Summary of Q&A - 2025-07-08

Goal

To predict wine quality using a subset of chemical features through regression models. The analysis also explores which features contribute most to the prediction and how model performance varies across different algorithms.

Questions and Answers Summary

1. Why were only 5 features used for prediction?

To simplify the model and focus on chemical variables that are often overlooked in basic correlations, we used:

- citric acid
- density
- fixed acidity
- total sulfur dioxide
- free sulfur dioxide

This reduced-feature setup helps analyze how well basic physicochemical properties predict sensory quality.

2. Why use log transformation on the target variable?

The wine quality variable is skewed and not normally distributed. By applying np.log() to the target during training (and np.exp() after prediction), the model:

- Handles variance more smoothly
- Produces more stable and accurate outputs, especially for tree-based models

3. Why did Random Forest perform best?

Model Performance Comparison:

Linear Regression:	MSE = 0.555, R^2 = 0.155
Random Forest:	MSE = 0.450, R^2 = 0.320
XGBoost:	MSE = 0.498, R^2 = 0.247

Random Forest worked best because:

- It captures nonlinear interactions between features like density and sulfur dioxide
- It is less sensitive to scaling and outliers than linear models

4. Why did XGBoost underperform compared to Random Forest?

Although XGBoost is a more advanced model:

- It wasn't hyperparameter-tuned in this experiment
- The input features were limited and may not have exposed its full potential
- Random Forest, by contrast, is robust even with minimal tuning

5. How to visualize model performance?

Model MSE and R^2 were compared using side-by-side bar charts. This clearly illustrated that Random Forest had the best trade-off between complexity and accuracy.

6. How was feature importance extracted?

From the fitted Random Forest:

```
  importances = rf_model.feature_importances_
```

This showed that density and total sulfur dioxide contributed most to predictions, even more than citric acid.

Final Tip

Even with only 5 features and basic preprocessing, Random Forest delivered meaningful insights. This highlights the importance of starting simple, iterating, and evaluating different models step by step.

References

- UCI Wine Quality Dataset: https://archive.ics.uci.edu/ml/datasets/wine+quality
- Scikit-learn RandomForestRegressor Docs
- XGBoost Documentation
- Matplotlib Bar Charts

Tools Used

- Python (pandas, numpy, matplotlib, seaborn, scikit-learn, xgboost)
- Jupyter Notebook
- Regression models: Linear, Random Forest, XGBoost
- Evaluation: Cross-validation (KFold), MSE, R^2
- Visualization: Correlation heatmap, bar plots for importance & comparison