УДК 303.722

Кластерный анализ алкогольных предпочтений по странам мира

М.И. Скоморохов✉

Уральский государственный Экономический Университет, Екатеринбург, Россия

В статье проводится анализ набора данных по потреблению различных алкогольных напитков по странам мира, разделенным по регионам: западный регион, центральный и восточный регионы. В ходе исследования была поставлена и проверена гипотеза, что у стран в одной группе имеются ярко выраженные схожие предпочтения в выборе спиртных напитков. Для проведения анализа была использована программа R Studio, а также различные сторонние библиотеки для этой программы. Кластерный анализ проводился по методу k-means, число кластеров было определено по методу «локтя». Результаты анализа подтвердили ранее поставленную гипотезу.

Ключевые слова: кластерный анализ, интеллектуальный анализ данных, алкогольные предпочтения, страны мира, R Studio.

Cluster analysis of alcohol preferences by countries

M.I. Skomorokhov✉

Ural State University of Economics, Yekaterinburg, Russia

The article analyzes a set of data on the consumption of various alcoholic beverages by countries of the world, divided by regions: the western region, the central and eastern regions. The study hypothesized and tested that countries in the same group have pronounced preferences when choosing alcoholic beverages. For the analysis, the R Studio program was used, as well as various third-party libraries for this program. Cluster analysis was carried out using the k-means method, the number of clusters was determined using the «elbow» method. The results of the analysis confirmed the previously posed hypothesis.

Keywords: cluster analysis, data mining, alcohol preferences, countries of the world, R Studio.

Анализ данных – это не просто обработка информации после ее сбора, это средство проверки гипотез. Цель любого анализа данных заключается в понимании исследуемой ситуации целиком (выявление тенденций, в том числе негативных отклонений от плана, прогнозирование и получение рекомендации). При постановке задачи необходимо сформировать гипотезу, которую и следует проверить, используя методы решения ИАД [1, 2].

В текущей работе в качестве набора данных решено было выбрать данные о потреблении различных спиртных напитков в мире, разделенные по странам и регионам (запад, центр и восток) [3]. Была выдвинута следующая гипотеза: страны одного региона имеют схожие предпочтения в выборе спиртных напитков. Гипотеза может показаться простой, но действительно ли страны одного региона предпочитают один и тот же вид напитка, узнаем в ходе решения задачи. Актуальность работы заключается в следующем: в последнее время стали открываться новые рынки сбыта для производителей разных отраслей, в том числе, для отрасли алкогольной продукции. При выходе на новый рынок производителям важно знать предпочтения жителей региона, чтобы не понести огромные потери и успешно провести экспансию. Данное исследование можно взять за основу в маркетинговом исследовании новых регионов сбыта продукции.

Цель текущей работы – выяснить, имеют ли страны одного региона схожие предпочтения в выборе спиртных напитков или не имеют. Для этого будет использоваться кластерный анализ [4, 5].

После постановки задачи и выбора метода решения, следует выбрать программное средство, которое поможет в достижении поставленной цели [6]. Тщательно проанализировав программное обеспечение для нужд интеллектуального анализа данных, было решено остановиться на программах, которые позволяют писать программный код для анализа данных. Это обусловлено тем, что задача имеет вид кластерного анализа, в данном случае такие программные средства как Yandex DataLens, Tableau и другие не подойдут, поскольку не предоставляют доступа к такому функционалу, который бы позволил решить задачу кластерного анализа.

Были выбраны два программных средства – Jupyter Notebook [7] и RStudio [8]. Первое позволяет удобно работать как с кодом, так и с его демонстрацией, не перекрывая другие графики, которые буду появляться во время решения задачи. Второе средство позволяет наиболее быстро решить задачу, но не имеет широких возможностей для демонстрации результатов этого решения. Однако, RStudio имеет несколько преимуществ по сравнению с Jupyter Notebook:

– RStudio является более легковесной и быстрой программой по сравнению с Jupyter Notebook;

– RStudio предоставляет больше функциональности для работы с R, имеет множество полезных вкладок, например, крайне полезную вкладку Environment, которая показывает все переменные, добавленные в окружение;

– RStudio не требует установки дополнительного ПО, а Jupyter Notebook требует установки «тяжелой» Anaconda, которая предоставляет множество, пускай и полезного, но ненужного функционала для решения текущей задачи.

Кластерный анализ представляет собой многомерную статистическую процедуру, выполняющую сбор данных, содержащих информацию о выборке объектов, и затем упорядочивающую объекты в сравнительно однородные группы [9]. Задача кластеризации относится к широкому классу задач обучения без учителя. Кластерный анализ применяется тогда, когда нужно либо породить гипотезу, либо проверить гипотезу, которую сформулировал ученый, а также при разработке типологии или классификации. Кластерный анализ проводится в несколько этапов [10]:

– отбор выборки для кластеризации. В данном случае, подразумевается, что имеет смысл кластеризовать только количественные данные;

– определение множества переменных, по которым будут оцениваться объекты в выборке;

– вычисление значений той или иной меры сходства между объектами;

– применение метода кластерного анализа для создания групп сходных объектов;

– проверка достоверности результатов кластерного решения.

Пройдём каждый этап кластерного анализа на примере нашей задачи. Данные, полученные из репозитория FiveThirtyEight, представляют собой ту самую выборку, которую делают на первом этапе кластерного анализа. В ней имеем только количественные данные (рис. 1), за исключением названия страны и группы, но страну будем использовать как название строки, а группу удалим.

	A	B	C	D	E	F
1	country	beer_servings	spirit_servings	wine_servings	tota	group
2	Afghanistan (east)	0	0	0	0.0	east
3	Albania (center)	89	132	54	4.9	center
4	Algeria (west)	25	0	14	0.7	west
5	Andorra (west)	245	138	312	12.4	west
6	Angola (west)	217	57	45	5.9	west
7	Antigua & Barbuda (west)	102	128	45	4.9	west
8	Argentina (west)	193	25	221	8.3	west
9	Armenia (east)	21	179	11	3.8	east
10	Australia (east)	261	72	212	10.4	east
11	Austria (west)	279	75	191	9.7	west
12	Azerbaijan (center)	21	46	5	1.3	center
13	Bahamas (west)	122	176	51	6.3	west
14	Bahrain (center)	42	63	7	2.0	center
15	Bangladesh (east)	0	0	0	0.0	east
16	Barbados (west)	143	173	36	6.3	west

Рисунок 1. Выборка данных по потреблению спиртных напитков

Множество переменных, которые будут использованы для проведения анализа следующие: beer_servings (порции пива), spirit_servings (порции крепких напитков), wine_servings (порции вина). В данном случае total (общее количество) использоваться не будет, поскольку для решения поставленной задачи это не требуется, а столбцы country (страна) и group (регион) являются вспомогательными, они при кластеризации будут убраны.

На третьем этапе вычисляется значение той или иной меры сходства между объектами. Для выполнения этого шага будет использован метод Варда. Этот метод отличается от всех других методов, поскольку он использует методы дисперсионного анализа для оценки расстояний между кластерами. Метод минимизирует сумму квадратов для любых двух кластеров, которые могут быть сформированы на каждом шаге. Метод представляется очень эффективным, однако он стремится создавать кластеры малого размера. Этот метод был выбран в виду его эффективности.

Последние два этапа кластерного анализа являются наиболее активными фазами. Для реализации метода кластерного анализа будет использован язык R в связке с RStudio. Сначала требуется загрузить данные в RStudio, это делается двумя способами – через команду «read.table» или с помощью кнопки «Import Dataset» в окне окружения. Поскольку текущий набор данных представлен в формате xlsx, то решено было выбрать импортирование через средства RStudio.

После импорта данных таблица откроется и появится в окружении. Далее были добавлены все требуемые библиотеки, которые будут использованы в ходе решения задачи кластеризации. Чтобы убедиться, что данные импортировались верно можно использовать просмотр таблицы через средства RStudio, а можно использовать команду head(), которая покажет первые несколько строк таблицы вместе с заголовками. Для более удобной работы с данными была создана переменная data, которой присвоили данные из xlsx файла. После этого можно изменять таблицу под нужды кластерного анализа.

Для анализа данных нам требуется несколько столбцов – beer_servings (порции пива), spirit_servings (порции крепких напитков), wine_servings (порции вина), потому остальные следует удалить. Уберём столбцы total_litres_of_pure_alcohol (общее количество чистого алкоголя), country (страна) и group (регион), предварительно переименовав строки в страны.

Сделав все приготовления с данными, можно непосредственно перейти к кластерному анализу. Сначала требуется определить оптимальное число кластеров. Для этого используется метод «локтя» и библиотека «factoextra» [11], написанная для языка R. На рисунке 2 можно заметить, что “локтем” является цифра 3, то есть оптимальным числом кластеров будет – 3.

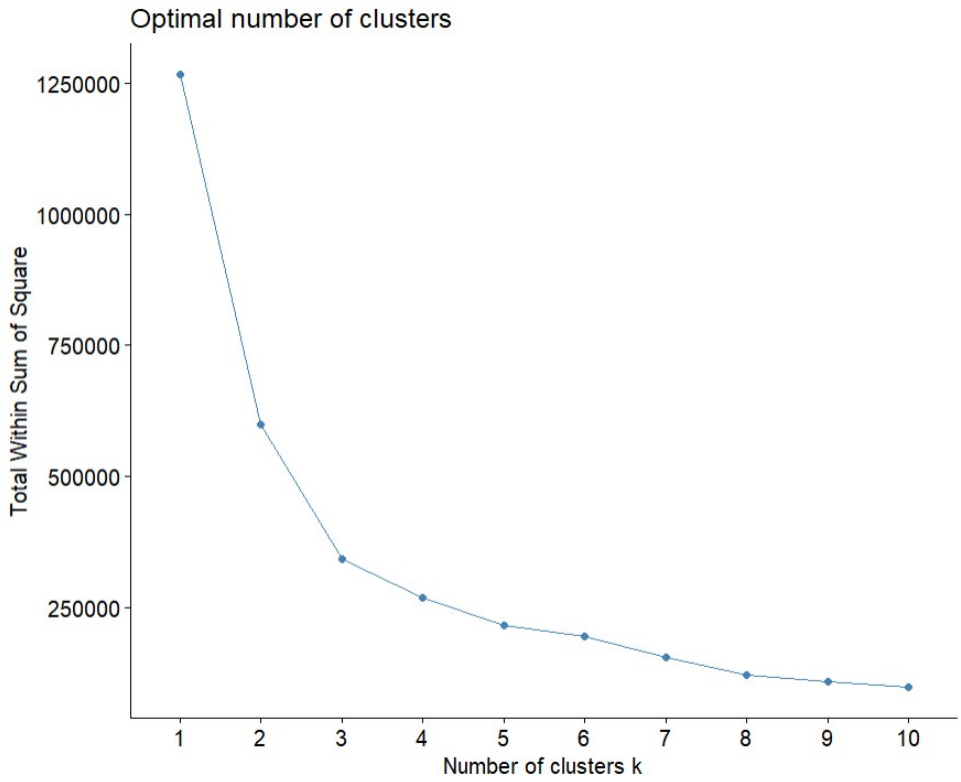

Рисунок 2. График определения числа кластеров

Далее следует либо изучить выборку всех стран, либо сделать ручную выборку, где будет пропорциональное распределение стран по регионам. Рассмотрим оба варианта, которые позволили бы подтвердить или опровергнуть гипотезу. Сначала выполним кластерный анализ на всей выборке стран. На рисунке 3 представлены результаты кластерного анализа. На рисунке мы можем видеть кластеры, каждый кластер представляет собой предпочтение в том или ином напитке, точки – это страны. Серая область представляет собой кластер крепких напитков, синий кластер – представляет вино, желтый кластер – пиво. Наиболее кучный кластер показывает, что в странах одного региона ярко выражено предпочтения населения к пиву. В остальных кластерах нет столь явно выраженного предпочтения.

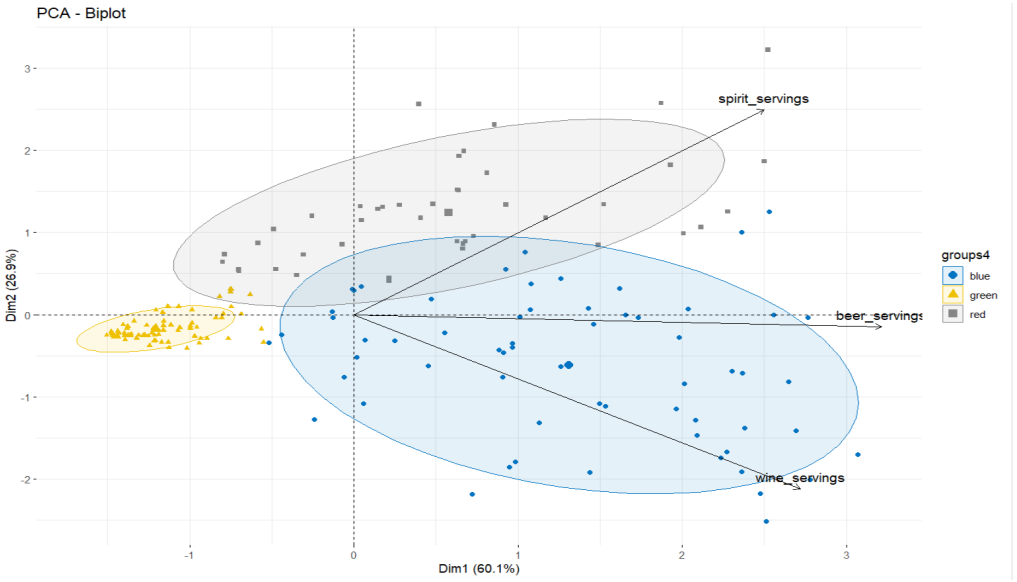

Рисунок 3. Результаты кластерного анализа на всей выборке стран

Следует сделать новую выборку, которая будет включать равное количество стран каждого региона. На рисунке 4 представлена кластеризация по новой выборке. В данном случае обозначение цветов таково – синее означает вино, красное обозначает пиво, а зеленое означает другие спиртные напитки.

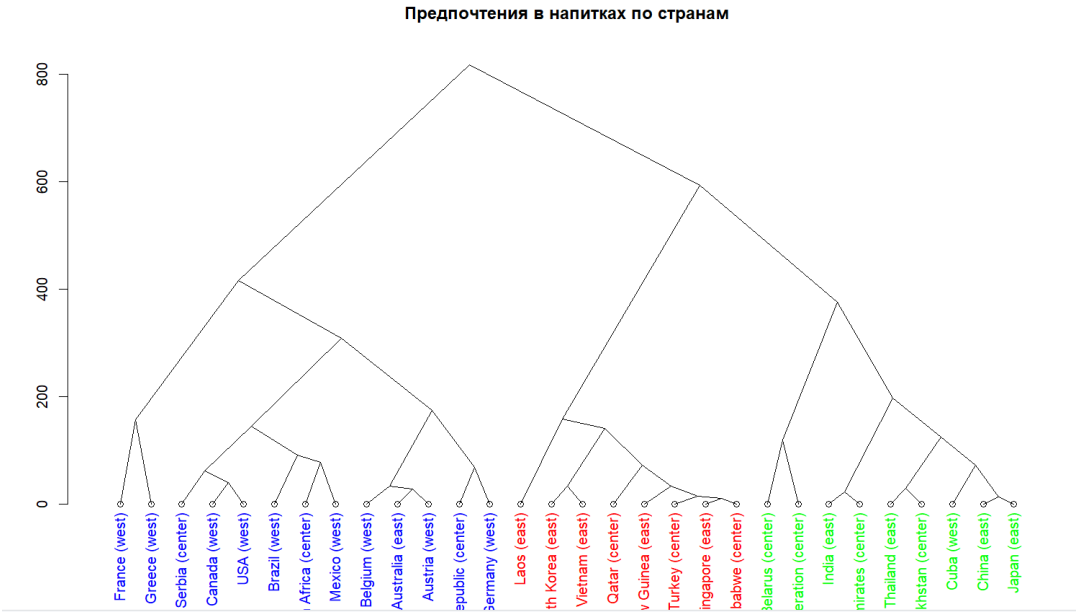

Рисунок 4. Результаты кластерного анализа на ручной выборке стран

Отчетливо видно, что вино предпочитают в 9 западных странах, в 3 центральных и в 1 восточной, причем, следует заметить, что Австралия, пусть и находится в восточной части мира, но имеет, скорее, западные предпочтения и привычки. Пиво предпочитают в 3 центральных странах и в 5 восточных, а иные спиртные напитки предпочитают в 4 центральных странах и в 4 восточных, а также в одной западной стране, которой является Куба.

Далее проверим, каково распределение стран в общей выборке со всеми странами. Пиво больше всего предпочитают центральные страны (42 страны), на втором месте восточные страны (25 стран), на третьем месте – западные страны (21 страна). Иные спиртные напитки больше всего предпочитают в западных странах (22 страны), на втором месте центральные страны (13 стран), замыкают тройку восточные (7 стран). Вино же больше всего предпочитают западные страны (46 стран), на втором месте центральные страны (11 стран), на третьем месте восточные (4 страны).

Следует заметить, что в двух группах есть явные лидеры, жителям западных странам свойственно пить вино, жителям центральных стран – пиво, что касается жителей восточных стран, они также больше предпочитают пиво. Иные же спиртные напитки делят между собой западные и центральные страны.

Вспомним выдвинутую ранее гипотезу, она звучала следующим образом: страны одного региона имеют схожие предпочтения в выборе спиртных напитков. Данные довольно ярко показали, что схожие предпочтения присутствуют, но нет строго распределения, при этом все западные страны предпочитают только один напиток.

СПИСОК ИСТОЧНИКОВ

1. Рубаков С.В. Современные методы анализа данных / С.В. Рубаков. – Москва: Наука, 2008. – 12 с.
2. Дадабаева Р.А. Методика оценки экономических показателей хозяйственной деятельности на основе ABC-анализа / Р.А. Дадабаева, А.В. Голубин // Цифровые модели и решения. – 2022. – Т. 1. – № 3. – С. 6.
3. Источник набора данных [Электронный ресурс]. – URL: https://github.com/fivethirtyeight/data/blob/master/alcohol-consumption/drinks.csv (дата обращения: 11.10.2023).
4. Ярош О.Б. Аромамаркетинг: асимметрия потребительского восприятия традиционных продуктов регионального происхождения / О.Б. Ярош, О.Б. Калькова // Управленец. – 2022. – Т. 13. – № 3. – С. 67-79.
5. Блусь П.И. Пространственная кластеризация как инструмент снижения внутрирегиональной неравномерности / П.И. Блусь // Journal of New Economy. – 2022. – Т. 23. – № 1. – С. 88-108.
6. Луньков А.Д. Интеллектуальный анализ данных / А.Д. Луньков, А.В. Харламов. – Саратов: 2014. – 96 с.
7. Jupyter Notebook [Электронный ресурс]. – URL: https://jupyter.org/ (дата обращения: 11.10.2023).
8. R Studio [Электронный ресурс]. – URL: https://posit.co/download/rstudio-desktop/ (дата обращения: 11.10.2023).
9. Ключников М.В. Технология кластерного анализа финансовых показателей банков / М.В. Ключников. – Москва: 2006. – 10 с.
10. Дюран Б. Кластерный анализ / Б. Дюран, П. Одел. – Москва: 1977. – 128 с.
11. Библиотека factoextra [Электронный ресурс]. – URL: https://www.rdocumentation.org/packages/factoextra/versions/1.0.7/topics/fviz_nbclust (дата обращения: 11.10.2023).

ИНФОРМАЦИЯ ОБ АВТОРАХ

Скоморохов Матвей Иванович, магистрант, Уральский государственный Экономический Университет, Екатеринбург, Россия.
e-mail: b.boy.matv@gmail.com