

Trabajo Práctico N° 1

Análisis Exploratorio

Teoría de Organización de Datos

Nombre grupo: "The data inception"

N° grupo: 38

Todo el trabajo realizado puede encontrarse en el siguiente repositorio de github:

<https://github.com/sebologue/tp1-datos.git>.

participantes	n° padrón	mail
LOIS, Lucas Edgardo	[COMPLETAR]	[COMPLETAR]
MARIJUAN, Magalí	100070	maguimar001@gmail.com
MARIANI, Santiago Tomás	[COMPLETAR]	[COMPLETAR]
LOGUERCIO, Sebastian Ismael	[COMPLETAR]	[COMPLETAR]

Índice

1. Introducción:	2
2. Análisis Preliminar	2
2.1. Eventos	2
2.2. Viewed product	4

1. Introducción:

Este trabajo práctico consistía en realizar un análisis sobre un conjunto de eventos de web analytics de usuarios que visitaron www.trocafone.com, su plataforma de ecommerce de Brasil.

Para realizar dicho trabajo utilizamos el lenguaje de programación *Python*. Para el análisis de datos usamos la librería *Pandas* y para la realización de gráficos utilizamos la librerías *Matplotlib* y *Sns*. Por otro lado, trabajamos usando el sistema de control de versiones *GIT*

En nuestra primera experiencia realizando un análisis exploratorio, nos dimos cuenta que una buena abstracción para realizar el trabajo práctico era: **preguntandole al dataset**. Es decir que en nuestro modelo siempre hacíamos preguntas que luego intentábamos responder con los datos que podíamos obtener del dataset.

2. Análisis Preliminar

En un primer análisis intentamos entender cómo estaba compuesto el data set. Es decir, cuáles son los eventos que los componen, cómo interactúan como un conjunto y cuáles son las características propias de cada evento.

2.1. Eventos

El dataset contenía los siguientes eventos:

- Viewed product: El usuario visita una página de producto.
- Brand listing: El usuario visita un listado específico de una marca viendo un conjunto de productos.
- Visited site: El usuario ingresa al sitio a una determinada url.
- Ad campaign hit: El usuario ingresa al sitio mediante una campaña de marketing online.
- Generic listing: El usuario visita la homepage.
- Searched products: El usuario realiza una búsqueda de productos en la interfaz de búsqueda del site.
- Search engine hit: El usuario ingresa al sitio mediante un motor de búsqueda web.
- Checkout: El usuario ingresa al checkout de compra de un producto.
- Staticpage: El usuario visita una página.
- Conversion: El usuario realiza una conversión, comprando un producto.
- Lead: El usuario se registra para recibir una notificación de disponibilidad de stock, para un producto que no se encontraba disponible en ese momento.

La cantidad total de eventos fue: 1011288. Ellos se distribuyeron de la siguiente manera:

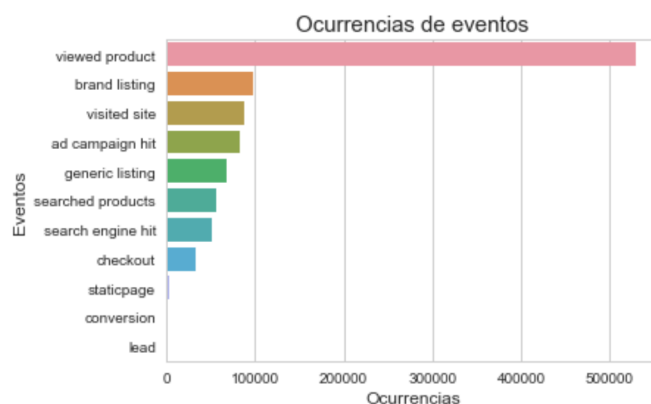


Figura 1: En este gráfico se puede ver los eventos y la cantidad de apariciones que tuvieron en el dataset.

Nos dimos cuenta que no todos los campos del data set eran relevantes para todos los eventos, por lo tanto, mediante un análisis de elementos nulos llegamos a las siguientes conclusiones:

- Todos los eventos tiene información temporal y sobre la persona que realizó dicho evento.
- Para el evento **viewed product** sus campos obligatorios son: timestamp , sku , model, condition, storage, color.
- Para el evento **brand listing** su campo obligatorio es: skus.
- Para el evento **visited site** sus campos obligatorios son: channel, new vs returning, city, region, country, device tipe , screen resolution, operating system version y browser version.
- Para el evento **ad campaing hit** sus campos obligatorios son: url y campaing source.
- Para el evento **generic listing** su campo obligatorios es: skus.
- Para el evento **serched product** sus campos obligatorios son: skus y search term.
- Para el evento **serched engine** su campo obligatorio es: search engine
- Para el evento **checked out** sus campos obligatorios son: sku, color, storage, model y condition
- Para el evento **static page** sus campo obligatorio es: satatic page
- Para le evento **conversion** sus campos obligatorios son: sku, model, color , condition y storage
- Para el evento **lead** su campo obligatorio es: model

Como todos los eventos tenían información temporal. Decidimos analizar la cantidad de eventos que se realizan por hora del día y el resultado fue:



Figura 2: En este gráfico se puede ver los eventos y su distribución a lo largo del día.

Por lo tanto, la mayoría de los eventos se efectúan entre las 10 am y 2 am.

2.2. Viewed product

La primera columna que decidimos analizar fue model. De ella obtuvimos los 15 celulares más vistos, ellos fueron:

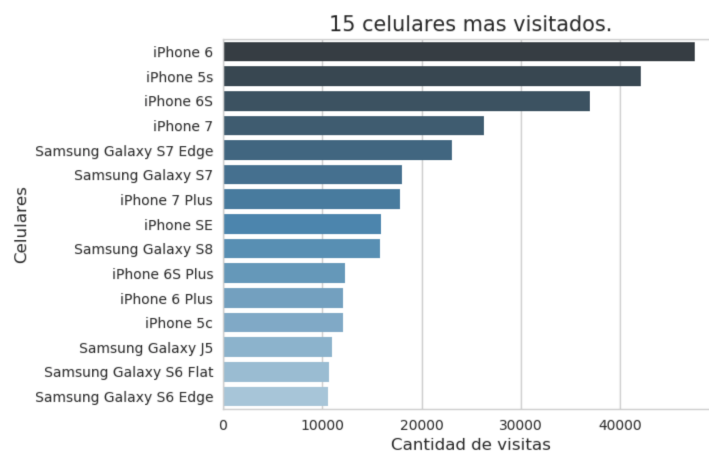


Figura 3: En este gráfico se pueden ver los 15 celulares más vistos.

Sobre las características de esos modelos primero analizamos el color de los productos vistos y concluimos que: El 50 % de de los usuarios concentró su elección en los siguientes colores:

- El 23 % de los productos son negro.
- El 20 % de los productos son dorados.
- El 11 % de los productos son gris espacial.

Luego analizamos el almacenamiento de los productos y concluimos que:

- El 33 % visita productos con almacenamiento de 16 GB.
- El 32 % visita productos con almacenamiento de 32 GB.
- El 17 % visita productos con almacenamiento de 64 GB.

- El 6 % visita productos con almacenamiento de 8 GB.

Por último analizamos las condiciones de los productos vistos y concluimos que:

- El 0,02 por ciento de los productos que visita el usuario son nuevos.
- El 2 % de los productos que tienen identificador de huella digital.
- El 42 % de los productos son de calidad buena.
- El 27 % de los productos son de calidad excelente .
- El 27 % de los productos son de calidad muy buena.

Realizamos un análisis temporal sobre el evento y obtuvimos que:



Figura 4: En este gráfico se puede ver como fueron las visitas según los días del año

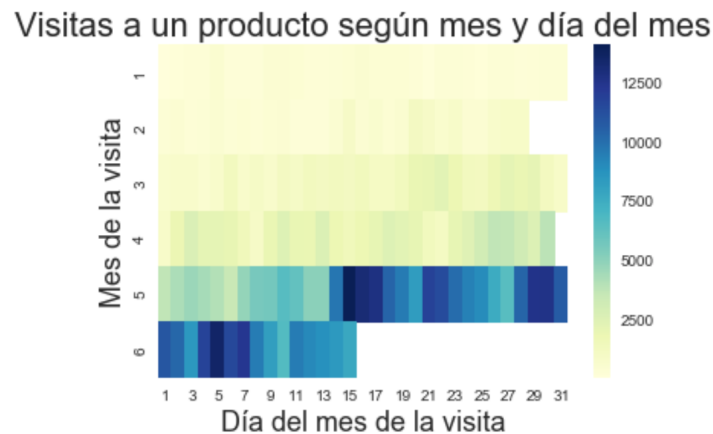


Figura 5: En este gráfico se pueden ver como fueron las visitas a un producto según mes y el día del mes.

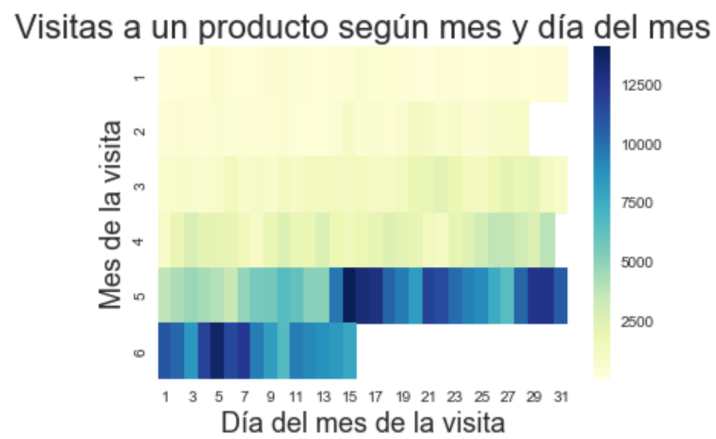


Figura 6: En este gráfico se pueden ver como fueron las visitas según el mes y el día de la semana.