Análisis Exploratorio de Datos
Trabajo Práctico 1 - Organización de Datos

Nombre de grupo

Poner fecha

Nombre	Padrón	Mail
Álvarez, Federico	99266	fede.alvarez1997@gmail.com
La Torre, Gabriel	87796	latorregab@gmail.com
Medrano, Lucas Nicolás	99247	lucasmedrano97@gmail.com
Piro Martino, Ariel	99469	ariel.piro@hotmail.com

Contents

1 Introduction

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, 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), reconocer el tipo de datos en cada columna, seguido de un análisis mas profundo para obtener mas información de los datos. Luego se hará un análisis global, buscando información relativa a los archivos en conjunto, permitiendo obtener otro tipo de información.

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 significante. '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, que se supone que debe ser Argentina. 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 Subastas por día de Marzo

Como primer acercamiento a este set de datos, 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 1 representa dicha distribución. En él se pueden observar varias cosas:

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

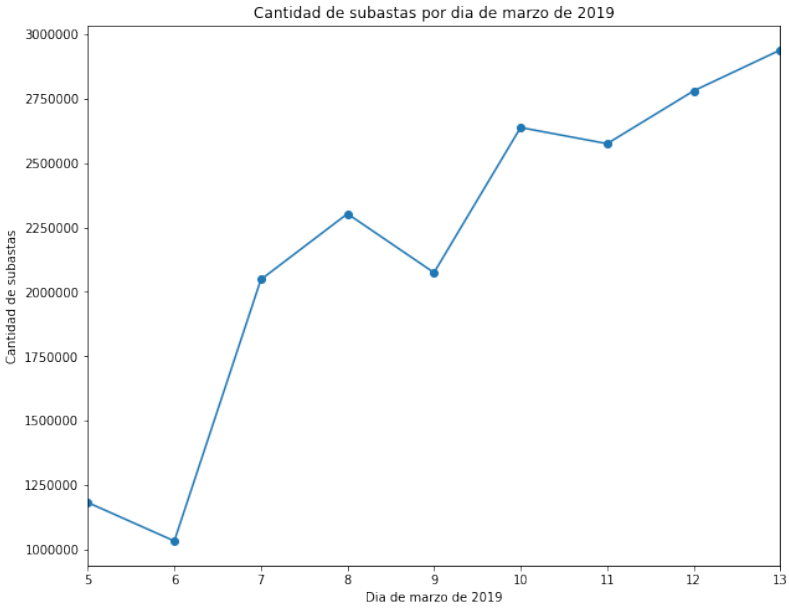

Figure 1: Cantidad de subastas por día de Marzo

2.1.3 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 2 se pueden observar las cantidades. Observese que llamamos '1' y '2' a las plataformas, ya que no sabemos cuál es Android y cuál es iOS. Puntos interesantes a reconocer:

- 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 mas "chata" que la de la plataforma '1', la cual representa mas picos y saltos.
- La figura de la plataforma '1' es muy parecida a la del gráfico 1, mientras que la de la plataforma '2' no lo es. Esto es resultado, principalmente, de lo indicado en el primer ítem. Este análisis puede llegar a ser muy útil para reconocer partes de los datos que son representativas del total.

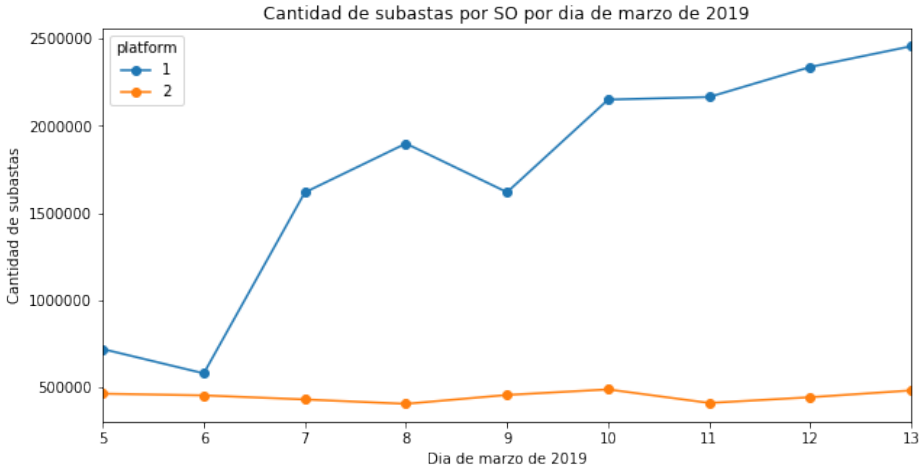

Figure 2: Cantidad de subastas por día de Marzo y por sistema operativo

2.1.4 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 3).

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.

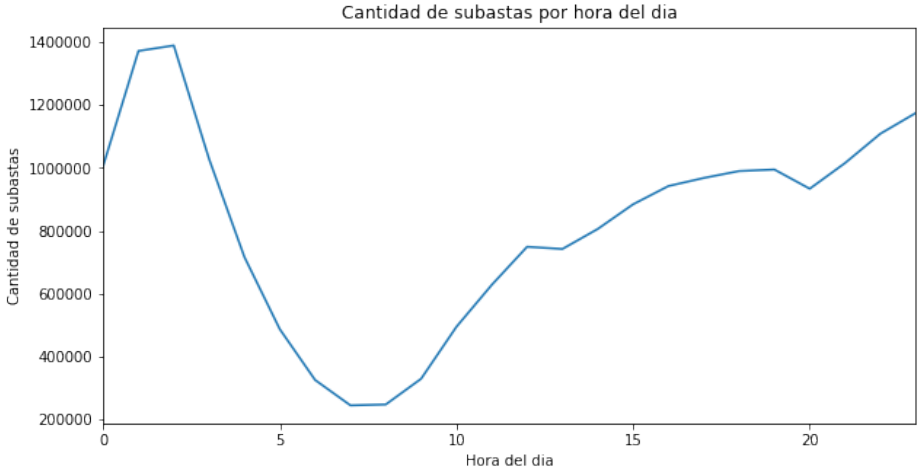

Figure 3: Cantidad de subastas por hora del día

2.1.5 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 mas chato, y con cantidades de subastas mucho mas chicas.

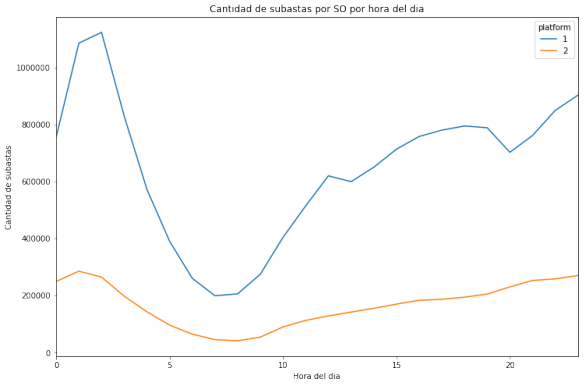

Figure 4: Cantidad de subastas por hora del día por sistema operativo

2.1.6 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 relacion de las cantidades de subastas de ambas plataformas. En las figuras que se ven a continuación podemos confirmar que lo que indicabamos en los gráficos anteriores era cierto. La cantidad de subastas es mucho mayor para la plataforma '1'. Además, ahora tenemos una visión mas cuantitativa de esta relación (en la imagen 5 vemos una relación porcentual, mientras que la imagen 6 nos da mas idea de las cantidades).

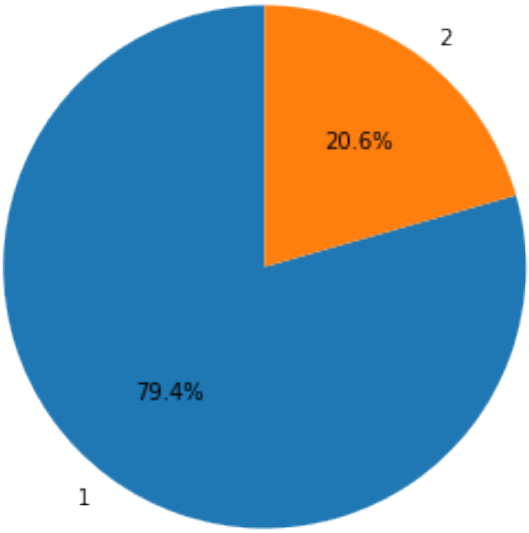

Figure 5: Porcentaje de subastas para cada plataforma

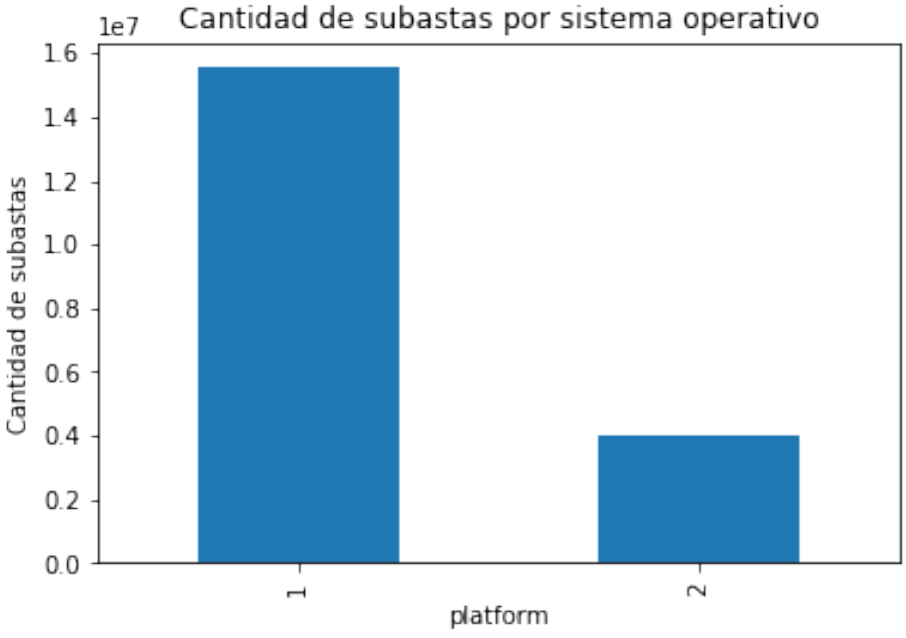

Figure 6: Porcentaje de subastas para cada plataforma

2.1.7 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 millon de subastas generadas.
  - El siguiente (source '1') aunque es mucho menor que el '0', sigue superando a los demas 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.

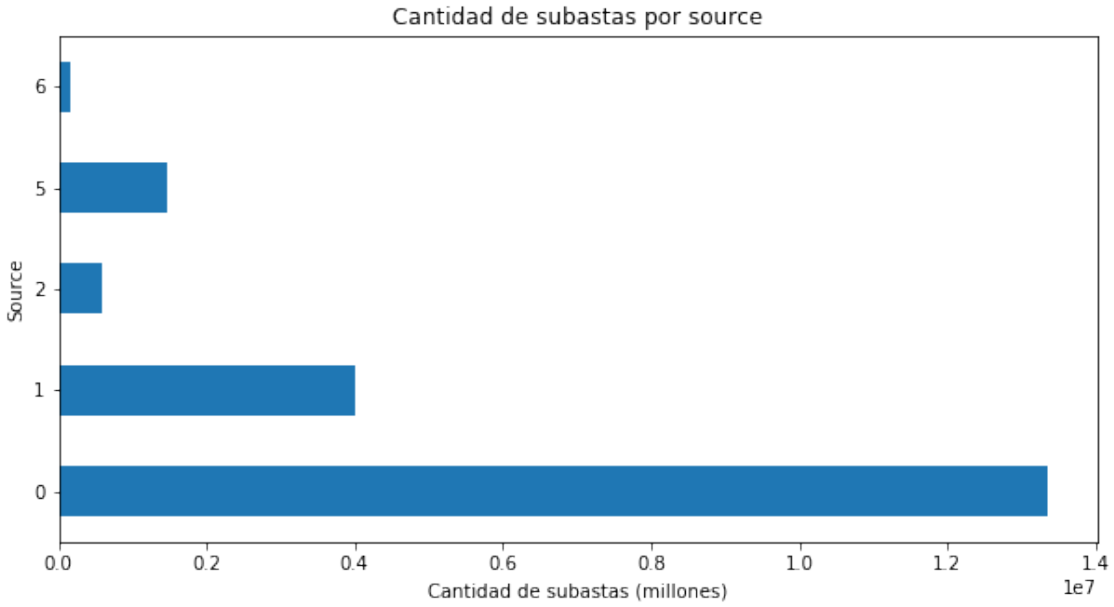

Figure 7: Cantidad de subastas para cada source

2.1.8 Análisis de los dispositivos en las subastas

Otro punto a analizar son los dispositivos. Es un punto que trajo un problema al analizar los datos, y que puede cambiar la forma en la que se entienden los datos anteriormente mencionados. Podemos, por ejemplo, ver el top diez de dispositivos de los cuales se generan mas subastas.

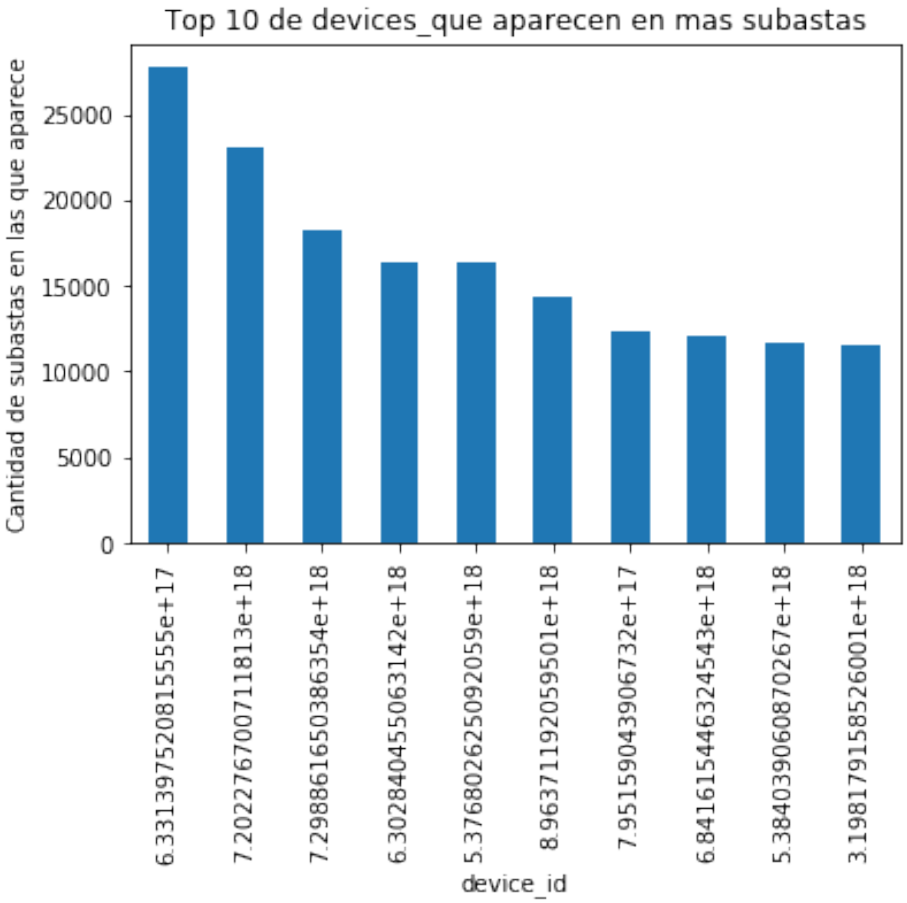

Figure 8: Top 10 de dispositivos con mayor cantidad de subastas

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.

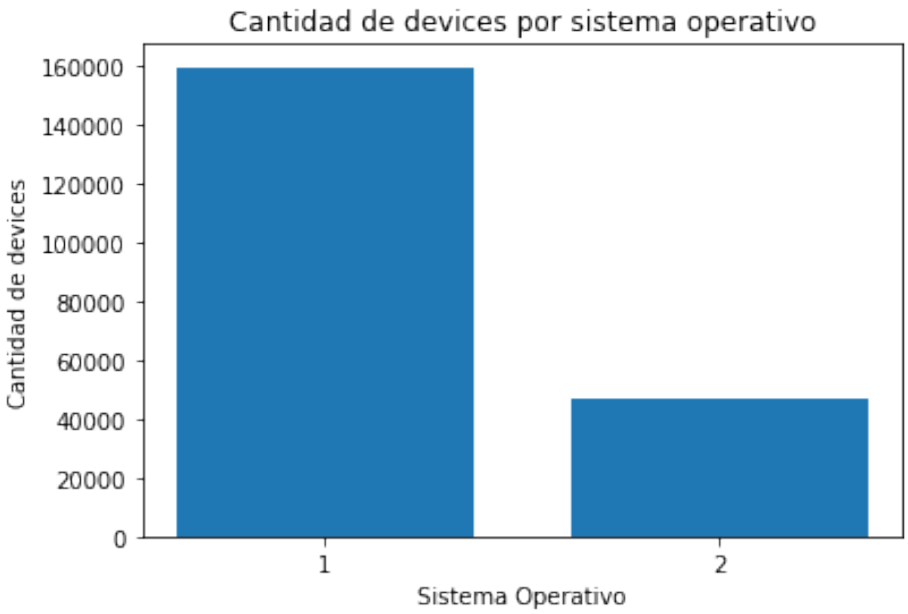

Figure 9: Cantidad de dispositivos para cada plataforma

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 problea por sistemas operativos. Por lo tanto se decidió dejar los datos, y saber que el análisis que se hace cuando se divide el problema, puede no ser tan preciso.

2.2 Clicks

2.3 Eventos

2.4 Instalaciones

2.4.1 Introducción

Los datos sobre instalaciones fueron provistos por Jampp en el archivo `installs.csv`, 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.

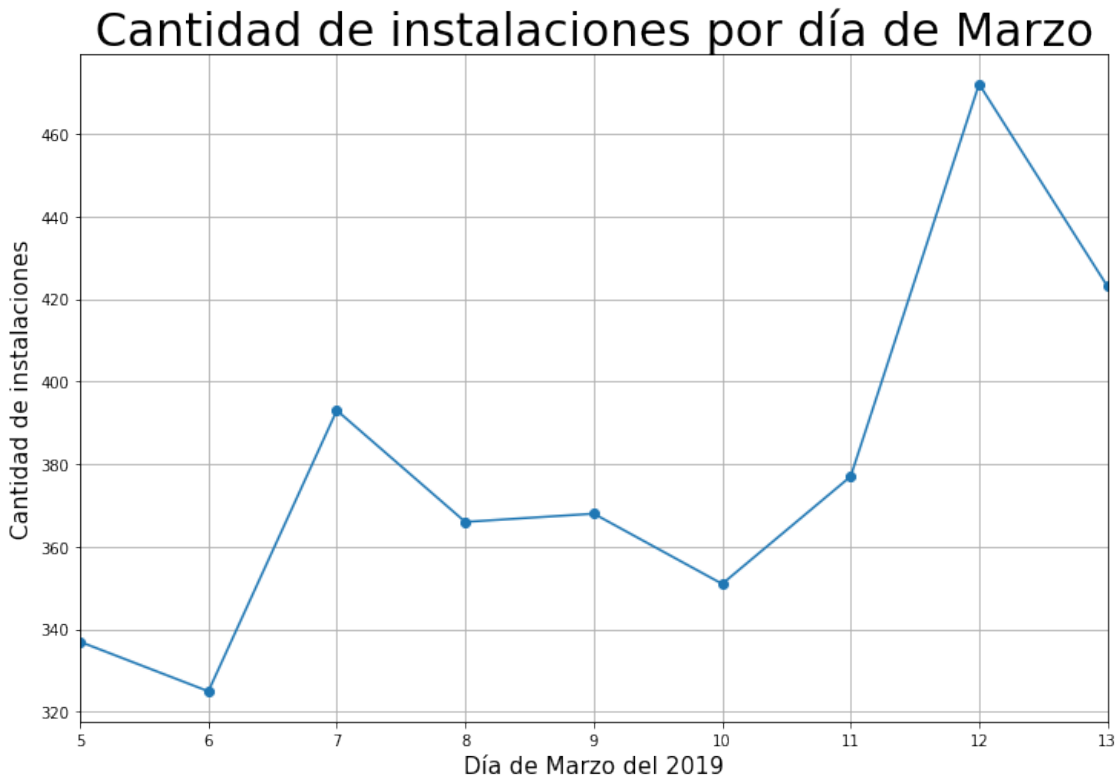

Figure 10: Instalaciones por día de Marzo

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.

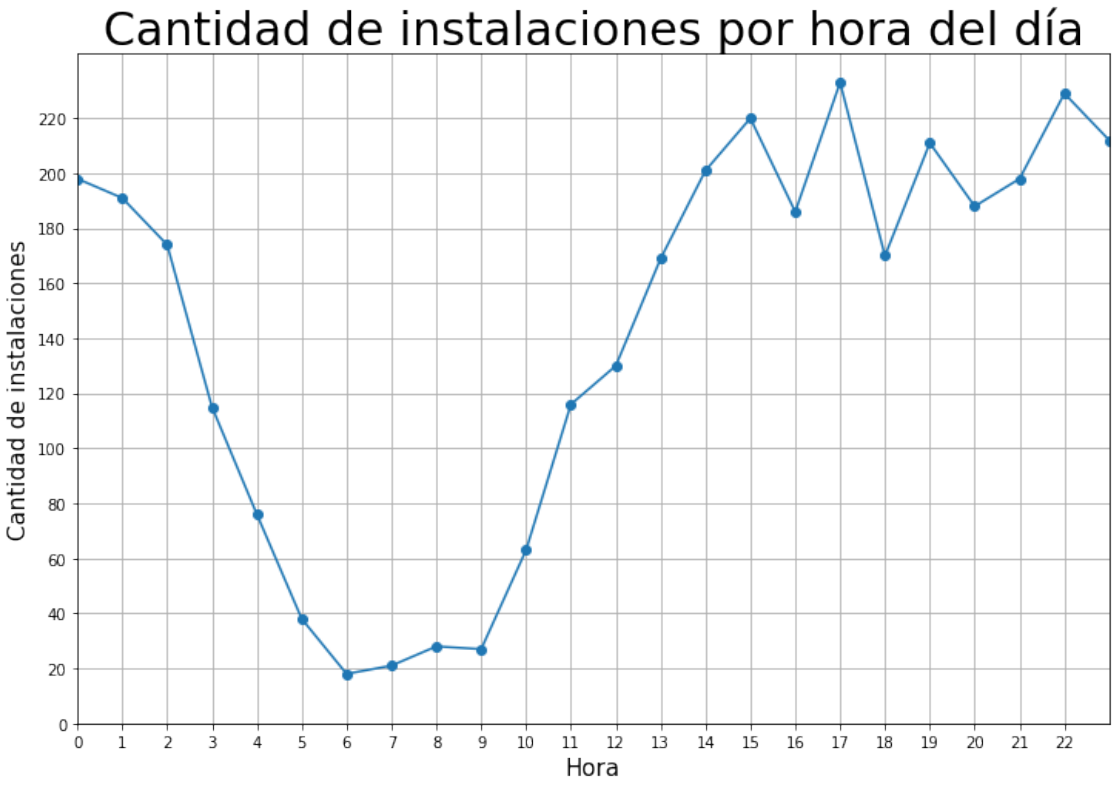

Figure 11: Instalaciones por hora del día

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.

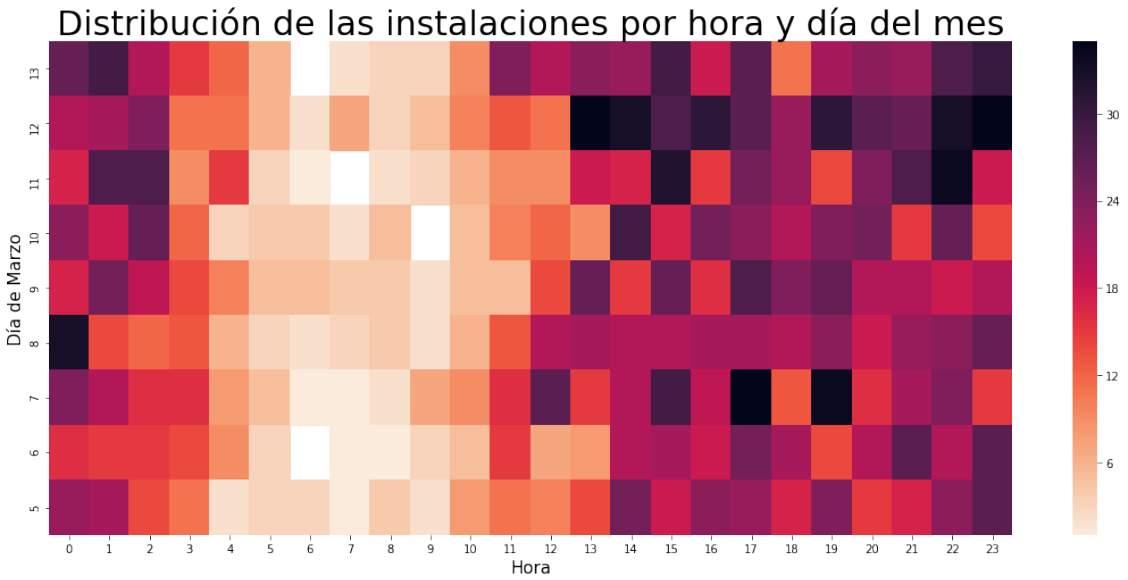

Figure 12: Instalaciones por fecha y hora

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.

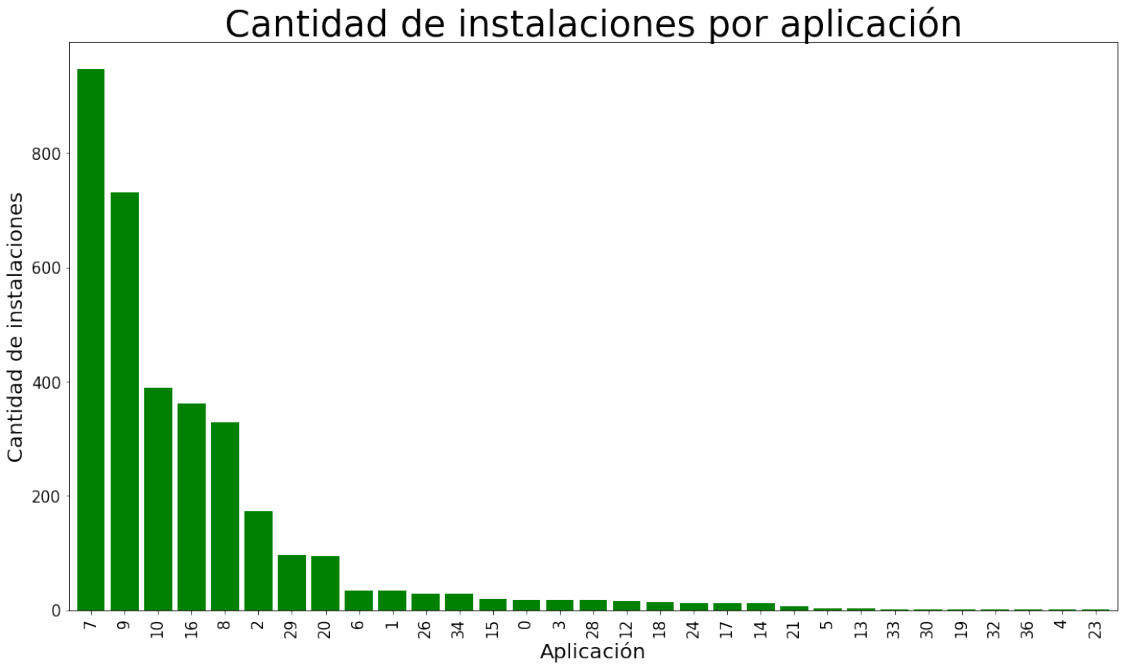

Figure 13: Instalaciones por aplicación

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*.

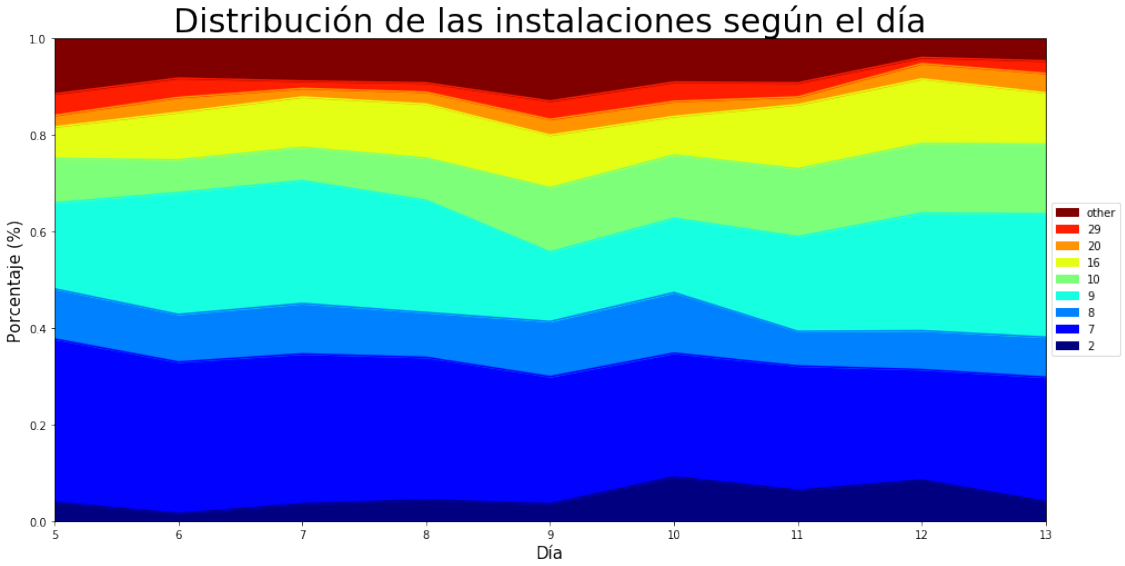

Figure 14: Incidencia de las aplicaciones según el día

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.

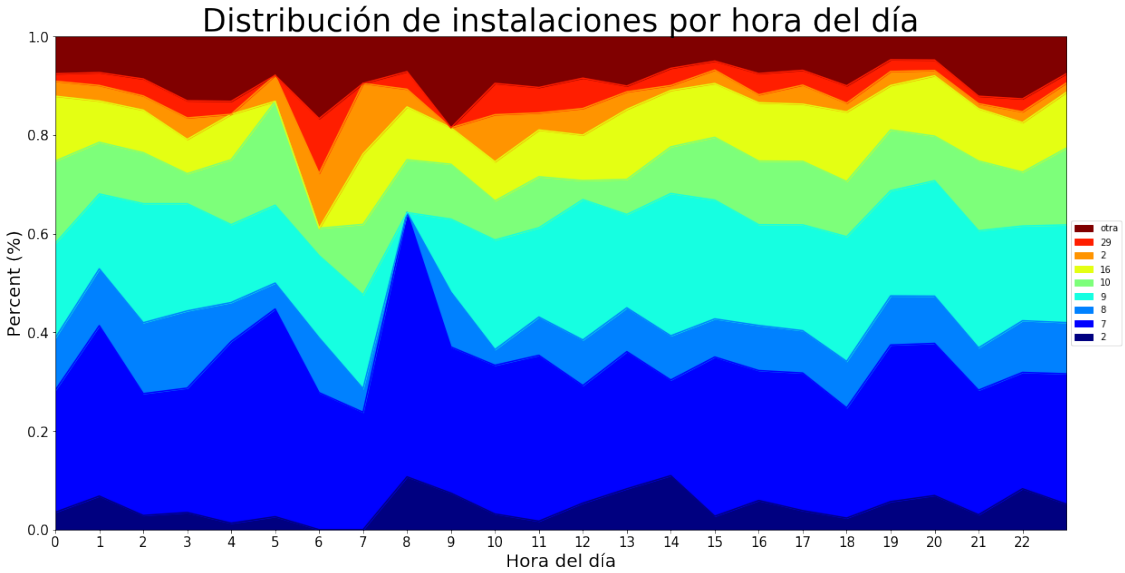

Figure 15: Incidencia de las aplicaciones según la hora del día

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.

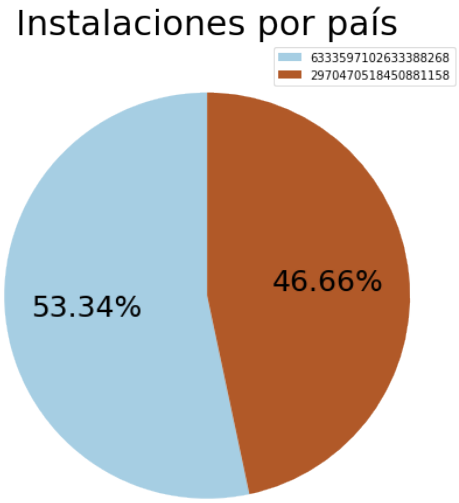

Figure 16: País de origen de las instalaciones

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.

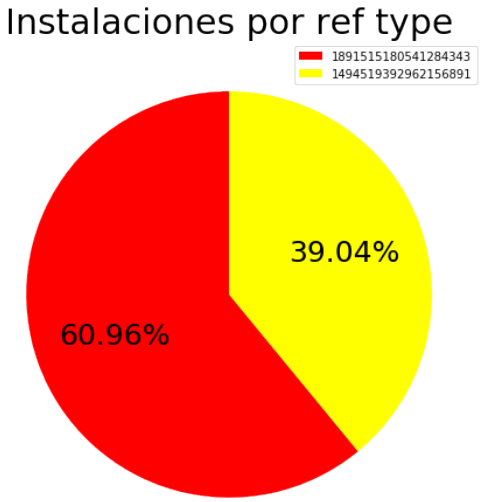

Figure 17: Tipo de referencia de las instalaciones

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.

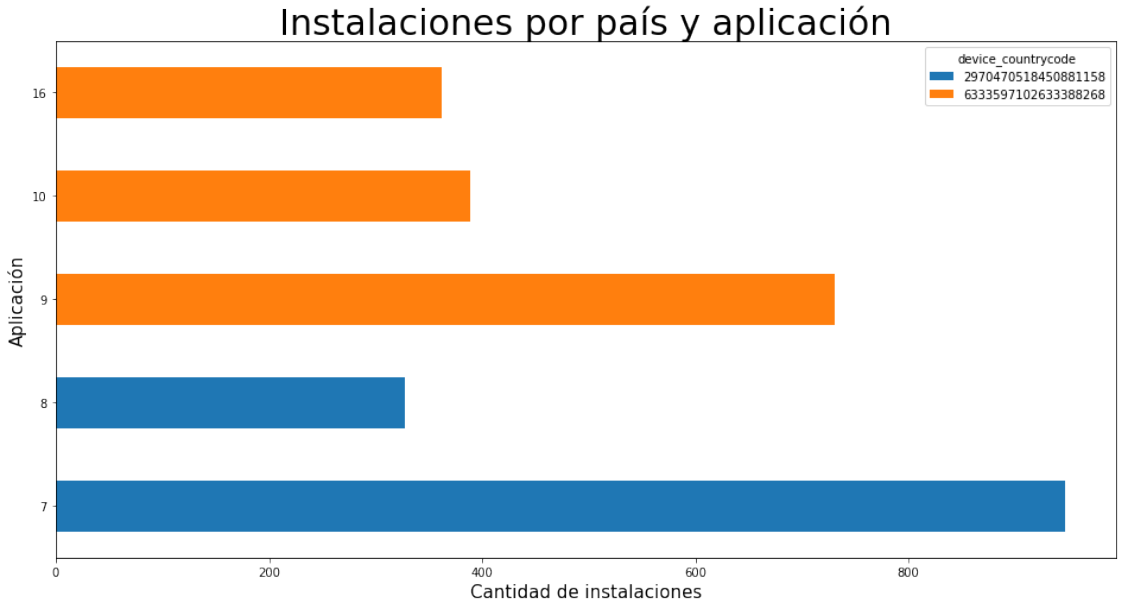

Figure 18: Distribución de países para las top 5 aplicaciones en instalaciones

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

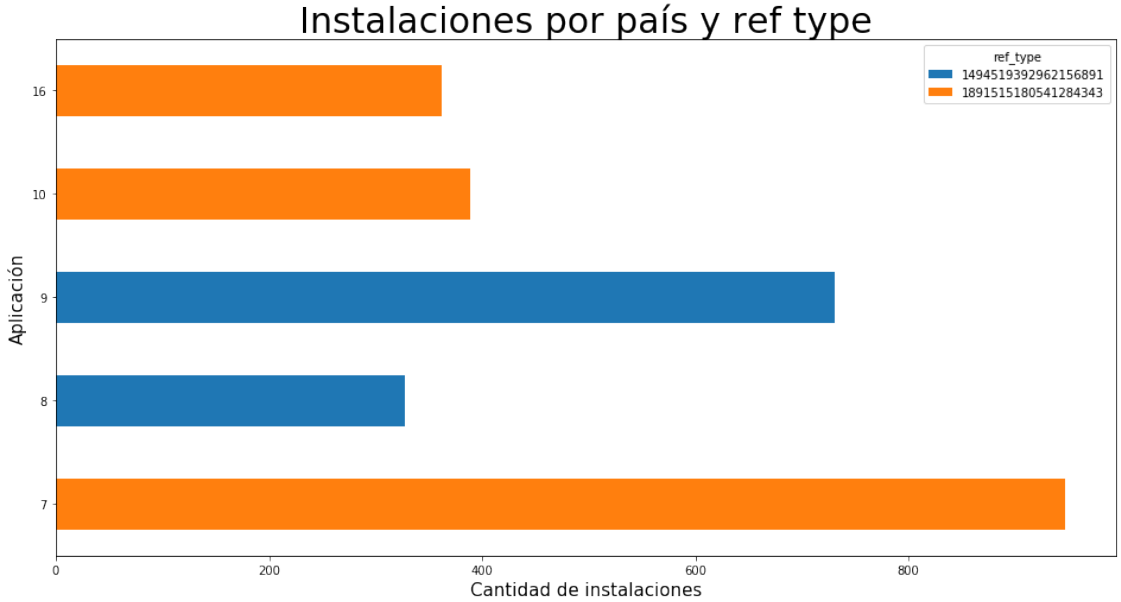

Figure 19: Distribución de ref types para las top 5 aplicaciones en instalaciones

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.

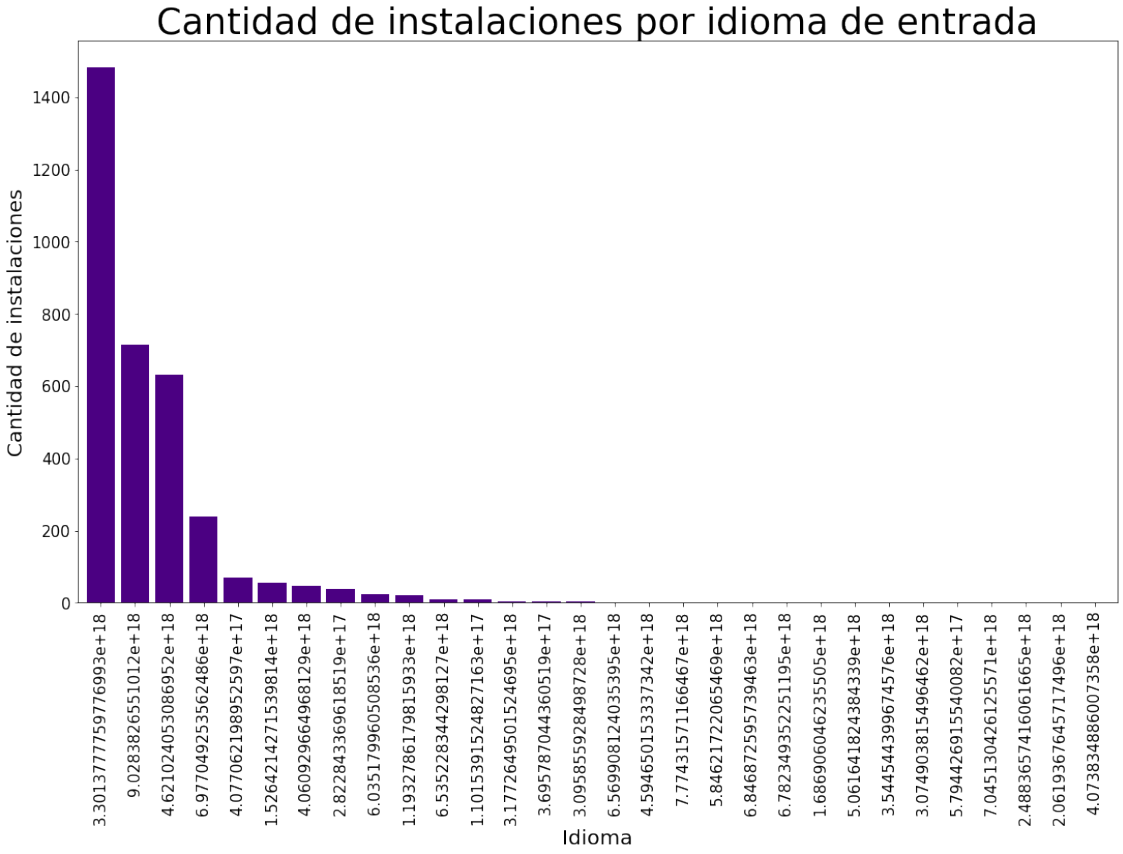

Figure 20: Idioma de entrada con mayor número de instalaciones

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 applicaciones más instaladas.

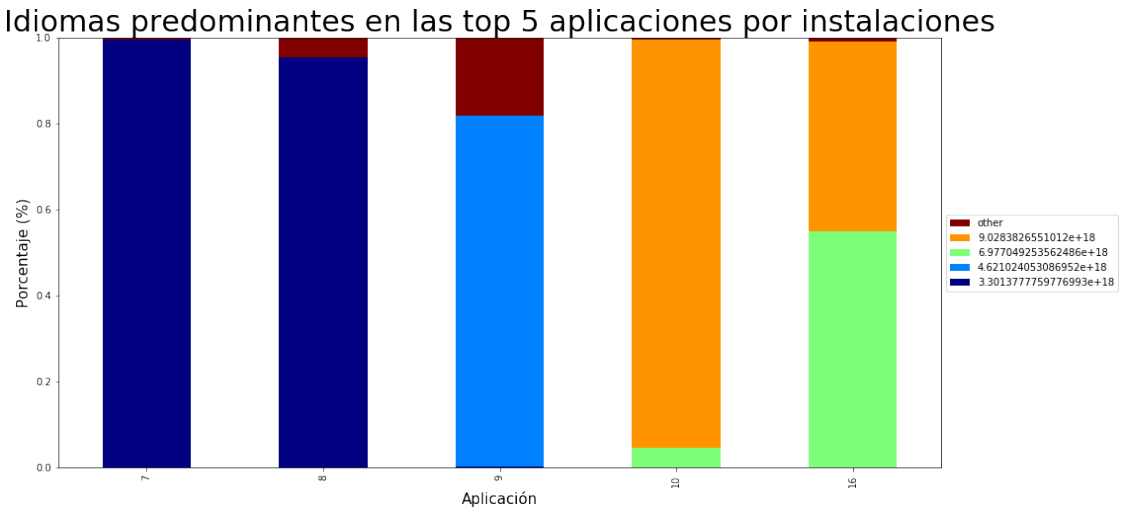

Figure 21: Idioma de entrada de los dispositivos para las 5 aplicaciones líderes en instalaciones

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 .....

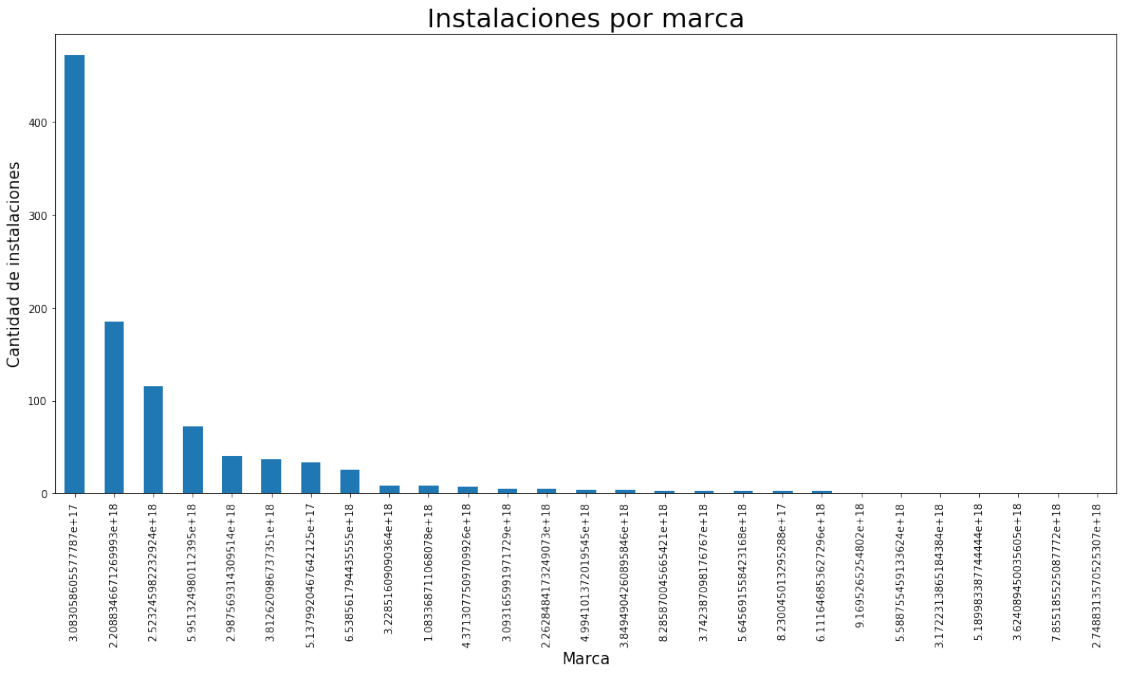

Figure 22: Marcas con más instalaciones

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.

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.

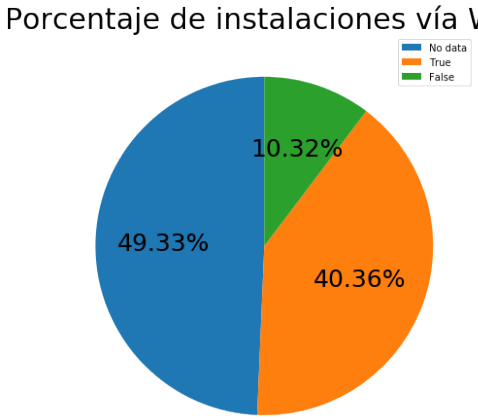

Figure 23: Instalaciones por conexión wifi.

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.

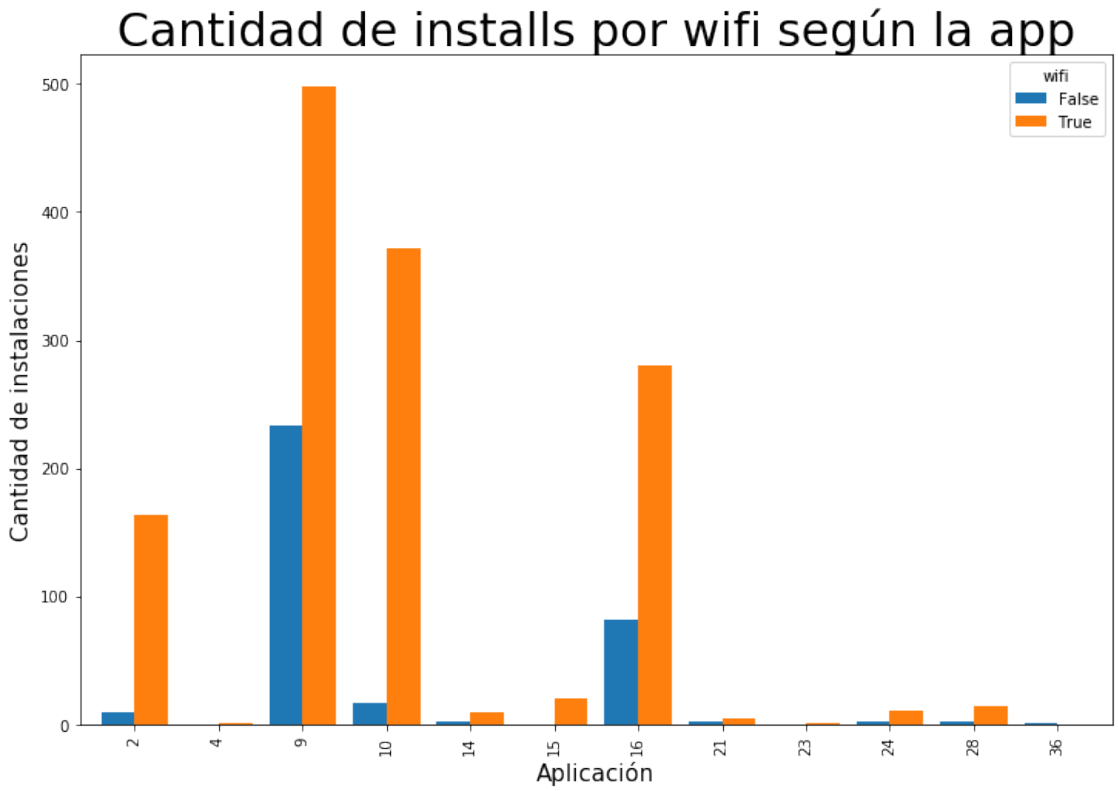

Figure 24: Instalaciones por conexión wifi según la aplicación.

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 gastarle 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.

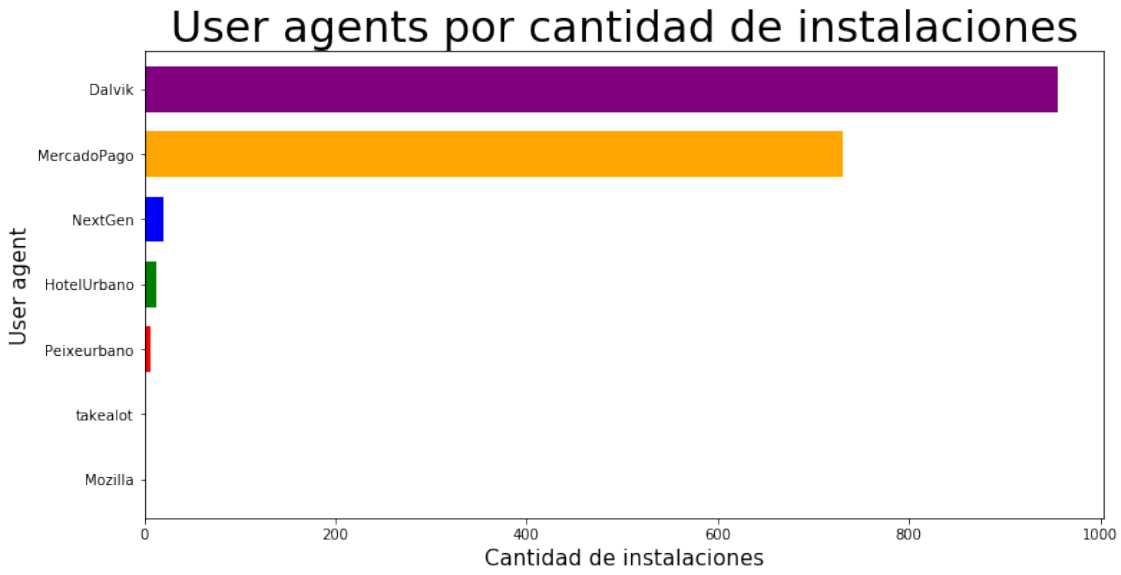

Figure 25: User agents por instalaciones

Resulta evidente que los dominadores absolutos de la categoría son Dalvik y MercadoPago, con agentes casi sin presencia, como Mozilla y takealot, que cuentan sólamente 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, le fue atribuida a una empresa de la competencia.

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.

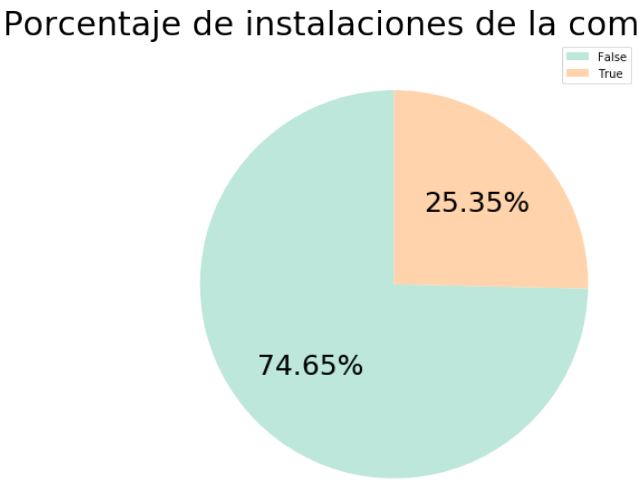

Figure 26: Porcentaje de instalaciones atribuidas a la competencia

Los datos muestran un *ratio* de algo más de 1 de cada 4 instalaciones atribuidas a empresas rivales.

Si lo trasladamos nuevamente a las 5 aplicaciones más instaladas.

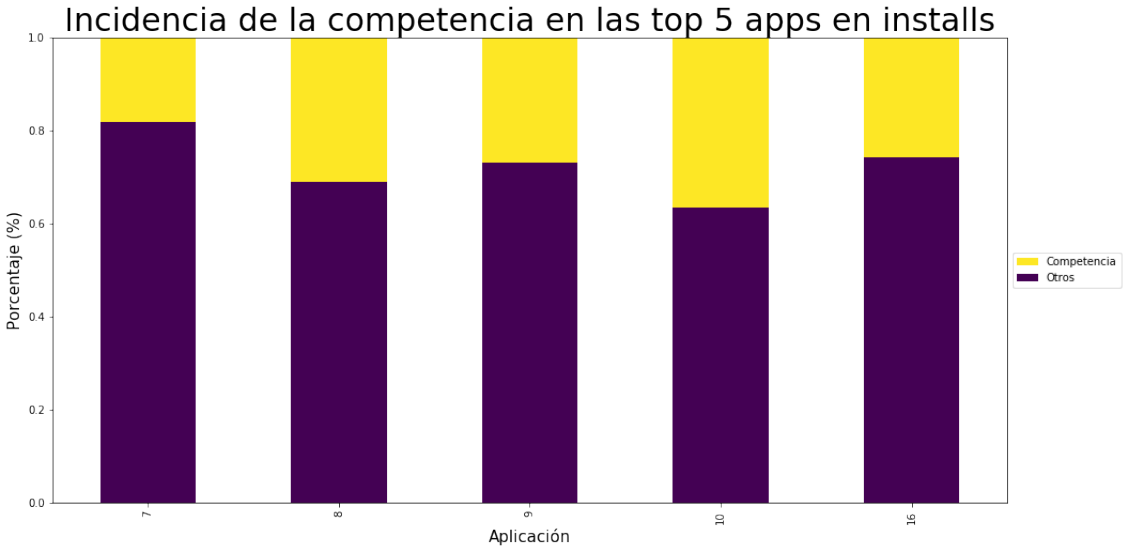

Figure 27: Porcentaje de instalaciones atribuidas a la competencia para las top 5 apps en descargas

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 le pertenecen a la competencia. 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 las empresas rivales tienen menos presencia.

3 Análisis de archivos en conjunto

4 Conclusion