Food Analysis

Robert Petit

May, 2022

Abstract

For this project, we will be attempting to suggest pairings with, replacements for, and additions to common ingredients in food based on the observed flavor compounds in the ingredients. To do this, we scrape a large set of ingredients and their associated chemical compounds from FlavorDB and investigate the cross-ingredient similarities through both clustering and principal analysis. We find success in pairing ingredients with the results of clustering, while principal component analysis provides colorful, if obscure, methods for comparing and contrasting ingredients for subjective interpretations.

Introduction

For a brief moment, consider deeply the process of cooking and eating a slice of bacon. There are a number of sensations that come to mind: the thin layer of fat that inevitable gets on your hand taking it out of the package followed by the sounds of it crackling when it hits the pan. Soon after, the kitchen is filled with the smell of maple and pork and you take the first little too-hot nibble. A small crunch followed immediately by saltiness then, slowly, the hearty pork and slightly-sweet maple fill your mouth.

Taste is one of the most complicated senses humans possess; partly due to the volume of information processed (though in that regards, no sense will compare to sight) and partly due to the wide variety of information considered. This variety of information is one of the most confounding parts of our food preferences. We can break it out into several parts, the simplest of which are texture, sight, sound, and memory of the food. Slightly more complicated are the distinctions between taste and flavor. Tastes will include the five very specific "core" of food: sweet, sour, salty, bitter, and umami. Flavor will include nearly everything else we experience after putting food in our mouth and is vastly more complicated than taste; typically, anything that you can link directly with a smell will be a taste, since the olfactory system processes both smell and taste. This distinction between taste and flavor is somewhat elusive in some cases: the physical sensation associated with sour and salty foods is obvious, but it may be more challenging to try and separately consider the umami and flavor in soy sauce, or to seperate the sour from the flavor in a lemon.

This project hones in on flavor specifically. There are a huge number of compounds that can contribute to flavor, each of which will have it's own flavor profile and personality. Typically, these are only given very basic consideration; either foods are looked at exclusively on a per-compound basis or ingredients are simply matched by the number of commonality they share with other individual ingredients. This work suffers extensively from the limited nature of the data, in particular the difference in observed molecules between different types of food, a topic which we briefly explore in the data. We will extend this by working with data on the compounds that are in a wide variety of ingredients and associating ingredients with one another in two ways: clustering on shared compounds and observing the principle components of foods and deciphering their various nuances.

Methods

The Data

The data set we will draw from is a list of observed flavor molecules in each of 707 recognizable ingredients, such as rye bread, garlic, and tangerines. There are 1772 tracked across these ingredients, with a huge number of overlap between various foods. Both of these totals are after all transformations and simplifications discussed in this section. We are not working with densities or total lists of all compounds, simply a present/not present list for 1772 compounds across all ingredients. There is some small heterogeneity in the tests performed, not all ingredients were tested for all molecules, but the vast majority of molecules have been correctly identified.

This data was scraped from FlavorDB's entity information which gives a large overview of any given ingredient and then cleaned to get the 'common' name (common to chemists, at least) of each flavor molecule it contains. We also build a complimentary data set that provides a basic flavor profile associated with each compound which, in turn, can be used to interpret the results of the clusters and principal components.

A quick glance at the counts of compounds (see appendix) in each ingredient shows that there are huge number of virtually unidentified ingredients. Many of these have just one or two compounds identified; since this is very uninformative, we drop all objects with fewer than 3 components. Since we don't make any inference that depends on a representative sample, this should not introduce any biases that we are worried about. Most of the dropped ingredients are things like walrus meat, which are not flavors that we can typically associate with anything else.

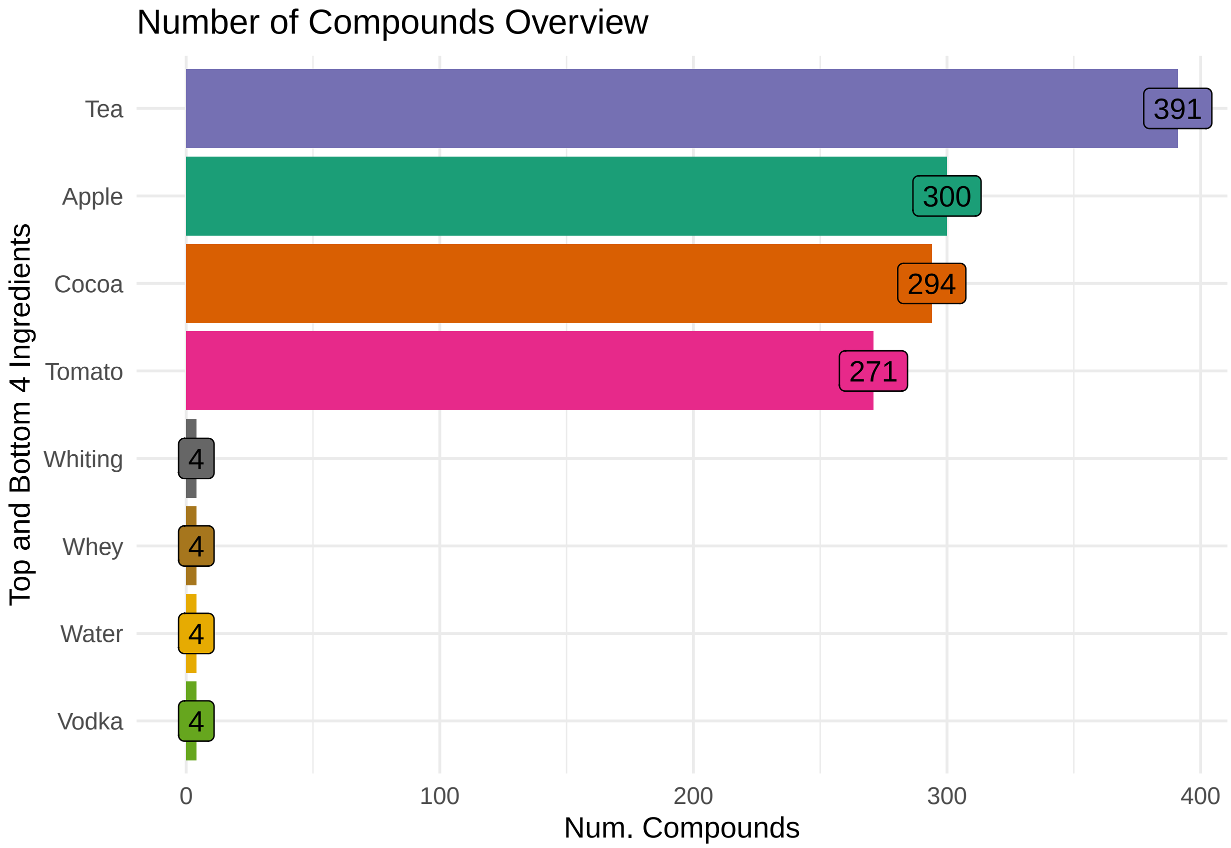

Figure 1: The 4 most and least compound-dense ingredients

As you can see in figure 1, this data has some important nuances to it that become apparent at a quick glance of the most and least compound-dense ingredients in the set. First, it is not a complete record of all compounds but rather a list of ‘common’ compounds in each ingredient. Because of this, ingredients that are more commonly consumed will likely have more compounds identified in them. There are two extremes to this: at low end very esoteric ingredients, which often only have a few compounds recorded whereas simpler but more common ingredients will have many more compounds identified. The other cause for the vast differences is in the somewhat arbitrary nature of how ingredients are grouped together; each variety of peach is represented only as ‘peach’ whereas there are dozens of different whisky and cheese variety. This generally does not create a huge number of problems, but for cases like tea, which is hugely varied, grouped into one stack, *and* very common, there are a disproportionate number of compounds associated with it. This causes significant issues for the ‘matching’ approach to food pairings, which matches purely on the number of similar molecules observed. Clustering in particular should provide *some* limits to that effect, since each ingredient is forced to exist in only one cluster.

On average, each ingredient will have an average of 84.0353607 molecules and a median of 93. This hints at a left-skew of the data, wherein a small number have disproportionately large numbers of compound Thankfully, these tend to occur in places where we can provide very solid subjective interpretation, such as with tea, as opposed to confusing or difficult to consider ingredients like walrus meat, which means we can still interpret and understand our data using these methods.

Additionally, there is the question of shared compound within our set. Since these are the source of our matching, extremely low or extremely high frequencies of any compound can be problematic for any clustering or component analysis.

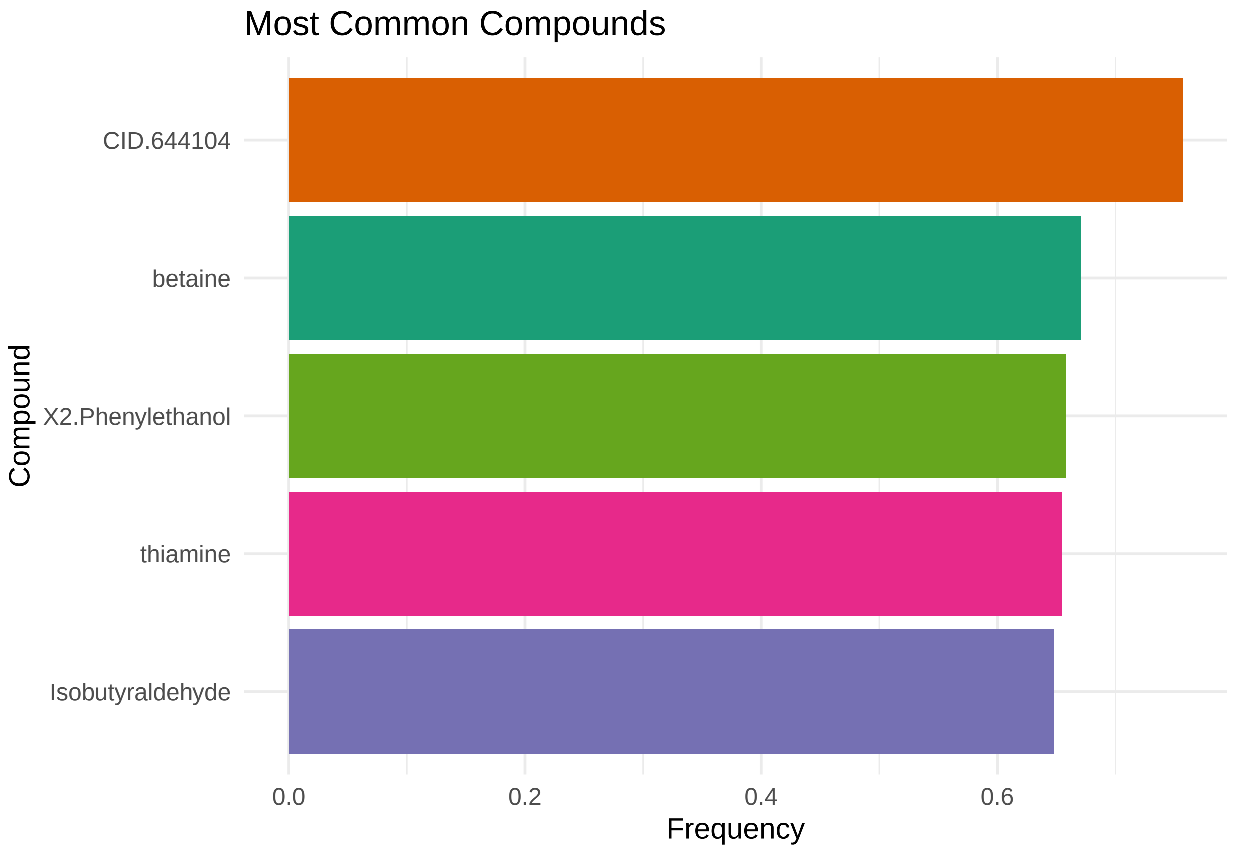

Figure 2: The most common compounds in our set of ingredients

With an average frequency of 0.047424 and a median of 0.0056577 there are certainly a number of compounds that only show up in a very select number of ingredients. Generally, since these are what we are clustering on, we are not overwhelmingly concerned with these occurrences. For some more simple understanding of the data, we can turn to our flavor profiles for each of the frequent molecules above. Each ingredient that contains these flavor compounds will include those flavors in their overall profile. Further intuition is provided by a random sample of ingredients that contain this compounds.

Compound	Flavor Profile	Sample of Foods
betaine	bland	Allspice, Banana, Tree fern,
CID.644104	grassy, very mild	Sparkleberry, Oyster mushro
Isobutyraldehyde	malt, fresh, floral, pungent, aldehydic, green	Jerusalem artichoke, Canola
thiamine	bitter	Margarine like spread, Milk
X2.Phenylethanol	spice, rose, lilac, rose dried, rose flower, bitter, floral, rose water, honey	Common persimmon, Cassa

Methods

The approach here will be twofold: first, we will cluster the data and observe the within-group characteristics for intuition. Second, we will perform basic analysis on the principal components of the ingredients, paying particular attention to ingredients that make up fairly unique branches of these components. Both of these approaches should provide some intuition beyond the matching approach since we infer the 'not present' chemicals in our data set and use those to shape both our clusters and our principal components.

The clustering method will produce single-membership sets of ingredients, where each cluster has significant commonalities with the other members of that cluster, even if they do not match exactly with a substantial number of them. The single-membership nature of the clusters will help to limit the impact of the extremely compound-dense ingredients in our data set, but comes with the trade off of not recognizing highly-versatile ingredients.

The principal component analysis approach will help to quickly identify fringe relationships; by continuing to find components based on the residual of the prior components, not only will we manage to use the information in the 'not present' observations, we will also rapidly distill information from the frequently observed variables, leaving space for the less common but contextually important compounds. The continuous nature of the principal components will allow us to calibrate our desired observations more closely, but will confound the interperetation of any single one substantially.

Results

The first task involved in clustering is to determine the optimal number of clusters. Working from the gap statistic (see appendix), using the 'lowest within 1 standard error of the first max" rule, we select a consistent K value of 14. Once we have a desired number of clusters, it is simple to turn to a k-means cluster. The sizes of the relative clusters vary substantially but they all take on non-trivial values.

Cluster Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Cluster Size	18	24	4	14	60	12	11	139	238	32	14	7	14	120

The principal component analysis is somewhat more difficult to judge in aggregate. 5 components seems to be the highest rank out of which we can determine any information. Given the massive size of the loadings and the esoteric nature of the chemical compounds, it is impossible to interpret these by way of their values. Instead, we can translate their values to simply showing the primary components that each emphasizes more of, and each emphasizes less of.

PCID	PC1	PC2
Less	alpha.L.Sorbopyranose, calcium.lactate, Farnesal, Allyl.Alcohol, alpha.Maltose	None
More	None	X1.octanol, octanoic.acid, X2.Me

Conclusion

Clustering

The clustering results are somewhat promising. To see them more clearly, we can look at the actual ingredients that appear in each cluster.

Cluster ID	ingredients
1	Tea, Cocoa, Tomato, Potato, Mushroom, Soybean, Peanut, Corn, Rice, Green Beans
2	Parmesan Cheese, Blue Cheese, Cheddar Cheese, Gruyere Cheese, Swiss Cheese, Camembert Cheese, Other C
3	Coffee, Beer, Cognac Brandy, Rum
4	Apple, Grape, Guava, Strawberry, Papaya, Mango, Passionfruit, Pineapple, Apricot, Melon
5	Raspberry, Cranberry, Onion, Cabbage, Lotus, Bilberry, Peas, Fig, Cherimoya, Olive
6	White Wine, Sherry, Red Wine, Wine, Port Wine, Sparkling Wine, Rose Wine, Botrytized Wine, Strawberry
7	Plum Brandy, Apple Brandy, Pear Brandy, Weinbrand Brandy, Cherry Brandy, Raspberry Brandy, Anise Bra
8	Honey, Wholewheat Bread, Vinegar, Mutton, Crab, Popcorn, Sake, Crayfish, Yogurt, Rye Bread
9	Durian, Safflower, Flaxseed, Kumquat, Prickly Pear, Chive, Fenugreek, Artemisia, Lemon Grass, Okra
10	Peppermint, Capsicum, Dill, Coriander, Carrot, Chamomile, Cinnamon, Clove, Lemon Balm, Thyme
11	Ginger, Orange, Spearmint, Pepper, Celery, Black Currant, Basil, Laurel, Rosemary, Lemon, Oregano
12	Malt Whisky, Scotch Whisky, Bourbon Whisky, Finnish Whisky, Japanese Whisky, Whisky, Canadian Whisk
13	Beef Processed, Pork, Butter, Chicken, Milk, Hops Oil, Malt, Filbert, Beef, Kidney Beans
14	Cider, Black Tea, Drumstick Leaf, Bonito, Colocasia, Citrus Peel Oil, Mountain Papaya, Kenaf, Mate, Bartle

Some of these clusters serve as important sanity checks for our results. Cluster 2, which contains 10 different types of cheese, is a good indicator that we are appropriately clustering by ingredients that would 'go well' together in the strictest sense. Cluster 6 provides the same information for wine, 7 pulls the brandy, 12 grabs all whisky, and 13 contains the high fat-content meats and oils.

Cluster 3 is notable for it's minuscule size; it is the smallest cluster and has grabbed four drinks which are all commonly mixed together, though coffee and beer is a somewhat divisive combination. Cluster 4 is a medley of fruits, and it is hard to consider any combination of fruit in that list that would be anything less than delicious. Less-obvious clusters that are still not quite surprising are the herbs that make up clusters 9, 10, and 11. While these are expected to be close to one another, some pairings suggested here may surprise nonetheless: Orange and pepper in particular is a pairing that, while strange to most western palettes, is common in Indian and African ingredients, whereas peppermint and capsicum, a spicy pepper, would quite likely be off putting if used together.

This leaves us with clusters 1, 5, 8, and 14. A close look at cluster 1 reveals a few curiosities; tea and tomato in particular is somewhat baffling. At first glance, the inclusion of cocoa may also seem strange, but it is important to consider that this is cocoa, not chocolate; in the U.S., cocoa and sugar are almost unanimous, but cocoa has a wide variety of applications, and can be an important source of bitterness and depth, Indian and Japanese cuisine in particular tend to pair cocoa with spice-heavy dishes, especially curry.

Cluster 5 is entirely plant-based foods, many of which are naturally sweet (with the exception of olives) though some, like onions, take a decent amount of cooking to yield their sweetness in a way that we can taste. Cluster 14 also contains a number of naturally sweet entries like cider, papaya, and Bartlett pears. The most confusing entry here, and perhaps in the entire result, is the inclusion of Bonito, which is a form of smoked and fermented tuna flakes common (and nearly exclusive to) Chinese and Japanese cuisine primarily used as a source of umami.

Last, and most interesting of all, is cluster 8. The initial three entries make sense, especially when you consider vinegar as a very diverse byproduct of wine. The other seven entries are odd at first glance, but crab is often paired with slightly sweet sauces and mixtures; Sake, an alcoholic beverage from Japan, is commonly paired with seafood such as crab and crayfish; and yogurt contains many of the same acids as vinegar. Out of the entire selection of ingredients from the clusters, it seems that cluster 8 has the most uncommon but potentially promising pairings.

Principal Component Analysis

The interpretation of the principal component analysis is much more limited than that of the clustering. The most sensible view is to look at the foods and flavors that make up the positive and negative directions for each principal component.

PC	Less Flavor
1	bitter, odorless, floral, minty, mustard, pungent, cherry, almond, balsam, balsamic
2	None
3	harsh, turpentine, minty, fresh, earthy, sweet, woody, terpene, pine, camphor, resinous, resin, green, hay, dry, lemon,
4	butterscotch, butter, tart, sharp, fruit, fruity, buttery, apple, ethereal, banana, bitter, sweet, tropical, spice, vanilla, s
5	sour, bitter, mild, caramel, spice, burnt sugar, maple, almond, butter, green bean, vegetable, earthy, beany, fruity, m

These are somewhat difficult to interpret; since the flavor palettes are not mutually exclusive, we often will see the same flavors in both the "more" and "less" columns. Additionally, it is not quite so straightforward as simply separating the ingredients from one another, in many cases (such as PC4 and PC6) we actually see the two extremes commonly paired together whereas other identified components (such as PC9) simply clash with one another.

One possible interpretation of these PCs is more as a comparing/contrasting axis than a set of pair/don't pair guidelines. By consulting these principal components, you can take a dish-wide approach and pick items that are on similar or opposite scales as you avoid or seek out contrasting flavors. This is not a blanket relationship, but rather a series of suggestions that may prove useful.

Results Comparison

Overall, the approach seems to have produced a number of pairings that are somewhat intuitive while providing new pairings that are uncommon and potentially fruitful; to fully test these will require an extraordinary amount of cooking, but there do seem to be some improvements, if only in select cases, to the typical method of simply matching ingredients by the most number of similarities. The main limitations of this method come from the data; the grouping of the observations causes significant, arbitrary loss of accuracy and the inconsistent nature of the compounds being observed may unfairly skew the data toward established combinations and comparisons.

Apendix

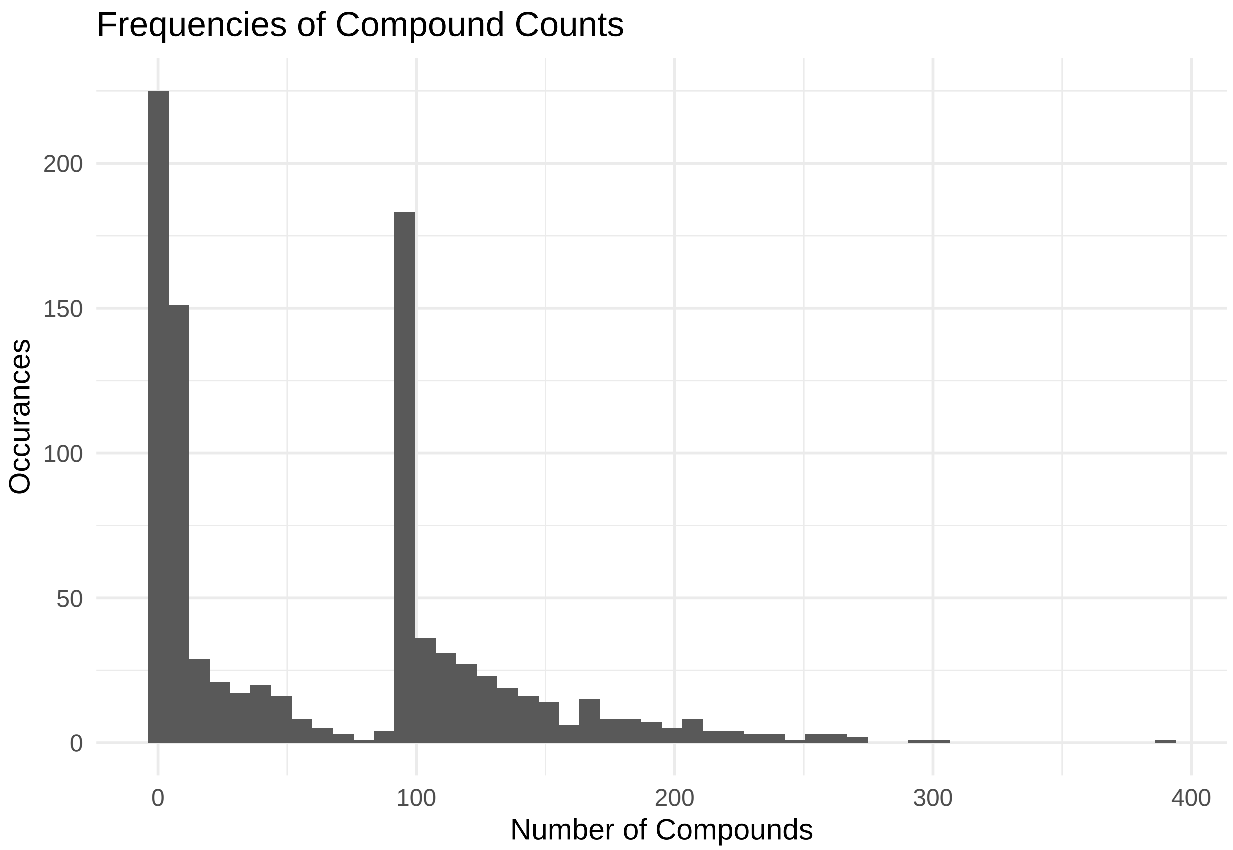

Figure 3: Figure 1: Compound Count Frequencies

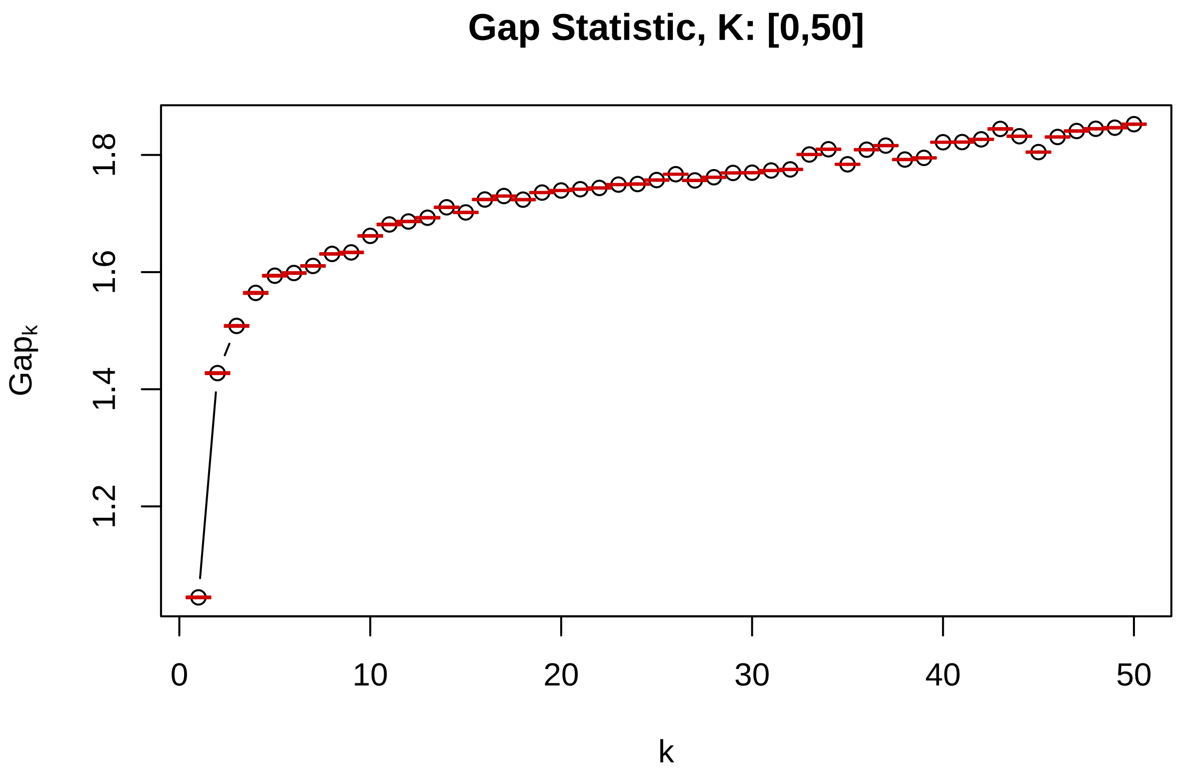

Figure 4: Gap statistics for each K value between 0 and 50