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**

**ЛАБОРАТОРНАЯ РАБОТА №3**

**Дисциплина: Обработка больших данных**

Работу выполнил: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ К. А. Корнилов

Направление подготовки: 02.03.02 Фундаментальная информатика и информационные технологии

Преподаватель: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_В. И. Шиян

**Тема:** Способы подготовки и отображения данных в R.

**Цель**: Изучить понятие нечеткой классификации предложений.

**Задание:**

1. Выполнить дескриптивный анализ данных из ЛР №2.
2. Выполнить сортировку наборов данных по выбранному признаку.
3. Сформировать отдельные наборы данных по одинаковому признаку (например, составить subdataset, из студентов, отдавших предпочтение по шкале > 0.7 определенной книге),вывести результат, выполнить подсчет размерностей новых таблиц, снова выполнить их анализ – гистограмма, боксплот, серединные меры (см п.2)
4. Результаты пояснить, сделать выводы (каждый график должен иметь подрисуночную подпись – Рис.1. Название графика. И пояснения к графику).

**Ход работы**:

1. Был выполнен соц. опрос с использованием Google Forms и Google Docs по теме: “Самый раздражающие вещи в видеоиграх”. Ссылка на опрос: <https://forms.gle/ZyHRigDZT4kANUvU9>.
2. Были поучены в виде Google table и проанализированы результаты опроса. Данные были загружены в среду RStudio.

Код загрузки данных:

library(openxlsx)

library(magrittr)

ex\_data <- read.xlsx("C:\\Users\\kiril\\Desktop\\Rdata\\lab2.xlsx", 1)

colnames(ex\_data)

colnames(ex\_data) = c( "ФИО", "Bugs", "Hitboxes", "AI", "Optimization", "Interface", "Allies", "Donate", "Cutscenes", "Random", "Levelling", "DLC", "Openworld" )

Результат загрузки данных представлен на рисунке 1:

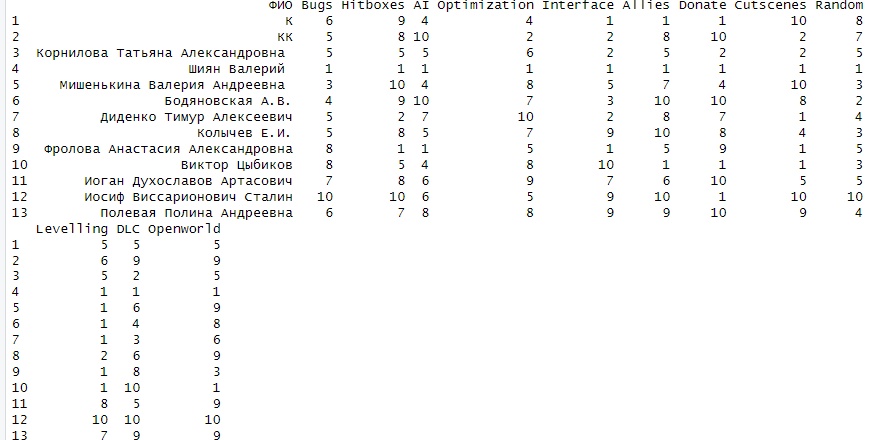


Рисунок 1 – Результат выгрузки данных.

* 1. Для дополнения к раннему выполненному анализу были вычислены квартилей и медиан, дисперсий, средних квадратичных отклонений, межквартильных размахов для каждого из столбцов.

Код для подсчета:

summary(ex\_data)

apply(ex\_data,2,sd,na.rm=TRUE)

apply(ex\_data,2,var,na.rm=TRUE)

apply(ex\_data,2,IQR,na.rm=TRUE)

Результат подсчета представлен на рисунке 2-3:

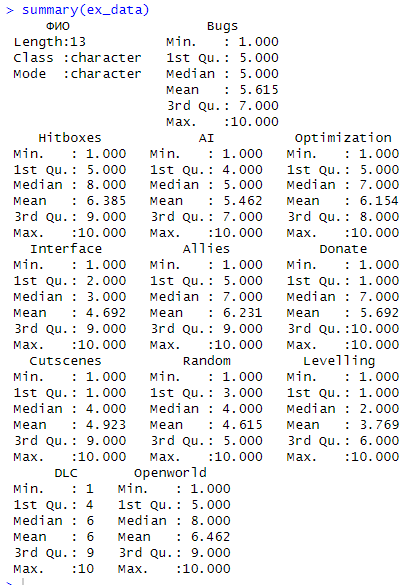


Рисунок 2 – Результат подсчета значений.

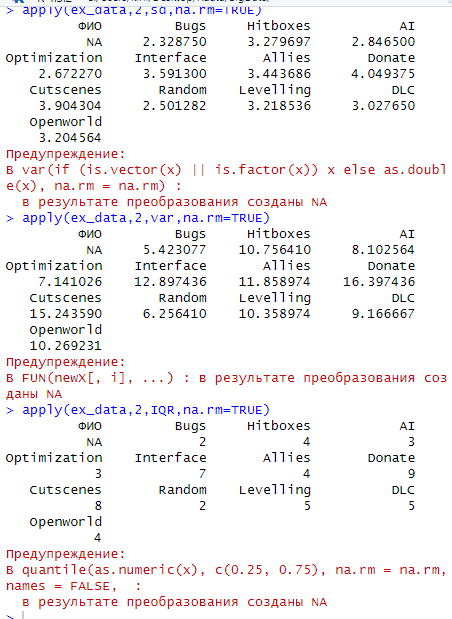


Рисунок 3 – Результат подсчета значений.

Также были построены блоки с усами для каждого из столбцов. Результат представлен на рисунке 4.

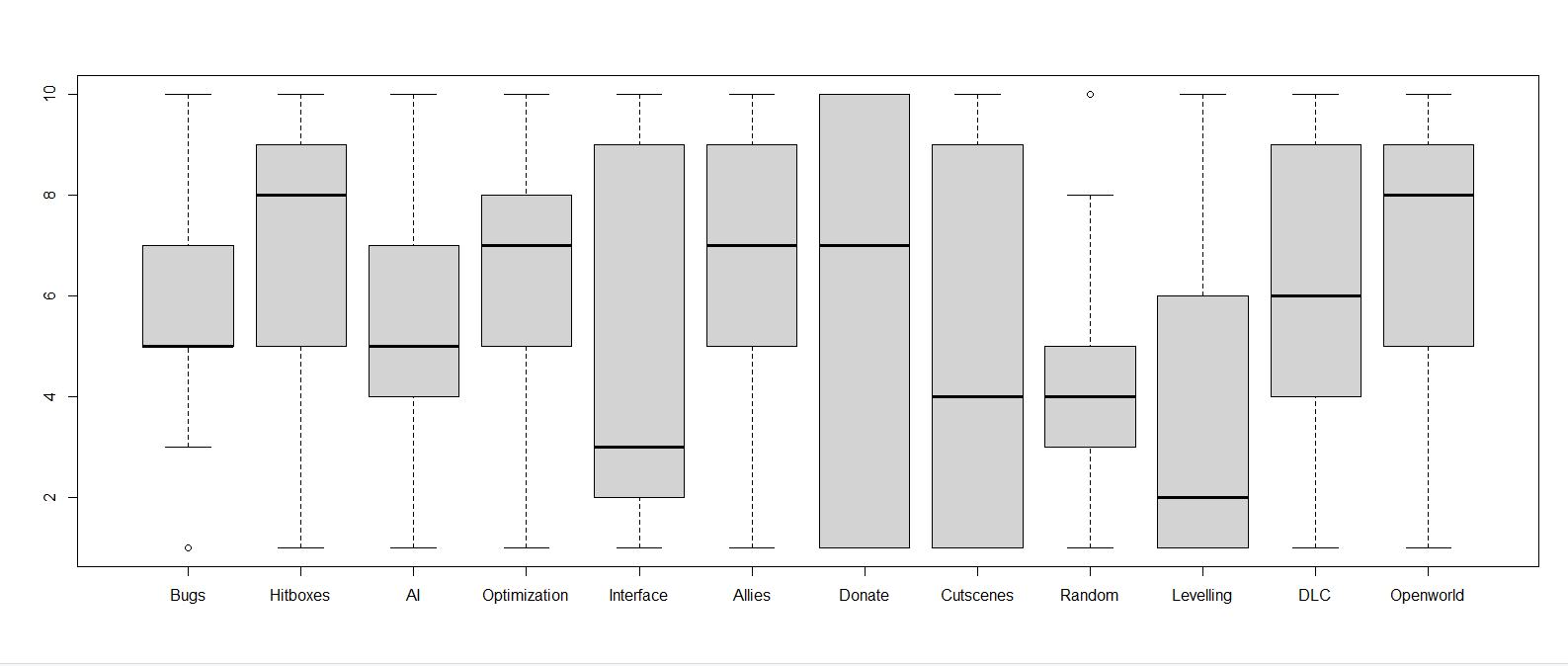


Рисунок 4 – Полученные блоки с усами.

* 1. Была проведена сортировка данных по нескольким признакам.

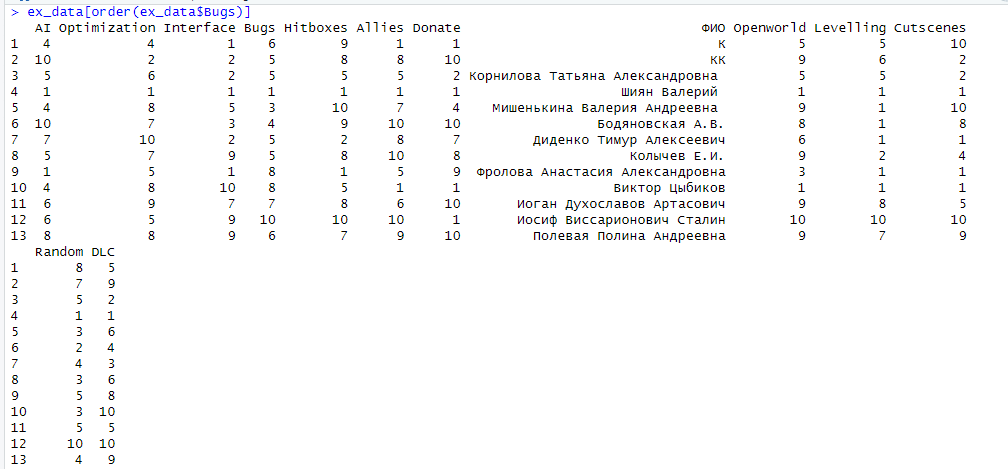
Код для подсчета:

ex\_data[order(ex\_data$Bugs)]

ex\_data[order(ex\_data$AI)]

ex\_data[order(ex\_data$Allies)]

Результат подсчета представлен на рисунке 5-7:





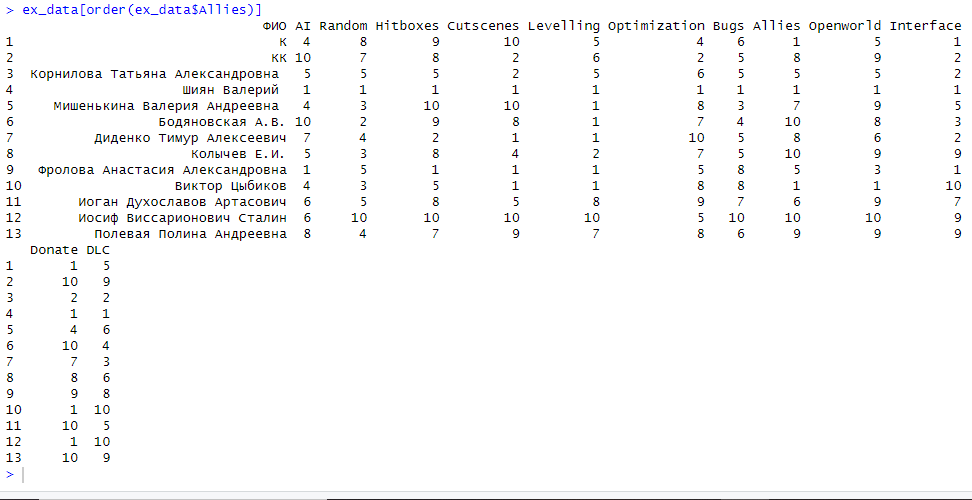


Рисунок 5-7 – Результат подсчета значений.

* 1. Были cформирован отдельные наборы данных по одинаковому признаку.

Код:

new\_data2 = subset(ex\_data,AI>=5 & Bugs<10 | Hitboxes < 9, select = ФИО:Cutscenes)

Результат создания данных представлен на рисунке 8:

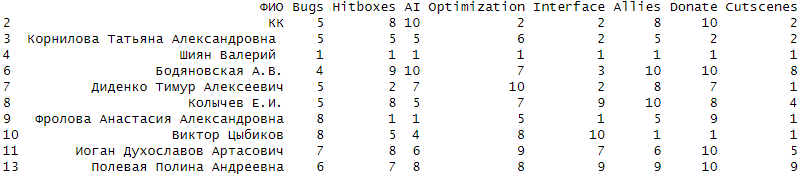


Рисунок 8 – Результат создания данных.

* 1. Были вычислены размерности полученных данных.

Код вычисления размерностей:

print(dim(new\_data2))

Результат построения представлен на рисунке 9:



Рисунок 9 – Результат вычисления размерностей.

* 1. Были подсчитаны серединные меры таблиц (максимум, минимум, медиана, среднее, 1 и 3 квантили, среднее квадратичное отклонение, дисперсия, размах квантилей).

Код:

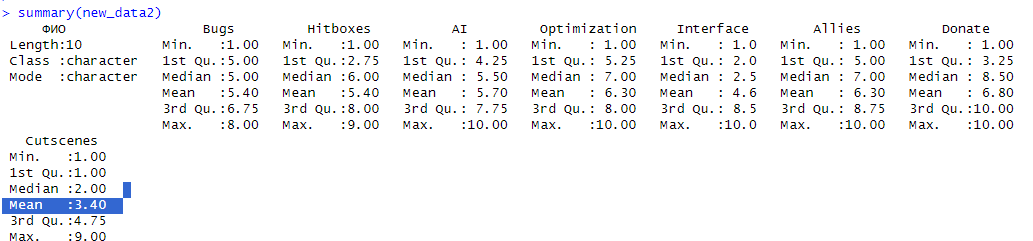
summary(new\_data2)

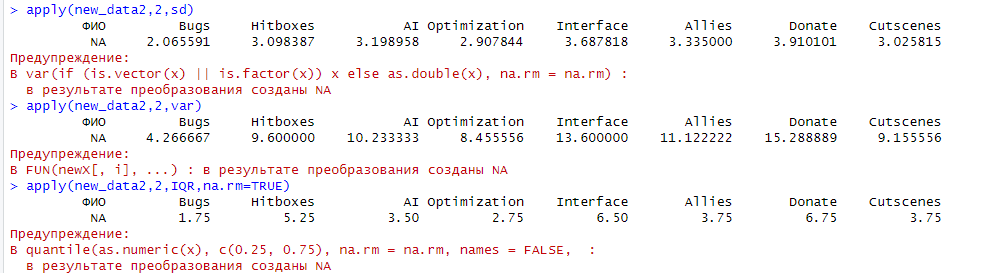
apply(new\_data2,2,sd)

apply(new\_data2,2,var)

apply(new\_data2,2,IQR,na.rm=TRUE)

Результаты подсчетов представлены на рисунке 10-11:





Рисунки 10-11 – Результаты подсчетов серединных мер для новых данных.

* 1. Были построены гистограмма и блоки с усами для полученных данных.

Код построения графиков:

barplot(apply(new\_data2[,c(2:length(colnames(new\_data2)))],2,mean),xlab="Объекты оценки",cex.names=0.6,ylab="Суммарная оценка")

boxplot(new\_data2[,c(2:length(colnames(new\_data2)))])

Результаты построения графиков представлены на рисунках 12-15:

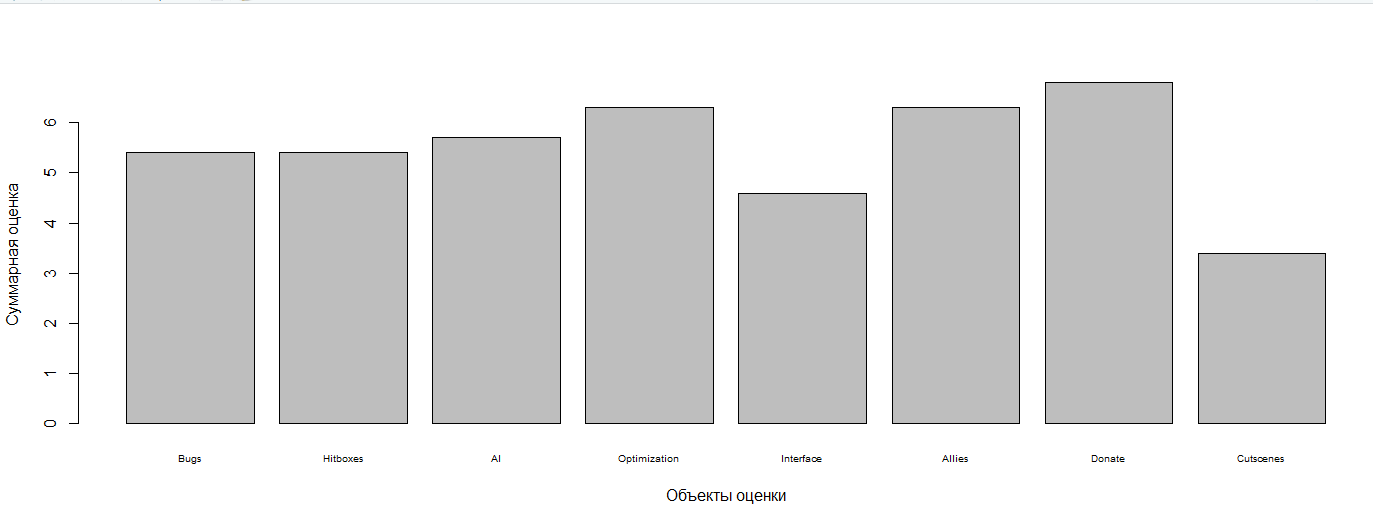


Рисунок 12 – Гистограмма средних оценок набора данных.

Как видно по графику 12 фильтрация данных по нескольким столбцам не поменяла сильно общую картину среднего значения оценок. Среднее значение в одних столбцах повысилось, в других понизилось. Но при этом данные изменения не слишком значительны.

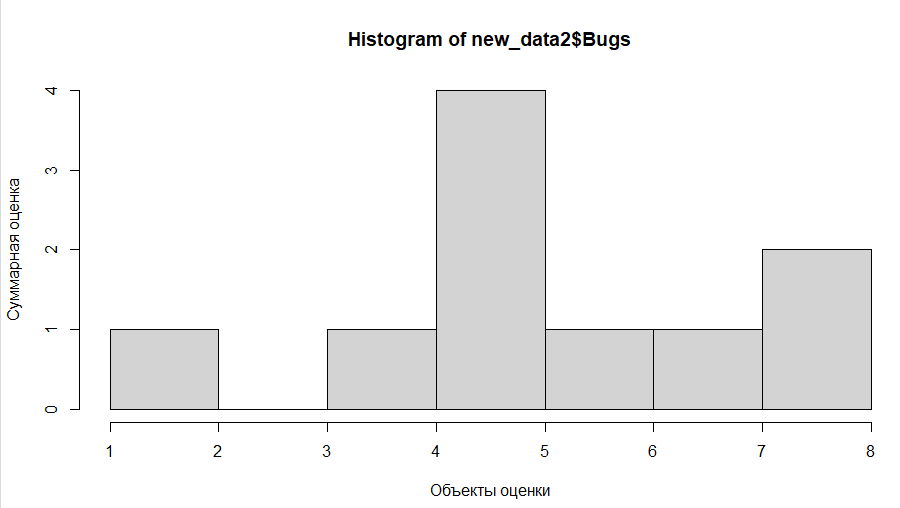


Рисунок 13 – Гистограмма распределения оценок столбца Баги.

Распределение оценок в столбце Баги поменялось после фильтрации. Исчезли значения больше 9 и остальных значений стало меньше.

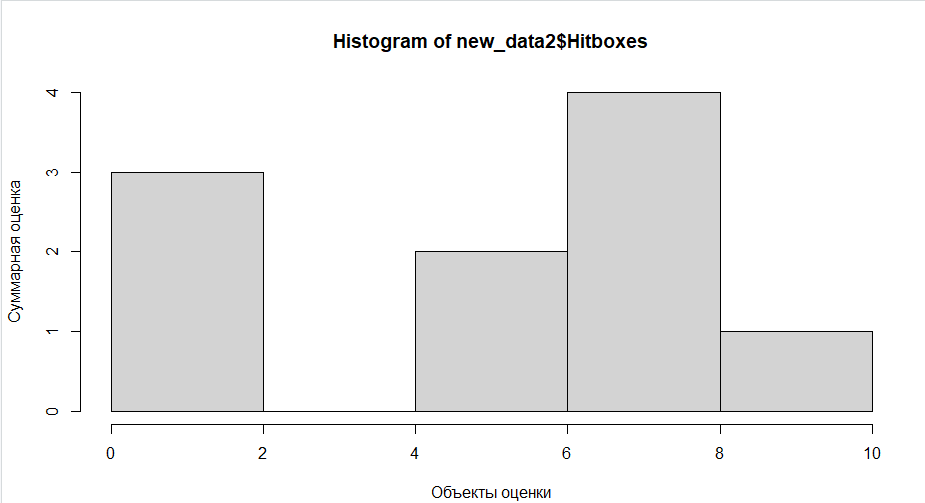


Рисунок 14 – Гистограмма распределения оценок столбца Хитбоксы.

Распределение оценок в столбце Хитбоксы поменялось после фильтрации. Значений больше 8 и от 4 до 6 заметно стало меньше.

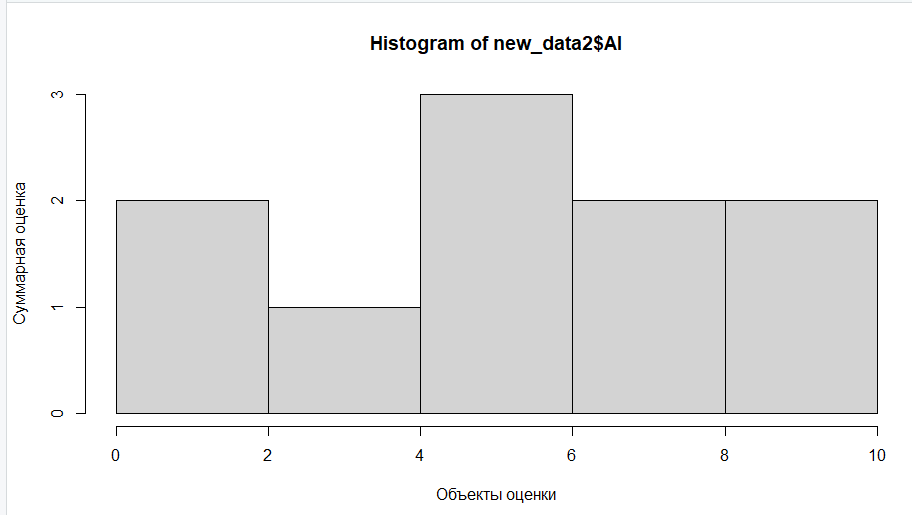


Рисунок 15 – Гистограмма распределения оценок столбца ИИ.

Распределение оценок в столбце ИИ поменялось после фильтрации. Стало заметно меньше оценок от 2 до 4. Остальных оценок стало чуть меньше.

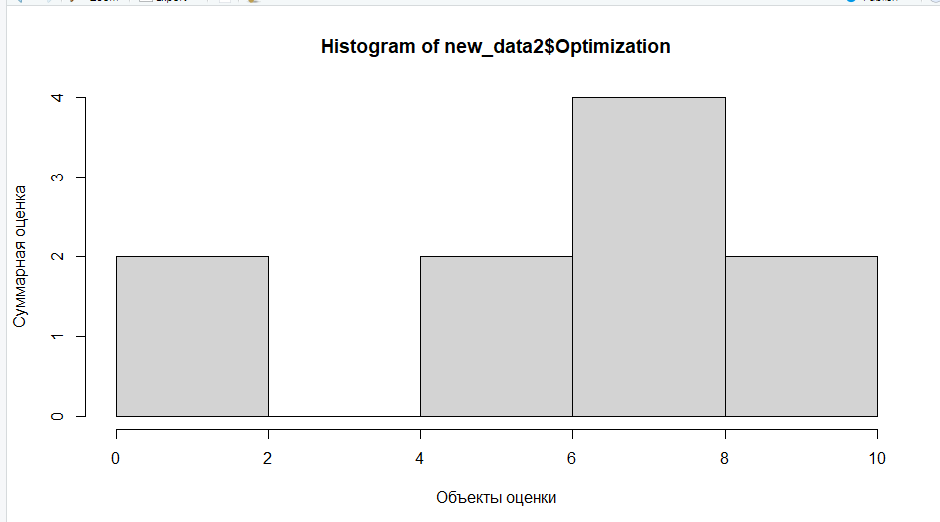


Рисунок 16 – Гистограмма распределения оценок столбца оптимизация.

Распределение оценок в столбце Оптимизация поменялось после фильтрации. Исчезли оценки от 2 до 4. Остальных оценок стало чуть меньше.

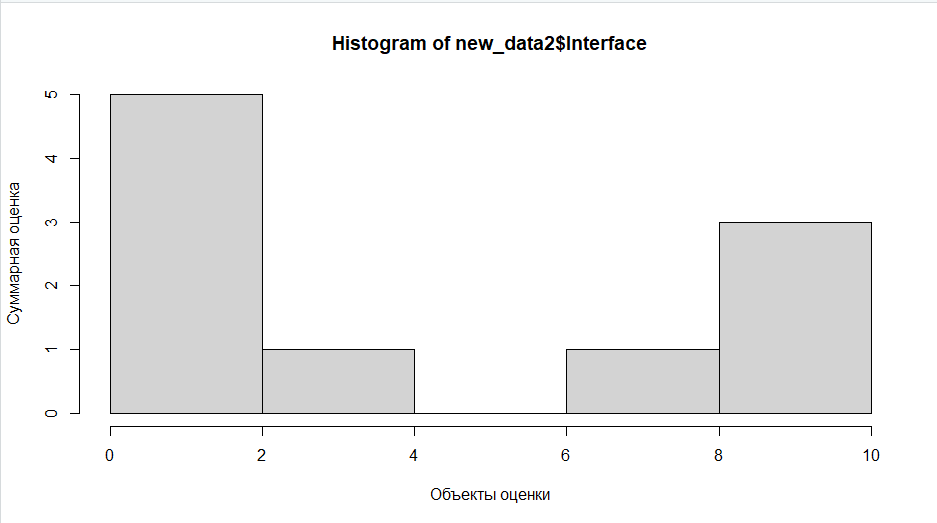


Рисунок 17 – Гистограмма распределения оценок столбца интерфейс.

Распределение оценок в столбце Интерфейс поменялось после фильтрации. Исчезли оценки от 4 до 6. Остальных оценок стало чуть меньше.

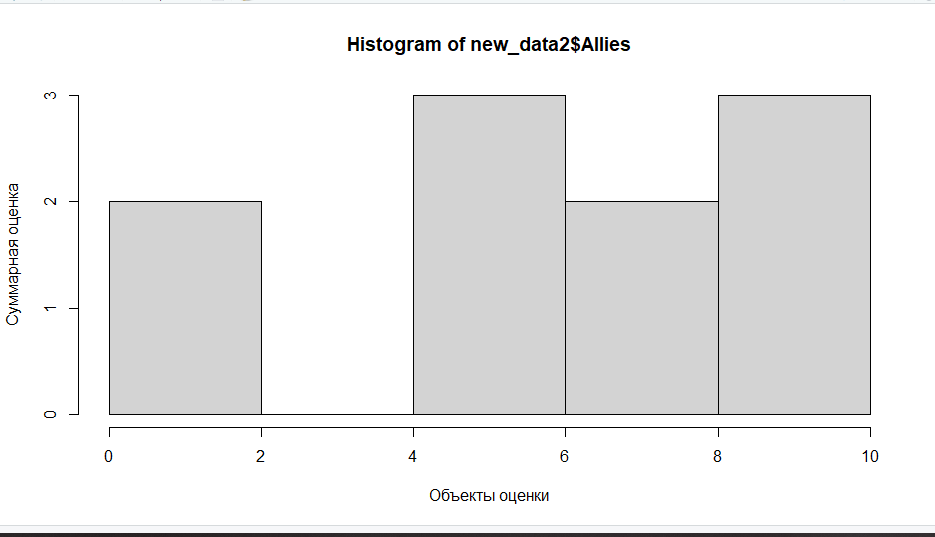


Рисунок 18 – Гистограмма распределения оценок столбца союзники.

Распределение оценок в столбце Союзники поменялось после фильтрации. Оценок от 0 до 2 и от 6 до 8 стало чуть меньше.

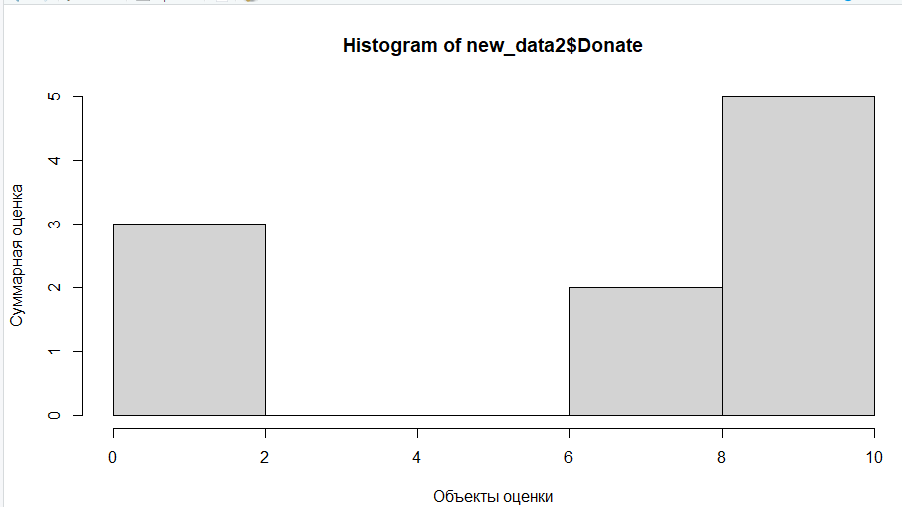


Рисунок 19 – Гистограмма распределения оценок столбца донат.

Распределение оценок в столбце Донат поменялось после фильтрации. Исчезли оценки от 2 до 6. Меньше стало оценок от 0 до 2.

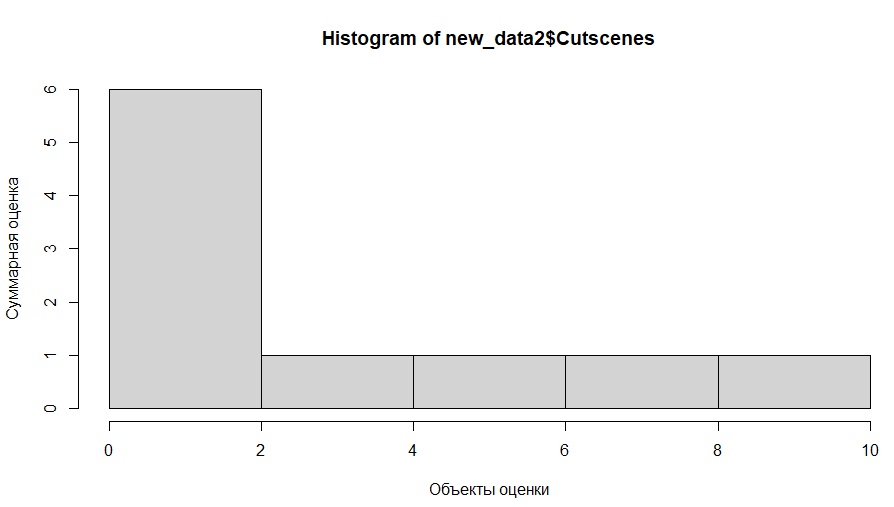


Рисунок 20 – Гистограмма распределения оценок столбца катсцены.

Распределение оценок в столбце ИИ поменялось после фильтрации. Стало заметно меньше оценок от 8 до 10. Остальные оценки не изменились.

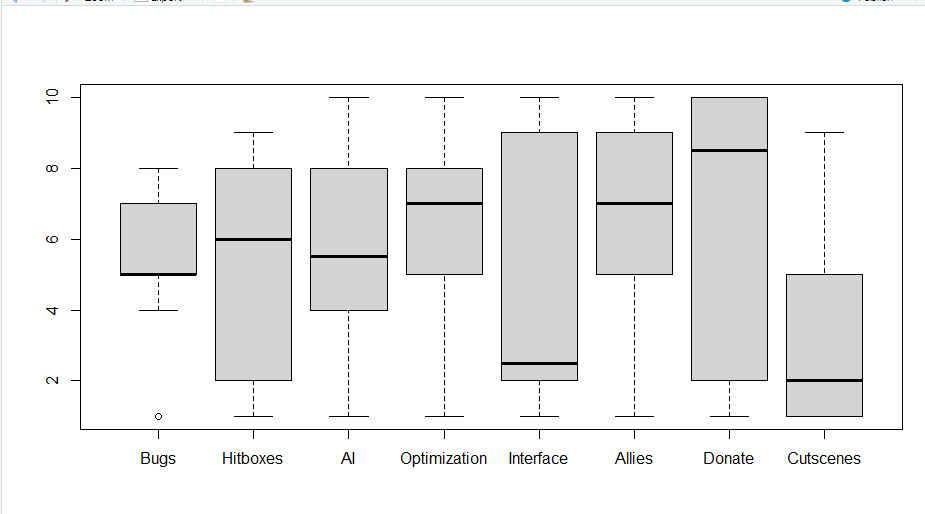


Рисунок 20 – Коробка с усами для набора данных.

Cогласно ящику с усами видны изменения в границах усов. В целом они сдвинулись ближе к 1 и 3 квантилям. Исчезли выбросы и у некоторых значений стали уже размахи квантилей. Также заметно небольшие изменения в средних значениях величин.

**Вывод:** Были получены начальные навыки анализа данных. С использованием данных навыков был проведен анализ данных на тему “Самый раздражающие вещи в видеоиграх”.