Organización de Datos (75.06):
Trabajo práctico n.1

Análisis exploratorio de datos

Caso de estudio: Jampp

Alumnos: Grupo 14

Apellido, nombre	Padrón	Correo electrónico
Aristegui, Facundo	90646	faristegui@fi.uba.ar
Castellanos, Cesar	81404	castellanos.cesar@gmail.com
Cozza, Fabrizio	97402	fabrizio.cozza@gmail.com
Rozanec,Matias	97404	rozanecm@gmail.com

Facultad de Ingeniería - Universidad de Buenos Aires

Introducción

Enunciado

El enunciado se encuentra en el siguiente link:
https://docs.google.com/document/d/1kZRZynbiVVR-DLtGVaKf7EvpXo5REZzM2Lga7Sdz7h0/edit

Acerca de Jampp

Jampp es una empresa centrada en una plataforma de Mobile App Marketing. La plataforma ayuda a anunciantes a promover sus apps a nivel global y también a recuperar a aquellos usuarios que ya instalaron la app pero están inactivos. Uno de los beneficios del servicio es que optimiza la compra de tráfico en base al nivel de actividad que los usuarios tienen en la app, de esa manera, sus clientes no obtienen solo instalaciones, sino que logran resultados concretos de negocio.
En palabras de su cofundador Diego Meller, *Jampp brinda una combinación única de volumen (logrado mediante la agregación de prácticamente todo el inventario de publicidad móvil disponible) y performance. La plataforma aprende qué señales producen los usuarios que mejor convierten para cada aplicación y, en función de eso, elige donde y que avisos mostrar a cada usuario. [...] El marketing de aplicaciones ya no se trata de conseguir instalaciones (más del 80% de los usuarios dejan de usar la aplicación luego de 6 meses), sino de lograr que esas instalaciones se conviertan en usuarios activos.*
Jampp hace esto posible a través de su plataforma propietaria de compra programática en tiempo real (Real Time Bidding), que está integrada a 18 exchanges de RTB y más de 150 ad networks móviles.

Challenge 7506

Según lo expuesto por Ignacio Javier Mermet, vocero de Jampp en la materia, los problemas con que enfrenta la compañía vienen definidos por la **saturación de usuarios** y saber **cuándo se podrá apostar** por un usuario.
Un usuario se maneja en varias aplicaciones durante el día. Algunas de esas aplicaciones son clientes de Jampp y permiten conocer cómo se comporta ese usuario dentro de la aplicación. Participa en varias subastas durante el día, en las que se invita a Jampp a participar.

Objetivo

El objetivo del presente trabajo práctico es obtener un entendimiento tal de los datos brindados por la empresa, que le permita al equipo sentirse cómodo con los mismos, entendiendo no solamente la información intrínseca de cada set de datos, sino también la relación existente entre los mismos, descubriendo detalles que no sean del todo evidentes mediante un método minucioso de análisis, comprobando cada hipótesis y dejando en claro los descubrimientos consiguientes para ver si efectivamente la hipótesis se cumplió según lo esperado o, si en cambio, los datos revelan un comportamiento anti-intuitivo, como suele suceder muchas veces durante este tipo de análisis.

Procesamiento y análisis

Información general sobre el análisis

Datos utilizados

La empresa proveyó los siguientes datasets:

- Datos de subastas RTB
- Instalaciones de aplicaciones
- Clicks de publicidades mostradas
- Eventos dentro de aplicaciones

Lenguaje y bibliotecas utilizadas

Para realizar el análisis exploratorio y sus correspondientes visualizaciones, se utilizó la web app *Jupyter Notebook*[1]. Algunos integrantes decidieron utilizarlo en su versión con *JupyterLab*[2].

Para la carga y manipulación de datos se utilizó la librería *Pandas*[3] (v. 0.24.1).

Para las visualizaciones se utilizaron las librerías *matplotlib*[4] y *seaborn*[5].

Para el cálculo de la matriz distancias se utilizó *scipy*[6].

Al contar con algunos datasets grandes, se consideró el uso de *Modin*[7], una librería que con cambiar una sola línea pasaría a utilizar todos los cores de la cpu obteniendo speedups muy significantes. Sin embargo, debido a que la herramienta presentó algunos comportamientos no esperados y que finalmente no hubo mayores problemas con el tamaño de los datasets, esta herramienta quedó descartada.

Repositorio de Github

Todo el desarrollo se encuentra subido a un repositorio de Github, donde se encuentran todos los análisis realizados necesarios para llevar a cabo la redacción del presente informe.
A continuación, el link a dicho repositorio: https://github.com/faristegui/tpdatos2019

[1] https://jupyter.org/
[2] https://jupyterlab.readthedocs.io/en/stable/
[3] https://pandas.pydata.org/
[4] https://matplotlib.org/
[5] https://seaborn.pydata.org/
[6] https://www.scipy.org/
[7] https://modin.readthedocs.io/en/latest/

Inspección de las columnas

A continuación se muestra un resumen del análisis previo de las columnas de los datos para lograr una eficiente lectura de los datos, motivado principalmente por el tamaño de los archivos Auction e Installs, de 1.4 Gb y 750 Mb respectivamente.

Auctions

Columna	Tipo por defecto	Análisis
auction_type_id	*float64*	Se puede omitir de la lectura, sus valores son nulos.
country	*int64*	Se puede omitir de la lectura ya que hay un único valor.
date	*object*	Se tendrá que convertir a datetime durante el análisis.
device_id	*int64*	Contiene pocos valores pero grandes, se opta por leerlo como category ya que ocupa menos espacio en memoria.
platform	*int64*	Contiene valores pequeños, se opta por leerlo como un int8.
ref_type_id	*int64*	Contiene valores pequeños, se opta por leerlo como un int8.
source_id	*int64*	Contiene valores pequeños, se opta por leerlo como un int8.

Events

Columna	Tipo por defecto	Análisis
date	*object*	Se tendrá que convertir a datetime durante el análisis.
event_id	*int64*	Contiene valores pequeños, se opta por leerlo como un int8.
ref_type	*int64*	Contiene valores pequeños, se opta por leerlo como un int8.
ref_hash	*int64*	Dejarlo como está.
application_id	*int64*	Contiene valores pequeños, se opta por leerlo como un int8.
attributed	*bool*	Dejarla como está.
device_countrycode	*int64*	Se puede omitir de la lectura ya que hay un único valor.

device_os_version	*float64*	Contiene pocos valores, se opta por leerlo como una category.
device_brand	*float64*	Se puede omitir de la lectura, sus valores son nulos.
device_model	*float64*	Contiene pocos valores, se opta por leerlo como una category.
device_city	*float64*	Contiene pocos valores, se opta por leerlo como una category.
session_user_agent	*float64*	Se puede omitir de la lectura ya que hay un único valor.
trans_id	*float64*	Se deja con propósitos de mergeo.
user_agent	*float64*	Se puede omitir de la lectura, sus valores son nulos.
event_uuid	*object*	Se puede omitir de la lectura, son todos valores únicos y en gran cantidad.
carrier	*float64*	Contiene pocos valores, se opta por leerlo como una category.
kind	*float64*	Contiene pocos valores, se opta por leerlo como una category.
device_os	*float64*	Se puede omitir de la lectura ya que hay un único valor.
wifi	*float64*	Se puede omitir de la lectura ya que hay un único valor.
connection_type	*object*	Contiene pocos valores, se opta por leerlo como una category.
ip_address	*int64*	Dejarlo como está.
device_language	*float64*	Contiene pocos valores, se opta por leerlo como una category.

Installs

Columna	**Tipo por defecto**	**Análisis**
created	*object*	Se tendrá que convertir a datetime durante el análisis.
application_id	*int64*	Contiene valores pequeños, se opta por leerlo como un int8.
ref_type	*int64*	Dejarla como está.

ref_hash	*int64*	Dejarla como está.
click_hash	*float64*	Se puede omitir de la lectura, sus valores son nulos.
attributed	*bool*	Se puede omitir de la lectura ya que hay un único valor.
implicit	*bool*	
device_countrycode	*int64*	Contiene pocos valores, se opta por leerlo como una category.
device_brand	*float64*	Contiene pocos valores, se opta por leerlo como una category.
device_model	*float64*	Dejarla como está.
session_user_agent	*object*	Contiene pocos valores, se opta por leerlo como una category.
user_agent	*object*	Contiene valores tipo string.
event_uuid	*object*	Si no existe la posibilidad de recuperarle datos, se puede omitir.
kind	*object*	Si no existe la posibilidad de recuperarle datos, se puede omitir.
wifi	*object*	Contiene valores tipo bool.
trans_id	*object*	Se deja con propósitos de mergeo.
ip_address	*int64*	Dejarla como está.
device_language	*float64*	Contiene pocos valores, se opta por leerlo como una category.

Clicks

Columna	**Tipo por defecto**	**Análisis**
advertiser_id	*int64*	Contiene pocos valores, se opta por leerlo como una category.
action_id	*float64*	Se puede omitir de la lectura, sus valores son nulos.
source_id	*int64*	Contiene pocos valores, se opta por leerlo como una category.
created	*object*	Se tendrá que convertir a datetime durante el análisis.

country_code	*int64*	Se puede omitir de la lectura ya que hay un único valor.
latitude	*float64*	Dejarla como está.
longitude	*float64*	Dejarla como está.
wifi_connection	*bool*	Se puede omitir de la lectura ya que hay un único valor.
carrier_id	*float64*	Contiene pocos valores, se opta por leerlo como una category.
trans_id	*object*	Se deja con propósitos de mergeo.
os_minor	*float64*	Contiene pocos valores, se opta por leerlo como una category.
agent_device	*float64*	Dejarla como está.
os_major	*float64*	Contiene pocos valores, se opta por leerlo como una category.
specs_brand	*int64*	Contiene pocos valores, se opta por leerlo como una category.
brand	*float64*	Contiene pocos valores, se opta por leerlo como una category.
timeToClick	*float64*	No hace falta tanta precisión, se opta por leerlo como un float16.
touchX	*float64*	No hace falta tanta precisión, se opta por leerlo como un float16.
touchY	*float64*	No hace falta tanta precisión, se opta por leerlo como un float16.
ref_type	*int64*	Contiene pocos valores, se opta por leerlo como una category.
ref_hash	*int64*	Contiene valores pequeños, se opta por leerlo como un int8.

Algunas observaciones a priori

Llamó la atención que no haya ningún valor en el campo *auction_id* de Auctions. No tanto por el valor que agregaría al análisis dicho campo, sino por la presencia del mismo. Se maneja la hipótesis de que en principio se iba a incluir, y que en la versión entregada al curso finalmente se omitió.

En cuanto al campo correspondiente al país, llamó la atención que 3 de los 4 datasets contengan solamente un valor, repetido en más de la mitad de los registros en el dataset restante. No parecería ser un campo que aporte mucha información a priori.

Similarmente a lo ocurrido en Auctions con *auction_id*, en el dataset Events se detectan todos valores nulos en el campo de *device_brand*. Ídem en *click_hash* de Installs y *action_id* de Clicks.

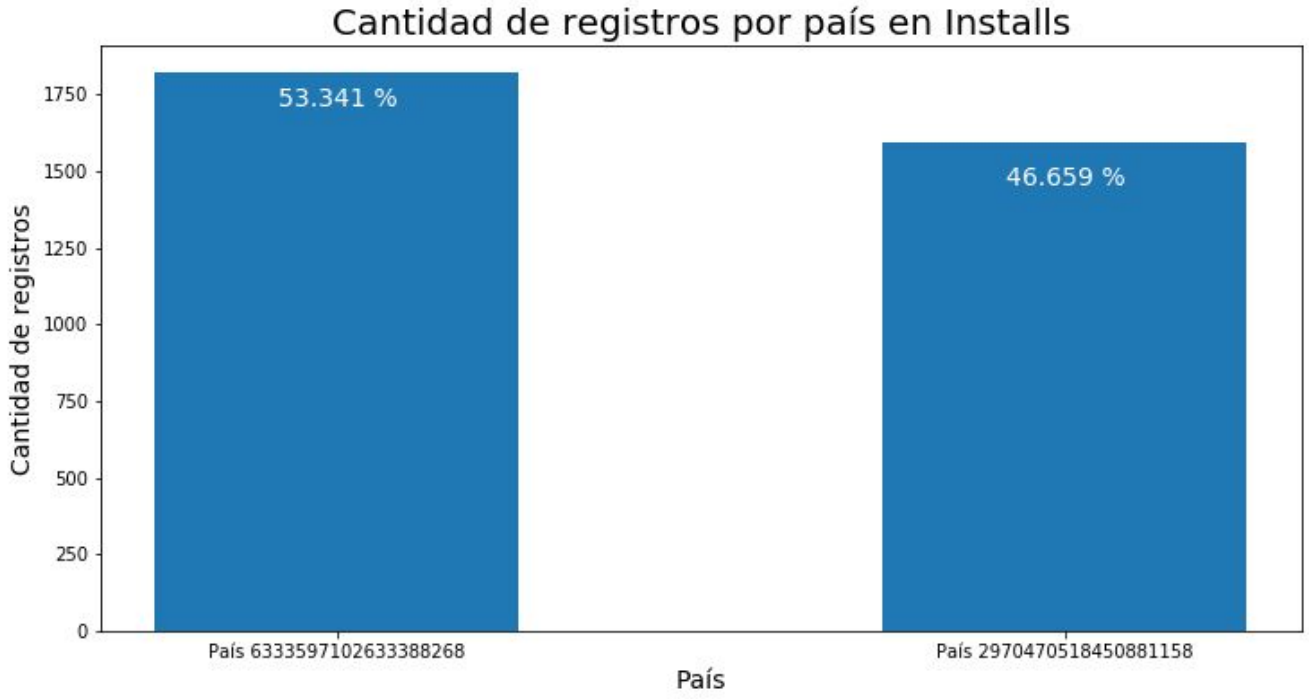
Lo que más poderosamente llamó la atención es que no existan registros con instalaciones atribuidas a Jampp.

Inspección de datos anonimizados

Lo primero que se observó en los datasets es que, tal como ya había sido anunciado, los datos fueron anonimizados. Si bien en principio esto no pareció ser de importancia, cambió un poco la forma de tratar los datos. Se listan a continuación los datos que fueron transformados durante dicho proceso, comentando en algunos casos qué impacto tuvieron durante el análisis.

Auctions

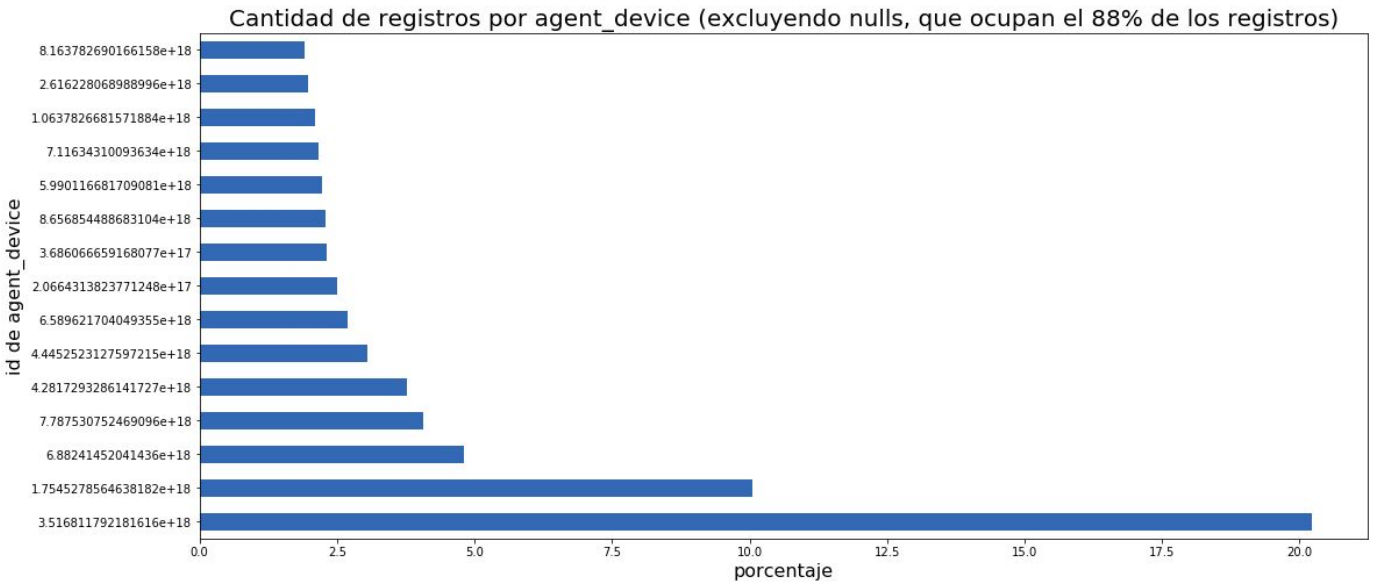
- country: si bien este dato pasó por *string hashing*, en el dataset toma un solo valor. Lo mismo pasa con los campos *country_code* y *device_countrycode* de Clicks y Events respectivamente. En los tres casos el valor existente es el mismo. No es así en el dataset Installs, donde en el campo *device_countrycode* aparecen dos valores con una distribución muy pareja, como se muestra en el gráfico a continuación. El valor de country de los otros tres datasets se corresponde con el valor mayoritario.



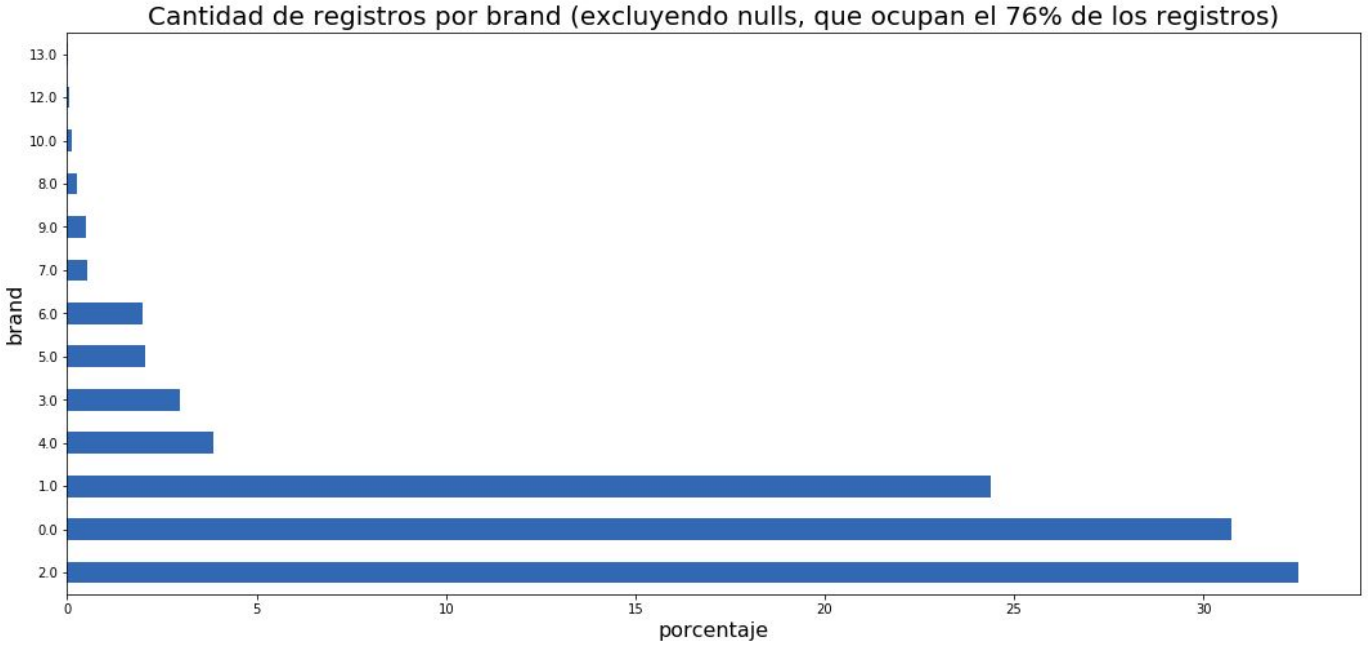
- device_id: como los ids no tienen información intrínseca, el hecho de que hayan sido hasheados no modificó sustancialmente el análisis.
- source_id:

Clicks

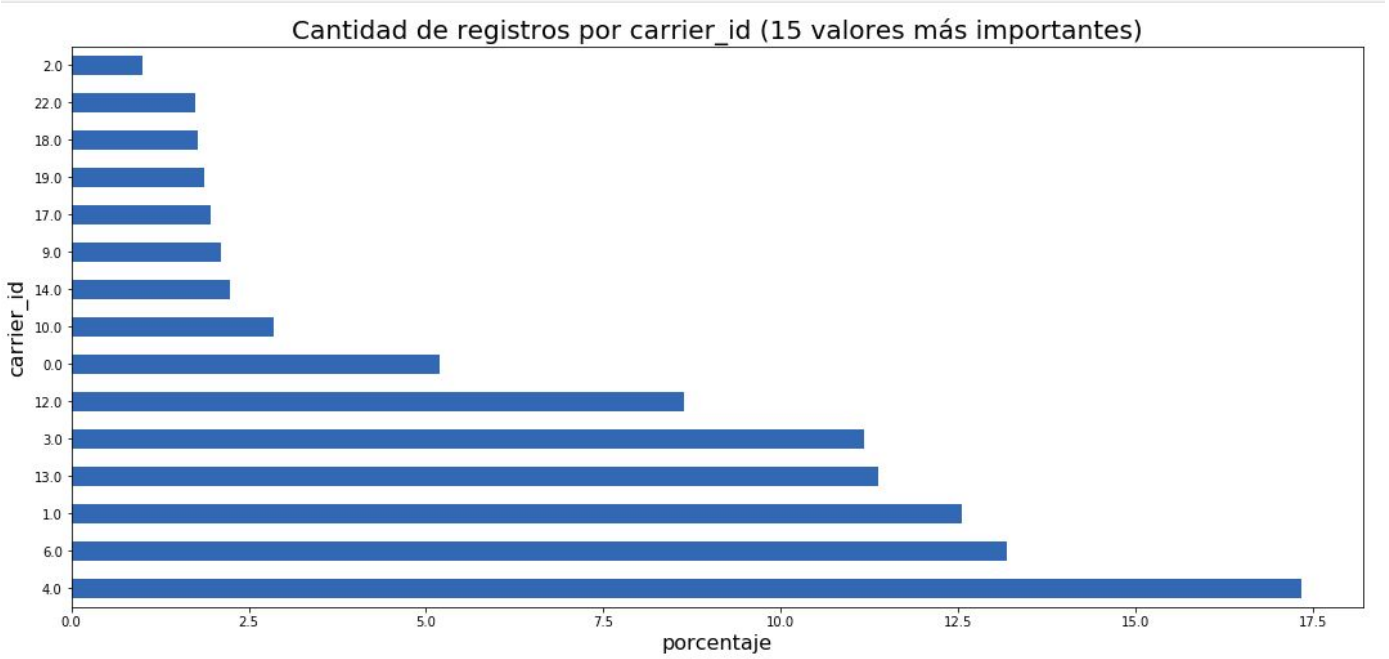
- advertiser_id: ídem caso anterior de id. En este caso se vio que no hay ningún campo nulo, y que se cuenta con un total de 7 ids únicos.
- agent_device: el encoding propiamente dicho no molesta en el análisis. Lo que sí se puede notar es que el 88% de los registros contiene valores nulos, y el resto sigue una ley de potencias, según se puede observar en el siguiente gráfico que muestra los 15 valores más frecuentes (de los 190 valores únicos que hay).



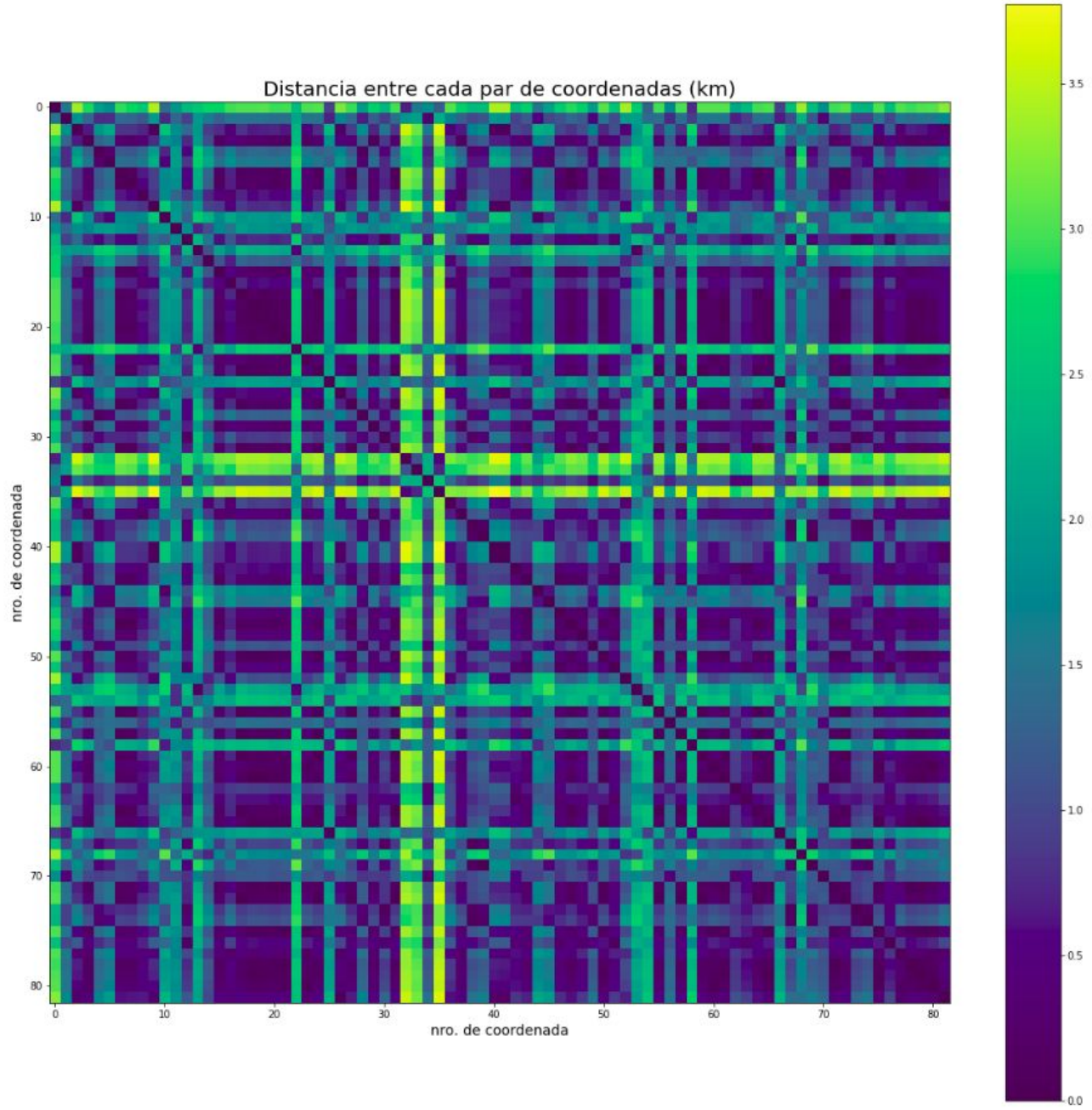
- brand: En un principio, desconocer explícitamente la marca de un dispositivo no debería traer problemas, aunque sí serviría percibir las marcas reales para poder poner un foco mayor en ellas, ya que aquellos usuarios que poseen esta marca serán los que más interactúen (por ejemplo la 1.0, 0.0 y 2.0 en el gráfico). Nuevamente se observa una gran cantidad de nulos, y luego una distribución similar al caso anterior.



- carrier_id: en cuanto a la anonimización, sucede como en el caso anterior, que no trae mayores problemas. La diferencia reside en la cantidad de campos nulos, que en este caso está por debajo del 0.1%, por lo que se termina obteniendo un panorama mucho más real.

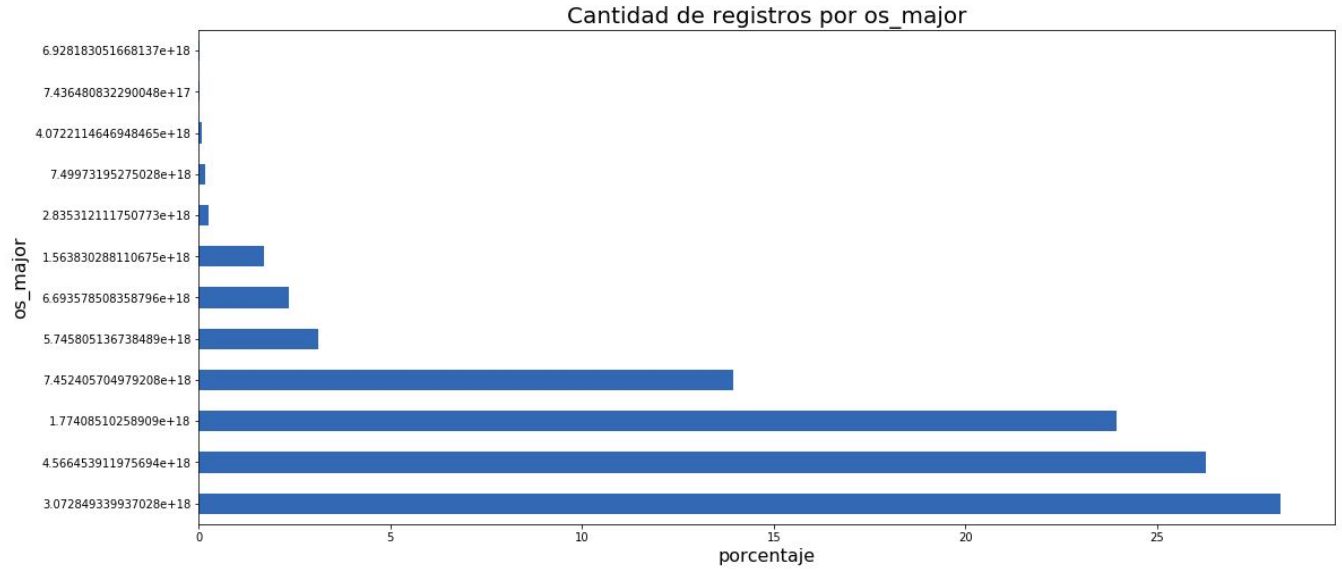


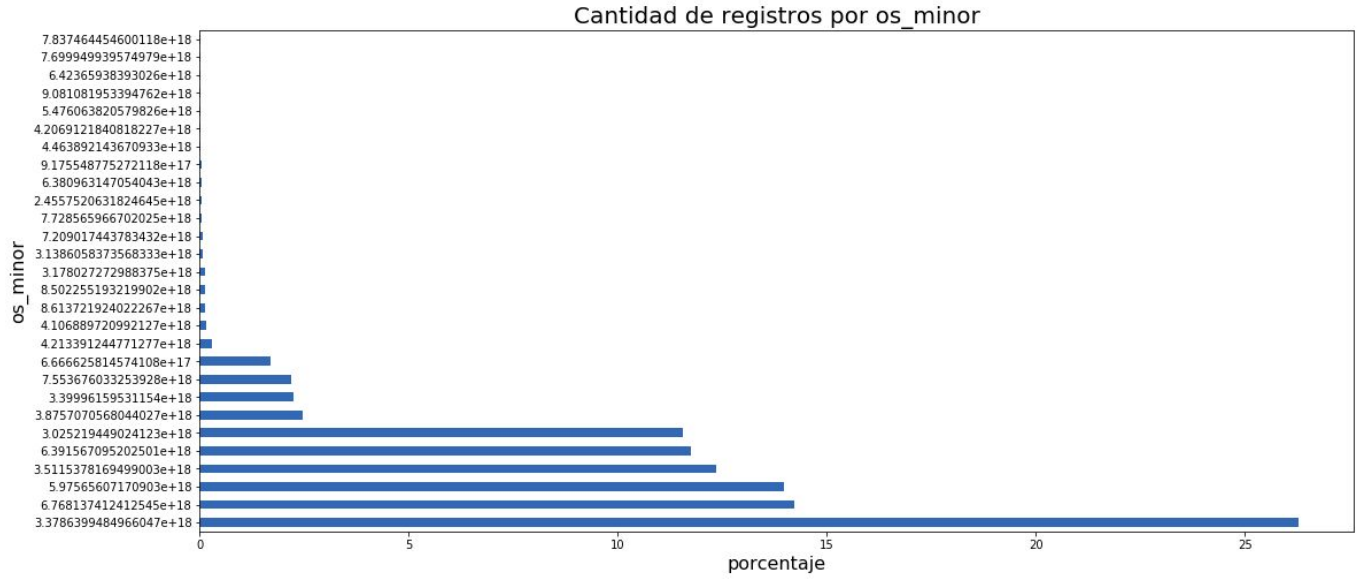
- latitud - longitud: la idea principal fue mostrar estos datos en un mapa, pero como se realizó una transformación lineal sobre estos datos, esto ya no será posible. Sin embargo, al ser la transformación lineal, se pudo llevar a cabo un análisis de distancias. A continuación se muestra un heatmap mostrando las distancias entre cada par de las 82 coordenadas únicas presentes en el dataset. Notar que en general las distancias son muy chicas, dándose las máximas entre un par muy reducido de coordenadas y con valores que rodean los 3.5 km.



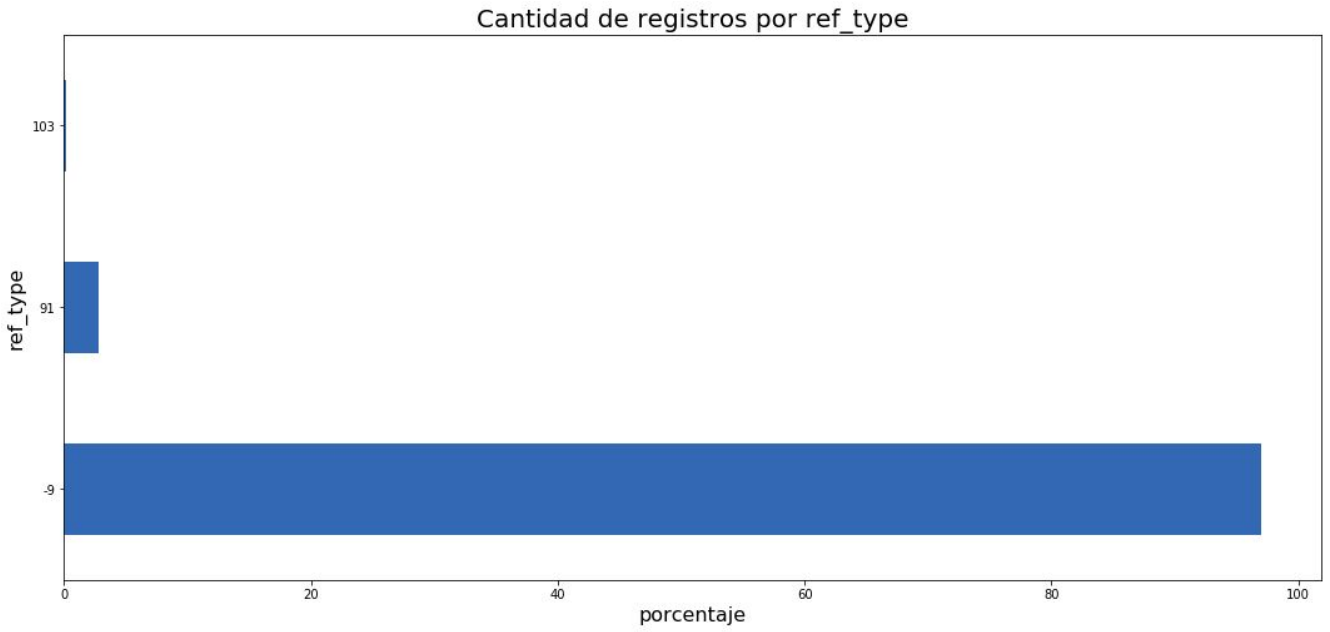
- os_major y os_minor: prácticamente no hay valores nulos. En os_major, si bien hay 12 valores únicos, 3 de ellos concentran el 75% de los registros; si se toman los 4 valores con mayor cantidad de registros ya se cubre el 90% de los registros.

  En os_minor hay 28 valores posibles, entre los cuales hay un valor que abarca el 26% de los registros no nulos; le siguen 5 valores que abarcan una cantidad aproximadamente pareja de registros; entre estos 6 valores ya se abarca el 90% del dataset.

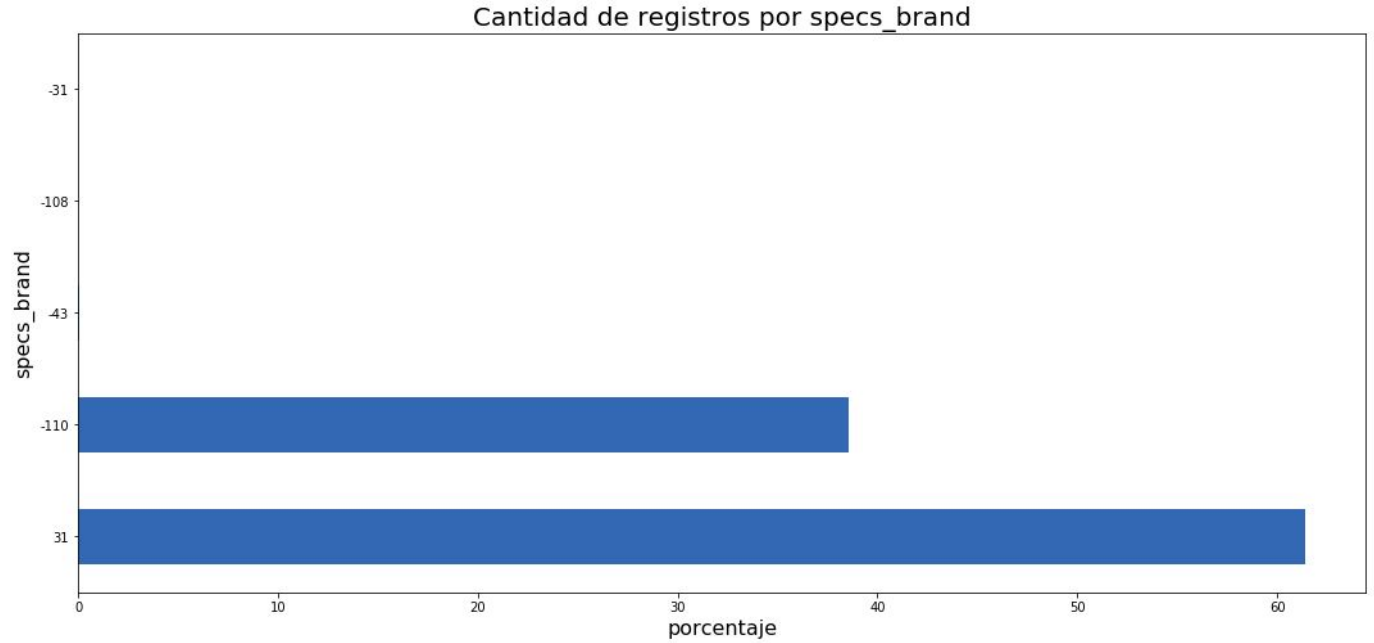




- ref_hash: este registro no importa que esté hasheado, ya que el campo en sí no tiene valor intrínseco. El propósito de este campo es poder cruzar datos entre los distintos datasets.
- ref_type: casi todos los registros contienen el mismo valor (no hay ninguno con valor nulo).

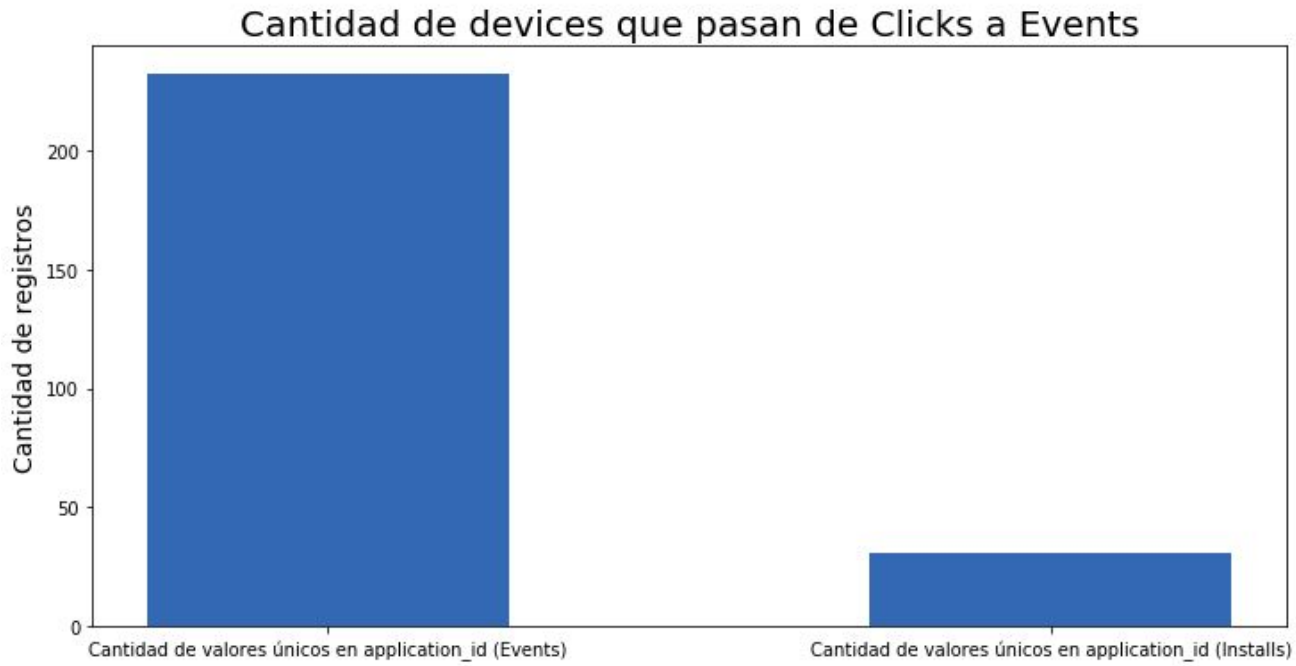


- specs_brand: no se llega a deducir qué es, y tampoco se proveyó una descripción del campo. Este campo toma solamente 5 valores posibles, pero casi la totalidad de los registros está dividida entre dos valores posibles.



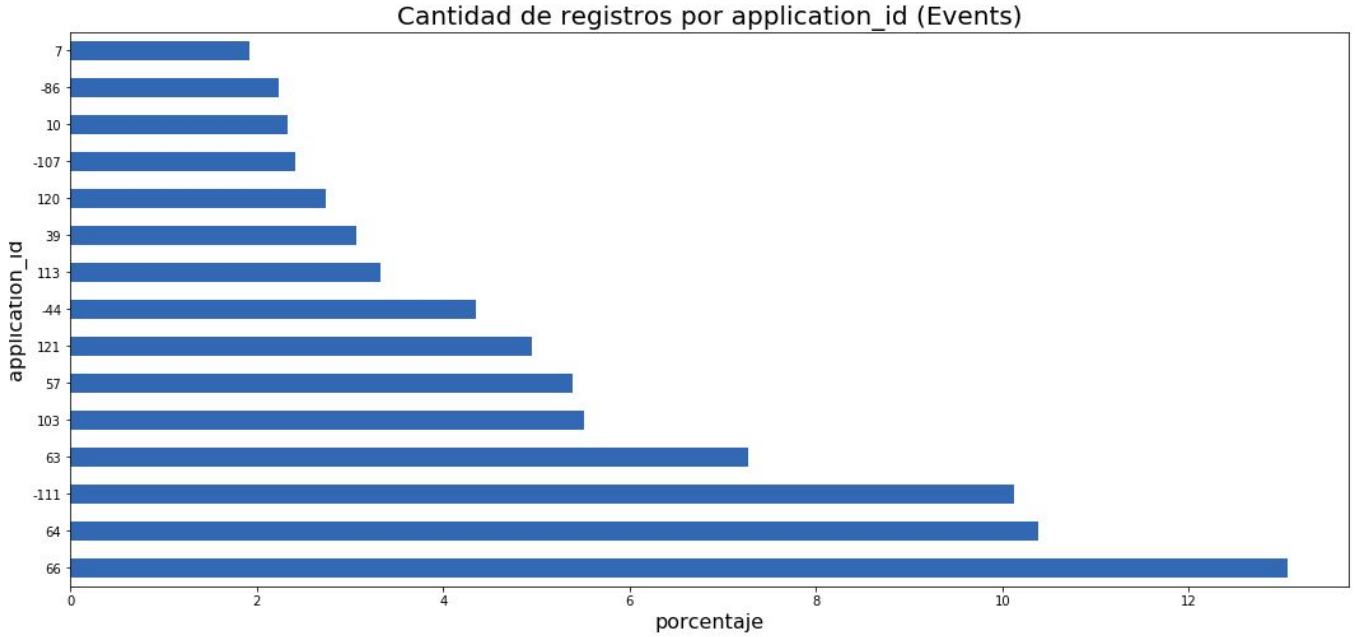
Events

- application id: registro compartido con Installs. Al no tener valor intrínseco, no cambia en nada la anonimización.
  En ambos datasets la cantidad de valores únicos presentes varía muchísimo.

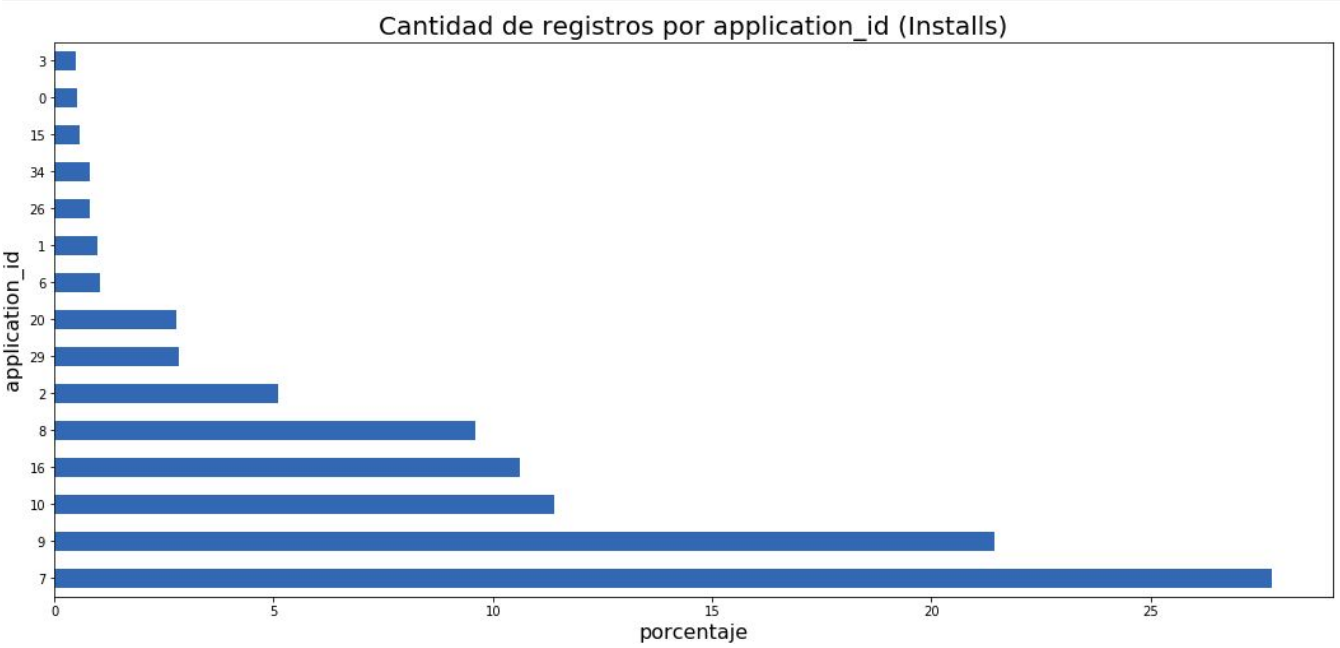


Por esta razón se dificulta bastante hacer una comparación a priori de ambos campos.

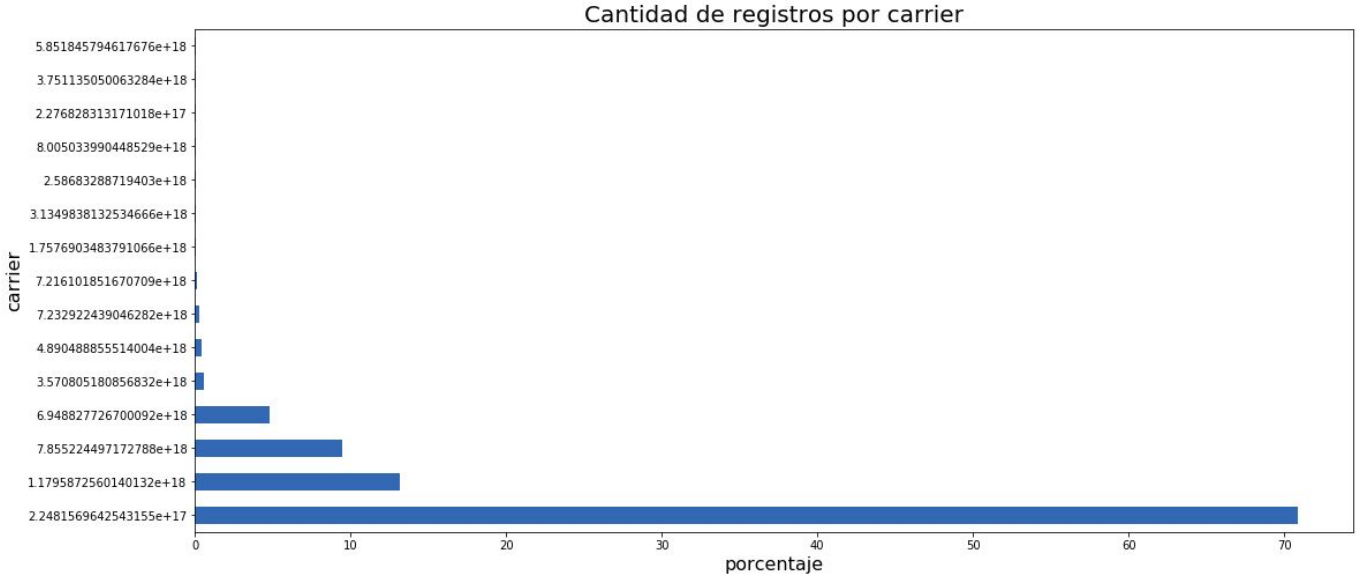
En el caso de Events hay 3 ids que toman por encima del 10% de los registros de Events cada uno.



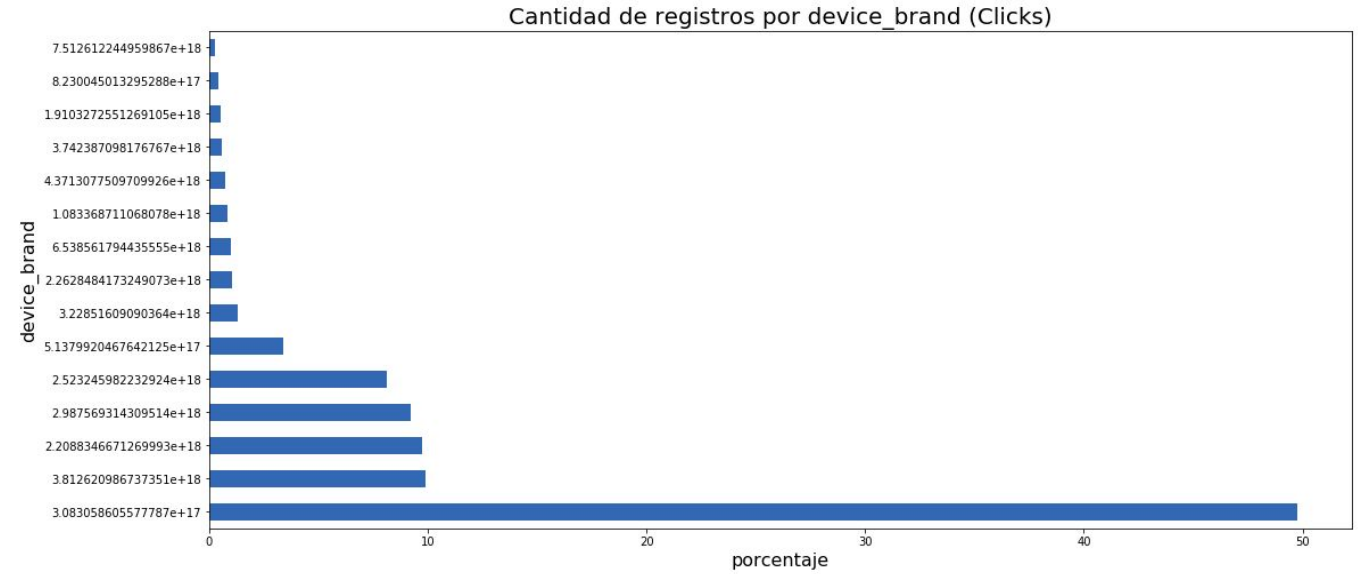
En Installs termina sucediendo algo similar: hay 4 valores que abarcan más del 10% de los valores no nulos cada uno.

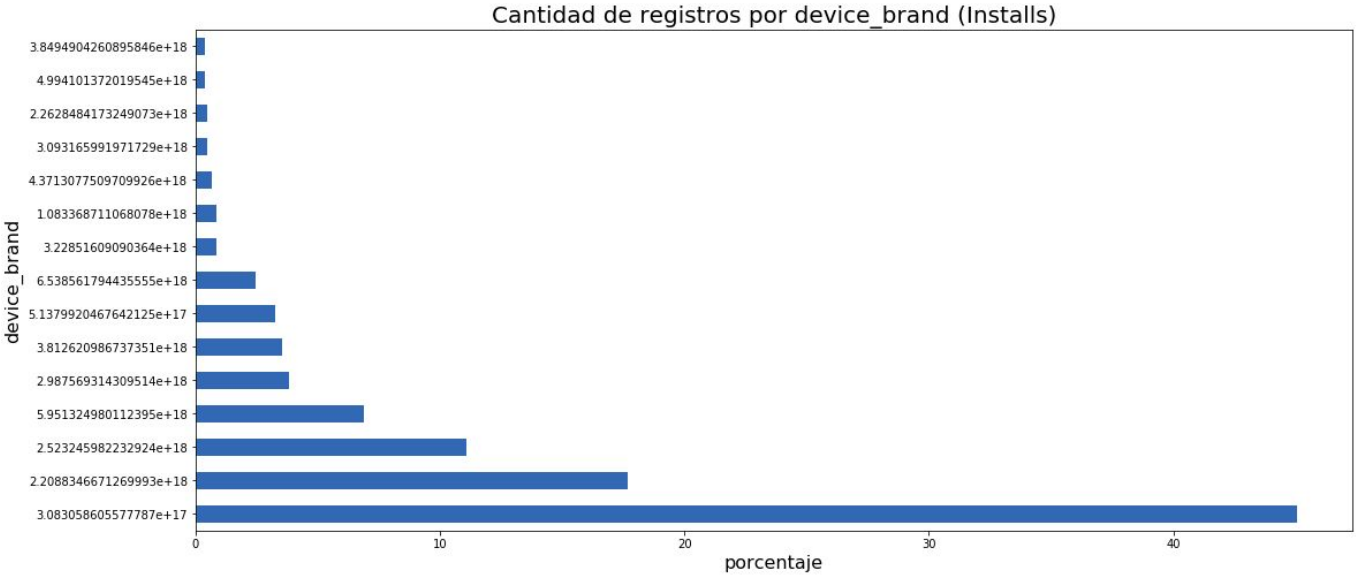


- carrier: se llegó a pensar que el hash sería compartido con el carrier_id de Clicks, sin embargo los valores no tienen nada que ver. En este caso, del 25% de registros que toman valores no nulos en este campo, el 75% toma un mismo valor (de 85 posibles).

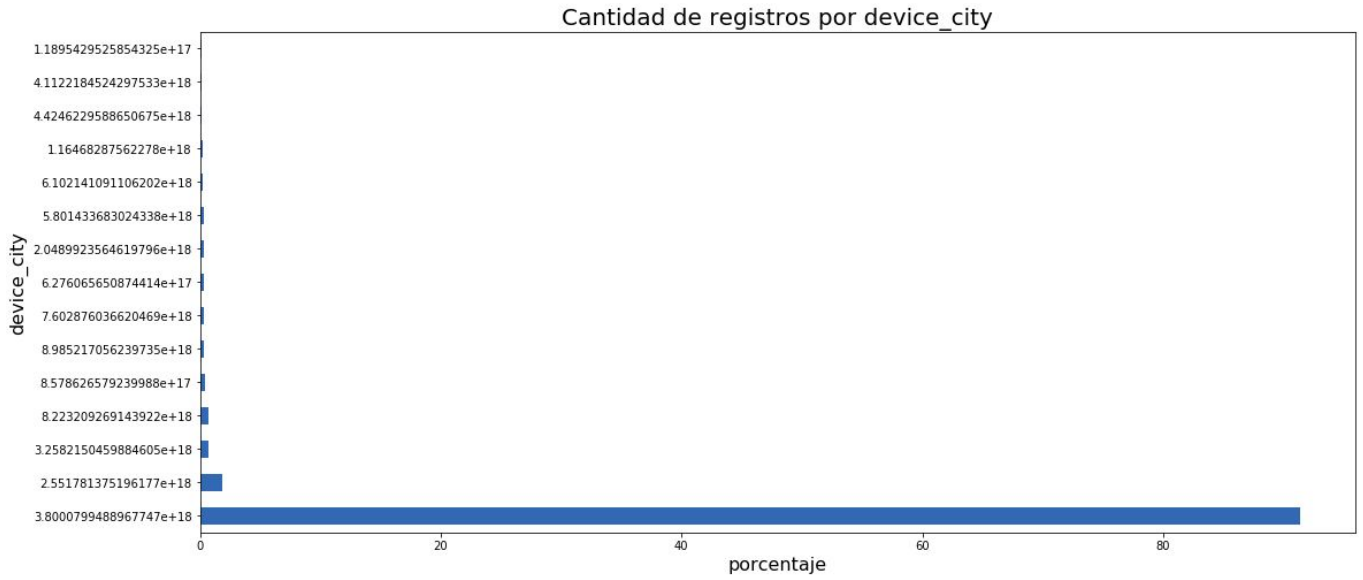


- device_brand: nuevamente se pensaría que habría valores en común con el dataset Clicks. Sin embargo tampoco tienen nada que ver, a pesar de que en ambos casos se trata de la marca del dispositivo. La distribución difiere del caso anterior también: en este caso, del 50% de campos no nulos, el 50% tiene un mismo valor, mientras que el resto de los valores ocupan cada uno menos del 10% del 50% restante.
  Hay que notar que sin embargo, este campo es compartido con Installs. En ambos casos se obtiene un distribución similar, aunque en un orden levemente cambiado, a excepción del caso mayoritario que en ambos casos ronda el 50% de los registros no nulos.

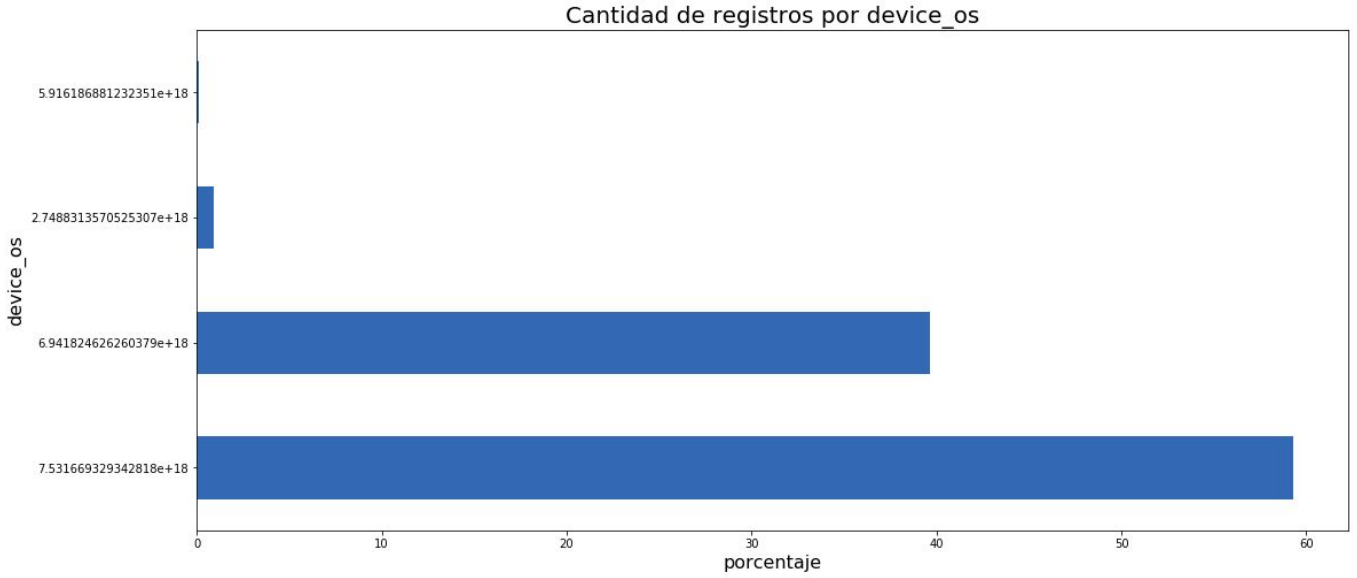




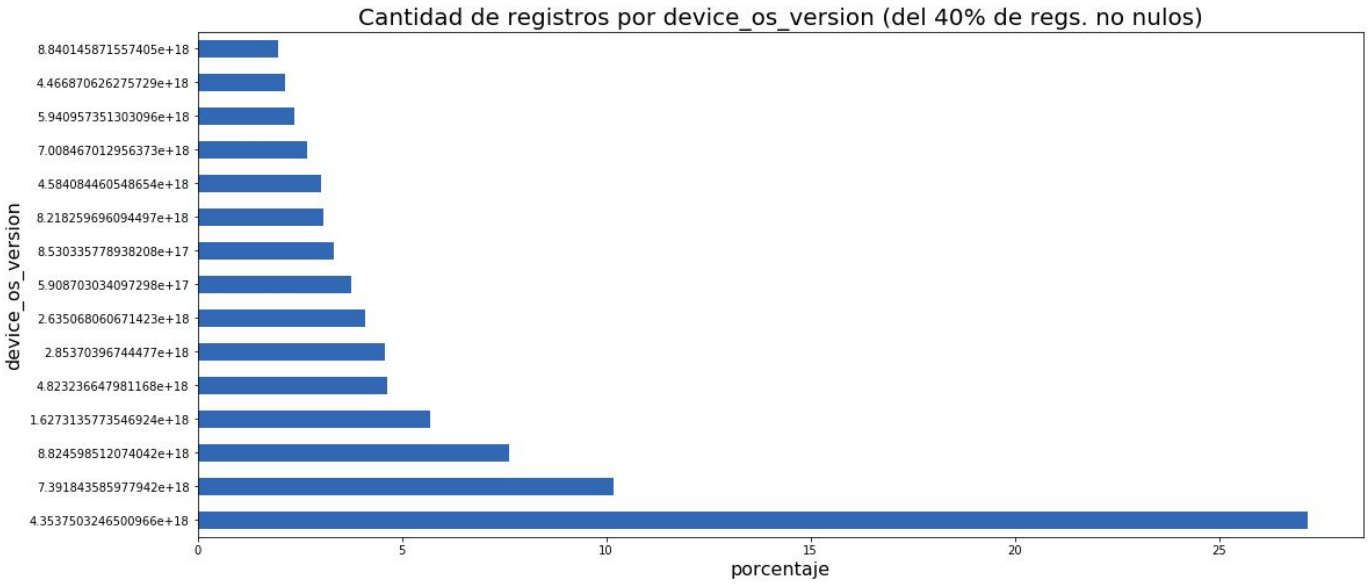
- device_city: al igual que pasó con las coordenadas, sería bueno poder hacer una representación visual de la distribución de ciudades. Así, lo que se puede rescatar es que del 25% de campos no nulos, más del 90% llevan el mismo valor.



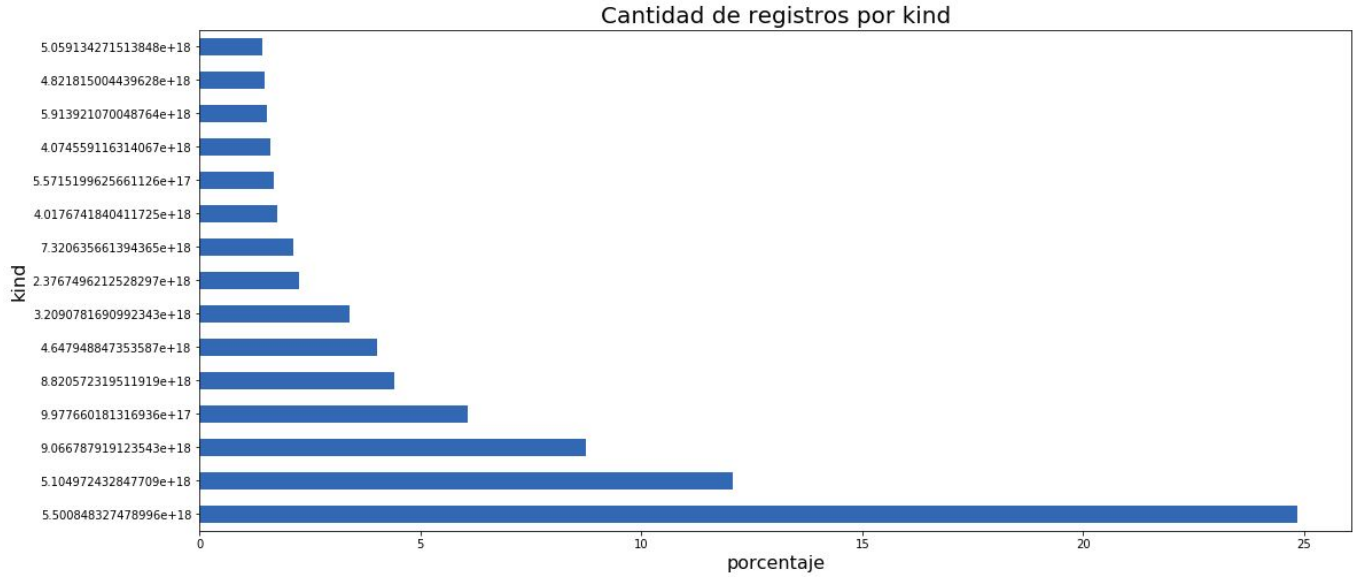
- device_os: casi el 75% de los registros no llevan valor. De los restantes, se supuso en principio habrá una gran cantidad con Android, una menor cantidad con iOS y una cantidad todavía menor con Windows Phone. Si bien no hay forma de verificar que efectivamente se trate de esos valores con las respectivas cantidades, el plot obtenido refleja lo esperado, agregando un 4to os.



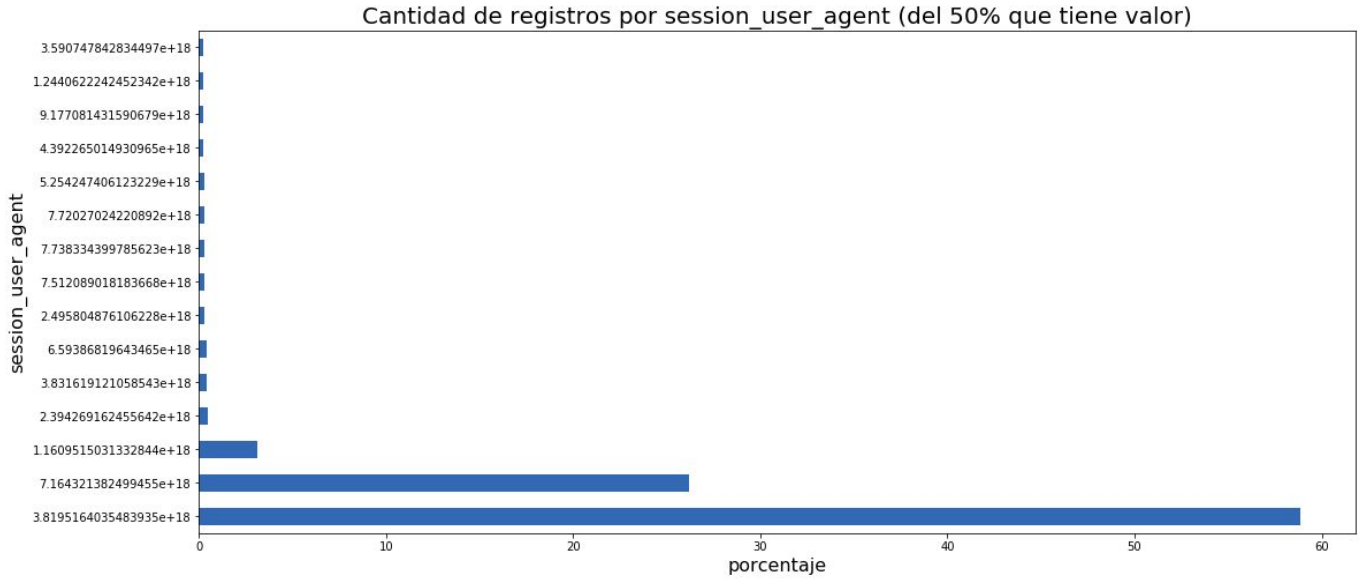
- device_os_version: la distribución sigue una ley de potencias.



- event_id: como el campo no contiene información intrínseca y además se cuenta con un identificador único de vento (event_uuid) que además es compartido entre datasets y no ha sufrido transformaciones, se decidió ignorar este campo.
- ip_address: es un dato muy ligado a la ubicación.
- kind: nuevamente la información sigue una ley de potencias. Casi no hay valores nulos, y el valor más abarcativo abarca el 25% de los registros.



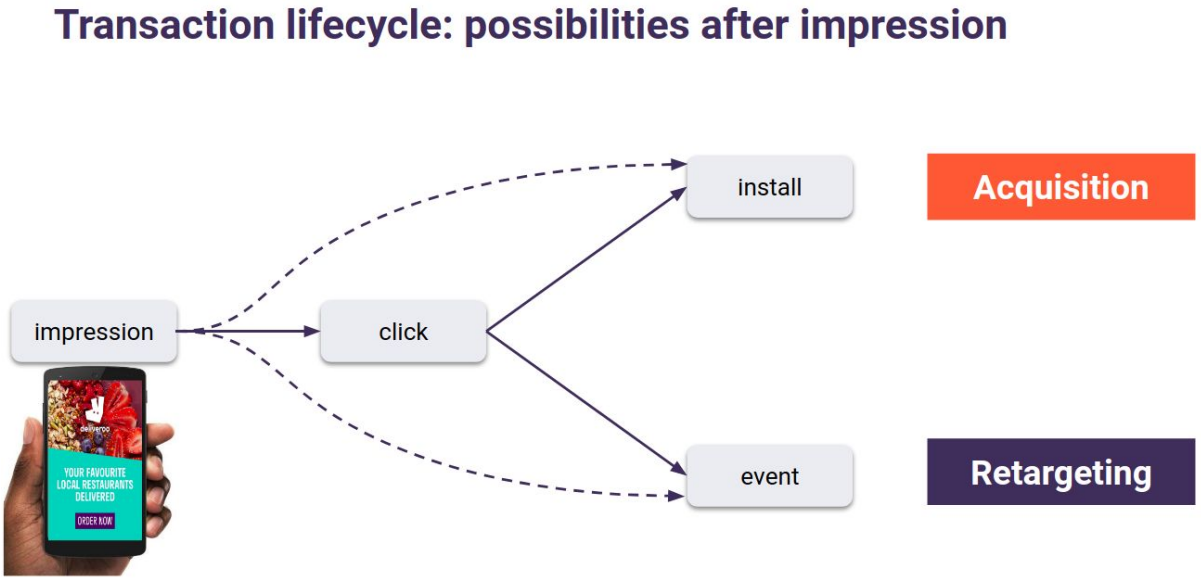
- session_user_agent: aproximadamente la mitad de los registros no tiene información. De lo que resta, toman uno de los 1460 valores posibles. Sin embargo, hay un valor que abarca el (casi) 60% de los registros no nulos, y otro que toma otro (casi) 30% de los registros no nulos.



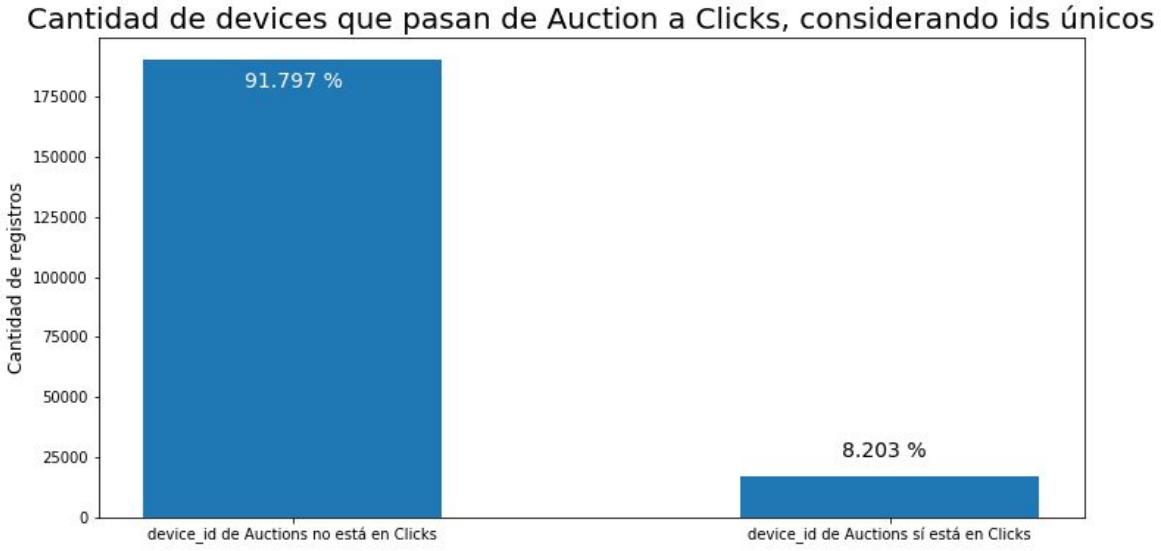
- user_agent: si bien también hay una completitud del 50% aprox., esta vez los campos toman uno de 5000 valores, y la distribución es mucho más pareja que del campo anterior. No hay ningún valor presente en más del 3% de los registros no nulos.

Análisis del flujo de datos

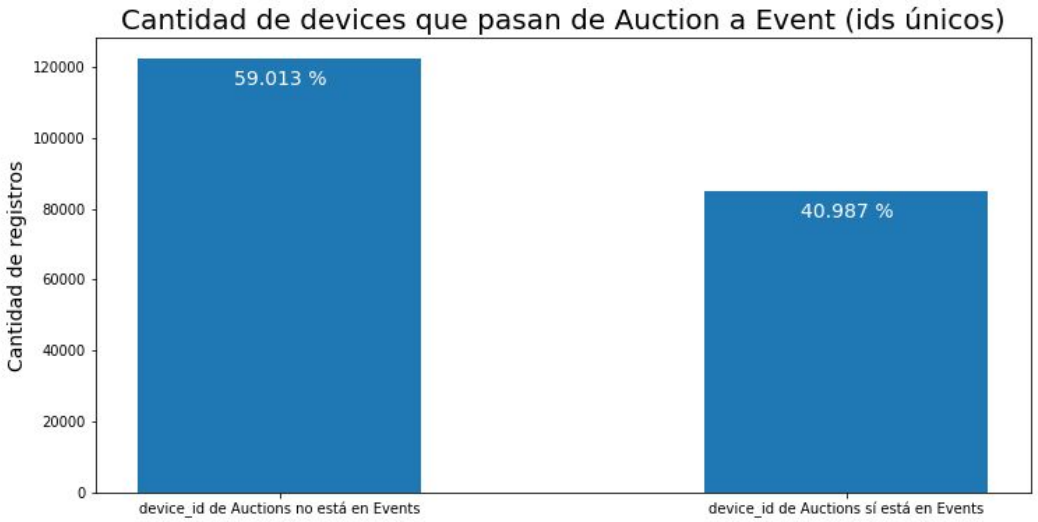
En la imagen a continuación se puede apreciar el flujo de datos según la presentación oficial de Jampp.



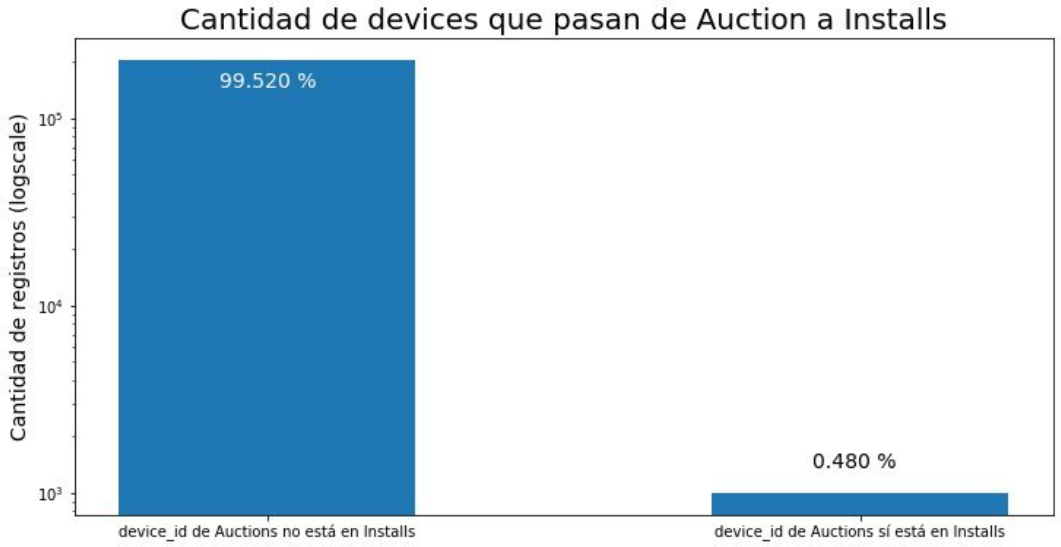
En primer lugar se observa la cantidad de dispositivos que pasaron de tener una impression (visto en el dataset Auctions bajo la columna device_id) a un click (visto en el dataset Clicks bajo la columna ref_hash). En el gráfico siguiente se puede ver la cantidad y el porcentaje respectivo de los ids en Auctions que se encuentran también en Clicks. Como se supuso, pasa un menor porcentaje. De todas formas se esperaba una diferencia muchísimo más marcada.



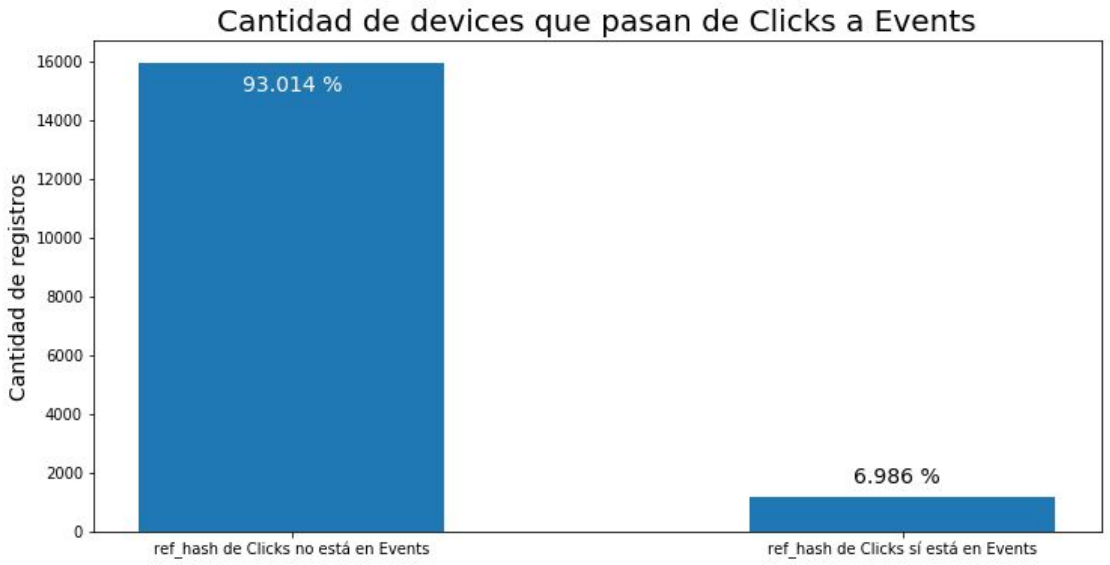
Como era de esperar, hay muchos más datos que pasan de Auctions a Events, considerando que un click es *un posible* evento.



Finalmente, de los dispositivos presentes en Auctions, no pasa casi ninguno a Installs, según se esperaba.



Devices de Clicks a Events:

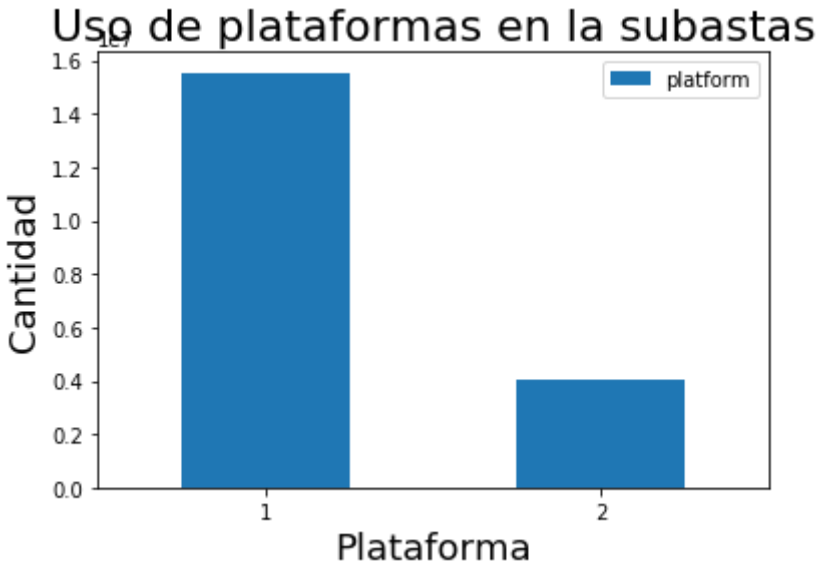


Devices que pasan de Clicks a Installs: no se grafica debido a que de 17119 valores únicos, pasan solamente 7, lo cual no se llega a apreciar en ningún gráfico.

De los más de 2 millones valores de uuid presentes en Events, solamente 860 se vuelven a repetir en Installs.

Análisis de Auctions

¿Cuáles son las plataformas que participan de las subastas (android / ios)? y ¿Cual es la plataforma con mayor participación en las subastas?

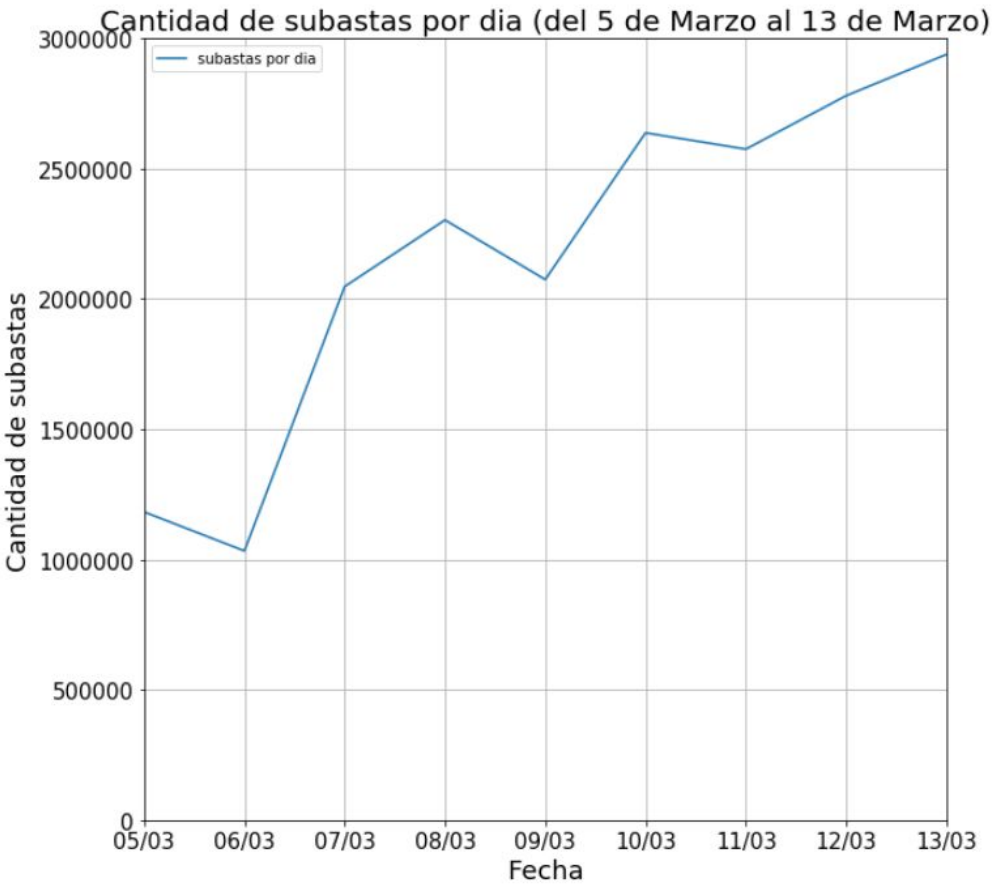


Vemos cuáles son las plataformas que participan de las subastas. Recordando que los datos se encuentran anonimizados (únicamente sabemos que existen ios y android) observamos dos plataformas participantes (1 y 2 respectivamente) siendo la 1 la que posee una mayor participación.

¿De qué fecha obtenemos la información? y ¿Cuántas subastas por día hay en el periodo?

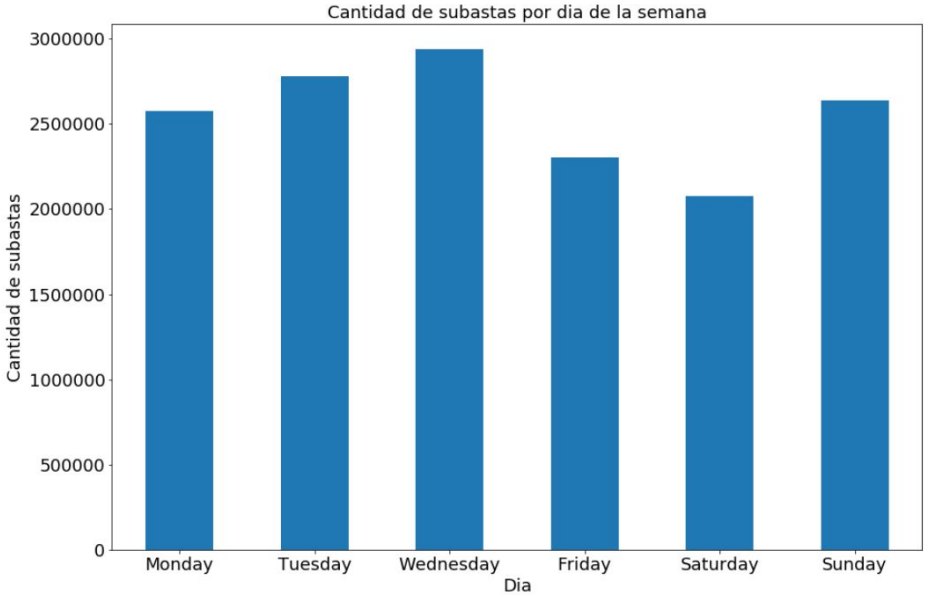
Observamos que tenemos información del 3 al 13 de marzo.

		total
dayofyear	**date_single**	
64	**2019-03-05**	1182401
65	**2019-03-06**	1032970
66	**2019-03-07**	2047661
67	**2019-03-08**	2303002
68	**2019-03-09**	2074552
69	**2019-03-10**	2637534
70	**2019-03-11**	2574916
71	**2019-03-12**	2779910
72	**2019-03-13**	2938373

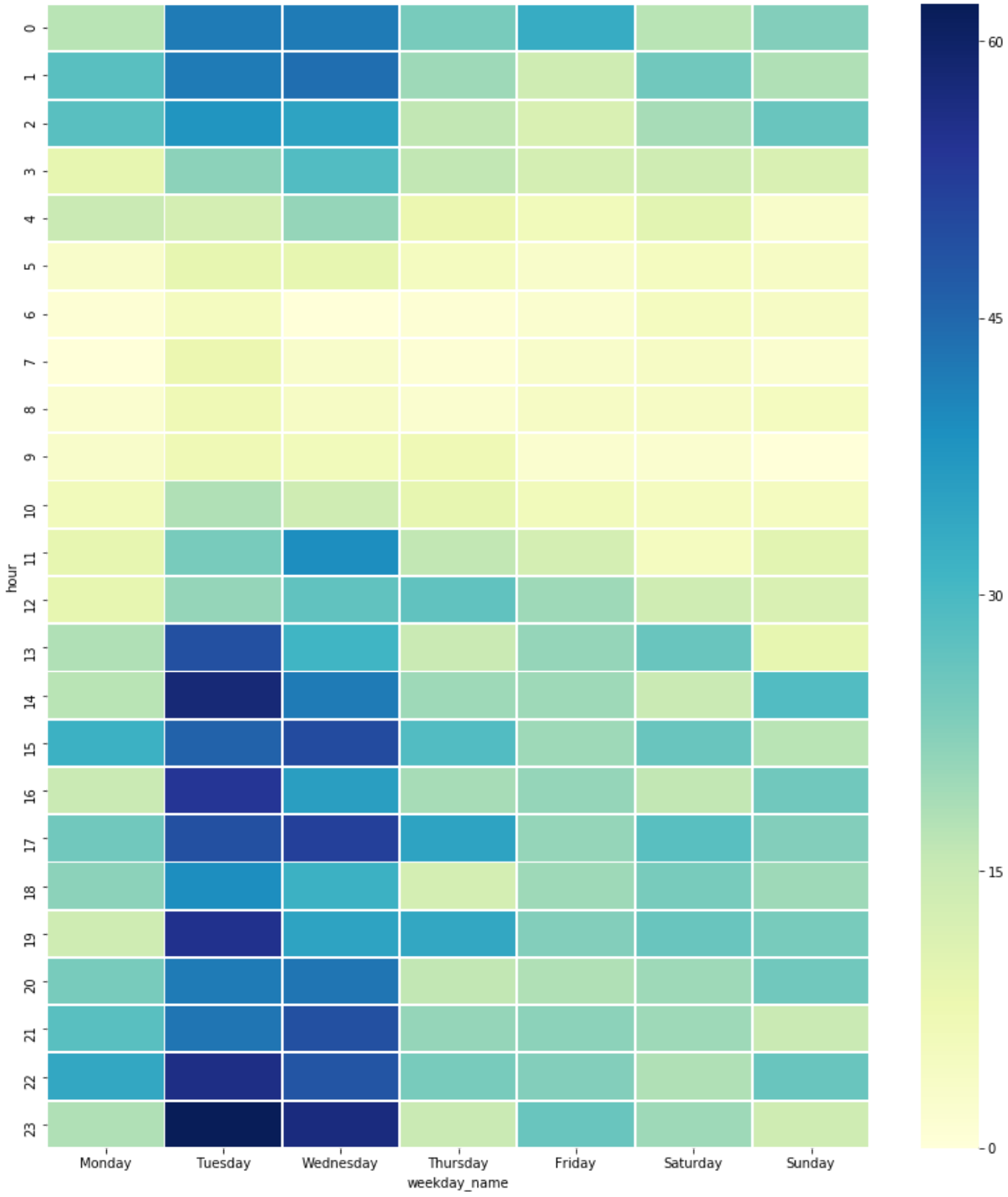


Observamos que se proyecta sobre la cantidad de subastas una tendencia alcista. Pero el periodo de fechas sobre el cual se trabaja es de una semana para decidir si esta tendencia permanecerá a lo largo del tiempo o sufrirá alguna fluctuación.

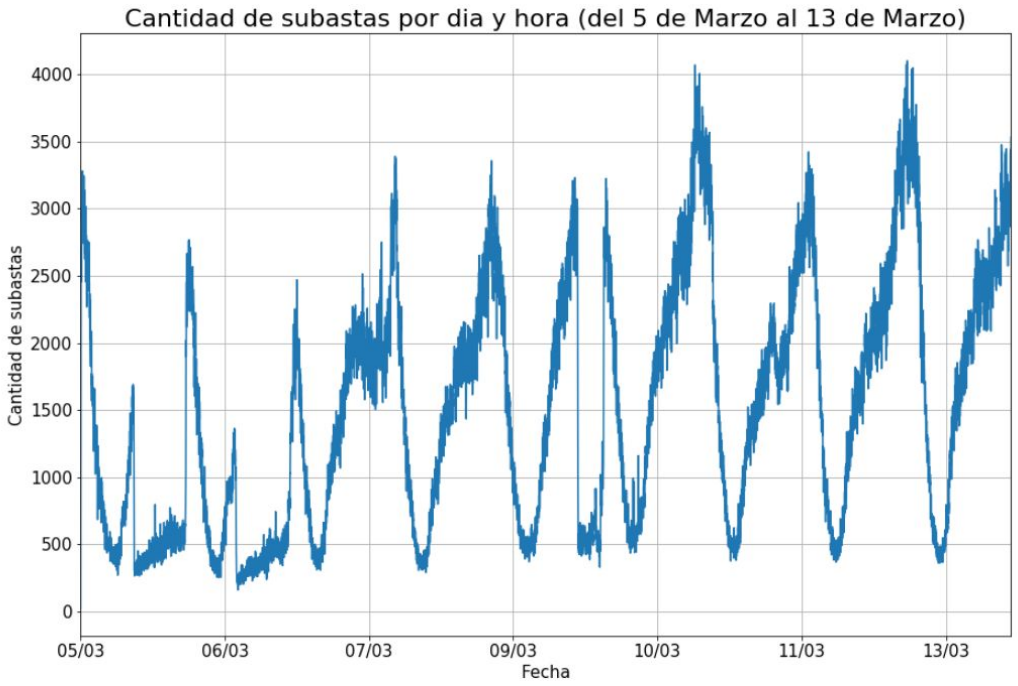
¿Cúal es el día que se producen más subastas en 1 semana?



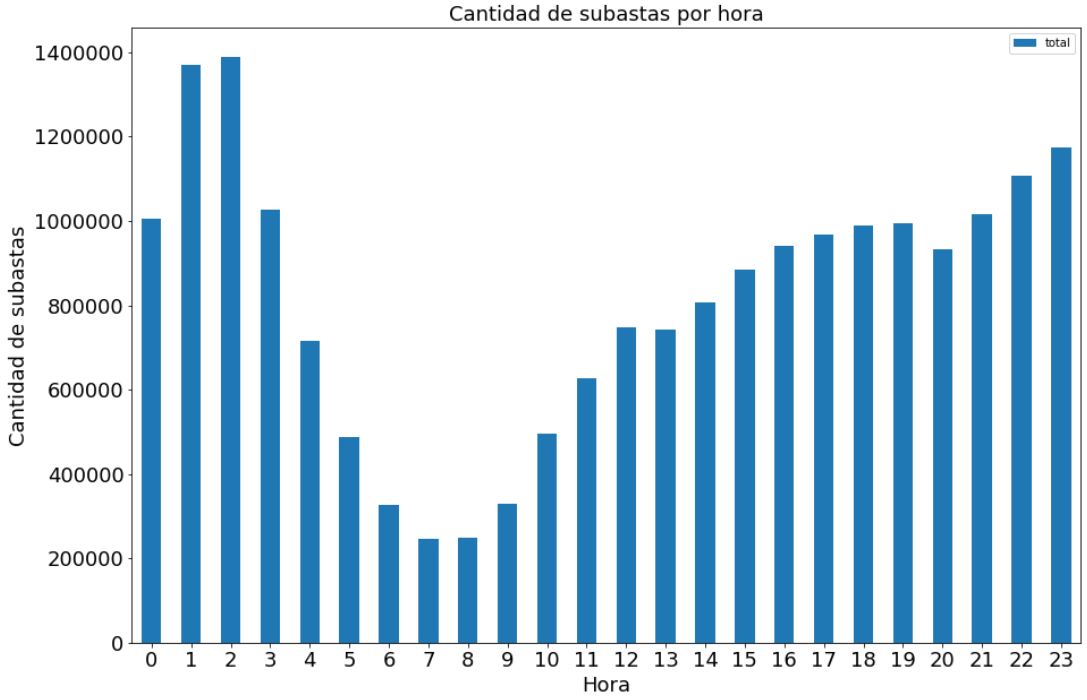
Heatmap por dia y cantidad de participaciones en subastas



¿Cual es la cantidad de subastas por hora y por día?

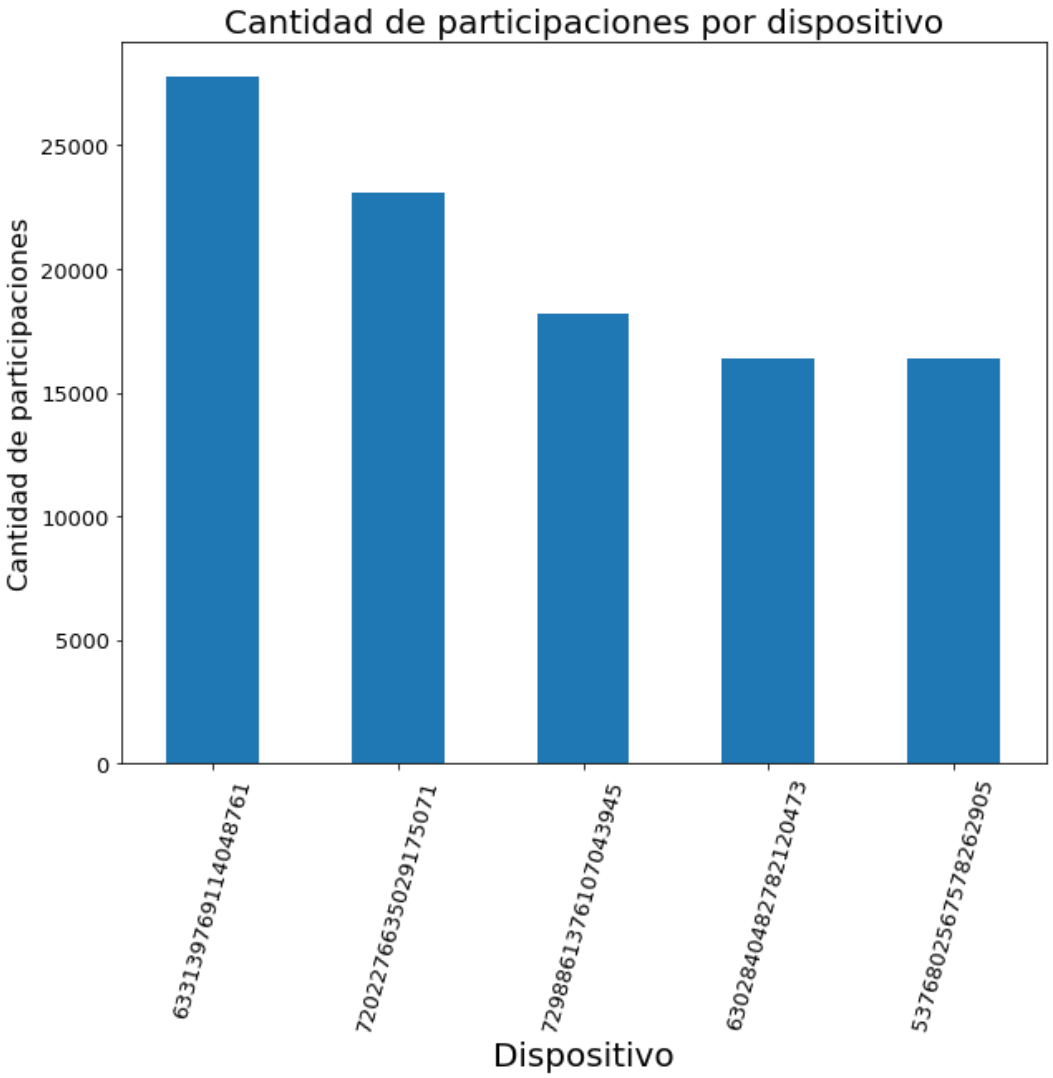


¿A qué hora se producen la mayor y la menor cantidad de subastas?

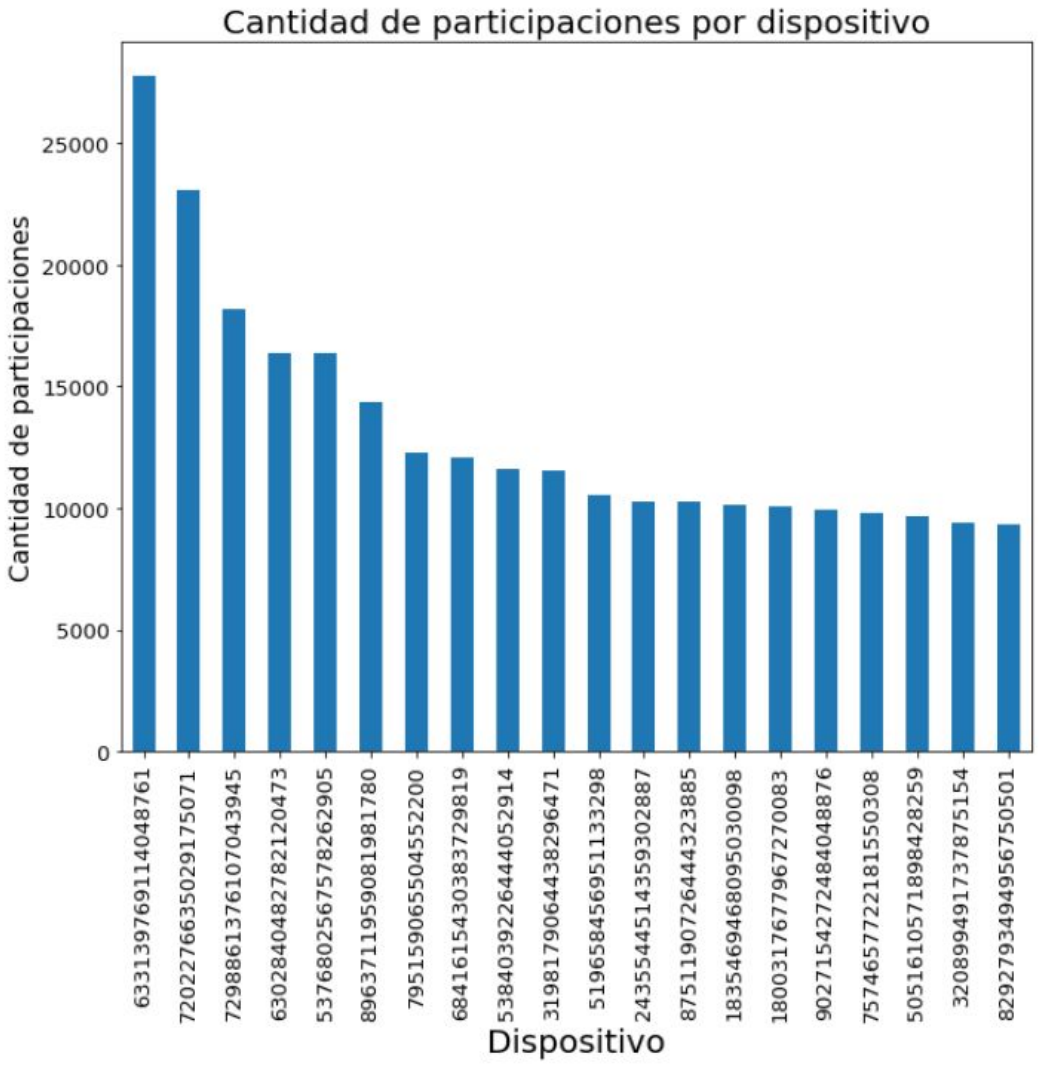


Observamos que las horas donde se encontraron mayor cantidad de subastas son entre la 1 y las 2 de la mañana y el periodo entre las 6 las 9 de la mañana donde se produjo la menor cantidad de subastas. Como observación vemos que desde la mañana a medida que suceden las horas aumenta la cantidad de subastas, llegando al pico máximo al inicio de la madrugada y luego decayendo hasta las primeras horas de la mañana.

¿Cuales fueron los 5 dispositivos que tuvieron mayores participaciones en las subastas?

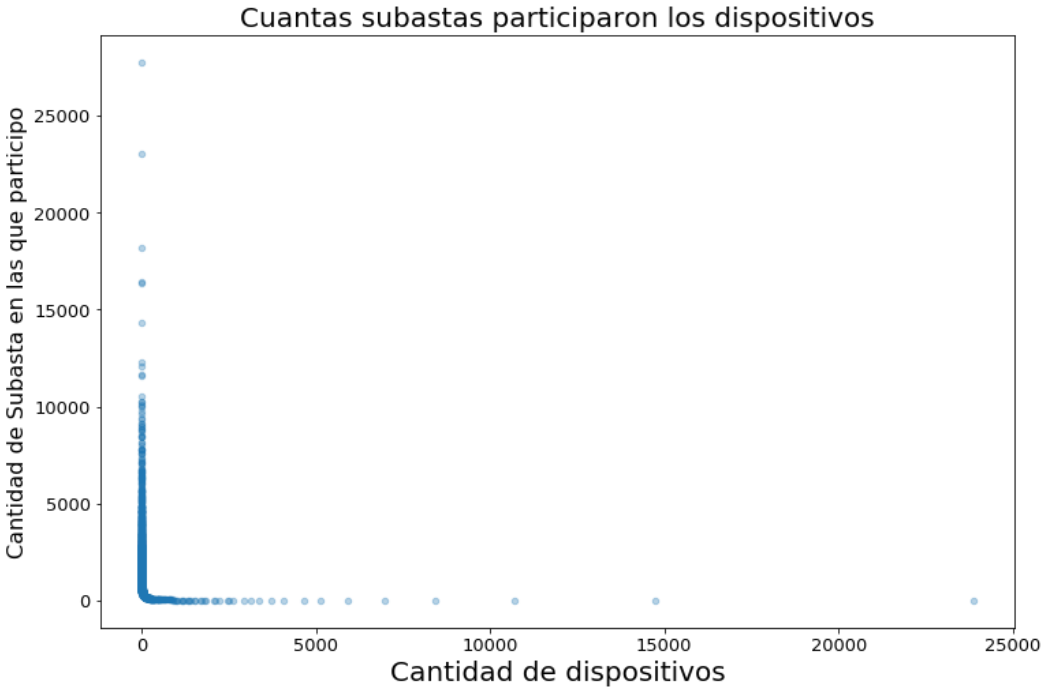


Aquí se observan los dispositivos que tuvieron mayor participación en la subastas. Estos números de participación son realmente muy altos. Estos nos podría indicar algún tipo de fraude o granja de instalación de apps.

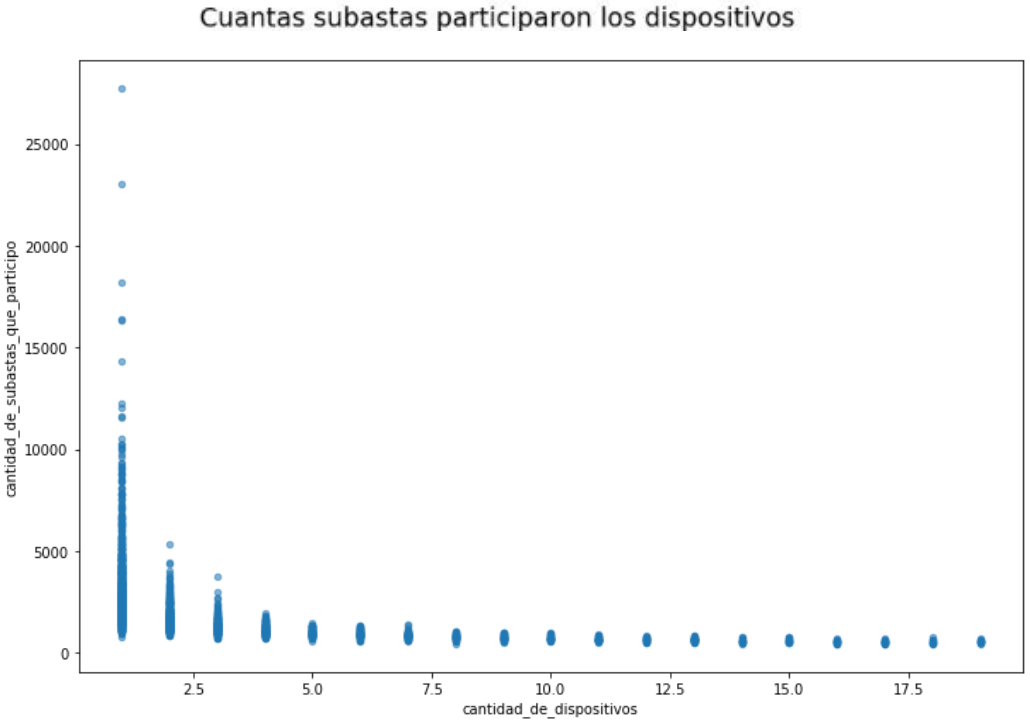


Viendo la gráfica anterior decidimos añadir un gráfico con los primeros 20 y vemos su participación en las subastas. Continúan siendo una participación excesiva en las subastas.

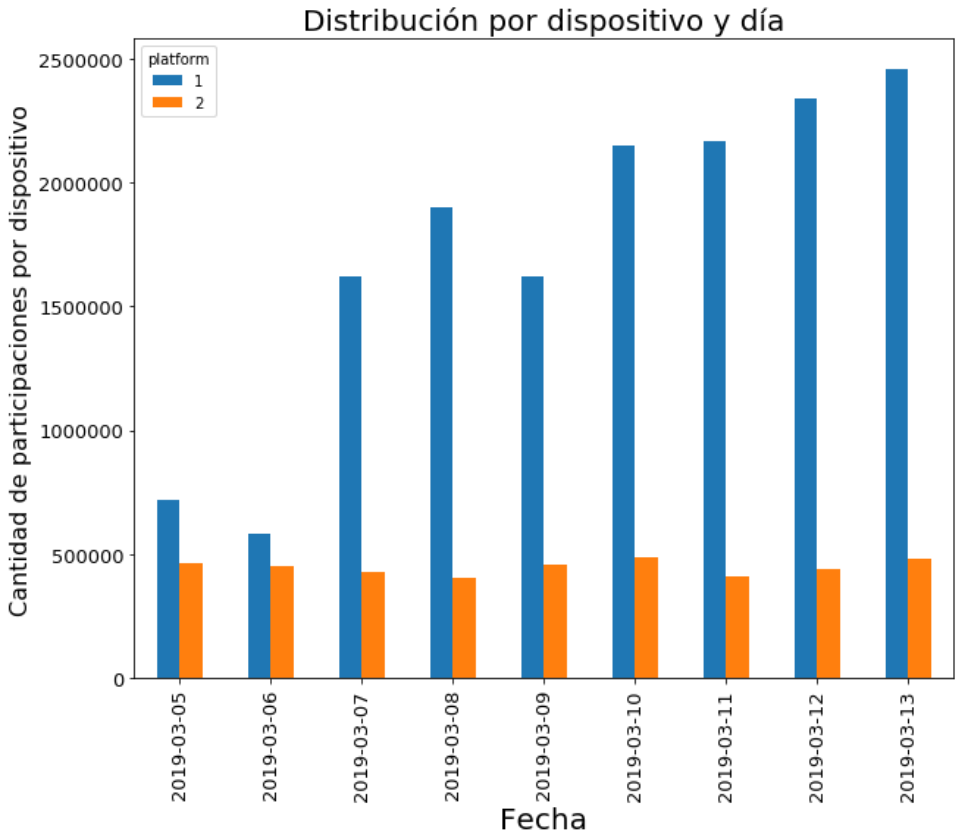
¿Cuántas participaciones en subastas por dispositivo hay en el periodo?



Aquí observamos que hay dispositivos que participan de subastas de manera excesiva para el periodo de tiempo del set de datos. Lo vimos en los gráficos anteriores que su cantidad de participación es muy alta. Reducimos la cantidad de dispositivos observados a 20.

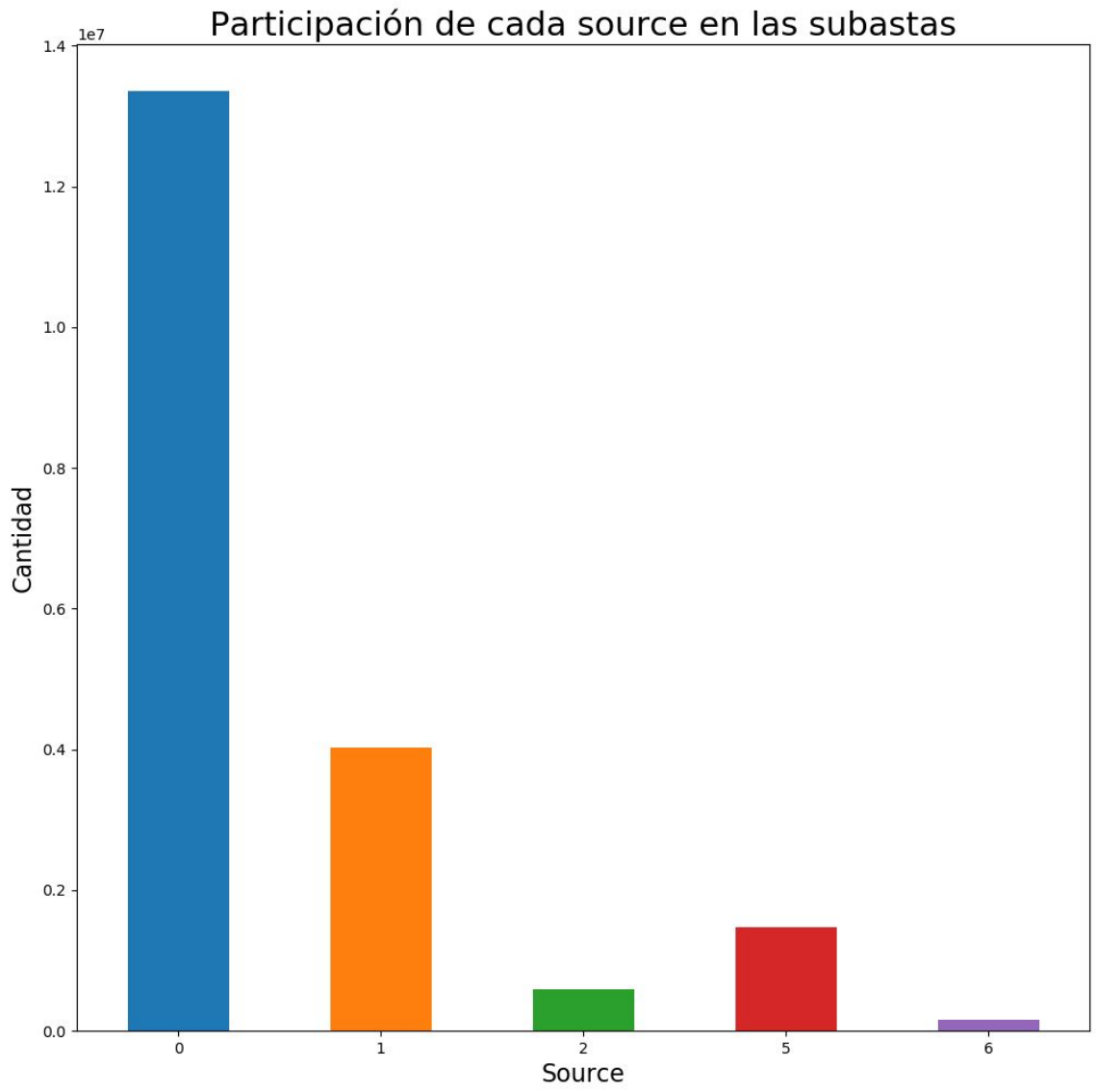


¿Como es la distribución del participación por día según la plataforma?

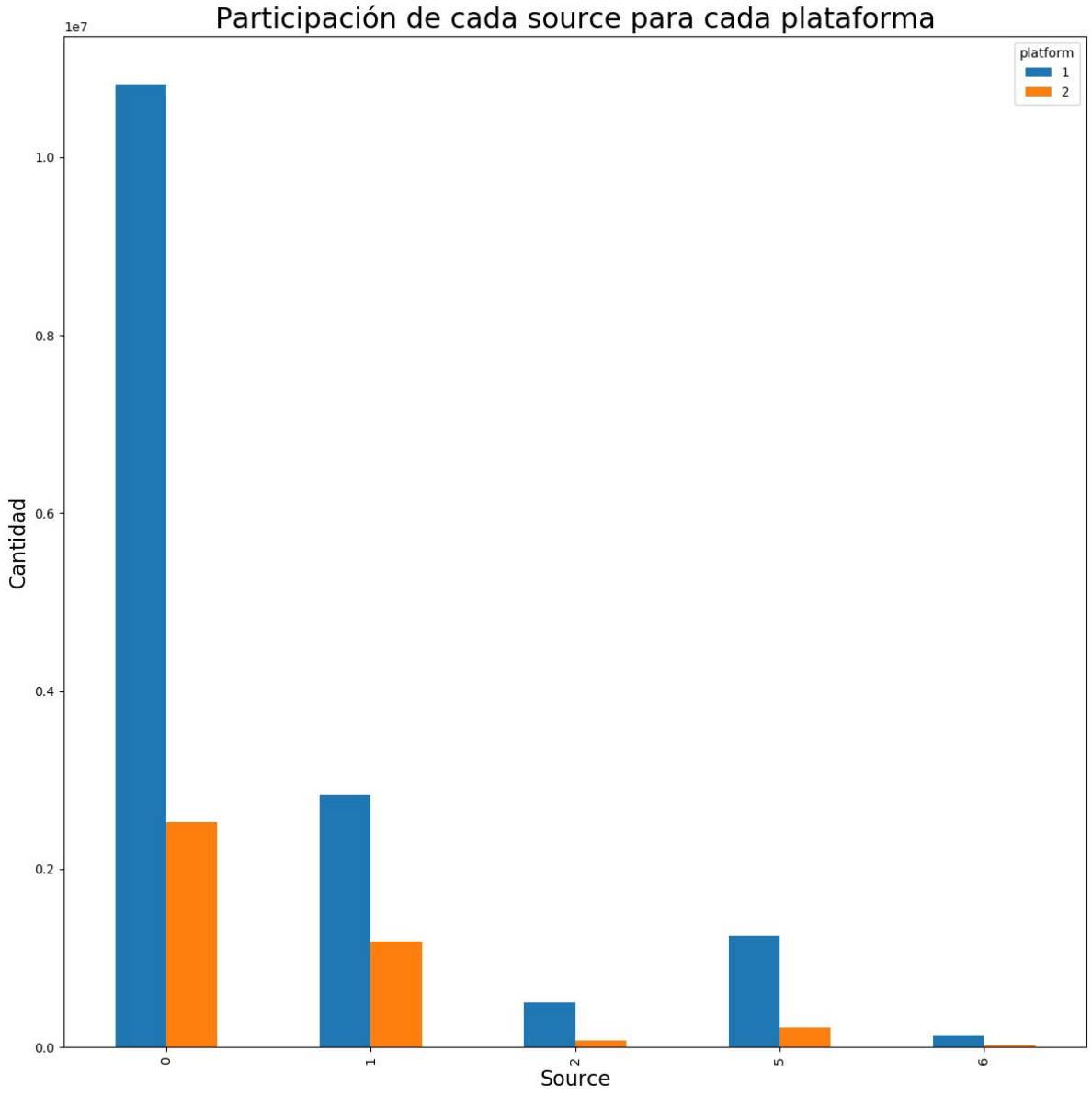


Aquí se ve que la plataforma 1 se observa un incremento diario de participación. En cuanto a plataforma 2 mantuvo una participación pareja durante el periodo siendo casi la misma participación por dia.

¿Cómo es la participación de los sources en las subastas?

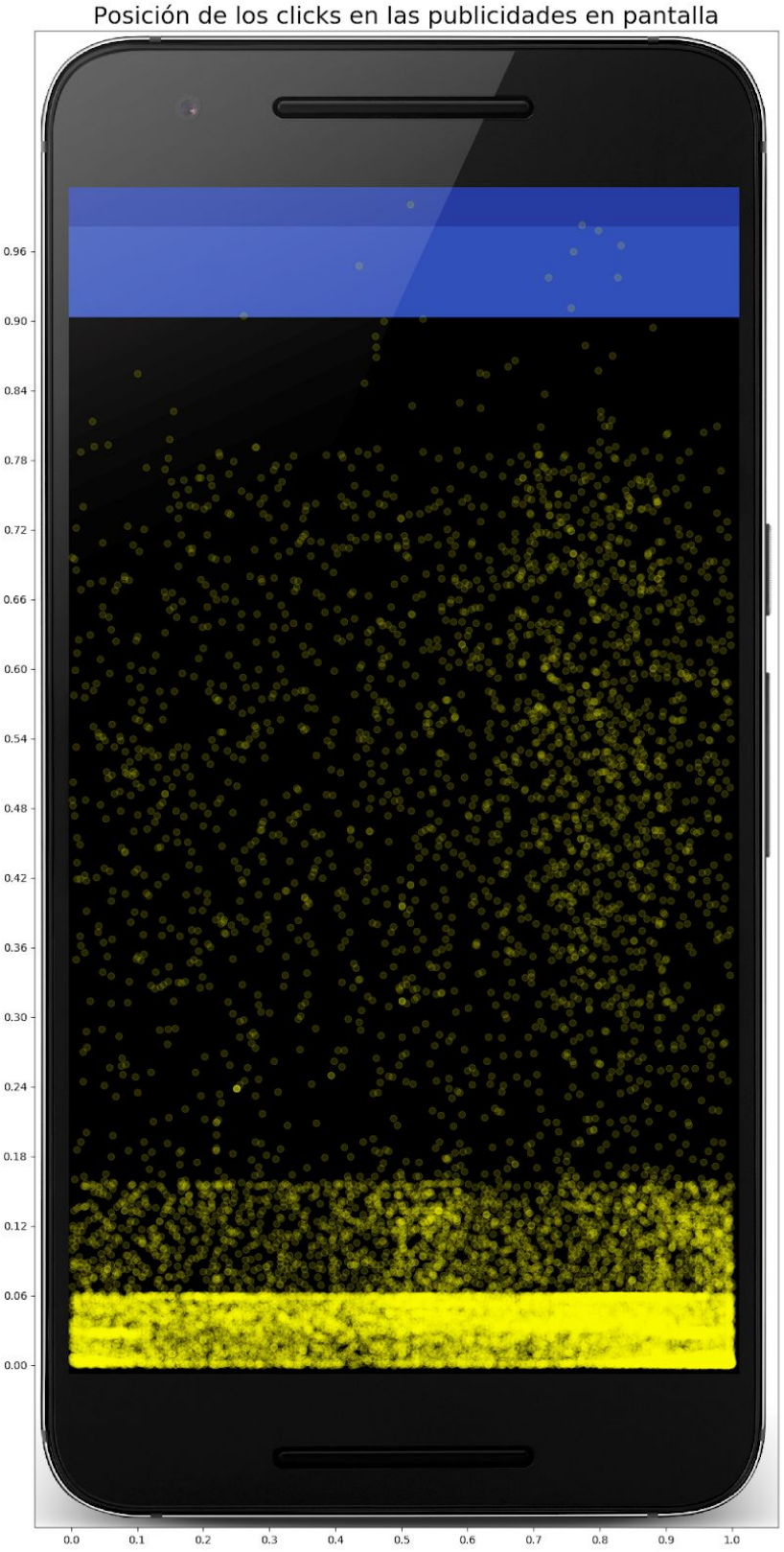


¿Cómo es la participación de los sources en las subastas para cada plataforma?



Análisis de Clicks

¿Cómo se distribuyen los clicks en la pantalla de los teléfonos?



Como se puede ver claramente en el gráfico, las publicidades suelen aparecer en la parte inferior de la pantalla de los teléfonos.

Investigando un poco acerca de anuncios en aplicaciones (App advertising[8]), en general existen dos tipos de anuncios: los de *pantalla completa* y los *banners*.

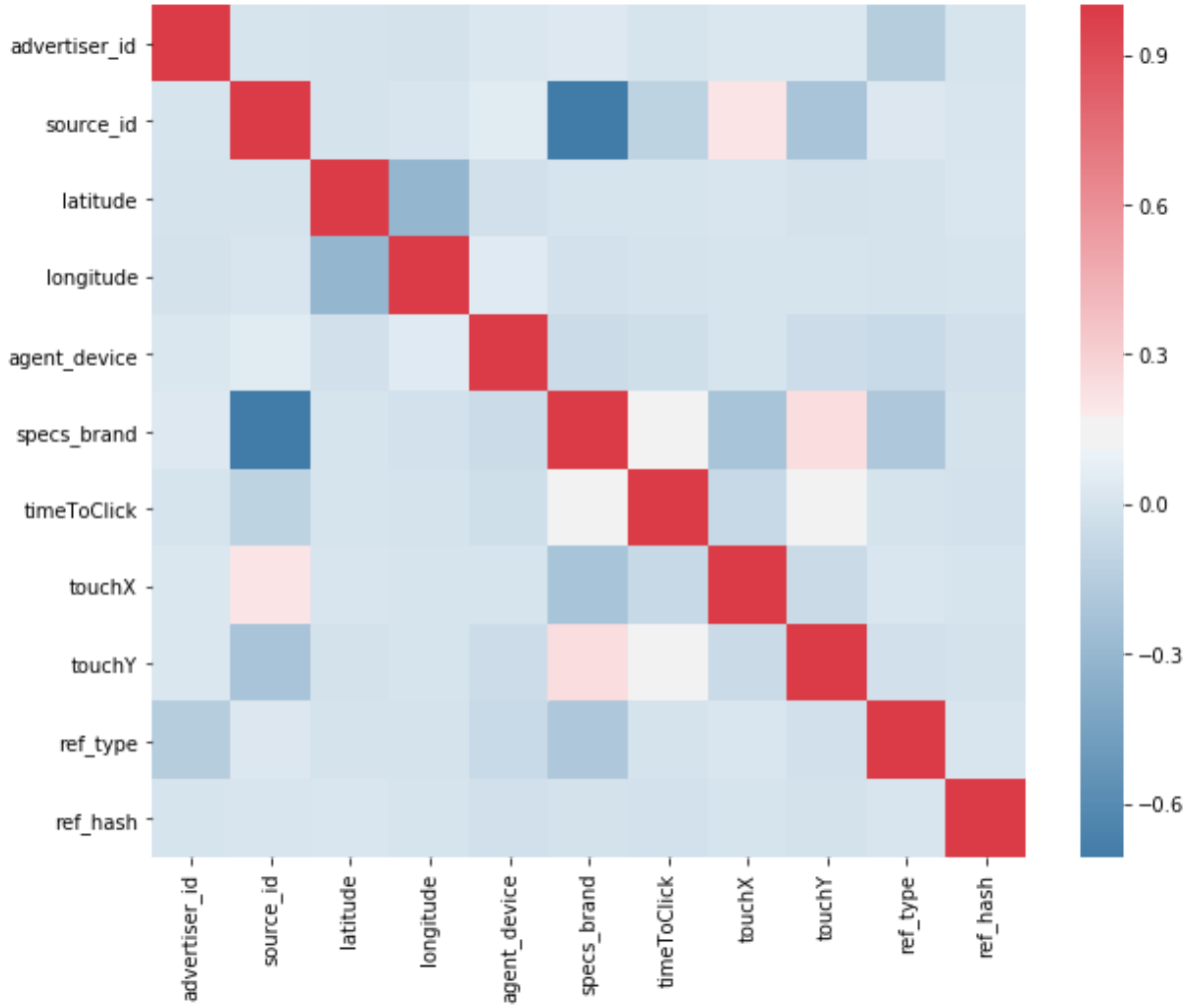
Los *anuncios de pantalla* completa brindan más espacio para un mensaje más amplio. Llaman la atención gracias al contenido de medios interactivos y visuales. Sin embargo, hay un inconveniente, este formato de publicidad en la aplicación es un poco intrusivo.

En cambio, el *banner* puede integrarse sin esfuerzos particulares. Es la opción más barata que puede usarse en cualquier pantalla. La única contra de este formato de anuncio puede ser presa del síndrome de la "ceguera del banner"; esto ocurre cuando el usuario ignora la información del mismo.

Dicho esto, si bien cada uno tiene sus problemas, posiblemente la opción del banner sea la más económica y fácil de integrar, y por este motivo, la más usada (en relación a los usuarios que los clientes de Jampp manejan).

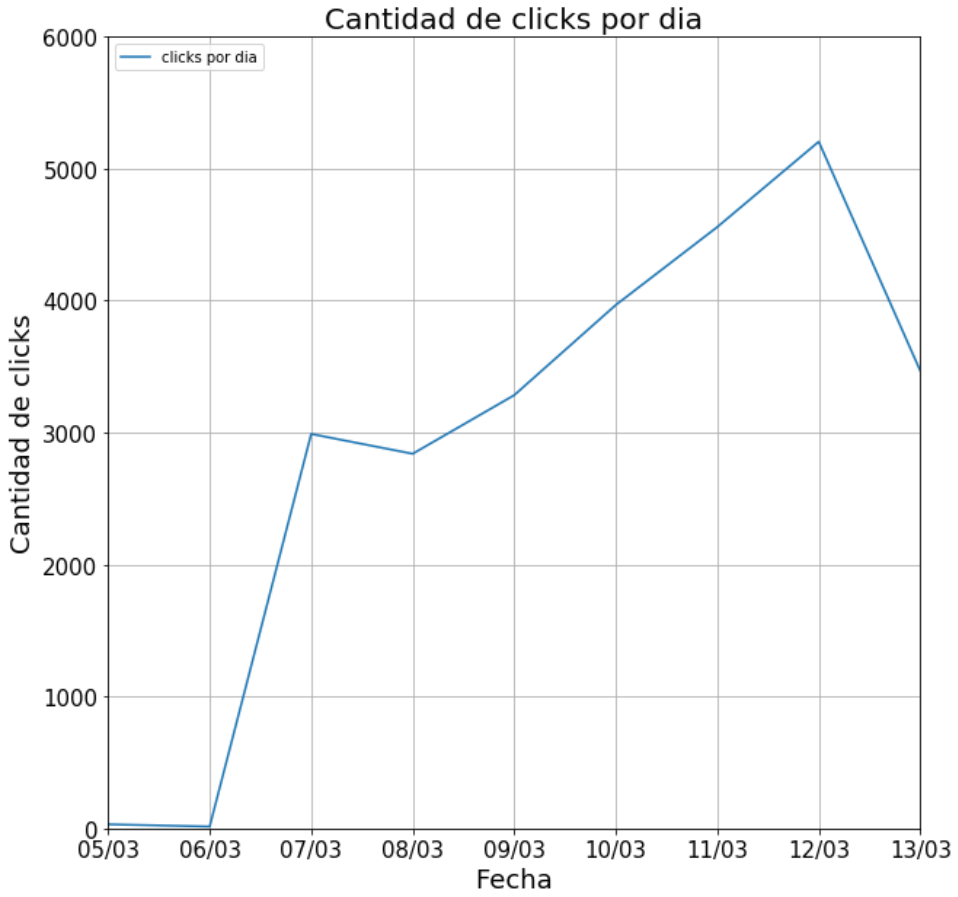
[8] https://splitmetrics.com/blog/guide-to-mobile-app-advertising/

Heatmap del set de datos



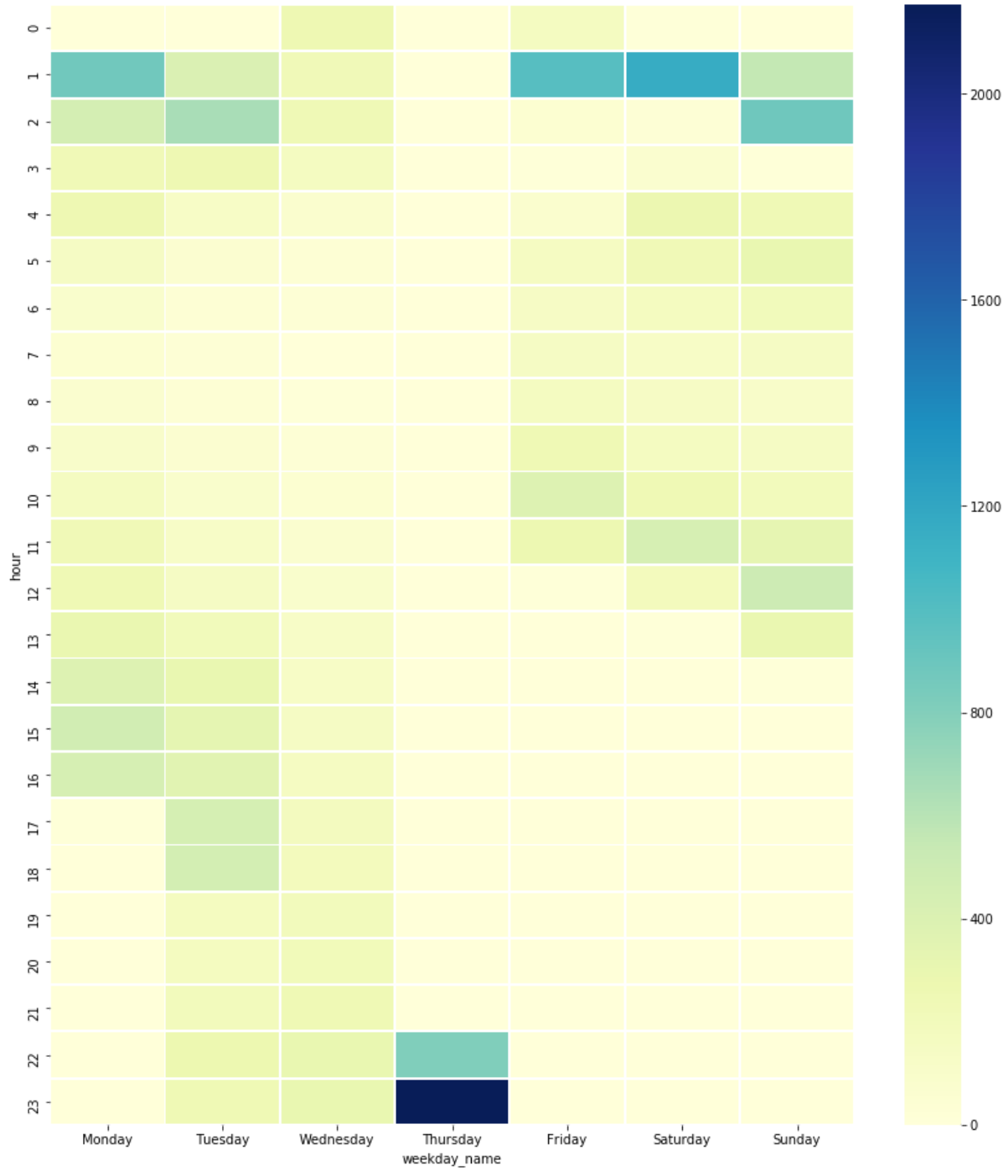
No se encuentran pares de features con correlación significante. La mayor correlación (negativa) se encuentra entre specs_brand y source_id.

¿De qué fecha obtenemos la información? y ¿Cuántos clicks por día hay en el periodo?



El dia Martes es el que se produce más clicks. Se mantiene la tendencia observada en los otros datasets. Crece a medida que surge el periodo pero en este caso el ultimo dia se produce un descenso.

Heatmap de clicks por hora y días de la semana

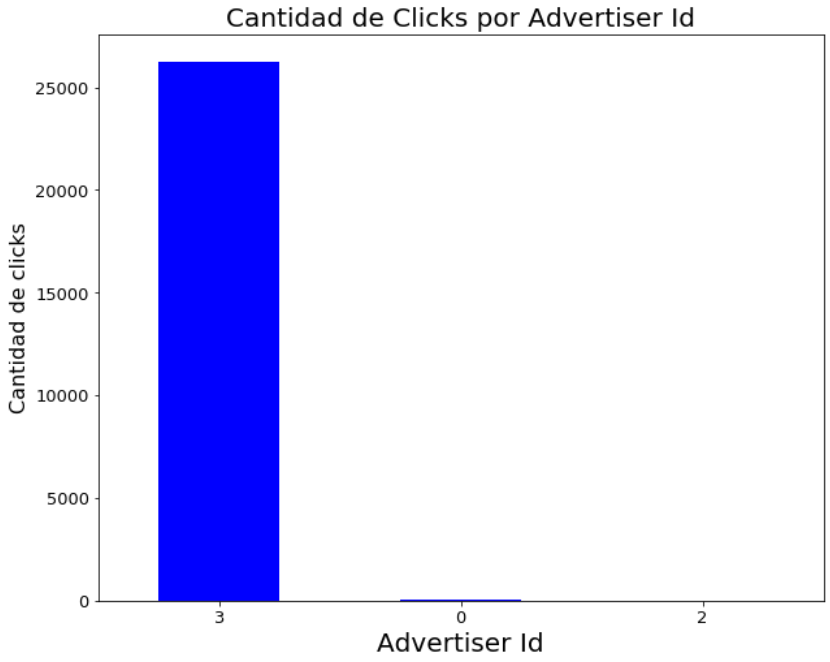


Se puede ver que los Lunes, Martes, Viernes y sábados, los horarios más frecuentados son a las 1 y 2 am. Los jueves a las 22 y más a las 23 están especialmente marcados. Habrá sido circunstancial de los días elegidos para formar el dataset, o será un comportamiento que se extiende en el tiempo?

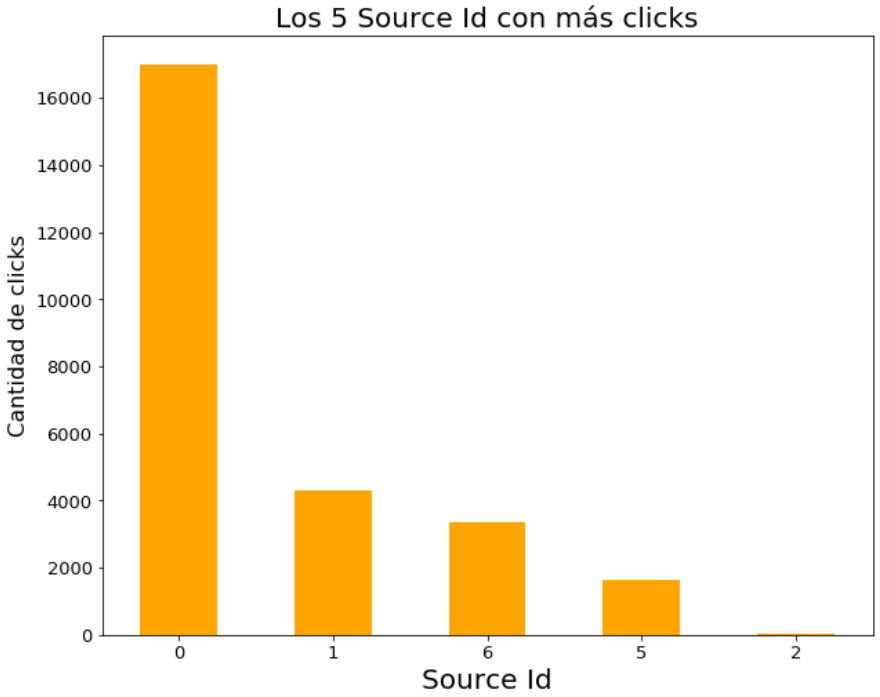
¿De qué anunciante provienen los clicks?

En este caso nos encontramos con que casi la totalidad provienen del mismo anunciante.

advertiser_id	total
3	26263
0	70
2	12



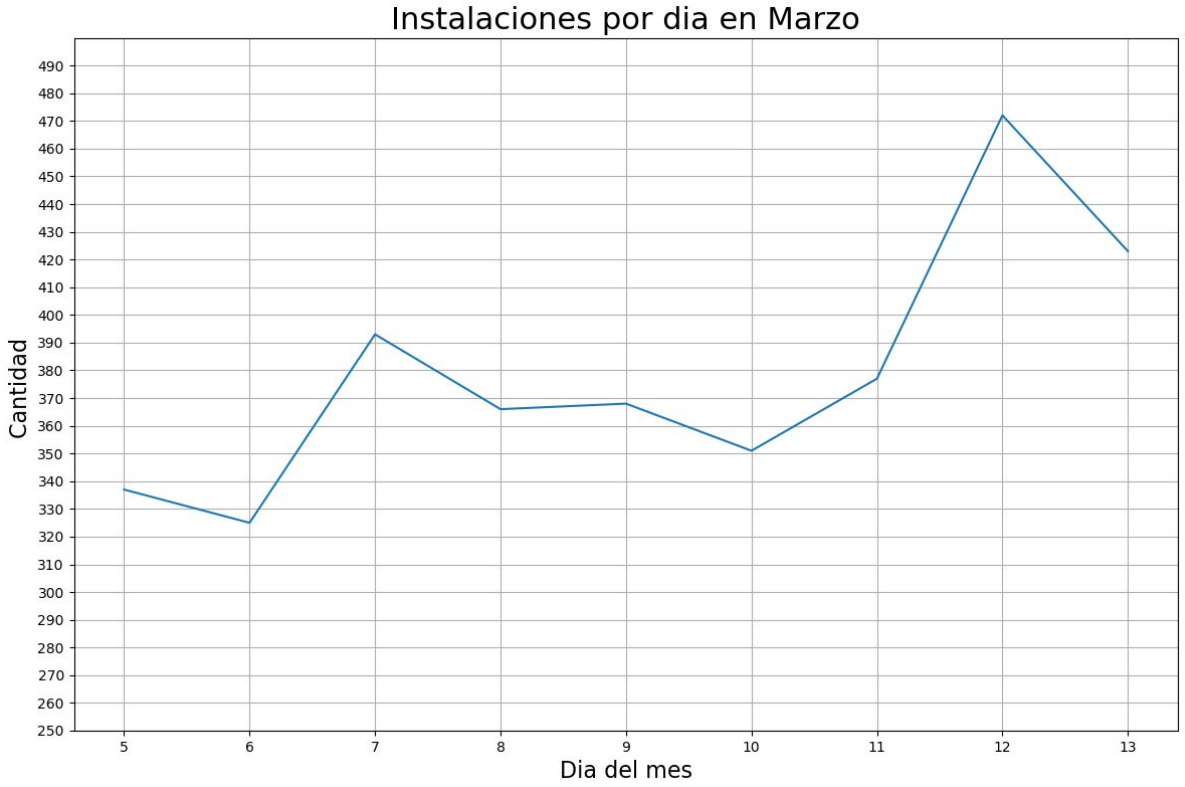
¿Cuales son las 5 fuente con más clicks?



Aquí se observa una predominancia del source 0 sobre los restantes.

Análisis de Installs

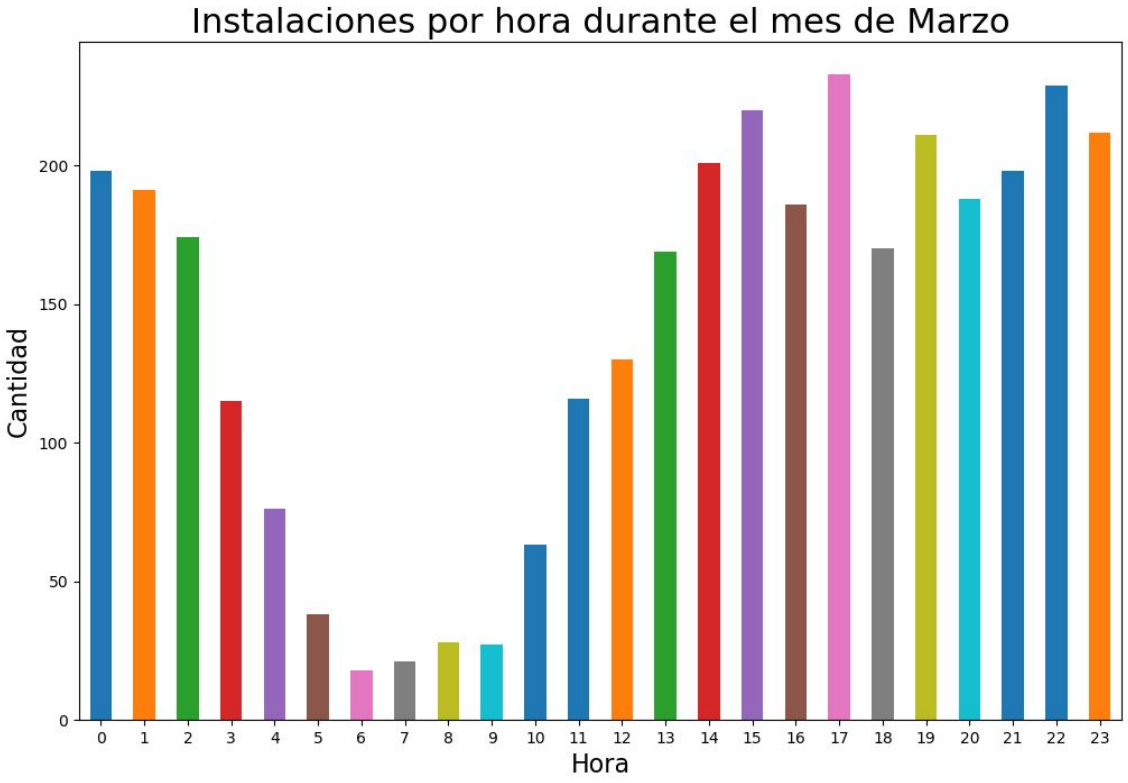
¿Cuantas instalaciones hubo en el periodo?



Aquí podemos ver que en el día doce es cuando más instalaciones se producen, esto no quiere decir que el próximo mes vaya a suceder lo mismo.

Si esto fuera en un lapso diferente y más amplio, podríamos darnos cuenta de alguna tendencia en la cual usuarios tienden a instalar más por ejemplo, pero lamentablemente aquí no es posible este tipo de análisis.

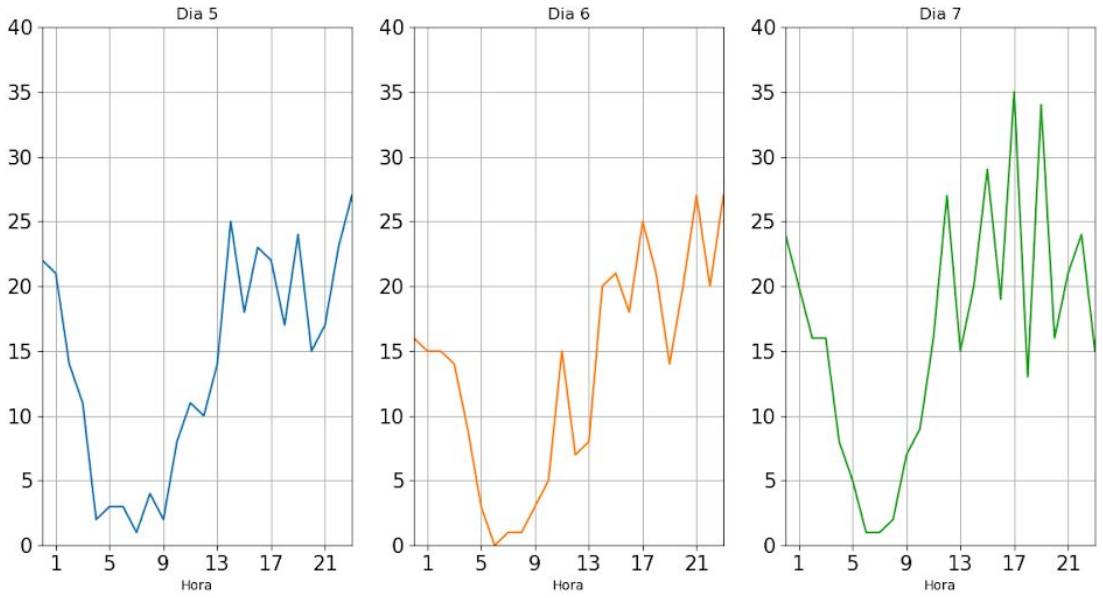
¿A qué hora se producen la mayor y la menor cantidad de instalaciones?

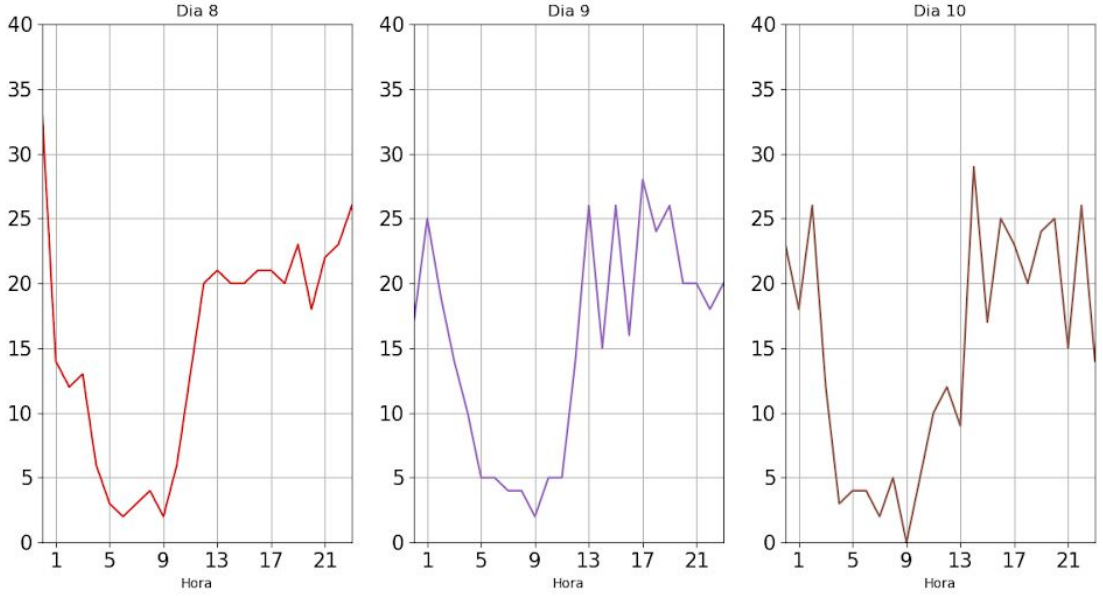


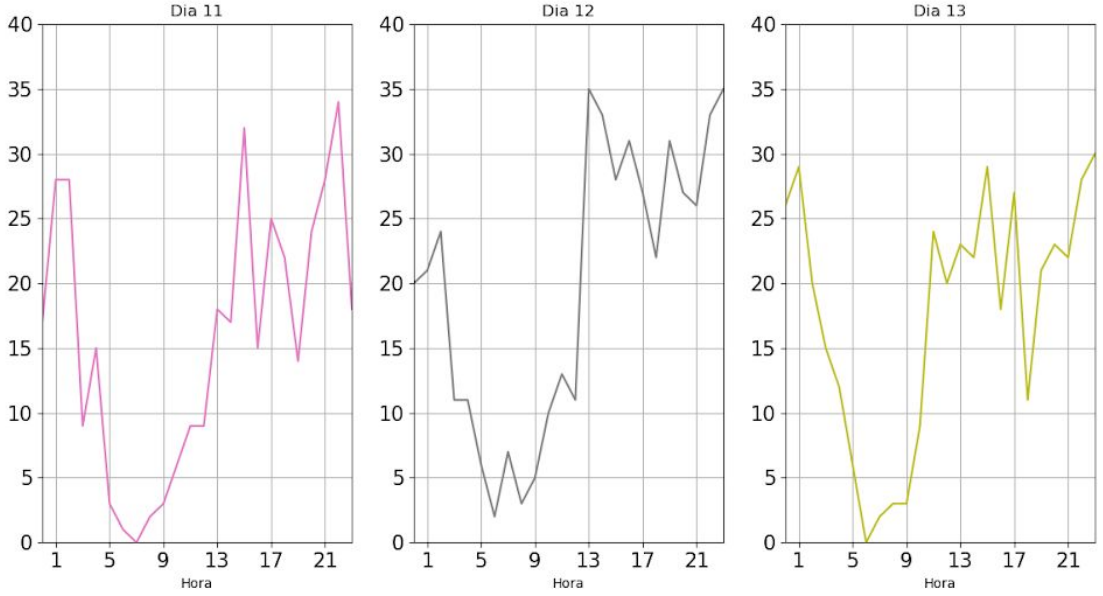
Volviendo un poco hacia lo mencionado de la representación de la realidad, en este gráfico, tanto como en aquellos a continuacion (horas para cada día de marzo), se ve que durante la noche la interacción de usuarios disminuye.

Lo único quizá notable aquí es que hasta las 2 AM, la interacción de usuarios continúa de manera no muy distinta a lo que puede ser por ejemplo la interacción a las 4 PM, lo que se podría uno imaginar que no debería suceder ya que es un horario en el que la gente generalmente duerme.

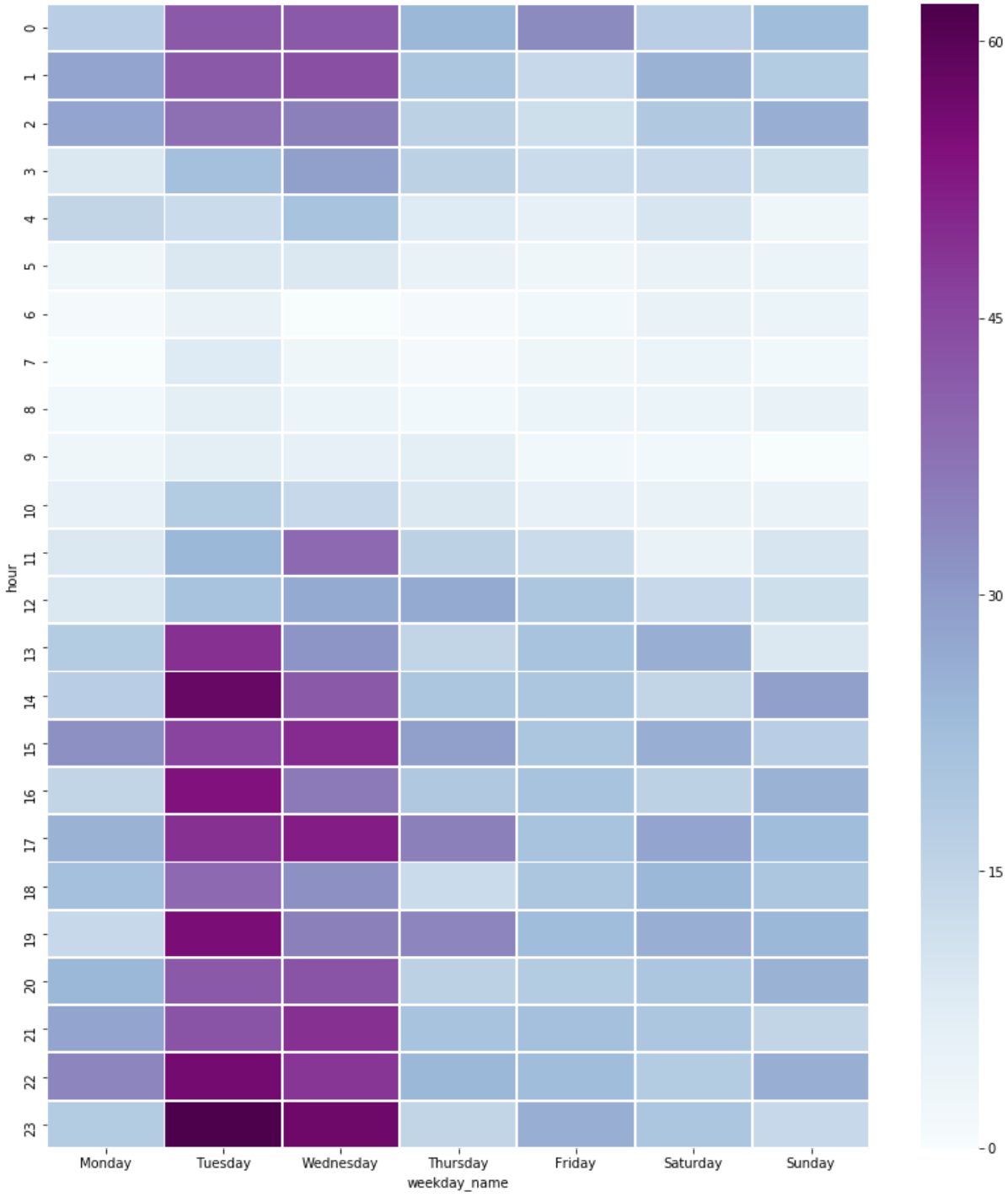
Además, vale la mención de que a las 5 PM es cuando más instalaciones se realizan.


Dia 5
Dia 6
Dia 7
Hora

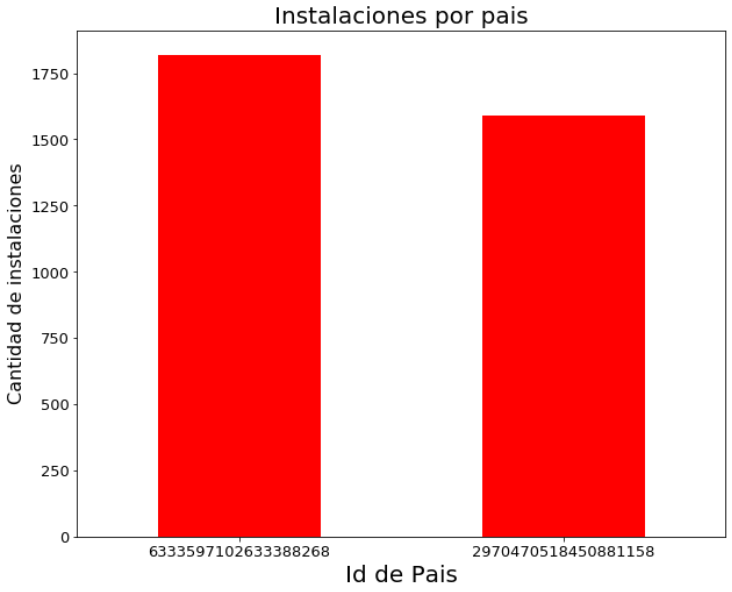

Dia 8
Dia 9
Dia 10
Hora


Dia 11
Dia 12
Dia 13
Hora

Heatmap de hora y días de la semana

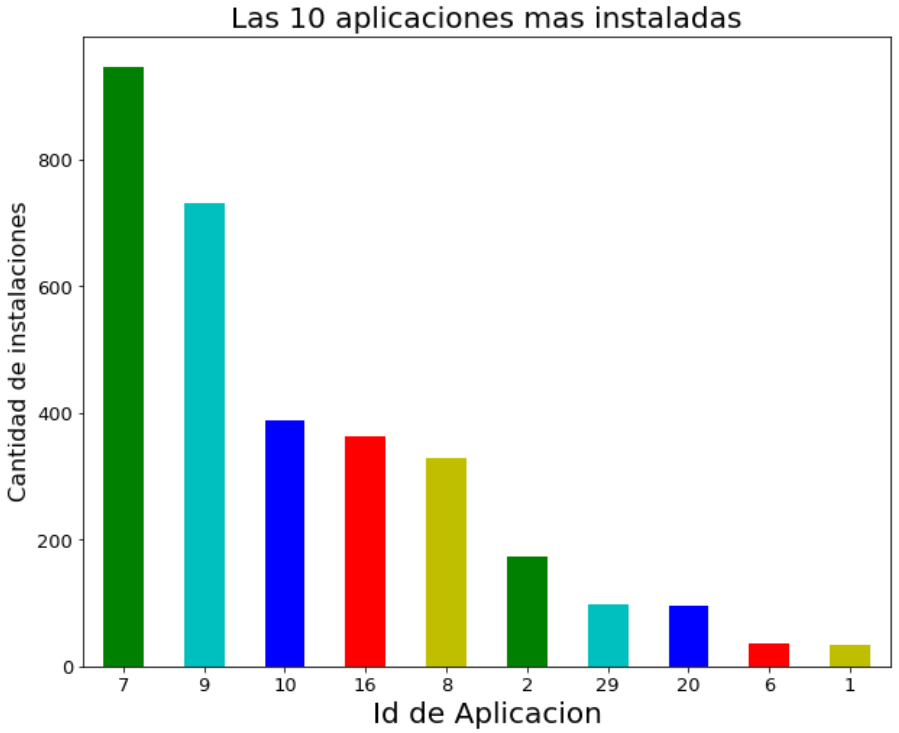


¿Cuantas instalaciones hay por país?



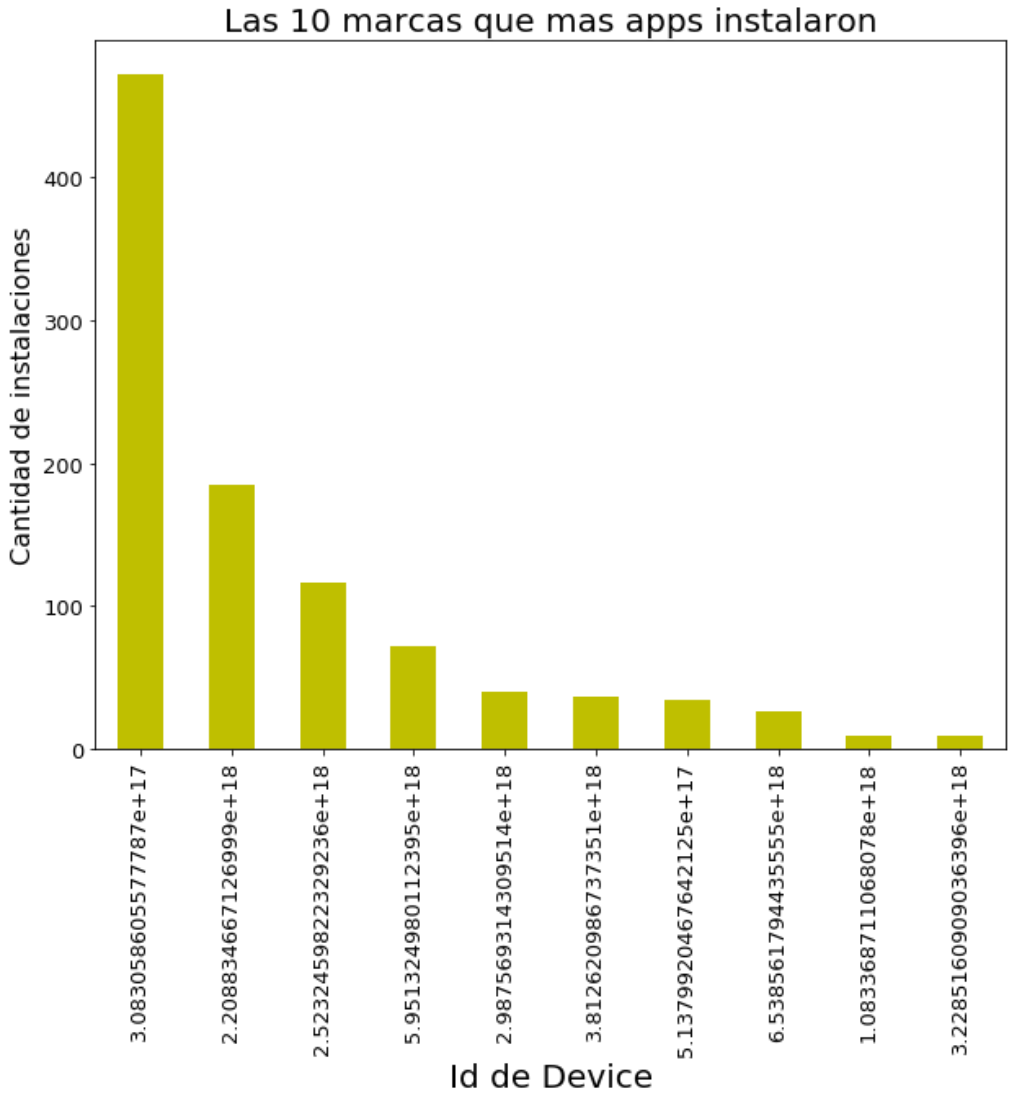
Aquí vemos que el set de datos nos muestra instalaciones que fueron realizadas desde 2 países, distinto por ejemplo a lo encontrado en el de Subastas que todas provienen del mismo país.

¿Cuales fueron las 10 apps más instaladas?



Obtenemos las apps que más instalaciones tuvieron. No nos es posible determinar su nombre o origen ya que están ocultos los datos pero observamos que de las 2 primeras hay una buena fuente de instalaciones.

¿Cuáles fueron las 10 en las cuales se instalaron más apps?



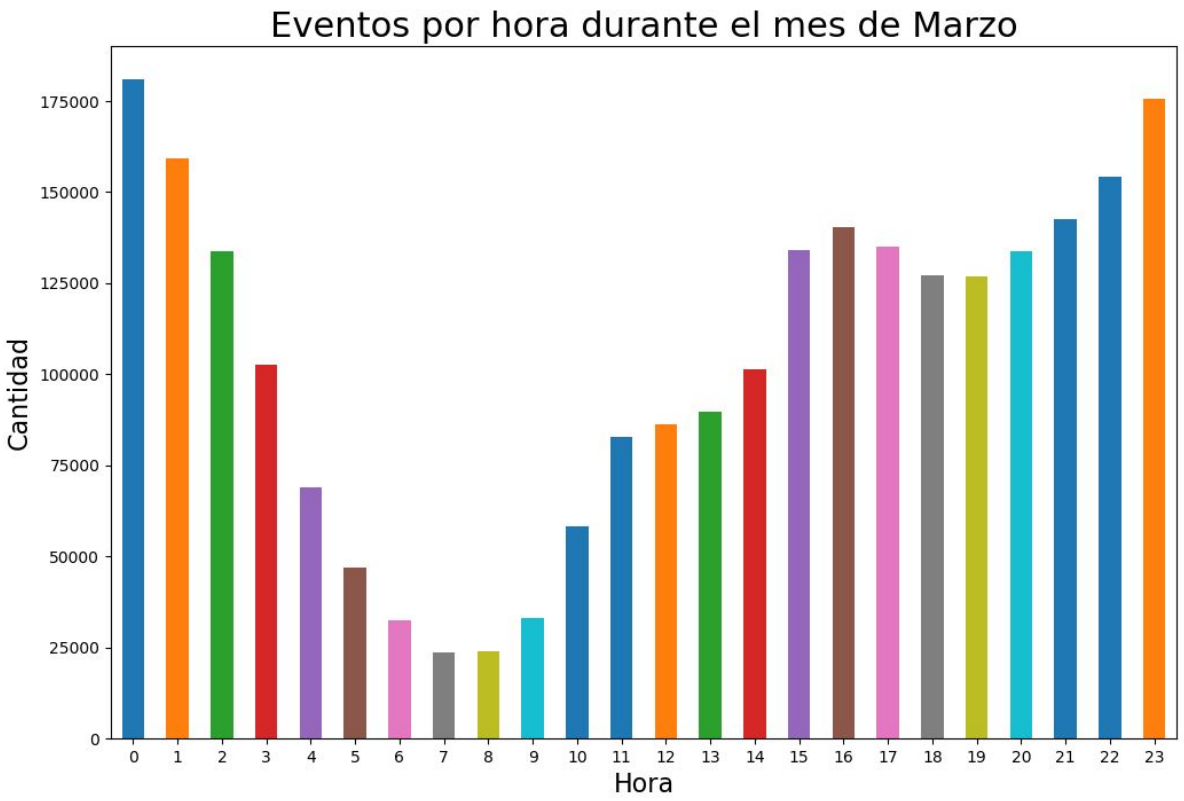
Definitivamente no nos es posible obtener la marca. Pero observamos el resultado y vemos que de la primera marca se produjeron muchas instalaciones con respecto a las siguientes.

Análisis de Events

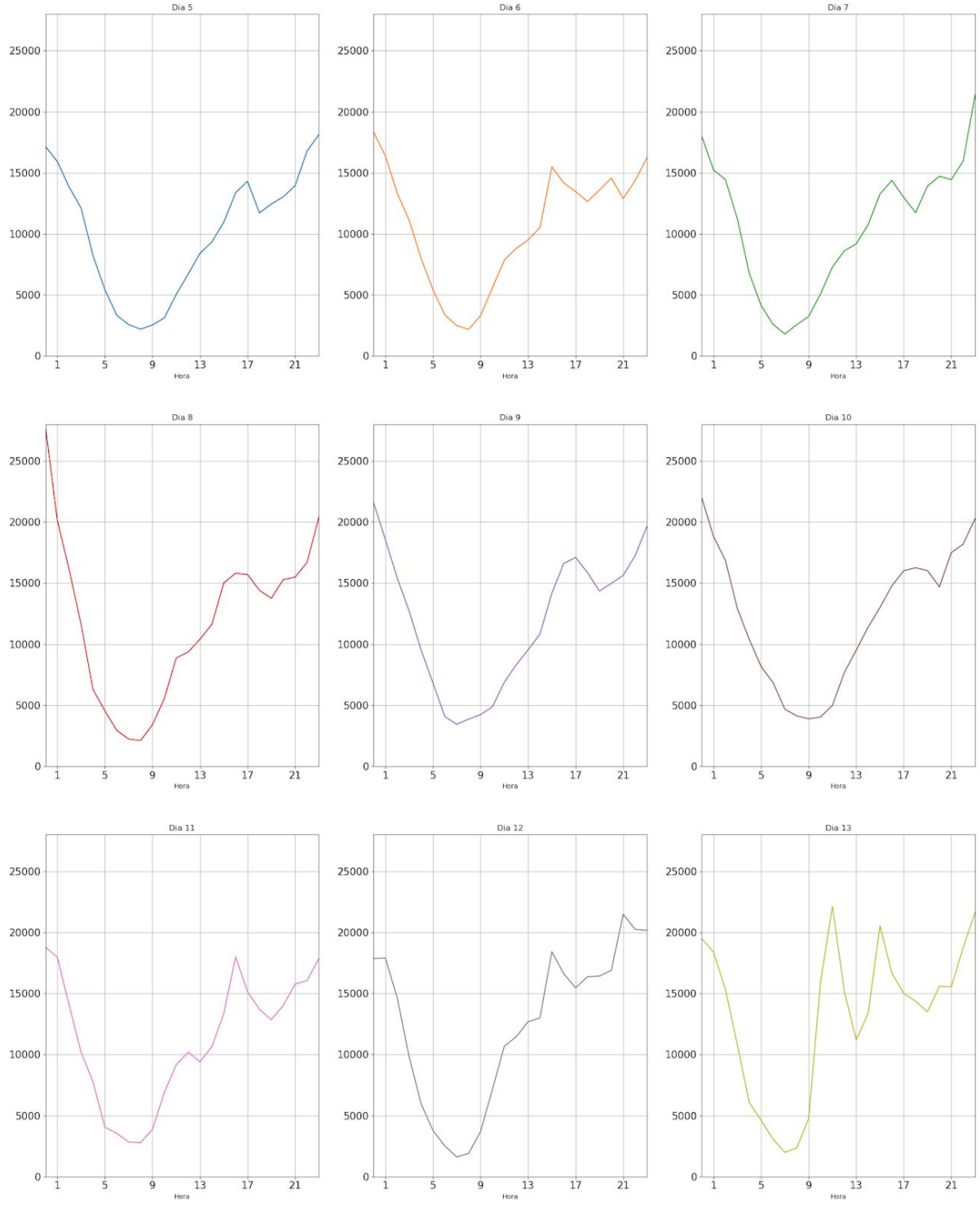
¿Cuantos eventos hubo en el periodo?

Aquí el análisis es análogo a las instalaciones, en el día trece es en el que más eventos se realiza pero no tiene mucho significado, incluso tiene menos significado por la naturaleza de lo que puede significar realizar un evento.

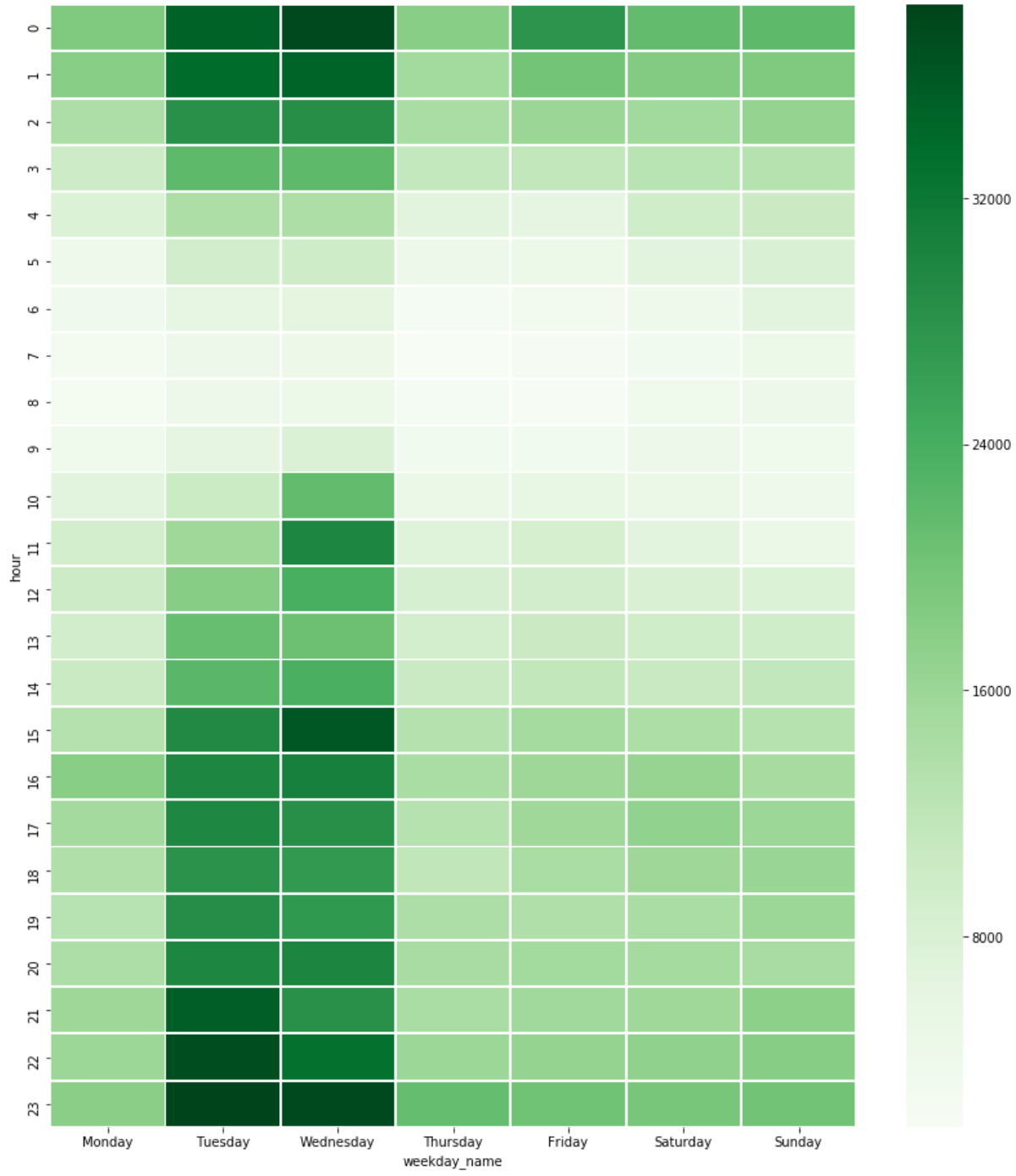
¿A qué hora se producen la mayor y la menor cantidad de eventos?



Análisis análogo a instalaciones por hora y por día, aquí a las 0 HS es cuando más eventos se realizan.

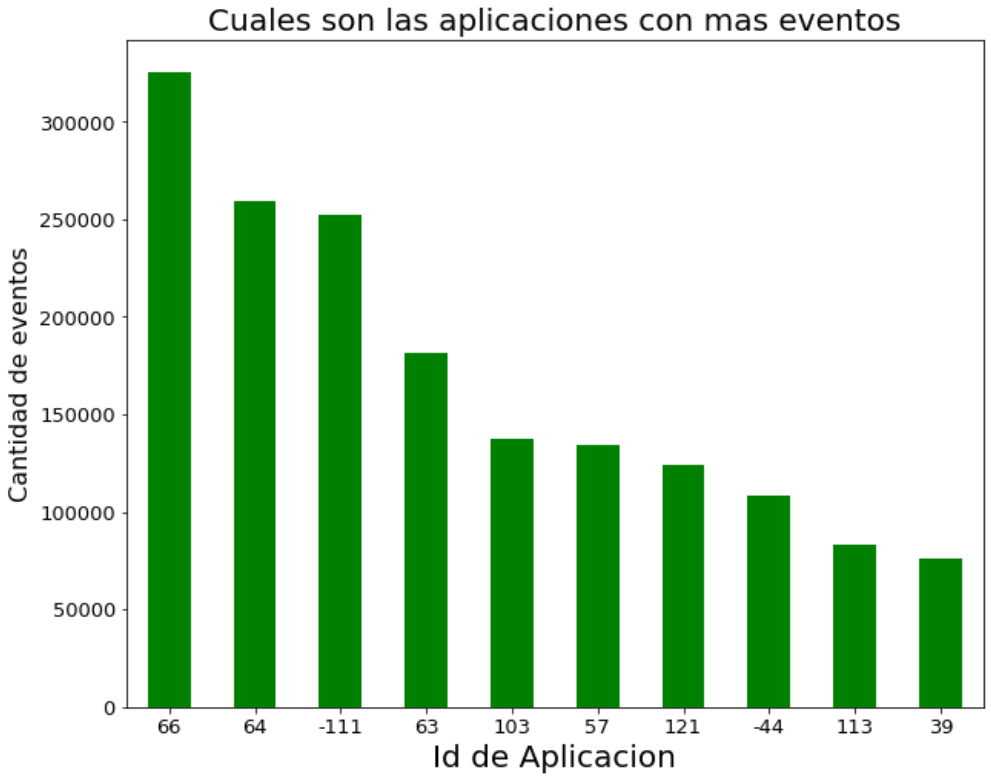

Dia 5
Dia 6
Dia 7
Dia 8
Dia 9
Dia 10
Dia 11
Dia 12
Dia 13
0
5000
10000
15000
20000
25000
1
5
9
13
17
21
Hora

Heatmap por dia de la semana y hora



Se observa una distribución mayor durante martes y miércoles.

¿Cuales son las apps que más eventos generaron?



Conclusión

A lo largo de este informe se realizaron visualizaciones y análisis partiendo de distintos puntos de vista con el fin de comprender, analizar y proponer mejoras al dominio en cuestión.

Tanto el modelo de negocio como los datos no resultaron amigables ni fácilmente entendibles a primera vista (algunos tampoco luego de recibir varias aclaraciones de la cátedra), por lo que requirieron también investigación previa al análisis de la información expuesta: algunos de sus resultados no son triviales.

El objetivo principal planteado al comienzo de la exploración de datos, más allá de analizar su información en sí, consistió en proponer alguna mejora de manera tal que Jampp pueda lograr una mejora en cuanto a las instalaciones atribuidas. Este objetivo hizo que el descubrimiento de la ausencia de registros con instalaciones atribuidas resultara un tanto chocante.

Fuera de la cantidad contada de inconvenientes, el grupo puede ahora afirmar que el entendimiento del set de datos es adecuado como para poder pasar a la segunda parte del tp.