EDA Documentation for Fashion Dataset

1. Data Type Analysis

Why: To understand what kind of data we're working with (numbers, text, categories).
Usage: Tells us which analysis methods to use and what tools we need. Most of our data is categorical (text) hence we don't use the techniques used for numerical data.

2. Missing Value Analysis

Why: To find gaps in our data where information is missing.
Usage:

a. Shows us data quality issues and helps decide how to handle missing information.
b. The 'usage' field has the most missing values (0.71%) which might need special attention.
c. We handled the missing data of target variables as they can give errors in training if a particular class of a target variable is present in the training set but not in the test/validation set.
d. **usage** has the highest missing values (317 records, 0.71%)
e. **season** has 21 missing values (0.05%)
f. **baseColour** has 15 missing values (0.03%)
g. **productDisplayName** has 7 missing values (0.02%)
h. **year** has 1 missing value (0.002%)
i. The remaining columns (id, gender, articleType, masterCategory, subCategory) have no missing values

3. Unique Value Analysis

Why: To see how many different values each column has.
Usage:

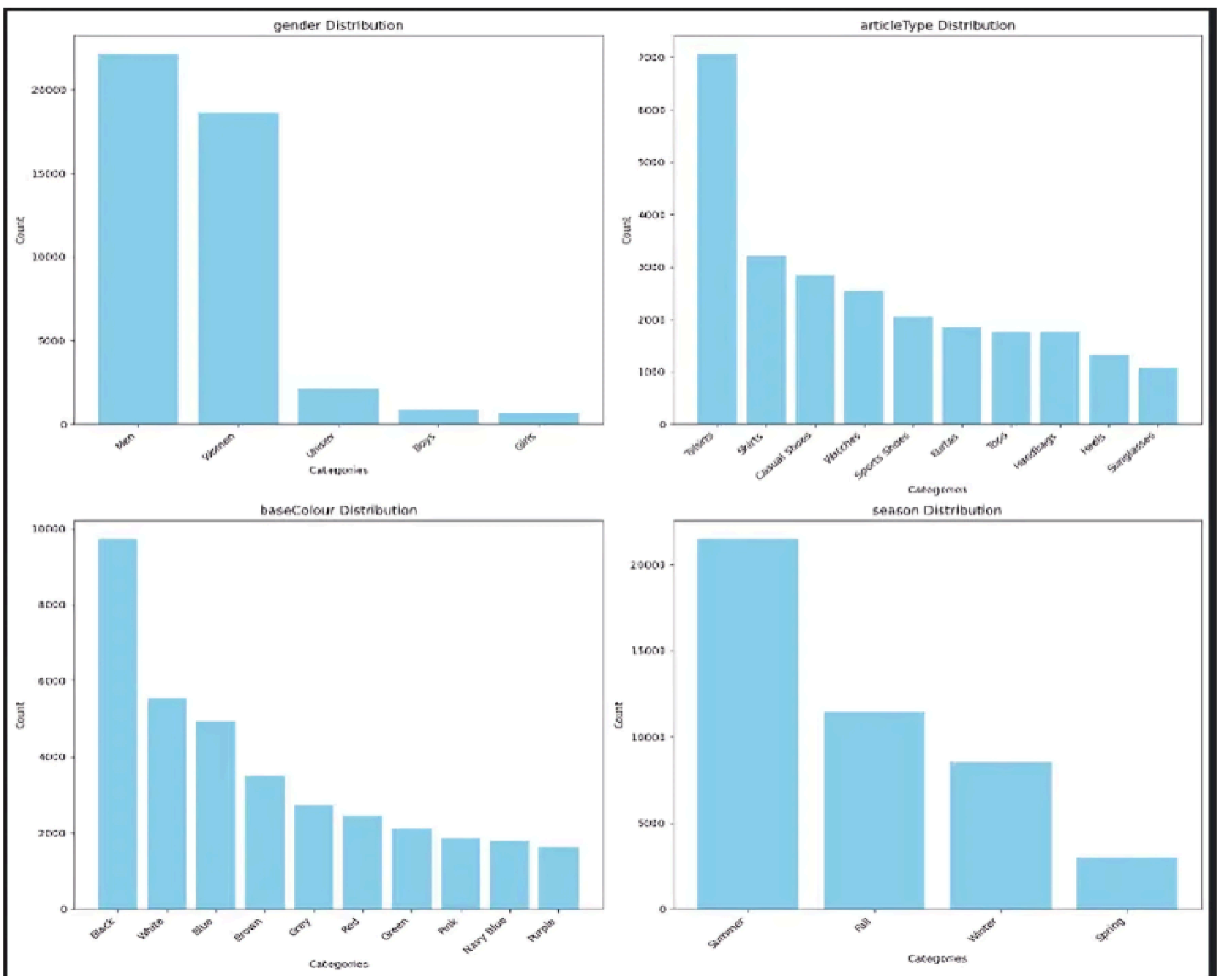
a. Helps us understand data diversity and complexity.
b. **id** has 44,424 distinct values (100% - all unique identifiers)
c. **productDisplayName** has 31,121 distinct values (70.05% - high variety in product names)
d. **articleType** has 143 distinct values (0.32% - good variety in article types)
e. **baseColour** has 46 distinct values (0.10% - moderate color options)
f. **subCategory** has 45 distinct values (0.10% - good subcategory diversity)
g. **year** has 13 distinct values (0.03% - covers 13 different years)
h. **usage** has 8 distinct values (0.02% - limited usage categories)
i. **masterCategory** has 7 distinct values (0.02% - few main categories)
j. **gender** has 5 distinct values (0.01% - likely Men/Women/Kids/Unisex/Other)

k. **season** has 4 distinct values (0.01% - Spring/Summer/Fall/Winter)

4. Target Feature Analysis

Why: To visualize how our data is distributed across different categories.
Usage:

a. Shows us patterns and imbalances in our data.
b. **Gender Distribution**: Shows Men have the highest count (~22k), followed by Women (~18k), with much smaller counts for Unisex, Boys, and Girls
c. **ArticleType Distribution**: Shows Shirts have the highest count (~6k), followed by Tshirts, Casual Shoes, Watches, Sports Shoes, Kurtas, Tops, Handbags, Heels, and Sunglasses in descending order
d. **BaseColour Distribution**: Shows Black has the highest count (~10k), followed by Blue, White, Navy Blue, Grey, Red, Brown, Green, Pink, Maroon, and Yellow
e. **Season Distribution**: Shows Summer has the highest count (~25k), followed by Fall (~12k), Winter (~8k), and Spring (~2k)

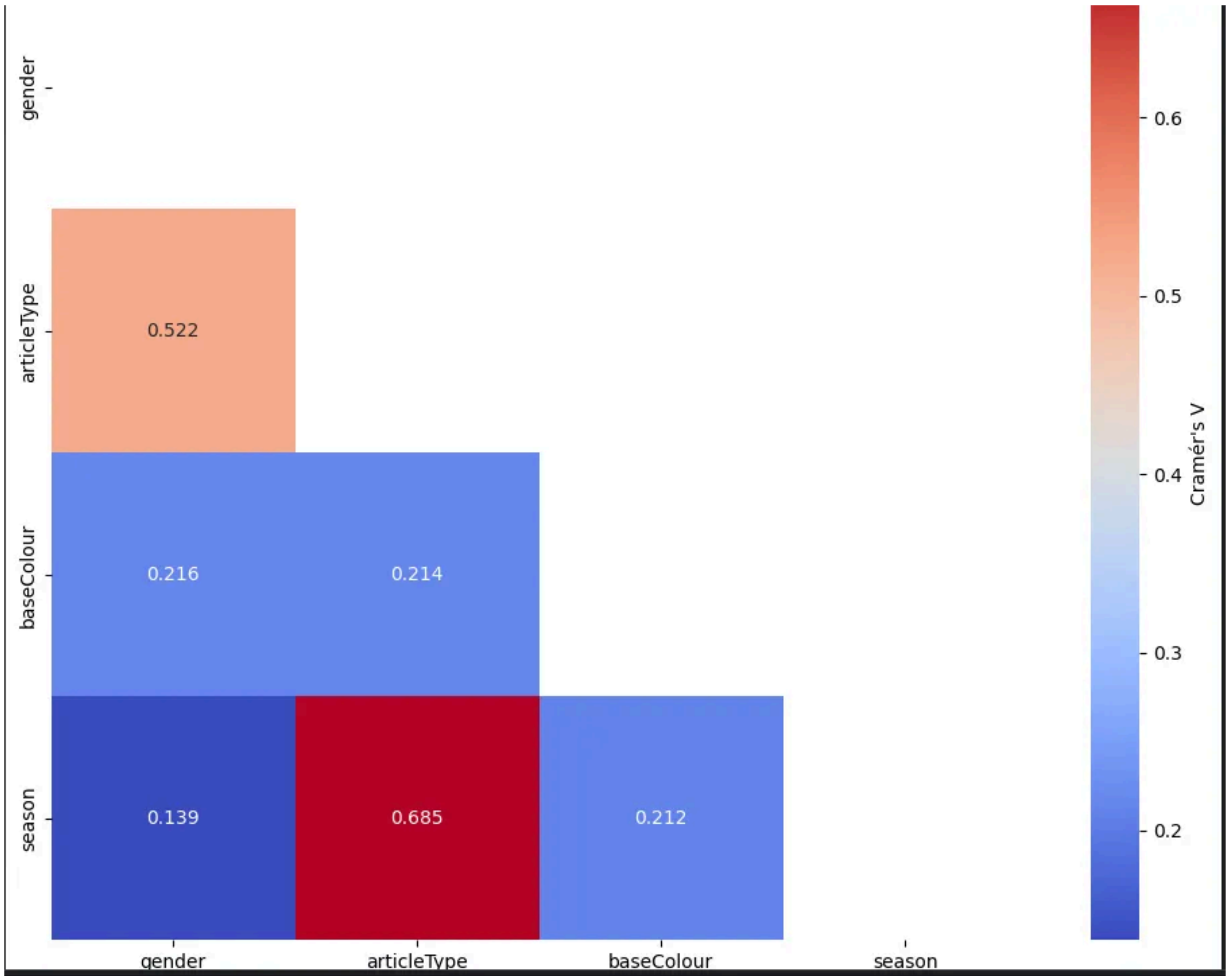


5. Cramer's V Correlation Analysis

Why: To measure how strongly different categorical variables are related to each other.

Usage:

a. Identifies which features influence each other.
b. Season and article type are strongly connected (0.685), meaning certain clothes are seasonal.



6. Chi-Square Independence Test

Why: To statistically prove whether relationships between variables are real or just coincidence.

Usage:

a. Confirms our correlation findings with statistical proof.
b. All relationships are significant, meaning they're genuine patterns, not random.
c. All relationships are statistically significant ($p < 0.001$)

Chi-square Test Results:

1. **gender ↔ articleType**:
   - Chi² = 48,981.826, $p < 0.001$, Cramer's V = 0.522 (Strong, Significant)
2. **gender ↔ baseColour**:
   - Chi² = 8,455.507, $p < 0.001$, Cramer's V = 0.216 (Weak, Significant)
3. **gender ↔ season**:
   - Chi² = 2,572.504, $p < 0.001$, Cramer's V = 0.139 (Weak, Significant)
4. **articleType ↔ baseColour**:
   - Chi² = 97,810.092, $p < 0.001$, Cramer's V = 0.214 (Weak, Significant)
5. **articleType ↔ season**:
   - Chi² = 63,002.398, $p < 0.001$, Cramer's V = 0.685 (Strong, Significant)
6. **baseColour ↔ season**:
   - Chi² = 6,112.383, $p < 0.001$, Cramer's V = 0.212 (Weak, Significant)

7. Image Quality Statistics

Why: To check if our product images are good quality and consistent.
Usage:

a. Ensures our visual data is reliable for analysis.
b. High brightness and low blur scores show professional product photography, making image analysis more accurate.

```
Image Quality Statistics:
Brightness: 0.842 ± 0.086
Contrast: 0.254 ± 0.082
Blur Score: 0.005 ± 0.003
```

8. Baseline Accuracy Metrics

Why: To establish minimum performance standards for any prediction models we build.
Usage:

a. Sets benchmarks for model evaluation.
b. Any AI model we create should perform better than these simple baselines (like 49.85% for gender prediction) to be considered useful.

```
Most common class: Men
Baseline Accuracy: 0.4985
Most common class: Tshirts
Baseline Accuracy: 0.1591
```

```
Most common class: Black
Baseline Accuracy: 0.2191
Most common class: Summer
Baseline Accuracy: 0.4838
```