Problem 1: Principal Component Analysis

A Kaggle user has created a dataset with over 20,000 recipes from the website Epicurious. Loosely speaking, there are a few groups of variables in the dataset:

- The outcome of interest (`cake`, whether a recipe is tagged as a cake)
- The nutritional variables (`calories`, `protein`, `fat`, `sodium`)
- Ingredient tags (`almond`, `amaretto`, `anchovy`, and so on)
- Place tags (`alabama`, `alaska`, `aspen`, `australia`, and so forth)
- Other tags (`advance.prep.required`, `anthony.bourdain`, etc.)

You may want to carry out your own exploratory analysis to a) become familiar with the dataset, and b) carry out some pre-processing. You will have to make choices during pre-processing–these choices are up to you. I am happy to offer suggestions, but as long as you feel that you have made a reasonable choice, you will not be penalized.

Load the dataset

```
import pandas as pd
epi = pd.read_csv("../../Data/epi_r.csv", index_col = 'title')
```

Once you have completed the pre-processing…

(a) Perform Principal Component Analysis (a.k.a., singular value decomposition) on all the features of the dataset.

(b) Create a scatter plot with the first component on the x-axis and the second component on the y-axis. Label some of the points with the variable names. Are you seeing any patterns in your scatterplot?

(c) Find the fraction of variance explained by each of the principal components, as well as the cumulative sum of the fraction of variance explained by each of the principal components. What fraction of the variance has been explained by PC1 and PC2? How many principal components would you pick?

Problem 2: ℓ_1-Regularized Logistic Regression

(a) Using all of the principal components as features, fit an ℓ_1-regularized logistic regression for several different values of the penalty/regularization parameter predicting whether or not a recipe is for cake.

(b) Find the optimum regularization parameter by maximizing AUC on the test set.

(c) Calculate the test set AUC for your classifier.

(d) Plot the feature coefficients as a function of the logarithm of the regularization parameter `C`.

Problem 3: K-Means Clustering

Let's move away from trying to predict whether a recipe is cake or not and simply try to identify patterns/clusters of recipes.

(a) Apply K-means clustering for a range of K to all of the Epicurious features.

(b) Make a plot of within cluster variance as a function of K, and choose the best K.

(c) Using the K that you found in the previous step, produce a visualization that explains how some of the clusters differ using the following steps:

- So that you can visualize the clusters in two dimensions at a time, perform Principal Component Analysis (a.k.a., singular value decomposition). This should be very similar to your analysis from Problem 1 (but it will include `"cake"` this time).
- Create scatter plots of the first two principal components, PC1 and PC2. Label the axes with the fraction of the variance explained by PC1 and PC2.
- Color each point according to the cluster labels.
- Repeat the previous two steps for the following pairs of components: PC3 and PC4, PC5 and PC6, PC7 and PC8, and PC9 and PC10. You should not have to re-perform PCA on the features.

(d) Try to come up with a "name" or description for at least a few of the clusters–are they easily interpretable?