FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Trabajo Práctico N° 1: Análisis Exploratorio de Datos Jampp

Organización de Datos [75.06 / 95.58]
Primer Cuatrimestre 2019

Grupo N° 15

Fecha de Entrega: 15 de Abril de 2019

Nombre	Padrón
Escobar Benitez Maria Soledad	97877
Elías Federico	96105
Jorge Luis Anca	82399
Diego Peyrano	94254

Índice

1 - Introducción

1.1 - Objetivo del Trabajo Práctico

El objetivo de este trabajo práctico es aplicar las herramientas vistas en clase para realizar un análisis exploratorio sobre los sets de datos de Jampp brindados por la cátedra.

La idea principal es investigar los datos y ver qué conclusiones interesantes se pueden deducir de los datos.

1.2 - Sobre la Empresa

Jampp es una plataforma líder para la promoción y remarketing de aplicaciones móviles.

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.

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.

Esto significa que no sólo se dedican a entregar instalaciones a sus clientes sino que además realizan optimizaciones para lograr aquello que los anunciantes buscan obtener a través de sus apps: compras, pedidos de comida, publicaciones de clasificados, pedidos de taxis, etc. El marketing de aplicaciones ya no se trata de conseguir instalaciones sino de lograr que esas instalaciones se conviertan en usuarios activos.

Para lograr maximizar el porcentaje de conversiones y minimizar su costo, Jampp utiliza algoritmos de Machine Learning para decidir en menos de 70 ms en cuáles de las más de 700.000 subastas por segundo va a participar y a qué precio máximo.

1.3 - Sobre los Datos

Entre los datos que nos proporciona podemos encontrar los siguientes set de datos:

- Installs (instalaciones): Registro de las instalaciones realizadas luego de una publicidad.
- Clicks: registro de los clicks realizados a las publicidades.
- Events (eventos): cualquier tipo de acción categorizada dentro de una aplicación. Por ejemplo, en una aplicación de e-commerce un funnel de eventos muy común puede ser del estilo “abrir_app” → “ buscar_producto” → “revisar_catalogo” → “agregar_a_carrito” → “ efectuar_compra”. Cada uno de estos pasos es un evento.
- Auctions (subastas): 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.

Cada uno de ellos contiene las siguientes columnas:

- Installs:
  - 'created': Fecha de instalación
  - 'application_id': Identificación de la aplicación instalada.
  - 'ref_type': Puede ser apple_ifa o google_advertising_id.
  - 'ref_hash': apple_ifa o google_advertising_id del dispositivo.
  - 'click_hash': Es una especie de identificación de la instalación.

- 'attributed': Indica si la instalación se atribuye a Jampp o no.
- 'implicit': Indica si la instalación es implícita (la instalación fue realizada por un dispositivo que no se ha instalado de acuerdo con la plataforma de seguimiento).
- 'device_countrycode': Código para el país donde se encuentra el dispositivo.
- 'device_brand': Marca del dispositivo.
- 'device_model': modelo del dispositivo.
- 'session_user_agent': agente de usuario que se utilizó para la instalación.
- 'user_agent': agente de usuario relacionado al dispositivo.
- 'event_uuid': uuid generado por el evento.
- 'kind': Tipo de instalación.
- 'wifi': marca booleana que indica si la instalación se realizó desde una conexión Wifi.
- 'trans_id': id de la transacción.
- 'ip_address': dirección ip através de la cual se realizo la instalación.
- 'device_language': Lenguaje relacionado al dispositivo.

- Clicks:
  - 'advertiser_id': Identificación interna del anunciante, el cliente de Jampp que paga por el anuncio.
  - 'action_id': Identificación interna de la acción.
  - source_id': Identificación interna de la fuente (intercambio) desde la cual se originó el click.
  - 'created': idem created de installs
  - 'country_code': código del país.
  - 'latitude': Latitud estimada desde donde se realizó el click.
  - 'longitude':Longitud estimada desde donde se realizó el click.
  - 'wifi_connection': Describe si el clic se realizó durante una conexión wifi.
  - 'carrier_id': id para el operador de telefonía móvil del dispositivo.
  - 'trans_id': identificación interna de la transacción.
  - 'os_minor': Versión menor del sistema operativo.

- 'agent_device': Agente para el dispositivo desde el que se realizó el click (modelo).
- 'os_major':versión principal para el sistema operativo.
- 'specs_brand':
- 'brand': Marca del dispositivo
- 'timeToClick': Momento en que se realizó el click, en segundos.
- 'touchX': Posición X para el click.
- 'touchY': Posición Y para el click.
- 'ref_type': idem installs
- 'ref_hash': idem installs

- Events:
  - 'date': Momento de creación del evento, fecha y hora.
  - 'event_id': Identificación del evento.
  - 'ref_type': idem installs
  - 'ref_hash': idem installs
  - 'application_id': idem installs
  - 'attributed': idem installs
  - 'device_countrycode': idem installs
  - 'device_os_version': Versión del sistema operativo.
  - 'device_brand': idem installs
  - 'device_model': idem installs
  - 'device_city': Ciudad en que se encuentra el dispositivo.
  - 'session_user_agent': agente de usuario que se utilizó para la instalación.
  - 'trans_id': idem installs
  - 'user_agent': agente de usuario relacionado al dispositivo.
  - 'event_uuid': Identificador del evento.
  - 'carrier': operador de telefonía movil.
  - 'kind': Tipo de evento.
  - 'device_os': sistema operativo del dispositivo
  - 'wifi': idem installs.
  - 'connection_type': Tipo de conexión.
  - 'ip_address': dirección ip desde la que se originó el evento.
  - 'device_language': idem installs

- Auctions:
    - 'auction_type_id’: identificador para el tipo de subasta.
    - 'country': país desde el que se originó la subasta.
    - 'date': tiempo de origen de la subasta, fecha y hora.
    - 'device_id': identificador del dispositivo desde donde se originó la subasta.
    - 'platform': si es android o ios.
    - 'ref_type_id': identificación interna para ref_type.
    - 'source_id': identificador de la fuente de la subasta.

2 - Análisis introductorio

Antes de comenzar con el análisis de los datos para cada set de datos específico nos tomamos el tiempo de reconocer los datos, observar tipos de datos, tipo de información que proveen, cantidad de memoria que ocupan y demás detalles que se describirán en esta sección. Los datos se corresponden al periodo comprendido entre el 5 de marzo de 2019 y el 13 de marzo de 2019.

2.1 - Conociendo los datos

La primera impresión que se tuvo al manipular los datos fue que resultan complejos de entender ya que la mayoría de la información almacenada consiste en datos transformados por medio de funciones de hashing. Esto se suma al desconocimiento previo del modelo de negocios de la empresa. Luego de leer la descripción y observar el contenido se logró comprender el significado de cada una de las columnas de los sets de datos.

La segunda impresión tuvo que ver con el volumen de los datos a analizar, se tuvieron que explorar alternativas para aminorar el consumo de memoria por parte de los mismos.

2.2 - Tratamiento de los datos en memoria

Dado el tamaño que ocupaban en memoria los datos y que no contábamos con sistema de procesamiento distribuido se llevó a cabo un estudio detallado de los tipos de datos de cada campo, sus valores y sus requerimientos de espacio. Haciendo este análisis para cada set de datos pudimos reducir en algunos casos hasta el 80 por ciento del tamaño original. Entre las acciones que tomamos se encuentran:

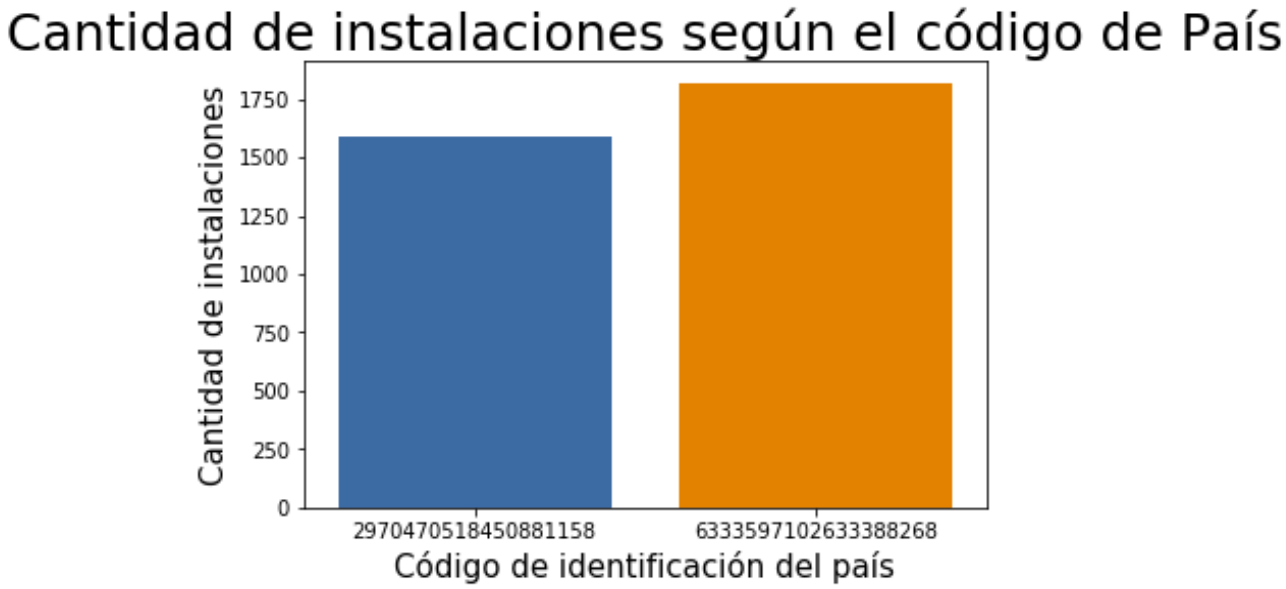
- Eliminación de columnas que no aportaban información ya sea porque éstas estuviesen pocos valores no nulos, porque el campo no se lo consideró relevante o porque todas las filas correspondientes tenían el mismo valor

- Ajuste del tipo de datos de cada campo al requerido según el tipo y rango de valores observado. Como ejemplo: pasar las columnas que contenían fechas y horas que eran almacenadas como tipo Object a tipo datetime, columnas con pocos valores distintos a tipo categóricos, o valores que estaban siendo almacenados como enteros de 64 bits se los definió como enteros de 8 bits.

2.3 - Análisis sobre la Ubicación

Comenzamos observando de qué países provienen nuestros datos, se observa que para los sets Auctions, Events y clicks sólo se tiene un único código de país, el cual corresponde a Uruguay, según nos informó luego el representante de la empresa. Luego observando las instalaciones (set de datos installs), se observan dos códigos de país, uno corresponde al código que encontramos en los otros sets de datos, luego el otro es completamente desconocido.

Al principio se pensó que podría tratarse de un evento aislado, por ejemplo alguna instalación realizada por alguien que viajó y estaba fuera de la ubicación inicial, código de país común en todos los sets de datos, pero luego al analizar la cantidad de instalaciones para cada país notamos que no es así. El el gráfico a continuación se puede observar que la cantidad de instalaciones para el código **6333597102633388268,** correspondiente a Uruguay, no difiere demasiado en la cantidad e instalaciones para el país desconocido, aunque sí se observa que es levemente mayor.

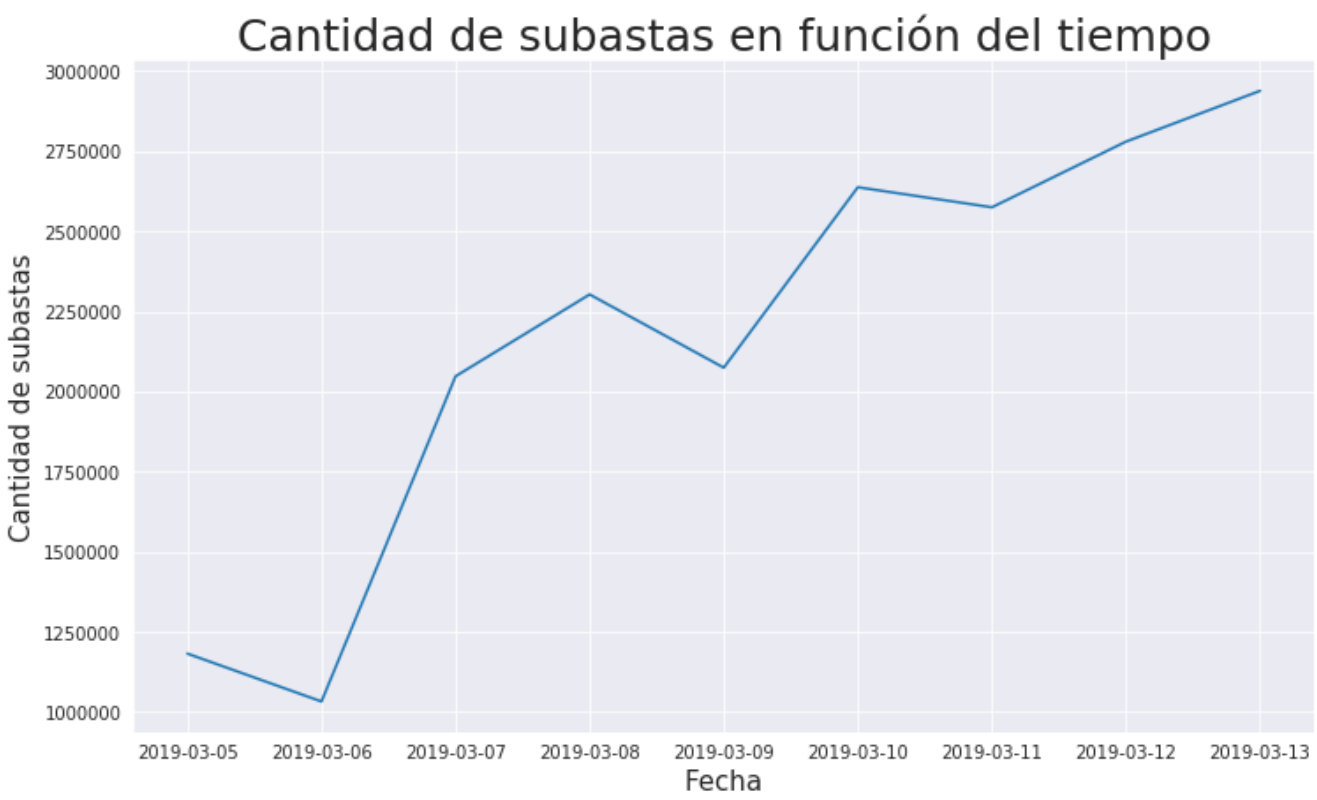


2.4 - Análisis temporal

Cuando se comenzó observando las fechas de los dato se observó que los datos corresponden a fechas entre el 5 de marzo y el 13 de marzo del 2019.

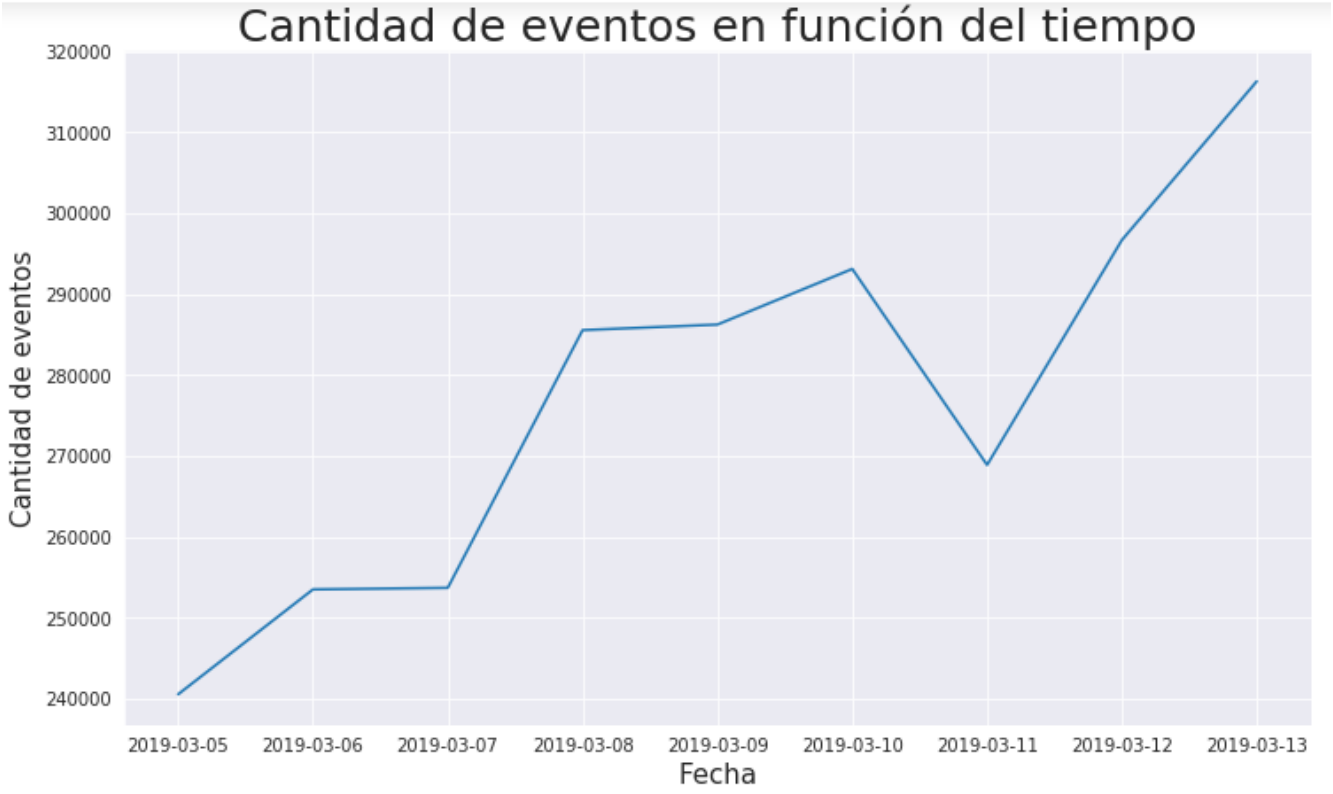
Se procede a realizar un análisis del comportamiento de los diferentes sets datos en función del tiempo. Para realizar este análisis lo que se hace es contar la cantidad de cada tipo, auctions, events, clicks e installs, en función de las fechas en que se suceden. Para ello sólo se toma en cuenta sólo la fecha, lo que permite dar un vistazo más general al comportamiento de los datos. Es importante destacar que el día martes 5 de marzo fue feriado en Uruguay.

Subastas en función del tiempo



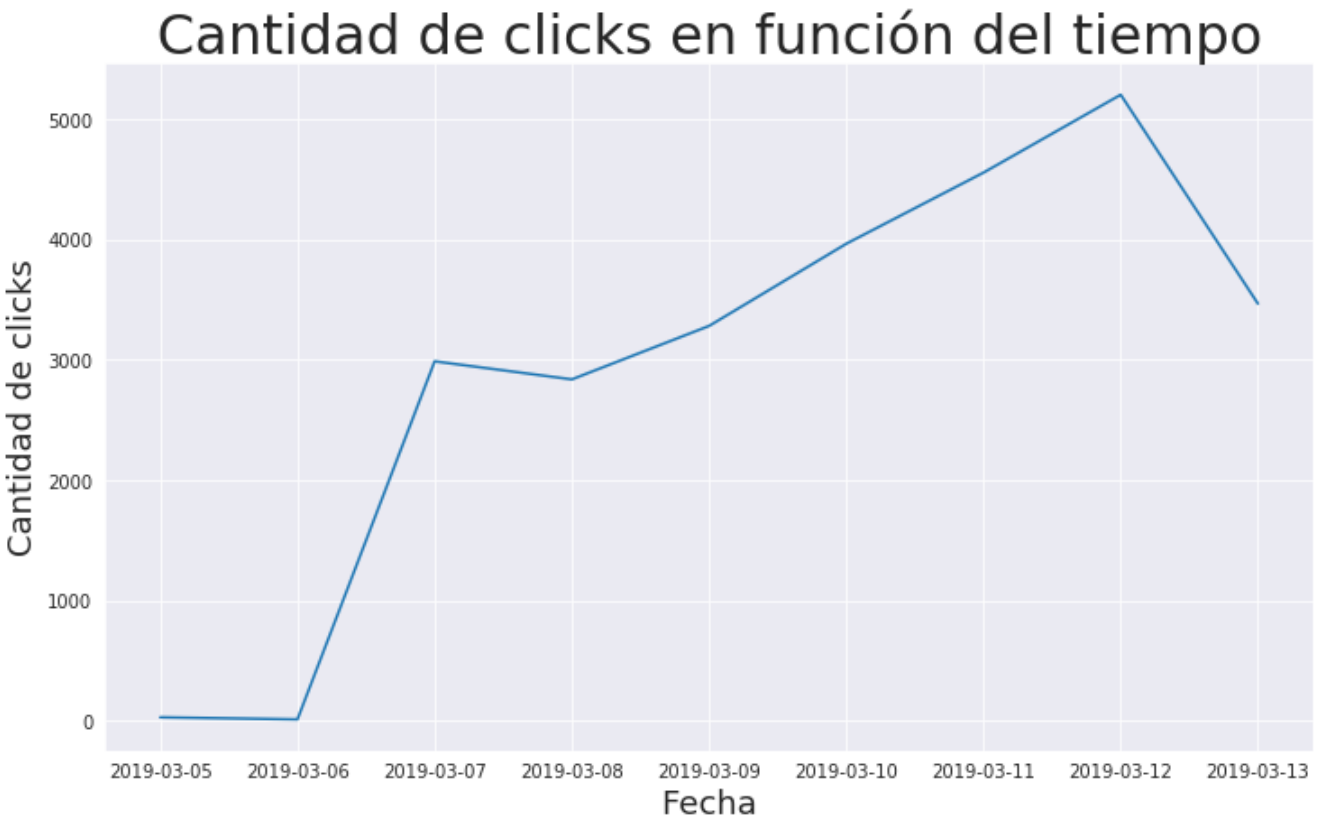
Se puede observar como la cantidad de subastas aumenta en función del tiempo, si bien hay pequeños picos que bajan en ciertos momentos, en un vistazo general se puede ver que las subastas contadas para el día 5 de marzo están por debajo de los 1.250.000 y van decayendo pero luego para el día 13 se cuenta un número cercano a los 3 millones y además se puede ver cómo va en aumento.

Eventos en función del tiempo



Se puede observar en el gráfico anterior como los eventos se comportan de manera similar a las subastas, pero con cantidades mucho más bajas que los observados en las Subastas. El número de eventos comienza con un valor relativamente bajo el día 5 de marzo, va en aumento hacia el final del periodo donde termina con un valor máximo para el período.

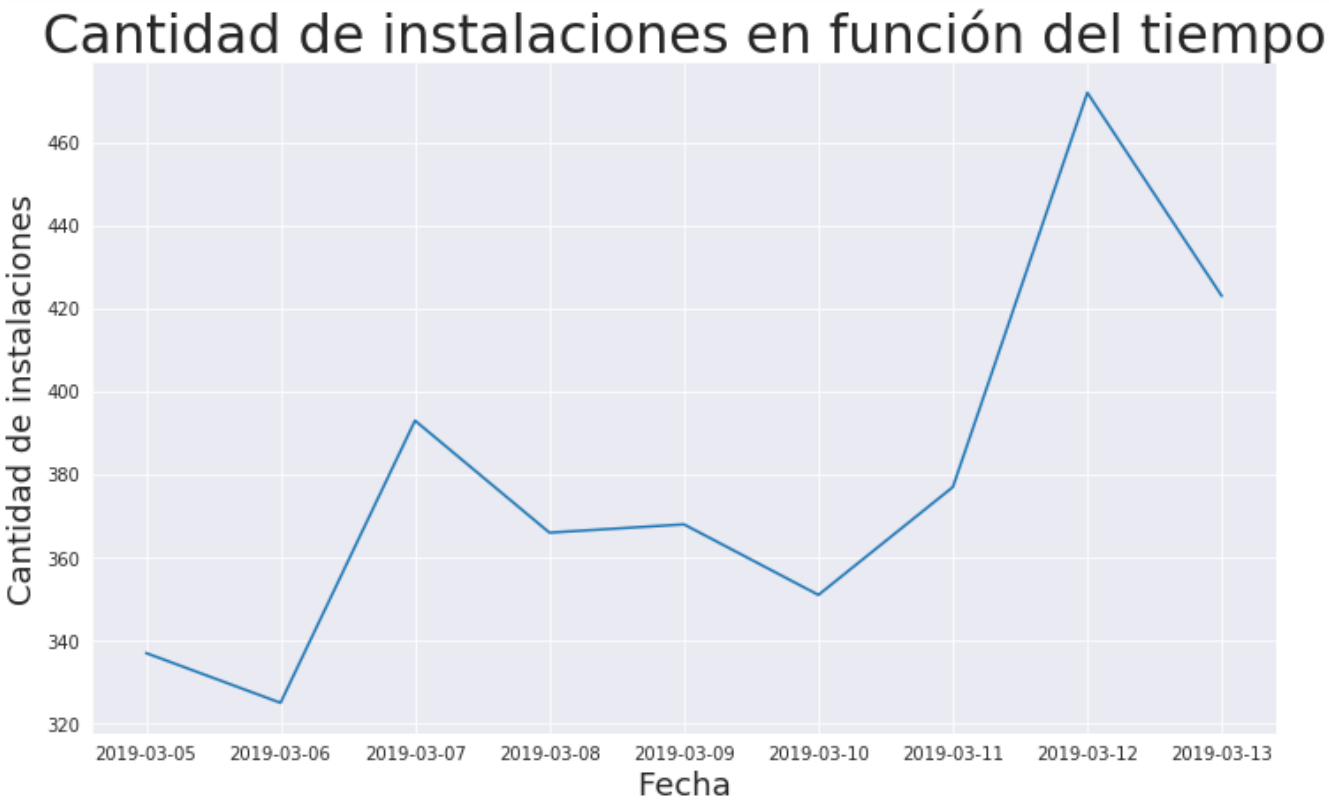
Cantidad de clicks en función del tiempo



En éste gráfico podemos observar un comportamiento un tanto diferente para los clicks si lo comparamos con los gráficos anteriores. Lo que se observa primero es que los números han disminuido bruscamente, lo cual tiene sentido, pues los datos se corresponden a datos de aplicaciones de clientes de Jampp. Por otra parte, se corresponde con la realidad cotidiana en la mayoría de las personas solemos ignorar las publicidades.

Algo más que se puede ver en el gráfico es que, aunque los números sean más bajos, igualmente van en aumento en relación al inicio del periodo, pero también está a la vista como a partir del día 12 los clicks disminuyen, ésta es la parte más inusual del gráfico puesto que el mismo no demuestra picos negativos tan bruscos en las fechas anteriores. Más adelante analizaremos a qué factor podría deberse esto.

Cantidad de instalaciones en función del tiempo



En el gráfico anterior se observa un comportamiento algo similar al de los clicks, y era de esperarse, pues luego de los clicks es que se puede desencadenar una instalación o no, por lo que observamos que generalmente van aumentando las instalaciones pero, al igual que los clicks, a partir del 12 comienzan a disminuir abruptamente. Otra cosa que se puede observar el la cantidad de instalaciones, pasando de unos miles de clicks a unas cientos de instalaciones, lo cual también tiene mucho sentido, pues muchas veces un click a una publicidad no significa que realmente vayamos a instalar el producto, por lo que finalmente sólo unos pocos clicks finalizan en instalaciones reales.

2.5 – Hipótesis de análisis

Para todos los análisis posteriores se tomará asumirá que el los datos del periodo descrito son representativos. Es de decir constituyen una muestra válida para obtener conclusiones cobre la misma. Entendemos que contar con datos de periodo corto y de un solo país obedece al hecho del gran volumen de datos generado por día. Para analizar mayor cantidad de datos deberíamos usar

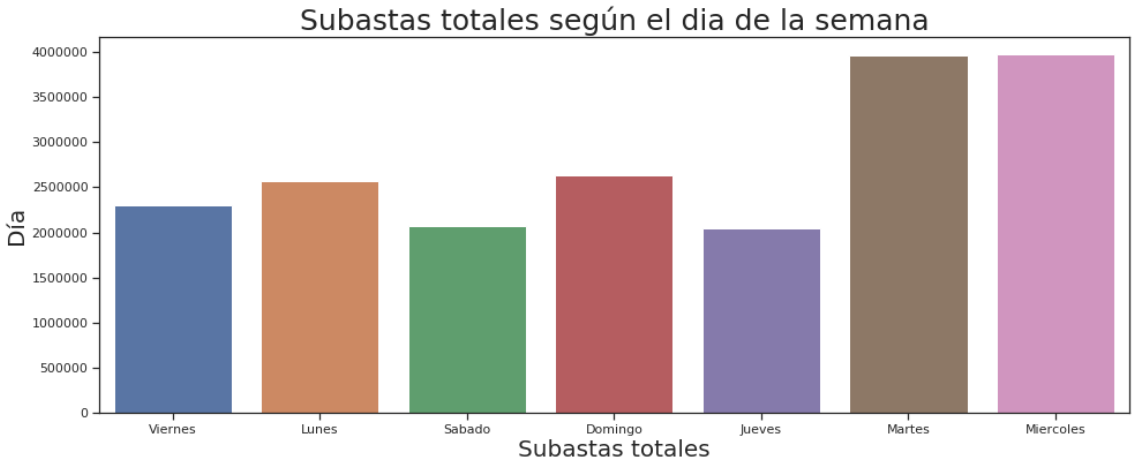
herramientas de procesamiento distribuido como Spark y contar con cluster ya que Pandas tiene limitaciones para analizar datos de más de algunos GBs.

3 - Análisis sobre las Subastas

3.1 - ¿En qué días y horas se generan más subastas?

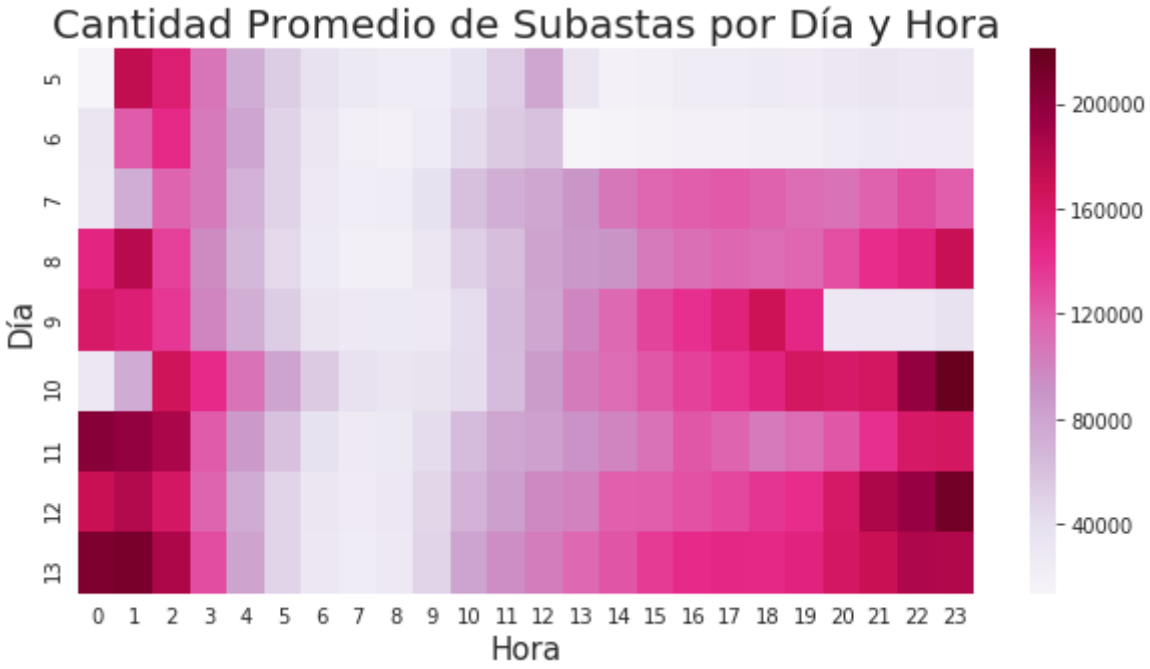
Luego del análisis temporal hecho en el Análisis Introductorio pudimos observar que las subastas van en aumento a medida que pasa el tiempo.

En esta sección intentaremos profundizar un poco más sobre la información que nos aportan las subastas y para esto parece interesante estudiar los momentos del día y también los días en que se realizaron las subastas, para visualizar la situación recurrimos al siguiente gráfico:



Vemos que los días Martes y Miércoles son los días que se realizan mayores cantidad de subastas.

Para analizar cómo se distribuyen las subastas por hora hicimos un heatmap:



Con esta visualización podemos observar fácilmente ciertos patrones.
Observamos como en un rango horario entre las 4 y las 12 hs. el número de subastas es realmente muy bajo en comparación al resto y en contraste observamos como la cantidad de subastas es realmente alto entre las 20 y 2 hs.

Algo muy interesante es el rango horario entre las 20 hs del día 9 y 2 hs. del día 10, del pasado mes de Marzo, el número de subastas tiene una baja abrupta, lo mismo se repite en los días 5, 6, y 7.

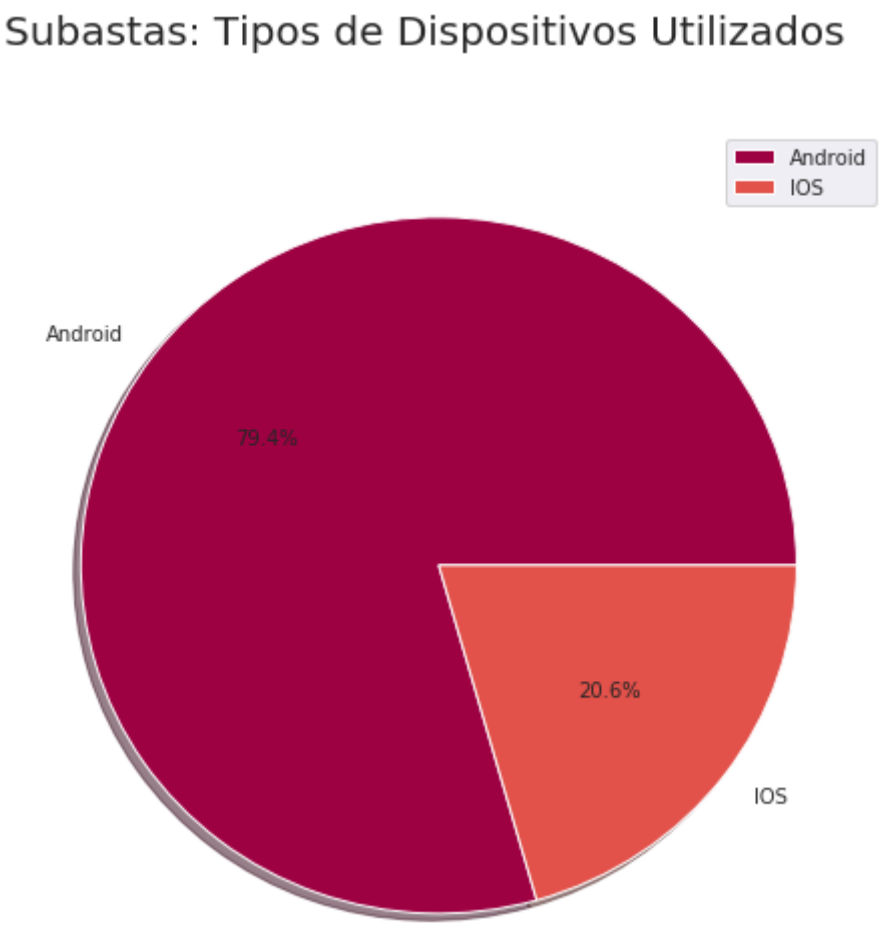
Viendo el lado opuesto tenemos que entre las 22 hs. del día 10 y las 2 hs del día 11 la concentración de subastas es muy alto, se repite lo mismo el día 12 y 13, observando esto podría pensarse que se trató de algún evento especial que desató en alto consumo de aplicaciones por parte de los usuarios y esto desembocó en un alto número de subastas.

3.2 - Tipos de plataformas desde las cuales se realizan las subastas

Otro de los datos interesantes con los que se cuenta en el set de datos de subastas es el de platform, el cual nos da información sobre la plataforma del dispositivo desde el cual se realizan las subastas.

El campo platform toma dos valores, 1 para la plataforma Android y 2 para la plataforma IOS.

Veamos desde cuál de estas plataformas se realiza un mayor número de subastas.



Observando el gráfico vemos que el tipo de plataforma más utilizada es Android, era de esperarse ya que es la plataforma más popular a nivel mundial.

4 - Análisis sobre los Eventos

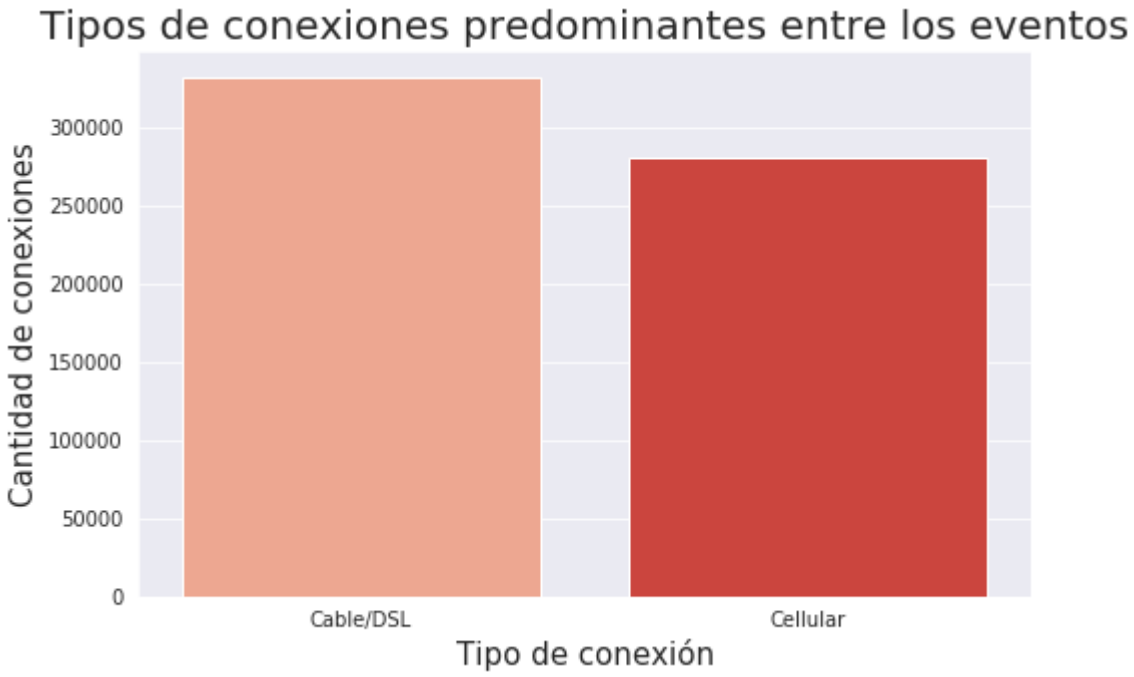
4.1- Tipos de conexión

Algo que puede resultar interesante es analizar el tipo de conexión utilizado en cada evento por lo que decidimos dedicar una parte de este informe a un análisis de los mismos.

Se observan tan sólo tres tipos de conexiones Cable/DSL, Cellular y Corporate. Haciendo un conteo de los mismos, para observar la proporción de uso de cada uno se obtuvo lo siguiente:

Tipo de Conexión	Cantidad de Eventos
Cable/DSL	331948
Cellular	280511
Corporate	4

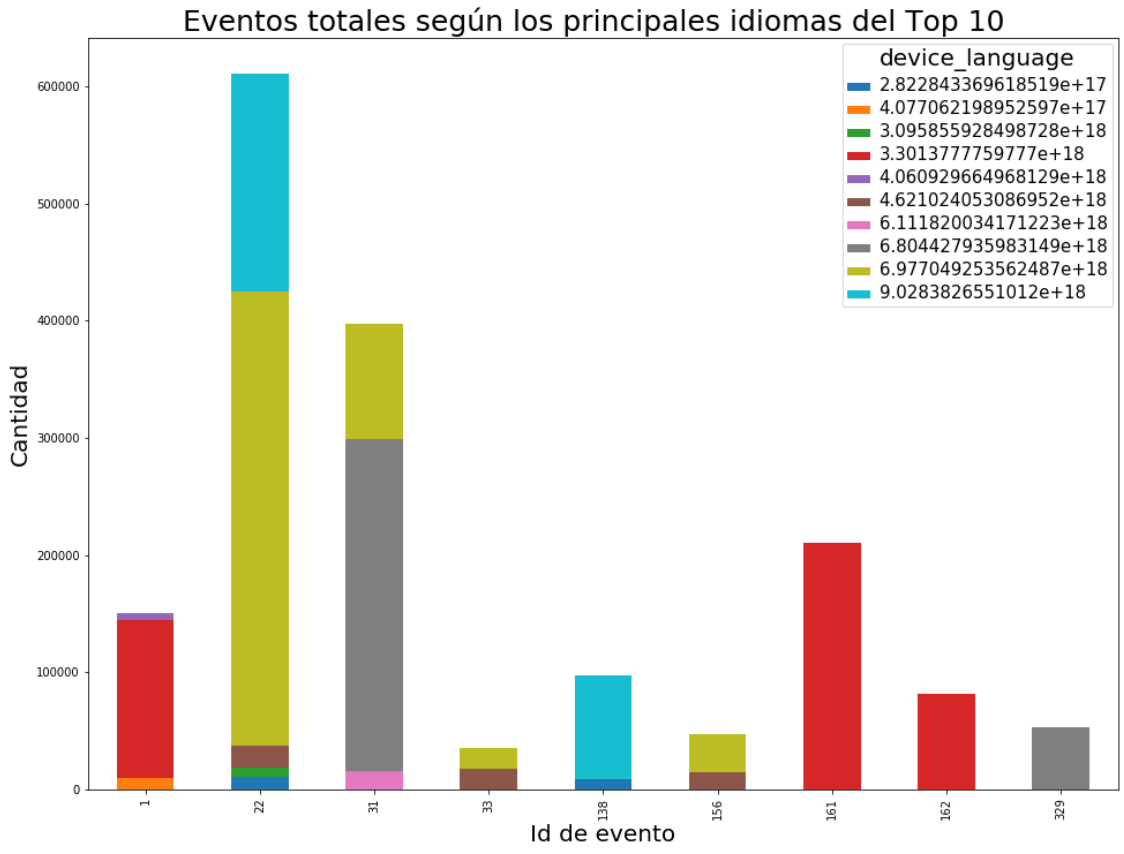
El tipo de conexión Corporate tiene tan solo 4 eventos y resulta que frente a los otros dos, que tienen más de 200.000 eventos, resulta un tanto despreciable por lo que realizamos un análisis de los tipos de conexiones predominantes en los eventos.



Finalmente en el gráfico anterior se puede observar cómo la conexión por Cable/DSL y Cellular difieren entre sí por muy poco. Por lo que no puede decirse demasiado sobre ellas, lo que sí se puede afirmar es que es muchísimo más probable que se genere un evento desde una conexión Cable/DSL o Celular que desde una conexión Corporate.

4.2- ¿Cuáles son los eventos más populares?

A continuación se muestra un gráfico de barras apiladas de la cantidad de ocurrencias de los top 10 eventos distinguidos por los principales lenguajes del dispositivo:

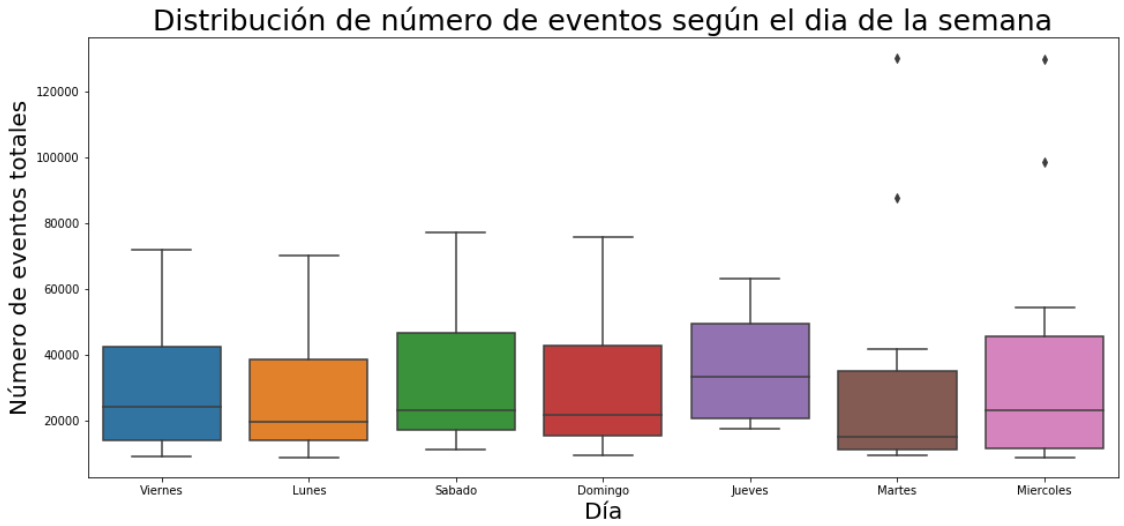


De la visualización anterior se puede concluir que el evento 22 es el más popular. El mismo ocurre en dispositivos configurados con los lenguajes cuyos

códigos aparecen en el gráfico. Vemos que para cada uno de los de los id de eventos hay distintos idiomas de dispositivo que predominan. Hay una fuerte correlación entre el tipo de evento y el leguaje del dispositivo. Por ejemplo, los eventos con id 161,162 y 329 solo se producen en dispositivos de un solo lenguaje (3,309…).

4.3- Distribución de los eventos en el tiempo

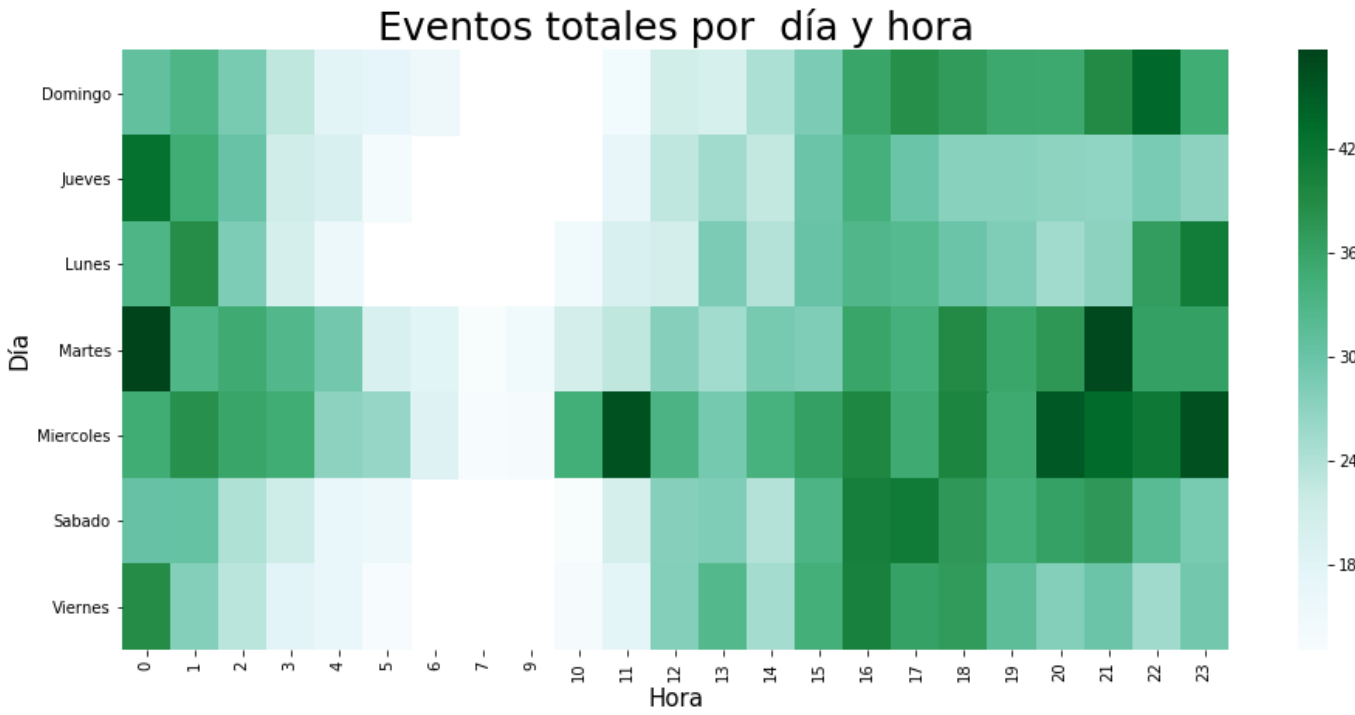
Si graficamos la distribución de los eventos según el día de la semana en un box plot:



Es interesante como se distribuyen la ocurrencia de eventos según el día. Vemos que los días martes tiene una mediana más baja que el resto de los días pero valores muy por encima definido por el máximo de la distribución (outliners). Esto indica una gran dispersión de los valores de ese día. El día miércoles presenta un comportamiento similar pero con un mediana superior.

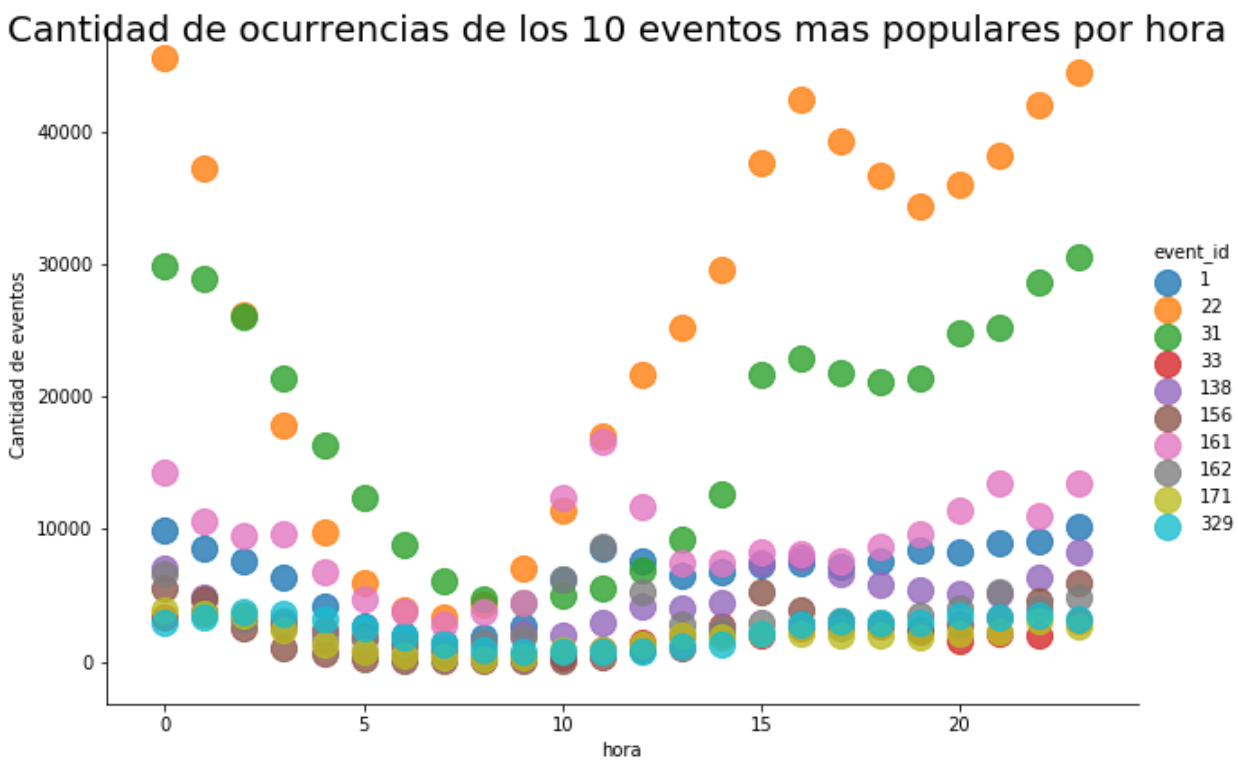
Esto significa el comportamiento de los usuarios de jueves a lunes es más parejo que para los días martes y jueves. Saber esto es importante para definir la importancia de cada subasta según el impacto que potencialmente pueda tener.

En el siguiente heatmap vemos cómo se distribuyen los eventos según el día y la hora.



De la visualización anterior podemos ver que los días martes y miércoles se registran mayor número de eventos, concentrados en horarios nocturnos. Asimismo notamos que en general de las 3 a las 10 am se registran pocos eventos en comparación. Es interesante notar que los miércoles se generan gran cantidad de eventos durante todo el día salvo en el horario de 5 a 8 de la mañana. Es un comportamiento no esperado de los usuarios.

Pare visualizar mejor el comportamiento de los usuarios según la hora del día a continuación graficamos como varías las ocurrencias de los 10 eventos más populares:



Vemos en el gráfico anterior que los eventos van aumentando a partir de las 10 de la mañana, teniendo un “máximo local ” alrededor de las 15 hrs . De las 15 hrs hasta las 20 horas el número de eventos disminuye un poco y luego de las 20 hasta las 0 horas vuelve crecer.

Este comportamiento se puede deber a que la gente utiliza más los celulares durante la hora del almuerzo y luego de terminada la jornada laboral o de estudio.

De estas visualizaciones se puede concluir que Jampp debería definir sus precios máximos según el horario y el día de la subasta. Especialmente en los días martes y miércoles donde las diferencia en la cantidad de eventos según el horario es mucho mayor que en el resto de los días.

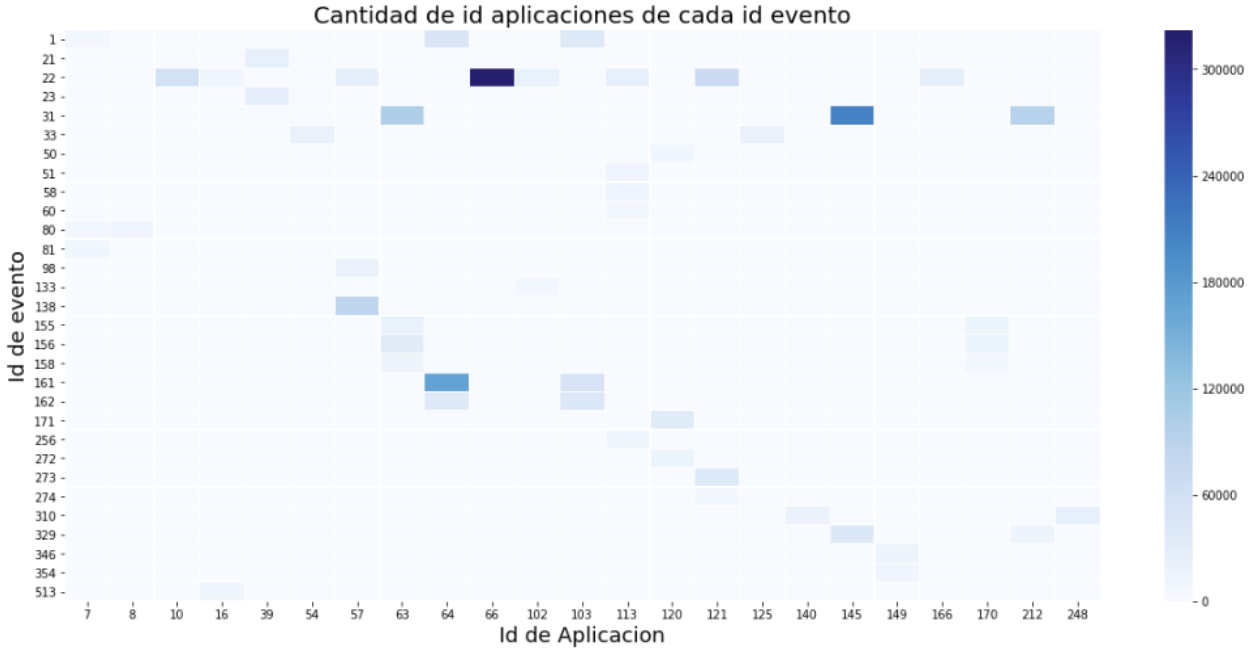
4.4- Aplicaciones con mayor cantidad de eventos

En este análisis se pretendió ver qué aplicaciones tienen más eventos, para descubrir cuál es la aplicación sobre la que se hacen la mayor cantidad de acciones, contando todas las categorías de las mismas.

Según el gráfico la aplicación más popular en los eventos es la de id 66 con bastante diferencia frente al segundo puesto, esto podría haber sido porque esta aplicación sea la que tenga mayor cantidad de opciones para los usuarios frente a las demás. Por otra parte vemos que se cumple el Principio de Pareto ya que pocas aplicaciones concentran la mayor parte de los eventos.

4.5- Cantidad de eventos y aplicaciones

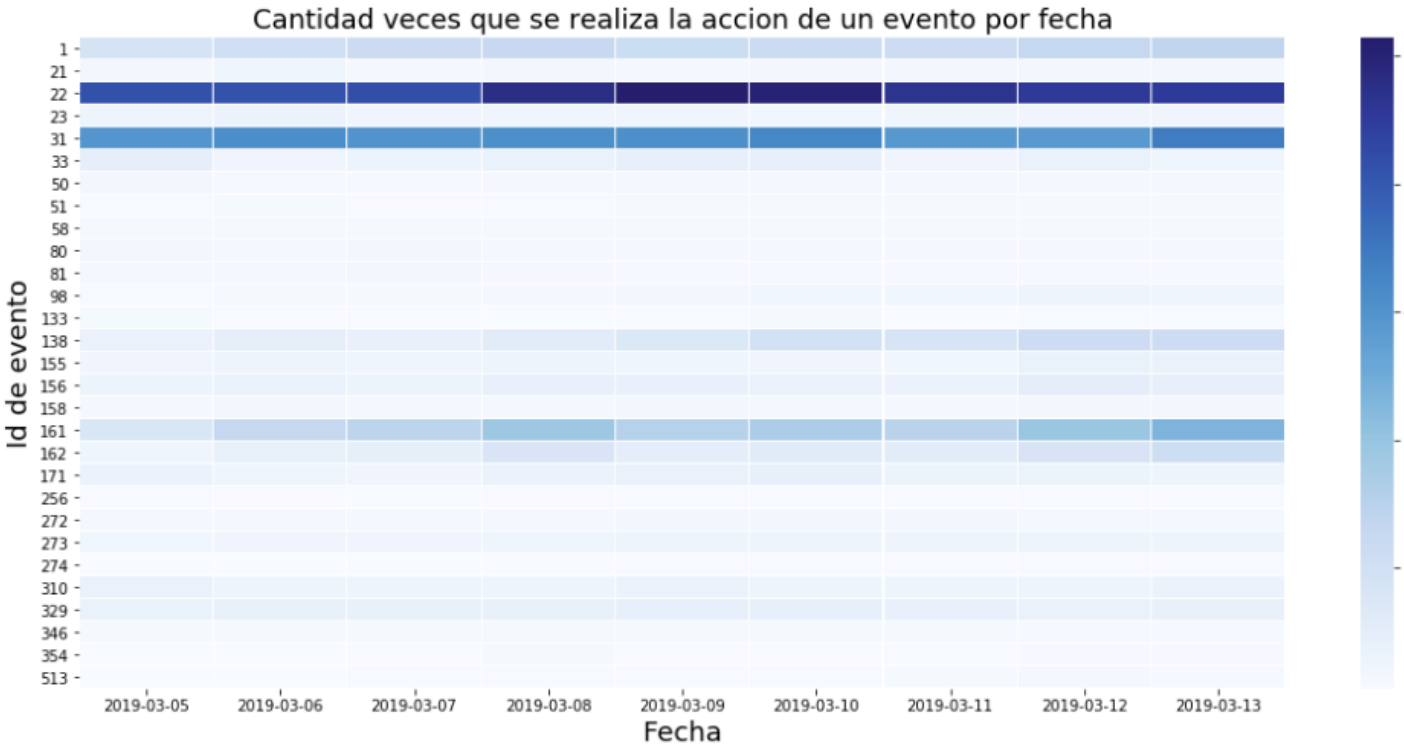
En este análisis se quiso hacer lo mismo de antes pero diferenciando las ids de los eventos, para ver la cantidad de cada evento que tiene cada aplicación



Con este plot se verifica que la aplicación 66 es la que más eventos tiene porque el evento 22 se da una gran cantidad de veces relacionado con esta aplicación, además la aplicación 66 es también la aplicación que más está relacionada con el evento 22, podría ser que la aplicación 66 sea la más popular para realizar la acción del evento 22, y que el evento 22 sea la razón principal para instalar la aplicación 66. La misma relación se observa con la aplicación id 145 y el evento 31, y con la aplicación 64 y el evento 161; ambas las aplicaciones que ocupan el segundo y tercer puesto respectivamente de aplicaciones con más acciones.

4.6- ¿Qué eventos predomina por día?

Acá se buscó ver cuáles eventos aparecen más por día, para ver qué es lo que más hacen los usuarios diariamente, independientemente de la aplicación en la que se realice.



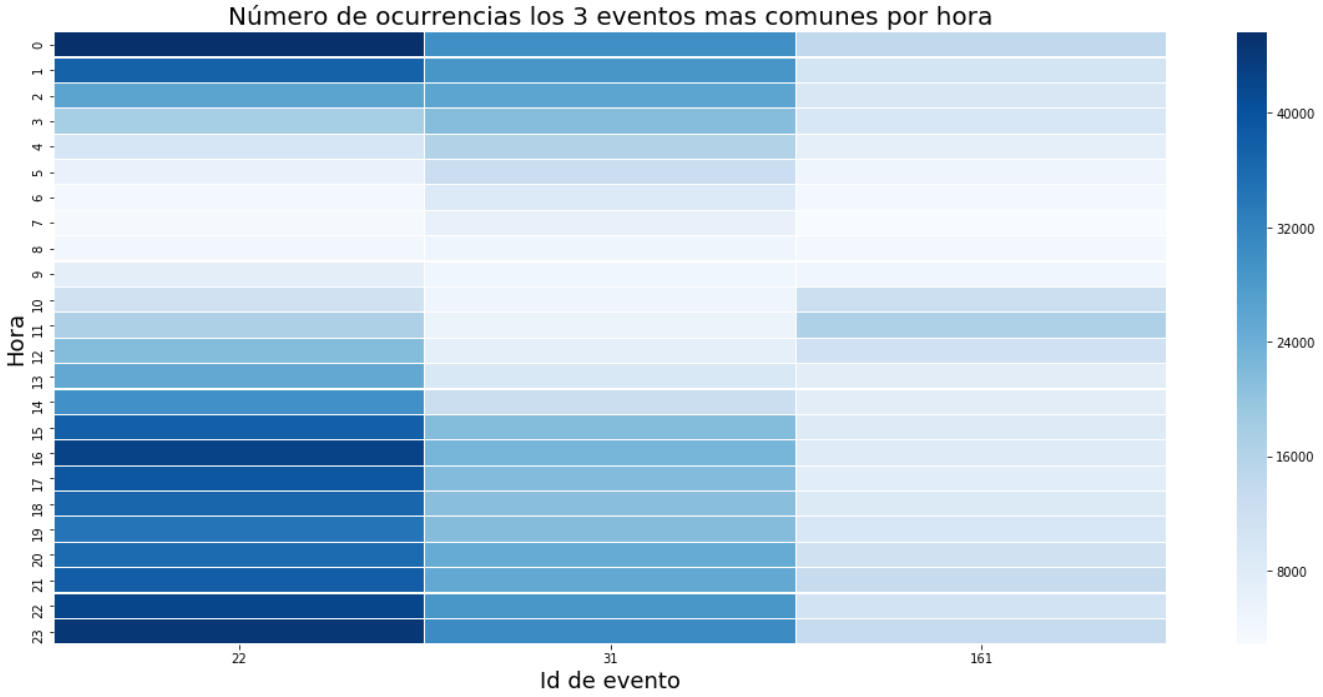
Como se vió en las visualizaciones anteriores, la acción más frecuente es la del event_id 22, pero no solamente es que más se realizó en total sino que además es la

que más se realiza por día, seguida de los eventos 31 y 161, que corresponden a los eventos asociados a las aplicaciones más usadas. Se entiende que las acciones de estos eventos son las acciones más comunes que realiza un usuario diariamente de entre las acciones que puedan realizarse a través de una aplicación.

4.7- Horarios en que se generan los eventos más comunes

Luego de obtener las tres acciones que más realizan los usuarios, se quiso ver en qué horas se usan y cuánto se usan más

.

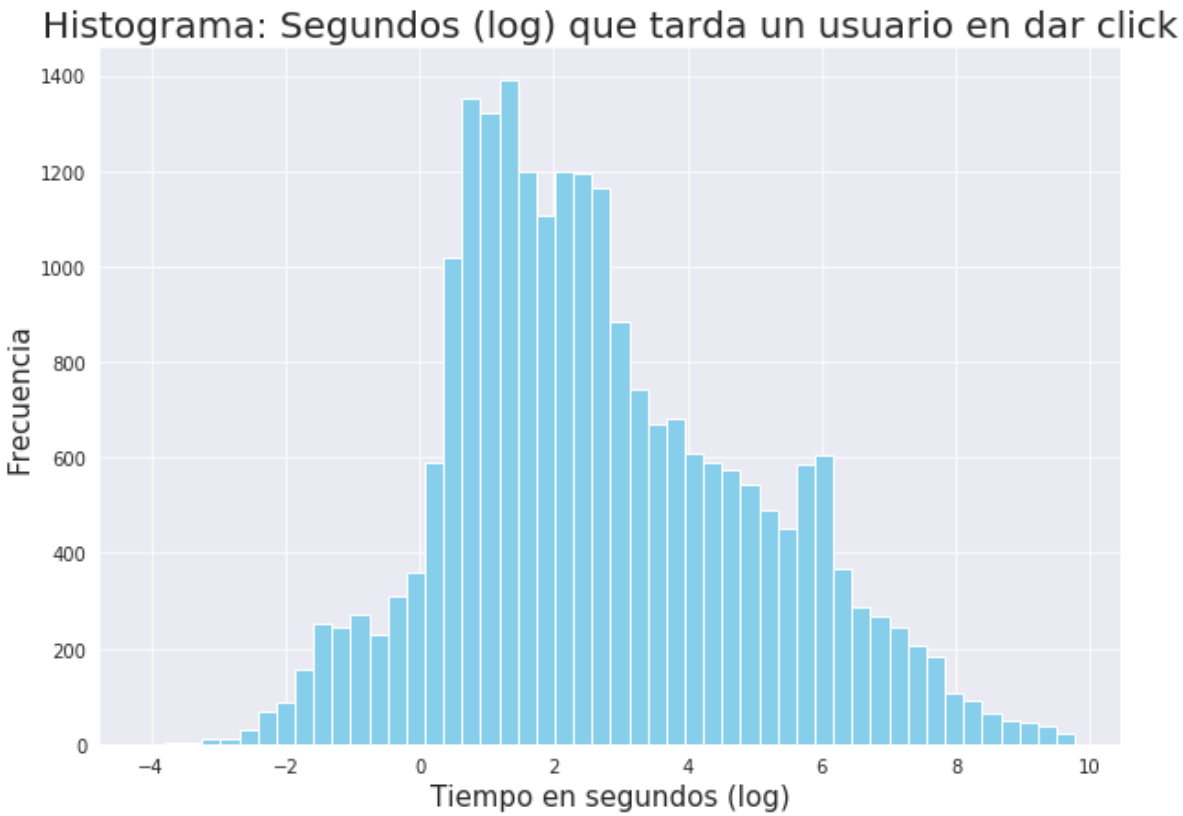


Se ve que los 3 se usan menos en los horarios donde la gente suele despertarse para trabajar por lo que no se puede concluir mucho de ello. Por otra parte, el evento 22 se usa más en los horarios más vespertinos, pero llega a un pico similar a las 16 horas que suelen estar dentro de horarios laborales, pero sin saber exactamente a qué acción corresponde el evento 22 no puede saberse a ciencia cierta si el evento 22 invita a realizarse más en esa hora o es una simple anomalía. El evento 161 curiosamente llega a su pico de frecuencia a las 11 horas, superando incluso los horarios vespertinos.

5 - Análisis sobre los Clicks

5.1- Distribución del tiempo que tarda un usuario en dar click a una publicidad

Para realizar este análisis utilizaremos la columna ‘timeToClick’ del set de datos Clicks haciendo una visualización de un Histograma y tomaremos los valores logarítmicos, ya que estos resultan más representativos que los valores reales.

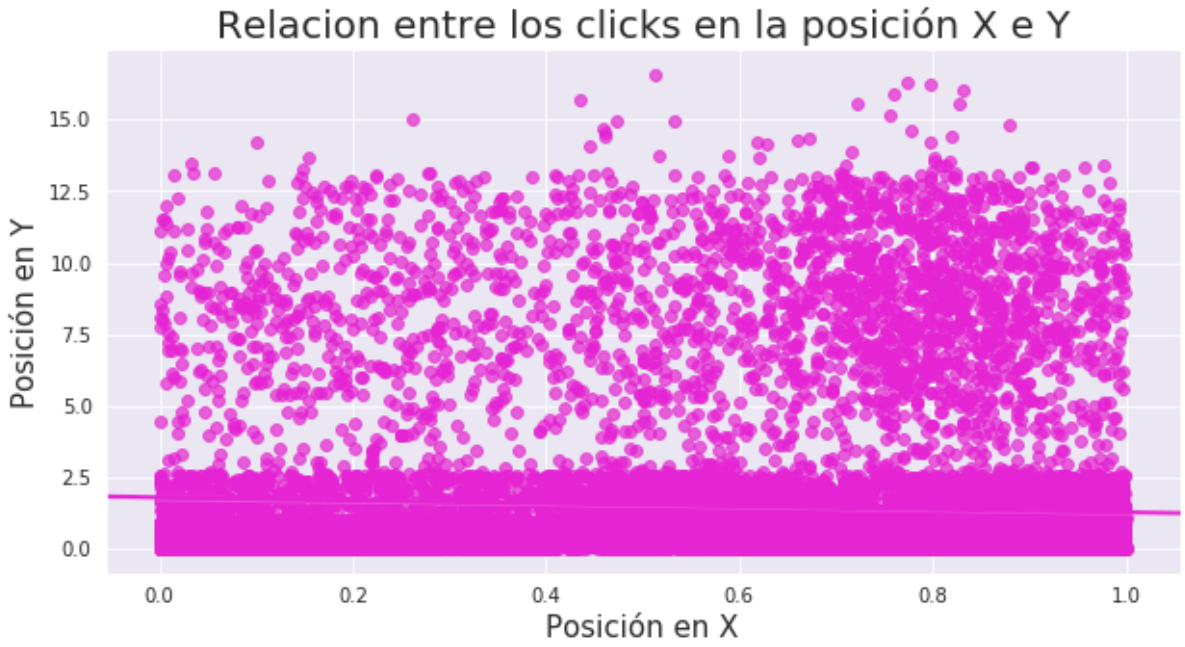


Como se puede observar los tiempos que están alrededor del rango entre 1 y 3 segundos (log) son los que tienen mayor frecuencia.

De esto podría deducirse que la mayoría de los usuarios que dan clicks se toman al menos más de 10 segundos en observar la publicidad, lo cual suena lógico pues se toman el tiempo para sentirse “interesados” en lo que les están ofreciendo.

5.2- Posición de pantalla desde donde se dan los clicks

Otro de los datos con los que contamos son las posiciones en donde se dan los clicks, nos resulta interesante analizar si hay algún patrón entre estos valores. Por ejemplo ver si hay una región dentro del cuadro en que se muestra la publicidad donde los usuarios se vean más atraídos a hacer clicks.

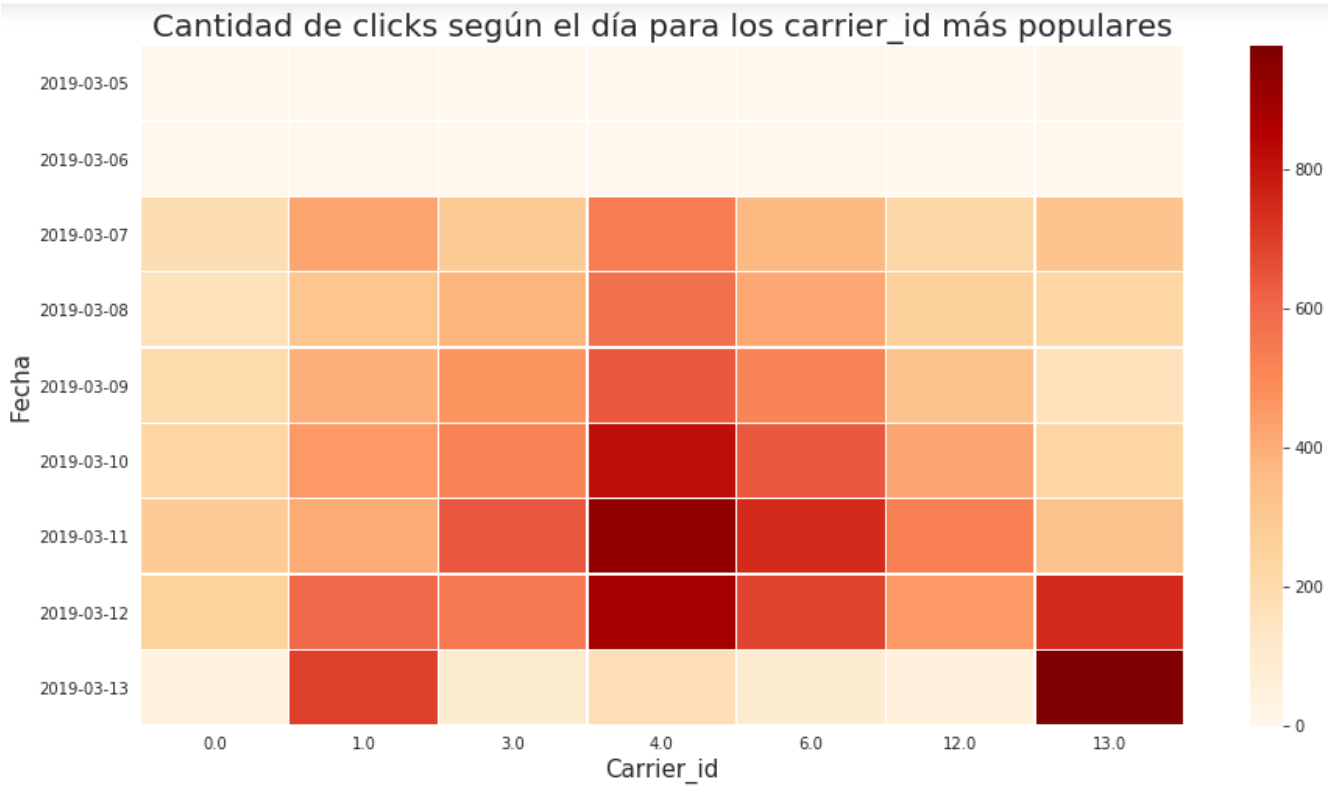


En el gráfico se puede observar como los clicks se acumulan alrededor de las posiciones entre 0 y 1 para X y entre 0 y 2.5 para Y, de lo que podemos deducir que la mayoría de los usuarios prefieren o tal vez se sienten más atraídos, a hacer click sobre esta región del cuadro de publicidad.

Este dato podría resultar de utilidad, por ejemplo si se intenta introducir algún nuevo método para atraer instalaciones por medio de promociones, ofertas o algo relacionado, podría probarse con colocarlo en esta posición de la pantalla.

5.3- Campo ‘carrier_id’ populares según el día

La idea de este análisis es detectar si se producen más clicks desde ciertos operadores móviles según las fechas registradas. Para realizar este análisis lo primero que hacemos es filtrar para quedarnos con los carrier_ids más populares, dado que hay muchos que tienen menos de 10 clicks, pero además tenemos unos pocos que acumulan más de 1000 clicks, por lo que se decide filtrar tomando los más populares, y analizamos sobre ellos.



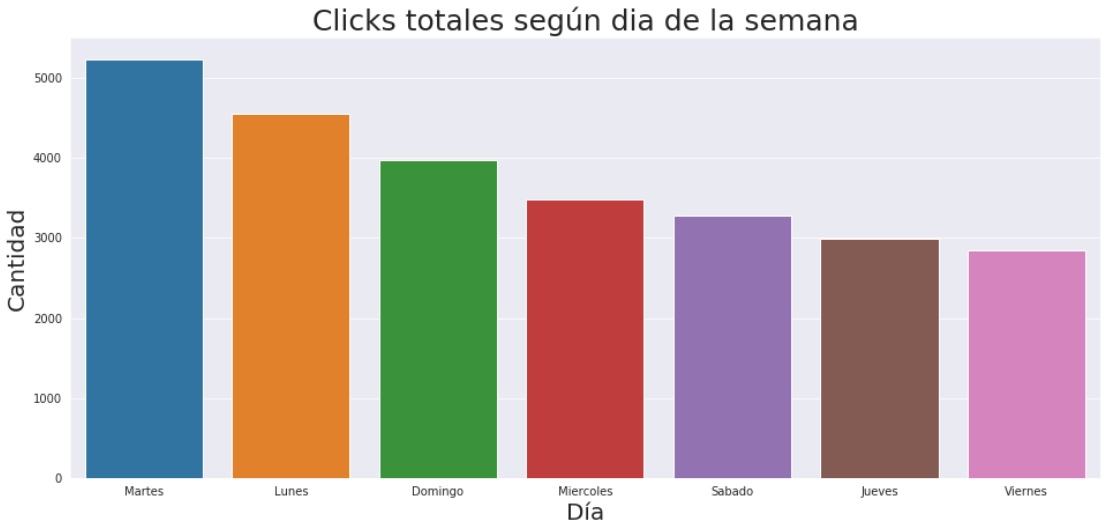
De este gráfico lo que logramos ver como el carrier_id 13 tiene alta cantidad de Clicks durante el día 13, y lo mismo se repite para el carrier_id 4 durante el día 11 y 12.

El resto de los carrier_ids son menos frecuentes, aunque también se denota un leve aumento, al menos hasta el día 12.

Otra cosa a remarcar del gráfico es que durante los días 5 y 6 la actividad registrada es muy baja, de hecho, esto ya lo hemos visto en otra sección del informe, donde podíamos ver que, a grandes rasgos, la cantidad de clicks iba aumentando en función del tiempo.

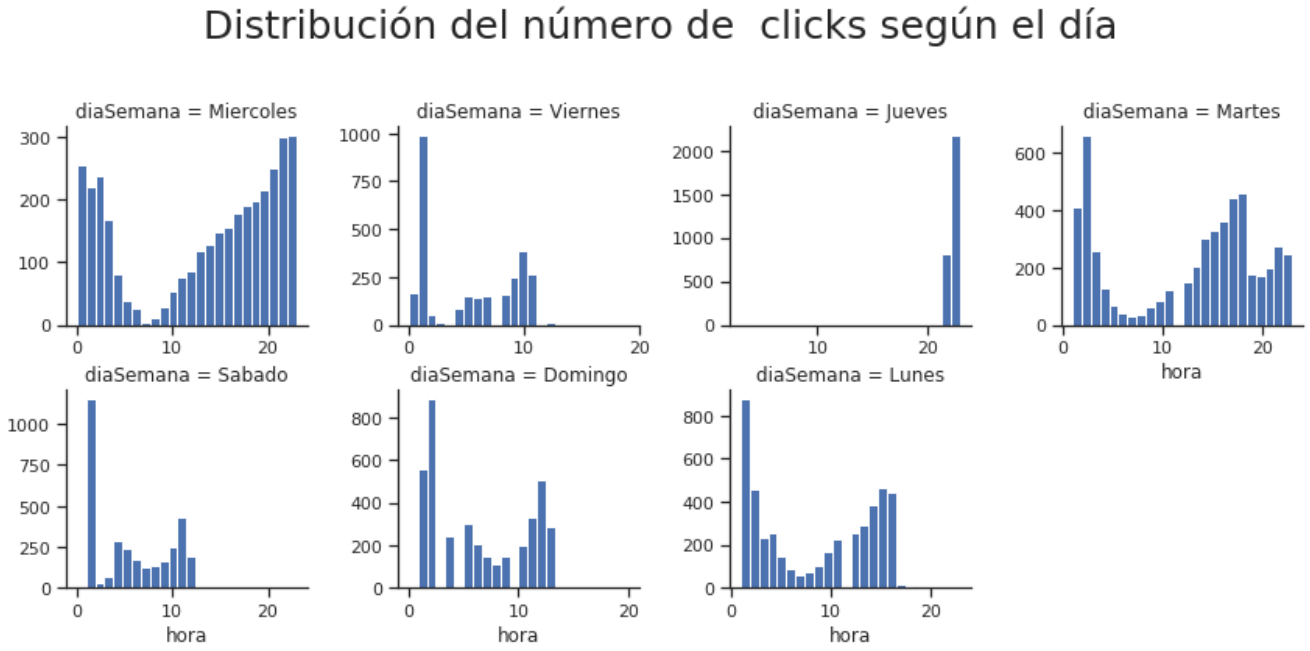
5.4-Distribución de clicks en tiempo

A continuación graficamos la distribución del número de clicks según el día de la semana.



Vemos que el día martes son los días que los usuarios de los clientes de Jampp realizan más cantidad de clicks.

Si graficamos la distribución de los clicks según la hora para cada día de la semana:



Para hacer los gráficos anteriores se usaron distintas escalas para lograr una visualización más clara dado la diferencia en los valores máximos entre los mismos.

De los gráficos anteriores notamos que hay ciertos horarios para los cuales no se registraron clicks y en el día jueves directamente todos los eventos se registraron desde las 22 hasta la 0 horas del día siguiente. Esto evidentemente no se corresponde con el comportamiento de los usuarios si cómo se registraron los datos en la base de datos de Jampp. Por lo tanto, no se pueden sacar conclusiones válidas con respecto a estos últimos gráficos.

6 - Análisis sobre las Instalaciones

6.1 - Relación entre Tipo de Instalación y Sesión de agente de Usuario

En el set de datos “Installs” encontramos los campos “kind” y “session-user_agent”, intentamos ver si se relacionan de alguna manera a la hora de “sumar” instalaciones, es decir, buscaremos ver si para cierto tipo de session_user_agent son más comunes ciertos tipos de instalación. Para ver esto, como ya es habitual, realizamos una visualización de los datos acumulados agrupándolos por “session_user_agent” y “kind”.

Lo primero que observamos en el gráfico es que hay dos tipos de session_user_agent que predominan, adjust.com (circulo azul) y http-kit/2.0 (circulo verde). Vemos que son los que abarcan una mayor variedad de tipos de instalaciones y también son los que generan a su vez mayor cantidad de instalaciones.

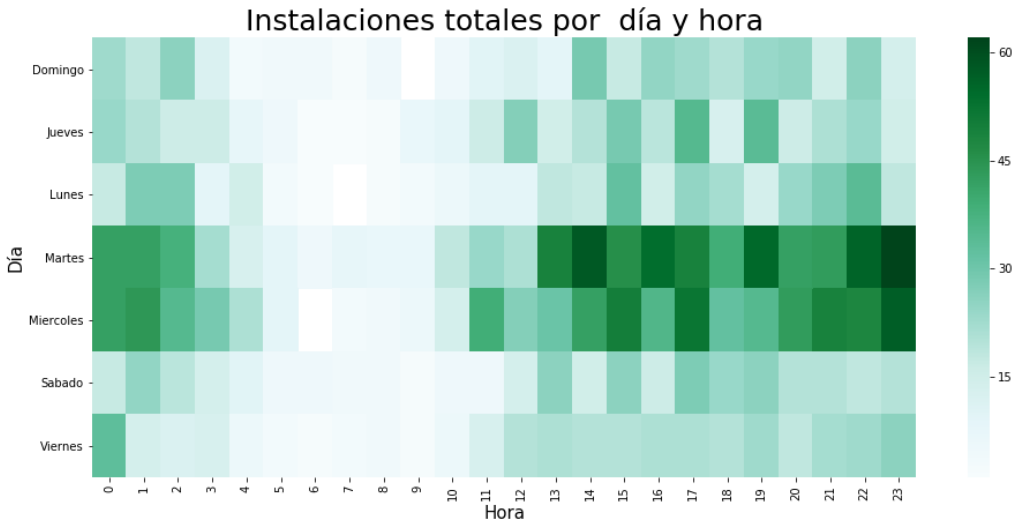
Los círculos azules son los que abarcan más tipos de instalaciones diferentes, con más del doble de los verdes y a su vez es el que tiene el pico más alto de instalaciones, por lo que denominamos que es el “session_user_agent” más popular entre las instalaciones

Luego tenemos 3 session_user_agent más, que son los menos populares, presentándose los mismos en un tipo de instalación determinada y con cantidades de instalaciones muy bajas, como por ejemplo desde el session_user_agent relacionado a Mozilla (círculo rojo) sólo se realizaron instalaciones de tipo desconocido y lo mismo sucede para los session_user_agent desconocidos (círculo violeta) y por último, tenemos los HasOffers Mobile AppTracking v1.0 (circulos Naranjas) desde los cuales se realizan sólo instalaciones del tipo Open.

Por lo tanto concluimos que los Session_user_agent más populares son también los que abarcan diferentes tipos de instalaciones.

6.2 - Número de instalaciones según día y hora

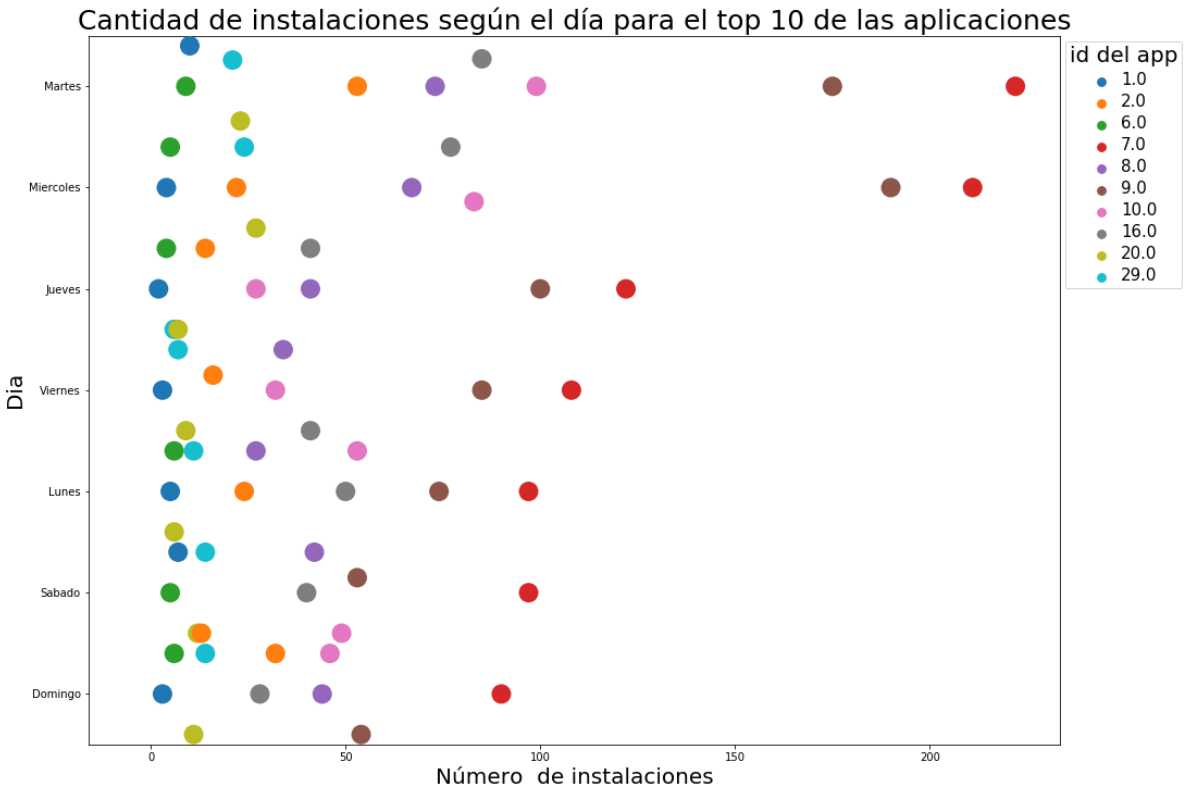
Es interesante analizar en qué momentos los usuarios realizan instalaciones. Para ello realizamos el siguiente heatmap:



De la visualización anterior se puede concluir que los días martes y miércoles son los días en los cuales se realizan mayores números de instalaciones. Asimismo vemos que de las 4 a las 11 de la mañana las mismas son menores y a partir de las de las 12 del mediodía empiezan a crecer hasta alcanzar un pico de instalaciones entre las 22 y 24 horas.

6.3 - Top 10 aplicaciones

A continuación graficamos el número de instalaciones de las 10 aplicaciones más instaladas según el día de la semana:

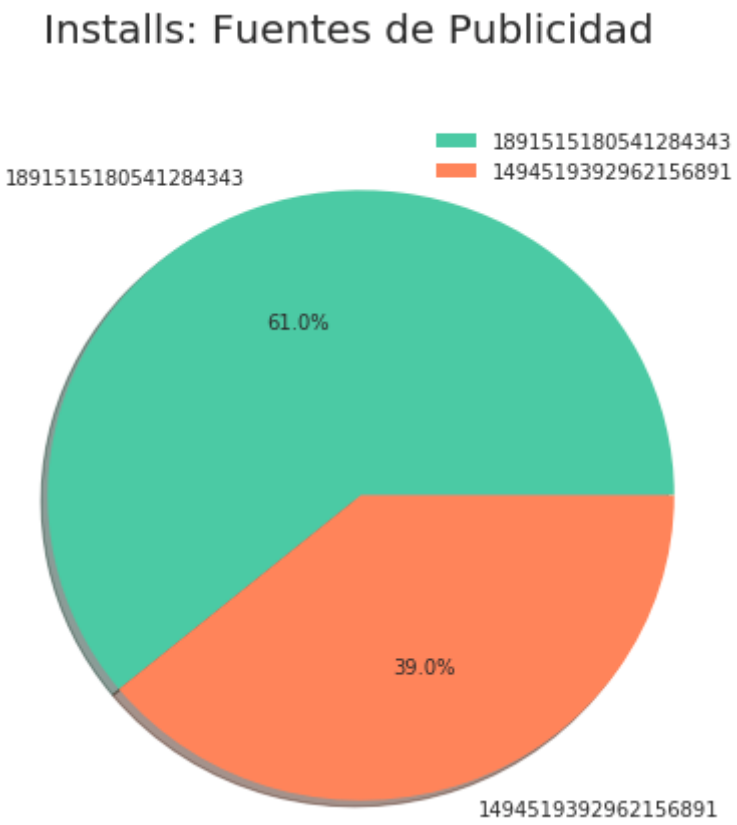


Es interesante notar que hay aplicaciones como la 6 y la 1 que no tiene gran variación en cuanto a las instalaciones según el día; mientras que otras como la 7 y la 30 si presentan variaciones notables. Esto denota que hay aplicaciones número de instalaciones es prácticamente constante mientras que otras aumentan mucho ciertos días (en general los martes y miércoles).

6.3 - Fuentes de Publicidad

Otro campo interesante del set de datos installs es el campo ‘ref_type’, el mismo asocia la instalación al tipo de fuente de publicidad, pudiendo ser publicidad de Google o de Apple.

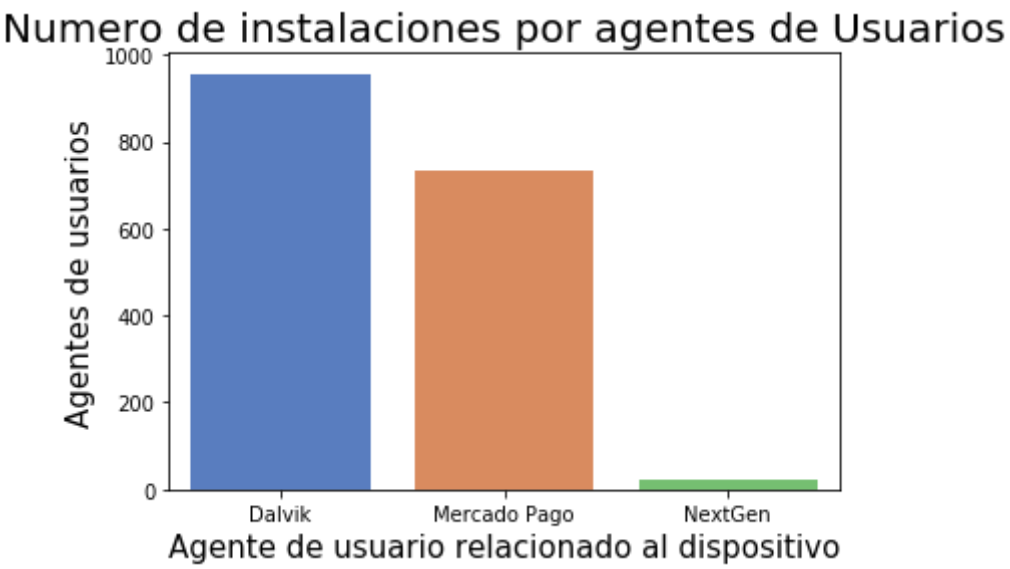
Como se sabe los datos están anonimizados, pero sabemos que solo hay dos focos posibles, publicidad de Apple (IFA) o publicidad de Google.



Como se puede observar en el gráfico anterior hay una fuente de publicidad predominante, asociada al id: 1891515180541284343.

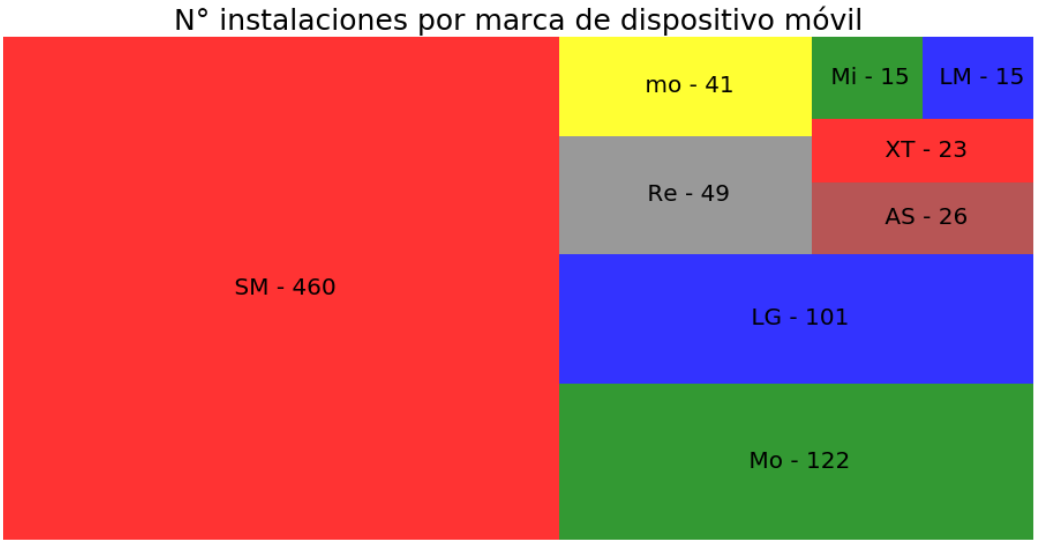
6.3- Análisis de Agentes de Usuarios

Del campo de ‘user_agent’ puede extraerse información de los agentes vinculados con los dispositivos de usuarios. Las referencias obtenidas corresponden principalmente con tres tipos de origen principales como se muestra en el gráfico a continuación, de acuerdo a la cantidad de instalaciones que tiene cada uno. Predomina **“Dalvik”** el cual tiene información de dispositivos móviles esencialmente. Le siguen Mercado pago y por última una minoría poco significativa a partir de NextGen, vinculada con navegadores y otras plataformas.



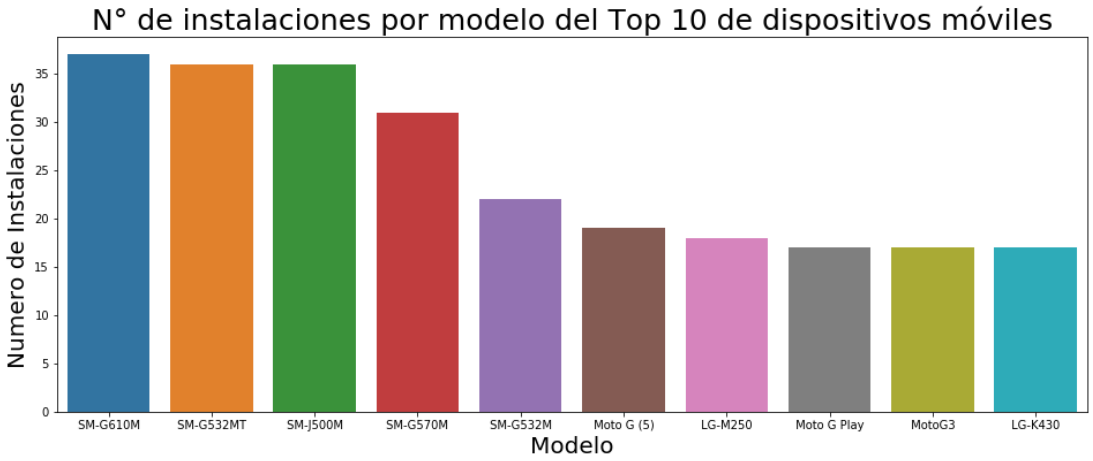
De aquellos que se corresponden con la identificación de “dalvik” pueden extraerse diversos modelos de dispositivos.

Con el objetivo de analizar la importancia de cada marca se hizo el siguiente tree map:



En el gráfico anterior notamos que la marca que predomina sobre todo es **Samsung** con 460 instalaciones. Le siguen Motorola con 122, Lg, Redmit y nuevamente Moto (aquellos motorola que con la identificación de moto cuya implementación le corresponden a Lenovo).

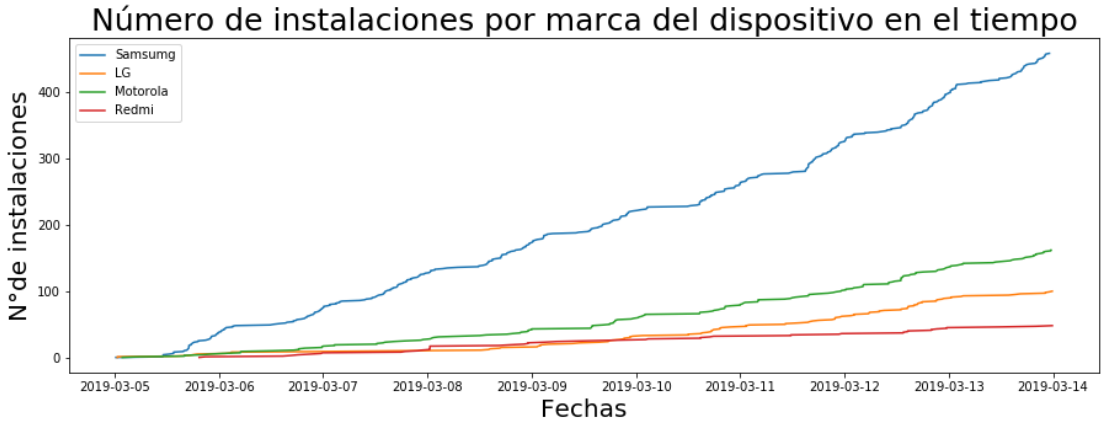
Si graficamos el número de instalaciones de los 10 modelos más populares obtenemos la siguiente visualización:



Vemos que predominan los modelos de Samsung: G610 (J7), G532 (J2) y J500 (J5).

Estos modelos juntos con el resto que aparecen en el gráfico se corresponden con la gama media y baja de celulares.

Si graficamos el número de instalaciones por modelo a lo largo de los 10 días vemos que el crecimiento es bastante regular a lo largo del tiempo :



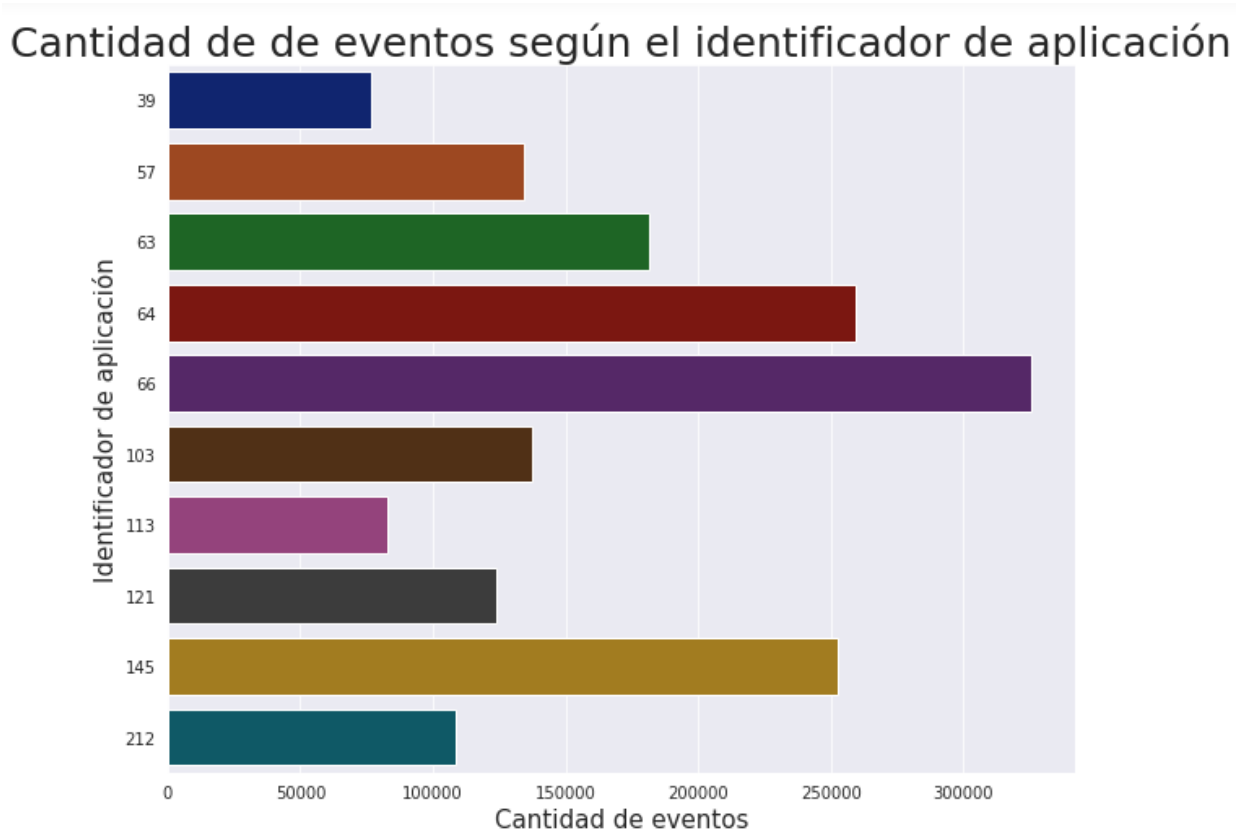
En la visualización anterior se nota que la pendiente de la curva de las instalaciones correspondientes a dispositivos de la marca Samsung es superior a la del resto.

7 - Análisis Varios

7.1 - Aplicaciones con más eventos vs instalaciones

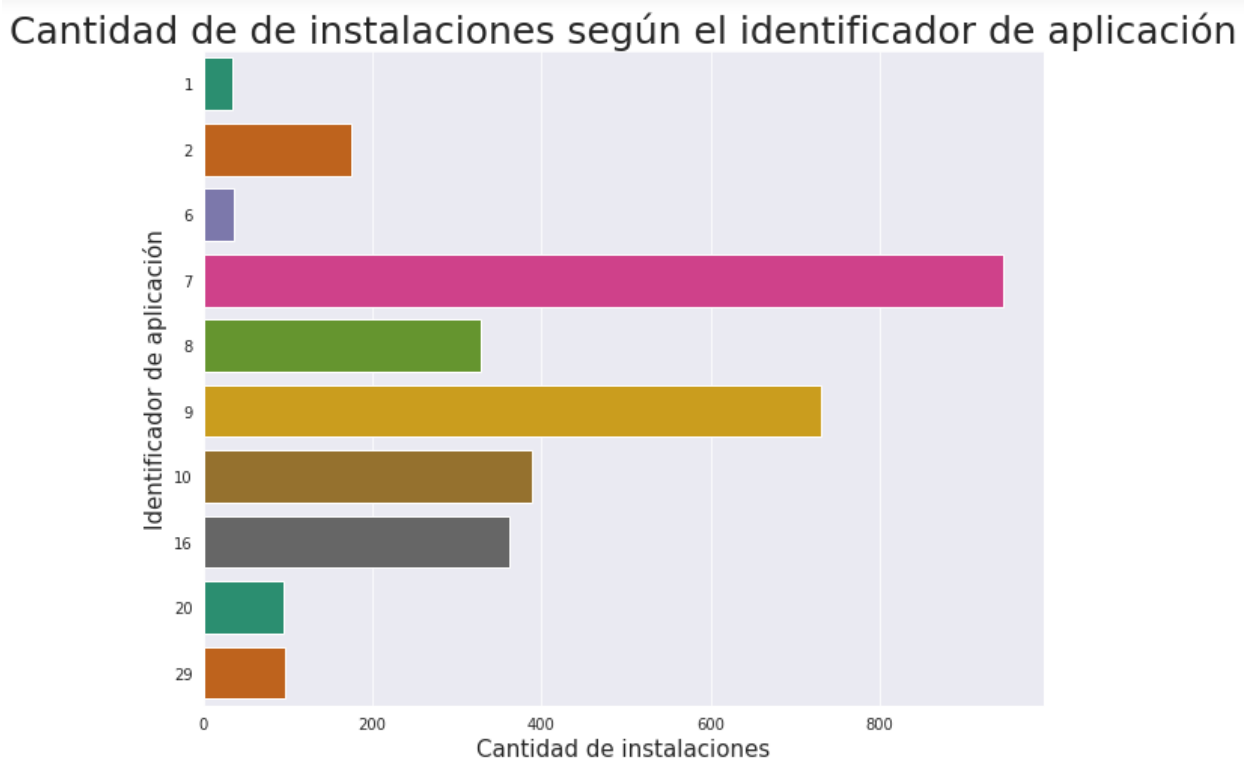
Nos interesa ver cuáles son las aplicaciones que generan más eventos y las que más se instalan, para esto analizamos lo que sucede con la columna 'application_id' y visualizaremos los identificadores de las aplicaciones más populares para ambos casos.

Primero veamos cuáles son las 10 aplicaciones que generaron más eventos:



Se puede observar que la aplicación que más eventos genera está identificada por el número 66 y generó más de 300.000 eventos.

Ahora veamos que sucede con las instalaciones:



Veamos ahora, la aplicación con más instalaciones está identificada por el número 7, generando más de 800 instalaciones.

¿Qué sucede aquí? No es la misma que genera más eventos, es más, la aplicación 66 ni siquiera figura entre las 10 aplicaciones más instaladas. Lo que sucede tiene que ver con algo ya mencionado anteriormente en la sección de Análisis introductorio, los eventos corresponden a clicks e instalaciones, pero obtener muchos clicks no garantiza muchas instalaciones. Los clicks a veces se generan por error o por curiosidad, pero sólo algunos pocos desembocan en una instalación y estas visualizaciones son el claro ejemplo de ello.

7.2 - Eventos vs. Instalaciones atribuidas a Jampp

En esta sección intentaremos analizar qué sucede con los eventos y las instalaciones atribuidas a Jampp. Parece interesante observar cómo funciona el negocio desde esta perspectiva, intentaremos ver cuáles de los eventos totales son atribuidos a Jampp y haremos lo mismo para las instalaciones, de esta manera intentaremos deducir cómo se desenvuelve la empresa entre la competencia.

Primero veamos los eventos, para esto realizaremos un simple conteo de la cantidad de eventos totales vs. los eventos atribuidos a Jampp, para este análisis nos valemos de la columna 'attributed' de los sets events e installs.

Cantidad de Eventos atribuidos a Jampp:

Attributed	Count
False	2489324
True	5099

Los números difieren tan significativamente que no tiene sentido realizar una visualización ya que estamos comparando poco más de 5 mil atribuidos contra casi 2.5 millones de no atribuidos, la diferencia es clara y nos deja una impresión de cómo resulta ser el mercado.

Por lo analizado en secciones anteriores sabemos que de las millones de subastas, unas 19.5 millones aproximadamente, menos del 12% generan eventos, pues el set events tiene aproximadamente 2.4 millones, y sólo el 0.2 % se atribuyen a Jampp, esto nos da la impresión de ser un mercado altamente competitivos y que además es difícil determinar a quién se le atribuye un evento realmente.

Ahora veamos qué sucede con las instalaciones.

Cantidad de instalaciones atribuidas a Jampp:

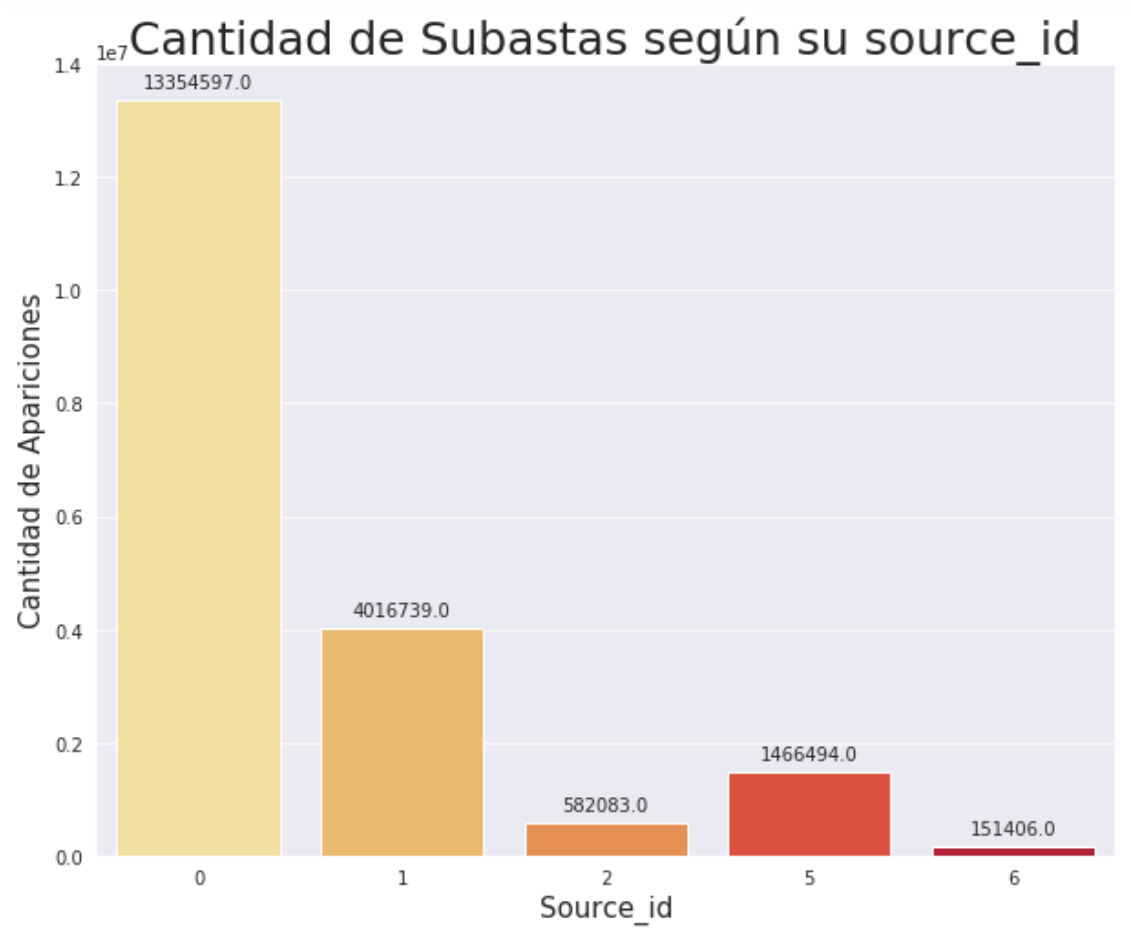
Attributed	Count
False	3412
True	0

Con estos últimos números podríamos concluir que la empresa se desarrolla en un mercado altamente competitivo, tenemos decenas de millones de subastas, unos pocos millones de eventos generados dónde sólo unos miles generan instalaciones y además vemos que es aún más difícil lograr que un evento se atribuya a la empresa y de los cuales no se atribuye ni siquiera una instalación.

7.3 - Análisis sobre el campo 'source_id'

Observando los diferentes sets de datos notamos que 'auctions', el set de datos relacionado a las subastas, cuenta con un campo source_id que refiere a la fuente de la subasta y el set de datos 'clicks' también cuenta con el mismo campo, así que veamos si hay algún tipo de relación entre ellos.

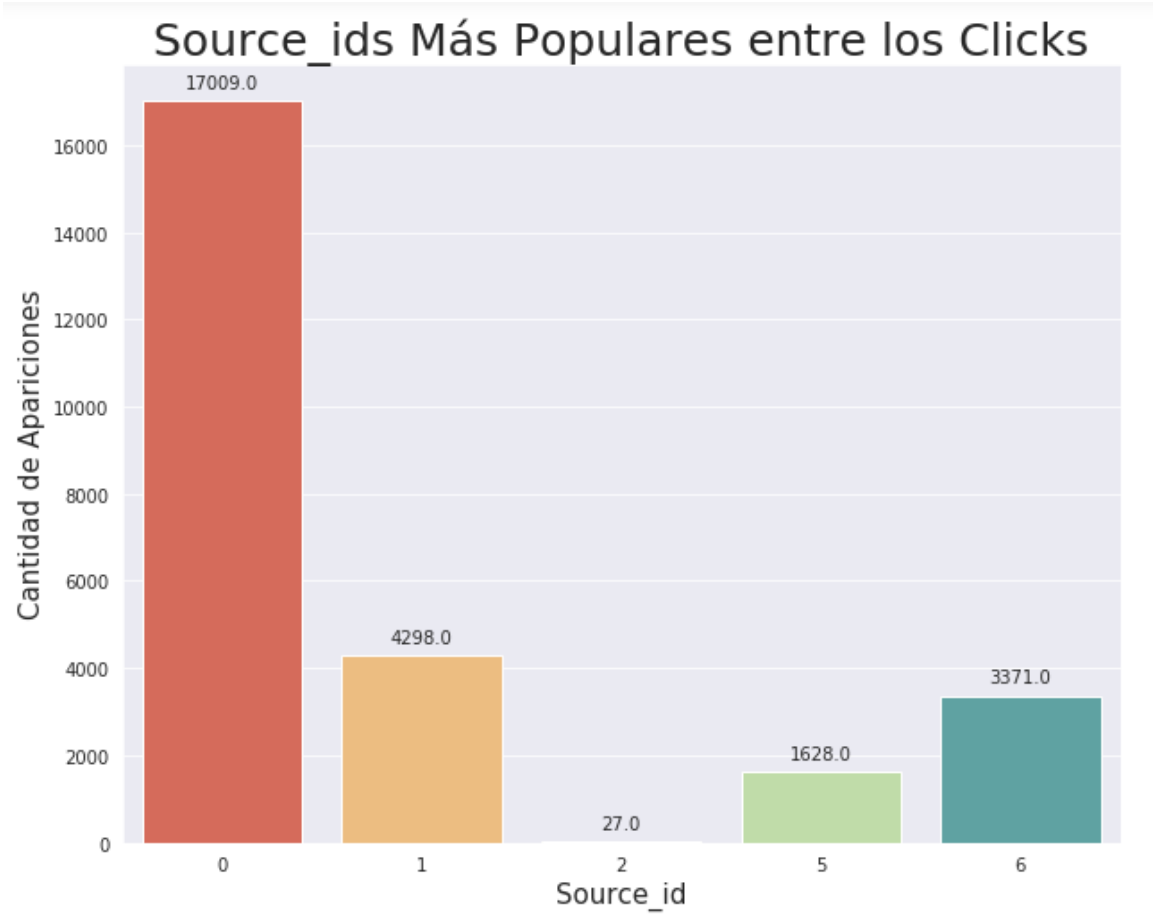
Primero observemos que sucede con las subastas:



En el gráfico observamos la totalidad de source_ids existentes en el set de datos. Como se puede apreciar el id de valor 0 (cero), es la fuente con mayor número

de subastas generadas, con más de 13 millones de subastas al cual le sigue el id 1 con 4 millones.

Veamos si esto se repite también en los datos sobre los clicks registrados:



Lo primero a destacar es que el gráfico no muestra en su totalidad a los source_ids hallados en el set de datos, en la siguiente tabla se puede observar cuales fueron descartados:

Source_id	Cantidad de Apariciones
3	9
10	3
7	2
4	2
9	1
8	1

Los ids que no se tomaron en cuenta son aquellos que tienen menos de 10 clicks, dado que contamos con ids con miles de clicks, no parecía relevante agregar estos ids, más aún los que tienen menos de 5 clicks, no aportan datos significativos ni aportan a la visualización ya que se vuelven insignificantes frente a los otros ids que tienen demasiados clicks.

Analizando el gráfico vemos que el id 0 (cero) es, aquí también, el que más clicks ha generado, este resultado no nos sorprende demasiado, ya que este id genera una gran cantidad de subastas por lo que es bastante lógico que a mayor números de subastas generadas, mayor posibilidades de obtener clicks se obtienen. En general los gráficos se parecen bastante, podemos pensar que la cantidad de subastas generadas por cada fuente repercute luego en la cantidad de clicks obtenidos para las mismas.

Por otro lado, algo que no se puede dejar de mencionar es la diferencia entre los valores, claramente de los 13 millones de subastas generadas por la fuente de id igual a 0 (cero), tan solo 17 mil generaron clicks, una diferencia abismal donde poco más del 0.1% de las subastas generaron clicks.

8 - Conclusiones

De los análisis realizados se pueden sacar las siguientes conclusiones:

- Los usuarios tiene un comportamiento que varía mucho según el día de la semana y el horario. Los días martes y miércoles son los que registran mayor cantidad de eventos , especialmente en el horario de 22 a 3 de la mañana. Por lo cual se debería priorizar subastas en esos días y horarios
- Las instalaciones siguen el mismo patrón anterior. Sin embargo hay aplicaciones que presentan un comportamiento más parejo durante el los días de las semana y otras que presentan grandes variaciones según el día.
- Las instalaciones en general tiene un pico en el horario de 22 a 24 horas al contario de eventos que se extiende hasta las 3 horas del siguiente día. Esto es importante dependiendo si el objetivo de ganar una subasta es maximizar la probabilidad de que el usuario conozca una aplicación o que la instale.
- Dado la variabilidad del comportamiento de los usuarios en el tiempo, se debe definir un monto máximo a ofrecer por cada subasta según el momento exacto de la misma.
- La mayor parte de los clicks en las publicidades mostradas desde las aplicaciones de los clientes de de Jampp realizan clicks en la parte baja del cuadro de publicidad.
- Los clientes con teléfonos de la marca Samsung tienen una mayor probabilidad de hacer instalaciones. Especialmente los que tiene los modelos:G610 (J7), G532 (J2) y J500 (J5).

9 - Algunos Análisis Descartados

A lo largo del desarrollo de este trabajo práctico nos hemos encontrado también con que al realizar un análisis sobre algún campo o conjunto de campos, de los sets de datos, no logramos concluir demasiado luego de invertir unas cuantas horas de estudio en ello, por lo que nos pareció también importante incluir un breve resumen sobre esos análisis y por qué no nos parecieron relevantes para el informe final. De todos modos estos análisis se encuentran a la vista en los notebooks del repositorio en Github del Trabajo Práctico.

9.1 - Merge de installs y clicks

Se buscó relacionar los dataframes de installs y clicks, para analizar la relación entre el entrar en una publicidad de una aplicación a través de un click con el hecho de instalación de dicha aplicación. Para eso se buscó sobre qué columnas se podría hacer merge entre ambos dataframes y asegurarse que en cada fila todas las columnas refirieron al mismo dispositivo, luego ver la diferencia de tiempo entre la fecha de click y la fecha de instalación para ver cuánta tardaba un usuario en instalar la aplicación después de hacer click en la publicidad.

La columna elegida para clicks fue advertiser_id y la columna elegida para installs fue application_id, suponiendo que ambas equivaldrían a la misma aplicación, luego se intentó filtrar por coincidencia de datos de las demás columnas tanto como sea posible, es decir que ref_type de clicks sea igual a ref_type de installs, sin embargo algunas columnas, como agent_device del dataframe de clicks, estaban vacías por lo que no se pudieron realizar un filtro confiable.

Aún así, al comparar la diferencia de tiempo entre click e instalación, se vieron aquellas diferencias de tiempo no mayores a 15 minutos y sólo se recuperaron 4 filas con datos, por lo que se entiende que o bien son muy pocos los usuarios que instalan las aplicaciones luego de hacer click en su respectiva publicidad o bien advertiser_id y application_id representan cosas distintas.

9.2 BaseMap Para Latitud y Longitud

Dentro del set de Datos de Clicks contábamos con los campos Latitude y Longitude, como al comenzar el análisis desconocíamos el país del cuál provenían los datos se pensó en intentar deducir la ubicación, a algo respecto de ella, a partir de estos campos.

Primero se observó el comportamiento tanto de la Latitud como la Longitud, concluyendo que la mayoría de los puntos caían en un radio muy pequeño, es decir, eran muy similares sus valores.

10 - Link a Repositorio de Github

https://github.com/EscobarMariaSol/Organizacion-de-datos