In [14]:
```
import pandas as pd
import numpy as np
```

In [2]:
```
df = pd.read_csv('recipes_w_search_terms.csv')
```

In [26]:
```
df.tail()
```

Out[26]:

	id	name	description	ingredients	ingredients_raw_str	s
494958	276465	Blackberry Orange Scones	The orange zest makes for a flavorful, not ove...	['unbleached flour', 'baking soda', 'butter', ...	["2 1/2 cups unbleached flour","2 teaspo...	
494959	257796	Slow Cooker Garlic Chicken With Rosemary	Delicious and easy!	['roasting chickens', 'lemons', 'rosemary spri...	["1 (5 lb) roasting chickens, rinsed and ...	
494960	78003	Pot Roast with Port (Stove Top)	This is a recipe from the Frugal Gourmet cooki...	['boneless beef chuck roast', 'olive oil', 'ta...	["2 -3 lbs boneless beef chuck roast","2 ...	
494961	328810	Kapusta (Cabbage and Kielbasa)	Cabbage and sausage in tomato sauce	['cabbage', 'condensed tomato soup', 'kielbasa...	["8 cups cabbage or 2 heads cabbage,...	
494962	9116	Yellow or Zucchini Squash Pie	This recipe was given to my mom from a lady wh...	['zucchini', 'onion', 'butter', 'oregano', 'ba...	["4 cups zucchini (can mix the squash if ...	

State your main research question

What patterns can I find in the ingredients of the recipes? Can I connect ingredients with tags or search terms? Can I generate good names for the recipes based on the other variables? What patterns might cluster analysis reveal about the different types of recipes and cuisines?

Brief summary of where your data came from

I got my data from a public Kaggle dataset of recipes collected from Food.com, one of the biggest recipe sites. Some work to clean the data has already been done, such as extracting the name of the ingredient from the raw list. The kaggle page (https://www.kaggle.com/datasets/shuyangli94/foodcom-recipes-with-search-terms-and-tags/data) doesn't describe the legality specifically, but they mention some studies that have used this dataset, and it seems to be okay to use.

Explanation/description (in words) of all the variables in your data (italicized are targets.)

- *Name*: a string which was the title of the recipe. It is of interest for text analysis and generation.
- Description: the string description proveded for the recipe. This could be of interest for text analysis and alsocould help to categorize the data better.
- Ingredients: this is the variable I'm the most excited about- I want to use the ingredients list to cluster recipes and predict their tags.
- Ingredients_raw_str: I may or may not even use this variable, it is the non-cleaned version of the ingredients. It has quantities and instructions.
- Serving_size: this variable has weight in grams for 1 serving of the recipe.
- Servings: the number of servings a recipe makes.
- Steps: plain text directions as an array of strings
- *Tags*: user created tags that describe the recipe. I want to try and predict these dags.
- *Search_terms*: these are values that would return the recipe if you searched them on the site. It could be useful to try and predict these as well.

Summary statistics for all variables

For numeric variables include: sample size, mean, standard deviation, and 5 number summary (min, q1, q2, q3, max)

For categorical variables include: sample size, category counts

Most of my variables are text-based, so it might take some improvization for this section.

In [8]:
```
print("Total number of observations:")
df.shape[0]
```

```
Total number of observations:
```

Out[8]: 494963

The name and description are text fields. Every recipe has a name, but some recipes have no description.

In [11]:
```
print("Sample size of descriptions:")
df["description"].dropna().shape[0]
```

```
Sample size of descriptions:
```

Out[11]: 485362

In [25]:
```
print("Description of the number of ingredients:")
num_ingredients = df["ingredients"].apply(lambda x: len(x))

num_ingredients.describe()
```

```
Description of the number of ingredients:
```

Out[25]:
```
count    494963.000000
mean        144.185139
std          64.664102
min           2.000000
25%          98.000000
50%         135.000000
75%         179.000000
max         843.000000
Name: ingredients, dtype: float64
```

In [35]:
```
print("Description of the serving size")
serving_size = df["serving_size"].apply(lambda x: int(x[3:-2]))
serving_size.describe()
```

```
Description of the serving size
```

Out[35]:
```
count    4.949630e+05
mean     3.750634e+02
std      2.702044e+03
min     -4.750000e+02
25%      1.220000e+02
50%      2.190000e+02
75%      3.810000e+02
max      1.595816e+06
Name: serving_size, dtype: float64
```

In [40]:
```
print("Description of the number of servings:")
df["servings"].describe()
```

```
Description of the number of servings:
```

Out[40]:
```
count    494963.000000
mean          7.063164
std          94.677417
min           1.000000
25%           4.000000
50%           4.000000
75%           8.000000
max       32767.000000
Name: servings, dtype: float64
```

In [39]:
```
print("The max number of servings is from a recipe for whale mea
df.iloc[df["servings"].idxmax()]
```

Description of the number of servings:

Out[39]:
```
id                                                          72549
name                                                  Alaskan Blue Stew
description           I copied this recipe off the wall of Sourdough...
ingredients           ['whale meat', 'unbleached flour', 'olive oil'...
ingredients_raw_str   ["1 (242000   lb)   blue whale meat, boned and...
serving_size                                                1 (199 g)
servings                                                        32767
steps                 ['Cut whale in bite size pieces (including blu...
tags                  ['weeknight', 'time-to-make', 'course', 'main-...
search_terms                                          {'stew', 'dinner'}
Name: 51114, dtype: object
```

In [41]:
```
print("Summary of number of steps")
df["steps"].apply(lambda x: len(x)).describe()
```

Summary of number of steps

Out[41]:
```
count    494963.000000
mean        598.236620
std         428.468252
min           2.000000
25%         320.000000
50%         501.000000
75%         757.000000
max       12688.000000
Name: steps, dtype: float64
```

In [48]:
```
print("this is an outlier for number of steps, it isn't written
df.iloc[df["steps"].apply(lambda x: len(x)).idxmax()]["steps"]
```

this is an outlier for number of steps, it isn't written in the usual format.

Out[48]: '[\'First of all: these are not typical directions, but you nee
d to know about needed equipment before attempting this cake. H
ere it is:\', \'8-inch round cake pan, at least 2 inches high.
\', \'8-inch round cake pan with removable bottom or 8-inch spr
ingform pan.\', \'untreated heavy-duty jelly-roll pans.\', \'ru
bber spatula, offset spatula, and flexible 8-inch metal icing s
patula.\', \'decorating turntable, lazy Susan, or inverted roun
d cake pan.\', \'ridged plastic shelf liner, freezer paper, or
055 Mylar (I used the plastic shelf liner).\', \'parchment pape
r and waxed paper.\', \'MAKING THE CAKE:\', \'Position a rack i
n the lower third of the oven or just below the center of the o
ven and preheat the oven to 350°F Fit the bottom of an 8-inch r
ound cake pan, one at least 2 inches high, with parchment paper
and set aside.\', "Pour the clarified butter into a 1-quart bow
l and stir in the vanilla extract, if you\'re using it. The but
ter must be hot when added to the batter, so either keep the bo
wl in a skillet of hot water or reheat at the last minute.",
\'Although the flour and cocoa were sifted before they were mea
sured, they need to be triple-sifted together. Sift or sieve th
e flour and cocoa together 3 times, then set sifter on a plate
or piece of waxed paper and return the dry ingredients to the s
ifter. Keep close at hand.\', \'Whisk the eggs and sugar togeth
er in a large heatproof bowl or the bowl of a heavy-duty mixer.
Set the bowl over direct heat or in a pan of barely simmering w
ater and heat the eggs, whisking constantly, until they are war
m to the touch. Remove the bowl from the heat and, working with
a heavy-duty mixer fitted with the whisk attachment (or using a
hand-held mixer), beat the eggs at high speed until they are co
ol, have tripled in volume, and hold a ribbon when the whisk is
lifted.\', \'Sift one third of the dry ingredients over the egg
s and, using a large rubber spatula, fold in gently but thoroug
hly. When the color of the batter is almost uniform, fold in th
e rest of the flour-cocoa mixture.\', \'Spoon about 1 cup of th
e batter into the hot clarified butter add fold together until
well blended. Spoon this over the batter and, using the large r
ubber spatula, gently fold into the batter.\', "Spoon the batte
r into the pan: there\'s no need to smooth the top or rap the p
an on the counter, as is sometimes done with foam-based cakes.
Bake the cake for 25-30 minutes, or until top of the cake sprin
gs back when pressed gently. Transfer the pan to a rack and let
the cake cool in the pan.", \'When the cake is completely cool,
run a small knife around the sides of the pan to release the ca
ke and unmold onto a rack; invert right side up onto a piece of
parchment paper. (The cake can be made ahead to this point, wra
pped well, and kept in the refrigerator for up to 2 days or fro
zen for up to 3 months. Thaw, still wrapped, at room temperatur
e.).\', \'PREPARING THE CHOCOLATE:\', "The chocolate is going t
o be spread and then scraped into ruffles from four baking pans

; if you don\'t have enough pans, you can make the ruffles in 2 batches. Choose heavy-duty jelly-roll pans that are neither war ped nor dented, neither nonstick non treated with special coati ngs. Keep them close at hand.", \'Melt the chocolate in a heatp roof bowl set in a skillet of barely simmering water, in the to p of a double boiler over an inch of simmering water, or in a m icrowave oven sat at medium power. Stir the chocolate regularly until it is fully melted. Smooth, and 115F to 120F (You can tes t the temperature with an instant-read thermometer or by puttin g a drop on your top lip - it should feel warm.).\', "Hold the bottom of one of the baking sheets over a burner (either gas or electric) and, moving it back and forth, heat it until it is wa rm but not hot enough to burn your fingers. Put the baking pan upside down on a flat surface and pour on about 1/3 cup of the chocolate. Use an offset spatula to spread the chocolate thinly and evenly over the bottom of the baking pan: the chocolate wil l only be about 1/16 inch thick. Refrigerate the pan for at lea st 30 minutes, or for as long as several hours, depending on yo ur schedule. (It is better to chill the pans for a long time an d let them come up to ruffling temperature - in which case they \'ll stay at temperature longer - than to catch them the moment they turn cool enough to ruffle.) Repeat with rest of the choco late and the other baking pans.", \'MAKING THE RUFFLES:\', \'To shape the ruffles, work with one baking pan of chocolate at a t ime. Remove a pan of chocolate from the refrigerator and leave it at room temperature to warm gradually until it is pliable en ough to be scraped.\', \'Place the baking pan on a counter in f ront of you, a short side braced against your body. Hold the en d of the blade of a then, flexible 8-inch metal icing spatula i n your left hand (reverse procedures if left-handed) and, with your right hand, grab the blade close to the handle. You should have 4 to 5 inches of blade exposed and available for ruffling. \', "Using the top left corner of the pan as your staring point and imagining that corner of the pan as 12 o\'clock, position y our left hand in that corner, and your right at 2 o\'clock. Pre ss the edge of the blade against the chocolate at a very shallo w angle, as if you were going to slide the spatula blade under the chocolate. Now slide the blade forward, moving your right h and down to 5 o\'clock and then pivoting the blade to the left, all the way to the edge of the pan. As your right hand is movin g down, so is your left, although not as far - your left hand w ill move down 4 to 5 inches. This is an important point - if yo u don\'t move your left hand down, you\'ll end up with tight cu rls of chocolate rather than ruffles. As you scrape and ruffle the chocolate against the blade and then make the pivot, the ch ocolate will gather against the blade -- use your left hand to pinch the chocolate so that the ruffles form a fan and the pinc hed part is a little handle. You\'ve completed one ruffle.", "A s you make each ruffle, place it on a parchment or waxed paper-

lined baking sheet and refrigerate. When the ruffles harden, yo
u can layer them between sheets of waxed paper. (Store them in
a container in the refrigerator; they\'ll keep for a few day
s.).", "Make 2 more ruffles across the top of the pan, using th
e previously Scraped area as your guide -- the left-hand corner
of chocolate will be your 12 o\'clock point and the cleaned-off
section of the pan your edge, or end point. Make the next three
ruffles just below; then turn the pan around to get to the choc
olate on the bottom and make three more. With practice -- and r
uffling takes lots of practice -- you\'ll get 9 ruffles from ea
ch pan. Don\'t worry if you get fewer at the start.", "If, as s
ometimes happens, your ruffles crack or you get rolls of chocol
ate, not ruffles, it might be because the chocolate is too cold
-- give it a few more minutes at room temperature before you tr
y again. If the chocolate melts and gets gooey next to the spat
ula, it\'s too soft and needs a minute or two more in the refri
gerator. When the temperature is just right -- smooth and pliab
le -- but you still can\'t get a nicely fanned ruffle, angle th
e blade differently as you scrape.", \'FOR THE SYRUP:\', "Bring
the water and sugar to the boil in a small saucepan, stirring t
o dissolve the sugar, and simmer for 2 to 3 minutes. Remove fro
m the heat and cool. Add 1/4 cup of the eau-de-vie. Taste the s
yrup and decide if you\'d like a little more of the liqueur; se
t aside.", \'FOR FILLING AND WRAP.\', \'Beat the creme fraiche
with the vanilla extract to soft peaks, then add 2 Tbsp of the
sugar, beating until thickened. Taste and add more sugar if you
want it, then continue to beat until the cream just begins to s
tiffen. Cover and keep refrigerated until needed.\', \'Assembli
ng the Cake -- Cut the cooled genoise into 3 even layers with a
long serrated knife. Fit one layer into the bottom of a high-si
ded 8-inch round cake pan with a removable bottom or an 8-inch
springform pan and brush the layer with syrup.\', \'Place the c
hopped chocolate in a small bowl and whisk in the boiling water
until the chocolate is fully melted and smooth. Switch to a rub
ber spatula and folds 1/4 cup of the creme fraiche into the cho
colate. Fold in another cup of the creme fraiche and then quick
ly, before it hardens, spread the chocolate creme fraiche evenl
y over the genoise layer in the pan.\', \'Moisten the second la
yer of genoise with syrup and set it, moistened side down, in t
he pan, pressing gently to level it on the chocolate creme frai
che. Moisten the top of the layer with some of the syrup and to
p with an even layer of fresh raspberries, leaving just a bit o
f space between each berry. Keep 1 perfect berry in reserve.\',
\'Beat the remaining creme fraiche until it holds its shape. Sp
oon 1 to 2 cups of the creme fraiche over the berries and, usin
g an offset spatula, delicately smooth the creme fraiche over a
nd between the berries.\', \'Moisten the remaining layer of gen
oise with syrup and set it, moistened side down, into the pan,
again pressing lightly to set it in place.\', \'Chilling the Ca

ke -- Cover the cake and the remaining creme fraiche with plast
ic and refrigerate for at least 2 to 3 hours, or up to 24 hour
s.\', \'Run a knife around the sides of the cake, then release
and remove the pan or the ring of the springform pan. Put the c
ake, still on its pan bottom, on a large piece of parchment pap
er and set the cake on a decorating turntable, a lazy Susan, or
a large inverted cake pan.\', "Making the Wrap -- Using ridged
plastic shelf liner (available in hardware and housewares store
s), freezer paper, or 500 Mylar (from an art supply store), cut
a strip 26 inches long and 3/8 inch wider than the height of th
e finished cake, about 3 inches. Place a larger piece of waxed
paper on the counter in front of you --this is your drip sheet
-- and put the strip on the waxed paper. (If you\'re using ridg
ed plastic or Mylar, put the smooth glossy side face up.).",
\'Melt the chocolate in the top of a double boiler set over an
inch of barely simmering water or in a microwave oven set at me
dium power, stirring chocolate once or twice until melted and s
mooth. The chocolate should be between 115F to 120°F Pour the c
hocolate down the center of the strip, spreading it with an off
set spatula across the entire strip and beyond -- let it run ov
er a bit onto the waxed paper. (You can scrape up the chocolate
from the waxed paper later and remelt it when you need a dollop
of chocolate to finish the cake.).\', \'Slip the point of a sma
ll knife under one edge of the chocolate-coated strip and grab
the edges of the strip with your fingers.\', \'Slide your free
hand under the strip and grab the other end. Lift the strip and
fit it neatly around the cake, positioning it so that the choco
late side is against the cake. Press one end against the cake a
nd leave the other end standing away from the cake at the point
where it would overlap if you pressed it closed. Slip a small p
iece of waxed paper into this spot, just to hold your place.\',
\'ASSEMBLY AND FINISHING.\', \'Chilling the Wrapped Cake -- Ref
rigerate the cake of at least 1 hour, until the chocolate harde
ns.\', \'Finishing the Wrapped Cake -- Place the cake on the de
corating turntable and spread the remained creme fraiche over t
he top, spreading it out to the edge of the band.\', "Remove th
e chocolate ruffles from the refrigerator and, beginning at the
outside edge, arrange the ruffles in a circle, planting them ge
ntly in the creme fraiche and allowing their frilly edges to ex
tend beyond the cake\'s rim. Continue to arrange the ruffles in
slightly overlapping concentric circles until the creme fraiche
is covered. Put the reserved perfect raspberry in the center of
the cake and chill the cake for about 15 minutes, until firm (o
r up to 6 hours, if necessary), before removing the plastic and
serving.", \'To remove the plastic on the chocolate band, disca
rd the waxed paper "place keeper" and peel away an inch of the
plastic from the end of the band attached to the cake. Put a do
llop of melted chocolate on that end to act as glue and overlap
the other end of the band, pressing lightly to seal it. Careful

```
ly remove the plastic. If the plastic sticks, put the cake back
in the refrigerator for about 10 minutes, then try again.\', "T
o cut the cake, dip a long sharp serrated knife into hot water,
wipe it dry, and cut straight down. Since the first piece is of
ten difficult to remove, it\'s best to make it a generous, easi
er-to-remove slice.", \'Storing -- Although the parts of the ca
ke can be made well in advance, the assembled cake should be se
rved the day it is made.\']'
```

In [49]:
```
print("Summary of number of tags")
df["tags"].apply(lambda x: len(x)).describe()
```

```
Summary of number of tags
```

Out[49]:
```
count    494963.000000
mean        242.596499
std          99.800796
min           4.000000
25%         168.000000
50%         230.000000
75%         304.000000
max        1029.000000
Name: tags, dtype: float64
```

In [50]:
```
print("Summary of number of search terms")
df["search_terms"].apply(lambda x: len(x)).describe()
```

```
Summary of number of search terms
```

Out[50]:
```
count    494963.000000
mean         32.087202
std          19.781546
min           7.000000
25%          19.000000
50%          28.000000
75%          43.000000
max         164.000000
Name: search_terms, dtype: float64
```

Two or three interesting graphs that start to address your main question of interest

I'm not certain that my question of interest can really be addressed with a graph. I'll need to do some more work to analyze the text here and start making predictions.

Answer these questions:

Were there any challenges or obstacles in finding the right dataset for your project?

It took a while to find a good dataset having to do with food, but that would present a significant research question I could answer. I like this one because of the ingredients list. I'm interested in analyzing the connections between ingredients, and this dataset is perfect for that.

Are there any other problems, concerns, or challenges that you are facing regarding your project?

I need to learn more about text analysis and pre-trained models, since that's going to be a big part of how I analyze this data.