CS_ELEC_3C | Oct 22, 2025

LAB ACTIVITY - TEXT ANALYTICS

Baylon, Karyle Zhienelle V.
4CSD

Learning Objectives

- Upload and explore Amazon review data.
- Produce descriptive statistics and visualizations
- Generate a word cloud and automated explanations.
- Create a sentiment analysis and a Topic Analysis.

PART 1. AMAZON DATASET

Purpose

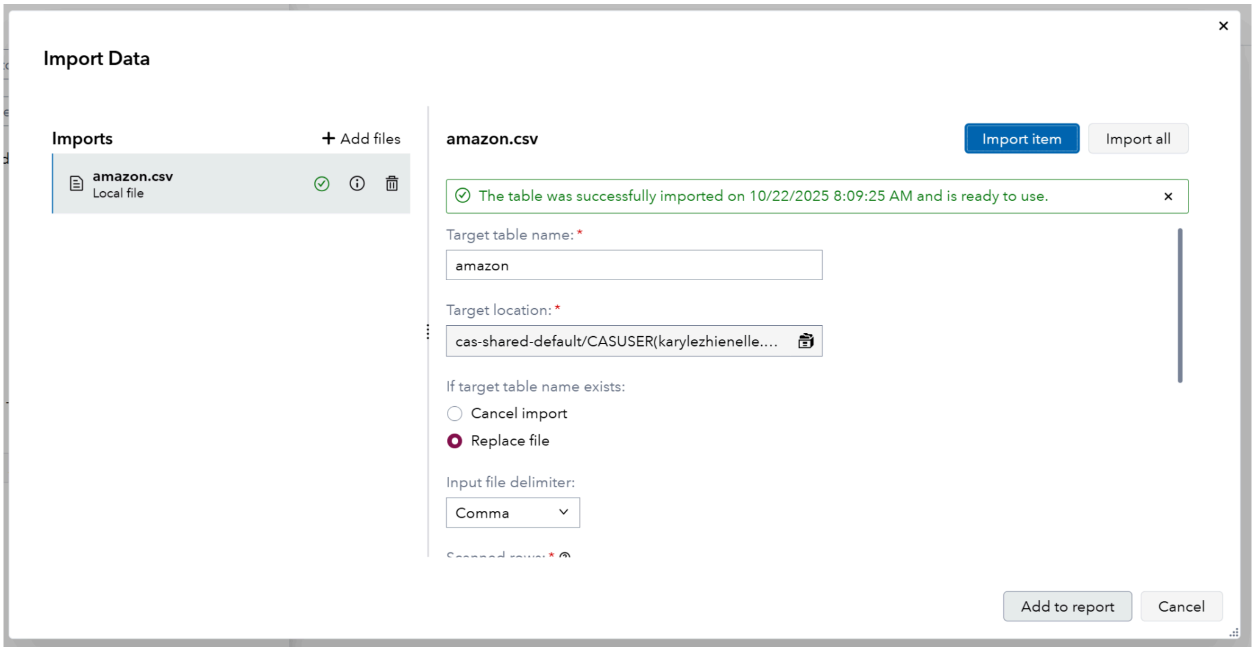
You've just been hired as a junior data analyst at Amazon Marketplace Analytics, and your first task is to explore customer reviews and uncover what drives positive product ratings.
In this SAS On-the-Job (OTJ) activity, you will use SAS Viya 4 Visual Analytics to:

- explore data visually, and
- perform text analytics on review text

1️⃣Upload the Dataset

1. From SAS Drive, open Applications ▸ Explore and Visualize.
2. Click New Report → Import data → Local files.
3. Upload your amazon.csv.
4. Target location: CASUSER.

5. If file exists → *Replace file.*
6. Click Import item → Add to report.



amazon.csv
Local file

amazon.csv

The table was successfully imported on 10/22/2025 8:09:25 AM and is ready to use.

Target table name: *
amazon

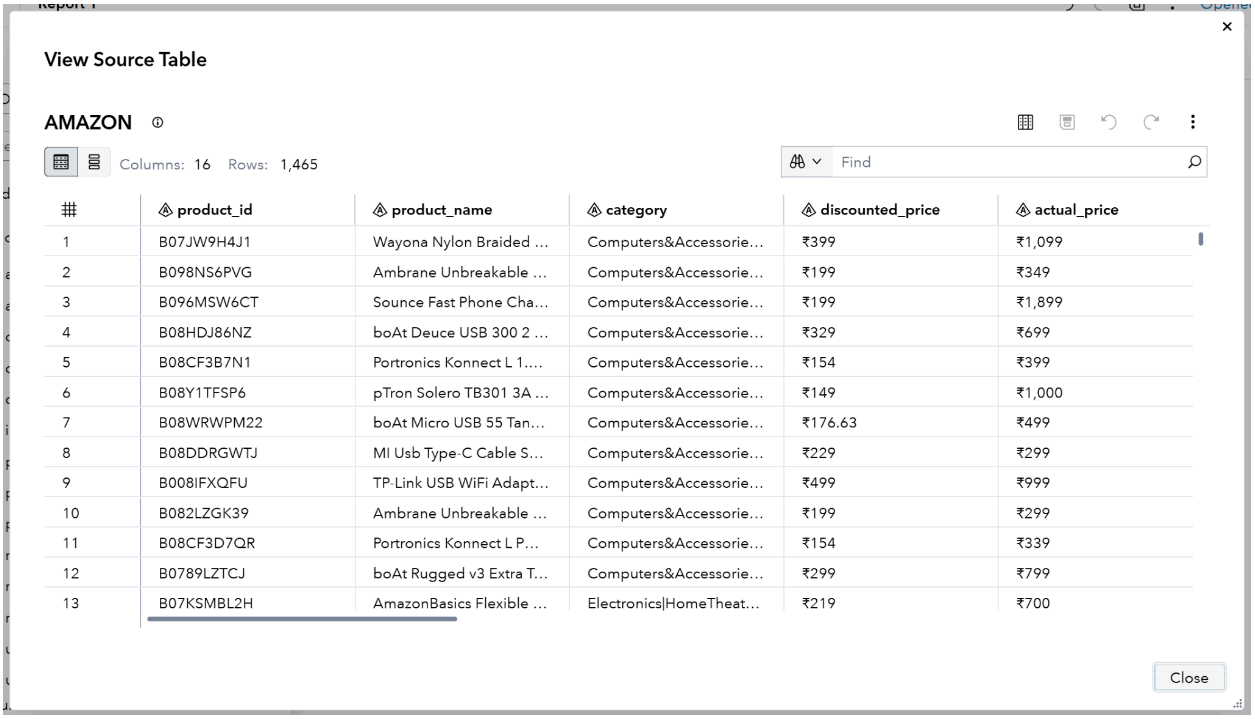
Target location: *
cas-shared-default/CASUSER(karylezhienelle....

If target table name exists:
○ Cancel import
● Replace file

Input file delimiter:
Comma

2️⃣Explore the Data

- Open the Data source pane ▸ View Source Table.
- Observe variables (e.g., reviewText, rating, price, category, brand).
- Identify:
  - The text field to analyze (reviewText).
  - The target variable (rating).



AMAZON

Columns: 16 Rows: 1,465

#	product_id	product_name	category	discounted_price	actual_price
1	B07JW9H4J1	Wayona Nylon Braided …	Computers&Accessorie…	₹399	₹1,099
2	B098NS6PVG	Ambrane Unbreakable …	Computers&Accessorie…	₹199	₹349
3	B096MSW6CT	Sounce Fast Phone Cha…	Computers&Accessorie…	₹199	₹1,899
4	B08HDJ86NZ	boAt Deuce USB 300 2 …	Computers&Accessorie…	₹329	₹699
5	B08CF3B7N1	Portronics Konnect L 1.…	Computers&Accessorie…	₹154	₹399
6	B08Y1TFSP6	pTron Solero TB301 3A …	Computers&Accessorie…	₹149	₹1,000
7	B08WRWPM22	boAt Micro USB 55 Tan…	Computers&Accessorie…	₹176.63	₹499
8	B08DDRGWTJ	MI Usb Type-C Cable S…	Computers&Accessorie…	₹229	₹299
9	B008IFXQFU	TP-Link USB WiFi Adapt…	Computers&Accessorie…	₹499	₹999
10	B082LZGK39	Ambrane Unbreakable …	Computers&Accessorie…	₹199	₹299
11	B08CF3D7QR	Portronics Konnect L P…	Computers&Accessorie…	₹154	₹339
12	B0789LZTCJ	boAt Rugged v3 Extra T…	Computers&Accessorie…	₹299	₹799
13	B07KSMBL2H	AmazonBasics Flexible …	Electronics|HomeTheat…	₹219	₹700

Insight:

- **Which variables might influence product ratings most (price, brand, sentiment, etc.)?**
  - Upon exploring the columns, I think that **price** and **sentiment** dominate the influence on product ratings. Because a lower discounted price may cause the consumers to give a higher rating, and the sentiments of others may also affect their own comments about a specific product.

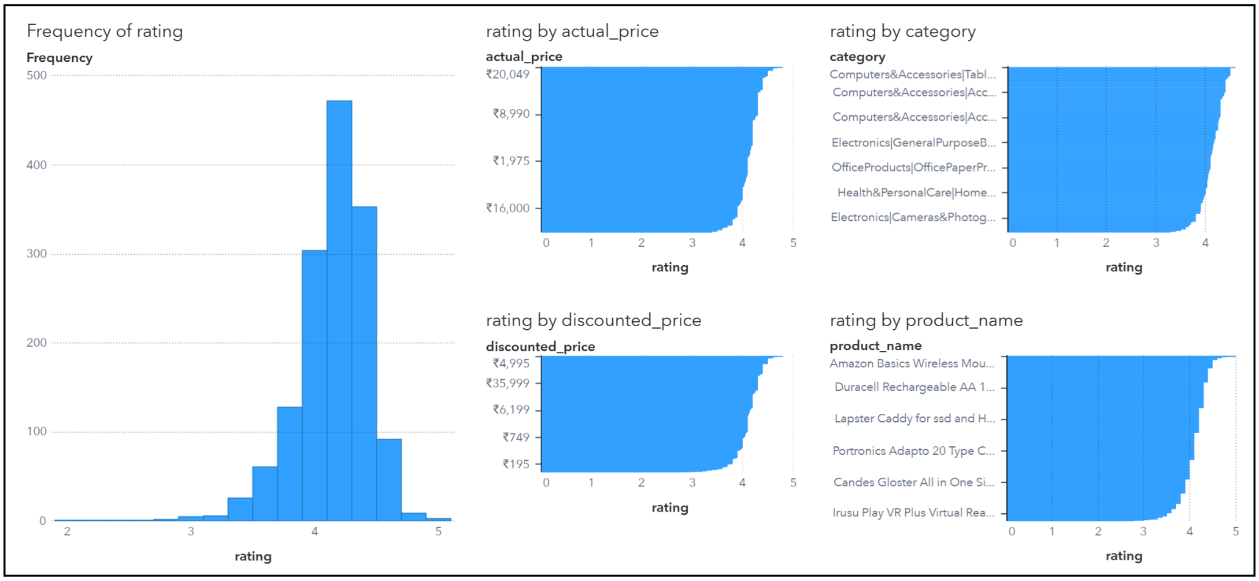
---

3 Descriptive Visuals

Create a dashboard page named DescriptiveStats:

- Drag rating → canvas (+ Auto Chart → Histogram).
- Add price, brand, and category → separate Auto Charts.

- Use bar charts or box plots to compare ratings across categories.



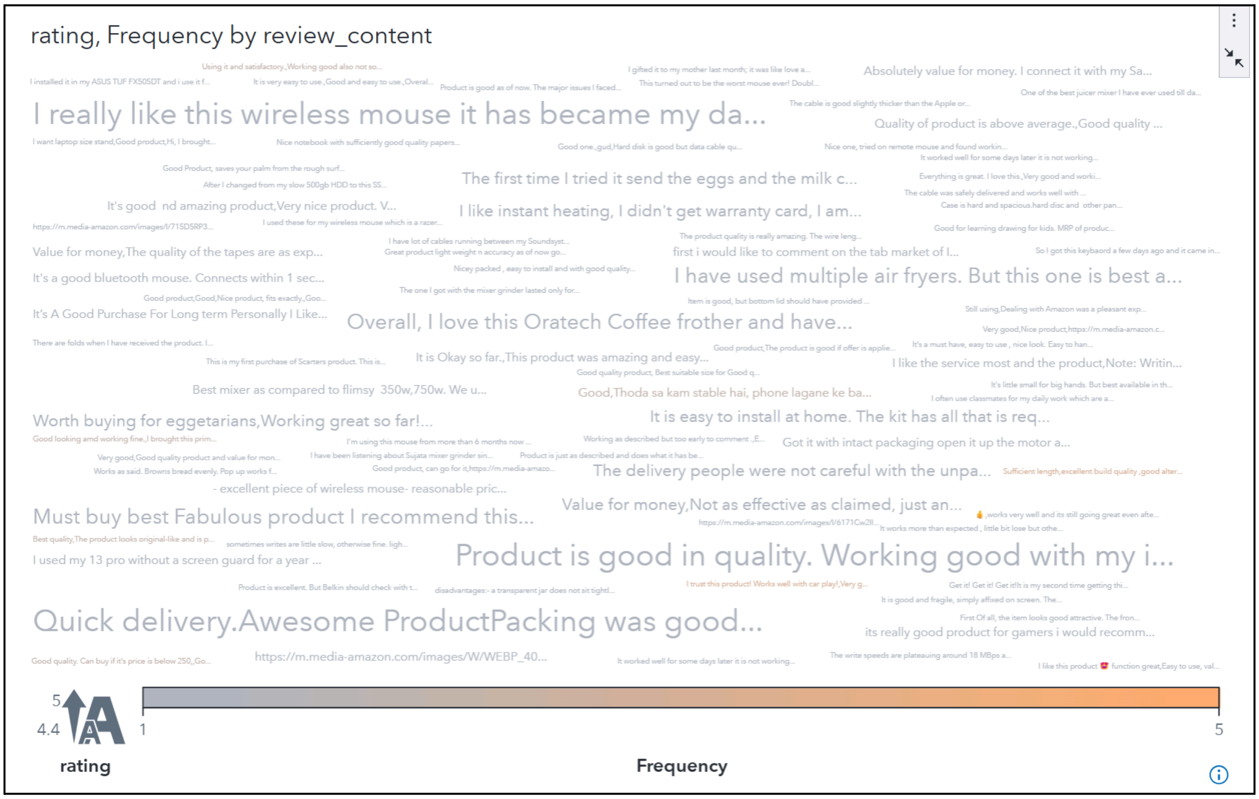
Insight:

- **Which product categories have the highest average ratings?**
  - Products under the **Computers and Accessories** category, especially Tablets, have the highest average ratings among the others, while several **Electronics** categories also show relatively high average ratings.
- **Do higher-priced items tend to receive better reviews?**
  - From the chart, there is a **slight positive relationship** showing that most ratings cluster around **4–5**, while very low ratings occur across a wide price range. Some higher-priced items are more likely to have higher ratings, but the relationship is not strong and should be quantified before making a firm conclusion.

4️⃣Word Cloud Page

- Add new page → drag Word Cloud object.

- Assign:
  - Word: reviewText
  - Size: rating (Aggregation = Average)
- Under Style ▸ Color Gradient → choose from gray to orange.



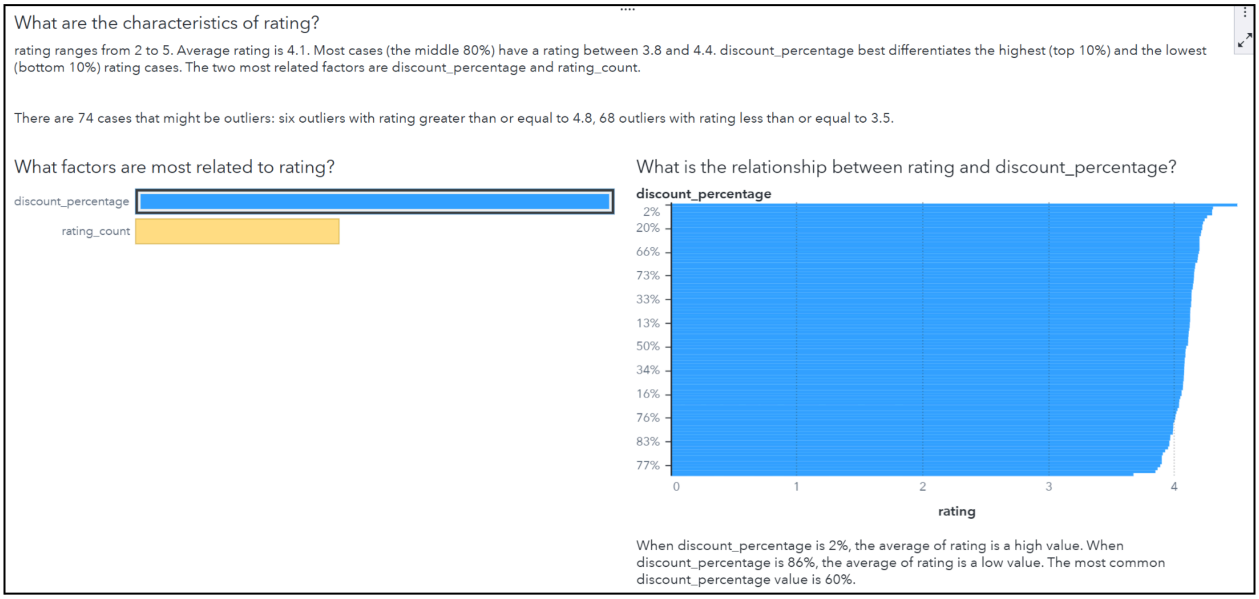
Insight:

- **Which words dominate positive reviews?**
  - Positive reviews frequently use phrases such as **"I love this"**, **"I like this"**, **"good quality"**, and **"quick delivery"**, indicating strong satisfaction and product approval.
- **Do negative reviews share common terms?**
  - Negative reviews often include phrases like **"product didn't work"** and **"costly"**, indicating functional failures and price dissatisfaction are common complaints. However, these

comments may not be visible on the chart since positive reviews dominate, showing that most customers express favorable opinions about their purchases.

5️⃣Automated Explanation

- Add new page → Automated Explanation object.
- Response variable: rating.
- Examine which predictors (price, brand, review length) best explain the rating.



What are the characteristics of rating?

rating ranges from 2 to 5. Average rating is 4.1. Most cases (the middle 80%) have a rating between 3.8 and 4.4. discount_percentage best differentiates the highest (top 10%) and the lowest (bottom 10%) rating cases. The two most related factors are discount_percentage and rating_count.

There are 74 cases that might be outliers: six outliers with rating greater than or equal to 4.8, 68 outliers with rating less than or equal to 3.5.

What factors are most related to rating?

What is the relationship between rating and discount_percentage?

When discount_percentage is 2%, the average of rating is a high value. When discount_percentage is 86%, the average of rating is a low value. The most common discount_percentage value is 60%.

Insight:

- **What top three variables explain customer satisfaction?**
  - Based on the graph, the top two variables explaining customer satisfaction are discount_percentage and rating_count. **Discount percentage** is the strongest predictor, but shows an inverse relationship in which when discounts are 2%, ratings are

high, but when discounts reach 83%, ratings are low, suggesting customers may perceive heavily discounted items as lower quality or clearance products. **Rating count** is the second most important factor, as products with more ratings typically indicate greater popularity, social proof, more established quality, and better visibility.

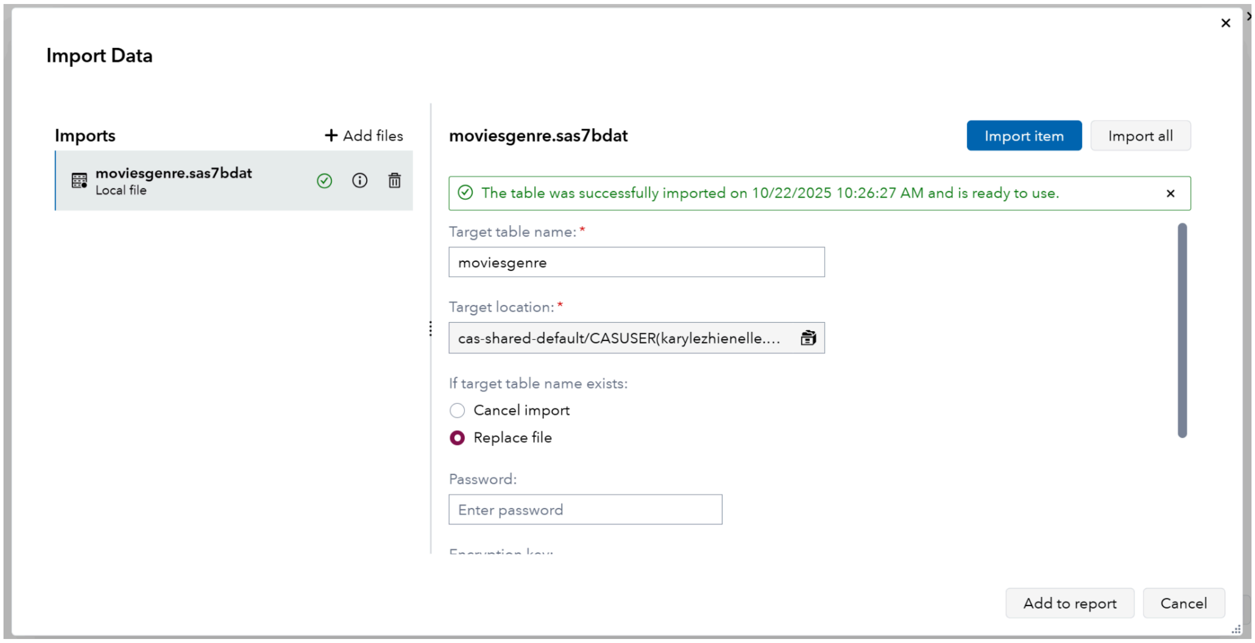
Save the report as AmazonTextAnalysis.

PART 2. MOVIES GENRE DATASET

For the movie genre dataset, we can perform a similar analysis, but here, we will choose different variables relevant to movies.

1️⃣Upload the Dataset

- Upload your movie genre dataset
  - Similar to the Amazon dataset, use the **Local files** option in SAS Viya.



Import Data

Imports

moviesgenre.sas7bdat
Local file

moviesgenre.sas7bdat

The table was successfully imported on 10/22/2025 10:26:27 AM and is ready to use.

Target table name: * moviesgenre

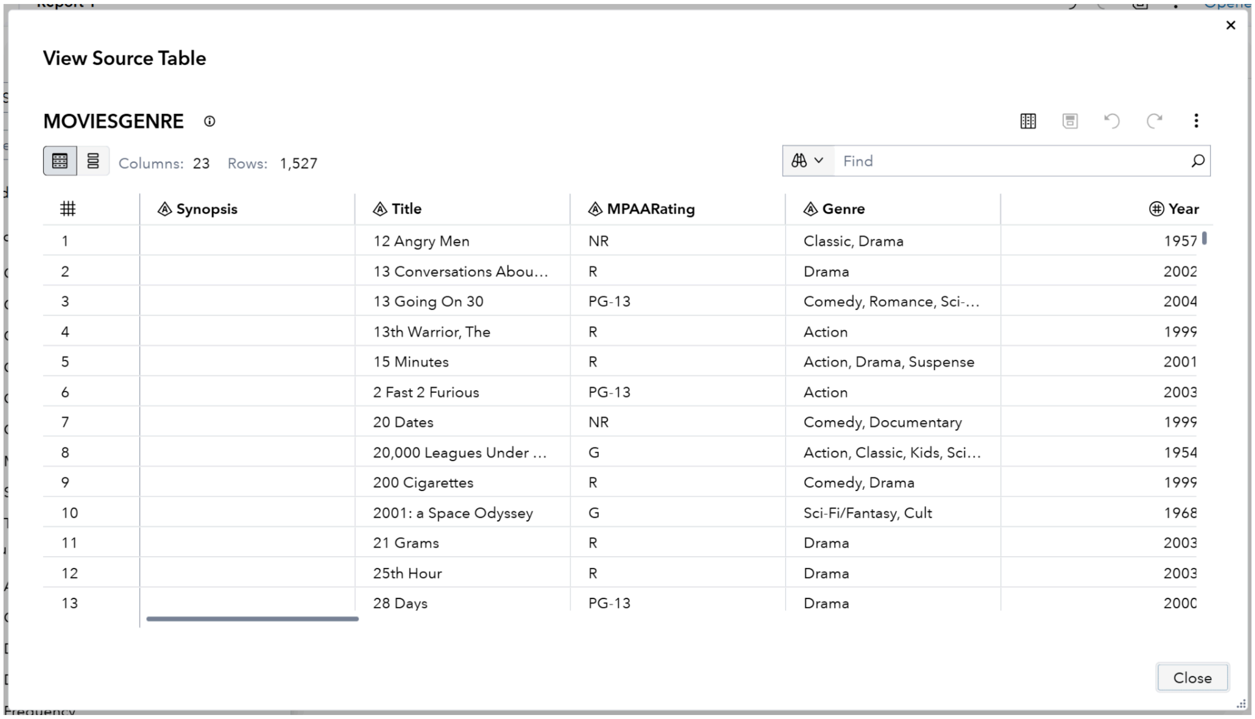
Target location: * cas-shared-default/CASUSER(karylezhienelle....

If target table name exists:
○ Cancel import
● Replace file

Password: Enter password

2️⃣Explore the Data

- **Step:** Identify key variables in the movie dataset. For movie ratings, we might be looking at:
  - **genre**: The movie genre (comedy, action, drama, etc.).
  - **ViewerRating**, **Size**, **NumGenres**, **Genre**, and **MPAARating**



View Source Table

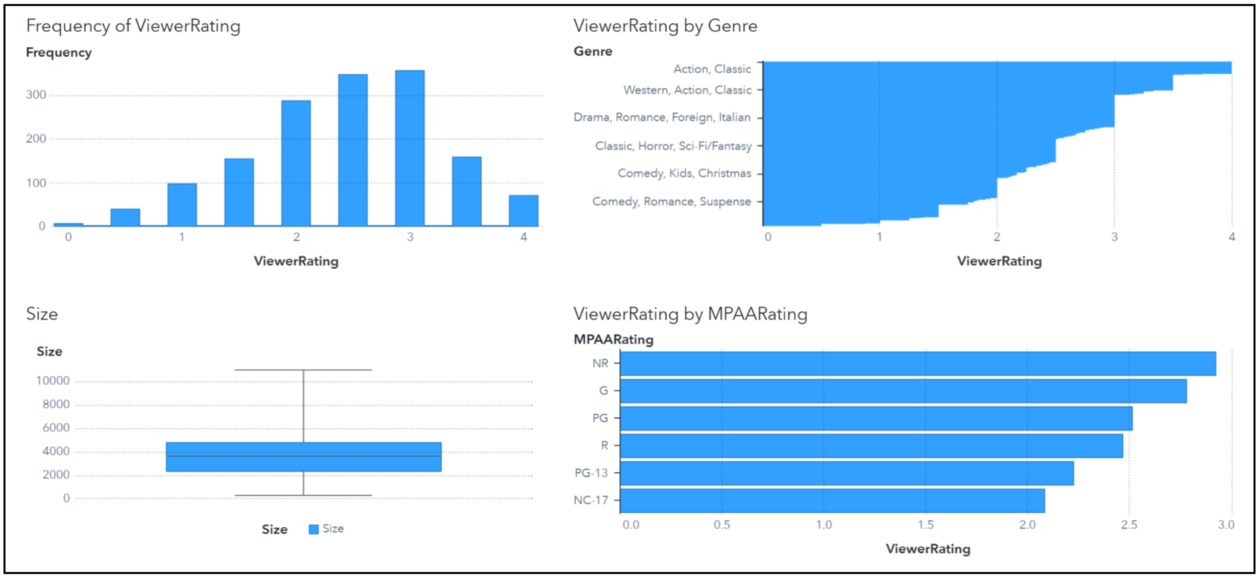
MOVIESGENRE

Columns: 23 Rows: 1,527

#	Synopsis	Title	MPAARating	Genre	Year
1		12 Angry Men	NR	Classic, Drama	1957
2		13 Conversations Abou…	R	Drama	2002
3		13 Going On 30	PG-13	Comedy, Romance, Sci-…	2004
4		13th Warrior, The	R	Action	1999
5		15 Minutes	R	Action, Drama, Suspense	2001
6		2 Fast 2 Furious	PG-13	Action	2003
7		20 Dates	NR	Comedy, Documentary	1999
8		20,000 Leagues Under …	G	Action, Classic, Kids, Sci…	1954
9		200 Cigarettes	R	Comedy, Drama	1999
10		2001: a Space Odyssey	G	Sci-Fi/Fantasy, Cult	1968
11		21 Grams	R	Drama	2003
12		25th Hour	R	Drama	2003
13		28 Days	PG-13	Drama	2000

3 Descriptive Visuals

- **Histogram** for **ViewerRating** to see the distribution of ratings across all movies.
- **Bar charts** for **Genre** and **MPAARating** to compare ratings by genre and movie rating.
- **Box plot** for **Size** to explore if larger movies (in terms of reach) tend to get better ratings



Frequency of ViewerRating

ViewerRating by Genre

Size

ViewerRating by MPAARating

Insights from the Visualization:

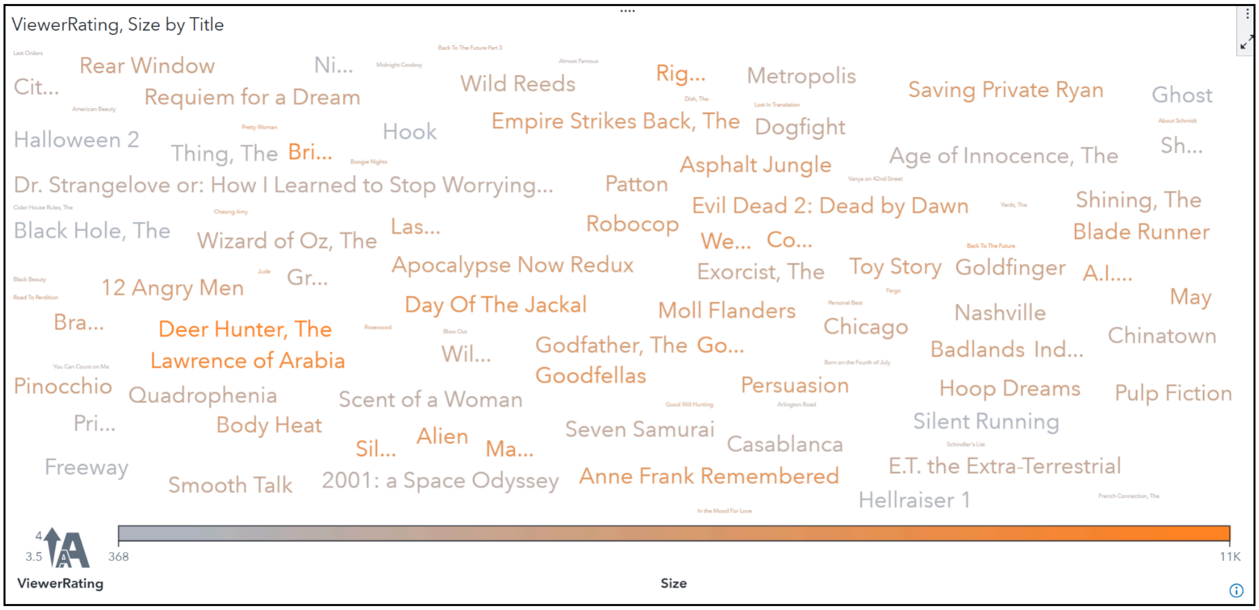
- **ViewerRating Insights**
  - Movie ratings follow a roughly normal pattern with a slight left skew. Most films fall in the **2–3 range**, showing that audiences generally rate movies as **average to good**. Very few receive near-zero scores, and perfect **4-star ratings** are rare, suggesting that outstanding films are uncommon.
- **Genre Insights**
  - Ratings vary widely by genre. **Action and Classic** films perform the best, often scoring between **3–4**, showing broad audience appeal. In contrast, **Comedy, Romance, and Suspense** movies tend to rate lower and show little variation, implying they struggle to impress viewers. **Drama, Romance, and Foreign** and **Comedy, Kids, and Christmas** genres perform moderately and show mixed results.
- **Size Insights**
  - Movie size (in terms of reach) shows a skewed pattern: most films fall between **~4,000**, with a few large-scale outliers above **10,000**.

The median is around **3,600**, suggesting most productions are modest, while a handful of **blockbusters** reach much wider audiences and often earn higher ratings.

- **MPAA Insights**
    - Viewer satisfaction differs by rating category. **NR** (Not Rated) and **G** films earn the highest average scores (around **3.0**), appealing to families or niche audiences. **PG** and **R** films rate moderately (**2.2–2.8**), while **PG-13** and **NC-17** movies perform worst (below **2.0**), suggesting teen and adult-only content often fails to meet viewer expectations.

4️⃣Word Cloud

- **Step:** Create a **Word Cloud/Topic Analysis/ Sentiment Analysis** visualization for Title with **ViewerRating as the size (average ViewerRating).**
    - Color gradient from **gray to orange**.

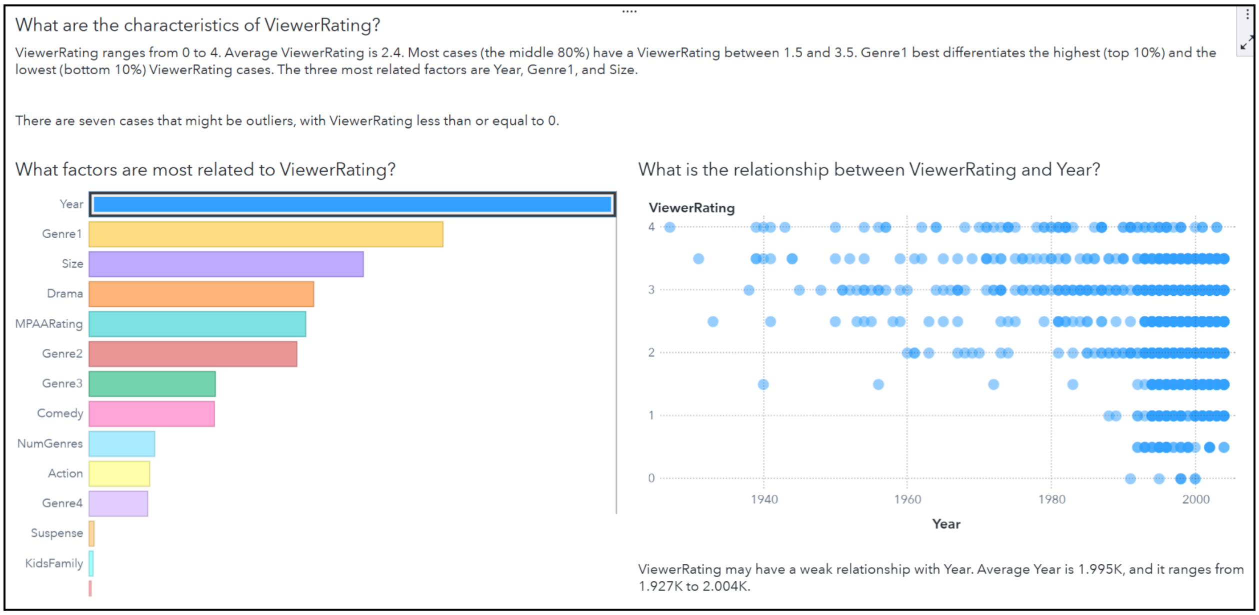


Insight: *(These questions cannot be answered as there is no column provided for movie reviews.)*

- What words tend to dominate in **positive movie reviews** (e.g., “amazing,” “thrilling”)?
- What words appear more often in **negative reviews** (e.g., “boring,” “predictable”)?

5️⃣Automated Explanation

- **Step:** Add an **Automated Explanation** for **rating**.
    - Find out which variables explain the ViewerRating.



What are the characteristics of ViewerRating?

ViewerRating ranges from 0 to 4. Average ViewerRating is 2.4. Most cases (the middle 80%) have a ViewerRating between 1.5 and 3.5. Genre1 best differentiates the highest (top 10%) and the lowest (bottom 10%) ViewerRating cases. The three most related factors are Year, Genre1, and Size.

There are seven cases that might be outliers, with ViewerRating less than or equal to 0.

What factors are most related to ViewerRating?

What is the relationship between ViewerRating and Year?

ViewerRating may have a weak relationship with Year. Average Year is 1.995K, and it ranges from 1.927K to 2.004K.

Insight:

- **What are the top 3 variables explaining movie ratings?**
    - Based on the graph, the top three factors influencing movie ratings (ViewerRating) are Year, Genre1, and Size, respectively. **Year** is the strongest predictor, as older films from the

1940s–1960s tend to have higher ratings (around 3–4 stars), while movies released after 2000 show a wider spread across the 0–4 range, possibly due to changing rating standards or survivorship bias, where only the best older films remain. **Genre1**, ranked second, shows that a movie's primary genre has a strong influence on ratings, such that some genres consistently earn higher scores due to fans and viewers liking them more, while others get mixed reactions. **Size**, the third most important factor, shows that films that reach larger audiences often earn higher ratings, likely because they gain more exposure, attract wider interest, and benefit from stronger marketing efforts.