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.

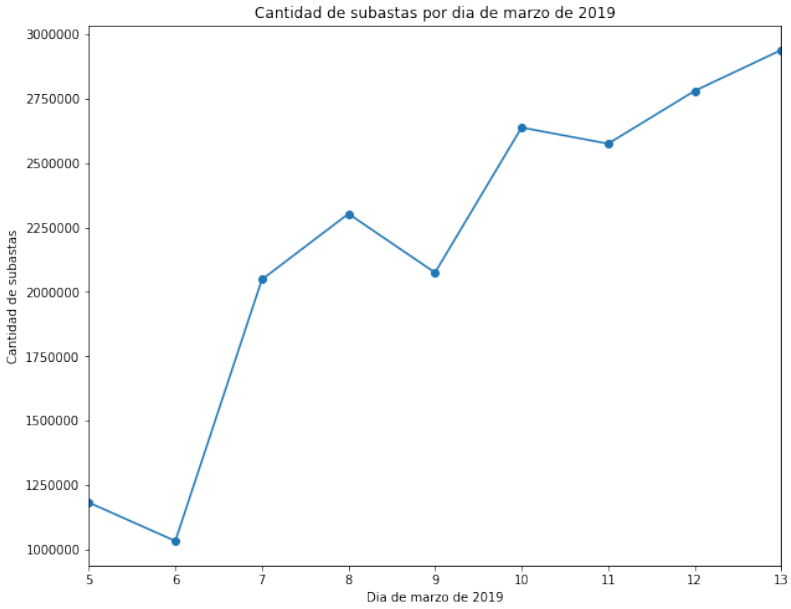

Figure 1: Cantidad de subastas por día de Marzo

Día de marzo	Cantidad de subastas
5	1182401
6	1032970
7	2047661
8	2303002
9	2074552
10	2637534
11	2574916
12	2779910
13	2938373

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.

Día de marzo	Sistema operativo	Cantidad de subastas
5	1	719286
	2	463115
6	1	579624
	2	453346
7	1	1617609
	2	430052
8	1	1898054
	2	404948
9	1	1618742
	2	455810
10	1	2149876
	2	487658
11	1	2165005
	2	409911
12	1	2337162
	2	442748
13	1	2456467
	2	481906

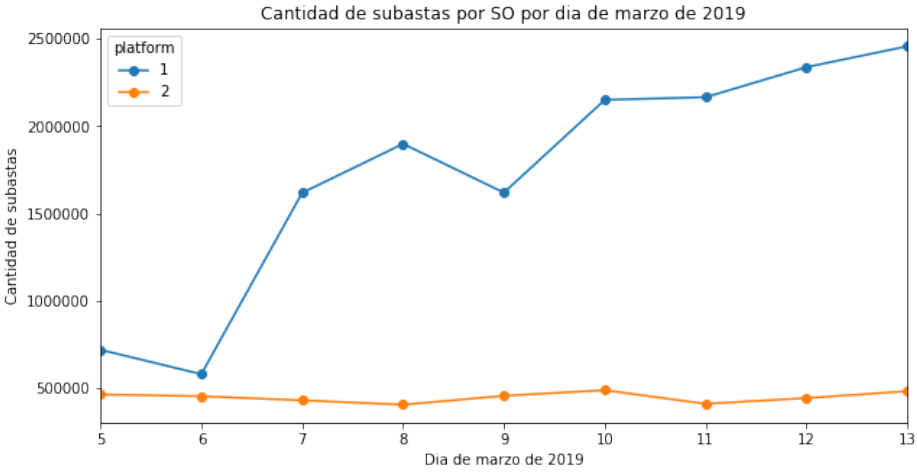

Figure 2: Cantidad de subastas por día de Marzo y por sistema operativo

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.

Hora del día	Cantidad de subastas
0	1005716
1	1371091
2	1388464
3	1027541
4	716194
5	487243
6	325730
7	245109
8	247915
9	329604
10	494726
11	627907
12	748935
13	741996
14	805579
15	883824
16	941866
17	967539
18	989528
19	994381
20	933318
21	1015053
22	1108219
23	1173841

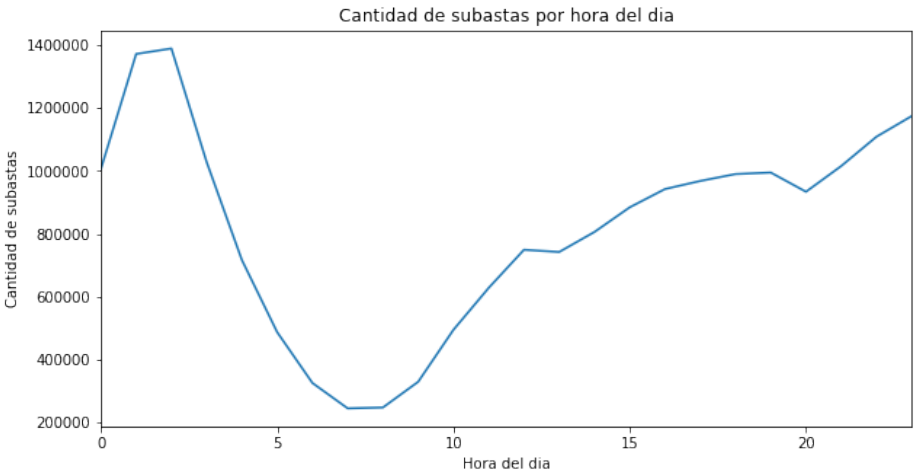

Figure 3: Cantidad de subastas por hora del día

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.

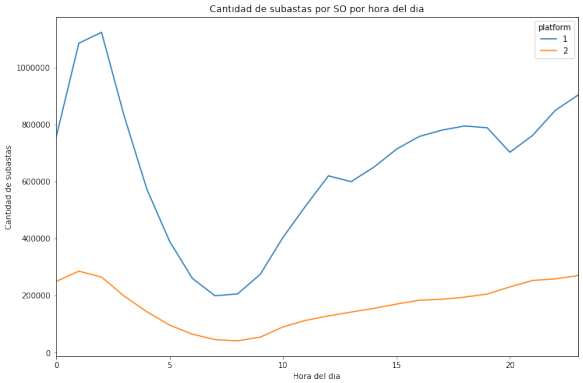

Figure 4: Cantidad de subastas por hora del día por sistema operativo

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).

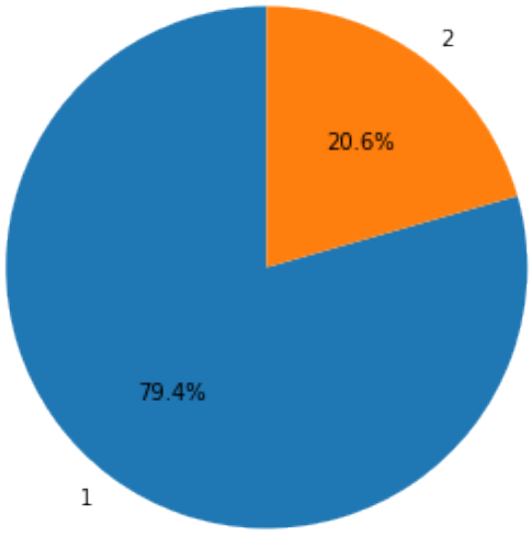

Figure 5: Porcentaje de subastas para cada plataforma

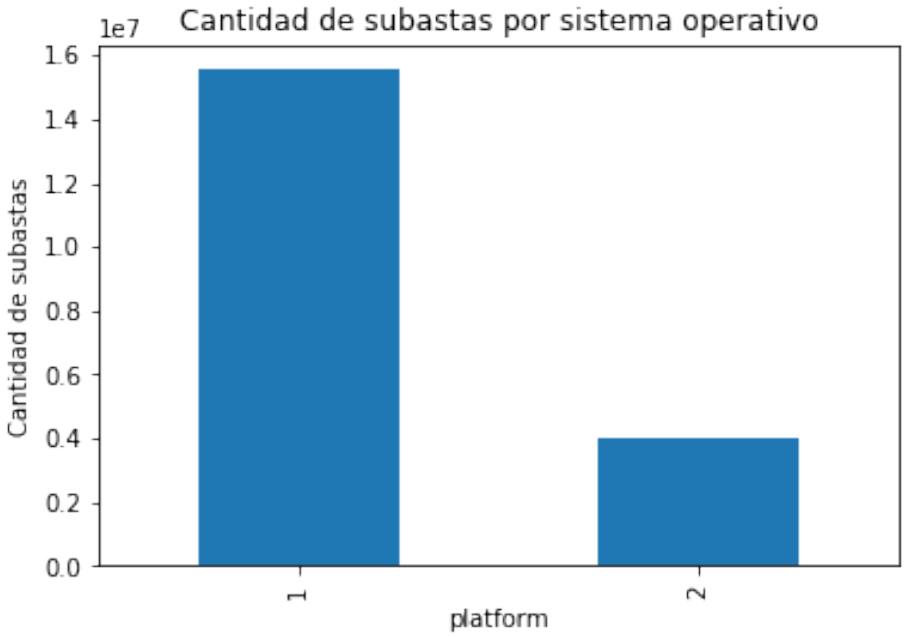

Figure 6: Porcentaje de subastas para cada plataforma

2.1.7 Subastas por source

Como indica la introducción a esta sección (vease Auctions/Análisis general), 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.

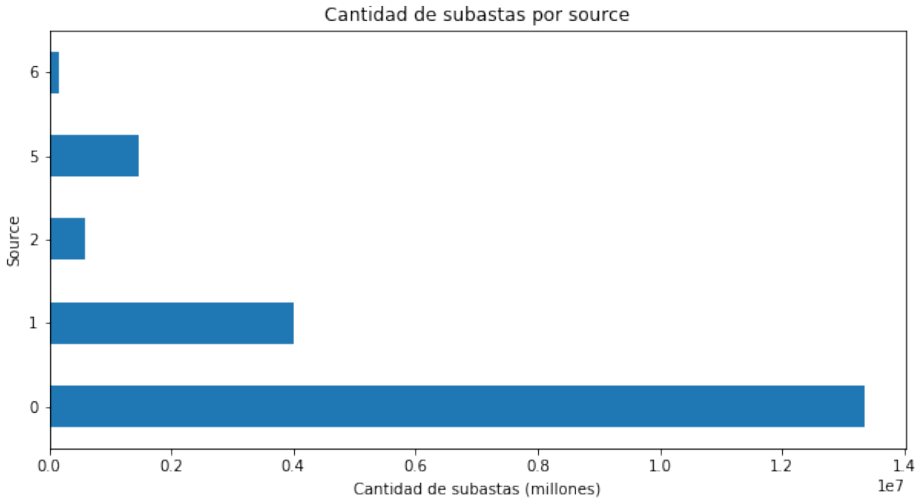

Figure 7: Cantidad de subastas para cada source

2.2 Clicks

2.3 Eventos

2.4 Instalaciones

3 Análisis de archivos en conjunto

4 Conclusion