

Avance 2 Evidencia 2

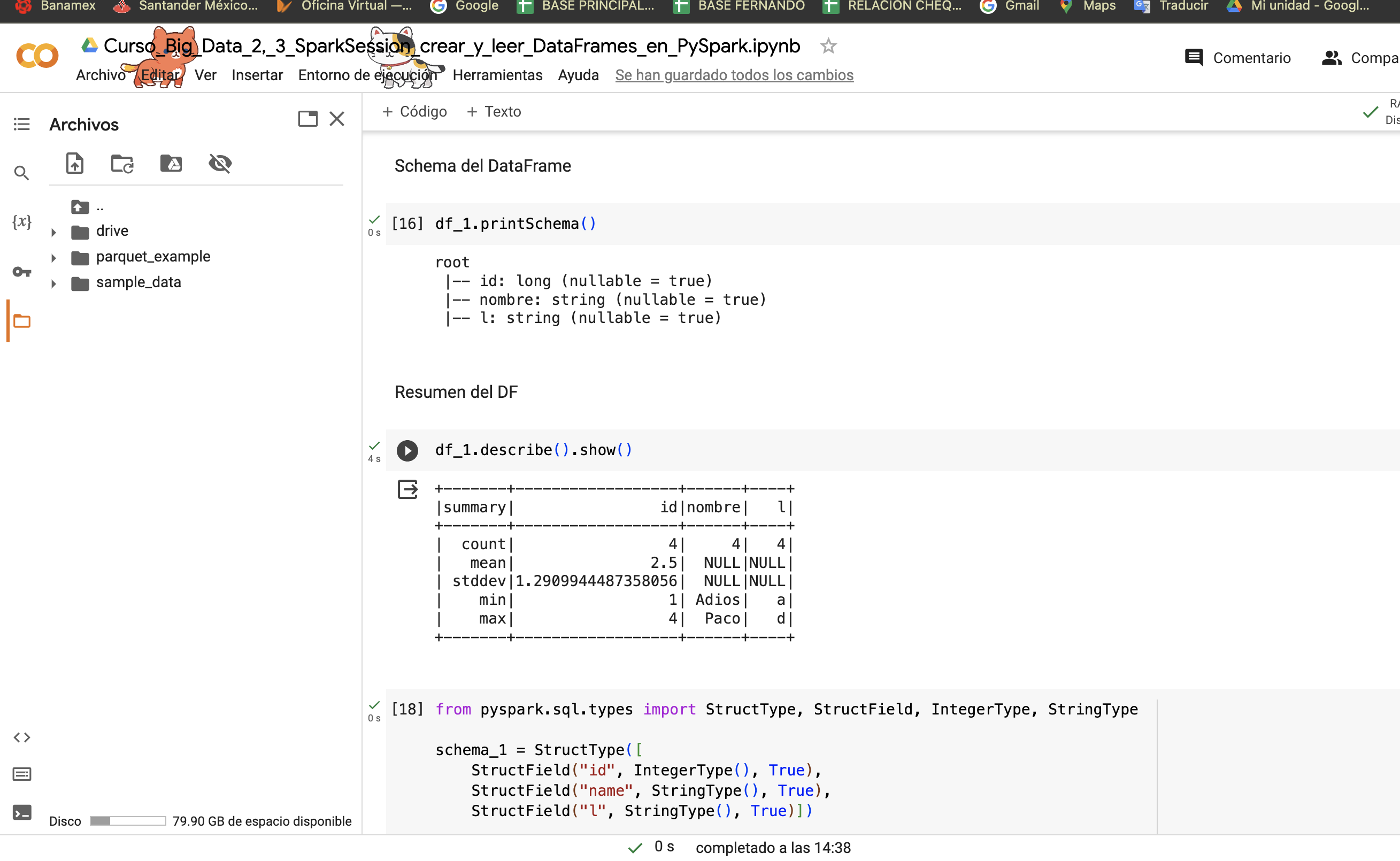
Prof: Miguel de Jesus Martinez Felipe

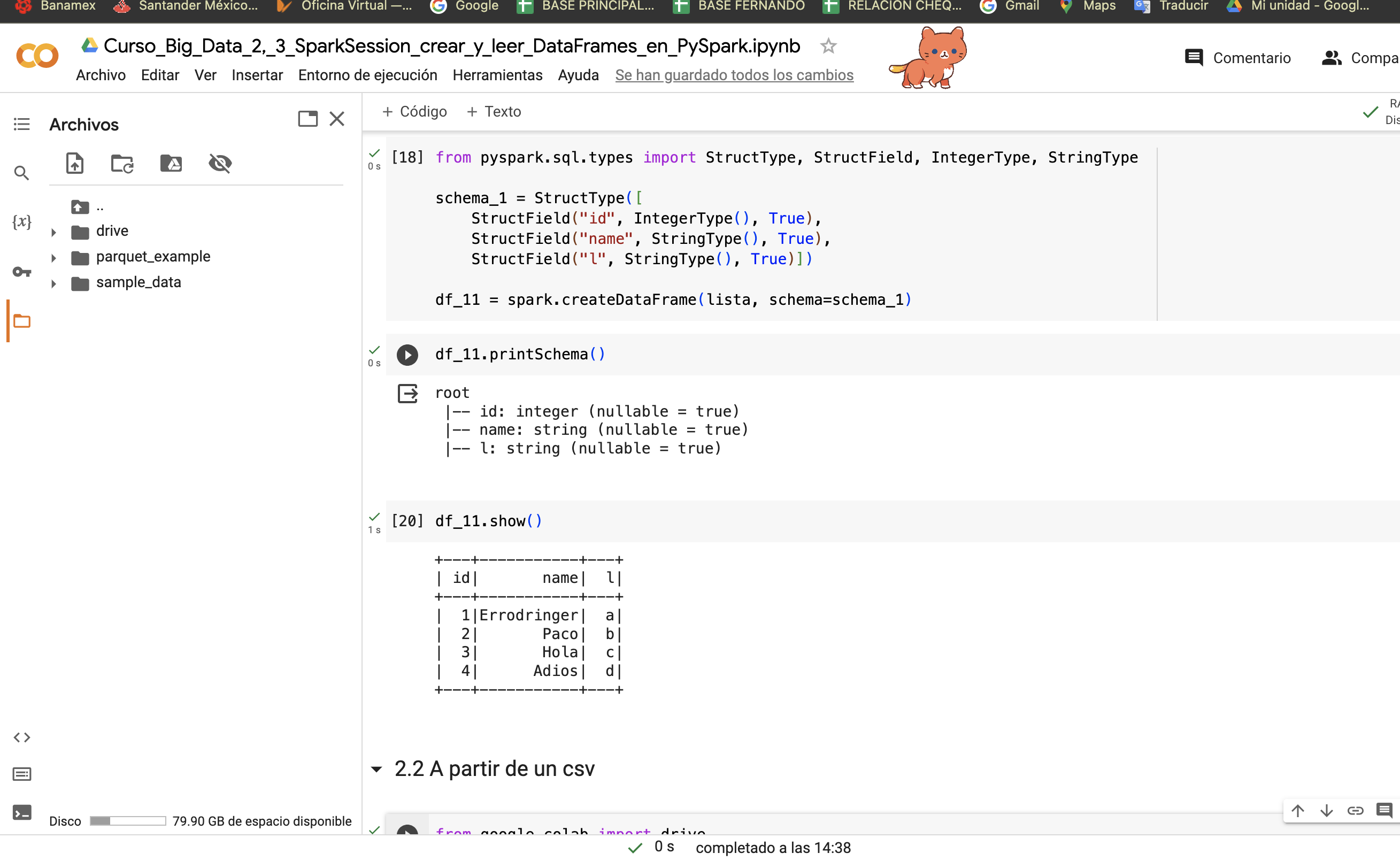
Yesica Butista Tapia | 2991269 | 27/11/23

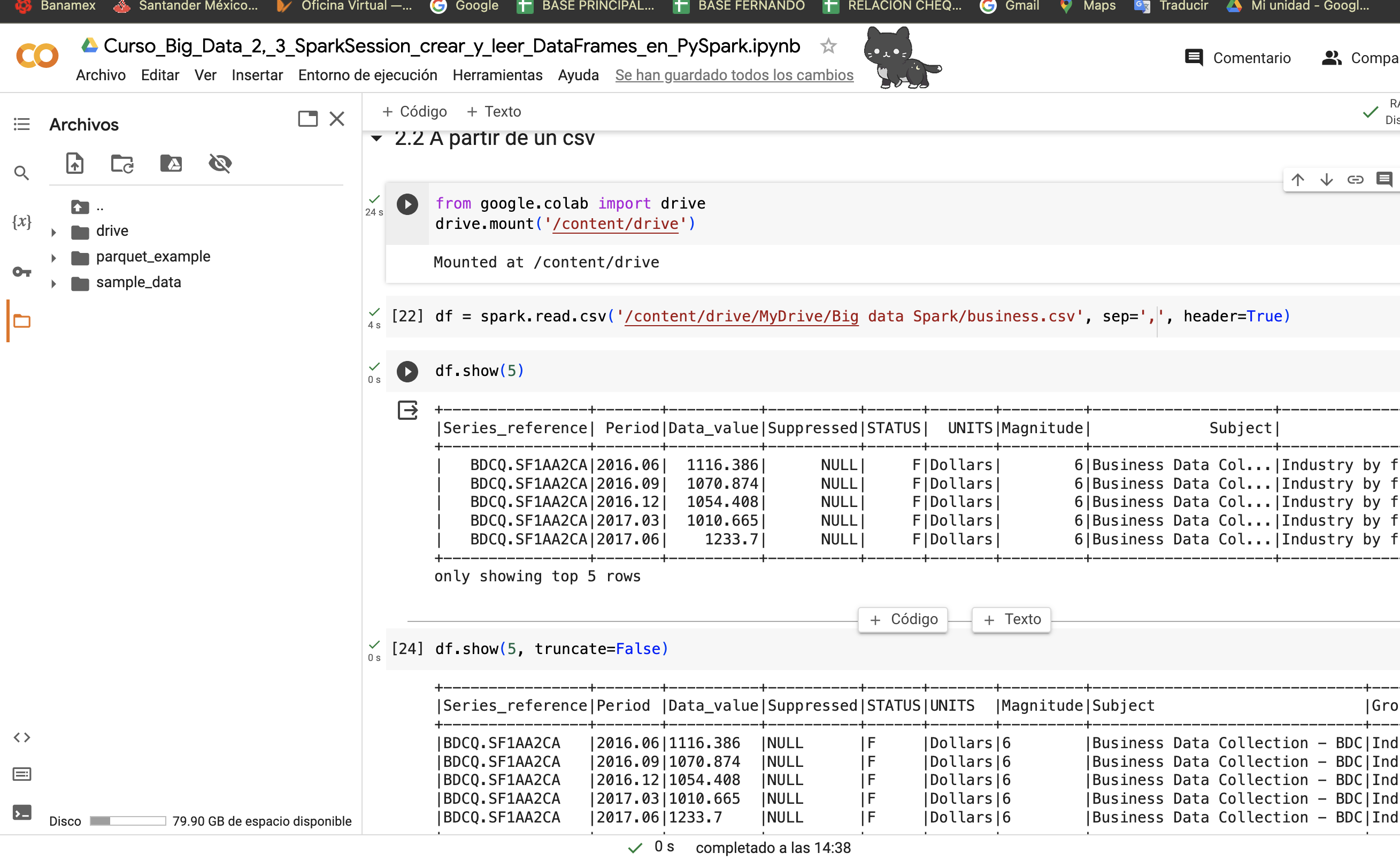
**Avance 2**

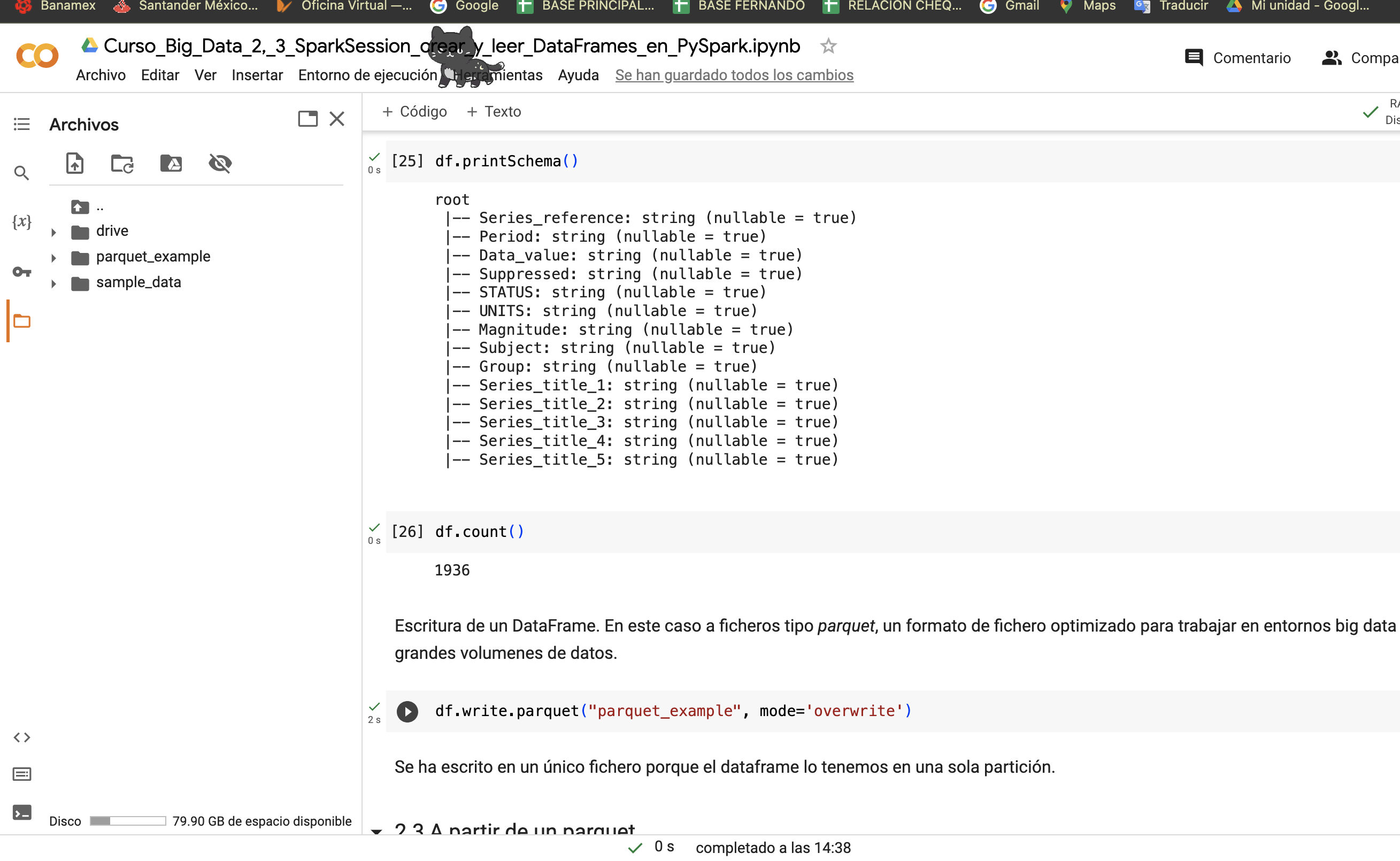
**Practica: gitHub:** [**https://github.com/YESICA287/Infraestructura-de-Big-Data/blob/main/Curso\_Big\_Data\_2%2C\_3\_SparkSession\_crear\_y\_leer\_DataFrames\_en\_PySpark.ipynb**](https://github.com/YESICA287/Infraestructura-de-Big-Data/blob/main/Curso_Big_Data_2%2C_3_SparkSession_crear_y_leer_DataFrames_en_PySpark.ipynb)

**Colab:** [**https://drive.google.com/file/d/1JSCrWktMvRxwYrJ8kqwaEgAzEL2wOKs9/view?usp=sharing**](https://drive.google.com/file/d/1JSCrWktMvRxwYrJ8kqwaEgAzEL2wOKs9/view?usp=sharing)

****

****

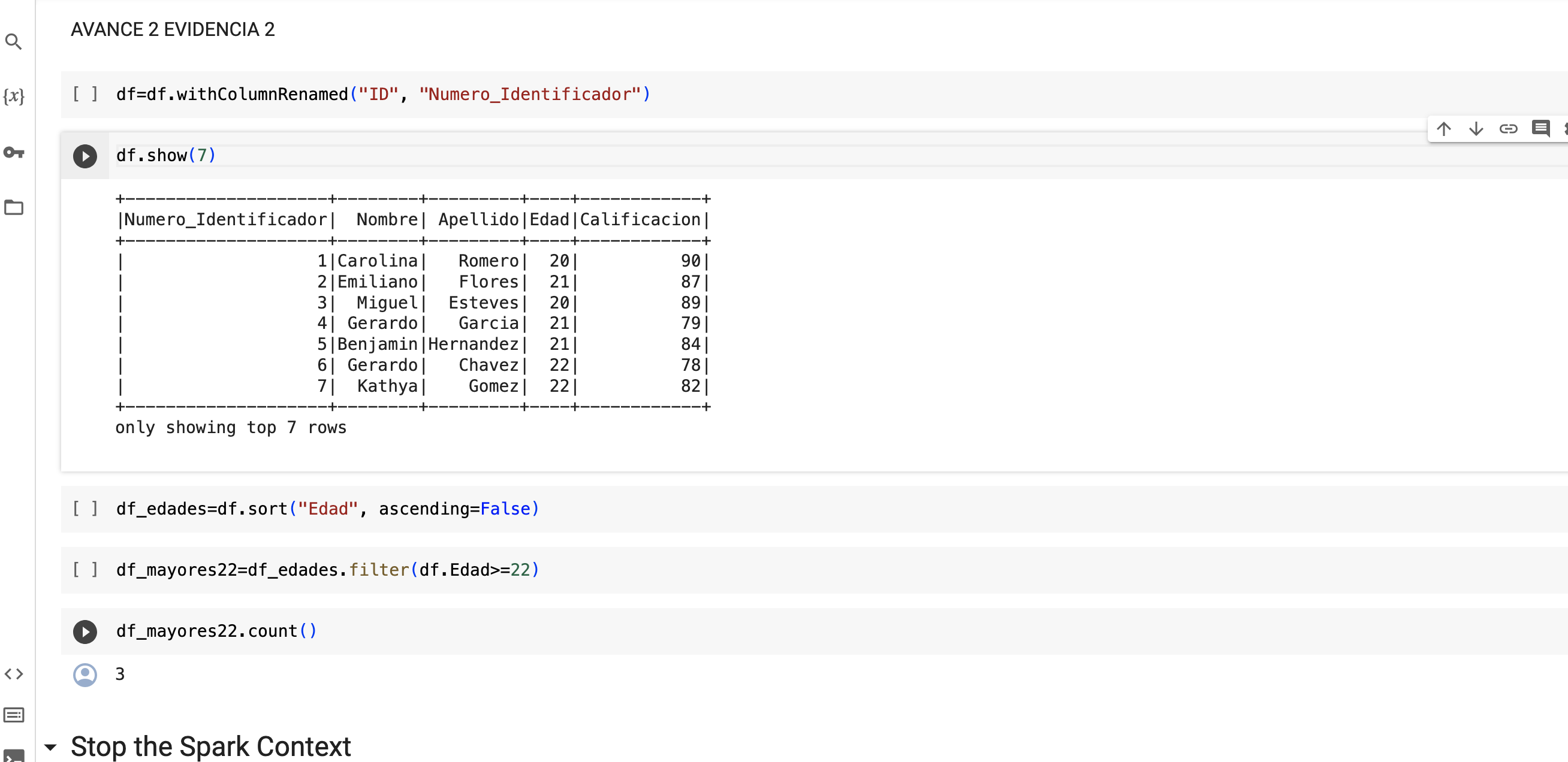
****

****

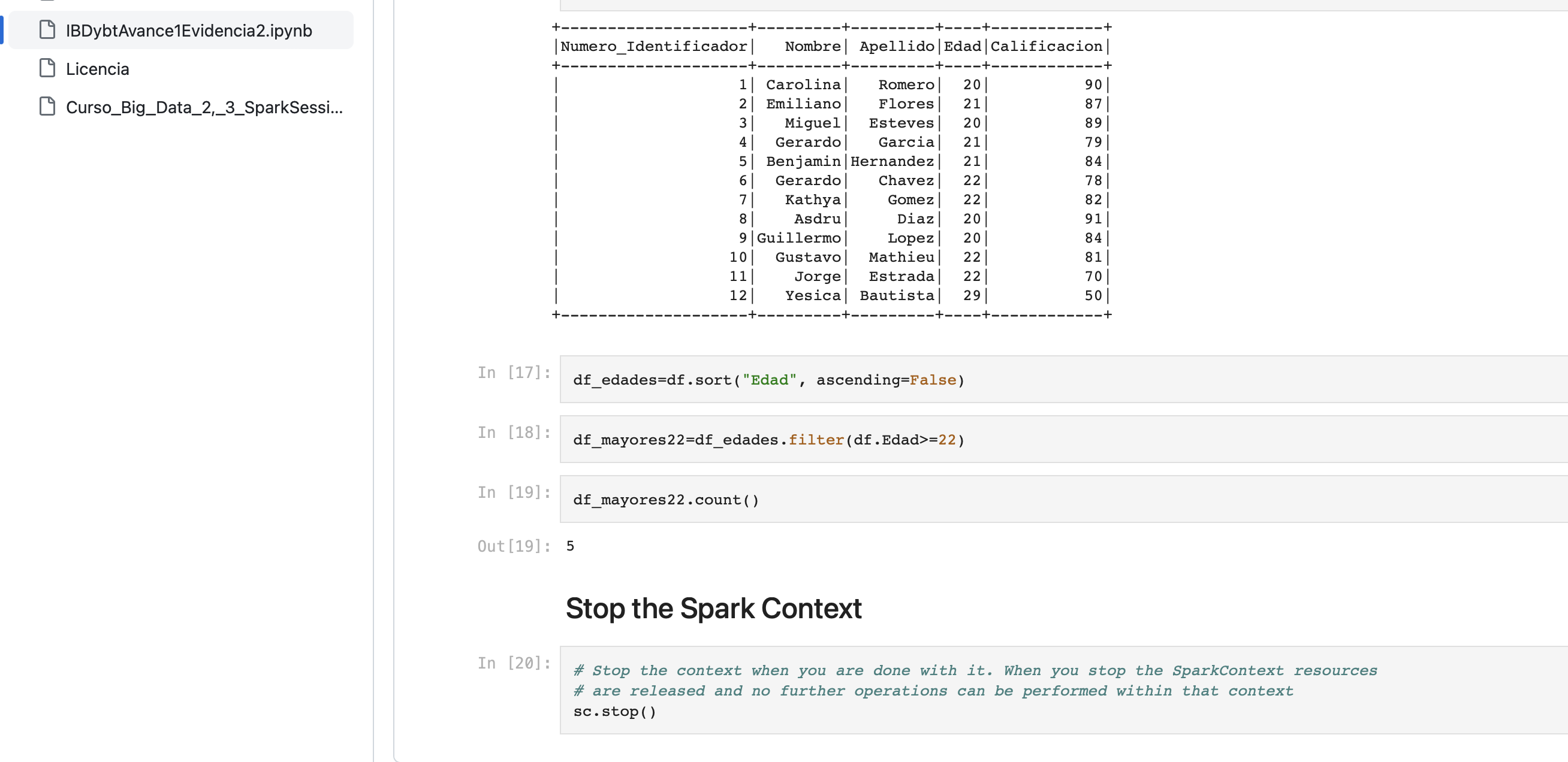
**Instrucciones canvas:**

Instrucciones con Pandas DataFrame:

1.- Modificar columna 'Id' por 'Numero\_Indetificador'  
2.- Mostrar los primero 7 datos.  
3.- Ocupar método de sort en su dataframe con la columna Edad  
4.- filtrar alumnos por edades con filter, edades mayores a 22 años.  
5.- Ocupar df.count(), df\_filter.count() para countar los estudiantes mayores a 22 años.

****

**Despues de imprimir los primero 7 datos, exploré cambiando la tabla para poder visualizar los 12 datos por eso el código al final quedo como la parte de abajo.**

****

**Reflexión:**

La realización de esta actividad ha sido sumamente valiosa para mi comprensión y destreza en el manejo de PySpark. A través de las diferentes etapas del proceso, he adquirido conocimientos significativos que quiero destacar.

Explorar la versatilidad de PySpark en el manejo de diversas fuentes de datos ha sido Interesante en mi punto de vista. Desde la carga de información desde archivos de distintos formatos hasta la creación de DataFrames a partir de diferentes fuentes, he comprendido la importancia de adaptarse a la diversidad de orígenes de datos. Esta capacidad de trabajar con datos de múltiples procedencias es esencial en un entorno donde la información proviene de diversas fuentes.

Asi tambié como reconocer la relevancia de configurar adecuadamente el objeto SparkSession como punto de entrada fue un hito clave en mi aprendizaje. Este componente no solo facilita la conexión con el clúster de Spark, sino que también permite establecer configuraciones esenciales para la aplicación. La comprensión de su manejo adecuado se traduce en un mejor rendimiento y eficiencia del proceso.