75.06/95.58 Organización de Datos - 1C 2019

Trabajo Práctico 1

Análisis Exploratorio

Grupo 34: "DataTravellers"

Integrantes:

- Andrés Pablo Silvestri: 85881 (silvestri.andres@gmail.com)
- Juan Manuel González: 79979 (juanmg0511@gmail.com)
- Patricio Pizzini: 97524 (pizzinipatricio@yahoo.com.ar)

Link a repositorio de GitHub:
https://github.com/silvahlaravel/Organizacion_Datos_1C2019/tree/master/TP1

Fecha de entrega: 22/04/2019

Contenido

1 - Introducción

Este trabajo está enfocado en hacer un primer análisis de los datos ofrecidos por Jampp, de manera que encontremos particularidades que puedan ser de interés para dicha empresa.

Siendo este el caso, lo primero que vamos a hacer es interiorizarnos de uno de los pilares que tiene la ciencia de datos, el negocio.

Se hará un análisis general para entender un poco más el negocio y luego un análisis más fino sobre los datos para contestar algunas preguntas que podrían resultar curiosas.

1.1 - ¿De qué se encarga Jampp?

Jampp es una plataforma para la promoción y remarketing de aplicaciones móviles. Se encarga de promover las apps de sus anunciantes a nivel global, y por otro lado de recuperar aquellos usuarios que ya instalaron la app pero no la usan actualmente.

Para lograr su objetivo, Jampp cuenta con una plataforma propietaria que participa de subastas en tiempo real (real time bidding o RTB), donde se ofrecen los espacios para colocar publicidad, y determina en forma programática en cuales de ellas participar, con la posibilidad de dirigirse a segmentos específicos de usuarios.

En las subastas ganadas por la plataforma, el usuario del espacio subastado verá la publicidad del anunciante de Jampp, sobre la cual podrá interactuar (hacer “click”) y finalmente optar por instalar la app promocionada.

En este caso, Jampp nos facilitó un set de datos que contiene distintos tipos de eventos que se registran en su plataforma, ya sea subastas, clicks, instalaciones u otros tipos de eventos que se detallarán luego.

2 - Estructura y manejo de los datos

A fin de mejorar la performance de ejecución de las operaciones realizadas, y por otro lado facilitar las tareas de análisis, se ha hecho un tratamiento previo sobre el set de datos proporcionado por la cátedra.

Se puede consultar el tratamiento completo, así como también el código utilizado para generar el análisis y los gráficos, en el notebook disponible en la carpeta del repositorio de Git presentado en la carátula de este informe.

Como primer medida, se procedió a tipar las columnas de los archivos en forma manual, a fin de mejorar el tiempo de carga y de uso de memoria. Con los cuatro archivos cargados en forma simultánea el consumo de RAM estaba en unos 8GB, esto mejoró notablemente al implementar la estrategia mencionada, según se recomendó en clase, bajando a unos 2GB.

Contamos con cuatro archivos para esta entrega (el archivo '*target_competencia.csv*' se utilizará en la segunda), que cubren un período de actividad de 8 días, más precisamente del 05/03/2019 al 13/03/2019:

- **EVENTS** - información sobre eventos de la plataforma en general.
- **CLICKS** - información sobre los eventos de click, es decir cuando un usuario interactúa con una publicidad.
- **INSTALLS** - información sobre los eventos de instalación, es decir cuando un usuario termina instalando la aplicación promocionada.
- **AUCTIONS** - información sobre las subastas disponibles.

Como primer medida pasamos a 'category' todas las columnas que nos resultan acordes para ser categorizadas, lo que nos permite darle otro tipo de manejo como así también mejorar el rendimiento de las diferentes operaciones que involucren estas columnas.

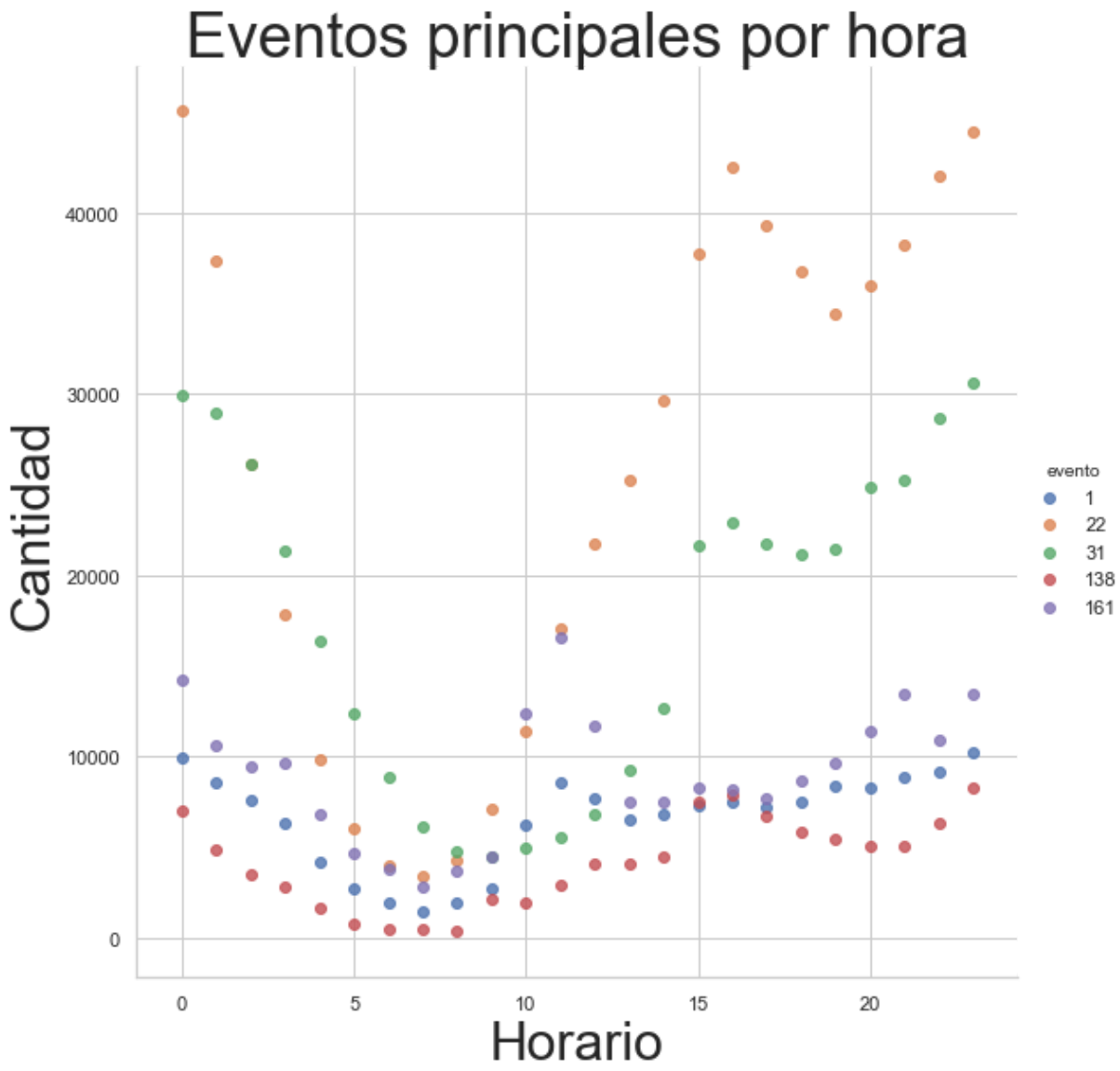
Asimismo, armamos algunas columnas auxiliares que nos permitan separar el tiempo en día, mes y hora para luego poder aplicar esta información como parámetros de búsqueda y organización en los diferentes gráficos que planteamos, como así también por ejemplo plantear el día de la semana con un formato personalizado que nos permita hacer más legible futuros gráficos.

Con respecto a los eventos tenemos que aclarar que no podemos hacer una distinción fidedigna sobre qué representa cada uno, o armar una cadena lógica de lo que pudiese llegar a representar un conjunto de eventos realizados por un celular en particular o para una aplicación en cuestión. De todas maneras este tipo de cosas es algo que se ha repetido en el análisis debido a que los datos están anonimizados, lo que implica que tengamos que tratar muchos de ellos sin saber lo que estamos expresando en última instancia.

Por último, vale agregar que en algunos casos nos hemos tomado la libertad de considerar algunos valores como despreciables en cuanto a relación de cantidad sobre el total, como así también hacer un análisis sobre los eventos que más cantidad de registros nos brindan lo que nos permite entender la situación con un volumen de datos mayor lo cual nos acerca un poco más a la realidad.

3 - Análisis sobre los horarios y la actividad

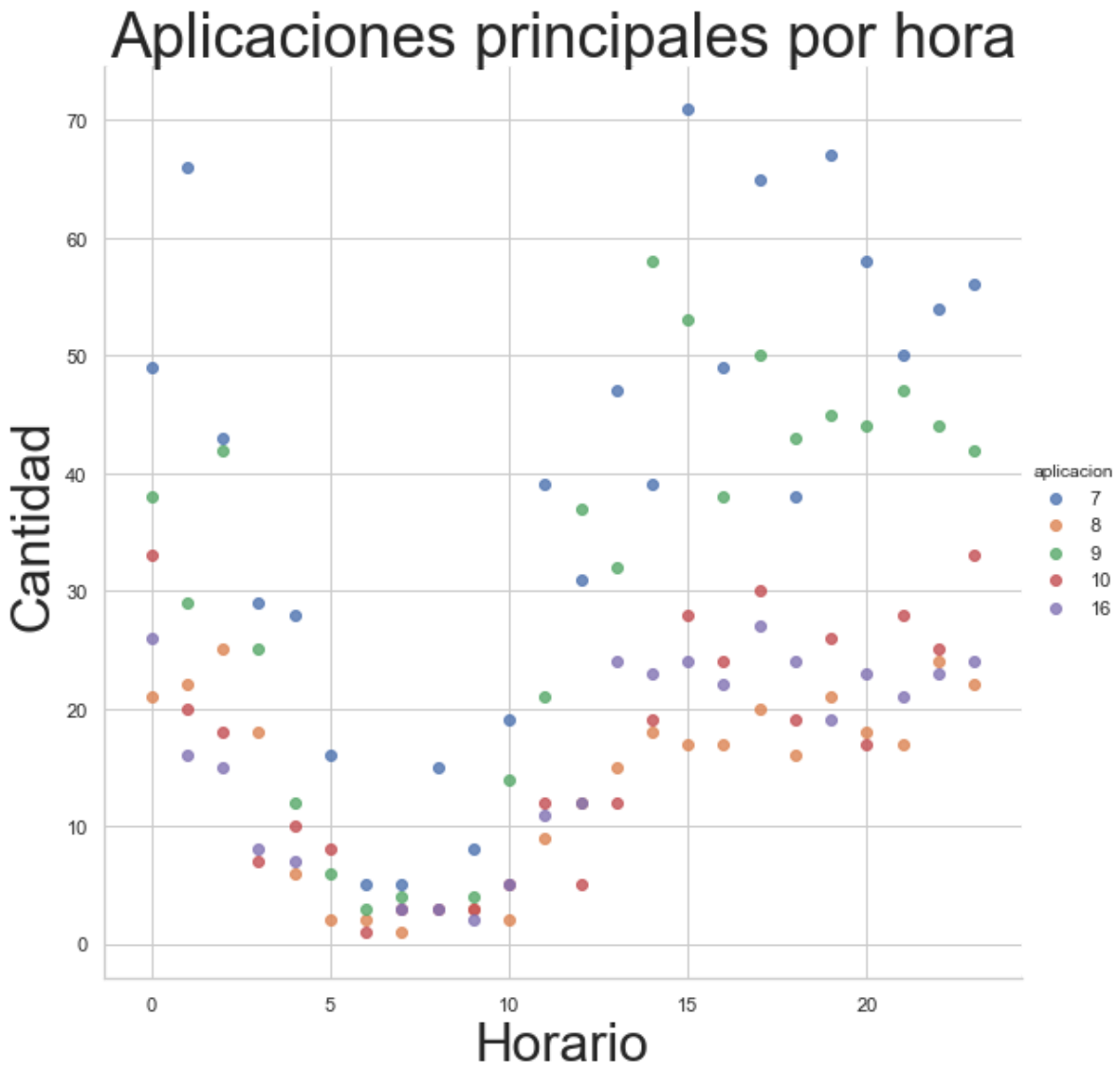
3.1 - ¿Cuáles son los horarios en los que ocurren los eventos más frecuentes?



Hemos tomado los cinco eventos principales, es decir aquellos que tienen más registros y hemos analizado como es el comportamiento a lo largo de las diferentes horas del día. Al tener los datos anonimizados, decidimos mostrar los ID de los eventos, lo cual permitirá al dueño del negocio entender esta información. Vemos que hay dos eventos que

tienen mucha interacción en los horarios de la tarde, noche y madrugada, pero que luego se acoplan al resto de los eventos, casi compartiendo la misma cantidad de apariciones que el resto en los horarios de las primeras horas del día y el medio día.

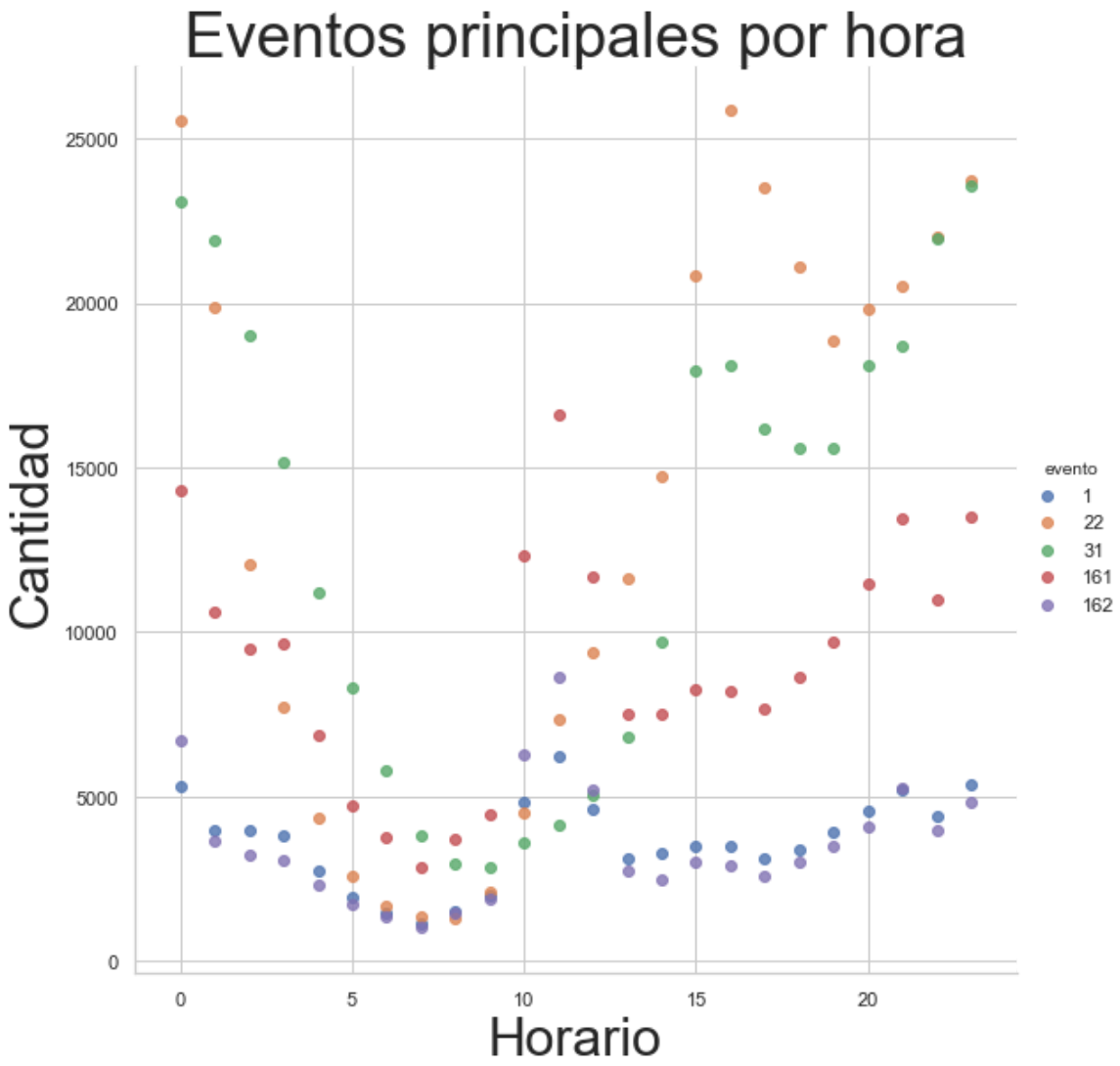
3.2 - ¿Cuáles son los horarios en los que se suelen instalar las aplicaciones principales?



Haciendo un análisis similar, hemos buscado cómo se maneja la instalación de aplicaciones a lo largo de la franja horaria que compone un día. Podemos ver que a diferencia de la ejecución de eventos, en la instalación fuera del horario de la primer hora de la mañana estas son mucho más dispersas, aunque vale aclarar que el valor de la

cantidad de instalaciones en general es mucho menor, lo que requiere tomar con mucho cuidado esta información. Asimismo, destacamos que hemos optado por escoger aquellas aplicaciones que más interacción tienen, es decir las que podrían ser consideradas más relevantes.

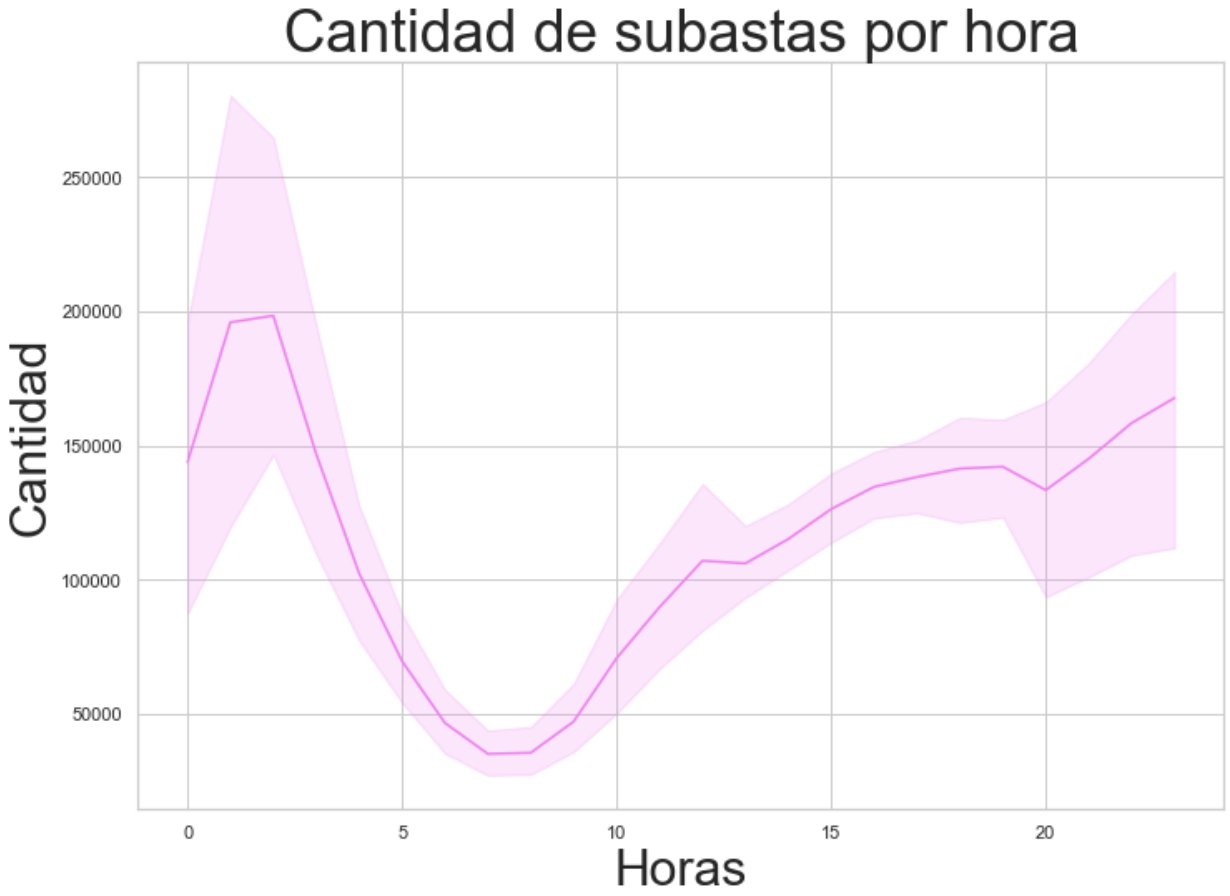
3.3 - ¿Cuáles son los horarios en los que ocurren eventos en las aplicaciones más populares?



Para este caso hemos decidido tomar las aplicaciones más relevantes, las mismas que tienen mayor interacción, y sobre estas averiguar cuáles son los eventos que más

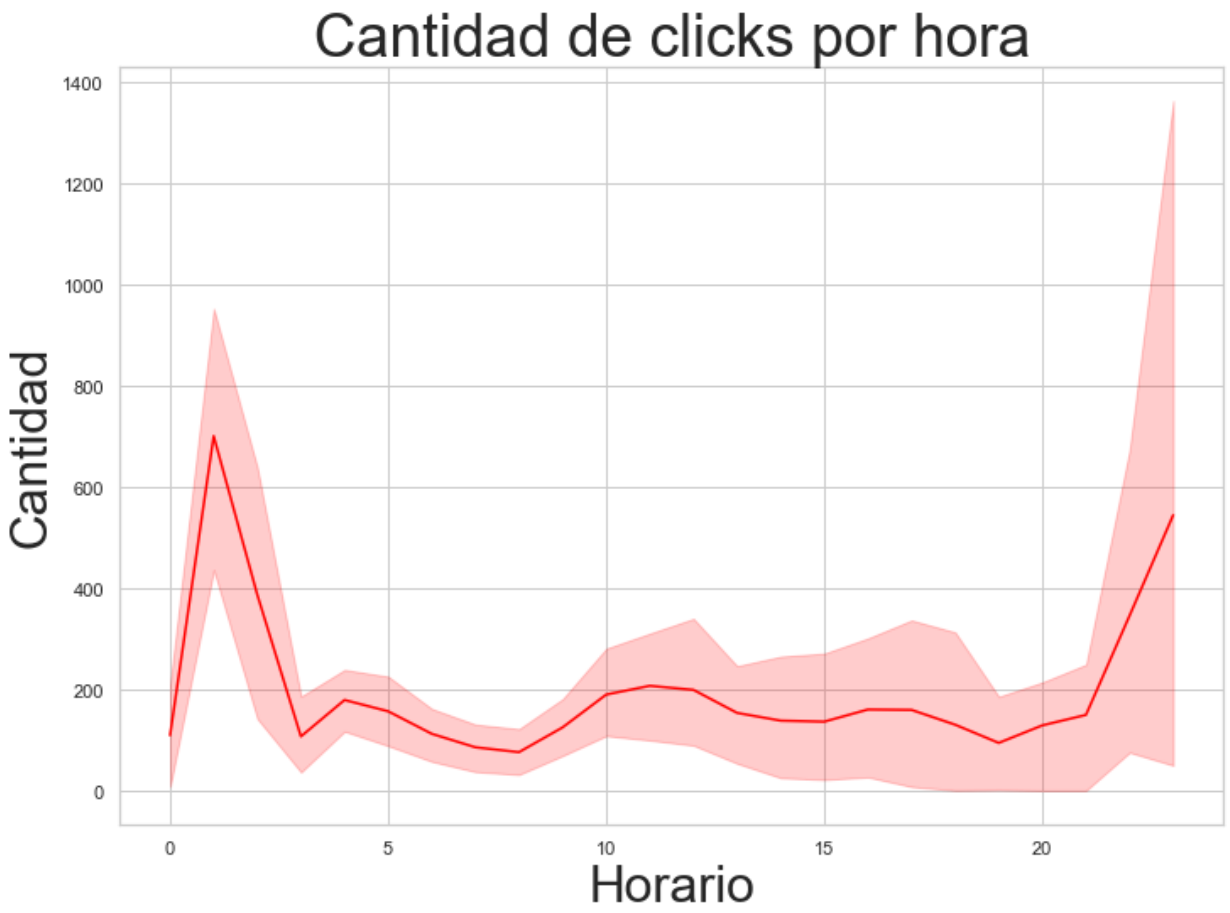
influyen. Sobre esta base, el gráfico nos muestra algo bastante similar a lo que vimos en el la visualización realizada en el punto 3.1.

3.4 - ¿Cómos se distribuyen las subastas a través de las horas del día?



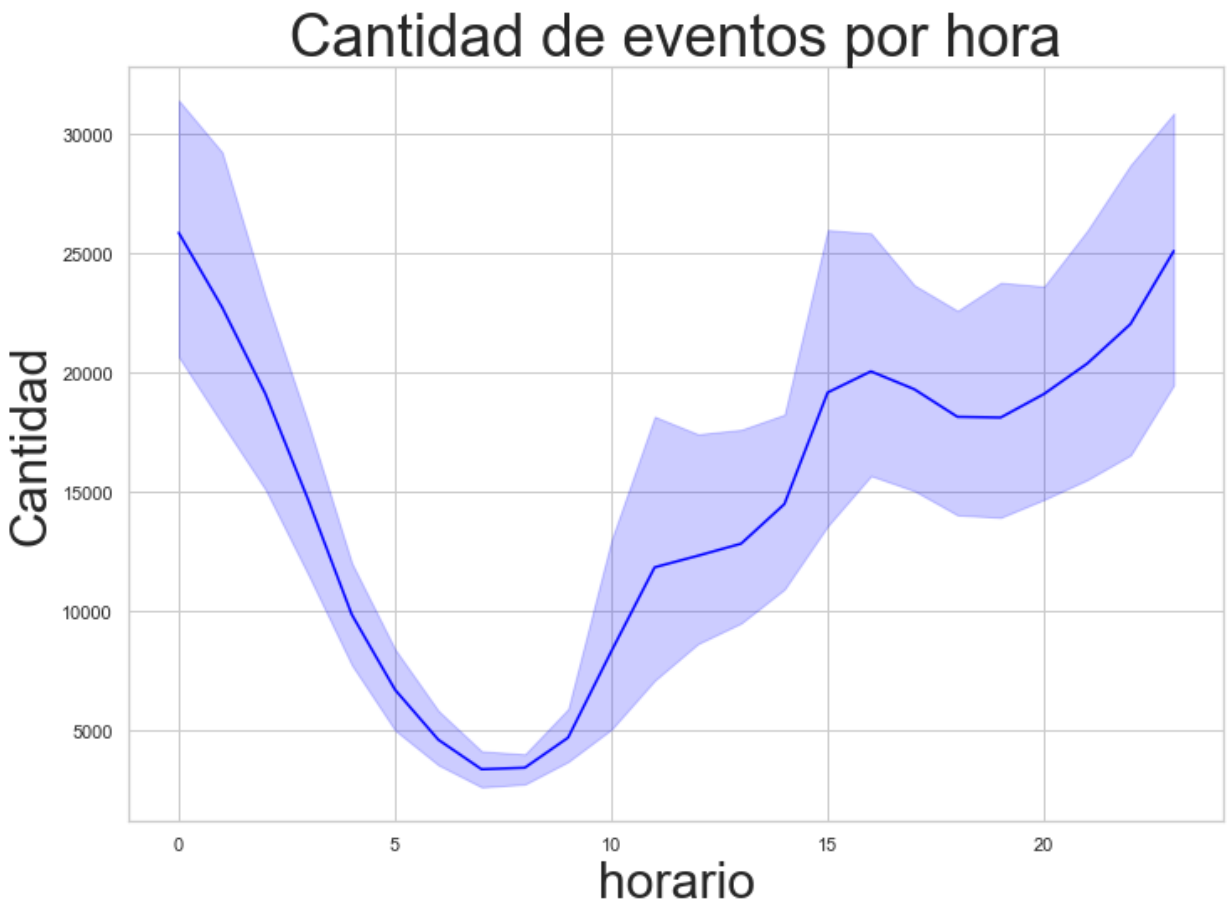
Aquí la idea es ver la distribución que hay entre las subastas y las horas del día, a sabiendas que el set de datos sólo abarca una semana de información. Lo que el gráfico nos muestra es que hay un margen bastante parejo a partir del mediodía hasta la noche y es ahí donde empieza un pico hasta las primeras horas de la madrugada, para luego tener una caída abrupta en el horario de las primeras horas de la mañana. Si bien este comportamiento es entendible, nos resultó interesante entender cuál es la hora que mayor interacción tiene o hay para los usuarios.

3.5 - ¿Cómos se distribuyen los clicks a través de las horas del día?



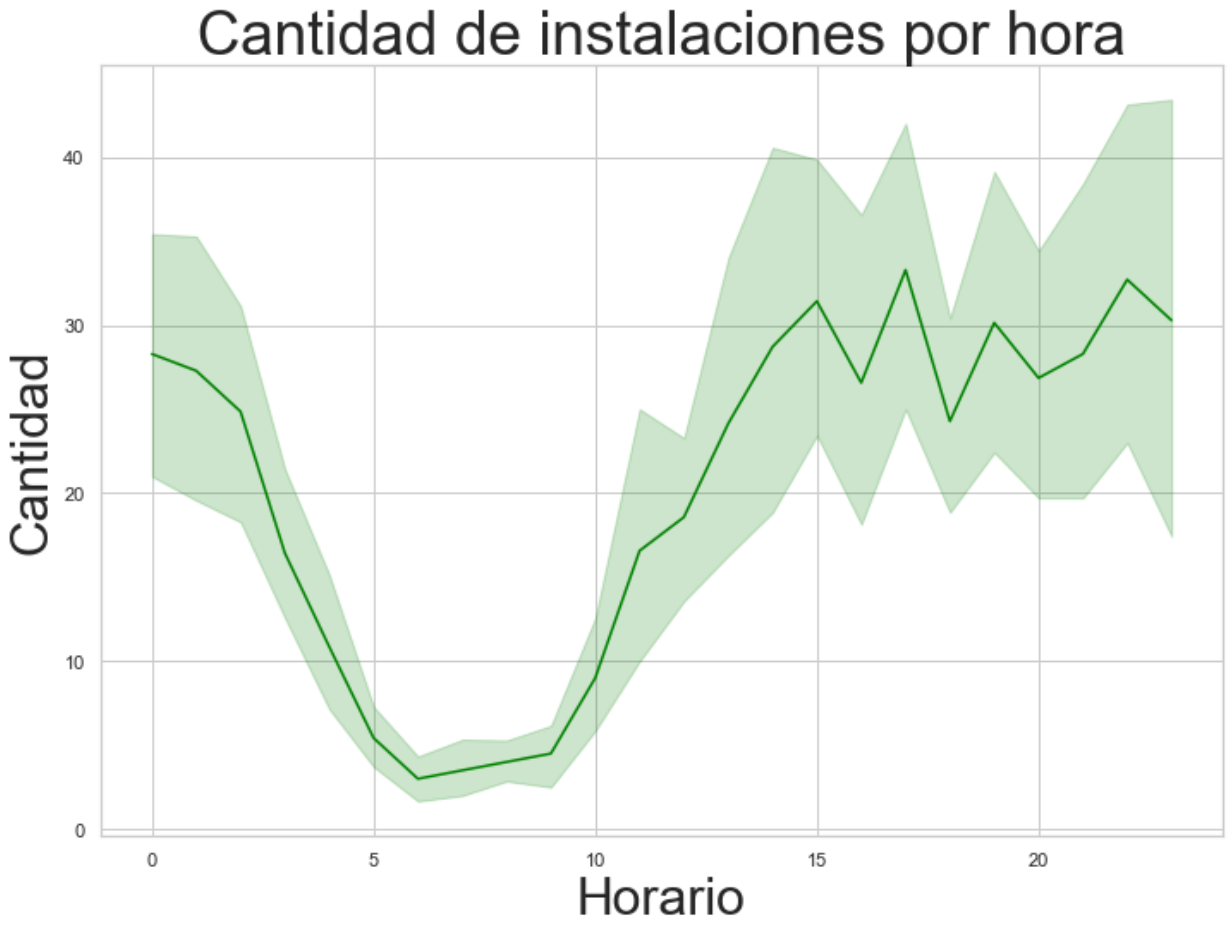
A diferencia de la visualización anterior, aquí uno puede ver que los clicks se distribuyen de una manera mucho más pareja a lo largo de todo el día, para pegar un salto vertiginoso en la madrugada, esto marca cierta uniformidad a lo largo del día para los clicks que se reciben.

3.6 - ¿Cómos se distribuyen los eventos a través de las horas del día?



En este caso para los eventos vemos una distribución muy similar a la que encontramos para las subastas, donde tenemos un bajón muy importante en las primeras horas de la mañana y luego comienza a crecer paulatinamente hasta llegar al pico que se da en las primeras horas del a madrugada. Esto nos deja bien en claro cuáles son los horarios que tienen mayor interacción por parte de los usuarios.

3.7 - ¿Cómos se distribuyen las instalaciones a través de las horas del día?



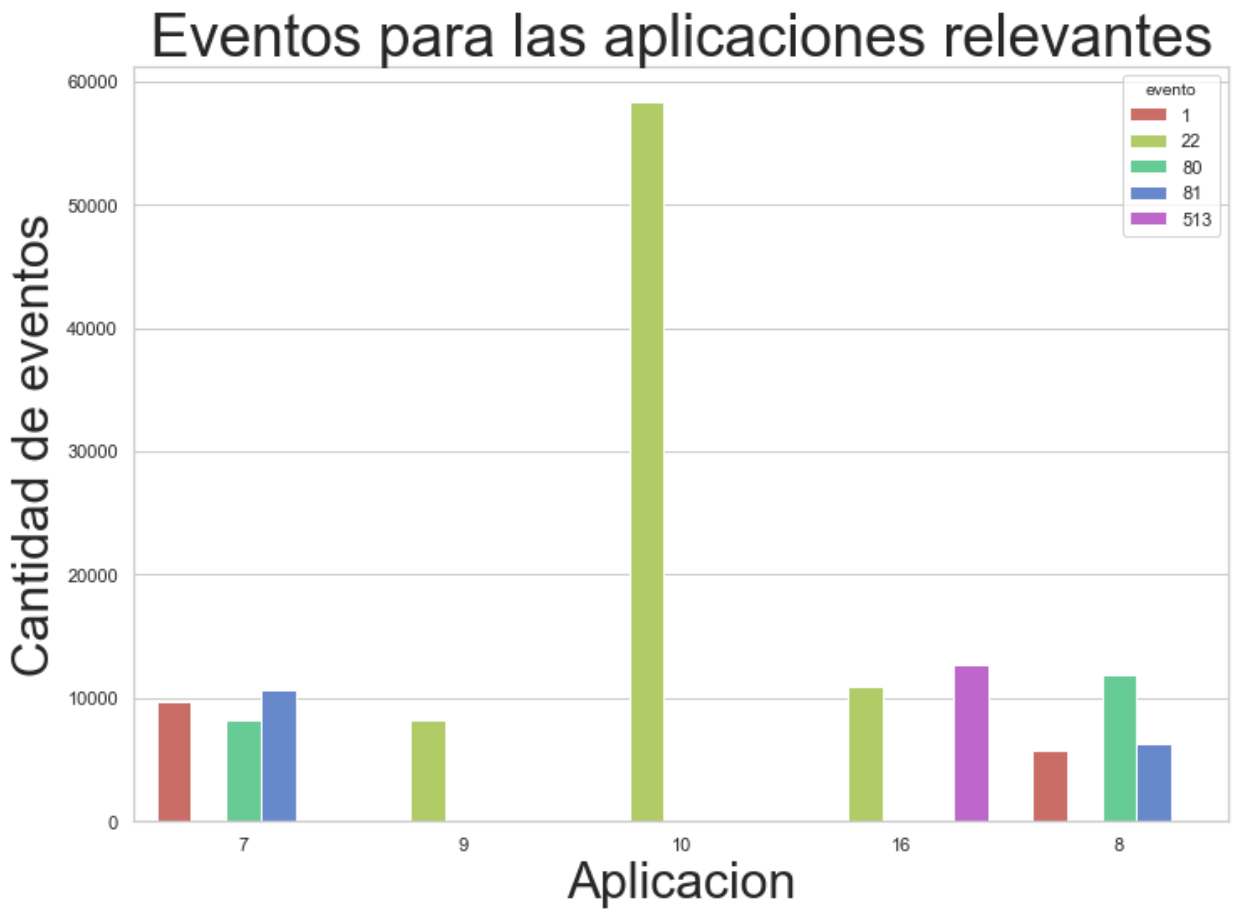
Ahora bien, sobre las instalaciones realizadas vemos que la distribución es mucho más pareja luego del horario del mediodía, ahí se mantiene con pequeñas subas y bajas pero dentro de un mismo rango, aunque vale aclarar que la cantidad de aplicaciones instaladas es mucho menor que las del resto de datos, sean subastas, eventos o clicks. Sin embargo, esta visualización nos da la pauta de que esta actividad tiende a ser pareja y que no hay un horario que predomine, salvo el de las primeras horas de la mañana, que por el contrario es en el que se ve poca interacción en todo sentido.

4 - Análisis sobre cantidades en general

4.1 - ¿Cuáles son las aplicaciones más relevantes en el marco de los eventos con más apariciones?

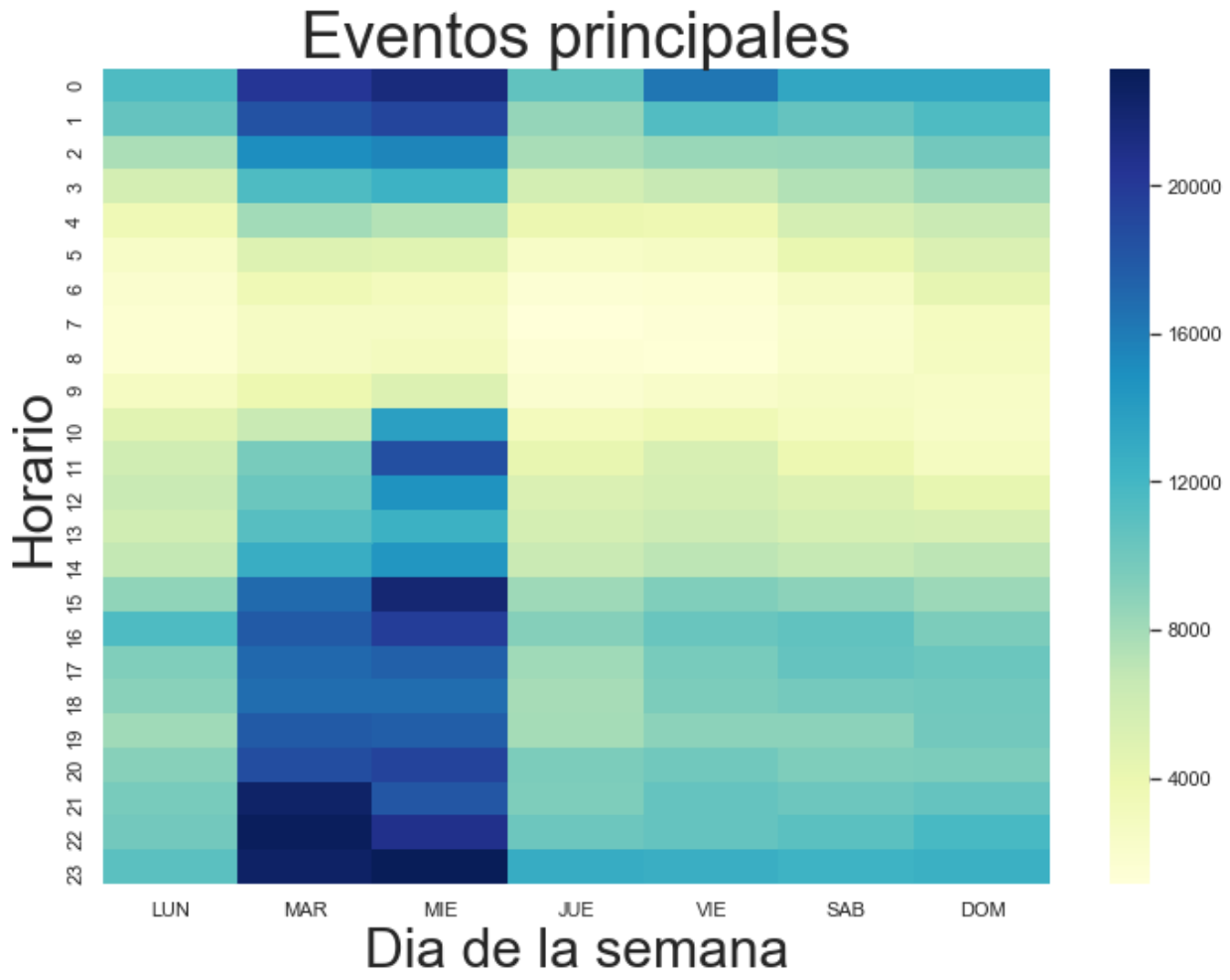
En este caso lo que buscamos entender es cómo se relacionan los eventos con las aplicaciones. En particular, de los eventos que más apariciones tienen, buscamos cuáles son las aplicaciones que más los han empleado, y sobre esto visualizamos cómo se distribuyen esas cantidades. Dado que para un evento en particular hay ciertas aplicaciones que los ejecutan, intentamos ver en qué cantidad sucede esto. Podemos ver que no existe una uniformidad con respecto a los eventos, si no que el resultado se encuentra bastante distribuido, sin existir un evento en común que sea empleado por todos.

4.2 - ¿Cuáles son los eventos más registrados en el marco de las aplicaciones más populares?



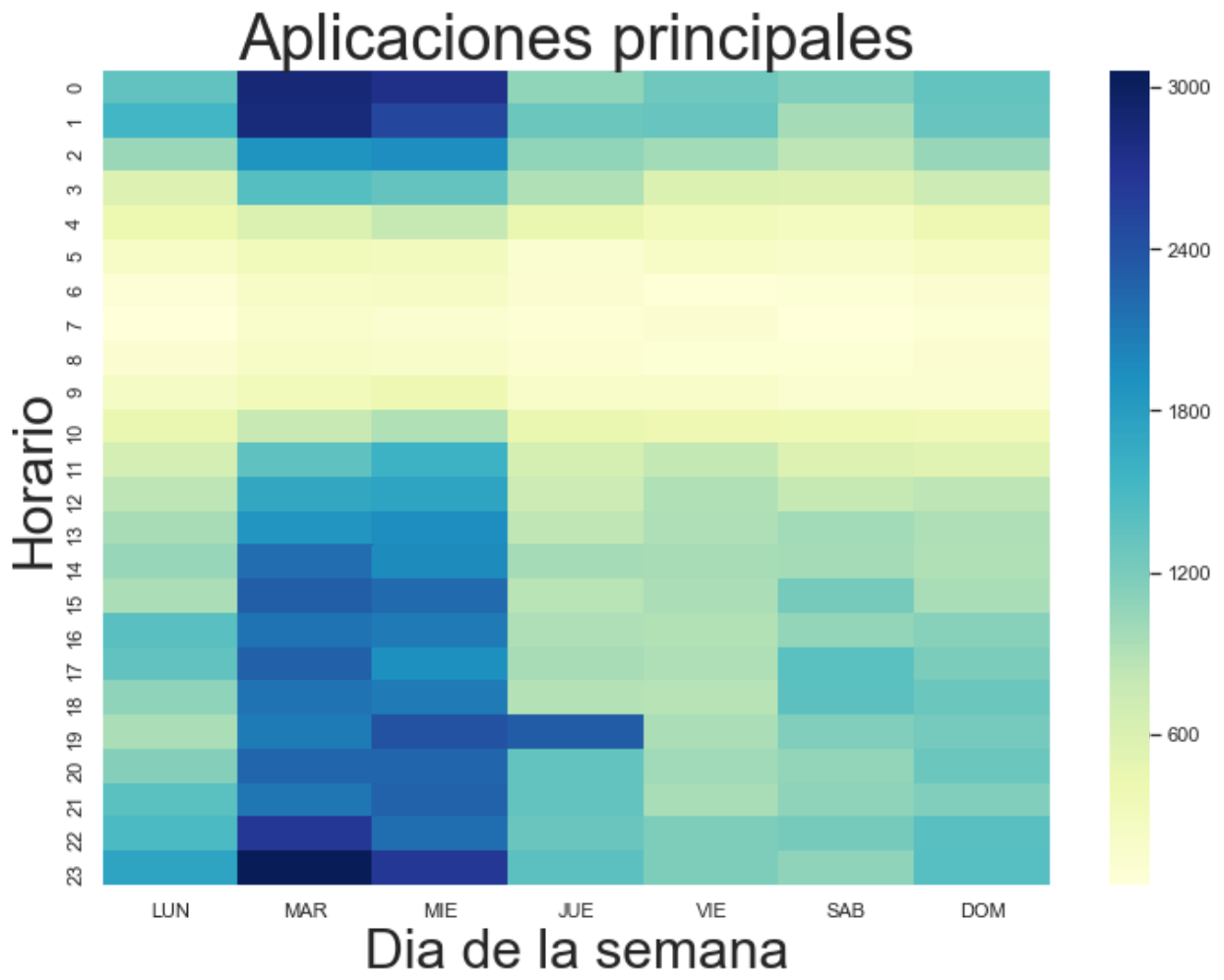
Análogo al análisis anterior, lo que buscamos con el siguiente gráfico es ver por ejemplo cómo se distribuyen dentro de las aplicaciones más importantes los eventos que más suelen suceder. En este caso vemos que para ciertas aplicaciones hay un evento predominante, que sería el 22, mientras que para la aplicación 7 y 8 hay un grupo de tres eventos repartidos de manera similar. Vale aclarar que al tener los datos anonimizados, utilizamos los identificadores pero se da por entendido que de tener los nombres el resultado podría ser un poco más enriquecedor.

4.3 - ¿Cómo se distribuye la ejecución de los eventos principales a lo largo de la semana?



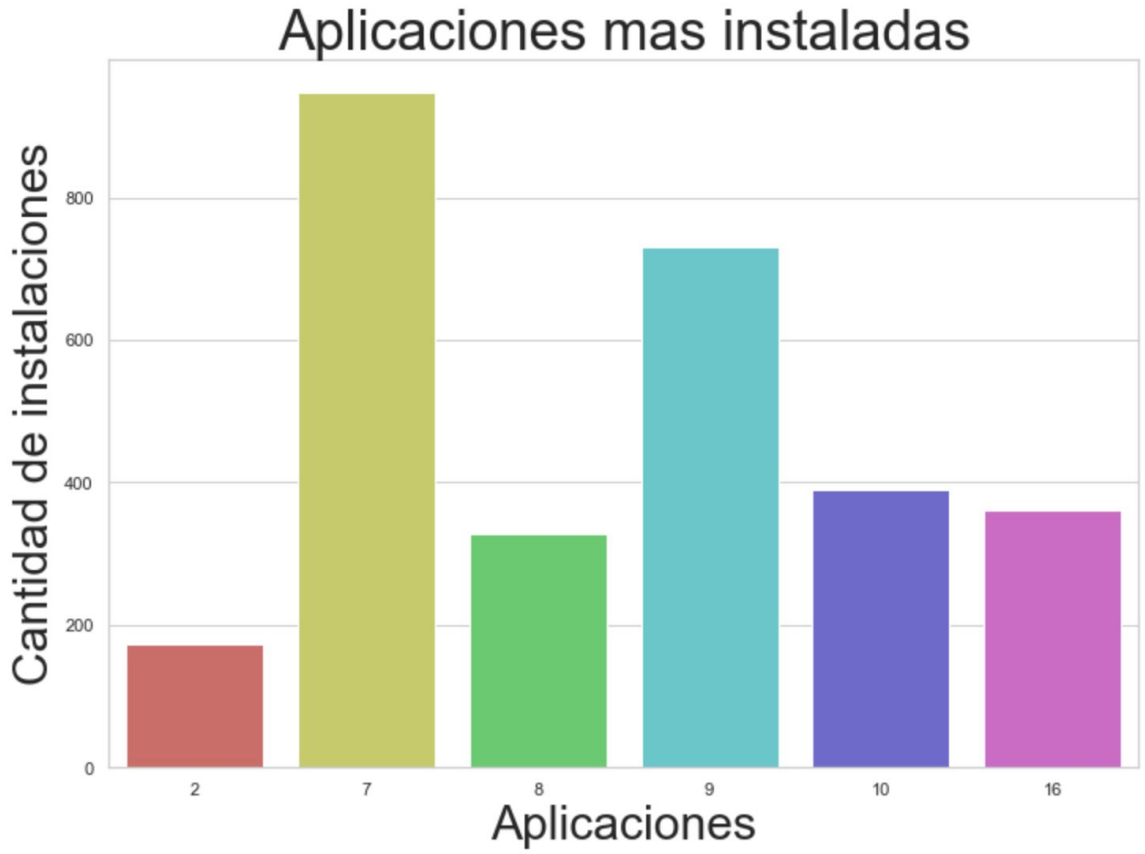
Habiendo tomado los cinco eventos principales, es decir aquellos que engloban la mayor cantidad de actividad, podemos ver con este mapa de calor que los días de la semana que tienen una mayor interacción son los días martes y miércoles en el horario de la tarde, incrementándose hacia la noche y madrugada, lo que nos lleva a remarcar la importancia de prestar atención a estos días y estos horarios.

4.4 - ¿Cómo se distribuye la ejecución de los eventos para las aplicaciones más utilizadas a lo largo de la semana?



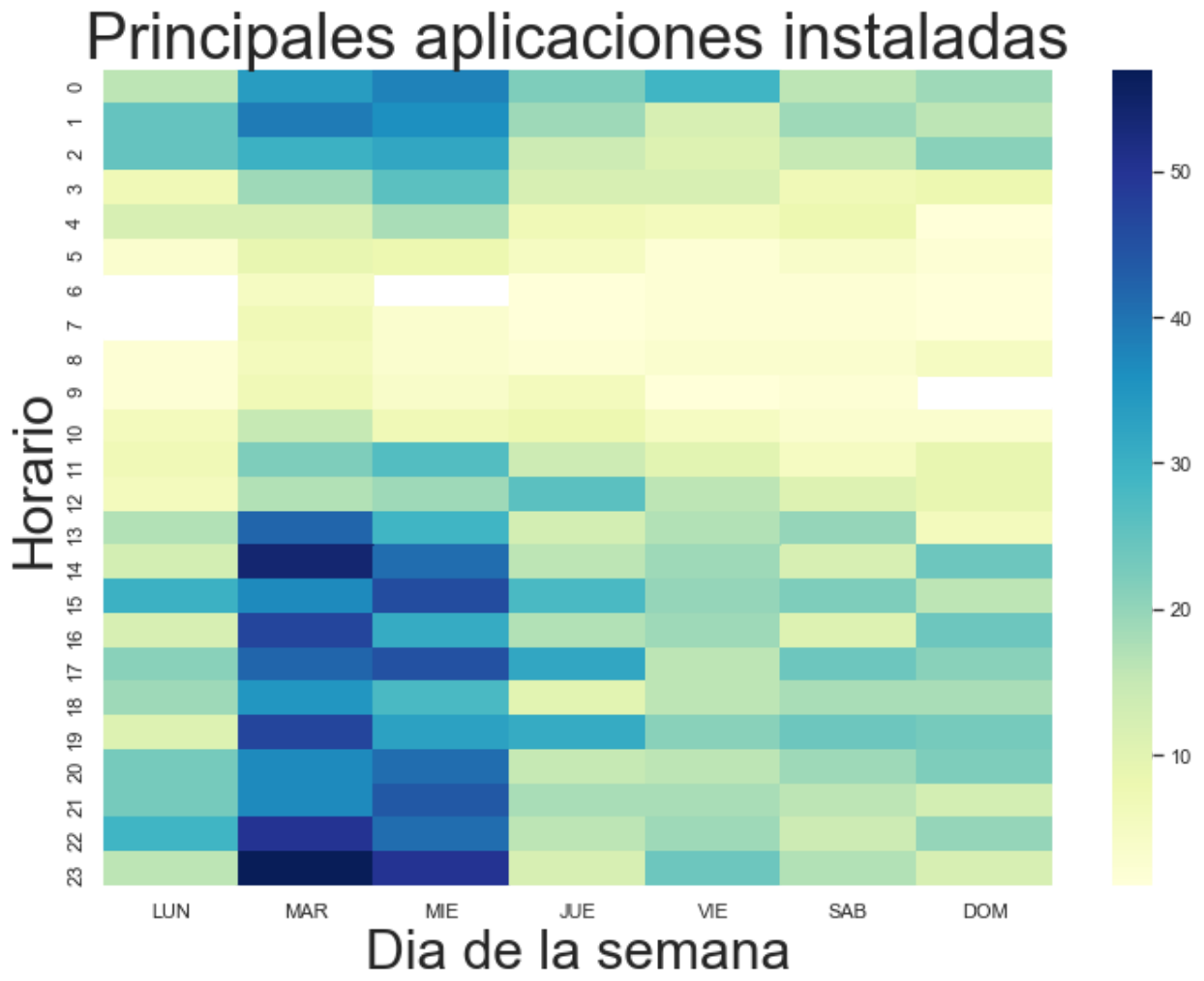
En este gráfico, que a priori es análogo al anterior, lo que buscamos entender es cómo es el comportamiento de los usuarios en relación a las aplicaciones que más uso tienen, es decir las aplicaciones más relevantes. El análisis realizado arroja que el comportamiento de los eventos es bastante similar a lo que se venía viendo en general, donde tenemos dos días con gran relevancia y con los horarios de la noche y madrugada como prioritarios, lo que refuerza la importancia de estos datos.

4.5 - ¿Cuáles son las aplicaciones más instaladas?



Otro dato que resulta de interés, es cuáles son las aplicaciones más instaladas. Dado que los datos se encuentran anonimizados, utilizamos los ID para la visualización, pero mediante la transformación inversa podría saberse de qué aplicaciones se tratan. Vemos que se destacan las aplicaciones 9 y 7, ya en segundo plano con aproximadamente la mitad de installs tenemos a las 8, 10 y 16.

4.6 - ¿Para las aplicaciones más instaladas, cómo se distribuye la cantidad en la semana?

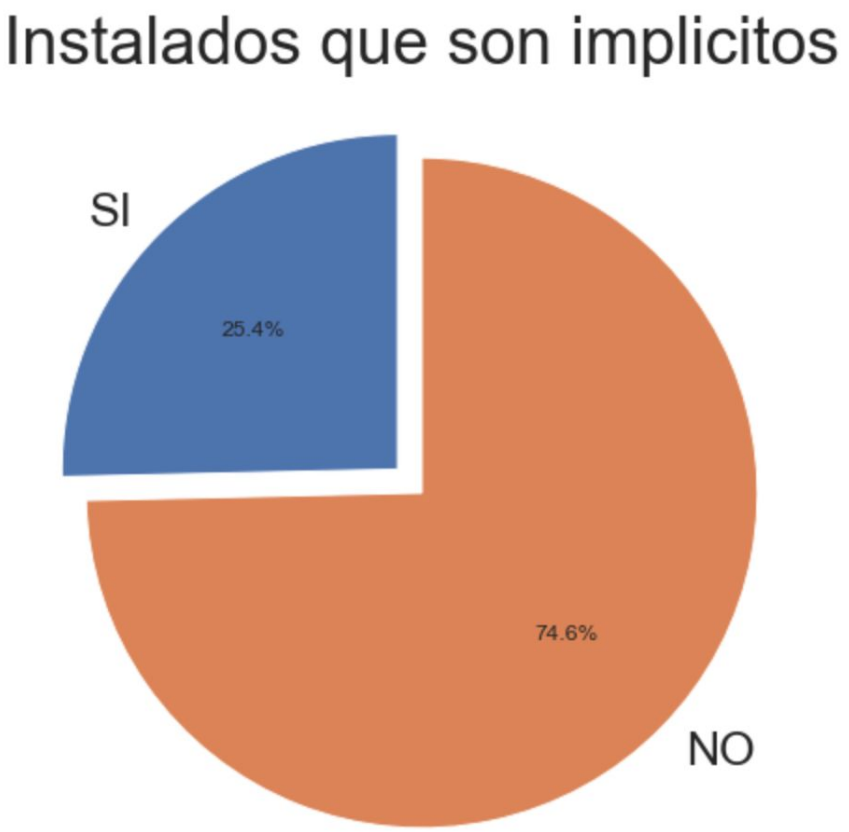


Partiendo de un universo menor, que son los datos de las aplicaciones instaladas, hacemos un análisis análogo a fin de ver cuáles son los momentos de la semana donde los usuarios son más propensos a instalar aplicaciones. Como podemos observar, no difiere de lo que esperábamos en base al análisis que veníamos haciendo, puesto que los días más influyentes son el martes y el miércoles, como así también el hecho de que los horarios más importantes son los de la tarde, noche y madrugada, siendo esto un dato relevante, que representa cuando es más importante presentarse al usuario.

4.7 - ¿Cuál es la proporción de aplicaciones instaladas que son atribuidos a Jampp?

En base a la información brindada en el set de datos, vemos que todos los casos analizados no son atribuidos a Jampp, esto nos deja poco margen para explorar o indagar sobre esta parte de la información. Puede que sea algo a mejorar, puede que sea una cuestión de la porción o tipos de datos que se brindaron para el análisis, pero a priori preferimos al menos hacer una pequeña mención sobre esto.

4.8 - ¿Cuál es la proporción de aplicaciones instaladas que son implícitas?



Entendemos que si la instalación es implícita significa que esta fue realizada por un dispositivo que no se ha instalado de acuerdo con la plataforma de seguimiento, pero como podemos ver casi las tres cuartas partes no son implícitas lo que fortalece el hecho de ser instalado a través de la plataforma.

4.9 - ¿Cuáles son los anunciantes (advertisers) con mayor cantidad de clicks?

Haciendo el análisis para responder esta pregunta, descubrimos que el 99% está relacionado al mismo anunciante, lo que nos deja muy poco margen para relacionar o aprovechar este dato, que a priori podría ser sumamente interesante, con lo cual solo decidimos mencionar el análisis básico que hicimos y la magnitud que ese anunciante tiene para este set de datos.

4.10 - ¿Cuáles son los advertisers con mayor cantidad de installs?

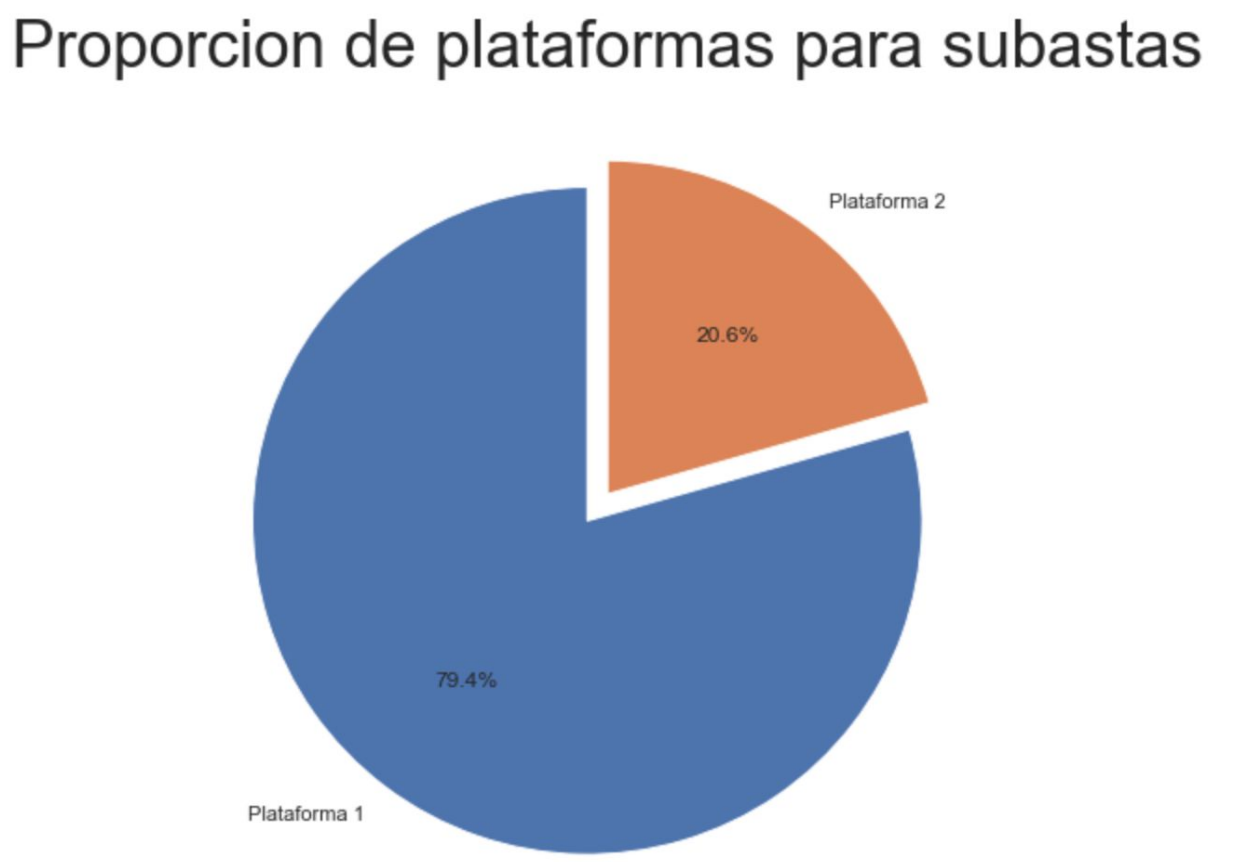
Para este caso, el análisis realizado es análogo al de la pregunta anterior.

5 - Análisis sobre las subastas

5.1 - ¿Cuáles son los países más populares donde se originan las subastas?

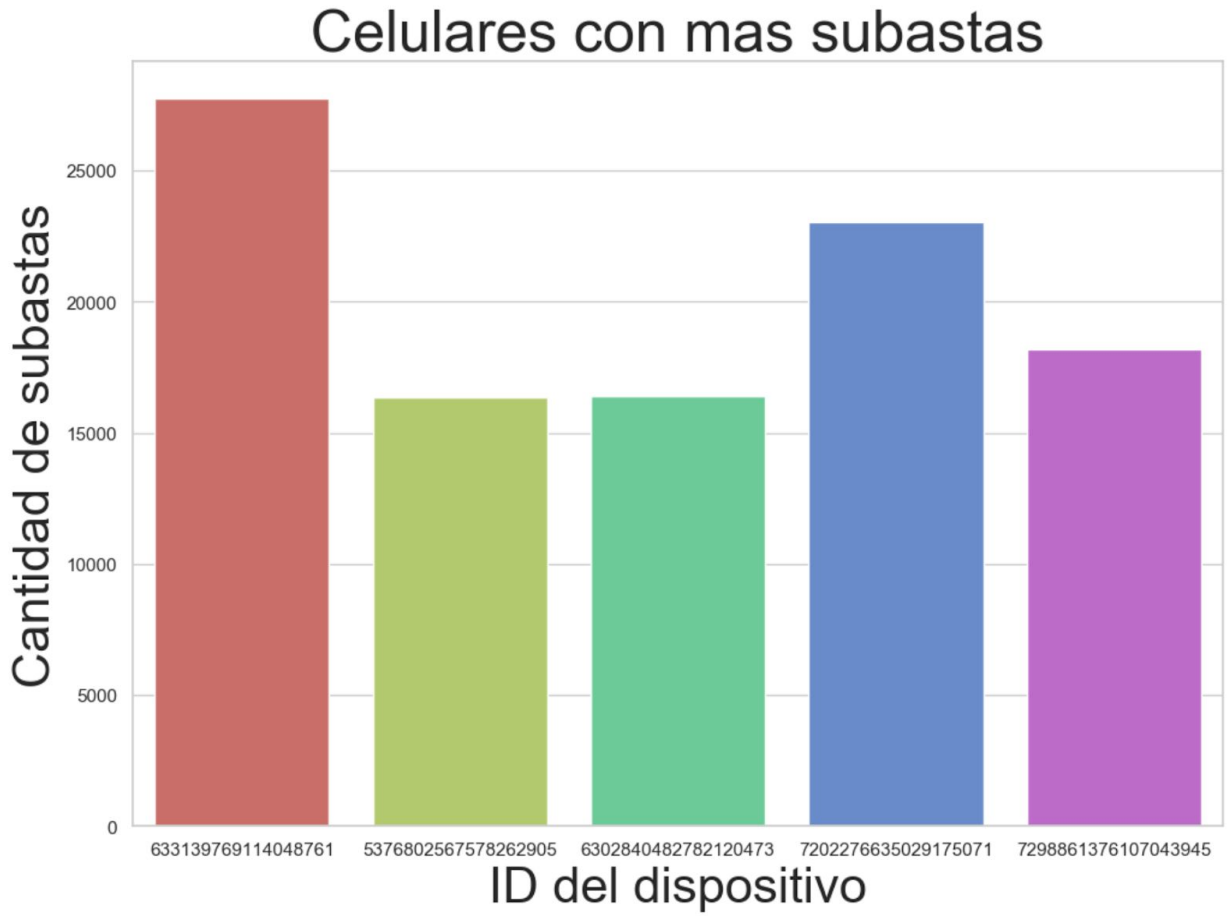
Aunque pudiésemos mostrar el detalle de los nombres que representan los países, notamos que todos los países tienen el mismo código con lo cual entendemos que esta información no es muy relevante o no brinda un plus al análisis hecho, dado que los datos vienen todos del mismo país.

5.2 - ¿Cúal es la proporción de plataformas usadas para las subastas?



A pesar de tener los datos anonimizados, el gráfico presentado en esta sección nos muestra un claro predominio por parte de una de las plataformas a la hora de llevarse a cabo una subasta, esto muestra la importancia que tiene esa plataforma.

5.3 - ¿Cuáles fueron los celulares sobre los que se hicieron más subastas?

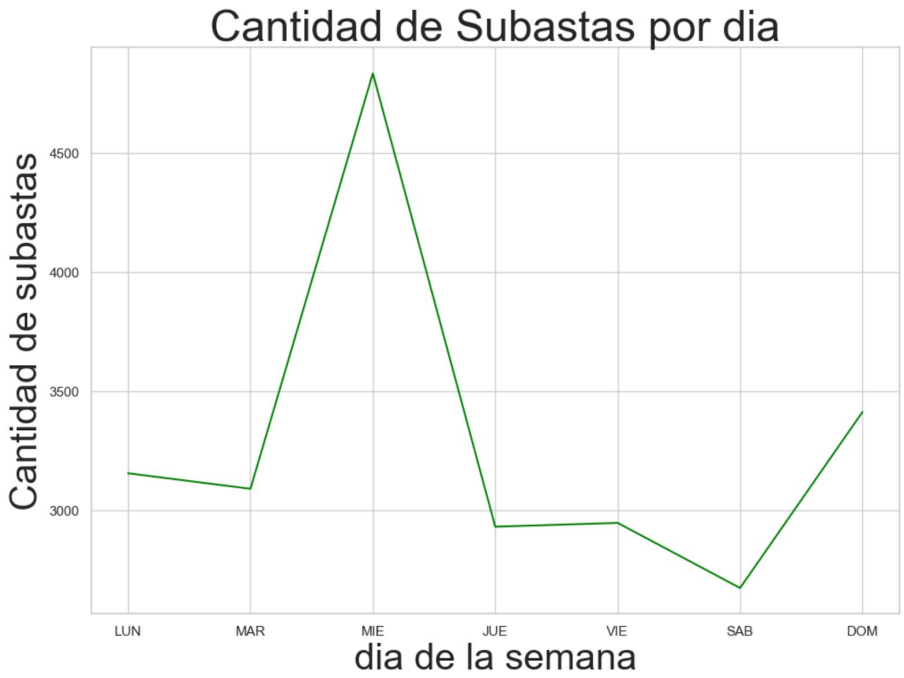


Un aspecto interesante que investigamos fue cuáles fueron los dispositivos que tuvieron la mayor cantidad de subastas. Inicialmente planteamos la pregunta respecto a la cantidad de installs, pero analizando el dataset descubrimos que los que tienen mayor cantidad de installs cuentan con sólo 4 instalaciones en la ventana de tiempo proporcionada por Jampp, por lo que decidimos verlo respecto de las subastas.

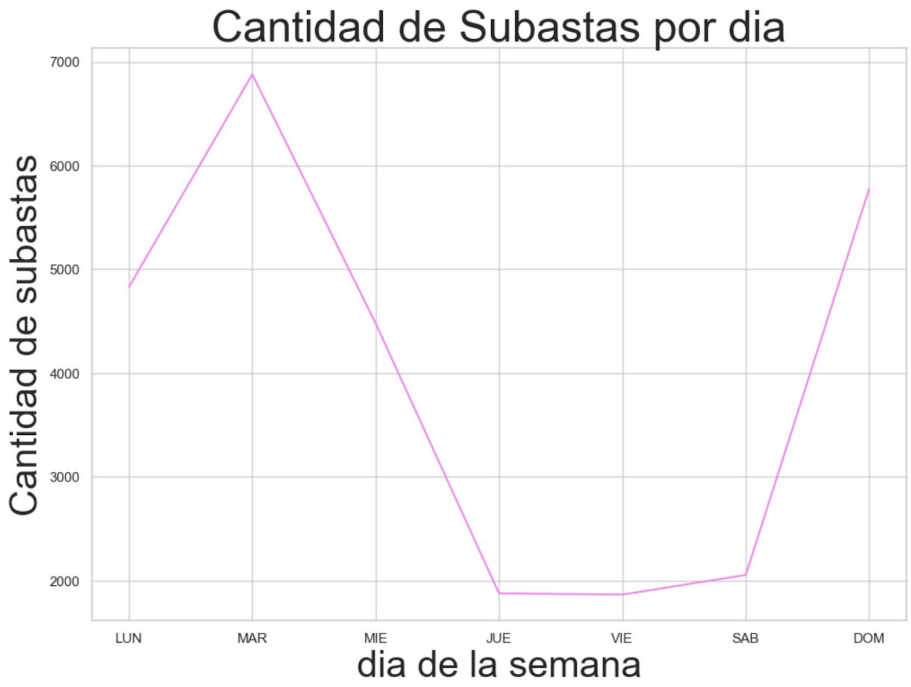
En el gráfico presentado arriba se puede apreciar que hay dos dispositivos que tuvieron más de 20000 subastas. Si tomamos una ventana de 8 días y contabilizamos las 24 horas de cada día, este número significa a priori que en promedio hubo una subasta cada aproximadamente 35 segundos para esos dispositivos.

Siendo este número por demás interesante, y ahondando en los dos IDs que registran la mayor actividad, graficamos la distribución de esas subastas según los días de la semana, obteniendo los siguientes gráficos:

Para el ID 633139769114048761



Para el ID 7202276635029175071



En ambos casos podemos apreciar días con una gran cantidad de actividad, mientras que en los demás el dispositivo es mucho menos activo.

Sin embargo, un número de 7000 subastas por día implica en promedio un evento cada 12 segundos, mientras que unas 2000 sugiere en promedio uno cada 28. Estos números reflejan a nuestro parecer una cantidad de subastas que no es normal, siendo importante investigar más a fondo su causa. Si bien no podemos afirmar que se trate de actividad fraudulenta, podría bien ser el caso. Por otro lado, en la clase del 15/04/2019 se mencionó que algunos eventos en el dataset eran internos de Jampp, lo que podría ayudar a comprender el comportamiento observado. El hecho de tener los datos anonimizados no nos permite explorar más a fondo estas posibilidades.

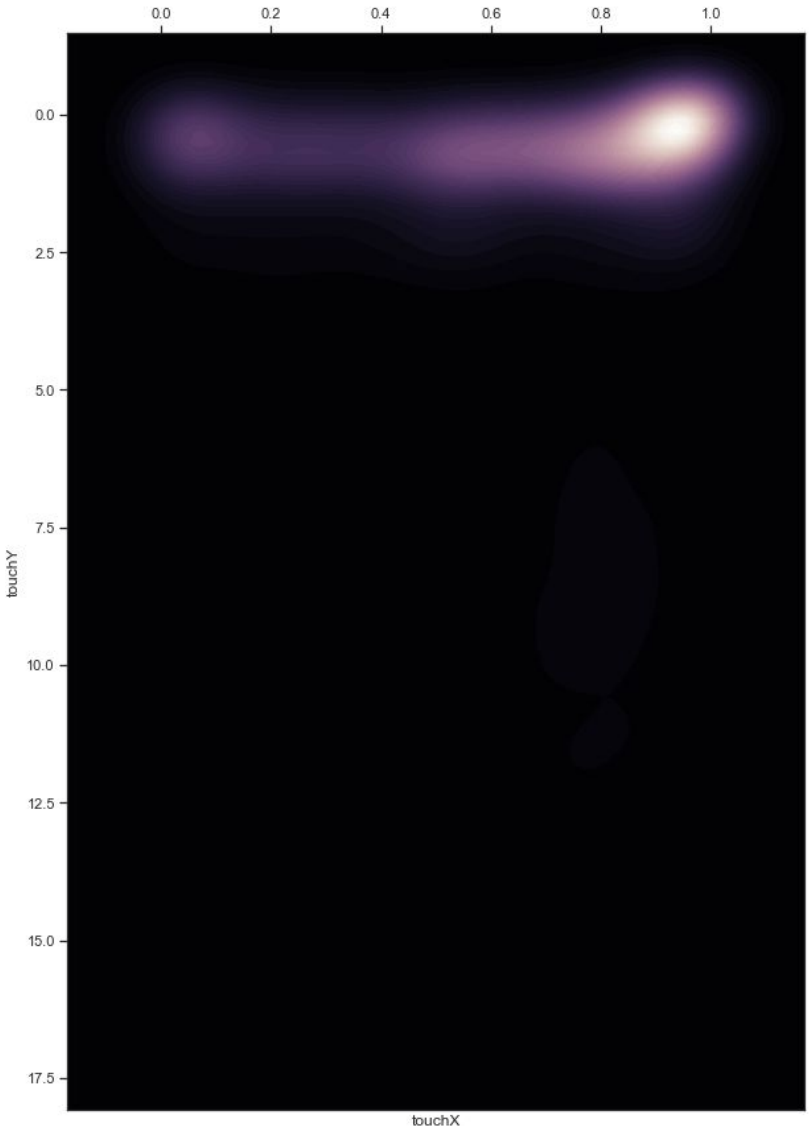
5.4 - ¿Qué proporción de las subastas terminan en un install?

SIn necesidad de mostrar gráfico alguno llegamos a la conclusión de que solo un 0.06% de las subastas terminan en una instalación, lo que representa un margen muy pequeño como para que tengan sentido expresarlo en un gráfico. De todas maneras, el dato muestra la importancia de afinar las subastas, para tratar de mejorar ese porcentaje, que entendemos a priori sería de gran provecho para Jampp.

6 - Análisis sobre el comportamiento de los usuarios

6.1 - ¿Cuál es la distribución de clicks en la pantalla?

Decidimos investigar para los eventos de click, en qué lugar físico de la pantalla los usuarios interactúan con las publicidades. Se asume que el borde superior izquierdo es el (0,0).



Si bien los datos incluyen seguramente muchos tipos de pantallas y de publicidades, en ambos casos de distintos tamaños y proporciones, podemos igualmente sacar algunas conclusiones útiles. El hecho de que el dominio de ambas dimensiones se encuentre acotado es indicativo de que los datos se encuentran normalizados, como se aclaró en la clase del 15/04/2019.

En primera instancia podemos ver que los clicks se concentran sobre el margen superior. En particular, hay una concentración muy fuerte de estos eventos sobre la derecha. Aunque no conocemos los tipos de publicidad que se muestran, esta información puede resultar importante para entender donde se concentra la actividad de los usuarios y así optimizar el diseño de las impresiones mostradas.

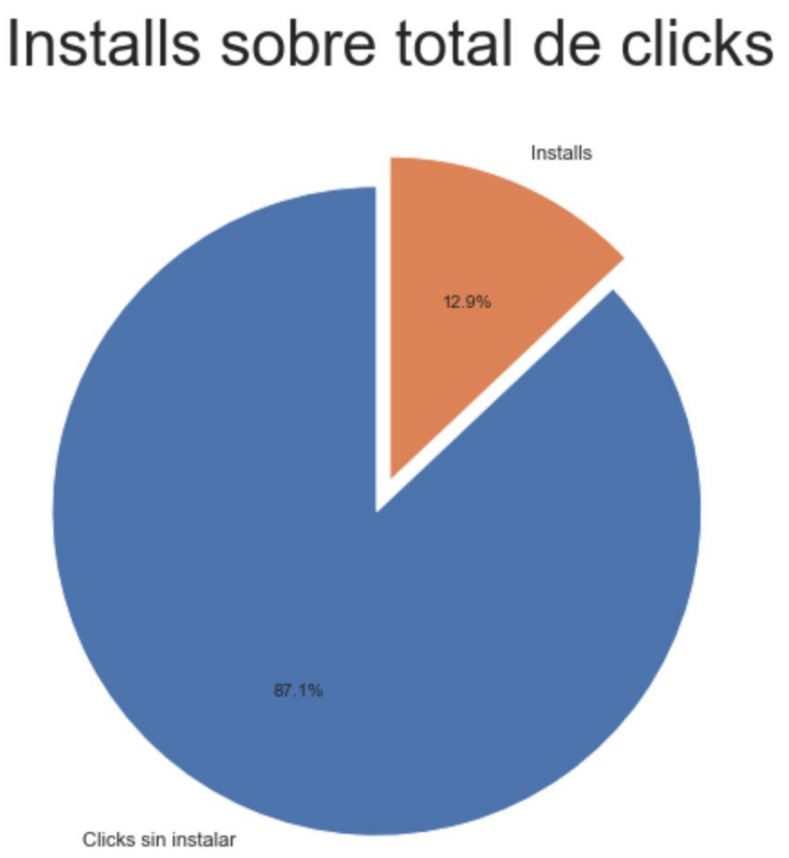
6.2 - ¿Cuánto tiempo tarda un usuario en hacer click en la pantalla?

Otro dato que resulta interesante explorar es la cantidad de tiempo que demora un usuario en hacer click en un anuncio.

Se descartaron los valores mayores a 60 segundos. Sobre la cantidad de clicks disponibles, vemos cómo el grueso de los usuarios interactúan con la publicidad en los primeros 15 segundos, por lo que se podría utilizar esta información para optimizar los costos de las publicidades a mostrar.

Entendiendo que a mayor tiempo menos relevancia tiene por parte del usuario, es decir que de hacer click (sea bien para instalar o para cerrar) el grueso está en los primeros 10 segundos y luego el valor es muy bajo y se mantiene en esa misma línea, con lo cual en un análisis apresurado uno podría pensar que publicaciones o publicidades extendidas en el tiempo son menos propensas a recibir interacción.

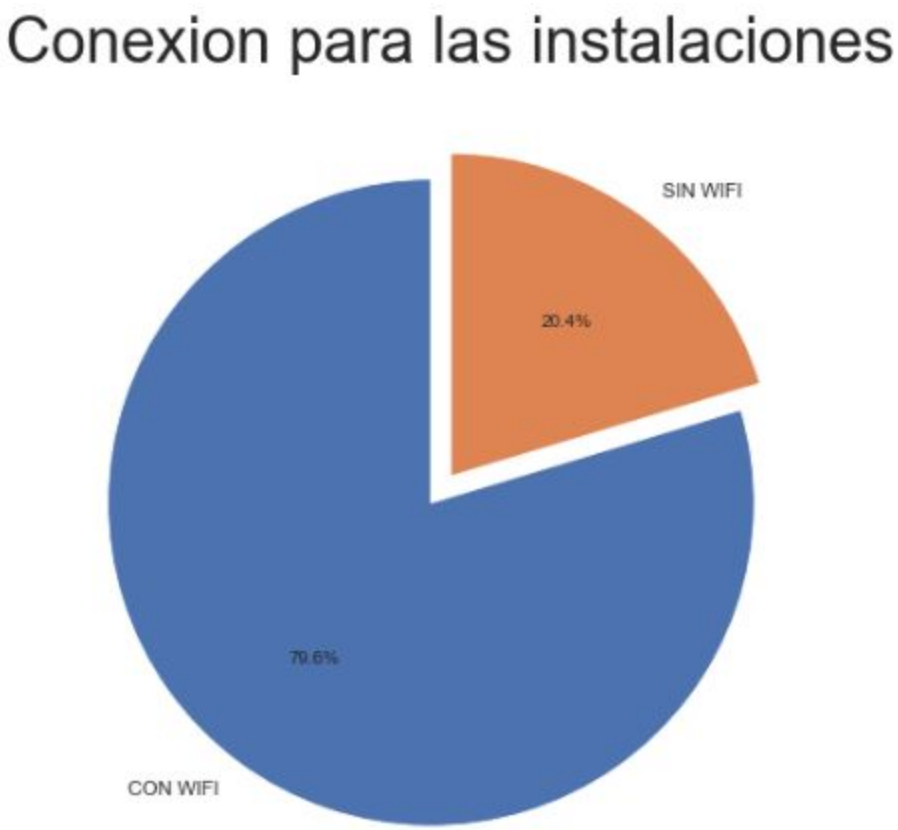
6.3 - ¿Cuál es la proporción de installs sobre el total de clicks?



Este gráfico nos muestra de manera certera la poca cantidad de veces que se instalan las aplicaciones a pesar de que se haga click, lo que viene de la mano con la relación porcentual que más arriba habíamos obtenido sobre las subastas y las aplicaciones instaladas. En este caso puntual podemos ver que a pesar de hacer click es

sólamente un 12% de las veces que se instalan las aplicaciones, esto puede deberse a que lo que ofrece la publicidad no es del gusto del usuario, o bien que el usuario tiende a errar el click al intentar cerrar la propaganda. De todas maneras, entendemos que si las publicidades fuesen más certeras esta proporción aumentaría.

6.4 - ¿Cuál es la proporción de installs realizados con conexión Wi-Fi?

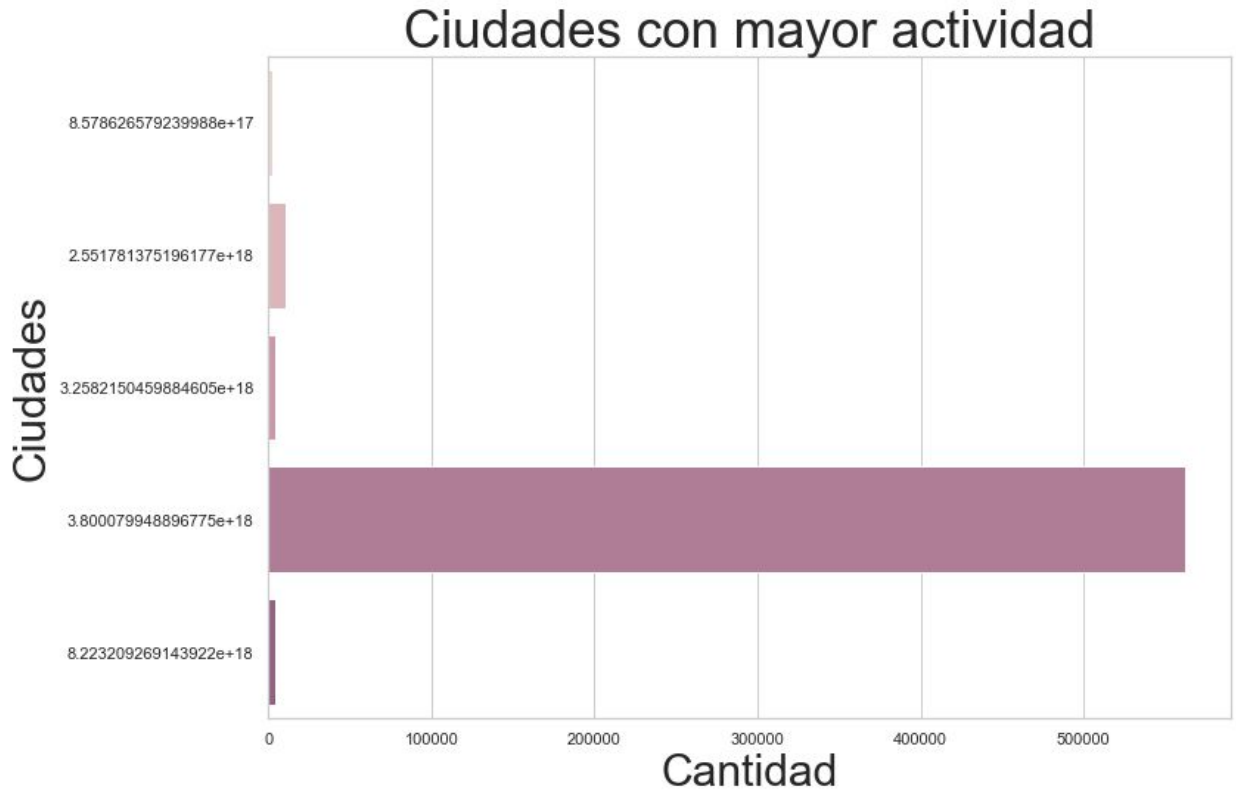


Si uno le pregunta a cualquier persona acostumbrada a usar un smartphone e instalar aplicaciones, de seguro tendremos la respuesta de que a la hora de instalar una aplicación lo tratará de hacer si tiene conexión wifi, lo cual es más que entendible. Este gráfico refleja justamente eso, aunque vale aclarar que hay muchos datos que no tenían un valor definido y decidimos omitir esa información a la hora de preparar la visualización.

7 - Análisis demográfico y características generales de los datos

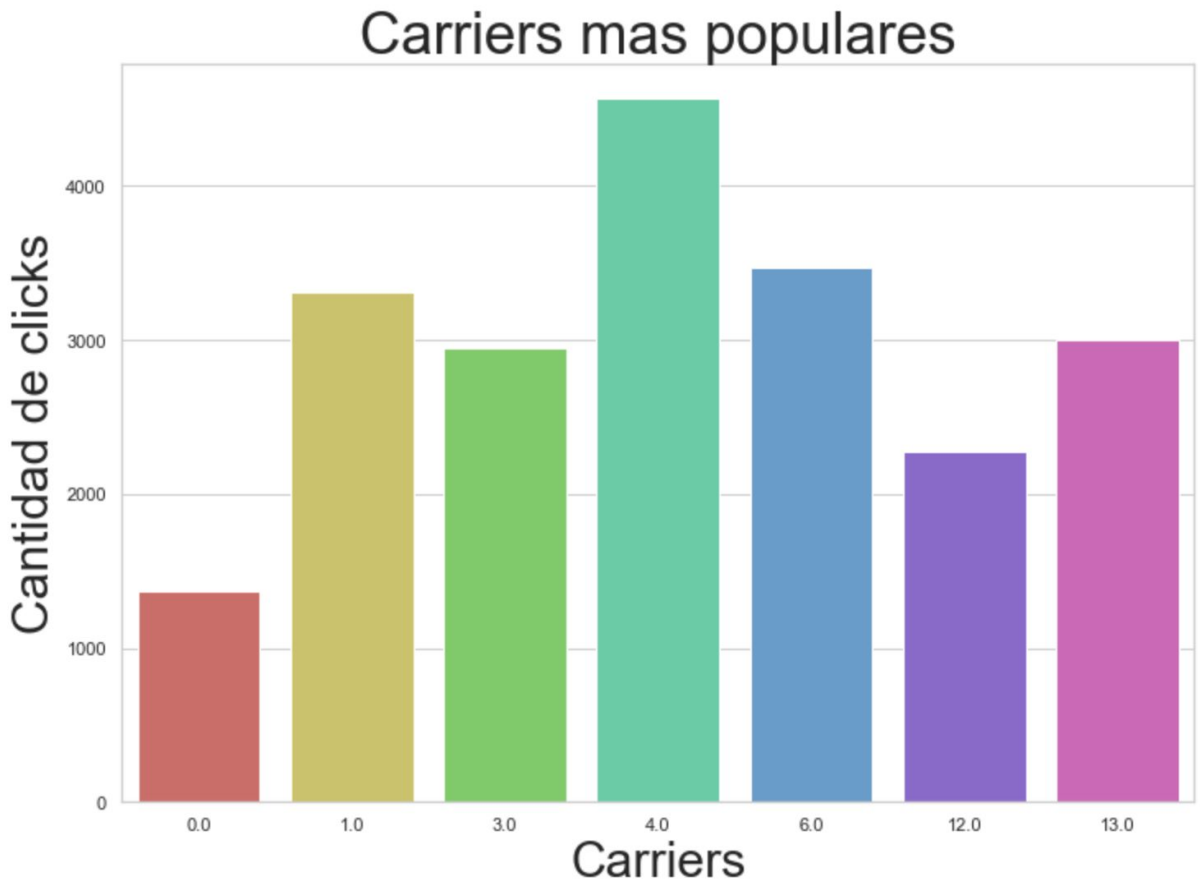
En esta sección analizaremos los datos demográficos y las características generales que nos brindan el set de datos. La idea es encontrar patrones que nos indiquen qué lenguajes, qué ciudades, qué características de los dispositivos, etc, hacen más propensa una posible instalación, o al menos saber cuáles tienen una mayor proporción en la interacción y actividad normal de los usuarios. Para esto lo primero a considerar es que vemos que hay un solo país a disposición para analizar, con lo cual optamos por no incluir un gráfico sobre esto y solo hacer una pequeña mención.

7.1 - ¿Cuáles son las ciudades más populares en las que se registran eventos?



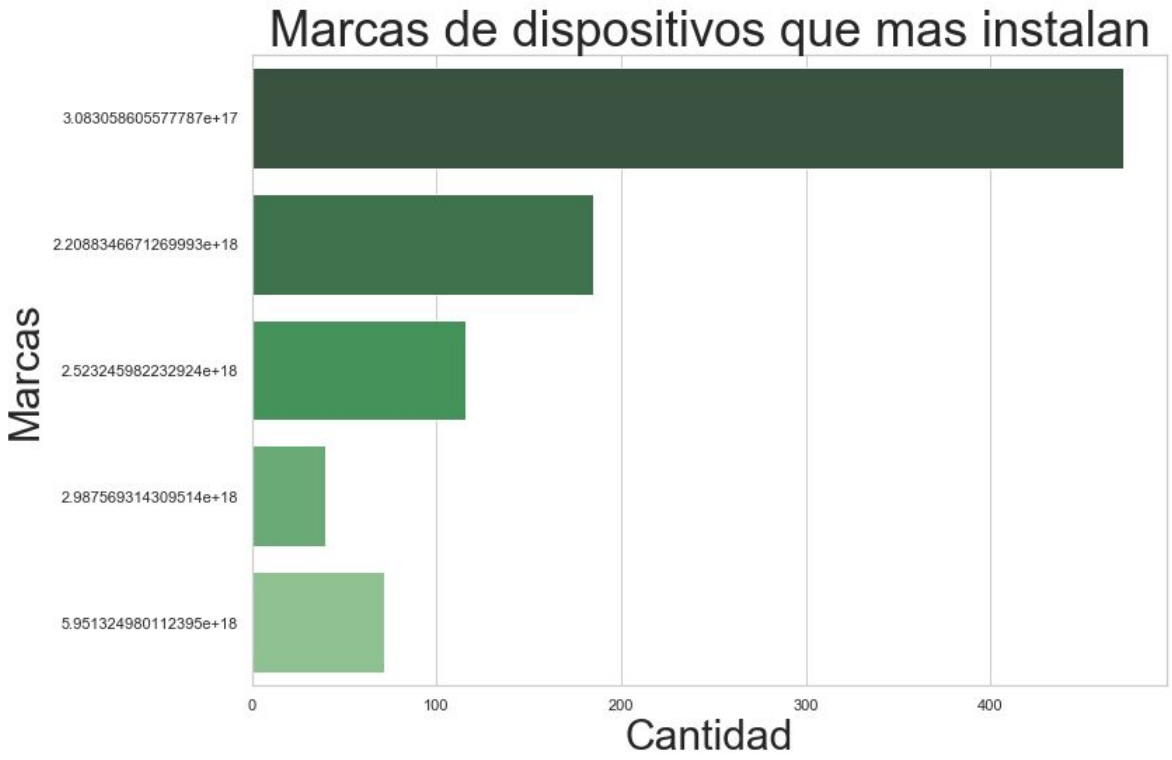
Al menos con el set de datos actual podemos ver que la información está muy restringida y la información de las ciudades es muy acotada, o más bien hay una ciudad que predomina ampliamente sobre las demás. Sería interesante saber si este resultado es relacionable a la distribución geográfica de población o riqueza del país analizado o responde a otras variables.

7.2 - ¿Cuáles son los carriers más populares?



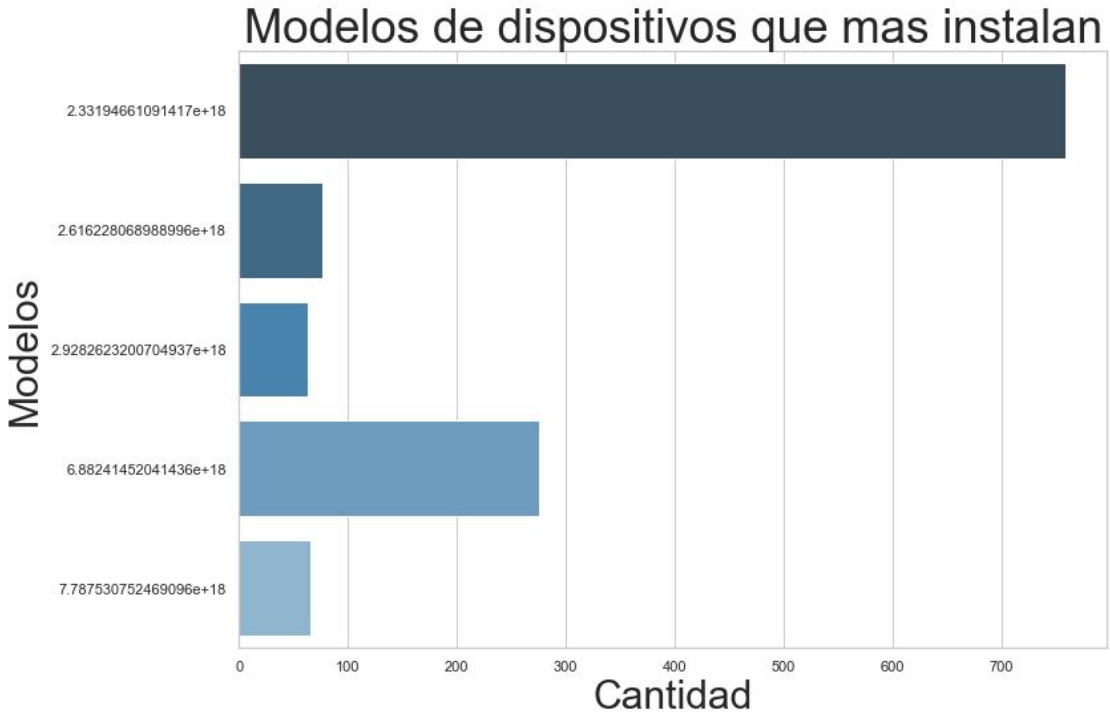
De manera similar a la pregunta anterior, obtenemos una visualización con las prestadoras de servicio (carriers) más populares en el dataset. Esta información nos permite saber cómo se distribuyen los usuarios de la plataforma en el mercado del país analizado.

7.3 - ¿Cuáles son las marcas de dispositivos que más instalaron?



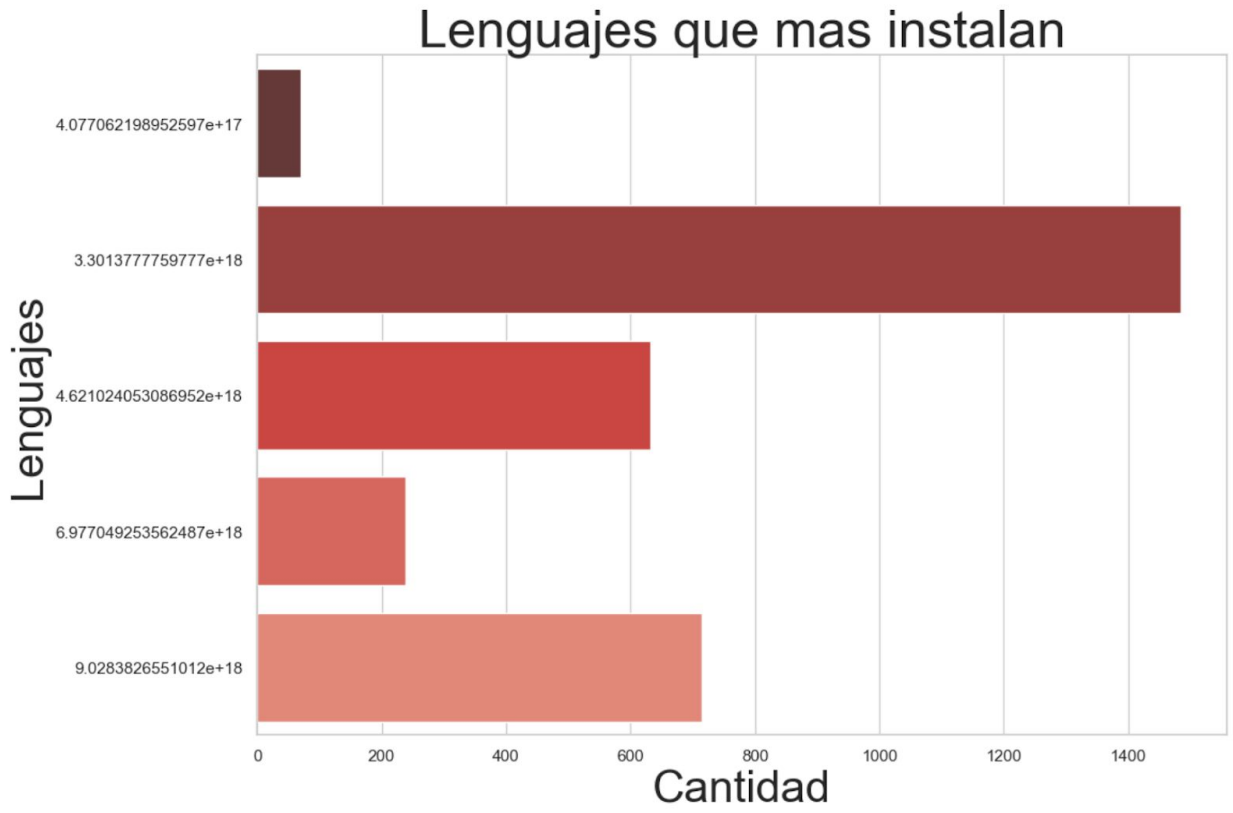
Podemos ver que hay una marca que está predominando en este set de datos, y supera en gran número al resto de las marcas en relación a la cantidad de instalaciones registradas. Esto puede deberse a que sea una marca puntera en el mercado, o bien que el recorte en el set de datos la haya favorecido. Asimismo, podría de tratarse de una marca con modelos que apuntan a un segmento de la población con mayor poder adquisitivo.

7.4 - ¿Cuáles son los modelos de dispositivos que más instalaron?



En este análisis, que es análogo al anterior, vemos que la brecha es aún mayor, donde hay un modelo con una gran cantidad de aplicaciones instaladas, un segundo que ni siquiera llega a la mitad y el resto que está muy por debajo. Siendo que hemos tomado los cinco modelos con mayor cantidad de instalaciones, de todas maneras al estar los datos anonimizados decidimos dejar el dato encriptado aunque con el dato real sería más visual.

7.5 - ¿Qué lenguajes tenían los dispositivos que más instalaron?



Decidimos realizar el análisis para ver cuáles son los idiomas de los dispositivos que predominan a la hora de instalar las aplicaciones, esto nos puede dar una pauta no solo de cuáles son los idiomas predominantes sino de cuáles pueden ser más propensos a una posible instalación.

8 - Conclusiones

A lo largo de este trabajo, hemos estudiado diversos aspectos del set de datos proporcionado por Jampp. Se han analizado los días y horarios en que ocurren los eventos más relevantes. Asimismo, hemos determinado aspectos cuantitativos como las aplicaciones y eventos más numerosos, cuales son las aplicaciones donde más publicidad se muestra, cuales son las apps que más se instalan entre otros. Siguiendo en la misma línea, hemos analizado la información general sobre las subastas, para entender su origen tanto geográfico como de plataforma, cuántas de ellas terminan en un click o un install. Otro aspecto interesante resultó ser el comportamiento de los usuarios, por ejemplo el área de la publicidad donde se registran los clicks, cuanto tiempo tarda en generarse dicho evento y por ejemplo si las instalaciones se realizan con conexión WiFi o no. Por último, estudiamos aspectos demográficos de los datos, como ser la distribución geográfica de los usuarios, los carriers y marcas de dispositivos por ellos utilizados.

En nuestra opinión, que los datos se encuentren anonimizados presenta una gran ventaja, que es la posibilidad de tratar información sensible o confidencial sin necesidad de firmar acuerdos de confidencialidad. De esta forma, un equipo ajeno a la empresa podría encargarse de realizar el análisis, incorporando nuevas técnicas o una visión diferente a la de un equipo propio. En contraste, esto dificulta en cierta medida el trabajo a realizar, ya que es complicado encontrar algunas relaciones o comprenderlas, por lo menos en esta etapa de aprendizaje.

Respecto al análisis realizado sobre los horarios y la actividad, vimos que la cantidad de subastas cae en horas de la madrugada y sigue una tendencia creciente a partir de las 10hs. Los clicks, por su parte, se mantienen con cierta uniformidad, registrando también un crecimiento a partir de las 20hs. En cuanto a los eventos, encontramos que la distribución es similar a la de subastas, mientras que la de las instalaciones sigue también esa tendencia, con una fuerte caída en la madrugada, y crecimiento a partir de las 9hs.

En cuanto a las cantidades analizadas, pudimos determinar cuáles son las aplicaciones más relevantes en el marco de los eventos con más apariciones. Pudimos determinar, a su vez, en que horario y día de la semana se registran dichos eventos, viendo que los días más activos son los martes y miércoles por la noche. También encontramos las aplicaciones más instaladas, notando que hay dos que se destacan por sobre el resto, y en qué día y horarios se dan dichas instalaciones, obteniendo un resultado similar al encontrado para los eventos. Curiosamente, encontramos que en el dataset no hay instalaciones atribuidas a Jampp, y que un 25% de los installs son implícitos. Por último, vimos que el 99% del dataset está atribuido al mismo anunciante, lo que no nos permite hacer un análisis muy profundo explotando esta veta.

Las subastas presentes en el dataset nos permiten determinar que hay dos plataformas donde se realizan, siendo una de ellas claramente dominante con el 80% de la actividad. Por otro lado, encontramos varios dispositivos que originaron una gran cantidad de subastas: para los dos casos con mayor cantidad sería en promedio una cada 35 segundos en el período de tiempo analizado. Analizando por día de la semana, obtuvimos un nivel de actividad dispar, en ambos casos fuertemente concentrado en dos días, con un pico máximo de una subasta cada 12 segundos. Si bien no podemos atribuir este inusual nivel de actividad a un fraude, si consideramos que debería estudiarse su causa con mayor detalle. Por último, pudimos determinar que una cantidad bajísima de subastas concluyen con un install, en este dataset el número es de 0.06%.

Sobre el comportamiento de los usuarios, pudimos determinar la zona de la pantalla donde se registra la mayor cantidad de clicks, así como también el tiempo que tarda un usuario desde que ve la impresión hasta que hace click. Ambos datos son importantes para optimizar el diseño de las impresiones mostradas por Jampp. Por otro lado, obtuvimos las proporciones de installs sobre clicks y la distribución de instalaciones con conexión WiFi vs. las hechas con datos celulares, viendo en ambos casos resultados lógicos: muy pocos usuarios que hacen click terminan instalando la aplicación, mientras que casi el 80% de las instalaciones se realizan con conexión WiFi.

Finalmente, los datos demográficos y características generales nos permitieron confirmar que el dataset pertenece a un solo país, y que la mayor cantidad de actividad se concentra en una de las ciudades, que supera por amplio margen al resto. Asimismo, obtuvimos la distribución de los usuarios respecto a la prestadoras de servicio, observando que es bastante uniforme, y las marcas y modelos de dispositivos utilizados, obteniendo en ambos casos dos valores que se destacan.

A modo de cierre, podemos decir que el análisis realizado nos ha permitido entender la estructura del set de datos y sus principales características, así como también obtener una noción del comportamiento de los usuarios respecto a las publicidades mostradas. De esta manera, un análisis de este tipo puede utilizarse para mejorar la oferta de productos, las prestaciones o características de la plataforma, o bien detectar zonas donde la compañía podría expandir su operación, entre otras aplicaciones.

Asimismo, nos gustaría agregar que el estudio de un set de datos de estas carácteristicas, que puede obtenerse en forma muy sencilla, por ejemplo en base a logs de actividad, resulta una herramienta poderosísima a la hora de tomar decisiones estratégicas para la compañía, en una amplia variedad de campos, a fin de administrar de la mejor forma posible los recursos disponibles.