Estimación de peso y dimensiones de los envíos de Mercado Libre

**Materia: Análisis y Curación de datos**

**Análisis del dataset.**

**Comunicación de resultados y conclusiones**

A partir de lo visto en la teoría de la materia y del segundo laboratorio, diagramar una comunicación en formato textual o interactivo describiendo la solución de las actividades propuestas a continuación. Al final de las mismas se proveen actividades opcionales (no obligatorias) que pueden resultar de interés.

**Actividades Propuestas:**

1. Eliminar valores cuyo status sea `404` , luego eliminar la columna `status` del dataset ya que solo es útil para limpieza.
2. Eliminar los valores NaN de las columnas con prefijo `SHP\_`. Estas son aquellas que representan o peso o dimensiones de un item.
3. Agrupar por item id y calcular mediana de peso y medidas. De esta forma debería quedar una única fila por cada item\_id.
4. Parsear la columna de atributos y extraer a columnas propias aquellos atributos cuyo `id` sea `BRAND` o `MODEL`. Estos atributos representan marca o modelo que el vendedor del item ingresó en la publicación. [Opcional] No es necesario limitarse a estos dos atributos, se puede probar quedarse con los N atributos más frecuentes.
5. Transformar variables categóricas en números (Se recomienda [OneHotEncoding](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html)) para las columnas (Sugerencia: arrancar con un sample de ~10K items)
   1. `CATALOG\_PRODUCT\_ID`
   2. `CONDITION`
   3. `DOMAIN\_ID`
   4. `SELLER\_ID`
   5. `BRAND` (extraída en 4)
   6. `MODEL`(extraída en 4)
6. En caso de tener alguna variable no medida (en nuestro caso `PRICE`) imputar sus valores utilizando [kNN](https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html).
7. Medir las distribuciones de las variables como histogramas, realizar normalizaciones e identificar outliers con los métodos vistos en clase. Hacer análisis de estos outliers y considerar si sería correcto o no eliminarlos del dataset. Sugerencia: Identificar outliers de las columnas `SHP\_WEIGHT` y `SHP\_VOLUME`, donde `SHP\_VOLUME` se define como el producto de las dimensiones.
8. [**Opcional**] Si están interesados en usar texto y técnicas de NLP sobre el título:
   1. Realizar preprocesamiento del texto, se recomienda leer [este artículo](https://medium.com/@datamonsters/text-preprocessing-in-python-steps-tools-and-examples-bf025f872908).
   2. Para convertir el título en vectores se recomienda ver tf-idf. [wiki](https://es.wikipedia.org/wiki/Tf-idf), [sklearn](https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html).
9. [**Opcional**] Aplicar PCA y decidir si existe un subespacio del espacio de features donde viven los datos y son bien descriptos. Graficar.
10. [**Opcional**] Aplicar Mixturas Gaussianas ([GMM](https://jakevdp.github.io/PythonDataScienceHandbook/05.12-gaussian-mixtures.html)) para hacer clustering.

La comunicación debe estar apuntada a un público técnico pero sin conocimiento del tema particular, como por ejemplo, sus compañeros de clase o stakeholders del proyecto. Idealmente, además del documento se debería generar una presentación corta para stakeholders explicando el análisis realizado sobre los datos y las conclusiones obtenidas de tal análisis.

Se evaluarán los siguientes aspectos:

* El informe debe contener un mensaje claro y presentado de forma concisa.
* Los gráficos deben aplicar los conceptos de percepción visual vistos en clase.
* Se debe describir o estimar la significancia estadística de su trabajo.

## Entrega

La entrega debe realizarse antes del **09-09**. Definan un repositorio en el cual podamos ver el notebook con el que hayan procesado los datos y manden el link al mismo. Pueden utilizar cualquier herramienta para generar el informe y la presentación (sugerimos Google Drive) y pasen el link a dónde podamos verlos.

## Presentación

A charlar. Estaría muy bueno tomarnos entre 30 y 60 minutos, aunque sea por videollamada, para que puedan exponer los resultados y charlamos sobre ellos. El objetivo de esto es hacer más fluída la devolución sobre el laboratorio.