(https://cognitiveclass.ai)

From Understanding to Preparation

Introduction

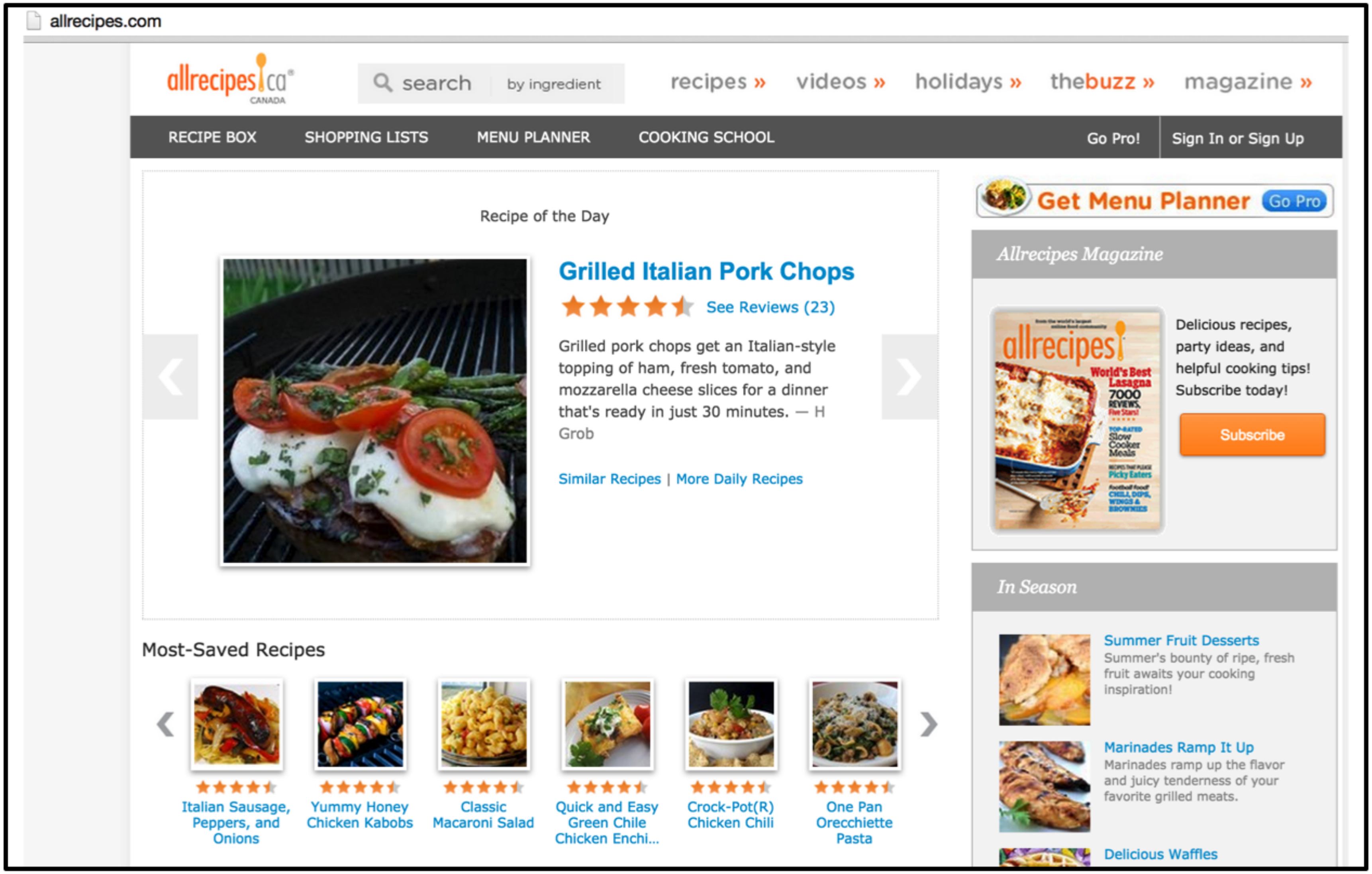
In this lab, we will continue learning about the data science methodology, and focus on the **Data Understanding** and the **Data Preparation** stages.

Table of Contents

</div>

Recap

In Lab **From Requirements to Collection**, we learned that the data we need to answer the question developed in the business understanding stage, namely *can we automate the process of determining the cuisine of a given recipe?*, is readily available. A researcher named Yong-Yeol Ahn scraped tens of thousands of food recipes (cuisines and ingredients) from three different websites, namely:



Recipe of the Day

Grilled Italian Pork Chops

See Reviews (23)

Grilled pork chops get an Italian-style topping of ham, fresh tomato, and mozzarella cheese slices for a dinner that's ready in just 30 minutes. — H Grob

Similar Recipes | More Daily Recipes

Allrecipes Magazine

Delicious recipes, party ideas, and helpful cooking tips! Subscribe today!

Most-Saved Recipes

Italian Sausage, Peppers, and Onions

Yummy Honey Chicken Kabobs

Classic Macaroni Salad

Quick and Easy Green Chile Chicken Enchi...

Crock-Pot(R) Chicken Chili

One Pan Orecchiette Pasta

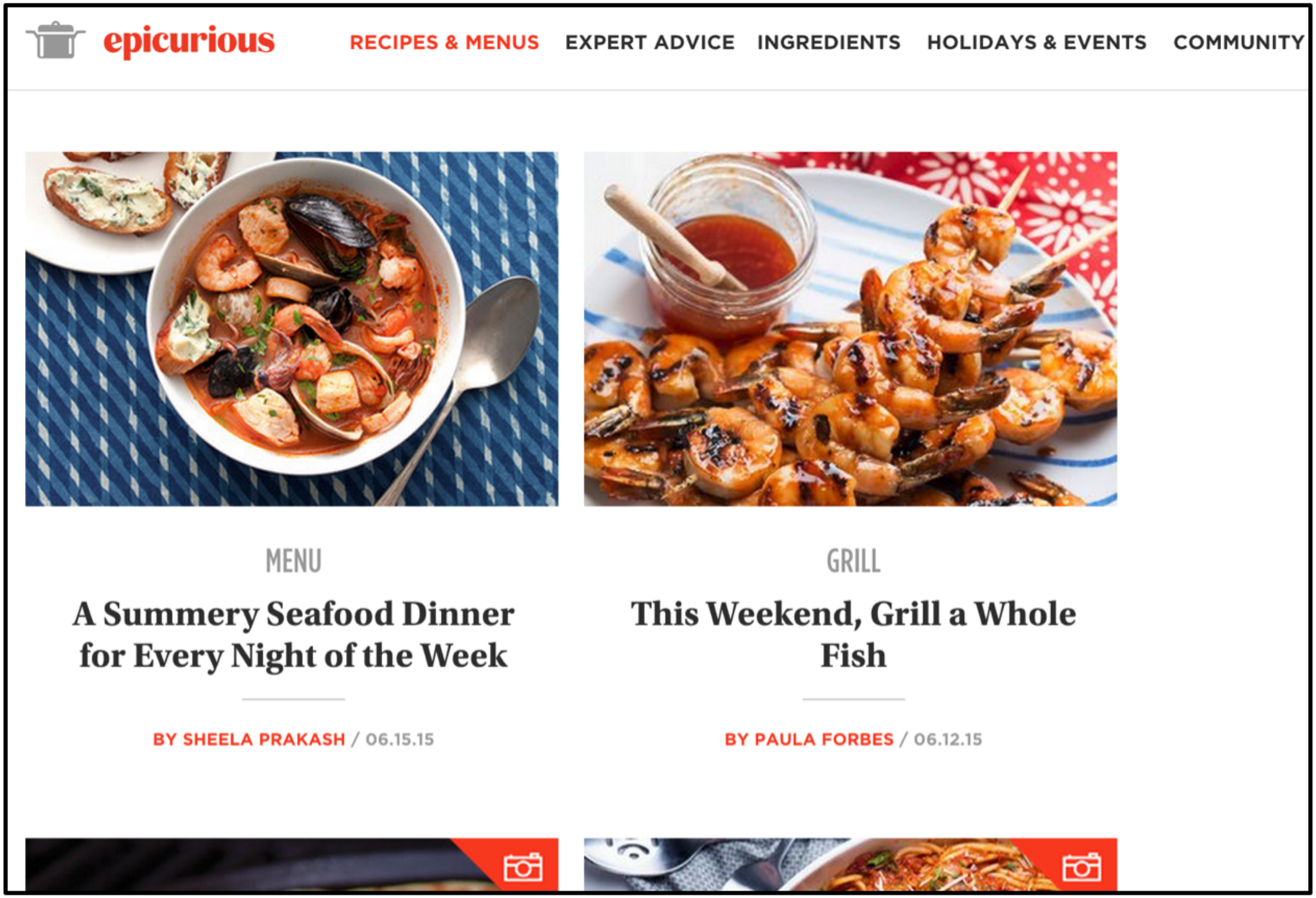
In Season

Summer Fruit Desserts
Summer's bounty of ripe, fresh fruit awaits your cooking inspiration!

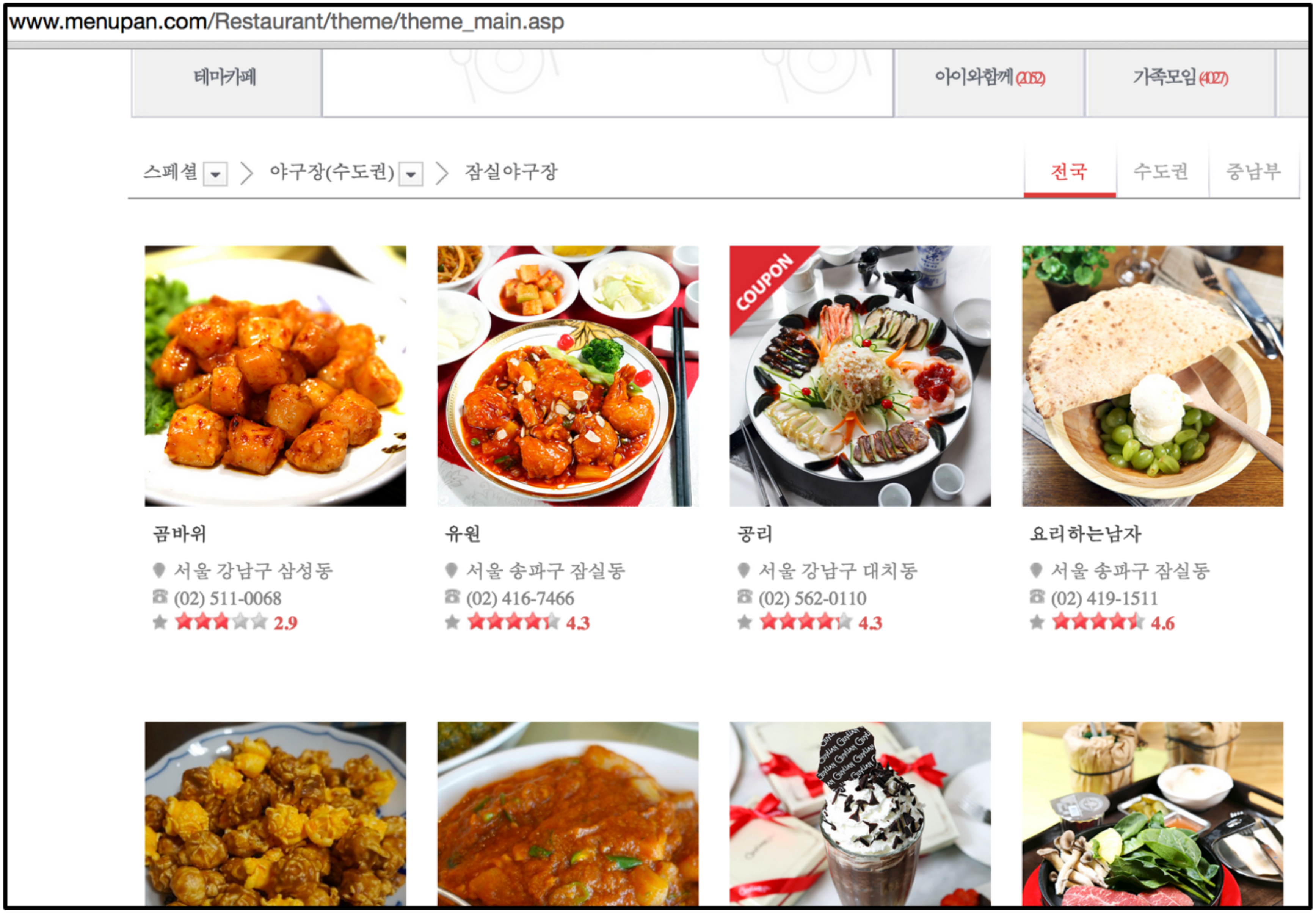
Marinades Ramp It Up
Marinades ramp up the flavor and juicy tenderness of your favorite grilled meats.

Delicious Waffles

www.allrecipes.com


epicurious
RECIPES & MENUS
EXPERT ADVICE
INGREDIENTS
HOLIDAYS & EVENTS
COMMUNITY
MENU
A Summery Seafood Dinner for Every Night of the Week
BY SHEELA PRAKASH / 06.15.15
GRILL
This Weekend, Grill a Whole Fish
BY PAULA FORBES / 06.12.15

www.epicurious.com



곰바위
서울 강남구 삼성동
☎ (02) 511-0068
★ 2.9

유원
서울 송파구 잠실동
☎ (02) 416-7466
★ 4.3

공리
서울 강남구 대치동
☎ (02) 562-0110
★ 4.3

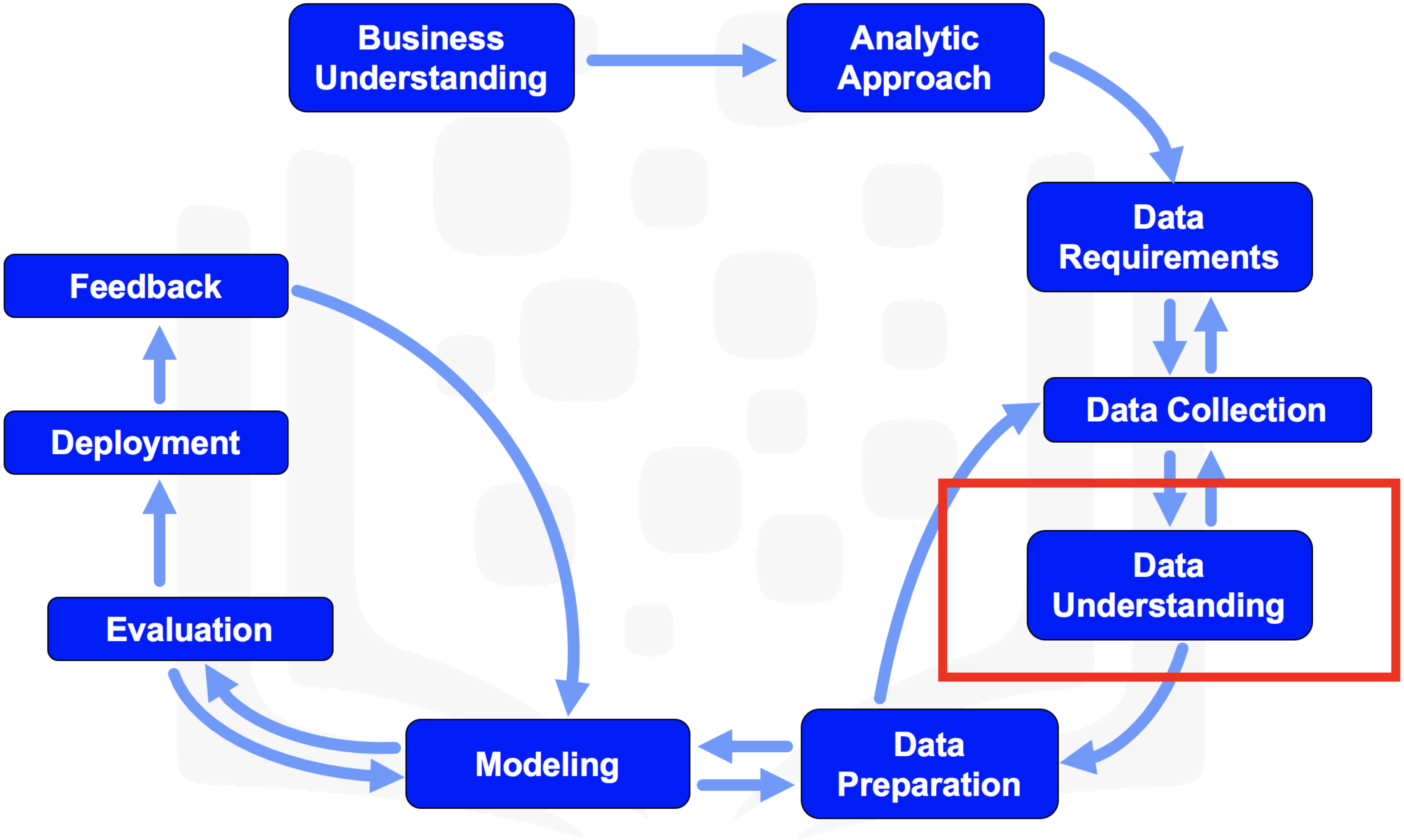
요리하는남자
서울 송파구 잠실동
☎ (02) 419-1511
★ 4.6

www.menupan.com

For more information on Yong-Yeol Ahn and his research, you can read his paper on Flavor Network and the Principles of Food Pairing (http://yongyeol.com/papers/ahn-flavornet-2011.pdf).

We also collected the data and placed it on an IBM server for your convenience.

Data Understanding



Important note: Please note that you are not expected to know how to program in python. The following code is meant to illustrate the stages of data understanding and data preparation, so it is totally fine if you do not understand the individual lines of code. We have a full course on programming in python, Python for Data Science (http://cocl.us/PY0101EN_DS0103EN_LAB3_PYTHON), so please feel free to complete the course if you are interested in learning how to program in python.

Using this notebook:

To run any of the following cells of code, you can type **Shift + Enter** to excute the code in a cell.

Get the version of Python installed.

In [1]:

```
Python 3.6.10
```

Download the library and dependencies that we will need to run this lab.

In [1]:

Download the data from the IBM server and read it into a *pandas* dataframe.

In [2]:

```
Data read into dataframe!
```

Show the first few rows.

In [3]:

Out[3]:

	country	almond	angelica	anise	anise_seed	a
0	Vietnamese	No	No	No	No	
1	Vietnamese	No	No	No	No	
2	Vietnamese	No	No	No	No	
3	Vietnamese	No	No	No	No	
4	Vietnamese	No	No	No	No	

Get the dimensions of the dataframe.

```
In [4]:
```

```
Out[4]:
```

```
(57691, 384)
```

So our dataset consists of 57,691 recipes. Each row represents a recipe, and for each recipe, the corresponding cuisine is documented as well as whether 384 ingredients exist in the recipe or not, beginning with almond and ending with zucchini.

We know that a basic sushi recipe includes the ingredients:

- rice
- soy sauce
- wasabi
- some fish/vegetables

Let's check that these ingredients exist in our dataframe:

In [6]:

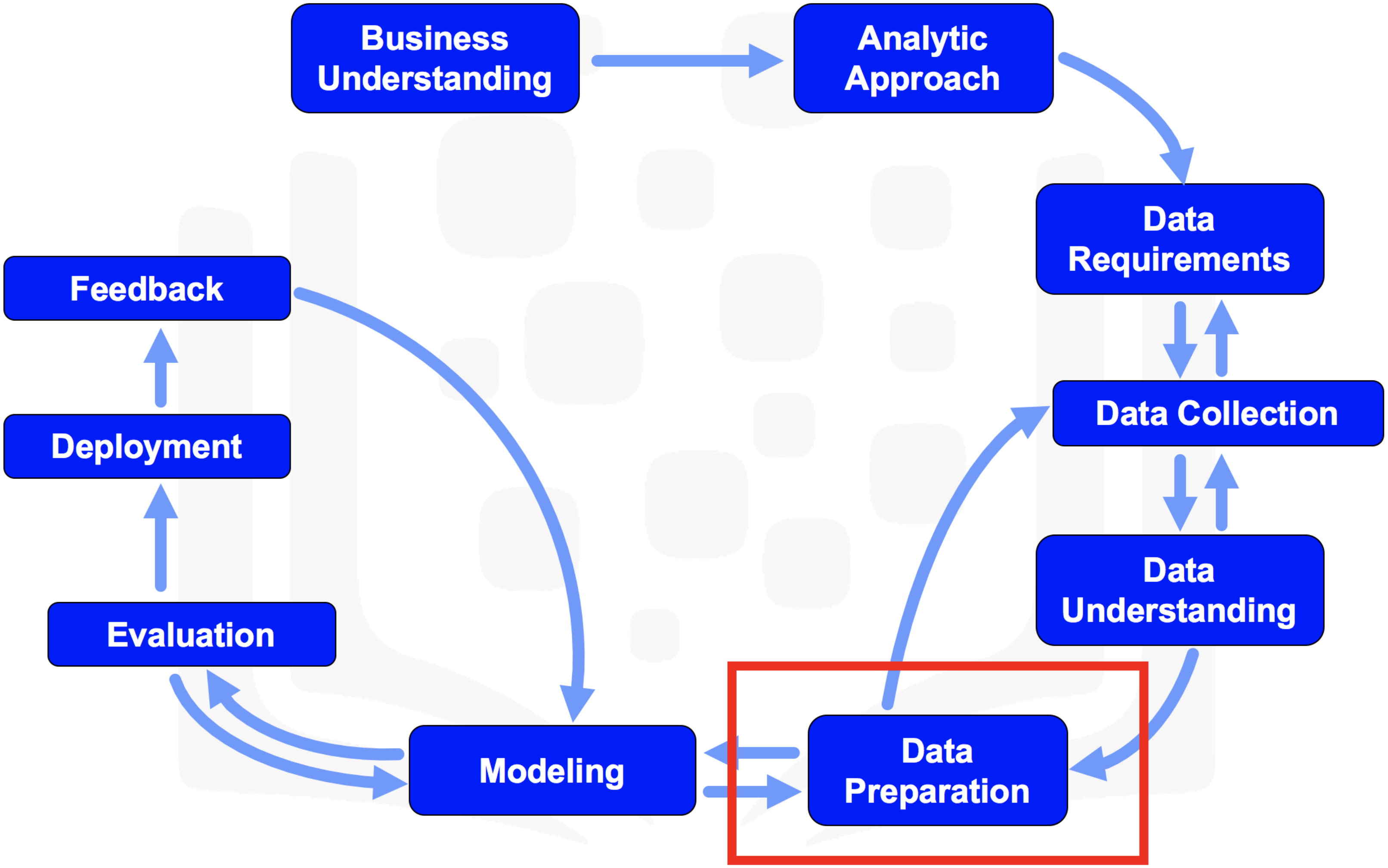
```
['brown_rice', 'licorice', 'rice']
['wasabi']
['soy_sauce', 'soybean', 'soybean_oil']
```

Yes, they do!

- rice exists as rice.
- wasabi exists as wasabi.
- soy exists as soy_sauce.

So maybe if a recipe contains all three ingredients: rice, wasabi, and soy_sauce, then we can confidently say that the recipe is a **Japanese** cuisine! Let's keep this in mind!

Data Preparation


Business Understanding
Analytic Approach
Data Requirements
Data Collection
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Feedback

In this section, we will prepare data for the next stage in the data science methodology, which is modeling. This stage involves exploring the data further and making sure that it is in the right format for the machine learning algorithm that we selected in the analytic approach stage, which is decision trees.

First, look at the data to see if it needs cleaning.

In [5]:

Out[5]:

```
American        40150
Mexico           1754
Italian          1715
Italy            1461
Asian            1176
                ...  
Indonesia          12
Belgium            11
East-African       11
Israel              9
Bangladesh          4
Name: country, Length: 69, dtype: int64
```

By looking at the above table, we can make the following observations:

1. Cuisine column is labeled as Country, which is inaccurate.
2. Cuisine names are not consistent as not all of them start with an uppercase first letter.
3. Some cuisines are duplicated as variation of the country name, such as Vietnam and Vietnamese.
4. Some cuisines have very few recipes.

Let's fixes these problems.

Fix the name of the column showing the cuisine.

In [6]:

Out[6]:

	cuisine	almond	angelica	anise	anise_see
0	Vietnamese	No	No	No	N
1	Vietnamese	No	No	No	N
2	Vietnamese	No	No	No	N
3	Vietnamese	No	No	No	N
4	Vietnamese	No	No	No	N
...	...	...	...	...	
57686	Japan	No	No	No	N
57687	Japan	No	No	No	N
57688	Japan	No	No	No	N
57689	Japan	No	No	No	N
57690	Japan	No	No	No	N

57691 rows × 384 columns

Make all the cuisine names lowercase.

In [7]:

Make the cuisine names consistent.

In [ ]:

Remove cuisines with < 50 recipes.

In [9]:

In [10]:

```
Number of rows of original dataframe is 576
91.
Number of rows of processed dataframe is 57
282.
409 rows removed!
```

Convert all Yes's to 1's and the No's to 0's

In [11]:

Let's analyze the data a little more in order to learn the data better and note any interesting preliminary observations.

Run the following cell to get the recipes that contain **rice** *and* **soy** *and* **wasabi** *and* **seaweed**.

```
In [12]:
```

```
Out[12]:
```

	cuisine	almond	angelica	anise	anise_seed	a
0	vietnamese	0	0	0	0	
1	vietnamese	0	0	0	0	
2	vietnamese	0	0	0	0	
3	vietnamese	0	0	0	0	
4	vietnamese	0	0	0	0	

```
In [13]:
```

Out[13]:

	cuisine	almond	angelica	anise	anise_seec
11306	japanese	0	0	0	(
11321	japanese	0	0	0	(
11361	japanese	0	0	0	(
12171	asian	0	0	0	(
12385	asian	0	0	0	(
13010	asian	0	0	0	(
13159	asian	0	0	0	(
13513	japanese	0	0	0	(
13586	japanese	0	0	0	(
13625	east_asian	0	0	0	(
14495	east_asian	0	0	0	(

Based on the results of the above code, can we classify all recipes that contain **rice** *and* **soy** *and* **wasabi** *and* **seaweed** as **Japanese** recipes? Why?

Your Answer: No

Double-click **here** for the solution.

Let's count the ingredients across all recipes.

```
In [14]:
```

```
In [15]:
```

```
          ingredient  count
0             almond   2306
1           angelica      1
2              anise    223
3         anise_seed     87
4              apple   2420
5       apple_brandy     37
6            apricot    619
7           armagnac     11
8          artemisia     13
9          artichoke    391
10         asparagus    459
11           avocado    660
12             bacon   2166
13      baked_potato      9
14              balm      3
15            banana    989
16            barley    266
17     bartlett_pear     23
18             basil   3833
19               bay   1457
```

```
20                  bean  1971
21                 beech     1
22                  beef  4877
23            beef_broth   842
24            beef_liver    10
25                  beer   307
26                  beet   233
27           bell_pepper  5957
28              bergamot     7
29                 berry   183
30         bitter_orange    85
31            black_bean   494
32         black_currant    11
33  black_mustard_seed_oil    30
34          black_pepper  9795
35       black_raspberry     8
36     black_sesame_seed    26
37             black_tea    44
38            blackberry   170
39     blackberry_brandy     4
40           blue_cheese   396
```

```
41                blueberry    466
42                 bone_oil     50
43          bourbon_whiskey    156
44                   brandy    395
45                 brassica    114
46                    bread   4567
47                 broccoli    929
48               brown_rice    345
49          brussels_sprout     92
50                buckwheat     90
51                   butter  20699
52               buttermilk   1634
53                  cabbage   1011
54  cabernet_sauvignon_wine     17
55                    cacao     35
56         camembert_cheese     12
57            cane_molasses   7735
58                  caraway    233
59                 cardamom    352
60                carnation      3
61                    carob      7
```

```
62              carrot  3673
63              cashew   208
64             cassava    19
65             catfish    71
66         cauliflower   332
67              caviar    28
68             cayenne  8225
69              celery  3621
70          celery_oil  1002
71              cereal   204
72           chamomile     3
73      champagne_wine   100
74             chayote    27
75      cheddar_cheese  3027
76              cheese  3278
77              cherry  1082
78       cherry_brandy    32
79             chervil    52
80             chicken  5425
81       chicken_broth  3598
82       chicken_liver    52
```

```
83            chickpea   401
84             chicory   156
85     chinese_cabbage   165
86               chive  1332
87               cider  1129
88            cilantro  2454
89            cinnamon  5589
90              citrus   167
91         citrus_peel     4
92                clam   472
93               clove    10
94               cocoa  4797
95             coconut  1800
96         coconut_oil    17
97                 cod   179
98              coffee   718
99              cognac    67
100      concord_grape    12
101          condiment     9
102          coriander  1646
103               corn  4824
```

```
104              corn_flake    225
105               corn_grit    163
106          cottage_cheese    347
107                    crab    571
108               cranberry    920
109                   cream  10169
110            cream_cheese   2840
111                cucumber   1888
112                   cumin   3270
113              cured_pork    315
114                 currant    240
115                    date    375
116                    dill   1105
117                  durian      0
118                     eel     20
119                     egg  20997
120              egg_noodle    316
121              elderberry      5
122        emmental_cheese      1
123                  endive    115
124               enokidake    106
```

```
125                 fennel    912
126              fenugreek    923
127            feta_cheese    623
128                    fig    139
129                   fish   2087
130                 flower     32
131            frankfurter     37
132                  fruit    479
133                galanga     49
134               gardenia      9
135                 garlic  17287
136                gelatin   1415
137               geranium      1
138                    gin     68
139                 ginger   4340
140            goat_cheese    260
141                  grape    346
142           grape_brandy      8
143            grape_juice    824
144             grapefruit    121
145      green_bell_pepper   2579
```

```
146                green_tea     35
147           gruyere_cheese     45
148                    guava     13
149                  haddock     31
150                      ham   1298
151                 hazelnut    284
152                  herring     10
153               holy_basil      3
154                    honey   2550
155                      hop      3
156              horseradish    396
157              huckleberry     10
158             jamaican_rum      1
159            japanese_plum     13
160                  jasmine      8
161              jasmine_tea      2
162            juniper_berry     33
163              kaffir_lime      1
164                     kale     96
165              katsuobushi     63
166                     kelp    179
```

```
167            kidney_bean   436
168                   kiwi   109
169               kohlrabi     6
170                kumquat    33
171                   lamb   481
172                   lard  3049
173                 laurel     2
174               lavender    62
175                   leaf     9
176                   leek   422
177                  lemon  3037
178            lemon_juice  5060
179             lemon_peel   728
180             lemongrass   211
181                 lentil   247
182                lettuce  1199
183               licorice    21
184       lilac_flower_oil     1
185              lima_bean   149
186                   lime  1152
187             lime_juice  1611
```

```
188    lime_peel_oil   108
189      lingonberry     9
190           litchi    12
191            liver    42
192          lobster   131
193      long_pepper     2
194           lovage   142
195    macadamia_nut   102
196         macaroni  3112
197             mace   117
198         mackerel    44
199             malt    37
200         mandarin   279
201    mandarin_peel    15
202            mango   418
203      maple_syrup   477
204         marjoram   527
205             mate     1
206        matsutake    57
207             meat   985
208            melon   163
```

```
209                  milk  12855
210              milk_fat    959
211                  mint   1004
212    mozzarella_cheese   1288
213             mung_bean     23
214        munster_cheese     27
215           muscat_grape     1
216              mushroom   3367
217                mussel    168
218               mustard   4118
219                mutton      3
220             nectarine     51
221                  nira     67
222                   nut   1254
223                nutmeg   2504
224                   oat   1265
225              oatmeal      61
226               octopus     45
227                  okra    102
228                 olive   1798
229             olive_oil   9855
```

```
230                    onion  18033
231                   orange   1721
232            orange_flower     17
233             orange_juice   1725
234              orange_peel    596
235                  oregano   3177
236                     ouzo      9
237                   oyster    404
238                     palm     46
239                   papaya     57
240          parmesan_cheese   3173
241                  parsley   5541
242                  parsnip    139
243            passion_fruit     20
244                      pea   1178
245                    peach    531
246                   peanut    505
247            peanut_butter   1014
248               peanut_oil    304
249                     pear    482
250              pear_brandy     11
```

```
251                    pecan   2176
252              pelargonium      1
253                   pepper   9200
254               peppermint    142
255           peppermint_oil      8
256                  pimenta      0
257                  pimento    270
258                pineapple   1637
259                pistachio    219
260                     plum    288
261                  popcorn     97
262                  porcini    106
263                     pork   2048
264               pork_liver      5
265             pork_sausage   1357
266                port_wine     48
267                   potato   3510
268              potato_chip     65
269                    prawn     24
270             prickly_pear     20
271         provolone_cheese    168
```

```
272           pumpkin   803
273            quince    28
274            radish   522
275            raisin  1889
276          rapeseed     3
277         raspberry   784
278          raw_beef     2
279         red_algae     2
280          red_bean    33
281   red_kidney_bean    59
282          red_wine  1393
283           rhubarb   169
284              rice  3829
285    roasted_almond     3
286      roasted_beef   226
287  roasted_hazelnut     1
288      roasted_meat    15
289       roasted_nut     1
290    roasted_peanut   201
291     roasted_pecan     1
292      roasted_pork   124
```

```
293   roasted_sesame_seed   593
294        romano_cheese    275
295                 root    101
296     roquefort_cheese     23
297                 rose     56
298             rosemary   1892
299                  rum    599
300             rutabaga     34
301            rye_bread     92
302            rye_flour    131
303              saffron    234
304                 sage    904
305                 sake    680
306               salmon    451
307           salmon_roe     15
308            sassafras     18
309           sauerkraut    185
310               savory    127
311             scallion   4760
312              scallop    300
313            sea_algae      4
```

```
314            seaweed   212
315               seed  1340
316         sesame_oil  1671
317        sesame_seed   764
318            shallot  1301
319       sheep_cheese     2
320          shellfish    27
321             sherry   705
322           shiitake   594
323             shrimp  1672
324              smoke   460
325        smoked_fish     6
326      smoked_salmon   100
327     smoked_sausage   267
328        sour_cherry    50
329          sour_milk    46
330          soy_sauce  3765
331            soybean  1184
332        soybean_oil     2
333          spearmint     6
334             squash   571
```

```
335                squid   237
336           star_anise   129
337               starch  2723
338           strawberry  1080
339       strawberry_jam     1
340     strawberry_juice     2
341      sturgeon_caviar     1
342                sumac    11
343        sunflower_oil     8
344         sweet_potato   527
345         swiss_cheese   519
346       tabasco_pepper   971
347             tamarind  1670
348            tangerine    52
349             tarragon   478
350                  tea   108
351              tequila   142
352          thai_pepper   136
353                thyme  3043
354               tomato  9902
355         tomato_juice   176
```

```
356                  truffle     52
357                     tuna    461
358                   turkey    900
359                 turmeric   1289
360                   turnip    188
361                  vanilla   9005
362                     veal    197
363                vegetable   1700
364            vegetable_oil  11078
365                  vinegar   8038
366                   violet      5
367                   walnut   2727
368                   wasabi    135
369               watercress    149
370               watermelon    110
371                    wheat  20757
372              wheat_bread     82
373                  whiskey    148
374              white_bread    370
375               white_wine   2196
376  whole_grain_wheat_flour    731
```

```
377          wine   1019
378          wood     33
379           yam     85
380         yeast   3377
381        yogurt   1033
382      zucchini   1100
```

Now we have a dataframe of ingredients and their total counts across all recipes. Let's sort this dataframe in descending order.

In [16]:

```
          ingredient  count
0                egg  20997
1              wheat  20757
2             butter  20699
3              onion  18033
4             garlic  17287
..               ...    ...
378  sturgeon_caviar      1
379      kaffir_lime      1
380            beech      1
381           durian      0
382          pimenta      0

[383 rows x 2 columns]
```

What are the 3 most popular ingredients?

Your Answer: 1. egg 2.wheat 3.butter

Double-click **here** for the solution.

However, note that there is a problem with the above table. There are ~40,000 American recipes in our dataset, which means that the data is biased towards American ingredients.

Therefore, let's compute a more objective summary of the ingredients by looking at the ingredients per cuisine.

Let's create a *profile* for each cuisine.

In other words, let's try to find out what ingredients Chinese people typically use, and what is **Canadian** food for example.

```
In [17]:
```

```
Out[17]:
```

	almond	angelica	anise	anise_see
cuisine				
african	0.156522	0.000000	0.000000	0.00000
american	0.040598	0.000025	0.003014	0.00057
asian	0.007544	0.000000	0.000838	0.00251
cajun_creole	0.000000	0.000000	0.000000	0.00000
canada	0.036176	0.000000	0.000000	0.00000

As shown above, we have just created a dataframe where each row is a cuisine and each column (except for the first column) is an ingredient, and the row values represent the percentage of each ingredient in the corresponding cuisine.

For example:

- *almond* is present across 15.65% of all of the **African** recipes.
- *butter* is present across 38.11% of all of the **Canadian** recipes.

Let's print out the profile for each cuisine by displaying the top four ingredients in each cuisine.

```
In [18]:
```

```
AFRICAN
onion (53%) olive_oil (52%) garlic (49%) cu
min (42%)

AMERICAN
butter (41%) egg (40%) wheat (39%) onion (2
9%)

ASIAN
soy_sauce (49%) ginger (48%) garlic (47%) r
ice (41%)

CAJUN_CREOLE
onion (69%) cayenne (56%) garlic (48%) butt
er (36%)

CANADA
wheat (39%) butter (38%) egg (35%) onion (3
4%)

CARIBBEAN
```

```
onion (51%) garlic (50%) vegetable_oil (3
1%) black_pepper (31%)

CENTRAL_SOUTHAMERICAN
garlic (56%) onion (54%) cayenne (51%) toma
to (41%)

CHINA
soy_sauce (70%) garlic (45%) scallion (43%)
egg (39%)

CHINESE
soy_sauce (67%) ginger (59%) garlic (56%) s
callion (50%)

EAST_ASIAN
garlic (55%) soy_sauce (50%) scallion (49%)
cayenne (47%)

EASTERN-EUROPE
wheat (53%) egg (52%) butter (48%) onion (4
5%)

EASTERNEUROPEAN_RUSSIAN
butter (60%) egg (50%) wheat (49%) onion (3
8%)

ENGLISH_SCOTTISH
butter (67%) wheat (62%) egg (53%) cream (4
1%)

FRANCE
butter (54%) egg (46%) wheat (43%) onion (3
2%)

FRENCH
butter (48%) egg (43%) wheat (35%) olive_oi
l (30%)

GERMAN
butter (55%) wheat (50%) onion (48%) egg (4
2%)
```

```
GERMANY
wheat (67%) egg (64%) butter (45%) onion (3
1%)

GREEK
olive_oil (76%) garlic (44%) onion (36%) le
mon_juice (33%)

INDIA
cumin (62%) onion (57%) turmeric (54%) garl
ic (50%)

INDIAN
cumin (58%) coriander (46%) turmeric (46%)
cayenne (45%)

IRISH
butter (59%) wheat (50%) egg (46%) cream (2
6%)
```

```
ITALIAN
olive_oil (65%) garlic (45%) tomato (30%) o
nion (28%)

ITALY
garlic (61%) olive_oil (55%) tomato (49%) b
asil (44%)

JAPAN
soy_sauce (55%) rice (42%) vegetable_oil (3
8%) vinegar (34%)

JAPANESE
soy_sauce (57%) rice (45%) vinegar (37%) ve
getable_oil (33%)

JEWISH
egg (59%) wheat (48%) butter (30%) onion (3
0%)

KOREAN
```

```
garlic (58%) scallion (52%) cayenne (52%) s
oy_sauce (48%)

MEDITERRANEAN
olive_oil (79%) garlic (50%) onion (38%) to
mato (34%)

MEXICAN
cayenne (70%) onion (60%) garlic (56%) toma
to (49%)

MEXICO
cayenne (74%) onion (71%) garlic (63%) toma
to (62%)

MIDDLEEASTERN
olive_oil (60%) garlic (46%) wheat (37%) le
mon_juice (35%)

MOROCCAN
olive_oil (72%) cumin (54%) onion (49%) gar
```

```
lic (45%)

NORTH-AFRICAN
onion (55%) olive_oil (50%) cumin (48%) gar
lic (46%)

SCANDINAVIA
wheat (74%) butter (70%) egg (59%) cream (2
7%)

SCANDINAVIAN
butter (53%) egg (41%) vinegar (31%) cream
(31%)

SOUTH-AMERICA
onion (42%) garlic (36%) egg (34%) milk (3
1%)

SOUTHERN_SOULFOOD
butter (57%) wheat (48%) egg (41%) corn (2
9%)
```

```
SOUTHWESTERN
cayenne (81%) garlic (62%) onion (61%) cila
ntro (51%)

SPAIN
onion (61%) olive_oil (57%) garlic (50%) to
mato (42%)

SPANISH_PORTUGUESE
olive_oil (62%) garlic (57%) onion (43%) be
ll_pepper (34%)

THAI
garlic (56%) fish (54%) cayenne (46%) coria
nder (42%)

THAILAND
garlic (64%) fish (51%) cayenne (47%) soy_s
auce (44%)
```

```
UK-AND-IRELAND
wheat (60%) butter (59%) egg (48%) milk (3
8%)

VIETNAMESE
fish (78%) garlic (72%) rice (47%) vegetabl
e_oil (47%)

WESTERN
egg (51%) wheat (46%) butter (46%) black_pe
pper (36%)
```

At this point, we feel that we have understood the data well and the data is ready and is in the right format for modeling!

Thank you for completing this lab!

This notebook was created by Alex Aklson (https://www.linkedin.com/in/aklson/). We hope you found this lab session interesting. Feel free to contact us if you have any questions!

This notebook is part of the free course on **Cognitive Class** called *Data Science Methodology*. If you accessed this notebook outside the course, you can take this free self-paced course, online by clicking here (https://cocl.us/DS0103EN_LAB3_PYTHON).