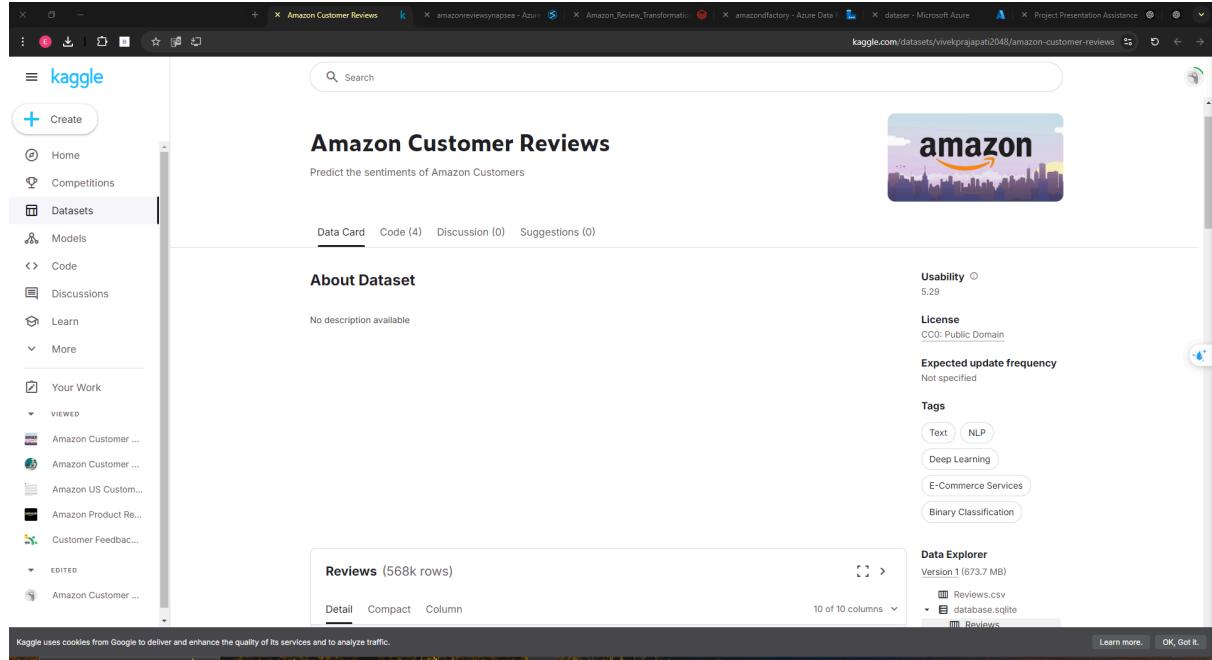


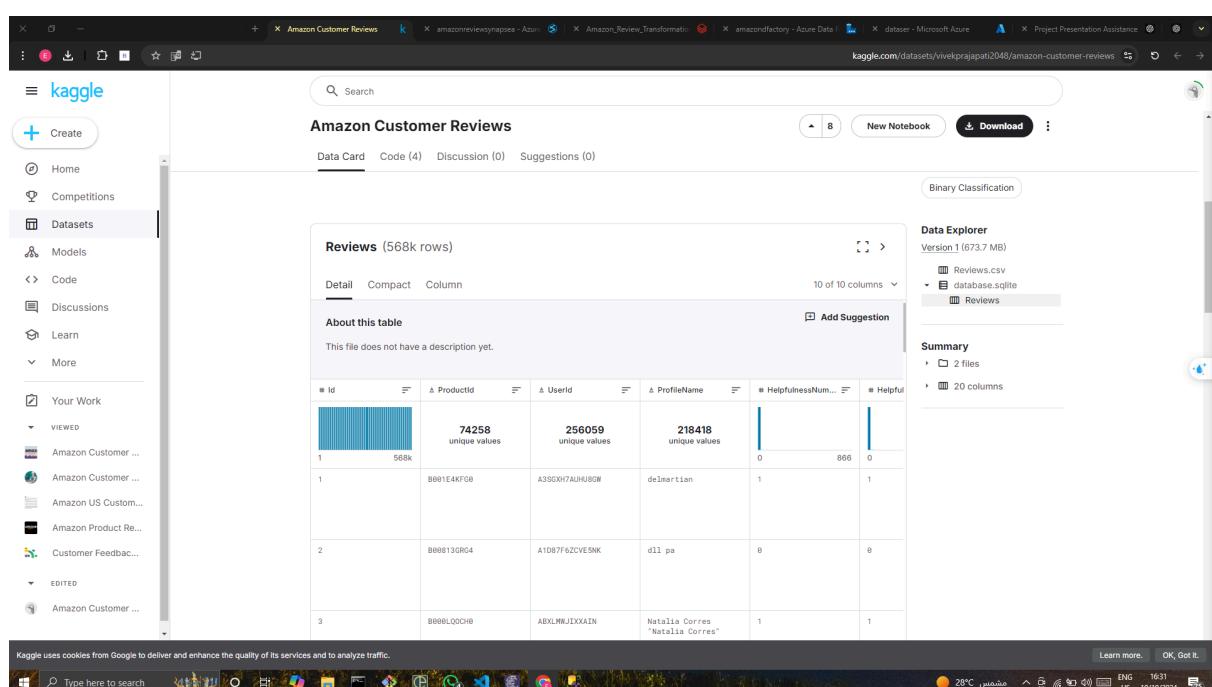
My part in the project : Mohamed Yaser karam abd Alshafy

Project title: Customer Feedback Analysis and Improvement

The data set set that I used



The screenshot shows the Kaggle dataset page for "Amazon Customer Reviews". The page title is "Amazon Customer Reviews" and the subtitle is "Predict the sentiments of Amazon Customers". On the left sidebar, there are links for Create, Home, Competitions, Datasets (selected), Models, Code, Discussions, Learn, and More. Under "Your Work", there are sections for VIEWED and EDITED datasets. The main content area displays the "Reviews" table with 568k rows. The table has columns: Id, ProductId, UserId, ProfileName, HelpfulnessNum..., and Helpful. The Data Explorer section shows the dataset version 1 (673.7 MB) containing files like Reviews.csv and database.sqlite. The Summary section indicates 2 files and 20 columns.



The screenshot shows the same Kaggle dataset page for "Amazon Customer Reviews". The main content area now displays the "About this table" section, which states: "This file does not have a description yet." Below this, the "Reviews" table is shown with the first few rows of data. The Data Explorer and Summary sections remain the same as in the previous screenshot.

The resources that I made on Azure

The screenshot shows the Microsoft Azure portal interface. At the top, there are several tabs: 'Create a resource', 'Cost Management...', 'Cost Management', 'Subscriptions', 'Storage accounts', 'Azure Synapse Analytics', 'App registrations', 'Microsoft Entra ID', 'Key vaults', and 'More services'. A search bar is located above the main content area. The main area is titled 'Resources' and has two tabs: 'Recent' (selected) and 'Favorite'. It lists various Azure resources with their names, types, and last viewed times. Below this is a 'See all' link. Further down are sections for 'Navigate' (with links to Subscriptions, Resource groups, All resources, and Dashboard) and 'Tools' (with links to Microsoft Learn, Azure Monitor, Microsoft Defender for Cloud, and Cost Management). At the bottom, there are 'Useful links' and 'Azure mobile app' sections, along with a taskbar at the very bottom.

The directories that I made inside the data lake

The screenshot shows the Microsoft Azure Storage Container blade for a container named 'dataser'. The left sidebar includes navigation links for Home, Storage, Container, and ContainerMenuBlade. The main area is titled 'Overview' and shows an authentication method of 'Access key' and a location of 'dataser'. It features a search bar and a 'Show deleted objects' button. Below this is a table listing blobs by prefix (case-sensitive), with three entries: 'row data', 'synapse', and 'transformd data'. The table columns include Name, Modified, Access tier, Archive status, Blob type, Size, and Lease state. At the bottom, there are tabs for 'Overview', 'Diagnose and solve problems', 'Access Control (IAM)', 'Settings', 'Shared access tokens', 'Manage ACL', 'Access policy', 'Properties', and 'Metadata'. A taskbar is visible at the bottom of the screen.

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

The screenshot shows the Microsoft Azure Data Factory interface. On the left, the 'Factory Resources' sidebar is open, showing a single pipeline named 'data-ingestion'. In the main workspace, a 'Copy data' activity is selected. The 'Source' tab is active, showing the 'AmazonDLS' dataset as the source. The 'Sink' tab is also visible. A preview window on the right shows a sample of the 'Reviews.csv' file data.

ID	ProductId	Userid	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	
1	1	8001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1

This screenshot shows the same Microsoft Azure Data Factory interface, but with a different configuration. The 'Sink' tab is now active, indicating the target dataset is 'transformed_data_sink'. Other sink settings like 'Copy behavior' (Flatten hierarchy), 'Max concurrent connections', and 'File extension' (txt) are visible. The rest of the pipeline structure and preview window remain the same as the first screenshot.

Azure data bricks and visualization using python

The screenshot shows the Azure Databricks interface with a Python notebook titled "Amazon_Review_Transformation". The notebook contains the following code:

```
1 Last execution failed
2 configs = [
3     "fs.azure.account.auth.type": "OAuth",
4     "fs.azure.account.oauth.provider.type": "org.apache.hadoop.fs.azurebfs.oauth2.ClientCredsTokenProvider",
5     "fs.azure.account.oauth2.client.id": "f8809d9c-e076-4033-9c33-5a834a0718ff", # Your Client ID
6     "fs.azure.account.oauth2.client.secret": "IB3Qo-91fHx0qv-eyXzv713db-jV8sytfdvubabou", # Your Client Secret
7     "fs.azure.account.oauth2.client.endpoint": "https://login.microsoftonline.com/
8         0bc92751-01a-4e2c-a48b-633280fe574/auth2/token" # Your Tenant ID
9 ]
10
11 dbutils.fs.mount(
12     source = "abfss://dataser@amazonstoreg.dfs.core.windows.net", # container@storageaccount
13     mount_point = "/mnt/amazonstoreg",
14     extra_configs = configs
15 )
16
17 > ExecutionError: An error occurred while calling o419.mount.
18 : java.rmi.RemoteException: java.lang.IllegalArgumentException: requirement failed: Directory already mounted: /mnt/amazonstoreg; nested exception is: ...
19
20 Diagnose error Debug Assistant Quick Fix ON
```

The notebook also displays a table view of data from the mounted storage:

#	path	name	size	modificationTime
1	dbfs/mnt/amazonstoreg/raw data/	row data/	0	1729092446000
2	dbfs/mnt/amazonstoreg/transforemd dat...	transforemd dat...	0	1729092466000

Below the table, it says "2 rows | 0.82 seconds runtime" and "Refreshed 2 days ago".

The right side of the interface includes an "Assistant" panel with some error messages and code snippets, and a "Comments" section at the bottom.

```
spark
SparkSession - hive
SparkContext
SparkUI
Version
v3.5.0
Master
local[*], 4
AppName
Databricks Shell
```

```
Reviews = spark.read.format("csv").option("header", "true").option("inferSchema", "true").load("/mnt/AmazonStore/reviews.csv")
```

```
Reviews.printSchema()
```

```
Reviews = Reviews.na.drop() # Drop rows with any null values
```

```
Reviews = Reviews.dropDuplicates() # remove duplicates
```

```
Reviews = Reviews.filter(Reviews['Score'] > 3) # filter out unwanted data
```

Amazon_Review_Transformation

```
Reviews = spark.read.format("csv").option("header", "true").option("inferSchema", "true").load("/mnt/amazonstoreg/reviews.csv")  
Reviews.show()
```

Review ID	Product ID	Score	Text
1	1 1339545600	1	y cats love thi...[One of my boys ne...]
2	2 1288915200	2	[...] 13 80000X1V50G A3ZPC1T23W9B LT Cats Are...[One of my boys ne...]
3	3 1268352000	3	[...] 14 80000X1V50M A1BECVXK3TAE "Willie ""roadie"" fresh and greasy! [good flavor! thes...]
4	4 1260440000	4	[...] 15 80000V153M A2PQMV2V2TQ04X7 "Yonic ""Oh HELL... raspberry Twizz...[The Strawberry Tw...]
5	5 1262044000	5	[...] 16 80000V153M A1C2Z3CP81QJ Brian A. Lee [...] 17 80000V153M A1KLWf64S9Bn [Brian loves...]
6	6 1264044000	6	[...] 18 80000V153M AFKX4U972Q0Q Erica Neethery poor taste!] I love eating the...[...]
7	7 1264044000	7	[...] 19 80000V153M A2A9X5862G1P Becca Love it! I am very satisfy...[...]
8	8 1264044000	8	[...] 20 80000V153M A13V1CLC1XK2W Wolfie [Becca] I love it! I am very satisfy...[...]
9	9 1264044000	9	[...] 21 80000V153M A13V1CLC1XK2W Greg we delivered tw...[Candy was deliver...]
10	10 1264044000	10	[...] 22 80000V153M A13V1CLC1XK2W Greg only showing top 20 rows

2 days ago (x1)

Run all Terminated Schedule Share

Workspace Recents Catalog Workflows Compute SQL Editor Queries Dashboards Alerts Query History SQL Warehouses Data Engineering Job Runs Data Ingestion Delta Live Tables Machine Learning Playground Experiments Features Models Serving Marketplace Partner Connect

Search data, notebooks, recent, and more... CTRL + P

amazon-databricks Assistant

28°C jesusus 15:41 US 19/10/2024

The screenshot shows a Microsoft Edge browser window with several tabs open. The active tab is a Jupyter Notebook titled "Amazon_Review_Transformation" running on a Databricks workspace. The notebook contains four code cells:

```
Reviews = spark.read.format("csv").option("header", "true").load("path/to/reviews.csv")
Reviews = Reviews.na.drop() # Drop rows with any null values
Reviews = Reviews.dropDuplicates() # Remove duplicates
Reviews = Reviews.filter(Reviews['Score'] > 3) # Filter out unwanted data
```

A sidebar on the left lists various Databricks workspace sections such as Workspace, Catalog, Workflows, Compute, SQL, and Machine Learning. A sidebar on the right features an "Assistant" section with a message about the "Reviews" DataFrame being undefined.

Amazon Review Transformation Python

```
# Summary statistics: Check average score
Reviews.groupby('ProductId').agg({'Score': 'avg'}).show()
```

(3) Spark Jobs

ProductId	avg(Score)
B00NEBLBNG	4.85
B0016GZ6ZG	5.0
B00KCFJOMY	4.714285714285714
B000SQNE	4.51
B0012XKQH	4.714285714285714
B0012YEKCH	4.818181818181818
B0016GOFQI	4.833333333333333
B001EL1L364	4.75
B00NPBRK8	4.9
B00NMFGSUY	5.0
B0016E3XG	4.731787317070707
B0011GOKWV	5.0
B0016GOKWV	4.76937069230769
B0018CE6Q0	4.925525525525526
B0052JNHR0	5.0
B0088LIVWAO	4.75
B0026ABHSM	4.555555555555555
B0015DYM0	4.8!

(2) Spark Jobs

```
# Save processed data
Reviews.write.format("csv").mode("overwrite").option("header", "true").save("/mnt/amazonstoreg/transfrend data/reviews_cleaned.csv")
```

(2) Spark Jobs

```
# 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)
```

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 Python

```
# Save processed data
Reviews.write.format("csv").mode("overwrite").option("header", "true").save("/mnt/amazonstoreg/transfrend 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 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

```
Reviews: pyspark.sql.DataFrame.DataFrame = [id: integer, ProductId: string ... 9 more fields]
```

Amazon Review Transformation Python

```

# average_sentiment = Reviews.groupby('ProductId').agg(['sentiment': 'avg']).alias("Average Sentiment")
average_sentiment.show()

```

(3) Spark Jobs

```

# average_sentiment = pyspark.sql.dataframe.DataFrame = [ProductId: string, avg(sentiment): double]
|B00001LBMG| 0.622664285742857 |
|B00001J0WY| 0.858864285742858 |
|B00012YKXCM| 0.8222242424242425 |
|B00010QFQ1| 0.8594 |
|B00011L364| 0.6950800000000001 |
|B0000PBRW8| 0.7726700000000001 |
|B0000WGSUY| 0.6244200000000001 |
|B0001E3XH| 0.7567756097569979 |
|B0001E6DXM| 0.6584557142857143 |
|B0001FG5XX| 0.7521 |
|B00018CEQQ| 0.7248555555555556 |
|B000523NVO| 0.9067666666666666 |
|B00083IVAO| 0.8011 |
|B0002G4B85M| 0.8721666666666666 |
|B0001E6XK| 0.753686 |
|B00015Q1AE| 0.7786599980000001 |
|B00020G89MD| 0.8527 |
+-----+
only showing top 28 rows

```

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 Python

```

display(average_sentiment)

```

(3) Spark Jobs

ProductId	1.2 avg(sentiment)
B00001LBMG	0.80662
B0001E6DXM	0.765
B00001J0WY	0.622664285742857
B0000PBRW8	0.8594
B00011L364	0.6950800000000001
B0000WGSUY	0.6244200000000001
B0001E3XH	0.7567756097569979
B0001FG5XX	0.7521
B00018CEQQ	0.7248555555555556
B000523NVO	0.9067666666666666
B00083IVAO	0.8011
B0002G4B85M	0.8721666666666666
B0001E6XK	0.753686
B00015Q1AE	0.7786599980000001
B00020G89MD	0.8527
+-----+	
only showing top 10000+ rows Truncated data due to row limit 1.32 minutes runtime	
Refreshed 2 days ago	

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 Python

```
# 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/amazonstore/transforemd_data/reviews_with_sentiment")

# (9) Spark Jobs
# 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)
```

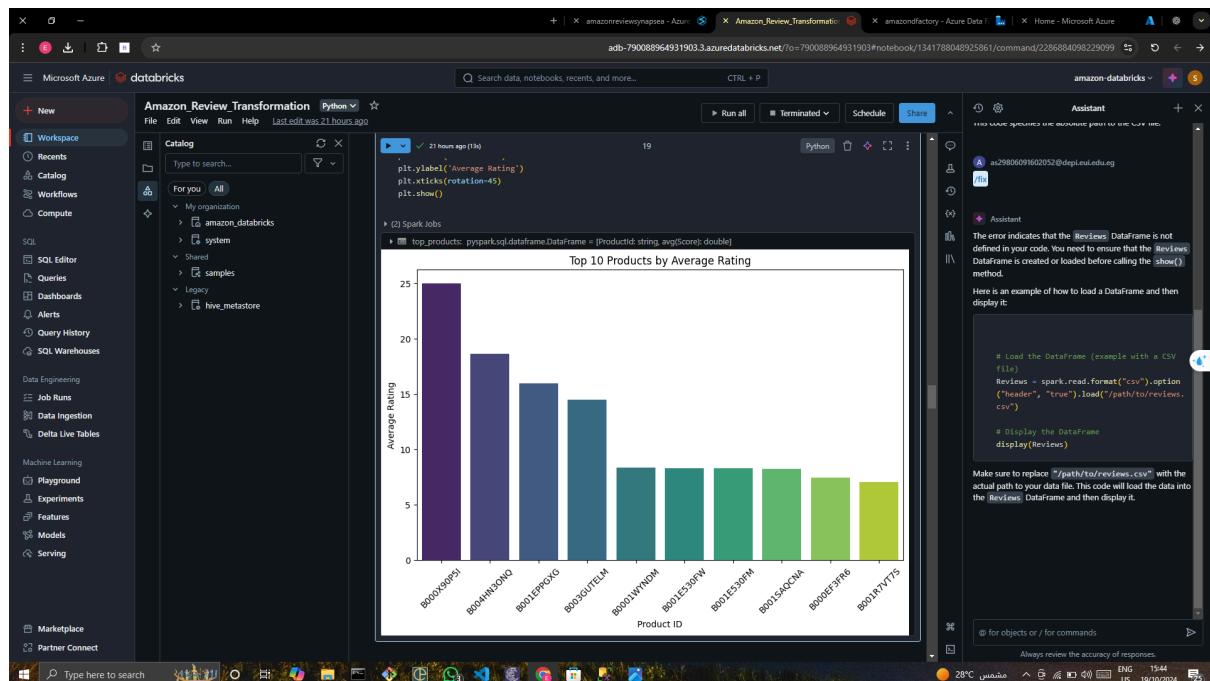
The error indicates that the `Reviews` DataFrame is not defined in your code. You need to ensure that the `Reviews` DataFrame is created or loaded before calling the `show()` method.

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 Python

File Edit View Run Help Last edit was 21 hours ago

Search data, notebooks, recents, and more... CTRL + P

Run all Terminated Schedule Share

Catalog Type to search... 21 hours ago (14) plt.show() Python

(1) Spark Jobs

Reviews: pyspark.sql.dataframe.DataFrame = [id: integer, ProductId: string ... 9 more fields]

Helpfulness Ratio Distribution

Frequency

Helpfulness Ratio

21 hours ago (9) 22

```
from pyspark.sql.functions import when, col, length # Import when and other required functions

# Create a new column for review sentiment (positive/negative)
```

Display the DataFrame

display(Reviews)

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.

for objects or / for commands Always review the accuracy of responses.

Amazon Review Transformation Python

File Edit View Run Help Last edit was 21 hours ago

Search data, notebooks, recents, and more... CTRL + P

Run all Terminated Schedule Share

Catalog Type to search... 21 hours ago (9) plt.show() Python

(2) Spark Jobs

top_users: pyspark.sql.dataframe.DataFrame = [UserId: string, count: long]

Top 10 Users by Number of Reviews

Number of Reviews

User ID

21 hours ago (10) 21

```
from pyspark.sql.functions import when, col, length # Import when and other required functions

# Create a new column for review sentiment (positive/negative)

# Display the DataFrame
```

display(Reviews)

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.

for objects or / for commands Always review the accuracy of responses.

Amazon Review Transformation Python

```

plt.title('Positive vs Negative Reviews')
plt.axis('equal')
plt.show()

```

Positive vs Negative Reviews

Category	Percentage
Positive	77.8%
Negative	22.2%

```

# Sort by helpfulness ratio and select the top 10 most helpful reviews
top_helpful_reviews = Reviews.orderBy("HelpfulnessRatio", ascending=False).limit(10)

```

Amazon Review Transformation Python

```

# 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)

```

Top 10 Users by Average Review Length

User ID	Average Review Length (Characters)
#sc-r1z5tL9rA5E	~12,000
A3C02JURWQ2F9L	~7,500
A1V7R04Z3jG6e6	~7,000
A1Gc0mRvZ2Wq9	~6,800
AshXiiWWL32L0	~6,500
AGENMPREATES4	~6,200
AhWw6M0Wk5Bw7	~6,000
A2H2E200rBC268	~5,800
#sc-90777BfRNk43	~5,500
A1f069jd9c0C8	~5,200

Amazon Review Transformation Python

File Edit View Run Help Last edit was 21 hours ago

Search data, notebooks, recents, and more... CTRL + P

(2) Spark Jobs

most_reviewed_products: pyspark.sql.DataFrame = [ProductID: string, count: long]

Top 10 Most Reviewed Products

Number of Reviews

Product ID

21 hours ago (1h)

Convert to Pandas DataFrame for plotting
Reviews_df = Reviews.toPandas()

26

Python

Run all Terminated Schedule Share

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.

for objects or / for commands Always review the accuracy of responses.

28°C 19/10/2024 ENG US

Amazon Review Transformation Python

File Edit View Run Help Last edit was 21 hours ago

Search data, notebooks, recents, and more... CTRL + P

(1) Spark Jobs

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)

21 hours ago (1h)

26

Python

Run all Terminated Schedule Share

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.

for objects or / for commands Always review the accuracy of responses.

28°C 19/10/2024 ENG US

Screenshot of a Microsoft Azure Databricks notebook titled "Amazon Review Transformation" running Python code.

The notebook displays the following Python code:

```
sns.barplot(data=avg_score_by_product_df, x='ProductID', y='avg(score)', palette='cividis')
plt.title('Top 10 Products by Average Rating')
plt.xlabel('Product ID')
plt.ylabel('Average Score')
plt.xticks(rotation=45)
plt.show()
```

The resulting bar chart is titled "Top 10 Products by Average Rating". The Y-axis is labeled "Average Score" and ranges from 0 to 25. The X-axis is labeled "Product ID" and lists ten product IDs: B000X9095J, B00MN3QNG, B001EP0XG, B003GUZELM, B001WYNDM, B001ES3PFW, B001E59RM, B00154CCNA, B000F3PR8, and B001A7T75. The chart shows that the highest average score is approximately 25.5 for Product ID B000X9095J, followed by B00MN3QNG at approximately 18.5.

The Assistant panel on the right provides a tip about loading DataFrames:

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)
```

A note below the code says: "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."