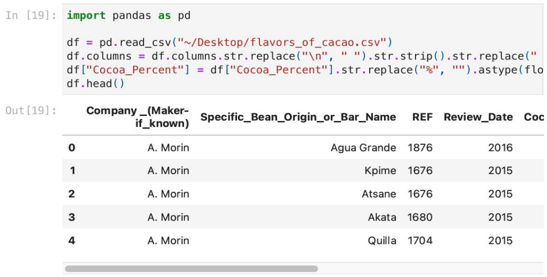
Group 7   STAT 451 Project Proposal

Unwrapping Chocolate: Modeling and Classifying Chocolate Bar Ratings

Emily Finkelmeyer, Jeff Zhang, Dianhao Sun, Yinuo Zhou, Tianshu Zhang

Reading the Data

We use the chocolate bar ratings dataset from Kaggle https://www.kaggle.com/datasets/rtatman/chocolate-bar-ratings   Below is our data processing code:



```
In [19]: import pandas as pd

         df = pd.read_csv("~/Desktop/flavors_of_cacao.csv")
         df.columns = df.columns.str.replace("\n", " ").str.strip().str.replace("
         df["Cocoa_Percent"] = df["Cocoa_Percent"].str.replace("%", "").astype(flo
         df.head()
```

Out[19]:

	Company _(Maker-if_known)	Specific_Bean_Origin_or_Bar_Name	REF	Review_Date	Coc
0	A. Morin	Agua Grande	1876	2016	
1	A. Morin	Kpime	1676	2015	
2	A. Morin	Atsane	1676	2015	
3	A. Morin	Akata	1680	2015	
4	A. Morin	Quilla	1704	2015	

Project Description

Our goal is to explore key factors influencing chocolate bar ratings and build predictive models and we address three main questions:

1. What features (e.g., cocoa percentage, ingredients, origin, manufacturer) are most associated with higher ratings?
2. Can we classify chocolate bars into high, medium, or low ratings using classification models?
3. Can we predict exact ratings using interpretable regression methods?

Variables Used

- **Rating**: Target variable (1.0–5.0), used for regression and classification.
- **Cocoa_Percent**: Numeric cocoa content (converted from % to float).
- **Company, Company_Location, Broad_Bean_Origin**: Categorical, one-hot encoded.
- **Ingredients**: Encoded as binary features indicating presence of each ingredient.
- **Review_Date**: Year of review, potentially used to explore trends over time.

Methods:

- **Exploratory Analysis**: Use visual tools (boxplots, scatterplots, heatmaps) to examine how features like cocoa percentage, origin, and ingredients relate to ratings.
- **Classification**: Convert ratings into three classes (High ≥ 4.0, Medium 3.0–3.99, Low < 3.0) and apply models like logistic regression, k-NN, decision trees, and SVM. Evaluate using accuracy, confusion matrices, and cross-validation.
- **Regression**: Predict exact ratings using interpretable models, starting with linear regression and comparing with random forest. Assess performance using RMSE, R^2, and residual plots.