Muhammad Zoraib Qadir

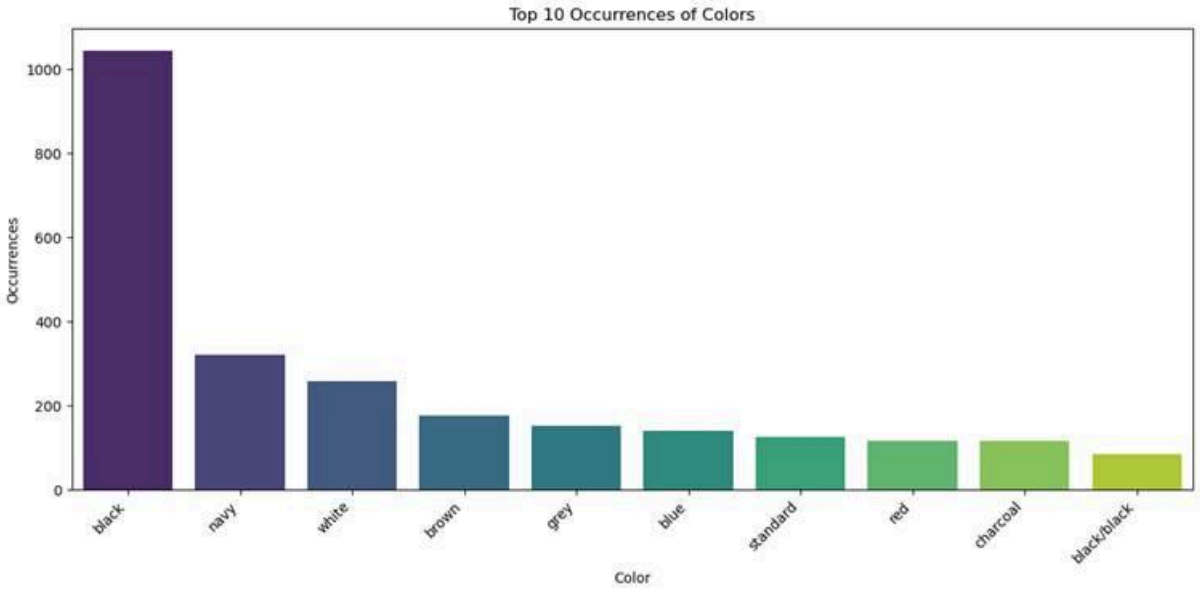
Data Mining

Table of Contents

Data Preprocessing:

Filling Null Counts

Checking the occurrence of the color and filling with that color



- **Color Data Handling:**
    - A function called get_color was created to identify and extract color lists within the 'Color' column.
    - This function was applied to the 'Color' column using df['Color'].apply(get_color).
- **Missing Value Handling:**
    - Missing values in the DataFrame were filled with NaN using df.fillna(value=np.nan, inplace=True).
- **Duplicate Identification and Handling:**
    - A function called list_duplicate was created to detect duplicate values in columns.
    - Columns containing lists were identified as having potential duplicates (list_colors).

    - These columns were converted to strings using a lambda function within a loop over list_colors.
- **Duplicate Removal:**
    - Duplicate rows within the DataFrame were removed using df.drop_duplicates(inplace=True).
- **'Average_Review' Cleaning:**
    - The 'Average_Review' column was cleaned by removing non-numeric characters and converting it to float data type using df['Average_Review'].str.replace('[^\d\.]', '', regex=True).astype(float).

Stopwords Removal and Lemmatization

- Defined a function clean_text to clean the text data:
- Converted text to lowercase.
- Removed punctuation using a list comprehension.
- Tokenized the text using word_tokenize from NLTK.
- Removed stopwords from the tokens.
- Lemmatized the tokens using the WordNet Lemmatizer.
- Join the lemmatized tokens back into a string and append it to cleaned_reviews.

Applying NLP to the 'Reviews' Column:

- Applied the clean_text function to each element of the 'Reviews' column using df['Reviews'].apply(lambda x: clean_text(x) if isinstance(x, list) else x).
- The lambda function checks if the element is a list and applies the clean_text function,

otherwise keeps the element unchanged.

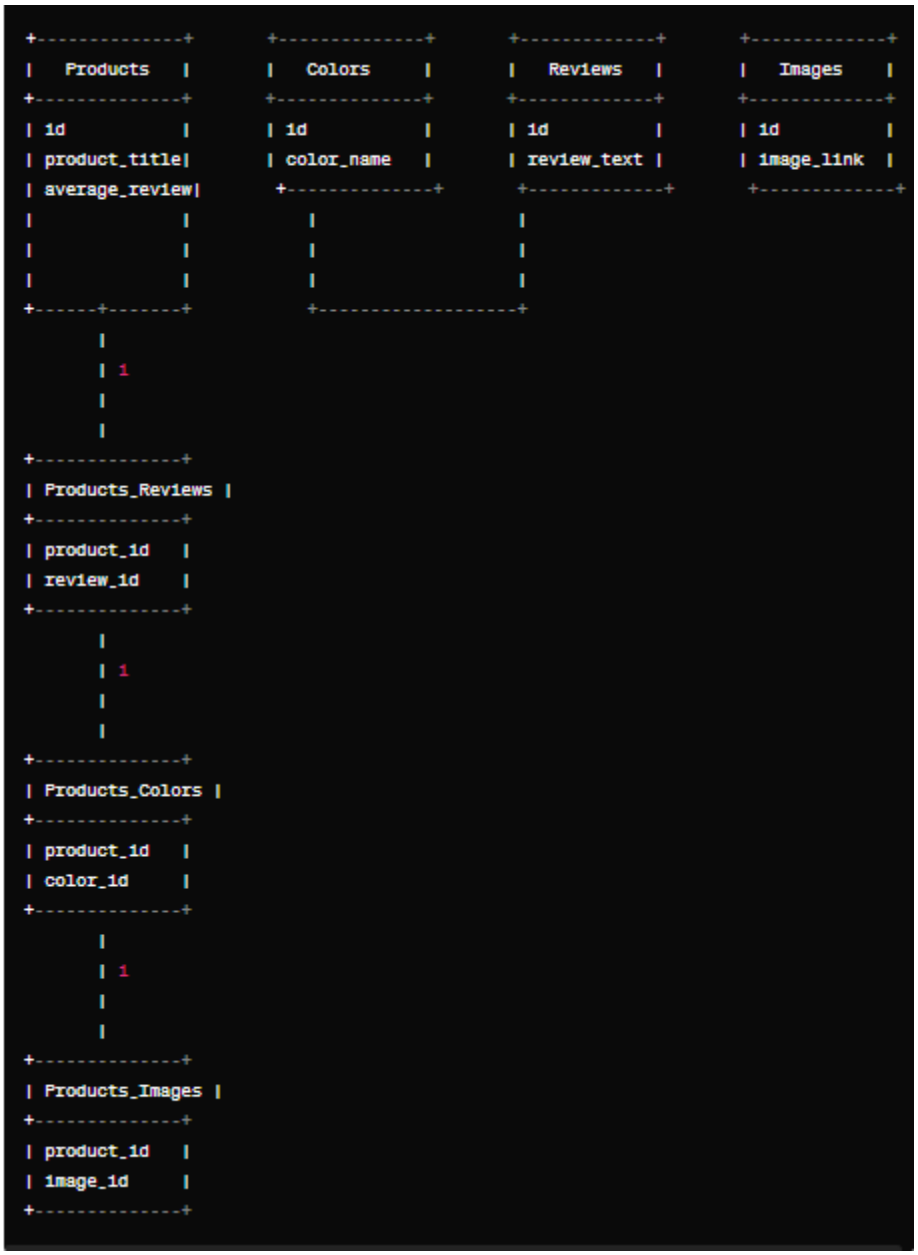
Outcome:

· The 'Reviews' column was processed to remove stopwords, punctuation, and lemmatize the words, resulting in a cleaner and more standardized format for text analysis.

Image Data Handling

- A custom function, extract_image_links, was designed to extract image links from the DataFrame's 'Color' column.
- The function demonstrates robust adaptability by handling both list and dictionary data structures within 'Color'.
- It prioritizes extracting "Landing_Image" links, followed by additional links from "Other Images" lists (if available).
- To maintain data organization, a new column named 'Image_Links' is strategically added to the DataFrame, storing the extracted links for subsequent use.
- For visual clarity and validation, a concise display of the 'Product_Title' and 'Image_Links' columns is presented.

Table Schema



Tables

1. Products:
    - **id (INT AUTO_INCREMENT PRIMARY KEY):** Unique identifier for each product (automatically incremented).

- **product_title (VARCHAR(255)):** Name of the product (limited to 255 characters).
- **average_review (VARCHAR(255) - NOTE: This appears to be an error)**: This column is initially defined as a decimal value with 3 digits total and 1 digit after the decimal point, but then altered to a string (VARCHAR) with a maximum length of 255 characters. It likely should be a decimal data type to store average review scores effectively.

2. Colors:
   - **id (INT AUTO_INCREMENT PRIMARY KEY):** Unique identifier for each color (automatically incremented).
   - **color_name (VARCHAR(255)):** Name of the color (limited to 255 characters).
3. Product_Colors:
   - **product_id (INT):** Foreign key referencing the "id" column in the "Products" table.
   - **color_id (INT):** Foreign key referencing the "id" column in the "Colors" table.
   - This table establishes a many-to-many relationship between products and colors. A single product can have multiple colors, and a single color can be associated with multiple products.
4. Reviews:
   - **id (INT AUTO_INCREMENT PRIMARY KEY):** Unique identifier for each review (automatically incremented).
   - **review_text (TEXT):** Stores the actual text content of the review.
5. Products_Reviews:
   - **product_id (INT):** Foreign key referencing the "id" column in the "Products" table.
   - **review_id (INT):** Foreign key referencing the "id" column in the "Reviews" table.
   - This table establishes a many-to-many relationship between products and reviews. A single product can have multiple reviews, and a single review can be associated

with multiple products (if applicable).

6. Images:
    - **id (INT AUTO_INCREMENT PRIMARY KEY):** Unique identifier for each image (automatically incremented).
    - **image_link (TEXT):** Stores the URL or link to the product image.
7. Products_Images:
    - **product_id (INT):** Foreign key referencing the "id" column in the "Products" table.
    - **image_id (INT):** Foreign key referencing the "id" column in the "Images" table.
    - This table establishes a many-to-many relationship between products and images. A single product can have multiple images, and a single image can be associated with multiple products (if applicable).

Tables of Database

Product Table:

168	Carhartt Men's Wb Suede Leather Waterproof ...	4.4 ratings
169	Carhartt Men's Winter Dex Cow Grain Leather T...	4.5 ratings
170	Coach Mens Tech Nappa Glove	4.5 ratings
171	Cold Snap Gloves, 7-Gauge, Hi-Vis Green, Brush...	4.3 ratings
172	Columbia Mens Blizzard Ridge Glove	4.8 ratings
173	Columbia Mens Men's Bugaboo™ Interchange Gl...	4.7 ratings
174	Columbia Mens Men's Bugaboo™ Interchange Gl...	4.0 ratings
175	Columbia Unisex Omni-Heat Touch Glove Liner, ...	4.1 ratings
176	Columbia unisex-adult Omni-heat Touch Glove Li...	4.1 ratings
177	Dickies Synthetic Leather Work Gloves Men, Im...	5.0 ratings
178	Eskimo 41592 Keeper™ Glove with Liner Glove, ...	5.0 rating
179	Eskimo Buffalo Plaid Cold Weather Glove	4.6 ratings
180	Eskimo unisex-adult Buffalo Chopper MittIce Fis...	4.3 ratings
181	Flylow Maine Line Synthetic Insulated Waterpro...	4.4 ratings
182	Fox River Men's Mid Weight Ragg Glove	4.5 ratings

Color Table:

Result Grid | Filter Rows: | Edit: | Export/Import:

color_id	color_name	price	product_id	path1
1	Black	88.56	1	images\B01DZQ9XSK\Black
2	Oxblood	84.43	1	images\B01DZQ9XSK\Oxblood
3	Black	83.69	2	images\B000MXVAIQ\Black
4	Chili	135.00	2	images\B000MXVAIQ\Chili
5	Black	17.00	3	images\B07FZBTJ81\Black
6	Tan	19.00	3	images\B07FZBTJ81\Tan
7	Dark Brown	19.00	3	images\B07FZBTJ81\Dark Brown
8	Black	19.00	4	images\B0BPRSFTKR\Black
9	Chestnut	15.90	4	images\B0BPRSFTKR\Chestnut
10	Tan	19.00	4	images\B0BPRSFTKR\Tan
11	Saddle	11.90	4	images\B0BPRSFTKR\Saddle

Image Table:

Result Grid | Filter Rows: | Edit: | Ex

image_id	image_url	color_id
1	https://m.media-amazon.com/images/I/710-SE...	1
2	https://m.media-amazon.com/images/I/31Omld...	1
3	https://m.media-amazon.com/images/I/71fwn+...	2
4	https://m.media-amazon.com/images/I/31BNWJ...	2
5	https://m.media-amazon.com/images/I/719V35r...	3
6	https://m.media-amazon.com/images/I/31UFpiK...	3
7	https://m.media-amazon.com/images/I/815OnZ...	4
8	https://m.media-amazon.com/images/I/41-9-t6...	4
9	https://m.media-amazon.com/images/I/717ION...	5
10	https://m.media-amazon.com/images/I/31C9Pq...	5
11	https://m.media-amazon.com/images/I/31tXDrI...	5

1.