UNIVERSIDAD DE BUENOS AIRES

FACULTAD DE INGENIERÍA

75.06 Organización de Datos

Trabajo Práctico 1
Primer Cuatrimestre de 2019

Grupo 24

Mansilla, Rodrigo
Rodriguez, Yanet

Link de GitHub: https://github.com/yanetrodriguez25/TpJampp

Introducción

El objetivo de este informe es mostrar el análisis exploratorio realizado sobre los datos provistos por la empresa Jampp, la cual tiene el objetivo de captar usuarios e incentivarlos a comprar o instalar ciertas aplicaciones, a través de publicidades expuestas dentro de otras aplicaciones.

Jampp nos provee cuatro sets de datos:

- Clicks.Csv: Contiene información sobre los clicks realizados en cada dispositivo donde se muestra una publicidad de un cliente de Jampp.
- Events.Csv: Cualquier evento realizado en el dispositivo.
- Installs.Csv: Instalaciones de aplicaciones, pudiendo ser o no incentivadas por Jampp.
- Auctions.Csv: Brinda información sobre las subastas.

El principal fin del informe es observar si existe alguna tendencia en los datos que fluya a través de los archivos y ver si hay información errónea o que no hace ningún aporte al análisis para poder brindarle a Jampp insights que pudieran llegar a proporcionarle información útil y generar estadísticas o conclusiones un poco más precisas sobre los resultados.

Herramientas utilizadas

Herramientas usadas para el análisis

Para realizar todo el procesamiento de los sets de datos provistos, se utilizó Python 3, junto con las librerías Pandas y Numpy.
Luego para la parte de visualización y representación de los datos, se utilizaron las librerías Matplotlib y Seaborn.
Por último, se usó Jupyter Notebook como entorno de trabajo.

Repositorio de GitHub

El link al repositorio de github es el siguiente:

https://github.com/yanetrodriguez25/TpJampp

Allí se encontrarán todos los archivos que conforman el informe completo, incluyendo los notebooks y sets de datos.

Procesamiento de Datos

Dataset Clicks

Introducción General

Este set de datos nos provee información acerca de los clicks realizados en los dispositivos donde se ha realizado una publicación de una aplicación de un cliente de Jampp. Realizando un primer análisis de los campos de este archivo, podemos ver lo siguiente:

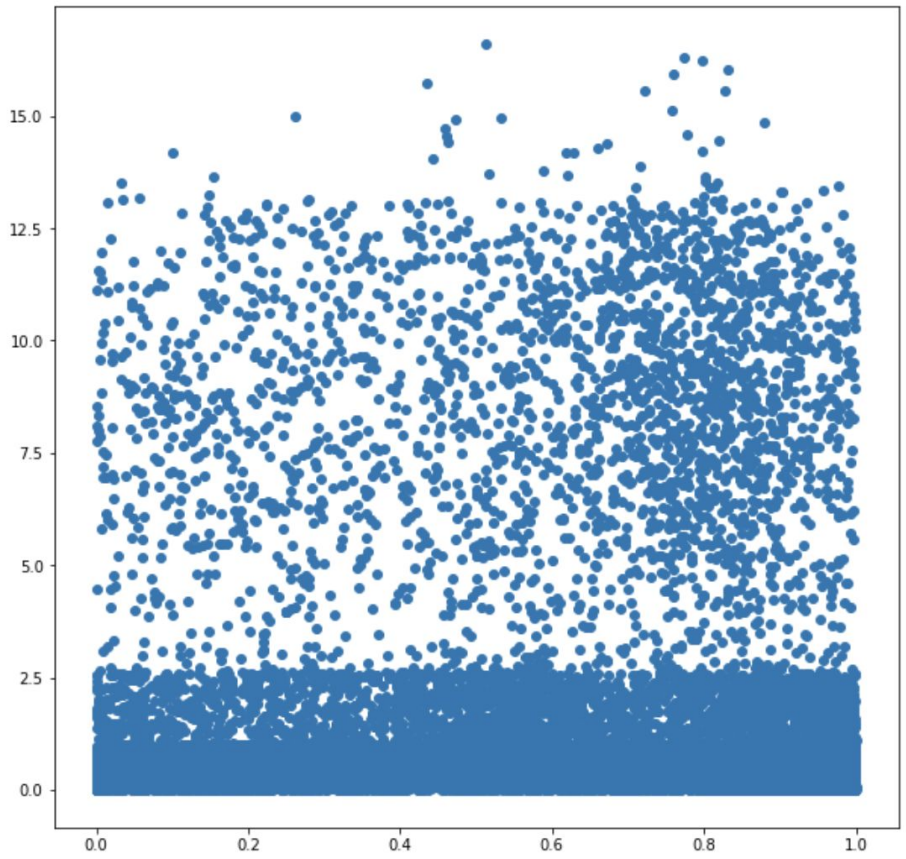
- Contamos con 26351 registros y 20 columnas.
- action_id tiene todos los registros nulos, por lo que no nos aporta información para analizar.
- country_code tiene solamente el código de un país, por lo que tampoco nos aporta algo en particular.
- brand tiene un 76% de valores nulos. No obstante, nos va a permitir ver una tendencia basado en las marcas por lo que lo utilizaremos en el análisis.
- timeToClick tiene un 12,8% de valores nulos. Asumimos que estos valores nulos pueden deberse a problemas en la inserción del registro o bien problemas del sistema, por lo que no los tendremos en consideración para el análisis.
- Tanto touchX como touchY tienen un 12,7% de valores nulos. Vamos a asumir lo mismo que en el punto anterior, y también los transformaremos en coordenadas 0,0.

Distribución de clicks en la pantalla

Utilizamos un gráfico de dispersión para ver la distribución de los clicks, utilizando los datos provistos por los campos touchX y touchY.

En este caso observamos que la mayoría de los clicks se distribuyen en dos grupos:

- X=(0, 1) e Y=(0, 2.5)
- X=(0.7, 0.9) e Y=(5, 10)



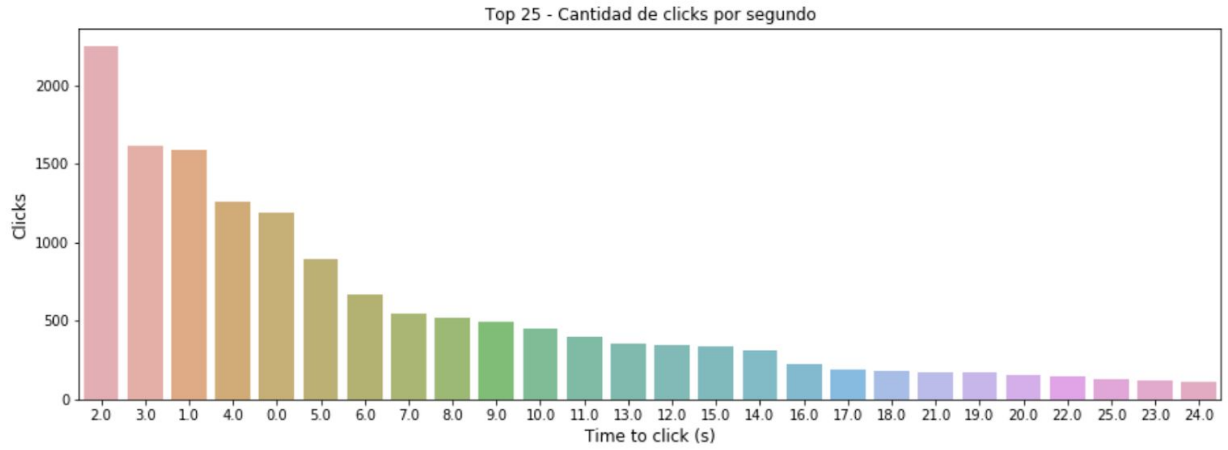
Tiempo promedio para realizar un click

Si bien se observa que alrededor del 57% de la cantidad de clicks (13191 / 22977) se encuentran distribuidos en los primeros 15 segundos, hay varios clicks que se dan pasados varios minutos y hasta horas. Puede deberse a algún tipo de error que desconocemos o realmente sucede así.

Nosotros decidimos tomarlos como válidos y, en base a esto, separamos en 3 escenarios:

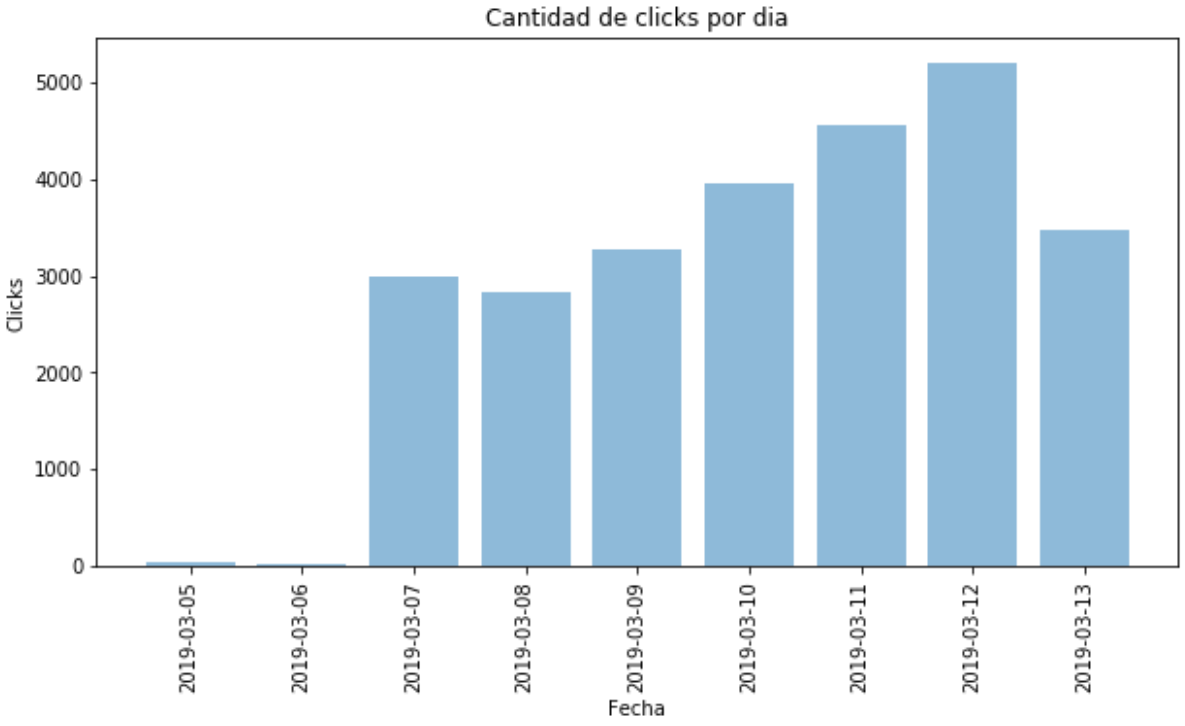
1. Tomando el total de registros (considerando los extremos más grandes): la media de tiempo hasta hacer click es de 3 minutos 20 segundos.
2. Tomando hasta los segundos que sumen al menos 30 cantidades: la media de tiempo hasta hacer click es de 44 segundos.
3. Tomando el top 25: la media de tiempo hasta hacer click es de 13 segundos.

Dado que alrededor del 57% de la cantidad de clicks se encuentran distribuidos en los primeros 15 segundos y si aumentamos a los primeros 25 segundos llegamos al 64% (14774 / 22977), creemos razonable tomar la media de tiempo hasta hacer click como 13 segundos.



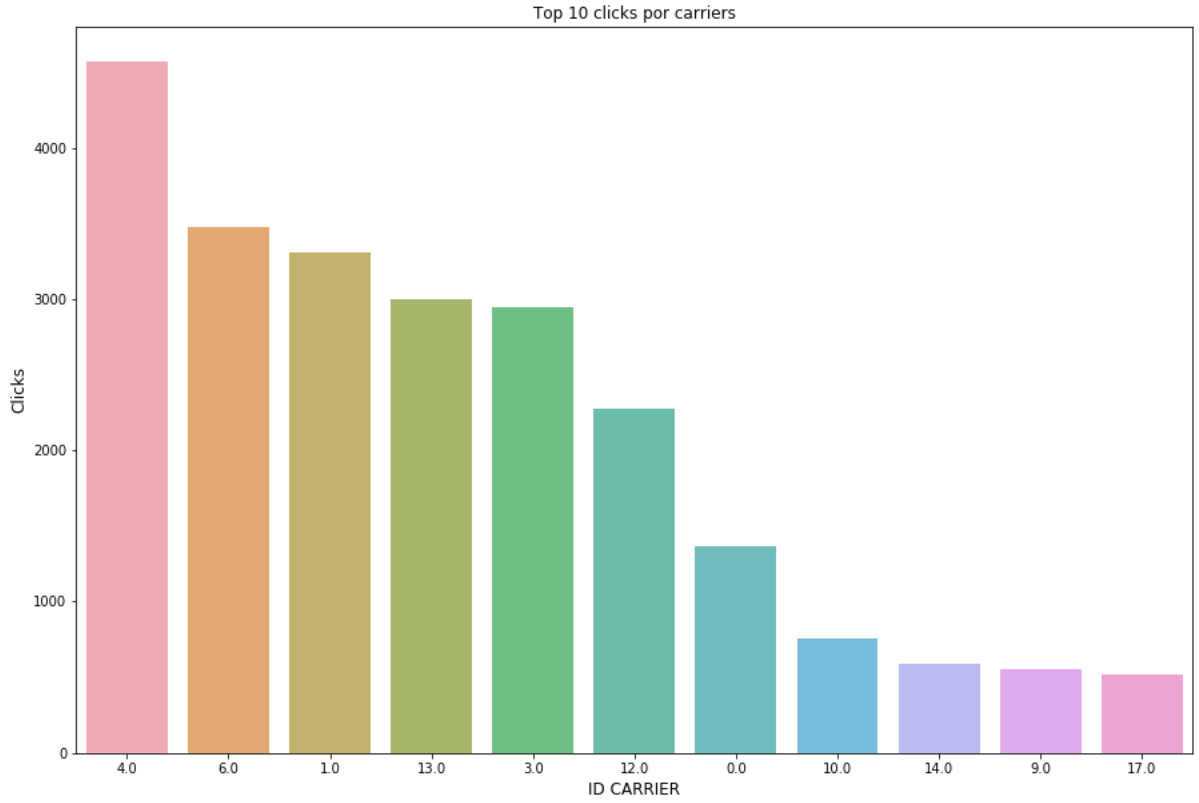
Cantidad de clicks por día

Utilizando la fecha de creación del click, podemos ver en este gráfico cómo se distribuyen a lo largo de los distintos días, observando que la mayor cantidad de clicks se produjo el día 12-03-2019 (martes). Siendo la cantidad de clicks bastante diferente de un día a otro, no se observa ningún patrón particular.



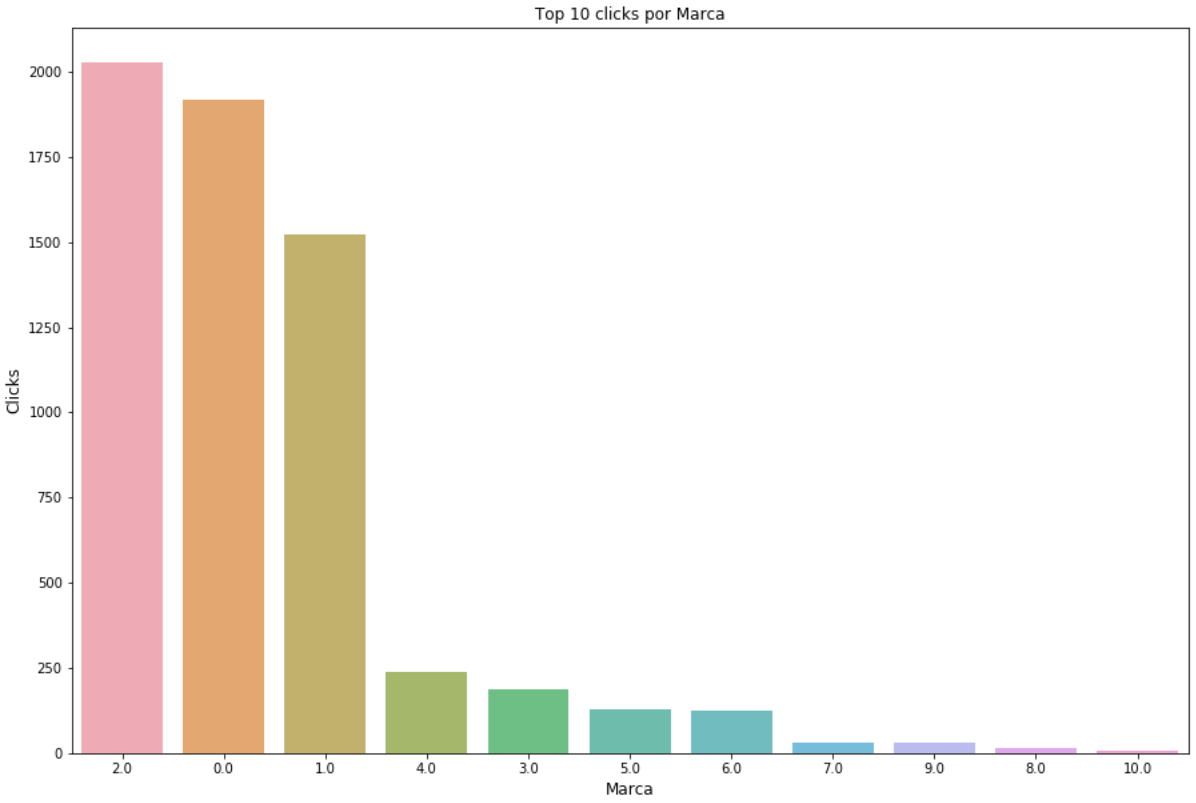
Clicks por carrier

Observando el campo carrier, podemos ver los clicks realizados y nos quedamos con el top 10. De acá se desprende que los usuarios del Carrier "4" son los que más hacen clicks por encima de los demás Carriers.



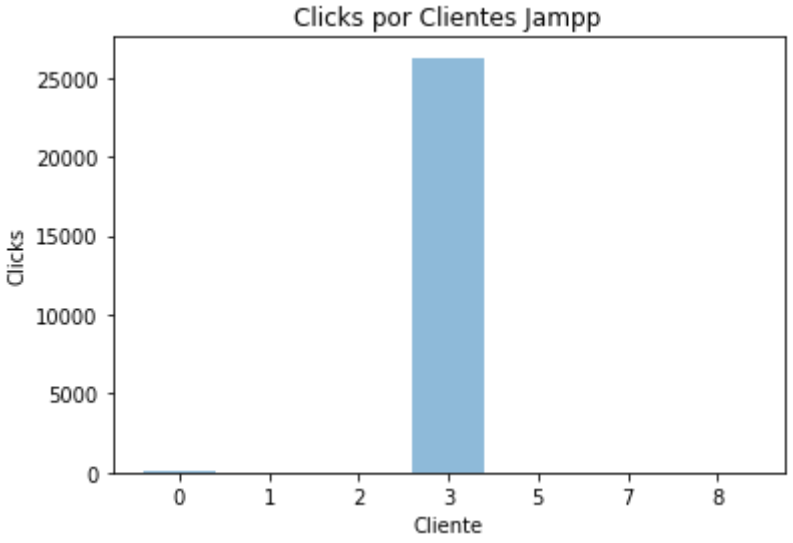
Clicks por Marca

Como vimos en el análisis inicial, en el campo de marca tenemos más del 75% de campos nulos, pero podemos utilizar la info existente para mostrar, tendencialmente, cuáles son las marcas en las cuales se hacen más clicks. Nos quedamos con el top 10 y vemos que las marcas “2” y “0”, son las que realizan la mayor cantidad de clicks.



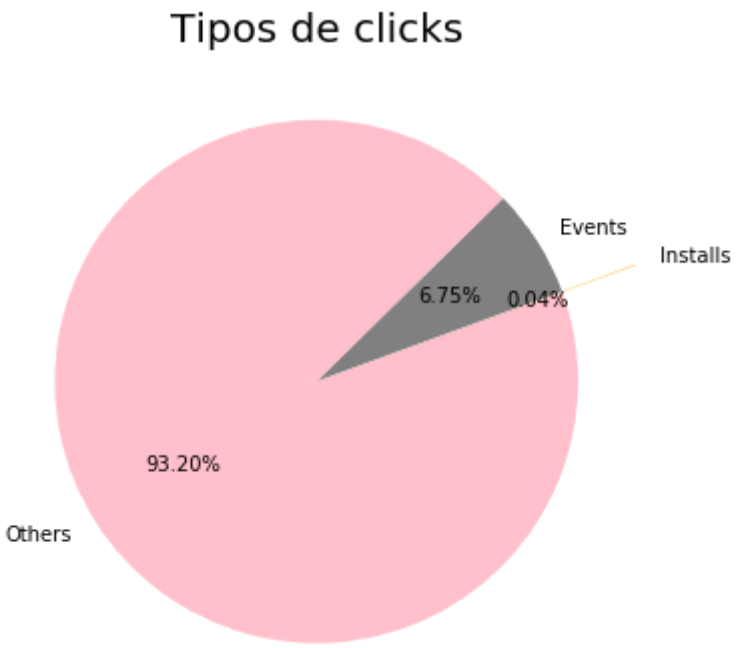
Clicks por cliente Jampp

A través de este análisis, se puede ser ver que el cliente con ID "3" es el que realiza la mayor cantidad de clicks. El resto de los demás clientes tienen valores muy pequeños.



¿Cuál es la distribución de tipos de clicks?

En este gráfico se puede observar el porcentaje de clicks que son eventos, clicks que terminan en instalaciones, y luego otro porcentaje que nos indican otros tipos de clicks, de los cuales no tenemos información. Se puede destacar que muy pocos clicks (0,04%) finalmente terminan en una instalación de la aplicación.



Dataset Events

Introducción general

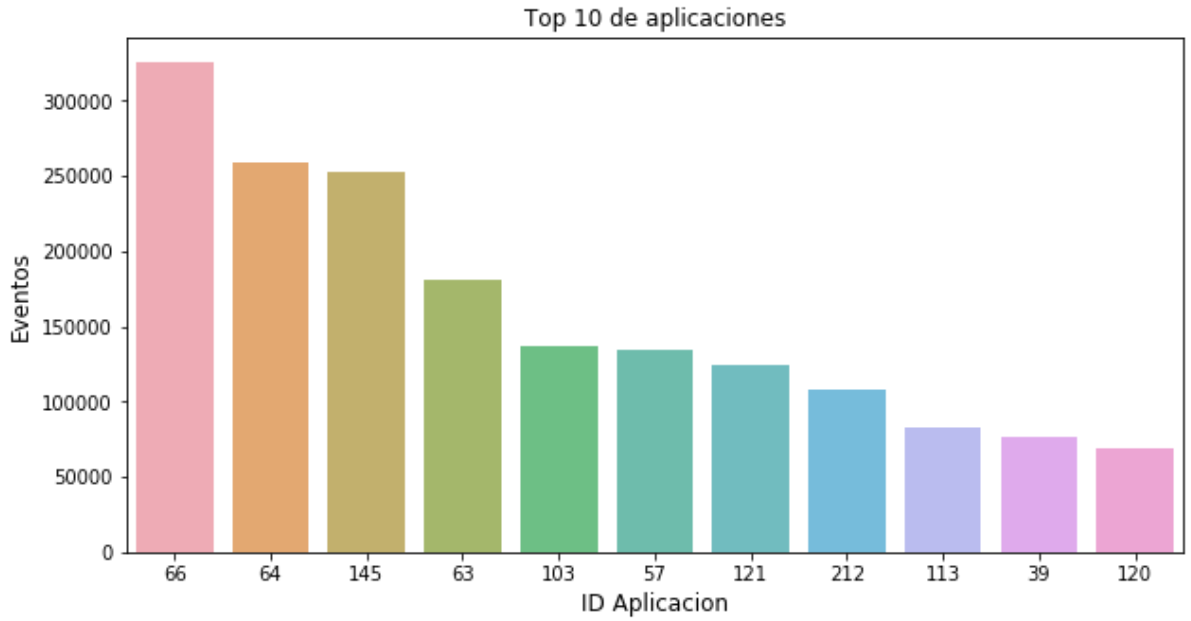
En este dataset podemos encontrar información acerca de cualquier tipo de evento que se haya producido en el dispositivo en el cual se muestra la subasta. Realizando un análisis general de los campos de este archivo podemos observar lo siguiente:

- Contamos con 2494423 registros y 22 columnas.
- trans_id tiene solamente 82 registros no nulos de un total de 2494423, por lo que no nos aporta información útil.
- Wifi tiene el 44% de los valores nulos. Del 56% restante, el 68% indica que el evento se realizó con una conexión wifi.
- connection_type tiene un 76% de valores nulos, pero con el resto de los datos podemos ver que hay tres tipos de conexiones.
- Con brand_device y device_os_version tenemos más de la mitad de los registros sin valor (53% y 59% respectivamente).

- Con device_countrycode, podemos observar que solamente tenemos datos de un único país, por lo que lo descartamos para los análisis.
- Con el campo attributed vemos que solamente el 0.20% corresponden a eventos que son atribuídos a Jampp.

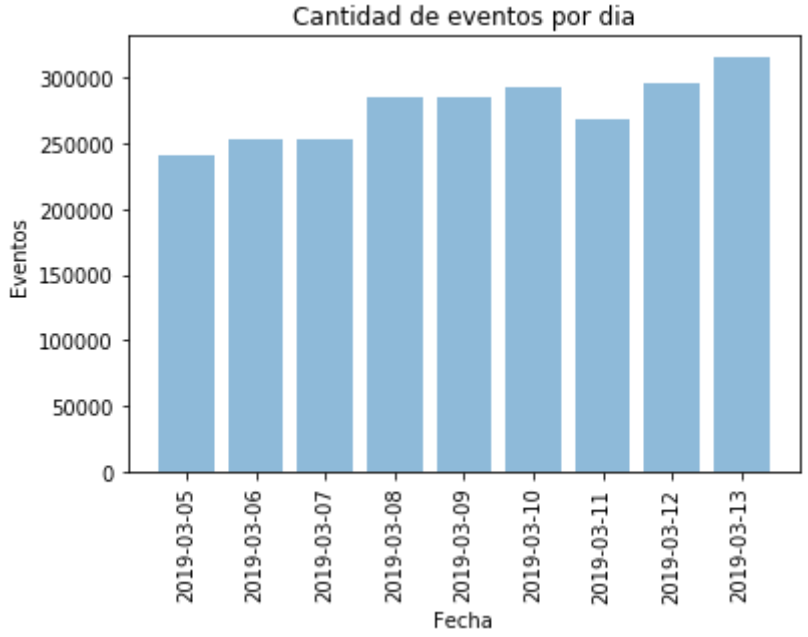
Aplicaciones dónde se realizaron más eventos

Agrupando la información por el campo application_id, podemos observar la cantidad de eventos que se produjeron por aplicación. Nos quedamos con el top 10 y de esto vemos que en las aplicaciones con ID 66, 64 y 145 son las que realizan la mayor cantidad de eventos totales.



Cantidad de eventos por día

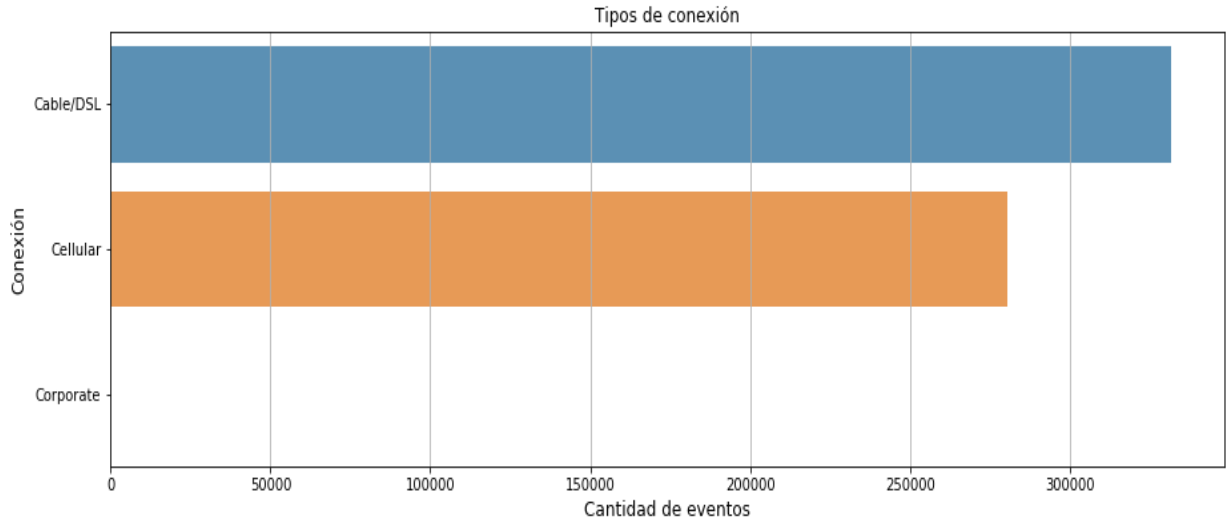
Observando la fecha de creación del evento, vemos cuántos de estos se realizan por día. En el gráfico se puede ver que la cantidad de eventos se mantiene constante a lo largo de la semana, sin picos particulares durante el fin de semana.



Cantidad de eventos por tipo de conexión

Como vimos en el análisis inicial, en el campo de connection_type tenemos 76% de campos nulos, pero podemos utilizar la información existente para mostrar los tipos de conexión utilizados al momento del evento.

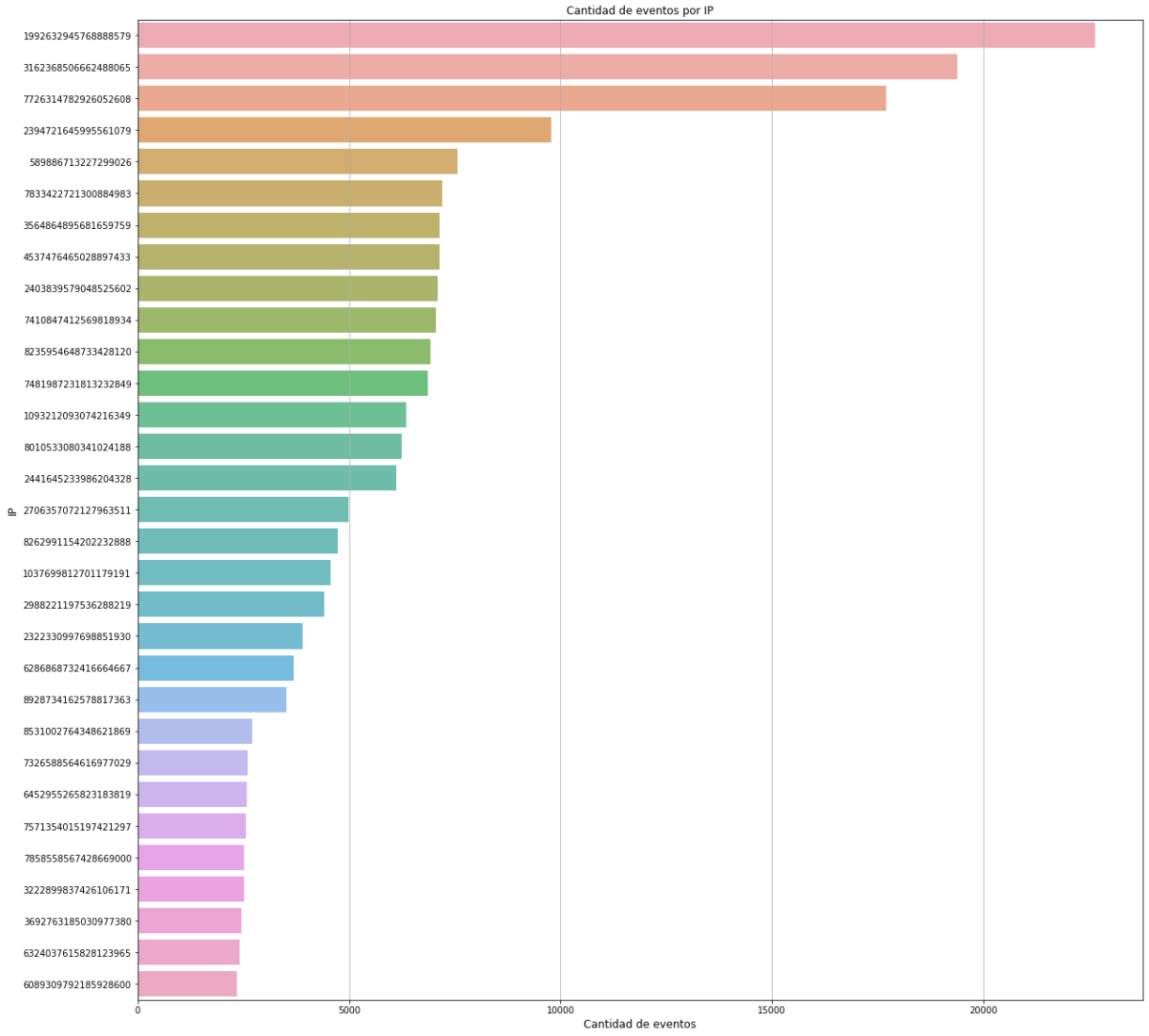
	conexion	cantidad
0	Cable/DSL	331948
1	Cellular	280511
2	Corporate	4



Cantidad de eventos por dirección IP

Agrupamos los eventos por la dirección IP del dispositivo. Podemos ver que la IP **1992632945768888579** es la que realizó la mayor cantidad de eventos.

	index	ip_address
0	1992632945768888579	22640
1	3162368506662488065	19379
2	7726314782926052608	17704
3	2394721645995561079	9777
4	589886713227299026	7561
5	7833422721300884983	7197
6	3564864895681659759	7142
7	4537476465028897433	7140
8	2403839579048525602	7100
9	7410847412569818934	7065
10	8235954648733428120	6933
11	7481987231813232849	6855
12	1093212093074216349	6347
13	8010533080341024188	6241
14	2441645233986204328	6108
15	2706357072127963511	4993
16	8262991154202232888	4736
17	1037699812701179191	4554
18	2988221197536288219	4408
19	2322330997698851930	3901
20	6286868732416664667	3698
21	8928734162578817363	3519
22	8531002764348621869	2710
23	7326588564616977029	2595
24	6452955265823183819	2571
25	7571354015197421297	2556
26	7858558567428669000	2520
27	3222899837426106171	2518
28	3692763185030977380	2450
29	6324037615828123965	2408
30	6089309792185928600	2340



¿Cuál es el porcentaje de conversión de eventos a instalaciones?

Relacionamos los datasets de eventos e instalaciones a través del campo ref_hash y obtuvimos que el 1,7% de los eventos son instalaciones, es decir 42474 eventos de un total de 2494423.

¿Cuál es el porcentaje de subastas sobre eventos?

Relacionamos los datasets de eventos y subastas a través del campo ref_hash y device_id y obtuvimos que el 41,9 % de los eventos son subastas, es decir 1047126 eventos de un total de 2494423.

Datasets Installs

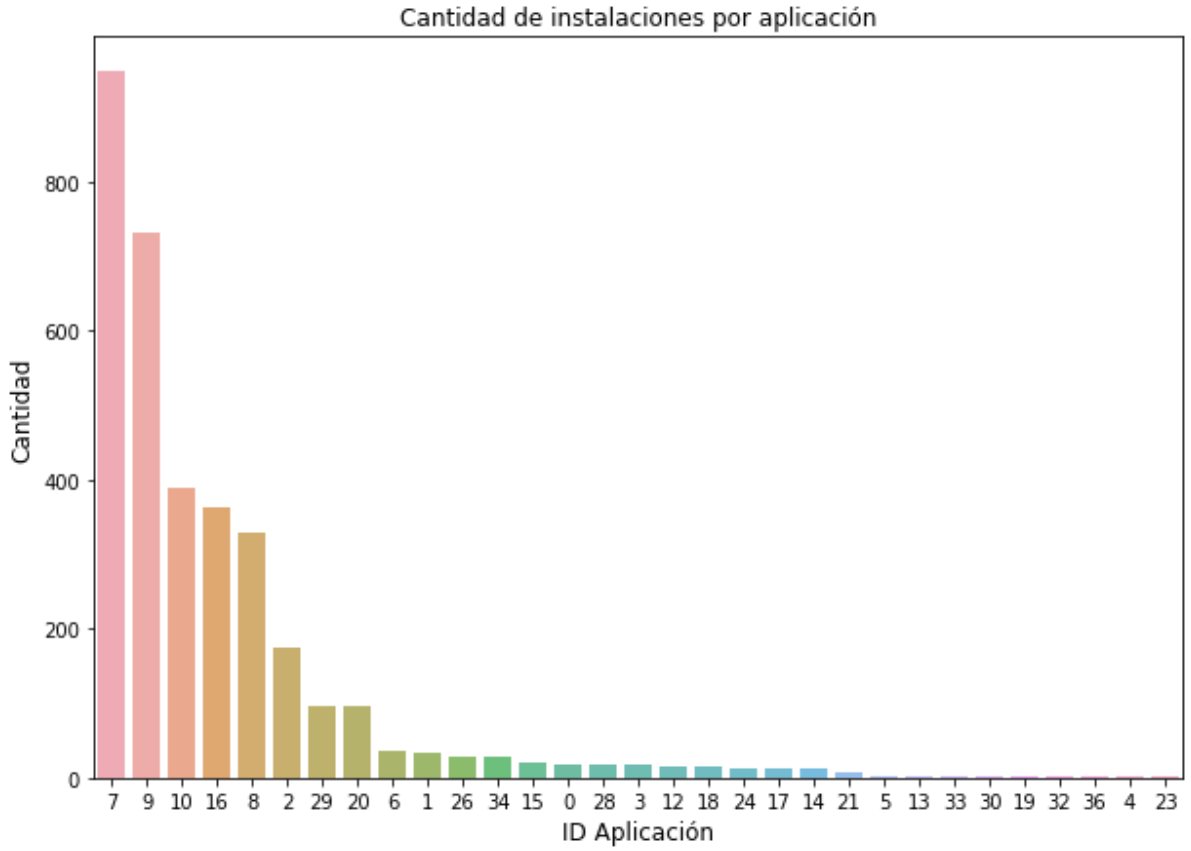
Introducción general

En el archivo installs.csv podemos encontrar datos acerca de las instalaciones de las aplicaciones realizadas de los clientes de Jampp, pudiendo ser o no a causa de la intervención de Jampp. Realizando un análisis inicial se puede ver lo siguiente:

- Contamos con 3412 registros y 18 columnas.
- Las instalaciones se realizaron en dos países diferentes.
- El 40% de las instalaciones se realizaron con una conexión a una red wifi.
- Ninguna instalación fue atribuida a Jampp.
- Sólo existe un registro sin valor de modelo, por lo que podremos analizar los modelos en los que se realizan la mayor cantidad de instalaciones.
- Sólo un tercio de los registros tienen datos acerca de las marcas de los dispositivos.

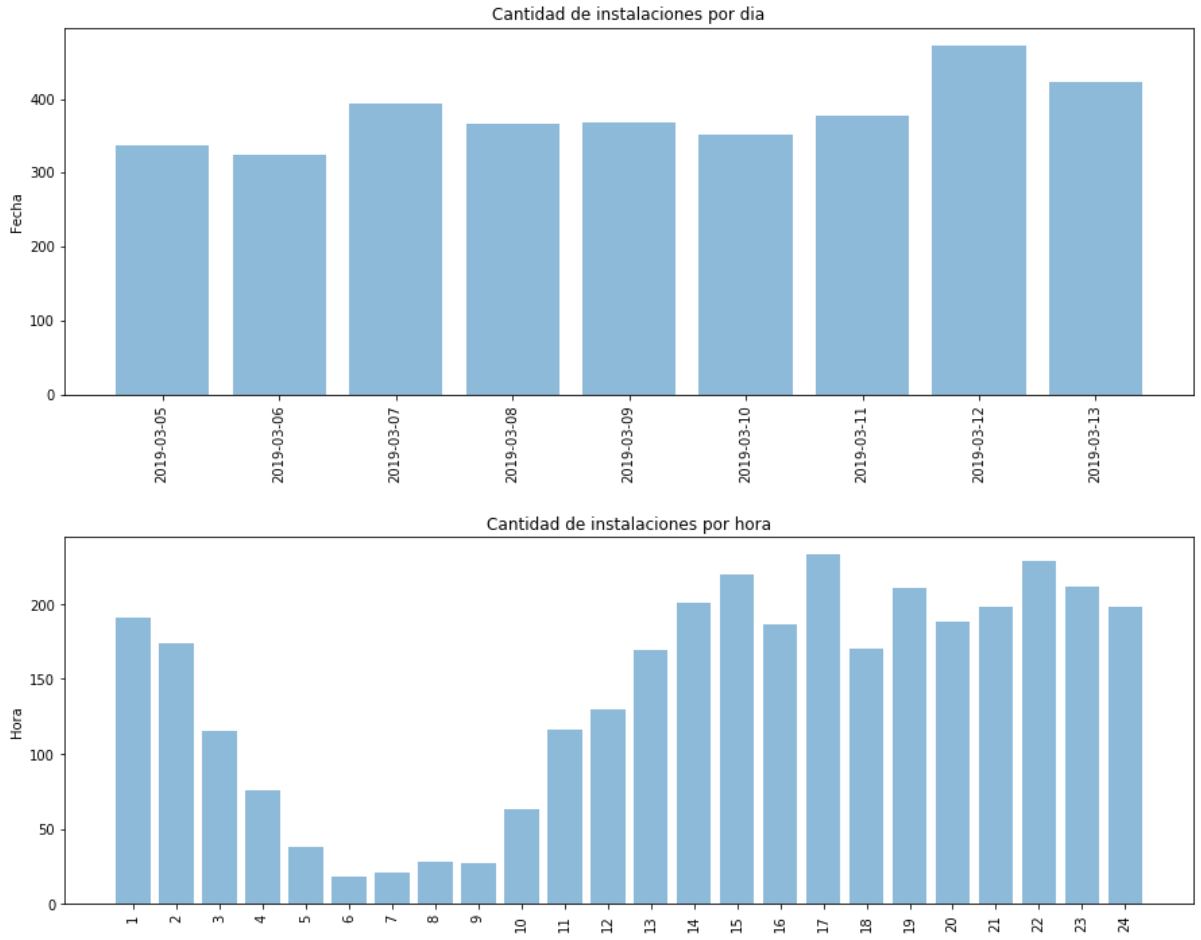
Cantidad de instalaciones por aplicación

Agrupando los datos por el campo application_id, podemos obtener cuáles son las aplicaciones que fueron más veces. En el gráfico podemos visualizar que las aplicaciones con ID 7 y 9 son las más instaladas.



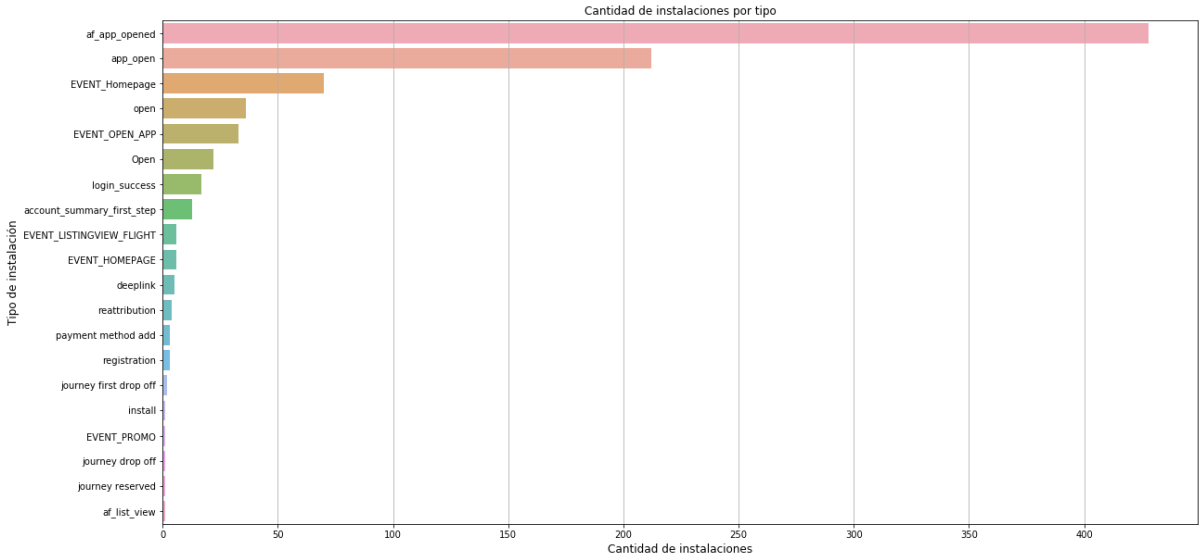
Cantidad de instalaciones por día y hora

La cantidad de instalaciones por día se mantiene constante durante todos los días de la muestra. Sin embargo, se puede observar que la mayor cantidad de instalaciones se producen en dos rangos de horarios : entre las 14 y 17 hs, y luego entre las 22 y 00 hs.
También podemos ver que a hora donde se produce la mayor cantidad de instalaciones es 17hs (233 installs) y la hora donde se produce la menor cantidad de instalaciones es 6hs (18 installs).



Cantidad de instalaciones por tipo

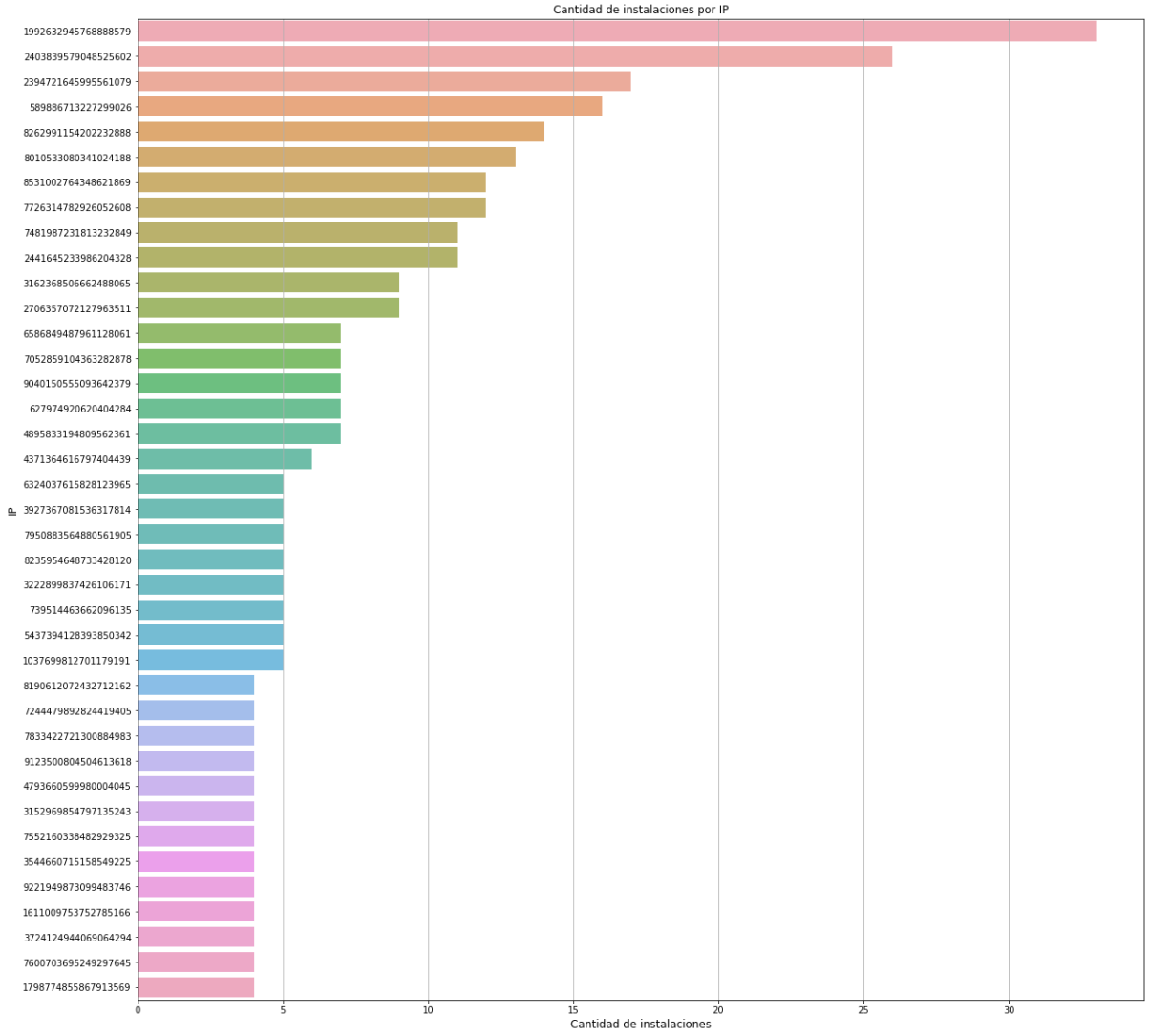
Visualizando el campo “kind”, podemos ver que existen distintos tipos de instalaciones, de las cuales no tenemos información acerca de las mismas o cuáles son sus características particulares, pero podemos ver que la mayor cantidad de instalaciones es del tipo “**af_app_opened**”.

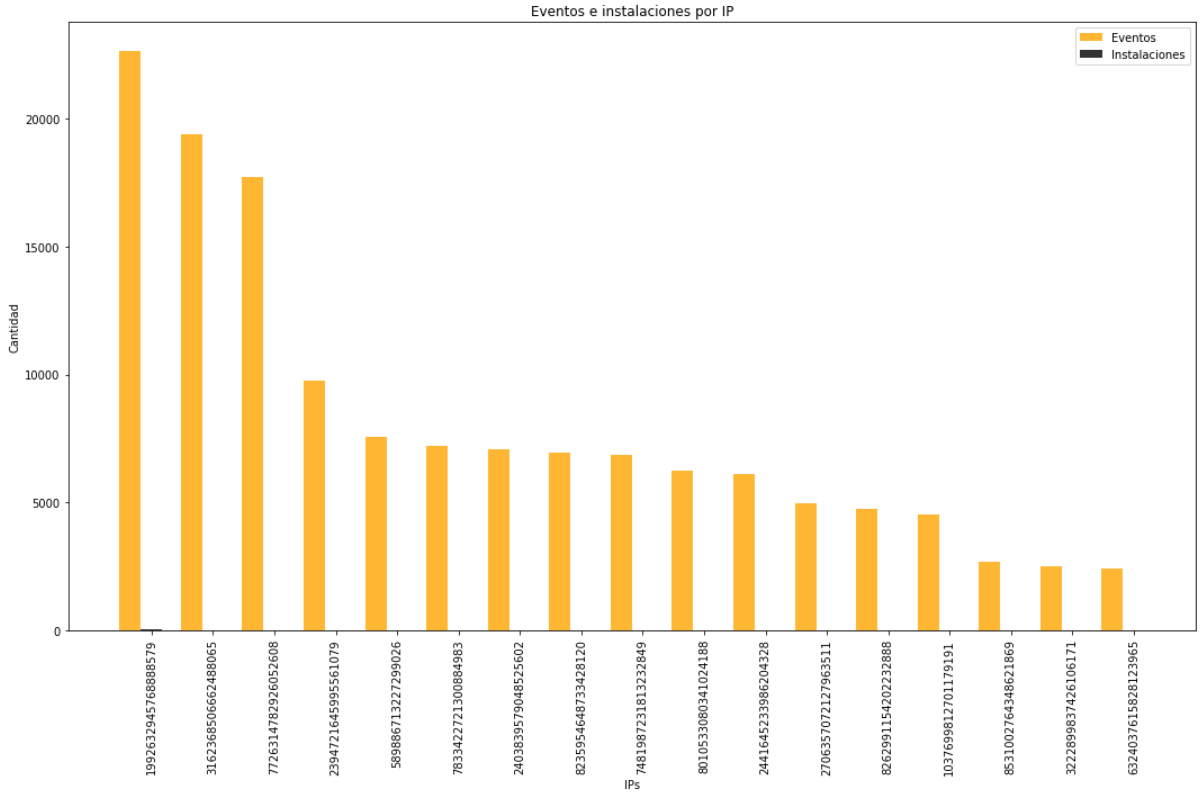


Cantidad de instalaciones por IP

Agrupamos las instalaciones por la dirección IP del dispositivo. Podemos ver que la IP **1992632945768888579** es la que realizó la mayor cantidad de instalaciones y de eventos, como vimos en el análisis del dataset anterior.

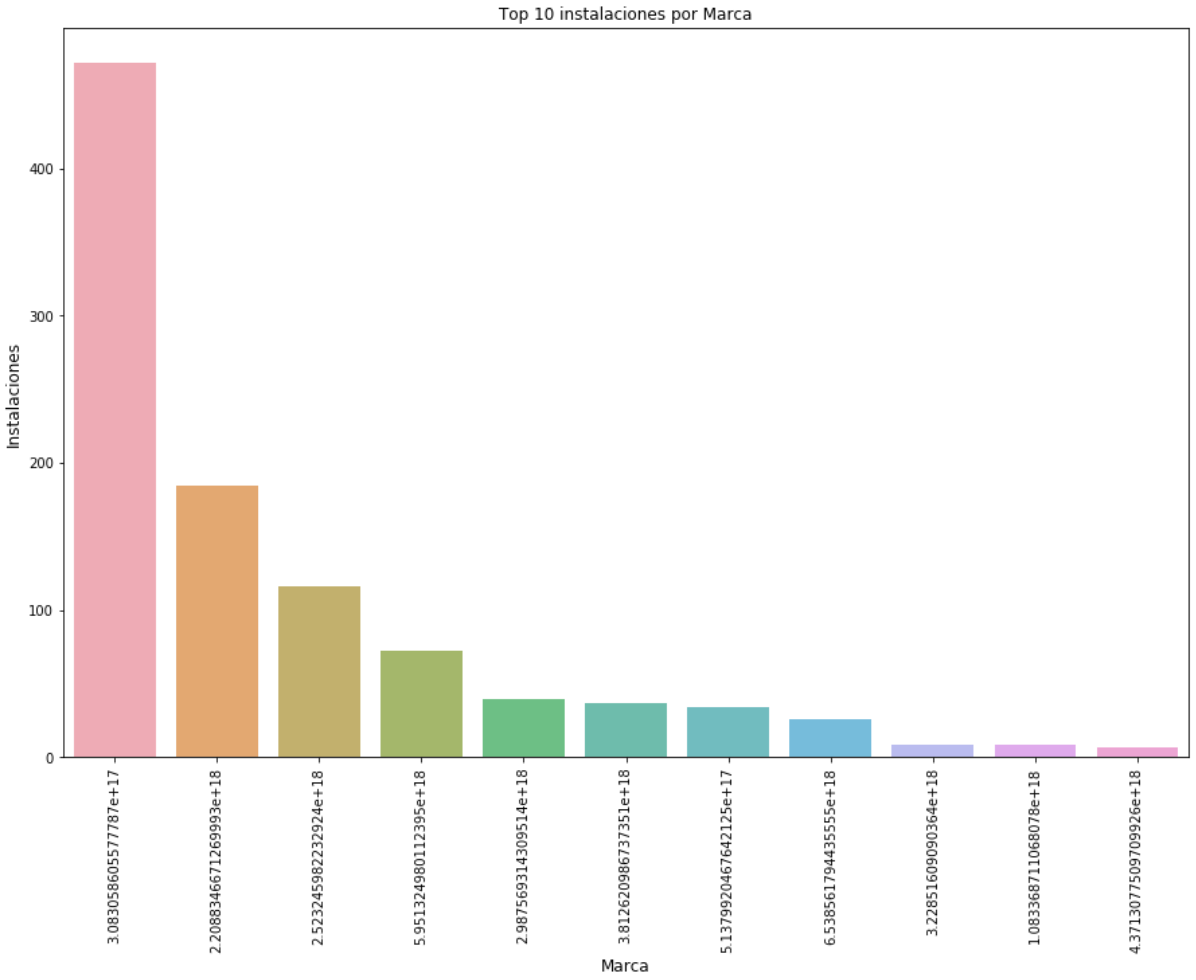
	index	ip_address
0	1992632945768888579	33
1	2403839579048525602	26
2	2394721645995561079	17
3	589886713227299026	16
4	8262991154202232888	14
5	8010533080341024188	13
6	8531002764348621869	12
7	7726314782926052608	12
8	7481987231813232849	11
9	2441645233986204328	11
11	3162368506662488065	9
10	2706357072127963511	9
12	6586849487961128061	7
13	7052859104363282878	7
14	9040150555093642379	7
15	627974920620404284	7
16	4895833194809562361	7
17	4371364616797404439	6
22	6324037615828123965	5
25	3927367081536317814	5
24	7950883564880561905	5
23	8235954648733428120	5
18	3222899837426106171	5
21	739514463662096135	5
20	5437394128393850342	5
19	1037699812701179191	5
37	8190612072432712162	4
46	7244479892824419405	4
45	7833422721300884983	4
44	9123500804504613618	4
43	4793660599980004045	4
42	3152969854797135243	4
41	7552160338482929325	4
40	3544660715158549225	4
39	9221949873099483746	4
38	1611009753752785166	4
31	3724124944069064294	4
36	7600703695249297645	4
30	1798774855867913569	4


Cantidad de instalaciones por IP
IP
1992632945768888579
2403839579048525602
2394721645995561079
58988671322729902 6
8262991154202232888
8010533080341024188
8531002764348621869
7726314782926052608
7481987231813232849
2441645233986204328
31623685066624880 65
27063570721279635 11
6586849487961128061
7052859104363282878
9040150555093642379
627974920620404284
4895833194809562361
43713646167974044 39
6324037615828123965
39273670815363178 14
7950883564880561905
82359546487334281 20
3222899837426106171
739514463662096135
5437394128393850342
10376998127011791 91
8190612072432712162
7244479892824419405
7833422721300884983
91235008045046136 18
4793660599980004045
31529698547971352 43
75521603384829293 25
35446607151585492 25
92219498730994837 46
16110097537527851 66
37241249440690642 94
76007036952492976 45
17987748558679135 69
0
5
10
15
20
25
30
Cantidad de instalaciones



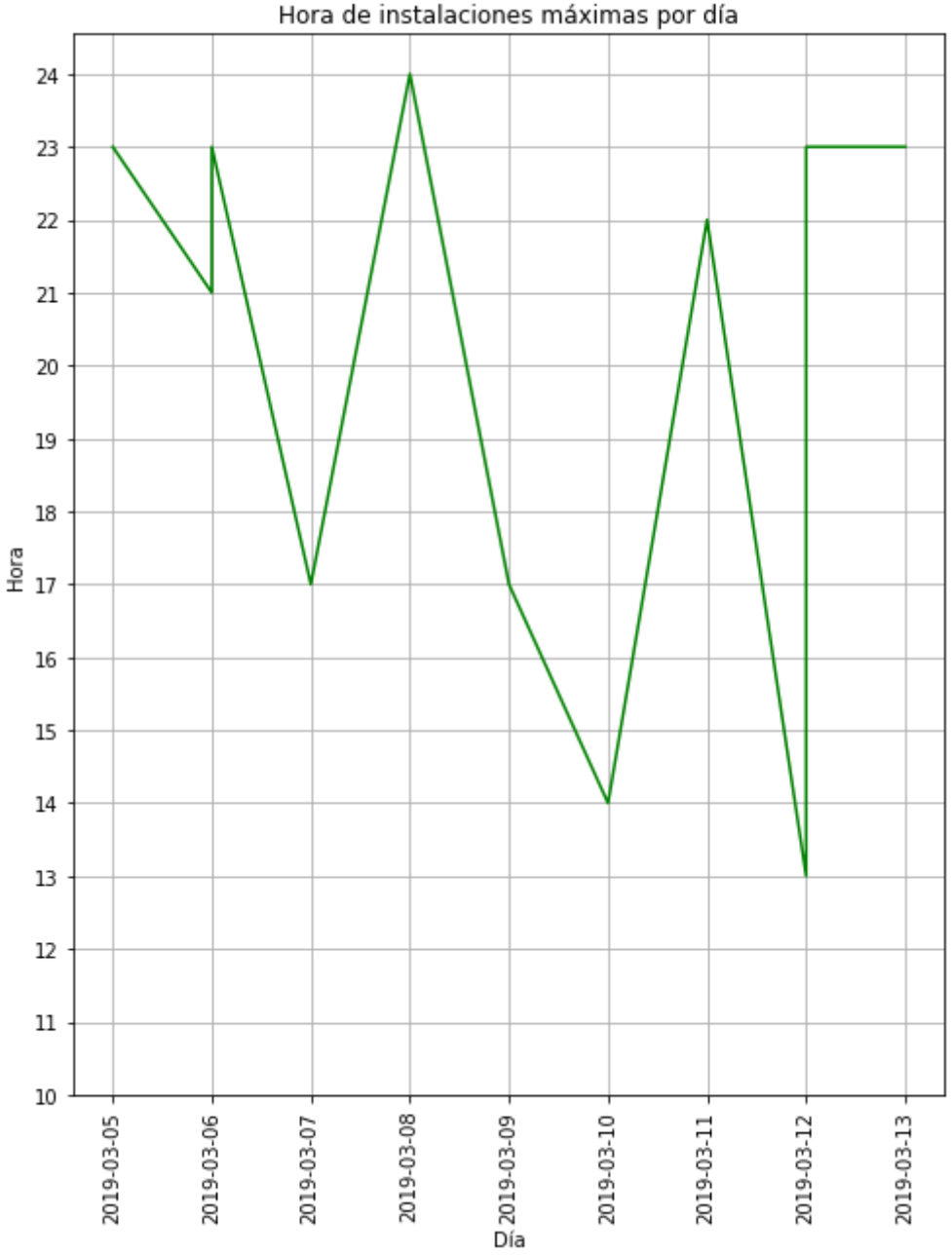
Cantidad de instalaciones por Marca

Realizamos el top 10 de instalaciones por marca, observando que las marcas 3.083 y 2.208 son en donde se realizan la mayor cantidad de instalaciones.



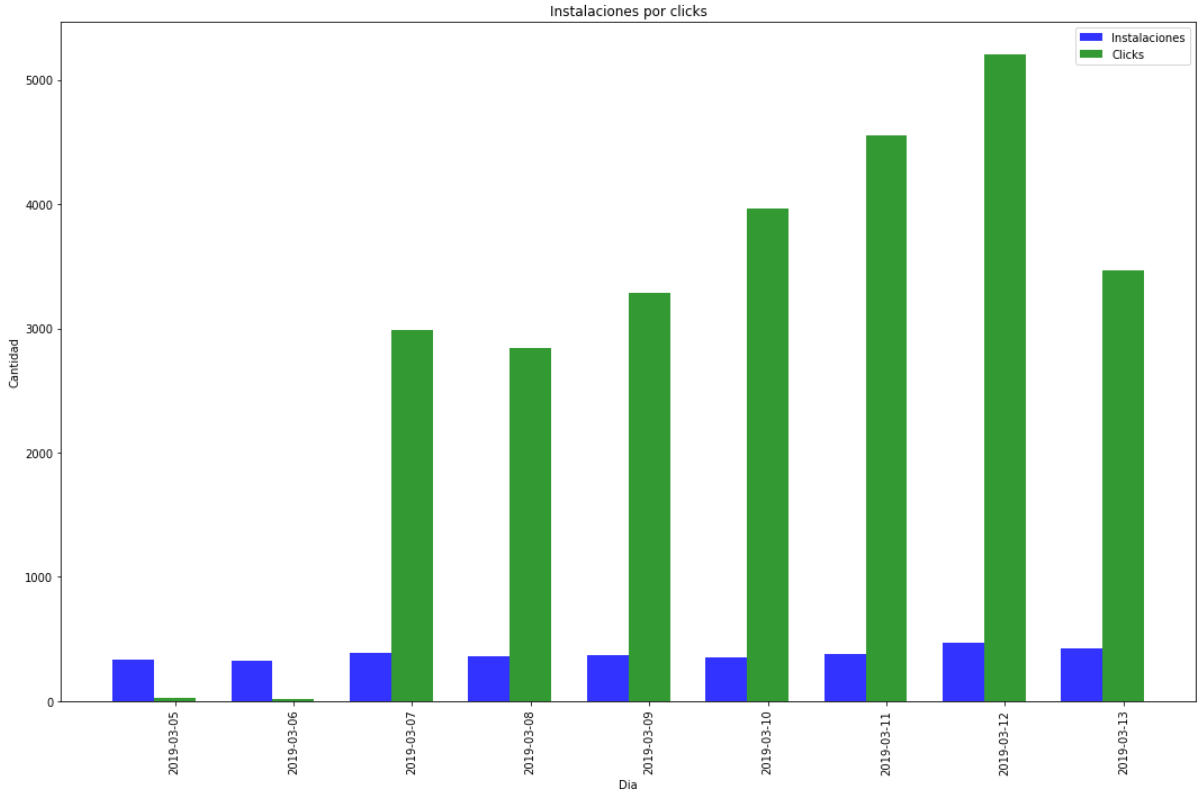
Hora de instalaciones máximas por día

Utilizando el campo de creación de las instalaciones, se generaron dos columnas nuevas al dataset, una correspondiente al día de instalación y otra haciendo referencia a la hora en que ocurrió la misma. Con esta información, se obtuvieron las horas en donde se produjeron las cantidades máximas de instalaciones por día.



Comparación cantidad de instalaciones/clicks por día

Como se puede observar en el gráfico, la proporción entre la cantidad de clicks e instalaciones que se realizan por día es muy diferente, ya que hay muy pocas instalaciones con respecto a la cantidad de clicks que se realizan. Además también se puede observar que en los días 5 y 6 ocurre lo opuesto, habiendo más instalaciones que clicks. Esto se puede deber a que se realizaron instalaciones de aplicaciones por otro medio, es decir, por fuera de la publicidad ofrecida en las subastas.



Datasets Auctions

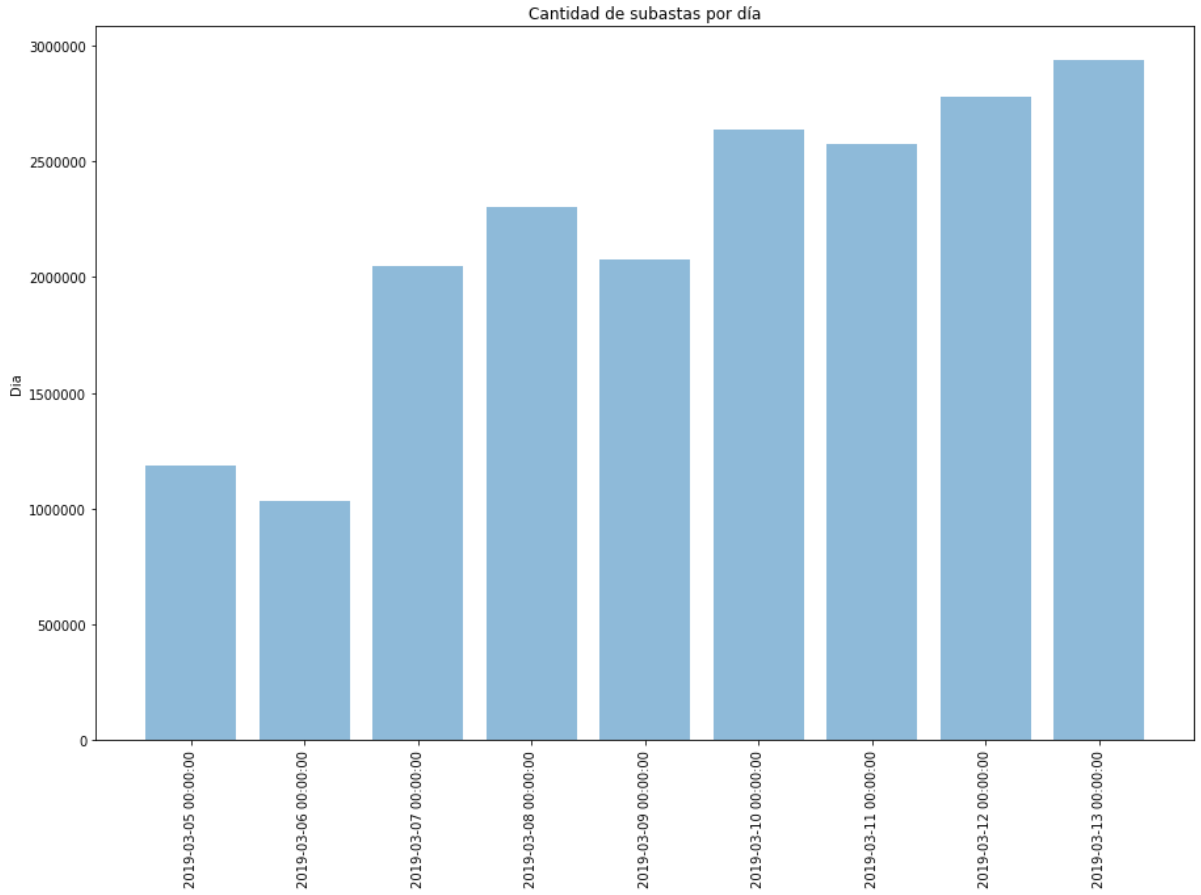
Introducción general

En ese set de datos tenemos información acerca de las subastas realizadas por Jampp. Analizando los campos de dicho dataset, podemos ver lo siguiente:

- Contamos con 19571319 registros y 7 columnas.
- auctions_type_id tiene todos los registros nulos, por lo que no nos aporta información útil.
- country tiene solamente datos de un único país, por lo que lo descartamos para los análisis.
- platform contiene 2 valores y suponemos que indican si la plataforma es Android o IOS.
- source_id contiene 5 códigos y los utilizaremos para hacer un ranking.

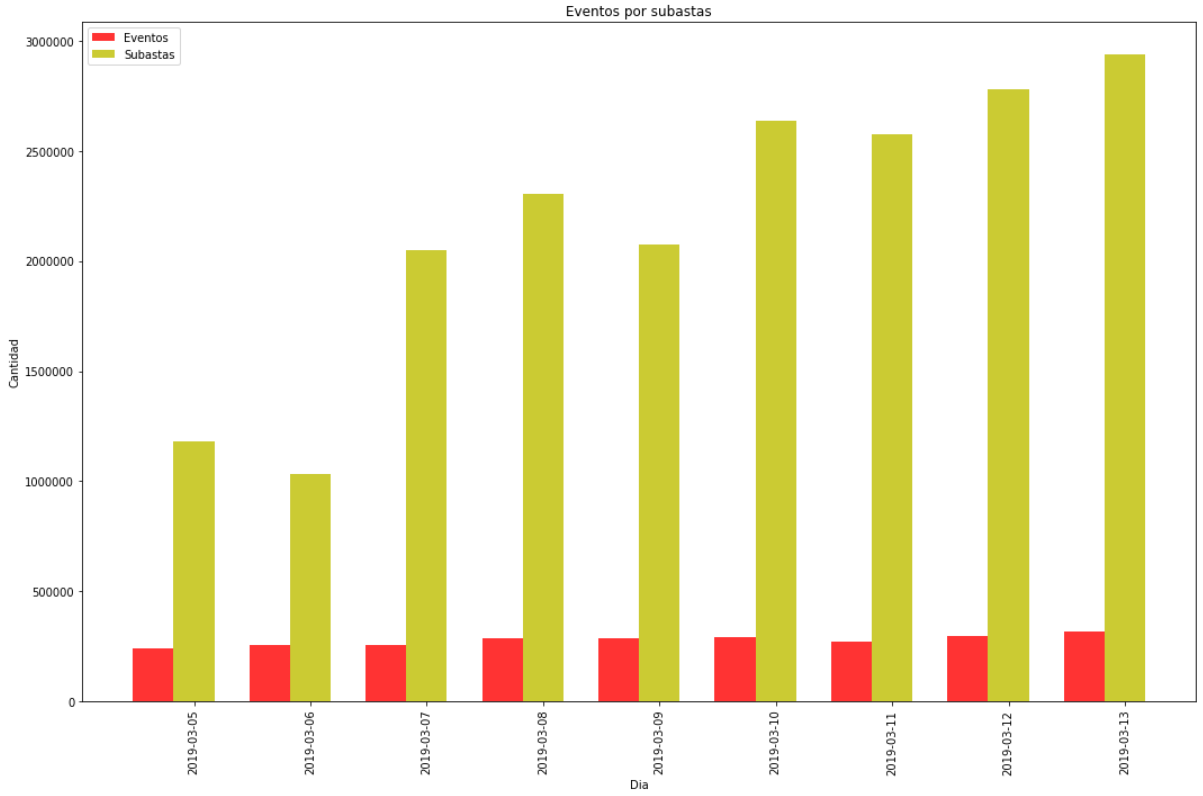
Cantidad de subastas por día

Analizando el campo de fecha, obtenemos la cantidad de subastas que se realizaron por día. Podemos observar que el día con mayor cantidad de subastas fue el martes 13 de marzo. No se observa ningún patrón particular respecto a las subastas realizadas por día, ya que en los primeros días de la primer semana (5,6 y 7) se ve que la cantidad es muy poca respecto a la segunda semana (11,12 y 13), donde aumenta casi al doble.

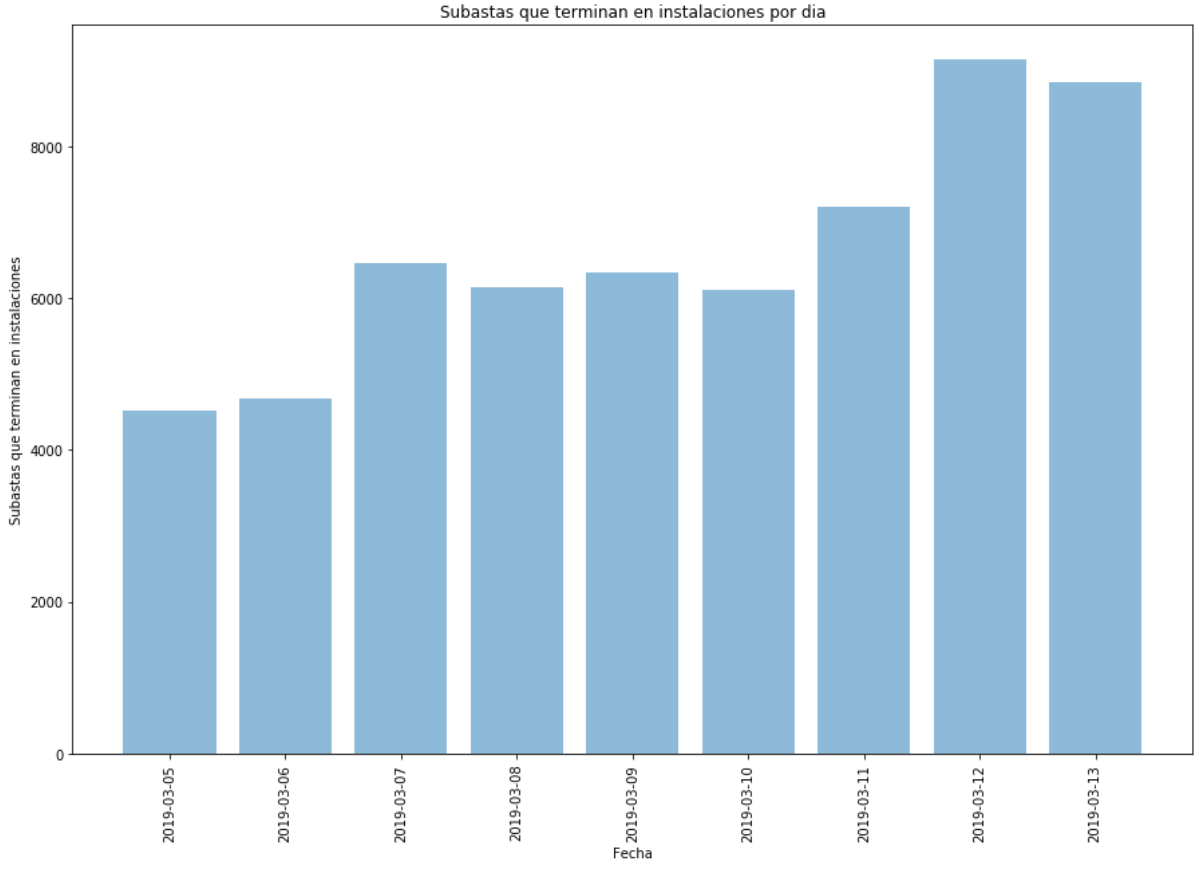


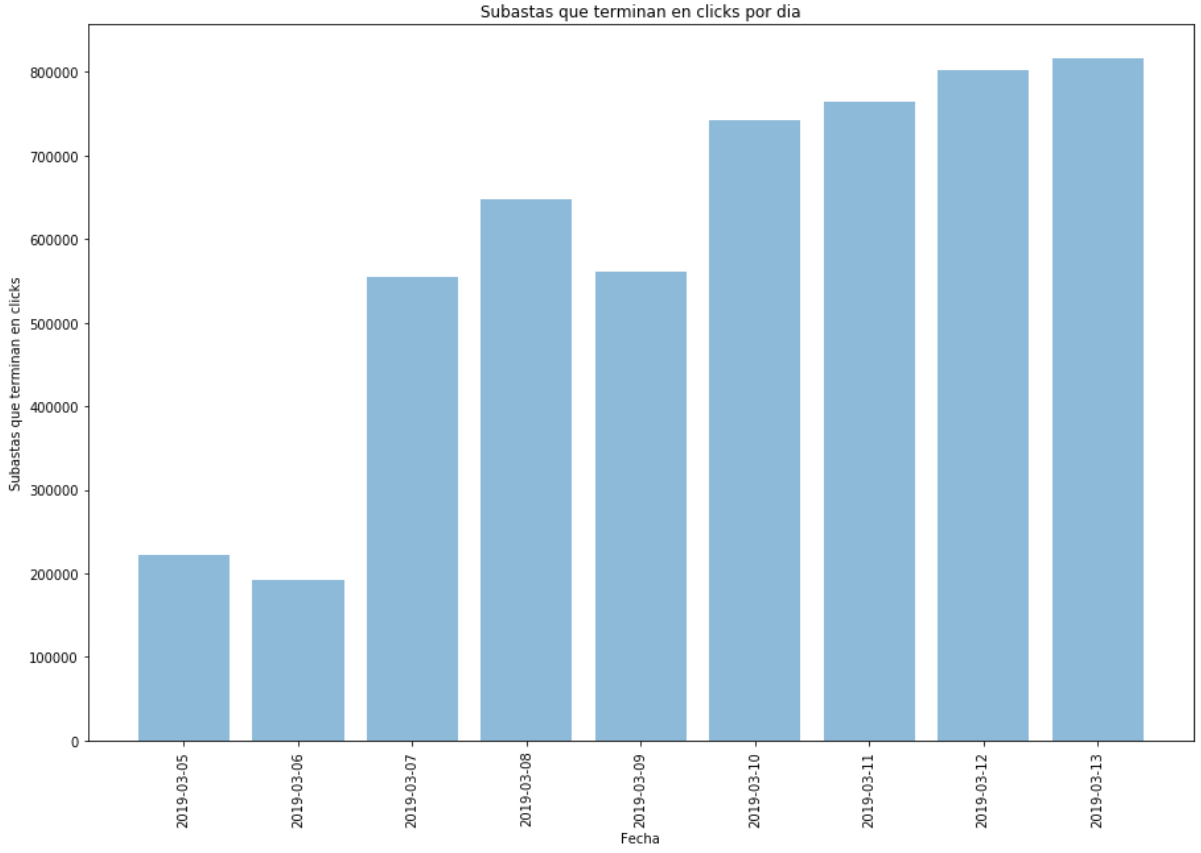
Comparación de eventos/subastas por día

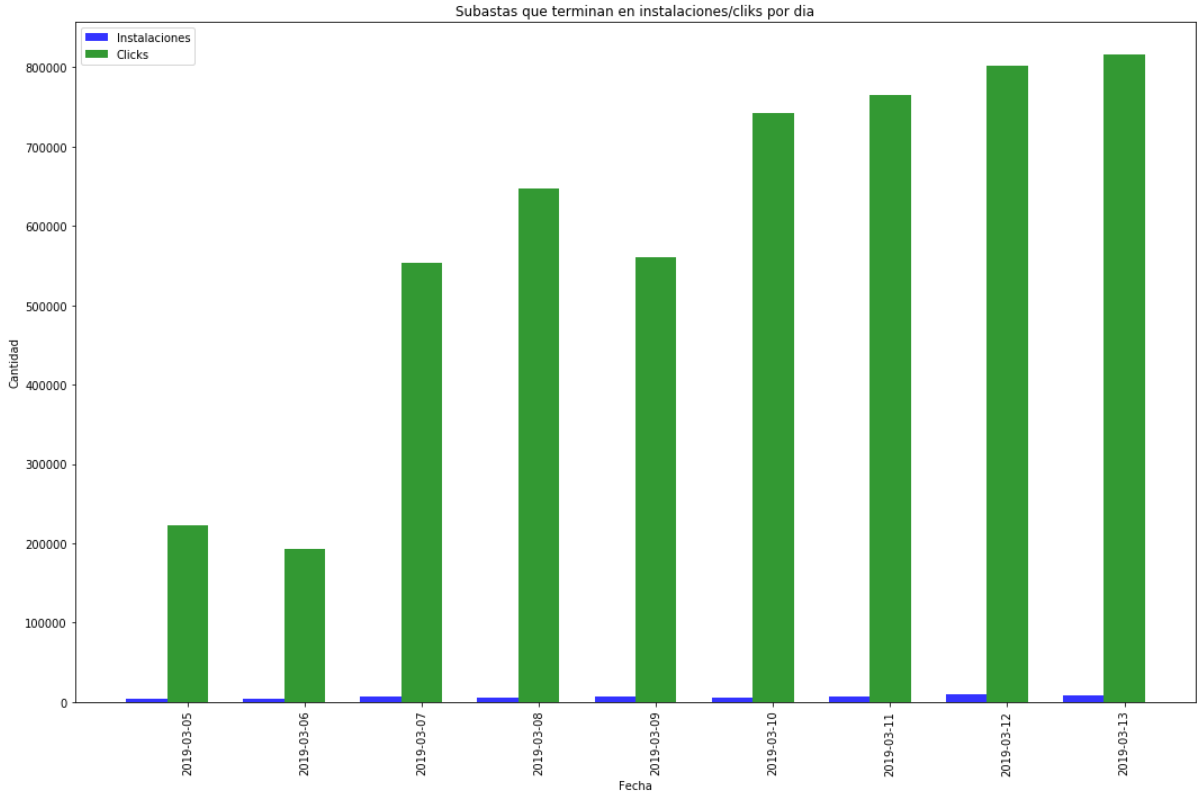
En este gráfico podemos observar que la cantidad de subastas es muchísimo mayor a la cantidad de eventos que se producen en un dispositivo por día. Mientras que la cantidad de eventos se mantiene bastante constante en el tiempo, la cantidad de subastas aumenta en los últimos días de la muestra.



¿Cuántas subastas terminan en instalaciones y cuántas en solamente clicks?



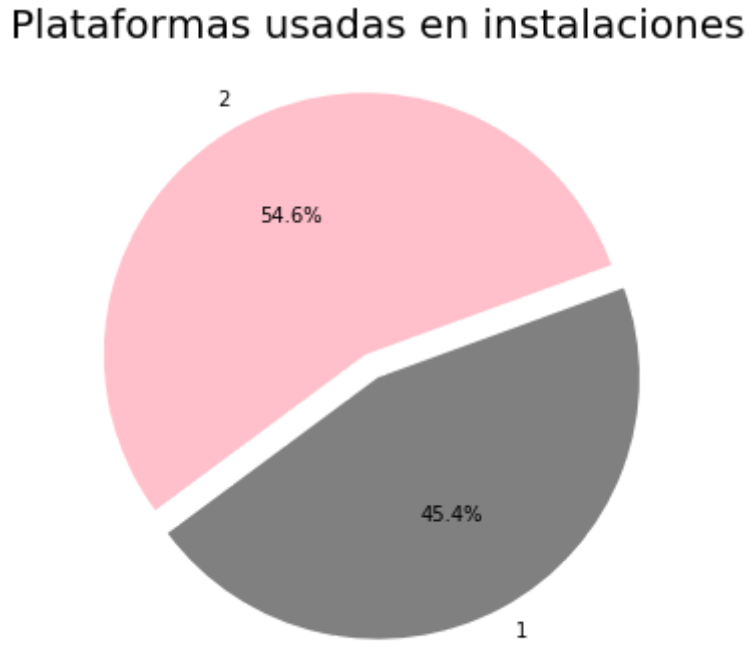




Por lo que podemos observar, el 27,1% de las subastas realizadas terminan en clicks pero sólo el 0.3% de las mismas terminan en instalaciones.

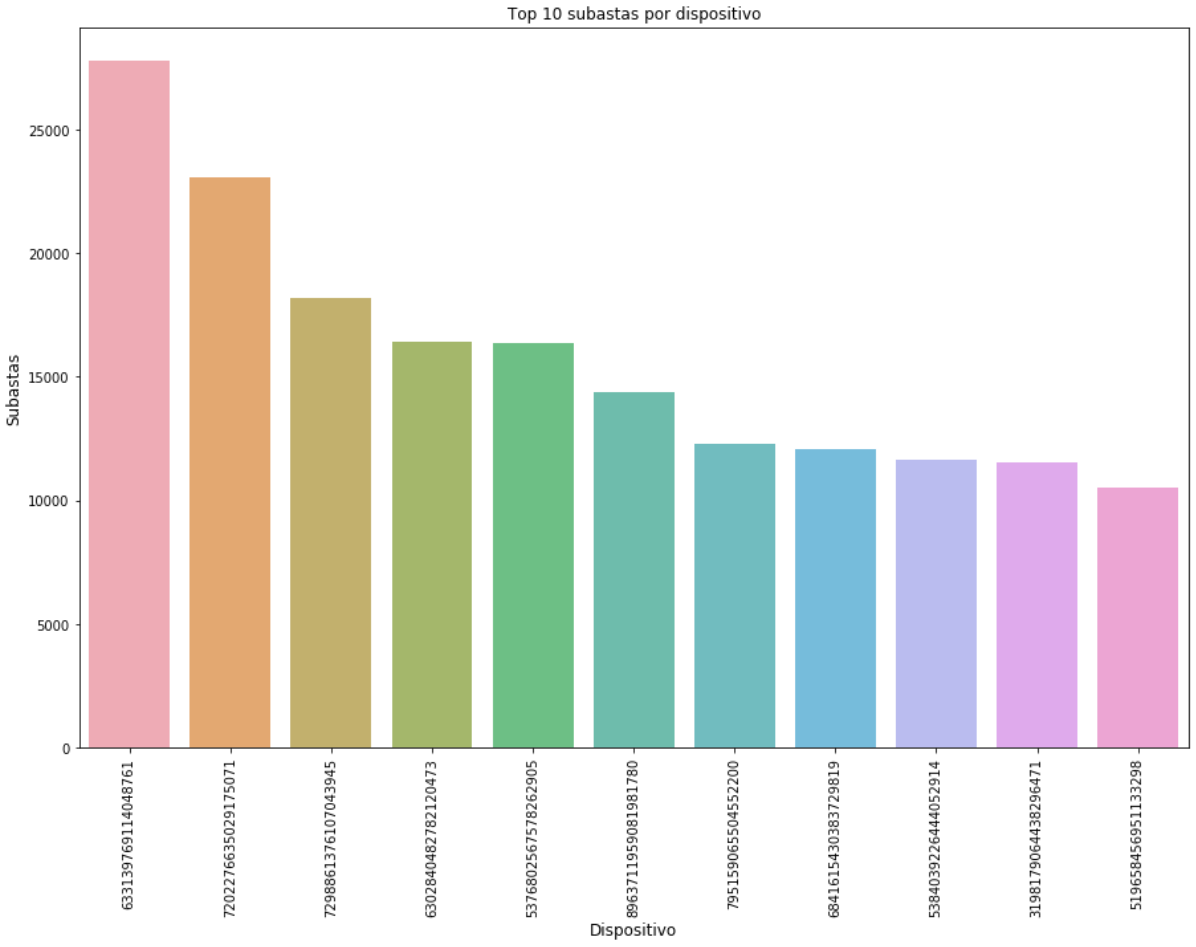
¿Sobre qué plataforma se realizan más subastas que terminan en instalaciones?

A través del campo platform, obtenemos datos sobre las plataformas utilizadas en las instalaciones. Se puede ver que sólo se utilizan dos plataformas, y la plataforma con ID “2” es en la que se realizan la mayor cantidad de instalaciones.



¿Sobre cuál dispositivo se realizaron más subastas?

Utilizando el campo device_id, contamos los dispositivos sobre los cuales se realizó la mayor cantidad de subastas y obtenemos el top 10. En el gráfico podemos ver que el dispositivo 633139769114048761 es en el cual más subastas se realizaron.



Conclusiones

Luego del análisis de cada dataset y de haberlos cruzados para ver cómo se relacionaban, en cada punto fuimos destacando algunos aspectos importantes que considerábamos que nos brindaban información relevante para cada uno. Si bien se tuvo en cuenta desde el principio y en cada parte que el porcentaje de conversión era muy bajo, pudimos replicar dicha hipótesis a lo largo del análisis, logrando ver que la tasa de instalaciones era de alrededor 0,04%.

No obstante, cabe destacar algunos insights que hemos mencionado y pueden aportarnos información importante de cara a mejorar el proceso de captura de Jampp.

- Teniendo en cuenta la dispersión de posiciones en donde se hacen clicks y el tiempo medio hasta hacer click, podríamos mejorar la eficiencia de visualización de las publicidades dirigiéndose a dichas coordenadas XY y que las mismas duren al menos 15 segundos.
- Los usuarios del Carrier 4 tienen mayor tasa de clicks, por lo que se podría apuntar a mostrarle más publicidad a los usuarios que tienen dicho proveedor.

- También podríamos utilizar como objetivo a las marcas con ID 2 y 3, las cuales son las que más click tiene.
- Teniendo en cuenta las ventanas de tiempo en donde se realizan la mayor cantidad de instalaciones, se puede lograr que las publicidades se realicen dentro de esos horarios para lograr una mayor tasa de conversión, es decir, de 14 a 17 y de 22 a 00.
- Se puede apuntar a la plataforma con ID 2, la cual tiene más porcentaje de instalaciones sobre subastas.
- Si se identifica a las conexiones de tipo Cable/DSL, podemos obtener mayor tasa de eventos sobre las aplicaciones, ya que creemos que es por la estabilidad de la conexión.
- Podemos identificar a los usuarios que más subastas reciben, más clicks hacen y más instalaciones realizan, por lo que suponemos es más fácil de predecir por quienes debemos biddear o no una subasta.

A grandes rasgos estos serían los insights más importantes que encontramos, pero, como se indicó anteriormente, tenemos más información a lo largo de todo el análisis que podría ser relevante.