FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Trabajo Práctico Nro. 1
Análisis Exploratorio

Eventos Jampp

[7542] Organizacion de Datos
Primer cuatrimestre de 2019

Alumno:	**Padrón**
ROTILI, German	99722
ALVAREZ JULIA, Santiago	99522
MONTES, Gaston	89397

Índice

1. Introducción

Jampp es una plataforma de **Performance Marketing**, pues su negocio está basado en el uso de tecnología predictiva para minimizar costos y maximizar las ganancias del **Advertiser**, en 2 tipos de campañas de marketing: **User Acquisition** y **Retargeting**. En el presente trabajo se realiza un análisis exploratorio de los datos que nos facilitó dicha plataforma, basados en eventos realizados por los usuarios que utilizan las aplicaciones partner de *Jampp* y la información que le proveen los Ad Exchanges.

2. Análisis Exploratorio

En los siguientes items analizaremos la información que a nuestro criterio es interesante según el archivo de datos en el que se encuentra. *Jampp* nos acercó 4 archivos:

2.1. Clicks

Como un primer análisis se da un 'vistazo general' al conjunto de datos. Éste contiene 26.351 registros y 20 atributos. Los datos no están ordenados cronológicamente (es decir según columna 'created'), pero se obtienen datos desde 2019-03-05 01:17:30 hasta 2019-03-13 23:59:59.

La frecuencia de los clicks registrados está detallado en la Figura 1, donde los primeros 2 dias casi no se registran clicks y el resto de los dias se mantiene la frecuencia entre 3000 y 5000 clicks aproximadamente. El dia con mayor cantidad de clicks fue el Martes 12 de marzo.

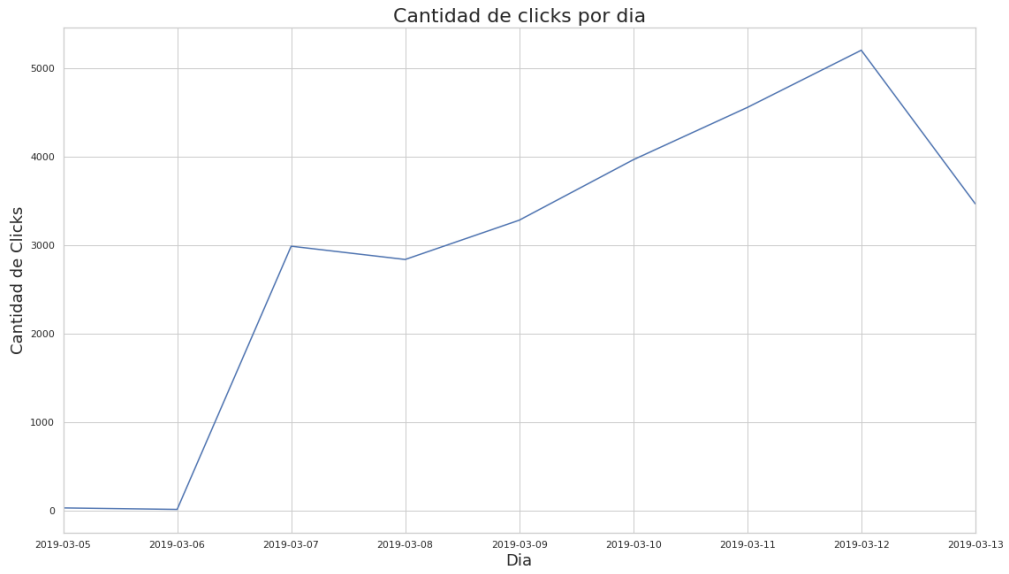

Figura 1: Cantidad de clicks según el dia del año.

A través de su ID, podemos obtener el detalle de las personas que se incluyen en este análisis. Por lo tanto, la Figura 2 muestra las personas con mayor cantidad de clicks en el set de datos. De esta manera, *Jampp* podría saber cuales son las personas con mayor conversión y posiblemente sus mejores clientes.

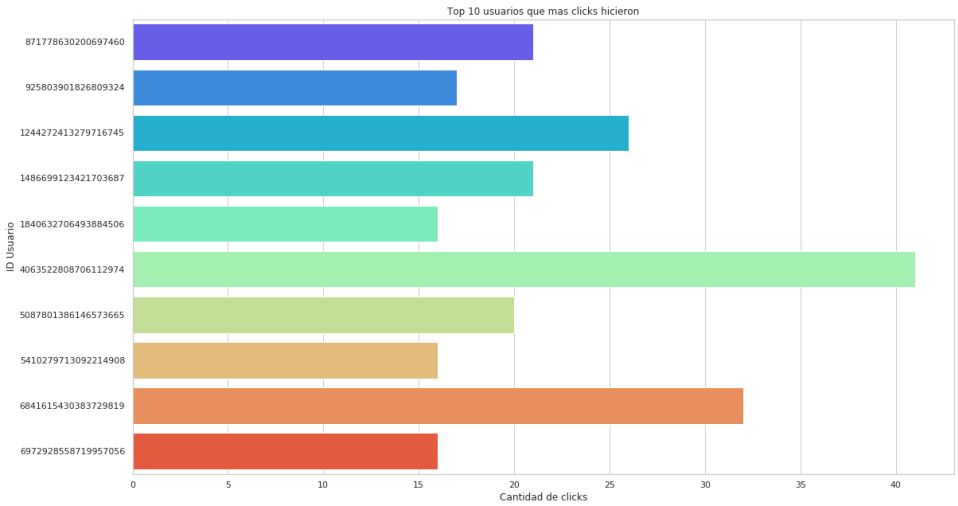

Figura 2: Las 10 personas con mayor cantidad de clicks a lo largo de todo el período de datos disponibles.

Sería interesante evaluar según el momento del día cuanto tarda en promedio un usuario en hacer click sobre una publicidad. El resultado es visible en la Figura 3, donde se observa que el menor promedio en cantidad de segundos se produce alrededor del mediodía, sin embargo no parece haber una clara relacion entre el momento del día y la cantidad de segundos que se tarda en hacer click sobre una publicidad.

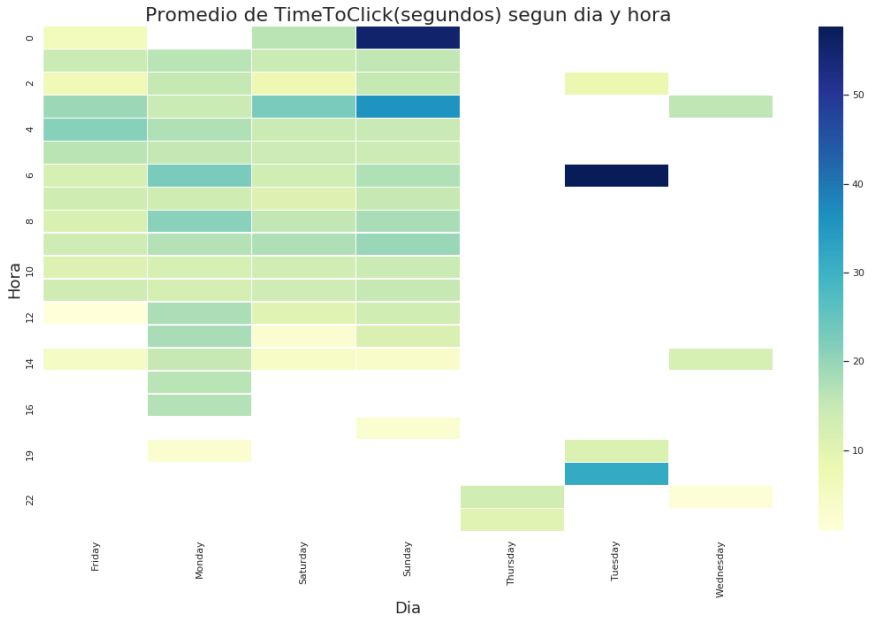

Figura 3: Promedio en segundos que tarda un usuario en hacer click sobre un ad.

Los Ad Exchanges proveen a Jampp de data del usuario, como por ejemplo la marca del teléfono. Por lo tanto analizamos la cantidad de clicks que se generan según la marca del dispositivo (para hacer mas claro el gráfico se realizó el análisis para un único Advertiser, el nombrado anteriormente cuyo ID es 3). En la Figura 4 se ve que la marca con mayor cantidad de clicks es la 2.

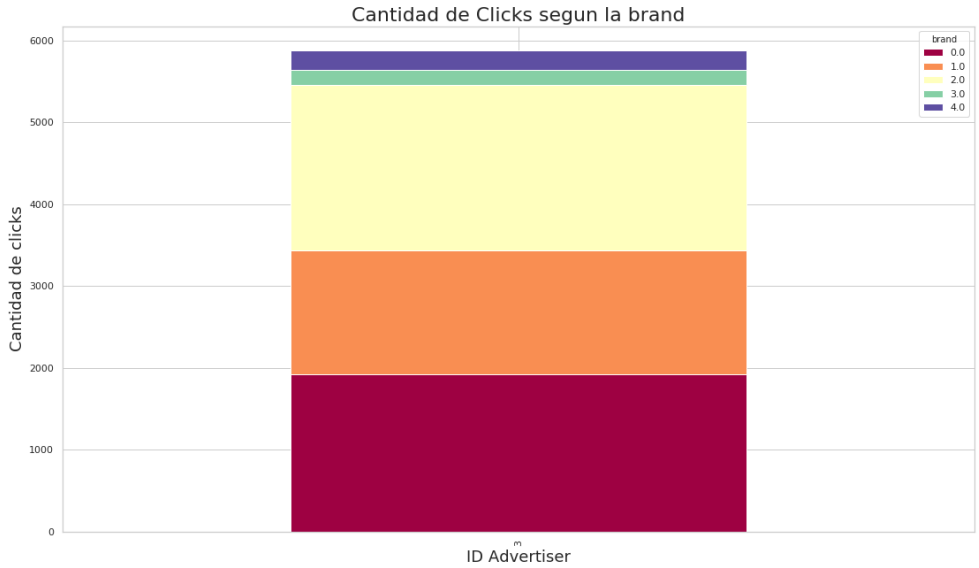

Figura 4: Cantidad de clicks según la marca del dispositivo del usuario.

Respecto a los Ad Exchange en si mismos, analizamos cuales son los que producen mayor cantidad de clicks según la hora del día. En la Figura 5 se aprecia un apabullante ganador en 20 de las 24 horas del día, el Ad Exchange cuyo ID es 0.

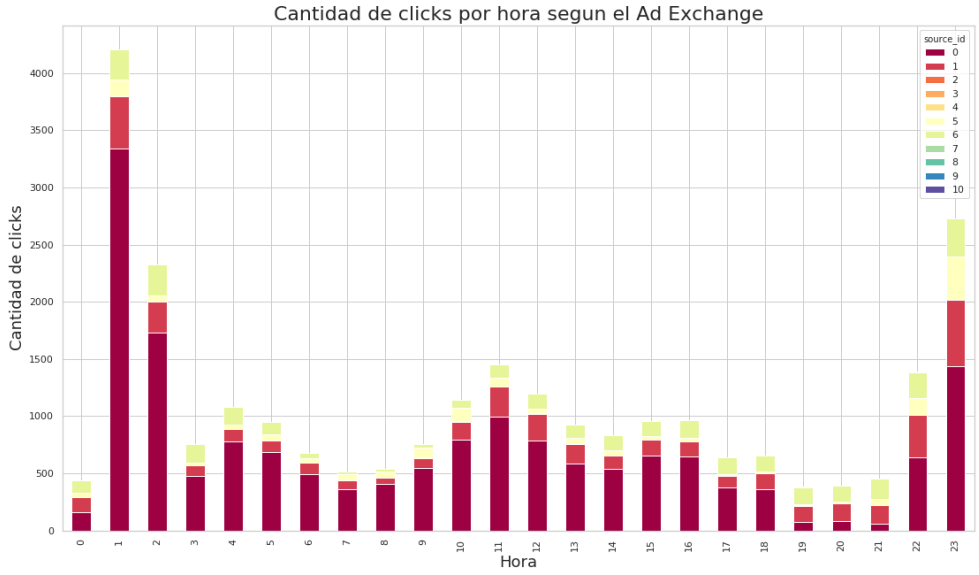

Figura 5: Cantidad de clicks por hora según el Ad Exchange.

A partir del atributo 'advertiser_id' del set de datos obtuvimos que un Advertiser sobresale por sobre el resto respecto a la cantidad de clicks que se hicieron sobre sus publicidades: 26262 clicks, casi la totalidad de clicks registrados en los datos que nos facilitó la gente de *Jampp*. El ID de dicho advertiser es el numero 3.

De manera similar al Advertiser, en la data sobresale un 'ref_type'(es un tipo de ID que representa al usuario y su dispositivo en advertising segun si es Apple o Google) = 1891515180541284343, presente en 25549 registros.

Un dato curioso se dió al querer analizar una posible relación entre el tipo de conexión a internet del teléfono a la hora de hacer click en la publicidad y la cantidad de clicks total en el set de datos. Uno podría pensar que la mayoria de usuarios hacen click en publicidades estando conectados a una red wifi porque en el caso contrario le consumiría el plan de datos del teléfono. Pero en todos los clicks del set el teléfono no se encontraba conectado a una red wifi. El hecho de que no alla ni siquiera uno podría significar que *Jampp* filtro los datos que nos entregó.

Para tener una idea de que parte del mundo provienen los datos, analizamos la columna 'country_code'. Todos los clicks provienen del mismo país que luego nos confirmo el representante de *Jampp* que se trata de Uruguay (también confirmo que filtraron su data para que nosotros solo tengamos datos de Uruguay).

2.2. Eventos

Como un primer análisis se da un 'vistazo general' al conjunto de datos. Éste contiene 2.494.423 registros y 22 atributos. Los datos no están ordenados cronológicamente (es decir según columna 'date'), pero se obtienen datos desde 2019-03-05 00:00:00 hasta 2019-03-13 23:59:59.

La frecuencia de los eventos registrados está detallado en la Figura 6, donde dia a dia se registran mas clicks excepto por el Lunes 11 de Marzo hasta finalmente llegar al 13 de Marzo, el dia con mayor cantidad de eventos.

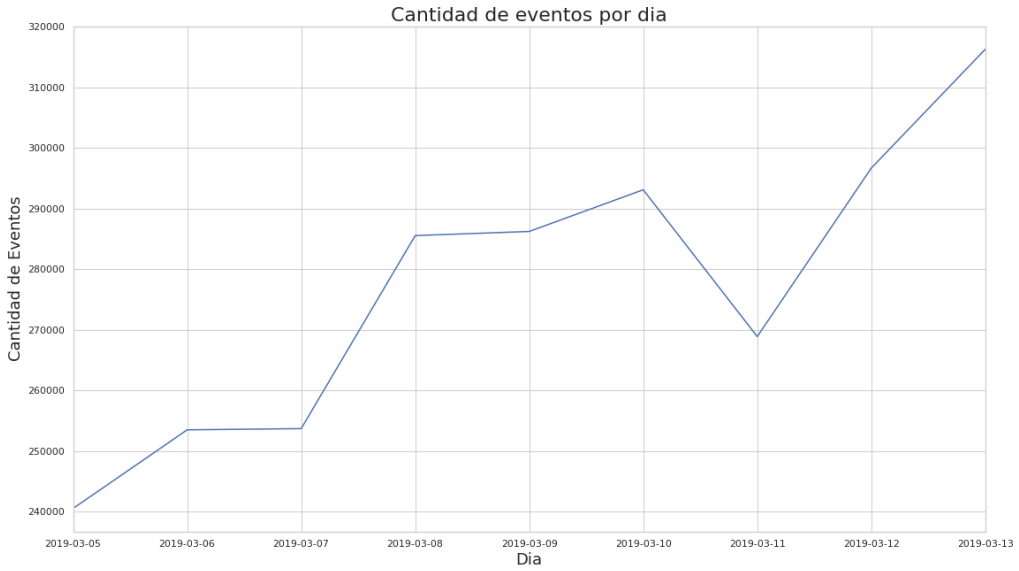

Figura 6: Cantidad de eventos según el dia del año.

A su vez las aplicaciones partners categorizan a los eventos y aunque no podamos saber con exactitud a que tipos de eventos se refieren, podría ser interesante saber cuales son los tipos de eventos mas registrados por estas aplicaciones.

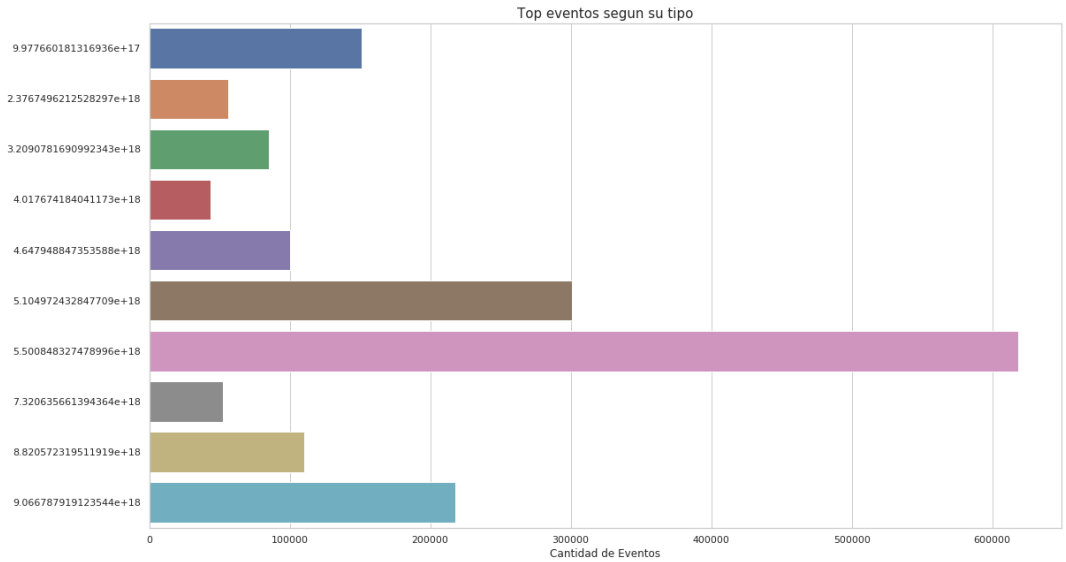

Figura 7: Frecuencia de los distintos eventos en el conjunto de datos.

Por otro lado, también sería interesante evaluar en qué momento del día suceden estos eventos durante la semana. Observar en qué horarios las aplicaciones partner tienen mayor tráfico de datos. Los eventos parecen estar distribuidos de manera uniforme en los distintos dias. En cuanto a la distribucion horaria, el rango de horarios con menor cantidad de eventos es de 4am a 9am. El de mayor cantidad de eventos es desde las 10pm hasta las 2am del siguiente dia.

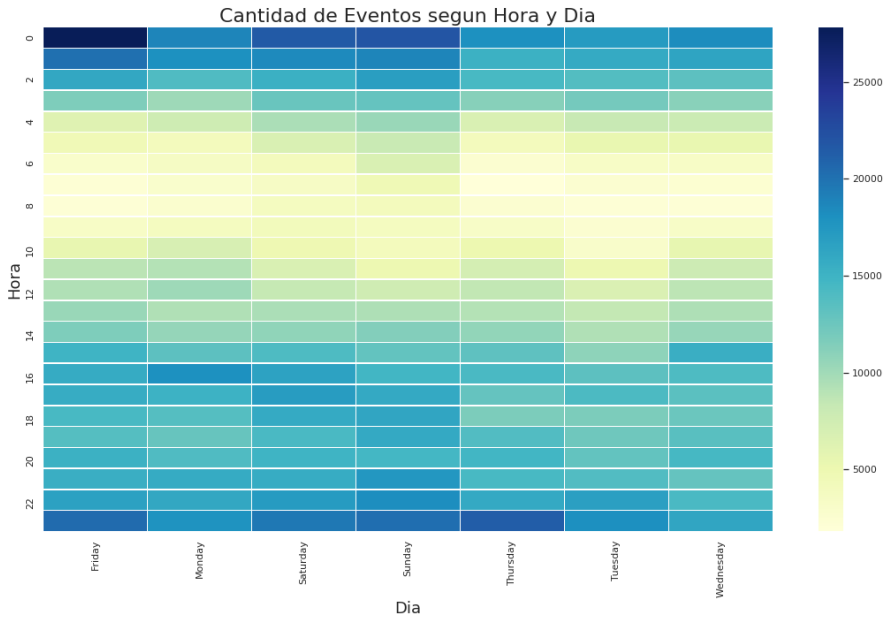

Figura 8: Cantidad de eventos en función de la hora y el día de la semana.

A partir de los datos que provee *Jampp* podemos obtener el detalle de las aplicaciones que se incluyen en este análisis. Por lo tanto, la Figura 9 muestra las aplicaciones con mayor cantidad de eventos registrados en el set de datos. De esta manera, *Jampp* podría saber cuales son las aplicaciones mas activas (que podrían estar de moda) e invertir en publicitarlas.

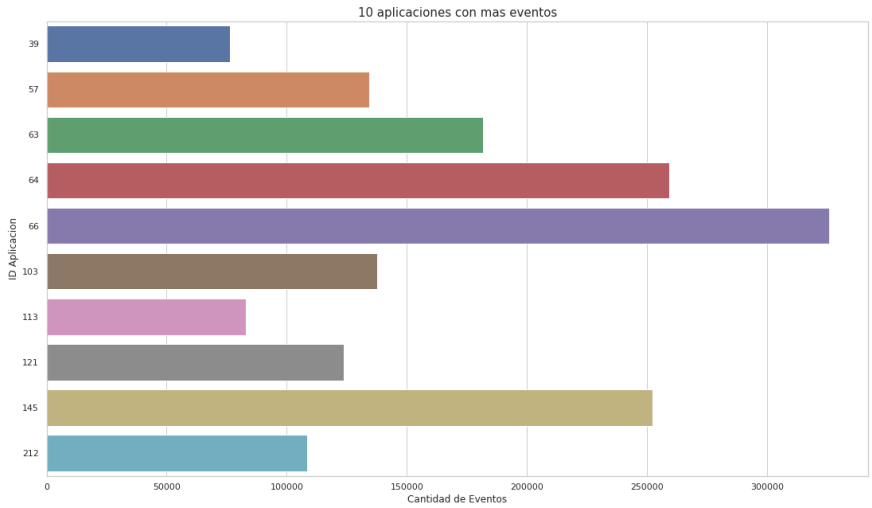

Figura 9: Las 10 aplicaciones con mayor cantidad de eventos a lo largo de todo el período de datos disponibles.

A través de su ID, podemos obtener el detalle de las personas que se incluyen en este análisis. Por lo tanto, la Figura 10 muestra las personas con mayor cantidad de clicks en el set de datos. De esta manera, *Jampp* podría saber cuales son las personas mas activas en la aplicación y posiblemente sus mejores clientes.

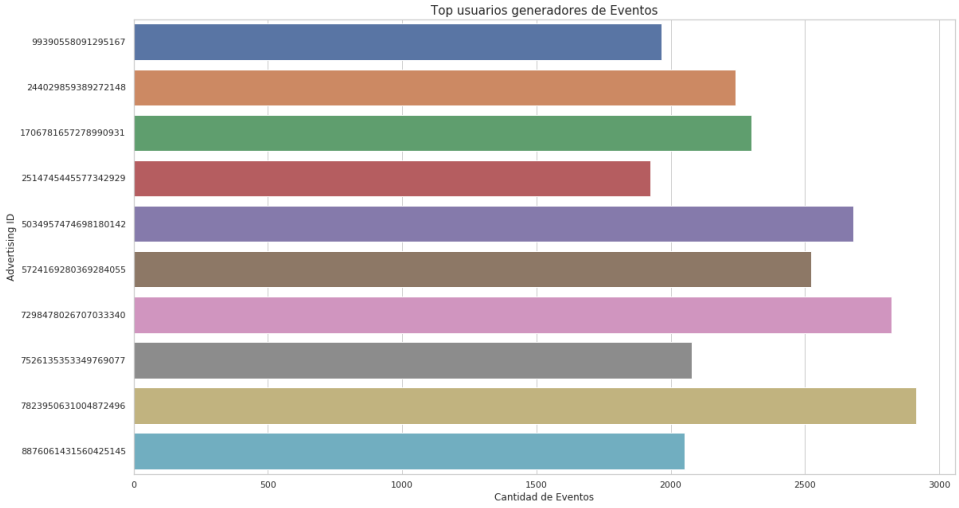

Figura 10: Las 10 personas que generaron mas cantidad de eventos a lo largo de todo el período de datos disponibles.

En el set de datos esta presente el atributo 'attributed' que es True cuando la aplicaciones

partners le atribuyen el evento a *Jampp*. A partir de la Figura 11. La aplicacion 63 es la que mas eventos le atribuye a Jampp, tambien era la 4ta que mas eventos habia generado. En el 2do puesto aparece la aplicacion 16 que no aparece entre las 10 apps que mas eventos generan.

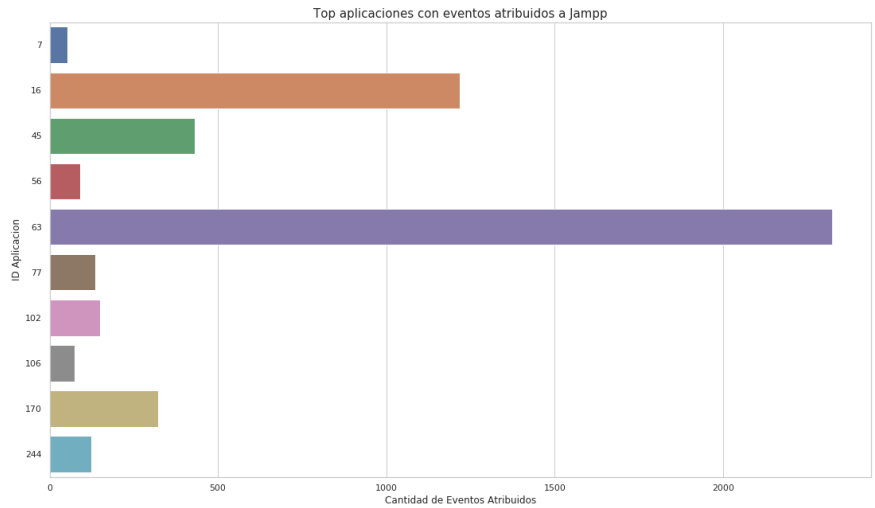

Figura 11: Las 10 aplicaciones con mayor cantidad de eventos atribuidos a Jampp.

Ahora queremos analizar si existe alguna relación entre el tipo de conexion y la cantidad de eventos registrados (especialmente los atribuidos a *Jampp*) según el día de la semana.

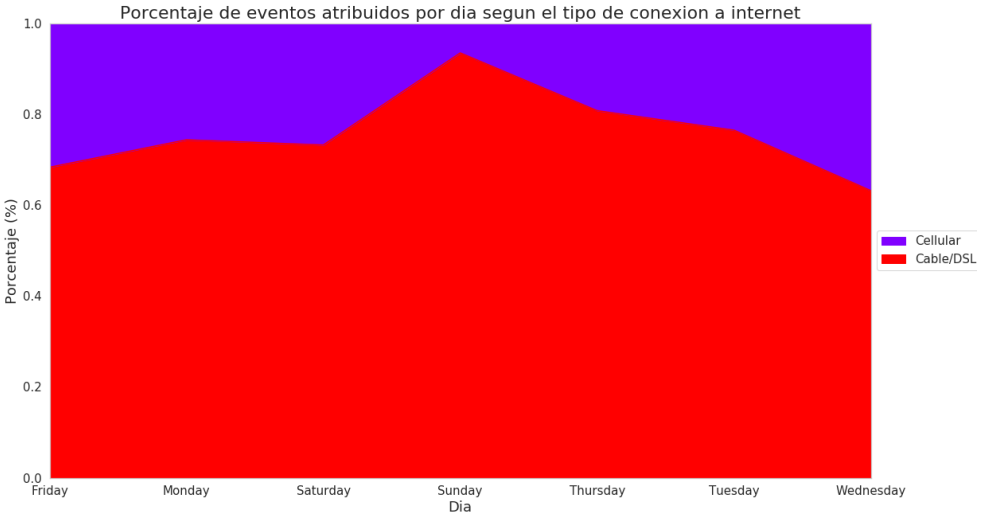

Figura 12: Porcentaje de eventos atribuidos a Jampp por dia segun el tipo de conexion a internet.

2.3. Instalaciones

Como un primer análisis se da un 'vistazo general' al conjunto de datos. Éste contiene 3412 registros y 18 atributos. Los datos no están ordenados cronológicamente (es decir según columna 'created'), pero se obtienen datos desde 2019-03-05 00:00:38 hasta 2019-03-13 23:54:00.

La frecuencia de los installs registrados está detallado en la Figura 13, donde los primeros 2 dias se realizaron menos de 350 installs cada dia y luego se ve una suba hasta casi los 400 installs el tercer dia para volver a bajar con un pico de installs el Martes 12 de marzo.

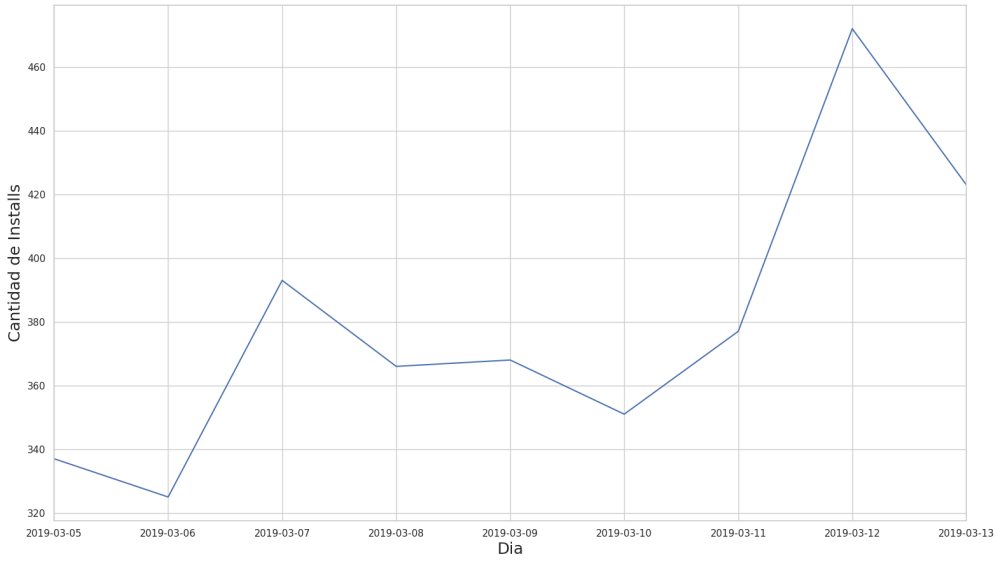

Figura 13: Cantidad de installs según el dia del año.

A través de su ID, podemos obtener el detalle de las personas que se incluyen en este análisis. Por lo tanto, la Figura 14 muestra las personas con mayor cantidad de installs en el set de datos. De esta manera, *Jampp* podría saber cuales son las personas con mayor conversión y posiblemente sus mejores clientes.

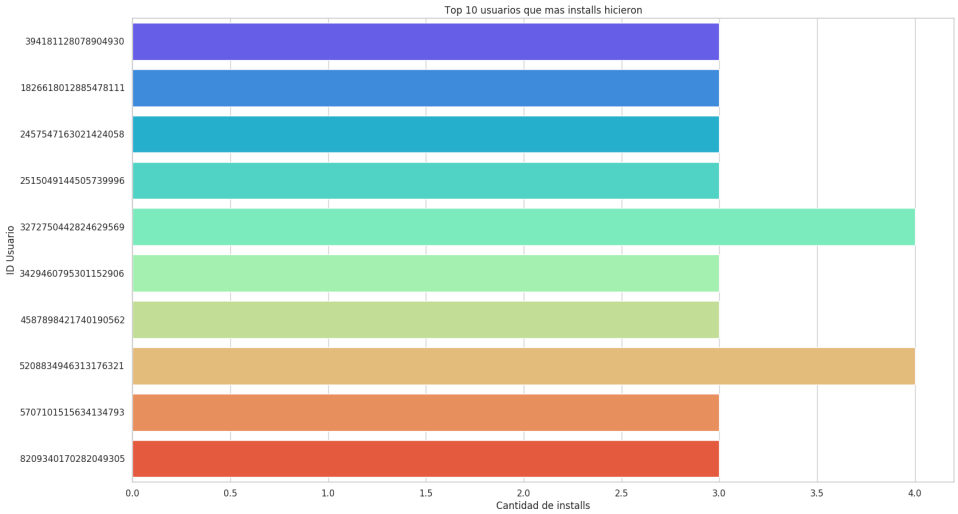

Figura 14: Las 10 personas con mayor cantidad de installs a lo largo de todo el período de datos disponibles.

Sin embargo es importante notar que a partir del atributo 'attributed', que si es True determina si el install fue atribuido a Jampp, podemos ver que 3412 de las 3412 instalaciones no se las atribuyeron a ellos. Esto nos puede estar diciendo que los datos fueron filtrados, o que las herramientas que se usan para determinar si deberian ser atribuidas, no son confiables. Aun asi, puede resultar bueno tener en cuenta a los 10 usuarios con mas instalaciones para armar un perfil de instaladores.

Ademas de ver la cantidad de instalaciones por dia, resulta interesante ver los horarios en los cuales se realizan mas instalaciones en una semana. La figura 15 muestra que hay una uniformidad en toda la semana, con unas pocas fluctuaciones segun el dia. Se ve que la franja horaria de menor trafico de installs es entre las 4 hs y las 10 hs con algunos dias extendiendose un pooco mas, la franja de mayor instalaciones es de 18 hs a 22 hs.

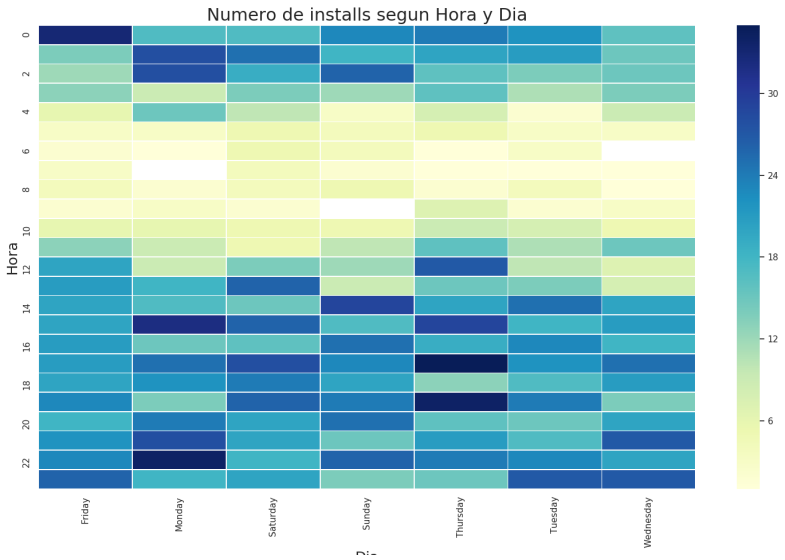

Figura 15: Installs segun la hora y el dia de la semana.

Tambien puede ser interesante saber cuales aplicaciones son las que mas installs tienen, para poder concentrarse en invertir en publicitar estas aplicaciones. En la figura 16 se puede ver que las aplicaciones 7 y 9 tienen muchos mas installs que el resto de las aplicaciones, con las aplicaciones 8, 10 y 16 en un segundo grupo, concentrar los recursos en estas aplicaciones puede ser util. Estaria bueno poder realizar un seguimiento de los installs por aplicacion a lo largo de un periodo mayor, ya que se podrian ver tendencias y aplicaciones que muestren indicios de despegarse de una franja de installs para caer en uno de los grupos que mencionamos antes, para aprovechar esa popularidad que puedan tener.

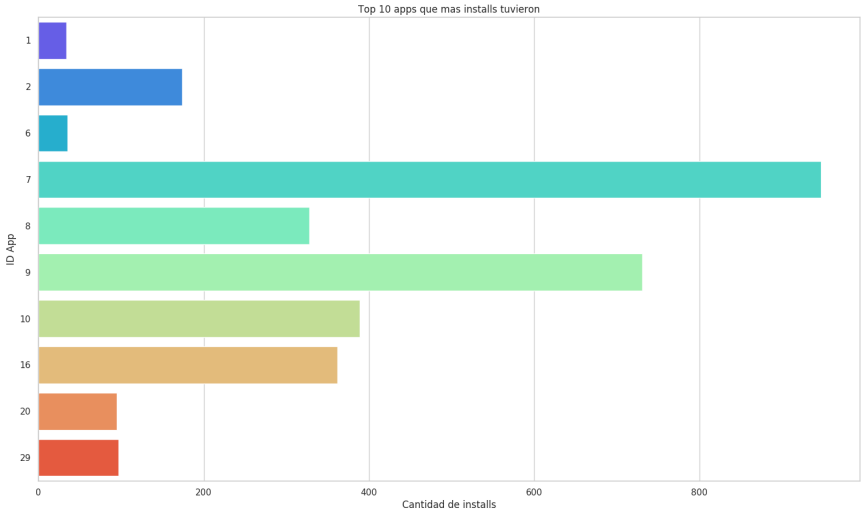

Figura 16: Top 10 apps con mas installs.

Por ultimo un analisis que se puede realizar es ver la cantidad de installs que se generan a partir de la marca de un telefono. La figura 17 muestra que hay una brand a la que se le atribuyen muchos mas installs que al resto, referenciandolo con el 'user_agent' podemos ver que se trata de un aparato que corre sistema operativo android. Esto puede ayudar a armar el perfil de users instaladores.

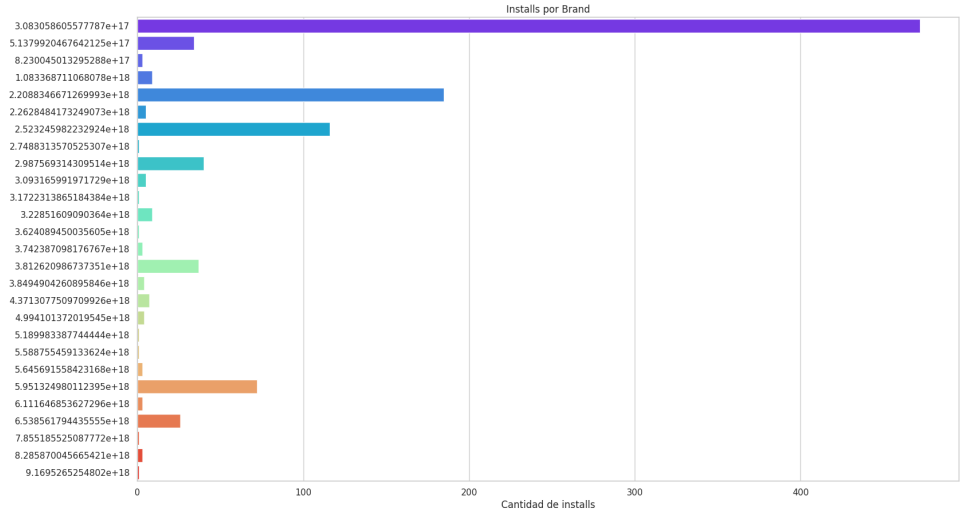

Figura 17: Installs por brand.

Es interesante notar que de los usuarios que mas installs realizaron solo dos utilizan la brand que mas installs realizo, uno de ellos estando entre los dos que mas installs realizaron, con 4.

2.4. Top usuarios

En esta sección se hace un análisis de los usuarios que mas interacciones tuvieron con la plataforma de JAMPP.

La idea de la sección es encontrar relaciones entre las cantidades de subastas, instalaciones, eventos y clicks que generó cada usuario (o device) con los diferentes adds que se le mostraron.

A su vez, se puso foco especial en los eventos e instalaciones que se le informa a JAMPP como propios (Atribuídos).

Se decidió utilizar tablas ya que se consideró que realizar gráficos con números muy dispares solo agregaría confusión a la visualización.

El siguiente gráfico muestra el top 10 de usuarios que generaron la mayor cantidad de subastas con la cantidad de instalaciones totales, instalaciones atribuidas, eventos totales, eventos atribuídos y clicks totales para cada usuario.

	ref_hash	auctionsCount	attributedInstall	installsCount	attributedEvent	eventsCount	clicksCount
0	633139769114048761	27762	0.0	0.0	0.0	0.0	10.0
1	7202276635029175071	23055	0.0	0.0	0.0	0.0	10.0
2	7298861376107043945	18188	0.0	0.0	0.0	0.0	12.0
3	6302840482782120473	16400	0.0	0.0	0.0	0.0	0.0
4	5376802567578262905	16367	0.0	0.0	0.0	0.0	9.0
5	8963711959081981780	14362	0.0	0.0	0.0	8.0	0.0
6	795159065504552200	12275	0.0	0.0	0.0	0.0	1.0
7	6841615430383729819	12077	0.0	0.0	0.0	0.0	32.0
8	5384039226444052914	11632	0.0	0.0	0.0	0.0	4.0
9	3198179064438296471	11565	0.0	0.0	0.0	3.0	0.0

Figura 18: Top 10 de usuarios que generaron la mayor cantidad de subastas.

Como se puede ver en la tabla, los datos generan una matriz muy dispersa ya que la mayoría de los valores son 0.

La conclusión que se puede llegar en este caso es que cada usuario (o device) genera muchas subastas de las cuales JAMPP participa en pocas, lo que se ve reflejado en la cantidad de 0 en los valores que indican la interacción del usuario con los ads.

El siguiente gráfico muestra el top 10 de usuarios que mas instalaciones generaron con con la cantidad de subastas en la que participaron, instalaciones atribuidas, eventos totales, eventos atribuídos y clicks totales para cada usuario.

	ref_hash	attributedInstall	installsCount	auctionsCount	attributedEvent	eventsCount	clicksCount
0	3272750442824629569	0	4	0.0	0	951	0.0
1	5208834946313176321	0	4	0.0	0	41	0.0
2	808602801225309575	0	3	0.0	0	19	0.0
3	1951826604052927528	0	3	63.0	0	3	0.0
4	5376622639905210608	0	3	54.0	0	15	0.0
5	2457547163021424058	0	3	0.0	0	64	0.0
6	4603142710878547974	0	3	0.0	0	9	0.0
7	8209340170282049305	0	3	0.0	0	49	0.0
8	1999001885916451618	0	3	0.0	0	3	0.0
9	2515049144505739996	0	3	236.0	0	46	0.0

Figura 19: Top 10 de usuarios que generaron la mayor cantidad de instalaciones.

En este caso, la cantidad de instalaciones producidas por usuario es realmente muy baja. El top de instalaciones para un mismo device es solamente de 4 y podemos ver que ninguna de estas instalaciones es atribuída a JAMPP.

Se puede notar que para la mayoría de los usuarios tops, las subastas en la que participaron son nulas, por lo que podemos inferir que las instalaciones se realizaron de una manera orgánica, es decir, el usuario entró al store a realizar la descarga de la misma y son los advertiser (clientes) los que nos indican que el usuario realizó la descargar de la aplicación que JAMPP esta promocionando.

Otra posibilidad es que los datos que los Ad Exchange (Sources) nos aportan sean realmente de mala calidad y es por ellos que vemos que los usuarios no hayan generado subastas.

En el siguiente gráfico podemos ver los 10 usuarios que más eventos generaron así como también la cantidad de subastas en la que participaron, instalaciones totales, instalaciones atribuídas, eventos atribuídos y clicks realizados.

	ref_hash	attributedEvent	eventsCount	attributedInstall	installsCount	auctionsCount	clicksCount
0	7823950631004872496	0	2913	0.0	0.0	0.0	0.0
1	7298478026707033340	0	2822	0.0	0.0	2.0	0.0
2	5034957474698180142	0	2681	0.0	0.0	43.0	0.0
3	5724169280369284055	0	2524	0.0	0.0	0.0	0.0
4	1706781657278990931	0	2300	0.0	0.0	3.0	0.0
5	244029859389272148	0	2240	0.0	0.0	70.0	0.0
6	7526135353349769077	0	2079	0.0	0.0	0.0	0.0
7	8876061431560425145	0	2053	0.0	0.0	0.0	0.0
8	99390558091295167	0	1966	0.0	0.0	6.0	0.0
9	2514745445577342929	0	1925	0.0	0.0	0.0	0.0

Figura 20: Top 10 de usuarios que generaron la mayor cantidad de eventos.

Al igual que en la gráfica de los top 10 usuarios con mas instalaciones, los top 10 usuarios con mas eventos generan una matriz dispersa en donde la cantidad de eventos atribuídos a JAMPP es nula en su totalidad.

En el siguiente gráfico se muestran el top 10 de usuarios que mas clikcs generaron.

	ref_hash	clicksCount	attributedInstall	installsCount	auctionsCount	attributedEvent	eventsCount
0	4063522808706112974	41	0.0	0.0	5596	0.0	0.0
1	6841615430383729819	32	0.0	0.0	12077	0.0	0.0
2	1244272413279716745	26	0.0	0.0	3297	0.0	0.0
3	1486699123421703687	21	0.0	0.0	6175	0.0	1.0
4	871778630200697460	21	0.0	0.0	1687	0.0	4.0
5	5087801386146573665	20	0.0	0.0	129	0.0	0.0
6	925803901826809324	17	0.0	0.0	8127	0.0	0.0
7	5141358577301248038	16	0.0	0.0	824	0.0	0.0
8	5410279713092214908	16	0.0	0.0	46	0.0	0.0
9	6972928558719957056	16	0.0	0.0	914	0.0	0.0

Figura 21: Top 10 de usuarios que generaron la mayor cantidad de clicks.

Salvo en la columna de subastas, las demás columnas tienen la mayoría de los valores en 0.

A pesar que la columna de subastas tiene todos valores no nulos, no se puede encontrar relación alguna entre la cantidad de clicks y las subastas que generó dicho device. Por lo que la única conclusión que podemos sacar es que para que haya algún click, el device debe haber generado alguna subasta.

Resumiendo: no existe click sin que el usuario haya generado alguna subasta.

2.5. Top aplicaciones

Al igual que en la sección donde se analizan los usuarios que mas subastas, eventos, clicks e instalaciones generaron, a continuación se realiza un análisis similar pero para las aplicaciones que mas eventos e instalaciones realizaron.

En la siguiente tabla podemos visualizar las aplicaciones que más eventos generaron.

	application_id	attributedInstall	installCount	attributedEvent	eventsCount
60	66	0.0	0.0	0.0	325696.0
58	64	0.0	0.0	0.0	259084.0
133	145	0.0	0.0	1.0	252431.0
57	63	0.0	0.0	2323.0	181555.0
94	103	0.0	0.0	0.0	137513.0
51	57	0.0	0.0	0.0	134498.0
110	121	0.0	0.0	0.0	123711.0
195	212	0.0	0.0	0.0	108489.0
103	113	0.0	0.0	0.0	82962.0
35	39	0.0	0.0	0.0	76448.0

Figura 22: Top 10 de aplicaciones que generaron la mayor cantidad de eventos.

Como en las tablas anteriores podemos observar que la cantidad de eventos es grande aunque los eventos atribuidos a JAMPP y las instalaciones son practicamente nulas. Solo una aplicación atribuye eventos a JAMPP y ninguna de las aplicaciones en cuestión logro obtener instalaciones mediante las subastas.

En la siguiente tabla visualizamos las aplicaciones que más eventos le atribuyen a JAMPP.

	application_id	attributedInstall	installCount	attributedEvent	eventsCount
57	63	0.0	0.0	2323.0	181555.0
15	16	0.0	362.0	1219.0	24365.0
41	45	0.0	0.0	431.0	14321.0
156	170	0.0	0.0	323.0	55820.0
93	102	0.0	0.0	150.0	32579.0
70	77	0.0	0.0	135.0	7810.0
222	244	0.0	0.0	125.0	1717.0
50	56	0.0	0.0	91.0	4587.0
97	106	0.0	0.0	75.0	1621.0
7	7	0.0	947.0	53.0	48005.0

Figura 23: Top 10 de aplicaciones que atribuyeron la mayor cantidad de eventos a JAMPP.

Los eventos atribuídos a JAMPP son realmente pocos en comparación a la cantidad total de eventos que genera una aplicación y, a pesar de tener eventos atribuídos, las instalaciones atribuídas a JAMPP son nulas para el top 10 de aplicaciones.

En la siguiente tabla podemos ver las aplicaciones que mas instalaciones tuvieron.

	application_id	attributedInstall	installCount	attributedEvent	eventsCount
7	7	0.0	947.0	53.0	48005.0
9	9	0.0	731.0	0.0	8198.0
10	10	0.0	389.0	0.0	58311.0
15	16	0.0	362.0	1219.0	24365.0
8	8	0.0	328.0	0.0	38972.0
2	2	0.0	174.0	0.0	5931.0
25	29	0.0	97.0	3.0	732.0
19	20	0.0	95.0	1.0	688.0
6	6	0.0	35.0	0.0	27.0
1	1	0.0	34.0	0.0	0.0

Figura 24: Top 10 de aplicaciones tuvieron la mayor cantidad de instalaciones.

Como podemos ver en la tabla, a pesar de que algunas aplicaciones atribuyen eventos a JAMPP, no existe ninguna instalación atribuída a JAMPP en este top 10.

En la siguiente tabla podemos ver un top 10 de aplicaciones que más instalaciones le atribuyen a JAMPP.

	application_id	attributedInstall	installCount	attributedEvent	eventsCount
0	0	0.0	18.0	0.0	56.0
183	199	0.0	0.0	8.0	195.0
189	206	0.0	0.0	0.0	2.0
188	205	0.0	0.0	0.0	9.0
187	204	0.0	0.0	0.0	1159.0
186	203	0.0	0.0	0.0	15.0
185	202	0.0	0.0	0.0	3644.0
184	200	0.0	0.0	0.0	41.0
182	198	0.0	0.0	1.0	9218.0
174	190	0.0	0.0	0.0	2005.0

Figura 25: Top 10 de aplicaciones que atribuyeron la mayor cantidad de instalaciones a JAMPP.

Como podemos ver, no existe aplicación alguna que le atribuya alguna instalación a JAMPP.

Como bien se mencionó mas arriba, si hacemos un describe de la columna .ªttributed"del CSV de instalaciones obtenemos que todos los registros se encuentran en "False". Por lo que no existe instalación alguna atribuída a JAMPP.

```
installs['attributed'].describe()
```

```
count       3412
unique         1
top        False
freq        3412
Name: attributed, dtype: object
```

Figura 26: Describe de la columna 'attributed' del archivo de instalaciones.

3. Conclusión

El análisis exploratorio sobre el conjunto de datos de la plataforma *Jampp* fue aplicado con diferentes enfoques: análisis temporal, geográfico, marca de dispositivos, ad exchanges, telefono del usuario, eventos atribuidos a *Jampp*, etc. Dentro de ellos se obtuvo resultados

tales como cuales son los usuarios y las aplicaciones que mas utilizan los servicios de *Jampp*, la cantidad de instalaciones segun la hora y el dia de la semana, la cantidad de eventos a lo largo del tiempo, los ad exchange mas utilizados segun la hora del dia, etc.

En base a los datos analizados en el presente informe llegamos a notar que JAMPP no cuenta con un método efectivo para medir la atribución de un install generado por su plataforma ya que los installs atribuídos a JAMPP son nulos.

JAMPP no participa en todas las subastas que se presentan, ya que los eventos, instalaciones y clicks atribuídos a JAMPP son realmente muy escasos o, como bien se indicó en el parrafo anterior, JAMPP no cuenta con un método eficaz para realizar estas mediciones.

La única relación encontrada entra las diferentes columnas de los set de datos es que para que existan clicks, el usuario debe haber generado alguna subasta.

4. Más información

Nuestro análisis fue realizado en Python Pandas. Utilizamos un repositorio público de github para juntar análisis y filtrar los que creaímos eran adecuados para incluir en este informe. El link del repositorio es:

https://github.com/GastonMontes/75.06-Datos-Grupo22-TP1

Dentro del repositorio se encuentra los notebooks. El set de datos no fue incluído en el repositorio por ser muy pesado, se puede encontrar dicho archivo en el siguiente link:

https://drive.google.com/drive/folders/1-7ACCWhS3sWVh0vU273YwbS74WmzEhLa?usp=sharing