**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.

Научиться выполнять классификацию на основе формулы Байеса и деревьев решений.

**Ход работы**

**Часть 1. Задача кластеризации и классификации.**

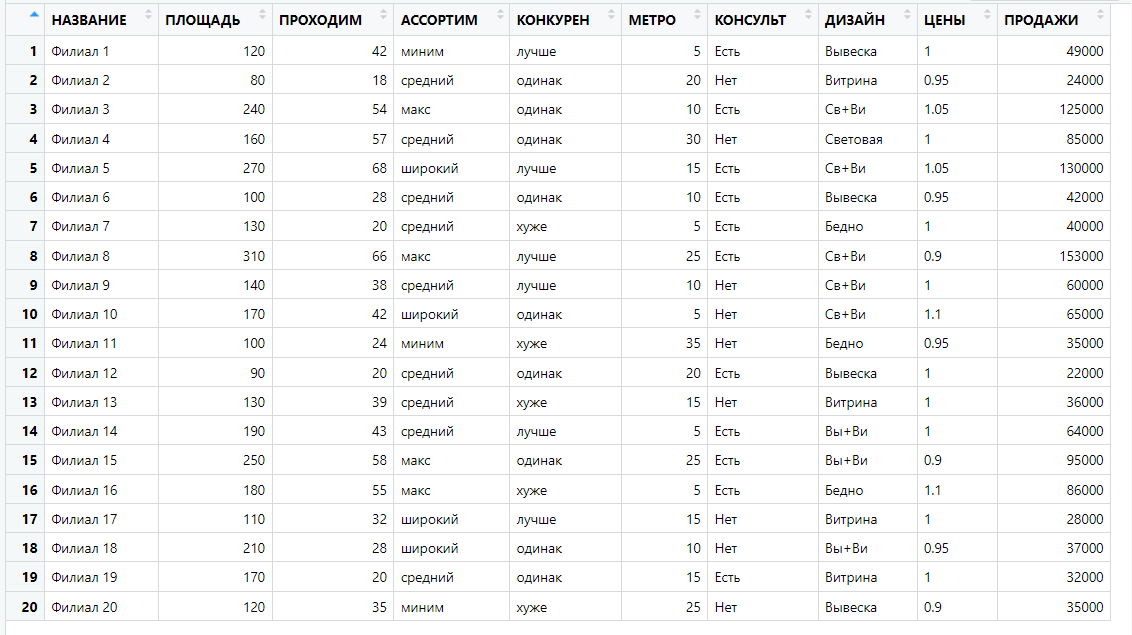


Рисунок 1 – Исходная таблица с данными

В данной задаче переменные существенно различны, поэтому необходимо выполнить нормализацию данных. На рисунке 2 изображена таблица после нормализации.

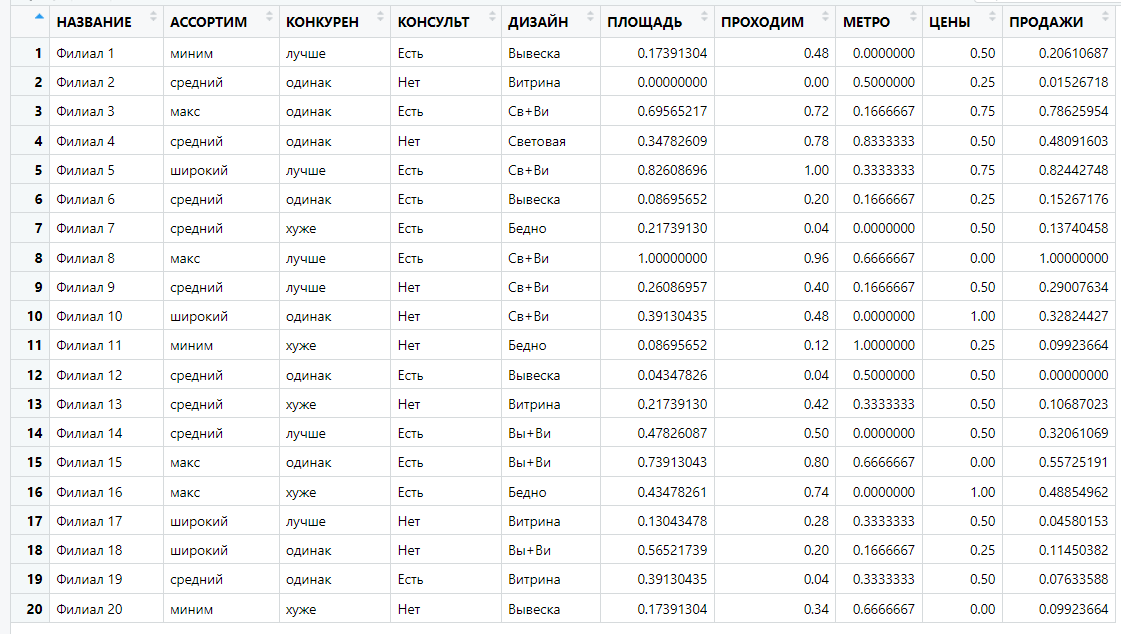


Рисунок 2 – Нормализированная таблица с данными

По полученным данным была построена дендрограмма. Она изображена на рисунке 3.

Изображение выглядит как диаграмма, зарисовка, Технический чертеж, План

Автоматически созданное описание

Рисунок 3 – Дендограмма, построенная по полученным данным

Увидев дендрограмму, очевидно, что целесообразнее выделить 4 кластера.

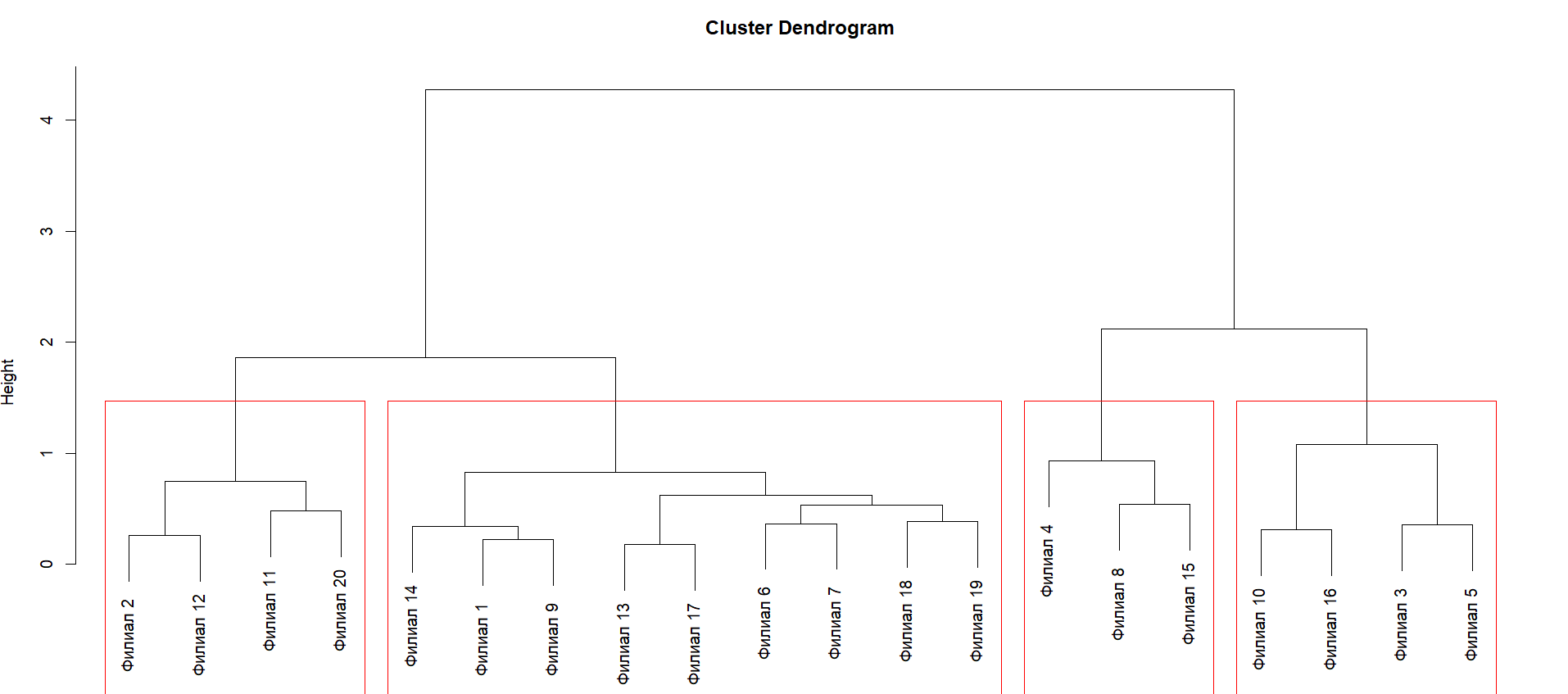


Рисунок 4 – Дендрограмма с кластерами

Далее для каждого кластера определяем средние значения характеристик. На рисунке 5 изображен график распределения всех характеристик соответственно.

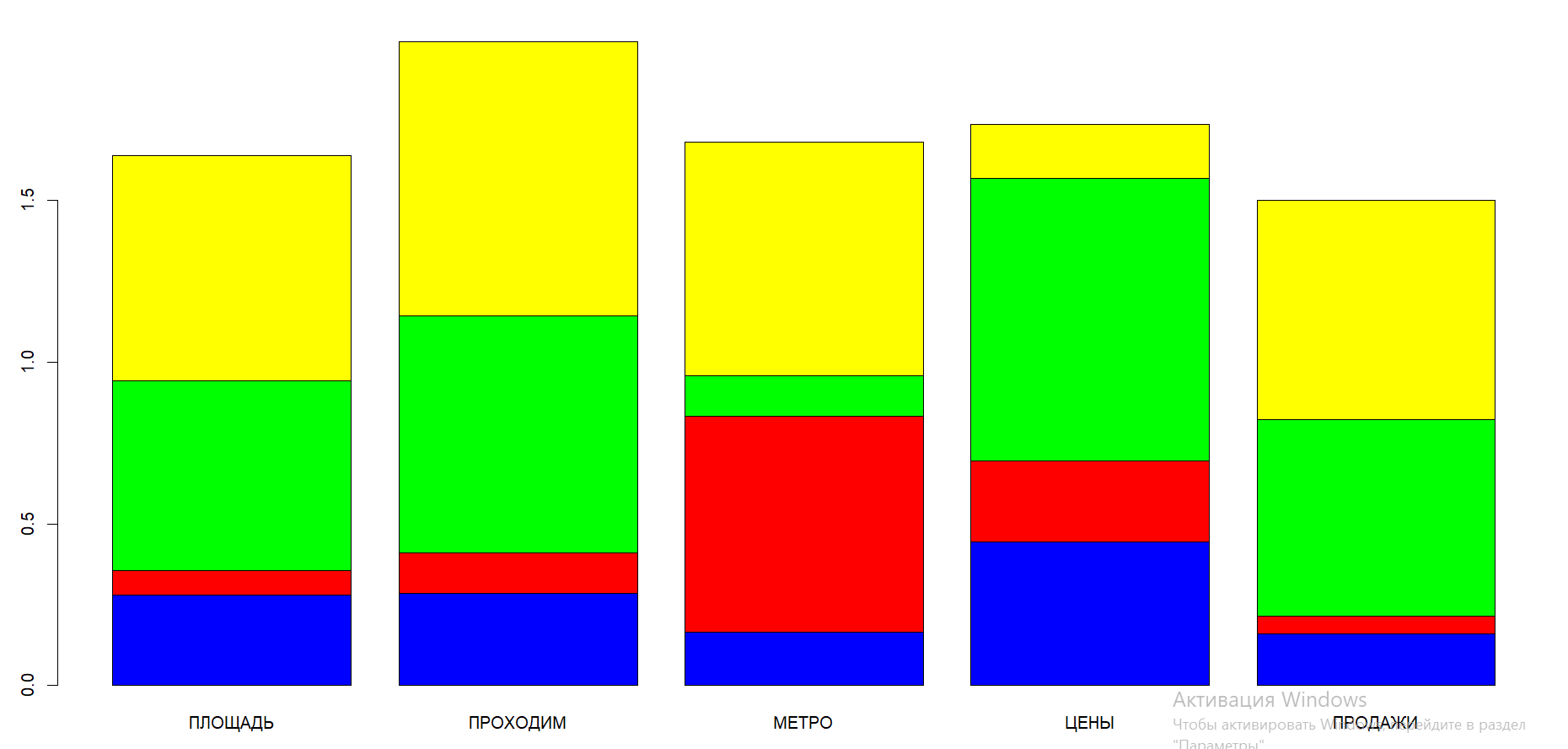


Рисунок 5 – Распределение характеристик

Кластер №4 выделяется большими значениями почти в каждой из характеристик. Это прибыльные, большие филиалы, которые точно не стоит закрывать. Кластер №2 выглядит невыгодным: довольно маленькая площадь, низкая проходимость, самые низкие цены и продажи, хотя он расположен ближе к метро. Вполне вероятно, близкое нахождение к метро не используется в маркетинговом продвижении, из-за чего покупатели просто не замечают эти филиалы. Остальные два кластера находятся на уровне между описанными выше и ничем особо не выделяются.

Затем, для анализа были построены графики изучения по каждой из характеристик на рисунках 6, 8, 9, 10, 11.

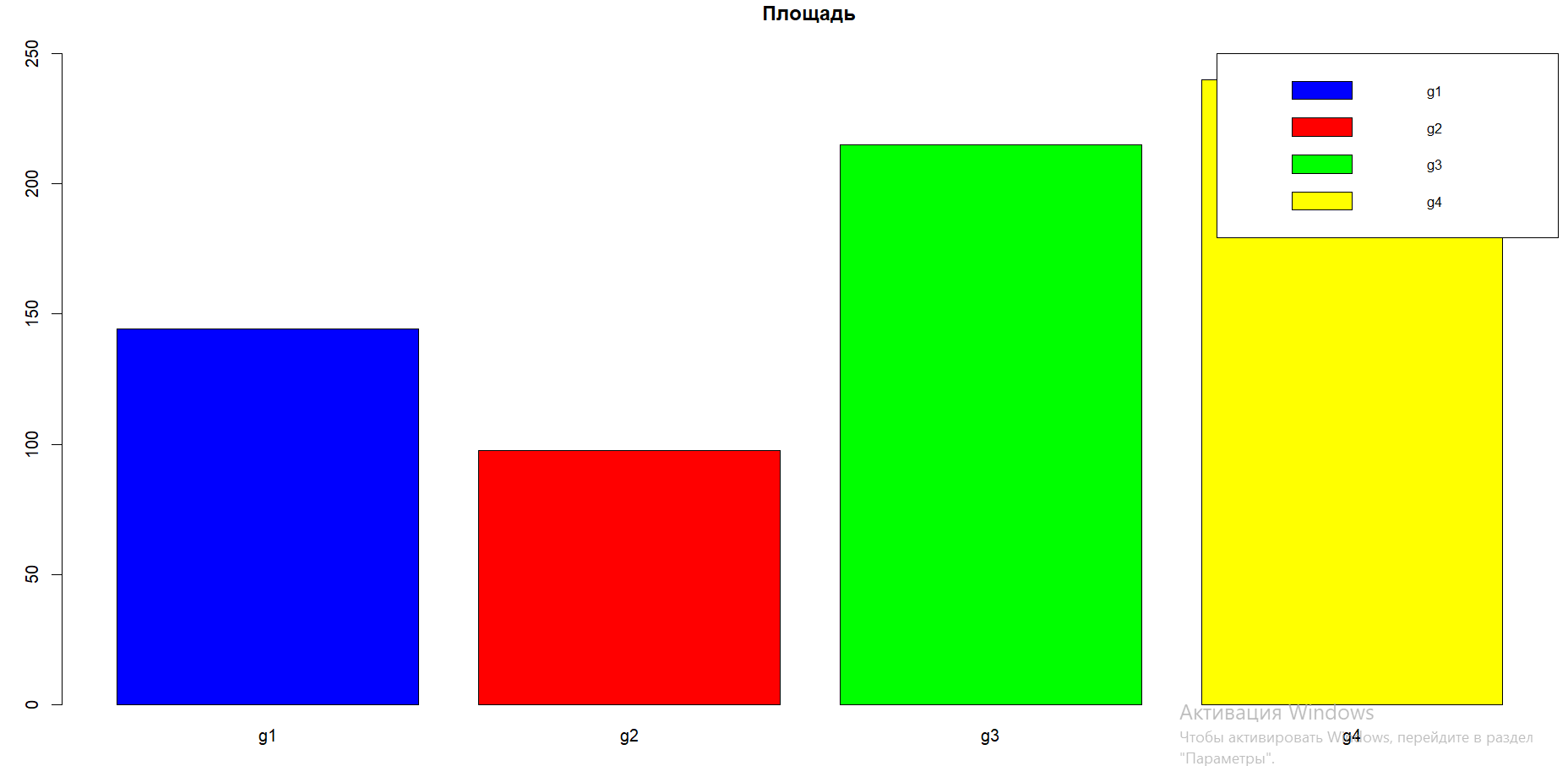


Рисунок 6 – Распределение характеристики ПЛОЩАДЬ

Исходя из графика, можно сделать вывод, что существует всего 2–3 типа площади для помещения филиалов.

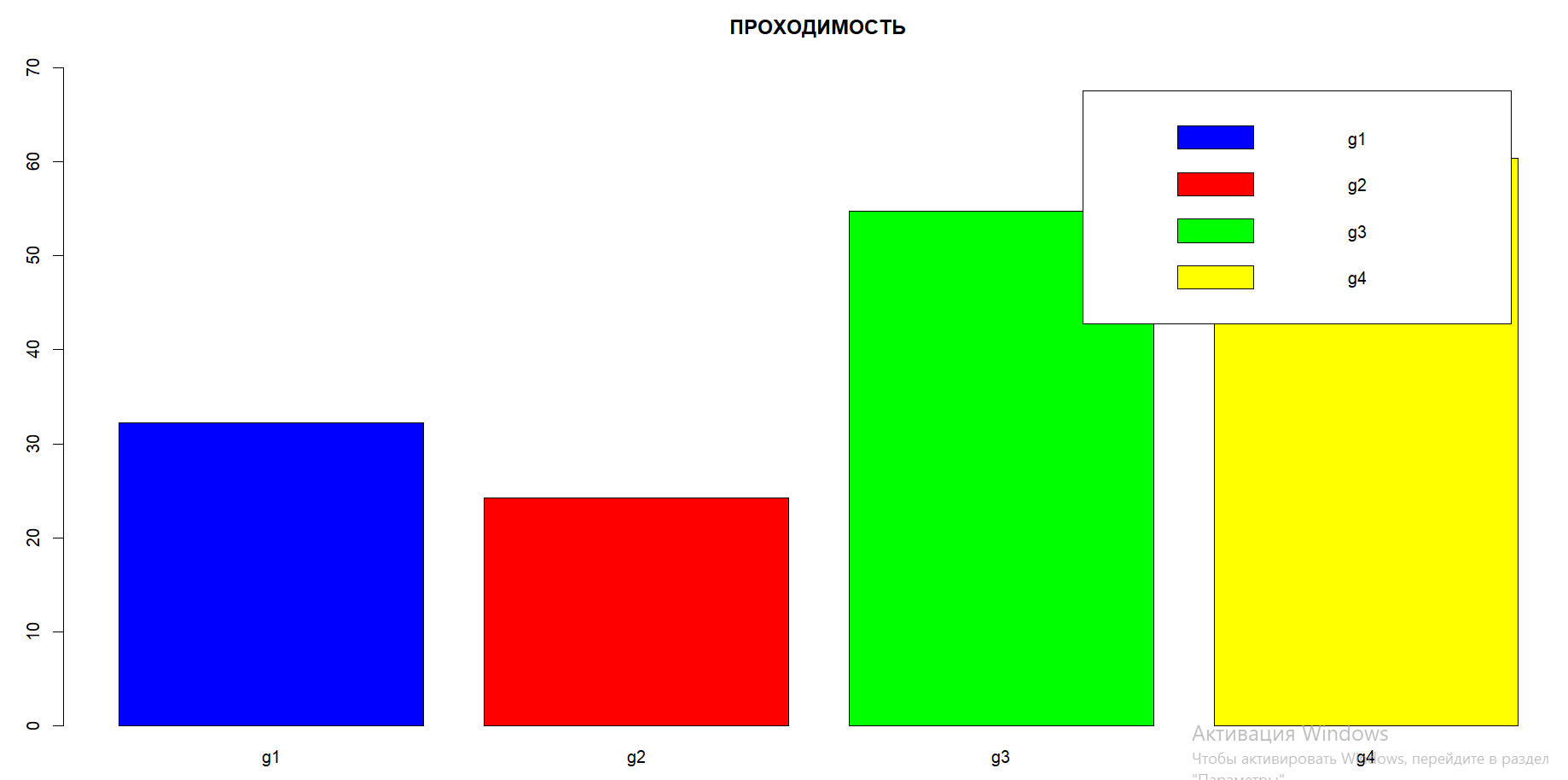


Рисунок 7 – Распределение характеристики ПРОХОДИМОСТЬ

Данный график схож с предыдущим графиков. Аналогично можно сказать, что для всех филиалов выделяются 2–3 типа среднего количества входящих в магазин.

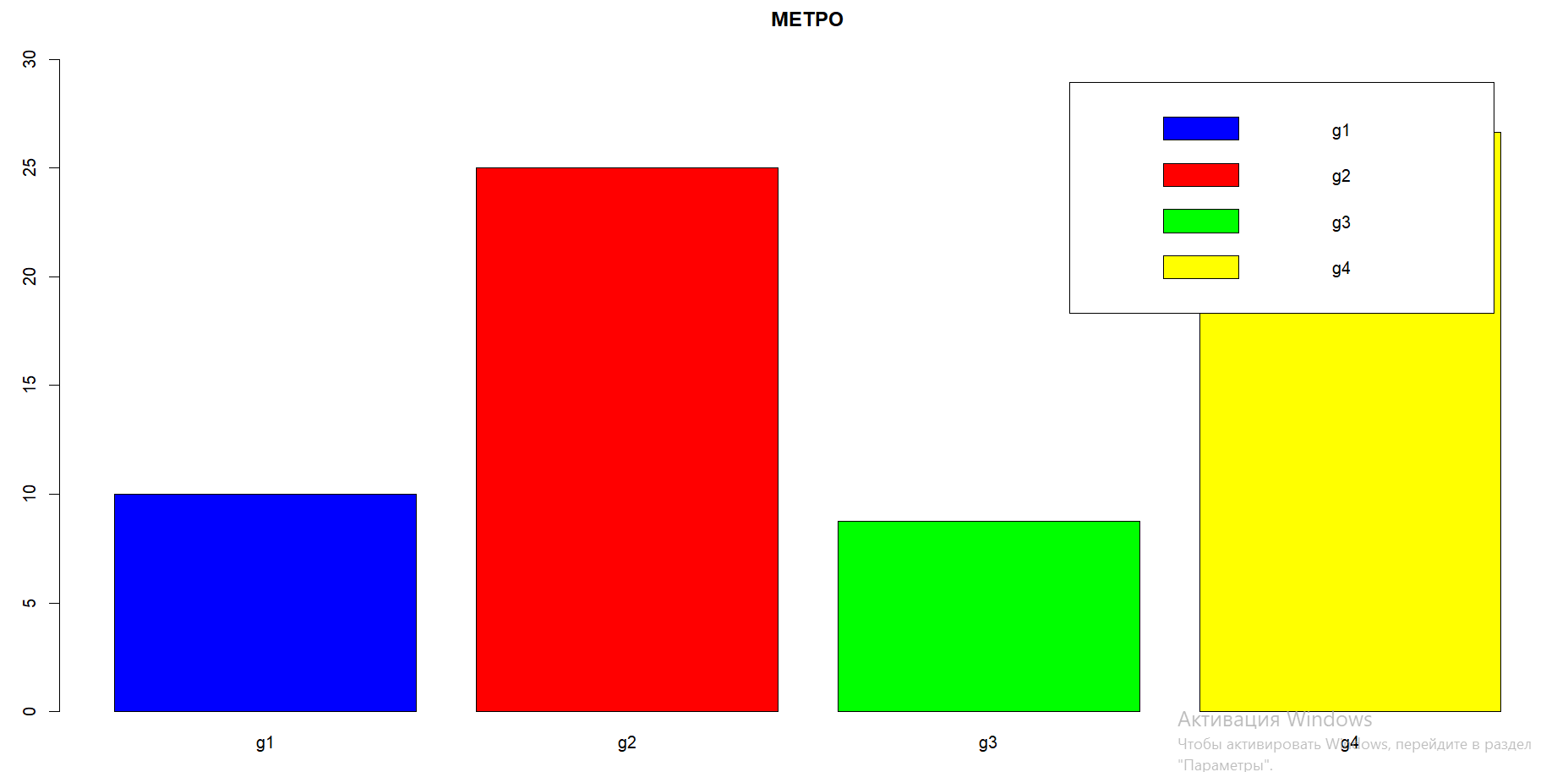


Рисунок 8 – Распределение характеристики МЕТРО

Анализ данной характеристики показывает, что филиалы помимо площади отличаются расположением у метро.

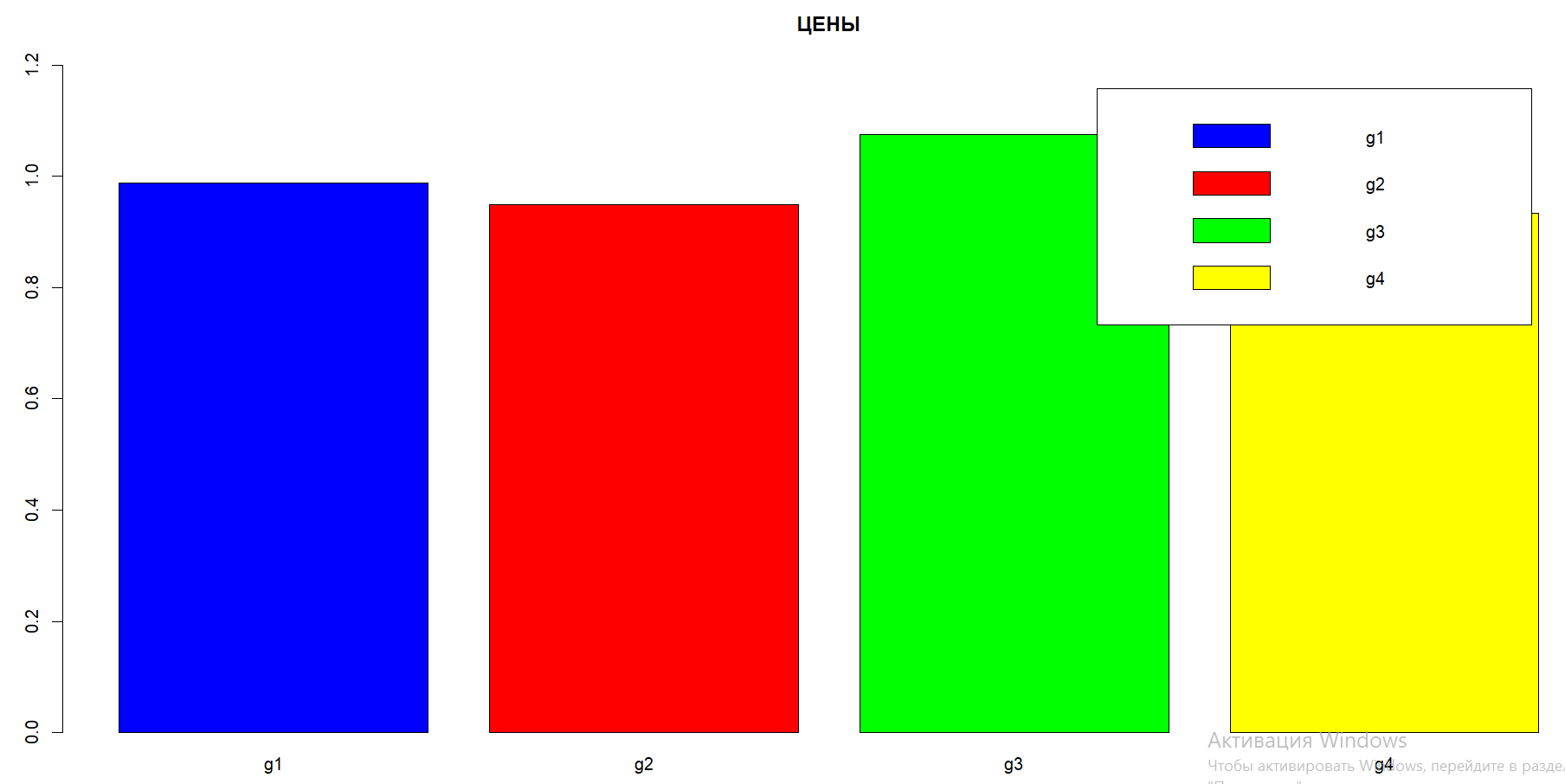


Рисунок 9 – Распределение характеристики ЦЕНЫ

Характеристика ЦЕНЫ показывает, что в филиалах все цены примерно на одном уровне.

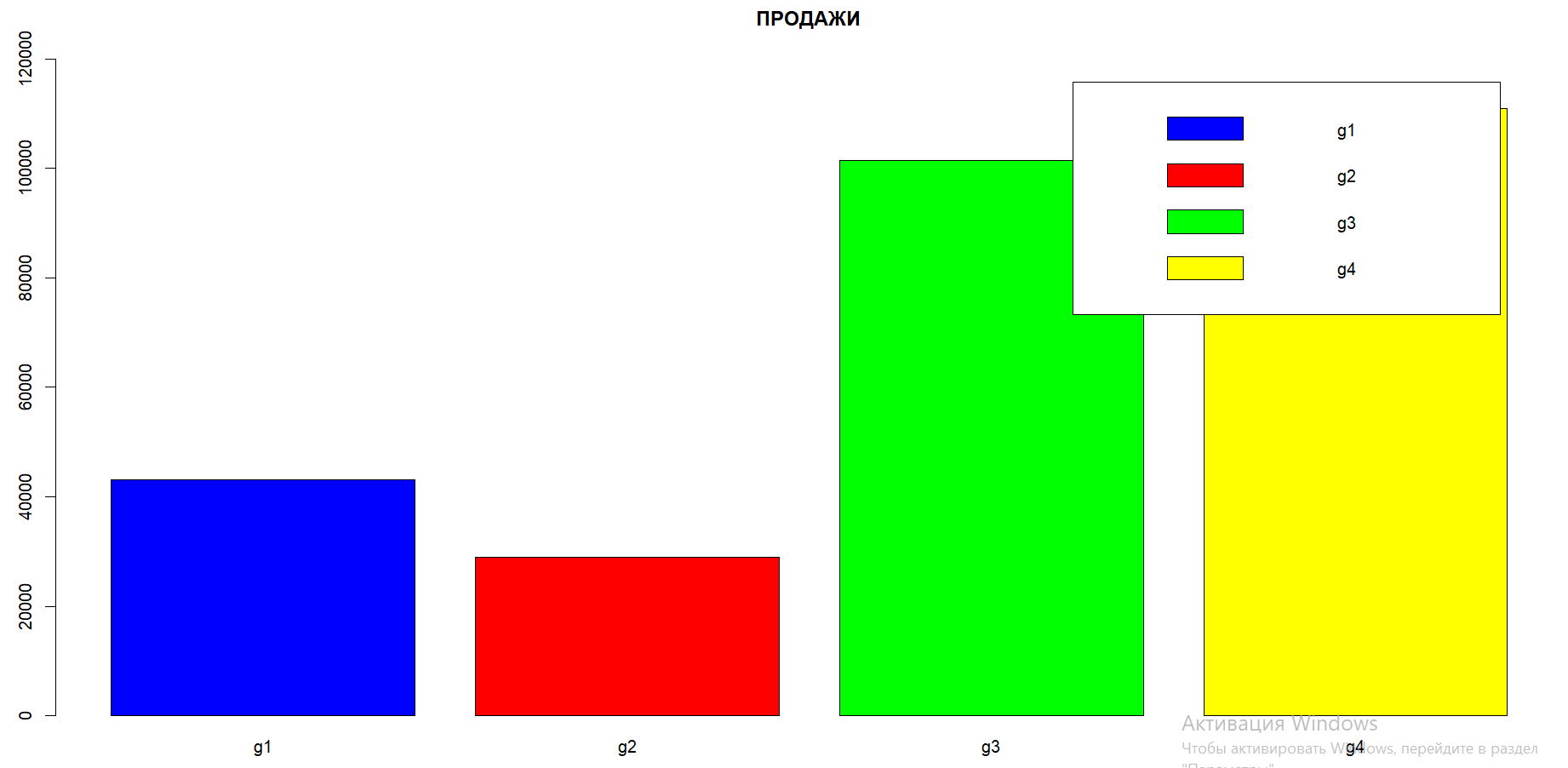


Рисунок 10 – Распределение характеристики ПРОДАЖИ

Анализ продаж в филиалах показывает, что они делятся на 2 типа: до 45000 и выше 45000.

Следующим заданием было построение диаграмма “Каменная осыпь”. Она представлена на рисунке 11.

Изображение выглядит как диаграмма, линия, зарисовка, График

Автоматически созданное описаниеРисунок 11 – Диаграмма “Каменная осыпь”

После этого была работа с построением scatterplot, используя ggplot2.

На рисунке 12 изображена диаграмма рассеивания, выражающая зависимость между продажами и проходимостью.

Изображение выглядит как снимок экрана, диаграмма, текст, линия

Автоматически созданное описание

Рисунок 12 – Диаграмма рассеивания для ПРОХОДИМОСТИ и ПРОДАЖ

Заметна тенденция: чем выше проходимость, тем выше продажи.

После были построены «ящики с усами», показанные на рисунке 13.

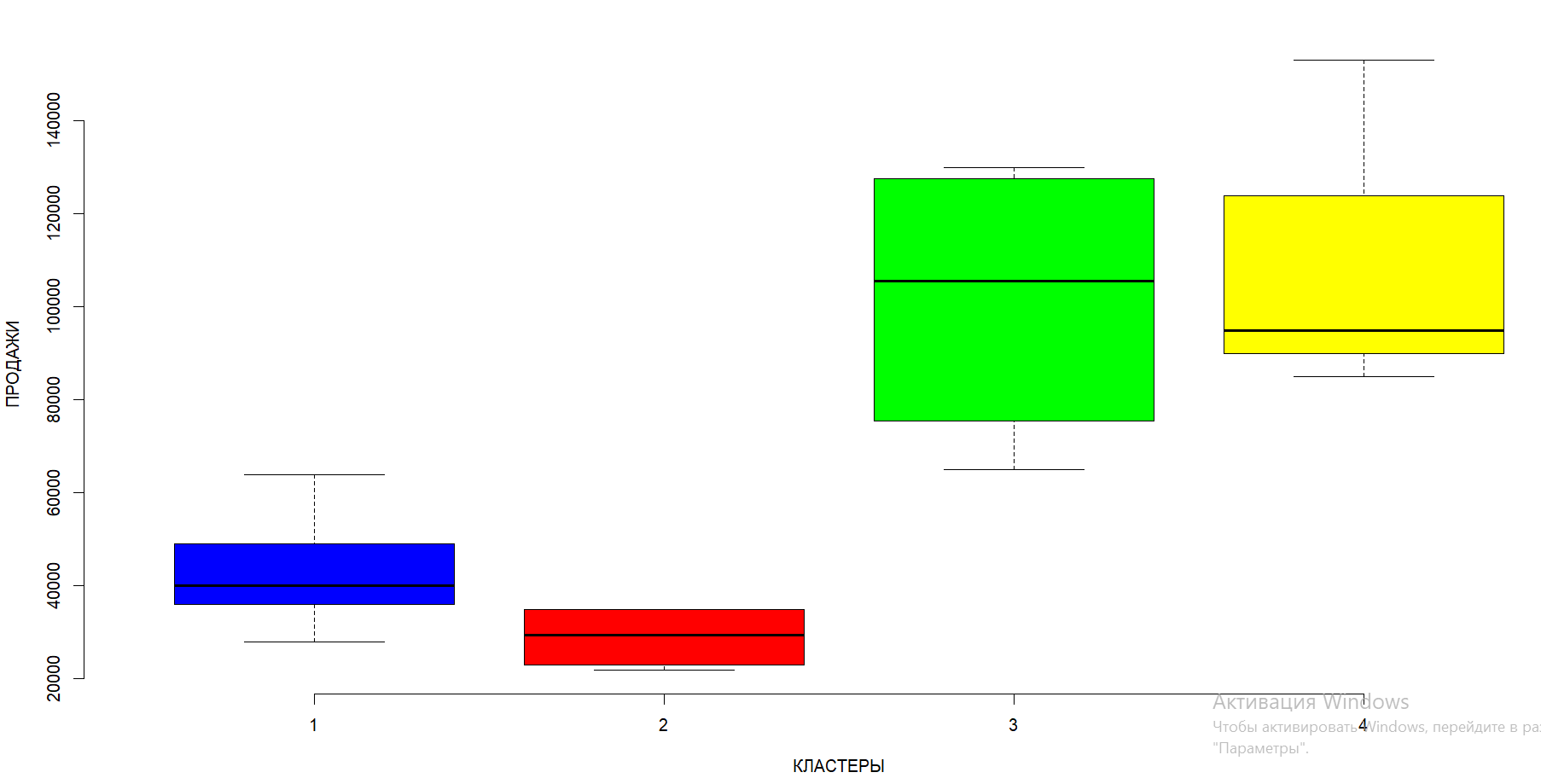


Рисунок 13 – Полученный Boxplot

На нем мы можем видеть, что в кластере №4 присутствуют высокие продажи, но не на постоянной основе. В кластере №2 выбросы отсутствуют, а в кластере №3 размах продаж невысок.

В завершении части 1 лабораторной работы был построен трехмерный график, классифицирующий филиалы по следующим характеристикам: продажи, метро и цены. Он изображён на рисунке 14.

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 14 – Трехмерный график

**Часть 2.** **Байесовская классификация и деревья принятия решений на R**

Сначала строятся функции плотности для характеристик на рисунке 15. Распределение на графике подтверждает правильность распределения на кластеры.

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 15 – Функции плотности

Далее проведем классификацию Decision Tree. Результаты можно увидеть на рисунке 17 и 18.

Изображение выглядит как текст

Автоматически созданное описание

Рисунок 17 – Распределение филиалов

Изображение выглядит как текст

Автоматически созданное описание

Рисунок 18 – Таблица с классификацией

Изображение выглядит как диаграмма

Автоматически созданное описание

Попробуем использовать более точный алгоритм Random Forest. На рисунке 19 изображена информация о модели, а именно об ошибках.

Изображение выглядит как текст

Автоматически созданное описание

Рисунок 19 – Результат работы алгоритма Random Forest

**Вывод:** В данной работе были получены знания об алгоритмах классификации и кластеризации данных, некоторых функциях языка R, осуществляющих этот вид анализа. Помимо этого, были получены навыки о визуализации результатов работы функций кластерного анализа и классификаторов, интерпретирования полученных результатов, а также выполнена классификацию на основе формулы Байеса и деревьев решений.