UNIVERSIDAD DE BUENOS AIRES

Facultad de Ingeniería

75.06/95.58 Organización de Datos

1°C 2019 - Trabajo Práctico 1

Grupo N° 8

Nombre	Padrón	Mail
Beroch, Santiago	101135	sberoch@gmail.com
Giordano, Franco	100608	francogior98@gmail.com
Nitz, Ignacio	100710	nachonitz@gmail.com
Rial Brandariz, Lucas	100778	lucasrialb@gmail.com

Índice

Introducción

En el siguiente trabajo se presenta un análisis a los datos brindados por la empresa Jampp, plataforma que ayuda a anunciantes a promocionar sus aplicaciones y a recuperar usuarios inactivos.

Se analizaron los siguientes datos transaccionales[1]:

- Las subastas en las que Jampp oferta por los espacios publicitarios en los dispositivos.
- Los clicks de los usuarios en las publicidades mostradas.
- Las instalaciones de las aplicaciones por parte de los usuarios.
- Eventos dentro de las aplicaciones.

Inicialmente se analizaron cada uno de los ítems nombrados por separado, de forma tal que se pueda explorar la forma de los datos y tomar decisiones tales como eliminar columnas que no se necesitaran para el análisis y también saber qué preguntas se pueden hacer con los datos.

Luego de la etapa de filtrado se realizaron algunos análisis, también de forma separada, de cada ítem, para obtener algunas visualizaciones interesantes como en qué lugar de la pantalla son más frecuentes los clicks en la publicidad, el dia y hora de mayor actividad en las subastas, o que tipo de instalación es más frecuente.

Y por último, y tal vez de donde se obtuvieron las conclusiones más interesantes, se analizó a los datos en conjunto. De esta forma se puede extraer conclusiones en el total del ciclo de vida de la transacción.

El análisis se llevó a cabo en la plataforma Jupyter Lab, utilizando el lenguaje de programación Python 3. Las librerías mayormente utilizadas fueron: Pandas, Matplotlib, Seaborn, Numpy y Pywaffle.

El trabajo colaborativo se benefició por el uso de un repositorio git donde se guardaron las versiones preliminares de los notebooks que concluyeron en la forma final que se puede consultar en:

https://github.com/sberoch/Datos.git

[1] Son datos transaccionales aquellos generados debido a la participación en el ambiente de Real-Time Bidding (RTB). Son muy voluminosos y de distribución de rápido cambio.

1 Auctions (subastas)

1.1 Análisis de los datos

Las subastas que se tratan en el siguiente apartado son las que ocurren cada vez que se genera un espacio de publicidad en un dispositivo. Para elegir cual de todas las publicidades posibles se va a mostrar, se realiza una subasta donde se queda con el espacio el mejor postor. Es importante aclarar que todo este proceso se realiza en fracciones de segundos, por lo que la velocidad es un factor determinante.

Se realizó un análisis preliminar de los datos y se obtuvieron las siguientes conclusiones:

- Contamos con aproximadamente 19.500.000 subastas
- La columna que indicaría el tipo de auction (auction_type_id) no tiene valores por lo que no se tendrá en cuenta de aquí en más.
- Hay 2 plataformas posibles, una de ellas tiene 4 veces más auctions que la otra.
- Todas las auctions provienen del mismo país.
- El campo ref_type_id está directamente asociado al campo platform, por lo que no aporta información adicional (platform 1 = ref_type 1 y platform 2 = ref_type 7).

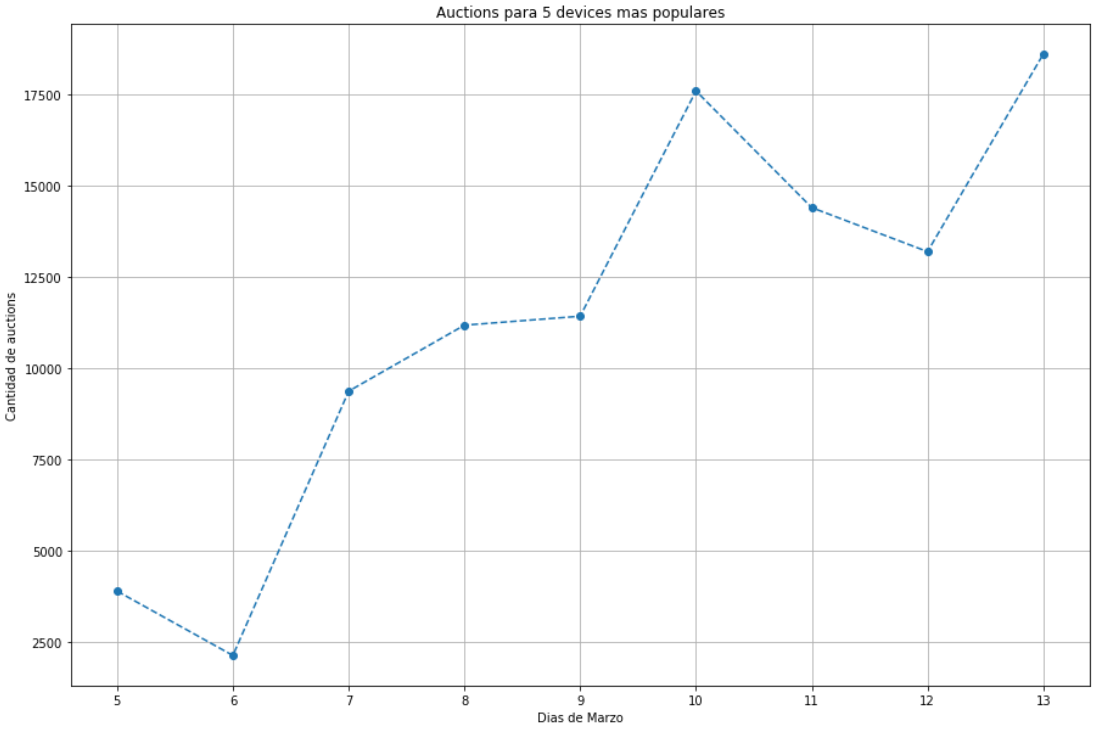
1.2 Los 5 usuarios con más participaciones

Se obtuvieron los 5 usuarios más frecuentes, los cuales son:

Posición	Usuario	Cantidad de auctions
1	633139769114048761	27762
2	7202276635029175071	23055
3	7298861376107043945	18188
4	7298861376107043945	16400
5	5376802567578262905	16367

Para cada uno de ellos se calculó la cantidad de auctions por día y se los representó en un mismo gráfico.

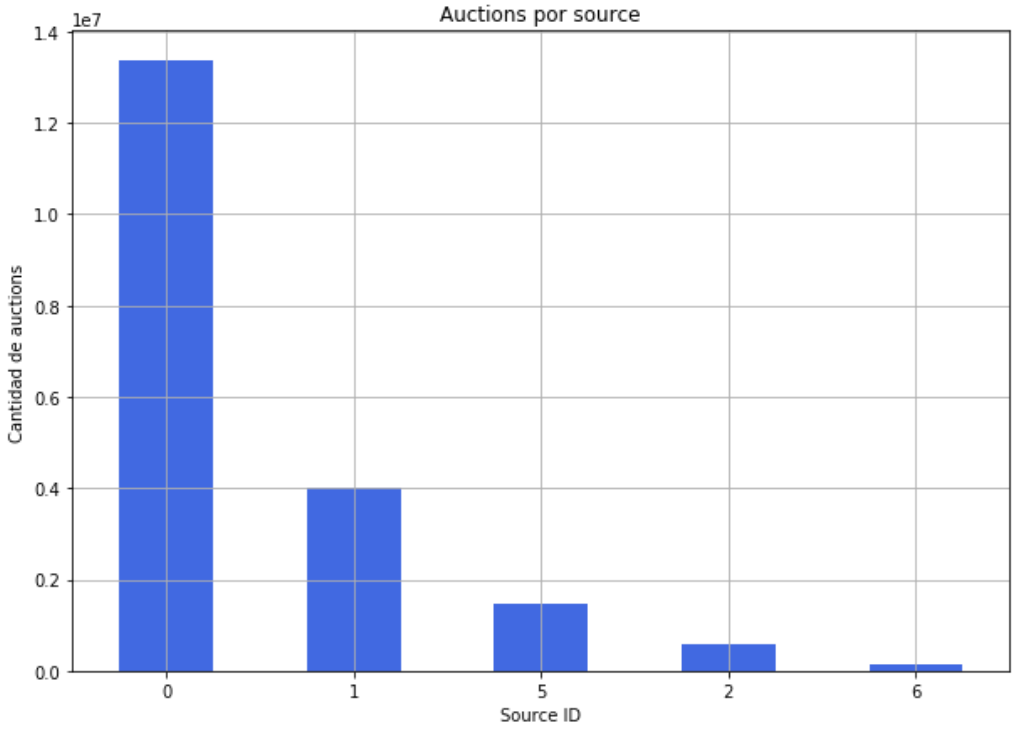
A continuación se realizó una suma de los valores diarios de dichos usuarios y se los graficó de la siguiente manera.



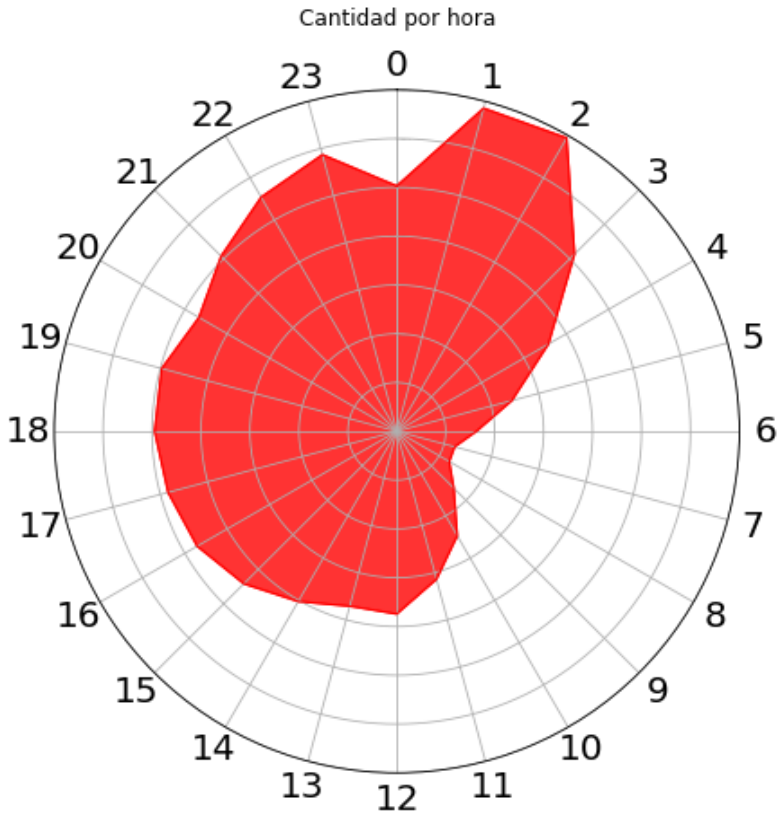
Como se puede ver, la cantidad de auctions en marzo del 2019 fue aumentando conforme pasaban los días, finalizando con valores muy superiores a los del comienzo (aproximadamente un 700% más)

1.3 Source (fuente)

Un source o fuente es el lugar del que proviene la subasta (también llamado exchange). Los sources son los que, de alguna manera, moderan la subasta.
Al hacer un análisis de los distintos sources contenidos en los datos se pudo observar que una gran cantidad provienen del source 0, repartiéndose las subastas restantes de manera más equitativa entre las demás fuentes.



1.4 Ocurrencias según hora del dia



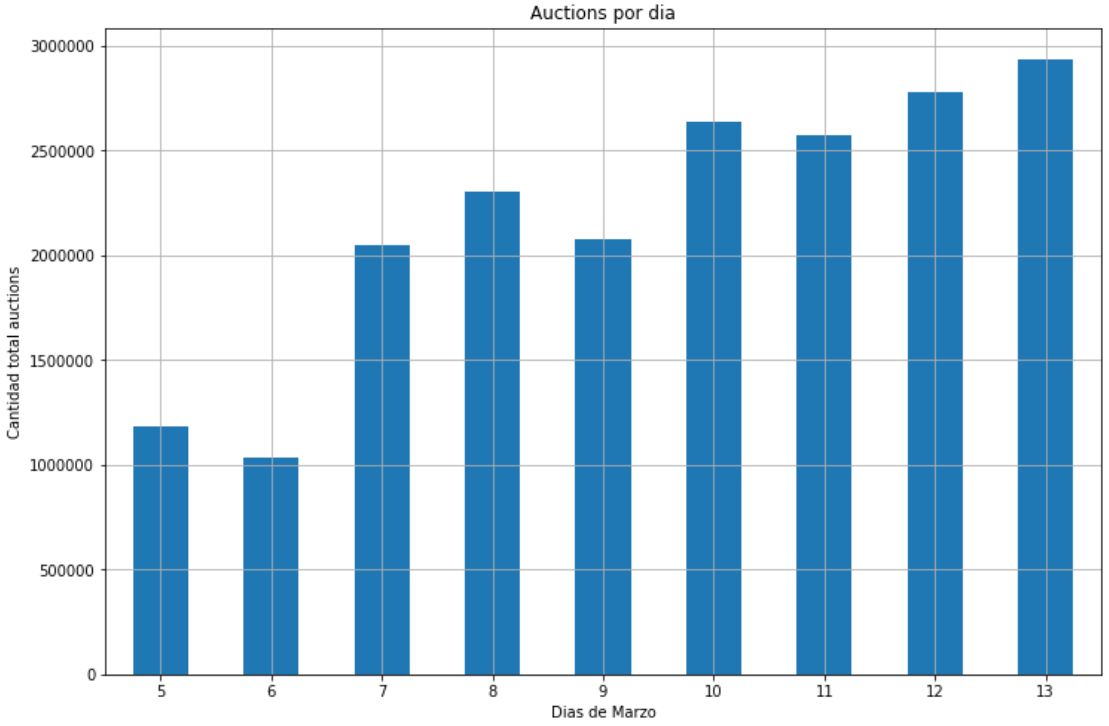
Escala radial: Desde 200000 hasta 1400000, paso de 200000.

Como se puede apreciar en el gráfico anterior, el mayor flujo de subastas ocurre entre la 1 y las 2 AM. El momento de menor cantidad se puede ver con igual facilidad, y este se encuentra entre las 9 y las 6.

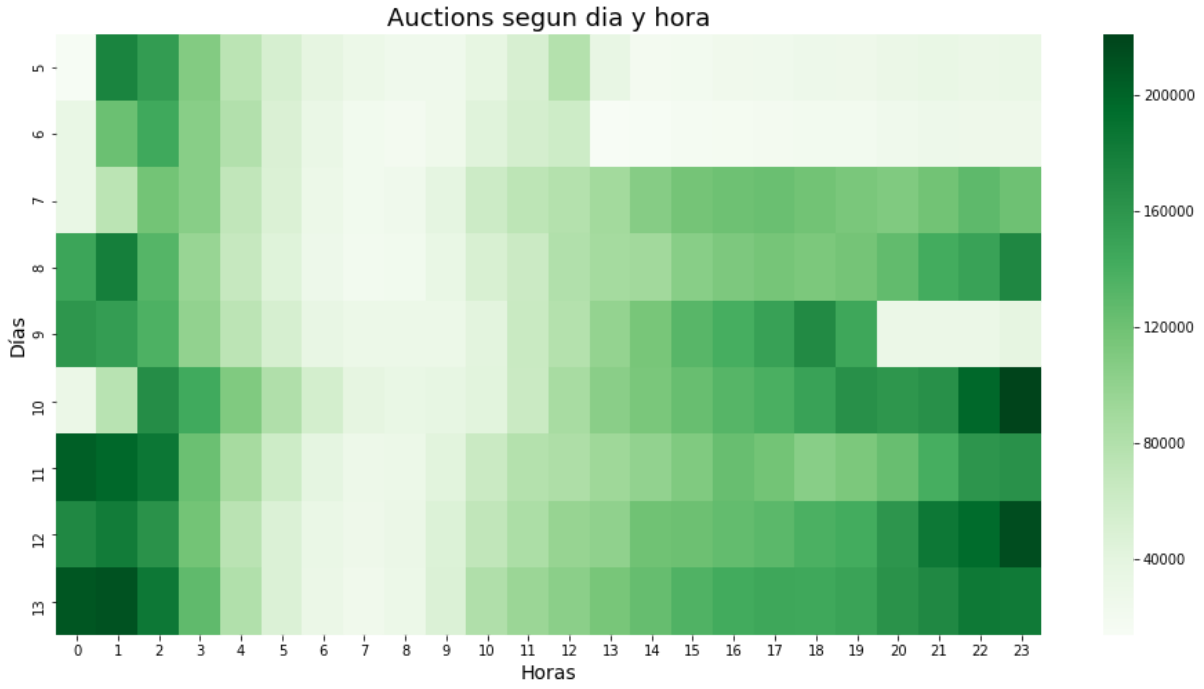
1.5 Ocurrencias por día en marzo

Al analizar la cantidad de subastas por día se intentaba encontrar algún evento importante que hubiera generado un cambio en los valores, ya sea un pico o un mínimo en los valores. Sin embargo, se pudo observar que las cantidades fueron aumentando constantemente en el tiempo, casi de manera lineal si se realizan algunas aproximaciones. Al tener los valores más altos y bajos en los extremos no se puede extrapolar de manera confiable los datos a otros intervalos de tiempo, debido a que no se sabe la causa de tal aumento en este corto intervalo de tiempo.

A continuación se anexa el gráfico que representa el comportamiento anteriormente enunciado.



1.6 Cantidad de auctions por día y hora



El gráfico anterior permite ver que casi todos los días se “comportan” de la misma manera en relación a las horas, por lo que el apartado 1.4 gana importancia y su conclusión se podría usar para cualquier día en particular.

Es necesario destacar la anomalía que se dió entre las 13 y 23 de los días 5 y 6. Sus valores son demasiado bajos (potencialmente nulos), lo que no se condice con lo que ocurre en el resto de los días.
La conclusión a la que se puede llegar es que se trató de algún tipo de error en el almacenamiento de los datos, o que la forma de reducir el tamaño de los datos originales filtró todos los que ocurrieron en ese intervalo de tiempo.

2 Clicks

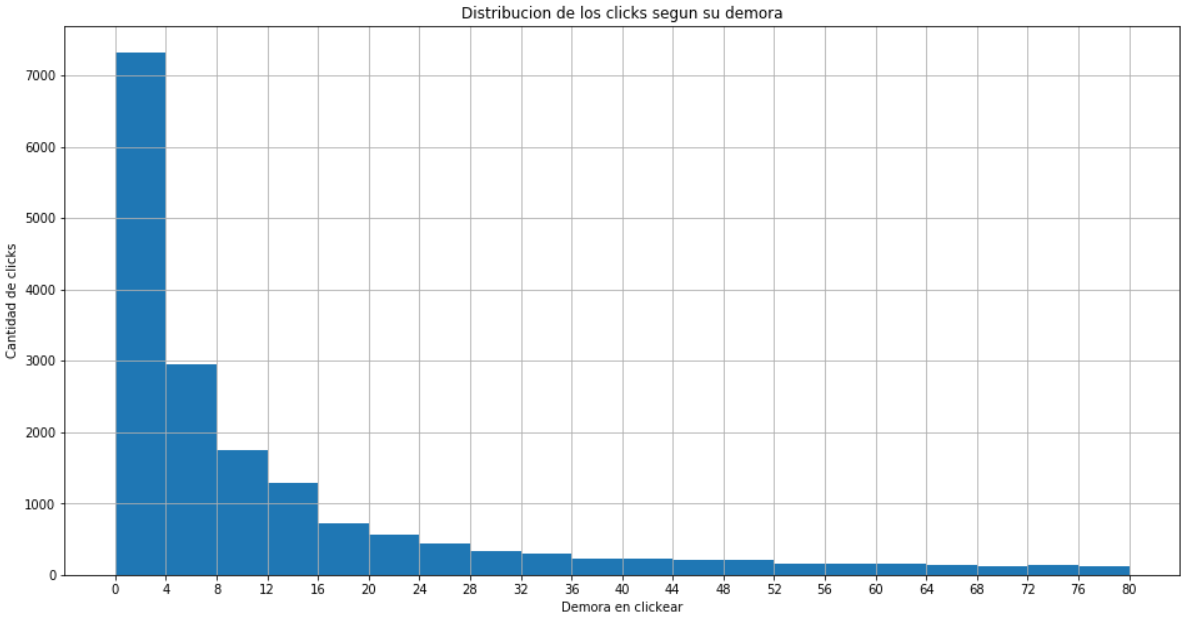
2.1 Análisis de los datos

Los clicks registrados en el siguiente set de datos son los que han sido producidos dentro del ámbito de la publicidad mostrada a un usuario.

Luego de inspeccionar los datos se pudo concluir:

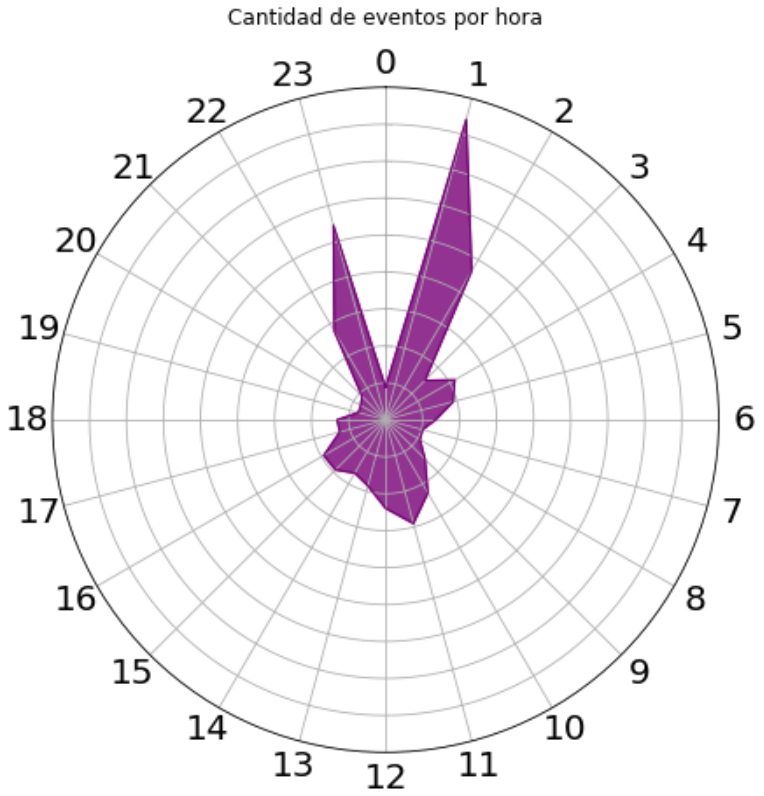
- Cada registro cuenta con 23 columnas asociadas, donde se cuenta con 26351 registros.
- Se encuentra que las columnas ‘action_id’, ‘agent_device’ y ‘brand’ presentan un alto porcentaje de nulos (100%, 87%, 76% respectivamente). Por ello serán descartadas en el análisis.
- La columna ‘wifi_connection’ está enteramente constituida de valores ‘False’, por lo que no aporta información extra. Será descartada.

2.2 Distribucion de los clicks, en función del tiempo hasta clickear



Se observa una clara tendencia entre los usuarios a rápidamente clickear el anuncio. A mayor tiempo hasta clickear, menor cantidad de ocurrencias. Curiosamente, este decrecimiento no presenta un comportamiento lineal, sino más bien uno exponencial.

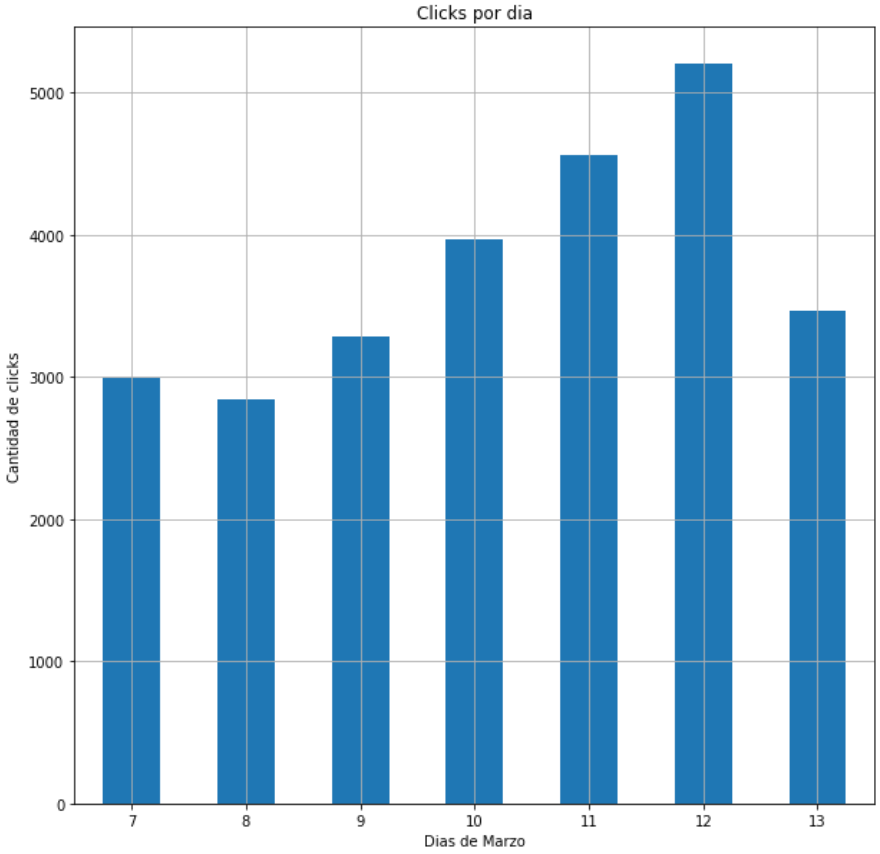
2.3 Ocurrencias según hora del dia



Escala radial: Desde 500 hasta 4500, paso de 500.

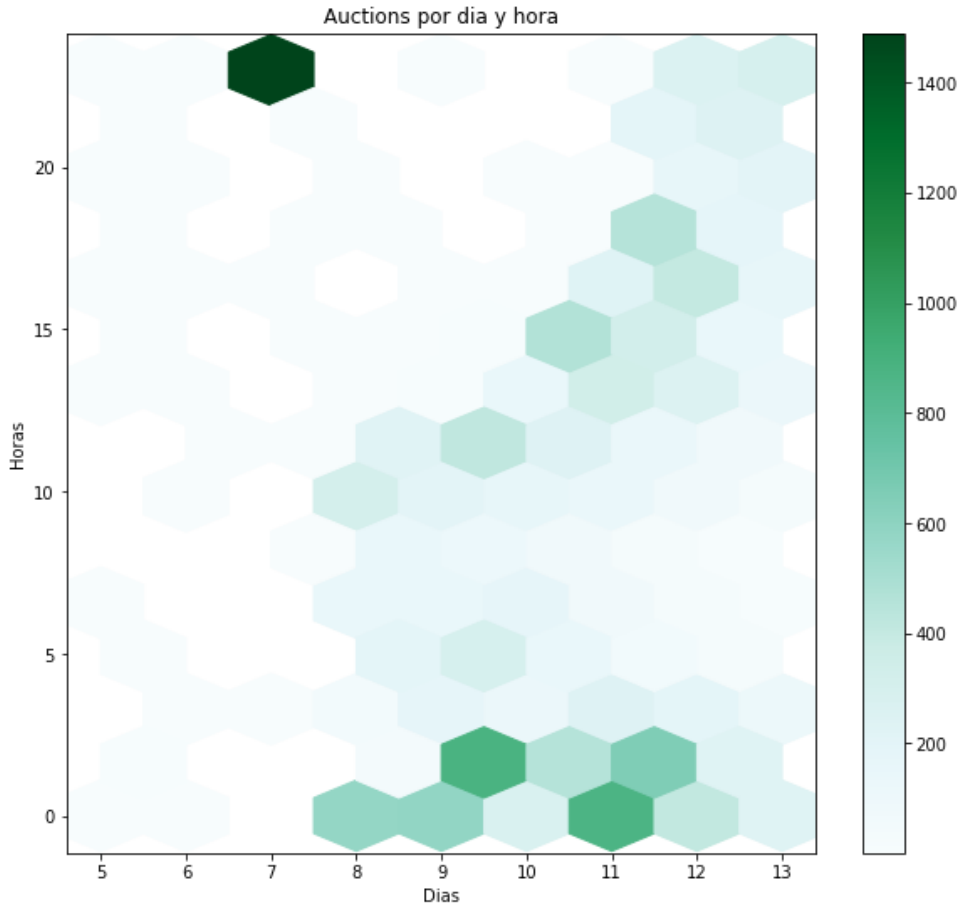
Contando ocurrencias según la hora asociada a cada registro, se halla que las horas nocturnas son más ‘efectivas’ (mayor cantidad de clicks). Cabe destacar que entre estas horas, 10pm y 2am, se halla un mínimo a las 0hs. Se podría atribuir este comportamiento a un funcionamiento interno del sistema, ya que no parece un actitud natural de los usuarios.
Por otro lado, se encuentra otra tendencia de crecimiento alrededor del mediodía.

2.4 Ocurrencias por día en marzo



Con los datos recibidos, se encontró que las ocurrencias registradas se distribuyen únicamente entre los días 5 y 13 de marzo 2019. Además, los días 5 y 6 presentaron cantidades de ocurrencias anormalmente bajas, por lo que fueron descartadas en el análisis.
Aclarado esto, se halla que los días 11 y 12 fueron los mas fuertes, mientras que los días 7 y 8 presentan menos ocurrencias.

2.5 Cantidad de clicks por día y hora



Al buscar la relación de ocurrencias según el día y la hora, se destaca la concentración de datos en el día 7, hora 22 aproximadamente. Resulta interesante ya que si bien la mayoría de ocurrencias se concentran en la hora ~22 (gráfico 2.3), no ocurre así para el dia 7 (gráfico 2.5). Se podría llegar a la conclusión que la gran aglomeración de clicks para el día 7 a las 24 se trató de un error interno del sistema de almacenamiento o bien del procesamiento de los datos al reducir las dimensiones del set original al acotado con el que se trabaja en este caso.
Por otro lado, se encuentra una esperada correlación entre los días 10 y 12, y las horas 1am y 2am. En otras palabras, muchos clicks distribuidos entre estas franjas.

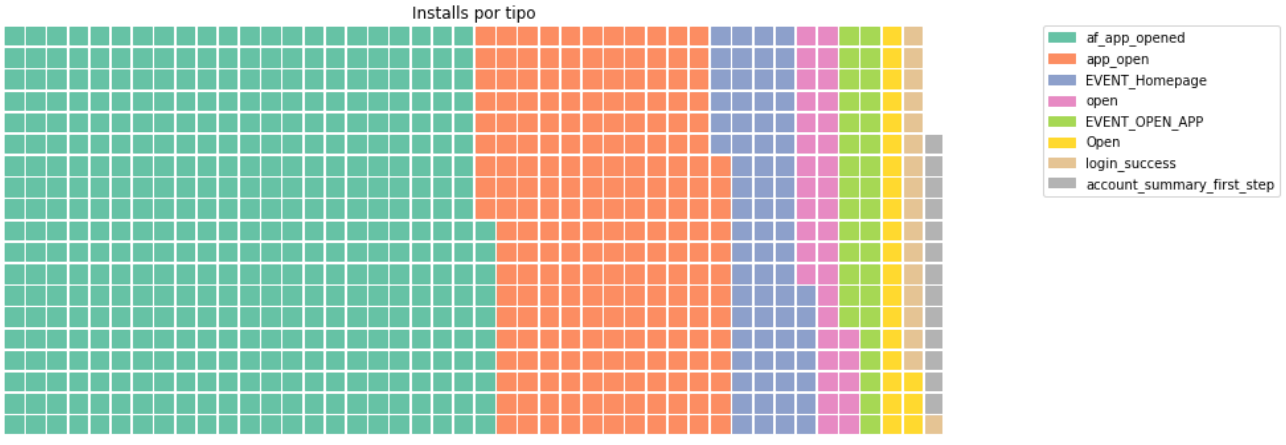
3 Installs (instalaciones)

3.1 Análisis de los datos

Una instalación, en este ámbito, se entiende por la descarga de una aplicación desde la tienda con la que cuente el dispositivo.
Se realizó un análisis inicial del set de datos y las conclusiones alcanzadas fueron las siguientes:

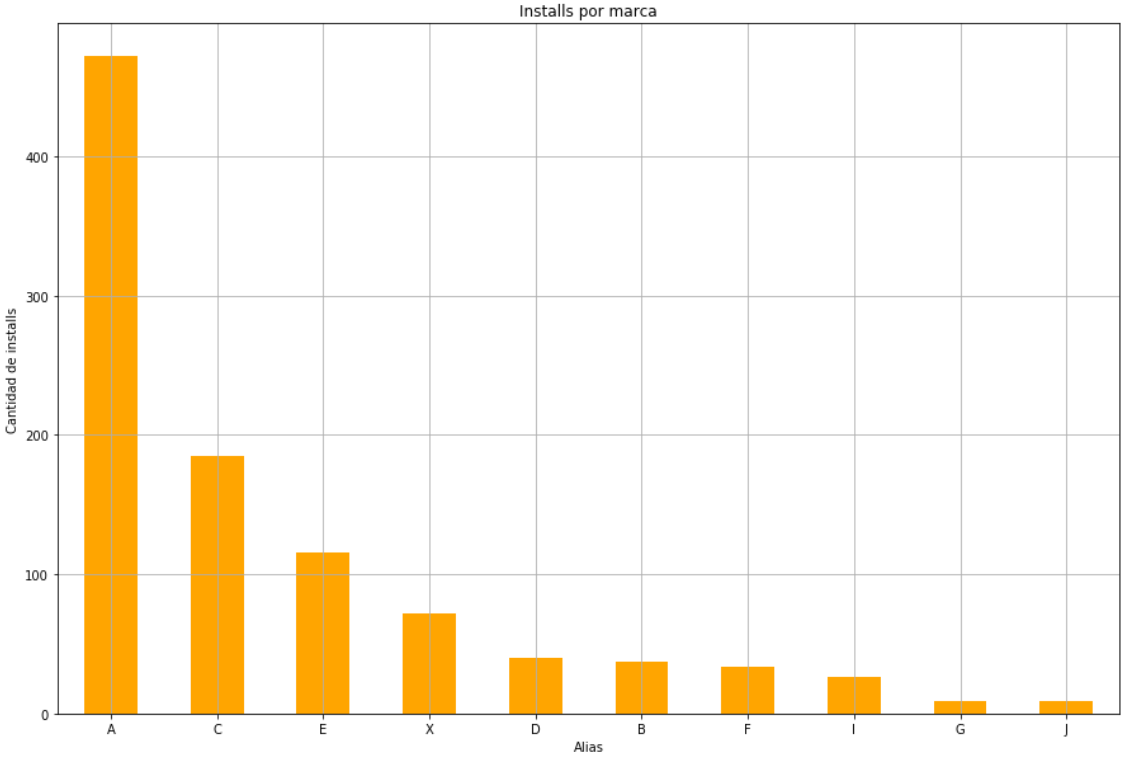
- Contamos con 3412 registros de instalaciones de 18 campos cada uno.
- Las columnas click_hash y trans_id tienen todos sus valores en null, por lo que no serán tenidas en cuenta.
- El campo attributed (que indica si la instalación fue atribuida a Jampp) se encuentra totalmente en False, y de esa manera no aporta información relevante.

3.2 Frecuencia de los distintos tipos de installs



Se graficó la distribución de instalaciones por tipo, donde un cuadrado representa una instalación. Para mantener el gráfico conciso se despreciaron los tipos con menos de diez instalaciones. Estudiando la representación se observa una clara mayoría en el ámbito de ‘abrir aplicacion’. Curiosamente, varios tipos comparten palabras claves similares (‘app_open’, ‘af_app_opened’, etc.) y todas estas resultan relacionarse con el evento ‘abrir’. Al no conocer los mecanismos de registro utilizados, no se puede asumir si se trata de tipos realmente iguales/equivalentes, o si son tipos completamente distintos.

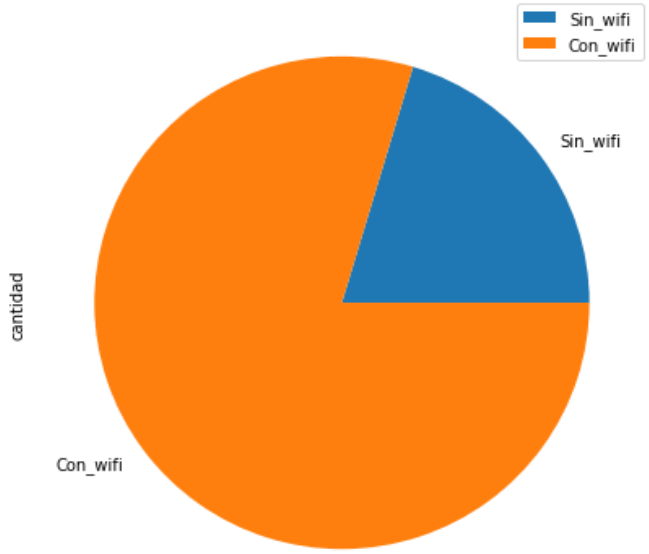
3.3 Diferencias entre las marcas



Alias	Valor	Alias	Valor
A	3.083059e+17	**F**	5.137992e+17
B	3.812621e+18	**G**	3.228516e+18
C	2.208835e+18	**X**	5.951325e+18
D	2.987569e+18	**I**	6.538562e+18
E	2.523246e+18	**J**	1.083369e+18

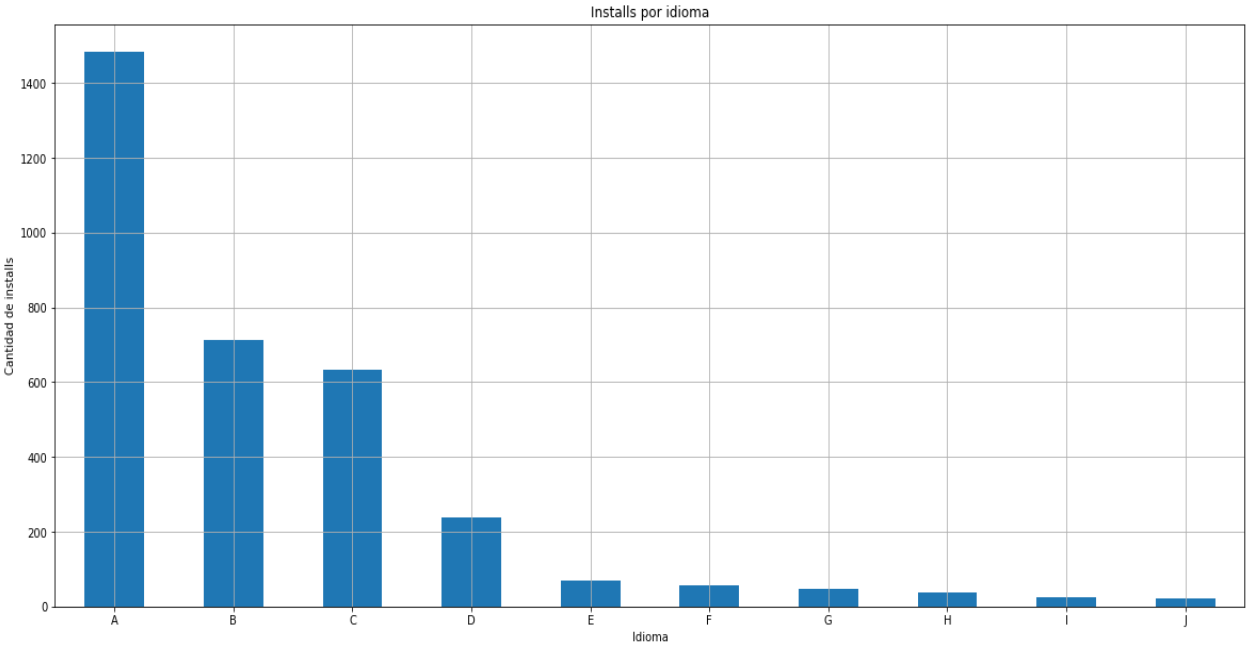
Al ser marcas hasheadas no puede realizarse un profundo análisis de la distribución del mercado. Lo que sí resulta interesante es el decrecimiento exponencial que evidencia el gráfico. Se encuentra que no son un par o un grupo de marcas las más influyentes, sino una en particular se destaca por sobre el resto (alias A).

3.4 Tipo de conexión (Wifi)



El análisis sugiere una conclusión un tanto esperada: la mayoría de los usuarios (~80%) opta por instalar mediante una conexión wifi.

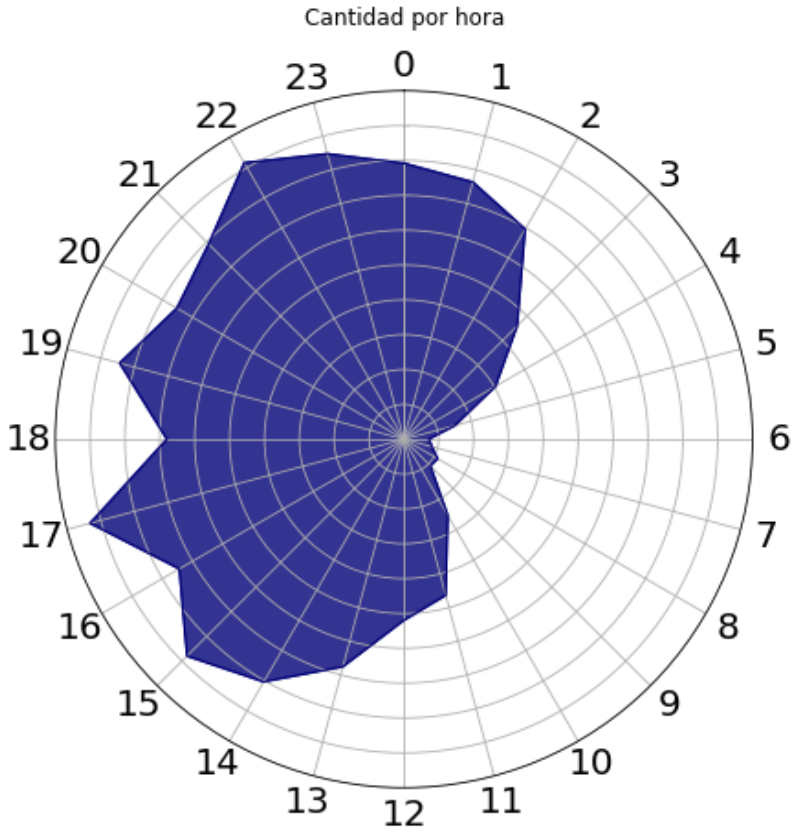
3.5 Clasificación por idioma



Alias	Valor	Alias	Valor
A	3.301378e+18	**F**	1.526421e+18
B	9.028383e+18	**G**	4.060930e+18
C	4.621024e+18	**H**	2.822843e+17
D	6.977049e+18	**I**	6.035180e+18
E	4.077062e+17	**J**	1.193279e+18

Nuevamente, al ser datos hasheados, no puede realizarse un análisis demográfico. Además, al encontrarse los datos sobre el país de origen también hasheados, no pueden cruzarse los registros y obtener conclusiones geográficas. De todas formas se encuentra que, si bien hay un idioma predominante, los siguientes dos idiomas más utilizados se encuentran a la par, lo cual no sigue la fuerte tendencia exponencial hallada en otros ámbitos (como en Installs por marca, Tiempo hasta click o Sources de auctions).

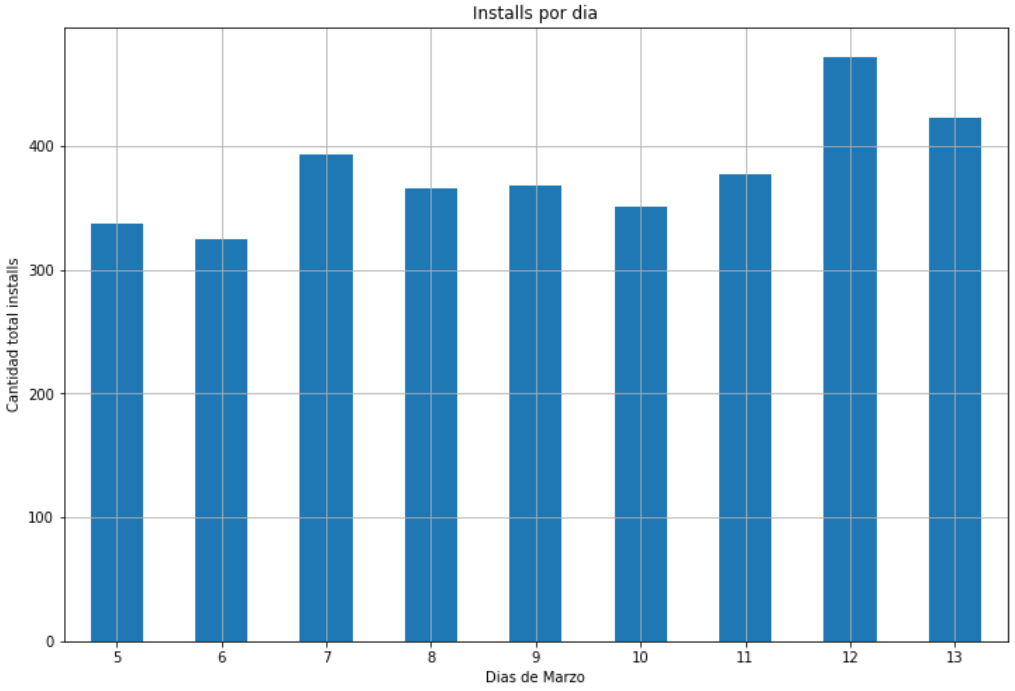
3.6 Ocurrencias según hora del día



Escala radial: Desde 25 hasta 250, paso de 25.

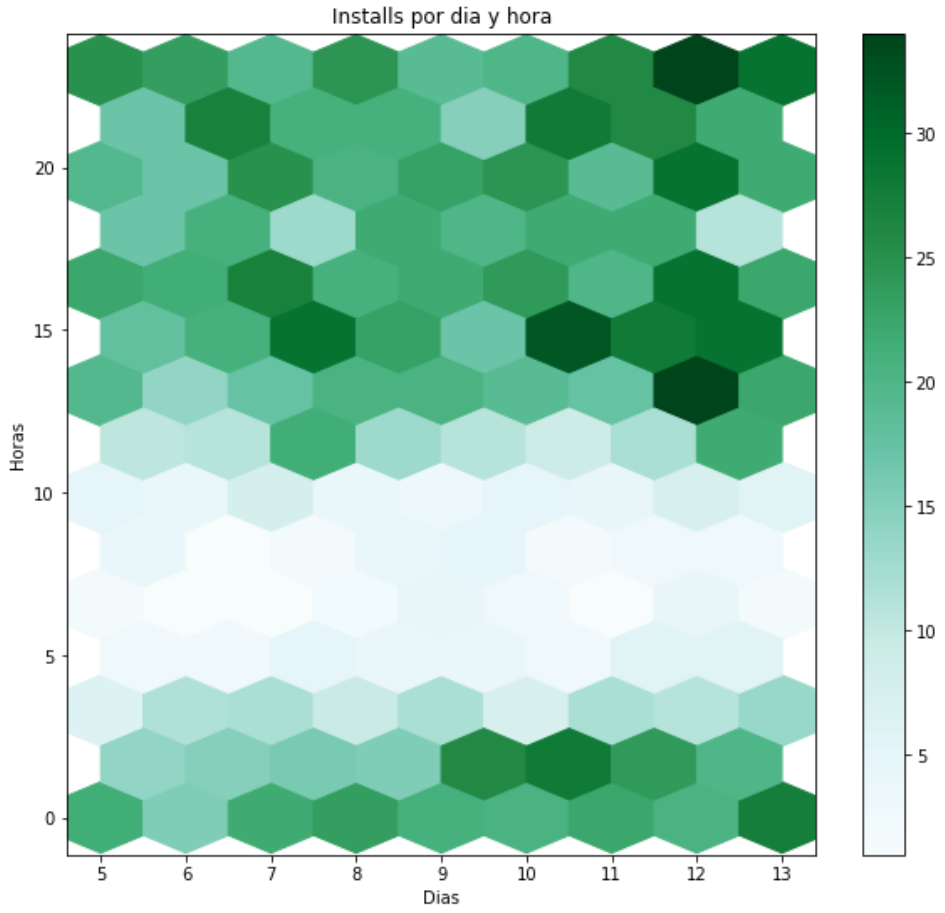
Más adelante al analizar la superposición de todos los datos se halla que el exhibido aquí es la tendencia global en toda la base de datos (en cantidad por hora). En otras palabras, se encuentran mínimos entre las 4am y 10am, y máximos en 15pm y 22pm.

3.7 Ocurrencias por día en marzo



Al ordenar las instalaciones por día del mes no se halla una tendencia muy marcada, más de que los últimos días presentan más apariciones que los primeros.

3.8 Cantidad de installs por día y hora



Al cruzar los datos de instalaciones en este heatmap resulta llamativa la franja de pocas ocurrencias entre las horas 5am y 10am. Es decir, si bien era esperado hallar pocas ocurrencias en estas horas (de gráfico 3.6), es interesante ver que el comportamiento *casi* constante según dia (de gráfico 3.7) se presenta aun analizando hora por hora.
Dicho de otra forma, se concluye que el día del mes influye *muy* poco en la cantidad de instalaciones registradas, al menos para esta base de datos.

4 Events (eventos)

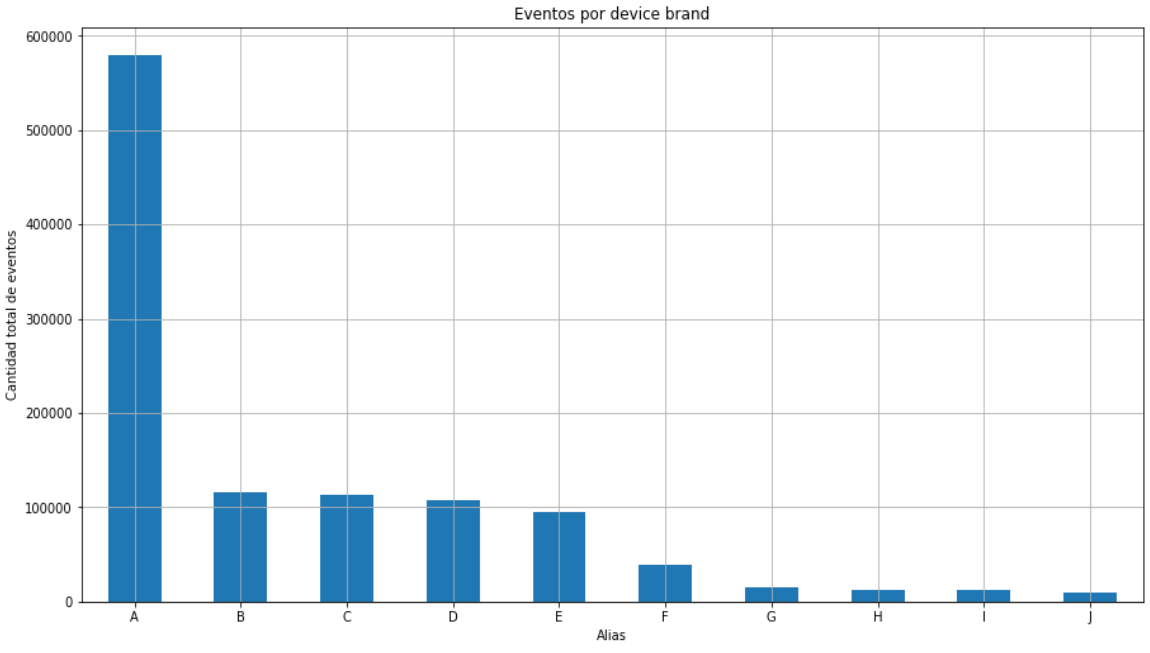
4.1 Análisis de los datos

Un event es cualquier tipo de acción tomada por el usuario dentro de la aplicación cliente de Jampp.
Para empezar a trabajar con estos datos se hizo un análisis inicial, donde se observó:

- trans_id no brindara información, es una columna con todos sus datos nulos.
- device_os_version, device_brand, device_city, user_agent, carrier, device_os, wifi y connection_type por su parte son columnas con un número muy elevado de datos nulos. Se podrán emplear para algún analisis particular pero no serán totalmente representativos de la muestra.

4.2 Diferencias del volumen total entre las marcas

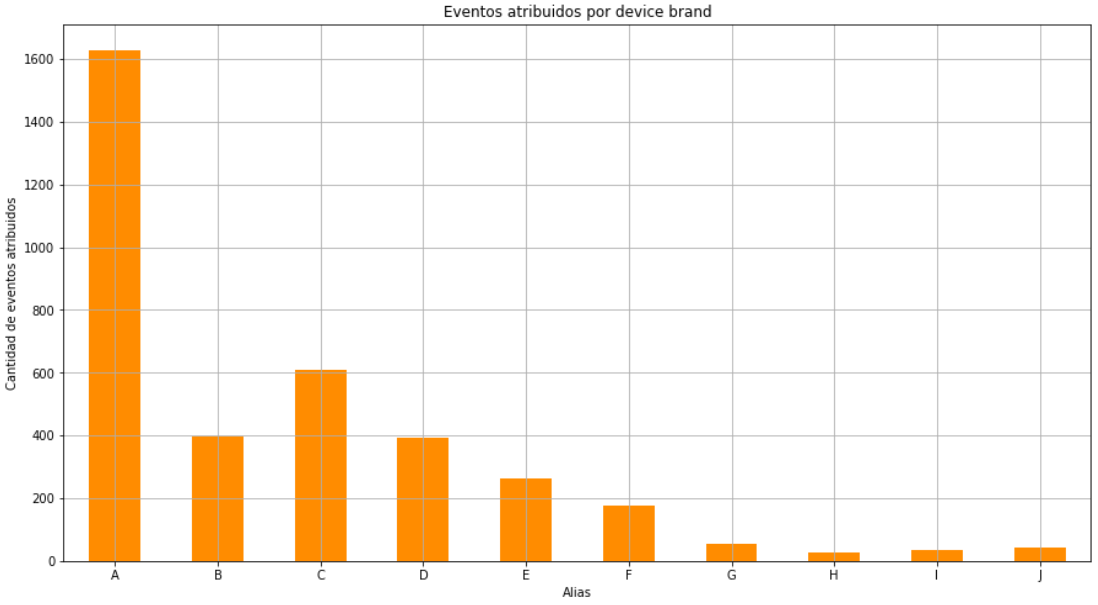
Se quiere responder a la pregunta de cómo responden la cantidad de eventos según la marca del dispositivo donde se muestra la publicidad. En el gráfico siguiente se puede ver que la mayoría de los eventos se concentran en el dispositivo “A” (ver tabla para ver el dispositivo correspondiente a cada alias), mientras que los siguientes 4 presentan aproximadamente 10000 eventos cada uno, y los últimos valores aún más pequeños. Podría concluirse que Jampp funciona mejor en dispositivos marca “A”, o que los usuarios de dispositivos marca “A” son más propensos a usar aplicaciones clientes de Jampp.



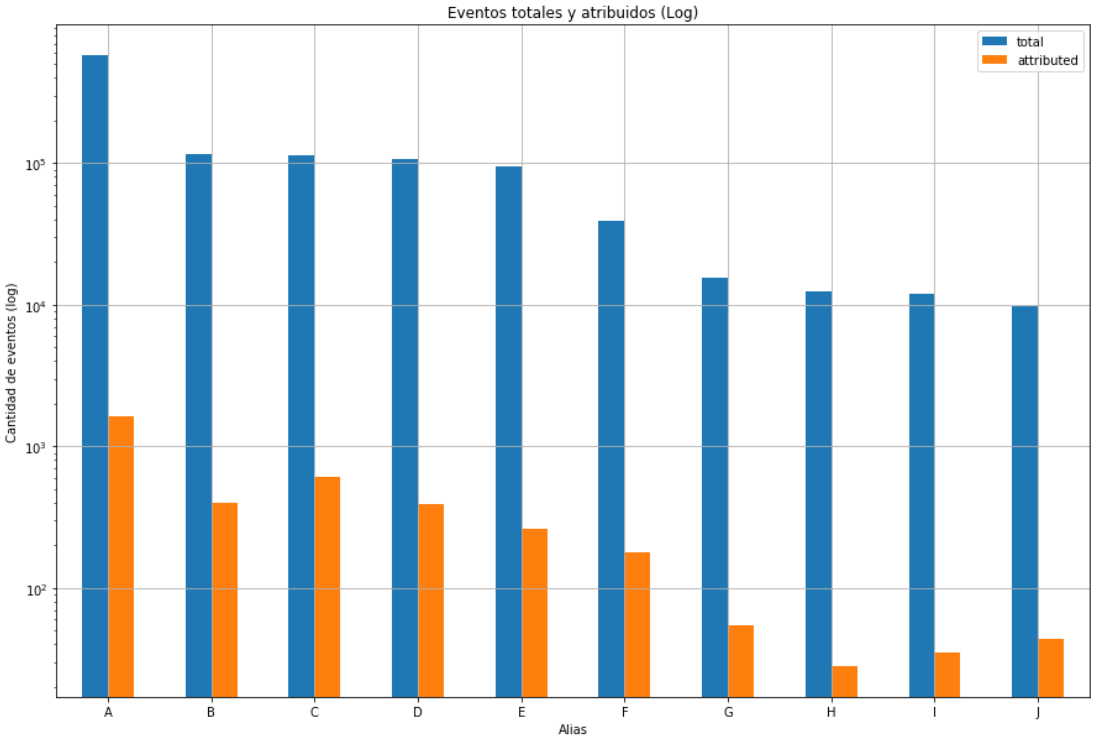
Alias	Valor	Alias	Valor
A	3.083059e+17	**F**	5.137992e+17
B	3.812621e+18	**G**	3.228516e+18
C	2.208835e+18	**H**	2.262848e+18
D	2.987569e+18	**I**	6.538562e+18
E	2.523246e+18	**J**	1.083369e+18

4.3 Comparación volumen total y atribuido a jampp por marca

Interesa, sin embargo, los eventos atribuidos a Jampp. Se realizó el mismo análisis solo para eventos atribuidos y se observa, aunque el conteo es mucho menor, que la tendencia es la misma.

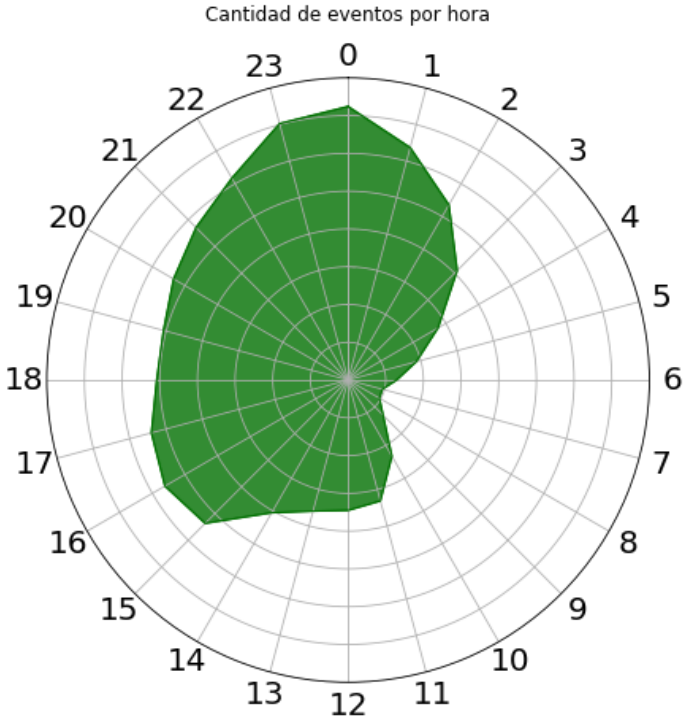


Se puede concluir entonces que a mayor número de eventos, mayor número de eventos atribuidos sin importar la marca.



4.4 Ocurrencias según la hora del día

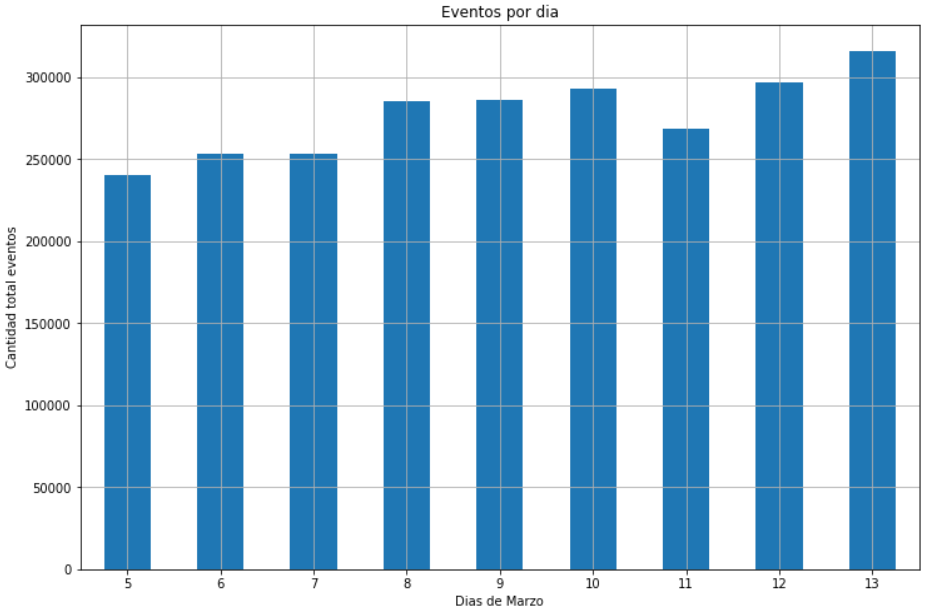
Por el lado del analisis hora por hora, se encuentra que la mayor parte de los eventos ocurren en horas nocturnas, con pico entre las 23 y las 0hs. Disminuye considerablemente para las primeras horas de la mañana.



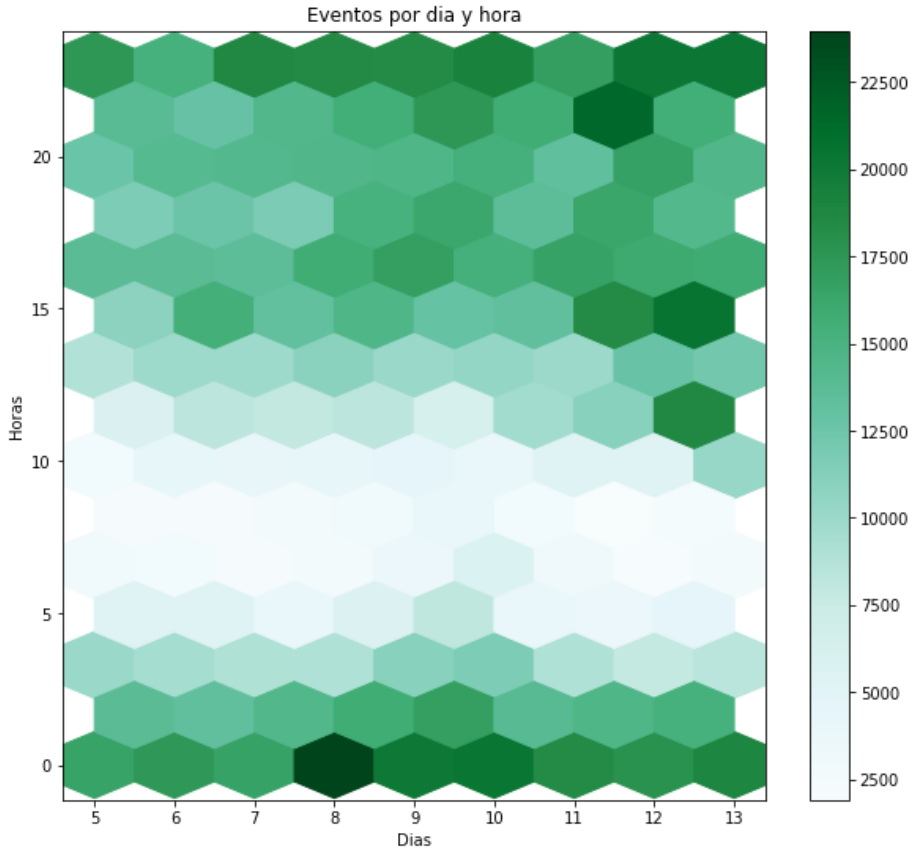
Escala radial: Desde 25000 hasta 200000, paso de 25000.

4.5 Ocurrencias por día de marzo

Se quiere ver en las próximas secciones como se distribuyen los eventos temporalmente. En el gráfico de cantidad por dia, lo se puede ver bastante uniforme, con cierta tendencia a aumentar hacia el final de la ventana de tiempo brindada por los datos.

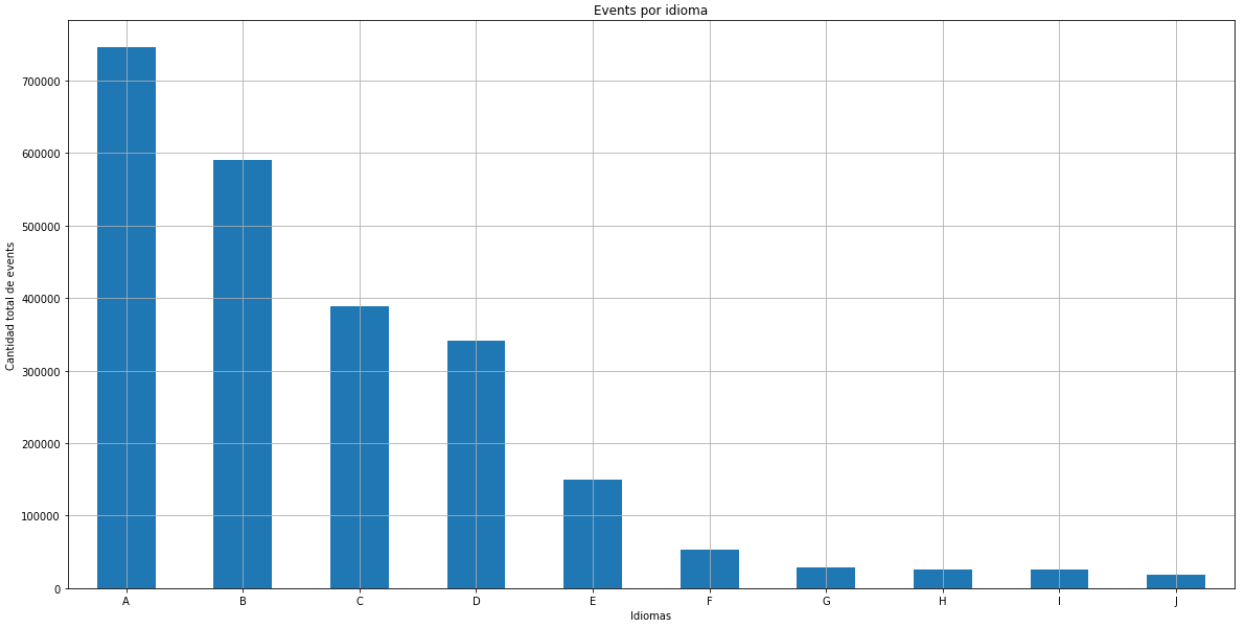


4.6 Cantidad de events por dia y hora



Visualizando ambas variables temporales en conjunto se observa la misma tendencia de alta actividad en horas nocturnas y muy poco en las primeras horas de la mañana. Además se mantiene constante según el día del mes.

4.7 Clasificación por idioma

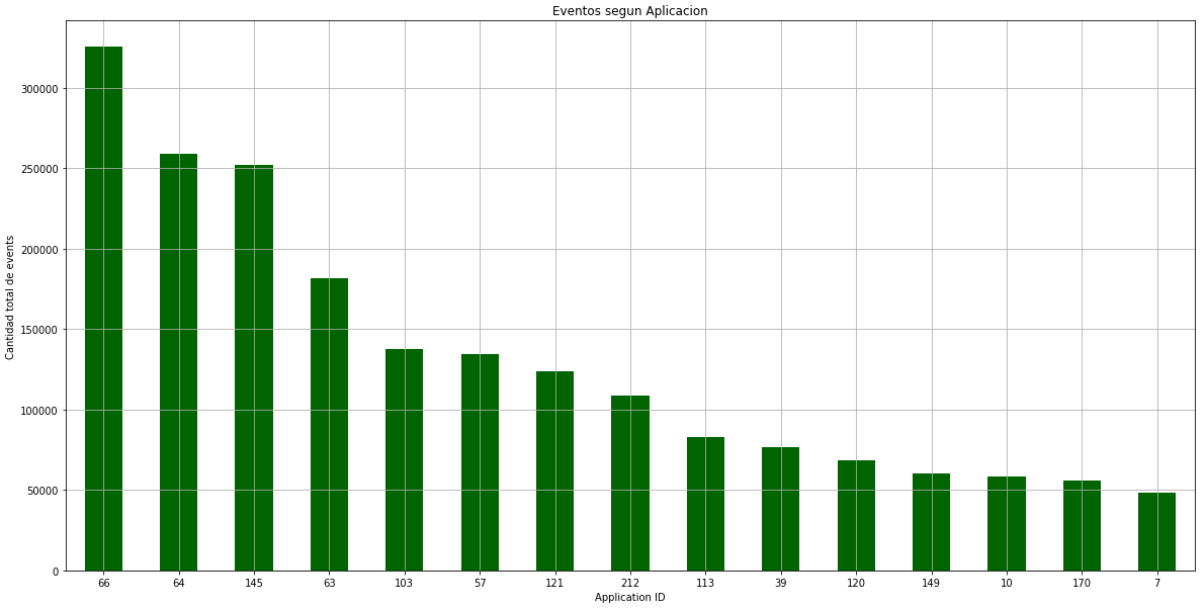


Alias	Valor	Alias	Valor
A	6.977049e+18	**F**	4.077062e+17
B	3.301378e+18	**G**	2.822843e+17
C	9.028383e+18	**H**	4.060930e+18
D	6.804428e+18	**I**	3.095856e+18
E	4.621024e+18	**J**	6.111820e+18

Mediante el gráfico se puede concluir que hay cuatro idiomas que son los más utilizados en los eventos, y a diferencia de los idiomas en las instalaciones, no hay uno predominante.

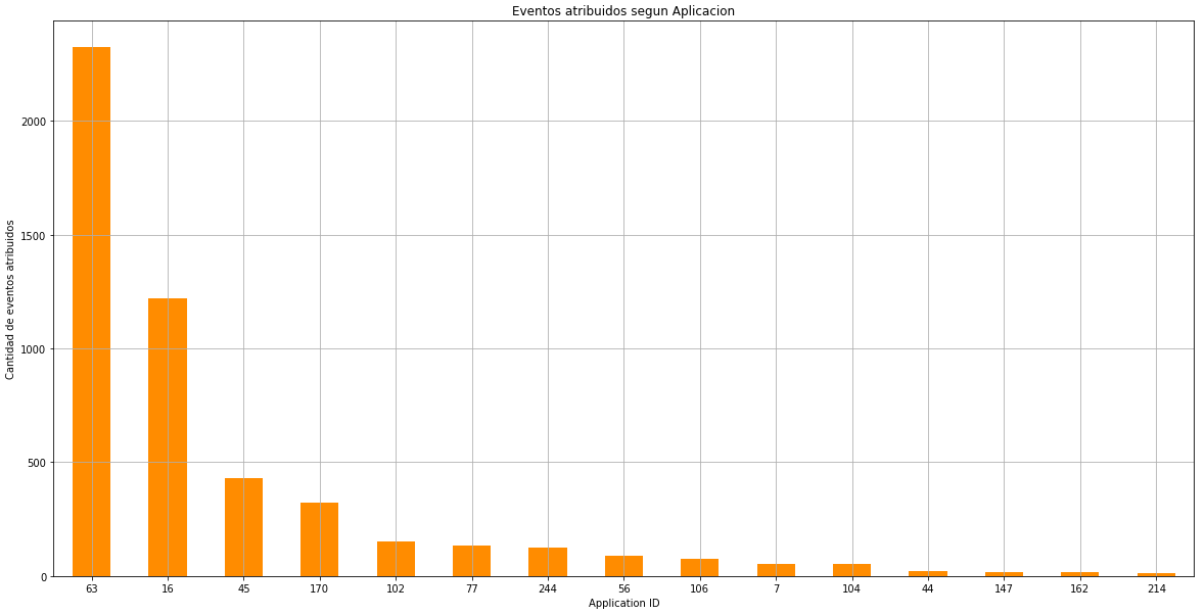
4.8 Volumen total por ID de aplicación

En este apartado se quiere estudiar la distribución de los eventos por ID de aplicación.



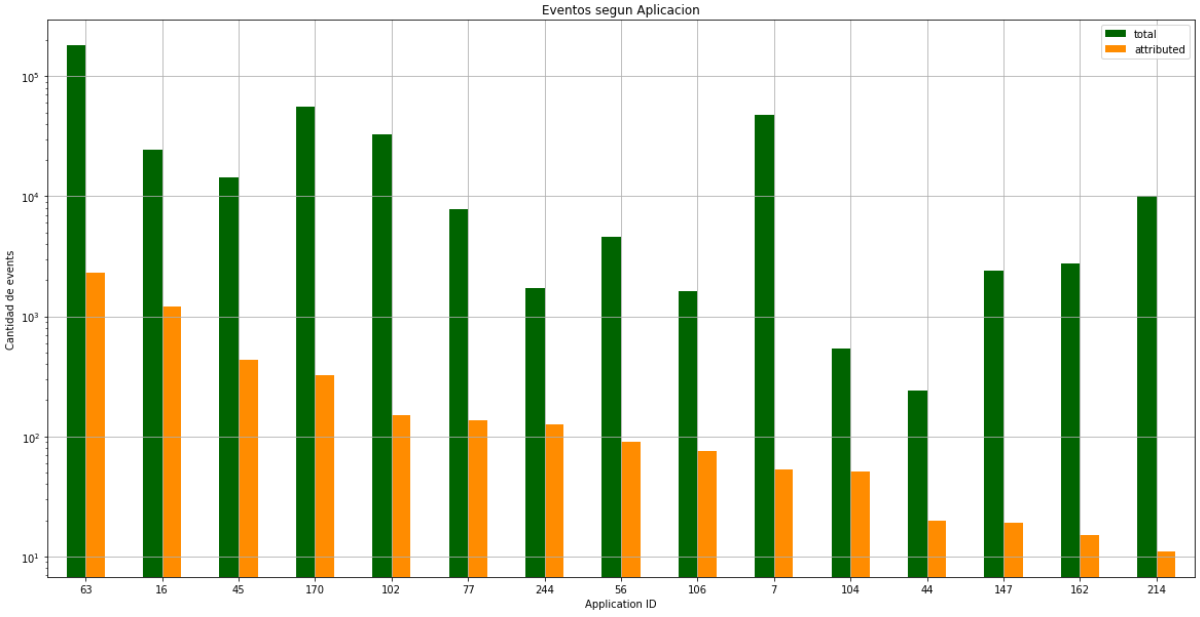
En el gráfico se puede observar que la mayoría de los eventos se realizan en unas pocas aplicaciones, pero no hay una que sea claramente predominante. En otras palabras, el decrecimiento no es *tan* marcado.

4.9 Comparación volumen total y atribuido a jampp por ID de aplicación



En este gráfico se puede observar la distribución de los eventos atribuidos a Jampp en función de la ID de aplicación. Curiosamente, a diferencia del apartado anterior, sí se puede ver una aplicación predominante.

A continuación un gráfico que combina los eventos totales y los atribuidos, ordenando de mayor a menor según atribuidos:



Cabe aclarar que solo se muestran los 15 con más atribuidos, lo cual resultó en que algunos con muchos totales queden fuera de la gráfica. Así, teniendo en cuenta que la aplicación que tiene más eventos atribuidos es la cuarta con más eventos totales, podemos concluir que las aplicaciones con más eventos totales no son necesariamente las que le atribuyen más eventos a Jampp. Dicho de otra forma, se ve que una gran interacción de usuarios no se traslada en un buen rendimiento (muchos atribuidos) para Jampp, lo cual no es un resultado inmediatamente obvio.

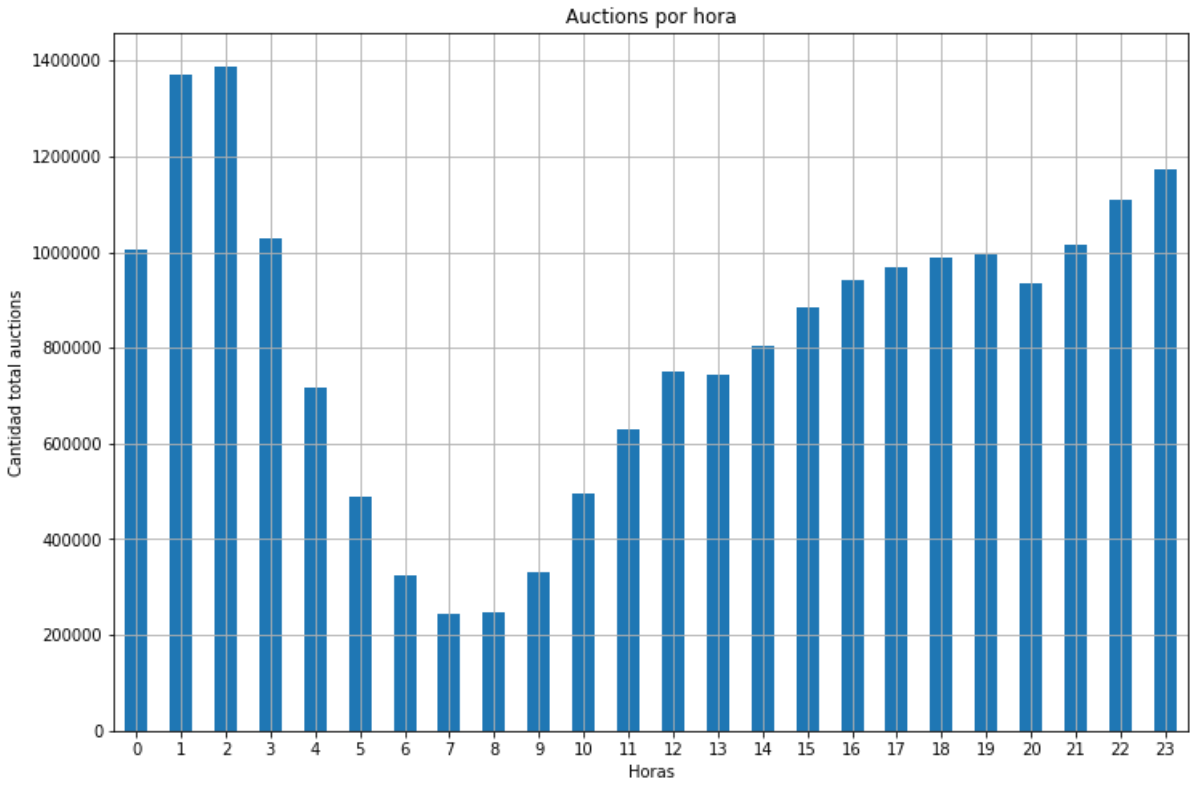
5 Combinación de los datos

5.1 Cantidad por hora

5.1.1 Introducción

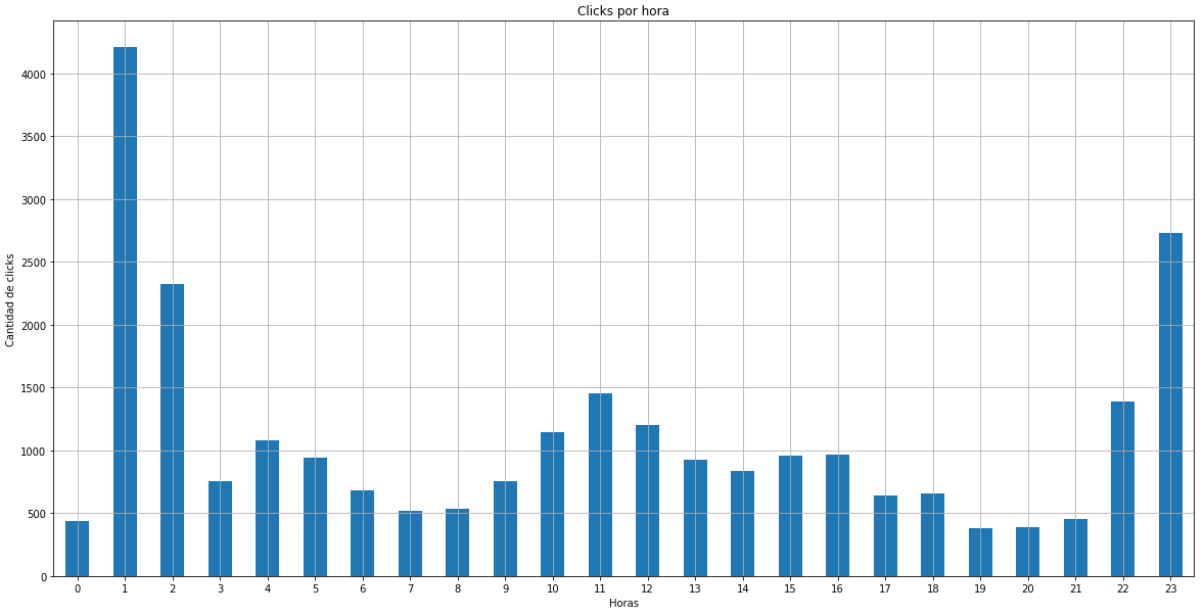
La motivación de esta sección es encontrar las horas de mayor y menor tráfico de usuarios para cada uno de los sets de datos, así como lograr establecer algún tipo de relación entre ellas.
Si bien se analizó por separado cada uno de los casos, se vuelve a realizar un resumen (acompañado por gráficos más cuantitativos que los anteriores) para ayudar a un mejor entendimiento de la conclusión.

5.1.2 Cantidad por hora para auctions



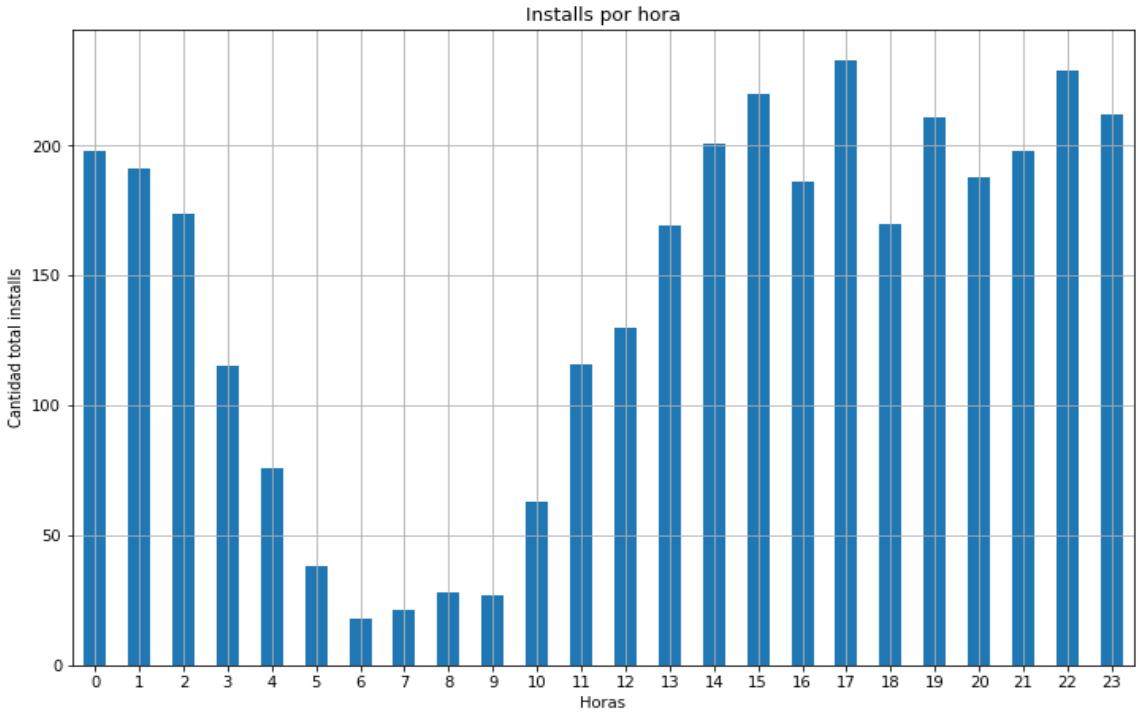
Se puede observar claramente que existe un mínimo muy marcado entre las 6 y las 9 (AM). El máximo se encuentra entre las 0 y las 3 (AM).

5.1.3 Cantidad por hora para clicks



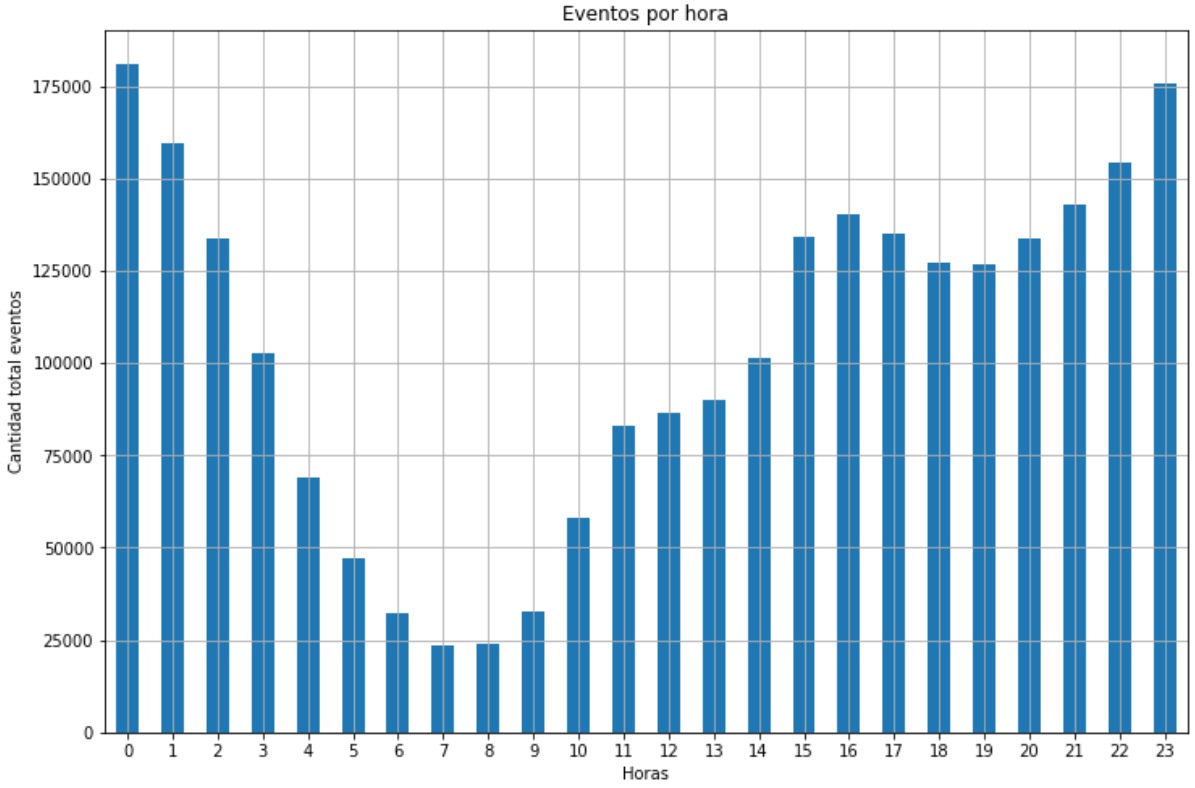
A diferencia de sus pares, el gráfico de clicks no representa una forma clara y bien definida. Se sospecha que este comportamiento se debe al menor volumen de datos con el que se cuenta, lo que posibilita que una diferencia de pocos clicks se muestre como un salto abrupto.
El mínimo puede encontrarse entre las 6 y las 9, o bien, entre las 18 y 22. La diferencia entre ambas bandas de valores es tan pequeña que no se puede discernir con total claridad. Sin embargo, viendo el resto de los gráficos se podría sugerir que el verdadero mínimo es el de las 6 - 9 (AM). El máximo está claramente situado entre la 1 y las 2 (AM).

5.1.4 Cantidad por hora para installs



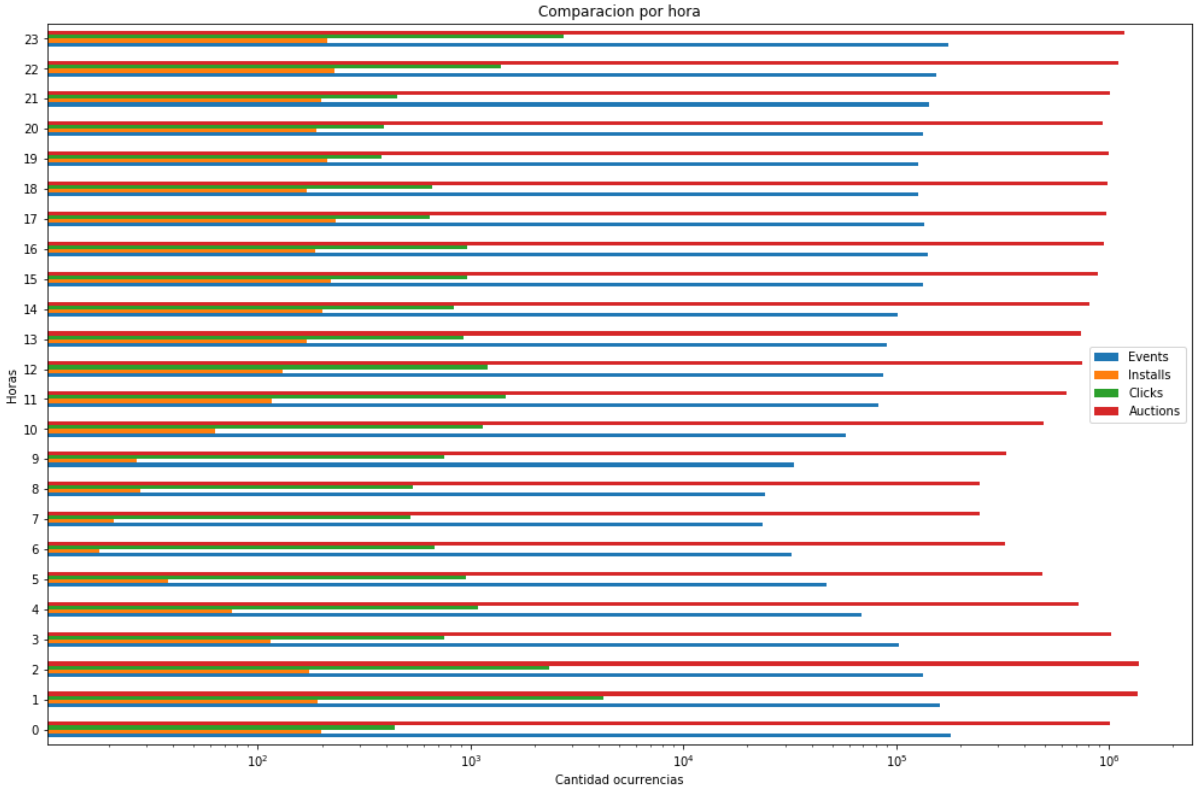
Existe un marcado descenso de los installs entre las 5 y las 10 (AM) que se puede ver con facilidad en el gráfico. Por otro lado, el máximo es más difícil de señalar ya que los valores se mantienen relativamente constantes en las horas que restan. En las próximas secciones se tomará una decisión en cuanto al máximo, pero por el momento se podría establecer entre las 13 y las 23.

5.1.5 Cantidad por hora para events

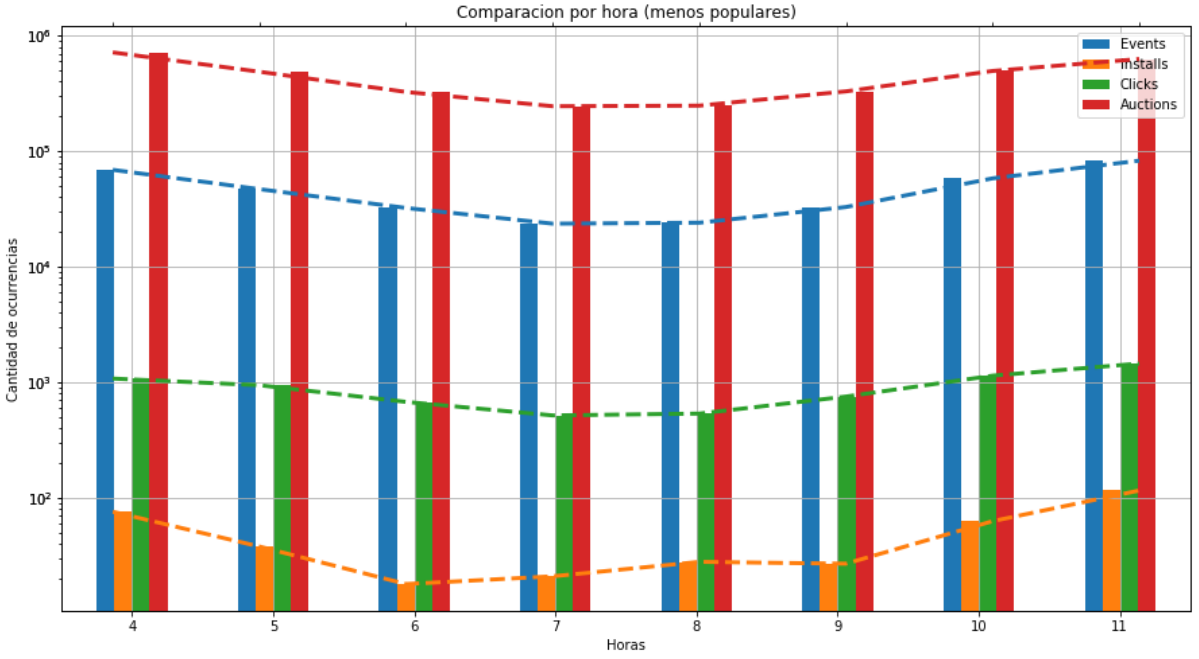


Se puede marcar un mínimo bastante pronunciado entre las 6 y las 9 (AM). El máximo se ubica entre las 22 y la 1.

5.1.6 Superposición de los gráficos



Si bien el gráfico anterior no es del todo claro para ver valores individuales, sirve para ver que los horarios con actividad mínima se encuentran todos en el mismo lugar. A continuación se expande dicha zona para una mejor visualización.



Se agrega una línea de tendencia entre las distintas barras para lograr ver que todas presentan una forma similar. De ninguna manera se pretende que esas líneas sean una

extrapolación de los datos. Los mismos fueron obtenidos de manera discreta y solamente de esa forma se los debería utilizar.

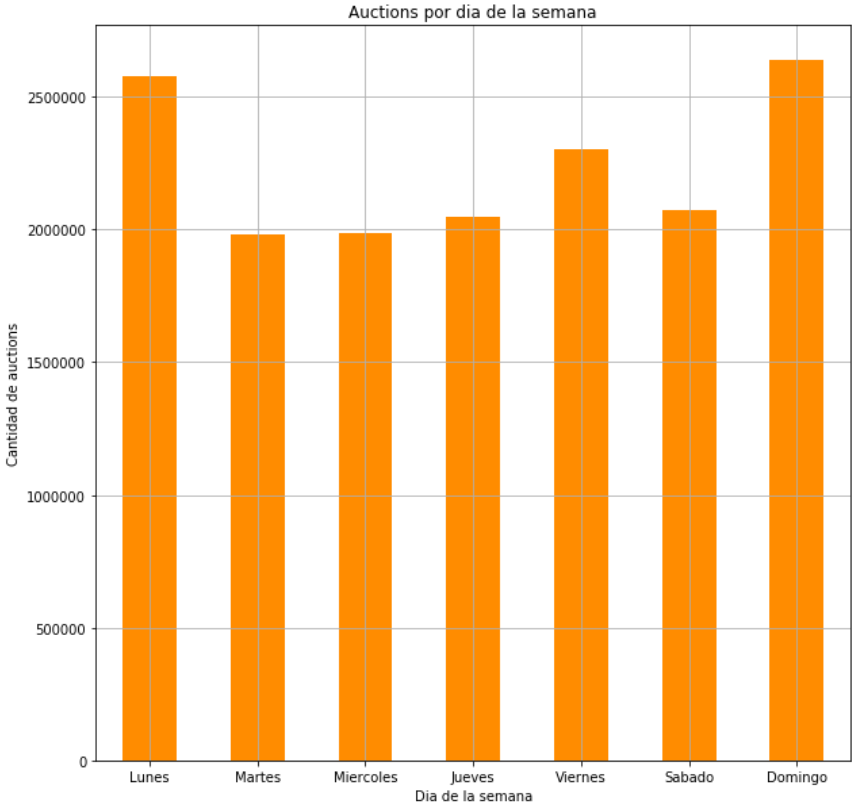
Por todo lo anteriormente mencionado se puede concluir que entre las 6 y las 9 (particularmente entre las 7 y las 8) se registra un descenso muy marcado de la actividad en los dispositivos.

5.2 Cantidad por día de la semana

5.2.1 Introducción

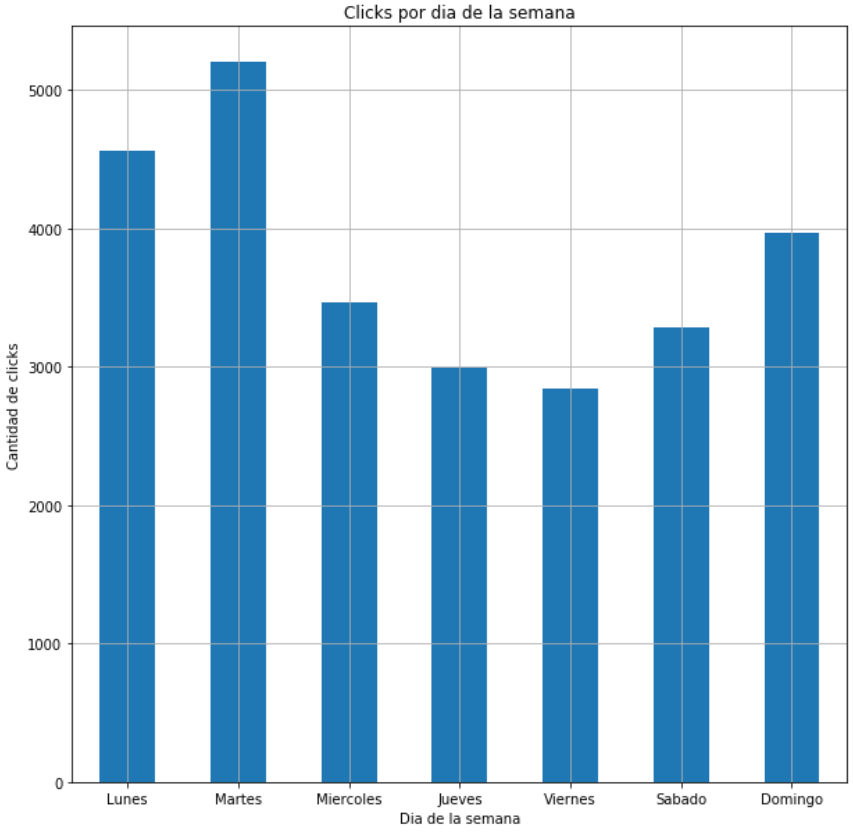
Este apartado tiene por objetivo determinar la relación de la cantidad de usuarios con el día de la semana. A priori se podría pensar que los usuarios son más propensos a usar el celular en determinados días en particular, por lo que se va a realizar un análisis para averiguarlo.
Nota: debido a que los datos pertenecen a 9 días, se decidió hacer un promedio de los días que se repiten. Sin embargo, en el apartado de clicks se encontraron valores anormalmente bajos para los dos primeros días de la semana, por lo que se resolvió dejar los valores más coherentes y descartar los primeros.

5.2.2 Cantidad por día de la semana para auctions



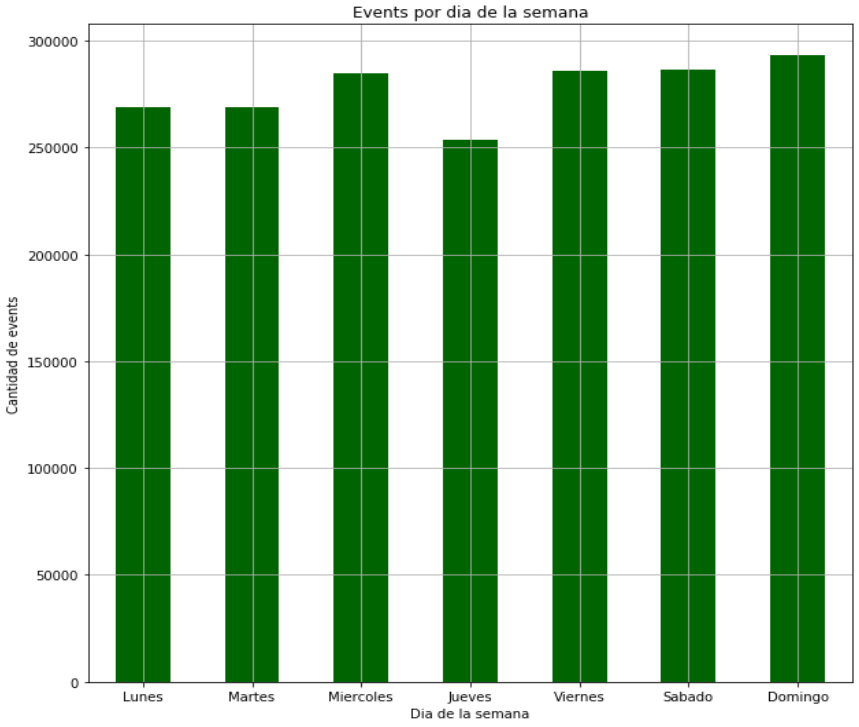
El gráfico muestra que existe un pico de subastas en los días domingo y lunes, que baja abruptamente al mínimo en el día martes. Desde el mínimo se puede ver un ascenso paulatino de las cantidades hasta llegar al máximo del domingo, lo que indicaría un comportamiento relativamente predecible.

5.2.3 Cantidad por día de la semana para clicks



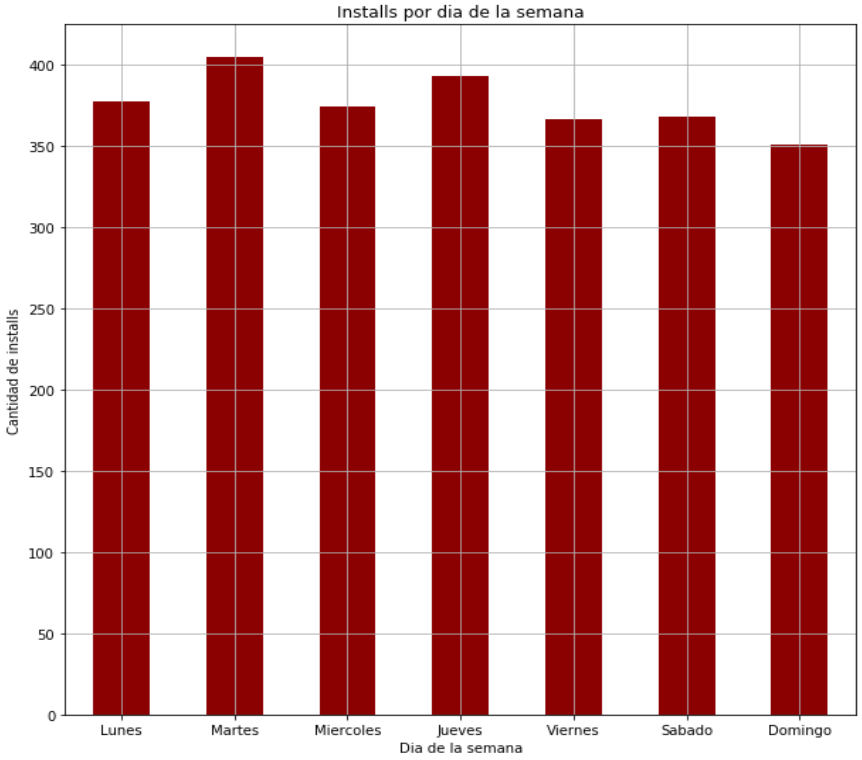
Curiosamente para los clicks se observa una tendencia casi opuesta que en las subastas. El mínimo del día martes se convierte en el máximo para clicks. Este comportamiento no favorece a la hipótesis inicial, pero se continuará con el análisis antes de sacar conclusiones.

5.2.4 Cantidad por día de la semana para events



Para events se puede ver que los valores son mucho más constantes que en sus pares. Si se quiere señalar un máximo se podría ubicar en el día domingo y el mínimo en el día jueves, pero no se trata de saltos muy marcados.

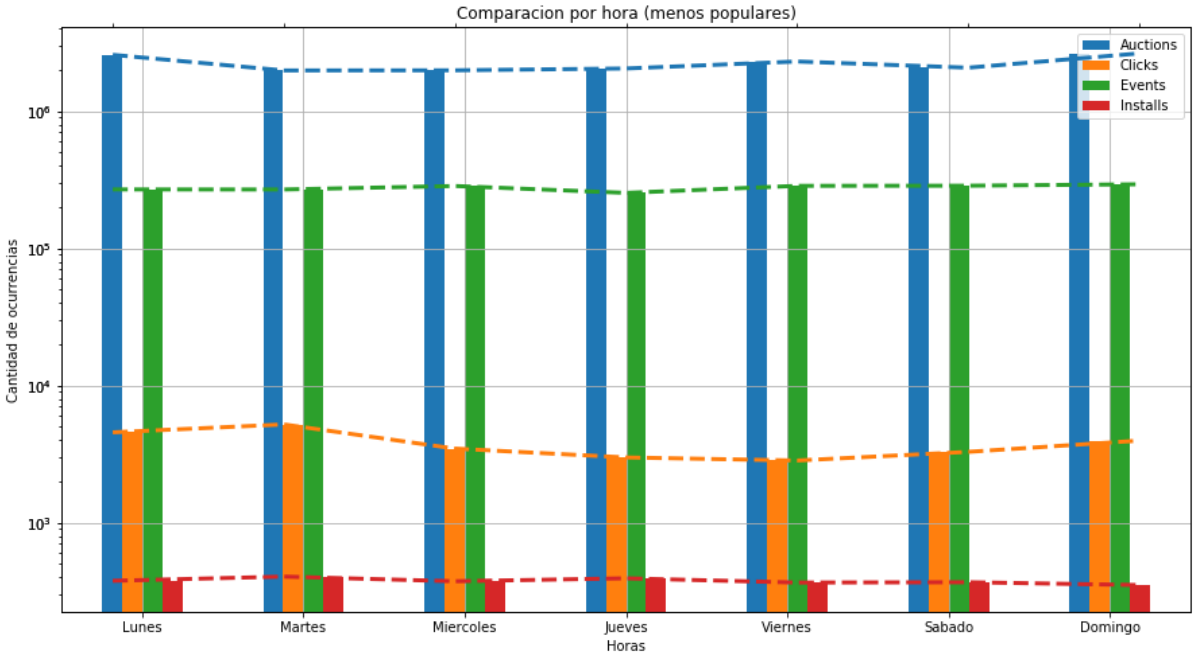
5.2.5 Cantidad por día de la semana para installs



En este caso, el mínimo se da para el domingo y el máximo para el martes.

5.2.6 Superposición de los gráficos

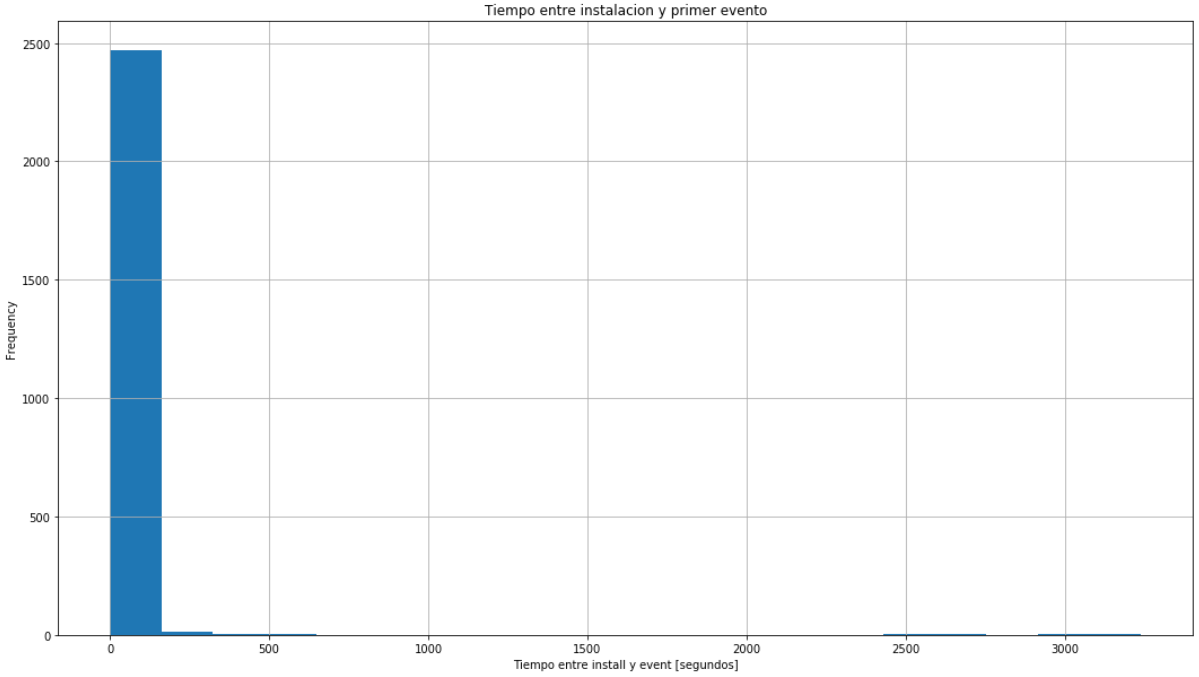
Como se fue evidenciando a lo largo del desarrollo de esta sección, los mínimos y máximos no se alinean para los diferentes gráficos. Por ese motivo, no podemos obtener información de este análisis más que decir que los días de la semana, en principio, no afectarían a la cantidad de interacciones. Para verlo más claramente se superponen los diferentes gráficos en una escala logarítmica.



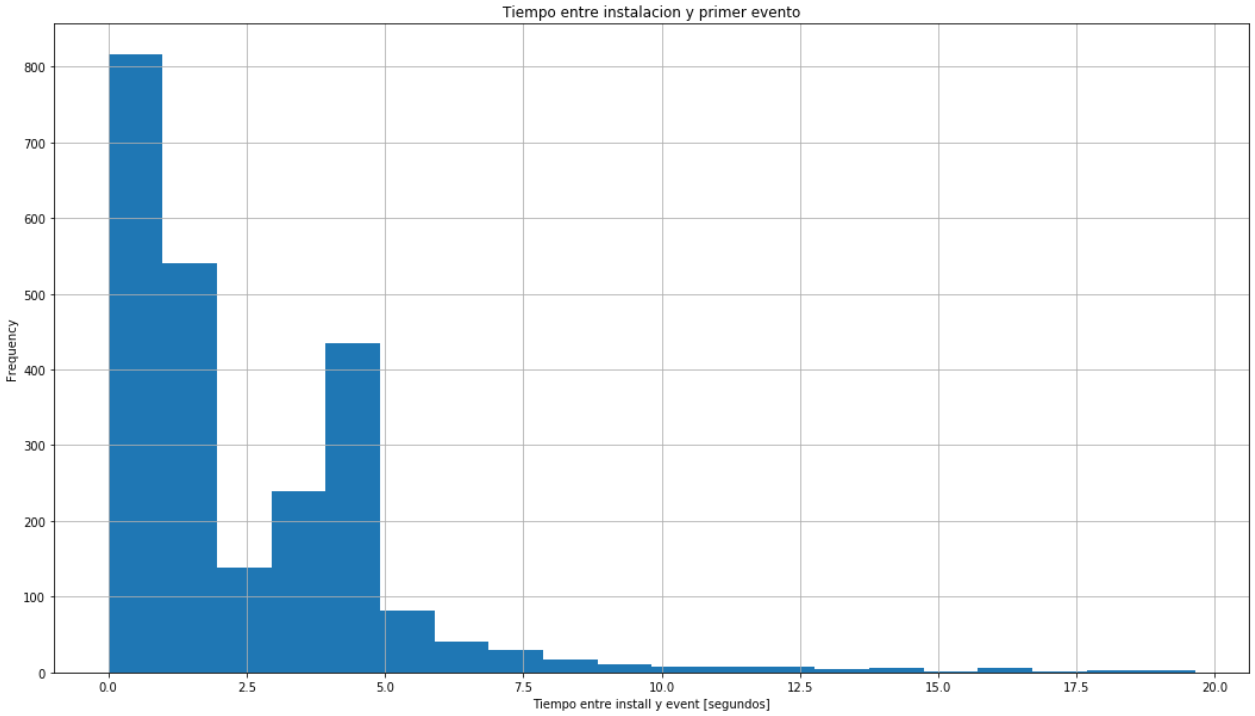
Se puede ver entonces que los valores son relativamente constantes, y que los pequeños mínimos no se alinean entre los diferentes niveles.
Por el momento se establece que el análisis en inconcluyente. Se podría repetir el mismo procedimiento pero para un período que abarque más semanas para asegurar que la tendencia se repite.

5.3 Distribución en función del tiempo entre que se instala una aplicación y se realiza un event

Este apartado tiene por objetivo buscar el tiempo que tardan las personas entre que instalan una aplicación y que realizan un evento en la misma. Es lógico pensar que luego de instalar una aplicación, la mayoría de las personas entran y realizan algún evento. Se decidió estudiar el tiempo entre la primera instalación realizada, y el primer evento.

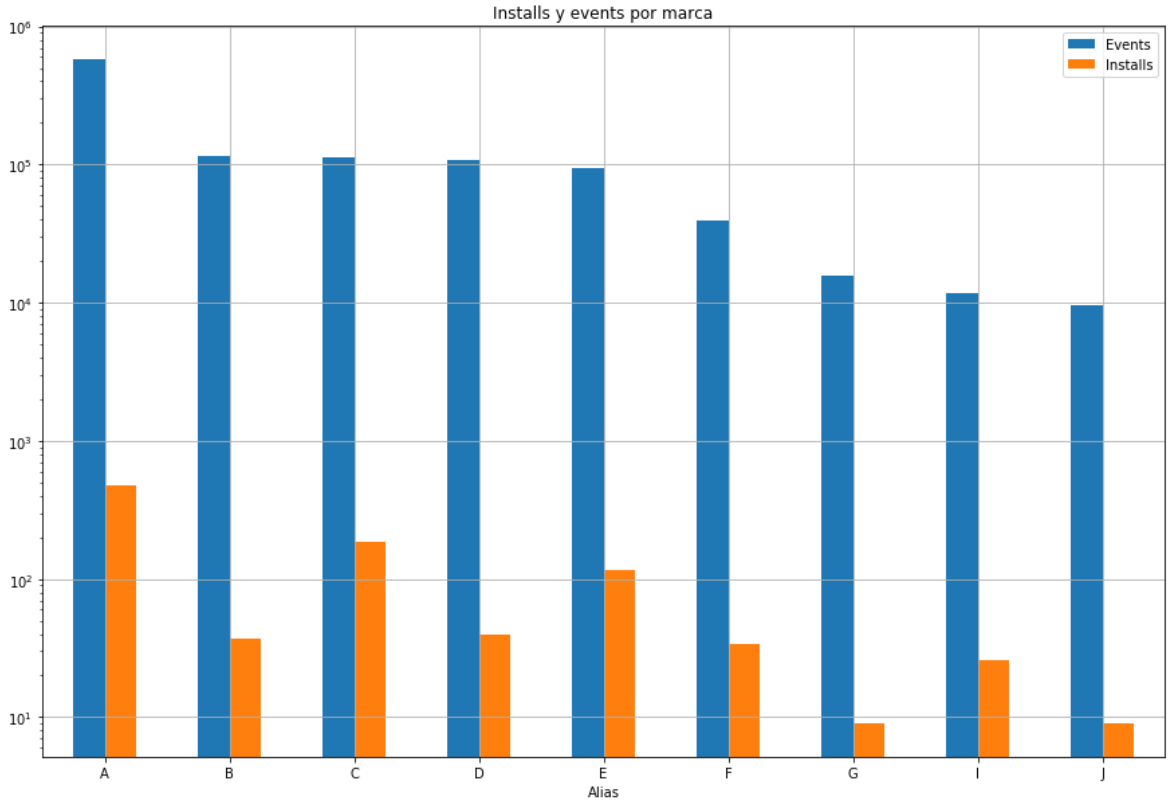


En el gráfico se puede observar que de la primera hora estudiada, la mayoría se encuentra en los primeros segundos, por lo tanto fue necesario un segundo gráfico que muestre más detalladamente la distribución.


Tiempo entre instalacion y primer evento
Frequency
0
100
200
300
400
500
600
700
800
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Tiempo entre install y event [segundos]

5.4 Comparación de events e installs por device brand

Se quiere analizar cómo se comportan en conjunto las instalaciones y los eventos, para cada marca de dispositivo.
Lo que nos indica el siguiente gráfico es que aquellas marcas que resultaban predominantes para los eventos son, aunque no necesariamente en el mismo orden, aquellas que son predominantes para las instalaciones.



Alias	Valor	Alias	Valor
A	3.083059e+17	**F**	5.137992e+17
B	3.812621e+18	**G**	3.228516e+18
C	2.208835e+18	**I**	6.538562e+18
D	2.987569e+18	**J**	1.083369e+18
E	2.523246e+18		

Conclusiones y Aprendizajes

- Es difícil trabajar con un volumen grande de datos. Los tiempos para cargar en memoria datasets de las dimensiones trabajadas, especialmente auctions y events, fueron un problema y por lo tanto es recomendable el correcto manejo de los tipos y las columnas para reducir dicho problema.

- Jampp debería enfocarse en las horas de mayor volumen de datos, que es donde se registran la mayor cantidad de eventos e installs (lo que debería aumentar las probabilidades de que le adjudiquen una conversión).

- Jampp no debería realizar mantenimientos en las horas cercanas a la medianoche, sino más bien en las primeras horas de la mañana ya que son las que menos movimiento presentan (evitando perder la menor cantidad posible de oportunidades de generar ingresos).

- El día de la semana no influye en el volumen de interacciones.

- La ventana de tiempo de los datos otorgados (9 días) es correcta para analizar los datos hora por hora. No es así para el análisis diario, ya que con un intervalo de tiempo mayor se podría haber evaluado otros comportamientos,como por ejemplo, el contraste fin de semana vs dia de semana.

- Se podría concluir que la cantidad de eventos atribuidos no está directamente relacionada con la cantidad de eventos totales, es algo que depende de la marca del dispositivo, de las cuales no se posee mayor informacion, pero seria interesante evaluar, por ejemplo, si son celulares de alta gama los que muestran más eventos atribuidos a Jampp, y se podría entonces evaluar cuales aplicaciones son más populares entre estos dispositivos. Se podría hacer un análisis similar para el tipo de aplicación donde se producen eventos.

- Los datos se presentaban en forma hasheada para evitar problemas de filtración de información sensible para los cliente de Jampp, pero esta decisión trajo como consecuencia que la mayoría de los campos categóricos sean numéricos, perdiéndose información y teniendo que trabajar de forma abstracta. En principio no debería afectar al desarrollo del trabajo, pero al tratarse de temas no del todo sencillos de entender, la complejidad del análisis aumentó significativamente. El entendimiento del significado de cada campo y la relación entre los distintos *data sets* llevó mucho más tiempo del que hubiese llevado si los datos no estuvieran transformados. Además, se podrían haber obtenido ilustraciones gráficas mucho más interesantes de haber sabido un poco más acerca de los datos con los que se trabajaba.