Características del Dataste -DS- de MLB.

* Se recibió un Dataset con 500mil envíos de productos vendidos en Mercado Libre Brasil

Datos del Producto:

-Peso

-Largo

-Ancho

-Alto

-Precio

-Marca

-Modelo

-Nuevo o Usado

380mil

Publicaciones

3mil

Categorías

61mil Vendedores

* Se detectaron muchos valores faltantes. El DS se reduce considerablemente.

*Envíos con Dimensiones válidas*

260mil

Envíos

500mil

envíos

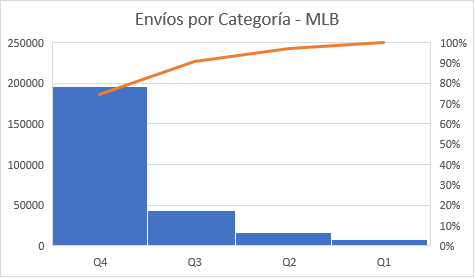
*Envíos con datos de Dimensiones y Categorías*

69%

52%

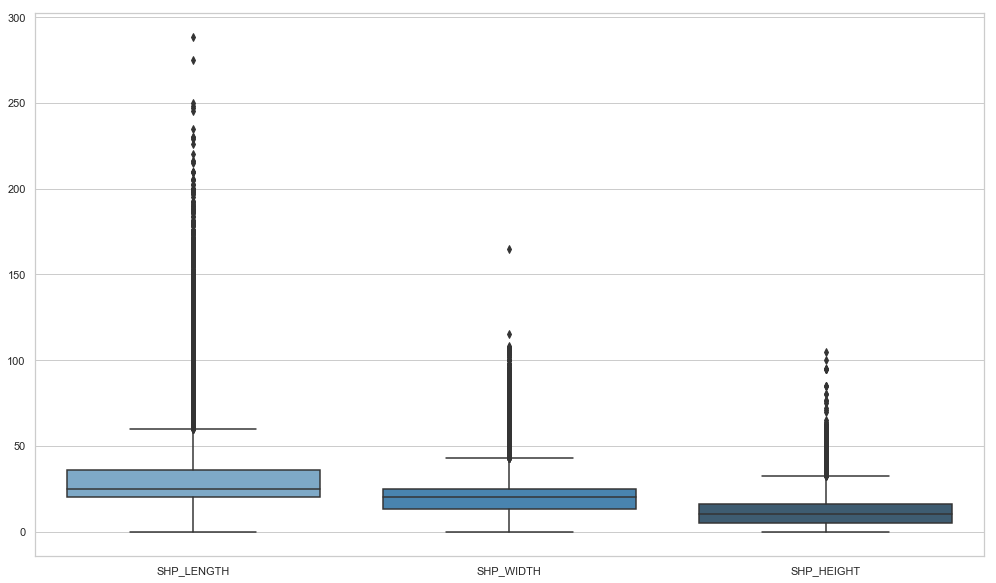
*Envíos con datos incompletos*

* Revisando Categorías, el 25% de las mismas genera el 75% de los envíos.

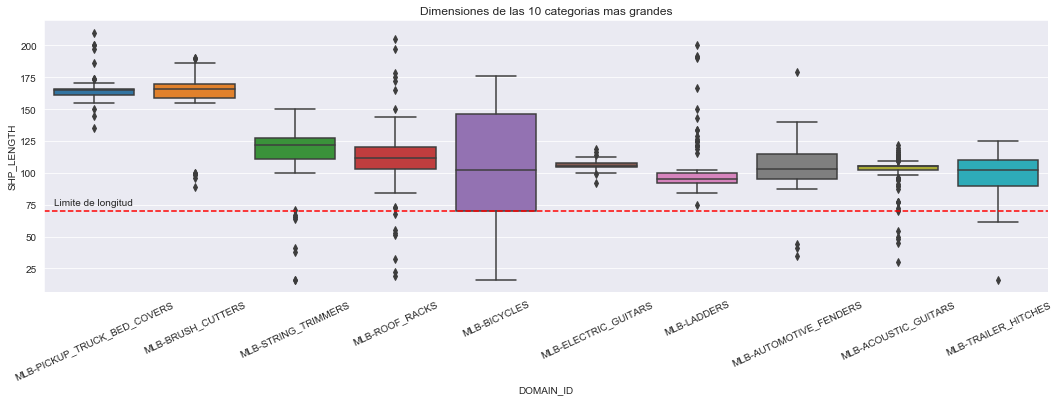


Comentarios sobre el análisis realizado

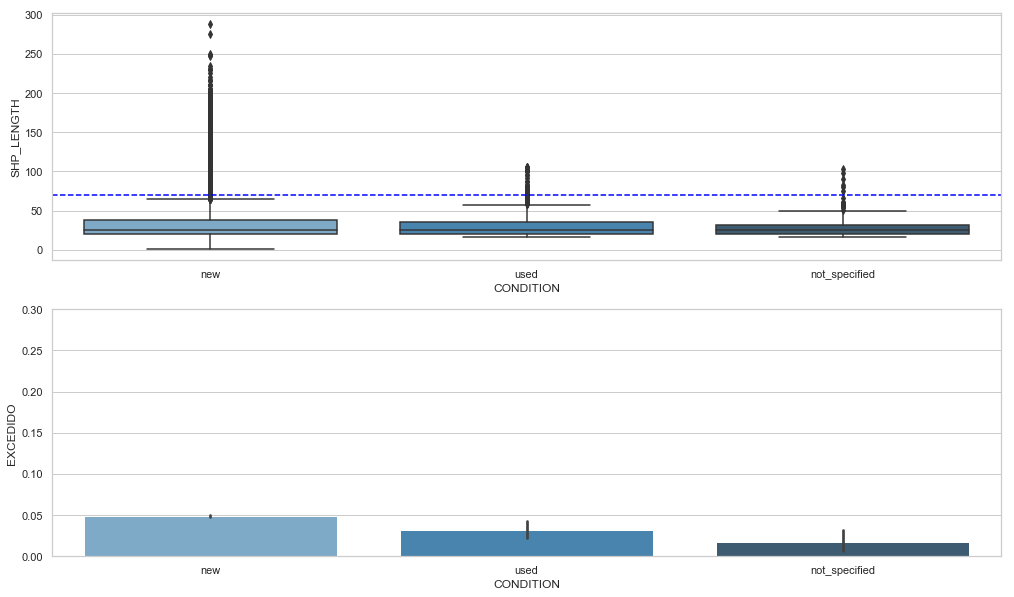
* Enfoque en datos faltantes:
  + Se verificó que los datos de Dimensiones y Pesos están “en equipo”, es decir se cuenta con todos los datos para el envío, o faltaban todos. Lo mismo ocurre con los “equipos” Condición y Precio, y Título y Atributos.
  + Se detectó que una misma publicación informa valores en Dimensiones para algunos envíos y para otros no. Lo mismo ocurre en otras variables.
  + En base a esta situación se trabajó para completar valores faltantes en algunos envíos tomando de base los datos mostrados en otros de la misma publicación. Se procedió con los datos de Dimensiones y Pesos, que fueron completados con la mediana, logrando recuperar el 0,3% de envíos. El mismo criterio se podría haber aplicado para las otras variables.
  + Para el resto del análisis a nivel general se descartaron los envíos con más de siete valores nulos.
* Análisis de Outliers:
  + Se verificó que los valores extremos para Dimensiones y Peso son coherentes con el producto que se envía. Por lo tanto, se mantienen dentro del dataset.
  + Para la variable Precio se encontró un valor extremadamente alto que se consideró basura.
* Atributos del Producto:
  + Los atributos están expresados como una lista de diferentes conceptos. Varía la cantidad de un producto a otro, habiendo algunos que se repiten en el resto de las variables como por ejemplo el Vendedor.
  + De estos conceptos se rescataron los datos de Marca y Modelo, que son comunes a todos.
* Situación:
  + El Correo multa a envíos con alguna dimensión mayor a 70cm
  + La variable LARGO de cada envío muestra siempre la mayor longitud.
  + En adelante el análisis se centrará en ese valor, evaluando si supera o no el tope definido por el correo.
  + Debajo se muestra como se distribuyen las Dimensiones.



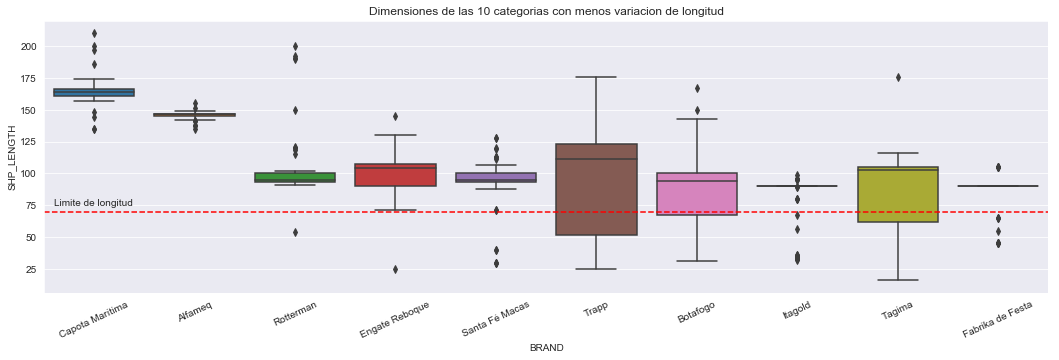
* Categoría del Producto:
  + Se analizaron las Categorías con mayor LARGO promedio verificando que, salvo ouliers, superan el tope impuesto por el correo. Esto muestra inicialmente una relación entre la Categoría del producto y la posibilidad de que ese envío sea multado.
  + Hay casos particulares, como la Categoría Bicicletas, que muestran envíos con un valor en LARGO que superan el tope y otros no.
  + Estos, en general, corresponden a Categorías con mayor Desviación respecto a la media.



* Condición de Venta del Producto:
  + Revisando la condición de venta del Producto, aquellos nuevos muestran una mayor casuística de envíos que superan el tope que el resto.
  + Esta correlación se validó con CHI2, obteniendo una significancia de 0.00132, menor al 0.05% de tolerancia.



* Marca y Modelo del Producto:
  + Se analizaron aquellas Marcas con mayor LARGO promedio observando que solo las primeras superan siempre el tope de multa, luego comienzan a mostrarse casos compartidos.
  + Para avanzar en la búsqueda de una correlación entre estas variables, recomendamos sumar el Modelo.



* Curiosidades del Título de la Publicación:
  + Se realizó un escaneo de las palabras usadas en el Título, quitando conectores. Luego se procedió a revisar las palabras que más se repiten, encontrando una correlación entre estas palabras y la posibilidad de superar el tope máximo. Revisando esas palabras, nombrarían productos con posibilidad de tener grandes dimensiones.

portas -> puerta

teto -> techo

lateral -> lado

mdf -> MDF

aluminio -> aluminio

mesa -> mesa

madeira -> madera

parede -> pared