**Exploración y Curación**

*Precios Claros*

A continuación se presenta un resumen de las principales decisiones tomadas para la exploración y curación de los dataset “precios”, “sucursales” y “productos” utilizados para el cumplimiento de nuestro objetivo de análisis: Comparar si un producto es más caro o mas barato que otro en términos de un bien numerario.

# Rutina de Curación

Si bien el checklist de curación estudiado en clase tiene varios incisos, algunos no son aplicables en nuestro caso de análisis, en consecuencia sólo serán mencionados.

## Importando los datos

### 1.1. Verificar si no hay problemas en la importación

### Algunos de los chequeos realizados al importar los registros son:

* Observar registros aleatorios en cada una de las bases de datos.
* Indagar que cada columna tenga el mismo formato ( en el caso de los producto\_id hay códigos de solo números y otros con guiones, se verificó que en el dataset de precios también existen productos id con guiones.

### 1.2. Asegurar de tener ids/claves únicas

* Se verifica que en el dataset de productos y sucursales existan ids únicas, y se indaga que no existan duplicados parciales. Es decir que tengan id distintos a mismos productos, mismas sucursales y mismos precios. Este punto se detalla en las inconsistencias encontradas en puntos posteriores.

### 1.3. Despersonalizar datos y guardarlos en un nuevo archivo

-No necesario

### 1.4. Nunca modificar los datos crudos u originales

-Los cambios y limpieza del dataset se guardan en los dataset precios\_limpios, sucursales\_limpios y productos\_limpios.

## Pasos necesarios

### 2.1. Etiquetas de variables/columnas: no usar caracteres especiales

Se realizó la comprobación correspondiente.

### 2.2. Tratar valores faltantes

#### Set de datos de precios

-Existen 7623 registros que no tienen precios (0.34% del dataset), estos registros son 2273 productos distintos de 5 sucursales distintas y de las 5 fechas distintas.

-Las demás variables del dataset no tienen valores nulos

¿qué hacer con estos valores?

Se ha decidido eliminar estos precios ya que no brindan informacion para ser analizados.

#### Set de datos de productos:

-Existen 2 productos que no tienen marca ni presentación. Si bien en el archivo precios no fueron reportados, igualmente serian excluidos en un futuro ya que no brindan información pertinente para ver si un producto es más caro o mas barato si no tiene identificación.

-Las variables Categorías 1, categorías 2 y categorías 3 tienen todos los datos en blanco, esas variables se eliminan.

¿qué hacer con estos valores?

Se ha decidido eliminar los 2 productos que no tienen ni marca ni presentación ya que no brinda ningún tipo de información para nuestro análisis, y las variables que tienen el 99% de los sus valores nulos se eliminarán en el paso de variables irrelevantes.

#### Set de datos de sucursales:

No hay valores faltantes

### 2.3. Codificar variables: las variables categóricas deben ser etiquetadas como variables

#### Set de datos de precios

Se han codificado las fechas y se han guardado en otro set llamado precios limpios

#### Set de datos de sucursales

Las variables aptas para ser codificadas son region tipoSucursal, región y provincia.

#### Set de datos de productos

Ninguna variable era apta para su codificación.

Si bien se realizaron las codificaciones, se tomó la decisión de que las variables queden en su estado original, ya que en pasos posteriores servirán para la preparación de nuestro set de datos para modelarlos a través de la creación de variables dummys.

### 2.4. No cambiar los nombres de las variables de la fuente de origen

### 2.5. Verificar la consistencia de las variables

Algunas de las inconsistencias encontradas:

Inconsistencia 1:

#### Set de datos de productos

-Hay productos id distintos para un misma misma marca, nombre y presentación

Para resolver este problema se creó una nueva variable llamada “id\_referencia” que adopta el mismo valor que el “id” cuando este es único (por las variables nombre marca y presentación) y será el mínimo valor de “id” si el producto está repetido.

#### Set de datos de precios

-Hay reportes de distintos precios para un mismo producto, sucursal y fecha.

¿Que hacer con estos valores?

Solo quedarán en el dataset un precio para cada producto, sucursal fecha. El precio eliminado es el que esté más lejos del valor promedio calculado por producto y fecha.

Inconsistencia 2:

Set de datos de productos

-Existen productos que tienen en su nombre una cantidad distinta a la expresada en la variable presentación.

En la segunda parte de nuestro trabajo se trabajara para salvar esta inconsistencia.

### 2.6. Identificar y documentar valores atípicos/outliers

Set de datos de productos

-Para identificar un producto atípico se definieron varios criterios de prueba:

1. SI su valor reportado es el precio de la media mas menos 3 desvíos estándares

En este caso se encontraron 16.000 productos reportados con la clasificación de extremo, ya sea inferior o superior.

normal 2214795

extremo inferior 12952

Extremo superior 3177

1. Si un precio supera 3 veces la media de ese producto teniendo en cuenta la fecha, se obtienen:

normal 2214775

extremo 20

Si bien quedaron pocos casos considerados como outliers se cree que es una decisión excesiva y quizás los outliers queden subestimados.

1. Se calculó la media, el cuartil 1 el cuartil 3 y el rango intercuartílico para identificar los valores extremos para cada tipo de producto. Se considera un outliers a cada uno de los precios que era menor al q1 -3 veces el rango intercuartilico o a los precios mayores que el q3 +3 veces el rango intercuartilico.

Mediante este método se encontraron:

normal 2174487

extremo inferior 29704

extremo superior 10604

Al analizar algunos casos particulares, no consideramos que esos valores extremos eran precios excesivos que generaban problemas en la distribución, y que era posible que ese precio observado sea mayor o menor de acuerdo a la zona geográfica en la que era reportado.

Para salvar dicho problema, se realizó el mismo cálculo, pero teniendo en cuenta las medidas media, cuartil 1, cuartil 3 y rango intercuartílico para cada producto en particular teniendo en cuenta no sólo la fecha que se reportaba sino también la región en la que se encuentra dicha

Sucursal.

De esta manera se han reportado:

normal 2147307

extremo inferior 31110 (0.01)

extremo superior 18380 (0.008)

Estos precios serán eliminados del análisis ya que se ha considerado todas las variables relevantes para su identificación.

¿Que podemos decir de estos valores? Set de datos adicional (Precios\_extremos)

-49490 registros son considerados extremos( ya sea inferiores o superiores)

-El menor precio reportado es 1.22, corresponde al id=’7622300871918’ que representa al Polvo para Preparar Postre Caramelo en Sobre Royal 75 Gr. Este producto promedio, para la región centro y la fecha 20200419, es de 41.45, lo cual se evidencia que hay un problema de carga.

-El precio maximo reportado es de 16995, para el producto id=’7792389079939’ que representa a los Espirales X5 4 Un. Dicho precio es reportado para las fechas 20200412 y 20200419 las dos en la zona norte grande.

¿Cuales son los productos que registran precios extremos?

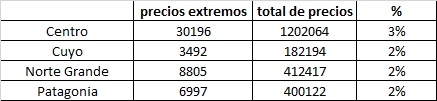
Hay un total de 456 de sucursales que reportan al menos un precio extremo. En comparación con el total de sucursales que han reportado en las fechas analizadas (488 aproximadamente), la mayoría de las mismas tienen algún precio extremo reportado.

Algunas de estas sucursales son el Super Mami, ubicado en Salsipuedes, el hipermercado Carrefour en Tierra del Fuego, entre otros.

Las principales fechas con valores extremos son las de mayo del 2020.



Los precios extremos están distribuidos en las distintas regiones de manera similar en término de porcentaje.



Aunque en valores absolutos la zona centro reporte la mayor cantidad de valores extremos (30196).

### 2.7. Evaluar cómo comprimir los datos para su almacenamiento más eficiente

Sin cambios

### 2.8. Guardar el set de datos con un nombre informativo.

Sin cambios

## Pasos deseables

### 3.1. Ordenar variables/columnas si es posible – primero ID, luego en el mismo orden que la fuente

Los archivos se encontraban ordenados

### 3.2. Quitar variables/columnas que no tienen información a analizar

Set de datos de productos

Se eliminaron las variables: categoria1, categoria2 y 'categoria3.

Set de datos de precios

No se elimino ninguna variable

Set de datos de sucursales

Se eliminaron las varialbes: comercioId, banderaId, comercioRazonSocial, provincia, localidad y direccion.

### 3.3. Renombrar variables de grillas

No hizo falta renombrar las variables

### 3.4. Categorizar resultados en “Otros”

No hay variables con categoria otros.

### 3.5. Agregar metadata a los datos: cuando y como fueron obtenidos, limpieza realizada

asunciones, etc

Vincular con etiquetas del código fuente y los datos. Al menos incluir un README

# Análisis en Profundidad del Contenido

## Todas las variables tienen el tipo apropiado? Analizar las features con tipo Objeto. Qué decisiones tomarán al respecto?

Todas las variables tienen un tipo apropiado, solo la variable fecha del dataset de precios se transformo en una variable tipo datetime.

## Todos los producto\_id en el dataset de precios se encuentran en la entidad de productos? Tomamos decisiones al respecto?

Existen 7 productos reportados por alguna sucursal que no se encuentran en el fichero original de productos. Frente a estos casos, al no poder distinguir que tipo de producto es, ni cual es su marca y presentación se ha decidido eliminar estos casos.

## Todos los sucursal id en el dataset de precios se encuentran en la entidad de sucursales? Tomamos decisiones al respecto?

No todos los las sucursales id que se encuentran en el dataset de precios se unen con el finchero original de sucursales. Estos casos ( 17.998 registros) al no tener informacion de cual es la sucursal que lo reporto en cuento a nombre, direccion, region, entre otros. Se decidio eliminarlos del analisis. Esta eliminacion se realiza en un paso previo de analizar los outliers ya que fue necesario unir los dataset para poder calcular las medidas descriptivas por producto, fecha y región.

## Cuántos precios reporta cada sucursal? Todas las sucursales reportan precios en todas las fechas?

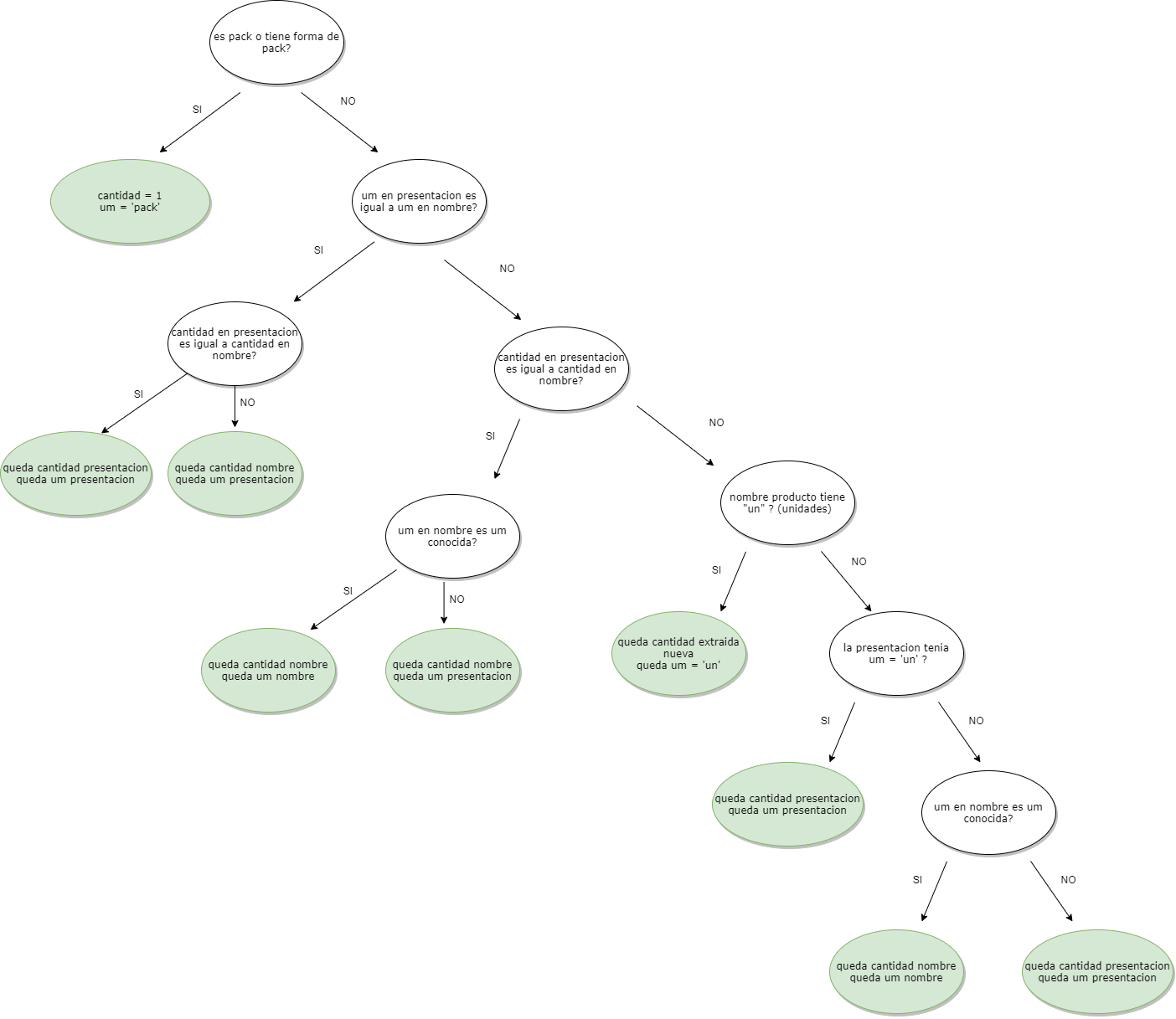
No todas las sucursales reportan precios, solo el 20% del total de sucursales que se encuentra en el dataset de sucursales reportan precios en estas fechas.

## En el dataset de productos:

### 5.1. presentación: a partir de esta característica, crear nuevas columnas que separan cantidad y un. Homogeneizar unidades de medida, al máximo posible, y crear nueva columna para indicar um homogénea y el factor que sería necesario para homogeneizar el precio de esa presentación. A partir de um homogénea, crear variables dummies (Tip: Pueden usar pd.get\_dummies()).

En el dataset de productos se encuentran distintas variables que representan la unidad y la cantidad del producto, en un primer plano, las unidades y cantidades que aparecen en el nombre del producto, en un segundo plano una variable de presentacion donde se encuentra la cantidad y la unidad para ese producto, y, por ultimo, en tercer lugar la presentacion depurada brindada por las mentoras.

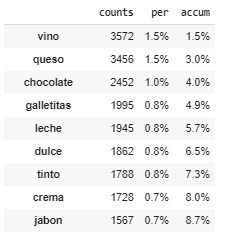
Al existir diferencias entre las tres opciones se decidió realizar una depuración detallada partiendo principalmente de las unidades y las medidas en el nombre comparandolas con lo encontrado. A continuación se presenta el arbol de decision que se tiene en cuenta para quedarnos por producto, con la unidad y medida limpias.



### 5.2. nombre: crear una nueva columna con el nombre\_depurado. Esto es, llevar todo a minúsculas, quitar signos de puntuación, quitar unidad de medida, identificar dígitos numéricos. Utilizando esta columna, nombre\_depurado, tokenizar los nombres y obtener las palabras más fecuentes. Eliminar palabras muy frecuentes en nombres que no nos dicen nada (‘de’, ‘en’, ‘con’, ‘para’, ‘la’, ‘el’, ‘&’, etc.). Convertir las n palabras más frecuentes en dummies. n es una decisión que deberán adoptar (Tip: Pueden usar Series.str.find()). Evaluar la posibilidad de crear una dummy otras palabras, para aquellos nombres que contienen palabras poco frecuentes no incluidas.

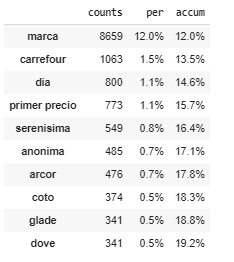
Luego de realizar la limpieza mencionada en la consigna, se crearon dummys para los 10 nombres principales encontrados en todo el dataset de precios reportados.

Si bien estas 10 palabras solo representan el 9,2% de los registros encontrados en nuestro dataset, se decidio partir por esa cantidad como una primera aproximacion. Estos valores pueden ser ajustados posteriormente.



### 5.3. marca: obtener las n marcas más frecuentes y crear un nuevo campo 'marca\_frecuente' que tenga la marca original si es de las más frecuentes y la categoría OTRA para aquellas menos frecuentes. Crear dummies a partir de este campo. (Una opción viable podría ser unir el nombre con la marca en un campo y luego depurar las palabras duplicadas que queden en el nombre+marca. Cualquier otra técnica que se les ocurra puede ser válida)

Se realizo el mismo tratamiento mencionado anteriormente, para el nombre, pero con el campo marca. Una vez que el campo quedo depurado, se realizaron dummys de las principales palabras obtenidas. En este caso se han tomado las 10 marcas principales que acumulan un 20% de lo reportado.



## En el dataset de sucursales:

### 6.1. provincia o nom\_provincia: a partir de esta característica, crear las variables dummies correspondientes.

Despues de la unificacion de nombres en el caso de provincias con mas de una palabra se creo una dummy para cada una de las provincias que aparecen en el dataset de sucursales y precios de manera conjunta.

### 6.2. tipoSucursal: a partir de esta característica, crear las variables dummies correspondientes.

Se realizo las dummys correspondientes por tipoSucursal.

### 6.3. banderaDescripcion: a partir de esta característica, crear las variables dummies correspondientes. La misma descripción de sucursal tiene siempre el mismo tipo? En qué afecta esto?

Como el tipo de sucursal no se coincide en todos los casos con la bandera descripción, fue necesario crear dummys para esta variable. Existen casos en la que un tipo de sucursal tiene dos banderaDescripción distintas, y tambien existen casos en la que una bandera de descripcion puede coincidir con mas de un tipo de sucursal.

Es por ello, que se decidió que la banderaDescripción es una variable adicional a tener en cuenta para incorporar al modelo.

## Unir los precios con las entidades de productos y sucursales para obtener un dataset unificado y limpio. A partir del precio y del factor, crear el precio\_homogeneo. Volver a analizar outliers.

Se creo la precio homogeneo de cada uno de los productos para poder analizar los mismos en sus unidades centrales. Es decir que todas las unidades analizadas serán un kilo, un litro, una unidad, entre otras, y se recalculara su precio en consecuencia.

## Crear el precio\_relativo, de acuerdo al producto seleccionado y la metodología implementada en el práctico de A&V.

Se calculo los precios relativos de cada producto, y con este precio se calculo la categoria de outliers.

Los outliers fueron seleccionados con el metodo explicado anteriormente, teniendo en cuenta la distribucion de cada uno de los prodcutos, por fecha y region.

De esta forma han quedado aproximadamente 8 mil outliers nuevos que se han excluido del analisis.