Chocolate Bar Ratings Predictor

Team Candy
Bruce Jilek, Tahereh Javaheripour, Yan Kin Pang, Travis Loseke

Topic Selection

Predicting chocolate bar ratings from variable such as:

- **Chocolateur**
- **Bean type**
- **Bean Source**
- **Cocoa Percent**

Reason for Topic Selection

- Examined several food and beverage topics
- Chocolate was universally loved by the team
- Dataset was expansive enough to cover final project requirement
- Team became genuinely interested in if cocoa bean types and their source drive chocolate quality.

Team Roles

- Square -(Bruce) Set up Repository and Branch structure
- Triangle - (Tahereh) Build additional ML models to see initial accuracy
- Circle - (Yan) Constructed EDR for PostGres SQL and begin building framework
- X- (Travis) Presentation and technology

Data sources

- Kaggle- Chocolate Bar ratings
  - https://www.kaggle.com/rtatman/chocolate-bar-ratings
- Kaggle - Countries and States Lat Long
  - https://www.kaggle.com/paultimothymooney/latitude-and-longitude-for-every-country-and-state
- Flavors of Cacao website
  - http://flavorsofcacao.com/chocolate_database.html

Four questions to answer

- Can we predict if a bar will be rated in the top 15%(>=3.75)?
  - 1 SD above the mean
- Where are the best cocoa beans grown?
- Which countries produce the highest-rated bars?
- What's the relationship between cocoa solids percentage and rating?

Data exploration phase

- Pandas and Matplot lib were used to review, evaluate and clean datasets
  - Examine null counts
  - Data continuity
  - Data types
- Tableau was used to make initial map plots to look at geographic distribution issues related to typographical errors and level of detail

Data Analysis phase

- Pandas and Matplot lib were used to review, evaluate and clean datasets
  - Initial plots of rating trends vs other features
  - Examine plots to look for skewness in the features and dataset
  - Quicklook plots for any trends or significant outlier anomalies
- Tableau was used to make initial map plots to look at geographic distribution of cocoa bean sources and data bias in geographic space

Data Processing and Evaluation Workflow

Raw CSV Data

Preliminary Data cleaning and formatting using Pandas to clean typos and other

Upload to Amazon S3

Send to PdAdmin

Upload to AWS RDS database

Data is Merged and additional formatting with Pandas

Cleaned data is visualized with Tableau

Loaded into Machine Learning model Utilizing SciKit Learn

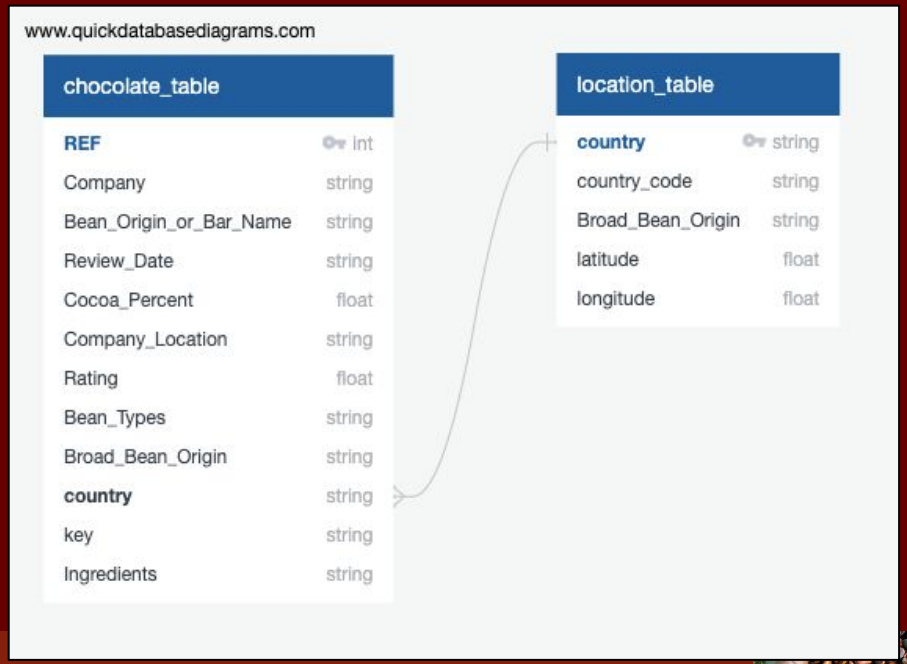
Data Preprocessing and Cleaning

- Pandas will be the primary avenue for cleaning and preparing the data
- Basic data editing of the csv will initially be performed
- Data will be uploaded to Amazon S3
- PDAdmin to create PostGres SQL with intake data and push to AWS RDS
- Python will merge data and format for use in Machine Learning and export out data to S3 for use in Tableau

Database Structure

AWS Database- Postgres SQL Database

Initial ERD:



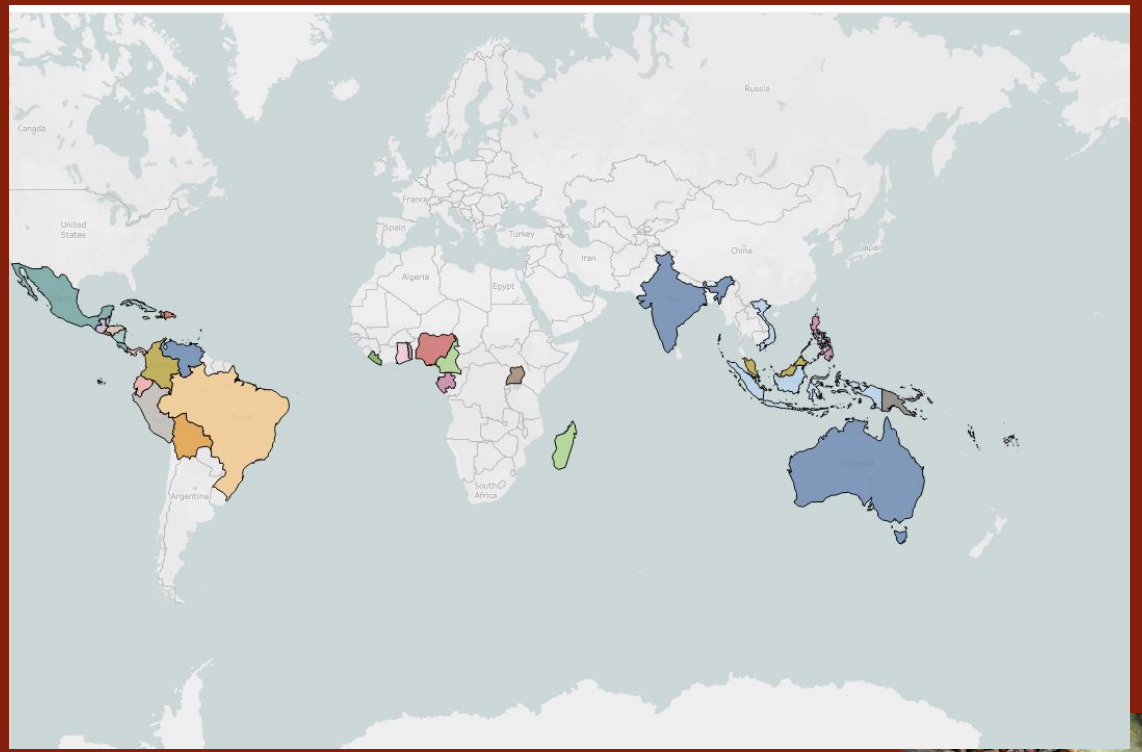
Machine Learning Model

- Review three ML classifiers
  - Random Forest
  - SVM
  - Neural Net
- Review model weighting and confusion matrices
- Choose simplest model that best describes and predicts the data

Dashboard

Tableau-

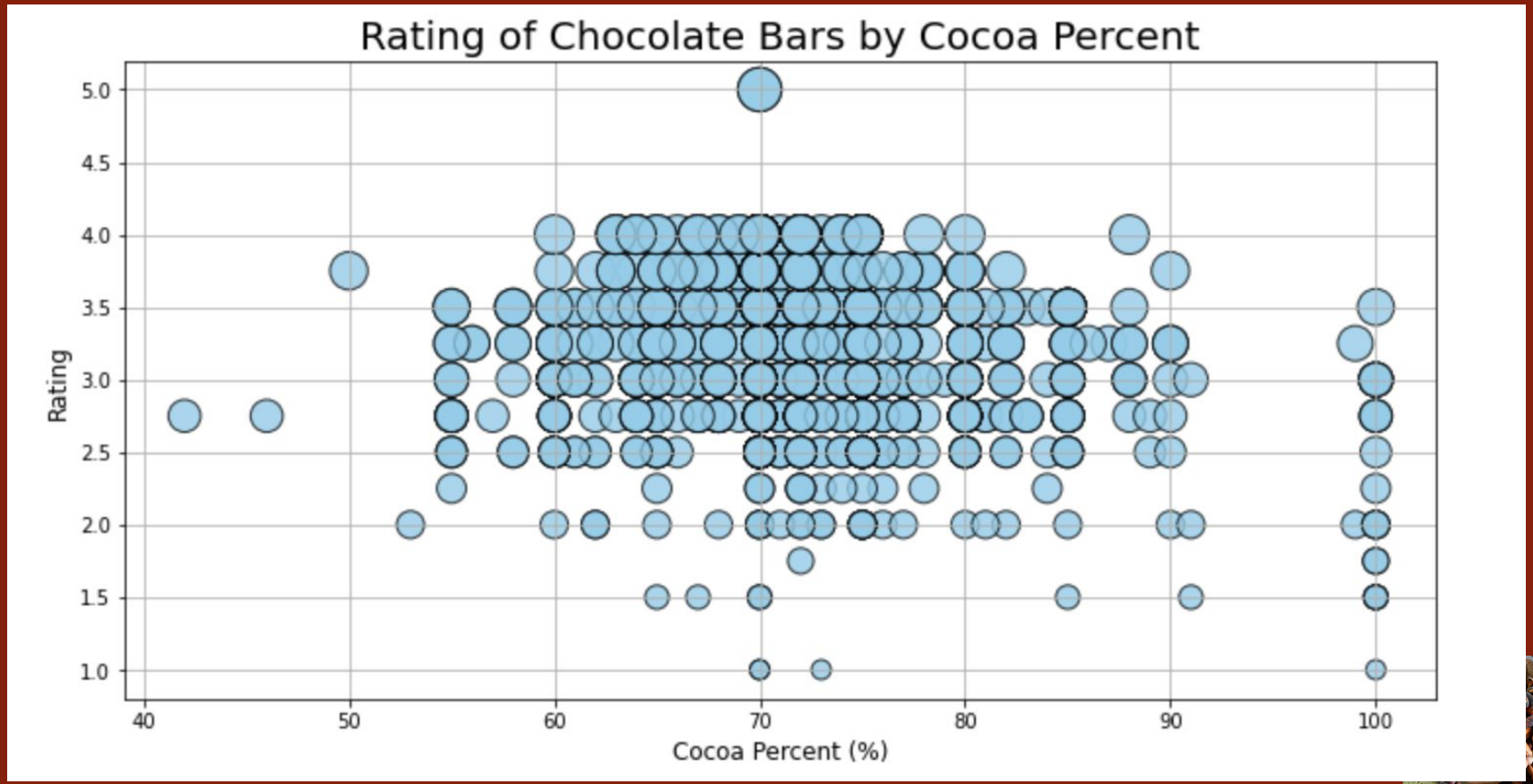
Global Chocolate production

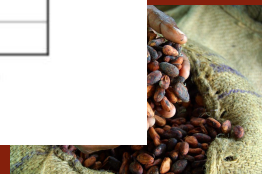


Where are the best beans grown?

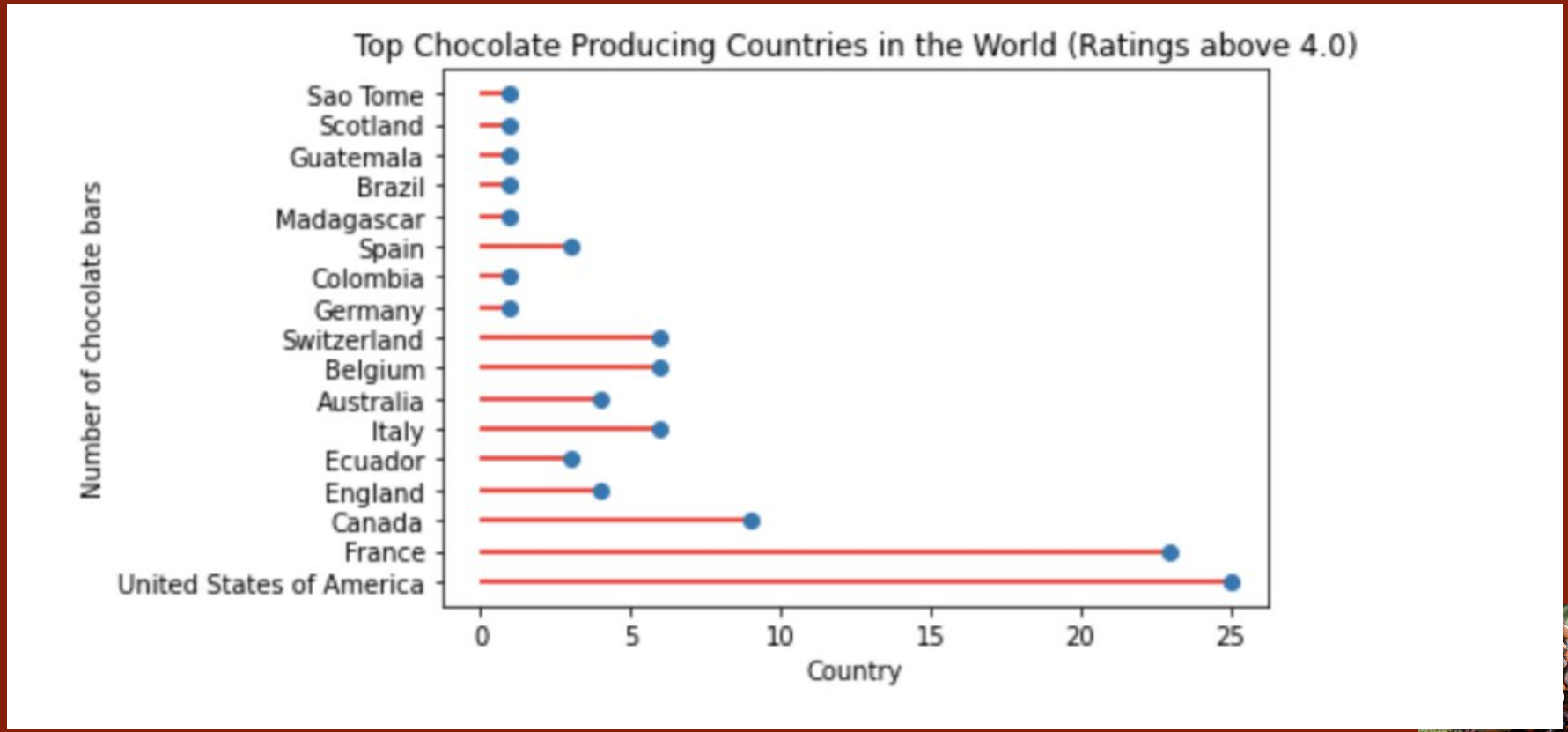
	Broad_Bean_Origin	REF	Review_Date	Cocoa_Percent	Rating
43	Solomon Islands	1811.000000	2016.250000	74.000000	3.437500
17	Haiti	1354.444444	2014.000000	71.333333	3.388889
19	Honduras	1478.666667	2014.533333	73.933333	3.350000
16	Guatemala	1352.758621	2013.896552	71.758621	3.344828
39	Republic of Congo	1091.600000	2012.600000	70.500000	3.325000

Does Cocoa % lead to better bar rating

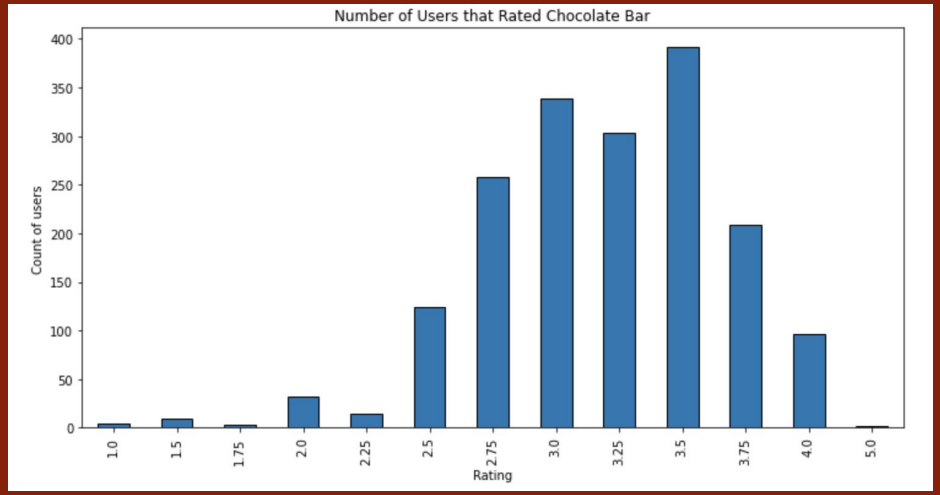


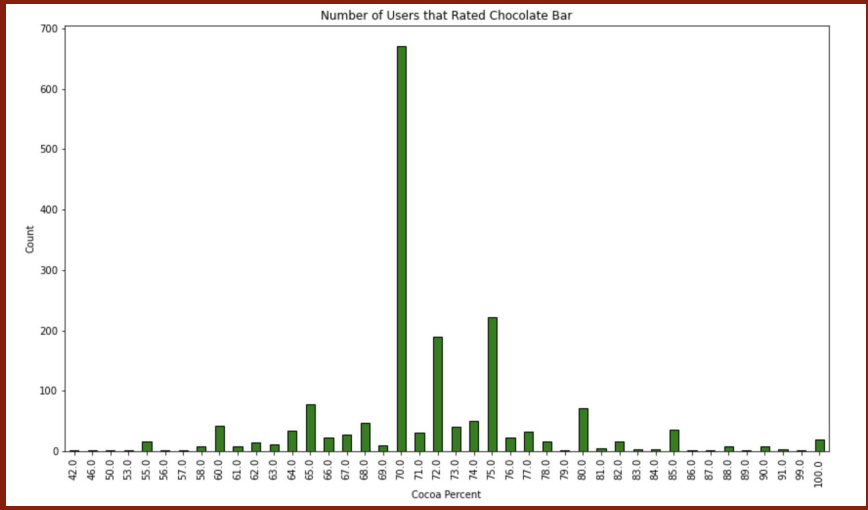


Best Chocolate Crafting Country



Histogram of Bar Ratings





Slight left skew on the chocolate bar ratings