**BWT Task-10 Exercise**

**Submitted By: Abduhu Khan**

**PySpark:**

**Code & Output:**

from pyspark.sql import SparkSession

from pyspark.sql.functions import col, to\_date, udf, hour, lower, to\_timestamp

from pyspark.sql.types import StringType

spark = SparkSession.builder \

    .appName("Data Cleaning") \

    .getOrCreate()

df = spark.read.csv('/content/drive/MyDrive/data.csv', header=True, inferSchema=True, quote='"', escape='"')

print("Initial Data:")

df.show(5)

# DATA Cleaning operations

# 1. Remove orders placed between 12am and 5am and convert timestamp to date

df = df.withColumn("order\_date", to\_timestamp(col("order\_date")))

# Apply the hour filter to exclude times between 0 and 5 AM

df = df.filter(~hour(col("order\_date")).between(0, 5))

# 2. Adding new column time\_of\_day

def get\_time\_of\_day(hour):

    if 5 <= hour < 12:

        return 'morning'

    elif 12 <= hour < 18:

        return 'afternoon'

    elif 18 <= hour < 24:

        return 'evening'

    else:

        return 'night'

time\_of\_day\_udf = udf(get\_time\_of\_day, StringType())

df = df.withColumn('time\_of\_day', time\_of\_day\_udf(hour(col('order\_date'))))

# 3. Remove rows containing "TV" in the product column

df = df.filter(~df.product.contains("TV"))

# 4. Ensure all product categories are in lowercase

df = df.withColumn("category", lower(col("category")))

# 5. Adding new column purchase\_state

def extract\_state\_and\_zip(address):

    parts = address.split(',')

    if len(parts) >= 3:

        return parts[-1].strip()

    else:

        return None

extract\_state\_and\_zip\_udf = udf(extract\_state\_and\_zip, StringType())

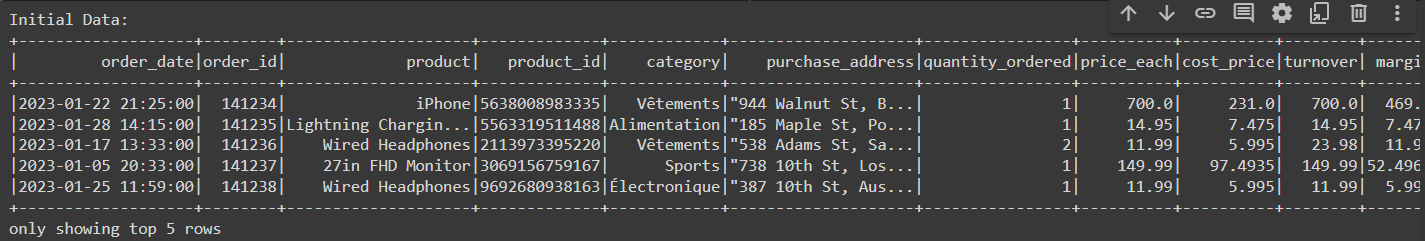
df = df.withColumn('purchase\_state', extract\_state\_and\_zip\_udf(col('purchase\_address')))

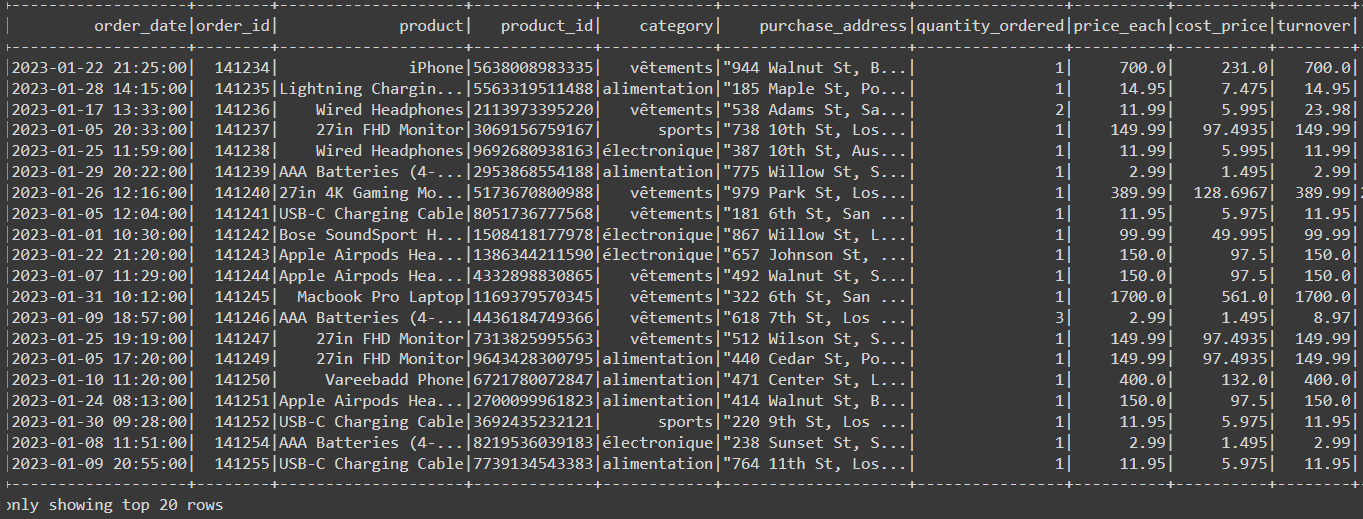
# Parquet file

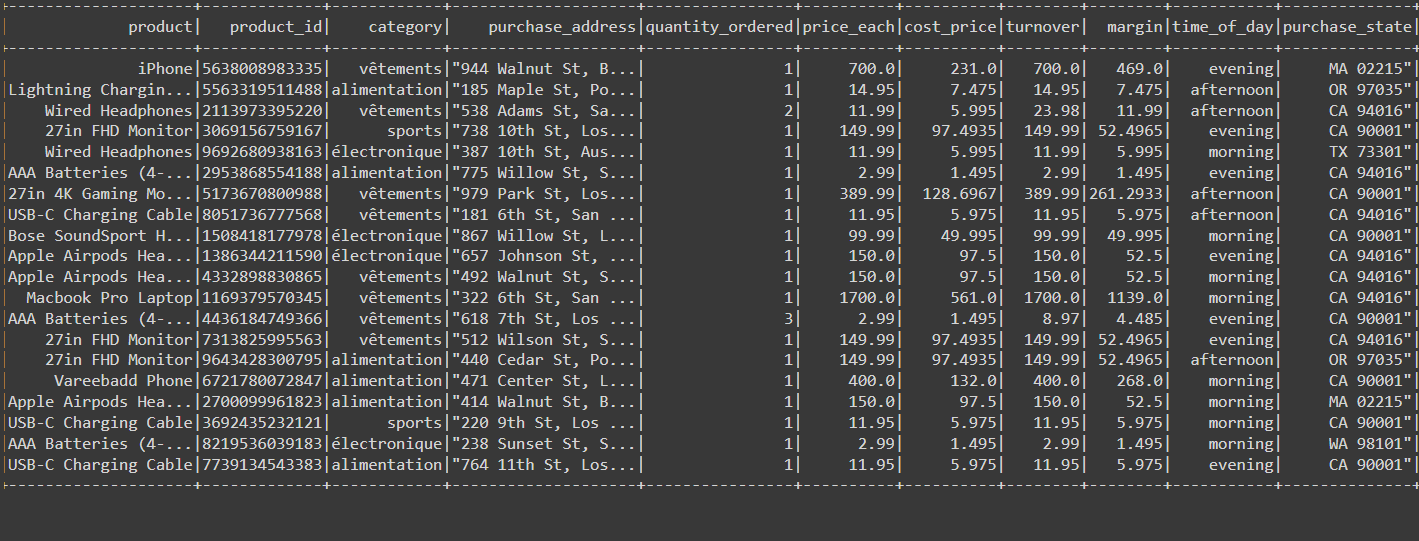
df.write.parquet('/content/drive/MyDrive/data.parquet')

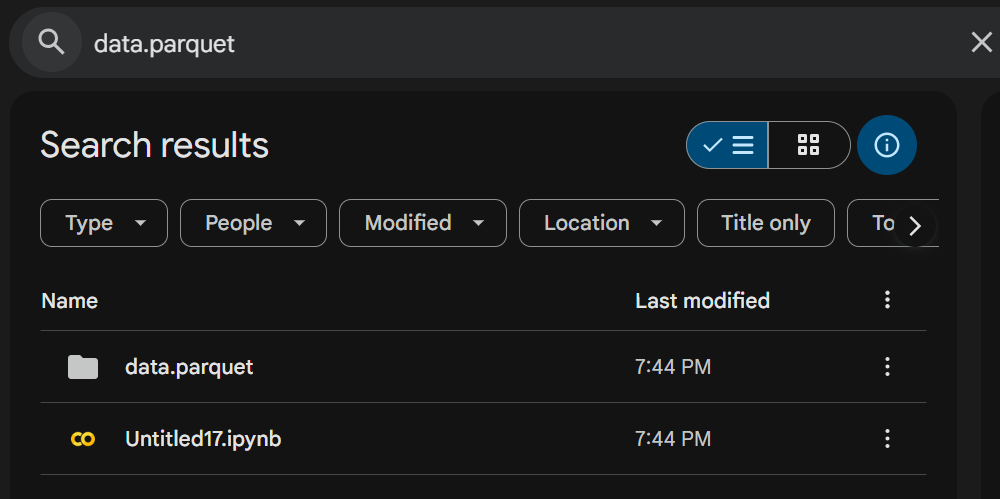
df.show()

spark.stop()









**Notebook Link:**

<https://colab.research.google.com/drive/1P02o2MJ3QCyQc1BKt9nfUa3pry2YS_HE#scrollTo=-PgsyoTsvQzg>