**Introducción:**

El presente informe consiste en el estudio y análisis de la venta bebidas alcohólicas, principalmente cervezas, y productos relacionados a las mismas a través de una plataforma de comercio electrónico. Se partirá de un set de datos con el registro de todas las compras efectuadas en un período determinado en dicha plataforma, al cual se le hará un tratamiento previo que tendrá como objetivo adecuar el set de datos al análisis extrayendo la información que no sea relevante. Luego, se realizará un análisis exploratorio de datos que permitirá visualizar la información contenida dentro del set y lograr comprender en gran medida el comportamiento de compra/venta dentro de dicha plataforma de Ecommerce. En tercer lugar, se aplicará un modelo de aprendizaje no supervisado con el fin de determinar grupos similares entre las muestras, a los cuales finalmente se les aplicará un último análisis exploratorio para obtener conclusiones.

**Descripción del dataset:**

El set de datos utilizado es de carácter privado, y consiste en compras efectuadas a través de un sitio web de comercio electrónico. Cada registro (de ahora en más: “sample”) del mismo representa una operación de compra, incluyendo en cada caso campos (de ahora en más: “features”) con información detallada sobre cada aspecto de ella (económico, transaccional, relacionado al producto, y relacionado a las características del usuario). Se han contabilizado un total de 359388 samples con 144 features cada una.

Sin embargo, durante el primer análisis se descubre que la gran mayoría de las features no cuenta con información (principalmente las asociadas al enriquecimiento de datos a partir de la información de la persona) o muestran información repetida, por lo que se procede a quitar las mismas quedandonos únicamente con aquellas que poseen información relevante sobre la operación de compra.

De esta manera, cada compra realizada tendrá la siguiente información asociada: id, nombre del producto comprado, cantidad comprada, marca del producto, categoría del producto (ej: cerveza, merchandising, otros), precio del producto, monto bruto de la compra, monto neto de la compra, descuento, cargo de delivery, sku del producto, estado de la compra, país, fecha de la compra, hora de la compra y hora en la que se cargó el registro.

**Análisis exploratorio de datos (EDA):**

El objetivo de dicho análisis consiste en comprender el comportamiento de las compras en Ecommerce, lo cual implica saber qué productos son los más adquiridos, de qué rubro, de qué marca, si presentan estacionalidad, cuanto se compra por usuario, etc. Para esto se procede a un proceso de limpieza del dataset en el que se eliminarán aquellas features que contengan valores nulos o una cantidad de valores nulos que justifique su extracción y aquellas que no aporten información relevante para el análisis posterior. De esta manera se extraen 114 features, quedando un dataset de 27 features, con la posibilidad de quitar otras a lo largo del análisis.

En segundo lugar, se analizará la existencia de outliers dentro de distintas features. En primer lugar, la feature denominada *‘quantity’*, la cual indica la cantidad comprada del producto correspondiente a la compra, permite observar compras con cantidades mayores a 60 (con un máximo de hasta 4000), las cuales representan menos del 0.1% del total de las operaciones, por lo que se procede a eliminarlas. En segundo lugar, quitaremos los outliers de precio, en los que encontramos un máximo de $999999 que se deduce como una operación de prueba por parte de los responsables del sitio. Finalmente, quitaremos los outliers de unit\_gross\_revenue\_local (gasto de la compra), ya que se observan gastos extraordinariamente elevados y gastos negativos.

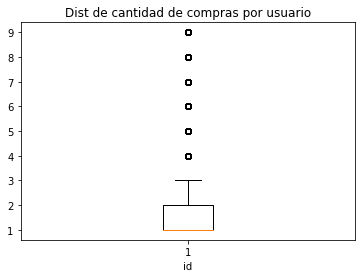
Luego, se procede a ordenar el dataset por fecha de compra, con lo cual puede observarse que el primer registro corresponde a febrero de 2016, y el último a agosto de 2018. Esto se debe a que el sitio comenzó su operación en 2016, mientras que el corte de datos en 2018 responde a limitaciones técnicas el momento de obtención del dataset original. Dado entonces que los registros de estos años no son representativos y desvirtuarían resultados a la hora de evaluar estacionalidad, y que 2017 es el único año disponible en su totalidad (con una cantidad adecuada de samples para aplicar algoritmos de Machine Learning), se realizará un filtrado del dataset para trabajar únicamente con registros de este ultimo.

Una vez definido el dataset para 2017, se realiza un filtrado por país, de los cuales se analizarán los correspondientes a Argentina. De esta manera, resulta en un dataset de 60896 samples y 30 features (3 más que antes dado que se han creado 3 features nuevas que corresponden a año, mes y día). Adicionalmente, se extraen dos nuevas variables del dataset ya que se observa que las mismas no proporcionan información relevante y poseen mucha cantidad de datos nulos. Ellas son: sku y subcategory. El dataset se reduce a 28 features.

Al comenzar con el análisis exploratorio de datos se analiza la cantidad de compras por usuario, lo cual se logra contabilizando la cantidad de apariciones de cada id. Puede observarse que una misma id ha aparecido, o lo que equivale a decir ha comprado, 3135 veces, siendo su siguiente 124, por lo cual se concluye que puede deberse a compras realizadas por la empresa dueña de la plataforma a modo de prueba. Dichos samples se extraen del dataset y este se reduce a 57761.

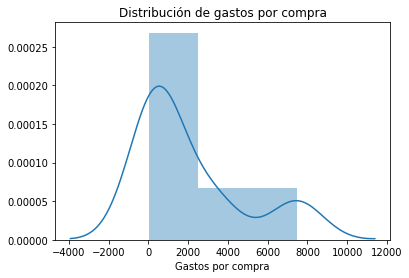
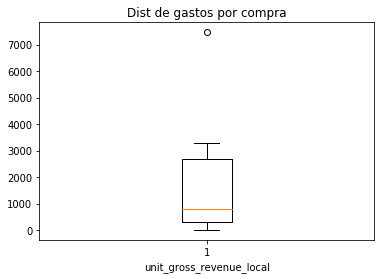
Luego, se analizan los productos más vendidos, pudiendo hallar que, de los primeros 20 registros del ranking, solo 3 corresponden únicamente a la categoría cervezas. Se puede observar, también, que, de las cervezas adquiridas la gran mayoría son de carácter importadas, por lo cual se procede a listar todas las marcas que abarca el dataset y a reemplazar a aquellas que sean importadas con el label “importada”, lo cual permitirá segmentar de una mejor manera para el análisis. De esta manera, las labels correspondiente a “Brand” son: Stella Artois, Patagonia Brewing Co., Corona, BevyBar, Cervecería y Maltería Quilmes, e Importada.

El siguiente paso consiste en analizar la cantidad de compras por usuario (id) para apariciones menores a 10, pudiendo observar lo siguiente:



La mediana de las compras es de una unidad, pudiendo ser hasta dos ó tres y, en muy pocos casos, mayor. Esto significa que **la mayoría de los usuarios solo ha realizado una compra**.

Luego, analizamos la distribución de los gastos por compra:

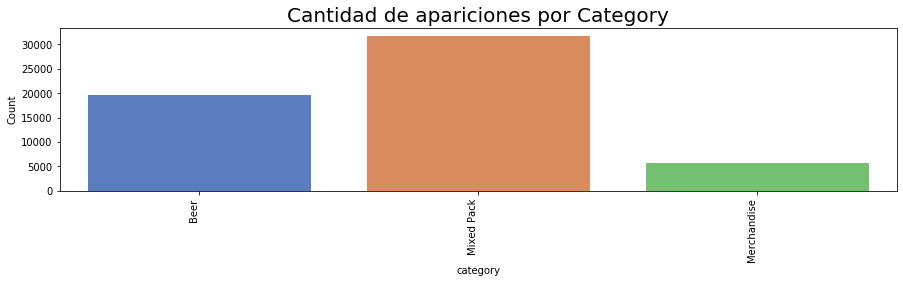
****

**std**: 714.26 **min**: 2.2517 **25%:** 190 **50%:** 650.176 **75%:** 955 **max**: 15732

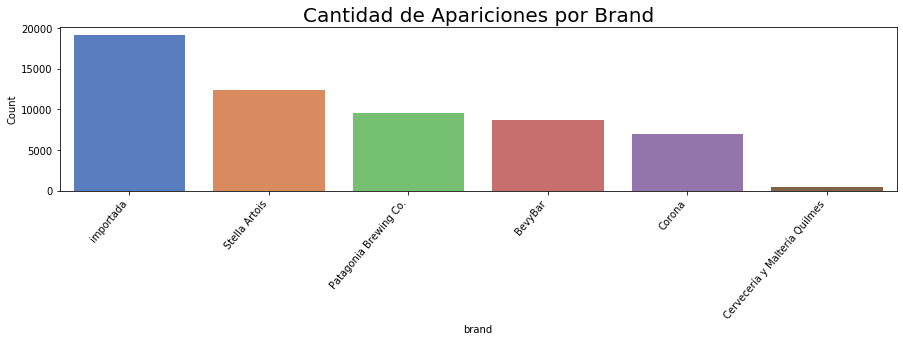
Debido a que ciertas compras se encuentran en estado *grand total is zero* (lo que implica que consisten en regalos u operaciones no habituales) se procede a la extracción de las mismas.

Observando las compras por producto (“name”), se puede verificar que los productos más comprados son “Pack Conservadora Corona” y “Copas Stella Artois”, lo que se corresponde con las grandes campañas publicitarias efectuadas por la empresa durante los ultimos años.

El EDA continúa con el análisis de las marcas más adquiridas, así también como las categorías más adquiridas. Para esta última variable, lo más adquirido son Mixed Packs (cerveza + merchandiseing), seguido por cerveza y merchandising. También, pero con muy baja cantidad, están las categorías Wine, Spirits, Unknown, y Home Breweing, las cuales representan menos del 1,8% del total, de modo que se procede a extraerlas y trabajar únicamente con las tres primeras. De esta manera, puede observarse lo siguiente:



Lo más adquirido en el Ecommerce son Mixed Pack (55.6%), seguido por Beer (34.58%) y Merchandise (9.8%).

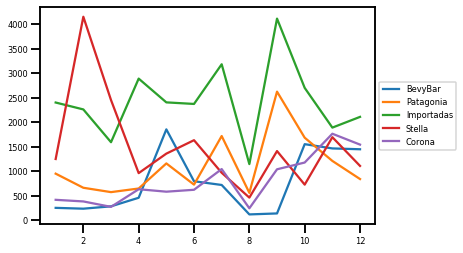


Lo más adquirido en el Ecommerce son marcas importadas (34.6%), seguido por Stella Artois (21.31%), Patagonia (16.6%), BevyBar (14.6%), Corona (12%) y Quilmes (0.82%).

Luego, nos facilitaremos de tablas pivot para analizar qué rubros son más adquiridos dentro de cada marca. Dicho análisis se realizará mensualmente lo cual, además, permitirá ver la evolución de las ventas. De dicha tabla podemos observar los siguientes puntos más relevantes:

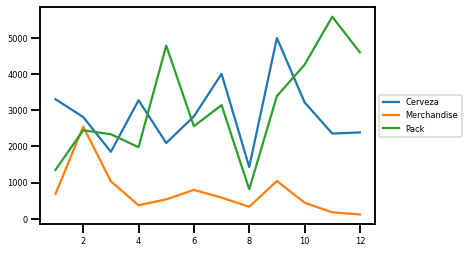
* De las marcas importadas, se compra, en su gran mayoría, cerveza.
* De las marcas nacionales, se compra, en su gran mayoría, mixed pack y merchandising.
* De Patagonia se adquiere una gran proporción de Mixed Packs, seguido por cerveza, y una escasa participación de Merchandising. Caso similar ocurre con Stella Artois, aunque este sí adquiere una significativa proporción de merchandising en lugar de cerveza.

A continuación, se analizará la evolución de las ventas tanto a nivel general como por distintos nichos. El objetivo es verificar la existencia (o no) de estacionalidad y poder comparar con registros de otros años. Quedará para un análisis posterior fuera de este informe la estacionalidad definitiva. Observaremos primero lo sucedido con las marcas:



A nivel general, podemos afirmar que las cantidades son similares en la última etapa del año. Por otro lado, Corona y BevyBar presentan una evolución similar entre sí. Stella Artois cuenta con un nivel elevado de ventas en la primer época del año, lo cual puede deberse a campañas de publicidad.

Vemos lo ocurrido para las diferentes categorías:



La categoría “Merchandise” tiene un comportamiento similar al de Stella en el gráfico anterior. Esto puede ser explicado por lo encontrado en las tablas pivot. Caso similar sucede con Cerveza e Importadas.

Modelo de aprendizaje:

Una vez finalizado el EDA se procede a aplicar un aprendizaje no supervisado (Clustering). Se aplicarán tres Clusterizaciones: uno para todo el dataset, uno para los productos nacionales (productos *core* del negocio) y otro para los productos importados.

Para ello se generan, previamente, dummies para las features de Category y Brand, que serán relevantes para este analisis. (Para el caso de las importadas, se procede a retirar aquellas que hayan tenido menos de 60 adquisiciones ya que la variedad de marcas es muy elevada).

En primer paso se procede a una reducción de la dimensionalidad de cada dataset vía PCA, un método de combinación lineal de variables que permite obtener nuevas features que expliquen en forma más adecuada la variabilidad de los datos.

Los resultados de las tres clusterizaciones son las siguientes:

Cluster total:

Cluster importadas:

Cluster nacional: