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 «**Интеллектуальные информационные системы**»**

**к лабораторной работе № 1**

**Темы: «**Анализ метода k-средних МОДЕЛЕЙ в программе R-studio**»**,

**«**Набор данных медицинское страхование**»,**

**«**Набор данных смс**»**

Выполнил:

Студент гр. №

Верохин Алексей Константинович

2721Б1ПИ

Проверил:

Ст.преподаватель, к.т.н.

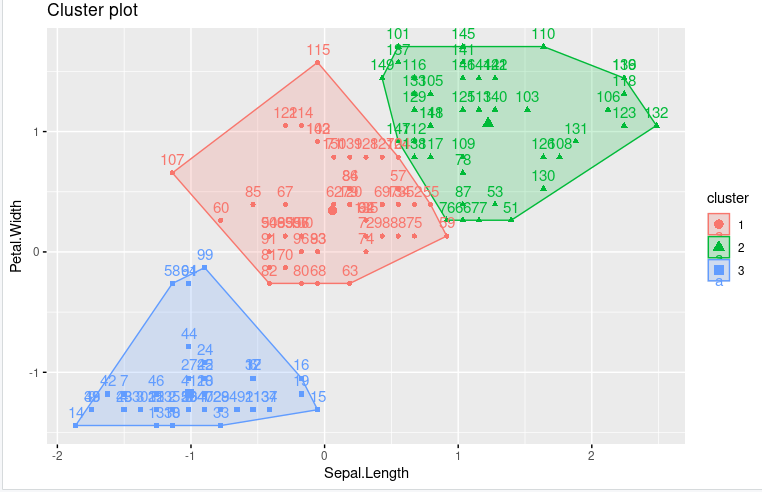
\_\_\_\_\_\_\_\_\_\_\_/Нажимов А.В./ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

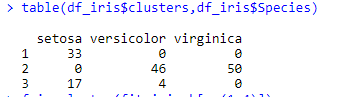
Дзержинск 2024

Лабораторная работа №1.1.

1. Отображение кластеров через команду fviz\_cluster()

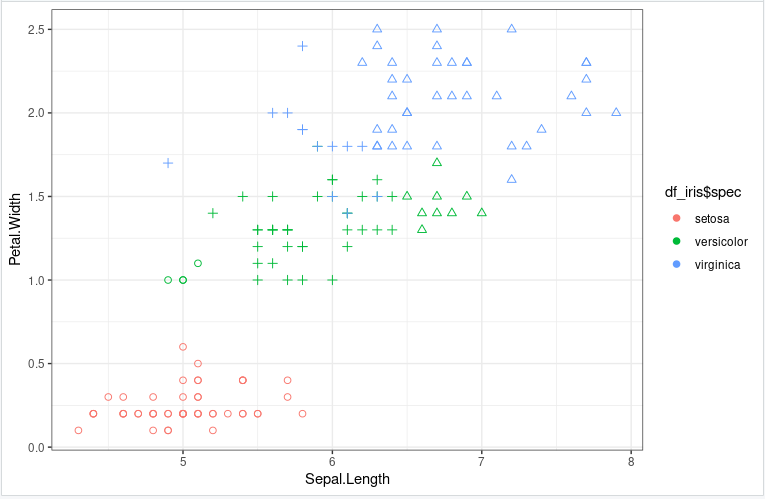
fviz\_cluster(fit,df\_iris[,c(1,4)])





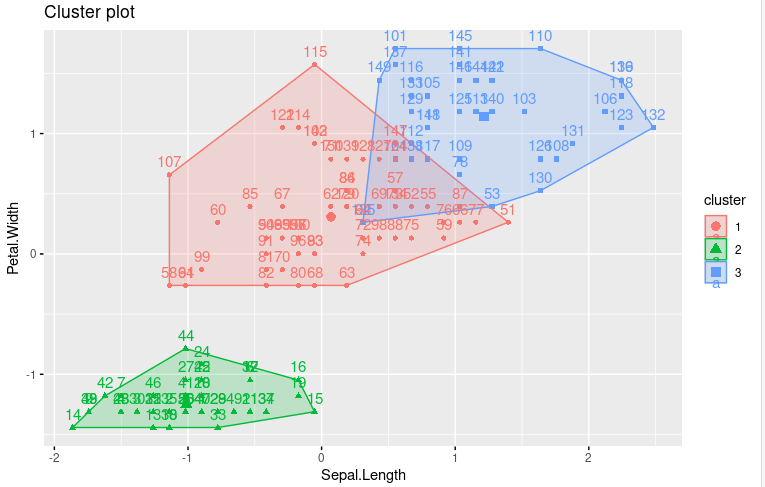
1. Получаем 3 кластера, сформированные функцией kmeans из набора данных df\_iris по 4ём столбцам.
2. Визуализация данных с помощью функции ggplot

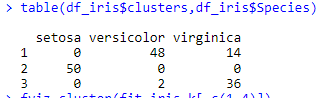
ggplot(df\_iris[,c(1,4)],aes(Sepal.Length,Petal.Width,col=df\_iris$spec))+geom\_point(pch=df\_iris$clusters,size=2)+theme\_bw()



1. Отображение обновленных кластеров с помощью функции fviz\_cluster()

fviz\_cluster(fit,iris\_k)



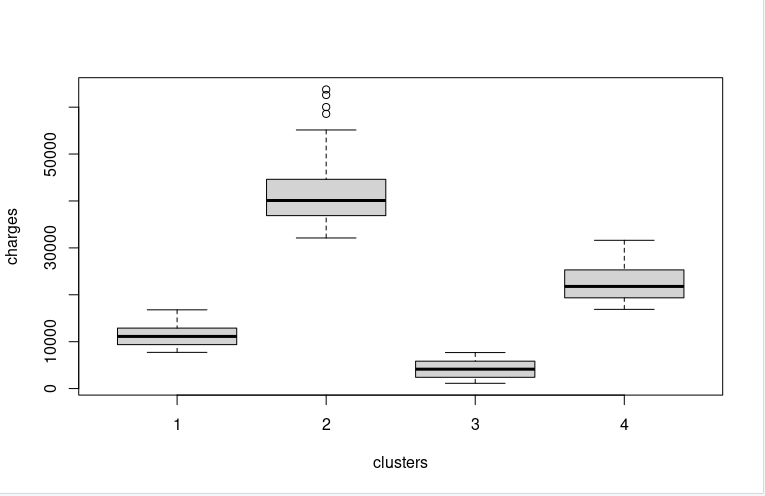


1. Описание результатов

Можно заметить, что некоторые данные из 2го кластера перешли в 3ий из-за улучшения качества кластеризации.

Лабораторная работа №1.2.

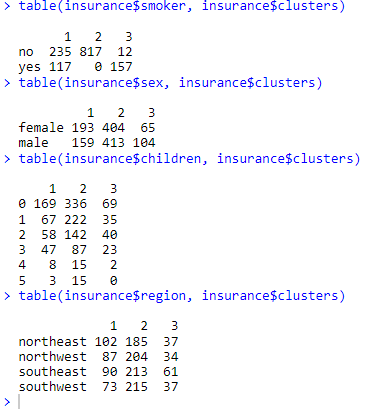
1. График распределения «ящик с усами» по медицинским расходам



1. Оценим распределение застрахованных курильщиков по трём кластерам с помощью команды table().

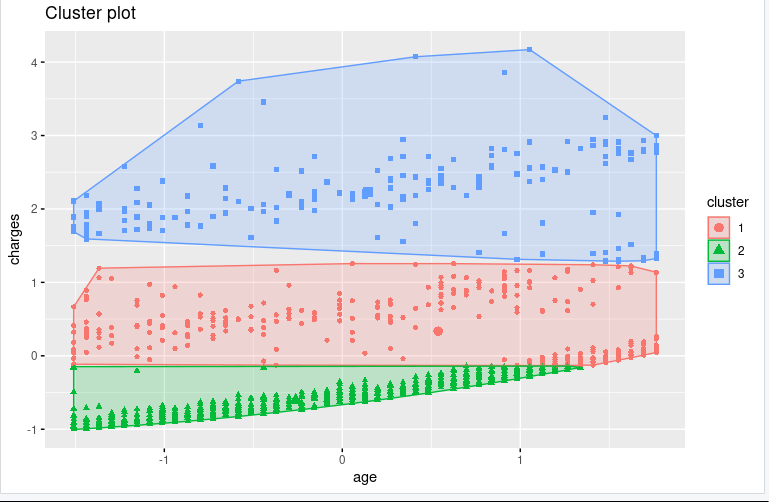


1. Также оценим распределение пола, наличия детей и регион.

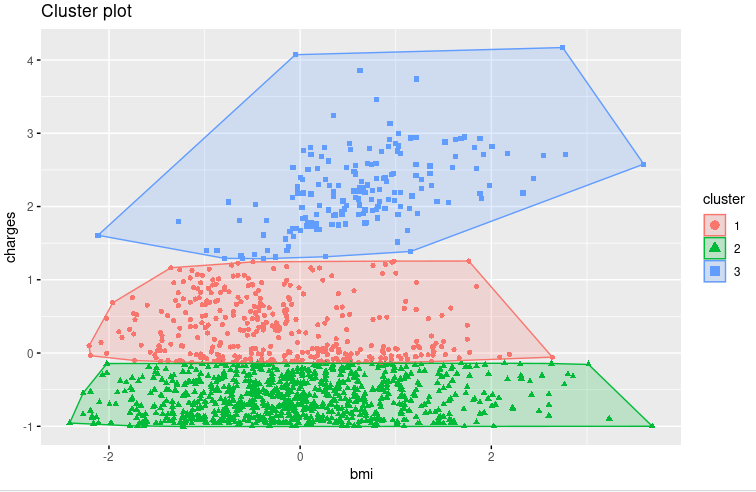


1. Построим график распределения курильщиков в мед. расходам с помощью функции fviz\_cluster()

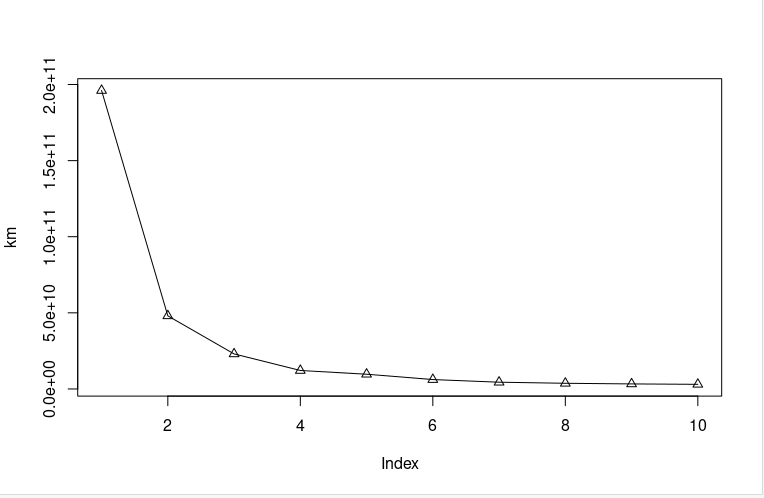
fviz\_cluster(fit, insurance\_num[,c(1,7)],geom="point")



Далее график индекса массы тела к мед страховке

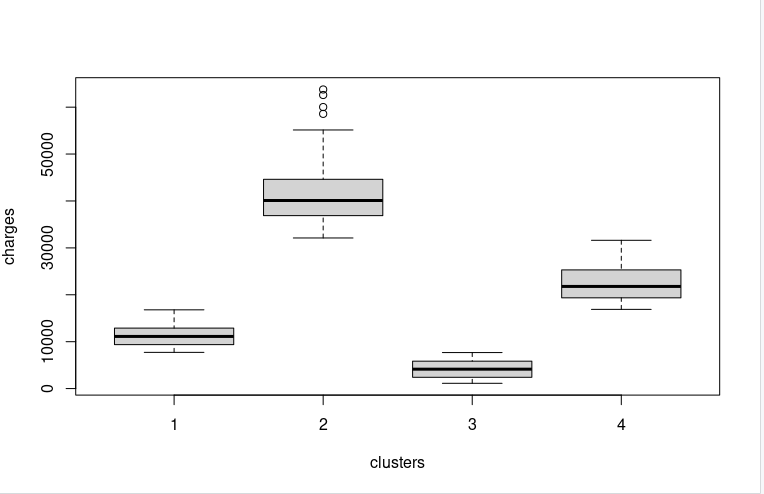


1. Построим график актуальности выбора количества кластеров

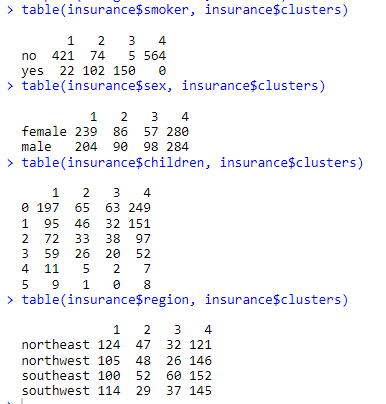


По графику видно, что можно разбить данные на 4 кластера.

1. График («Ящик с усами») распределения данных по медицинским расходам, используя 4 кластера

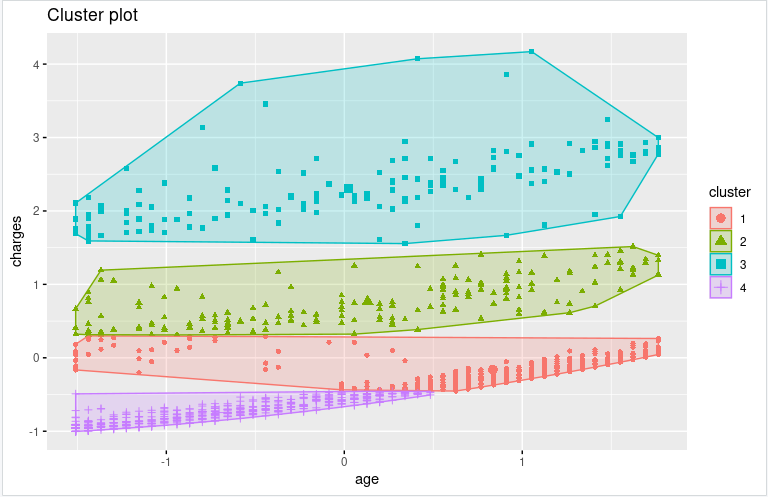


1. Оценим распределение застрахованных по четырем кластерам по переменным smoker, sex, children, region

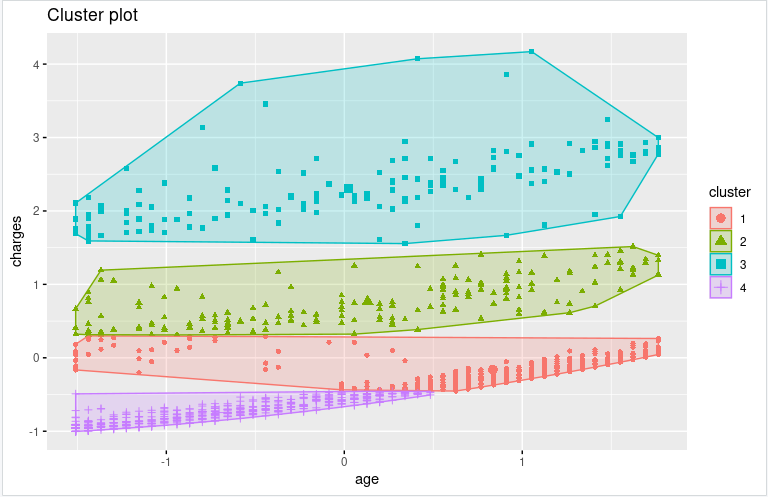
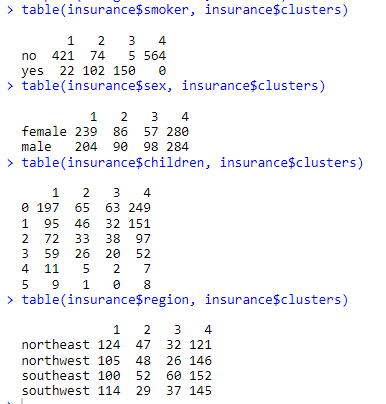


1. Распределение кластеров по возрасту и мед. Расходы

fviz\_cluster(fit,insurance\_num[,c(1,7)],geom = "point")



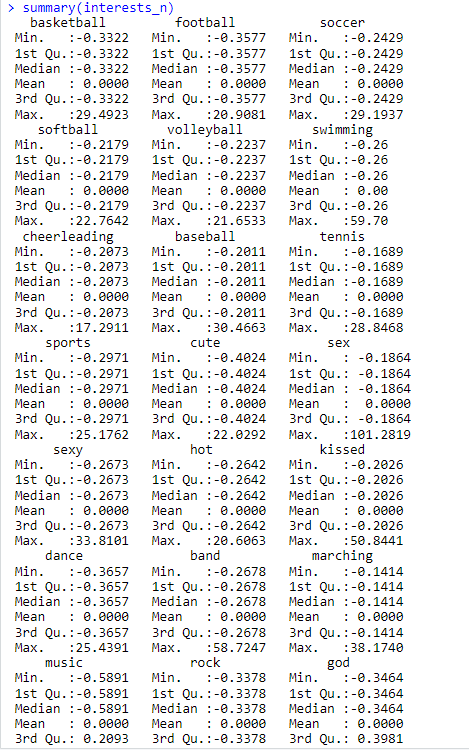
Можно заметить, что распределение стало более точным.

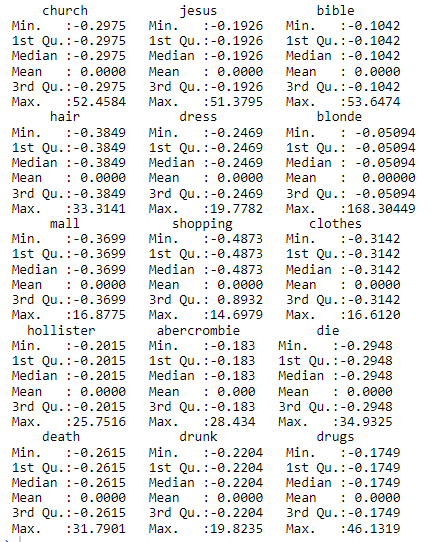


Если взять таблицу курильщиков и график возрастов, то можно заметить, что чем больше люди курят, тем больше они тратят на мед. страховку.

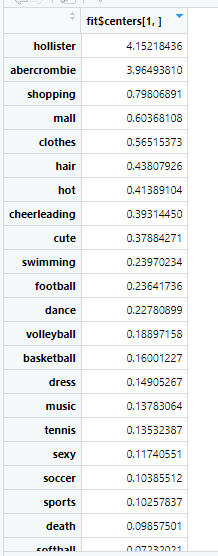
Лабораторная работа №1.3.

1. Описательная статистика для стандартизированного набора данных interests\_n



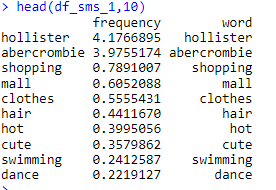


1. Набор данных df\_sms\_1, отсортированный по убыванию по столбцу fit$centers

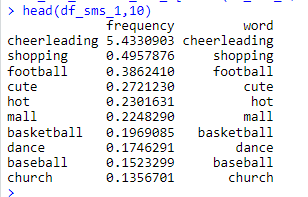


1. Вывод десяти строк упорядоченных данных пяти кластеров данных

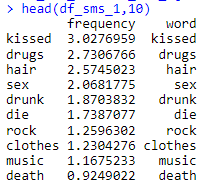
1-ый кластер



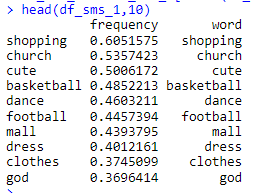
2-ой кластер



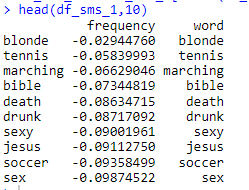
3-ий кластер



4-ый кластер



5-ый кластер



Из представленных данных можно сделать вывод, что самыми популярными словами являются: Hollister, cheerleading, kissed, shopping, blonde.