varían aproximadamente entre 100 y 200 eventos.

Figura 11: Dispositivos que más eventos realizaron.

1.7. Top Aplicaciones

Del set de datos pudimos obtener que hay eventos registrados para 269 aplicaciones distintas, vamos a ver cuales son las más populares.

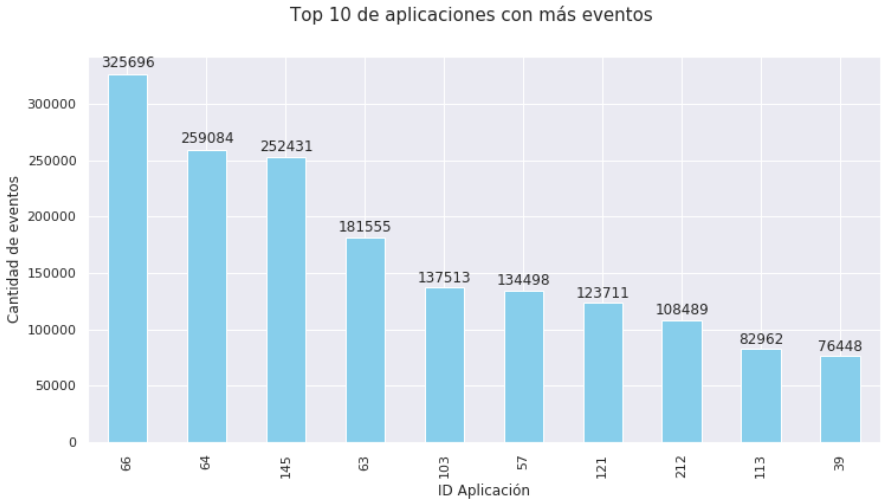

Figura 12: Aplicaciones que más eventos tuvieron.

El gráfico nos muestra que los primeros tres puestos para las aplicaciones tienen una gran diferencia con el resto, estas, con ID: 66, 64 y 145 tienen más

de 250 mil eventos registrados, la primera esta rondando los 325 mil. Estas tres juntas abarcan aproximadamente el 35 % de los eventos.

1.8. Eventos y clicks

En el siguiente apartado se especificará el procedimiento usado para unir los datasets 'clicks' y 'events', y cuales fueron los resultados obtenidos.

Para comenzar se comenzó por consultar cuantos de los dispositivos (a través de su *ref_hash* del dataset 'events' se encontraban dentro de los datos del dataset 'clicks', esto nos arrojo un total de 1196 dispositivos.

Finalmente nos propusimos encontrar cuales eran los clicks que llevaron a *Jammp* a conseguir una atribución de evento, luego de filtrar los datos para dispositivos que poseían un evento atribuido, nos quedo un total de:

- 26 dispositivos.
- 52 clicks.
- 173 eventos.

Finalmente, para encontrar la forma de relacionar un click con un evento se utilizo la hipótesis de que el click que llevaba a un evento atribuido debería ser aquel que tenga la fecha anterior más cercana al evento.

Utilizando este criterio se obtuvo:

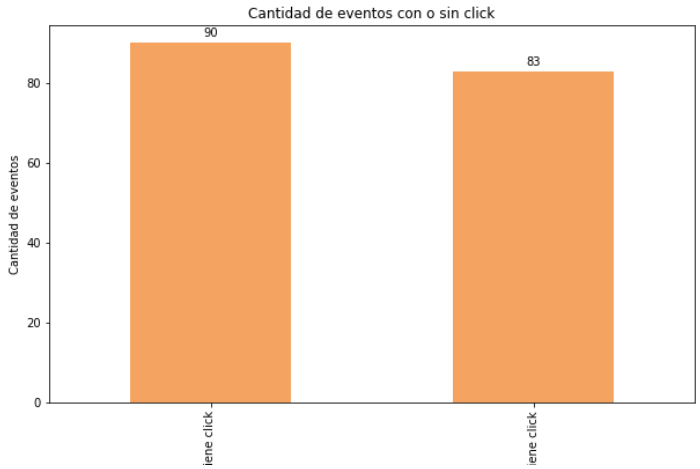

Figura 13: Cantidad de eventos filtrados que poseen click

Finalmente para los eventos que no poseen click la razón más logica es que los clicks que llevaron a ese evento no estan dentro del periodo de tiempo estudiado.

2. Installs

En la siguiente sección se analizará el dataset de 'installs' tanto de forma aislada como conjunta con los otros datos del dominio. Los datos de esta sección provienen de las instalaciones de clientas de Jammp y son brindados por ellos. Es sumamente importante entender que esto no significa que las mismas sean instalaciones provocadas por la plataforma de marketing.

2.1. Contexto

En los 9 días de datos brindados se hicieron un total de 3412 instalaciones de las cuales ninguna se atribuyó a Jammp; ya sea porque el usuario la instaló sólo, se logró mediante alguna publicidad de la competencia o porque no se notificó la atribución. Sin embargo 865 instalaciones fueron implícitas, es decir instalaciones en las cuales Jampp mostró publicidad y un usuario clickeó. Después, el usuario instaló la aplicación. No sé la dieron atribuida a Jampp pero se toma como propia porque se mostró publicidad antes que ocurra

2.2. Variación en el tiempo

Las instalaciones variaron significativamente según el día de la semana y la hora. Esto se puede observar en las figuras de la sección.

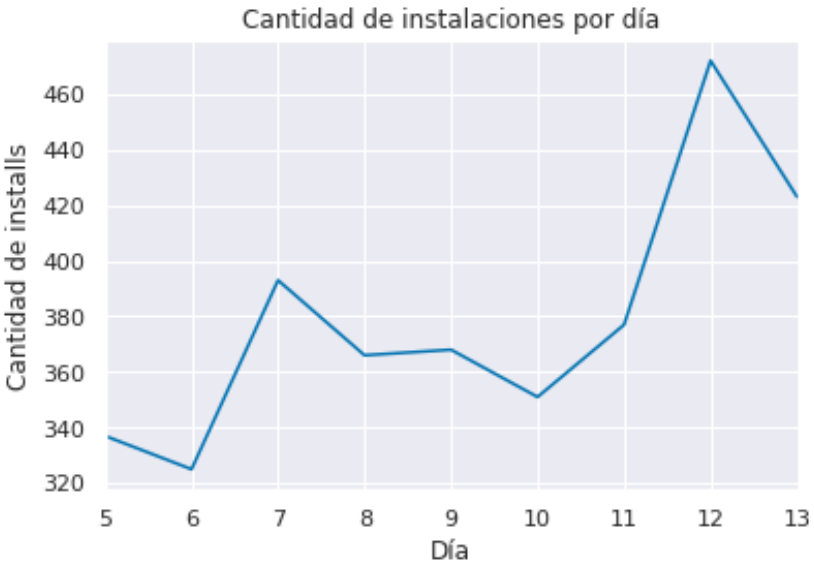

Figura 14: Cantidad de instalaciones por día

En la figura de arriba se ve que desde el día 11 las instalaciones crecen hasta llegar a un máximo de 472. Este día contrario a lo naturalmente pensado fue un día de semana, en específico el martes. El día con menos instalaciones fue el martes 6. Mientras que durante el fin de semana se mantuvo casi constante rondando las 360 instalaciones por día.

Las instalaciones también variaron según la hora del día pero de forma consistente durante toda la semana, siendo el rango desde las 15hs hasta las 00hs las más congestionadas mientras que desde las 5hs hasta las 10hs las instalaciones fueron virtualmente nulas.

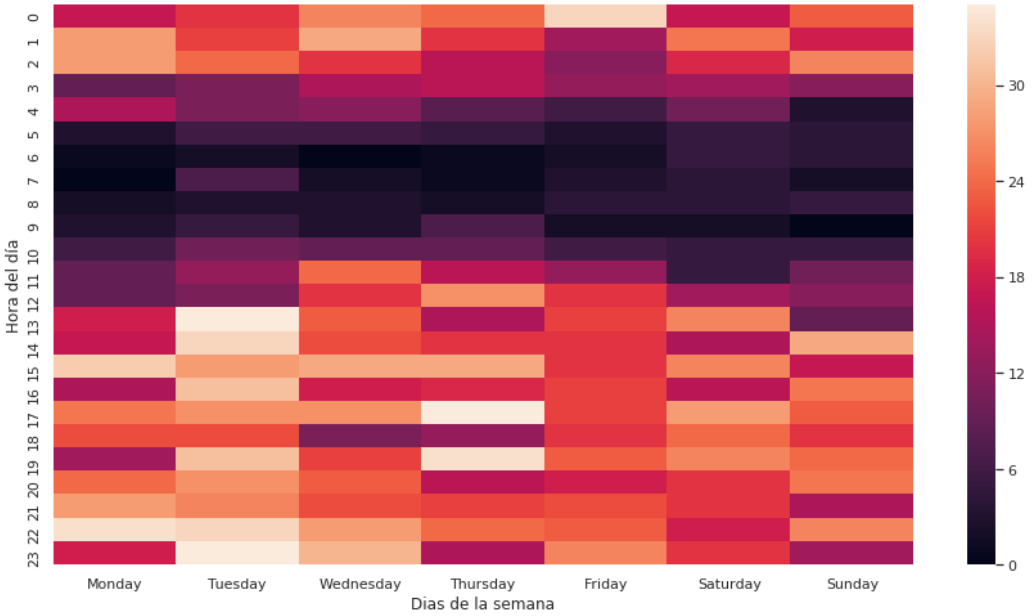

Figura 15: Cantidad de instalaciones por hora y día

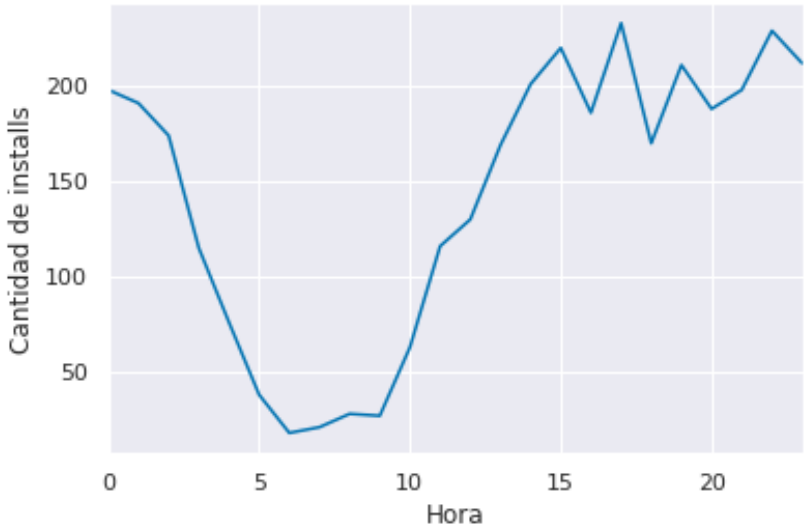

Figura 16: Cantidad de instalaciones por hora

Si se especifica la variación en el tiempo con instalaciones implícitas se observa que la tendencia es igual que tomando en cuenta todas. Respecto a

las instalaciones por hora son dos gráficos muy parecidos. Lo único a destacar como una diferencia son las dos picos de instalaciones diarias. Si bien estos corresponden al mismo día, la diferencia entre ellos es más pronunciada tomando las instalaciones en su totalidad.

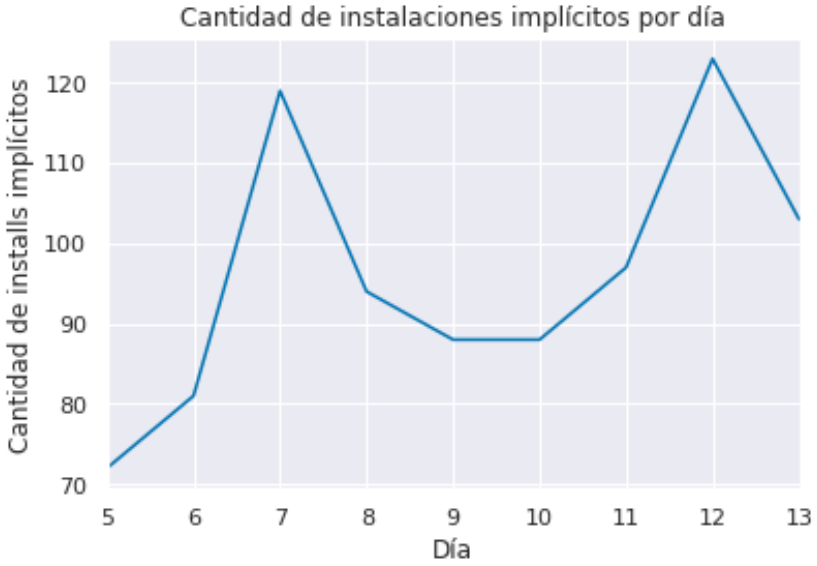

Figura 17: Cantidad de instalaciones implícitas por día

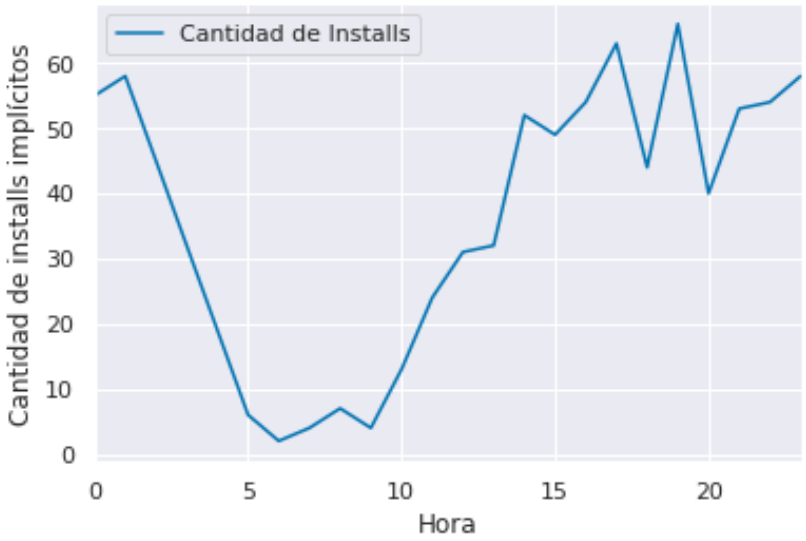

Figura 18: Cantidad de instalaciones implícitas por hora

2.3. Estado del dispositivos

Si se analizan los dispositivos en donde ocurrieron las instalaciones se ve muy claramente que una marca de dispositivos domina en la cantidad de instalaciones respecto al resto teniendo más del doble de instalaciones. Es necesario ponderar con la cantidad de dispositivos de esta marca en el país analizado para sacar conclusiones correctas.

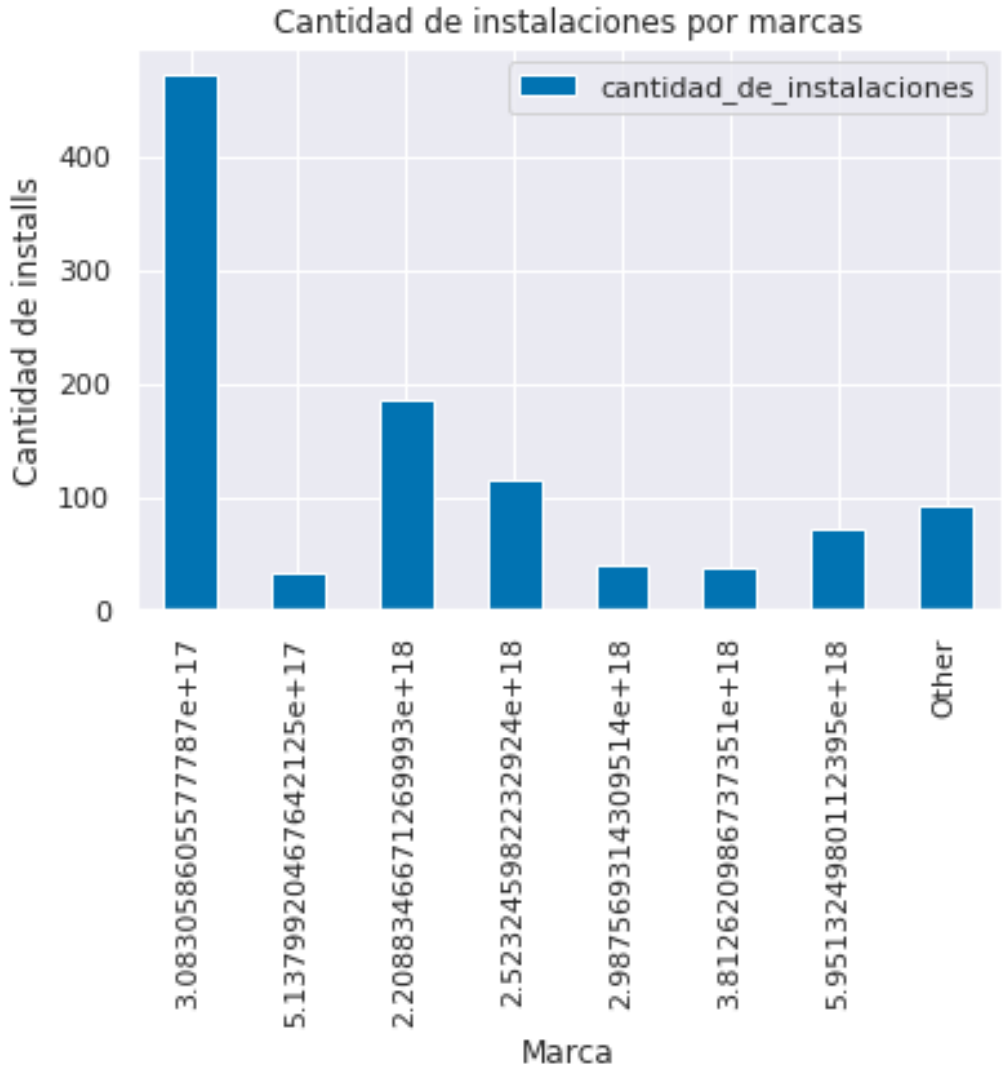

Figura 19: Instalaciones por marca

Analizando por sistema operativo (OS), se ve como se divide en un 60 % para el sistema con más instalaciones y un 40 % para el que menos instala-

ciones tiene. Al igual que en el análisis por marcas es necesario ponderar con la cantidad de dispositivos.

Una tendencia clara se puede ver en la figura 2.3. El 80 % de las instalaciones se dan cuando el usuario está conectado a una red de wifi.

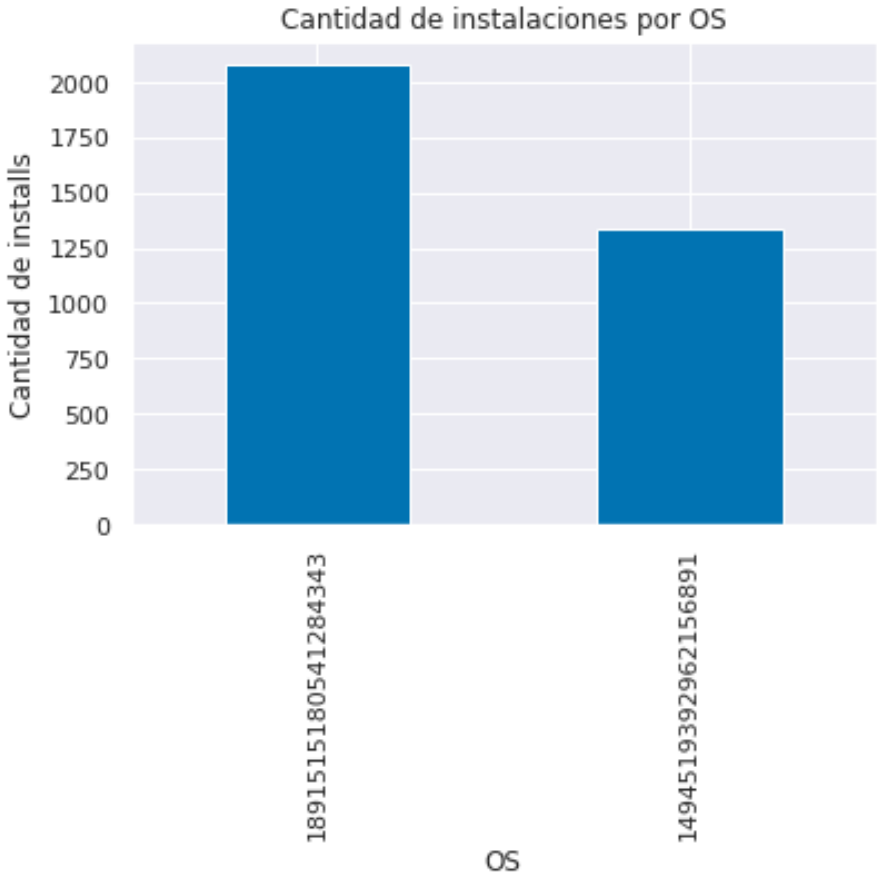

Figura 20: Instalaciones por OS

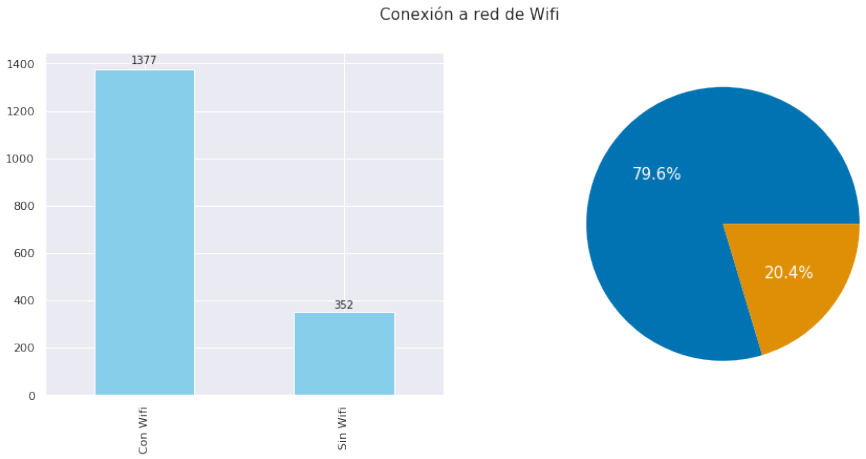

Figura 21: Instalaciones con WIFI

2.4. Eventos e Installs

En el gráfico inferior se muestra el evento anterior a un install, es decir el que hizo el usuario antes de instalar la aplicación. Se puede considerar que este lo llevó a instalarla. Estos eventos son dentro de las aplicaciones clientas de Jampp. Se ve que dos son los mayores causantes de instalaciones al tener una cantidad consecuente exponencialmente más grandes que el resto.

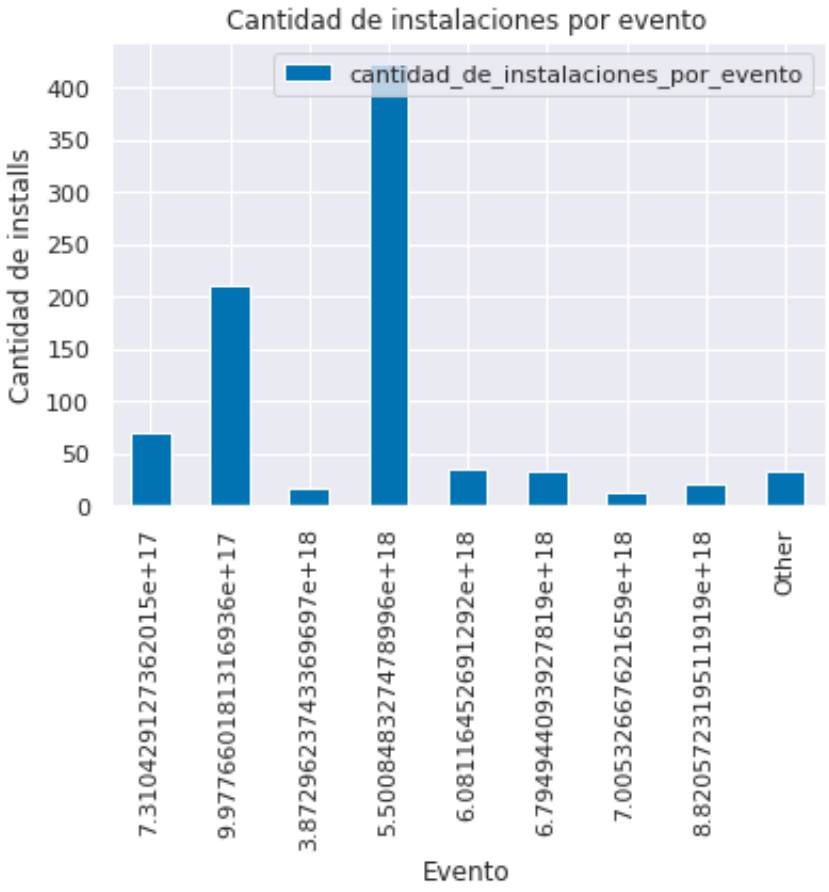

Figura 22: Evento anterior al install

2.5. Aplicaciones

Analizando las aplicaciones instaladas se ve como son dos las dominantes respecto al resto. La aplicación más instalada es la de ID número 7 con más de 800 instalaciones, seguida de la número 9 rondando las 700. El resto de aplicaciones no pasan de las 400.

Esta diferencia no se ve tanto en las instalaciones implícitas ya que la diferencia entre las 5 más instaladas no es tan grande.

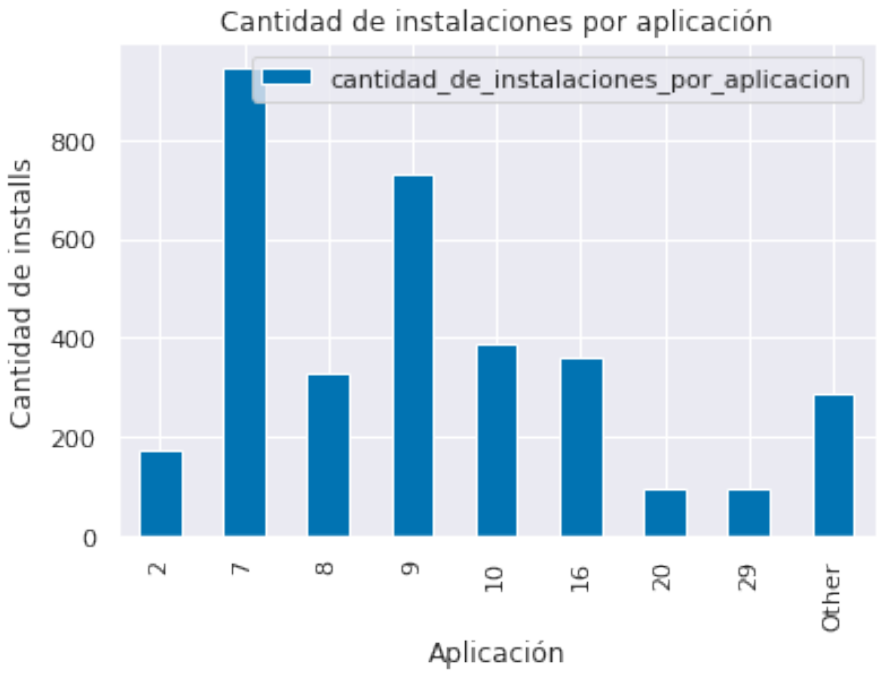

Figura 23: Instalaciones por aplicaciones

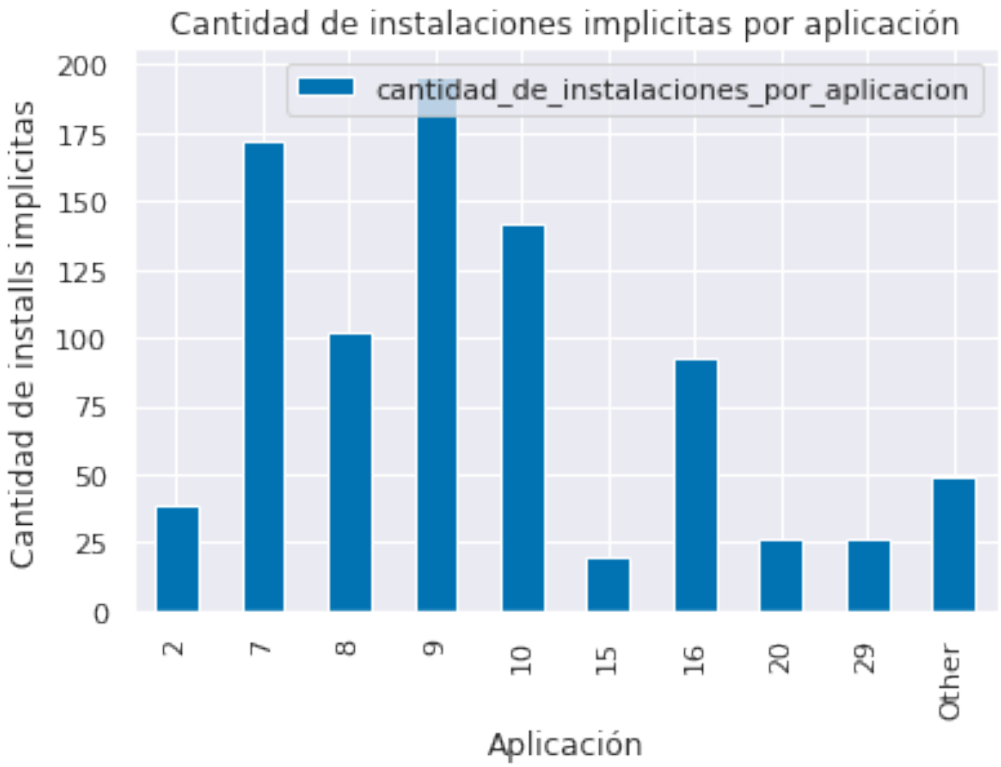

Figura 24: Instalaciones implícitas por aplicaciones

3. Clicks

A continuación se analizan los clicks de los usuarios en la plataforma correspondiente. Estos representan el punto de acceso de los usuarios a los anuncios publicados.

3.1. Posiciones de los clicks en la pantalla del dispositivo

Se puede observar que la mayoría de los clicks se dieron en la parte inferior del dispositivo donde se mostraba el anuncio, tanto para casos en los que se sostenía el dispositivo horizontalmente como verticalmente. para la posición vertical la región inferior y para la posición horizontal la parte derecha del gráfico, siendo la posición vertical la mas utilizada.

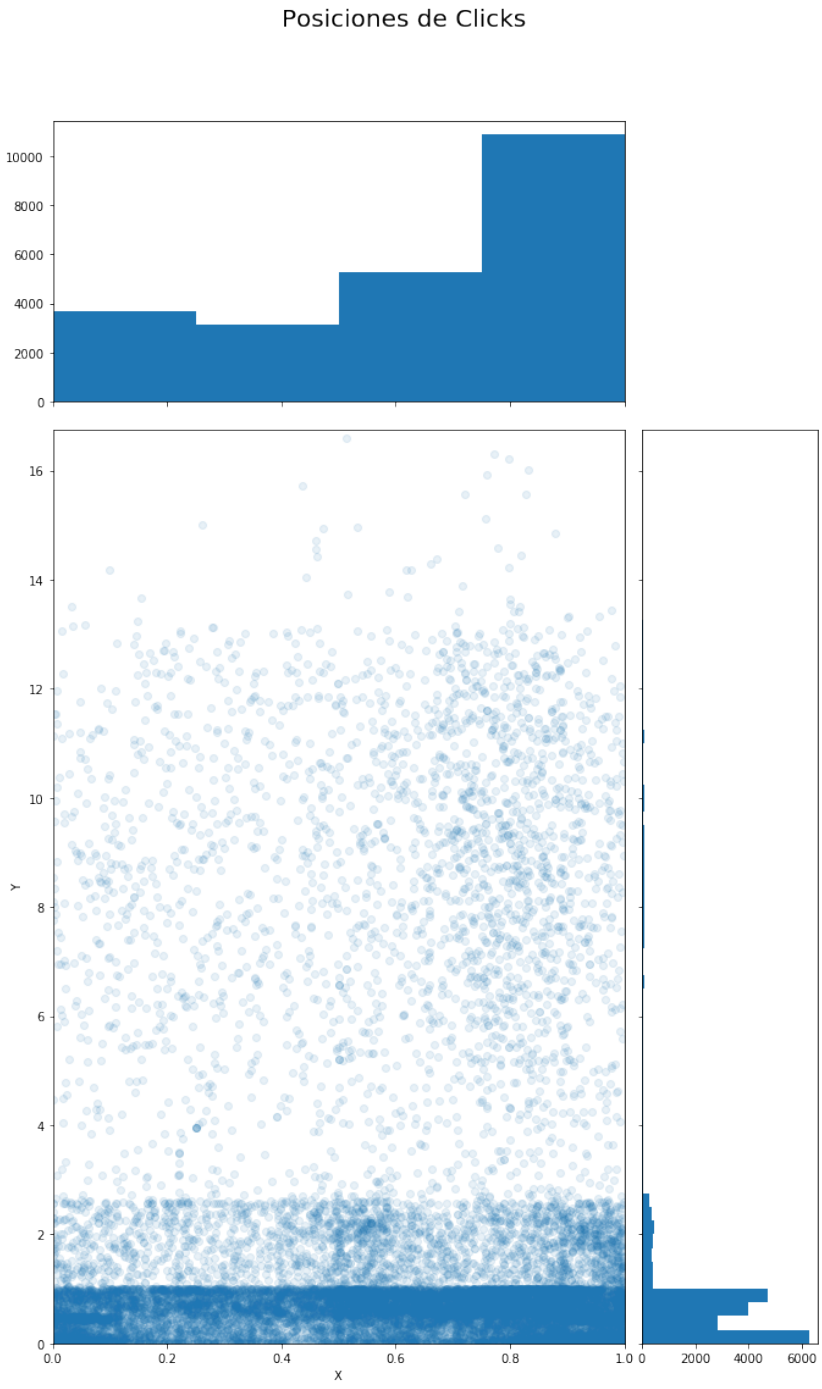

Figura 25: Posiciones de clicks

3.2. Posiciones geográficas de los clicks

La distribución de las posiciones geográficas aproximadas de los clicks se condice con las ubicaciones mas pobladas de Uruguay. Teniendo la mayoría de los clicks concentrados en su capital.

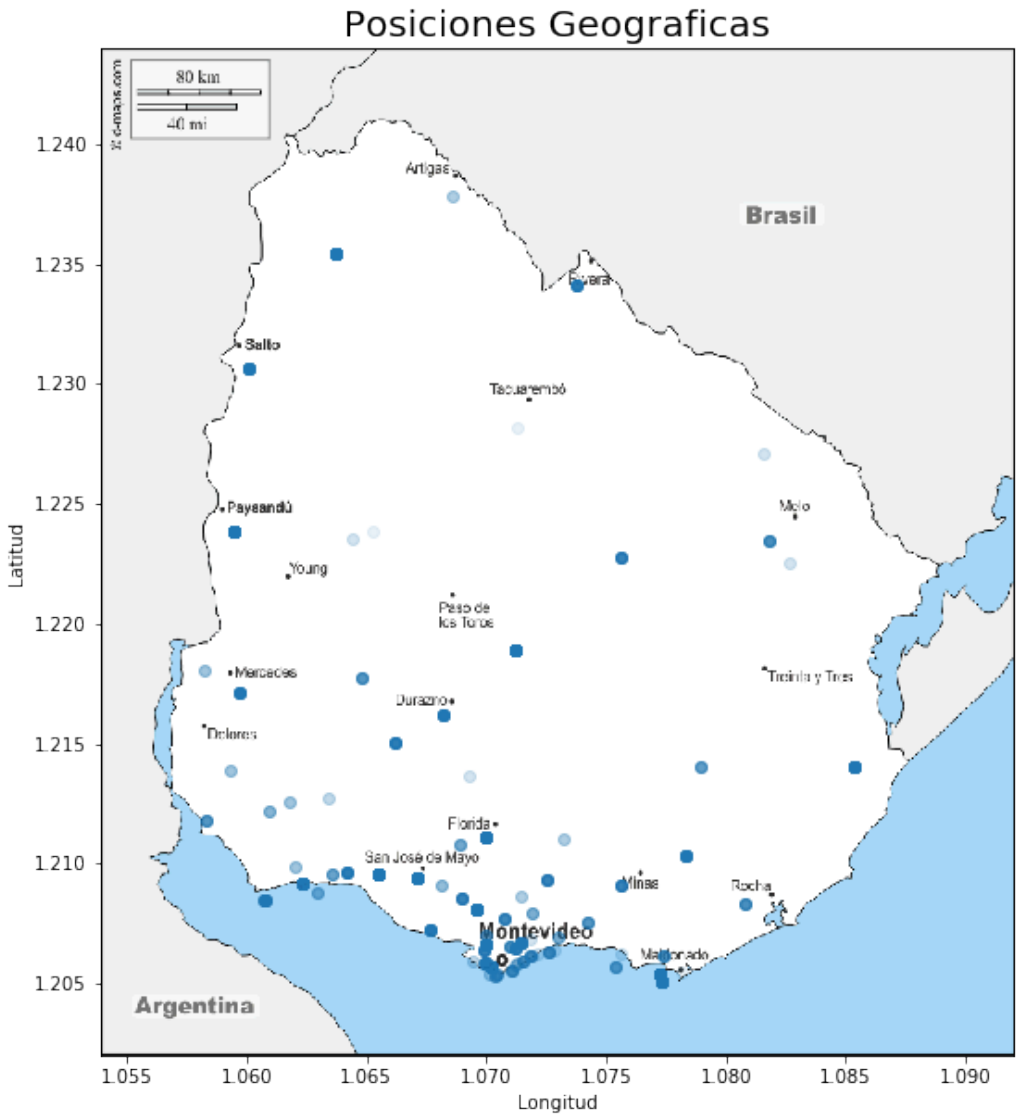

Figura 26: Posiciones de geográficas

3.3. Top carriers

Se obtuvieron una cantidad mayor a la esperada en cuanto al numero de proveedores de red celular, dado que en Uruguay solamente hay 3, la hipótesis es que, sumado al hecho de que hay una alta cantidad de lenguajes en los dispositivos, la procedencia puede estar relacionada al turismo del país.

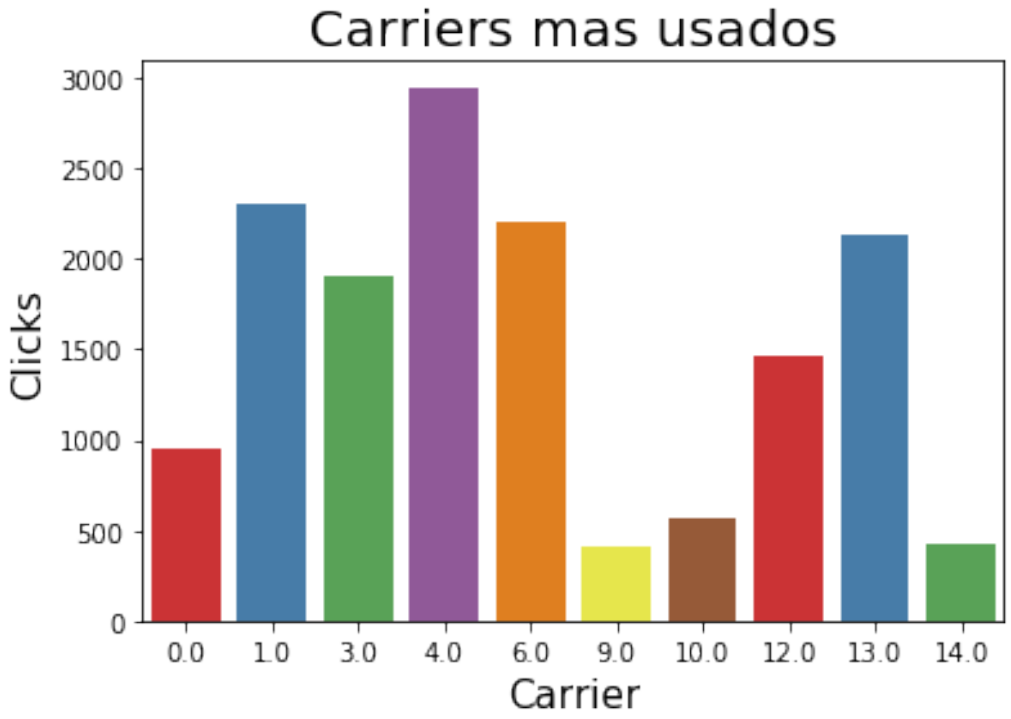

Figura 27: Top 10 carriers

3.4. Top Anunciantes

El siguiente gráfico muestra a que anunciantes pertenecen los clicks realizados del set de datos. Casi en su totalidad pertenecen a un solo anunciante (3), y solamente 88 clicks para el resto de los anunciantes. Esto puede ser tanto por que se tomaron muestras muy chicas o por que el interés esta puesto en un anunciante en particular.

Figura 28: Anunciantes mas clickeados

3.5. Top Fuentes

En el próximo gráfico se muestra como hay una predominancia de una de las fuentes en las que se iniciaron los clicks por sobre las demás.

Figura 29: Fuentes mas clickeadas

4. Auctions

4.1. Dispositivos que más aparecen en subastas

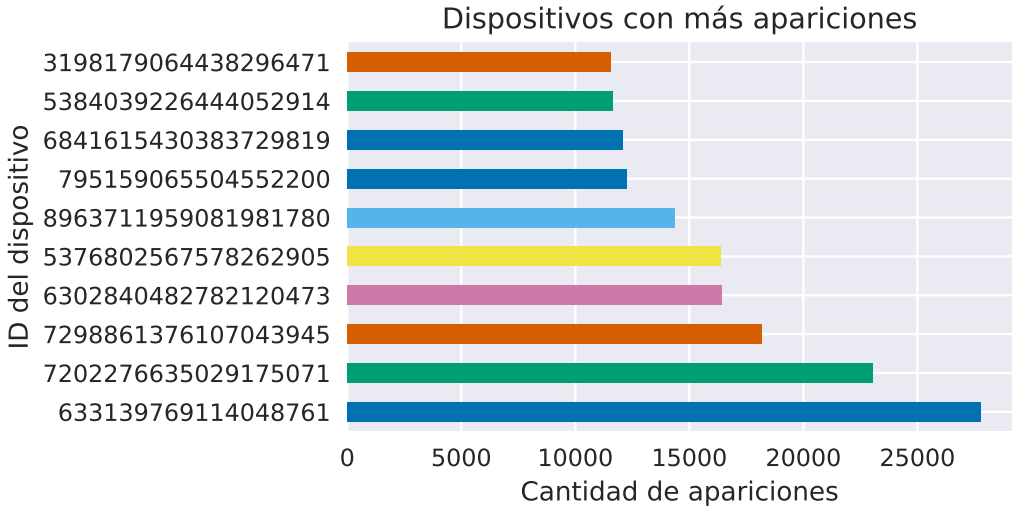

Figura 30: Dispositivos que más aparecieron en las subastas.

En la figura 30 puede observarse que los dispositivos que más aparecen tienen una cantidad considerablemente grande de apariciones en subastas. Por ejemplo, para el dispositivo top se tienen más de 25.000 apariciones, lo que implica una media de más de 2.700 apariciones por día. Por otro lado, puede verse que todos los dispositivos poseen una cantidad de algún modo "exagerada"de apariciones: para el dispositivo que menos aparece entre los diez primeros, se tienen más de 1.000 apariciones por día. Lo anterior podría justificarse de varias formas, pues podría tratarse de dispositivos que no comparten su información y se interpretan como si fueran el mismo. Otra causa posible sería la de fraudes, aunque confirmar esta suposición requeriría mucho cuidado y análisis.

4.2. Cantidad de subastas según el día de la semana

Como se contaba con datos de días consecutivos, se consideró interesante analizar cómo variaba la cantidad de subastas para cada día de la semana. Primero se analizó la cantidad de subastas para la semana desde el 6/3 hasta el 12/3 inclusive. Sin embargo, se observó que los primeros dos días para los que se tienen datos (Martes y Miércoles) tuvieron una cantidad de subastas

mucho menor (como se observa en las figuras 32 y 33). Por lo tanto, se optó por promediar la cantidad de subastas para esos días. Los resultados de dicho análisis se muestran en la figura 31, y muestran una distribución prácticamente uniforme para la cantidad de subastas por día.

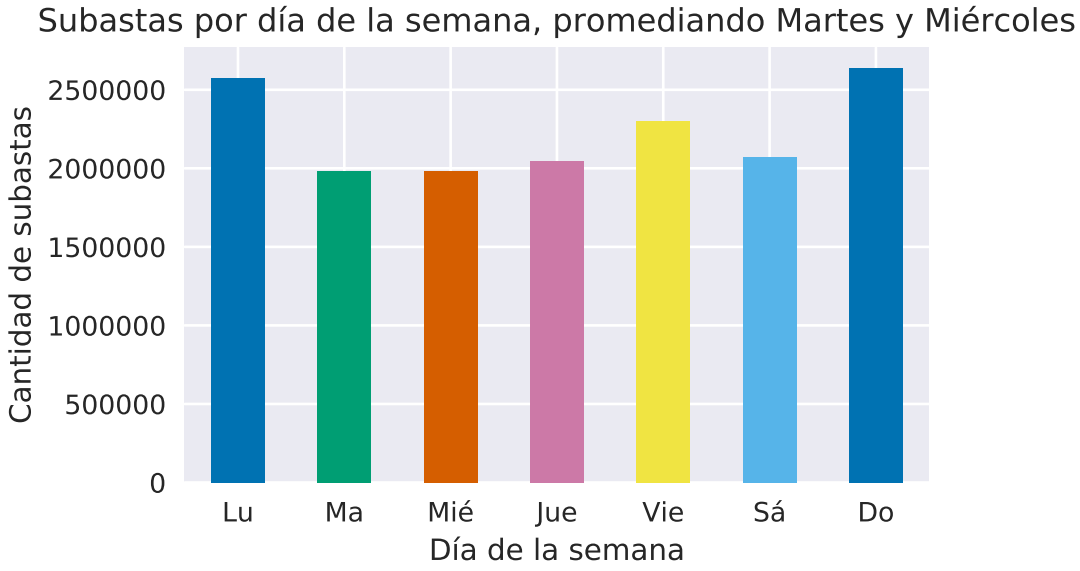

Figura 31: Comparación de la cantidad de subastas para cada día de la semana.

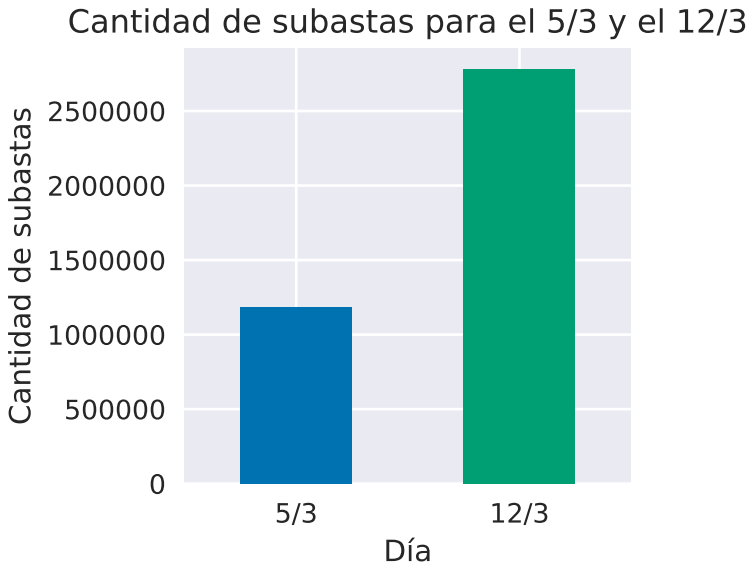

Figura 32: Comparación de la cantidad de subastas para los días 5/3 y 12/3.

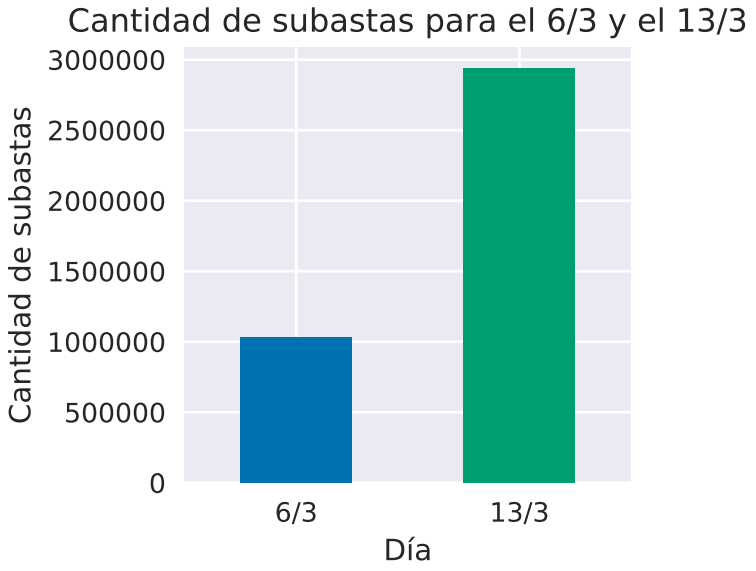

Figura 33: Comparación de la cantidad de subastas para los días 6/3 y 13/3.

4.3. Frecuencia de apariciones por dispositivo

Teniendo en cuenta lo observado al analizar los diez dispositivos que más aparecieron en subastas, se decidió investigar si la tendencia observada se mantenía para todos los dispositivos. En la figura 34 puede observarse que muy pocos dispositivos tienen una cantidad considerablemente grande da apariciones. Esto podría reforzar la idea de que se trata de valores anómalos en el set de datos.

En la figura 35 se muestra que si no se consideran los dispositivos que aparecen una gran cantidad de veces en las subastas, la distribución parece seguir una tendencia exponencial. Nuevamente se observa que muy pocos dispositivos tienen una gran cantidad de apariciones, algo que parece lógico a simple vista.

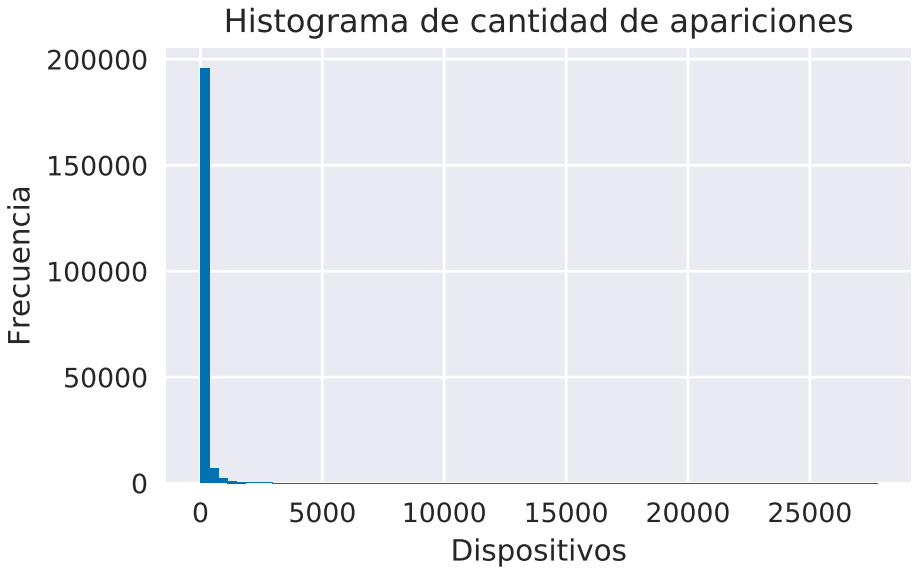

Figura 34: Frecuencia de apariciones.

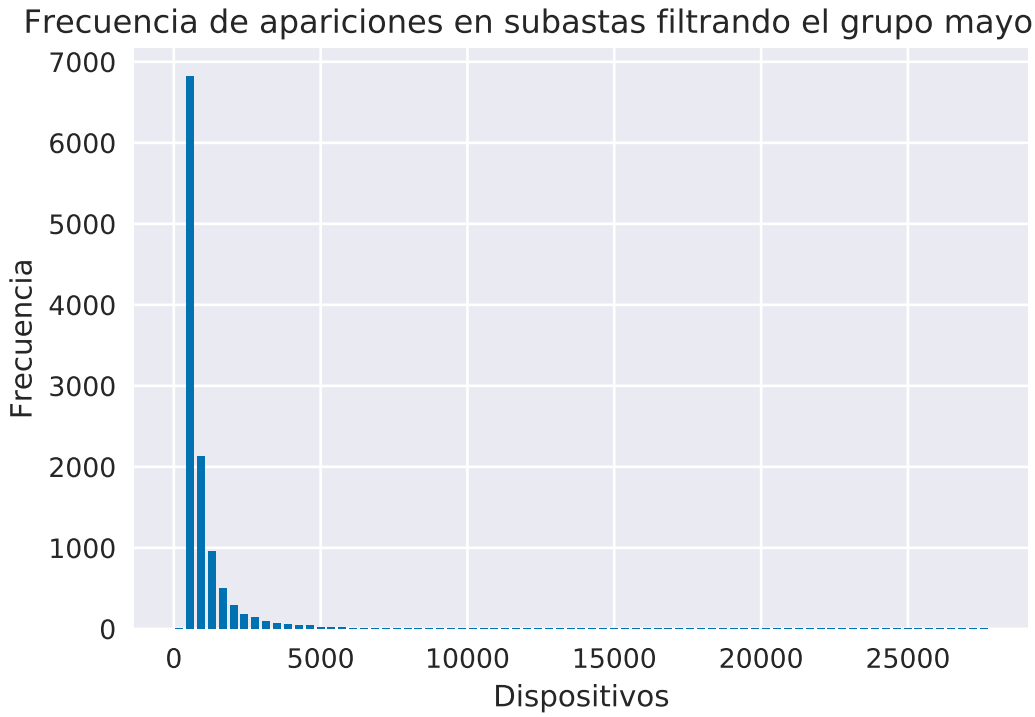

Figura 35: Frecuencia de apariciones, excluyendo al grupo de dispositivos con más apariciones.

4.4. Cantidad de subastas según el día del mes

Aquí se muestran los resultados de analizar la cantidad de subastas por día según su orden cronológico. La figura 36 exhibe una tendencia más o menos lineal de crecimiento en la cantidad de subastas, aunque podría considerarse un quiebre a partir del día Jueves 7 de Marzo pues los días 5/3 y 6/3 difieren poco en la cantidad de subastas, comparados a los demás (incluso la cantidad decrece).

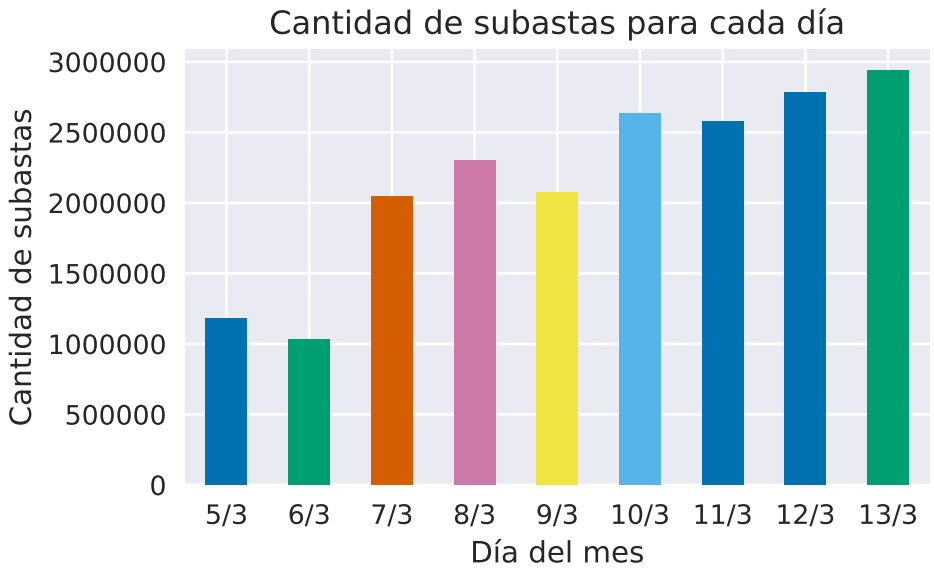

Figura 36: Cantidad de subastas para cada día del dataset.

4.5. Cantidad de subastas según el horario

Al analizar la cantidad de subastas por horario, se encuentra que hay un mínimo en la cantidad de subastas entre las 7 y 8 de la mañana. A partir de allí, se observa un crecimiento aparentemente lineal hasta las 3 de la mañana donde hay un máximo. Por otro lado, desde las 3 de la mañana hasta las 7 de la mañana, la cantidad de subastas decrece pronunciadamente, pues lo más natural es que las personas estén durmiendo en dichos horarios. Los resultados mencionados pueden verse en la figura 37.

Si se realiza un análisis por plataforma, resulta evidente que la distribución de la cantidad de subastas es independiente de la misma. Esto se observa en la figura 38

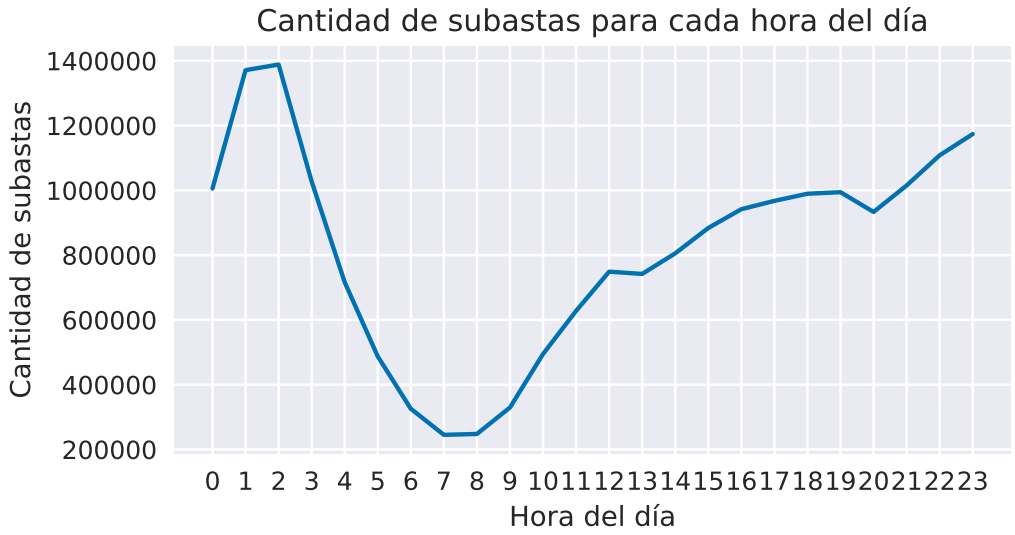

Figura 37: Cantidad de subastas para cada hora.

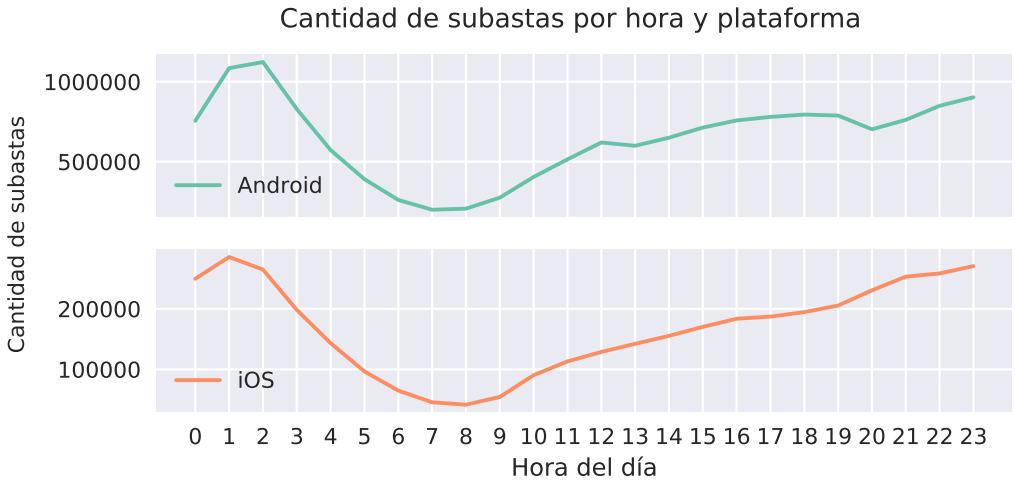

Figura 38: Cantidad de subastas para cada hora.

4.6. Cantidad de subastas para días de semana y fin de semana

Para los datos con los que se cuenta, puede asegurarse que no hay más subastas durante el fin de semana que en los días de semana, incluso si se considera al Viernes (día laboral) como parte del fin de semana. Se registraron

aproximadamente 12 millones de subastas de Lunes a Jueves, mientras que hay aproximadamente 7 millones de subastas de Viernes a Domingo.

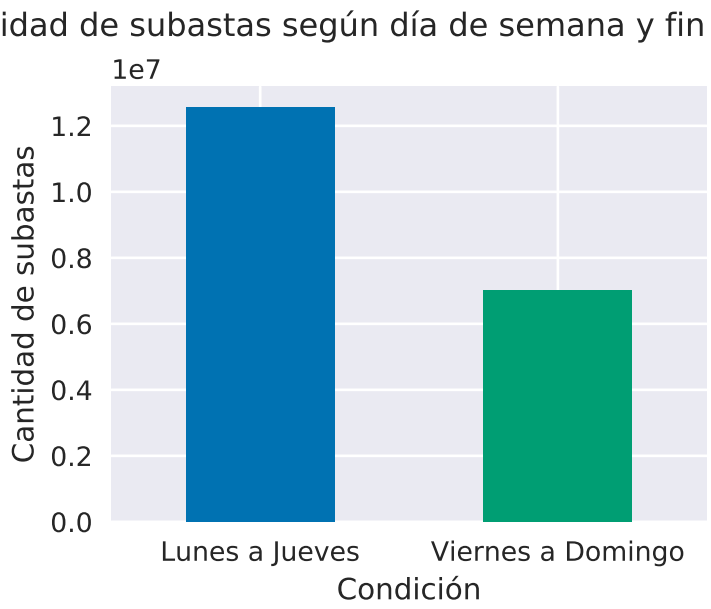

Figura 39: Cantidad de subastas según se trata de un día de semana o fin de semana.

4.7. Cantidad de dispositivos por plataforma

Como se tienen datos de la plataforma del dispositivo para todas las subastas, se decidió analizar la cantidad de dispositivos por plataforma. En la figura 40 se muestra la gran predominancia de Android sobre iOS, con más del triple de dispositivos.

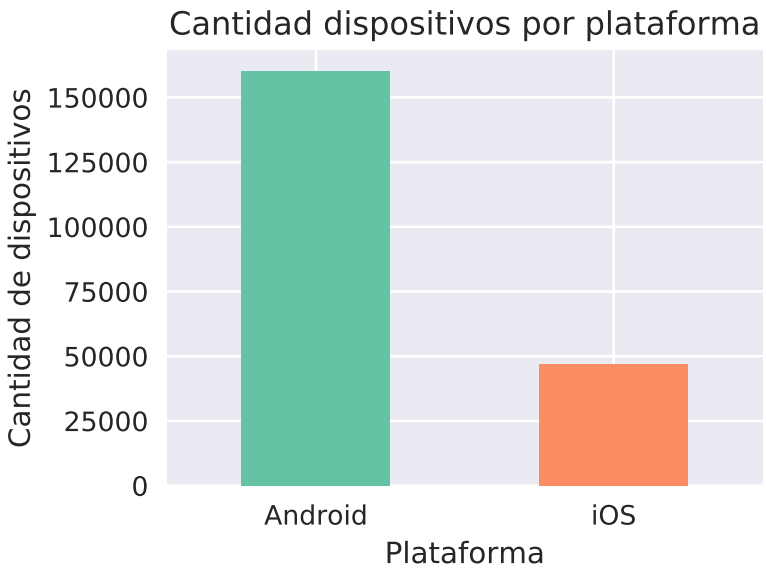

Figura 40: Cantidad de dispositivos para cada plataforma.

4.8. Proporción en la cantidad de subastas por plataforma

Para el total de subastas, se analizó la proporción según la plataforma. Nuevamente aparece Android como plataforma dominante sobre iOS, aunque la diferencia se hace notoria a partir del día Viernes 8 de Marzo. Para los dispositivos que cuentan con iOS, la participación crece hasta el día 10/3, donde parece estancarse.

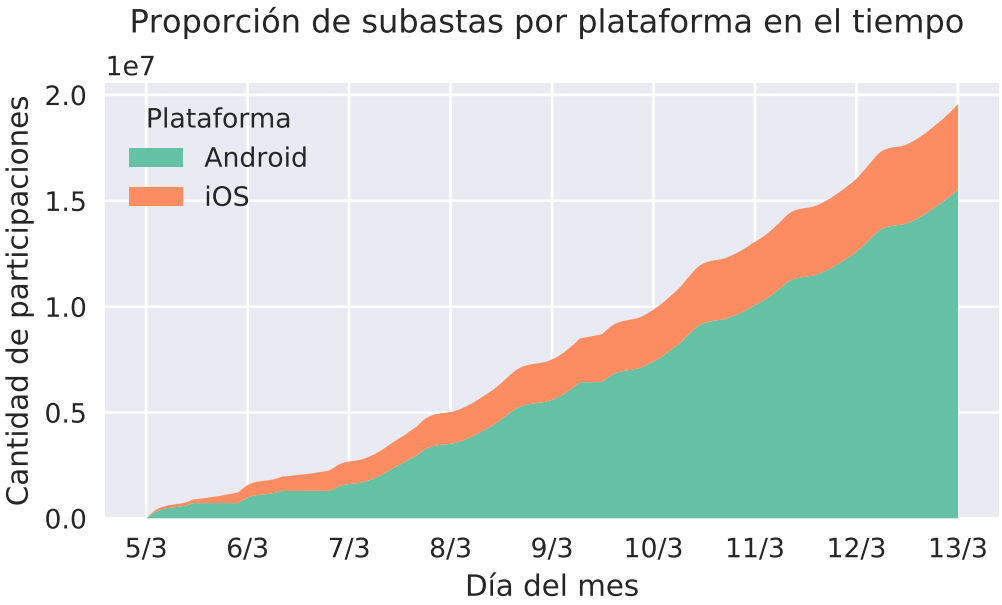

Figura 41: Evolución en la cantidad total de subastas por plataforma.

Parte IV
Conclusiones

1. Auctions

A modo de conclusión, se destaca que la cantidad de subastas a lo largo de las horas del día no depende de la plataforma. Esto resulta interesante siendo que se tienen muchísimos más dispositivos Android que iOS. Por otro lado, resulta también interesante observar que existe aproximadamente el doble de cantidad de subastas en los días de semana en comparación al fin de semana, donde podría esperarse una mayor cantidad debido a que en general hay más tiempo libre. Además, el hecho de tener una tendencia lineal creciente en la cantidad de subastas para los días desde el 5/3 al 13/3 debería analizarse en el contexto del país (Uruguay) en el cual fueron originadas. Si bien nuestro país Vecino también celebra carnaval los días 4 y 5 de Marzo, no se observa que haya una mayor cantidad de subastas en esos días, por lo que podría atribuirse el crecimiento a un aumento de la participación de Jampp.

2. Events

Cómo conclusión de los datos obtenidos de este dataset nos parece interesante la baja tasa de eventos dentro de las aplicaciones que realmente son atribuidos al trabajo de marketing de *Jammp*, este porcentaje fue del 0.2 % del total.

3. Clicks

Algo interesante de este dataset fue haber podido encontrar la procedencia geográfica aproximada de los clicks y poder situarlos en el mapa, verificando que las zonas en las que se realizaban los clicks eran también a su vez los lugares mas poblados o de mayor actividad del país. De la misma manera poder ubicar en la pantalla la posición de los clicks ayudo a entender de que manera se están mostrando los anuncios a los usuarios. También se pudieron extraer otros resultados que concordaban con el resto de los datasets en el informe, lo que ayudo a ver la consistencia de la información.

4. Installs

Si bien los datos hubiesen sido mucho más interesantes de analizar si existiesen más instalaciones atribuidas, se logró determinar con claridad los periodos de tiempo en los cuales se concretaron las instalaciones así como el estado del celular (Modelo, OS, WIFI). También se pudo determinar que evento principalmente llevó a esta instalación.