FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Trabajo Práctico Nº 1

Análisis Exploratorio

Teoría de Organización de Datos

Nombre grupo: "The data inception"

Nº grupo: 38

Todo el trabajo realizado puede encontrarse en el siguiente repositorio de github:
https://github.com/sebalogue/tp1-datos.git.

participantes	nº padrón	mail
LOIS,Lucas Edgardo	[COMPLETAR]	[COMPLETAR]
LOGUERCIO, Sebastian Ismael	[COMPLETAR]	[COMPLETAR]
MARIANI, Santiago Tomás	100516	santiagomariani2@gmail.com
MARIJUAN, Magalí	100070	maguimar001@gmail.com

Índice

1. Introducción

Este trabajo práctico consite 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 utilizaremos el lenguaje de programación *Python*. Para el análisis de datos usaremos la librería *Pandas* y para la realización de gráficos utilizaremos las librerías *Matplotlib* y *Seaborn*. Por otro lado, trabajaremos usando el sistema de control de versiones *GIT*

En nuestra primera experiencia decimos usar la siguiente abstracción para realizar el trabajo práctico: **preguntar al dataset**, es decir, que en nuestro modelo siempre hacemos preguntas que luego intentaremos responder con los datos que obtendremos del dataset.

2. Análisis Preliminar

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

2.1. Eventos

El dataset contiene los siguiente 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 campana de marketing online.
- Generic listing: El usuario visita la homepage.
- Serched 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 es: 1011288. Los cuales están distribuidos de la siguiente manera.

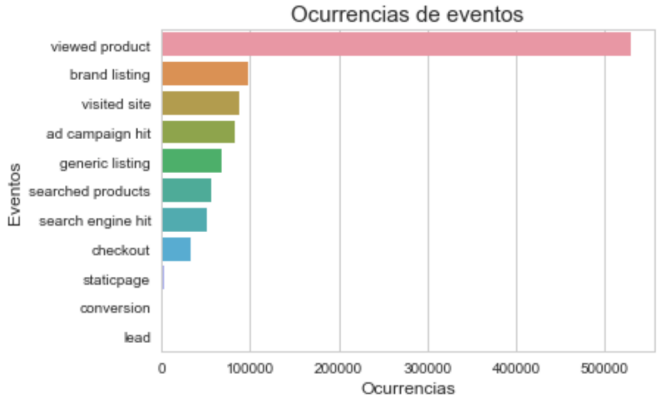

Figura 1: *Cantidad de apariciones de cada evento.*

Nos dimos cuenta que no todos los campos del data participan en todos los eventos. Cada evento utiliza una determinada cantidad de columnas del data set. Por lo tanto, mediante un análisis de elementos nulos llegamos a las siguiente conclusiones:

- Todos los eventos tienen 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 tienen información temporal. Decidimos análizar la cantidad de eventos que se realizan por hora del día y el resultado fue:

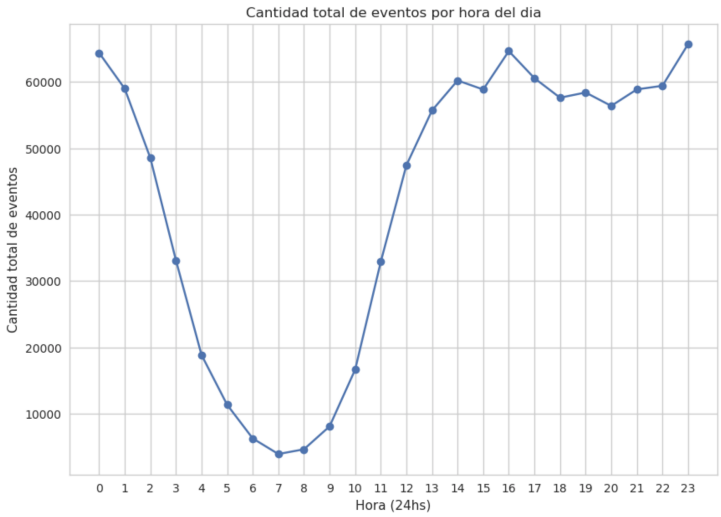

Figura 2: *Distribución de los eventos a lo largo del día.*

Por lo tanto, la mayoría de los eventos se efectúan entre las 14 am y 2 am.

2.2. Viewed product

2.2.1. Análisis individual de las características

La primera columna que decicimos analizar de éste evento es model. De ella obtuvimos los 15 celulares más vistos, ellos son:

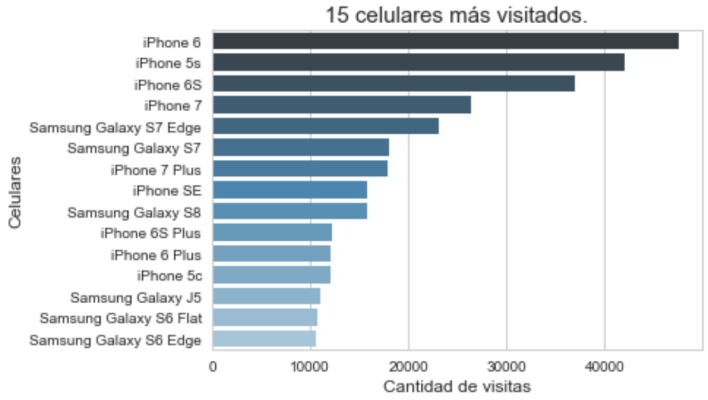

Figura 3: *Cantidad de visitas por celular.*

Sobre las características de esos modelos primero analizamos el color de los productos vistos y concluímos que: el 50 % de las visitas se concentraron en los siguientes colores: Las visitas según el color de los celulares fue:

- el 23 % de las visitas es hacia productos negros.
- el 20 % de las visitas es hacia productos dorados.
- el 11 % de de las visitas es hacia productos gris espacial.
- el 10 % de de las visitas es hacia productos blancos.
- el 9 % de de las visitas es hacia productos plateados.

- el 6 % de de las visitas es hacia productos rosas.
- Despreciamos el resto de los colores ya que el porcentaje de visitas es poco significativo.

Luego analizamos las visitas hacia el almacenamiento de los productos y concluímos que:

- El 33 % de las visitas es hacia productos con almacenamiento de 16 GB.
- El 32 % de las visitas es hacia productos con almacenamiento de 32 GB.
- El 17 % de las visitas es hacia productos con almacenamiento de 64 GB.
- El 6 % de las visitas es hacia productos con almacenamiento de 8 GB.

Por último analizamos las condiciones que más visita la gente y concluímos que:

- El 42 % de las visitas es hacia productos que son de calidad buena.
- El 27 % de las visitas es hacia productos que son de calidad excelente .
- El 27 % de las visitas es hacia productos que son de calidad muy buena.
- El 2 % de las visitas es hacia productos que tienen identificador de huella digital.
- El 0,02 % de las visitas es hacia productos nuevos.

Por otra parte, realizamos un análisis temporal sobre el evento y obtuvimos que:

2.2.2. Análisis temporal

Figura 4: *Cantidad de visitas según día del año.*

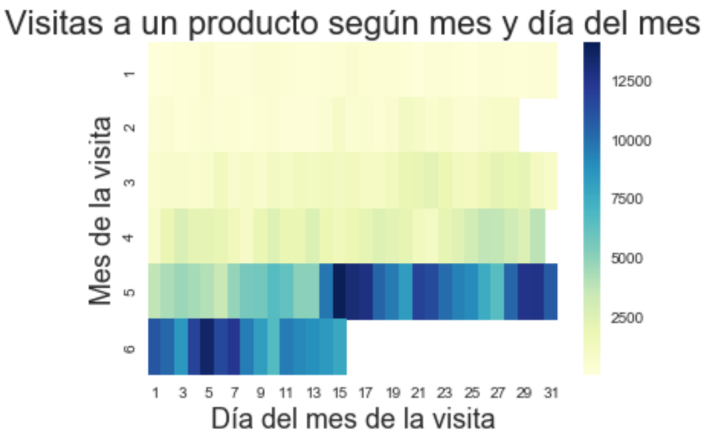

Figura 5: *Cantidad de visitas por cada día de cada mes.*
Podemos notar que hay un fuerte incremento de actvidad a partir del 15 de mayo.

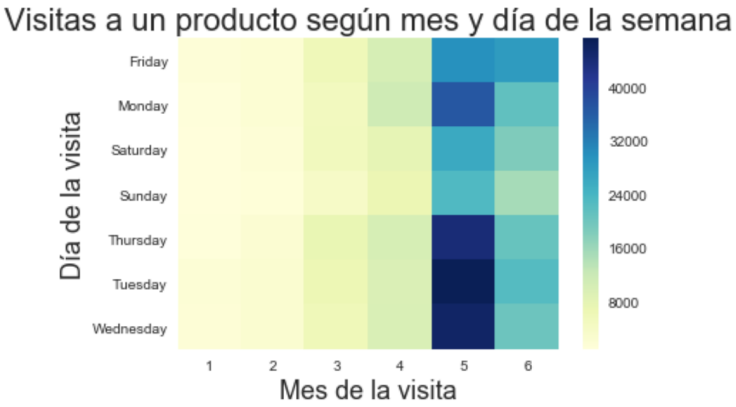

Figura 6: *Cantidad de visitas por cada día de semana de cada mes.*

Nuevamente hay un fuerte incremento a partir del mes de mayo. Sin embargo, no parece haber una predominancia cuando hablamos de los días de la semana. Lo que podemos concluir es que los domingos hay menos actividad.

Es importante destacar en el análisis temporal que el mes de junio no está completo. Sólo tenemos datos de su primera quincena.

2.2.3. Análisis cruzado

Por último relacionaremos todas las careacterísticas analizadas anteriormente entre sí. Primero entrelazaremos la inforamción entre los cinco modelos más vistos y su color. Obtuvimos lo siguiente:

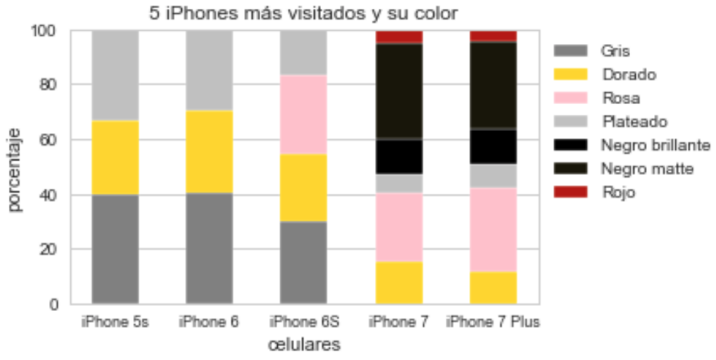

Figura 7: *Top 5 de celulares según su color.*

Por último, entrelazremos información entre los cinco modelos más visitos y su almacenamiento. Obtuvimos lo siguiente:

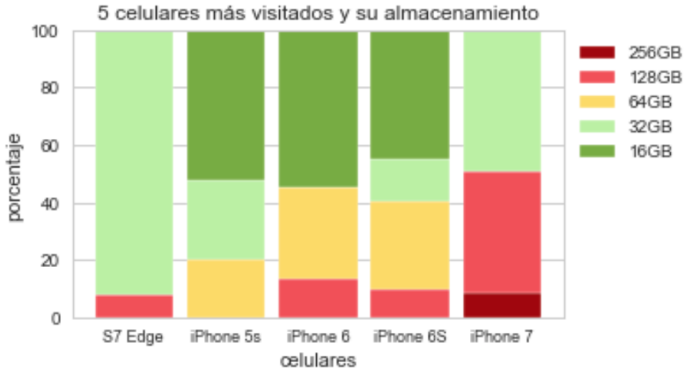

Figura 8: *Top 5 de celulares según su almacenamiento.*

[¿AGREGAR INFORMACIÓN SOBRE 5 TOP DE PRODUCTOS Y CONDICIÓN? agregar explicación de porqué sólo se analizaron iphones en color.]

2.3. Ad campaign Hit

2.3.1. Análisis individual de las caracterísiticas

Comenzamos analizando este evento fijandonos cuáles son las 10 campañas publicitarias más visitadas. Obtuvimos lo siguiente:

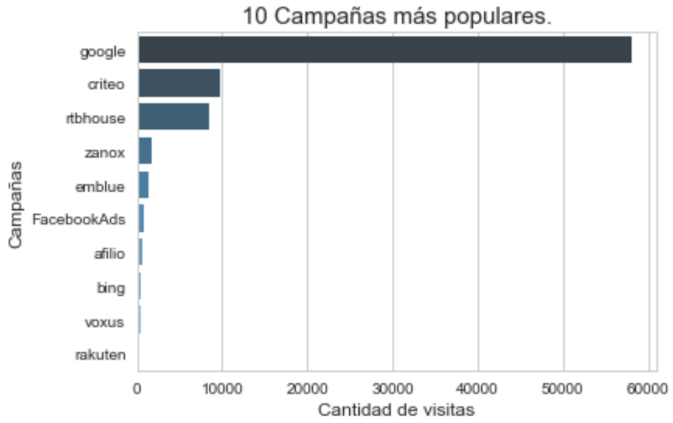

Figura 9: *En este gráfico se muestran las 10 campañas publicitarias más visitadas.*

Ad campaign hit se divide en tres tipos de accesos: página principal, ventas y compras. Obtuvimos que:

- El 65 % de los accesos es a compras.
- El 34 % de los accesos es a la página principal.
- El 0,26 % de los accesos es a ventas.

El acceso según ventas puede clasificarse según la marca. Obtuvimos lo siguiente:

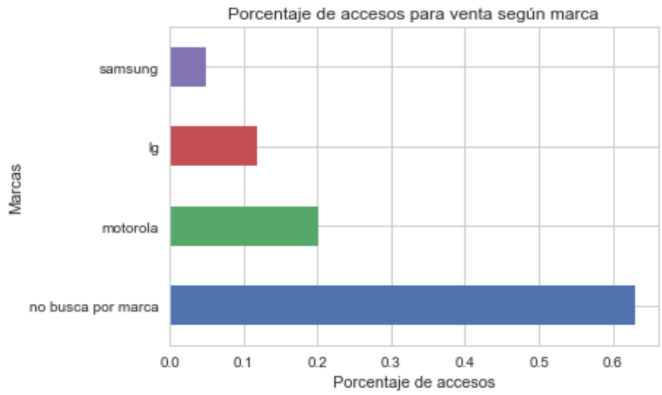

Figura 10: *Porcentaje de accesos para venta según la marca.*

Este acceso tambien puede clasificarse según el modelo. Obtuvimos lo siguiente:

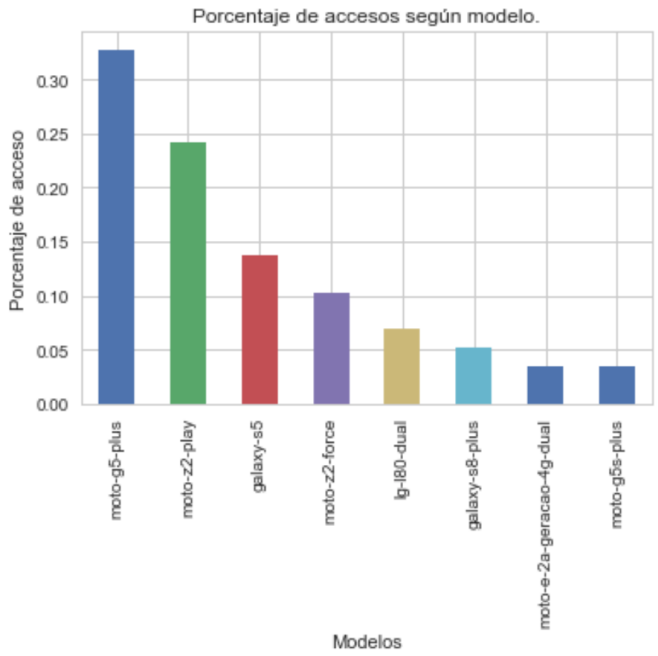

Figura 11: *Porcentaje de accesos para venta según el modelo.*

El acceso según compras puede clasificarse según la marca. Obtuvimos lo siguiente:

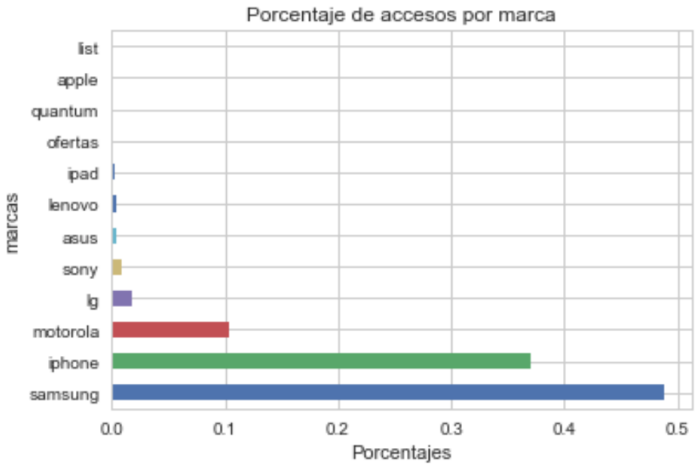

Figura 12: *Porcentaje de accesos para compra según la marca*

El acceso de compras por oferta es poco significativo. Por eso no lo excluyo del gráfico. Hacer un análisis de las compras según modelo es muy complicado y no creemos que valga la pena. Esto se debe a que un mismo modelo puede tener urls distitntos.

2.3.2. Análisis temporal

Luego, analizamos el top 5 de las campañas publicitarias más clickeadas, es decir, más visitadas según el mes. Obtuvimos lo siguiente:

Figura 13: *Cantidad de clicks según campaña por mes.*

Analizamos así tambien el top5 de publicidades según la cantidad de clicks según el día del año. Obtuvimos lo siguiente:

Figura 14: *Cantidad de clicks según campaña por día del año.*

2.4. Checkout

2.4.1. Análisis individual de las caracterísiticas

Empezamos analizando los 10 modelos que más llegaron al checkout. Ellos fueron:

- iPhone 6 con el 10 % de los checkouts.
- iPhone 5s con el 8 % de los checkouts.
- iPhone 6S con el 7 % de los checkouts.
- Samsung Galaxy J5 con el 6 % de los checkouts.
- Samsung Galaxy S7 con el 4 % de los checkouts.
- iPhone 7 con el 4 % de los checkouts.
- Samsung Galaxy S8 con el 3 % de los checkouts.
- iPhone 7 Plus con el 3 % de los checkouts.
- Samsung Galaxy J7 Prime con el 3 % de los checkouts.
- Samsung Galaxy S6 Flatcon el 3 % de los checkouts.

Analizamos la cantidad de checkouts según el almacenamiento. Lo obtenido fue:

- El 37 % de los checkouts fue de dispositivos de 16GB
- El 29 % de los checkouts fue de dispositivos de 32GB
- El 16 % de los checkouts fue de dispositivos de 64GB
- El 11 % de los checkouts fue de dispositivos de 8GB
- El 5 % de los checkouts fue de dispositivos de 128GB
- El 1 % de los checkouts fue de dispositivos de 4GB
- El 1 % de los checkouts fue de dispositivos de 256GB
- El 0,1 % de los checkouts fue de dispositivos de 512MB

Analizamos la cantidad de checkouts según el color. Lo obtenido fue:

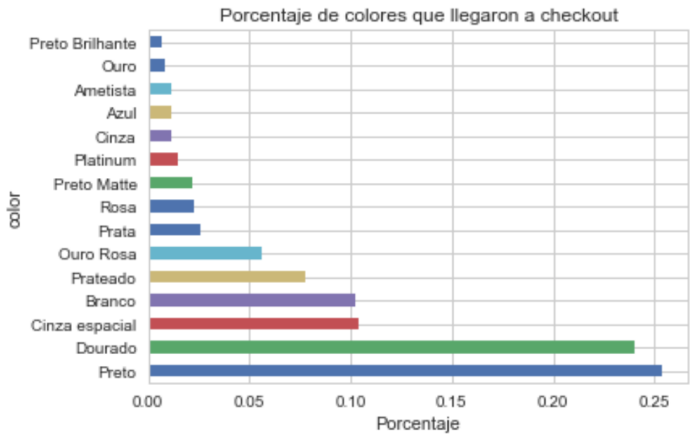

Figura 15: *Porcentaje de colores que llegaron al checkout.*

Por último analizamos la calidad de los productos que llegaron al checkout. Lo obtenido fue:

- El 45 % de los productos son de calidad buena.
- El 27 % de los productos son de calidad excelente.
- El 24 % de los productos son de calidad muy buena.
- El 3 % de los productos son de calidad buena con touch id.
- No es representativo la cantidad de productos nuevos.

2.4.2. Análisis temporal

Analizamos la cantidad de checkouts según los días del año. Obtuvimos lo siguiente:

Figura 15: *Cantidad de checkouts según día del año.*

2.4.3. Análisis cruzado