

UNIVERSIDAD DE BUENOS AIRES

FACULTAD DE INGENIERÍA

**75.06 Organización de Datos Trabajo Práctico 1**

**Primer Cuatrimestre de 2019**

Grupo 18

|  |  |
| --- | --- |
| Camussi, Alan |  |
| Catolino, Lucas | 95560 |
| Zugna, Federico | 95758 |

**Link de GitHub:** <https://github.com/LucasCatolino/Orgnizacion-de-Datos> **(Privado)**

# Índice

# Introducción

En este trabajo se nos pidió analizar un set de datos provisto por la empresa Jammp, plataforma enfocada en la promoción y remarketing de aplicaciones móviles. La empresa busca promover apps, lograr que quien instaló la app efectúe una primera acción, y quien la usa que la siga usando.

Desde la empresa pusieron a nuestra disposición cuatro archivos .csv comprimidos, y un archivo .json explicando los datos. Estos eran:

* auctions.csv.gzip
* clicks.csv.gzip
* events.csv.gzip
* installs.csv.gzip
* desc.json

La finalidad del trabajo es realizar un análisis exploratorio de los datos, familiarizarnos con las herramientas vistas en clase, investigar sobre la materia para así poder visualizar resultados y analizar conclusiones que podamos encontrar.

# Análisis de clicks

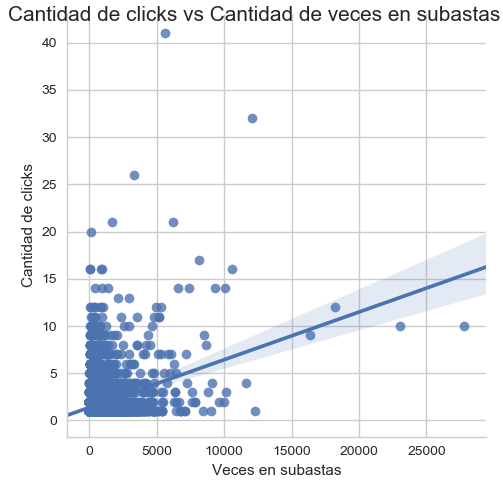
## Análisis de horarios

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

|  |
| --- |
|  |
|  |
|  |
|  |
|  |
|  |
|  |

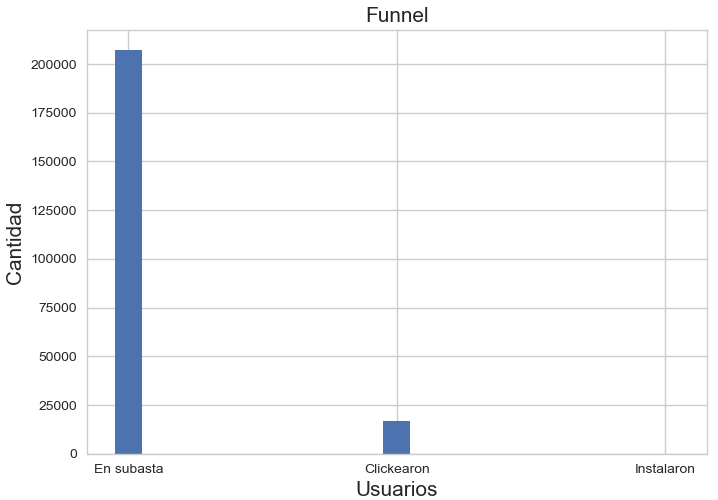
## Cantidad de clicks vs cantidad de veces en subasta

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

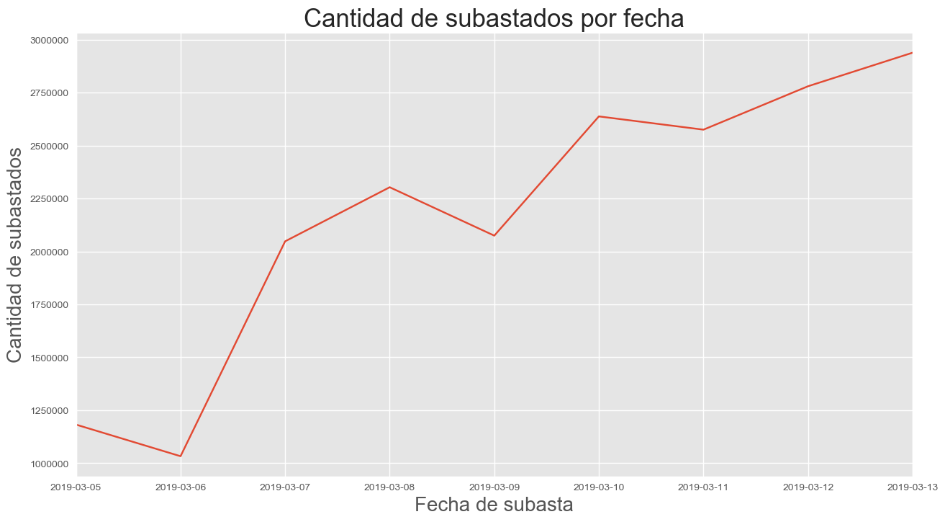


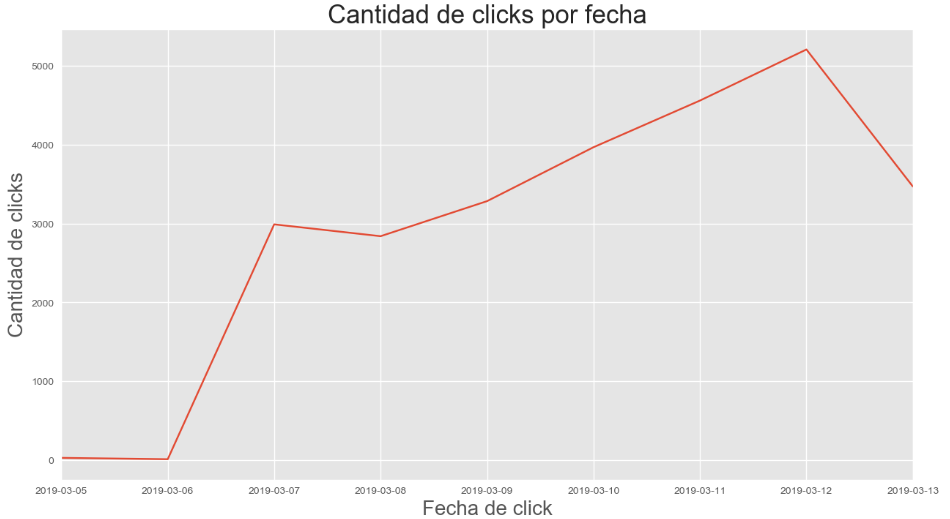
## Funnel de conversión

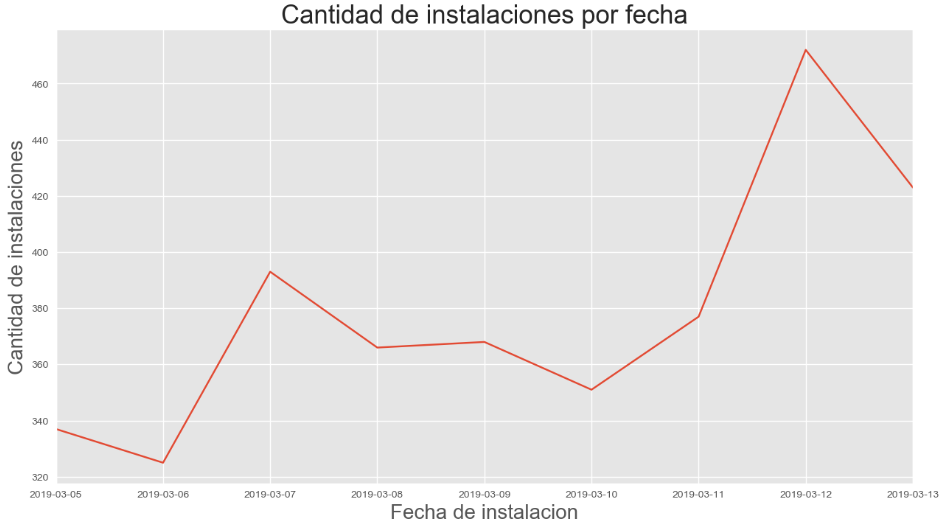
Tomando como objetivo la instalación de una app, se tomaron los siguientes pasos para la conversión de un usuario: el usuario aparece en subasta, el usuario clickea una publicidad, el usuario instala la aplicación. Del total de los usuarios poco más de 200.000 usuarios que entraron en subasta en los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



Analizando por día cada una de las variables por separado:

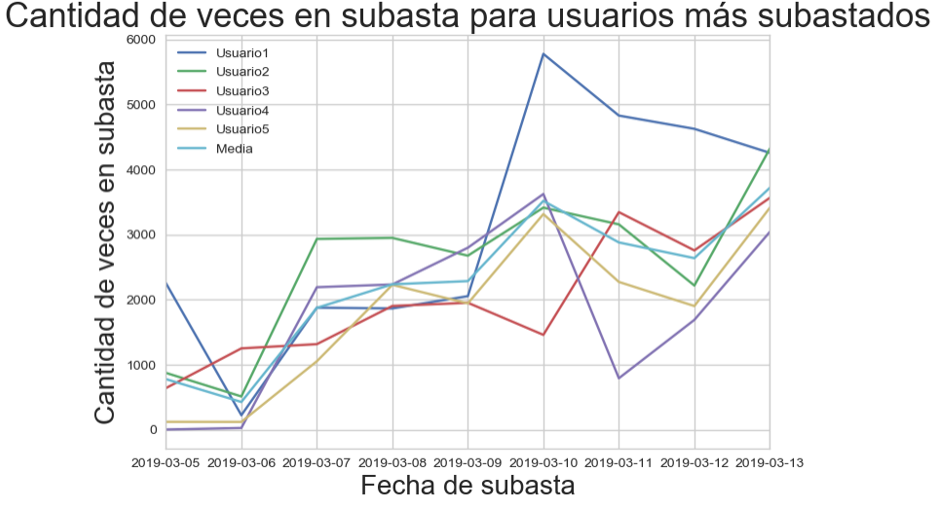






## Análisis de los usuarios que más aparecen en subasta

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



## Conclusión

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

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

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

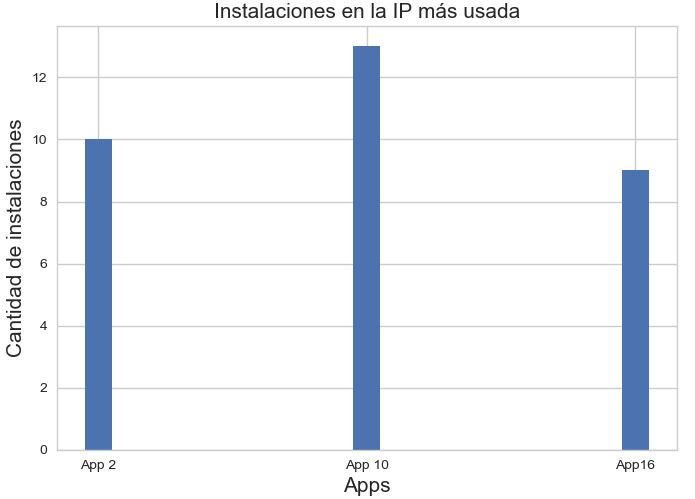
En cuanto a cada proceso por separado, el mejor día para subastas, clicks e instalaciones fue el martes 12 de marzo. También puede verse una conclusión anterior: no por haber muchas subastas tiene que haber muchos clicks o instalaciones. El dia con mayor cantidad de subastas fue el miércoles 13 de Marzo,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

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

# Análisis sospechoso

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

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



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