Executive Summary

Data Set and Problem

- Goal is to predict the cuisine type of a recipe based on the list of ingredients
- Data set has ingredient lists and cuisine types for various dishes. It is from an old Kaggle competition https://www.kaggle.com/competitions/whats-cooking. About Kaggle's provided training data:
    - The data is in a json file and contains a table with three columns: id, cuisine and ingredients.
    - The cuisine column contains strings. There are 20 different cuisines of which italian is listed the most (7,838 times) and brazilian is listed the least (467 times).
    - The ingredients column contains lists of strings. Specifically, each entry is a list of ingredients for one recipe. In total, there are 6,714 distinct ingredients.
    - The ingredients contain lowercase and uppercase letters, misspelled words, different spellings of the same word, special characters (e.g. the trademark symbol), brand names, and extra descriptors (e.g. drain and flake).
    - There are 39,774 rows.
- We created two cleaned training data sets: key_words and train_trimmed.

Key Performance Indicators

- Overall Accuracy

Planned Modeling Approach

- Use stratified k-fold cross-validation with 5 splits.

- Look at multiple models:
    - k-nearest Neighbors
    - Logistic Regression
    - Decision Trees and Random Forests
    - Linear/Quadratic Discriminant Analysis and Naive Bayes
    - Support Vector Machines
- Compare the models' performances on the two cleaned data sets.

Results and Conclusions

- Accuracies for the two best performing models:
    - Linear Support Vector Machines - train_trimmed data set: 74.9%
    - Logistic Regression - train_trimmed data set: 74.8%
- Project Recommendations:
    - Additional cleaning of the data set (i.e. new approaches).
    - Using more performance indicators (e.g. precision).
    - Testing the models on new data.