From Bean to Brew: Analyzing Coffee Ratings and Market Trends

Anumol Issac

Objectives

❖**Understand the Dataset:**

- Explore the structure and quality of the data.
- Identify data issues (e.g., missing values, duplicates).
- Prepare a clean and structured dataset.

❖**Answer Key Business Questions:**

- Identify popular products, roasters and get insights about seasonal trends.
- Compare average ratings across regions.
- Identify popular roast types and their regional distribution.
- Analyze correlations between sensory attributes and ratings.

❖**Create Visualizations for Data Interpretation:**

- Develop interactive dashboard using tableau to visualize the key findings.

❖**Develop and Deploy Predictive Models :**

- Develop machine learning models to predict ratings and classify products into popularity tiers.

❖**Provide Insights for Business Strategy:**

- Offer actionable recommendations for marketing, sales, and product strategies.

Analytical Approach

❖**Data Exploration and Quality Check**

- Understand the structure and identify data quality issues.
- Explore relationships between columns across datasets and verify unique identifiers.

❖**Data Cleaning and Transformation**

- Prepare a clean and structured dataset
- Handle missing data and impute values where necessary.
- Standardize ratings, add derived columns (review_year, review_month, normalized_rating, popularity_tier) and calculate averages.

❖**Exploratory Data Analysis (EDA)**

- Analyze trends and distributions (ratings, product popularity, roast preferences, etc.).

❖**Time-Series Analysis**

- Analyze seasonal or yearly trends based on ratings.

❖**Regional Analysis**

- Compare ratings and preferences across different regions.

Analytical Approach(Cont.)

❖**Roast Preference Analysis**

- Determine the most popular roast types and their distribution.

❖**Correlation Analysis**

- Investigate relationships between sensory attributes (e.g., flavor, aroma etc.) and ratings.

❖**Predictive Modeling**

- Build regression models to predict ratings based on roast type and region.
- Use decision trees to classify products into popularity tiers.

❖**Advanced Reporting and Visualization**

- Create interactive dashboards to visualize trends and forecast ratings.

❖**Business Strategy Recommendations**

- Provide actionable insights and recommendations for marketing and sales strategies.

Data Summary

Data schemas

coffee_id.csv

- **slug**: Product identifier
- **name**: Product name
- **roaster**: Coffee roaster
- **rating**: Product rating
- **review_date**: Review date

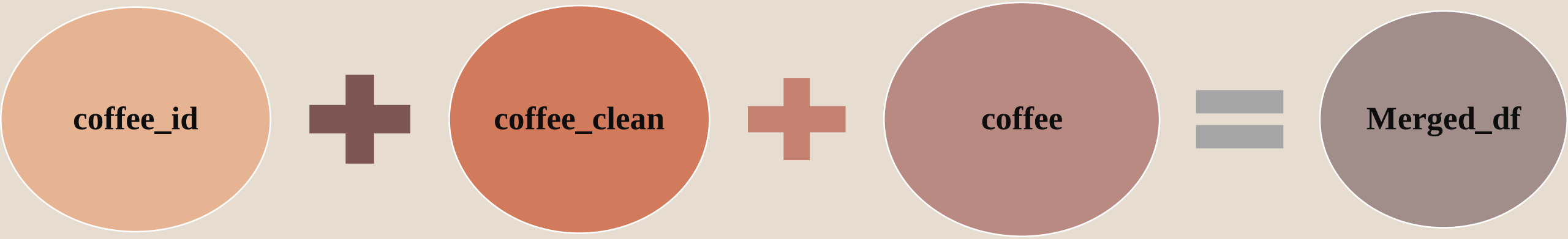
coffee_clean.csv

- **slug**: Product identifier
- **sensory attributes**: Sensory details like aroma and aftertaste
- **roast types**: (medium-light, medium, dark)
- **regions**: Africa, Asia pacific etc,
- **clean_text**
- **type attributes**: Organic, Fair Trade, Decaffeinated, etc.

coffee.csv

- **all_text**: Web-scraped text
- **name**: Product name
- **rating**: Product ratings
- **roaster**: Coffee roaster
- **slug**: Product identifier
- **regions**: Africa, Asia pacific etc.
- **type attributes:** Organic, Fair Trade etc.
- **location and origin**
- **est_price:** estimated price
- **review_date**: Date of the review
- **roast**: (medium-light, dark etc.)
- **sensory attributes**: Sensory details like aroma and aftertaste

Create a Structured Dataset



❖ **Data Preparation:**

- Cleaned slug column in ‘coffee’ to remove unnecessary text.
- Removed duplicate column (type_with_milk.1) and unwanted columns from ‘coffee_clean’.

❖**Data Merging**:

- Merged ‘coffee_id’ and ‘coffee_clean’ on slug.
- Filtered and cleaned extra entries in ‘coffee’.
- Combined these datasets into a single structured ‘merged_df’.

❖**Final Output**:

- structured dataset for analysis by integrating and cleaning all three datasets.

Data Cleaning and Transformation

Handle Missing Data

- Impute missing ratings using average rating by roaster
- Impute missing values in 'aftertaste' column with mean
- Drop rows where 'roast' column has missing values

Normalize ratings

- Normalize ratings to a 1-10 scale using min-max normalization
- normalized_rating = ((rating - min_rating) / (max_rating - min_rating)) * (new_max - new_min) + new_min
- new_max=10 , new_min=1

Derived Columns

- normalized_rating based on min-max normalization
- review_month and review_year form review date
- popularity_tier based on normalized rating
- origin_derived from origin

Final Dataset Overview

```
Data columns (total 42 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   slug                    4252 non-null   object
 1   name                    4252 non-null   object
 2   roaster                 4252 non-null   object
 3   rating                  4252 non-null   float64
 4   review_date             4252 non-null   datetime64[ns]
 5   aroma                   4252 non-null   float64
 6   acid_or_milk            4252 non-null   float64
 7   body                    4252 non-null   float64
 8   flavor                  4252 non-null   float64
 9   type_with_milk          4252 non-null   int64
 10  roast_dark              4252 non-null   int64
 11  roast_light             4252 non-null   int64
 12  roast_medium            4252 non-null   int64
 13  roast_medium_dark       4252 non-null   int64
 14  roast_medium_light      4252 non-null   int64
 15  roast_very_dark         4252 non-null   int64
 16  roast_nan               4252 non-null   int64
 17  region_africa_arabia    4252 non-null   int64
 18  region_caribbean        4252 non-null   int64
 19  region_central_america  4252 non-null   int64
 20  region_hawaii           4252 non-null   int64
```

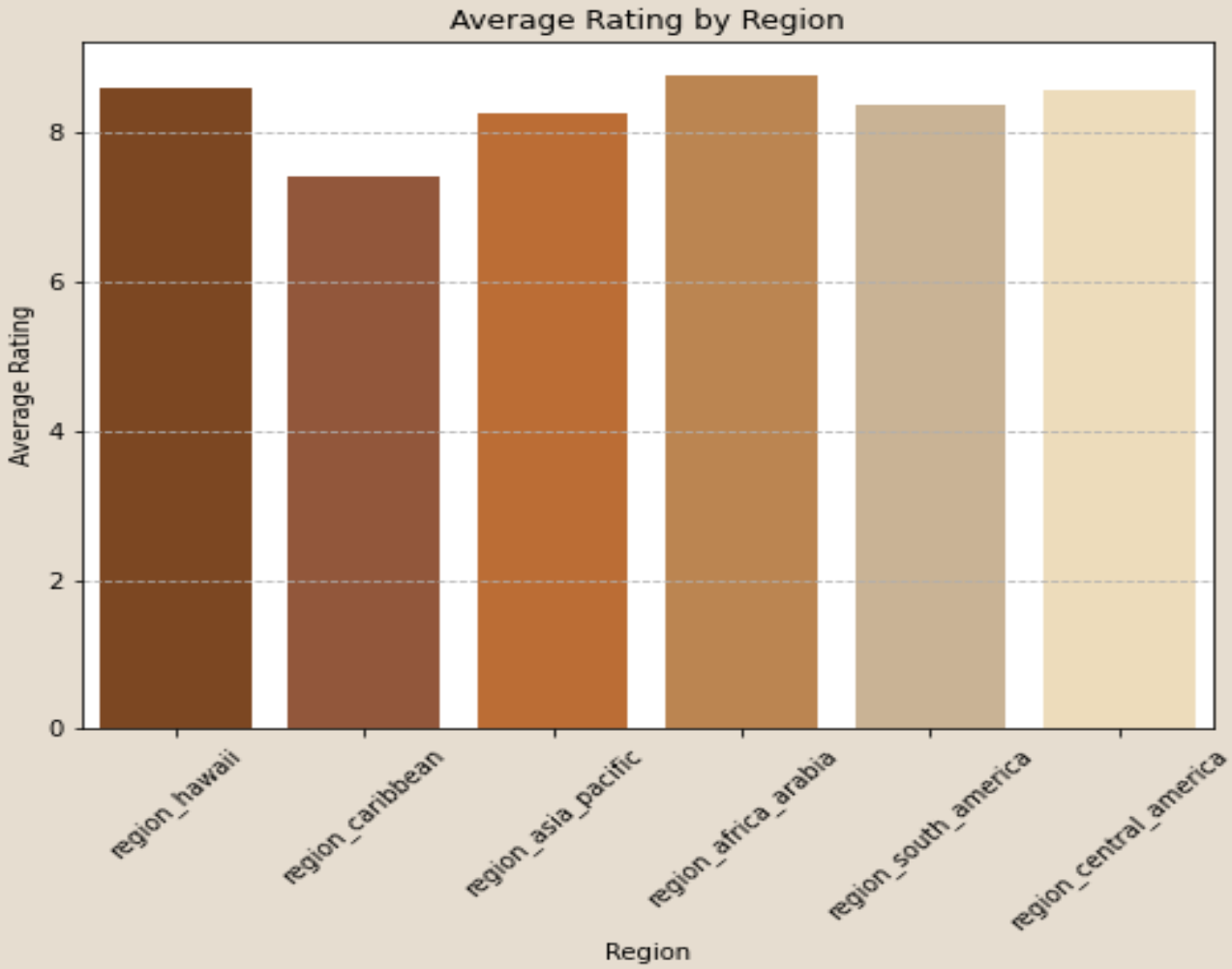
```
 21  region_asia_pacific     4252 non-null   int64
 22  region_south_america    4252 non-null   int64
 23  type_espresso           4252 non-null   int64
 24  type_organic            4252 non-null   int64
 25  type_fair_trade         4252 non-null   int64
 26  type_decaffeinated      4252 non-null   int64
 27  type_pod_capsule        4252 non-null   int64
 28  type_blend              4252 non-null   int64
 29  type_estate             4252 non-null   int64
 30  location                4251 non-null   object
 31  origin                  4252 non-null   object
 32  roast                   4252 non-null   object
 33  est_price               2922 non-null   object
 34  agtron                  4252 non-null   object
 35  acid                    3598 non-null   float64
 36  aftertaste              4252 non-null   float64
 37  with_milk               670 non-null    float64
 38  normalized_rating       4252 non-null   float64
 39  origin_derived          4252 non-null   object
 40  review_year             4252 non-null   int32
 41  review_month            4252 non-null   object
types: datetime64[ns](1), float64(9), int32(1), int64(21), object(10)
```

Total records: 4252

Total variables: 42

Business Questions and Insights

Average Rating Per Region

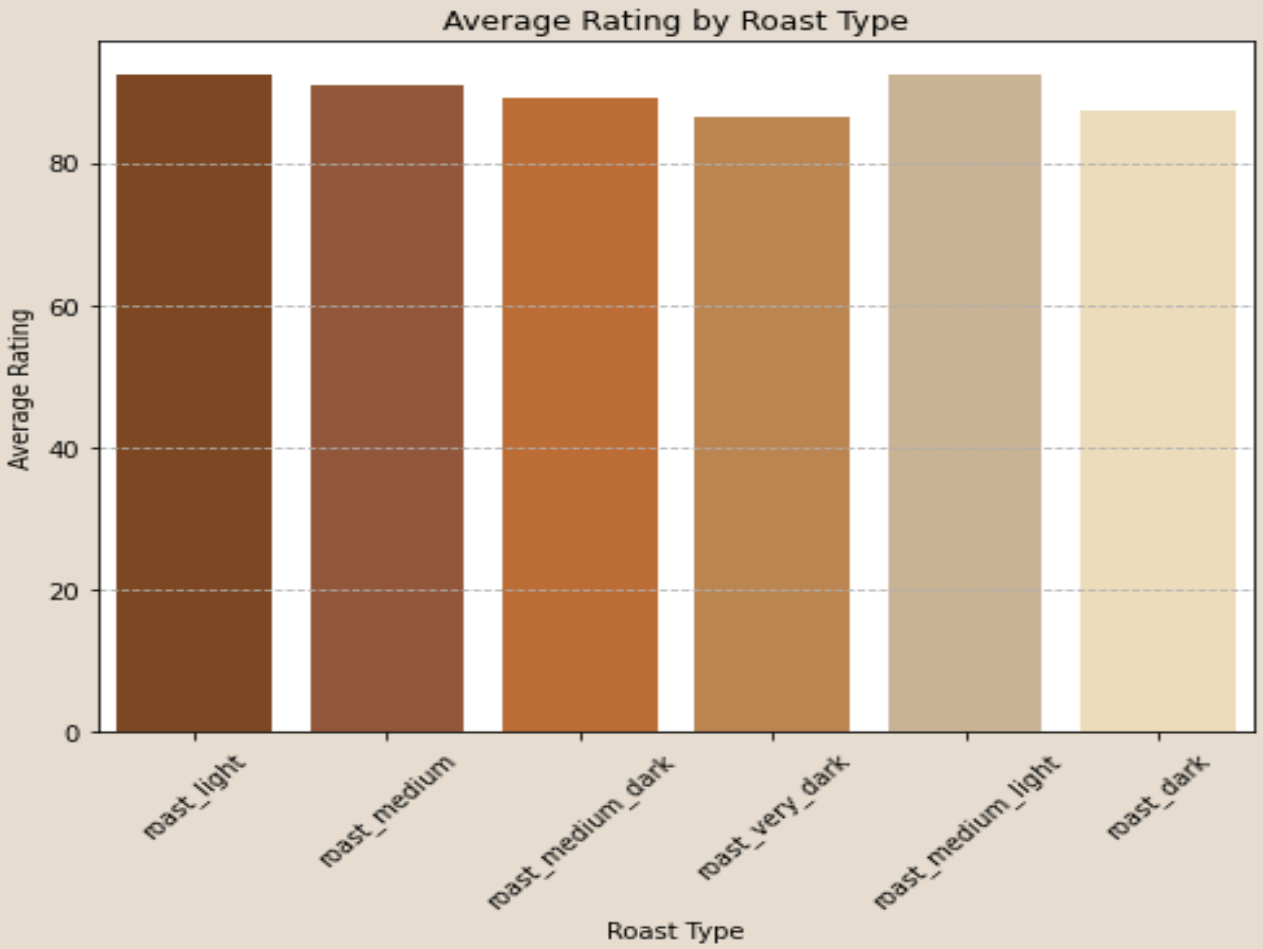


Africa Arabia has highest average rating of 8.78.

Hawaii comes next with an average rating of 8.59.

Central America follows with a rating of 8.59.

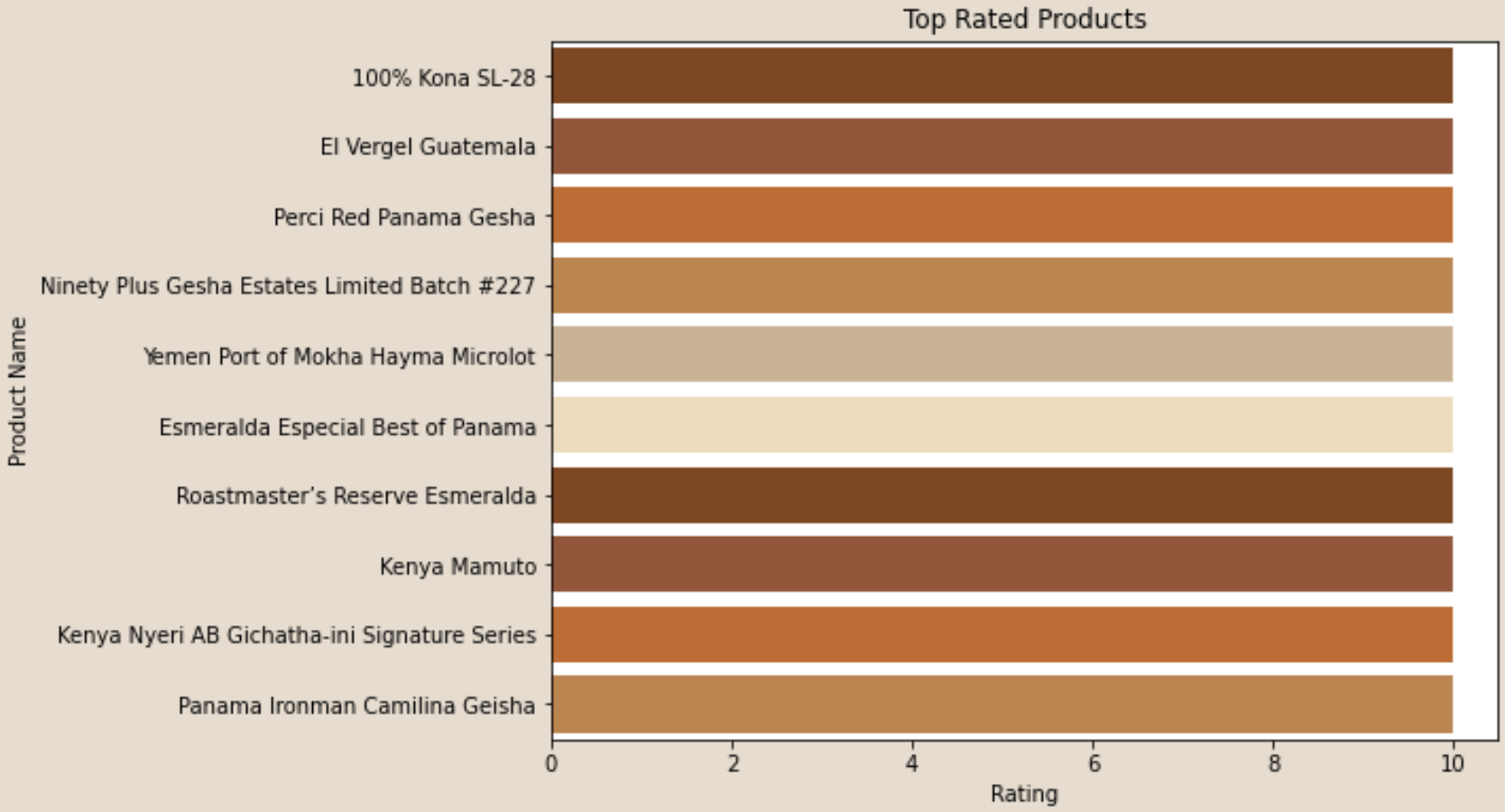
Average Rating Per Roast Type



Light roast type has the highest rating of 92.42

Medium-light roast comes next with a rating of 92.38

Top Rated Products



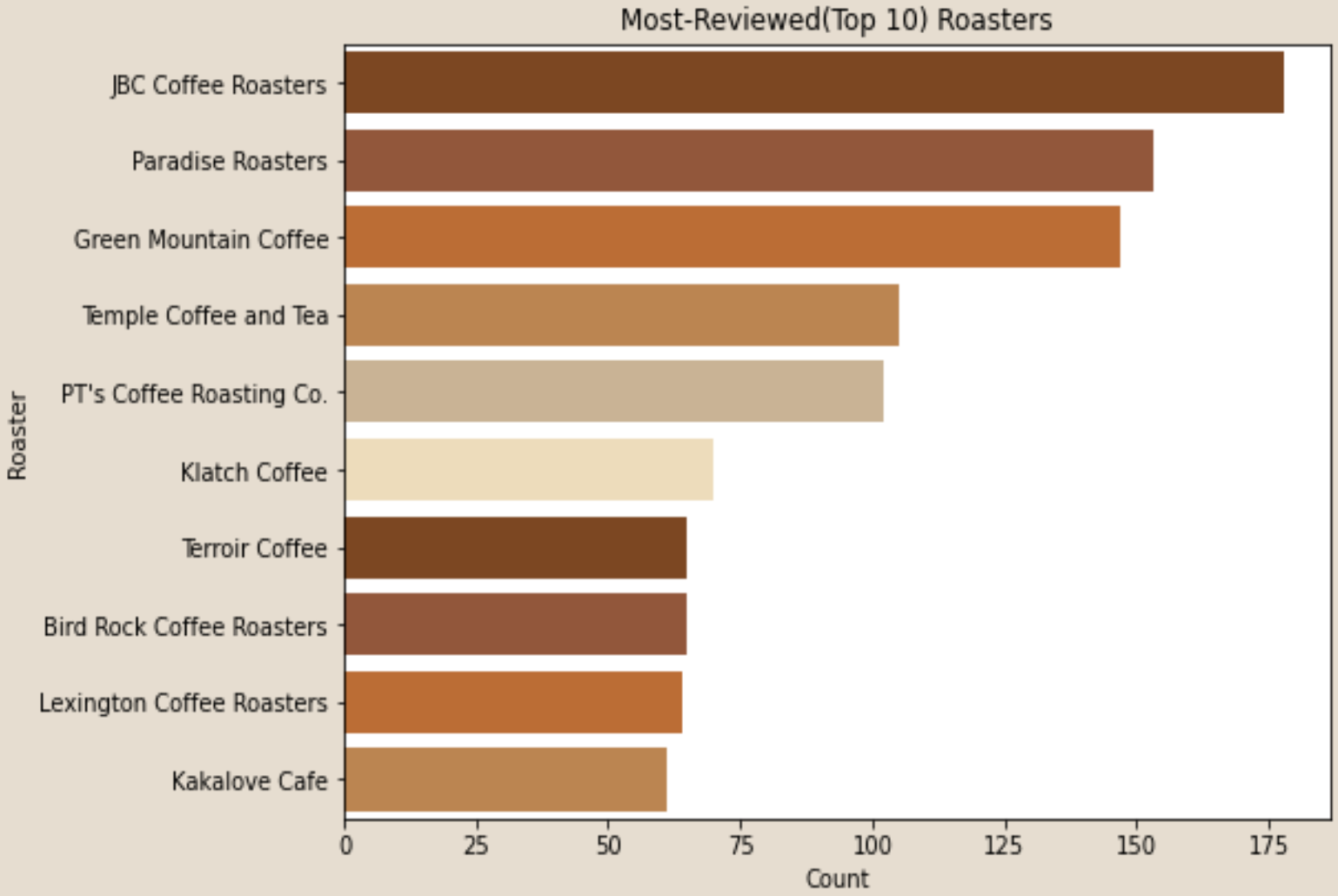
Top-rated products with a perfect rating of 10 include 100% Kona SL-28, El Vergel Guatemala, Perci Red Panama Gesha, and others, reflecting their exceptional quality.

Most Reviewed Roasters

JBC Coffee Roasters leads with the highest number of reviews (178).

Paradise Roasters got second highest reviews (153).

Green Mountain Coffee comes next with 147 reviews.

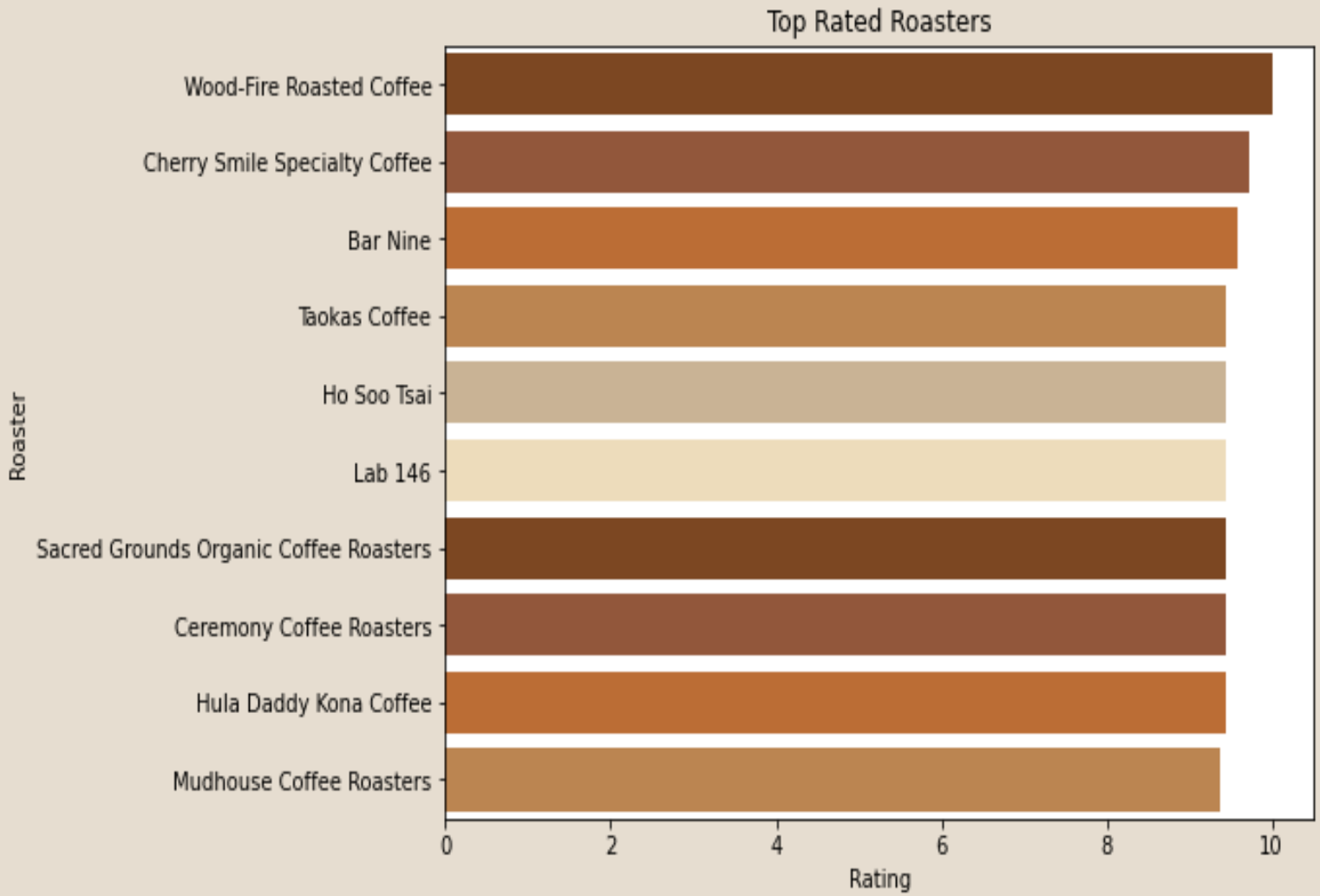


Top Rated Roasters

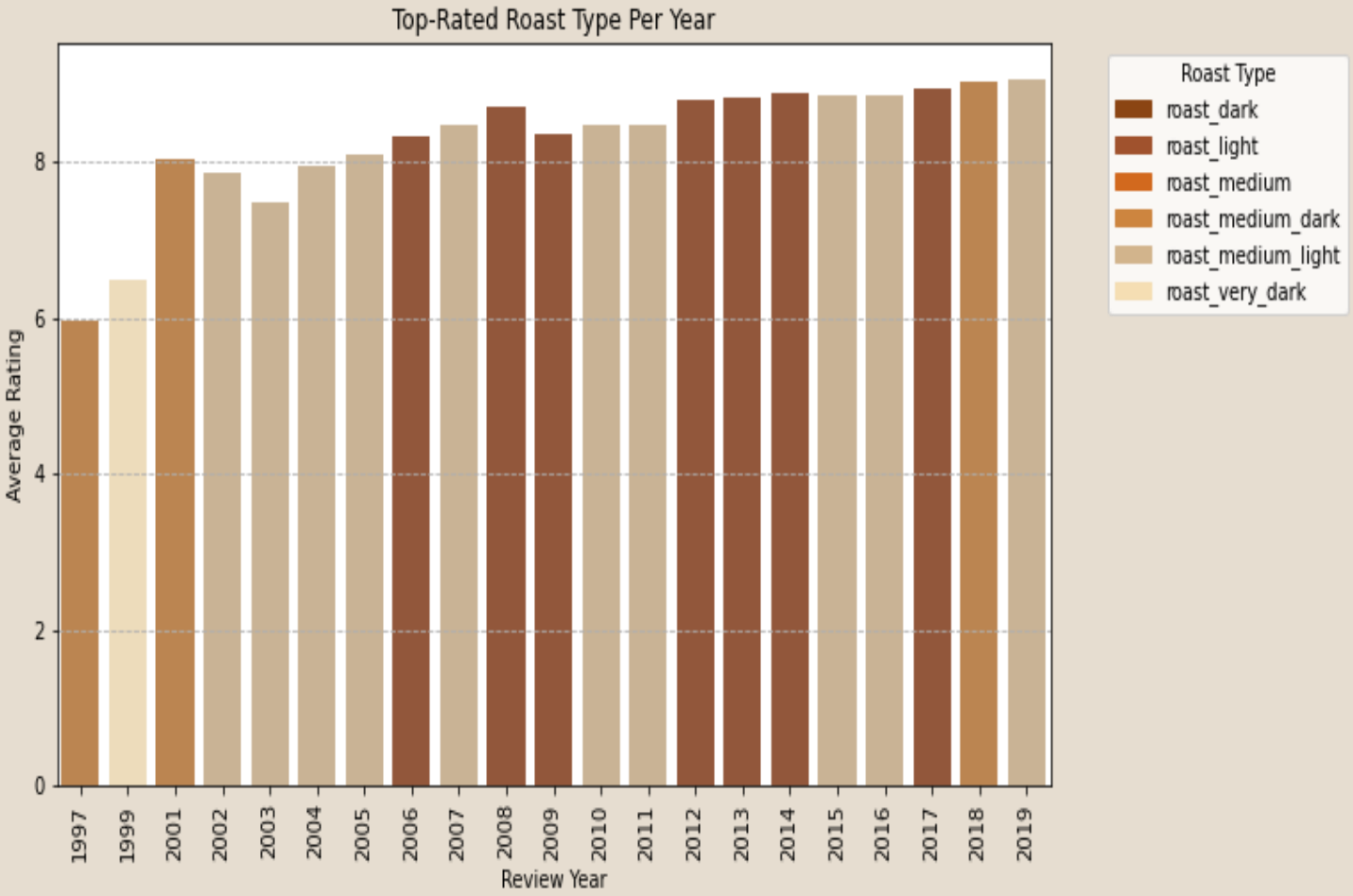
Wood-Fire Roasted Coffee achieved the highest average rating of 10.

Cherry Smile Specialty Coffee got second highest rating, 9.71.

Bar Nine comes next with 9.57 as average rating.



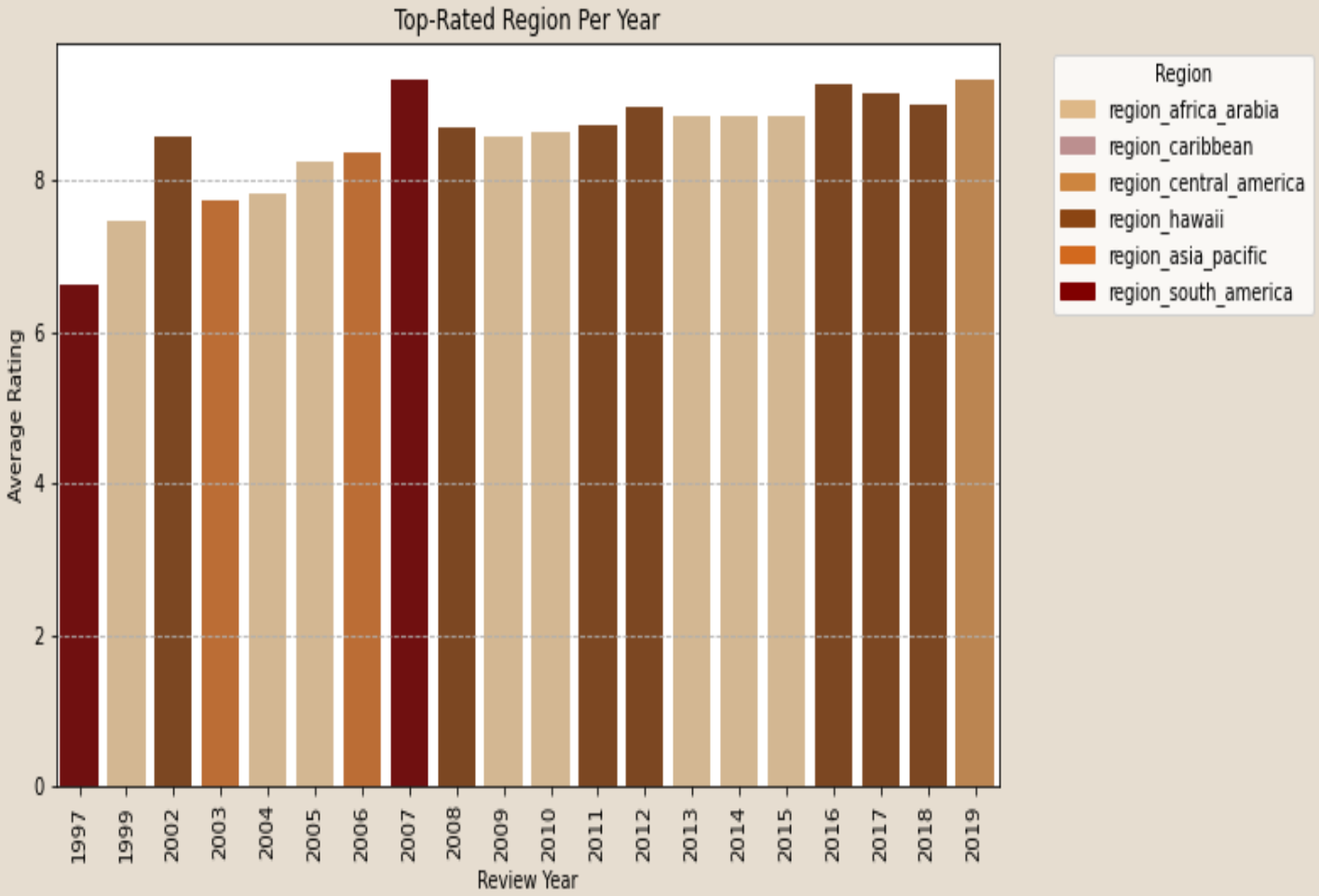
Top Rated Roast Type Per Year



Medium Light Roast has consistently received the highest ratings in most years.

Light Roast follows closely as the second most highly rated roast type over the years.

Top Rated Region Per Year



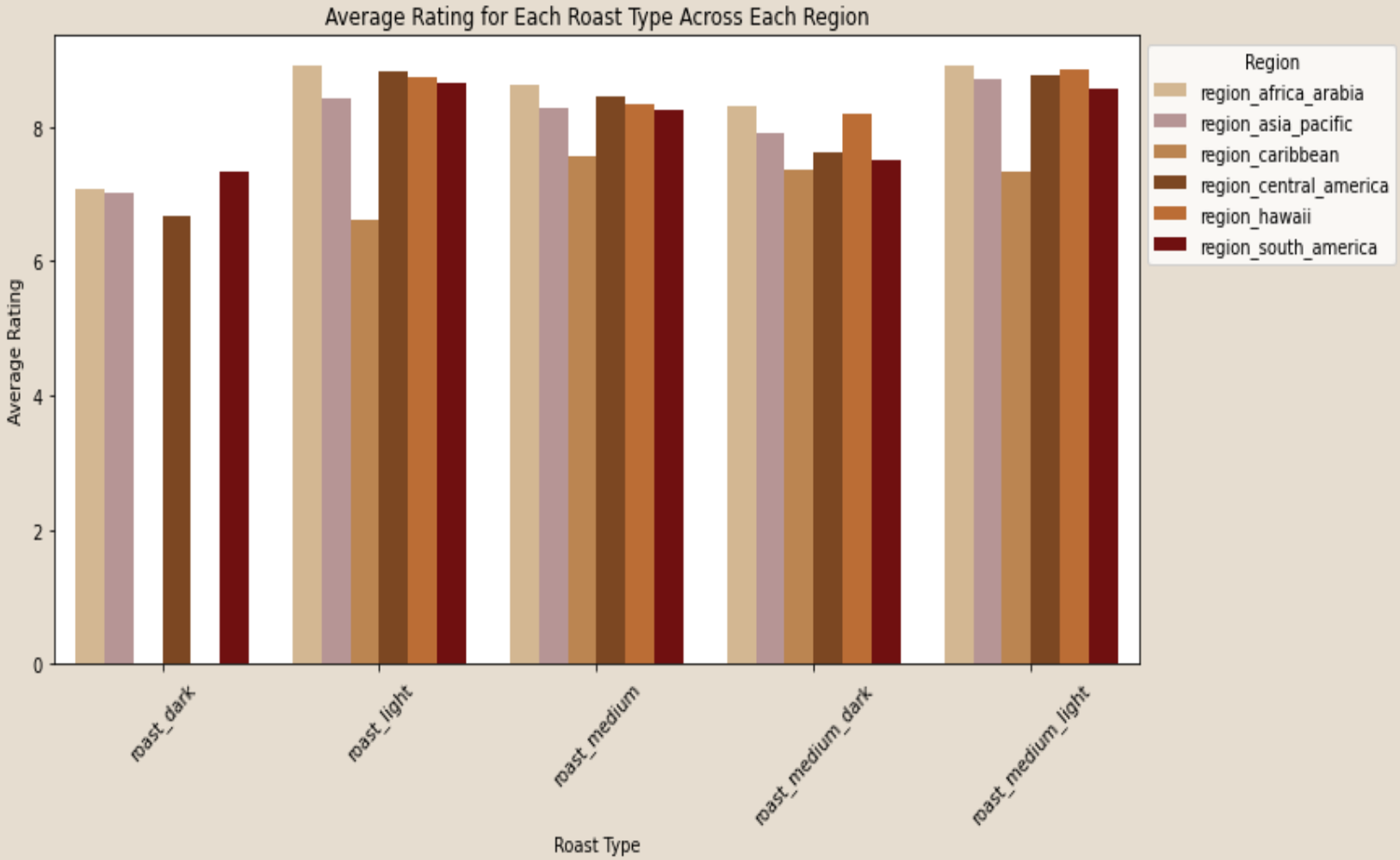
Africa-Arabia region has maintained the highest ratings in most years.

Hawaii region ranks as the second highest-rated over the years.

Average Rating For Each Roast Type Across Each Region

Africa-Arabia region: All roast type from this region has exceptionally high rating.

Caribbean region: All roast types from this region got comparatively lower rating.



Top Rated Roast Type For Each Region

Africa-Arabia: This region got light roast as the best roast type with an average rating of 8.91.

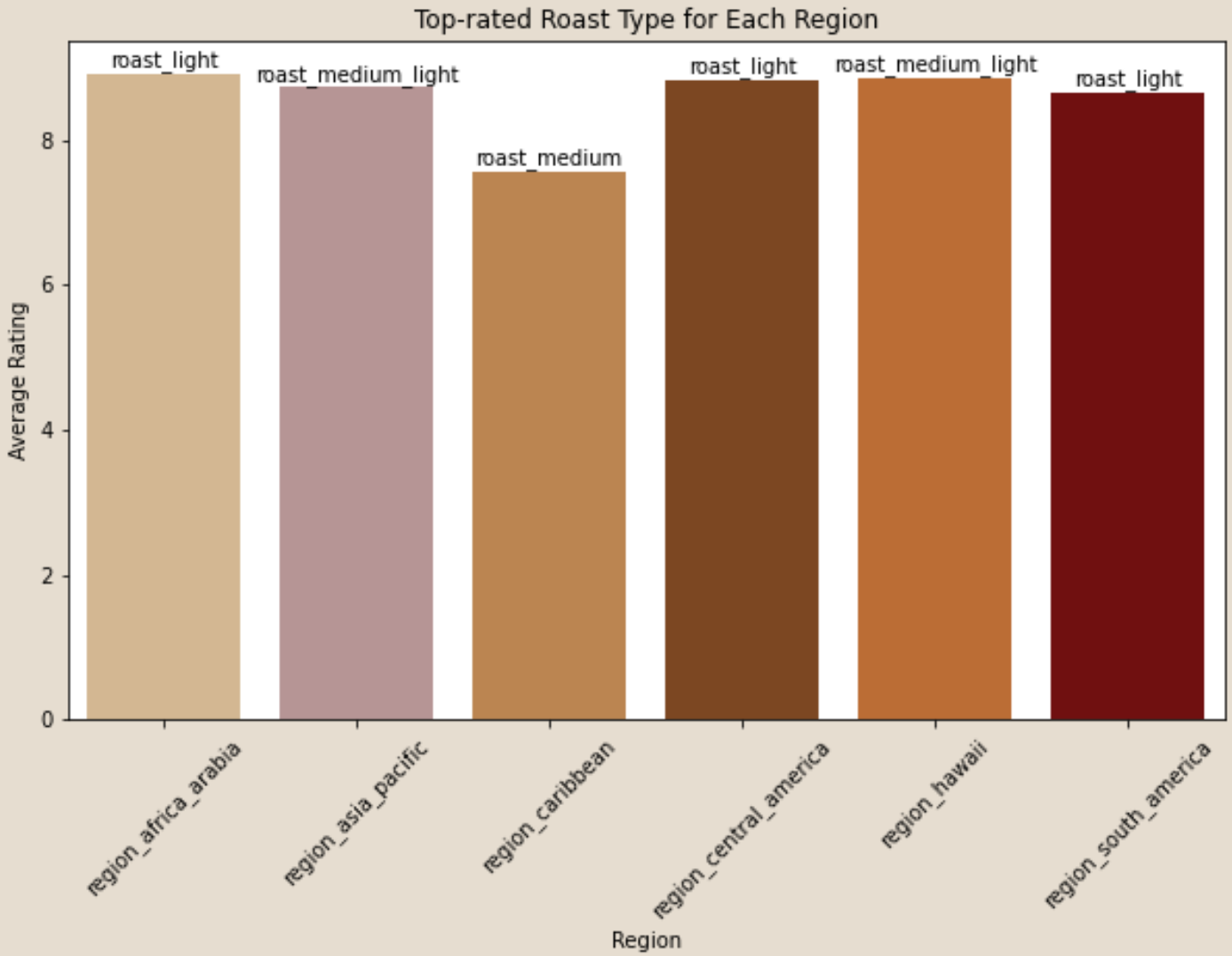
Asia-Pacific: This region got medium light as the best roast type with an average rating of 8.71.

Caribbean: Medium roast is the best roast type for this region.

Central America: Light roast is the top-rated roast with an average rating of 8.82.

Hawaii: Medium light is the best roast type**.**

South America: Light roast is the top-rated roast type for South America.



Top Rated Region For Each Roast Type

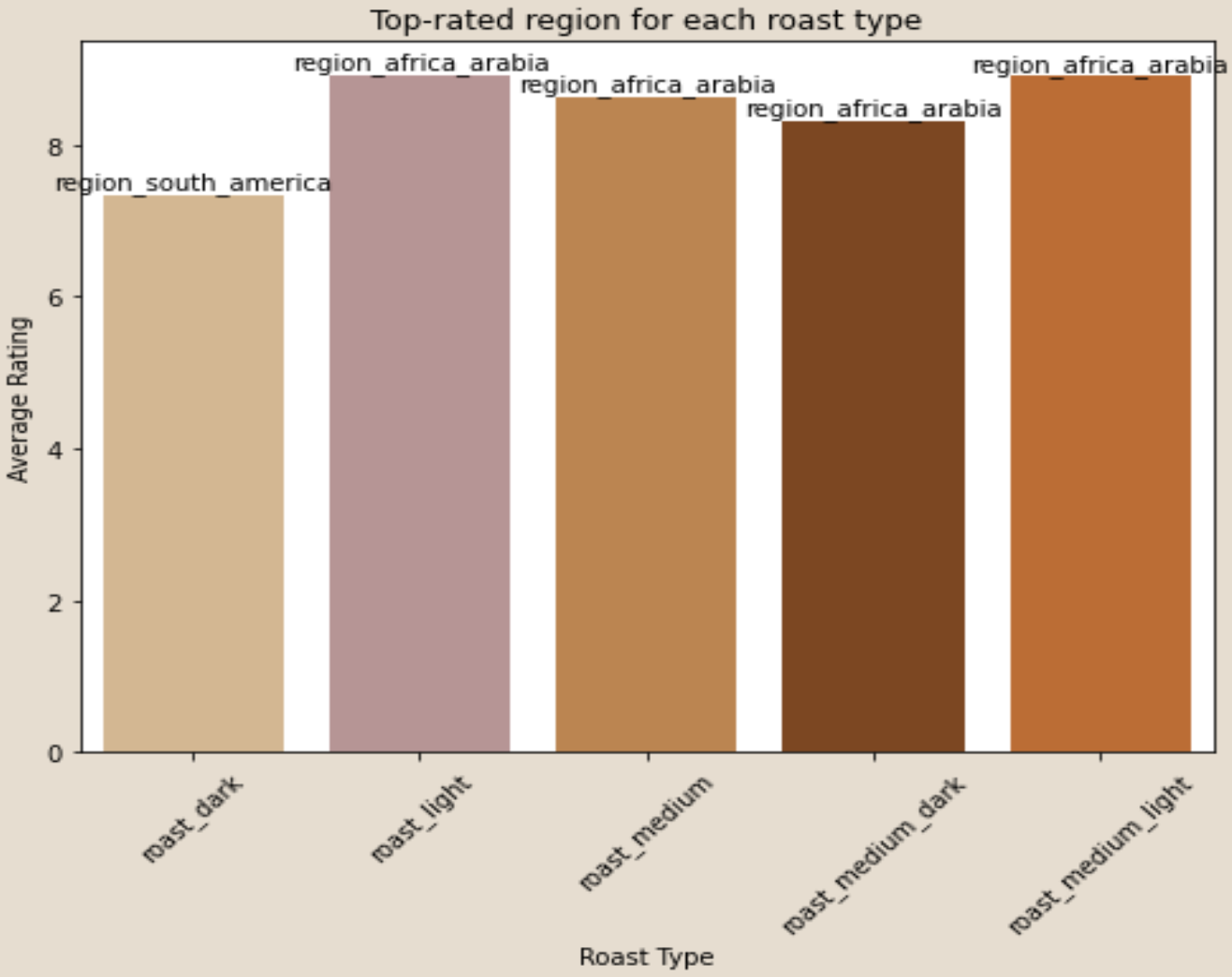
Light Roast: The top-rated is from Africa Arabia region (8.91).

Medium Roast: Africa Arabia is again the top-rated region (8.63).

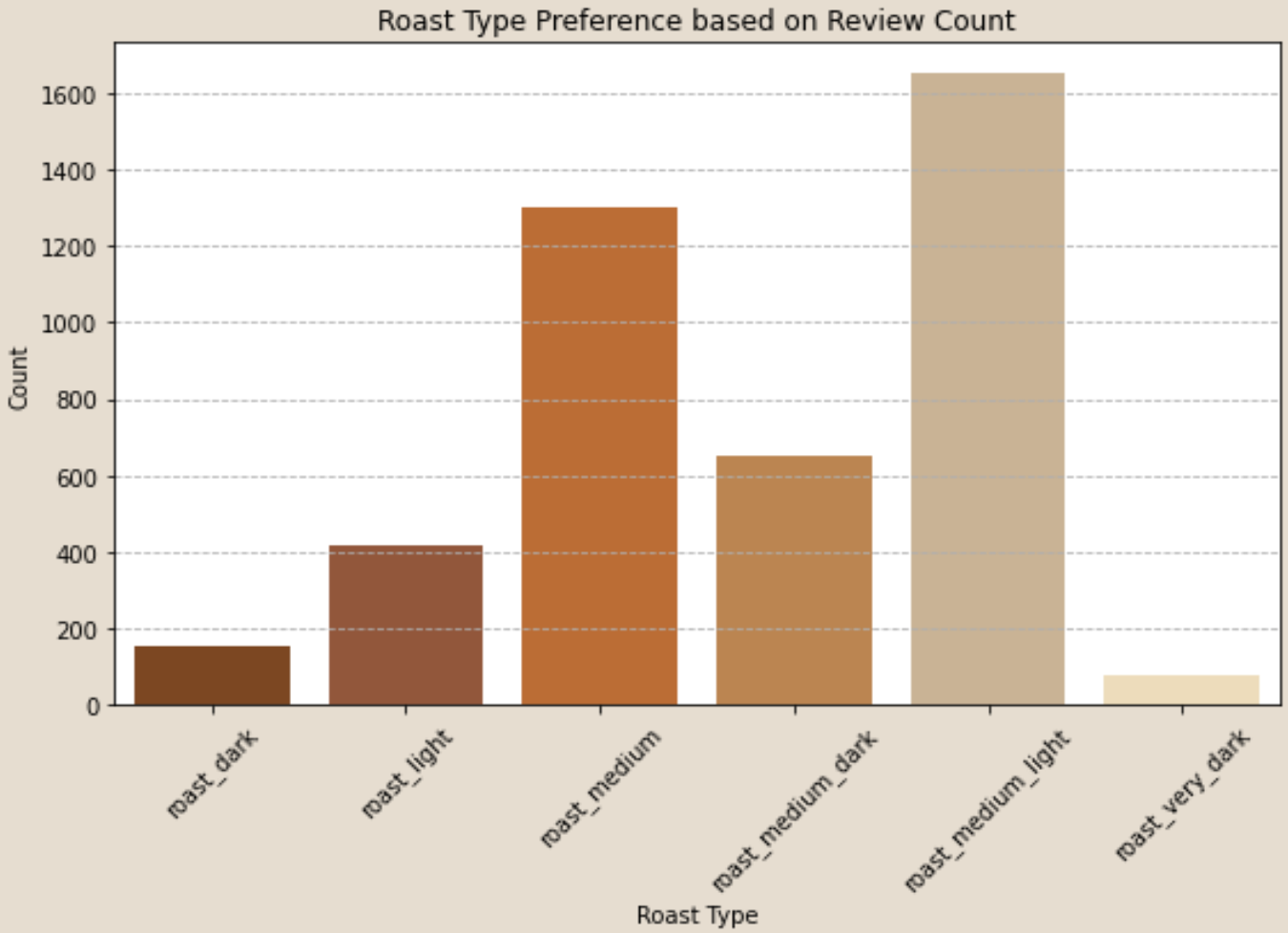
Medium-Dark Roast: Africa Arabia leads with a rating of (8.31).

Medium-Light Roast: Africa Arabia tops the list with a rating of (8.90).

Dark Roast: The top-rated is from South America region with a rating of (7.32)

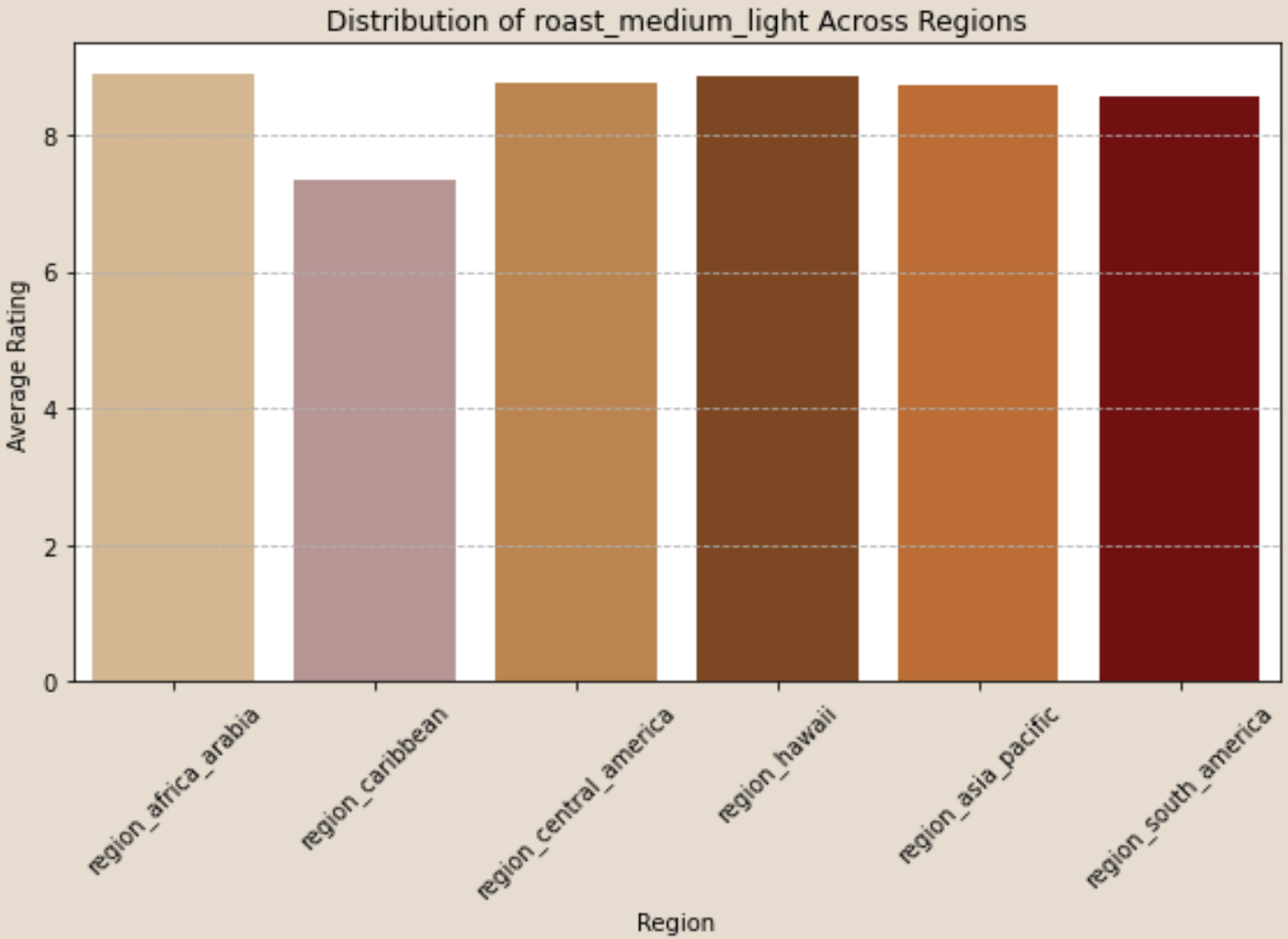


Most Popular Roast Type based on Review Count



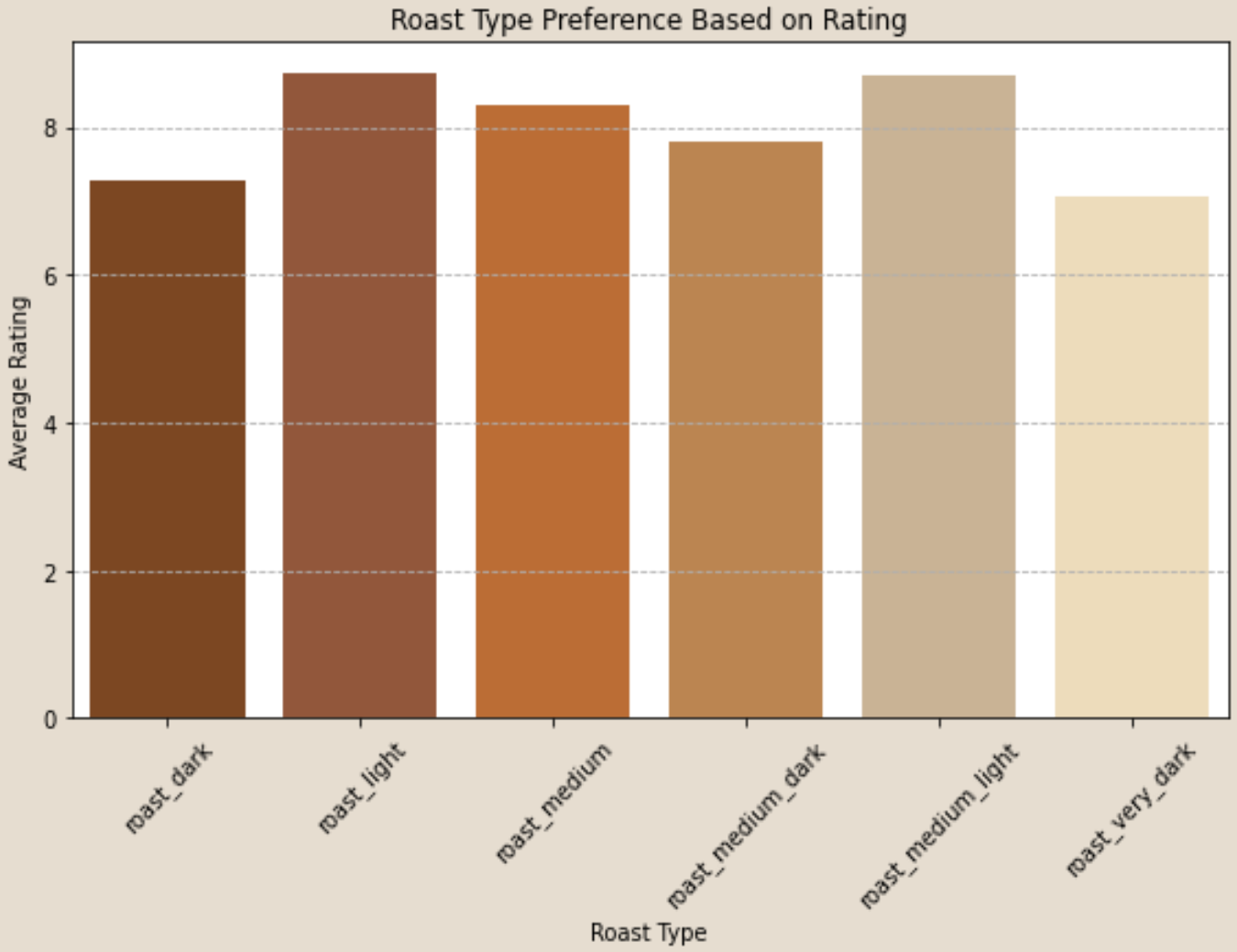
Medium Light Roast is the most popular roast type based on review count.

Distribution of Most Popular (Review Count) Roast Type



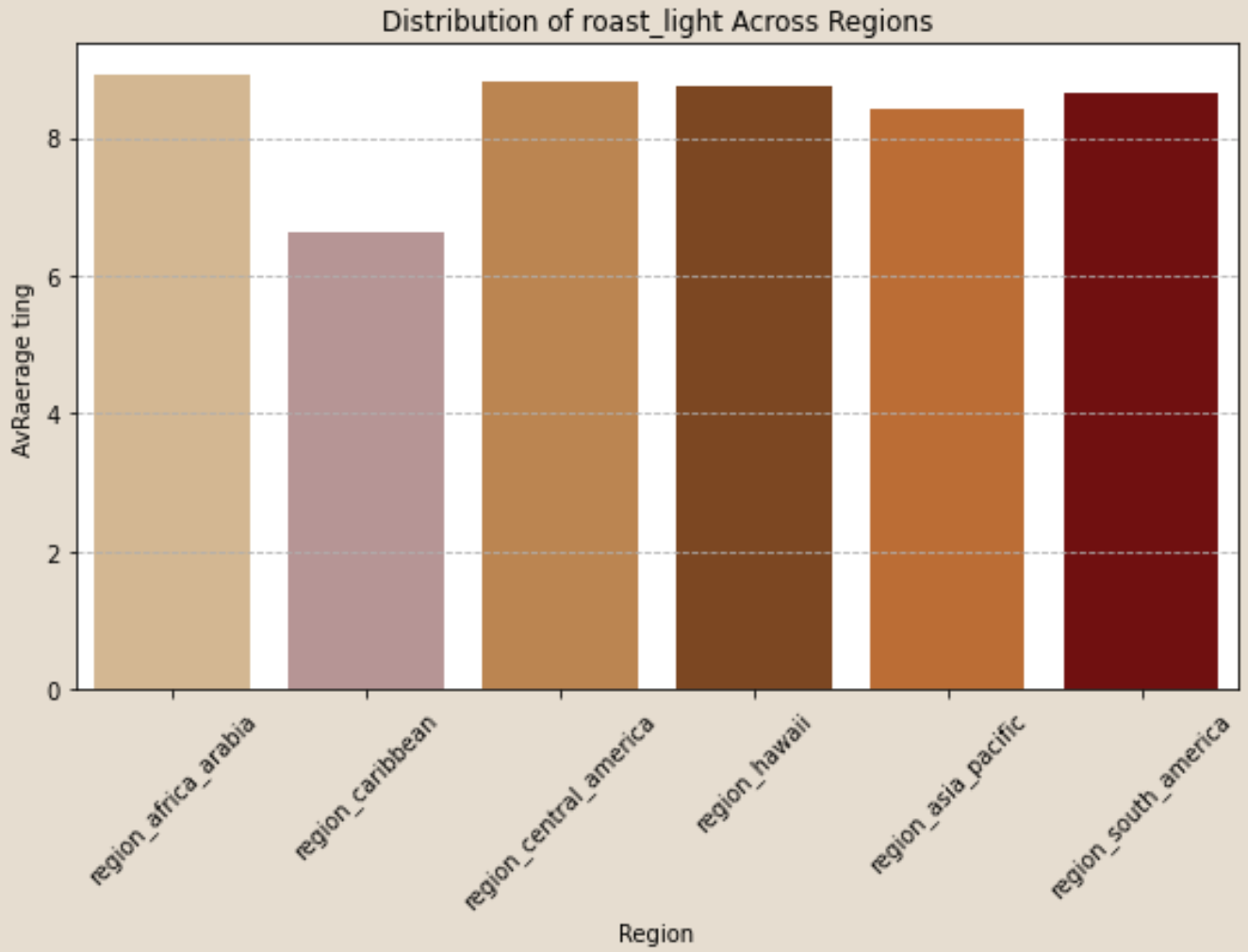
Medium Light Roast got highest average rating(8.89) from Africa-Arabia region and the lowest (7.32) from Caribbean region.

Most Popular Roast Type based on Rating



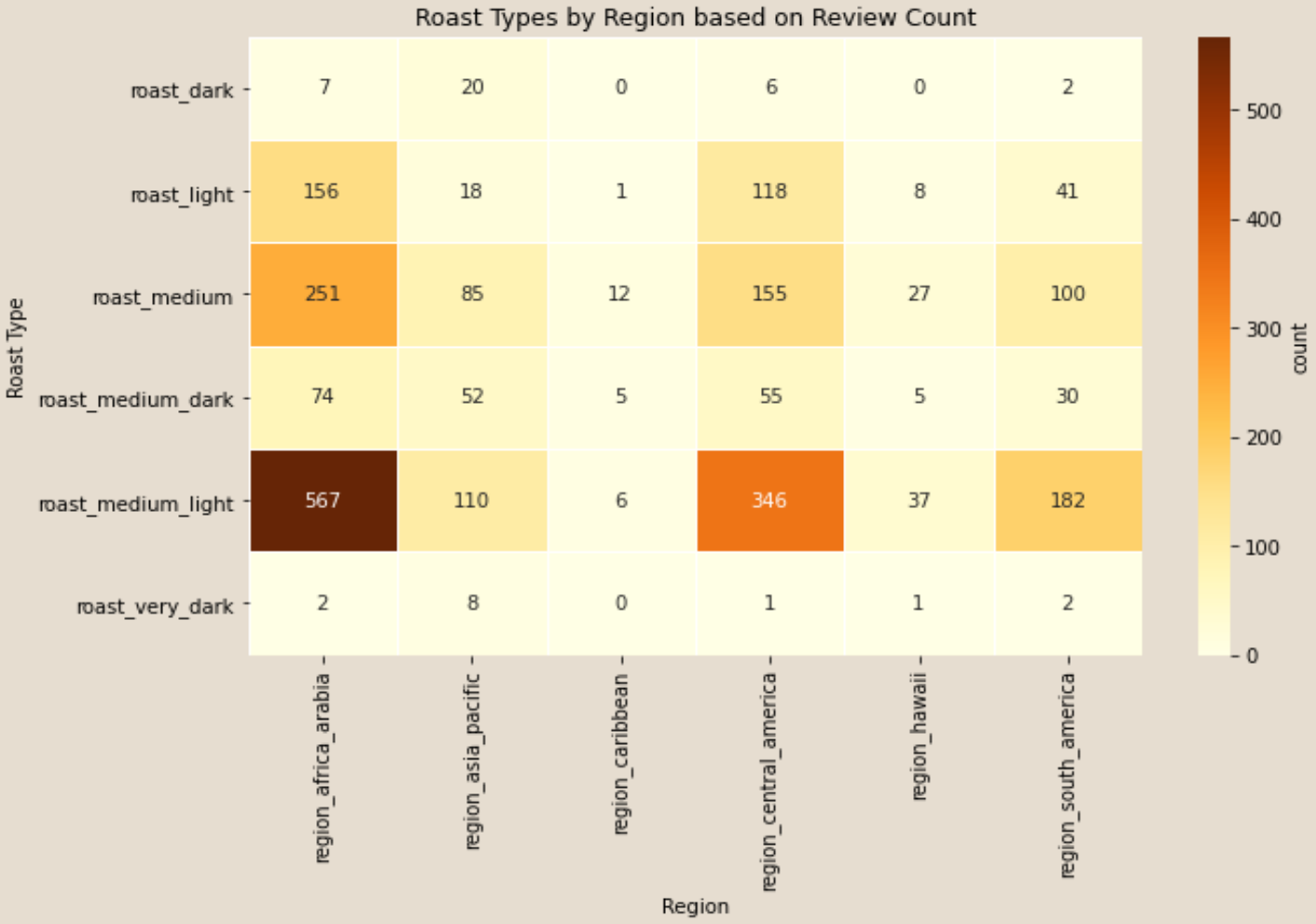
Light Roast is the most popular roast type based on rating.

Distribution of Most Popular (Rating) Roast Type



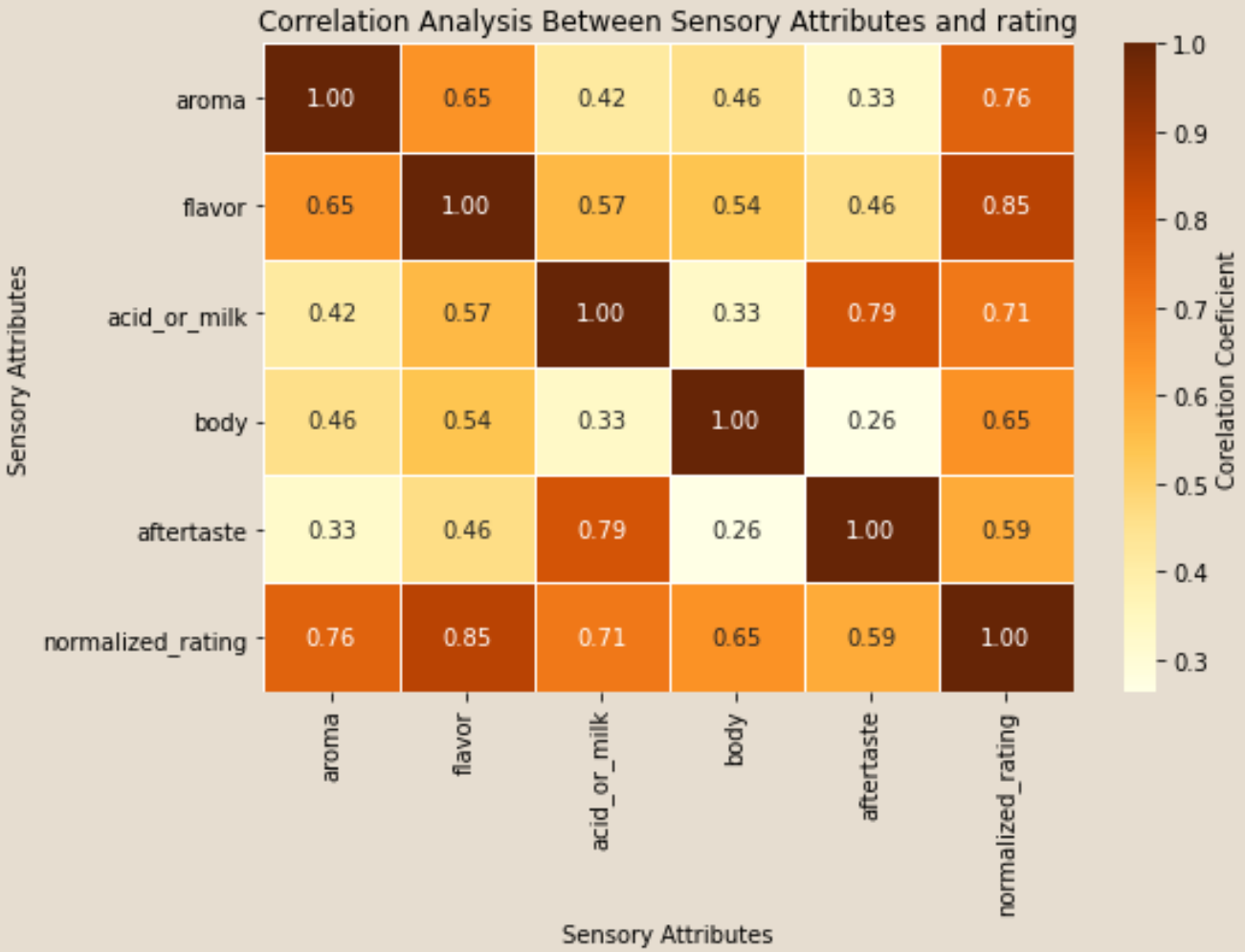
Light Roast got highest average rating(8.91) from Africa-Arabia region and the lowest (6.62) from Caribbean region.

Distribution of Roast Types by Region



Medium Light Roast from africa-arabia region got highest number of reviews, followed the by Medium Light Roast from central America.

Relationship Between Sensory Attributes And Rating



Sensory attributes showing a strong correlation with rating, exceptionally aroma and flavor.

Aroma has a correlation of 0.76, while flavor exhibits the highest correlation at 0.85.

Regression Model To Predict Ratings

Feature Set And Target

Feature Set

- Roast Types
- Regions
- Sensory attributes
- roaster

Target

- normalized_rating

Encoding Techniques

One Hot Encoder

- Roaster

- **One-hot encoding** is a method used to convert categorical data into a numerical format by creating binary columns for each category, where 1 represents the presence of a category and 0 represents its absence.

- Applied One-Hot Encoding to convert categorical 'roaster' values into binary features for machine learning model input.

Models And Parameters

Linear Regression	Kneighbors Regressor	Decision Tree Regressor	Random Forest Regressor
• No parameters	• n_neighbors = 5	• max_depth = 5 • min_samples_split = 10 • min_samples_leaf = 2 • random_state = 42	• n_estimators=100 • random_state=42

Regression Model Evaluation

Linear Regression

- Mean squared error: 0.077
- R2: 0.879

Random Forest

- Mean squared error: 0.082
- R2: 0.871

Decision Tree

- Mean squared error: 0.113
- R2: 0.822

Knn

- Mean squared error: 0.232
- R2: 0.637

❖ **Linear Regression:** lowest MSE (0.077) and highest R2 (0.879) indicating the best predictive accuracy.

❖ **Random Forest Regressor:** shows strong performance with a high R2 (0.871) and relatively low MSE (0.082)

❖ **Decision Tree Regressor:** Moderate performance with an R2 of 0.822 and an MSE of 0.113

❖ **Kneighbors Regressor:** The lowest performance with an R2 of 0.637 and a an MSE of 0.232.

Classification Model to Categorize Products into Popularity Tiers

Feature Set And Target

Feature Set

- Roast Types
- Regions
- Sensory attributes
- roaster

Target

- popularity_tier

Techniques For Class Balancing In Classification Model

❖**Data Stratification:**

- Ensured balanced class representation across the popularity tiers (Highly Popular, Moderately Popular, Less Popular) by setting stratify=Y during data splitting.

❖**Addressing Class Imbalance:**

- Applied SMOTE (Synthetic Minority Oversampling Technique) to oversample minority classes, improving model performance.

❖**Implementation**:

- Used SMOTE(random_state=42) for synthetic data generation.
- Resampled training data: X_resampled and Y_resampled.

Encoding Techniques and Model Parameters

❖**One-Hot Encoding:**

- Applied One-Hot Encoding to convert categorical 'roaster' values into binary features for machine learning model input.

❖**Decision Tree Classifier** :

- Employed decision tree classifier to categorize coffee products into ‘Highly Popular’ ,’Moderately Popular’, and ‘Less Popular’ tiers based on ratings.

- **Decision Tree Classifier**

•max_depth=10

•min_samples_split=10

•min_samples_leaf=2

•random_state=42

•class_weight='balanced'

Classification Model Evaluation

Precision

- The percentage of correctly predicted positive cases out of all predicted positive cases.
- Focuses on minimizing false positives.

Recall

- The percentage of correctly predicted positive cases out of all actual positive cases.
- Focuses on minimizing false negatives.

F1 Score

- The harmonic mean of precision and recall.

Classification Model Evaluation (Cont.)

Classification Report

Class	Precision	Recall	F1-Score	Support
Highly Popular	0.96	0.94	0.95	665
Less	0.71	0.83	0.77	6
Moderately	0.79	0.87	0.83	180
Accuracy	0.92			
Macro avg	0.82	0.88	0.85	851
Weighted avg	0.93	0.92	0.92851	851

❖**Highly Popular :** Excellent performance with high precision (0.96), recall (0.94) and F1 score (0.95)

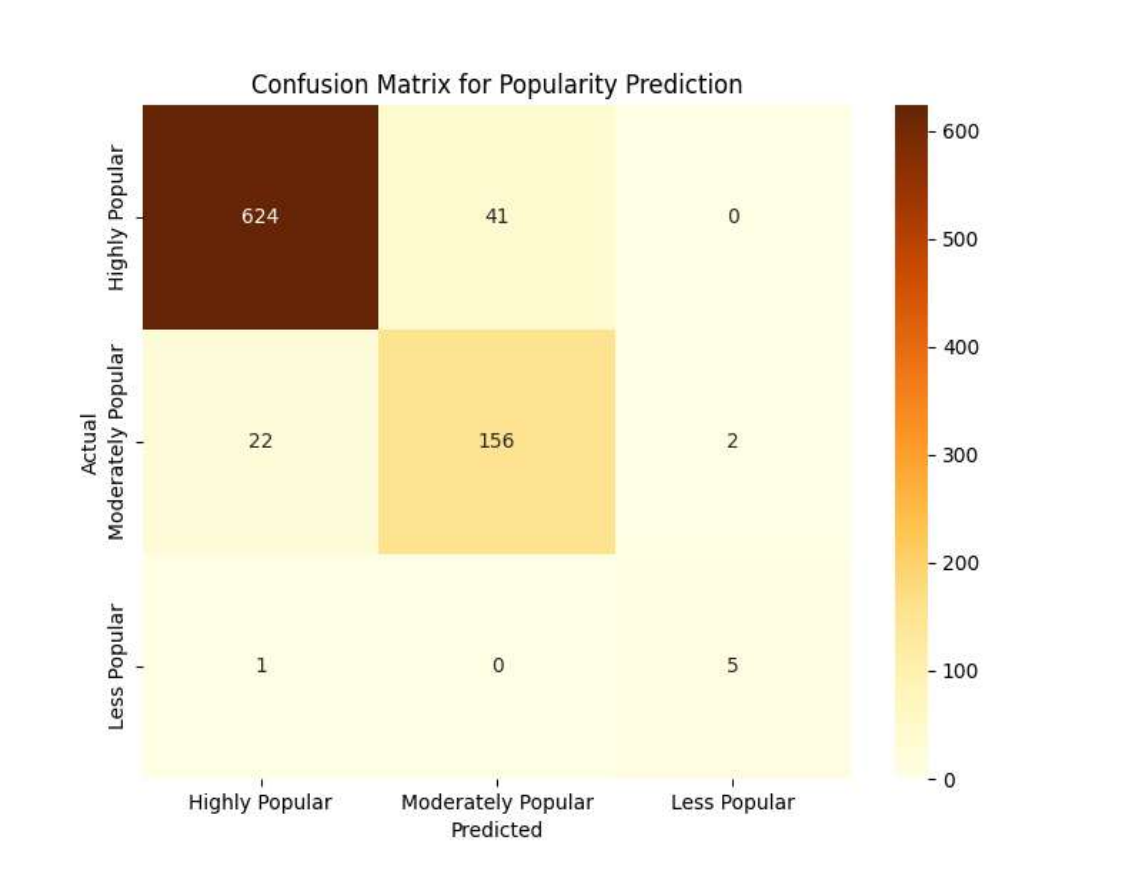
❖**Moderately Popular :** Strong recall (0.87), good precision (0.79) and F1 score (0.83)

❖**Less Popular :** Moderate precision (0.71),and and F1 score (0.77), but good recall (0.83)

❖**Model Summary**: High accuracy (92%) and strong performance (ROC-AUC: 0.94)

Classification Model Evaluation (Cont.)

Confusion Matrix

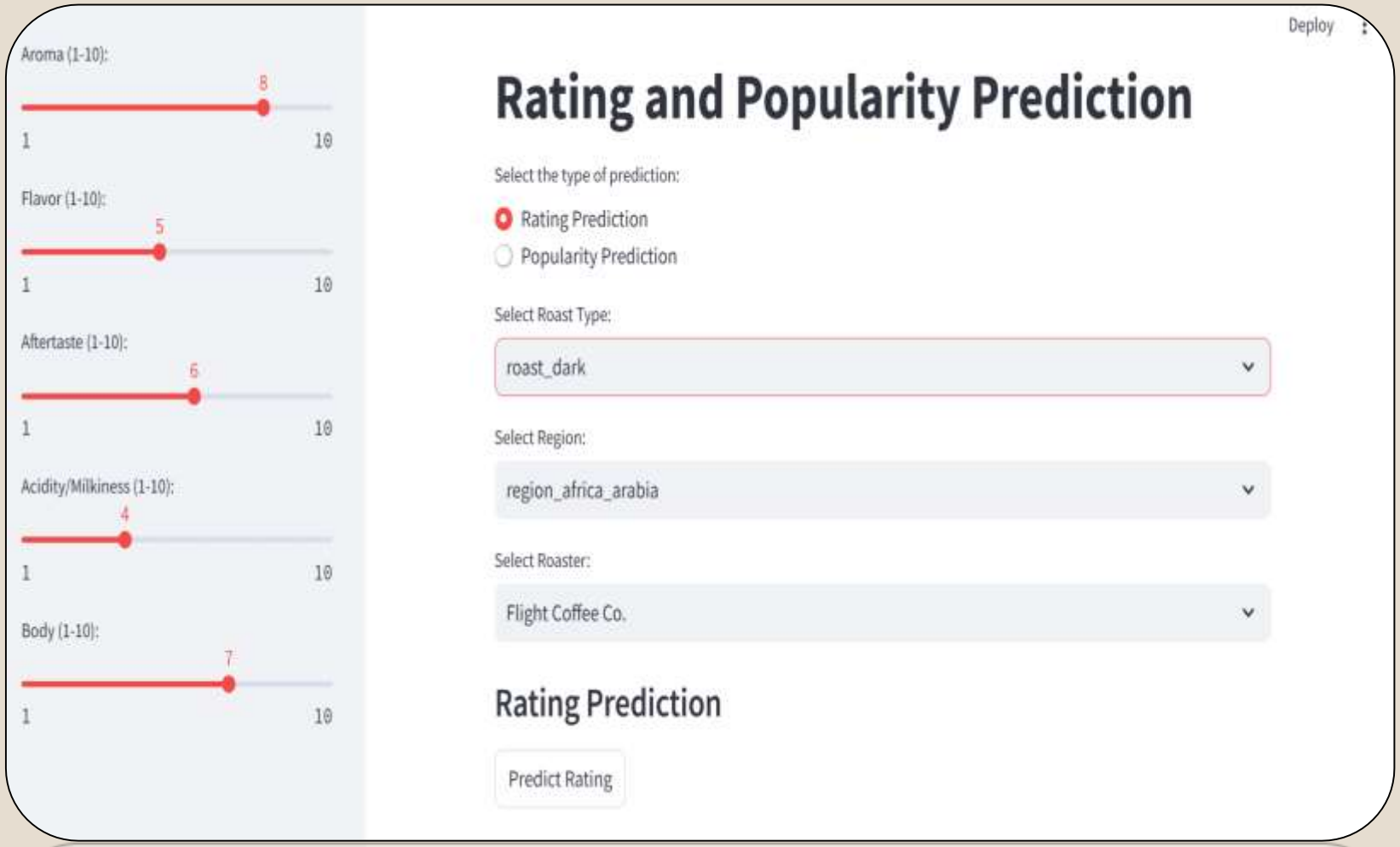


- **Highly Popular** : Most instances correctly predicted (624), minor misclassifications as Medium (41).
- **Moderately Popular** : Predicted well (156), small misclassifications into High (22) and Low (2).
- **Less Popular** : Correctly predicted 5 instances, with negligible misclassifications.
- **Summary**: Correctly predicted most instances with some minor misclassifications.

Model Deployment

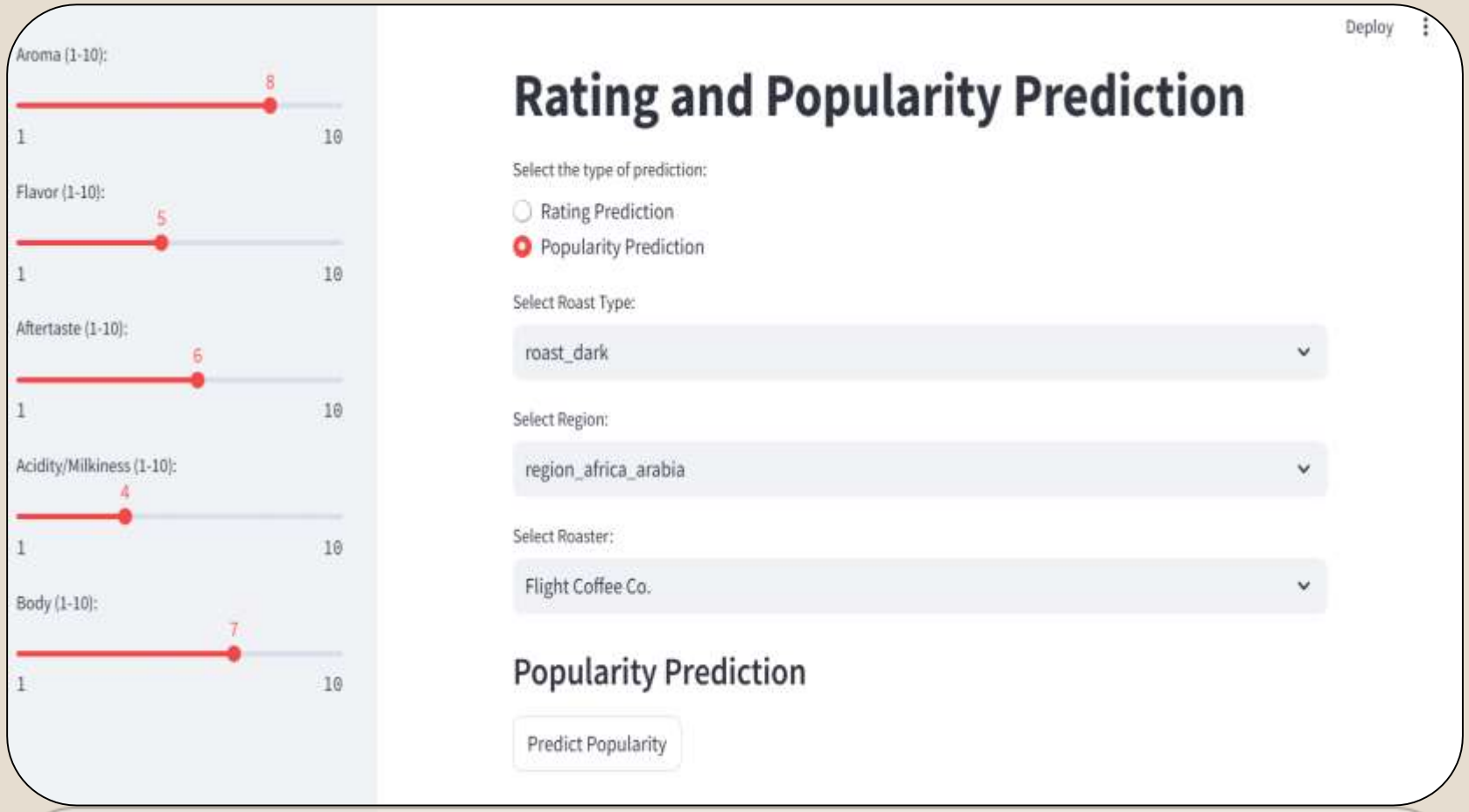
Regression Model To Predict Rating

Deployed the best model (Linear regression) for real-time predictions using Streamlit



Classification Model to Categorize Products into Popularity Tier

Decision tree classifier is deployed for real-time predictions using Streamlit



Key Insights And Recommendations

Key Insights

- Top-rated products include 100% Kona SL-28, El Vergel Guatemala, and Perci Red Panama Gesha.
- Attributes such as high-quality sourcing, unique flavor profiles, and exclusive origins drive high ratings.
- Light and medium-light are the most highly rated roast types, indicating strong consumer preference.
- Coffee from Africa-Arabia and Hawaii regions consistently holds the top spot for highest ratings.

Product Recommendations

❖**Replicate success attributes:**

- Focus on replicating the characteristics of top-rated products, such as premium sourcing, unique flavor profiles, and origins.

❖**Prioritize medium-light and light roasts**:

- Prioritize medium-light and light roasts to align with consumer preferences and capitalize on their popularity.

❖**Explore blends inspired by top-rated regions**:

- Develop special blends using beans from high-rated regions like Africa-Arabia and Hawaii to offer unique experiences that reflect the top-rated roast types.

❖**Introduce limited-edition products:**

- Offer exclusive, limited-edition coffee releases with medium –light and light roast options, maintaining a sense of rarity and desirability that attracts niche coffee enthusiasts.

Regional Focus

- **Focus on Africa-Arabia Region:**
  - Since the Africa-Arabia region consistently receives the highest ratings, coffee businesses should prioritize sourcing and promoting beans from this region to appeal to quality-conscious consumers.
- **Highlight Hawaiian Coffee:**
  - The Hawaii region follows closely in ratings, indicating a strong preference for coffee from this origin. Marketers can position Hawaiian coffee as a premium alternative to Africa-Arabia varieties.
- **Promote Regional Blends**:
  - Create unique blends using beans from Africa-Arabia and Hawaii region to offer a curated selection that showcases the strengths of each region.
  - Regional blends can provide consumers with a taste experience that combines the best attributes of each origin, providing both variety and exclusivity.

Marketing Campaigns

❖**Target Popular Roast Types:**

- Focus on promoting light and medium-light roasts, which consistently receive higher ratings.

❖**Target Key Regions:**

- Considering the strong ratings for coffee from Africa-Arabia and Hawaii origins, specialty coffee brands should focus on premium markets that prioritize high-quality beans from these regions.

❖**Leverage storytelling:**

- Highlight the unique origin, cultivation process, and roast profile of coffees in marketing campaigns to emphasize their exclusivity and appeal.

❖**Leverage seasonal offerings**:

- Introduce seasonal or limited-edition releases based on regional harvests, offering customers something unique at specific times of the year, such as holiday blends or harvest-specialty coffees.

Thank You !