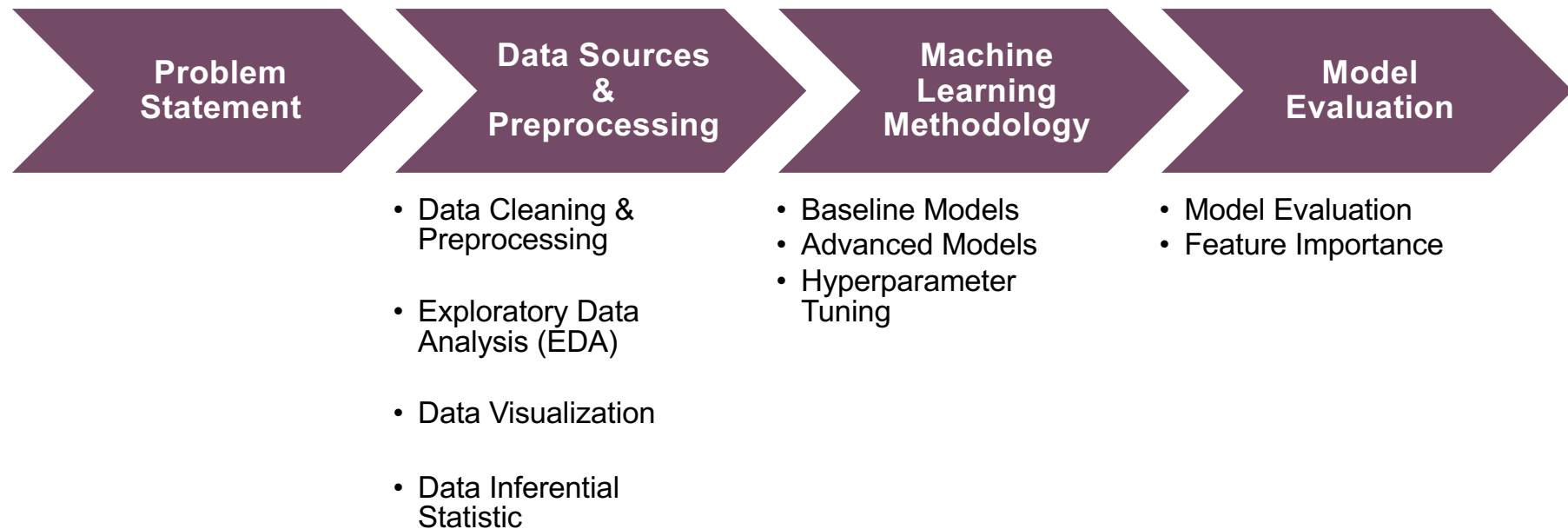

COFFEE REVIEW ANALYSIS & RATING PREDICTION

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 CLEANING & PREPROCESSING

- **Handling Missing Values**
 - Missing values were removed if necessary
- **Encoding Categorical Variables**
 - Convert Categorical
 - Encoded region data by using one-hot encoding
- **Outlier Detection**
 - Define the price over $1.5 \times \text{IQR}$ as outliers
 - Deal with outliers
- **Feature Scaling**
 - Normalization the features as Standard Scales



EXPLORATORY DATA ANALYSIS (EDA)

Data Visualization

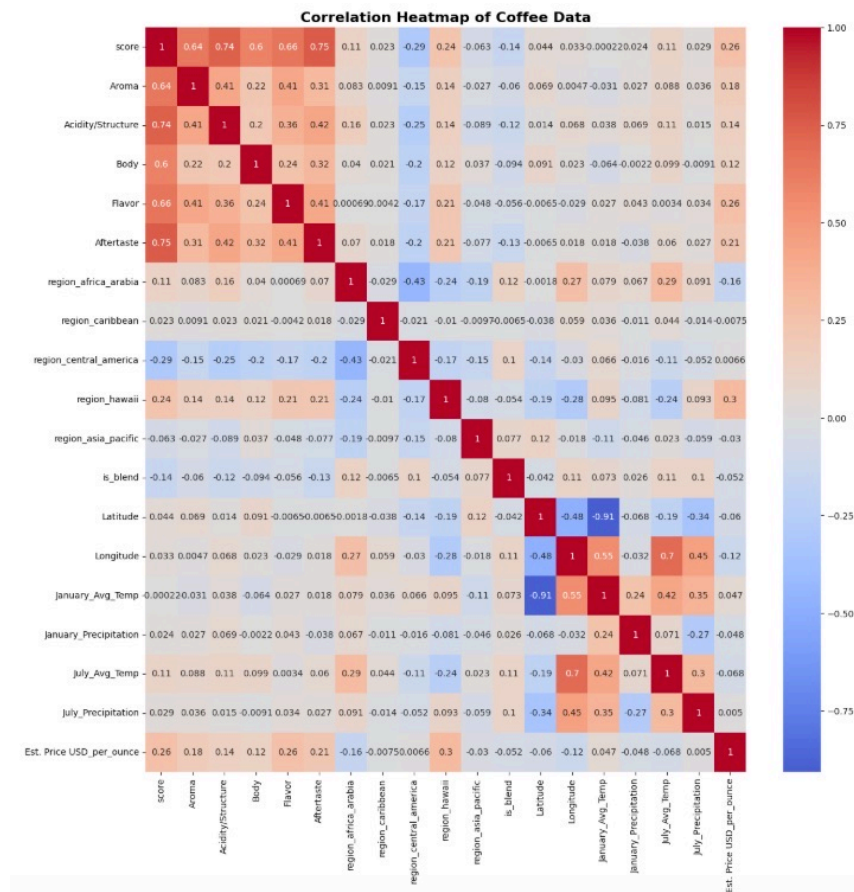
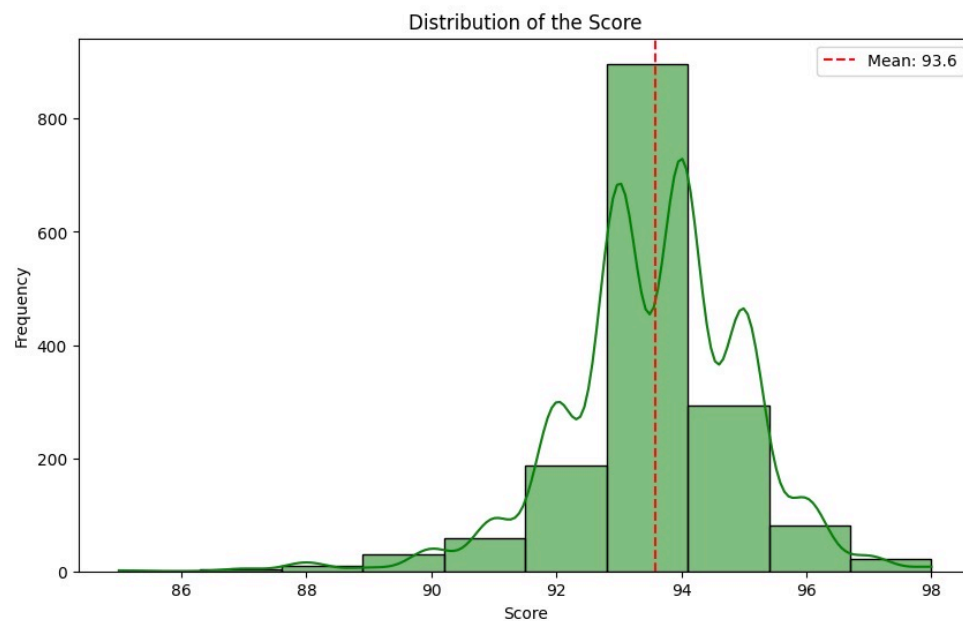
- Sensory features such as Aroma, Acidity had a strong positive correlation with overall coffee ratings.
- Coffee beans from Hawaii tended to receive higher median ratings compared to other regions.

Data Inferential Statistic

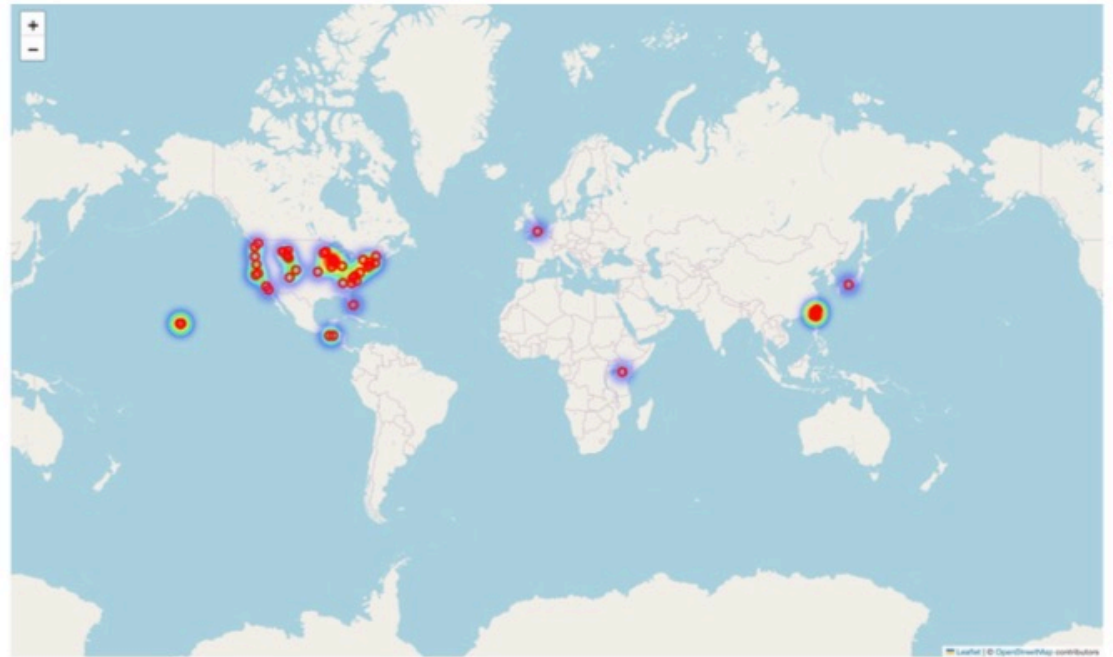
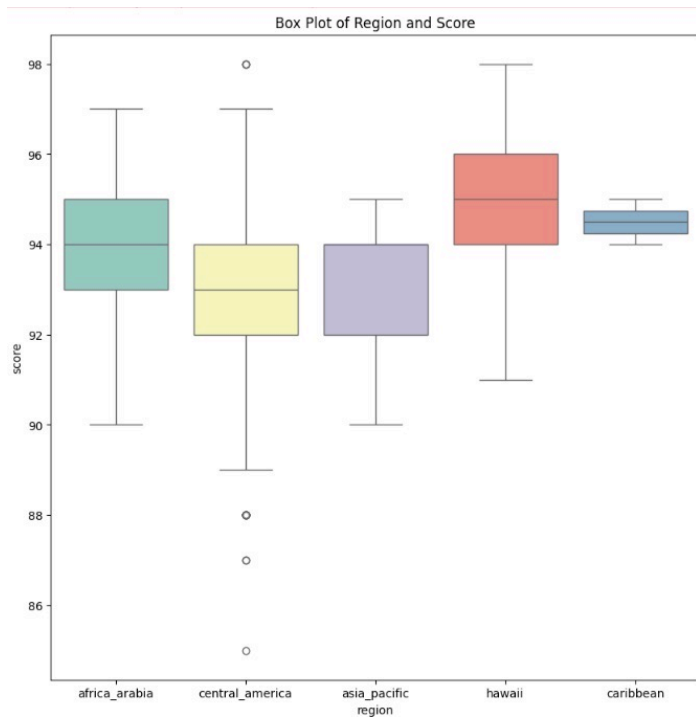
- Medium roast coffees demonstrated the highest average scores and dark roast coffees showed significantly lower scores compared to medium roasts
- blending coffee beans is significantly affects quality ratings



DATA VISUALIZATION



DATA VISUALIZATION



DATA INFERENCE STATISTICS

Statistic test	Examination	Result
ANOVA	coffee origin and the score	p-value <0.05, indicating that the geographical source of coffee beans significantly affects quality ratings.
t-test	coffee blended (Y/N) and the score (continuous numeric)	p-value <0.05, which means the analysis showed significant differences in coffee scores across is_blend or not, indicating if blending coffee beans significantly affects quality ratings.
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 difference. In addition, after post hoc test (tukey HSD), it shows that : <ul style="list-style-type: none">- Medium roast coffees demonstrated the highest average scores.- Dark roast coffees showed significantly lower scores compared to medium roasts

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:

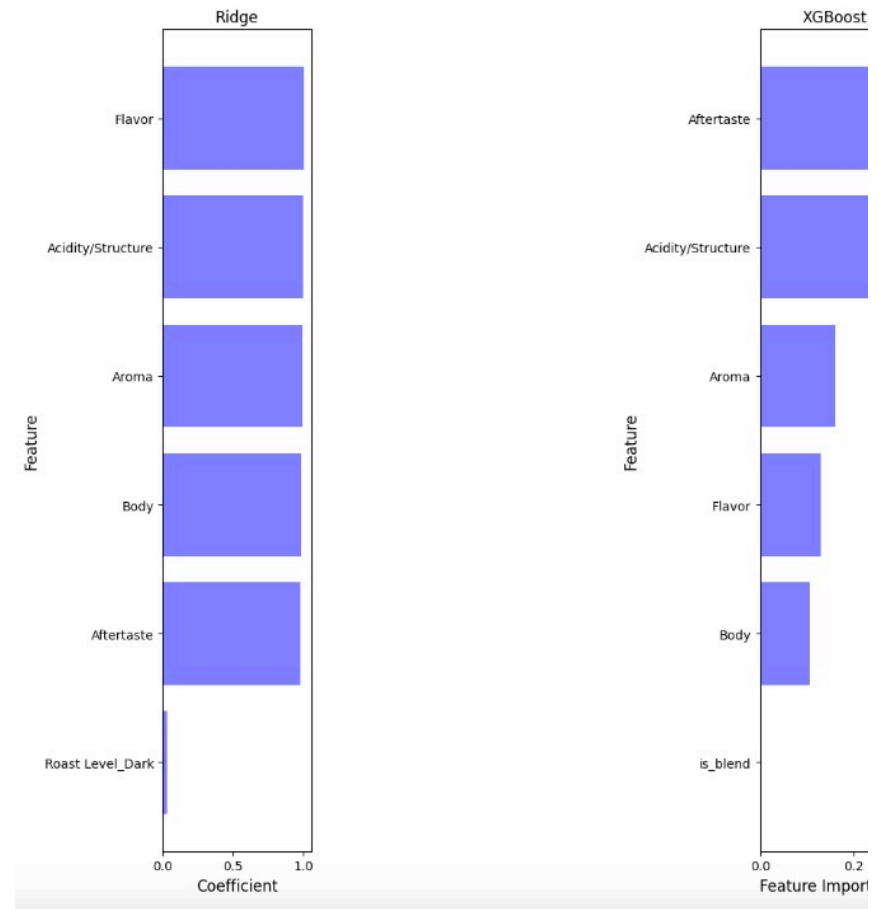
- Number of trees in Decision Forest and XGBoost
 - Maximum depth of trees
 - Learning rate for XGboost and MLP model
-

MODEL EVALUATION

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

FEATURE IMPORTANT

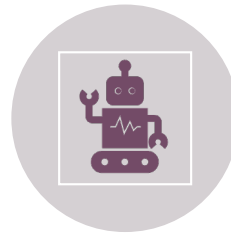
- Compare with Ridge regression and the XGBoost model and see their feature importance. The top 5 features in both models are the same, though they are in a different order.



FUTURE WORK



Sentiment Analysis: Using Natural Language Processing (NLP) to analyze customer reviews alongside numerical ratings.



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>
-