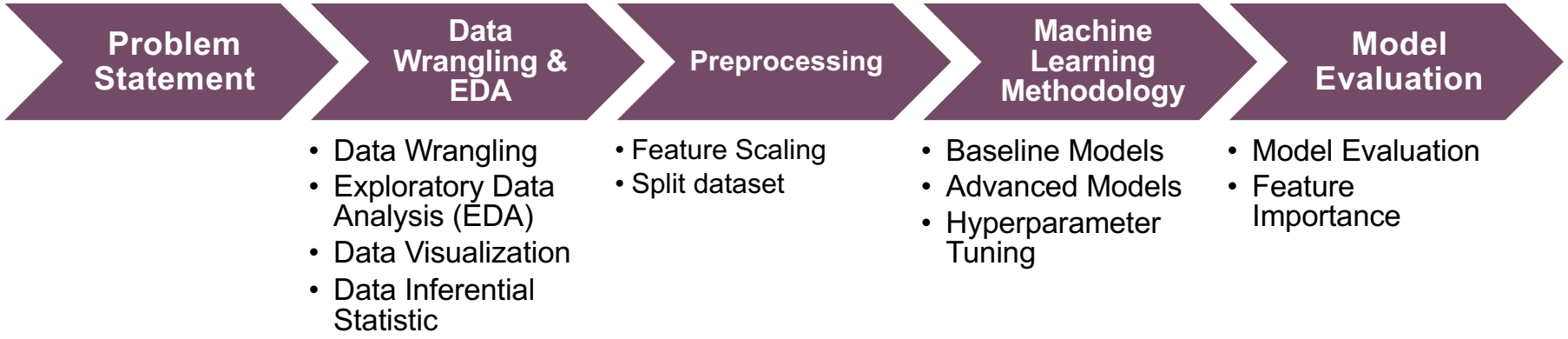
COFFEE REVIEW ANALYSIS & RATING PREDICTION

Wenshan, Liu

PROJECT OBJECTIVES



PROBLEM STATEMENT

Challenges in Coffee Quality Assessment

- Coffee quality evaluation is highly **subjective**, often based on **expert cupping scores**.
- Key attributes include: **aroma, acidity, aftertaste, body, balance**, and **flavor**.
- These evaluations are **complex** and can lead to **inconsistencies**.
- It's often difficult for **producers, roasters, and distributors** to pinpoint what exactly drives high scores.
- The scoring system may vary due to **human bias** or **inconsistent grading standards** across different experts.

DATA SOURCES

Scape coffee review data from *Coffee Review* website

Weather data of coffee region from *Open Weather API*

Geography data from Python package

DATA WRANGLING
BASIC DATA CLEANING

- **Handling Missing Values**
  - Missing values were removed if necessary

- **Encoding Categorical Variables**
  - Convert Categorical
  - Encoded region data by using one-hot encoding

- **Deal with the coffee price in the same unit**
  - Convert various currency and weight units to USD per pound.

EXPLORATORY DATA ANALYSIS (EDA)

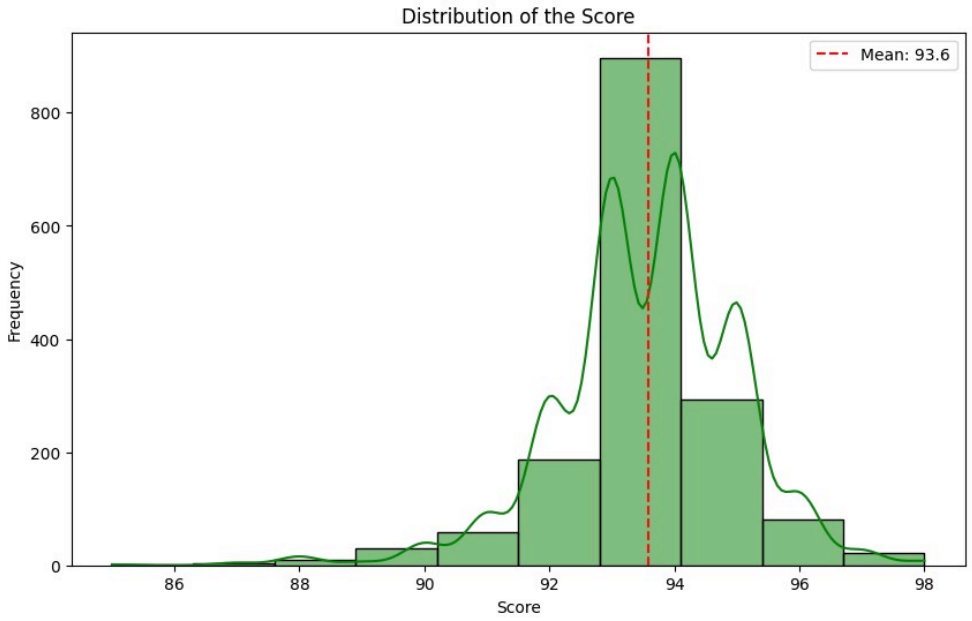
Data Visualization

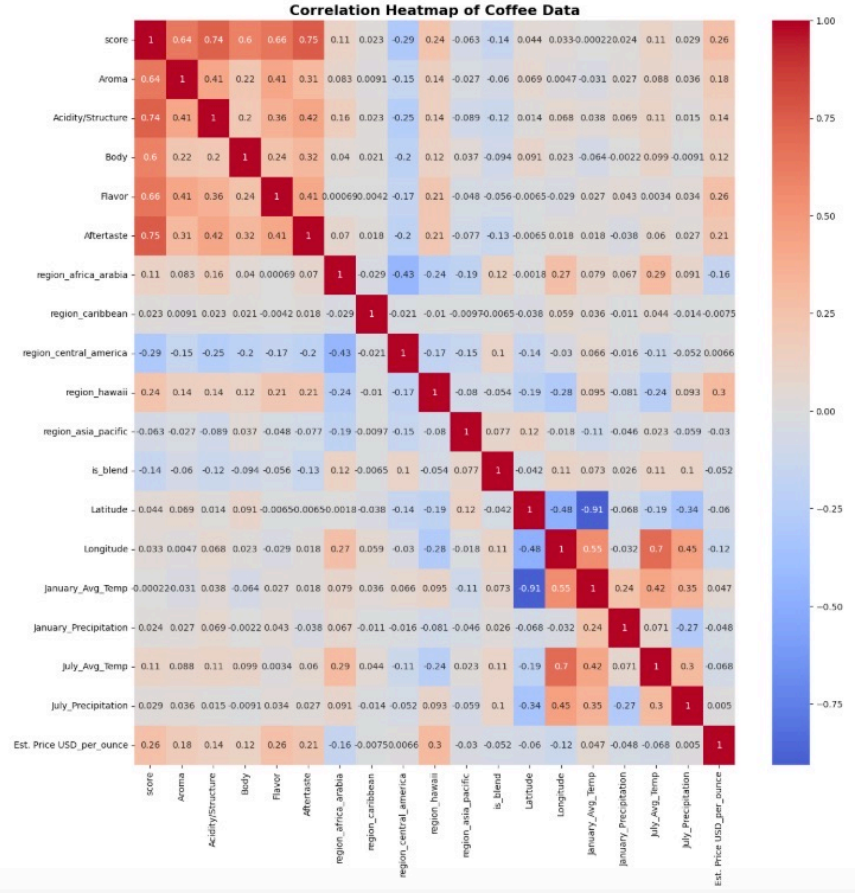
- Sensory features such as Aroma, Acidity had a strong positive correlation with overall coffee ratings.
- Coffee beans from Hawaii tended to receive higher median ratings compared to other regions.

Data Inferential Statistic

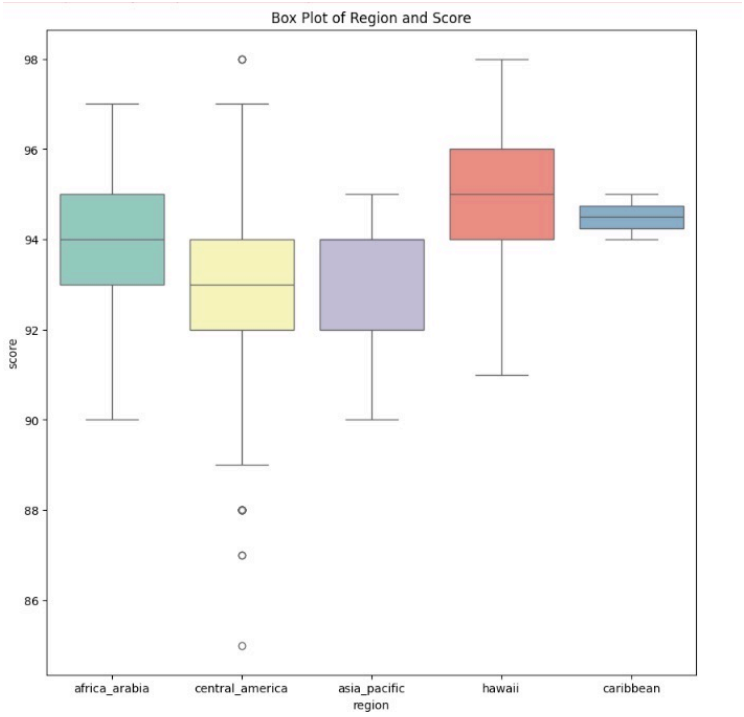
- Medium roast coffees demonstrated the highest average scores and dark roast coffees showed significantly lower scores compared to medium roasts
- blending coffee beans is significantly affects quality ratings

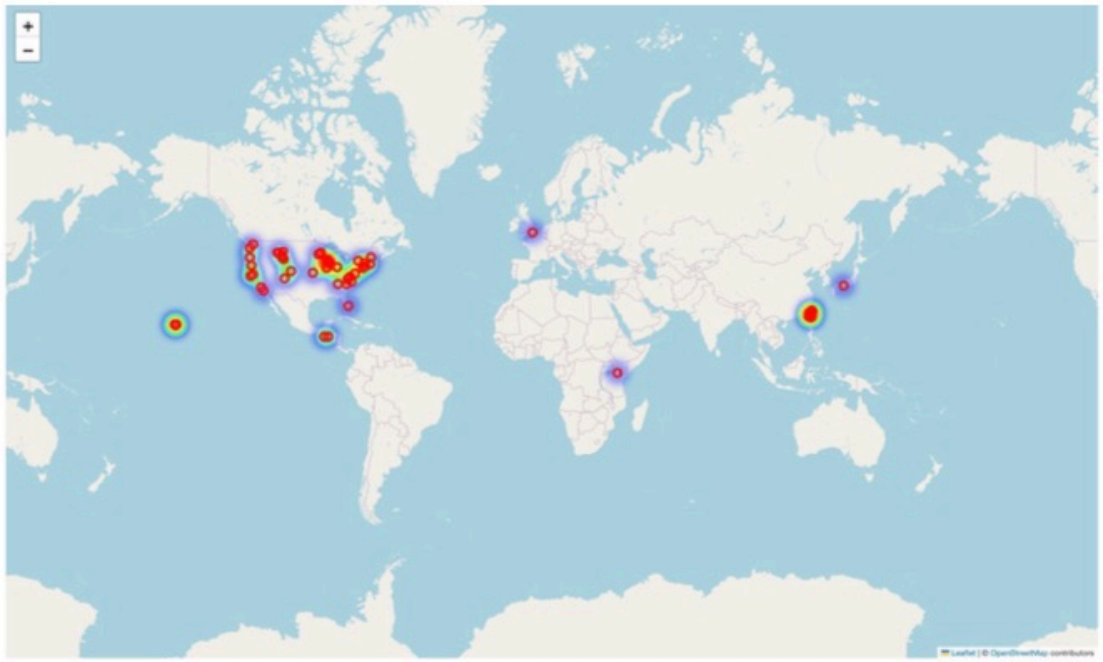
DATA VISUALIZATION





DATA VISUALIZATION





DATA INFERENTIAL STATISTIC

Statistic test	Examination	Result
ANOVA	coffee origin and the score	p-value < 0.05, indicating that the region of coffee beans significantly affects the score.
t-test	coffee blended (Y/N) and the score (continuous numeric)	p-value < 0.05, meaning the analysis showed significant differences in coffee scores depending on whether the coffee is blended or not. This indicates that blending coffee beans significantly affects the scores.
ANOVA	Roasters and prices	p-value <0.05, which means the analysis showed significant differences in coffee price across different roasters.
ANOVA	Roast levels and scores	p-value<0.05, which is significant different . In addition, after post hoc test (tukey HSD), it shows that : - Medium roast coffees demonstrated the highest average scores. - Dark roast coffees showed significantly lower scores compared to medium roasts

PRE-PROCESSING

- **Outlier Detection**
  - Define the price over 1.5* IQR as outliers
  - Deal with outliers

- **Feature Scaling**
  - Normalization the features as Standard Scales

- **Splitting dataset**
  - Splitting dataset for training / testing dataset(80/20)

MACHINE LEARNING METHODOLOGY

Modeling:

- **Baseline Models**: Ridge Regression(L2 regression), Decision Tree
- **Advanced Models**: XGBoost, Multi-Layer Perceptron (MLP)

Tuning:

GridSearchCV and optimizing key parameters such as:

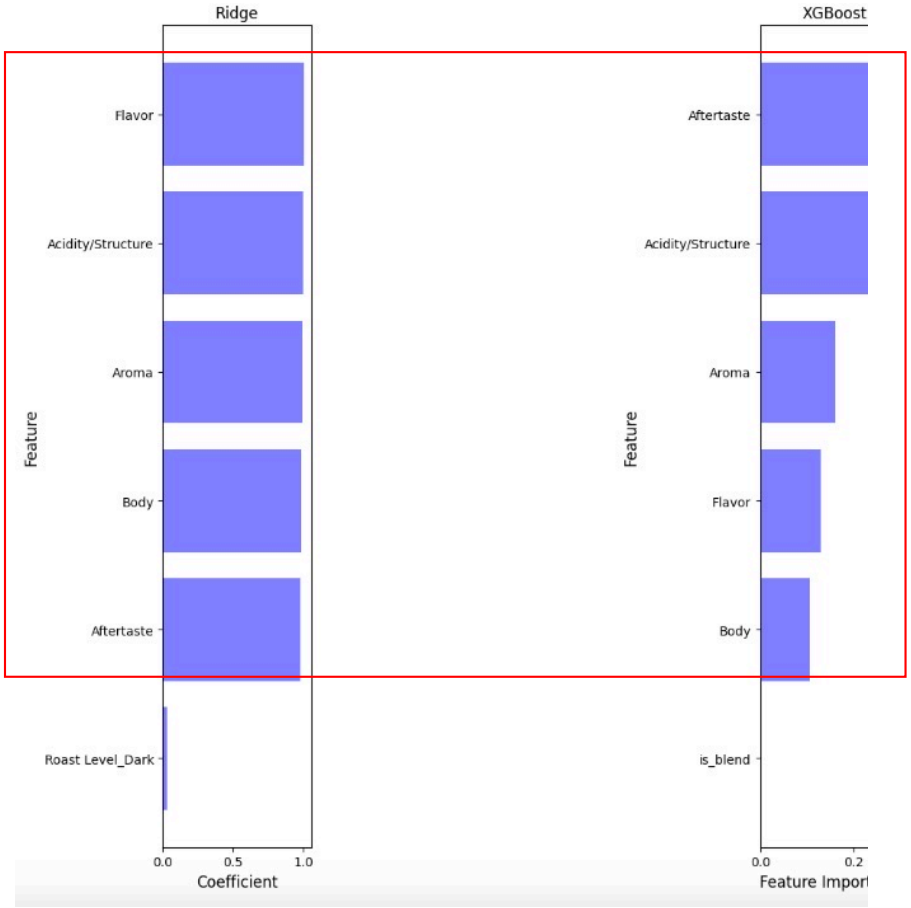
- Number of trees in Decision Forest and XGBoost
- Maximum depth of trees
- Learning rate for XGboost and MLP model

MODEL EVALUATION

Algorithm	Model accuracy score(R2)	Mean Squared Error(MSE)
Ridge linear Regression	**0.994493**	**0.012673**
Decision Tree Regressor	0.983122	0.038837
XGBoost	0.986694	0.030617
MLP	0.563319	1.004845

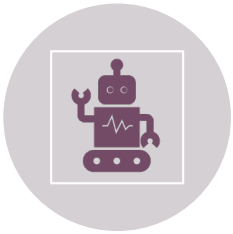
FEATURE IMPORTANT

- Compare with Ridge regression and the XGBoost model and see their feature importance. The top 5 features importance in both models are the same, though they are in a different order.



FUTURE WORK

Sentiment Analysis: Using Natural Language Processing (NLP) to analyze customer reviews as well as numerical scores.



Deep Learning Models: Exploring Convolutional Neural Networks (CNNs) to analyze coffee bean images for quality assessment.

REFERENCE

[1] Rodgers, E. (2025, February 25). *Coffee Statistics: Consumption, Preferences, & Spending*.

Drive Research. https://www.driveresearch.com/market-research-company-blog/coffee-survey/

[2] *Coffee Review - the world's leading coffee guide*. (2025, February 13). Coffee Review. https://www.coffeereview.com/

[3] *geopy*. (2023, November 23). PyPI. https://pypi.org/project/geopy/

[4] OpenWeatherMap.org. (n.d.). *Weather API - OpenWeatherMap*. https://openweathermap.org/api