Universidad de Buenos Aires

Facultad de Ingeniería

(75.06) Organización de Datos

Cátedra: Luis Argerich

TP 1

https://github.com/freedocx/7506Jampp

1° Cuatrimestre 2019

Alumnos:

GARATE, Julián Matías (93043)

CAI, Ana Maria (102150)

ORTIZ, Javier (96598)

FERNÁNDEZ, Andrés (102220)

Introducción

El presente trabajo, se trata sobre la realización de un análisis exploratorio sobre el set de datos brindado por la empresa JAMPP.
El dominio en cuestión es el marketing digital. La empresa tiene aplicaciones clientes las cuales comparten información, de sus customers, sobre estos datos, sumada a la de RTB (Real-Time Bidding), JAMPP trata de discriminar a cuáles usuarios, que son potenciales customers de las aplicaciones clientes, se le muestra una publicidad, comprando ese espacio en una subasta digital.

Para el análisis de utilizar el lenguaje de programación **Python** con las librerías de **PANDAS** y para las visualizaciones **SEABORN**, **MATPLOTLIB**, **SQUARFY, WORDCLOUD, BOKEH** y **HOLOVIEWS y Graph-tools**.
En el repositorio de Github se colocaron las correspondientes notebooks con los códigos fuente que validan los procedimientos y el análisis.

Set de datos

La estructura en la que se nos presentan los datos, es de cuatro archivos .CSV:

1) Clicks: datos enviados a JAMPP desde las Apps clientes de JAMPP.i

   - advertiser_id: id cliente que paga el aviso.
   - action_id: id del tipo de click.
   - source_id: id de los publishers que subastan el inventario de las apps.
   - created: timestamp.
   - country_code: no se uso ya que es el mismo, Uruguay.
   - latitude y longitude : posición transformada desde donde se hizo el click.
   - wifi_connection: sus valores son todos falsos, por lo tanto no lo utilizamos.
   - carrier_id: operador del celular..
   - trans_id: id interno de la transacción
   - os_minor, os_mayor: versión del OS.
   - brand: marca del dispositivo.
   - timeToClick: segundos que tardó en clickear.
   - touckX y touchY: posición de la pantalla en dónde se clickeó.
   - ref_type: si es iOS o Android.
   - ref_hash: id de la aplicación cliente de JAMPP

2) Installs: installs de las Apps clientes de JAMPP

   - created: timestamp.
   - application_id: id de la App.
   - ref_type: si es Iphone o Android.
   - ref_hash: id del dispositivo desde el cual se realizó el evento.

- click_hash: id de la instalación.
- attributed: si la instalación fue atribuida a JAMPP.
- implicit: sí la instalación se hizo de tal manera que no pudo ser trackeada por la plataforma (Todos los valores son falsos).
- device_brand: marca del dispositivo.
- device_model: modelo del dispositivo.
- session_user_agent: user agent utilizado para la instalación.
- user_agent: user agent del dispositivo.
- event_uuid: id del evento.
- kind: tipo de instalación.
- wifi: si el dispositivo utilizaba conección wifi.
- trans_id: id de la transacción.
- ip_address: dirección ip dónde se realizó la instalación.
- device_language: lenguaje del dispositivo.
    -

3) Auctions: subastas en las que participó JAMPP

- auction_type_id: columna nula.
- date: timestamp
- device_id: id dónde se inició la subasta
- platform: si es iOS o Android.
- ref_type_id: id interna de JAMPP para la plataforma (otro bool de iOS o Android)
- source_id: id de los publishers que subastan el inventario de las apps.

4) Eventos: datos sobre cómo interactúan los customers con las Apps clientes de JAMPP

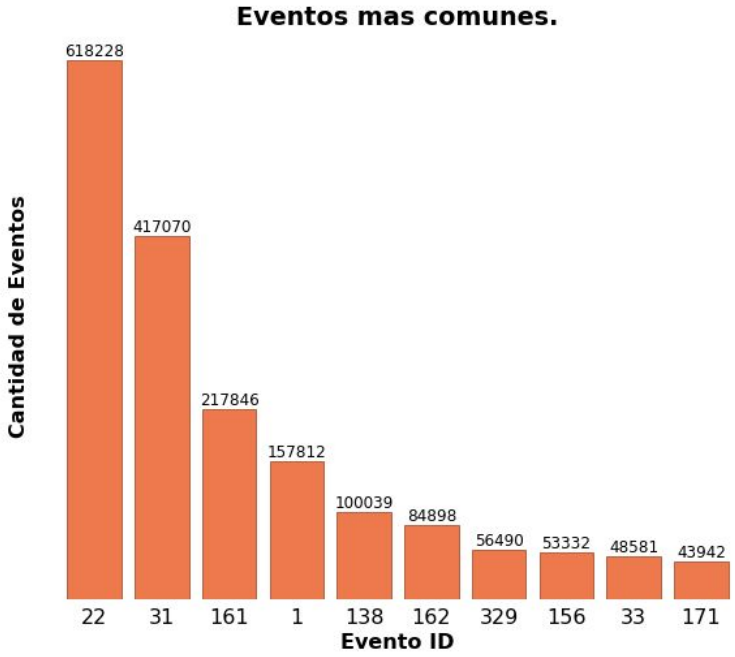
- date: momento en el que se registró el evento.
- events: acción de un usuario dentro de una determinada aplicación.
- ref_hash: id del dispositivo desde el cual se realizó el evento.
- ref_type: si es iOS o Android.
- application_id: id de la aplicación cliente de JAMPP, la cual comparte los datos.
- Ip_address: ip pública desde la cual se realizó el evento.

Debido a la corta ventana de tiempo, algunas columnas no son relevantes al análisis, por tener gran cantidad de campos vacíos y variables muy predominantes sobre otras, son retiradas.

Analisis Exploratorio

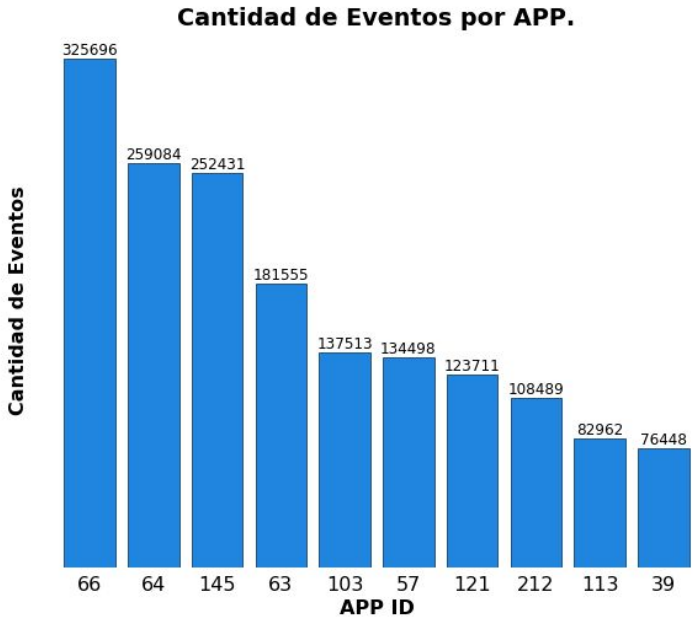
Eventos

Hay una gran cantidad de eventos, de los cuales como se ve en el gráfico, por ejemplo el evento 22 acumula 600 mil registros.



Parece ser que las aplicaciones comparten eventos, puede ser el caso de que, por ejemplo, todos los e-commerce tengan un evento del estilo ‘brand_list_view’.

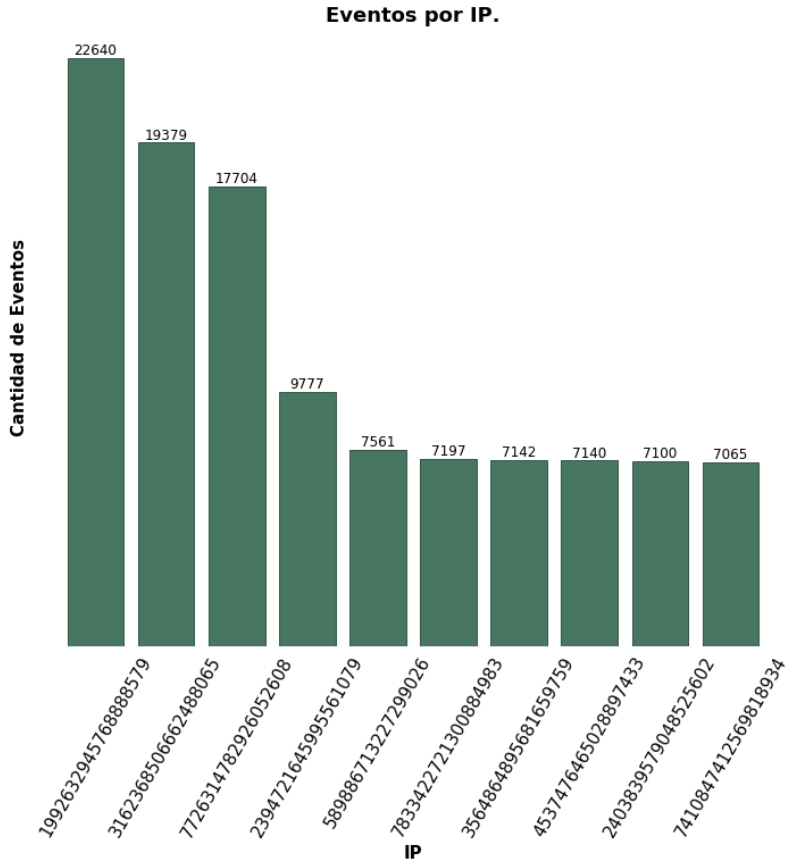
Se cuentan a continuación la cantidad de eventos de cada una de las aplicaciones clientes

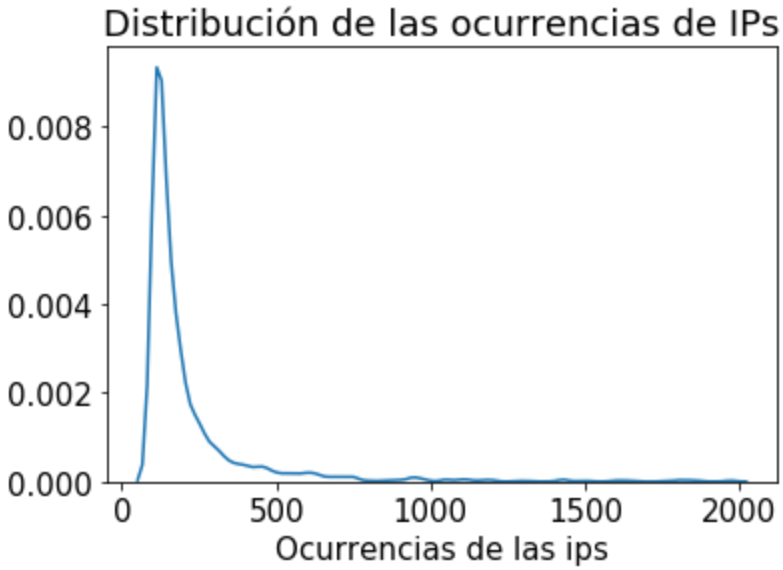


La aplicación 66 tiene 325 mil eventos, denotando que hay diferencia de popularidad.

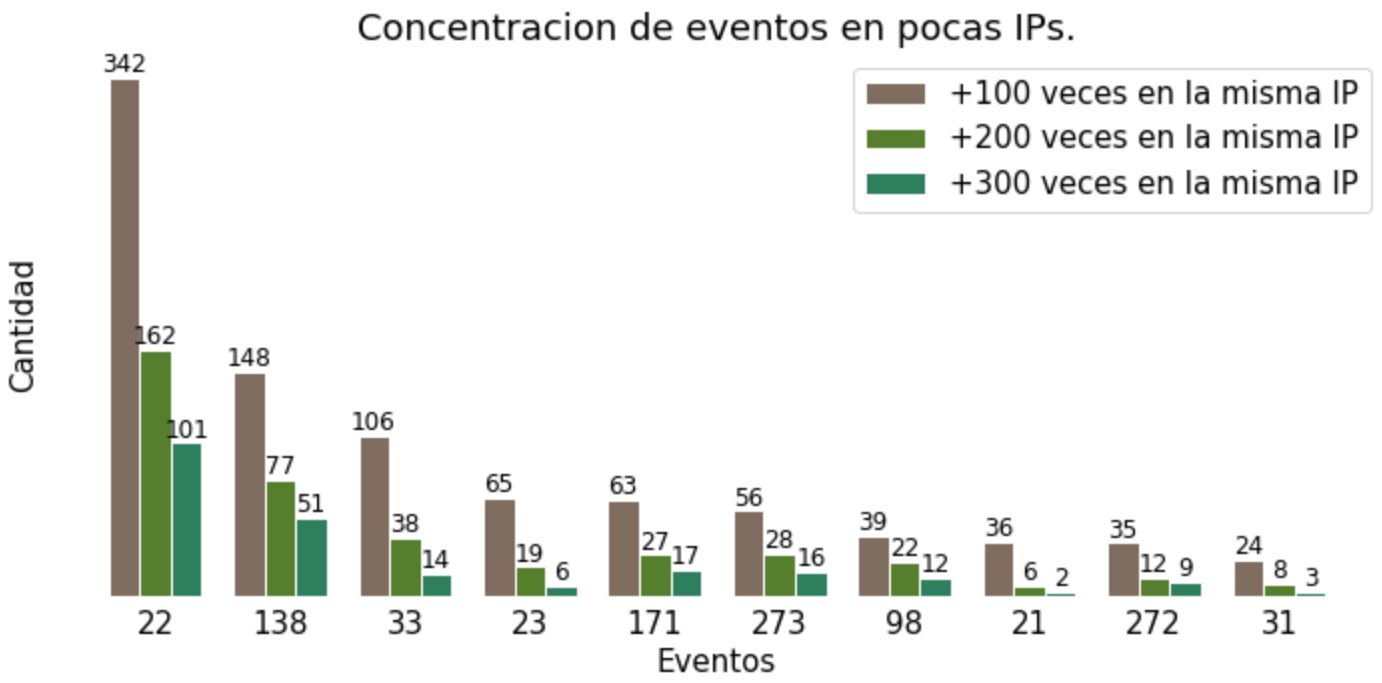
Ip address

Hay ips que concentran muchos eventos incluso, para la ventana de tiempo dada, arriba de 20 mil.

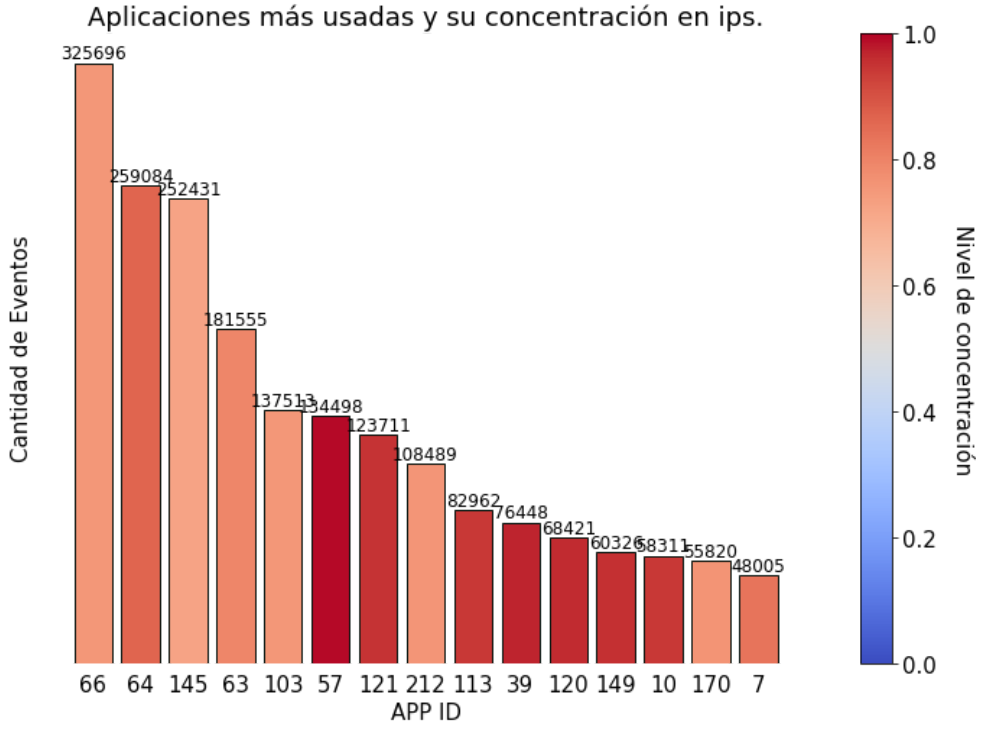




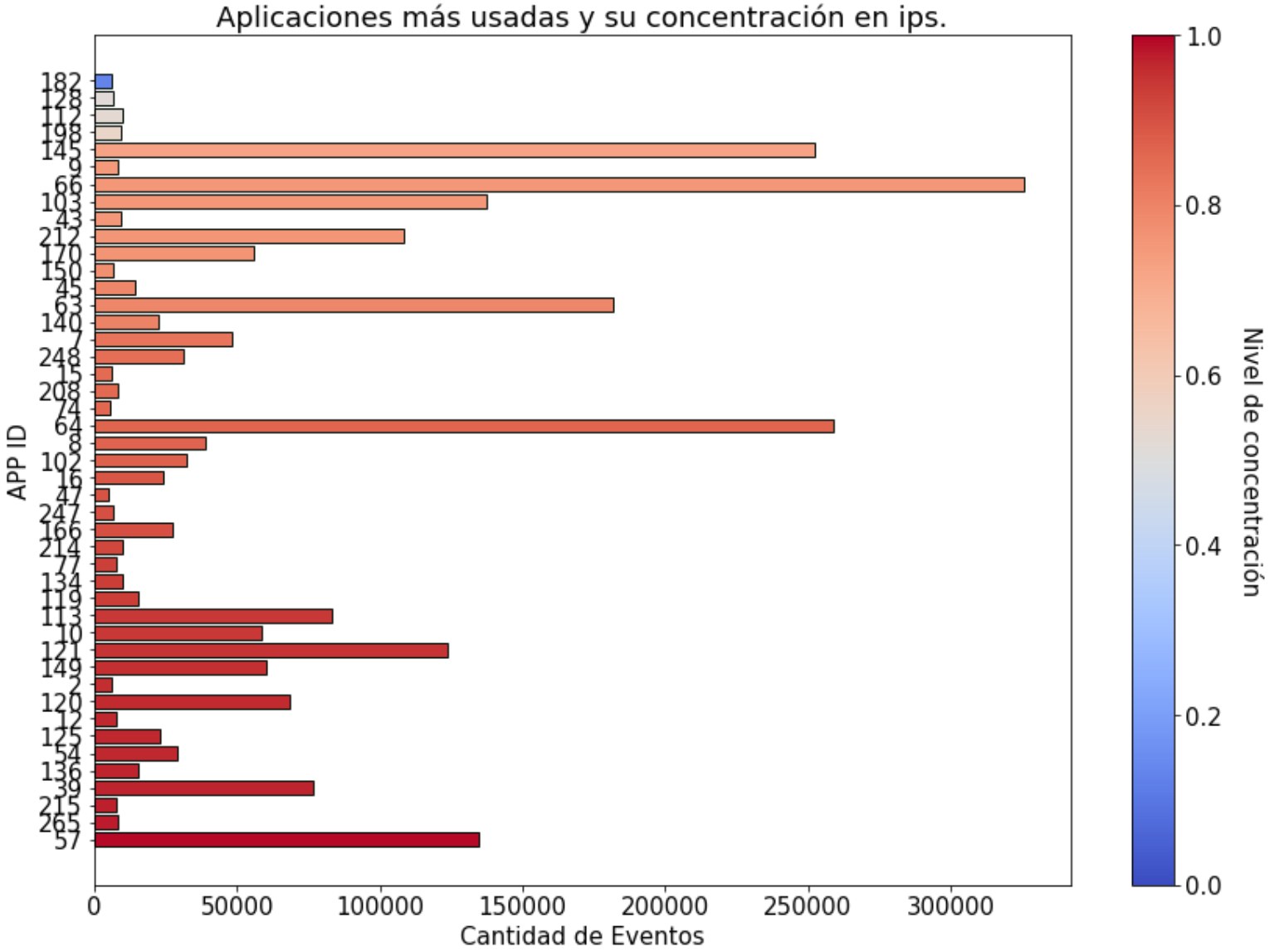
La gran mayoría tienen pocas ocurrencias pero hay algunas pocas que concentran un gran número



La concentración de eventos, más allá de su gran cantidad de ocurrencias, por ips nos muestra que hay algunos que son específicamente dispersos. Se procede a hacer un análisis de las aplicaciones en las cuales estos eventos son realizados



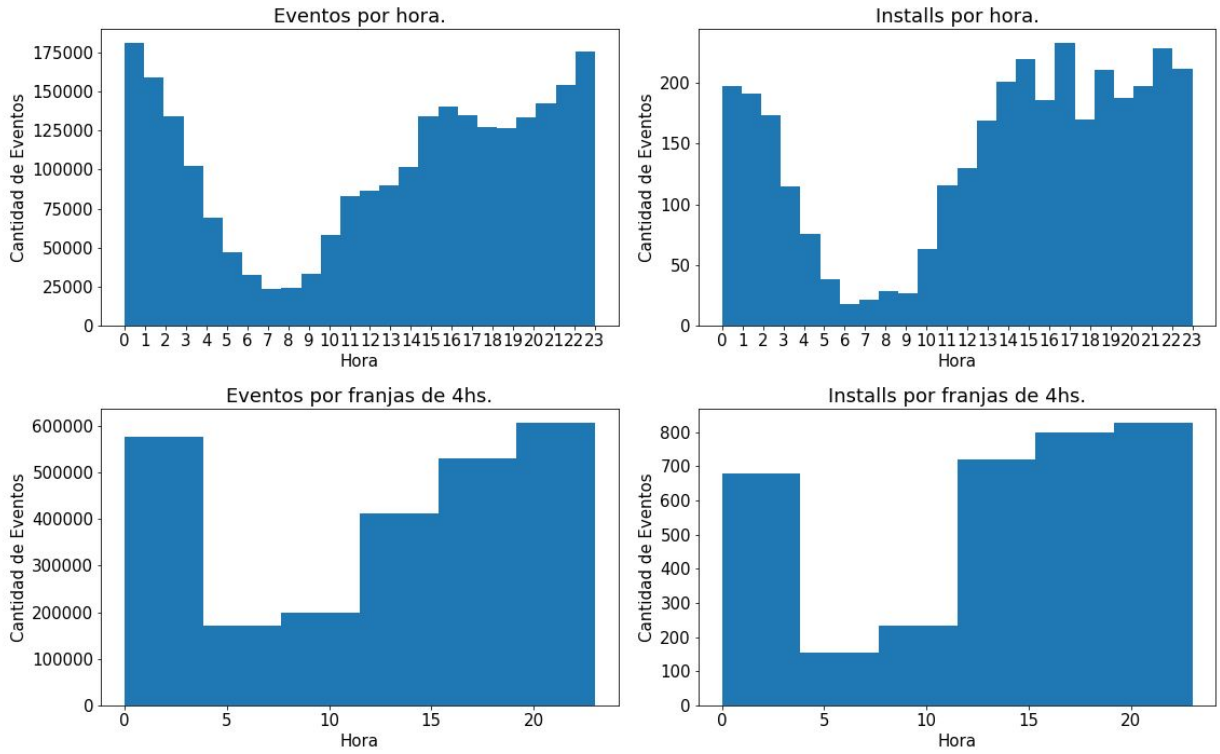
La diferencia del tono de rojo muestra el nivel de concentración en pocas ips de las aplicaciones más comunes. Ampliando el número entonces.



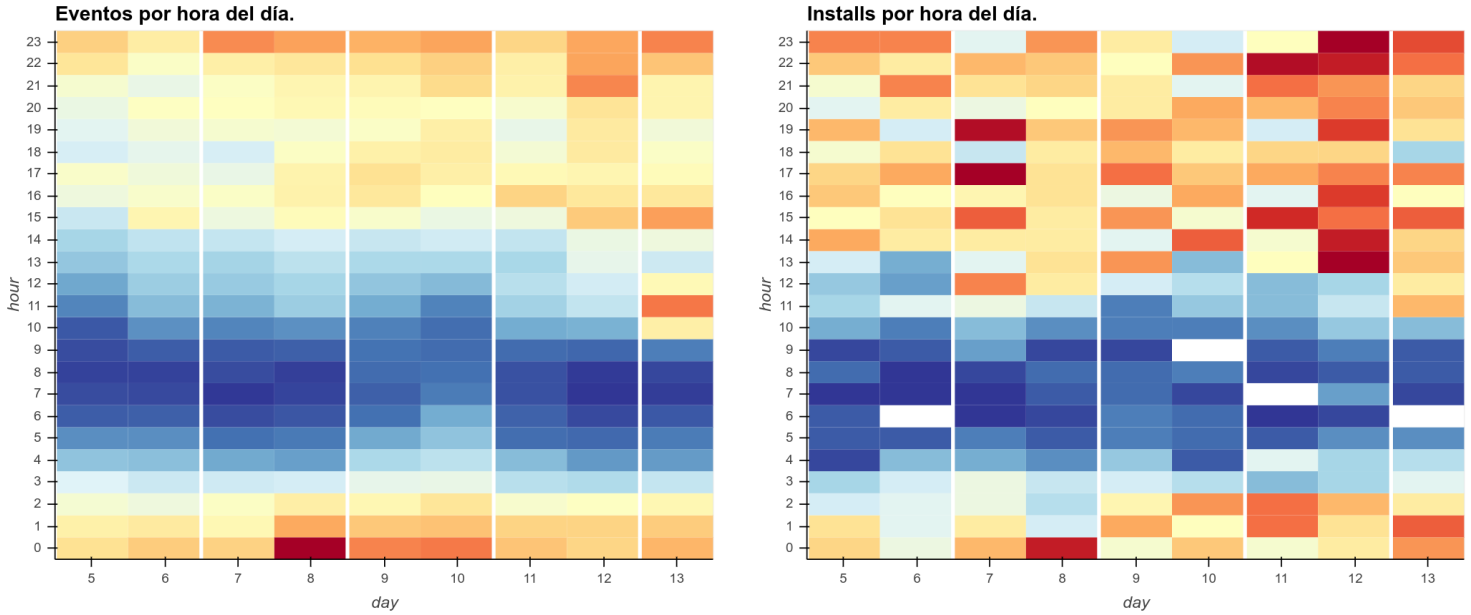
Los niveles de concentración son relativamente elevados, pero hay algunas apps que son bastante dispersas en ips, por ejemplo la 182.

Installs vs Eventos

Las aplicaciones son usadas durante diferentes momentos en el dia, se pueden ver franjas horarias bien marcadas según la hora. Pasa lo mismo con las instalaciones de aplicaciones, se procede a comparar ambos dataset y a analizar cómo se distribuyen.

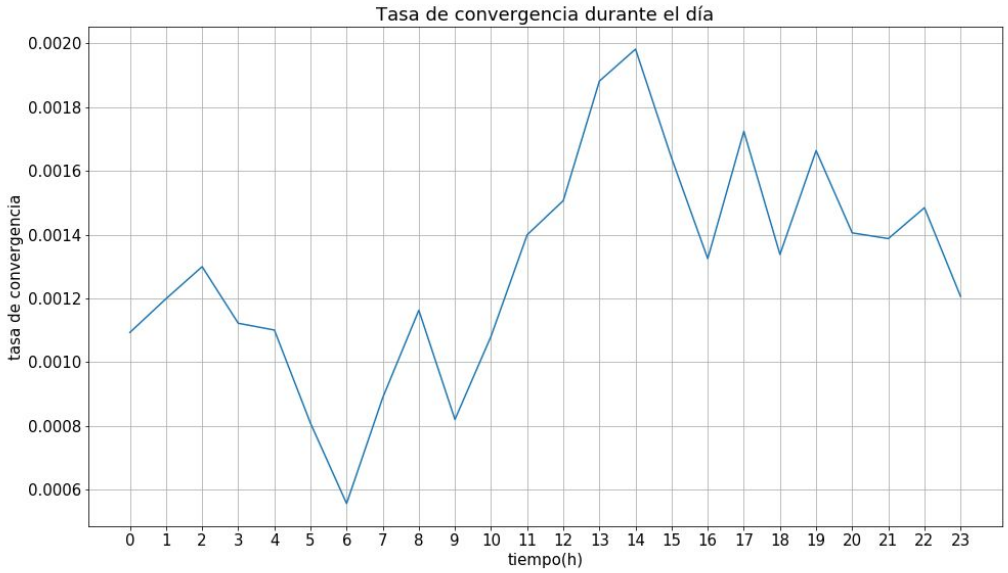


La columna de la derecha son instalaciones y la de la izquierda son eventos, se puede ver que hay momentos en el dia, que a primera vista, tienen diferente proporción de instalaciones y eventos.



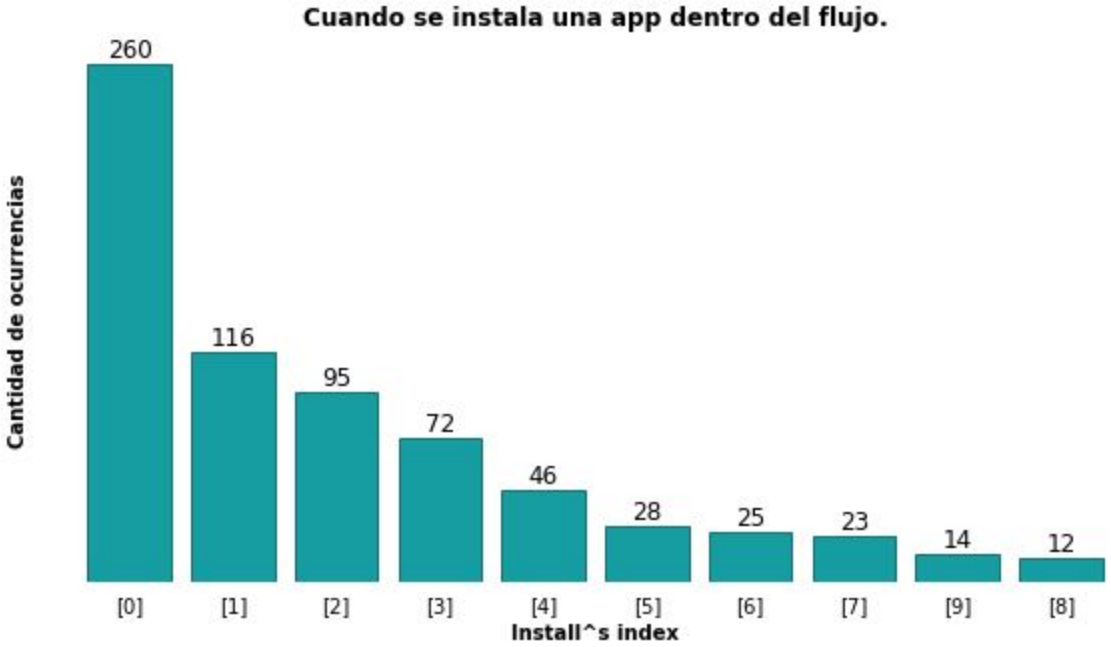
Se pueden observar franjas de baja temperatura (azules) donde la ocurrencia de eventos (installs respectivamente) es muy baja.

Profundizando el análisis anterior, se calcula una relación entre cantidad de eventos e instalaciones de aplicaciones exponiendola a lo largo del tiempo como una tasa de convergencia.



A la primera hora de la tarde se puede ver un pico en donde la tasa de installs-events es la más elevada y la inversa sucede temprano a la mañana. Igualmente la tasa es muy baja.

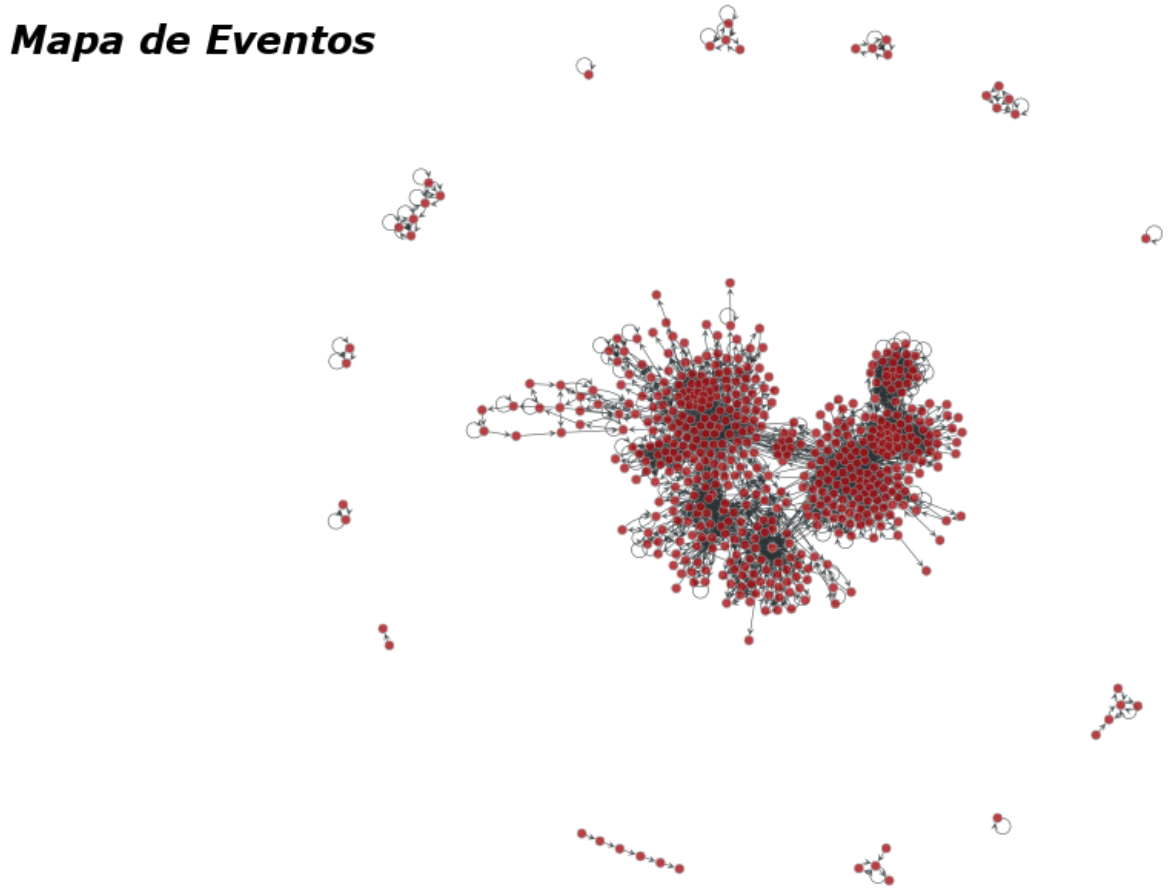
Los installs y los eventos están relacionados, cruzando los set de datos se puede apreciar que los installs son atribuidos a eventos, al parecer no siempre el mismo. Pero la pregunta es **cuándo** sucede la instalación de la App. Se arman listas con la cadena de eventos que realiza un usuario dentro de cada App, y se cuentan donde ocurre la instalación.



El resultado es que la mayoría suceden en la primer evento, pero hay varias secuencias con repeticiones de installs. Esto puede deberse a que hay usuarios que tuvieron problemas y reinstalaron la aplicación varias veces o a algún tipo de fraude en el contador provocado por reinstalaciones programadas.

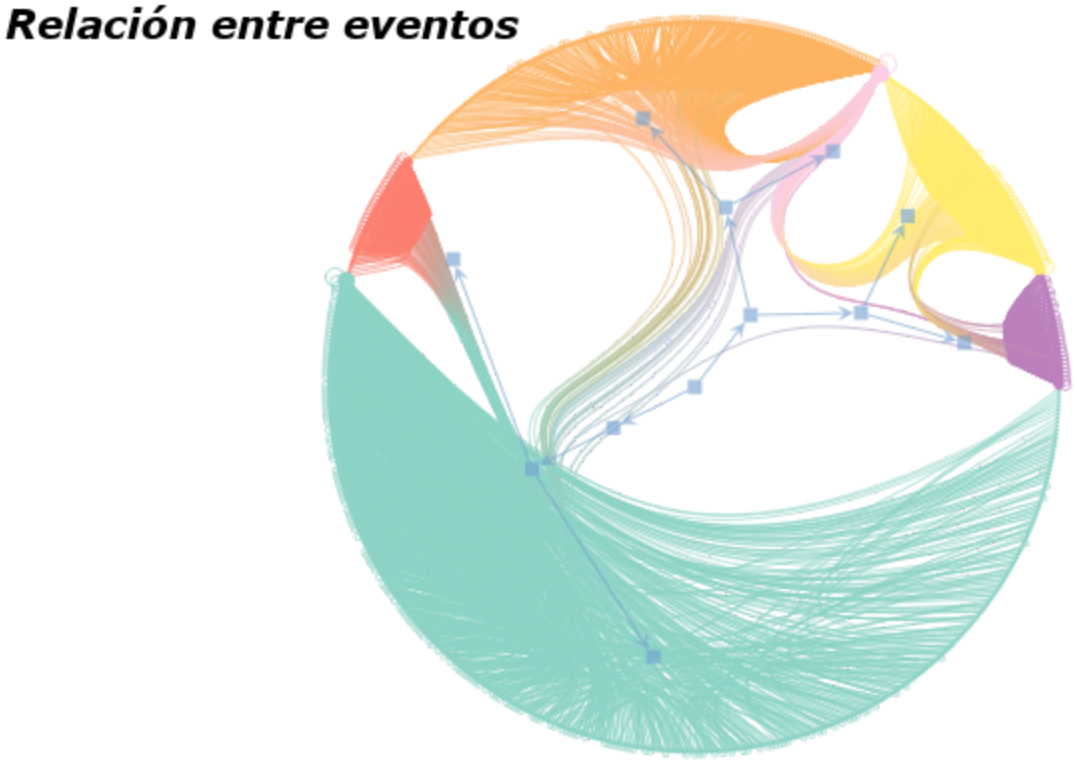
Movimiento de usuarios en las apps

Los usuarios a lo largo del tiempo van moviéndose dentro de las apps registrándose de esta manera eventos.
Se puede apreciar entonces un flujo que empieza por lo general cuando un usuario instala la app (como se vio previamente).
Se arma un grafo donde cada vértice es un evento y las aristas nos dan la relación que hay entre estos (cantidad de veces que los usuarios desencadenaron estos eventos de forma consecutiva dentro de una misma app).



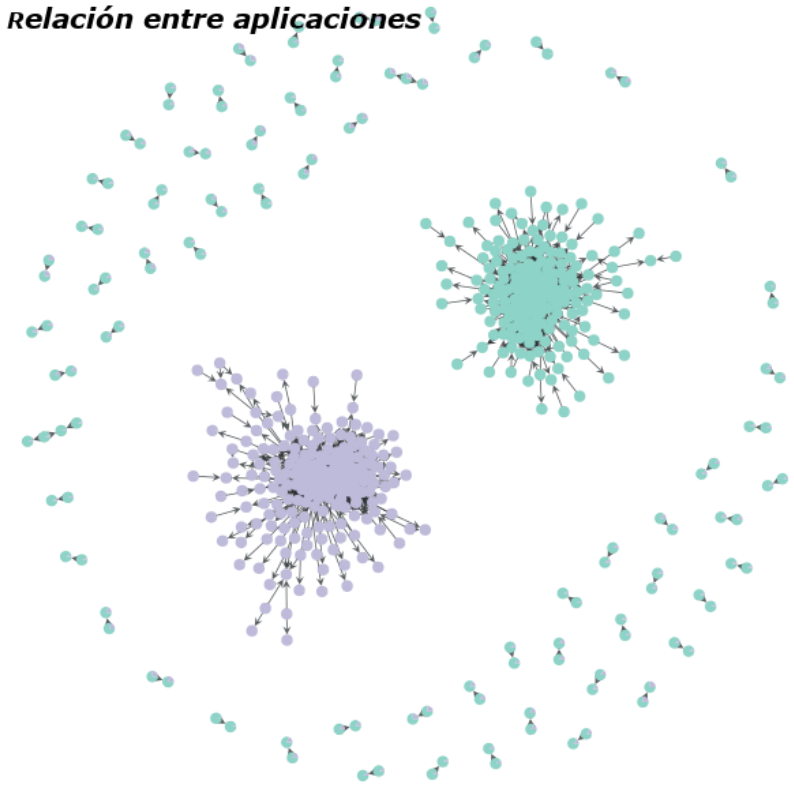
En el grafo se pueden ver dos grandes grupos de eventos interconectados y un órbita de eventos aislados. Esto se puede deber a algunas aplicaciones con eventos muy característicos.

Armando comunidades definiendo el número en seis (en base a las agrupaciones visibles en el gráfico anterior), se disponen los colores. El algoritmo utilizado es (*) Nonparametric Statistical Inference, el cual es, en simples términos, un procedimiento basado en el cálculo de probabilidad de pertenencia de un vértice a un grupo.



Se pueden ver, más allá de los grupos de colores, dos grandes bloques.

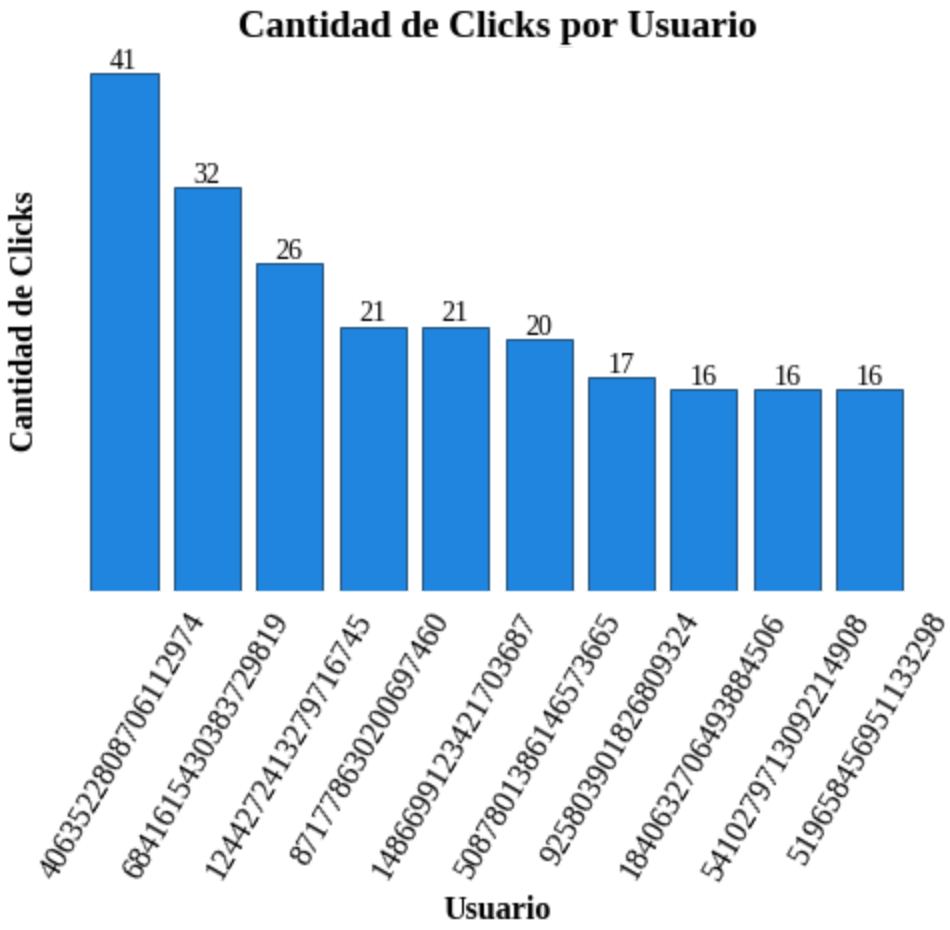
De la misma manera que se armó el grafo con el flujo de eventos que hacen los usuarios en las distintas apps clientes, también se elabora un grafo con los saltos que hacen los usuarios de aplicación en aplicación.



La división que se veía entre los eventos parece deberse a que hay dos grande grupos de aplicaciones. Debido a que los datos son anónimos no se puede saber a que hace referencia. El cálculo de probabilidades asignó los subgrafos disconectos al turquesa.

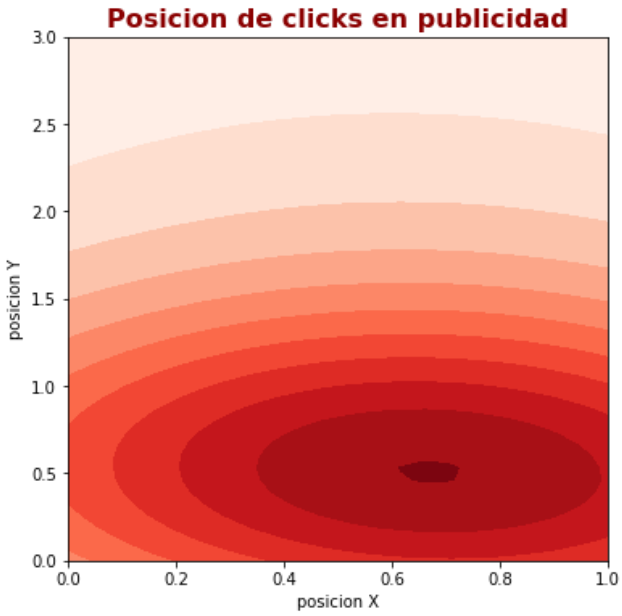
Análisis de los Clicks

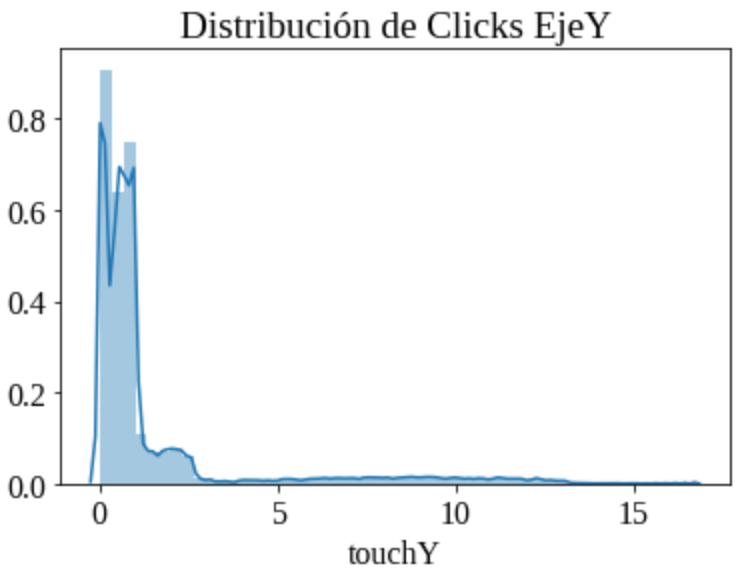
Hay usuarios que son más propensos a hacer clicks en publicidades

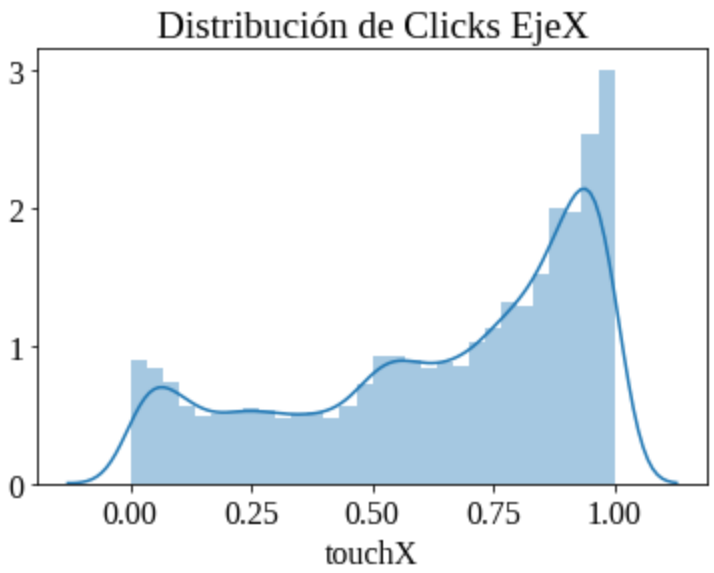


Cuánto tarda un usuario en hacer click. Se filtran, los usuarios con pocos clicks para evitar volatilidad en el promedio.

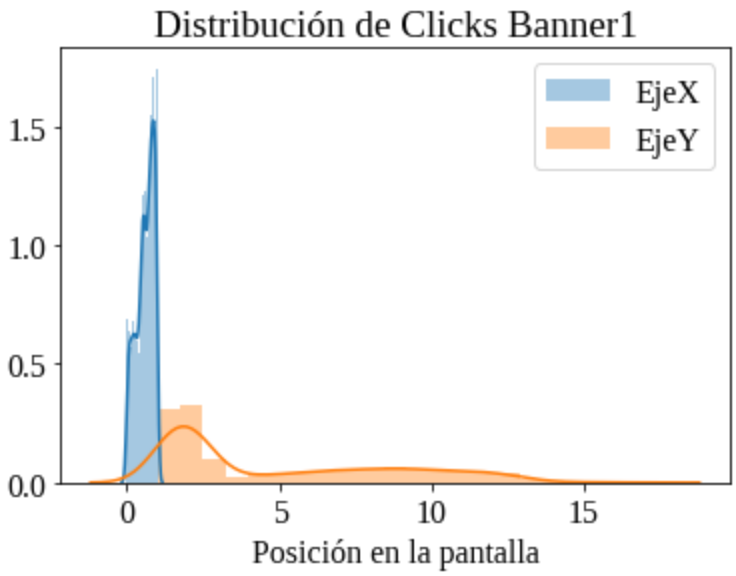
Dónde es que los usuarios hacen clicks. En el dataset se dispone la posición del píxel en el que se clickea la publicidad, el siguiente gráfico que muestra la concentración de los clicks en la pantalla. Cabe destacar que se acotaron los valores a mostrar a aquellos menores que tres ya que representan la mayoría

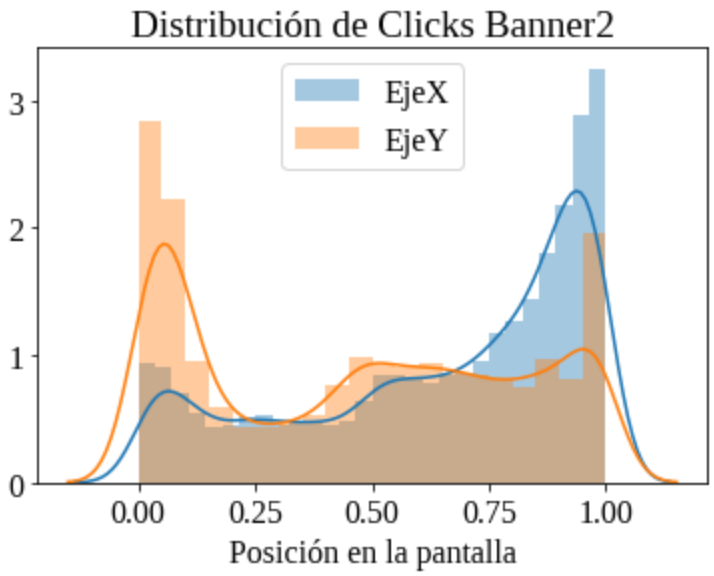




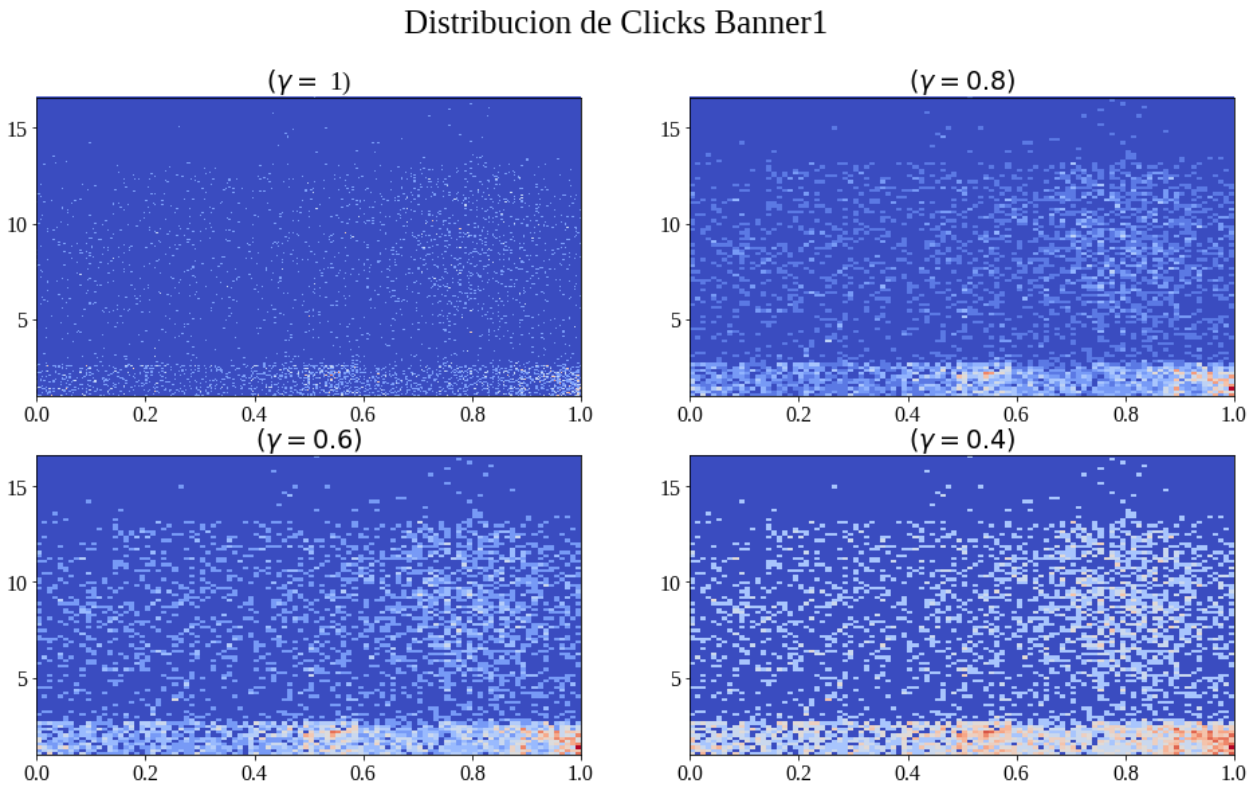


Pareciera ser que el rango del Eje Y, a diferencia del X, no está entre 0 y 1. Esto podría deberse a diferentes tipos de banners publicitarios. Se procede a agrupar en dos tipos.

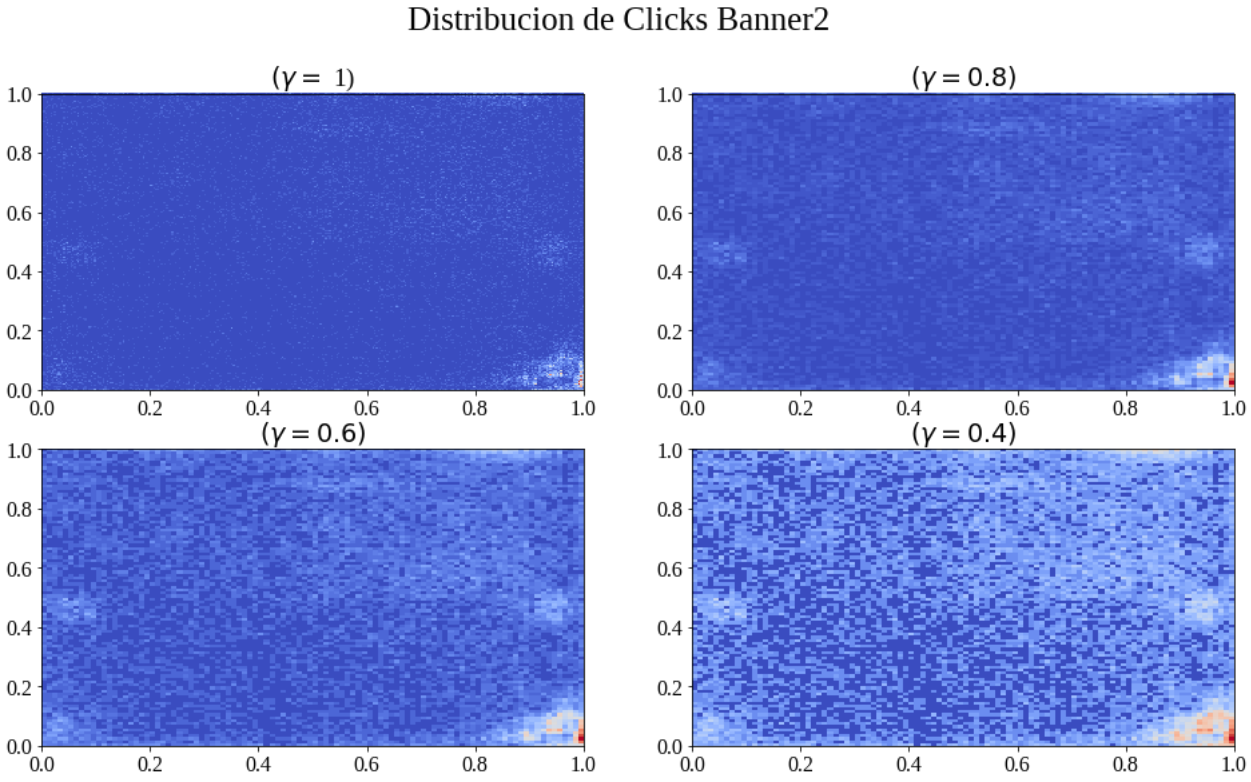




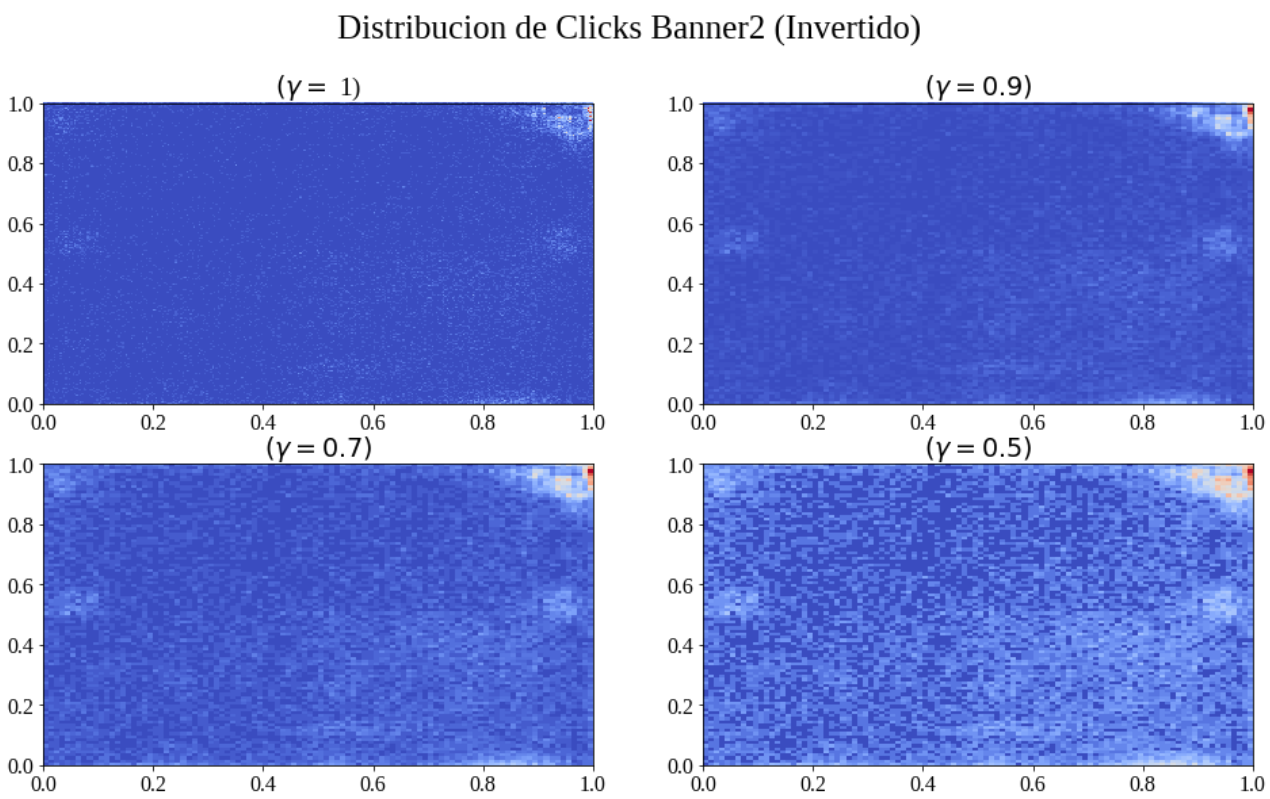
En el banner1 se supone una dimensión de 1x16 y en el banner2 se supone una dimensión de 1x1. Lamentablemente va a haber algunos puntos que estén entre {0,1}x(0,1} que posiblemente pertenezcan al banner1, pero debido a que la mayoría son de 1x1 se optó por imputarlos a esta franja.



Dentro de este grupo de 1x16 también se observa un posible subgrupo más por la franja caliente en la base de los recuadros.

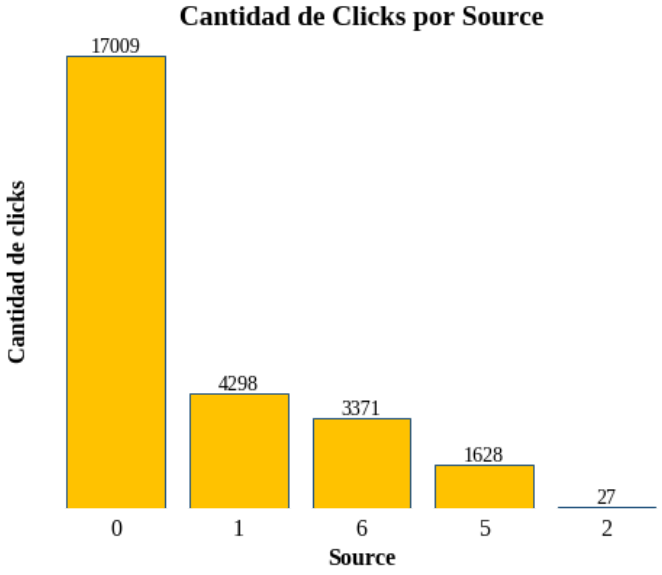


En el caso del grupo Banner2 se puede ver claramente una tendencia a clickear sobre la esquina y un zona caliente del lado derecho de la pantalla. Esto podría deberse a que se invirtieron los valores del Eje Y en la transformación que se le aplicó.



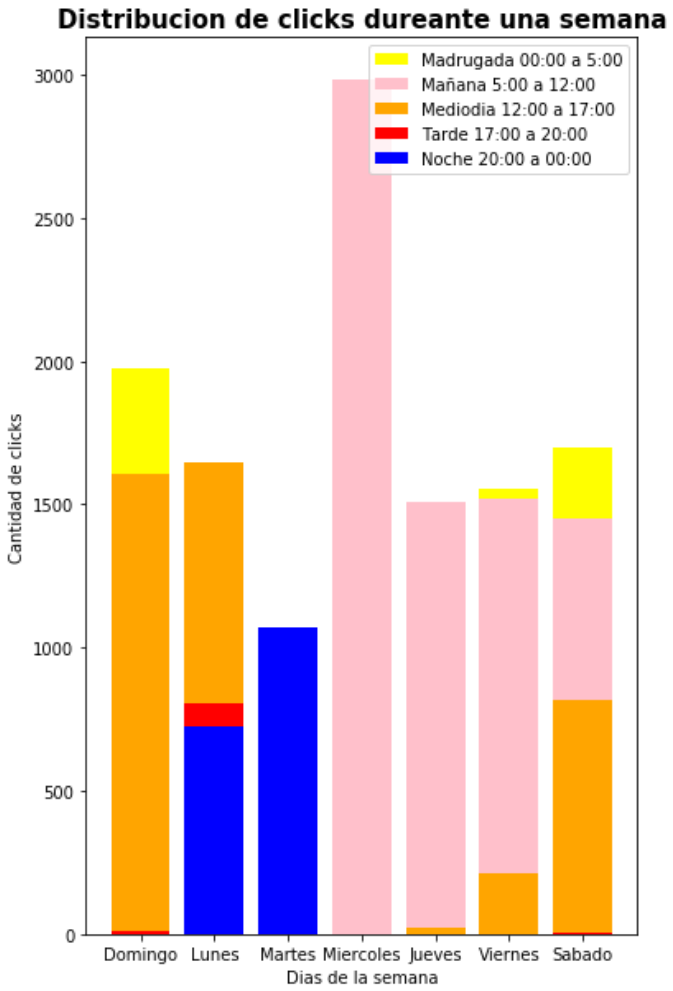
Con el eje Y invertido, la disposición de clicks tiene pinta al touch de un celular (manejado con la mano derecha) y la acumulacion de puntos en la esquina superior podría darse por clicks, no intencionados, al tratar de cerrar la publicidad.

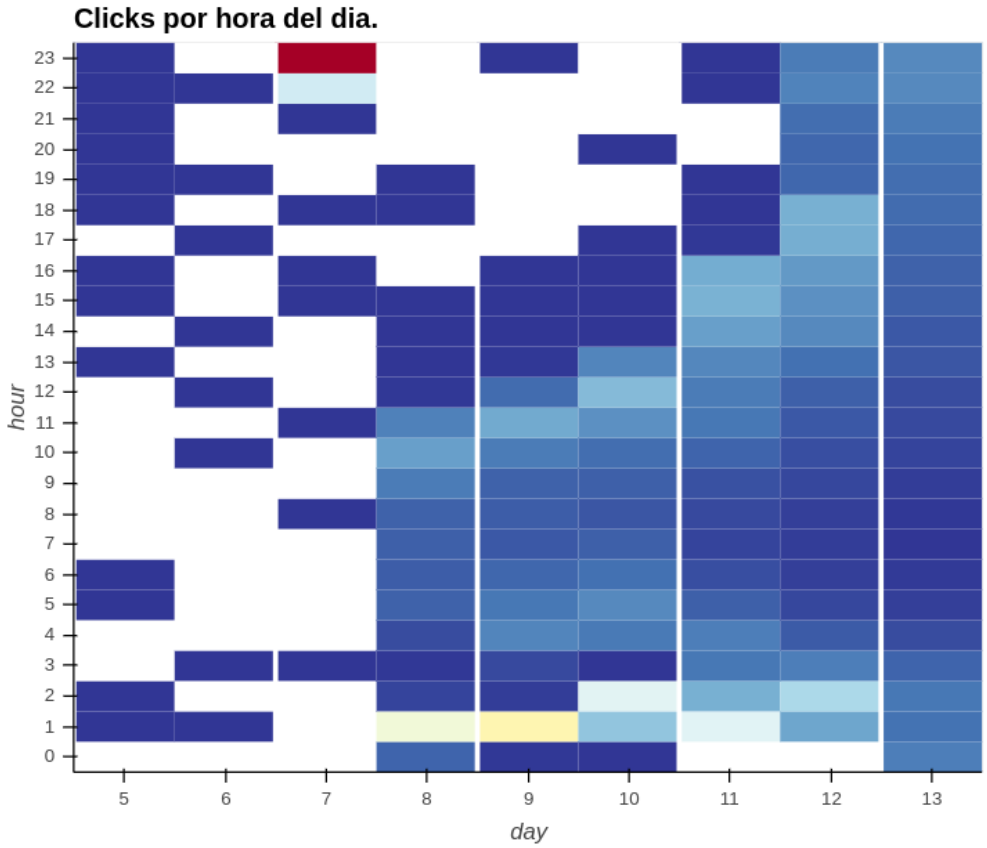
Tanto los advertisers como los sources estan muy concentrados en pocos ids.



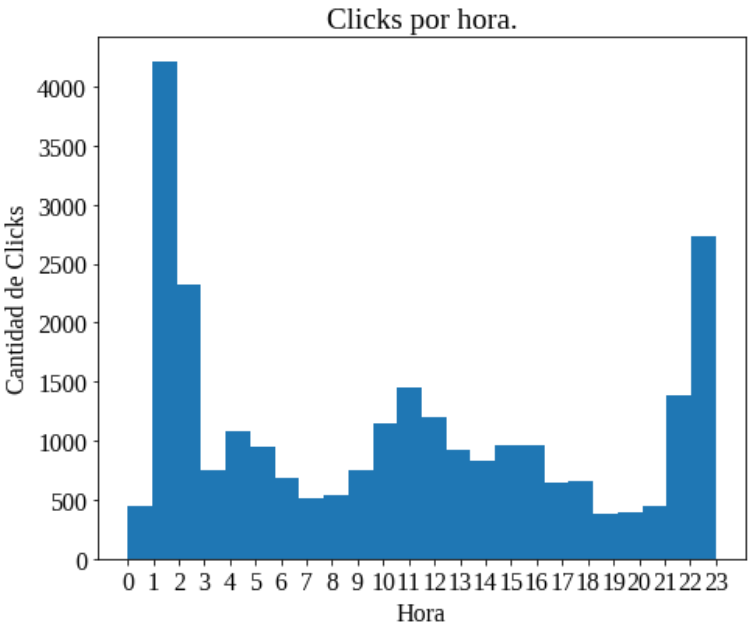
La mayoría de los clicks son de un solo advertiser.

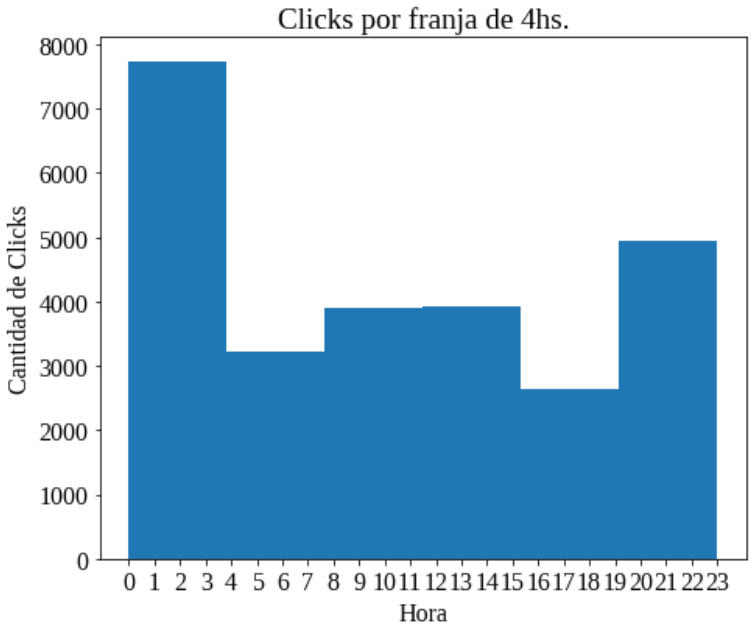
La distribución de clicks a lo largo de la ventana de tiempo dada es muy distinta a lo que se pudo ver en el caso de los eventos e instalaciones.



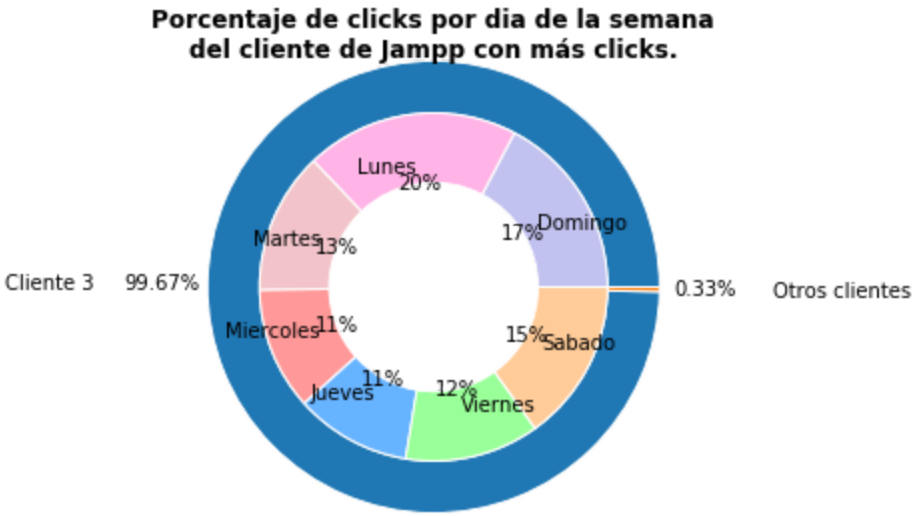


Los clicks son muy dispersos en el tiempo y hay momentos precisos en los cuales la cantidad de clicks se dispara, posiblemente tenga que ver también la metodología con la se realizó el sample de datos.



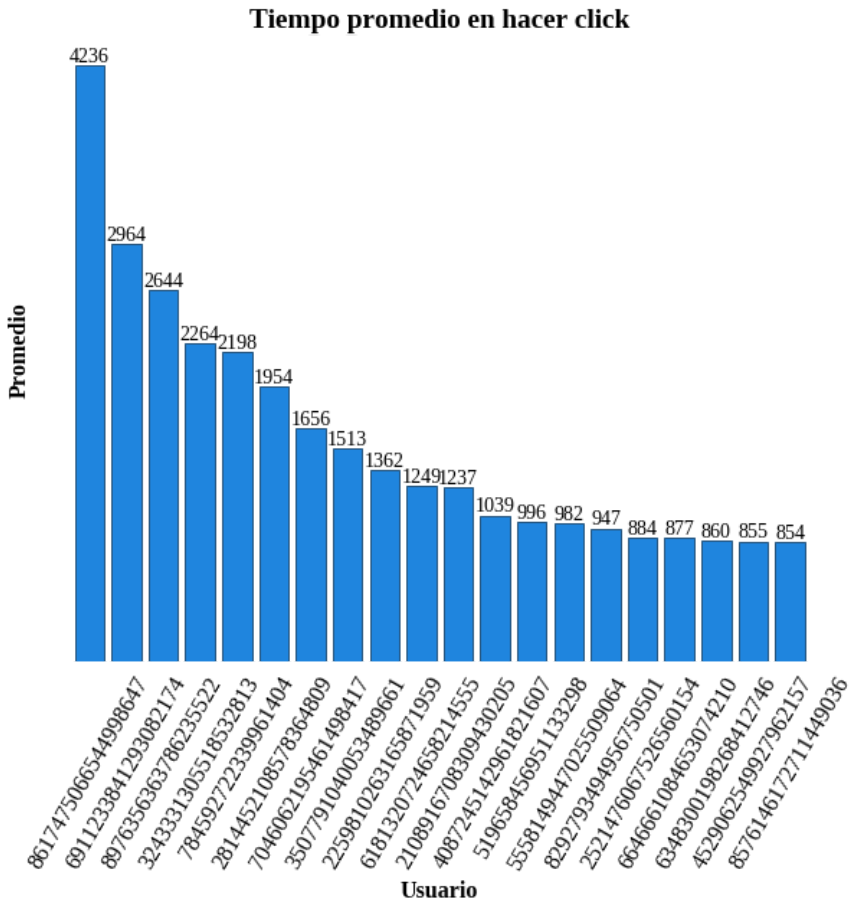


Dentro de los clientes de jampp se analiza al que tiene mayor porcentaje de clicks. Hay una diferencia muy grande con el resto, la mayoría pertenece a un solo advertiser.

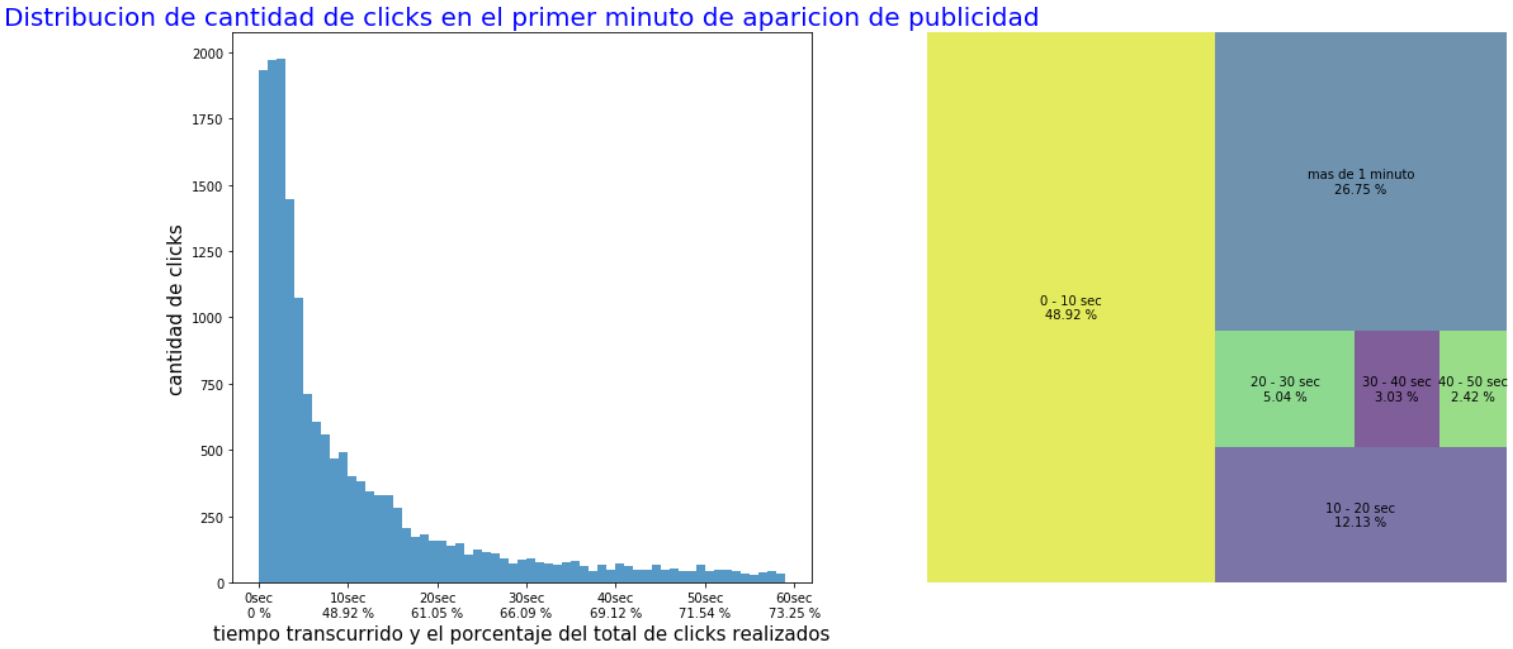


Se observa que o bien los clicks realizados son dominados por un solo cliente, o bien los datos con los que se trabajaron perteneces solo a los de tal cliente.

Analizamos cuánto se tarda en realizar un click.



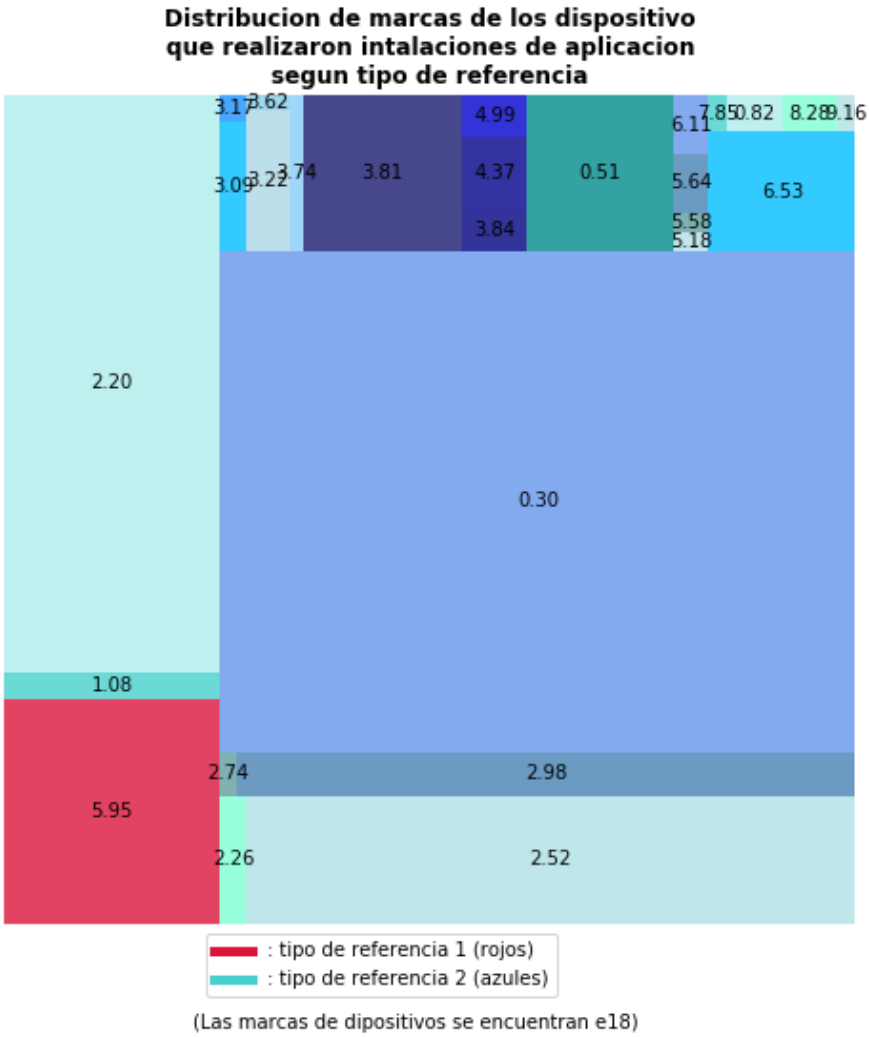
Para el cálculo del promedio se tomaron usuarios con repetidas ocurrencias



En el gráfico de la derecha, se analiza los datos del primer minuto ya que notamos que en ese periodo se realizan el 75% de los clicks y que va decreciendo la cantidad en función del tiempo.
En el segundo gráfico tenemos un análisis general en el cual hacemos énfasis en los porcentajes de clicks realizados en intervalos de 10 segundos hasta llegar al minuto. Se consideró agrupar los clicks realizados después de un minuto ya que sus valores no son uniformes.

Estos datos resultan interesantes, ya que si no se quiere saturar al consumidor de la publicidad parece innecesario que la publicidad dure más de un minuto.

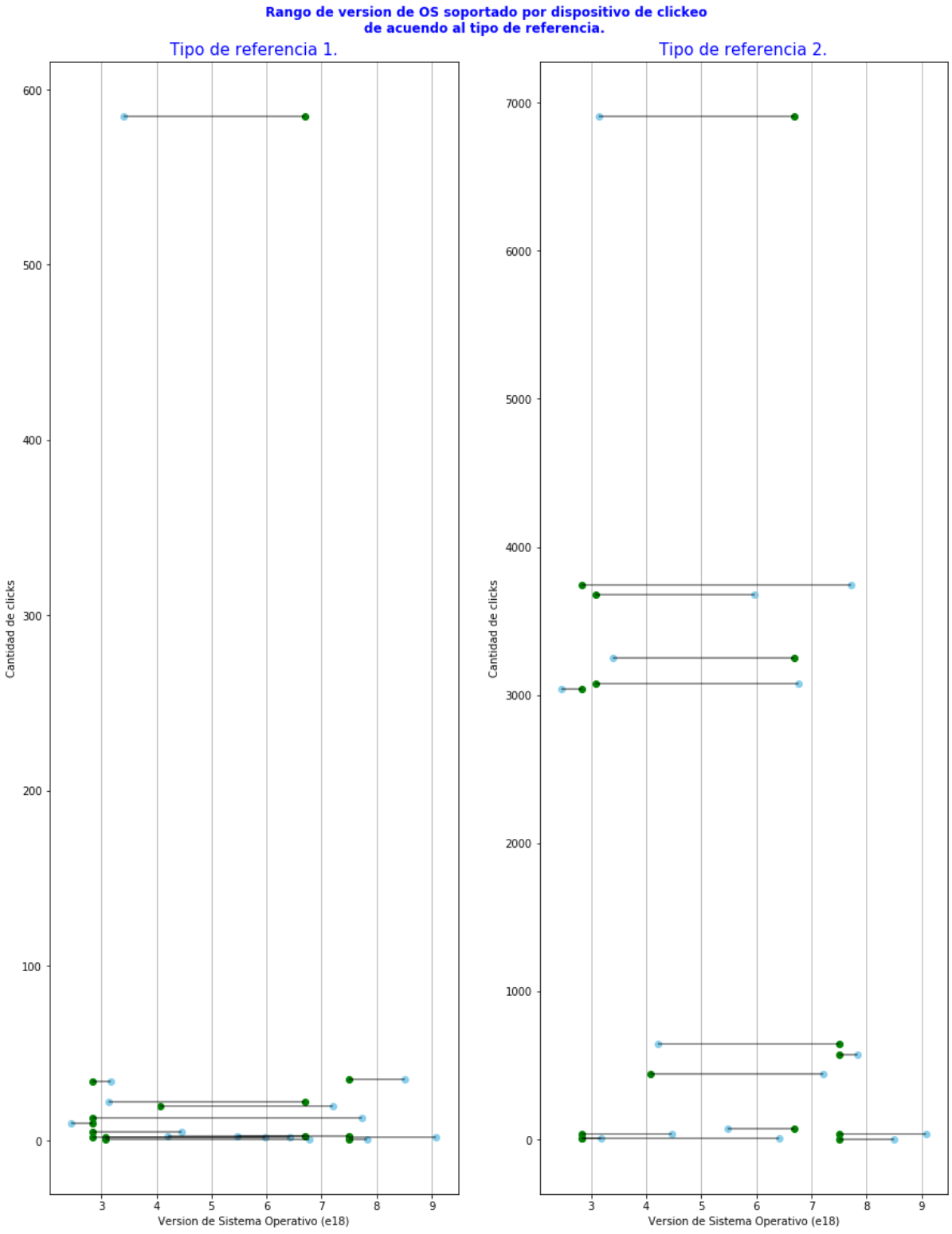
Luego, se buscó analizar la información de los dispositivos que realizaron los clicks. Se realizó un gráfico que analiza los modelos de dispositivos más comunes que realizaron clicks. Pero, además, como se sabe que la columna con los datos “type_ref” hace referencia a que los clicks fueron realizado por motores de publicidad de Google o de Apple. Por ende, se coloreó de rojo los modelos de un tipo de referencia y de azules/celestes los del otro tipo. Más aún, para una mejor visualización de los datos, nos quedamos con solo 3 cifras significativas del código hasheable que representan los distintos modelos de dispositivo.



(Las marcas de dipositivos se encuentran e18)

En el gráfico, se observa que el tipo de referencia 2 predomina sobre el tipo de referencia 1. Además, se observa que 4 modelos (2.20e18, 0.30e18, 2.52e18 y 5.95e18) abarcan la mayoría.

Además, analizamos el rango de versión del sistema operativo de los dispositivos que realizaron los clicks con el objetivo de ver que tipo de formato debería tener una publicidad para que la soporte la mayoría de los dispositivos pertenecientes a futuros consumidores. Para un análisis aprovechable, dividimos los datos en dos gráficos según “ref_type” ya que las versiones sistema operativo de Google y de Apple suelen variar.



Se destaca en las visualizaciones que el rango más común es el que abarca un rango intermedio: ni versiones muy viejas ni versiones muy nuevas.

Por otro lado, observamos que ningún clicks fue realizado con wifi y por eso se analizo las compañías móviles utilizadas para realizar los clicks.

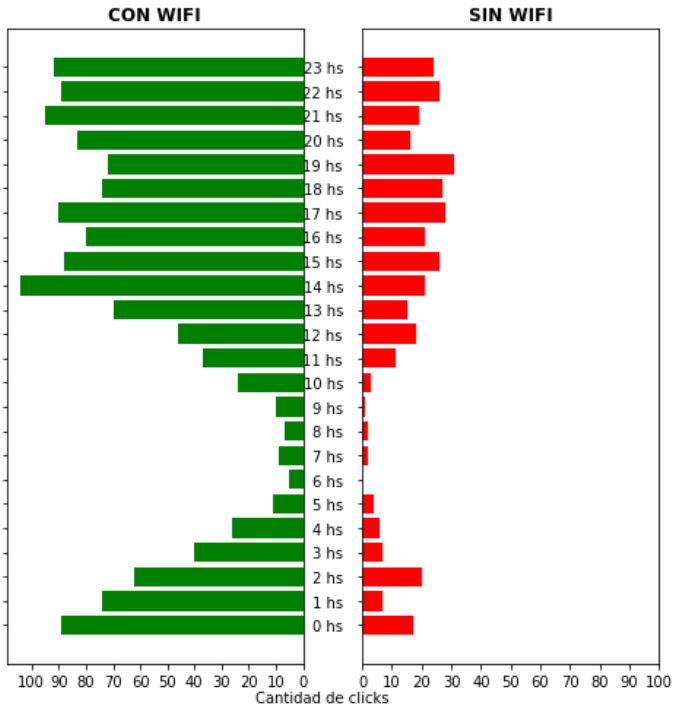
Compañias moviles mas utilizadas
por usuarios que clickearon las publicidades

Se observó que la distribución de clicks por empresa es bastante uniforme y, por lo tanto, no tiene utilidad.

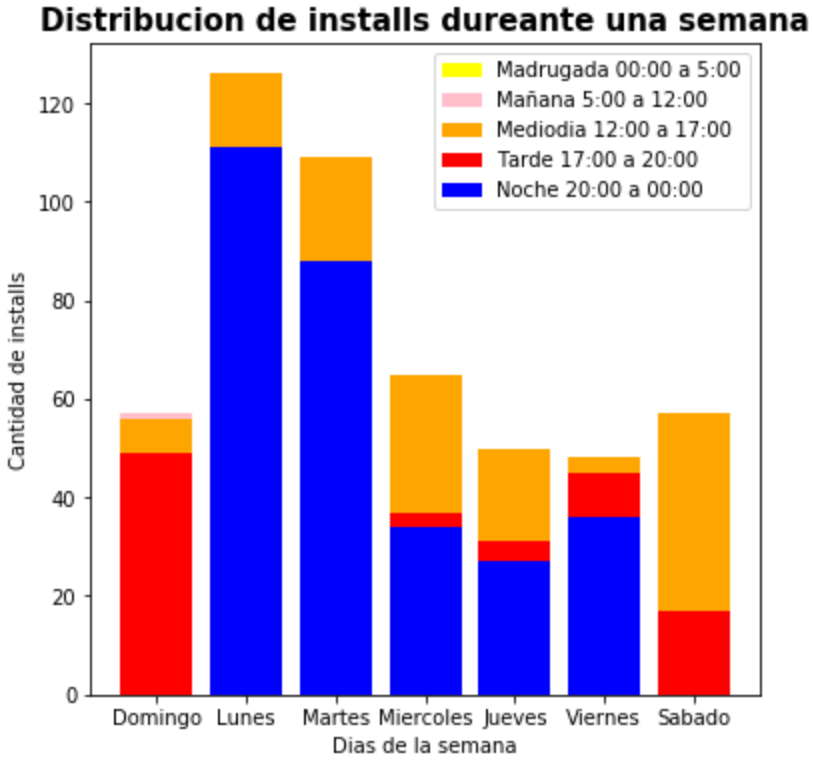
Análisis de los Installs

Analizamos los installs que representan otro servicio que ofrece la empresa Jampp. Se muestran la cantidad de instalaciones que se realizan por hora e identificamos si aquellas instalaciones se realizaron con o sin wi-fi.

Cantidad de instalaciones por hora del dia.

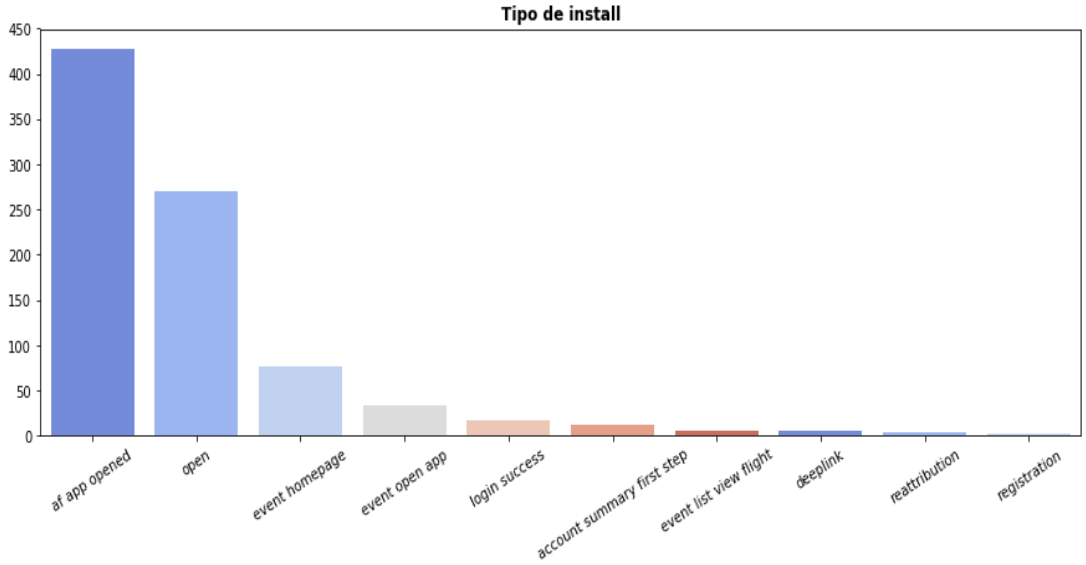


Se observa que la mayoría de los clicks se realizan con wi-fi y además que de las 4 am a 11 am es la franja horaria con menos instalaciones tanto con como sin wi-fi. Asimismo analizamos los installs por día de la semana y turno del día.



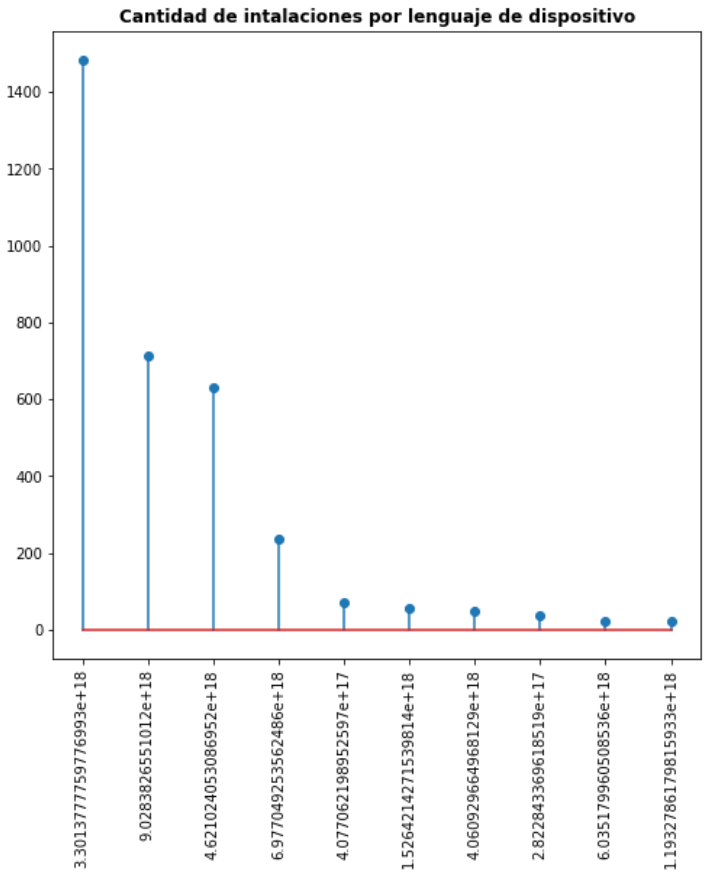
Se observa, que la gran parte de las instalaciones se realizan en general de noche y al mediodía.

Por otro lado, analizamos los tipos de instalaciones y realizamos un gráfico de los 10 mayores tipos.



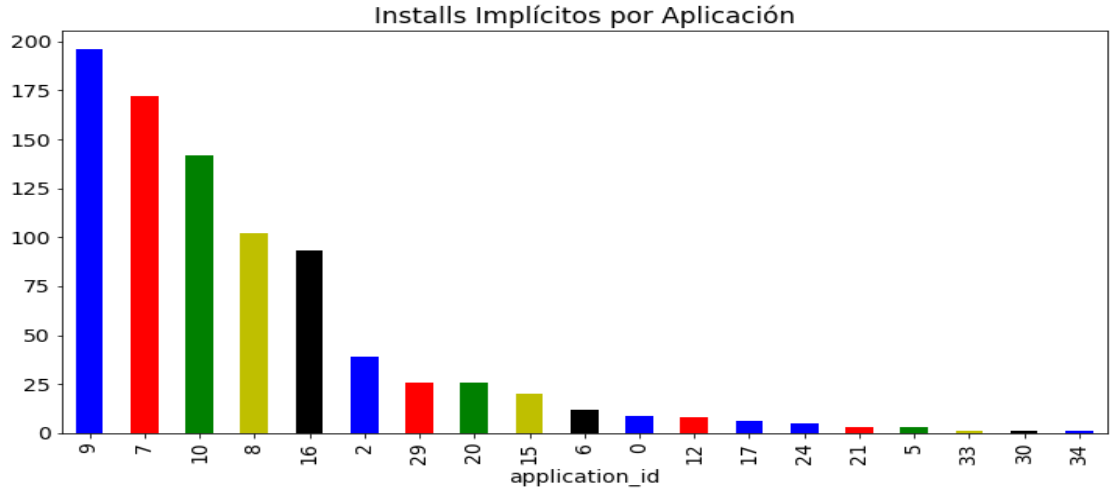
Observamos que el tipo de install más común es el relacionado en abrir la aplicación o instalación.

Más aún, analizamos los lenguajes de los dispositivos que realizaron la instalación.



Notamos que son 4 lenguajes que predominan las instalaciones lo que es lógico ya que anteriormente se verificó que todas las instalaciones proviene de un mismo país.

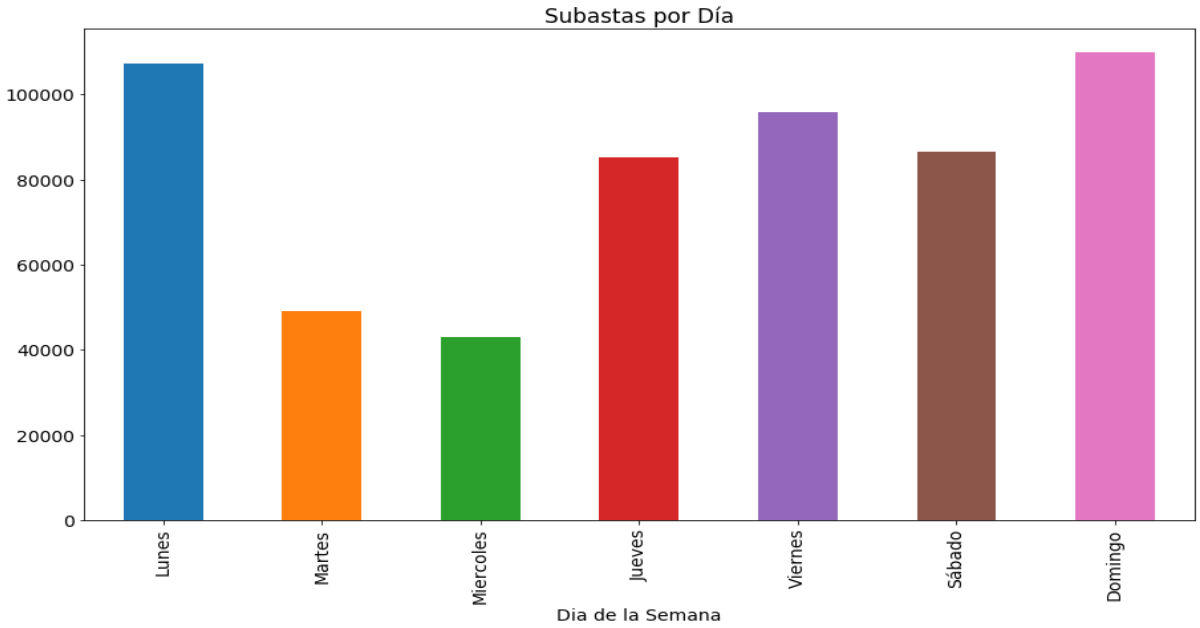
Analizamos también cuántos installs, según cada aplicación, fueron implícitamente atribuidos a Jampp (es decir, el install provino desde una subasta ganada por Jampp, pero la misma no le fue atribuida).



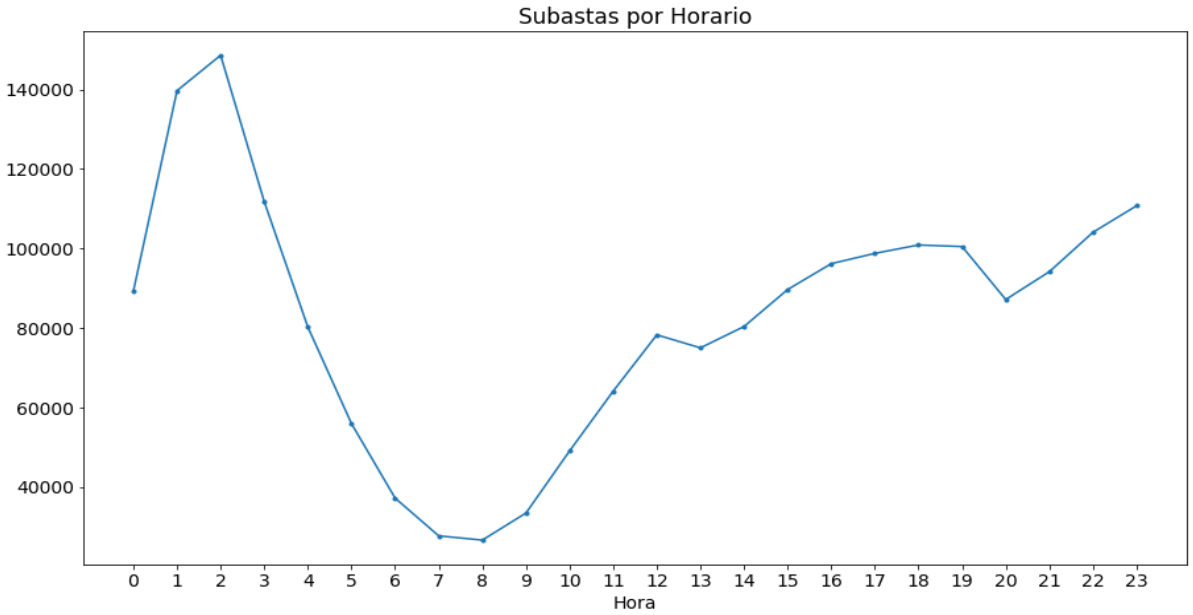
Se destacan claramente cinco aplicaciones que se llevan casi la totalidad de las aplicaciones que se realizaron desde Jampp.

Análisis de las Subastas

Pensamos que sería interesante conocer la distribución de las subastas según el día de la semana y el horario en el que fueron generadas, con el objetivo de destacar los días y horas más relevantes.

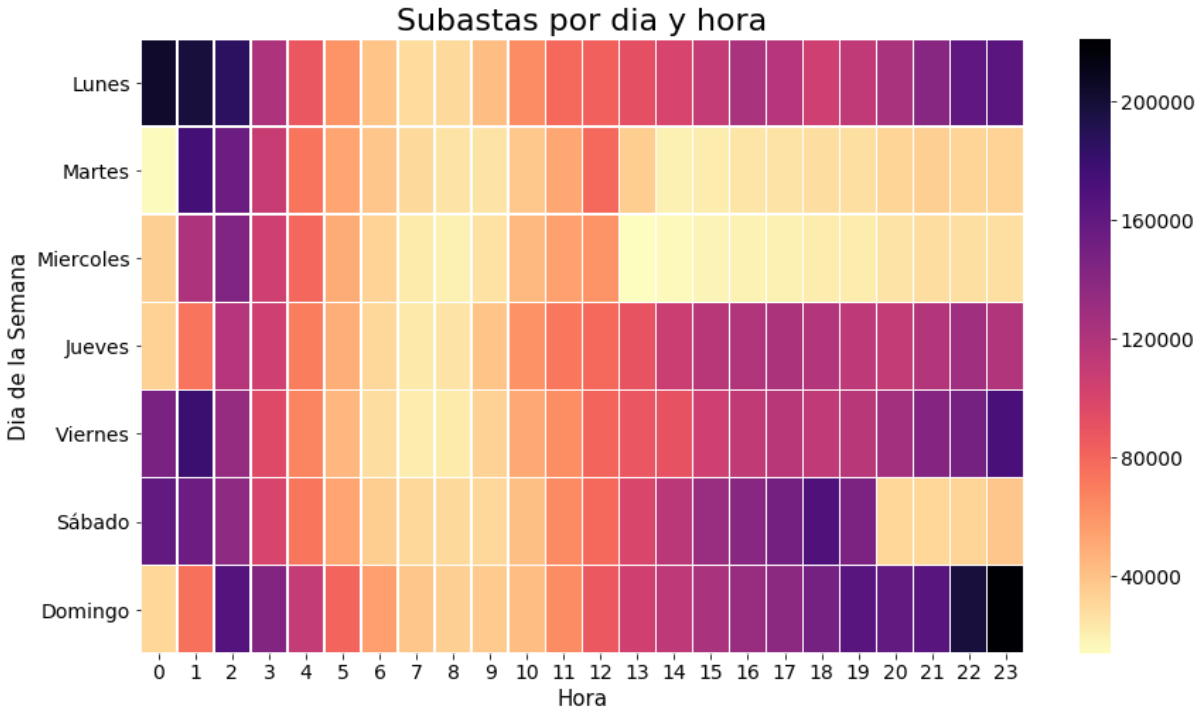


Hay un marcado descenso de generación de eventos los días martes y miércoles, siendo los domingos y los lunes claramente los días de mayor movimiento en cuanto a la generación de subastas.



En cuanto a los horarios, se marca un claro pico de generación de subastas entre la una y las tres de la mañana, lo que cobra un mayor sentido si se toma la información de que los días más relevantes son los fines de semana, lo que explica la mayor actividad nocturna.

Para reflejar esto, pensamos en el siguiente mapa de calor, que combina la información de los dos cuadros anteriores.



La franja horaria que se mantiene relativamente estable en cuanta a la poca aparición de subastas es la que va entre las 6 y las 10 de la mañana, lo cual es lógico para esos horarios. Se marca un claro incremento de subastas la mayoría de los días, con una excepción difícil de analizar como es la del sábado entre las 20 y las 24 (o incluso, la 1 del domingo), que podría llegar a estar relacionada a algún fallo en el envío o en la recolección de información.

Conclusiones

1) Los usuarios con la actividad dentro de las aplicaciones arman un flujo de eventos. Este flujo es lógico a las aplicaciones preferidas por el usuario.
2) De la misma manera que los usuarios registran actividades dentro de las aplicaciones, también van saltando de aplicación en aplicación armando una correlación entre las mismas. De este estudio podrían derivar posibles recomendaciones a customers siempre teniendo en cuenta la competencia, la cual por el anonimato de los datos no se puede salvar.
3) Las ips, al ser públicas, nos brindan información interesante de cómo se comportan los usuarios en diferentes ambientes. Por ejemplo en un shopping un customer no tiene el mismo comportamiento que en su residencia privada. Hay aplicaciones, muy populares, que tienden a ser más concentradas en pocas ips.
4) Los eventos, installs, clicks y auctions muestran franjas horarias de actividad alta y baja. Los clicks son los más volátiles, posiblemente se deba a comportamientos compulsivos atribuibles a campañas publicitarias.
5) Los installs indican que es su mayoría se realizaron utilizando wi-fi, de tipo: ‘open’ y por la noche, pero la mayor proporción de installs vs eventos se da durante el mediodía.

Bibliografía y Links a las herramientas

http://holoviews.org/
https://bokeh.pydata.org/en/latest/
https://graph-tool.skewed.de/
https://matplotlib.org/
https://seaborn.pydata.org/
https://docs.python.org/3/
(*) https://graph-tool.skewed.de/static/doc/demos/inference/inference.html#inference-howto