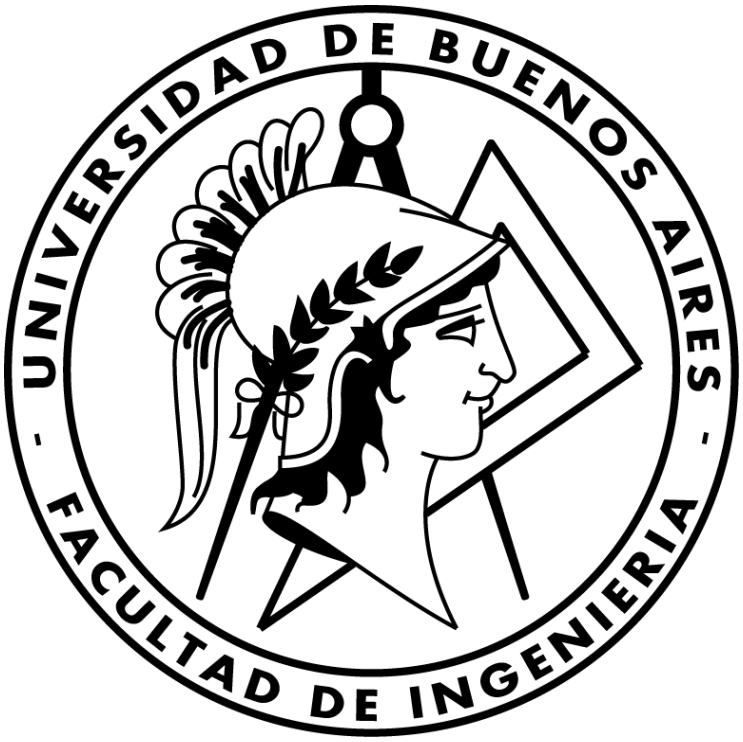


UNIVERSIDAD DE BUENOS AIRES

FACULTAD DE INGENIERÍA

75.06 Organización de Datos

Trabajo Práctico 1

Primer Cuatrimestre de 2019

Grupo 18

Camussi, Alan	95903
Catolino, Lucas	95560
Zugna, Federico	95758

Link de GitHub: https://github.com/LucasCatolino/Orgnizacion-de-Datos

Índice

1. Introduccion

Este informe se encarga de analizar los datos sobre un conjunto de set de datos provisto por la empresa Jampp.

Dentro del directorio del set de datos podemos encontrar:

- installs
- clicks
- events
- auctions

El objetivo del TP es realizar un análisis exploratorio del set de datos.

1. ¿Qué es Jampp?

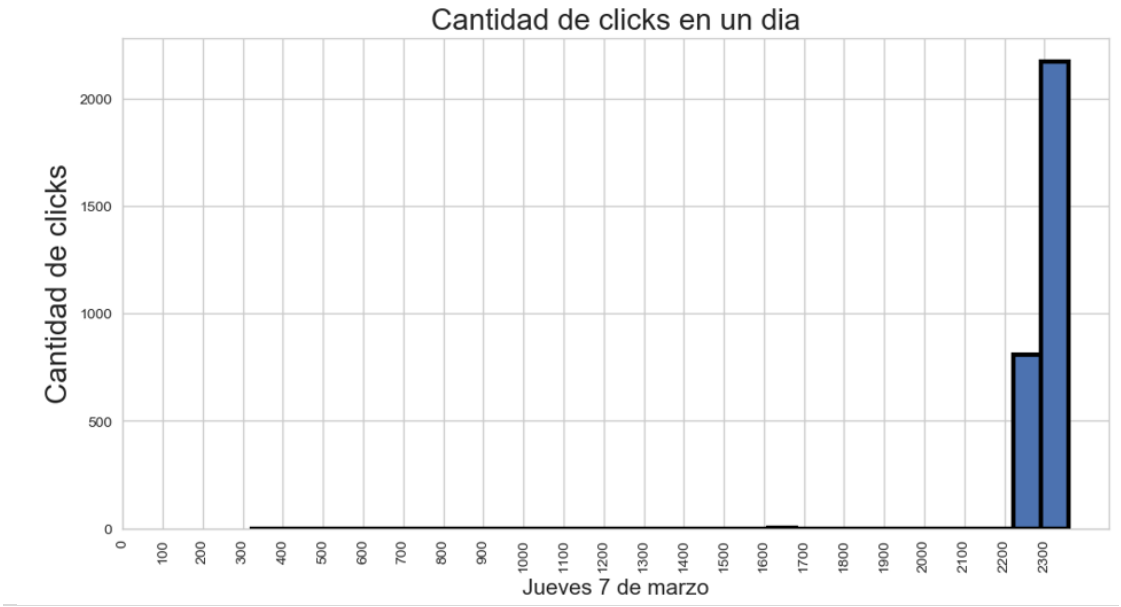
JAMPP es una plataforma de marketing de rendimiento para adquirir y atraer clientes móviles. La compañía combina datos de comportamiento con tecnología predictiva y programática para generar ingresos para los anunciantes al mostrar anuncios personales y relevantes que instan a los consumidores a comprar por primera vez o más a menudo.

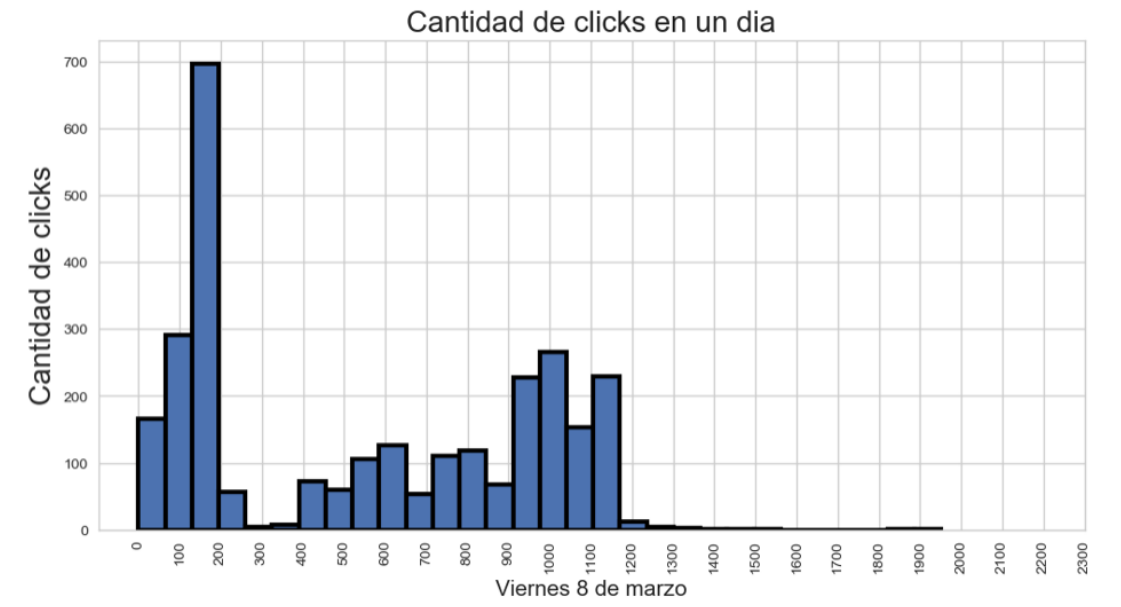
Fundado en 2003, nuestro equipo atiende a una base global de clientes desde oficinas en San Francisco, Londres, Berlin, San Pablo, Singapur, Ciudad del cabo y Buenos Aires.

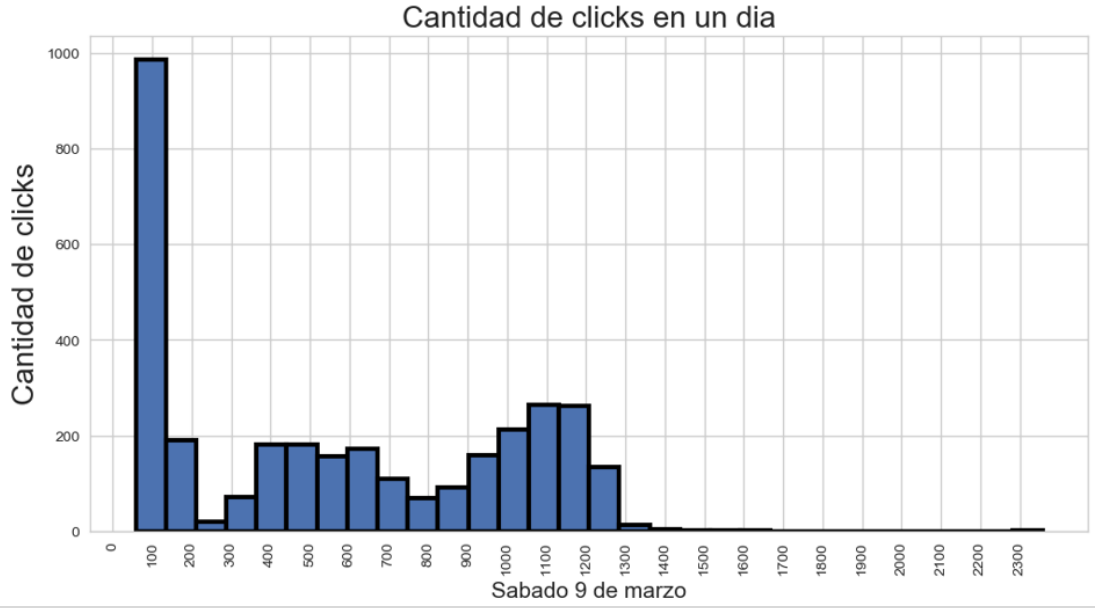
2. Análisis Set de Datos Clicks

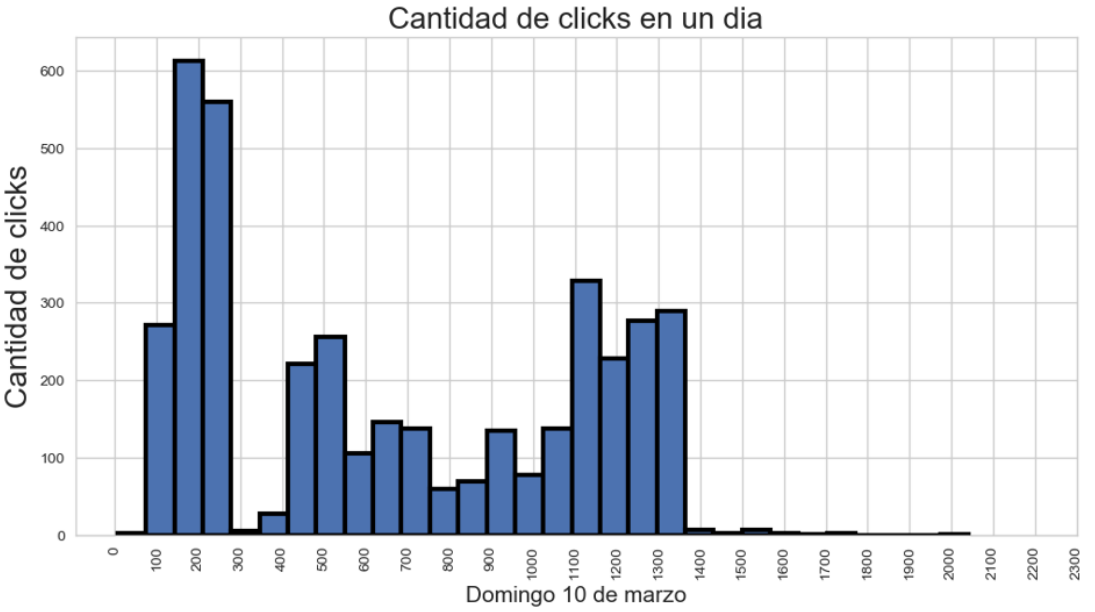
1. Análisis de horarios

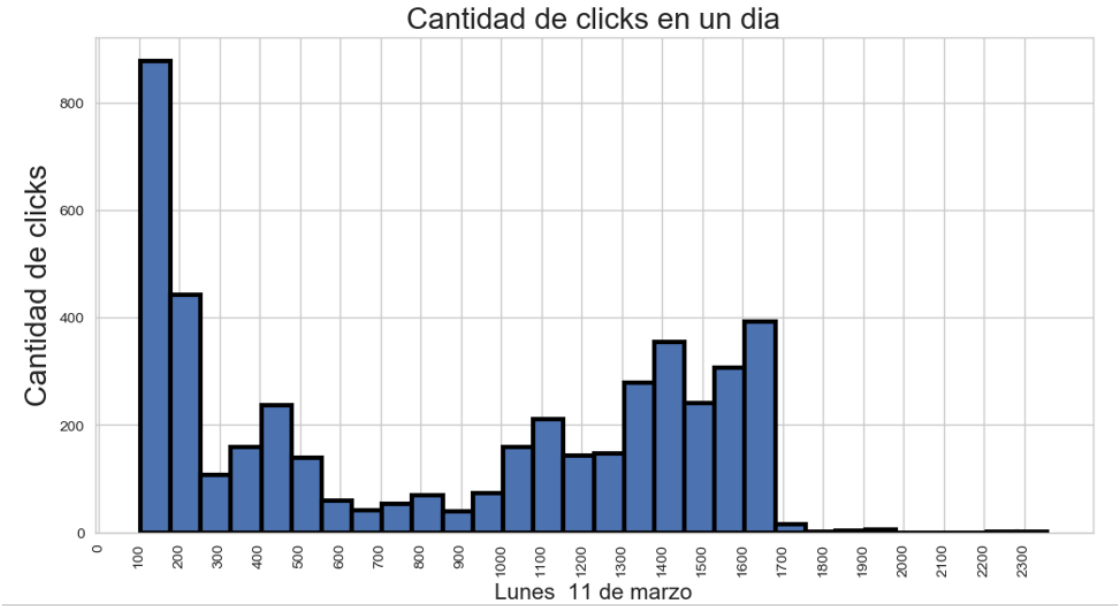
Depurando un poco los datos, puede verse que los datos correspondientes a los primeros días (es decir, el 5 y 6 de marzo), presentan muy pocos valores. Esto puede hacer que las conclusiones no sean representativas, por lo cual no se los tomó en cuenta. Por lo tanto, graficando la cantidad de clicks por día y hora, podemos ver los siguientes gráficos:

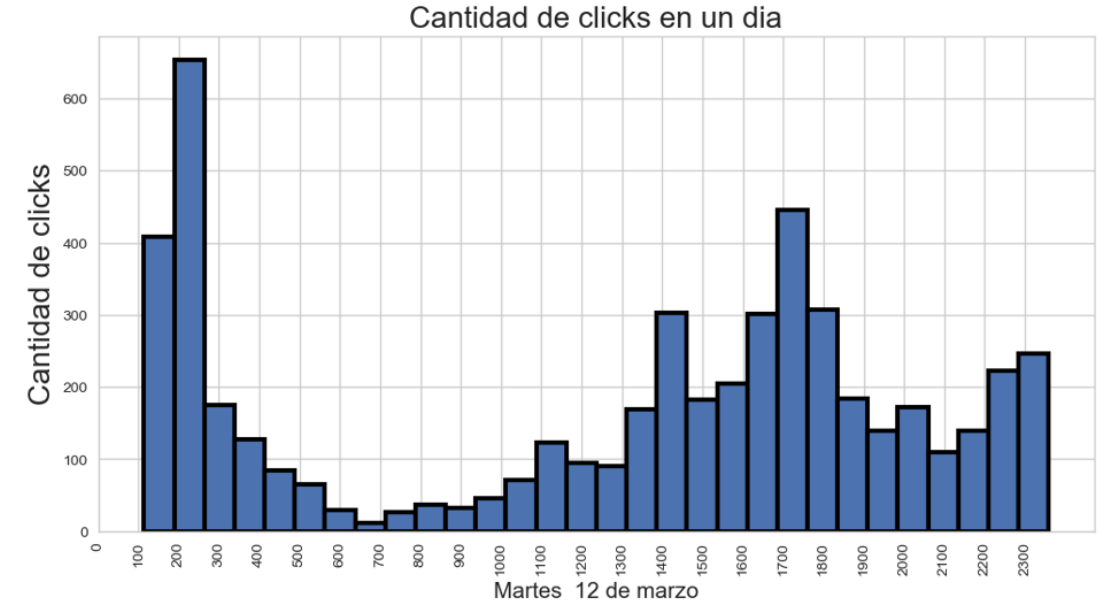


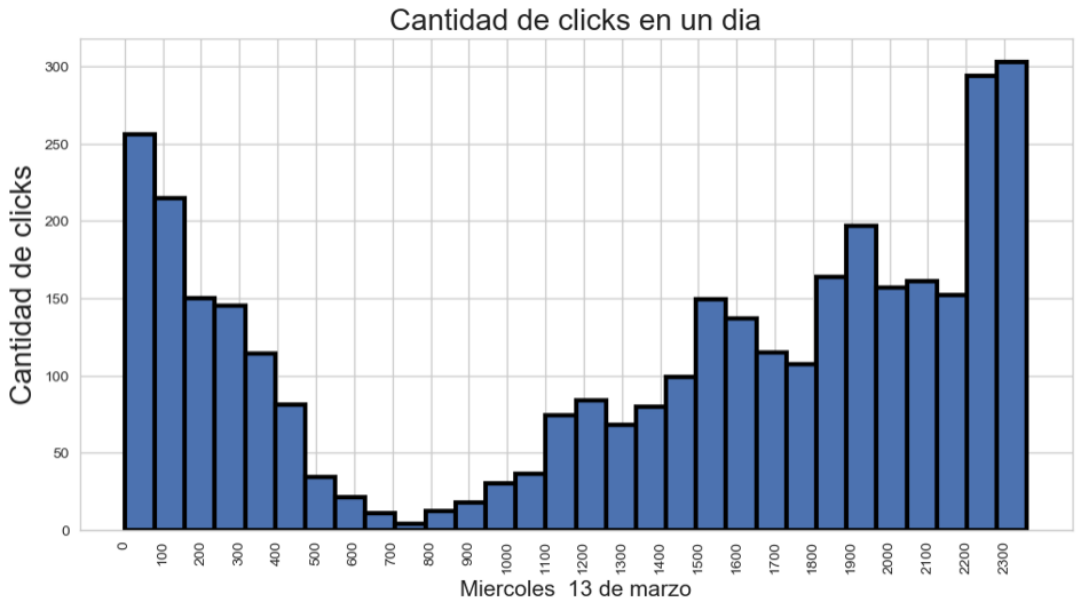



Cantidad de clicks en un dia
Cantidad de clicks
Sabado 9 de marzo


Cantidad de clicks en un dia
Cantidad de clicks
Domingo 10 de marzo

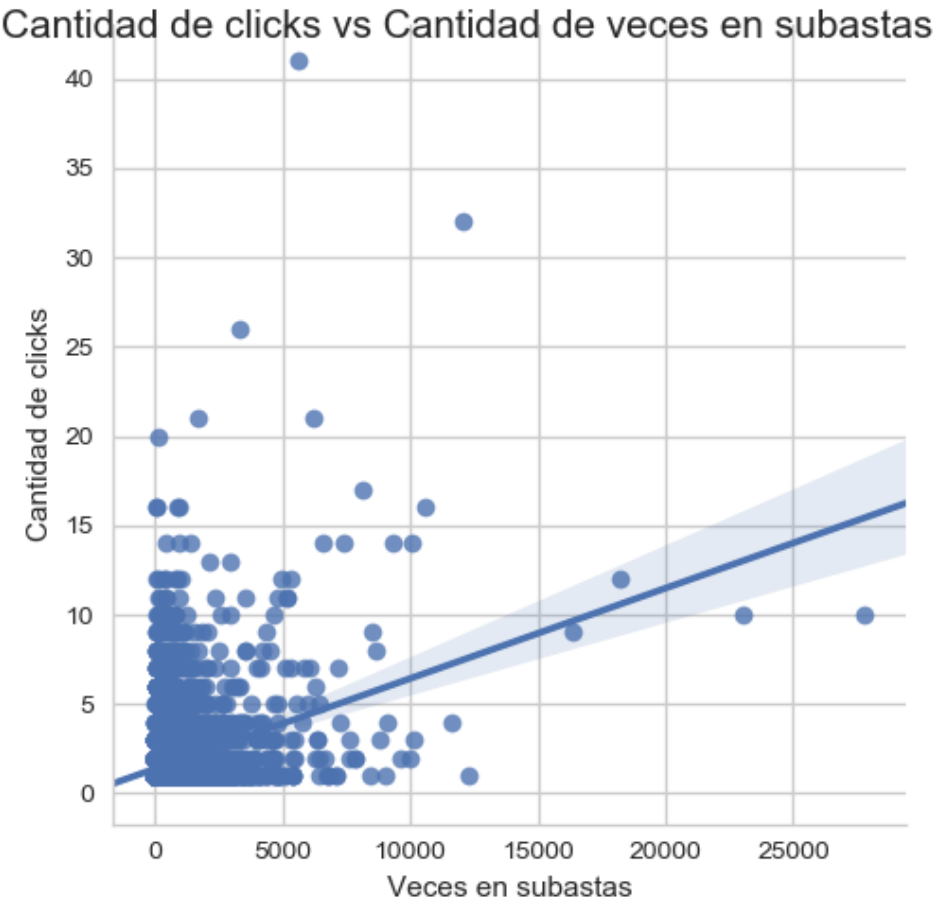

Cantidad de clicks en un dia
Cantidad de clicks
0
200
400
600
800
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900
2000
2100
2200
2300
Lunes 11 de marzo


Cantidad de clicks en un dia
Cantidad de clicks
0
100
200
300
400
500
600
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900
2000
2100
2200
2300
Martes 12 de marzo



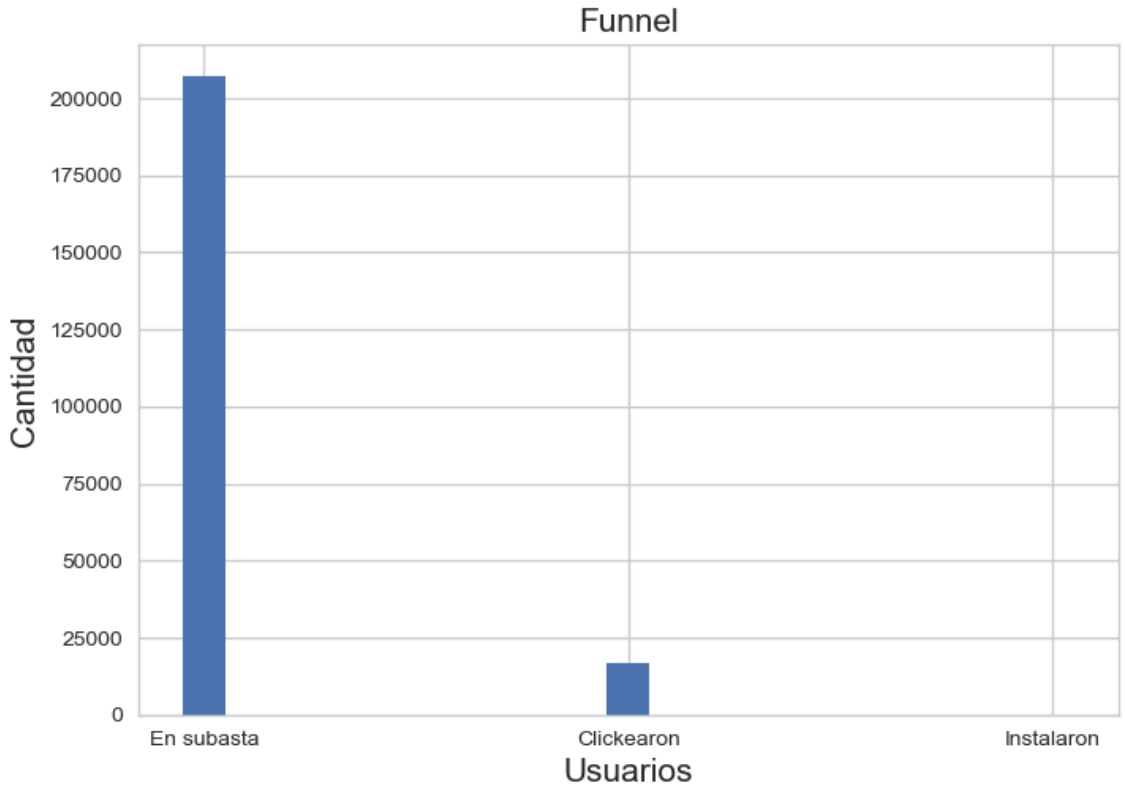
2. Cantidad de clicks vs cantidad de veces en subasta

De los datos provenientes de los sets de subastas y clicks, se agrupó por id de usuario y luego se hizo un inner join para obtener la relación entre la cantidad de veces que un usuario aparece en una subasta, y la cantidad de veces que ese mismo usuario hizo click en la publicidad, obteniendo el siguiente gráfico:

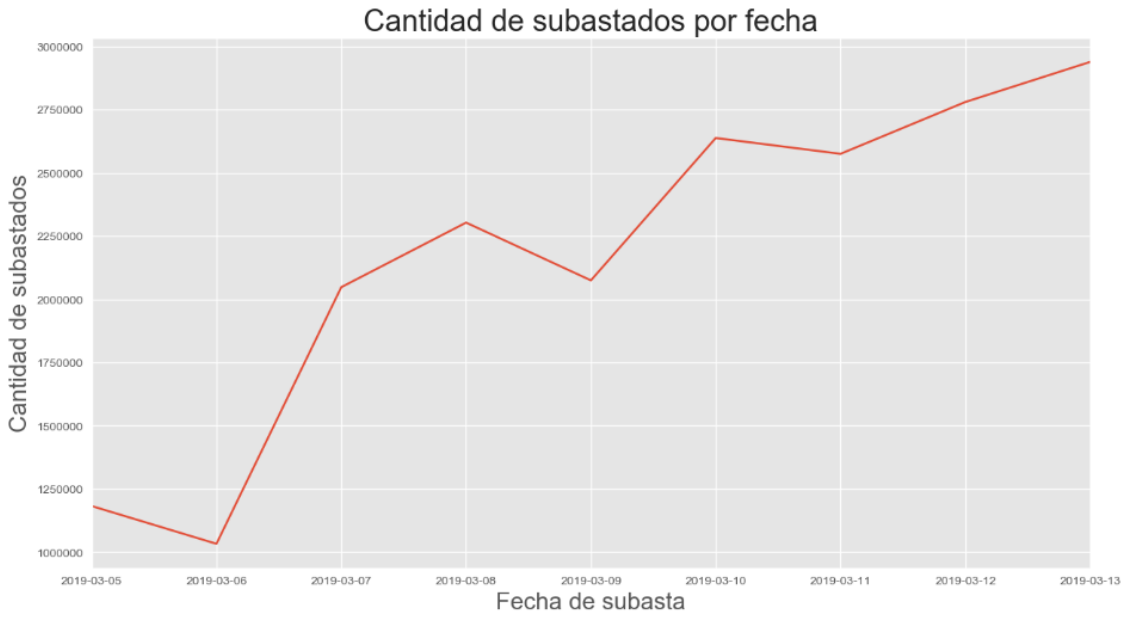


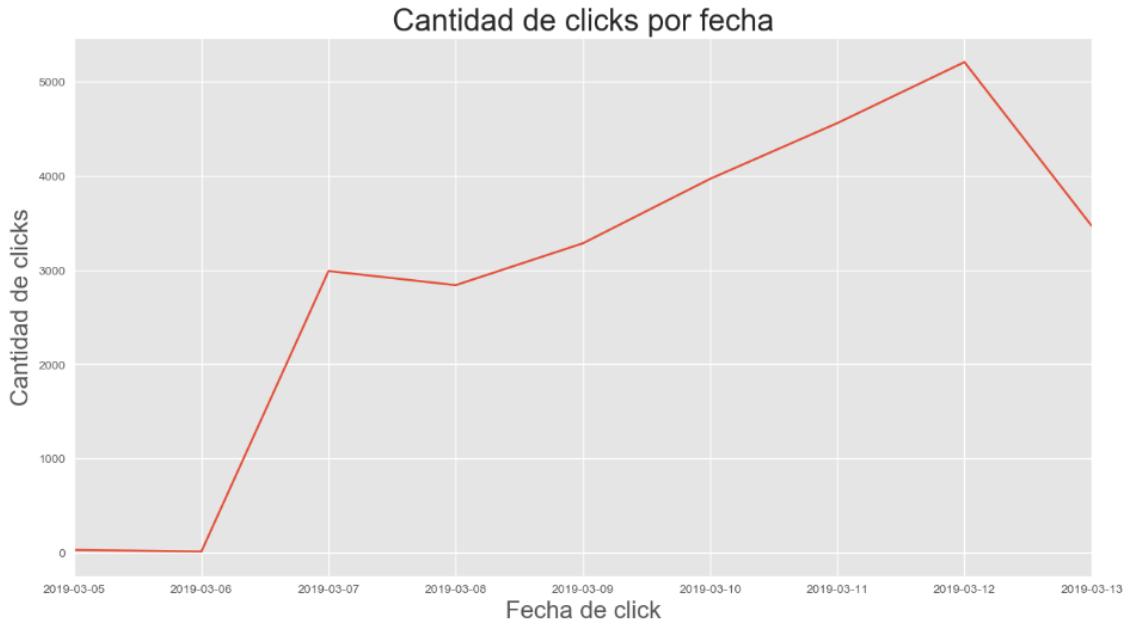
3. Funnel de conversión

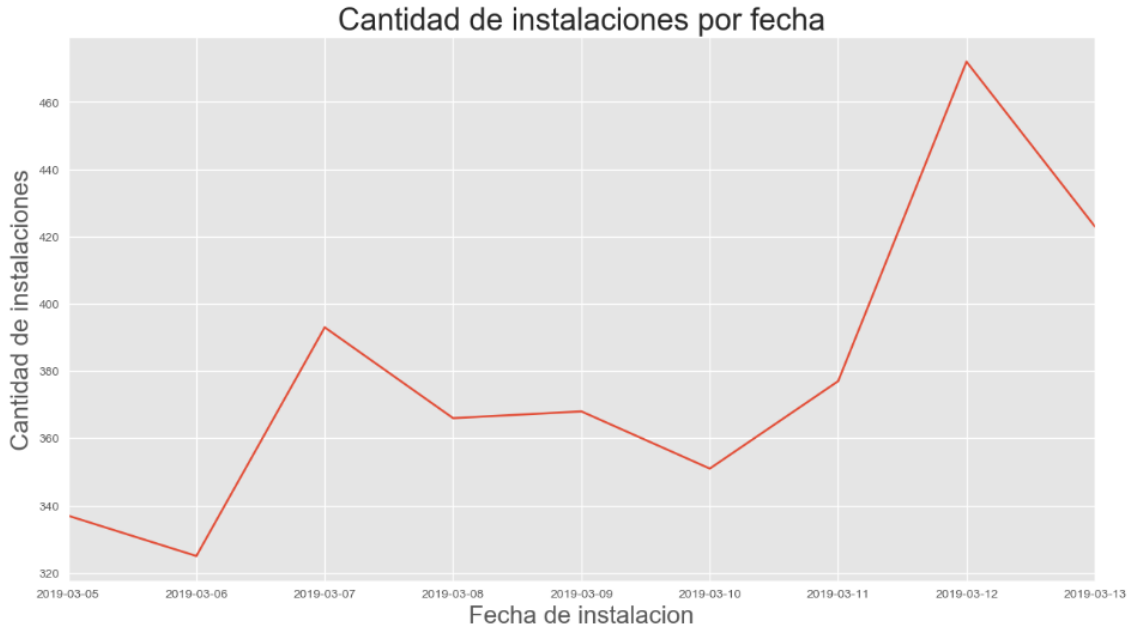
Tomando como objetivo la instalación de una app, se tomaron los siguientes pasos para la conversión de un usuario: el usuario aparece en subasta, el usuario clickea una publicidad, el usuario instala la aplicación. Del total de los usuarios poco más de 200.000 usuarios que entraron en subasta en los días analizados, se encontró la siguiente relación:



Analizando por día cada una de las variables por separado:

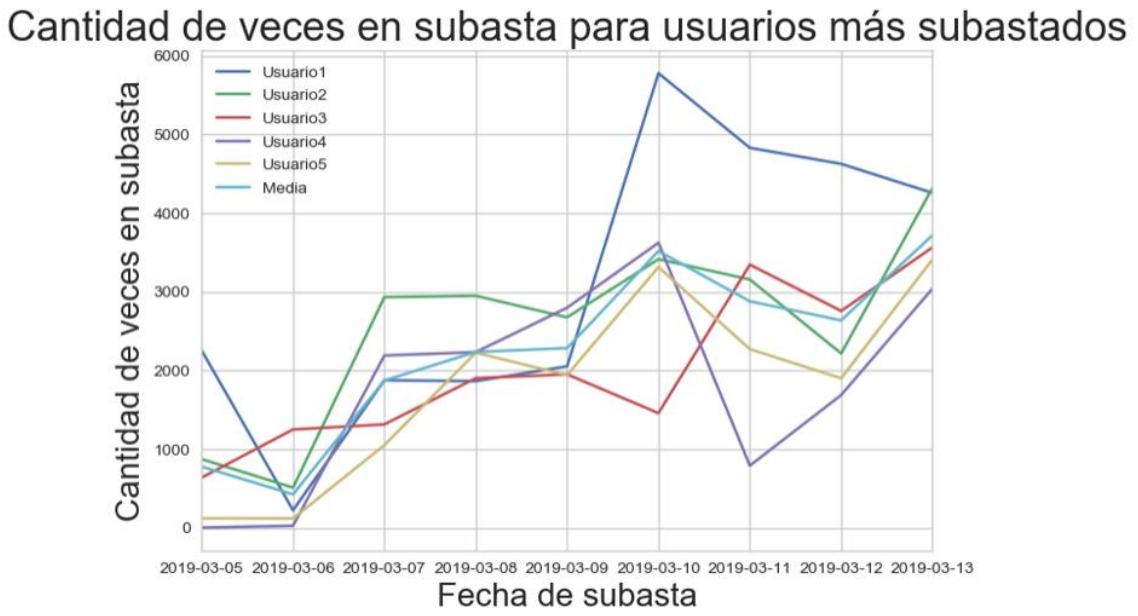






4. Análisis de los usuarios que más aparecen en subasta

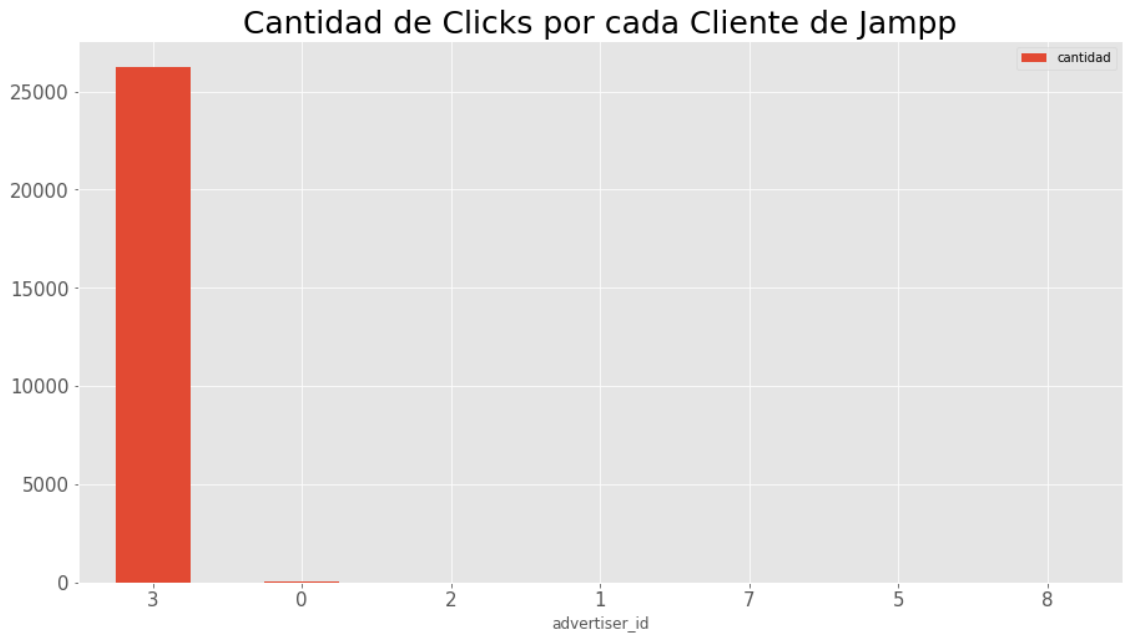
Puede verse en el set de datos que hay usuarios que aparecen muchas más veces que otros usuarios en las subastas. Estos parecen ser usuarios potencialmente activos. Analizando su aparición en subastas por fecha, y dibujándolo en una línea de tiempo:



5. Cantidad de Clicks por Cliente

Se realiza un análisis de la cantidad de clicks que logró obtener cada cliente de Jampp a través de sus publicidades. El advertiser_id es lo que identifica a cada cliente.

advertiser_id	cantidad
3	26263
0	70
2	12
1	2
7	2
5	1
8	1

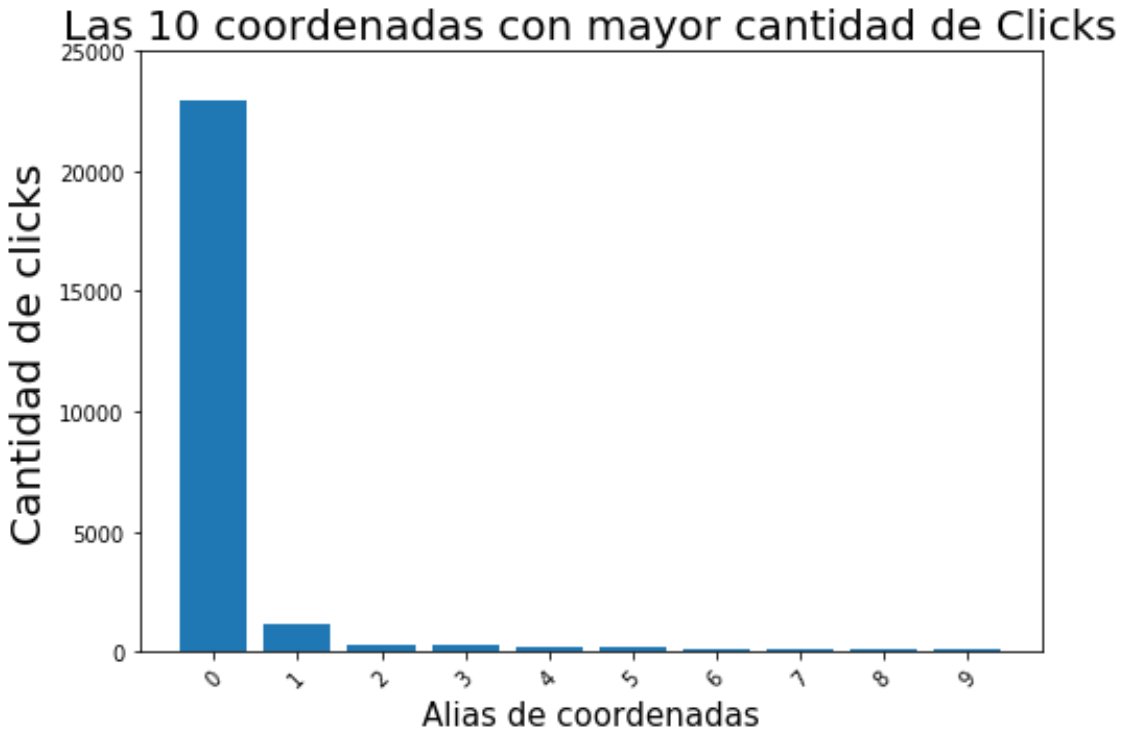


El cliente de Jampp con advertiser_id = 3 fue quien logró obtener mayor cantidad de clicks a través de sus publicidades.

6. Top 10 de coordenadas con más clicks.

Se le atribuye a cada coordenada un alias para poder graficarlo mejor

latitude	longitude	cantidad	alias
1.205689	1.070234	22949	0
1.218924	1.071209	1105	1
1.235406	1.063737	281	2
1.205393	1.077238	250	3
1.208059	1.069624	186	4
1.205058	1.077332	171	5
1.223819	1.059475	100	6
1.206592	1.069958	90	7
1.209520	1.065525	85	8
1.209372	1.067147	83	9



La mayoría de los clicks fueron producidos en las coordenadas 1.205689 y 1.070234

7. Conclusión

De los gráficos puede verse que en general los picos de actividad suelen darse al final y al inicio del día. Es decir, estamos hablando de usuarios en con fuerte actividad noctámbula. También puede verse, como era de esperar, muy baja actividad en los horarios de madrugada.

Además, parecería ser que el hecho de que un usuario aparezca muchas veces en una subasta no implica que vaya a hacer muchos clicks. Aunque el gráfico tenga tendencia alcista, hay que pensar en no saturar al usuario.

En cuanto al proceso de conversión, llama la atención que de los usuarios que entraron a la subasta, aproximadamente el %8 pasó al siguiente paso de clickear en la imagen. Se puede aproximar una pérdida del %90 de los usuarios que entran a la subasta para que hagan click. Esto es lógico, ya que no siempre entramos a todas las publicidades que nos muestran. En cuanto a las instalaciones, fueron 7. Si bien parece infinitamente poco comparado con la cantidad de usuarios que entraron a subastas, tenemos que preguntarnos: ¿Cuántas veces instalamos una aplicacion a partir de una publicidad?

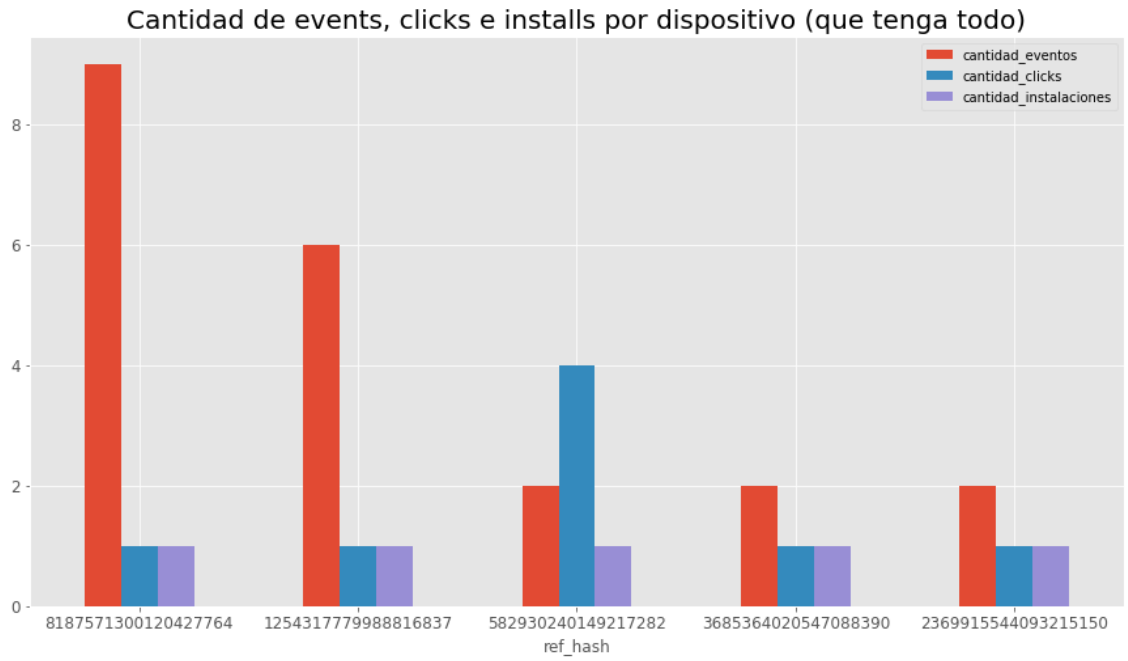
En cuanto a cada proceso por separado, el mejor día para subastas, clicks e instalaciones fue el martes 12 de marzo. También puede verse una conclusión anterior: no por haber muchas subastas tiene que haber muchos clicks o instalaciones. El dia con mayor cantidad de subastas fue el miércoles 13 de marzo, y ese día con respecto al anterior hubo una caída de clicks e instalaciones. Además, el domingo 10 de marzo fue uno de los días con mayor cantidad de subastas, y al mismo tiempo uno de los de menor cantidad de instalaciones.

Finalmente, en cuanto a los usuarios que más aparecieron en subastas, si bien su comportamiento es aleatorio, puede observarse que la media sigue la curva del gráfico de subastas.

3. Análisis Set de Datos Events

1. Análisis de cantidad de eventos con sus respectivos clicks e instalaciones

ref_hash	cantidad_eventos	cantidad_clicks	cantidad_instalaciones
8187571300120427764	9	1	1
1254317779988816837	6	1	1
582930240149217282	2	4	1
3685364020547088390	2	1	1
2369915544093215150	2	1	1



Para este análisis, no se tuvo en cuenta si el evento fue atribuido a Jampp.

2. Conclusión

Como podemos observar, solo hubo 5 casos de dispositivos diferentes que realizaron alguna instalación producida por alguna serie de eventos y clicks.

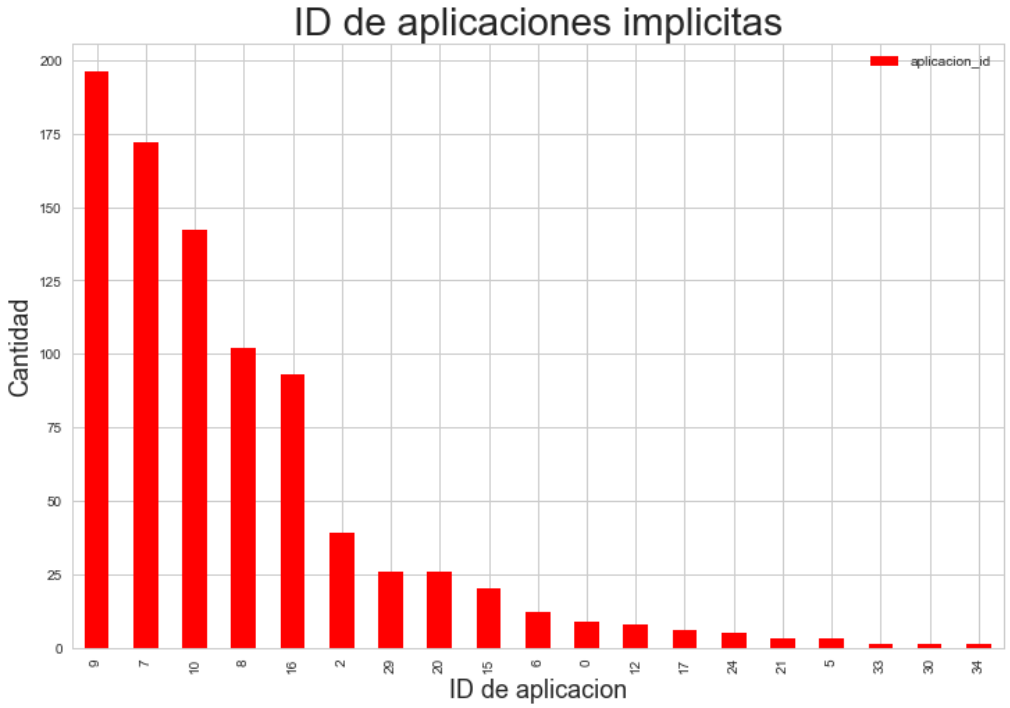
Para el caso de eventos atribuidos por Jampp, no hubo algún caso en el cual se haya producido una instalación (se comenta en el análisis del set de datos Installs)

4. Análisis Set de Datos Installs

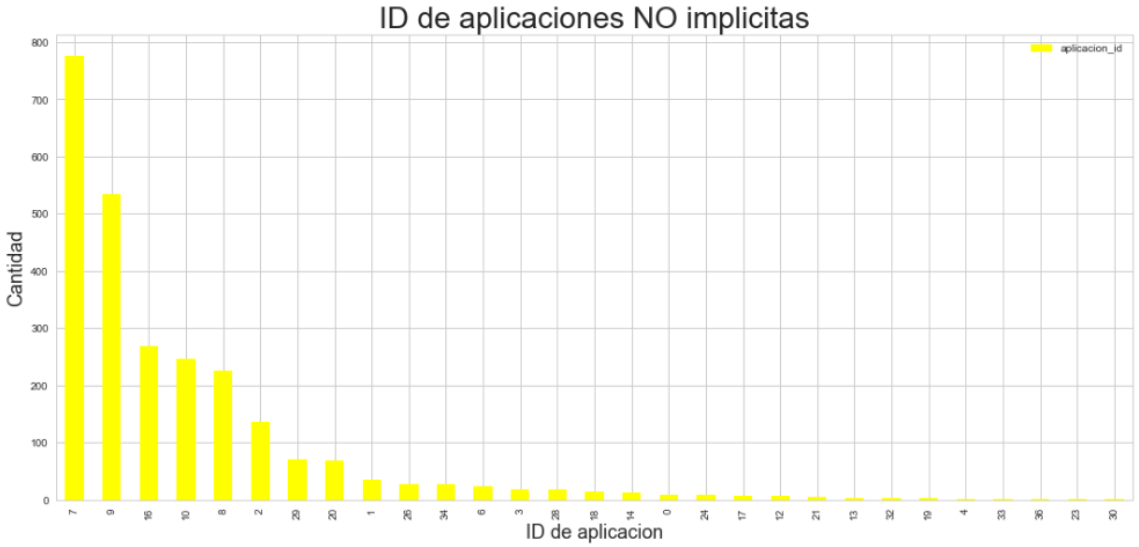
Para comenzar, se carga el archivo con los datos que contiene los datos de las instalaciones de aplicaciones hechas por los usuarios y se estudia que campos tenemos y verificamos las descripciones de cada campo con el archivo de descripciones que fue provisto especialmente para el trabajo práctico.

1. Implicit Install

Lo primero que se realiza es un filtrado de todas las aplicaciones que se instalaron de forma implícita, esto significa que la instalación fue realizada por un dispositivo que no se ha instalado de acuerdo con la plataforma de seguimiento.



Luego, se realiza el mismo análisis para las instalaciones que no fueron de forma implícita.

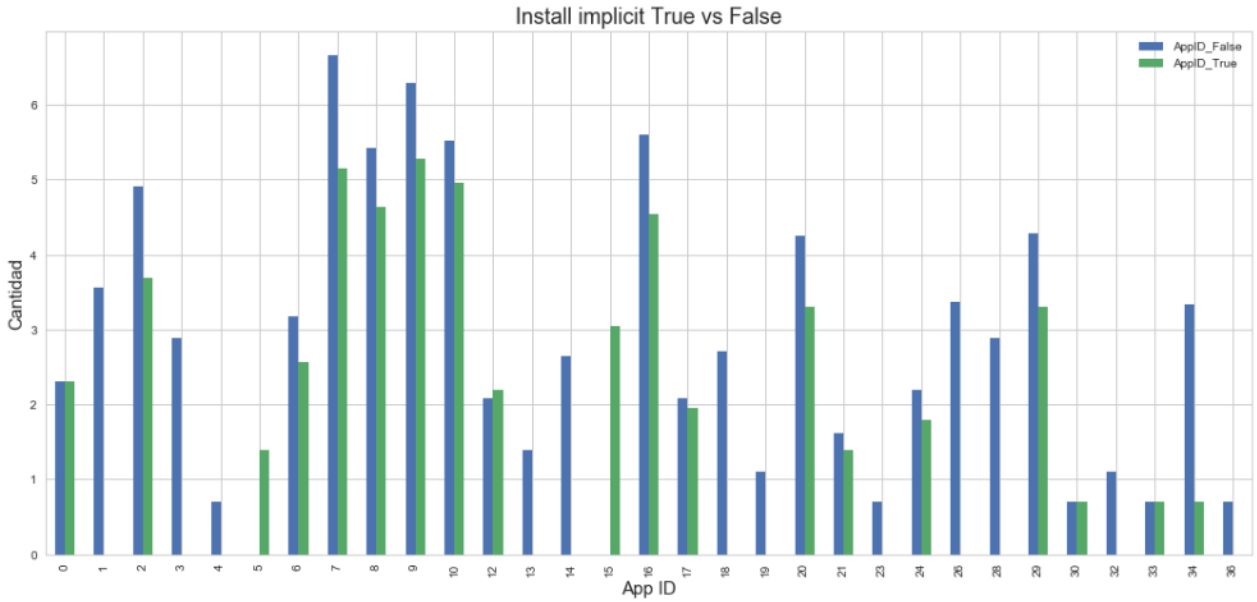


En este segundo gráfico, vemos que hay mayor cantidad de aplicaciones instaladas de forma no implícita, hay 29 aplicaciones instaladas sobre un total de 30 estudiadas en el set de datos.

Haciendo una comparación del total de las aplicaciones que tanto las aplicaciones n°7 y n°9 son las aplicaciones más instaladas en ambas formas.

	AppID_False	AppID_True
0	2.30	2.30
1	3.56	nan
2	4.91	3.69
3	2.89	nan
4	0.69	nan
5	nan	1.39
6	3.18	2.56
7	6.65	5.15
8	5.42	4.63
9	6.28	5.28
10	5.51	4.96
12	2.08	2.20
13	1.39	nan
14	2.64	nan
15	nan	3.04
16	5.60	4.54

	AppID_False	AppID_True
17	2.08	1.95
18	2.71	nan
19	1.10	nan
20	4.25	3.30
21	1.61	1.39
23	0.69	nan
24	2.20	1.79
26	3.37	nan
28	2.89	nan
29	4.28	3.30
30	0.69	0.69
32	1.10	nan
33	0.69	0.69
34	3.33	0.69
36	0.69	nan

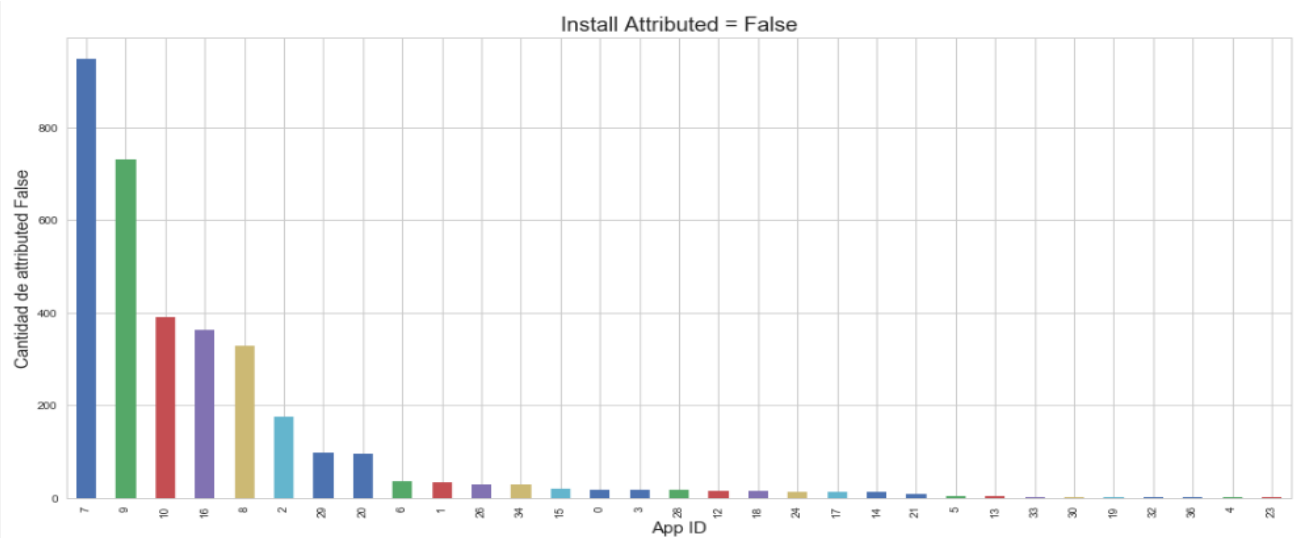


Hubo más instalaciones de más aplicaciones de forma no implícita, por no ser dispositivos que hicieron las instalaciones pero no por la plataforma de seguimiento, o sea por las publicidades que aparecen y el usuario le hace click, sino que lo hace por otra vía.

2. Atributed Install

Las instalaciones hechas pueden ser atribuidas a Jampp o no.

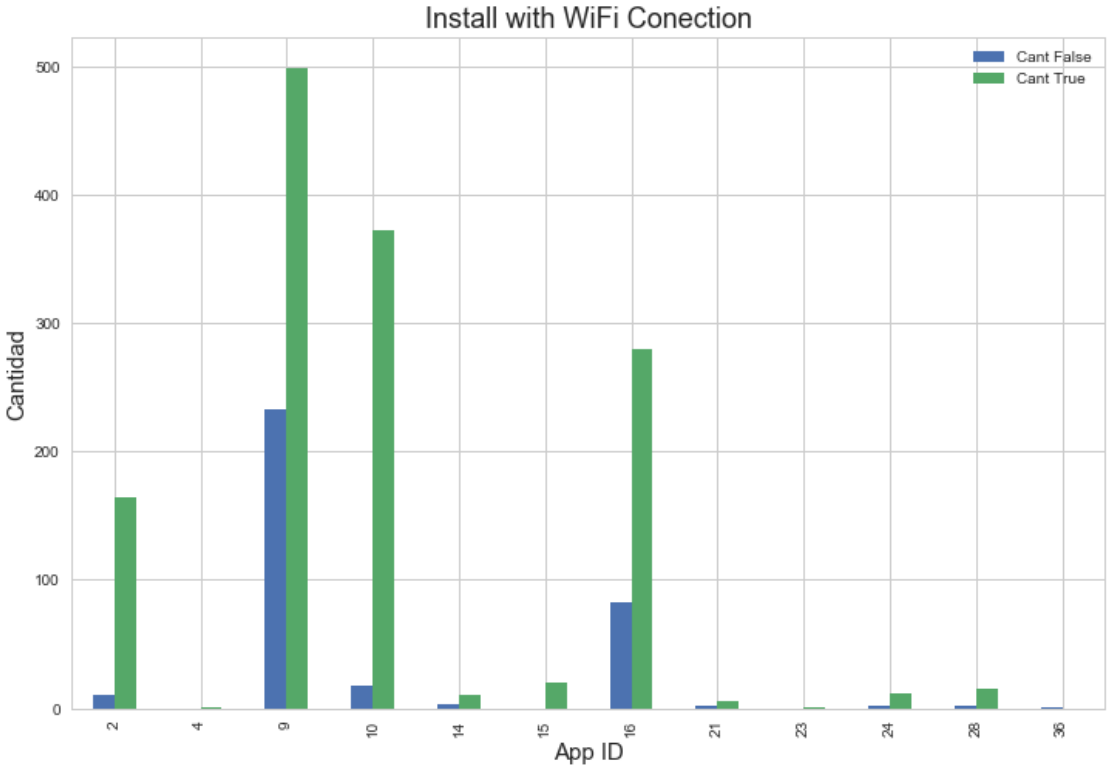
Al realizar el estudio de los datos, se verifica que ninguna instalación fue atribuida a Jampp, por los que todas las instalaciones hechas de todas las aplicaciones fueron hechas por fuera de las plataformas de Jampp.



El gráfico está ordenado de mayor cantidad a menor cantidad de instalaciones sin atribuir a Jampp y seguimos viendo como la aplicación 7 y 9 son las que lideran las instalaciones.

3. Install with WiFi

Para este análisis, lo que tenemos es varios datos de si la instalación se hizo en dispositivos con conexión de WiFi o no.



Como se puede apreciar en el gráfico, hay muy pocos datos explícitos de las instalaciones con dispositivos que permiten acceder a la información dela conexión de WiFi, tanto si está conectado o no a la red.

De igual manera, en todos los casos la cantidad de instalaciones hechas con WiFi activado son mayores que las que no, por lo que el usuario aprovecha para descargar e instalar aplicaciones cuando está conectado a una red inalámbrica para no hacer uso de sus propios datos móviles, y poder hacer uso de los mismos en su rutina diaria y también poder reducir sus gastos.

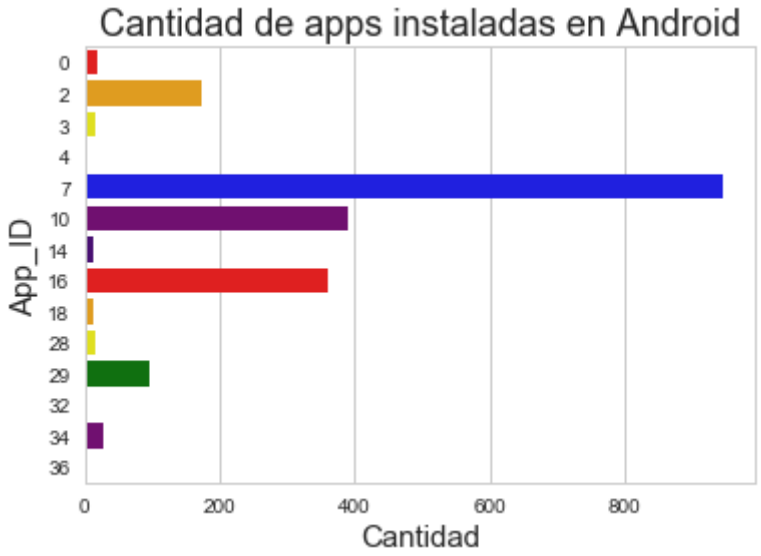
4. Instalación por Sistemas Operativos

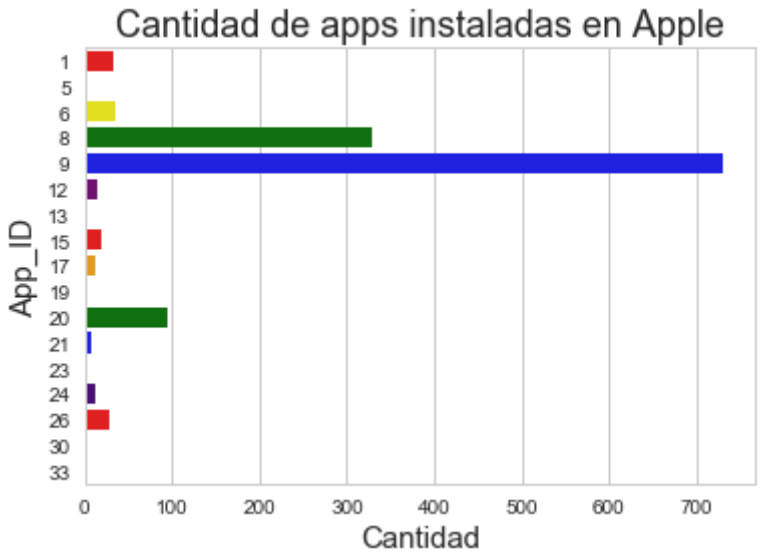
En la columna del set de datos “ref_Type” donde hay un número que puede ser tanto Apple_ifa o google_advertising_id según el archivo de descripciones.

Al juntar y contar los valores, se observa que solo hay dos tipos de “ref_type” por lo que se supone que el de mayor valor es el de google_advertising_id por lo que son dispositivos con sistema operativo Android y el otro tipo corresponde a Apple_ifa entonces deducimos que son dispositivos con iOS como sistema operativo.

```
Android    2080
Apple      1332
```

Del total de instalaciones en el período estudiado (9 días) se notan que las instalaciones en dispositivos Android supera ampliamente a las de iOS.



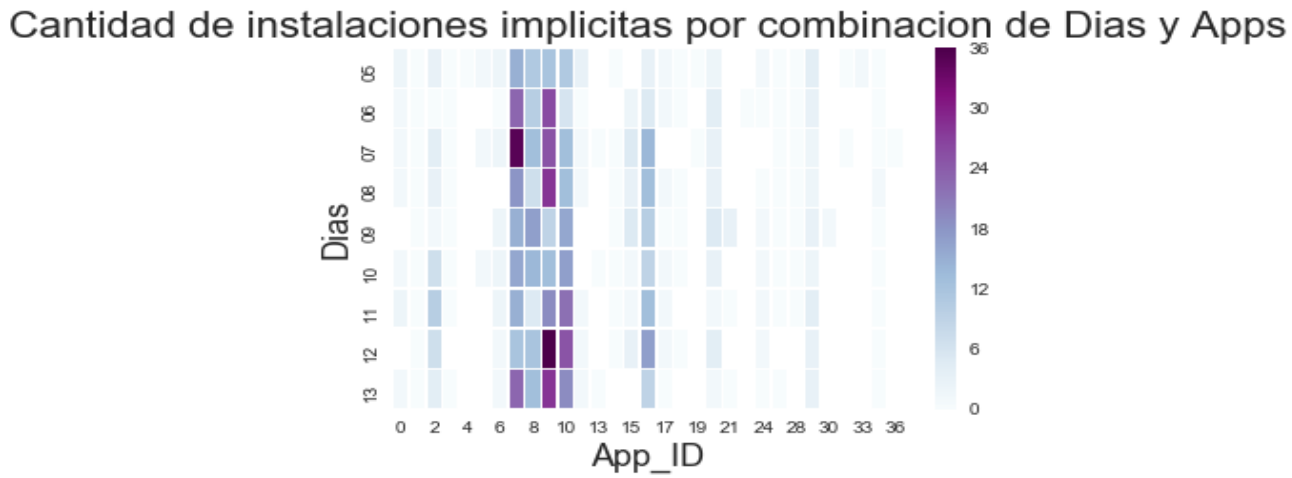


Lo que se observa de estos gráficos, es que las dos aplicaciones más instaladas, la n°7 y la n°9, son instaladas solamente en dispositivos con Android y con Apple respectivamente. Lo mismo sucede con el resto de las aplicaciones, solo son instaladas en cada dispositivo y nada más.

5. Instalaciones por día

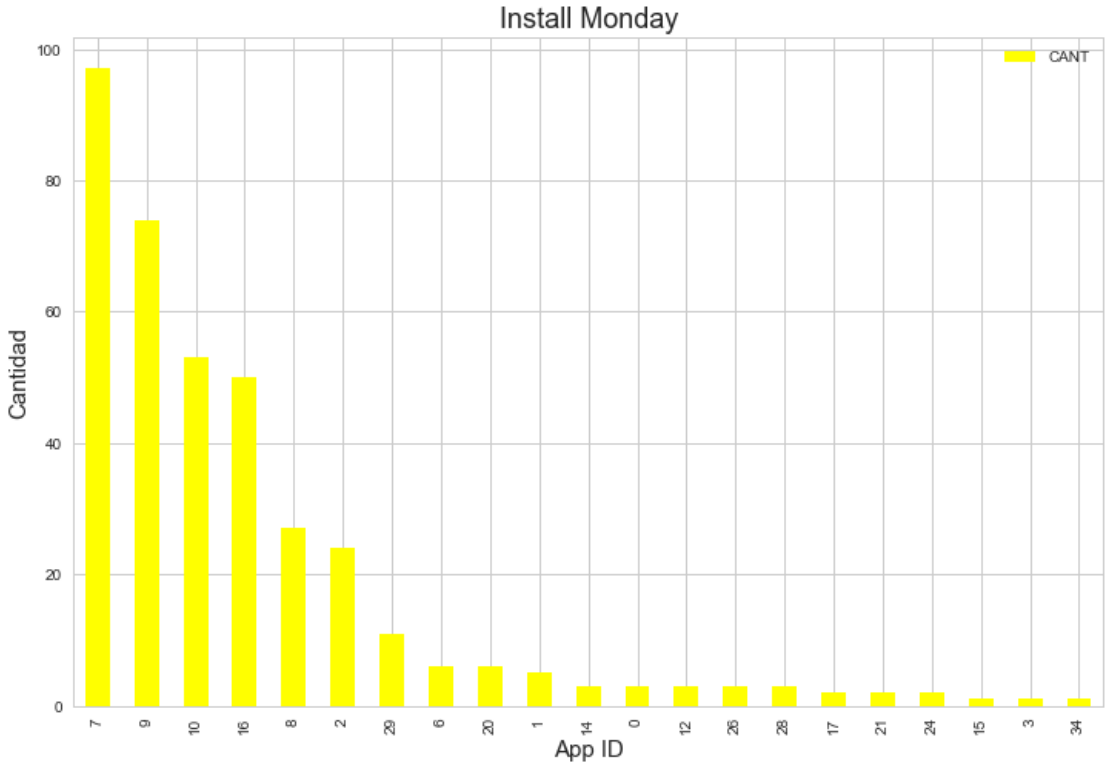
Todas las instalaciones vienen en el set de datos con un campo de creación. Ese campo viene en el fomato de aaaa-mm-dd. Por lo que se separa la fecha y se divide en nuevas columnas donde se guardan solamente el año, el mes y el día por separado.

Al tener estos datos por separado y sabiendo que el año y el mes no cambian en todos los datos, solo tomo los días para analizar.



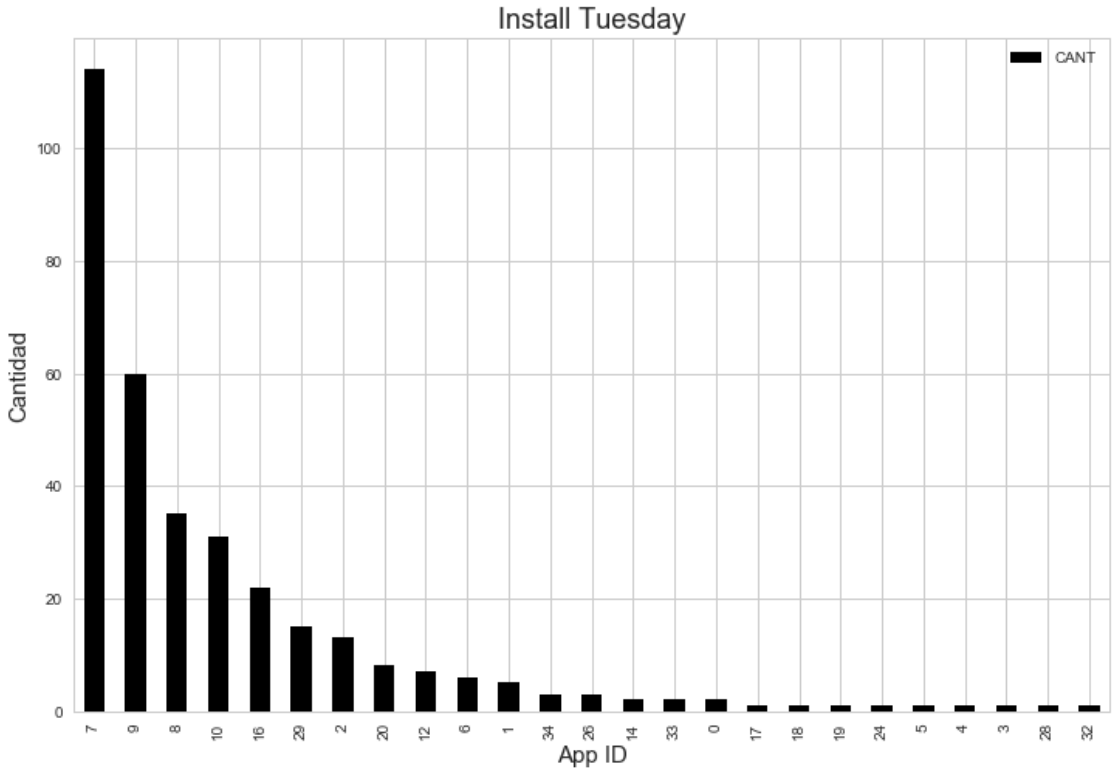
Viendo este HeatMap se observa que de las dos aplicaciones más instaladas, la n° 7 es instalada con mayor fuerza en el día 7 del mes y la n° 9 el día 12 del mes. Luego, el resto de las aplicaciones se van instalando durante todo el ciclo estudiado.

a. Lunes



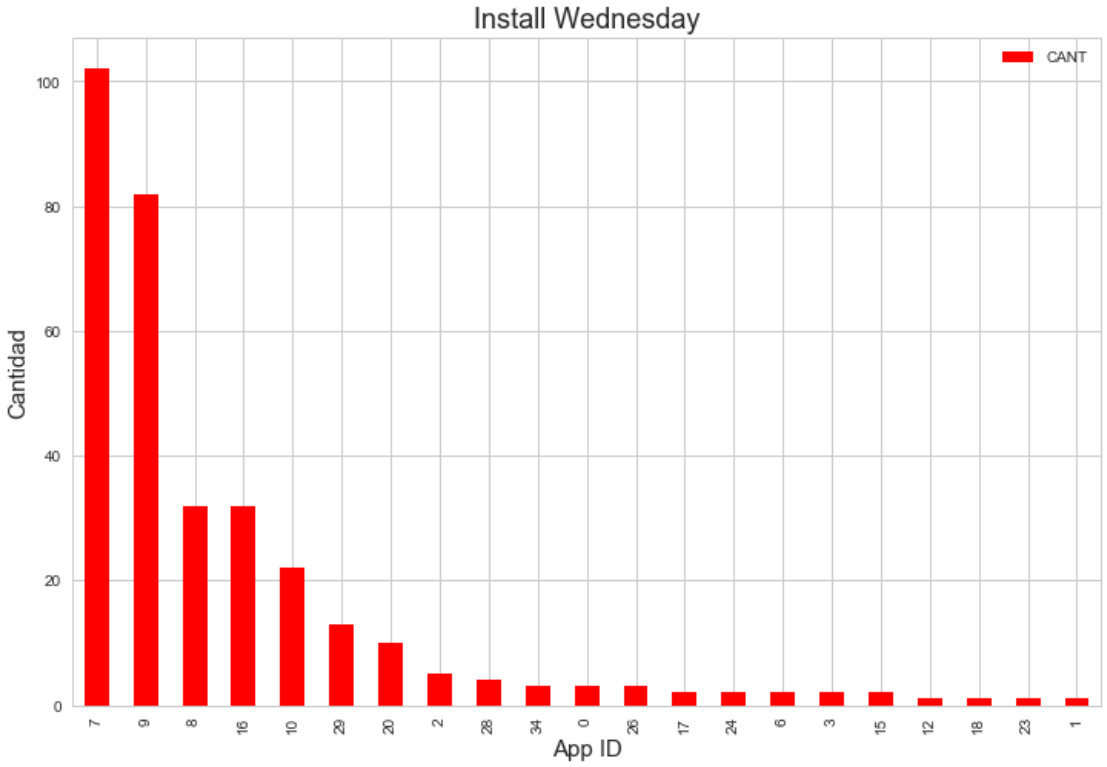
b. Martes

Al ser el ciclo de 9 días, hay dos martes dentro de del período, el 5 y 12 de Marzo del 2019. Por lo que al filtrar los días por separado, luego los junto para analizar en conjunto los martes.

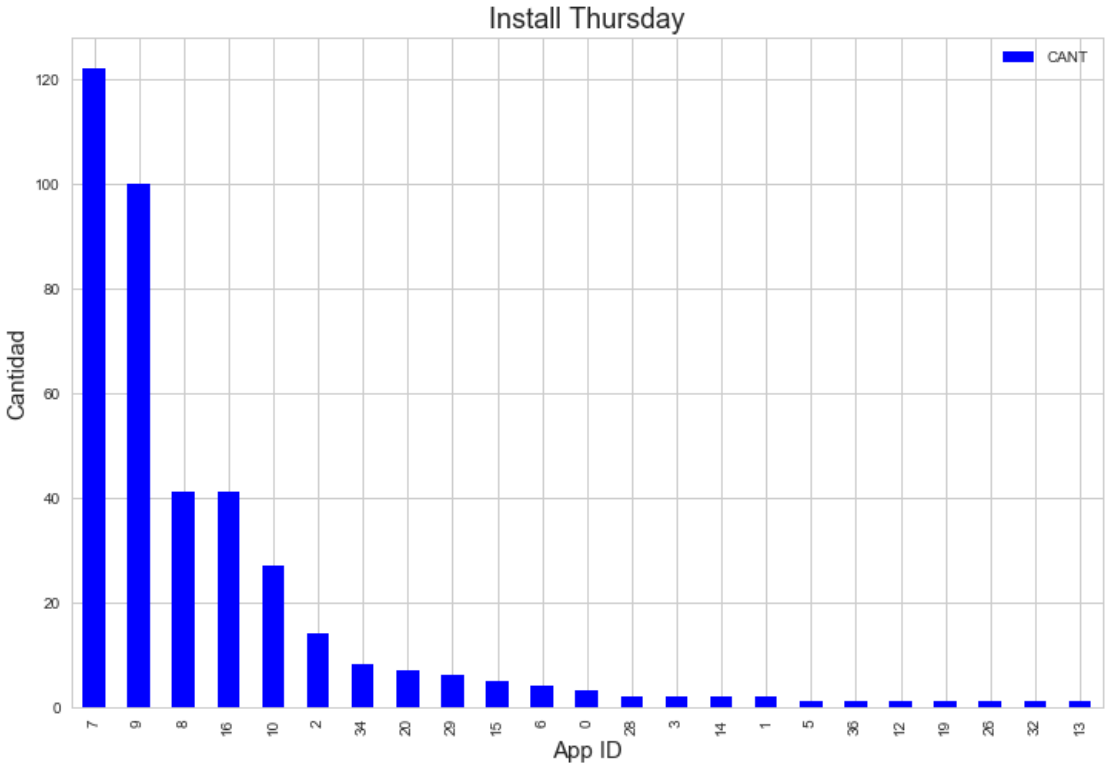


c. Miércoles

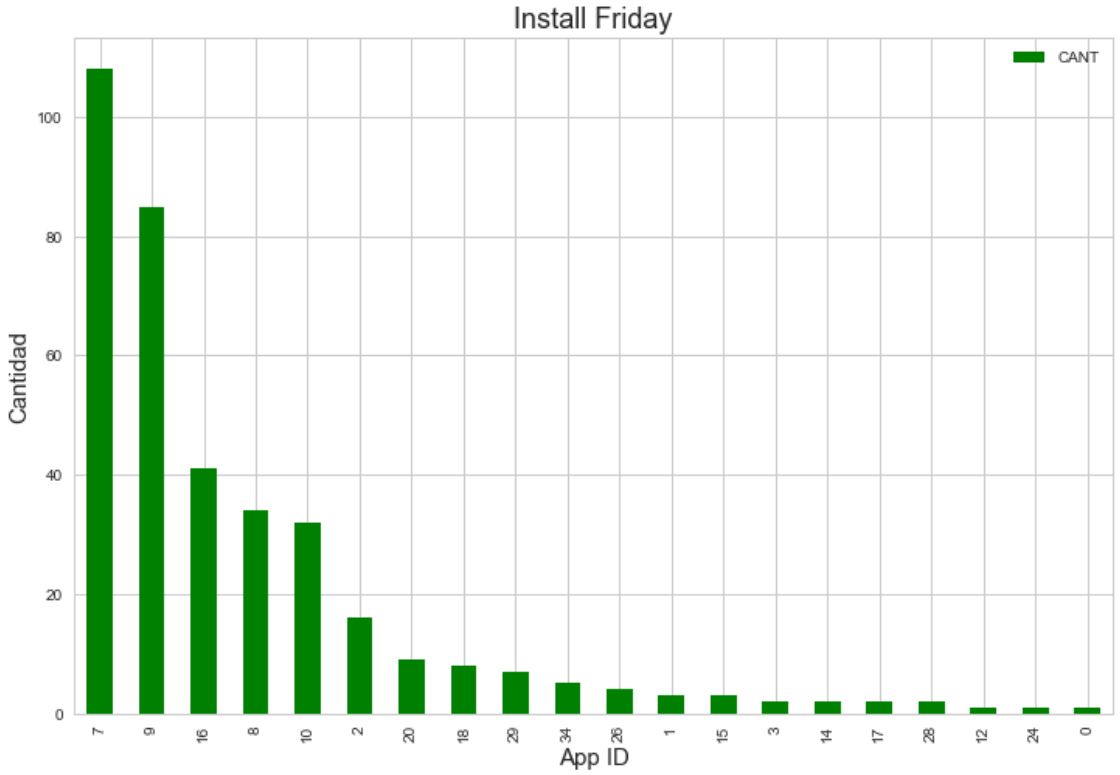
Al igual que el martes, es el otro día que se repite dentro del período a analizar. Se actuó de la misma manera que con el día martes.



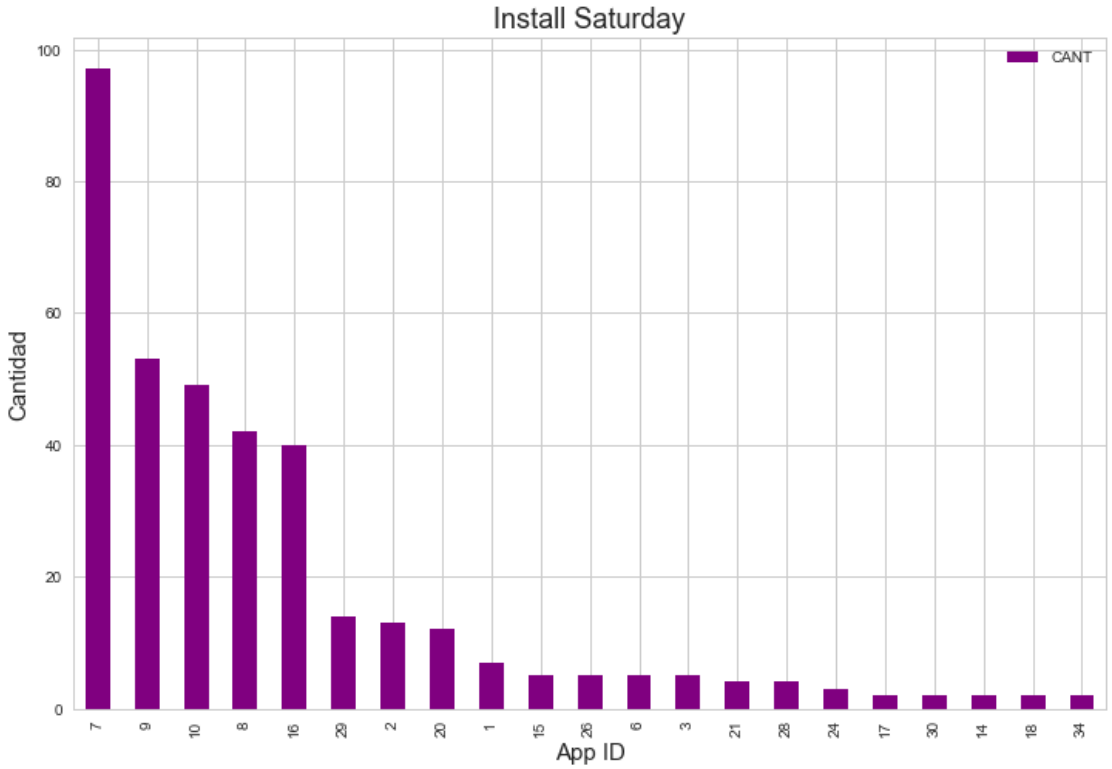
d. Jueves



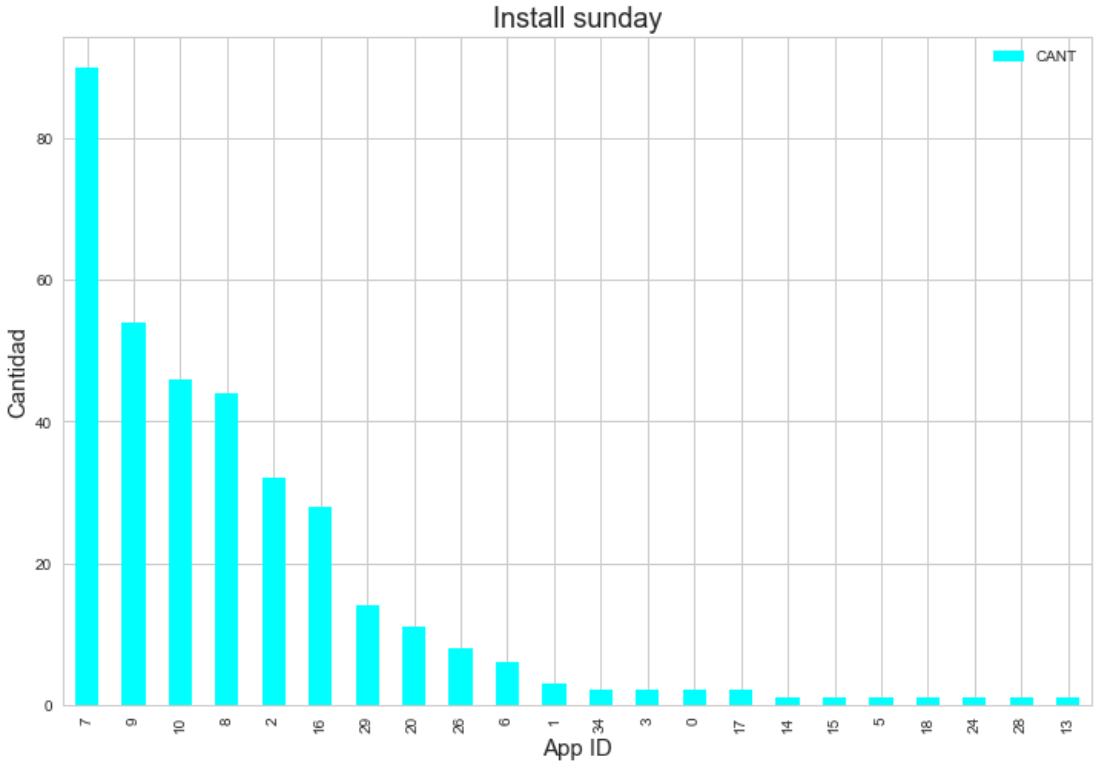
e. Viernes



f. Sábado



g. Domingo



Según los gráficos de cada día, los jueves es cuando hay mayor cantidad de instalaciones y en el fin de semana, sábado y domingo, es cuando hay menos cantidad. Esto puede ser porque baja la actividad en el uso del celular entonces hay menos instalaciones.

6. Instalaciones por cliente

	advertiser_id	application_id	ref_hash
0	3	7	582930240149217282
1	3	8	1254317779988816837
2	3	7	2369915544093215150
3	3	7	3685364020547088390
4	3	7	7190737170444985036
5	3	7	7759178785240189555
6	3	7	8187571300120427764

Solamente el advertise_id 3 logró obtener 7 instalaciones correspondientes a 2 aplicaciones (7 y 8).

advertiser_id	application_id	cantidad_instalaciones
3	7	6
3	8	1

De las 7 instalaciones que tuvo el advertiser_id 3, 6 de ellas corresponden al application_id 7 y 1 al application_id 8.

advertiser_id	application_id	cantidad_instalaciones	cantidad de clicks
3	7	6	26263
3	8	1	

7. Conclusión

A la hora de las instalaciones hechas por los dispositivos, las no implícitas superan a las que sí lo son, esto puede ser porque el usuario llega a la aplicación por otro medio que los eventos que aparecen en un juego u otro lado. Esta puede ser por publicidad gráfica, la televisión o la recomendación de otra persona. Lo importante para el cliente es que la aplicación sea instalada. A su vez, ninguna es atribuida a Jampp, por lo que habría que mejorar las formas para que sea rentable el trabajo.

Es común que la mayoría sean en dispositivos con Android, ya que son muchas las marcas que eligen el sistema operativo de Google, en cambio solo los productos Apple tienen a iOS como sistema operativo, y son muchos menos comparados con la gran variedad del resto.

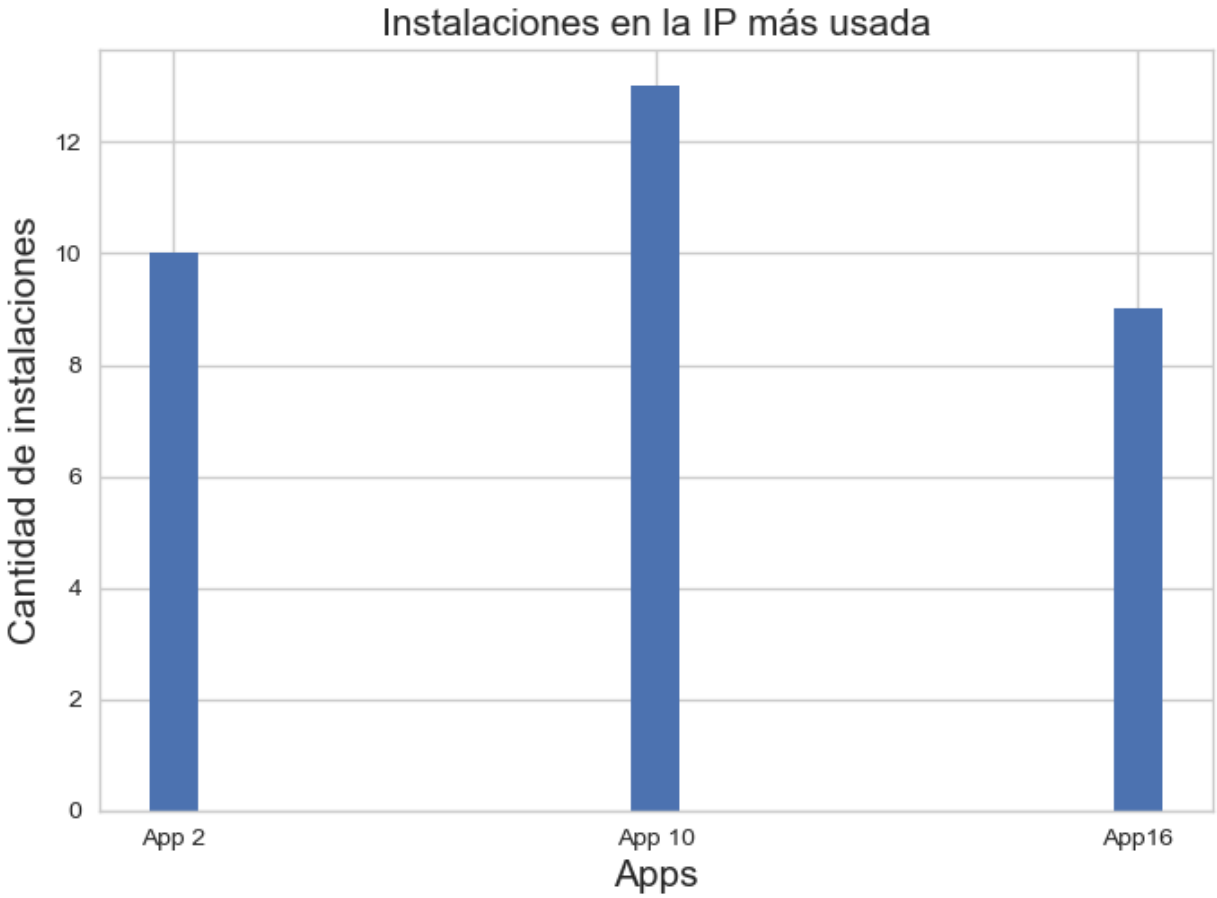
Los días jueves son los de mayor descarga, puede ser porque es un día próximo al fin de semana donde el usuario común disminuye el uso de los dispositivos.

Finalmente, el cliente de Jampp con advertiser_id = 3 fue el único que logró obtener 7 instalaciones correspondientes a 2 aplicaciones (7 y 8) de los 26.263 clicks que logró en sus publicidades.

5. Análisis sospechoso

Analizando los usuarios que más instalaron apps, se encontró que algunas aplicaciones aparecen duplicadas con segundos (o incluso microsegundos) de diferencia. Pero llama la atención el usuario con device id 2515049144505739996, quien instaló la aplicación 9 tres veces en días distintos. ¿Por qué alguien instalaría y desinstalaría la aplicación en días distintos?

Además, analizando por dirección IP, se observa que la que presenta mayor cantidad de instalaciones, repite en gran parte tres aplicaciones: la 2, la 10 y la 16.



1. Conclusión

En la industria parece haber un fraude por redireccionamiento de ID de usuario. Resulta sospechoso que se repitan tantas veces la misma app con tan poco tiempo de diferencia bajo una misma dirección IP.

6. Conclusión Final

- Mayor instalaciones NO implícitamente hechas
- Mayor cantidad de instalaciones en dispositivo Android, porque tiene más variedad de dispositivos disponibles con ese sistema operativo
- Los fin de semana disminuye la cantidad de instalaciones
- NINGUNA instalación se le atribuye a Jampp
- De los datos que tenemos, predominan las instalaciones hechas con el WiFi conectado, esto puede ser porque el usuario prefiere no gastar sus propios datos móviles para hacerla descarga y la instalación de las aplicaciones.
- Baja actividad en horarios de madrugada.
- No encontramos relación que demuestre que a mayor cantidad de publicidad a un usuario, mayores instalaciones.
- Creemos conveniente destinar tiempo al análisis de las cantidades de clicks o instalaciones para poder evitar fraudes.