**Datathon-2024**

**Problem Statement:**

In the problem statement, given the task of predicting the length of various entities (likely products) using provided textual information and metadata, it is crucial in e-commerce for optimizing logistics, storage, and customer satisfaction.

The dataset contains both structured features (e.g Category) and unstructured textual data (e.g., product descriptions), which need to be pre-processed and transformed to develop a predictive model. Accurate length prediction can help with packaging and shipping optimization, ensuring better warehouse management and reducing product returns.

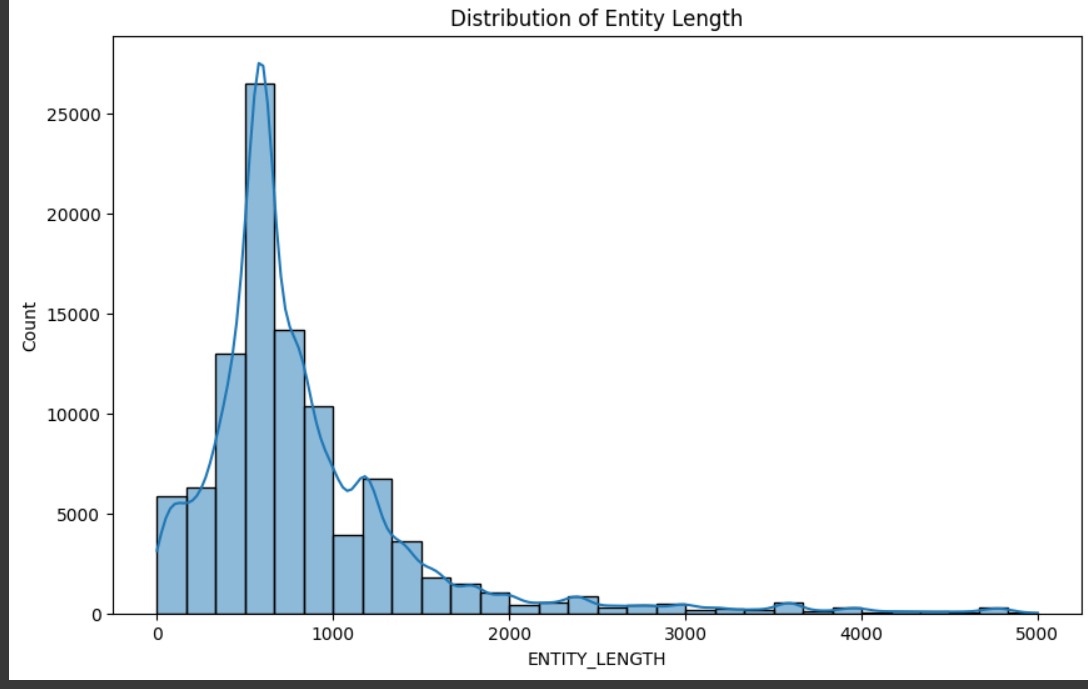
The challenge involves cleaning and processing the data, engineering relevant features from both text and structured data, selecting an appropriate machine learning model, and evaluating its performance using suitable regression metrics like RMSE or MAE. Success in this task will enable businesses to streamline logistics and improve customer experience by ensuring accurate size information for products.

**Objective:**

The objective of this task is to develop a machine learning model that accurately predicts the length of various entities (likely products) using a combination of structured metadata and unstructured textual information. Achieving this objective will enable businesses to optimize logistics, improve warehousing efficiency, and enhance customer satisfaction by providing accurate length predictions, which are critical for packaging, shipping, and minimizing product returns.

**Data visualisation:**

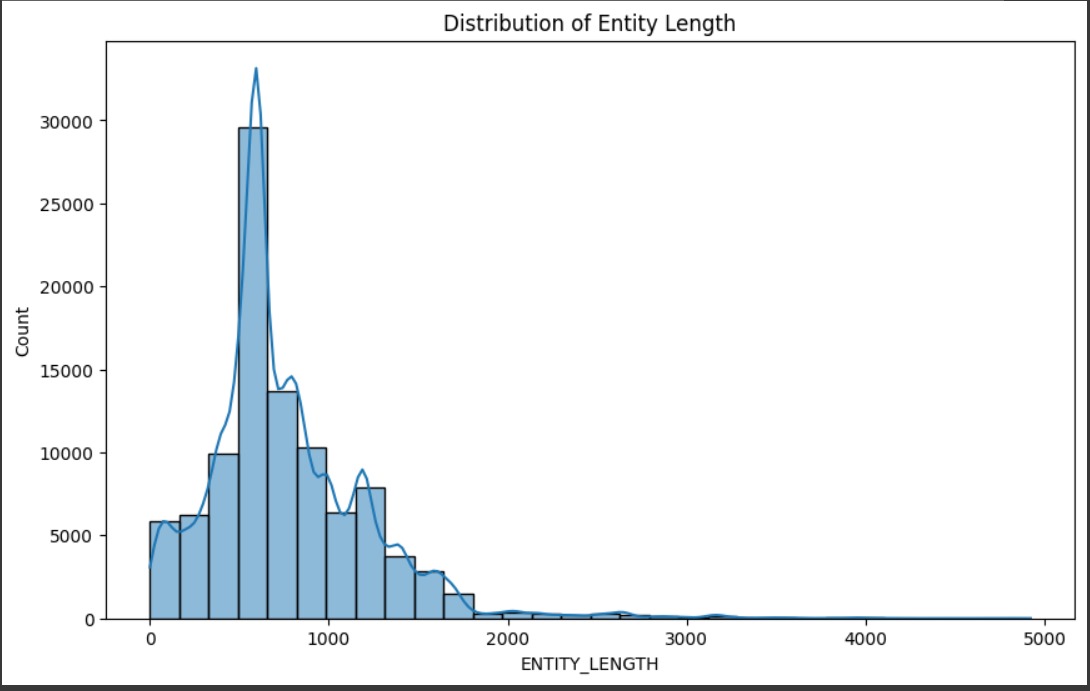
* Right Skewed:



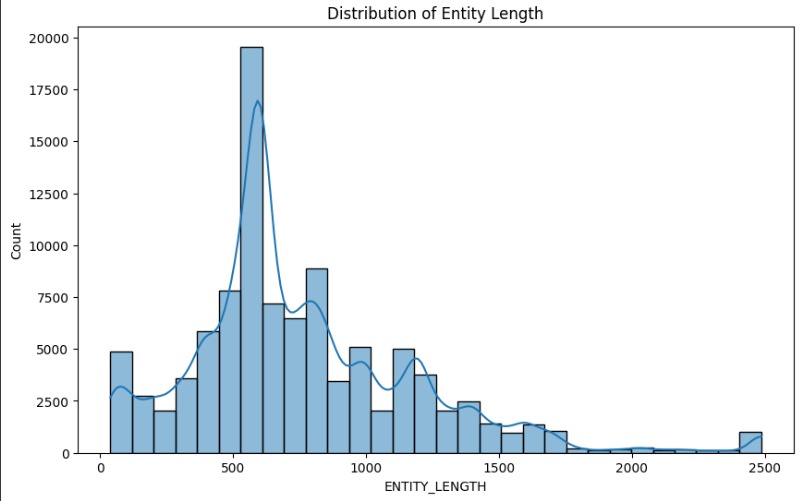
**Interpretation:**

The distribution is right-skewed, with most entities having lengths between 200 and 1000, and a few much longer outliers extending up to 5000.

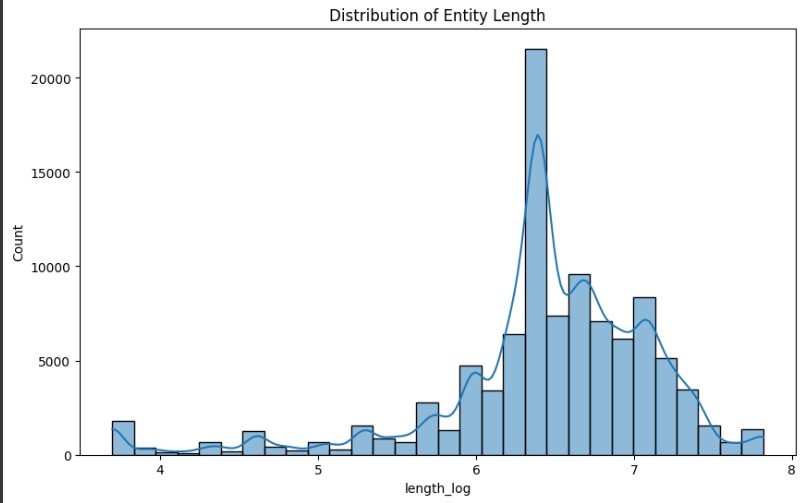
* Outlier treatment :



* Outlier capping

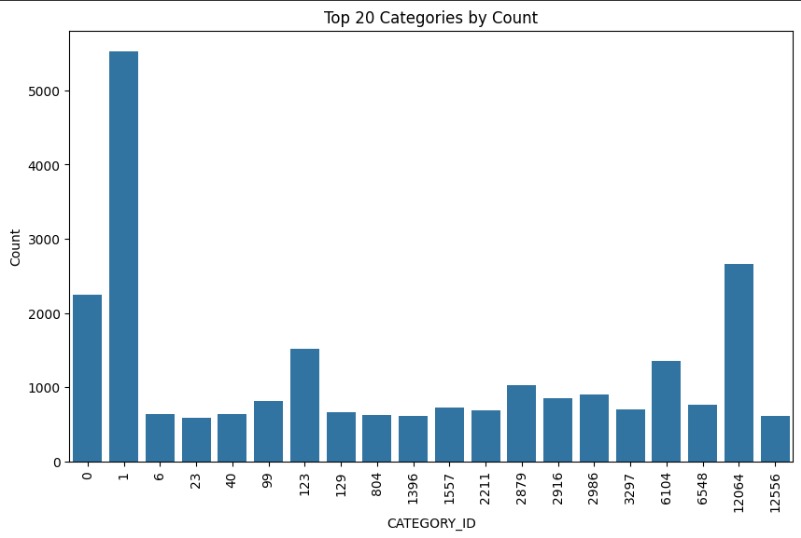


* Log normal

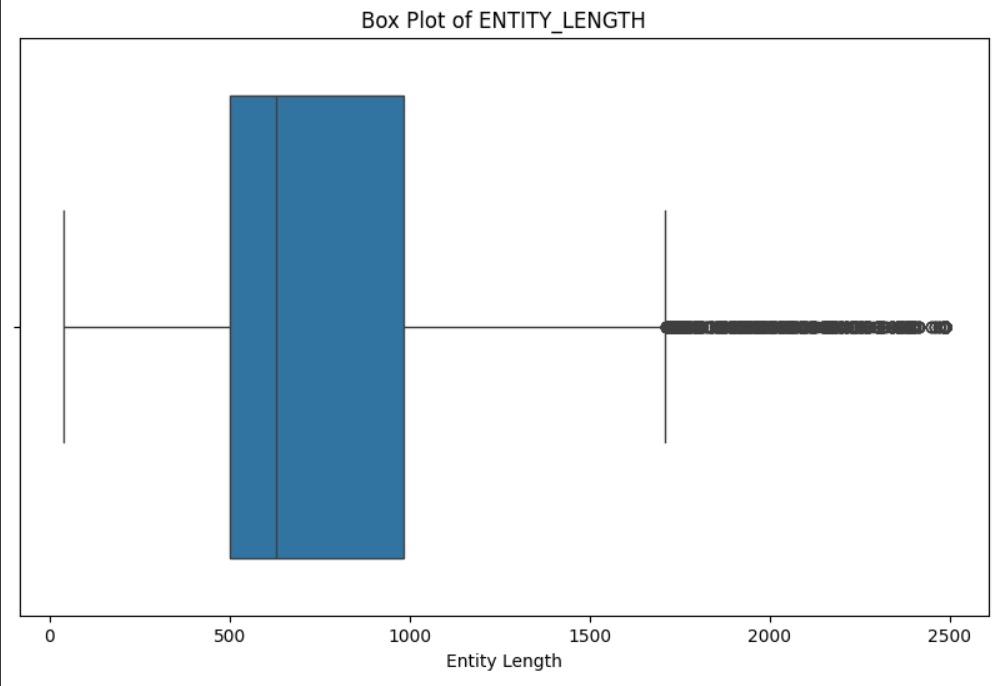


Interpretaion:

* Top 20 categories :



* With outliers:



* Without outliers:

