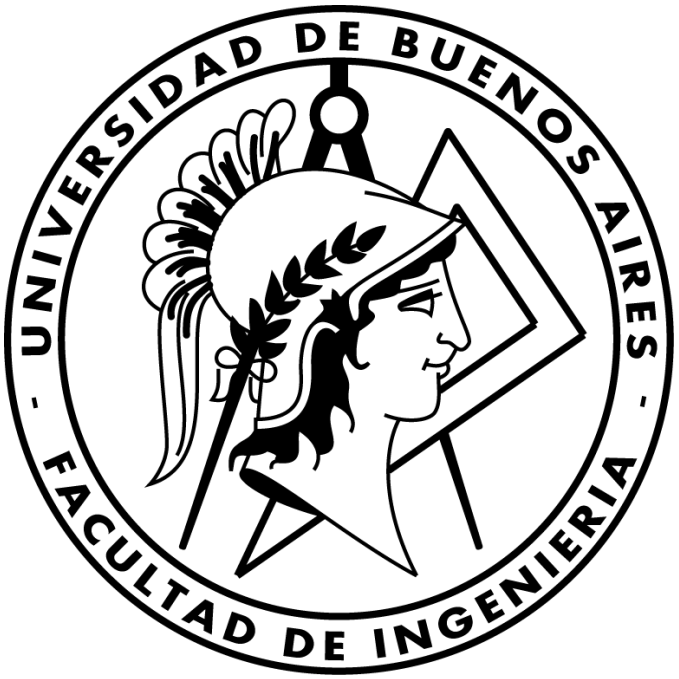
Trabajo Práctico 1
75.06 - Organización de Datos



Grupo 31
“El mejor equipo de los últimos 50 años”

Alumnos:

Nombre	Padrón
Gatti Nicolás	93570
Del Carril Manuel	100772
Verón, Lucas	89341

Índice

Introducción

El presente trabajo tiene como objetivo realizar un análisis exploratorio de datos sobre la información provista por la plataforma de Jampp.

Se analizará la información provista y consolidada en 4 archivos diferentes:

- Auctions: Información de subastas en las que Jampp participó.
- Clicks: Clicks de usuarios sobre aplicaciones instaladas de clientes de Jampp.
- Installs: Instalaciones de aplicaciones publicitadas por Jampp.
- Events: Eventos que suceden sobre aplicaciones instaladas de clientes de Jampp.

Se realizará un análisis con las siguientes características: Cada apartado intentará responder una pregunta sobre el set de datos, para lo cual se realizarán operaciones de filtrado y corrección de los datos a fin de preparar la información para realizar el análisis y responder lo solicitado.

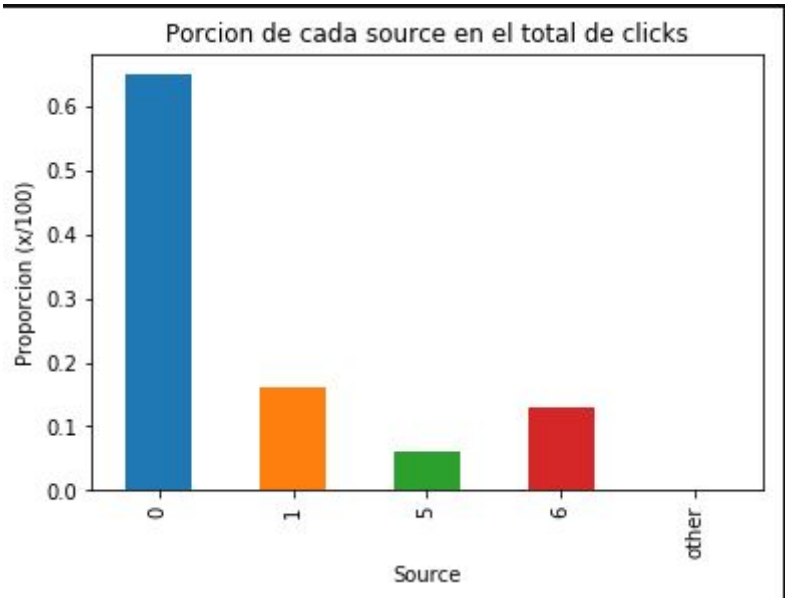
Análisis Exploratorio

¿Qué sources distintos existen?

Dentro del archivo de clicks cada registro está identificado con el proveedor de subastas (el source) desde donde se consiguió el usuario.

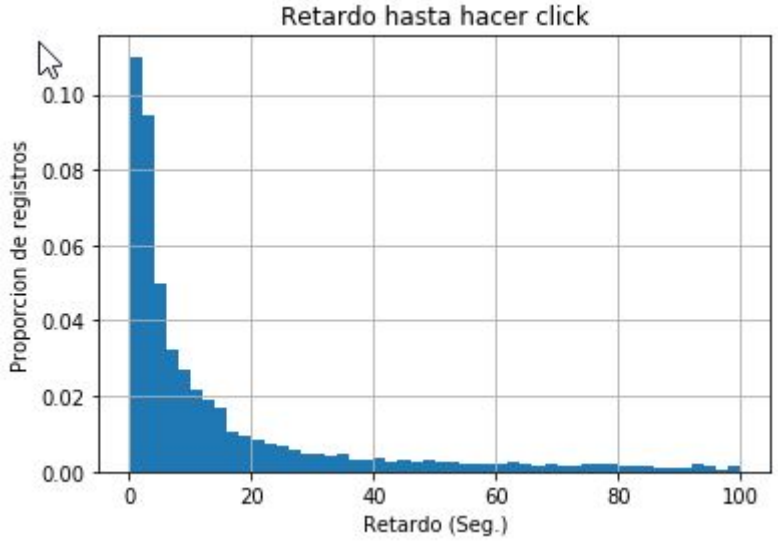
El objetivo es identificar si hay un source que generan más clicks que otro, esto podría indicar que las publicidades realizadas por este tienen más llegada entre los usuarios.

Para evidenciar esto, se separaron los cuatro sources principales y se muestra qué proporción del total representan. El resto de los sources que existen en el set de datos no representan un volumen de tráfico considerable y son agrupados todos juntos como ‘other’:



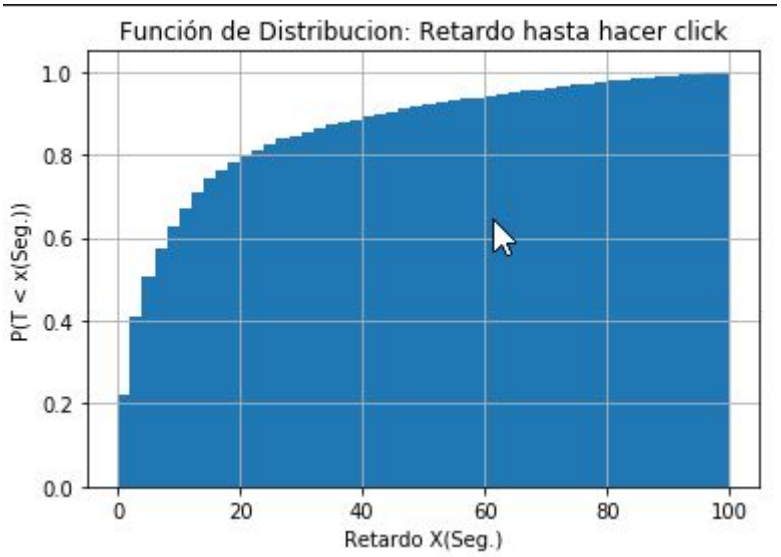
¿Cuánto tiempo tardan los usuarios en hacer clicks?

Del set de datos de clicks se desea saber cuanto tarda un usuario en hacer click, se analiza esta variable y se obtiene la siguiente distribución de los tiempos:



Se puede observar que los clicks ocurren en su mayoría antes de los 20 segundos, concentrándose mayormente entre los 0 y los 10 segundos.

Esta observación indica que es altamente probable que los clicks se observen en muy poco tiempo o que no se observen si se superó este umbral. A continuación se muestra la función de distribución acumulada de esta misma variable:



¿Cómo se distribuyen los click(x,y) en el marco de la publicidad?

La idea plotear los puntos de los clicks para cada registro y ver la densidad de puntos que encontramos.

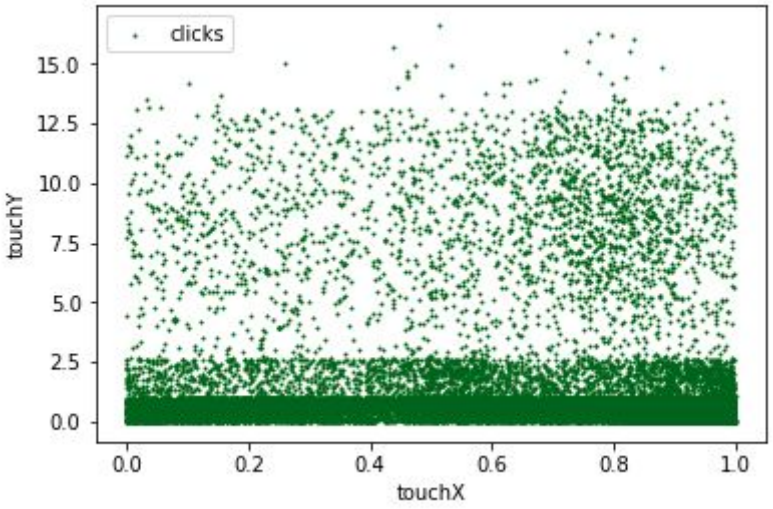


Fig 1. touchX vs touchY para todos los registros no nulos de clicks

Achicando la resolución de la imagen se puede apreciar mejor donde se encuentra la mayor densidad.

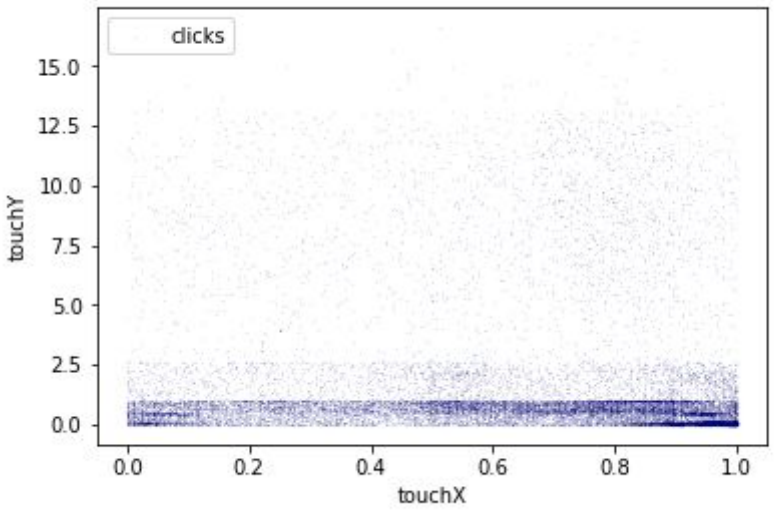


Fig 2. touchX vs touchY para todos los registros no nulos de clicks

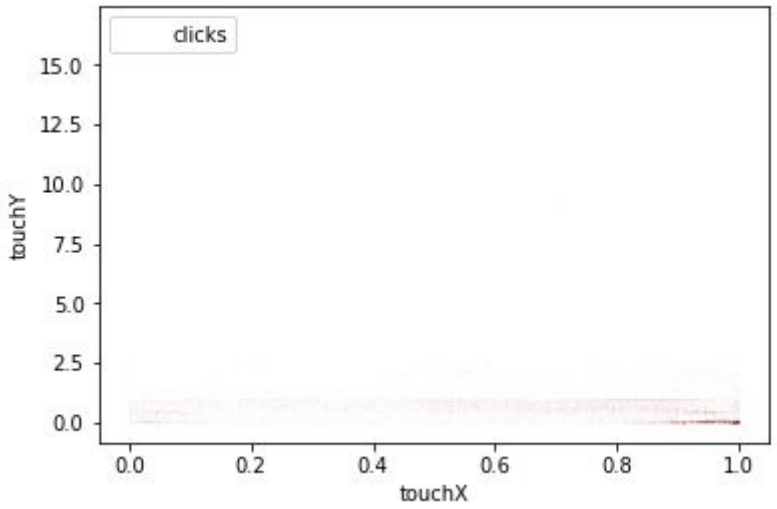


Fig 3. touchX vs touchY para todos los registros no nulos de clicks

Parece que la gran mayoría de los usuarios clickean en la parte inferior de la publicidad.
En la parte inferior derecha existe la mayor cantidad, densidad, de clicks.

¿Qué se puede decir de las zonas(de la publicidad) respecto al tiempo en hacer clicks?

Si separamos los clicks de los usuarios en grupo de acuerdo a las franjas de densidades que aparecen, vemos cuánto demoran los usuarios en promedio en hacer clicks en la pantalla.
Tomamos un límite a partir del cual no tiene sentido la espera(el dato): 20 segundos.
Separamos los clicks en 6 rangos: **r1: 0~1; r2: 1~2.5; r3:2.5~5.0 ; r4: 5.0~7.5 ; r5: 7.5~10.0 ; r6: 10.0~15.0** y vemos en promedio como fue la duración de la publicidad para cada grupo.

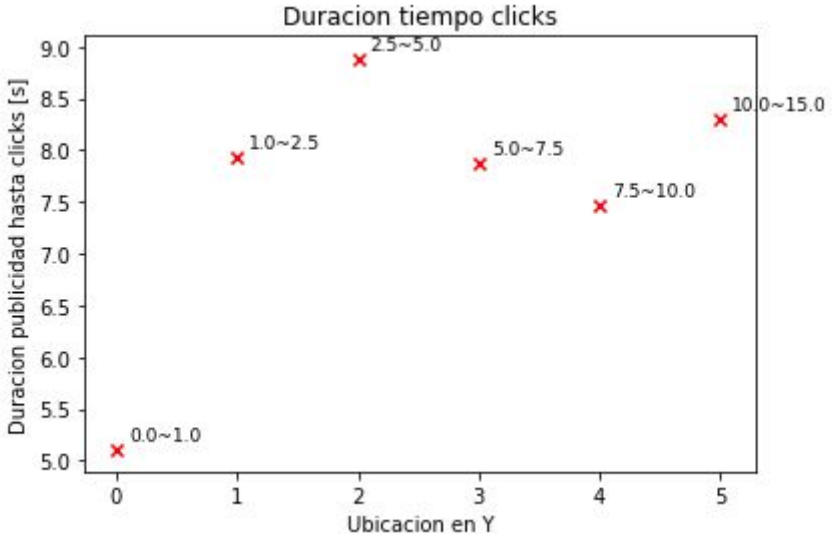


Fig4. duración de publicidad en función de la posición Y

Podemos ver que de los usuarios que clickean entre 0~1/15 de la pantalla(Y) lo hacen muy rápido(la publicidad dura bastante menos) que en el resto de los casos.

¿Cómo se distribuyen los clicks según el ref_type(id publicidad)?

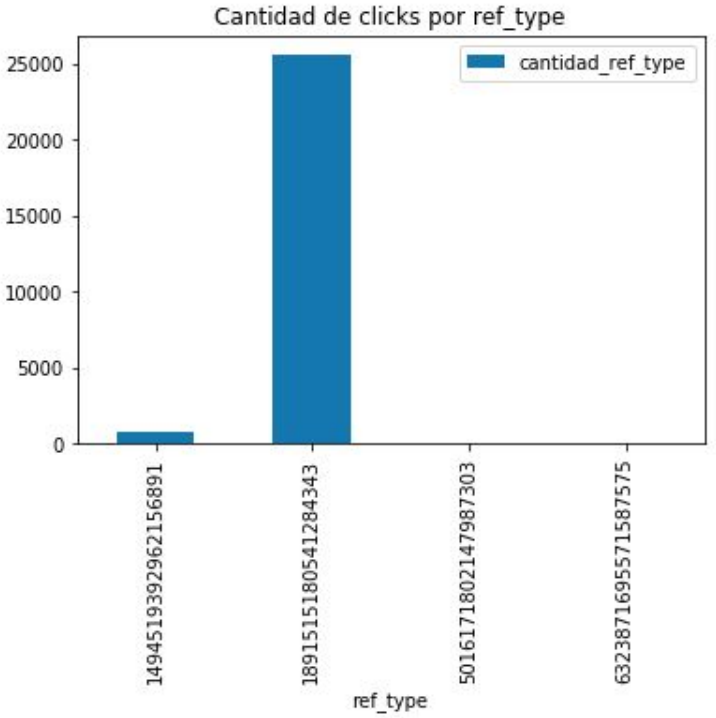


Fig 5. ref_type por cantidad de clicks

La mayoría de clicks perteneces a una sóla publicidad.

¿Cómo se distribuyen las subastas según los días?

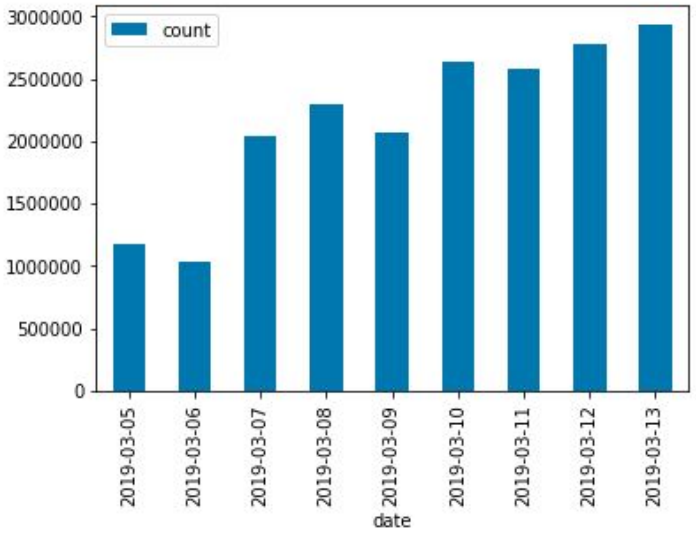


Fig 6. distribución de subastas por día

M	X	J	V	S	D	L	M	X
5/3	6/3	7/3	8/3	9/3	10/3	11/3	12/3	13/3

Mirando las cantidades de subastas por día vemos que para el mismo día, en semanas diferentes existe una variación alta. No parece haber una relación entre la cantidad de subastas y los días en donde se producen.

¿Cuáles son las aplicaciones que más eventos registraron?

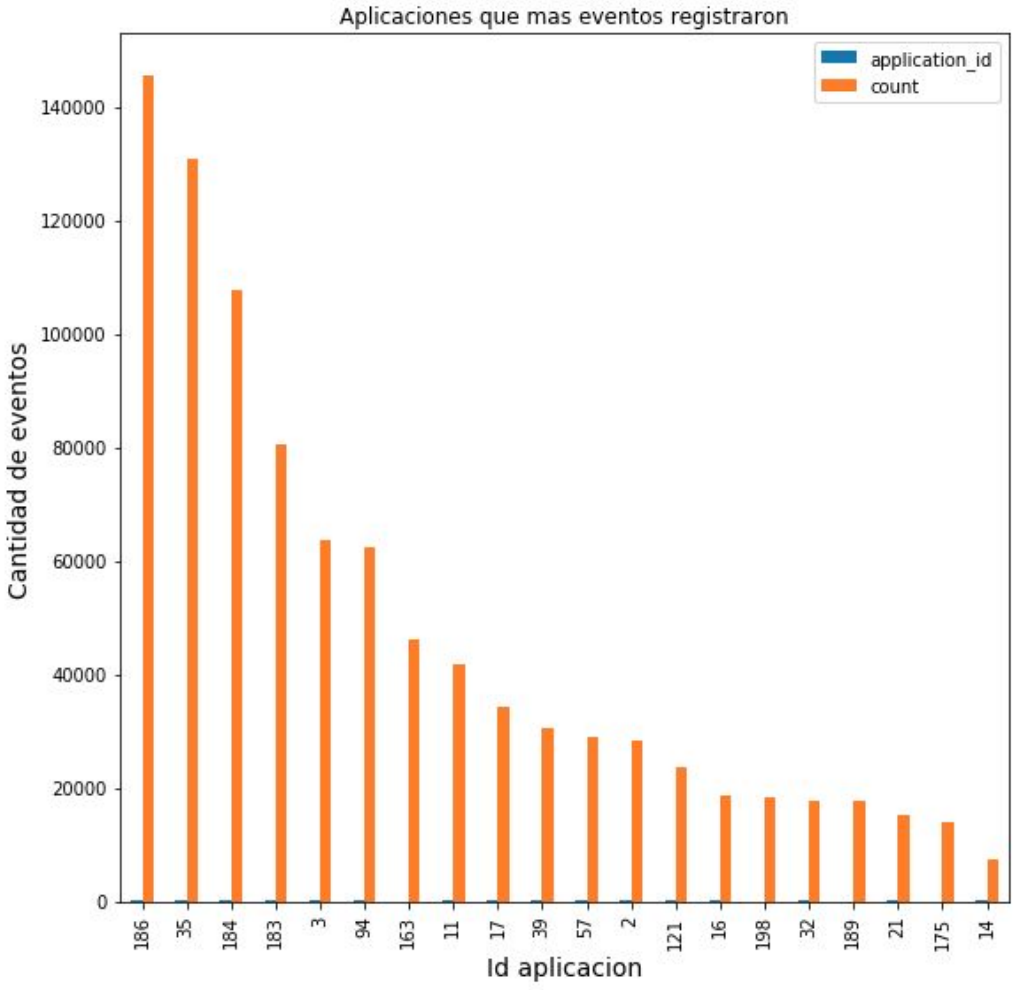


Fig 6. app que más eventos registraron

Podemos ver las app que más eventos tienen y participaron de subastas. Existen 3 aplicaciones que son las que más participación tienen. A partir de la app 5,6… etc más que se duplica la participación de la mayor(186) respecto del resto.

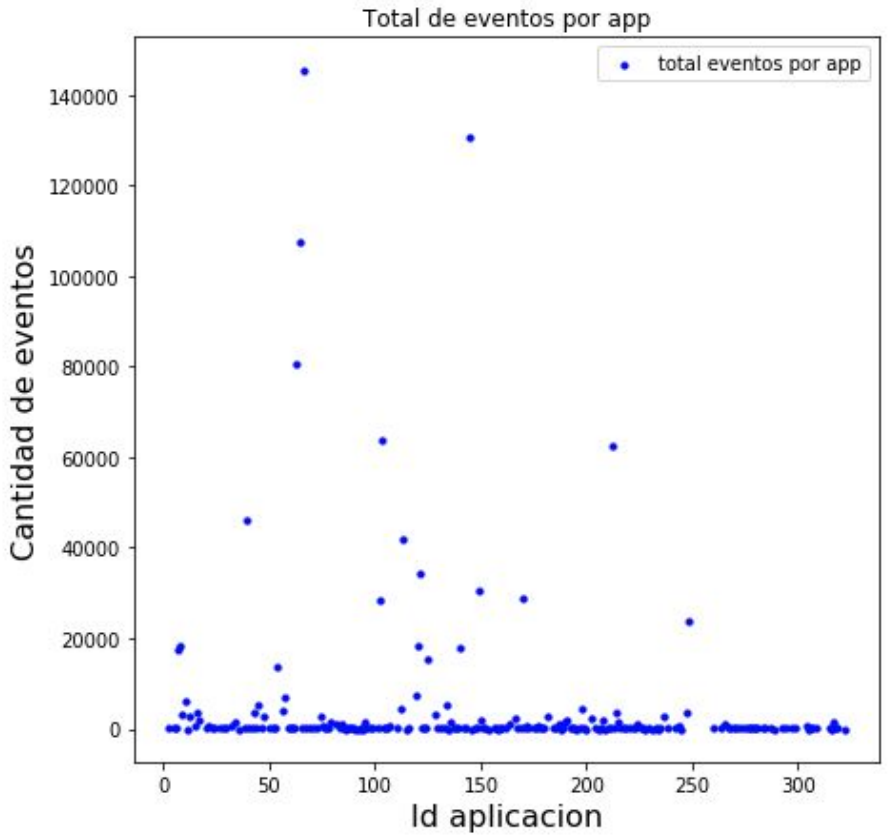


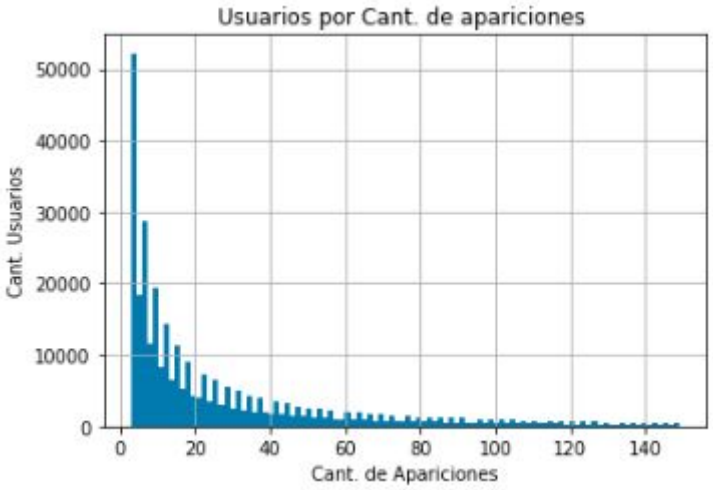
Fig 7. app que más eventos registrados

Se puede notar que existe una gran dispersión entre las aplicaciones líderes en eventos. Hay no más de 20 app que se llevan la mayor cantidad. Por otro lado en relación a estos últimos, el resto de las aplicaciones tiene una cantidad de eventos similares.

¿Cuántas veces aparece un usuario en una encuesta?

Para analizar la cantidad de veces que un dispositivo aparece en una encuesta se tuvieron en cuenta todos los sources a la vez, es decir, sin importar de qué proveedor provino el usuario.

Adicionalmente se descartaron usuarios que aparecieron una sola vez y usuarios que aparecieron un número exagerado (muy por encima de la mediana) de veces, este valor se tomó en 150 apariciones. Se puede ver en la distribución de la cantidad de apariciones que con estos valores se puede apreciar claramente la tendencia de los valores:



Algunos valores de esta distribución se resumen en la siguiente tabla:

Mín.	3
Prom.	31.33
25%	6
50%	14
75%	35

Se puede observar que en el periodo muestreado un usuario aparecerá en promedio 31 veces mientras que la mediana de los usuarios aparece 14 veces.

¿Cual es el tiempo promedio entre apariciones de un usuario?

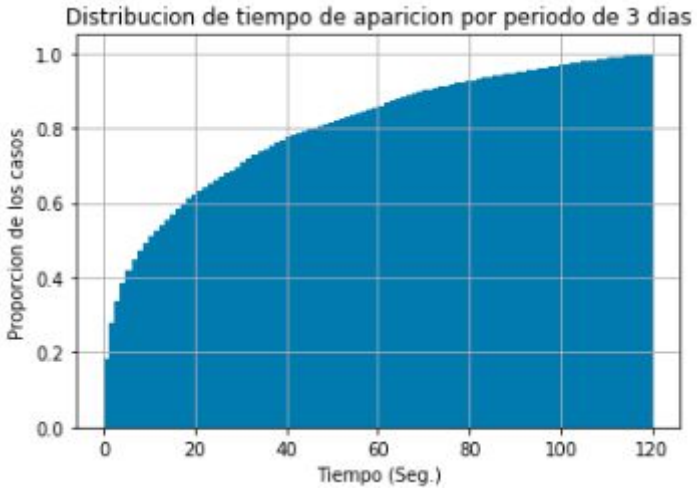
Para medir el tiempo promedio entre apariciones se ordena cada dispositivo por fecha de aparición y se calcula el tiempo entre una aparición y la siguiente, luego se eliminan todos los registros que no tengan fecha siguiente.

En una primer pasada se encuentra que el 75% de los casos no superan los 450 segundos, sin embargo se encuentran casos muy por encima de este valor hasta un máximo de 603942 segundos. Dado esto, se filtran los datos hasta llegar a un descarte de valores por encima de 120 segundos.

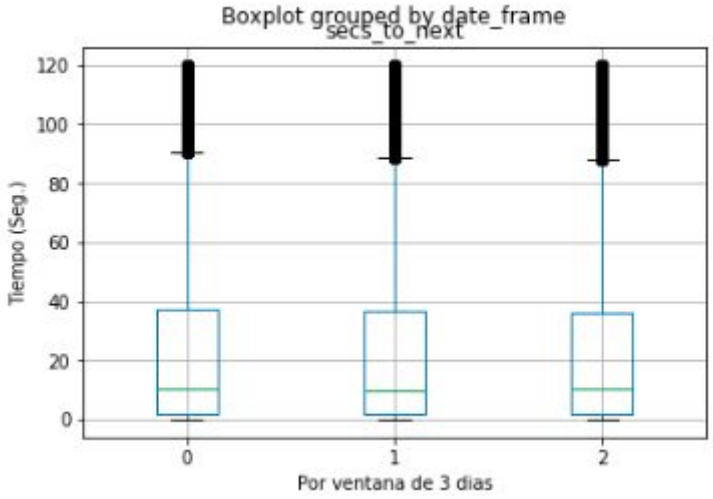
Además, se limita el muestreo a periodos de solo 3 días. Más adelante se muestra que en todos los periodos de 3 días formados la tendencia se mantiene.

Se puede observar que la variable no parece ser exponencial, sino que posee un pendiente mucho más pronunciada. Se puede decir que es muy probable que si un usuario no es visto nuevamente dentro del primer minuto, es muy probable que no aparezca.

A continuación se muestra la distribución acumulada de esta variable:

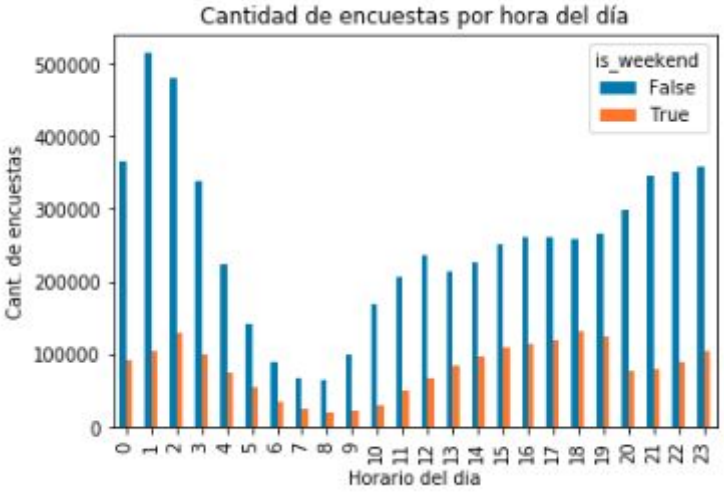


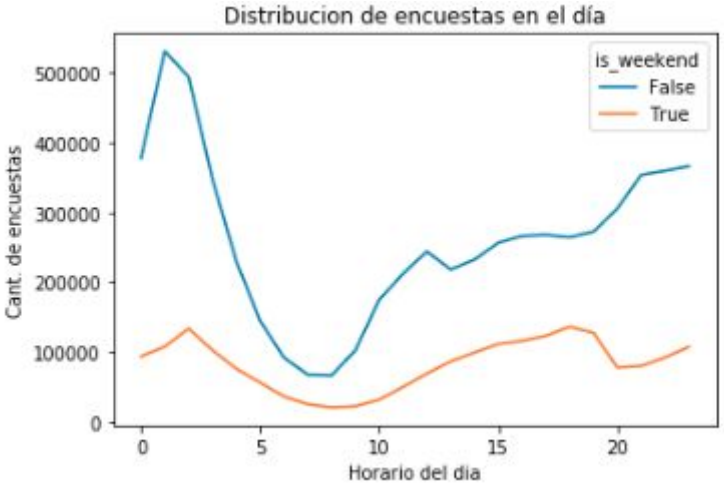
Para finalizar, se demuestra que es válido filtrar los períodos de tres días, ya que como se puede ver en el siguiente gráfico, los momentos de las variables no cambian:



¿En qué horarios se realizan más subastas?

Del total de encuestas se analiza la cantidad de encuestas que se realizan por hora del día, distinguiendo las que se realizan un día del fin de semana de un día laboral normal:



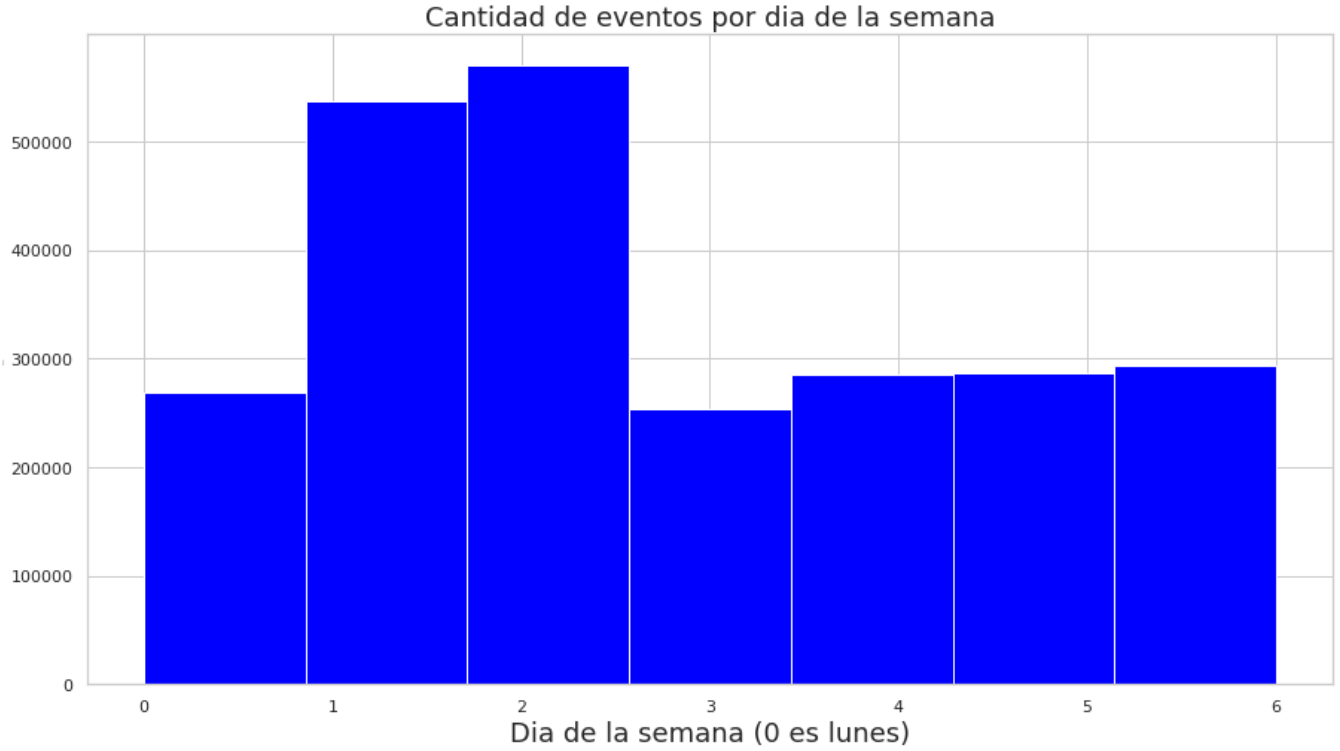


Se puede observar que para los días laborales existen picos de encuestas a la madrugada y en horarios del mediodía, con un valle en las primeras horas de la mañana.

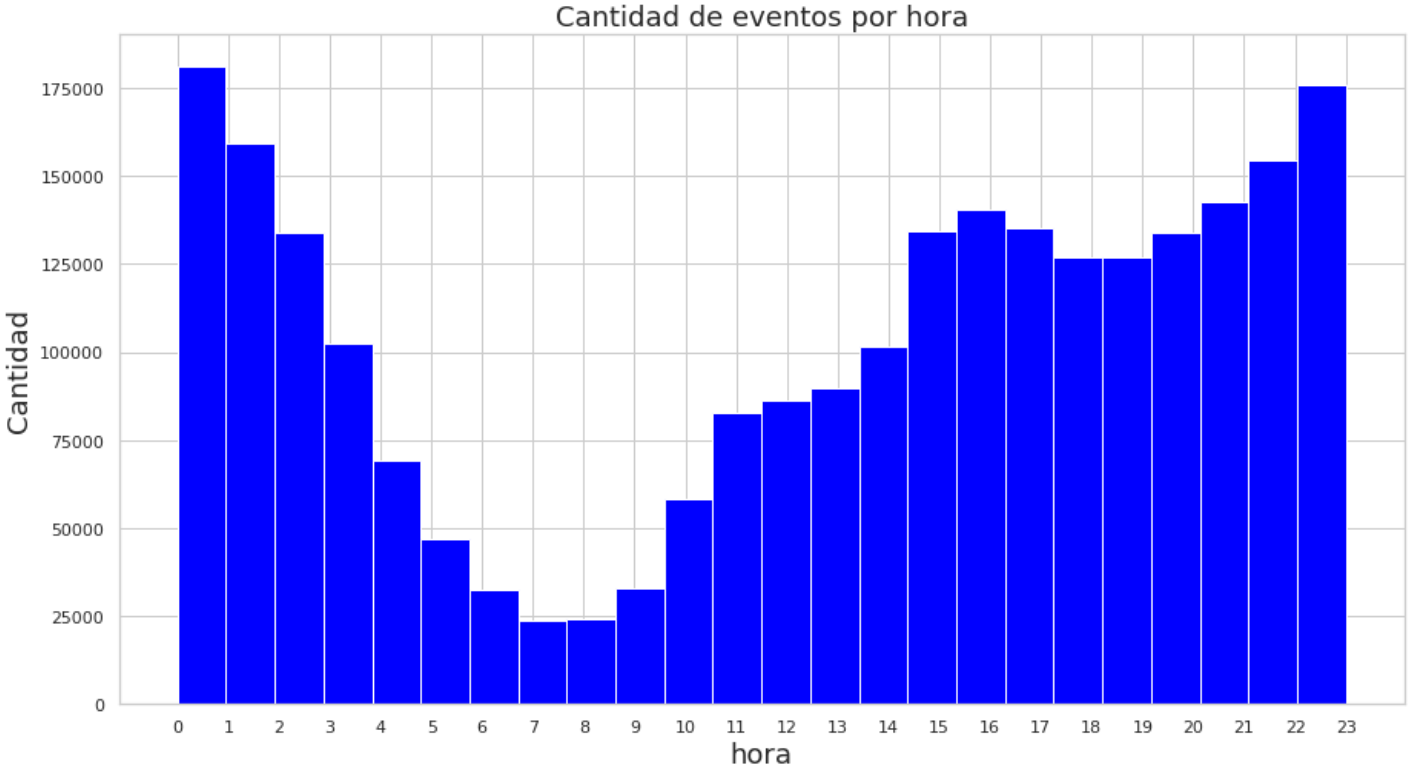
En cambio, para el fin de semana, el pico se encuentra solo en horas de la tarde.

¿Depende un evento o una conversión de un usuario del día de la semana? ¿Y de la hora?

Del set de datos de **eventos** se obtuvo la distribución de la cantidad de eventos a lo largo de la semana y de la hora. Los resultados obtenidos fueron los siguientes:

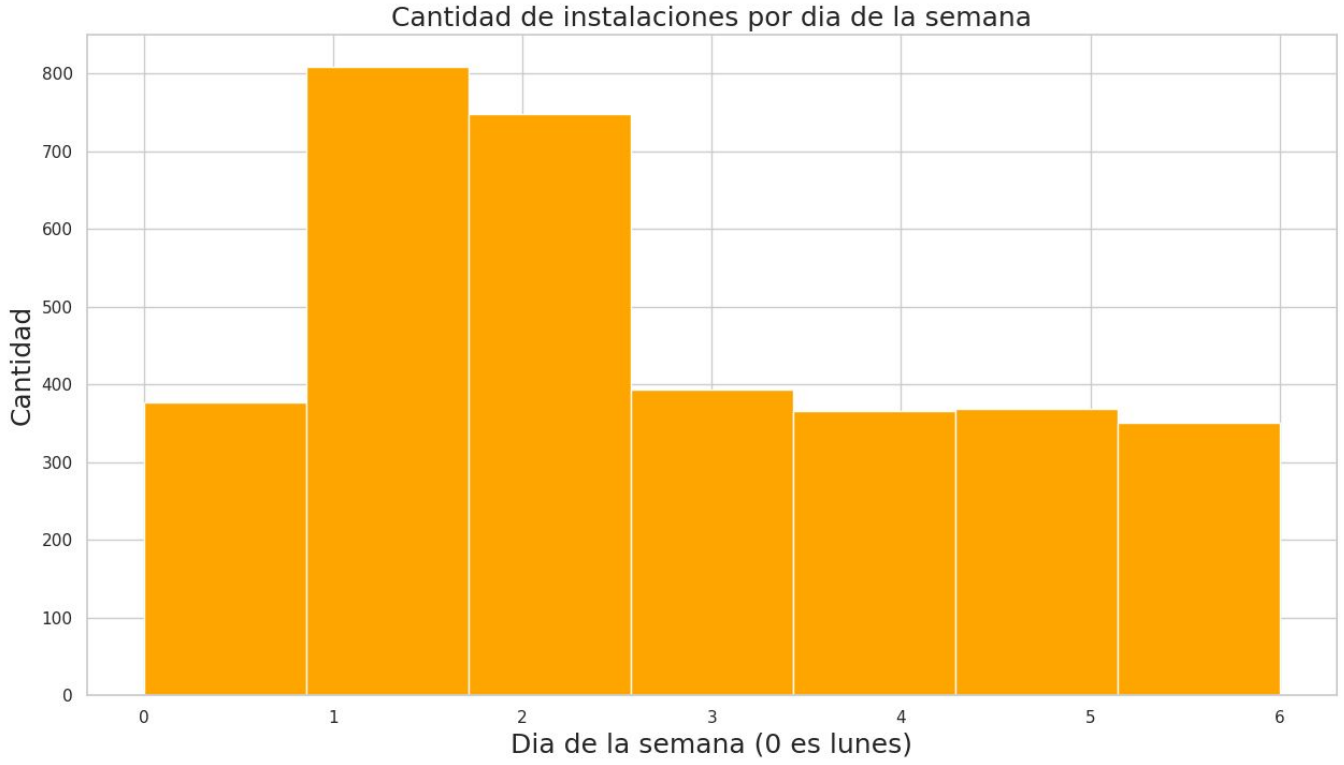


La cantidad de eventos se mantiene constante excepto por un pico producido el martes y el miércoles.
Durante el día:



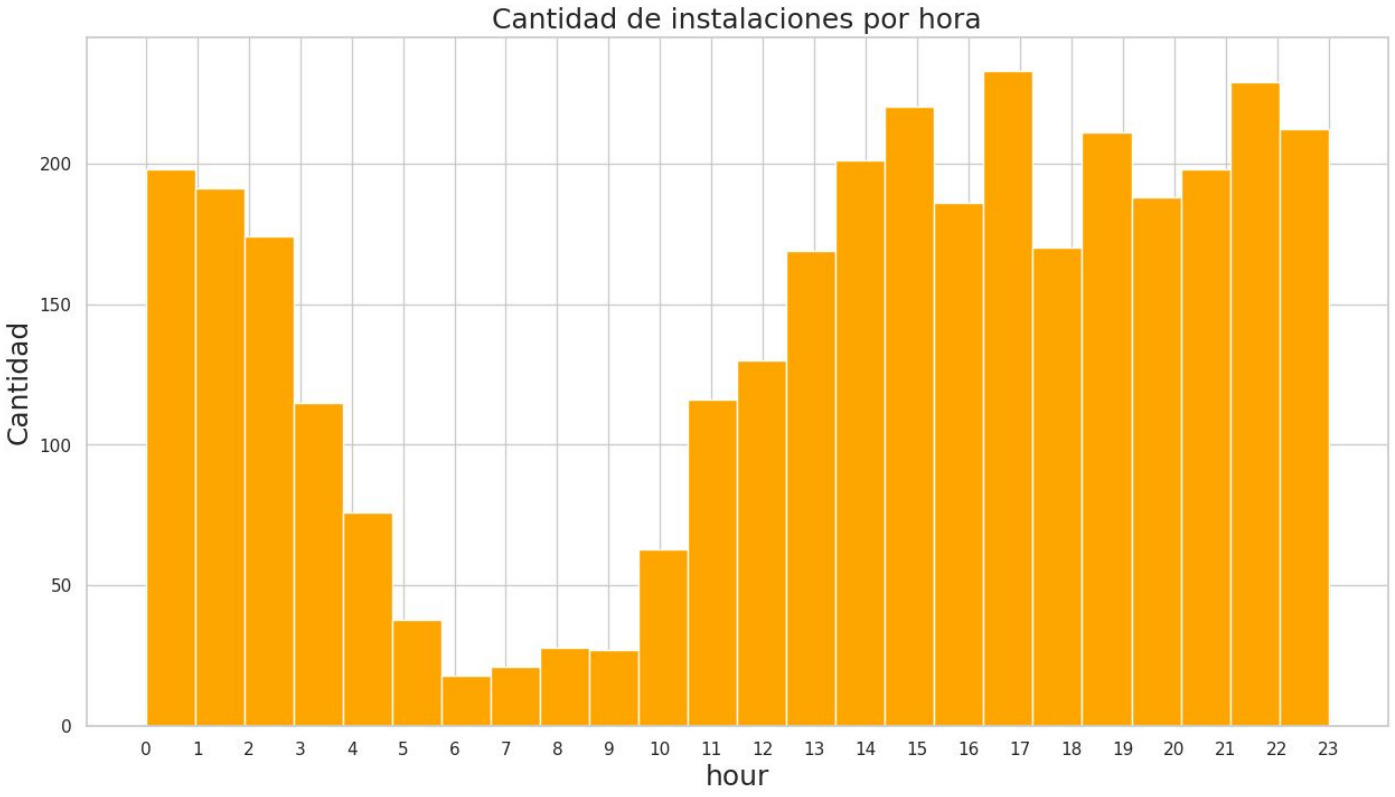
Vemos que la mayor cantidad de eventos suceden durante la tarde noche. Por la madrugada y la mañana temprana no suceden muchos eventos. Probablemente debido a que durante esa hora la mayoría de la gente está durmiendo.

Por otro lado, del set de datos de **instalaciones** se obtuvieron los siguientes resultados:



El resultado es similar al de eventos: constante durante la semana con un pico los **martes** y **miércoles**.

Durante el dia:

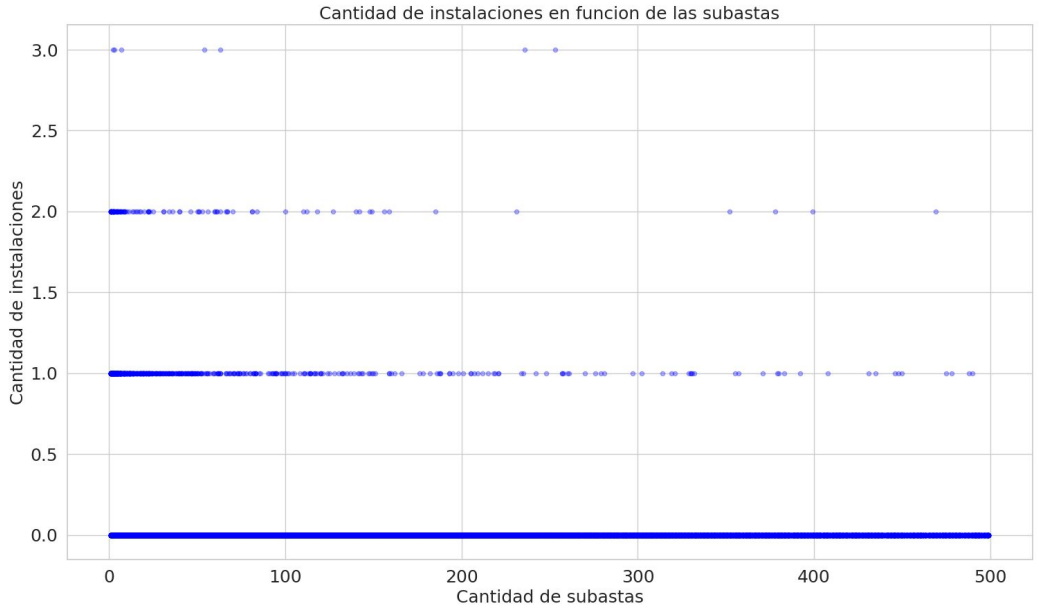


Otra vez, vemos una distribución muy similar a la de eventos, tenemos una mayor cantidad de instalaciones en los horarios de la tarde y noche y casi nula en los horarios de la mañana.

Esta pregunta es interesante debido a que podría ayudar a decidir si en un horario/día conviene apostar por un usuario o no.

¿Hay correlación entre cantidad de encuestas e instalaciones por dispositivo?

Nos interesa esta correlación para saber si más subastas sobre un dispositivo (que implican más publicidad consumida por el usuario), implica una mayor probabilidad de instalación. Es decir, estaremos observando si aquellos usuarios que instalaron, participaron a su vez en muchas subastas o no. Además observaremos si los usuarios que no instalaron, pero si participaron en subastas, tienen bajos números en lo que a subastas respecta. Los resultados fueron los siguientes:
Podemos ver que no hay correlación entre la cantidad de instalaciones con la cantidad de subastas participadas. Ya de entrada tengo muchísimos casos donde se participa de subastas y no se produce ninguna instalación, incluso en zonas de muchas subastas. Además se ve que en los casos donde se produjo una instalación, hay mayor densidad en la zona de pocas subastas.



¿Afecta la presencia de wifi a la cantidad de installs?

Es una pregunta interesante puesto que es común que la gente no esté dispuesta a instalar aplicaciones estar conectado a una red WiFi. Los resultados arrojados por el dataset de installs fueron que **~80%** de las instalaciones son realizadas estando conectados a una red WiFi. Esto puede ser bastante determinante a la hora de apostar por un dispositivo o no.

¿Cuales son las aplicaciones más populares?

Vamos a ver cuáles son las aplicaciones preferidas de los usuarios basándonos en las que tengan **mayor cantidad de usuarios distintos** y las que **más eventos generan**. Las aplicaciones serán representadas por su id único.

Las aplicaciones con **mayor cantidad de usuarios** fueron:

application_id	total_users
66	70312
63	18419
145	17576
64	17007
103	15053

Las aplicaciones que **más eventos generaron** fueron:

application_id	**total_events**
66	325696
64	259084
145	252431
63	181555
103	137513

Podemos ver que en ambos casos, las aplicaciones líderes son las mismas, con un ranking que difiere en una posición.

Conclusiones

Gracias al arduo trabajo de exploración de datos, se logró responder diversas preguntas que consideramos potencialmente útiles para la empresa. Las mismas fueron de diversa índole y apuntaban a identificar distintos patrones de comportamiento de los usuarios.

Se obtuvieron resultados en formato numérico y gráfico, para mejor visualización del lector.

Se lograron identificar intervalos temporales donde los usuarios están más predispuestos a generar eventos o convertir.

Se descubrieron zonas que son preferentes por los usuarios para clickear publicidades y qué aplicaciones son las más usadas.

A su vez, se determinó la duración de las publicidades para los usuarios, lo cuál podría ser útil para calcular los tiempos estimados hasta la conversión.

De esta manera, fuimos cubriendo varios aspectos de los datasets con el fin de aportar diversos “insights” que puedan ser información útil de manera directa o para lograr el proyecto propuesto para el próximo trabajo que involucra la posibilidad de estimar los tiempos de aparición en subastas y de conversión.

Apéndice

Link al repositorio:
https://github.com/gatti2602/7506-datos-tp1.git