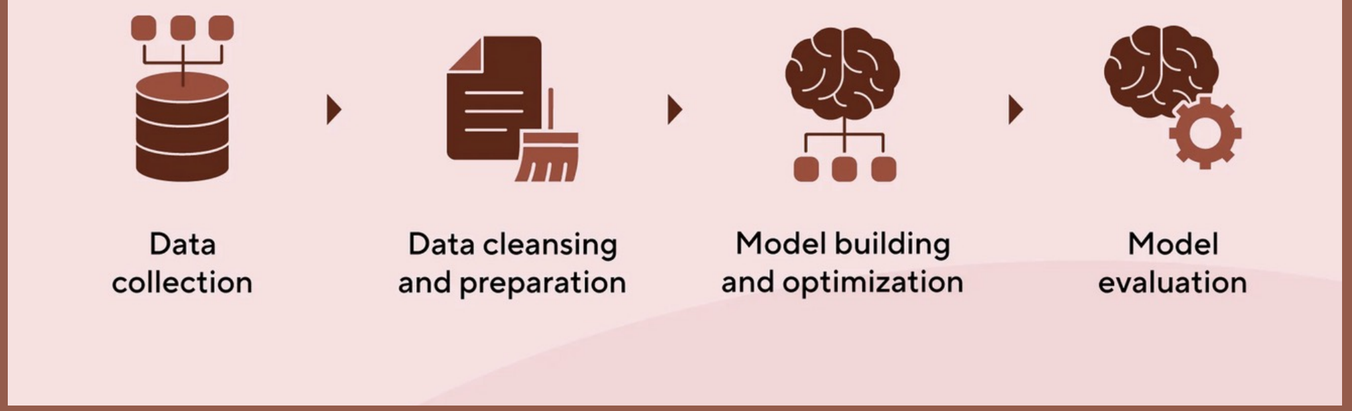
CHOCOLATE CLASSIFICATION AND QUALITY PREDICITION

Using Machine Learning Techniques

Presented By : Anshu Reddy Ashanna
(G38094812)

PROBLEM STATEMENT

- Classify and assess the quality of chocolate products based on various features such as ingredients, cocoa composition, manufacturer, and consumer ratings.
- Build a predictive system to categorize chocolate products.
- Identify factors that contribute to high-quality chocolates.


Data collection
Data cleansing and preparation
Model building and optimization
Model evaluation

DATA SET

Size: 2,789 distinct chocolate entries

Key Attributes:

- Manufacturing Company
- Company Location
- Review Date
- Country of Bean Origin
- Specific Bean Origin
- Cocoa Percentage
- Ingredient list
- Most Memorable Characteristics (tasting notes)
- Consumer Rating

```python
#Importing DataSet
df = pd.read_csv('/content/chocolate_data.csv')
```

```python
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2789 entries, 0 to 2788
Data columns (total 10 columns):
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   Reference                       2789 non-null   float64
 1   Manufacturing Company           2789 non-null   object 
 2   Company Location                2789 non-null   object 
 3   Review Date                     2789 non-null   float64
 4   Country of Bean Origin          2789 non-null   object 
 5   Specific Bean Origin            2789 non-null   object 
 6   Cocoa Percent                   2789 non-null   object 
 7   Ingredients                     2702 non-null   object 
 8   Most Memorable Characteristics  2789 non-null   object 
 9   Rating                          2789 non-null   float64
dtypes: float64(3), object(7)
memory usage: 218.0+ KB
```

```python
df.head()
```

	Reference	Manufacturing Company	Company Location	Review Date	Country of Bean Origin	Specific Bean Origin	Cocoa Percent	Ingredients	Most Memorable Characteristics	Rating
0	2454.0	5150	U.S.A.	2019.0	Madagascar	Bejofo Estate, batch 1	76%	3- B,S,C	cocoa, blackberry, full body	3.75
1	2458.0	5150	U.S.A.	2019.0	Dominican\nRepublic	Zorzal, batch 1	76%	3- B,S,C	cocoa, vegetal, savory	3.50
2	2454.0	5150	U.S.A.	2019.0	Tanzania	Kokoa Kamili, batch 1	76%	3- B,S,C	rich cocoa, fatty, bready	3.25
3	2542.0	5150	U.S.A.	2021.0	India	Anamalai, batch 1	68%	3- B,S,C	milk brownie,\nmacadamia,chewy	3.50
4	2546.0	5150	U.S.A.	2021.0	Uganda	Semuliki Forest, batch 1	80%	3- B,S,C	mildly bitter, basic cocoa, fatty	3.25

DATA PREPROCESSING

- Check and handle missing values: For missing values in Ingredients, compute the mode (most frequent ingredient string) and replace with the mode.
- Drop Unnecessary Columns: Remove fields like Reference that are no longer useful.
- Review Date Cleanup: Convert Review Date from float (e.g., 2018.0) to integer year
- Cocoa Percent Parsing: Remove “%” and cast to float for numerical analysis.

```python
#Checking for null values
print(df.isnull().sum())
```

```
Reference                          0
Manufacturing Company              0
Company Location                   0
Review Date                        0
Country of Bean Origin             0
Specific Bean Origin               0
Cocoa Percent                      0
Ingredients                       87
Most Memorable Characteristics     0
Rating                             0
dtype: int64
```

```python
# Handling missing Values using Mode
df_cleaned = df.copy()
most_common_ingredient = df['Ingredients'].mode()[0]

# Fill missing values with that value
df_cleaned['Ingredients'] = df['Ingredients'].fillna(most_common_ingredient)
```

```python
print(df_cleaned.isnull().sum())
```

```
Reference                         0
Manufacturing Company             0
Company Location                  0
Review Date                       0
Country of Bean Origin            0
Specific Bean Origin              0
Cocoa Percent                     0
Ingredients                       0
Most Memorable Characteristics    0
Rating                            0
dtype: int64
```

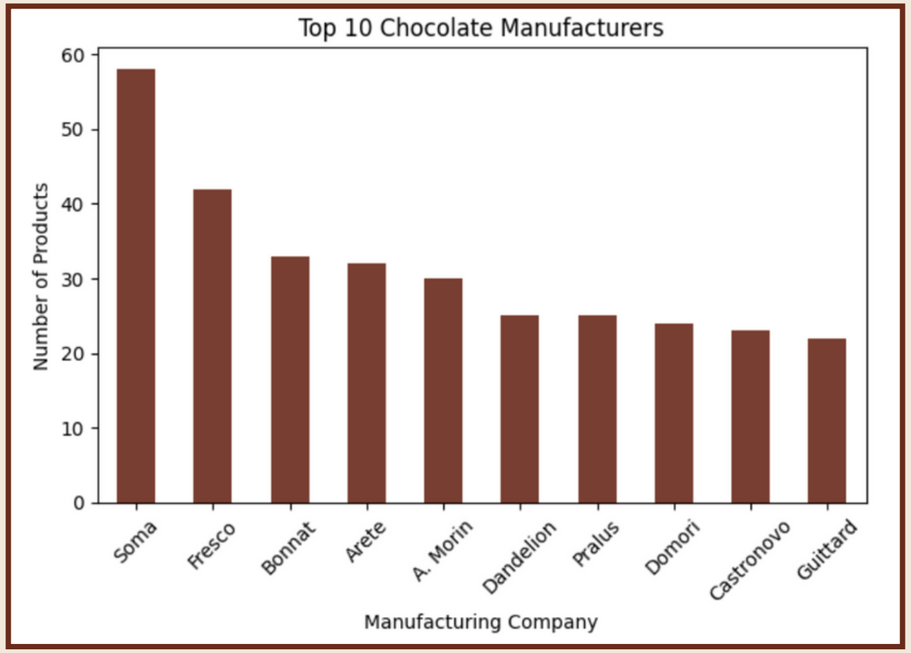
```python
# Removing Unnecessary Columns
df_cleaned = df_cleaned.drop('Reference', axis=1)

# Converting Float to Integer
df_cleaned['Review Date'] = df['Review Date'].fillna(0).astype(int)

#Remove the '%' symbol and convert to float
df_cleaned["Cocoa Percent"] = df['Cocoa Percent'].str.strip('%').astype(float)
```

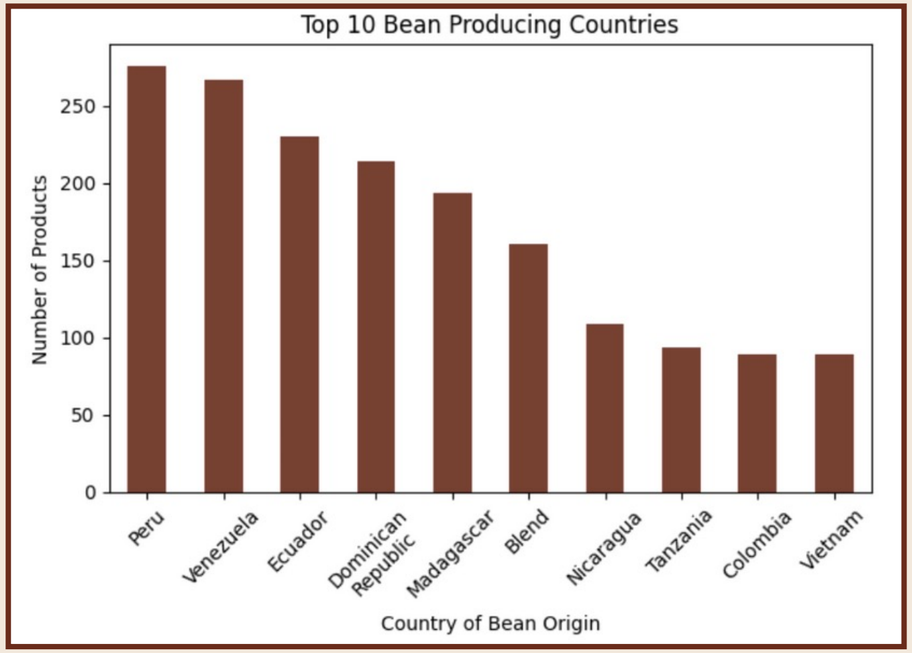
```python
df_cleaned.info()
```

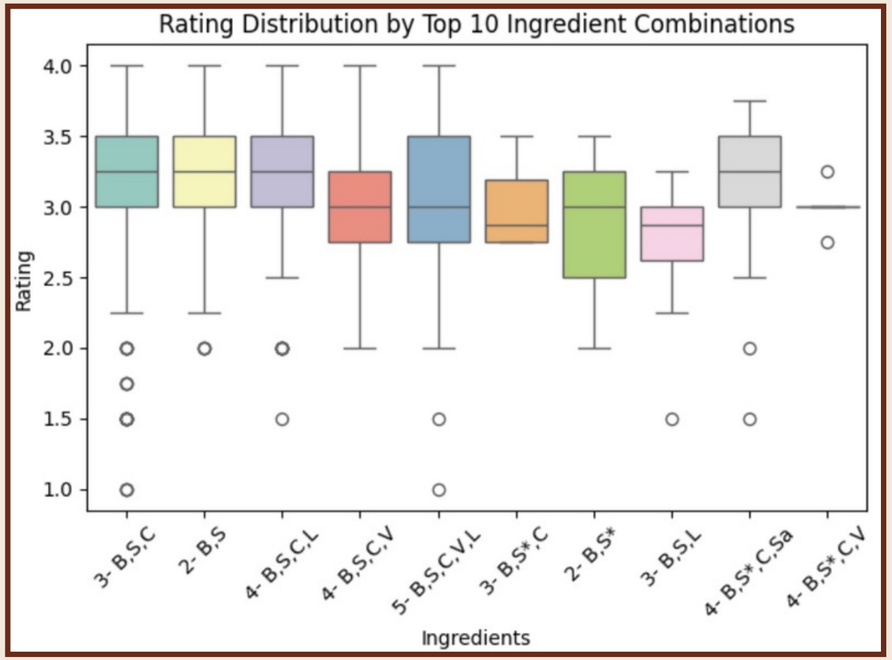
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2789 entries, 0 to 2788
Data columns (total 9 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Manufacturing Company           2789 non-null   object
 1   Company Location                2789 non-null   object
 2   Review Date                     2789 non-null   int64
 3   Country of Bean Origin          2789 non-null   object
 4   Specific Bean Origin            2789 non-null   object
 5   Cocoa Percent                   2789 non-null   float64
 6   Ingredients                     2789 non-null   object
 7   Most Memorable Characteristics  2789 non-null   object
 8   Rating                          2789 non-null   float64
dtypes: float64(2), int64(1), object(6)
memory usage: 196.2+ KB
```

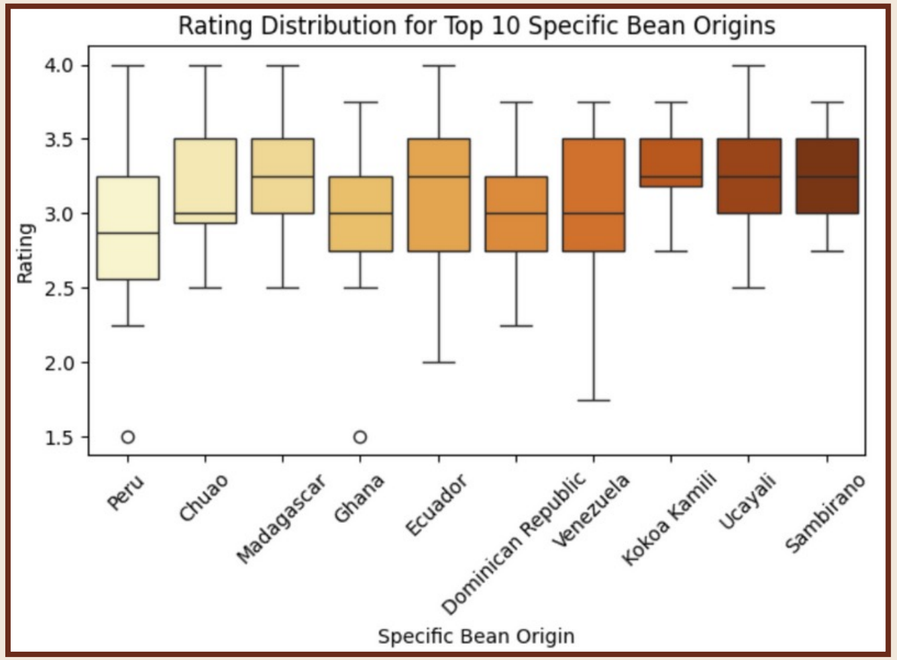
DATA ANALYSIS.....

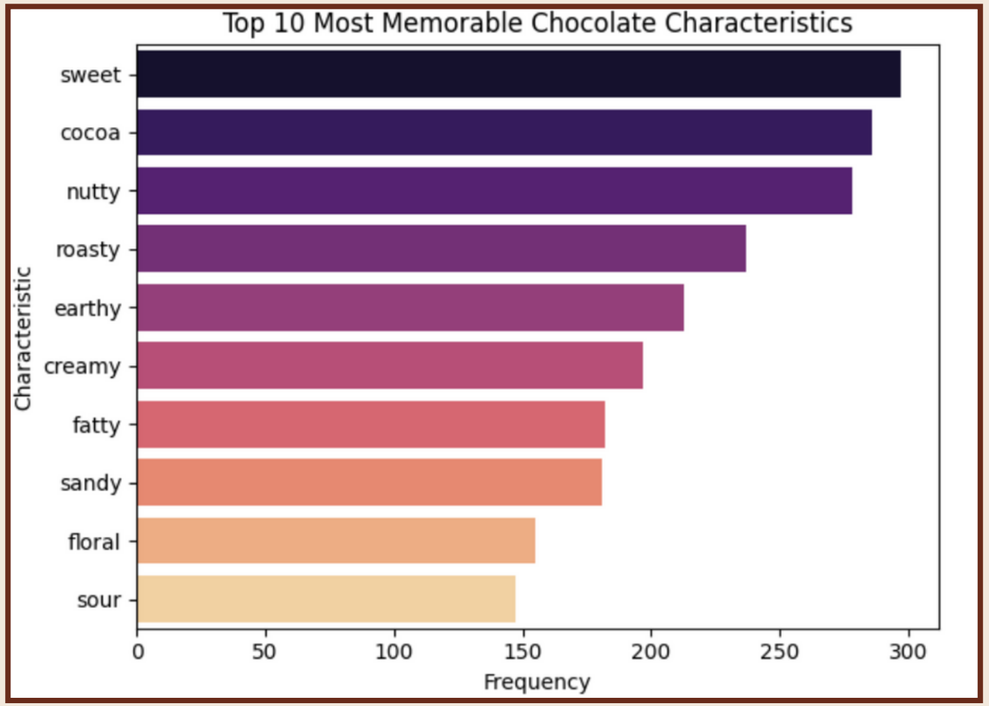
- Bar plot of the 10 most frequent Manufacturing Company values
- Bar plot showing the countries contributing the most chocolate entries
- Box plot of average consumer rating across the 10 most common Specific Bean Origin values
- Box plots comparing rating distributions for the ten most-used ingredients
- Bar chart of the top 10 most frequently mentioned tasting-note tokens in Most Memorable Characteristics
- Correlation Heatmap showing pairwise correlations among Cocoa Percent, Rating, and Review Year
- Scatter plot of Cocoa Percent (x) vs. Rating (y), colored by Review Year to reveal temporal trends

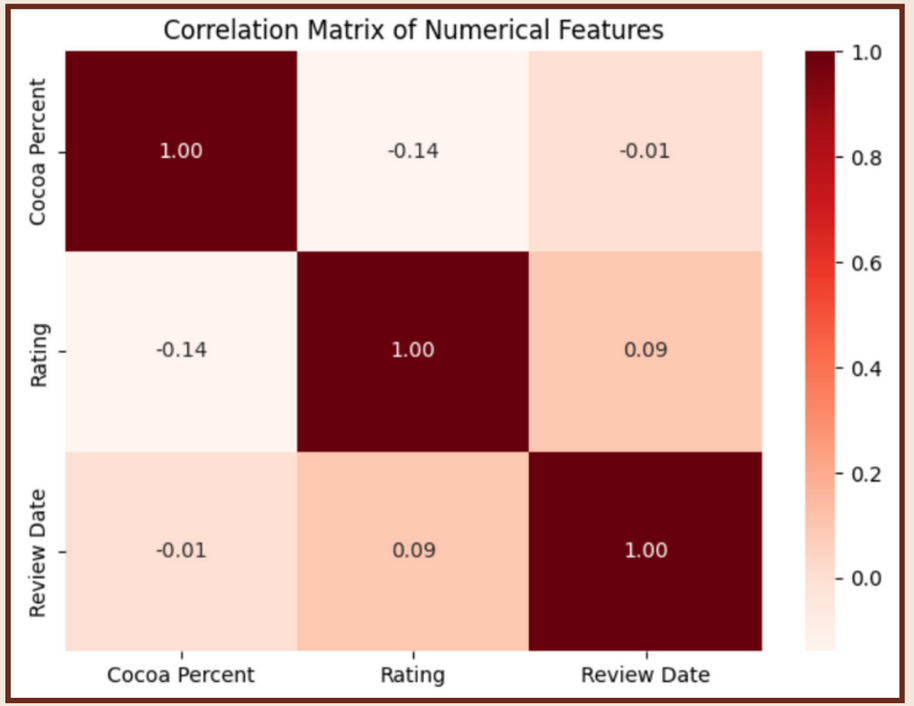

Top 10 Chocolate Manufacturers
Number of Products
0
10
20
30
40
50
60
Soma
Fresco
Bonnat
Arete
A. Morin
Dandelion
Pralus
Domori
Castronovo
Guittard
Manufacturing Company

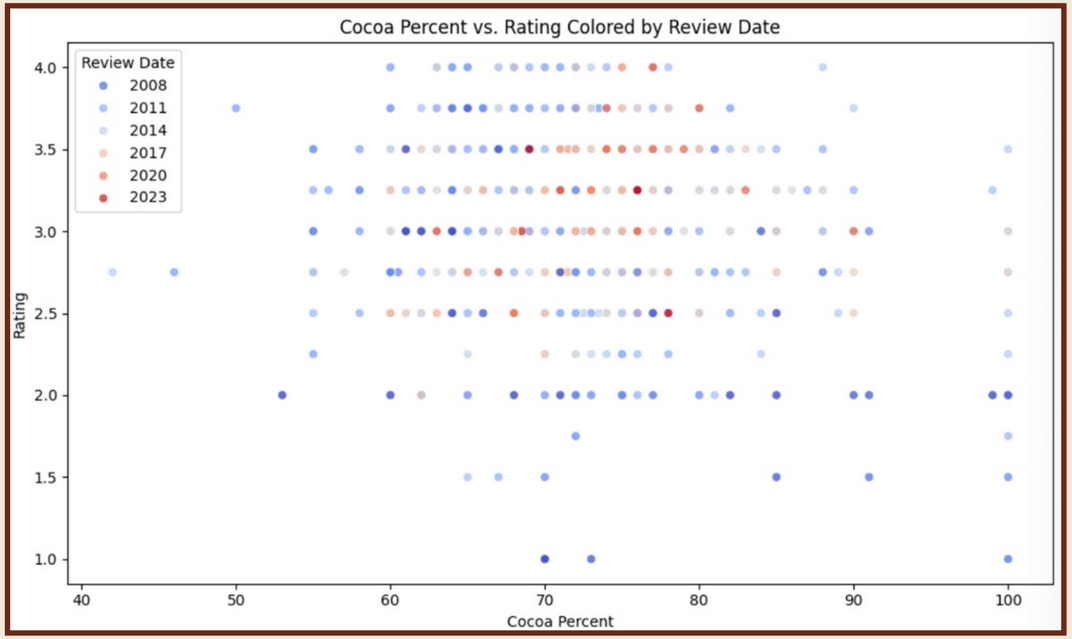

Top 10 Bean Producing Countries
Number of Products
0
50
100
150
200
250
Peru
Venezuela
Ecuador
Dominican Republic
Madagascar
Blend
Nicaragua
Tanzania
Colombia
Vietnam
Country of Bean Origin

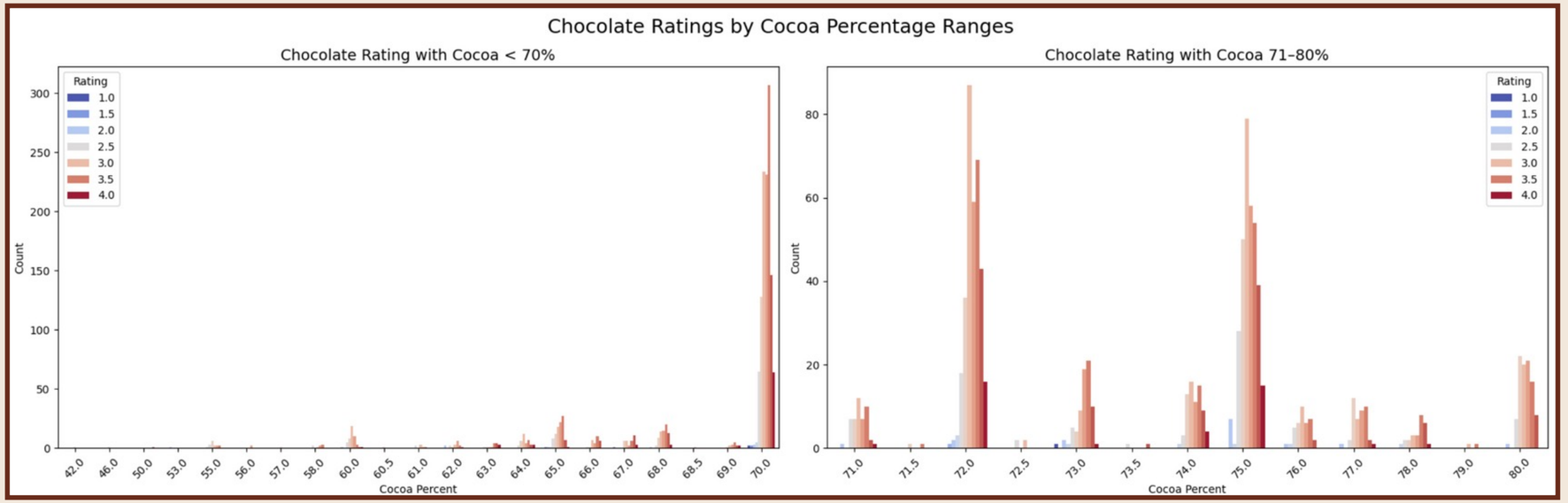

Rating Distribution by Top 10 Ingredient Combinations
Rating
1.0
1.5
2.0
2.5
3.0
3.5
4.0
3- B,S,C
2- B,S
4- B,S,C,L
4- B,S,C,V
5- B,S,C,V,L
3- B,S*,C
2- B,S*
3- B,S,L
4- B,S*,C,Sa
4- B,S*,C,V
Ingredients

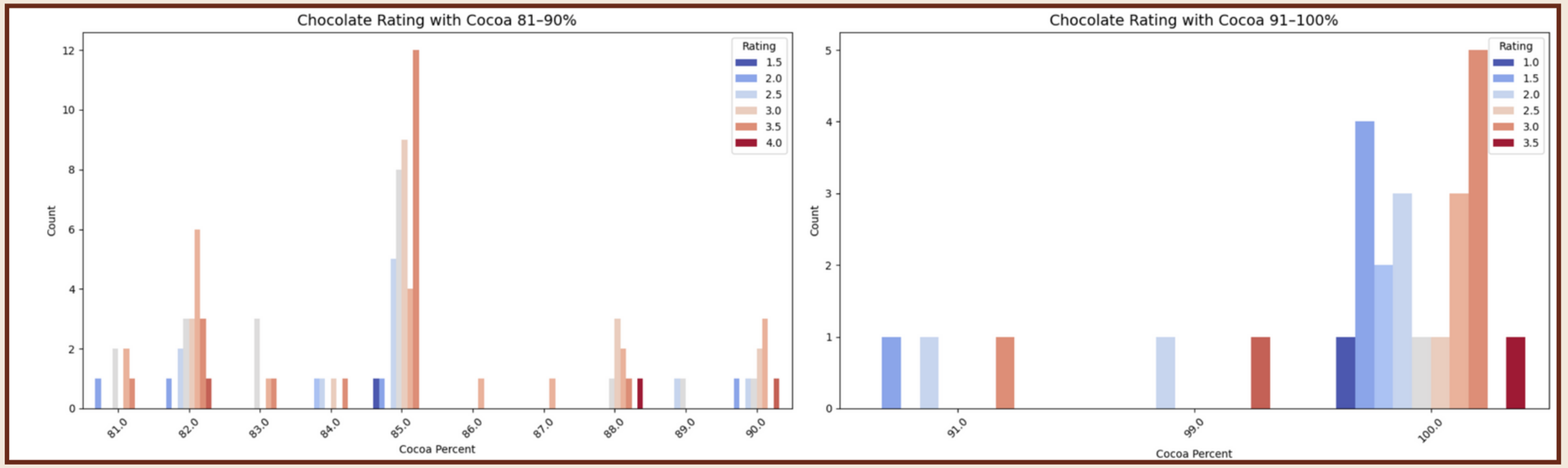

Rating Distribution for Top 10 Specific Bean Origins
Rating
1.5
2.0
2.5
3.0
3.5
4.0
Peru
Chuao
Madagascar
Ghana
Ecuador
Dominican Republic
Venezuela
Kokoa Kamili
Ucayali
Sambirano
Specific Bean Origin

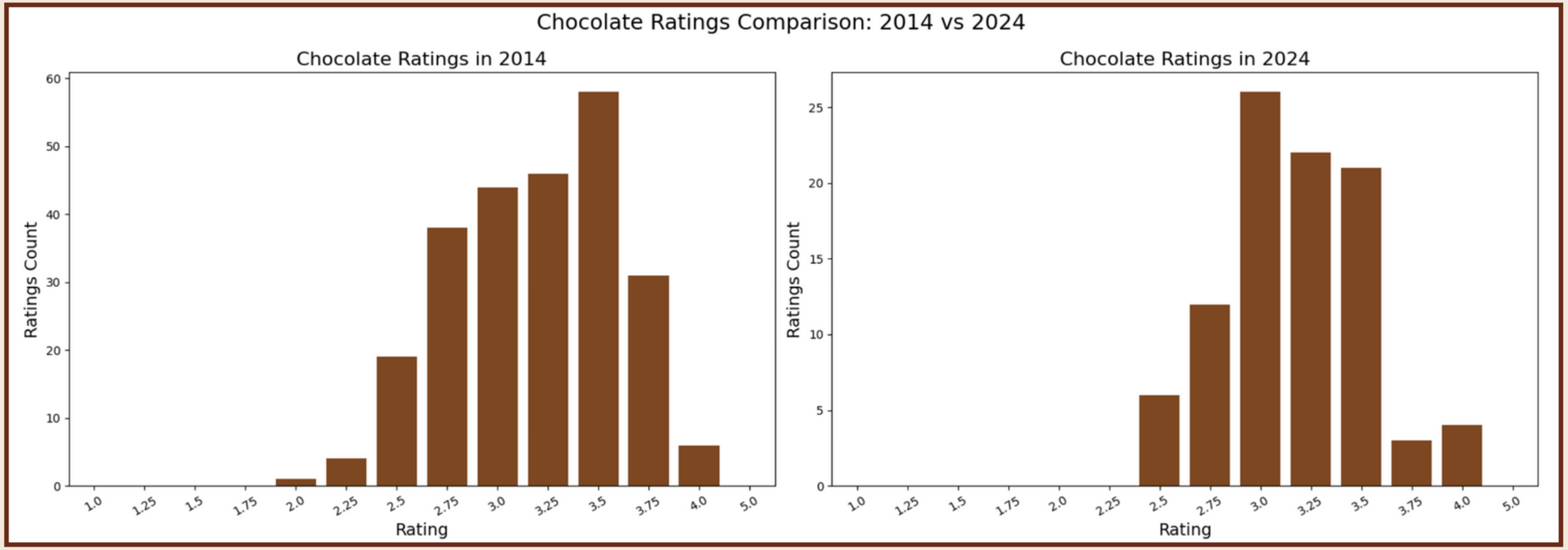

Top 10 Most Memorable Chocolate Characteristics
sweet
cocoa
nutty
roasty
earthy
creamy
fatty
sandy
floral
sour
Characteristic
0
50
100
150
200
250
300
Frequency


Correlation Matrix of Numerical Features
Cocoa Percent
Rating
Review Date
1.00
-0.14
-0.01
-0.14
1.00
0.09
-0.01
0.09
1.00
Cocoa Percent
Rating
Review Date
1.0
0.8
0.6
0.4
0.2
0.0


Cocoa Percent vs. Rating Colored by Review Date
Review Date
2008
2011
2014
2017
2020
2023
Rating
4.0
3.5
3.0
2.5
2.0
1.5
1.0
40
50
60
70
80
90
100
Cocoa Percent

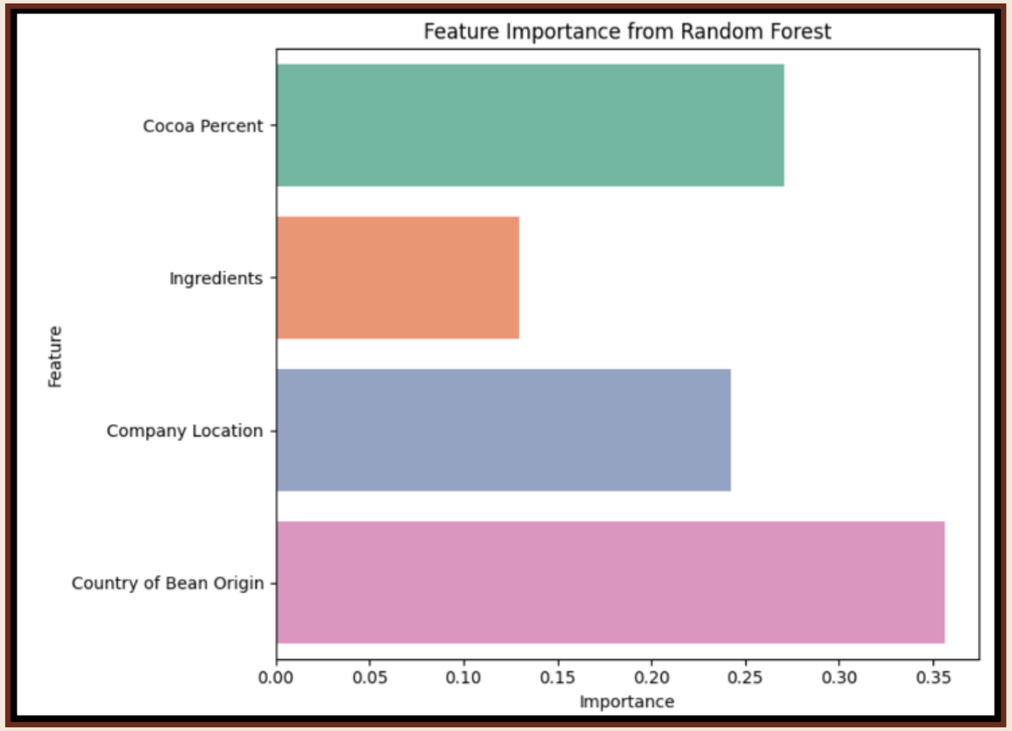

Chocolate Ratings by Cocoa Percentage Ranges
Chocolate Rating with Cocoa < 70%
Chocolate Rating with Cocoa 71–80%
Rating
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Count
Cocoa Percent

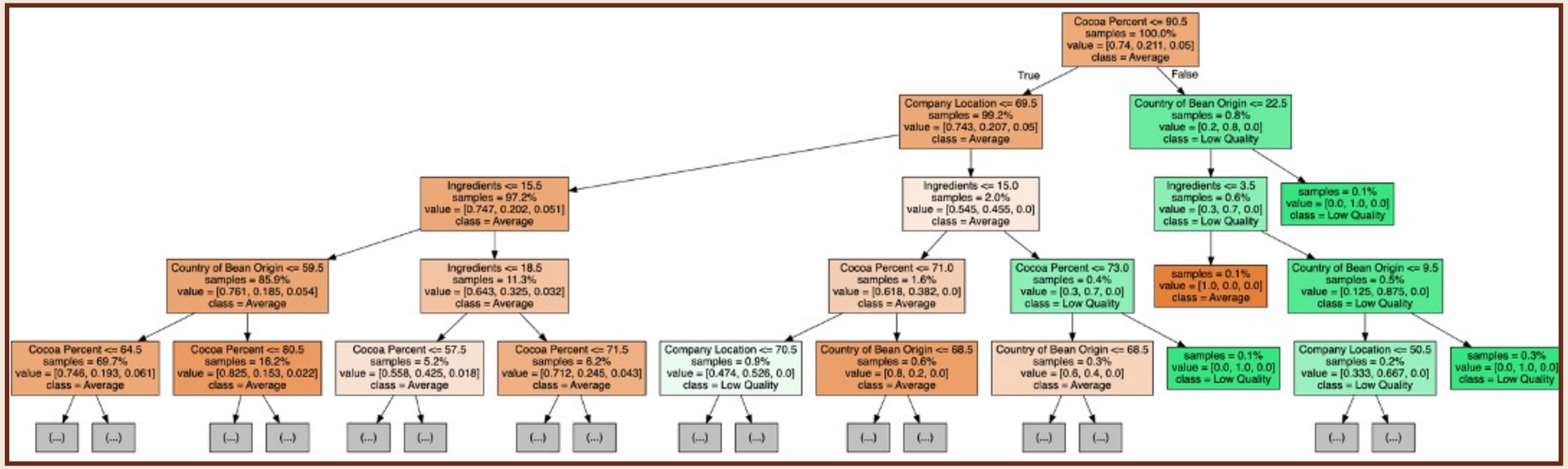

Chocolate Rating with Cocoa 81–90%
Chocolate Rating with Cocoa 91–100%
Rating
Count
Cocoa Percent


Chocolate Ratings Comparison: 2014 vs 2024
Chocolate Ratings in 2014
Ratings Count
Rating
0
10
20
30
40
50
60
1.0
1.25
1.5
1.75
2.0
2.25
2.5
2.75
3.0
3.25
3.5
3.75
4.0
5.0
Chocolate Ratings in 2024
Ratings Count
Rating
0
5
10
15
20
25
1.0
1.25
1.5
1.75
2.0
2.25
2.5
2.75
3.0
3.25
3.5
3.75
4.0
5.0

RANDOM FOREST CLASSIFER

- Used to classify chocolates as Premium, Average, and Low Quality
- Trained on cocoa percentage, ingredients, and origin
- Achieved ~66% accuracy on 80/20 split
- Ranked features by importance
- Reduced overfitting using bootstrapped trees
- Country of Bean Origin and Cocoa Percentage ranked highest by feature importance

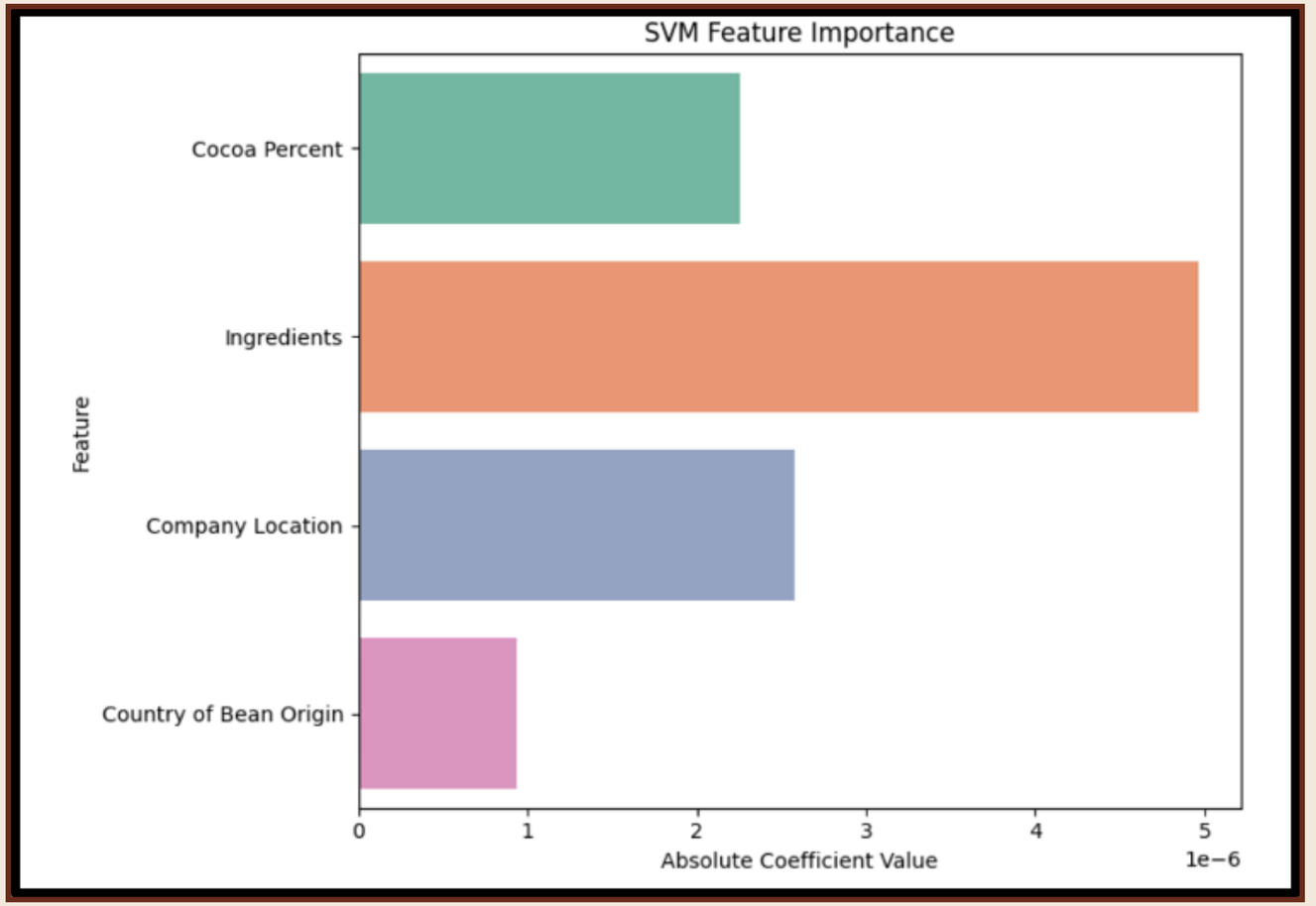




Decision Tree

SUPPORT VECTOR MACHINE

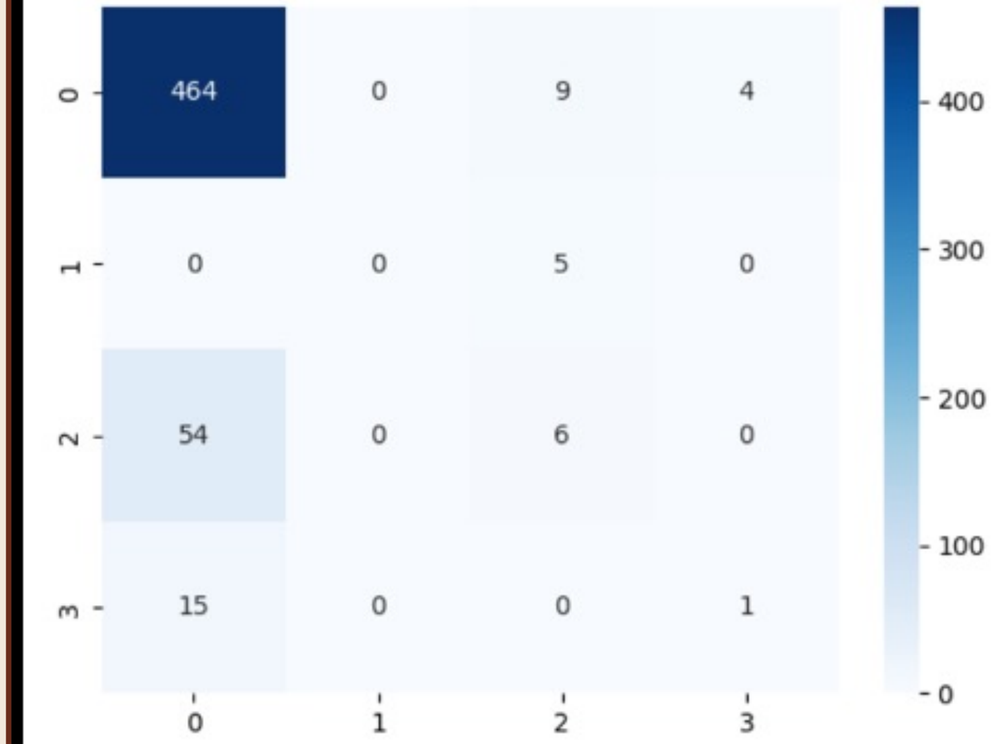
- Linear kernel SVM used with label-encoded numeric and categorical features
- Accuracy ~72.4%, but biased toward majority class “Average”
- Showed class imbalance issues
- Feature weights highlighted ingredients, company location and cocoa percentage


SVM Feature Importance
Cocoa Percent
Ingredients
Company Location
Country of Bean Origin
Feature
0
1
2
3
4
5
Absolute Coefficient Value
1e−6

Naïve Bayes Classification

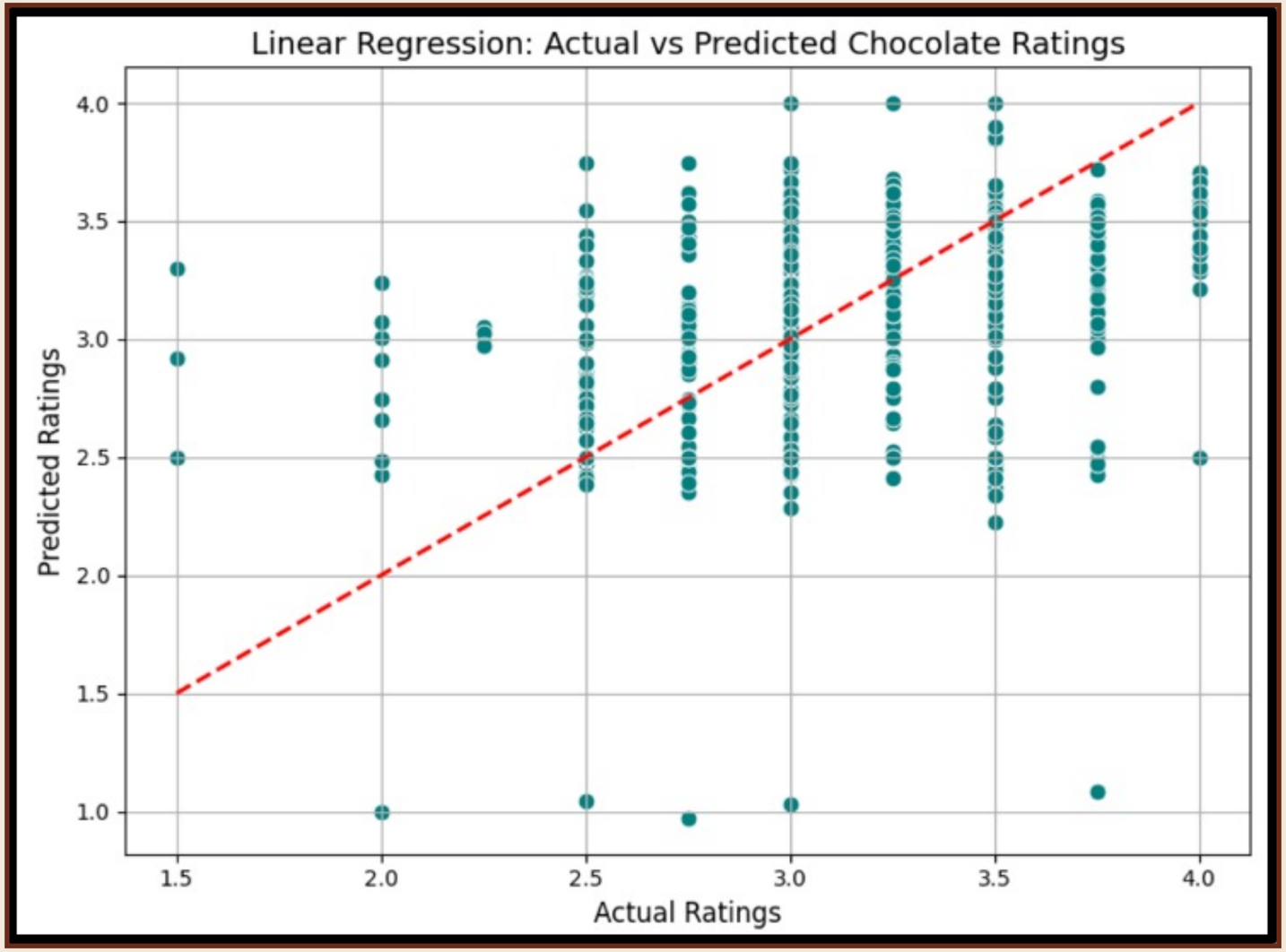
- Classified chocolate type based on tasting notes
- Used CountVectorizer for text features
- MultinomialNB on bag-of-words tasting-note vectors
- Accuracy ~84%, but dominated by “Bittersweet Chocolate” (recall 97 %)
- Top tokens per class (e.g., “cocoa,” “sweet,” and “nutty” for Bittersweet; “bitter,” “intense,” and “floral” for Liquor)

```
Top tokens per class:
  Bittersweet Chocolate: cocoa, sweet, nutty, fruit, roasty, sour, earthy, mild, creamy, sandy
  Chocolate Liquor: bitter, fruit, intense, floral, coffee, strong, fatty, earthy, smooth, pastey
  Dark Chocolate: bitter, mild, intense, fruit, fatty, roasty, nutty, cocoa, sour, tart
  Milk Chocolate: sweet, cocoa, vanilla, banana, gritty, caramel, nutty, intense, floral, creamy
```



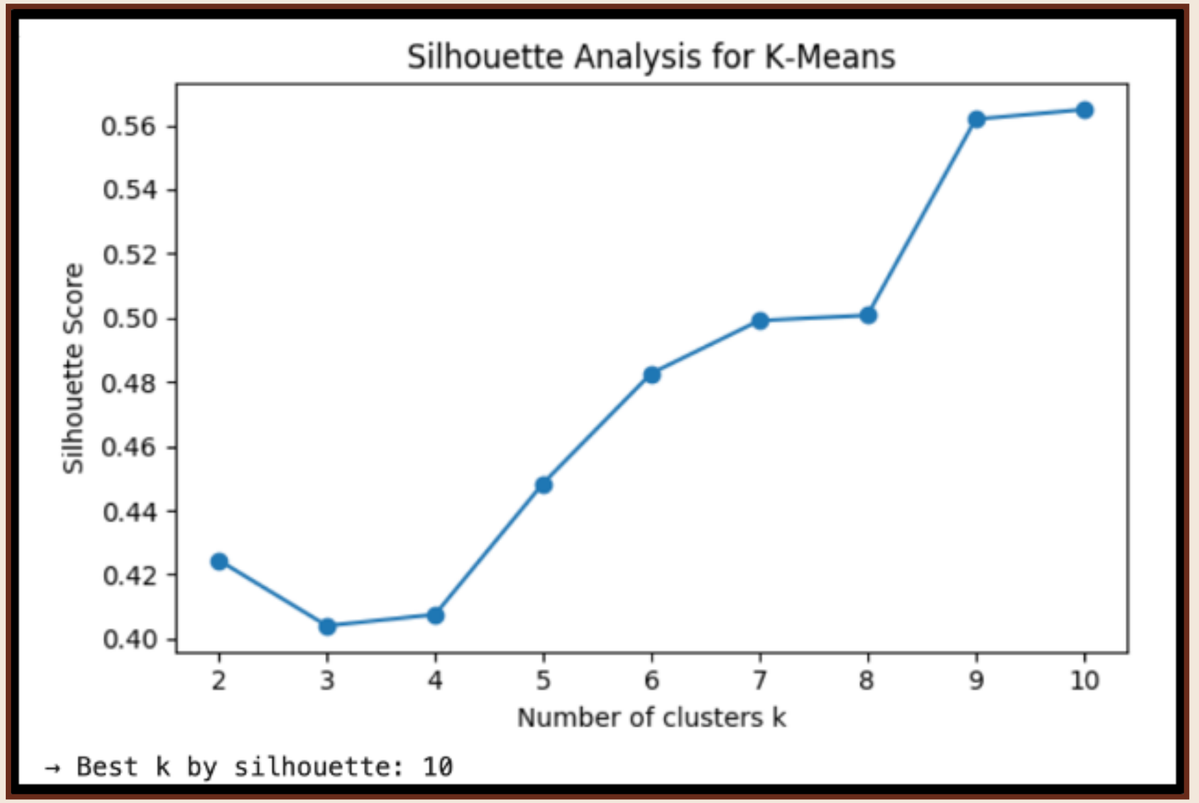
LINEAR REGRESSION

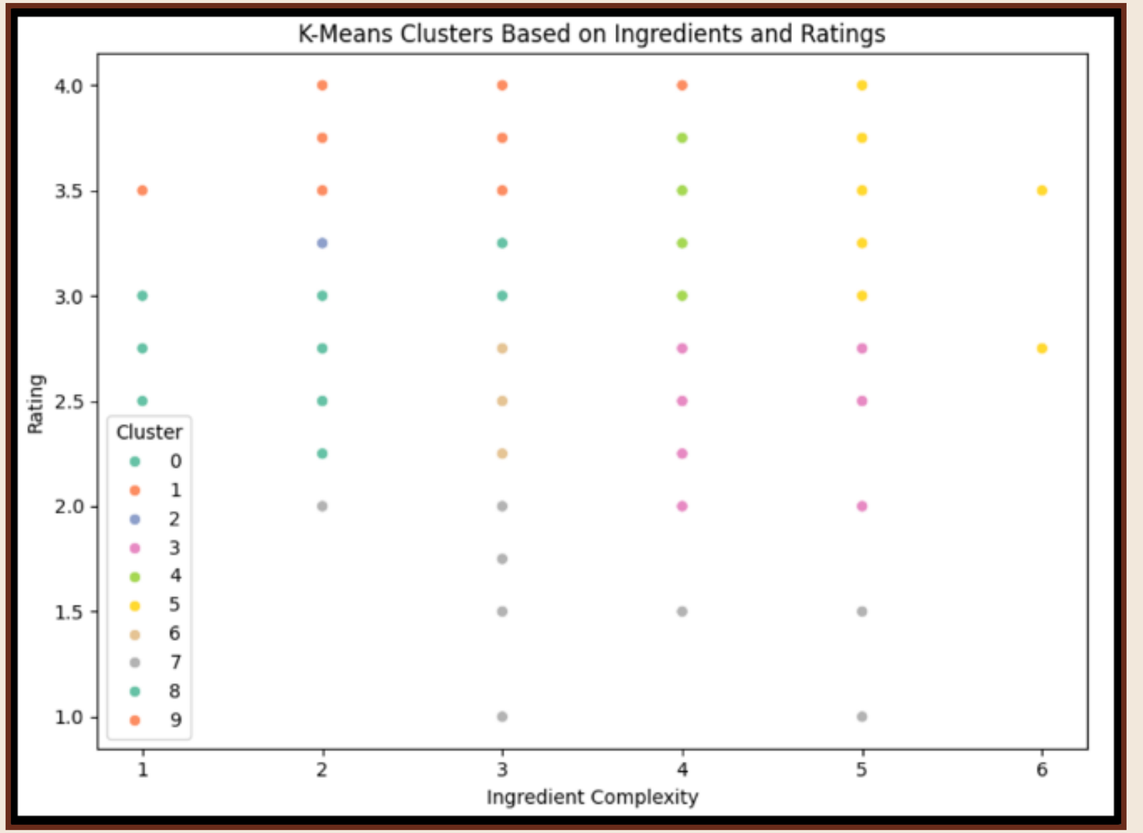
- Predicted continuous ratings from Cocoa Percentage and the manufacturer
- Performance: RMSE ≈ 0.492; R^2 ≈ −0.207 (20% variance explained)
- Partial Linear correlation found
- Actual vs. Predicted shows a linear trend with residual dispersion


Linear Regression: Actual vs Predicted Chocolate Ratings
Predicted Ratings
Actual Ratings
4.0
3.5
3.0
2.5
2.0
1.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0

K-MEANS CLUSTERING

- Unsupervised clustering using rating and ingredient complexity
- Silhouette analysis identified the best K
- Coherent clusters; cluster-mean RMSE ≈ 0.039 as a baseline rating predictor
- Revealed natural groupings (e.g., high-complexity, high-rating)


Silhouette Analysis for K-Means
Silhouette Score
0.56
0.54
0.52
0.50
0.48
0.46
0.44
0.42
0.40
2
3
4
5
6
7
8
9
10
Number of clusters k
→ Best k by silhouette: 10


K-Means Clusters Based on Ingredients and Ratings
Rating
4.0
3.5
3.0
2.5
2.0
1.5
1.0
Cluster
0
1
2
3
4
5
6
7
8
9
1
2
3
4
5
6
Ingredient Complexity

MODEL SELECTION & EVALUATION

- Classification Metrics:
  - Accuracy
  - Precision, Recall, F1-Score
- Confusion Matrix: detailed error analysis
- Regression Metrics
  - RMSE: root-mean-square error (penalizes large errors)
  - R^2: R Squared (variance in the dependent variable)

```
Confusion Matrix:
[[341  45  18]
 [105  23   3]
 [ 20   1   2]]
Accuracy:
0.6559139784946236

Classification Report:
              precision    recall  f1-score   support

     Average       0.73      0.84      0.78       404
 Low Quality       0.33      0.18      0.23       131
     Premium       0.09      0.09      0.09        23

    accuracy                           0.66       558
   macro avg       0.38      0.37      0.37       558
weighted avg       0.61      0.66      0.63       558
```

Results of Random Forest Classifier

```
Confusion Matrix:
[[404   0   0]
 [131   0   0]
 [ 23   0   0]]

Accuracy:
0.7240143369175627

Classification Report:
              precision    recall  f1-score   support

     Average       0.72      1.00      0.84       404
 Low Quality       0.00      0.00      0.00       131
     Premium       0.00      0.00      0.00        23

    accuracy                           0.72       558
   macro avg       0.24      0.33      0.28       558
weighted avg       0.52      0.72      0.61       558
```

Results of Support Vector Machine

```
Confusion Matrix:
[[464   0   9   4]
 [  0   0   5   0]
 [ 54   0   6   0]
 [ 15   0   0   1]]
Accuracy: 0.844
Classification Report:
                     precision    recall  f1-score   support

Bittersweet Chocolate       0.87      0.97      0.92       477
     Chocolate Liquor       0.00      0.00      0.00         5
       Dark Chocolate       0.30      0.10      0.15        60
       Milk Chocolate       0.20      0.06      0.10        16

             accuracy                           0.84       558
            macro avg       0.34      0.28      0.29       558
         weighted avg       0.78      0.84      0.80       558
```

Results of Naïve Bayes Classifier

```
Cluster summary:
         avg_cocoa  avg_complex  avg_rating  count
cluster
0        71.666342     3.000000    3.126459    514
1        71.409091     3.031540    3.647495    539
2        72.064327     2.000000    3.250000    171
3        70.745902     4.289617    2.553279    183
4        70.355422     4.000000    3.304970    332
5        68.892857     5.028571    3.332143    140
6        72.360619     3.000000    2.662611    226
7        85.800000     3.080000    1.610000     25
8        72.630564     1.985163    2.820475    337
9        71.625776     1.996894    3.615683    322

Clustering-based Root Mean Squared Error: 0.039
```

Results of K-Means Clustering

CONCLUSION

- Machine Learning Models effectively classify and predict chocolate quality
- Preprocessing and Exploratory Data Analysis provided deep insights
- Random Forest & Naïve Bayes delivered the strongest results
- SVM revealed the need for imbalance handling
- Linear Regression pointed to missing features
- K-Means offered meaningful market segments and a solid rating baseline
- Future work: enrich features, balance classes, deploy model for real-time quality control

THANK YOU