FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Trabajo Práctico 1: Análisis Exploratorio de Datos

[75.06 / 95.58] Organización de Datos
Segundo cuatrimestre de 2018

Grupo Datatouille

Alumno	Padrón	Mail
Bojman, Camila	101055	camiboj@gmail.com
del Mazo, Federico	100029	delmazofederico@gmail.com
Hortas, Cecilia	100687	ceci.hortas@gmail.com
Souto, Rodrigo	97649	rnsoutob@gmail.com

https://github.com/FdelMazo/7506-Datos/
https://kaggle.com/datatouille2018/7506-TP1/

Curso 01

- Argerich, Luis Argerich
- Golmar, Natalia
- Martinelli, Damina Ariel
- Ramos Mejia, Martín Gabriel

Índice

1. Introducción

Se propone analizar en el presente informe los datos obtenidos de usuarios que visitaron `www.trocafone.com`, un sitio de e-commerce de compra y venta de celulares reacondicionados, con operaciones principalmente en Brasil. Para ello, la empresa Trocafone nos proporcionó acceso a los datos a través del archivo `events.csv`.

El objetivo principal de este informe es poder realizar un análisis exploratorio abarcativo donde a medida que se exploren los datos se vayan encontrando tanto las preguntas como las respuestas a hacerse. Se propone específicamente:

- Descubrir features en el campo `model`
- Identificar patrones de usuarios que realizan checkouts y conversiones
- Analizar las búsquedas que realizan los usuarios y las *keywords* utilizadas
- Analizar los distintos lugares de dónde se originan las visitas a Trocafone
- Descubrir features jerarquizando alguno de los campos disponibles
- **agregar más items a medida que desarrollamos el informe**

Finalmente, entre lo descubierto en el análisis exploratorio y los ites marcados, se busca obtener un listado de *insights* aprendidos sobre los mismos y con ellos realizar un aporte a la empresa Trocafone con datos que sirvan para mejorar sus servicios.

2. Información general sobre los datos

Lo primero y básico a analizar es la estructura general de los datos proporcionados, para comenzar a tener una idea de que se tiene y que se puede hacer con ello. Se observa que:

- Estos datos corresponden al período de tiempo comprendido entre el 1 de enero del 2018 al 16 de junio del 2018.
- Son 1011288 registros con 23 atributos, no siempre todos completos.
- Son 1011288 registros con 23 atributos, no siempre todos completos.

Los atributos son:

- **timestamp**: Fecha y hora cuando ocurrió 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**: Téŕminos de búsqueda utilizados en el evento.
- **static_page**: Identificador de página estática visitada.
- **campagn_source**: Origen de campaña, si el tráfico se originó de una campaña de marketing.
- **text_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 (New) o por un usuario que previamente había visitado el sitio (Returning) según 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 genero 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.

Es en este momento del análisis donde se tienen que hacer las configuraciones necesarias sobre el set de datos para poder trabajar mejor más tarde. Las operaciones realizadas incluyen:

- **Conversión de tipo de datos**: Teniendo en cuenta que al cargar el set original no se infiere el tipo de cada dato (que atributo es numérico, que atributo es categórico, etc), se convierten los datos para tratarlos por su tipo original. Esto tiene como ventaja principal el ahorrar memoria, ya que en vez de tener variables que almacenan objetos genéricos (y ocupan un bloque genérico de memoria) ahora se pueden tener especificamente categorías, números, valores booleanos y más. Un particular caso que es de gran ayuda es el de tratar el atributo ‘timestamp‘ como una variable del tipo ‘datetime‘.

- **Lidiar con los nulos**: No todos los registros tienen todos los atributos completos, por motivos obvios (por ejemplo, un evento de compra de producto no tiene asociado una búsqueda de palabras). En la transformación de tipos hay que lidiar con estos, y se tomaron decisiones como que el SKU de un producto 'Not a Number' es el SKU '0.0', así permitiendo que el atributo SKU sea numérico.
- **Data Mining**: Se generan nuevos sets de datos y se extraen atributos importantes de los proporcionados. Por ejemplo, dividir el atributo de tiempo en atributos de mes, día y hora.
- **Limpieza de datos**: Una decisión importante es la de cuando un dato es invalido de entrada. En este caso tomamos como un error de tracking cuando la misma venta es registrada dos veces por el mismo usuario en un corto plazo de tiempo, ya que se toma como algo muy improbable. Estos registros son eliminados.

2.1. Hipótesis sobre el truncamiento de los datos

Se sabe que el dataset proporcionado no representa el conjunto total de datos de todos los eventos realizados por los usuarios en el período de tiempo determinado. Es por esta razón que se busca elaborar una hipótesis en base al criterio con el que se fijó la selección de los datos. A partir de un análisis de los mismos se observó que todos los usuarios registrados en el dataset realizaron al menos un checkout, por lo que la base de datos original se truncó. Sin ir más lejos es evidente que no todos los usuarios que ingresan al sitio de Trocafone van a realizar un checkout.

A esta información se le adiciona que los datos de entrada son solamente el tráfico del *sitio web* de Trocafone, y no de la empresa entera, que tiene más actividad que la del sitio. Por ejemplo, venderle a sitios terceros [1].

Con estos dos datos en mente, es importante remarcar que las conclusiones a las que se llegará en el desarrollo del Trabajo se basan en un sector segmentado de los datos, y que estos datos son un sector segmentado de la empresa, por lo que en ciertos aspectos del análisis no se podrá arribar a conclusiones fundadas sobre la totalidad de los servicios de Trocafone, y que si por momentos la información parece poca para el tamaño de la empresa, esta solo representa el sitio web.

3. Análisis de eventos

En esta sección se propone analizar los distintos tipos de eventos realizados por los usuarios de Trocafone.

El campo `event` puede adquirir distintos tipos de valores categóricos que se describen como sigue:

- **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.

[1] https://medium.com/trocafone/el-maravilloso-mundo-de-trocafone-5bdc5761856b

- **ad campaign hit**: El usuario ingresa al sitio mediante una campana 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.
- **static page**: 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.

3.1. Conversion rate

En primer lugar se analiza la métrica llamada *conversion rate* o tasa de conversión debido a su importancia en cualquier negocio de e-commerce.

La tasa de conversión es el porcentaje de visitantes que completan un objetivo deseado (en este caso realizar una compra de celular) sobre el total de visitantes. En otras palabras es la razón entre las conversiones y el total de eventos.

Tomando todos los datos del dataset dicha tasa de conversión es de `0,096`.

Se busca analizar la evolución de la *conversion rate* a lo largo del tiempo. Para ello se realiza un gráfico de la conversion rate a lo largo de las semanas del año en la figura 1.

Las mayores tasas de conversiones se registran en las semanas 1,4 y 11, es decir, enero y marzo. Para observar mejor este fenómeno se realiza otro gráfico (figura 2)que muestre la evolución de la tasa de conversión a lo largo de los meses.

Se refuerza la teoría de que enero y marzo fueron los meses de mayor tasa de conversión. Así mismo, dicha tasa se mantiene estable por 4 meses para luego tener una baja en los meses de mayo y junio. Más adelante se analizarán estos dos meses en profundidad.

3.2. Frecuencia de eventos

Se analiza qué tipo de evento es el más frecuente en el dataset. Para ello se grafica la cantidad registrada de eventos en función de los distintos tipos de eventos.

Se observa en el gráfico 3 que la mayor cantidad de eventos se relacionan a la vista de un producto, lo cual era previsible ya que Trocafone es una plataforma de e-commerce y ver productos constituye su principal función como sitio.

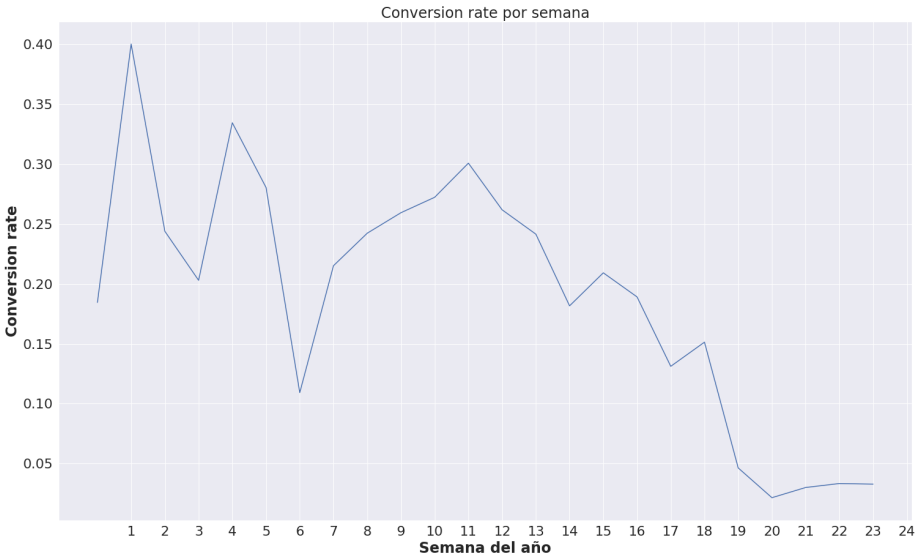

Figura 1: Tasa de conversión a lo largo de las semanas del año

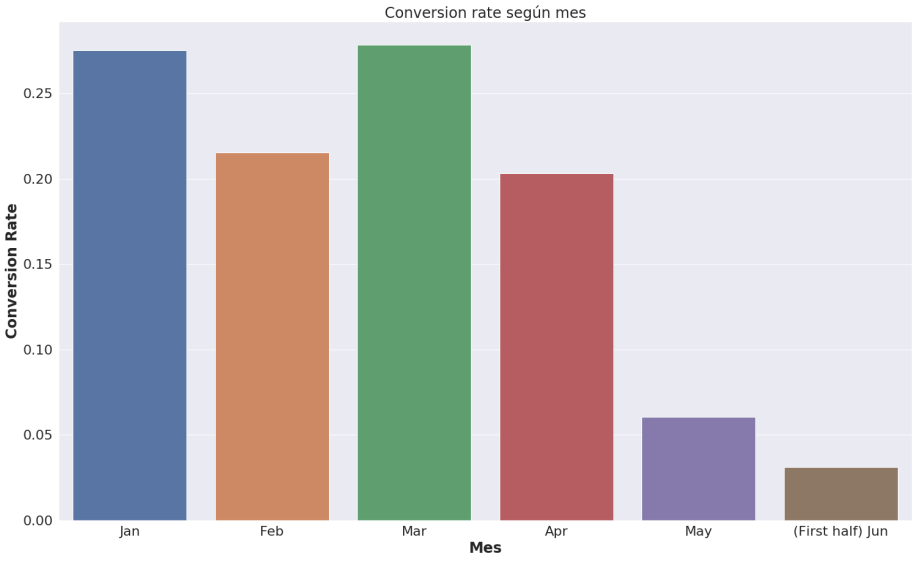

Figura 2: Tasa de conversión a lo largo de los meses del año

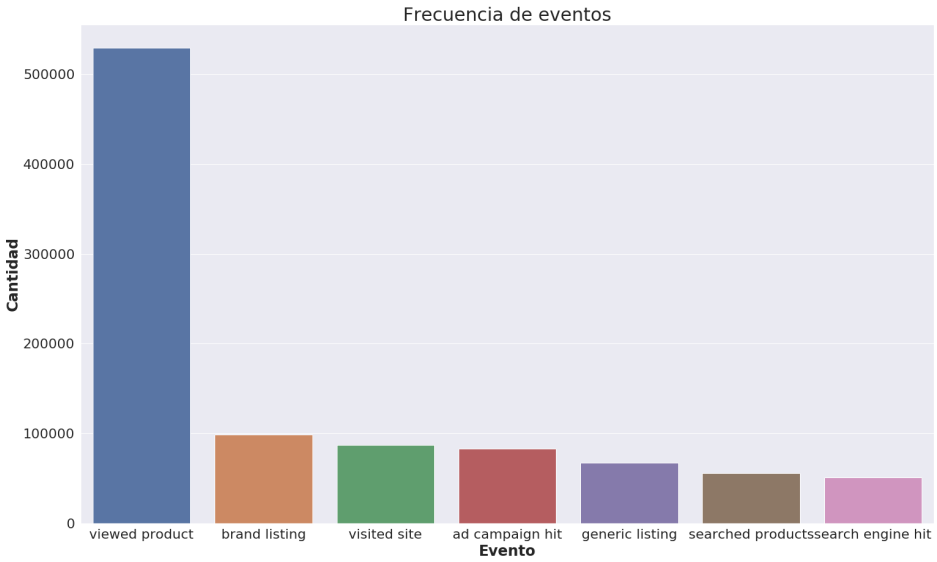

Figura 3: Frecuencia de eventos

3.3. Evolución de los eventos a través del tiempo

3.3.1. Tráfico del sitio de acuerdo al mes y al día

Otro aspecto a analizar es la cantidad de eventos producidos en cada día de la semana y del mes. Se busca detectar si se mantiene algún comportamiento específico a lo largo de los meses o si la cantidad de eventos registrada depende de algún factor temporal. Un patron esperado a encontrar es el de si hay más visitas o compras de celular en las primeras semanas del mes, lo cual coincidiría con el pago de sueldos mensuales.

Para realizar este gráfico los datos fueron normalizados para evitar llegar a la conclusión que el mes con una mayor cantidad de eventos es el mes con más eventos por día [2].

Se desprende del gráfico **??** que la cantidad de eventos registrada no presenta ningún comportamiento específico. Se observa que dicha cantidad aumenta en la segunda quincena de cada mes pero se considera que la diferencia con el resto de los días no tiene la magnitud suficiente como para extraer alguna conclusión fundada.

3.3.2. Tráfico del sitio de acuerdo al mes y al día de la semana

Sin haber encontrado nada acerca del número de día, se buscar ahora analizar si algún día de la semana se registra una mayor cantidad de eventos.

[2] Lo cual sería un error muy grave en el análisis, famosamente conocido gracias a la ecuación de de Moivre

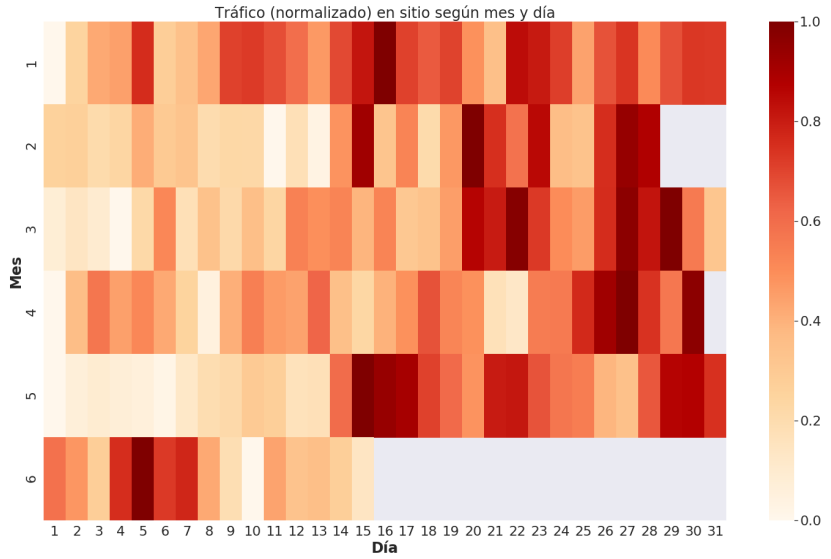

Figura 4: Eventos segun mes y día

Se realiza el gráfico 5 del mismo estilo que el anterior y se normaliza por las mismas razones.

Es notable que durante los días hábiles de la semana el tráfico es mucho mayor que al fin de semana. Esto puede deberse a que los fines de semana suelen ser días de descanso, donde la gente puede no estar pensando en realizar una compra, además de no poder retirarla. En la semana aumenta el tráfico debido a que el envío o el retiro del celular puede realizarse en el momento.

3.3.3. Tráfico del sitio según mes

Habiendo analizado las semanas, ahora se hace un enfoque más global, buscando patrones de tráfico según el mes. En este apartado se busca analizar si en algún mes se registró una mayor cantidad de eventos o si la distribución de las visitas fue uniforme a lo largo del tiempo.

La figura 6 muestra que los meses de mayo y junio registraron una cantidad notablemente mayor de eventos. Este resultado llama la antención por lo que se analiza en mayor profundidad en la próxima sección. Es importante destacar que esta evolución es inversa a la de la tasa de conversión. De estos dos datos se concluye que en mayo si bien no hubo tantas ventas, sí aumentó mucho la cantidad de eventos.

3.3.4. ¿Por qué mayo y junio registran una mayor cantidad de eventos?

Para tratar de encontrar una respuesta a esta pregunta se centra el análisis en estos meses y se estudian los tres eventos principales del dataset: `conversion`,

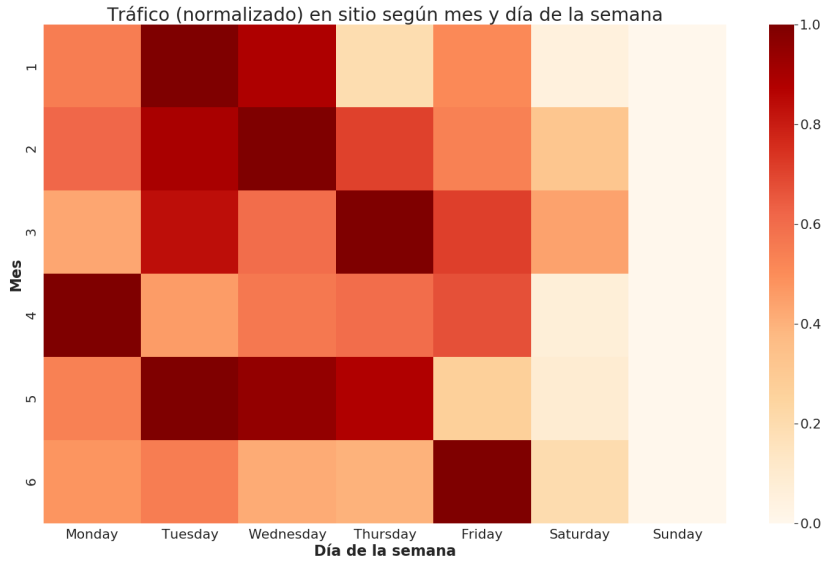

Figura 5: Eventos segun mes y día

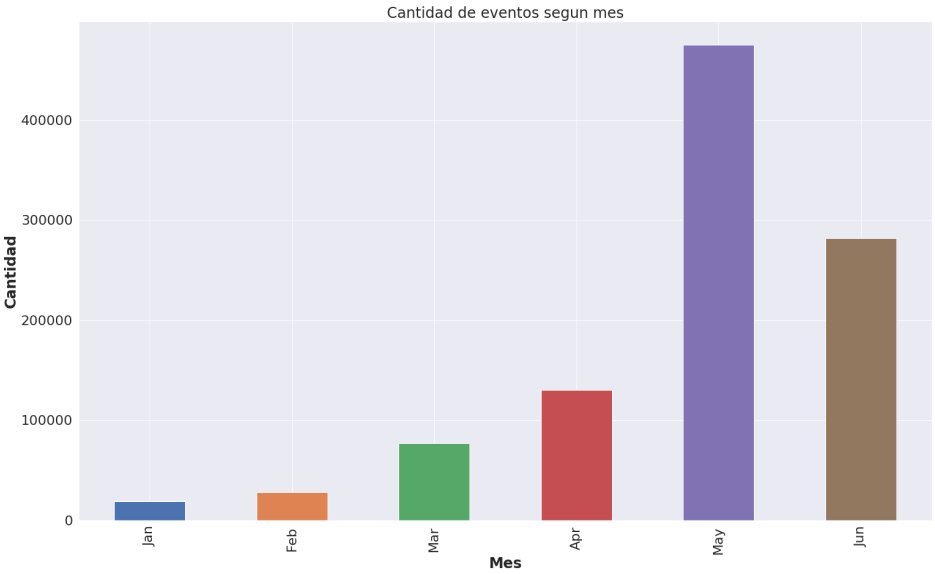

Figura 6: Eventos segun mes

`checkout` y `viewed products`.

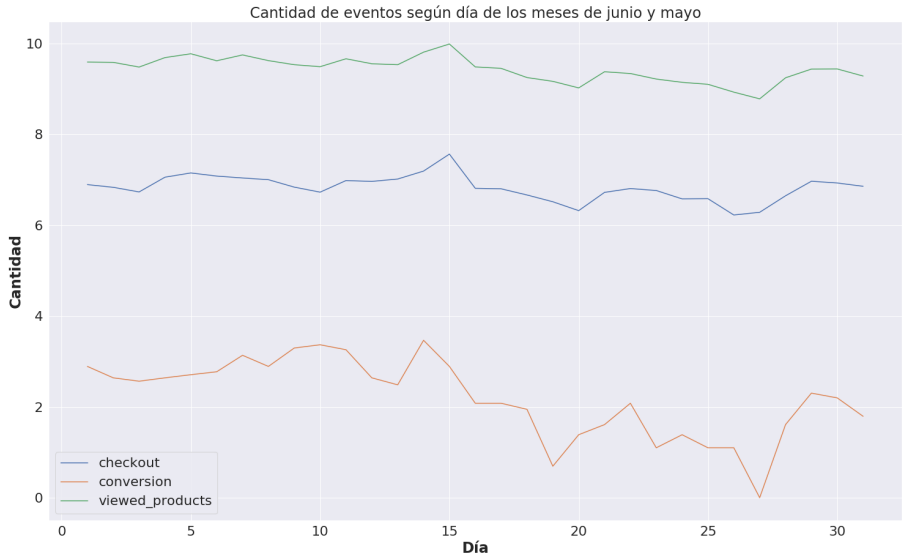

Figura 7: Conversiones, checkouts y viewed products a lo largo de los días del mes de mayo y junio en éscala logarítmica

Los tres eventos presentan su máximo alrededor de los días 14 a 16. Como Trocafone es del país de Brasil y el mayor tráfico proviene de allí, lo cual será verificado posteriormente, se infiere que puede deberse a alguna promoción lanzada en la plataforma o en el mismo país. Esto no puede concluirse con certeza debido a la falta de información en internet y en la plataforma de promociones pasadas.

3.3.5. Hora de mayor cantidad de conversiones y checkouts

En un intento de encontrar un patrón por parte de los clientes se grafica la cantidad de conversiones y de checkouts en función de las horas del día. Se busca determinar la hora en la que ambas confluyan en su máximo para analizar el motivo por el que dicha hora registra mayor tráfico.

Se observa que en ambos gráficos confluyen en su máximo a las 19 hs. Esto puede significar que la mayoría de los usuarios cuando vuelven del trabajo o están finalizando su día deciden realizar conversiones o checkouts. La diferencia entre ambos gráficos es que la cantidad de checkouts realizados se mantiene relativamente constante en las segundas 12 hs del día mientras que las conversiones son mucho menores y presentan picos más marcados en los horarios de la tarde-noche.

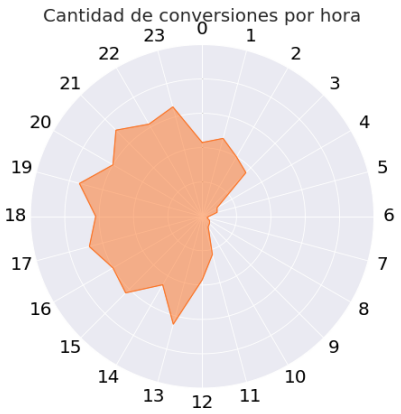

Figura 8: Conversiones a lo largo de las horas del día

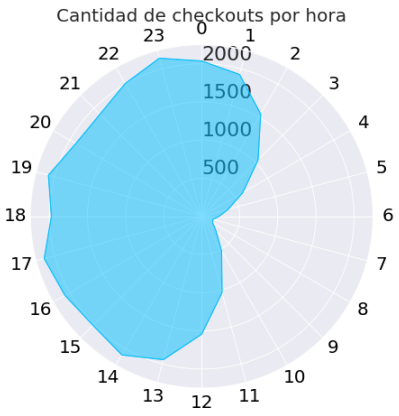

Figura 9: Conversiones a lo largo de las horas del día

4. Análisis geográfico

En este apartado se busca analizar las ciudades, países y regiones de dónde provienen los distintos tipos de eventos. Trocafone es una empresa que inició en Brasil y expandió sus comercios a Argentina en el 2015, por lo que se deduce que probablemente Brasil sea la zona de mayor influencia en los eventos.

4.1. Países que registran mayor cantidad de eventos

Se corrobora en 10 la teoría inicial por lo que se procedió a eliminar a Brasil del gráfico, generando 11, para poder observar qué otros países intervienen en la página de Trocafone y en que diferencia de magnitud y orden lo hacen. Estados Unidos supera en una amplia cantidad la influencia en la página a Argentina, a pesar de ser una de las sedes de la empresa. Esto puede explicarse debido a que la gran mayoría de los eventos no son conversiones, por lo tanto es factible que cualquier persona de los Estados Unidos busque celulares en la plataforma, sin llegar a registrar un evento de tipo `checkout` o `conversión`.

4.2. Regiones y ciudades de Brasil que registran mayor cantidad de eventos

Se procede a analizar qué ciudades y regiones de Brasil son las que registran mayor cantidad de eventos. Para ello se grafica las regiones con una mayor cantidad de visitas. Se observa que las tres regiones con la mayor cantidad de eventos (San Pablo, Minas Gerais y Rio de Janeiro) están sobre la costa del sudeste. Para visualizar esto de una mejor manera se realiza un gráfico que muestra las ciudades de Brasil más visitadas y se verifica que la mayoría de los eventos se producen sobre la costa sudeste.

Lo que presentan las figuras 12 y 13 tiene sentido, considerando que sobre el mayor blanco del gráfico es donde esta la selva brasileña.

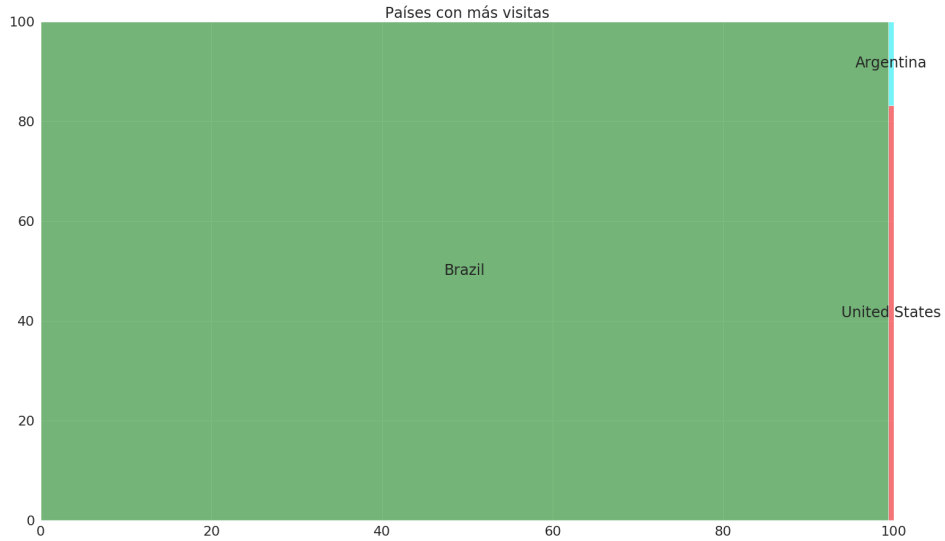

Figura 10: Países de mayor tráfico de la página

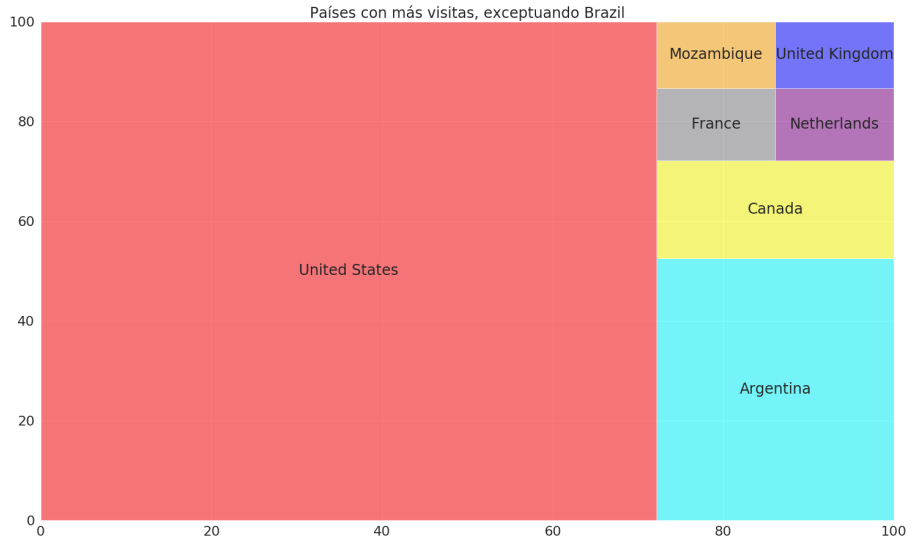

Figura 11: Países de mayor tráfico de la página sin contar a Brasil

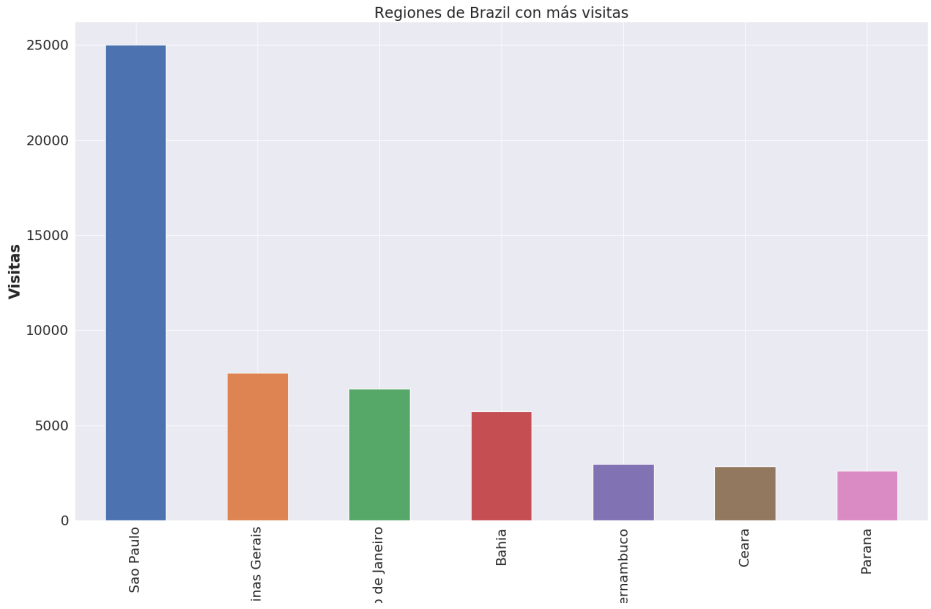

Figura 12: Regiones de Brasil con mayor cantidad de eventos

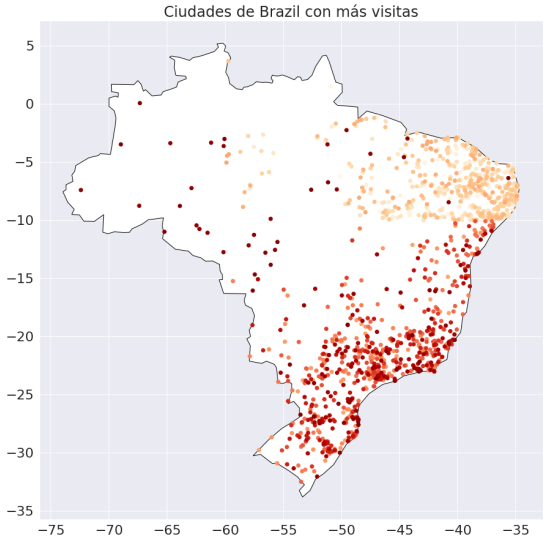

Figura 13: Ciudades de Brasil con mayor cantidad de eventos

5. Análisis de búsquedas

La idea de este apartado radica en analizar los términos que buscan los usuarios en la plataforma y así identificar ciertos patrones de búsqueda como

por ejemplo cuál es el modelo de celular más buscado. Este análisis se dividirá en dos:

- Términos ingresados en el buscador: se utiliza la columna `search_term` del dataframe.
- Productos buscados en la plataforma: se utiliza la columna `event` del dataframe para buscar aquellos eventos que correspondan a `searched_product`.

5.1. Términos ingresados en el buscador

Se realiza un gráfico para visualizar a grandes rasgos los términos más buscados por los usuarios. Se busca tener una idea aproximada de los modelos de celular más requeridos o deseados por los usuarios. Figuran en el gráfico los términos que fueron buscados como mínimo 300 veces, un número impuesto para fijar un límite mínimo de búsquedas para ser considerado de los más buscados. De no fijar este límite el gráfico estaría sobrecargado y sería difícil de interpretar.

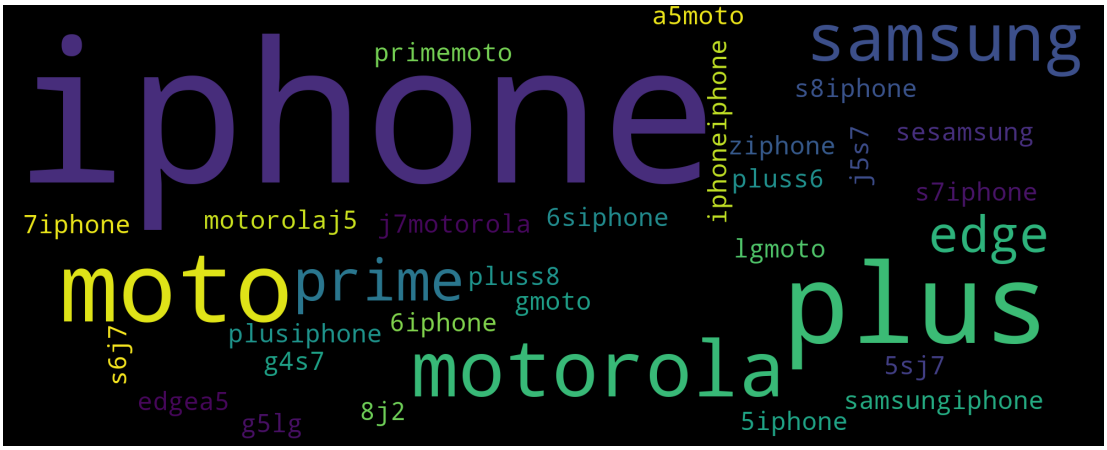

Figura 14: Términos más buscados por los usuarios de Trocafone

Los términos más buscados son iPhone, Motorola y Samsung. Esto era completamente esperable debido a que son las marcas que dominan el sector tecnológico.

5.2. Productos buscados en la plataforma

En esta sección se busca obtener los productos más buscados. Esta búsqueda es más específica que la anteriormente mencionada debido a que corresponde a un producto puntual, no el nombre de su marca, buscado por la interfaz del sitio. De esta manera los productos más buscados son los que se detallan en la tabla 1 y se representan en el gráfico que le sigue.

Se concluye con el gráfico 15 en esta sección que si bien `iPhone` y `Motorola` eran los términos más buscados en la plataforma, no sucede lo mismo con los productos buscados ya que todos corresponden a la marca `Samsung`. Esto puede

sku	sku_name
3371	Samsung Galaxy S6 Flat 32GB Dourado (Bom)
2777	Samsung Galaxy S4 i9505 16GB Preto (Bom)
6357	Samsung Galaxy J5 16GB Preto (Bom)
6413	Samsung Galaxy J7 16GB Dourado (Bom)
6371	Samsung Galaxy J5 16GB Dourado (Bom)

Cuadro 1: SKUs más buscados y su nombre

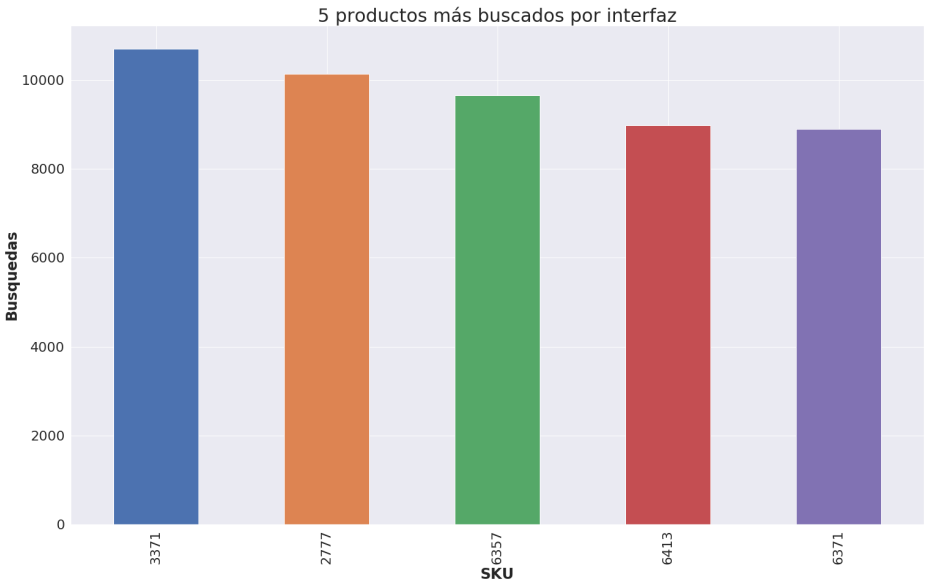

Figura 15: Productos más buscados por los usuarios de Trocafone

deberse a un tema de la calidad que ofrece dicha marca o su precio, probablemente más conveniente. Los iPhone se caracterizan por tener un precio difícil de acceder por lo que es probable que sea buscado como término para ver las diferentes opciones globalmente pero que no muchas veces se busque un producto determinado de dicha marca.

A. Ejecución

El trabajo fue realizado en Anaconda [3]. Para poder replicar el trabajo, hay que también instalar las siguientes librerías adicionales:

- Squarify [4]: Para los treemaps.
- Geopandas [5]: Para poder graficar sobre mapas geográficos.
- Wordcloud [6]: Para poder visualizar los términos más buscados.

Estos pueden ser instalados con los siguientes comandos:

```
pip install squarify
conda install -c conda-forge geopandas
conda install -c conda-forge wordcloud
```

B. Datasets adicionales incoporados para el análisis

Se utiliza adicionalmente un dataset del mapa de Brasil y sus ciudades, para el análisis geográfico. Este fue sacado de Geonames [7], una base de datos publica de paises y regiones del mundo.

[3] https://anaconda.org/
[4] https://github.com/laserson/squarify
[5] http://geopandas.org/
[6] https://github.com/amueller/word_cloud/
[7] http://www.geonames.org/