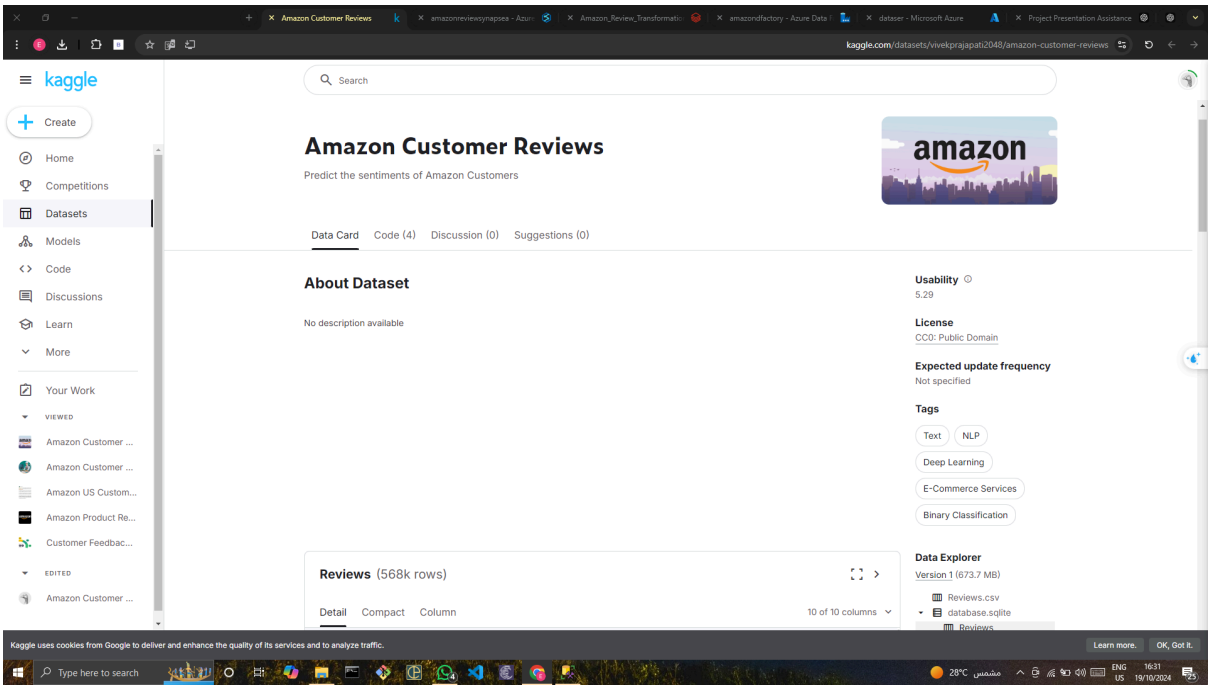
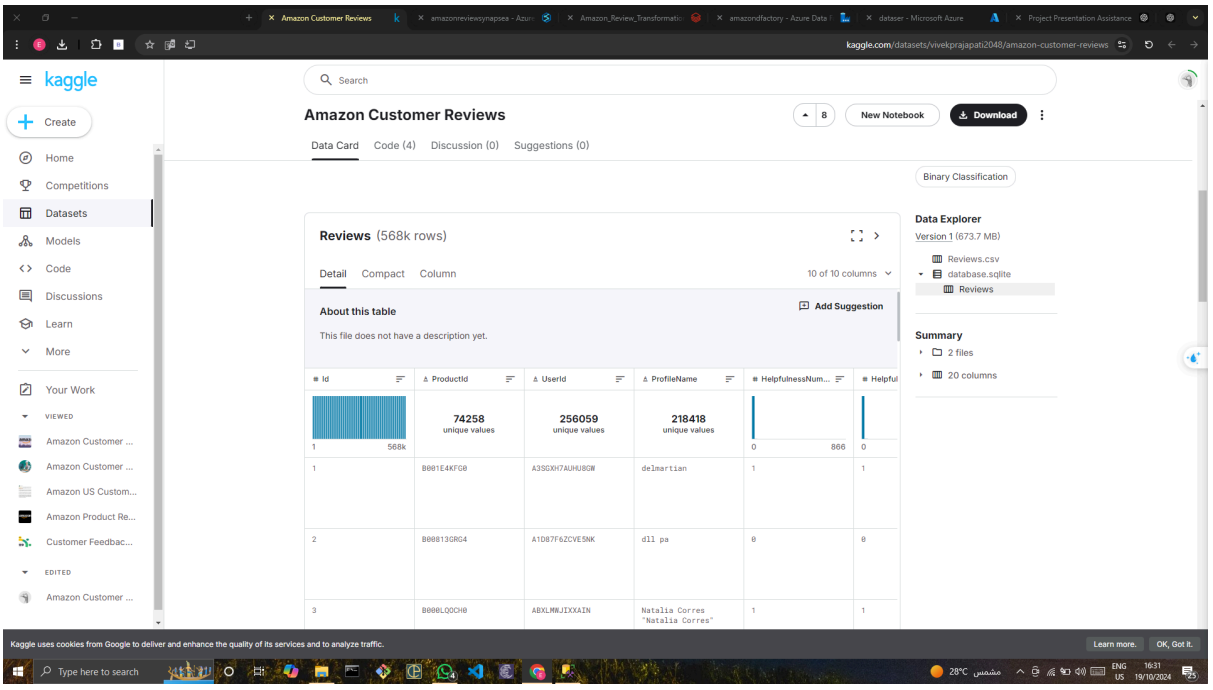
My part in the project : Mohamed Yaser karam abd Alshafy

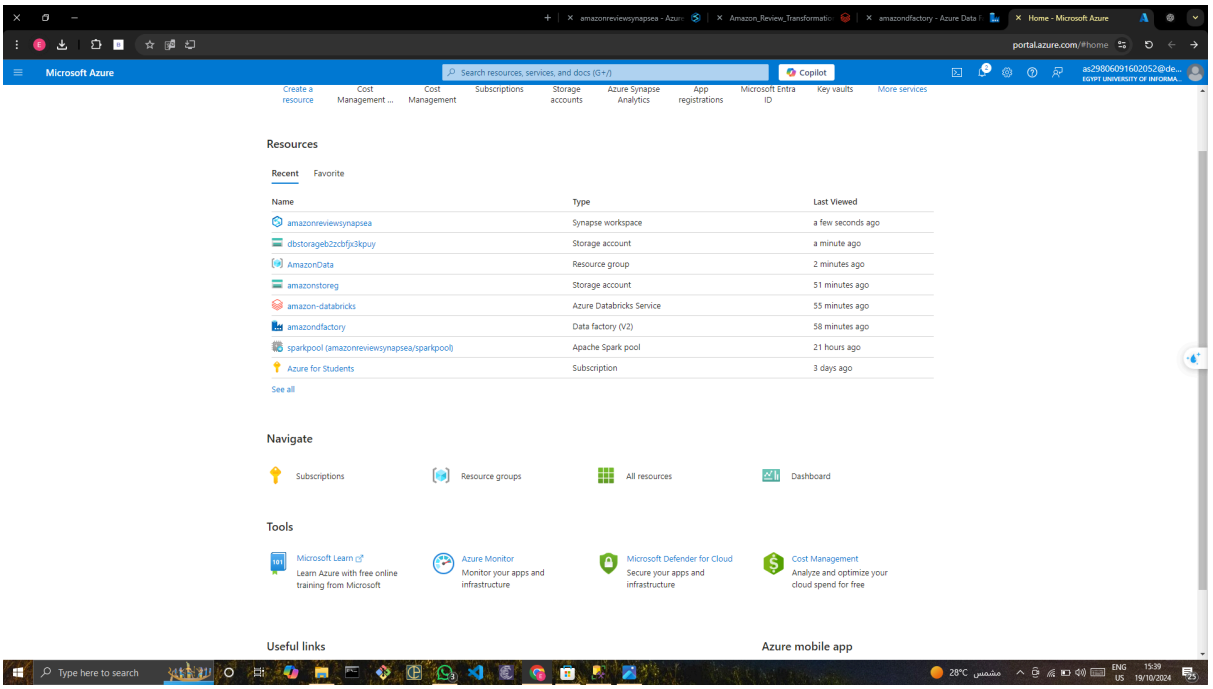
Project title: Customer Feedback Analysis and Improvement

The data set set that I used

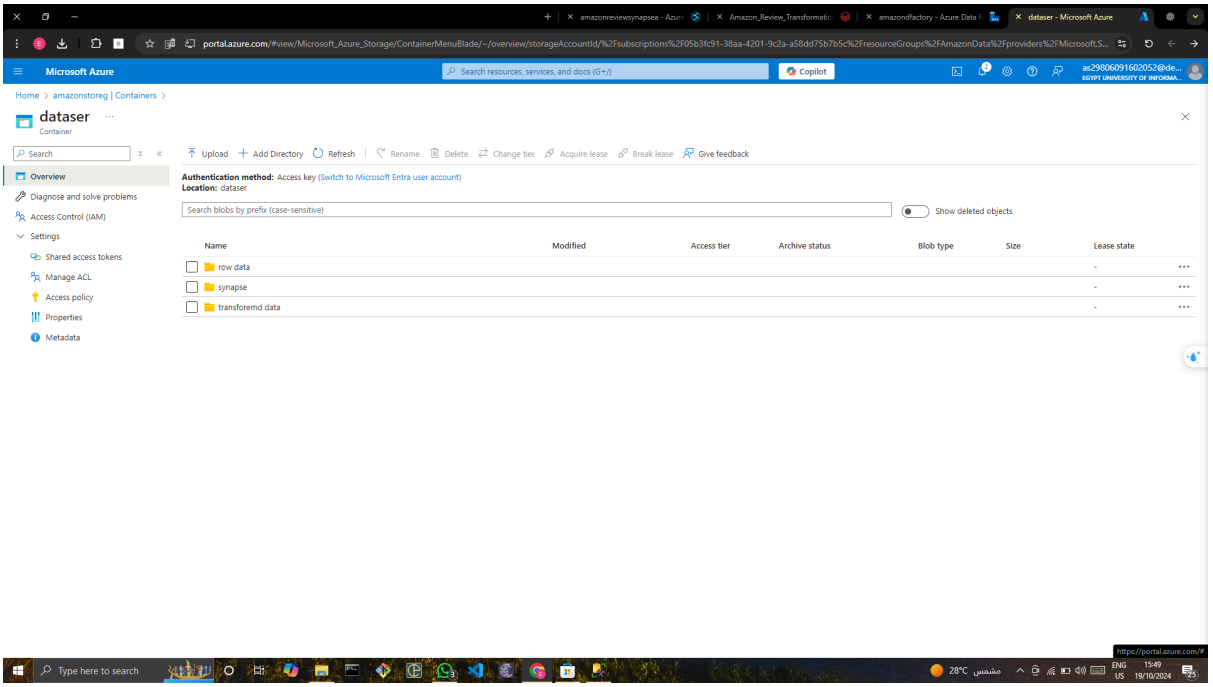




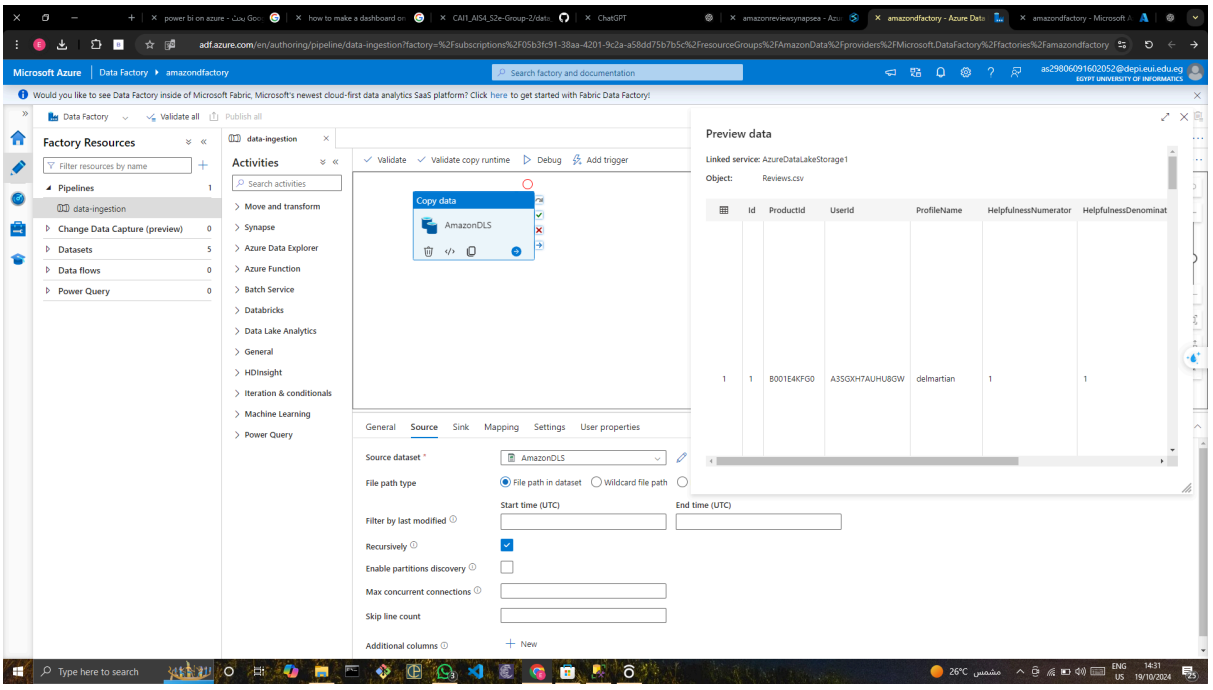
The resources that I made on Azure

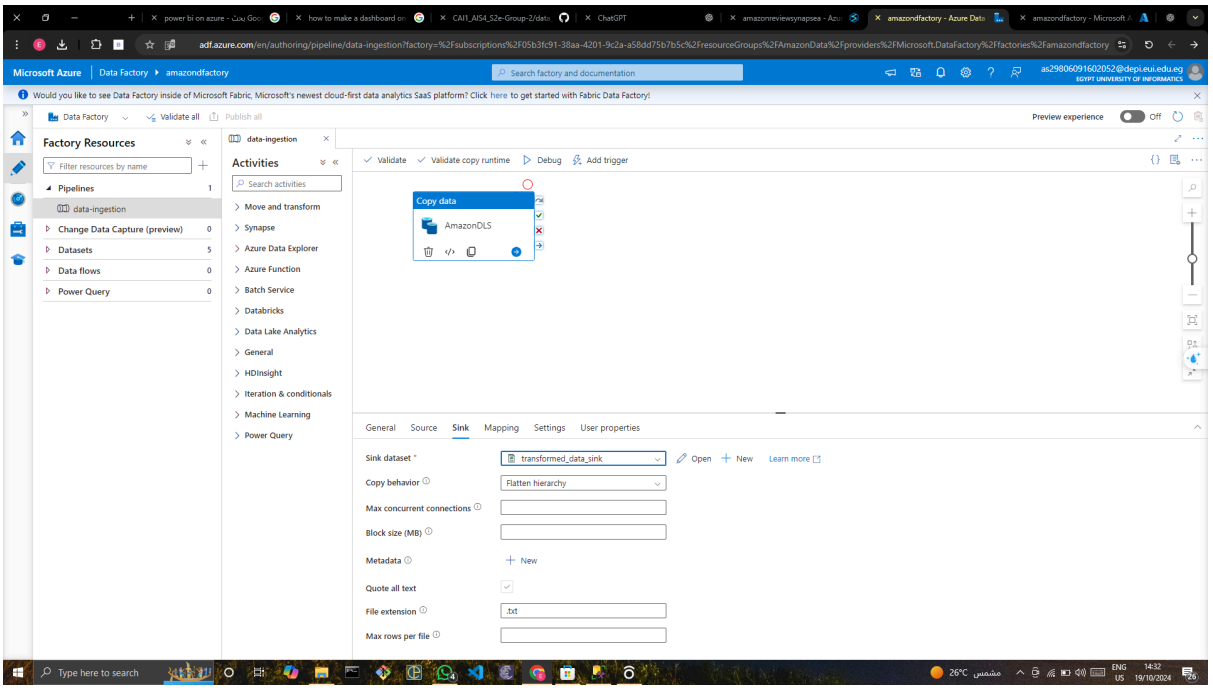


The directories that I made inside the data lake

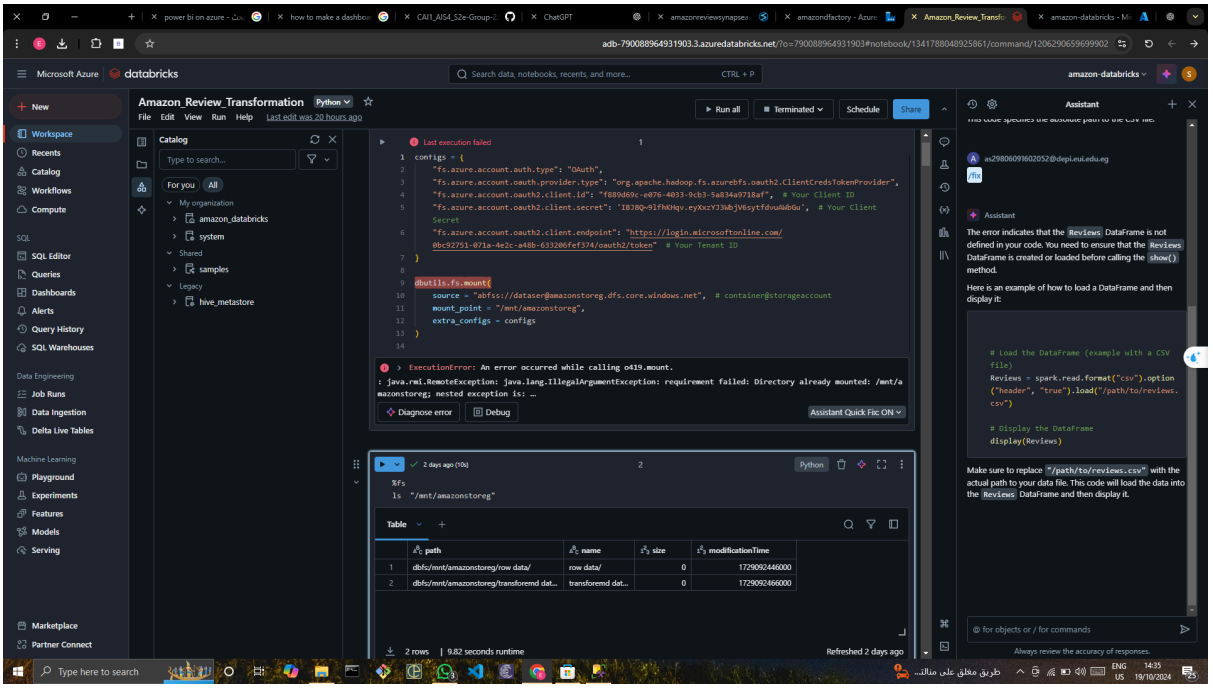


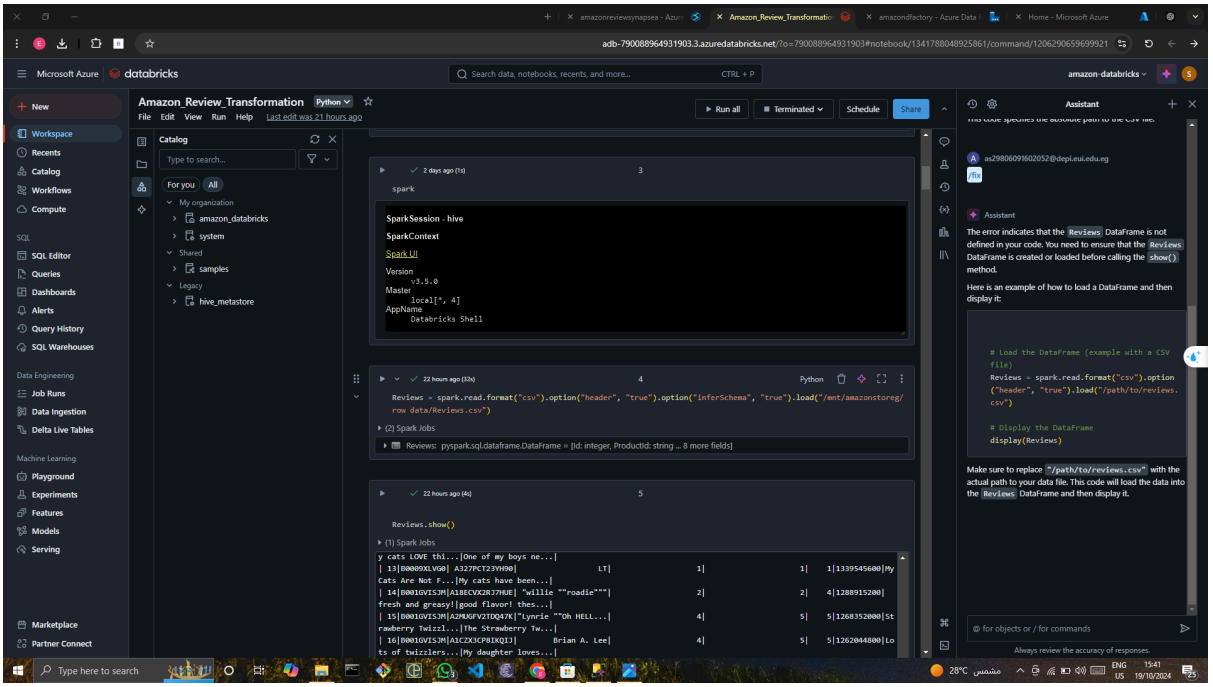
Here I uploaded the data to the data factory and I made the pipeline

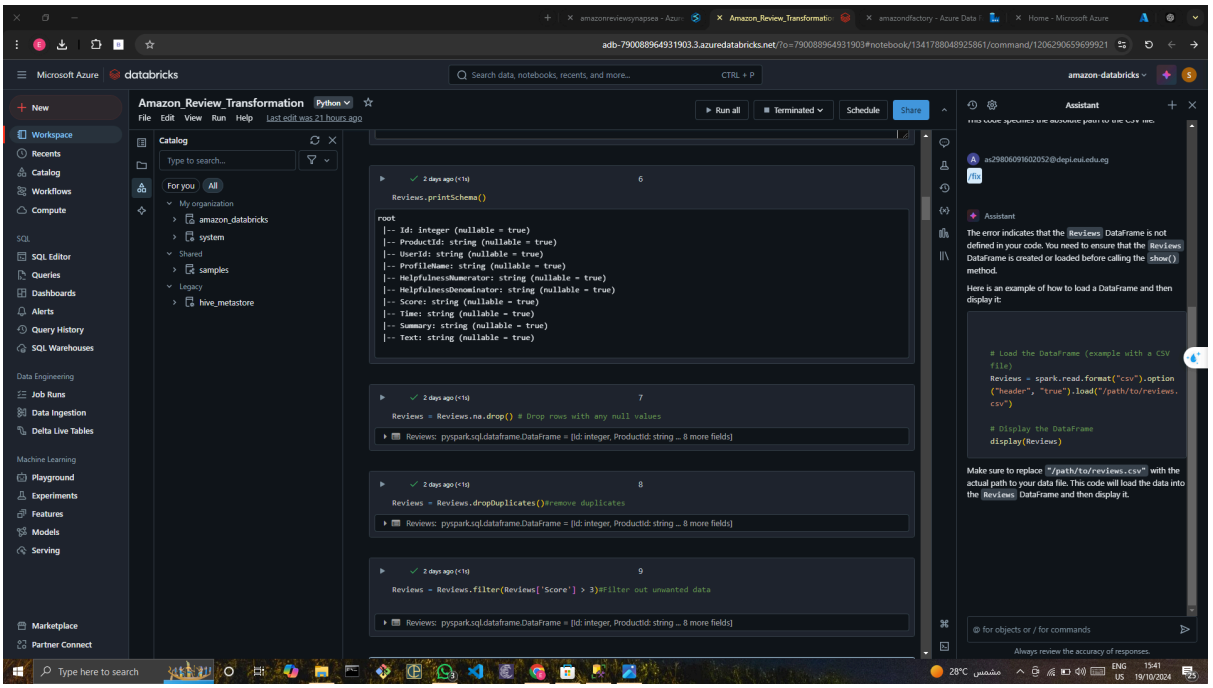


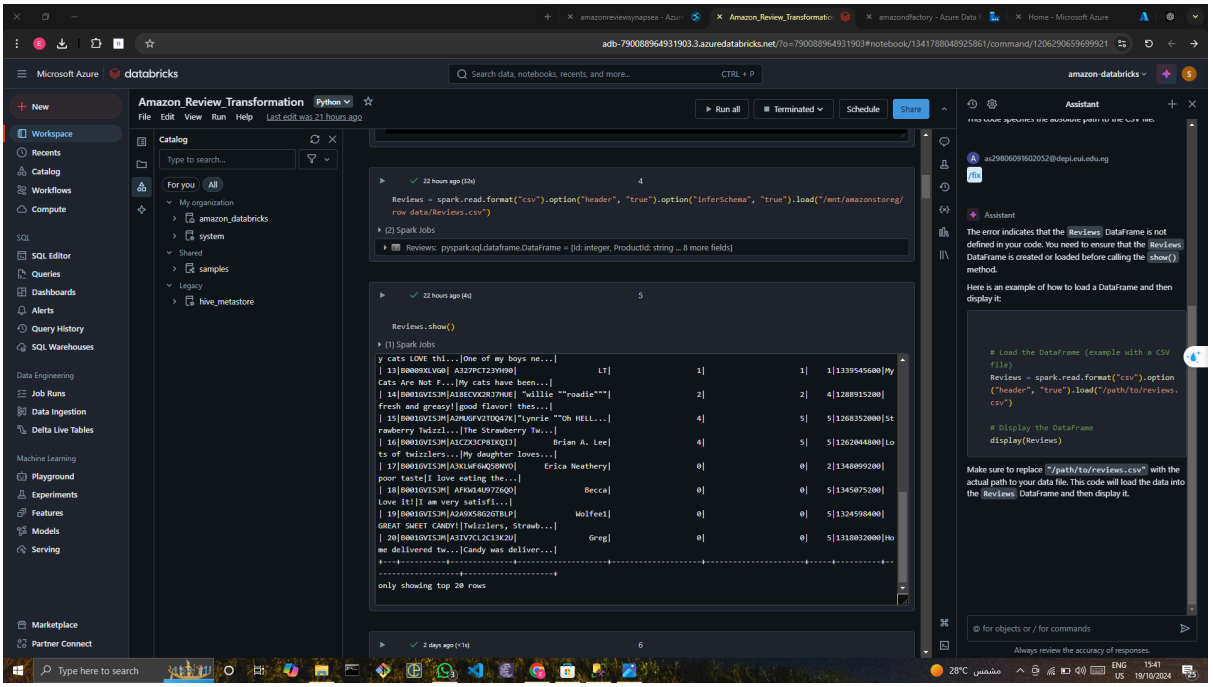


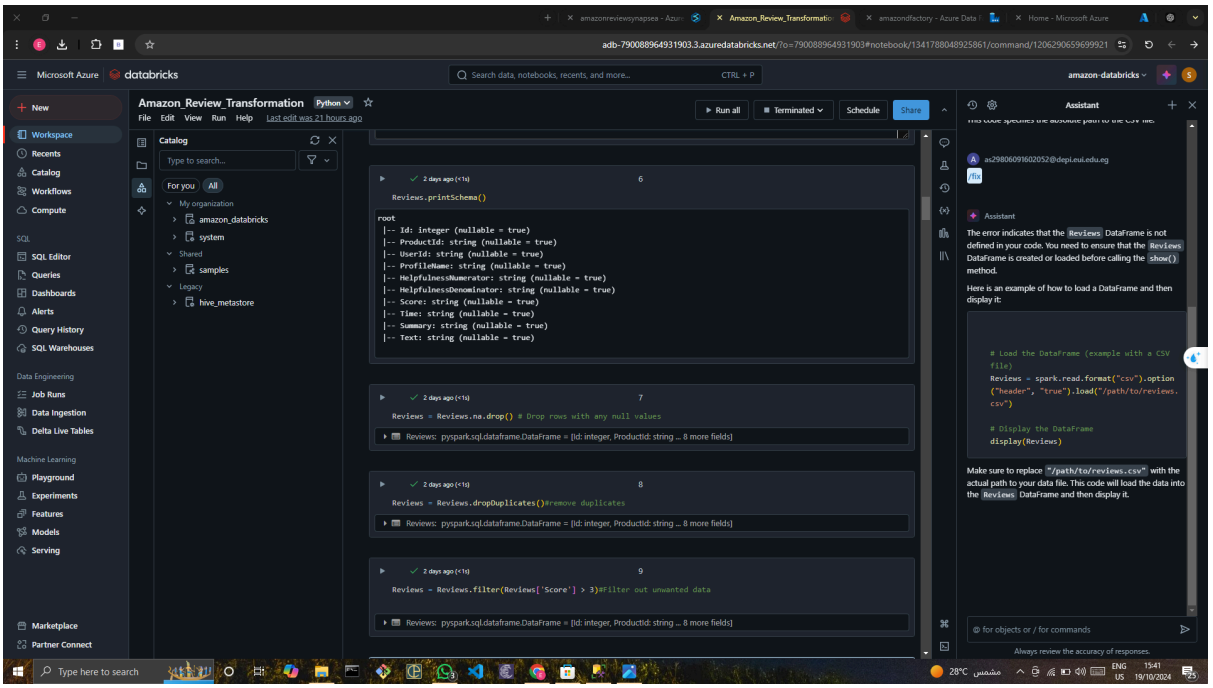
Azure data bricks and visualization using python

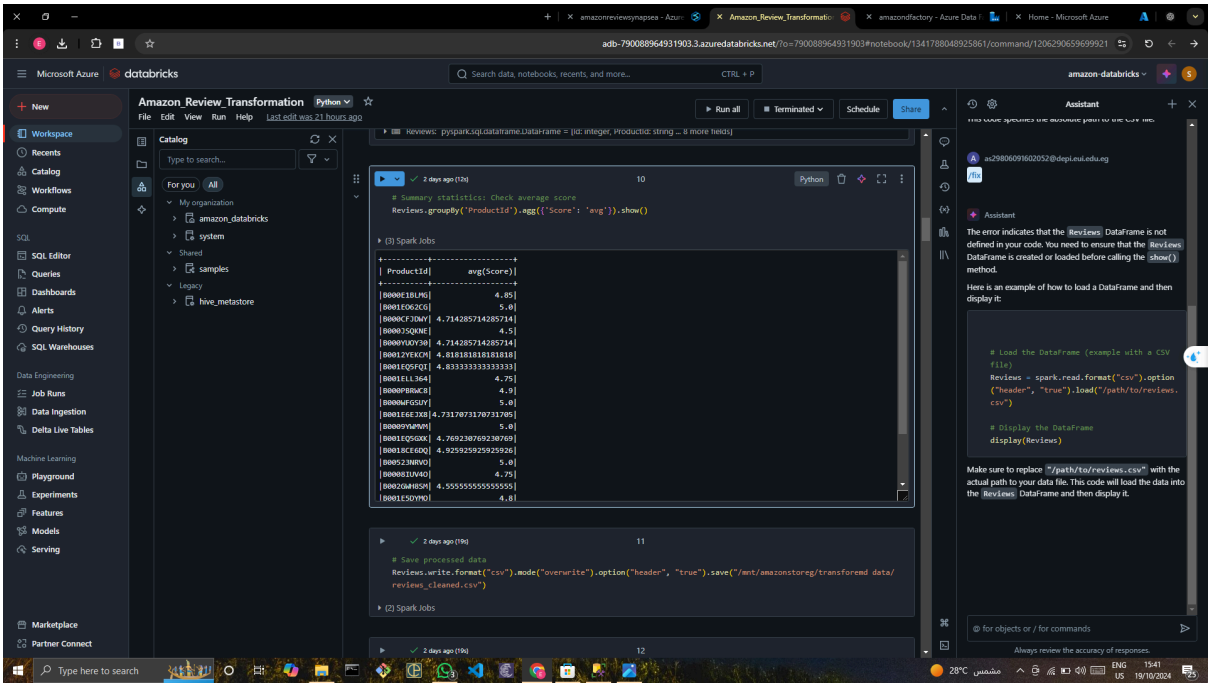


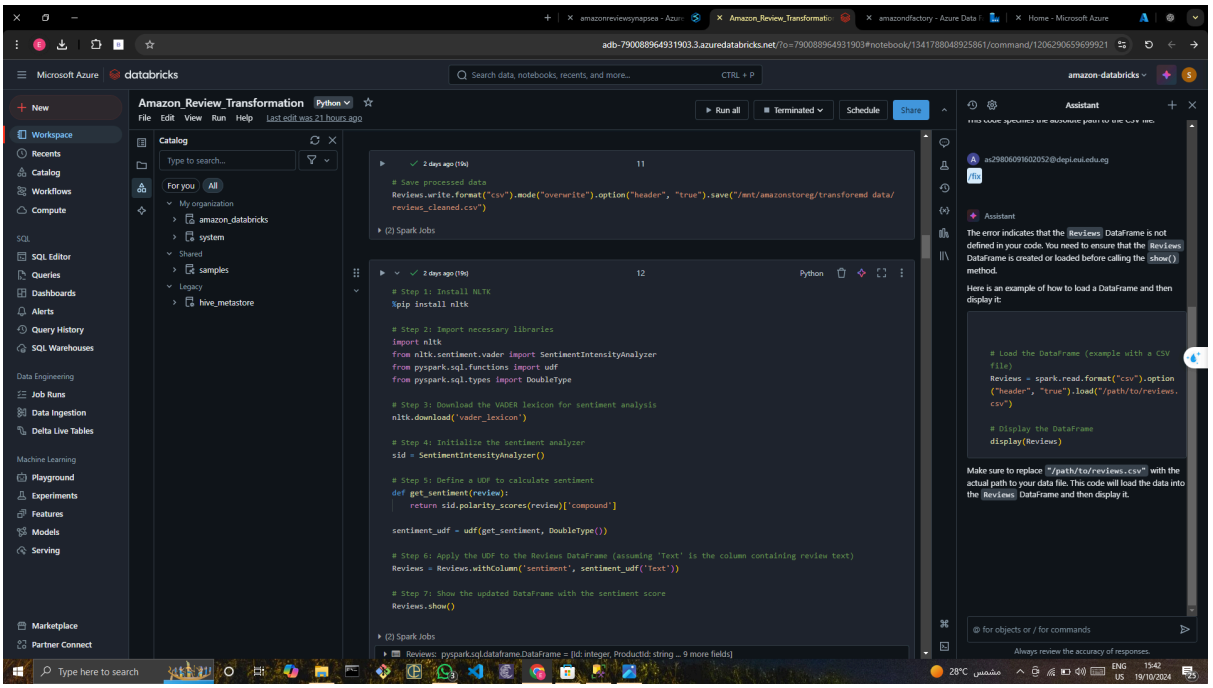

Amazon_Review_Transformation
Python
File Edit View Run Help Last edit was 21 hours ago
Run all
Terminated
Schedule
Share
Catalog
Assistant
spark
SparkSession - hive
SparkContext
Spark UI
Version
v3.5.0
Master
local[*, 4]
AppName
Databricks Shell
Reviews = spark.read.format("csv").option("header", "true").option("inferSchema", "true").load("/mnt/amazonstoreg/row data/Reviews.csv")
(2) Spark Jobs
Reviews: pyspark.sql.dataframe.DataFrame = [Id: integer, ProductId: string ... 8 more fields]
Reviews.show()
(1) Spark Jobs
y cats LOVE thi...|One of my boys ne...|
| 13|B0009XLVG0| A327PCT23YH90| LT| 1| 1| 1|1339545600|My
Cats Are Not F...|My cats have been...|
| 14|B001GVISJM|A18ECVX2RJ7HUE| "willie ""roadie"""| 2| 2| 4|1288915200|
fresh and greasy!|good flavor! thes...|
| 15|B001GVISJM|A2MXGFV2TDQ47K|"Lynrie ""Oh HELL...| 4| 5| 5|1268352000|St
rawberry Twizzl...|The Strawberry Tw...|
| 16|B001GVISJM|A1CZX3CP8IKQIJ| Brian A. Lee| 4| 5| 5|1262044800|Lo
ts of twizzlers...|My daughter loves...|
# Load the DataFrame (example with a CSV file)
Reviews = spark.read.format("csv").option("header", "true").load("/path/to/reviews.csv")
# Display the DataFrame
display(Reviews)
Make sure to replace "/path/to/reviews.csv" with the actual path to your data file. This code will load the data into the Reviews DataFrame and then display it.
Always review the accuracy of responses.

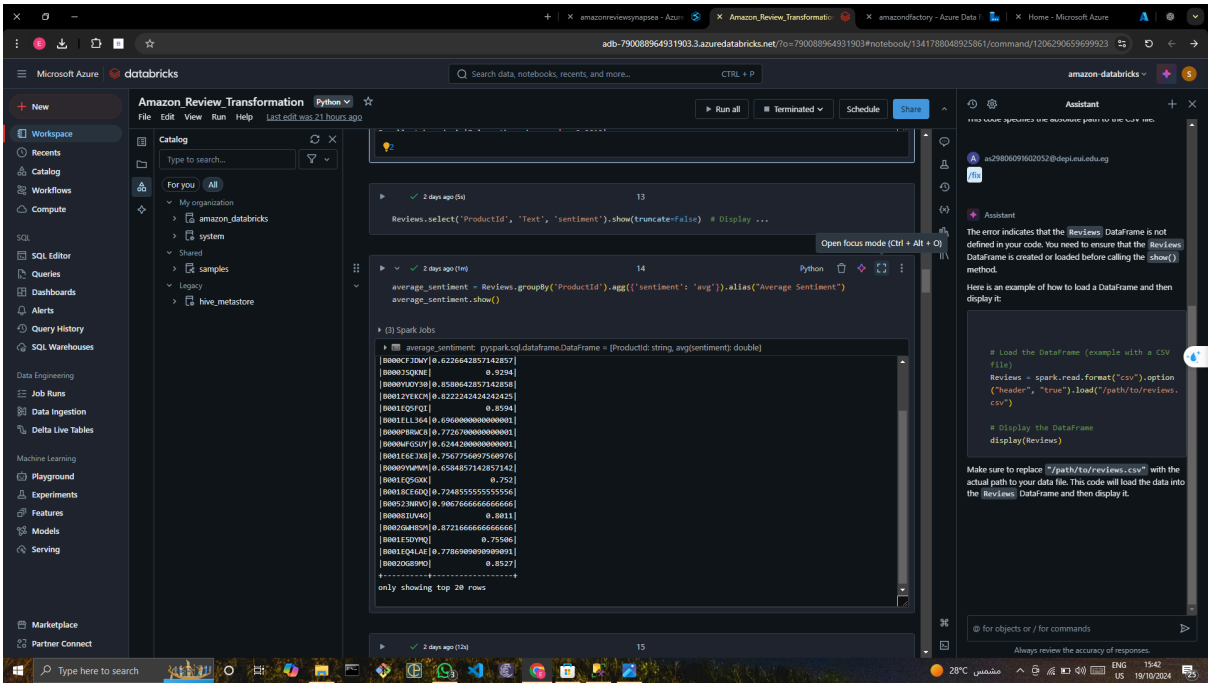

Amazon_Review_Transformation
Python
File Edit View Run Help Last edit was 21 hours ago
Run all
Terminated
Schedule
Share
Catalog
Assistant
Reviews.printSchema()
root
|-- Id: integer (nullable = true)
|-- ProductId: string (nullable = true)
|-- UserId: string (nullable = true)
|-- ProfileName: string (nullable = true)
|-- HelpfulnessNumerator: string (nullable = true)
|-- HelpfulnessDenominator: string (nullable = true)
|-- Score: string (nullable = true)
|-- Time: string (nullable = true)
|-- Summary: string (nullable = true)
|-- Text: string (nullable = true)
Reviews = Reviews.na.drop() # Drop rows with any null values
Reviews: pyspark.sql.dataframe.DataFrame = [Id: integer, ProductId: string ... 8 more fields]
Reviews = Reviews.dropDuplicates()#remove duplicates
Reviews: pyspark.sql.dataframe.DataFrame = [Id: integer, ProductId: string ... 8 more fields]
Reviews = Reviews.filter(Reviews['Score'] > 3)#filter out unwanted data
Reviews: pyspark.sql.dataframe.DataFrame = [Id: integer, ProductId: string ... 8 more fields]
# Load the DataFrame (example with a CSV file)
Reviews = spark.read.format("csv").option("header", "true").load("/path/to/reviews.csv")
# Display the DataFrame
display(Reviews)
Make sure to replace "/path/to/reviews.csv" with the actual path to your data file. This code will load the data into the Reviews DataFrame and then display it.
Always review the accuracy of responses.

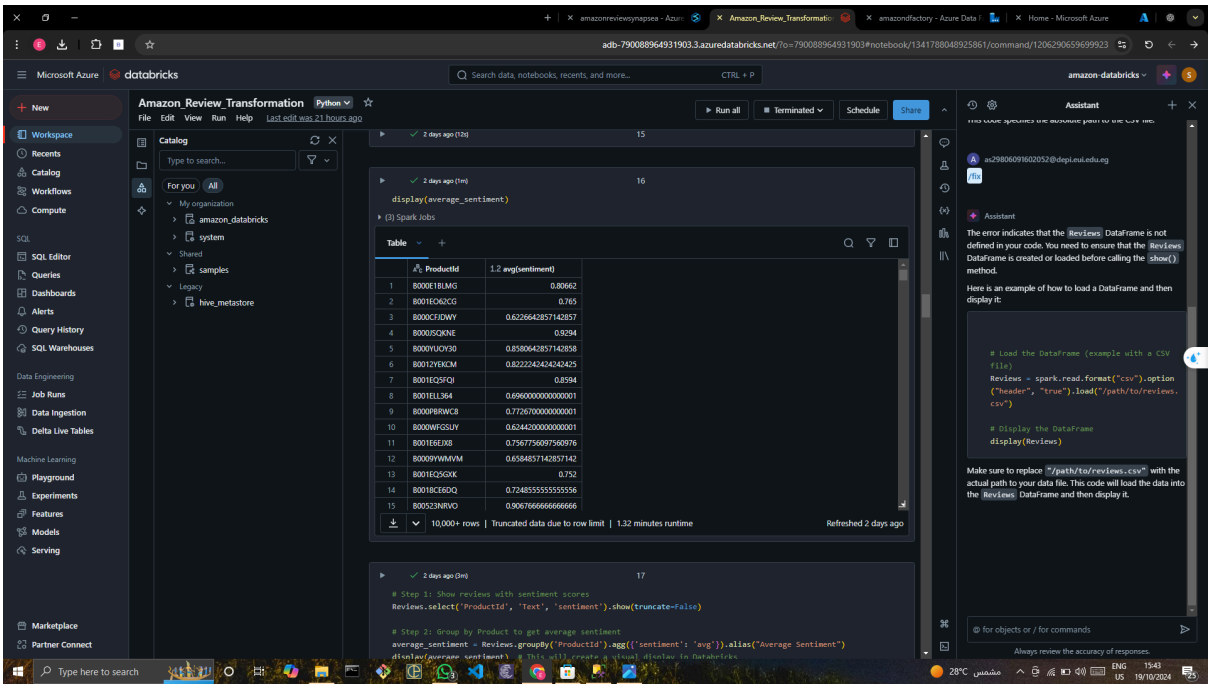

Amazon_Review_Transformation
Python
File Edit View Run Help Last edit was 21 hours ago
Run all
Terminated
Schedule
Share
Catalog
Type to search...
For you All
My organization
amazon_databricks
system
Shared
samples
Legacy
hive_metastore
22 hours ago (32s) 4
Reviews = spark.read.format("csv").option("header", "true").option("inferSchema", "true").load("/mnt/amazonstoreg/row data/Reviews.csv")
(2) Spark Jobs
Reviews: pyspark.sql.dataframe.DataFrame = [Id: integer, ProductId: string ... 8 more fields]
22 hours ago (4s) 5
Reviews.show()
(1) Spark Jobs
only showing top 20 rows
2 days ago (<1s) 6
Assistant
The error indicates that the Reviews DataFrame is not defined in your code. You need to ensure that the Reviews DataFrame is created or loaded before calling the show() method.
Here is an example of how to load a DataFrame and then display it:
# Load the DataFrame (example with a CSV file)
Reviews = spark.read.format("csv").option("header", "true").load("/path/to/reviews.csv")
# Display the DataFrame
display(Reviews)
Make sure to replace "/path/to/reviews.csv" with the actual path to your data file. This code will load the data into the Reviews DataFrame and then display it.
Always review the accuracy of responses.

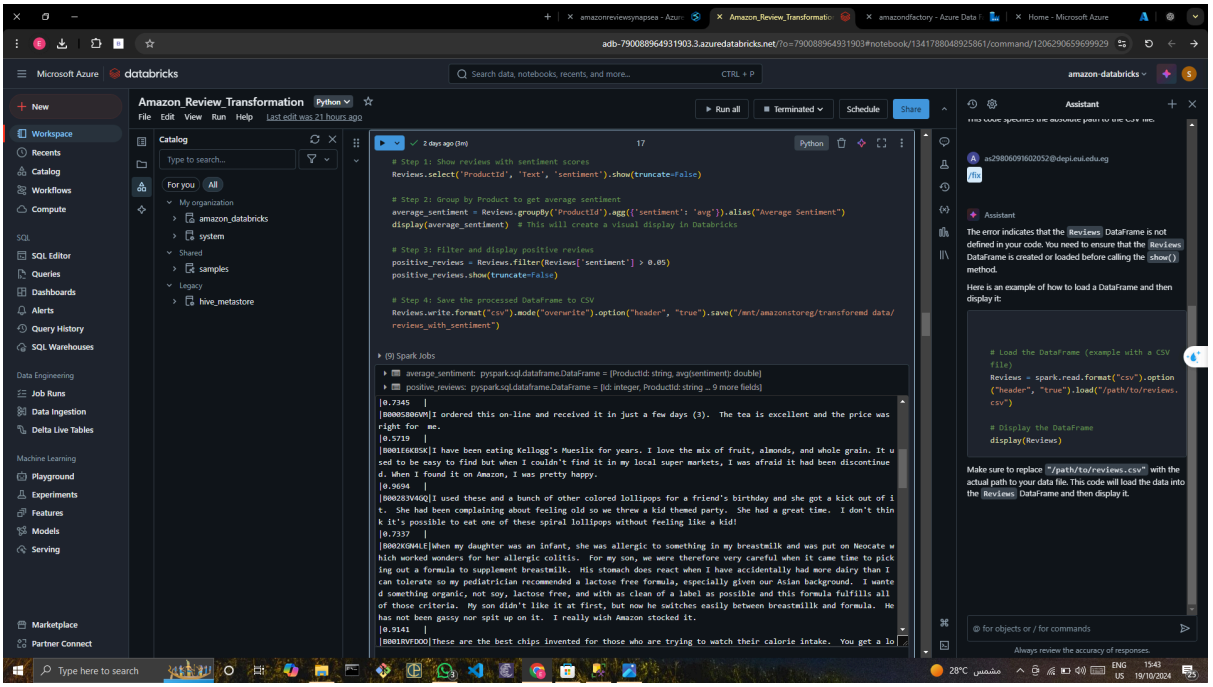

Amazon_Review_Transformation
Python
File Edit View Run Help Last edit was 21 hours ago
Run all
Terminated
Schedule
Share
Catalog
2 days ago (<1s) 6
Reviews.printSchema()
root
|-- Id: integer (nullable = true)
|-- ProductId: string (nullable = true)
|-- UserId: string (nullable = true)
|-- ProfileName: string (nullable = true)
|-- HelpfulnessNumerator: string (nullable = true)
|-- HelpfulnessDenominator: string (nullable = true)
|-- Score: string (nullable = true)
|-- Time: string (nullable = true)
|-- Summary: string (nullable = true)
|-- Text: string (nullable = true)
2 days ago (<1s) 7
Reviews = Reviews.na.drop() # Drop rows with any null values
Reviews: pyspark.sql.dataframe.DataFrame = [Id: integer, ProductId: string ... 8 more fields]
2 days ago (<1s) 8
Reviews = Reviews.dropDuplicates()#remove duplicates
Reviews: pyspark.sql.dataframe.DataFrame = [Id: integer, ProductId: string ... 8 more fields]
2 days ago (<1s) 9
Reviews = Reviews.filter(Reviews['Score'] > 3)#filter out unwanted data
Reviews: pyspark.sql.dataframe.DataFrame = [Id: integer, ProductId: string ... 8 more fields]
Assistant
The error indicates that the Reviews DataFrame is not defined in your code. You need to ensure that the Reviews DataFrame is created or loaded before calling the show() method.
Here is an example of how to load a DataFrame and then display it:
# Load the DataFrame (example with a CSV file)
Reviews = spark.read.format("csv").option("header", "true").load("/path/to/reviews.csv")
# Display the DataFrame
display(Reviews)
Make sure to replace "/path/to/reviews.csv" with the actual path to your data file. This code will load the data into the Reviews DataFrame and then display it.
Always review the accuracy of responses.

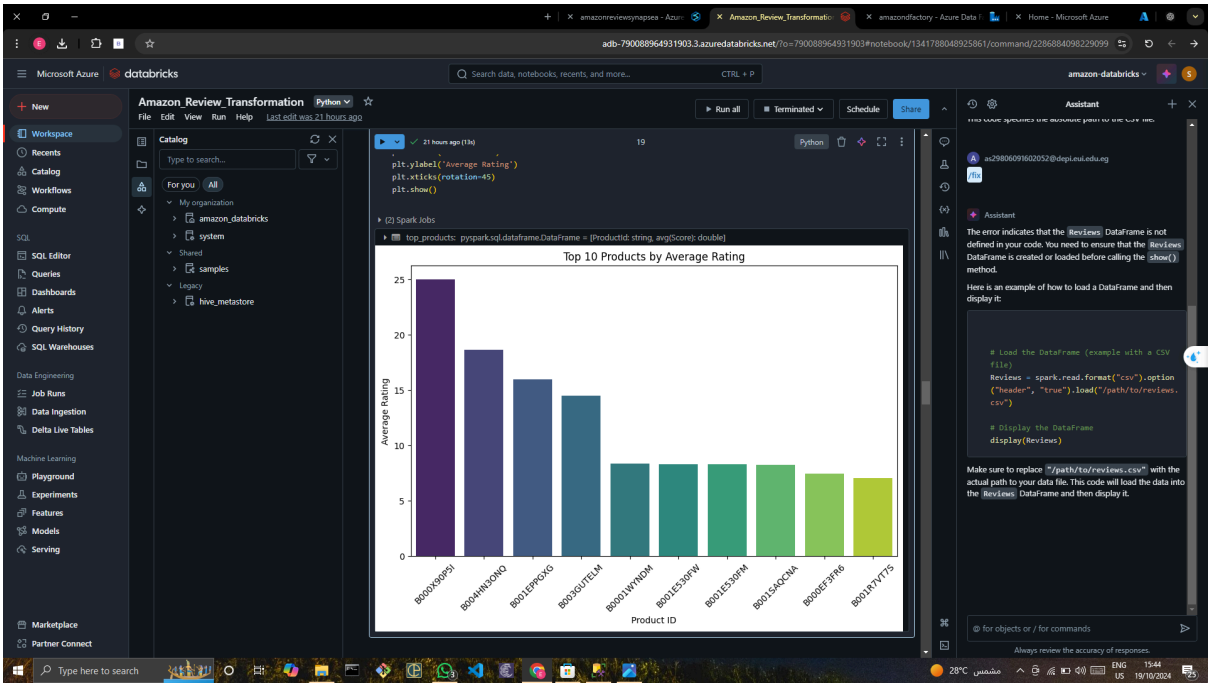

Amazon_Review_Transformation
Python
File Edit View Run Help Last edit was 21 hours ago
Run all
Terminated
Schedule
Share
Catalog
Assistant
# Summary statistics: Check average score
Reviews.groupBy("ProductId").agg({'Score': 'avg'}).show()
(3) Spark Jobs
| ProductId| avg(Score)|
|B000E1BLMG| 4.85|
|B001E06ZCG| 5.0|
|B000CFJDWY| 4.714285714285714|
|B000JSQKNE| 4.5|
|B000YUOY30| 4.714285714285714|
|B0012YEKCM| 4.818181818181818|
|B001EQSFQI| 4.833333333333333|
|B001ELL364| 4.75|
|B000P8RWC8| 4.9|
|B000WFGSUY| 5.0|
|B001E6EIXB|4.7317073170731705|
|B0009YWMVM| 5.0|
|B001EQSGGX| 4.769230769230769|
|B001ECI4OQ| 4.925925925925926|
|B00523NRVO| 5.0|
|B000BIUV4O| 4.75|
|B002GWHBSM| 4.555555555555555|
|B001E5DYMO| 4.8|
# Save processed data
Reviews.write.format("csv").mode("overwrite").option("header", "true").save("/mnt/amazonstoreg/transforemd data/reviews_cleaned.csv")
(2) Spark Jobs
Make sure to replace "/path/to/reviews.csv" with the actual path to your data file. This code will load the data into the Reviews DataFrame and then display it.
Always review the accuracy of responses.

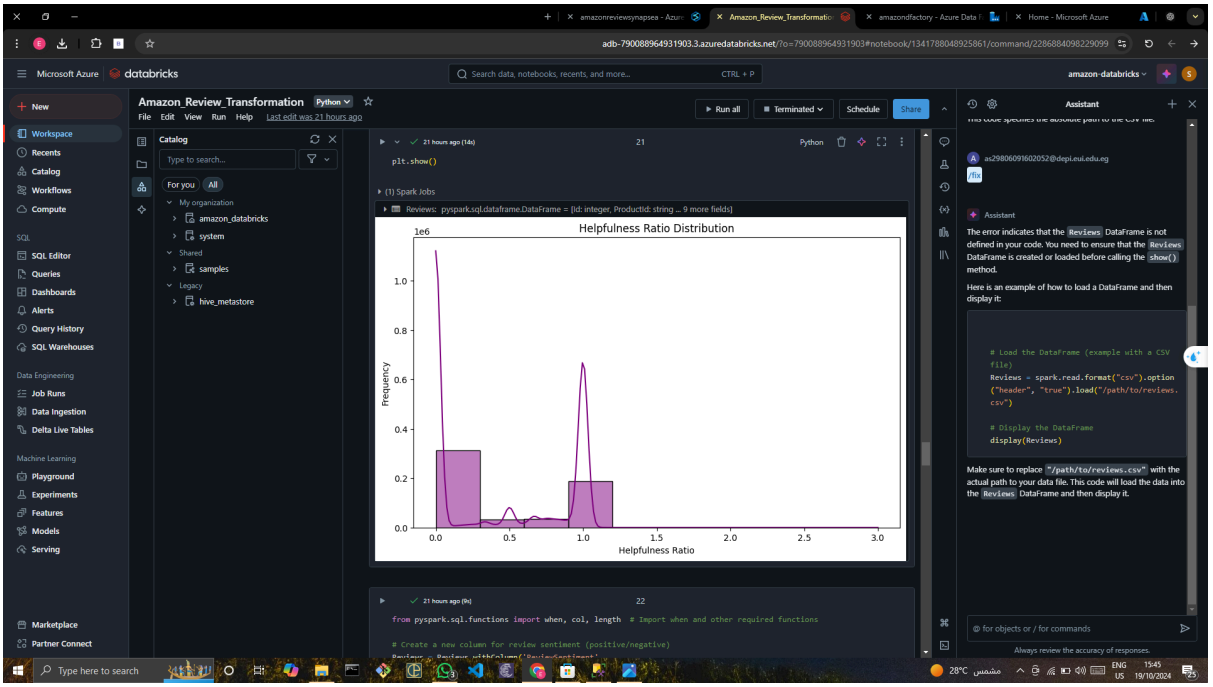

Amazon_Review_Transformation
Python
File Edit View Run Help Last edit was 21 hours ago
Run all
Terminated
Schedule
Share
Catalog
Assistant
# Save processed data
Reviews.write.format("csv").mode("overwrite").option("header", "true").save("/mnt/amazonstoreg/transforemd data/reviews_cleaned.csv")
(2) Spark Jobs
Step 1: Install NLTK
%pip install nltk
# Step 2: Import necessary libraries
import nltk
from nltk.sentiment.vader import SentimentIntensityAnalyzer
from pyspark.sql.functions import udf
from pyspark.sql.types import DoubleType
# Step 3: Download the VADER lexicon for sentiment analysis
nltk.download('vader_lexicon')
# Step 4: Initialize the sentiment analyzer
sid = SentimentIntensityAnalyzer()
# Step 5: Define a UDF to calculate sentiment
def get_sentiment(review):
return sid.polarity_scores(review)['compound']
sentiment_udf = udf(get_sentiment, DoubleType())
# Step 6: Apply the UDF to the Reviews DataFrame (assuming 'Text' is the column containing review text)
Reviews = Reviews.withColumn('sentiment', sentiment_udf('Text'))
# Step 7: Show the updated DataFrame with the sentiment score
Reviews.show()
(2) Spark Jobs
Make sure to replace "/path/to/reviews.csv" with the actual path to your data file. This code will load the data into the Reviews DataFrame and then display it.
Always review the accuracy of responses.

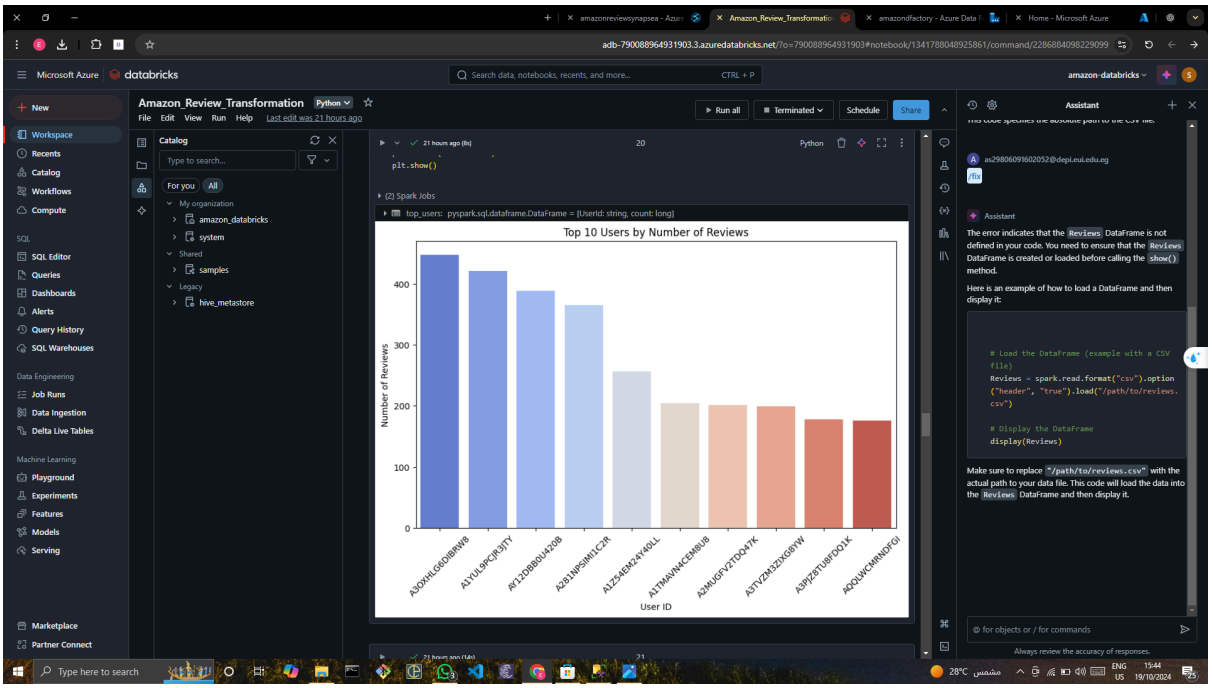

Amazon_Review_Transformation
Python
File Edit View Run Help Last edit was 21 hours ago
Run all
Terminated
Schedule
Share
Catalog
Assistant
13
Reviews.select('ProductId', 'Text', 'sentiment').show(truncate=False) # Display ...
14
average_sentiment = Reviews.groupBy('ProductId').agg({'sentiment': 'avg'}).alias("Average Sentiment")
average_sentiment.show()
(3) Spark Jobs
B000CFJDWY	0.6226642857142857
B000JSQKNE	0.9294
B000YUOY30	0.8580642857142858
B0012YEKCM	0.8222242424242425
B001EQSFQI	0.8594
B001ELL364	0.6960000000000001
B000PBRWC8	0.7726700000000001
B000WFGSUY	0.6244200000000001
B001E6EJX8	0.7567756097560976
B0009YWMVM	0.6584857142857142
B001EQSGXK	0.752
B0018CE6DQ	0.7248555555555556
B00523NRVO	0.9067666666666666
B000BIUV4O	0.8011
B002GWH8SM	0.8721666666666666
B001ESDYMQ	0.75506
B001EQ4LAE	0.7786909090909091
B002OG89MO	0.8527
only showing top 20 rows
15
Reviews = spark.read.format("csv").option("header", "true").load("/path/to/reviews.csv")
display(Reviews)
Make sure to replace "/path/to/reviews.csv" with the actual path to your data file. This code will load the data into the Reviews DataFrame and then display it.
Always review the accuracy of responses.

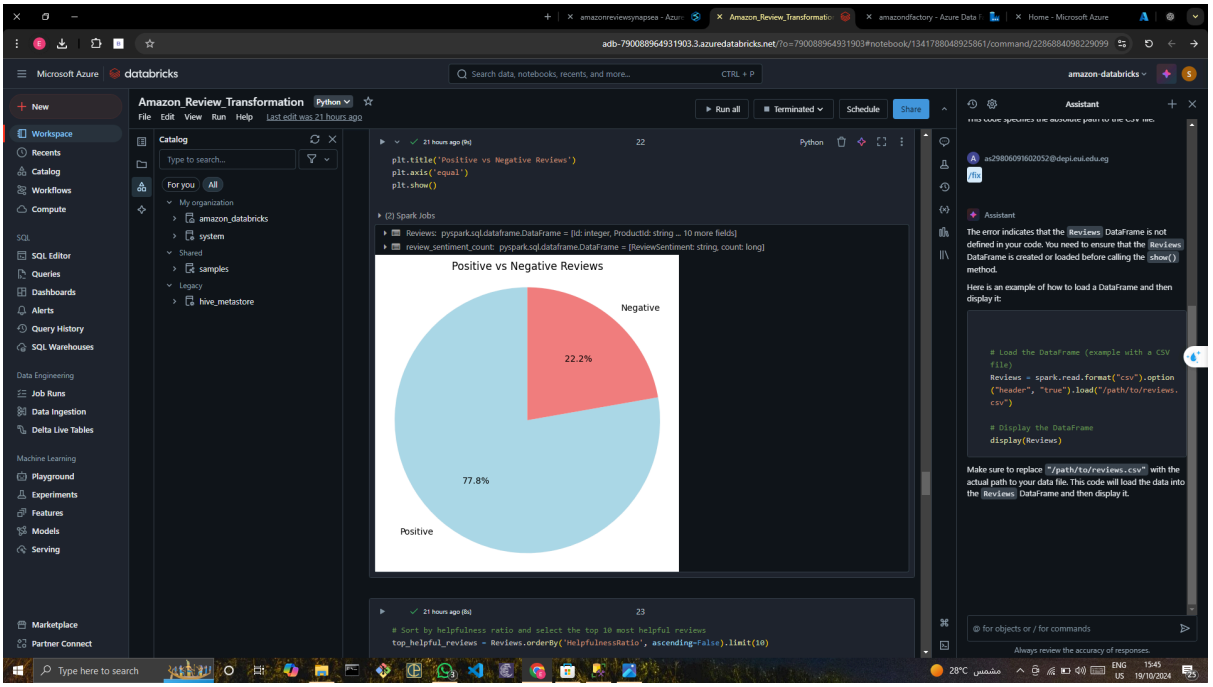

Amazon_Review_Transformation
Python
File Edit View Run Help Last edit was 21 hours ago
Run all
Terminated
Schedule
Share
Catalog
Assistant
15
16
display(average_sentiment)
(3) Spark Jobs
Table
ProductId
avg(sentiment)
1 B000E1BLMG 0.80662
2 B001EO6ZCG 0.765
3 B000CFJDWY 0.6226642857142857
4 B000JSQKNE 0.9294
5 B000YUOY30 0.8580642857142858
6 B0012YEKCM 0.8222242424242425
7 B001EQSFQI 0.8594
8 B001ELL364 0.6960000000000001
9 B000PBRWC8 0.7726700000000001
10 B000WFGSUY 0.6244200000000001
11 B001E6EJX8 0.7567756097560976
12 B0009YWMVM 0.6584857142857142
13 B001EQSGXK 0.752
14 B0018CE6DQ 0.7248555555555556
15 B00523NRVO 0.9067666666666666
10,000+ rows | Truncated data due to row limit | 1.32 minutes runtime
Refreshed 2 days ago
17
# Step 1: Show reviews with sentiment scores
Reviews.select('ProductId', 'Text', 'sentiment').show(truncate=False)
# Step 2: Group by Product to get average sentiment
average_sentiment = Reviews.groupBy('ProductId').agg({'sentiment': 'avg'}).alias("Average Sentiment")
Reviews = spark.read.format("csv").option("header", "true").load("/path/to/reviews.csv")
display(Reviews)
Make sure to replace "/path/to/reviews.csv" with the actual path to your data file. This code will load the data into the Reviews DataFrame and then display it.
Always review the accuracy of responses.

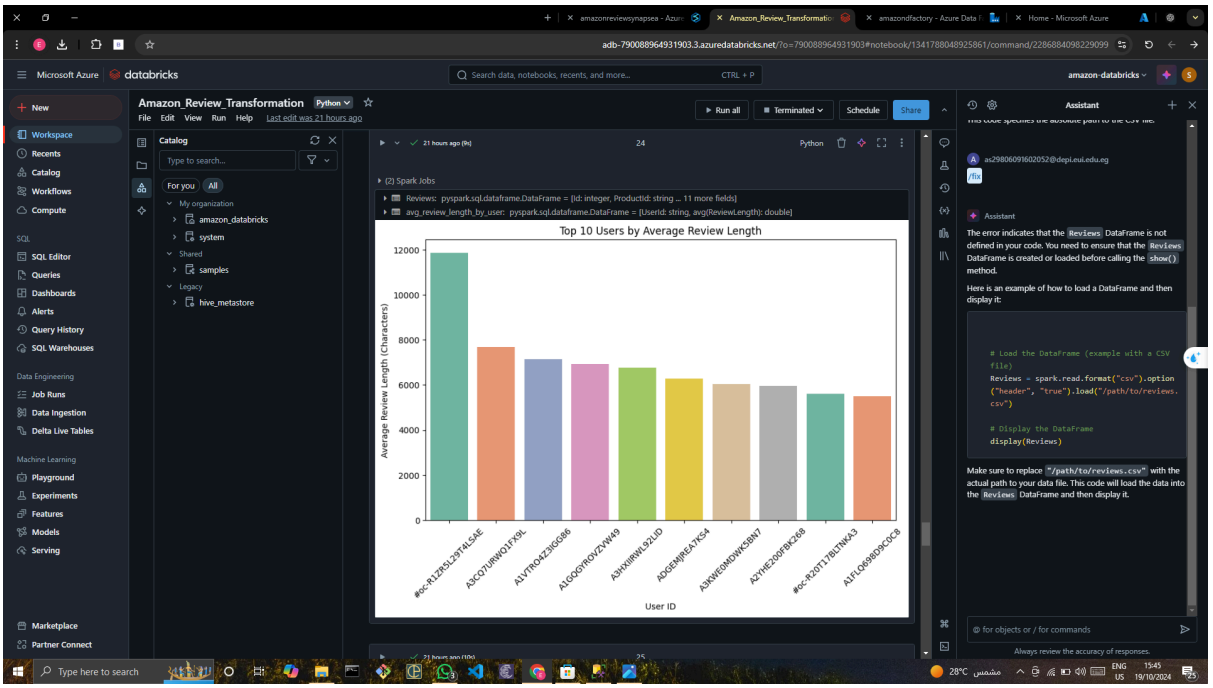

Amazon_Review_Transformation
Python
File Edit View Run Help Last edit was 21 hours ago
Run all
Terminated
Schedule
Share
Catalog
Type to search...
For you All
My organization
amazon_databricks
system
Shared
samples
Legacy
hive_metastore
2 days ago (3m)
17
# Step 1: Show reviews with sentiment scores
Reviews.select('ProductId', 'Text', 'sentiment').show(truncate=False)
# Step 2: Group by Product to get average sentiment
average_sentiment = Reviews.groupBy('ProductId').agg({'sentiment': 'avg'}).alias("Average Sentiment")
display(average_sentiment) # This will create a visual display in Databricks
# Step 3: Filter and display positive reviews
positive_reviews = Reviews.filter(Reviews['sentiment'] > 0.05)
positive_reviews.show(truncate=False)
# Step 4: Save the processed DataFrame to CSV
Reviews.write.format("csv").mode("overwrite").option("header", "true").save("/mnt/amazonstoreg/transforemd data/reviews_with_sentiment")
(9) Spark Jobs
average_sentiment: pyspark.sql.dataframe.DataFrame = [ProductId: string, avg(sentiment): double]
positive_reviews: pyspark.sql.dataframe.DataFrame = [Id: integer, ProductId: string ... 9 more fields]
Assistant
The error indicates that the Reviews DataFrame is not defined in your code. You need to ensure that the Reviews DataFrame is created or loaded before calling the show() method.
Here is an example of how to load a DataFrame and then display it:
# Load the DataFrame (example with a CSV file)
Reviews = spark.read.format("csv").option("header", "true").load("/path/to/reviews.csv")
# Display the DataFrame
display(Reviews)
Make sure to replace "/path/to/reviews.csv" with the actual path to your data file. This code will load the data into the Reviews DataFrame and then display it.
Always review the accuracy of responses.

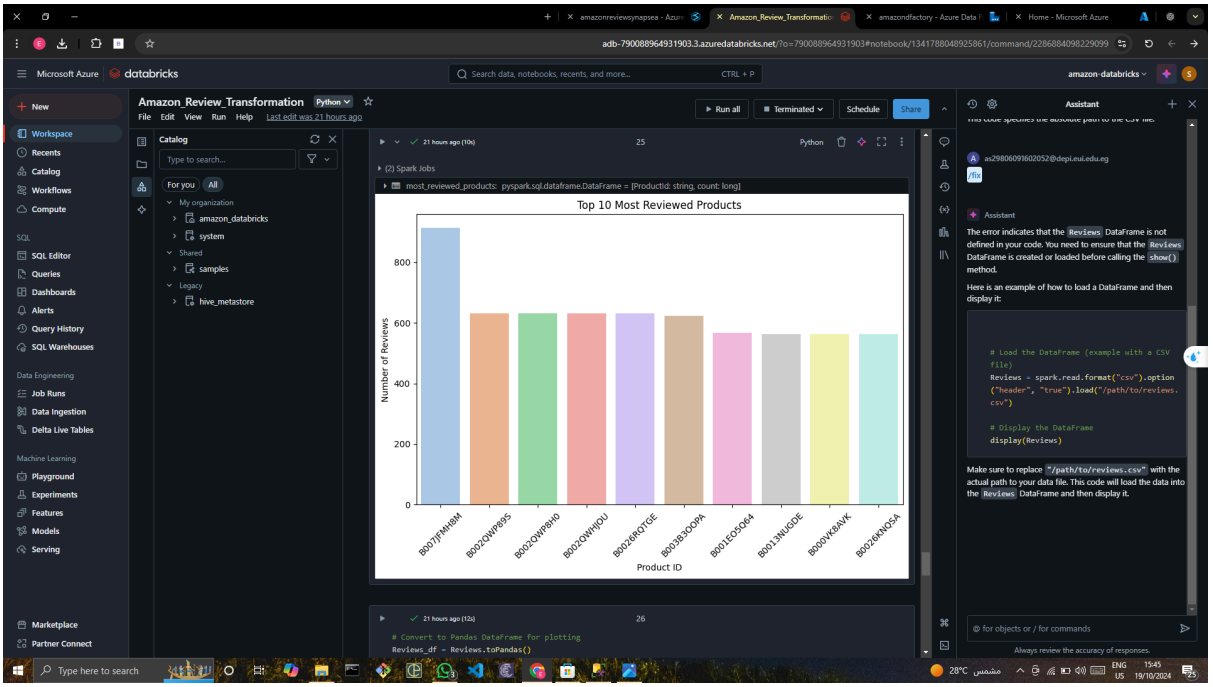

Amazon_Review_Transformation
Python
File Edit View Run Help Last edit was 21 hours ago
Run all
Terminated
Schedule
Share
Catalog
21 hours ago (13s)
19
plt.ylabel('Average Rating')
plt.xticks(rotation=45)
plt.show()
(2) Spark Jobs
top_products: pyspark.sql.dataframe.DataFrame = [ProductId: string, avg(Score): double]
Top 10 Products by Average Rating
Average Rating
Product ID
B000X90PSI
B004HN3ONO
B001EPPOXG
B003GUTELM
B0001WYNDM
B001E530FW
B001ES30FM
B0015AOCNA
B000EF3FR6
B001R7VT7S
Assistant
The error indicates that the Reviews DataFrame is not defined in your code. You need to ensure that the Reviews DataFrame is created or loaded before calling the show() method.
Here is an example of how to load a DataFrame and then display it:
# Load the DataFrame (example with a CSV file)
Reviews = spark.read.format("csv").option("header", "true").load("/path/to/reviews.csv")
# Display the DataFrame
display(Reviews)
Make sure to replace "/path/to/reviews.csv" with the actual path to your data file. This code will load the data into the Reviews DataFrame and then display it.
Always review the accuracy of responses.

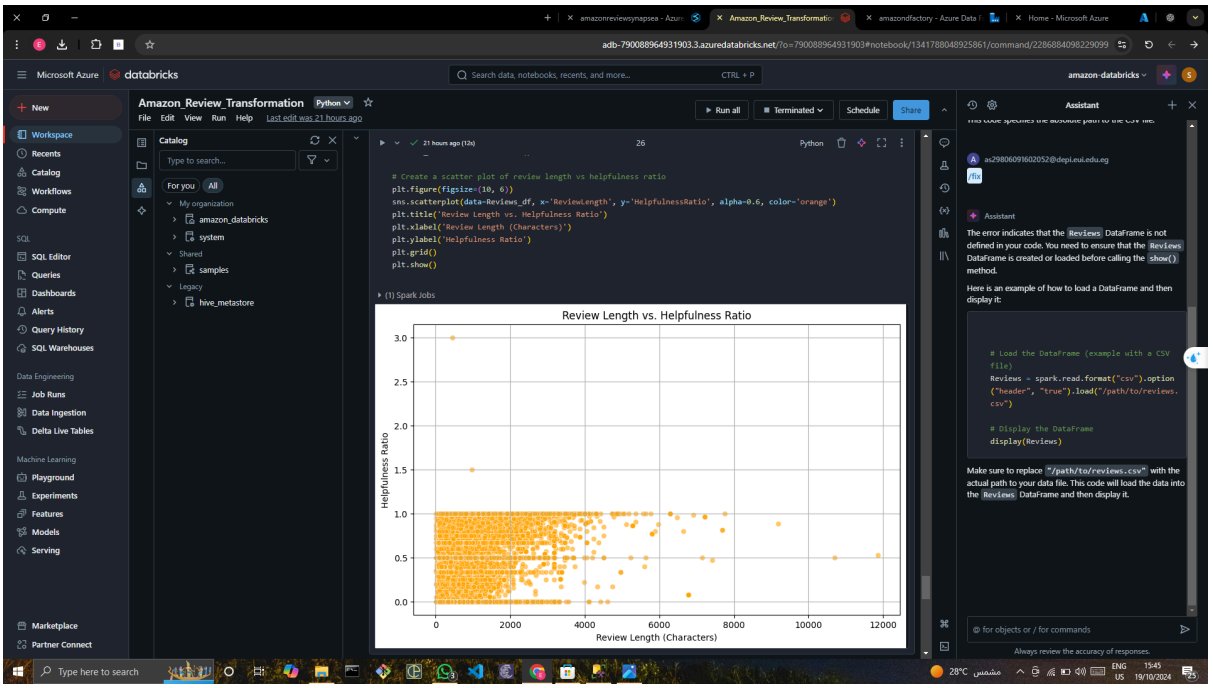

Amazon_Review_Transformation
Helpfulness Ratio Distribution
Frequency
Helpfulness Ratio
plt.show()
from pyspark.sql.functions import when, col, length # Import when and other required functions
Assistant
The error indicates that the Reviews DataFrame is not defined in your code. You need to ensure that the Reviews DataFrame is created or loaded before calling the show() method.
Here is an example of how to load a DataFrame and then display it:
# Load the DataFrame (example with a CSV file)
Reviews = spark.read.format("csv").option("header", "true").load("/path/to/reviews.csv")
# Display the DataFrame
display(Reviews)
Make sure to replace "/path/to/reviews.csv" with the actual path to your data file. This code will load the data into the Reviews DataFrame and then display it.

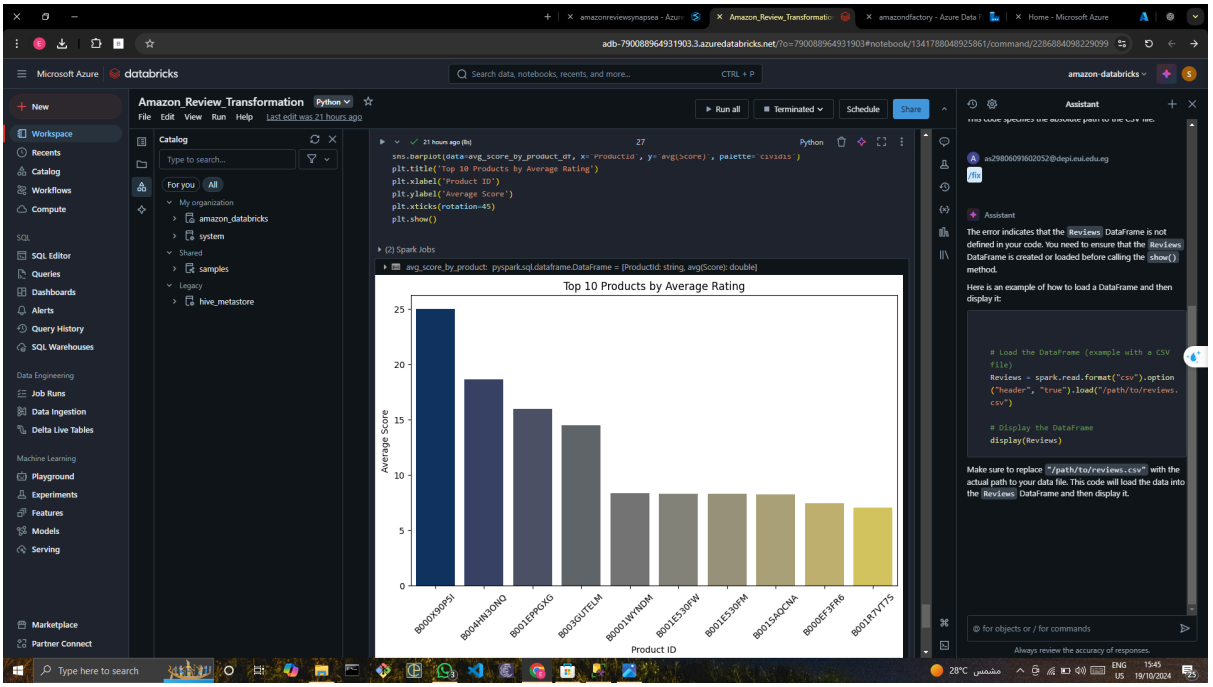

Amazon_Review_Transformation
Top 10 Users by Number of Reviews
Number of Reviews
User ID
A3OXHLG6DIBRW8
A1YUL9PCJR3JTY
AF13DBB0U420B
A2B1WPSIM1CZR
A1Z54EM24Y40LL
A1TMAVN4CEMBU8
A3MUGFV2TDQ47K
A3TVZM3ZIXG8YW
A3PJZ8TU8FDQ1K
ADQLWCMRNDFGI
plt.show()
Assistant
The error indicates that the Reviews DataFrame is not defined in your code. You need to ensure that the Reviews DataFrame is created or loaded before calling the show() method.
Here is an example of how to load a DataFrame and then display it:
# Load the DataFrame (example with a CSV file)
Reviews = spark.read.format("csv").option("header", "true").load("/path/to/reviews.csv")
# Display the DataFrame
display(Reviews)
Make sure to replace "/path/to/reviews.csv" with the actual path to your data file. This code will load the data into the Reviews DataFrame and then display it.


Amazon_Review_Transformation
Positive vs Negative Reviews
Negative
22.2%
77.8%
Positive
plt.title('Positive vs Negative Reviews')
plt.axis('equal')
plt.show()
# Sort by helpfulness ratio and select the top 10 most helpful reviews
top_helpful_reviews = Reviews.orderBy('HelpfulnessRatio', ascending=False).limit(10)
Assistant
The error indicates that the Reviews DataFrame is not defined in your code.
# Load the DataFrame (example with a CSV file)
Reviews = spark.read.format("csv").option("header", "true").load("/path/to/reviews.csv")
# Display the DataFrame
display(Reviews)
Make sure to replace "/path/to/reviews.csv" with the actual path to your data file.


Amazon_Review_Transformation
Top 10 Users by Average Review Length
Average Review Length (Characters)
User ID
#oc-R12R5L29TALS4E
A3CQ7URWO1FX9L
A1VTRO4Z3GG86
A1GGOYRGVZVW49
A3HXIRWL92LID
ADGEMJPEATKS4
A3KWEOMDWKSBN7
A27HE200FBK268
#oc-RJOT17BLTNKA3
A1FLO6986D9COC8
Assistant
The error indicates that the Reviews DataFrame is not defined in your code.
Make sure to replace "/path/to/reviews.csv" with the actual path to your data file.


Amazon_Review_Transformation
Top 10 Most Reviewed Products
Number of Reviews
Product ID
B007JFMH8M
B002QWP89S
B002QWP8H0
B002QWPHJQU
B00268OTGE
B003B3OOPA
B001EO5Q64
B0013NUGDE
B000VK8AVK
B0026KNQSA
# Convert to Pandas DataFrame for plotting
Reviews_df = Reviews.toPandas()
Assistant
The error indicates that the Reviews DataFrame is not defined in your code. You need to ensure that the Reviews DataFrame is created or loaded before calling the show() method.
Here is an example of how to load a DataFrame and then display it:
# Load the DataFrame (example with a CSV file)
Reviews = spark.read.format("csv").option("header", "true").load("/path/to/reviews.csv")
# Display the DataFrame
display(Reviews)
Make sure to replace "/path/to/reviews.csv" with the actual path to your data file. This code will load the data into the Reviews DataFrame and then display it.


Amazon_Review_Transformation
# Create a scatter plot of review length vs helpfulness ratio
plt.figure(figsize=(10, 6))
sns.scatterplot(data=Reviews_df, x='ReviewLength', y='HelpfulnessRatio', alpha=0.6, color='orange')
plt.title("Review Length vs. Helpfulness Ratio")
plt.xlabel('Review Length (Characters)')
plt.ylabel('Helpfulness Ratio')
plt.grid()
plt.show()
Review Length vs. Helpfulness Ratio
Helpfulness Ratio
Review Length (Characters)
Assistant
The error indicates that the Reviews DataFrame is not defined in your code. You need to ensure that the Reviews DataFrame is created or loaded before calling the show() method.
Here is an example of how to load a DataFrame and then display it:
# Load the DataFrame (example with a CSV file)
Reviews = spark.read.format("csv").option("header", "true").load("/path/to/reviews.csv")
# Display the DataFrame
display(Reviews)
Make sure to replace "/path/to/reviews.csv" with the actual path to your data file. This code will load the data into the Reviews DataFrame and then display it.


Amazon_Review_Transformation
Top 10 Products by Average Rating
Average Score
Product ID
B000X90PSJ
B004HN3ONQ
B001EPPOXG
B003GUTELM
B0001WFNDM
B001E530FW
B001E530FM
B0015AQCNA
B000EF3FR6
B001R7V775