МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКО ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ

УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ НИЖЕГОРОДСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМ. Н.И. ЛОБАЧЕВСКОГО»**

Дзержинский филиал

**Кафедра** Прикладная информатика

**ОТЧЕТ**

**по дисциплине «Интеллектуальный анализ данных»**

**к лабораторной работе № 1**

**Тема «R»**

Выполнил:

Студент гр. № 2721Б1ПИ

Погодин Дмитрий Евгеньевич

Проверил:

Ст.преподаватель, к.т.н.

\_\_\_\_\_\_\_\_\_/Нажимов А.В./ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Дзержинск 2024

Оглавление

[**Лабораторная работа №1.1** 3](#_Toc135056069)

[**Лабораторная работа №1.2** 6](#_Toc135056070)

[**Лабораторная работа №1.3** 14](#_Toc135056071)

[**Вывод** 23](#_Toc135056072)

# **Лабораторная работа №1.1**

**Метод k-средних**

**Цель**

Построить график, импортировать библиотеки для построения, применить метод k-средних на нескольких примерах.

**Ход выполнения работы**

1. Добавим библиотеки cluster, ggplot2, factoextra. Для этого в пакетах найдём их по поиску и загрузим
2. Код для графика:

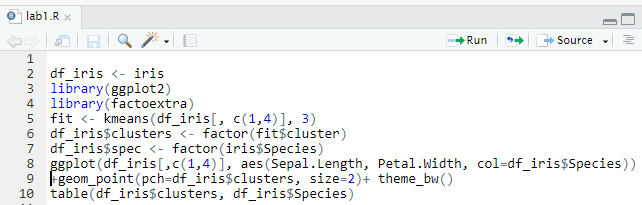


Рис. 1. Код для графика

1. Выделим весь код и получим график:

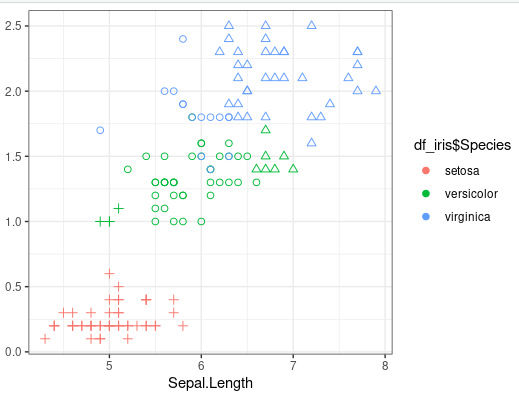


Рис. 2. Визуализация графика

1. Построим отображение кластеров используя команду fviz\_cluster(fit, df\_iris[,c(1,4)]):

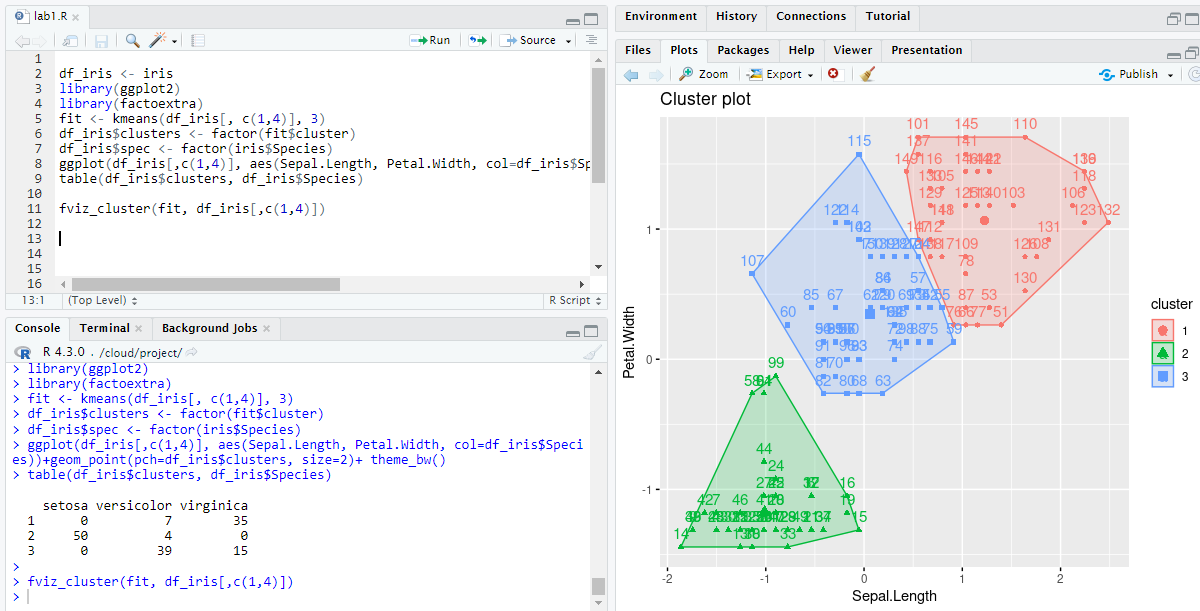


Рис. 3. Отображение кластеров

1. Опишем индивидуальные результаты. Попробуем использовать для кластеризации всю доступную информацию, но график построим для двух переменных Sepal.Length, Petal.Width:

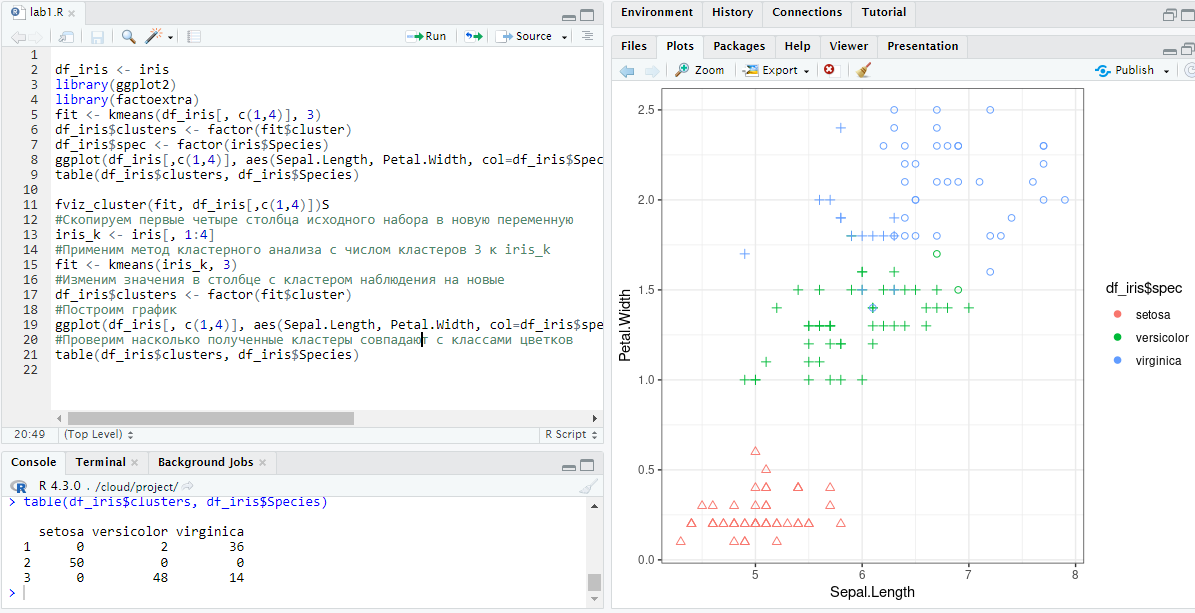


Рис. 4. Распределение ирисов по классам и кластерам

1. Строим Fviz\_cluster:

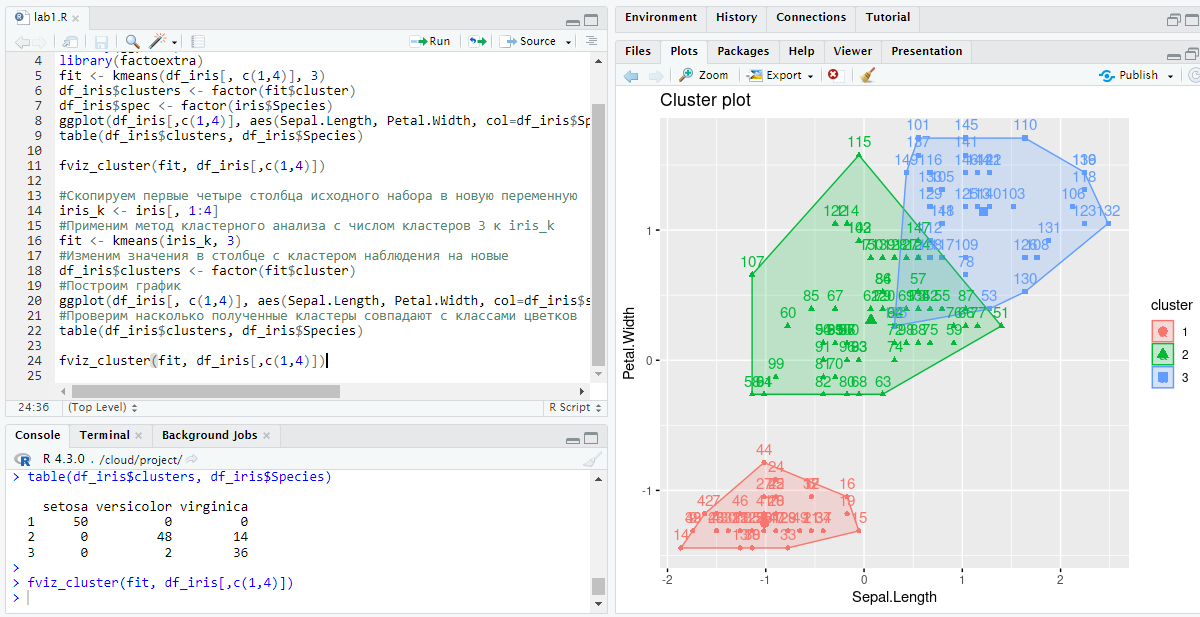


Рис. 5. Отображение кластеров с командой fviz\_cluster

# **Лабораторная работа №1.2**

**Набор данных медицинское страхование**

**Цель**

Рассмотрим теперь более корректный случай, когда изначально разметки у данных по классам нет, и мы не знаем, какое количество кластеров нужно выбрать.

Данные, которые мы используем – расходы на медицинское обслуживание тех, кто имеет медицинскую страховку: insurance.csv

**Ход выполнения работы**

1. Загрузим набор данных:

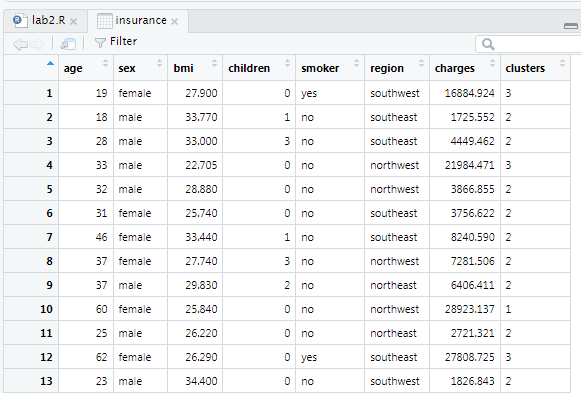


Рис. 6. Набор данных insurance

1. Построим график:

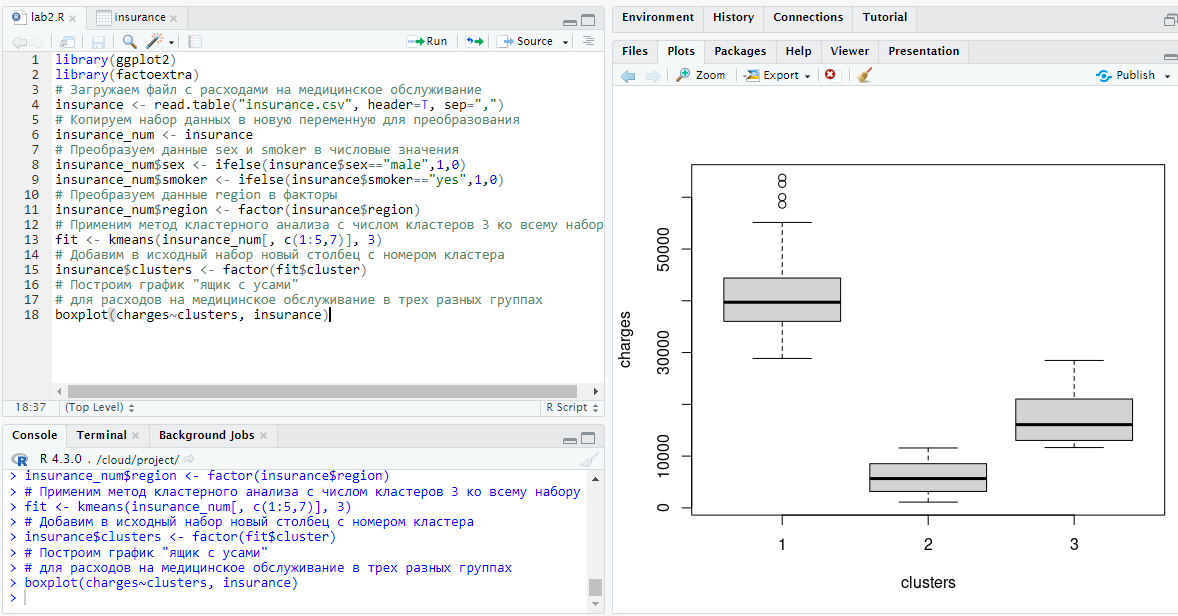


Рис. 7. Распределение застрахованных по кластерам в зависимости от расходов на медицинское обслуживание (3 кластера)

1. Оценить, как распределились застрахованные по трём кластерам по переменной smoker используя команду: table(insurance$smoker, insurance$clusters)

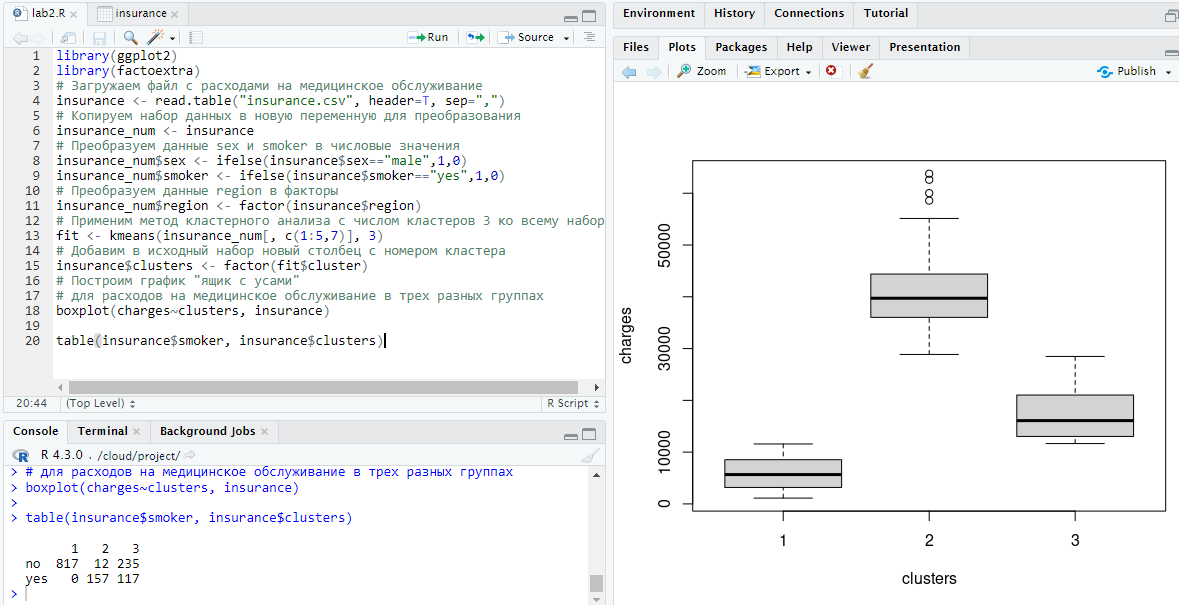


Рис. 8. Распределение застрахованных по трём кластерам переменной smoker

Из распределения видно, что

1. Распределение по sex:

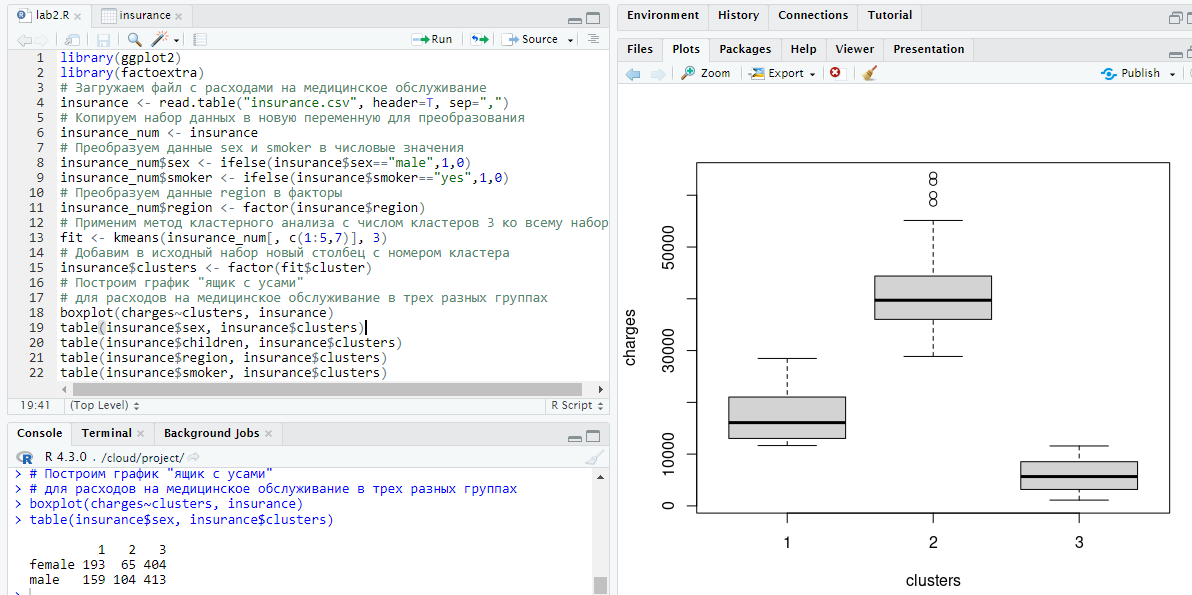


Рис. 9. Распределение застрахованных по трём кластерам переменной sex

1. Распределение по children:

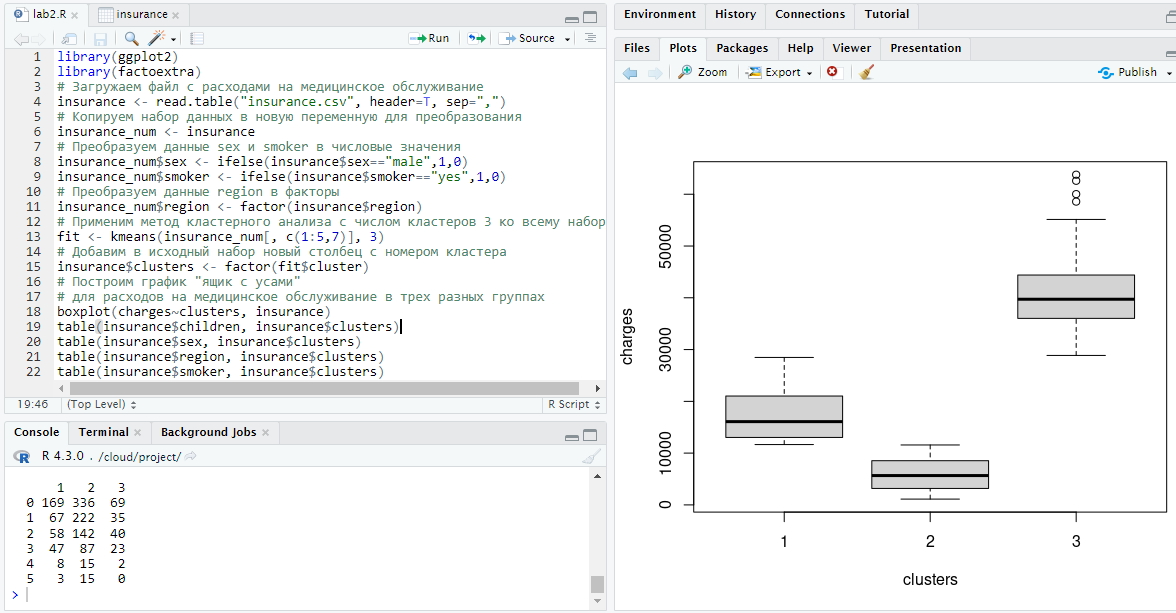


Рис. 10. Распределение застрахованных по трём кластерам переменной children

1. Распределение по region:

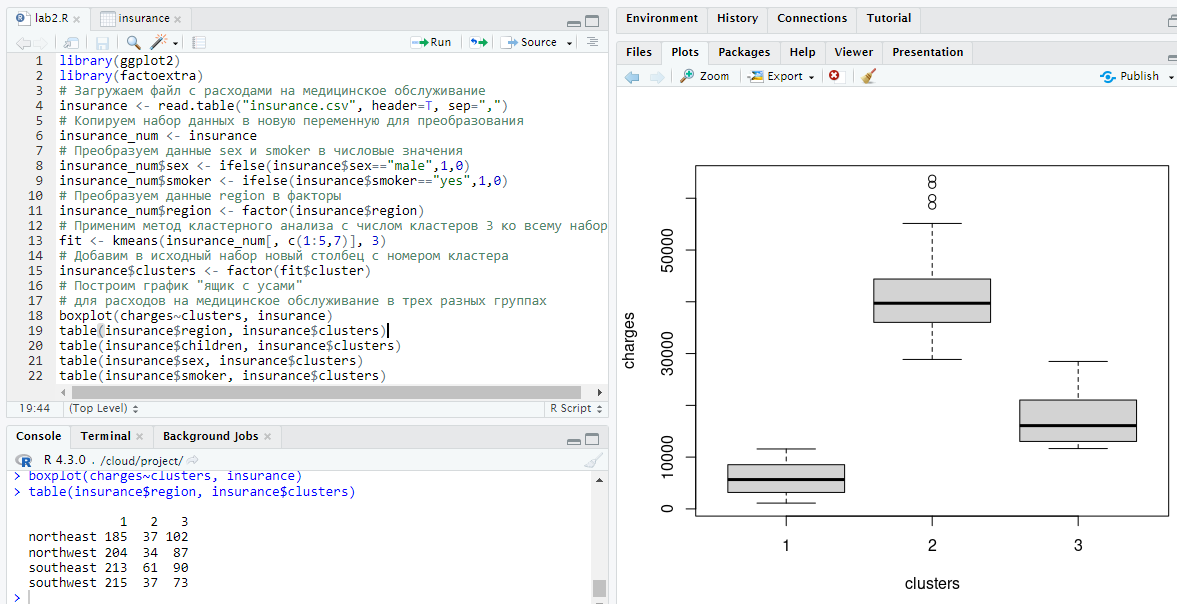


Рис. 11. Распределение застрахованных по трём кластерам переменной region

1. Построение отображений используя fviz\_cluster:

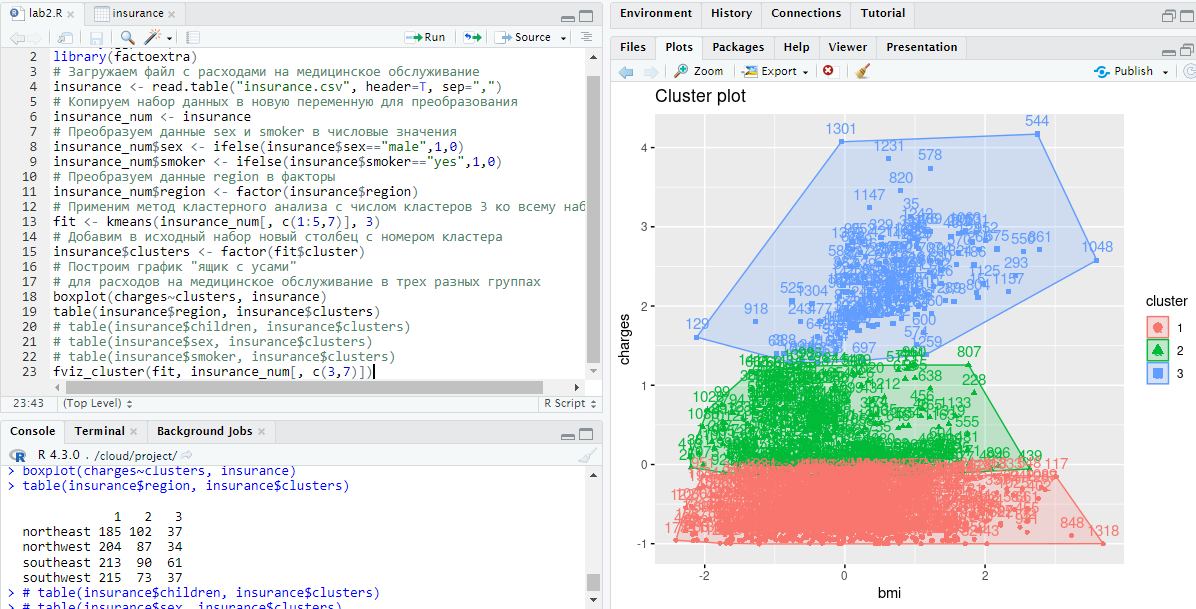


Рис. 12. Распределение по region столбцов bmi и charges

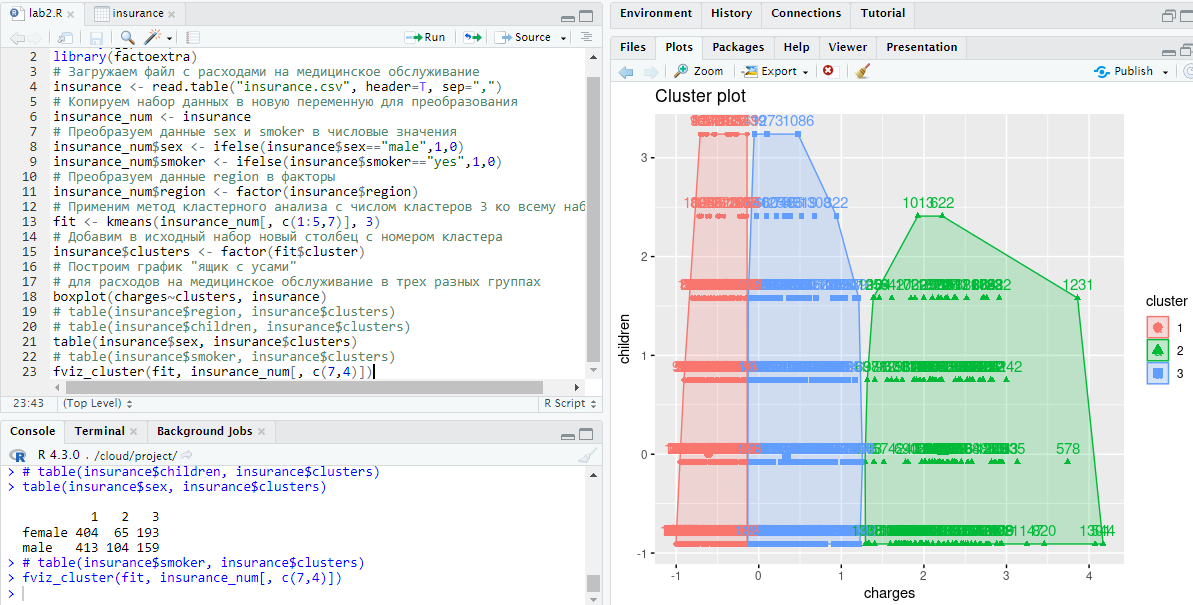


Рис. 13. Распределение по sex столбцов charges и children

1. Выясним лучшее количество кластеров с помощью метода локтя

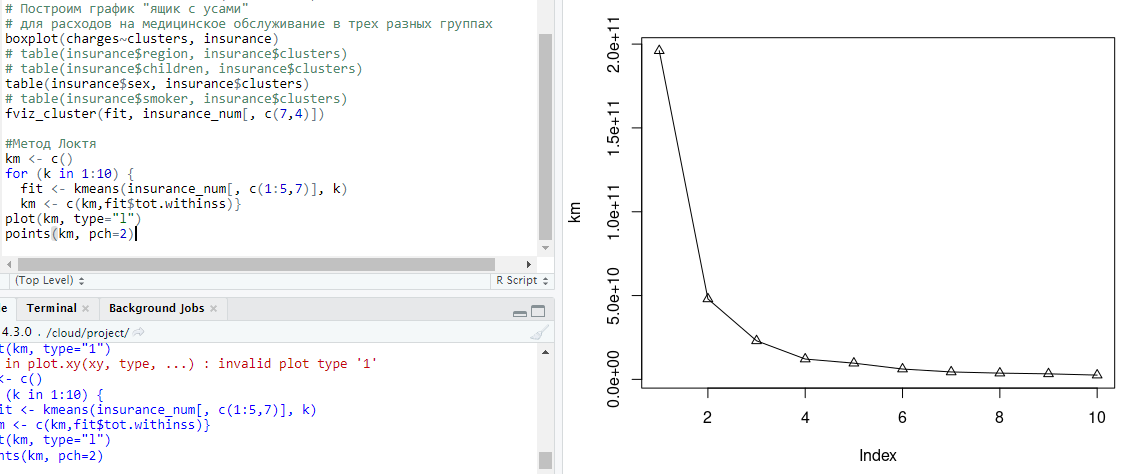


Рис. 14. График для выявления оптимального числа кластеров (insurance)

1. Можно попробовать разбиение на 4 кластера:

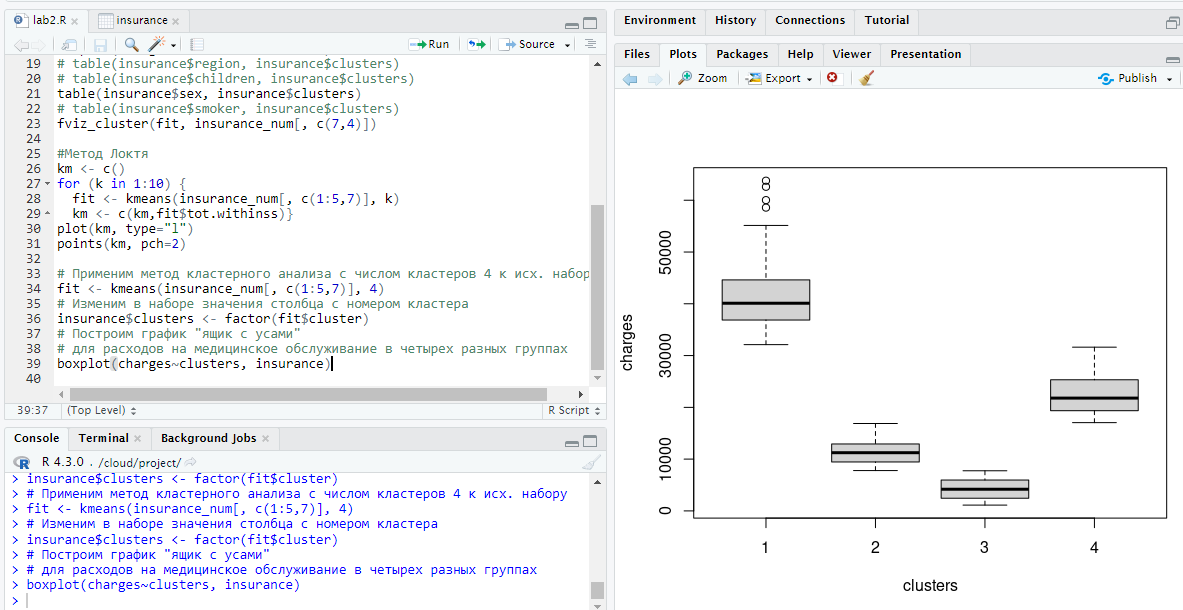


Рис. 15. Распределение застрахованных по кластерам в зависимости от расходов на медицинское обслуживание (4 кластера)

1. Оценим, как распределились застрахованные по 4 кластерам по переменным smoker, sex, children, region.

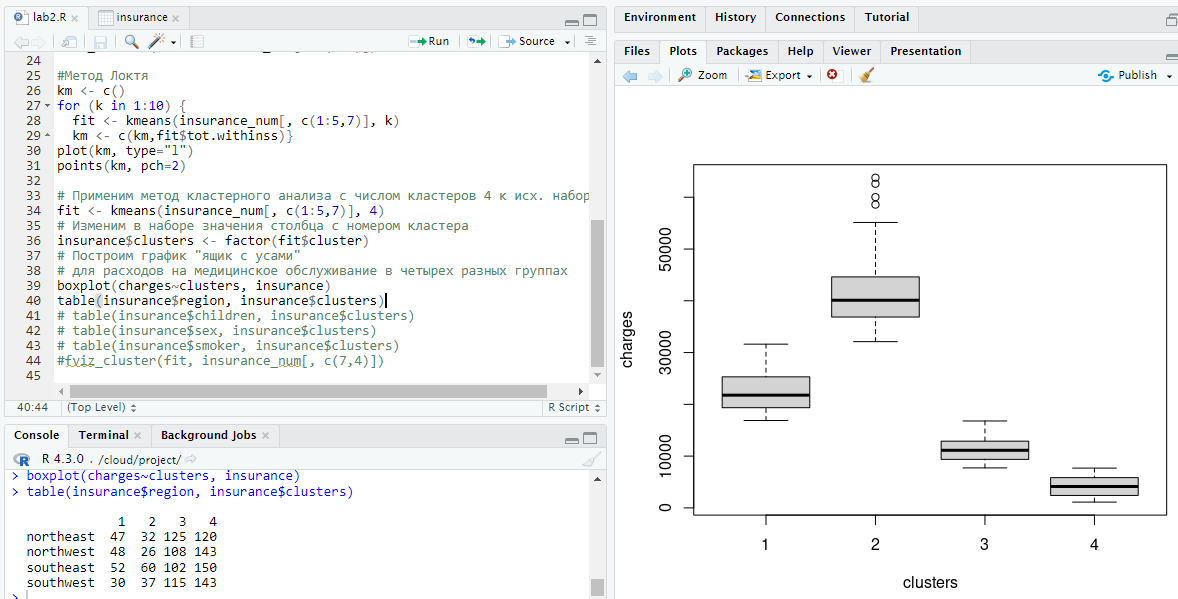


Рис. 16. По region

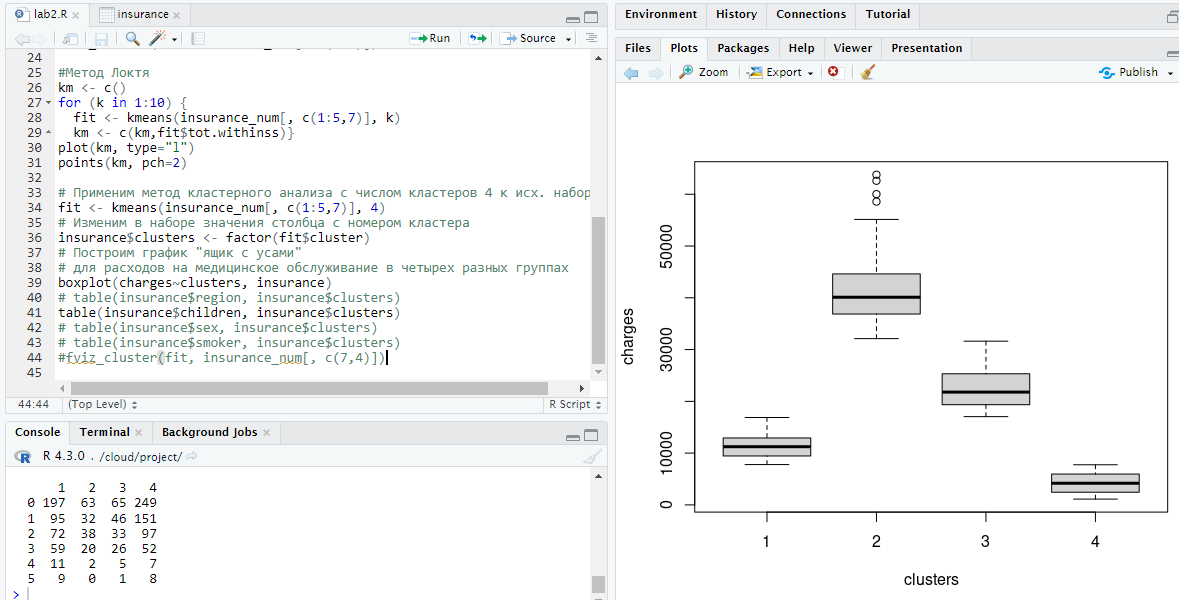


Рис.17. По Children

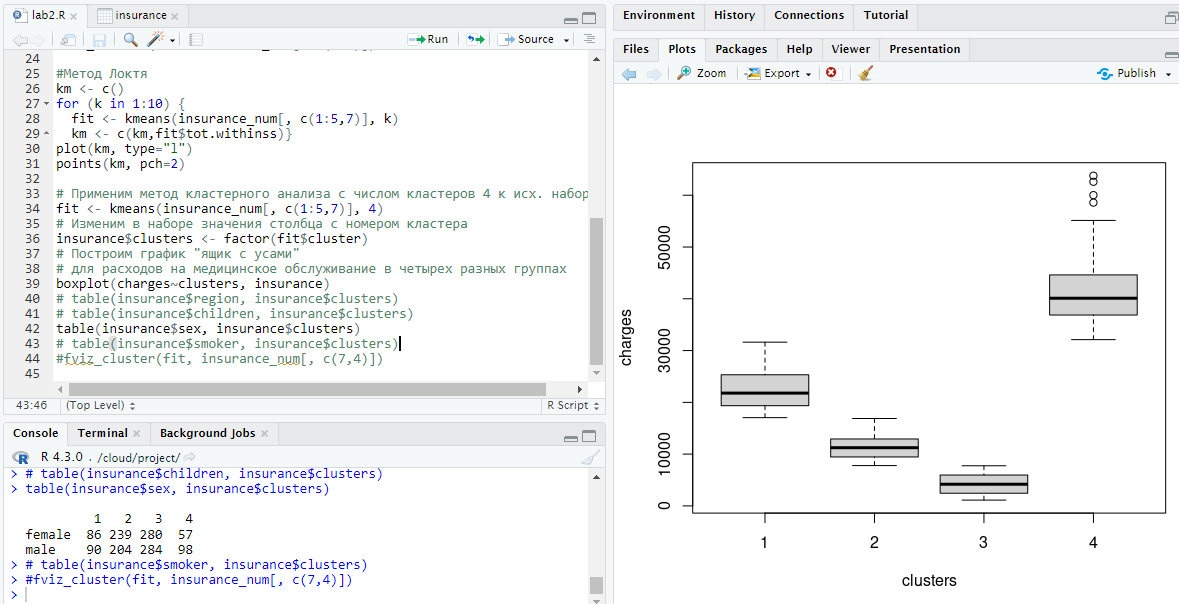


Рис.18. По sex

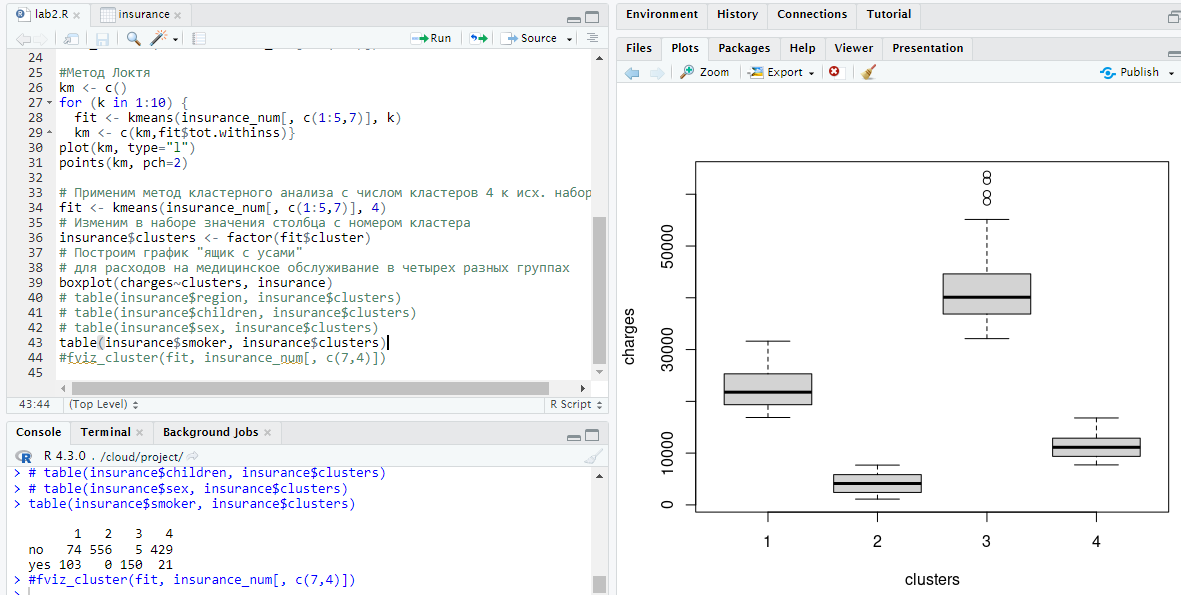


Рис.19. По smoker

1. Построение отображений используя Fviz\_cluster

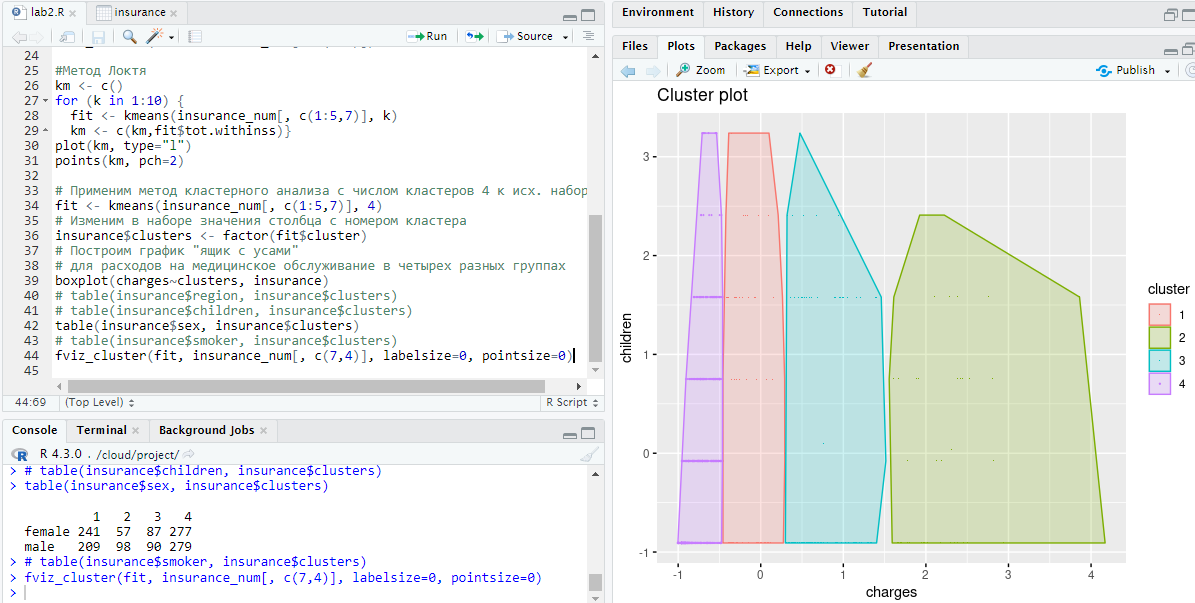


Рис. 20. Распределение по sex столбцов charges и children

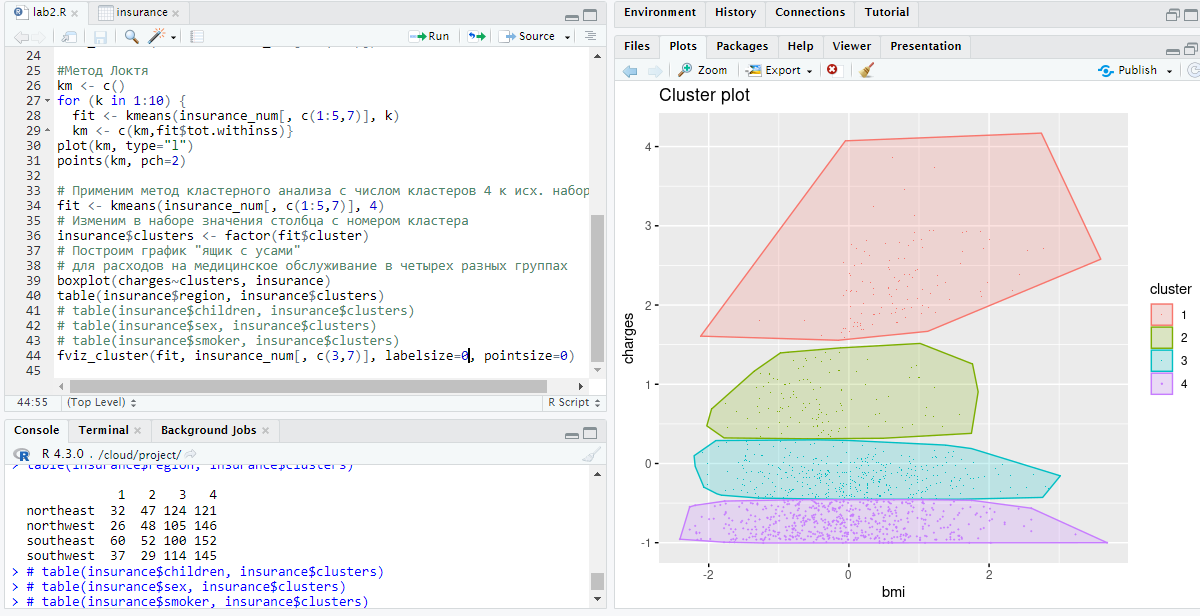


Рис.21. Распределение по region столбцов bmi и charges

# **Лабораторная работа №1.3**

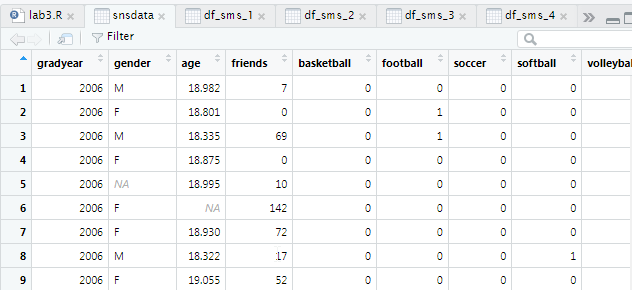
**Набор данных медицинское страхование**

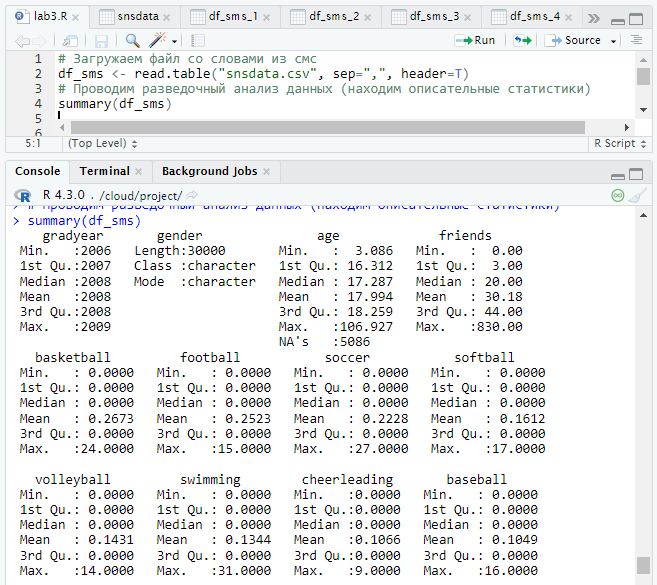
**Цель**

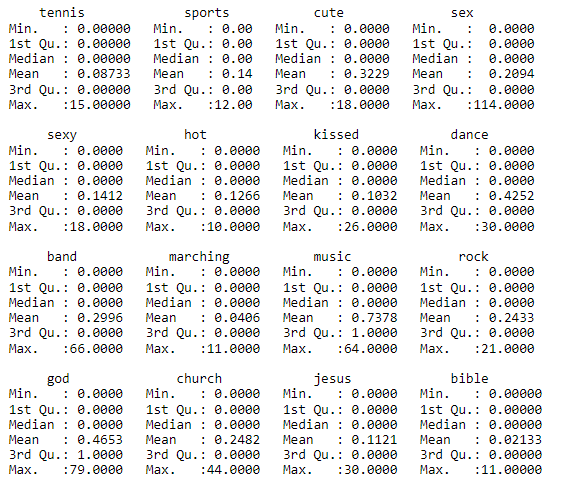
Рассмотрим задачу кластерного анализа методом k-средних для sms. Это данные, полученные на основе методов ОЕЯ – обработки естественного языка. Подготовленные данные возьмем из файла snsdata.csv.

**Ход выполнения работы**

1. Загрузим набор данных:

Рис.22. DataSet snsdata.scv





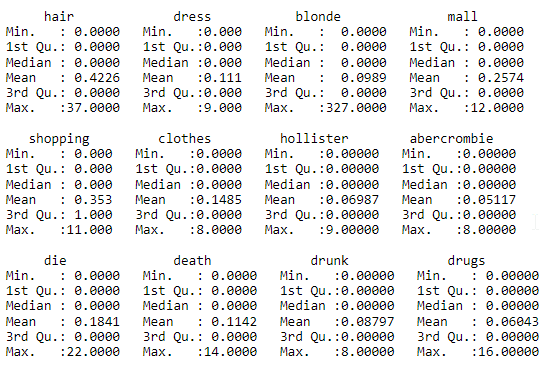


Рис.23. Описательные статистики для набора данных df\_sms

Для переменной возраст (age) обнаружились явно аномальные значения: минимальное значение 3 года, максимальное значение 106 лет и у 5086 наблюдений значение соответствующей переменной NA.

1. Проведём кластеризацию с использованием метода средних.

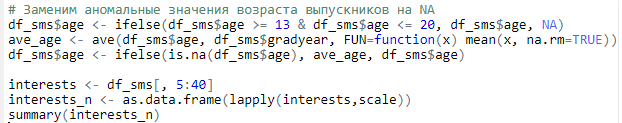


Рис.24. Код описательной статистики для стандартизированного набора данных interests\_n

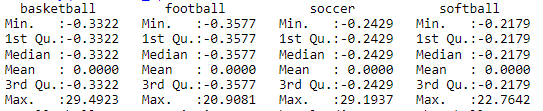


Рис.25. Описательные статистики для стандартизованного набора данных interests\_n

1. Применим метод кластерного анализа с числом кластеров 5:



Рис.26. Метод кластерного анализа

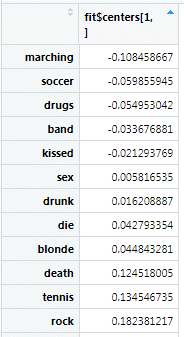


Рис.27. Частотность слов в первом кластере, упорядоченная по убыванию частоты

1. Выполним операции упорядочивания данных для центроида первого кластера:

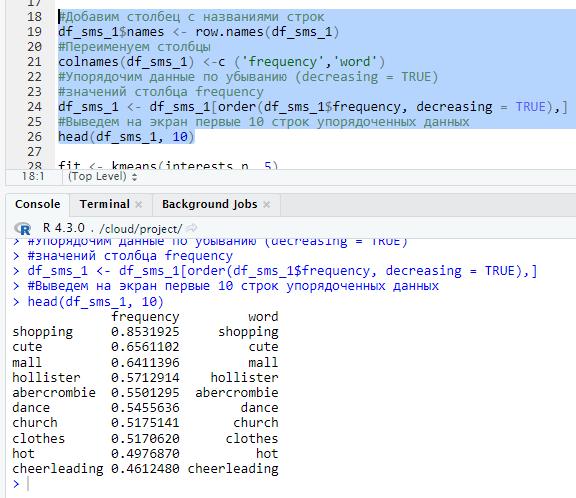


Рис.28. Частотность слов в первом кластере, упорядоченная по убыванию частоты(первые 10 строк)

1. Проанализировать 2-5 кластеров, сделать выводы о возможности выявить значимые интересы у представителей 1-5 кластеров.
2. Для 2 кластера:

На 2 кластере видно, что самую меньшую частоту по модулю имеет blonde

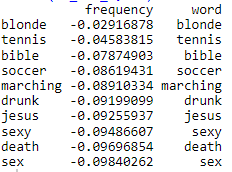


Рис.29. 2 Кластер первые 10 строк

А самое большее по модулю Shopping:

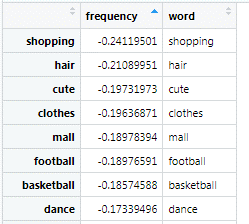


Рис.30. 2 Кластер наибольшее по модулю

На 3 кластере видно, что самую большую частоту по модулю имеет Shopping

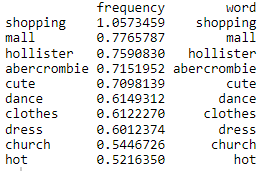


Рис.31. 3 Кластер первые 10 строк

А самое меньшее по модулю:

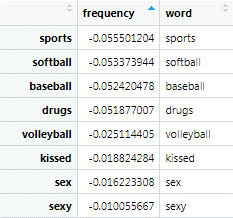


Рис.32. 3 кластер меньшее

На 4 кластере видно, что самую большую частоту имеет cheerleading

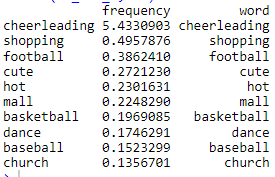


Рис.33. 4 Кластер первые 10 строк

А самую меньшую по модулю marching:

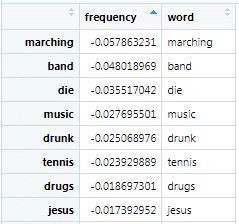


Рис.34. 4 кластер меньшее по модулю

На 5 кластере видно, что самую большую частоту имеет kissed

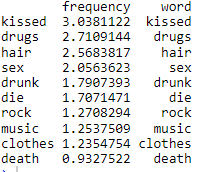


Рис.35. 5 Кластер первые 10 строк

А самую меньшую по модулю bible:

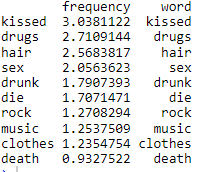


Рис.36. 5 кластер меньшее

1. Проанализируем полученные кластеры по таким переменным, как возраст, количество друзей и пол.

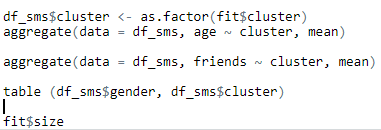


Рис.37. Код

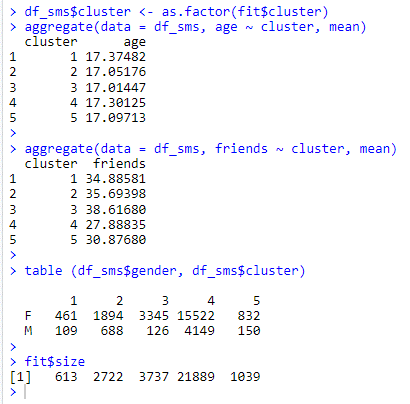


Рис.38. Вывод в консоль

Исходя из выведенных данных можно сделать выводы:

1. Во всех кластерах средний возраст равен 17 годам.
2. Количество друзей варьируется от 27 до 38 человек у людей.
3. Соотношение мужчин и женщин различается в разы, на каждого мужчину приходится примерно от 4 до 7 девушек.
4. Есть неопределённые пола из представленных данных.

# **Вывод**

В ходе выполнения всех лабораторных работ мною были получены навыки технологии работы с кластерами, графиками, анализа данных. Была проведена работа с разными данными, построены графики, описывающие различные данные, также построены графики, описывающие максимально наглядно максимальное отображение.

- выбирали количество кластеров

- построили график методом локтя

- проанализирован набор данных медицинское страхование

- научились делать выводы по представленным данным в кластерах

- строили частоты от большей к меньшей по модулю и определяли самое популярное и менее популярное в определённом кластере