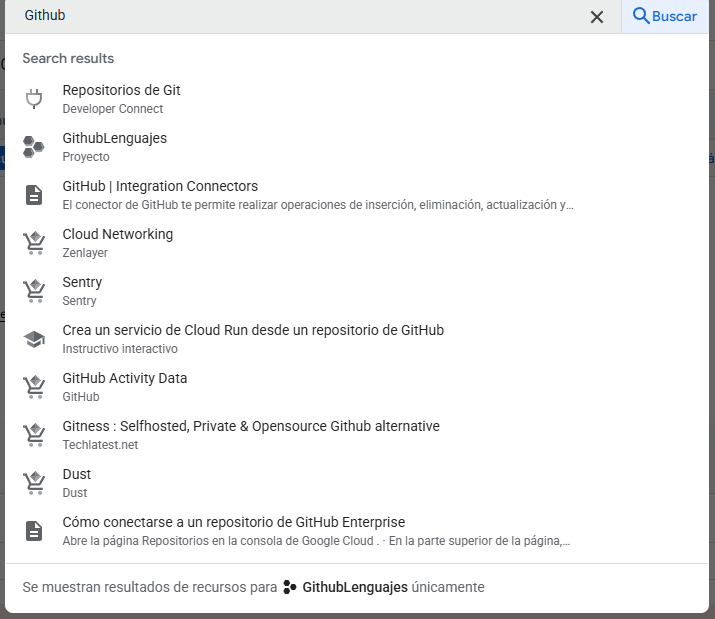
***Proyecto final de análisis de datos Lenguajes mas usados en Github***

Para este tema, se recolectará datos de Google BigQuery, que funcionará como base de datos relacional y al mismo tiempo nos proporcionará el data set que buscamos

<https://console.cloud.google.com/bigquery?p=bigquery-public-data&d=github_repos&page=dataset&inv=1&invt=Ab4c2w&project=ambient-core-467817-a9&ws=!1m10!1m4!4m3!1sbigquery-public-data!2sgithub_repos!3slanguages!1m4!1m3!1sambient-core-467817-a9!2sbquxjob_7c75101d_1986db2b8da!3sUS>

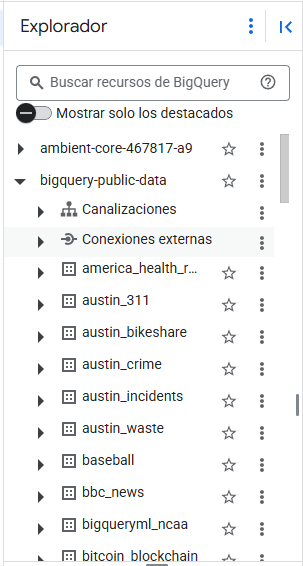
1. Registrarnos en Google BigQuery, una vez adentro, buscaremos en la barra de búsqueda Github y seleccionaremos activity data



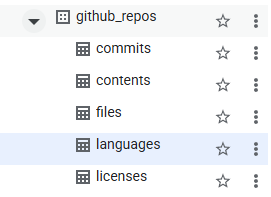
1. Vamos a conjunto de datos



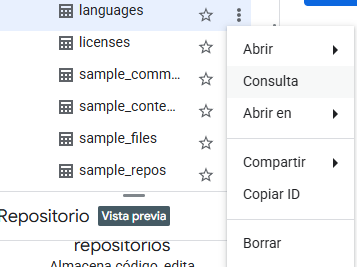
1. Desplegaremos bigquerydata-public-data en el explorador



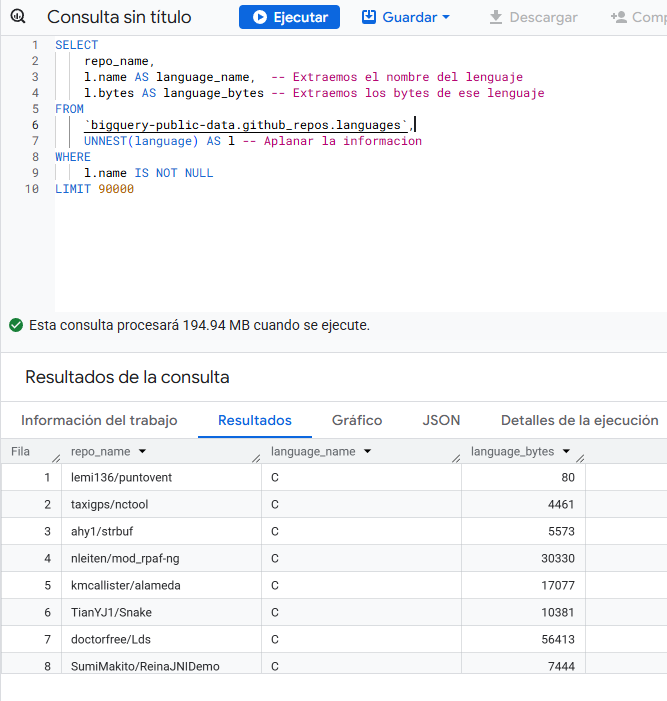
1. Buscaremos github\_repos y lo desplegaremos, aquí estará nuestro tema de interés



1. Presionamos los 3 puntos y seleccionaremos consulta

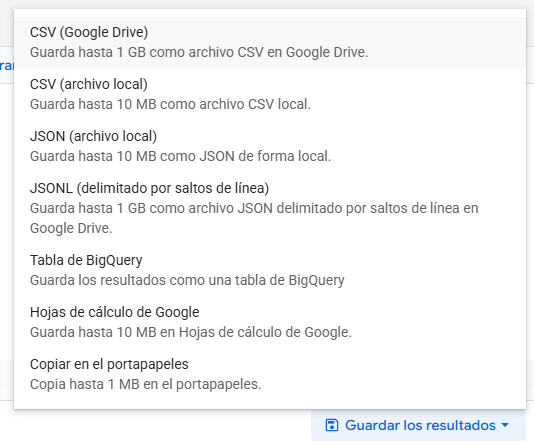


1. Esto nos enviara a una hoja de consultas, donde pegaremos la siguiente consulta, se recomienda aplanar la información para poder guardar nuestra consulta después.



En la parte inferior visualizaremos nuestra consulta

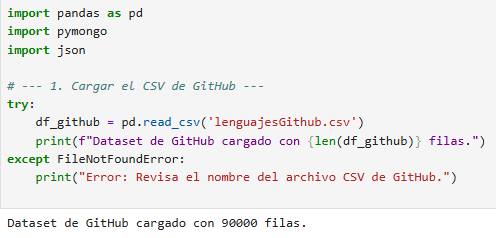
1. En la parte derecha, encontraremos el botón de Guardar los resultados, de preferencia se recomienda hacerlo en csv de OneDrive, ya que nuestra consulta es pesada



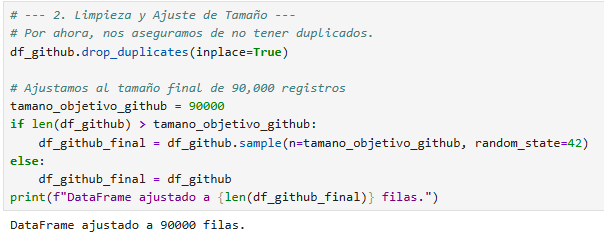
Al guardarla, podremos pasarla a jupyter para hacer su respectiva limpieza

***LIMPIEZA E IMPORTACION A MONGODB ATLAS***

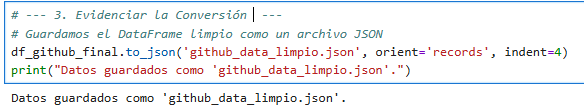
1. Cargaremos el csv previamente guardado



1. Realizaremos la respectiva limpieza de duplicados, ajustar el tamaño nos sirve para eliminar datos extra que no necesitamos o no tomaremos en cuenta



1. Convertimos nuestro dataframe ya limpio a JSON



1. El siguiente código nos ayudará a exportar nuestro JSON a mongo atlas

MONGO\_URI = "mongodb+srv://juan:12345@cluster0.agcdm8h.mongodb.net/"

try:

client = pymongo.MongoClient(MONGO\_URI)

client.admin.command('ping')

print("¡Conexión a MongoDB Atlas exitosa!")

db = client['proyecto\_futbol']

collection = db['repositorios\_github'] # creacmos una nueva coleccion

# Convertimos el DataFrame a un formato compatible con Mongo

datos\_para\_mongo = df\_github\_final.to\_dict(orient='records')

print(f"Limpiando la colección '{collection.name}'...")

collection.delete\_many({})

# Insertamos los datos en lotes

batch\_size = 10000

print(f"Insertando {len(datos\_para\_mongo)} documentos en la nueva colección...")

for i in range(0, len(datos\_para\_mongo), batch\_size):

batch = datos\_para\_mongo[i:i + batch\_size]

collection.insert\_many(batch)

print(f" -> Lote {i//batch\_size + 1} insertado.")

print("¡Carga de datos de GitHub completada!")

except Exception as e:

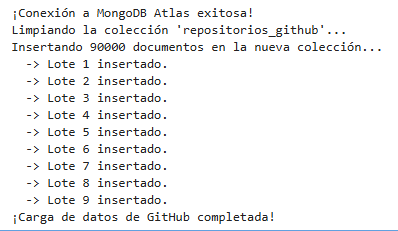
print(f"Error durante el proceso de carga a MongoDB: {e}")

finally:

if 'client' in locals():

client.close()

El resultado debería ser algo así:



Al igual que el dataset de deportes, este también lo hicimos mediante lotes, para no perder la conexión con mongo atlas