TELEDUC
Educación a Distancia

Herramientas avanzadas de programación en *Python* para procesamiento de datos

Actividad Final: Base de datos de clientes

En esta actividad, recibirás datos de las boletas de la papelería “DM”. Debes ser capaz de procesar estos datos, e integrarlos a los datos de los productos de esta empresa para poder extraer información valiosa.

Puedes ocupar todos los contenidos que se han visto en el curso. Como recomendación, revisa los videos de la actividad práctica y los códigos que están subidos.

Pasos para resolver la actividad final

Considera la siguiente secuencia de tareas:

1. Debes descargar el archivo “detalle_boletas.csv” en la plataforma y cargarlo en Python. Para esto, crea un Data Frame de nombre “detalle_boletas” con los datos de este archivo.

   Este archivo contiene la información de diversas boletas emitidas por DM desde el 2016. En particular, cada fila contiene la siguiente información:

   a) Fecha: Fecha de emisión de la boleta, en formato “año/mes/día”.
   b) ID: Identificador de uno de los productos vendidos en esa boleta.
   c) NXXX: Identificador de cada boleta.
   d) Cantidad: Cantidad del producto vendido.
   e) Precio_prod: Corresponde al precio de cada producto.

2. Antes de comenzar el análisis, es necesario modificar la base de datos “detalle_boletas”. En particular, hay que:

   a) Eliminar la columna Precio_prod, ya que estos precios están incorrectos.
   b) Crear una columna “Pais_Venta”, dado que DM quiere internacionalizarse en el futuro y contar con esta información será clave. Los valores de esta columna son todos “Chile”, ya que por ahora solo hay ventas en este país.
   c) Cambiar el nombre de la columna “NXXX” por “Num Boleta”.

Pueden notar que, por la disposición de datos de este archivo, cuando una boleta tiene más de un producto entonces puede haber varias filas de esa boleta. Por ejemplo:

Fecha	ID	Num Boleta	Cantidad	País
2016/01/01	400009	554170000002	3	Chile
2016/01/01	400007	554170000002	2	Chile
2016/01/02	400005	554170000003	2	Chile
2016/01/02	400005	554170000004	2	Chile
2016/01/02	400001	554170000004	2	Chile
2016/01/02	400002	554170000004	2	Chile

En este caso:

- La boleta 554170000002 tiene dos productos, de ID 400009 y 400007. Se vendieron 3 y 2 unidades de cada producto respectivamente. Fue emitida el 01 de enero de 2016.
- La boleta 554170000003 tiene un producto, de ID 400005. Se vendieron 2 unidades de este producto. Fue emitida el 02 de enero de 2016.
- La boleta 554170000004 tiene tres productos, de ID 400005, 400001 y 400002. Se vendieron 2 cada producto. Fue emitida el 02 de enero de 2016.

Algunas consideraciones importantes:

- En un día se pueden emitir una o más boletas.
- Una boleta nunca tendrá dos fechas distintas de emisión.
- Nunca habrá dos filas de la misma boleta con el mismo producto. Por ejemplo (esto nunca pasará en la base de datos):

Fecha	ID	Num Boleta	Cantidad	País
2016/01/01	400009	554170000002	3	Chile
2016/01/01	400009	554170000002	2	Chile

3. El archivo CSV “detalle_boletas.csv” anterior venía sucio y hay que limpiarlo.
   a) Hay productos que tienen ID “4XXXXX” y Num Boletas “55417XXXXXXX”. Debes eliminar cualquier fila del Data Frame detalle_boletas que contenga **alguno de los dos** (es decir, que el identificador ID del producto sea “4XXXXX” o bien el número de boleta Num Boleta sea “55417XXXXXXX”), ya que son datos que se generaron de forma errónea por el sistema y no deben ser considerados.
   b) La columna Fecha tiene caracteres extra. Debes limpiarla de tal manera que el formato sea el original, es decir, “año/mes/día” (sin ningún otro carácter extra). En particular, identifica qué caracteres extra hay aparte de los "/" y números y eliminalos.

4. A continuación, calcule estadísticos descriptivos de la columna Cantidad para cada uno de los productos que existen e imprímalos en la consola. Los estadísticos descriptivos que debes calcular son: media, desviación estándar, mínimo y máximo. Tu resultado debería verse de la siguiente manera:

ID	amax	amin	mean	std
400001	5.0	1.0	3.017323	1.392588
400002	5.0	1.0	3.020270	1.386025
400003	5.0	1.0	3.008947	1.406157
400004	5.0	1.0	3.003133	1.398234
400005	5.0	1.0	3.002101	1.413469
400006	5.0	1.0	2.963720	1.384331
400007	5.0	1.0	2.957627	1.415452
400008	5.0	1.0	3.061611	1.435621
400009	5.0	1.0	3.005735	1.424676
400010	5.0	1.0	3.017857	1.397655

5. Ahora que la información del Data Frame detalle_boletas está limpia, debes generar una columna Anho (que contenga el año de la columna Fecha), una columna Mes (que contenga el mes de la columna Fecha), y una columna Dia (que contenga el día de la columna Fecha). Estas columnas debes agregarlas al Data Frame detalle_boletas. Luego, elimina la columna Fecha para que no haya información redundante.
6. Descarga el archivo “Lista productos.csv” de la plataforma y cárgalo a Python. Para esto, crea un Data Frame de nombre “lista_productos” con los datos de este archivo.

Este archivo contiene el detalle de los 10 productos que tiene el inventario de la empresa DM. Específicamente, es la siguiente información:

ID	Nombre	Descrip	Precio Unitario
400001	Alerce A4C	Resma A4 Carta 500 hojas	2250
400002	Alerce A4O	Resma A4 Oficio 500 hojas	2500
400003	Alerce A4C XL	Resma A4 Carta 1000 hojas	4200
400004	Alerce A4O XL	Resma A4 Oficio 1000 hojas	4700
400005	Alerce Kraft	Pliego papel kraft 90x60cm	500
400006	Alerce Kraft XL	Pliego papel kraft 150x90cm	750
400007	Alerce PreCuad Oficio	Block prepicado, tamaño oficio, cuadriculado, 80 hojas	1300
400008	Alerce PreCuad Carta	Block prepicado, tamaño carta, cuadriculado, 80 hojas	1100
400009	Alerce PreComp Oficio	Block prepicado, tamaño oficio, composición, 80 hojas	1200
400010	Alerce PreComp Carta	Block prepicado, tamaño carta, composición, 80 hojas	1000

Donde:

- ID: Corresponde al identificador de cada producto.
- Nombre: Corresponde al nombre de cada producto.
- Descrip: Corresponde a la descripción de cada producto.
- Precio Unitario: Corresponde al precio de cada unidad de ese producto.

7. Une el Data Frame lista_productos con el Data Frame detalle_boletas, en base a la información de la columna ID. El Data Frame resultante de esta unión debe contener la misma información que el Data Frame detalle_boletas, pero ahora cada fila además debe tener el nombre del producto, la descripción, y el precio unitario. El Data Frame debes llamarlo detalle_boletas2. Imprime este Data Frame en la consola. Pon atención con el tipo de datos, porque para hacer esta unión, la columna donde se busquen los valores en común en ambos Data Frames **deben tener el mismo tipo**.

8. Calcula cuántos ingresos significaron la venta de los productos por boleta. Para eso, agrega una nueva columna de nombre “Ingreso total” al Data Frame detalle_boletas2. Esta columna debe tener como valores la multiplicación entre la columna “Precio Unitario” y “Cantidad”. Imprime el Data Frame detalle_boletas2 con esta nueva columna en la consola.

9. Finalmente, calcule estadísticos descriptivos de la columna “Ingreso total” para cada uno de los productos que existen. Los estadísticos descriptivos que debes calcular son: media, desviación estándar, mínimo y máximo. Tu resultado debería verse de la siguiente manera:

ID	amax	amin	mean	std	sum
400001	11250.0	2250.0	6788.976378	3133.322406	12933000.0
400002	12500.0	2500.0	7550.675676	3465.062916	14527500.0
400003	21000.0	4200.0	12637.578947	5905.857753	24011400.0
400004	23500.0	4700.0	14114.725849	6571.699430	27029700.0
400005	2500.0	500.0	1501.050420	706.734328	2858000.0
400006	3750.0	750.0	2222.789985	1038.248231	4350000.0
400007	6500.0	1300.0	3844.915254	1840.086976	7259200.0
400008	5500.0	1100.0	3367.772512	1579.182886	6395400.0
400009	6000.0	1200.0	3606.882169	1709.610924	6918000.0
400010	5000.0	1000.0	3017.857143	1397.654528	5746000.0

Nota: Los nombres de Data Frames y resultados de ejemplos entregados se muestran como guía para facilitar la realización del ejercicio. Si por alguna razón estiman conveniente cambiarlos, o mostrar los resultados en otro formato ***SIN CAMBIAR EL OBJETIVO INICIAL DE LO PEDIDO****, pueden hacerlo.*

Un ejemplo de lo anterior sería el cálculo de los estadísticos descriptivos, donde ustedes podrían calcular los estadísticos pedidos por ID de productos, pero no formatearlos en la tabla que se adjunta como ejemplo. No obstante, siempre deben calcular el mínimo, máximo, media y desviación estándar por ID de producto.

IMPORTANTE: Las imágenes incluidas en este documento son solamente referenciales, y no necesariamente representan exactamente lo que ustedes deben mostrar.