МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ

УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Кафедра вычислительной техники

**Лабораторная работа №2**

**по дисциплине «Методы анализа данных»**

**Решение задач интеллектуального анализа данных (ИАД): кластеризация объектов средствами интегрированной системы Statistica, языка R**

|  |  |
| --- | --- |
| Студент | Великжанин С.И. |
| Вариант | 3 |
| Группа | АВТ-010 |
| Преподаватель | Альсова О.К. |
| Дата сдачи | 27.11.2023 |

Новосибирск, 2023 г.

1. **Цели**

* изучить алгоритмы и методы кластерного анализа данных на примере решения конкретной задачи ИАД;
* исследовать эффективность использования различных алгоритмов и методов кластерного анализа данных для решения прикладной задачи;
* ознакомиться и получить практические навыки работы с модулями интегрированной статистической системы Statistica, языка R, реализующими решение задачи кластеризации объектов.

1. **Постановка задачи ИАД**

Изучаются показатели работы программистов крупной организации.Необходимо выделить группы (классы) программистов в соответствии с рядом устойчивых признаков. Разделение программистов на группы проводится с целью установления и обоснования тарифных разрядов и размера заработной платы для каждой выделенной группы. Также требуется построить правило отнесения программиста к одной из выделенных групп (классов).

Исходные данные для проведения статистического анализа представлены в табл. 1. Рассматриваются следующие показатели (признаки) для каждого программиста:

* возраст;
* время написания первой тестовой программы, в час.;
* время написания второй тестовой программы, в час.;
* стаж работы по специальности в данной организации;
* образование (непрофильное – 0; профильное - 1);

наличие сертификатов о повышении квалификации в области программирования и программного обеспечения (0 –сертификаты отсутствуют; 1- сертификаты 1-го уровня; 2 - сертификаты 2-го уровня, 3-  сертификаты 3-го уровня).

**вариант 3.**

*Таблица 1.*

***Исходные данные (фрагмент)***

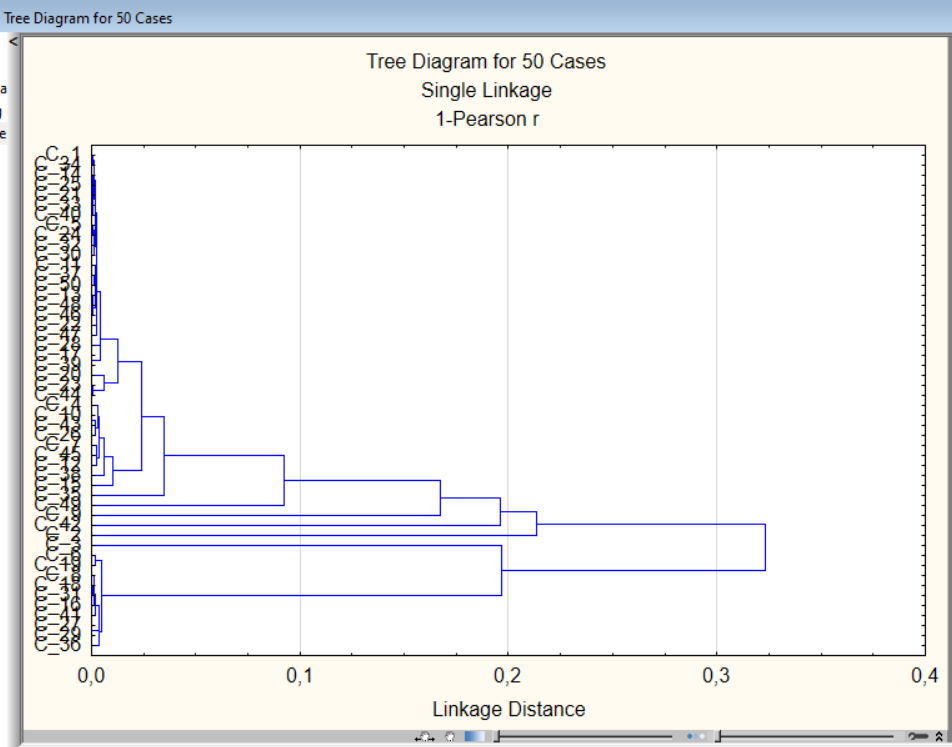
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **номер** | **возраст** | **время   тест 1** | **время   тест 2** | **стаж** | **образование** | **сертификат** |
| 1 | 39 | 5,4 | 2,8 | 0,8 | 0 | 0 |
| 2 | 31 | 2,0 | 1,7 | 3,5 | 1 | 2 |
| 3 | 32 | 2,4 | 1,8 | 6,1 | 1 | 3 |
| 4 | 36 | 3,4 | 1,8 | 3,3 | 1 | 2 |
| 5 | 29 | 5,4 | 2,7 | 0,7 | 0 | 0 |
| ... | … | … | … | … | … | … |
| 50 | 39 | 5,4 | 2,6 | 0,7 | 1 | 0 |

Таблица 2.

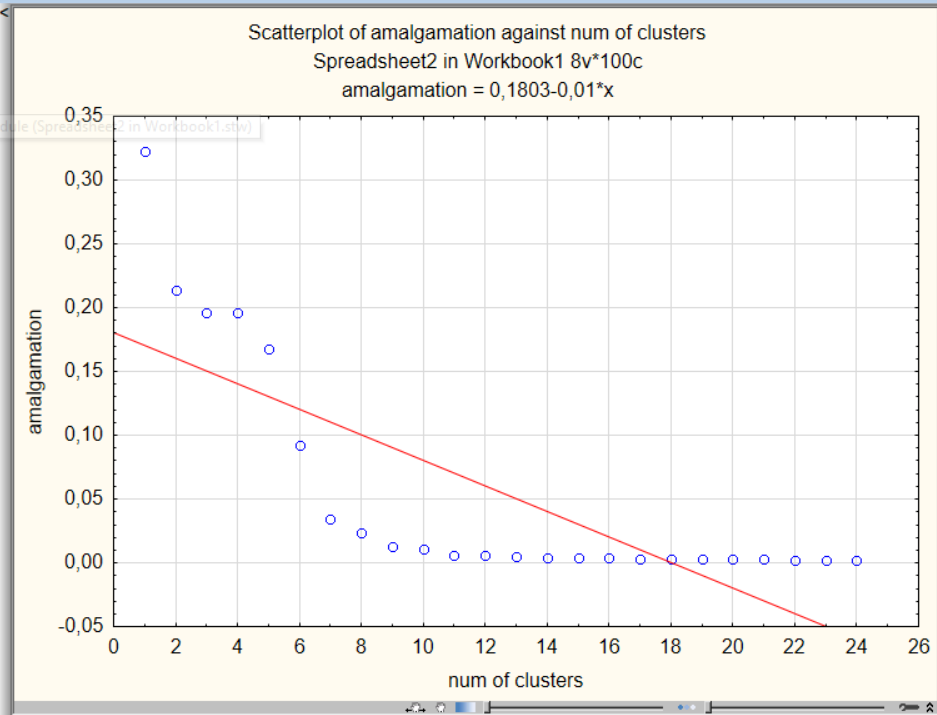
|  |  |  |
| --- | --- | --- |
| **Номер варианта** | **Иерархический метод** | **Мера расстояния** |
| 3. | *1. Single Linkage – одиночная связь*  *2. Unweighted pair-group average – не взвешенное попарное среднее* | *1-Pearson r – коэффициент корреляции Пирсона*  *2. Power: SUM(ABS(x-y)p)*1/r *степенное расстояние* |

1. **Ход работы**

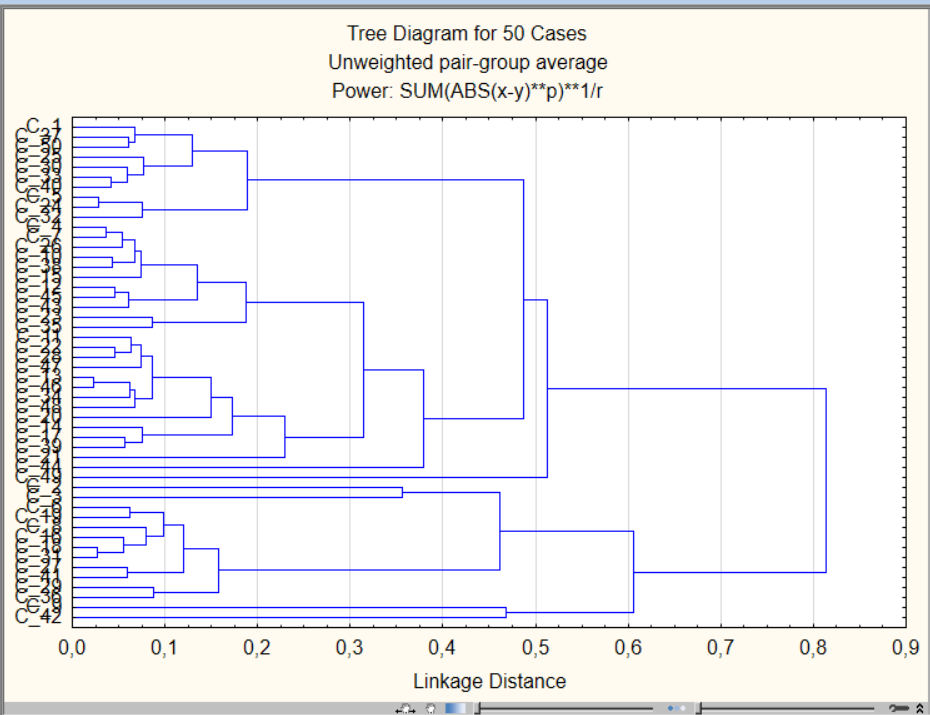
Результаты применения иерархических методов кластеризации:



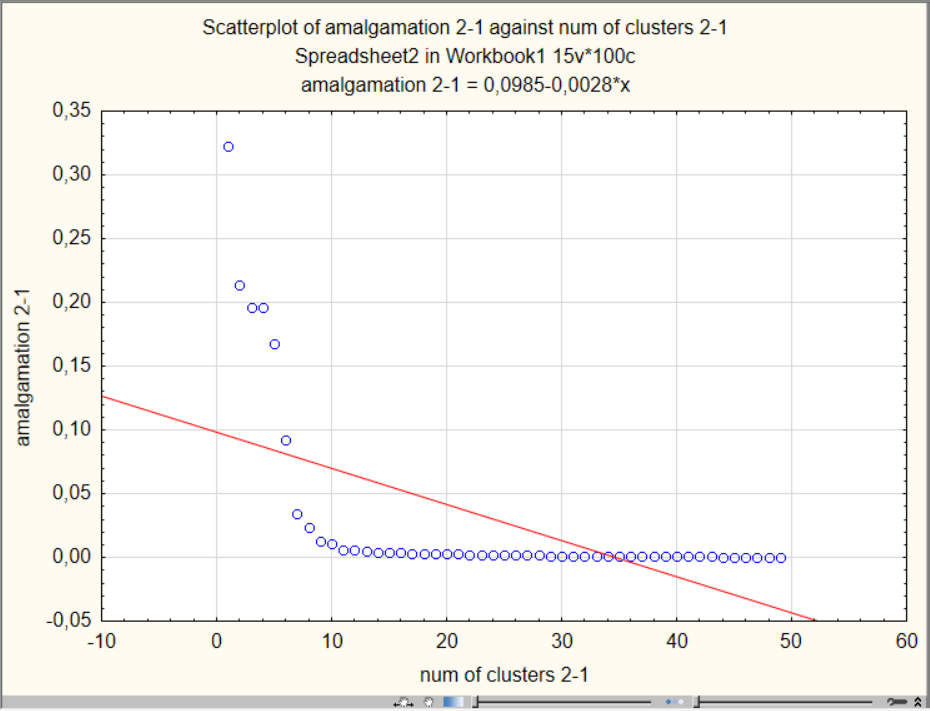
*Рис. 1 Дендрограмма метода одиночной связи с коэф. корр. Пирсона как мера расстояния*



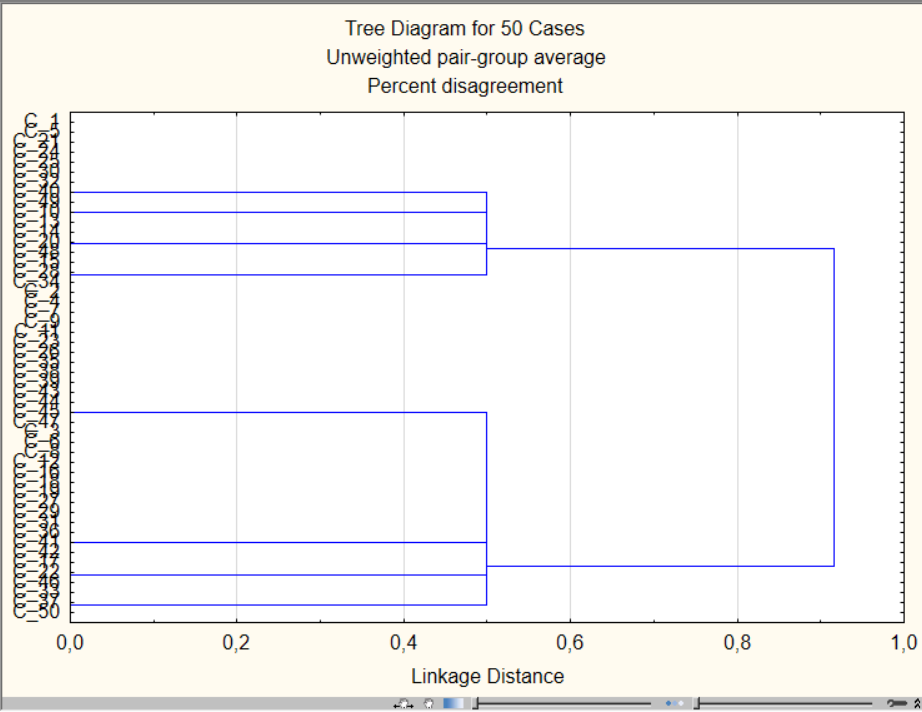
*Рис. 2 График зависимости числа кластеров от коэф. слияния для первой комбинации*

**

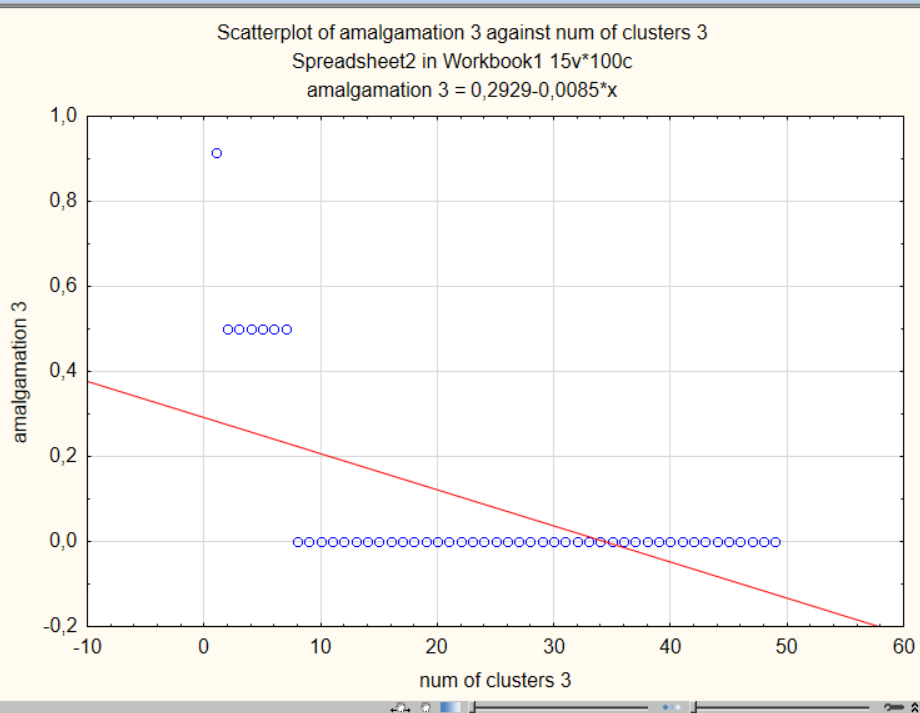
*Рис. 3 Дендрограмма метода невзвешенное попарное среднее с степенным расстоянием*

**

*Рис. 4 График зависимости числа кластеров от коэф. слияния для второй комбинации*

**

*Рис. 5 Дендрограмма метода невзвешенное попарное среднее с процентом несогласия (качественные признаки)*

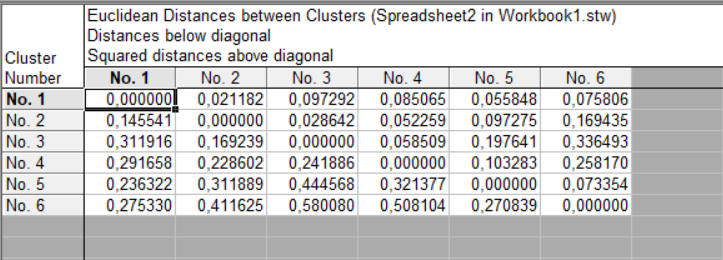
**

*Рис. 6 График зависимости числа кластеров от коэф. слияния для качественных признаков*

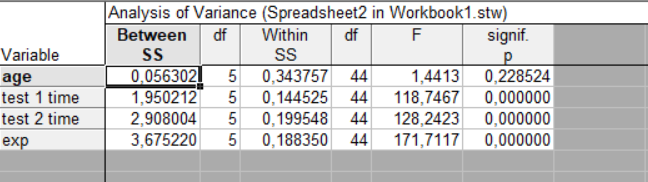
*Таблица 3.*

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **метод** | **мера расстояния** | **количество кластеров** | **Количество объектов в кластере** | | | | | | | | | **достоинства, недостатки метода и меры расстояния** |
| **кластер 1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| Одиночная связь | Коэф. Корр. Пирсона | 6 | 12 | 14 | 10 | 1 | 2 | 11 | - | - | - |  |
| Невзвешенное попарное среднее | Степенная | 9 | 11 | 1 | 2 | 13 | 1 | 2 | 6 | 4 | 10 | Увеличилось число кластеров с малым кол-вом объектов |
| Невзвешенное попарное среднее | Процент несогласия | - | - | - | - | - | - | - | - | - | - | Нет закономерностей, кластеризация по качественным признакам не сработала |

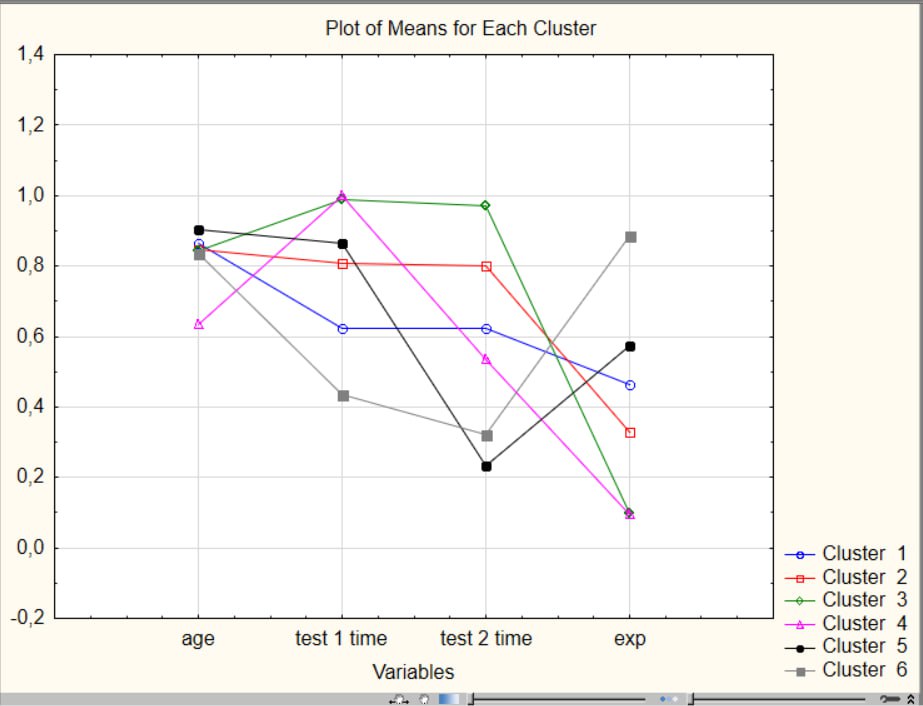
Кластеризация объектов по методу К-средних:



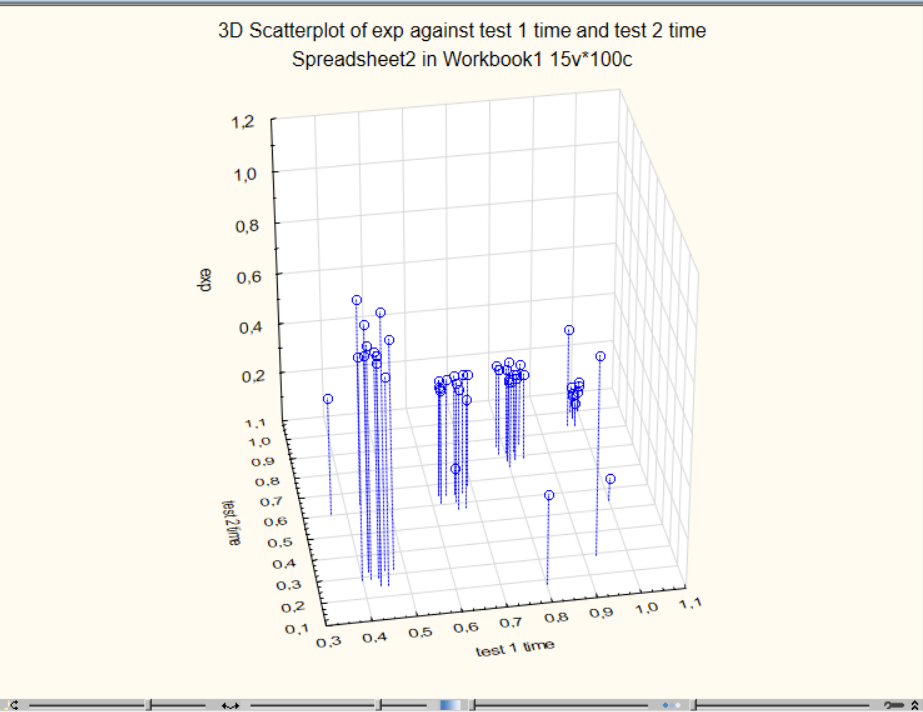
*Рис. 7 Матрица евклидовых расстояний между кластерами*

**

*Рис. 8 Значимость влияния каждого из признаков*

**

*Рис. 9 График средних значений признаков по каждому кластеру*

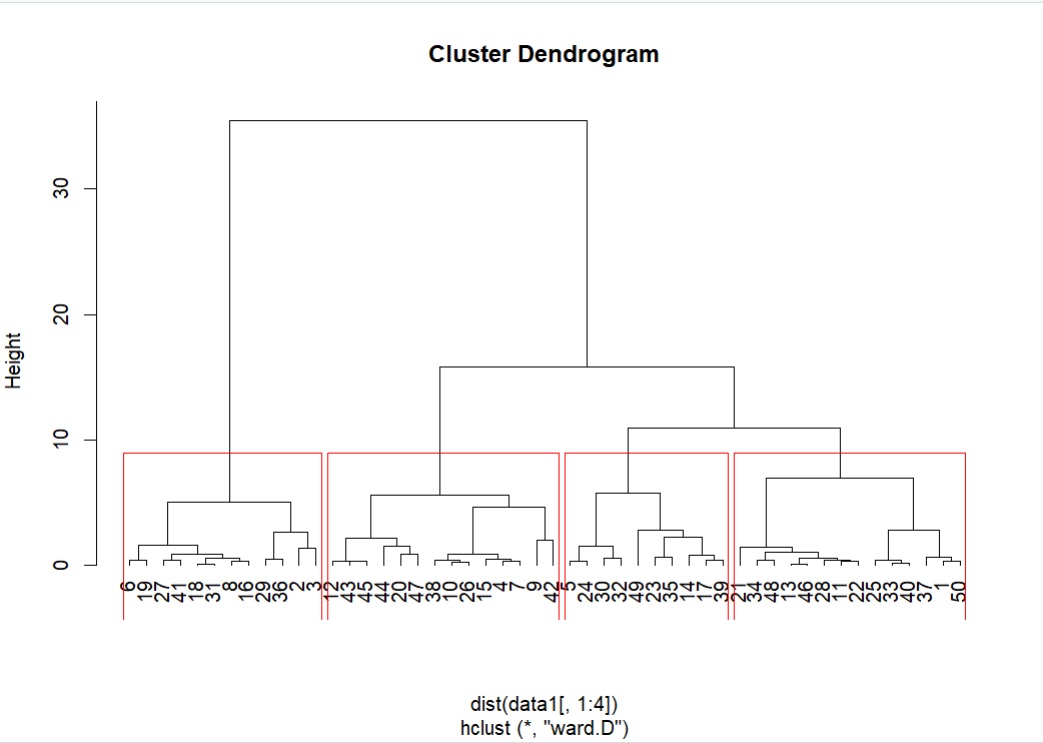
**

*Рис. 10 Диаграмма рассеяния в пространстве трех признаков*

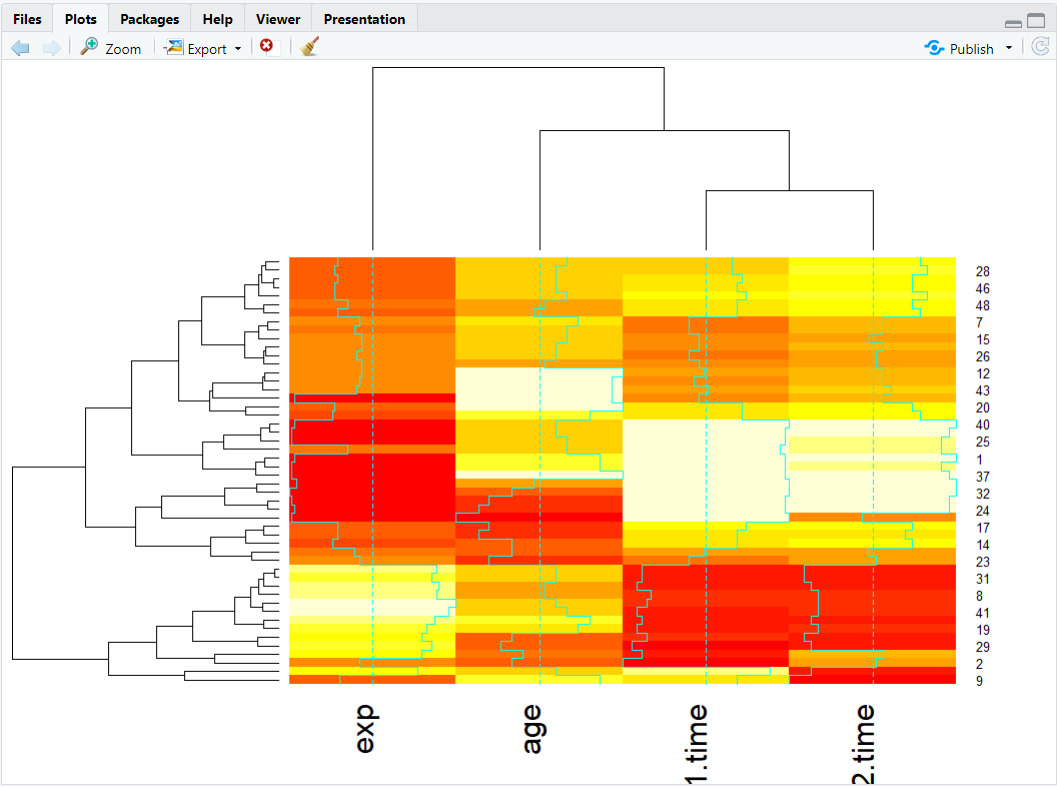
*Таблица 3.*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **метод** | **мера расстояния** | **количество кластеров** | **Количество объектов в кластере** | | | | | | **достоинства, недостатки метода и меры расстояния** |
| **кластер 1** | **2** | **3** | **4** | **5** | **6** |
| К-средних | Евклидово расстояние | 6 | 12 | 14 | 10 | 1 | 2 | 11 | - |

Кластеризация в R

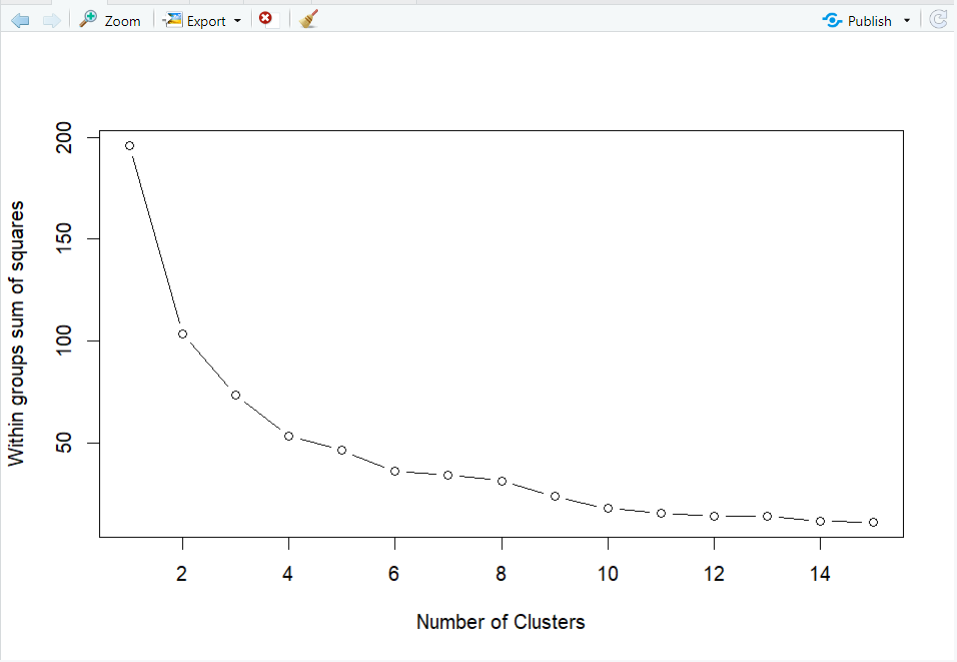


*Рис. 11 Дендрограмма иерархического метода*

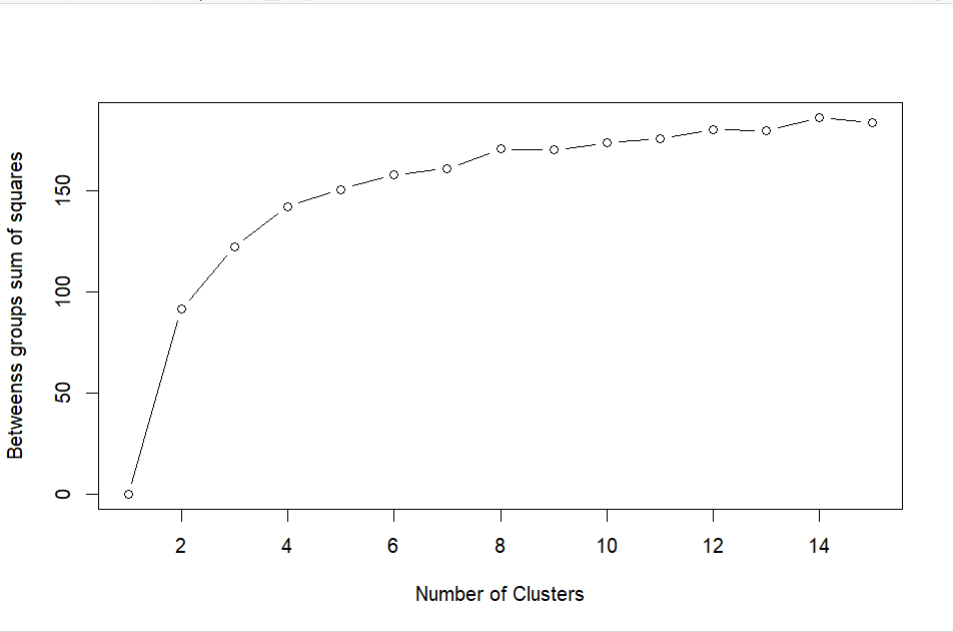
**

*Рис. 12 Тепловая карта*

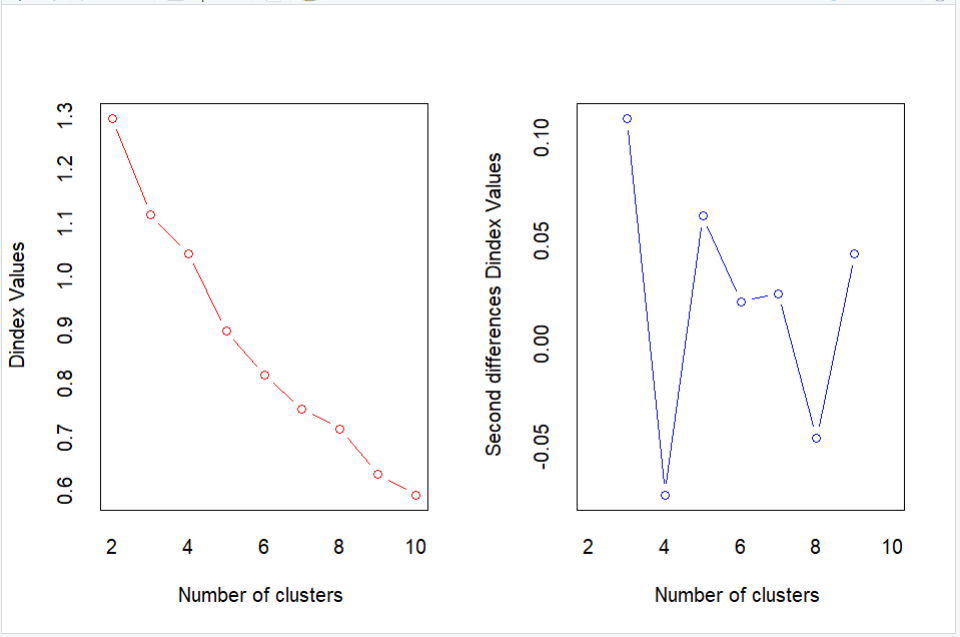
*C:\Users\magaz\Desktop\архив\5 kofen corr.png*

**

*Рис. 13 График зависимости суммы квадратов отклонений внутри кластера от кол-ва кластеров*

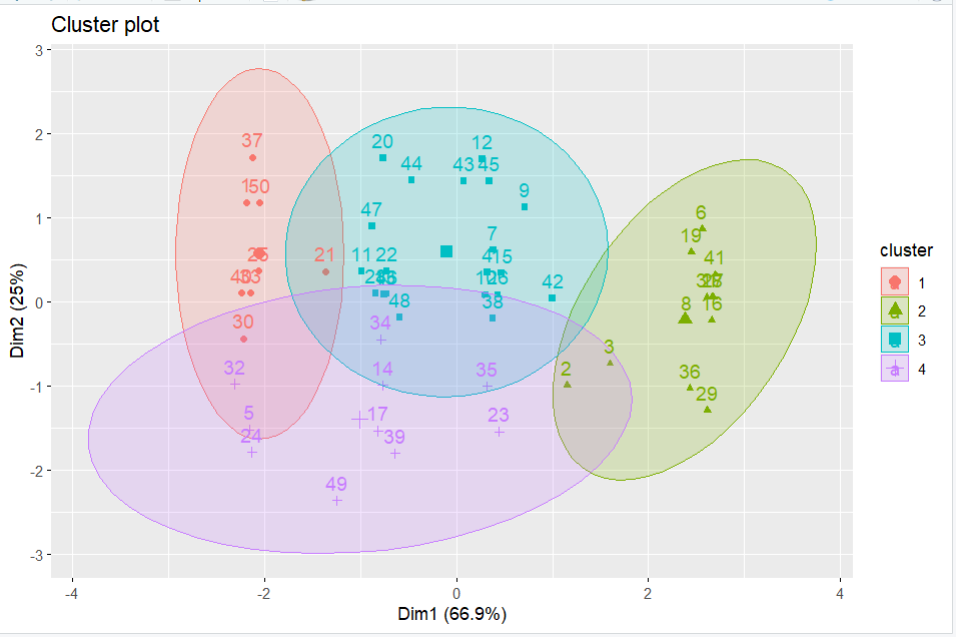
**

*Рис. 14 График зависимости суммы квадратов отклонений между кластерами от кол-ва кластеров для метода К-средних*

**

*Рис. 15 Графики изменения индексов качества кластеризации от кол-ва кластеров*

Анализ полученных результатов:



*Рис. 16 Оптимальное кол-во кластеров - 4*

В ходе выполнения работы были выявлены 4 кластера в данных варианта:

1. Красный – старшие программисты с малым стажем и наибольшим временем выполнения тестовых заданий. Большинство не имеют сертификатов о повышении квалификации.
2. Зеленый – старшие программисты с большим стажем и меньшим временем выполнения тестовых заданий. Большинство имеют сертификаты о повышении квалификации 3 уровня (максимальный).
3. Голубой – старшие программисты с средним стажем и средним временем выполнения тестовых заданий. Большинство имеют сертификаты 2 уровня.
4. Фиолетовый – молодые программисты (<30) с малым стажем работы и большим временем выполнения заданий. Несмотря на слабые показатели при тестах, большинство опережают программистов красного кластера.
5. **Вывод**

В ходе выполнения работы были изучены алгоритмы и методы кластерного анализа данных, исследована их эффективность при решении прикладной задачи. Получены навыки работы с модулями решения задачи кластеризации объектов в Statistica и языке R.

* Наиболее точным методом кластеризации при решении данной задачи был метод К – средних, который позволил четко объединить всех программистов в 4 группы по явным признакам.
* Иерархические методы одиночной связи и невзвешенного попарного среднего дали большее число кластеров, состоящих из 1-4 элементов, так как некоторые количественные значения сильно отличаются от средних.
* Сделан вывод, что кластерный анализ лучше проводить на количественных признаках, так как качественные не дают нужного разброса данных для выявления закономерностей. Однако качественные признаки помогают при описании кластеров, созданных при анализе по количественным признакам.