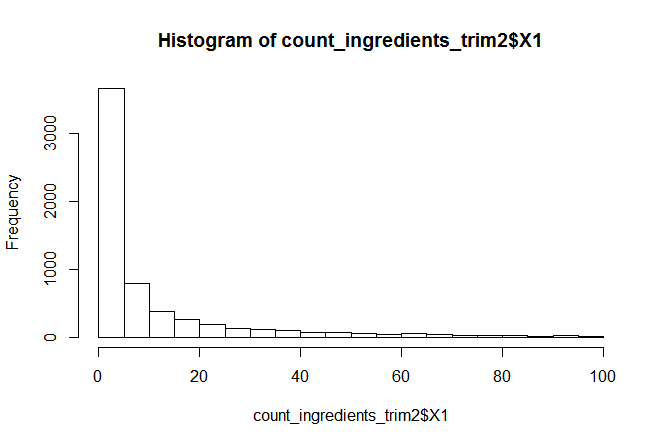
Data Science 450 Final Project – Part I – Define objective

Adrian Guerin

Michael Patterson

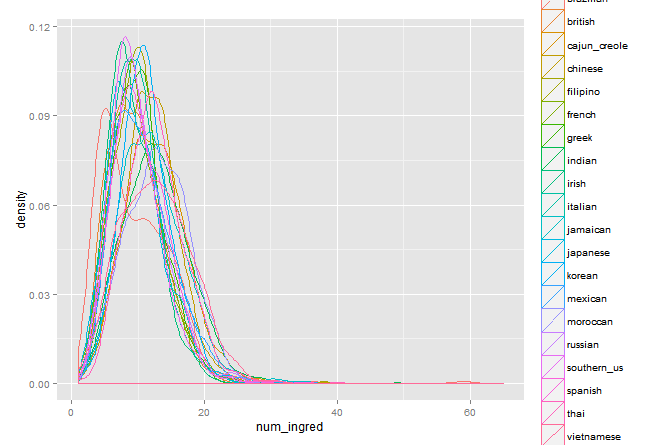
We have chosen to work on the Kaggle recipe categorization challenge: <https://www.kaggle.com/c/whats-cooking>. In this challenge, we are given a JSON containing 39,000 recipes from twenty different cuisines. Each recipe contains a list of ingredients. The goal of the challenge is to categorize recipes based on their ingredients; as such it is a supervised learning problem.

The data is highly structured. Each row contains a cuisine, and then a list of ingredients in the cuisine. The ingredients are stored as a list of strings. There are approximately 100 duplicate recipes that contain the same cuisine and ingredients list, which we will likely remove. There are ~6,700 different ingredients in the dataset. We plotted a histogram of ingredient frequency (that is, how many times an ingredient shows up in different recipes):

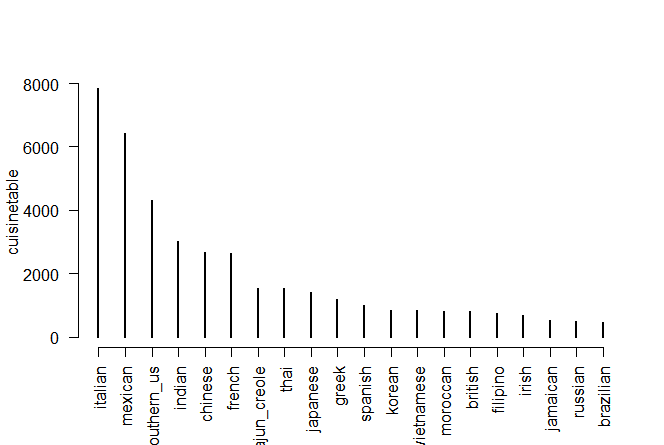


Almost half the ingredients are used in < 5 recipes. The distribution is highly skewed, with a small number of ingredients being used in dozens to hundreds of recipes. The most commonly used ingredients were: sugar (6000 recipes), garlic, water, olive oil, onions, and salt (18,000 recipes).

We also plotted a histogram of how many ingredients are in each recipe, for each cuisine:



There was no appreciable difference in the number of ingredients per cuisine. Finally, we plotted how often each cuisine occurs in the dataset:



Italian and Mexican were the most common recipes.

Now that we have a handle on the basics of the dataset we are engineering features. An easy feature is the number of ingredients in the recipe. We have also created a large binary indicator matrix for each ingredient in each recipe, although this may be too many features, and lead to overfitting. We will also create features for whether recipes contain spices, are vegetarian, and use some sort of NLP information criterion to rank features. It is an open question to us right now whether common features will be more informative than rare features: while rare features may only be in certain recipes, they occur infrequently, and thus may not be useful.