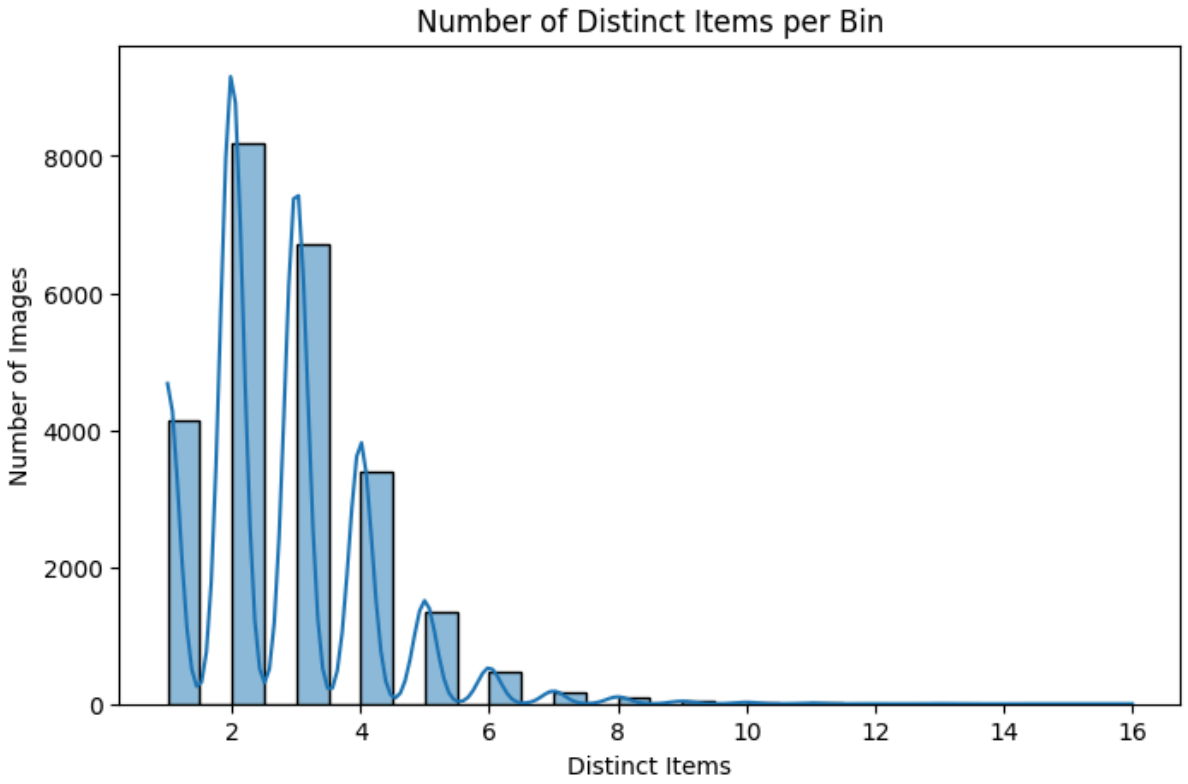
Exploratory Data Analysis

Dataset Scale & Diversity:

- The subset of the dataset contains **25,000 total images** and **66,297 total item entries**.
- There are **38,721 unique ASINs** (product classes) present in the data.
- The image count is 25,000, and the metadata file count is 24,999.
- Images without items are: **433**

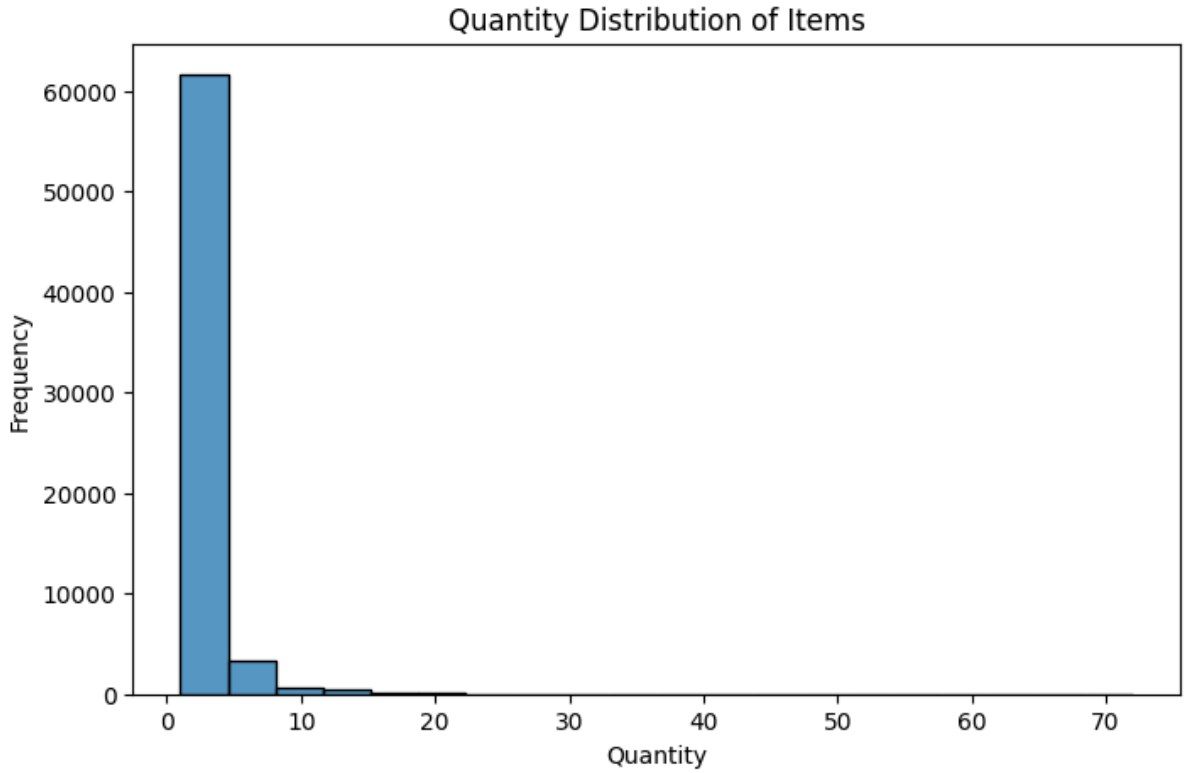
Items per Bin:

- The distribution of **distinct items per bin** is heavily concentrated at **1, 2, 3 and 4 distinct ASINs**.
- The frequency of bins containing 5 or more distinct ASINs drops off sharply.



ASIN Frequency per bin-image:

- The distribution of product frequency is **extremely skewed**, confirming a **long-tail distribution**.
- Most items in bins appear in **very small quantities** mostly between 1 and 3



Top 20 Most Frequent ASIN in bins:

- The distribution is **highly unbalanced**: a small subset of ASINs appears much more often than the rest.
- The most frequent item is appearing in **more than 35 bin images.**
- There are more than 38000 items in the dataset, but only around 66000 unique image-item pairs exist. Hence, the subsequent items appear in only maybe **1 or 2 images.**

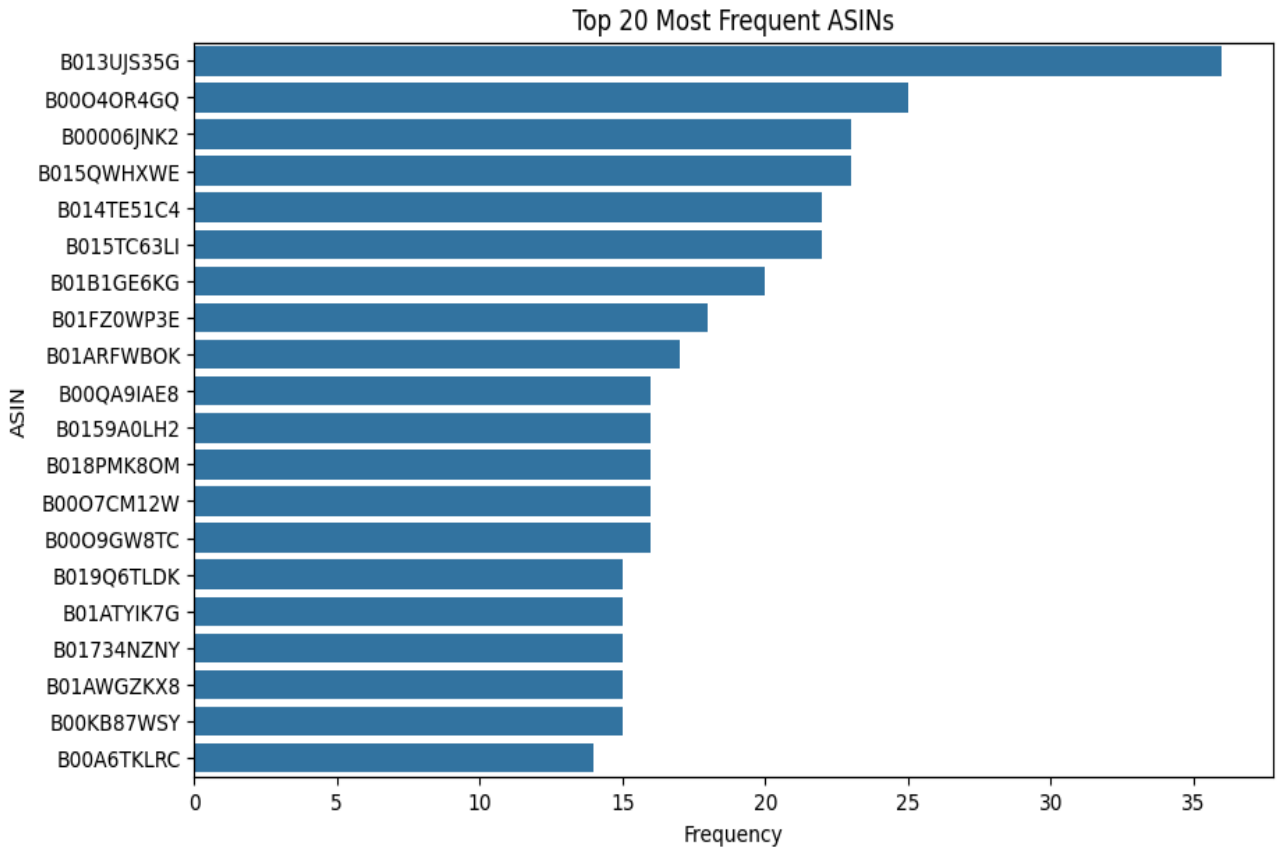
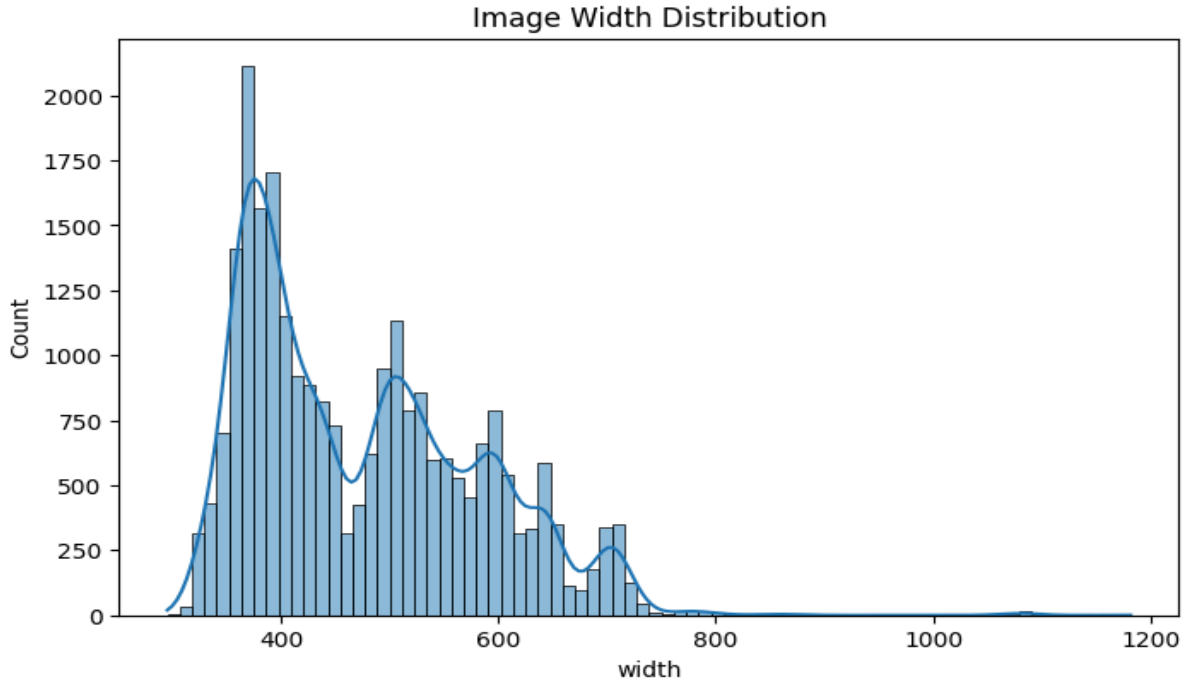
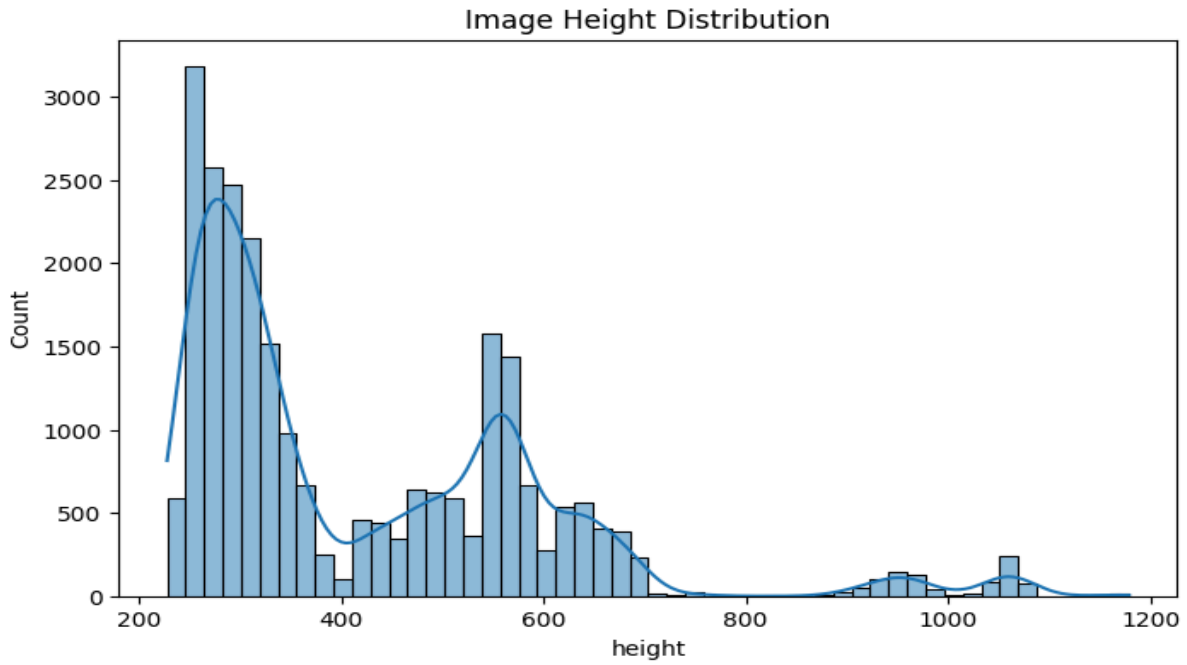


Image Resolution (Statistics & Histograms):

- Image resolutions are not uniform, ranging from a minimum of 295 x 228 pixels to a maximum of 1182 x 1179 pixels.
- The **Standard Deviation** for **height (≈ 177.71) is significantly higher** than for width (≈ 106.52).
- The **width** and **height** distribtions are **multi-modal** and **skewed**.





Data Alignment:

- A check for file consistency found **2 images without corresponding metadata files**.

Total Quantity per Bin:

- The distribution of **total unit quantity per bin** is **severely right-skewed**.
- The highest frequency is clustered around bins containing **1, 2, or 3 total units**.
- Rare outlier bins contain 40 total units.