TP 1: Análisis Exploratorio

Organización de Datos 2019

Integrantes: Nilda Dombrovsky, José Rivas y Florencia D'Andrea - Grupo 36 (Oyentes)
Repositorio de Github:
https://github.com/flor14/Organizacion_de_datos_2019/blob/master/Tp1_datos.ipynb

---

1. Introducción

El mundo de la publicidad móvil tiene contacto directo con mucha información, esto es debido a que hoy en días las aplicaciones que instalamos y utilizamos en nuestros dispositivos forman parte de micromomentos diarios, donde estamos generando datos importantes que son son utilizadas en el marketing digital.

Jampp es una empresa que en 2013 asumió el reto de explotar el mercado de la publicidad móvil, con la visión de ayudar a los anunciantes a crear clientes leales a la marca, considerando que la inversión a dicha publicidad no puede ser mayor a las ventas de las aplicaciones móviles (apps), en ese sentido consiguen espacios publicitarios en sitios de subastas y para impulsar las ventas de lo que ofrece cada apps hacen foco en el retargeting, estos desafíos son alcanzados con el uso de técnicas de aprendizaje automático y el enfoque que se necesita en cada industria.

La materia Organización de Datos de la FIUBA logró una alianza con Jampp para que se haga este primer trabajo práctico, donde el objetivo es realizar un análisis exploratorio sobre cuatro datasets, a saber: Auctions, Clicks, Installs y Events. A continuación, un breve resumen de la información contenida en cada dataset correspondiente a un periodo acotado de tiempo, comprendido entre las fechas 2019-03-05 y 2019-03-13, en Uruguay:

- Auctions: generado por el RTB (real-time bidding), cada fila representa una subasta;
- Clicks: cada fila representa un click realizado por el usuario en la impresión de una publicidad;
- Installs: cada fila representa una instalación realizada por un usuario, ya sean atribuidas o no a las impresiones de publicidad de Jampp;
- Events: muestra una observación de la tabla en relación a las acciones que hace cada usuario dentro de la aplicación ya instalada.

A partir de este análisis pretendemos descubrir patrones que puedan resultar interesantes para, en una segunda parte, lograr determinar para un instante dado, el tiempo hasta que un dispositivo d aparezca de vuelta en una subasta RTB, y para un instante dado, el tiempo hasta que un dispositivo d instale una aplicación.

2. Análisis Exploratorio

2.1 Dataset Clicks

Cada fila del dataset clicks representa un click realizado por el usuario en la impresión de una publicidad, contando Clicks con 26351 filas y 20 columnas. Las columnas que no presentaron valores nulos fueron “advertiser_id”, “source_id”, “created”, “country_code”, “latitude”, “longitude”, “wifi_connection”, “trans_id”, “specs_brand”, “ref_type” y “ref_hash” (Tabla 1). Mientras que la columna “action_id” sólo posee nulos. Los valores de las columnas “timeToClick”, “touchX” y “touchY” presentaron una cantidad de valores nulos similar, alrededor de 3300 sobre el total de la base de datos.

|

Tabla 1. Cantidad de nulos por columna para el dataset Clicks

columnas	nulos
advertiser_id	0
action_id	26351
source_id	0
created	0
country_code	0
latitude	0
longitude	0
wifi_connection	0
carrier_id	11
trans_id	0
os_minor	12
agent_device	23108
os_major	12
specs_brand	0
brand	20116
timeToClick	3374
touchX	3340

touchY	3340
ref_type	0
ref_hash	0

El 99.66% de los clicks de esta base de datos estuvieron relacionados con publicidades del “advertiser 3”. La figura 1 permite observar que: (1) sólo hubo clicks todos los días del período estudiado para el advertiser 3 y el 0, (2) de ambos, el advertiser 3 presenta un aumento de tres órdenes de magnitud en la cantidad de clicks a partir del día 7 de marzo, lo cual no ocurre para el advertiser 0, (3) el día máximo de clicks ocurrió el día 12 para el advertiser 3, que obtuvo 5198 clicks y (4) los días 5 y 6 de marzo de 2019 presentaron considerablemente menos cantidad de clicks. Eliminar los valores correspondientes a los advertisers distintos al 3 no modifica el patrón general de datos que implica pocos valores los días 5 y 6 de marzo, y luego un importante aumento en la cantidad de clicks.

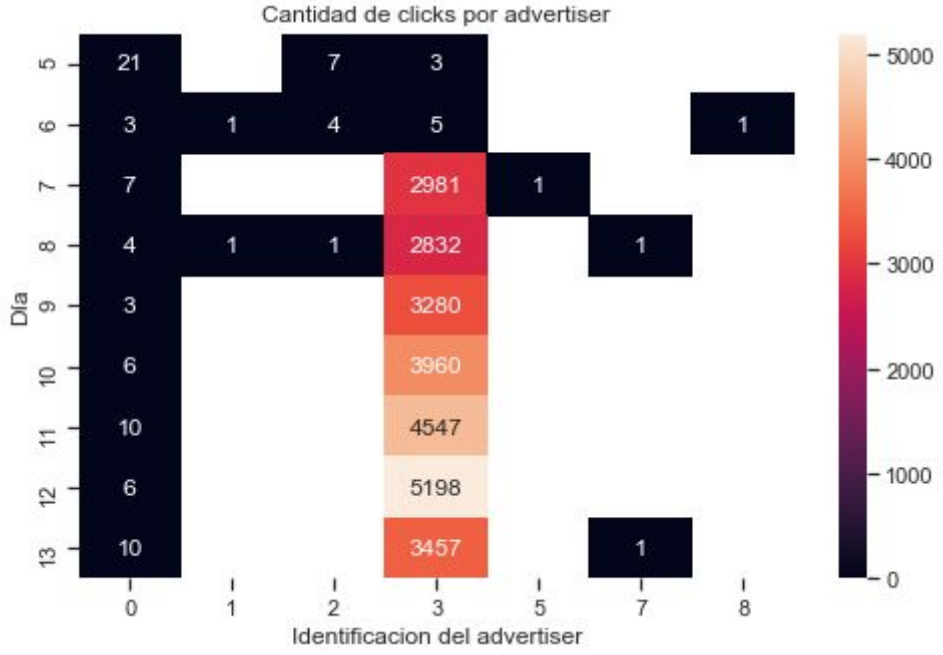


Figura 1. Cantidad de clicks por día para publicidades correspondientes a cada advertiser.

Los valores de “advertiser_id” no presentaban nulos en esta base de datos, por lo tanto en el análisis de la figura 1 estamos considerando la totalidad de los clicks. Resulta particularmente llamativo reconocer que la variable con valores nulos “tiempo hasta el click” (“timeToClick”) (Tabla 1), parecería tener una gran cantidad de faltantes el día 7 de marzo para el cual se habían observado 2981 reportes de clicks (Fig. 1), y apenas se observan algunos puntos con “tiempo hasta el click” reportado (Fig 2). Si bien se tratará más adelante ya se observa que la mayor cantidad de observaciones se encontró ligada a valores de “tiempo hasta el click” más bajos (Fig 2). Además se observaron cuatro franjas horarias en las cuales no se reportó el valor de tiempo hasta el click. Un patrón similar se observó de graficar “touchY” y “touchX” a lo largo del tiempo (datos no mostrados).

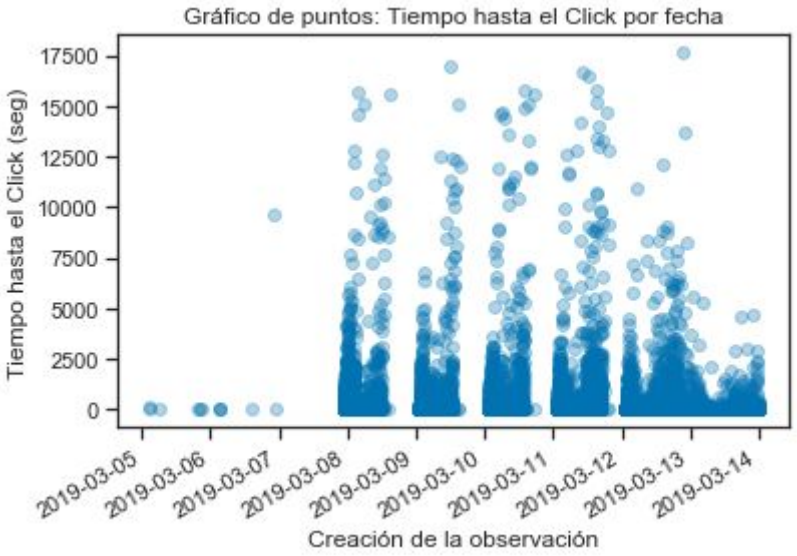
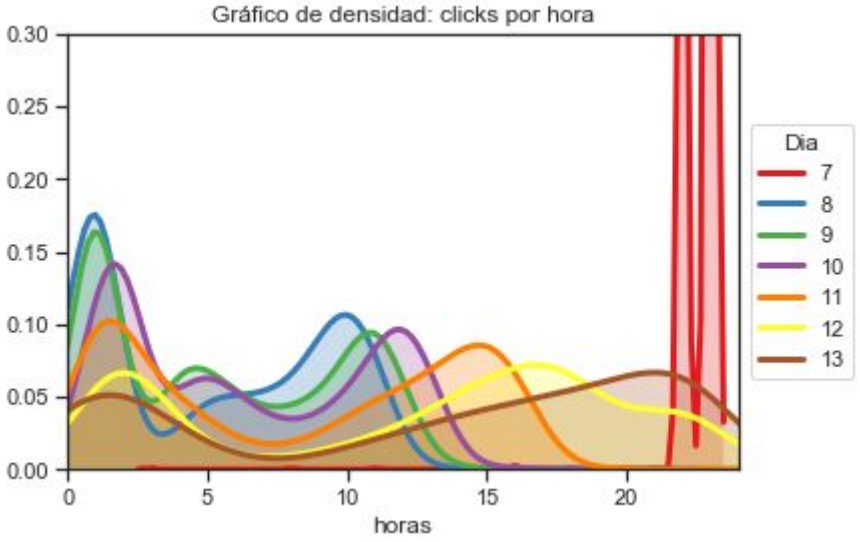


Figura 2. Tiempo hasta el click para cada observación realizada en el período de análisis.

Para determinar cómo se distribuyen los clicks a lo largo de cada día, se procedió a evaluar la densidad de clicks por hora, incluyendo o no incluyendo el día 7 (Fig. 3 a y b). Los días 5 y 6 fueron removidos del análisis ya que consideramos no muestran ningún patrón específico. Se observa en la figura 3a que el día 7 presenta la totalidad de los 2981 clicks en un rango muy corto de horas, a diferencia de lo que ocurre el resto de los días elegidos en este análisis. Como generalidad para el resto de los días (Fig 2b) se observa que la mayoría de los clicks ocurren durante la madrugada, existiendo un segundo pico alrededor de las 10 de la mañana para todos los días, si bien este no se da a la misma hora. Es importante recordar que al estar observando la densidad de los datos no estamos evaluando la cantidad de clicks por día sino simplemente su distribución.

a)



b)

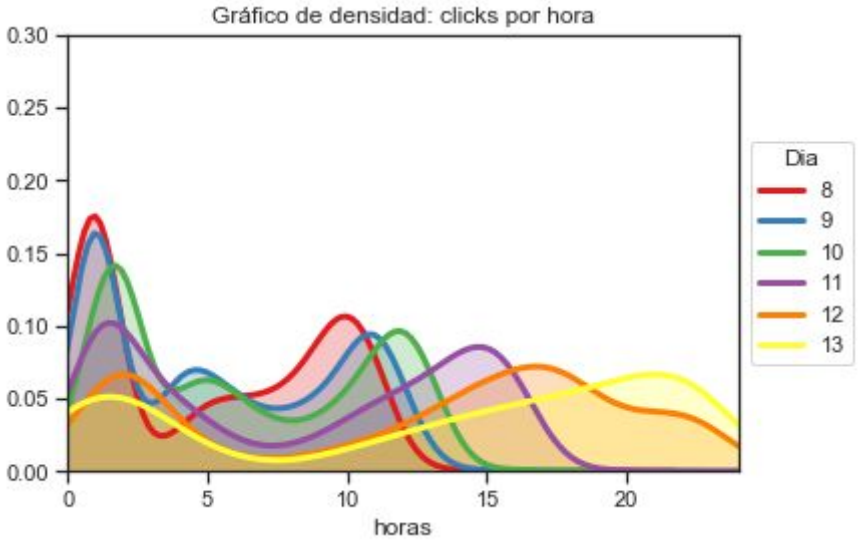


Figura 3. Densidad de clicks por hora para cada uno de los días del período (a) 7 al 13 de marzo de 2019 y (b) 8 al 13 de marzo de 2019.

La distribución de datos el día 7 resulta particular (Fig. 3a) por lo que realizamos algunas caracterizaciones extras para intentar entender qué ocurría este día que lo diferenciaba del resto. Los valores de las coordenadas geográficas para los clicks del día 7 mostraron que 2586 clicks (86,7% del total) provinieron de un teléfono en la coordenada geográfica representada por los valores de latitud 1.20568 y longitud 1.07023 (Fig. 4). Para obtener más detalles sería de interés conocer a qué coordenadas reales (no estandarizadas) corresponden estos valores.

Considerando que la mayoría de estos clicks se realizó desde las mismas coordenadas, lo que podría indicar que provienen de un único teléfono, y en un rango de horario muy limitado nos lleva a pensar la posibilidad de encontrarnos frente a un caso de fraude. Esto implica que un único dispositivo envía una cantidad de clicks muy alta en un tiempo muy corto, tal como observamos que ocurrió en el día 7.

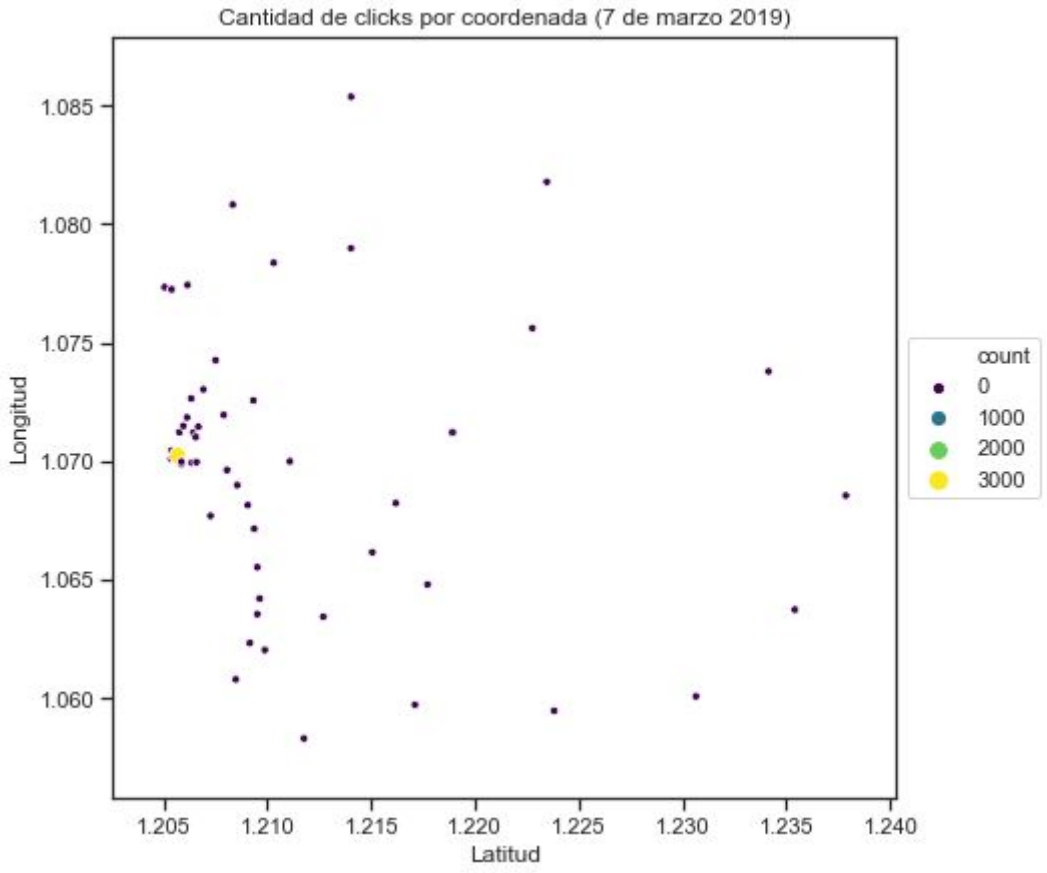
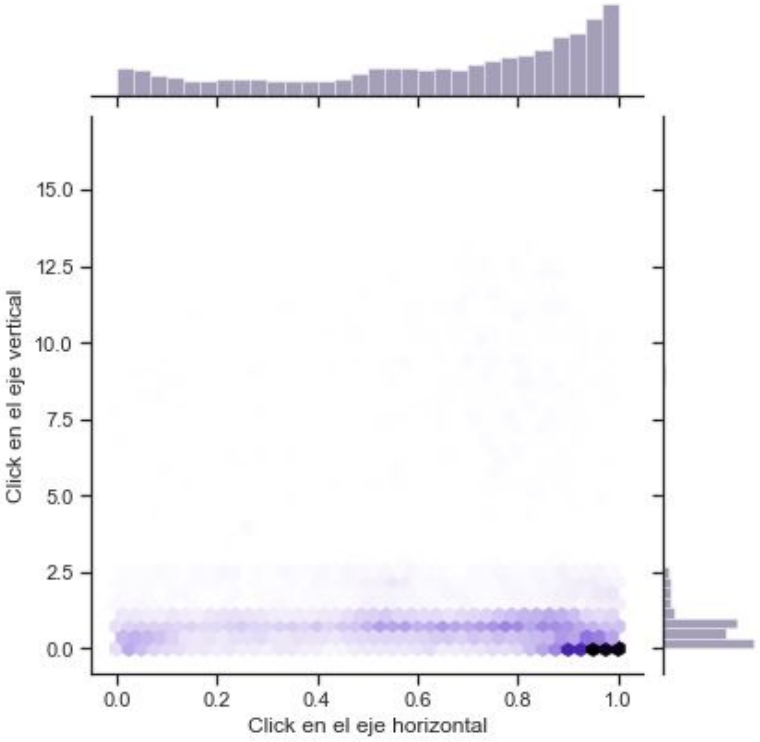


Figura 4. Cantidad de clicks provenientes de distintas coordenadas geográficas para el día 7 de marzo.

Si consideramos la densidad de clicks en relación a las dos dimensiones de la pantalla del teléfono observamos que la gran mayoría de los mismos se encuentra centrada en un punto correspondiente a los mayores valores del eje X y menores valores del eje Y, en escala adimensional y dentro de una franja de mayor densidad (Fig. 5 a). Como no contábamos con información de dónde se encontraba la publicidad situada en la pantalla, se generó otro dónde se amplió el área de mayor densidad de puntos comprendida entre los valores de 0 y 1 del eje Y (Fig 5b). Si bien no son notorios por su baja cantidad en la Fig. 5a, hubo clicks dispersos por fuera de este área de mayor densidad.

En particular el punto de color de mayor intensidad que se observa en la figura 5a y b podrían indicar que en esta zona se encuentra el botón de cierre de la publicidad. Realizar un análisis de los clicks que efectivamente producen ingresar a la publicidad implicaría eliminar de la base de datos los que ocurran en las coordenadas del botón de cierre de la aplicación. Sería de utilidad tener acceso a las impresiones que reciben los usuarios para realizar mayores análisis en esta dirección.

a)



b)

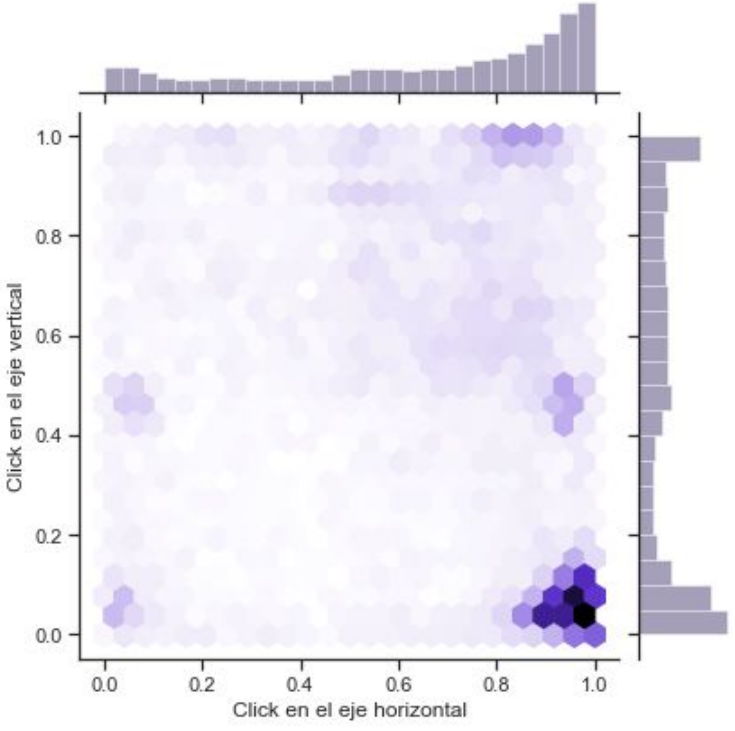


Figura 5. Densidad de clicks realizados en la pantalla del teléfono para el eje Y (a) considerado de 0 a 15 (b) de 0 a 1. Cuanto mayor la intensidad del color de los hexágonos mayor la cantidad de clicks en esa área. Las frecuencias de datos se muestran en relación a cada uno de los ejes.

El 75% de los datos de clicks presentaron un valor de tiempo hasta el click menor a 71 segundos (Tabla 2), siendo el mayor valor registrado 17616 segundos. La figura 6 muestra el histograma para “Tiempo hasta el click” donde se puede observar que la distribución de los valores son marcadamente asimétricos, indicando que los clicks tienden a realizarse a los pocos segundos de mostrarse la impresión en la pantalla. Para obtener una visualización representativa se eligió como límite máximo en el eje x el cuartil que representa el 90% de los datos, dejando fuera del histograma solo los valores más extremos.

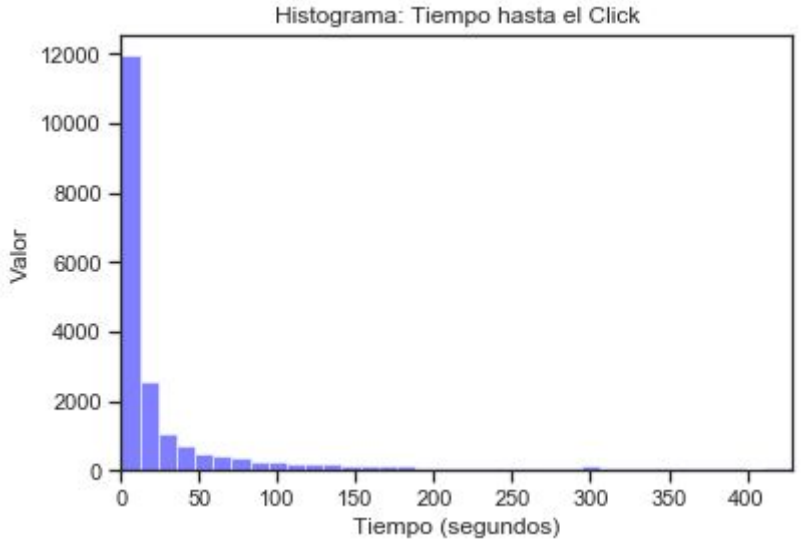


Figura 6. Histograma para “tiempo hasta el click”

Tabla 2. Estadística descriptiva para los valores de Tiempo hasta el Click

	timeToClick
count	22977
mean	230.4033095
std	976.8491489
min	0.017
25%	2.915
50%	10.588
75%	71.703
max	17616.188

Si analizamos los clicks por día en relación al tiempo hasta el click observamos que los valores más extremos se encontraron entre el 8 y el 12 de marzo de 2019 (Fig 7).

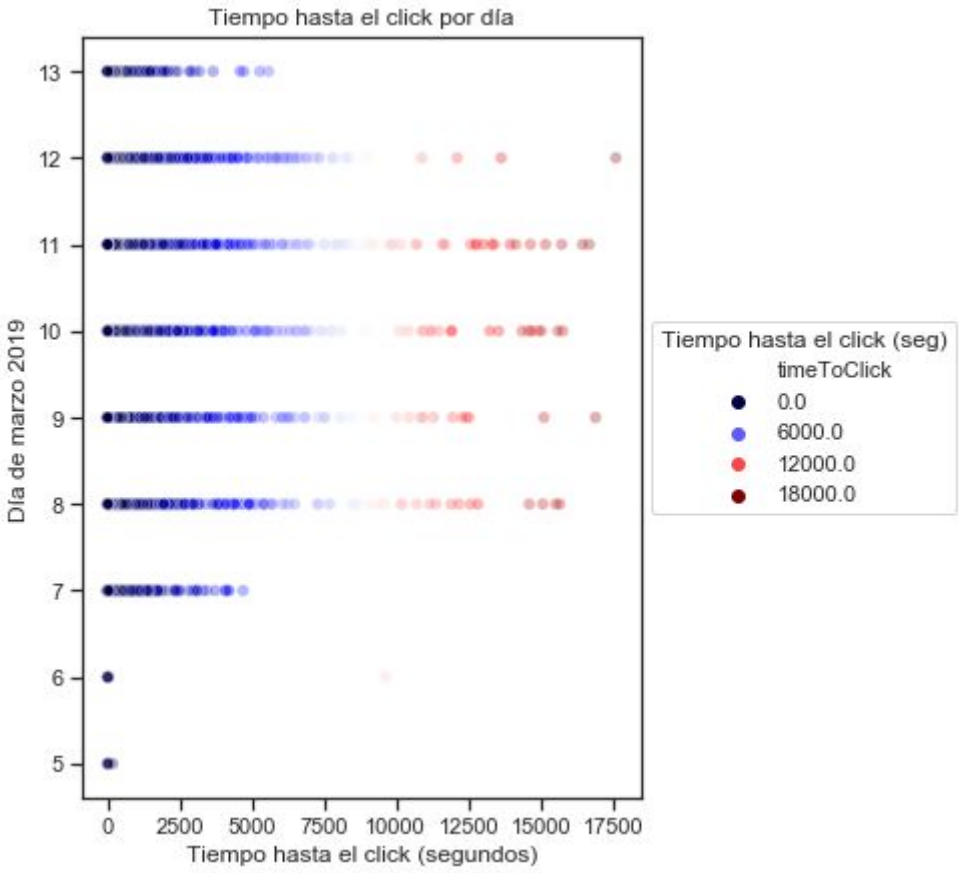


Figura 7. Tiempo hasta el click para los clicks realizadas de forma diaria.

2.2 Dataset Installs

El dataset Installs cuenta con 3412 filas y 18 columnas. En particular, las columnas que no presentaron valores nulos fueron “created”, “application_id”, “attributed”, “implicit”, “device_countrycode” e “ip_adress”. La columna “click_hash” no presentó ningún valor reportado. De las instalaciones en este período estudiado ninguna fue atribuida a Jampp, mientras que un 25% fue considerada una instalación implícita. El 75% fue de las observaciones de este dataset no corresponden ni a una instalación atribuida a Jampp ni implícita.

En la figura 7 se comparó la densidad de observaciones en el tiempo para el dataset clicks con las instalaciones. Se observa que el dataset de instalaciones presenta registros previos al comienzo de la mayor cantidad de clicks, que como ya observamos en la figura 1, recién ocurrió el 7 de marzo. Se puede observar un leve aumento de la cantidad de instalaciones el día 12 que se corresponde con el de mayor cantidad de clicks el mismo día (Fig 7). El desfase de las instalaciones con los valores de clicks los primeros días resultan interesantes de observar, más aún sabiendo que los valores de clicks del 7 pueden ser fraudulentos, y no por clicks que vayan a producir instalaciones reales. Tal vez el aumento de instalaciones del día 12 sea el único que se corresponde con un aumento real sobre clicks que llevarían a la instalación de la publicidad.

Se probó graficar la densidad separando las instalaciones implícitas y no implícitas pero no se observaron resultados concluyentes por lo cual no fue incluído como figura en el informe (gráfico en la notebook).

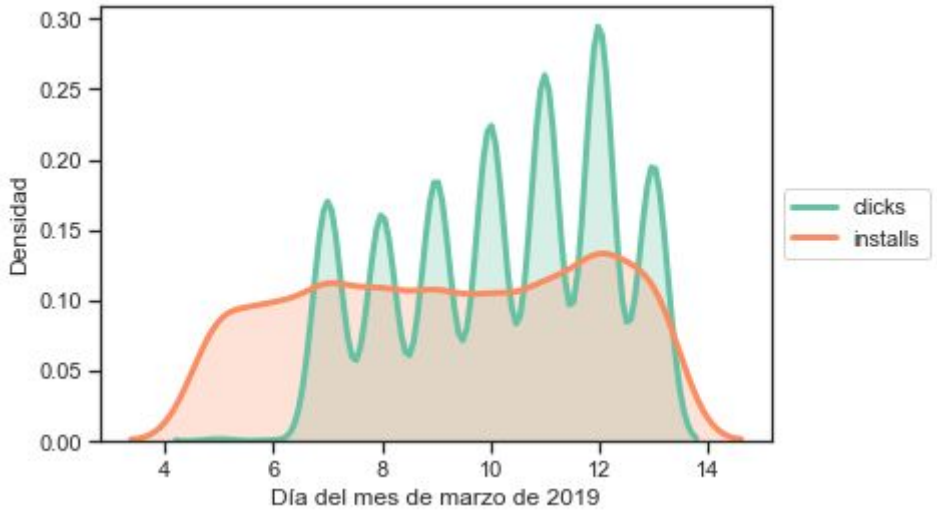


Figura 7. Densidad de clicks y de instalaciones por día para el período comprendido entre 2019-03-05 y 2019-03-13.

La cantidad de instalaciones es considerablemente menor a la cantidad de clicks (Fig 8), lo cual no se observa en el gráfico de densidad (Fig. 7).

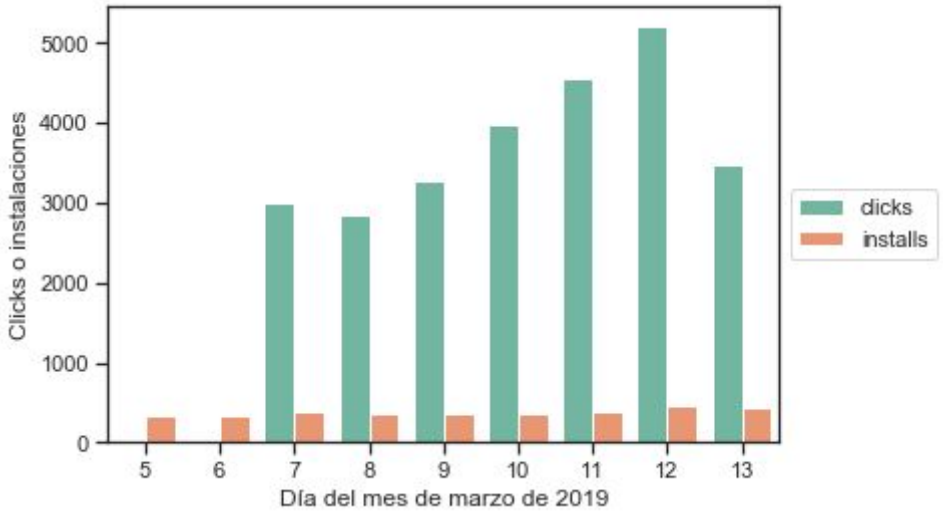


Figura 7. Cantidad de clicks y de instalaciones por día para el período comprendido entre 2019-03-05 y 2019-03-13.

2.3 Dataset Events

El dataset Events cuenta con 2494423 filas y 22 columnas. Las columnas que no presentaron valores nulos fueron “date”, “event_id”, “ref_type”, “ref_hash”, “application_id”, “attributed”, “device_countrycode” y “ip_address”.

Estos eventos están asociados al comportamiento de los usuarios con la aplicación instalada, donde algunos pueden ser atribuidos a Jampp.

Al parecer por temas

Evaluación de la cantidad de valores nulos:

Tabla 3. Cantidad de nulos por columna para el dataset Events

columnas	nulos
date	0
event_id	0
ref_type	0
ref_hash	0
application_id	0
attributed	0
device_countrycode	0
device_os_version	1472357
device_brand	1329460
device_model	87967
device_city	1879725
session_user_agent	11786
trans_id	2494341
user_agent	1102896
event_uuid	5099
carrier	1877989
kind	5099
device_os	1836756
wifi	1115551
connection_type	1881960
ip_address	0
device_language	87819

En la Tabla3, se pueden ver los eventos que son atribuidos a Jampp, estos se pueden observar en la columna “attributed” pero hay muy pocos.

Tabla 4. Cantidad de eventos atribuidos a Jampp.

attributed	cuenta
FALSE	2489324
TRUE	5099

En la Tabla 4, los registros asociados a TRUE sólo conforman el 0.2%.

La columna “application_id”, nos puede dar una idea las aplicaciones con mayor cantidad de eventos, lo que intentamos explorar son las atribuidas a Jampp, la siguiente tabla representan el 92,9% de las mismas.

Tabla 5. Cantidad de eventos por “application_id” atribuidos a Jampp

application_id	cuenta eventos
63	2323
16	1219
45	431
170	323
102	150
77	135
244	125

En la tabla 5 se puede ver un top de “event_id” por “application_id”, por lo que estos serían los eventos que Jampp estaría impulsando con su retargeting.

Tabla 6. Top de “event_id” por “application_id” de eventos atribuidos a Jampp

application_id	event_id	cuenta
77	204	100
63	31	1253
45	41	213
244	1	46
170	155	121

16	513	1004
102	133	129

En la Tabla 6, se pueden observar cuales fueron los eventos más predominantes, por “application_id”.

Conociendo los datos reales de estos eventos se pueden tomar para

2.4 Dataset Auctions

Cada observación o fila del dataset de Auctions presenta una subasta, seleccionada de la base de datos de Jampp entre la fecha 2019-03-05 y 2019-03-13 para Uruguay. Es por ello que la columna "country" solo presenta un único valor correspondiente al código de dicho país y no consideramos que brinde información relevante.

Evaluación de la cantidad de valores nulos:

```
auction_type_id          0
country           19571319
date              19571319
device_id         19571319
platform          19571319
ref_type_id       19571319
source_id         19571319
date2             19571319
dtype: int64
```

Sólo “auction_type_id” tiene nulos en todos sus registros, el resto de las columnas están completas.

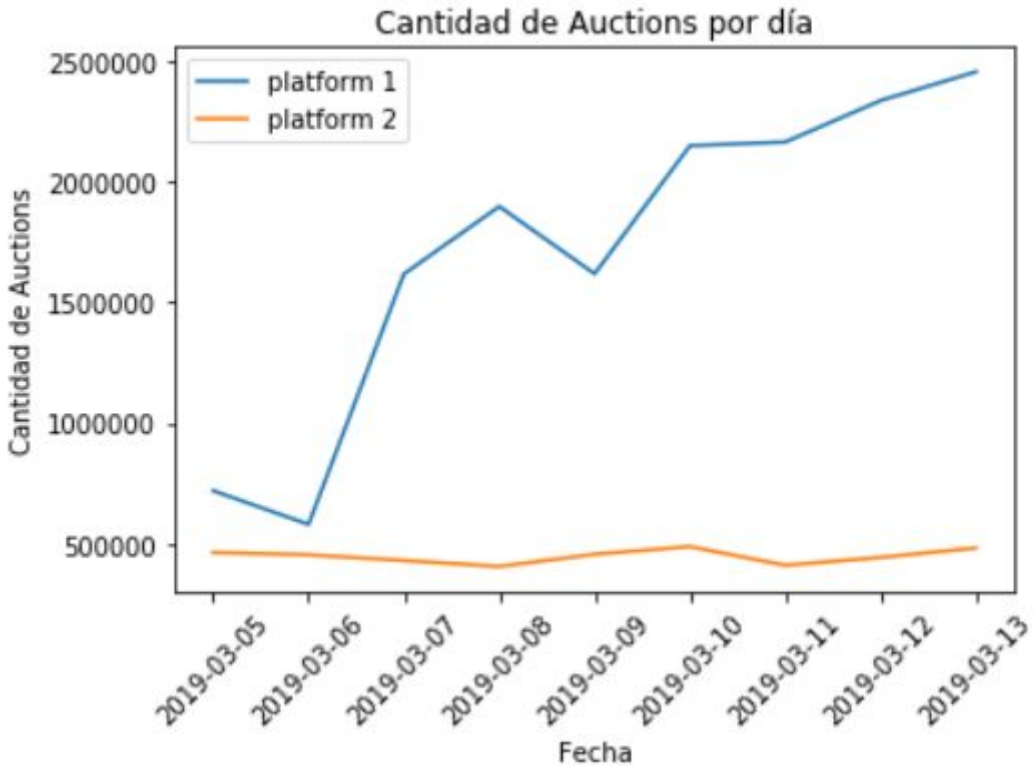


Figura 8. Cantidad de auctions por día por plataforma.

En la figura 8, se puede observar que en el proceso de subasta se observa un aumento en el movimiento de las subasta para la plataforma 1, se considera que se presentaron grandes oportunidades para que el sistem de Jampp inicie un proceso de auctions para conseguir alguna adjudicación.

Al no tener algo que nos indique que tipo de acción fue cada registro, nos imposibilita poder hacer un tracking del proceso de subasta, de igual forma no se puede determinar de cuántas subastas realizadas se logró alguna adjudicación.

3. Conclusiones

Las plataformas digitales han crecido de forma exponencial, trayendo consigo una gigantesca cantidad de datos que ameritan ser analizados, por lo que se vuelve importante el manejo de nuevas tecnología que permitan manipular estos datasets. Este trabajo práctico permitió obtener información de valor lo cual es una necesidad fundamental, debido a que hoy en día las empresas toman decisiones basadas en los datos.

El análisis del dataset clicks permitió comprender la dinámica temporal de las variaciones de las interacciones de los usuarios con las impresiones de la publicidad y también el patrón espacial de los clicks sobre la pantalla del teléfono. Estudiando los clicks realizados se logró determinar un caso de fraude observado inicialmente el día 7 de marzo debido a la alta cantidad de clicks en una misma franja horaria y provenientes de una misma ubicación. Por otro lado, el estudio del dataset installs reveló que no hubo durante este período instalaciones atribuidas a Jampp. Tampoco se observó una complementariedad al momento de comparar las instalaciones con los clicks, excepto tal vez el 12 de marzo donde se observa un aumento tanto de los clicks como de las instalaciones, siendo esto esperable considerando la posibilidad de fraude y la baja tasa de instalaciones atribuidas. La cantidad de clicks puede estar relacionada con eventos de cierre de la aplicación, haciendo más pronunciada aún la diferencia con las instalaciones. Con respecto al dataset Events se pudo observar que existen diferentes tipos de eventos que son realizados por los usuarios cuando interactúan con una aplicación móvil, pero los casos de éxitos son muy bajos para ser atribuidos a Jampp, sabiendo cuáles serían esos eventos probablemente se puede idear una estrategia para retargeting por último con el dataset auctions se observa un comportamiento a una plataforma eso quiere decir que la plataforma de Jampp consiguió una oportunidad para subastar para hacerse de un espacio publicitario.