Universidad de Buenos Aires
Facultad de Ingeniería

75.06 - Organización de Datos

Trabajo Práctico 1

Primer cuatrimestre 2019

Grupo 4

René Mauricio Villegas Paredes

Federico Flores

Nadia Kazlauskas

Índice

1. Introducción general

El objetivo de este trabajo es realizar un análisis exploratorio sobre los datos facilitados por la empresa Jampp. Esta es una empresa de marketing en la plataforma mobile, que utiliza anuncios programáticos para obtener nuevos usuarios, o promover cierto comportamiento de los usuarios dentro de las aplicaciones.

Para esto participa en subastas en tiempo real, donde se calcula la probabilidad de que un usuario convierta, es decir, que instale una aplicación o realice el comportamiento deseado.

En líneas generales, nuestro objetivo es buscar en los datos ciertos patrones de comportamiento de los usuarios, que de existir, podrían ayudarnos a predecir una conversión.

2. Procesamiento y datos utilizados

El análisis exploratorio se realizó sobre los datos provistos por la empresa Jampp. Se recibieron 4 datasets en formato csv, conteniendo los datos correspondientes a 9 días consecutivos:

- Subastas: contiene información de las subastas en las que Jampp participó, ya sea que haya resultado ganador o no.

- Clicks: información sobre los clicks que realizó el usuario sobre las impresiones.

- Eventos: datos sobre el comportamiento del usuario dentro de las aplicaciones clientes de Jampp.

- Instalaciones: contiene información sobre las aplicaciones que han sido instaladas, ya sea debido a Jampp o de manera implícita, donde el usuario instala una aplicación de manera independiente de Jampp.

En primer lugar se revisó la integridad de los datos, de manera de ver si era necesario realizar una limpieza de los datos. Se buscó la existencia de valores nulos o NaN, pero no se encontraron este tipo de valores dentro de las columnas de interés, por lo que no fue necesario sanitizar los datos.

El procesamiento de los datos se realizó utilizando Python 3 y la librería Pandas. Para lograr las visualizaciones utilizamos MatPlotLib y Seaborn.

Los notebooks con el código del análisis pueden encontrarse en el siguiente repositorio: https://github.com/nkazlauskas/7506-tp1

3. Análisis exploratorio

3.1. Subastas

En la exploración del dataset auctions brindado por la empresa Jampp encontramos:

- Subastas recibidas: 19.571.319
- Ventana temporal 10 días (5 al 14 de Marzo 2019).
- 100% Provenientes de un único país.
- Sistemas Operativos: 79.4 % Android ~ 20. 6% IOS

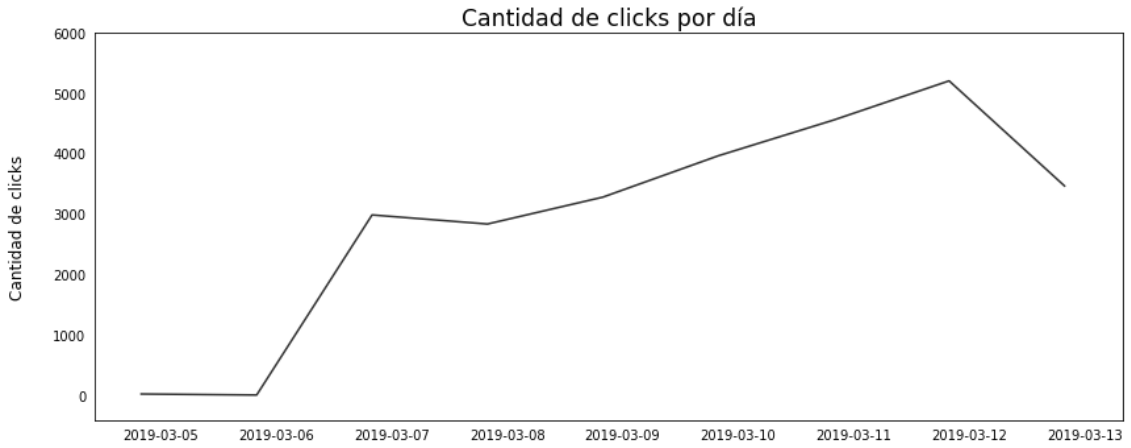
3.2. Clicks

3.2.1 Introducción

Los datos sobre clicks se refieren a los clicks realizados por los clientes potenciales luego de visualizar el advertise, esto puede derivar en un install o event dependiendo de la decisión que tome el potencial.
Los datos se presentan con una ventana aproximada de 10 días, en un mismo país. Se proporcionan datos adicionales, como la ubicación relativa del click en el dispositivo, ubicación geográfica (anonimizada), carrier, etc. En este primer apartado analizamos a fondo las ocurrencias y peculiaridades de los datos.

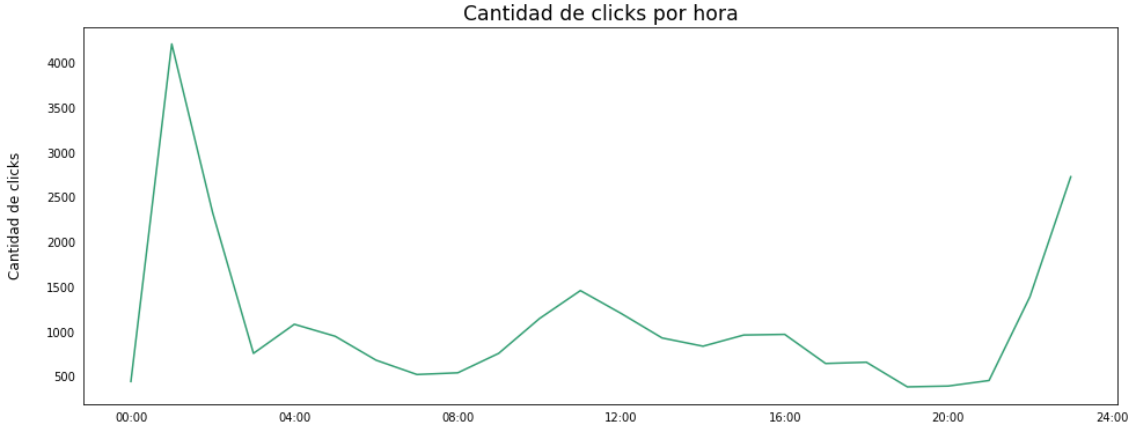
3.2.2 Distribución de los clicks temporalmente

La ventana de tiempo presentada tiene los siguientes valores:



Vemos la concentración de clicks entre el 07 y 13 de marzo, con una tendencia creciente a partir del 07 y el pico máximo el día 12. Esto puede indicar que se realizó previendo alguna fecha o evento específico(s) para realizar un advertising masivo.

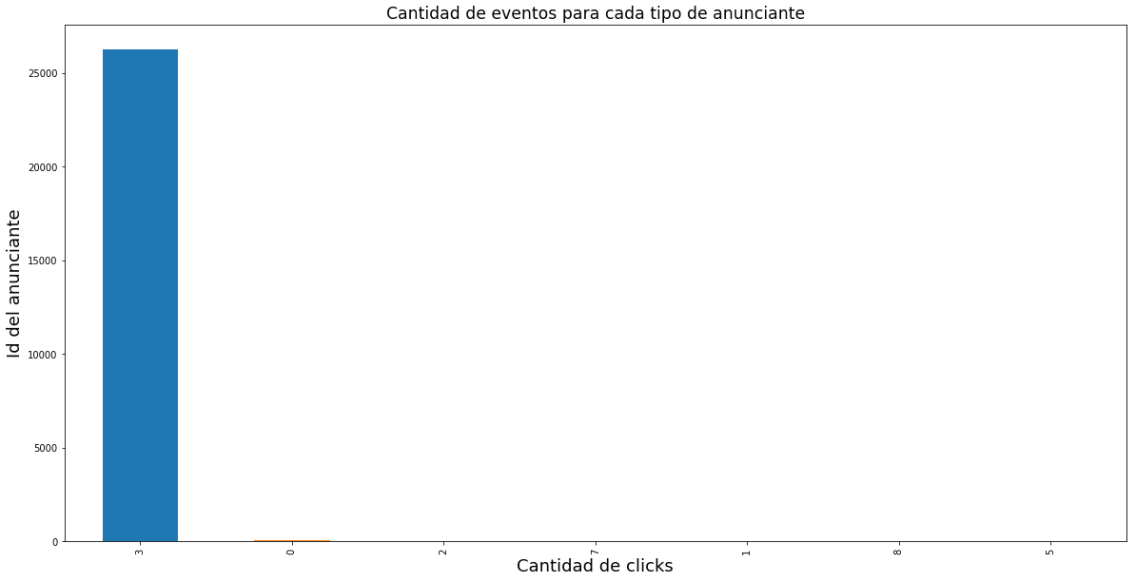
Intentamos observar un mayor detalle con las distribuciones por horas:



Se puede apreciar una tendencia oscilante entre las 04 y 20 hrs. Y creciente luego de las 20 hrs., con un pico máximo entre las 1 y 2 de la madrugada.

3.2.3 Advertisers

Los clientes o “advertisers”, son quienes por medio de Jampp intentan causar las impresiones a los potenciales clientes. Los datos asociados a cada uno por su ID o identificador interno dentro de Jampp nos muestran los siguientes datos:

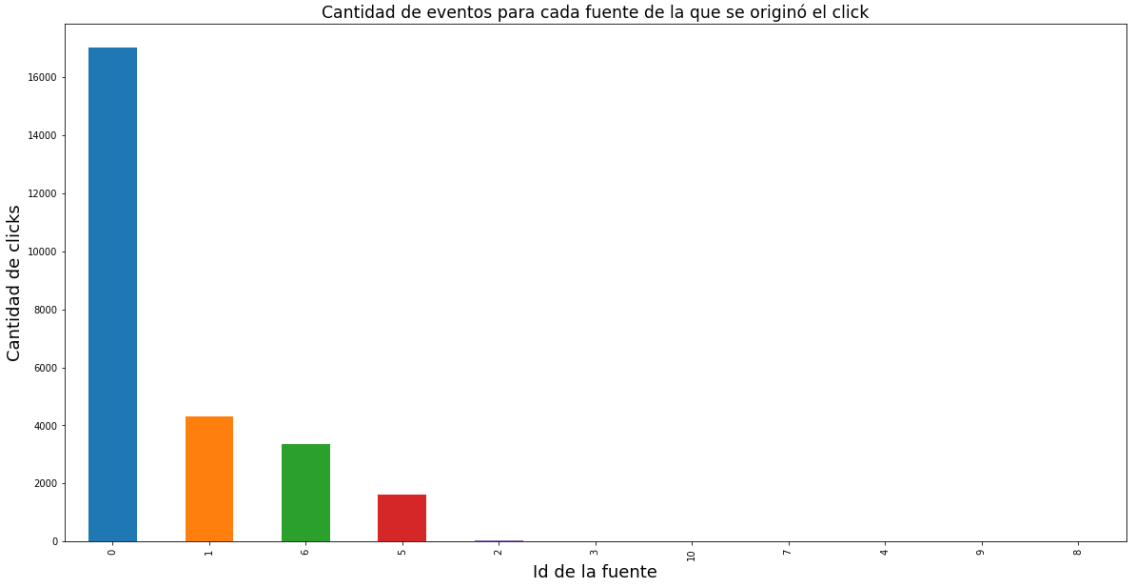


Siendo estos 7 Ids de clientes quienes componen el 99.67% de la muestra dada.

Sin embargo, esto simplemente muestra que el cliente identificado como 3, tiene alrededor de 95% de clicks dentro la muestra. Podemos ampliar el análisis y ver la fuente desde la que se originan los clicks.

3.2.4 Exchangers

JAMPP identifica internamente a las “fuentes” donde se originan los clicks, es decir quienes proporcionan los “ad spots” (espacios de advertising) disponibles a biddear. Esto generará un auction (subasta) a la que JAMPP está invitada a participar y pugnar por mostrar la impresión al potencial cliente. Los datos muestran el siguiente comportamiento:



Observamos que las fuentes con Ids 0, 1, 6, 5 y 2 son las que predominan.

3.2.5 Conclusiones

Los datos presentados, muestran las ocurrencias de clicks de acuerdo a distintos valores, muchos de ellos internos y otros hasheados, esto puede resultar en un sesgo para un análisis más profundo, por ejemplo si la finalidad es saber acerca de una ubicación geográfica, un carrier específico, etc.

Sin embargo, los datos proporcionados muestran tendencias interesantes como las ocurrencias en función del tiempo para la ventana de tiempo presentada y que puede ayudar a JAMPP a establecer horarios adecuados de advertising a potenciales segmentos.

3.3. Instalaciones

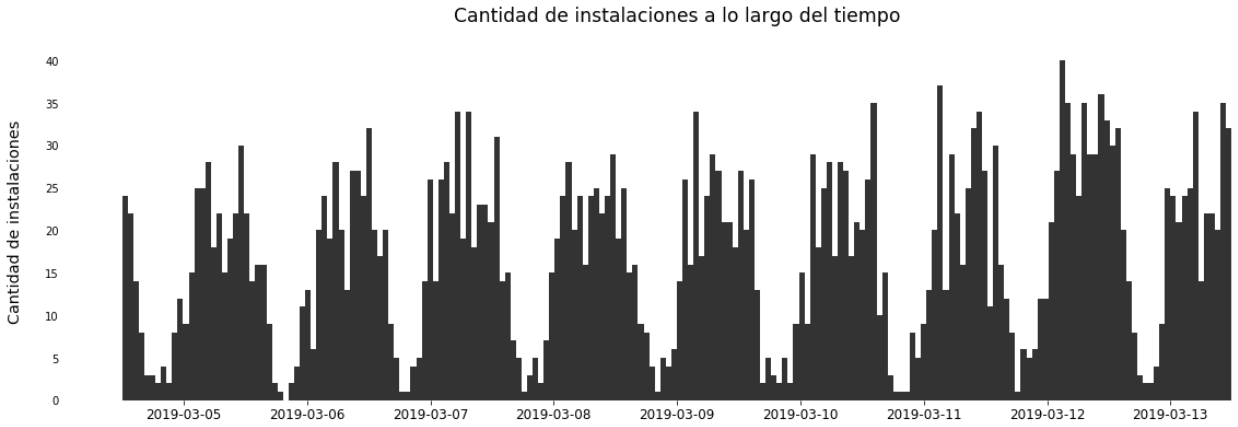
3.3.1. Introducción

Estos datos se corresponden con las instalaciones de las diferentes aplicaciones por parte de los usuarios (devices). Estas instalaciones pueden haber sido atribuidas o no a Jampp en base al atributo implicit.

A modo de comenzar con el análisis hacemos una indagación de la integridad de los datos principales que vamos a necesitar a modo de realizar una posterior depuración en caso de ser necesario. En un principio, los datos relacionados a la fecha de creación de las instalaciones (created), atribución a Jampp de la instalación (implicit), identificador de la aplicaciçon (application_id), e identificador del device (ref_hash).

3.3.2. Distribución de las instalaciones a lo largo del tiempo

En una primera instancia determinamos la extensión temporal del dataframe, el valor mínimo (05/03/2019 a las 00:00:38) y el valor máximo (13/03/2019 a las 23:54:00). Asimismo utilizamos un histograma para visualizar la distribución de eventos a lo largo de este marco temporal.



Se puede observar que no parecen haber variaciones muy considerables en la distribución de instalaciones, siendo que cada día posee un momento en cual las instalaciones bajan considerablemente para ir progresivamente en aumento hasta alcanzar un pico. Si bien, con algunas variaciones, la forma de cada día presenta regularidades.

Nos proponemos entonces profundizar en este aspecto y buscamos determinar en términos absolutos cuáles fueron los días con más instalaciones. Para esto utilizaremos un gráfico de líneas en donde compararemos el aumento o la disminución de instalaciones en base a cada valor.

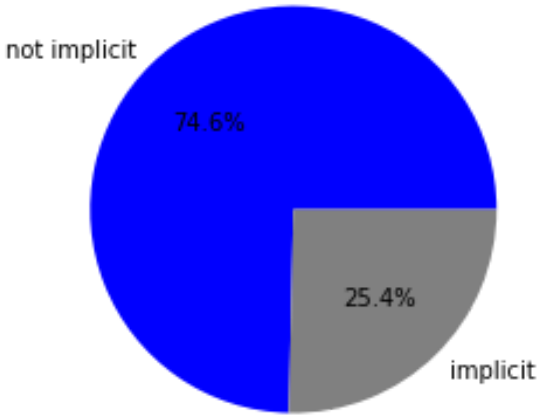
Vemos así que desde el 10 de marzo se establece un progresivo aumento de las instalaciones que culmina en un pico el día 12 de marzo para luego ir progresivamente disminuyendo.

Por otro lado, buscamos establecer las instalaciones por hora para ver en qué momentos del día hay mayor recurrencia de instalaciones y en qué momento menor. Para esto utilizamos nuevamente un gráfico de líneas.

Observamos entonces que el pico de instalaciones se da a las 17 hs con un promedio de 233 instalacione seguido por las 22 hs con 229 instalaciones. Asimismo existen varios picos con más de 200 instalaciones (15 hs, 23 hs, 19 hs y 14 hs). A partir de las 12 hs comienza a bajar el número de instalaciones para llegar a su punto más bajo a las 15 hs.

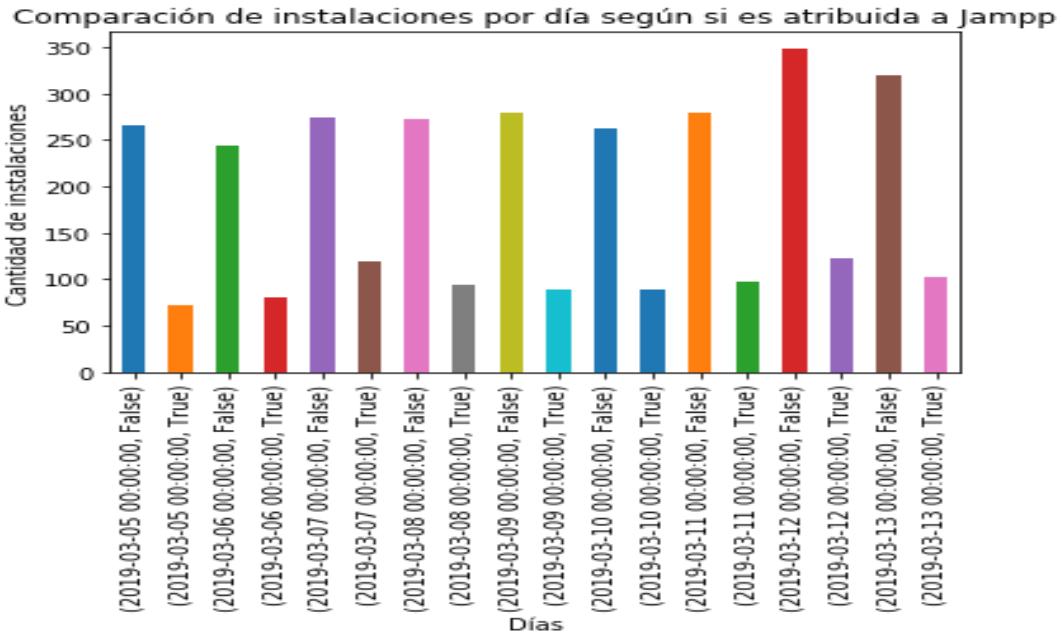
3.3.3. Aplicaciones atribuidas a Jampp.

Un elemento a tomar en cuenta es la cantidad de instalaciones que fueron atribuidas a Jamp. Para esto realizamos un gráfico de tortas en el que se establezcan los porcentajes para visualizarlo más fácilmente.



Podemos observar de esta manera que sólo el 25 % de las instalaciones son atribuidas directamente a Jampp. El esto de las mismas puede se atribuida a la competencia o a que el usuario realizó la instalación a partir de flujos no directamente relacionados con Jampp.

Profundizando en este aspecto nos interesa analizar cómo se distribuyen la cantidad de instalaciones a lo largo de cada día del rango temporal del dataframe en relación a si dichas instalaciones fueron atribuidas o no a Jampp.

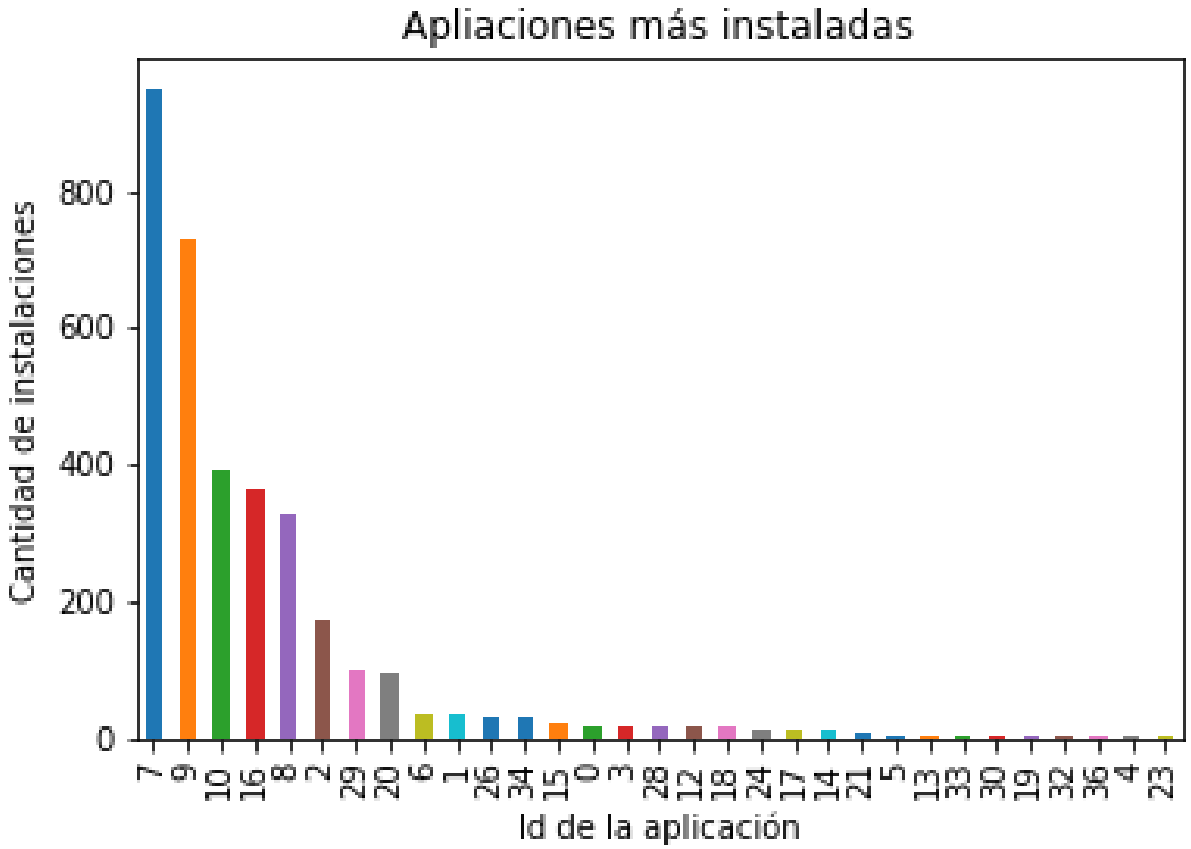


Se puede observar nuevamente que existe a simple vista una cierta homogeneidad en la forma en la que se desarrollan los datos, manteniéndose relativamente constante la diferencia entre las instalaciones atribuidas a Jampp y las que no.

3.3.4. Distribuciones de las instalaciones según las distintas aplicaciones.

Otro punto a analizar consiste en ver cuáles son las aplicaciones que interactúan en este dataset y ver cómo están distribuidas. Para esto analizamos que el dataset consta de 31 aplicaciones distintas que fueron instaladas.

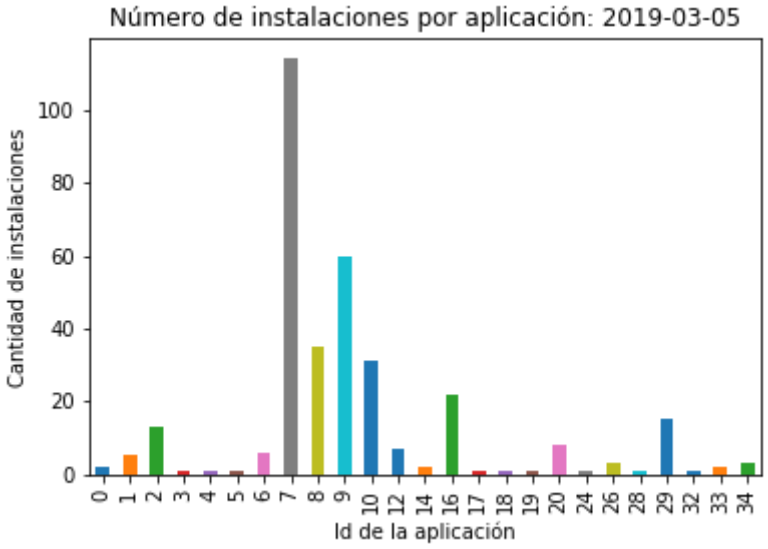
Consecuentemente queremos averiguar cuáles fueron las más instaladas y cómo fue su distribución. Para esto realizamos un gráfico de barras que nos permita comparar la cantidad total de instalaciones para cada aplicación.

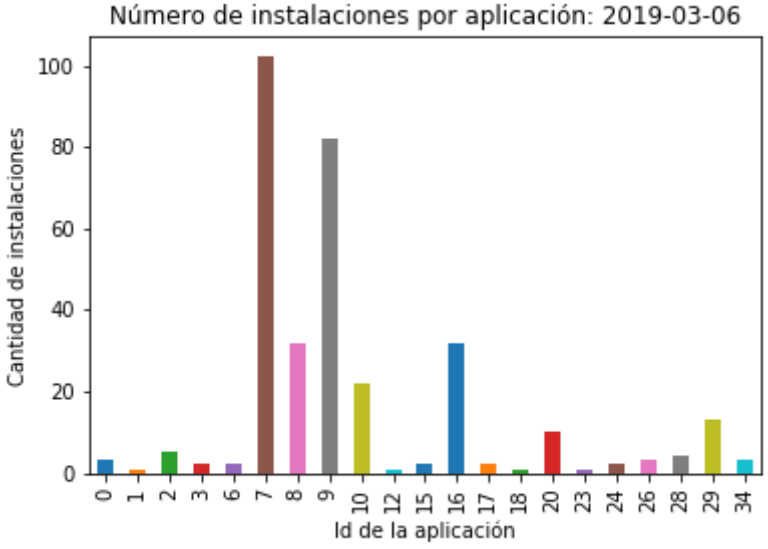


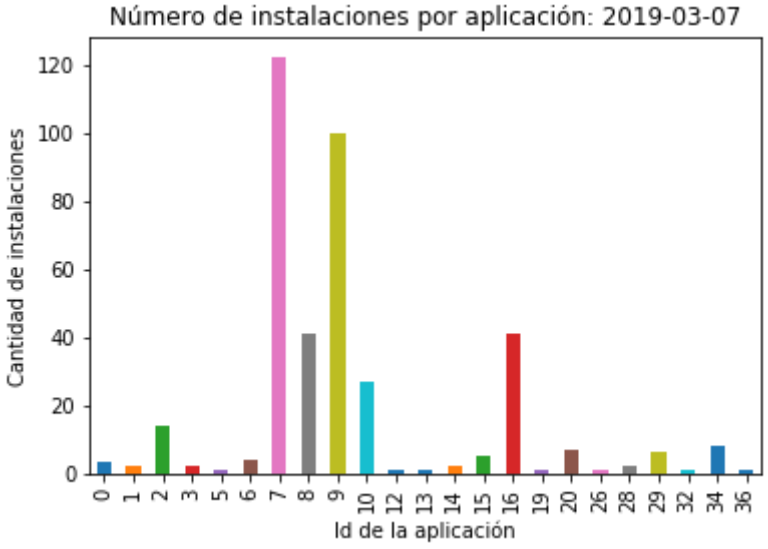
Podemos observar que la aplicación más instalada son: 7, 9, 10, 16 y 8; todas ellas con más de 200 instalaciones.
En este sentido, las apps con más instalaciones son las 7 y la 9 liderando donde la app 7 consta de 947 instalaciones, seguida de la 9 que tiene 731. Ya irremediablemente más abajo se encuentran la 10 con 389 instalaciones, la 16 con 362 y la 8 con 328.

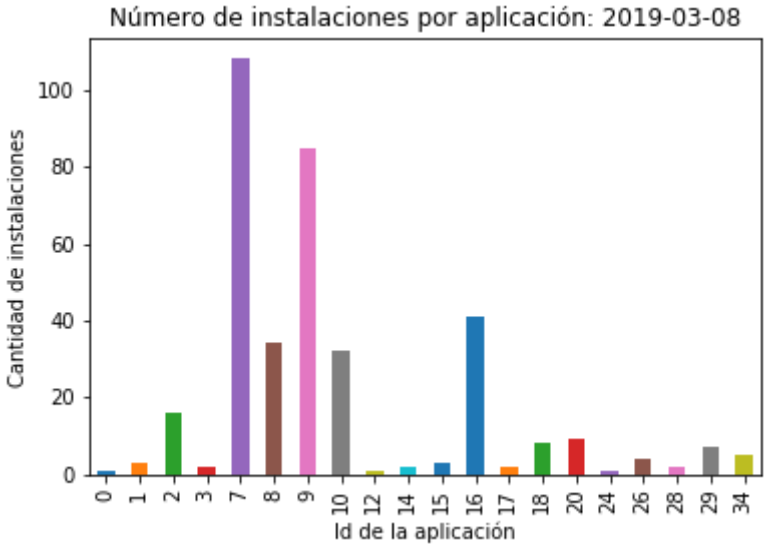
Asimismo, la suma de las instalaciones de estas 5 apps constituyen más del 80 % del total de las instalaciones de la muestra.

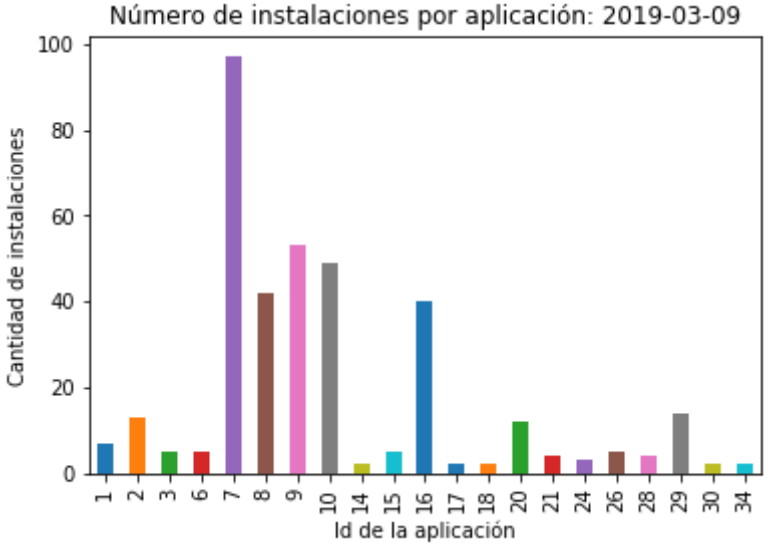
Quisiéramos ver ahora la distribución de la instalación de las distintas apps a lo largo del tiempo, centrándonos en cada día en particular. Para esto nuevamente realizaremos gráficos de barra para cada día en particular.

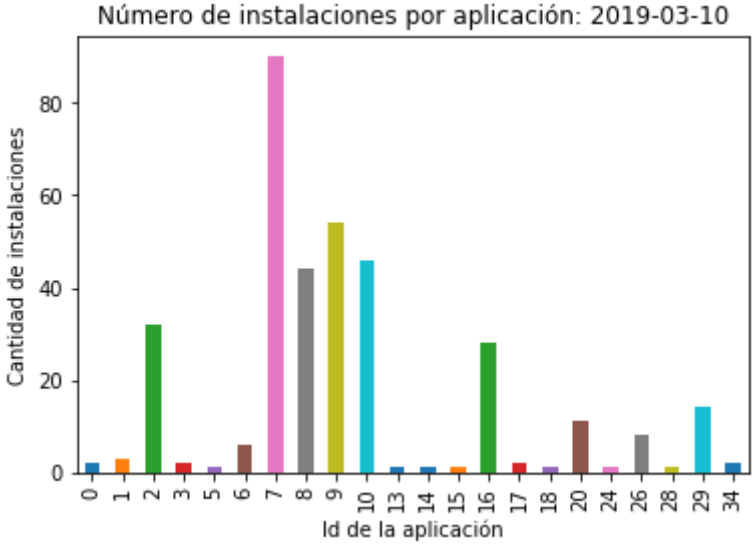

Número de instalaciones por aplicación: 2019-03-05
Cantidad de instalaciones
0
20
40
60
80
100
0 1 2 3 4 5 6 7 8 9 10 12 14 16 17 18 19 20 24 26 28 29 32 33 34
Id de la aplicación

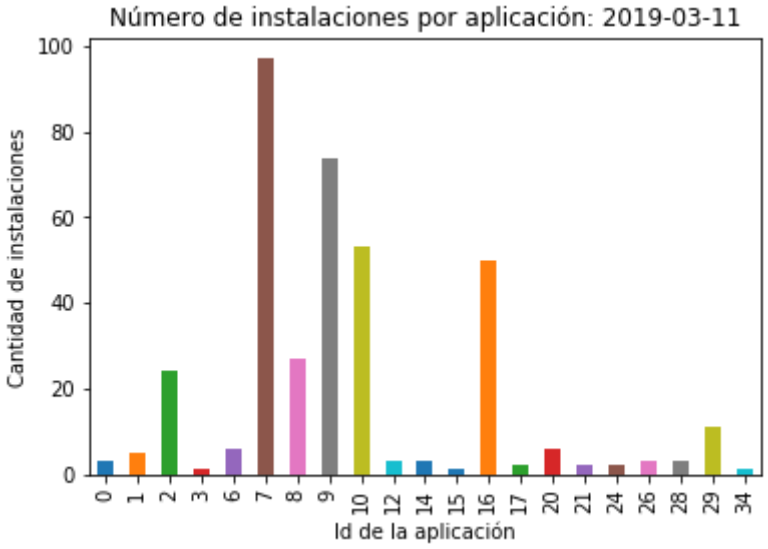

Número de instalaciones por aplicación: 2019-03-06
Cantidad de instalaciones
0
20
40
60
80
100
0 1 2 3 6 7 8 9 10 12 15 16 17 18 20 23 24 26 28 29 34
Id de la aplicación

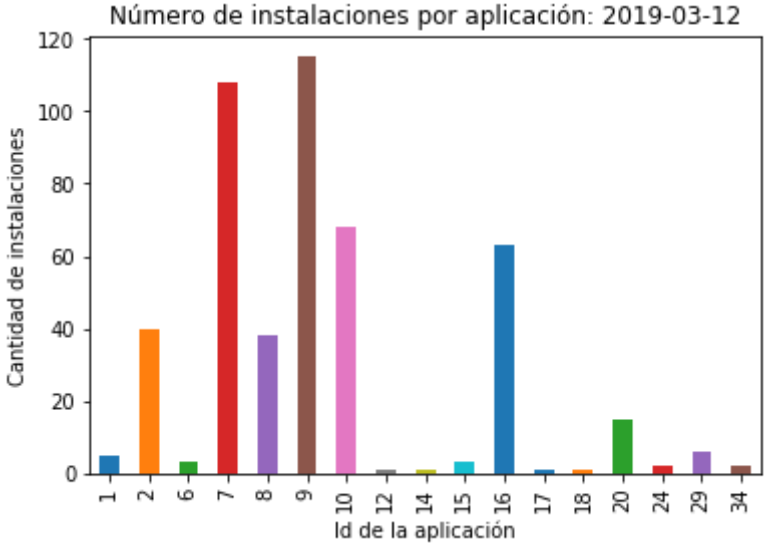

Número de instalaciones por aplicación: 2019-03-07
Cantidad de instalaciones
0
20
40
60
80
100
120
0 1 2 3 5 6 7 8 9 10 12 13 14 15 16 19 20 26 28 29 32 34 36
Id de la aplicación

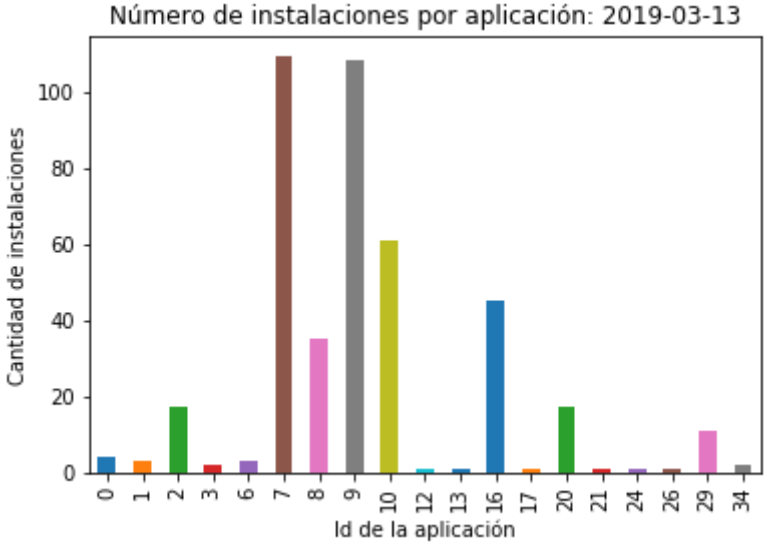

Número de instalaciones por aplicación: 2019-03-08
Cantidad de instalaciones
0
20
40
60
80
100
0 1 2 3 7 8 9 10 12 14 15 16 17 18 20 24 26 28 29 34
Id de la aplicación


Número de instalaciones por aplicación: 2019-03-09
Cantidad de instalaciones
0
20
40
60
80
100
1 2 3 6 7 8 9 10 14 15 16 17 18 20 21 24 26 28 29 30 34
Id de la aplicación


Número de instalaciones por aplicación: 2019-03-10
Cantidad de instalaciones
0
20
40
60
80
0 1 2 3 5 6 7 8 9 10 13 14 15 16 17 18 20 24 26 28 29 34
Id de la aplicación


Número de instalaciones por aplicación: 2019-03-11
Cantidad de instalaciones
0
20
40
60
80
100
0 1 2 3 6 7 8 9 10 12 14 15 16 17 20 21 24 26 28 29 34
Id de la aplicación


Número de instalaciones por aplicación: 2019-03-12
Cantidad de instalaciones
0
20
40
60
80
100
120
1 2 6 7 8 9 10 12 14 15 16 17 18 20 24 29 34
Id de la aplicación


Número de instalaciones por aplicación: 2019-03-13
Cantidad de instalaciones
0
20
40
60
80
100
0 1 2 3 6 7 8 9 10 12 13 16 17 20 21 24 26 29 34
Id de la aplicación

Si bien existe una fuerte regularidad dentro de la cantidad de aplicaciones más instaladas, siendo las 5 anteriormente mencionadas las que priman en todo momento, en los últimos 2 días se observa un fuerte incremento en la cantidad de instalaciones de la app 9, la cual en el día 12-03 supera a la 7 y al día siguiente obtiene casi la misma cantidad de instalaciones (107 instalaciones de la 7 contra 108 de la 9).

Asimismo, si bien a lo largo de todos los días se ve que las aplicaciones 7 y 9 ocupan el 1er y 2do puesto en tanto más instaladas. La tercera posición se encuentra disputada entre la 8, la 10 y la 16 con un margen bastante pequeño de diferencia entre estas 3.

3.3.5. Conclusiones

Al analizar el Dataset de Instalaciones podemos ver que la distribución en el tiempo resulta bastante homogénea a los largo de los distintos días así como de las horas dentro de estos. Presenta un pico de instalaciones durante la tarde, entre las 17 y las 23 hs, y teniendo su punto más bajo entre las 1 am y las 5 am.

Las aplicaciones más instaladas también se mantienen relativamente constantes a lo largo de estos días siendo que son 5 de ellas las que representan el 80 % del total de las instalaciones. Por su parte las aplicaciones 7 y 9 son las 2 que presentan mayor número de instalaciones a lo largo de todo el rango temporal.

Por su parte, las instalaciones atribuidas a Jampp constituyen sólo una pequeña parte del total de las mismas, un 25 % aproximadamente, y de estas su distribución a lo largo del tiempo es también homogénea a lo largo de los distintos días.

3.4 Eventos

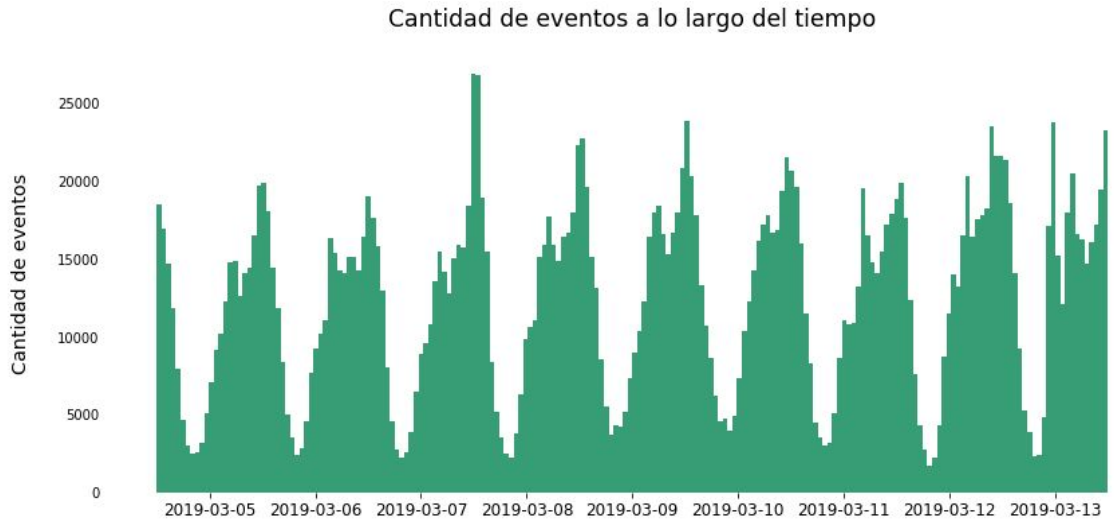
3.4.1. Introducción

Los datos orgánicos son generados por el comportamiento de los usuarios dentro de las aplicaciones clientes de Jampp. Cada una de estas acciones son denominadas eventos.
En primer lugar, al comenzar el análisis de este dataset, realizamos un chequeo sobre la integridad de los datos y sobre qué información teníamos disponible para realizar el análisis exploratorio.

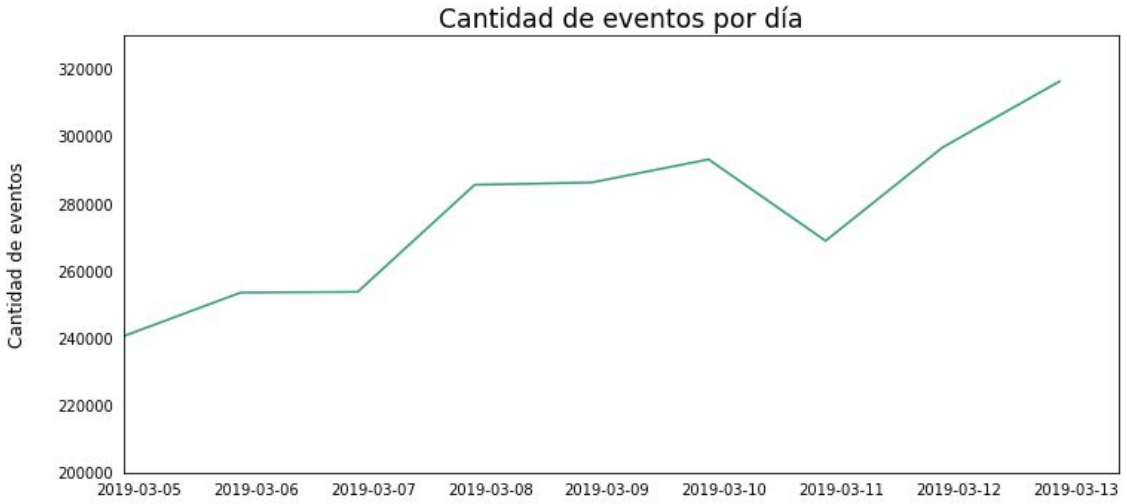
Principalmente contábamos con información sobre la fecha de creación del evento, el código del mismo y desde qué aplicación se generó; además de información sobre el dispositivo (modelo del dispositivo, tipo de conexión, sistema operativo, etc.).

3.4.2. Distribución de los eventos a lo largo del tiempo

El dataset cuenta con información de varios días, por lo que en primera instancia nos pareció conveniente realizar un histograma, de manera de poder observar el patrón de distribución de los eventos a lo largo del tiempo.



Vemos que la generación de eventos parece seguir un patrón cíclico, repitiéndose a lo largo de los días analizados. Parece haber un momento del día en el cual se registran la mayor cantidad de eventos, y cualitativamente el área bajo la curva de cada día se ve de similar tamaño.
Para poder observar esto en más detalle, graficamos la cantidad de eventos registrados en cada día, y la cantidad de eventos en función de la hora del día.



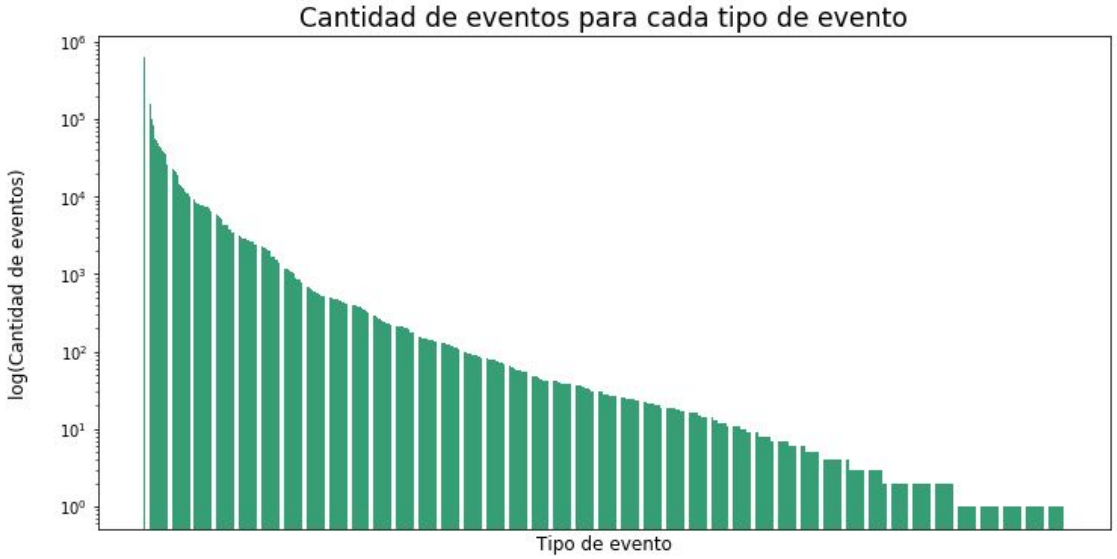
Vemos que la cantidad de eventos recibidos por día parece mantenerse dentro de ciertos valores, aunque necesitaríamos los datos de un período mayor de tiempo para lograr detectar alguna tendencia.

Confirmando lo que se observaba en el histograma, vemos que los eventos se distribuyen de manera heterogénea a lo largo del día: presentan un mínimo alrededor de las 7-8 hs y un máximo cercano a las 23-24 hs.

3.4.3. Tipos de eventos

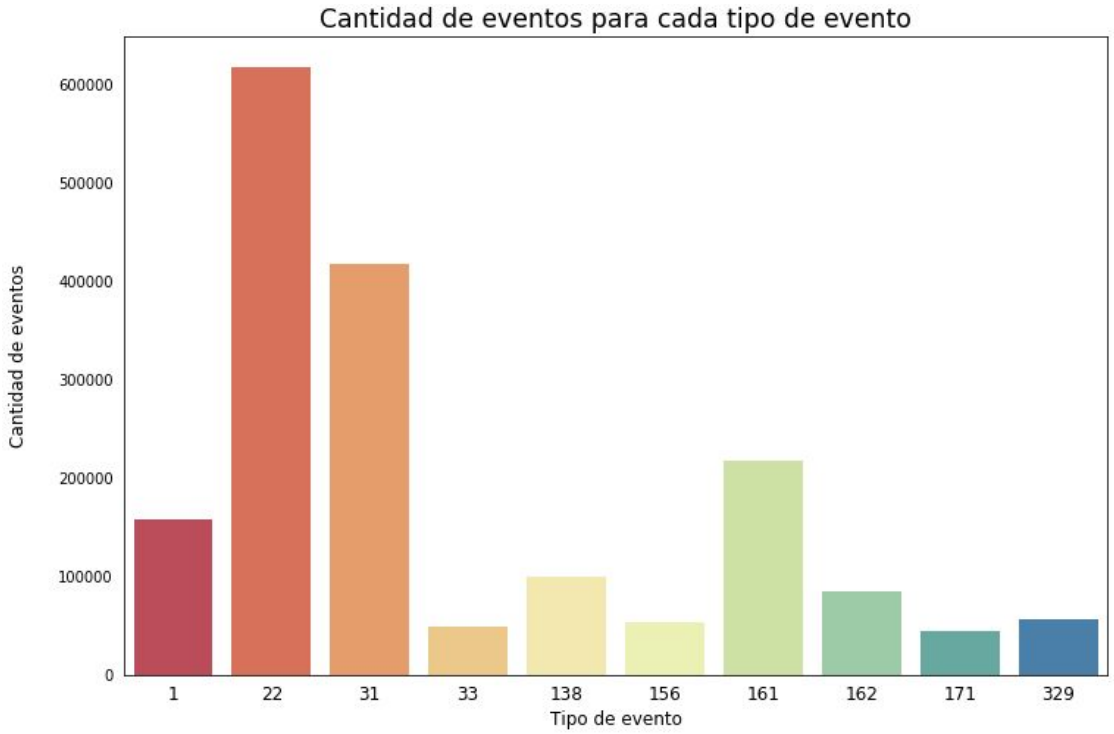
El dataset cuenta con información de múltiples eventos. Dado que esta variable se encuentra codificada, no conocemos de qué eventos se trata, pero igualmente

entendemos que el patrón observado de manera general, podría no ser el mismo para todos los tipos de evento.
Quisimos analizar entonces la distribución de los 568 tipos de eventos que se encuentran reportados en el dataset.



La cantidad de eventos reportados para los distintos tipos de eventos es muy variable, existen eventos con una gran cantidad de ocurrencias mientras que también existen otros muy poco frecuentes. Para poder visualizar en el mismo gráfico estos valores tan diversos, decidimos hacerlo en escala logarítmica.

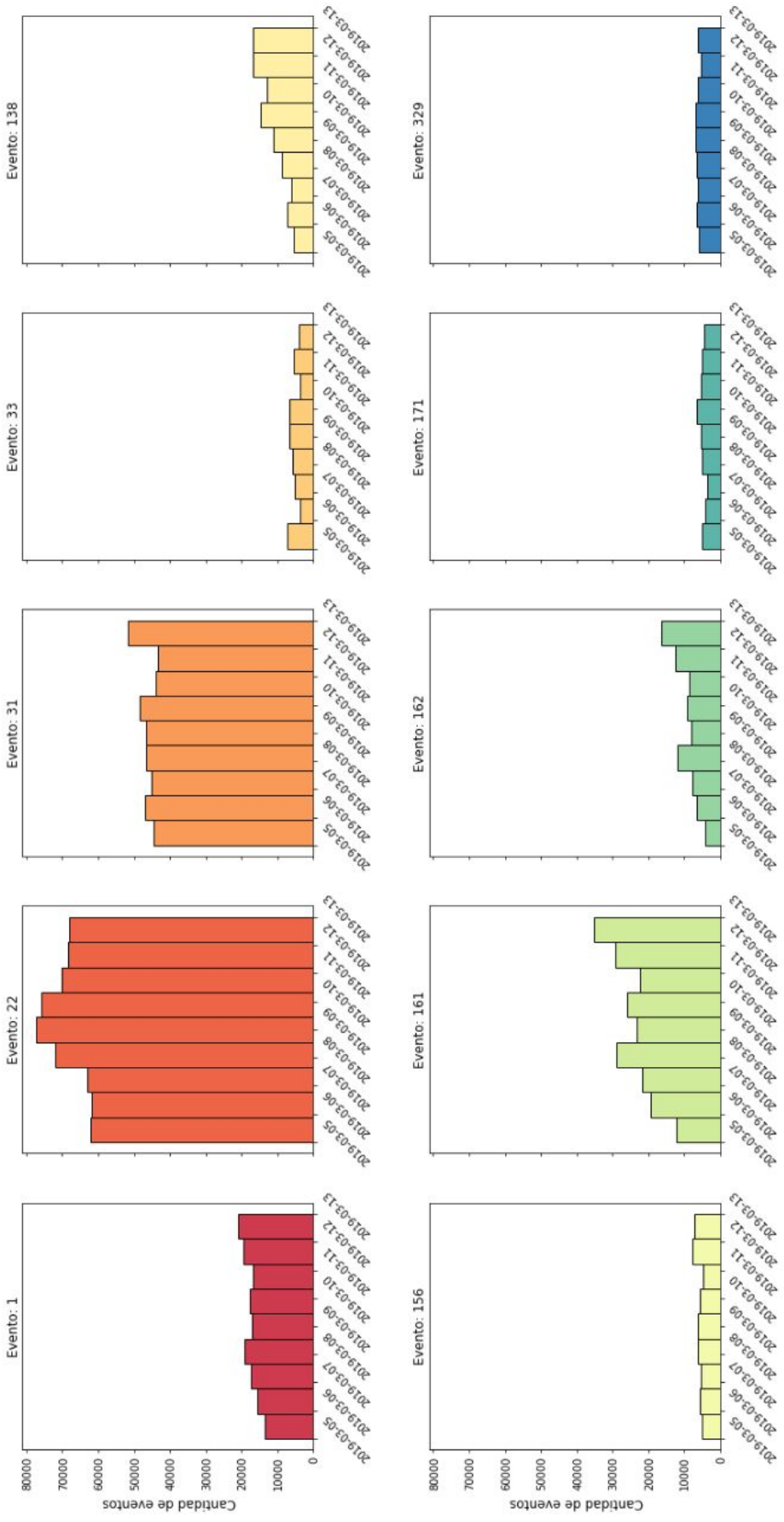
Aunque sabemos que los eventos con mayor cantidad de ocurrencias podrían no ser los más relevantes, a fines de simplificar el análisis, y como primera aproximación, nos quedaremos con los 10 eventos más frecuentes.
Estos 10 eventos mayoritarios corresponden a más del 70% de los eventos reportados en el dataset.



Incluso tomando únicamente los 10 eventos mayoritarios, la distribución de las ocurrencias es muy irregular.

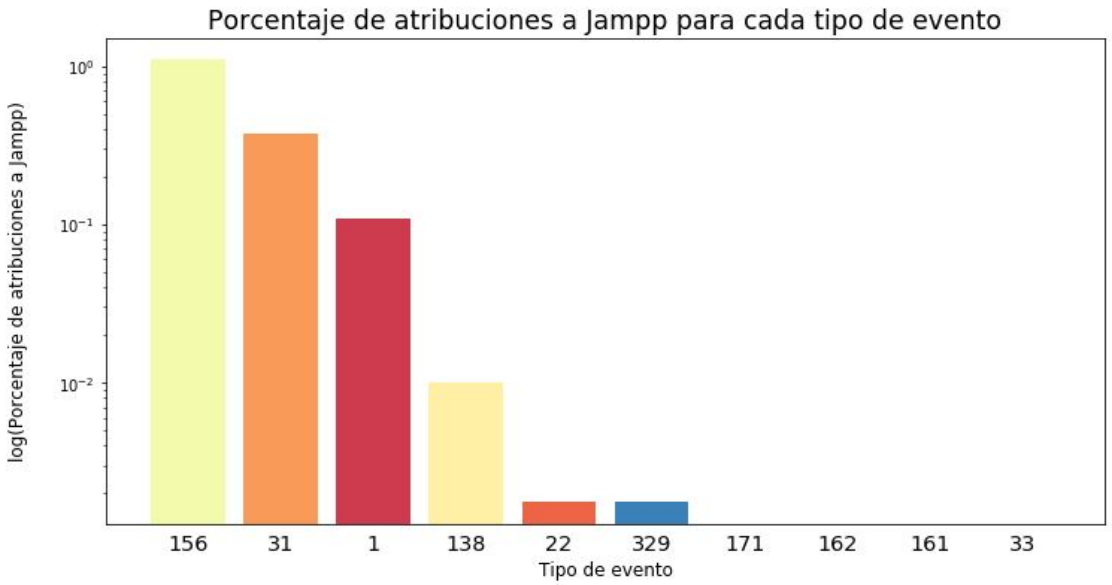
Dada esta variabilidad en la cantidad de cada tipo de evento reportado, el patrón de eventos con un gran número de ocurrencias podrían estar enmascarando patrones más sutiles, más difíciles de detectar, de eventos con menor cantidad de reportes. Debido a esto, analizamos la distribución de cada uno de estos eventos a lo largo de los 9 días de estudio.

No observamos patrones diferenciales para la distribución de los 10 eventos mayoritarios, vemos que las ocurrencias para cada tipo de evento se distribuyen homogéneamente a lo largo del período analizado.


Evento: 1
Evento: 22
Evento: 31
Evento: 33
Evento: 138
Evento: 156
Evento: 161
Evento: 162
Evento: 171
Evento: 329
Cantidad de eventos
0
10000
20000
30000
40000
50000
60000
70000
80000
2019-03-05
2019-03-06
2019-03-07
2019-03-08
2019-03-09
2019-03-10
2019-03-11
2019-03-12
2019-03-13

3.4.4. Atribuciones a Jampp

Cuando un evento ocurre como consecuencia de una impresión mostrada al usuario por Jampp, se dice que este evento ha sido atribuído.
Nos preguntamos cómo son las atribuciones a Jampp para cada tipo de evento.



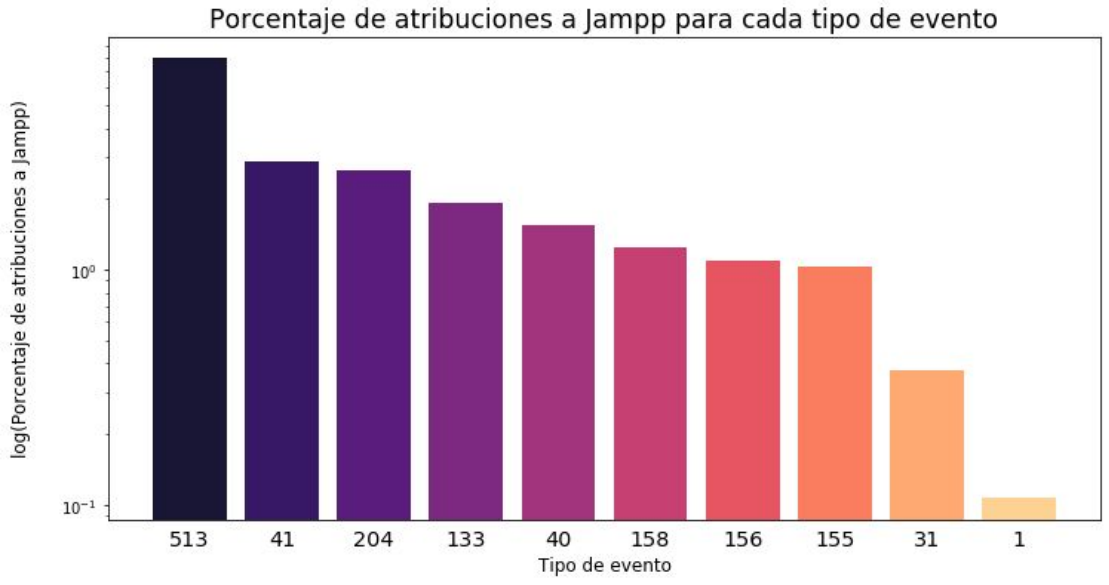
Vemos que los porcentajes varían para cada tipo de evento, y dado que hay una amplia diferencia entre ellos decidimos graficarlos en escala logarítmica para poder visualizarlo de mejor manera.

Código de evento	Porcentaje de atribución
1	0.108
22	0.002
31	0.373
33	0
138	0.01
156	1.095
161	0
162	0
171	0
329	0.002

Los porcentajes de atribución son pequeños, y además algunos eventos nunca son atribuídos. El evento con más porcentaje de atribuciones es aquel con id 156, siendo de un orden de magnitud mayor que el resto de los eventos.

Dado que el evento con mayor porcentaje de atribución no es uno de los eventos con mayor cantidad de ocurrencias, ¿podrían existir eventos con menor cantidad de ocurrencias, pero con mayores porcentajes de atribución a Jampp? Quizás al tomar los 10 eventos más frecuentes pudimos haber dejado de lado eventos con más atribuciones.

Buscamos entonces los eventos con mayores atribuciones a Jampp: decidimos enfocarnos en los eventos con mayor frecuencia de atribución, en lugar de aquellos con mayor porcentaje de atribución. Eventos con altos porcentajes de atribución muchas veces correspondían a eventos muy poco frecuentes (por ejemplo, un evento con 100% de atribución ya que sólo se registró una vez), y debido a que esto podría guiarnos a tomar conclusiones equivocadas, preferimos trabajar con las muestras más grandes posibles.



Podemos observar que mientras que algunos de los eventos con más ocurrencias son también los que brindan mayores porcentajes de atribución, también existen otros eventos minoritarios que son mayormente atribuídos a Jampp.

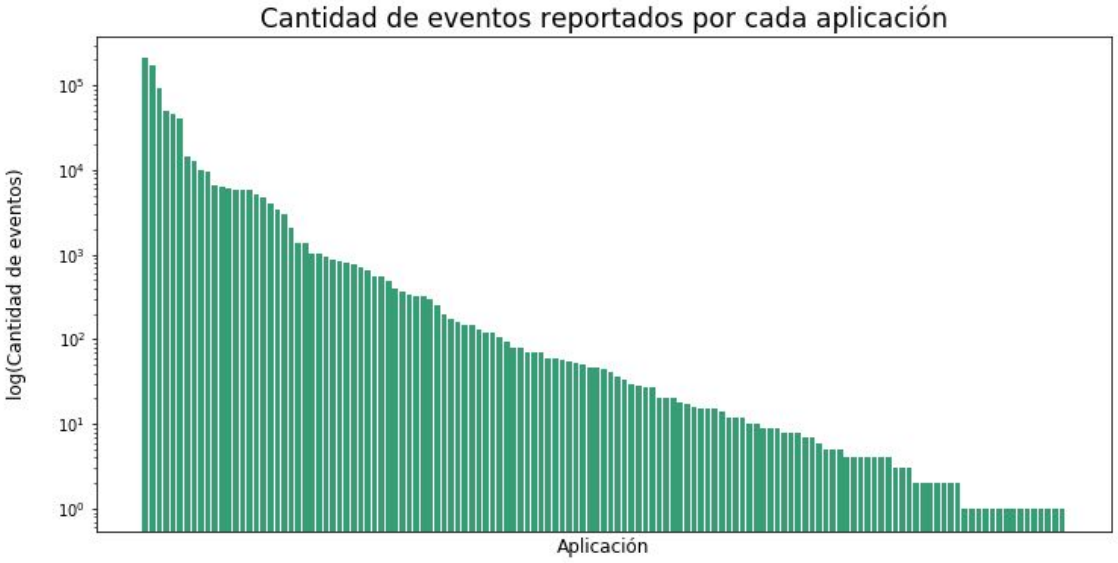
Dado que nos interesan principalmente el subset de eventos que han sido atribuídos, no nos centraremos en los tipos de eventos más frecuentes, sino en aquellos que presentan mayores porcentajes de atribución.

Código de evento	Porcentaje de atribución
1	0.108
31	0.373
40	1.554
41	2.902
133	1.932
155	1.035
156	1.095
158	1.239
204	2.657
513	7.917

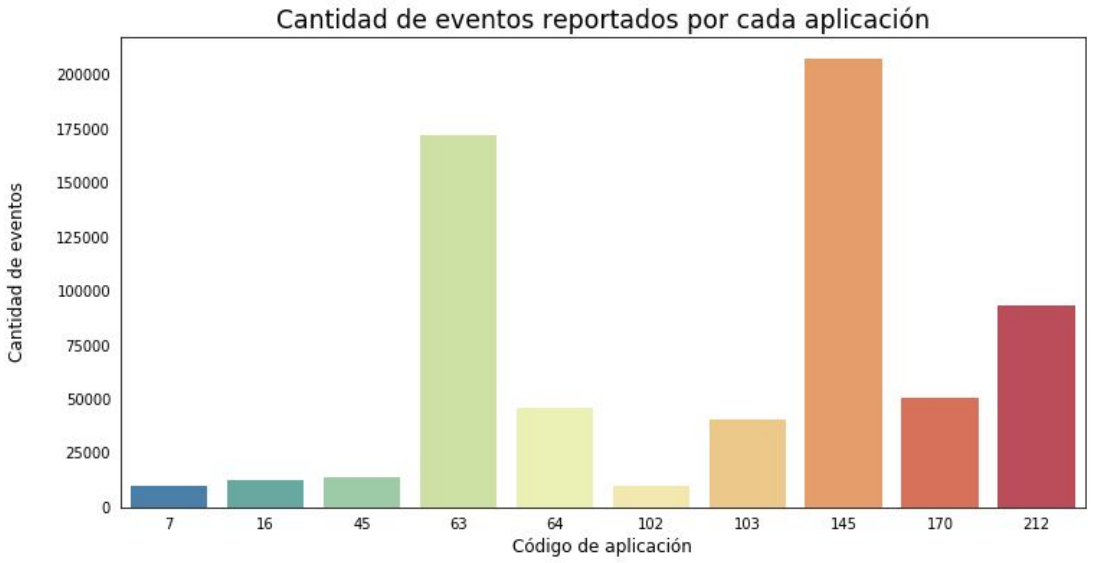
Para los análisis siguientes, utilizaremos el dataset con la información de los 10 eventos con mayores atribuciones a Jampp.

3.4.5. Aplicaciones que reportan datos a Jampp

Como siguiente paso, nos propusimos analizar un poco a las distintas aplicaciones que reportan datos a Jampp. En el dataset se encuentran datos otorgados por 133 aplicaciones, pero queremos ver cómo se distribuye la cantidad de eventos reportados por cada una de ellas.



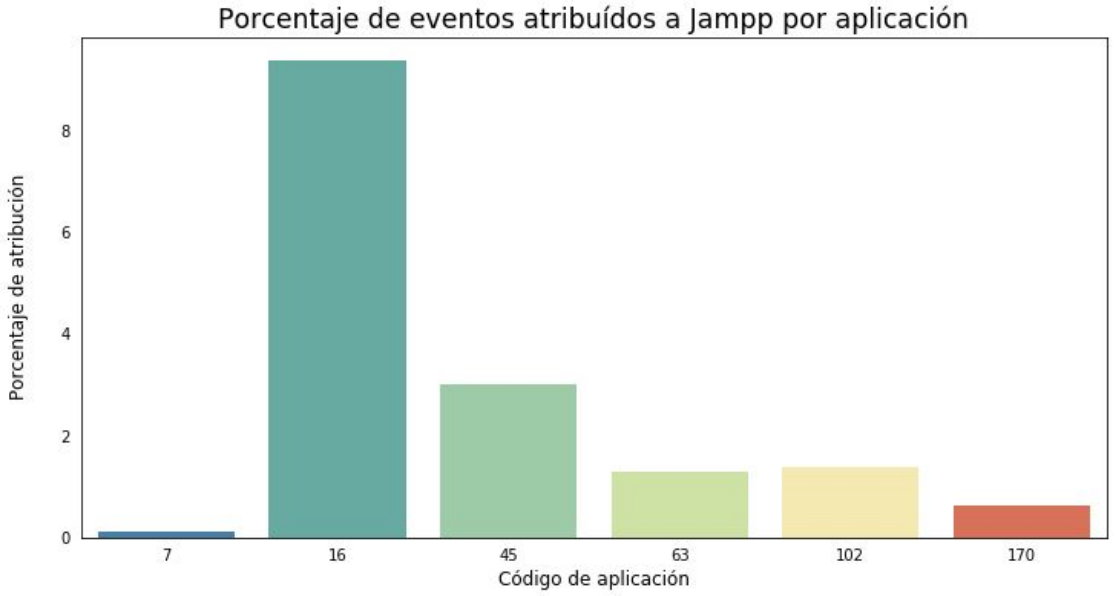
De la misma manera que para los tipos de eventos, la cantidad de eventos reportados por cada aplicación se distribuye de manera heterogénea. Para simplificar el análisis, y trabajar con el tamaño muestral más grande posible, nos quedaremos con las 10 aplicaciones con mayor cantidad de eventos reportados. A través de estas 10 aplicaciones se reportan un 26% de los eventos totales.



La cantidad de eventos reportados varía según la aplicación. ¿Cómo se distribuyen las atribuciones de los eventos a Jampp para cada aplicación?

Código de aplicación	Porcentaje de atribuciones
7	0.103
16	9.375
45	3.023
63	1.302
64	0
102	1.393
103	0
145	0
170	0.631
212	0

Vemos que no todas las aplicaciones con mayor cantidad de reportes presentan eventos atribuídos.



3.4.6. Conclusiones

Del análisis exploratorio del dataset con la información del comportamiento de los usuarios dentro de las aplicaciones surge que los eventos se distribuyen de manera bastante homogénea a lo largo de las fechas analizadas, y presentan un patrón cíclico. La mayor cantidad de ocurrencias suceda cercana a la medianoche, mientras que el mínimo se presenta alrededor de las 8 AM.

El dataset contiene información de múltiples eventos, entre ellos hay algunos con una gran cantidad de ocurrencias y otros muy poco frecuentes. Analizando un subset de los 10 eventos más frecuentes, encontramos el mismo patrón que para la totalidad de los eventos: las frecuencias son muy variables, y se reportan de manera homogénea a través de todo el período temporal analizado.

Dentro de los 10 eventos mayoritarios, los porcentajes de atribución a Jampp son pequeños, y existen eventos sin ninguna atribución. Si observamos únicamente los eventos con mayor cantidad de atribuciones a Jampp del total de los eventos, podemos encontrar que algunos de los eventos más frecuentes son también los que otorgan mayores porcentajes de atribución. Sin embargo, existen también eventos con menor cantidad de ocurrencias, pero que son mayormente atribuídos a Jampp.

Cuando analizamos las distintas aplicaciones de las que surgen los datos sobre los eventos, observamos que los eventos reportados se distribuyen de manera muy variable entre las aplicaciones. Además, de las aplicaciones con mayor cantidad de eventos reportados, sólo algunas presentan eventos atribuídos a Jampp.

3.5. Eventos e instalaciones

3.5.1. Introducción

En primer lugar nos propusimos analizar a los usuarios cuya información estuviera presente tanto en el dataset con información sobre los clicks como en el dataset con información de las instalaciones. Estos usuarios, serían aquellos que al ver una publicidad presentada por Jampp clickearan en ella, y posteriormente instalaran la aplicación sugerida en la publicidad.

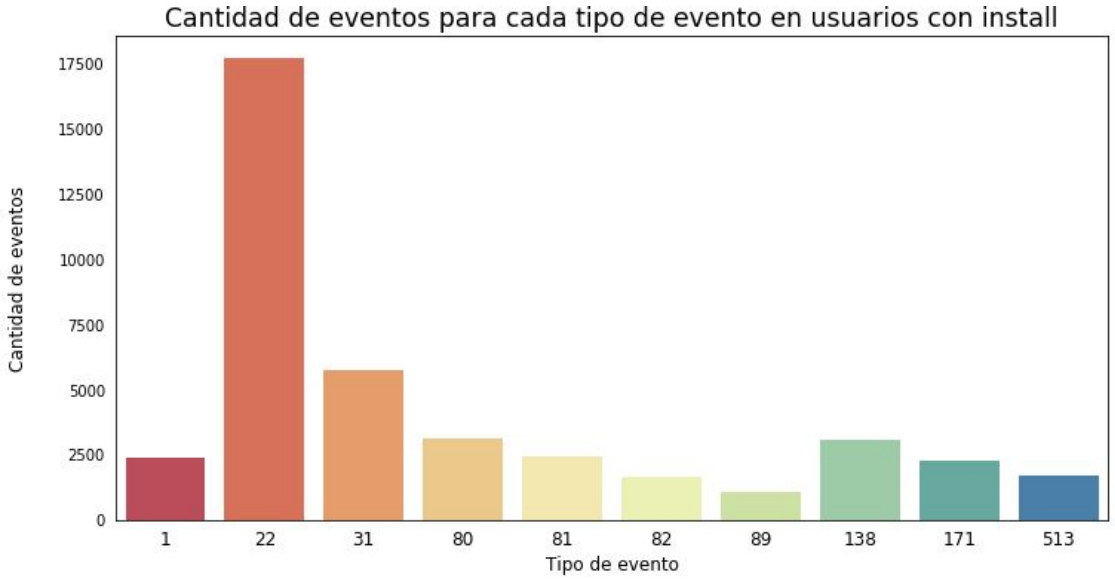
Sin embargo, la unión de estos dos datasets a través del id del usuario da como resultado sólo 11 coincidencias. Es decir, que sólo 11 instalaciones podrían ligarse a un click en una publicidad. El resto de las instalaciones son del tipo implícito, es decir que el usuario instaló la aplicación por sus propios medios, sin mediar publicidades, o lo hizo pasada la ventana temporal de atribución.

Decidimos entonces unir la información de las instalaciones con el dataset con información del comportamiento de los usuarios dentro de las aplicaciones, con el objetivo de detectar patrones de comportamiento de los usuarios.

3.5.2. Eventos de usuarios con instalación

Trabajaremos entonces con un subset de usuarios que presentan datos comportamentales otorgados por las aplicaciones, pero también instalaron alguna aplicación.

Nos propusimos en primer lugar analizar los tipos de eventos mayoritarios. De manera análoga al dataset de eventos, nos quedamos con los 10 eventos mayoritarios, que dentro de este subset de usuarios representan el 77% de los eventos totales.



Vemos que algunos de los 10 eventos mayoritarios en usuarios con instalaciones concuerdan con eventos mayoritarios del caso general. Sin embargo, otros sólo aparecen en este subset de usuarios.

Eventos presentes en ambos casos	Eventos presentes sólo en caso general	Eventos presentes sólo en usuarios con install
1	33	80
22	156	81
31	161	82

138	162	89
171	329	513

3.5.3. Atribuciones a Jampp de usuarios con instalación

Realizamos en este caso un análisis sobre los eventos con mayor cantidad de atribuciones a Jampp dentro de este subset de usuarios. Buscamos los 10 eventos con mayor cantidad de atribuciones, sin embargo, sólo encontramos 6 eventos con alguna atribución.

Código de evento	Cantidad de atribuciones
513	87
239	8
31	7
238	4
153	2
518	1

Podemos observar que el número de atribuciones en este grupo de usuarios es pequeño, y con una gran diferencia entre el evento con más ocurrencias y el que le sigue, siendo ésta de un orden de magnitud. Tomaremos al evento con código 513 como el evento de mayor atribución y descartaremos el resto, ya que los números son demasiado pequeños y podríamos cometer errores relacionados al tamaño muestral.

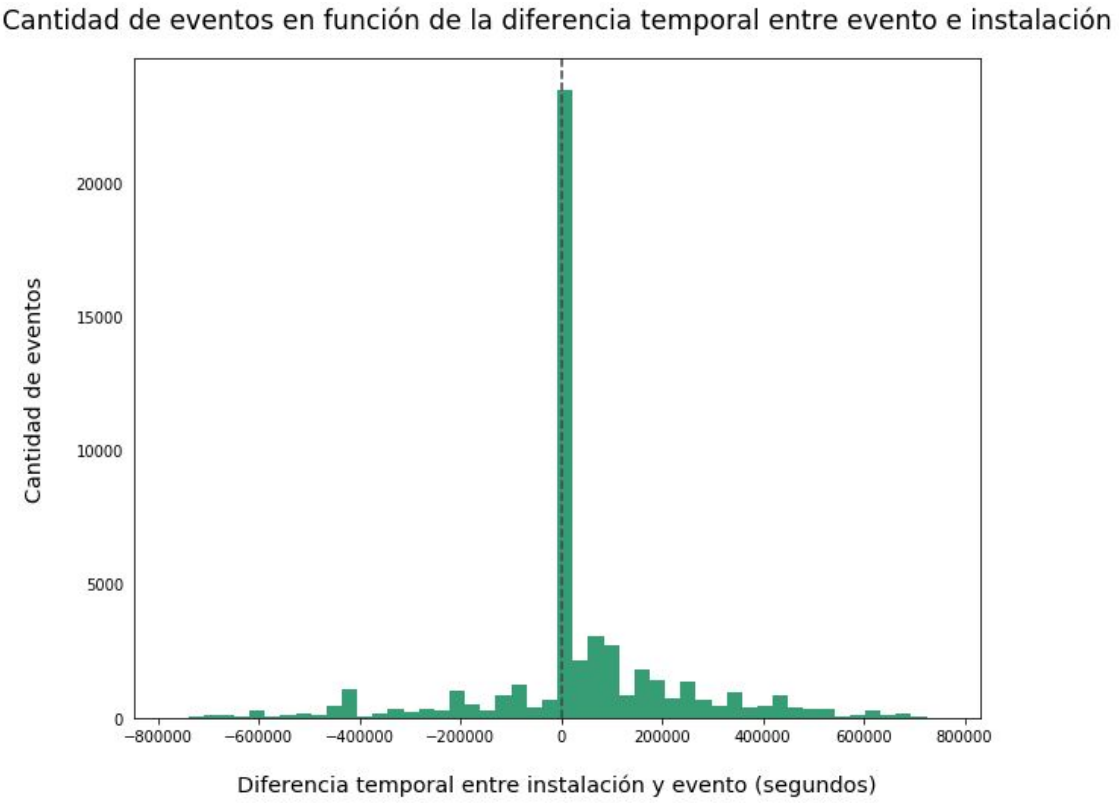
En la sección 3.4.4 vimos que el evento 513 es también en el caso general el evento con mayor porcentaje de atribuciones a Jampp. Podemos concluir que este es un evento importante debido a que es el más atribuído, pero dado que se encuentra en ambos grupos, no podemos relacionarlo directamente con las instalaciones.

3.5.4. Temporalidad de los eventos

Si bien vimos que el evento de código 513 es frecuente entre los usuarios que instalaron, nos gustaría tener una idea de la temporalidad de los eventos. Dado que no conocemos el código del evento de instalación, pero si conocemos el momento en que

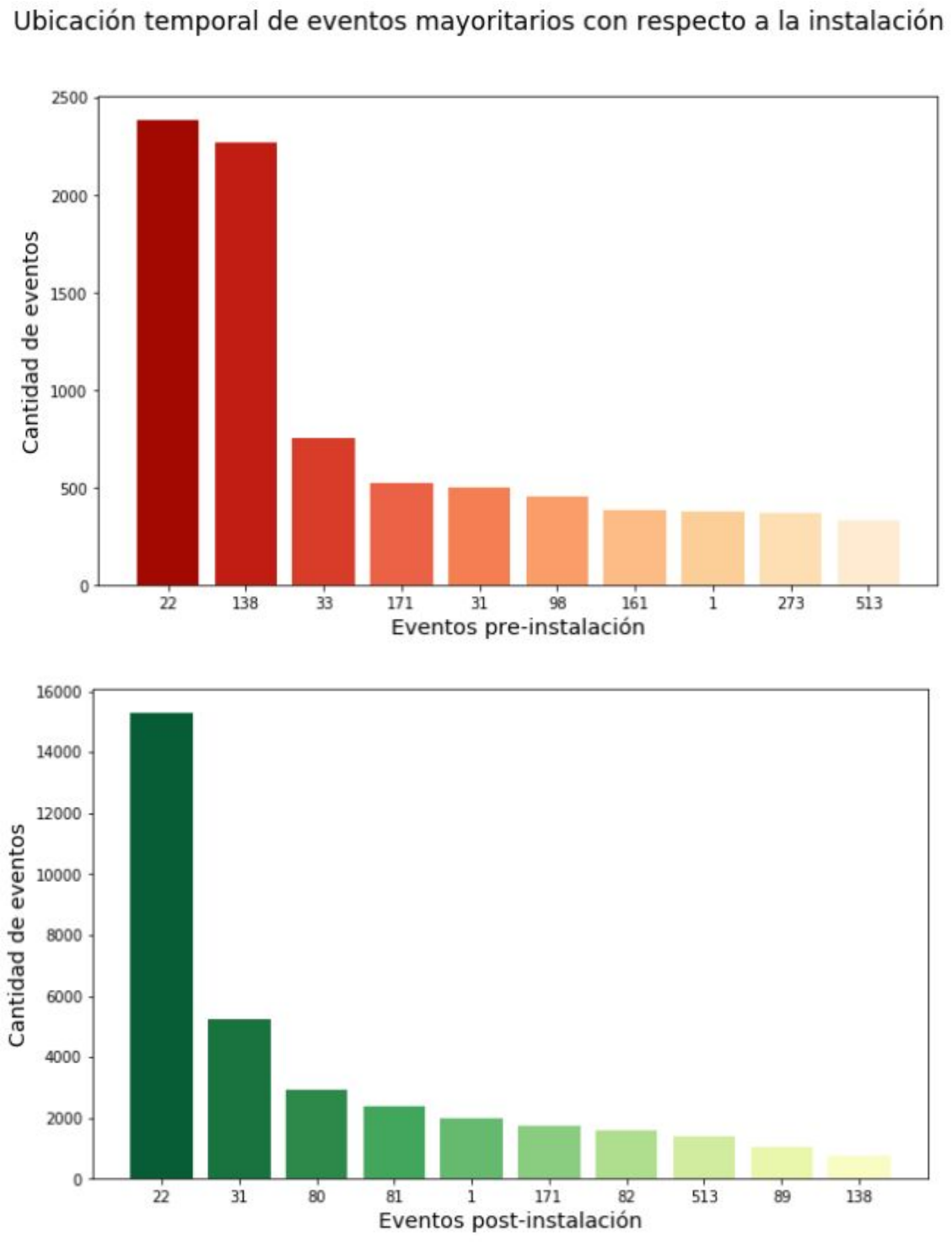
se realizó un install, nos propusimos identificar los eventos que ocurren próximos a una instalación.

Queremos ver cómo es la distribución de los eventos alrededor del evento de instalación. Para esto, calculamos la diferencia temporal entre el evento de instalación y el resto de los eventos.



La línea punteada representa el evento de instalación, vemos que hay ocurrencia de eventos pre y post-instalación. Sin embargo, la mayor cantidad de eventos se dan muy cercanos a la instalación. Nos interesa saber cuáles son estos eventos.

Los eventos previos a la instalación son aquellos que presentan un delta temporal negativo, mientras que los eventos posteriores a la instalación son los que poseen un delta temporal positivo.



Podemos observar que hay eventos que se encuentran en ambos grupos, es decir que presentan ocurrencias a lo largo de todo el período temporal, mientras que otros eventos se encuentran mayoritariamente pre o post-instalación.

Eventos que ocurren pre y post instalación	Eventos que ocurren sólo pre-instalación	Eventos que ocurren sólo post-instalación
1	33	80
22	98	81
31	161	82
33	273	89
80		
81		
82		
89		
98		
138		
161		
171		
273		
513		

3.5.5. Comportamiento pre-instalación

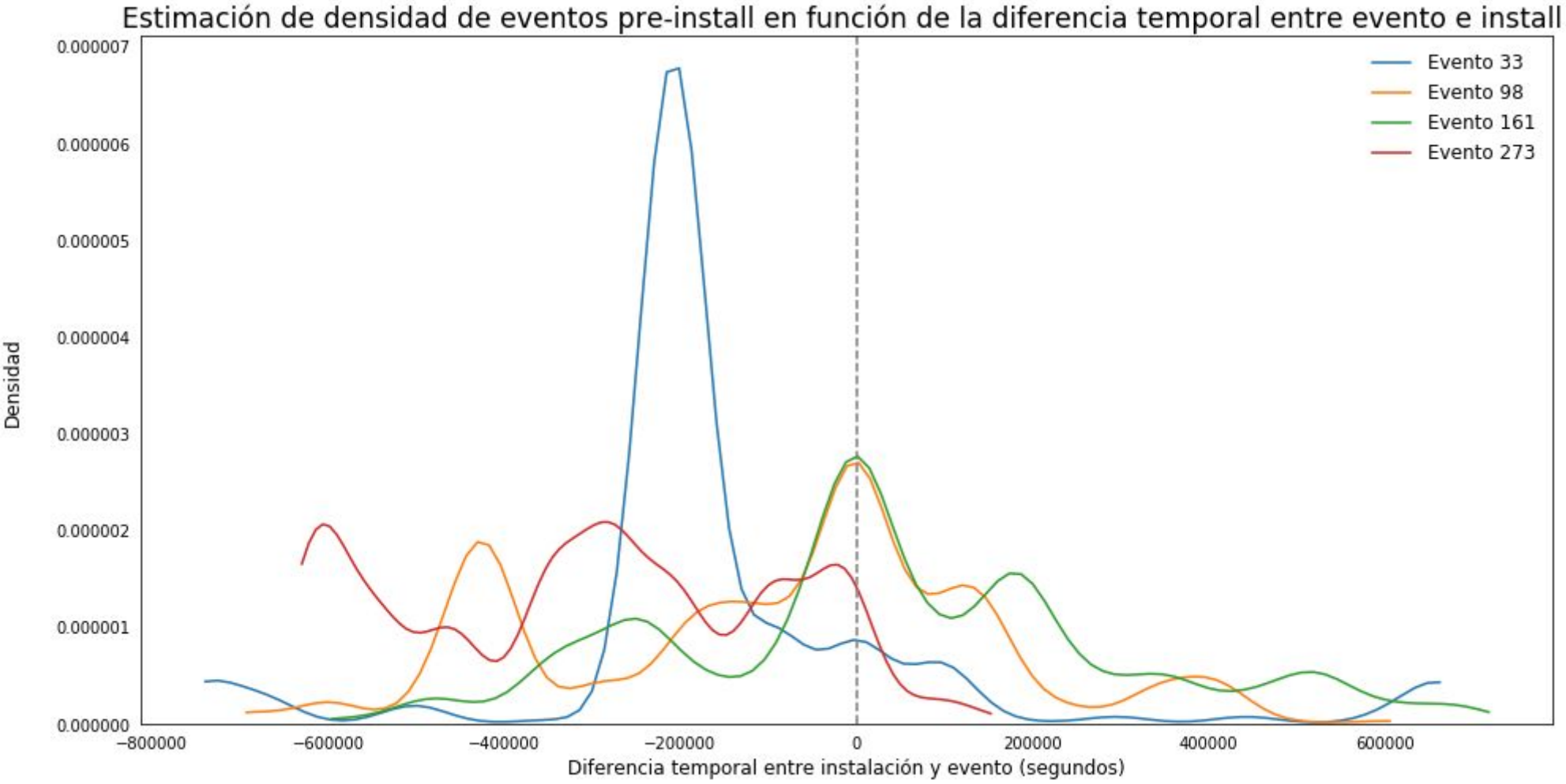
Dado que los eventos que ocurren a lo largo de todo el período temporal y los eventos que ocurren post-instalación no nos sirven para predecir un install, nos centraremos únicamente en los eventos previos. Queremos conocer la distribución de estos eventos, de manera de poder analizar si alguno se encuentra ligado con el evento de instalación.

Los eventos de interés son los de código 33, 98, 161 y 273. Si graficamos la densidad de eventos en función del delta temporal con la instalación, podremos observar el patrón temporal de los eventos, en caso de existir alguno.

En la visualización siguiente, podemos observar que antes de la instalación (tiempo = 0, marcado con una línea punteada) se observa un pico muy notorio del evento 33, seguido tiempo después de un progresivo aumento de los eventos 98 y 161.

La presencia de este patrón de eventos en un usuario podría ser útil para predecir una futura instalación. Dado que el evento 33 se encuentra más alejado temporalmente del evento de instalación, probablemente los eventos más significativos sean el 161 y 98, que se encuentran mucho más cercanos temporalmente y alcanzan su máximo en el momento de la instalación.

Cabe destacar que, a menos que sean mutuamente excluyentes, la presencia de dos eventos distintos que presentan un máximo de ocurrencias muy cercano al evento de instalación favorece aún más la detección de las posibles instalaciones.


Estimación de densidad de eventos pre-install en función de la diferencia temporal entre evento e install
Densidad
0.000007
0.000006
0.000005
0.000004
0.000003
0.000002
0.000001
0.000000
−800000
−600000
−400000
−200000
0
200000
400000
600000
Diferencia temporal entre instalación y evento (segundos)
Evento 33
Evento 98
Evento 161
Evento 273

3.5.6. Conclusiones

Analizamos un subset de los datos totales, los usuarios que presentan eventos reportados así como una instalación, con el objetivo de encontrar diferencias en el comportamiento respecto al análisis general de todos los usuarios.

En primer lugar buscamos los eventos con mayor cantidad de ocurrencias, y los eventos con mayores atribuciones a Jampp. Vimos que el evento más atribuído, el de código 513, es también el más atribuído en el análisis general, por lo que no podemos ligarlo directamente al evento de instalación.

Posteriormente, nos propusimos analizar la temporalidad de los eventos. ¿Cuándo ocurren los eventos mayoritarios en este grupo de usuarios, que también presenta un evento de instalación? Calculamos la diferencia temporal entre la instalación y los distintos tipos de eventos, logrando identificar un subgrupo de eventos que ocurren mayoritariamente previos a la instalación.

Logramos, además, identificar un patrón temporal de ocurrencia de ciertos eventos previos a una instalación. Dado que al momento de decidir si participar en una subasta para un cierto dispositivo o no, Jampp recibe información del usuario, creemos que esta información podría ser útil para predecir una posible instalación.

3.6. Eventos y clicks

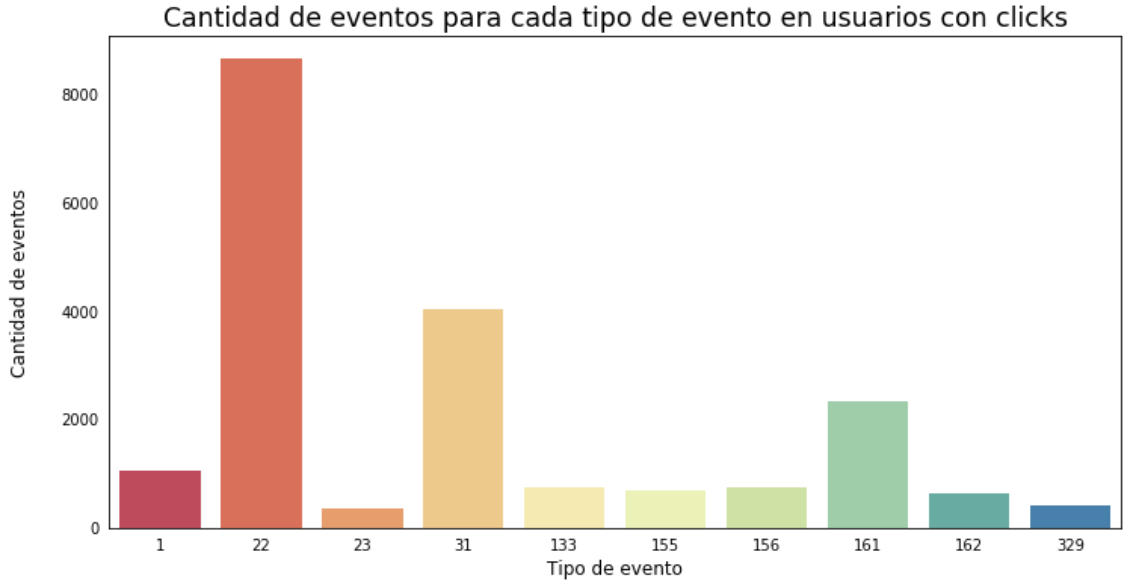
3.6.1. Introducción

Otro eje que nos resultó relevante para el análisis remite a aquellos usuarios cuya información estuviera presente tanto en el dataset con información sobre los clicks como en el dataset con información de los eventos. Estos usuarios, serían aquellos que realizan un click ante un evento determinado.
De este modo nos pareció que podíamos profundizar a la relación existente entre los distintos eventos de los que tenemos información en los que participaron los diferentes usuarios y los clicks que realizan estos usuarios realizan.

3.6.2. Eventos de usuarios con clicks

El eje de este análisis se trata de un subset de aquellos usuarios que de los cuales se pueden observar comportamientos específicos de las aplicaciones a la par que realizaron un click.

Nos propusimos en primer lugar analizar los eventos que tienen mayor recurrencia de clicks dentro del dataset. En este sentido, nos decidimos por un recorte de los 10 mayores, lo cual representa un porcentaje del 83 % del total de la muestra.



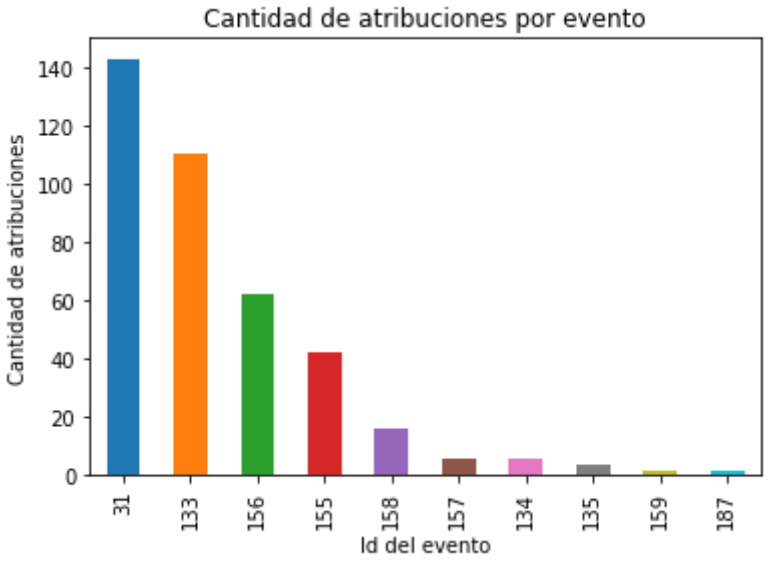
En este sentido podemos observar que el evento con más eventos asociados es el 22 con 8645. Lo sigue el evento 31 con 4040 eventos asociados y en tercer lugar el 161 con 2345 eventos.

Vemos entonces la existencia de eventos que están presentes dentro de las 10 aplicaciones con mayor cantidad de eventos por clicks

Eventos presentes en ambos casos	Eventos presentes sólo en caso general	Eventos presentes sólo en usuarios con clicks
1	33	23
22	138	133
31	171	155
156	-	-
161	-	-
162	-	-
329	-	-

3.6.3. Atribuciones a Jampp de usuarios con clicks

En este caso buscamos indagar en la influencia de Jampp dentro de los eventos con clicks asociados. Para este caso buscamos los 10 eventos con mayor cantidad de atribuciones a Jampp.



Podemos observar así que el evento con mayor cantidad de atribuciones a Jampp es el 31 con 143. Seguido por el evento 133 que cuenta con 110 atribuciones. Finalmente el evento 156 posee 62 atribuciones.
Nos interesa profundizar en esto y establecer la relación de estos eventos con la totalidad de eventos. Para esto estableceremos el porcentaje de cada uno de estos dentro del total.

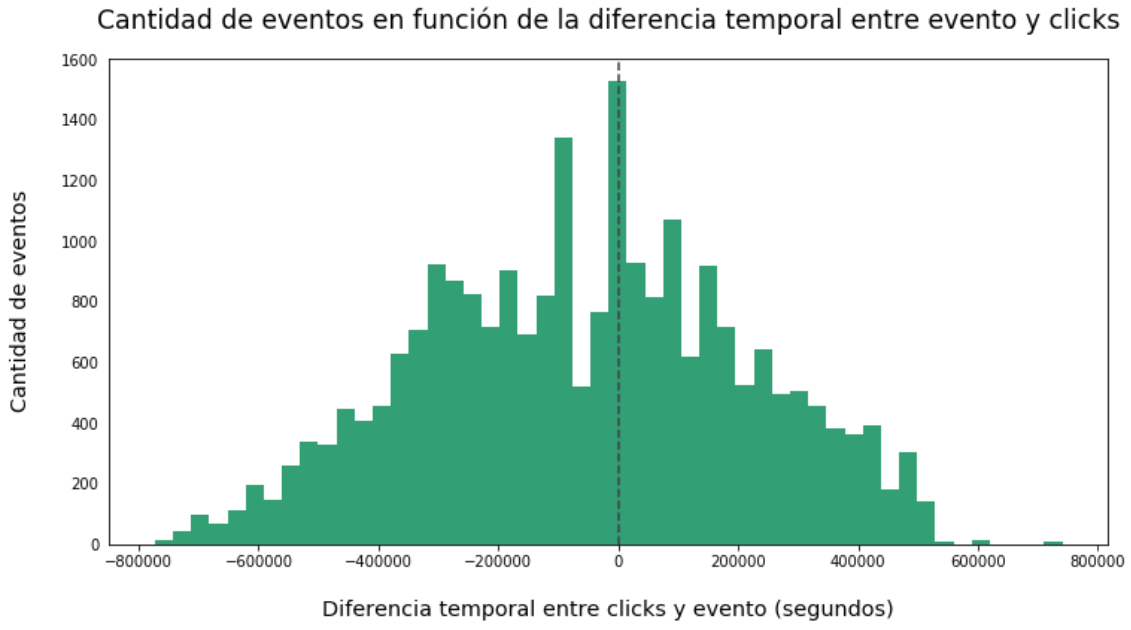
Código de evento	Porcentaje en relación al total de eventos
31	3.53 %
133	15.00 %
134	4.58 %
135	37.50 %

155	6.04 %
156	8.22 %
157	8.77 %
158	5.36 %
159	0.77 %
187	14.28 %

3.6.4. Distribución temporal de los eventos

Nos interesa ahora conocer el modo en que se distribuyen estos eventos temporalmente. Nos interesa ahora determinar los eventos que se encuentran próximos a cada click.

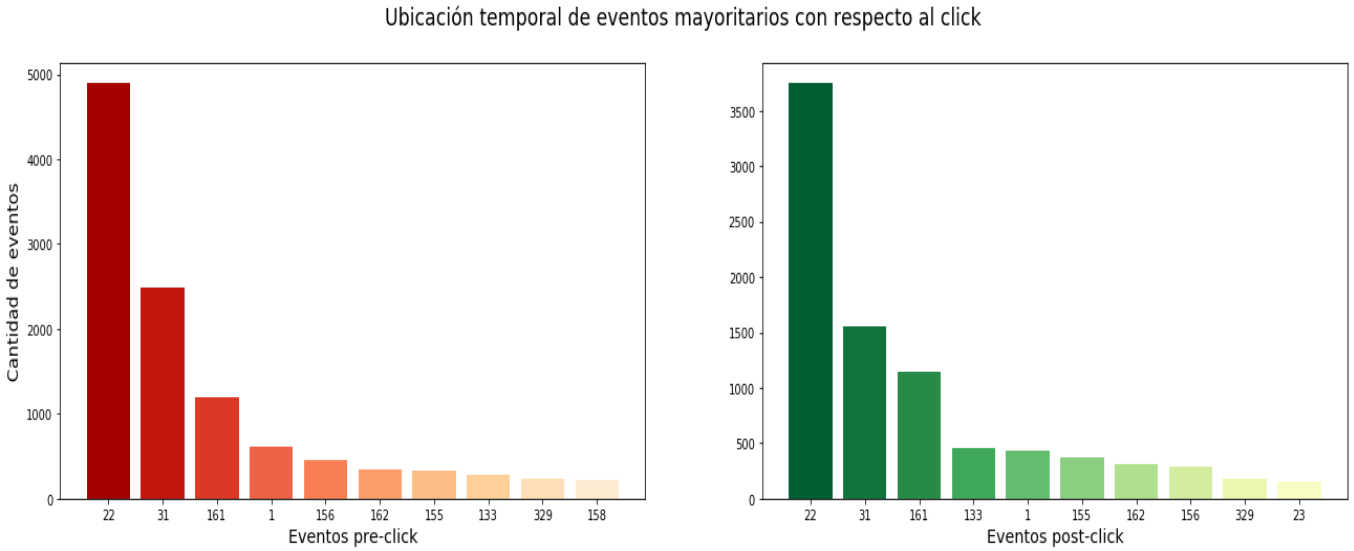
Queremos ver cómo es la distribución de los eventos alrededor del evento de click. Calculamos entonces la diferencia temporal entre el evento de click y el resto de los eventos.



La línea punteada representa el evento del click. Si bien hay una ligera variación post y pre click se puede establecer que la mayor cantidad de eventos se dan muy cercanos al click y que el gráfico es entonces relativamente homogéneo para ambos casos.

De modo similar que las instalaciones, los eventos previos al click son aquellos que presentan un delta temporal negativo, mientras que los eventos posteriores a la instalación son los que poseen un delta temporal positivo.

Nos determinamos a identificar los eventos que interactúan en estos casos y diferenciarlos según sean pre o post click.



Podemos visualizar de este modo cuáles son los eventos que se dan en cada momento, determinando sus recurrencias y sus singularidades.

De este modo, podemos establecer que el único evento que se da antes del click manifestarse luego del click es el evento 158; mientras que a la inversa, el único evento que se da después del click sin manifestarse antes es el evento 23.

3.6.5. Conclusiones

A lo largo de este apartado, pudimos observar como dentro de los eventos con un clic asociado, el evento con más eventos asociados es el 22, el cual representa el 36 % del total de la muestra. Posteriormente lo siguen el evento 31 y el 161. Asimismo, esto significa que más del 63 % de la totalidad de los eventos se encuentra nucleada en estos 3 eventos.