**Progress Report: FAQ Dataset Cleaning and Exploratory Analysis**

**Introduction**

The goal of this project is to prepare and analyze a multilingual Frequently Asked Questions (FAQ) dataset for downstream tasks such as modeling or chatbot development. The dataset contains questions and answers from various e-commerce sources, primarily in English, with translations into Swahili and Sheng. This report summarizes the data cleaning, preprocessing, and exploratory data analysis (EDA) steps completed so far, along with key preliminary findings.

**Methods**

**1. Data Loading and Initial Inspection**

* The original dataset was loaded using pandas from a CSV file.
* Initial inspection included checking column names, data types, and viewing the first few rows to understand the structure.

**2. Data Cleaning**

* **Missing Values:**  
  Rows with missing values in either the 'Question' or 'Answer' columns were identified and removed to ensure data completeness.
* **Duplicates:**  
  Duplicate rows were checked and could be removed if found, ensuring each FAQ entry is unique.
* **Text Normalization:**  
  All text in the 'Question' and 'Answer' columns was converted to lowercase to maintain consistency.
* **Punctuation Removal:**  
  Punctuation was stripped from the text to simplify further processing and analysis.

**3. Multilingual Translation and Shengification**

* **Swahili Translation:**  
  The Google Translate API was used to translate questions and answers into Swahili, with error handling for missing or problematic entries.
* **Sheng Conversion:**  
  A custom function replaced common Swahili words with their Sheng equivalents, creating additional columns for Sheng questions and answers.

**4. Text Preprocessing**

* **Stopword Removal:**  
  Stopwords for English, Swahili, and Sheng were compiled and removed from the text.
* **Lemmatization:**  
  English words were lemmatized using NLTK to reduce them to their base forms.
* **Final Cleaned Dataset:**  
  The cleaned and preprocessed data was saved to a new CSV file for further analysis.

**5. Exploratory Data Analysis (EDA)**

* **Word Cloud:**  
  A word cloud was generated to visualize the most frequent words in the cleaned questions.
* **Question Length Distribution:**  
  A histogram was plotted to show the distribution of question lengths (in words).
* **Top Words Bar Chart:**  
  The ten most common words in the questions were identified and visualized.
* **Question vs. Answer Length:**  
  A scatter plot was created to explore the relationship between question and answer lengths.
* **Source Frequency:**  
  A bar chart displayed the number of questions from each data source.
* **Missing Data Heatmap:**  
  A heatmap confirmed the absence of missing values in the cleaned dataset.

**Preliminary Findings**

**Data Quality**

* **Completeness:**  
  After cleaning, the dataset contains no missing or duplicate entries, as confirmed by both code checks and visualizations.
* **Dominant Source:**  
  The majority of FAQ entries originate from "[https://www.myntra.com/faqs](vscode-file://vscode-app/c:/Users/Admin/AppData/Local/Programs/Microsoft%20VS%20Code/resources/app/out/vs/code/electron-browser/workbench/workbench.html)," with other sources contributing far fewer questions.

**Content Analysis**

* **Frequent Topics:**  
  The most common words in questions are "myntra," "order," "card," "product," "credit," "return," "account," "delivery," "payment," and "use." This indicates that user concerns are mainly about orders, products, payments, returns, and account management.
* **Question Length:**  
  Most questions are concise, typically containing 3–7 words. Longer questions are rare.
* **Answer Length:**  
  Answers vary widely in length, with some short and others very detailed, even for brief questions.
* **Question vs. Answer Length:**  
  There is no strong correlation between the length of a question and the length of its answer. Short questions can have both short and long answers, suggesting that even simple queries may require detailed explanations.

**Data Validation**

* **Missing Values:**  
  All missing values were removed, and a heatmap confirmed the dataset is now complete.
* **Duplicates:**  
  Duplicate entries were checked and removed as needed.
* **Ready for Modeling:**  
  The dataset is now clean, consistent, and suitable for further analysis or machine learning tasks.

**Next Steps**

* **Advanced Text Analysis:**  
  Further analysis such as topic modeling or clustering to uncover deeper patterns.
* **Model Training:**  
  Use the cleaned dataset to train FAQ retrieval or chatbot models.
* **Evaluation:**  
  Assess model performance and iterate on preprocessing if necessary.

**Conclusion**

Significant progress has been made in cleaning, translating, and analyzing the FAQ dataset. The data is now validated and ready for advanced modeling, with clear insights into user concerns and data structure. The next phase will focus on leveraging this clean dataset for machine learning and natural language processing applications.