

Avance 1 Evidencia 2

Prof: Miguel de Jesus Martinez Felipe

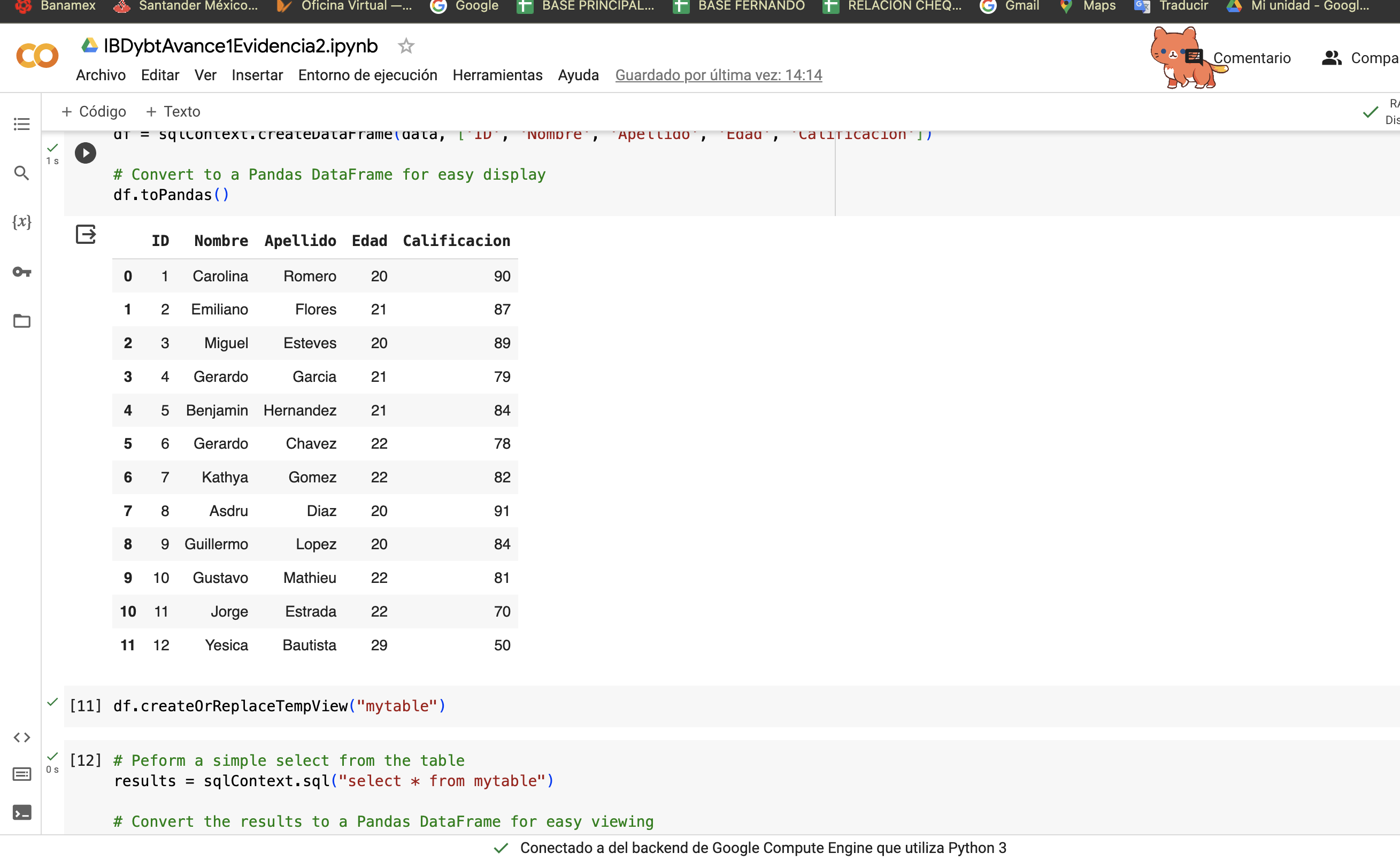
Yesica Butista Tapia | 2991269 | 27/11/23

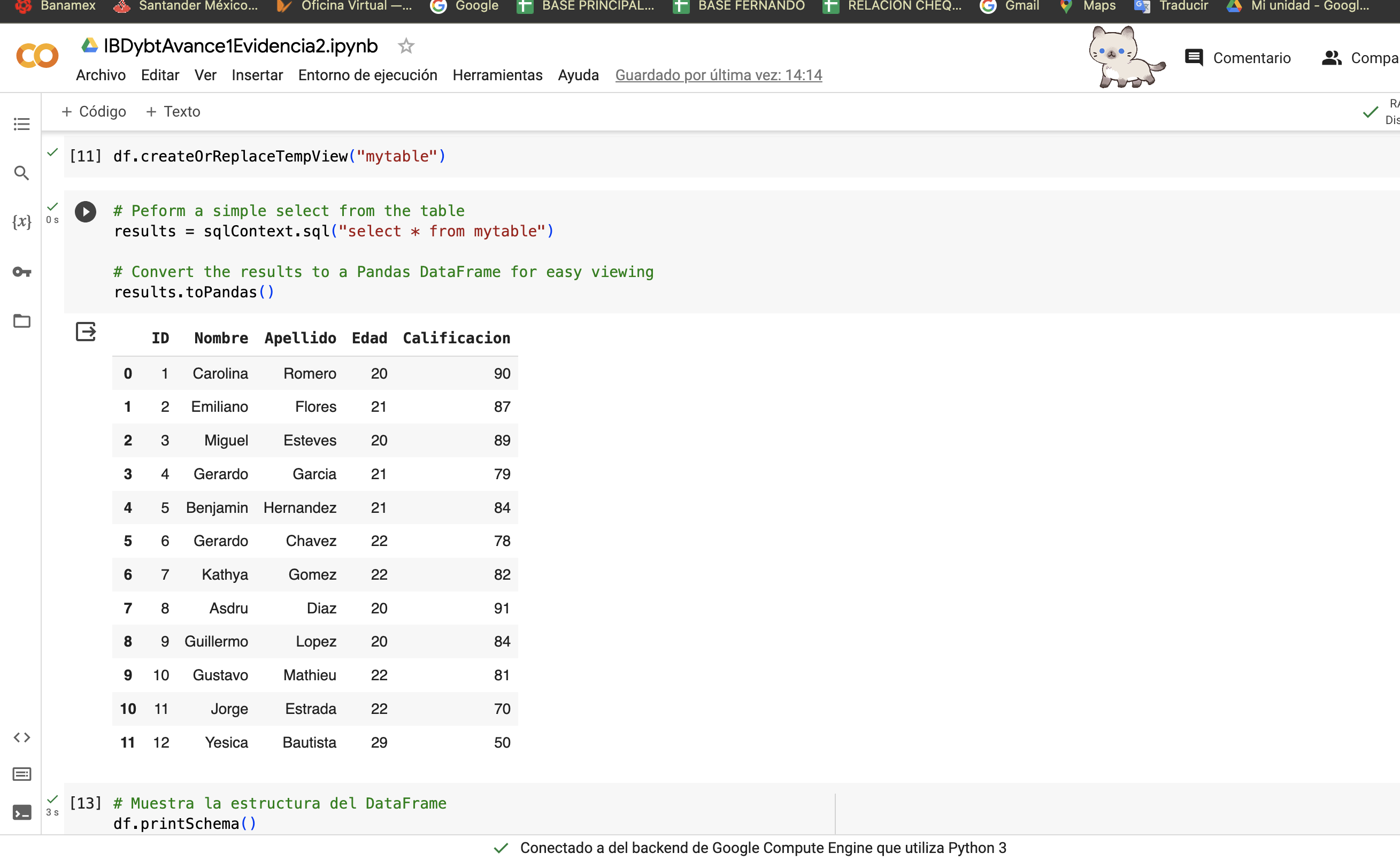
**Avance 1**

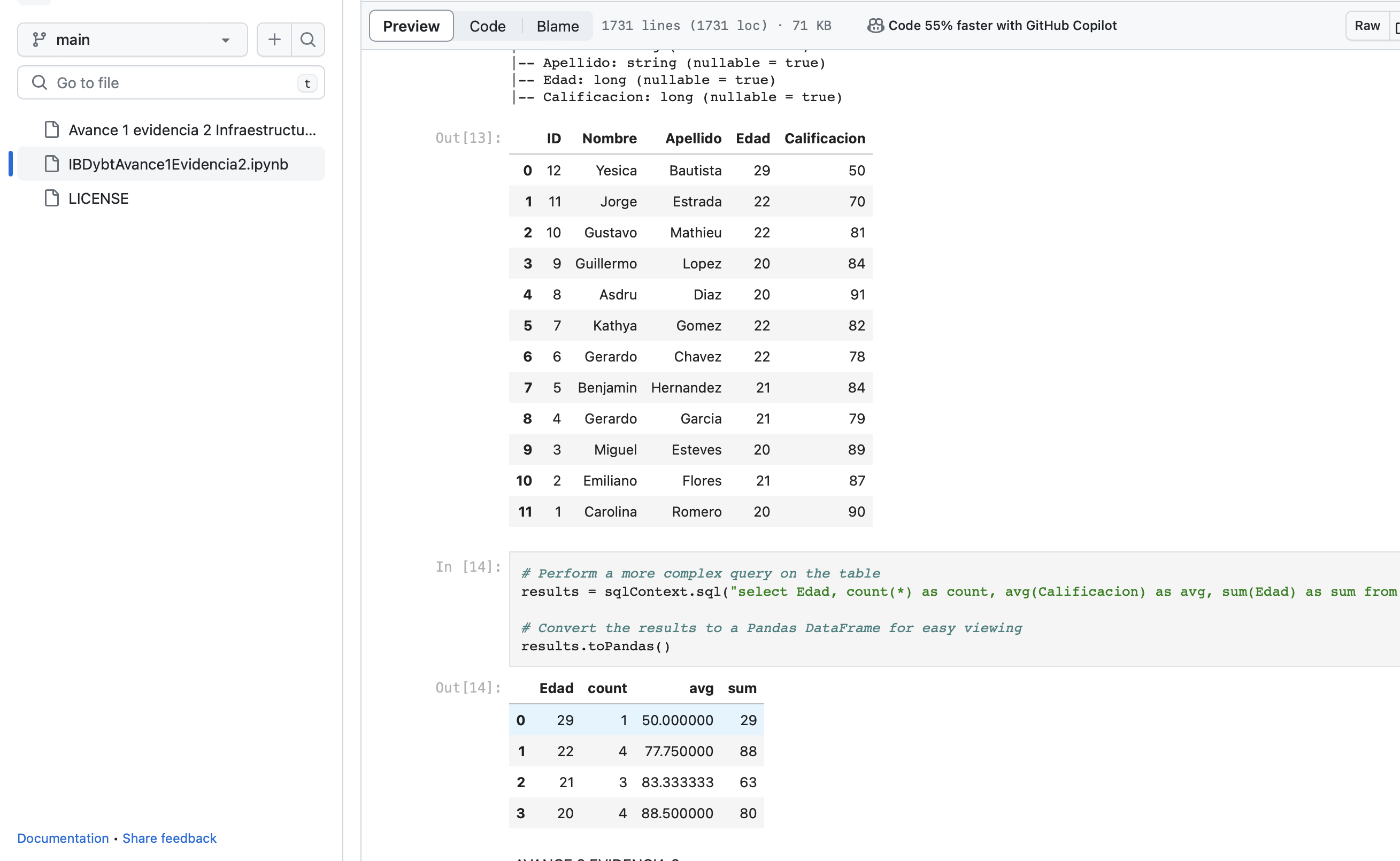
**Práctica:GitHub:** [**https://github.com/YESICA287/Infraestructura-de-Big-Data/blob/main/IBDybtAvance1Evidencia2.ipynb**](https://github.com/YESICA287/Infraestructura-de-Big-Data/blob/main/IBDybtAvance1Evidencia2.ipynb)

**Colab:** [**https://colab.research.google.com/drive/1EBr4Pu3FbgvwncFdAMEkoUM7BYZWRNWc?usp=sharing**](https://colab.research.google.com/drive/1EBr4Pu3FbgvwncFdAMEkoUM7BYZWRNWc?usp=sharing)

**Evidencias**

****

****

****

**Importancia de los Data Lakes en Hadoop:**

**1. Almacenamiento Versátil:**

* La flexibilidad de los data lakes para almacenar datos en diversos formatos, como texto, imágenes y datos no estructurados, simplifica la gestión de información heterogénea y favorece su análisis.
* **2. Economía de Escala:**
* Al aprovechar tecnologías de código abierto como Hadoop, los data lakes pueden almacenar grandes volúmenes de datos de manera eficiente desde el punto de vista de los costos, crucial en un entorno donde la generación de datos es constante.
* **3. Flexibilidad:**
* La ausencia de una estructura rígida permite a las organizaciones adaptarse fácilmente a nuevas fuentes de datos sin necesidad de rediseñar la arquitectura de almacenamiento.
* **4. Análisis Avanzado:**
* Con herramientas como Apache Spark y Hadoop, los data lakes ofrecen una plataforma robusta para realizar análisis profundos de conjuntos extensos de datos.

**Ventajas de los Data Lakes Multifuentes o Multiformato en un Ecosistema Hadoop:**

**1. Escalabilidad:**

* La alta escalabilidad de los data lakes les permite crecer para cumplir con las crecientes necesidades de almacenamiento de datos.
* **2. Integración de Datos:**
* Facilitan la consolidación de datos provenientes de diversas fuentes, fomentando así la integración de información y contribuyendo a la toma de decisiones empresariales más informadas.
* **3. Acceso a Datos Históricos:**
* La retención eficiente de datos en un data lake permite un acceso sencillo a información histórica, propiciando análisis retrospectivos.

**Desventajas de los Data Lakes Multifuentes o Multiformato en un Ecosistema Hadoop:**

**1. Gestión de Calidad de Datos:**

* La diversidad de formatos y fuentes puede desencadenar problemas de calidad de datos si no se implementan procesos adecuados de limpieza y transformación.
* **2. Costos de Almacenamiento y Administración:**
* Aunque el almacenamiento de datos en data lakes puede ser económico, los costos asociados con la gestión, seguridad y acceso pueden ser significativos.
* **3. Complejidad:**
* La administración de un data lake implica complejidades y demanda habilidades técnicas avanzadas, ya que la falta de una estructura rígida puede complicar la organización y búsqueda de datos.

**Reflexión:**

En resumen, los data lakes multifuentes son un componente esencial en la evolución de las arquitecturas de datos. Su capacidad para adaptarse a la diversidad de datos y su papel central en el análisis avanzado los posicionan como una herramienta invaluable para organizaciones que buscan extraer conocimientos significativos de sus activos de datos. No obstante, el éxito en su implementación requiere una cuidadosa consideración de los desafíos asociados y un enfoque estratégico para maximizar su potencial.

**Referencias:**

* **Stedman, C., & Lutkevich, B. (2021, octubre 13). *What is a data lake? Definition from SearchDataManagement*. Data Management; TechTarget.** [**https://www.techtarget.com/searchdatamanagement/definition/data-lake**](https://www.techtarget.com/searchdatamanagement/definition/data-lake)
* Harris, J. (2017, junio 15). *On the way to data lakes and Hadoop – Part 2*. The Data Roundtable. <https://blogs.sas.com/content/datamanagement/2017/06/15/on-the-way-data-lakes-and-hadoop2/>
* *Data lake service*. (2020, diciembre 2). Cloudera. <https://www.cloudera.com/products/sdx/data-lake-service.html>
* (S/f-d). Techopedia.com. Recuperado el 28 de noviembre de 2023, de <https://www.techopedia.com/definition/34874/data-lakehouse>
* Envizi, I. B. M. (2020, mayo 4). *Data lakes and analytics: powerful alone, but better together*. IBM Blog. <https://www.ibm.com/blog/data-lakes-and-analytics/>