МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение

высшего образования

**«КУБАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»**

**(ФГБОУ ВО «КубГУ»)**

**Факультет компьютерных технологий и прикладной математики**

**Кафедра вычислительных технологий**

**ЛАБОРАТОРНАЯ РАБОТА №5.**

**Дисциплина: Обработка больших данных**

Работу выполнил:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Лотарев С.Ю.

Направление подготовки: 02.03.02 Фундаментальная информатика и информационные технологии

Преподаватель: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Т. А. Приходько

Краснодар

2025

**Тема**: Задачи классификации и кластеризации. Байесовская классификация и деревья принятия решений в R.

**Цель работы**: Закрепить знания об алгоритмах классификации и кластеризации данных, ознакомиться с некоторыми функциями языка R, осуществляющими этот вид анализа, принципами их работы. Научиться визуализировать результаты работы функций кластерного анализа и классификаторов, интерпретировать полученные результаты. Научиться выполнять классификацию на основе формулы Байеса и деревьев решений.

Вариант 5:

|  |  |
| --- | --- |
| Лотарев Сергей Юрьевич | Туристические поездки |

**Ход работы**

1. Приведём фрагмент начальных данных и решаемую задачу.

Фрагмент исходного датасета:

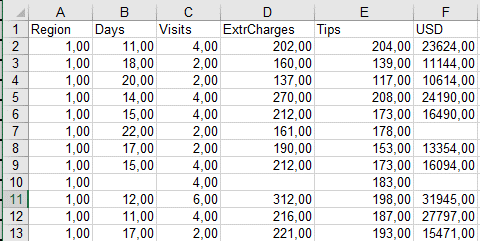


Рисунок 1. Фрагмент исходного датасета

Задача: Провести кластерный анализ наблюдений, представляющих туристов из различных регионов. Цель — выявить устойчивые группы туристов по поведенческим и финансовым признакам (продолжительность проживания, число визитов, дополнительные расходы, чаевые, доход). Предполагается, что полученные кластеры могут в определённой степени соответствовать регионам происхождения туристов. Однако при проведении кластерного анализа признак "регион" исключается из анализа, чтобы не вносить априорную информацию о принадлежности к группам. Регион будет использоваться только для последующего анализа и интерпретации кластеров.

2. Выполним дескриптивный анализ по каждому из числовых атрибутов датасета. Он будет включать минимальное и максимальное значение и соответствующие им страны, среднее арифметическое, медиану, моду, а также боксплот.

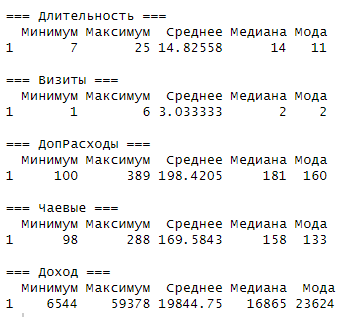


Рисунок 2. Числовые результаты дескриптивного анализа

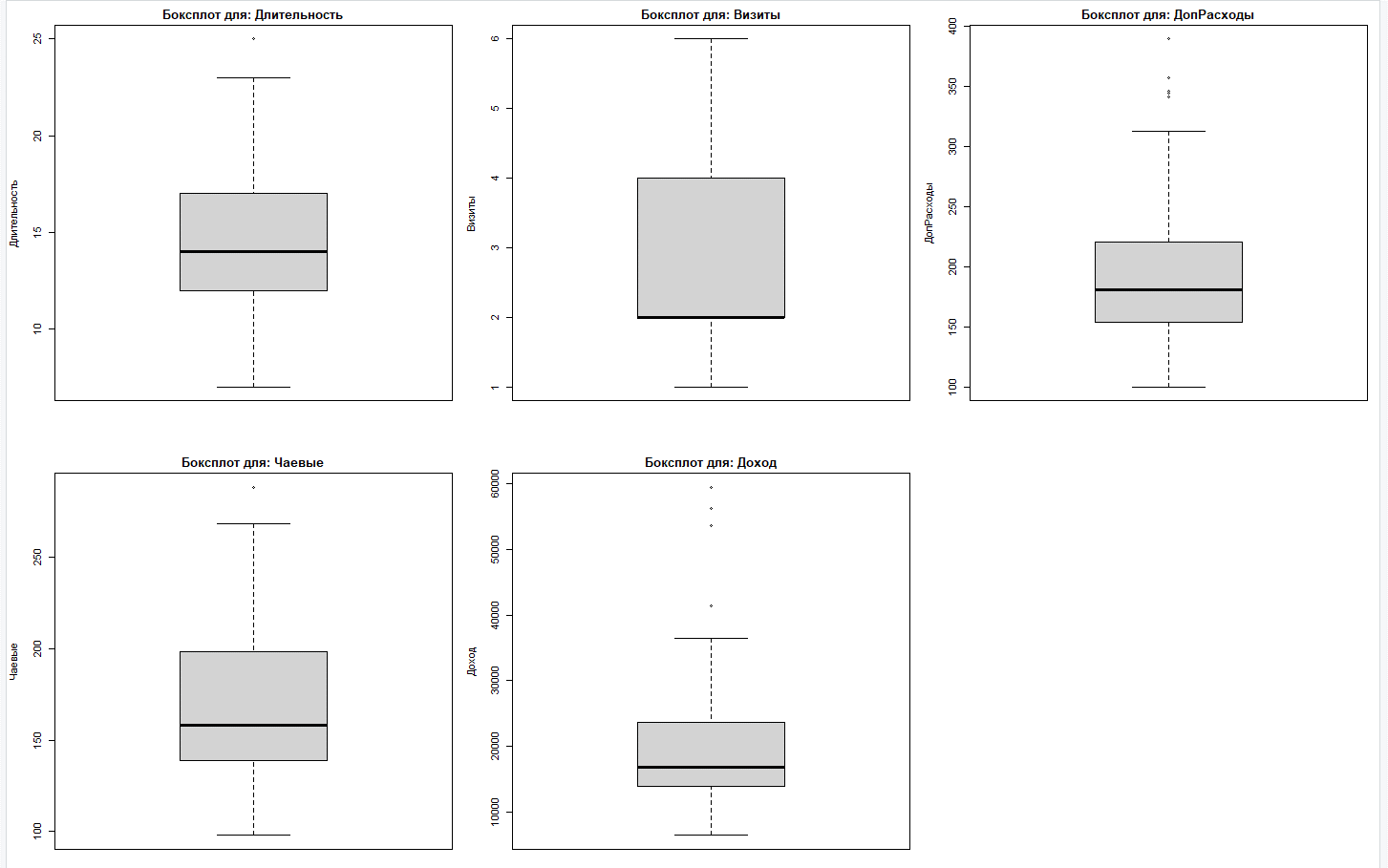


Рисунок 3. Графические результаты дескриптивного анализа

3. Оценим оптимальное количество кластеров. Для этого воспользуемся четырьмя различными методами:

1) Метод локтя: рассматривает общую сумму квадратов внутри кластера относительно количества кластеров.

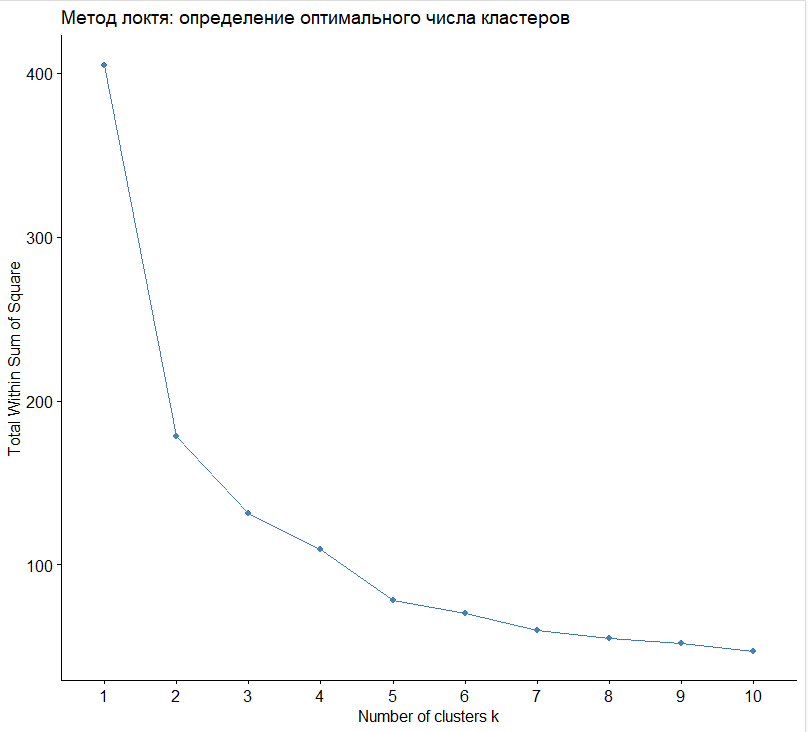


Рисунок 4. График внутригруппового разброса по методу локтя

Согласно графику метода локтя, наибольшее снижение внутрикластерной дисперсии наблюдается при увеличении количества кластеров с 1 до 3. После этого добавление кластеров приводит к менеевыраженному уменьшению ошибки, что указывает на наличие "точки локтя" при k = 3. Таким образом, наиболее обоснованное количество кластеров — 3.

2) Метод среднего силуэта: определяет насколько хорошо каждая точка лежит в пределах своего кластера.

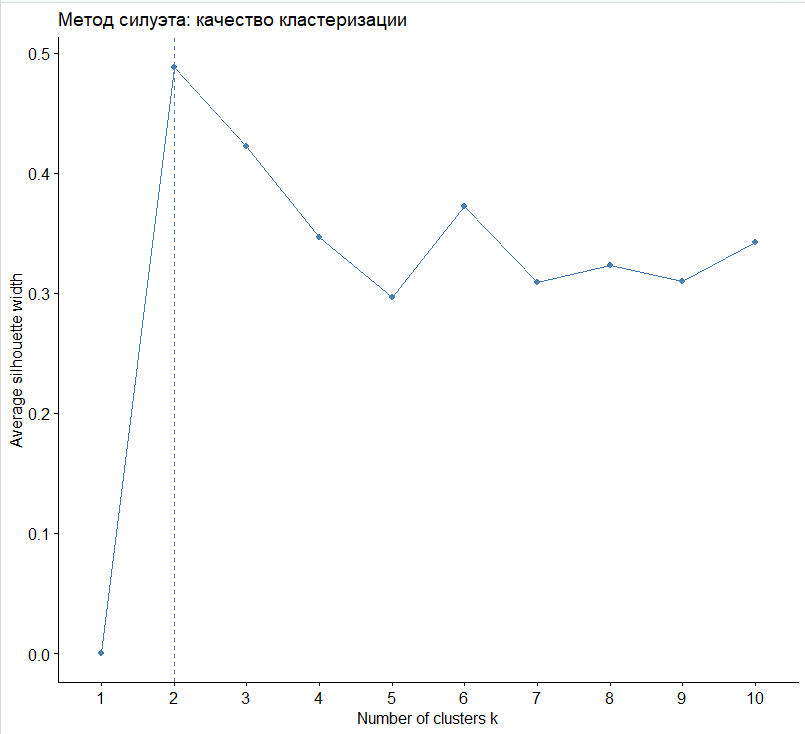


Рисунок 5. График ширины силуэта по методу среднего силуэта

Согласно графику метода силуэта, наилучшее качество кластеризации достигается при k = 2, что соответствует максимальному значению средней ширины силуэта. При k = 3 метрика остаётся высокой, что также делает это значение обоснованным. Начиная с k = 4, качество кластеризации начинает снижаться, а при k = 5 наблюдается резкий спад. При дальнейшем увеличении числа кластеров (k ≥ 6) показатель частично восстанавливается, но не достигает прежнего уровня, оставаясь на среднем уровне. Таким образом, оптимальными являются 2–3 кластера, остальные варианты уступают по качеству.

3) Статистика разрыва: сравнивает общую внутрикластерную дисперсию для различных значений k с их ожидаемыми значениями для распределения без кластеризации.

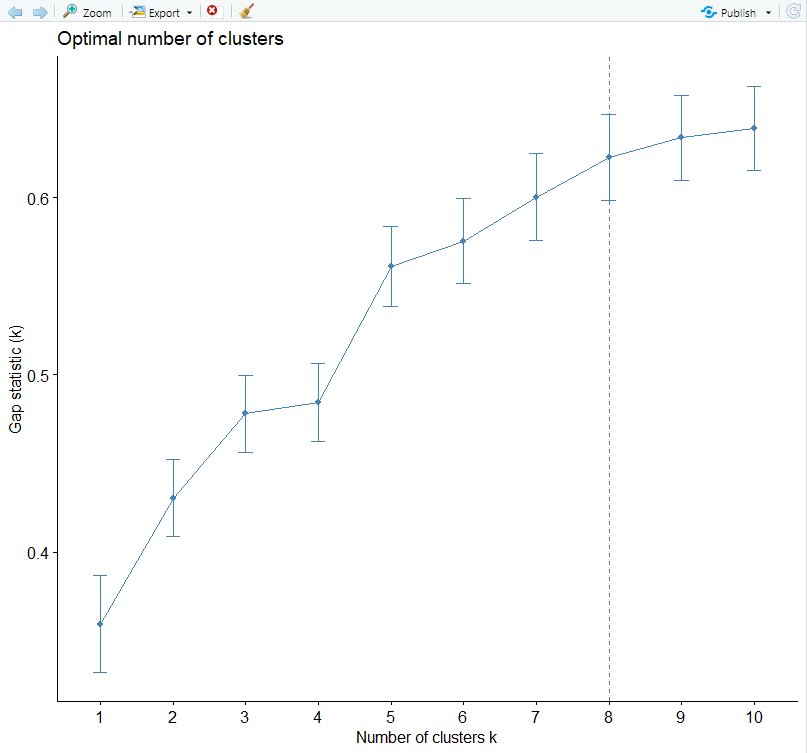


Рисунок 6. График статистики разрыва

Согласно графику статистики разрыва, максимальное значение достигается при k = 10, однако начиная с k = 8 прирост метрики замедляется и находится в пределах стандартной ошибки. Таким образом, оптимальным числом кластеров по методу разрыва является k = 8.

4) Алгоритм на основе консенсуса: выполняет множество различных методов выбора количества кластеров и собирает по ним статистику, показывая наиболее согласованное число.

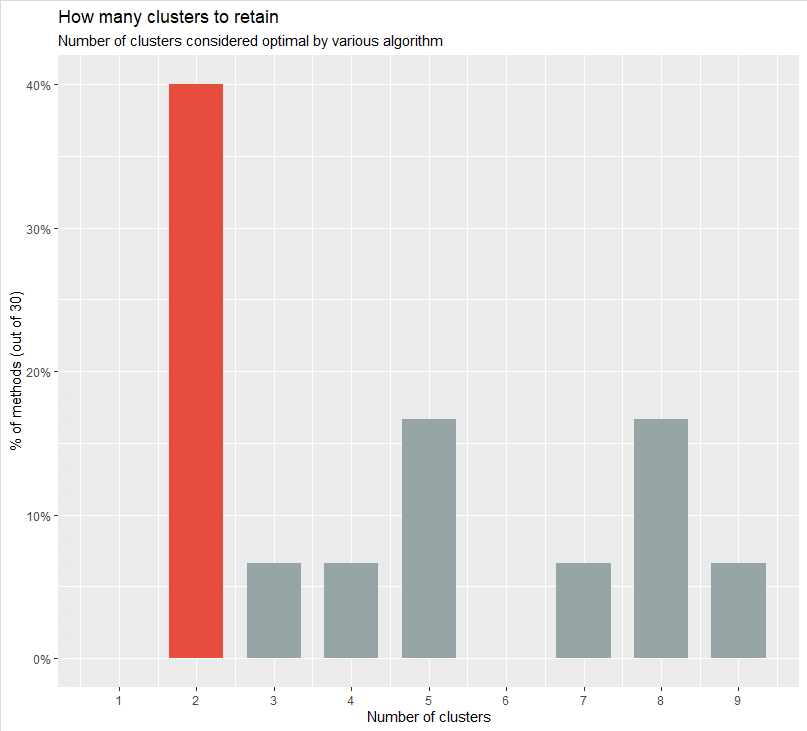


Рисунок 7. Диаграмма консенсуса методов выбора оптимального количества кластеров

Согласно алгоритму консенсуса, наибольшее число методов (около 40%) считают оптимальным числом кластеров k = 2. Также с заметной частотой встречаются значения k = 5 и k = 8, однако они существенно уступают по количеству голосов. Таким образом, наиболее согласованное значение — 2 кластера.

4. Выполним кластеризацию набора данных, построив дендрограмму.

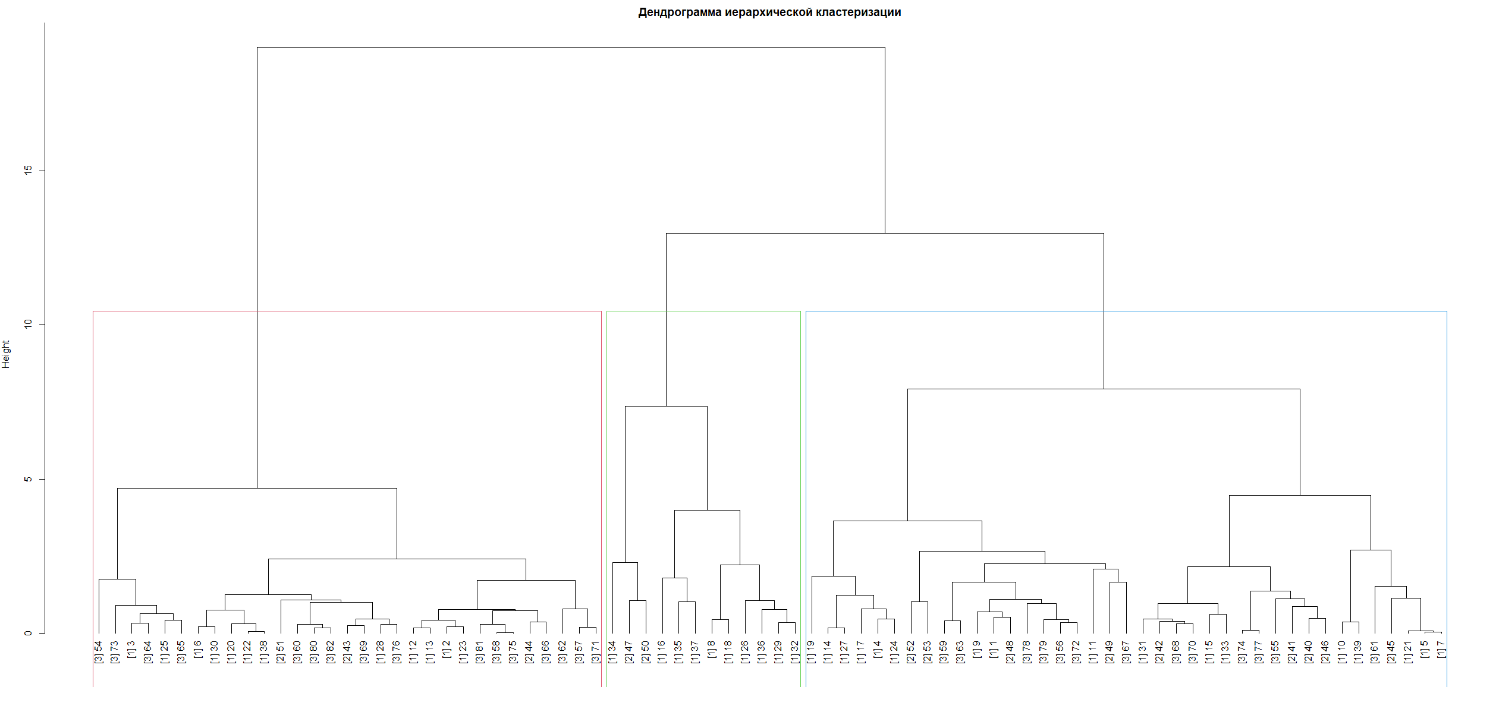


Рисунок 8. Дендрограмма иерархической кластеризации набора данных

При иерархической кластеризации по методу Ward получена дендрограмма, визуально делимая на 3 крупные группы на высоте около 10. Однако состав этих кластеров не строго совпадает с регионами: в каждом из них есть туристы из разных частей Европы. Это говорит о том, что кластеры формируются не по географии, а по поведению и финансовым показателям. Несмотря на это, деление на 3 кластера выглядит осмысленным и удобным для дальнейшего анализа, так как отражает структуру данных и поддерживается методом локтя и силуэта.

5. Построим столбчатые диаграммы и боксплоты средних значений параметров для всех кластеров, полученных в дендрограмме.

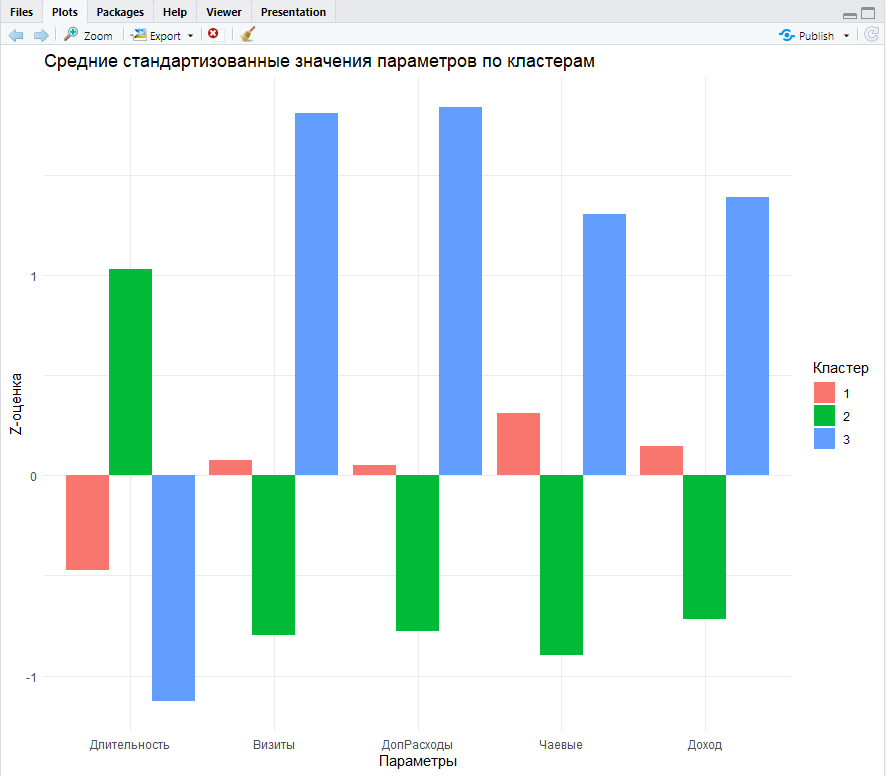


Рисунок 9. Столбчатая диаграмма кластеров дендрограммы

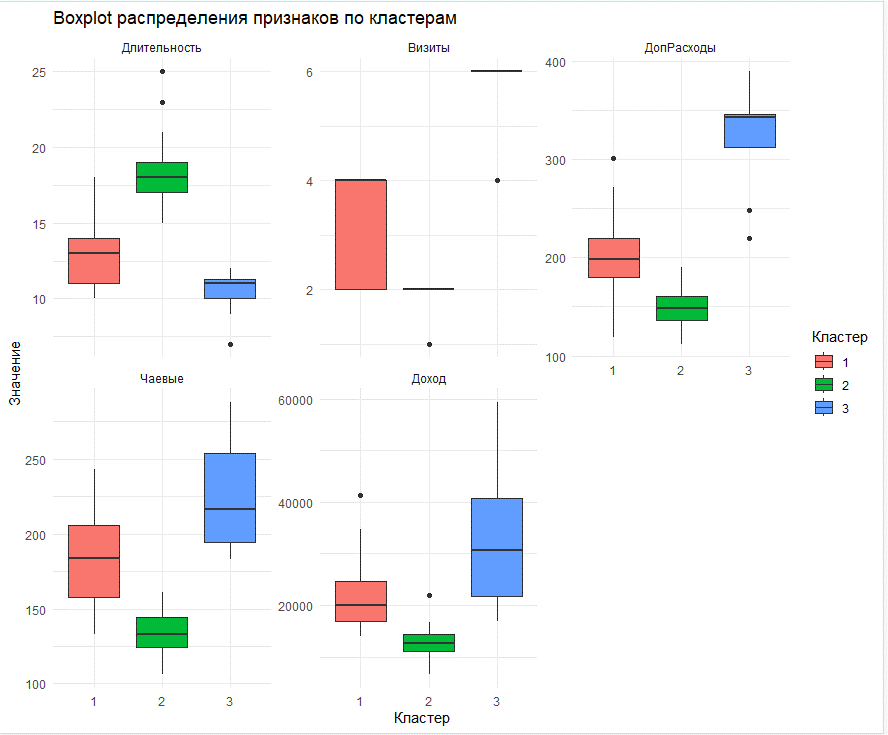


Рисунок 10. Боксплоты кластеров дендрограммы

Проведённый сравнительный анализ кластеров по всем пяти признакам (длительность проживания, число визитов, дополнительные расходы, чаевые и доход) позволил выявить выраженные различия в профилях групп.

Кластер 3, куда в основном вошли туристы из Скандинавии (регион 3), характеризуется наивысшими показателями по доходу, чаевым, дополнительным расходам и числу посещаемых стран. При этом длительность их проживания — наименьшая среди всех групп. Это позволяет охарактеризовать данный кластер как группу финансово обеспеченных и активно перемещающихся туристов, предпочитающих короткие, но насыщенные поездки с высокими расходами.

Кластер 2, включающий преимущественно туристов из Восточной Европы (регион 2), напротив, демонстрирует самые низкие значения по доходу, расходам и числу визитов. Однако при этом в нём зафиксирована наибольшая продолжительность проживания. Это может указывать на более экономный, но длительный отдых, характерный для туристов с ограниченным бюджетом.

Кластер 1, состоящий в основном из представителей Западной Европы (регион 1), занимает промежуточную позицию между двумя другими. Эти туристы демонстрируют сбалансированный стиль путешествий с умеренными расходами, средней продолжительностью проживания и количеством визитов.

6. Выполним кластеризацию набора данных методом К-средних.

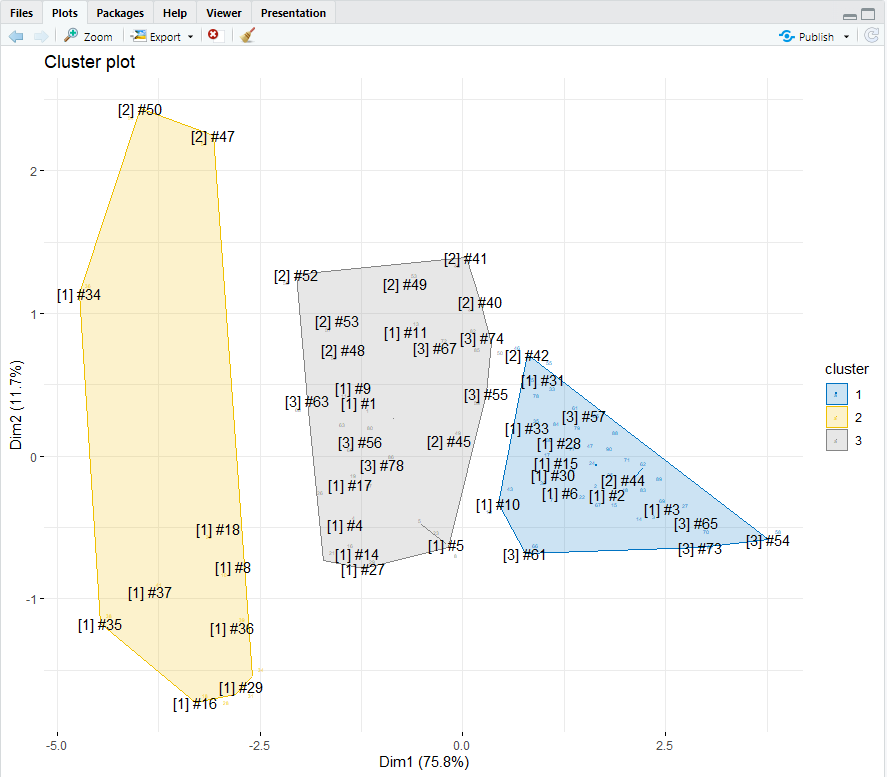


Рисунок 11. Кластеризация методом k-means

При визуализации результатов кластеризации методом k-means с подписями регионов стало очевидно, что состав кластеров не подчиняется жёсткому делению по регионам. Во всех трёх группах наблюдаются представители всех трёх регионов (Западная, Восточная Европа и Скандинавия). Это подтверждает, что кластеризация происходит на основе поведенческих и финансовых характеристик, а не географической принадлежности. Однако, ранее выполненный анализ средних значений показал, что определённые поведенческие типы всё же чаще встречаются в тех или иных регионах.

7. Построим scatterplot для кластеров, полученных методом К-средних.

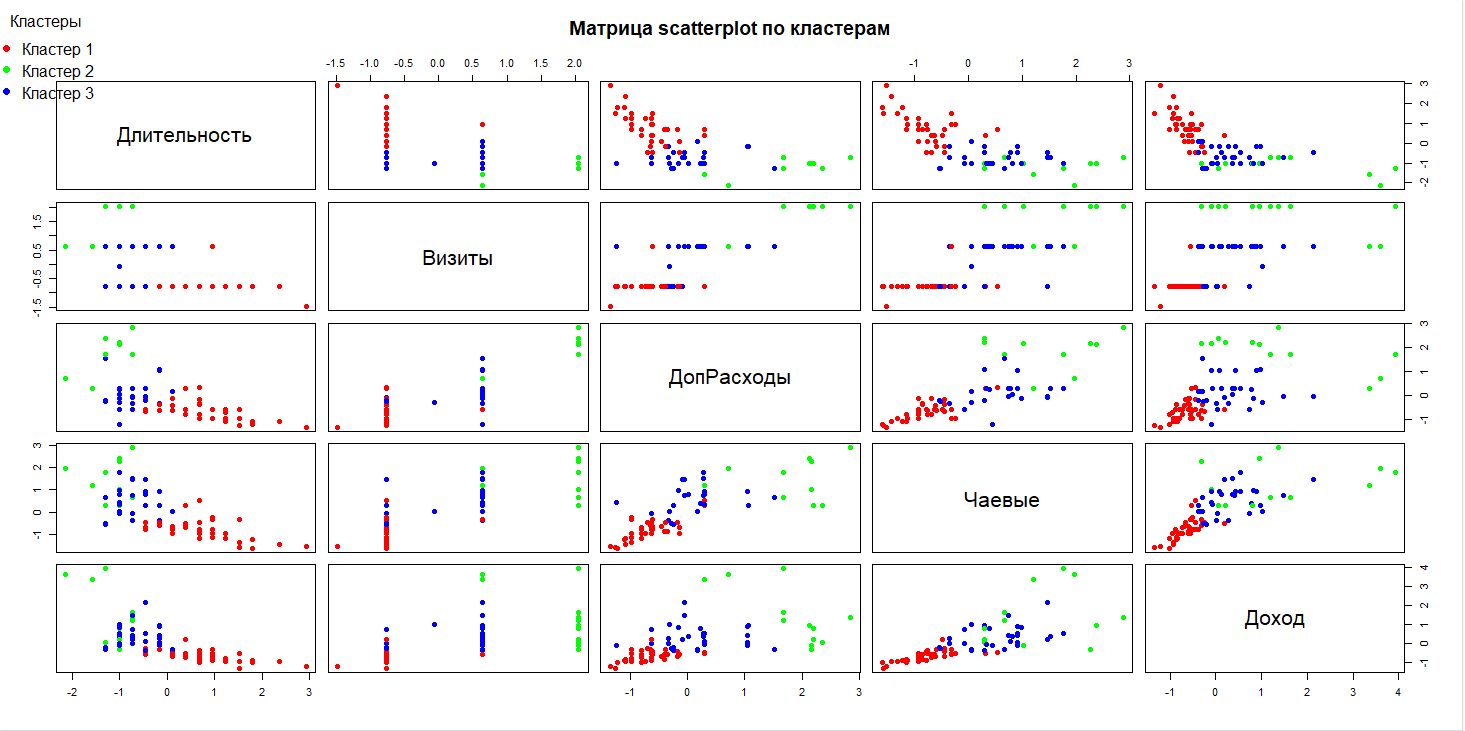


Рисунок 12. Диаграмма рассеяния для кластеров

Анализ диаграмм рассеяния показывает, что кластеры, выделенные методом k-means, демонстрируют различающиеся профили по ключевым признакам.  
Кластер 1 (красный) характеризуется наиболее низкими значениями дохода, чаевых и дополнительных расходов, что указывает на ограниченные финансовые возможности туристов этой группы.  
Кластер 2 (зелёный) демонстрирует наибольшую продолжительность проживания и наибольшую вариативность по финансовым показателям — в том числе присутствие туристов с высокими значениями доходов и расходов.  
Кластер 3 (синий) занимает промежуточное положение между двумя другими и может быть интерпретирован как умеренно обеспеченная и более сбалансированная группа.  
Разделение между группами наиболее чётко прослеживается в проекциях Доход – Чаевые и Доход – Доп. расходы.

8. Модифицируем начальный датасет, добавив в него вектор с полученными в результате кластеризации кластерами, и разделим его на обучающие и тестовые данные в приблизительной пропорции 7 к 3, которые будем в дальнейшем использовать при классификации.

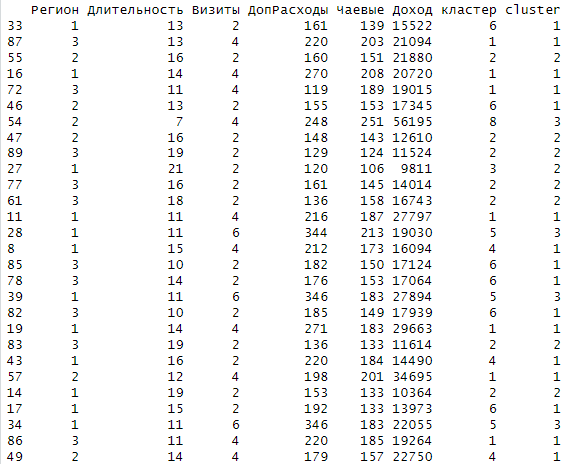


Рисунок 13. Фрагмент обучающих данных датасета

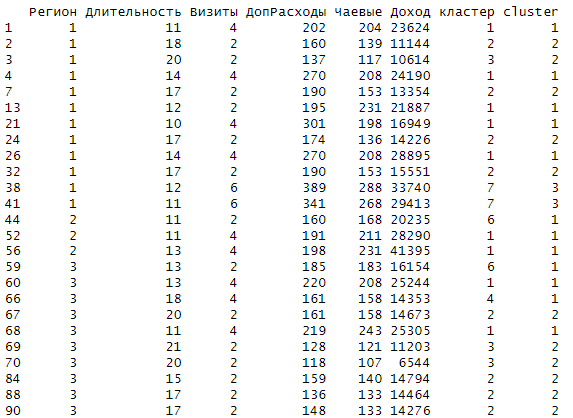


Рисунок 14. Тестовые данные датасета

9. Обучим Байесовский классификатор на обучающих данных. Далее выведем одну из полученных таблиц, например, по параметру *доход*

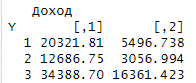


Рисунок 15. Таблица Байесовского классификатора для параметра *длит\_муж*

Первый столбец в таблице отвечает за среднее значение параметра, второй – за его стандартное отклонение. Байесовский классификатор дал следующие оценки распределения дохода по кластерам: в первом кластере средний доход составляет примерно 20 322 USD при стандартном отклонении около 5 497 USD, что указывает на относительно узкий разброс финансовых возможностей участников этой группы. Во втором кластере средний доход заметно ниже — около 12 687 USD при σ ≈ 3 057 USD, что отражает наиболее экономный сегмент туристов. Третий кластер демонстрирует максимальный средний доход (≈ 34 389 USD) с очень высоким стандартным отклонением (≈ 16 361 USD), свидетельствующим о значительной неоднородности в уровне дохода этой группы. Такой результат подтверждает выделение трёх поведенческих сегментов: с низкими, средними и высокими доходами.

10. Проведём прогноз тестовых данных обученным классификатором и оценим полученный результат.

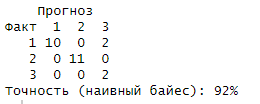


Рисунок 16. Результаты прогноза Байесовского классификатора

Классификатор надёжно распознаёт вторую и третью группы, тогда как в первой группе встречаются единичные ошибки, когда объекты перепутаны с третьим кластером. Это говорит о высокой разделимости данных, но также указывает на близость между кластерами 1 и 3 по ряду признаков.

11. Выполним обучение классификатора на основе дерева решений и построим график полученного дерева.

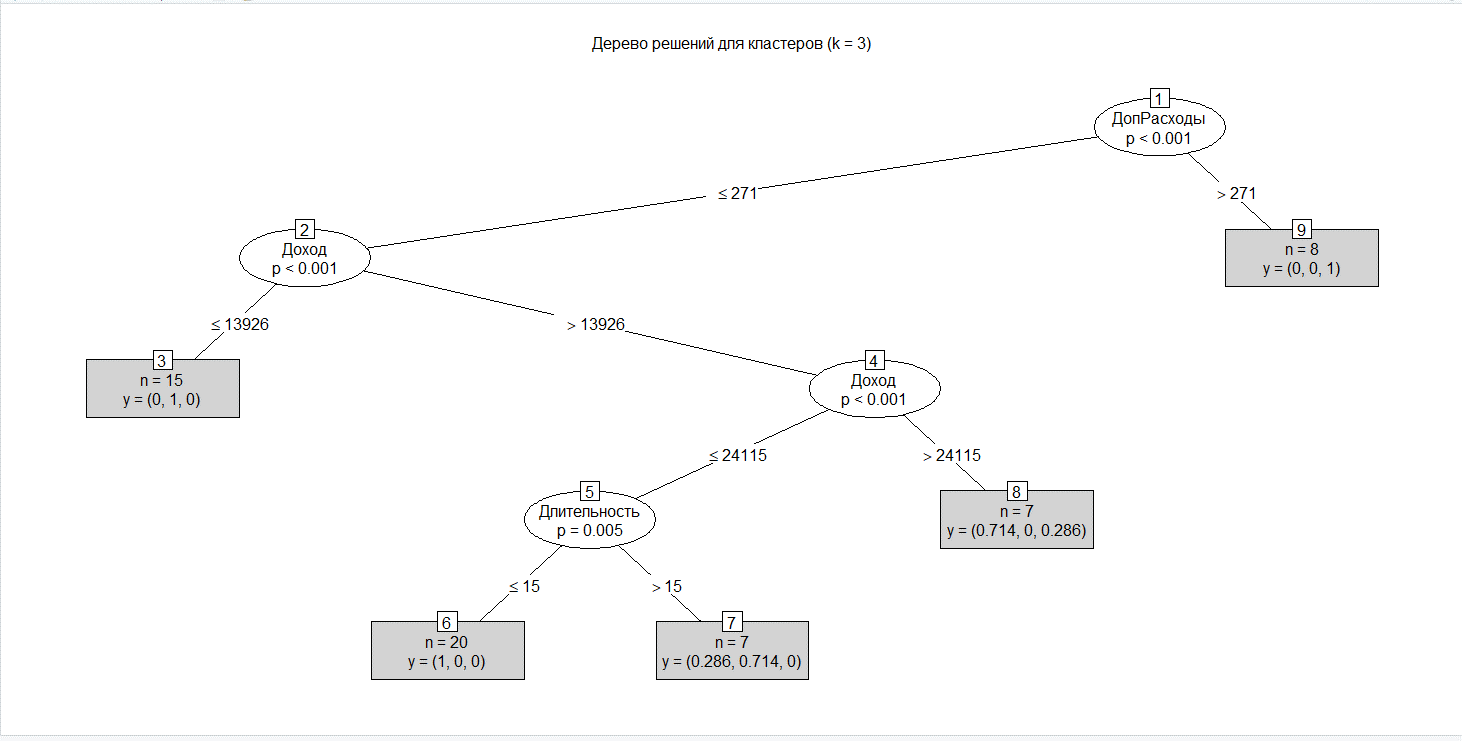


Рисунок 17. График дерева решений

Дерево решений для трёх кластеров показало, что первым и самым информативным признаком при разделении туристов является доход: все наблюдения с доходом до 13 926 USD сразу отнесены к кластеру «экономных» туристов, а при более высоком доходе происходит дальнейшее деление.

В правой ветви дерево снова сравнивает доход с уровнем 24 115 USD: наиболее обеспеченные туристы с доходом выше этой границы оказываются в «высоком» кластере, тогда как при среднем уровне дохода дальше учитывается длительность проживания. Туристы, живущие до 15 дней, относятся к «среднему» кластеру, а более длительное пребывание возвращает их в «экономный» сегмент. Кроме того, для группы с доходом выше 13 926 USD дерево вводит отдельное разбиение по дополнительным расходам (порог 271 USD): высокие траты однозначно классифицируются как «высокий» сегмент. Таким образом, дерево использует всего три признака — доход, дополнительные расходы и длительность — и описывает три чётко различающихся профиля путешественников.

12. Cделаем прогноз на основе классификатора дерева решений и проверим его точность.

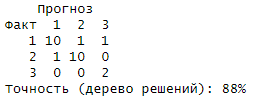


Рисунок 18. Результаты прогноза дерева решений

Дерево решений показало высокую эффективность в разделении трёх кластеров туристов: из 12 объектов первого кластера 10 были классифицированы верно, по одному наблюдению ошибочно отнесено к кластерам 2 и 3; в группе 2 модель допустила одну ошибку, перепутав объект с кластером 1; все объекты третьего кластера определены без ошибок. В результате общая точность составила 88%, что свидетельствует о надёжном распознавании основных поведенческих сегментов на основе выбранных признаков.

13. Выполним обучение классификатора случайного леса для 100 деревьев и проведём на его основе классификацию тестовых данных. Далее проанализируем полученные результаты и сравним с результатами предыдущих методов.

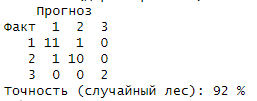


Рисунок 19. Результаты прогноза случайного леса

Сопоставим все полученные результаты классификаторов в таблице:

|  |  |  |  |
| --- | --- | --- | --- |
| Классификатор | Байесовский | Дерево решений | Случайный лес |
| Точность | 92% | 88% | 92% |

В результате сравнения методов классификации наилучшие показатели точности продемонстрировали байесовский классификатор и случайный лес — оба дали результат 92%. Дерево решений показало немного меньшую точность — 88%, но всё же сохранило высокую эффективность. Интересно, что несмотря на простоту, байесовский метод сработал не хуже более сложных моделей. Это говорит о том, что выбранные признаки хорошо разделяют кластеры, и даже базовые алгоритмы справляются с задачей классификации достаточно точно.

**Вывод**: в процессе выполнения данной лабораторной работы я изучил методы кластеризации и подбора оптимального числа кластеров, методы классификации и способы их оценки, а также освоил инструменты языка R для кластеризации и классификации и применил их для анализа набора данных с информацией о параметрах, связанных с уровнем жизни различных стран, выполнив в результате кластеризацию исходного датасета, обучение классификаторов и классификацию данных.