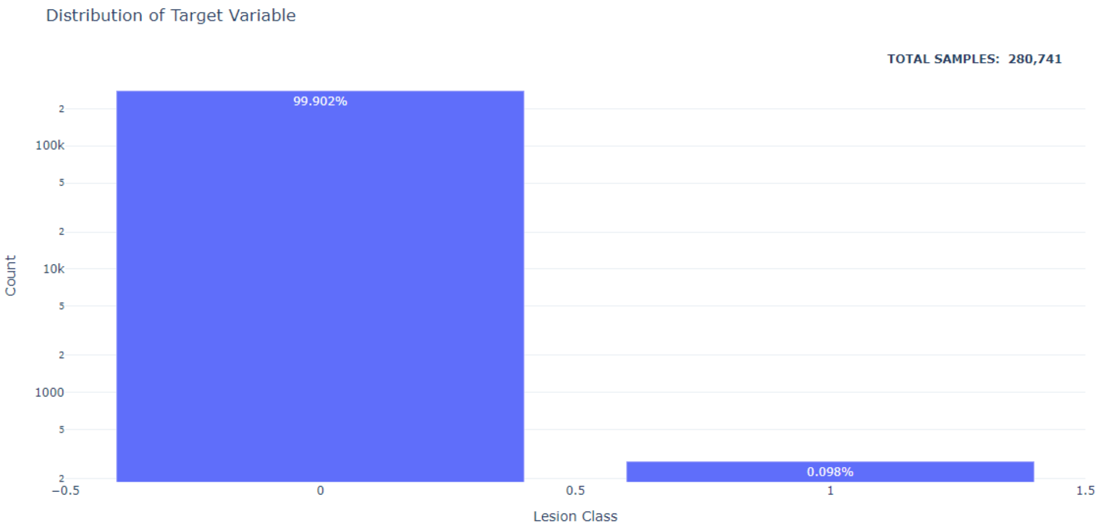
Data Partition

The dataset partition strategy used here divides the data into three sets: training, validation, and test, using a 70%-15%-15% split. First, 70% of the data is allocated to the training set, which allows the model to learn from the majority of the data. The remaining 30% is split evenly into validation and test sets, each receiving 15% of the original data. This allocation ensures a balanced approach where the model has enough data to train, while also preserving separate datasets for tuning and evaluation. Importantly, stratified splitting is applied, meaning the class distribution of the target variable is maintained across the splits. This is crucial when dealing with imbalanced datasets, such as those often found in medical tasks like skin cancer detection, where one class (e.g., benign cases) might dominate. This partition strategy effectively balances the needs for sufficient training data and unbiased model evaluation.

EDA

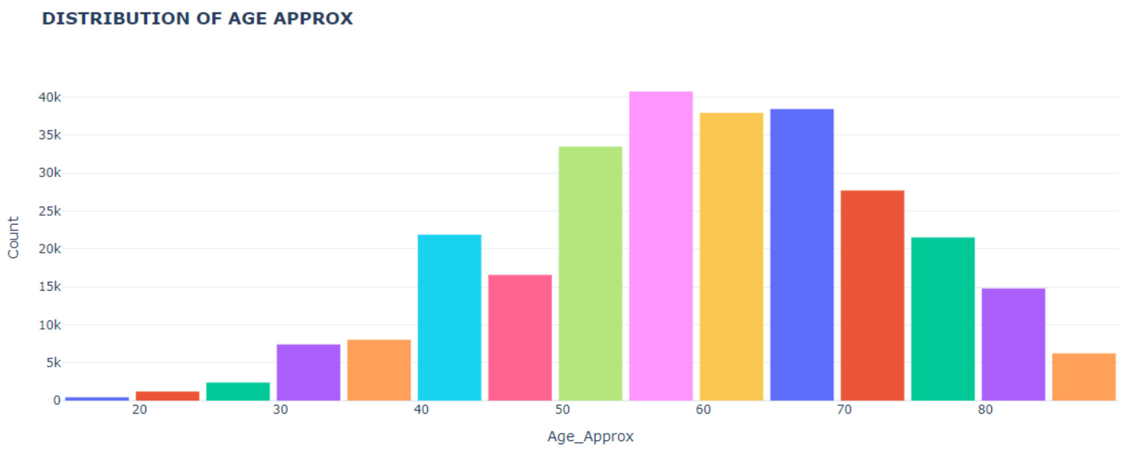
Target Variable Distribution

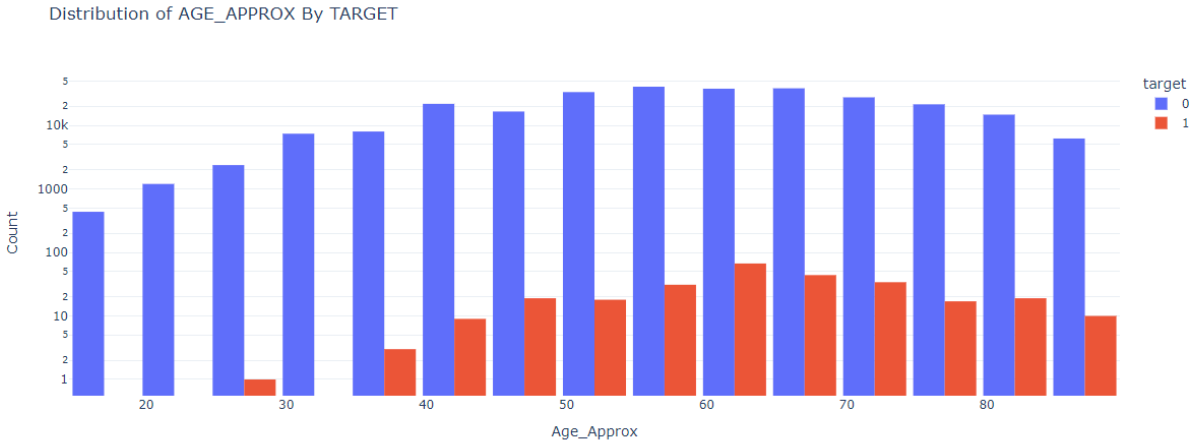


Insight: The target variable (0 for benign and 1 for malignant) is significantly imbalanced, with benign lesions comprising 99.902% of the dataset, while malignant lesions account for only 0.098%. This substantial class imbalance presents a challenge for the classification problem, as models may become biased toward predicting the majority class (benign lesions) and may struggle to accurately identify the minority class (malignant lesions). Addressing this imbalance will be crucial for improving the model's performance and ensuring reliable predictions.

Numerical Feature Analysis

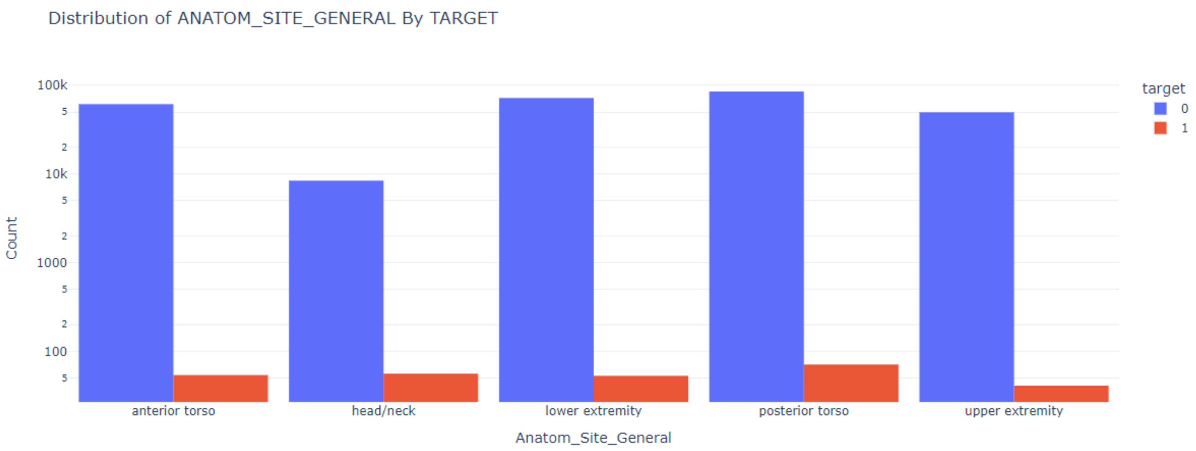
Age





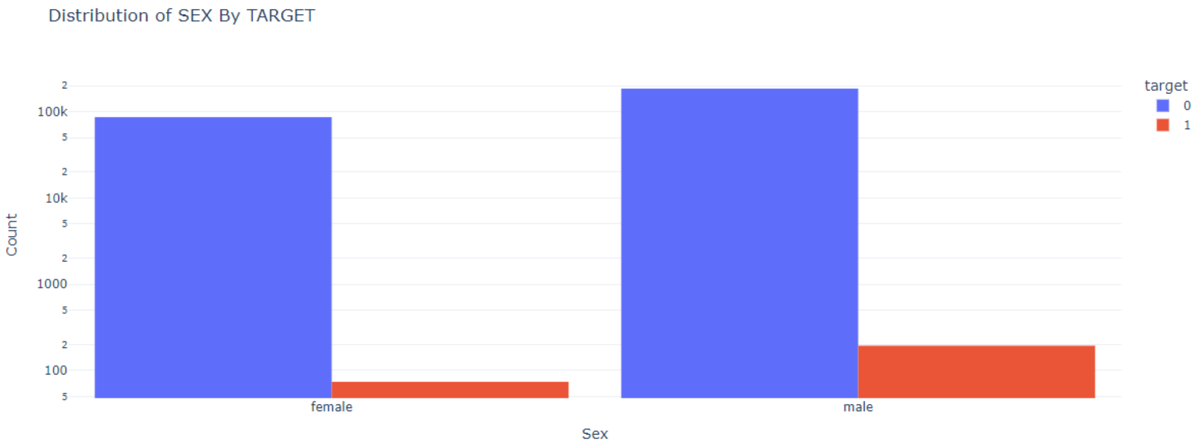
Insight: The majority of patients in the dataset are between the ages of 50 and 70, with patients aged 65 having the highest rate of malignant lesions. There are no malignant lesion cases for patients under the age of 25 and at the age 30.

Anatomical location of the lesion



Insight: The majority of malignant lesions are located on the posterior torso which is the back, indicating that this region may be particularly prone to skin cancers in the dataset.

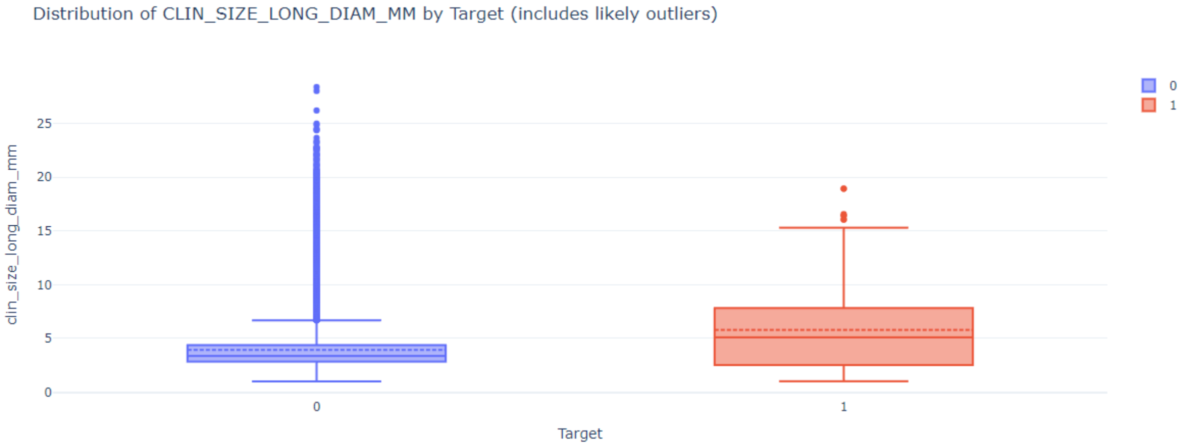
Sex

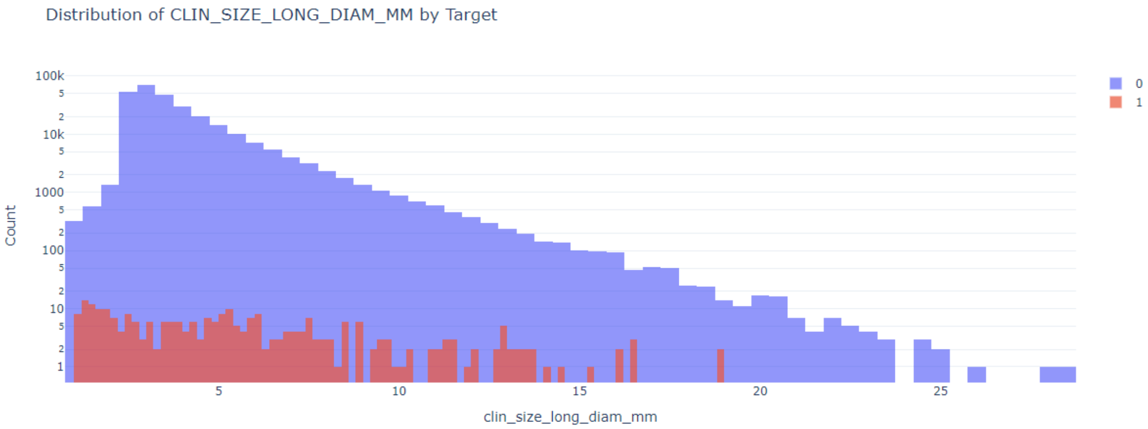


Insight: In this dataset, the majority of patients with malignant lesions are male, indicating a potential gender-based disparity in the occurrence of malignant skin lesions.

Categorical Feature Analysis

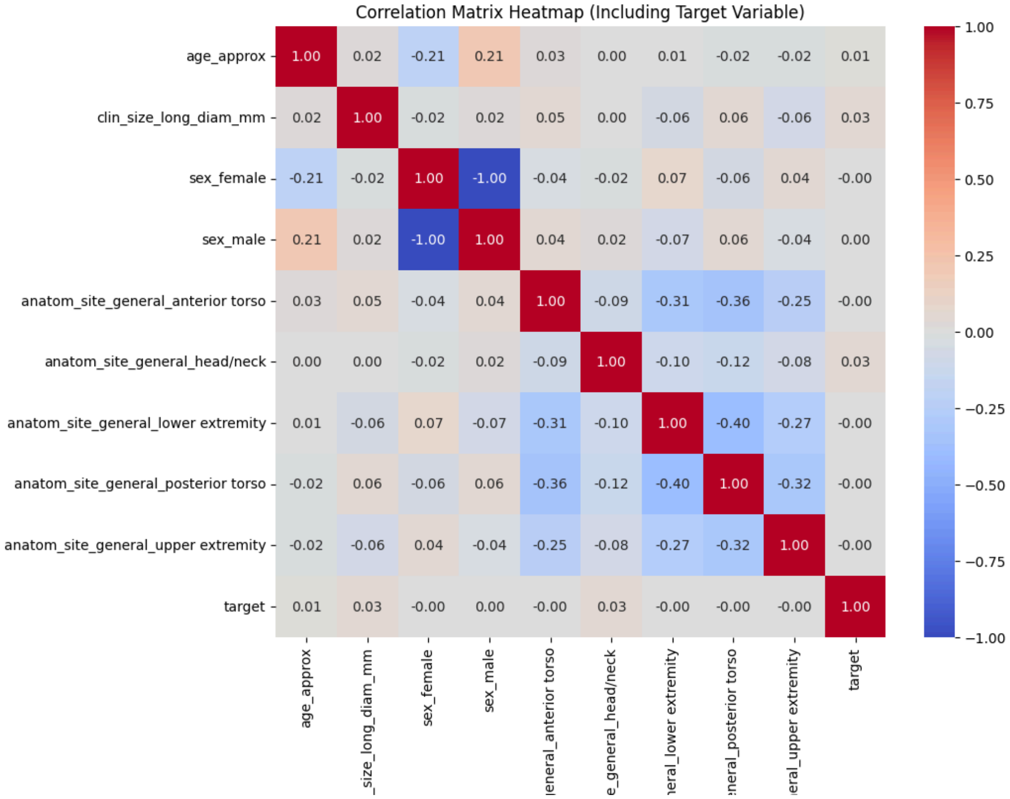
Maximum diameter of the lesion in millimeters


Distribution of CLIN_SIZE_LONG_DIAM_MM by Target (includes likely outliers)


Distribution of CLIN_SIZE_LONG_DIAM_MM by Target

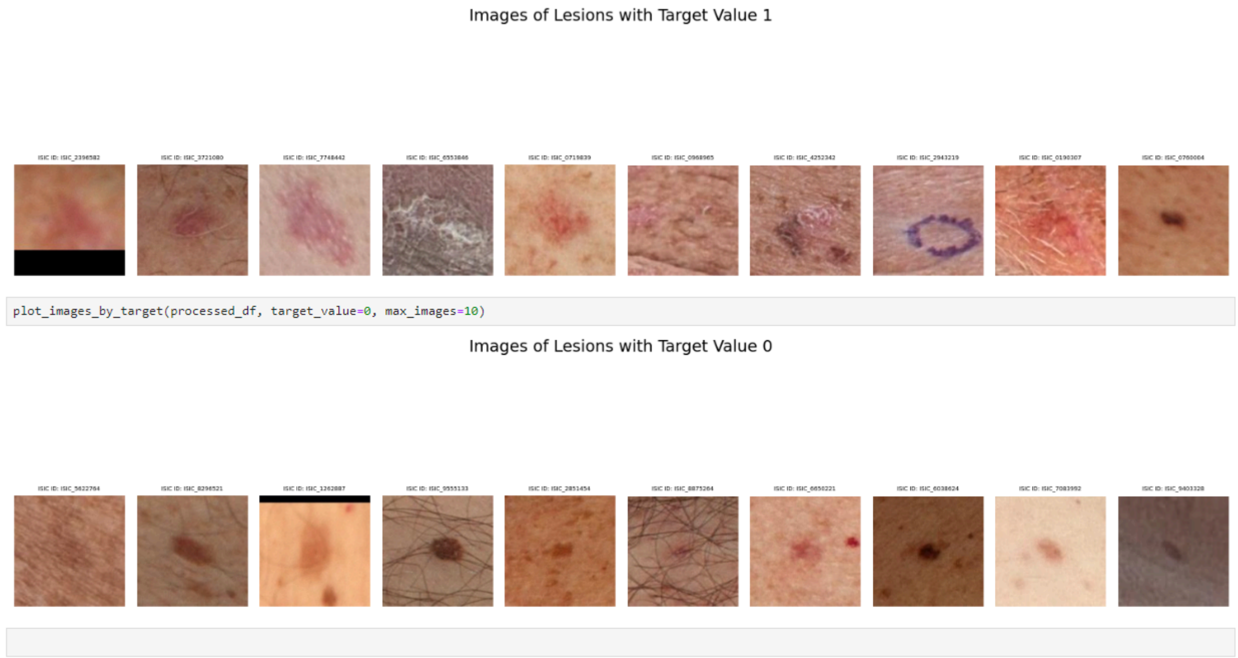
Insight: There are a large number of outliers in this variable, and the second graph of the distribution shows a right-skewed pattern. This indicates that most values are concentrated between 1 mm and 14 mm, with a few extreme values over 25 mm.

Correlation Analysis



Insight: There's no strong correlation between the variables in the dataset. We can disregard the male and female variables, as they are encoded and do not provide additional insight into the relationships among the other features.

Image Visualization


Images of Lesions with Target Value 1

```
plot_images_by_target(processed_df, target_value=0, max_images=10)
```

Images of Lesions with Target Value 0

Data Problems

Imbalance target variable

In the dataset, the primary data problem identified is the imbalanced target variable. The target variable is heavily skewed, with benign lesions making up 99.902% of the dataset, while malignant lesions account for only 0.098%. This imbalance can lead to biased model predictions, where the model may favor the majority class and struggle to generalize well for the minority class.

Outliers in continuous variable

Another significant data problem identified is the presence of outliers in certain numerical variables. These outliers can distort statistical analyses and adversely affect model training.

Recommendations

To address imbalance target variable, I recommend implementing upsampling techniques to increase the number of samples in the minority class. This can be achieved by either replicating existing samples or generating synthetic samples using methods such as SMOTE (Synthetic Minority Over-sampling Technique). Additionally, downsampling the majority class may also be considered to balance the dataset, although it may lead to the loss of valuable information. It is also crucial to use appropriate evaluation metrics that are sensitive to class imbalance, such as precision, recall, and F1-score, rather than relying solely on accuracy.

Besides that ,I recommend applying transformations, such as logarithmic or Box-Cox transformations, to reduce the skewness of the data and mitigate the impact of outliers.