Recipes Dataset EDA

제로베이스 딥러닝 2차 프로젝트

출처 - **Food.com Recipes and Interactions (kaggle)**

Food.com Recipes and Interactions

Crawled data from Food.com (GeniusKitchen) online recipe aggregator

https://www.kaggle.com/datasets/shuyangli94/food-com-recipes-and-user-interactions/code

프로젝트 목표

- 자취생과 주부들을 위한 냉장고에 있는 식품을 처리할 수 있는 레시피 찾기

EDA 목표

- DL 학습에 최적화된 클래스 찾기
- output으로 낼 수 있는 컬럼 생성

DATA 형태

- 총 230,319개의 행과 12개의 컬럼을 보유
- ingredients는 리스트의 형태로 갖고 있습니다.
- Columns
  - **name - 요리 이름**
  - **id - id**
  - **minutes - 요리에 걸리는 시간**
  - **contributor_id - 작성자 id**
  - **submitted - 작성 시간**
  - **tags - tag**
  - **nutrition - 영양소**
  - **n_steps - 요리 단계 수**
  - **steps - 요리 방법**
  - **description - 요리에 대한 설명**
  - **ingredients - 재료**
  - **n_ingredients - 재료의 수**

컬럼 정리

전체 컬럼

	name	id	minutes	n_steps	steps	description	ingredients	n_ingredients	calories	total fat	sugar	sodium	protein	saturated fat	carbohydrates
0	arriba baked winter squash mexican style	137739	55	11	['make a choice and proceed with recipe', 'dep...	autumn is my favorite time of year to cook! th...	['winter squash', 'mexican seasoning', 'mixed ...	7	51.5	0.0	13.0	0.0	2.0	0.0	4.0
1	a bit different breakfast pizza	31490	30	9	['preheat oven to 425 degrees f', 'press dough...	this recipe calls for the crust to be prebaked...	['prepared pizza crust', 'sausage patty', 'egg...	6	173.4	18.0	0.0	17.0	22.0	35.0	1.0
2	all in the kitchen chili	112140	130	6	['brown ground beef in large pot', 'add choppe...	this modified version of 'mom's' chili was a h...	['ground beef', 'yellow onions', 'diced tomato...	13	269.8	22.0	32.0	48.0	39.0	27.0	5.0
3	alouette potatoes	59389	45	11	['place potatoes in a large pot of lightly sal...	this is a super easy, great tasting, make ahea...	['spreadable cheese with garlic and herbs', 'n...	11	368.1	17.0	10.0	2.0	14.0	8.0	20.0
4	amish tomato ketchup for canning	44061	190	5	['mix all ingredients& boil for 2 1 / 2 hours ...	my dh's amish mother raised him on this recipe...	['tomato juice', 'apple cider vinegar', 'sugar...	8	352.9	1.0	337.0	23.0	3.0	0.0	28.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
231632	zydeco soup	486161	60	7	['heat oil in a 4-quart dutch oven', 'add cele...	this is a delicious soup that i originally fou...	['celery', 'onion', 'green sweet pepper', 'gar...	22	415.2	26.0	34.0	26.0	44.0	21.0	15.0
231633	zydeco spice mix	493372	5	1	['mix all ingredients together thoroughly']	this spice mix will make your taste buds dance!	['paprika', 'salt', 'garlic powder', 'onion po...	13	14.8	0.0	2.0	58.0	1.0	0.0	1.0
231634	zydeco ya ya deviled eggs	308080	40	7	['in a bowl , combine the mashed yolks and may...	deviled eggs, cajun-style	['hard-cooked eggs', 'mayonnaise', 'dijon must...	8	59.2	6.0	2.0	3.0	6.0	5.0	0.0
231635	cookies by design cookies on a stick	298512	29	9	['place melted butter in a large mixing bowl a...	i've heard of the 'cookies by design' company,...	['butter', 'eagle brand condensed milk', 'ligh...	10	188.0	11.0	57.0	11.0	7.0	21.0	9.0
231636	cookies by design sugar shortbread cookies	298509	20	5	['whip sugar and shortening in a large bowl , ...	i've heard of the 'cookies by design' company,...	['granulated sugar', 'shortening', 'eggs', 'fl...	7	174.9	14.0	33.0	4.0	4.0	11.0	6.0

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 231637 entries, 0 to 231636
Data columns (total 12 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   name            231636 non-null  object
 1   id              231637 non-null  int64
 2   minutes         231637 non-null  int64
 3   contributor_id  231637 non-null  int64
 4   submitted       231637 non-null  object
 5   tags            231637 non-null  object
 6   nutrition       231637 non-null  object
 7   n_steps         231637 non-null  int64
 8   steps           231637 non-null  object
 9   description     226658 non-null  object
 10  ingredients     231637 non-null  object
 11  n_ingredients   231637 non-null  int64
dtypes: int64(5), object(7)
memory usage: 21.2+ MB
```

- nuturition 컬럼
  - 'calories','total fat','sugar','sodium','protein','saturated fat','carbohydrates'
  - 리스트에 담겨있는 순서대로 위와 같은 컬럼을 생성하여 값 출력
- minutes 컬럼
  - 컬럼을 확인을 하였을 때 0인 값이 있어
  - 15분을 최저로 잡아 15분으로 모두 대체 하였습니다.
- 필요없는 컬럼 삭제
  - 'contributor_id','submitted','tags','nutrition'
  - 위의 4개의 컬럼은 삭제 하였습니다.

Name 컬럼

```
name
crock pot lemon garlic chicken                  3
gluten free chocolate chip cookies              3
chocolate peanut butter cookies                 3
three bean chili                                3
pop up rolls                                    3
                                                ..
easy pineapple cake                             1
easy pineapple chicken                          1
easy pineapple chili                            1
easy pineapple dessert                          1
cookies by design   sugar shortbread cookies    1
Name: count, Length: 230185, dtype: int64
```

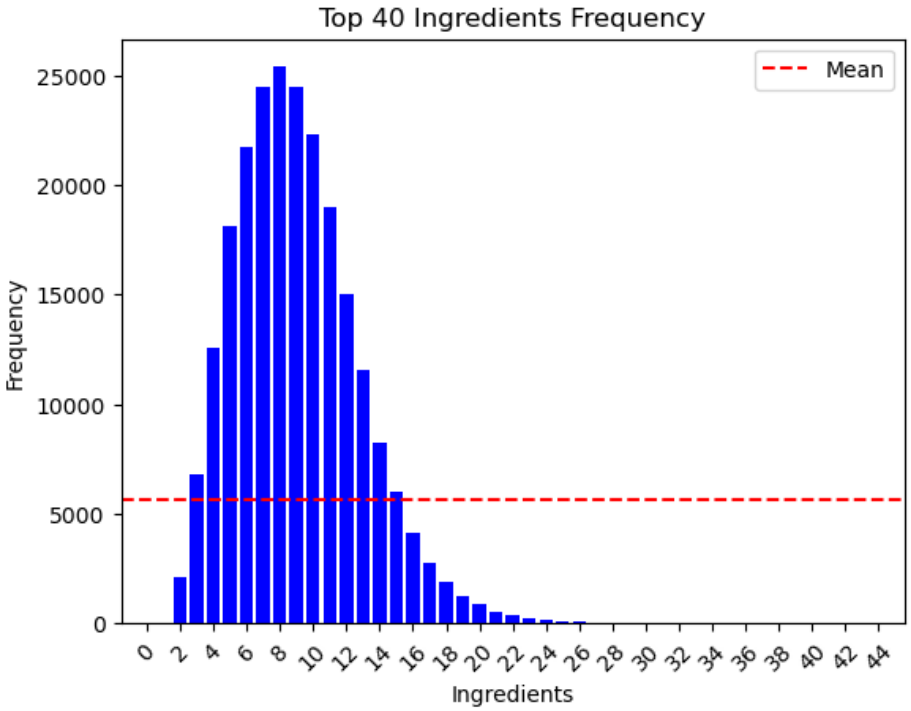
똑같은 요리 이름이 존재를 하여 확인 시 거의 동일한 것을 볼 수 있었기에, 요리 순서(n_steps)가 가장 짧은 것으로 남겨두었습니다. 그럼에도 동일한 순서를 갖고 있는 요리가 존재하여 가장 위에 있는 요리만 남겨두고 삭제 하였습니다.

Ingredients 컬럼

df4

0	1	2	3	4	5	6	7
winter squash	mexican seasoning	mixed spice	honey	butter	olive oil	salt	
prepared pizza crust	sausage	eggs	milk	salt	cheese		
beef	onion	tomatoes	tomato paste	soup	tomatoes	kidney beans	water
garlic	potatoes	shallots	parsley	tarragon	olive oil	vinegar	salt
tomato juice	vinegar	sugar	salt	pepper	clove oil	cinnamon oil	mustard
milk	vanilla ice cream	frozen apple juice concentrate	apple				
fennel seeds	green olives	ripe olives	garlic	pepper	orange rind	orange juice	red chile
pork	soy sauce	garlic	ginger	chili powder	pepper	salt	cilantro leaves
chocolate sandwich style cookies	chocolate syrup	vanilla ice cream	bananas	strawberry ice cream	whipped cream		
sugar	salt	bananas	eggs	lemon juice	orange rind	flour	baking soda
whole berry cranberry sauce	sour cream	radish					
vanilla wafers	butter	sugar	eggs	whipping cream	strawberry	walnuts	
great northern bean	cube	sugar	molasses	cornstarch	onion	garlic	mustard powder
collard greens	sugar	molasses	hot sauce	whiskey	ham hock	salt	
gentian root	scullcap herb	burnet root	wood bethony	spearmint			

리스트의 형태를 풀어 숫자로 만들어진 컬럼에 1개씩 재료를 담았다.



최대 요리 재료의 갯수가 44개가 되어 재료 수(n_ingredients)가 동일한 값들을 카운트하여 평균값을 기준으로 재료가 14개가 들어가는 행만 두고 나머지 행은 삭제를 하였습니다.

각 재료들은 여러가지 수식어와 조합이 되어 있어 동일한 제품들은 클래스화 시키기 위해 간단하게 통일하게 만들어야 했다.

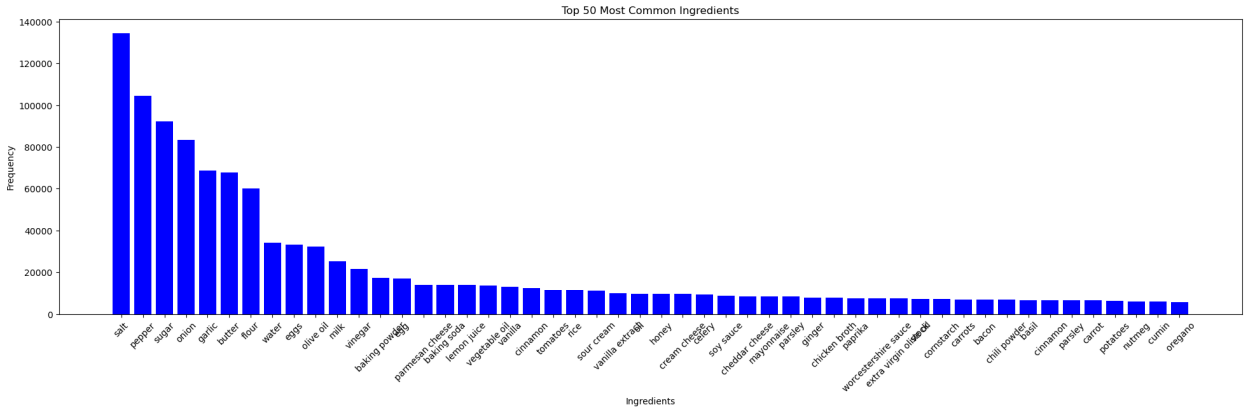
총 19개의 재료들을 키워드를 포함하고 있다면 각 키워드의 값으로만 넣었습니다.

- 공통 재료
  - salt, flour, vinegar, butter, garlic, pepper, cubes, stock, sugar, noodle, rice

    - milk, tofu, oil
- 육류
    - rib, cheese

또한 전처리시 의미없는 수식어가 들어간 단어에서 수식어만 제거를 했습니다.

- fresh, ground, dry, stem, bulb, floret



이후 가장 많이 나온 값들의 분포도를 살폈을 때 x축의 값들이 한층 정리가 되어 보였습니다.

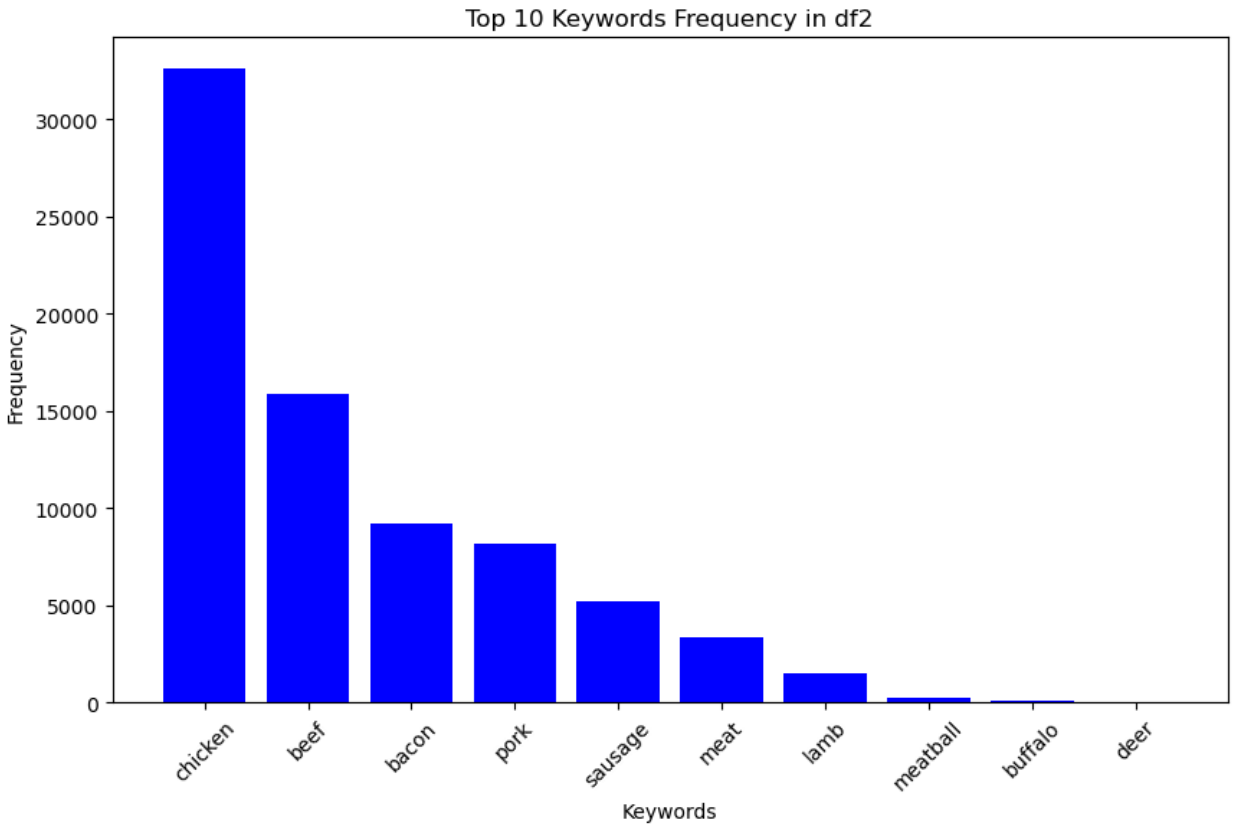
---

herb, oil, powder, paste

데이터에서 주 재료는 아니지만 많은 부분들을 차지하고 있는 4가지의 재료 분류가 있어 확인을 위해 각 컬럼에 고유값을 확인했습니다.

따로 처리는 하지 않았으며, 클래스는 주 재료(육류, 채소, 과일, 디저트)로만 지정했습니다.

Meat



육류 별 카운트를 하여 분포를 확인 했을 시에 chicken이 가장 많았습니다.

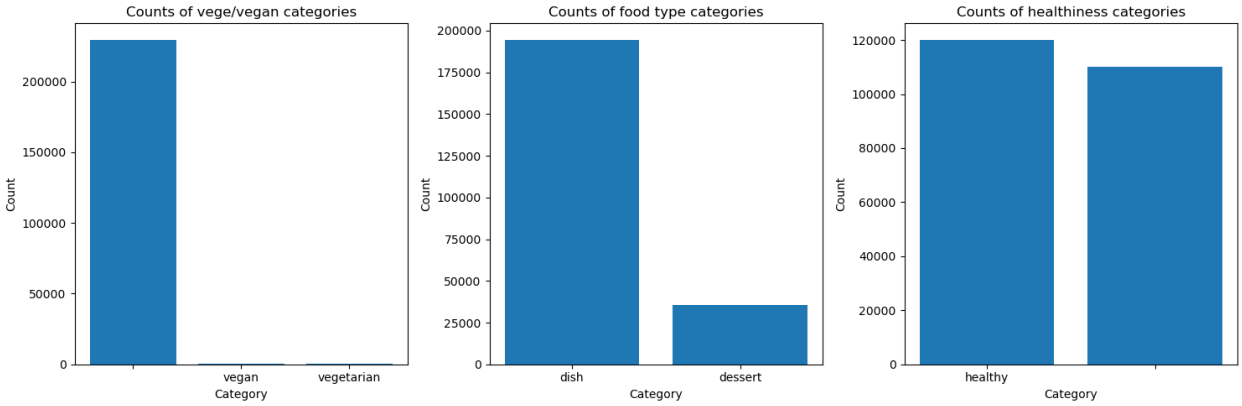
다만, chicken broth, chicken stock, chicken noodle 등 클래스화 하는 육류 chicken이 아닌 재료가 있어 추가적으로 하나의 값으로 통일 했습니다.

- 추가 통일 값
  - broth, soup, meatball, rib

Type 설정

- output에 도움이 될 수 있다고 생각이 될 수 있는 컬럼을 설정 하였습니다.
- vege/vegan
  - 재료에 vegan과 vegetarian이 수식어로 붙어있습니다.
  - 해당 컬럼에 vegan과 vegetarian이 있다면 컬럼에 해당 값을 집어 넣었습니다.
  - 해당 되지 않는 행은 빈값으로 두었습니다.
- food type
  - dessert에 흔히 사용되는 단어들을 리스트에 담았습니다.
  - 해당이 되는 단어가 있다면 컬럼에 dessert 아니라면 dish로 구분했습니다.
- healthiness

- 칼로리 컬럼이 300이 넘지 않는다면 건강한 음식으로 분류
- healthy 그렇지 않다면 빈값으로 두었습니다.



Class

모든 전 처리는 동일합니다.

- keyword를 해당 육류로 설정하여 데이터프레임에 있는 값들을 찾기
- part들을 모은 리스트를 만들어 같은 값이 있다면 새로운 리스트에 담기
- 데이터프레임에는 하나의 값만 담기

Meat 종류

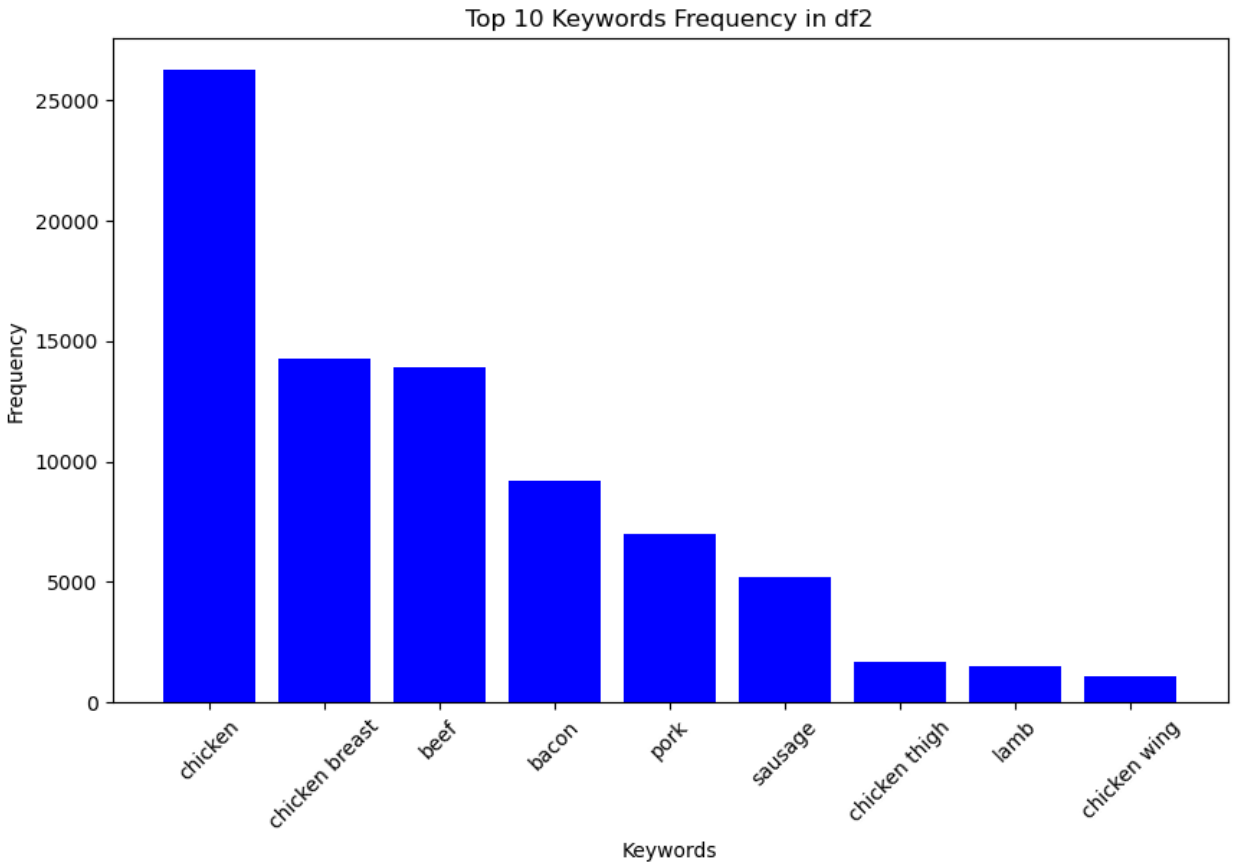
종류를 확인 해봤을시 pork, beef, lamb, sausage, bacon, deer, alligator, turtle 등 다양하였지만 가장 많이 분포되어 있는 pork, beef, lamb, sausage, bacon 만 사용하기로 했습니다.

또한 육류의 경우 생고기는 형태가 비슷하여 구분하기 힘들어 하나의 종류만 사용하였고 rib은 육류 종류 상관없이 rib으로 하나의 클래스를 만들었습니다.

총 10개의 클래스

- sausage → sausage
- bacon → bacon
- chicken → chicken breast, chicken thigh, chicken wing
- beef → beef
- pork → pork
- lamb → lamb
- rib → rib
- ham → ham

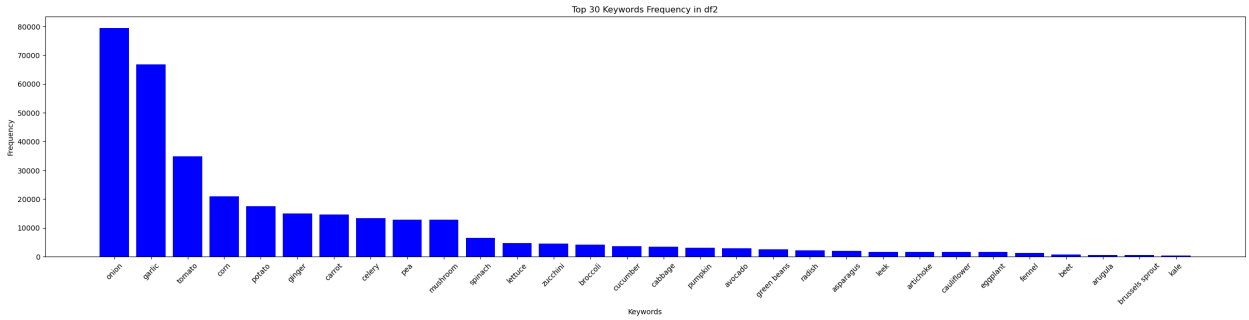
- meat
  - 나머지의 전처리를 거치고 난 이후 meat를 가지고 있는 행의 갯수를 확인했을 시 2791개의 행을 확인했습니다.
  - 고윳값들이 위에 언급한 소수의 종류의 고기들이었기 때문에 행 삭제를 하였습니다.
  - 행은 총 230,319의 행이 남았습니다.



불 필요한 수식어와 단어들을 삭제를 하고 난 이후 chicken의 값이 많이 줄었습니다. 육류로만 분류가 되어 있는게 아닌 broth, stock 처럼 다양한 재료가 있었다는 것을 알 수 있습니다.

주 재료 (Vegetable)

> 💡 채소, 과일, 디저트 재료들은 따로 분류할 수 있는 방법이 없어, 재료에 가장 많이 사용되는 분류별 값들을 리스트에 담아 데이터프레임의 값들과 동일한 값들만 남기어 확인했습니다.



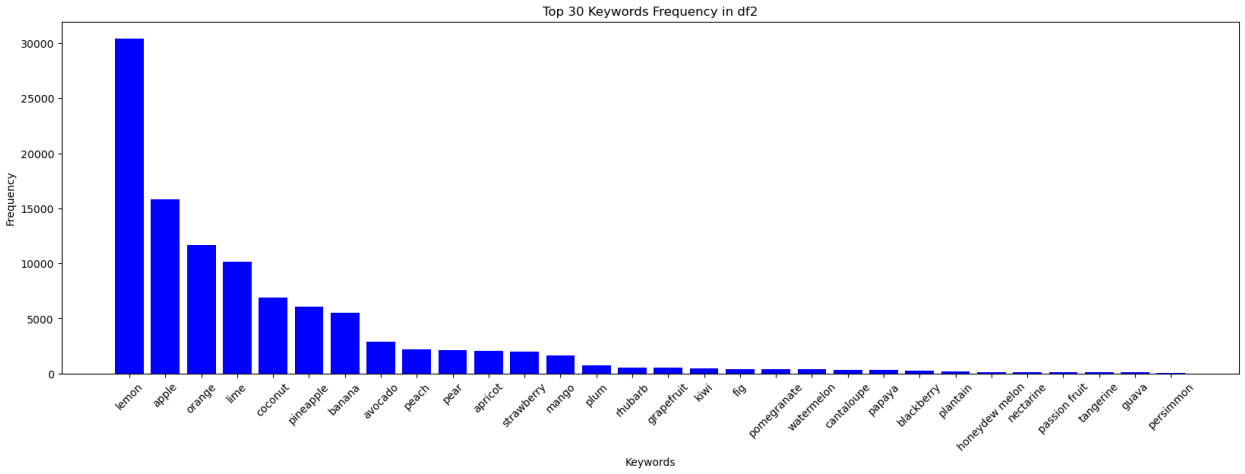
해당 키워드들을 모두 담아 roboflow의 클래스들과 비교하며 선정하였습니다.

총 27개의 클래스

- Onion
- Garlic
- Tomato
- Corn
- Potato
- Ginger
- Carrot
- Pea
- Mushroom
- Celery
- Spinach
- Zucchini
- Brocoli
- Cucumber
- Cabbage
- Pumpkin
- Avocado
- Green beans
- Radish
- Asparagus
- Leek
- Cauliflower
- Eggplant
- Fennel

- Cabbage
- Lettuce
- Chilli
- Sweet Potato

주 재료 (Fruits)



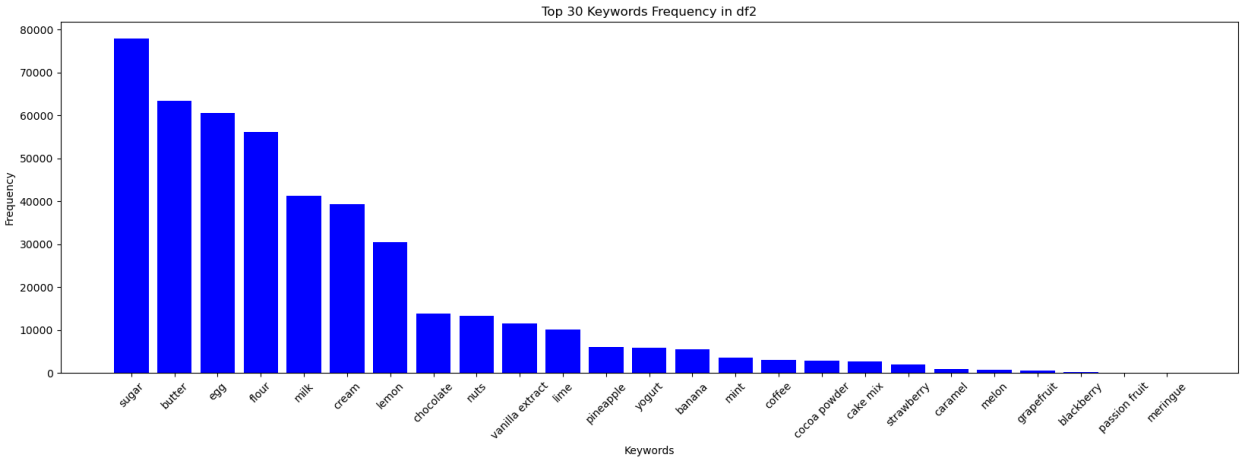
해당 키워드들을 모두 담아 roboflow의 클래스들과 비교하며 선정하였습니다.

총 17개 클래스

- Lemon
- Apple
- Orange
- Lime
- Coconut
- Pineapple
- Banana
- Avocado
- Peach
- Pear
- Apricot
- Strawberry
- Mango
- Grapefurit

- Watermelon
- Plantain
- Tangerine

주 재료 (Dessert)



해당 키워드들을 모두 담아 roboflow의 클래스들과 비교하며 선정하였습니다.

총 3개 클래스

- Cream
- Chocolate
- Yogurt

공통 재료

공통으로 들어가는 재료

총 11개 클래스

- Sugar
- Butter
- Cheese
- Egg
- Flour
- Milk
- Rice
- Tofu
- Noodle

- Oil
- Salt

최종 정리

- 데이터의 최종 형태는 아래와 같습니다.
  - 행 - 230,185
  - 컬럼 (33개)
    - **name - 요리 이름**
    - **id -id**
    - **minutes - 요리에 걸리는 시간**
    - **n_steps - 요리 단계 수**
    - **steps - 요리 방법**
    - **description - 요리에 대한 설명**
    - **n_ingredients - 재료의 수**
    - **8~14번째 - 영양소들**
    - **15~30번째 - 재료들**
    - **vege/vegan - 채식주의 음식 구분**
    - **food type - 요리, 디저트 구분**
    - **healthiness - 건강식 구분**
    - **ingredients - 재료 (원본)**
- 클래스
  - 총 68개의 클래스 입니다.

```
final_class = [
    "sausage", "bacon", "chicken breast", "chicken thigh", "chicken wing", "beef", "pork", "lamb", "rib", "ham",
    "onion", "garlic", "tomato", "corn", "potato", "ginger", "carrot", "pea", "mushroom", "celery", "spinach",
    "zucchini", "broccoli", "cucumber", "cabbage", "pumpkin", "avocado", "green beans", "radish", "asparagus",
    "leek", "cauliflower", "eggplant", "fennel", "lettuce", "chilli", "sweet potato", "lemon", "apple", "orange",
    "lime", "coconut", "pineapple", "banana", "avocado", "peach", "pear", "apricot", "strawberry", "mango",
    "grapefruit", "watermelon", "plantain", "tangerine", "cream", "chocolate", "yogurt", "sugar", "butter",
    "cheese", "egg", "flour", "milk", "rice", "tofu", "noodle", "oil", "salt"
]
```

이미지 데이터셋 (Roboflow)

ingredients Object Detection Dataset (v1, 2023-07-25 6:38pm) by Wonkeun Jung

8866 open source ingredients images and annotations in multiple formats for training computer vision models. ingredients (v1, 2023-07-25 6:38pm), created by Wonkeun Jung

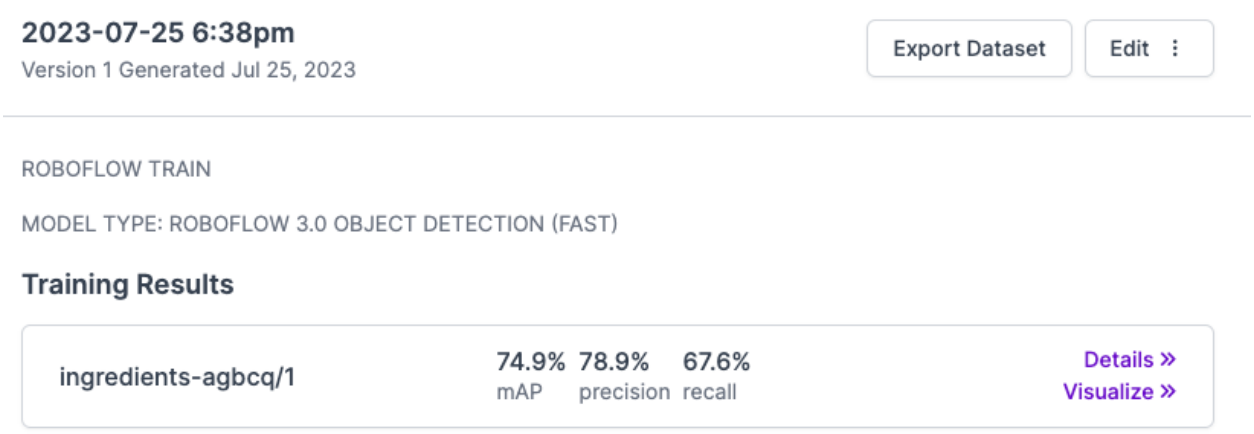
https://universe.roboflow.com/wonkeun-jung-vfcwn/ingredients-agbcq/dataset/1

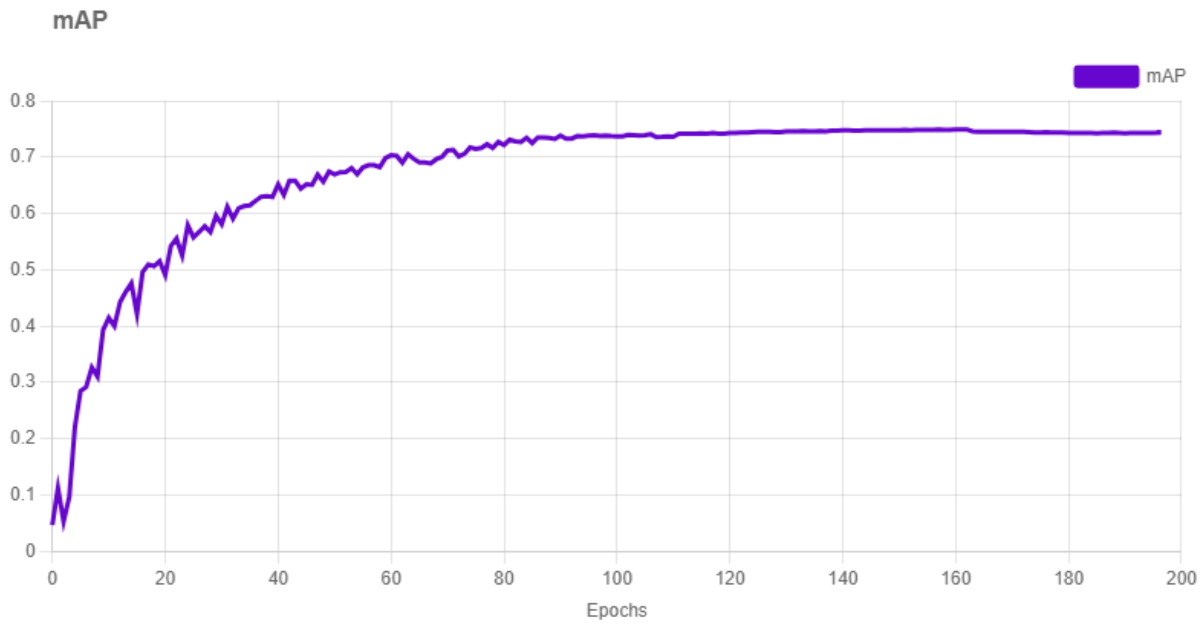
Roboflow와 kaggle데이터의 클래스 빈도수와 비교하며 정하여 새로 만들었습니다.

Roboflow에서 class명이 다른 것들 전처리를 통하여 위의 클래스 리스트와 동일시 하였습니다.

아래는 사이트에서 제공하는 프로그램을 돌린 결과입니다. 전처리를 클래스명 제외하고 하지 이미지에 대한 처리는 하지 않았습니다.

- **Modify Classes:** 90 remapped, 12 dropped
- **총 이미지 수 : 8866장**
- **Training Set : 6.4k (72%)**
- **Validation Set : 1.4k (16%)**
- **Testing Set : 1.1k (13%)**




mAP
mAP
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
20
40
60
80
100
120
140
160
180
200
Epochs


Box Loss
Box Loss
1.9
1.8
1.7
1.6
1.5
1.4
1.3
1.2
1.1
0
50
100
150
200
Epochs

Class Loss
Class Loss
8
7
6
5
4
3
2
1
0
50
100
150
200
Epochs

Object Loss
Object Loss
2.1
2.0
1.9
1.8
1.7
1.6
1.5
1.4
1.3
0
50
100
150
200
Epochs