Trabajo Práctico 1 — Análisis Exploratorio

[7506/9558] Organización de Datos
Primer cuatrimestre de 2019

FRITZ, Lautaro Gastón	102320
ROLDÁN, María Cecilia	102999

Índice

1. Introducción

El siguiente trabajo práctico fue realizado para la asignatura Organización de Datos [75.06/95.58] de la Facultad de Ingeniería de la Universidad de Buenos Aires. El objetivo general fue analizar los sets de datos provistos por la empresa Jampp y en base a eso, alcanzar conclusiones sobre los datos.

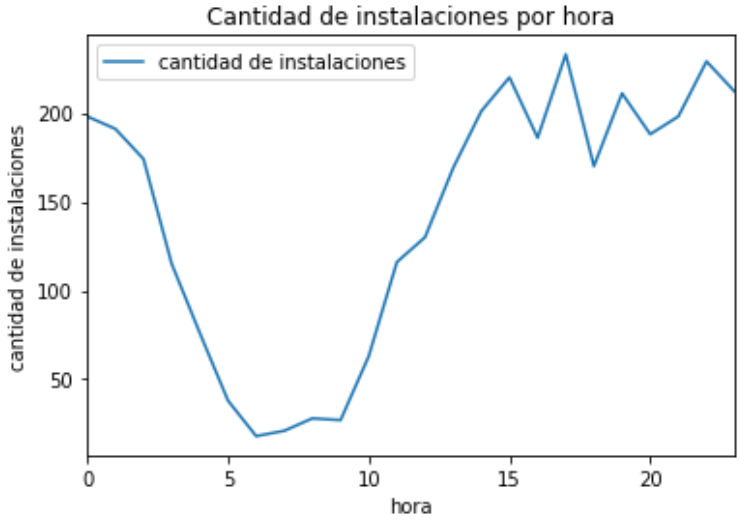
Se disponía de los siguientes archivos, en formato .csv:

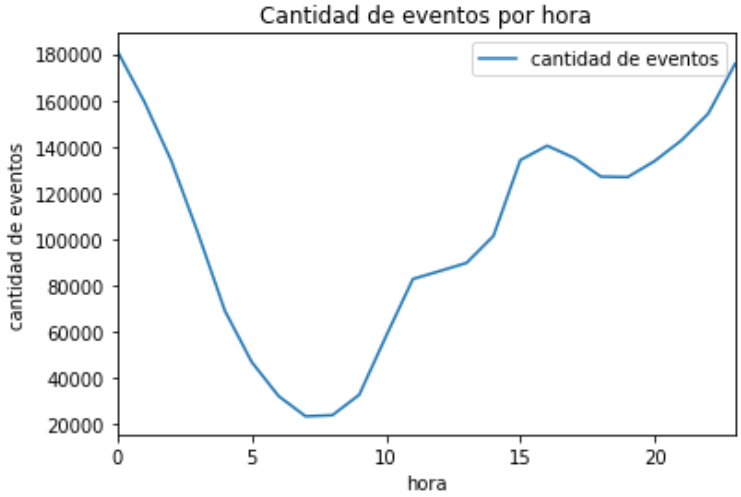
- **installs.csv:** Contiene información acerca de las instalaciones de las aplicaciones de empresas clientes de Jampp.
- **clicks.csv:** Posee información sobre los clicks efectuados en publicidades de aplicaciones de clientes de Jampp.
- **events.csv:** Tiene información acerca de eventos generados por usuarios dentro de aplicaciones de clientes de Jampp.
- **auctions.csv:** Posee información sobre subastas que le llegaron a Jampp, sin importar si participaron o no.

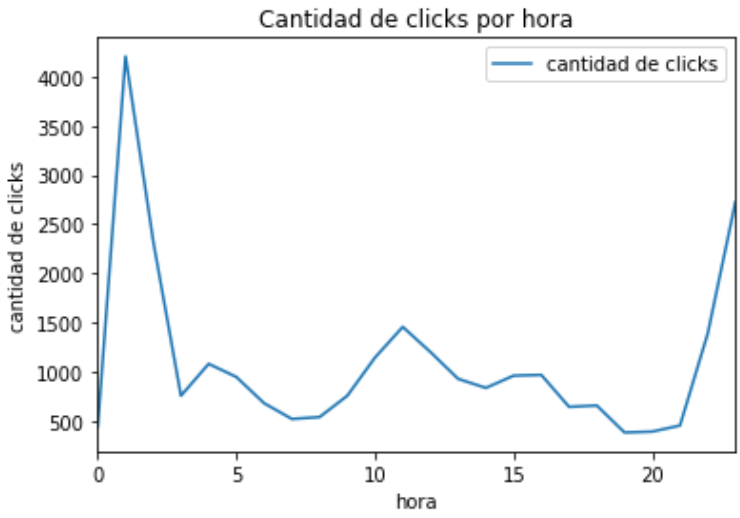
2. Análisis de cantidades por hora

Se dividió este análisis en dos secciones: analizando cada archivo por separado y luego juntando los datos obtenidos en los de instalaciones y clicks. A priori no parece correcto relacionar los datos de estos dataframes con los del de eventos, debido a la naturaleza de cada uno.

2.1. Análisis por separado





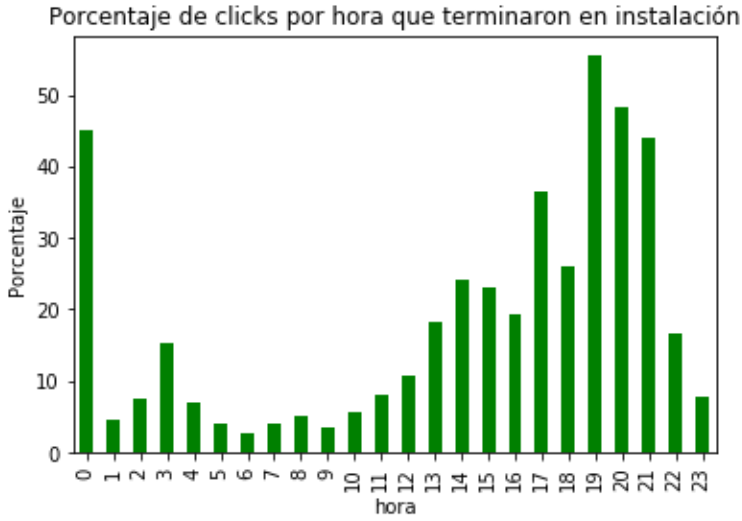


No se pusieron todos los datos en el mismo gráfico ya que los dataframes manejan volúmenes muy disímiles, lo que ocasiona inconvenientes con la escala.

A primera vista, las instalaciones y los eventos son similares: disminuyen notablemente en las primeras horas del día, y alrededor de las 10 comienzan a ascender marcadamente, un ascenso que no disminuye hasta el inicio de día siguiente.

Por el contrario, los clicks no parecen seguir esa tendencia: se disparan hacia la 1 de la mañana, y luego disminuyen súbitamente. Hay algunos ascensos a lo largo del día, pero ninguno tan marcado como el primero. Al igual que los otros gráficos, se ve un incremento en la cantidad a partir de las 10 de la mañana, pero este no dura mucho y cae bastante regularmente durante la tarde, sólo volviendo a subir a la noche.

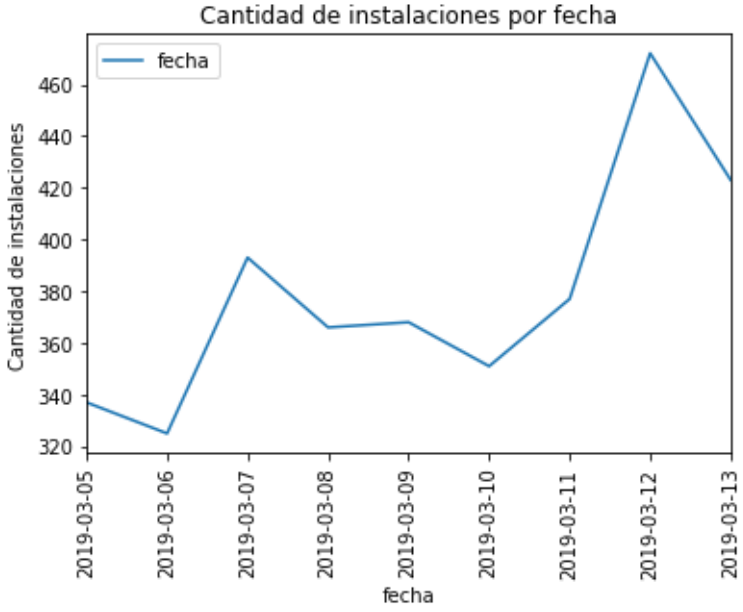
2.2. Análisis en conjunto

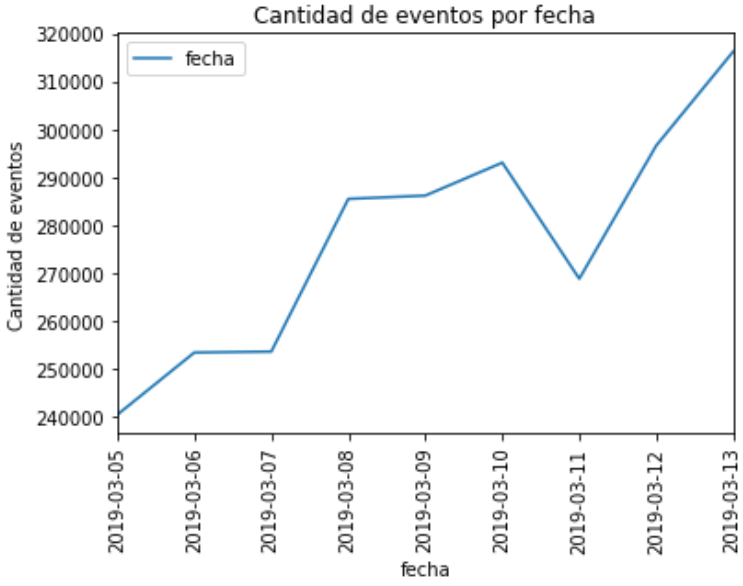


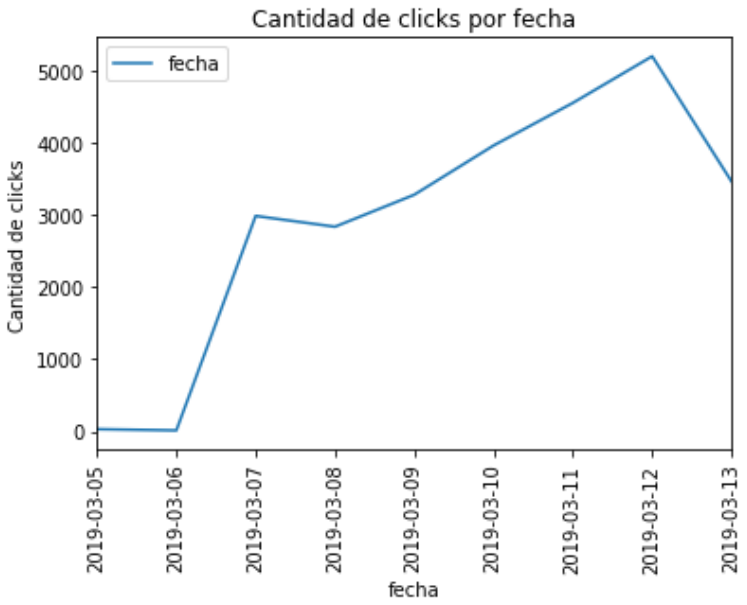
Se aprecia que la franja horaria de la 1 (la que más clicks tiene) posee un porcentaje de éxito similar a la banda horaria de las 7, la quinta con menos clicks. Por el contrario, las cuatro con menor número de clicks (19, 20, 0 y 21 respectivamente) son las franjas con mayor porcentaje de éxito.

3. Análisis de cantidades por día

3.1. Análisis por separado







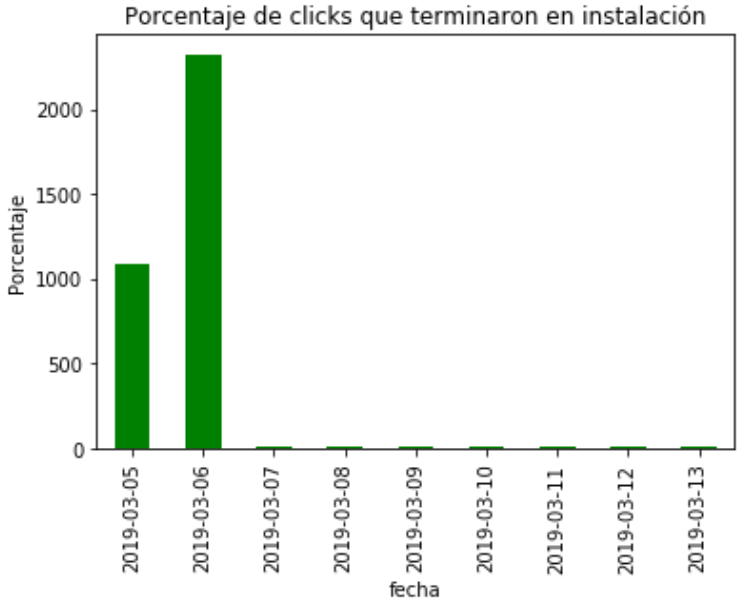
Nuevamente, no se juntaron los gráficos por temas de escala.

Estos gráficos son más disímiles que los anteriores, siendo la única similitud un ascenso en las cantidades a partir del segundo o tercer día. De ahí, cada uno parece comportarse de forma diferente: las instalaciones disminuyen, pero luego alcanzan un pico varios días después, los eventos bajan un día pero vuelven a subir, y los clicks no paran de subir hasta el último día, donde vuelven a disminuir.

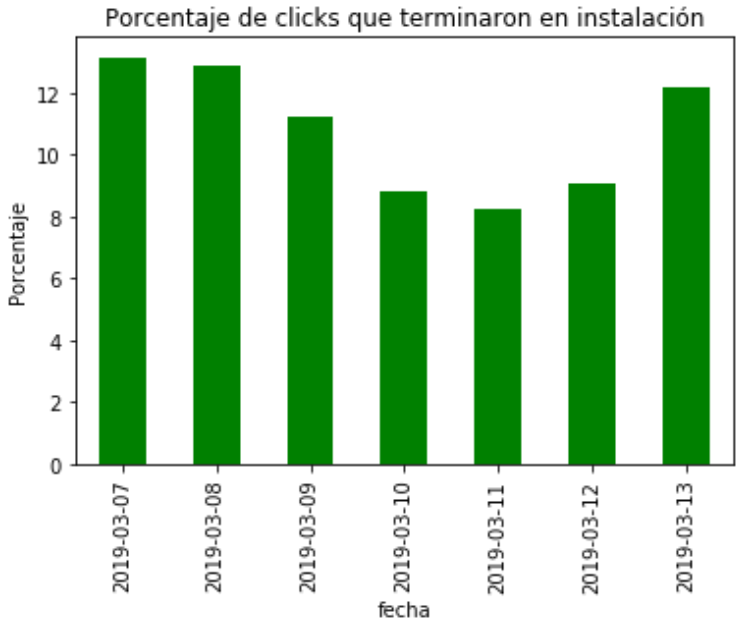
Se ve algo interesante en el de clicks: los primeros dos días tienen muy pocos clicks, cercanos a 0.

También, las instalaciones se distribuyen de forma relativamente pareja a lo largo de los días, habiendo una diferencia de 140 instalaciones aproximadamente entre los días con mayor y menor número de instalaciones.

3.2. Análisis en conjunto



No tiene sentido plantear los porcentajes de éxito por día con estos datos, ya que los dos primeros días tienen una cantidad de clicks cercana a 0 y un número de instalaciones mucho mayor. Por eso, para realizar este análisis, se excluirá a los primeros dos días.



Se ve que los días que más clicks tuvieron son los que menos porcentaje de instalación tienen. Esto es una consecuencia de la distribución pareja de las instalaciones a lo largo de los días.

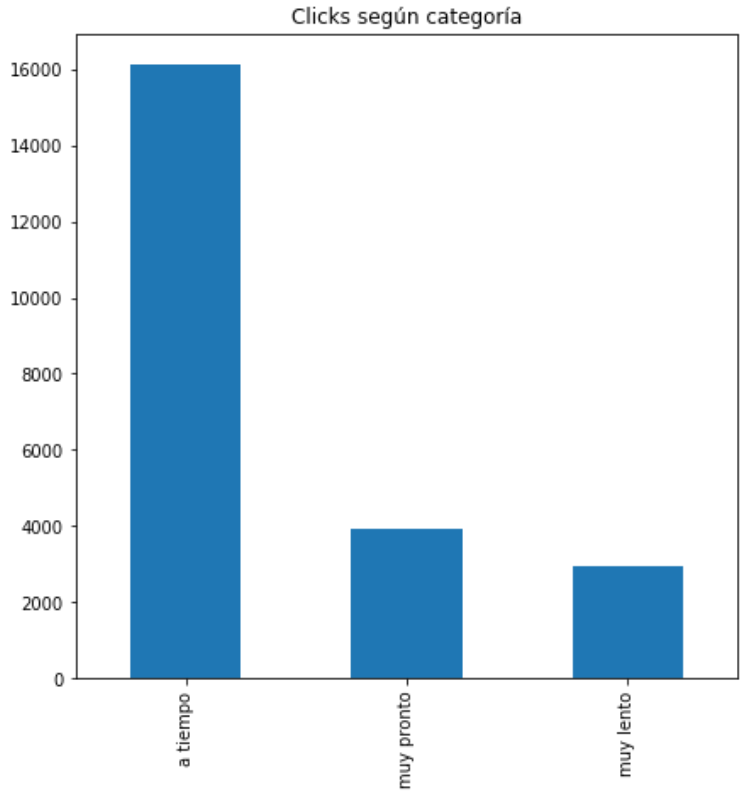
La conclusión inmediata de estos análisis es que ciertamente una mayor cantidad de clicks no implica un mayor número de instalaciones.

4. Análisis de cantidades

4.1. Clicks según el tiempo de demora en clickear por parte del usuario

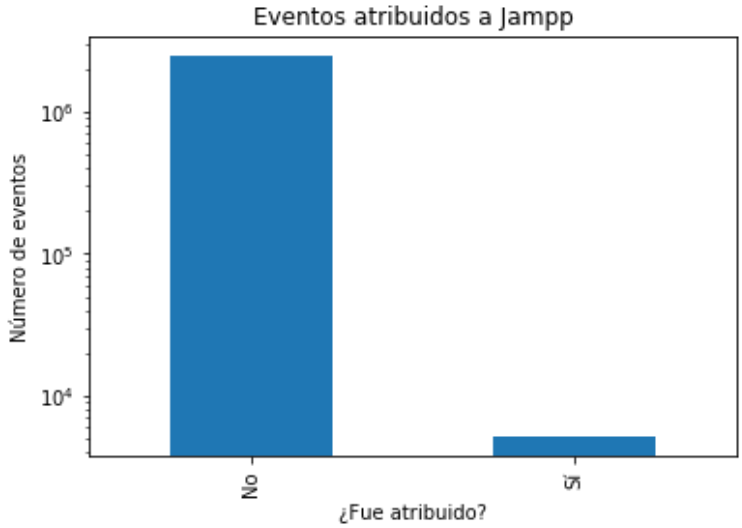
Se dividieron los clicks en tres intervalos:

- El intervalo 'muy pronto' tiene clicks con un tiempo de demora entre los 0 y los 2 segundos. A nuestro criterio, es un tiempo muy corto como para decidir si se desea acceder al servicio publicitado. Los clicks de esta categoría podrían provenir, por ejemplo, de usuarios que quieren pasar rápidamente al próximo nivel del juego que se encuentren jugando, sin saber que les aparecerá una publicidad. Al no esperar esto, tocan la pantalla a los pocos segundos de haber perdido, justo cuando aparece el anuncio.
- La categoría 'a tiempo' contiene clicks con un tiempo de demora entre los 2 y los 300 segundos. Parece un tiempo razonable para determinar si se desea acceder al servicio en cuestión.
- El intervalo 'muy lento' posee clicks con un tiempo de demora entre los 300 y los 18000 segundos (valor máximo del campo). Si una persona está interesada en un servicio es poco probable que tarde más de cinco minutos para decidir si quiere acceder a él. Podrían ser clicks realizados por error.
- No se tomaron en cuenta los clicks cuyo campo 'timeToClick' es NaN. (unos 3374 clicks)



Se observa que si bien la mayoría de los clicks caen en la categoría 'a tiempo', más de seis mil figuran en las otras dos categorías (sobre un total de 22977 clicks).

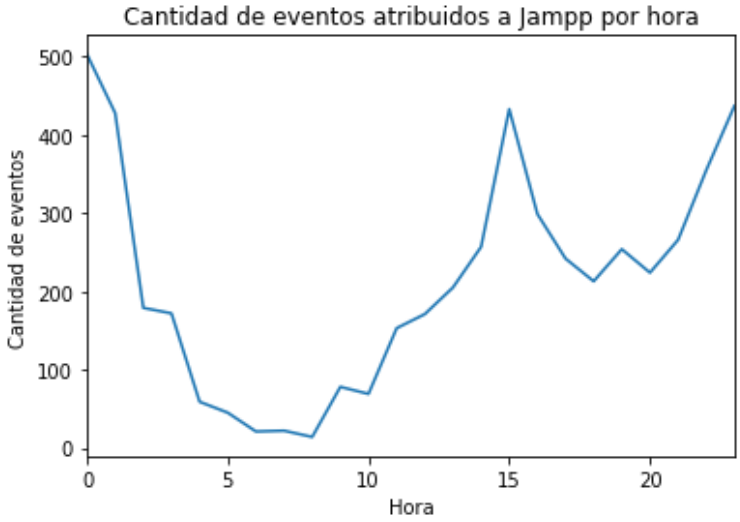
4.2. Cantidad de eventos atribuidos a Jampp



Se aprecia que muy pocos eventos fueron atribuidos a Jampp (5099 sobre una base de 2500000 eventos aproximadamente, un 0,2 % del total).

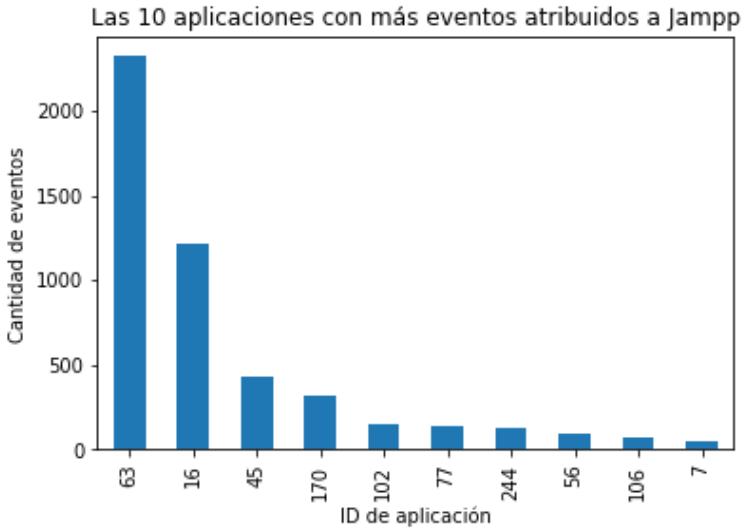
De aquí en adelante se hará una comparación de ciertos aspectos entre el total de los eventos atribuidos a Jampp y los eventos generados por los 10 usuarios que más eventos atribuidos a Jampp generaron.

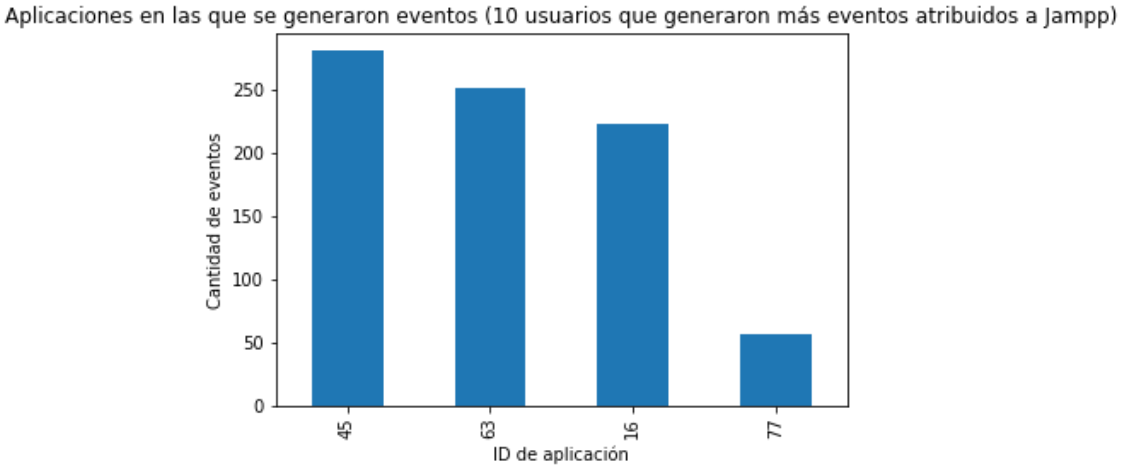
4.2.1. Eventos atribuidos a Jampp por hora



Los gráficos parecen seguir tendencias similares, con algunas diferencias. Ambos caen a partir de la medianoche y suben alrededor de las 10 de la mañana, alcanzando un pico a las 15. Sin embargo, el segundo gráfico tiene una caída antes del pico, cosa que el primero no posee. Luego, ambos caen durante la noche, repuntando alrededor las 23.

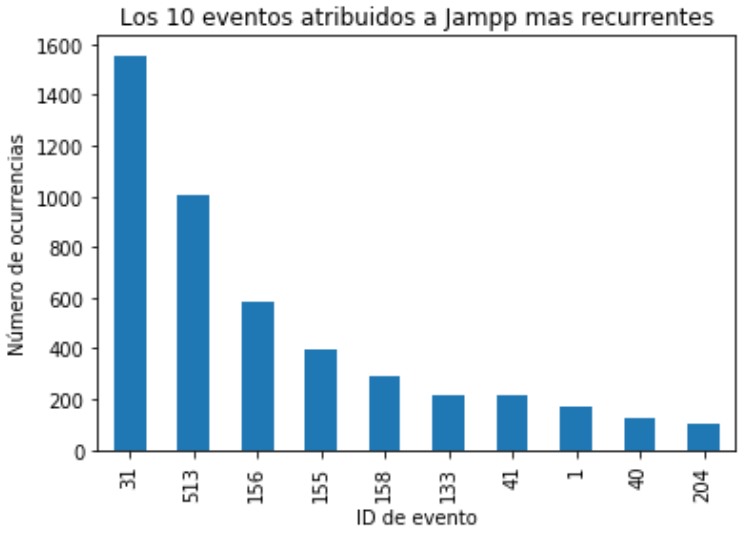
4.2.2. Aplicaciones en las que se generaron los eventos

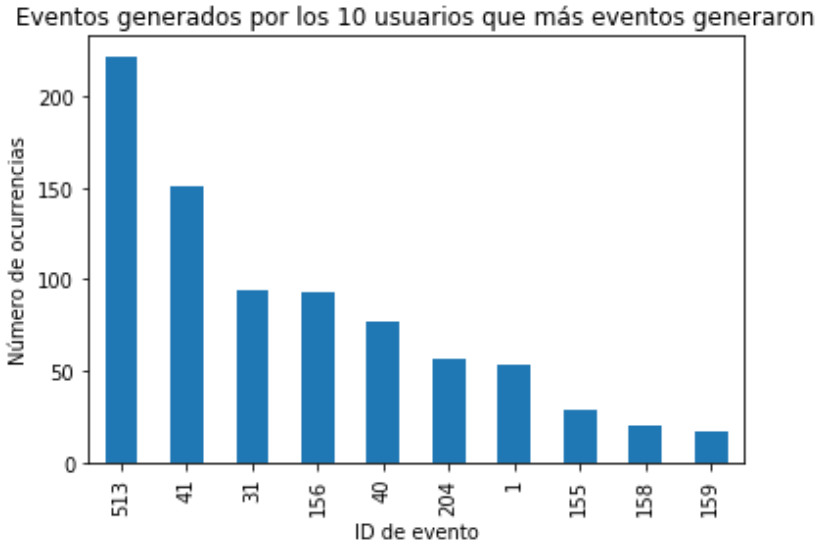




Se puede ver que las cuatro aplicaciones del segundo gráfico están también en el primero. Lo interesante es que los 10 usuarios que más eventos atribuidos a Jampp generaron lo hicieron sólo en 4 aplicaciones (812 de los 5099 en total atribuidos fueron generados en estas 4 aplicaciones). Además, la aplicación 63, que en el primer gráfico tiene muchos más eventos que el resto, es la segunda en el otro gráfico, esto quiere decir que tuvo muchos usuarios que no generaron tantos eventos en ella, a diferencia de la 45, que tuvo menos usuarios que le generaron muchos eventos.

4.2.3. Eventos atribuidos a Jampp

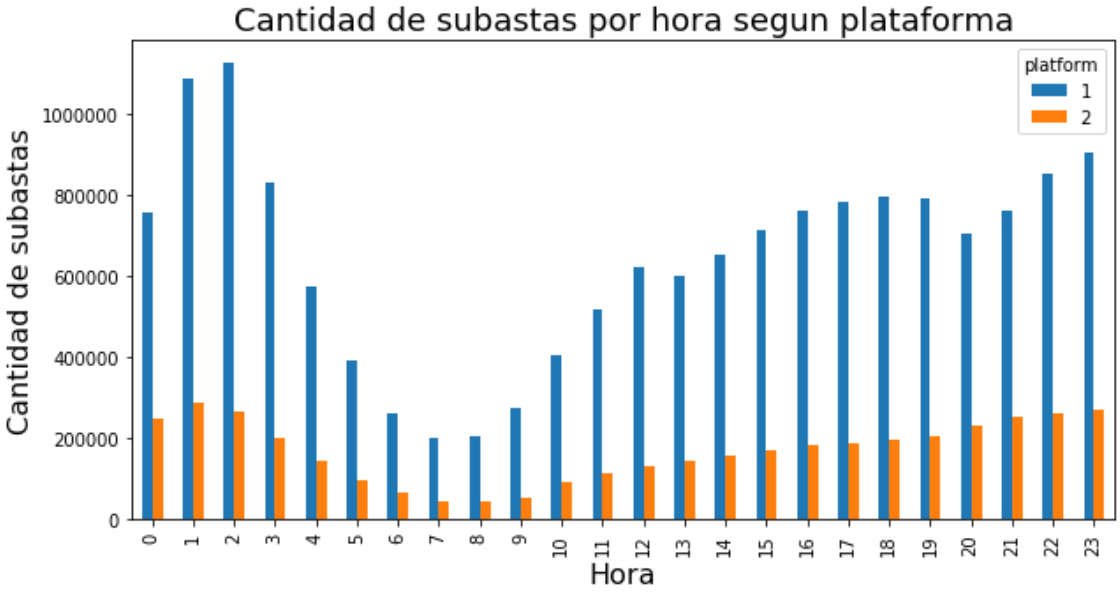


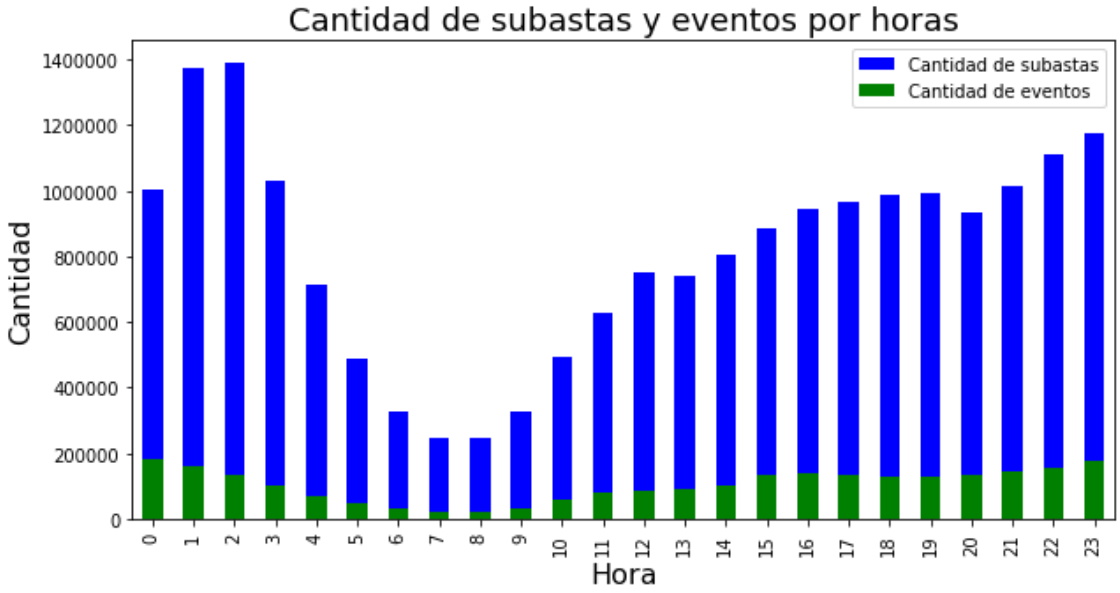


Puede observarse que 9 de los 10 están en ambos gráficos: esto significa que estos eventos ocurrieron muchas veces con los 10 usuarios que más eventos generaron, pero también con el resto de los usuarios. Algunos, como el 31, parecen haberse generado mucho más con el resto de los usuarios que con los 10 primeros, al contrario de otros, por ejemplo, el 41.

5. Subastas

Por último, se generaron dos visualizaciones sobre las subastas.





Estas dos visualizaciones permiten ver que, a lo largo de las horas, las subastas siguen un patrón no muy distinto al que se muestra en los primeros gráficos de las cantidades de eventos e instalaciones por hora.