Trabajo Práctico 1

[7506/9558] Organización de Datos
Primer cuatrimestre de 2019

Grupo 27

Alumno	Número de Padrón	Email
Minino, Nahuel	99599	nahuel.minino@gmail.com
Alvarez, Gonzalo A.	98359	gonza2703@gmail.com
Aparicio, Axel	96283	axelaparicior@gmail.com
Peña, Alejandro	98529	alee.pena.94@gmail.com

Repositorio GitHub: https://github.com/alepenaa94/Datos_2019

Índice

1. Glosario

- *Convertir*: el objetivo de mostrar publicidad es que un dispositivo instale una aplicación, a ese evento se le llama **conversión**.
- *Retargeting*: es una estrategia de marketing que vuelve a atraer a los usuarios, que han mostrado interés en sus productos o servicios, mediante la publicación de anuncios personalizados en sus dispositivos móviles.
- *Instalación implícita*: Instalaciones atribuidas a jampp pero que no fueron contadas como conversión mediante el flujo habitual (Ej: Usuario instala la aplicación publicitada buscándola en el application store).
- *Dispositivo/Device*: entidad con un id de publicidad asociado. Por ejemplo: un celular Samsung J6 con Android tiene un id único, un Apple iPhone tiene un identificador único.
- *Evento*: cualquier tipo de interacción categorizada dentro de una aplicación.
- *Subasta/Auction*: en el momento que una aplicación quiere mostrar una publicidad, ese espacio se vende en una subasta (generalmente de segundo precio) donde todos los interesados en mostrar una publicidad ofertan un precio y gana quién más ofrece.
- *Agente de usuario/User Agent*: un agente de usuario es un software que actúa en nombre de un usuario. Un uso común del término se refiere a un navegador web que recupera, procesa y facilita la interacción del usuario final con el contenido web".

2. Introducción

En el siguiente informe se detalla un estudio de datos que la empresa Jampp pudo proporcionar, empresa encargada de la promoción de aplicaciones por medio de publicaciones para que instalen la misma.

El objetivo es realizar un análisis de los datos que fueron proporcionados buscando principalmente información útil para poder predecir posibles usuarios que instalen la aplicación, evitando "sobrecargar"a los usuarios con publicidades que no tengan una conversión.

El análisis presentado se hizo utilizando Python3 con las siguientes librerías:

- Pandas
- Numpy
- Matplotlib
- Seaborn

A lo largo del informe se pueden ver distintas visualizaciones de los datos, las cuales fueron hechas con la intención de facilitar la comprensión de la información disponible.

Se utilizó un repositorio en GitHub: https://github.com/alepenaa94/TP1-Datos, como herramienta de integración, donde se pueden encontrar los notebooks utilizados tanto para la limpieza de los datos como para el análisis exploratorio.

La manera en la que realizamos el trabajo fue primero analizando cada dataset por separado, viendo aspectos básicos de los mismos, para luego determinar que preguntas interesantes podíamos formular a partir de esos datos.

3. Consultas a Jampp

Tuvimos un espacio de consultas con el representante de Jampp y pudimos aclarar muchas dudas respecto de los datasets. A continuación listaremos cada una de ellas ya que creemos que ayudarán a entender algunos fenómenos que se presentaron en los análisis realizados.

3.1. El dataset de Auctions es un muestreo

El dataset de Auctions representa un 25 % de las subastas reales que suceden en tiempo real.

3.2. Los datos son de dispositivos en Uruguay

Todos los datos corresponden a los de un único país, el cual es Uruguay.

3.3. Horarios en timezone UTC

Todos los campos correspondientes a fechas con horarios están representados en la zona horaria UTC-0. A lo largo del trabajo hablaremos de horarios en esta zona horaria.

3.4. Eventos de Aplicaciones de Jampp

Los eventos recaudados en el dataset de events, sólo pertenecen a aplicaciones clientes de Jampp ya que solo pueden recaudar eventos provistos por los mismos clientes.

3.5. Los horarios no son exactos

Los horarios en los datasets no representan exactamente el instante en el que sucedió el acontecimiento registrado (Ej.: Horario de click, Creación del evento, Horario de Instalación).

4. Análisis de los datos

4.1. Primera aproximación a los datos

La información proporcionada está compuesta por los archivos:

- Auctions.csv
- Clicks.csv
- Installs.csv
- Events.csv

En el dataset de *Auctions* se puede encontrar los dispositivos en los que se realizan subastas. Luego en *Clicks* se tienen los registros de los clicks que los dispositivos realizaron sobre las publicaciones, coordenadas de los mismos, tipo de conexión, entre otros. *Installs* posee registros de instalaciones realizadas por los dispositivos, si son atribuidas a la empresa, etc. Por último, *Events* contiene registros de los eventos que ocurren dentro de las aplicaciones clientes de Jampp; se tiene información sobre el id de la aplicación, el tipo de evento, características del dispositivo, etc.

Previo a realizar algún análisis, se comenzó por limpiar los datos recibidos. Se determinó que tipo de datos debería contener cada columna, si se los puede categorizar o no, se descartaron columnas de datos triviales o que no aporten al análisis, se redefinieron ciertas columnas para facilitar el entendimiento de su contenido. Esto se encuentra en los notebooks de "Limpieza_ %NOMBRE%"

A continuación se mostrará como el resultado final los datos y sus tipos inferidos.

Clicks Types	
atributo	**tipo de dato**
advertiser_id	int8
source_id	int8
created	datetime64[ns]
country_code	category
latitude	float64
longitude	float64
wifi_connection	bool
carrier_id	int16
trans_id	object
os_minor	float32
agent_device	float16
os_major	float32
specs_brand	category
brand	int8
timeToClick	float64
touchX	float32
touchY	float32
ref_type	category
ref_hash	int64

Auctions Types	
atributo	**tipo de dato**
country	category
date	datetime64[ns]
device_id	int64
platform	category
ref_type_id	int8
source_id	int8

Events Types	
atributo	**tipo de dato**
date	datetime64[ns]
event_id	int8
ref_type	category
ref_hash	int64
application_id	int16
attributed	bool
device_countrycode	category
device_os_version	float64
device_brand	float64
device_model	float64
device_city	float64
session_user_agent	float64
user_agent	float64
event_uuid	object
carrier	float64
kind	float64
device_os	float64
connection_type	category
ip_address	int64
device_language	float64
Wifi_cat	category

Installs Types	
atributo	**tipo de dato**
created	datetime64[ns]
application_id	int8
ref_type	category
ref_hash	int64
attributed	bool
implicit	bool
device_countrycode	category
device_brand	float64
device_model	float64
session_user_agent	object
user_agent	object
event_uuid	object
kind	object
ip_address	int64
device_language	float64
Wifi_cat	category

4.2. General

4.2.1. Cantidad de registros en cada dataset dependiendo de la hora

Como un primer análisis general de los datos, buscamos ver: Qué sucede según el horario para cada set de datos?

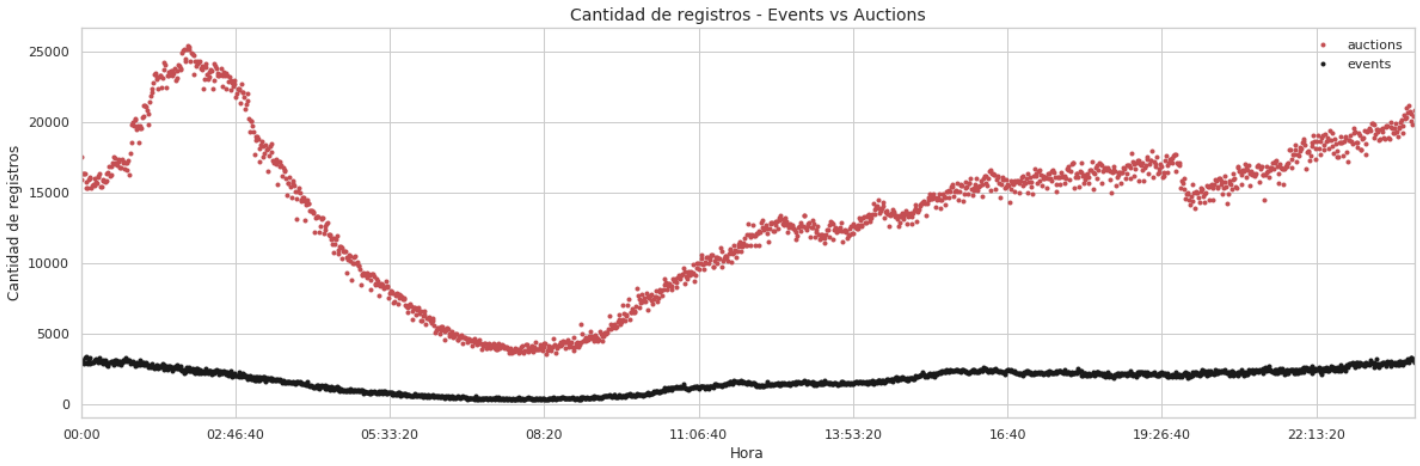

Figura 1: Cantidad de eventos vs subastas dependiendo la hora

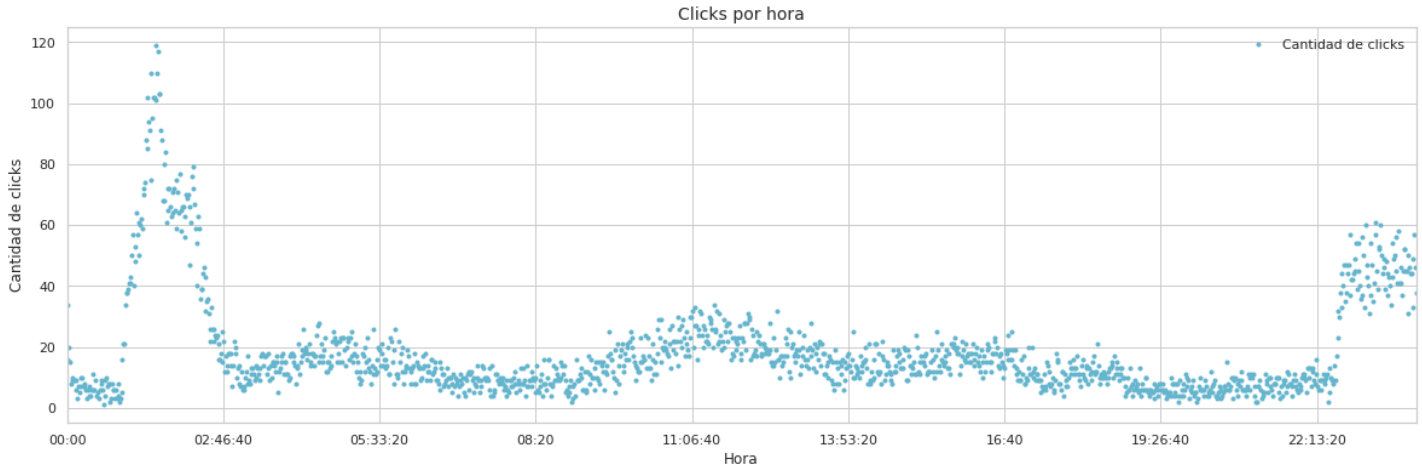

Figura 2: Cantidad de clicks por hora

Podemos observar que la cantidad de clicks aumenta durante la noche alcanzando el pico entre la 1 y 2 de mañana con alrededor de 100 clicks registrados, mientras que durante el resto del día no llega a alcanzar los 40.

Esto se puede deber a que durante la noche la gente está más libre y tiende a usar más las aplicaciones del celular que durante el día, resultando en que salgan más publicidades en las cuales clickean.

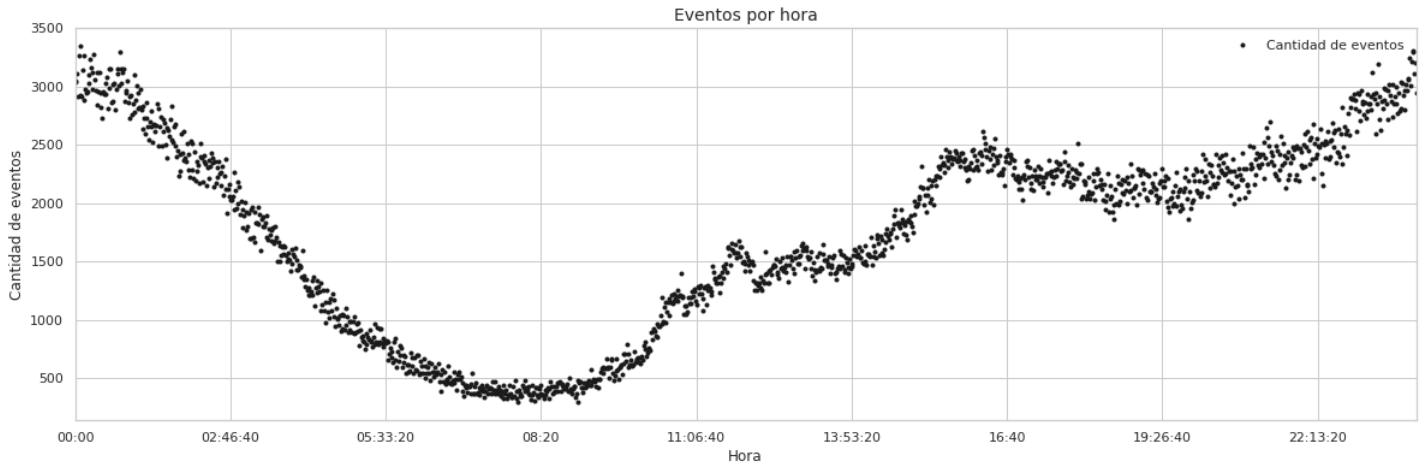

Figura 3: Cantidad de eventos por hora

Durante la medianoche es cuando más eventos se registran, afirmando más aún lo que vimos en el primer gráfico. Sin embargo, como los eventos se relacionan más con el horario de actividad habitual de una persona, no encontramos un pico tan notable como en los clicks del usuario. Además se tiene el comportamiento esperado ya que, como se mencionó previamente, todos los eventos son de una misma zona horaria; lo cual se refleja viendo lo pareja que es la curva y que en la madrugada la actividad va en caída;.

Figura 4: Cantidad de instalaciones por hora

Las instalaciones, tienen una regularidad similar a la de los eventos, están fuertemente relacionados. Aunque hay que tener en cuenta que tenemos mucho menos volumen de datos en las instalaciones y tal vez no sea lo mas correcto decir que se asemejan.

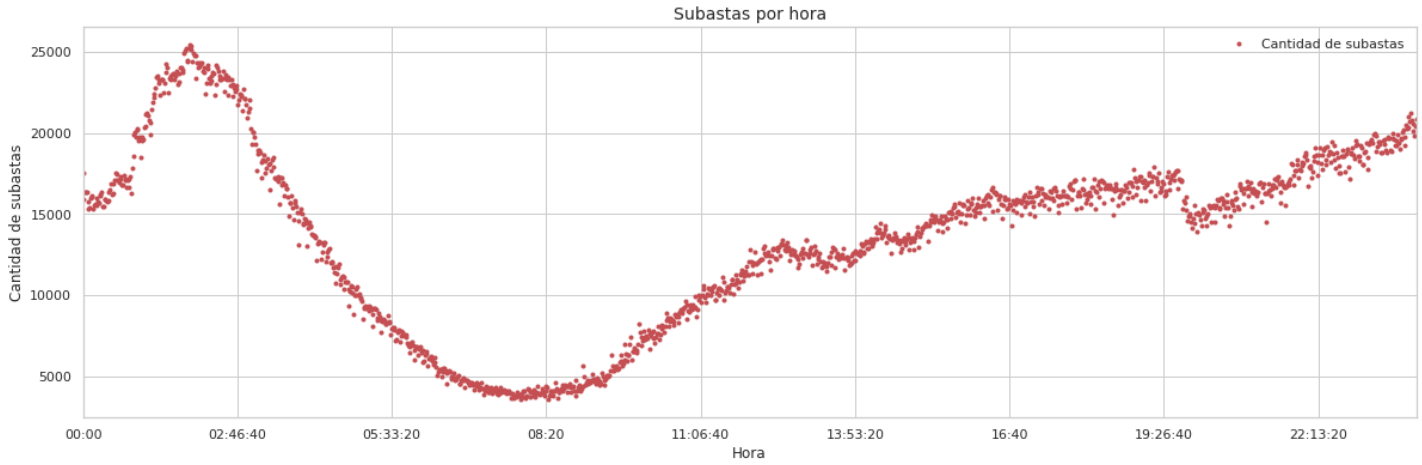

Figura 5: Cantidad de subastas por hora

La cantidad de subastas se relaciona bastante en el mismo rango horario superando las 25000 subastas. Lo cual tiene sentido porque deberíamos apostar más en el momento en que el usuario es más activo.

4.2.2. Análisis por dispositivos

En esta sección vamos a realizar un análisis particular y general de los dispositivos y su ”actividad” registrada por Jampp.

Comenzamos analizando los clicks realizados por dispositivo.

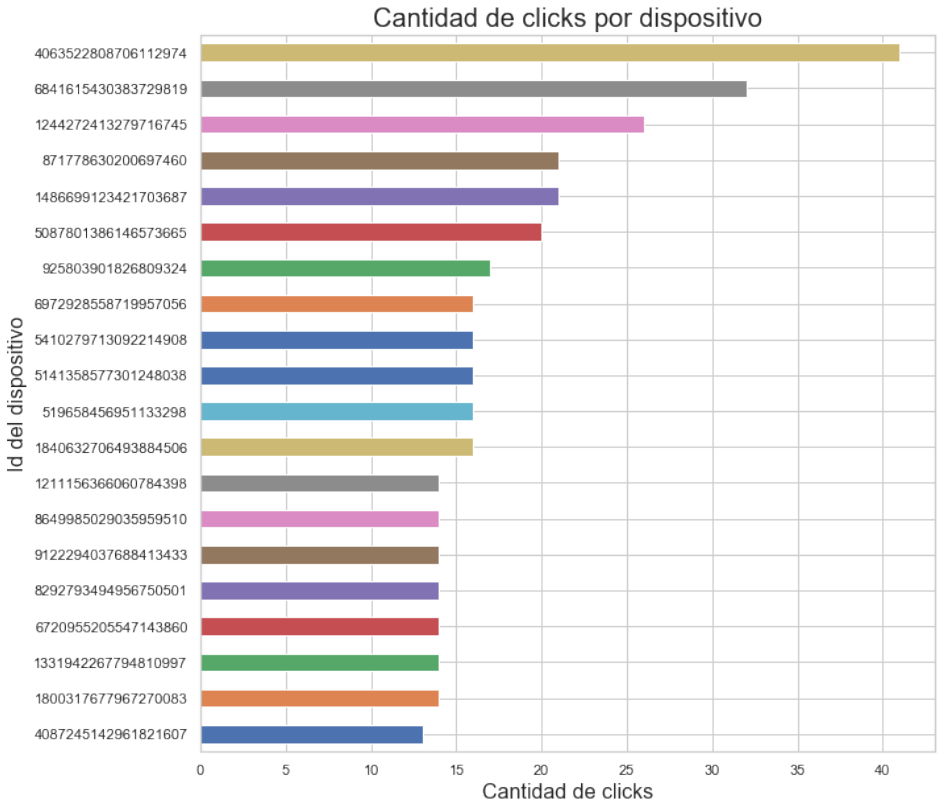

Figura 6: Cantidad de clicks por dispositivo

Podemos ver qué dispositivos son los que realizan más clicks en las publicaciones y que en promedio los dispositivos rondan los **15** clicks independientes de la publicación.

Por otro lado en las dos siguientes figuras vemos en cuantas subastas ”participa el dispositivo” y la cantidad de eventos que realiza.

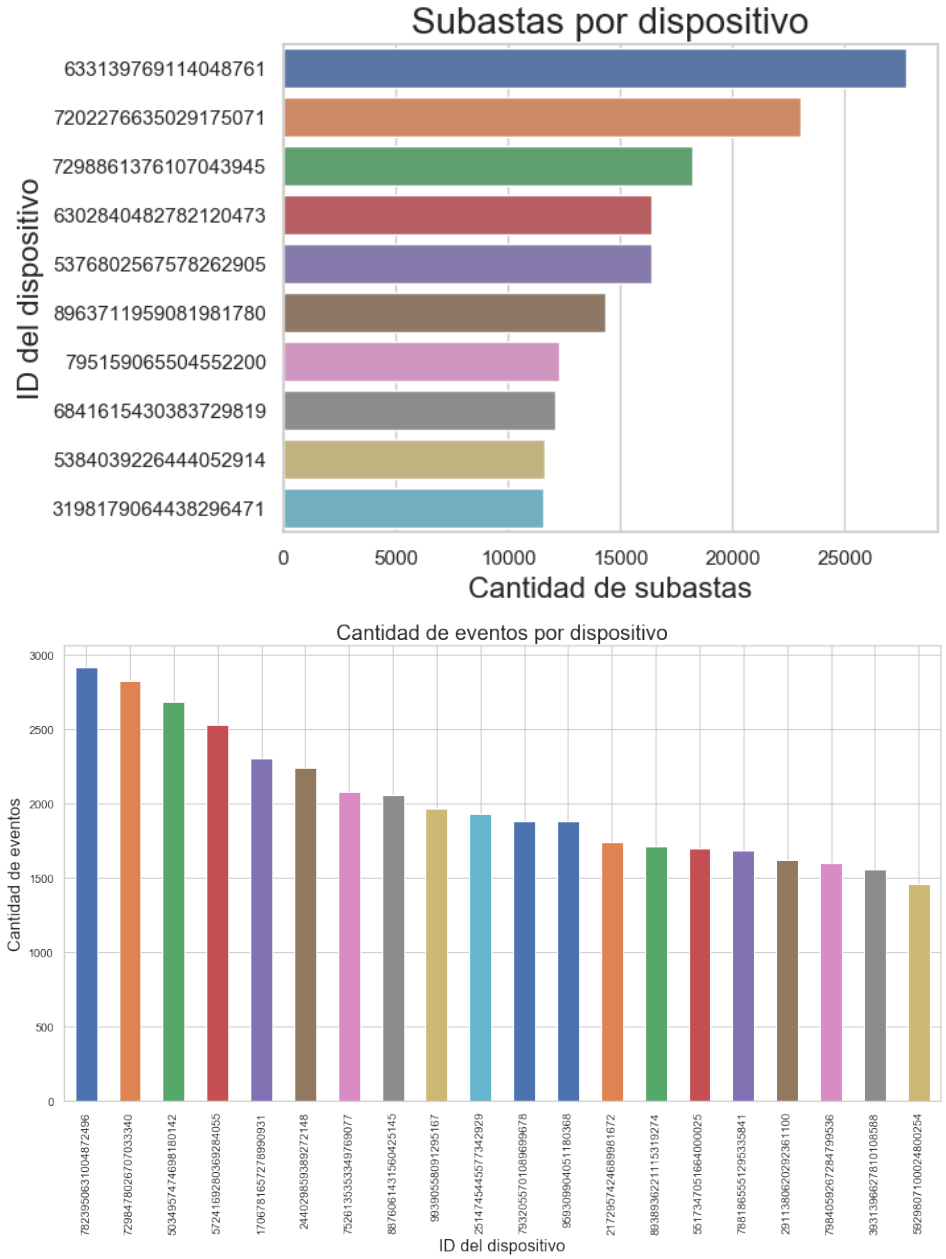

Figura 7: Cantidad de subastas y eventos por dispositivo

Y por último vemos las instalaciones por dispositivo.

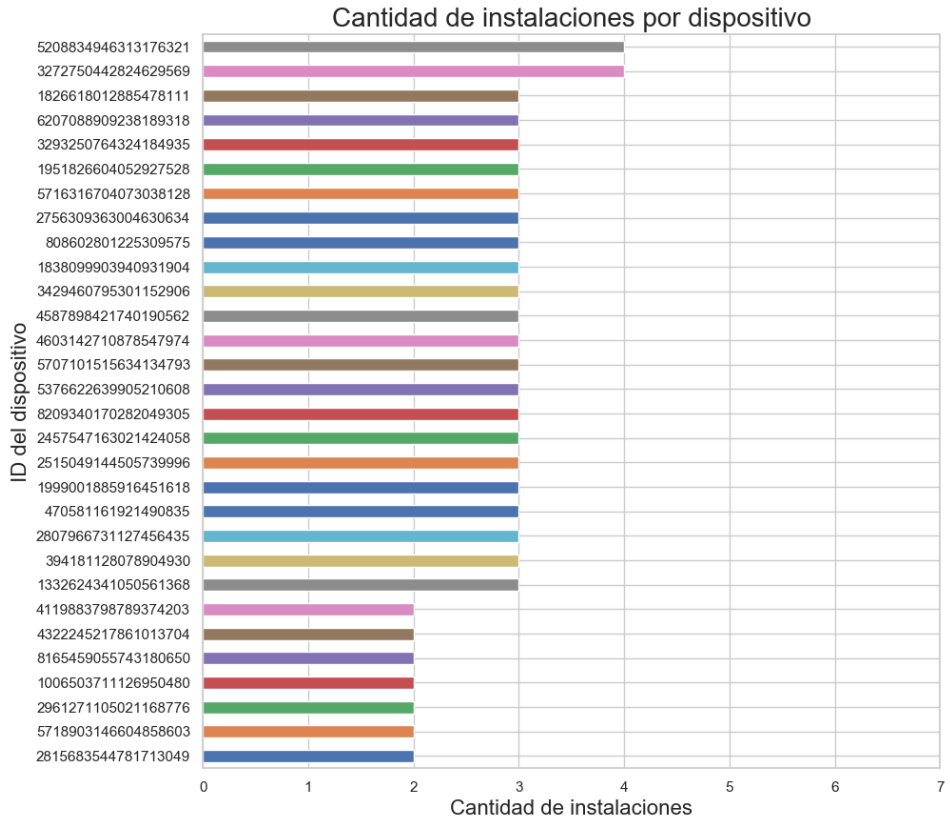

Figura 8: Cantidad de installs por dispositivo

Algo importante a observar es que pocos dispositivos instalan más de dos aplicaciones y que son pocos(371 de 3412) los que instalan más de una vez la misma aplicación.

De estos análisis de forma individual por dispositivos, si los juntamos en un único gráfico para realizar una mejor comparación y observación tenemos la siguiente visualización.

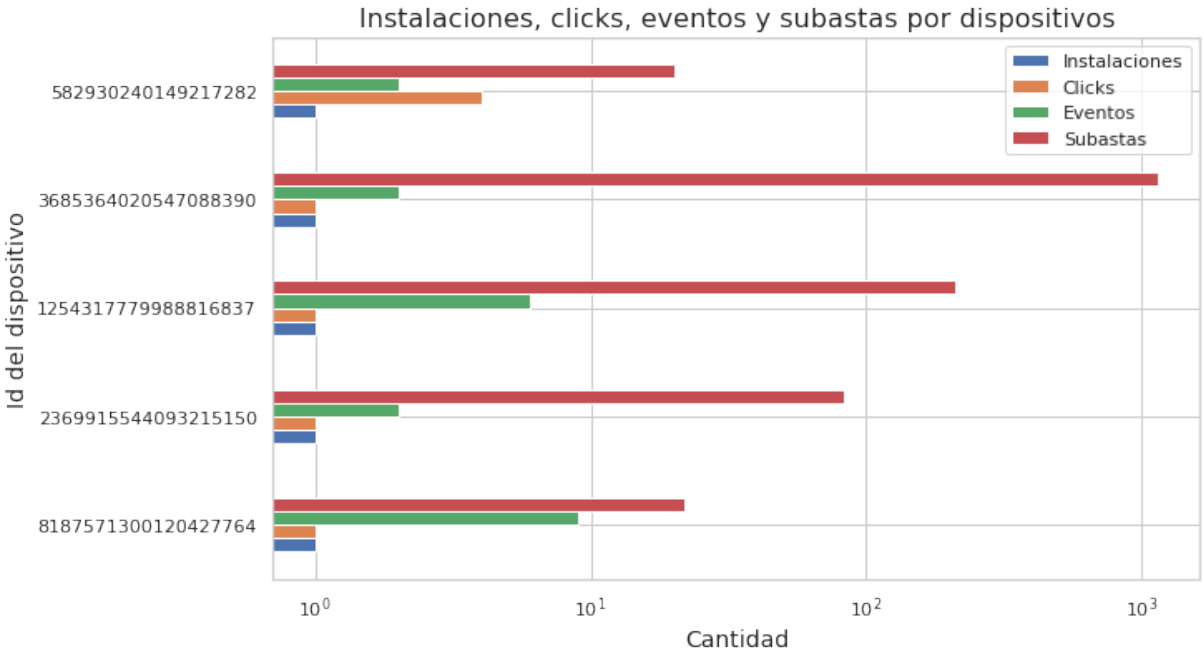

Figura 9: Comparación eventos, clicks, subastas e instalaciones por dispositivo

De los set de datos dados no se tiene suficiente información como para poder vincular todos los set de datos para varios dispositivos, únicamente se pudo con los que se muestran en la figura anterior resultando muy difícil explicar el

comportamiento y decir si es o no esperado.

4.3. Clicks

4.3.1. Clicks en pantalla y su ubicación

A continuación se visualizan los clicks que se realizaron y donde fueron efectuados según las coordenadas que se poseen.

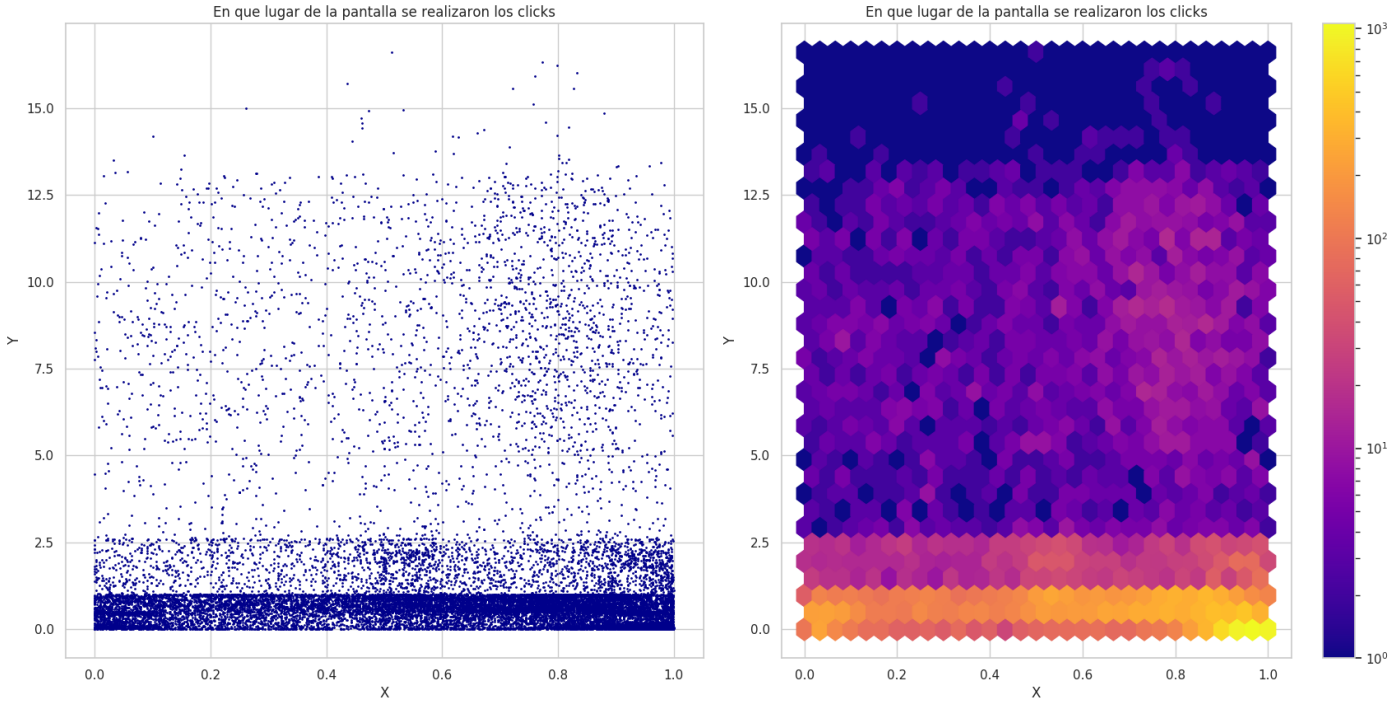

Figura 10: Lugar de la pantalla donde se realizaron clicks

De estos gráficos podemos destacar varias cosas:

- Se puede observar que la mayor concentración de clicks es en la parte inferior de la pantalla. Podemos inferir que las publicidades en su mayoría fueron de banners al pie de la pantalla.
- Los clicks se ven con una predilección mayoritaria hacia la derecha. Un motivo podría ser que los clicks fueron hechos en su mayoría por personas diestras.
- Casi no hay clicks en la parte superior de la pantalla.

Podemos decir que es un gráfico bastante acertado con respecto al uso habitual de un celular. Muestra claramente el comportamiento de un usuario en un smartphone.

4.3.2. Cantidad de clicks vs. tiempo en realizarlo

En los datos de Clicks poseemos la información de cuanto el usuario estuvo viendo la publicación y cuanto tardó en hacer click, ya sea para realizar una conversión o no.

Figura 11: Tiempo en realizar el click en el aviso

En el gráfico se aprecia a simple vista que la mayoría de los clicks fueron realizados con no demasiada espera. Los usuarios no superan el minuto visualizando la publicación y vemos una notable diferencia que interactúa de manera casi instantánea.

Lo que podemos rescatar de esta visualización es que las probabilidades de que un usuario se interese por un anuncio y decida interactuar con el, son mayores al principio y disminuyen conforme pasa el tiempo, por lo que podríamos asumir que una buena publicidad debería ser aquella que logre cautivar a un usuario durante los primeros segundos.

En estos casos tenemos información de cuando se realizo el click, pero no de cuando no están interesados y cuanto tardan en dejar de estarlo. En consecuencia no podríamos medir efectividad de una publicación, sino medir cuanto interactúan con ellas.

4.3.3. Clientes de Jampp con más clicks

Un punto importante a analizar es cuales de los clientes de Jampp tienen más clicks. Más adelante se podrá encontrar una relación entre clicks, subastas e instalaciones, de la cual una posibilidad es suponer que a mayor cantidad de clicks, mayor es la probabilidad de instalaciones.

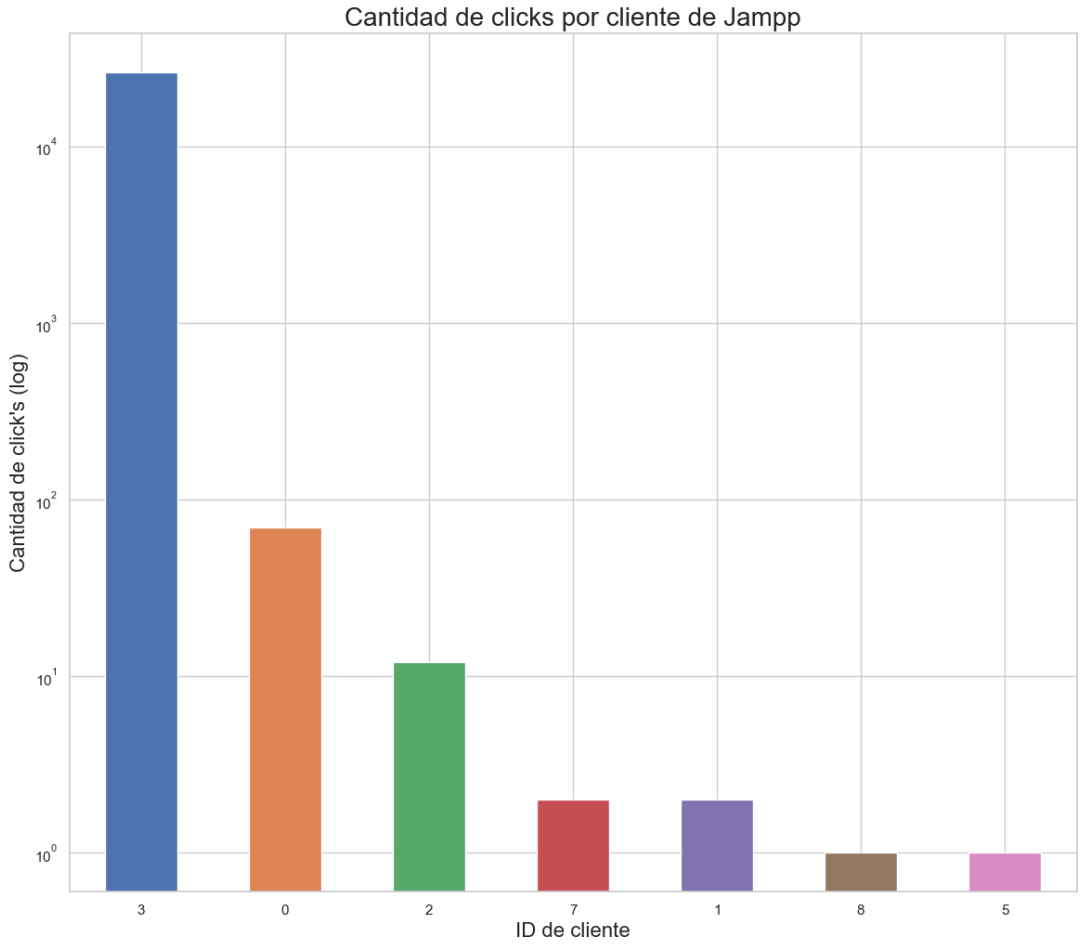

Figura 12: Clicks según cliente

Lo importante a observar es cual cliente produce o puede producir mayor cantidad de conversiones para Jampp, lo que es uno de los objetivos principales de esta empresa.

4.3.4. Cantidad de clicks por aviso/publicación de Apple o Google.

A continuación se puede visualizar los tipos de avisos que presenta cada cliente y cuales obtuvieron más clicks. El nombre del tipo de publicación fue inferido gracias al set de datos de Installs que proporciona dicha información en la columna *session_user_agent*

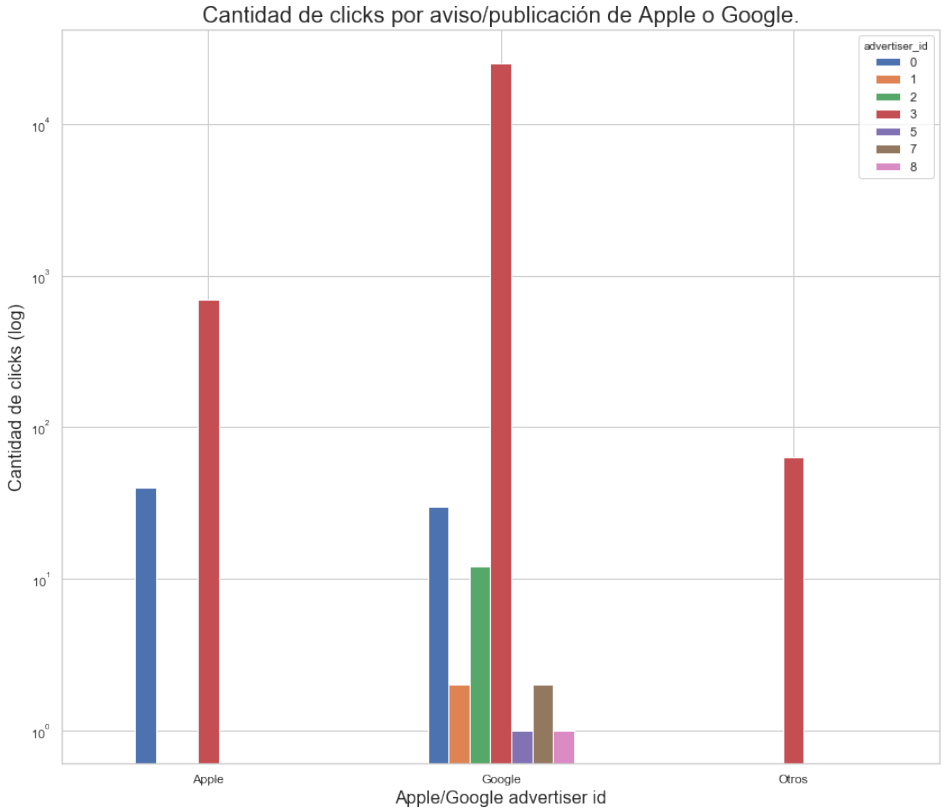

Figura 13: Clicks y tipo de aviso según cliente

Se puede observar que la mayoría de los clientes poseen clicks en avisos de tipo Google, y clicks de publicaciones tipo Apple son los clientes más destacadas en la Figura 12. Esto no significa que todos los clientes no tengan avisos Apple, sino que solo se puede asegurar que no tuvieron clicks en ellos.

4.3.5. Ubicación geográfica de los clicks

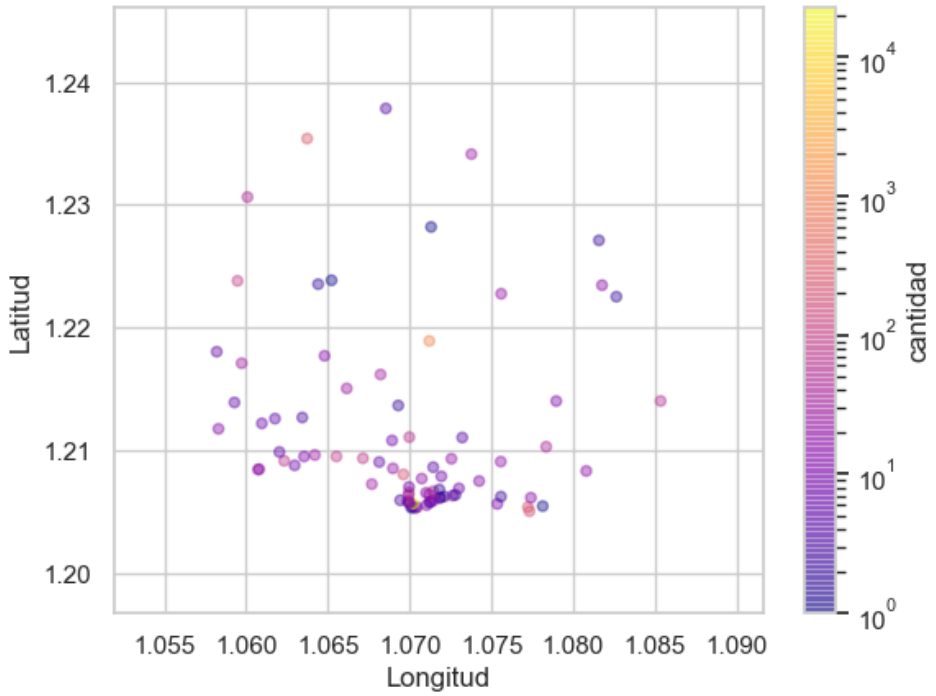

Figura 14: Ubicación geográfica de los clicks

Si bien la ubicación geográfica tiene una transformación lineal para ofuscar los datos, este gráfico nos sirve para darnos una idea de la distribución geográfica. Podemos apreciar que hay un gran agrupamiento en (1.21, 1.070). Esto nos permite reafirmar el hecho de que el lote de clicks que se nos proporcionó era de un único país.

4.4. Installs

4.4.1. Instalaciones por hora y día de la semana

La siguiente figura muestra las instalaciones que se realizan, contabilizando por día de semana y horario.

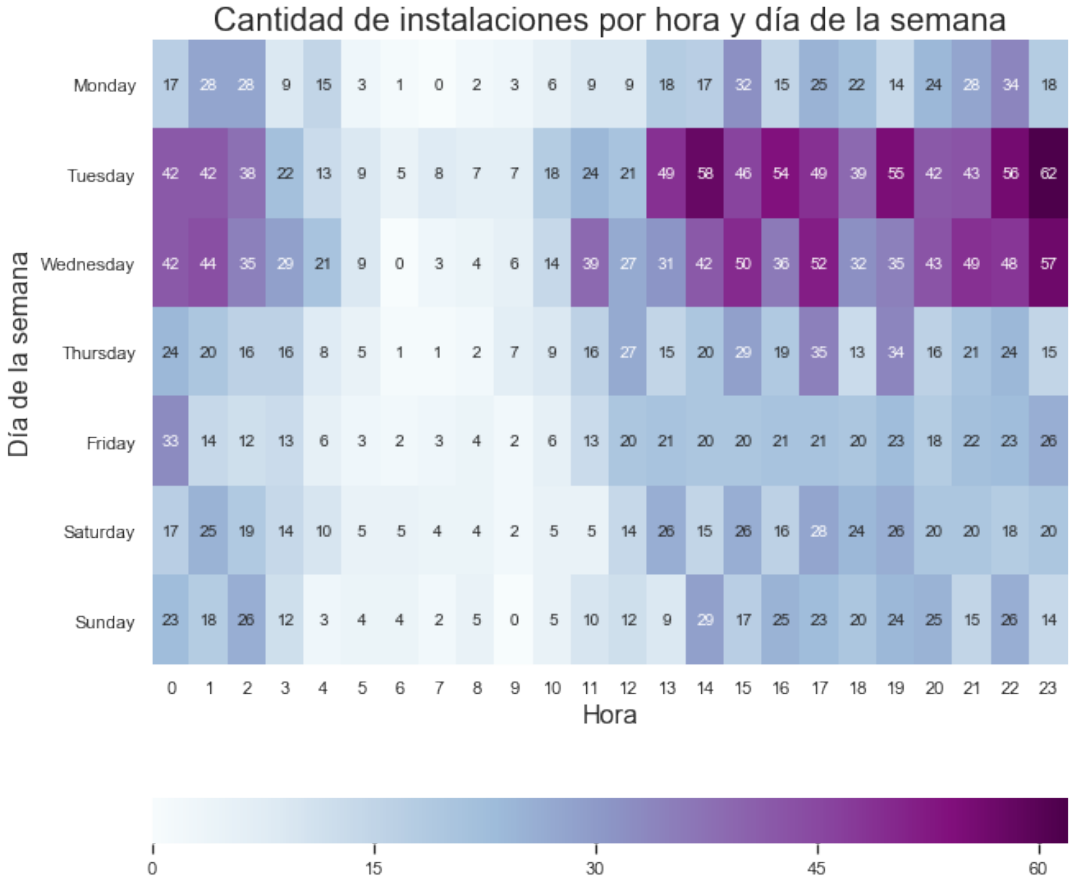

Figura 15: Instalaciones según día de semana y horario

Como se observó en la Figura 4, la mayor actividad se encuentra entre las 13 y 2hs del siguiente día, es decir en horario nocturno. Además en este caso vemos que los días que se realizan más instalaciones son Martes y Miércoles sobresaltando del resto.

Mientras se realizaba este análisis se pudo observar algo no menor, que de los datos proporcionados de Jampp, ninguna de las instalaciones se las atribuye a esta. Es decir que todas las instalaciones que se tienen no son ”gracias a” Jampp.

4.4.2. Instalaciones por hora

En este caso puntualmente nos focalizamos a visualizar cual es la franja horaria en que se producen más conversiones, el cual es el objetivo de Jampp.

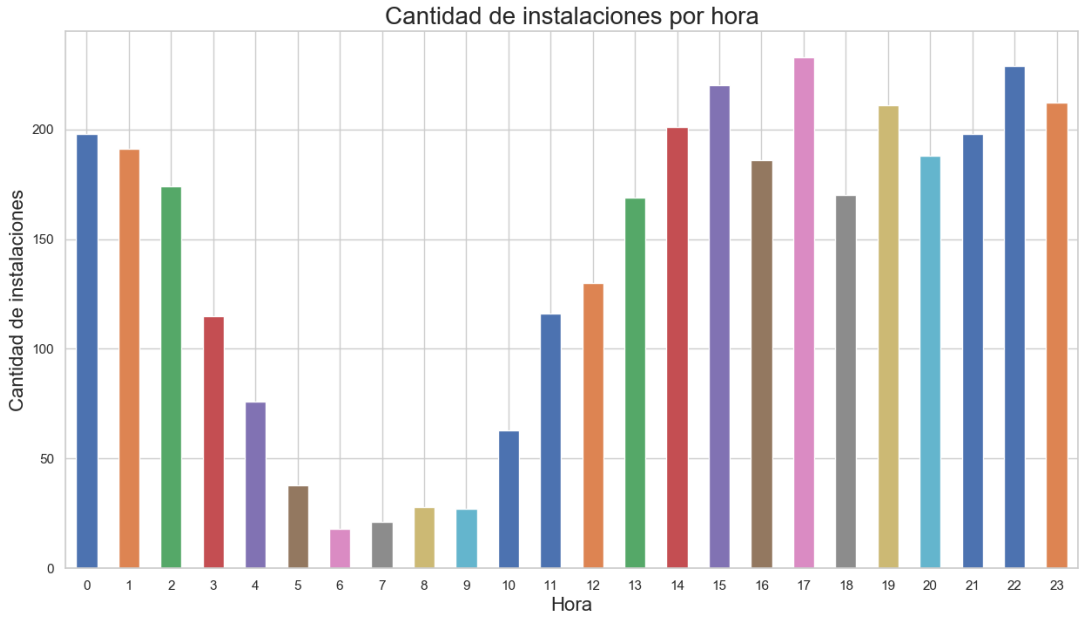

Figura 16: Instalaciones por hora

Visualizando la figura anterior, se puede concluir que es posible concretar una conversión en la franja horaria de 13hs a 02hs del día siguiente, con picos a las 17 y 22hs, donde el primero se puede suponer que se debe al mayor uso del dispositivo ya que finalizó su horario laboral. Esto a su vez marca un horario en el cual la mayoría de los usuarios trabaja y no hace uso de los dispositivos, este mismo va de 6/7 a 17/18hs.

4.4.3. Cantidad de instalaciones por IP

Algo interesante por visualizar es la cantidad de instalaciones que se realizan por IP, ya que es posible realizar una conversión por medio wifi, dando para una misma IP varias conversiones.

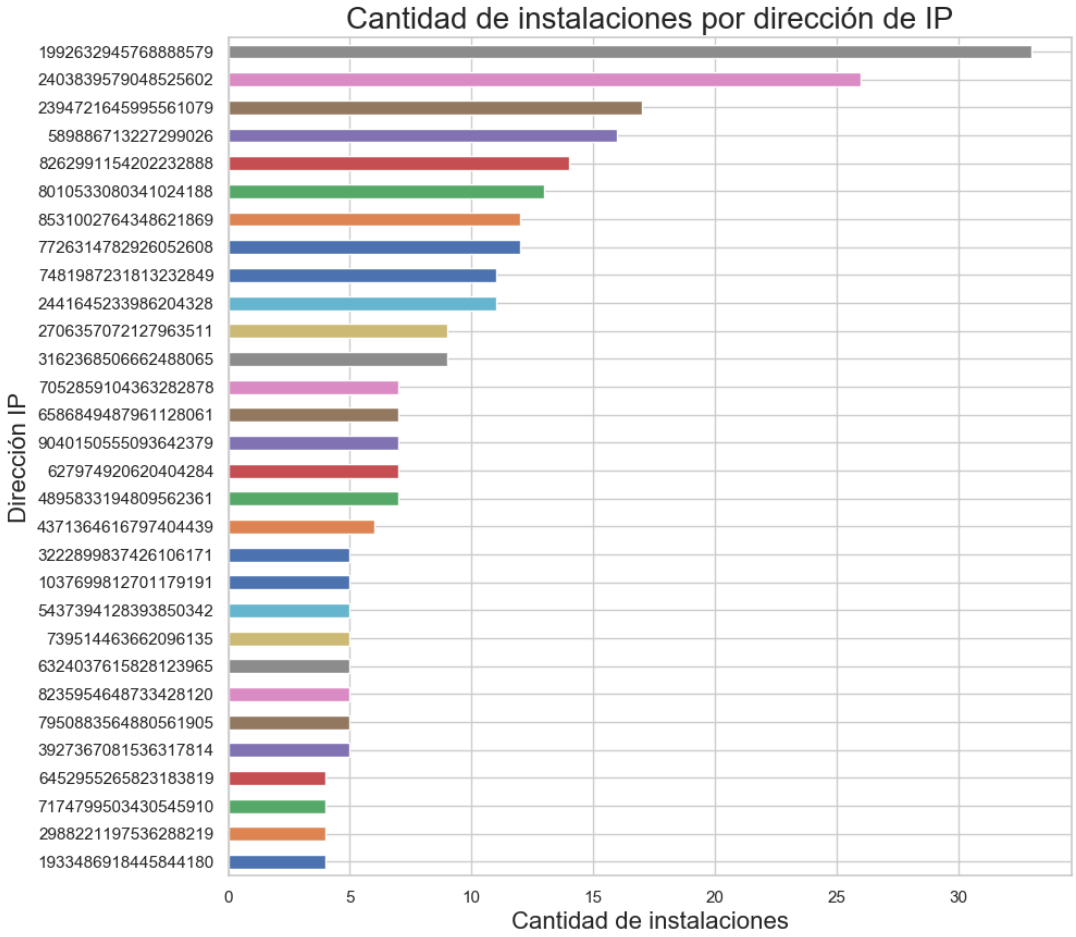

Figura 17: Cantidad instalaciones por IP

Esto puede inferir que hay instalaciones que se realizan por medio de una red wifi o por una local, ya que la ip pública es única y a varios dispositivos se le adjudican estas instalaciones. El dispositivo con mayor cantidad de instalaciones es 4, como se puede ver en la figura 8. Lo que nos lleva al siguiente análisis, instalaciones según la conexión.

Otra cosa que podemos plantear a partir de este gráfico es que varias de las instalaciones fueron hechas en un punto de reunión común de los usuarios. Un lugar con wifi publico (Ej.: bares, patio de comidas, shoppings). En la figura 14 vimos que los clicks no varían mucho en cuanto a la ubicación geográfica por lo que podemos considerar aún más esto.

4.4.4. Instalaciones por conexión

Como vimos anteriormente podemos encontrar que hay instalaciones que se generan por Wifi y otras no.

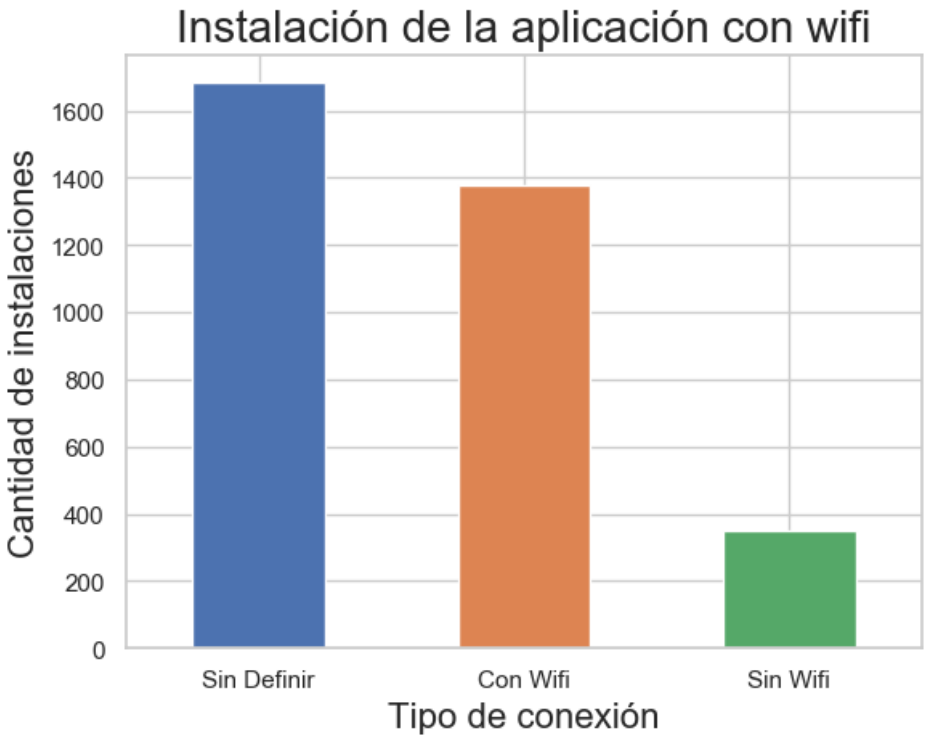

Figura 18: Instalaciones por conexión

Algo a destacar es que la mayoría de las instalaciones no poseen información acerca de haber utilizado o no wifi y lo segundo a destacar es que la gran mayoría utiliza wifi. Dentro de esta primer categoría de "Sin definir" no se sabe si involucra conexiones de redes móviles o estas están dentro de la categoría "Sin Wifi". Debido a la cantidad es posible considerar que "Sin Definir" tenga involucrados las instalaciones con uso de redes móviles, ya que hoy en día el uso de un dispositivo móvil utilizando redes móviles es muy común y en cierto punto podría ser mas que utilizar wifi.

4.4.5. Tipos de instalación

En *Installs* se posee el tipo de instalación, columna *kind*, con el cual analizamos cuantas instalaciones según el tipo se realizaron, es otras palabras, se contaron cuantos registros se poseen agrupando por tipo de instalación.

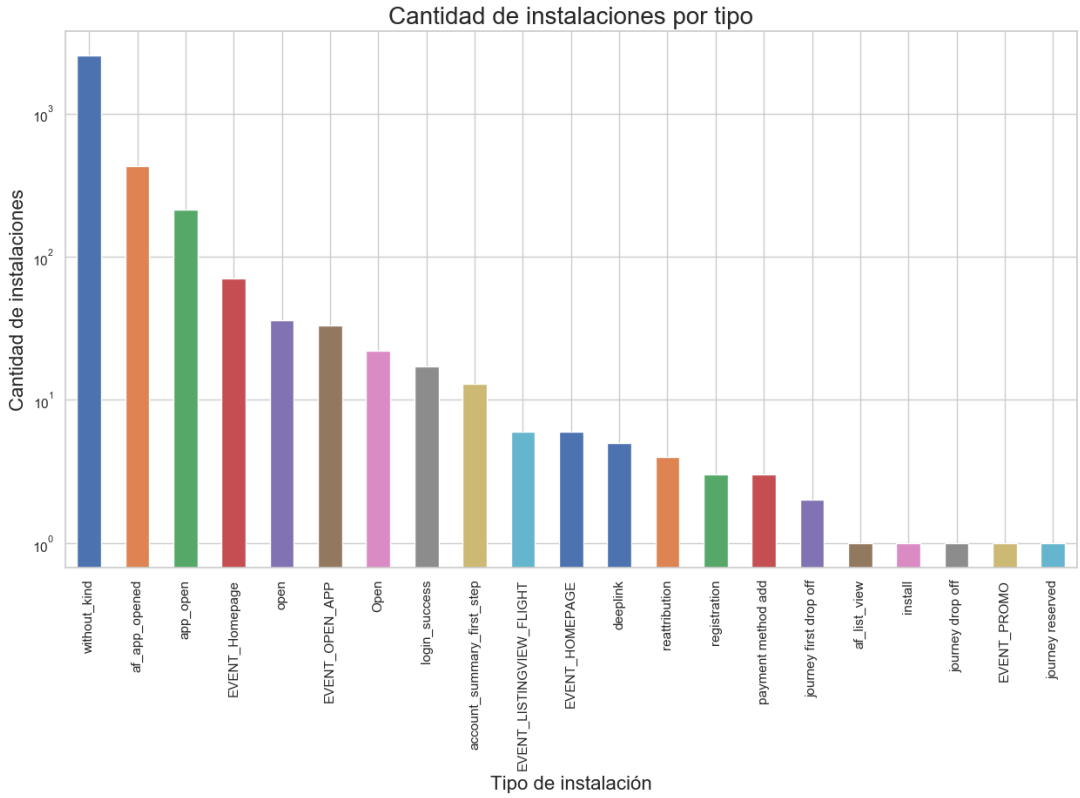

Figura 19: Cantidad tipos de instalación

Observando detenidamente vemos que el tipo con más instalaciones es en realidad cuando no se sabe el tipo de instalación, siguiendo como segundo af_app_oppened de AppsFlye que atribuye de donde proviene y de qué anuncio se produjo la instalación.

Otra observación es que es posible que estos tipos de instalaciones sean mas bien tipos de eventos de una instalación

4.4.6. Aplicaciones instaladas

Esta sección puede resultar importante a Jampp ya que podría influir en que clientes ”apostar” según que aplicación se presenta en la publicación. Lo más razonable es hacer foco en las aplicaciones que más se instalan, porque aumenta la probabilidad que al realizar una aviso de esta termine en una conversión de Jampp. Por otro lado, puede tomarse como objetivo o desafío buscar mejorar la publicación o en que medios y para que tipos de usuarios se muestran y generar mas instalaciones en aquellas que no tuvo conversiones.

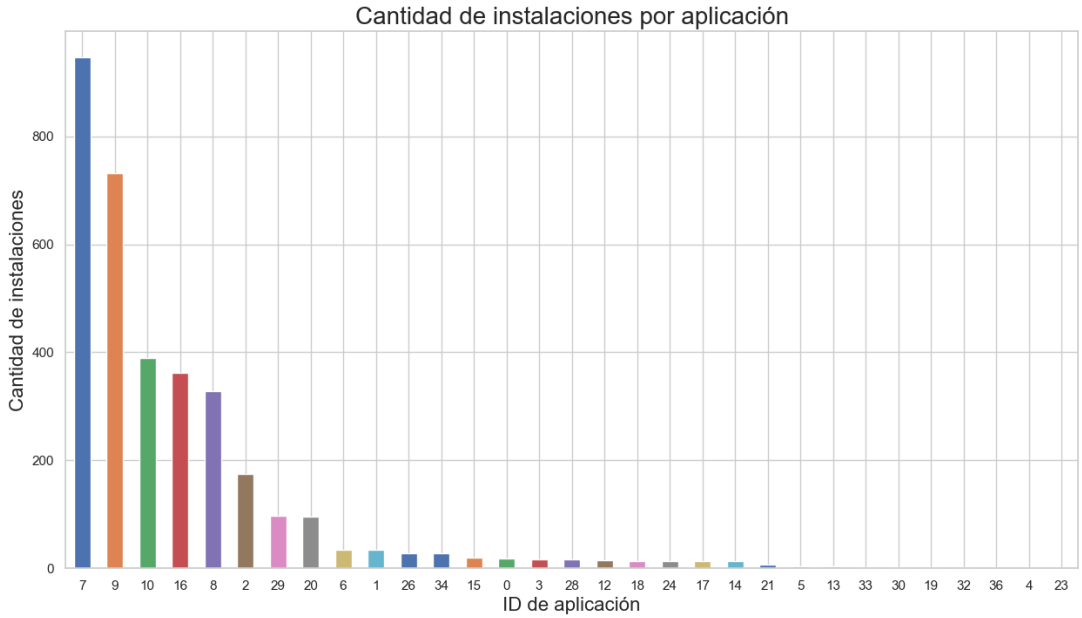

Figura 20: Aplicaciones instaladas

4.4.7. Aplicaciones instaladas implícitas

De las aplicaciones mostradas previamente, veamos cuales son instalaciones implícitas.

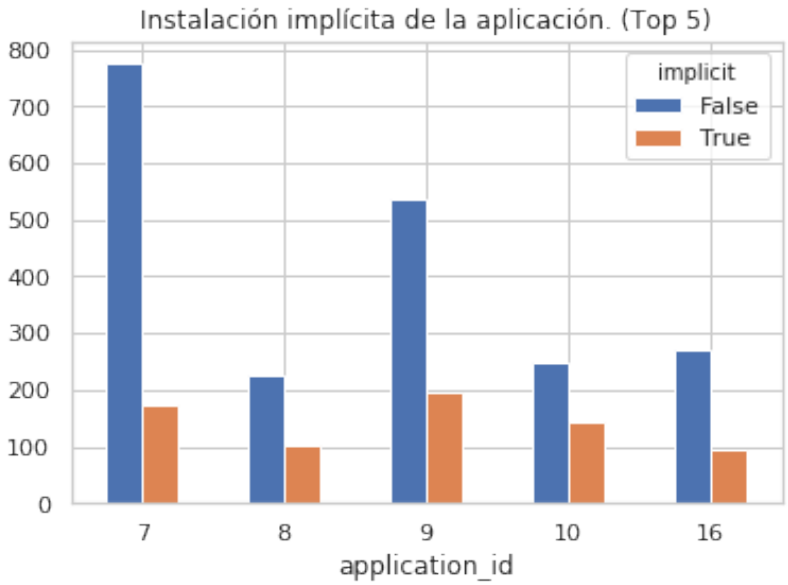

Figura 21: Aplicaciones instaladas implícitamente

Podemos ver que las aplicaciones que tienen más instalaciones implícitas coincide con las mismas que tienen más instalaciones.

4.4.8. Instalaciones por tipo de aviso

En esta sección buscamos analizar cual es la tendencia de instalaciones que se realizan según el tipo de la publicación.

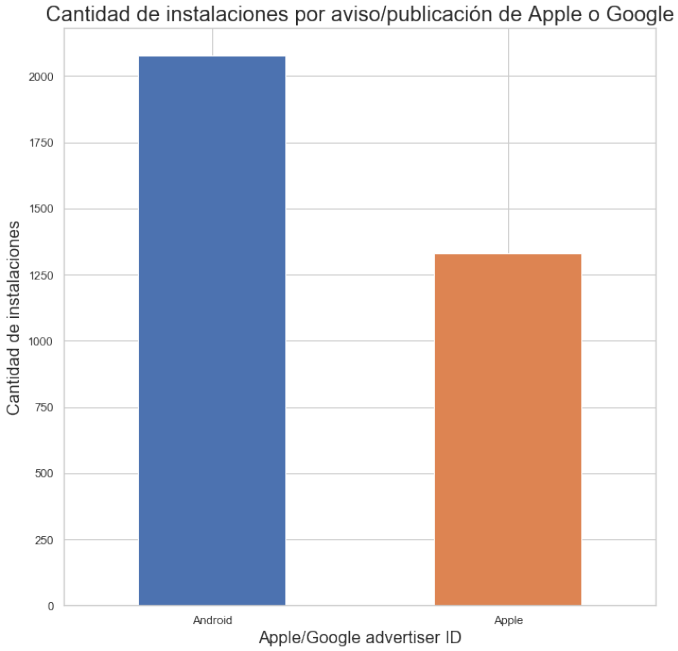

Figura 22: Cantidad instalaciones por tipo de aviso

Al ver la visualización anterior, vemos que los avisos de Android superan a los de Apple, algo que no es raro ya que Android tiene la mayoría del mercado de los smartphones y Apple solo cubre sus productos, ej.: iPhone, iPad, etc

4.4.9. Instalaciones - Subastas

Nos interesa mostrar cual es el porcentaje de las instalaciones que poseen un registro en las subastas proporcionadas por Jampp y cuales no. Igualmente, tenemos que tener en cuenta que las subastas son un muestreo de 25 % de las subastas reales que suceden a lo largo del tiempo, por lo que tendría sentido que hayan instalaciones que no se puedan asociar a subastas.

Figura 23: Instalaciones - Subastas

4.4.10. Tipos Eventos y mayores instalaciones

Algo que se consideró importante de analizar es que tipos de eventos se efectuó por los dispositivos con mayor cantidad de instalaciones

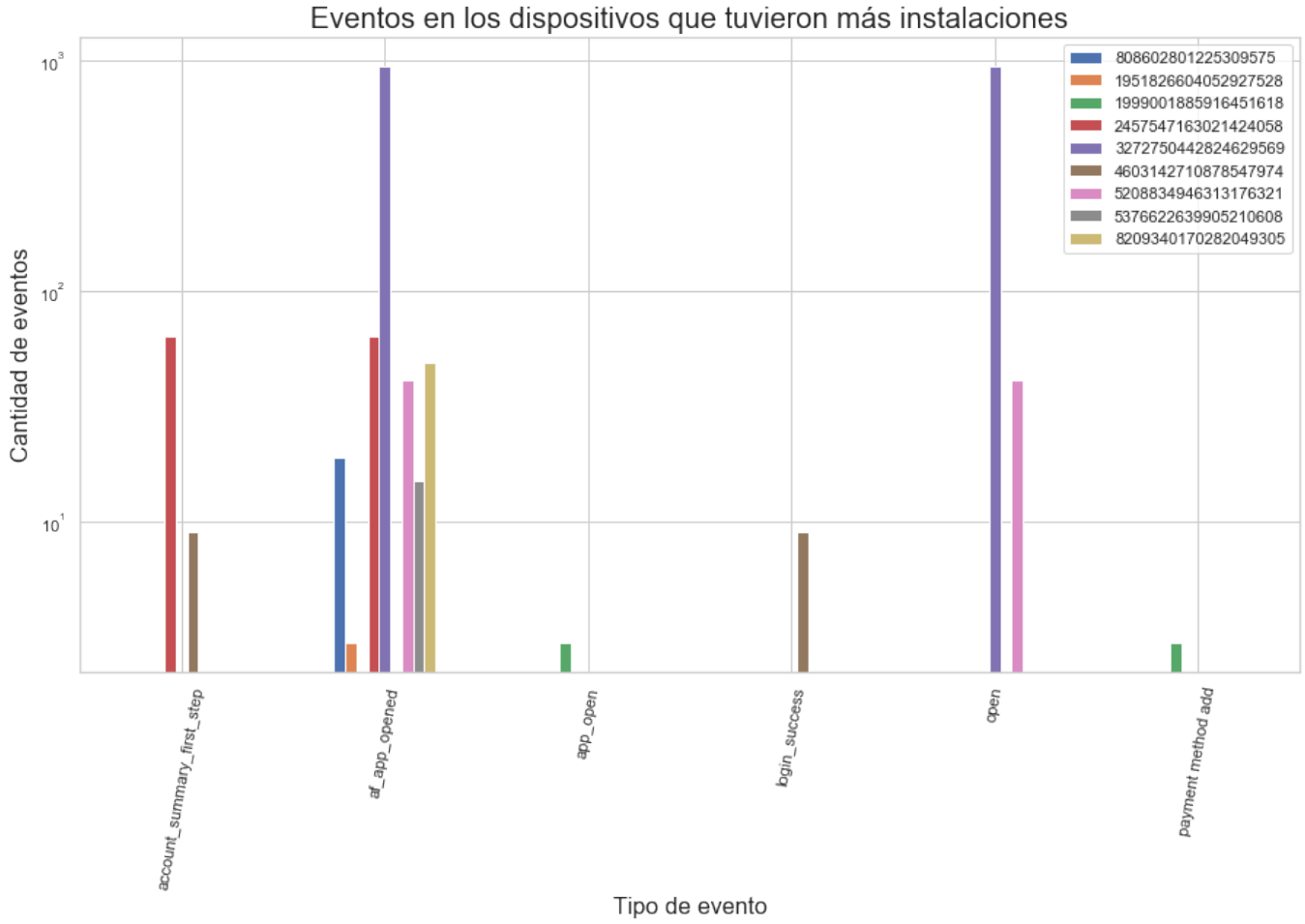

Figura 24: Eventos y mayor instalaciones

Para poder apreciar más el gráfico, solo se realizo con algunos dispositivos y no su totalidad (específicamente top mayor cantidad instalaciones). Podemos interpretar y observar que los que representan mayor cantidad de eventos se encuentran en el top 20 de los dispositivos con mas instalaciones, lo que nos puede dar un indicio de que tipos de eventos esperar de un usuario que tenga varias conversiones, y si se desea que tenga mayor probabilidad de convertir, se desea que tenga similarmente este comportamiento.

4.4.11. Session User Agent

El *Session User Agent* contiene información sobre el usuario que origina la solicitud de conexión, que a menudo es utilizada por los servidores para ayudar identificar el alcance de los problemas de interoperabilidad reportados y para analíticas de que navegador o sistema operativo utiliza.

Vamos a realizar una comparación de las plataformas utilizadas para distintas instalaciones.

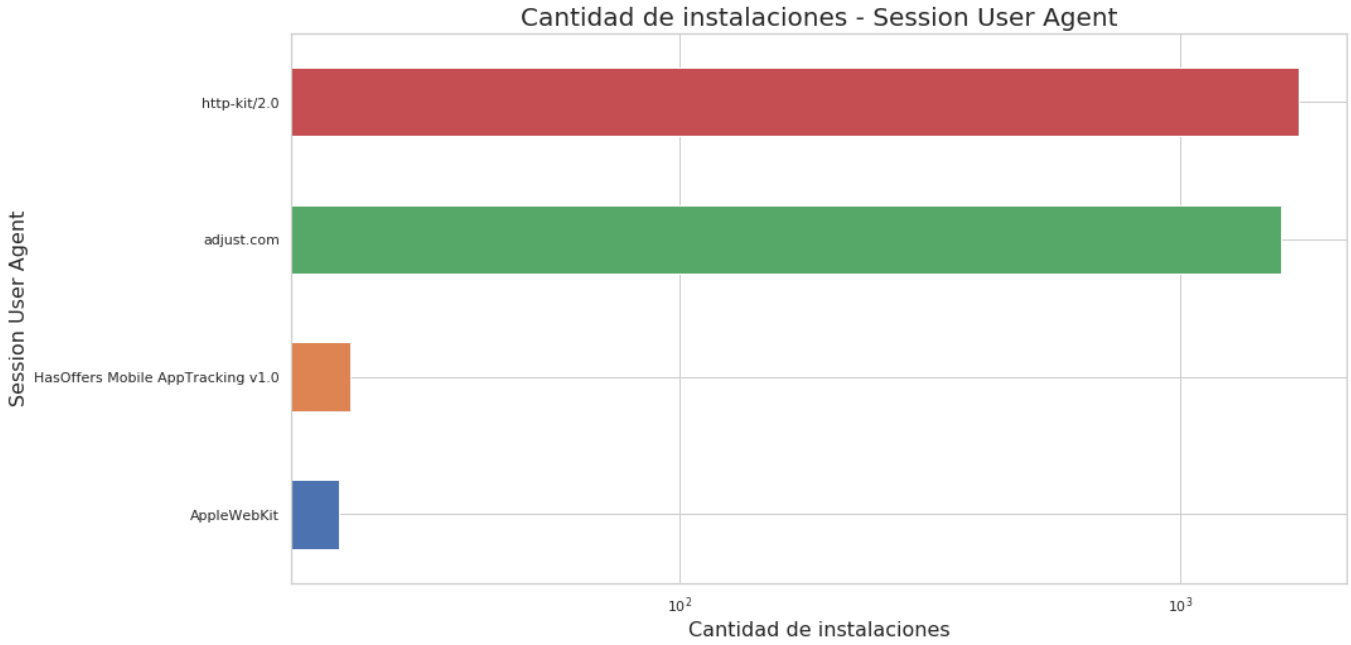

Figura 25: Session user agent

Buscando información de que nos podría proporcionar un ”user_agent” encontramos que el formato en el que pueden estar presentados podría ser el siguiente:

Mozilla/[version] ([system and browser information]) [platform] ([platform details]) [extensions]

en el cual observamos que tiene una estructura muy similar a lo proporcionado, encontrando información de la plataforma.

En primer lugar tenemos ***http-kit/2.0*** el cual es un servidor/cliente HTTP de alto rendimiento controlado por eventos para Clojure. En segundo lugar ***Adjust*** que es una compañía de medición móvil donde unifican actividades de marketing en una plataforma poderosa, brindando información necesaria escalar un negocio. En todos los casos este análisis nos brinda información de como se realizó una instalación, ya sea qué plataforma usó, si tiene o no compatibilidad con determinado engine como el de Mozilla, como fue la ”negociación” de la conexión realizada por medio de http y más que nada que la mayoría utiliza servicios http-kit o de la compañía Adjust.

4.4.12. Sistema operativo de las instalaciones

Con información obtenida de *Session User Agent* se pudo interpretar que tipo de sistema operativo utilizan los dispositivos que realizan instalaciones.

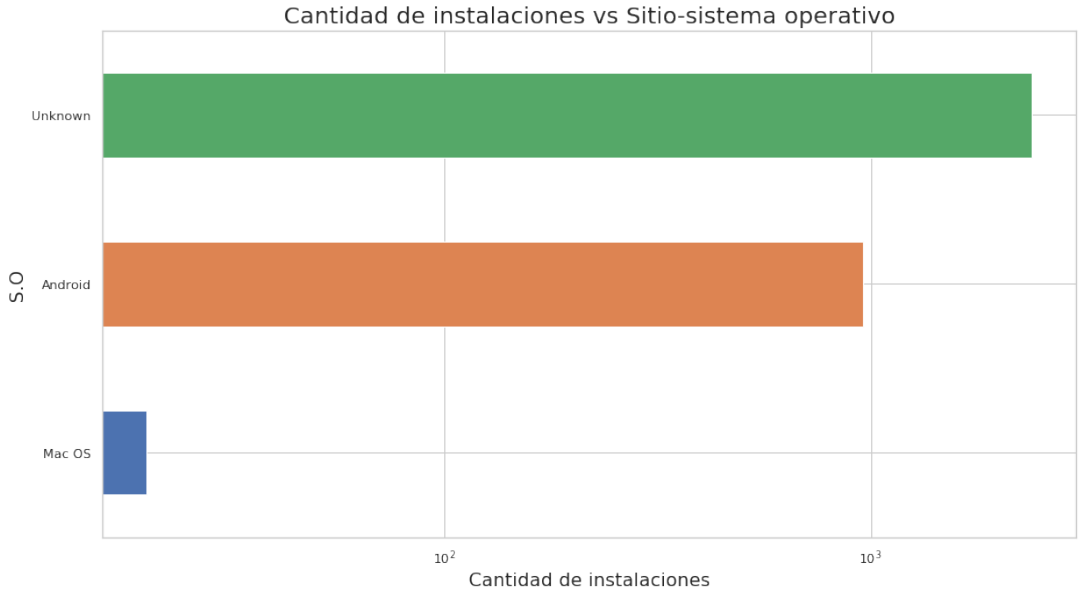

Figura 26: Instalaciones - S.O

Dentro de la categoría Unknown se pueden encontrar algunos que utilizan HTTP-kit y/0 Adjust, ya que no nos proveen información del sistema utilizado a diferencia del resto, donde vemos una clara diferencia entre Andriod e IOS.

4.4.13. Tiempo entre instalaciones

Algo interesante de analizar es el tiempo promedio que tardan los usuarios en realizar una conversión, analizando cuantos días y horas demora en convertir nuevamente, ya que esto puede dar conocimiento a Jampp de cuanto tiempo debe esperar aproximadamente para una nueva posible conversión.

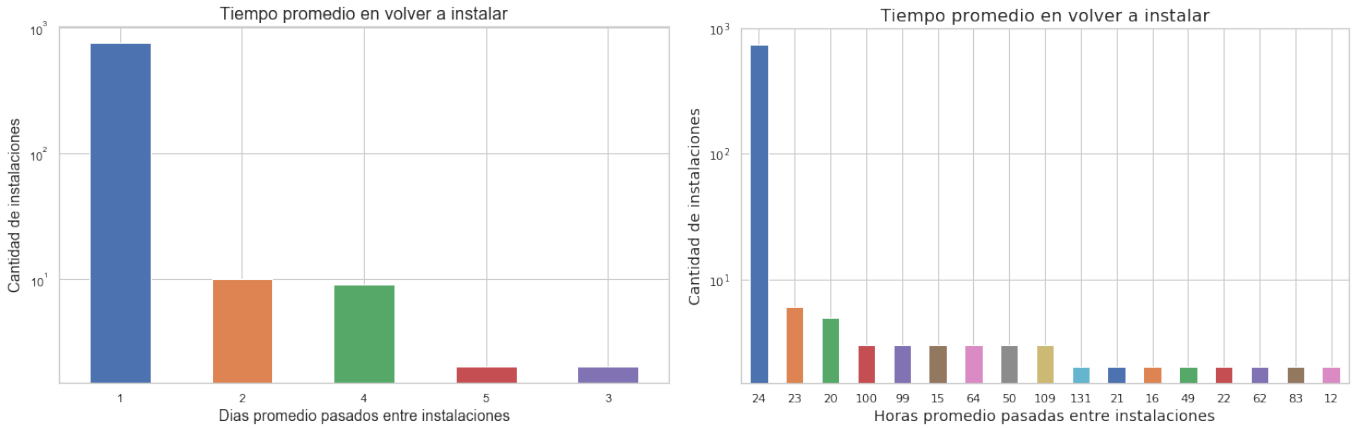

Figura 27: Tiempo entre instalaciones

Como se puede ver en los gráficos anteriores la gran mayoría se concentra dentro de los primeros dos días, puntualmente no pasadas las 24 hs convierten nuevamente. Esto se puede deber a que el usuario puede instalar la aplicación, abrirla y realizar determinadas acciones que se consideran un *install*. Para que se pueda apreciar de forma un poco más precisa, se realizó el mismo análisis pero por horas, para tener un "desglose" de día en horas, obteniendo así por horas cuanto en promedio se tarda en convertir.

4.5. Events

En está sección analizaremos el set de datos *Events* y su relación con los otros set de datos. Este set contiene propiamente dicho eventos registrados por las aplicaciones, puede contener instalaciones, logeos a la app, acciones dentro de la app y demás.

4.5.1. Horario de los eventos

Algo importante es ver el horario de los eventos, que día suceden y con que cantidad para poder observar un comportamiento en el usuario, cuando es más típico que utilice la app.

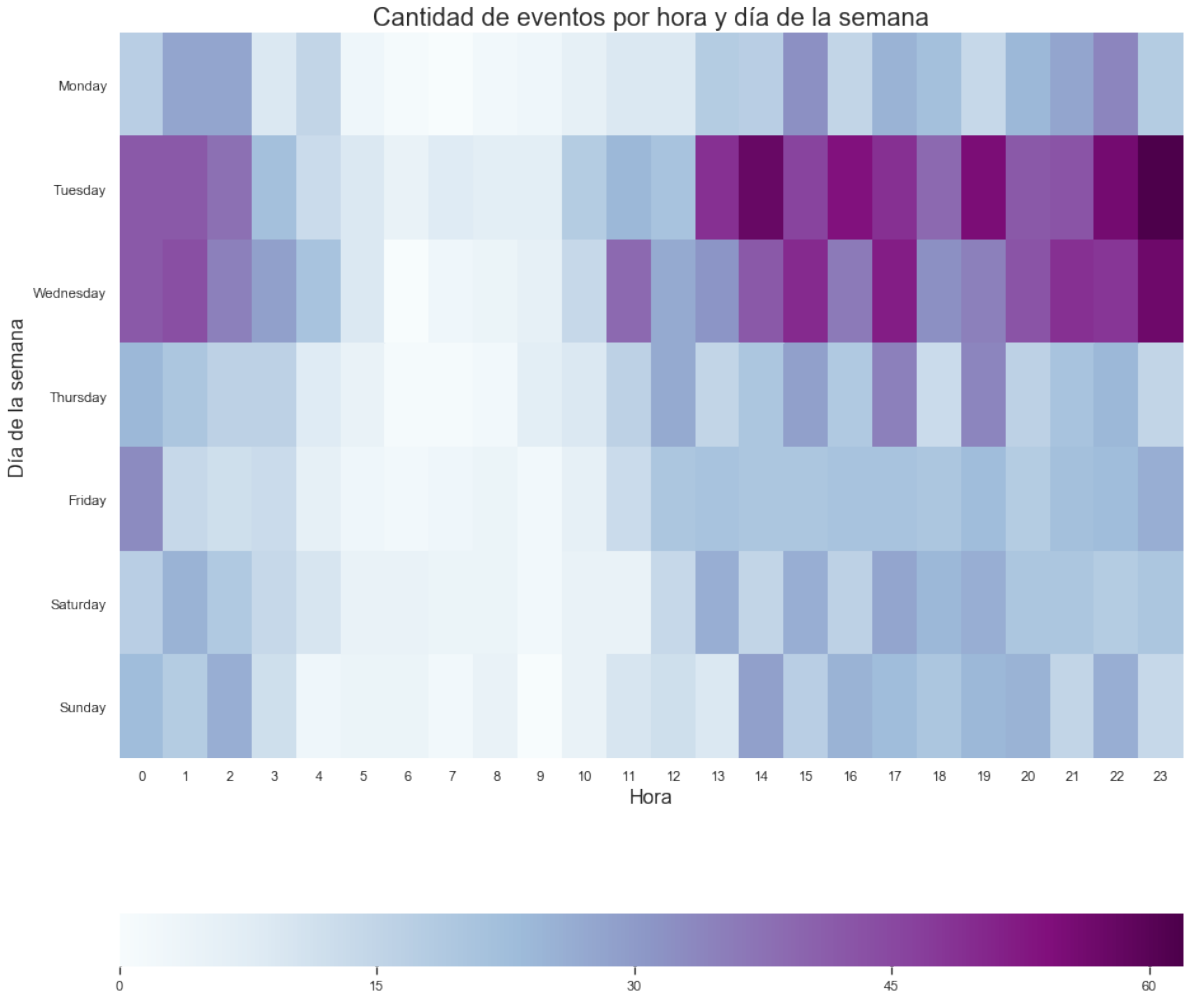

Figura 28: Horario eventos

En la figura anterior podemos ver que hay una gran similitud con la figura 15, en la cual se analizó por día de semana y horario la cantidad de conversiones que se realizan. Podemos observar también que hay un gran uso de los dispositivos y en particular de las aplicaciones que registran eventos, los días Martes y Miércoles y en la franja horaria de 22 a 2 hs.

En la siguiente figura analizamos por hora independiente de que día de la semana se realizó el evento.

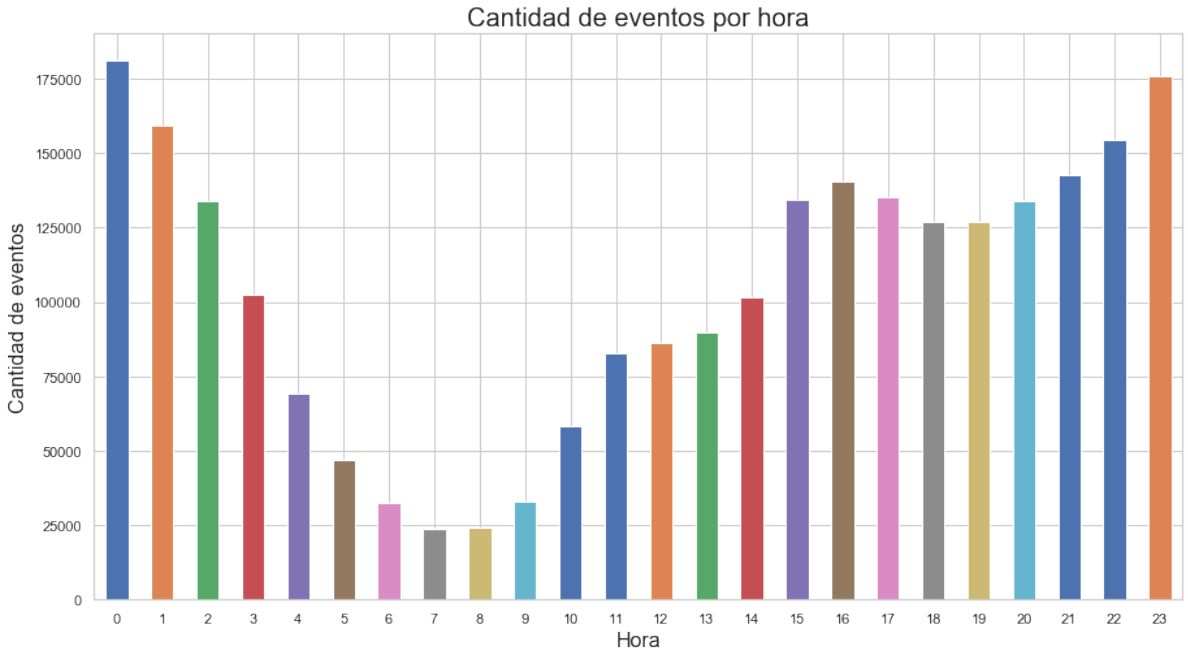

Figura 29: Eventos por hora

Como se menciona anteriormente vemos con claridad que entre las 22 y las 2 hs es cuando más eventos se producen y vemos que cuando es de esperarse entre las 3 y 9 hs es cuando menos se producen, ya que es de suponer que la mayoría de las personas duermen.

4.5.2. Eventos por ciudad

Cada registro de evento posee una columna que indica en que ciudad se efectuó el evento, por lo que se considera interesante encontrar que ciudades tienen más actividad.

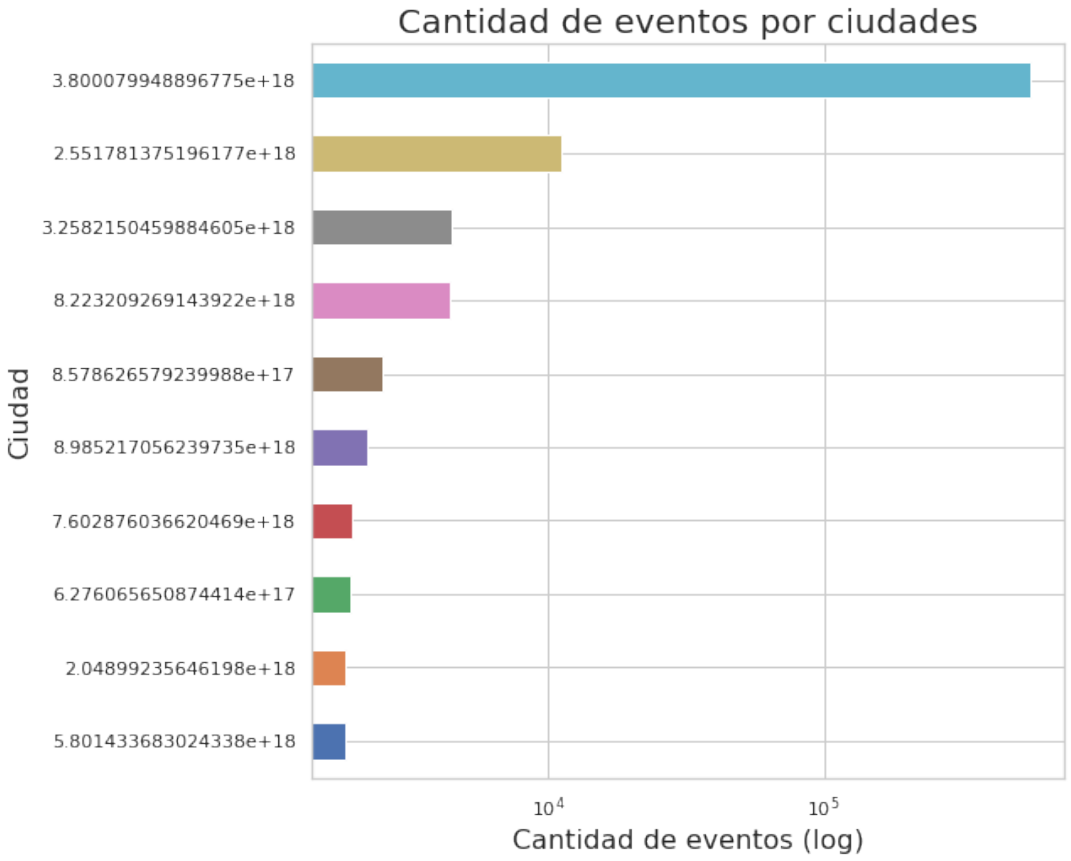

Figura 30: Eventos por ciudad

Como fueron anonimizados los datos no se tienen nombre de las ciudades pero si un hash que las representa. Podemos observar una alta diferencia de eventos entre las primeras dos ciudades y el resto, sería interesante comparar con la ubicación geográfica de clicks figura 14 pero no fue posible debido a la transformación lineal de latitud y longitud y la desconocida ubicación de cada ciudad.

Top 2 Ciudades con más eventos

A continuación se comparan los eventos de las primeras dos ciudades de forma individual. Se utiliza el nombre ”Ciudad 1” y ”Ciudad 2” a la primer y segunda ciudad con más eventos respectivamente, con fines de hacer más legible la comparación.

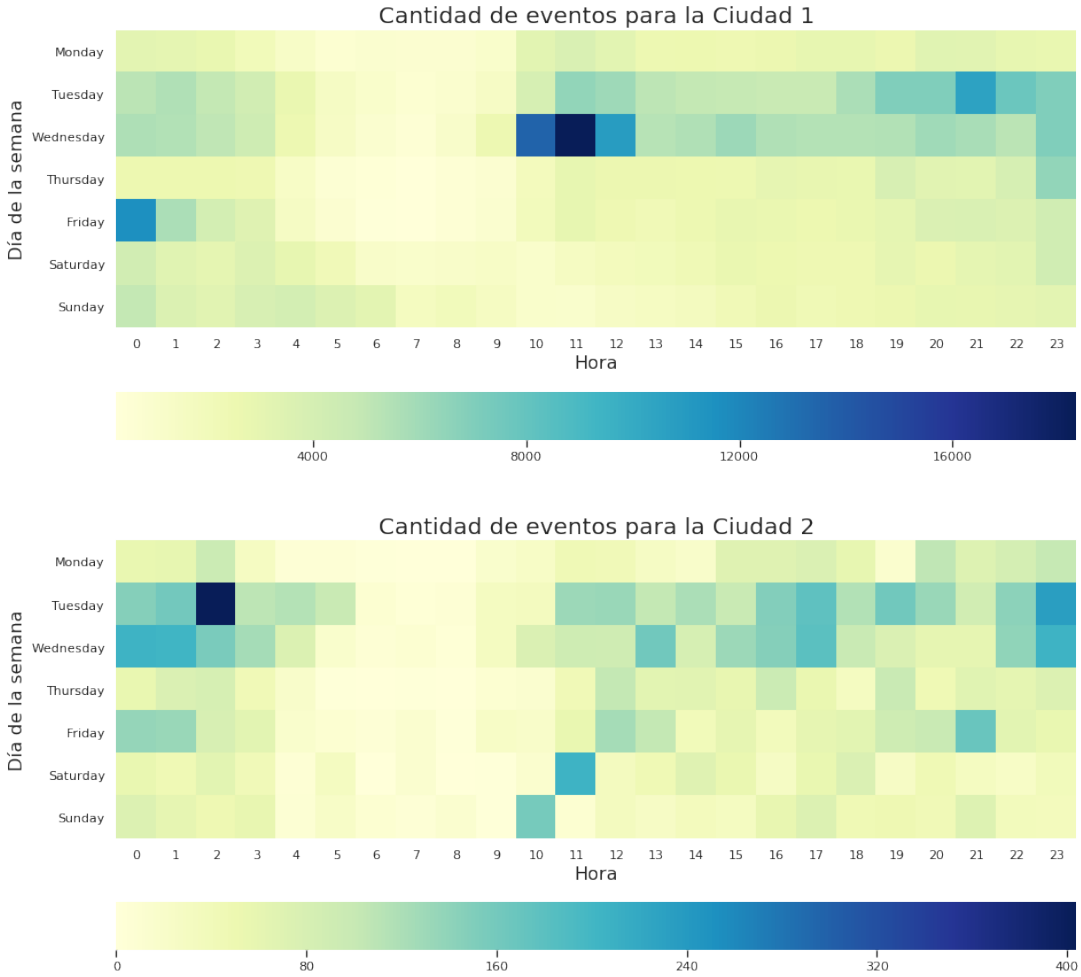

Figura 31: Eventos - Ciudad 1 vs. Ciudad 2

Se puede observar que difieren de lo mostrado por la figura 28, pero a su vez se nota la franja horaria. Por otro lado vemos que entre la primer ciudad y la segunda hay una gran diferencia de cantidad de eventos en el mediodía del Miércoles y el jueves a las 2 am, puede deberse a determinadas actividades que se realicen en cada ciudad.

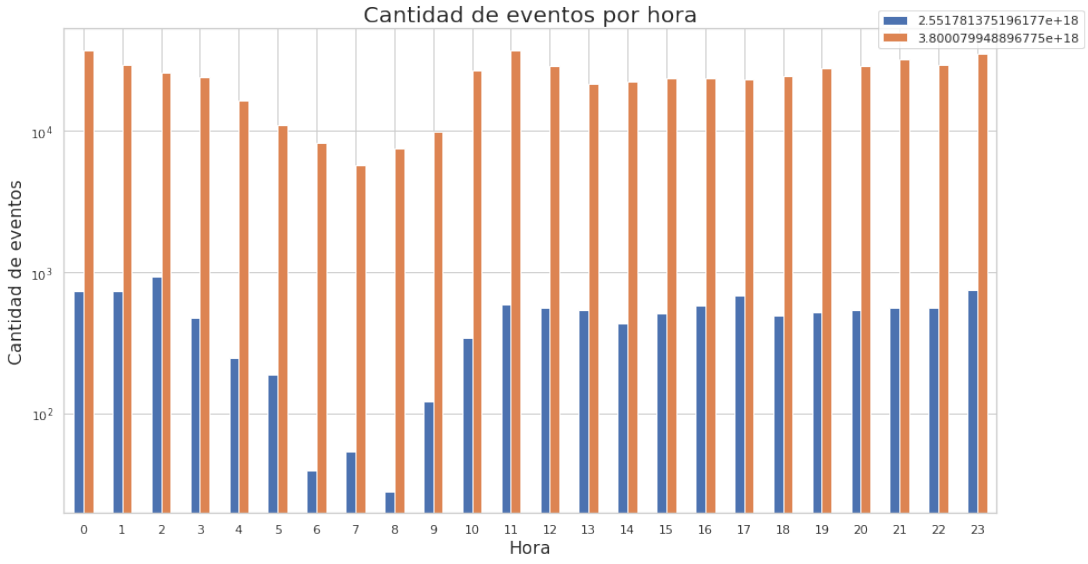

Figura 32: Eventos por hora - Ciudad 1 vs Ciudad 2

Como se observo en la figura 31 se ve de forma más notoria la diferencia de cantidad de eventos por hora, pero a

pesar de todas estas diferencias se nota con diferencia de cual posee más cantidad de eventos totales.

Dispositivos por Ciudad

Se analiza a continuación la cantidad de dispositivos que posee cada ciudad.

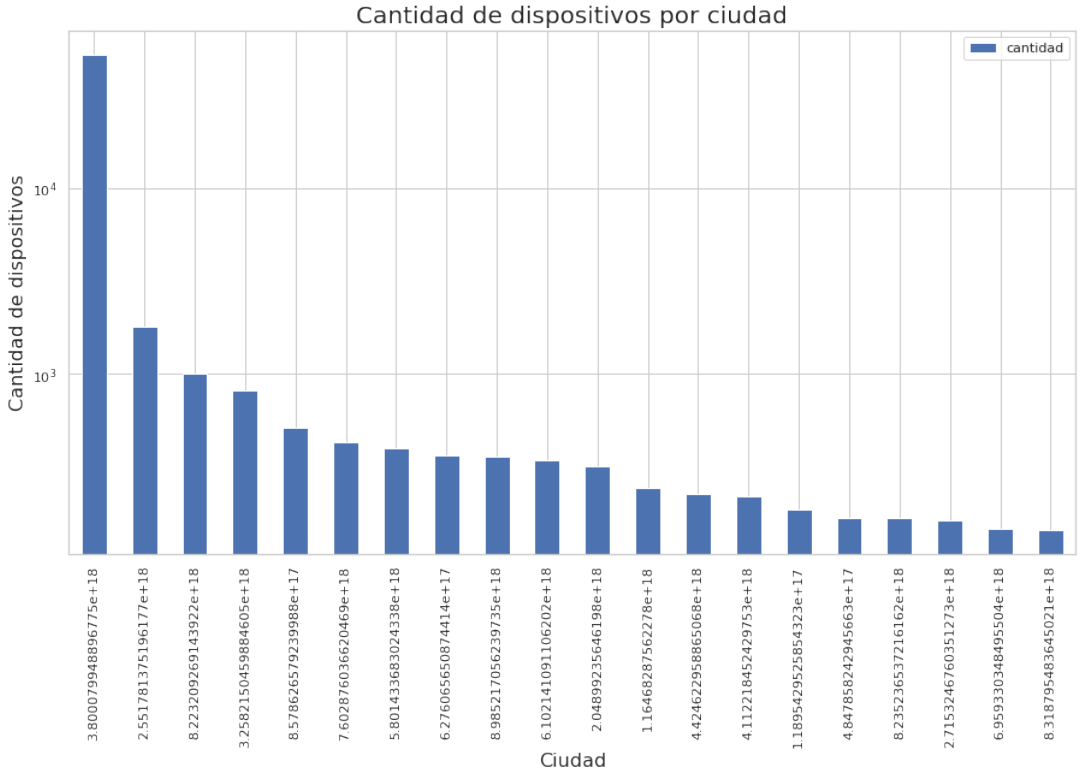

Figura 33: Dispositivos por ciudad

Vemos que la cantidad de dispositivos por ciudad es proporcional a la cantidad de eventos por ciudad, dejando las mismas dos primeras ciudades con mayor cantidad de dispositivos que las primeras dos ciudades con mayor cantidad de eventos

4.5.3. Eventos - Sistema Operativo de los dispositivos

A continuación se realiza un análisis por sistema operativo de los dispositivos.

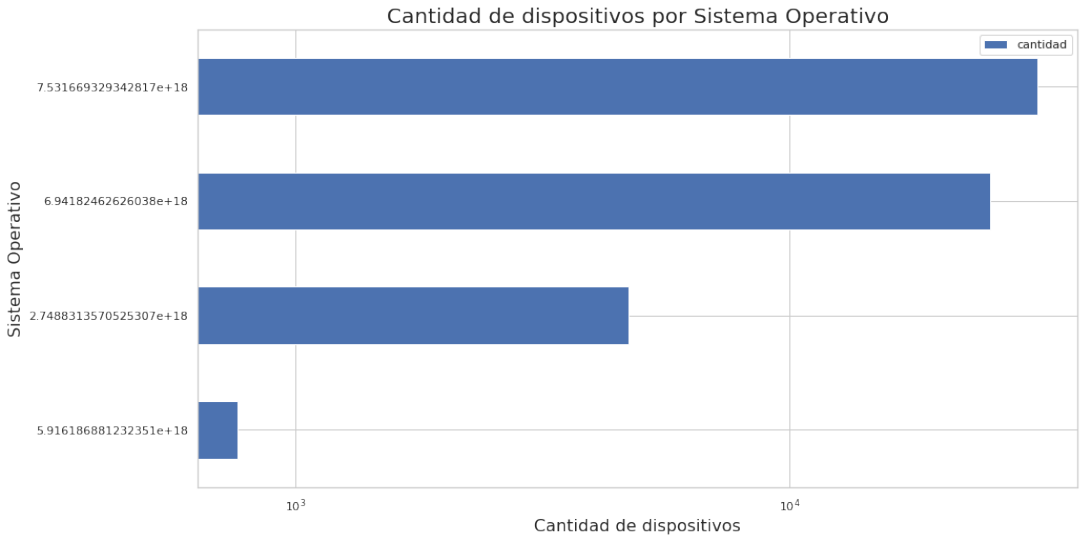

Figura 34: Eventos - S.O

En la figura 26 se realizó un análisis sobre el S.O con información obtenida por el session_user_agent , del cual podría verse una semejanza entre Andriod y Mac OS con los representados por el hash 5.916 y 2.748 respectivamente, el resto representaría la categoría UnKnown.

4.5.4. Eventos y Publicidades

Vamos a analizar la cantidad de eventos según el tipo de publicación (Apple o Google).

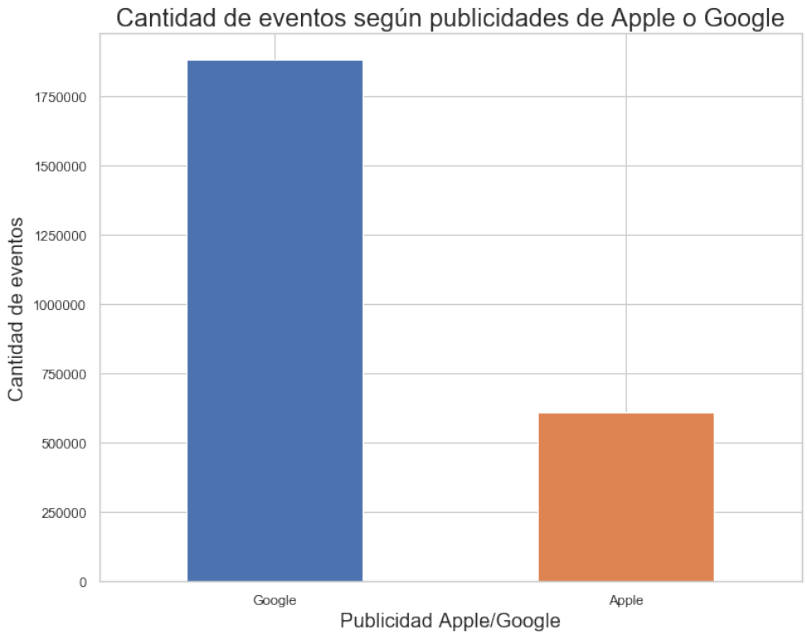

Figura 35: Cantidad eventos y tipo de aviso

En el gráfico anterior se puede observar una gran diferencia de cantidad de eventos que se producen por publicaciones de tipo Google que las de tipo Apple, lo cual no implica que hayan o no más publicaciones de un tipo u otro, sino que simplemente se produjo más eventos para las de tipo Google. Algo similar se puede observar en la figura 13, donde también se aclara que son registros representativos de la cantidad de clicks o en este caso eventos, no la cantidad de publicaciones , sino donde sucedió el evento.

4.6. Auctions

4.6.1. Mayor promedio de instalaciones por subastas

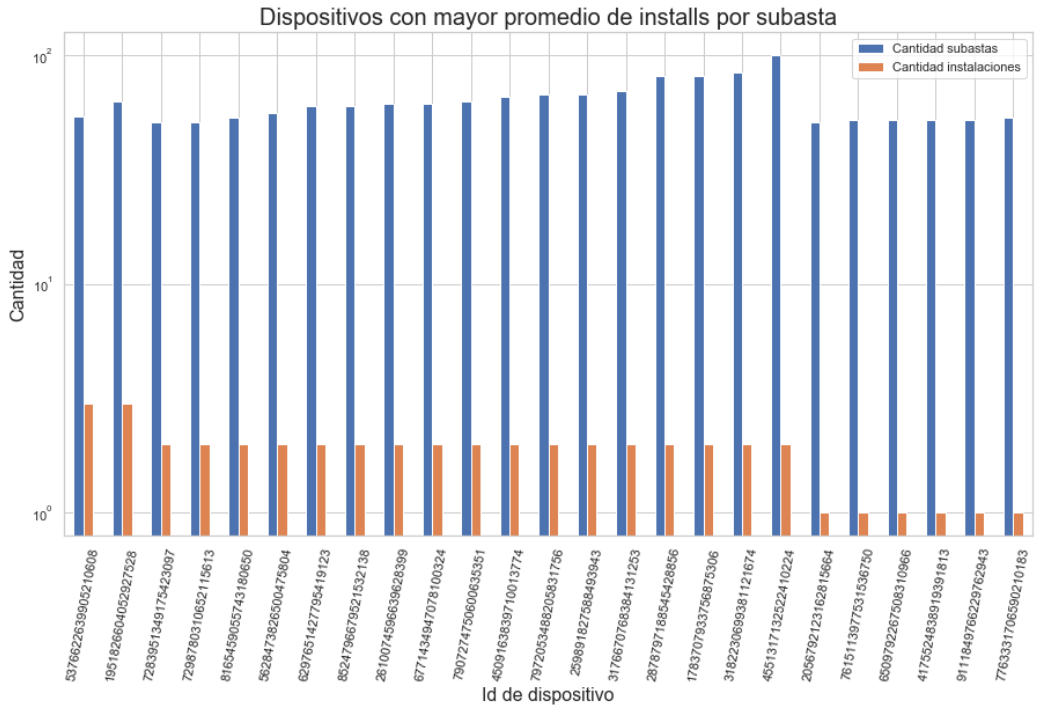

Figura 36: Mayor promedio de instalaciones por subastas

Con esta visualización podemos ver los dispositivos que tienen un mayor promedio de instalación por subastas registradas, esto podría llegar a servir para ver que dispositivos son mas propensos a instalar aplicaciones y así saber a quien darle mas importancia a la hora de apostar en una subasta, pero dado que la cantidad de instalaciones por dispositivo es muy poca, es difícil encontrar algún dato que sea muy relevante

4.6.2. Menor promedio de instalaciones por subastas

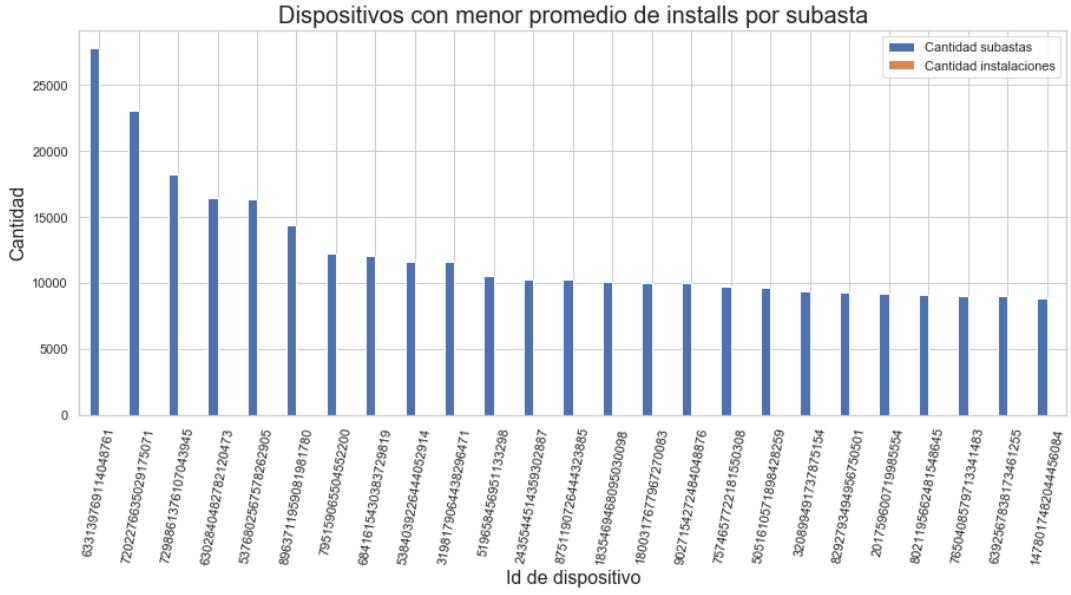

Figura 37: Menor promedio de instalaciones por subastas

En esta visualización podemos ver aquellos dispositivos que tuvieron el menor promedio de subastas, en este caso aquellos que estuvieron en la mayor cantidad de subastas pero nunca realizaron una instalación, de esta visualización se podría ver a que dispositivos darle menor importancia a la hora de apostar en una subasta. De esta visualización también se puede ver que quien tiene menor promedio es el dispositivo que a su vez es el que mas subastas realizo en total.

5. Conclusión

Luego de todo el análisis que se pudo realizar con el set de datos, se pueden destacar las siguientes observaciones que se hicieron a lo largo de todo este trabajo práctico.

Realizando un análisis de horarios en el que se registran las actividades, se pudo observar que **los usuarios realizan un mayor uso de las aplicaciones en horario nocturno y fuera del horario laboral**. Esto puede indicarnos que es un buen horario para apostar por usuarios, mostrándole publicidades que transmitan la información deseada de manera inmediata ya que otro comportamiento que resalta de los usuarios es que **dentro de los primeros segundos de visualización del aviso es cuando más clicks se hacen**. De esta forma sería ideal llamar la atención del usuario y brindarle la información deseada lo antes posible, evitando que llegue a perder su interés.

Otro aspecto que pudimos analizar (figura 20) fue qué clientes obtenían la mayor cantidad de instalaciones. De este análisis concluimos los que pueden más fácilmente obtener instalaciones y, por lo tanto, por cuales conviene apostar. A su vez un posible ”desafío” es mejorar las publicaciones de los que tienen menos clicks y proponer un mejor diseño para obtener más clicks y, posiblemente, más conversiones.

Siguiendo con el análisis del comportamiento del usuario, es importante poder estimar cuantos días podrían llegar a pasar hasta que genere una próxima conversión. Con la ayuda de la figura 27 podemos ver la cantidad de días promedio que tarda en volver a convertir, dándonos una idea de que **se obtiene otro install con una demora menor a 2 días**.

Teniendo en cuenta el comportamiento de los usuarios podemos establecer categorías para los mismos en base a su historial:

- **Usuario que interactúa con las publicidades y convierte**: Son los usuarios que regularmente hacen instalaciones y prestan atención en las publicidades. Estos usuarios son los que más nos interesan a la hora de apostar por un espacio publicitario aunque habría que tener la consideración de no saturarlo.
- **Usuario que interactúa con las publicidades y no convierte**: Si bien, es un usuario en el que sirven las publicidades, habría que analizar que tanto vamos a considerarlo para el retargeting. Si es un usuario que luego de mostrarle mucha publicidad no logra convertir no nos beneficia. Sin embargo, si luego del retargeting si convirtió entonces si fue remunerativo y nos beneficia.
- **Usuario que no interactúa con las publicidades**: Teniendo en cuenta que la subasta que se realiza en tiempo real es para determinar si mostraremos el anuncio o no, este tipo de usuarios generaría grandes perdidas. Tal vez es un tipo de usuario al que no querríamos mostrarle publicidad ya que no generara una conversión de manera explícita que se nos atribuya.