COGNITIVE CLASS.ai

(https://cognitiveclass.ai)

From Understanding to Preparation

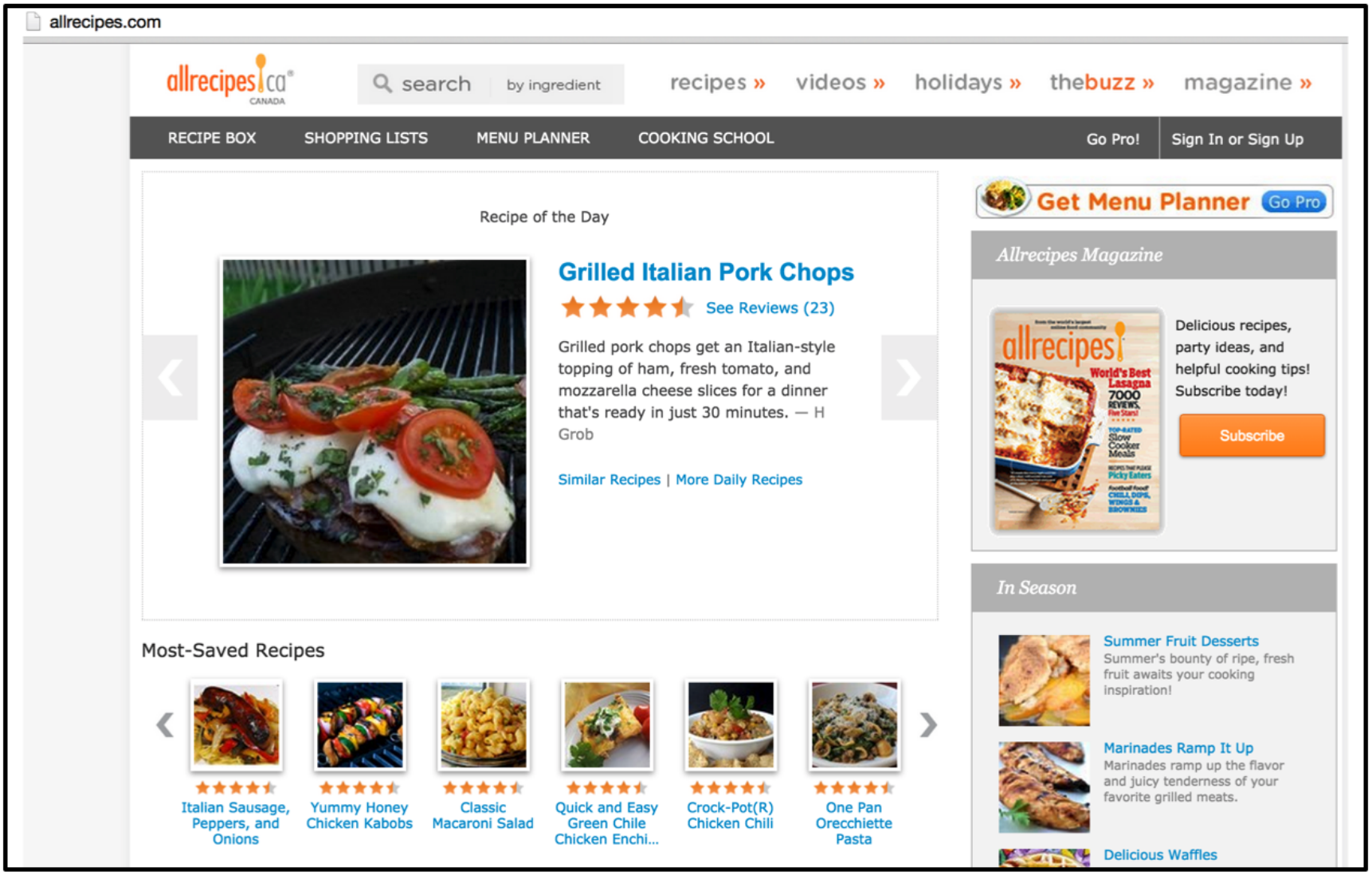
Introduction

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

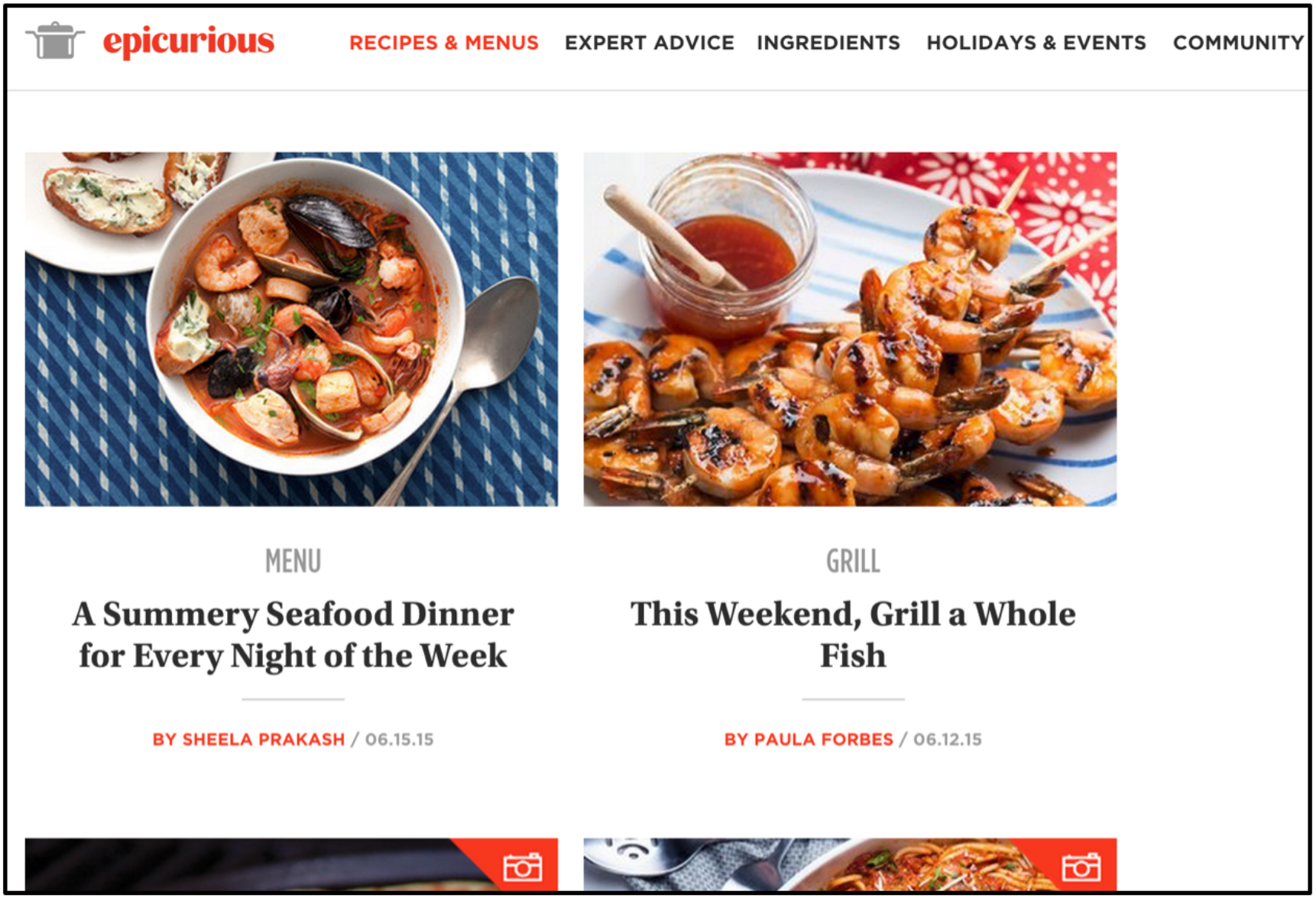
Table of Contents

Recap

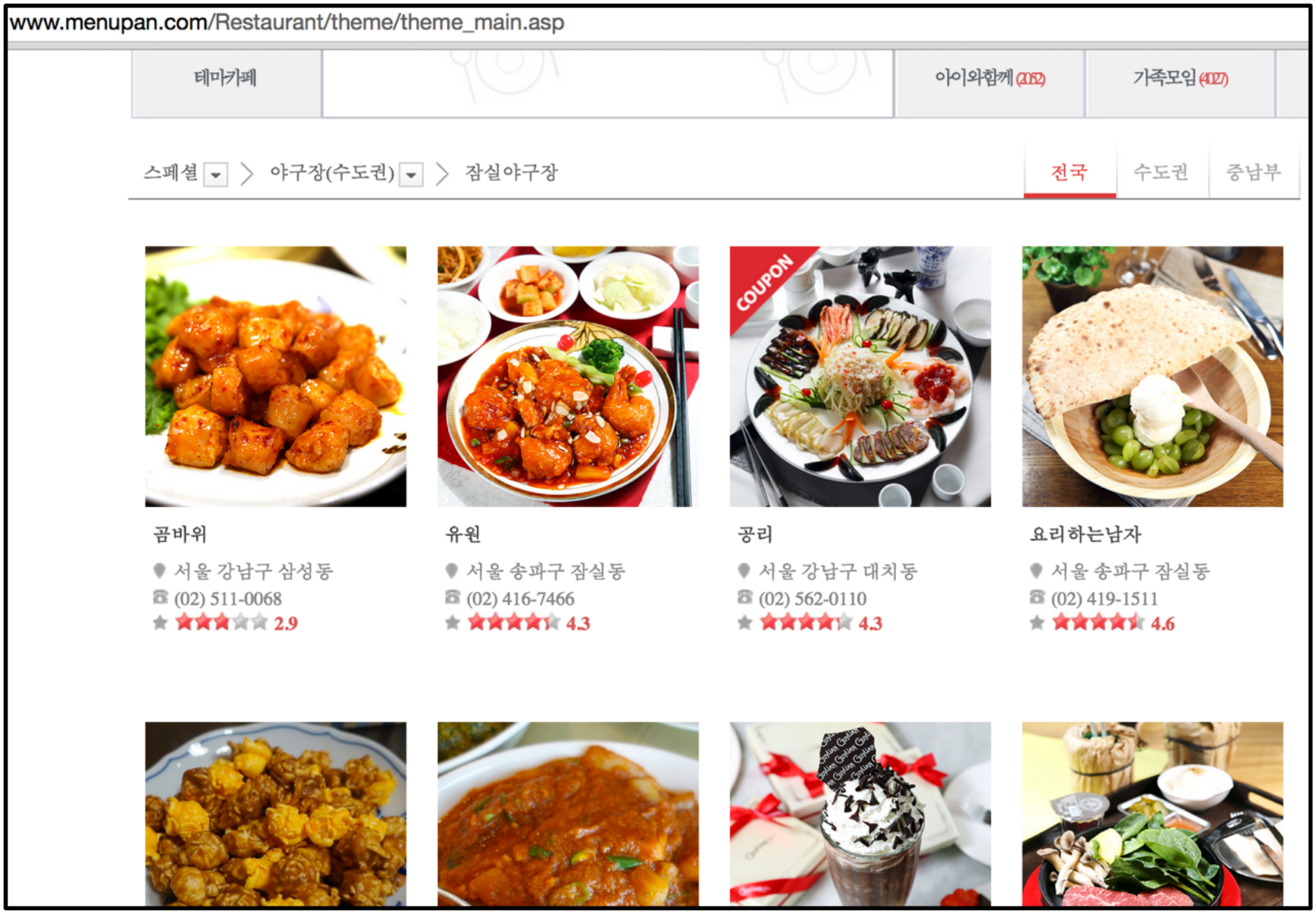
In Lab **From Requirements to Collection**, we learned that the data we need to answer the question developed in the business understanding stage, namely *can we automate the process of determining the cuisine of a given recipe?*, is readily available. A researcher named Yong-Yeol Ahn scraped tens of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



www.allrecipes.com



www.epicurious.com

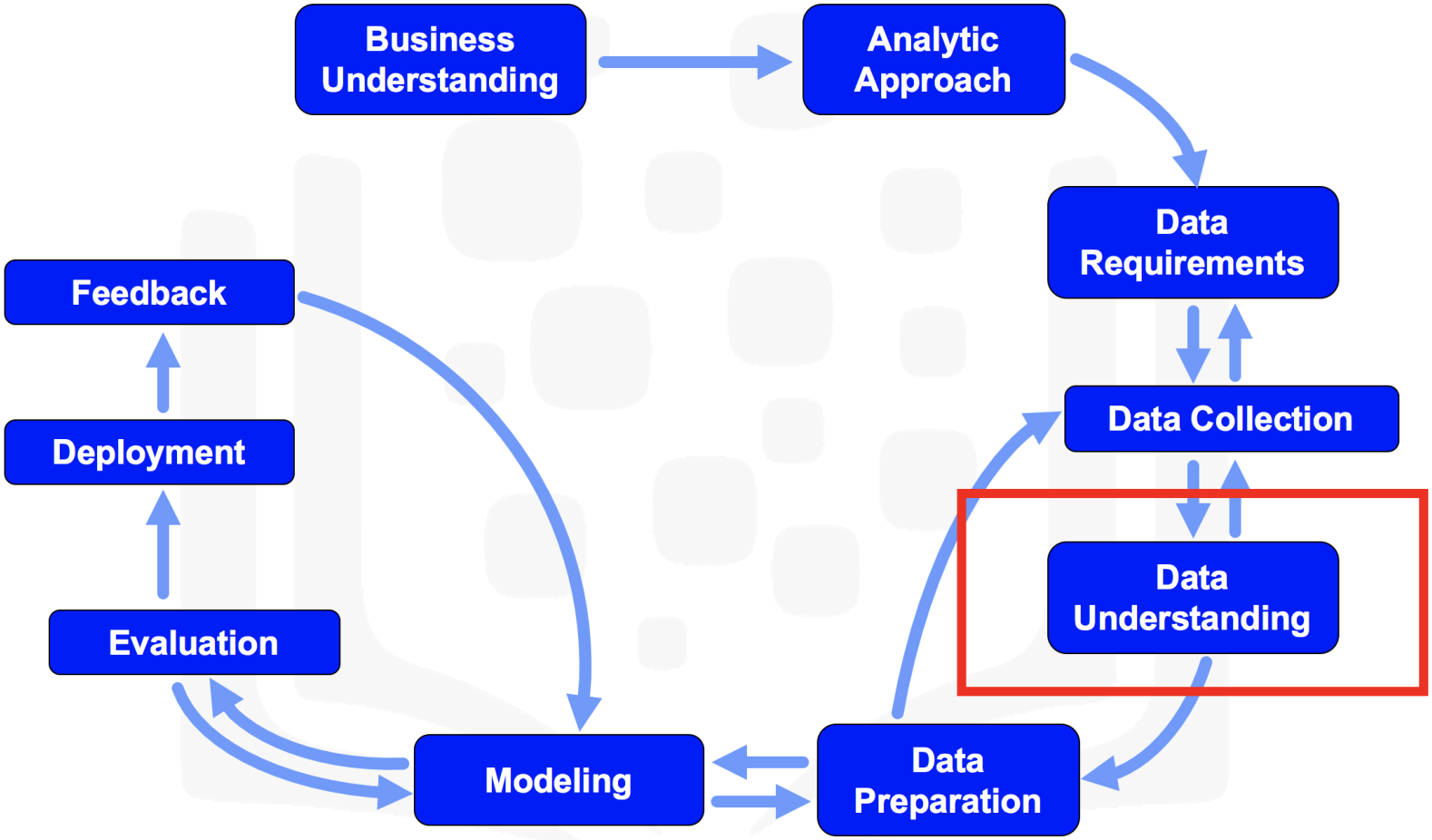


www.menupan.com

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 R. 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 R, R101 (http://cocl.us/RP0101EN_DS0103EN_LAB3_R), so please feel free to complete the course if you are interested in learning how to program in R.

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 R installed.

In [1]:

'R version 3.5.1 (2018-07-02)'

Download the data from the IBM server.

In [2]:

Show the first few rows.

In [3]:

A data.frame: 6 × 384

	country	almond	angelica	anise	anise_seed	a
	<fct>	<fct>	<fct>	<fct>	<fct>	<
1	Vietnamese	No	No	No	No	
2	Vietnamese	No	No	No	No	
3	Vietnamese	No	No	No	No	
4	Vietnamese	No	No	No	No	
5	Vietnamese	No	No	No	No	
6	Vietnamese	No	No	No	No	

Get the dimensions of the dataframe.

In [4]:

57691

In [5]:

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'

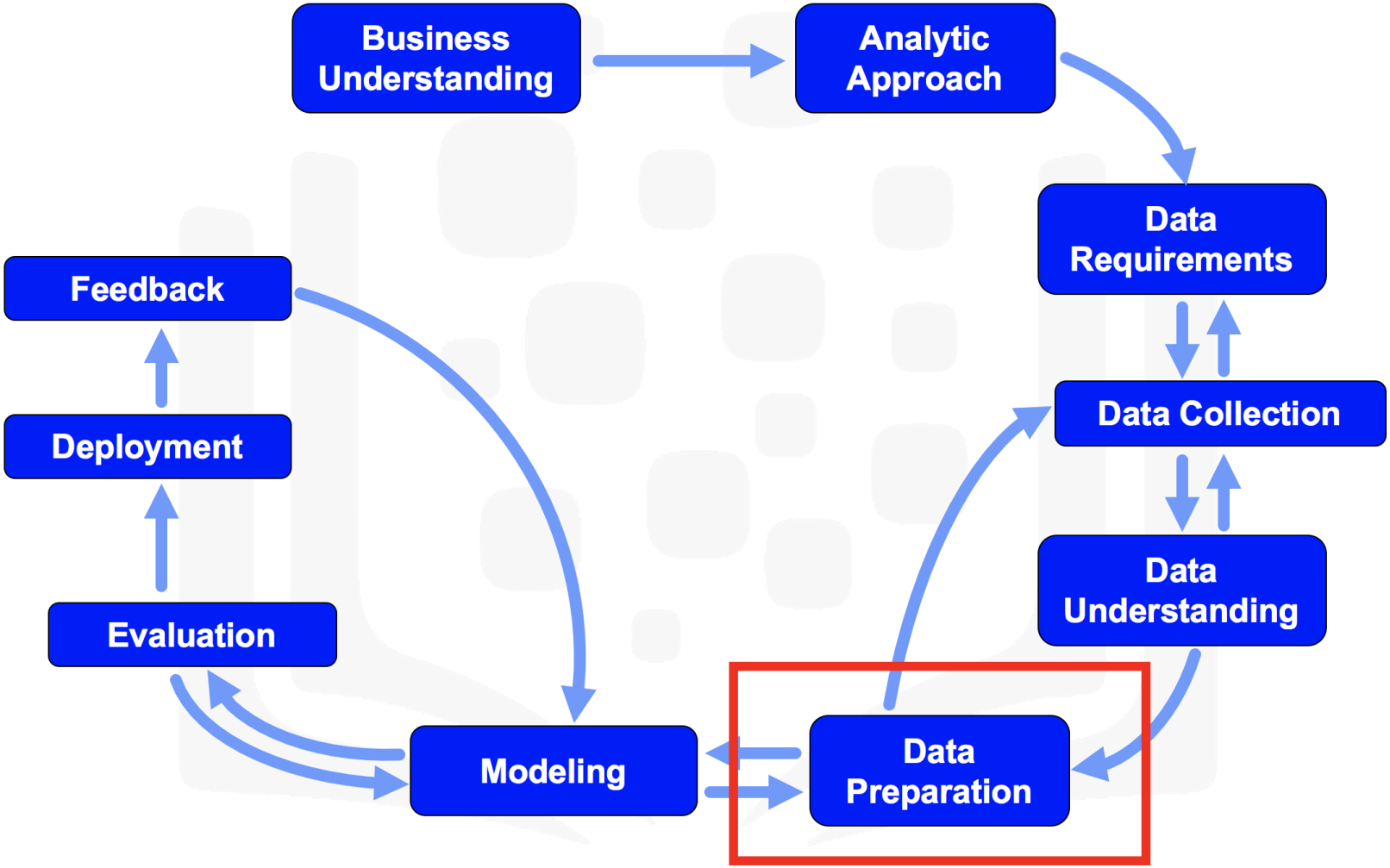
'soy_sauce' · 'soybean' · 'soybean_oil'

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 the 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 [7]:

```
                African                 Amer
ican                asian
                   115                   4
0150                  17
                  Asian                 Aus
tria              Bangladesh
                  1176
21                     4
                 Belgium            Cajun_Cr
eole                Canada
                     11
146                   774
             Caribbean   Central_SouthAmer
ican                China
                   183
241                   130
                 chinese                Chi
nese               east_asian
                     86
226                   951
           East-African          Eastern-Eu
```

```
rope EasternEuropean_Russian
                      11
235                      146
       English_Scottish                    Fr
ance                    French
                     204
268                      996
                  German                   Ger
many                    Greek
                      52
237                      225
                   India                    In
dian                Indonesia
                     324
274                       12
                    Iran                     I
rish                   Israel
                      21
86                        9
                 italian                  Ita
lian                    Italy
```

```
                    74
1715                  1461
                Japan                japa
nese              Japanese
                    85
99                    136
              Jewish                   K
orea               korean
                   320
32                    767
             Lebanon                Mala
ysia        Mediterranean
                    31
18                    289
             Mexican                  me
xico               Mexico
                   622
14                    1754
        MiddleEastern                Moro
ccan           Netherlands
                   248
```

```
137                    32
        North-African                 Paki
stan          Philippines
                   60
19                    43
             Portugal             Scandin
avia          Scandinavian
                   50
158                    92
        South-African          South-Ame
rica      Southern_SoulFood
                   16
103                   346
         Southwestern                    S
pain     Spanish_Portuguese
                  108
75                   291
           Switzerland
Thai             Thailand
                   20
164                   125
```

```
                    Turkey          UK-and-Ire
land                    Vietnam
                        16
282                         30
              Vietnamese            West-Afr
ican                    western
                        65
13                        450
```

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 [9]:

Make all the cuisine names lowercase.

In [10]:

A data.frame: 57691 × 384

cuisine	almond	angelica	anise	anise_seed	appl
<chr>	<fct>	<fct>	<fct>	<fct>	<fct
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N

cuisine	almond	angelica	anise	anise_seed	appl
<chr>	<fct>	<fct>	<fct>	<fct>	<fct
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N

cuisine	almond	angelica	anise	anise_seed	appl
<chr>	<fct>	<fct>	<fct>	<fct>	<fct
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
⋮	⋮	⋮	⋮	⋮	
japan	No	No	No	No	N

cuisine	almond	angelica	anise	anise_seed	appl
<chr>	<fct>	<fct>	<fct>	<fct>	<fct
japan	No	No	No	No	N
japan	No	No	No	No	N
japan	No	No	No	No	N
japan	No	No	No	No	N
japan	No	No	No	No	N
japan	No	No	No	No	N
japan	No	No	No	No	N
japan	No	No	No	No	N
japan	No	No	No	No	N
japan	No	No	No	No	N
japan	No	No	No	No	N

cuisine	almond	angelica	anise	anise_seed	appl
<chr>	<fct>	<fct>	<fct>	<fct>	<fct
japan	No	No	No	No	N
japan	No	No	No	No	N
japan	No	No	No	No	N
japan	No	No	No	No	N
japan	No	No	No	No	N
japan	No	No	No	No	N
japan	No	No	No	No	N
japan	No	No	No	No	N
japan	No	No	No	No	N
japan	No	No	No	No	N
japan	No	No	No	No	N

cuisine	almond	angelica	anise	anise_seed	appl
<chr>	<fct>	<fct>	<fct>	<fct>	<fct
japan	No	No	No	No	N
japan	No	No	No	No	N
japan	No	No	No	No	N
japan	No	No	No	No	N
japan	No	No	No	No	N
japan	No	No	No	No	N
japan	No	No	No	No	N

Make the cuisine names consistent.

In [11]:

A data.frame: 57691 × 384

cuisine	almond	angelica	anise	anise_seed	appl
<chr>	<fct>	<fct>	<fct>	<fct>	<fct
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N

cuisine	almond	angelica	anise	anise_seed	appl
<chr>	<fct>	<fct>	<fct>	<fct>	<fct
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N

cuisine	almond	angelica	anise	anise_seed	appl
<chr>	<fct>	<fct>	<fct>	<fct>	<fct
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
vietnamese	No	No	No	No	N
⋮	⋮	⋮	⋮	⋮	
japanese	No	No	No	No	N

cuisine	almond	angelica	anise	anise_seed	appl
<chr>	<fct>	<fct>	<fct>	<fct>	<fct
japanese	No	No	No	No	N
japanese	No	No	No	No	N
japanese	No	No	No	No	N
japanese	No	No	No	No	N
japanese	No	No	No	No	N
japanese	No	No	No	No	N
japanese	No	No	No	No	N
japanese	No	No	No	No	N
japanese	No	No	No	No	N
japanese	No	No	No	No	N
japanese	No	No	No	No	N

cuisine	almond	angelica	anise	anise_seed	appl
<chr>	<fct>	<fct>	<fct>	<fct>	<fct
japanese	No	No	No	No	N
japanese	No	No	No	No	N
japanese	No	No	No	No	N
japanese	No	No	No	No	N
japanese	No	No	No	No	N
japanese	No	No	No	No	N
japanese	No	No	No	No	N
japanese	No	No	No	No	N
japanese	No	No	No	No	N
japanese	No	No	No	No	N
japanese	No	No	No	No	N

cuisine	almond	angelica	anise	anise_seed	appl
<chr>	<fct>	<fct>	<fct>	<fct>	<fct
japanese	No	No	No	No	N
japanese	No	No	No	No	N
japanese	No	No	No	No	N
japanese	No	No	No	No	N
japanese	No	No	No	No	N
japanese	No	No	No	No	N
japanese	No	No	No	No	N

Remove cuisines with < 50 recipes:

In [12]:

```
                american                    ita
lian                mexican
                 40150
3250                    2390
               french                        a
sian             east_asian
                  1264
1193                     951
               korean                    cana
dian                indian
                   799
774                     598
              western                     chi
nese      spanish_portuguese
                   450
442                     416
          uk-and-irish          southern_soul
food                jewish
                   368
346                     329
               japanese                     ge
```

```
rman            mediterranean
                          320
289                       289
                     thai            scandina
vian            middleeastern
                          289
250                       248
  central_southamerican          eastern-eu
rope                    greek
                          241
235                       225
       english_scottish                carib
bean             cajun_creole
                          204
183                       146
easterneuropean_russian                 moro
ccan                   african
                          146
137                       115
            southwestern           south-ame
rica                vietnamese
```

```
                    108
103                  95
         north-african            philip
pine                dutch
                    60
43                  32
              lebanese               aust
rian                iranian
                    31
21                  21
                 swiss              pakis
tani              malaysian
                    20
19                  18
         south-african                tur
kish           west-african
                    16
16                  13
            indonesian                bel
gian           east-african
                    12
```

```
11
```

```
11
```

```
bangladesh
```

```
4
```

In [13]:

'american' · 'italian' · 'mexican' · 'french' · 'asian' · 'east_asian' · 'korean' · 'canadian' · 'indian' · 'western' · 'chinese' · 'spanish_portuguese' · 'uk-and-irish' · 'southern_soulfood' · 'jewish' · 'japanese' · 'german' · 'mediterranean' · 'thai' · 'scandinavian' · 'middleeastern' · 'central_southamerican' · 'eastern-europe' · 'greek' · 'english_scottish' · 'caribbean' · 'cajun_creole' · 'easterneuropean_russian' · 'moroccan' · 'african' · 'southwestern' · 'south-america' · 'vietnamese' · 'north-african'

In [14]:

```
[1] "Number of rows of original dataframe i
s 57691"
[1] "Number of rows of processed dataframe
is 57403"
[1] "288 rows removed!"
```

Convert all of the columns into factors. This is to run the classification model later.

In [15]:

A data.frame: 57403 × 384

	cuisine	almond	angelica	anise	anise_see
	<fct>	<fct>	<fct>	<fct>	<fct
1	vietnamese	No	No	No	N
2	vietnamese	No	No	No	N
3	vietnamese	No	No	No	N
4	vietnamese	No	No	No	N
5	vietnamese	No	No	No	N
6	vietnamese	No	No	No	N
7	vietnamese	No	No	No	N
8	vietnamese	No	No	No	N
9	vietnamese	No	No	No	N
10	vietnamese	No	No	No	N

	cuisine	almond	angelica	anise	anise_see
	<fct>	<fct>	<fct>	<fct>	<fct
11	vietnamese	No	No	No	N
12	vietnamese	No	No	No	N
13	vietnamese	No	No	No	N
14	vietnamese	No	No	No	N
15	vietnamese	No	No	No	N
16	vietnamese	No	No	No	N
17	vietnamese	No	No	No	N
18	vietnamese	No	No	No	N
19	vietnamese	No	No	No	N
20	vietnamese	No	No	No	N
21	vietnamese	No	No	No	N

	cuisine	almond	angelica	anise	anise_see
	<fct>	<fct>	<fct>	<fct>	<fct
22	vietnamese	No	No	No	N
23	vietnamese	No	No	No	N
24	vietnamese	No	No	No	N
25	vietnamese	No	No	No	N
26	vietnamese	No	No	No	N
27	vietnamese	No	No	No	N
28	vietnamese	No	No	No	N
29	vietnamese	No	No	No	N
30	vietnamese	No	No	No	N
⋮	⋮	⋮	⋮	⋮	
57662	japanese	No	No	No	N

	cuisine	almond	angelica	anise	anise_see
	<fct>	<fct>	<fct>	<fct>	<fct
57663	japanese	No	No	No	N
57664	japanese	No	No	No	N
57665	japanese	No	No	No	N
57666	japanese	No	No	No	N
57667	japanese	No	No	No	N
57668	japanese	No	No	No	N
57669	japanese	No	No	No	N
57670	japanese	No	No	No	N
57671	japanese	No	No	No	N
57672	japanese	No	No	No	N
57673	japanese	No	No	No	N

	cuisine	almond	angelica	anise	anise_see
	<fct>	<fct>	<fct>	<fct>	<fct
57674	japanese	No	No	No	N
57675	japanese	No	No	No	N
57676	japanese	No	No	No	N
57677	japanese	No	No	No	N
57678	japanese	No	No	No	N
57679	japanese	No	No	No	N
57680	japanese	No	No	No	N
57681	japanese	No	No	No	N
57682	japanese	No	No	No	N
57683	japanese	No	No	No	N
57684	japanese	No	No	No	N

	cuisine	almond	angelica	anise	anise_see
	<fct>	<fct>	<fct>	<fct>	<fct
57685	japanese	No	No	No	N
57686	japanese	No	No	No	N
57687	japanese	No	No	No	N
57688	japanese	No	No	No	N
57689	japanese	No	No	No	N
57690	japanese	No	No	No	N
57691	japanese	No	No	No	N

In R, you can check the structure of your data using the **str** function.
Let's check the structure of our dataframe **recipes**.

In [16]:

```
'data.frame':   57403 obs. of  384 variable
s:
 $ cuisine                : Factor w/ 34 le
vels "african","american",..: 33 33 33 33 3
3 33 33 33 33 33 ...
 $ almond                 : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ angelica               : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ anise                  : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ anise_seed             : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ apple                  : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ apple_brandy           : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ apricot                : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ armagnac               : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
```

```
 $ artemisia               : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ artichoke               : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ asparagus               : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ avocado                 : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ bacon                   : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ baked_potato            : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ balm                    : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ banana                  : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ barley                  : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ bartlett_pear           : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ basil                   : Factor w/ 2 lev
```

```
els "No","Yes": 2 1 1 2 1 2 2 1 1 1 ...
 $ bay                  : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ bean                 : Factor w/ 2 lev
els "No","Yes": 1 1 1 2 1 1 1 2 1 1 ...
 $ beech                : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ beef                 : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 2 1 1 1 1 ...
 $ beef_broth           : Factor w/ 2 lev
els "No","Yes": 1 1 1 2 1 1 1 1 1 1 ...
 $ beef_liver           : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ beer                 : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ beet                 : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ bell_pepper          : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ bergamot             : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
```

```
 $ berry                  : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ bitter_orange          : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ black_bean             : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ black_currant          : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ black_mustard_seed_oil : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ black_pepper           : Factor w/ 2 lev
els "No","Yes": 1 2 1 1 1 1 2 2 1 2 ...
 $ black_raspberry        : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ black_sesame_seed      : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ black_tea              : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ blackberry             : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ blackberry_brandy      : Factor w/ 2 lev
```

```
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ blue_cheese             : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ blueberry               : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ bone_oil                : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ bourbon_whiskey         : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ brandy                  : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ brassica                : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ bread                   : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 2 1 1 ...
 $ broccoli                : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ brown_rice              : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ brussels_sprout         : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
```

```
 $ buckwheat               : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ butter                  : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 2 1 1 ...
 $ buttermilk              : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ cabbage                 : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ cabernet_sauvignon_wine: Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ cacao                   : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ camembert_cheese        : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ cane_molasses           : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ caraway                 : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ cardamom                : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ carnation               : Factor w/ 2 lev
```

```
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ carob                 : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ carrot                : Factor w/ 2 lev
els "No","Yes": 2 1 1 1 1 1 1 2 1 1 ...
 $ cashew                : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ cassava               : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ catfish               : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ cauliflower           : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ caviar                : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ cayenne               : Factor w/ 2 lev
els "No","Yes": 2 2 1 2 2 1 2 2 2 2 ...
 $ celery                : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ celery_oil            : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
```

```
 $ cereal                  : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ chamomile               : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ champagne_wine          : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ chayote                 : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ cheddar_cheese          : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ cheese                  : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ cherry                  : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ cherry_brandy           : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ chervil                 : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ chicken                 : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 2 1 1 2 ...
 $ chicken_broth           : Factor w/ 2 lev
```

```
els "No","Yes": 1 1 1 1 1 1 1 1 1 2 ...
 $ chicken_liver          : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ chickpea               : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ chicory                : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ chinese_cabbage        : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ chive                  : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ cider                  : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ cilantro               : Factor w/ 2 lev
els "No","Yes": 2 1 1 2 1 1 1 1 1 1 ...
 $ cinnamon               : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ citrus                 : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ citrus_peel            : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
```

```
 $ clam                    : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ clove                   : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ cocoa                   : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ coconut                 : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 2 ...
 $ coconut_oil             : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ cod                     : Factor w/ 2 lev
els "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
  [list output truncated]
```

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 [17]:

A data.frame: 11 × 384

	cuisine	almond	angelica	anise	anise_seed
	<fct>	<fct>	<fct>	<fct>	<fct>
11307	japanese	No	No	No	No
11322	japanese	No	No	No	No
11362	japanese	No	No	No	No
12172	asian	No	No	No	No
12386	asian	No	No	No	No
13011	asian	No	No	No	No
13160	asian	No	No	No	No
13514	japanese	No	No	No	No
13587	japanese	No	No	No	No
13626	east_asian	No	No	No	No

	cuisine	almond	angelica	anise	anise_seed
	<fct>	<fct>	<fct>	<fct>	<fct>
14496	east_asian	No	No	No	No

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 [18]:

A data.frame: 383 × 2

	ingredient	count
	<fct>	<dbl>
2	almond	2306
3	angelica	1
4	anise	223
5	anise_seed	87
6	apple	2422
7	apple_brandy	37
8	apricot	620
9	armagnac	11
10	artemisia	13
11	artichoke	391

	ingredient	count
	<fct>	<dbl>
12	asparagus	460
13	avocado	660
14	bacon	2169
15	baked_potato	9
16	balm	3
17	banana	989
18	barley	266
19	bartlett_pear	23
20	basil	3842
21	bay	1463
22	bean	1992
23	beech	1

	ingredient	count
	<fct>	<dbl>
24	beef	4902
25	beef_broth	845
26	beef_liver	10
27	beer	307
28	beet	233
29	bell_pepper	5979
30	bergamot	7
31	berry	183
⋮	⋮	⋮
355	thyme	3043
356	tomato	9920
357	tomato_juice	176

	ingredient	count
	<fct>	<dbl>
358	truffle	52
359	tuna	463
360	turkey	901
361	turmeric	1291
362	turnip	188
363	vanilla	9010
364	veal	197
365	vegetable	1703
366	vegetable_oil	11105
367	vinegar	8060
368	violet	5
369	walnut	2729

	ingredient	count
	<fct>	<dbl>
370	wasabi	135
371	watercress	150
372	watermelon	110
373	wheat	20781
374	wheat_bread	82
375	whiskey	148
376	white_bread	370
377	white_wine	2205
378	whole_grain_wheat_flour	731
379	wine	1026
380	wood	33
381	yam	85

	ingredient	count
	<fct>	<dbl>
382	yeast	3385
383	yogurt	1033
384	zucchini	1102

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

In [19]:

A data.frame: 383 × 2

	ingredient	count
	<fct>	<dbl>
1	egg	21025
2	wheat	20781
3	butter	20719
4	onion	18080
5	garlic	17353
6	milk	12870
7	vegetable_oil	11105
8	cream	10171
9	tomato	9920
10	olive_oil	9876

	ingredient	count
	<fct>	<dbl>
11	black_pepper	9828
12	pepper	9230
13	vanilla	9010
14	cayenne	8254
15	vinegar	8060
16	cane_molasses	7741
17	bell_pepper	5979
18	cinnamon	5594
19	parsley	5552
20	chicken	5436
21	lemon_juice	5065
22	beef	4902

	ingredient	count
	<fct>	<dbl>
23	corn	4828
24	cocoa	4799
25	scallion	4782
26	bread	4571
27	ginger	4358
28	mustard	4119
29	rice	3857
30	basil	3842
⋮	⋮	⋮
354	holy_basil	3
355	hop	3
356	mutton	3

	ingredient	count
	<fct>	<dbl>
357	rapeseed	3
358	roasted_almond	3
359	jasmine_tea	2
360	laurel	2
361	long_pepper	2
362	pimenta	2
363	raw_beef	2
364	red_algae	2
365	sheep_cheese	2
366	soybean_oil	2
367	strawberry_juice	2
368	angelica	1

	ingredient	count
	<fct>	<dbl>
369	beech	1
370	emmental_cheese	1
371	geranium	1
372	jamaican_rum	1
373	kaffir_lime	1
374	lilac_flower_oil	1
375	mate	1
376	muscat_grape	1
377	pelargonium	1
378	roasted_hazelnut	1
379	roasted_nut	1
380	roasted_pecan	1

	ingredient	count
	<fct>	<dbl>
381	strawberry_jam	1
382	sturgeon_caviar	1
383	durian	0

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 [20]:

A data.frame: 6 × 384

	cuisine	almond	angelica	anise	anise_seed
	<fct>	<dbl>	<dbl>	<dbl>	<dbl>
1	african	0.1565	0	0.0000	0.0000
2	american	0.0406	0	0.0030	0.0006
3	asian	0.0075	0	0.0008	0.0025
4	cajun_creole	0.0000	0	0.0000	0.0000
5	canadian	0.0362	0	0.0000	0.0000
6	caribbean	0.0164	0	0.0109	0.0000

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 [21]:

```
AFRICAN
onion (53%)  olive_oil (52%)  garlic (50%)
cumin (43%)

AMERICAN
butter (41%)  egg (41%)  wheat (40%)  onion
(29%)

ASIAN
soy_sauce (50%)  ginger (49%)  garlic (48%)
rice (41%)

CAJUN_CREOLE
onion (70%)  cayenne (56%)  garlic (49%)  b
utter (36%)

CANADIAN
wheat (40%)  butter (38%)  egg (35%)  onion
(34%)

CARIBBEAN
```

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

CENTRAL_SOUTHAMERICAN
garlic (57%)  onion (54%)  cayenne (52%)  t
omato (41%)

CHINESE
soy_sauce (69%)  ginger (53%)  garlic (53%)
scallion (48%)

EAST_ASIAN
garlic (55%)  soy_sauce (50%)  scallion (5
0%)  cayenne (48%)

EASTERN-EUROPE
wheat (53%)  egg (52%)  butter (48%)  onion
(45%)

EASTERNEUROPEAN_RUSSIAN
butter (60%)  egg (51%)  wheat (49%)  onion
```

```
(38%)

ENGLISH_SCOTTISH
butter (67%)  wheat (62%)  egg (53%)  cream
(41%)

FRENCH
butter (50%)  egg (44%)  wheat (37%)  olive
_oil (28%)

GERMAN
wheat (65%)  egg (61%)  butter (47%)  onion
(35%)

GREEK
olive_oil (76%)  garlic (44%)  onion (36%)
lemon_juice (34%)

INDIAN
cumin (60%)  turmeric (51%)  onion (50%)  c
oriander (48%)
```

```
ITALIAN
olive_oil (61%)  garlic (53%)  tomato (39%)
onion (33%)

JAPANESE
soy_sauce (57%)  rice (44%)  vinegar (37%)
vegetable_oil (35%)

JEWISH
egg (59%)  wheat (49%)  butter (31%)  onion
(30%)

KOREAN
garlic (59%)  scallion (52%)  cayenne (52%)
soy_sauce (49%)

MEDITERRANEAN
olive_oil (80%)  garlic (51%)  onion (39%)
tomato (35%)
```

```
MEXICAN
cayenne (74%)  onion (68%)  garlic (62%)  t
omato (59%)

MIDDLEEASTERN
olive_oil (60%)  garlic (47%)  wheat (38%)
lemon_juice (36%)

MOROCCAN
olive_oil (73%)  cumin (55%)  onion (50%)
garlic (46%)

NORTH-AFRICAN
onion (55%)  olive_oil (50%)  cumin (48%)
garlic (47%)

SCANDINAVIAN
butter (64%)  wheat (58%)  egg (53%)  cream
(29%)

SOUTH-AMERICA
```

```
onion (43%)  garlic (37%)  egg (35%)  milk
(31%)

SOUTHERN_SOULFOOD
butter (58%)  wheat (49%)  egg (42%)  corn
(30%)

SOUTHWESTERN
cayenne (81%)  garlic (62%)  onion (61%)  c
ilantro (52%)

SPANISH_PORTUGUESE
olive_oil (58%)  garlic (54%)  onion (47%)
bell_pepper (35%)

THAI
garlic (60%)  fish (53%)  cayenne (47%)  ci
lantro (42%)

UK-AND-IRISH
butter (60%)  wheat (58%)  egg (48%)  milk
```

```
(33%)

VIETNAMESE
fish (74%)  garlic (73%)  rice (49%)  cayen
ne (43%)

WESTERN
egg (51%)  wheat (46%)  butter (46%)  black
_pepper (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 Polong Lin (https://ca.linkedin.com/in/polonglin) and revised 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 (http://cocl.us/DS0103EN_LAB3_R).