Codes and screenshot of cleaning dataset using spark with the help of dataproc(\* done in gcs with sdk)

# cleaning code using spark

from pyspark.sql import SparkSession

from pyspark.sql.functions import trim, col

def main():

    spark = SparkSession.builder.appName("CleanDataset").getOrCreate()

    # Input and output paths - update file name if needed

    parquet\_output\_path = "gs://dataset\_nutrifuson/cleaned-data/cleaned\_dataset"

    csv\_output\_path = "gs://dataset\_nutrifusion/cleaned-data/csv\_output"

    input\_path = "gs://dataset\_nutrifusion/input-data/70000\_recipes.csv"

"

    df = spark.read.option("header", "true").option("inferSchema", "true").csv(input\_path)

    # Trim whitespace from string columns

    string\_cols = [f.name for f in df.schema.fields if f.dataType.simpleString() == "string"]

    for c in string\_cols:

        df = df.withColumn(c, trim(col(c)))

    # Drop rows with any nulls

    df\_clean = df.na.drop()

    # Drop duplicates

    df\_clean = df\_clean.dropDuplicates()

    # Save cleaned dataset as Parquet

    df\_clean.write.mode("overwrite").parquet(parquet\_output\_path)

    # Also save cleaned dataset as CSV (with header)

    df\_clean.write.mode("overwrite").option("header", "true").csv(csv\_output\_path)

    print(f"Cleaned data saved as Parquet to: {parquet\_output\_path}")

    print(f"Cleaned data saved as CSV to: {csv\_output\_path}")

    spark.stop()

if \_\_name\_\_ == "\_\_main\_\_":

    main()

"""1st created a project in gcs as Nutrifusion\_project

then created a bucket named dataset\_nutrifusion

"""

# authenticate and onnect set the project using this

gcloud auth login

gcloud config set project nutrifusion-drive-2025

#set the region

gcloud config set dataproc/region us-central1

#checked the cluster status

gcloud dataproc clusters list

#this was my specific cluster

gcloud dataproc clusters describe recipe-cluster --region=us-central1

# uploaded dataset the cleanig code to gcs

gsutil cp "D:\Downloads\sparkjob\_2.py" gs://dataset\_nutrifusion/scripts/

#uploaded the dataset to the gcs using sdk

gsutil cp D:\OneDrive\Documents\DSMM\TERM 3\capstone [project\70000\_recipes.csv gs://dataset\_nutrifusion/input-data/

# give me list of files in cluster

gsutil ls gs://dataset\_nutrifusion/cleaned-data/

#sparkjob in dataproc

gcloud dataproc jobs submit pyspark D:\Downloads\sparkjob\_2.py --cluster=recipe-cluster --region=us-central1

Job [4c2fa67d5c664c50a9ff022de222e5a1] submitted.

#code to convert the parquet to csv

from pyspark.sql import SparkSession

# Step 1: Start Spark session

spark = SparkSession.builder.appName("ParquetToCSV").getOrCreate()

# Step 2: Read Parquet data from GCS

df = spark.read.parquet("gs://dataset\_nutrifusion/cleaned-data/cleaned\_dataset/")

# Step 3: Convert to CSV and save back to GCS (single file)

df.coalesce(1).write.csv("gs://dataset\_nutrifusion/cleaned-data/csv\_output", header=True, mode="overwrite")

#parquet in dataproc

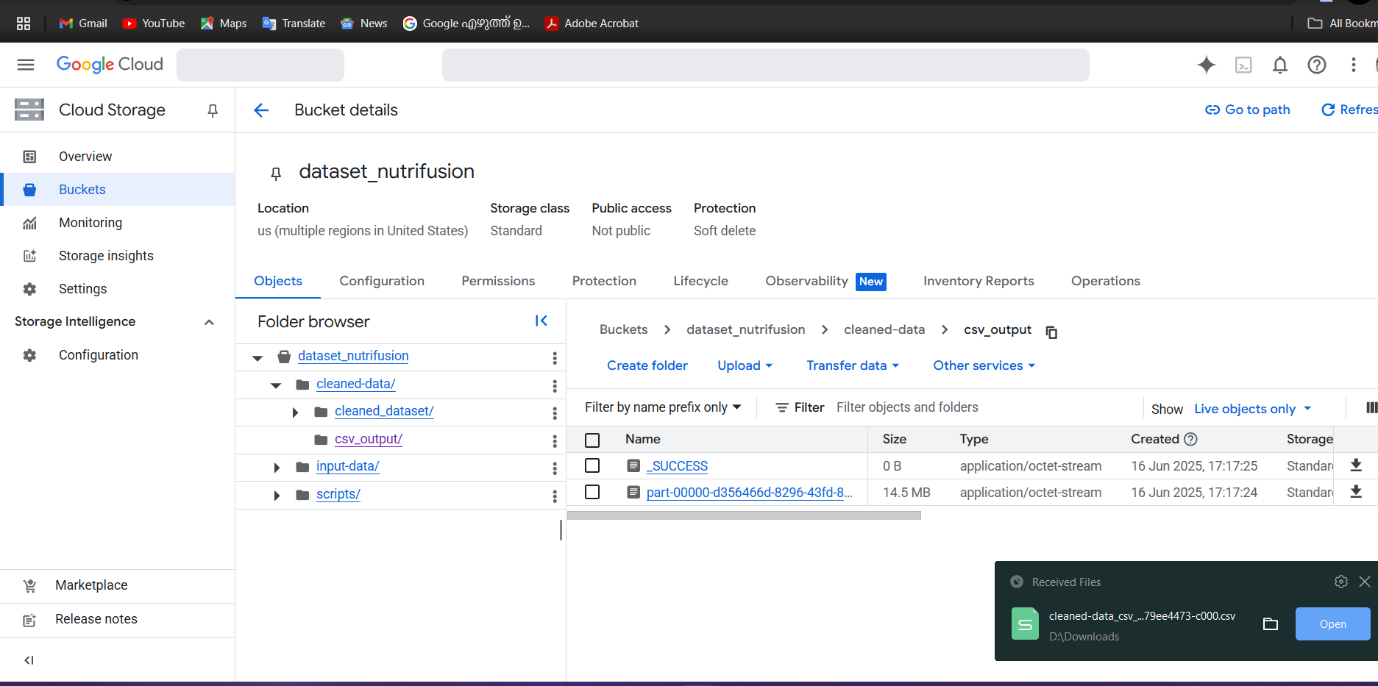
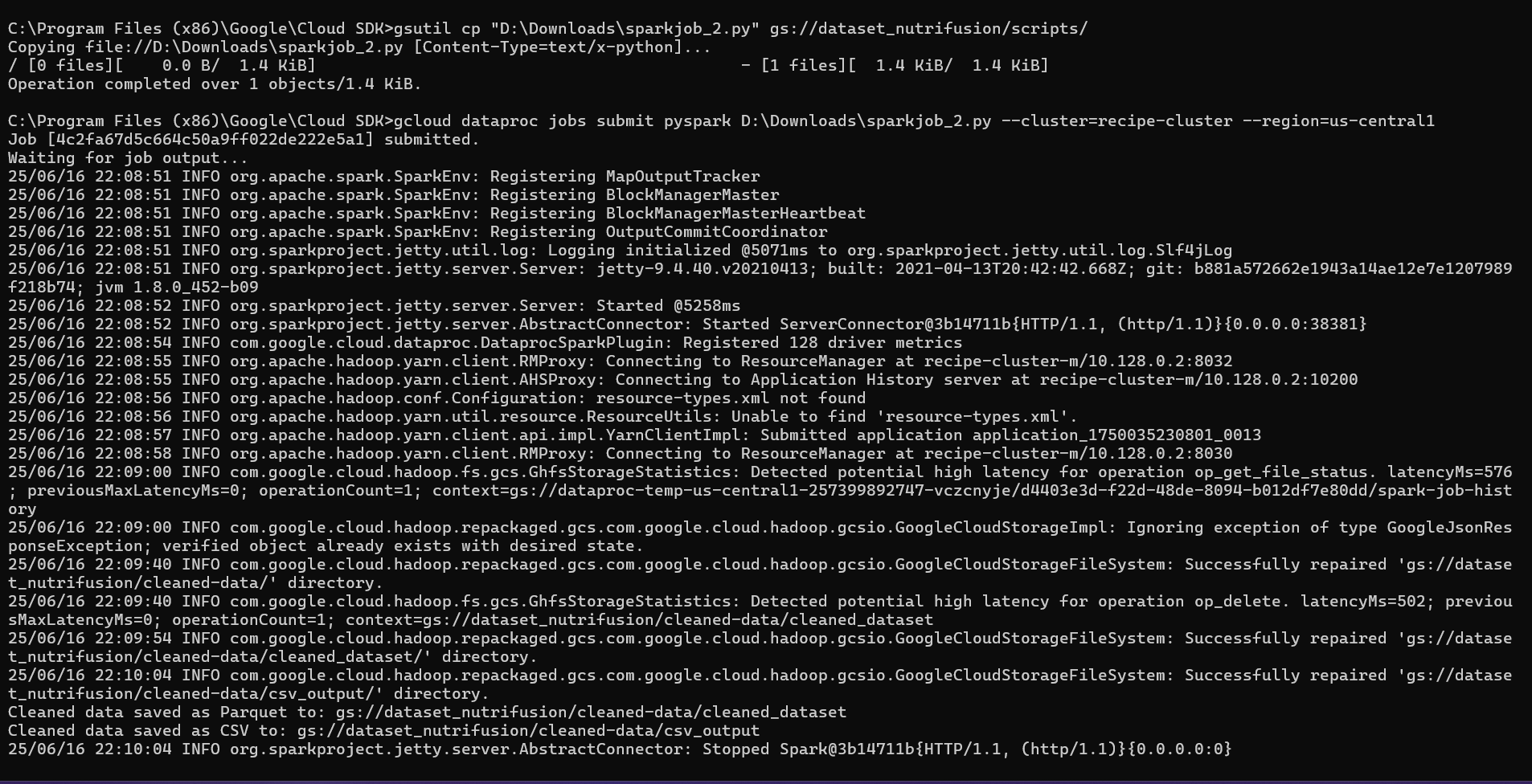
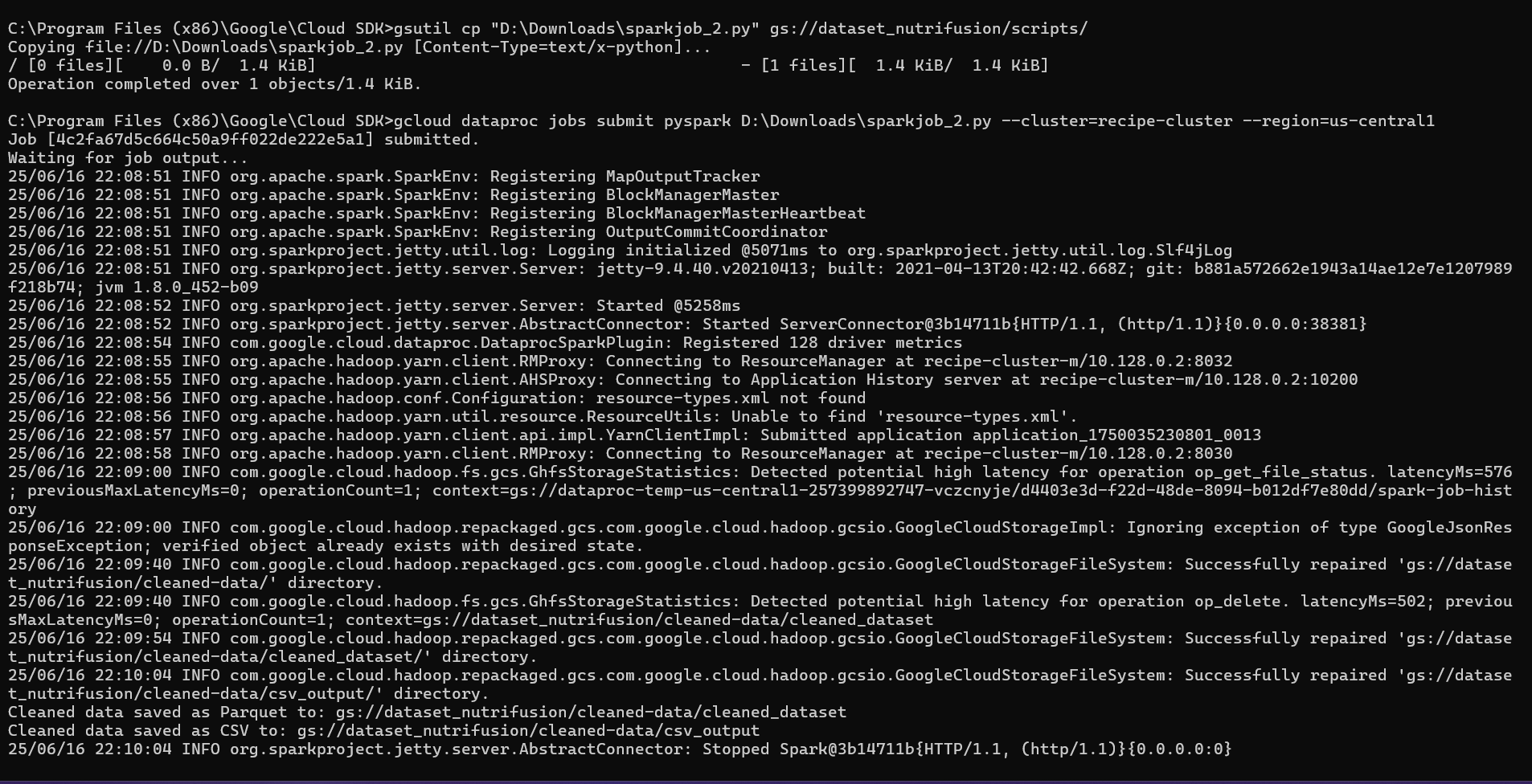
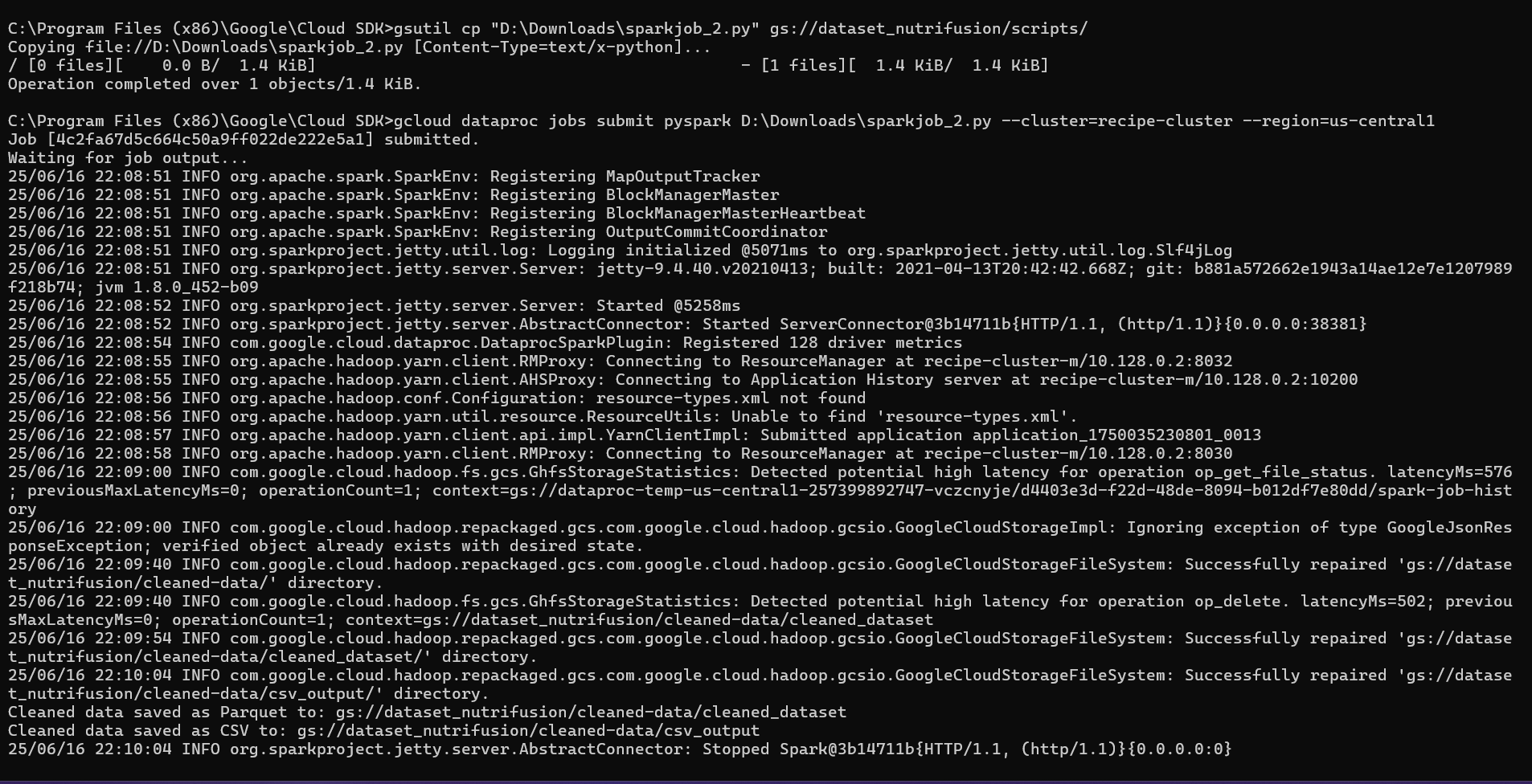
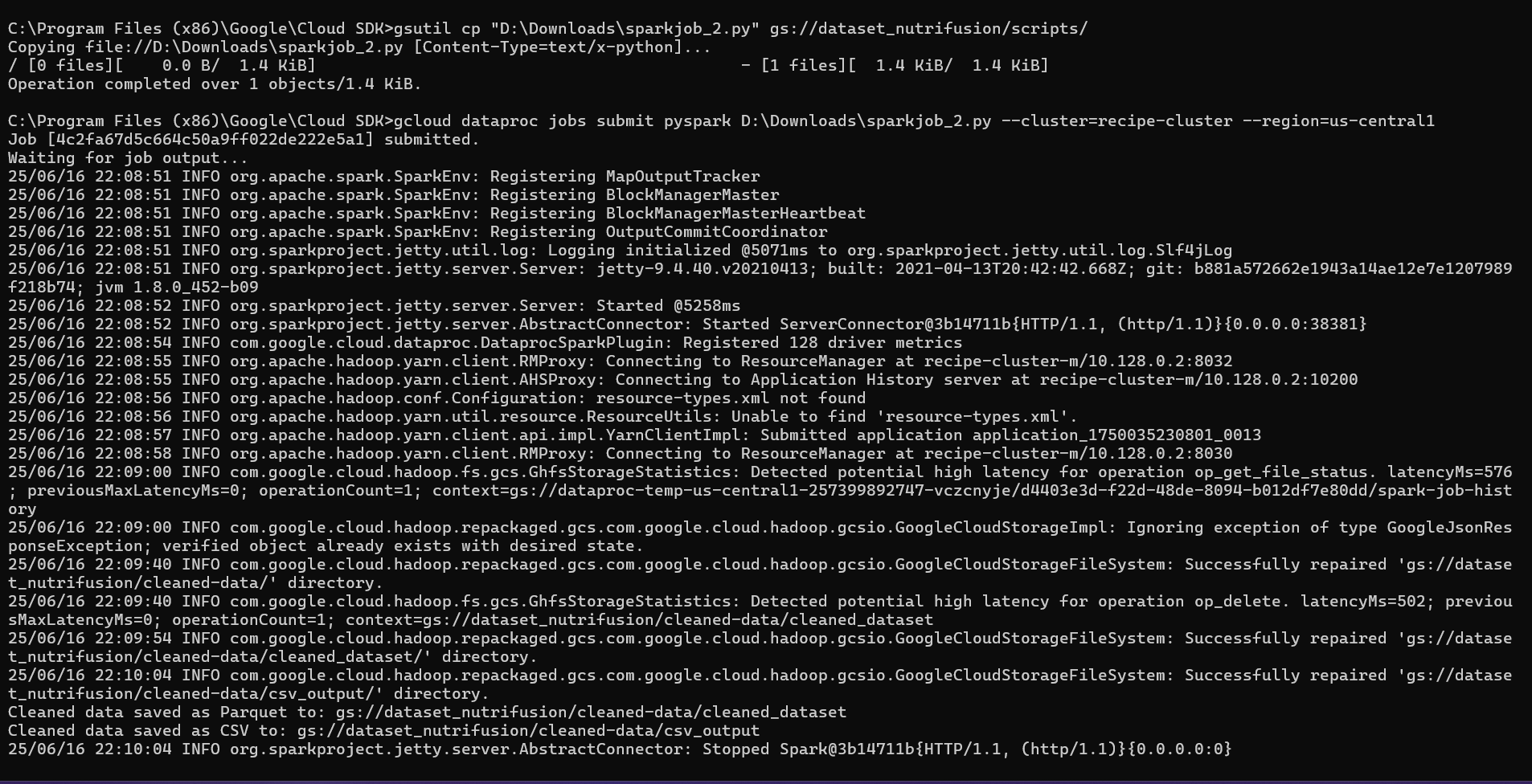
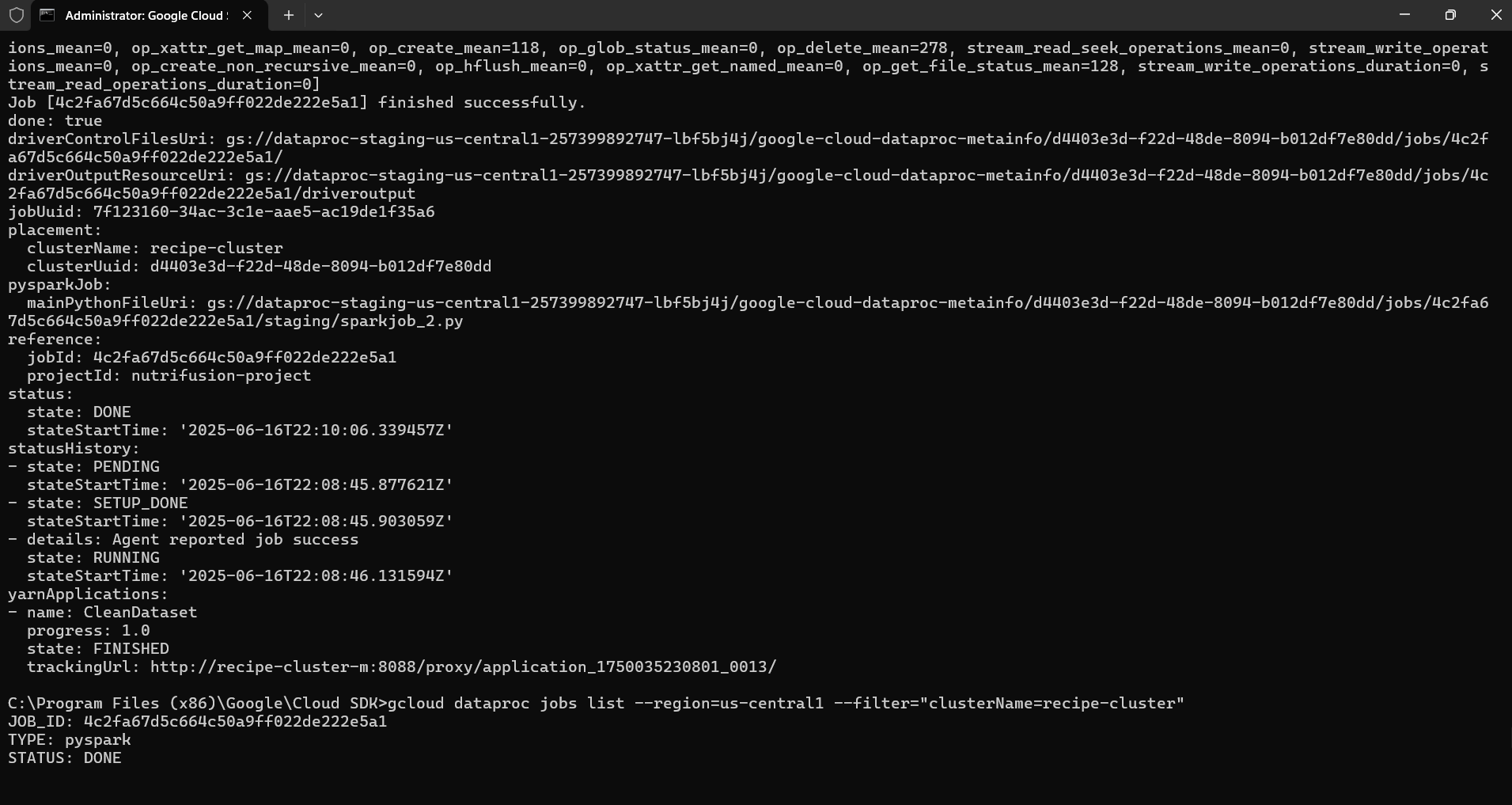
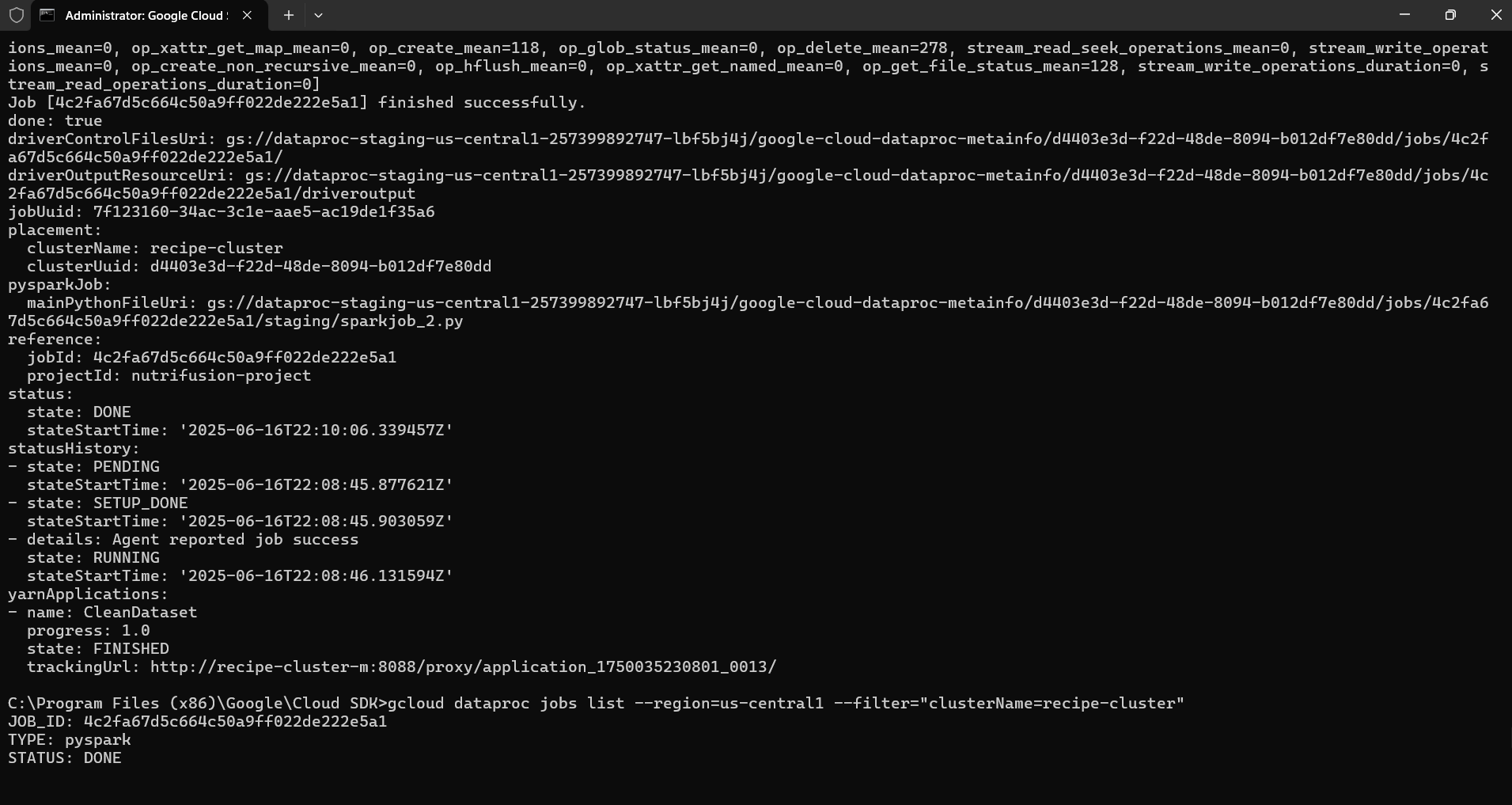
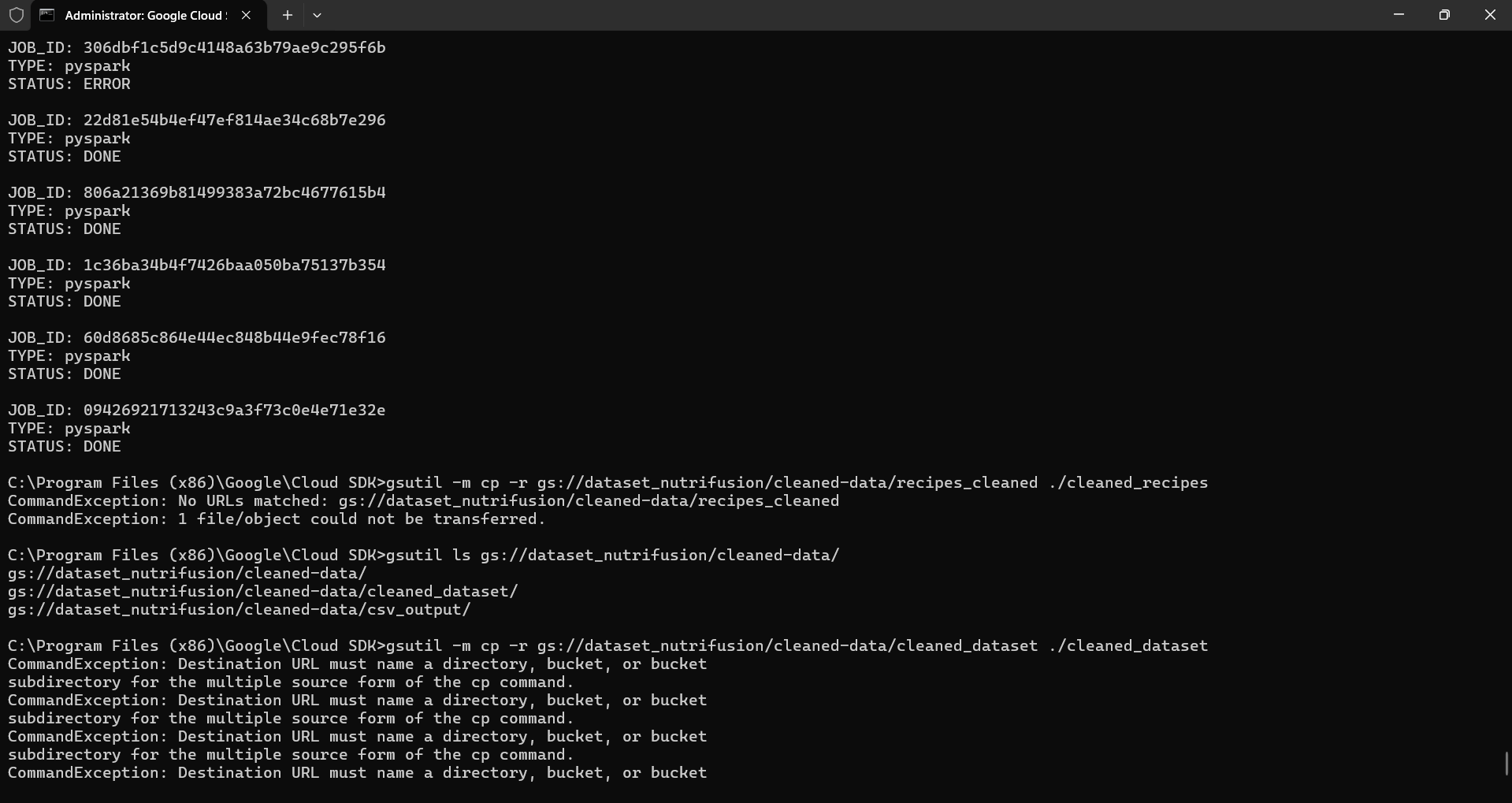
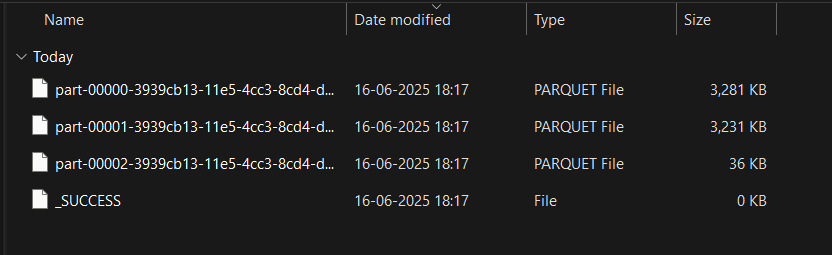
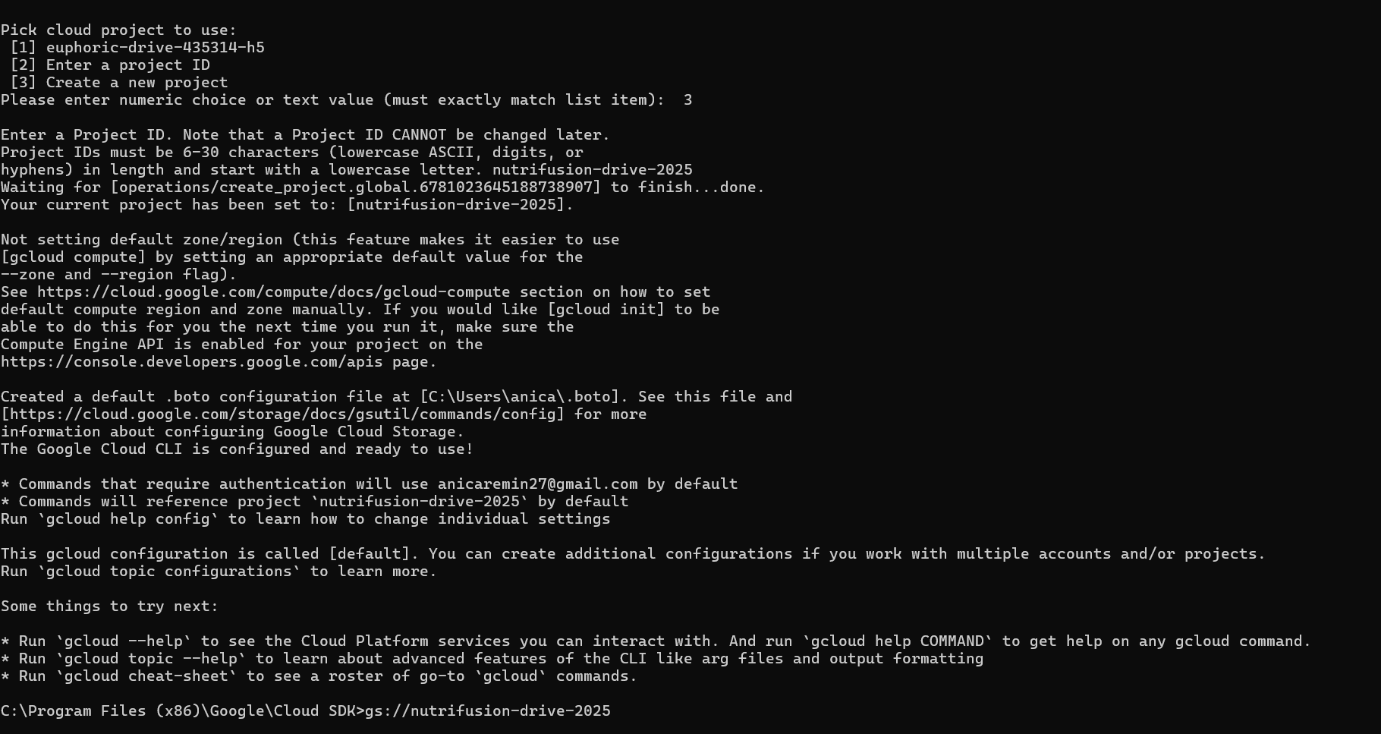
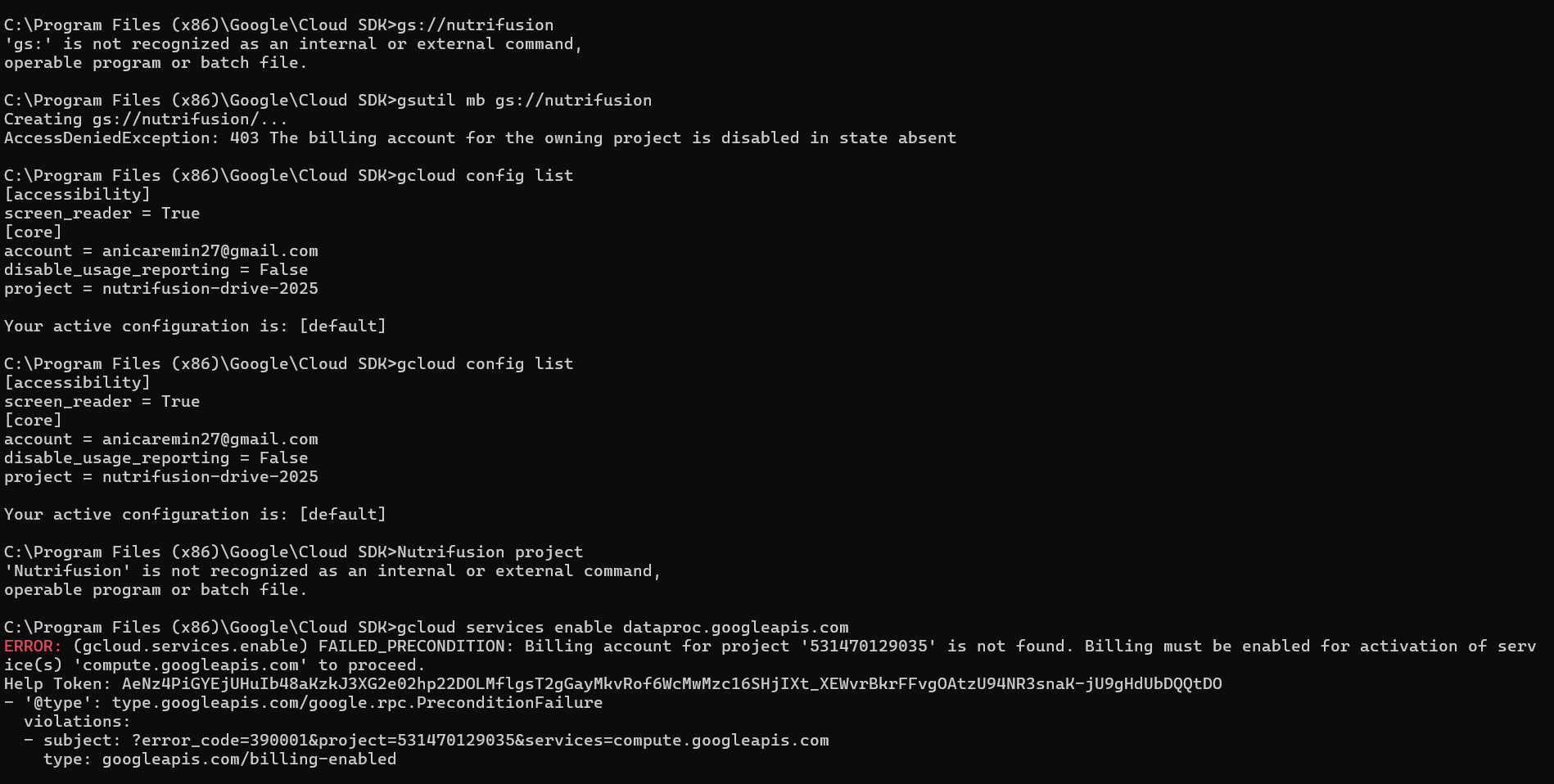
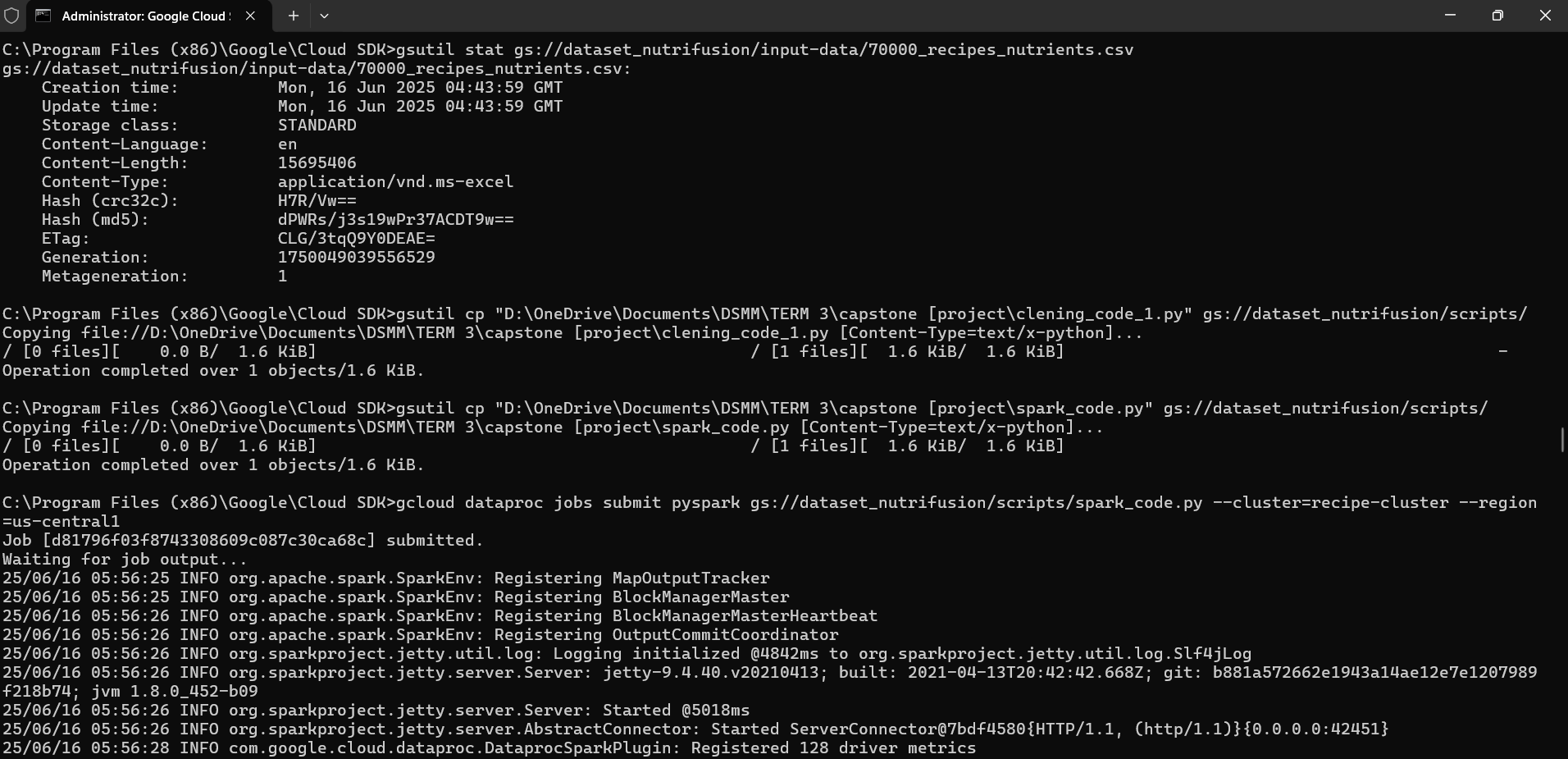
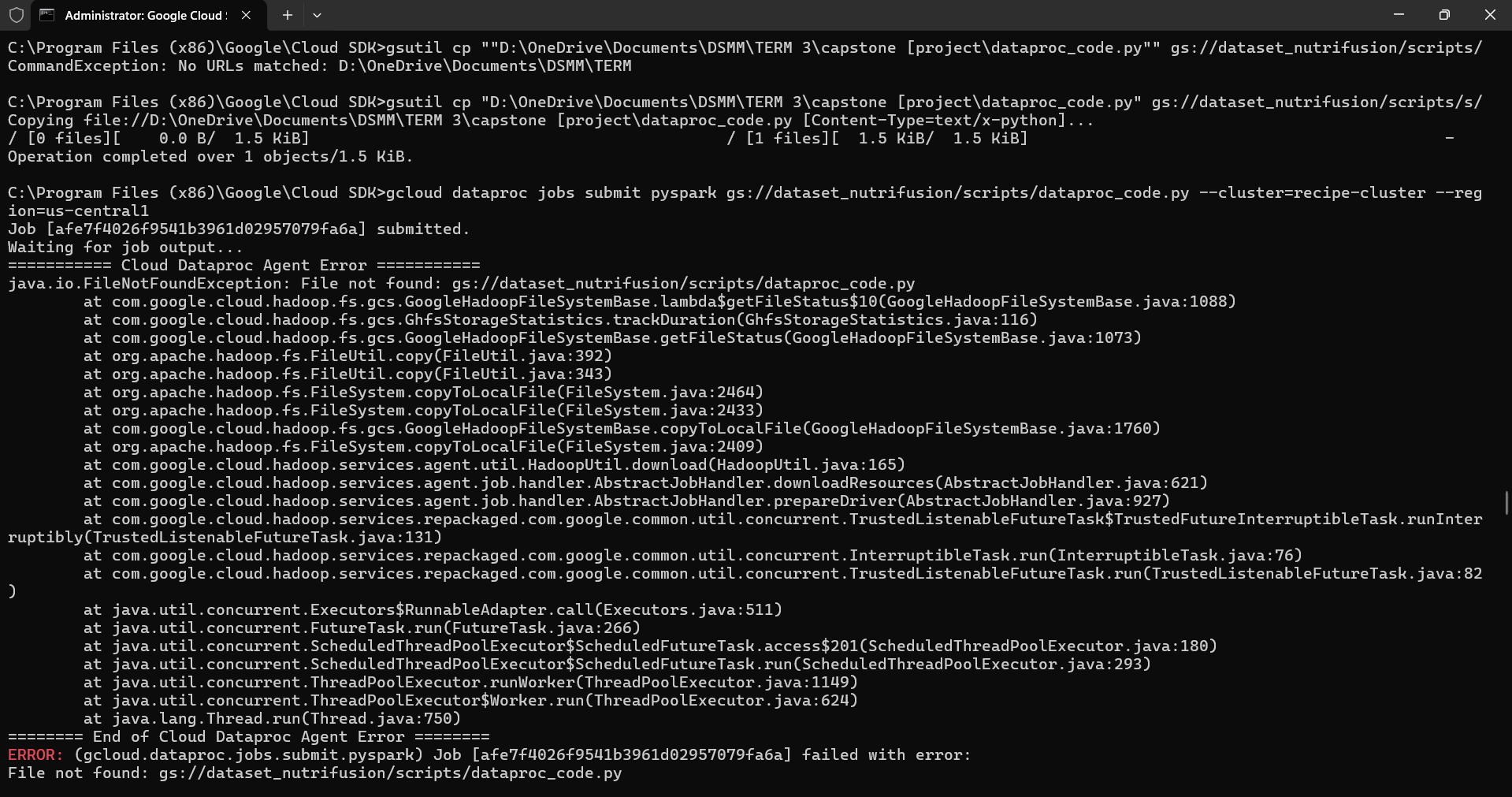
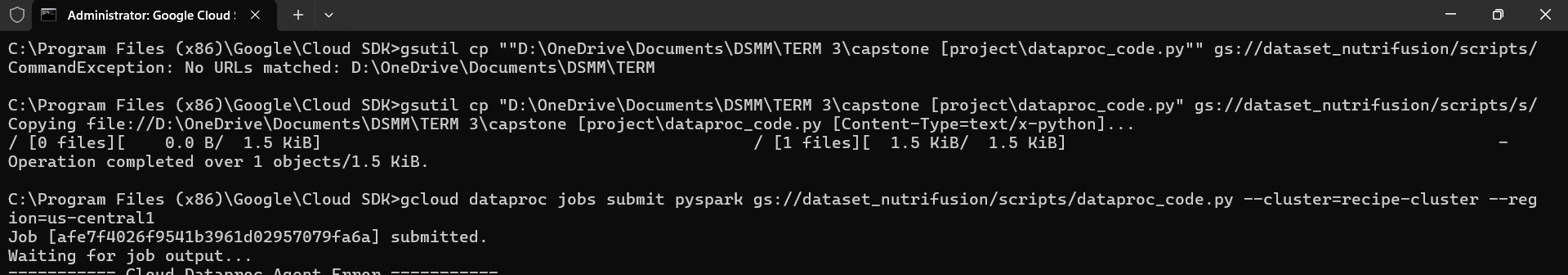
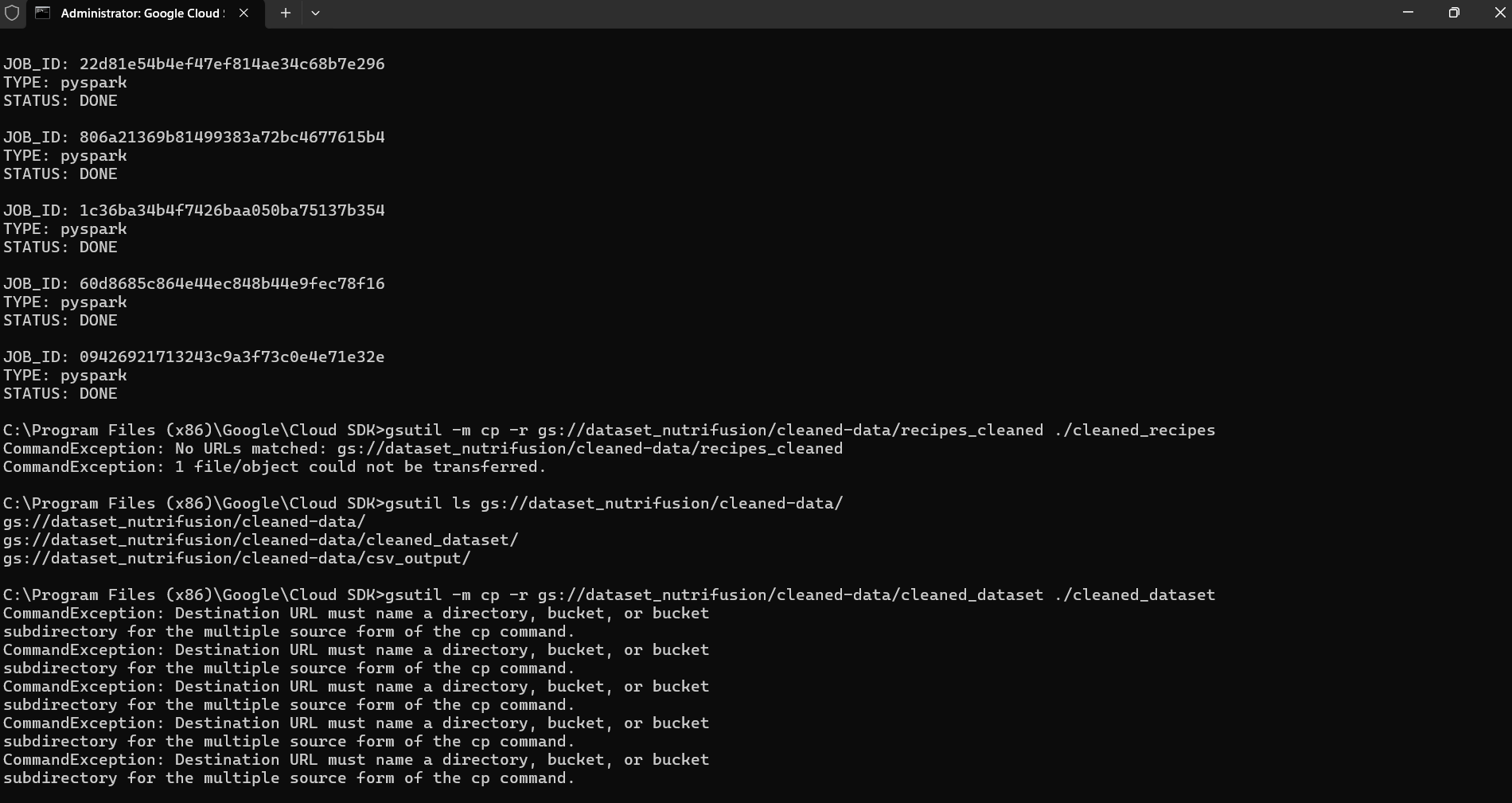
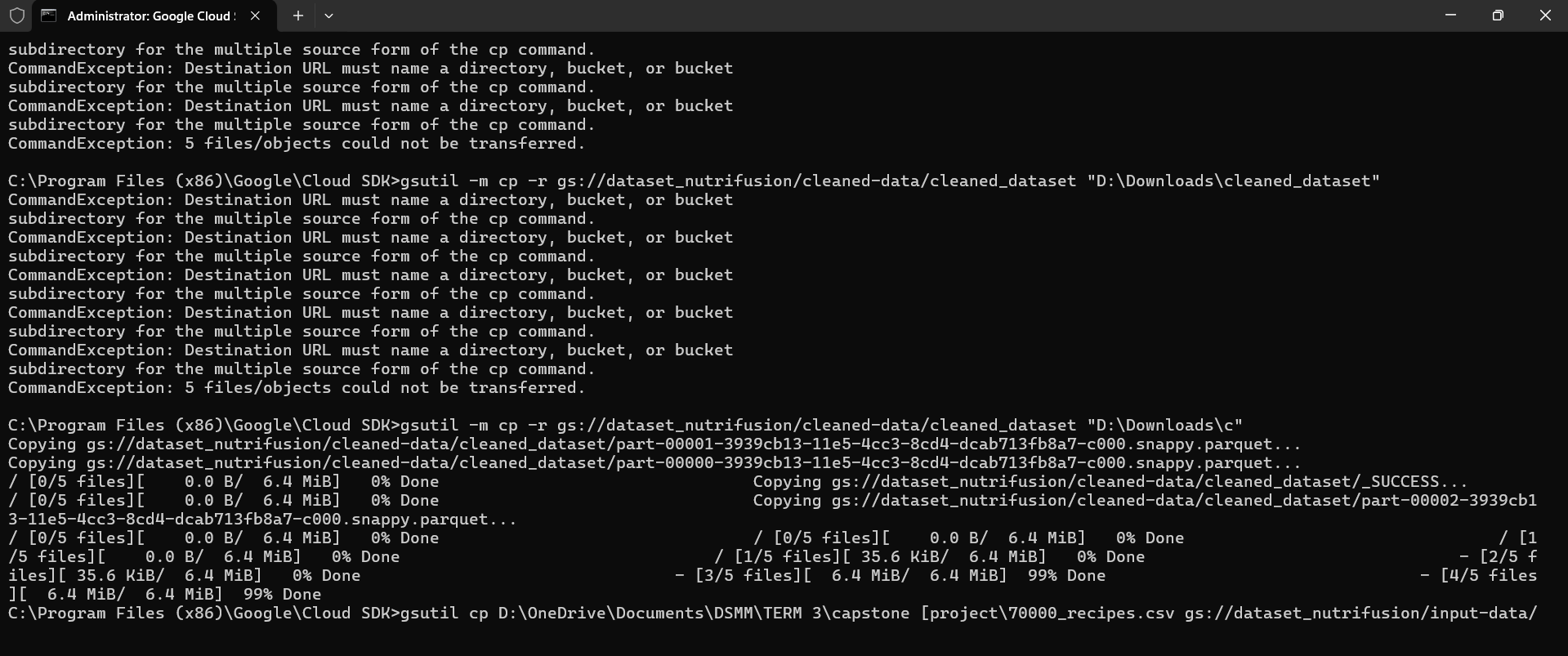
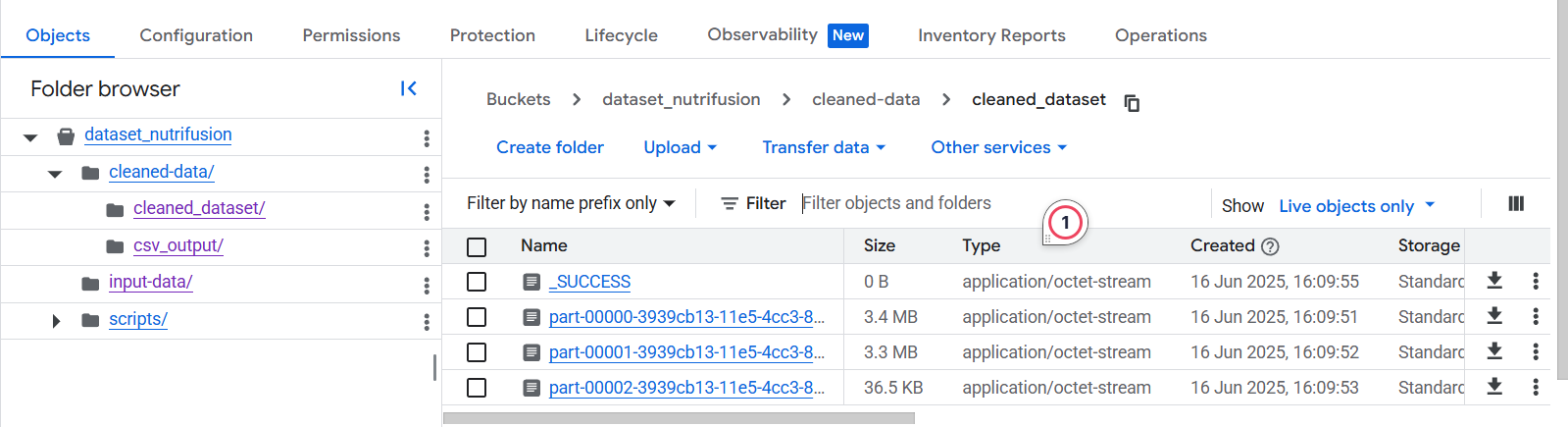
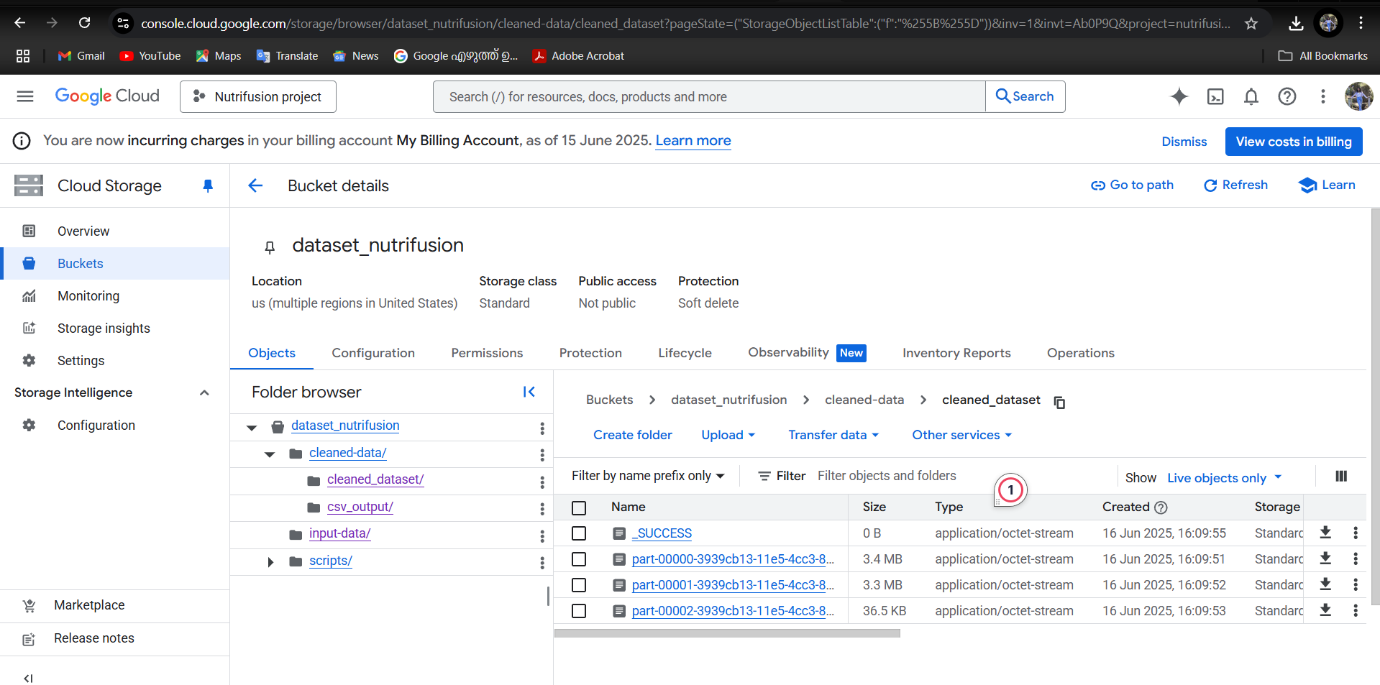
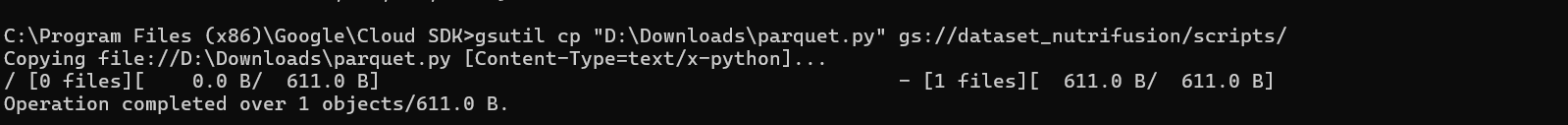
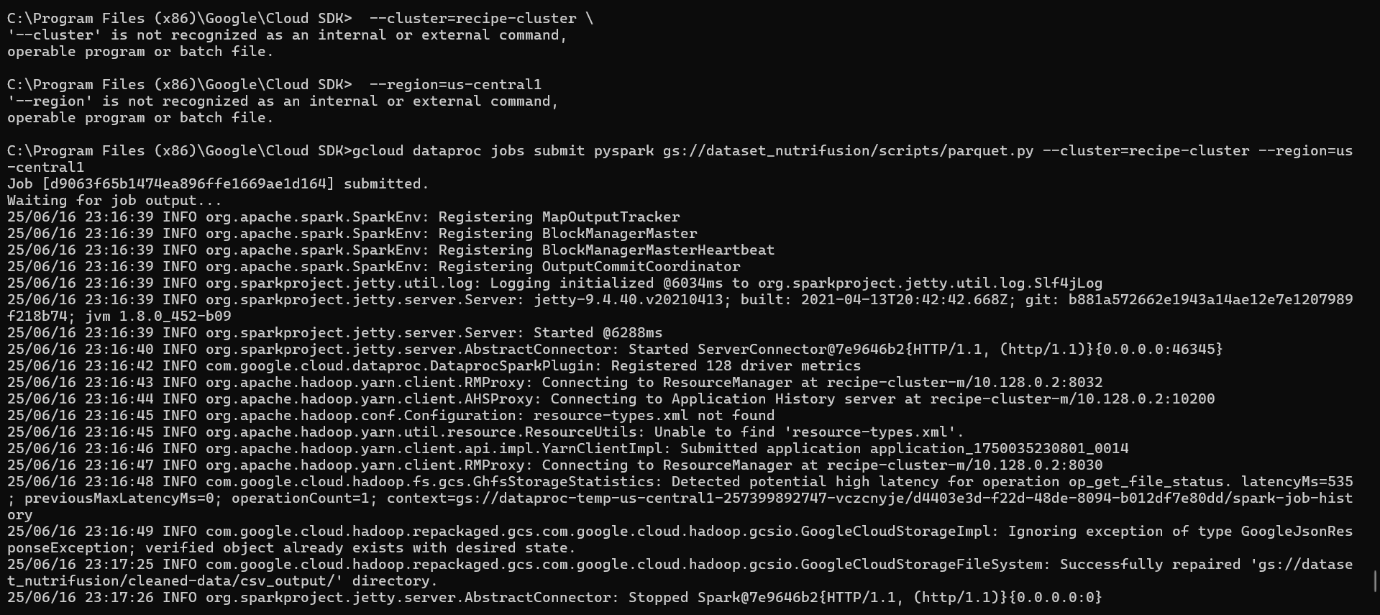
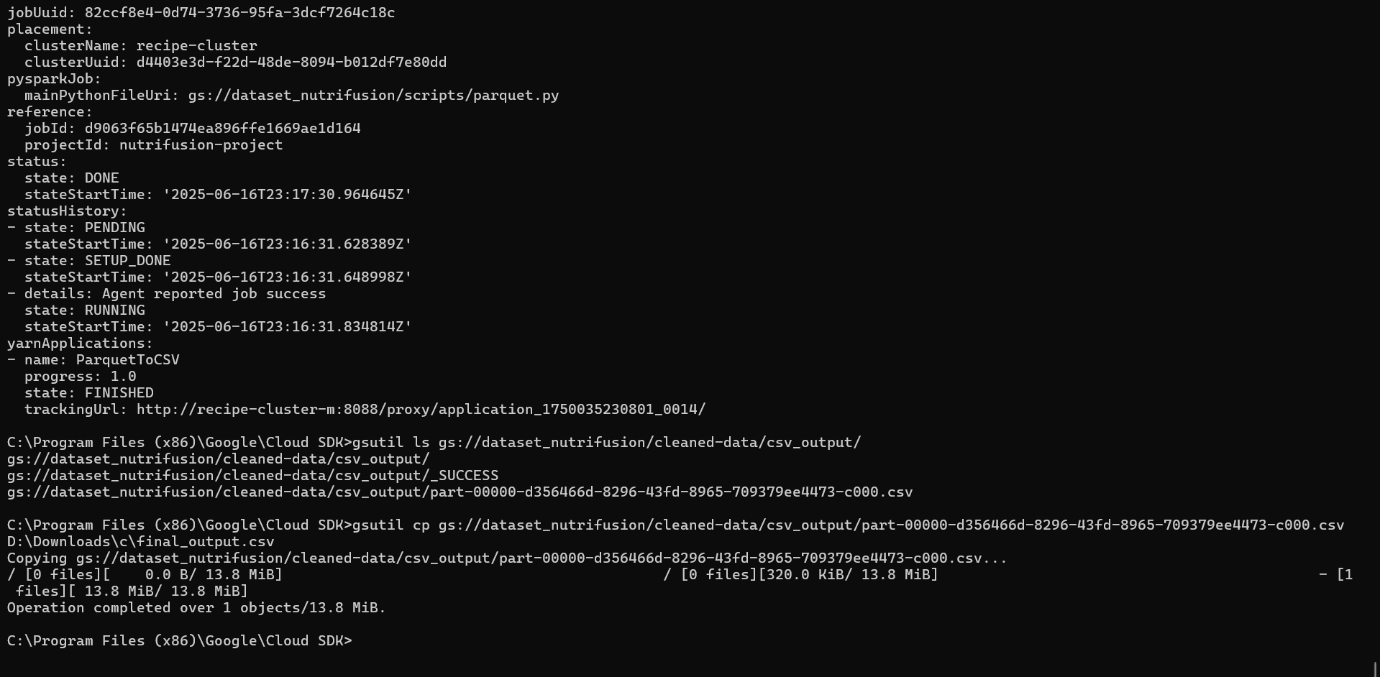
gcloud dataproc jobs submit pyspark gs://dataset\_nutrifusion/scripts/parquet.py --cluster=recipe-cluster --region=us-central1

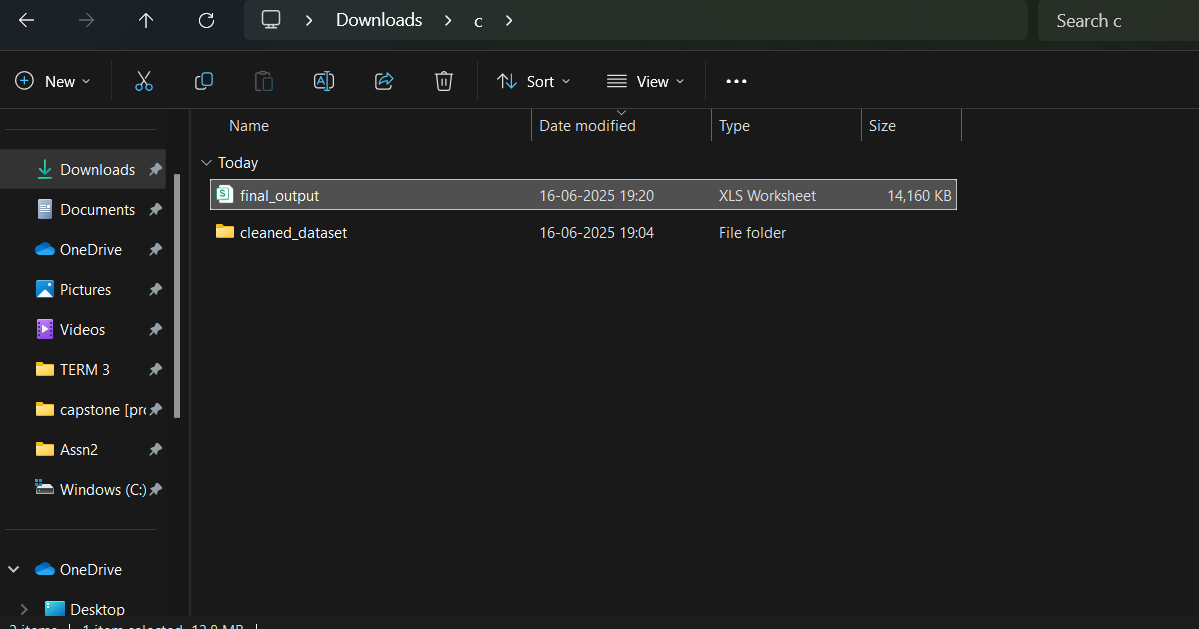
# list of datasets after runing the above

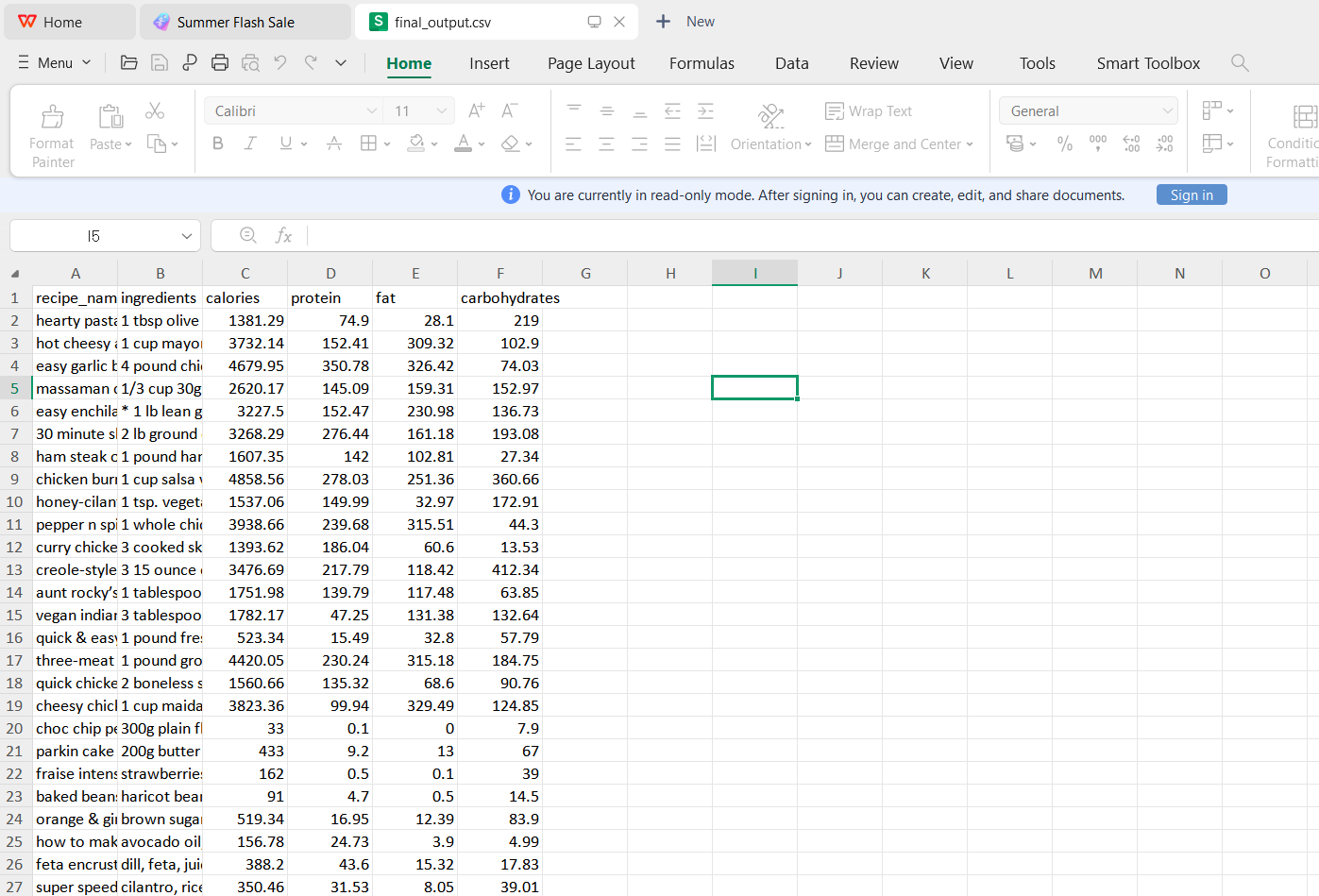
gsutil ls gs://dataset\_nutrifusion/cleaned-data/csv\_output/

#code to download the csv file

gsutil cp gs://dataset\_nutrifusion/cleaned-data/csv\_output/part-00000-d356466d-8296-43fd-8965-709379ee4473-c000.csv







**Summary of Cleaning Applied**

| **Step** | **Description** |
| --- | --- |
| Trim Whitespace | Removes leading/trailing spaces from strings |
| Drop Nulls | Deletes any row that has missing (null) values |
| Drop Duplicates | Deletes fully duplicated rows |