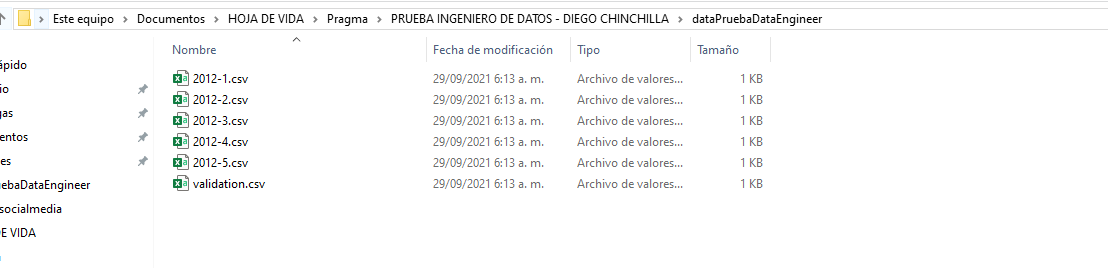
**Solución prueba pragma - Diego chinchilla- Pragma**

Requerimientos:

1. Descarga la carpeta comprimida que contiene los datos y déjalos en una carpeta exclusiva para este reto que vas a realizar.

**Respuesta:**



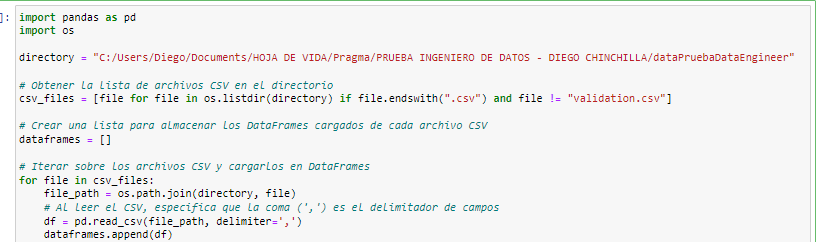
2. Construye un pipeline que sea capaz de:

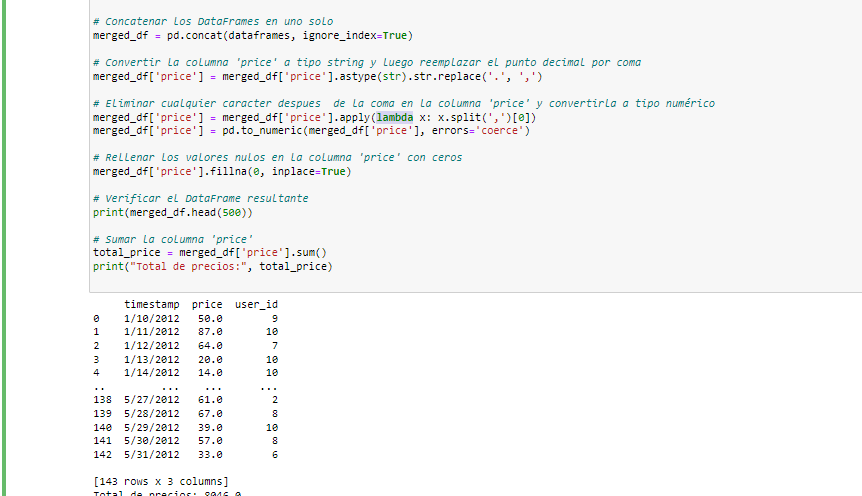
a. Cargar todos los archivos .CSV excepto el llamado “validation.csv” (El pipeline no debe contener todo el conjunto de datos, es decir, los 5 archivos .CSV al mismo tiempo en memoria en cualquier momento).

**Respuesta:**

Importación de las librerías pandas y os para el proceso de pipeline de carga de los archivos csv

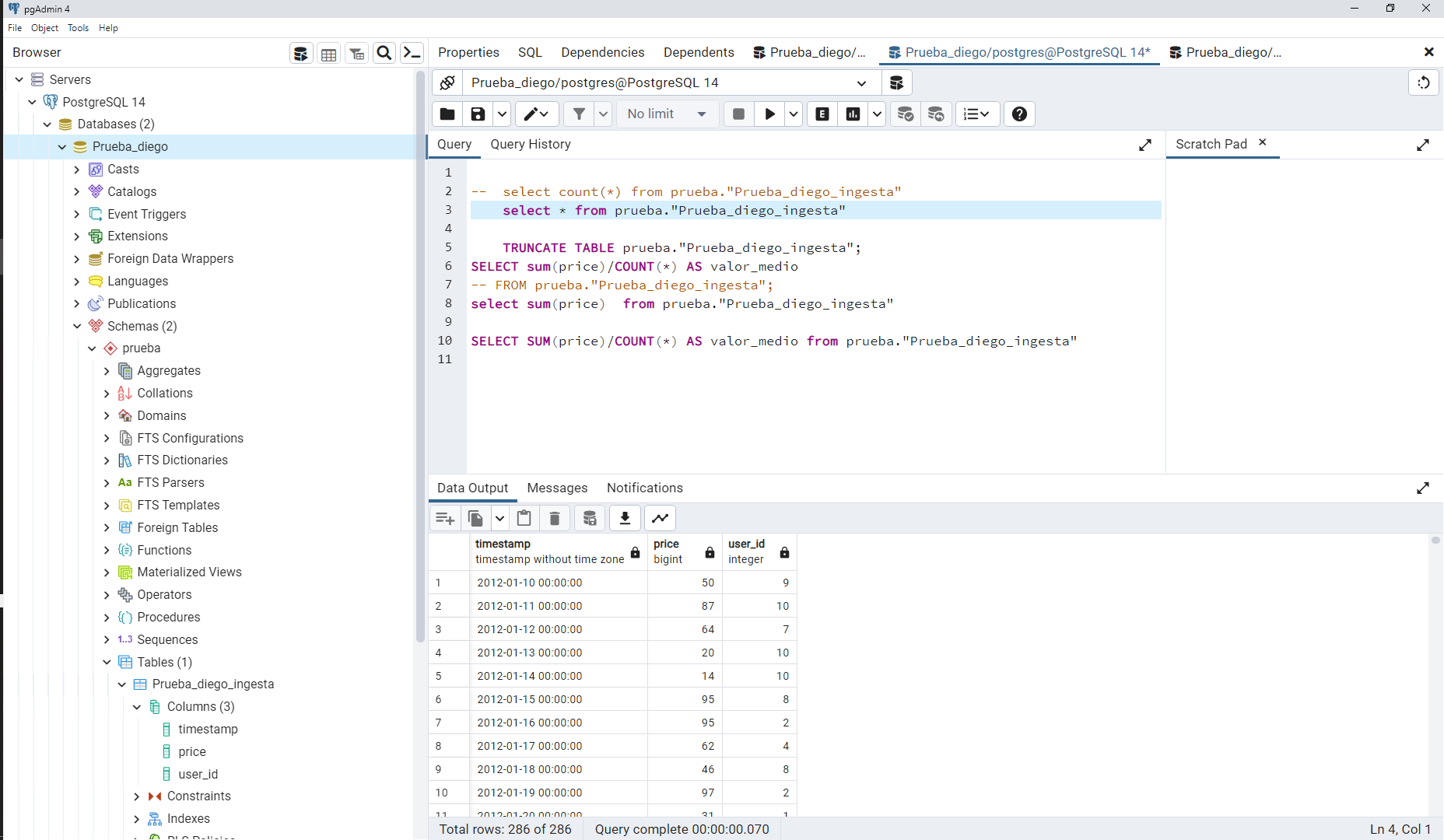
ubicados en la ruta C:\Users\Diego\Documents\HOJA DE VIDA\Pragma\PRUEBA INGENIERO DE DATOS - DIEGO CHINCHILLA\dataPruebaDataEngineer, de donde se lee el conjunto de archivos y se genera la excepción del archivo con el uso del operador de desigualdad !**=** donde nombramos el archivo a excluir , luego se procede a pasar la lista vacía para luego almacenar los dataframes usando el bucle que recorre cada archivo csv leyendo el contenido y agragandolo al df , una vez terminado el proceso se procede a concatenar los df para dejar solo uno seguido de la impresión de los resultados como test , al igual el uso de la función lambda para separar los valores después del .



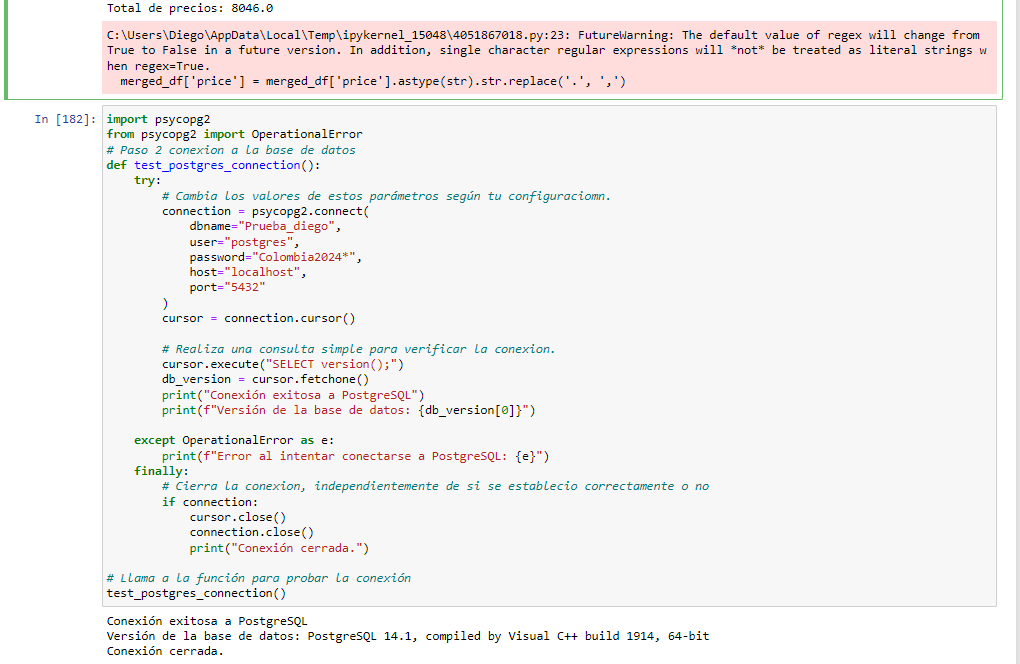


b. Almacena los datos de los archivos .CSV, en una base de datos de tu elección (ejemplo: PostgreSQL, MySQL, etc). El diseño de esta base dedatos dependerá de ti, crea la tabla o tablas que creas necesarias con el esquema que creas es adecuado, pero ten presente que todos los .CSV deben ir en la misma base de datos.

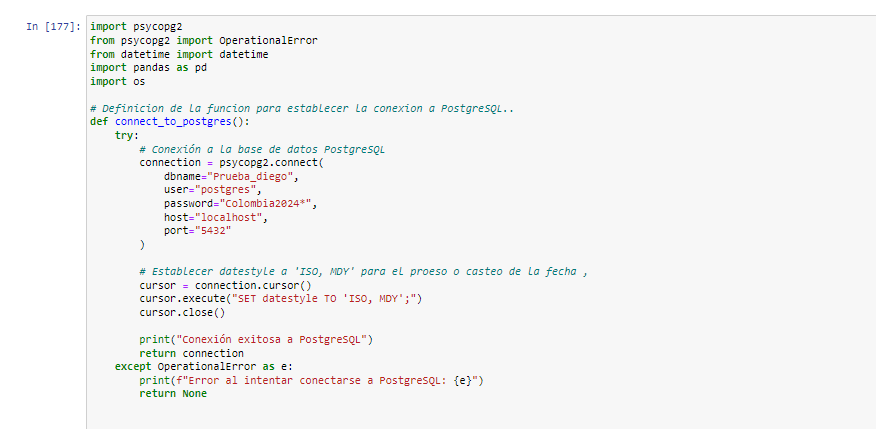
Use postgres, y el editor de pgAdmin4, donde se creó el la bd Prueba\_diego y el schema prueba para crear la tabla Pueba\_diego\_ingesta



**Respuesta:** Se efectúa proceso de conexión a la bd de postgres con los parámetros de establecidos.



Valido que este ok para dar continuidad al proceso e inicio el proceso de aplicación de máximo mínimos y valor medio o promedio solicitado



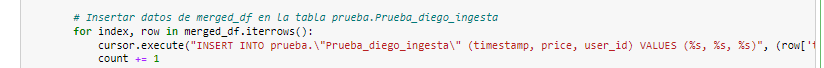


Como vemos, se genera el proceso de lógica donde tomo el df mergef\_df y ejecuto el proceso de sumatoria, conteo, máximos y minis y el promedio solicitado sin usar la función AVG que se indica. Al generar el proceso de salida de los datos, vemos las campos calculados para retornar la salida esperada.

c. A medida que los datos son cargados, realiza un seguimiento

de:

• Recuento (Count) de filas cargadas a la base de datos.



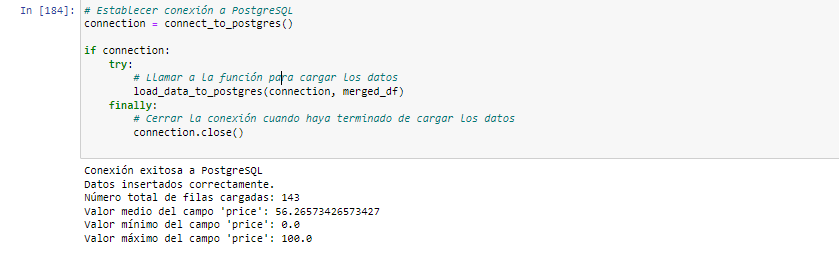
• Valor medio, mínimo y máximo del campo “price”.Se espera que en la ejecución de este pipeline al menos después de que se cargue cada .CSV (pero idealmente después de la inserción de cada



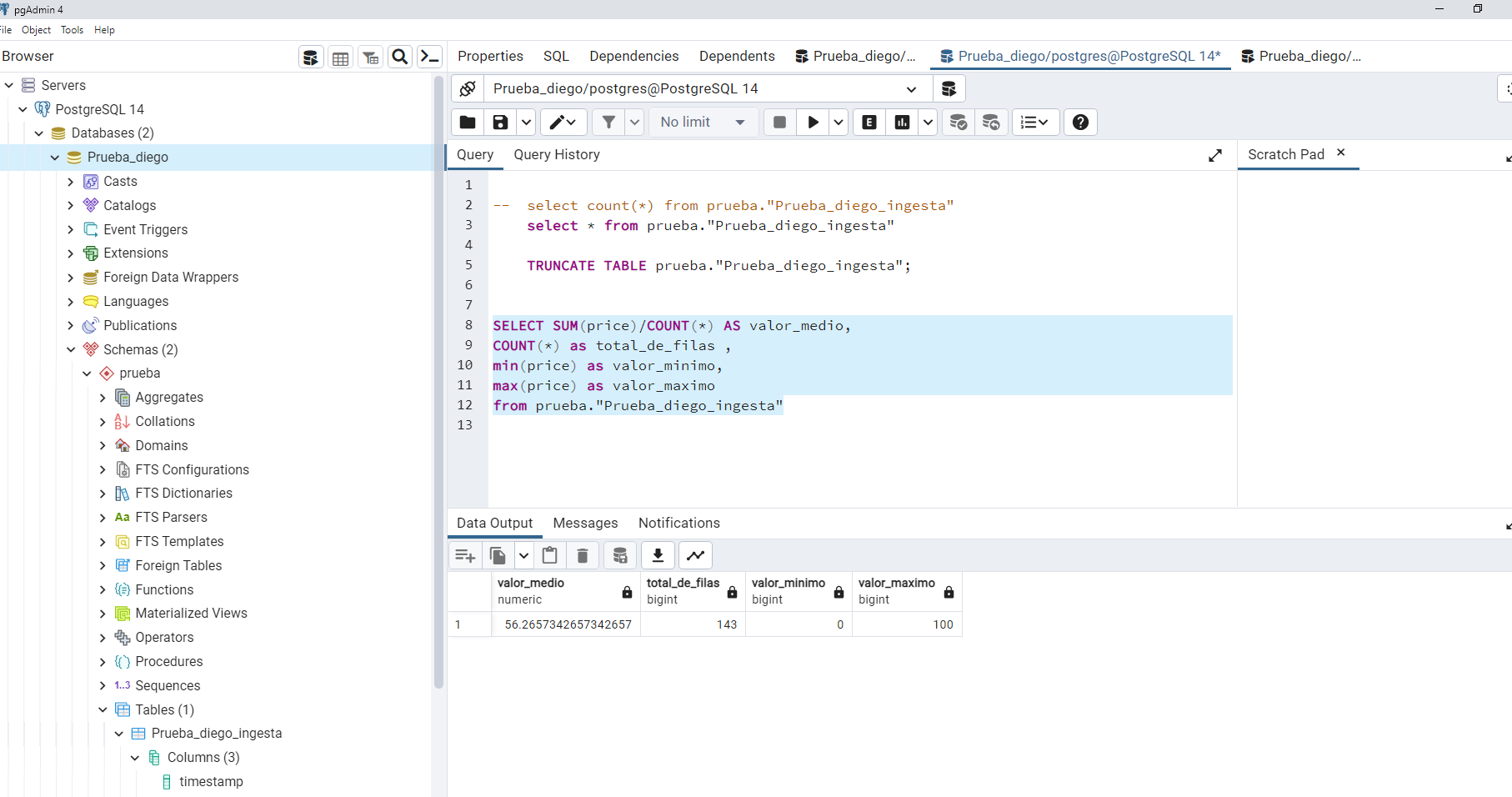
fila), estas estadísticas se actualicen para reflejar los nuevos datos. Las actualizaciones del pipeline NO deben tocar los datos ya cargados, es decir, hacer “SELECT avg (price) ...” para cada actualización no es una buena solución al problema.

3. Comprobación de resultados:

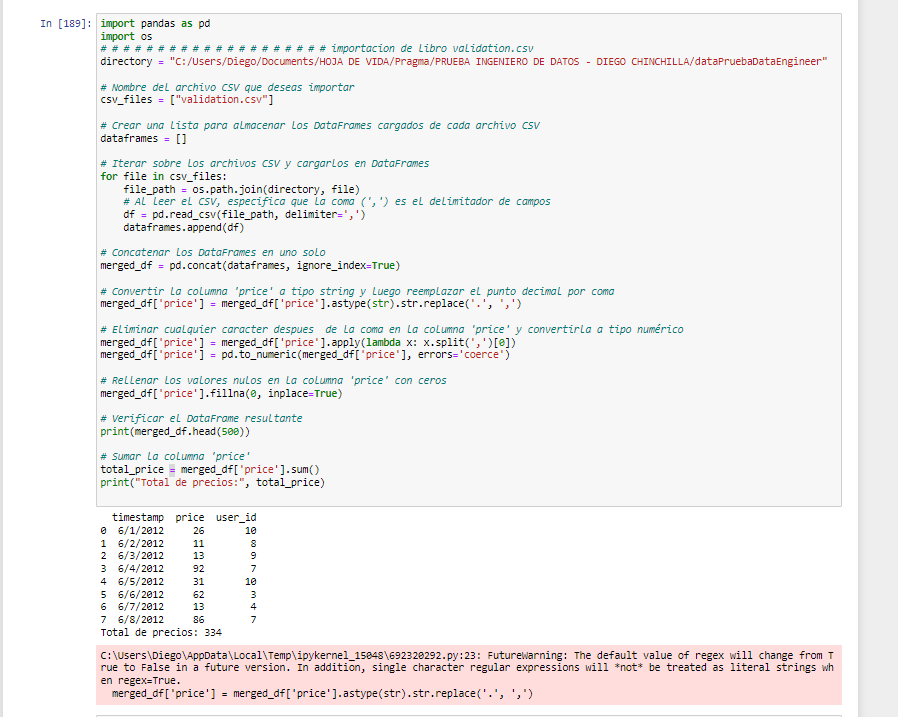
● Imprime el valor actual de las estadísticas en ejecución.



● Realiza una consulta en la base de datos del: recuento total de filas, valor promedio,

valor mínimo y valor máximo para el campo “price”. 

● Ejecuta el archivo “validation.csv” a través de todo el pipeline y muestra el valor de

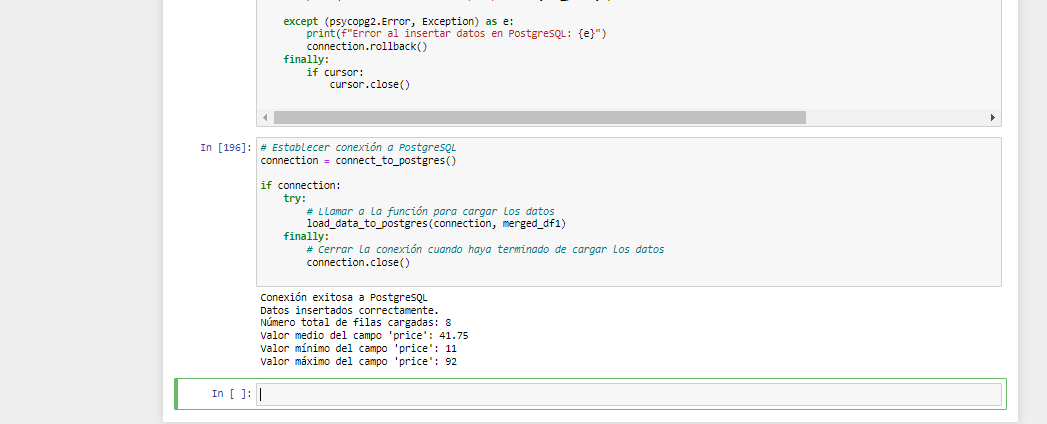
las estadísticas en ejecución. 

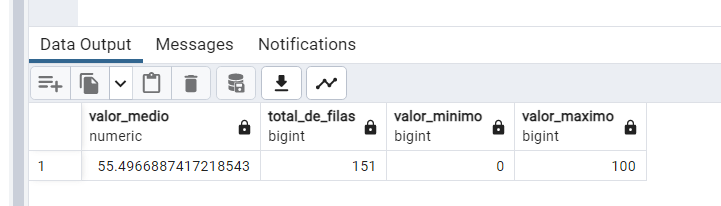
● Realice una nueva consulta en la base de datos después de cargar

“validation.csv”, para observar cómo cambiaron los valores del: recuento

total de filas, valor promedio, valor mínimo y valor máximo para el

campo “price”.





NOTA: se debe instalar la Liberia psycopg2 para el proceso de conexión a bd