

# Презентация к проекту на тему: What`s cooking?

Выполнила: Вдовина Юлия




## Цели

- Разработать модель для решения задачи многоклассовой классификации.

## Задачи


- Изучить и проанализировать данные
- Выбрать модель, удовлетворяющую условию и данным
- Получить предсказания модели для тестовой выборки
- Сформулировать выводы

В задании использовался датасет из соревнования с платформы Kaggle (Соревнование)

 Playground Code Competition

## What's Cooking? (Kernels Only)

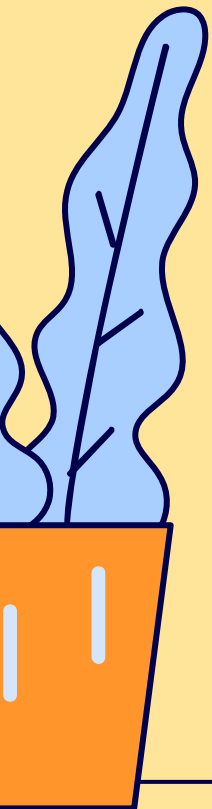
Use recipe ingredients to categorize the cuisine

 Kaggle · 520 teams · 3 years ago

[Overview](#) [Data](#) [Code](#) [Discussion](#) [Leaderboard](#) [Rules](#) [Team](#) [My Submissions](#) [Late Submission](#)

Overview

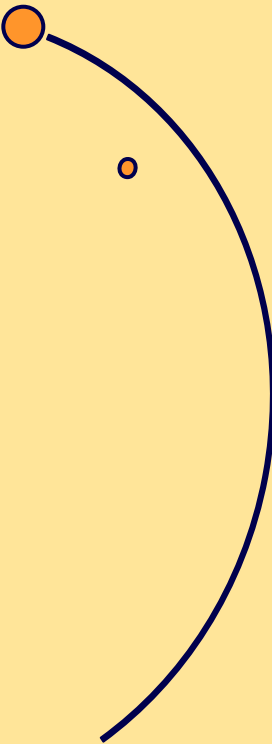
Description	<p><i>Picture yourself strolling through your local, open-air market... What do you see? What do you smell? What will you make for dinner tonight?</i></p>
Evaluation	<p>If you're in Northern California, you'll be walking past the inevitable bushels of leafy greens, spiked with dark purple kale and the bright pinks and yellows of chard. Across the world in South Korea, mounds of bright red kimchi greet you, while the smell of the sea draws your attention to squids squirming nearby. India's market is perhaps the most colorful, awash in the rich hues and aromas of dozens of spices: turmeric, star anise, poppy seeds, and garam masala as far as the eye can see.</p> <p>Some of our strongest geographic and cultural associations are tied to a region's local foods. This playground competitions asks you to predict the category of a dish's cuisine given a list of its ingredients.</p>

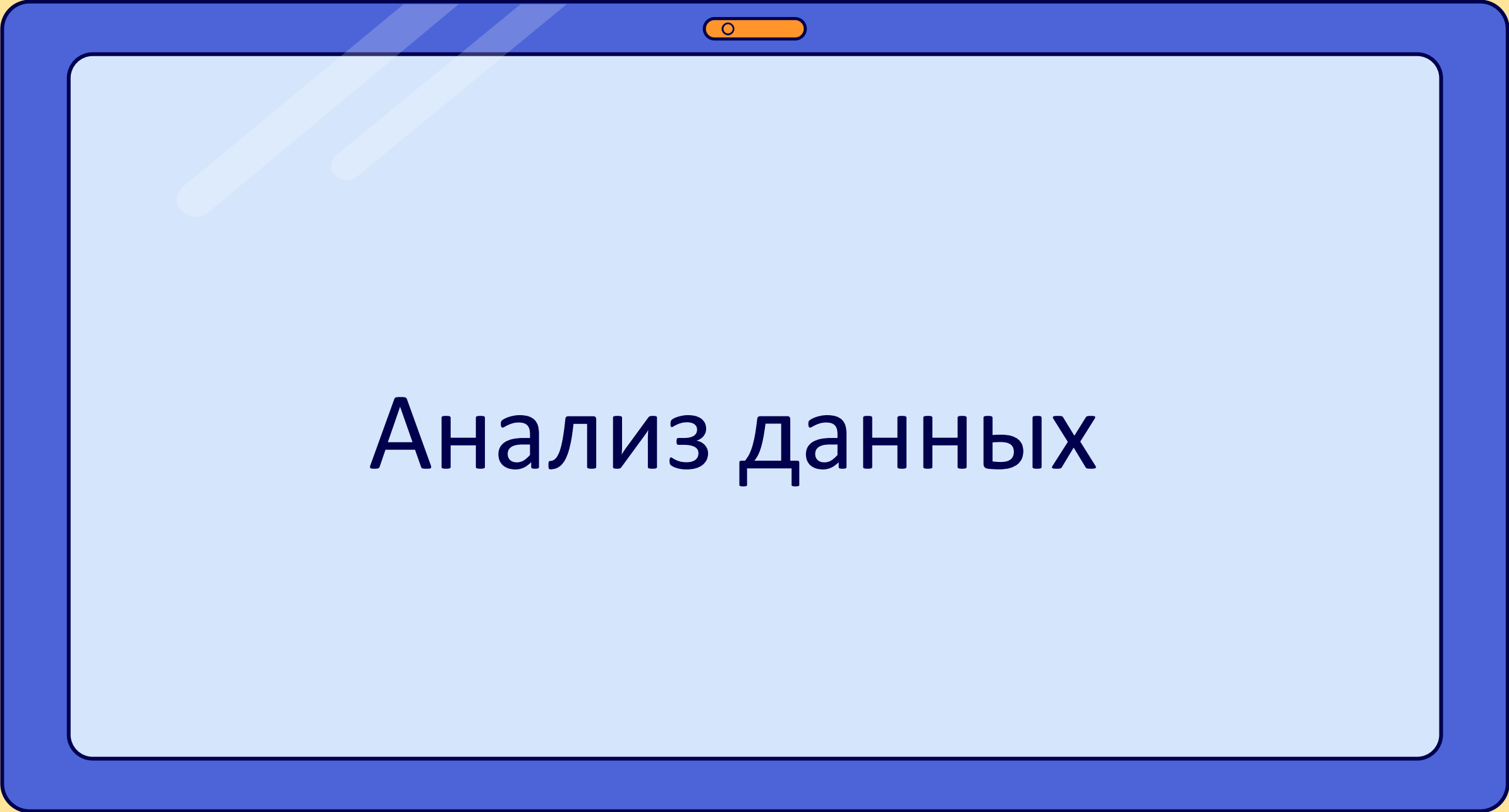


```
[ ] train_df.head()
```

	id	cuisine	ingredients
0	10259	greek	[romaine lettuce, black olives, grape tomatoes...
1	25693	southern_us	[plain flour, ground pepper, salt, tomatoes, g...
2	20130	filipino	[eggs, pepper, salt, mayonaise, cooking oil, g...
3	22213	indian	[water, vegetable oil, wheat, salt]
4	13162	indian	[black pepper, shallots, cornflour, cayenne pe...

- Данные представлены в формате JSON (вывод данных из обучающей выборки показан на рисунке)
- Ингредиенты записаны в список, длина рецепта варьируется, поэтому нужно решить, как эффективнее работать с такими данными





# Анализ данных

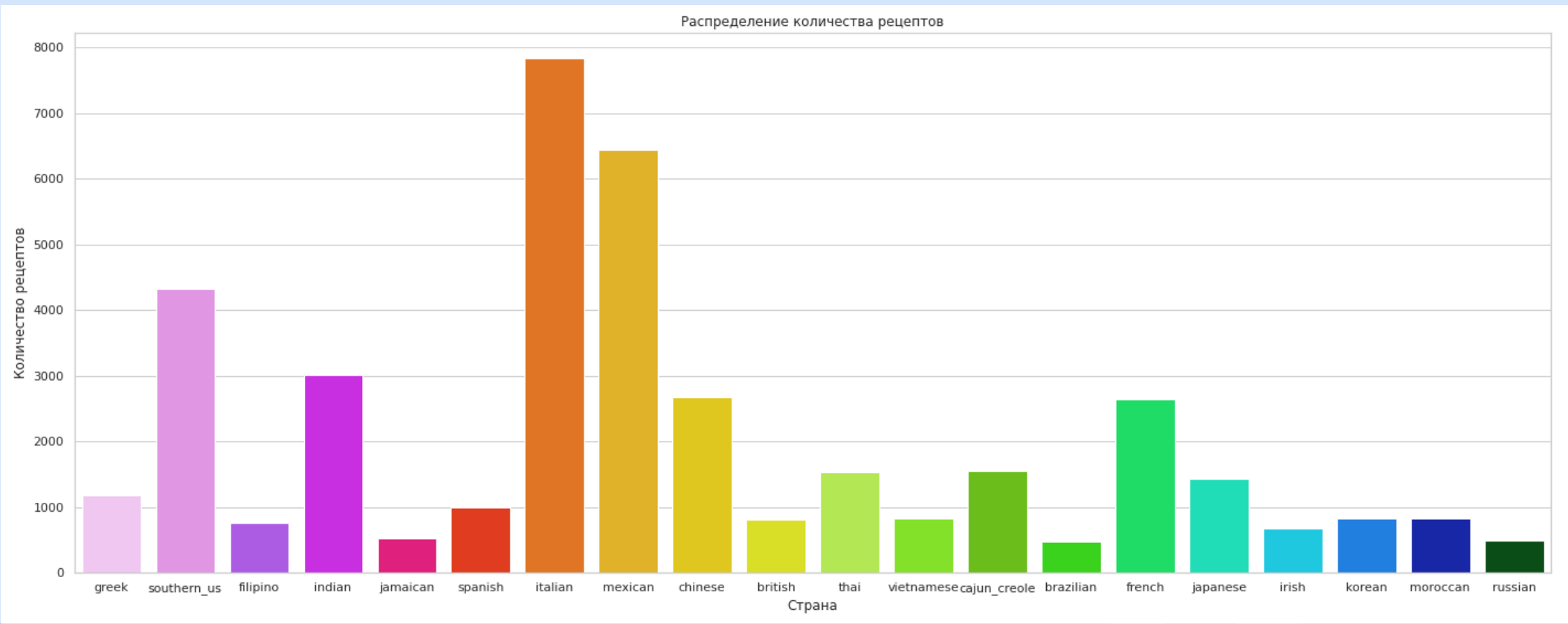


Целевая переменная – страна. По ингредиентам рецепта надо предсказать одну из 20 стран, для которой характерно такое блюдо.

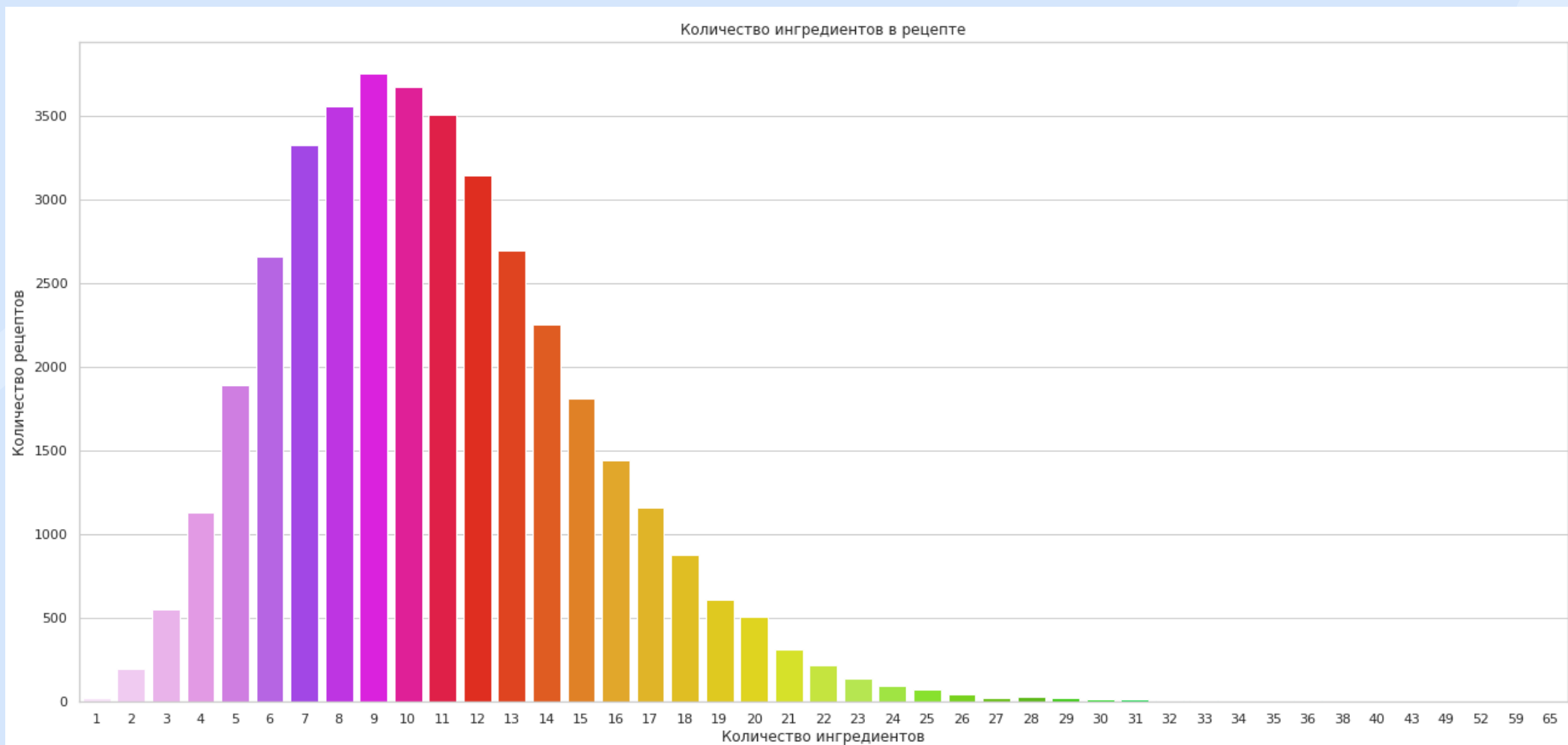
Данные довольно большие (39774 рецепта в обучающей выборке и 9944 рецепта в тестовой).



По данному графику видно, что в выборке больше всего рецептов относятся к итальянской и мексиканской кухням; меньше всего рецептов – это рецепты из России, Ямайки и Бразилии.



По данному графику видно, что большая часть рецептов состоит из 6-14 ингредиентов, но также есть какие-то странные рецепты из 1, или, наоборот, 65 ингредиентов.





Даже по рецептам, состоящим только из одного ингредиента, видно, что в данных есть повторы, поэтому перед построением модели нужно их удалить.

	id	cuisine	ingredients
940	4734	japanese	[sushi rice]
2088	7833	vietnamese	[dried rice noodles]
6787	36818	indian	[plain low-fat yogurt]
7011	19772	indian	[unsalted butter]
8181	16116	japanese	[udon]
8852	29738	thai	[sticky rice]
8990	41124	indian	[butter]
10506	32631	mexican	[corn tortillas]
13178	29570	thai	[grained]
17804	29849	southern_us	[lemonade concentrate]
18136	39186	thai	[jasmine rice]
18324	14335	indian	[unsalted butter]
21008	39221	italian	[cherry tomatoes]
22119	41135	french	[butter]
22387	36874	indian	[cumin seed]
23512	35028	french	[haricots verts]
26887	18593	mexican	[vegetable oil]
29294	7460	spanish	[spanish chorizo]

Обучающая выборка

Если посмотреть на тестовую выборку, то можно заметить, что в ней тоже встречаются рецепты из одного ингредиента, поэтому отбрасывать из обучающей выборки границы нет смысла.

	id	ingredients
544	36822	[plain low-fat yogurt]
3248	34002	[glutinous rice]
3444	28414	[pimentos]
3621	10077	[sweetened condensed milk]
4021	32883	[unsalted butter]
7417	45798	[chiles]
8081	45398	[parmesan cheese]
9407	32743	[shiitake]

## В обучающей выборке:

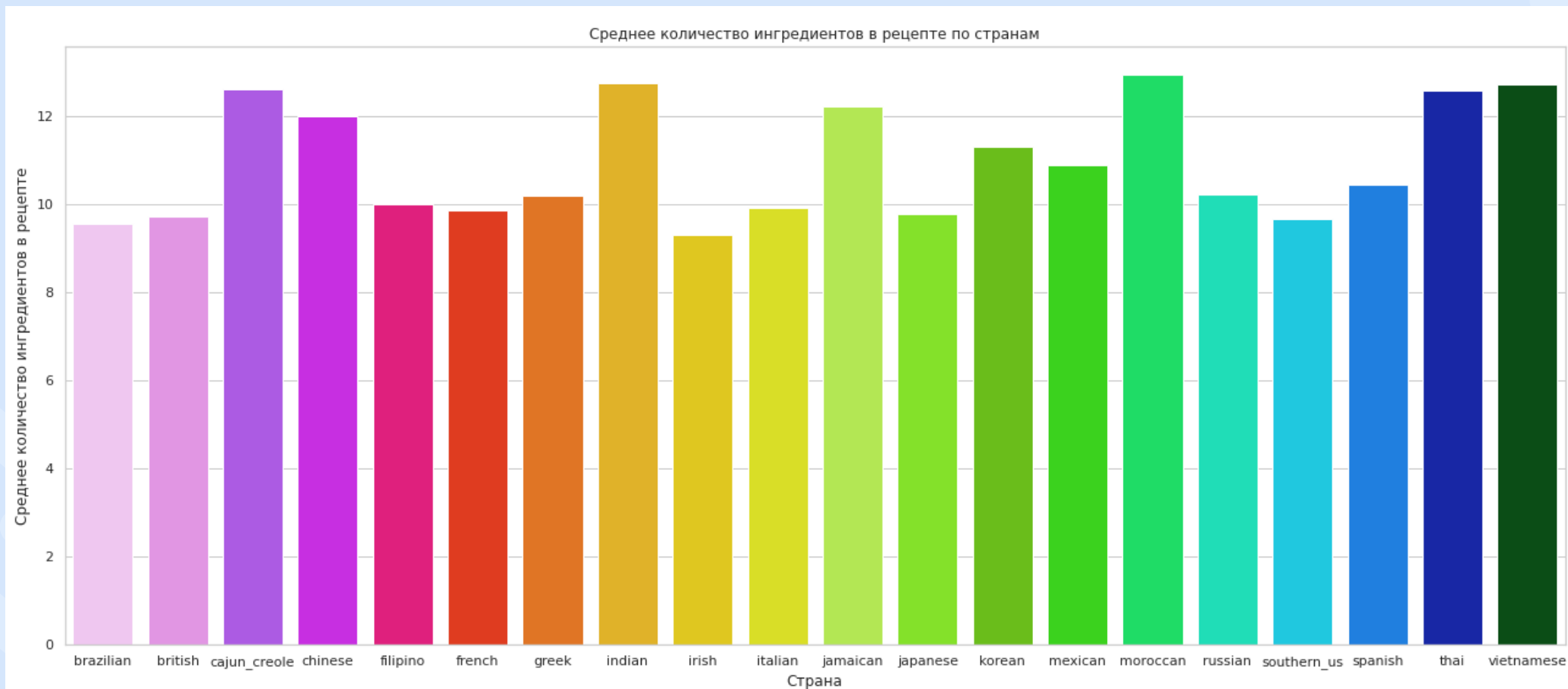
	id	cuisine	ingredients	ingredients_count
10513	49282	mexican	[condensed cream of chicken soup, pepper, refr...	49
15289	3885	italian	[fettucine, fresh marjoram, minced garlic, oli...	65
22906	2253	indian	[white vinegar, sparkling lemonade, coconut su...	49
26103	13049	mexican	[vanilla ice cream, lime, garlic powder, zucch...	52
30350	13430	brazilian	[marshmallows, fresh corn, cheddar cheese, shr...	59
31250	29216	italian	[eggs, warm water, pepper, dried basil, unsalt...	43

## В тестовой:

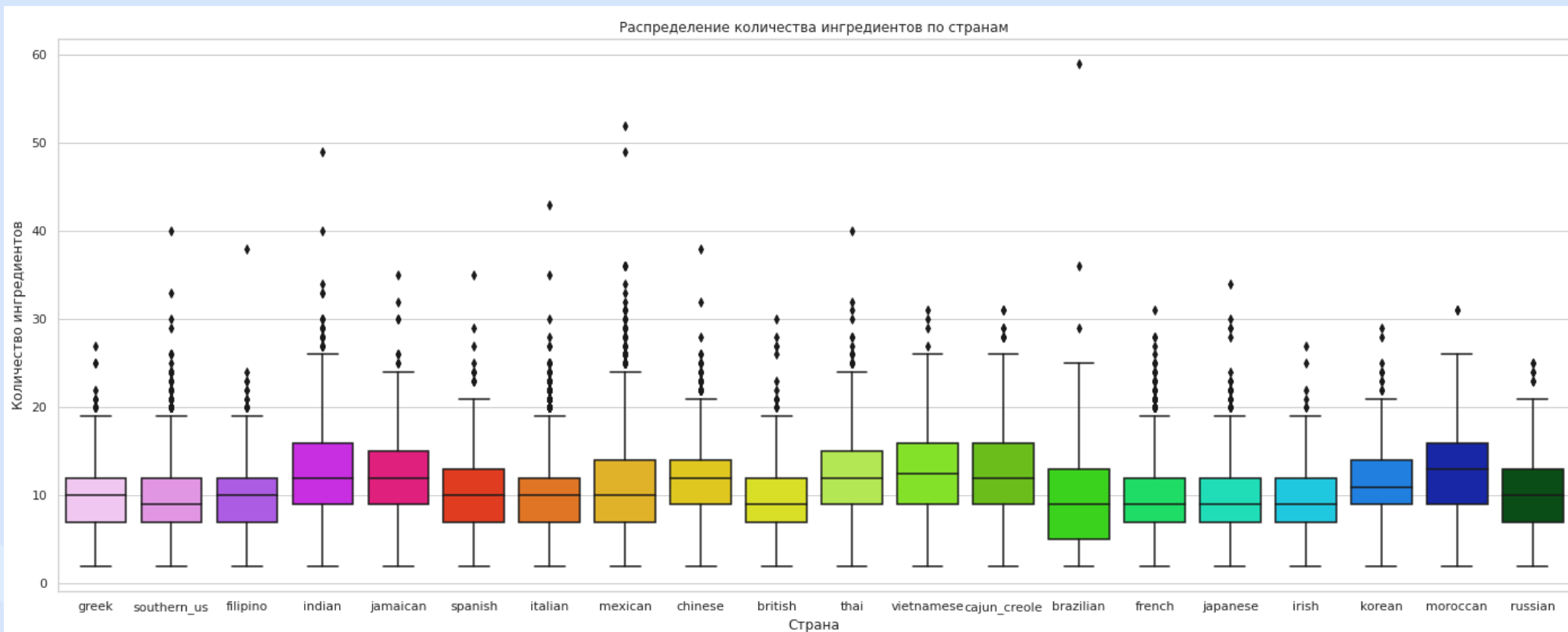
	id	ingredients	ingredients_count
4338	39167	[pico de gallo, slaw, orange, coriander seeds,...	41
4809	526	[diced onions, yellow mustard seeds, chili pep...	50

Сказать такое же про рецепты с большим количеством ингредиентов (больше 55) нельзя. Стоит их отбросить, чтобы не допустить переобучение модели.

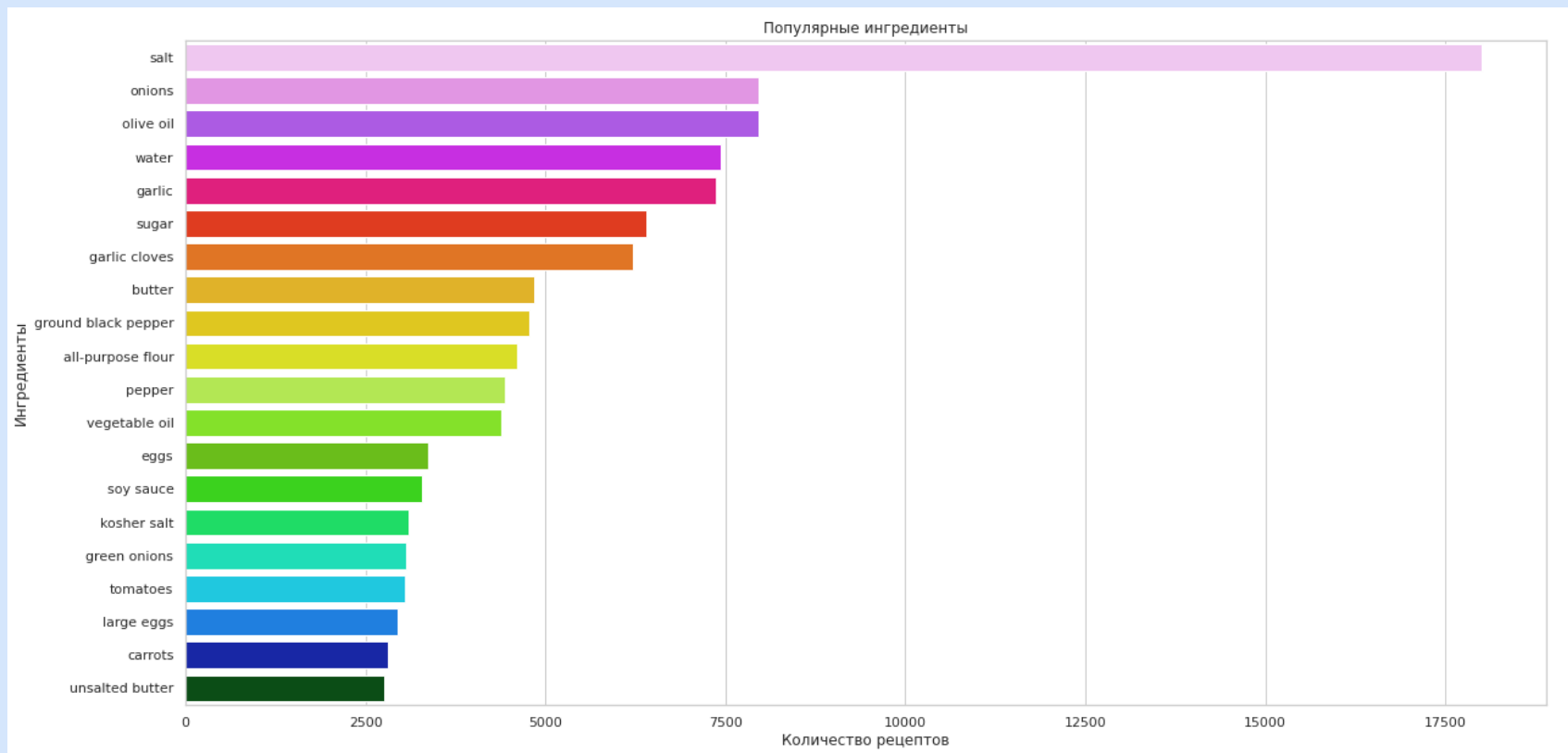
На графике показано среднее количество ингредиентов в рецепте в зависимости от страны. Здесь нет каких-то выделяющихся стран.





По распределению количества ингредиентов по странам видно, что опять же нет каких-то выделяющихся стран и использовать этот признак нелогично. На графике видно, что выбросы не зависят от страны и присутствуют во всех странах.



Если посмотреть на самые популярные ингредиенты, то можно заметить, что даже самый часто встречающийся из них (соль) появляется только примерно в половине рецептов, поэтому отбрасывать его из выборки нельзя.

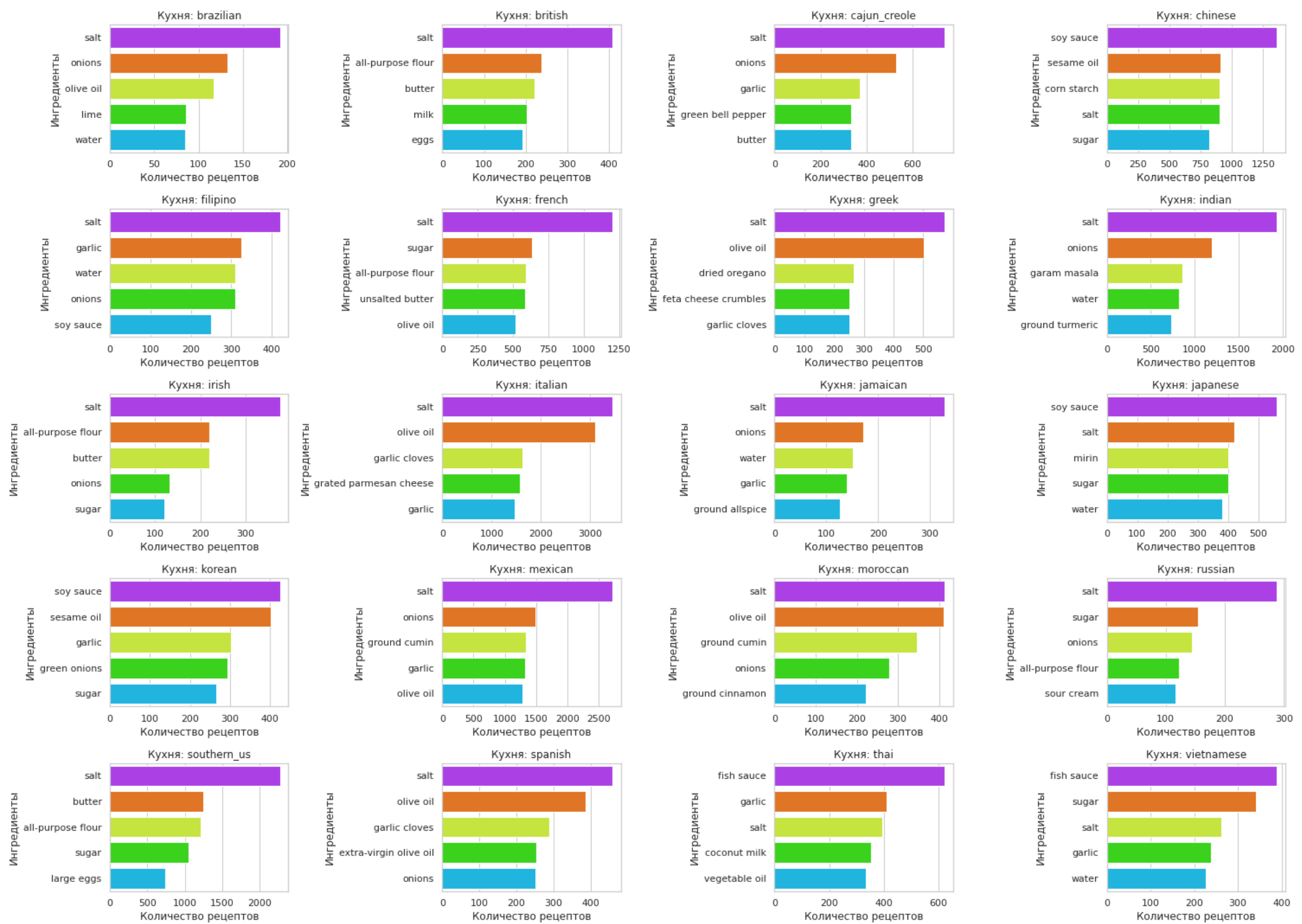




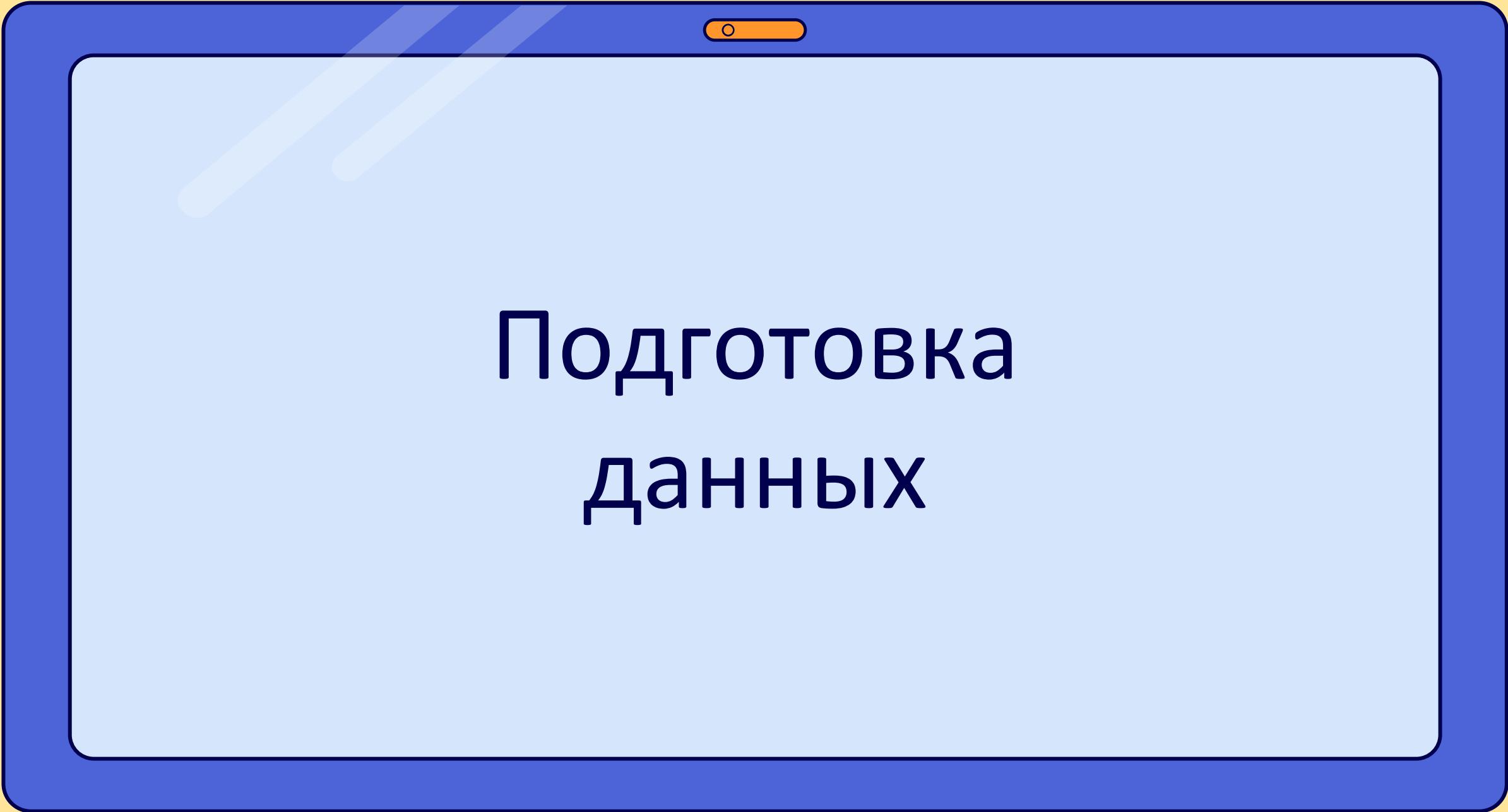
На следующем графике показаны популярные ингредиенты для каждой кухни в отдельности.

Можно предположить, что наиболее значимое влияние на предсказание окажут такие ингредиенты: **olive oil, fish sauce, soy sauce.**

Видно, что соль хоть и оказывает влияние на большинство рецептов, для некоторых стран (Вьетнам, Корея, Китай) большее значение имеют другие ингредиенты.





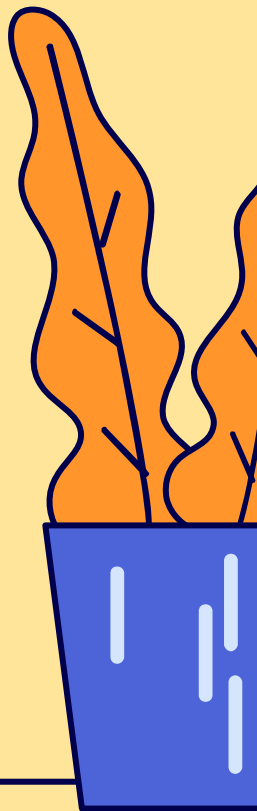


# Подготовка данных

Что нужно изменить в данных:

- Привести все слова к строчным буквам
- Удалить единицы измерения (кг, унции)
- Удалить дополнительные символы (скобки, цифры, проценты)
- Удалить лишние пробелы
- Из длинных словосочетаний оставить только существительные

После выполнения очистки, в обучающей выборке осталось 39672 рецепта.



С помощью метода CountVectorizer данные приводятся к виду, показанному на рисунке.

Нужно изменить параметр токенизации, т.к. по умолчанию она происходит по пробелам, а для сохранения словосочетаний токенизация должна происходить по запятым.

```
train_features.head()
```

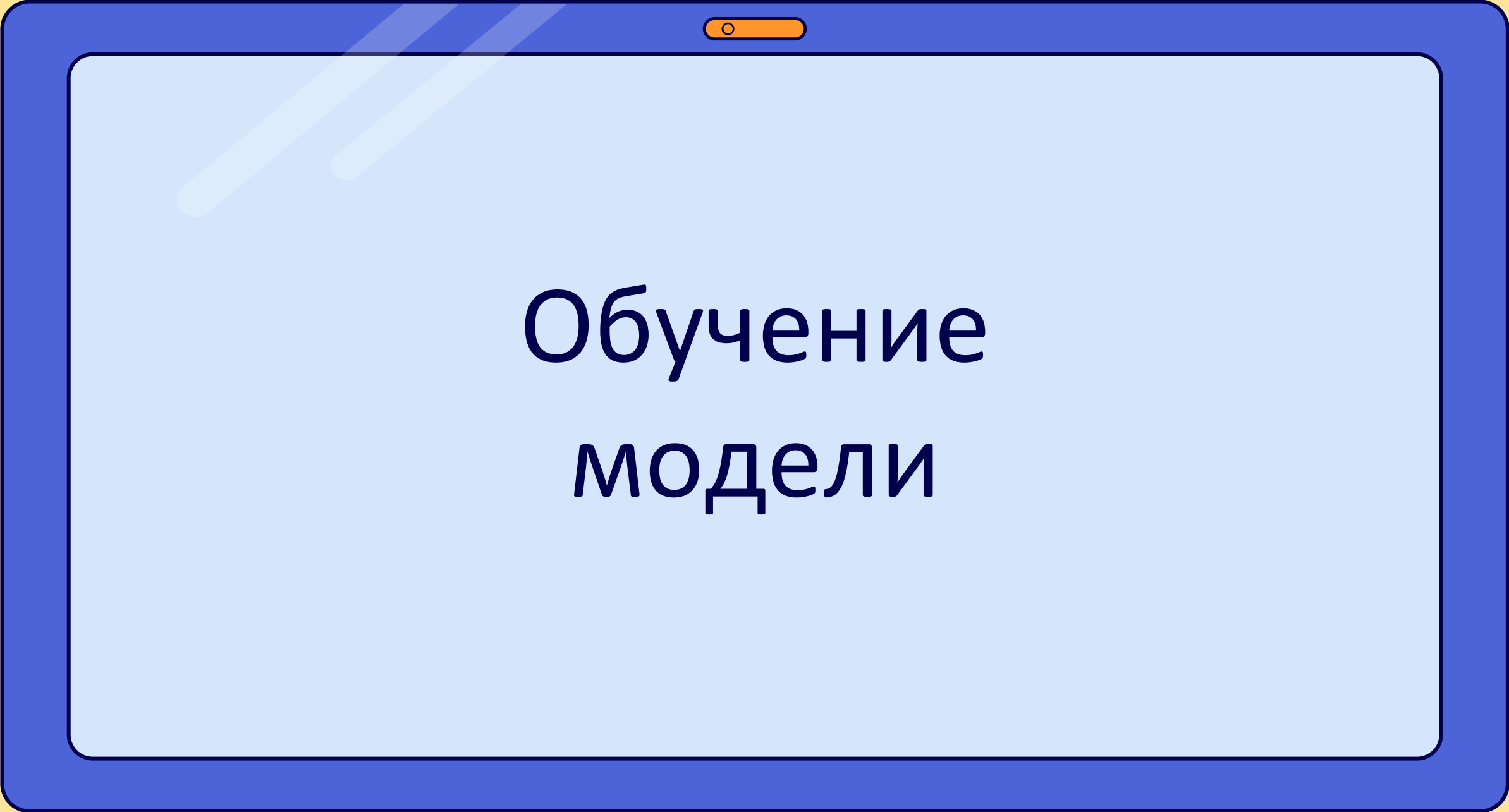
	acorn squash	active dry yeast	adobo	adobo sauce	agave nectar	ai	ajwain	alfredo sauce	all purpose flour	all purpose unbleached flour	allspice	allspice berries	almond extract	almond flour	almond milk	almond paste	almonds	amaretto	amchur	american cheese
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

5 rows × 2000 columns



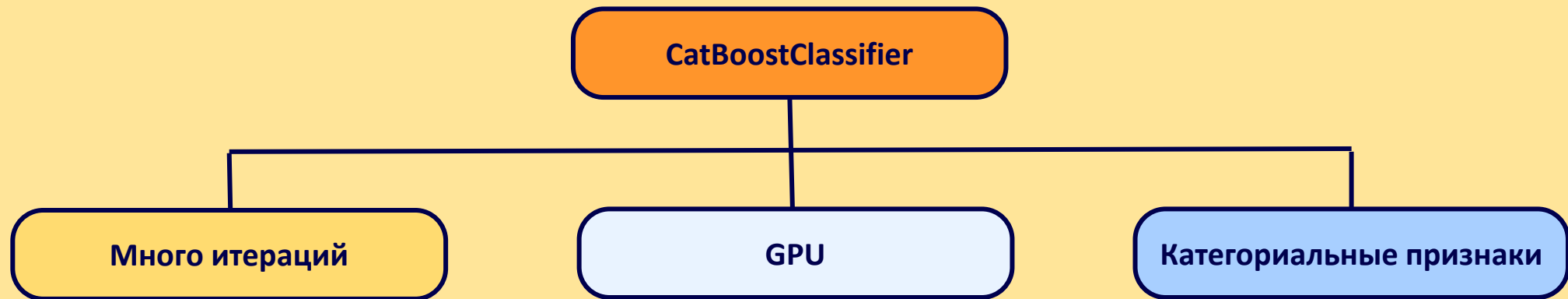
Для параметра `max_features`, который определяет размер словаря, считая самые часто встречающиеся ингредиенты, в ходе экспериментов было установлено значение 2000, потому что при меньшем размере теряется полнота охвата переменных, а при большем размере происходит переобучение

После этого данные разделяются на обучающую и валидационную выборку; обучающая выборка состоит из 80% данных (31737 рецептов), а валидационная из 20% рецептов (7935).

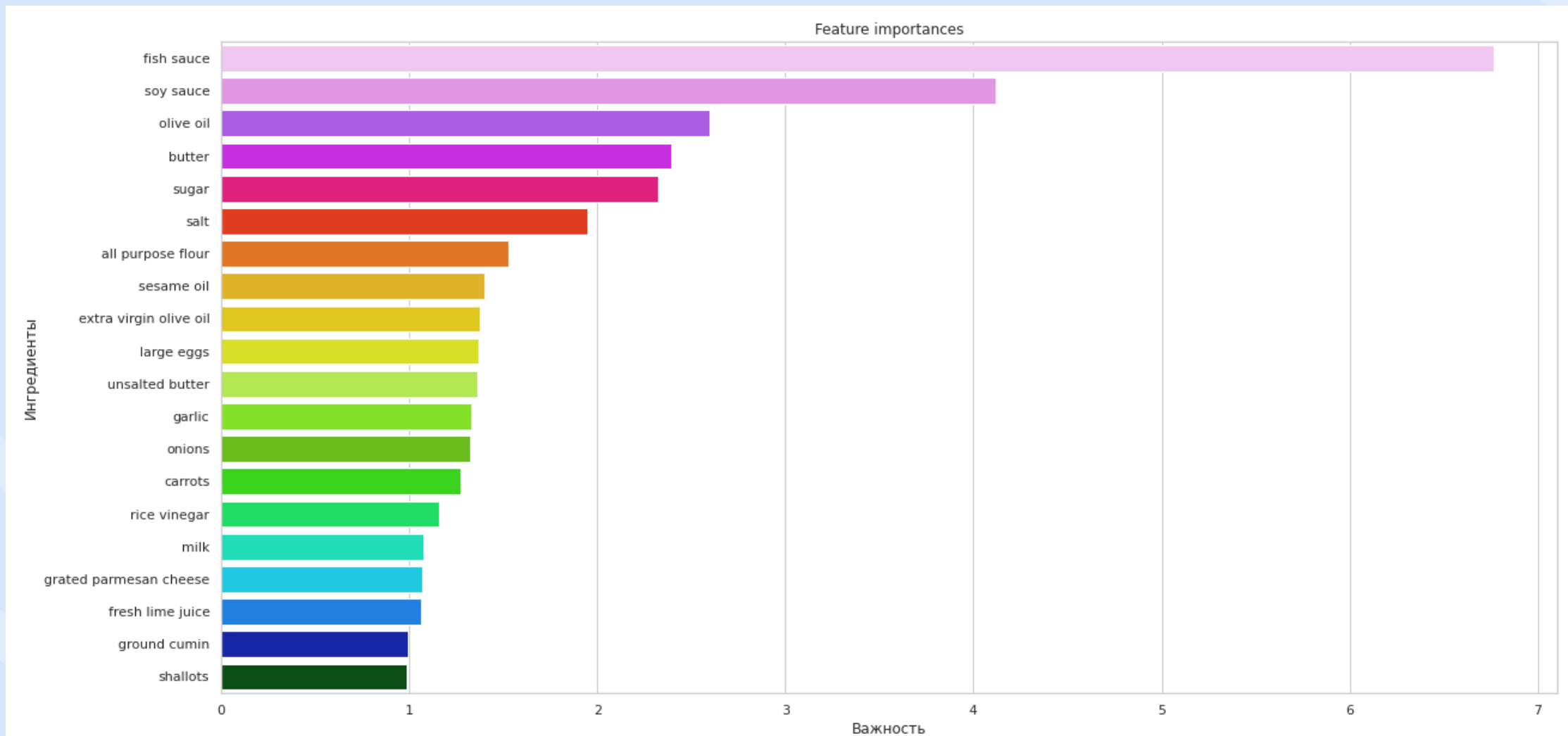




# Обучение модели

Для обучения используется класс **CatBoostClassifier** из библиотеки **catboost**, потому что целевая переменная представлена в категориальном виде. Также плюс в том, что обучение можно производить на GPU, а для больших размеров обучающей выборки это особенно важно.



На графике показаны признаки, которые оказывают наибольшее влияние на предсказание.





Особое влияние на предсказание оказали такие ингредиенты:  
**fish sauce, soy sauce, butter.**

Если еще раз посмотреть на график популярных ингредиентов в зависимости от кухни, то видно, что масло встречается сразу в нескольких странах, поэтому это неудивительно.





Спасибо за внимание

The image is a stylized illustration of a laptop. The laptop's body is dark blue. The screen is a light blue rectangle with rounded corners, containing the Russian text 'Спасибо за внимание' (Thank you for attention) in a dark blue, sans-serif font. To the left of the screen, there are three stylized, overlapping shapes representing books or folders in blue and orange. To the right, there is a stylized orange rectangular object with a blue wavy line above it, resembling a plant or a decorative element. The bottom of the laptop is a dark blue keyboard area with a grid of orange lines representing keys. The entire background is a solid light yellow color.