DS0103EN-2-2-1-From-Requirements-to-Collection-v1.0

December 20, 2018

From Requirements to Collection

0.1 Introduction

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

0.2 Table of Contents

1 Data Requirements

In the videos, we learned that the chosen analytic approach determines the data requirements. Specifically, the analytic methods to be used require certain data content, formats and representations, guided by domain knowledge.

In the **From Problem to Approach Lab**, we determined that automating the process of determining the cuisine of a given recipe or dish is potentially possible using the ingredients of the recipe or the dish. In order to build a model, we need extensive data of different cuisines and recipes.

Identifying the required data fulfills the data requirements stage of the data science methodology.

---

2 Data Collection

In the initial data collection stage, data scientists identify and gather the available data resources. These can be in the form of structured, unstructured, and even semi-structured data relevant to the problem domain.

Web Scraping of Online Food Recipes 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.

Luckily, we will not need to carry out any data collection as the data that we need to meet the goal defined in the business understanding stage is readily available.

We have already acquired the data and placed it on an IBM server. Let's download the data and take a look at it. Important note: Please note that you are not expected to know how to program in Python. The following code is meant to illustrate the stage of data collection, 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, which is also offered on Coursera. So make sure to complete the Python course if you are interested in learning how to program in Python.

2.0.1 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]: # check Python version
        !python -V
```

```
Python 3.6.6 :: Anaconda, Inc.
```

Read the data from the IBM server into a *pandas* dataframe.

```
In [2]: import pandas as pd # download library to read data into dataframe
        pd.set_option('display.max_columns', None)

        recipes = pd.read_csv("https://ibm.box.com/shared/static/5wah9atr5o1akuuavl2z9tkjzdinr1l

        print("Data read into dataframe!") # takes about 30 seconds
```

```
Data read into dataframe!
```

Show the first few rows.

```
In [3]: recipes.head()
```

```
Out[3]:       country almond angelica anise anise_seed apple apple_brandy apricot  \
        0  Vietnamese     No       No    No          No    No           No      No
        1  Vietnamese     No       No    No          No    No           No      No
        2  Vietnamese     No       No    No          No    No           No      No
        3  Vietnamese     No       No    No          No    No           No      No
        4  Vietnamese     No       No    No          No    No           No      No

          armagnac artemisia artichoke asparagus avocado bacon baked_potato balm  \
        0       No        No        No        No      No    No           No   No
```

```
1       No        No        No        No      No    No           No   No
2       No        No        No        No      No    No           No   No
3       No        No        No        No      No    No           No   No
4       No        No        No        No      No    No           No   No

  banana barley bartlett_pear basil bay bean beech beef beef_broth beef_liver  \
0     No     No            No   Yes  No   No    No   No         No         No
1     No     No            No    No  No   No    No   No         No         No
2     No     No            No    No  No   No    No   No         No         No
3     No     No            No   Yes  No  Yes    No   No        Yes         No
4     No     No            No    No  No   No    No   No         No         No

  beer beet bell_pepper bergamot berry bitter_orange black_bean black_currant  \
0   No   No          No       No    No            No         No            No
1   No   No          No       No    No            No         No            No
2   No   No          No       No    No            No         No            No
3   No   No          No       No    No            No         No            No
4   No   No          No       No    No            No         No            No

  black_mustard_seed_oil black_pepper black_raspberry black_sesame_seed  \
0                     No          No              No                No
1                     No         Yes              No                No
2                     No          No              No                No
3                     No          No              No                No
4                     No          No              No                No

  black_tea blackberry blackberry_brandy blue_cheese blueberry bone_oil  \
0        No         No                No          No        No       No
1        No         No                No          No        No       No
2        No         No                No          No        No       No
3        No         No                No          No        No       No
4        No         No                No          No        No       No

  bourbon_whiskey brandy brassica bread broccoli brown_rice brussels_sprout  \
0              No     No       No    No       No         No              No
1              No     No       No    No       No         No              No
2              No     No       No    No       No         No              No
3              No     No       No    No       No         No              No
4              No     No       No    No       No         No              No

  buckwheat butter buttermilk cabbage cabernet_sauvignon_wine cacao  \
0        No     No         No      No                       No    No
1        No     No         No      No                       No    No
2        No     No         No      No                       No    No
3        No     No         No      No                       No    No
4        No     No         No      No                       No    No

  camembert_cheese cane_molasses caraway cardamom carnation carob carrot  \
```

```
0              No            No      No       No        No    No    Yes
1              No            No      No       No        No    No     No
2              No            No      No       No        No    No     No
3              No            No      No       No        No    No     No
4              No            No      No       No        No    No     No

  cashew cassava catfish cauliflower caviar cayenne celery celery_oil cereal  \
0     No      No      No          No     No     Yes     No          No     No
1     No      No      No          No     No     Yes     No          No     No
2     No      No      No          No     No      No     No          No     No
3     No      No      No          No     No     Yes     No          No     No
4     No      No      No          No     No     Yes     No          No     No

  chamomile champagne_wine chayote cheddar_cheese cheese cherry cherry_brandy  \
0        No             No      No             No     No     No            No
1        No             No      No             No     No     No            No
2        No             No      No             No     No     No            No
3        No             No      No             No     No     No            No
4        No             No      No             No     No     No            No

  chervil chicken chicken_broth chicken_liver chickpea chicory  \
0      No      No            No            No       No      No
1      No      No            No            No       No      No
2      No      No            No            No       No      No
3      No      No            No            No       No      No
4      No      No            No            No       No      No

  chinese_cabbage chive cider cilantro cinnamon citrus citrus_peel clam clove  \
0              No    No    No      Yes       No     No          No   No    No
1              No    No    No       No       No     No          No   No    No
2              No    No    No       No       No     No          No   No    No
3              No    No    No      Yes       No     No          No   No    No
4              No    No    No       No       No     No          No   No    No

  cocoa coconut coconut_oil cod coffee cognac concord_grape condiment  \
0    No      No          No  No     No     No            No        No
1    No      No          No  No     No     No            No        No
2    No      No          No  No     No     No            No        No
3    No      No          No  No     No     No            No        No
4    No      No          No  No     No     No            No        No

  coriander corn corn_flake corn_grit cottage_cheese crab cranberry cream  \
0        No   No         No        No             No   No        No    No
1        No   No         No        No             No   No        No    No
2        No   No         No        No             No   No        No    No
3        No   No         No        No             No   No        No    No
4       Yes   No         No        No             No   No        No    No
```

```
  cream_cheese cucumber cumin cured_pork currant date dill durian eel egg  \
0           No      Yes    No         No      No   No   No     No  No  No
1           No       No    No         No      No   No   No     No  No  No
2           No       No    No         No      No   No   No     No  No  No
3           No       No    No         No      No   No   No     No  No  No
4           No      Yes    No         No      No   No   No     No  No  No

  egg_noodle elderberry emmental_cheese endive enokidake fennel fenugreek  \
0         No         No              No     No        No     No        No
1         No         No              No     No        No     No        No
2         No         No              No     No        No     No        No
3         No         No              No     No        No     No        No
4         No         No              No     No        No     No        No

  feta_cheese fig fish flower frankfurter fruit galanga gardenia garlic  \
0          No  No  Yes     No          No    No      No       No    Yes
1          No  No  Yes     No          No    No      No       No    Yes
2          No  No   No     No          No    No      No       No    Yes
3          No  No  Yes     No          No    No      No       No     No
4          No  No  Yes     No          No    No      No       No    Yes

  gelatin geranium gin ginger goat_cheese grape grape_brandy grape_juice  \
0      No       No  No     No          No    No           No          No
1      No       No  No     No          No    No           No          No
2      No       No  No     No          No    No           No          No
3      No       No  No    Yes          No    No           No          No
4      No       No  No     No          No    No           No          No

  grapefruit green_bell_pepper green_tea gruyere_cheese guava haddock ham  \
0         No               No        No             No    No      No  No
1         No               No        No             No    No      No  No
2         No               No        No             No    No      No  No
3         No               No        No             No    No      No  No
4         No               No        No             No    No      No  No

  hazelnut herring holy_basil honey hop horseradish huckleberry jamaican_rum  \
0       No      No         No    No  No          No          No           No
1       No      No         No    No  No          No          No           No
2       No      No         No    No  No          No          No           No
3       No      No         No    No  No          No          No           No
4       No      No         No    No  No          No          No           No

  japanese_plum jasmine jasmine_tea juniper_berry kaffir_lime kale  \
0            No      No          No            No          No   No
1            No      No          No            No          No   No
2            No      No          No            No          No   No
3            No      No          No            No          No   No
4            No      No          No            No          No   No
```

```
  katsuobushi kelp kidney_bean kiwi kohlrabi kumquat lamb lard laurel  \
0          No   No          No   No       No      No   No   No     No
1          No   No          No   No       No      No   No   No     No
2          No   No          No   No       No      No   No   No     No
3          No   No          No   No       No      No   No   No     No
4          No   No          No   No       No      No   No   No     No

  lavender leaf leek lemon lemon_juice lemon_peel lemongrass lentil lettuce  \
0       No   No   No    No          No         No         No     No     Yes
1       No   No   No    No          No         No         No     No      No
2       No   No   No    No          No         No         No     No      No
3       No   No   No    No          No         No         No     No      No
4       No   No   No   Yes          No         No         No     No      No

  licorice lilac_flower_oil lima_bean lime lime_juice lime_peel_oil  \
0       No               No        No   No        Yes            No
1       No               No        No   No         No            No
2       No               No        No   No        Yes            No
3       No               No        No  Yes        Yes            No
4       No               No        No   No        Yes            No

  lingonberry litchi liver lobster long_pepper lovage macadamia_nut macaroni  \
0          No     No    No      No          No     No            No       No
1          No     No    No      No          No     No            No       No
2          No     No    No      No          No     No            No       No
3          No     No    No      No          No     No            No       No
4          No     No    No      No          No     No            No       No

  mace mackerel malt mandarin mandarin_peel mango maple_syrup marjoram mate  \
0   No       No   No       No            No    No          No       No   No
1   No       No   No       No            No    No          No       No   No
2   No       No   No       No            No    No          No       No   No
3   No       No   No       No            No    No          No       No   No
4   No       No   No       No            No    No          No       No   No

  matsutake meat melon milk milk_fat mint mozzarella_cheese mung_bean  \
0        No   No    No   No       No  Yes               No        No
1        No   No    No   No       No   No               No        No
2        No   No    No   No       No   No               No        No
3        No   No    No   No       No  Yes               No        No
4        No   No    No   No       No  Yes               No        No

  munster_cheese muscat_grape mushroom mussel mustard mutton nectarine nira  \
0             No           No       No     No      No     No        No   No
1             No           No       No     No      No     No        No   No
2             No           No       No     No      No     No        No   No
3             No           No       No     No      No     No        No   No
```

```
4             No           No       No     No      No     No        No   No

  nut nutmeg oat oatmeal octopus okra olive olive_oil onion orange  \
0  No     No  No      No      No   No    No       Yes    No     No
1  No     No  No      No      No   No    No        No   Yes     No
2  No     No  No      No      No   No    No        No    No     No
3  No     No  No      No      No   No    No        No    No     No
4  No     No  No      No      No   No    No        No    No     No

  orange_flower orange_juice orange_peel oregano ouzo oyster palm papaya  \
0            No           No          No      No   No     No   No     No
1            No           No          No      No   No     No   No     No
2            No           No          No      No   No     No   No     No
3            No           No          No      No   No     No   No     No
4            No           No          No      No   No     No   No     No

  parmesan_cheese parsley parsnip passion_fruit  pea peach peanut  \
0              No      No      No            No   No    No     No
1              No      No      No            No   No    No     No
2              No      No      No            No   No    No     No
3              No      No      No            No  Yes    No     No
4              No      No      No            No   No    No    Yes

  peanut_butter peanut_oil pear pear_brandy pecan pelargonium pepper  \
0            No         No   No          No    No          No     No
1            No         No   No          No    No          No     No
2            No         No   No          No    No          No     No
3            No         No   No          No    No          No     No
4            No         No   No          No    No          No     No

  peppermint peppermint_oil pimenta pimento pineapple pistachio plum popcorn  \
0         No             No      No      No        No        No   No      No
1         No             No      No      No        No        No   No      No
2         No             No      No      No        No        No   No      No
3         No             No      No      No        No        No   No      No
4         No             No      No      No        No        No   No      No

  porcini pork pork_liver pork_sausage port_wine potato potato_chip prawn  \
0      No   No         No           No        No     No          No    No
1      No   No         No           No        No     No          No    No
2      No   No         No           No        No     No          No    No
3      No   No         No           No        No     No          No    No
4      No   No         No           No        No     No          No    No

  prickly_pear provolone_cheese pumpkin quince radish raisin rapeseed  \
0           No               No      No     No     No     No       No
1           No               No      No     No     No     No       No
2           No               No      No     No     No     No       No
```

```
3           No               No      No     No     No     No       No
4           No               No      No     No     No     No       No

  raspberry raw_beef red_algae red_bean red_kidney_bean red_wine rhubarb rice  \
0        No       No        No       No              No       No      No  Yes
1        No       No        No       No              No       No      No   No
2        No       No        No       No              No       No      No   No
3        No       No        No       No              No       No      No  Yes
4        No       No        No       No              No       No      No  Yes

  roasted_almond roasted_beef roasted_hazelnut roasted_meat roasted_nut  \
0             No           No               No           No          No
1             No           No               No           No          No
2             No           No               No           No          No
3             No          Yes               No           No          No
4             No           No               No           No          No

  roasted_peanut roasted_pecan roasted_pork roasted_sesame_seed romano_cheese  \
0             No            No           No                  No            No
1             No            No           No                  No            No
2             No            No           No                  No            No
3             No            No           No                  No            No
4             No            No           No                  No            No

  root roquefort_cheese rose rosemary rum rutabaga rye_bread rye_flour  \
0   No               No   No       No  No       No        No        No
1   No               No   No       No  No       No        No        No
2   No               No   No       No  No       No        No        No
3   No               No   No       No  No       No        No        No
4   No               No   No       No  No       No        No        No

  saffron sage sake salmon salmon_roe sassafras sauerkraut savory scallion  \
0      No   No   No     No         No        No         No     No       No
1      No   No   No     No         No        No         No     No       No
2      No   No   No     No         No        No         No     No       No
3      No   No   No     No         No        No         No     No       No
4      No   No   No     No         No        No         No     No      Yes

  scallop sea_algae seaweed seed sesame_oil sesame_seed shallot sheep_cheese  \
0      No        No      No  Yes         No          No      No           No
1      No        No      No  Yes         No          No      No           No
2      No        No      No   No         No          No      No           No
3      No        No      No   No         No          No     Yes           No
4      No        No      No   No         No          No      No           No

  shellfish sherry shiitake shrimp smoke smoked_fish smoked_salmon  \
0        No     No      Yes    Yes    No          No            No
1        No     No       No     No    No          No            No
```

```
2        No     No       No     No    No          No            No
3        No     No       No     No    No          No            No
4        No     No       No    Yes    No          No            No

  smoked_sausage sour_cherry sour_milk soy_sauce soybean soybean_oil  \
0             No          No        No        No      No          No
1             No          No        No        No      No          No
2             No          No        No       Yes      No          No
3             No          No        No        No      No          No
4             No          No        No        No      No          No

  spearmint squash squid star_anise starch strawberry strawberry_jam  \
0        No     No    No         No     No         No             No
1        No     No    No         No     No         No             No
2        No     No    No         No     No         No             No
3        No     No    No         No     No         No             No
4        No     No    No         No     No         No             No

  strawberry_juice sturgeon_caviar sumac sunflower_oil sweet_potato  \
0               No              No    No            No          No
1               No              No    No            No          No
2               No              No    No            No          No
3               No              No    No            No          No
4               No              No    No            No          No

  swiss_cheese tabasco_pepper tamarind tangerine tarragon tea tequila  \
0           No             No       No        No       No  No      No
1           No             No       No        No       No  No      No
2           No             No       No        No       No  No      No
3           No             No       No        No       No  No      No
4           No             No       No        No       No  No      No

  thai_pepper thyme tomato tomato_juice truffle tuna turkey turmeric turnip  \
0          No    No     No           No      No   No     No       No     No
1          No    No     No           No      No   No     No       No     No
2         Yes    No     No           No      No   No     No       No     No
3          No    No     No           No      No   No     No       No     No
4          No    No     No           No      No   No     No       No     No

  vanilla veal vegetable vegetable_oil vinegar violet walnut wasabi  \
0      No   No        No            No     Yes     No     No     No
1      No   No        No            No      No     No     No     No
2      No   No        No            No      No     No     No     No
3      No   No        No           Yes      No     No     No     No
4      No   No        No            No     Yes     No     No     No

  watercress watermelon wheat wheat_bread whiskey white_bread white_wine  \
0         No         No    No          No      No          No         No
```

```
1          No          No     No           No       No           No          No
2          No          No     No           No       No           No          No
3          No          No     No           No       No           No          No
4          No          No     No           No       No           No          No

  whole_grain_wheat_flour wine wood yam yeast yogurt zucchini
0                      No   No   No  No    No     No       No
1                      No   No   No  No    No     No       No
2                      No   No   No  No    No     No       No
3                      No   No   No  No    No     No       No
4                      No   No   No  No    No     No       No
```

Get the dimensions of the dataframe.

```
In [4]: recipes.shape

Out[4]: (57691, 384)

In [ ]:
```

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.

Now that the data collection stage is complete, data scientists typically use descriptive statistics and visualization techniques to better understand the data and get acquainted with it. Data scientists, essentially, explore the data to:

- understand its content,
- assess its quality,
- discover any interesting preliminary insights, and,
- determine whether additional data is necessary to fill any gaps in the data.

2.0.2 Thank you for completing this lab!

This notebook was created by Alex Aklson. I hope you found this lab session interesting. Feel free to contact me if you have any questions!

This notebook is part of a course on **Coursera** called *Data Science Methodology*. If you accessed this notebook outside the course, you can take this course, online by clicking here.