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)

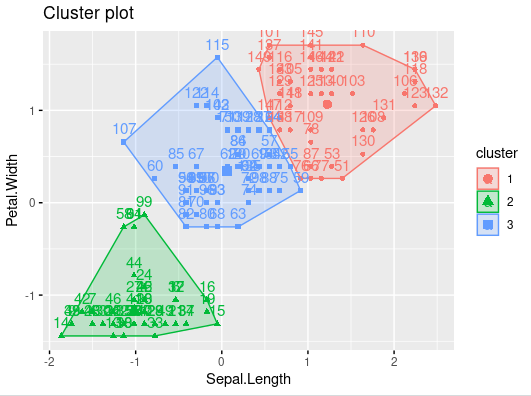
ggplot(df\_iris[, c(1,4)], aes(Sepal.Lenght, Petal.Width, col = df\_iris$spec))+theme\_bw()

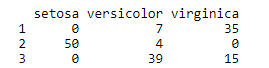
table(df\_iris$clusters, df\_iris$Species)

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

### **График№1:**

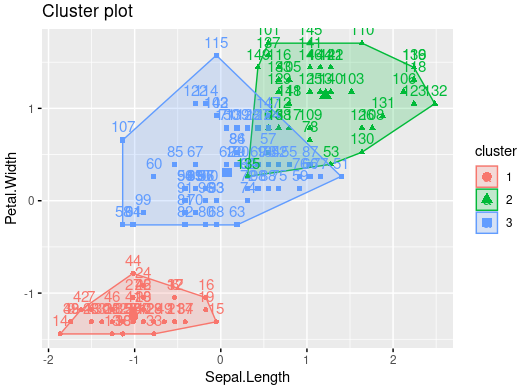
(Первый ввод кластеров)

****

****

### **График№2:**

(Второй ввод кластеров)





# **ЛАБОРАТОРНАЯ РАБОТА №1.2.**

**НАБОР ДАННЫХ МЕДИЦИНСКОЕ СТРАХОВАНИЕ**

## **Код:**

library(ggplot2)

library(factoextra)

insurance <- read.table("insurance.csv", header=T, sep=",")

insurance\_num <- insurance

insurance\_num$sex <- ifelse(insurance$sex=="male",1,0)

insurance\_num$smoker <- ifelse(insurance$smoker=="yes",1,0)

insurance\_num$region <- factor(insurance$region)

fit <- kmeans(insurance\_num[, c(1:5,7)], 3)

insurance$clusters <- factor(fit$cluster)

boxplot(charges~clusters, insurance)

table(insurance$smoker, insurance$clusters)

table(insurance$sex, insurance$clusters)

table(insurance$children, insurance$clusters)

table(insurance$region, insurance$clusters)

fviz\_cluster(fit, data=insurance\_num[, c(1:5,7)])

km <- c()

for (k in 1:10){

fit <- kmeans(insurance\_num[, c(1:5,7)], k)

km <- c(km,fit$tot.withinss)}

plot(km, type="l")

points(km,pch=2)

fit <- kmeans(insurance\_num[, c(1:5,7)],4)

insurance$clusters <- factor(fit$cluster)

boxplot(charges~clusters, insurance)

table(insurance$smoker, insurance$clusters)

table(insurance$sex, insurance$clusters)

table(insurance$children, insurance$clusters)

table(insurance$region, insurance$clusters)

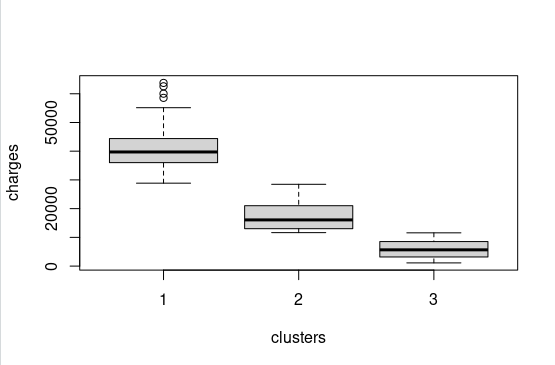
ggplot(insurance\_num[, c(1,7)], aes(age, charges))+

geom\_point(aes(shape=factor(insurance$clusters)), size=2)+

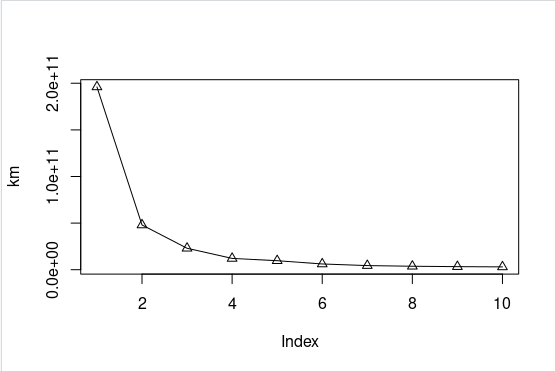
theme\_bw()

fviz\_cluster(fit, insurance\_num[, c(1,7)])

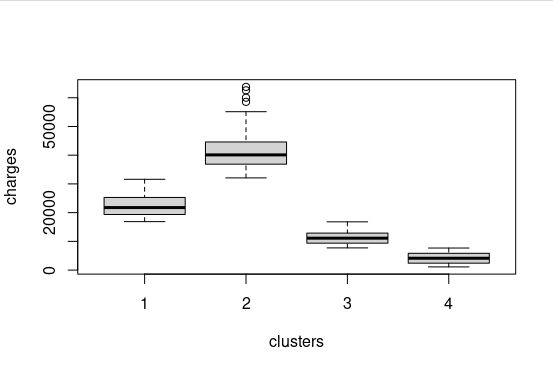
**График№1:**

****

**График№2:**

****

**График№3:**

****

# **ЛАБОРАТОРНАЯ РАБОТА №1.3.**

НАБОР ДАННЫХ СМС

## **Код:**

df\_sms <- read.table("snsdata.csv", sep=",", header=T)

summary(df\_sms)

df\_sms$age <-ifelse(df\_sms$age >= 13 & df\_sms$age <=20, df\_sms$age, NA)

ave\_age <-ave(df\_sms$age, df\_sms$gradyear, FUN=function(x) mean(x, na.rm=TRUE))

df\_sms$age <- ifelse(is.na(df\_sms$age), ave\_age, df\_sms$age)

interests <- df\_sms[,5:40]

interests\_n <- as.data.frame(lapply(interests, scale))

summary(interests\_n)

fit <- kmeans(interests\_n, 5)

df\_sms\_1 <- as.data.frame(fit$centers[1,])

df\_sms\_1$names <- row.names(df\_sms\_1)

colnames(df\_sms\_1) <-c ('frequency', 'word')

df\_sms\_1 <- df\_sms\_1[order(df\_sms\_1$frequency, decreasing=TRUE),]

head(df\_sms\_1, 10)

df\_sms\_2 <- as.data.frame(fit$centers[2,])

df\_sms\_2$names <- row.names(df\_sms\_2)

colnames(df\_sms\_2) <-c ('frequency', 'word')

df\_sms\_2 <- df\_sms\_2[order(df\_sms\_2$frequency, decreasing=TRUE),]

head(df\_sms\_2, 10)

df\_sms\_3 <- as.data.frame(fit$centers[3,])

df\_sms\_3$names <- row.names(df\_sms\_3)

colnames(df\_sms\_3) <-c ('frequency', 'word')

df\_sms\_3 <- df\_sms\_3[order(df\_sms\_3$frequency, decreasing=TRUE),]

head(df\_sms\_3, 10)

df\_sms\_4 <- as.data.frame(fit$centers[4,])

df\_sms\_4$names <- row.names(df\_sms\_4)

colnames(df\_sms\_4) <-c ('frequency', 'word')

df\_sms\_4 <- df\_sms\_4[order(df\_sms\_4$frequency, decreasing=TRUE),]

head(df\_sms\_4, 10)

df\_sms\_5 <- as.data.frame(fit$centers[5,])

df\_sms\_5$names <- row.names(df\_sms\_5)

colnames(df\_sms\_5) <-c ('frequency', 'word')

df\_sms\_5 <- df\_sms\_5[order(df\_sms\_5$frequency, decreasing=TRUE),]

head(df\_sms\_5, 10)

df\_sms$cluster <- as.factor(fit$cluster)

aggregate(data = df\_sms, age ~ cluster, mean)

aggregate(data = df\_sms, friends ~ cluster, mean)

table(df\_sms$gender, df\_sms$cluster)

# **Выводы:**

**Лабораторная часть 1.1:**

По итогу лабораторной работы я выяснил, что используя два столбца, на график №1 области не пересекаются, а используя четыре столбца график №2 области пересекаются.

**Лабораторная часть 1.2:**

 Разбиение на кластеры: Применяется алгоритм кластеризации K-means для разделения застрахованных на кластеры на основе их характеристик. В данном случае было выбрано 3 кластера.

Распределение кластеров по переменным: были проведены анализы распределения кластеров по переменным "smoker" (курящий или нет), "sex" (пол), "children" (количество детей) и "region" (регион). Таблицы, полученные с помощью функции `table()`, показывают количество застрахованных в каждой категории переменной для каждого кластера.

Визуализация кластеров: для визуализации кластеров были построены диаграммы размаха (boxplot) и графики рассеяния (scatter plot). Диаграммы размаха показывают распределение затрат на страховку (charges) в каждом кластере, а графики рассеяния отображают зависимость между возрастом (age) и затратами на страховку для каждого кластера.  
  
Анализ кластеров по переменной "smoker": Для оценки распределения застрахованных по трем кластерам по переменной "smoker" можно использовать функцию `table()`. Эта функция позволяет подсчитать количество застрахованных в каждой категории "smoker" для каждого кластера. Результаты этого анализа помогут понять, как кластеры различаются по проценту курящих и некурящих застрахованных.

Максимально наглядным отображением является график под номером три так как на нём отображены четыре кластера показывали результаты курящий или нет, пол, количество детей, а также регион.

**Лабораторная часть 1.3:**

 Возраст: При анализе среднего возраста по кластерам (`aggregate(data = df\_sms, age ~ cluster, mean)`) мы можем наблюдать различия в возрасте между кластерами. Кластеры могут быть характеризованы разными возрастными группами или иметь различную среднюю возрастную характеристику.

Количество друзей: Также было произведено агрегирование данных для вычисления среднего количества друзей по кластерам (`aggregate(data = df\_sms, friends ~ cluster, mean)`). Это позволяет оценить различия в социальной активности или популярности между кластерами на основе количества друзей.  
  
Пол: Используя функцию `table(df\_sms$gender, df\_sms$cluster)`, была создана таблица сопряженности, которая позволяет оценить распределение пола в каждом кластере. Это помогает определить, есть ли различия в соотношении мужчин и женщин между кластерами.

# **Список литературы:**

1. Лекции по Интеллектуальному анализу данных к.т.н Нажимов А.В.