Capstone Project Milestone Report:

*Prediction of Cuisines from given set of ingredients*

# Introduction

When we go for lunch or dinner we would sense aromas of different dishes. It always intrigued me how change in ingredients would affect the aroma of a dish. Especially coming from India, I have observed a small variation in ingredients can change the aromas drastically.

I have chosen this interesting project from Kaggle which is also on similar lines to predict cuisines based on given set of the ingredients.

# Data Set

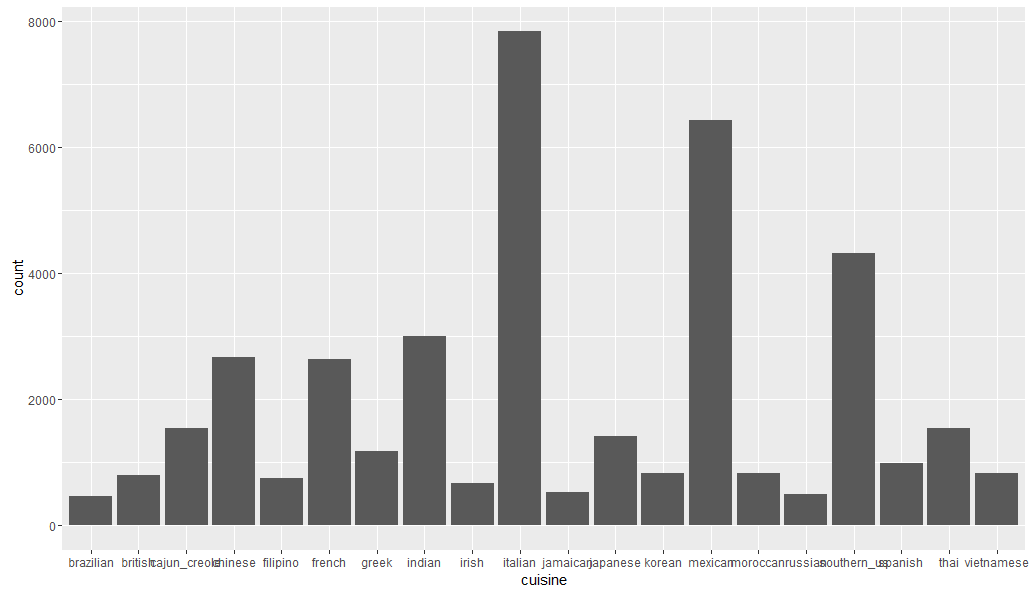
The data set used in this study is from “What's Cooking?” Kaggle Data competition <https://www.kaggle.com/c/whats-cooking/data>

Two datasets are given for the study, train and test. Train dataset for creating the model and test set for predicting the cuisine. Each dataset includes the recipe id, the type of cuisine, and the list of ingredients of each recipe (of variable length). The data is stored in JSON format. An example of JSON node is given below

{  
 "**id**": 24717,  
 "**cuisine**": "indian",  
 "**ingredients**": [  
 "tumeric",  
 "vegetable stock",  
 "tomatoes",  
 "garam masala",  
 "naan",  
 "red lentils",  
 "red chili peppers",  
 "onions",  
 "spinach",  
 "sweet potatoes"  
 ]  
 },

# Preliminary findings

Preliminary investigations showed that in the given train dataset has lots of Italian cuisines followed by Mexican and southern\_us



Also ingredients data doesn’t look clean, words are identical but not same. For example, plural forms and lower /upper case differences as shown below.

|  |
| --- |
| braeburn apple |
| Braeburn Apple |
| Bragg Liquid Aminos |
| braggs liquid aminos |

# Analysis Plan

Based on the findings described above, the analysis plan is as follows:

1. Clean the ingredients data.
2. Do feature engineering to understand the ingredients impact on cuisines.
3. Perform Cart analysis.
4. Select the best model
5. Write and post final report