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. №

Ишеков Н.А

2721Б1ПИ

Проверил:

Ст.преподаватель, к.т.н.

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

Дзержинск 2024

Оглавление

[**Цель работы:** 3](#_Toc135328904)

[**ЛАБОРАТОРНАЯ РАБОТА №1.1.** 4](#_Toc135328905)

[**Код:** 4](#_Toc135328906)

[**График№1:** 5](#_Toc135328907)

[**График№2:** 6](#_Toc135328908)

[**ЛАБОРАТОРНАЯ РАБОТА №1.2.** 7](#_Toc135328909)

[**Код:** 7](#_Toc135328910)

[**ЛАБОРАТОРНАЯ РАБОТА №1.3.** 10](#_Toc135328911)

[**Код:** 10](#_Toc135328912)

[**Выводы:** 12](#_Toc135328913)

[**Список литературы:** 14](#_Toc135328914)

# **Цель работы:**

Цель лабораторной работы: провести анализ на примере метода k-средних МОДЕЛЕЙ, набор данных медицинское страхование, набор данных смс, получение практических навыков работы с R в программе R-studio.

# **ЛАБОРАТОРНАЯ РАБОТА №1.1.**

**РЕАЛИЗАЦИЯ МЕТОДА K-СРЕДНИХ С ПОМОЩЬЮ НАБОРА ДАННЫХ ИРИСОВ ФИШЕРА ПОСТРОЕНИЕ ГРАФИКА**

## **Код:**

df\_iris <- iris

library(ggplot2)

library(factoextra)

fit <- kmeans(df\_iris[, c(1,4)],3)

df\_iris$clusters <-factor(fit$cluster)

df\_iris$spec <- factor(iris$Species)

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

table(df\_iris$clusters, df\_iris$Species)

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

iris\_k <-iris[, 1:4]

fit <- kmeans(iris\_k, 3)

df\_iris$clusters <- factor(fit$cluster)

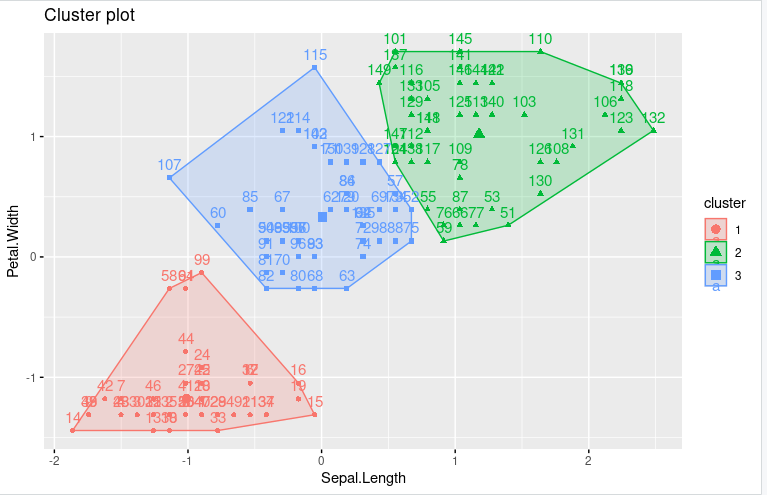
ggplot(df\_iris[, c(1,4)], aes(Sepal.Lenght, Petal.Width, col = df\_iris$spec))+theme\_bw()

table(df\_iris$clusters, df\_iris$Species)

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

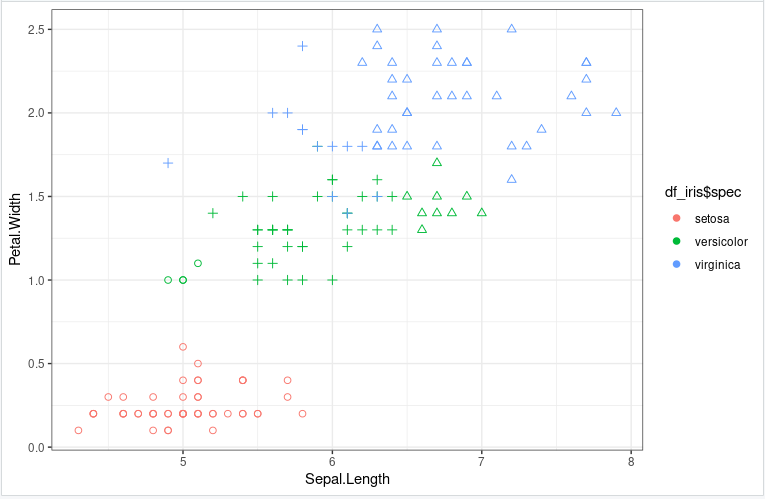
1. Отображение кластеров через команду fviz\_cluster()

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



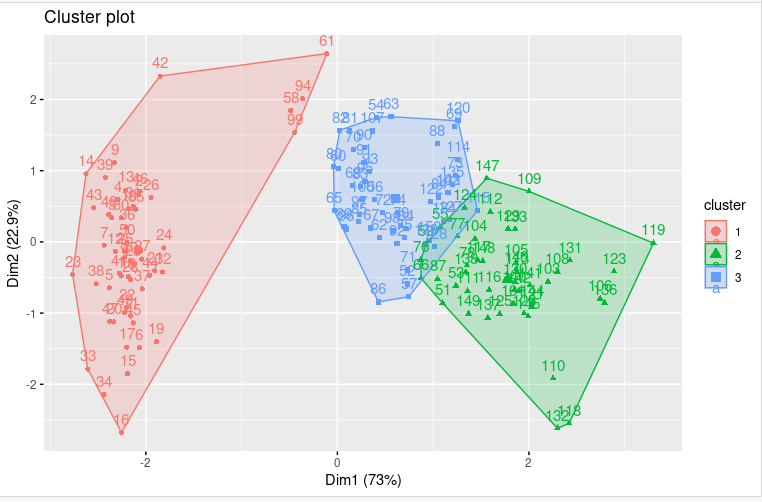
1. Получаем 3 кластера, сформированные функцией kmeans из набора данных df\_iris по 4ём столбцам.
2. Визуализация данных с помощью функции ggplot

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)



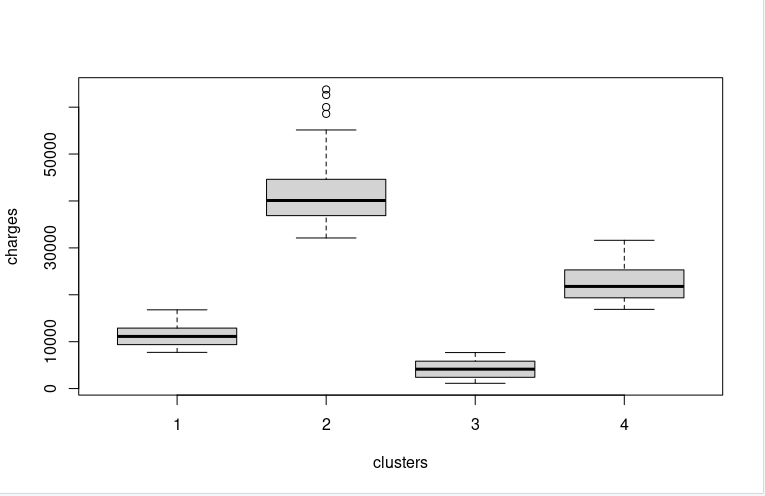
1. Описание результатов

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

# **ЛАБОРАТОРНАЯ РАБОТА №1.2.**

**НАБОР ДАННЫХ МЕДИЦИНСКОЕ СТРАХОВАНИЕ**

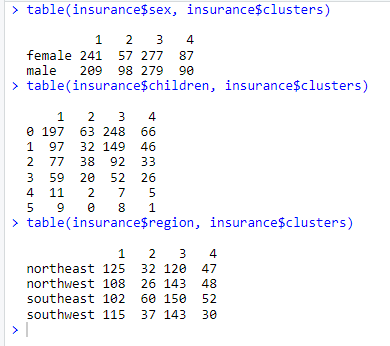
1. График распределения «ящик с усами» по медицинским расходам



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

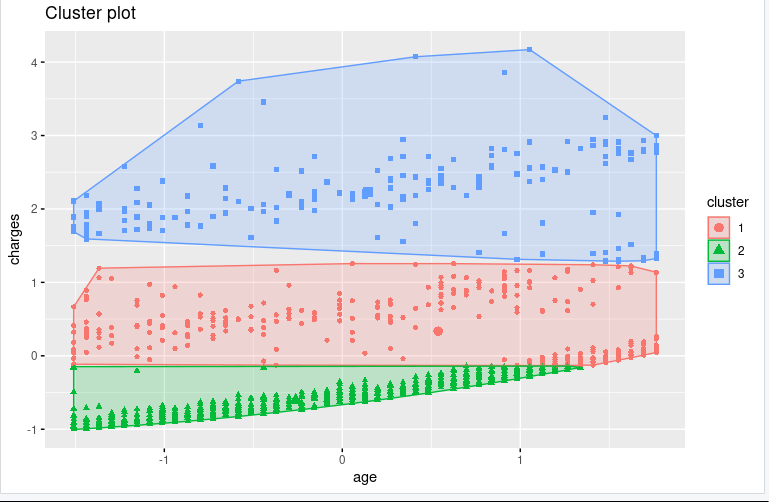


1. Также оценим распределение пола, наличия детей и регион.

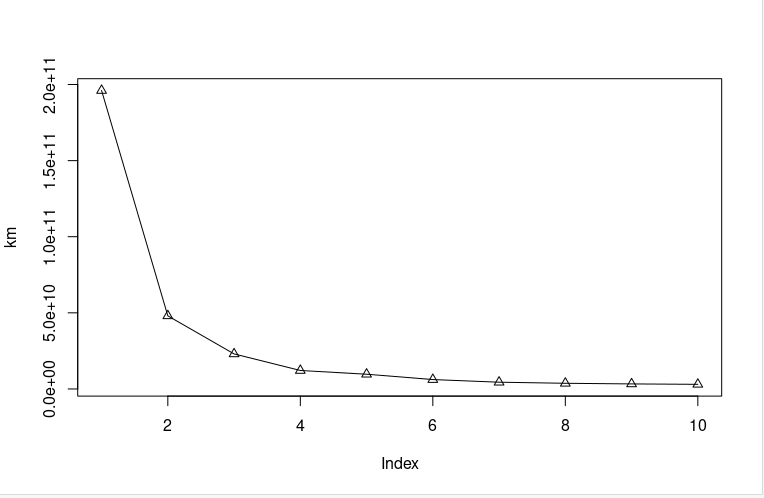


1. Построим график распределения курильщиков в мед. расходам с помощью функции fviz\_cluster()

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

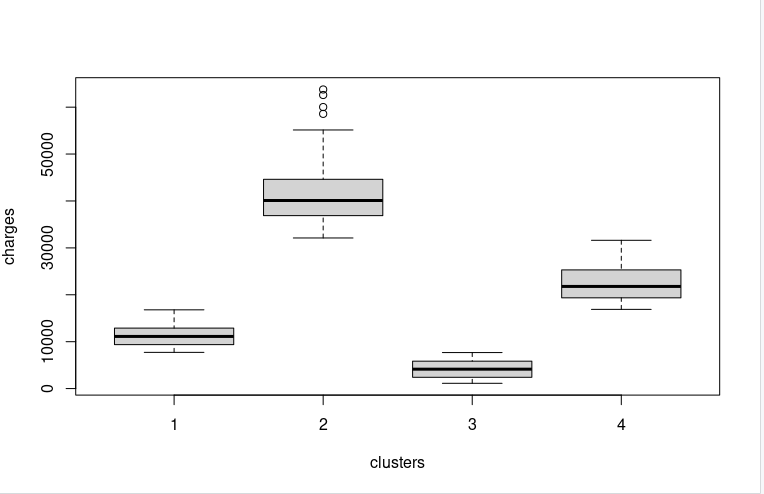


1. Построим график актуальности выбора количества кластеров

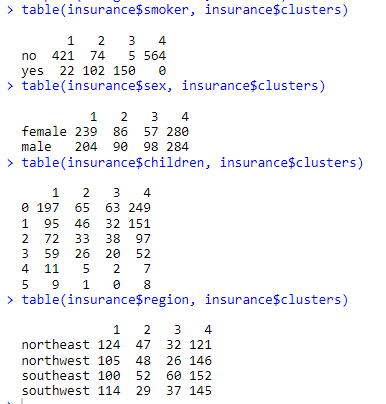


По графику видно, что можно разбить данные на 4 кластера.

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

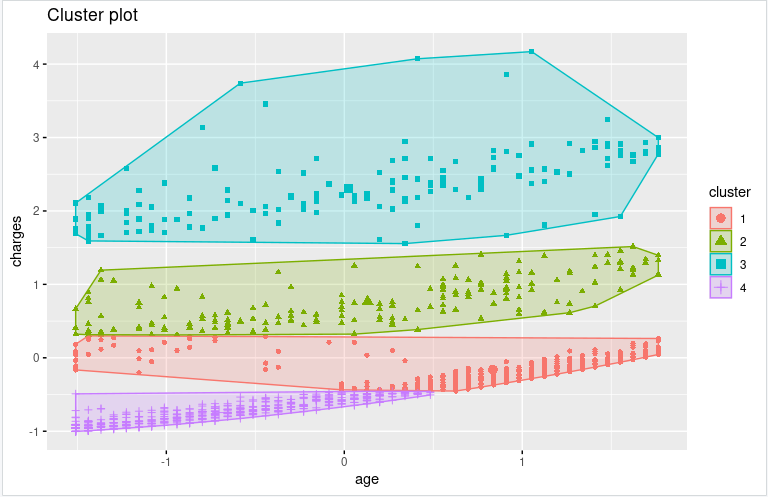


1. Оценим распределение застрахованных по четырем кластерам по переменным smoker, sex, children, region



1. Распределение кластеров по возрасту и мед. Расходы

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

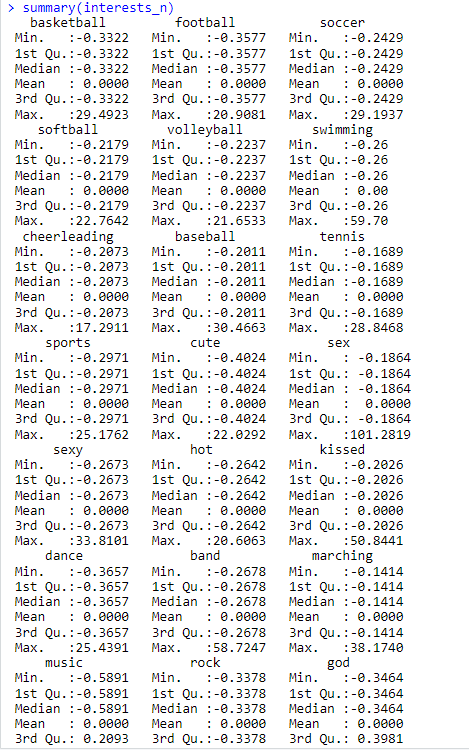


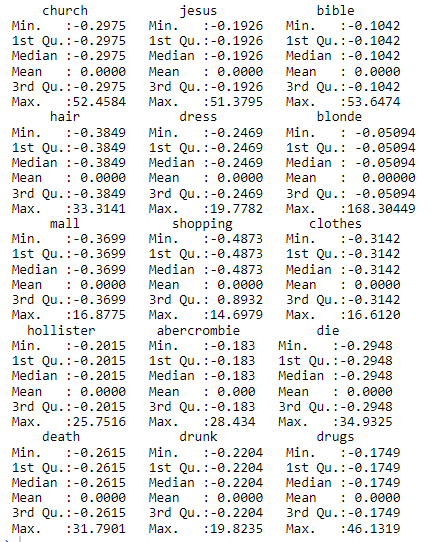
Можно заметить что распределение стало более точным.

# **ЛАБОРАТОРНАЯ РАБОТА №1.3.**

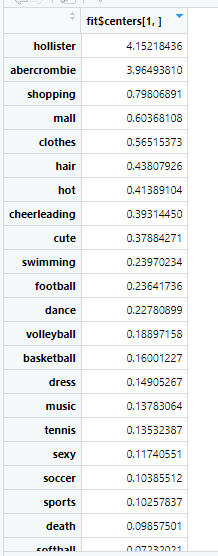
НАБОР ДАННЫХ СМС

1. Описательная статистика для стандартизированного набора данных interests\_n



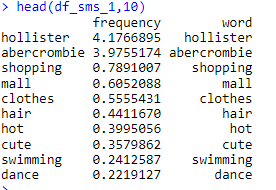


1. Набор данных df\_sms\_1, отсортированный по убыванию по столбцу fit$centers

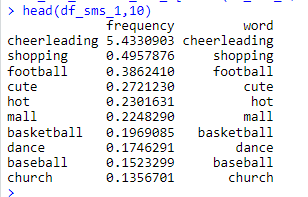


1. Вывод десяти строк упорядоченных данных пяти кластеров данных

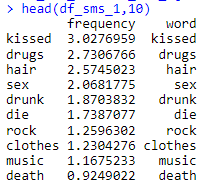
1-ый кластер



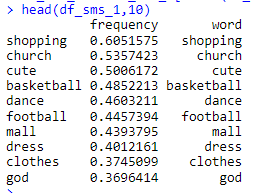
2-ой кластер



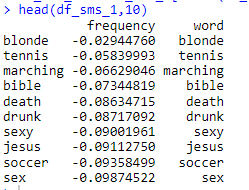
3-ий кластер



4-ый кластер



5-ый кластер



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