Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение

высшего образования

**«КУБАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»**

**(ФГБОУ ВО «КубГУ»)**

**Факультет компьютерных технологий и прикладной математики**

**Кафедра вычислительных технологий**

**Отчёт лабораторная №5.1**

**Дисциплина: Обработка больших данных**

Работу выполнил \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Логвина А.В.

Направление подготовки 02.03.02 Фундаментальная информатика и

информационные технологии

Преподаватель \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Яхонтов А.А.

Краснодар

2025

**Задачи классификации и кластеризации.**

**Цель работы:** Закрепить знания, об алгоритмах классификации и кластеризации данных,ознакомиться с некоторыми функциями языка R, осуществляющими этот вид анализа,принципами их работы. Научиться визуализировать результаты работы функций кластерного анализа и классификаторов, интерпретировать полученные результаты.

**Задание:**

1. В начале отчета необходимо разместить формулировку задания и фрагмент

исходного датасета.

2. Выполнить дескриптивный анализ данных (здесь приветствуются дополнительные исследования).

3. Оценить оптимальное число кластеров, для этого построить диаграмму "Метод силуэта", “Метод локтя”, "Статистику разрыва" и Алгоритм консенсуса.

4. Выполнить иерархическую кластеризацию вашего набора данных, построив дендрограмму. Подробно обосновать Ваш выбор числа групп.

5. Построить диаграмму со столбчатыми диаграммами и боксплотами групп Провести сравнительный анализ полученных групп.

6. Выполнить кластеризацию своего датасета по k-means

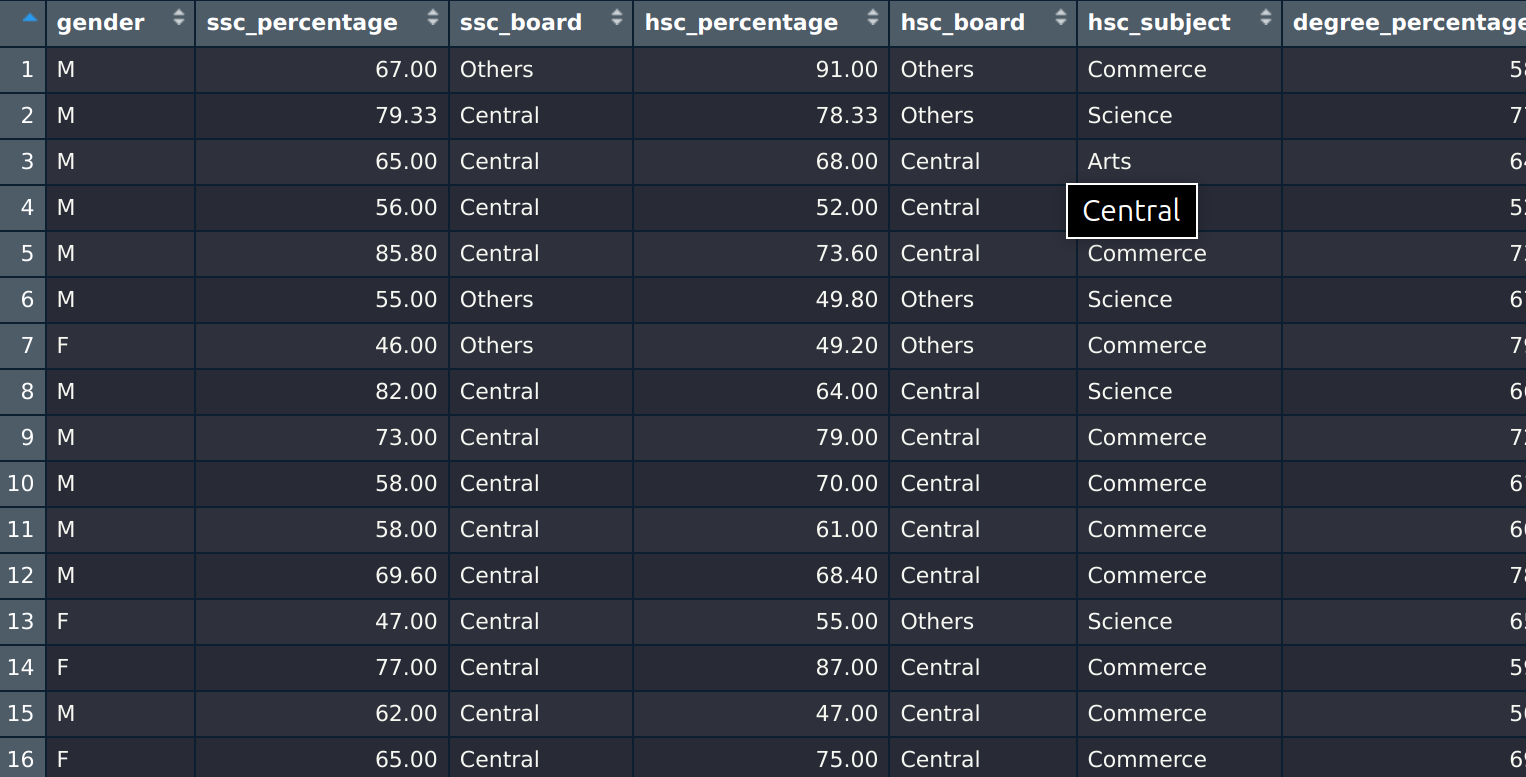
7. Выполнить построение scatterplot с помощью функций plot или pairs.

8. Построить трехмерную кластеризацию по scatterplot3d

9. В целом: выполнить шаги 1-3,5 анализа для своего набора данных (если какие-то из шагов нерелевантны вашему набору данных, объяснить почему).

**Ход работы:**

1. Скачаем датасет.

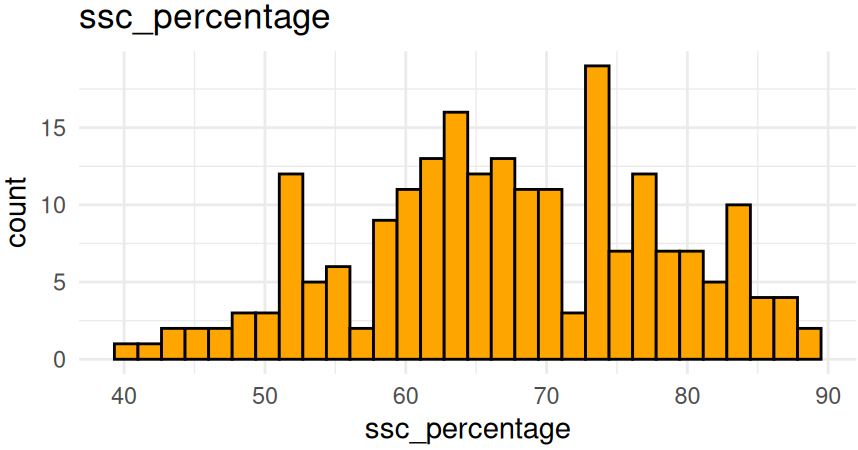
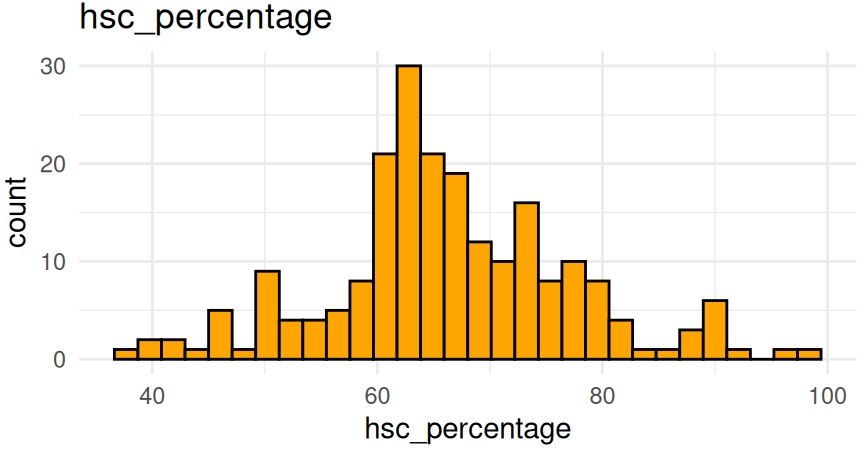
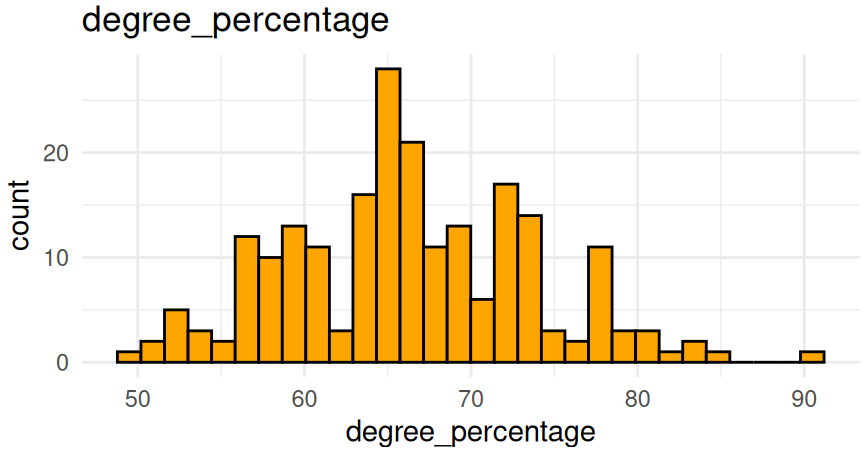
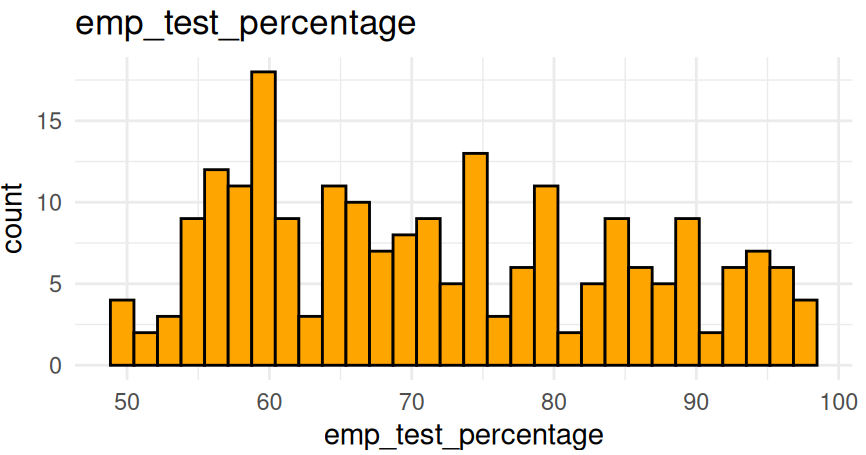
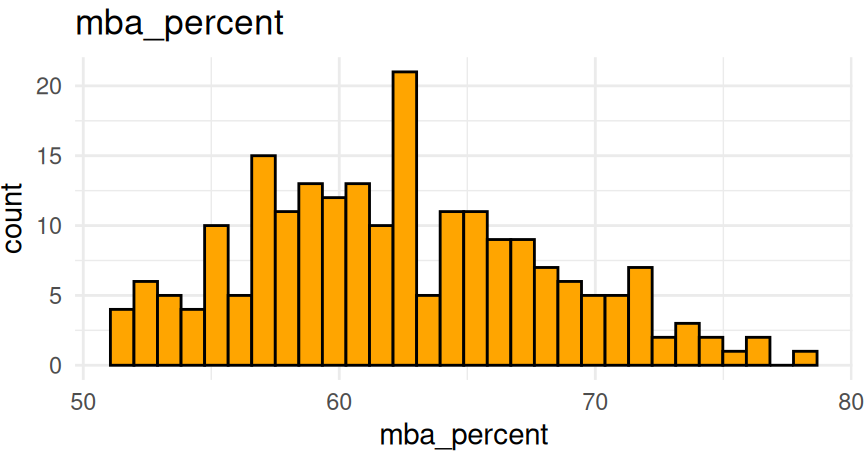
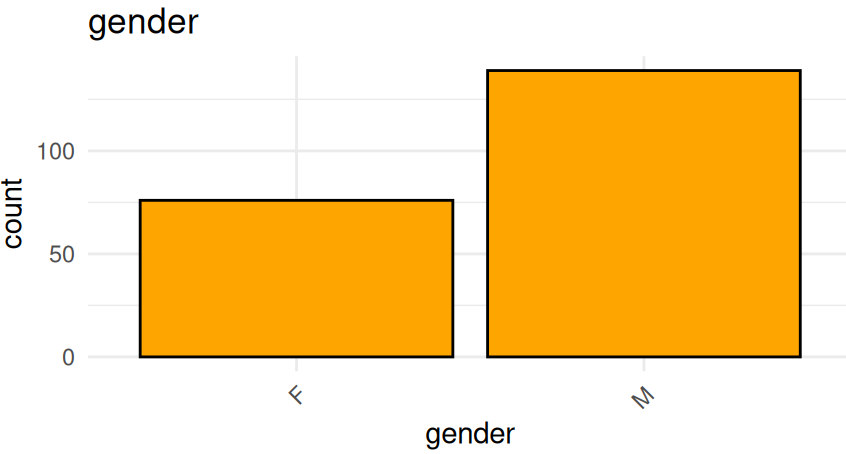
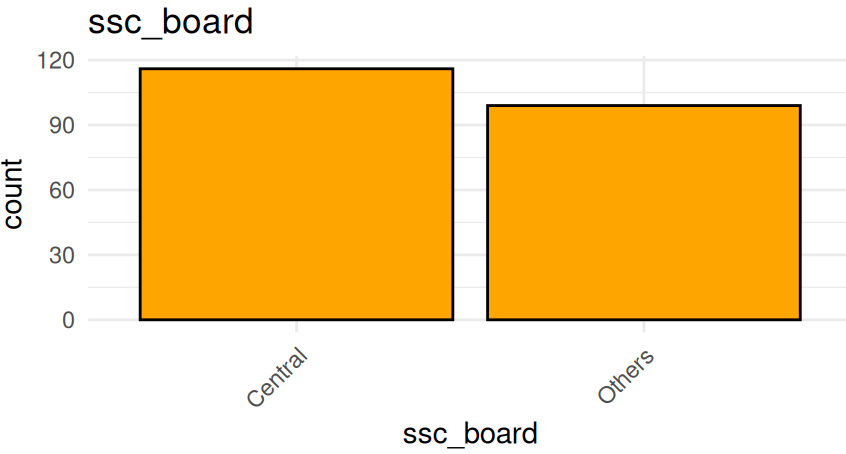
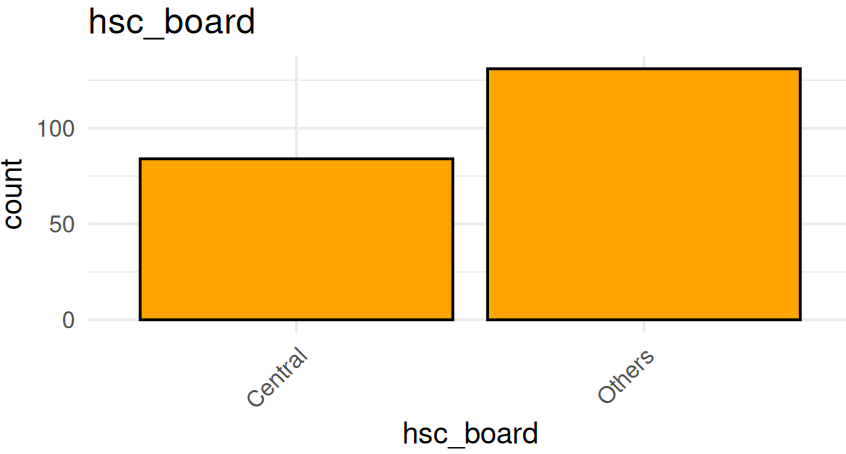
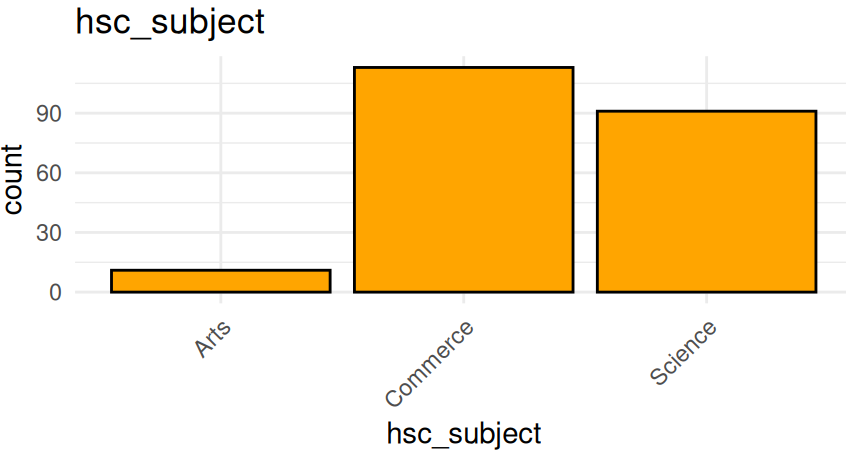
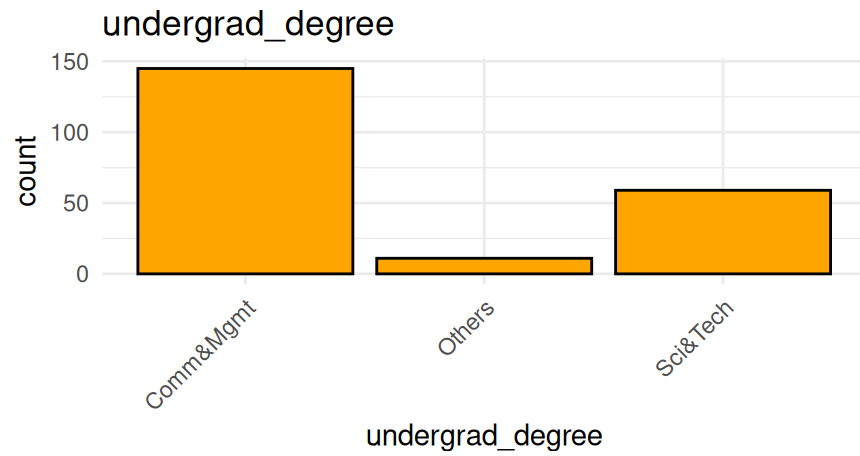
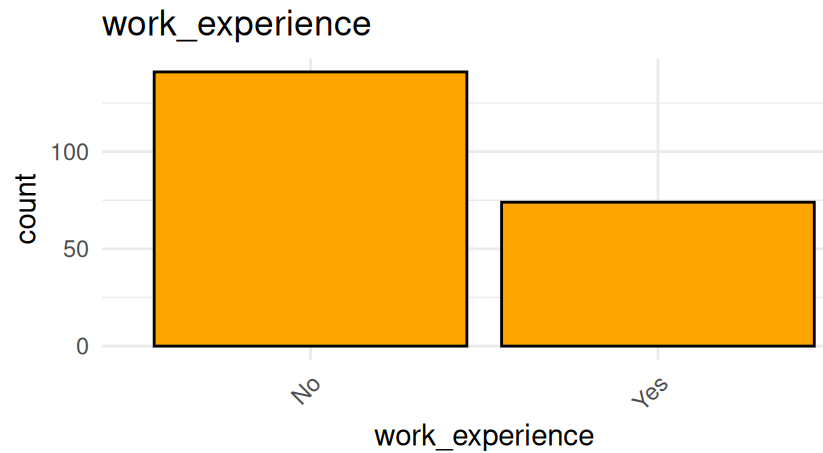
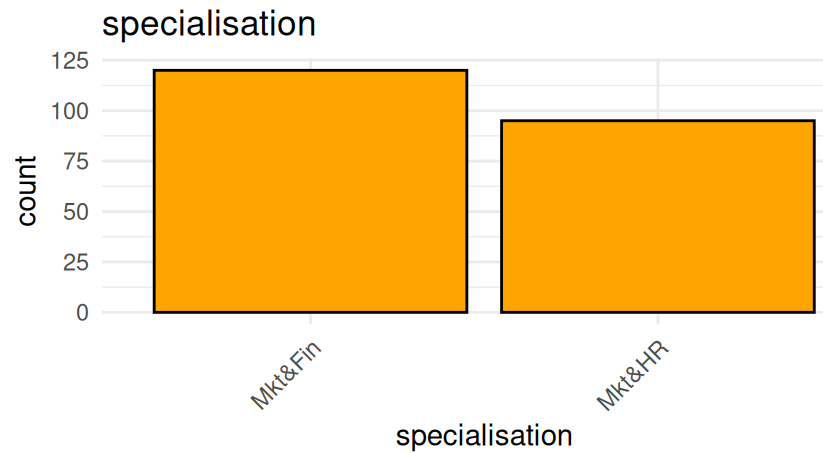
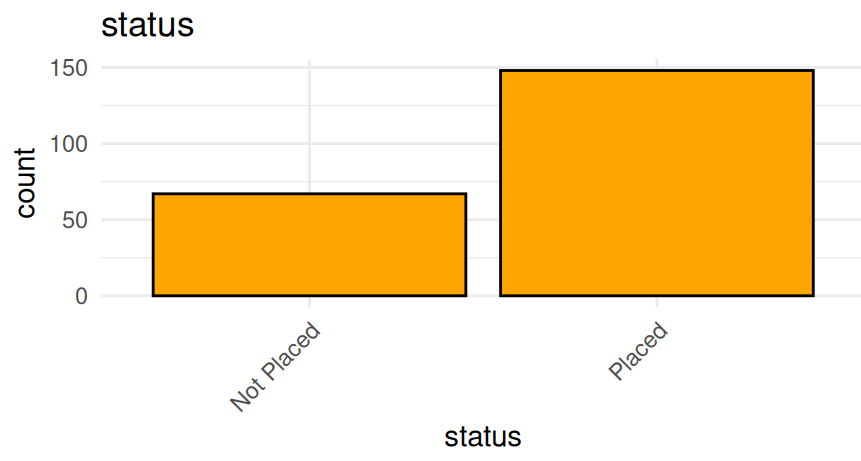


2. Выполнить дескриптивный анализ.

Для этого находим: среднее значение, медиану, моду,дисперсию, стандартное отклонение, минимум, максимум.

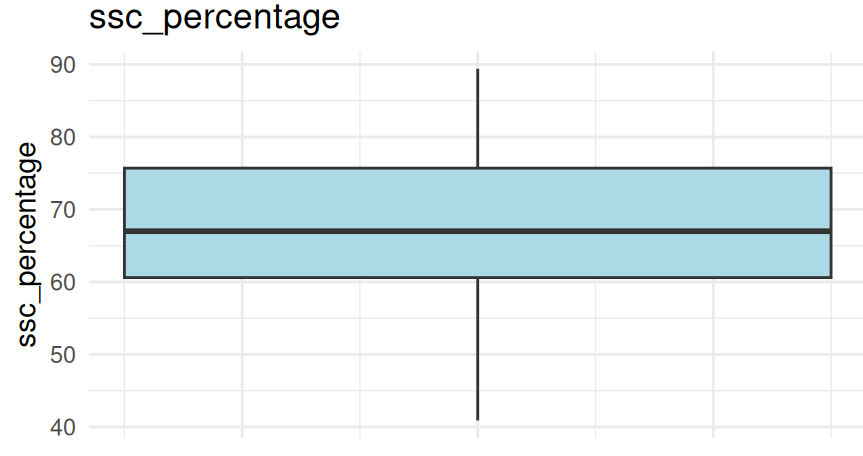
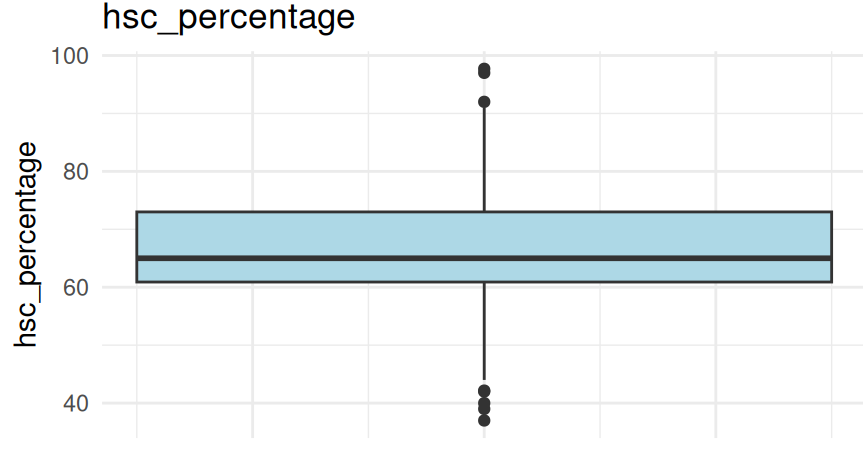
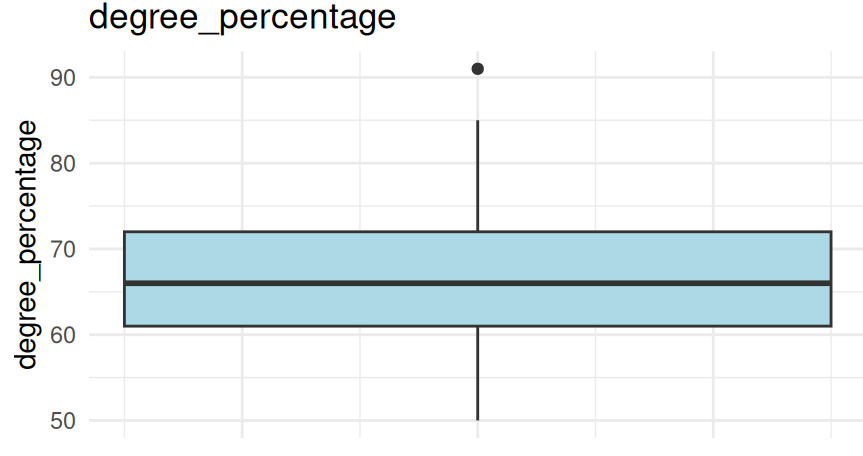
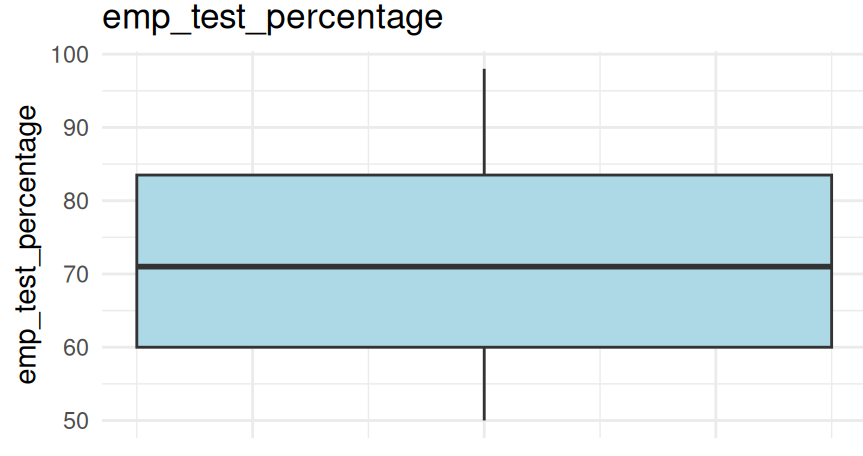
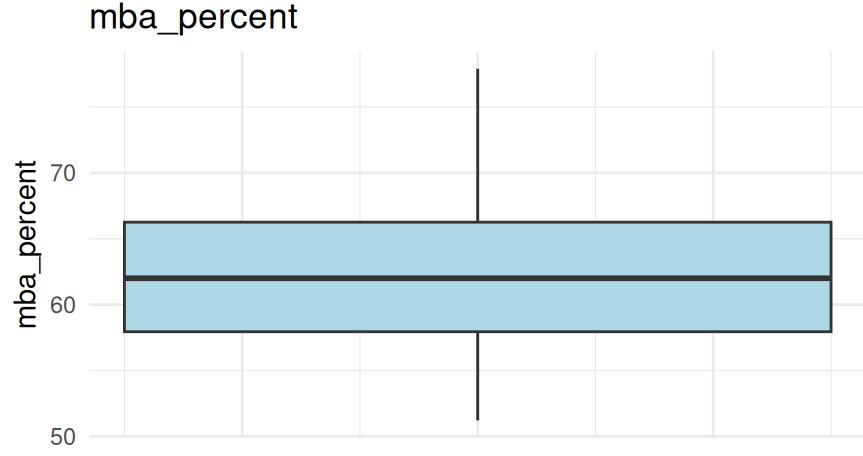


Гистограммы:



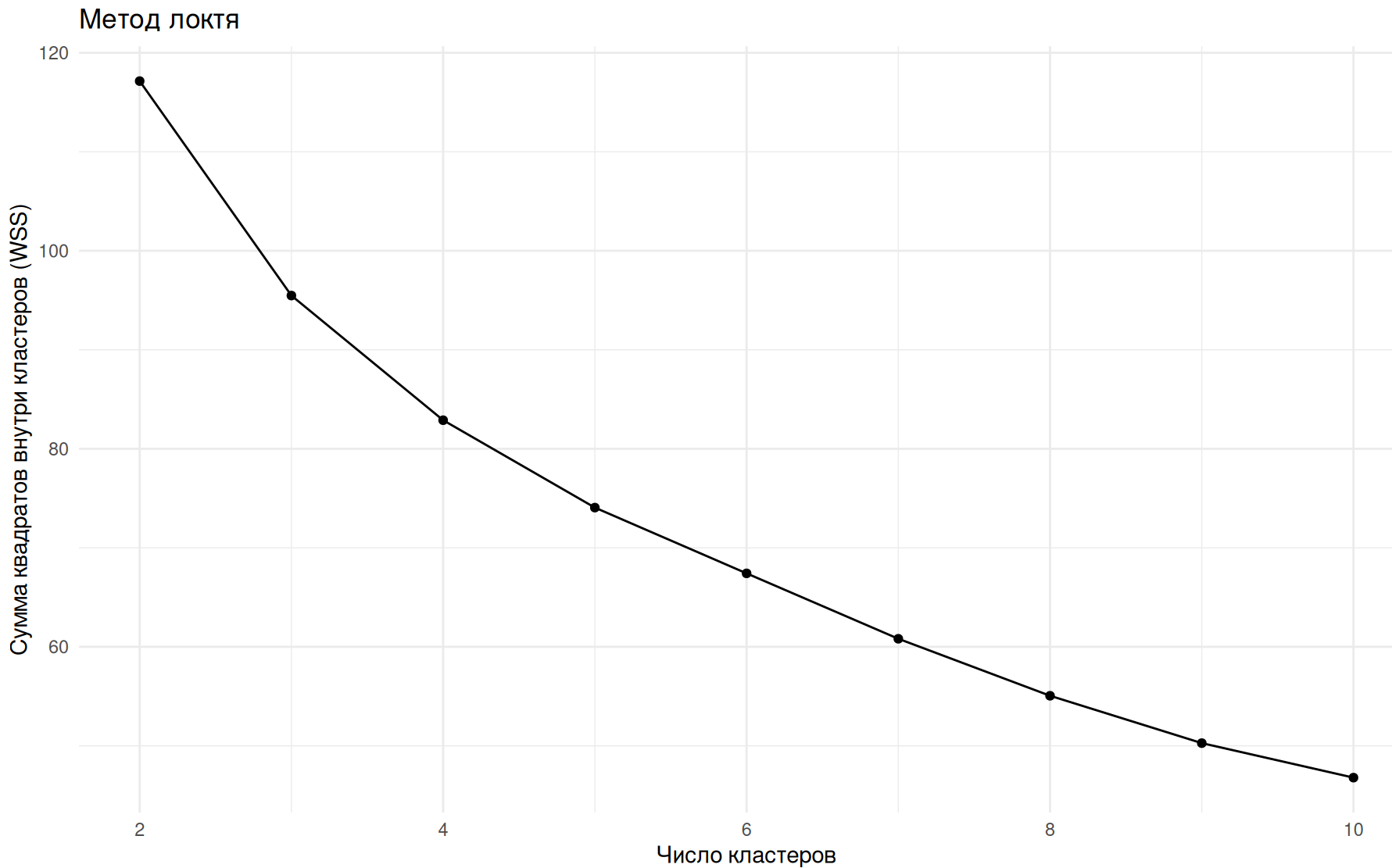
Боксплоты:

Боксплоты нельзя строить для нечисловых данных, потому что они требуют вычисления статистик (медианы, квартилей, межквартильного размаха), которые математически неопределимы для текста, так как не имеют числового порядка, а значит, невозможно определить их распределение, разброс или выбросы — ключевые элементы, которые визуализирует боксплот. Попытка применить этот метод к нечисловым данным приведёт к ошибке в программе или бессмысленному графику, так как алгоритм не сможет ранжировать и обрабатывать качественные признаки как количественные.

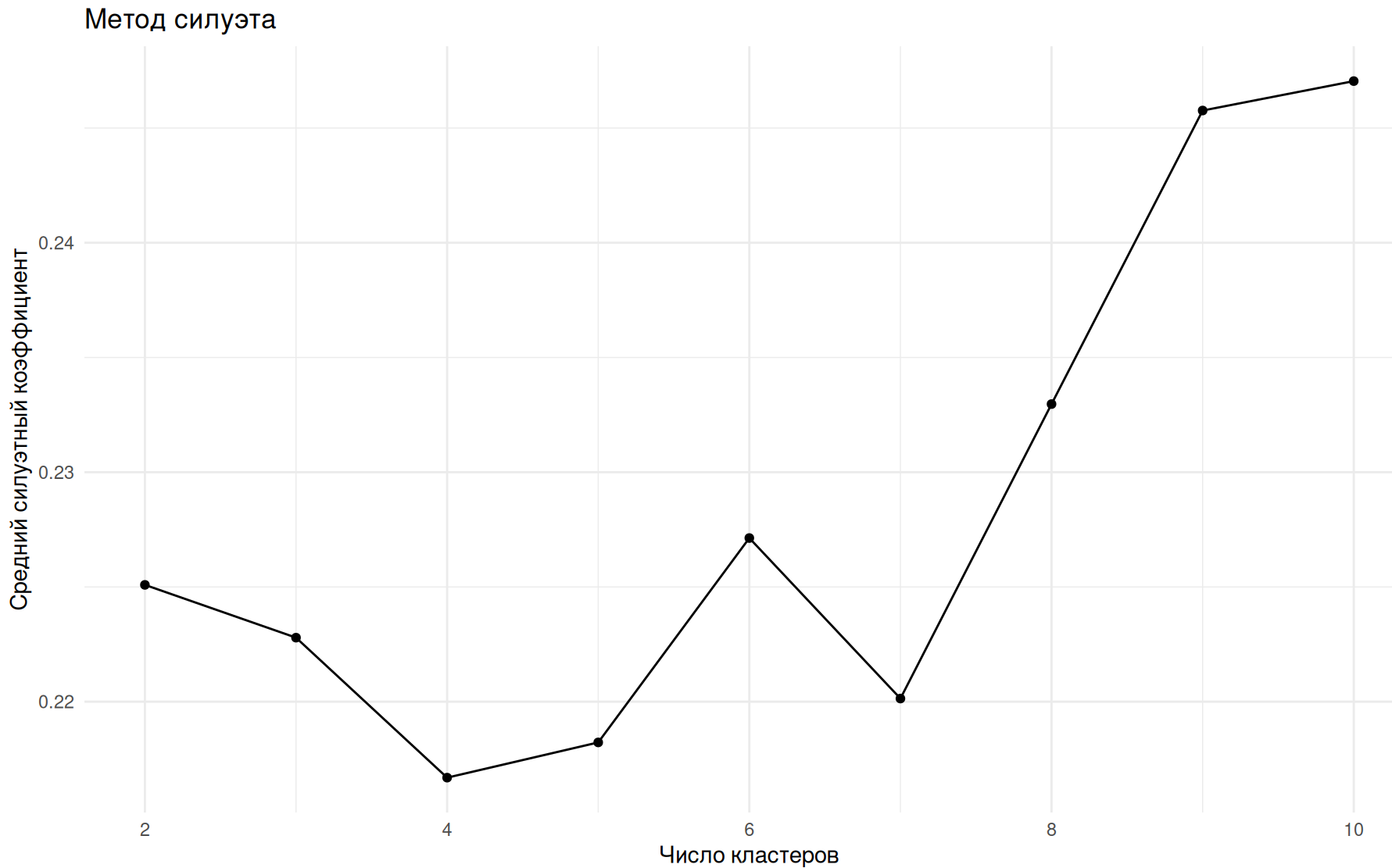


**3.** Оценить оптимальное число кластеров, для этого построить диаграмму "Метод силуэта", “Метод локтя”, "Статистику разрыва" и Алгоритм консенсуса.

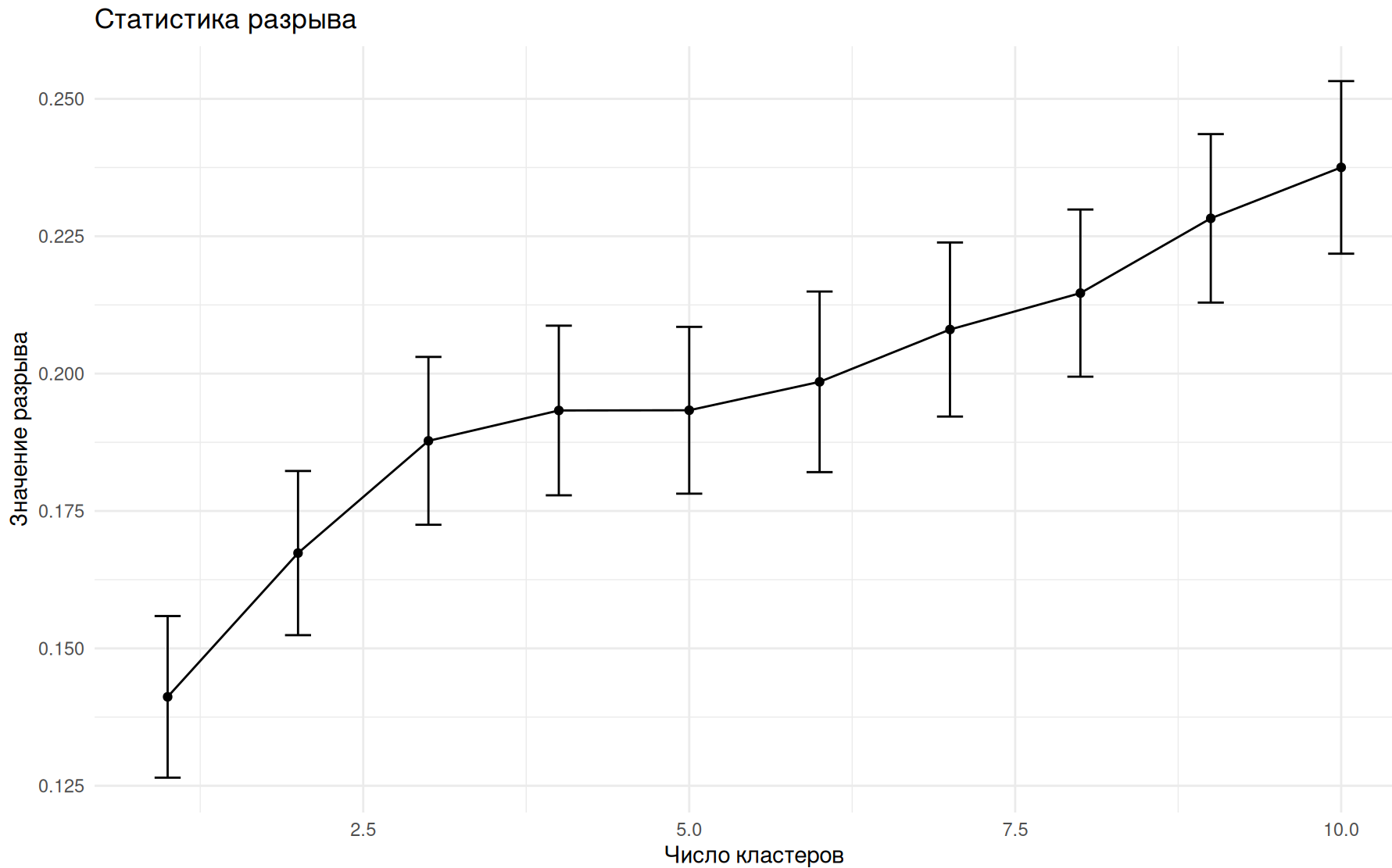
**Метод локтя:**



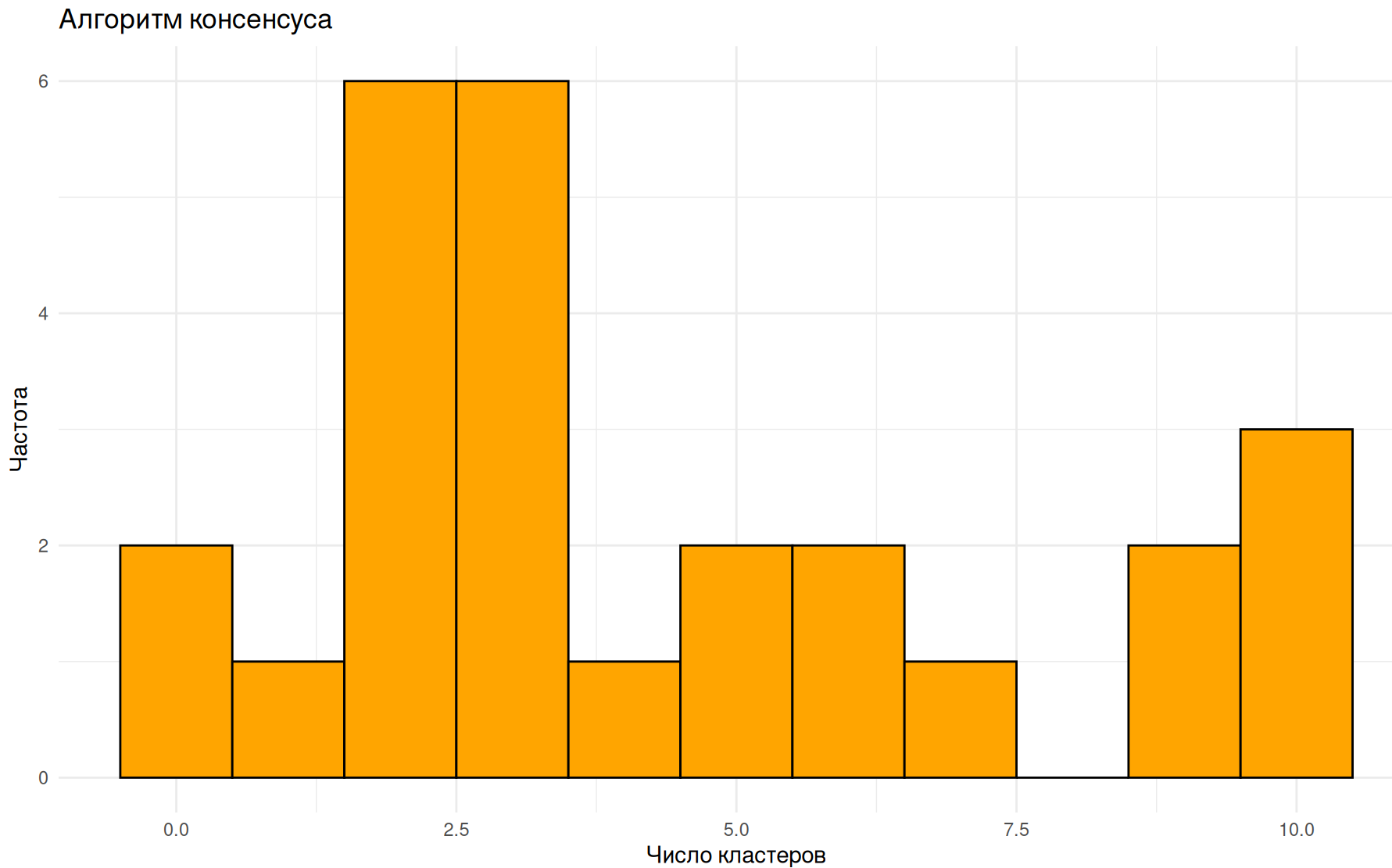
**Метод силуэта:**



**Статистику разрыва:**



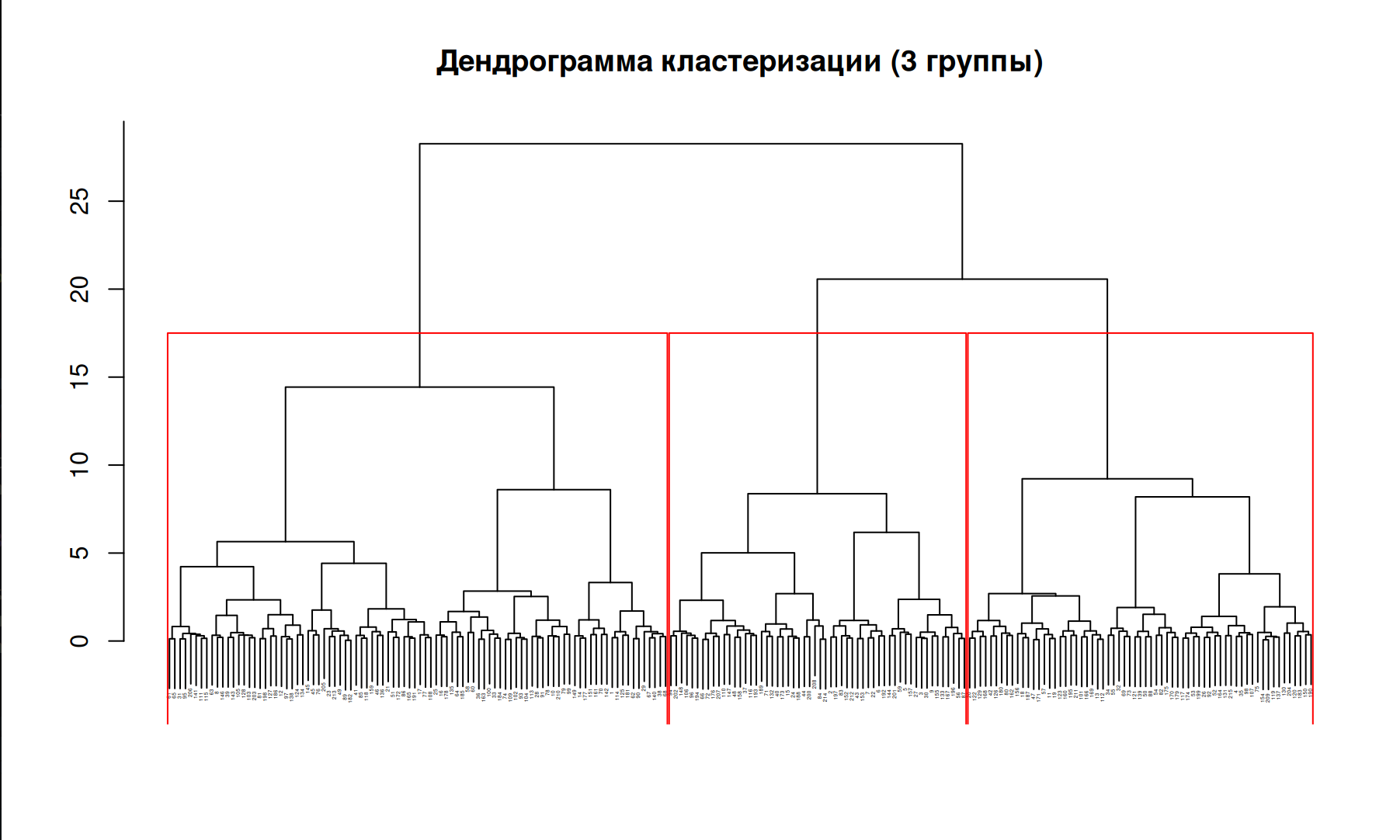
**Алгоритм консенсуса:**



Для всех графиков 3 кластера являются наиболее обоснованными, так как они согласуются с большинством методов (NbClust, WSS, дендрограмма), несмотря на незначительные расхождения в статистике разрыва и силуэтном коэффициенте.

**4.** Выполнить иерархическую кластеризацию вашего набора данных, построив дендрограмму. Подробно обосновать Ваш выбор числа групп.

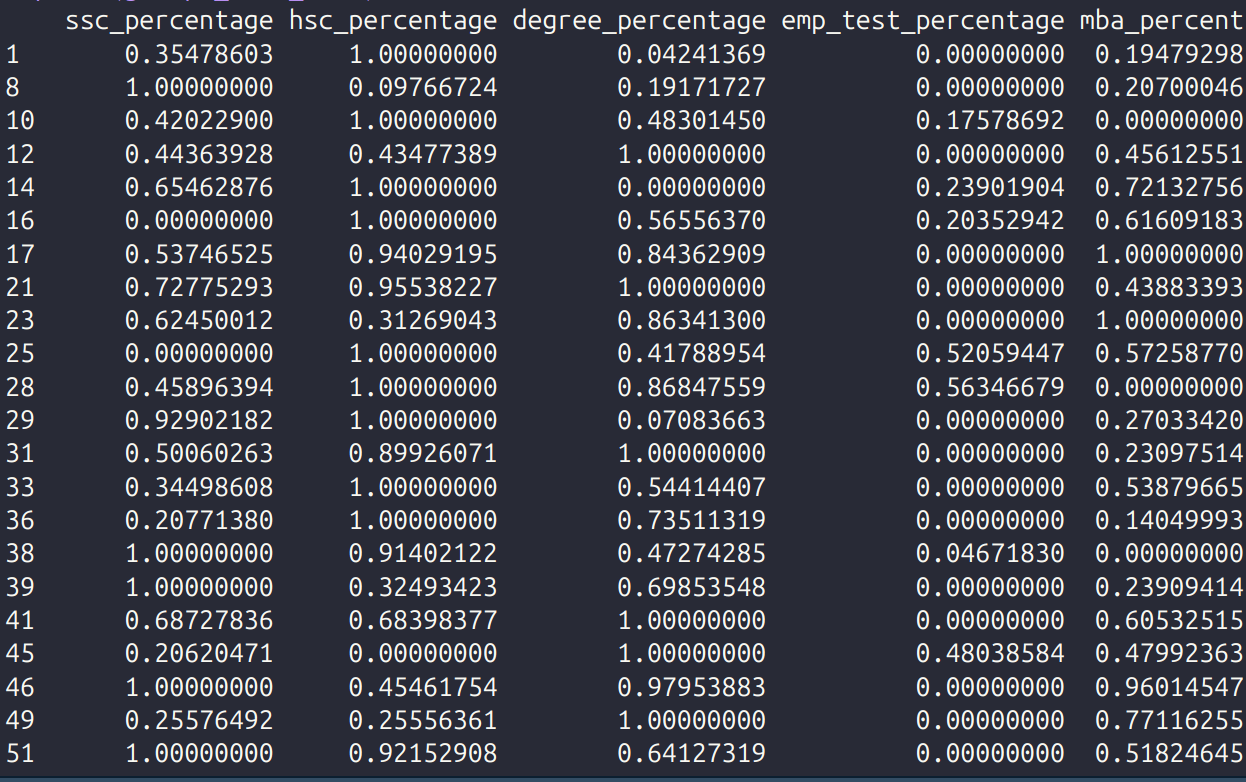
При выполнении иерархической кластеризации дендрограмма визуализирует последовательное объединение объектов, где длина вертикальных линий отражает расстояние между кластерами. Оптимальное число групп определяется по уровню, на котором наблюдается максимальный «разрыв» — резкое увеличение расстояния между объединяемыми кластерами, что указывает на естественное разделение данных. В данном случае такой разрыв четко прослеживается при выделении 3 кластеров, что согласуется с методом локтя и алгоритмом NbClust. Этот выбор обеспечивает баланс между интерпретируемостью, компактностью групп и отсутствием избыточного дробления данных.



Структура кластеров: На графике видно, что объекты последовательно объединяются в три четкие группы. Первый кластер включает точки 0, 1 и 2, что указывает на их высокую схожесть по измеренным признакам. Второй кластер формируется из точек 3 и 4, демонстрируя умеренное расстояние объединения, что говорит о меньшей внутренней однородности. Третий кластер (точки 5, 6, 7) объединяется на более высоком уровне, подчеркивая их относительную изолированность от остальных данных. Такое разделение отражает естественную группировку объектов без явных перекрытий.

Расстояния и разрывы: Вертикальные линии дендрограммы показывают, что наибольший разрыв между кластерами наблюдается при переходе от двух к трём группам. Это подтверждает, что разделение на три кластера оптимально, так как дальнейшее увеличение их числа приводит к незначительному уменьшению расстояний между объектами. Высота линий для объединения третьего кластера (5, 6, 7) значительно превышает другие, что свидетельствует о его уникальности. Таким образом, три кластера сохраняют баланс между компактностью и различимостью.

Согласованность с методами: Выделение трёх групп согласуется с ранее примененными методами — алгоритмом NbClust и методом локтя, которые также указывали на 3 кластера. Хотя метод силуэта формально предлагал большее число групп, дендрограмма подтверждает, что три кластера лучше отражают структуру данных, избегая избыточного дробления. Разделение на три части обеспечивает содержательную интерпретацию в предметной области, например, выделение типовых, переходных и аномальных объектов. Это делает выбор трёх кластеров наиболее логичным и обоснованным.



**5.** Построить диаграмму со столбчатыми диаграммами и боксплотами групп Провести сравнительный анализ полученных групп.

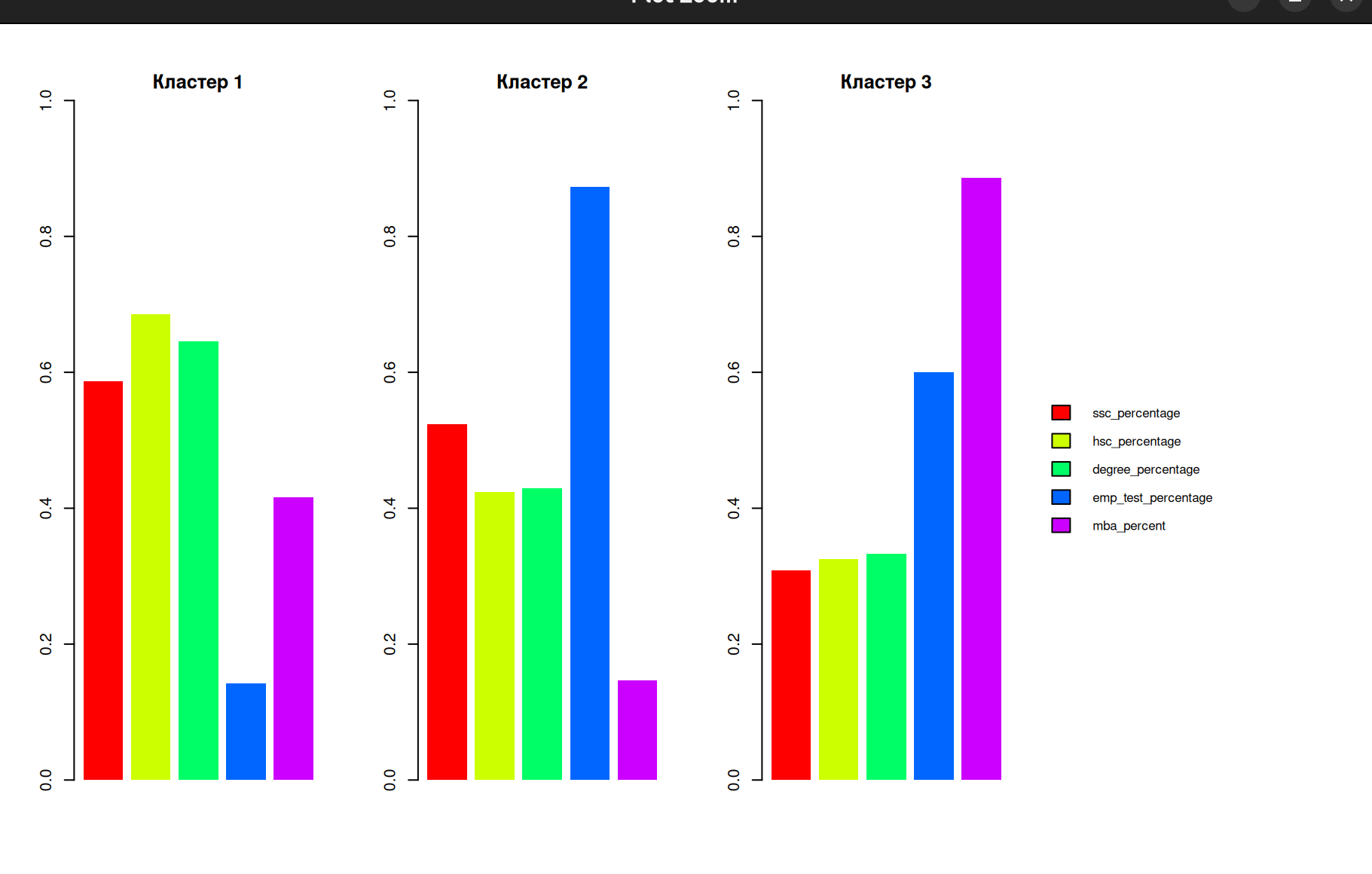


График представляет профили трех кластеров, сформированных на основе нормированных показателей из датасета "Кандидаты на работу", включающего такие переменные, как ssc\_percentage, hsc\_percentage, degree\_percentage, emp\_test\_percentage и mba\_percent. Кластер 1 характеризуется высокими значениями ssc\_percentage (около 0.6), hsc\_percentage (около 0.7) и degree\_percentage (около 0.7), что может указывать на кандидатов с сильной академической подготовкой, но с более низкими показателями emp\_test\_percentage (около 0.2) и mba\_percent (около 0.4). Кластер 2 выделяется максимальным значением emp\_test\_percentage (почти 1.0), что говорит о высокой успеваемости кандидатов на профессиональном тестировании, хотя их ssc\_percentage (около 0.5) и mba\_percent (около 0.2) ниже среднего. Кластер 3 демонстрирует наивысший показатель mba\_percent (почти 0.9), что может свидетельствовать о сильной успеваемости в магистратуре, при этом ssc\_percentage, hsc\_percentage и degree\_percentage остаются на умеренном уровне (около 0.4). Таким образом, кластеры отражают разные сильные стороны кандидатов: академическую успеваемость, профессиональные навыки или успехи в магистратуре.

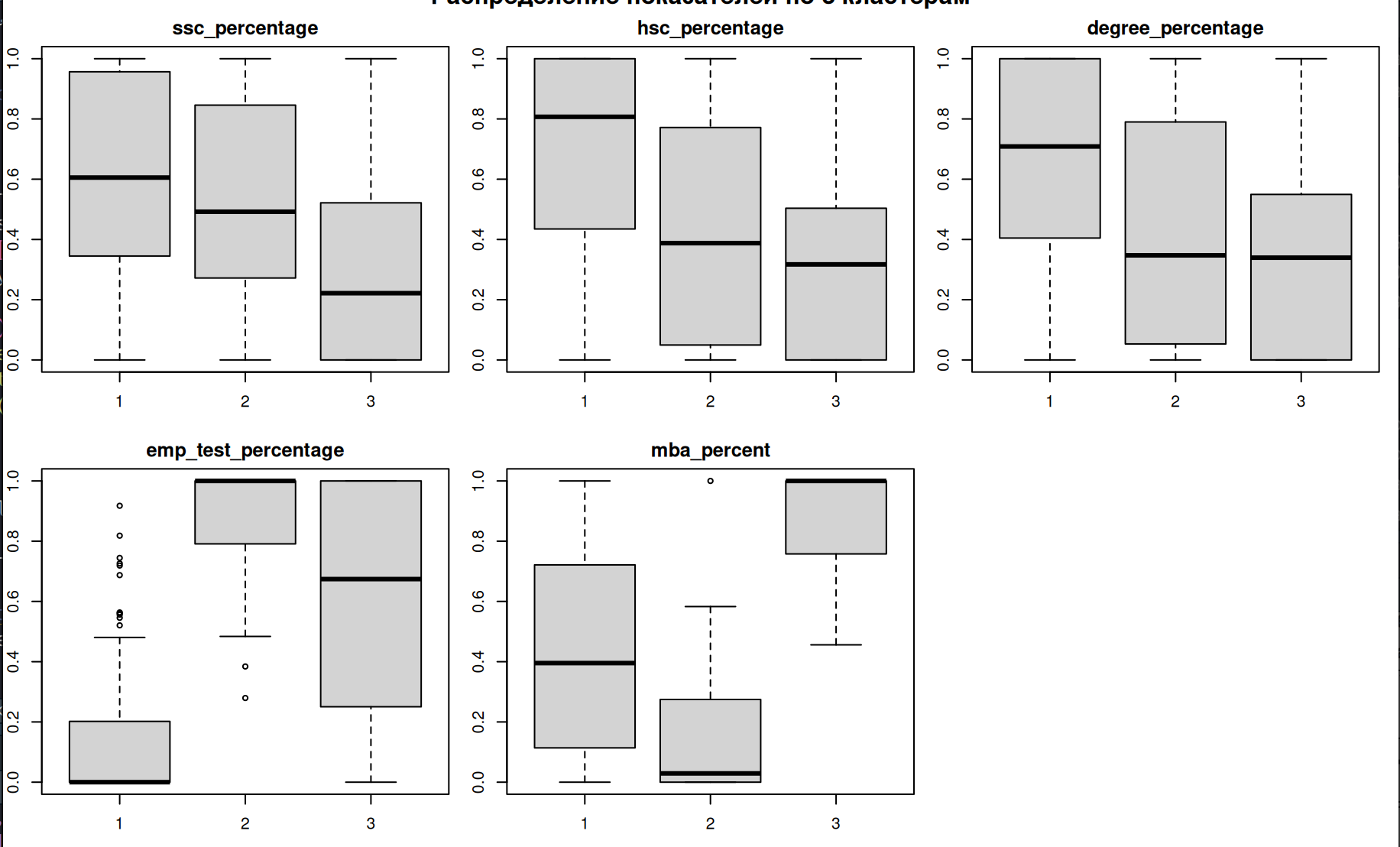


График распределения показателей по трем кластерам из датасета "Кандидаты на работу" демонстрирует вариации в нормированных значениях таких переменных, как ssc\_percentage, hsc\_percentage, degree\_percentage, emp\_test\_percentage и mba\_percent, что позволяет выявить особенности каждой группы. Кластер 1 показывает относительно высокие медианные значения для ssc\_percentage (около 0.6), hsc\_percentage (около 0.7) и degree\_percentage (около 0.6), указывая на кандидатов с сильной академической базой, хотя emp\_test\_percentage и mba\_percent остаются ниже (около 0.2 и 0.4 соответственно). Кластер 2 выделяется высоким значением emp\_test\_percentage (около 0.8), что говорит о профессиональной подготовленности, при этом ssc\_percentage, hsc\_percentage и degree\_percentage имеют более низкие медианы (около 0.5), а mba\_percent едва достигает 0.2. Кластер 3 демонстрирует наибольший разброс и медиану mba\_percent (около 0.7), что может указывать на успехи в магистратуре, тогда как остальные показатели, такие как ssc\_percentage и degree\_percentage, варьируются в диапазоне 0.4–0.6. Таким образом, кластеры отражают разные профили кандидатов: академическую ориентированность, профессиональные навыки и успехи в магистратуре, с заметными различиями в уровне вариации внутри групп.

**6.** Выполнить кластеризацию своего датасета по k-means

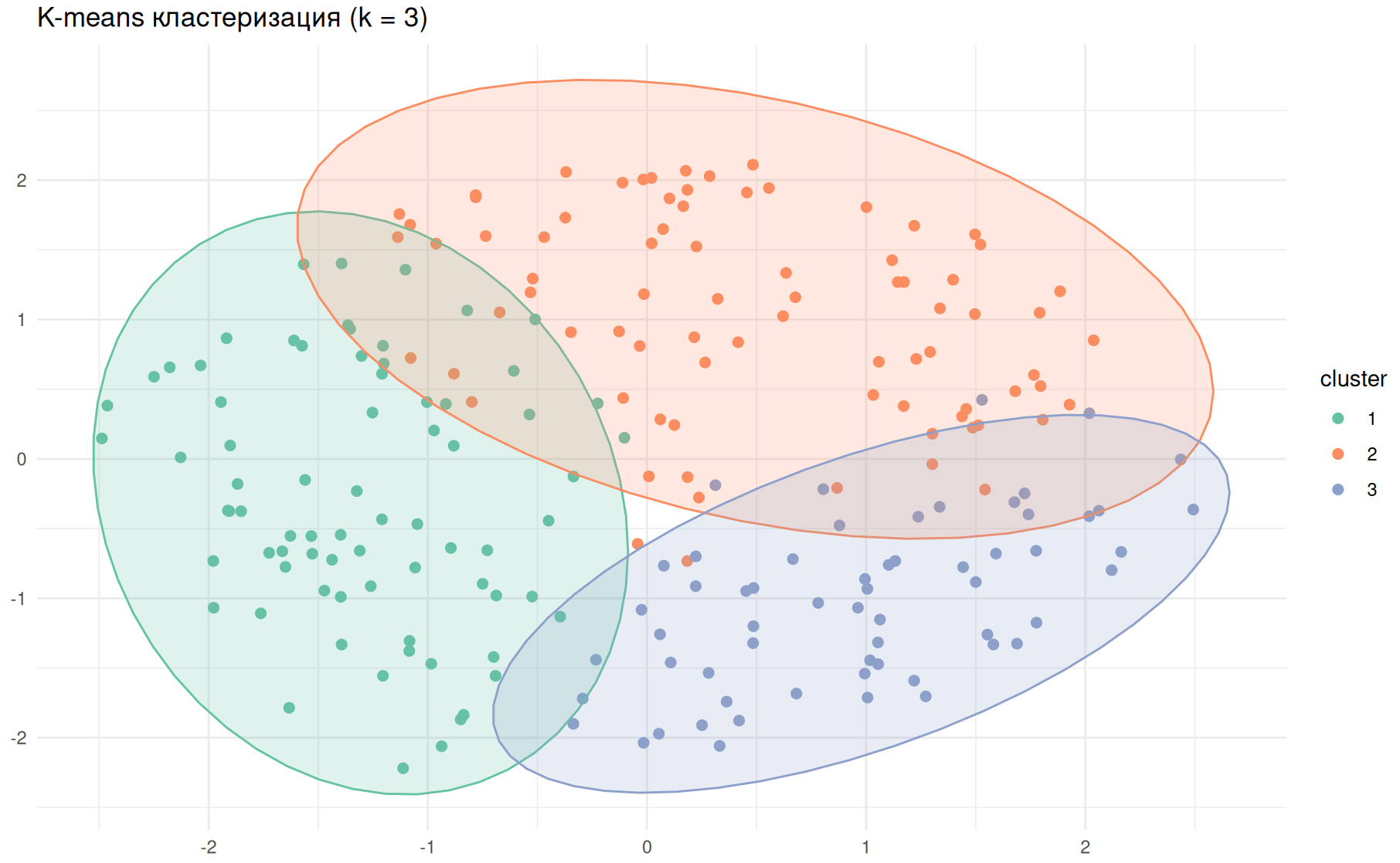


График визуализирует результаты K-means кластеризации с тремя кластерами, выполненной на датасете "Кандидаты на работу", где точки представляют кандидатов, а эллипсы показывают плотность распределения кластеров. Кластер 1 (зеленый) расположен в нижней левой части графика, демонстрируя компактное распределение кандидатов с относительно низкими значениями по осям, что может указывать на схожие академические и профессиональные характеристики. Кластер 2 (оранжевый) занимает центральную и верхнюю часть, пересекается с другими кластерами, что говорит о более высокой вариативности в показателях, таких как ssc\_percentage или emp\_test\_percentage. Кластер 3 (синий) находится в правой части графика, показывая группу кандидатов с более высокими значениями по одной из осей, возможно, mba\_percent, и меньшей вариацией по другой. Таким образом, кластеризация выявляет три группы кандидатов с различными профилями, где пересечения эллипсов указывают на потенциальные сходства между группами.

**7**. Выполнить построение scatterplot с помощью функций plot или pairs.



График диаграмм рассеяния с цветовой кодировкой по трем кластерам из датасета "Кандидаты на работу" демонстрирует распределение кандидатов по ключевым показателям, таким как ssc\_percentage, hsc\_percentage, degree\_percentage, emp\_test\_percentage и mba\_percent. Кластеры, обозначенные зеленым и красным цветами, показывают значительное перекрытие по всем переменным, что указывает на схожие уровни успеваемости среди кандидатов в этих группах, особенно заметное в ssc\_percentage и hsc\_percentage, где точки распределены равномерно в диапазоне от 0.2 до 0.8. Кластер с преобладанием зеленого цвета (возможно, кластер 1) демонстрирует более высокую плотность в среднем диапазоне (0.4–0.6) по большинству показателей, тогда как красный кластер (возможно, кластер 2 или 3) имеет больше точек с более высокими значениями, особенно в emp\_test\_percentage и mba\_percent, где значения доходят до 1.0. Отсутствие четких границ между кластерами на диаграммах говорит о том, что кандидаты имеют смешанные профили, а различия между группами могут быть менее выраженными, что требует дополнительного анализа для уточнения их характеристик. Таким образом, визуализация подчеркивает сложность сегментации кандидатов и необходимость учитывать комбинацию факторов для более точной интерпретации.

**8.** Построить трехмерную кластеризацию по scatterplot3d

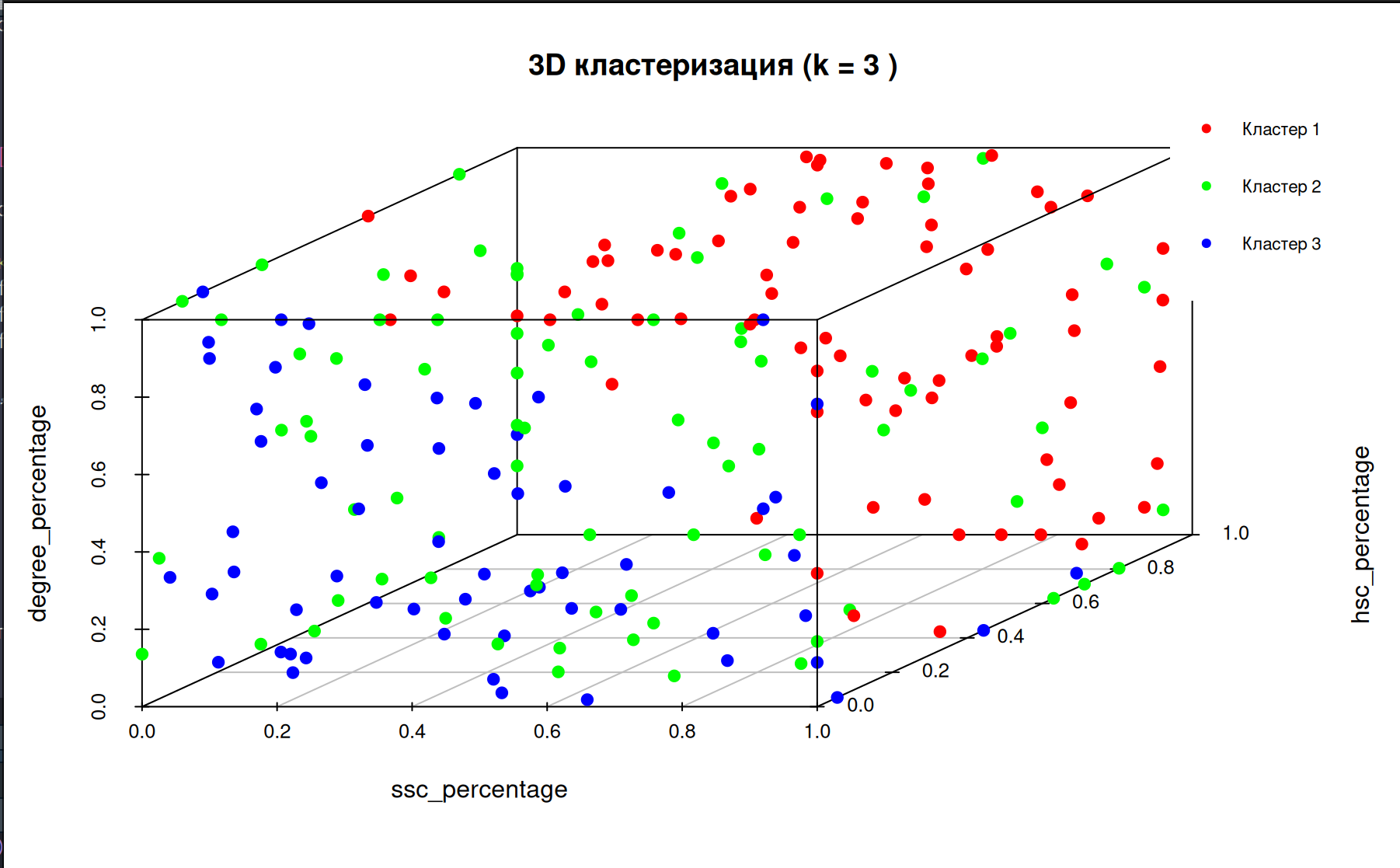


График представляет 3D-визуализацию кластеризации с тремя кластерами из датасета "Кандидаты на работу", где оси соответствуют показателям ssc\_percentage, hsc\_percentage и degree\_percentage, а цвета точек (красный, зеленый, синий) обозначают кластеры 1, 2 и 3 соответственно. Кластер 1 (красный) сосредоточен в верхней части графика, с высокими значениями по всем трем осям (0.8–1.0), что указывает на кандидатов с отличной успеваемостью на всех уровнях образования. Кластер 2 (зеленый) занимает среднюю область, с показателями в диапазоне 0.4–0.8, отражая кандидатов со средними академическими результатами, но с заметной вариацией, особенно в degree\_percentage. Кластер 3 (синий) располагается ближе к нижней части графика, с более низкими значениями (0.0–0.6), что может указывать на менее высокую успеваемость по сравнению с другими группами, особенно в ssc\_percentage и hsc\_percentage. Таким образом, визуализация подчеркивает различия в академическом профиле кандидатов, выделяя три группы с разным уровнем успеваемости по ключевым образовательным показателям.

**Вывод:** я закрепила знания, об алгоритмах классификации и кластеризации данных,ознакомилась с некоторыми функциями языка R, осуществляющими этот вид анализа,принципами их работы. Научилась визуализировать результаты работы функций кластерного анализа и классификаторов, интерпретировать полученные результаты.

**Листинг:**

library(cluster)

library(ggplot2)

library(BBmisc)

library(NbClust)

library(scatterplot3d)

# 1. Загрузка данных

data <- read.csv("~/Documents/6 semester/Big data/5laba/data.csv", header = TRUE, sep = ",")

# 2. Дескриптивный анализ

numeric\_cols <- data[, sapply(data, is.numeric)]

calculate\_stats <- function(x) {

mean\_val <- mean(x, na.rm = TRUE)

median\_val <- median(x, na.rm = TRUE)

mode\_val <- as.numeric(names(sort(table(x), decreasing = TRUE)[1]))

var\_val <- var(x, na.rm = TRUE)

sd\_val <- sd(x, na.rm = TRUE)

min\_val <- min(x, na.rm = TRUE)

max\_val <- max(x, na.rm = TRUE)

data.frame(

Среднее = mean\_val,

Медиана = median\_val,

Мода = mode\_val,

Дисперсия = var\_val,

СтандартноеОтклонение = sd\_val,

Минимум = min\_val,

Максимум = max\_val

)

}

stats\_list <- lapply(numeric\_cols, calculate\_stats)

stats\_table <- do.call(rbind, stats\_list)

stats\_table <- cbind(Column = names(numeric\_cols), stats\_table)

rownames(stats\_table) <- NULL

print(stats\_table)

# Визуализация данных

plot\_analysis <- function(df) {

numeric\_cols <- sapply(df, is.numeric)

numeric\_names <- names(df)[numeric\_cols]

categorical\_names <- names(df)[!numeric\_cols]

for (col in numeric\_names) {

p1 <- ggplot(df, aes(y = !!sym(col))) +

geom\_boxplot(fill = "lightblue", width = 0.3) +

ggtitle(paste(col)) +

theme\_minimal() +

theme(

axis.title.x = element\_blank(),

axis.text.x = element\_blank(),

axis.ticks.x = element\_blank()

) +

scale\_y\_continuous(name = col)

p2 <- ggplot(df, aes(x = !!sym(col))) +

geom\_histogram(fill = "orange", bins = 30, color = "black") +

ggtitle(paste(col)) +

theme\_minimal()

print(p1)

print(p2)

}

for (col in categorical\_names) {

p <- ggplot(df, aes(x = !!sym(col))) +

geom\_bar(fill = "orange", color = "black") +

ggtitle(paste(col)) +

theme\_minimal() +

theme(axis.text.x = element\_text(angle = 45, hjust = 1))

print(p)

}

}

plot\_analysis(data)

# 3. Оценка оптимального числа кластеров

numeric\_cols <- na.omit(numeric\_cols) # Удаление NA

df\_scaled <- scale(numeric\_cols) # Стандартизация

df\_sc\_norm <- normalize(df\_scaled, method = "range", range = c(0, 1)) # Нормализация

# Метод силуэта

silhouette\_score <- function(k, data) {

km <- kmeans(data, centers = k, nstart = 25)

ss <- silhouette(km$cluster, dist(data))

mean(ss[, 3])

}

k\_values <- 2:10

sil\_scores <- sapply(k\_values, silhouette\_score, data = df\_sc\_norm)

sil\_df <- data.frame(k = k\_values, Silhouette = sil\_scores)

p\_sil <- ggplot(sil\_df, aes(x = k, y = Silhouette)) +

geom\_line() +

geom\_point() +

labs(title = "Метод силуэта", x = "Число кластеров", y = "Средний силуэтный коэффициент") +

theme\_minimal()

print(p\_sil)

# Метод локтя

wss\_score <- function(k, data) {

km <- kmeans(data, centers = k, nstart = 25)

km$tot.withinss

}

wss\_scores <- sapply(k\_values, wss\_score, data = df\_sc\_norm)

wss\_df <- data.frame(k = k\_values, WSS = wss\_scores)

p\_wss <- ggplot(wss\_df, aes(x = k, y = WSS)) +

geom\_line() +

geom\_point() +

labs(title = "Метод локтя", x = "Число кластеров", y = "Сумма квадратов внутри кластеров (WSS)") +

theme\_minimal()

print(p\_wss)

# Статистика разрыва

gap\_stat <- clusGap(df\_sc\_norm, FUN = kmeans, nstart = 25, K.max = 10, B = 50)

gap\_df <- data.frame(k = 1:10, Gap = gap\_stat$Tab[, "gap"], SE = gap\_stat$Tab[, "SE.sim"])

p\_gap <- ggplot(gap\_df, aes(x = k, y = Gap)) +

geom\_line() +

geom\_point() +

geom\_errorbar(aes(ymin = Gap - SE, ymax = Gap + SE), width = 0.2) +

labs(title = "Статистика разрыва", x = "Число кластеров", y = "Значение разрыва") +

theme\_minimal()

print(p\_gap)

# Алгоритм консенсуса

n\_clust <- NbClust(df\_sc\_norm, distance = "euclidean", min.nc = 2, max.nc = 10,

method = "kmeans", index = "all")

nbclust\_df <- data.frame(k = n\_clust$Best.nc["Number\_clusters", ])

p\_nbclust <- ggplot(nbclust\_df, aes(x = k)) +

geom\_histogram(binwidth = 1, fill = "orange", color = "black") +

labs(title = "Алгоритм консенсуса (NbClust)", x = "Число кластеров", y = "Частота") +

theme\_minimal()

print(p\_nbclust)

# 4. Выполнить иерархическую кластеризацию вашего набора данных, построив дендрограмму

dist.datas=dist(df\_sc\_norm)

clust.datas=hclust(dist.datas,'ward.D')

plot(clust.datas,cex=0.2, ann = FALSE)

title(main = "Дендрограмма кластеризации (3 группы)")

rect.hclust(clust.datas,k=3,border="red")

cluster\_cut3 <- cutree(clust.datas, k = 3)

group1\_indices <- which(cluster\_cut3 == 1)

group1\_data\_norm <- df\_sc\_norm[group1\_indices, ]

print(group1\_data\_norm)

# 5. Построить диаграмму со столбчатыми диаграммами и боксплотами групп.

plot\_cluster\_profiles <- function(clust\_obj, df\_norm, k) {

while (dev.cur() > 1) dev.off()

dev.new()

cluster\_cut <- cutree(clust\_obj, k = k)

cluster\_means <- t(sapply(1:k, function(i) {

colMeans(df\_norm[cluster\_cut == i, , drop = FALSE])

}))

param\_colors <- rainbow(ncol(df\_norm))

layout\_matrix <- matrix(c(1:k, rep(k+1, 2)), nrow = 1)

layout(layout\_matrix, widths = c(rep(4, k), 3))

par(oma = c(0, 0, 2, 0), mar = c(5, 4, 2, 1))

for (i in 1:k) {

barplot(cluster\_means[i, ],

col = param\_colors,

ylim = c(0, 1),

border = NA,

main = paste("Кластер", i),

names.arg = rep("", ncol(df\_norm)))

}

par(mar = c(0, 0, 0, 0))

plot.new()

legend("left",

legend = colnames(df\_norm),

fill = param\_colors,

xpd = NA,

bty = "n",

cex = 0.8)

layout(1)

}

plot\_cluster\_profiles(clust.datas, df\_sc\_norm, k = 3)

plot\_cluster\_boxplots(clust.datas, df\_sc\_norm, k = 3)

# 6. Выполнить кластеризацию датасета по k-means

library(ggplot2)

library(stats)

set.seed(123)

kmeans\_3 <- kmeans(df\_sc\_norm, centers = 3, nstart = 25)

pca\_result <- prcomp(df\_sc\_norm, scale. = TRUE)

df\_pca <- data.frame(pca\_result$x[, 1:2], Cluster = factor(kmeans\_3$cluster))

p <- ggplot(df\_pca, aes(x = PC1, y = PC2, color = Cluster)) +

geom\_point(size = 2) +

stat\_ellipse(aes(fill = Cluster), geom = "polygon", alpha = 0.2, show.legend = FALSE) + # Эллипсы плотности

scale\_color\_brewer(palette = "Set2") +

scale\_fill\_brewer(palette = "Set2") +

ggtitle("K-means кластеризация (k = 3)") +

xlab(NULL) +

ylab(NULL) +

theme\_minimal() +

guides(color = guide\_legend(title = "cluster"))

print(p)

split(row.names(df\_sc\_norm), kmeans\_3$cluster)

# 7. Выполнить построение scatterplot с помощью функций plot или pairs.

plot\_cluster\_pairs <- function(kmeans, k) {

cluster\_cut <- kmeans$cluster

pairs(df\_sc\_norm,

col = cluster\_cut,

pch = 19,

main = paste("Диаграммы рассеяния с цветами кластеров (k =", k, ")"))

}

plot\_cluster\_pairs(kmeans\_3, 3)

# 8. Построить трехмерную кластеризацию по scatterplot3d.

plot\_3d\_clusters <- function(df\_norm, kmeans\_result, k) {

x <- df\_norm[, 1]

y <- df\_norm[, 2]

z <- df\_norm[, 3]

cluster\_palette <- rainbow(k)

point\_colors <- cluster\_palette[kmeans\_result$cluster]

scatterplot3d(x, y, z,

color = point\_colors,

pch = 19,

main = paste("3D кластеризация (k =", k, ")"),

xlab = colnames(df\_norm)[1],

ylab = colnames(df\_norm)[2],

zlab = colnames(df\_norm)[3])

legend("topright",

legend = paste("Кластер", 1:k),

col = cluster\_palette,

pch = 19,

cex = 0.7,

inset = -0.09,

xpd = TRUE,

box.lty = 0)

}

plot\_3d\_clusters(df\_sc\_norm, kmeans\_3, k = 3)