

Amazon Products

Data Integration Spring 2025

Theme

The goal of the project is to analyze the Amazon Products and Reviews Data to understand consumer behavior and sentiment in online shopping.

Row	rating	title	text	images	asin
772	5.0	4 Milk Jugs In the Side Door! Spacious, Versatile and Attractive	My wife and I actually had a fridge that was working perfectly fine, but we just had our fourth kid in five years and suffice it to say	{small_image_url': 'https://images-na.ssl-images- amazon.com/images/I/61I8ih+uVSL._SL256_.jpg',	B003M5L284

Row	rating	title	text	images	asin
42	5.0	nice product	Have always used burner covers on my electric stove. These are nice and give a clean look! Nice quality and couldn't	0 rows	B0000CFPK8

Row	rating	title	text	images	asin
1	4.0	Four Stars	Just what was needed.	0 rows	B00004YWK2

Data Sources

Product Reviews

McAuley-Lab (UCSD)
/Amazon-Reviews-2023
571.54M reviews

Prime User

Kaggle Dataset

Product MetaData 2

Julian McAuley, UCSD
143.7 million reviews
spanning May 1996 - July
2014.

Product MetaData

McAuley-Lab (UCSD)
/Amazon-Reviews-2023
Richer Metadata

Product Sales

Kaggle Dataset

Categories

Google search

Data Sources

Product Reviews

Field name	Type	Mode
rating	FLOAT	NULLABLE
title	STRING	NULLABLE
text	STRING	NULLABLE
images	STRING	REPEATED
asin	STRING	NULLABLE
parent_asin	STRING	NULLABLE
user_id	STRING	NULLABLE
review_date	DATE	NULLABLE
helpful_vote	INTEGER	NULLABLE
verified_purchase	BOOLEAN	NULLABLE
details	STRING	NULLABLE
videos	STRING	REPEATED
_data_source	STRING	NULLABLE
_load_time	TIMESTAMP	NULLABLE

Product Meta Data

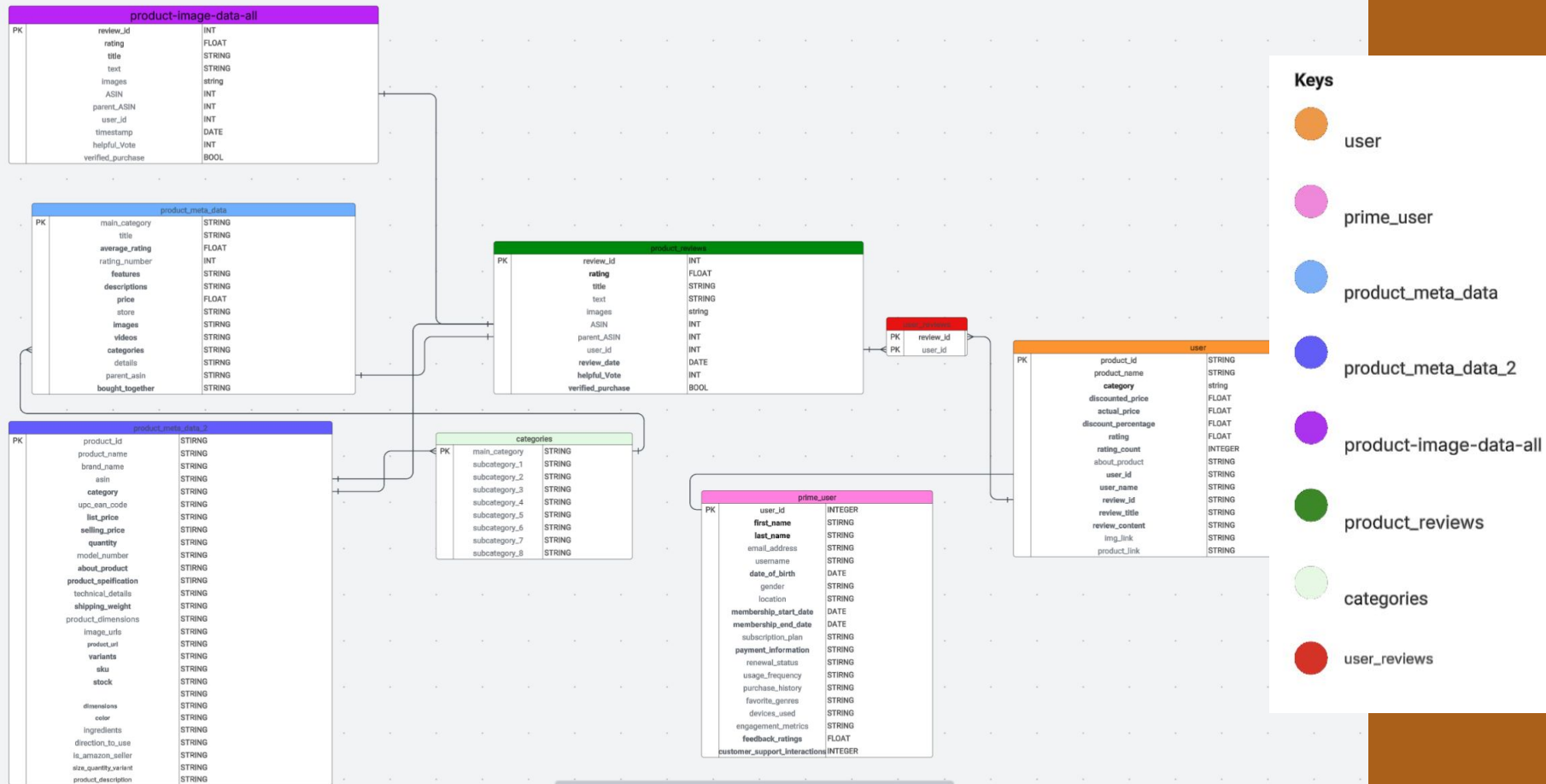
Field name	Type	Mode
main_category	STRING	NULLABLE
title	STRING	NULLABLE
average_rating	FLOAT	NULLABLE
rating_number	INTEGER	NULLABLE
features	STRING	NULLABLE
description	STRING	NULLABLE
price	FLOAT	NULLABLE
store	STRING	NULLABLE
images	STRING	REPEATED
videos	STRING	REPEATED
categories	STRING	REPEATED
details	STRING	NULLABLE
parent_asin	STRING	NULLABLE
bought_together	STRING	NULLABLE
_data_source	STRING	NULLABLE
_load_time	TIMESTAMP	NULLABLE

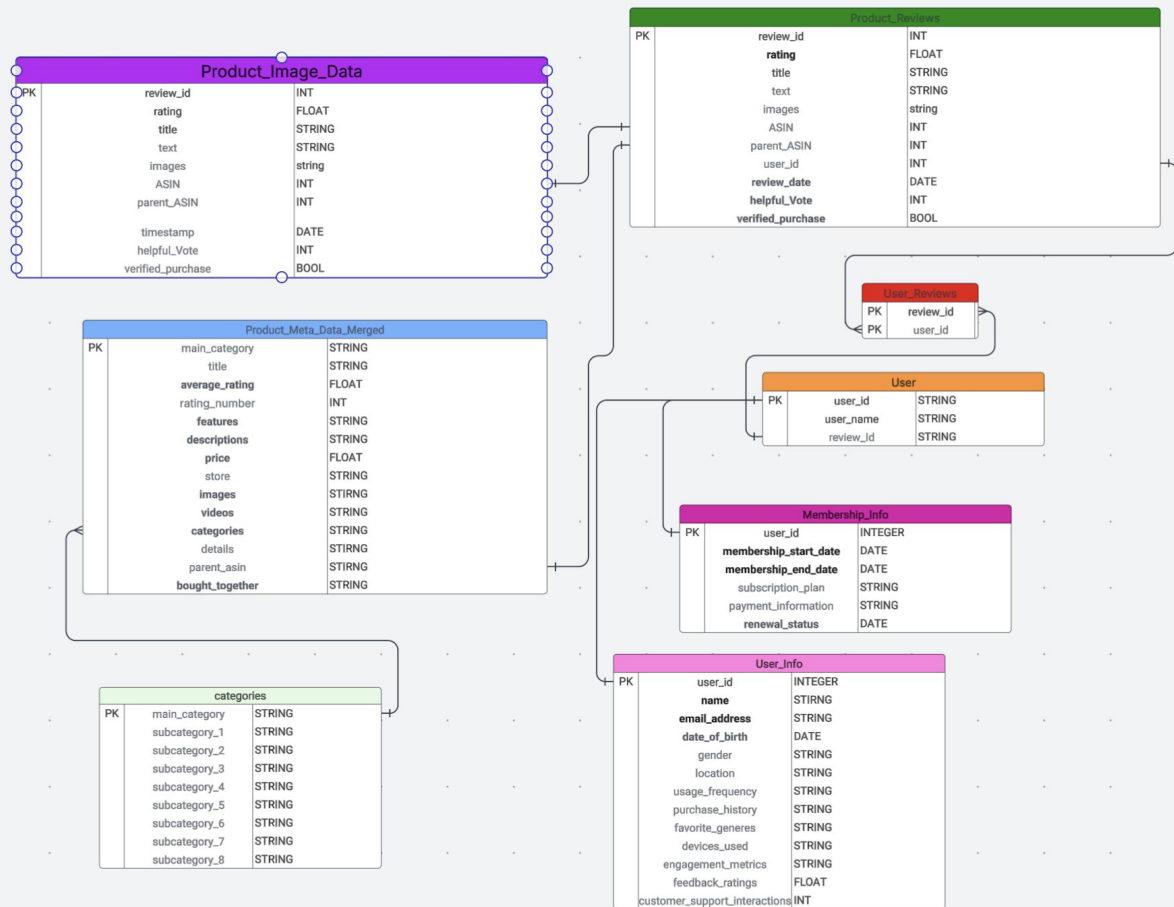
User

Field name	Type	Mode
user_id	INTEGER	NULLABLE
name	STRING	NULLABLE
email_address	STRING	NULLABLE
username	STRING	NULLABLE
date_of_birth	DATE	NULLABLE
gender	STRING	NULLABLE
location	STRING	NULLABLE
membership_start_date	DATE	NULLABLE
membership_end_date	DATE	NULLABLE
subscription_plan	STRING	NULLABLE
payment_information	STRING	NULLABLE
renewal_status	STRING	NULLABLE
usage_frequency	STRING	NULLABLE
purchase_history	STRING	NULLABLE
favorite_genres	STRING	NULLABLE
devices_used	STRING	NULLABLE
engagement_metrics	STRING	NULLABLE
feedback_ratings	FLOAT	NULLABLE

Sales

Field name	Type	Mode
product_id	STRING	NULLABLE
product_name	STRING	NULLABLE
category	STRING	NULLABLE
discounted_price	STRING	NULLABLE
actual_price	STRING	NULLABLE
discount_percentage	STRING	NULLABLE
rating	FLOAT	NULLABLE
rating_count	INTEGER	NULLABLE
about_product	STRING	NULLABLE
user_id	STRING	NULLABLE
user_name	STRING	NULLABLE
review_id	STRING	NULLABLE
review_title	STRING	NULLABLE
review_content	STRING	NULLABLE
img_link	STRING	NULLABLE
product_link	STRING	NULLABLE
_data_source	STRING	REQUIRED
_load_time	TIMESTAMP	REQUIRED





Keys



user_reviews



User



User_Info



Membership_Info



Product_Meta_Data



Product_Meta_Data_2



Product_Image_Data_All



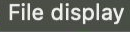
Categories



Product_Reviews

Image Data Extraction

Initial Generation

```
prompt = """
Analyze  product image and extract relevant metadata attributes. Identify and categorize unique features:

Extract the following attributes (if applicable):

Product Identification: Extract product_id, ASIN, UPC (if visible).
Brand & Naming: Identify brand_name, product_name, model_number.
Categorization: Identify Main_Category, Subcategory_1, Subcategory_2, Subcategory_3.
Pricing Details: If visible, extract list_price and selling_price.
Technical Details: Extract dimensions, weight, product_specs, product_technical information.
Text Extraction: If the image contains any text, extract the words accurately.
Visual Features: Identify relevant object features such as materials, packaging details, logos, and labels.
Color Information: Identify dominant colors present in the image.
Return the output in JSON format with this structure:
```

Image Data Extraction

Version 1

```
prompt = """
```

```
Analyze the given product image and extract relevant metadata attributes. Identify and categorize unique features, colors, and text present
```

```
Extract the following attributes (if applicable):
```

```
Product Identification: Extract product_id, ASIN, UPC (if visible).
```

```
Brand & Naming: Identify brand_name, product_name, model_number.
```

```
Categorization: Identify Main_Category, Subcategory_1, Subcategory_2, Subcategory_3.
```

```
Pricing Details: If visible, extract list_price and selling_price.
```

```
Technical Details: Extract dimensions, weight, product_specs, product_technical information.
```

```
Text Extraction: If the image contains any text, extract the words accurately.
```

```
Visual Features: Identify relevant object features such as materials, packaging details, logos, and labels.
```

```
Color Information: Identify dominant colors present in the image.
```

```
Return the output in JSON format.
```


Image Data Extraction

Version 1

Return the output in JSON format.

```
{
  "business": "product_data_analysis",
  "category": "product_metadata",
  "product_attributes": {
    "product_id": "STRING",
    "brand_name": "STRING",
    "product_name": "STRING",
    "ASIN": "STRING",
    "UPC": "BOOL",
    "Main_Category": "STRING",
    "Subcategory_1": "STRING",
    "Subcategory_2": "STRING",
    "Subcategory_3": "STRING",
    "list_price": "FLOAT",
    "selling_price": "FLOAT",
    "quantity": "INT",
    "model_number": "STRING",
    "about_product": "STRING",
    "product_specs": "STRING",
    "product_technical": "STRING",
    "weight": "STRING",
    "dimensions": "STRING",
    "url": "STRING"
  },
  "image_analysis": {
    "detected_text": ["TEXT"],
    "dominant_colors": ["COLOR1", "COLOR2"],
    "object_features": ["FEATURE1", "FEATURE2"]
  }
}
```

Do not include any extra details outside this format. Ensure extracted attributes are precise and

11/11/2024

Image Data Extraction

Version 2

```
prompt = """
```

```
Analyze the given product image and extract relevant metadata attributes. Identify and categorize unique features, colors, and text present in the image.
```

```
Extract the following attributes (if applicable):
```

- Brand & Naming: Extract brand_name, product_name, model_number.
- Categorization: Identify Main_Category, Subcategory_1, Subcategory_2, Subcategory_3.
- Pricing Details: If visible, extract list_price and selling_price.
- Technical Details: Extract product_specs, product_technical information.
- Text Extraction: If the image contains any text, extract the words accurately.
- Visual Features: Identify relevant object features such as materials, packaging details, logos, and labels.
- Color Information: Identify dominant colors present in the image.

```
Additionally, consider extracting:
```

- Logo Detection: Recognize brand logos if visible.
- Text Sentiment: Analyze text on packaging for positive or negative language.
- Object Detection: Identify key objects (e.g., accessories, packaging).
- Image Quality: Assess clarity, lighting conditions, and presence of watermarks.

```
Return the output in JSON format.
```

Image Data Extraction

Version 2

Return the output in JSON format.

```
{
  "business": "product_data_analysis",
  "category": "product_metadata",
  "product_attributes": {
    "brand_name": "STRING",
    "product_name": "STRING",
    "Main_Category": "STRING",
    "Subcategory_1": "STRING",
    "Subcategory_2": "STRING",
    "Subcategory_3": "STRING",
    "list_price": "FLOAT",
    "selling_price": "FLOAT",
    "model_number": "STRING",
    "about_product": "STRING",
    "product_specs": "STRING",
    "product_technical": "STRING",
    "url": "STRING"
  },
  "image_analysis": {
    "detected_text": ["TEXT"],
    "dominant_colors": ["COLOR1", "COLOR2"],
    "object_features": ["FEATURE1", "FEATURE2"],
    "logo_detection": ["LOGO1", "LOGO2"],
    "text_sentiment": "STRING",
    "object_detection": ["OBJECT1", "OBJECT2"],
    "image_quality": {
      "clarity": "STRING",
      "lighting": "STRING",
      "watermarks": "BOOL"
    }
  }
}
```

Do not include any extra details outside this format. Ensure extracted attributes are precise and relevant to the specific product image.

Image Data Extraction

Version 3

```
prompt = ""
```

```
Analyze the given product image and extract relevant metadata attributes. Identify and categorize unique features, colors, and text present in the image.
```

```
Extract the following attributes (if applicable):
```

- Brand & Naming: Extract brand_name, product_name.
- Categorization: Identify Main_Category, Subcategory_1, Subcategory_2, Subcategory_3, Subcategory_4, Subcategory_5.
- Pricing Details: If visible, extract list_price and selling_price.
- Technical Details: Extract product_technical information.
- Text Extraction: If the image contains any text, extract the words accurately.
- Visual Features: Identify relevant object features such as materials, packaging details, logos, and labels.
- Color Information: Identify dominant colors present in the image.
- Logo Detection: Recognize brand logos if visible.
- Text Sentiment: Analyze text on packaging for positive or negative language.
- Object Detection: Identify key objects (e.g., accessories, packaging).
- Image Quality: Assess clarity, lighting conditions, and presence of watermarks.
- Product Condition: Assess whether the product appears new, used, or refurbished.
- Packaging Type: Identify retail box, bulk packaging, eco-friendly packaging, etc.
- Target Demographic: Identify likely target audience based on visual cues.
- Competitor Products: Identify if competitor products are visible in the image.

```
Return the output in JSON format.
```

Image Data Extraction

Version 3

Return the output in JSON format.

```
{
  "business": "product_data_analysis",
  "category": "product_metadata",
  "product_attributes": {
    "brand_name": "STRING",
    "product_name": "STRING",
    "Main_Category": "STRING",
    "Subcategory_1": "STRING",
    "Subcategory_2": "STRING",
    "Subcategory_3": "STRING",
    "Subcategory_4": "STRING",
    "Subcategory_5": "STRING",
    "list_price": "FLOAT",
    "selling_price": "FLOAT",
    "about_product": "STRING",
    "product_technical": "STRING",
    "url": "STRING"
  },
  "image_analysis": {
    "detected_text": ["TEXT"],
    "dominant_colors": ["COLOR1", "COLOR2"],
    "object_features": ["FEATURE1", "FEATURE2"],
    "logo_detection": ["LOGO1", "LOGO2"],
    "text_sentiment": "STRING",
    "object_detection": ["OBJECT1", "OBJECT2"],
    "image_quality": {
      "clarity": "STRING",
      "lighting": "STRING",
      "watermarks": "BOOL"
    }
  },
  "product_condition": "STRING",
  "packaging_type": "STRING",
  "target_demographic": ["DEMOGRAPHIC1", "DEMOGRAPHIC2"],
  "competitor_products": ["COMPETITOR1", "COMPETITOR2"]
}
```

Do not include any extra details outside this format. Ensure extracted attributes are precise and relevant to the specific product image.

Image Data Extraction

Version 4

```
prompt = """
Analyze the given product image and extract relevant metadata attributes. Identify and categorize unique features, colors, and t
Extract the following attributes (if applicable):
- Brand & Naming: Extract brand_name, product_name.
- Categorization: Identify Main_Category, Subcategory_1, Subcategory_2, Subcategory_3, Subcategory_4, Subcategory_5.
- Pricing Details: If visible, extract list_price and selling_price.
- Technical Details: Extract product_technical information.
- Text Extraction: If the image contains any text, extract the words accurately.
- Visual Features: Identify relevant object features such as materials, packaging details, logos, and labels.
- Color Information: Identify dominant colors present in the image.
- Logo Detection: Recognize brand logos if visible.
- Text Sentiment: Analyze text on packaging for positive or negative language.
- Object Detection: Identify key objects (e.g., accessories, packaging).
- Image Quality: Assess clarity, lighting conditions, and presence of watermarks.
- Product Condition: Assess whether the product appears new, used, or refurbished.
- Packaging Type: Identify retail box, bulk packaging, eco-friendly packaging, etc.
- Target Demographic: Identify likely target audience based on visual cues.
- Competitor Products: Identify if competitor products are visible in the image.
- Image Presentation: Analyze background type (studio, lifestyle, environmental), product completeness, viewing angles, scale re
- Marketing Elements: Identify promotional badges, comparative elements, certifications, seasonal themes, special editions.
- User Experience: Assess interface visibility, controls shown, ergonomic highlights, accessories shown, usage environment.
- Visual Metrics: Calculate color harmony score (0-10), style classification, visual complexity score (0-10), symmetry score (0-
```


Image Data Extraction

Version 4

```
Return the output in JSON format.
{
  "business": "product_data_analysis",
  "category": "product_metadata",
  "product_attributes": {
    "brand_name": "STRING",
    "product_name": "STRING",
    "Main_Category": "STRING",
    "Subcategory_1": "STRING",
    "Subcategory_2": "STRING",
    "Subcategory_3": "STRING",
    "Subcategory_4": "STRING",
    "Subcategory_5": "STRING",
    "list_price": "FLOAT",
    "selling_price": "FLOAT",
    "about_product": "STRING",
    "product_technical": "STRING",
    "url": "STRING"
  },
  "image_analysis": {
    "detected_text": ["TEXT"],
    "dominant_colors": ["COLOR1", "COLOR2"],
    "object_features": ["FEATURE1", "FEATURE2"],
    "logo_detection": ["LOGO1", "LOGO2"],
    "text_sentiment": "STRING",
    "object_detection": ["OBJECT1", "OBJECT2"],
    "image_quality": {
      "clarity": "STRING",
      "lighting": "STRING",
      "watermarks": "BOOL"
    },
    "product_condition": "STRING",
    "packaging_type": "STRING",
    "target_demographic": ["DEMOGRAPHIC1", "DEMOGRAPHIC2"],
    "competitor_products": ["COMPETITOR1", "COMPETITOR2"]
  }
}
```

```
},
"image_presentation": {
  "background_type": "STRING",
  "product_completeness": "STRING",
  "viewing_angles": ["ANGLE1", "ANGLE2"],
  "scale_reference": "BOOL",
  "assembly_stage": "STRING"
},
"marketing_elements": {
  "promotional_badges": ["BADGE1", "BADGE2"],
  "comparative_elements": "BOOL",
  "certifications": ["CERT1", "CERT2"],
  "seasonal_theme": "STRING",
  "special_edition": "BOOL"
},
"user_experience": {
  "interface_visibility": "STRING",
  "controls_shown": ["CONTROL1", "CONTROL2"],
  "ergonomic_highlights": ["HIGHLIGHT1", "HIGHLIGHT2"],
  "accessories_shown": ["ACC1", "ACC2"],
  "usage_environment": "STRING"
},
"visual_metrics": {
  "color_harmony_score": "FLOAT",
  "style_classification": "STRING",
  "visual_complexity": "FLOAT",
  "symmetry_score": "FLOAT",
  "focal_area": "STRING"
}
```