Login Datathon

Team Name: Algorithm Army

Coimbatore Institute of Technology

Team Members: Srimathy Balakrishnan, Harinne R Kumar

Problem Statement

The task involves predicting the **entity length** (denoted as ENTITY_LENGTH) for a set of entities using categorical data and textual descriptions Given the vast amount of categorical data and various entity descriptions, the goal is to predict the length of the entity descriptions accurately

Business Context

In many data-driven industries, entities (such as products, documents, or media files) need to be classified or processed based on their textual metadata and associated categorical information The ability to predict properties like the length of an entity's description can be crucial for tasks like automatic summarization, resource allocation, and content management

The problem is approached as a **supervised machine learning problem** where the objective is to build a predictive model that accurately estimates the length of an entity description given its categorical and descriptive information

Problem Understanding

Input Data

- **CATEGORY_ID**: A numerical or categorical identifier representing a category or class to which an entity belongs
- **ENTITY_DESCRIPTION**: A text field that contains a description of the entity, which varies in length and content

The task involves predicting the length of an entity’s description (ENTITY_LENGTH), which is not directly provided but needs to be inferred from the provided features

Output

- **ENTITY_LENGTH**: The predicted length of the entity’s description (a numerical value)

The challenge is to leverage the categorical information from CATEGORY_ID and the textual properties of ENTITY_DESCRIPTION to predict this numerical value

Objectives

The key objectives of this project are:

1. **Data Exploration**: Analyse the dataset to understand the distribution, unique values, and key characteristics of CATEGORY_ID and ENTITY_DESCRIPTION

2. **Data Preprocessing**:
    - Encode categorical variables to be used in the model
    - Engineer features such as the length of the description text
3. **Modelling**: Train a machine learning model using the processed data to predict the length of the description (ENTITY_LENGTH)
4. **Model Prediction**: Use the trained model to predict ENTITY_LENGTH on unseen test data
5. **Submission Generation**: Generate a CSV file containing the entity IDs along with their predicted description lengths for evaluation

Packages Used

The following Python packages were utilized in this notebook:

1. **pandas**:
    - Used for data manipulation and analysis
    - Helps with reading and writing CSV files, handling dataframes, and performing basic data exploration
2. **NumPy**:
    - Used for numerical operations
    - Essential for performing mathematical operations and handling arrays
3. **sklearn (scikit-learn)**:
    - Provides tools for preprocessing (like Label Encoding and Scaling) and building machine learning models
    - Used for label encoding (Label Encoder), data scaling (StandardScaler), and model training/predictions
4. **TensorFlow / keras**:
    - Used for deep learning models and model prediction tasks
    - The notebook may involve pre-trained models from this library for predicting ENTITY_LENGTH
5. **matplotlib / seaborn**:
    - May be used for visualization (though not seen explicitly in the output) These are popular for generating visual plots to explore the data

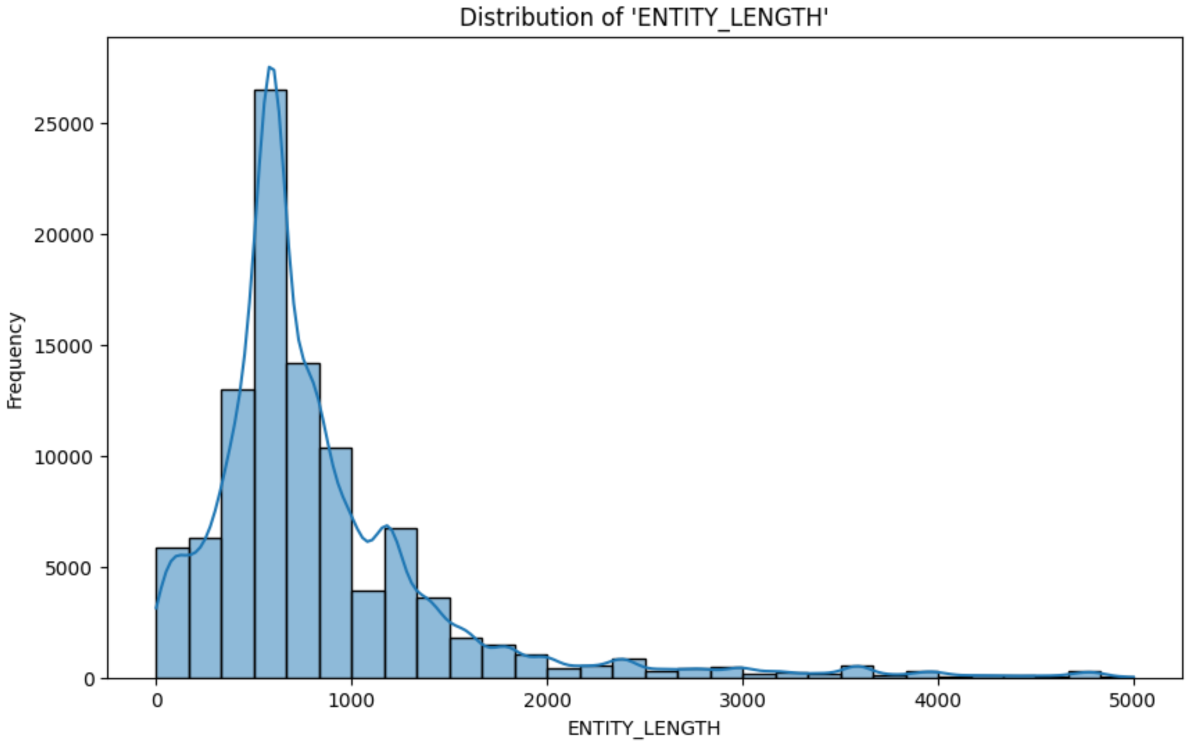
Method of Solving

Data Exploration

The first step involves exploring the dataset to understand its structure, distribution, and key characteristics Insights include:

- The CATEGORY_ID column is a categorical variable with 6,7 unique values and no missing values The distribution of this column is highly skewed, with CATEGORY_ID = being the most frequent
- The ENTITY_DESCRIPTION column contains text-based descriptions with 99,856 unique values, and common entries include "Unknown Title" and "Greatest Hits"

Data Visualization Using Charts:



Distribution of ENTITY_LENGTH

Chart Type: Histogram with Kernel Density Estimate (KDE) overlay

- **X-axis**: ENTITY_LENGTH – This represents the length of the entity descriptions, typically measured by the number of characters or words in the description.
- **Y-axis**: Frequency – This shows how many entities have descriptions of a particular length.

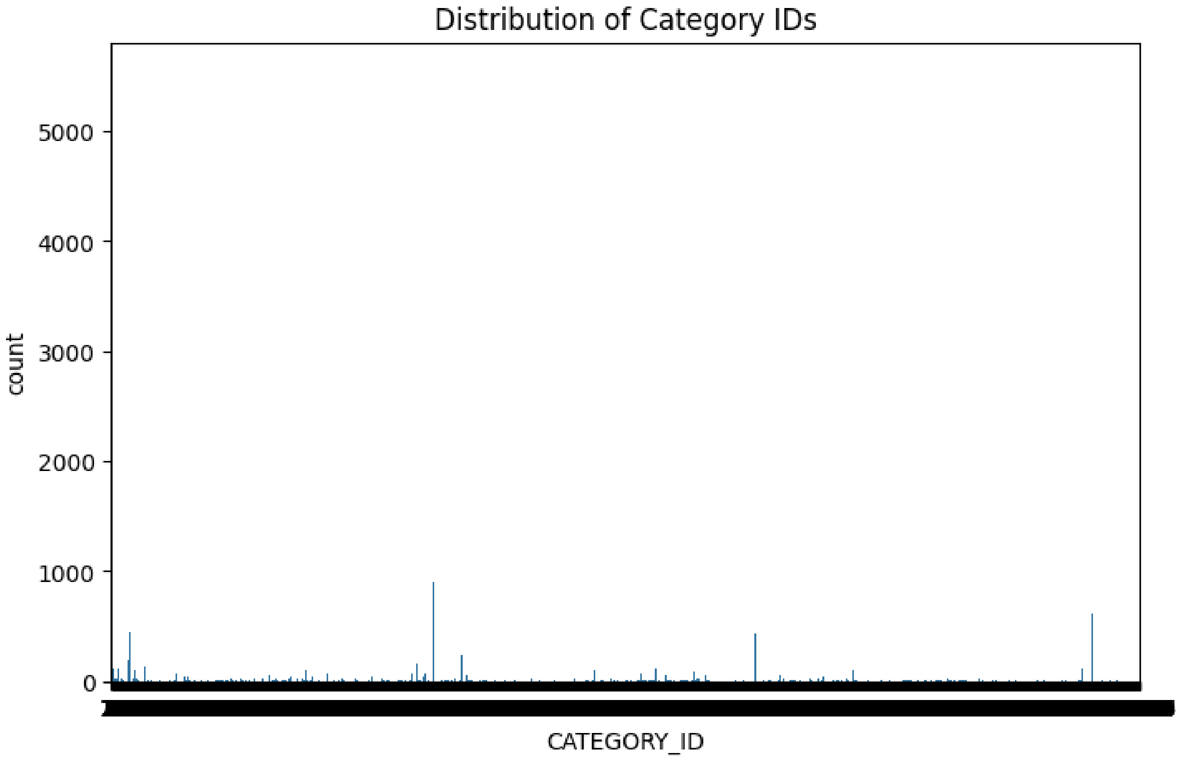
Description:

- The histogram shows a **right-skewed distribution** of the ENTITY_LENGTH, meaning most entity descriptions are relatively short, with fewer longer descriptions.

- The peak (mode) of the distribution occurs around **600-800 characters**, meaning most entity descriptions tend to fall within this range.
- The **tail of the distribution** extends toward much larger lengths (up to 5000 characters), but very few entities have descriptions longer than 2500 characters.

Key Insights:

- The majority of entity descriptions are shorter, likely containing only a few sentences or phrases.
- The presence of a long tail suggests that there are some outliers with extremely lengthy descriptions, which might need special handling or truncation in preprocessing.
- Understanding this distribution is critical, as it informs the model about the typical range of entity lengths and the potential outliers that could affect performance.



Distribution of CATEGORY_ID

Chart Type: Bar Plot (or Count Plot)

- **X-axis**: CATEGORY_ID – This represents unique category identifiers for different entities in the dataset.
- **Y-axis**: Count – This shows how many entities belong to each CATEGORY_ID.

Description:

- The distribution of CATEGORY_ID is **highly imbalanced**. Some category IDs occur far more frequently than others.
- A small number of CATEGORY_ID values (such as those represented by taller bars) have very high counts (e.g., over 5000), while the majority of categories have much lower counts, close to 0.
- Many CATEGORY_IDs are either rarely represented or sparsely populated.

Key Insights:

- This skewed distribution highlights a **class imbalance problem**, where a few categories dominate the dataset while most categories are underrepresented.
- Such imbalances can cause the model to be biased toward the most frequent categories unless steps like **resampling, class weighting, or regularization** are applied.

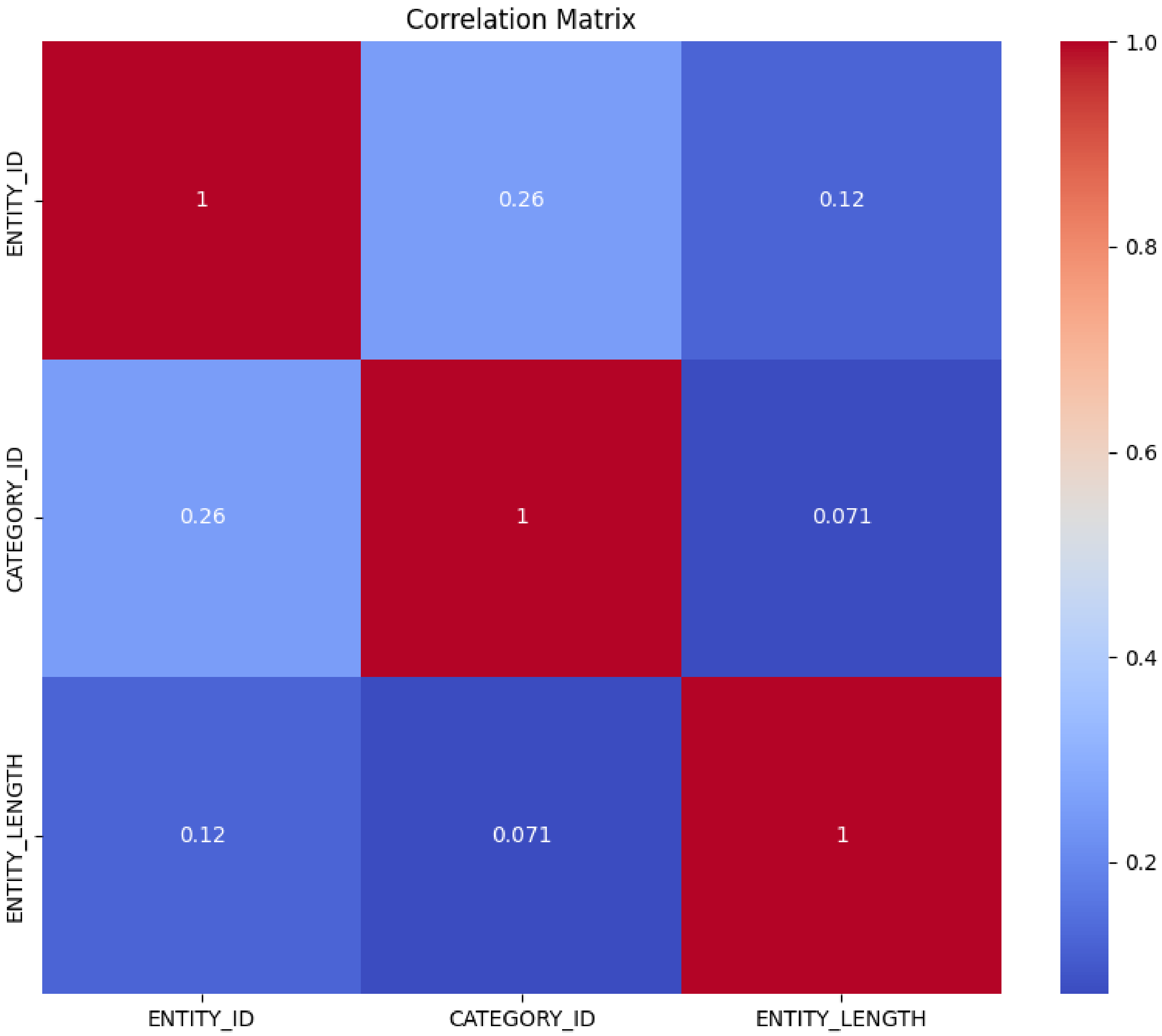


Chart Type: Heatmap

X-axis and Y-axis: Represent the variables ENTITY_ID, CATEGORY_ID, and ENTITY_LENGTH that were analysed to find correlations between them.

Colour Legend: The colour gradient from blue to red represents the correlation coefficients between the variables, with red indicating high positive correlations (close to 1), and blue indicating low or no correlation (close to 0).

Description:

- ENTITY_ID and CATEGORY_ID show a moderate correlation (0.26), suggesting a weak but notable relationship between these two variables.
- ENTITY_ID and ENTITY_LENGTH show a weaker correlation (0.12), meaning ENTITY_ID has a minimal association with the length of the entity.
- CATEGORY_ID and ENTITY_LENGTH have the lowest correlation (0.071), indicating almost no relationship between the entity's category and its length.
- Diagonal values (1) represent the correlation of each variable with itself, which is always perfect.

Key Insights:

- ENTITY_ID has a stronger relationship with CATEGORY_ID than it does with ENTITY_LENGTH, indicating that the categorization of entities might depend more on the entity identifier than its length.
- There is little to no correlation between ENTITY_LENGTH and either of the other two variables, implying that entity length is not a determining factor for entity classification or identification.

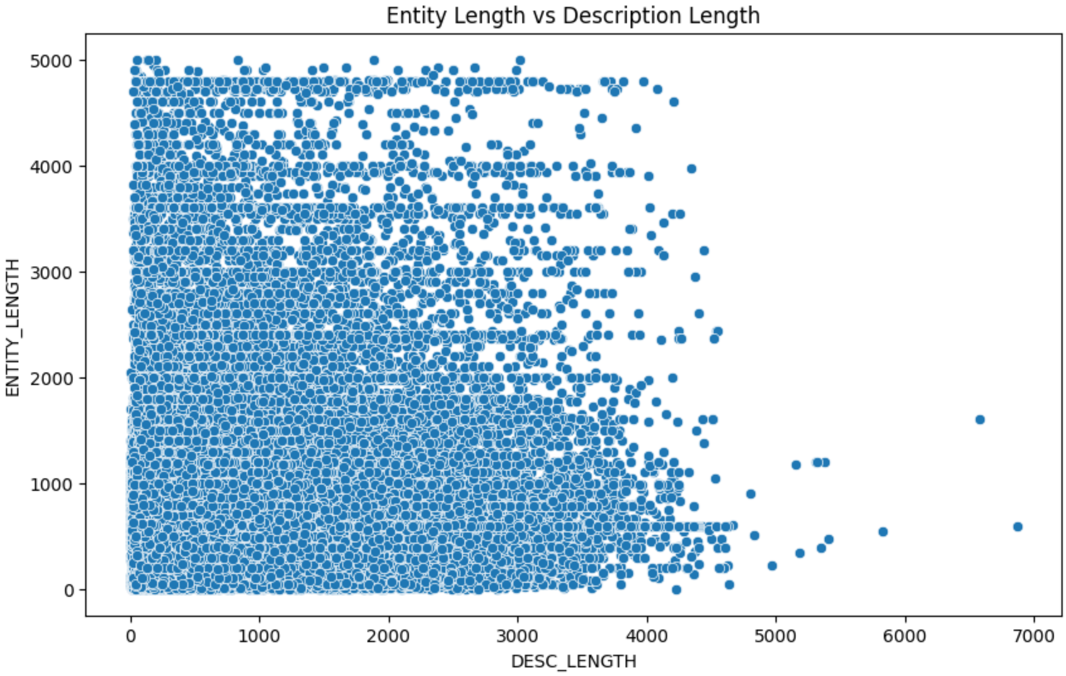


Chart Type: Scatter Plot

- X-axis: DESC_LENGTH – This represents the length of descriptions for entities.
- Y-axis: ENTITY_LENGTH – This represents the length of the entities.

Description:

- The plot shows the relationship between entity length and description length for various data points.
- There is a high density of points clustered towards the bottom-left corner, with both the entity length and description length being relatively small in most cases.
- A smaller number of points spread out across the higher ranges, indicating that some entities have both long descriptions and long entity lengths.
- Outliers exist, particularly where either description length or entity length is significantly larger than the general population.

Key Insights:

- No strong correlation: Based on the scatter plot, there seems to be no clear linear relationship between entity length and description length.
- Clusters: A large portion of the entities appears to have both short entity lengths and short descriptions, potentially indicating more standardized data in this range.
- Outliers: Some entities have much larger descriptions or entity lengths, representing exceptions to the general trend.

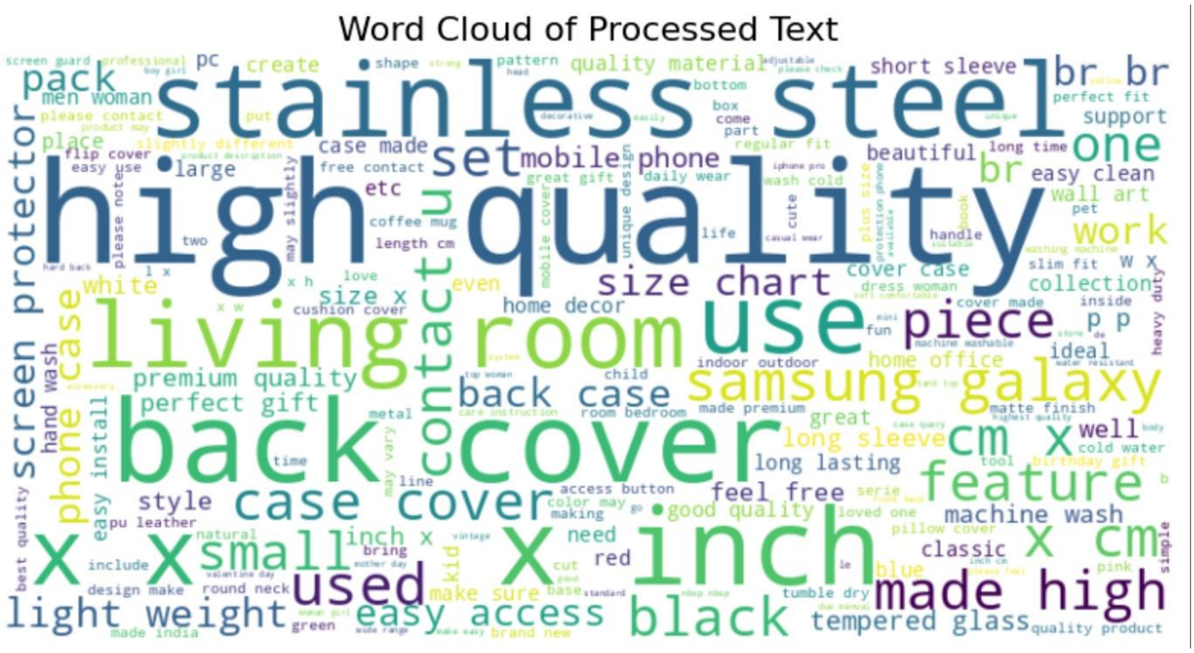


Chart Type: Word Cloud

Description:

- The word cloud displays the most frequent words from a processed text dataset, with the size of each word indicating its frequency or prominence in the text.
- Larger words such as "high", "quality", "stainless", "steel", "back", and "cover" are the most commonly used terms, suggesting these are key topics or product descriptions in the dataset.
- Words like "living", "room", "use", "case", "phone", and "contact" also appear frequently but are slightly smaller, indicating a slightly lower frequency compared to the largest words.
- The cloud includes terms related to product characteristics (e.g., "premium", "lightweight", "perfect gift", "made high", "stainless steel"), item types (e.g., "back cover", "case", "screen protector"), and measurements (e.g., "inch", "cm", "size").

Key Insights:

- The dominant focus appears to be on product quality and material, with "high quality", "stainless steel", and "lightweight" as standout phrases.
- Words such as "living room" and "piece" may indicate common product types or use cases.
- This word cloud highlights popular product features, materials, and sizes, potentially reflecting the dataset's focus on consumer goods, electronics, and home décor.

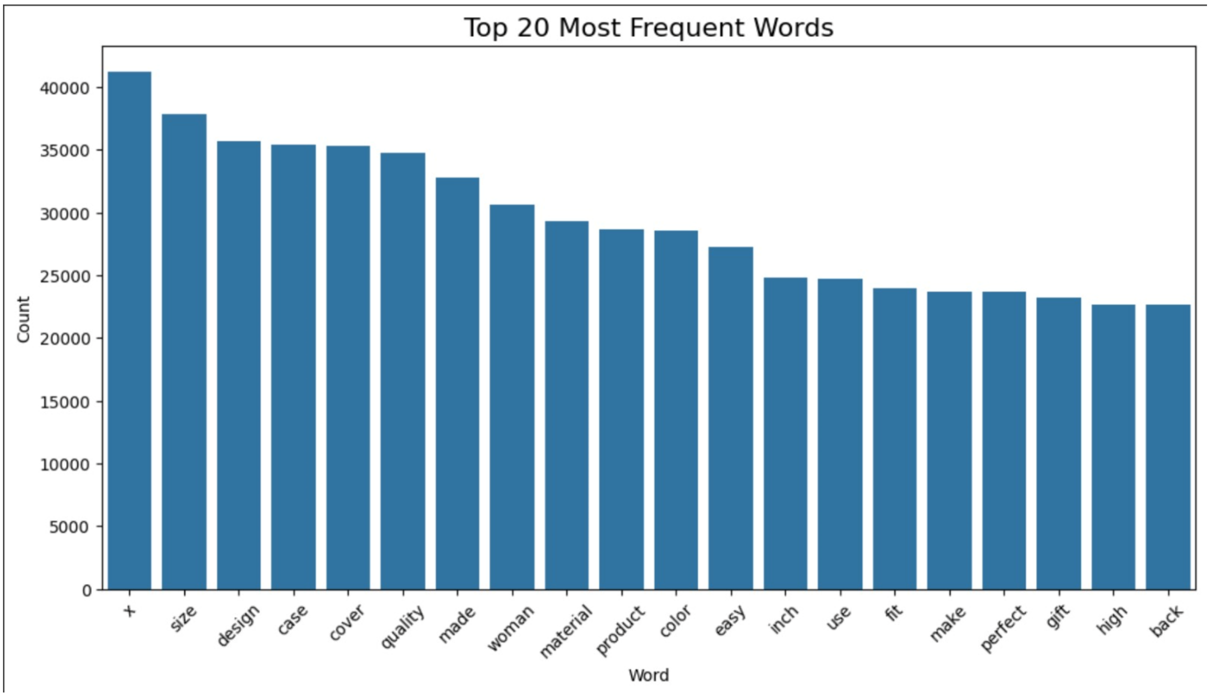


Chart Type: Bar Plot

X-axis: Word – This represents the top 20 most frequent words found in the processed text data.

Y-axis: Count – This shows the frequency of each word's appearance in the dataset.

Description:

- The most frequent word is represented by the character "x," followed by words like "size," "design," "case," and "cover," which appear over 35,000 times.
- The words related to product descriptions such as "quality," "material," and "product" are also highly frequent.
- Words like "easy," "inch," and "fit" indicate a focus on product specifications, ease of use, and dimensions, which may point to reviews or product data.

Key Insights:

- The frequent words suggest the dataset is likely related to product reviews, where users comment on product features such as size, design, material, and quality.
- The appearance of high-frequency terms like "woman" and "gift" may indicate that the products are often intended for specific demographics or as gifts.
- The commonality of basic, short, and non-contextual words may necessitate further text preprocessing, such as stop-word removal, for more meaningful insights in future analyses.

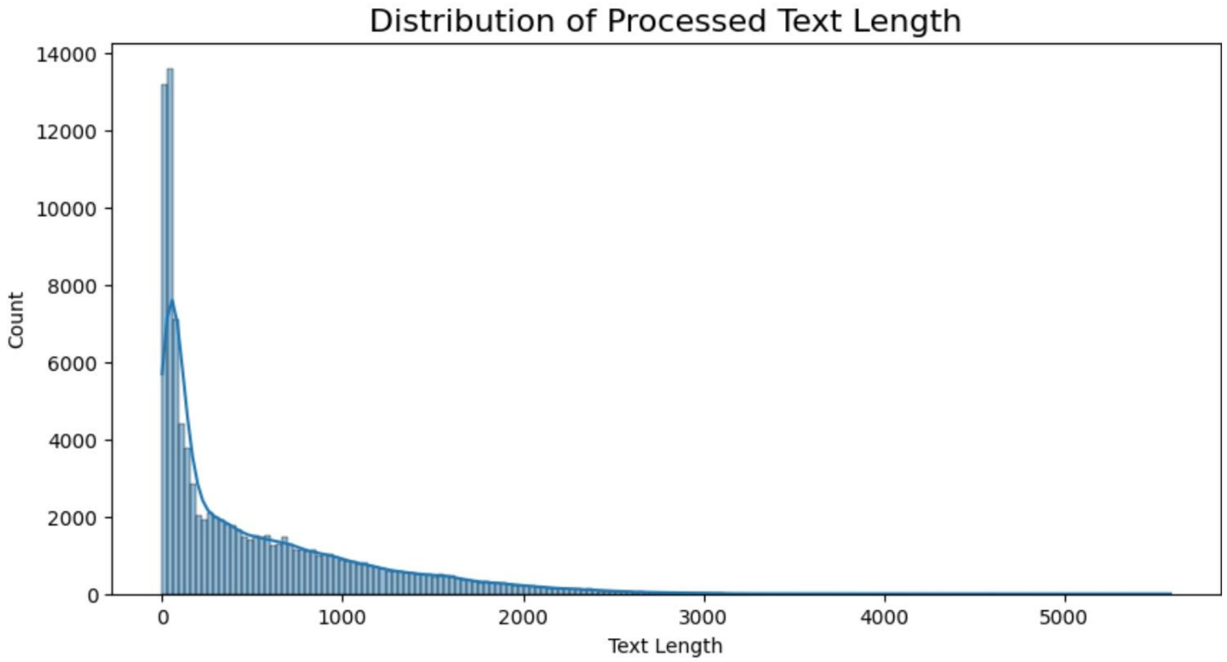


Chart Type: Histogram with KDE (Kernel Density Estimate)

X-axis: Text Length – This represents the length of the processed text for each entity or document in the dataset.

Y-axis: Count – This shows the number of occurrences of each text length.

Description:

- The distribution of text length is highly right-skewed, with a large concentration of shorter texts.
- The majority of text lengths are clustered around small values, with the highest peak occurring between 0 and 500 characters.
- There is a sharp decline after the initial peak, with very few texts exceeding lengths of 1000 characters. Some rare instances extend as far as 5000+ characters, but they are uncommon.

Key Insights:

- Most processed texts are relatively short, indicating that the dataset predominantly contains brief textual information.
- The heavy skewness might indicate a need for normalization or scaling when used in machine learning models, as longer texts could disproportionately influence model training.
- Depending on the analysis objective, such as sentiment analysis or text classification, strategies like padding or truncating text lengths may be necessary to handle varying text lengths consistently.