

Análisis Exploratorio de Datos

Trabajo Práctico 1 - Organización de Datos

Grupo "Back to the Data"

22/04/2019

| Nombre | Padrón | Mail |
|------------------------|--------|----------------------------|
| Álvarez, Federico | 99266 | fede.alvarez1997@gmail.com |
| La Torre, Gabriel | 87796 | latorregab@gmail.com |
| Medrano, Lucas Nicolás | 99247 | lucasmadrano97@gmail.com |
| Piro Martino, Ariel | 99469 | ariel.piro@hotmail.com |

Índice

| | |
|--|----------|
| 1. Introducción | 4 |
| 1.1. Repositorio | 4 |
| 2. Análisis individual de archivos | 4 |
| 2.1. Subastas | 4 |
| 2.1.1. Análisis general | 4 |
| 2.1.2. Análisis de los dispositivos en las subastas | 4 |
| 2.1.3. Subastas por día de Marzo | 7 |
| 2.1.4. Subastas por día de Marzo por sistema operativo | 7 |
| 2.1.5. Subastas por hora del día | 8 |
| 2.1.6. Subastas por hora del día por sistema operativo | 9 |
| 2.1.7. Subastas por sistema operativo | 10 |
| 2.1.8. Subastas por source | 11 |
| 2.2. Clicks | 13 |
| 2.2.1. Análisis General | 13 |
| 2.2.2. Clicks en los distintas publicidades | 13 |
| 2.2.3. Clicks en los source_id | 14 |
| 2.2.4. Clicks en los carrier_id | 14 |
| 2.2.5. Clicks en los os_minor | 15 |
| 2.2.6. Clicks en los os_major | 16 |
| 2.2.7. Clicks en los agent_device | 16 |
| 2.2.8. Clicks en los spec_brands | 17 |
| 2.2.9. Clicks en las distintas marcas | 17 |
| 2.2.10. Clicks en Android e IOS | 18 |
| 2.2.11. Clicks en ref_hash | 19 |
| 2.2.12. Clicks en los días y horas | 20 |
| 2.2.13. Posición geográfica del click | 24 |
| 2.2.14. Clicks en la pantalla del dispositivo | 25 |
| 2.2.15. Tiempo en realizar el click | 25 |
| 2.3. Eventos | 28 |
| 2.3.1. Introducción | 28 |
| 2.3.2. Eventos por fecha y hora | 28 |
| 2.3.3. Eventos por aplicación y detalles de los dispositivos | 30 |
| 2.3.4. Eventos más frecuentes por tipo | 32 |
| 2.3.5. Eventos por tipo de conexión | 35 |
| 2.3.6. Marcas y eventos | 38 |
| 2.3.7. Modelos y eventos | 40 |
| 2.3.8. Ciudades y eventos | 41 |
| 2.4. Instalaciones | 43 |
| 2.4.1. Introducción | 43 |
| 2.4.2. Instalaciones por día y hora | 43 |
| 2.4.3. Instalaciones por aplicación | 45 |
| 2.4.4. Instalaciones por fecha según la aplicación | 45 |
| 2.4.5. Instalaciones por país | 47 |
| 2.4.6. Instalaciones por tipo | 47 |

| | | |
|-----------|---|-----------|
| 2.4.7. | Aplicaciones por país y tipo | 48 |
| 2.4.8. | Idiomas y aplicaciones | 49 |
| 2.4.9. | Instalaciones por marca | 50 |
| 2.4.10. | Instalaciones wifi y user agents | 51 |
| 2.4.11. | Instalaciones atribuidas a Jampp | 53 |
| 3. | Análisis de archivos en conjunto | 55 |
| 3.1. | Clicks y subastas | 55 |
| 3.1.1. | Análisis general | 55 |
| 3.2. | Clicks e instalaciones | 58 |
| 3.2.1. | Análisis general | 58 |
| 3.2.2. | Clicks e instalaciones en advertiser_id | 59 |
| 3.2.3. | Clicks e instalaciones en source_id | 59 |
| 3.2.4. | Posicion geografica de clicks e instalaciones | 60 |
| 3.2.5. | Clicks en la pantalla que terminan en instalaciones | 62 |
| 3.2.6. | Tiempo en realizar el click | 62 |
| 3.2.7. | Clicks e instalaciones en Android e IOS | 63 |
| 3.2.8. | Clicks e instalaciones en carreirs_id | 63 |
| 3.2.9. | Clicks e instalaciones en os_minor | 64 |
| 3.2.10. | Clicks e instalaciones en os_minor | 65 |
| 3.2.11. | Clicks e instalaciones en specs_brand | 65 |
| 3.2.12. | Clicks e instalaciones en país de origen | 66 |
| 3.2.13. | Clicks e instalaciones en application_id | 66 |
| 3.2.14. | Tipo de instalaciones | 67 |
| 3.2.15. | Clicks e instalaciones en device_model | 68 |
| 3.2.16. | Clicks e instalaciones en device_language | 68 |
| 3.2.17. | Clicks e instalaciones en session_user_agent | 69 |
| 3.2.18. | Clicks e instalaciones en ip_address | 69 |
| 3.2.19. | Clicks e instalaciones en los días y horas | 70 |
| 4. | Conclusiones | 74 |

1. Introducción

En el trabajo se hace un análisis exploratorio de un set de datos provistos por la empresa Jampp. En el mismo se encuentra información de subastas, instalaciones, clicks, entre otros.

Primero se hará una visión general de los archivos `installs.csv`, `clicks.csv`, `auctions.csv` y `events.csv` para entender la distribución y la cantidad de datos, el significado de las columnas, reconocer las columnas que no aportan información, por ejemplo, las que tienen todos sus valores nulos, y reconocer el tipo de datos en cada columna, seguido de un análisis más profundo para obtener más información de los datos.

Luego se hará un análisis global, buscando información relativa a los archivos en conjunto, permitiendo obtener otro tipo de información.

1.1. Repositorio

El análisis exploratorio realizado con pandas se puede visualizar en el siguiente link de github: <https://github.com/shizus/7506-1c2019>.

2. Análisis individual de archivos

2.1. Subastas

2.1.1. Análisis general

El archivo `'auctions.csv'` contiene información acerca de subastas.

Hay dos columnas que no nos aportan información significativa. `'auction_type_id'` tiene todos sus valores nulos, por lo que no fue tomada en cuenta para el análisis. `'country'` informa un solo valor posible.

La columna `platforms` tiene dos valores posibles (1 y 2) que se supone son Android e iOS. Va a ser importante para el análisis que hagamos más adelante. De ahora en más, `platform` y sistema operativo serán sinónimos en este informe.

Por último, `'source'` nos da información acerca del exchange de donde surge la subasta.

Además vemos que ningún valor de este archivo, excluyendo la columna `'auction_type_id'`, es nulo, por lo que no es necesario tomar ninguna decisión respecto a eso.

2.1.2. Análisis de los dispositivos en las subastas

Un punto muy importante para analizar son los dispositivos. Esto trajo un problema al analizar los datos y puede cambiar la forma en la que se entienden los mismos.

Podemos, por ejemplo, ver el top diez de dispositivos de los cuales se generan más subastas.

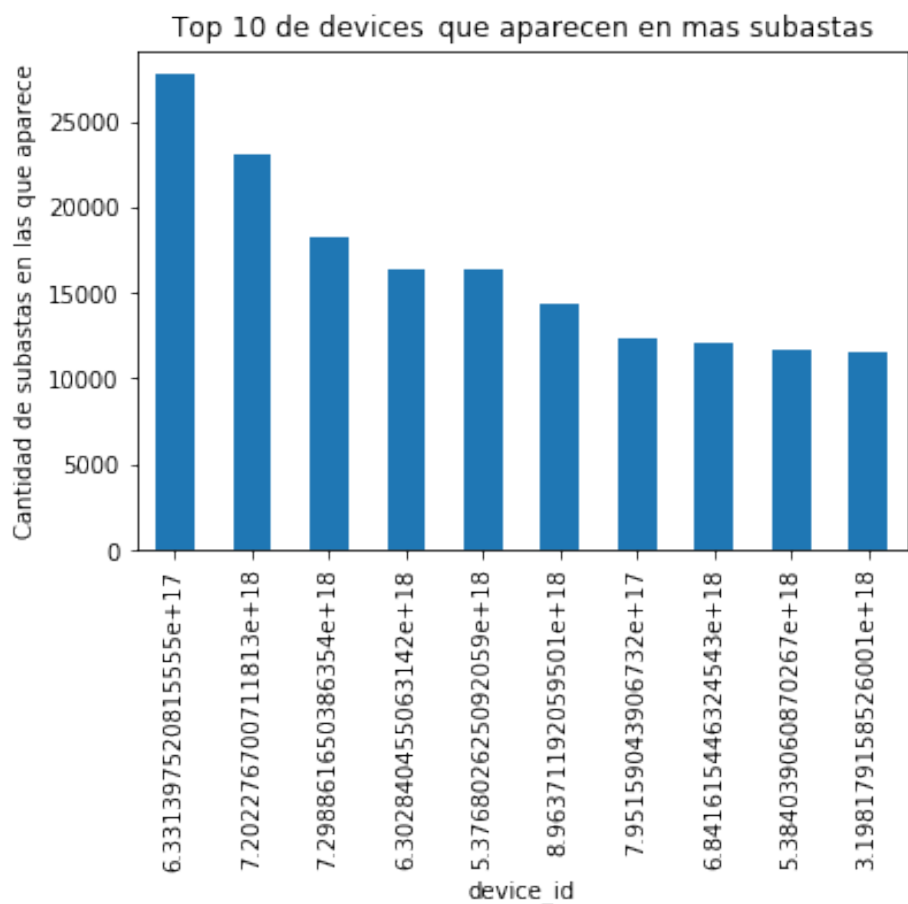


Figura 1: Top 10 de dispositivos con mayor cantidad de subastas

Ahora el problema.

En total, en las 19571319 subastas, aparecen 206171 dispositivos. Algo interesante sería conocer cuántos tienen cada plataforma. Por lo que hacemos un gráfico.

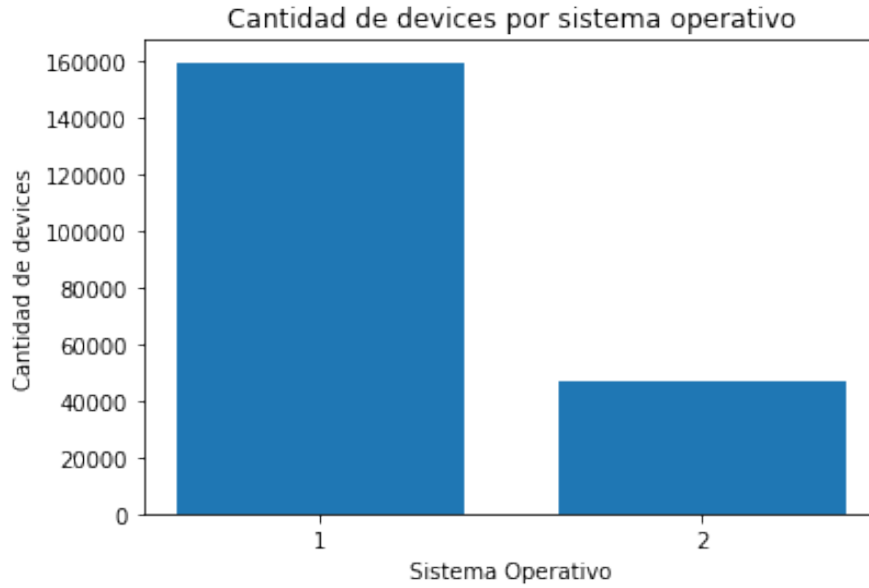


Figura 2: Cantidad de dispositivos para cada plataforma

Al obtener los datos del gráfico, observamos que la suma de las cantidades de dispositivos para cada plataforma es 206453. Esto es mayor a la cantidad total de dispositivos (206171).

La causa de este problema se entiende al conocer cómo se obtuvieron las cantidades. Primero se contó la cantidad de dispositivos total. Luego se separaron los que tuvieran plataforma '1' de los que tuvieran plataforma '2' (tener en cuenta que como hay una fila por subasta, cada dispositivo puede aparecer en filas distintas). Al sumar ambas cantidades obtenemos un número mayor que el total, por lo que es evidente que hay dispositivos que tienen ambas plataformas, es decir, en algunas filas aparecen con un valor de la columna 'platform' y en otras filas con otro.

Haciendo una simple resta, se puede ver que la cantidad de dispositivos que aparecen con plataformas distintas es 282, y que en total participan en 47862 subastas. De hecho hay un dispositivo ($5.292967062497395e+18$) que aparece en el top 300 de cantidad de subastas.

Esto puede alterar los análisis que incluyen división en plataformas. Sin embargo, no es trivial que haya que borrar estos datos, porque se estarían borrando casi 50000 subastas. Además, dejar los datos como están no altera a los estudios que se hagan sin dividir el problema por sistemas operativos. Por lo tanto se decidió trabajar de la siguiente manera.

- En aquellos análisis que, a priori, no parezcan ser afectados por estos errores en las plataformas de los dispositivos usaran el dataset entero.
- En los restantes, se usará un dataframe filtrado, en el que no están los dispositivos conflictivos.

Igualmente, se compararon los gráficos conflictivos para el dataframe original y el filtrado, y no había diferencias substanciales cualitativas ni cuantitativas.

2.1.3. Subastas por día de Marzo

Es interesante ver cómo se distribuye la cantidad de subastas en los días que incluye el archivo (05/03/19 al 13/03/19). El gráfico 3 representa dicha distribución. En él se pueden observar varias cosas:

- La cantidad de subastas parece, en general, aumentar con el paso de los días.
- El valor del último día es más del doble del valor del primer día.
- El mayor aumento se da del sexto al séptimo día.
- La cantidad de subastas varía entre unos valores extremos, aproximados, de un millón y 3 millones.

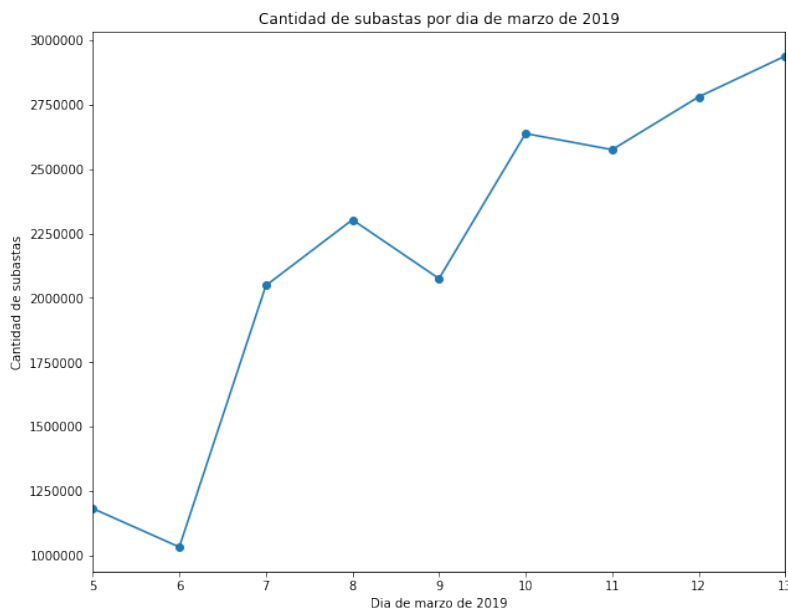


Figura 3: Cantidad de subastas por día de Marzo

2.1.4. Subastas por día de Marzo por sistema operativo

Otro punto interesante es dividir el problema. Obtener la distribución de subastas en los días con datos disponibles para cada plataforma (Android e iOS). En la imagen 4 se pueden observar las cantidades. Obsérvese que llamamos '1' y '2' a las plataformas, ya que no sabemos cuál es Android y cuál es iOS.

Puntos interesantes a reconocer:

- Para este análisis se usaron los datos filtrados, ya que los devices que tienen ambas plataformas pueden generar ruido.
- La cantidad de subastas para la plataforma '1' es, salvo en el cuarto y el quinto día, considerablemente mayor a la cantidad para la plataforma '2'.
- La figura de la plataforma '2' es mucho más “chata” que la de la plataforma '1', la cual representa más picos y saltos.

- La figura de la plataforma '1' es muy parecida a la del gráfico 3, mientras que la de la plataforma '2' no lo es. Esto es resultado, principalmente, de lo indicado en el segundo ítem. Este análisis puede llegar a ser muy útil para reconocer partes de los datos que son representativas del total.

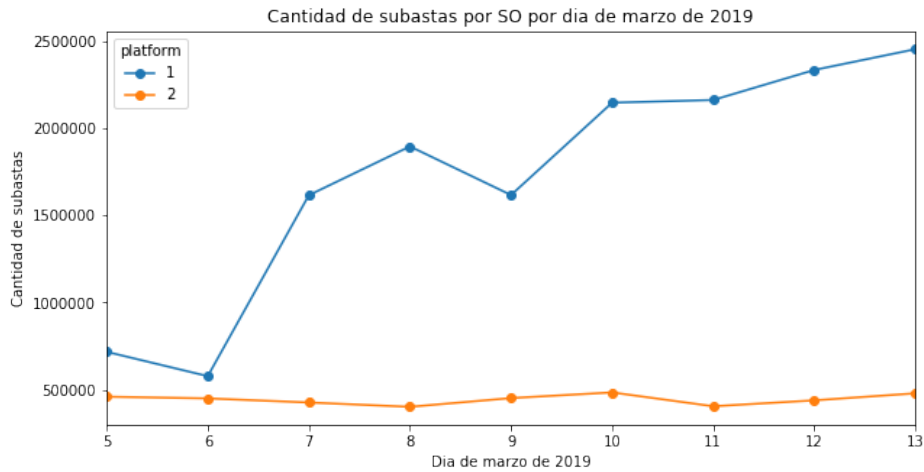


Figura 4: Cantidad de subastas por día de Marzo y por sistema operativo

2.1.5. Subastas por hora del día

Ahora vamos a analizar cómo se distribuyen las subastas a lo largo del día. Para esto hacemos un gráfico de hora del día contra cantidad de subastas (Gráfico 5).

Desde un análisis cualitativo se pueden observar algunos puntos:

- Parece ser que la mayor cantidad de subastas se distribuyen por la noche y la madrugada.
- La cantidad de subastas es poca en horas de la mañana y el mediodía. En el gráfico se puede ver un gran valle en esa parte del día.

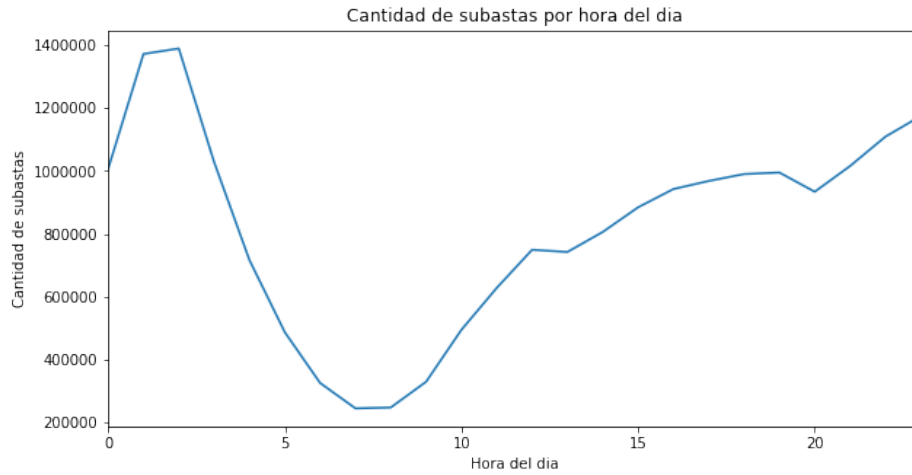


Figura 5: Cantidad de subastas por hora del día

2.1.6. Subastas por hora del día por sistema operativo

Al igual que se hizo antes podemos dividir el gráfico para ambas plataformas. Vemos que pasa algo muy parecido que lo que pasaba para la cantidad de subastas por día. El gráfico de la plataforma '1', al tener una cantidad mucho mayor de subastas, es la que predomina en el gráfico de la sección "Subastas por hora del día", y por eso sus gráficos son tan parecidos. El gráfico de la plataforma '2' es bastante más chato, y con cantidades de subastas mucho más chicas.

En esta subsección y en la siguiente usaremos los datos filtrados.

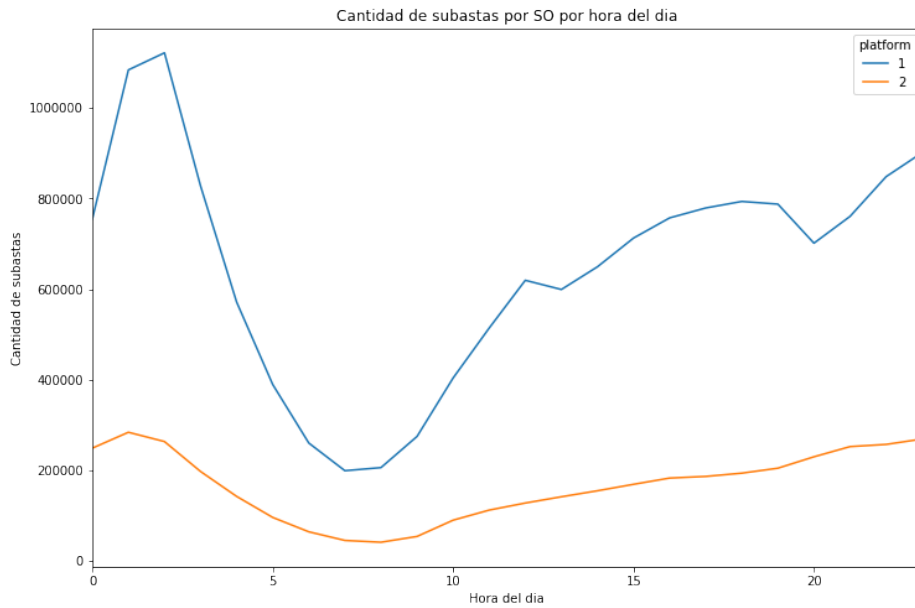


Figura 6: Cantidad de subastas por hora del día por sistema operativo

2.1.7. Subastas por sistema operativo

Resulta interesante conocer cuál es el sistema operativo para el cual se generan más subastas. Es algo que ya se venía viendo en la forma y nivel (cantidad de subastas) de los gráficos anteriores. Sin embargo, los gráficos vistos hasta ahora no dan un conocimiento directo de la relación de las cantidades de subastas de ambas plataformas.

En las figuras que se ven a continuación podemos confirmar que lo que indicábamos en los gráficos anteriores era cierto. La cantidad de subastas es mucho mayor para la plataforma '1'. Además, ahora tenemos una visión más cuantitativa de esta relación, en la imagen 7 vemos una relación porcentual, mientras que la imagen 8 nos da más idea de las cantidades.

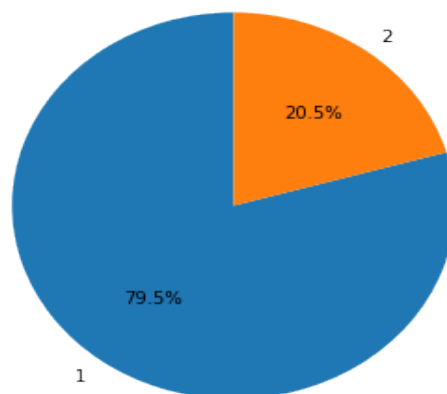


Figura 7: Porcentaje de subastas para cada plataforma

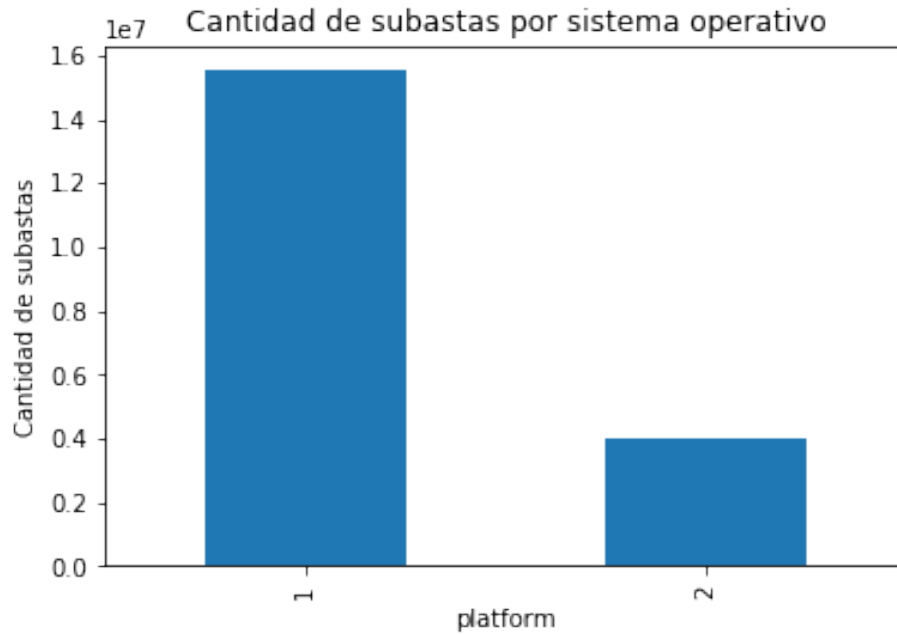


Figura 8: Porcentaje de subastas para cada plataforma

2.1.8. Subastas por source

Como indica la introducción a esta sección (vease 2.1.1), source nos indica el exchange que generó la subasta. Se puede obtener, a partir de los datos, cuáles son los exchanges principales, y cuántas subastas generan.

A tener en cuenta:

- Los que se muestran son todos los exchanges que aparecen en el archivo.
- Al igual que en las plataformas, los exchanges se toman por un id, y no por su nombre.
- Hay una clara diferencia en las cantidades.
 - El exchange '0' es predominante, superando ampliamente el millón de subastas generadas.
 - El siguiente, source '1', aunque es mucho menor que el '0', sigue superando a los demás exchange por una gran cantidad. Llegando a las 400000 subastas generadas.
 - Los exchanges '2', '5' y '6' parecen no tener mucho peso en el gráfico. Aunque quizás podría tomarse la cantidad del '5' como significativa.

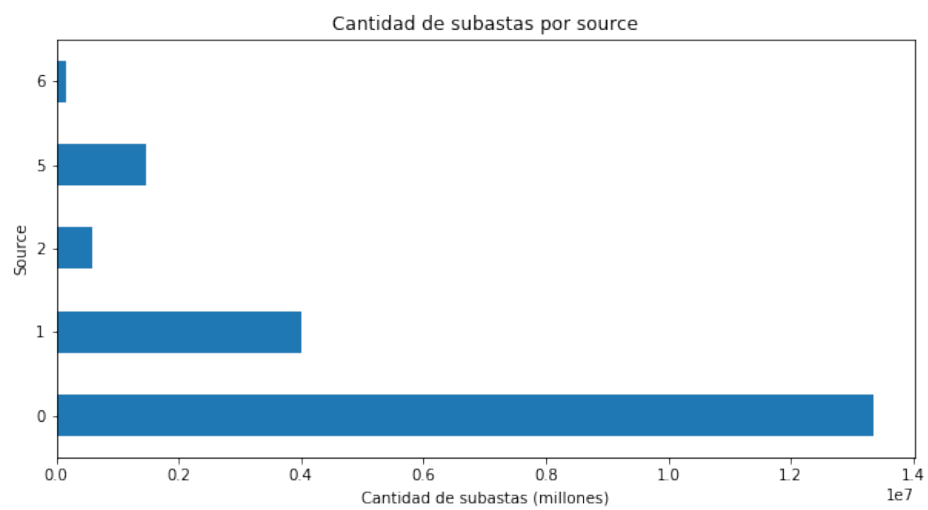


Figura 9: Cantidad de subastas para cada source

2.2. Clicks

2.2.1. Análisis General

En el archivo `clicks.csv` se nos muestran los clicks que se realizaron entre el 5 y el 13 de Marzo de este año. Este csv cuenta con 20 columnas y 26000 filas aproximadamente. En las columnas tenemos datos como en qué latitud y longitud se hizo el click, cuánto tardó el usuario en hacer el click, de qué país es el dispositivo, en qué parte de la pantalla del celular se realizó click, entre otras cosas.

En primera medida sacamos las columnas que no nos aportaban nada. La columna `action_id` tenía todos NaN así que fue removida. La columna que indicaba el país solo daba como resultado un país únicamente así que nos pareció irrelevante y la quitamos. Así también eliminamos la columna de `wifi_connection` ya que todas daban False, lo cual es un poco raro, porque esto indicaría ninguno estaba conectado a wifi. Finalmente se sacó la columna `trans_id` ya que era un identificador único para cada uno de los clicks y para analizar este archivo solo no nos aportaba nada.

También, se tomó la decisión de cambiar los tipos de algunas columnas para ahorrar más espacio. Las columnas `advertiser_id` y `source_id` se las cambió a `category` ya que tenían pocos valores y siempre eran los mismos. Y a varias de las columnas se las pasó de `float64` a `float32`, ya que no necesitaban tantos bits para representar sus datos.

2.2.2. Clicks en los distintas publicidades

Se van a observar la cantidad de clicks en las distintas publicidades

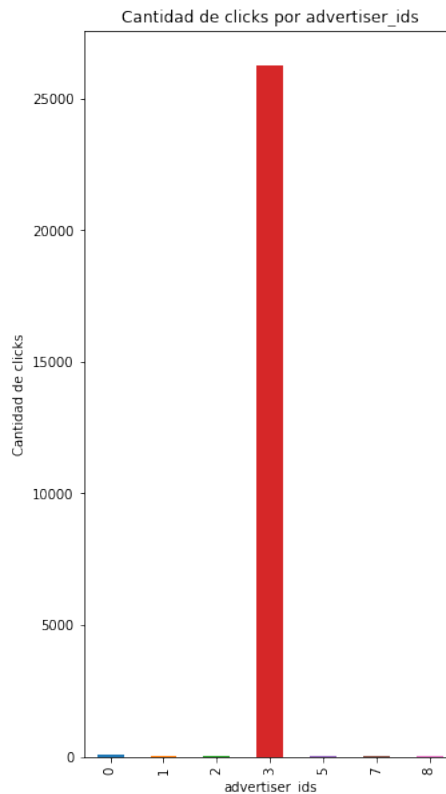


Figura 10: Clicks en los advertiser_id

Se puede observar que prácticamente todos son clicks en la publicidad con el id 3. El porcentaje de clicks en este id es de 99.7. Al parecer, le sirve mucho más a Jampp mostrar las publicidades del id 3 o es una publicidad de una app que es más común que todos usen.

2.2.3. Clicks en los source_id

En este gráfico vamos a mostrar la cantidad de clicks que se hacen en los distintos source_id que hay en el archivo

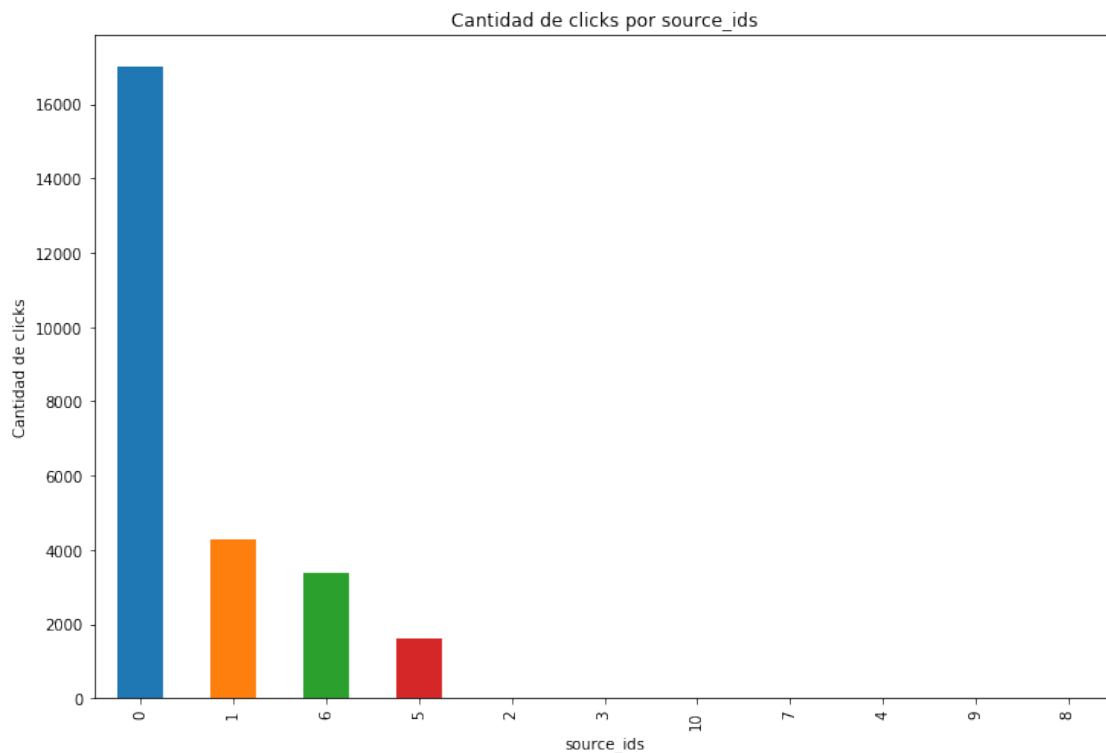


Figura 11: Clicks en los source_id

Como se puede observar la mayoría de los clicks estan distribuidos entre los id 0, 1, 6 y 5 siendo el source_id 1 el que más cantidad de clicks tiene.

2.2.4. Clicks en los carrier_id

En el siguiente gráfico se mostrará la cantidad de clicks que tiene cada carrier_id

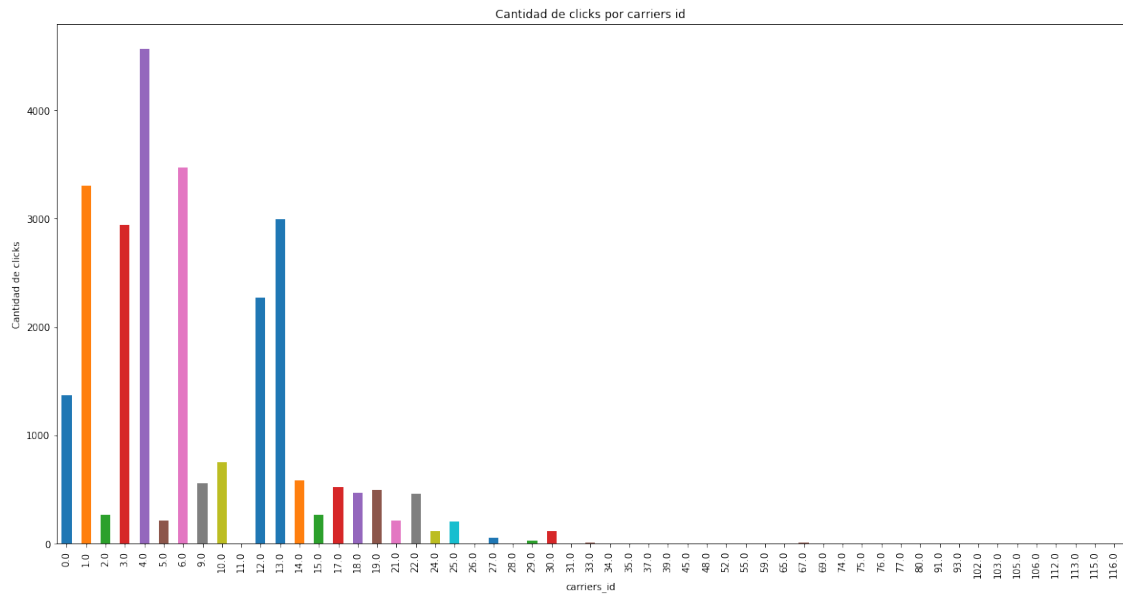


Figura 12: Clicks en los carrier_id

Por más que hay varios carrier_id se ve que la mayoría de clicks se distribuyen en aproximadamente 20 carriers_id. El carrier_id que más realiza clicks es el 4.0.

2.2.5. Clicks en los os_minor

Se mostrará en este gráfico la cantidad de clicks que realizan los dispositivos que tienen ciertas mínimas versiones de sus sistemas operativos.

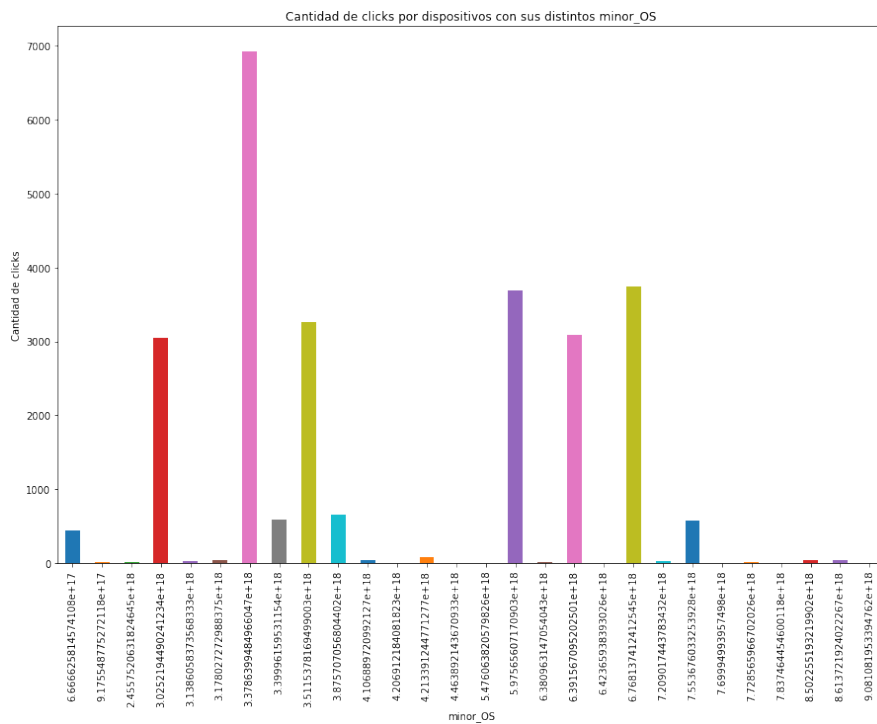


Figura 13: Clicks en los os_minor

Vemos que la cantidad de clicks es liderada por los dispositivos con un `os_minor` 3.378640e+18.

2.2.6. Clicks en los `os_major`

Se mostrará en este gráfico la cantidad de clicks que realizan los dispositivos que tienen ciertas máximas versiones de sus sistemas operativos.

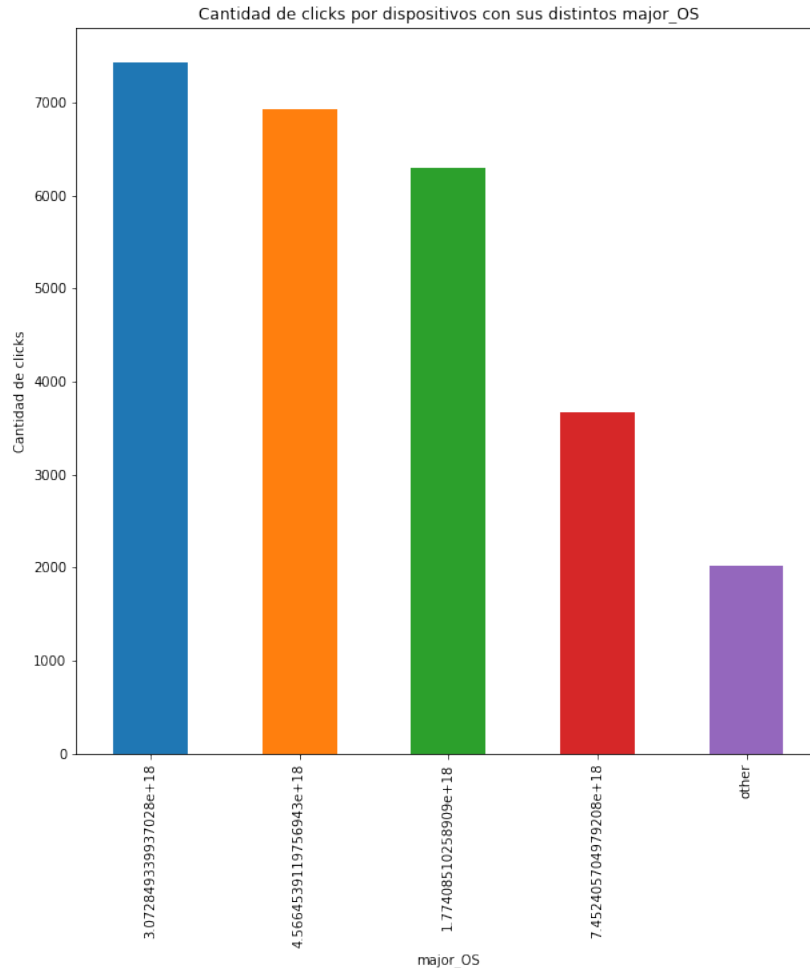


Figura 14: Clicks en los `os_major`

En este casos se reparte bastante equitativamente la cantidad de clicks en cuatro distintos `os_major`. No podemos saber cuanto es la cantidad total de los `os_major` debido a que una categoría es “others” y no sabemos cuántos distintos sistemas operativos tiene encerrado ese término.

2.2.7. Clicks en los `agent_device`

En este caso notamos que la cantidad de `agent_device` es mucha. Hay un total de 190 diferentes `agent_device` y la mayoría de sus datos son NaN. Solamente posee alrededor de 3000 datos. Por lo tanto, llegamos a la conclusión de que no aporta ningún dato importante ni tampoco es confiable la información que dé ya que la cantidad de datos que tiene es muy poca.

2.2.8. Clicks en los spec_brands

A continuación mostraremos un gráfico donde podemos ver la cantidad de clicks que tienen los distintos spec_brands.

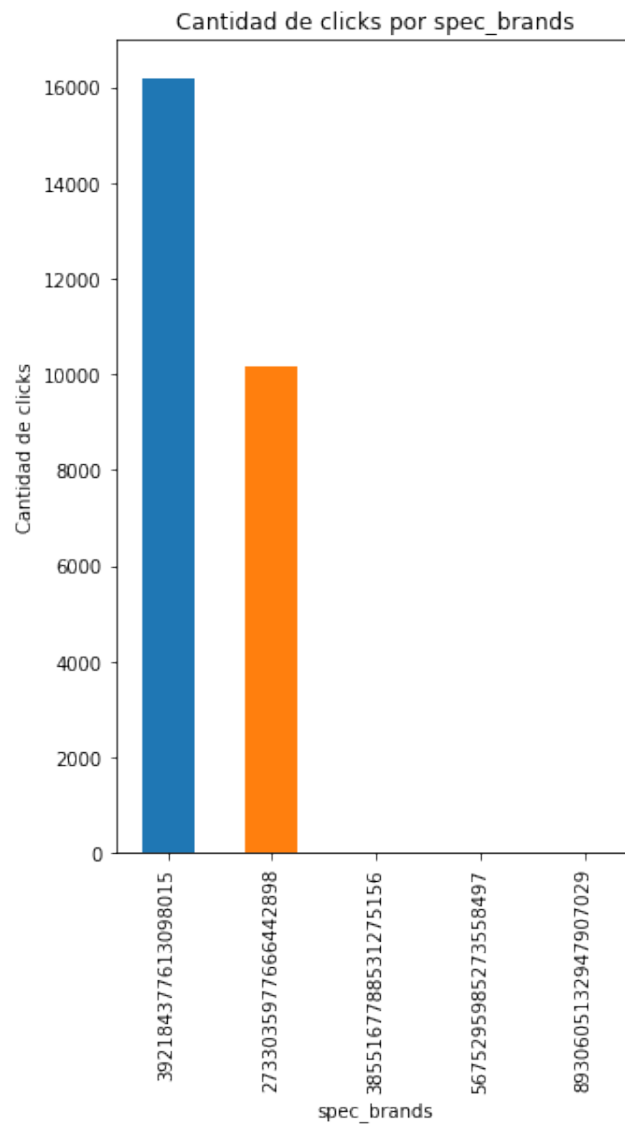


Figura 15: Clicks en los spec_brands

Por más que hay 5 spec_brands, se ve claramente que los clicks se realizan prácticamente dos spec_brands. Estos son 392184377613098015 y 2733035977666442898.

2.2.9. Clicks en las distintas marcas

Este gráfico nos muestra la cantidad de clicks en las distintas marcas de celular.

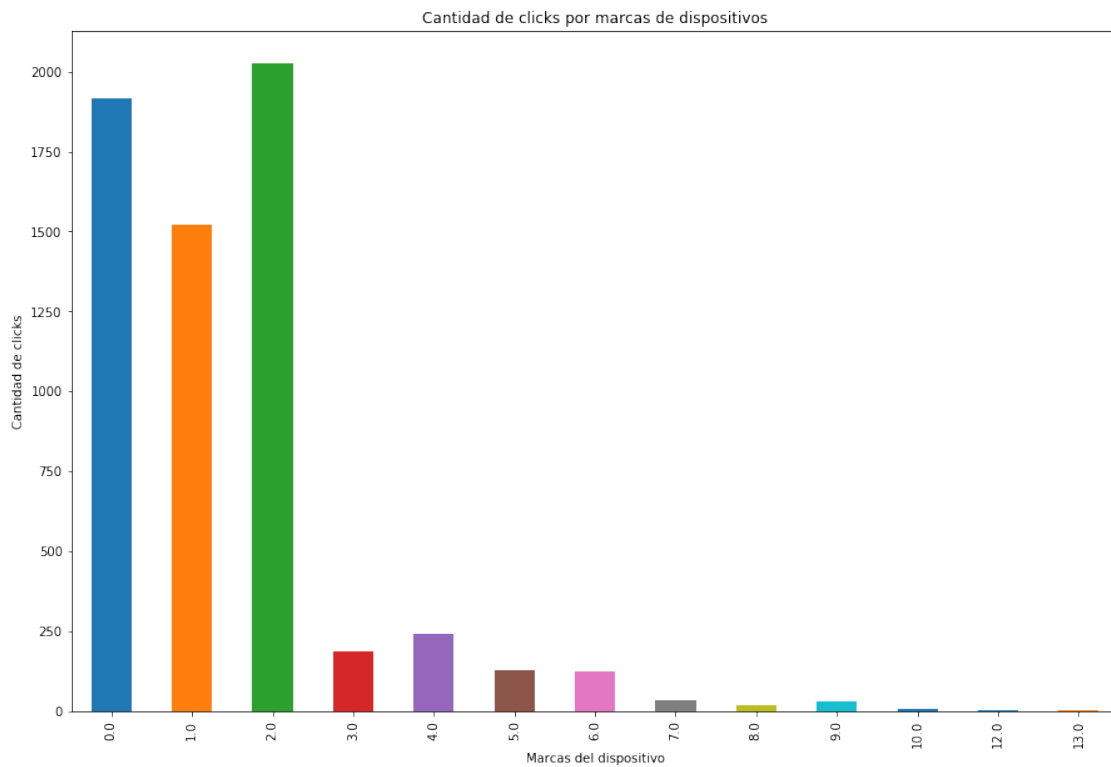


Figura 16: Clicks en las distintas marcas de teléfono

Como se puede observar, hay un total de 13 marcas distintas de teléfonos y la mayoría de clicks la realizan celulares de la marca 2.0, 0.0 y 1.0.

2.2.10. Clicks en Android e IOS

En el siguiente gráfico se muestra la cantidad de clicks que tienen los celulares con Android o IOS.

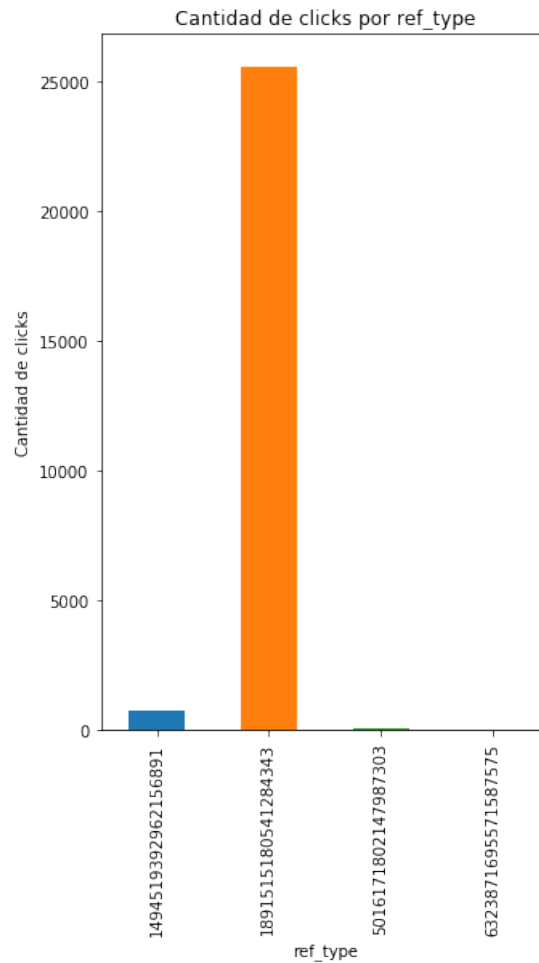


Figura 17: Clicks en las distintos sistemas operativos

Algo raro que se observa es que hay un total de cuatro sistemas operativos. Esto esta mal debido a que solo debería haber Android o IOS. Por lo tanto, se puede concluir que los dos valores que menos clicks tienen son datos erróneos. Es decir, solo hay que tener en cuenta los que clicks que tienen como sistema operativo los dos que más clicks concentran en este gráfico.

2.2.11. Clicks en ref_hash

El ref_hash nos muestra un id único de cada dispositivo que realiza click. Encontramos que hay ciertos ref.hash que se repiten varias veces. Por ejemplo, el que más veces lo hace, realiza un total de 41 clicks. La verdad es que es muy raro que una persona toque click 41 veces en 9 días. Por lo tanto, no podemos descartar que esos casos con muchos clicks sea un fraude. Hay muchísimos que tienen más de 10 clicks. Es información que hay que manejar con cuidado ya que uno no sabe si son usuarios de verdad. Son números difíciles de creer a priori, basándonos en la experiencia personal y la intuición, pero, aunque improbables, no son imposibles.

2.2.12. Clicks en los días y horas

En este primer gráfico se puede ver la evolución de la cantidad de clicks que se hicieron entre los días 5 y 13 de Marzo

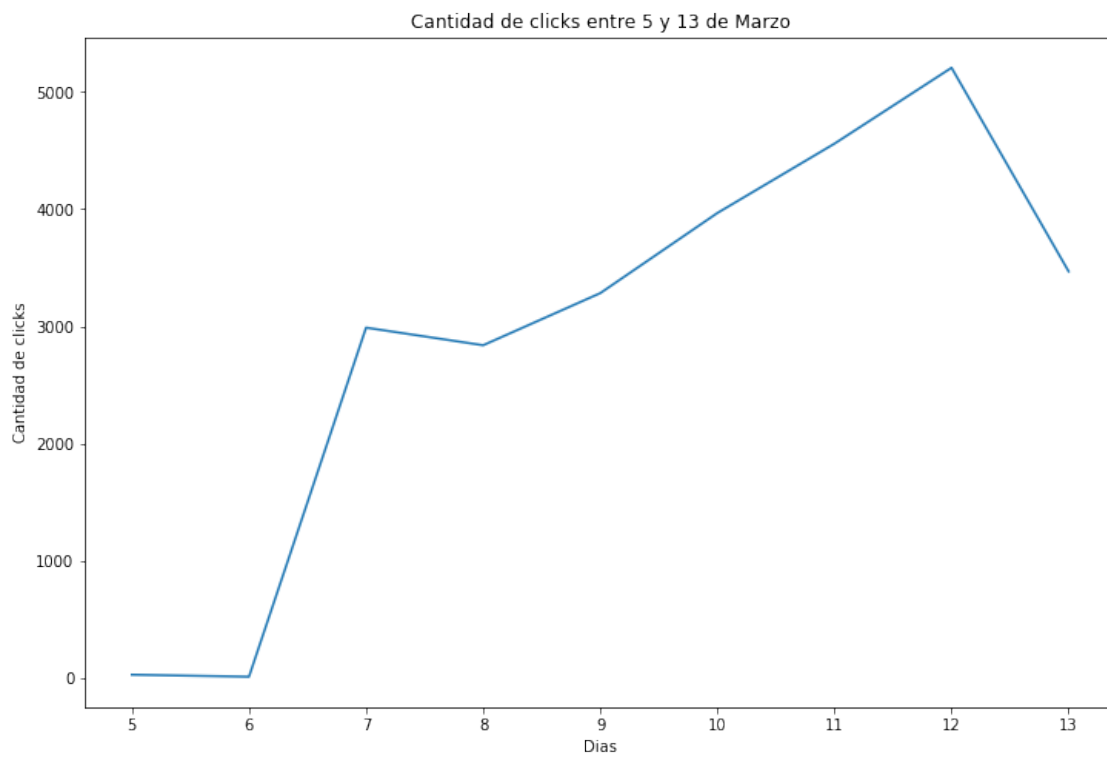


Figura 18: Clicks entre 5 y 13 de Marzo

Se observa que la cantidad de clicks empieza a aumentar a partir del día 8 y sube hasta alcanzar su máximo el día 12 y luego baja.

Ahora vamos a mostrar un gráfico que indica la cantidad de clicks distribuidos en las distintas horas del día.

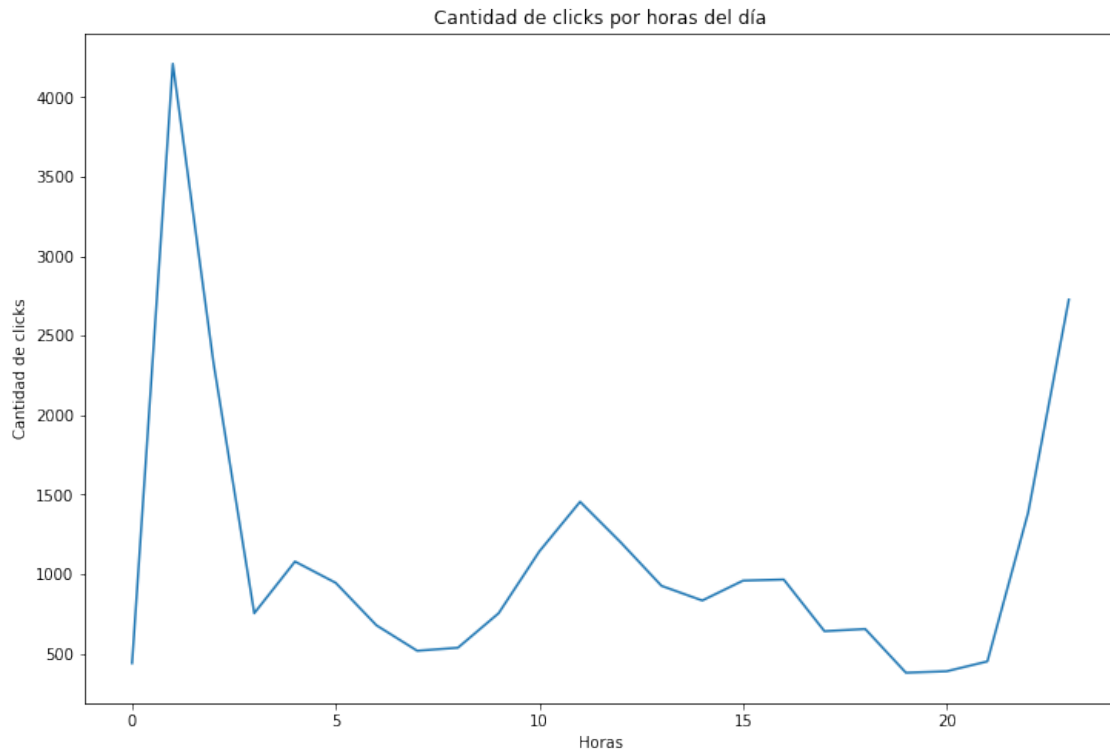


Figura 19: Clicks en las horas del día

En este gráfico se ve claramente que la cantidad de clicks suben exageradamente cuando llega la noche, más precisamente entre las 8pm y las 2am. Se podría decir que la gente utiliza más el celular a estas horas.

En este gráfico de áreas se muestran la cantidad de clicks por hora en el día de los distintos advertiser_id



Figura 20: Clicks en las horas del día de los advertiser_id (Porcentaje)

Se puede ver, como se esperaba, que prácticamente toda el área está pintada del advertiser id 3 que es el que más abunda en este archivo.

El siguiente gráfico muestra la cantidad de clicks en las distintas horas del día de los source_id que se encuentran en este archivo.

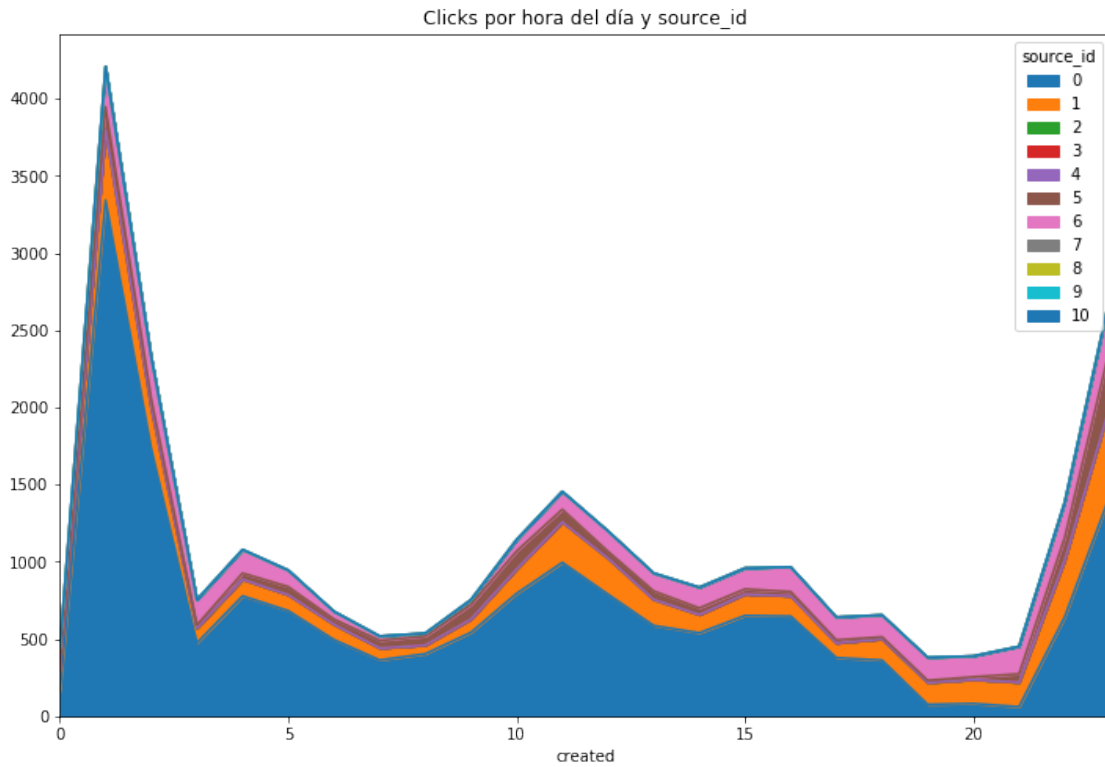


Figura 21: Clicks en las horas del día de los source_id

No se ve diferencia en cómo evoluciona la cantidad de clicks de los distintos source_id dependiendo la hora del día. No obstante, se puede observar que la mayoría de los clicks son de los source_id 0, 1, 5, 6.

Este nuevo gráfico nos mostrará los clicks en las horas del día distribuidos en los distintos specs_brand existentes.

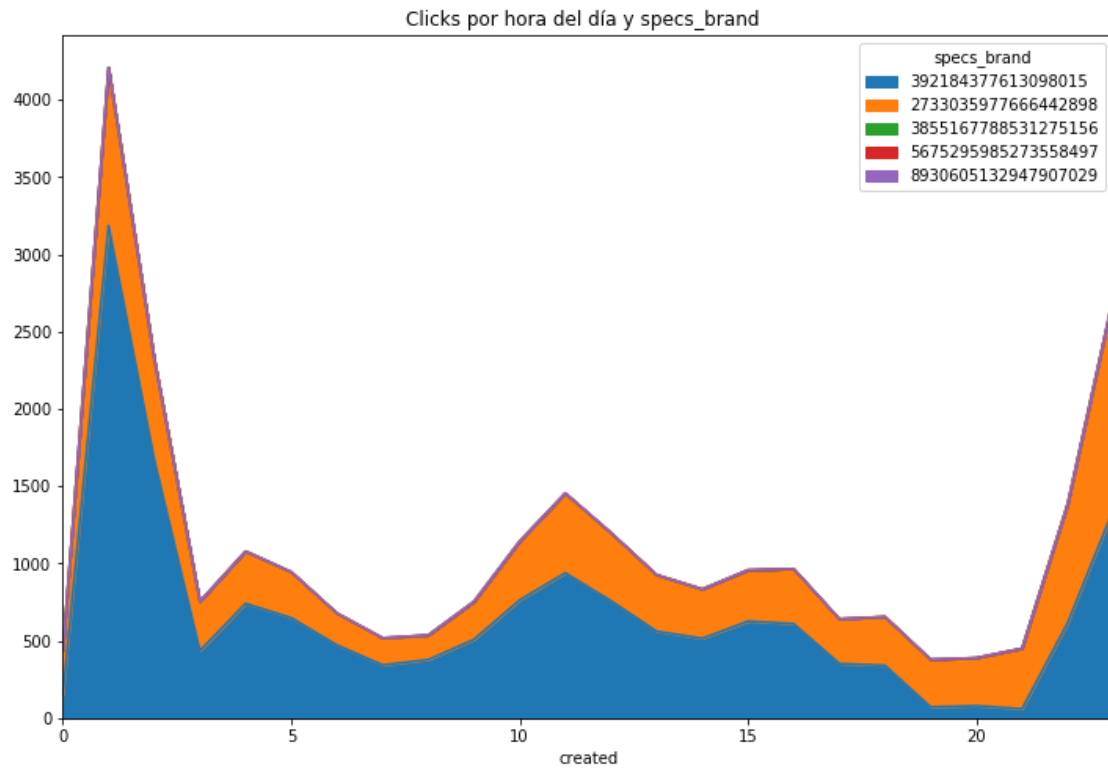


Figura 22: Clicks en las horas del día de los specs_brand

La cantidad de clicks se concentra en los specs_brand 392184377613098015 y 2733035977666442898 que son los que predominan en el csv.

2.2.13. Posición geográfica del click

En estos gráficos mostraremos las latitudes y longitudes de los clicks efectuados.

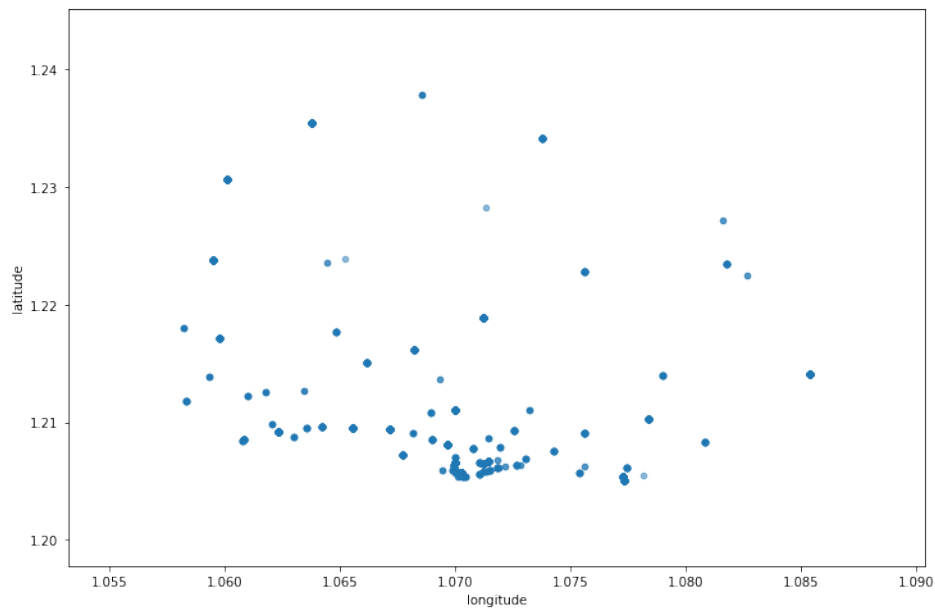


Figura 23: Latitud y longitud del click

Se ve que se concentran muchos clicks entre las latitudes 1.21 y 1.20 que intersecan con la longitud 1.07 y alrededores. Posiblemente sea una ciudad más poblada.

En este gráfico se vuelve a mostrar la latitud y la longitud de los clicks pero en este caso identificaremos los clicks con la mayor version de sistema operativo que tiene el dispositivo

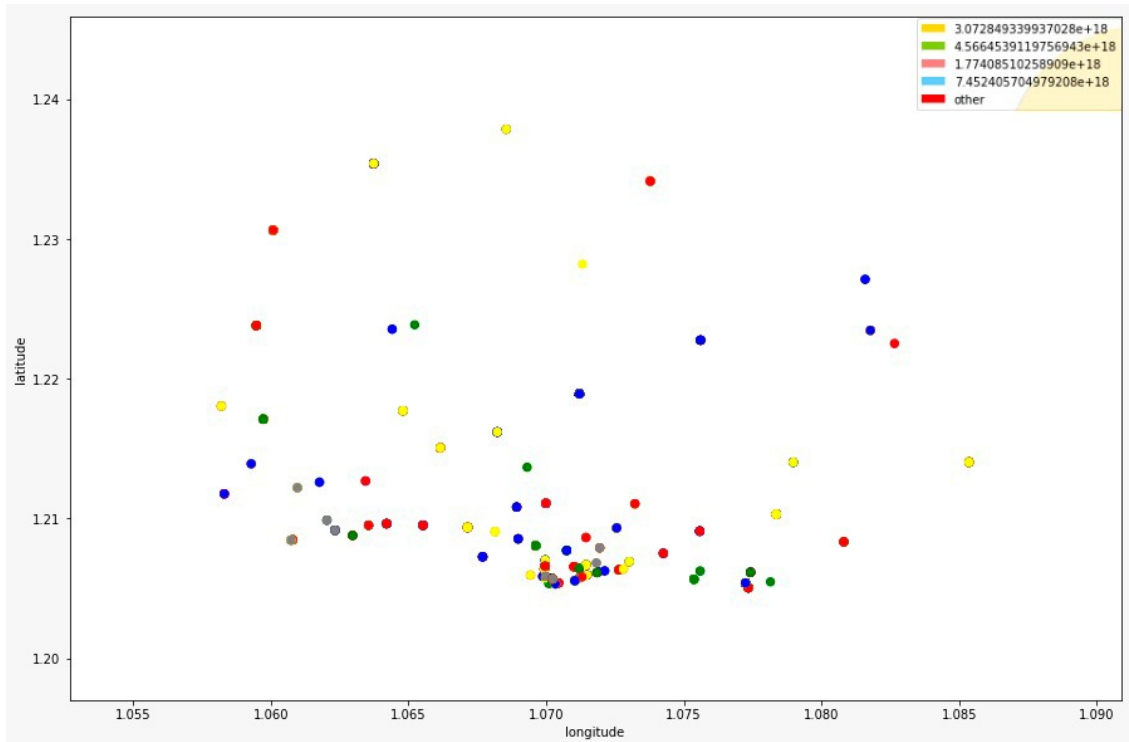


Figura 24: Latitud y longitud del click de los dispositivos con distintos os_major

No se observa que haya una concentracion de algún tipo de dispositivo con la mayor versión de sistema operativo en ningun lugar. De hecho, estan bastantes repartidos.

2.2.14. Clicks en la pantalla del dispositivo

Los siguientes gráficos muestran en que posición del celular se realiza el click

Se puede ver claramente que la mayor cantidad de clicks se concentra en la parte inferior del teléfono y mayormente, en la parte inferior derecha.

2.2.15. Tiempo en realizar el click

Se nos ocurrió analizar el promedio de tiempo que tarda una persona en realizar click. Primero decidimos filtrar los clicks en el que la persona tardaba más de 10 minutos en hacer click. Esto lo hicimos debido a que había clicks que tardaban más de un año en realizarse y supusimos que es un error en los datos. Llegamos a que una persona tarda 55.5921630859375 segundos en promedio en tocar el anuncio.

En los siguientes gráficos se mostrará la distribución en el tiempo que tardan los distintos advertiser.id en hacer click.

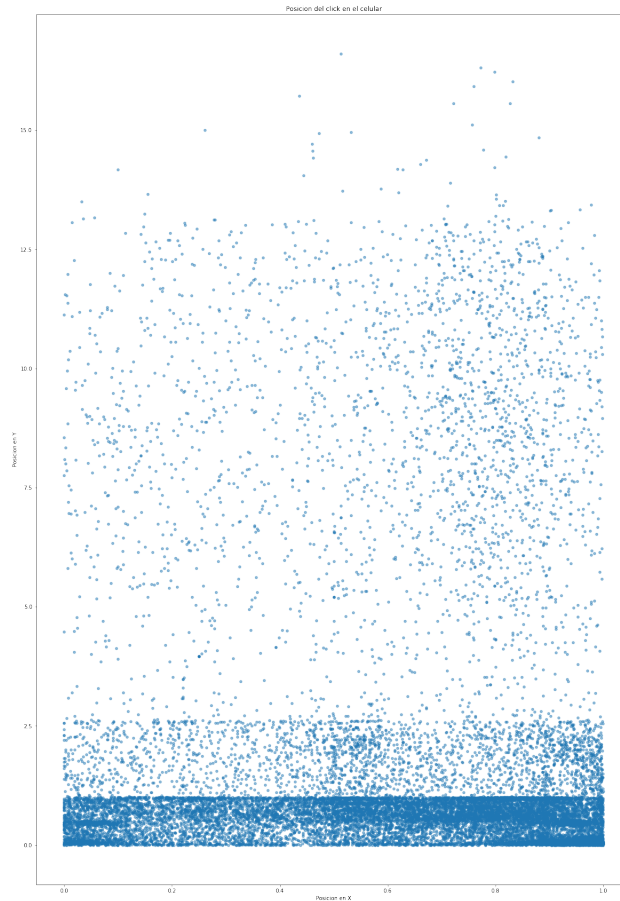


Figura 25: Posicion del click en pantalla



Figura 26: Clicks en el tiempo de los advertiser_id

Se observa que la mayoría de clicks se realizan en los primeros segundos y prácticamente todos por el advertiser_id 3. Esto es algo totalmente esperable debido a lo que se analizó previamente.

En el siguiente gráfico se mostrará la distribución en el tiempo que tardan los distintos spec_brands en hacer click.

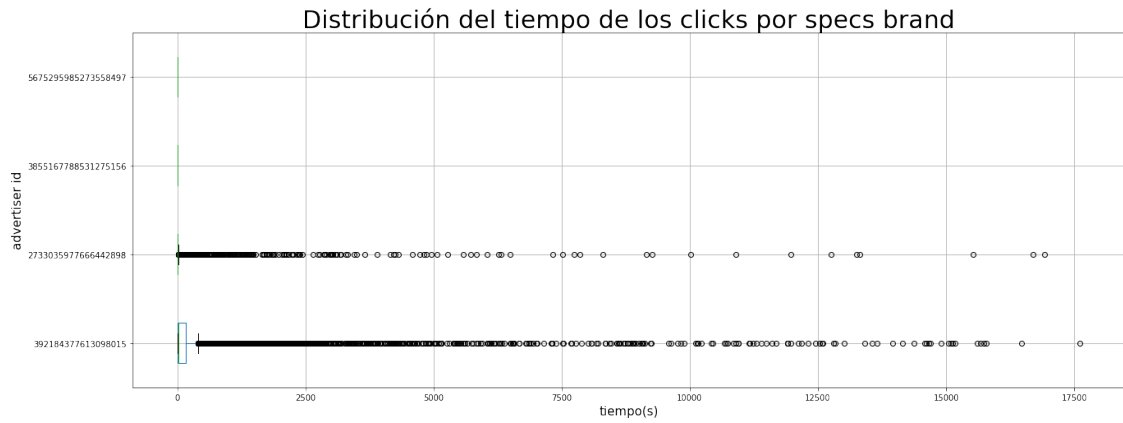


Figura 27: Clicks en el tiempo de los spec_brands

El siguiente gráfico nos muestra la distribución en el tiempo que tardan los distintos sistemas operativos en hacer click.



Figura 28: Clicks en el tiempo de los sistemas operativos

Se observa lo normal, en ambos sistemas operativos los clicks se concentran en los primeros segundos y luego se van diluyendo con el tiempo. Igualmente se puede notar que el sistema operativo que más clicks tiene, tiene clicks que tardaron bastante en hacerse.

2.3. Eventos

2.3.1. Introducción

Los datos sobre eventos se encontraron en el archivo `events.csv.gz`, que contenía información sobre los eventos registrados entre el 5/3/19 y el 13/3/19. Esta información incluía el tipo y la fecha del evento, la aplicación que lo generó, datos sobre el dispositivo y sobre el tipo de conexión e información sobre el user agent. A su vez, se tomó la decisión de descartar columnas como el id del dispositivo, el id del evento, la información del país y el id de la transacción, ya que no aportaban información relevante a este análisis.

2.3.2. Eventos por fecha y hora

Lo primero que analizaremos será cómo se distribuyeron los eventos cronológicamente, para eso los siguientes gráficos nos muestran esa información de formas distintas.

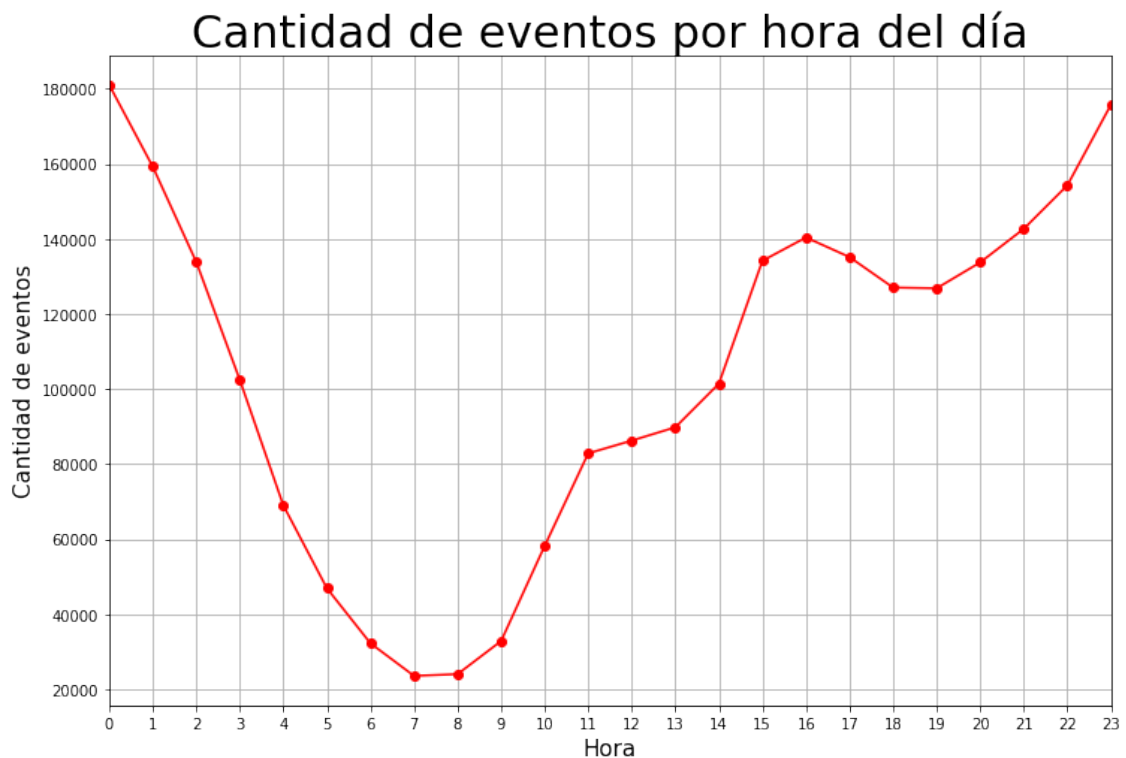


Figura 29: Cantidad de eventos por hora del día

Como se aprecia, se registra una mayor cantidad de eventos en horas de la noche/madrugada y un mínimo absoluto en horas de la mañana. Sin embargo, es de destacar el pico y posterior descenso en horas de la tarde, ya que los valores de las 16 superan incluso a los de horas normalmente no laborables, como las 19 o 20. Añadamos ahora a este análisis la fecha de los eventos.

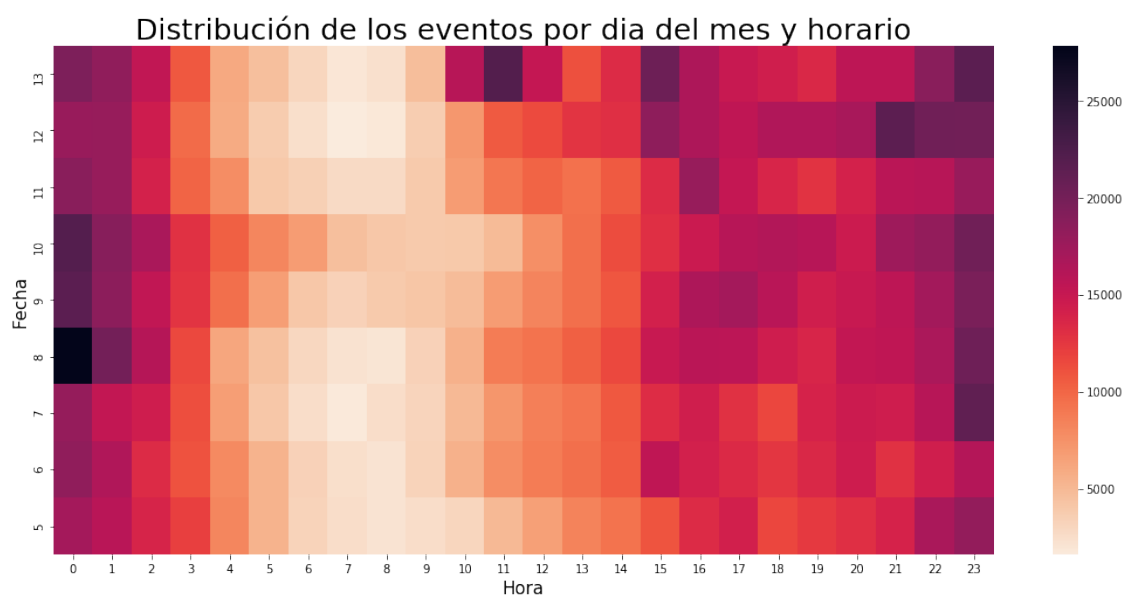


Figura 30: Cantidad de eventos por fecha y hora del día

En la figura se puede apreciar claramente el valle de las horas de la mañana, pero además se ve que el día con más eventos fue el 13, con un llamativo pico a las 11 a.m., mientras que el día 5 fue el de menos ocurrencias. El máximo número de eventos, sin embargo, se registró a comienzos del día 8 y el gráfico muestra un notorio corrimiento de la franja para el 9 y 10 de Marzo, lo que tiene sentido ya que son los días del fin de semana.

Si quisiéramos ver la distribución por días de la semana, primero debemos tener en cuenta que contamos con días repetidos, por lo que será necesario utilizar el promedio.

Aclaración: Para calcular el promedio, identificamos aquellos días que se repiten, como sabemos que se repiten solo dos veces cada día repetido, y que estos días también coinciden con ser los únicos que tienen más de 40.000 filas, calculamos el promedio dividiendo por 2 aquellos días que cumplen esta condición. Sin este análisis previo, calcular el promedio hubiese requerido un poco más de trabajo.

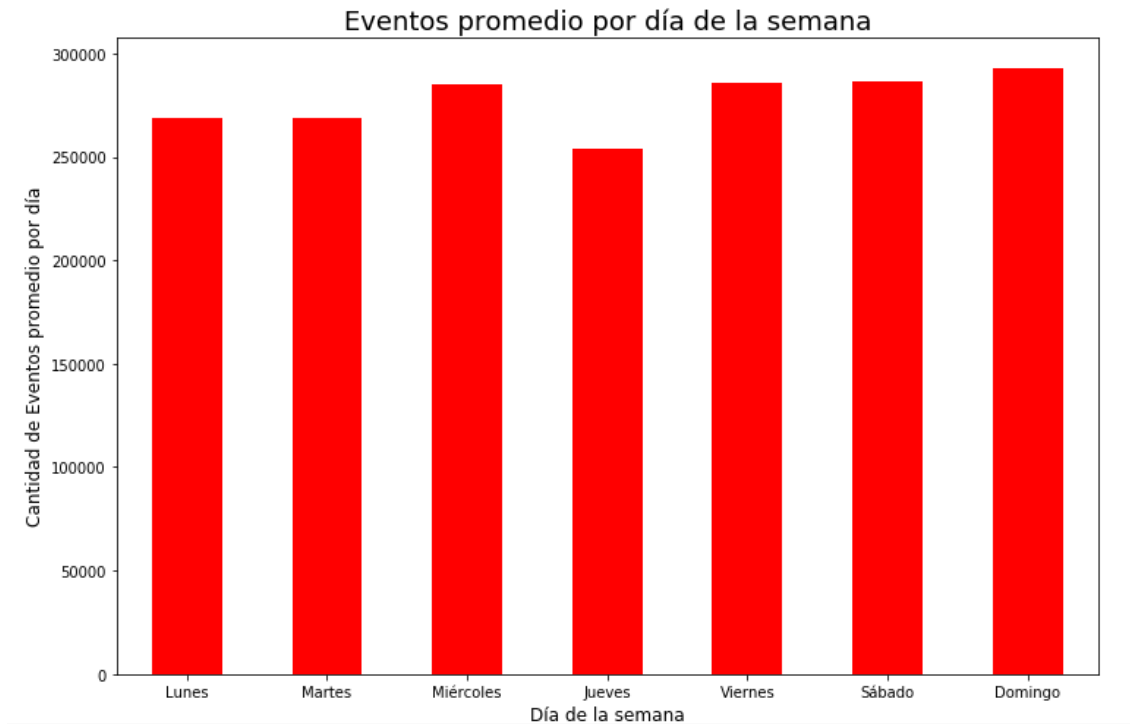


Figura 31: Promedio de eventos por día de la semana

Observamos que es bastante pareja, con un leve aumento los domingos, el día que menos gente trabaja.

2.3.3. Eventos por aplicación y detalles de los dispositivos

Al evaluar la cantidad de eventos por aplicación, se debe tener en cuenta primero que se tienen eventos registrados en 269 aplicaciones distintas, por lo que se hace imposible graficarlas todas. Para ello, se decidió tomar las diez que mayor cantidad de eventos generaron.

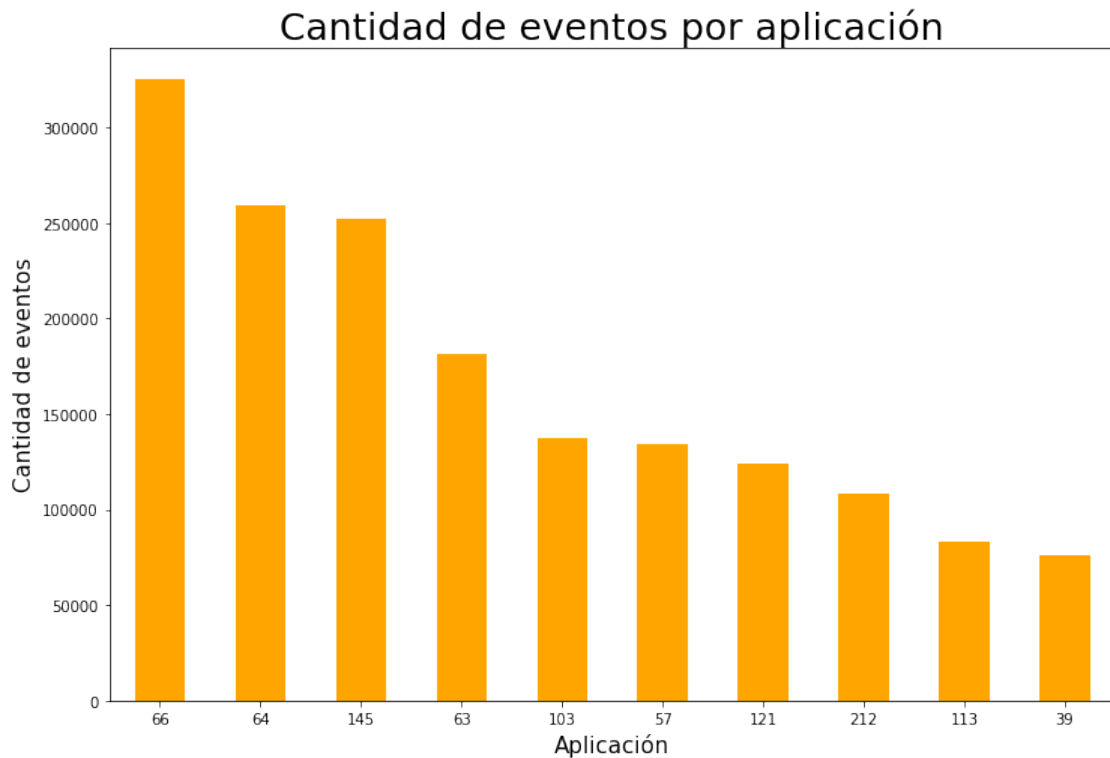


Figura 32: Aplicaciones con mayor cantidad de eventos

Teniendo éstas, podemos observar, por ejemplo, cuáles son los idiomas predominantes en cada una.

Idiomas predominantes en las top 10 apps que generan eventos

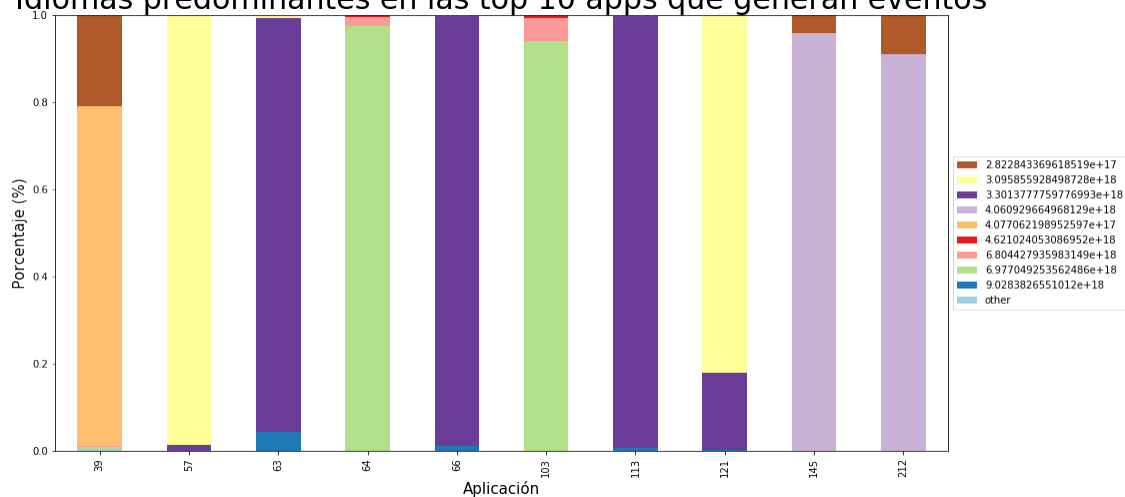


Figura 33: Distribución de idiomas para las diez aplicaciones con más eventos

Como se observa, cada app tiene un idioma que es muy predominante, lo que sugiere que cada una apunta a un público diferente, y cada idioma se distribuye de forma bastante pareja, siendo $3.3013777759776993e+18$ el más utilizado, ya que domina en tres de las diez líderes.

Analicemos ahora el ref type para cada una de las diez.

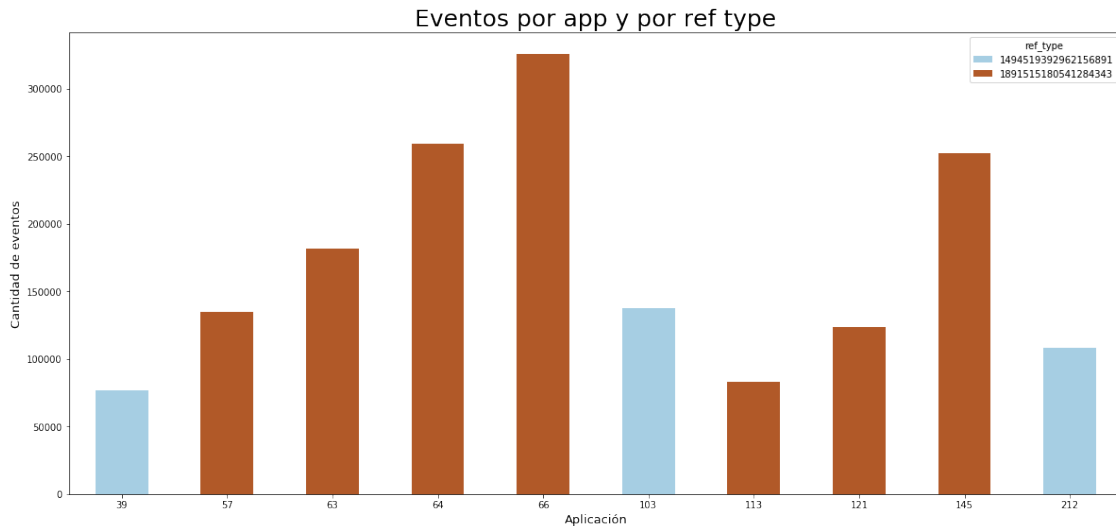


Figura 34: Cantidad de eventos por aplicación del top 10 y ref type

Como se ve, cada aplicación opera con un solo ref type y el predominante es 1891515180541284343.

También se puede analizar qué *session user agent* utiliza cada aplicación.

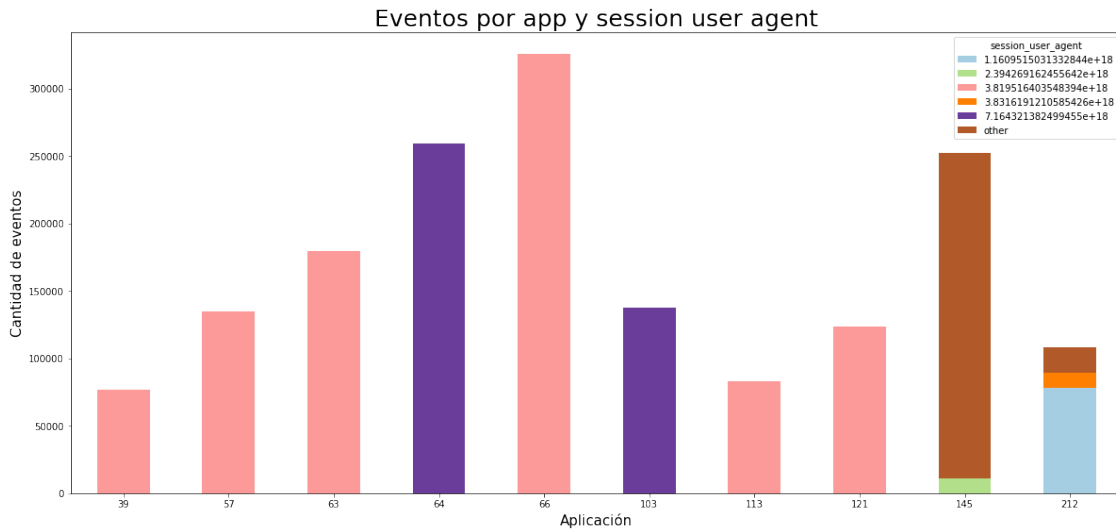


Figura 35: Cantidad de eventos por aplicación del top 10 y session user agent

A excepción de dos aplicaciones, todas operan con un único session user agent y el que predomina es 3.819516403548394e+18, presente en seis. Cabe aclarar que la categoría *other* está compuesta por 1455 session user agents que aparecían en cantidades muy pequeñas.

2.3.4. Eventos más frecuentes por tipo

Otro aspecto interesante a analizar será el tipo de los eventos registrados, el cual nos permite saber cuál era el objetivo del usuario en ese momento. Como se registraron en total 589 tipos distintos de eventos, y como la cantidad de eventos

de cada tipo sigue una distribución bastante acorde a la ley de Zipf, se tomarán los primeros diez para este análisis.



Figura 36: Tipos de eventos con más cantidad de apariciones

Como se puede apreciar, el segundo tipo más común aparece casi exactamente la mitad de las veces que el primero, el tercero casi exactamente un tercio y así sucesivamente, lo que se condice con la ley de Zipf.

Analicemos ahora de dónde provinieron dichos eventos.

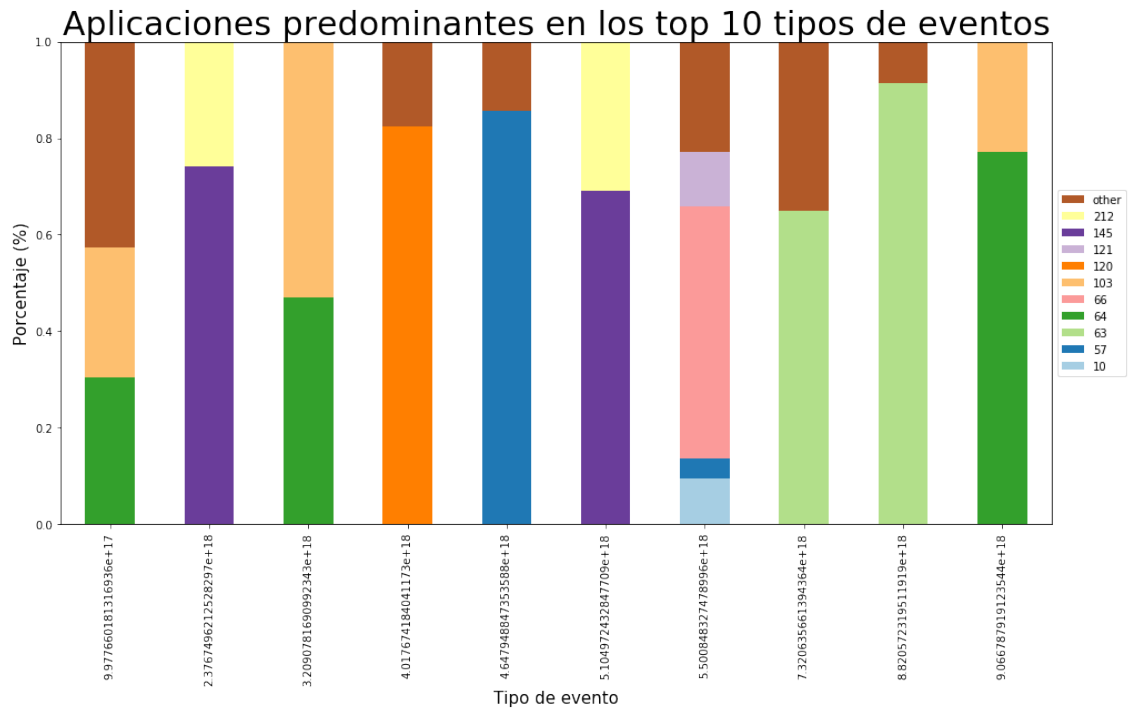


Figura 37: Distribución de las aplicaciones para los diez tipos de evento más comunes

Podemos también contrastarlos con las aplicaciones que más eventos generan.

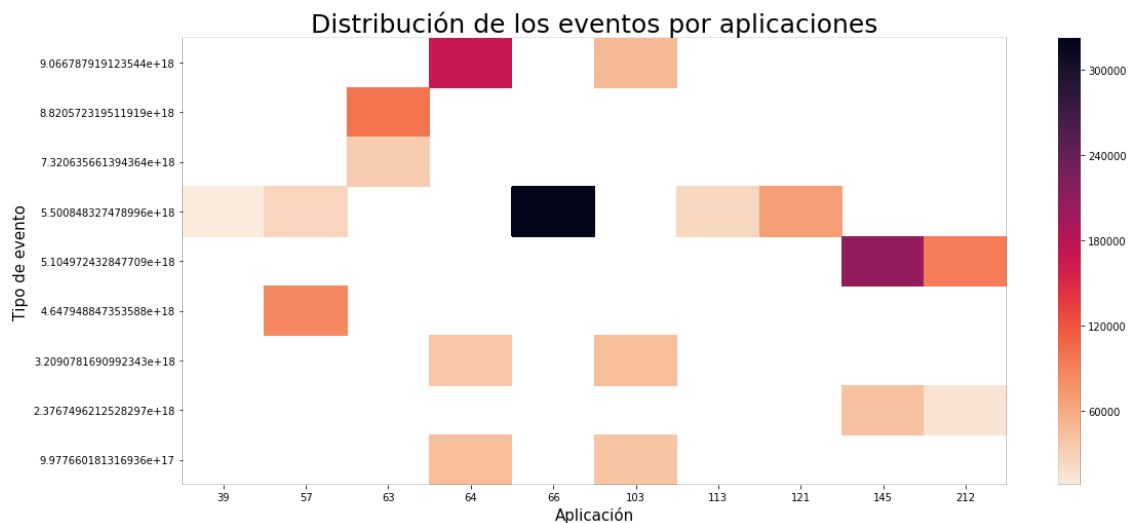


Figura 38: Distribución de las diez aplicaciones con más eventos para los diez tipos de evento más comunes

Como se puede ver, la distribución es bastante heterogénea. Los eventos que genera la app líder son todos del mismo tipo, que a su vez es el tipo de evento más común, por lo que si se busca que los usuarios hagan lo indicado por el evento 5.500848e+18, la mejor opción es mostrarles publicidad a través de la app 66. Esto también aplica para las demás, ya que la mayoría solo generan dos o tres tipos de evento distintos y puede utilizarse la misma lógica.

2.3.5. Eventos por tipo de conexión

El archivo nos provee información acerca de la conexión con la que contaban los usuarios a la hora de generar cada evento, ya sea wifi o datos móviles, y en caso de los datos, el tipo de conexión.

En el caso del wifi, vemos que la distribución es la siguiente.

Porcentaje de eventos vía Wifi

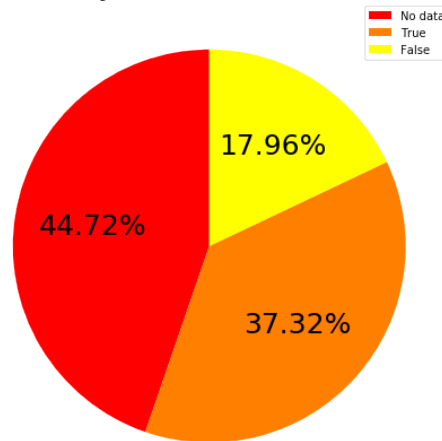


Figura 39: Distribución de eventos generados vía wifi

Como se puede observar, no contamos con datos en la mayoría de los eventos, aunque en los que sí se brindan la tendencia es claramente favorable al wifi, como sería esperable.

Sin embargo, para la gran mayoría de los eventos de los que no se tienen datos de conexión wifi sí se tiene información relacionada a conexiones por datos móviles, como la empresa que los provee o el tipo de conexión.

Tipos de conexión por datos

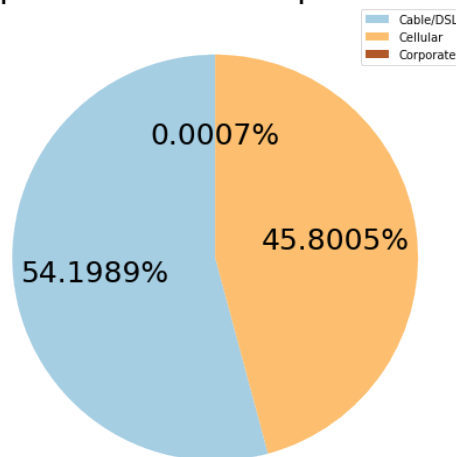


Figura 40: Tipo de conexión para los eventos generados con datos móviles

La distribución es pareja con una leve ventaja para las conexiones por cable, mientras que las corporativas son prácticamente despreciables.

Analicemos ahora las empresas que proveen dichas conexiones.

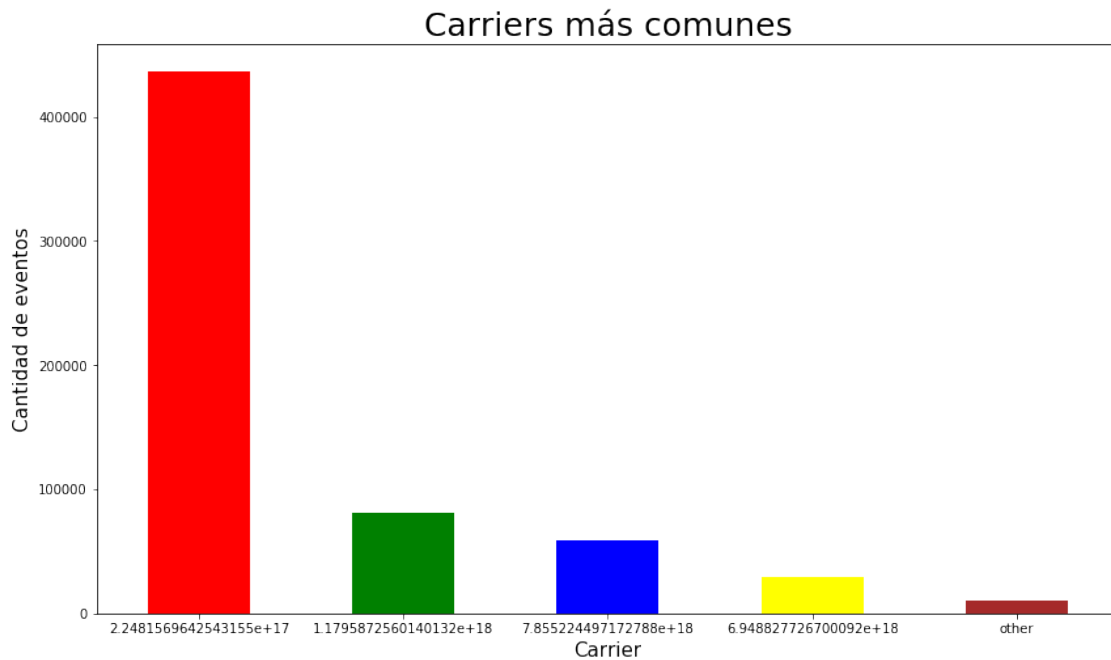


Figura 41: Proveedores de servicio más comunes para conexiones por datos móviles

Se ve que hay un carrier dominante con mucha diferencia con el resto y cabe aclarar que la categoría *other* engloba a 79 carriers distintos con valores muy bajos.

Si combinamos ambos, podemos observar qué tipo de conexión utiliza cada una de las empresas.

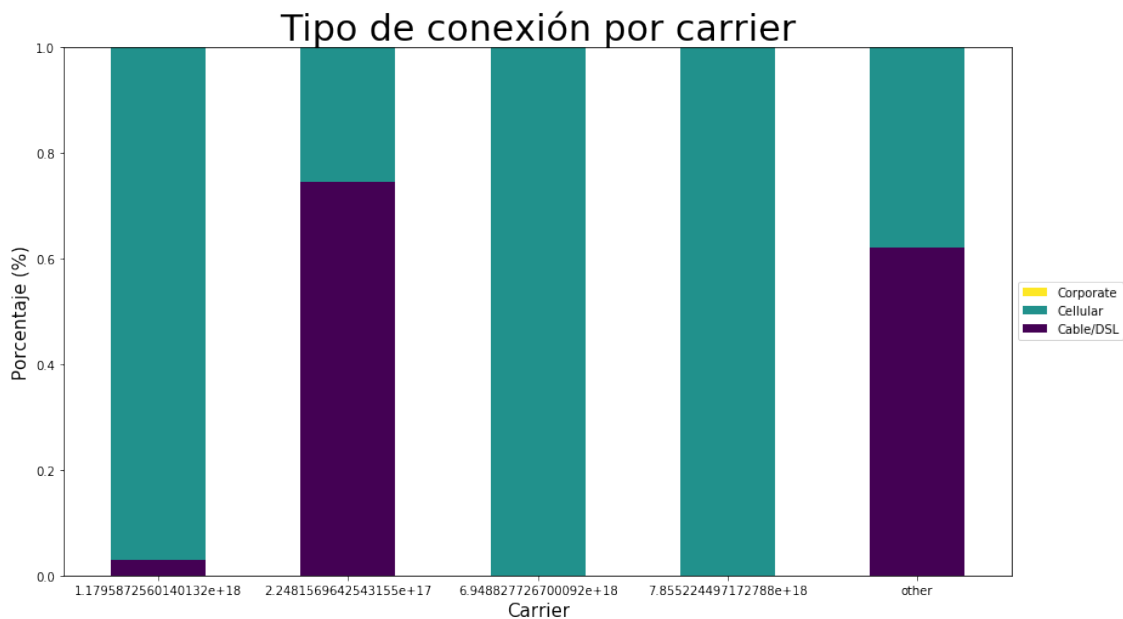


Figura 42: Tipo de conexión por empresa proveedora

La mayoría usa Cellular mientras que el carrier líder utiliza mayormente Cable/DSL.

Eventos más frecuentes para cada tipo de conexión

El tipo de conexión con el que se cuenta es un aspecto importante a tener en cuenta a la hora de buscar que los usuarios adopten un comportamiento determinado, por ello analizaremos qué eventos tienden a generarse con los siguientes dos tipos de conexiones.

■ Wifi



Figura 43: Tipos de evento más frecuentes en conexiones wifi

■ Datos móviles

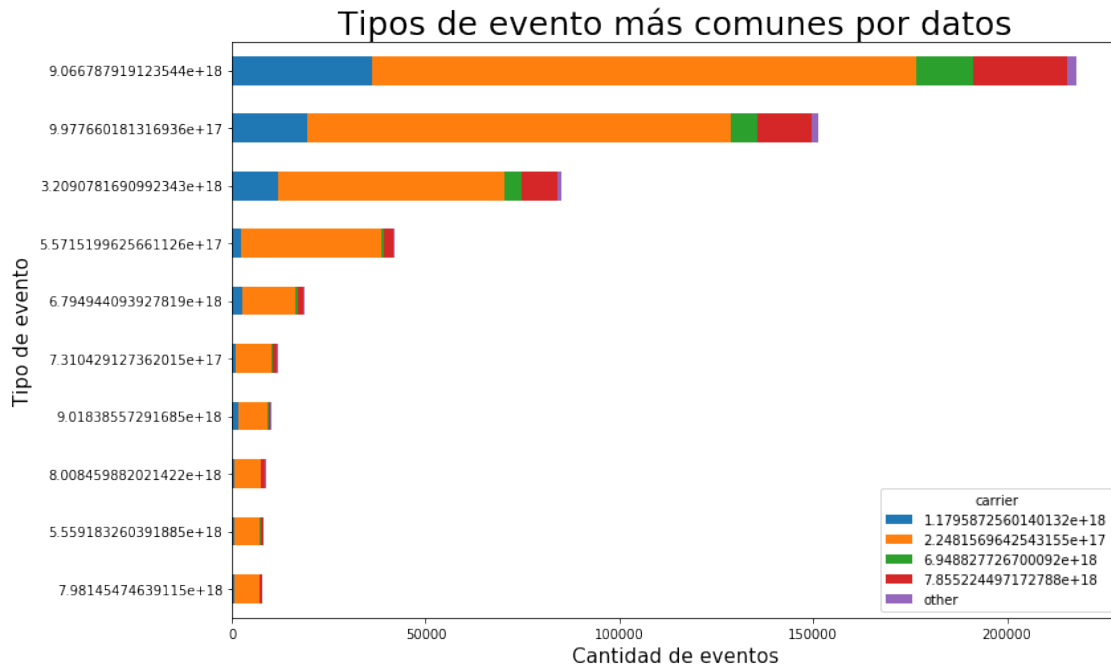


Figura 44: Tipos de evento más frecuentes para conexiones móviles por carrier

Se puede apreciar que cada tipo de conexión genera eventos distintos, por lo que si se quisiera lograr que los usuarios generen el evento 5.500848327478996e+18, será importante verificar que cuenten con conexión wifi, mientras que si el que se buscara generar fuera el 9.066787919123544e+18, la mejor opción será verificar que estén usando datos móviles y preferiblemente con el carrier 2.2481569642543155e+17.

2.3.6. Marcas y eventos

Otra cuestión interesante de analizar es el comportamiento de los usuarios de las diferentes marcas de dispositivos. Si bien no contamos con los datos de la marca para todos los eventos, contamos con suficientes datos para evaluar. En ellos encontramos 250 marcas distintas, aunque muchas de ellas de muy bajo impacto. Observemos primero cuáles son las marcas más populares, se muestran las primeras diez, ya que concentran el 94 % de los eventos.



Figura 45: Marcas que más eventos generan

Se aprecia una notable diferencia entre la marca $3.083058605577787e+17$ y sus competidoras.

Analizando los datos, vemos que tenemos un 53% de filas donde la marca del celular figura en NaN, pero vemos que hay otras filas donde el mismo modelo de celular figura y tenemos su marca. Así que usaremos eso para reparar el set de datos con esta información para posteriormente volver a graficar y ver qué cambió.

Lo primero que vemos es que recuperamos bastantes filas, ahora solo un 8,8% no tiene datos sobre la marca del celular. Descartamos esos valores y volvemos a graficar.

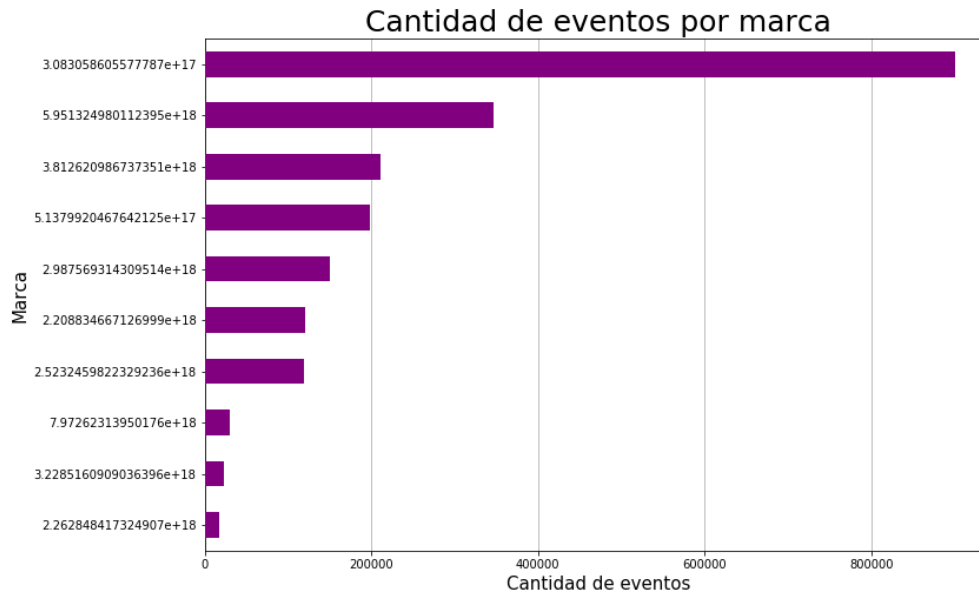


Figura 46: Marcas que más eventos generan con set de datos reparado

Vemos que la tendencia se inclina aún por la marca $3.083058605577787e+17$ pero el segundo lugar ahora lo ocupa $5.951324980112395e+18$ y $2.208834667126999e+18$, el 3ro antes de la reparación ahora cayó a un 6to puesto.

Trabajar un poco con la información que teníamos cambió bastante la visualización y nos previno de sacar conclusiones que podían ser erróneas.

2.3.7. Modelos y eventos

En la sección anterior analizamos las diferentes marcas y su asociación con los eventos que se producen en los ads. Veamos como se distribuye la relación en los modelos.

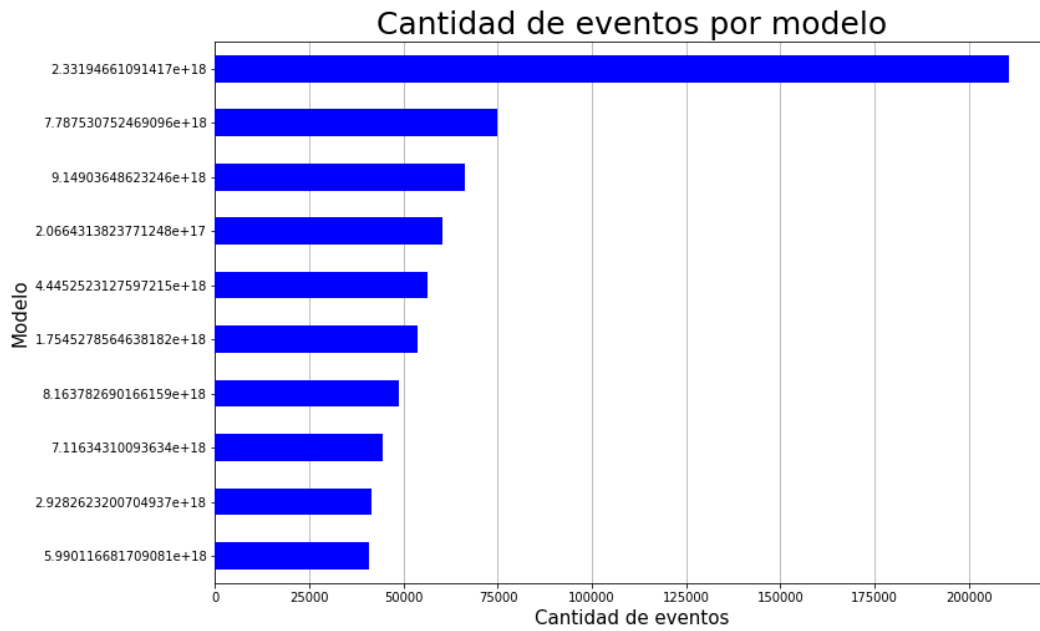


Figura 47: Distribución de eventos por modelos

Es interesante analizar los modelos que más eventos provocan. Quizás al momento de decidir ofertar en una subasta, conviene enfocarse en algún modelo de celular determinado en lugar de diversificarse. Lamentablemente, no conocemos las características de los eventos, necesitaríamos saber qué tipos de eventos son positivos para la empresa a fin de tomar una decisión que maximice la ocurrencia de ese tipo de eventos.

2.3.8. Ciudades y eventos

El lugar de origen de los usuarios de cada evento también será provechoso de analizar para poder distinguir los intereses de los diversos lugares. Si bien en este caso contamos con datos parciales. Lo primero que salta a la vista es que la ciudad $3.8000799488967747e+18$ tiene mucha preponderancia, concentrando el 91,41 % de los eventos. Utilicemos esto para contrastar los eventos más populares y las aplicaciones que más eventos generaron en los distintos lugares, haciendo la distinción entre la ciudad líder y el resto.

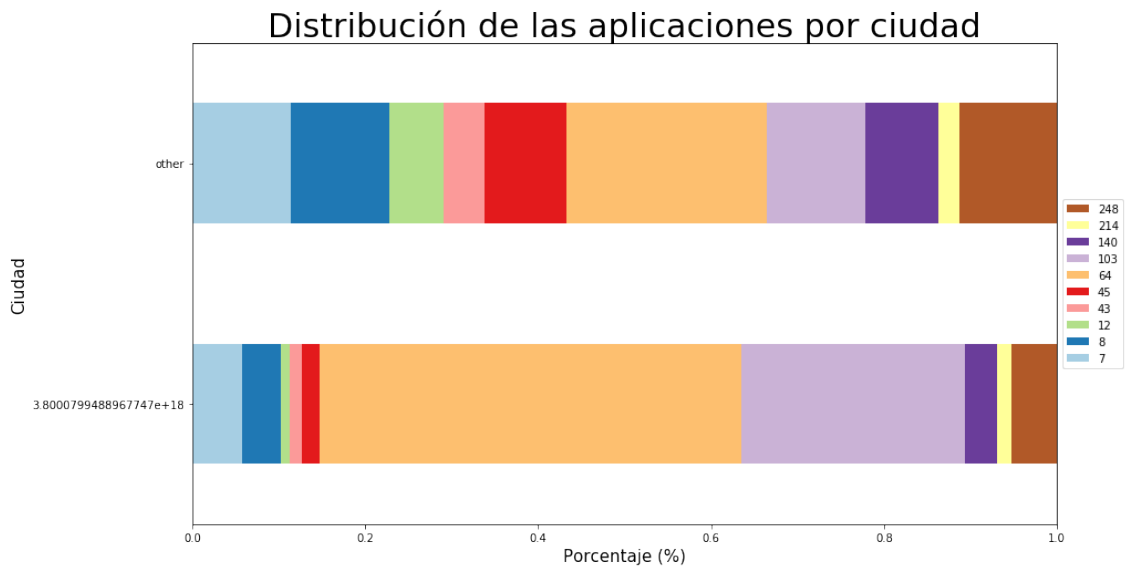


Figura 48: Distribución de aplicaciones por ciudad

Notamos que aplicaciones como la 64 y la 103 son mucho más preponderantes en la ciudad $3.8000799488967747e+18$, mientras que en el resto de lugares, si bien la 64 sigue siendo la más usada, el segundo lugar es mucho más parejo. Esto resulta útil ya que si se quisiera lograr que los usuarios del resto del país generen un evento, será provechoso mostrarles los ads mediante las aplicaciones 64, 8, 7, 248 o 103, cuando si se tratase de la ciudad anterior mencionada la mejor elección sería o la 64 o la 103.

2.3.9. Eventos atribuidos a Jampp

Será importante conocer la proporción de los eventos registrados que le fueron oficialmente atribuidos a Jampp, así como su composición.



Figura 49: Eventos atribuidos a Jampp

Como se aprecia, el porcentaje es muy bajo. Veamos ahora de dónde provienen dichos eventos.

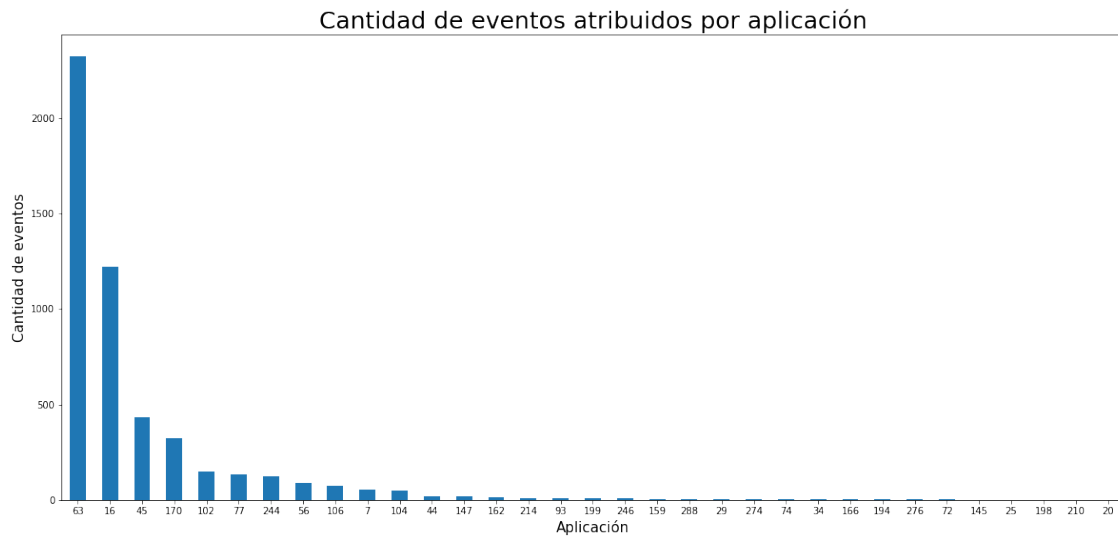


Figura 50: Eventos atribuidos a Jampp por aplicación

Las aplicaciones 63 y 16 son las que más eventos generaron para Jampp, con una amplia diferencia. Si analizamos el ref type de estos obtenemos.

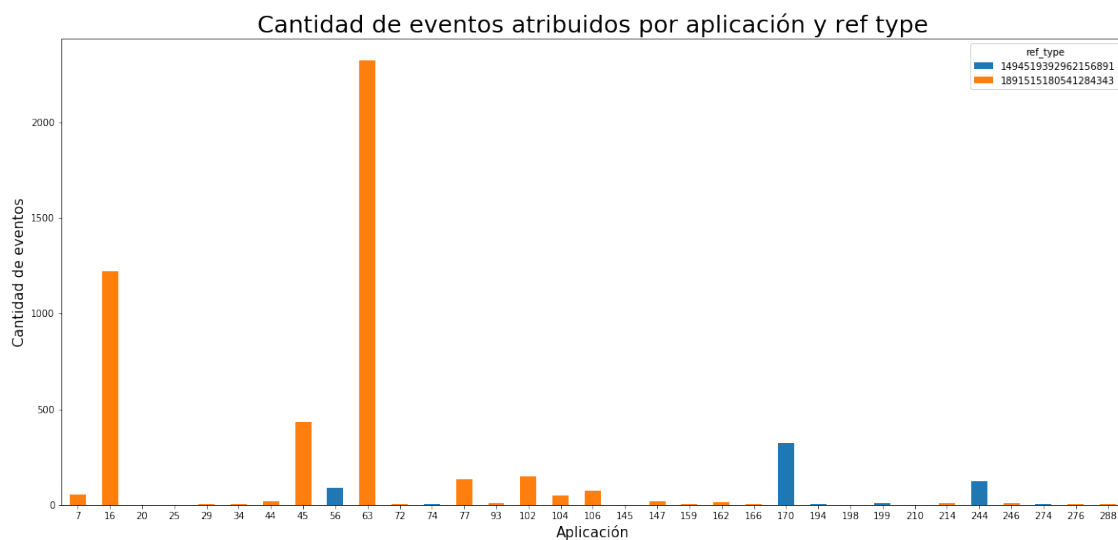


Figura 51: Eventos atribuidos a Jampp por aplicación y ref type

Donde se observa una predominancia marcada del tipo 1891515180541284343. De los eventos atribuidos, la distribución según conexión wifi es la siguiente.

Eventos atribuidos vía Wifi

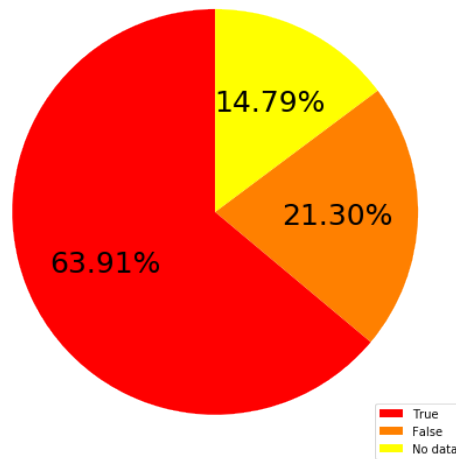


Figura 52: Porcentaje de eventos atribuidos a Jampp generados con conexión wifi

Como era de esperarse, la mayoría fueron generados vía wifi

2.4. Instalaciones

2.4.1. Introducción

Los datos sobre instalaciones fueron provistos por Jampp en el archivo `installs.csv.gz`, el cual contenía información acerca de todas las instalaciones registradas entre los días 5 y 13 Marzo del corriente año, indicando el tipo de aplicaciones descargadas, su fecha de descarga, país de origen, modelo, marca e idioma del dispositivo, entre otras cosas.

Cabe destacar que se descartaron datos como las direcciones ip y los varios id únicos generados para cada instalación, puesto que no aportaban información relevante al análisis que se pretende hacer en este trabajo, como también los datos del *session user agent*, ya que la misma empresa informó que no los consideran de importancia y pudieron haberse visto modificados por los propios agentes que les proveyeron los datos.

2.4.2. Instalaciones por día y hora

Para comenzar, lo primero que haremos será ver cómo se distribuyen las instalaciones en el periodo dado.

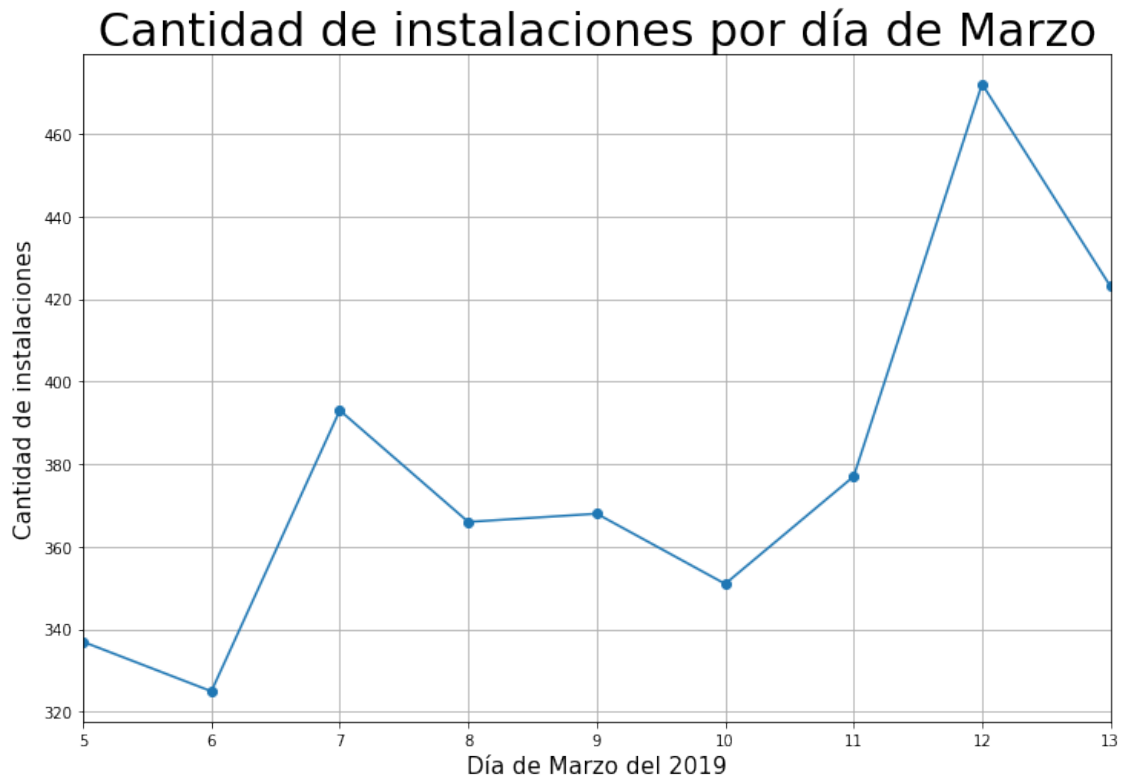


Figura 53: Instalaciones por día de Marzo

Como se puede observar, se registran ascensos considerables entre los días 6 y 7 y 11 y 12, con respectivas caídas al día siguiente, pero manteniendo una tendencia general al alza.

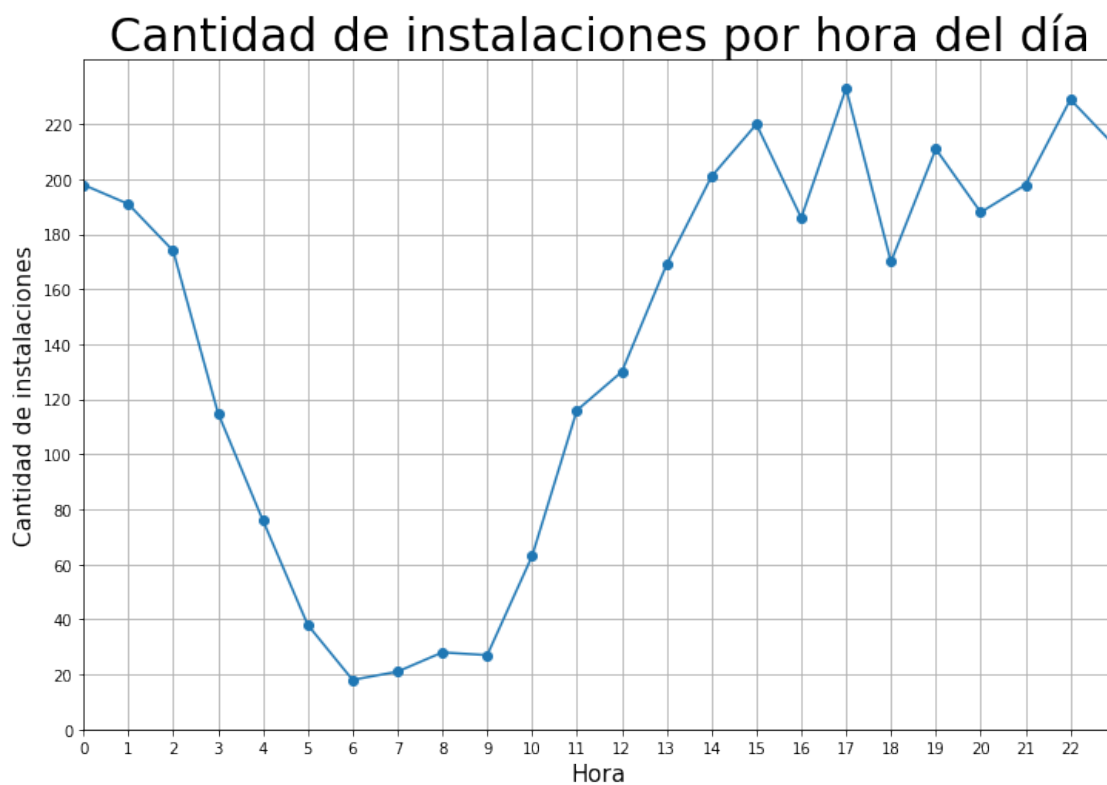


Figura 54: Instalaciones por hora del día

El gráfico anterior nos indica que la gran mayoría de las instalaciones se registran en horas de la tarde y la noche, con un pico a las 5 de la tarde. Cabe destacar a su vez el notorio valle que se da en horas de la mañana, donde el número es hasta diez veces menor que en el punto máximo.

Para un análisis más general, la siguiente figura engloba los dos puntos mencionados anteriormente.

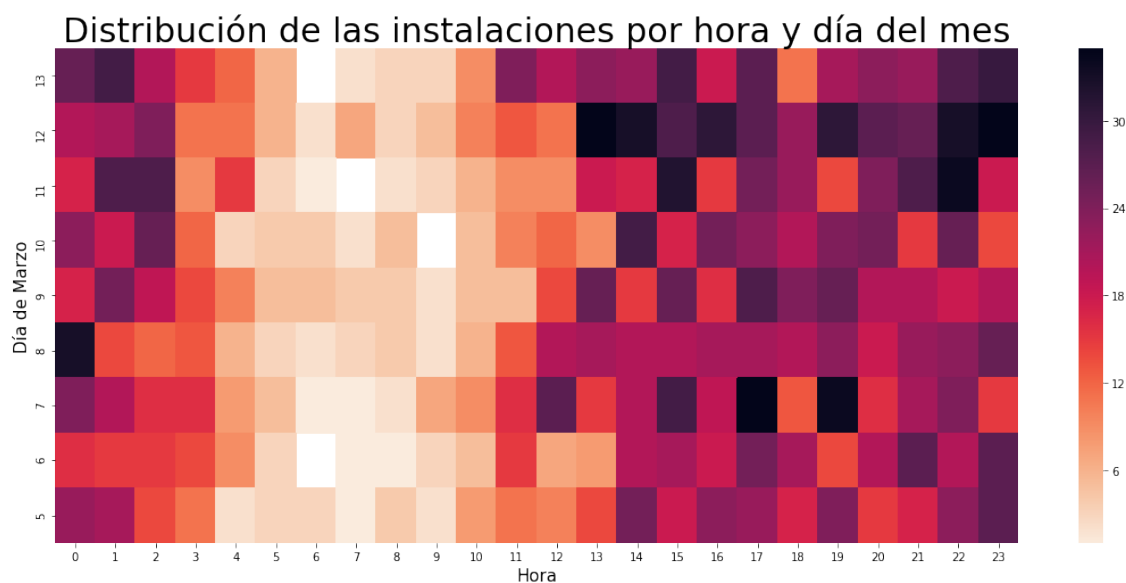


Figura 55: Instalaciones por fecha y hora

Se puede observar claramente el valle de las horas de la mañana, como así también el pico que se da en el día 12. Sin embargo, este gráfico resulta útil ya que permite notar que tanto en los días 6 y 13 a las 6 de la mañana, el día 11 a las 7 y el 10 a las 9 no se produjo ninguna instalación.

2.4.3. Instalaciones por aplicación

Resultará de utilidad conocer de que aplicación provienen las instalaciones registradas y observar cual es la tendencia en ese aspecto para determinar en cuales es mejor colocar la publicidad.

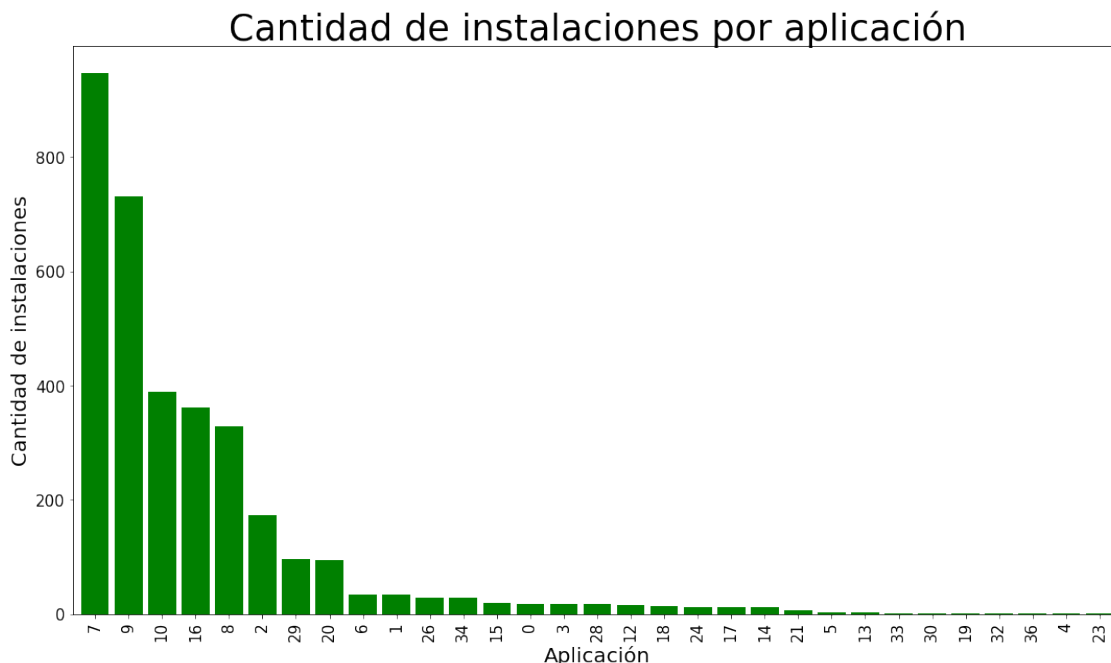


Figura 56: Instalaciones por aplicación

La figura anterior muestra un dominio claro de las aplicaciones 7 y 9 por sobre las demás, ya que la segunda casi duplica en cantidad a la tercera. Además, se puede ver otra diferencia importante—otra vez, de casi el doble—entre la quinta y la sexta, lo que deja en evidencia cuales son las que dominan en este campo, puesto que las primeras cinco aplicaciones concentran más del 80 % de las instalaciones.

2.4.4. Instalaciones por fecha según la aplicación

A continuación veremos a qué aplicaciones pertenecen las instalaciones según el día y la hora del día, lo que puede servir para determinar en qué momento apostar por una u otra aplicación. Cabe destacar que, para ello se agruparon todas aquellas aplicaciones que generaron un número muy bajo (menos de 90) de instalaciones en la categoría *other*.

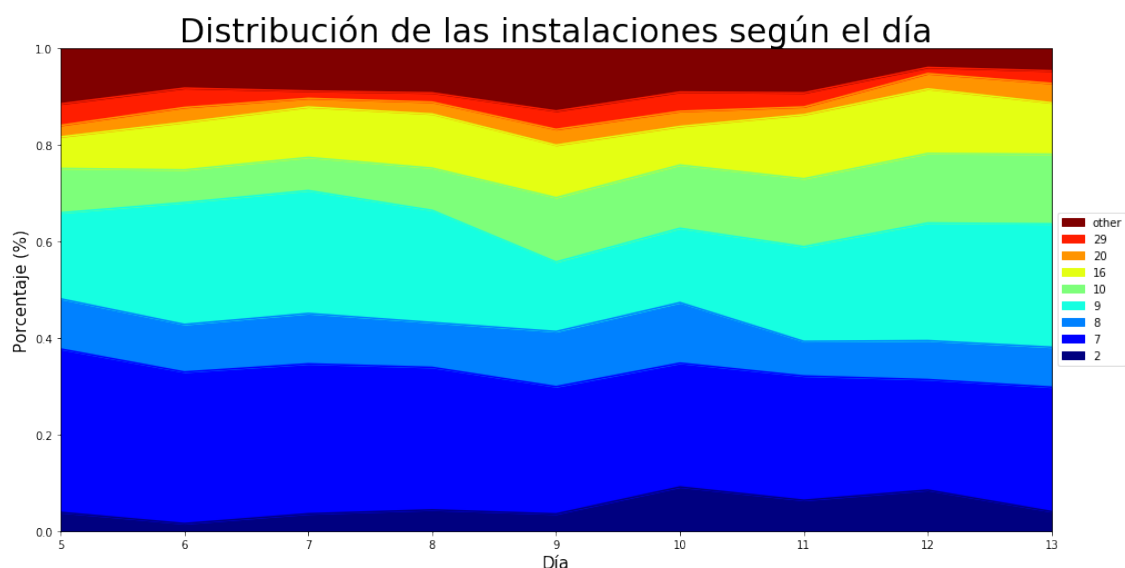


Figura 57: Incidencia de las aplicaciones según el día

Como se puede observar, la aplicación 7 dominó la mayoría de los días aunque disminuyendo hacia el 12 y el 13, donde fue superada por la 9. La tercera aplicación con más instalaciones, la 10, se mantuvo bastante estable durante los últimos cinco días, mientras que los días 10 y 12 la número 2, de poca preponderancia en otros días, obtuvo su mejor resultado. Cabe aclarar además, que si bien en días como el 9 se ve bastante incidencia de la categoría *other*, ésta engloba los resultados de 23 aplicaciones distintas.

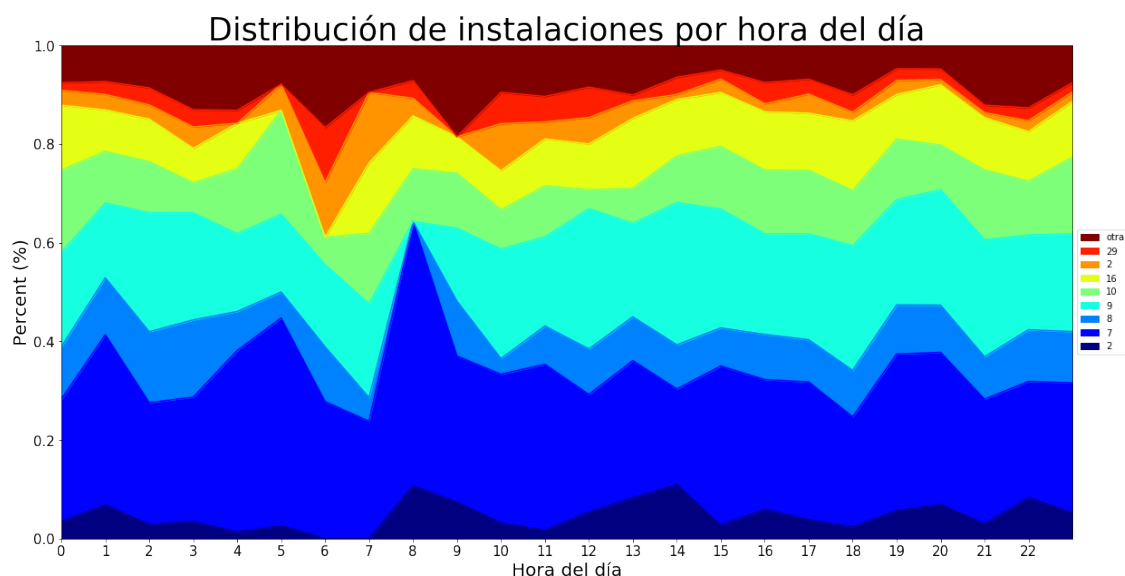


Figura 58: Incidencia de las aplicaciones según la hora del día

En cuanto a la hora cabe destacar que la segunda aplicación en instalaciones no tiene presencia alguna a las 8 a.m., que es a su vez el momento del día donde más incidencia tiene una aplicación de poca presencia como la 2. La tendencia sigue mostrando a la número 7 como dominadora absoluta en todos los horarios.

2.4.5. Instalaciones por país

El país de origen es un aspecto a considerar a la hora de decidir qué advertisers priorizar para cada situación. En este caso Jampp nos provee información de instalaciones en dos países distintos.

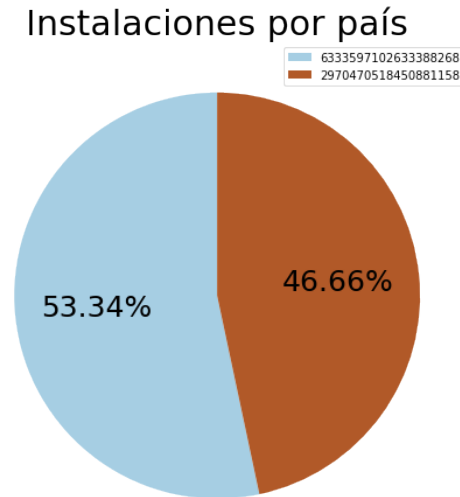


Figura 59: País de origen de las instalaciones

La proporción es bastante pareja, con leve mayoría para el país 6333597102633388268.

2.4.6. Instalaciones por tipo

Los dispositivos se clasifican en dos grandes grupos, aquellos que provienen de Apple y los que provienen de Google.

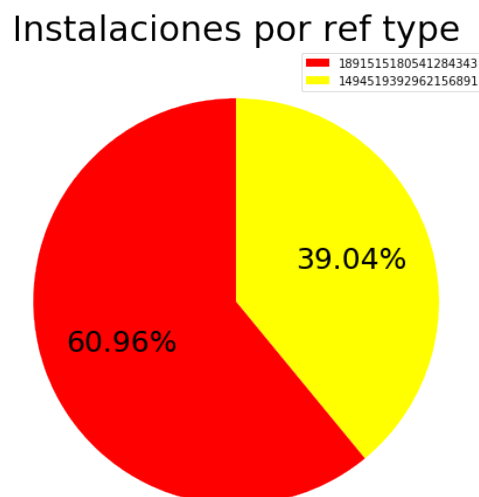


Figura 60: Tipo de referencia de las instalaciones

Como se observa, el claro dominador es 1891515180541284343.

2.4.7. Aplicaciones por país y tipo

Analicemos ahora de donde provienen las instalaciones de las 5 aplicaciones con más installs.

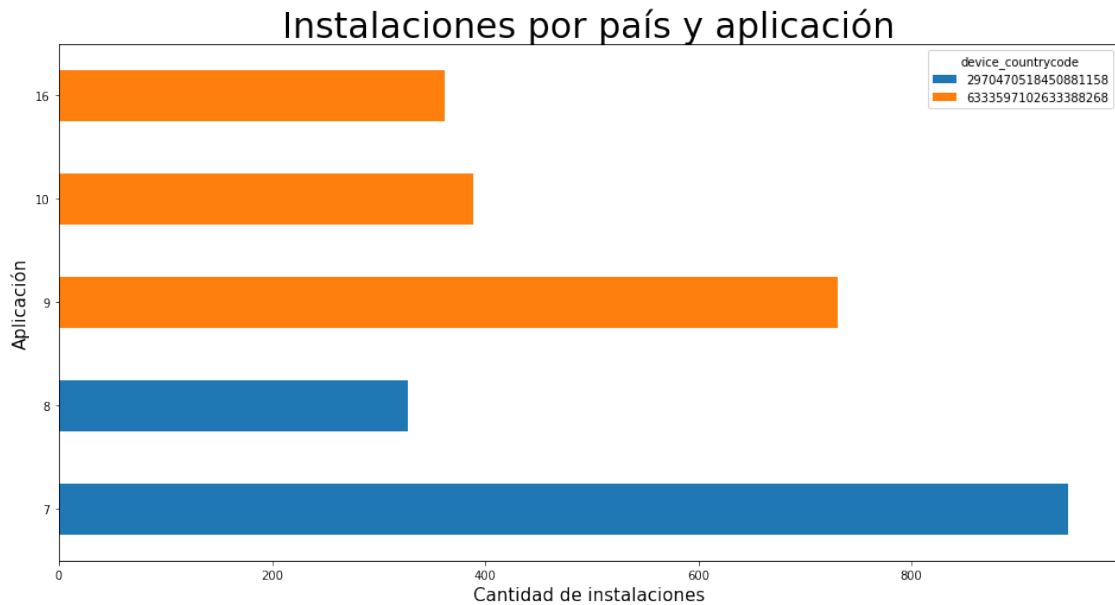


Figura 61: Distribución de países para las top 5 aplicaciones en instalaciones

Sorprendentemente el gráfico es determinante, las aplicaciones provienen o de un país o del otro, ninguna está presente en ambos.

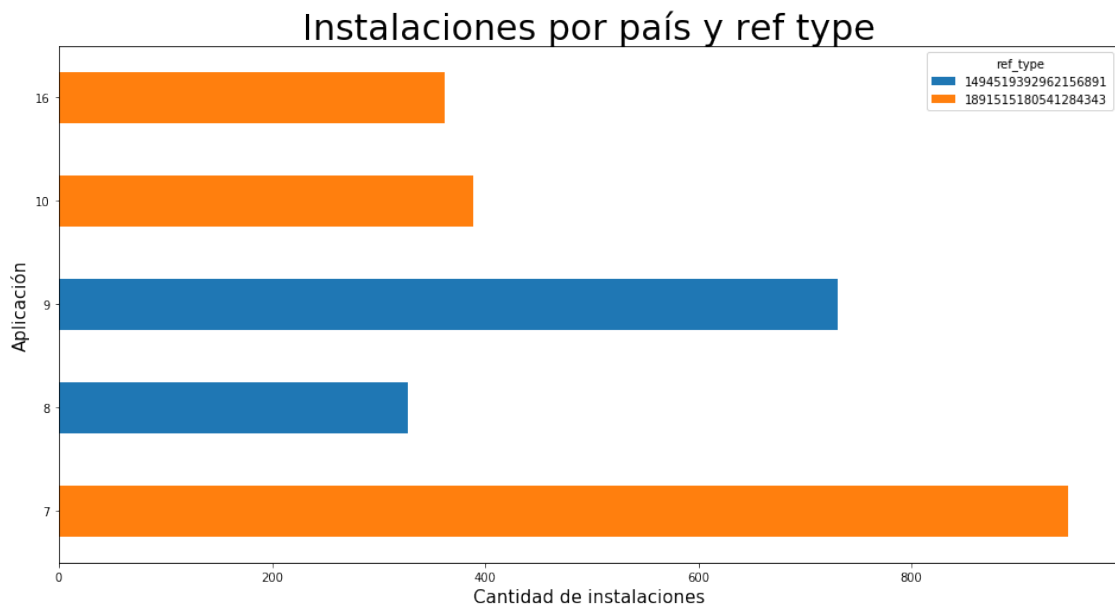


Figura 62: Distribución de ref types para las top 5 aplicaciones en instalaciones

En el caso de los tipos sucede lo mismo, aunque esto es más esperable, ya que existen muchas aplicaciones que son exclusivas de Apple o de Google. Esta información es útil a la hora de determinar el advertiser a seleccionar en cada subasta,

ya que la aplicación o servicio a mostrar en la publicidad debe estar disponible para ese tipo.

2.4.8. Idiomas y aplicaciones

Otro aspecto importante a considerar será el idioma del dispositivo al que se le mostrará la publicidad, ya que el usuario debe entender el mensaje.



Figura 63: Idioma de entrada con mayor número de instalaciones

Vemos que hay un idioma claramente predominante, luego dos en nivel parejo y una brecha considerable. Por esta razón, y para mayor simplicidad de los gráficos, se agruparán los 26 idiomas menos predominantes en la categoría *other*.

Analicemos ahora como se distribuyen los idiomas en las 5 aplicaciones más instaladas.

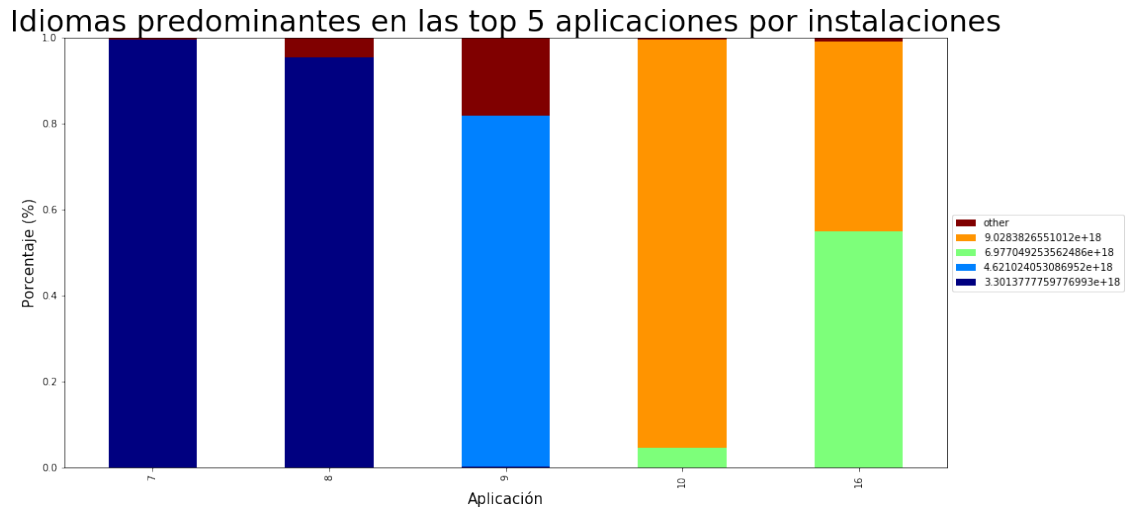


Figura 64: Idioma de entrada de los dispositivos para las 5 aplicaciones líderes en instalaciones

Como se puede ver, las aplicaciones líderes (7 y 9 respectivamente) se utilizan con idiomas distintos, mientras que en 10 y 16 predominan otros dos lenguajes diferentes.

2.4.9. Instalaciones por marca

La información de marca y modelo de los dispositivos que instalaron.

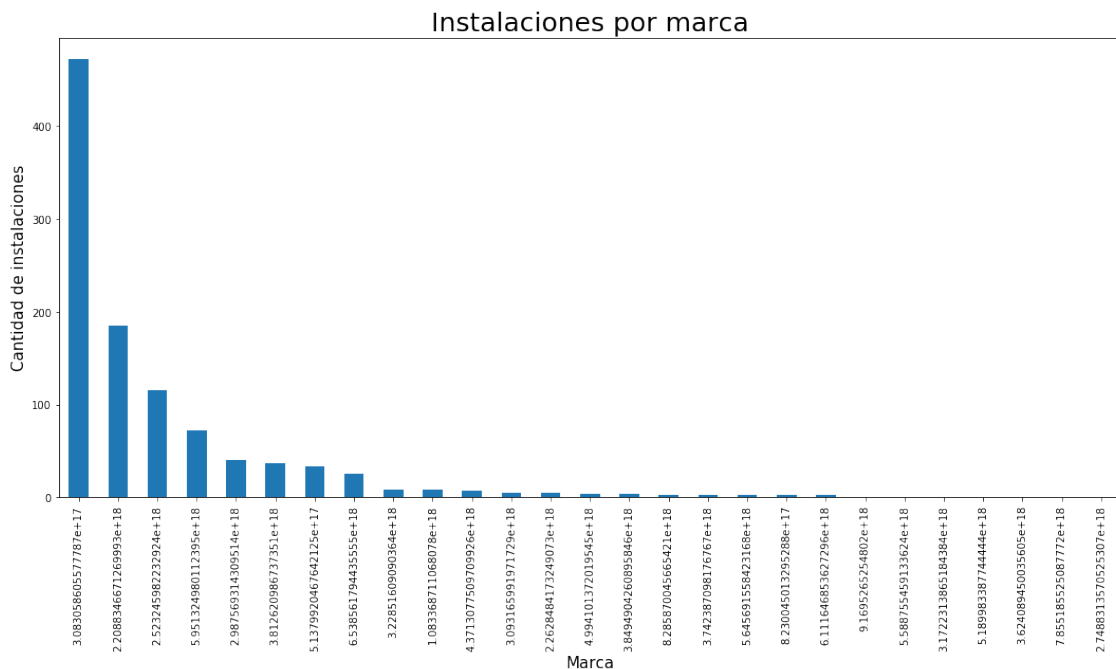


Figura 65: Marcas con más instalaciones

Como se aprecia, hay una marca que lidera absolutamente con una cantidad de installs dos veces y media mayor a su inmediata competidora.

Lo interesante será conocer además el tipo de dispositivo que producen esas marcas, para así determinar el tipo de publicidades que soportarán y así elegir mejor al advertiser.

Del gráfico vemos que todas las marcas de las que se tiene registro utilizan el tipo 1891515180541284343 a excepción de una, la 5.951324980112395e+18, que utiliza el otro.

2.4.10. Instalaciones wifi y user agents

Otro aspecto a considerar es la conexión con la que cuentan los usuarios al momento de decidir si hacer caso o no a alguna publicidad. Los datos que nos proporciona Jampp nos indican si las instalaciones fueron hechas vía conexión wifi.

Inicialmente se esperaría que sea mayor el número de instalaciones vía wifi, ya que los datos móviles suelen tener un límite de uso bastante bajo.

Porcentaje de instalaciones vía Wifi

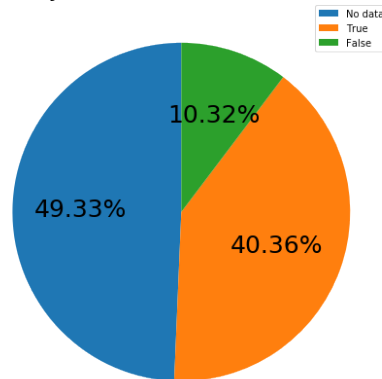


Figura 66: Instalaciones por conexión wifi.

Sorprendentemente la figura muestra que para una gran mayoría de las instalaciones no se proporcionó data respecto de la conexión. Sin embargo, y en concordancia con lo mencionado previamente, en los casos en los que sí se tiene información resulta claro ver que la mayor parte de dichas instalaciones sí fueron realizadas vía conexión wifi.

Trasladando estos datos a las aplicaciones vemos.

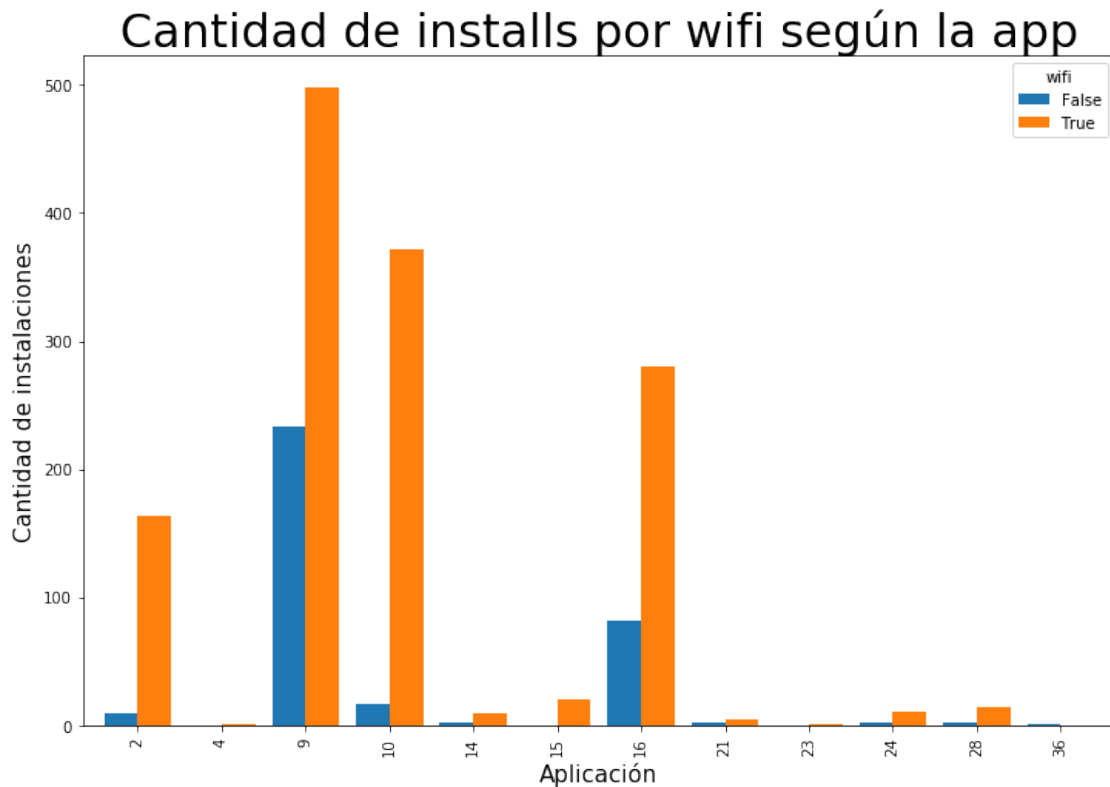


Figura 67: Instalaciones por conexión wifi según la aplicación.

Lo primero que observamos es que la cantidad de aplicaciones de las que se tienen datos de conexión es reducida en comparación al total, ya que, por ejemplo, la aplicación líder en instalaciones no especifica tipo de conexión en ninguna de sus installs. Sin embargo, de este gráfico puede tomarse el hecho de que, si bien la tendencia indica que la mayoría de las instalaciones de las que se tienen datos se realizaron vía wifi, hay aplicaciones como 21, donde la proporción es más pareja, o la 36, donde se invierte. Esto puede servir para determinar qué tipo de publicidad mostrar, puesto que en los casos donde no se cuenta con wifi hay que tener en cuenta el consumo, para no gastar los datos al usuario.

Por otra parte, el hecho de que un install se haya realizado por wifi nos permite conocer otro dato importante, el *user agent* relacionado a la acción.

Análogamente a lo hecho al comienzo de la sección 2.4.3, se analizaron los user agents.

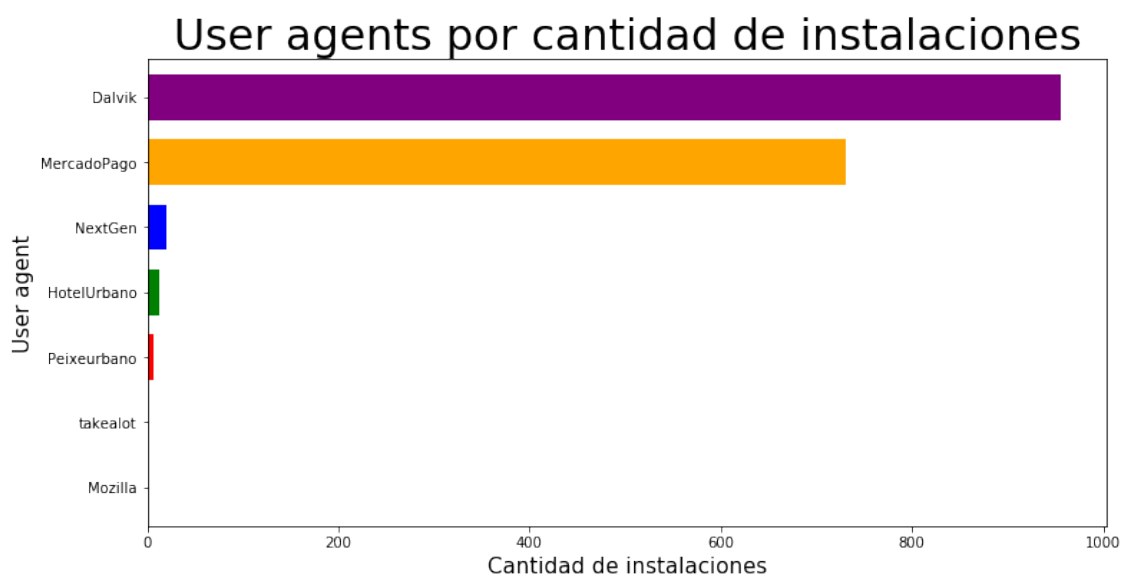


Figura 68: User agents por instalaciones

Resulta evidente que los dominadores absolutos de la categoría son Dalvik y MercadoPago, con agentes casi sin presencia, como Mozilla y takealot, que cuentan sólo con una instalación.

2.4.11. Instalaciones atribuidas a Jampp

Será de particular importancia saber cuantas de las instalaciones se le atribuyeron oficialmente a Jampp, es decir, cuantas de ellas fueron realmente obra de la empresa. Los datos proporcionados hacen dos tipos de distinciones.

- **Attributed:** Indica si la instalación le fue reconocida oficialmente a Jampp.
- **Implicit:** Indica si la instalación se registró de manera implícita, es decir, se realizó por otro canal.

En este aspecto, los datos indican que absolutamente ninguna de las instalaciones le fue atribuida a Jampp, lo que indica un porcentaje de efectividad del 0%. Sin embargo, para la categoría *implicit*, los resultados fueron los siguientes.

Porcentaje de instalaciones externas

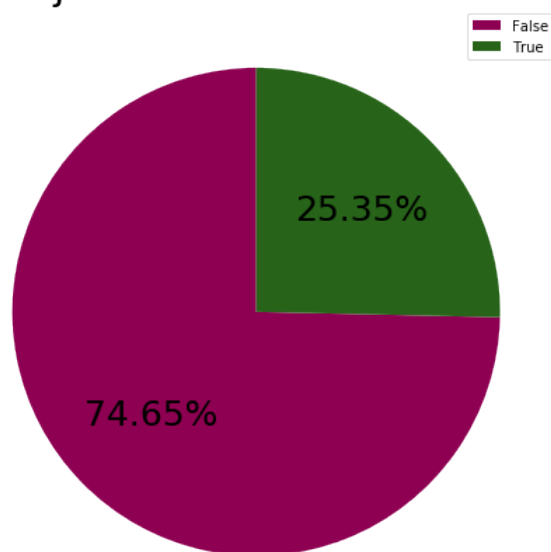


Figura 69: Porcentaje de instalaciones registradas por un canal distinto a Jampp

Los datos muestran un *ratio* de algo más de 1 de cada 4 instalaciones se realizaron por fuera de la influencia de Jampp.

Si lo trasladamos nuevamente a las 5 aplicaciones más instaladas.

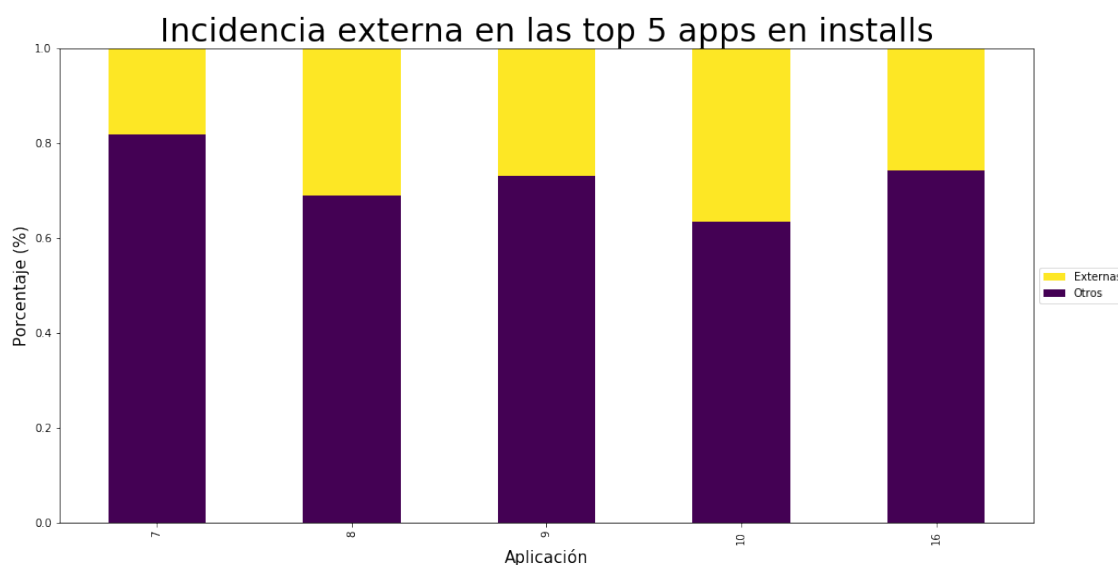


Figura 70: Porcentaje de instalaciones por algún otro método para las top 5 apps en descargas

Como podemos ver, los resultados van acorde a lo mostrado en la figura anterior, con casos como el de la aplicación 10, donde casi el 40 % de las instalaciones son externas. Sin embargo, cabe destacar que en la aplicación líder en descargas, la cual supera a su perseguidora por más de 200 instalaciones (véase sección 2.4.3), es donde estos métodos tienen menos presencia.

3. Análisis de archivos en conjunto

3.1. Clicks y subastas

3.1.1. Análisis general

Hay dos temas que se van a analizar en estos dos archivos en conjunto:

1. Relación entre cantidades de clicks y de subastas.
2. Advertisers

Empezamos mostrando gráficos representativos de los datos.



Figura 71: Relación clicks/subastas

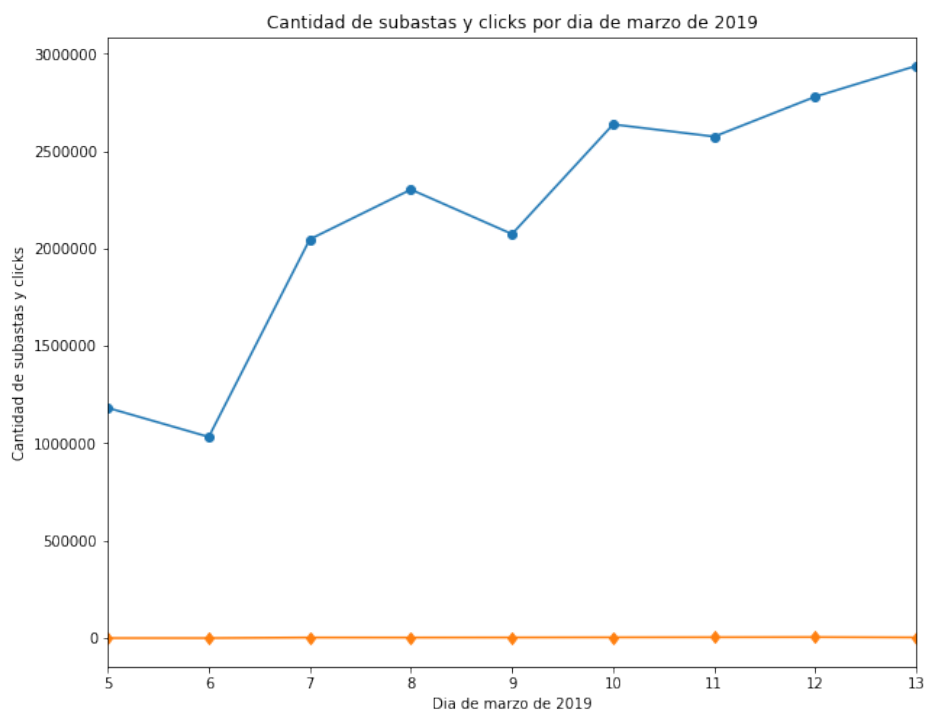


Figura 72: Cantidad de clicks y subastas por día

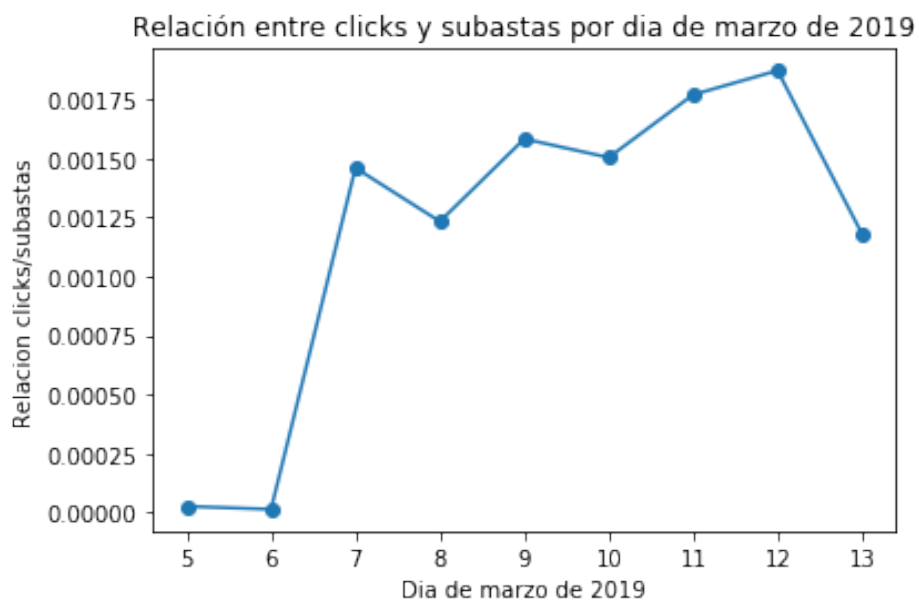


Figura 73: Relación clicks/subastas por día

Podemos obtener conclusiones importantes en base a estos gráficos, teniendo en cuenta que cuando se habla de relación clicks/subastas o de subastas que terminan en clicks no se hace referencia a subastas, o clicks, individuales sino a una relación en conjunto, ya que no podemos identificar que subasta termino en cada click. Por lo tanto cuando digamos relación clicks/subastas estaremos hablando de $\frac{n_{clicks}}{n_{subastas}}$, siendo n_{clicks} la cantidad de clicks y $n_{subastas}$ la cantidad de subastas.

- La cantidad de clicks es ínfima en relación a la cantidad de subastas. Se pueden perder cantidades en dos partes del proceso (pocas subastas se ganan y/o poca gente hace clicks una vez que se gana la subasta y se muestra un anuncio). Gráfico 67.
- No tiene mucho sentido el gráfico 68 ya que la cantidad de subastas es tan grande en comparación con la cantidad de clicks que la figura de clicks se achata cercana a cero.
- El gráfico 69 nos permite ver mejor lo confirmado en el ítem anterior. Día a día la relación $\frac{n_{clicks}}{n_{subastas}}$ es menor al 0,005

Por otra parte, vimos en la sección "Subastas" que cada device_id tenía asociado una plataforma. Los datos obtenidos nos permiten saber quién es el advertiser predominante en cada plataforma. Para ello vemos dos gráficos (uno para la plataforma '1' y otro para la plataforma '2').

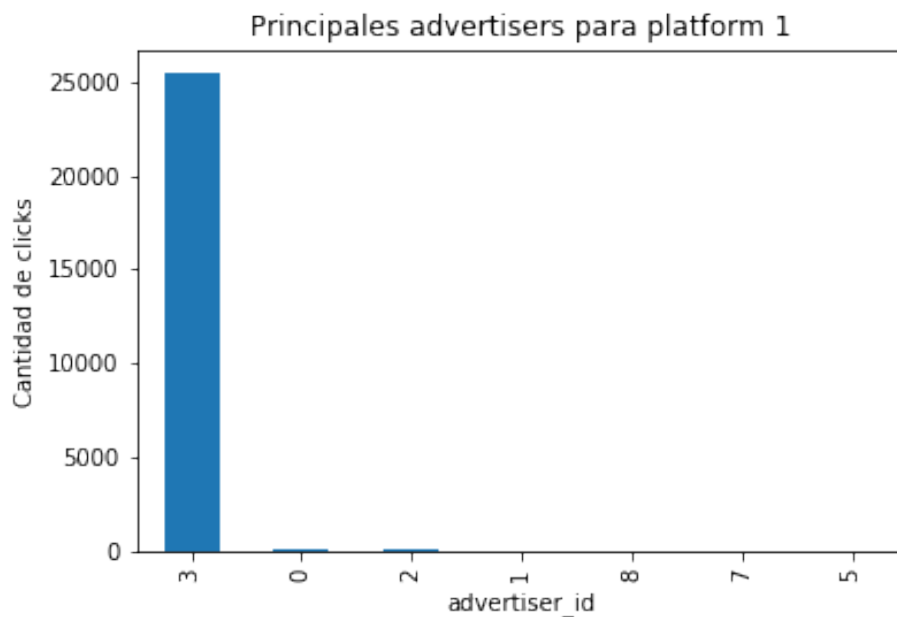


Figura 74: Cantidad de clicks en dispositivos con plataforma '1' para cada advertiser

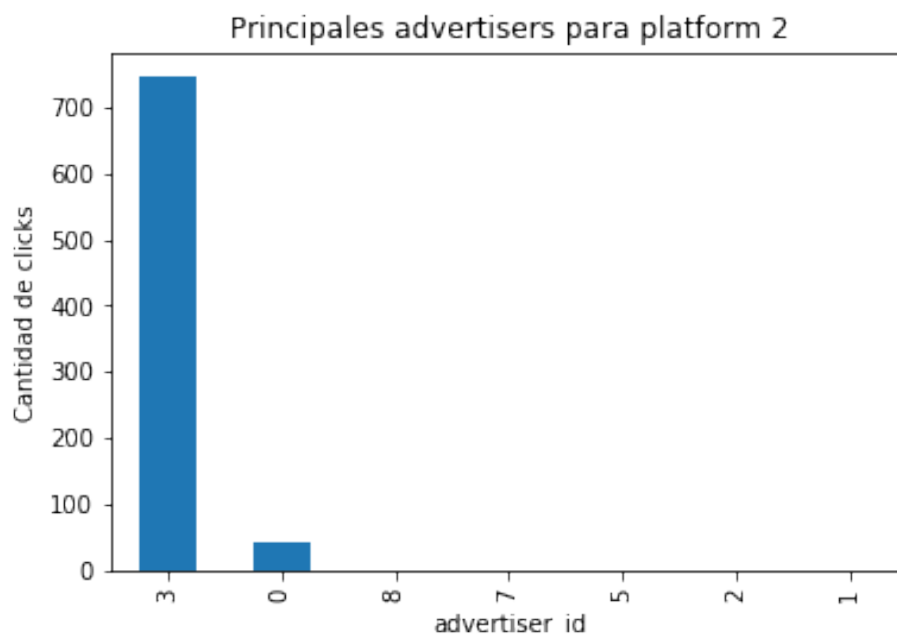


Figura 75: Cantidad de clicks en dispositivos con plataforma '2' para cada advertiser

Se puede ver, claramente, que el advertiser '3' es predominante para ambas plataformas. Pero además, un dato interesante es la diferencia en la cantidad de clicks entre las plataformas. Cerca de 25000 para la plataforma '1' y cerca de 700 para la plataforma '2' (en el anexo 'tablas de datos' se podrán comprobar los números exactos). Esta diferencia entre los sistemas operativos ya la habíamos visto en la sección "Subastas".

3.2. Clicks e instalaciones

3.2.1. Análisis general

Decidimos hacer un join entre installs y clicks para poder saber de cuantas personas hacen click cauntas installan. Para poder hacer el join primero se nos ocurrió hacerlo por la fecha y hora. Pero nos dimos cuenta que si los uniamos por esta columna no necesariamente el click iba a ser correspondiente al install. Esto se debe a que cuando una persona hace click, no instala en ese mismo instante, sino, que pasan unos segundos o hasta minutos. Luego tuvimos la idea de unir los archivos por medio del transaction id ya que este era único para cada uno de los click e installs. Sin embargo, esta idea no funcionó debido a que en installs solo habian 4 trans_id y el resto nulls y ninguno de esos 4 coincidía con algún transaction_id de clicks. Finalmente, decidimos unirlos por el ref.hash ya que este es un id unico de cada celular.

Al juntarlos por ref.hash, solamente obtuvimos, de los más de 26000 clicks que había, un csv con 11 filas. Cabe destacar que de estos 11 que quedaron el ref.hash 582930240149217282 se repite un total de 4 veces y el ref.hash 7759178785240189555 se repite 2 veces. Ambos tienen esa misma cantidad de repeticiones en el archivo clicks mientras que solo están una vez en el archivo installs. Es decir, que de los 4 clicks que hizo el teléfono 582930240149217282 solo uno fue una instalación y de los

2 clicks que hizo 7759178785240189555 solo uno fue una instalación. Por lo tanto, se podrá decir que de los ams de 26000 clicks solo se realizaron 7 installs. Este número es muy bajo.

Antes de realizar los gráficos se optó por sacar columnas que no tenían ningún tipo de información útil ya que o tenían muchos nulls o no daban información relevante. Algunas de esas columnas fueron `action_id`, `agent_device`, `brand`, `kind`, entre varias otras. Los siguientes gráficos se van a realizar a las 11 filas que quedaron del join entre clicks e installs.

3.2.2. Clicks e instalaciones en `advertiser_id`

Este gráfico no mostrará la cantidad de `advertiser_id` que quedaron después de quedarnos con los clicks que fueron installs.

Cantidad de clicks e installs en común por `ref_hash` en los distintos `advertiser_id`

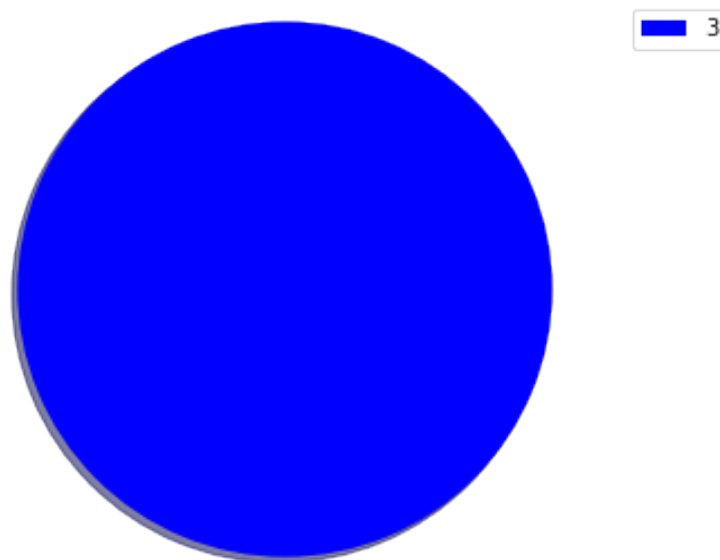


Figura 76

Finalmente, como se esperaba, quedó solo el `advertiser_id` 3 que era el que predominaba en el archivo de clicks.

3.2.3. Clicks e instalaciones en `source_id`

En los siguientes gráficos se notará la cantidad y el porcentaje de clicks que terminaron en installs de los distintos `source_id`.

Cantidad de clicks e installs en comun por ref_hash en los distintos source_id

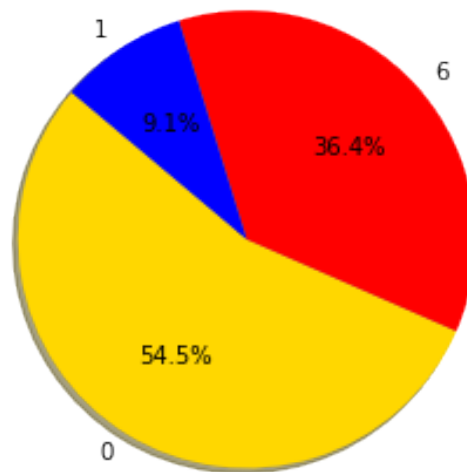


Figura 77

Cantidad de clicks e installs en comun por ref_hash en los distintos source_id

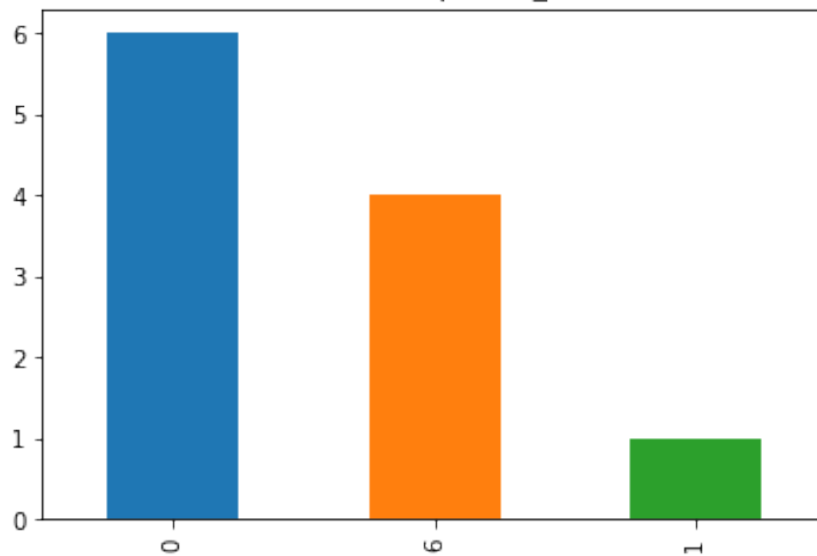


Figura 78

Quedaron simplemente tres source_id que eran los que más abundaban en el csv de clicks.

3.2.4. Posicion geografica de clicks e instalaciones

Este heatmap nos muestra la latitud y longitud de donde se realizó el click que terminó en instalación.

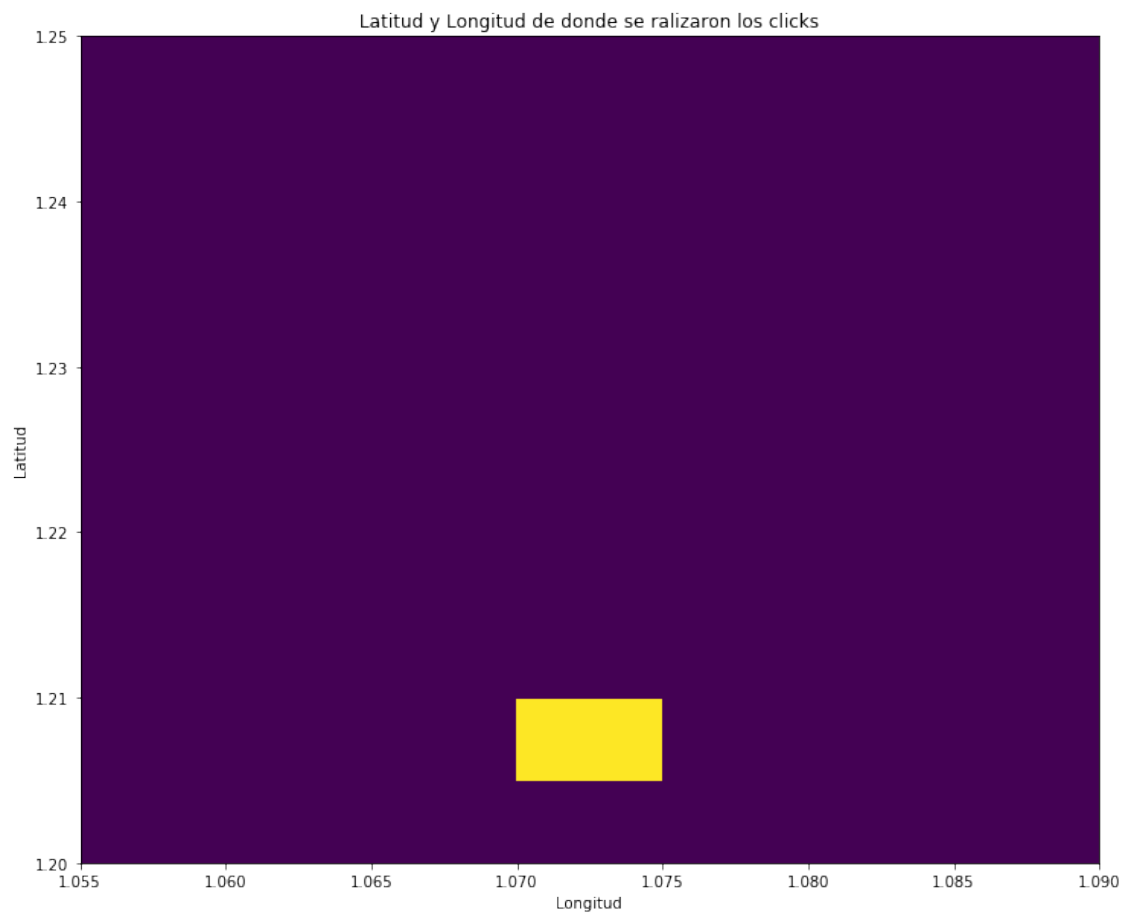


Figura 79

Todas las instalaciones se realizaron entre las latitudes 1.21 y 1.20 que intersecan con la longitud 1.07 y alrededores. Esto era de esperarse ya que la mayoría de clicks se habían consentrado en ese mismo sitio.

3.2.5. Clicks en la pantalla que terminan en instalaciones

El siguiente scatter plot nos indica en que lugar de la pantalla del teefono celular se realizó el click que luego terminó en una instalación de la aplicación.

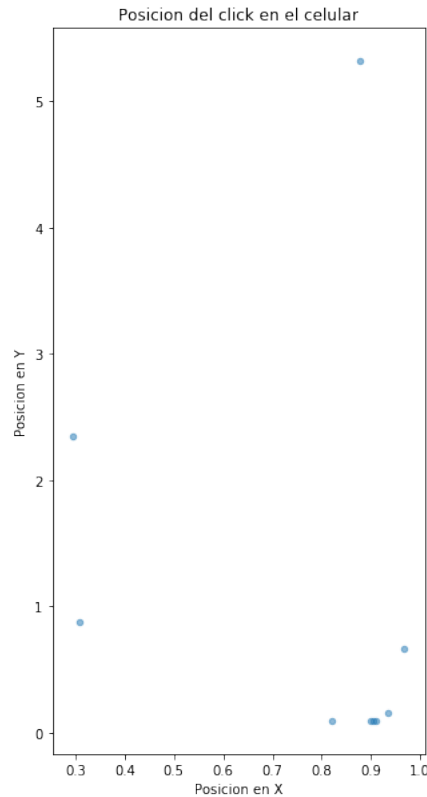


Figura 80

Se puede observar que la mayoría de los clicks fueron en la parte inferior derecha de la pantalla. Esto era de esperarse por lo ya previamente analizado en el inciso de clicks.

3.2.6. Tiempo en realizar el click

Queríamos saber cual era el promedio de tiempo que tardaba una persona en hacer click que luego terminaa instalando la aplicación. Para esto primeros filtramos el dataframe y nos quedamos con los que tenían un tiempo menor a 10 minutos y, obviamente, con los que no eran nulls. Después procedimos a calcular el promedio que nos dio como resultado 19.230125 segundos.

En esta oportunidad no nos pareció importante analizar la distribución en el tiempo de los clicks que termianron en instalaciones de los disitintos application_id o de los distintos source_id como se hizo previamente en el archivo de clicks ya que hicimos los gráficos y estos, con tan poca cantidad de clicks que quedaron, no nos mostraban nada relevante.

3.2.7. Clicks e instalaciones en Android e IOS

Vale aclarar, que tanto en clicks como en installs, se mostraba una columna con el sistema operativo del celular llamada `ref_type`. Antes de realizar algún gráfico, nos fijamos que ambas columnas que quedaron después de hacer el join llamadas `ref_type_x` y `ref_type_y` tuvieran el mismo sistema operativo. Por suerte, así lo fue y no se encontró un error en ese aspecto y procedimos a hacer el gráfico con una de estas dos columnas que nos mostraban exactamente lo mismo.

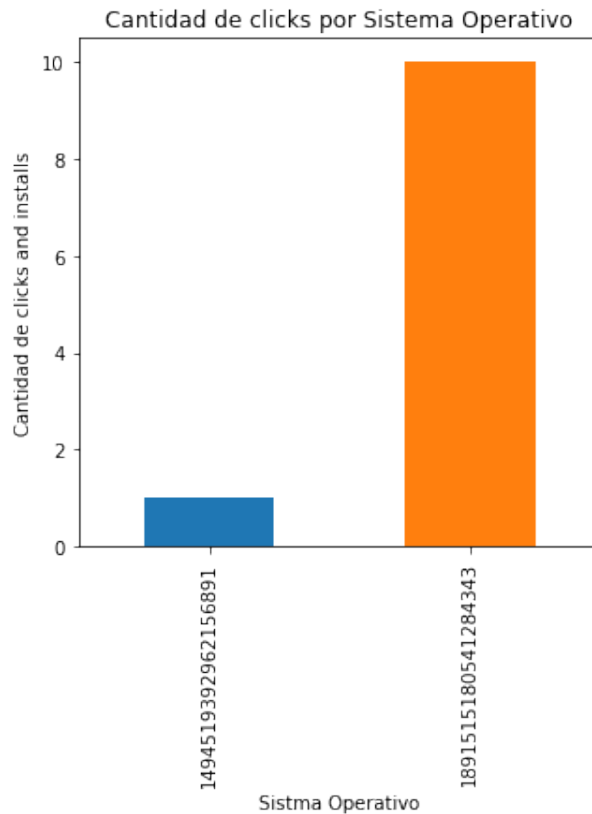


Figura 81

Se puede apreciar que al igual que en clicks el sistema operativo que más aparece es el que comienza con 18 y el que le sigue es el que comienza con 14.

3.2.8. Clicks e instalaciones en `carriers_id`

Vamos a mostrar en el siguiente gráfico cuales son los `carriers_id` que quedaron y en que cantidad después de elegir los clicks que terminaron en instalaciones.

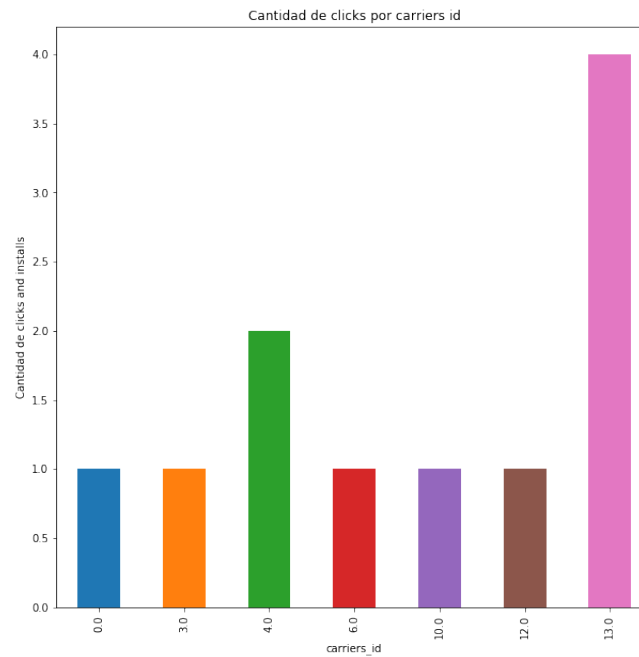


Figura 82

Sorpresivamente, el carrier_id predominante es el 13.0 que no era mayoría en el archivo de clicks.

3.2.9. Clicks e instalaciones en os_minor

En el siguiente gráfico se enseña las menores versiones del sistema operativo de cada celular que hizo click e instalo la aplicación.

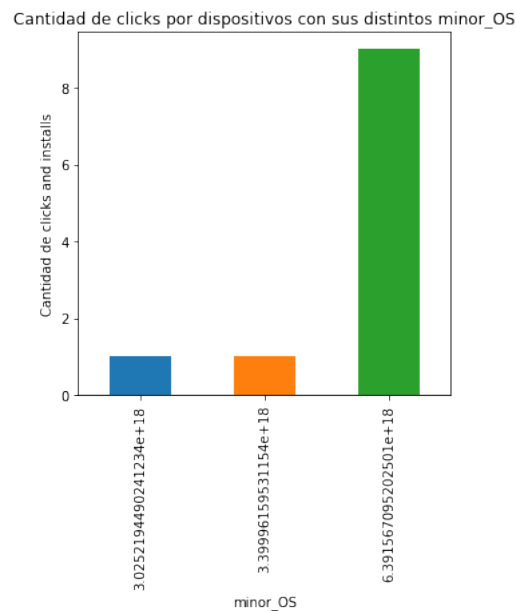


Figura 83

En este caso, solo quedaron 3 tipos de os_minor que no eran los que más pre-

dominaban en clicks.

3.2.10. Clicks e instalaciones en os_minor

En el siguiente gráfico se enseña las mayores versiones del sistema operativo de cada celular que hizo click e instalo la aplicación.

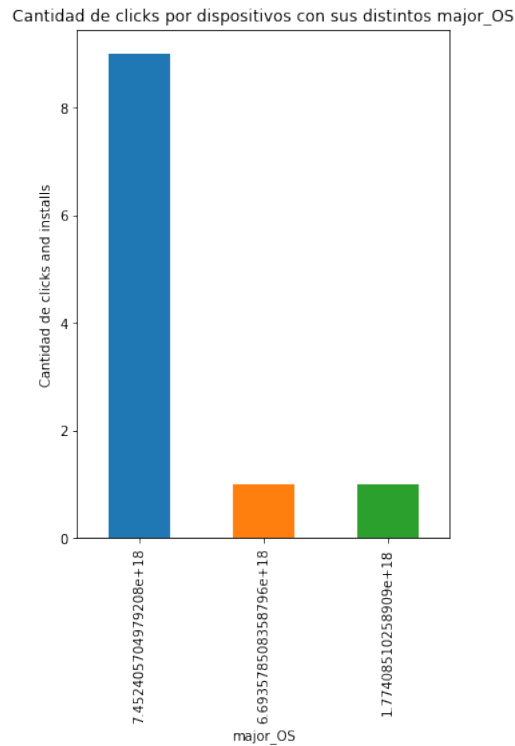


Figura 84

Solamente quedaron 3 os_major y el más repetitivo es el que comienza con 7.452 que era el que menos aparecía en el archivo de clicks.

3.2.11. Clicks e instalaciones en specs_brand

Este gráfico nos muestran los specs_brand que quedaron y en que cantidad teniendo en cuenta solo los clicks que instalaron la aplicación.

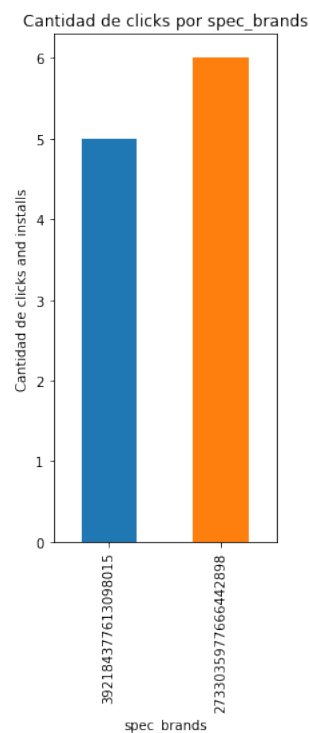


Figura 85

Como se esperaba, quedaron los dos specs_brand que más aparecían en el archivo de clicks.

3.2.12. Clicks e instalaciones en país de origen

En el archivo de clicks solo aparecía un solo país mientras que en el archivo de installs aparecían dos distintos. Luego de hacer el join solo quedó un país como se suponía.

3.2.13. Clicks e instalaciones en application_id

En el siguiente gráfico veremos las aplicaciones_id que quedaron luego de elegir los clicks que instalaron.

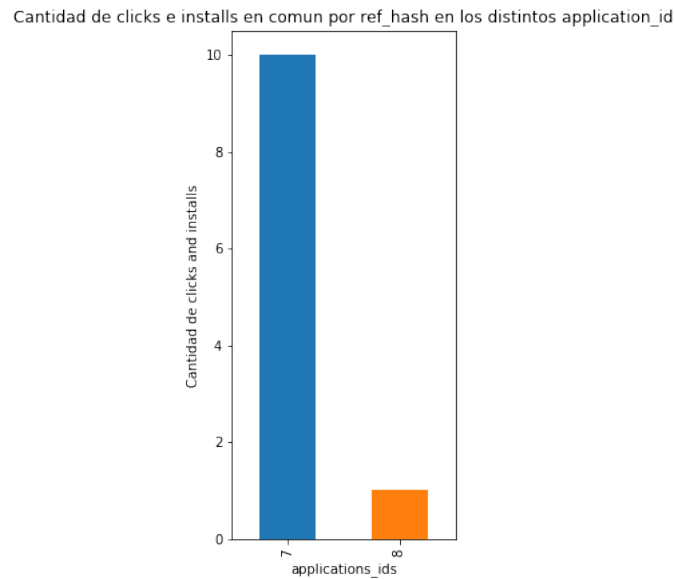


Figura 86

Casi todas las application_id son del tipo 7 que era lo más lógico ya que esa era la application_id que más abundaba en el archivo de installs.

3.2.14. Tipo de instalaciones

Dentro de los tipos de instalaciones tenemos las atribuidas y las implícitas. Como ya se mencionó anterior mente las atribuidas son todas false, lo que es muy raro ya que estas serían las que se le darían a Jampp.

El siguiente gráfico nos muestra cuantas de las instalaciones fueron implícitas.

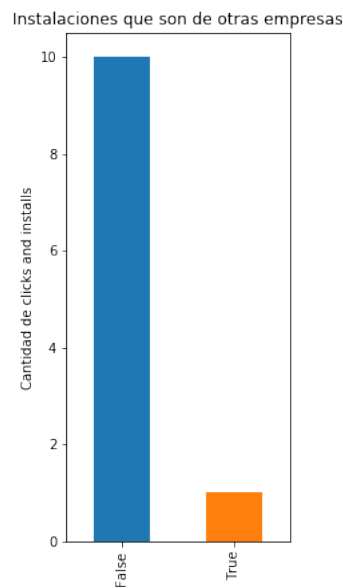


Figura 87

Se observa que una de ellas es implícita. El resto de las instalaciones no se sabe con certeza a quien le corresponden.

3.2.15. Clicks e instalaciones en device_model

El gráfico mostrado a continuación nos enseñará que modelos de celular quedaron y en que cantidad luego de quedarnos con los installs que les hicieron click previamente.

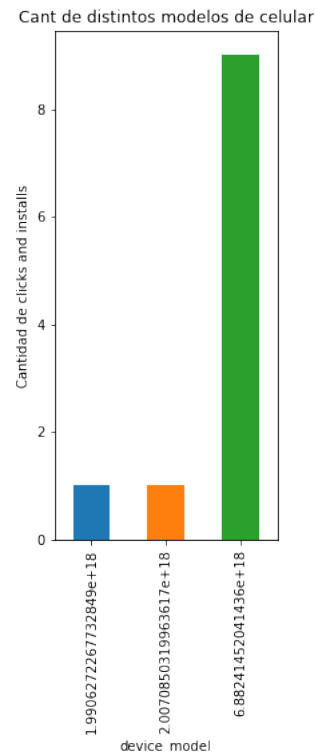


Figura 88

Se puede notar que solamente quedaron 3 tipos de modelos de celular de los cuales el más repetitivo es el comenzado con 6.88.

3.2.16. Clicks e instalaciones en device_language

En el siguiente gráfico mostraremos los lenguajes que tenían los celulares al momento de la instalación de la aplicación.

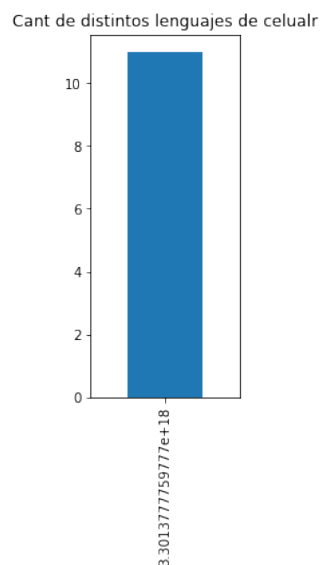


Figura 89

Solamente quedó un lenguaje.

3.2.17. Clicks e instalaciones en session_user_agent

En este gráfico que se enseñará a continuación se notará los distintos session_user_agent.

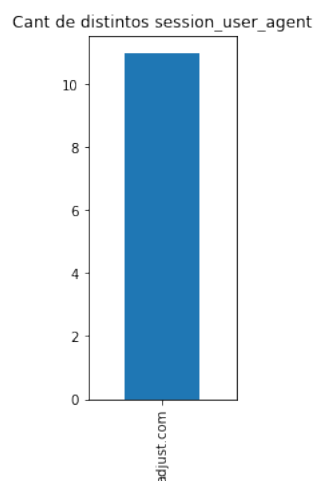


Figura 90

Se puede observar que el único session_user_agent que quedó es adjust.com

3.2.18. Clicks e instalaciones en ip_address

El siguiente gráfico nos mostrará los ip_address de los clicks que terminaron en instalación.

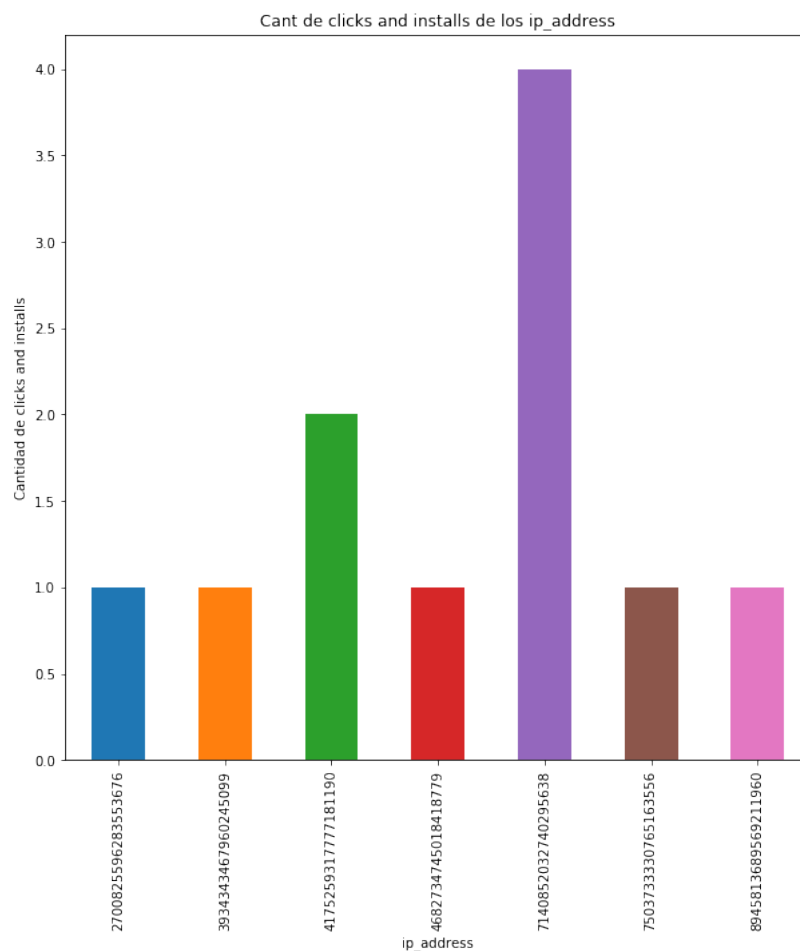


Figura 91

El ip_address que predomina es el comanzado en 7140 con un total de 4 apariciones. Esto igual es de esperarse porque hay una ip_address por cada celular y como en total tenemos 7 ya que uno se repite 4 veces y otro 2 veces. Por lo tanto se obtienen 7 dsitintas ip_address.

3.2.19. Clicks e instalaciones en los días y horas

Cabe destacar que cada uno de los csv tenia su fecha y hora. En estos dos siguientes gráficos vamos a mostrar los días en que hubo más clicks y los días que más se instaló.

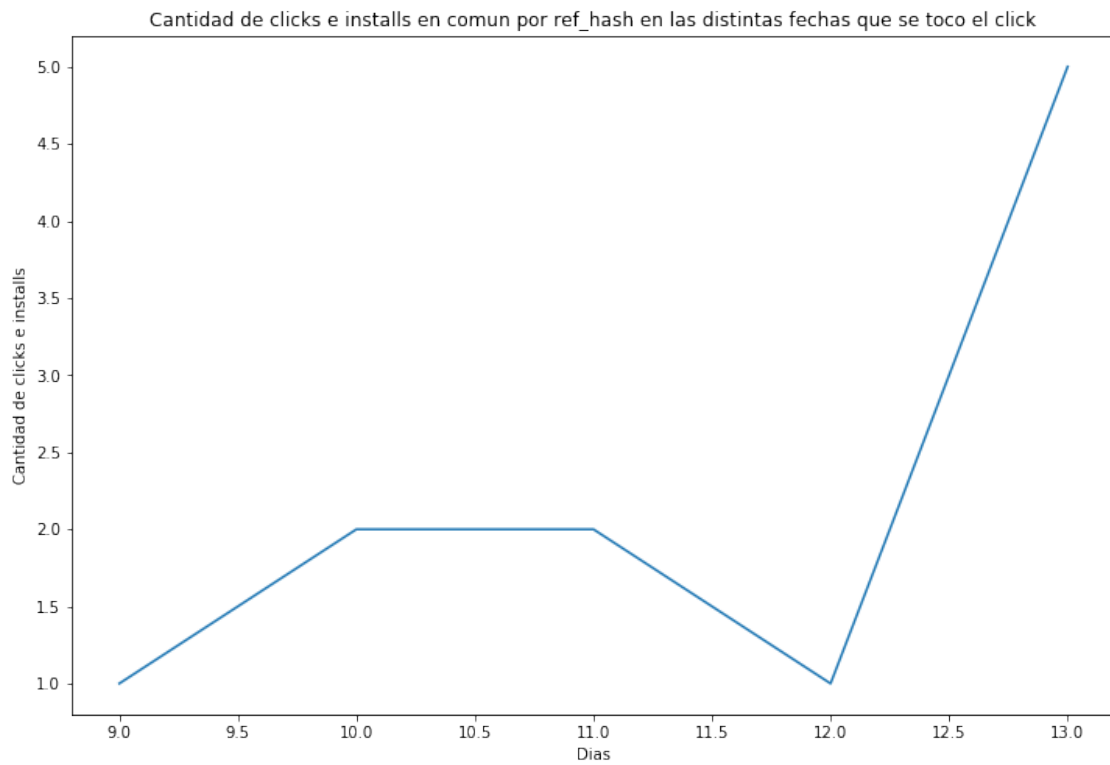


Figura 92

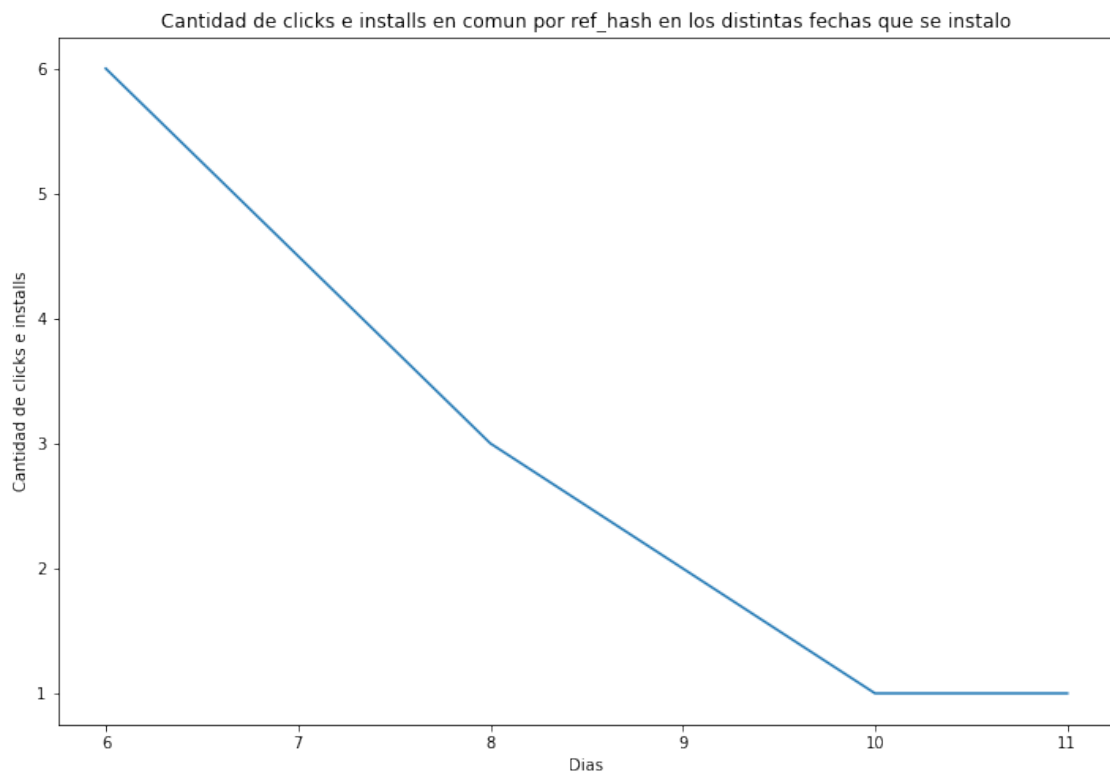


Figura 93

Viendo estos gráficos notamos algo muy raro. La instalación que se realizó luego de tocar click debería ser uno segundos o minutos después de hacer el click, no obstante, se puede ver que los clicks se realizaron entre el 9 y el 13 de Marzo y los installs entre el 6 y el 11 de Marzo. Eso no tienen ningún sentido. Se puede deber a algún error en los datos, ya sea en la fecha y el horario o en los ref_hash por los que unimos los archivos clicks e installs.

En los siguientes gráficos se enseñará los clicks y las instalaciones en las distintas horas del día.

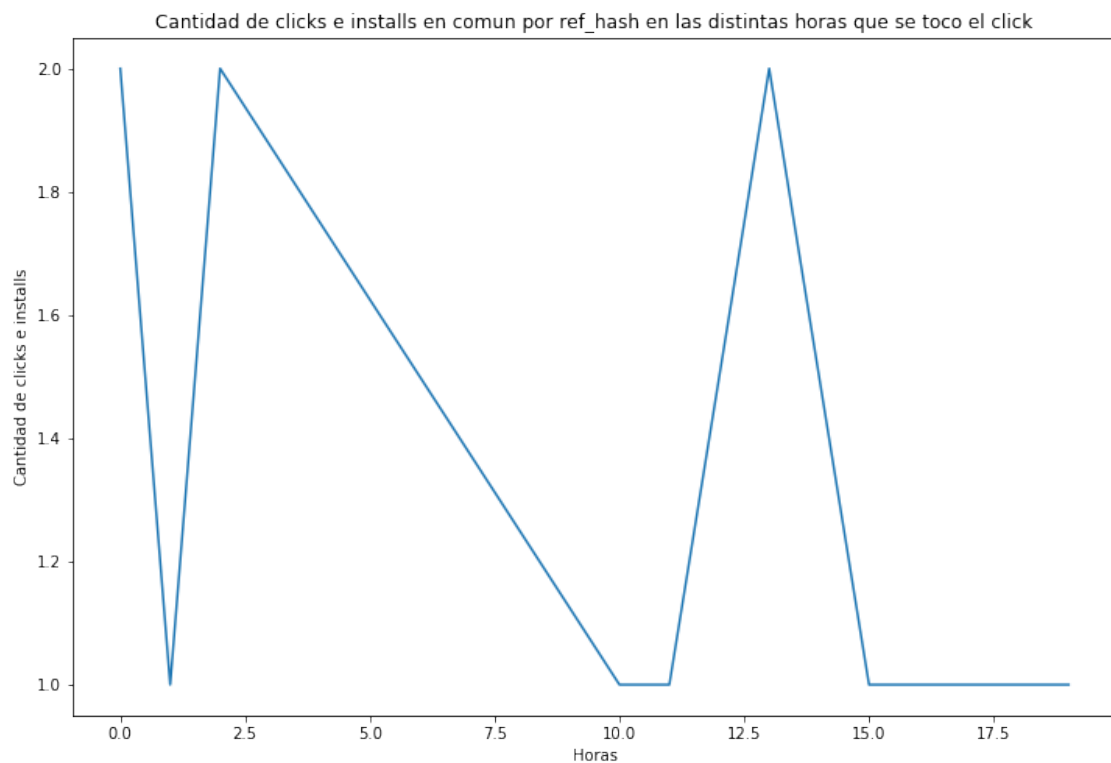


Figura 94

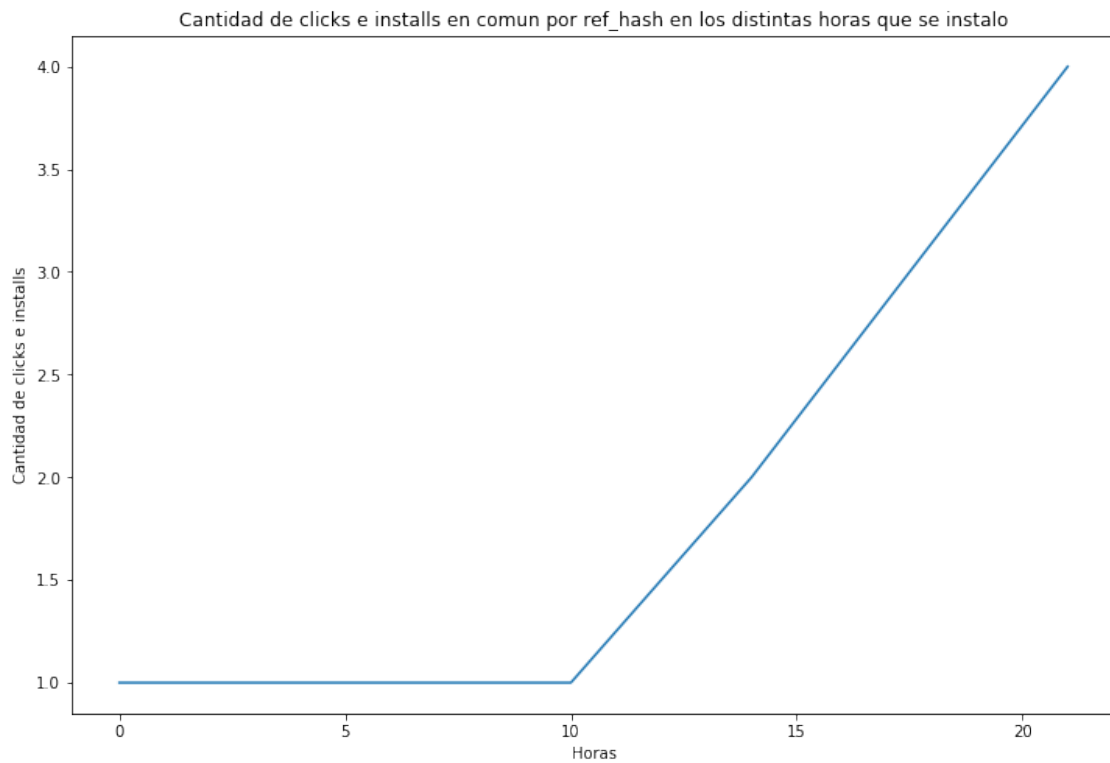


Figura 95

En este caso también pasa lo mismo. No se ve que haya coherencia entre ambos gráficos. Las horas del día en la que se hace click y las que se instala son muy diferentes. Como se mencionó previamnte esto puede ser un error en las fechas y horas o también en el ref_hash.

4. Conclusiones

Luego de realizar un análisis de todos los datos nos enfocamos en las conclusiones que generen valor para la empresa. Intentando encontrar patrones que nos acerquen a predecir instalaciones que sean asignadas en el futuro y detectar posibles clientes.

- La mayoría de las subastas se ofrecen desde las 3 PM hasta la 1 AM. Esta tendencia se mantiene para ambos SO.
- Hay un SO que ofrece hasta 5 veces más subastas.
- La mayoría de las subastas provienen de una misma fuente, es necesario buscar más fuentes de subastas. Al parecer esta fuente solo está funcionando para un solo ad, que de seguro solo lleva a una aplicación.
- El carrier parece no ser una variable muy importante al momento de determinar el éxito de un ad. Puesto que la mayoría de los clicks está distribuida entre 7 carriers que lideran. Aunque si podemos determinar que hay algunos carriers que implicarían una certeza en que no vamos a recibir un click. Esto también es útil, evitará en el futuro que ofertemos en subastas donde éstos se vean implicados.
- Al no saber a qué versiones corresponden las variables `os_minor` y `os_major` es difícil sacar conclusiones, pero ciertamente esta es una variable a la que hay que prestarle atención, no tiene sentido participar de subastas donde el sistema operativo del celular en el que se mostrará la publicidad no se corresponde con los requerimientos de la app que queremos publicar.
- En cuanto a las marcas, hay 3 marcas que lideran los clicks. Enfocarse en ellas mejorará el rendimiento de las ofertas al participar de las subastas. Aunque, otra vez, queda por encontrar formas de mejorar el rendimiento de los otros ads, porque lo anterior dicho es cierto solo para 1 ad.
- Hay un sistema operativo que destaca en la cantidad de clicks logrados. Quedará pendiente para la empresa analizar si la publicidad que se está queriendo exponer debe mejorarse para el otro SO que queda resagado. O tal vez es un problema de compatibilidad de las apps que se ofrecen y su disponibilidad en el otro SO.
- La cantidad de clicks aumenta entre las 8 PM y las 2 AM. Pero hay que tener cuidado porque las subastas también se ofrecen más en esos horarios. De todas formas es un dato muy útil, las subastas entre las 3 PM y las 8 PM, mencionadas unos puntos atrás, no están rindiendo bien, conviene no ofertar en esos horarios.
- La ubicación del dispositivo parece influenciar, hay posiciones donde claramente es más probable que recibamos un click, entre las latitudes 1.21 y 1.20 que intersecan con la longitud 1.07 y alrededores. Sería necesario cruzar esta información con algún análisis de densidad de población de esa zona para

estar seguros de si no se trata, simplemente, de que en esa zona habita más gente y por ende hay más dispositivos.

- La posición del click en la pantalla nos ofrece una muestra de gente haciendo click principalmente abajo, lo que sugiere una llamada a la acción en ese sector de la publicidad. Debe haber un dibujo de un botón en la imagen o una flecha. Esto puede ser especialmente útil para detectar diferencias de diseño entre la publicidad que está funcionando y las que no.
- El problema con los eventos es que no sabemos si todos indican algo positivo para la empresa. Asumiremos en primera medida que si. Con esto en mente, podemos analizar que se cumple nuevamente el patrón de interacción entre las 3 PM y las 2 AM.
- No hay un día donde rindan mejor las publicidades según los eventos.
- Los idiomas son bastante variados, esto puede indicarnos que necesitamos hacer los anuncios más específicos para cada cliente, también nos habla un poco del público que tiene cada aplicación. Es imprescindible determinar si los ads están en el idioma 3.3013777759776993e+18 que predomina en los eventos.
- Si bien los eventos están bastante parejos entre el uso de Wifi o datos del celular, hay una ligera ventaja para el uso de Wifi, para ver esto fue más fácil determinarlo por el tipo de conexión que por el campo wifi. A veces los campos de detalle nos permiten analizar y completar información que por alguna razón llega de manera errónea.
- Hay un carrier que destaca entre los demás. Investigar la estrategia de marketing de ese carrier nos puede ayudar a conocer a nuestro público objetivo.
- El wifi predomina en todos los tipos de evento. Es crucial determinar qué tipos de eventos son positivos para favorecer la aparición de eventos beneficiosos.
- Luego de reconstruir la data de marcas, vemos que tenemos nuevos líderes en el gráfico, 3.083058605577787e+17 y, seguido de lejos, 5.951324980112395e+18
- Las instalaciones muestran un tendencia alcista. Nuevamente el horario de 3 PM a 2 AM es beneficioso, pero, otra vez, también es la franja de más oferta.
- Las aplicaciones 7 y 9 son las más exitosas en las campañas. Al parecer no influye demasiado de qué país es el usuario, solo vemos 2 países en el set de datos y las instalaciones se reparten entre ellos.
- Solo 25 % de las instalaciones son atribuidas a Jampp, si bien esto es un valor relativamente bajo, pareciera ser aceptable. En realidad, deberíamos saber cuánto se está cobrando por cada atribución implícita para saber si es un buen valor en la industria o necesita mejorarse, a priori, pareciera que es bastante bueno, considerando que las conversiones son bastante bajas.
- Vemos que un porcentaje muy bajo de las subastas terminan en click.