Organizacion de Datos 75.06/95.58

Trabajo práctico N°1

September 24, 2018

Alumno	Padrón
Ivo, Biaus	98708
Toscano, Miguel	98385
Labaisse, Hugo	103624
Perrone, Patricio	98230

1 Introduccion

El trabajo consistió en lo siguiente: Se pusieron a disposición datos de usuarios que visitaron el sitio web www.trocafone.com, una plataforma de e-commerce. Los mismos debieron ser analizados y ver que relaciones interesantes entre sus datos se pueden encontrar que le fueran de interés a la empresa.

Como herramientas se usó Jupyterlab, un entorno de desarrollo para Data Science, particularmente en lenguaje Python.

2 Objetivo

Mediante los análisis que se presentan a continuación, se buscó contestar a las siguientes preguntas:

¿Vale la pena invertir en la creación de una aplicación móvil?

¿Hay algún sector en el cual se podría ahorrar presupuesto?

¿Se podría mejorar algún aspecto de la página web de la empresa?

3 Desarrollo

3.1 Una primera mirada a los datos

Como primer paso, se cargaron los datos desde un archivo de formato *csv* a un *DataFrame* (estructura propia de pandas) para un primer análisis.

En dicho *DataFrame*, se encontraban las siguiente columnas:

- timestamp: Fechar y hora de cuando ocurrio el evento.

- event: Tipo de evento.
- person: Identificador de cliente que realizó el evento.
- url: Url visitada por el usuario.
- sku: Identificador de producto relacionado al evento.
- model: Nombre descriptivo del producto incluyendo marca y modelo.
- condition: Condición de venta del producto
- storage: Cantidad de almacenamiento del producto.
- color: Color del producto.
- skus: Identificadores de productos visualizados en el evento.
- search_term: Términos de búsqueda utilizados en el evento.
- staticpage: Identificador de página estática visitada.
- campaign_source: Origen de campaña, si el tráfico se originó de una campaña de marketing.
- search_engine: Motor de Búsqueda desde donde se originó el evento, si aplica.
- channel: Tipo de canal desde donde se originó el evento.

- new_vs_returning: Indicador de si el evento fue generado por un usuario nuevo o por un usuario que previamente habia visitado el sitio segun el motor de analytics.

- city: Ciudad desde donde se originó el evento.

- region: Región desde donde se originó el evento.

- country: País desde donde se originó el evento.

- device_type: Tipo de dispositivo desde donde se generó el evento.

- screen_resolution: Resolución de pantalla que se está utilizando en el dispositivo desde donde se generó el evento.

- operating_system_version: Versión de sistema operativo desde donde se originó el evento.

- browser_version: Versión del browser utilizado en el evento.

3.2 Limpieza de datos

Antes de empezar a realizar algún tipo de análisis, hubo que filtrar aquellos datos que no aportaran información interesante. Tambien hubo que eliminar aquellos que en su mayor parte eran valores inválidos, a menos que el mismo nos resultara interesante de entrada.

Teniendo en cuenta que el *DataFrame* otorgado tenia 1011288 filas, la siguiente tabla nos muestra lo previamente mencionado:

Columna	Valores inválidos
timestamp	0
event	0
person	0
url	928532
sku	447450
model	447004
condition	447452
storage	447452
color	447452
skus	789589
search_term	962321
staticpage	1007690
campaign_source	928492
search_engine	960331
channel	923910
new_vs_returning	923910
city	923910
region	923910
country	923910
device_type	923910
screen_resolution	923910
operating_system_version	923910
browser_version	923910

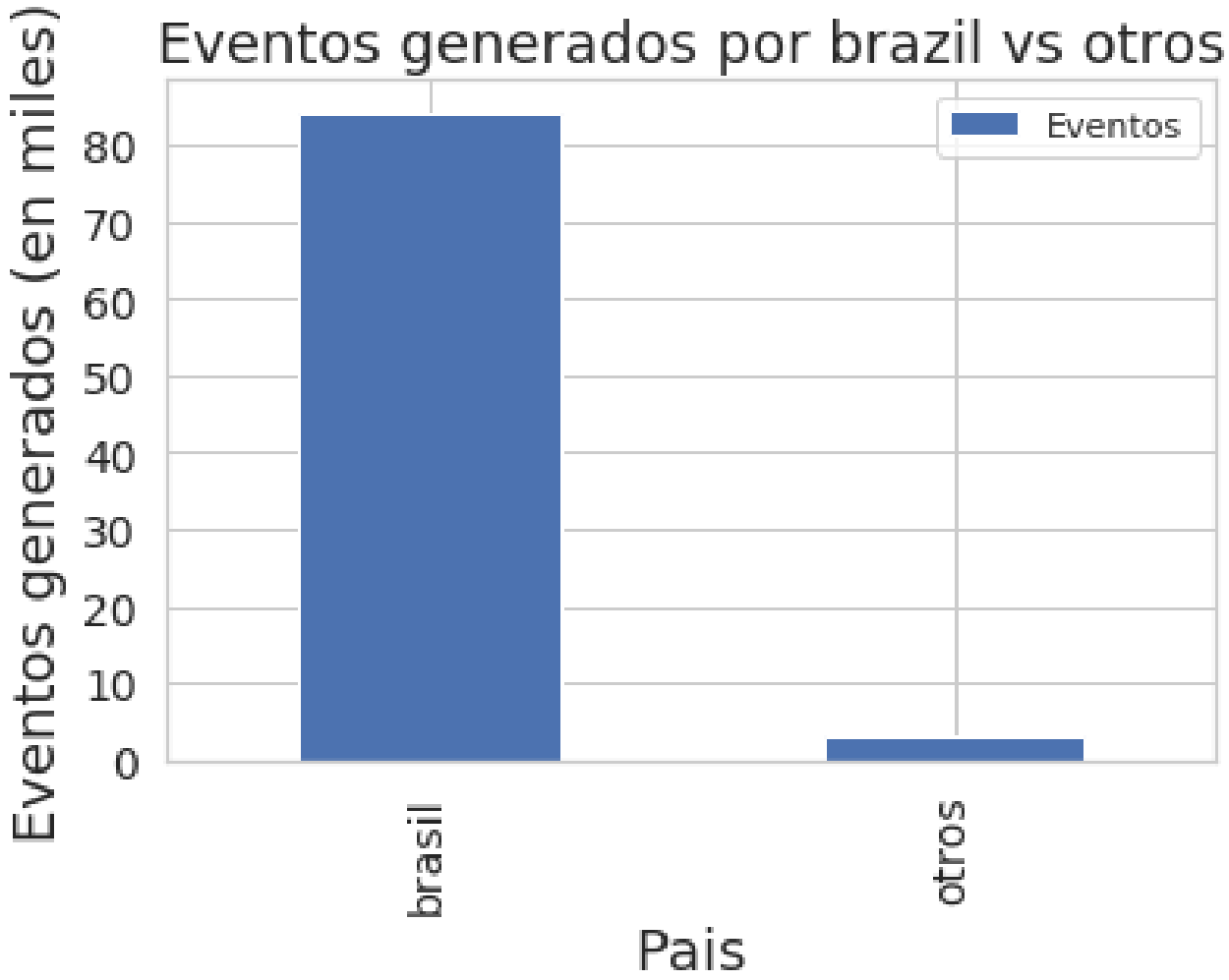
Tabla 1. Cantidad de datos inválidos por columna

En un prinicpio se decidieron eliminar columnas que consideramos de poca importancia para las preguntas planteadas tales como:

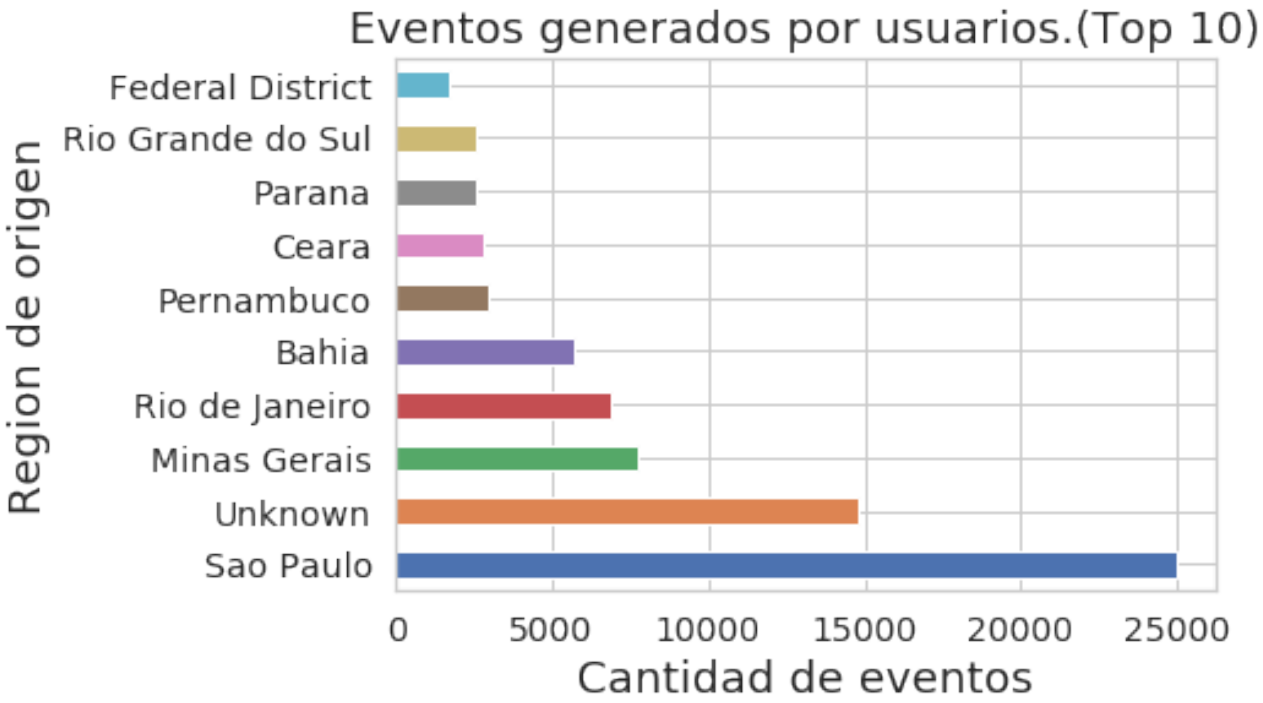
- screen_resolution
- browser_version
- color
- skus
- staticpage

3.3 Análisis demográfico

En el siguiente gráfico se muestra que la mayor parte de usuarios que generan eventos son de Brasil, por lo que nos concentramos en ellos:



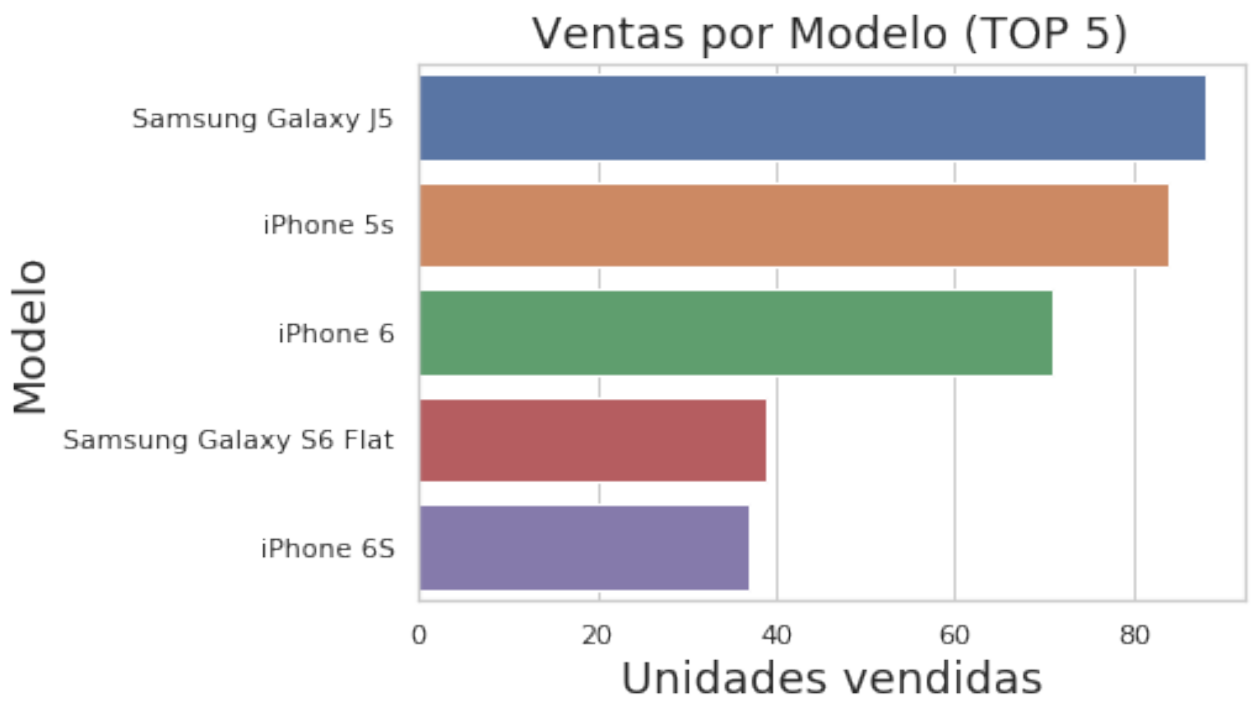
También se analizó la cantidad de eventos generados en cada región de Brasil, donde Sao Paulo es indiscutiblemente la región en donde mas ocurrieron:



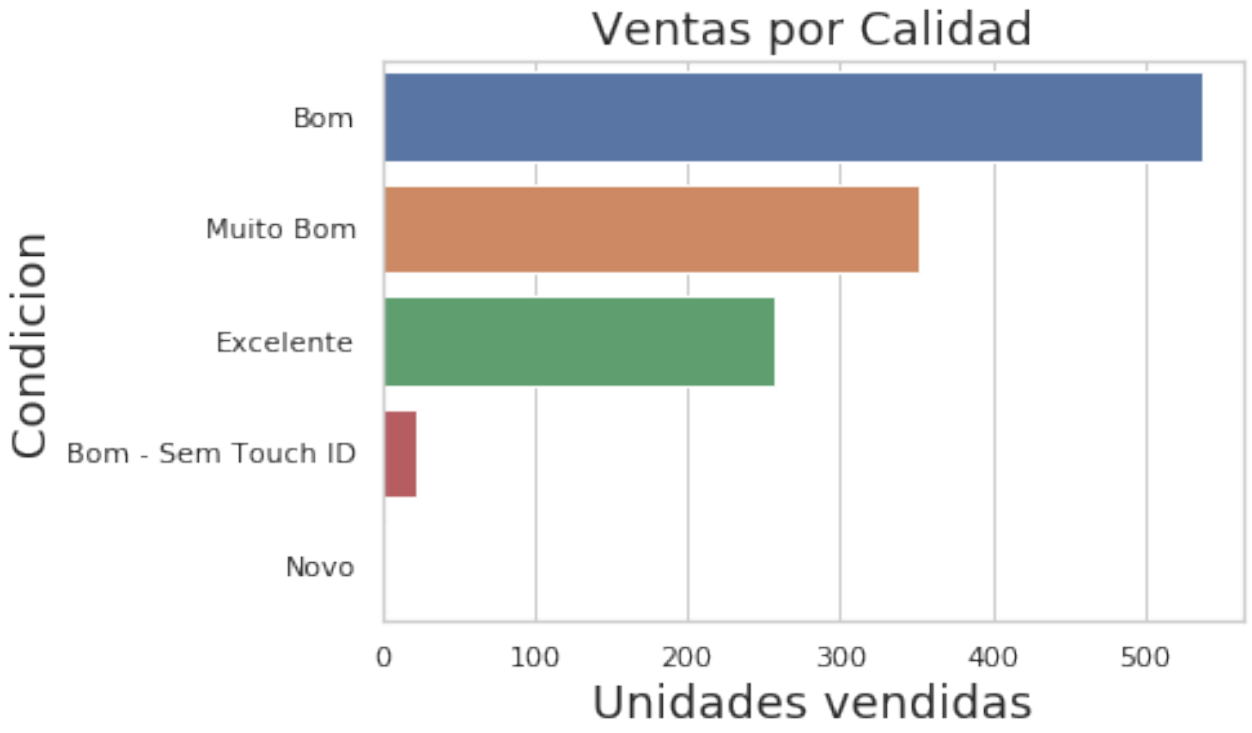
4 Análisis de ventas

4.1 Modelos mas vendidos

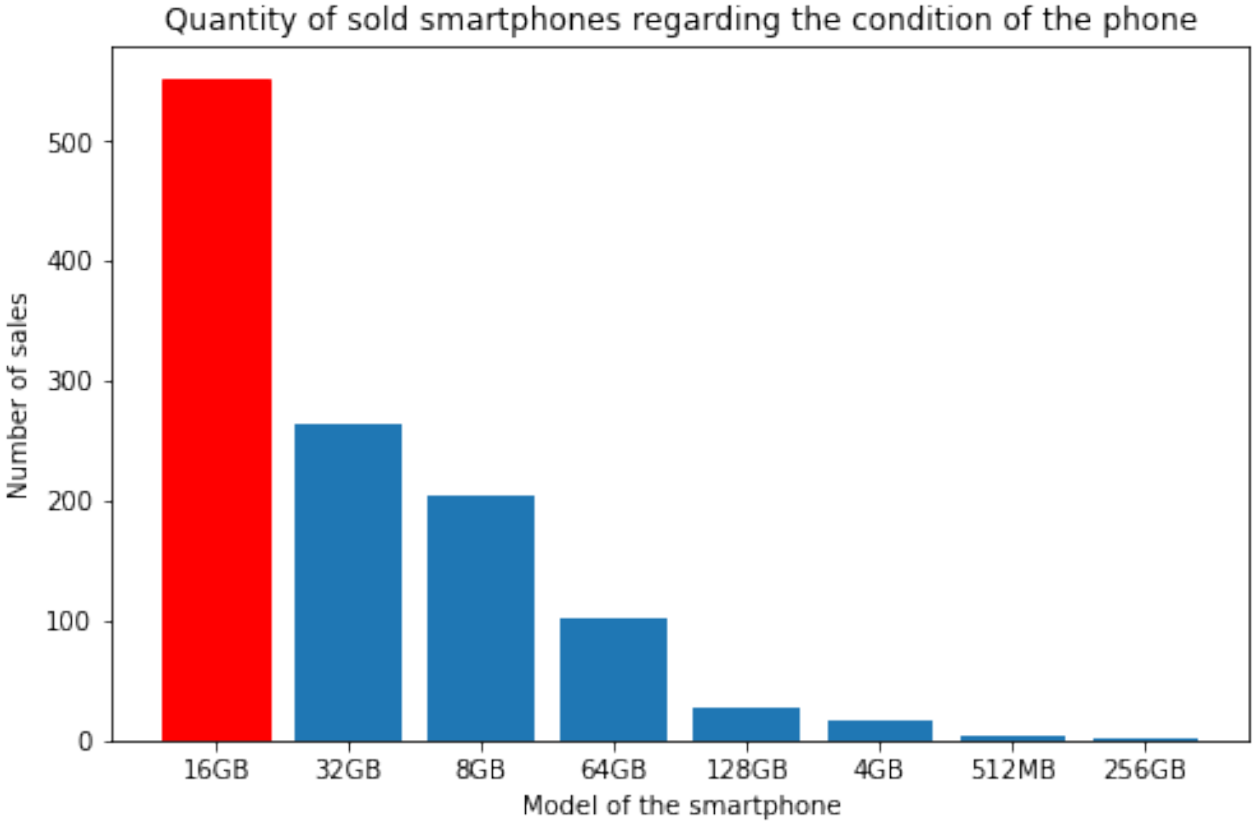
A continuación se realizó un análisis de los modelos mas vendidos, algo de principal interés para la empresa:



4.2 Ventas por calidad



De ambos gráficos se puede ver que los modelos que más circularon en el mercado son: *Samsung Galaxy J5, iPhone 5s y iPhone 6*. Sorpresivamente, los que más se vendieron son los que se identificaron en "buen" estado.



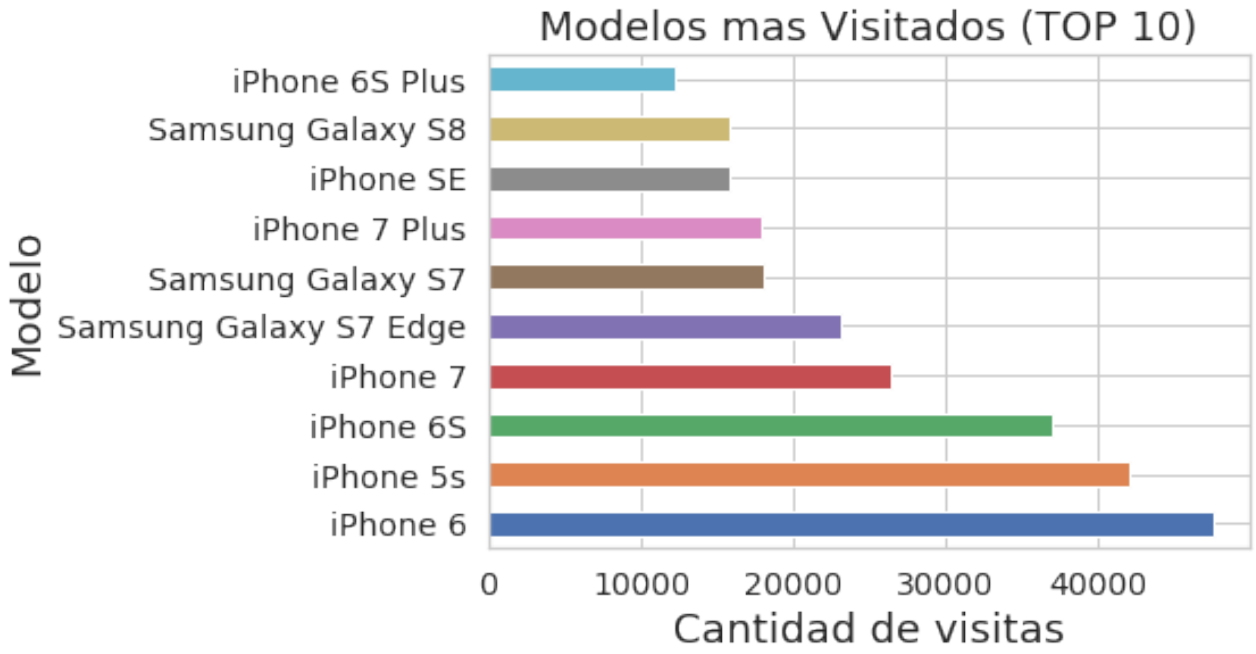
Es interesante ver como los dispositivos con capacidad de almacenamiento de 16GB superan ampliamente a los demás.

También vale destacar que no se registraron dispositivos en estado "nuevo", lo cual indica que las ventas se mantuvieron fieles al propósito de la empresa y no hubo gente que haya utilizado la misma para otros fines.

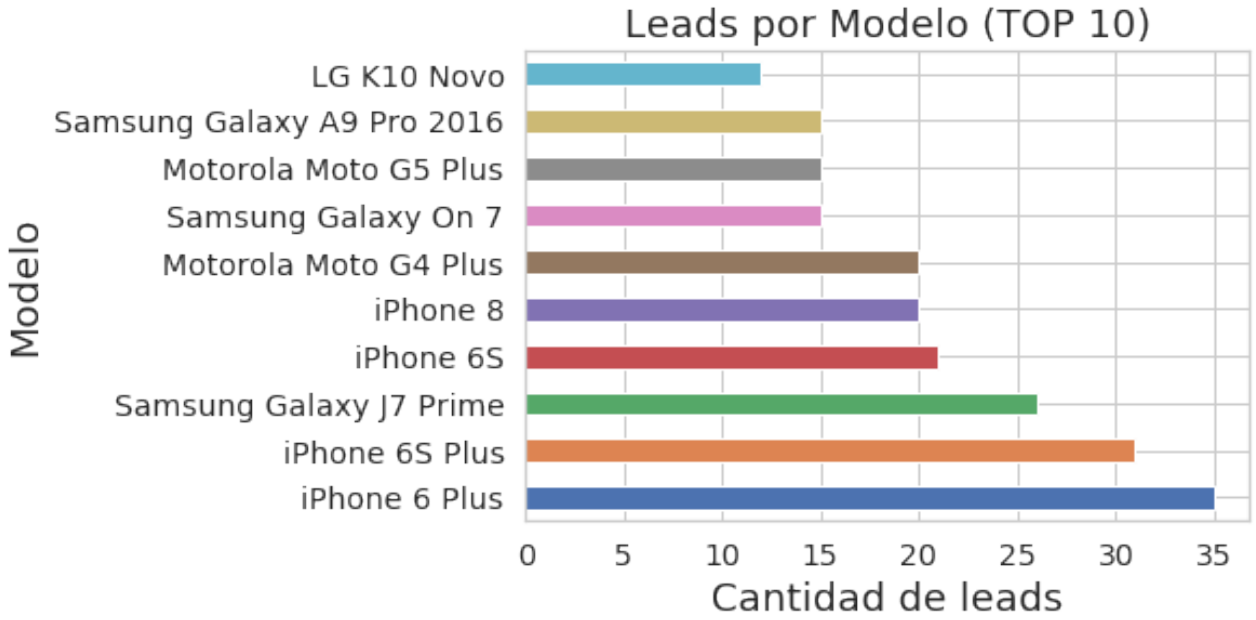
5 Análisis de popularidad

5.1 Modelos más visitados

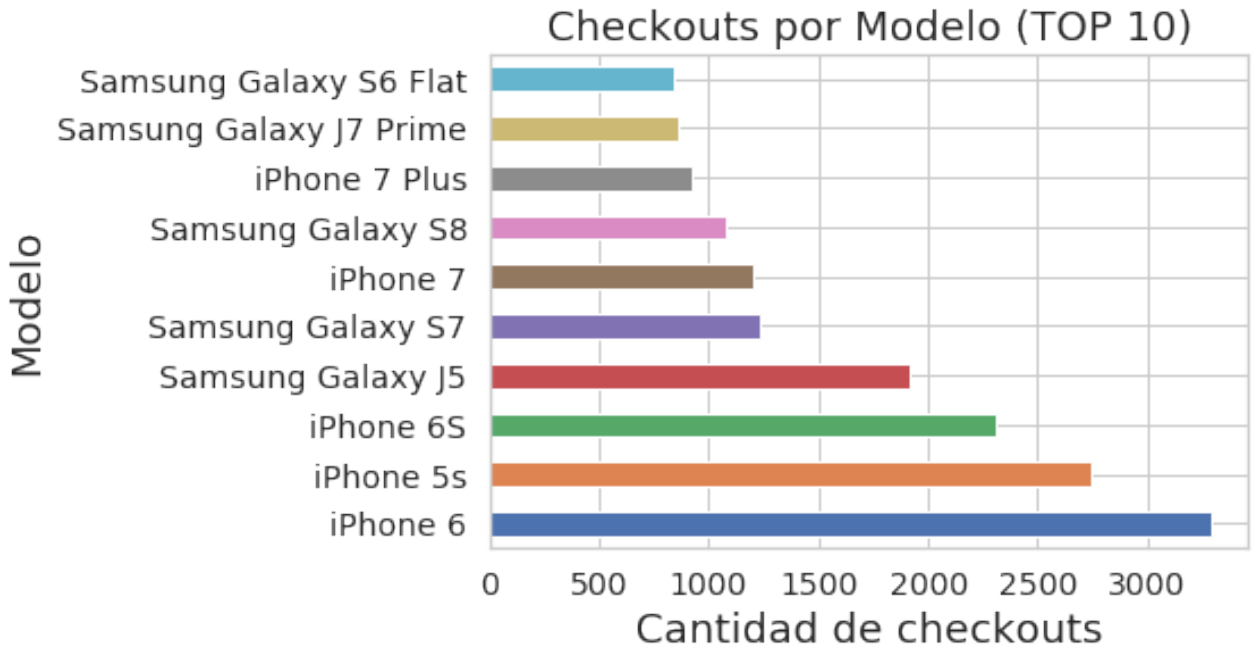
Se analizó la popularidad de ciertos modelos que destacaron entre los demás para poder decidir de que manera éstos se



5.2 Leads por modelo

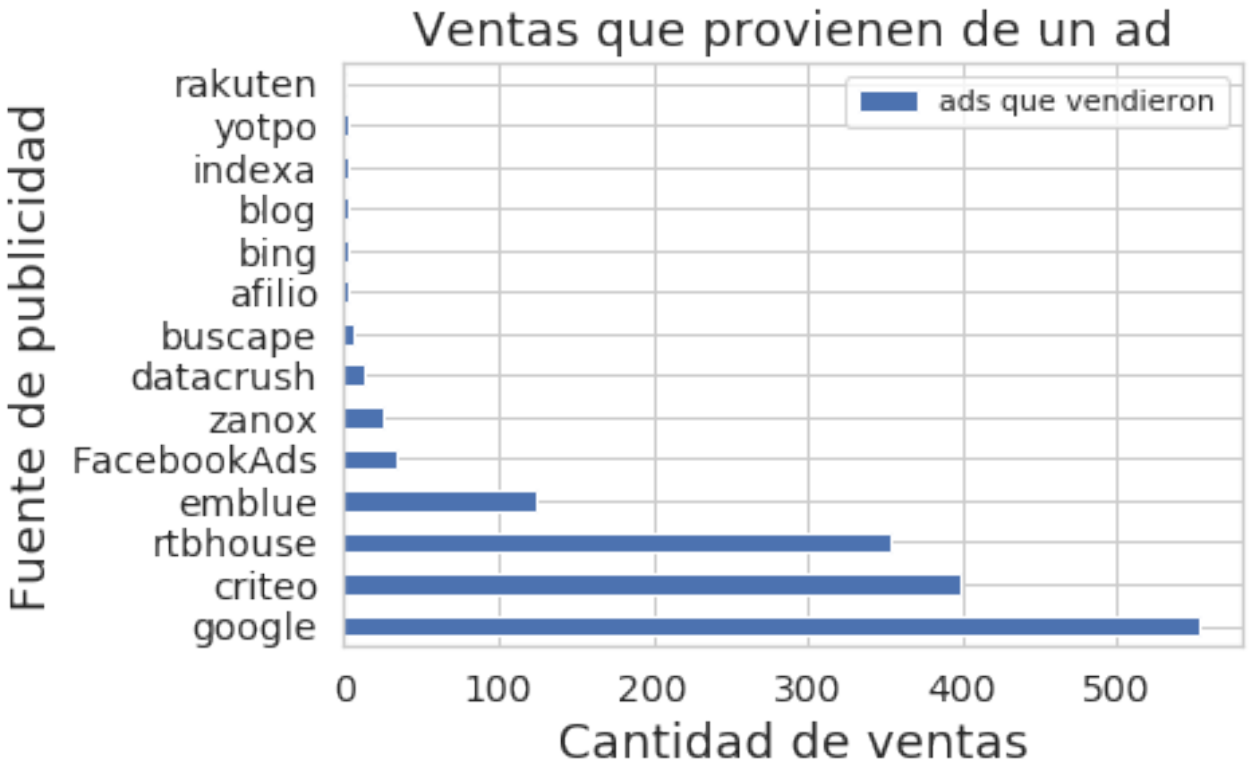


A partir de éstos gráficos, se pudo ver como particularmente el modelo *iPhone 6* fue muy codiciado por los usuarios, lo cual coincide con el análisis de ventas previo (aún si la venta no fue concretada).



6 Efectividad de las publicidades

En el siguiente segmento, se analizaron a los usuarios que entraron mediante *adds* y que efectivamente realizaron una compra gracias a la misma.

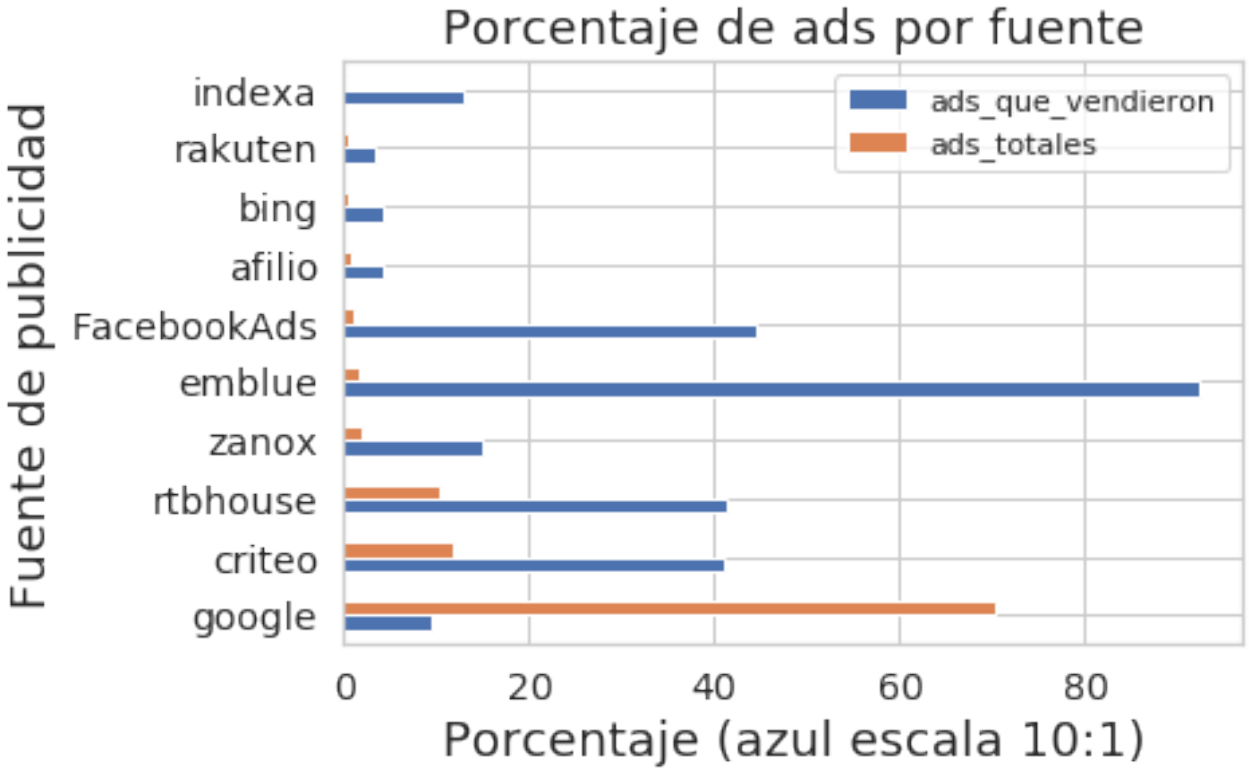


Se busco analizar una correlacion entre las compras de los usuarios con su ingroso a la pagina a travez de una publicidad. Se intento filtrar por eventos relacionados a ventas y luego ver los datos en la columna "campaign source" pero por alguna razon para todos los eventos que sean ventas esta informacion, asi como muchas otras figura como nula,cuando en realidad la informacion del usuario,como pudes ser el pais, se puede encontrar facilmente en la tabla. Teniendo en cuenta que la informacion puede simplemente no estar asignada a la venta, no se descarta que el comprador venga por medio de un ad.

Procedemos entonces a separar la informacion en dos partes, por un lado una lista de todos los compradores junto con el producto involucrado (la llamaremos Tabla A), y por otro lado todas las entradas de usuarios a la pagina por medio de un ad (la llamaremos Tabla B). La mayor parte del analisis se lleva a cabo en la Tabla B, como filtrar que las publicidades solo conduzcan a una compra, y no al inicio de la pagina o a ventas. Luego agrupo los eventos por usuario y marca del producto. A la Tabla A la organizamos de la misma manera, por usuario y marca de producto. Luego se procedio a hacer un Merge de ambas tablas , con las columnas de usuario y marca (Tabla A a izquiera) descartando a todos los usuarios que no habian realizado ninguna compra y a los compradores cuyos productos comprados no coincidian con algun producto visto en una publicidad. De esta manera, si ahora agrupamos los eventos por "campaign source" , sumando la cantidad de productos comprados en cada uno de esos eventos obtenemos la cantidad de ventas que fueron apoyadas por cada fuente de publicidad y asi tener una idea de que fuentes generan mas ventas, asi poder decidir mejor en que campañas publicitarias invertir mas y en cuales menos.

Vale tener en cuenta que si por ej 1 usuario compro 1 samsung , pero entro previamente a travez de 1 publicidad de google y criteo , ambas publicidades sobre samsung, la venta se cuenta como que fue gracias a ambas publicidades. Por lo tanto si sumamos la cantidad de " ventas totales por fuente de publicidad" (2) va a ser mayor que "las ventas totales " (1). Lo mismo ocurre para el "porcentaje de colaboración a las ventas"

Se pudo ver que la mayor eficacia la tuvieron aquellos que entraron mediante google, lo cual era muy esperable dada su popularidad. Sin embargo hubo otros portales que también mostraron ser efectivos.

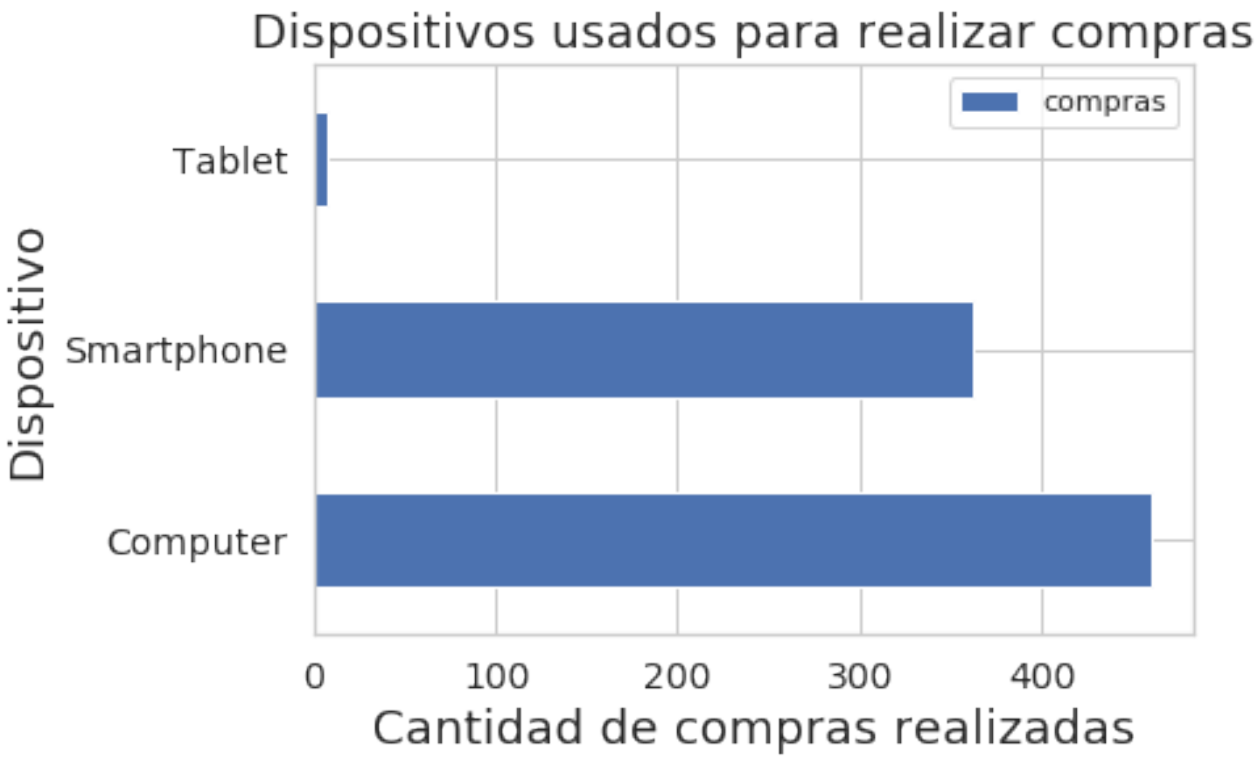


Siguiendo los pasos del analisis anterior, se decidio tener en cuenta aparte de los ads que colaboraron con una venta (analizado previamente), la cantidad de ads generados por la fuente de publicidad; y asi poder no solo saber que fuente apoyo en mas cantidad las ventas, sino tambien que porcentaje de sus ads.

De esta manera podriamos ver si hay alguna empresa A que genera 5000 ads de los cuales 500 terminan en ventas, versus una empresa B que genere 400 de los cuales 200 terminen en ventas. La empresa A podra tener mayor cantidad pero no habria que descartar invertir en la B dado que su porcentaje de hits podria es mayor.

7 Dispositivos de acceso

Una de las principales preguntas de este análisis apunta a que si una versión móvil de la plataforma tendría sentido. Para poder contestar eso se analizó la cantidad de personas que efectúan compras contra los dispositivos desde las que lo hacen:



8 Conclusiones

- **¿Vale la pena invertir en la creación de una aplicacion móvil?**

  Sí. Si bien la mayor cantidad de usuarios que efectúan compras lo hacen a través de una computadora, la misma no es mucho más grande que la que lo hace mediante un dispositivo móvil, por lo que dedicarle una plataforma a aquellos que lo hacen podría ser una buena inversión.

- **¿Hay algún sector en el cual se pueda ahorrar presupuesto?**

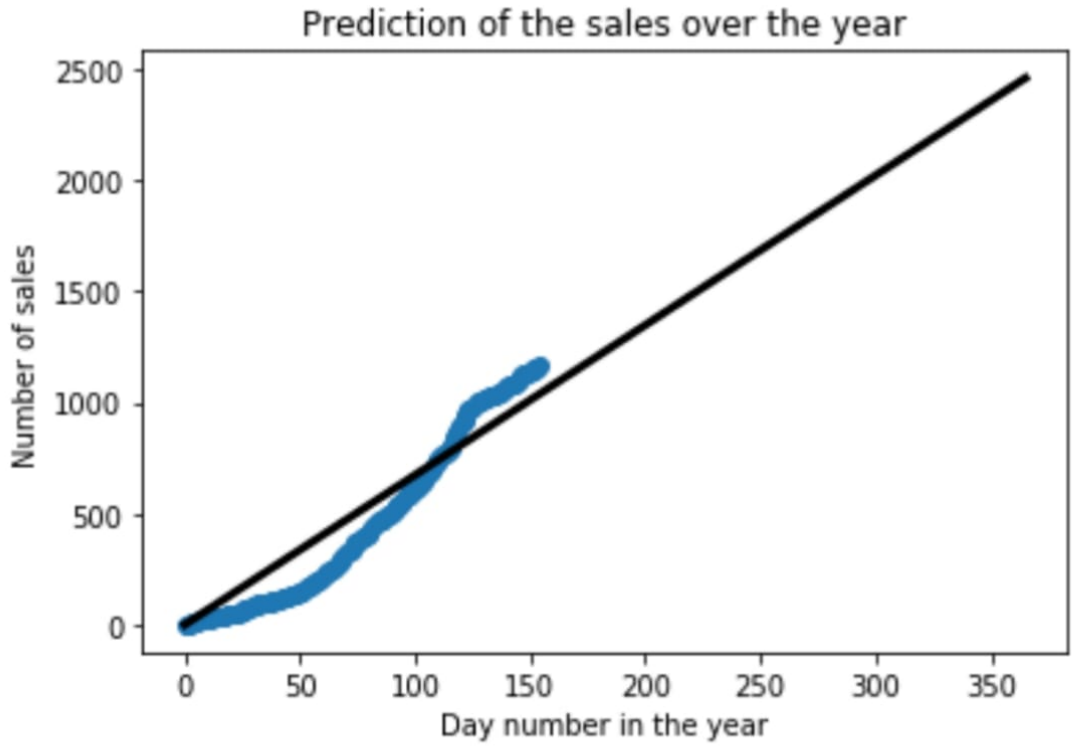
  Sí. Particularmente en la sección donde se analizó la efectividad de las publicidades, se vió que hay campañas de marketing cuya efectividad es una mínima fracción de las que son más efectivas. Dichas campañas representan una pérdida para la compañía.

- **¿Se podría mejorar algún aspecto de la página web de la empresa?**

  En los análisis de ventas y de popularidad, los modelos iPhone demostraron ser de los mas solicitados. Sin embargo, en la barra de navegación de la página web, no existe ningún filtro que nos permita acceder rápidamente a los modelos que la gente más parece querer. Dicha cuestión se podría corregir.

9 Análisis predictivo

En esta parte quisimos tratar de encontrar un modelo de ventas. Escogimos la regresión lineal porque la curva parece un poco una línea, y es un modelo simple. Es más, creemos que si utilizariamos otra regresión polinomica nos apegariamos mucho a la curva y no generalizaria. Elegimos arreglar la intersección en 0, porque al comienzo del año la cantidad total de ventas era 0. Por lo tanto quisimos extender la curva en orden que podamos predecir las ventas para el final del año. En efecto, para predecir la cantidad es necesario tener una visión interna de las ventas del año. Es también importante para definir objetivos para la compañía.




```
Estimated number of product sales over the year: 2461
```