Overview

This project analyzes a dataset containing reviews of musical instruments from Amazon. It provides insight into the distribution of ratings, recommendations, and feedback count. The scripts are composed of data loading, cleaning, processing, analysis, and visualization tasks to understand patterns within the data. The dataset has several features which are: reviewerID, asin, reviewerName, helpful, reviewText, overall, summary, unixReviewTime, and reviewTime. Here is a summary of the features and preprocessing steps taken on them:

- **reviewerID**: A unique identifier for each reviewer.
- **asin**: The Amazon Standard Identification Number - a unique identifier for products on Amazon.
- **reviewerName**: The name of the reviewer. This column has a few missing values which were filled with 'Anonymous'.
- **helpful**: Originally contained the count of people who found the review helpful and total feedback. This was separated into 'helpful_votes' and 'total_votes' columns.
- **reviewText**: The text content of the review. Missing values were filled with 'No review text provided'.
- **overall**: The overall rating given to the product by the reviewer. Ratings range from 1 to 5 and were used to derive a 'recommended' column, where ratings of 4 and above were coded as 'Recommended' and below 4 as 'Not Recommended'.
- **summary**: A brief summary of the review.
- **unixReviewTime**: The time of the review in Unix time. This was converted to datetime format in 'reviewTime'.
- **reviewTime**: The time of the review in a more human-readable format.
- **year_month**: A derived column which contains the year and month of the review.
- **recommended**: A derived column indicating whether the product is recommended or not based on the overall rating.
- **total_votes** and **helpful_votes**: These columns were derived from the 'helpful' column, indicating the total feedback count and the count of people who found the review helpful, respectively.
- **Feedback Category**: A derived column categorizing 'total_votes' into 'Low Count', 'Medium Count', and 'High Count'.

Dependencies

This script uses the pandas library for data handling, as well as seaborn and matplotlib libraries for visualization. Custom classes **DataProcessor**, **DataAnalyzer**, **DataVisualizer**, and **ColumnAnalyzer** are used to develop needed methods separately.

Key Functions, Operations and sample outputs

1. **Data Loading:** The '**DataProcessor'** class is used to load the dataset. It reads a CSV file and returns a pandas DataFrame.
2. **Data Cleaning:** '**DataProcessor'** handles missing values, removes duplicates after initial preprocessing, and converts specific columns to the appropriate format for analysis. It also splits the 'helpful' column into two separate 'helpful_votes' and 'total_votes' columns. Missing values are 27 reviewer names and 7 review text. Rows with missing values were dropped using dropna() method.

```
Missing Values:
 reviewerID         0
asin                0
reviewerName       27
helpful             0
reviewText          7
overall             0
summary             0
unixReviewTime      0
reviewTime          0
dtype: int64
Duplicates:  0
```

Figure 1 Missing Values and Duplicates analysis

3. **Recommendations Column Creation:** A new column named 'recommended' is created based on the 'overall' rating of each review in the main script. A rating of 4 or above is considered a recommendation.
4. **Time Conversion and Grouping:** The 'unixReviewTime' column is converted to datetime format and a new 'year_month' column is created for analyzing average ratings per month.

5. **Feedback Categories:** The '**ColumnAnalyzer'** class handles the analysis of the 'total_votes' column. It generates a histogram and box plot for the column, checks its skewness, and categorizes it into three feedback categories.

```
Summary Statistics for total_votes:
count    10227.000000
mean         1.828982
std          9.245187
min          0.000000
25%          0.000000
50%          0.000000
75%          1.000000
max        300.000000
Name: total_votes, dtype: float64

Skewness of total_votes: 16.0763993842465
```

Figure 2 total_values column Statistics

6. **Data Analysis:** Various data analysis operations are performed using the '**DataAnalyzer'** class. These include calculating the average rating per month, getting the distribution, median, and mode of ratings and recommendations, and identifying the distribution of feedback categories.

```
Average Rating per Month:
 reviewTime
2004-09-18    4.0
2004-09-29    5.0
2004-11-29    5.0
2004-12-01    5.0
2005-01-28    5.0
              ...
2014-07-16    4.5
2014-07-19    5.0
2014-07-20    5.0
2014-07-21    4.0
2014-07-22    4.0
Name: overall, Length: 1569, dtype: float64
Median of Recommendations:  1.0
Mode of Recommendations:  1
Mode of Feedback Count:  0
```

Figure 3 Key statistics on Ratings, Recommendations and Feedback Count

7. **Data Visualization:** The '**DataVisualizer'** class provides methods to plot a line chart, a histogram, a pie chart, and a scatter plot. In this script, it is used to visualize the distribution of ratings, recommendations, and feedback categories.

Main Findings

- **Average Rating per Month:** The average rating was consistently high, often above 4.0 as shown in figure 3, from September 2004 to July 2014.
- **Recommendations:** More than half of the products were recommended, as shown by a median and mode of 1.0 as shown in figure 3. This indicates that most reviewers recommend the products.
- **Feedback Count:** Most of the reviews had a low feedback count, as indicated by the mode of the 'feedback_count' column being 0 shown in figure 3. This suggests that while many reviews are written, not many receive feedback. This is supported by the fact that the skewness of 'total_votes' is 16.0763 as shown in figure 2, indicating a right-skewed distribution. This suggests most reviews receive little feedback, but a few have exceptionally high feedback counts.

Main Visualizations

- **Distribution of Ratings:** A bar chart is used to represent the distribution of ratings, with each bar indicating a different rating value. The length of the bars indicates the frequency of each rating in the dataset. The longest bar corresponds to the most common rating, suggesting that most customers gave a rating of this value. This distribution of ratings can provide useful insights into overall customer satisfaction and product quality. The chart indicates that most users gave a higher rating and are pleased with their purchase.

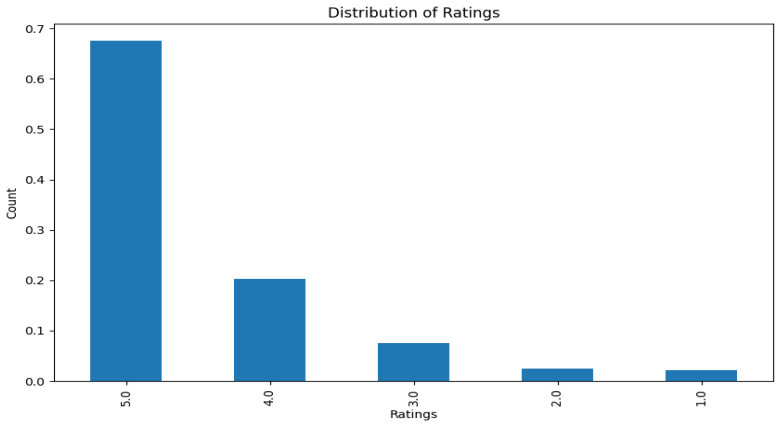

Figure 4

- **Distribution of Recommendations:** A pie chart represents the distribution of recommendations, showing that 87.9% of products were recommended by the reviewers.

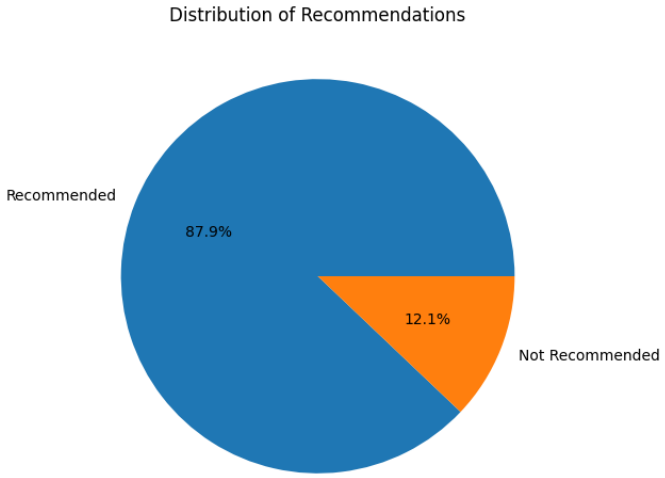

Figure 5

- **Distribution of Feedback Categories:** A pie chart displays the distribution of feedback counts, categorized into 'Low Count', 'Medium Count', and 'High Count'. This provides a clearer understanding of how much feedback each review typically receives.

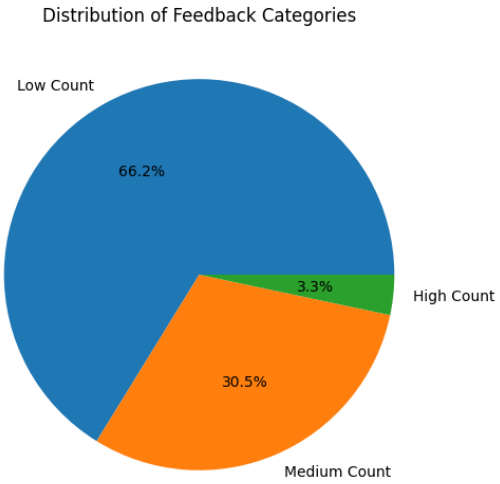

Figure 6