75.06/95.58 Organización de Datos

Primer Cuatrimestre de 2019

Trabajo Práctico 1: Enunciado

GRUPO 12

Alumnos Oyentes:

Ximena Lisouski

Antonella Mazzeo

Objetivos del trabajo

En primer lugar se realizará un análisis exploratorio de los data sets recibidos para comprender mejor su composición. Se recibieron cuatro sets de datos:

- Auctions: Contempla una muestra de las subastas publicitarias en las que Jampp pudo o no haber participado.
- Events: Contiene el detalle de eventos que se realizan en las aplicaciones pertenecientes a clientes de Jampp.
- Clicks: Contempla el detalle de los clicks recibidos en las publicidades mostradas por los clientes de Jampp.
- Installs: Contiene el detalle de instalaciones de las aplicaciones correpsondientes a clientes de Jampp.

Luego se intentarán responder algunas preguntas interesantes, como las siguientes:

- ¿Los set de datos se encuentran relacionados?
- ¿La distribución de determinadas variables se mantiene en el tiempo?
- ¿Existe un conjunto de dispositivos con gran cantidad de apariciones en la muestra de subastas? ¿Eso podría atribuirse a fraude de la industria?
- ¿El tiempo entre subastas de un mismo dispositivo puede explicarse con alguna variable disponible en el data set de subastas?
- ¿Los eventos atribuidos cuentan con algunas características particulares, distintas al resto de los eventos?
- ¿Existen subastas que luego pudieron convertirse en instalaciones? ¿Qué características tienen?
- ¿Existen clicks que luego se convirtieron en instalaciones? ¿Qué características tienen?
- ¿Existen dispositivos con muchas instalaciones? ¿Eso podría asociarse a fraude?
- ¿Existe un tipo de aplicaciones que suelen ser las más instaladas?
- ¿Existen eventos que terminaron en instalaciones? ¿Qué características tienen?

Revisión de la información recibida

Conteos Iniciales

Data_set	Cantidad_Columnas	Cantidad_Registros	Cantidad_Dispositivos	Prom_Repeticiones_Disp
Auctions	7	19571319	206977	94.6
Clicks	20	26351	17119	1.5
Events	22	2494423	196049	12.7
Installs	18	3412	3008	1.1

Contenido de cada set de datos

Auctions

variable	class	distinctValues	missingValues	perMissingValues	sampleValue
auction_type_id	logical	1	19571319	100	NA
country	numeric	1	0	0	6333597102633388032
date	factor	19570963	0	0	18128643
device_id	numeric	206977	0	0	2744133326590046720
platform	integer	2	0	0	1
ref_type_id	integer	2	0	0	7
source_id	integer	5	0	0	0

Clicks

variable	class	distinctValues	missingValues	perMissingValues	sampleValue
advertiser_id	integer	7	0	0.00	3.000000
action_id	logical	1	26351	100.00	NA
source_id	integer	11	0	0.00	0.000000
created	factor	26347	0	0.00	26238.000000
country_code	numeric	1	0	0.00	6333597102633388032.000000
latitude	numeric	78	0	0.00	1.205689
longitude	numeric	81	0	0.00	1.070234
wifi_connection	factor	1	0	0.00	1.000000
carrier_id	numeric	56	11	0.04	4.000000
trans_id	factor	26351	0	0.00	19335.000000
os_minor	numeric	29	12	0.05	4213391244771276800.000000
agent_device	numeric	191	23108	87.69	NA
os_major	numeric	13	12	0.05	3072849339937028096.000000
specs_brand	numeric	5	0	0.00	392184377613097984.000000
brand	numeric	14	20116	76.34	NA
timeToClick	numeric	17295	3374	12.80	NA
touchX	numeric	1002	3340	12.68	0.632000
touchY	numeric	3915	3340	12.68	0.665000
ref_type	numeric	4	0	0.00	1891515180541284352.000000
ref_hash	numeric	17119	0	0.00	4938201217282053120.000000

Events

variable	class	distinctValues	missingValues	perMissingValues	sampleValue
date	factor	2488829	0	0.00	1871549
event_id	integer	568	0	0.00	112
ref_type	numeric	2	0	0.00	1891515180541284352
ref_hash	numeric	196049	0	0.00	336696677951535936
application_id	integer	269	0	0.00	120
attributed	factor	2	0	0.00	1
device_countrycode	numeric	1	0	0.00	6333597102633388032
device_os_version	numeric	82	1472357	59.03	NA
device_brand	numeric	251	1329460	53.30	308305860557778688
device_model	numeric	2625	87967	3.53	5990116681709080576
device_city	numeric	128	1879725	75.36	NA
session_user_agent	numeric	1461	11786	0.47	3819516403548393984
trans_id	factor	14	0	0.00	1
user_agent	numeric	5112	1102896	44.21	NA
event_uuid	factor	2489325	0	0.00	2440827
carrier	numeric	85	1877989	75.29	NA
kind	numeric	584	5099	0.20	4647948847353586688
device_os	numeric	5	1836756	73.63	NA
wifi	factor	3	0	0.00	1
connection_type	factor	4	0	0.00	1
ip_address	numeric	285212	0	0.00	726770255695587456
device_language	numeric	187	87819	3.52	6977049253562485760

Installs

variable	class	distinctValues	missingValues	perMissingValues	sampleValue
created	factor	3412	0	0.00	336
application_id	integer	31	0	0.00	1
ref_type	numeric	2	0	0.00	1891515180541284352
ref_hash	numeric	3008	0	0.00	4130243476992130048
click_hash	logical	1	3412	100.00	NA
attributed	factor	1	0	0.00	1
implicit	factor	2	0	0.00	2
device_countrycode	numeric	2	0	0.00	2970470518450881024
device_brand	numeric	28	2365	69.31	NA
device_model	numeric	416	1	0.03	6882414520414359552
session_user_agent	factor	13	0	0.00	4
user_agent	factor	335	0	0.00	258
event_uuid	factor	866	0	0.00	703
kind	factor	21	0	0.00	1
wifi	factor	3	0	0.00	3
trans_id	factor	5	0	0.00	1
ip_address	numeric	2717	0	0.00	7833422721300884480
device_language	numeric	31	34	1.00	6977049253562485760

El set de datos con más cantidad de registros es “Auctions” (casi 20 millones de filas), dado que comtempla todas las subastas abiertas por advertisers. Sin embargo, cuenta con la menor cantidad de columnas, de las cuales una está completamente vacía (“auction_type_id”), una contiene siempre el mismo valor (“country”), y hay dos columnas que contienen exactamente la misma información (“platform” y “ref_type_id”). Luego de esta limpieza de columnas, restan sólo 4 para analizar. La columna que identifica a cada dispositivo con un id único es “device_id”.

El set de datos que sigue en cuanto a volumen de registros es “Events”, con casi 2.5 millones de filas y 22 columnas. La columna que identifica a cada dispositivo con un id único es “ref_hash”.

“Clicks” e “Installs” son los datasets con menor volumen de datos, y cuentan con 20 y 18 columnas respectivamente. En ambos casos, la columna que identifica a cada dispositivo con un id único es “ref_hash”

Relación entre device_id y ref_hash en los distintos set de datos

Analizando la cantidad de dispositivos que surgen del set de datos “Auctions” que se encuentran en el resto de los data sets, se observan los siguientes resultados:

device_id de Auctions en ref_hash de Clicks	Freq	Freq_Rel
FALSE	189998	0.92
TRUE	16979	0.08

device_id de Auctions en ref_hash de Events	Freq	Freq_Rel
FALSE	122143	0.59
TRUE	84834	0.41

device_id de Auctions en ref_hash de Installs	Freq	Freq_Rel
FALSE	205983	0.9952
TRUE	994	0.0048

El data set “Events” es el que contiene mayor cantidad de dispositivos que han surgido de “Auctions”. El set de datos “Installs” es el que menor cantidad de dispositivos de “Auctions” tiene.

Analizando la combinación de los cuatro data sets, se observan las siguientes cantidades y frecuencias relativas:

Data_set	Cantidad_Dispositivos	Cant_En_Auctions	Cant_En_Clicks	Cant_En_Events	Cant_En_Installs
Auctions	206977	206977	16979	84834	994
Clicks	17119	16979	17119	1196	7
Events	196049	84834	1196	196049	2708
Installs	3008	994	7	2708	3008

Data_set	Cant_En_Auctions	Cant_En_Clicks	Cant_En_Events	Cant_En_Installs
Auctions	1.0000000000	0.0820332694	0.4098716282	0.0048024660
Clicks	0.9918219522	1.0000000000	0.0698638939	0.0004089024
Events	0.4327183510	0.0061005157	1.0000000000	0.0138128733
Installs	0.3304521277	0.0023271277	0.9002659574	1.0000000000

Algunas conclusiones / interrogatorios:

- El 99% de los dispositivos de “clicks”, 43% de “events” y 33% de “installs” se encuentran en “auctions”. Es decir, el complemento de esos porcentajes (1%, 57% y 67% respectivamente) corresponde a dispositivos que no se encuentran contemplados en la base de “Auctions”.
- El 7% de los dispositivos que se encuentran en “Clicks”, también se encuentran en “Events”.
- El 1,3% de los dispositivos de “Events” también se encuentran en “Installs”. ¿Hay alguna relación?
- Sólo el 0,04% de los dispositivos que se encuentran en “Clicks”, también están en “Installs”. ¿Esto no es raro?

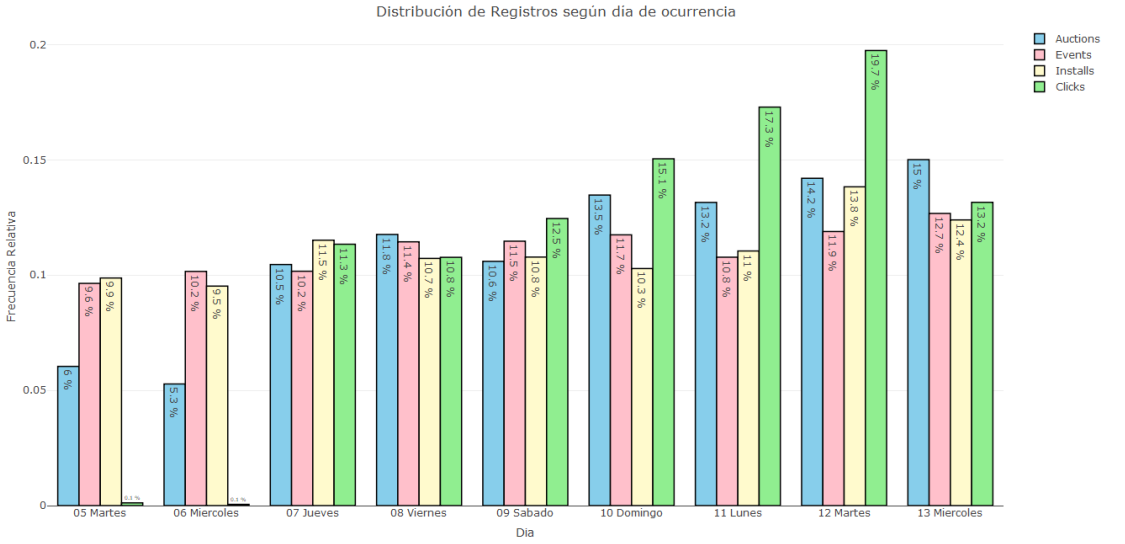
Distribución de registros según día y hora de ocurrencia

Todos los registros de los cuatro data sets tienen fecha desde el 5/3/2019 hasta el 13/03/2019.

Distribución por día:

A continuación se expone la distribución de los registros por día en cada dataset:

Nro_Dia	Freq_auctions	Freq_events	Freq_installs	Freq_clicks
05 Martes	1182401	240549	337	31
06 Miercoles	1032970	253505	325	14
07 Jueves	2047661	253706	393	2989
08 Viernes	2303002	285535	366	2839
09 Sabado	2074552	286221	368	3283
10 Domingo	2637534	293091	351	3966
11 Lunes	2574916	268884	377	4557
12 Martes	2779910	296665	472	5204
13 Miercoles	2938373	316267	423	3468

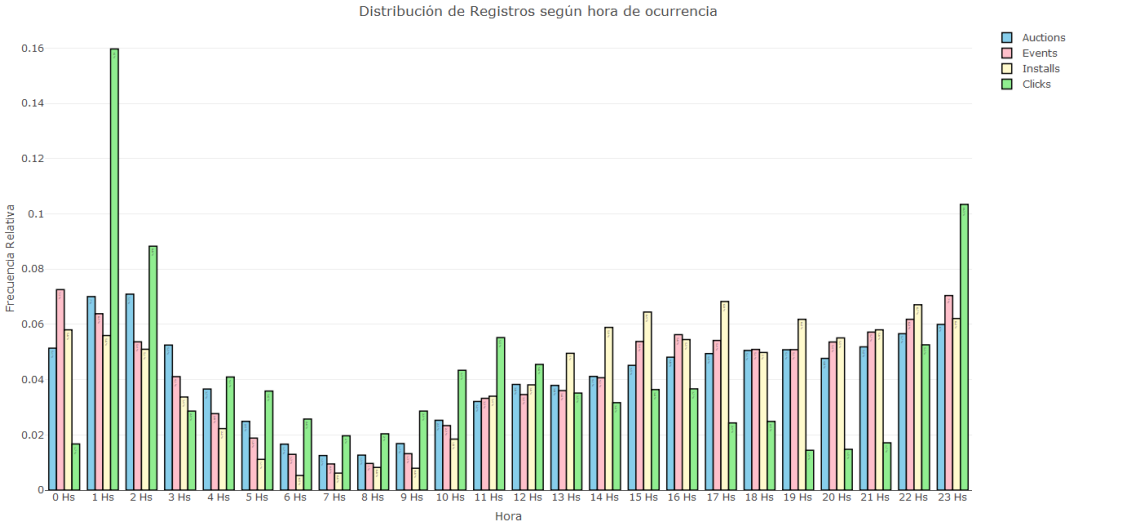


Algunas conclusiones iniciales:

- La frecuencia de los primeros dos días en “Clicks” es extremadamente baja. ¿No es raro?
- “Events” e “Installs” cuentan con distribuciones de registros muy similares en función al día de ocurrencia, siendo siempre bastante pareja la cantidad de registros por día. ¿Están relacionados?
- “Auctions” cuenta con una tendencia creciente, casi triplicando la frecuencia relativa en los días más recientes (15% último día, vs 6% el día inicial). Esto indicaría que un incremento en las subastas no necesariamente genera eventos o instalaciones.

Distribución por hora:

Hora	Freq_auctions	Freq_events	Freq_installs	Freq_clicks
0 Hs	1005716	181072	198	440
1 Hs	1371091	159288	191	4209
2 Hs	1388464	133891	174	2327
3 Hs	1027541	102469	115	754
4 Hs	716194	69027	76	1079
5 Hs	487243	46961	38	945
6 Hs	325730	32295	18	678
7 Hs	245109	23623	21	518
8 Hs	247915	24076	28	537
9 Hs	329604	32899	27	754
10 Hs	494726	58244	63	1143
11 Hs	627907	82888	116	1455
12 Hs	748935	86290	130	1199
13 Hs	741996	89836	169	926
14 Hs	805579	101452	201	834
15 Hs	883824	134192	220	959
16 Hs	941866	140399	186	966
17 Hs	967539	135168	233	641
18 Hs	989528	127056	170	655
19 Hs	994381	126855	211	380
20 Hs	933318	133759	188	389
21 Hs	1015053	142695	198	451
22 Hs	1108219	154274	229	1386
23 Hs	1173841	175714	212	2726

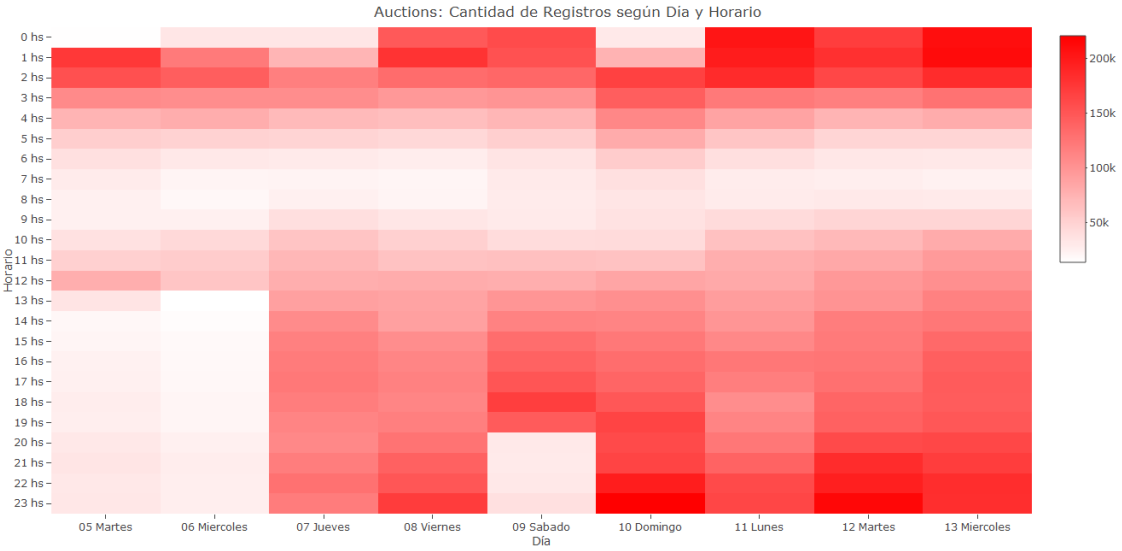


El data set “Clicks” es el que cuenta con mayor diferencia en la distribución de registros por hora respecto al resto de los set de datos, contando con los picos máximos de frecuencia relativa a las 23hs, 1hs y 2hs. Algo extraño que ocurre es que a las 0hs se observa una baja importante en la frecuencia relativa. Quizás esto se encuentre relacionado con la baja frecuencia en los primeros días de análisis, tal como se muestra en el apartado anterior.

Los horarios de 4 a 10 am suelen ser los menos frecuentes.

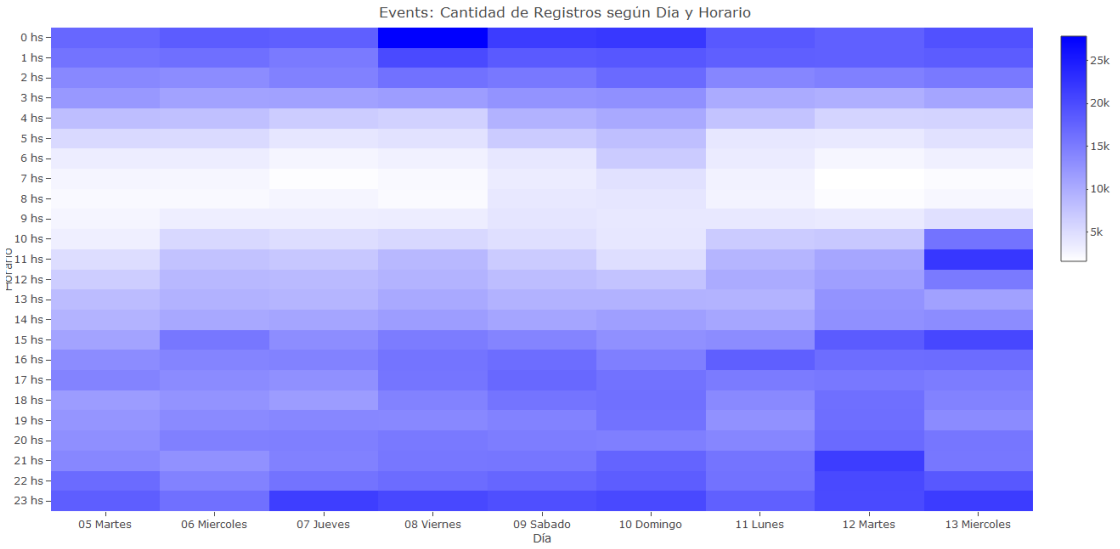
Evolución de cantidad de registros por día y horario:

Auctions:



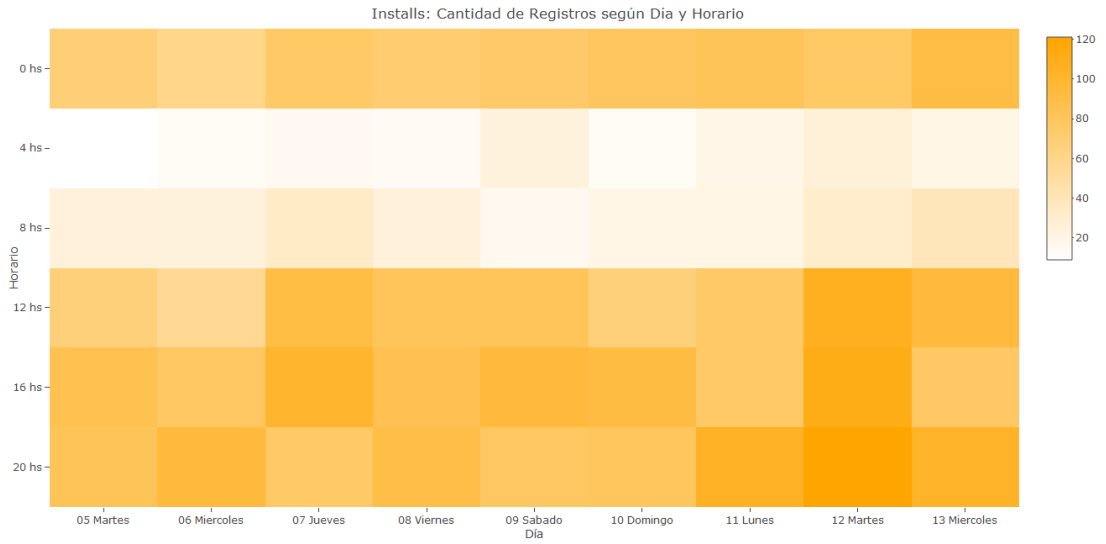
La menor cantidad de subastas en general es entre las 4am y las 12am. Sin embargo, el domingo 10/04 a las 4 am hubo una gran cantidad de subastas, dado que se trata de un fin de semana.

Events:



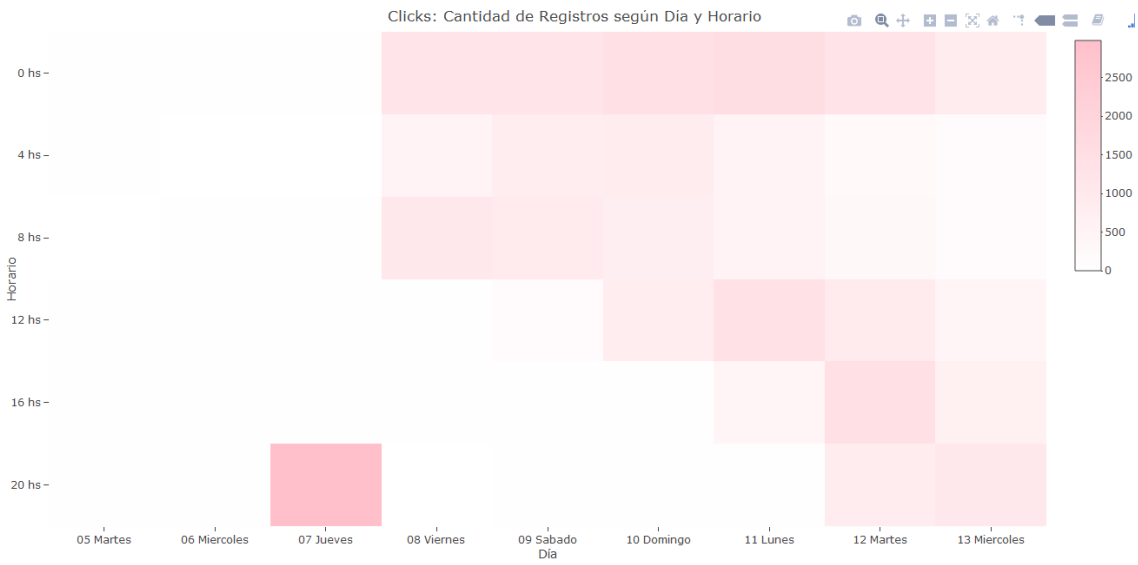
La menor cantidad de events se suelen dar entre las 5 am y las 9 am. Los sábados y domingos en la madrugada se puede observar que se incrementa un poco en la cantidad de eventos respecto del resto de los días (entre 4am y 6 am).

Installs:



Dado que se cuenta con una cantidad de registros mucho menor, se agrupan los horarios en 6 bloques de 4 hs cada uno (0-3hs / 4-7 hs / 8-11 hs / 12-15 hs / 16-19 hs / 20-23hs). En la franja de 4 a 7 hs se observa siempre la menor cantidad de installs, seguida por la franja de 8 a11 hs..

Clicks:



Dado que se cuenta con una cantidad de registros mucho menor, se agrupan los horarios en 6 bloques de 4 hs cada uno (0-3hs / 4-7 hs / 8-11 hs / 12-15 hs / 16-19 hs / 20-23hs).

Los primeros dos días casi no hay registros.

Se empieza a observar un volumen interesante de clicks a partir del jueves 7 de marzo en la franja 20-23 hs.

El viernes y sábado la menor cantidad de clicks se concentra en la franja de 12 a 23 hs (esto pareciera ser raro). Domingo y lunes se observa la menor frecuencia de clicks en la franja 16 a 23 hs.

Exploración de data set “Auctions”

Análisis de dispositivos distintos (columna “device_id”)

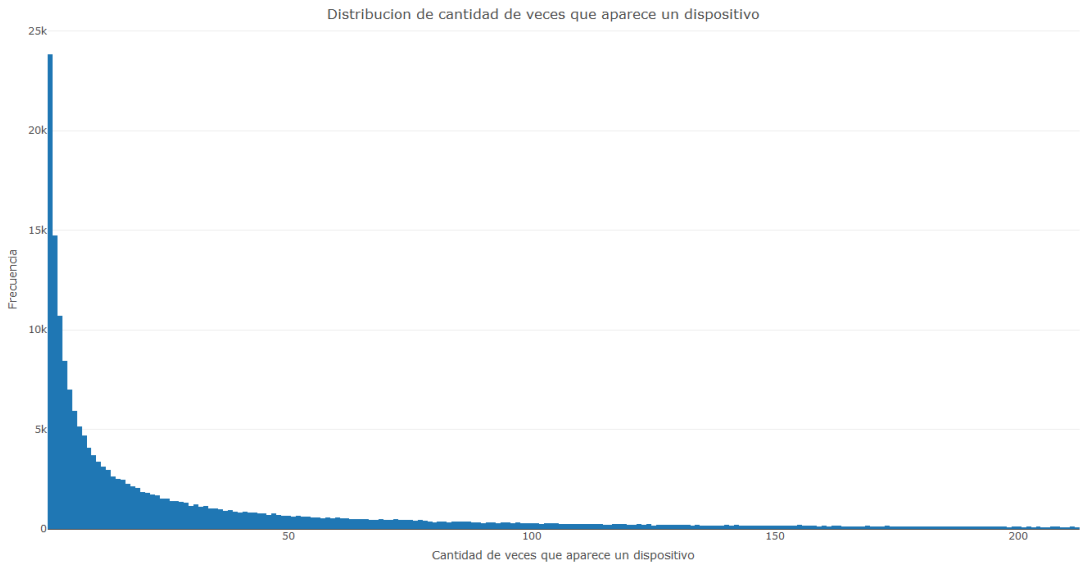
Haciendo un conteo por id de dispositivos (existen 206.977 dispositivos distintos), se observa que existen dispositivos que aparecen una única vez (con estos no podríamos calcular la diferencia de tiempo entre una subasta y otra), así como también hay dispositivos que aparecen más de 27.000 veces en el set de datos:

Metrica	Valores
Minimo	1.00
Perc_25	4.00
Perc_50	16.00
Promedio	94.56
Perc_75	67.00
Maximo	27762.00
Desvio_Std	326.85
Coef_Variac	3.46

El coeficiente de variación de la frecuencia asociada a cada dispositivo es de 346%, lo cual indica que la cantidad de veces que se repite cada dispositivo en el set de datos es muy variada.

Además, existen algunos outliers que generan que el promedio sea incluso más elevado que el percentil 75. Esto indicaría que unos pocos dispositivos aparecen una cantidad de veces muy amplia respecto a la mayoría.

A continuación se expone un histograma de las frecuencias observadas para cada dispositivo (se grafica hasta el percentil 90 para evitar que el gráfico se aplane por los outliers), observando que la mayor parte de los registros cuentan con frecuencias menores a 100 (analizando los percentiles de la primera tabla, surge que el 75% de los dispositivos tienen una frecuencia menor o igual a 67):



A continuación se muestra una tabla que muestra la cantidad de dispositivos que cuentan con determinadas frecuencias en el data set de auctions, para profundizar la información que revela el histograma:

Cant_Dispositivos	Frec_Absoluta	Frec_Relativa
1	23859	11.5 %
>1 & <=5	40857	19.7 %
>5 & <=10	23561	11.4 %
>10 & <=20	25377	12.3 %
>20 & <=50	32640	15.8 %
>50 & <=100	21627	10.4 %
>100 & <=250	21428	10.4 %
>250 & <=500	9730	4.7 %
> 500	7898	3.8 %

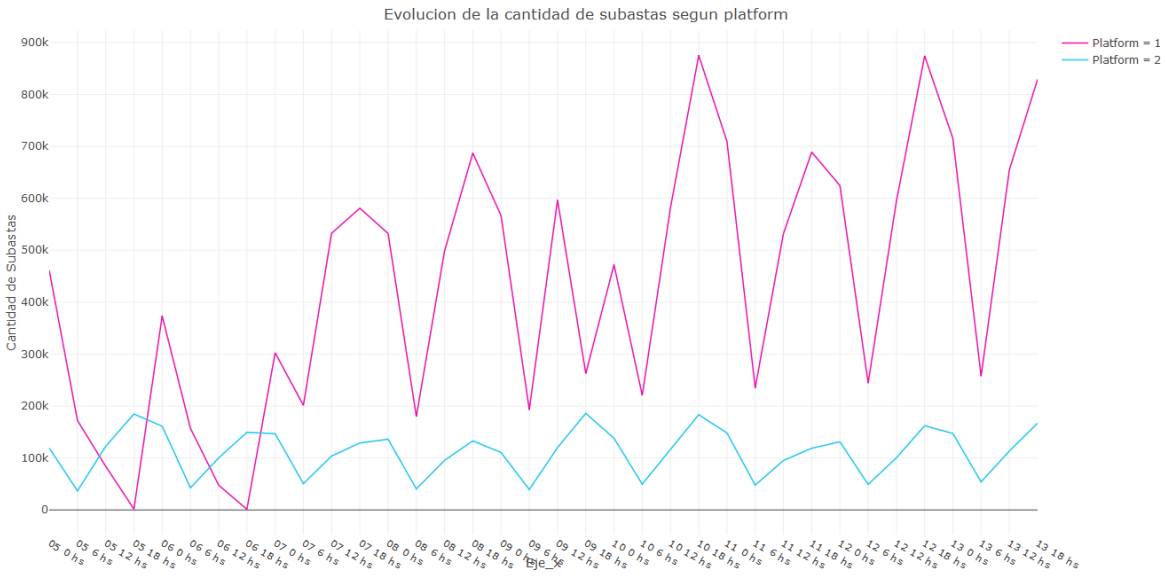
De este análisis surge que el 70% de los dispositivos aparecen en el data set 50 veces o menos. Sin embargo, hay un 3,8% de los dispositivos que aparecen más de 500 veces, lo que podría estar relacionado con eventos de fraude de la industria.

Análisis de columna “Platform”

A continuación se presentan las frecuencias de cada una de las categorías que conforman la columna de referencia:

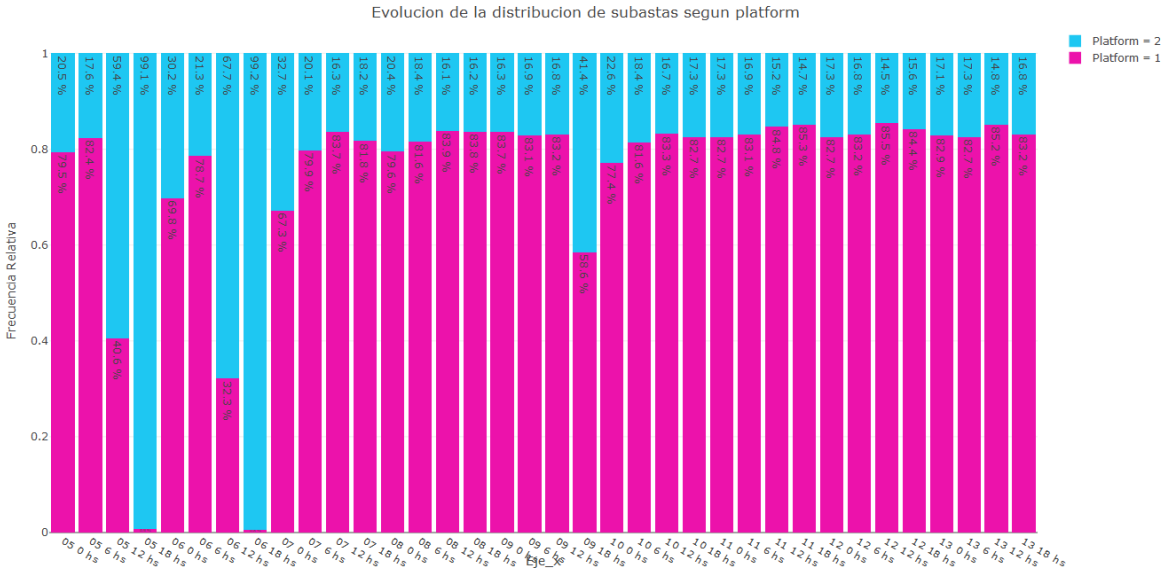
La mayor parte de los registros tienen asignado “Platform=1”.

Se analizó la cantidad de registros a través del tiempo en el dataset auctions para cada una de las categorías de Platform. Para ello, partió cada día en 4 grupos de 6 hs cada uno, a los efectos de detectar alguna tendencia temporal. Los resultados son los siguientes:



Tal como se verificaba en el mapa de calor presentado con anterioridad, los picos mínimos de cantidad de registros se observan por la mañana, y los máximos por la tarde. En general ambas categorías cuentan con la misma tendencia, excepto en algunos horarios de los primeros días en donde se observa que la tendencia cambia (lo cual puede indicar algo raro en los datos).

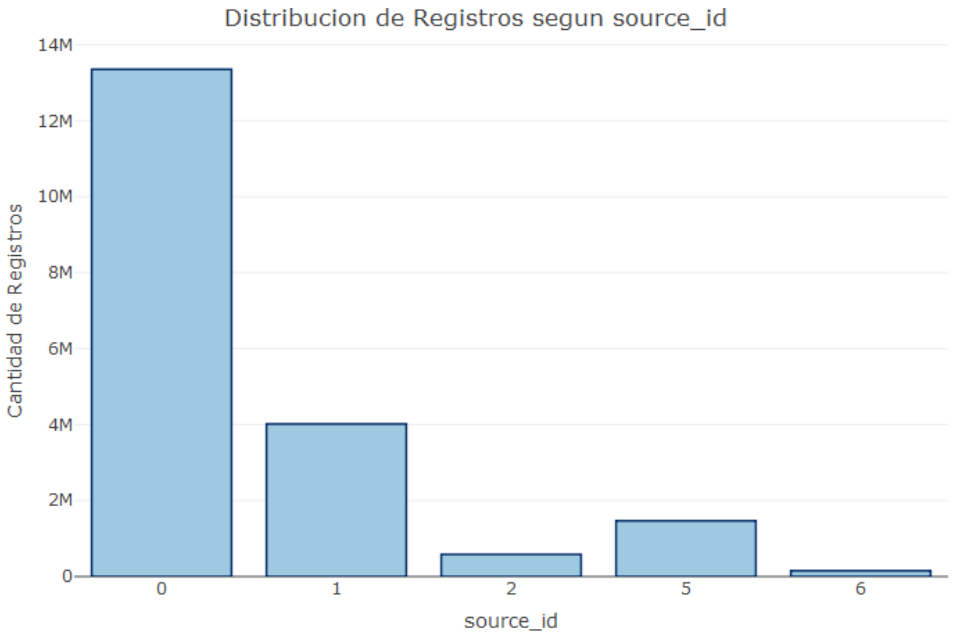
Para profundizar el análisis temporal, se graficó la distribución de los registros en función a la columna “platform” para cada bucket de tiempo predefinido, y se observan los siguientes resultados:



Si bien en general la proporción de registros con platform=2 suele ser menor a la proporción con platform=1, el 5 y 6 de marzo entre las 12 y las 23 hs esta relación se invierte completamente, llegando a valores extremos en el bucket de 19 a 23 hs. Esto podría indicar algún tipo de sesgo en la selección de la muestra que hizo que esta relación se invierta, por lo que se sugiere revisar el origen del dataset.

Análisis de columna “source_id”

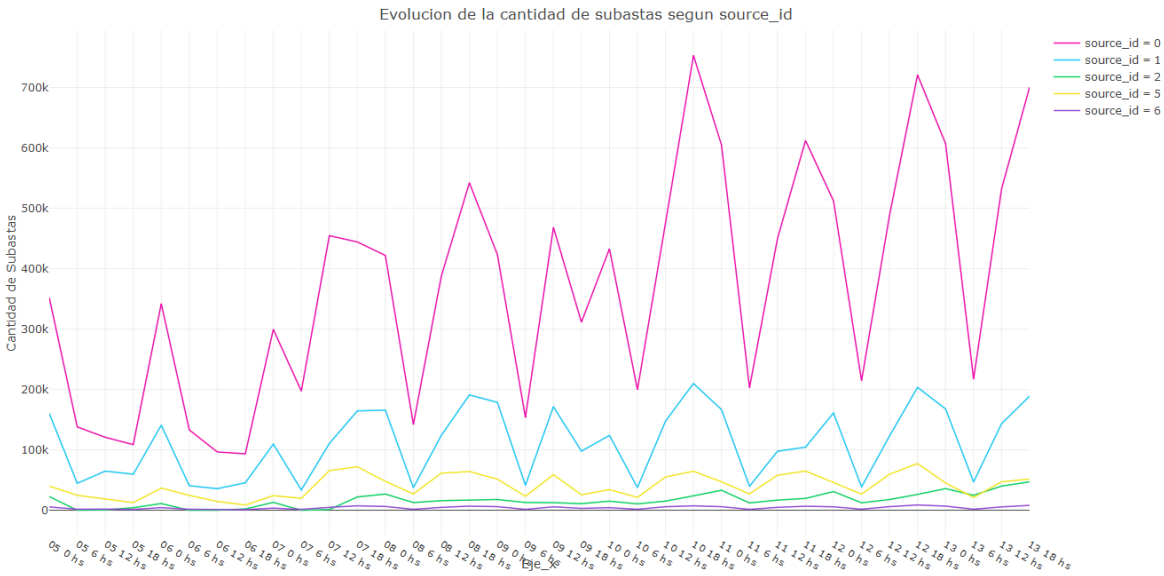
A continuación se presentan las frecuencias de cada una de las categorías que conforman la columna de referencia:



La categoría más frecuente en los registros es la “0”, seguida por la “1”.

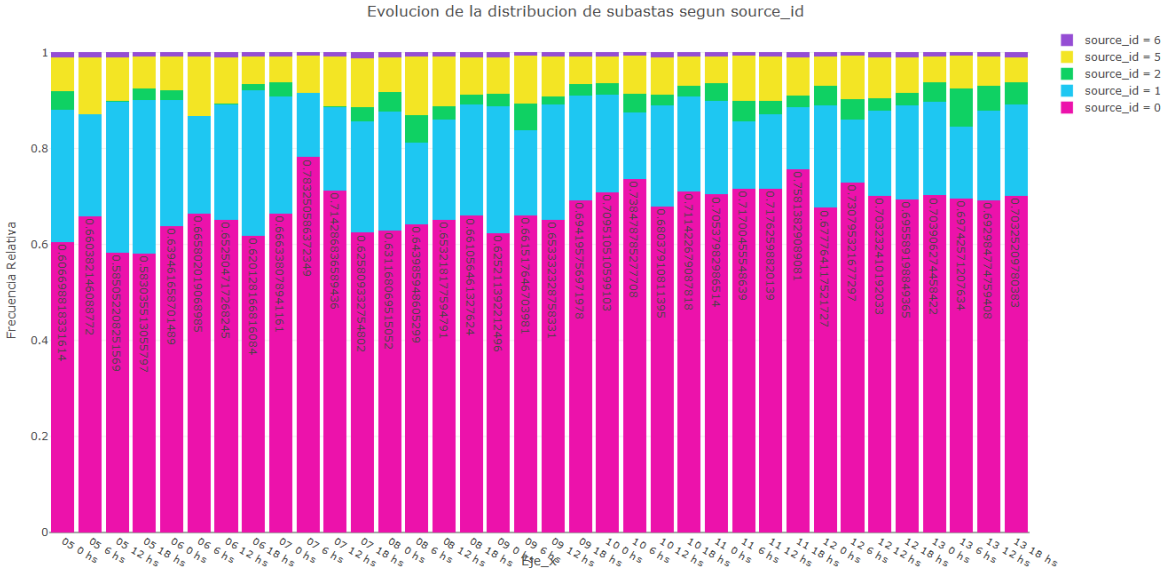
Las categorías “2” y “6” no cuentan con mucho peso relativo en el set de datos (representan un 3,7% del total de registros)

Se analizó la cantidad de registros a través del tiempo en el dataset auctions para cada una de las categorías de source_id. Para ello, partió cada día en 4 grupos de 6 hs cada uno, a los efectos de detectar alguna tendencia temporal. Los resultados son los siguientes:



Al igual que al analizar a la columna platform, los picos mínimos de cantidad de registros se observan por la mañana, y los máximos por la tarde. En general todas las categorías cuentan con la misma tendencia.

Para profundizar el análisis temporal, se graficó la distribución de los registros en función a la columna “source_id” para cada bucket de tiempo predefinido, y se observan los siguientes resultados:



En general la proporción de registros con cada una de las categorías de source_id suele ser similar en el transcurso del tiempo.

Análisis de tiempo entre subastas

Dado que el set de datos tiene estructura de tipo transaccional (un registro, una subasta), hay que realizar algún tipo de manejo para poder calcular el tiempo entre una subasta y la próxima, correspondiente al mismo dispositivo.

Para calcular esta variable se llevó a cabo el siguiente procedimiento:

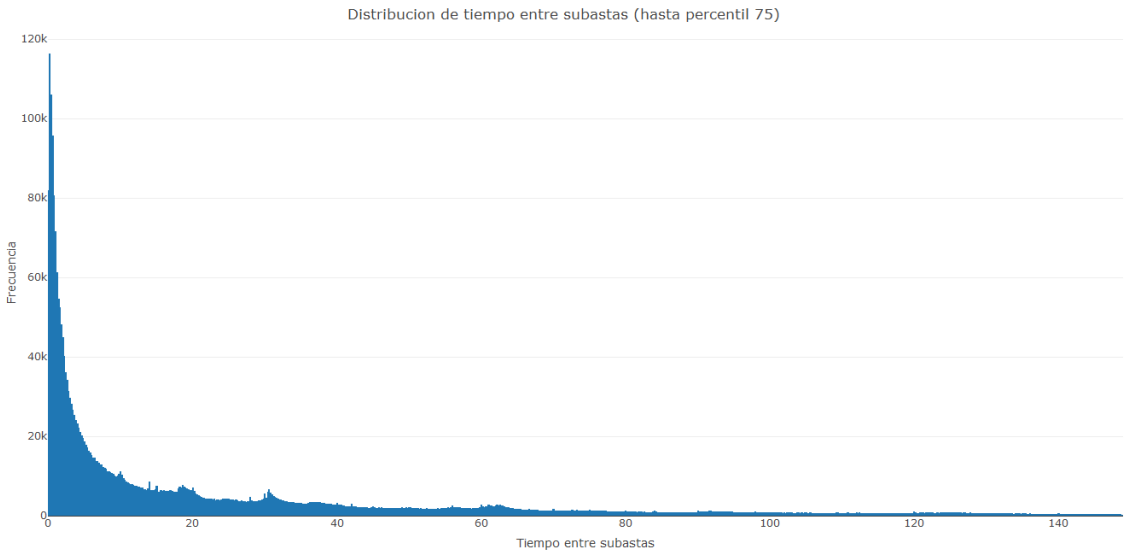
1- Se ordena el data frame de auctions de menor a mayor fecha (columna “date”), para asegurar el orden cronológico de todos los registros.
2- Se ordena el data frame de menor a mayor código de dispositivo (columna “device_id”), para asegurar que los dispositivos iguales queden juntos en el data frame, y además ordenados cronológicamente por el paso anterior.
3- Se calculan dos vectores: device_id y date, ambos corridos una posición para adelante, y se agrega al data frame de auctions, para tener en formato de columna el dato inmediatamente posterior.
4- Se eliminan del data frame aquellos registros en que no se puede calcular la diferencia de tiempo (sea porque el dispositivo aparece una única vez, o porque es el último dispositivo que aparece en el set de datos).
5- Se calcula la diferencia de tiempo entre la fecha original y la fecha corrida una posición, en segundos.

Por cuestiones de falta de tiempo de procesamiento, se tomó una muestra aleatoria de aproximadamente el 20% del total de registros del data frame de Auctions (contó con 3.875.182 registros) para realizar algunos análisis iniciales. Como punto pendiente quedaría replicar este estudio en el set de datos completo. A continuación se comparten algunos resultados:

Metrica	Valores_DifTime_EnSegundos
Minimo	0.000001
Perc_25	2.984298
Perc_50	21.955905
Promedio	3929.704736
Perc_75	148.718067
Maximo	769662.171305
Desvio_Std	23919.876120
Coef_Variac	6.086940

Algunas conclusiones preliminares:

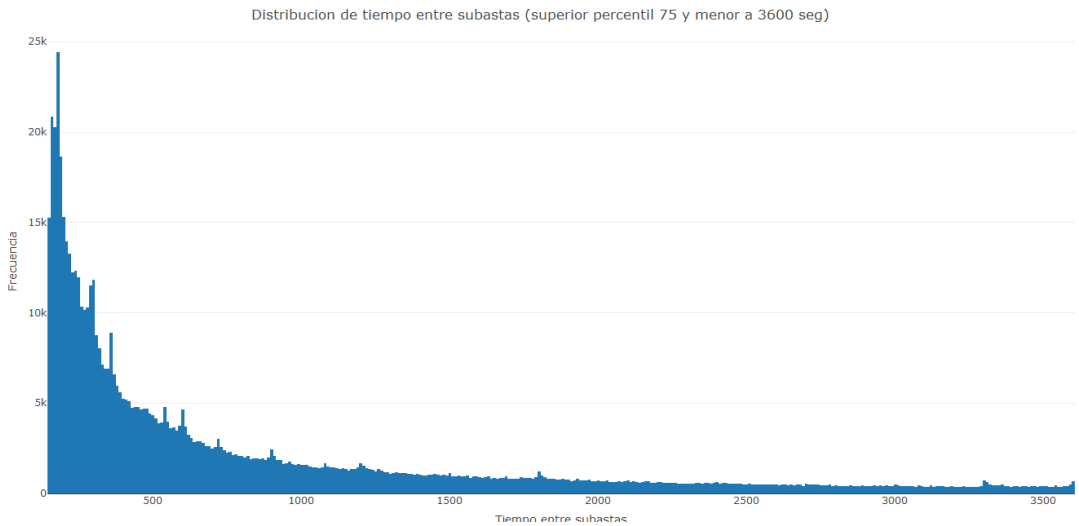
- La diferencia entre el valor mínimo y máximo de la diferencia entre subastas en segundos es muy grande (80704928093725% ¡!).
- La mitad de las subastas tienen un tiempo promedio menor a 22 segundos.
- El 75% de las subastas tienen un tiempo promedio menor a 149 segundos (aprox 2 minutos y medio).
- El coeficiente de variación es muy grande porque existen diferencias abismales (outliers).
- El 90% de las subastas ocurre en menos de 1943 segundos (aprox 25 minutos).



Algunas conclusiones del histograma:

- La mayor parte de las subastas cuentan con poco tiempo de distancia entre las mismas.

¿Cómo se distribuyen los “outliers” (tiempo > percentil 75 y tiempo < 3600 segundos)?:



Analizando los tiempos entre subastas superiores al percentil 75 (149 segundos) y menores a 3600 segundos, se observa una tendencia decreciente en las frecuencias, a medida que se incrementa el tiempo.

Aproximadamente un 8% de la muestra cuenta con un tiempo superior a los 3600 segundos (1 hora).

Sólo un 1% de los registros cuenta con un tiempo superior a las 24 hs entre subastas.

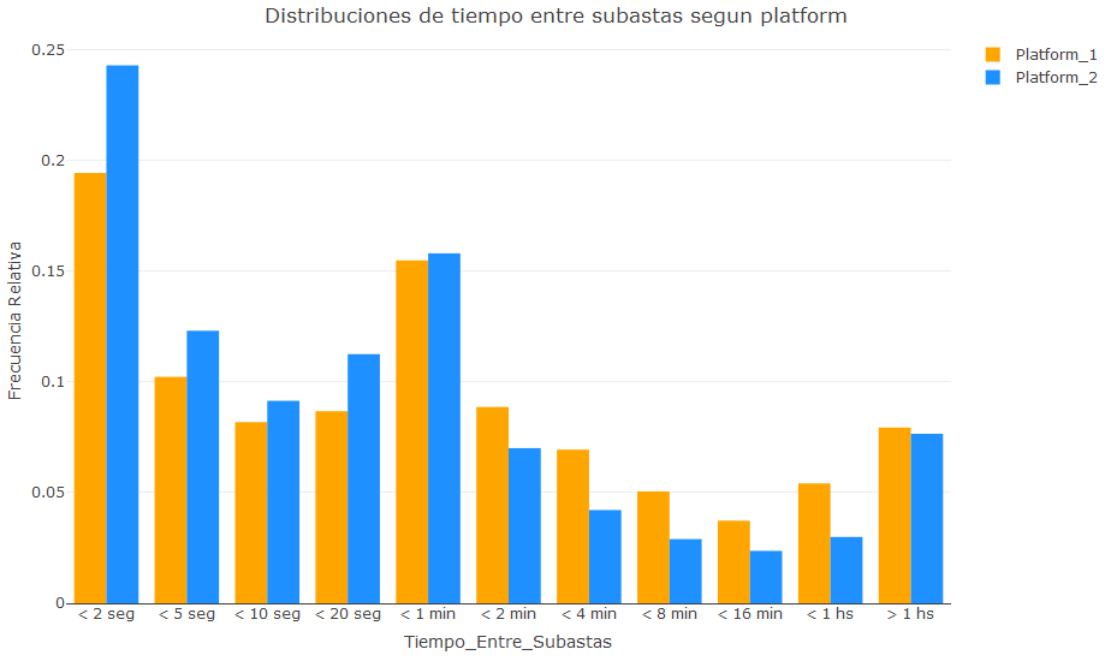
Tiempo entre subastas según Platform:

Platform	Cantidad	Minimo	Media	Pct_25.25.	Mediana	Pct_75.75.	Maximo	Desvio	Coef_Var
1	3077750	0.0000019073486	3822.796	3.327769	26.00594	176.4612	769662.2	23267.58	6.086536
2	797432	0.0000009536743	4342.329	2.099419	13.25966	69.9290	768446.3	26282.02	6.052518

En ambas subpoblaciones se cuenta con amplia variabilidad en el tiempo que transcurre entre subastas (coeficiente de variación arriba de 600%!).

Analizando las medidas de posición:

- La mediana del tiempo entre subastas en platform 2 es la mitad de la mediana en platform 1 (es decir, la mitad de los registros con platform 1 tardó menos de 26 segundos en aparecer en otra subasta, mientras que la mitad de los registros de platform 2 tardó menos de 13 segundos).
- Los percentiles 25 y 75 del tiempo transcurrido hasta la próxima subasta también son inferiores en platform 2 vs platform 1. El 25% de los registros con platform 1 tarda menos de 3,33 segundos en aparecer en una nueva subasta, mientras que en platform 2 tarda menos de 2 segundos.
- A grandes rasgos esta variable parece marcar una diferencia en la definición del tiempo hasta una nueva subasta.



Analizando las distribuciones:

- En general se observa una tendencia decreciente en ambas poblaciones (a mayor tiempo entre subastas, menor frecuencia relativa).
- Sin embargo, vemos que el peso de los registros con tiempo menor a 2 segundos en platform_2 es más elevado que en platform_1.
- Las frecuencias relativas son mayores en platform_2 cuando el tiempo entre subastas es menor (en 1 minuto se igualan), y son menores cuando el tiempo es mayor.
- Podríamos afirmar que hay mayor frecuencia de registros cuyo tiempo entre subastas es menor a un minuto en platform_2 que en platform_1.

Tiempo entre subastas según source_id:

Source_id	Cantidad	Minimo	Media	Pct_25.25.	Mediana	Pct_75.75.	Maximo	Desvio	Coef_Var
0	2652221	0.0000009536743	3100.610	2.988331	21.84486	132.97263	769662.2	20475.44	6.603680
1	793632	0.0000681877136	4568.487	2.497309	13.19172	99.45958	762875.4	27072.23	5.925863
2	110906	0.0000360012054	13518.882	8.453589	104.03835	3392.01118	760800.4	44939.09	3.324172
5	290984	0.0000078678131	3560.683	3.991866	33.26762	389.34254	726783.4	21637.77	6.076860
6	27439	0.0001590251923	30747.938	96.213611	1345.72992	27991.20564	768446.3	69179.28	2.249884

En los source_id 6 y 2 se observan los menores coeficientes de variación (225% y 332% respectivamente), lo que indicaría que pertenecen a la subpoblación con tiempo entre subastas más estable.

Analizando las medidas de posición:

- La mediana del tiempo entre subastas en source_id 6 es muy superior al resto. El 75% de los registros con dicho identificador cuentan con tiempo entre subastas mayor a 96 segundos.
- La siguiente categoría con mayor mediana es source_id = 2. El 50% de dichos registros tiene tiempo entre subastas mayor a los 104 segundos.

- La categoría con menos mediana en el tiempo entre subastas es source_id 1. El 25% de dichos registros cuentan con tiempo entre subastas menor a 4,5 segundos.

Esta variable pareciera ser importante al momento de determinar el tiempo entre subastas, dado que se encuentran amplias diferencias en los tiempos medianos.

Exploración de data set “Installs”

Algunas aclaraciones preliminares:

- La columna “click_hash” se encuentra completamente vacía, por lo que no se va a analizar.
- La columna “atributed” tiene un solo valor posible (“1”), por lo que no se va a analizar.
- La columna “device_brand” cuenta con un alto nivel de vacíos (69%).
- La columna “trans_id” se encuentra vacía en el 99,82% de los registros, por lo que no se va a analizar.

Análisis de dispositivos distintos (columna “ref_hash”)

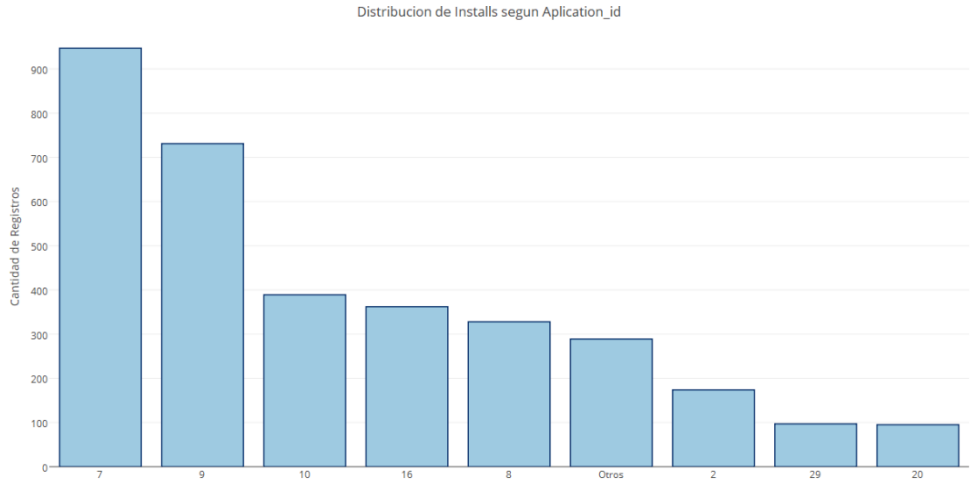
Haciendo un conteo por id de dispositivos (existen 3.008 dispositivos distintos, en 3.412 registros, no hay tantos dispositivos repetidos), se observa que la gran mayoría de los dispositivos se encuentran una única vez en el set de datos:

Metrica	Valores
Minimo	1.00
Perc_25	1.00
Perc_50	1.00
Promedio	1.13
Perc_75	1.00
Maximo	4.00
Desvio_Std	0.37
Coef_Variac	0.32

Sólo 379 dispositivos aparecen en el set de datos más de una vez, de los cuales 2 aparecen 4 veces.

Análisis de la columna “aplication_id”

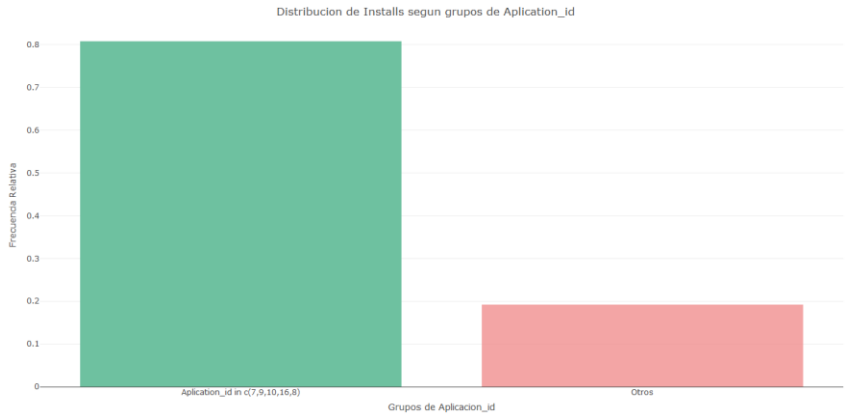
A continuación se presentan las frecuencias de cada una de las categorías que conforman la columna de referencia, agrupando en “Otros” a aquellas categorías con frecuencia menor a 50:



A continuación se expone el top 10 de las categorías agrupadas en “Otros” más frecuentes:

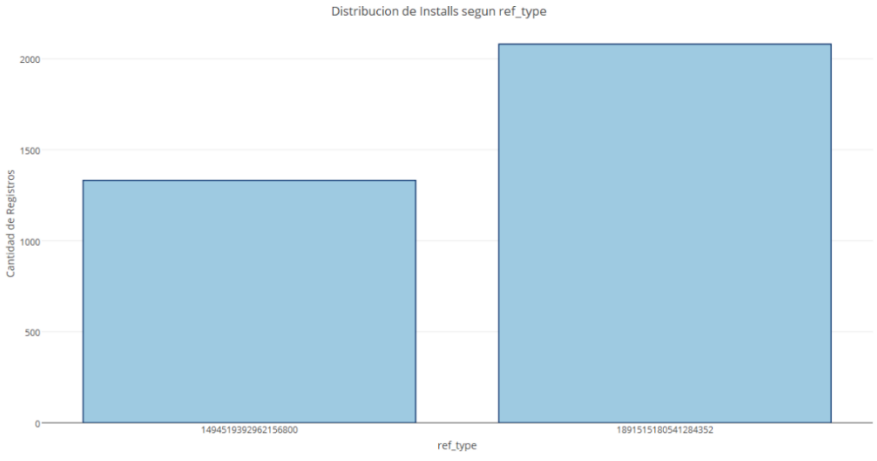
Aplication_id	Freq
6	35
1	34
26	28
34	28
15	20
0	18
3	17
28	17
12	15
18	14

La categoría con mayor peso en el set de datos es la “7”, representando un 28% del total de los registros. En cuestión de relevancia le sigue la categoría “9”, representando un 21% del total de registros, y luego la categoría “10” con un 11%, al igual que la categoría “16”. La categoría “8” representa un 10% del total de registros. Estas cuatro categorías (7, 9, 10, 16 y 8) representan el 81% del total de registros del set de datos:

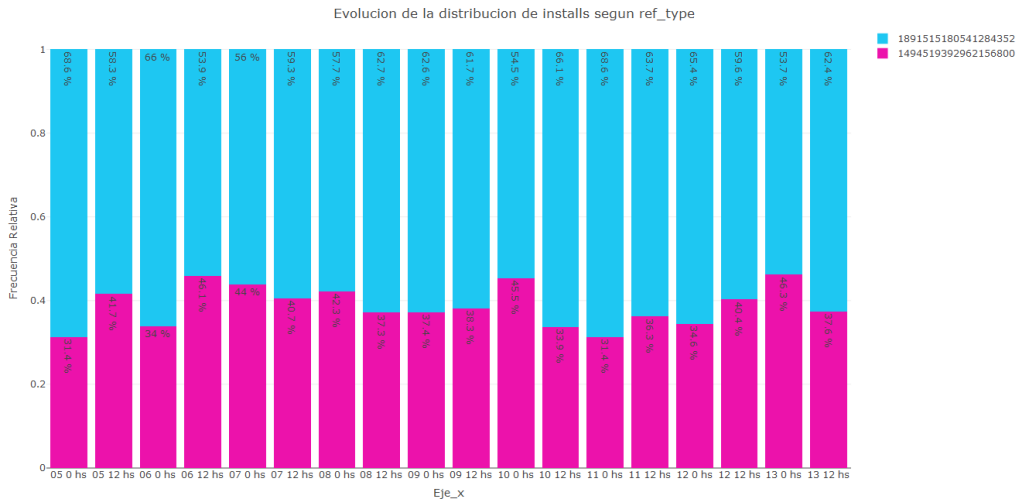


Análisis de la columna “ref_type”

A continuación se presentan las frecuencias de cada una de las categorías que conforman la columna de referencia:



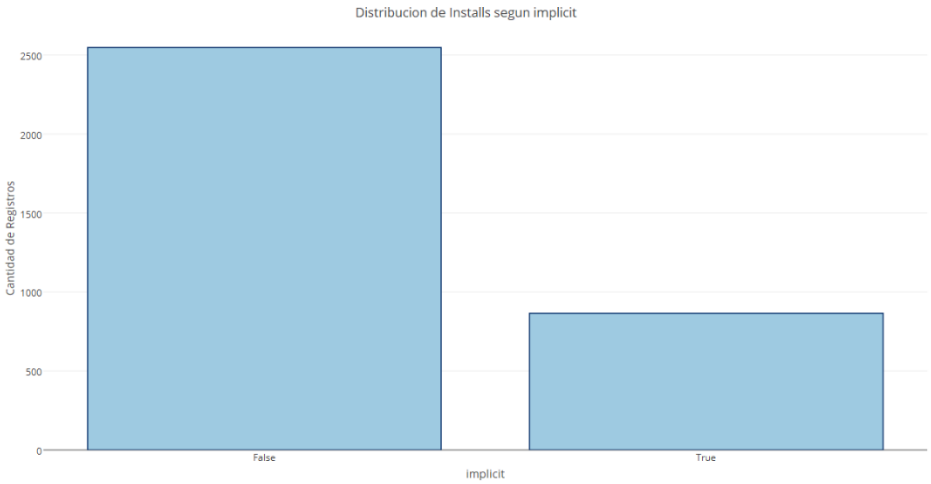
Se analizó la evolución a través del tiempo de la distribución de registros en función a cada categoría que conforma la variable “ref_type”. Para ello, partió cada día en 2 grupos de 12 hs cada uno, a los efectos de detectar alguna tendencia temporal. Los resultados son los siguientes:



En general, se observa que la distribución es similar a lo largo del tiempo.

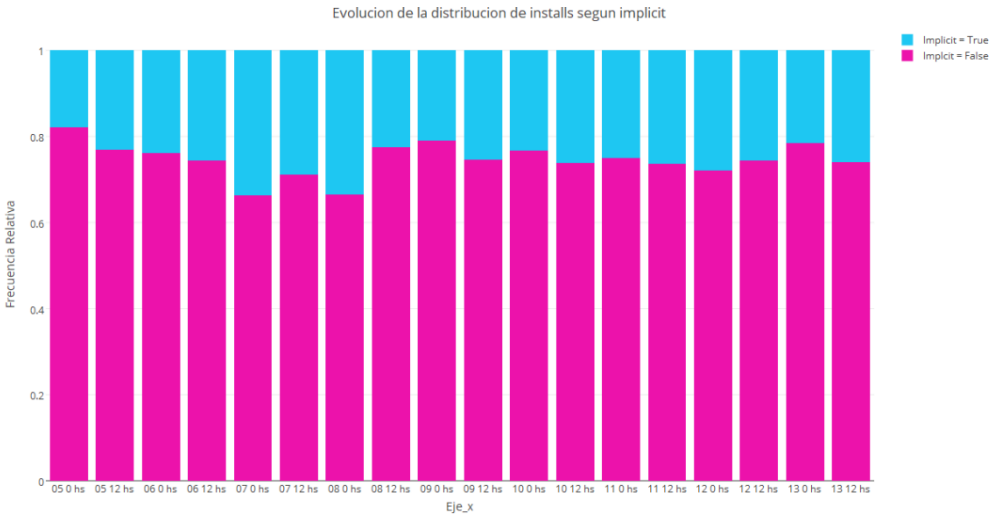
Análisis de la columna “implicit”

A continuación se presentan las frecuencias de cada una de las categorías que conforman la columna de referencia:



La mayoría de los registros se encuentran en categoría “False” (75%), lo que indicaría que la mayoría de las instalaciones no son implícitas.

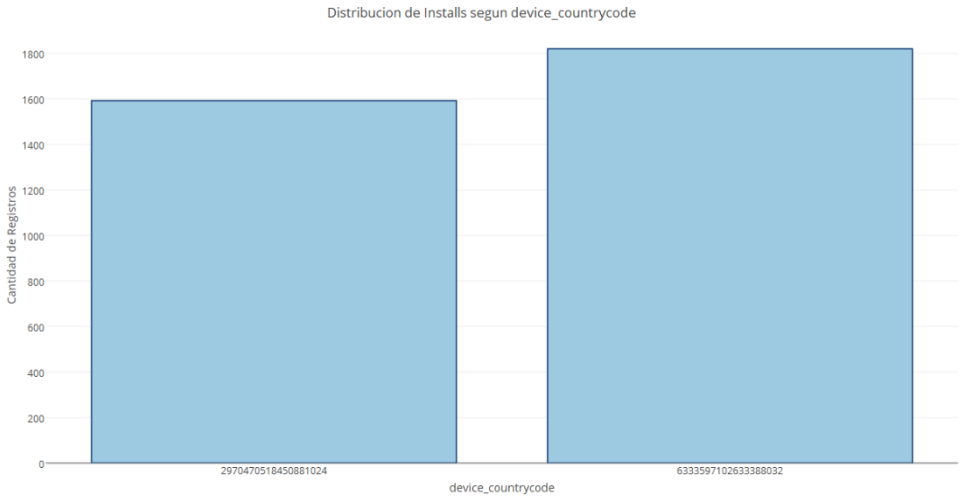
Se analizó la evolución a través del tiempo de la distribución de registros en función a cada categoría que conforma la variable “implicit”. Para ello, partió cada día en 2 grupos de 12 hs cada uno, a los efectos de detectar alguna tendencia temporal. Los resultados son los siguientes:



En general, la proporción de registros con aplicaciones implícitas se mantiene estable en el transcurso del tiempo.

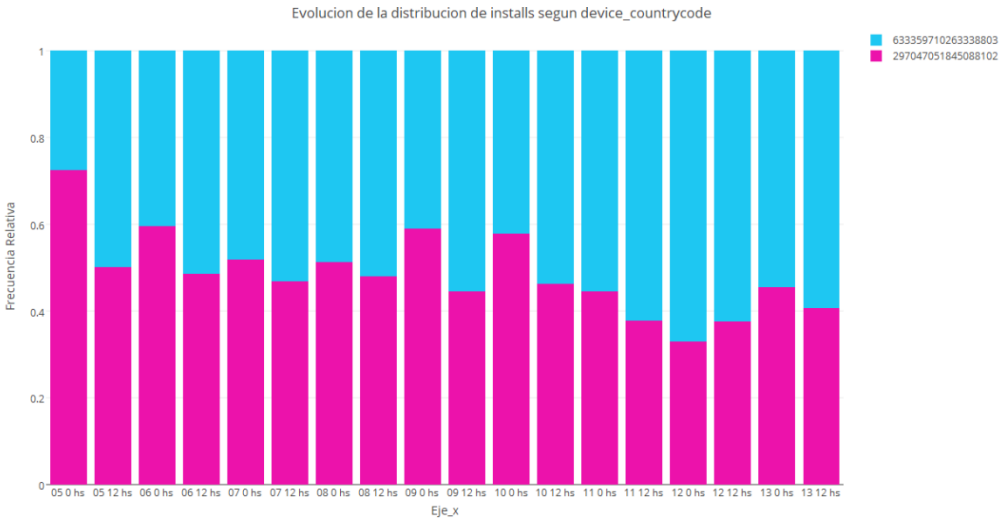
Análisis de la columna “device_countrycode”

A continuación se presentan las frecuencias de cada una de las categorías que conforman la columna de referencia:



Se observa una cantidad de registros similar en cada una de las dos categorías posibles.

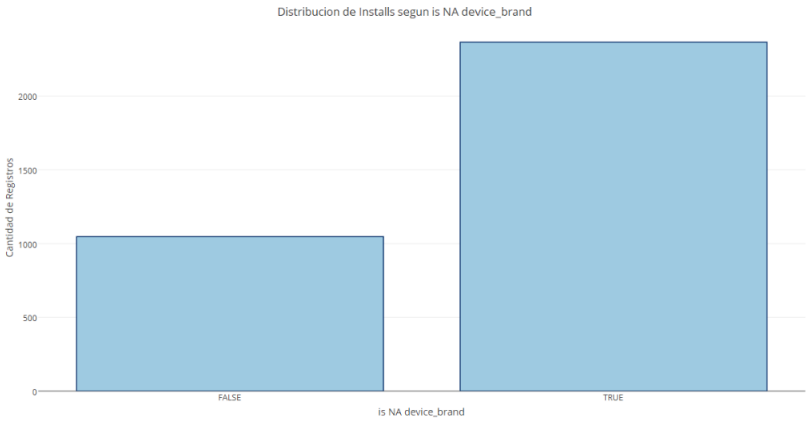
Se analizó la evolución a través del tiempo de la distribución de registros en función a cada categoría que conforma la variable “device_countrycode”. Para ello, partió cada día en 2 grupos de 12 hs cada uno, a los efectos de detectar alguna tendencia temporal. Los resultados son los siguientes:



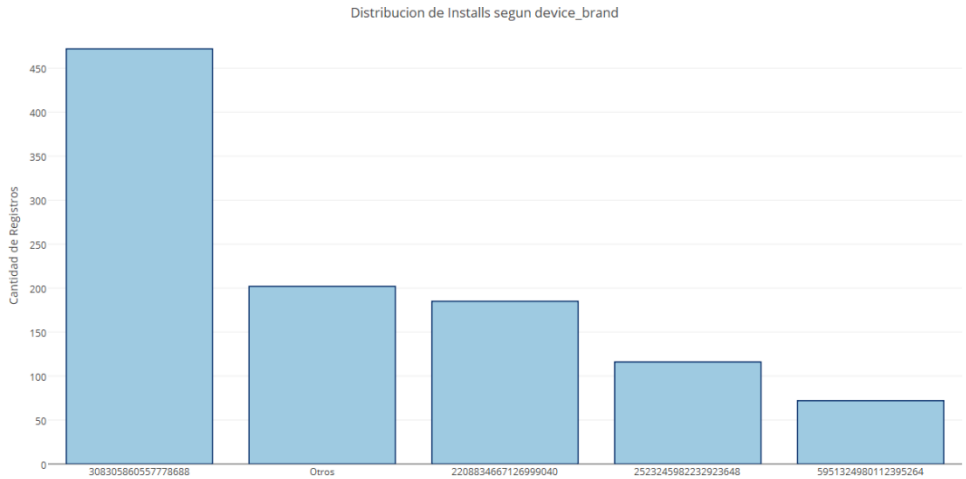
Pareciera que a medida que avanza el tiempo, existe una tendencia creciente en la participación del código que inicia en “6”.

Análisis de la columna “device_brand”

Esta variable cuenta con un elevado nivel de vacíos (69%):



Analizando al 31% de registros con el campo no vacío, se obtienen los siguientes resultados:



En la categoría “Otros” se incluyen aquellas categorías cuya frecuencia es menor a 50. Si bien la categoría más frecuente es “308305860557778688”, cabe señalar que solo aparece en 472 registros (14% del total del set de datos).

Top 10 de las categorías más frecuentes incluidas en “Otros”:

device_brand	Freq
2987569314309514240	40
3812620986737351168	37
513799204676421248	34
6538561794435555328	26
1083368711068077952	9
3228516090903639552	9
4371307750970993152	7
2262848417324907008	5
3093165991971728896	5
3849490426089584640	4

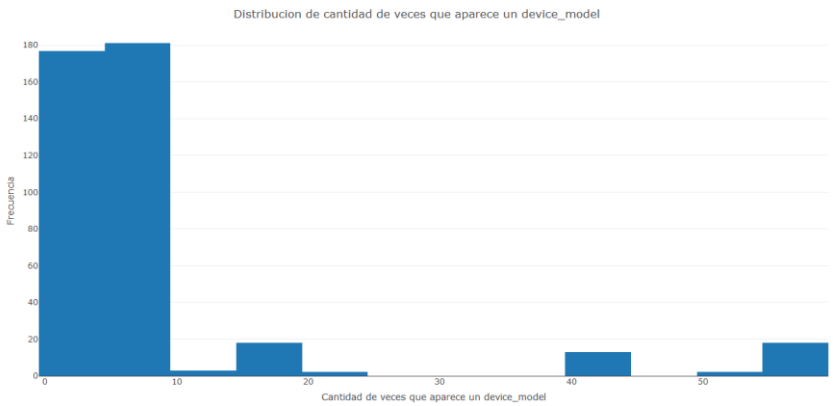
Análisis de la columna “device_model”

Esta variable cuenta con 416 valores distintos, de los cuales el 75% aparecen 6 veces o menos en el set de datos. A continuación se exponen las características de la frecuencia de cada una de las categorías de esta variable:

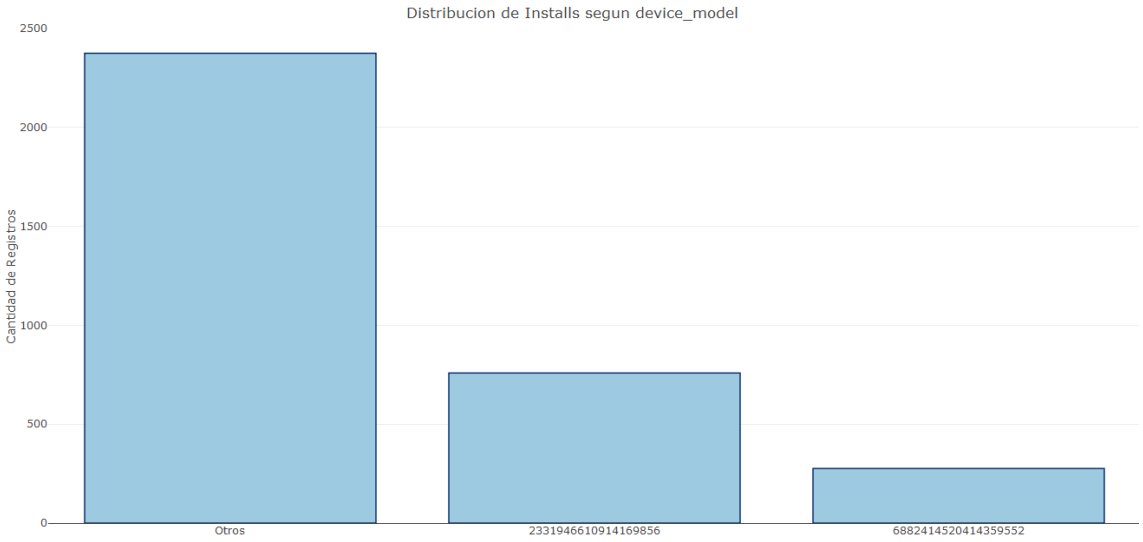
Metrica	Valores
Minimo	1.00
Perc_25	1.00
Perc_50	2.00
Promedio	8.22
Perc_75	6.00
Maximo	759.00
Desvio_Std	40.50
Coef_Variac	4.93

Si bien la gran mayoría de los valores posibles se encuentra en el set de datos muy pocas veces, existe al menos una categoría que aparece 759 veces.

A continuación se presenta el histograma sobre la cantidad de veces en que aparece cada device_model en el set de datos, evidenciando lo mencionado anteriormente en cuanto a qua la mayor parte de las categorías aparecen muy pocas veces en la muestra:



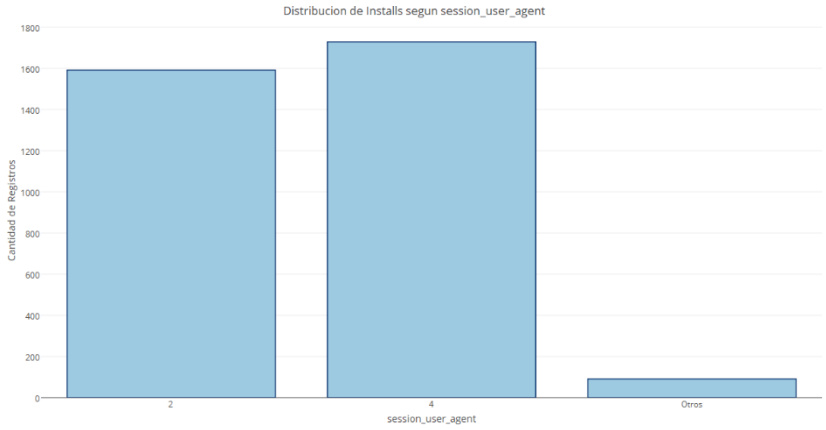
Por otra parte, se expone la cantidad de registros por cada categoría de la variable, siempre y cuando dicha categoría aparezca en al menos 100 registros:



El 70% de los registros cuenta con un device_model que aparece menos de 100 veces en el set de datos. Entre los dos códigos más frecuentes se obtiene el 31% de los registros.

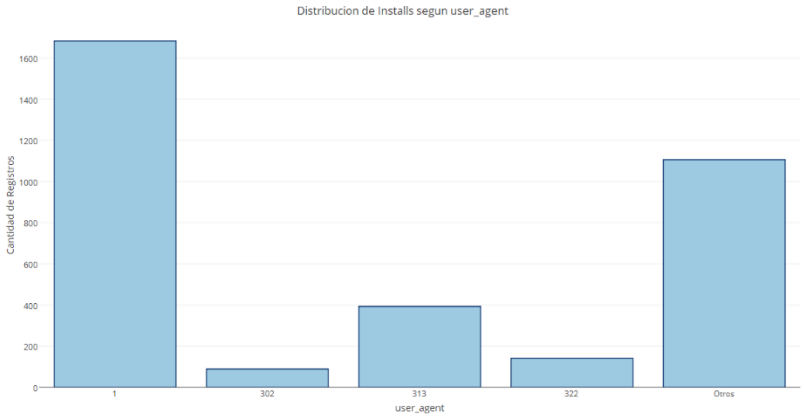
Análisis de la columna “session_user_agent”

A continuación se presentan las frecuencias de cada una de las categorías que conforman la columna de referencia, agrupando en “Otros” a aquellas categorías con frecuencia menor a 50:



Análisis de la columna “user_agent”

A continuación se presentan las frecuencias de cada una de las categorías que conforman la columna de referencia, agrupando en “Otros” a aquellas categorías con frecuencia menor a 50:



Gran parte de los registros (49%) se concentran en la categoría “1”.

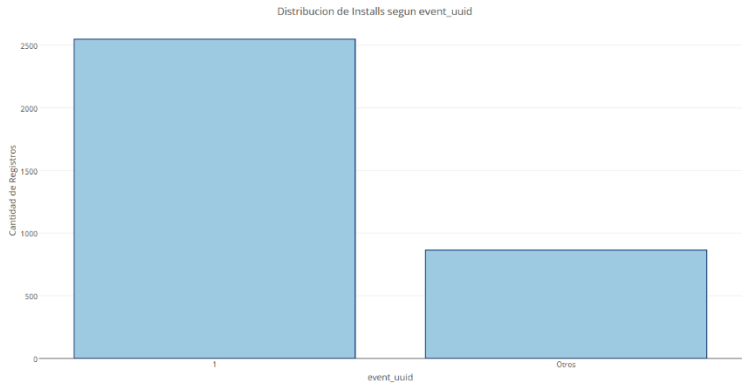
Por otra parte, cruzando la información se “sesión_user_agent” y “usser_agent” se obtienen los siguientes resultados:

	session_user_agent 2	session_user_agent 4	session_user_agent Otros
user_agent 1	1592	0	91
user_agent 302	0	89	0
user_agent 313	0	393	0
user_agent 322	0	141	0
user_agent Otros	0	1106	0

Es decir, la mayor parte de los registros correspondientes a “user_agent 1” tienen como valor de “sesión_user_agent” a la categoría 2. Además, dicha categoría de “sesión_user_agent” sólo se presenta ante el “user_agent 1”.

Análisis de la columna “event_uuid”

A continuación se presentan las frecuencias de cada una de las categorías que conforman la columna de referencia, agrupando en “Otros” a aquellas categorías con frecuencia menor a 2:

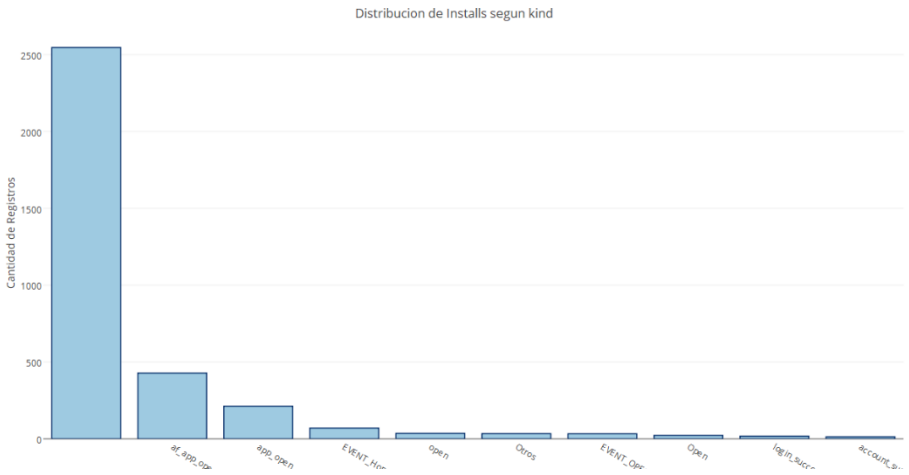


La mayor parte de los registros se concentran en la categoría “1” (75%). Sin embargo, las categorías agrupadas en “Otros” sólo aparecen una vez en el dataset:

	Var1	Freq
2	001b1f1c-3a0c-4668-aa61-5ace47667735	1
3	007ec71f-d7b5-4fb2-811f-21dbe161178f	1
4	0090f110-02e4-4c3d-a514-d95a713793c6	1
5	009d3d77-acba-4cab-a910-aa1229df51cf	1
6	013727e9-9ff4-4f58-a896-39ab88f6e11c	1
7	01614548-ec76-4a69-ace6-ff70b0d4c279	1

Análisis de la columna “kind”

A continuación se presentan las frecuencias de cada una de las categorías que conforman la columna de referencia, agrupando en “Otros” a aquellas categorías con frecuencia menor a 10:

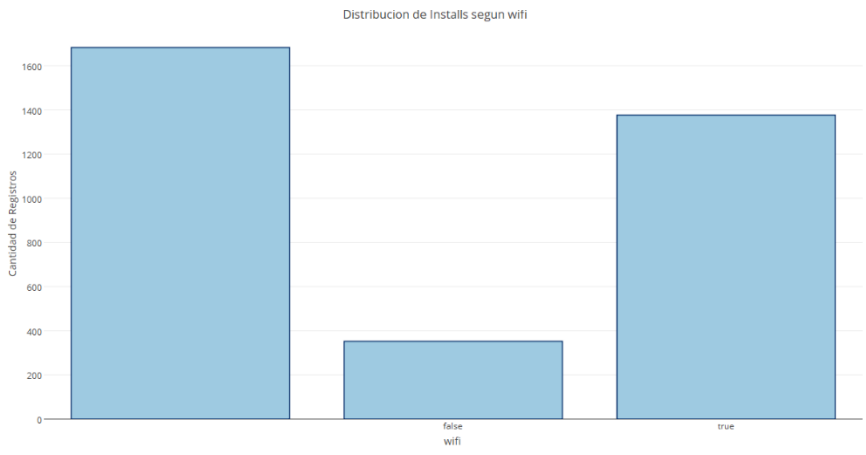


binning	Freq	Freq_Rel
	2547	0.75
af_app_opened	428	0.13
app_open	212	0.06
EVENT_Homepage	70	0.02
open	36	0.01
Otros	34	0.01
EVENT_OPEN_APP	33	0.01
Open	22	0.01
login_success	17	0.00
account_summary_first_step	13	0.00

Sin embargo, la mayor parte de los registros (75%) cuentan con este dato vacío.

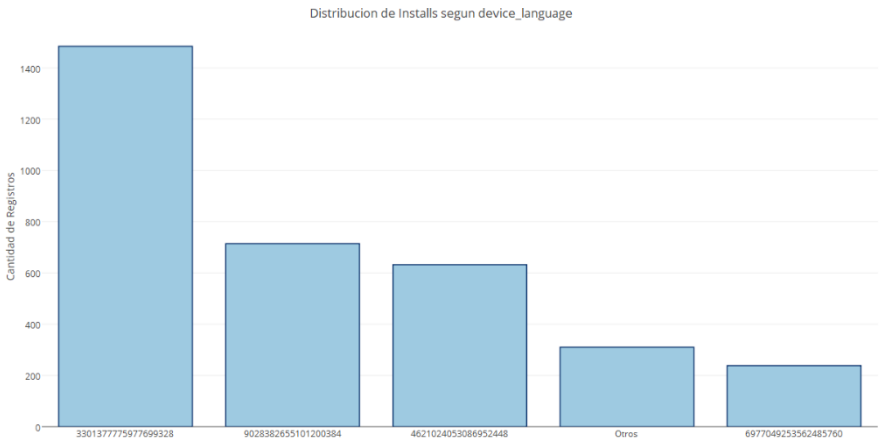
Análisis de la columna “wifi”

A continuación se presentan las frecuencias de cada una de las categorías que conforman la columna de referencia:



Análisis de la columna “device_language”

A continuación se presentan las frecuencias de cada una de las categorías que conforman la columna de referencia, agrupando en “Otros” a aquellas categorías con frecuencia menor a 100:



La categoría más frecuente se encuentra presente en el 44% de los registros.

Top 10 de las categorías agrupadas en “Otros” más frecuentes:

device_language	Freq
407706219895259712	70
1526421427153981440	56
406092966496812902 4	48
282284336961851904	39
6035179960508535808	24
1193278617981593344	21
6535228344298127360	10
110153915248271632	9
369578704436051904	5
3177264950152469504	5

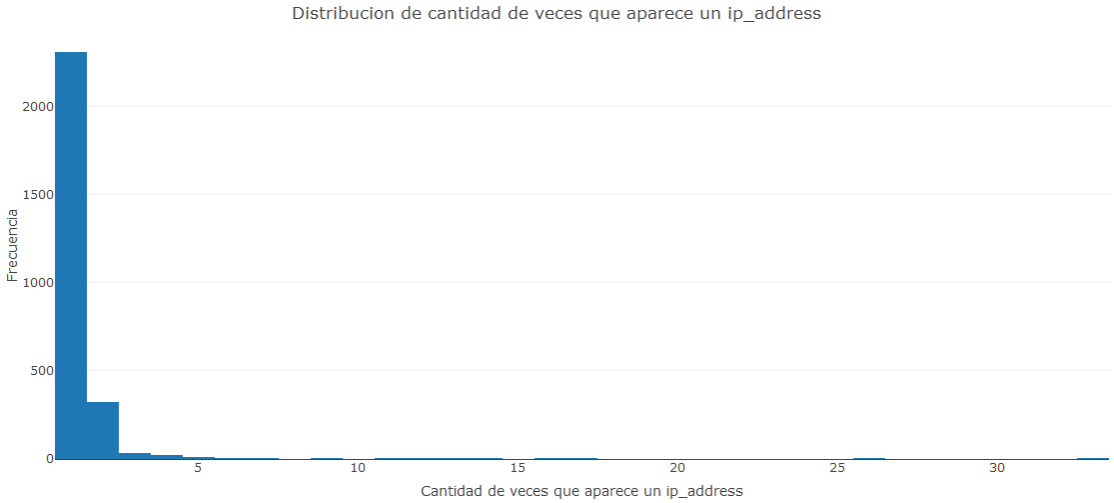
Análisis de la columna “ip_address”

Esta variable cuenta con 2.717 valores distintos, de los cuales al menos el 75% aparecen una única vez en el set de datos. A continuación se exponen las características de la frecuencia de cada una de las categorías de esta variable:

Metrica	Valores
Minimo	1.00
Perc_25	1.00
Perc_50	1.00
Promedio	1.26
Perc_75	1.00
Maximo	33.00
Desvio_Std	1.19
Coef_Variac	0.94

Si bien la gran mayoría de los valores posibles se encuentra en el set de datos muy pocas veces, existe al menos una categoría que aparece 33 veces.

A continuación se presenta el histograma sobre la cantidad de veces en que aparece cada ip_address en el set de datos, evidenciando lo mencionado anteriormente en cuanto a qua la mayor parte de las categorías aparecen muy pocas veces en la muestra:



Exploración de data set “Events”

El análisis se basará en comparar la distribución de las diversas variables del set de datos, teniendo en cuenta las siguientes variables:

- Columna “attributed”: Marca a los eventos atribuidos a Jampp.
- Marca “Install_Post”: Marca calculada. Indica un si un dispositivo del data set de events aparece en el data set de installs, posteriormente al de events.

El objetivo será comparar la distribución del total de registros de events, con las dos subpoblaciones.

Conteos

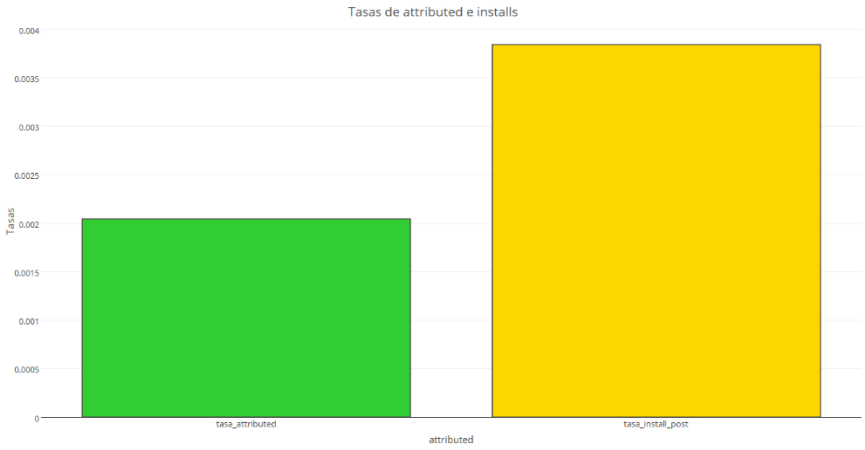
	Install_Post 0	Install_Post 1
Attributed False	2479822	9502
Attributed True	5016	83

	Install_Post 0	Install_Post 1
Attributed False	0.994146542	0.00380929778
Attributed True	0.002010886	0.00003327423

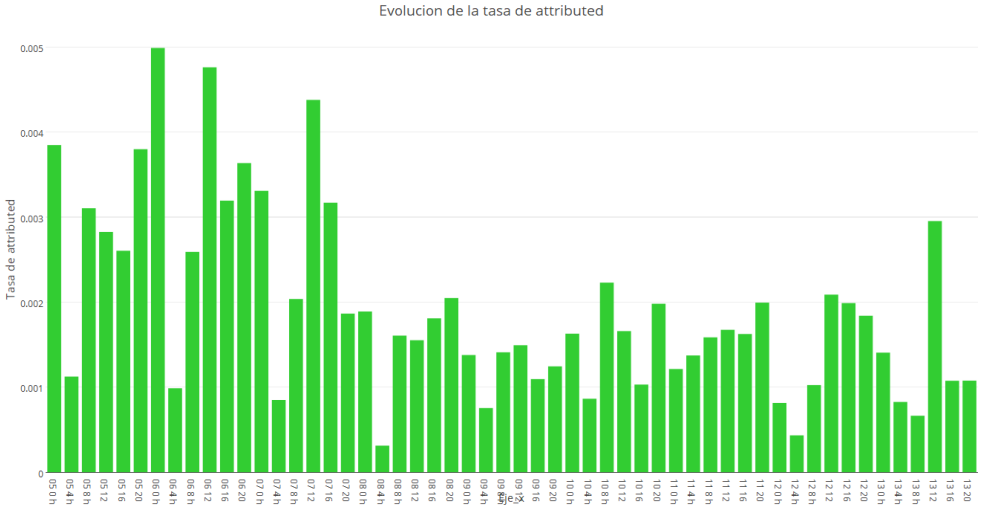
El 99,41% de los registros no corresponde a eventos atribuidos, ni obtuvo instalaciones luego del evento.

Sin embargo, 9502 registros correspondientes a eventos no atribuidos aparecieron en la base de installs con posterioridad al evento que le da origen. En total, fueron atribuidos 5099 eventos.

A continuación se exponen las tasas de éxito de los dos eventos analizados:

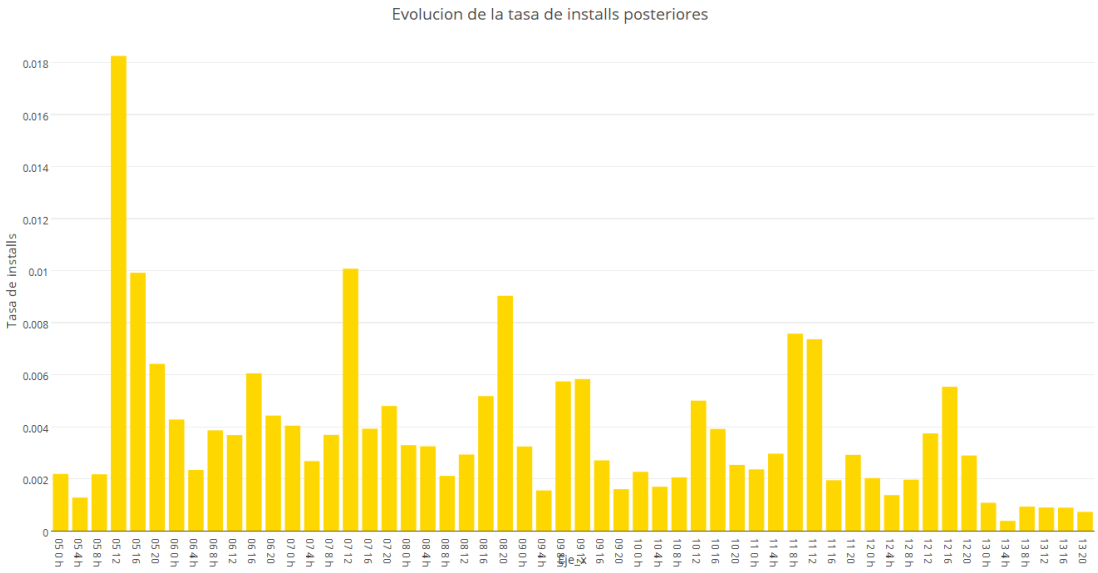


A continuación se muestra la evolución de la tasa de attributed:



En general, las menores tasas se observan en el bucket que va de 4 a 7 am, y las máximas de varían, ya que a veces se dan entre las 12 y 15 hs y a veces entre las 20 a 23 hs. La tasa mínima (menos del 0.01%) se vio el 8 de marzo en el bucket de 4 a 7 am, y la máxima el 6 de marzo, en el bucket de 0 a 3 am. La tasa máxima (0.5%) fue el 6 de marzo, en el bucket de 0 a 3 am.

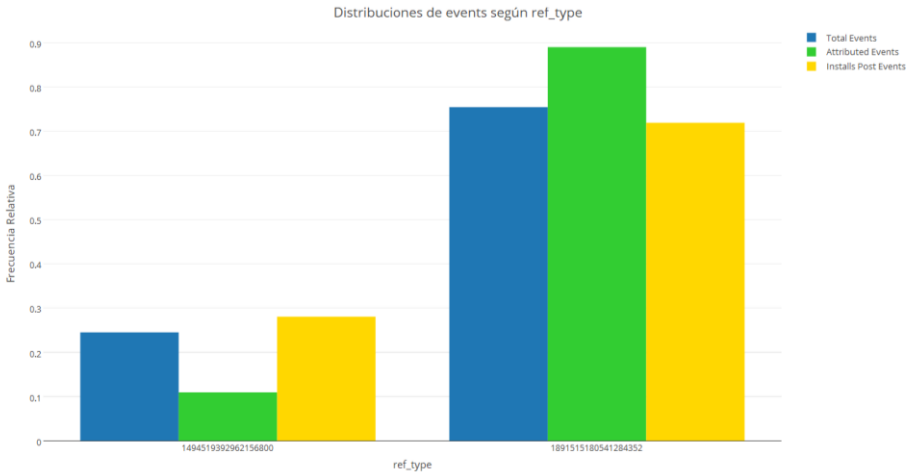
A continuación se muestra la evolución de la tasa de installs:



Se observa un comportamiento similar al de atribuciones, salvo que pareciera existir una tendencia a la baja en la tasa. La tasa máxima (1,8%) se vio el 5 de marzo en el bucket de 12 a 15 hs. La mínima (menos del 0.1%) se vio el 13 de marzo, en el bucket de 4 a 7 am.

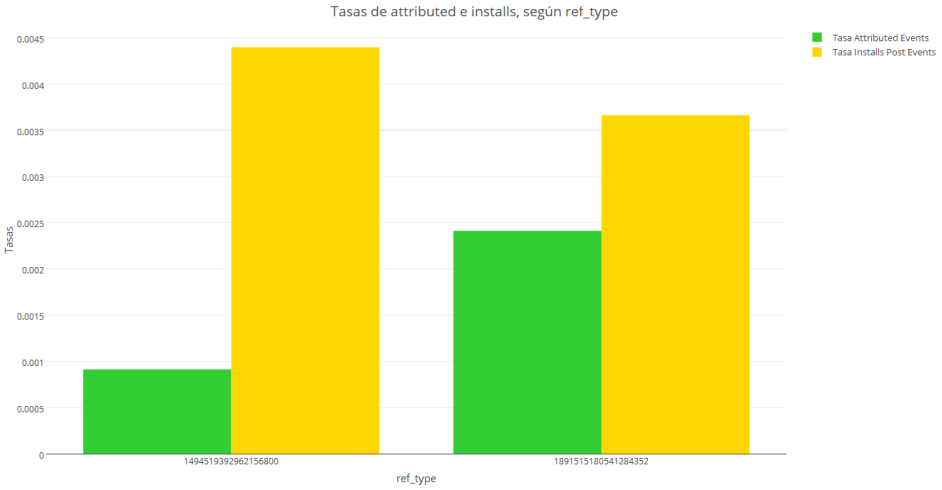
Análisis de la columna “ref_type”

A continuación se muestran las frecuencias relativas de cada categoría, teniendo en cuenta el total de eventos, los eventos atribuidos a Jampp, y los eventos detectados en installs:



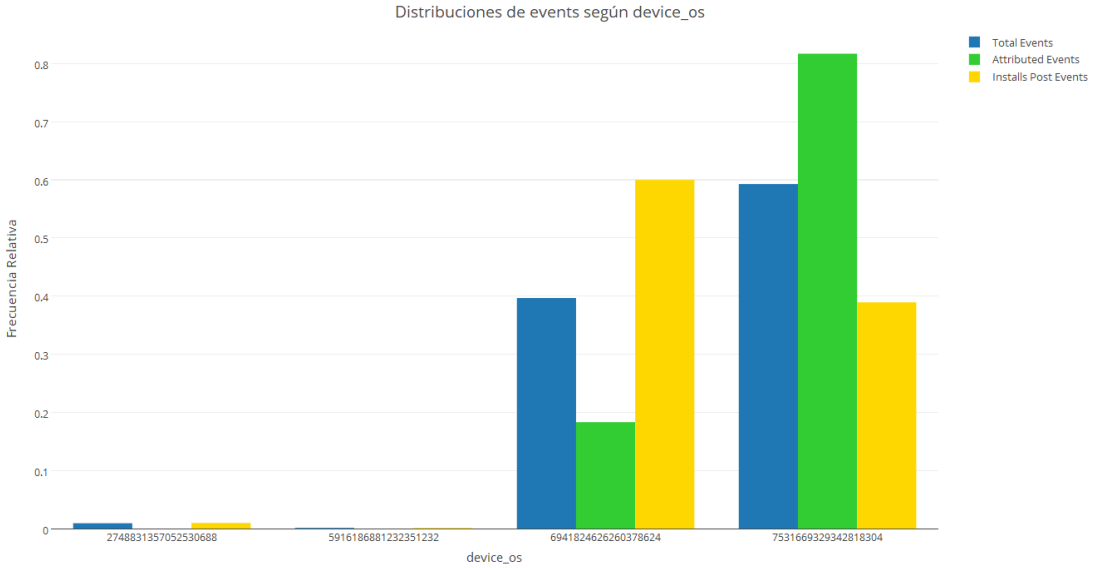
Comparando la distribución del total de eventos y los atribuidos, se observa que el primer tipo disminuye su participación en los atribuidos. Esto indicaría que, los ref_type correspondientes

al segundo código son más propensos a ser atribuidos. Lo inverso ocurre en las instalaciones. Analizando las tasas de éxito de ambos eventos, se observan los siguientes resultados:



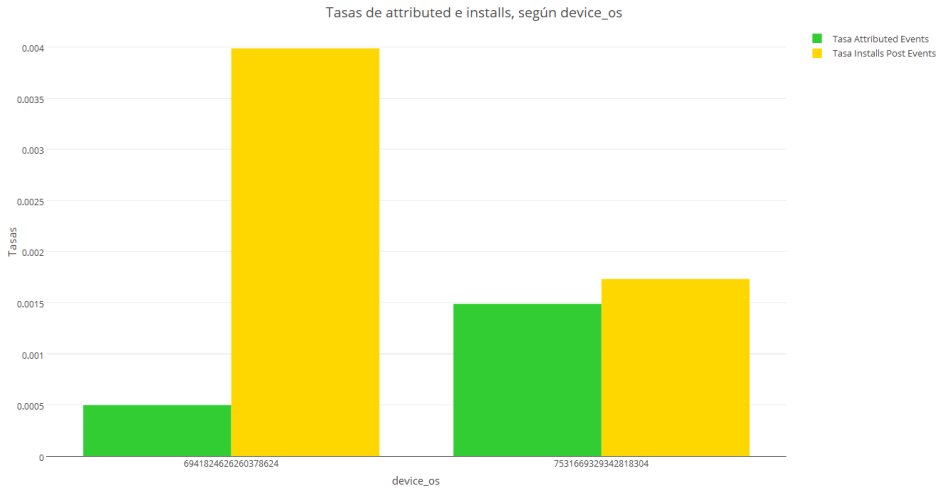
Las conclusiones son similares que en el análisis anterior. El primer código de ref_type cuenta con la tasa más baja de attributed, pero con la tasa más alta de installs.

Análisis de la columna “device_os”



En primer lugar hay que señalar que los primeros dos códigos son insignificativos en cuanto a frecuencia relativa.

El último código muestra una frecuencia relativa de attributed events superior al total de events, lo que indicaría que esta subpoblación es más propensa a ser un evento atribuido. Analizando las instalaciones posteriores a los eventos, ocurre todo lo contrario. Si analizamos las tasas de éxito de los eventos de interés podemos arribar a las mismas conclusiones:

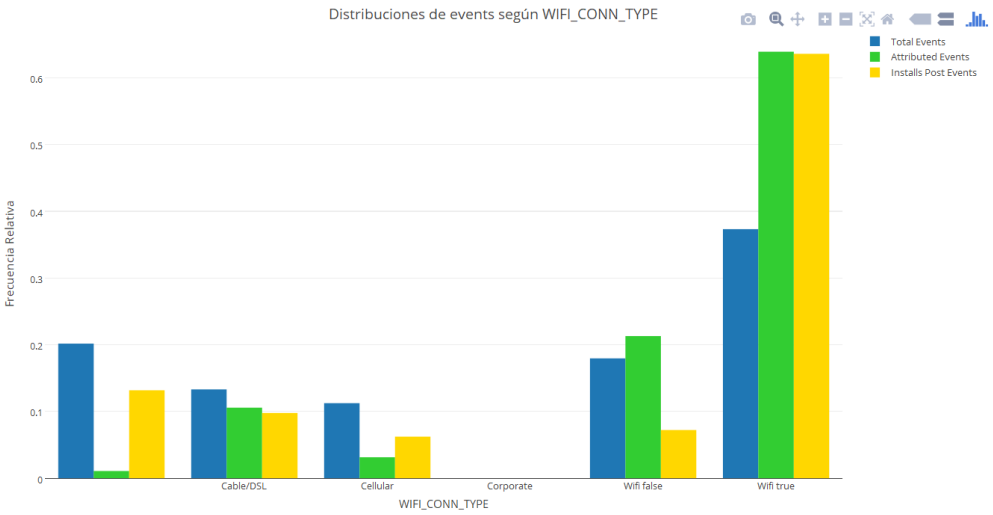


Análisis de las columnas “wifi” y “connection_type”

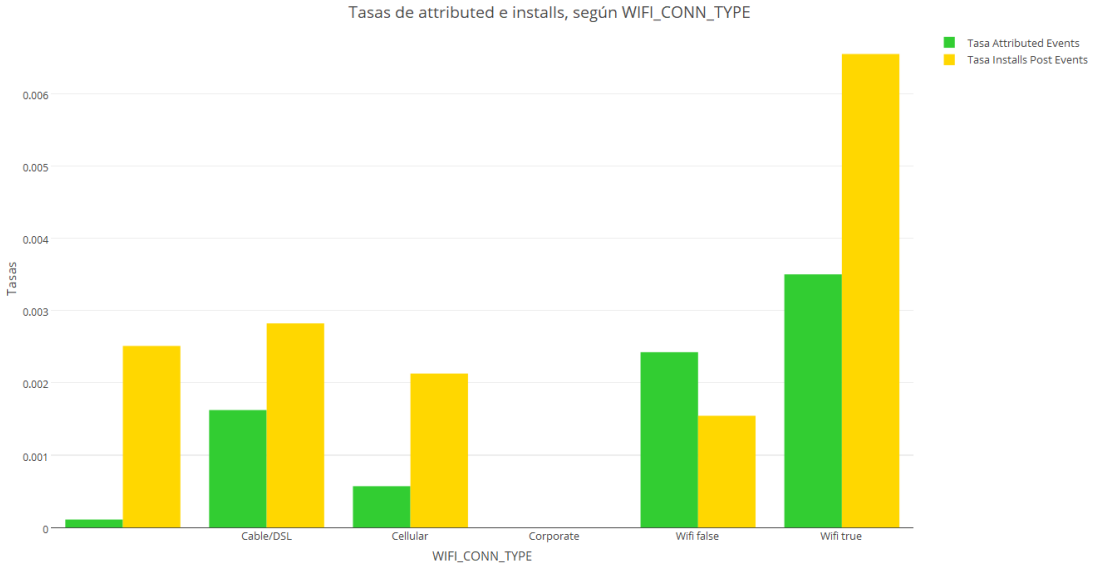
Cruzando la información de estas dos variables podemos darnos cuenta que se encuentran completamente relacionadas:

	Connect_Type: V1	Connect_Type: Cable/DSL	Connect_Type: Cellular	Connect_Type: Corporate
wifi:	503088	331948	280511	4
wifi: false	447970	0	0	0
wifi: true	930902	0	0	0

Solo aquellos registros con el campo “wifi” vacíos contienen información del tipo de conexión. Por lo tanto, se va a crear una variable nueva para analizar, en donde si el campo wifi se encuentra vacío se completará con la información del tipo de conexión. A continuación se expone el análisis de frecuencias relativas del total de eventos y los atribuidos vs los instalados:



La categoría corporate solo cuenta con 4 events, por lo que no vale la pena analizarla. En el resto de las categorías se puede observar que la frecuencia relativa de eventos atribuidos en “Wifi False” aumenta respecto de la frecuencia relativa del total de eventos. Lo contrario ocurre con la subpoblación de installs posteriores. Este incremento es mucho mayor en “Wifi True”, y ocurre en ambas subpoblaciones, por lo que se puede asumir que el tener wifi true es indicador de que el evento es más propenso a ser atribuido, y a ser instalado:



Por el contrario, aquellos eventos con wifi vacío y connect_type vacío son los que cuentan con menor tasa de atribuciones, lo que indica que esta característica denota a los eventos con menor propensión a ser atribuidos. Sin embargo, en términos de tasas de installs posteriores, el evento con menor tasa es el wifi false, lo cual tiene sentido dado que la gente no suele instalarse aplicaciones sin wifi.

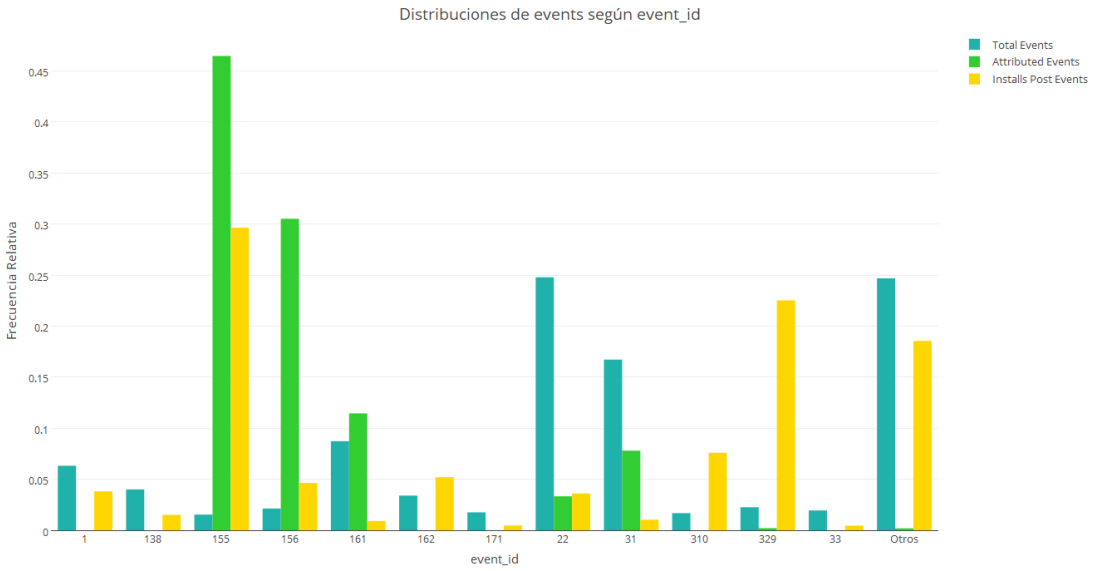
Análisis de columna "event_id"

Para analizar esta variable que tiene 568 valores posibles, se calcularon las frecuencias de cada event_id, y se analizaron las 12 categorías más frecuentes, las cuales representan un 75% del total de eventos disponibles. El 25% restante se agrupó en una categoría "Otros".

En primera instancia se puede observar que los eventos más frecuentes en el set de datos no son necesariamente los más frecuentes en el subconunto de eventos atribuidos, o de eventos que aparecen en installs:

event_id	Freq	Freq_attributed	Freq_Installs_Post
22	618228	170	346
Otros	615619	10	1778
31	417070	398	100
161	217846	584	88
1	157812	0	366
138	100039	0	145
162	84898	0	499
329	56490	11	2159
156	53332	1556	445
33	48581	0	44
171	43942	1	46
310	42101	0	728
155	38465	2369	2841

Analizando las frecuencias relativas se puede observar que el evento más frecuente dentro de los eventos atribuidos es el 155 (al igual que las instalaciones posteriores a los eventos):



De este análisis surge que el evento 155 es muy propenso a ser atribuido a Jampp, y también a que posteriormente se realice algún tipo de instalación.

En general, la distribución en las tres poblaciones (total, attributed, install posterior) según el event_id resulta muy diferente, por lo que posiblemente sea una variable fuerte al momento de explicar la conversión.

Análisis de columna “aplication_id”

Se realizará un análisis comparativo en la distribución de registros en events e installs según la columna aplication_id. Para poder realizar el análisis comparativo, sólo se tendrán en cuenta los códigos que aparecen en ambos data sets.

Comparando las frecuencias obtenidas en events e installs del mismo código de aplication_id, se obtienen los siguientes resultados:

cation_id	Freq	Freq_Rel	Freq_Installs	Freq_Installs_Rel
Otros	2279225	0.913728345192455	112	0.032825322391 5592
10	58311	0.0233765484041 8	389	0.11400937866354
7	48005	0.0192449315933984	947	0.277549824150059
8	38972	0.0156236532456604	328	0.096131301289 5662
16	24365	0.00976778998590055	362	0.106096131301 29
9	8198	0.00328653159468142	731	0.214243845252052
12	7882	0.00315984899112941	15	0.004396248534583 82
2	5931	0.0023777041824903	174	0.050996483 0011723
15	5801	0.00232558792153536	20	0.005861664712778 43
17	4795	0.00192228824060715	13	0.003810082063305 98
34	4025	0.0016135996180279	28	0.008206330597889 8
32	3602	0.00144402132276683	2	0.000586166471277843
21	2858	0.00114575595237857	7	0.002051582649472 45
29	732	0.000293454638607806	97	0.028429073856 9754
20	688	0.000275815288746135	95	0.027842907385 6975
28	513	0.000205658783614487	17	0.004982415005861 66
24	300	0.000120268294511396	13	0.003810082063305 98
23	100	0.000040089431503798 7	1	0.000293083235638921
0	56	0.000022450081642127 3	18	0.005275 49824150059
6	27	0.000010824146506025 6	35	0.010257913247362 3

Algunas conclusiones:

- La gran mayoría (91%) de los application_id del set de datos events no se encuentran en el set de datos installs.
- La gran mayoría de los application_id del set de datos de installs si se encuentra contemplado en events (97%).
- Los application_id más frecuentes en installs, no son significativos en el set de datos de events.
- Pareciera que en events se contemplan muchas más aplicaciones que las que se capturan como instaladas.

Exploración de data set “Clicks”

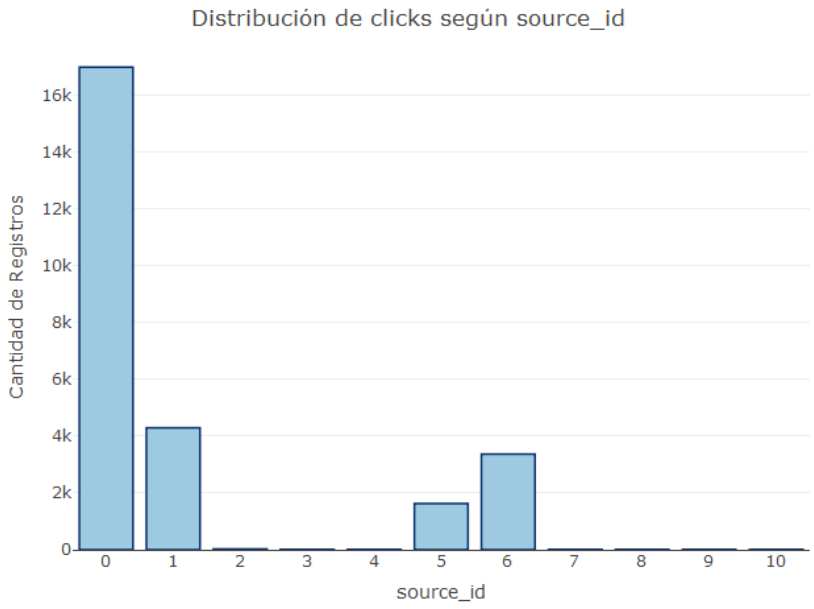
Algunas aclaraciones preliminares:

- La columna “action_id” se encuentra completamente vacía, por lo que no se va a analizar.
- La columna “Country_code” está completa con el código de un solo país, por lo que no tiene sentido analizarla.
- La columna “wifi_conections” posee un único valor igual a “False”, por lo tanto no será analizada.
- Las columnas “agent_device” y “brand” poseen un 88% y 76% de celdas vacías respectivamente, por lo que tampoco serán analizadas.
- La columna “Specs_brand”

Análisis de “advertiser_id”, Identificación interna del anunciante, el cliente de Jampp que paga por el anuncio.

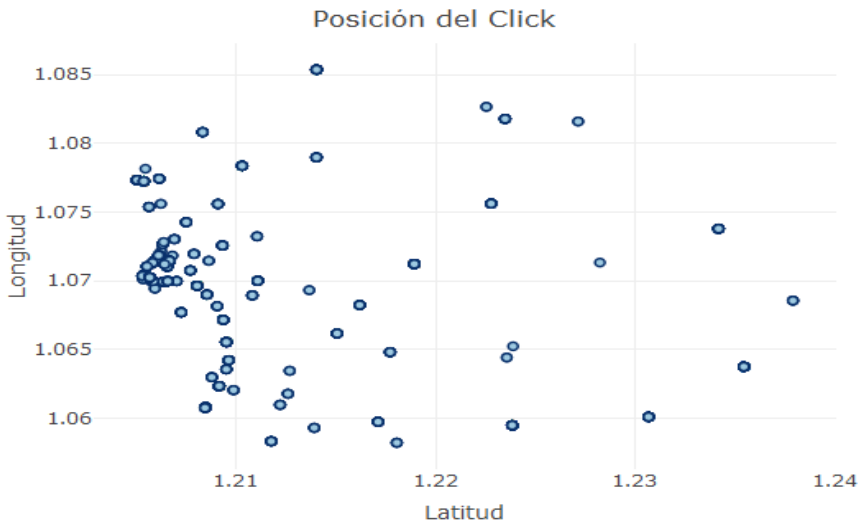
En este caso se observa que un solo cliente (3) contiene el 99,66% de los clicks totales.

Análisis de “source_id”: Fuente desde la cual se originó el clic.



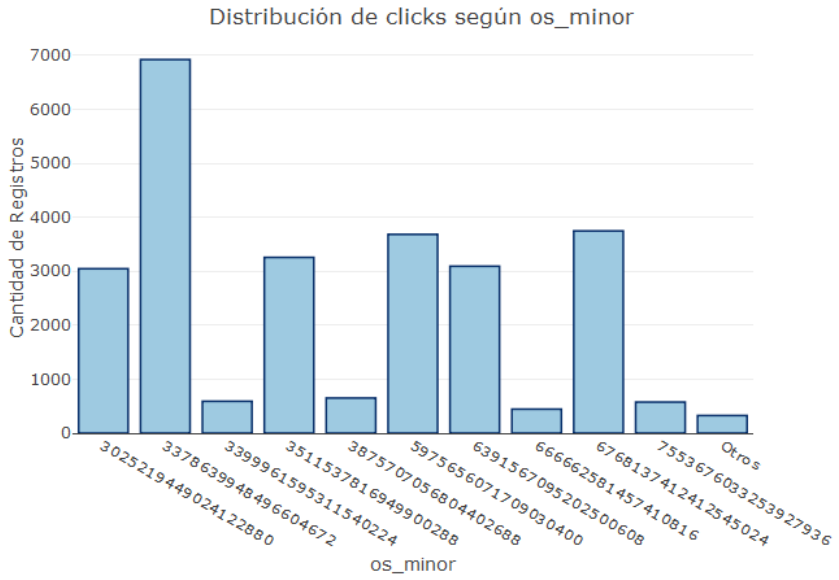
El 99,8% de los clicks tienen source_id 0, 1, 5 o 6.

Análisis de “Latitud” vs “Longitud”: Latitud y longitud estimadas donde se realizó el clic



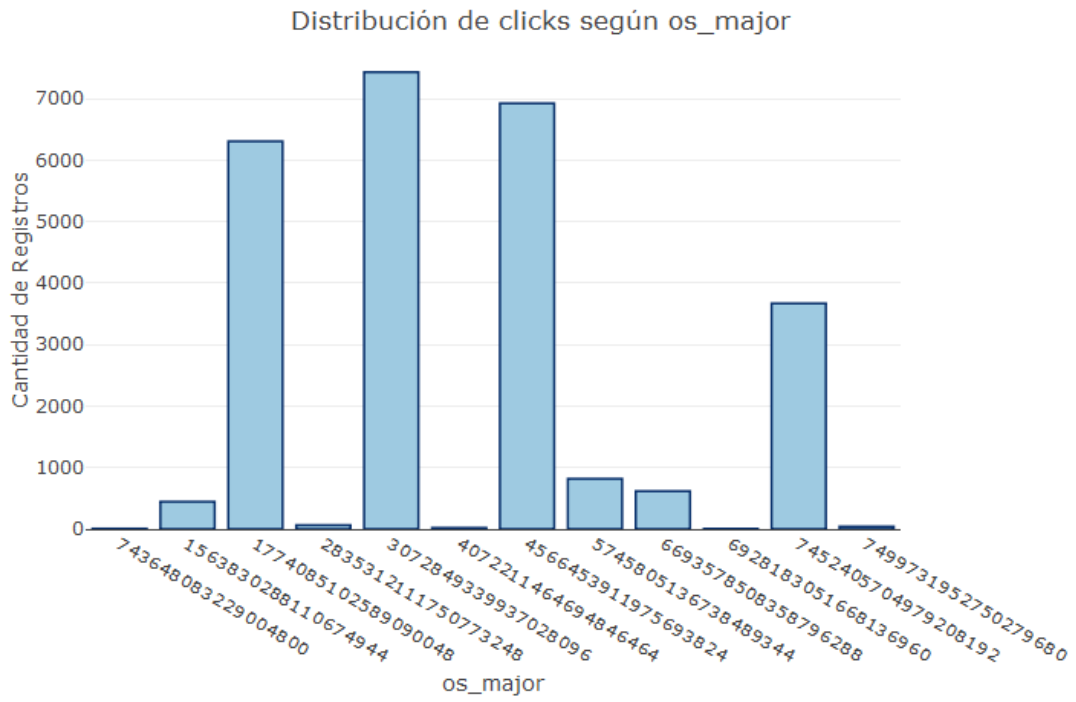
El 91% de los clicks, se encuentran concentrados entre los valores de longitud menor a 1.21 y longitud menor a 1.075.

Análisis de “OS_minor” menor versión para el sistema operativo



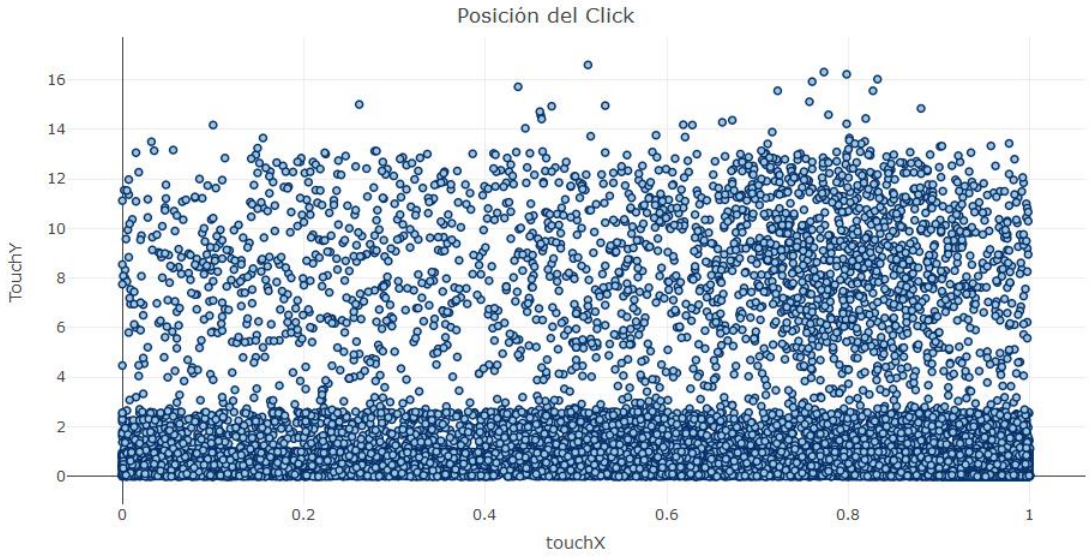
El 26% de los clicks provienen de la versión 3378639948496604672, que es la que concentra la mayoría en esta clase.

Análisis de "OS_major" mayor versión para el sistema operativo



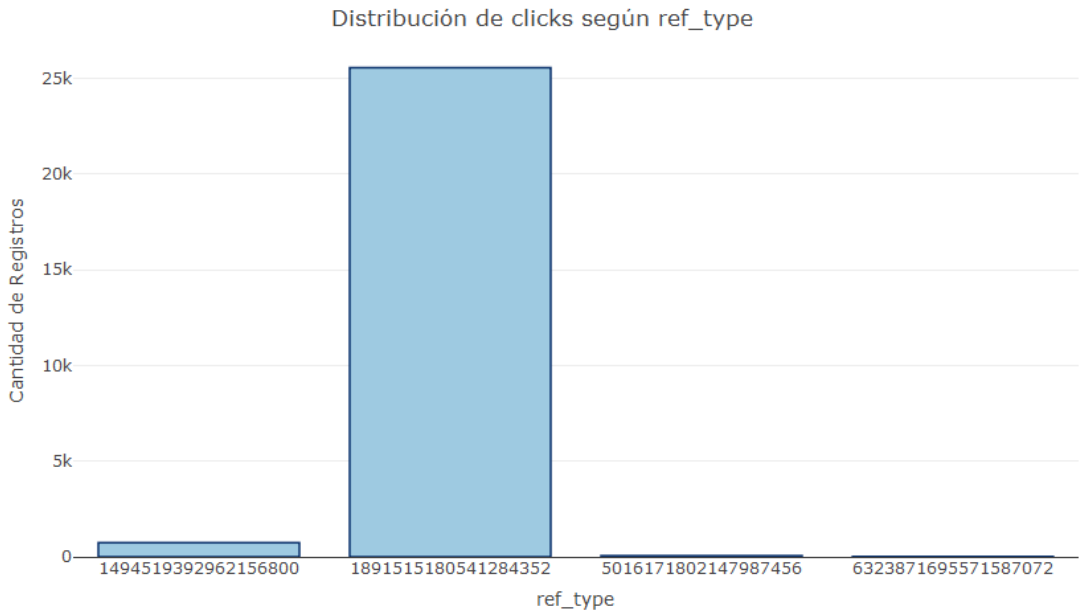
En este caso, la distribución de los clicks en los dispositivos de mayor SO, se concentra principalmente en 3 de ellos: un 28% en 3072849339937028096, 26% en 4566453911975693824 y 24% en 1774085102589090048.

Análisis de "TouchX" y "TouchY" : Posición X e Y del click



La mayoría de los clicks se concentran en los valores inferiores a 3 de la posición Y, para todos los valores de X.

Análisis de “ref_type”



El 97% de los clicks, provienen de dispositivos con el sistema 1891515180541284352.

Algunas conclusiones interesantes

- El 99% de los dispositivos que aparecen en “Clicks”, también están en “Auctions”.
- El 90% de los dispositivos que aparecen en “Installs” también están en “Events”.
- La distribución de registros por día es similar en “Events” e “Installs”, pero es muy distinta en “Clicks” (hay muy poquitos en los primeros dos días).
- Hay una tendencia ascendente en la frecuencia de registros en “Auctions” a medida que pasa el tiempo.
- Los horarios con menor frecuencia de registro suelen ser por la mañana en todas las tablas (4 a 10 hs aprox). En “Clicks” también hay una frecuencia muy baja en horarios de la tarde /noche que suelen ser mucho más frecuentes en el resto de los datasets (19-22 hs aprox).
- En “Auctions”, hay un 38% de dispositivos que aparecen más de 500 veces en el set de datos (es un promedio de 55 veces por día, lo cual podría resultar elevado). Incluso hay al menos un dispositivo que aparece 27.762 veces (en promedio, 3085 veces por día, 128 por hora, 2 veces por minuto, es muy raro). Esto podría asociarse a temas de fraude.
- La distribución de registros en “Auctions” según platform tiene saltos extraños en los primeros días. Habría que revisar si el set de datos se encuentra sesgado, o si pasó algo particular en esos momentos.
- Analizando el tiempo entre subastas, el 75% de los registros cuenta con menos de 149 segundos entre una subasta y la otra (2,5 minutos aprox), por lo que la mayor parte de los registros se concentra en valores bajos.
- Sólo un 1% de la muestra cuenta con tiempo superior a las 24 hs.
- La variable platform y source_id parecen buenas predictoras del tiempo entre subastas, ya que presentan medianas y percentiles del tiempo muy distintos en sus categorías.
- En “Installs”, las aplicaciones más frecuentes son las 7, 9, 10, 16 y 8, que representan el 81% del total de registros.
- La mayor parte de las instalaciones NO son implícitas.
- La mayor parte de las instalaciones se realizan con el sesión_user_agent 2 o 4.
- En “Events”, un 0.2% de los registros se marca como atribuido a Jampp, y un 0.4% de los eventos aparece con posterioridad en el data set “Installs”. Estos valores se mantienen bastante estables en el tiempo, aunque los picos máximos se suelen ver entre las 12 y 15hs, o 20 y 23hs.
- Los “event_id” más frecuentes en la base de “Events” no son los más frecuentes en el subconjunto de eventos atribuidos. Esta variable puede resultar muy importante si se intenta predecir la atribución de un evento a Jampp, o la aparición de una instalación con posterioridad a un evento, ya que las distribuciones varían mucho.
- La gran mayoría (91%) de los application_id del set de datos events no se encuentran en el set de datos installs, pero la gran mayoría de los application_id del set de datos de installs si se encuentra contemplado en events (97%).
- Los eventos con “wifi false” son los que menor tasa de instalaciones posteriores al evento tienen, lo que tiene sentido dado que la gente no suele instalarse aplicaciones sin wifi.
- En “Clicks” un solo cliente (3) contiene el 99,66% de los clicks totales.
- El 99,8% de los clicks tienen source_id 0, 1, 5 o 6.
- La mayor parte de los clicks se concentran en una franja inferior (posición Y menor a 3).
- El 97% de los clicks, provienen de dispositivos con el sistema 1891515180541284352 (ref_type).

Algunos puntos interesantes que quedaron pendientes de investigación

- ¿Cómo es la distribución de los auctions que se encontraron después en la base de clicks? ¿Cambia respecto de la distribución total?
- ¿Cómo impacta el tiempo entre subastas en la realización de un click, o install?
- ¿Existen events o instalaciones anteriores al momento en que se habilita la subasta? ¿Eso podría explicar el tiempo entre subastas?
- ¿Existen eventos anteriores al evento atribuido? ¿Eso impacta en la atribución del evento? ¿Y en la realización de instalaciones en un momento posterior?
- ¿La cantidad de clicks antes de un evento atribuido puede impactar en la probabilidad de convertir? ¿Y el tiempo entre estos clicks?
- ¿Cuántas subastas se habilitaron para un dispositivo antes de lograr un evento atribuido? ¿Y antes de lograr un click? ¿Y antes de que se genere una aplicación? ¿Qué características tenían estas subastas?

Por falta de tiempo no se ha llegado a explorar todos estos interrogatorios, que serán tenidos en cuenta al momento de realizar el próximo trabajo.