**“1. Cargar los datos**

* **Importa el archivo Excel proporcionado por el tutor (disponible en los recursos del curso).**
* **Visualiza las primeras filas para identificar la estructura general de las columnas.”**

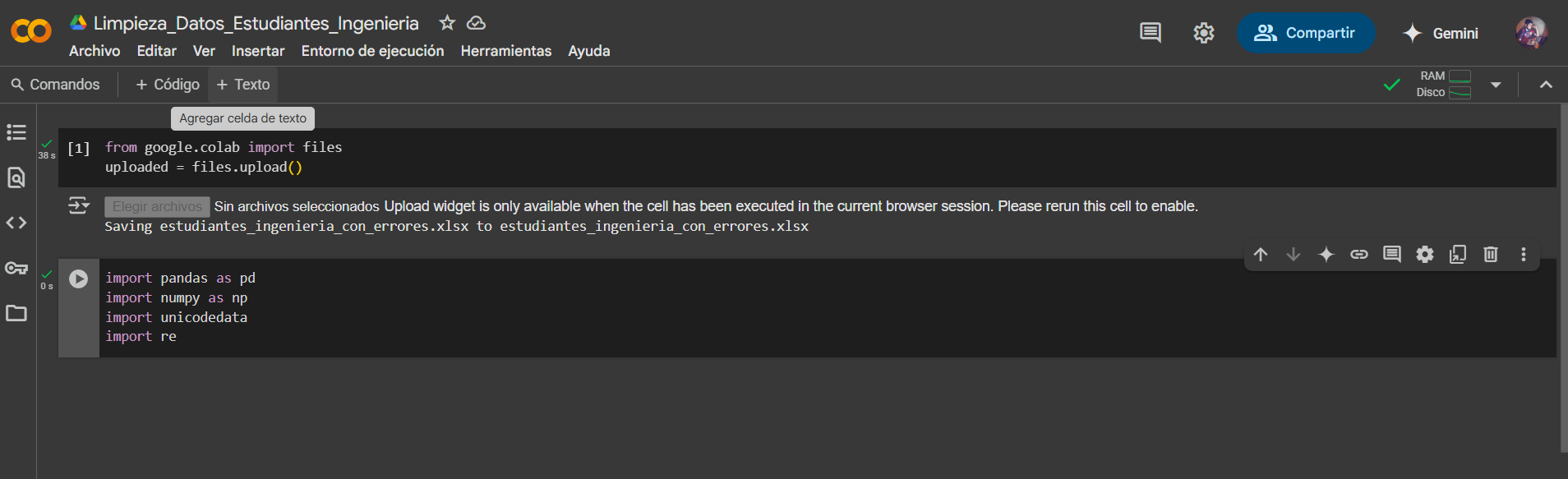
1. función para adjunta documento

Interfaz de usuario gráfica, Texto

El contenido generado por IA puede ser incorrecto.

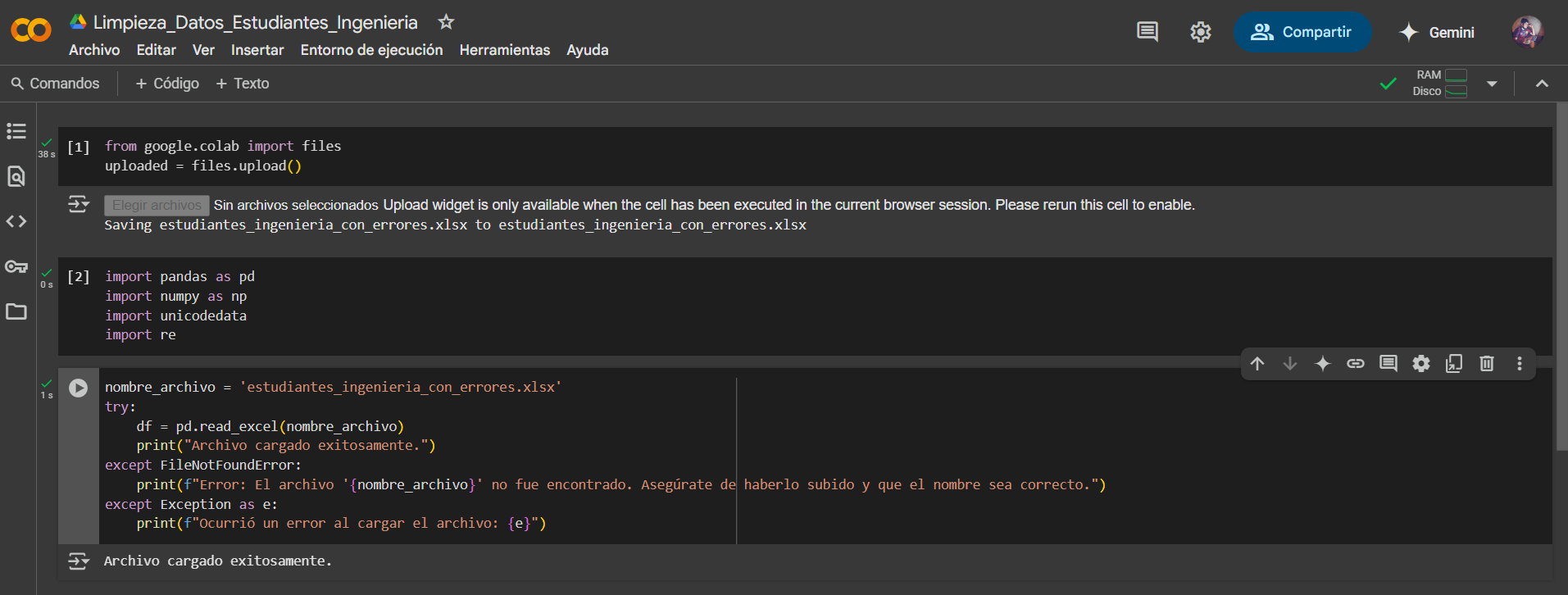
1. Cargo librerías para limpieza según el requerimiento necesitado

* Pandas: Analizar datos en tablas
* Numpy: Para realizar cálculos
* Unicodedata: Trabajar con caracteres especiales para normalizar información o eliminar
* Re: Manejar patrones de texto.

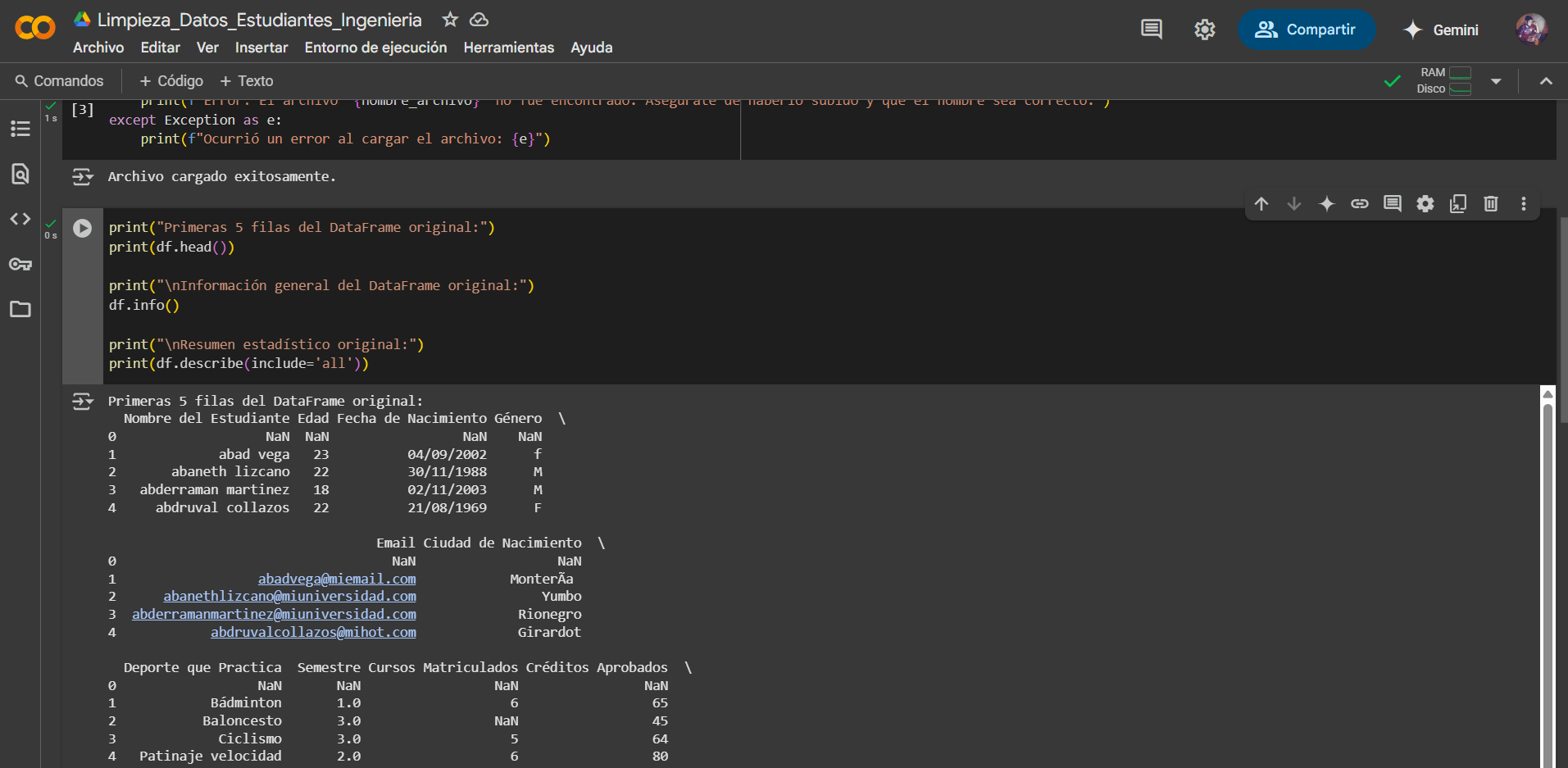


1. Crear DataFrame

Con el archivo adjunto, se realizara un DataFrame para que la librería de Panda pueda leerlo, en donde se están colocando condicionales de nombre del documento para hacer validación del nombre correcto, sino existe el archivo, lectura y resultado de esta misma lectura



1. Revisión de la estructura del documento mediante Python para identificar las columnas (cabeceras) y que información contienen.



Interfaz de usuario gráfica, Texto

El contenido generado por IA puede ser incorrecto.

**2. Identificar y eliminar registros no válidos**

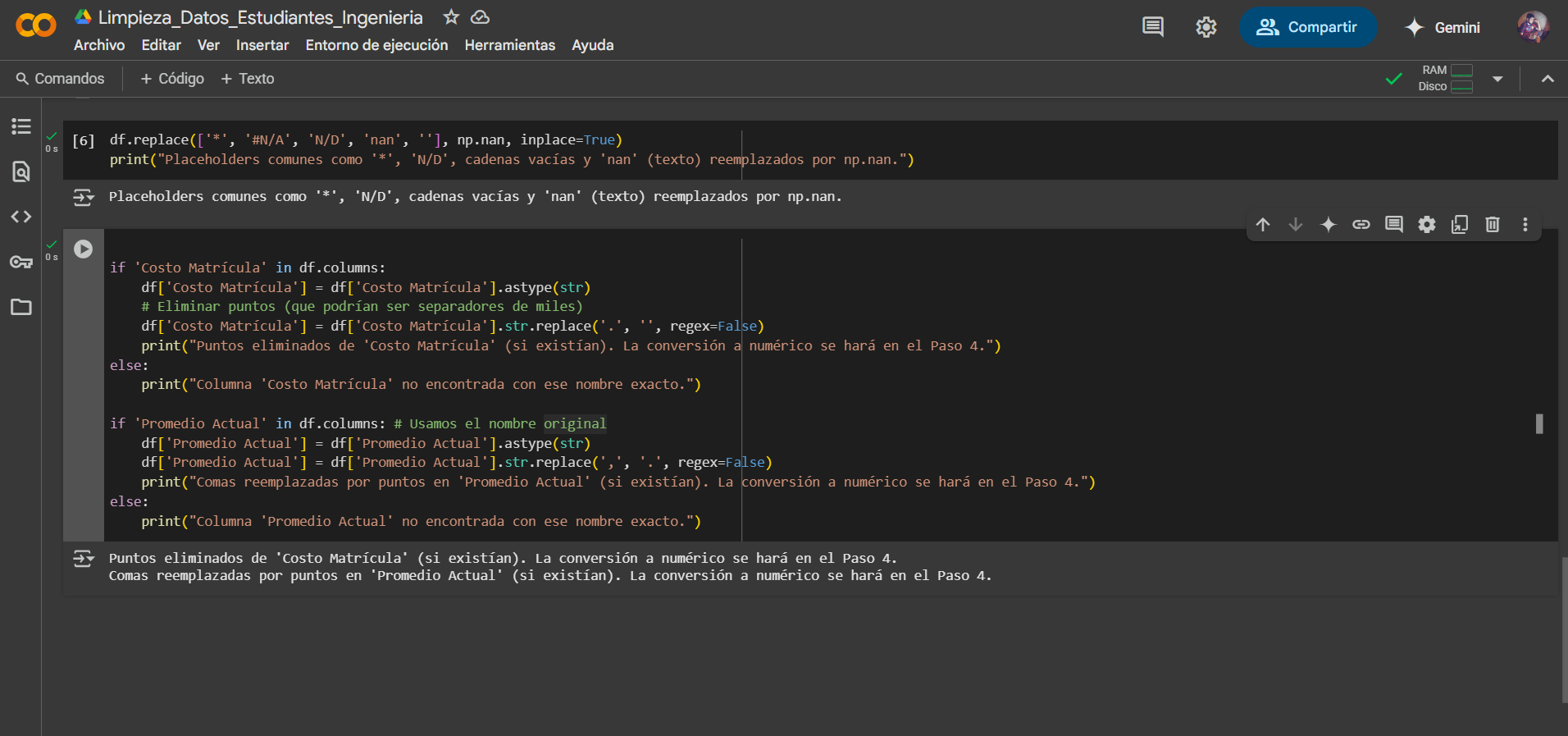
* Elimina filas completamente vacías.
* Detecta y corrige valores con errores como asteriscos, palabras en campos numéricos (ej. "cuarenta"), formatos no estándar (ej. "2.300.000", "tres punto siete").

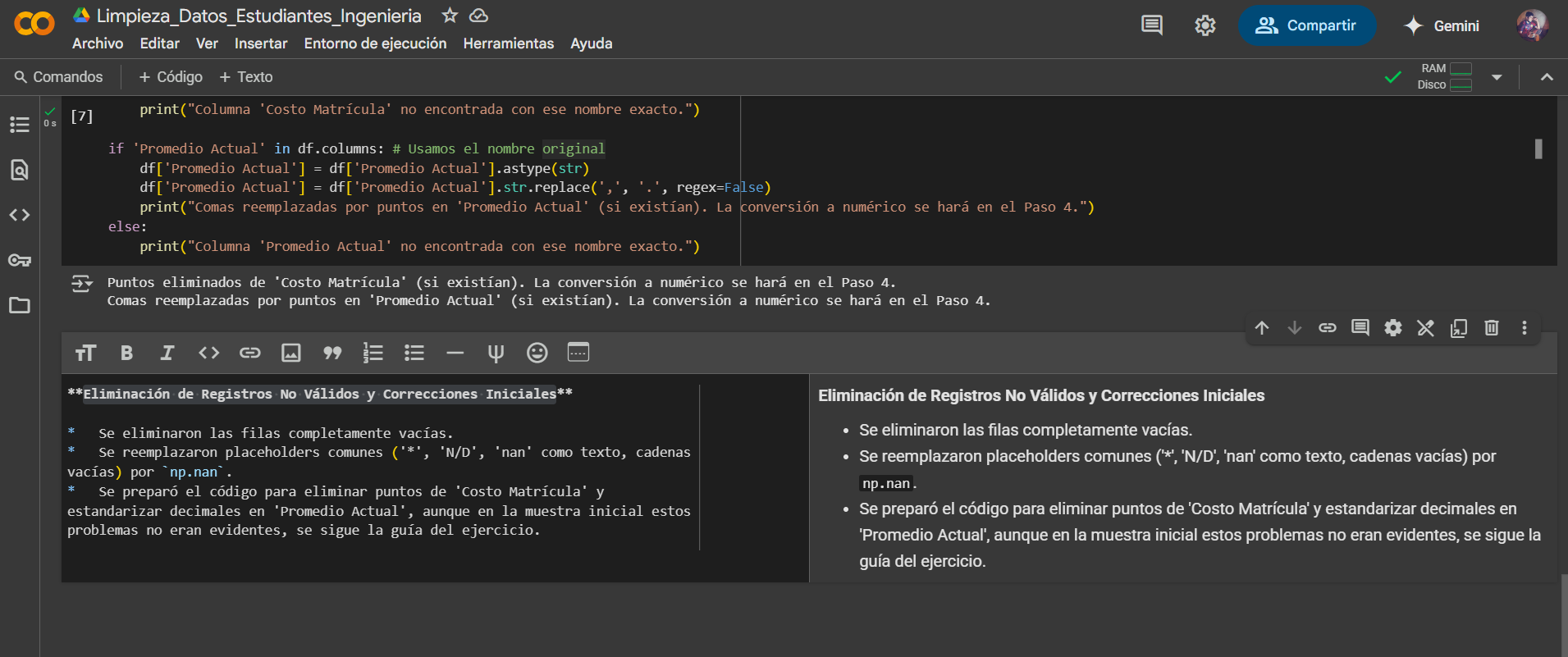
1. Primero se debe identificar cuantas filas hay vacías. En la cual se usa len(df), que me devuelve el numero de filas del Dataframe. Y el df(dropna) me ayuda a eliminar tanto filas como columnas con valores NaN, y el howw=all que valide y ejecute la eliminación siempre y cuando los resultados de los valores son NaN

Interfaz de usuario gráfica, Texto, Aplicación

El contenido generado por IA puede ser incorrecto.

1. Inicio depuración de (.) (,) (caracteres especiales) en columnas de valores numéricos para identificar y hacer la operación.

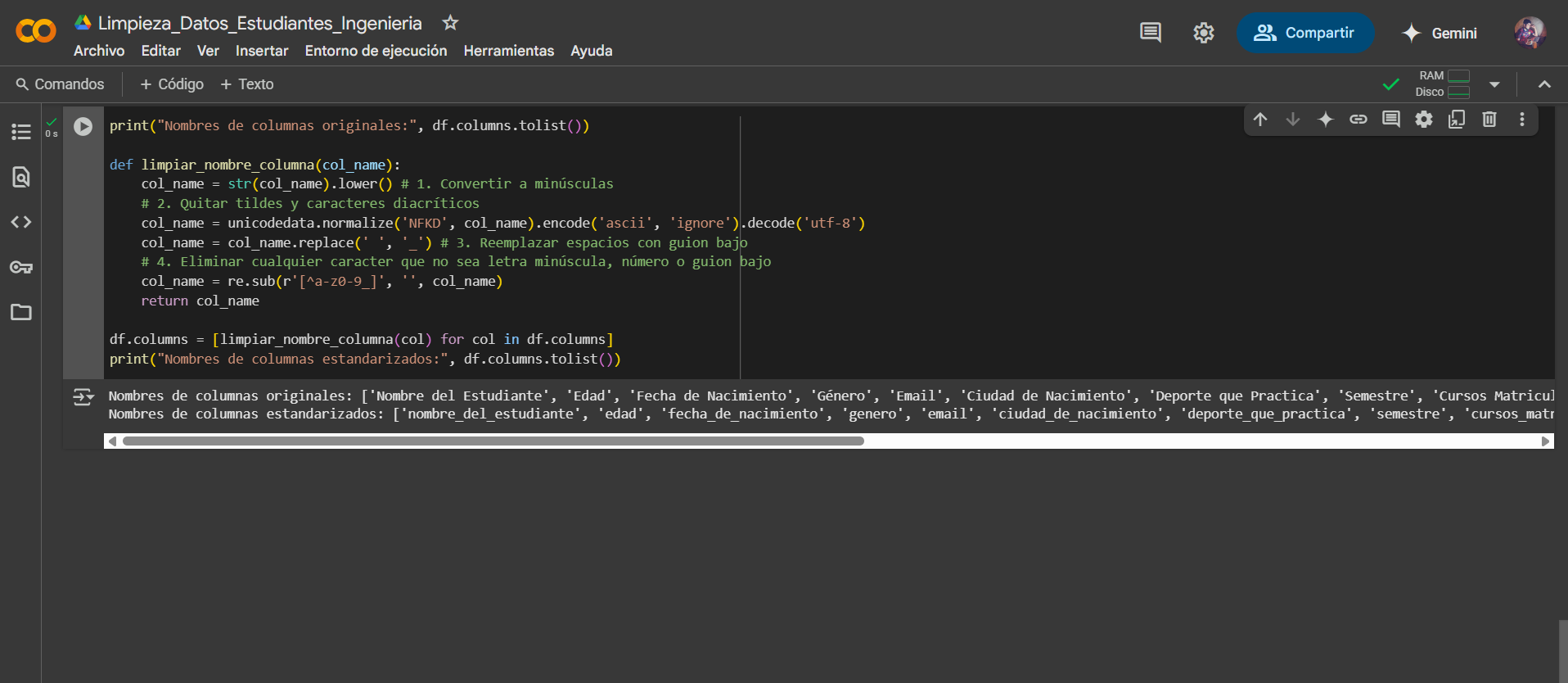




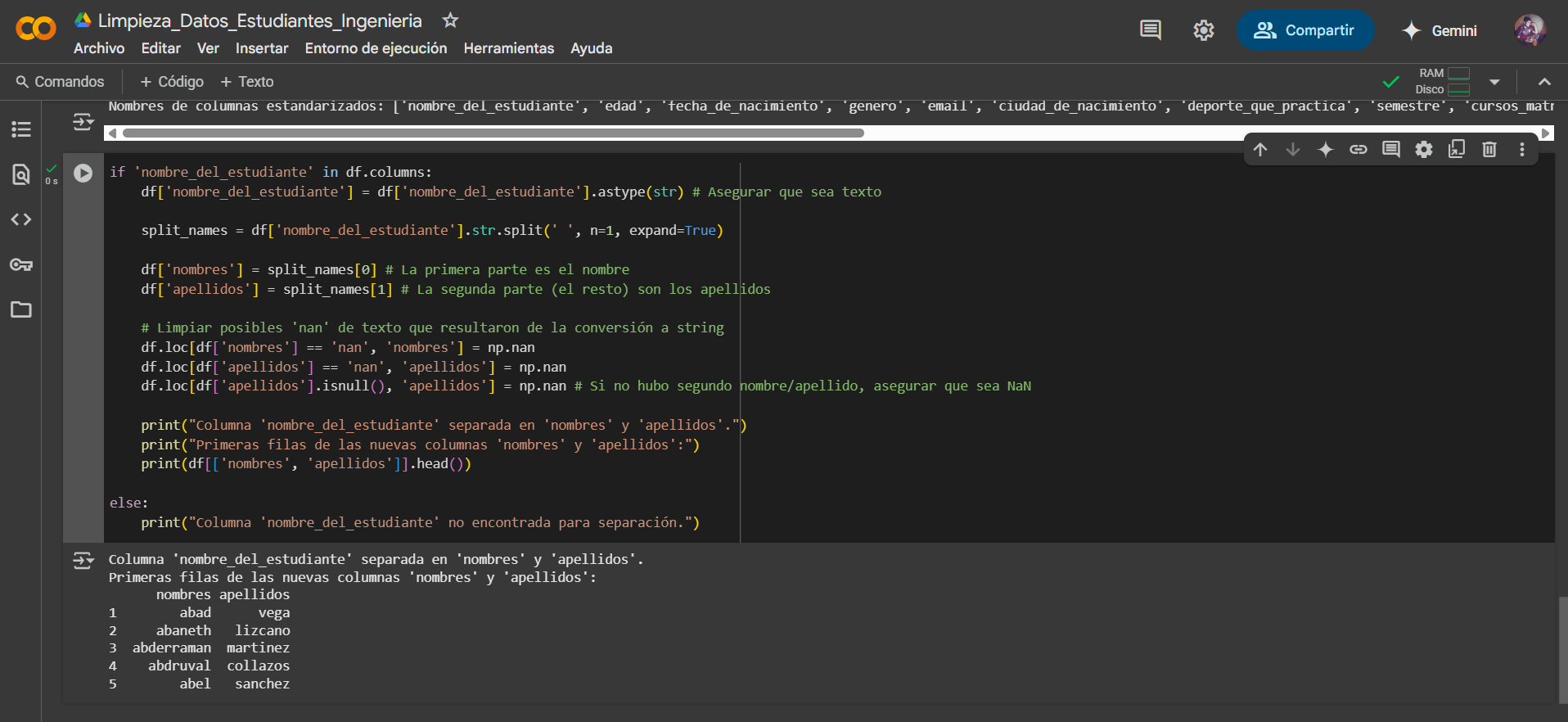
**3. Estandarización de columnas**

* Asegúrate de que los nombres de columnas tengan formato uniforme.
* Si es necesario, **separa nombres y apellidos** en columnas distintas.
* Ajusta los tipos de datos de cada columna (números como int o float, texto como string).

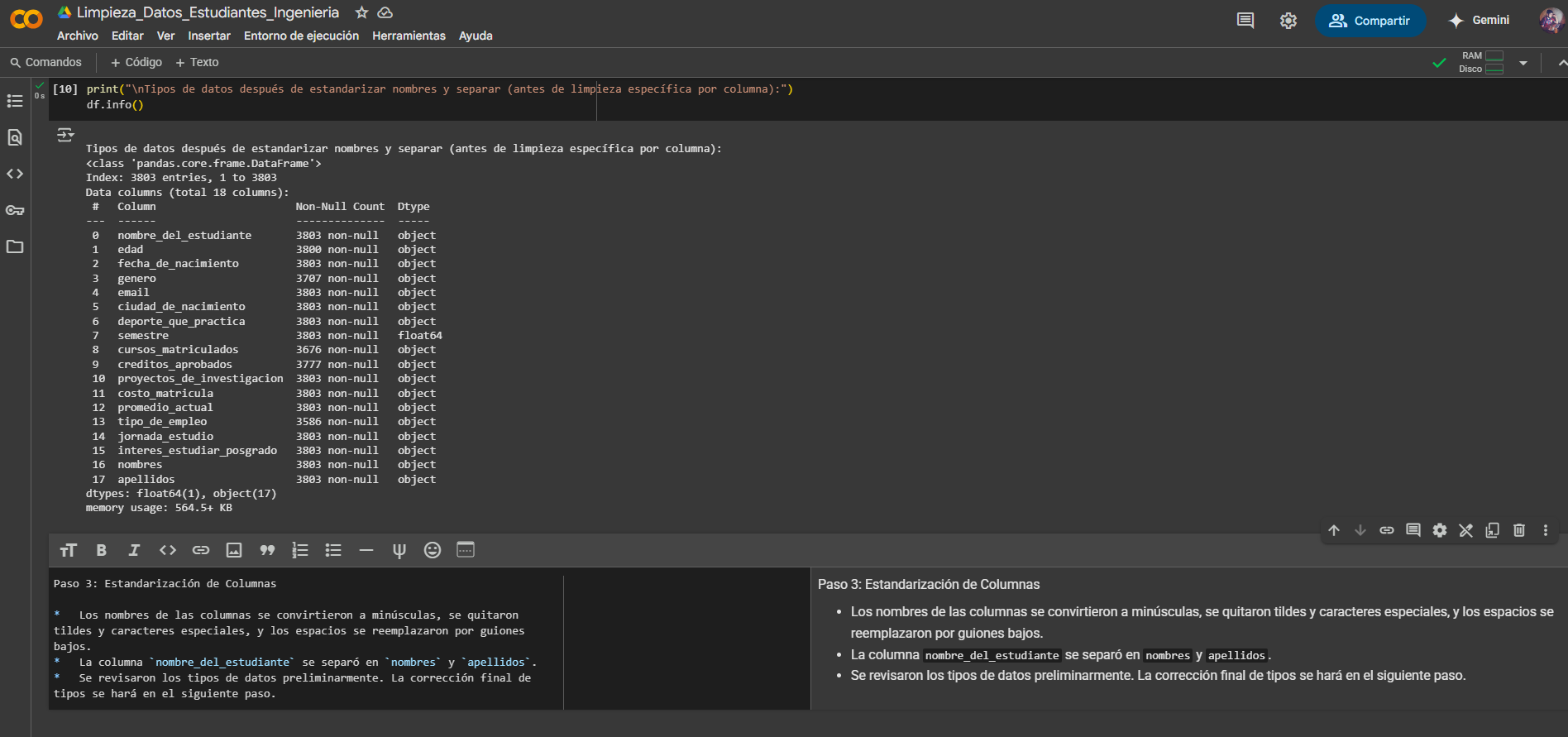
1. Debo encontrar los nombres de las columnas originales, para eso uso Python para llamar cada columna con la función col\_name, y comienzo a definicar que los nombres sean tratados como textos, valido temas de mayúscula, minúscula, y además quitar tildes como caracteres especiales. De esta manera convierto texto a formato ASCII, para después convertirlo en formato UTF-8, reemplazo todos los espacios por (\_) “Guion bajo”.



1. La columna nombre\_del\_estudiante (que originalmente era "Nombre del Estudiante") probablemente contiene tanto el nombre como el apellido. Por lo cual se hace la segmentación independiente de cada dato



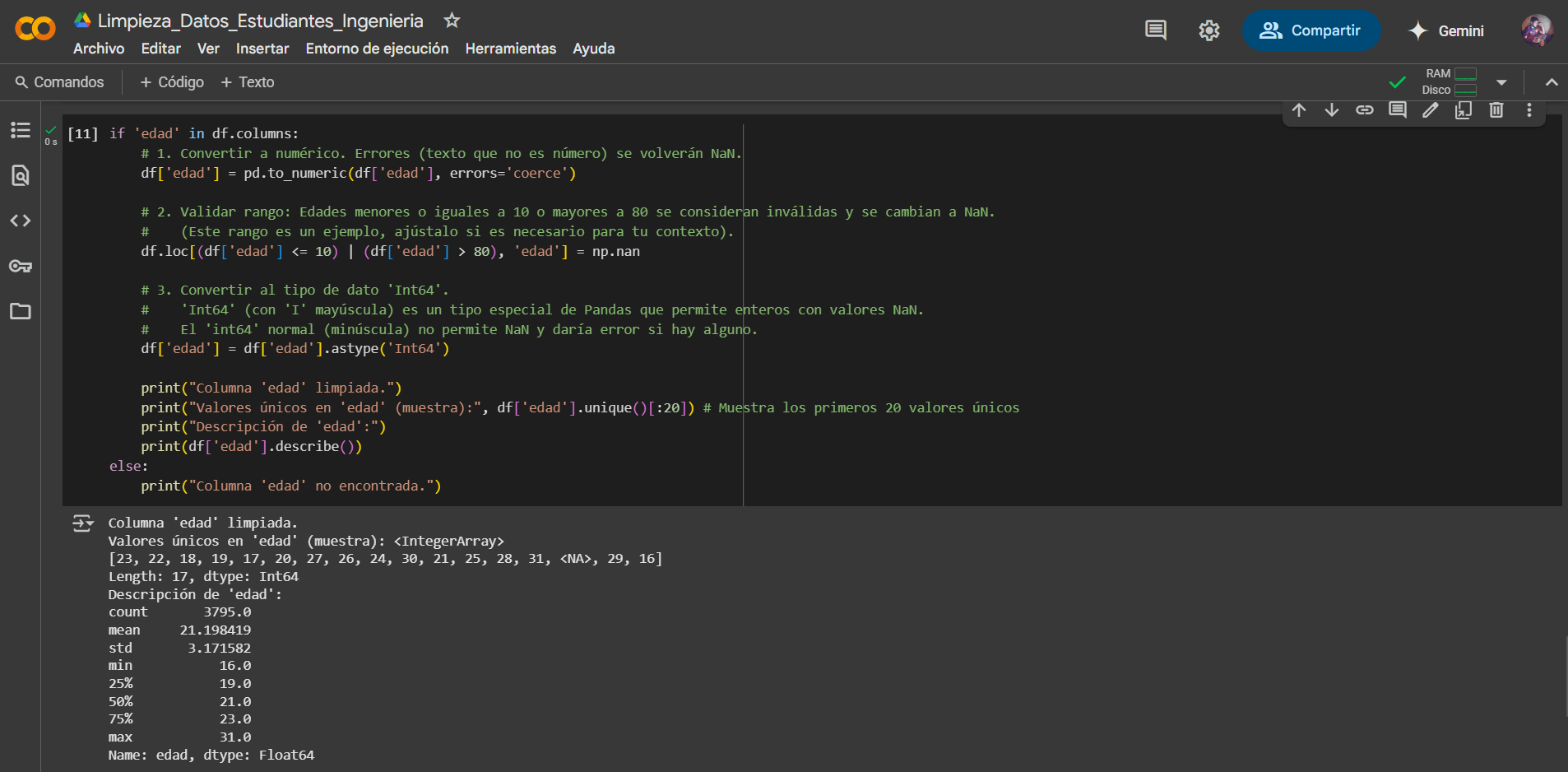
1. Se hace una vista de las columnas que se han estandarizado como muestra de que los script fueron funcionales



**4. Limpieza específica por columna**

* **Edad**: Convertir a número. Eliminar o corregir valores fuera de rango (por ejemplo, mayores de 80 años).
* **Género**: Validar que solo existan las categorías “M” o “F”. Corregir inconsistencias como “f”, “x”, “otro”.
* **Ciudad de nacimiento**: Corregir errores ortográficos y caracteres no alfabéticos.
* **Semestre**: Validar que esté entre 1 y 10. Eliminar o corregir valores inválidos como 13 o 20.
* **Promedio actual**: Convertir a formato decimal uniforme (usar puntos, no comas).
* **Costo matrícula**: Eliminar separadores de miles (ej. puntos) y convertir a tipo numérico.

1. Iniciamos con Edad en la limpieza de información



1. Seguimos con Genero

Captura de pantalla de computadora

El contenido generado por IA puede ser incorrecto.

1. Ciudad de nacimiento

Texto

El contenido generado por IA puede ser incorrecto.

1. Semestre

Interfaz de usuario gráfica, Texto, Aplicación

El contenido generado por IA puede ser incorrecto.

1. Promedio

Interfaz de usuario gráfica, Texto, Aplicación

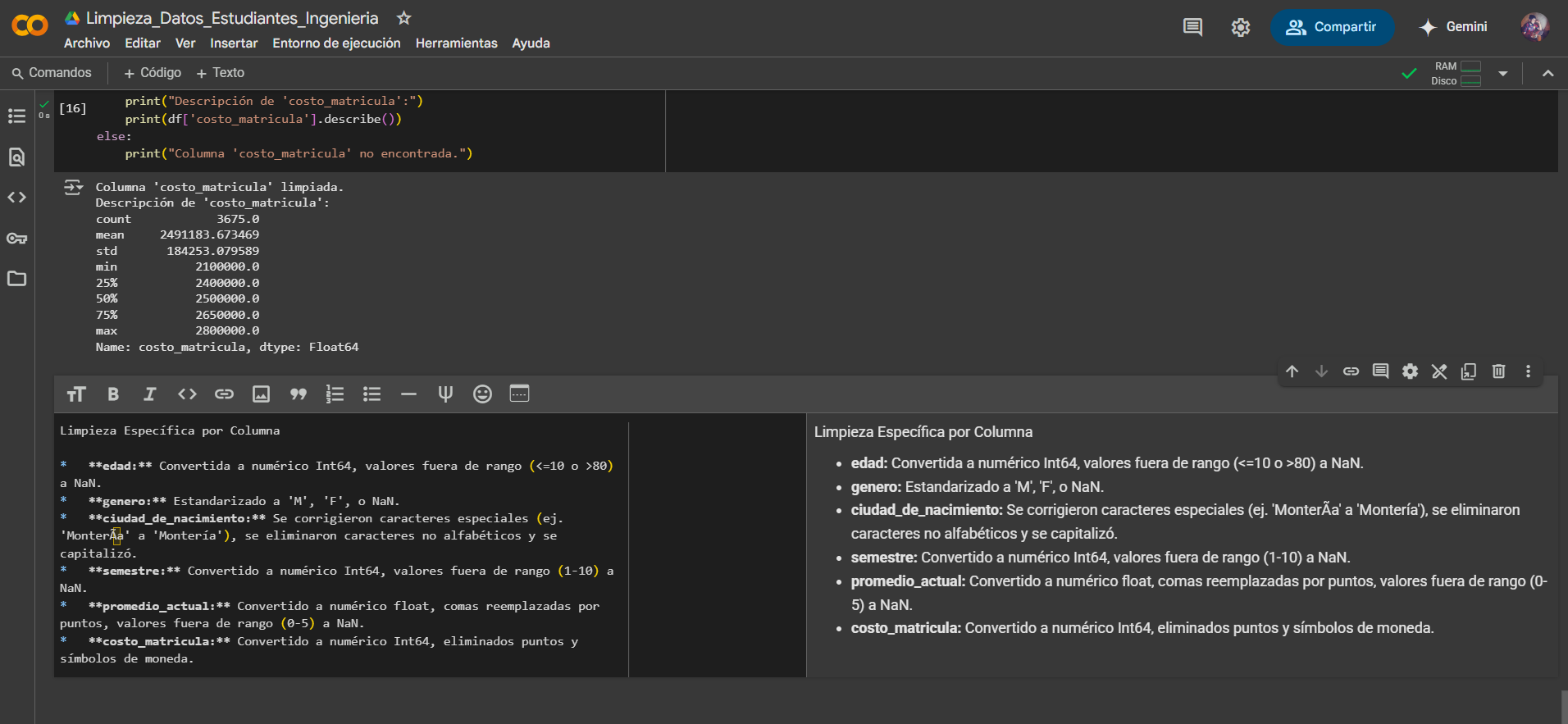
El contenido generado por IA puede ser incorrecto.

1. Costo de matriculas

Captura de pantalla de un celular

El contenido generado por IA puede ser incorrecto.

Justificación de cada limpieza de los 6 items



**5. Generar nuevo archivo limpio**

* Guarda el archivo final con el nombre: estudiantes\_ingenieria\_limpios.xlsx.

Genero una vista rápida de todas las tablas y los datos que quedaron con NAN o NA. Que son aquellos datos equivocados o erróneos.

Texto

El contenido generado por IA puede ser incorrecto.

Guardo toda la limpieza en un nuevo Excel, con el nombre recomendado según la guía

Texto

El contenido generado por IA puede ser incorrecto.

