Importing Libraraies

In [1]:

```
import pandas as pd
import numpy as np

import matplotlib.pyplot as plt
import seaborn as sns
import matplotlib.cm as cm

from sklearn.feature_extraction.text  import TfidfVectorizer
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_samples, silhouette_score

from collections import defaultdict
```

Reading the dataset

In [35]:

```
df = pd.read_csv('./indian_food.csv')
df.head()
```

Out[35]:

	name	ingredients	diet	prep_time	cook_time	flavor_profile	course	state	region
0	Balu shahi	Maida flour, yogurt, oil, sugar	vegetarian	45	25	sweet	dessert	West Bengal	East
1	Boondi	Gram flour, ghee, sugar	vegetarian	80	30	sweet	dessert	Rajasthan	West
2	Gajar ka halwa	Carrots, milk, sugar, ghee, cashews, raisins	vegetarian	15	60	sweet	dessert	Punjab	North
3	Ghevar	Flour, ghee, kewra, milk, clarified butter, su...	vegetarian	15	30	sweet	dessert	Rajasthan	West
4	Gulab jamun	Milk powder, plain flour, baking powder, ghee,...	vegetarian	15	40	sweet	dessert	West Bengal	East

EDA

In [3]:

```
df.describe()
```

Out[3]:

	prep_time	cook_time
count	270.000000	270.000000
mean	30.988889	34.796296
std	70.762311	46.990539
min	-1.000000	-1.000000
25%	10.000000	20.000000
50%	10.000000	30.000000
75%	20.000000	40.000000

max	500.000000	720.000000

Findind out the null values and replacing them with actual value

In [4]:

```
df.isna().sum()
```

Out[4]:

```
name             0
ingredients      0
diet             0
prep_time        0
cook_time        0
flavor_profile   0
course           0
state            0
region           1
dtype: int64
```

In [5]:

```
df.loc[df['region'].isna(),'region'] = 'North'
```

In [6]:

```
df.loc[df['state']== '-1']
```

Out[6]:

	name	ingredients	diet	prep_time	cook_time	flavor_profile	course	state	region
7	Kaju katli	Cashews, ghee, cardamom, sugar	vegetarian	10	20	sweet	dessert	-1	-1
9	Kheer	Milk, rice, sugar, dried fruits	vegetarian	10	40	sweet	dessert	-1	-1
10	Laddu	Gram flour, ghee, sugar	vegetarian	10	40	sweet	dessert	-1	-1
12	Nankhatai	Refined flour, besan, ghee, powdered sugar, yo...	vegetarian	20	30	sweet	dessert	-1	-1
94	Khichdi	Moong dal, green peas, ginger, tomato, green c...	vegetarian	40	20	spicy	main course	-1	-1
96	Kulfi falooda	Rose syrup, falooda sev, mixed nuts, saffron, ...	vegetarian	45	25	sweet	dessert	-1	-1
98	Lauki ki subji	Bottle gourd, coconut oil, garam masala, ginge...	vegetarian	10	20	spicy	main course	-1	-1
109	Pani puri	Kala chana, mashed potato, boondi, sev, lemon	vegetarian	15	2	spicy	snack	-1	-1
111	Papad	Urad dal, sev, lemon juice, chopped tomatoes	vegetarian	5	5	spicy	snack	-1	-1
115	Rajma chaval	Red kidney beans, garam masala powder, ginger,...	vegetarian	15	90	spicy	main course	-1	North
117	Samosa	Potatoes, green peas, garam masala, ginger, dough	vegetarian	30	30	spicy	snack	-1	-1
128	Dosa	Chana dal, urad dal, whole urad dal, blend ric...	vegetarian	360	90	spicy	snack	-1	South

130	Idli	Split urad dal, urad dal, idli rice, thick poh...	vegetarian	360	90	spicy	snack	-1	South
144	Masala Dosa	Chana dal, urad dal, potatoes, idli rice, thic...	vegetarian	360	90	spicy	snack	-1	South
145	Pachadi	Coconut oil, cucumber, curd, curry leaves, mus...	vegetarian	10	25	-1	main course	-1	South
149	Payasam	Rice, cashew nuts, milk, raisins, sugar	vegetarian	15	30	sweet	dessert	-1	South
154	Rasam	Tomato, curry leaves, garlic, mustard seeds, h...	vegetarian	10	35	spicy	main course	-1	South
156	Sambar	Pigeon peas, eggplant, drumsticks, sambar powd...	vegetarian	20	45	spicy	main course	-1	South
158	Sevai	Sevai, parboiled rice, steamer	vegetarian	120	30	-1	main course	-1	South
161	Uttapam	Chana dal, urad dal, thick poha, tomato, butter	vegetarian	10	20	spicy	snack	-1	South
162	Vada	Urad dal, ginger, curry leaves, green chilies,...	vegetarian	15	20	spicy	snack	-1	South
164	Upma	Chana dal, urad dal, ginger, curry leaves, sugar	vegetarian	10	20	spicy	snack	-1	-1
231	Brown Rice	Brown rice, soy sauce, olive oil	vegetarian	15	25	-1	main course	-1	-1
248	Red Rice	Red pepper, red onion, butter, watercress, oli...	vegetarian	-1	-1	-1	main course	-1	-1

Data Engineering

Here, we are filtering out the state and the ingredients of their corresponding food

In [7]:
```
x = df.groupby('state')['ingredients'].apply(list).reset_index(name='ingredients')
```

In [8]:
```
# function to get the uniqueue values from each string
def get_unique_ingred(ingred):
    for i in ingred:
        word = i.lower().split(',')
    return set(word)
```

In [9]:
```
x['ingredients'] = x['ingredients'].apply(get_unique_ingred)
x
```

Out[9]:

	state	ingredients
0	-1	{red pepper, olive oil, red onion, butter, ...
1	Andhra Pradesh	{green moong beans, rice flour}
2	Assam	{ gur, glutinous rice, black sesame seeds}
3	Bihar	{sattu, dough, atta, filling, mustard oil}

4	Chhattisgarh	{ garam masala powder, arhar dal, white urad...
5	Goa	{ ginger powder, brown rice, fennel seeds, b...
6	Gujarat	{ peas, ridge gourd, sugar, baking soda, gr...
7	Haryana	{ curry leaves, garam masala powder, besan, ...
8	Himachal Pradesh	{ cinnamon, lentils, cloves, salt, yogurt,...
9	Jammu & Kashmir	{ pistachio, badam, cottage cheese, dry date...
10	Jharkhand	{ jaggery, cardamom powder, whole wheat flour...
11	Karnataka	{ curry leaves, thin rice flakes, black sesam...
12	Kerala	{ whole red beans, tamarind, coconut, sesame...
13	Ladakh	{ turmeric powder, salt, tomato, garlic, ch...
14	Madhya Pradesh	{milk powder, arrowroot powder, dry fruits, ...
15	Maharashtra	{ beans, potato, gobi, khus khus, coconut}
16	Manipur	{ slivered almonds, forbidden black rice, gar...
17	NCT of Delhi	{ garam masala powder, cashew nuts, greek yo...
18	Nagaland	{ salt, pork, chillies, axone, water, rice}
19	Odisha	{ curry leaves, dry chilli, cooked rice, curd}
20	Punjab	{ biryani masala powder, yogurt, chickpea flo...
21	Rajasthan	{ khus khus, sesame seeds, whole wheat flour,...
22	Sikkim	{ salt, dried chili, garlic, radish, oil, ...
23	Tamil Nadu	{ cinnamon, tomato, meat curry powder, chick...
24	Telangana	{rose water, white bread slices, milk, saff...
25	Tripura	{ ginger and garlic, boiled pork, onions, ch...
26	Uttar Pradesh	{ musk melon seeds, edible gum, whole wheat f...
27	Uttarakhand	{ coconut, khoa, molu leaf}
28	West Bengal	{ bitter gourd, brinjal, green beans, ridge ...

In [10]:

```
x.drop(0,inplace=True)
```

In [11]:

```
x['ingredients'] = x['ingredients'].apply(' '.join)
x
```

Out[11]:

	state	**ingredients**
1	Andhra Pradesh	green moong beans rice flour
2	Assam	gur glutinous rice black sesame seeds
3	Bihar	sattu dough atta filling mustard oil
4	Chhattisgarh	garam masala powder arhar dal white urad da...
5	Goa	ginger powder brown rice fennel seeds black...
6	Gujarat	peas ridge gourd sugar baking soda grated ...
7	Haryana	curry leaves garam masala powder besan gram...
8	Himachal Pradesh	cinnamon lentils cloves salt yogurt cumi...

9	Jammu & Kashmir	pistachio badam cottage cheese dry dates d...
10	Jharkhand	jaggery cardamom powder whole wheat flour c...
11	Karnataka	curry leaves thin rice flakes black sesame s...
12	Kerala	whole red beans tamarind coconut sesame oil...
13	Ladakh	turmeric powder salt tomato garlic chhurpi...
14	Madhya Pradesh	milk powder arrowroot powder dry fruits all...
15	Maharashtra	beans potato gobi khus khus coconut
16	Manipur	slivered almonds forbidden black rice garlic...
17	NCT of Delhi	garam masala powder cashew nuts greek yogur...
18	Nagaland	salt pork chillies axone water rice
19	Odisha	curry leaves dry chilli cooked rice curd
20	Punjab	biryani masala powder yogurt chickpea flour ...
21	Rajasthan	khus khus sesame seeds whole wheat flour dr...
22	Sikkim	salt dried chili garlic radish oil soy s...
23	Tamil Nadu	cinnamon tomato meat curry powder chicken c...
24	Telangana	rose water white bread slices milk saffron ...
25	Tripura	ginger and garlic boiled pork onions chillies
26	Uttar Pradesh	musk melon seeds edible gum whole wheat flou...
27	Uttarakhand	coconut khoa molu leaf
28	West Bengal	bitter gourd brinjal green beans ridge gour...

In [12]:

```
corpus = x['ingredients'].tolist()
corpus[4][:36]
```

Out[12]:

```
' ginger powder brown rice  fennel se'
```

In [13]:

```
```

In [14]:

```
tfidf = TfidfVectorizer()
vec = tfidf.fit_transform(corpus)
final_df = pd.DataFrame(data=vec.toarray(),columns=tfidf.get_feature_names_out())
final_df.T.nlargest(5, 0)
```

Out[14]:

	0	1	2	3	4	5	6	7	8	9	...	18	19
moong	0.585267	0.000000	0.0	0.000000	0.000000	0.0	0.000000	0.000000	0.0	0.000000	...	0.00000	0.000000
green	0.474853	0.000000	0.0	0.000000	0.000000	0.0	0.000000	0.000000	0.0	0.000000	...	0.00000	0.282721
beans	0.439308	0.000000	0.0	0.000000	0.000000	0.0	0.000000	0.000000	0.0	0.000000	...	0.00000	0.000000
flour	0.345678	0.000000	0.0	0.149386	0.000000	0.0	0.253189	0.000000	0.0	0.258388	...	0.00000	0.205812
rice	0.345678	0.306994	0.0	0.000000	0.240347	0.0	0.000000	0.217426	0.0	0.000000	...	0.26067	0.000000

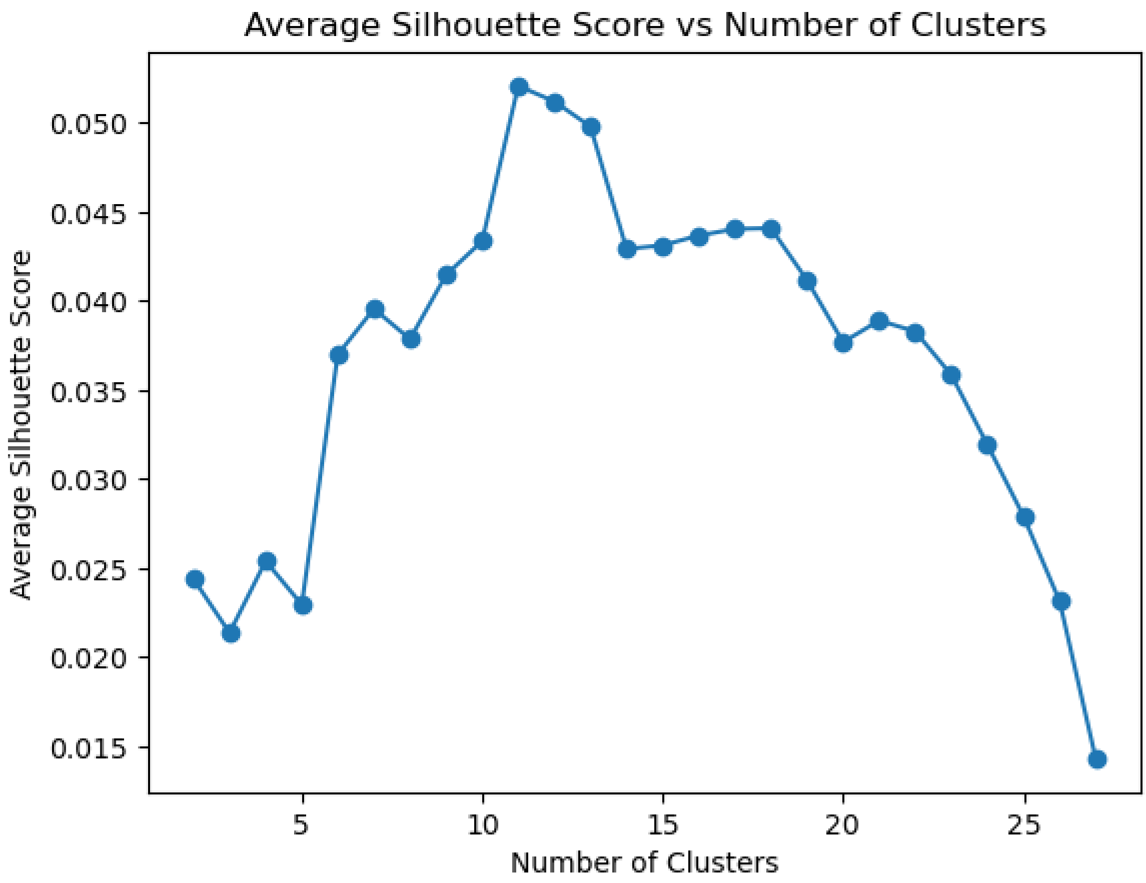
5 rows × 28 columns

K-Means

Calculating the silhouette scores of different cluster to find out the optimal number of cluster

In [16]:

```
max_clusters = 27
silhouette_scores = []

for n_clusters in range(2, max_clusters + 1):
    # Fit K-means to the TF-IDF matrix
    kmeans = KMeans(n_clusters=n_clusters, random_state=42)
    kmeans.fit(final_df)

    # Calculate the silhouette score
    silhouette = silhouette_score(final_df, kmeans.labels_)
    silhouette_scores.append(silhouette)

# Plot the silhouette scores
plt.plot(range(2, max_clusters + 1), silhouette_scores, marker='o')
plt.xlabel('Number of Clusters')
plt.ylabel('Average Silhouette Score')
plt.title('Average Silhouette Score vs Number of Clusters')
plt.show()
```



12 gives us the best silhoutte score

In [17]:

```
no_cluster = 12

kmeans = KMeans(n_clusters=no_cluster,init='k-means++', random_state=36)
y_pred = kmeans.fit_predict(vec)
```

In [18]:

```
x['cluster'] = y_pred
x
```

Out[18]:

	state	ingredients	cluster
1	Andhra Pradesh	green moong beans rice flour	2
2	Assam	gur glutinous rice black sesame seeds	10
3	Bihar	sattu dough atta filling mustard oil	9
4	Chhattisgarh	garam masala powder arhar dal white urad da...	8
5	Goa	ginger powder brown rice fennel seeds black...	10
6	Gujarat	peas ridge gourd sugar baking soda grated ...	2
7	Haryana	curry leaves garam masala powder besan gram...	0
8	Himachal Pradesh	cinnamon lentils cloves salt yogurt cumi...	8
9	Jammu & Kashmir	pistachio badam cottage cheese dry dates d...	3
10	Jharkhand	jaggery cardamom powder whole wheat flour c...	8
11	Karnataka	curry leaves thin rice flakes black sesame s...	10
12	Kerala	whole red beans tamarind coconut sesame oil...	1
13	Ladakh	turmeric powder salt tomato garlic chhurpi...	0
14	Madhya Pradesh	milk powder arrowroot powder dry fruits all...	5
15	Maharashtra	beans potato gobi khus khus coconut	6
16	Manipur	slivered almonds forbidden black rice garlic...	7
17	NCT of Delhi	garam masala powder cashew nuts greek yogur...	11
18	Nagaland	salt pork chillies axone water rice	4
19	Odisha	curry leaves dry chilli cooked rice curd	3
20	Punjab	biryani masala powder yogurt chickpea flour ...	11
21	Rajasthan	khus khus sesame seeds whole wheat flour dr...	6
22	Sikkim	salt dried chili garlic radish oil soy s...	4
23	Tamil Nadu	cinnamon tomato meat curry powder chicken c...	0
24	Telangana	rose water white bread slices milk saffron ...	7
25	Tripura	ginger and garlic boiled pork onions chillies	4
26	Uttar Pradesh	musk melon seeds edible gum whole wheat flou...	8
27	Uttarakhand	coconut khoa molu leaf	1
28	West Bengal	bitter gourd brinjal green beans ridge gour...	2

Plotting

In [19]:

```
import geopandas as gpd
```

In [20]:

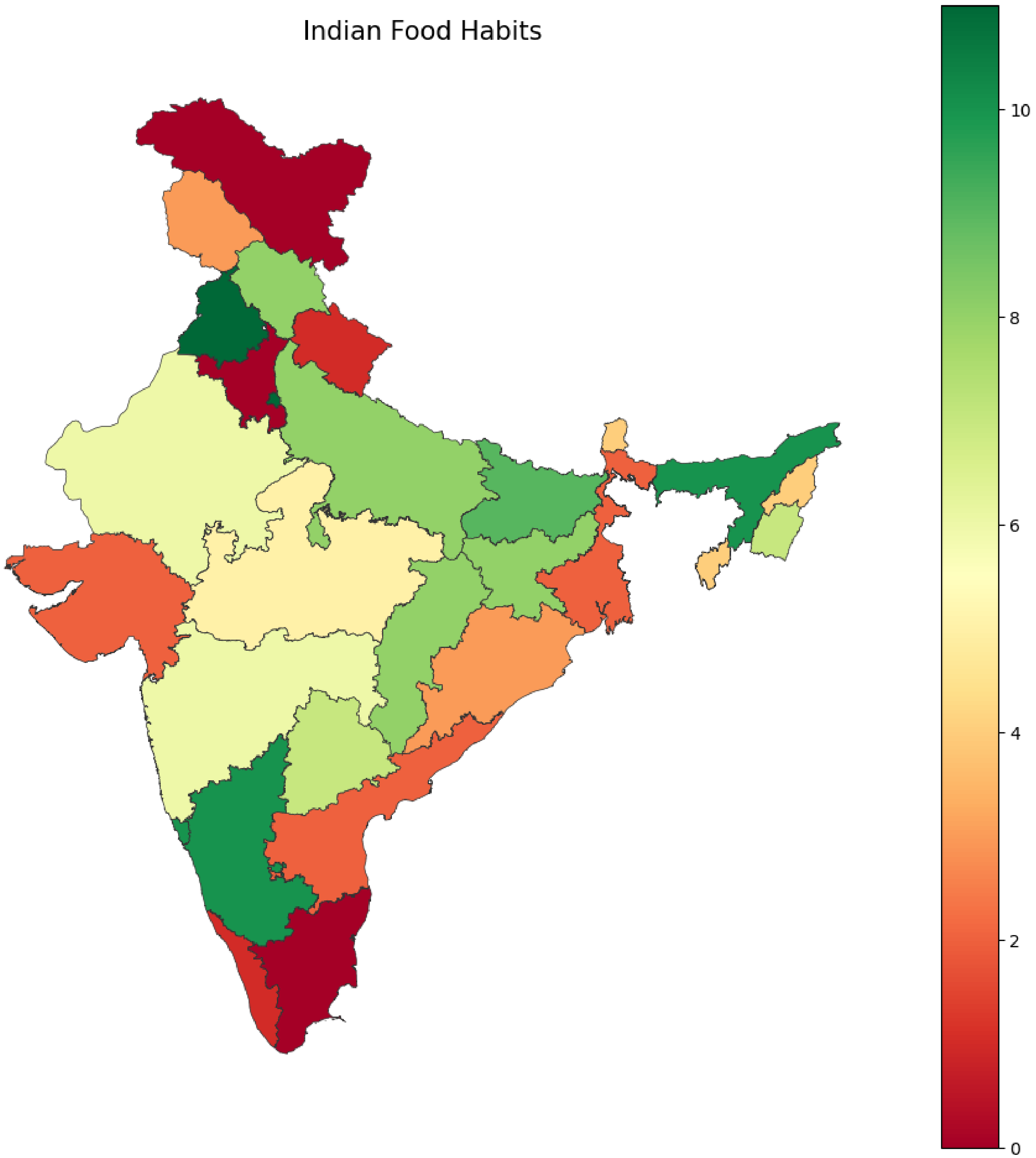
```
out_res = pd.concat([x['state'],x['cluster']], axis=1)
out_res.replace('NCT of Delhi','Delhi',inplace=True)
```

In [21]:

```
indian_map = gpd.read_file('./India-State-and-Country-Shapefile-Updated-Jan-2020-master/

indian_map.replace('Jammu and Kashmir','Jammu & Kashmir',inplace=True)
indian_map.replace('Telengana','Telangana',inplace=True)
indian_map.replace('Tamilnadu','Tamil Nadu',inplace=True)
indian_map.replace('Chhattishgarh','Chhattisgarh',inplace=True)
```

In [22]:
```
merged = indian_map.set_index('State_Name').join(out_res.set_index('state'))
```

In [23]:
```
fig, ax = plt.subplots(1, figsize=(12, 12))
ax.axis('off')
ax.set_title('Indian Food Habits',
             fontdict={'fontsize': '15', 'fontweight' : '3'})
fig = merged.plot(column='cluster', cmap='RdYlGn', linewidth=0.5, ax=ax, edgecolor='0.2'
```

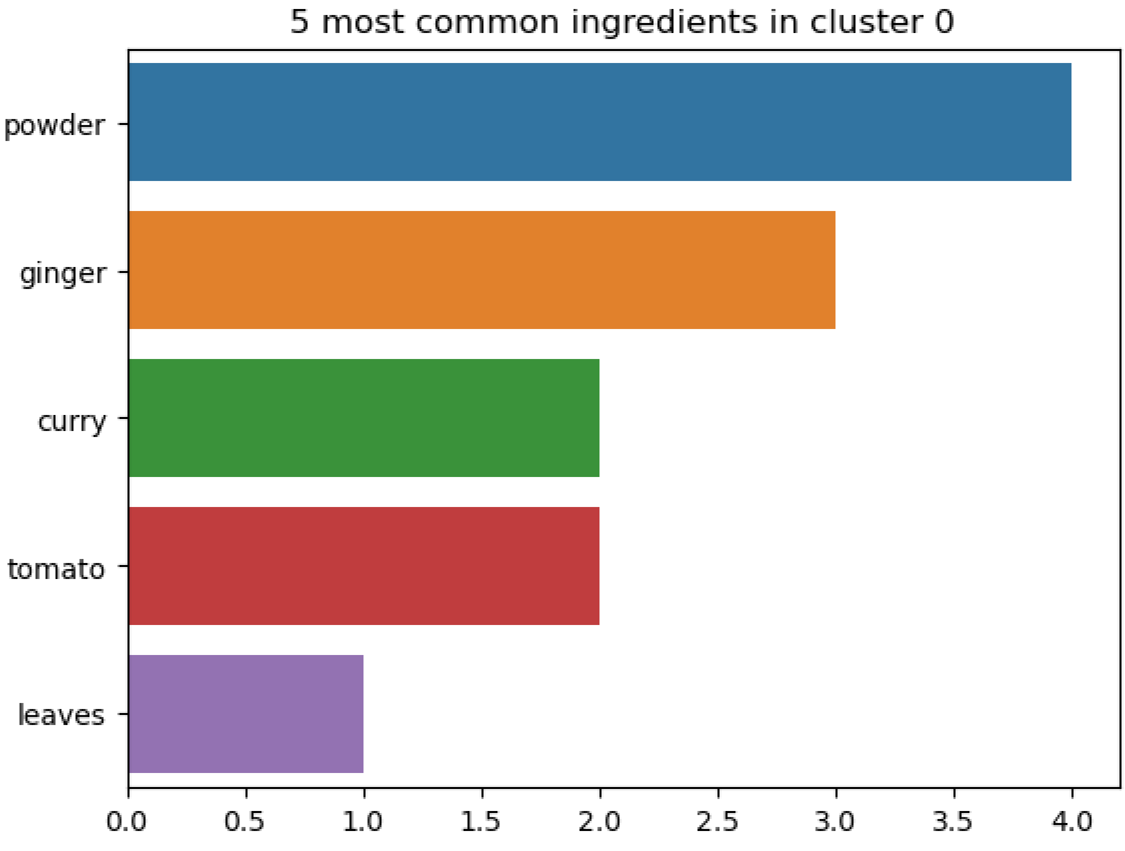


Cluster Analysis

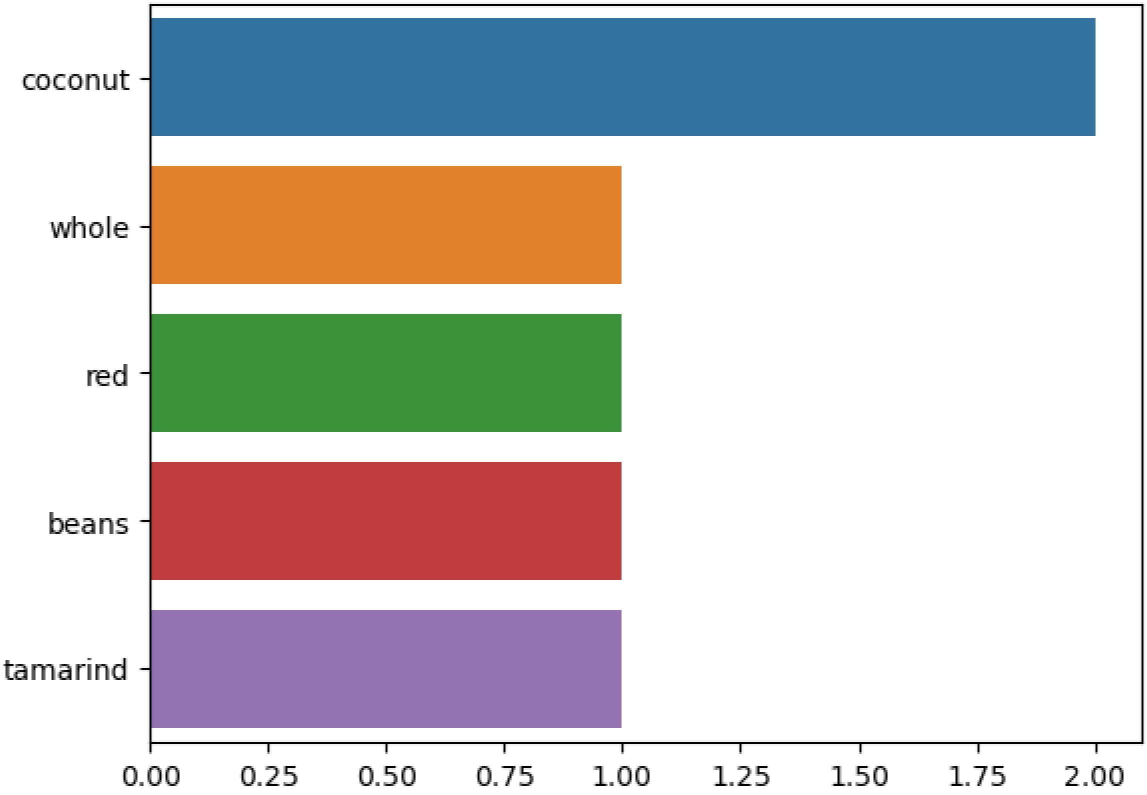
In [33]:
```
def get_top_features_cluster(tf_idf_array, prediction, n_feats):
    cluster_word_count = defaultdict(dict)

    # interate over each cluster
    for cluster_label in  range(no_cluster):
        cluster_df = x[x['cluster'] == cluster_label]
        cluster_text = ' '.join(cluster_df['ingredients'])
        words = cluster_text.split()

        word_count = defaultdict(int)
        for word in words:
            word_count[word] += 1

        cluster_word_count[cluster_label] = word_count

#          return cluster_word_count
        top_cluster_words = defaultdict(dict)

        for cluster_label, word_count in  cluster_word_count.items():
            top_cluster_words[cluster_label] = dict(sorted(word_count.items(), key=lambd


    return top_cluster_words

def plotWords(dfs, n_feats):
    for i in range(no_cluster):
        plt.title((f'Top {n_feats} ingredients in cluster {i}'), fontsize=10, fontweight
        key = list(dfs[i].keys())
        value = list(dfs[i].values())
        sns.barplot(x=value[:n_feats], y=key[:n_feats],orient='h')
        plt.title(f'{n_feats} most common ingredients in cluster {i}')
        plt.show()
```

In [34]:

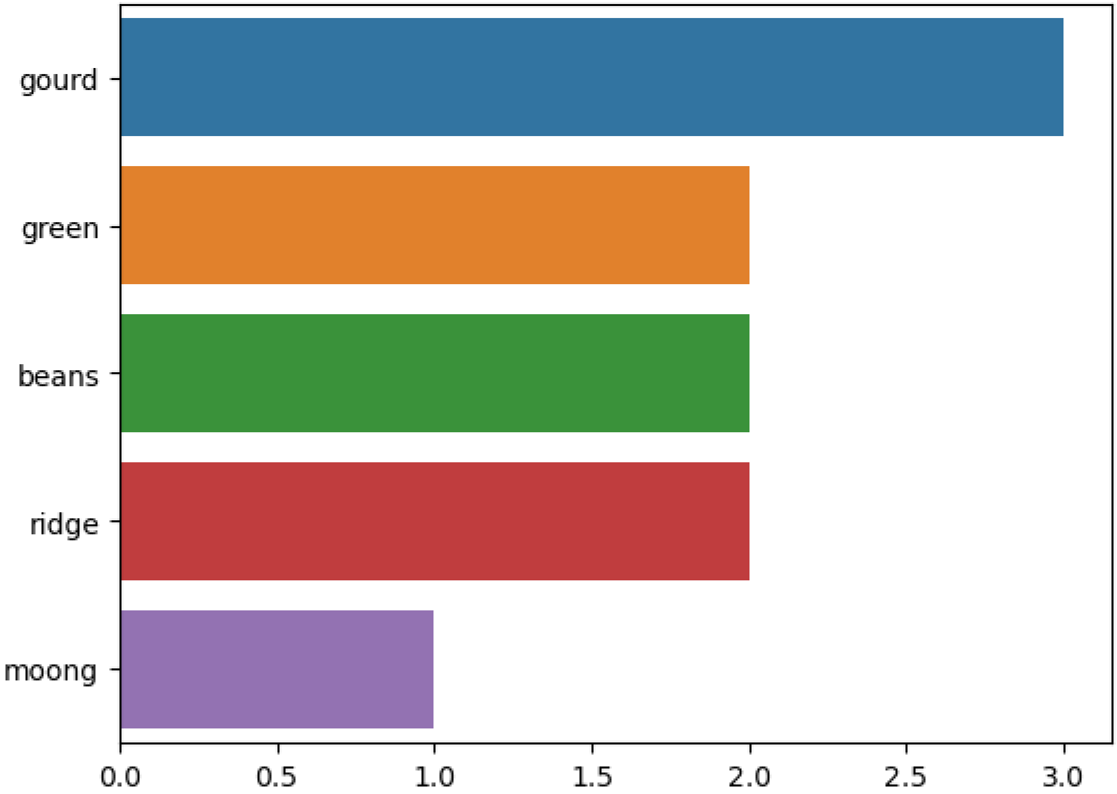
```
final_df_array = final_df.to_numpy()
n_feats = 5
dfs = get_top_features_cluster(final_df_array, y_pred, n_feats)
plotWords(dfs,n_feats=n_feats)
```

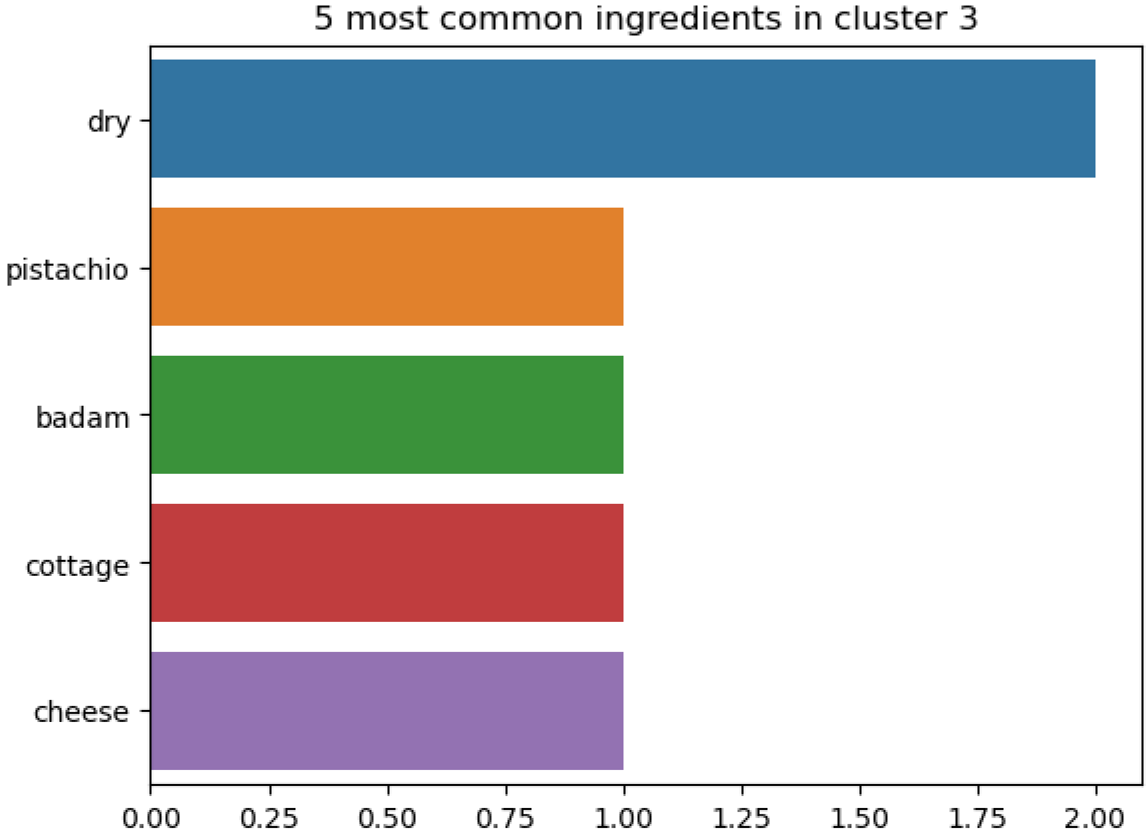


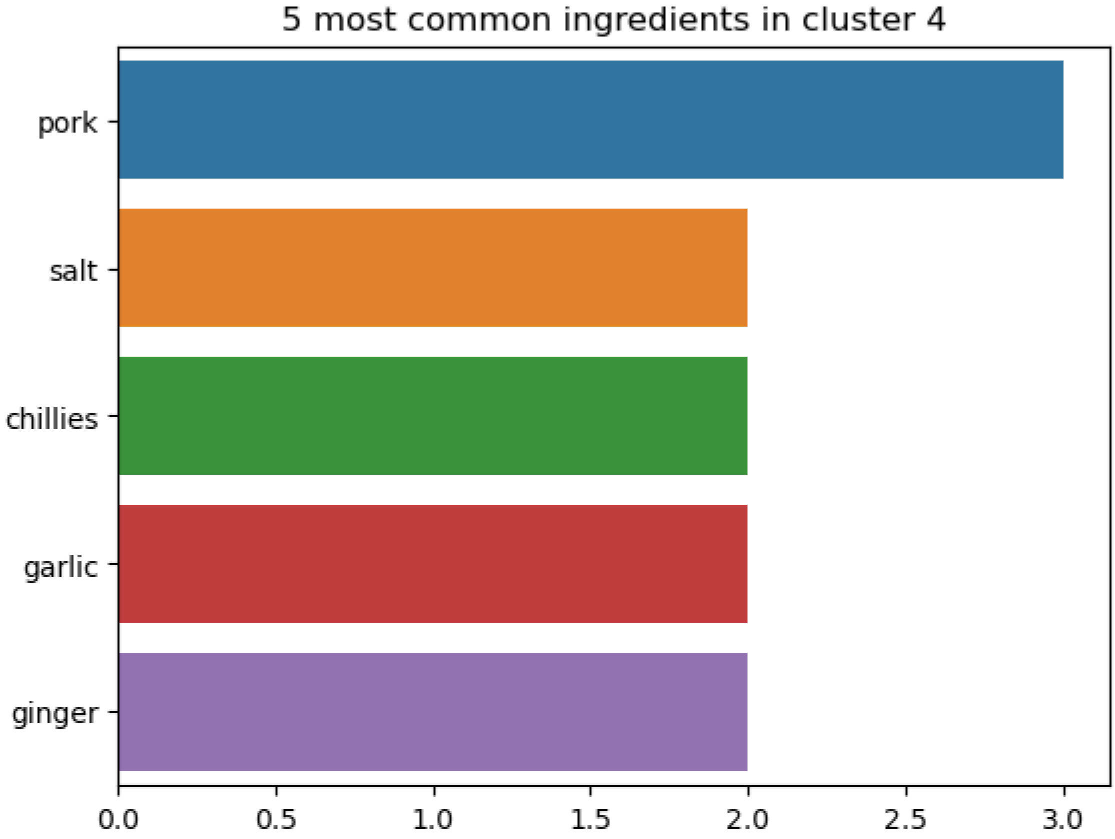
5 most common ingredients in cluster 1

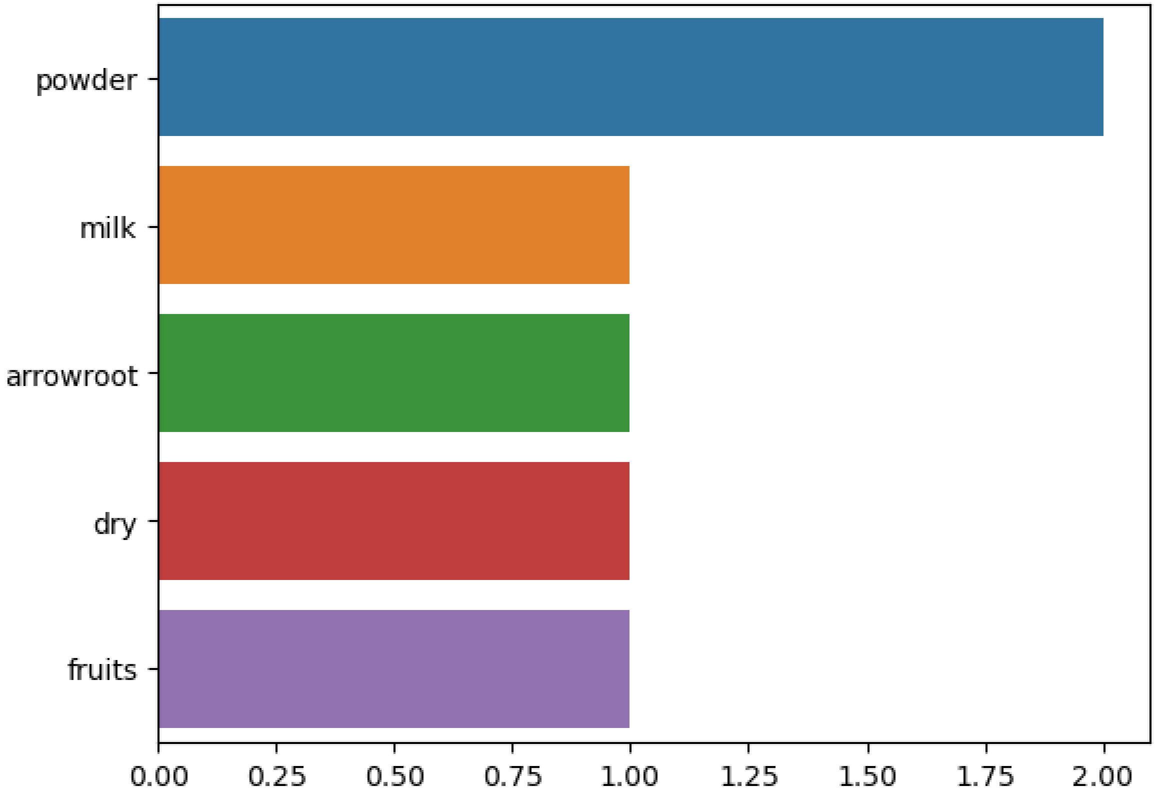


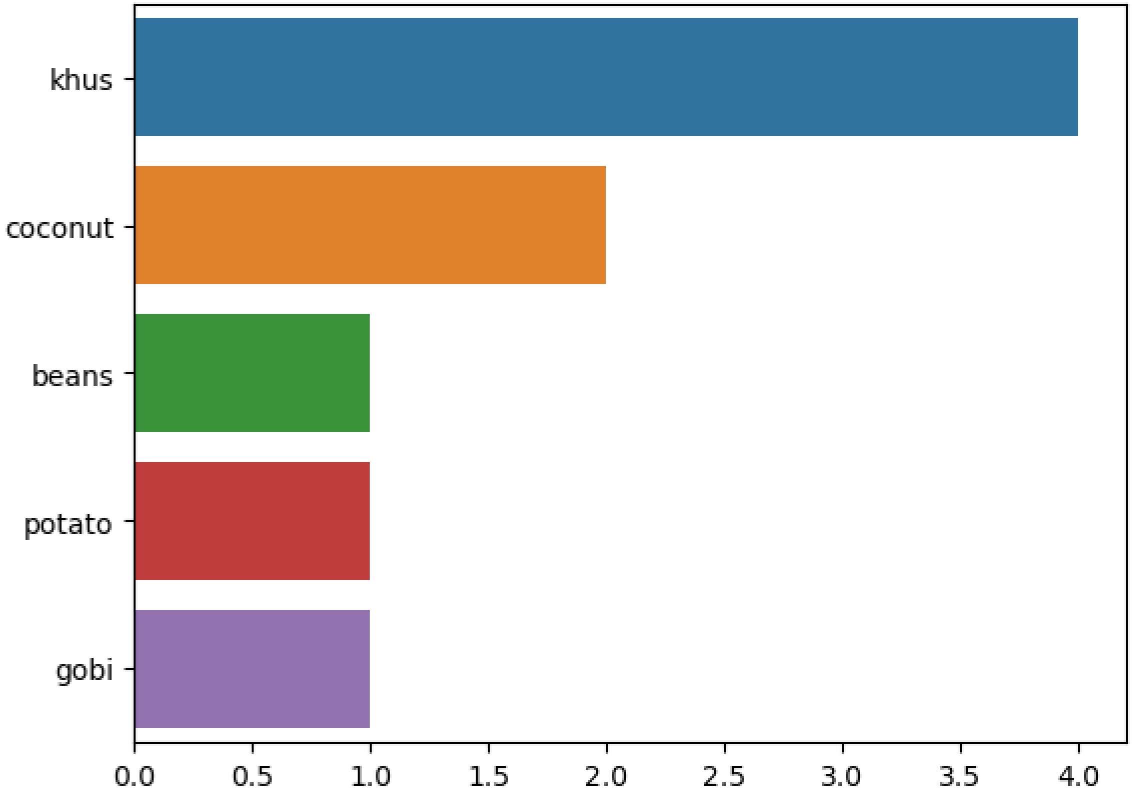
5 most common ingredients in cluster 2

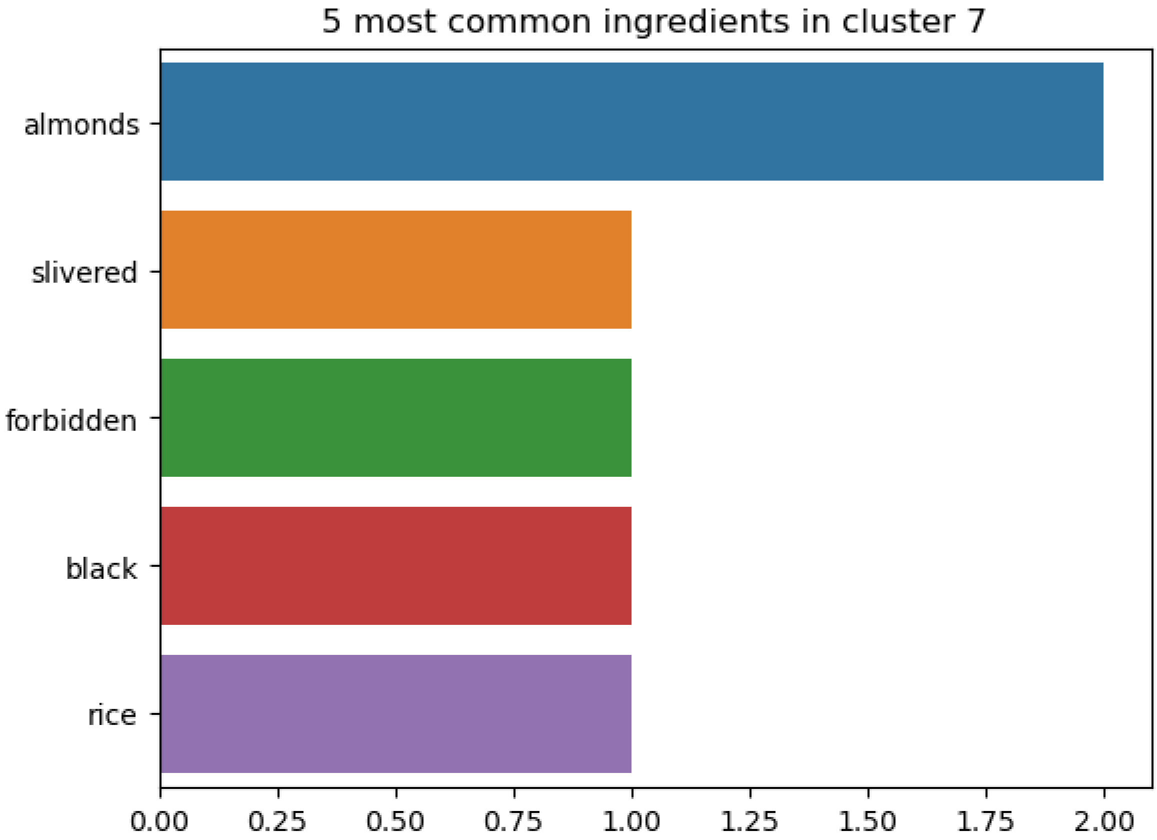


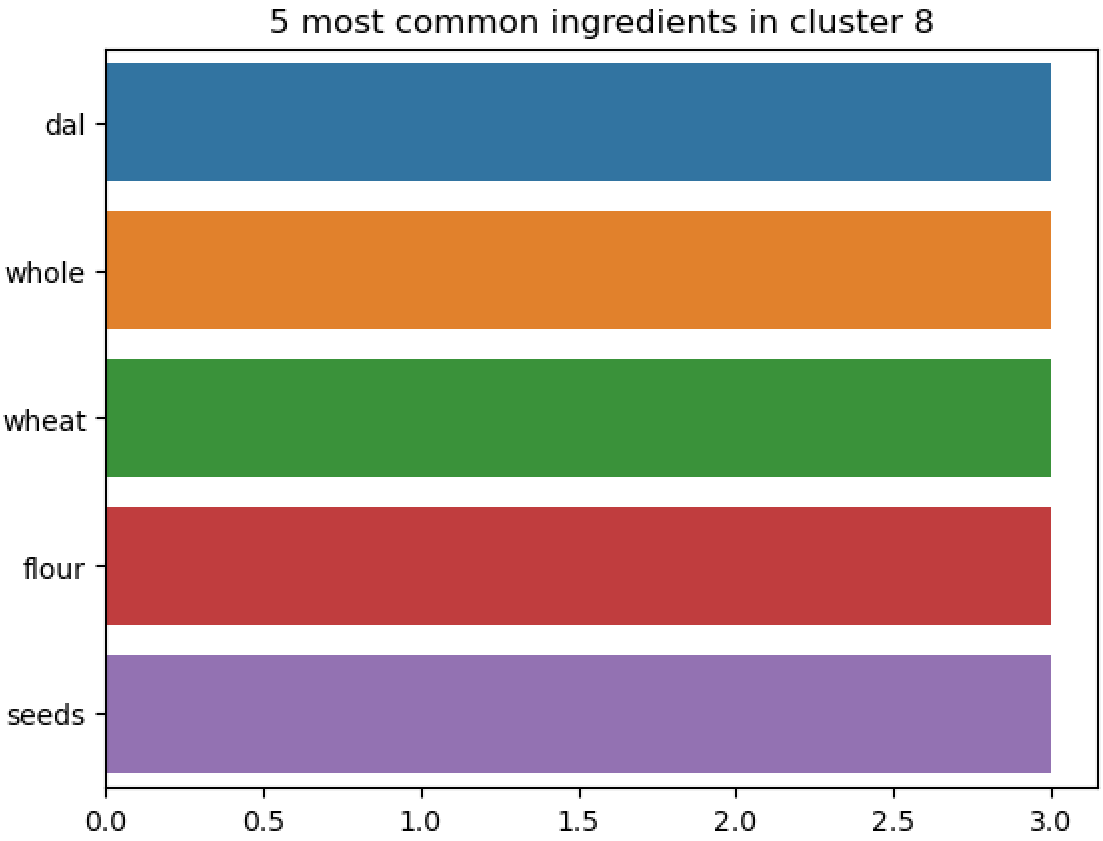

5 most common ingredients in cluster 3

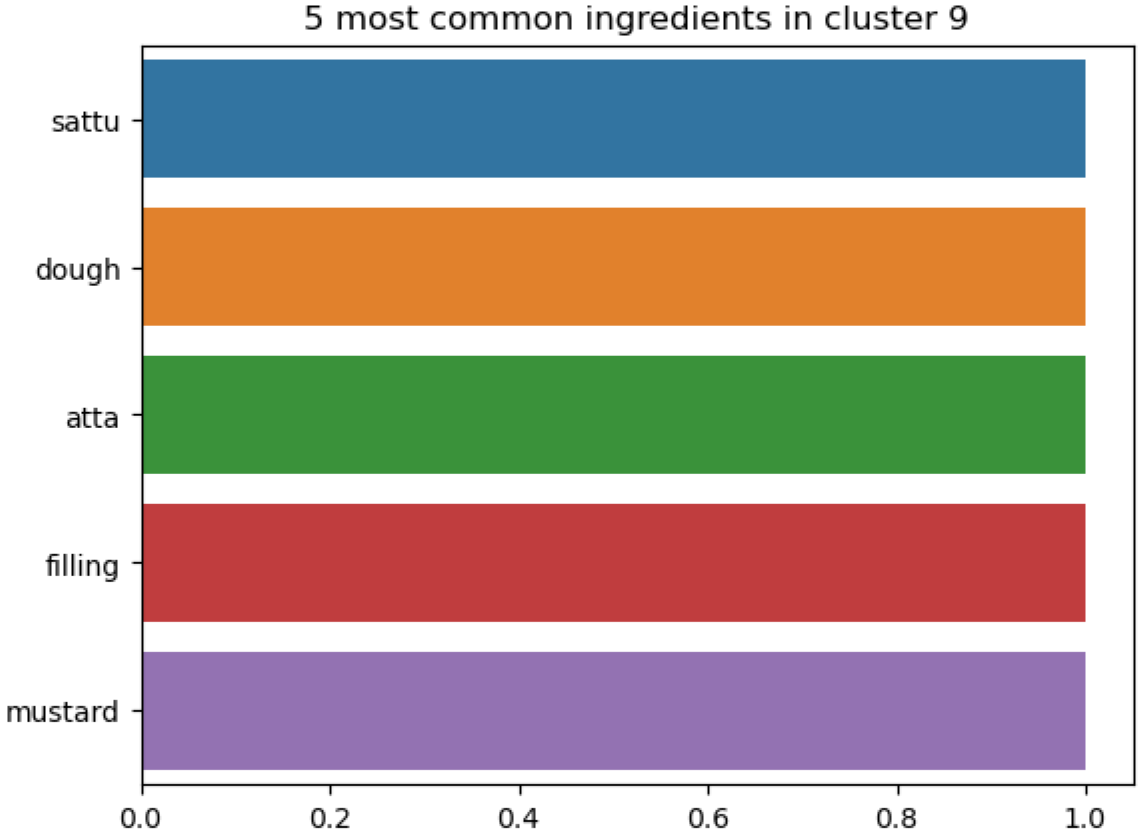

5 most common ingredients in cluster 4

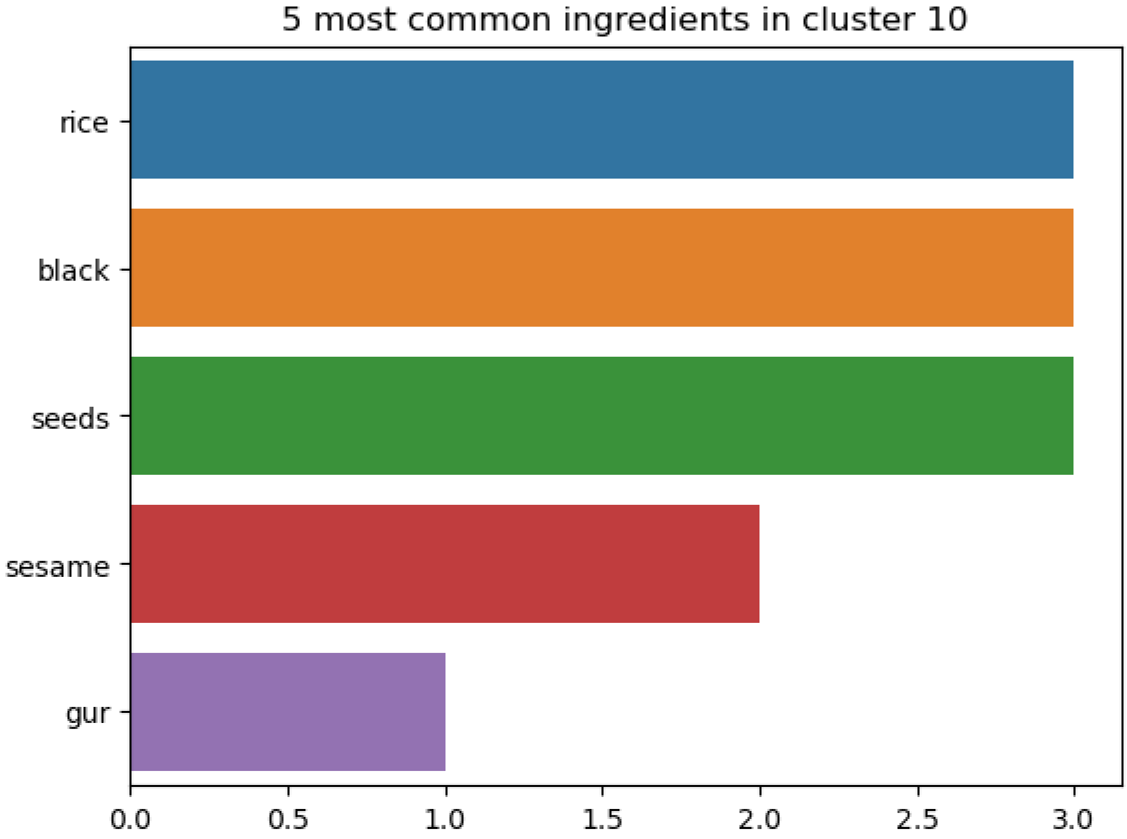

5 most common ingredients in cluster 5

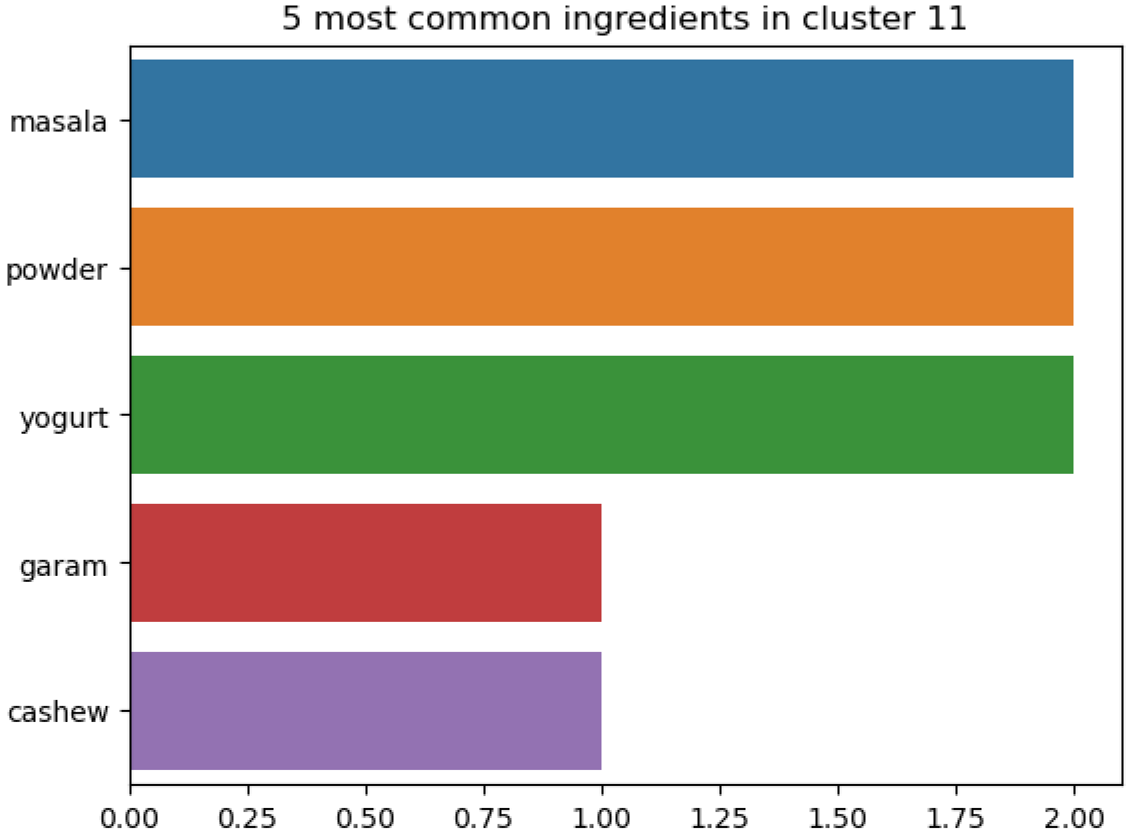

5 most common ingredients in cluster 6


5 most common ingredients in cluster 7
almonds
slivered
forbidden
black
rice
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00


5 most common ingredients in cluster 8
dal
whole
wheat
flour
seeds
0.0
0.5
1.0
1.5
2.0
2.5
3.0


5 most common ingredients in cluster 9


5 most common ingredients in cluster 10


5 most common ingredients in cluster 11
masala
powder
yogurt
garam
cashew
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00